Radiant – интерактивное статистическое веб-приложение

Автор: Владимир Шитиков
https://stok1946.blogspot.com/

Вызывает искреннее недоумение трогательное пристрастие аспирантов и соискателей к такому громоздкому и неповоротливому продукту как *Statistica*. В авторефератах диссертаций непременно указывается, что весь статистический анализ выполнен с использованием этой программы, даже если расчеты сводились к вычислению простых средних. Напомним, однако, что лицензия на использование, скажем, *Statistica Ultimate Academic Bundle* стоит 12 тыс. руб. (150$) и только традиционная вороватость российских юзеров спасает их от финансовых проблем.

Альтернатива же очевидна – использование многочисленных бесплатных статистических программ, лучшие из которых нисколько не уступают по функциональности *Statistica*, а по удобству использования даже превосходят. Я здесь представлю *Radiant*, заслуживающее внимание приложение, использующее веб-интерфейс для проведения статистического анализа, независимое от используемой платформы и операционной системы пользователя (*Windows*, *Mac* или *Linux*), написанное в кодах R и легко интегрируемый в ее среду. Оно не требует никакой дополнительной инсталляции и может быть запущено в любом интернет-браузере (*Explorer*, *Google Chrome*, *FireFox*, *Opera*) путем ввода адресной строки

https://vnijs.shinyapps.io/radiant

Впервые *Radiant* был представлен в феврале 2015 г. его разработчиком Винсентом Найджем (*Vincent Nijs*), профессором *Rady School of Management* в Сан-Диего, Калифорния. Его шутливый портрет мы заимствовали с его страницы в Твитере (https://twitter.com/vrnijs). С той поры статистическое приложение, естественно, постоянно расширялось и усовершенствовалось.

Наше описание *Radiant* мы выполним в семи разделах, первые шесть из которых адресованы статистическим аналитикам и другим пользователям, не знакомым с программированием, а седьмой – активным разработчикам скриптов R, желающим добраться до самой сути и модифицировать представляемый пакет под свои нужды.

1. Охватываемые разделы статистики и система меню *Radiant*

После ввода в адресную строку браузера ссылки на *Radiant* открывается страница с иерархической системой меню, самая верхняя (черная) горизонтальная строка которого определяет 5 основных групп функций статистического анализа. Раздел ***Data*** ("Данные") включает интерфейсы для загрузки, сохранения, просмотра, визуализации, обобщения, преобразования и объединения данных. Он также содержит функциональные возможности для создания тиражируемых отчетов по результатам анализа, выполненного в веб-приложении. Раздел ***Design*** ("План") включает в себя инструменты для проектирования эксперимента, рандомизации и расчета необходимого объема выборки. Меню ***Basics*** ("Основной") содержит функции для расчета вероятностей, моделирования центральной предельной теоремы, сравнения

средних и пропорций, тестирования соответствия, перекрестных связей и корреляции. Раздел ***Model*** ("Модели") включает интерфейсы для создания различных моделей: линейной и логистической регрессии, нейронных сетей, иерархических моделей, анализа решений и имитации. Меню ***Multivariate*** ("Многомерный") включает интерфейсы для различных вариантов факторного и кластерного анализа.

При выборе раздела ***Data*** главного меню появляется второй уровень иерархии, который располагается в горизонтальной строке ниже и состоит из совокупности вкладок (на рис. 1 ***Manage***, ***View***, ***Visualize*** и т.д.).

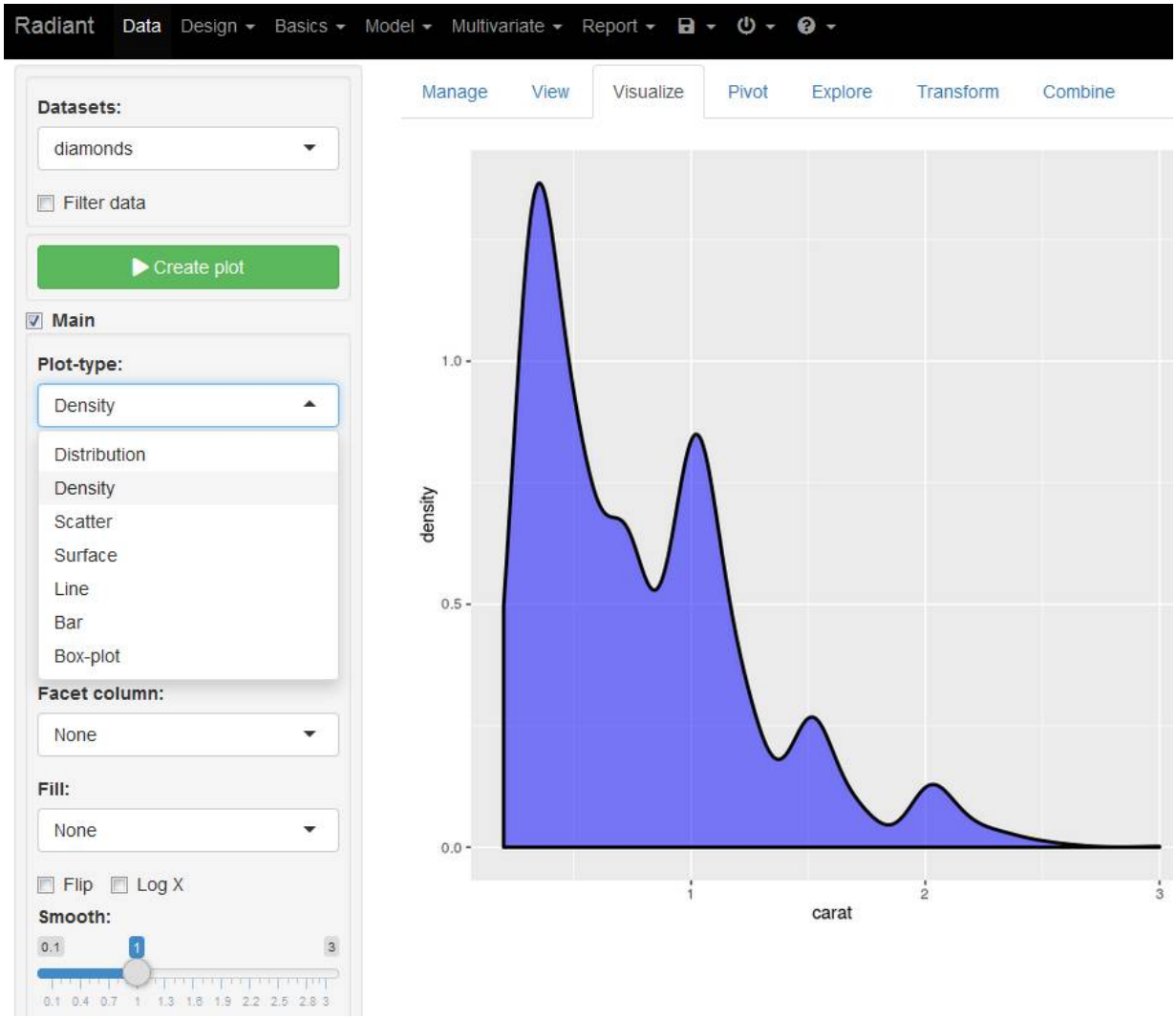


Рис. 1

При выборе конкретной вкладки на боковой панели интерфейса появляется набор графических объектов (или виджетов – численных и текстовых полей, выпадающих списков, флажков, ползунков, радио-боксов и др.) для ввода параметров и условий, задаваемых пользователем. В частности, на рис. 1 рассматривается поле `carat` (т.е. вес алмазов в каратах) базовой таблицы R `diamonts`, которая была выбрана в пункте меню ***Manage***, определен тип графика *Densiny* (плотность вероятностей эмпирического распределения) и его параметр сглаживания (*Smooth*). На главной панели экрана компьютера прорисовывается график с заданными параметрами.

При выборе остальных разделов главного меню (***Design***, ***Basics***, ***Model*** и проч.) появляются выпадающие списки предлагаемых функций и отдельные пункты запускаются непосредственно из них.

Рассмотрим далее работу с основными функциями перечисленных разделов главного меню.

2. Меню *Data* ("Данные")

Перечень вкладок меню изображен на рис. 1 и включает следующие подразделы: ***Manage*** – загрузка данных из различных источников, ***View*** – просмотр таблицы данных, ***Visualize*** – прорисовка различных графиков, ***Pivot*** – создание сводных таблиц для анализа данных, ***Explore*** – обобщение и анализ данных, ***Transform*** – преобразование данных, ***Combine*** – объединение двух наборов данных.

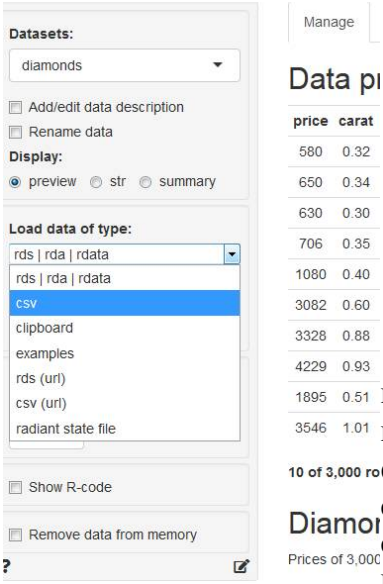


Рис. 2

Manage ("Управление данными") – включает загрузку данных в *Radiant*, сохранение обработанных таблиц на диск и управление состоянием. При первом запуске *Radiant* отображает набор данных `diamonts` с информацией о ценах на алмазы.

Отображать данные на главной панели можно тремя способами: *preview* (первые 10 строк таблицы), *str* (список полей таблицы) и *summary* (статистики для каждой загруженной переменной).

Лучший способ загрузить и сохранить данные для последующего использования в *Radiant* (и R) – это использовать формат *Rdata* (`rds` или `rda`). Это двоичные файлы, которые можно компактно хранить и быстро считывать в R. Выберите *rds* (или *rda*) в раскрывающемся списке *Load data of type* ("Тип загружаемых данных") и нажмите кнопку *Browse* ("Обзор"), чтобы найти файлы, которые вы хотите загрузить на свой компьютер.

Если данные `csv` или `rda` доступны онлайн, то выберите *csv (url)/rda (url)* из выпадающего списка *Load*..., вставьте URL-адрес в показанный текстовый ввод и нажмите кнопку *Load* ("Загрузить"). Можно также получить доступ к различным примерам таблиц данных, используемых в комплекте с *Radiant*: выберите *examples* из раскрывающегося списка *Load*... и нажмите кнопку *Load*. Эти файлы используются для иллюстрации различных данных и инструментов анализа, приводимых ниже.

Загрузить в *Radiant* данные из электронной таблицы (*Excel* или *Google sheets*) можно двумя способами: через файл в формате *csv* или через буфер обмена (*clipboard*):

1. Сохраните данные из электронной таблицы в *csv*-файле (с разделением чисел запятыми, точками с запятой или символами табуляции), а затем в *Radiant* выберите *csv* из раскрывающегося списка *Load*.... Скорее всего, первая строка (заголовок) *csv*-файла будет содержать имена переменных. Чтобы найди *csv*-файл на своем компьютере, нажмите кнопку "Обзор".
2. Выделите блок необходимых данных электронной таблицы и скопируйте их в буфер обмена с помощью CTRL-C, перейдите в *Radiant*, выберите *clipboard* из раскрывающегося списка *Load*... и нажмите кнопку *Paste* ("Вставить"). Этот путь может быть удобен для небольших наборов данных, которые четко отформатированы.

Рекомендуется добавлять описание данных и переменных в каждый используемый файл. Для файлов, которые поставляются в комплекте с *Radiant*, можно увидеть краткий обзор таких описаний. Чтобы добавить описание для ваших собственных данных, установите флажок *Add/edit data description* ("Добавить/изменить описание данных"). Под таблицей откроется окно ввода текста, в которое можно добавить текст в формате `markdown`. После добавления или редактирования описания нажмите кнопку *Update description* ("Обновить описание"). При сохранении данных в виде файла `rds` (или `rda`) созданное (или отредактированное) описание будет автоматически добавлено в файл в качестве атрибута.

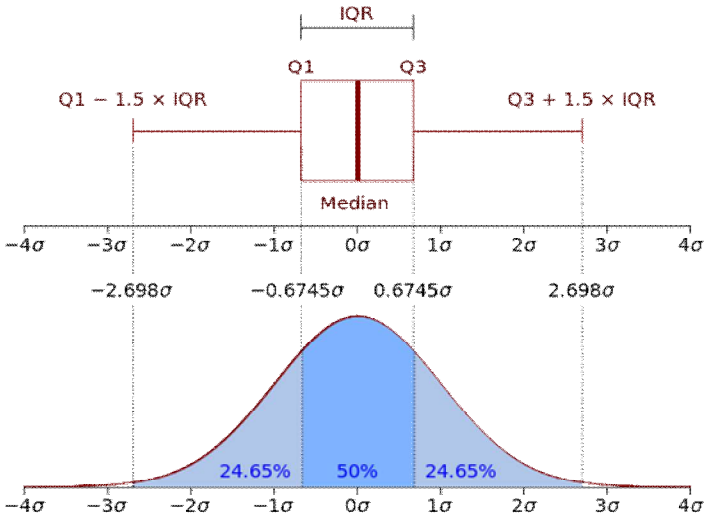
Переместить данные из *Radiant* в произвольную электронную таблицу можно теми же двумя способами, что и при их загрузке: т.е. сохранить данные в формате `csv` и загрузить файл в электронную таблицу, либо скопировать данные в буфер обмена, открыть электронную таблицу и вставить данные с помощью CTRL-V. Во всех перечисленных случаях следует выбрать необходимый пункт из выпадающего списка *Save data to type* ("Тип сохраняемых данных") и нажать кнопку *Save* ("Сохранить").

Если работа не закончена и будет завершена позднее, то удобно работать с файлами состояний (*state of Radiant*). Можно таким образом повторить все ранее выполненные расчеты или осуществить их на другом компьютере. Сохранять и загружать состояние приложения *Radiant* можно точно так же, как и файл данных. Файл состояния (с расширением *.state.rda*) будет содержать: (1) данные, загруженные в *Radiant*, (2) настройки для анализа, с которым выполнялись вычисления, и (3) любые отчеты или код из меню *Report* ("Отчет").

View **(**"Просмотр таблицы данных") позволяет просматривать таблицу данных заданными порциями, выбрать необходимые столбцы для просмотра, осуществить сортировку строк, поиск данных по шаблону и их фильтрацию с помощью системы логических выражений.

Visualize ("Визуализация данных) выполняет формирование графиков из фиксированного набора на основе переменных загруженной таблицы (***Manage***) с учетом установленных фильтров (***View***). Этот пункт требует задания двух ключевых параметров: *Plot-type* ("Тип графика") и *X/Y-variable* ("Список переменных для графика"). Тип графика определяет количество и природу визуализируемых переменных, т.е. (1) одна числовая переменная *X*, (2) одна категориальная переменная, называемая далее фактором, (3) две числовых переменных *X* и *Y*, (4) числовая переменная и фактор. Обратим внимание, что если категориальная переменная (фактор) выбрана в качестве *Y*-переменной, она будет преобразована в числовую переменную, если это необходимо для выполнения выбранной функции. Поскольку среднее значение, стандартное отклонение и т.д. не имеют отношения к недвоичным категориальным переменным, то они будут преобразованы в 0-1 (двоичные) переменные *D*, где один из уровней кодируется как 1, а все остальные уровни – как 0.

Несколько диаграмм оперируют только с одной переменной определенного типа. Например, графики плотности *Densiny* (рис. 1) можно использовать только для числовых переменных. Если выбрать все столбцы таблицы `diamonts` и тип графика *Distribution*, то сформируются гистограммы для каждой численной переменной и столбчатые диаграммы (*Bar plots*) для всех категориальных переменных в наборе данных.



Диаграммы размахов, или "ящики с усами" (*Box-plots*) в своей центральной оси соответствуют медианному значению, а верхняя и нижняя "петли" коробки – первому и третьему квартилям (25-й и 75-й процентили) в данных. Верхний и нижний усы ограничивают максимальное и минимальное допустимые значения, а их длина составляет 1.5 межквартильного диапазона (IQR). Точки за пределами усов считаются выбросами.

Точечные диаграммы (*Scatter plots*) используются для отображения взаимосвязи между двумя переменными. Для их визуализации нужно выбрать одну или несколько переменных для построения графика по оси *Y* и одну или несколько переменных для построения графика по оси *X*. Если одна из переменных является категориальной (т. е. `factor`), она должна быть определена как *X*-переменная. Линейные графики (*Line plots*) похожи на точечные графики, но они соединяют точки и особенно полезны для визуализации временных рядов. Трехмерные графики или графики поверхности (*Surface plots*) аналогичны тепловым картам и требуют задания 3 входных переменных: *X*, *Y* и *Fill* (ось *z* или заполняющая переменная).

Строка фасета и столбец фасета (*Facet row* / *Facet column*) могут использоваться для разделения данных на несколько различных групп и создания отдельных графиков для каждой группы. Если выбирается точечный или линейный график, то появится раскрывающийся список *Color* ("Цвет") и каждой из групп будет соответствовать линия соответствующего цвета. Выбор *Color variable* (цветовой переменной) задаст тип тепловой карты (*heat-map*), на которой цвета связаны со значениями этой переменной.

Чтобы добавить линейную или нелинейную линию регрессии к точечной диаграмме, нужно установить флажки *Line* ("Линия") или *Loess* ("Сглаживание"). Флажок *Jitter* может быть полезен, чтобы понять, как распределены по шкале переменной сгущения точек данных. Для проверки предположения о нелинейном характере зависимостей можно воспользоваться флажками логарифмического преобразования *LogX* и/или *LogY* (например, для точечной или столбчатой диаграммы).

Поменять местами переменные по осям X и Y можно с помощью флажка *Flip*. Чтобы сделать графики больше или меньше, следует отрегулировать значения в полях высоты *height* и ширины *width* в левом нижнем углу экрана. Для сохранения графика на диске в виде *png*-файла, нужно нажать на значок ↓ в правом верхнем углу экрана.

Pivot – создание сводных таблиц для анализа данных по заданной комбинации категориальных переменных. Например, если после загрузки данных `diamonds` выбрать пункт `clarity` и `cut` в раскрывающемся списке *Categorical variables* и `price` из списка *Numeric variables*, то после нажатия кнопки *Create pivot table* ("Создать сводную таблицу") отображается таблица средней стоимости алмазов с различными уровнями четкости и качества огранки.

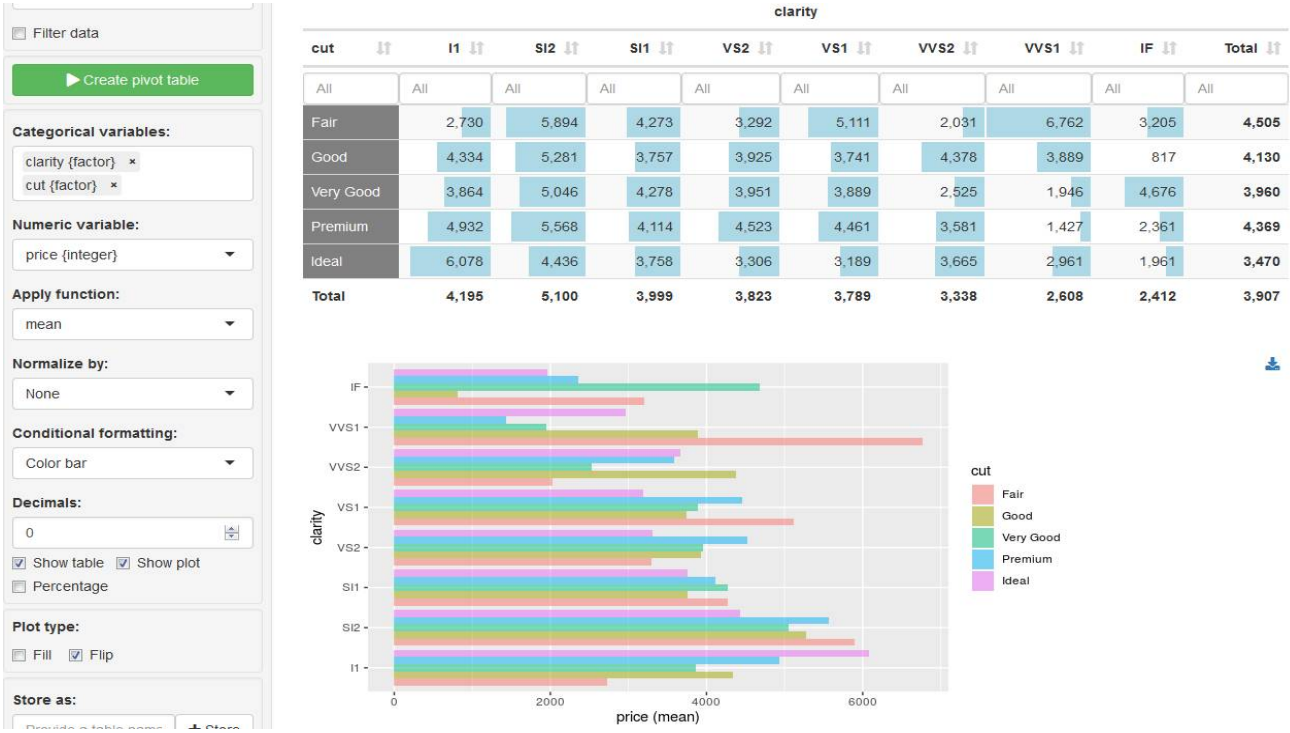


Рис. 3

Можно отсортировать таблицу, щелкнув по заголовкам столбцов, показать только выбранные категории или ограниченные диапазоны данных с использованием ползунков. Можно также выбрать из раскрывающегося списка *Conditional formatting* ("Условное форматирование") опции *Choose Color bar* ("Цветные полосы") или *Heat map* ("Тепловая карта"), чтобы подчеркнуть самые высокие значения итога.

Кроме средних (*means*) для подсчета итога можно использовать самые различные функции из *Apply function*:

- *n* – вычисляет количество наблюдений в данных или в группе, если выбрана переменная *Group by*;
- *n_distinct* – вычисляет количество различающихся значений;
- *n_missing* – вычисляет количество пропущенных значений;
- *sd* и *var* – вычисляют выборочное стандартное отклонение и дисперсию для числовых данных;
- *cv* – это коэффициент вариации, т. е. `mean(x)/sd(x)` ;
- *me* – вычисляет погрешность для числовой переменной с использованием 95%-ного доверительного уровня;
- *prop* – вычисляет пропорцию (для переменной, имеющей только значения 0 или 1, она эквивалентно среднему значению; для других числовых переменных она фиксируется по отношению к максимальному значению; для фактора она фиксируется по отношению к первому уровню) ;
- *sdprop* и *varprop* – вычисляют выборочное стандартное отклонение и дисперсию для пропорции;
- *meprop* – вычисляет погрешность для пропорции, используя 95% - ный доверительный уровень;
- *sdpop* и *varpop* – вычисляют стандартное отклонение и дисперсию популяции,

Если *Numeric variables* не задана, то подсчитываются частоты. Их можно нормализовать по маргинальным суммам строки, столбца или общему итогу из списка *Normalize by*. Если выбрана опция нормализации, то удобно установить флажок *Percentage*, чтобы выразить частоты в процентах.

Созданная сводная таблица может быть сохранена для дальнейшего использования.

Explore – обобщение и анализ данных. Этот раздел очень напоминает описанное выше формирование сводных таблиц. Однако, если вкладка ***Data*** > ***Pivot*** лучше подходит для частотных таблиц и анализа одной числовой переменной, то более мощные функции на вкладке ***Data*** > ***Explore*** позволяют обобщить несколько переменных одновременно, используя различные статистические функции, представленные выше.

Transform - преобразование данных. Два главных выпадающих списка на боковой панели задают процессы трансформации: какие переменные участвуют в преобразовании (*Select variables*) и что с ними нужно сделать (*Transformation type*). Кнопка *Store* ("Сохранить") позволяет зафиксировать выполненные операции.

A) Опции раскрывающегося списка *Transformation type* по изменению переменных.

При выборе *Type* в *Transformation type* отображается другое раскрывающееся меню, которое позволяет изменить тип (или класс) одной или нескольких переменных. Например, можно изменить переменную типа *integer* на переменную типа *factor* или наоборот. При преобразовании переменной в тип `ts` (т. е. временной ряд) необходимо, по крайней мере, указать начальный период и частоту наблюдений. В частности, для еженедельных данных, которые начинаются с 4-й недели года, нужно ввести в качестве *Start period* = 4 и *Frequency* = 52.

Выбор опции *Normalize* позволяет нормализовать одну или несколько переменных по отношению к другой. Например, в данных по алмазам можно выбрать `carat` как

Normalizing variable и `price` как *Select variable(s)* и получить новую переменную (например, `price_carat`), выражающую цену алмаза за один карат. На главной панели для нее тут же отображается сводная статистика. Примечание: для больших наборов данных или когда итоговые статистики не нужны, полезно нажать кнопку *Hide summaries* ("Скрыть сводки") для новой переменной.

При выборе пункта *Transform* в *Transformation type* отображается другое раскрывающееся меню, которое позволяет выполнить общие функциональные преобразования одной или нескольких переменных. Преобразованная переменная будет иметь соответствующее расширение к имени переменной (например, *_ln*). Ниже приводится описание функций преобразования, включенных в *Radiant*.

- *Ln*: создается логарифмически преобразованная версия выбранной переменной, например, `log(x)` или `ln(x)`.
- *Square*: переменная умножается на саму себя, т. е. x^2.
- *Square-root*: извлекается квадратный корень из переменной, т. е. $x^{0.5}$.
- *Absolute*: абсолютное значение переменной, т. е. `abs(x)` .
- *Center*: новая переменная со средним значением, равным нулю, т.е. `x-mean(x)`.
- *Standardize*: новая переменная со нулевым средним значением и единичным стандартным отклонением, т. е. `(x - mean(x))/sd(x)` .
- *Inverse*: 1/x.

Опция *Bin* выполняет дискретизацию числовой переменной в категориальные ячейки одинаковой длины. Если задать число ячеек *Nr bins* = 5, то создастся новая переменная – вектор порядковых номеров квинтилей, соответствующих исходной выборке. Возможен и обратный порядок дискретизации (опция *reverse*).

Можно использовать функцию перекодирования (*Recode*), чтобы, например, произвольным образом выполнить дискретизацию данных и создать факторную переменную, либо определить предполагаемые выбросы как NA. Для этого вводятся одна или несколько команд перекодирования, разделенных символом "`;`", таких как

`1:12 = 'A'; 13:24 = 'B'; else = 'C'` или `400 = NA`

Если в поле *Select variable(s)* выбрана единственная переменная типа `factor`, то можно выполнить операцию *Reorder/Remove levels* ("Переупорядочить или удалить уровни"). Вы можете перетаскивать уровни, чтобы изменить их порядок, или нажать кнопку × , чтобы удалить их. Обратите внимание, что по умолчанию удаление одного или нескольких уровней приведет к появлению пропущенных значений в данных. Поэтому лучше перекодировать удаляемые уровни в новый уровень.

Выберите пункт *Rename* (**"**Переименовать**"**) для задания новых имен одной или нескольких выбранных переменных.

Б) Опции *Transformation type* по созданию новых переменных.

Пункт *Create* из раскрывающегося списка *Transformation type* в принципе не отличается от описанных выше *Transform, Normalize* или *Recode* (во всех этих случаях создается новая переменная). Но это наиболее гибкая команда, требующая, однако, некоторых базовых знаний R-синтаксиса. Новая переменная может быть любой функцией R или выражением с другими переменными из (активного) набора данных. Некоторые примеры приведены ниже, в которых новая переменная с именем слева от знака равенства = создается на основе различных выражений:

```
z = x - mean(x) #  Центрированная переменная
z = ifelse(x < y, 'smaller', 'bigger')  #  Фактор с двумя уровнями
tdiff = as_duration(time2 - time1) # Разность двух дат в сек.
d = as_distance(lat1, long1, lat2, long2) # Расстояние между двумя точками
```

Выше упоминалась возможность обмена данными с любой программой обработки электронных таблиц через буфер обмена во вкладках ***Data > Manage*** или ***Data > View***. В разделе ***Data > Transform*** также можно выделить один или несколько столбцов,

например, на листе *Excel* и скопировать новые переменные с их заголовками в буфер обмена (CTRL-C). После этого нужно выбрать пункт *Clipboard* (Буфер обмена) в *Transformation type* и вставить новые данные в поле *Paste*. Важно, чтобы новые переменные имели такое же количество наблюдений, как и данные в *Radiant*.

Примечание: Все операции, примененные на вкладке ***Data > Transform***, записываются в журнал команд преобразования (*Transform command log*). Это означает, что выполненная работа воспроизводима и может быть повторена с теми же или с новыми, но похожими данными. Однако операции с буфером обмена не являются воспроизводимыми.

В) Опции *Transformation type* по очистке данных.

Если выбрать в раскрывающемся списке типа преобразования опцию *Remove missing*, то можно удалить строки с одним или несколькими отсутствующими значениями для выбранных переменных. Если имелись отсутствующие значения, то можно наблюдать, как количество наблюдений *n* в сводке данных изменится. Нажмите кнопку "Сохранить", чтобы сохранить измененные данные.

Часто в наборе данных есть одна или несколько переменных, которые должны иметь только уникальные значения (например, порядковый номер наблюдения). Чтобы проверить, есть ли в таких данных дубликаты, выберите одну или несколько переменных для оценки их уникальности и используйте *Show duplicates*. Чтобы удалить дубликаты, выберите *Remove duplicates* и, если есть повторяющиеся строки, то количество наблюдений *n* и *n_distinct* в сводке данных изменится.

Г) Опции *Transformation type* по переформированию таблицы данных.

Чтобы удалить ненужные переменных или изменить порядок их следования, выберите *Reorder or remove variables.* Перетаскивайте переменные, чтобы изменить их порядок в данных, а чтобы удалить переменную, нажмите кнопку × (символ рядом с названием столбца).

Опция *Expand grid* создает новую таблицу со всеми комбинациями значений выбранных переменных. Например, нужно создать набор данных со всеми возможными комбинациями значений огранки `cut` (5 уровней фактора) и цвета алмаза `color` (7 уровней фактора). Выбрав эти переменные в таблице `diamonds`, мы получим таблицу с 35 возможными комбинациями уровней факторов. Если задать имя для нового набора данных (например, `diamonds_expand`) и сохранить, он появится везде в раскрывающемся списке *Datasets***.**

Используя опцию *Gather columns,* можно объединить несколько переменных в один столбец. Например, если загружен набор данных `diamonds` и выбраны `cut` и `color` в поле *Select variable(s)* , то эта функция создаст новые переменные `key` и `value`. `key` , имеющий два уровня (т. е. `cut` и `color`), а `value` захватывает все значения в `cut` и `color`. *Spread column* по смыслу противоположен *Gather columns*.

Пункт *Table-to-data* позволяет сохранить в новом наборе данных таблицы частот.

Опция *Holdout sample* применяется для таблиц с наложенным активным фильтром и создает набор данных с **"**оппозицией**"** (т.е. со всеми строками, не отобранными при фильтрации).

Чтобы создать переменную, которая может использоваться для случайной группировки набора данных при обучении и тестировании модели, выберите опцию *Training variable*. Укажите либо количество наблюдений, используемых для обучения (например, 2000), либо их долю (например, 0.7). Новая переменная будет иметь значение 1 для обучающих и 0 для тестовых данных. Можно также выбрать одну или несколько переменных для регулировки процесса случайного присвоения (например, чтобы достигалась одинаковая доля частот по заданному фактору в обучающей и тестовой выборках).

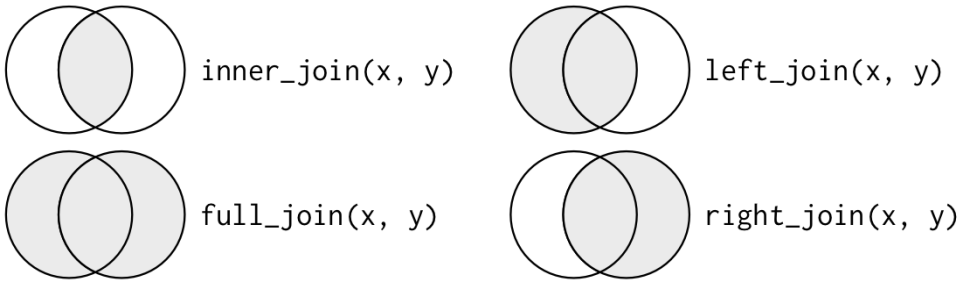
Combine – объединение двух наборов данных. Предположим, что у нас есть две таблицы `Dataset1` и `Dtaset2`, у которых есть одно или несколько полей для связи (т.е. ключи с отчасти совпадающими значениями). Ставится задача объединить по столбцу *Join by* оба набора данных, в частности, к первой таблице добавить переменные из списка *Variables to add* второй таблицы, определенной в поле *Combined dataset*. Механизм объединения задается в выпадающем списке *Combine type.*

Есть несколько вариантов создания связей между ключами реляционных таблиц (см. обзор на нашем блоге https://stok1946.blogspot.com/2019/08/blog-post.html). Существует два типа ключей:

- Первичный (уникальный) ключ однозначно идентифицирует объект в таблице, представляющей список этих объектов. Отсутствие в таблице первичного ключа или (еще хуже) его дублирование – подлинный кошмар администратора базы данных.
- Вторичный (внешний) ключ однозначно ссылается на первичный ключ другой таблицы.

Первичный ключ и соответствующий ему вторичный ключ в другой таблице образуют отношения между таблицами типа "один-ко-многим". В некоторых случаях можно увидеть отношение типа "один-к-одному", хотя оно трактуется как частный случай "один-ко-многим". Можно построить, например, модель реляционных отношений "многие ко многим" плюс "один-ко-многим".

Осуществить выборку записей из связанных таблиц можно четырьмя возможными способами, представленными для записей двух таблиц следующими диаграммами Венна:



В частности, операция `inner_join(x, y)` выводит все записи из обеих связываемых таблиц, у которых совпадают значения ключевых полей. Аналогично `left_join(x, y)` выводит все записи из таблицы `x` и те записи из таблицы `y`, для которых совпадают значения ключевых полей.

Обнаружить и устранить ошибки в объединении таблиц помогают два оператора:

- `anti_join(x, y)` удаляет все наблюдения из `x`, которые имеют совпадение в `y`.
- `semi_join(x, y)` сохраняет все наблюдения в `x`, которые имеют совпадение в `y`.

Все эти варианты представлены в списке *Combine type.*.

3. Меню *Design* ("План")

Выпадающий список меню этого раздела состоит из двух подразделов:

- ***Design > DOE > Design of Experiments*** ("План эксперимента") и
- ***Design > Sample*** ("Операции с выборками").

Второй подраздел состоит в свою очередь из четырех пунктов:

- ***Random sampling*** – Формирование простой случайной выборки заданного объема из полного списка наблюдаемых объектов.
- ***Random assignment --*** случайное назначение объектов наблюдений, распределяя их в заданной пропорции в зависимости от условий эксперимента.

° ***Sample size (single)*** – Определение необходимого размера выборки для тестирования статистической значимости среднего значения или пропорции, рассчитанных на основе выборочных данных.

° ***Sample size (compare)*** – Определение необходимого размера выборки при сравнении средних или пропорций.

Design > Design of Experiments. После выбора этого пункта меню, для того, чтобы построить план эксперимента, достаточно определить уровни варьируемых факторов и их взаимодействия и нажать кнопку *Create design* (рис. 4).

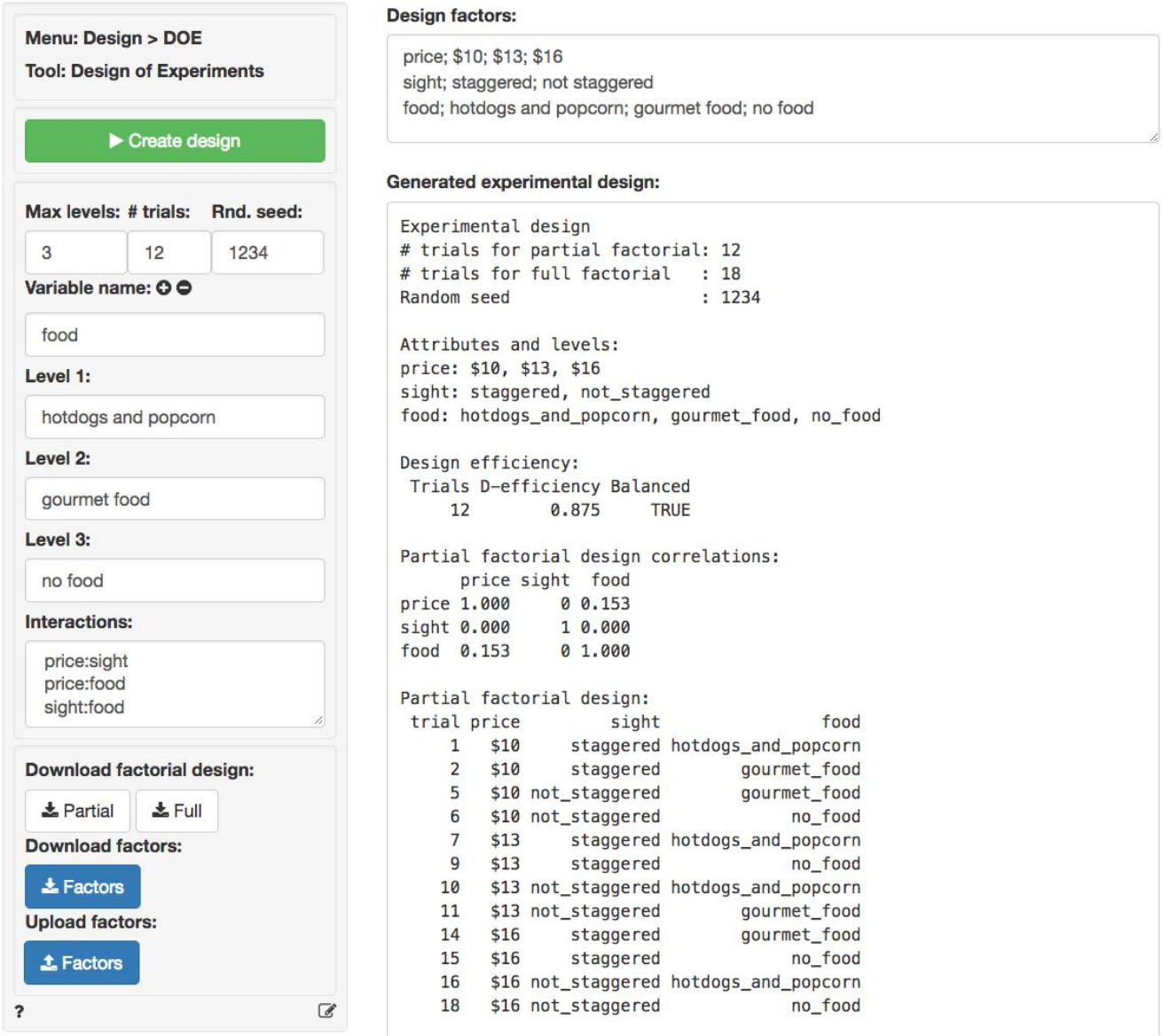


Рис. 4

На рис. 4 представлен план трехфакторного эксперимента по изучению посещаемости кинотеатров в зависимости от цены билета (3 уровня), зрелищности фильма (Да/Нет) и наличия подкормки ("хот-доги и попкорн", "изысканная еда" или "вообще ничего").

Вначале устанавливают максимальное число уровней *Max levels* , стартовое число для генератора случайных чисел *Rnd.seed*, после чего вводят уровни исследуемых факторов, оперируя кнопками +/-.

В этом примере полный план состоит из 18 испытаний. Однако обычно стараются выяснить, нельзя ли сократить количество наблюдений. Можно заранее ввести их число в поле ввода # *trials* (например, 12) и создать дробный факторный план. Эти

входные данные можно использовать для управления количеством генерируемых испытаний. Если оставить *# trials* пустым, то *Radiant* попытается найти соответствующее количество испытаний плана, который является сбалансированным и будет иметь минимальную корреляцию между факторами (например, оценка *D*-эффективности выше 0,8). Обратите внимание, что мы уже не сможем оценить все возможные взаимодействия между ценой, зрелищностью и пищей, если будем использовать план с 12 испытаниями.

Нажмите на кнопку *Partial* ("Частичный") или *Full* ("Полный"), чтобы загрузить ранее сохраненный необходимый вариант плана в формате `csv` .

Design > Random sampling. Пусть имеется большой набор данных (*sampling frame*), предназначенных для последующего наблюдения (например, список предполагаемых респондентов или точек на карте), в котором есть столбец, содержащий уникальный идентификатор объектов. При выборе опции ***Random sampling*** предварительно задаются значения *Variable*, *Sample size* (необходимый объем выборки) и *Rnd.seed.* Каждому объекту данных присваивается случайное число от 0 до 1 из равномерного распределения, после чего строки сортируются, и отбирается *Sample size* объектов с наибольшим случайным баллом. Созданную выборку можно сохранить в *Radiant*: указать имя набора данных и нажать кнопку "Сохранить"

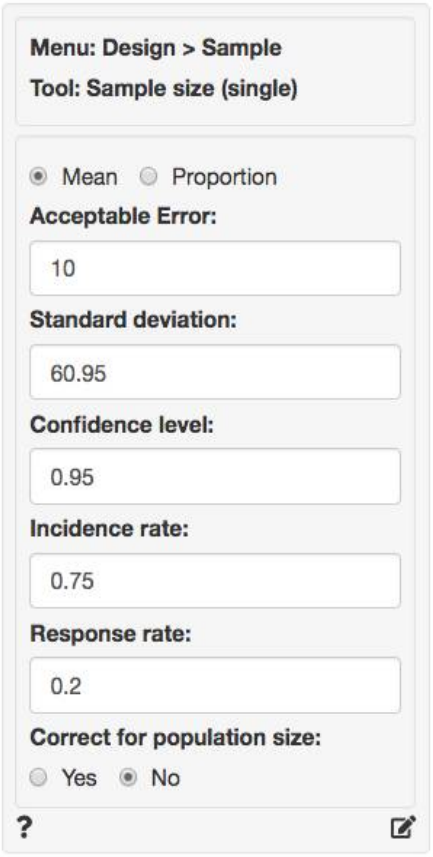
Design > Random assignment. Часто требуется выполнить случайное назначение объектов наблюдений, распределяя их в заданной пропорции *Probabilities* в зависимости от условий эксперимента. Пусть, например, совокупность животных нужно распределить на две группы: "опыт" (30%) и "контроль" (70%). Если в исходном наборе есть поле уникального идентификатора объекта, то это легко сделать с использованием инструмента случайного назначения.

Пусть также в совокупности животных есть самцы и самки и это будет *Blocking variable* или блокирующая переменная. Тогда функция случайного назначения должна использовать равные вероятности для каждого уровня этой переменной (т.е. для опыта будет отобрано ровно 30% самцов и ровно 30% самок). Выражаясь научно, при блочном случайном назначении (или стратифицированном случайном назначении) субъекты сначала сортируются на блоки (или страты) на основе одной или нескольких характеристик, после чего случайным образом назначаются в пределах каждого блока.

Design > Sample size (single). Для определения необходимого размера выборки n при тестировании среднего значения или пропорции задаются следующие исходные показатели: уровень доверительной вероятности (*Confidence level*), требуемая допустимая ошибка E (*Acceptable Error*), доля охвата и доля отклика (*Incidence rate, Response rate*), численность популяции N (*Population size*), выборочное стандартное отклонение s (*Sample standard deviation*) для тестирования среднего и выборочная популяция p (*Population*).

Расчеты необходимого объема выборки ведутся по формулам: при тестировании средних $n = \frac{z^2 \times s^2}{E^2}$, для пропорций $n = \frac{z^2 \times p(1-p)}{E^2}$. Здесь z-статистика связана с необходимой доверительной вероятностью (т.е. равна 1.96 при 95% уровне доверительности). Формула коррекции на численность популяции: $n^* = \frac{nN}{n-1+N}$.

На рис. 5 показаны результаты расчета необходимого объема выборки пользователей Интернета, при которой выборочная оценка среднего с вероятностью 95% не будет отклоняться от истинного значения более, чем на 10 мин. Предположим, что предыдущие исследования дали оценку стандартного отклонения в 60.95 мин.




```
Sample size calculation
Calculation type     : Mean
Acceptable Error     : 10
Standard deviation   : 60.95
Confidence level     : 0.95
Incidence rate       : 0.75
Response rate        : 0.2
Population correction: None

Required sample size     : 143
Required contact attempts: 954
```

Рис. 5

Из расчетов видно, что необходимый размер выборки n = 143. Предположим, однако, что только 75% квартир имеют доступ к Интернету и только 20% жителей захотят ответить на наши вопросы. Поэтому исследователям необходимо связаться с n = 143 / 0.75 / 0.2 = 954 респондентами.

Design > Sample size (compare) – Чтобы определить необходимые размеры выборки при сравнении двух средних или пропорций, предъявляются к заполнению следующие входные значения:

n1: Объем выборки для контрольной группы;

n2: Объем выборки для тестовой группы;

Confidence level ("Уровень доверия"): 1 – α (например, 0.95 = 1 - 0.05) ;

Power ("Мощность"): 1 – β (например, 0.8 = 1 - 0.2).

Входные данные для сравнения средних:

Delta: Разность между групповыми средними, которую надеемся обнаружить;

Std. deviation: Предполагаемое стандартное отклонение.

Входные данные для сравнения пропорций:

Proportion 1: Предполагаемая пропорция в группе 1 (например, 0.1) ;

Proportion 2: Пропорция 1 плюс разность между ними, которую надеемся обнаружить (например, .1 + .05 = .15).

По умолчанию входные значения размера выборки (*n1* и *n2*) остаются пустыми, и вычисляется необходимый размер выборки для обеих групп. Если значения указаны как для *n1*, так и для *n2*, то будет вычислено значение для любого другого входа, оставленного пустым. Если вводится только одно значение *n1* (или *n2*), то для *n2* (или *n1*) должны быть предоставлены все остальные входные данные для определения требуемого объема выборки

4. Меню *Basics* ("Базовая статистика")

Этот раздел меню состоит из следующих подразделов и пунктов:

- ***Probability*** **("Вероятность"):**
  - *Probability calculator* ("Вероятностный калькулятор");
  - *Central Limit Theorem* ("Центральная предельная теорема");
- ***Means*** **("Средние"):**
  - *Single mean* ("Простое средние");
  - *Compare means* ("Сравнение средних");
- ***Proportions*** **("Пропорции"):**
  - *Single proportion* ("Простая пропорция");
  - *Compare proportions* ("Сравнение пропорций");
- ***Tables*** **("Таблицы"):**
  - *Goodness of fit* ("Тест согласия Пирсона");
  - *Cross-tabs* ("Таблицы сопряженности");
  - *Correlation* ("Корреляции").

Basics > Probability > Probability calculator – вычисляет вероятности или частоты, основанные на *Binomial* (биномиальном), *Chi-squared* (χ^2), *Discrete* (Дискретном), *F* (Фишера), *Exponential* (экспоненциальном), *Normal* (нормальном), *Poisson* (Пуассона), *t* (Стьюдента), или *Uniform* (равномерном) распределениях.

Тип распределения задается в поле *Distribution*, его нижняя и верхняя границы – в полях *Lower bound* и *Upper bound*, число необходимых значащих цифр – в поле *Decimals*. Значения остальных полей зависят от воспроизводимого типа распределения. Например, для нормального распределения входными параметрами являются среднее и стандартное отклонение (см. рис. 6).

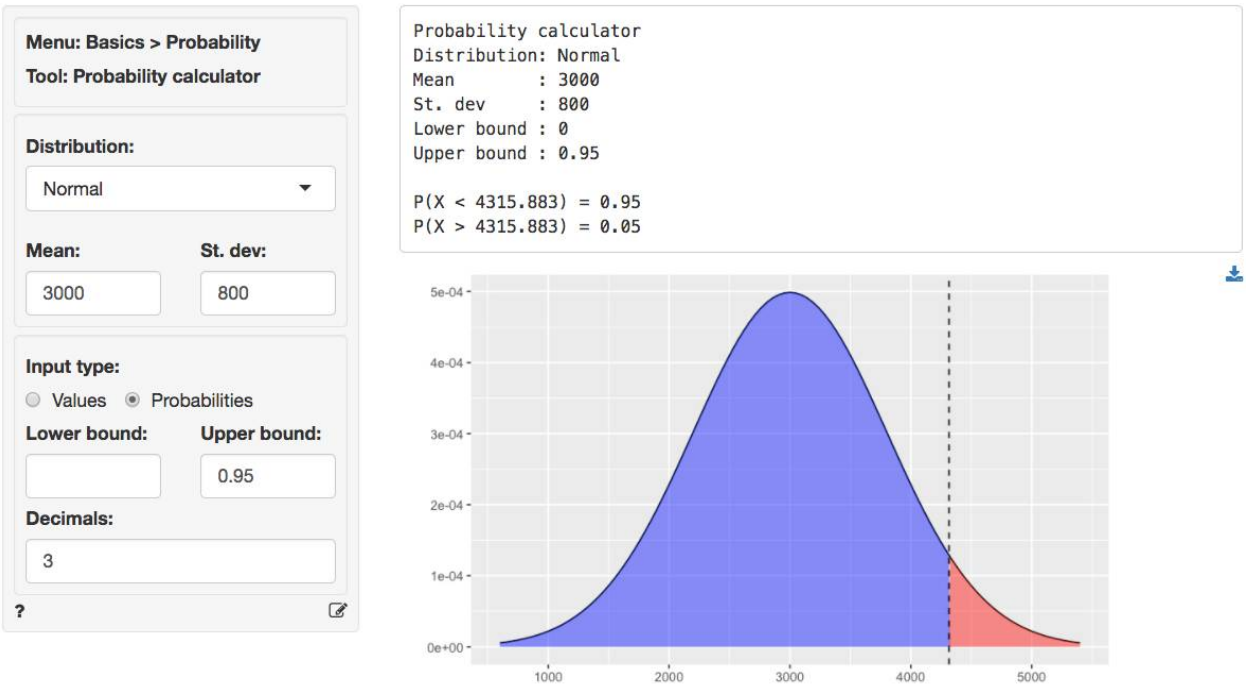


Рис. 6.

С помощью вероятностного калькулятора можно посчитать, например, *p*-значения, т.е. достигнутые критические вероятности при тестировании статистических гипотез. С другими его возможностями можно познакомиться на видео-уроках Radiant Tutorial Series (на английском)

Basics > Probability > Central Limit Theorem – Использование случайной выборки для иллюстрации Центральной предельной теоремы.

Предположим, что получена выборка, содержащая большое количество реализаций изучаемого процесса, причем каждое наблюдение генерируется случайным образом, не зависящим от значений других реализаций, и вычисляется среднее арифметическое наблюдаемых значений. Если эта процедура выполняется много раз, Центральная предельная теорема говорит, что вычисленные значения среднего будут сходиться к нормальному распределению (обычно известному как "колоколообразная кривая"). Этот процесс можно наблюдать в ходе имитации.

Basics > Means > Single mean – выполняется одновыборочный тест с использованием *t*-критерия для сравнения среднего значения переменной в выборке данных с (гипотетическим) средним значением во всей популяции, из которой взяты выборочные данные. Выборочные данные задаются в выпадающем списке *Variable* (только одна переменная), а значение для сравнения – в поле *Comparison value.*

В раскрывающемся списке *Alternative hypothesis* ("Альтернативные гипотезы") можно выбрать либо *two-sided test* (двусторонний тест), либо односторонний тест *less than / greater than* (т. е. меньше / больше). С помощью ползунка задается доверительная вероятность (*confidence level* = 1 - α).

Результаты отображаются на двух вкладках: *Summary* и *Plot*. Сводная статистика включает выборочные оценки параметров (среднее значение *mean*, стандартное отклонение *sd*, стандартная ошибка среднего *se*, погрешность *me*), а также информацию по существу теста:

- *diff* – разница между выборочным средним значением и базой сравнения;
- *se* – стандартная ошибка (т. е. стандартное отклонение выборочного распределения *diff*) ;
- *t.value* – *t*-статистика, связанная с *diff* (т. е. *diff* / *se*), которую можно оценить на основе *t*-распределения;
- *p.value* – если нулевая гипотеза верна, то это вероятность появления значения *diff*, более экстремального, чем рассчитанное ;
- *df* – число степеней свободы, связанное со статистическим тестом (т. е. *n* - 1) ;
- *5% 100%* – показывают 95% - ный доверительный интервал вокруг выборочного среднего значения, т.е. диапазон, в пределах которого, вероятно, находится истинное среднее значение популяции.

Проверить истинность нулевой гипотезы можно с использованием любого из трех условий:

- значение *p.value* должно превышать уровень значимости (т. е. α = 0,05);
- сравниваемое значение должно входить в доверительный интервал выборочного среднего;
- рассчитанное значение *t.value* не должно превышать критическое значение, найденное по *t*-распределению с *df* степенями свободы и заданной верхней границей вероятности (то есть 1 - 0,05).

В дополнение к числовым выводам, можно визуализировать данные на вкладке *Plot* ("График"), например, в виде гистограммы с проведенными линиями, соответствующими среднему значению выборки (сплошная), доверительному интервалу (пунктир) и базе сравнения (красная линия).

Basics > Means > Compare means – осуществляет сравнение средних значений двух или более переменных или групп данных. Отметим, что многие положения аналогичны описанному выше одновыборочному тесту и повторно не обсуждаются.

Рассмотрим выполнение теста на примере сравнения зарплаты за 9 месяцев для различных преподавательских должностей (*Assistant Professors*, *Associate Professors* и

Professors) в колледжах США за 2008-2009 гг. В качестве исходных данных задается фактор `rank` с тремя уровнями должностей и числовая переменная `salary`. В поле *Choose combinations* выбираются все возможные комбинации для проведения парных сравнений на всех трех уровнях.

Проверим гипотезу, получают ли преподаватели более низкого ранга более низкую зарплату по сравнению с профессорами более высокого ранга, т.е. рассматривается нулевая гипотеза *Null hyp.* о равенстве зарплат против односторонней альтернативной гипотезы *Alt. hyp.* что *less than.* Обратим также внимание, что данные в группах являются независимыми, а поправку Бонферрони мы использовать не будем. Результаты тестирования представлены на рис. 7.

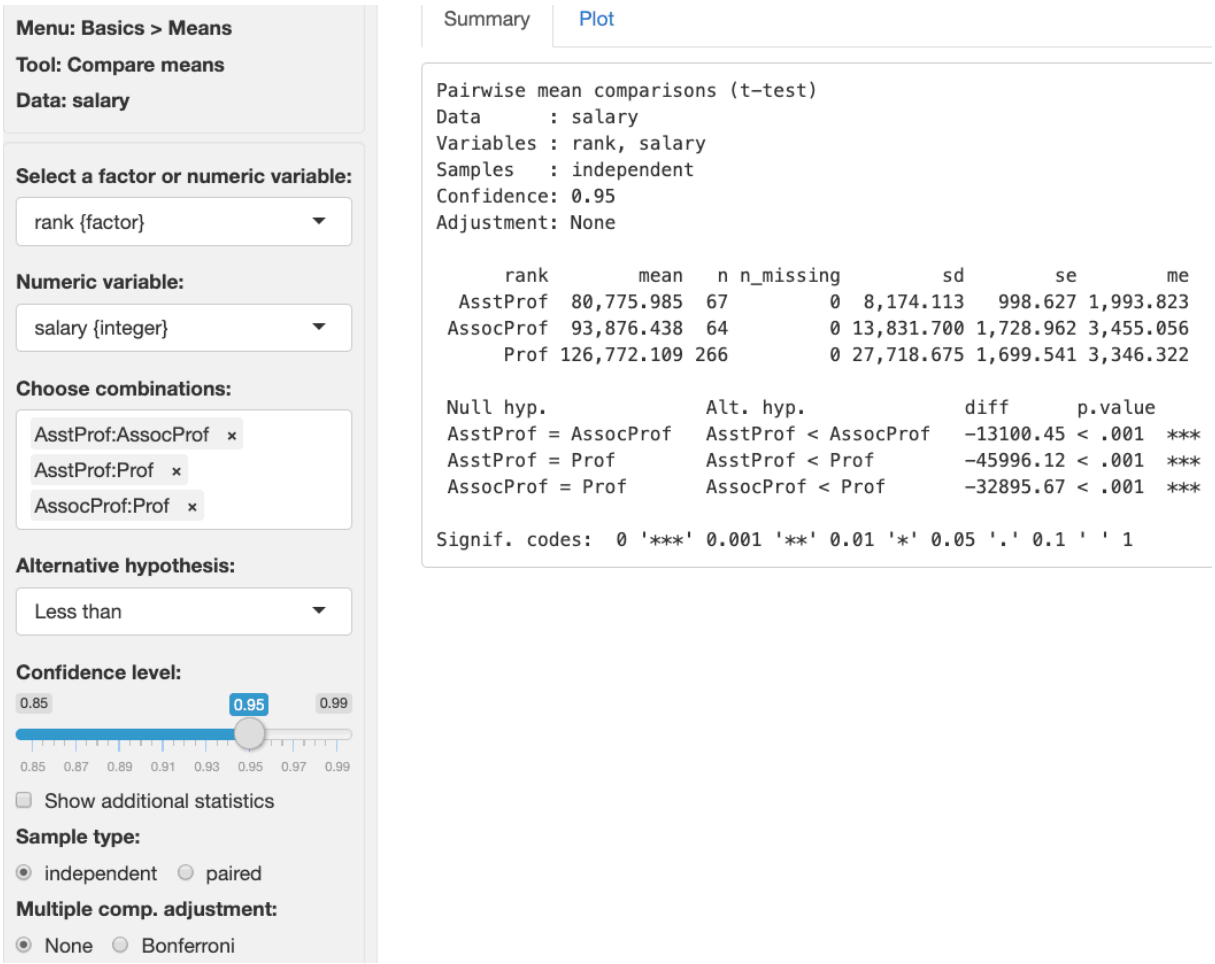


Рис. 7

Как и в одновыборочном тесте, результаты включают набор традиционных статистик для каждой из трех групп данных. Блок проверки гипотез *Null hyp.* против *Alt. hyp.* для каждой пары комбинаций групп включает разность между групповыми средними *diff* и соответствующее *p.value*. При установленном флажке *Show additional statistics* выводится более полная информация: дополнительно выводятся *t*-статистика, стандартное отклонение и доверительные интервалы для разности групповых средних.

Тест показал, что для любых парных комбинаций должностей нулевая гипотеза о равенстве зарплат отвергается в пользу альтернативы. Этот вывод подкрепляется графиками, представленными на вкладке *Plot* – комбинация графиков Scatter + Bar представлена на рис. 8. Обратим внимание, что множественная гипотеза о различии зарплат в целом для всех категорий нами не рассматривалась, поскольку поправка Бонферрони была отключена. Отметим также, что зарплата российских преподавателей примерно в 15 раз меньше, чем у американских.

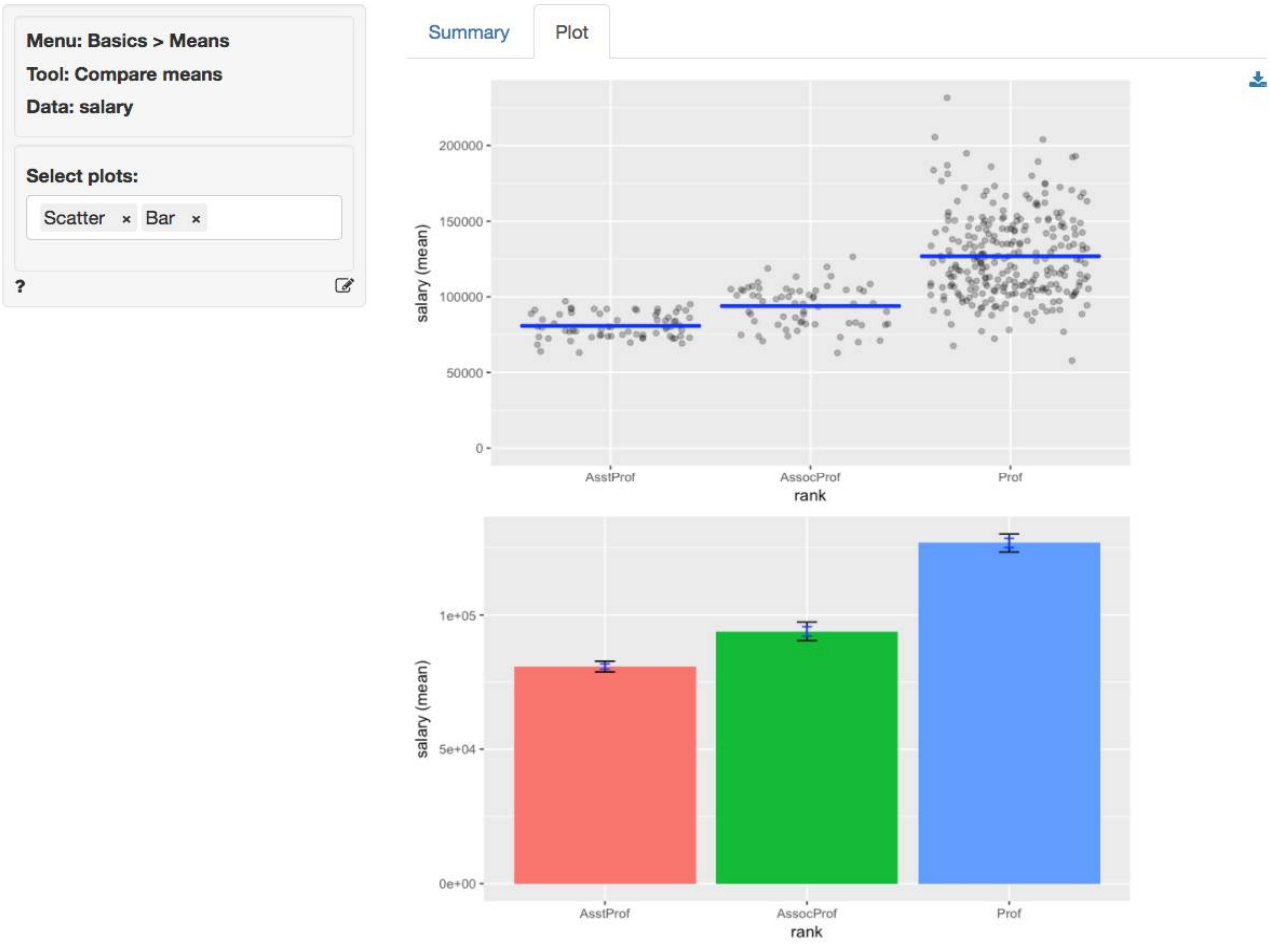


Рис. 8

Basics > Proportions > Single proportion – используется биномиальный тест для долей по результатам одной выборки при сравнении с (гипотетической) долей во всей популяции. Обычно выборочные пропорции рассчитываются для столбца анализируемого фактора, один из уровней *Choose level* которого принимается за базовый, и вычисляется доля его значений по отношению к объему всей выборки. Проверка гипотез о значимости различий выполняется на основе χ^2-критерия. В остальном реализация этого пункта меню не отличается от одновыборочного теста для средних.

Basics > Proportions > Compare proportions. Тест сравнения пропорций используется для оценки того, однородны ли в группах частоты возникновения какого-либо события, поведения, намерения и т. д. Обычно исходные данные представлены в виде двух факторов: первый является группообразующим, в второй – для вычисления долей одного из уровней для каждой группы. Нулевая гипотеза утверждает, что разности в пропорциях между группами равны нулю. В целом техника применения теста не отличается от сравнения групповых средних.

Basics > Tables > Goodness-of-fit. Тест согласия Пирсона (или критерий Хи-квадрат) является непараметрическим методом, который позволяет оценить значимость различий между фактическим (выявленным в результате исследования) количеством исходов и теоретическими частотами, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы. Для выполнения теста задается анализируемый фактор *categorical variable* с произвольным числом уровней и список ожидаемых долей *Probabilities* для каждого из них. Выводятся наблюдаемые и ожидаемые частоты, χ^2-статистика и соответствующее ей *p*-значение.

Basics > Tables > Cross-tabs - или анализ таблиц сопряженности используется для оценки того, связаны ли между собой категориальные переменные. Для выполнения анализа задаются две переменные *categorical variable* – факторы с произвольным числом уровней каждый и состав необходимых блоков для вывода результатов (рис. 9).

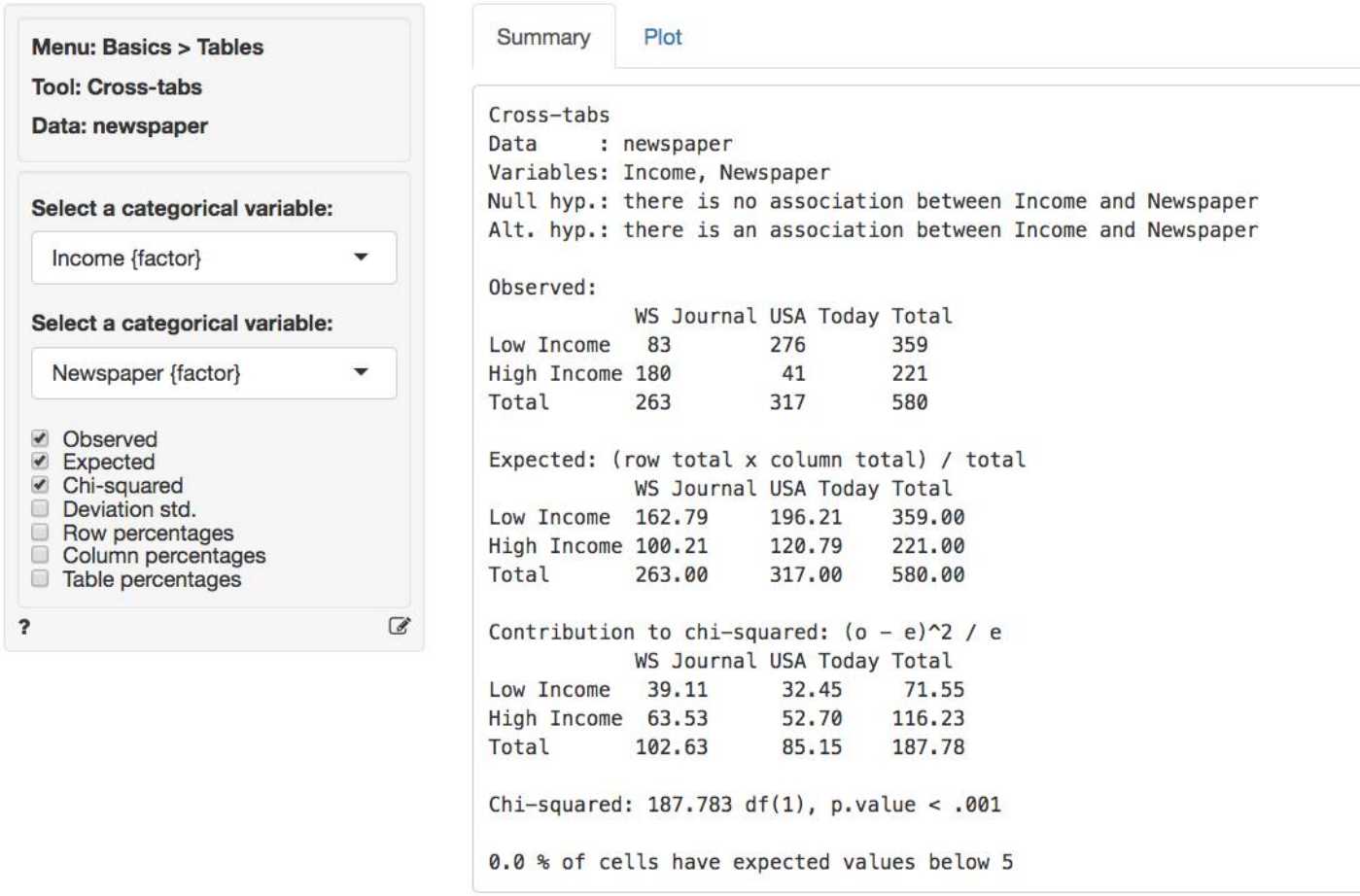


Рис. 9

Здесь проверяется, существует ли связь между уровнем дохода американцев и выбором читаемой ими газеты. В частности, рассматриваются следующие нулевая и альтернативная гипотезы:

- H_0: Нет никакой связи между уровнем дохода и выбором газеты;
- H_a: Существует связь между уровнем дохода и выбором газеты.

Анализ включает, во-первых, сравнение наблюдаемых (*observed*) и ожидаемых (*expected*) частот. Ожидаемые частоты вычисляются при справедливости H_0 (т.е. без учета связи) как *Row Total* × *Column Total*) / *Total Total*.

Далее вычисляется значение хи-квадрат для каждой ячейки, а его общее значение получается суммированием по всем ячейкам. Выходные результаты на вкладке "Сводка" включают значение χ^2-статистики, число степеней свободы и *p*-значение, связанные с тестом. Не лишним будет проверить, чтобы все ожидаемые частоты были больше 5, и тогда значение *p* в тесте считается корректным.

В дополнение к численной проверке гипотез, представленной на вкладке *Summary*, сделать выводы можно и визуально (см. вкладку *Plot*). Например, можно построить график стандартизированных отклонений, т.е. оценок того, насколько различны наблюдаемые и ожидаемые частоты в каждой ячейке. Если стандартизированное отклонение по абсолютному значению больше 1,96, то эта ячейка имеет значимое отклонение от модели независимости (см. рис. 10).

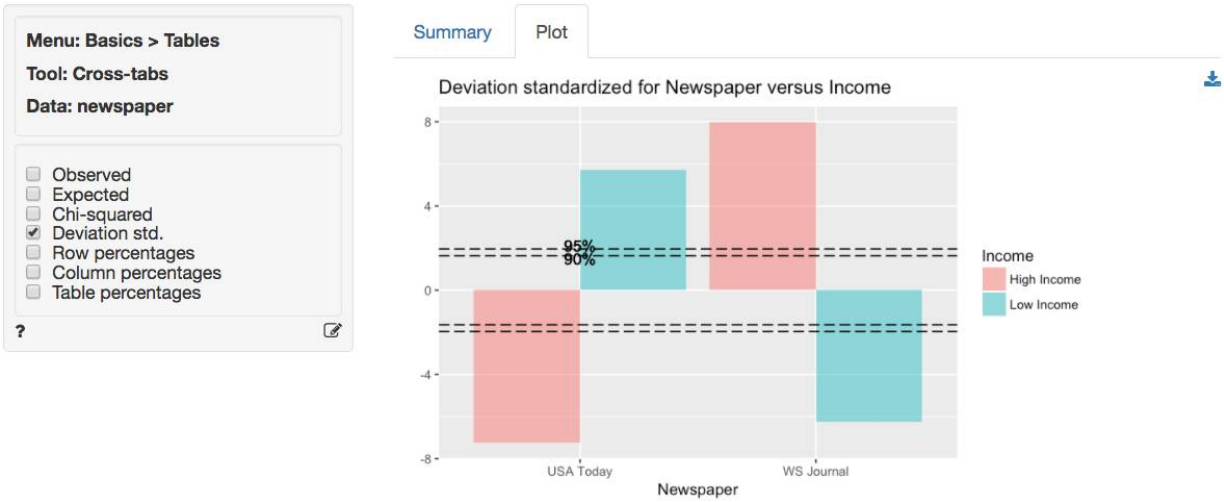


Рис. 10

На графике видно, что все ячейки вносят свой вклад в ассоциацию между доходом и читательской аудиторией, поскольку все стандартизированные отклонения больше 1,96 (то есть столбики выходят за пределы внешней пунктирной линии на графике).

Basics > Tables > Correlation. Создает корреляционную матрицу выбранных переменных, в которой приведены коэффициенты корреляции и *p*-значения для каждой пары переменных. Корреляции могут быть вычислены для переменных типа `numeric`, `integer`, `date` и `factor`. При включении переменных типа `factor` следует установить флажок *Adjust for {factor} variables*. Чтобы показать только те корреляции, которые превышают определенный абсолютный уровень, используется поле отсечки малой корреляции (*Correlation cutoff*).

Визуально матрица может быть показана на вкладке *Plot* – см. рис. 11.

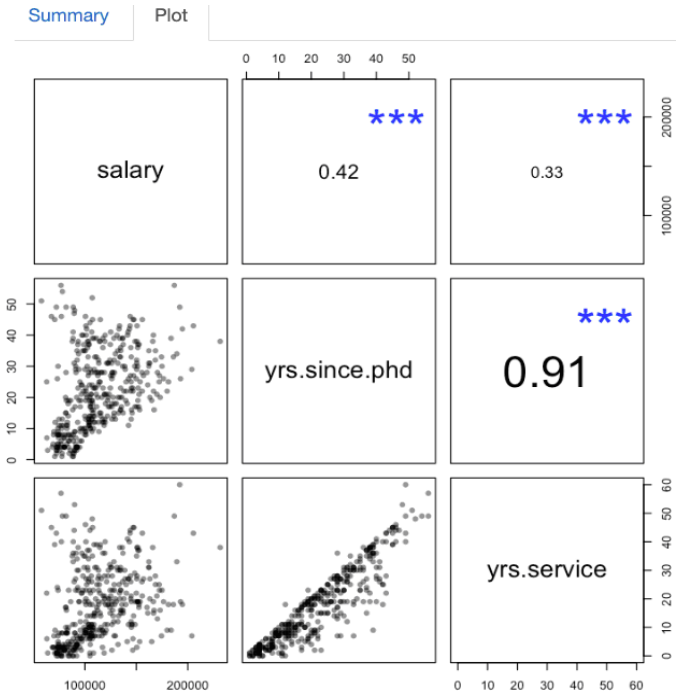


Рис 11

Если установлен флажок *Show covariance matrix*, то все сказанное относится к ковариационной матрице.

Корреляционная или ковариационная матрицы могут быть сохранены для последующего использования как таблицы данных (*data.frame*).

5. Меню *Models* (Статистические модели)

Этот раздел меню состоит из следующих подразделов и пунктов:

- ***Estimate* – модели на основе оценки параметров:**
  - *Linear regression (OLS)* – линейная регрессия (обыкновенный метод наименьших квадратов) ;
  - *Logistic regression (GLM)* – логистическая регрессия (обобщенная линейная модель) ;
  - *Multinomial logistic regression (MNL)* – мультиномиальная логистическая регрессия ;
  - *Naive Bayes* – "наивный" байесовский классификатор;
  - *Neural Network* – нейронные сети.
- ***Trees* – деревья:**
  - *Classification and regression trees* – деревья классификации и регрессии;
  - *Random Forest* – "случайный лес";
  - *Gradient Boosted Trees* – деревья градиентного бустинга.
- ***Evaluate*:**
  - *Evaluate regression* – оценка эффективности прогнозов моделей регрессии;
  - *Evaluate classification* – оценка эффективности прогнозов моделей классификации.
- ***Recommend:***
  - *Collaborative Filtering* – коллаборативная фильтрация.
- ***Decide* :**
  - *Decision analysis* – анализ решений;
  - *Simulation* – имитационные модели.

Model > Estimate > Linear regression (OLS). Построение и интерпретация модели линейной регрессии – наиболее распространенный метод анализа эмпирических данных во многих отраслях наук.

Перед построением модели определяются следующие функциональные условия оценки ее коэффициентов:

° Выбирается переменная отклика (*Response variable*) и одна или несколько объясняющих переменных (*Explanatory variables*).

° Если в модель включаются две или более объясняющих переменных, то можно учесть парные (или тройные) эффекты их взаимодействия (*Interactions 2-way or 3-way*). Взаимодействие существует, если влияние объясняющей переменной на переменную отклика определяется, по крайней мере, частично, уровнем другой объясняющей переменной. Например, увеличение цены алмаза на 1 карат может зависеть от его чистоты.

° Для проведения стандартного дисперсионного анализа полученной модели выбранные переменные отмечаются в раскрывающемся списке *Variables to test*.

° Коэффициенты модели трудно сопоставимы, если объясняющие переменные измеряются в разных масштабах. Стандартизация (*Standardize*) переменной отклика и объясняющих переменных позволяет оценить степень относительного влияния переменных. При интерпретации эффектов взаимодействия полезно выполнить хотя бы центрирование (*Center*) переменных.

° *Stepwise*: Запуск пошаговой процедуры для выбора наилучшей модели с минимальным числом переменных.

° *Robust standard errors* ("Робастные стандартные ошибки"): При выборе робастности оценки коэффициентов совпадают с оценками *OLS*. Однако стандартные ошибки корректируются с учетом (незначительной) неоднородности и ненормальности.

Изменение перечисленных условий приводит к повторному построению модели. На основе полученной модели при необходимости могут быть рассчитаны следующие дополнительные показатели:

- *RMSE*: среднеквадратичная ошибка и остаточное стандартное отклонение.
- *Sum of Squares* ("Разложение суммы квадратов"): т.е. общая дисперсия переменной отклика раскладывается на составляющие – дисперсию, объясняемую регрессией, и дисперсию, которая остается необъясненной (т. е. случайную ошибку).
- *VIF* ("Фактор инфляции дисперсии"): показатель мультиколлинеарности объясняющих переменных.
- *Confidence intervals*: доверительные интервалы для коэффициентов модели.

В состав предикторов модели (т.е. объясняющих переменных) могут входить как числовые, так и категориальные переменные (`factor`). На примере уже знакомой таблицы `diamonds` рассмотрим зависимость стоимости бриллиантов `price` от их веса в каратах `carat`, максимальной ширины `table` и факторов `clarity` и `color`, связанных с прозрачностью и цветом алмазов. После того, как выбраны все перечисленные условия, можно приступить к оценке параметров модели и нажать кнопку *Estimate model* (см. рис. 12).

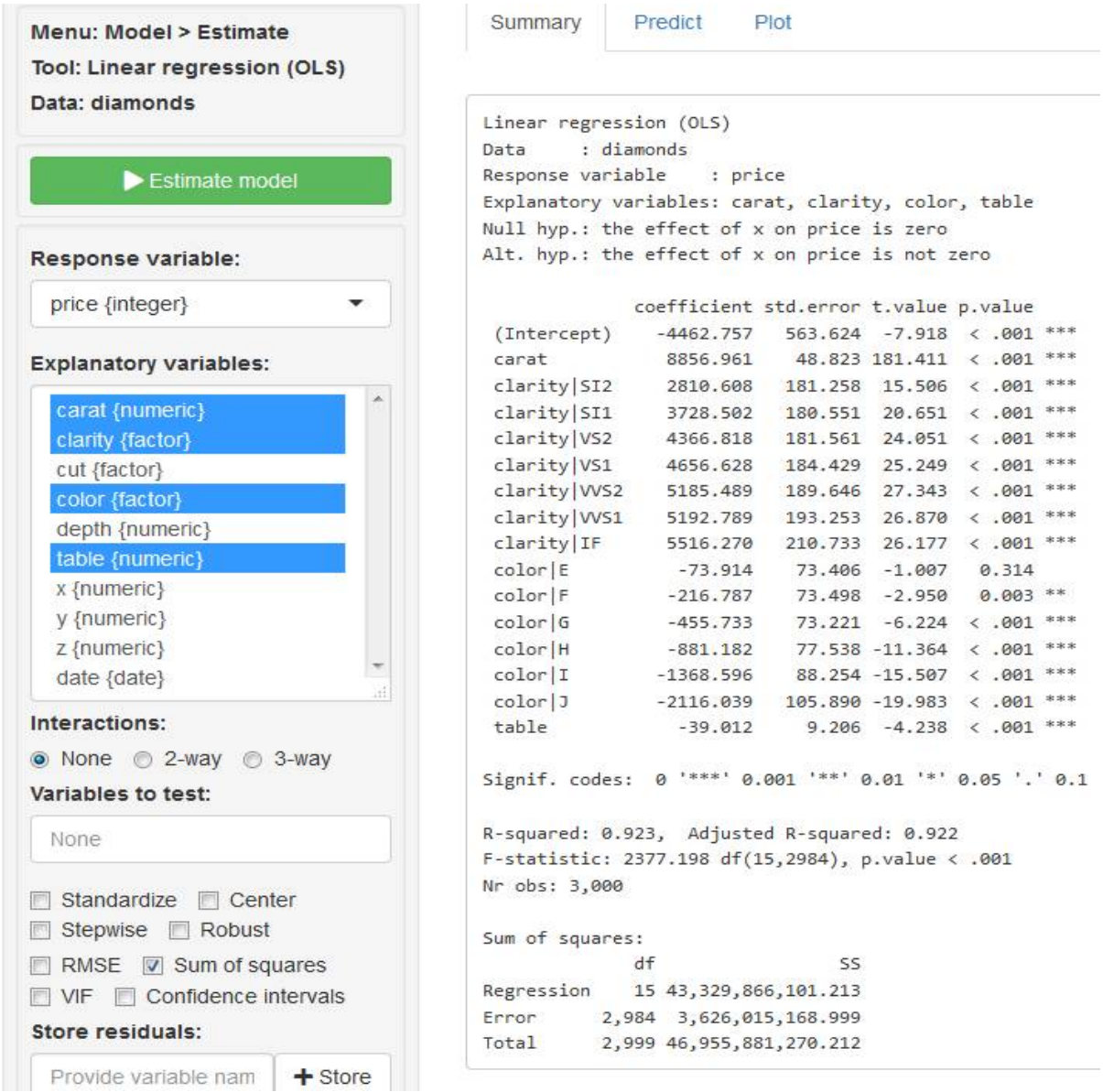


Рис. 12

В стандартный набор выводимых результатов вкладки *Summary* входят:

- Данные об оцененных коэффициентах модели для каждой из числовых объясняемых переменных или для каждого уровня факторов: стандартная ошибка, *t*-статистика и соответствующее ей *p*-значение. Если *p*-value > α (α = 0.05), то коэффициент считается незначимым, а влияние соответствующей переменной – несущественным.
- Обычный (*R-squared*) и приведенный (*Adjusted R-squared*) коэффициенты детерминации.
- *F*-статистика, которая является отношением суммы квадратов отклонений отклика нуль-модели без параметров к оцененной модели с учетом степеней свободы, а также соответствующее ей *p*-значение. Если все объясняющие переменные включить в *Variables to test* , то выводится дополнительно стандартная таблица дисперсионного анализа, а если установлен флажок *Sum of Squares* – просто таблица сумм квадратов.

Интерпретация коэффициентов при категориальных переменных осуществляется следующим образом. Свободный член модели *Intercept* рассчитывается из условия, что коэффициенты для первых по списку уровней факторов равны 0. Если, например, вместо цветности `color|D` алмаз будет иметь категорию `color|H`, то его цена упадет на 881$.

Вкладка *Predict* ("Прогноз") позволяет рассчитать прогнозируемые значения отклика по регрессионной модели. Предсказать результаты можно только на основе переменных, включенных в модель. Их значения могут быть представлены в виде таблицы или заданы командой (выражением синтаксиса R). Источник данных, по которым нужно сделать прогноз, выбирается из раскрывающегося списка *Prediction input* ("Ввод прогноза"). Как только нужные прогнозы будут сгенерированы, их можно сохранить в `csv`-файл, или добавить прогнозы в набор данных, используемый для построения модели, нажав кнопку *Store*.

Вкладка *Plot* ("График") используется для визуализации вероятностных диаграмм исходных переменных и диагностических графиков проверки предположений регрессионной модели. В их комплект входит такие традиционные или представленные ранее графики, как *Distribution* (по каждой из переменной выводится график распределения в виде гистограммы), *Correlations* (см. рис. 11), *Scatter* (см. рис. 8), *Coefficient plot* (диаграмма коэффициентов и их доверительных интервалов на одной шкале), *Residuals vs Explanatory* (зависимость остатков модели для каждой переменной). Рассмотрим подробнее графики *Dashboard* и *Influential observations*, которые для построенной модели показаны на рис. 13-14.

Панель *Dashboard* состоит из шести графиков остатков модели, позволяющих выполнить диагностику линейной модели и проверку ее исходных предпосылок:

- нормальность распределения остатков и постоянство их дисперсии;
- независимость остатков от значений предикторов X и прогнозов отклика $\hat{Y}$;
- адекватность линейной формы модели по отношению к обработанным данным.

График зависимости наблюдаемых значений отклика от его предсказаний моделью (*Actual vs Fitted values*) в идеале должен выглядеть (почти) как прямая линия со случайным разбросом, то есть по мере роста фактических значений переменной отклика пропорционально растут и предсказанные по модели значения. На графике *Residuals vs Fitted* значения остатков (т. е. разностей между наблюдаемыми величинами отклика и значениями, предсказанными регрессией) не должны демонстрировать никакой закономерности и быть случайно и равномерно разбросанными по горизонтальной линии, параллельной оси прогнозов отклика. Аналогично сглаженная кривая остатков относительно порядка строк набора данных *Residuals vs Row order* не должна серьезно отклоняться от горизонтальной прямой. Зависимость квантилей стандартизованных остатков от теоретических квантилей нормального распределения (*Normal Q-Q*) должна быть максимально приближена к диагональной линии, что

свидетельствует о нормальном распределении остатков. Этот вывод подтверждается гистограммой остатков (*Histogram of residuals*) и графиком (*Residuals vs Normal Density*) плотности распределения остатков (зеленое заполнение) по отношению к теоретической плотности нормально распределенной величины (синяя линия).

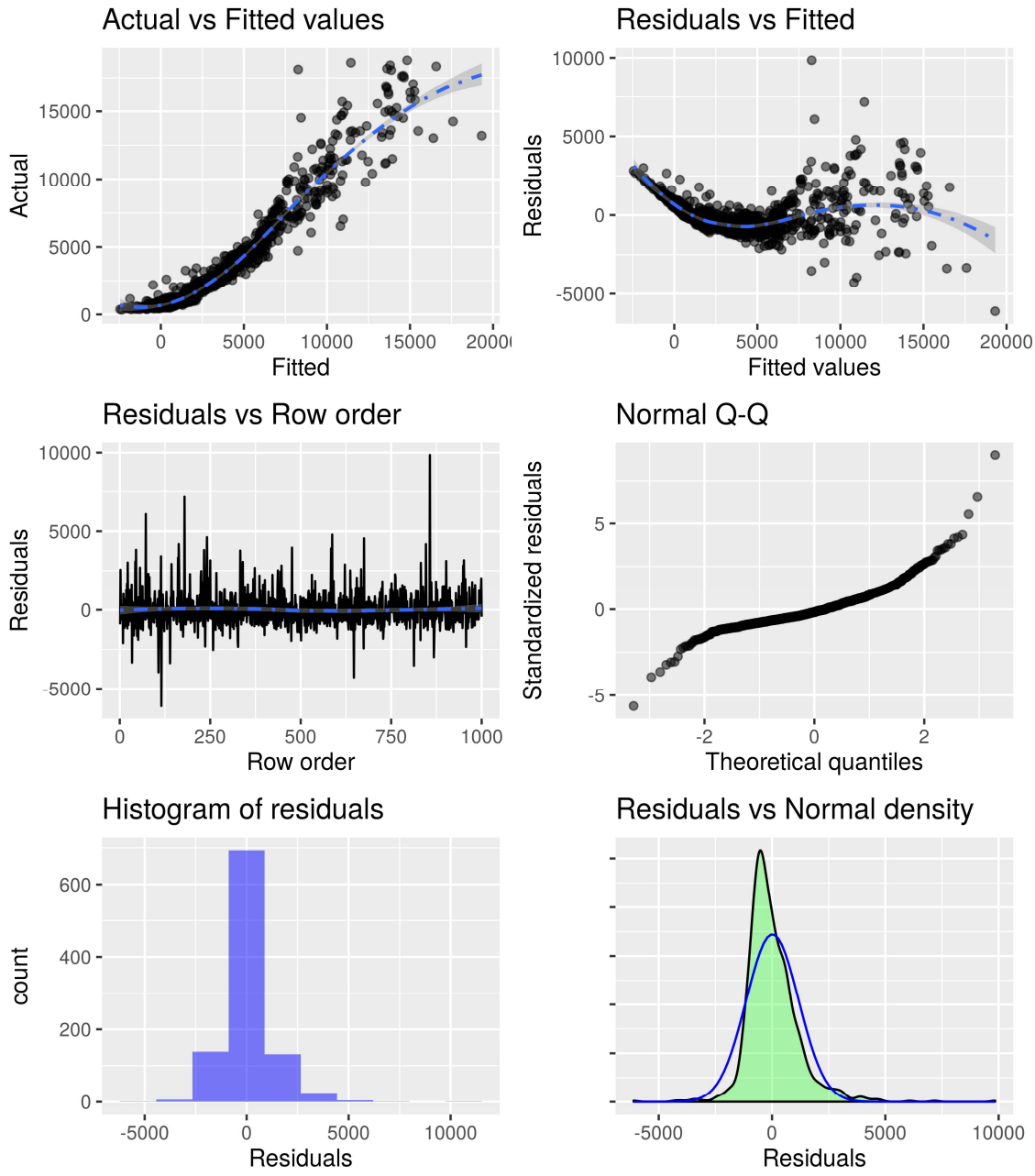


Рис. 13

Представленные графики наглядно демонстрируют, что предпосылки модели не соблюдаются и зависимость стоимости алмазов от их веса носит отчетливо нелинейный характер.

График *Influential observations* представляет собой диаграмму для обнаружения выбросов с использованием стандартизированных остатков (т.е. остатков, деленный на оценку стандартного отклонения). Диаметр кружков соответствует дистанции Кука. Номерами строк относительно исходного набора данных отмечены кружки с аномально высокими и низкими значениями выбросов.

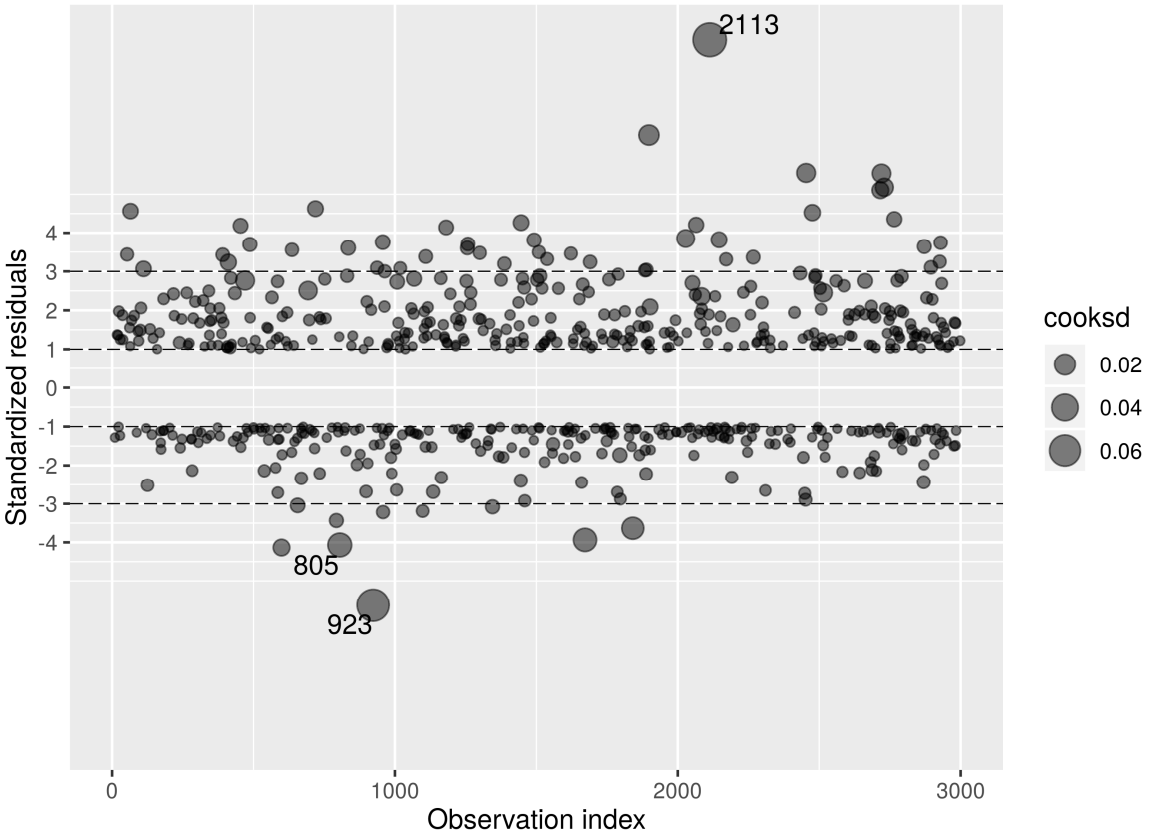


Рис. 14

Model > Estimate > Logistic regression (GLM) выполняет построение модели логистической регрессии для прогнозирования отклика в форме бинарной переменной. Например, можно рассмотреть зависимость числа выживших в крушении «Титаника» (уровень `Yes` фактора `survived` таблицы `titanic`) от совокупности иных численных и категориальных переменных (класса каюты `class`, пола `sex`, и возраста пассажиров `age`). Имеются лишь незначительные отличия в процедурах построения модели логита по сравнению с линейным случаем *OLS,* поэтому приведем ниже только некоторые характерные особенности.

В блоке коэффициентов появился новый столбец *OR* (*odds-ratios* или соотношение шансов), определяющий в данном случае оценку вероятности выжить у пассажира. Поскольку подгонка модели выполнялась методом максимального правдоподобия, приведена достигнутая оценка *Log-likelihood*, а также информационные критерии *AIC* и *BIC*. Гипотеза об отсутствии отличий между нулевой и оцененной моделями проверяется с помощью теста Хи-квадрат.

Вкладку *Predict* ("Прогноз") можно использовать для оценки распределения вероятностей отклика при различных значениях объясняющих переменных (вероятности более удобны для интерпретации, чем коэффициенты *GLM* или отношения шансов). Сначала с помощью раскрывающегося списка *Prediction input type* выбирается тип данных для прогнозирования: существующий набор данных *Data* ("Данные"), либо *Command* (“Команда”) для генерации входных данных..

Если вводится команда, нужно указать, по крайней мере, одну переменную и одно значение в поле *Prediction command*, чтобы получить прогноз. Для каждой незаданной переменной оценка прогноза будет использовать либо среднее значение, либо наиболее частый уровень. Если для прогноза выбрать саму таблицу `titanic`, то можно сформировать и сохранить не только полную таблицу вероятностей, но и график, который дает четкое представление о том, как вероятность выживания меняется с возрастом, полом и классом каюты (рис. 15):

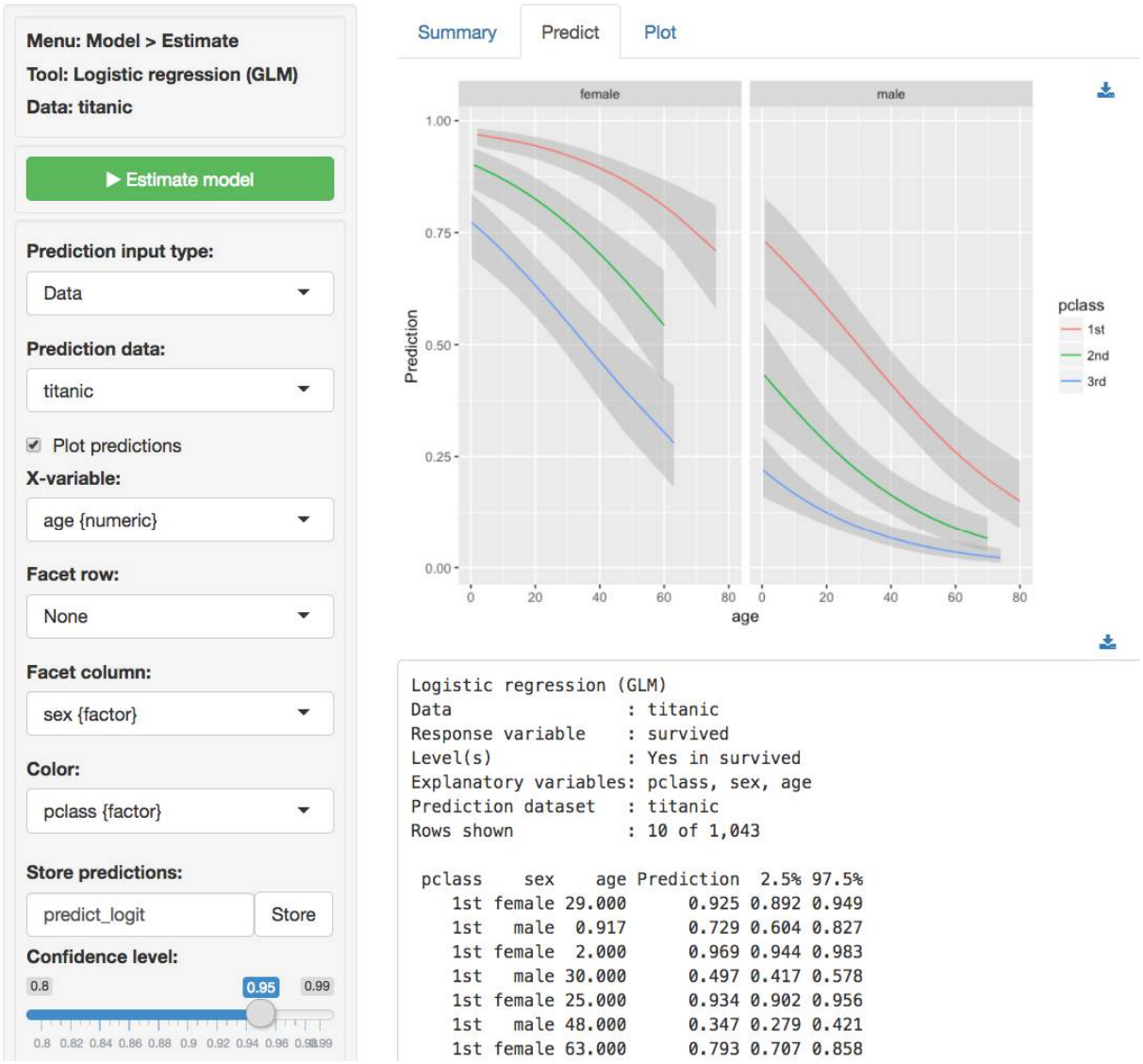


Рис. 15

Model > Estimate > Multinomial logistic regression (MNL) или мультиномиальный логит используется, когда необходимо сделать прогноз категориальной переменной отклика с тремя или более уровнями от одной или нескольких объясняющих переменных. Если имеется J уровней отклика, то при независимости всех альтернатив многомерное распределение вероятности представлено вектором $\{\pi_1, ..., \pi_J\}$. Для его моделирования достаточно сформировать систему $J - 1$ логит-уравнений, коэффициенты которых оцениваются совместно путем максимизации общего критерия правдоподобия. В остальном имеются лишь некоторые незначительные отличия в процедурах построения модели по сравнению с простой логистической регрессией.

К сожалению, коэффициенты мультиномиальной модели логистической регрессии трудно интерпретировать напрямую. Легче работать с оценками относительных соотношений шансов (*Relative-Risk-Ratios*, сокращенно *RRR*), которые численно являются функциями экспоненты от коэффициентов регрессии.

Результаты построения модели, как обычно, представлены блоками *Summary*, либо на графиках вкладки *Plot* для визуального анализа. На рис. 16 представлен график коэффициентов модели (вернее, относительных отношений шансов) с доверительными интервалами, с помощью которого можно проанализировать относительную важность цены при выборе того или иного сорта кетчупа.

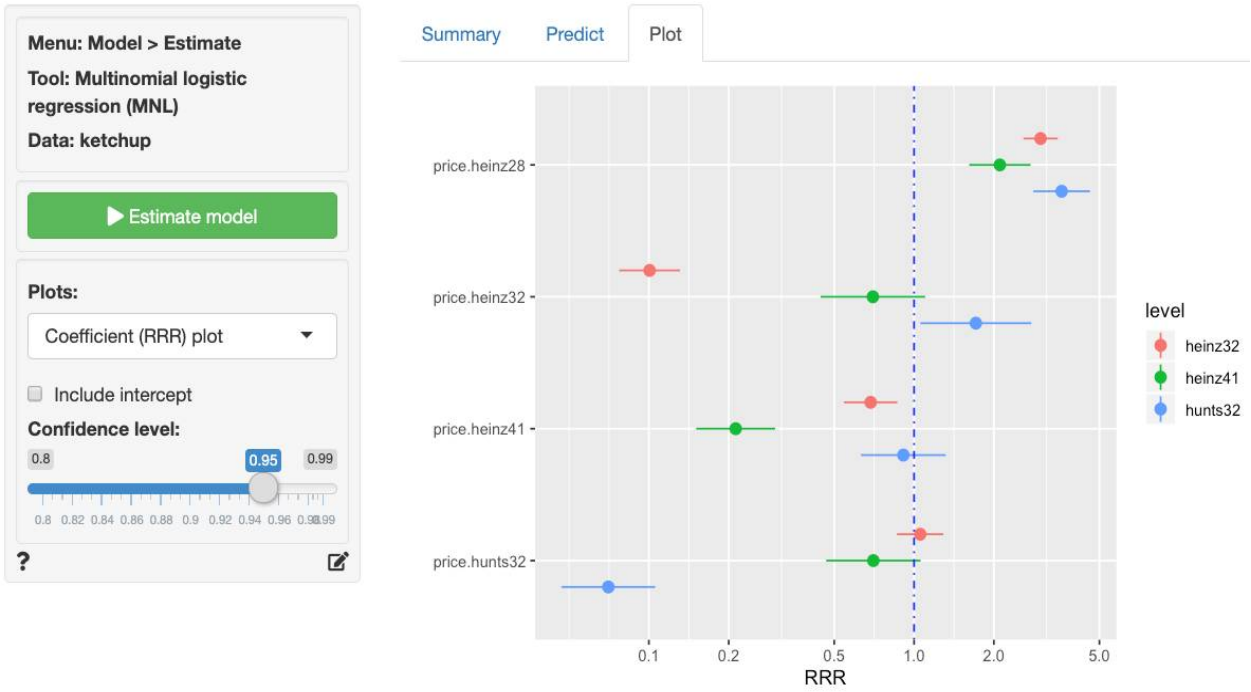


Рис. 16

Model > Estimate > Naive Bayes . Модель "наивного" байесовского классификатора также применяется, когда необходимо сделать прогноз категориальной переменной отклика от одной или нескольких объясняющих переменных. По технике использования не отличается от простого или мультиномиального логита.

Рассмотрим снова пример с выжившими при крушении «Титаника» (коррекцию вероятностей по Лапласу проводить не будем). Задав на вкладке *Predict* имена соответствующих переменных, получим график, представленный на рис. 17, из которого видно соотношение выживших/погибших мужчин и женщин разных возрастов:

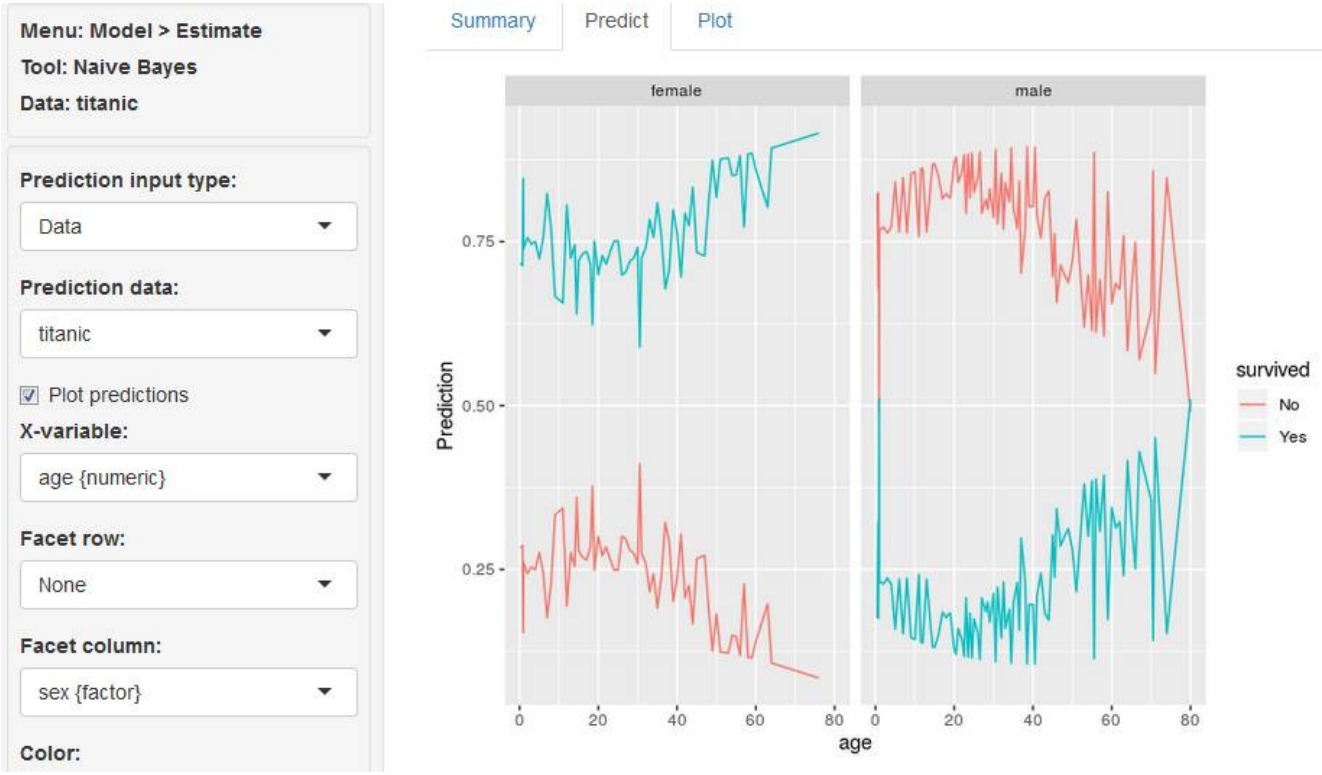


Рис. 17

Model > Estimate > Neural Network – выполняет построение многослойного персептрона Розенблата, который, в принципе, аналогичен нелинейной регрессии или логиту, но добавление произвольного числа промежуточных (скрытых) слоев и применение нелинейных функций активации позволяет существенно усложнить структуру модели. Перед построением модели необходимо выбрать тип (*Classification* или *Regression*), переменную отклика и одну или несколько объясняющих переменных.

Основная проблема обучения ИНС заключается в необходимости предварительно исследовать поверхность ошибок и задать архитектуру сети и параметры оптимизации, приводящие, по возможности, в окрестность глобального минимума. К ним относятся число слоев скрытых нейронов *size* и параметр "затухания весов" *decay*, который осуществляет регуляризацию точности подстройки коэффициентов (при *decay* = 0 стремление к точности может перерасти в эффект переусложнения модели). Оптимальные значения параметров оцениваются путем перекрестной проверки.

Построим персептрон (рис. 18) с одним промежуточным слоем для моделирования зависимости стоимости бриллиантов `price` от веса в каратах `carat`, максимальной ширины `table` и глубины `depth` алмазов (численные предикторы) и факторов `cut` и `clarity`, связанных с качеством огранки и прозрачностью (см. таблицу `diamonds`).

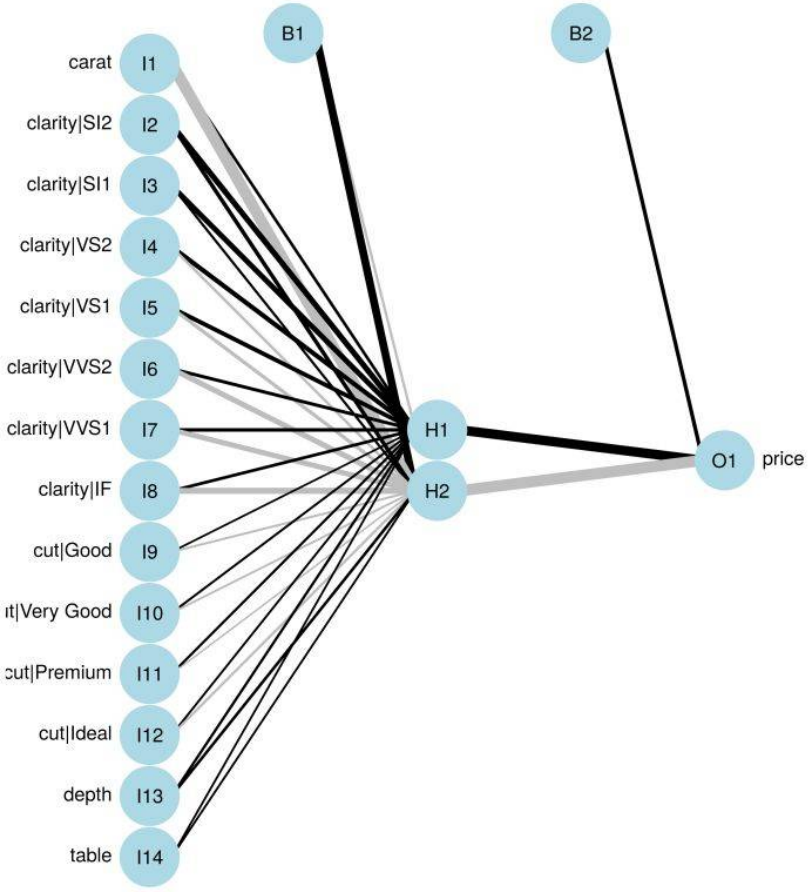


Рис. 18

На вкладке *Plot* можно построить графики, касающиеся связи отдельных переменных с откликом (рис. 19) и сравнительной важности предикторов по *Olden* (рис. 20). В частности, для нейрона, связанного с весом алмаза, выбрана сигмоидная функция активации, для ширины основания `table` - линейная, а для прозрачности `clarity` выходной сигнал задается дискретным набором значений. Можно отметить также, что из численных переменных только вес алмаза имеет решающее значение.

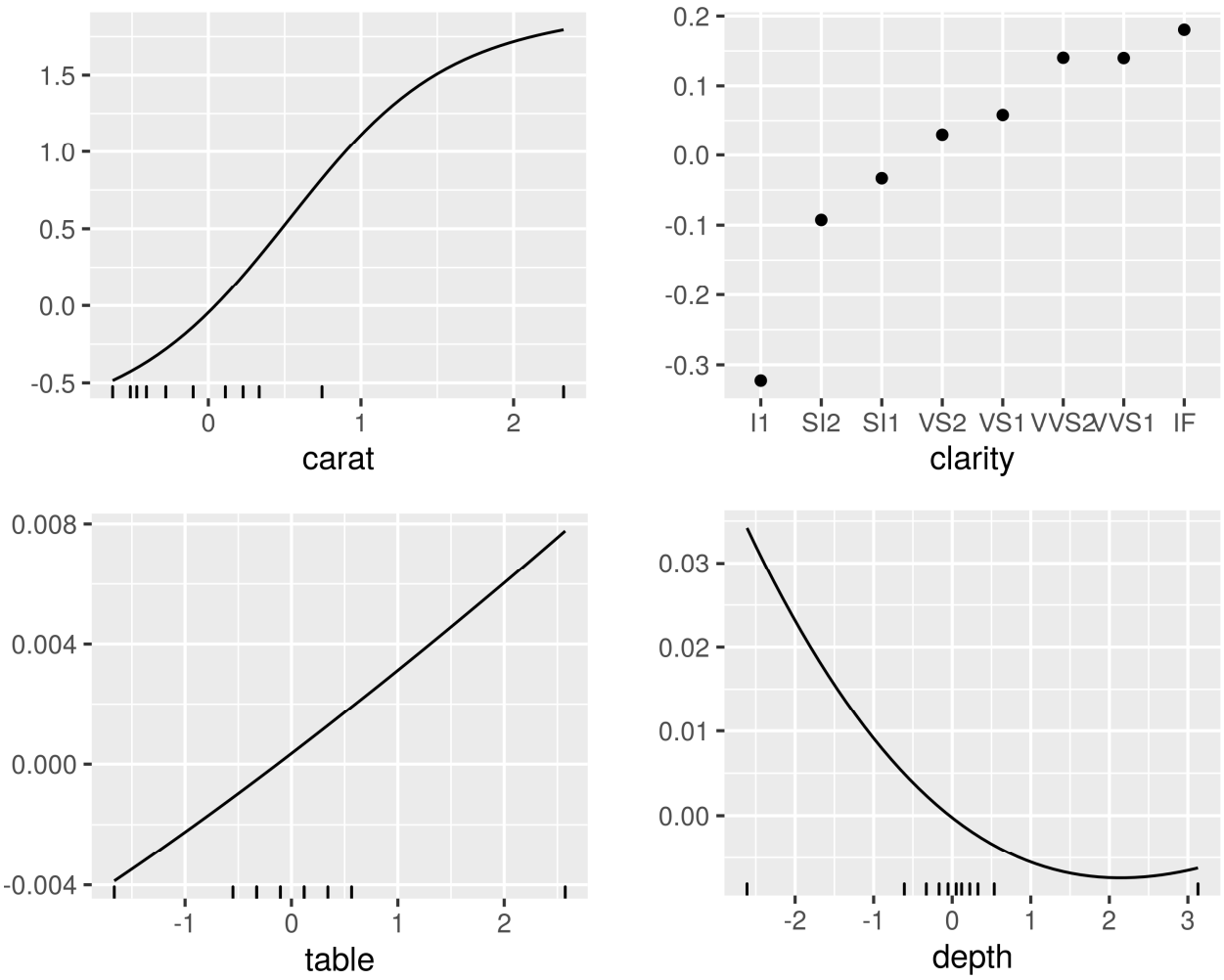

1.5
1.0
0.5
0.0
-0.5
0
1
2
carat
0.2
0.1
0.0
-0.1
-0.2
-0.3
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
clarity
0.008
0.004
0.000
-0.004
-1
0
1
2
table
0.03
0.02
0.01
0.00
-2
-1
0
1
2
3
depth

Рис. 19

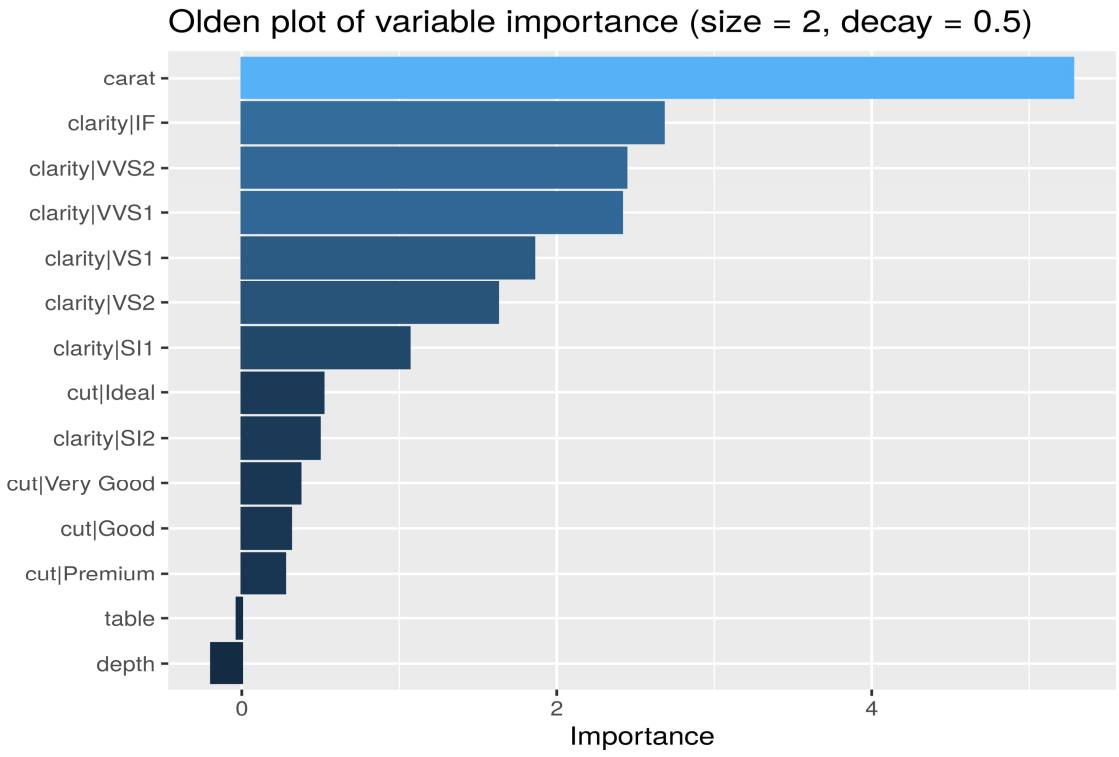

Olden plot of variable importance (size = 2, decay = 0.5)
carat
clarity|IF
clarity|VVS2
clarity|VVS1
clarity|VS1
clarity|VS2
clarity|SI1
cut|Ideal
clarity|SI2
cut|Very Good
cut|Good
cut|Premium
table
depth
0
2
4
Importance

Рис. 20

Model > Trees > Classification and regression trees осуществляет построение деревьев классификации и регрессии, которые становятся одним из наиболее популярных методов решения многих практических задач. При использовании этого метода в соответствии с некоторым набором правил разбиения исходный набор данных рекурсивно делится на подмножества, которые становятся все более и более гомогенными относительно отбираемых на каждом шаге признаков, в результате чего формируется древовидная иерархическая структура.

Перед построением модели необходимо выбрать тип (*Classification* или *Regression*), переменную отклика и одну или несколько объясняющих переменных. Если выбрать те же переменные из таблицы `diamonds`, что и при построении персептрона, то сформированное дерево будет иметь вид, представленный на рис. 21.

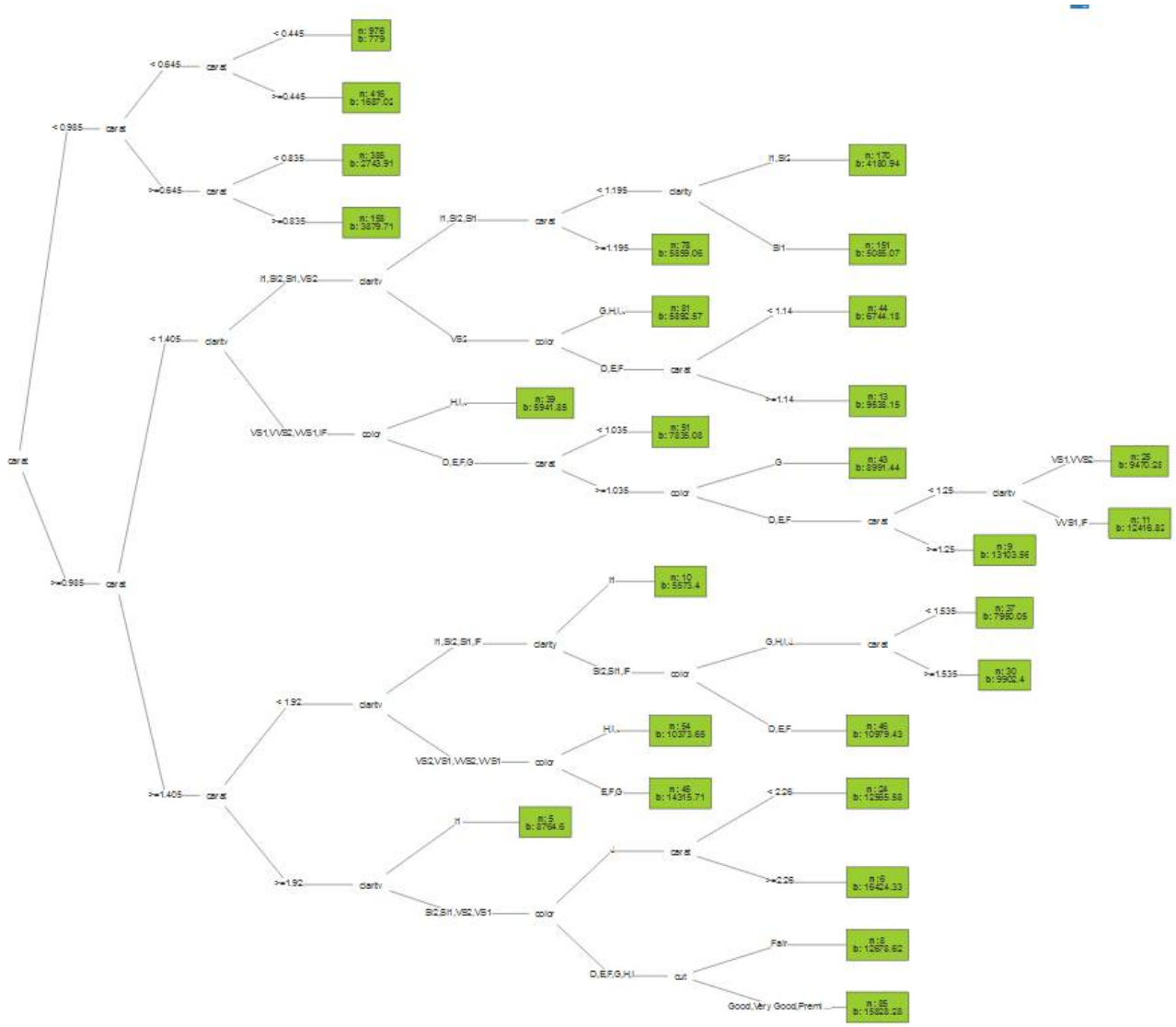


Рис. 21

Визуализировать это дерево можно, выбрав тип графика *Tree* на вкладке *Plot*. Проблема построения деревьев заключается в "жадности" алгоритма, который в стремлении получить более точное решение начинает наращивать дерево избыточными узлами и ветвями. Регулировать "кустистость" можно, задав исходные параметры: штраф за сложность модели (*Complexity*), максимальное число узлов (*Max.nodes*) и штраф за обрезку (*Prune complexity*). Выбор оптимальных значений этих параметров можно выполнить с использованием кросс-проверки – см. график типа *Prune* (рис. 22).

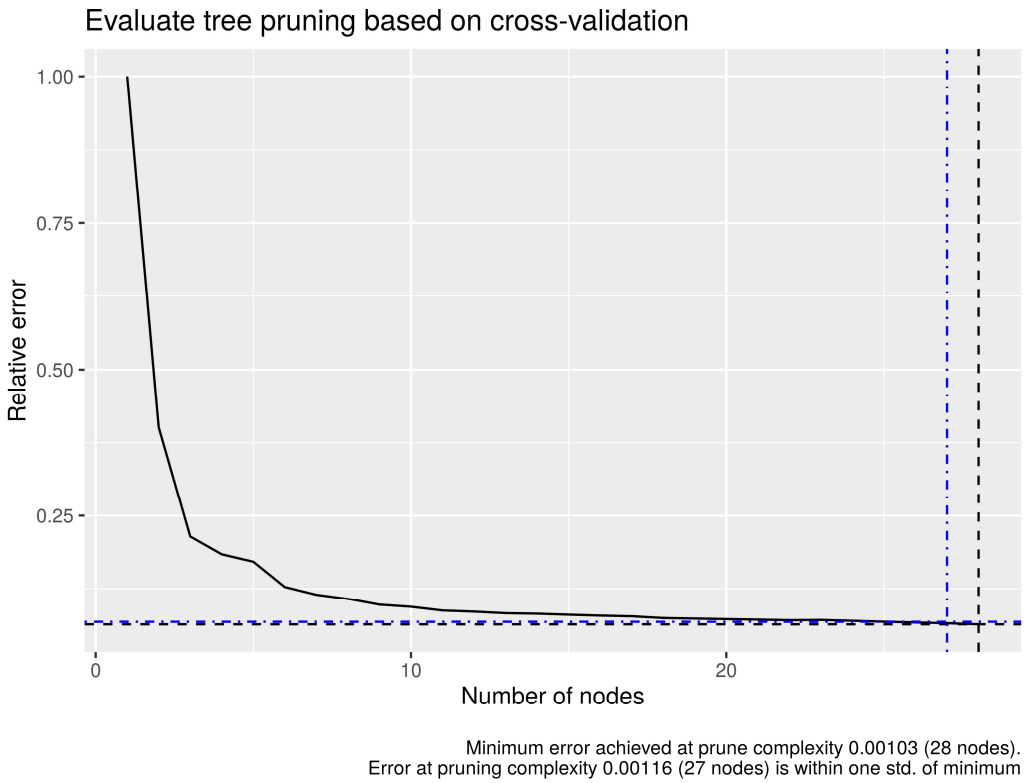


Рис. 22

Вид остальных графиков идентичен приведенным ранее: *Dashboard* (рис. 13), *Partial Dependence* (рис. 19), *Importance* (рис. 20).

Model > Trees > Random Forest. Метод "случайного леса" реализует идею, что если построить ансамбль из несколько сотен (*# trees*) деревьев решений, то эффективность прогнозирования может быть существенно выше, поскольку ослабляется влияние случайных ошибок. Обучение ведется по случайным бутстреп-выборкам и на каждой итерации построения дерева разбиение разрешается выполнять только по небольшому набору *mtry* случайно назначенных переменных.

Перед построением модели необходимо выбрать тип (*Classification* или *Regression*), переменную отклика, одну или несколько объясняющих переменных и задать настроечные параметры алгоритма *# trees, mtry*, *Min node size* и *Sample fraction inputs*. Наилучший путь оценить значения параметров состоит в кросс-проверке.

Model > Trees > Gradient Boosted Trees. Как и *Random Forest*, градиентный бустинг формирует ансамбль деревьев, но их отбор основывается несколько на иной идее. Запускается итеративный процесс последовательного построения частных моделей, но каждое новое дерево обучается в направлении отрицательного градиента ошибок с использованием информации, полученном на предыдущем этапе. Результирующая функция представляет собой линейную комбинацию всего ансамбля моделей с учетом минимизации назначенной штрафной функции.

Перед построением модели необходимо выбрать тип (*Classification* или *Regression*), переменную отклика и одну или несколько объясняющих переменных. Список и смысл настроечных параметров алгоритма представлен на https://xgboost.readthedocs.io/en/latest/parameter.html. Наилучший путь оценить значения параметров состоит в кросс-проверке. В остальном техника построения и форма графической интерпретации моделей мало отличается от ранее представленных методов.

Model > Evaluate > Evaluate regression - сравнительная оценка качества предсказаний моделей. Точность предсказаний каждой модели оценивается по трем показателям: среднему абсолютному отклонению (*MAE*), корню из среднеквадратичного отклонения (*RSME*) и квадрату коэффициента детерминации *Rsq* = 1 - *NSME*, где *NSME* – относительная ошибка, равная отношению средних квадратов отклонений от регрессии к отклонениям от общего среднего.

Вернемся опять к зависимости стоимости бриллиантов `price` от веса в каратах `carat`, глубины `depth` и факторов `cut` и `clarity`, связанных с качеством огранки и прозрачностью алмазов (см. таблицу `diamonds`). Выполним построение моделей регрессии четырьмя описанными выше методами и сохраним результаты прогноза в следующих таблицах: `pred_reg` для обычной регрессии, `pred_nn` для нейронной сети, `pred_rf` для случайного леса и `pred_gbt` для градиентного бустинга.

Нажатие кнопки *Evaluate models* позволит выполнить расчет критериев оценки и построить график сравнения эффективности прогноза для этих моделей – рис. 23.

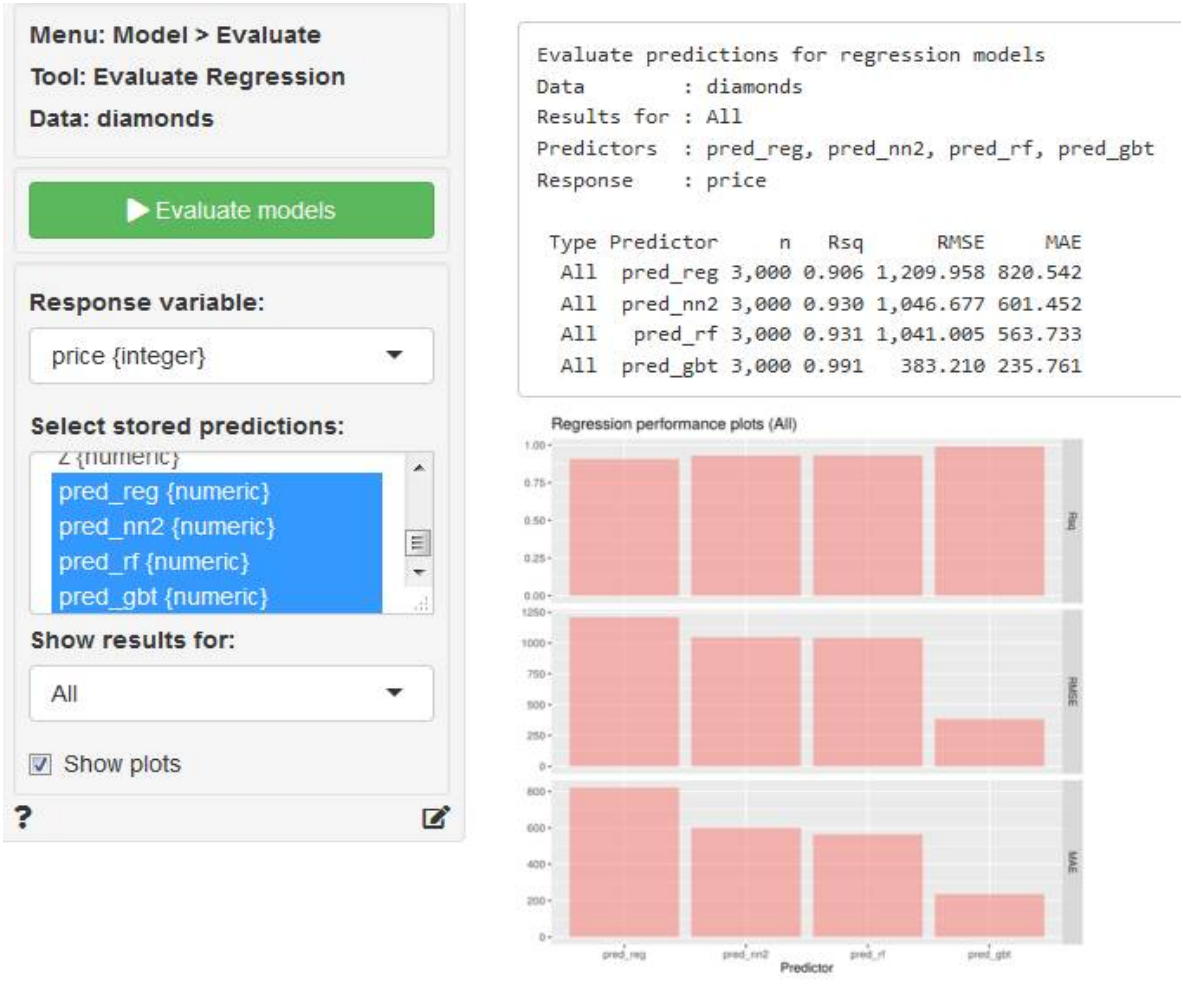


Рис. 23

Очевидно значительное преимущество модели градиентного бустинга над всеми остальными моделями по всем использованным критериям. Представленные результаты следует рассматривать как учебный пример и не принимать за аксиому необходимость повсеместного использования бустинга. Во-первых, при построении моделей необходимо проводить вдумчивую настройку входных параметров, во-вторых, оценка эффективности прогноза на той же выборке, по которой строилась модель – плохая практика. Истинная эффективность метода оценивается только на данных независимых выборок или имитационными методами.

Model > Evaluate > Evaluate classification **-** оценка качества выполняемых прогнозов уровней категориальной переменной отклика. Эффективность прогноза каждой модели классификации оценивается по 15 показателям, основанным на матрице неточностей (*Confusion matrix*), и 4 графикам.

Вернемся опять к зависимости числа выживших в крушении "Титаника" (уровень `Yes` фактора `survived` таблицы `titanic`) от класса каюты `class`, пола `sex`, и возраста пассажиров `age`. Выполним построение моделей классификации тремя методами и сохраним результаты прогноза в следующих таблицах: `pred_reg` для логистической регрессии, `pred_rf` для случайного леса и `pred_gbt` для бустинга.

Отметим, что, если активизирован фильтр (например, установлен в ***Data > View***), то используя раскрывающийся список *Show results*, мы можем генерировать результаты тестирования для различных подмножеств строк *All*, *Training*, *Test* и *Both*, т.е. всей, обучающей и тестовой выборок.

Нажатие кнопки *Evaluate models* позволит выполнить расчет критериев оценки эффективности прогноза для этих моделей во вкладке *Confusion* – рис. 24.

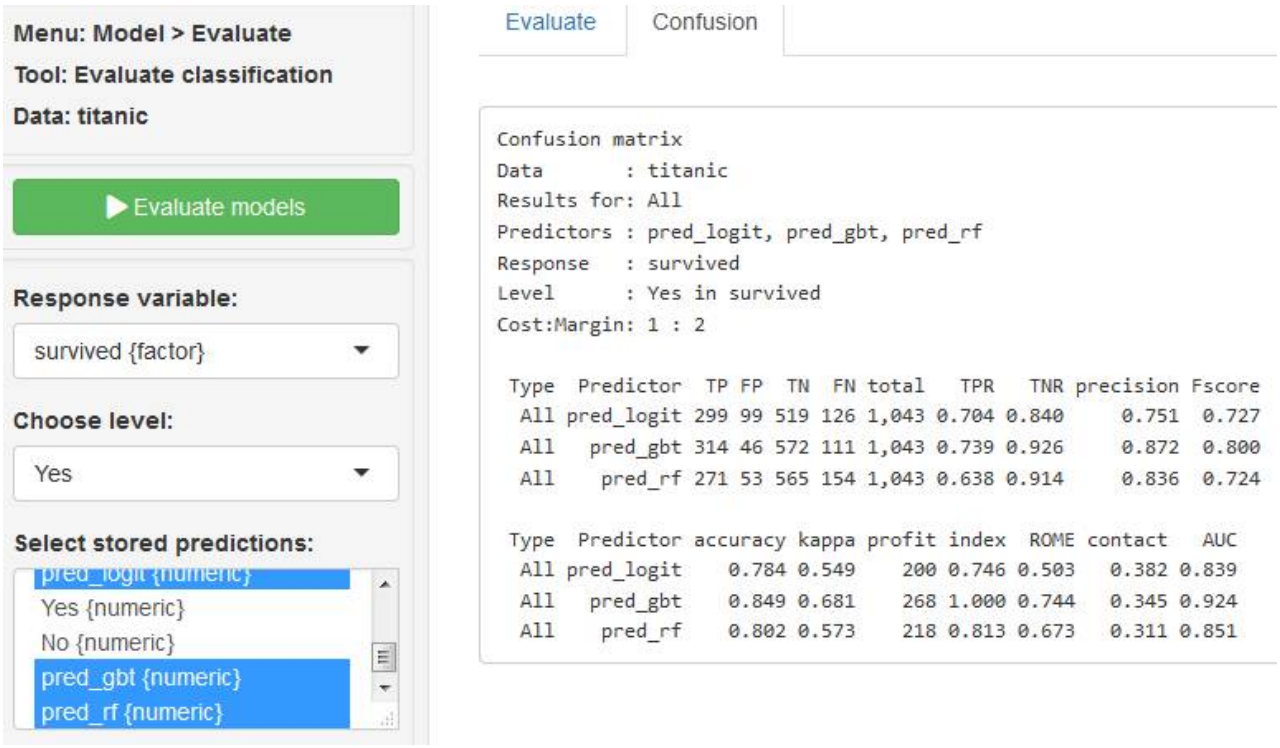


Рис. 24

Установив соответствующие флажки, можно построить графики сравнения моделей типа *Lift*, *Gains*, *Profit* и *ROME* (см. рис. 25). Подробный рассказ о смысле этих показателей и графиков потребует расширенного повествования, поэтому предоставим читателям разобраться с ними самостоятельно. Отметим только, что, по нашему мнению, наиболее полезными являются кривая *Cumulative Gains* и численная оценка площади под ROC-кривой *AUC* (*Area Under Curve*). Практически *AUC* изменяется от 0.5 ("бесполезный" классификатор) до 1.0 ("идеальная" модель). По этим показателям наилучшую точность прогноза опять имеет градиентный бустинг.

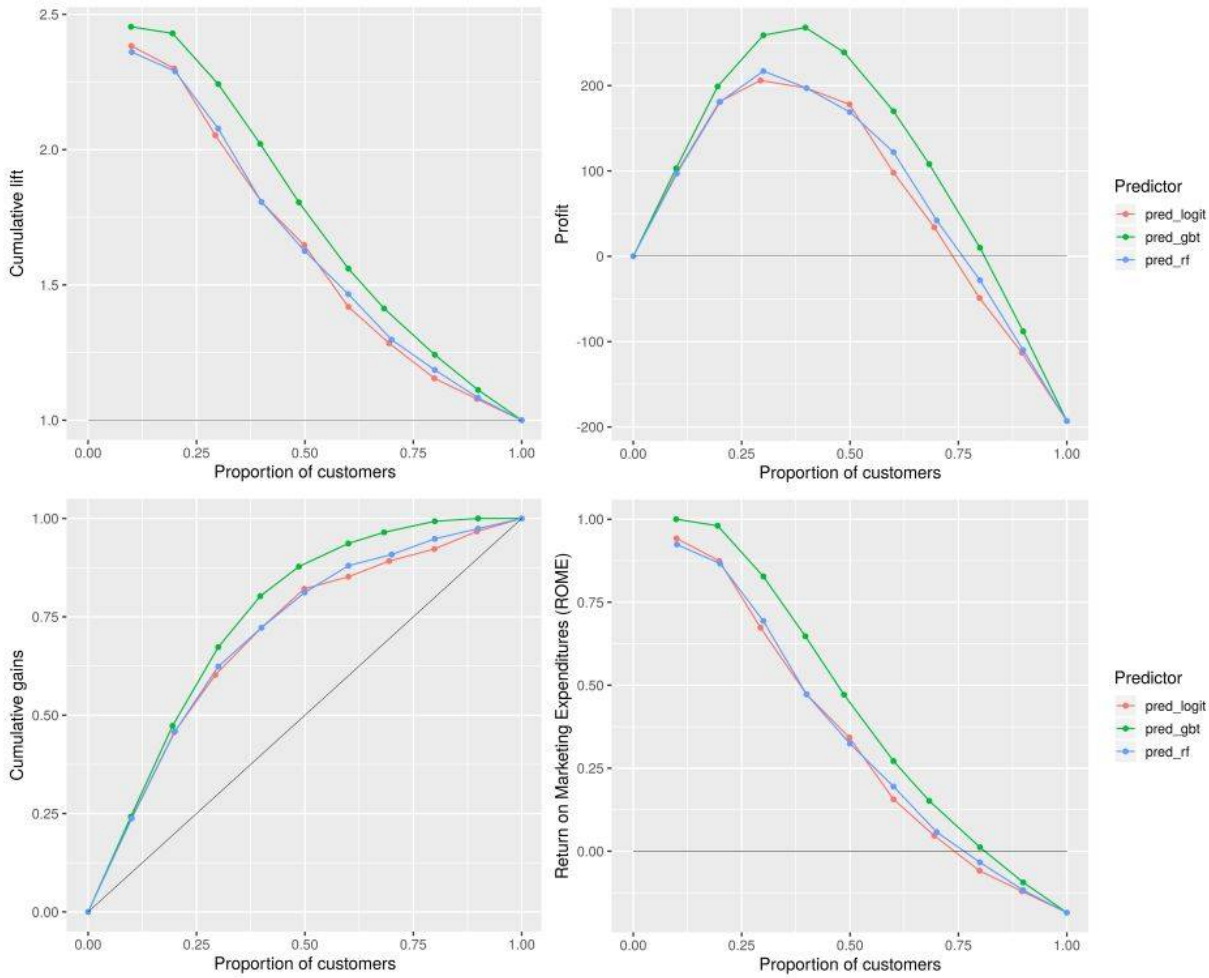


Рис. 25

Model > Recommend > Collaborative Filtering. Коллаборативная (или совместная) фильтрация – это один из методов построения прогнозов в "рекомендательных" системах, использующий известные предпочтения (оценки) группы пользователей для прогнозирования неизвестных предпочтений другого пользователя. Его основное допущение состоит в следующем: те, кто одинаково оценивал какие-либо объекты в прошлом, склонны давать похожие оценки другим предметам и в будущем.

В *Radiant* для построения моделей можно использовать примеры файлов, загружаемых с вкладки о страницы ***Data > Manage*** (нажмите переключатель *examples* и нажмите кнопку *Load*). Воспользуемся этим и загрузим пример `ratings`, в котором 10 независимых пользователей `Users` дали оценки привлекательности `Ratings` по 5-бальной шкале 10 видам товара `Movies`. Поле `training` == 1 с установленным фильтром (в ***Data > View***) показывает, что отфильтрованные пользователи образуют обучающую выборку.

Перед началом генерации рекомендаций необходимо задать имена полей *User id* <– `Users`, а *Product id* <- `Movies`, *Ratings variable* <- `Ratings`, *Choose products recommendations* <- одно или несколько уровней переменной *Product id* , по которым ведется прогнозирование. Рекомендации по выбору товаров даются по пользователям с `training` == 0, т.е. пользователю `U11`, в форме, представленной на рис. 26.

```
Collaborative filtering
Data       : ratings
Filter     : training==1
User id    : Users
Product id : Movies
Predict for: M6, M7, M8, M9, M10

Recommendations:

 Users product rating average   cf ranking avg_rank cf_rank
   U11      M6           3.30 4.10                3       1
   U11      M7           2.70 2.08                5       4
   U11      M8           3.50 1.70                2       5
   U11      M9           2.90 2.13                4       3
   U11     M10           4.10 2.71                1       2
```

Рис. 26

Model > Decide > Decision analysis - Создание и оценка иерархических деревьев для анализа решений.

Чтобы создать и оценить дерево решений, нужно ввести описание его ветвей и узлов в редакторе входных данных или загрузить иерархическую структуру из файла. При первом переходе на вкладку ***Decision analysis*** загружается шаблон древовидной структуры, основанной на примере Кристофа Глура, разработчика библиотеки `data.tree`. Подробные правила ввода новой структуры и заполнения отдельных полей дерева решений приведены на экране помощи (нажмите **?**).

Полученные решения и соответствующие им вероятности представлены во вкладках *Model* (см. рис. 27) и *Plot* в виде древовидной структуры, подобной рис. 20.

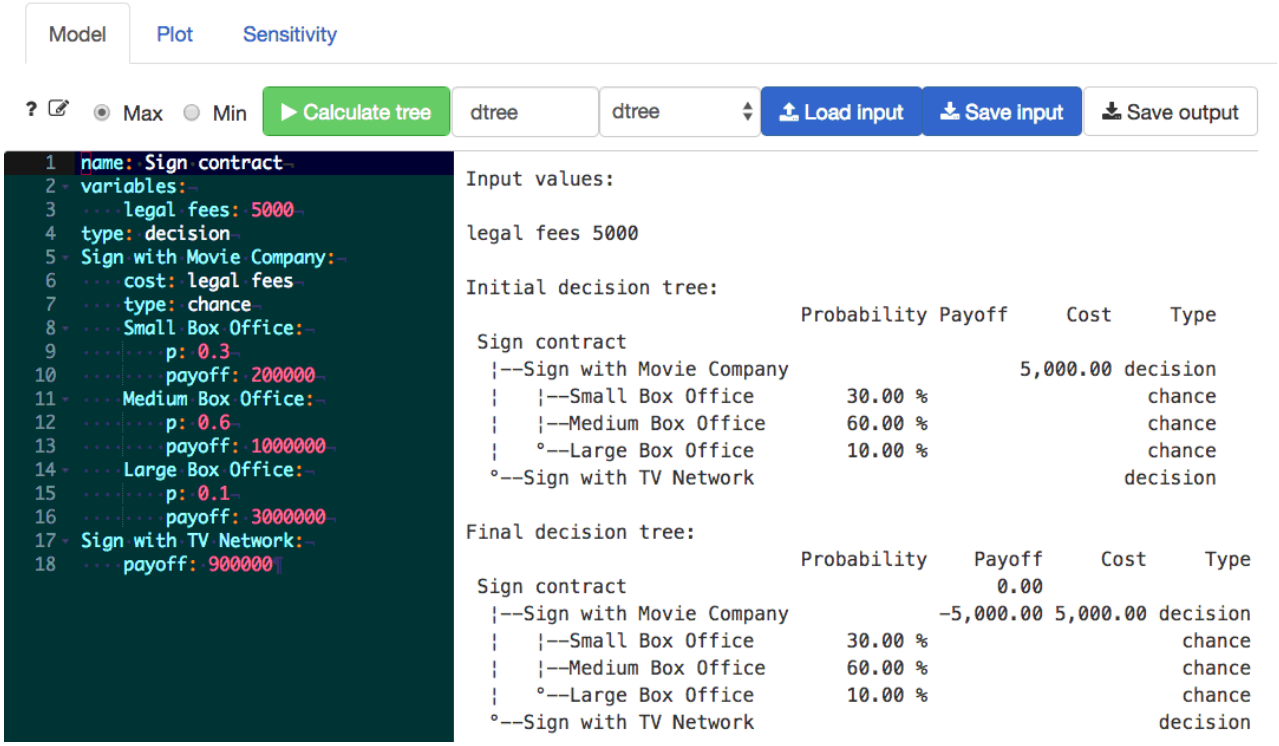


Рис. 27

После нажатия кнопки *Evaluate sensitivity* ("Оценить чувствительность") будет показан график, иллюстрирующий, как меняются издержки и доход в результате принятых решений.

Model > Decide > Simulate **-** использование имитационных методов для обоснования экономических решений. В документации приводится подробный разбор построения различных имитационных моделей для исследования зависимости спроса и прибыли от цены (моделируемая переменная). Окончательная формула используется для определения количества (и доли) случаев, когда прибыль составит величину ниже заданного предела.

6. Меню *Multivariate* (Многомерный анализ)

Этот раздел меню состоит из следующих подразделов и пунктов:

- ***Maps*** – карты:
  - (*Dis)similarity* – карты сходства/несходства;
  - *Attributes* – атрибутивные карты.
- ***Factor*** - факторный анализ:
  - *Pre-factor* – предварительные тесты;
  - *Factor* – факторный анализ.
- ***Cluster*** - кластерный анализ:
  - *Hierarchical* – иерархическая кластеризация;
  - *K-clustering* – группировка по методу *k*-средних.
- ***Conjoint*** – метод изучения эластичности спроса:
- *Conjoint* – так он и есть.

Многомерный анализ включает два основных направления снижения размерности данных: **ординацию**, использующую различные методы проецирования, и **кластеризацию**, выполняющую разбиение объектов на группы. Оптимальное целенаправленное проецирование (*projecting pursuit*) облака точек из многомерного пространства в пространство малой размерности заключается в представлении исходной матрицы данных **X** в виде совокупности *p* латентных переменных **F**:

$$X_1, X_2, \ldots, X_m \Rightarrow F_1, F_2, \ldots, F_p,$$

которые и являются осями ординационной диаграммы (двумерной при $p = 2$ или трехмерной с тремя такими латентными переменными). Выбор осей ординации осуществляется с использованием принципа оптимальности: т.к. стремления достичь минимума потерь содержательной информации, имеющейся в исходных данных.

Непрямая ординация, которая отчасти представлена разделами меню ***Maps*** и ***Factor***, представляет собой совокупность методов обучения без учителя, основанных, как правило, на анализе расстояний между всеми возможными парами объектов в пространстве наблюдаемых независимых признаков. Используемые алгоритмы (методы главных компонент и главных координат, анализ соответствия, многомерное неметрическое шкалирование и др.) выполняют редукцию данных с учетом минимально возможного искажения исходной взаимной упорядоченности точек, что обеспечивает наглядное графическое представление геометрической метафоры исследуемых объектов.

Multivariate > Maps > (Dis)similarity – выполняет многомерное неметрическое шкалирование (*Multi-Dimensional Scaling - MDS*) матрицы дистанций и визуализацию ординационной диаграммы (*Map*). К сожалению, меры расстояний в этом пункте меню рассчитываются как функции только одной переменной. Например, таблица `city` из раздела *examples* содержит расстояния `distance` между каждой парой городов США `from` и `to`. Зададим эти условия на боковой панели и выполним анализ, основные результаты которого показаны на рис. 28.

```
Original distance data:
        Boston   NY   DC Miami Chicago Seattle   SF   LA
NY         206
DC         429  233
Miami     1504 1308 1075
Chicago    963  802  671  1329
Seattle   2976 2815 2684  3273    2013
SF        3095 2934 2799  3053    2142     808
LA        2979 2786 2631  2687    2054    1131  379
Denver    1949 1771 1616  2037     996    1307 1235 1059

Recovered distance data:
         Boston      NY      DC   Miami Chicago Seattle      SF      LA
NY       216.17
DC       424.34  209.27
Miami   1481.04 1320.47 1159.80
Chicago  964.17  781.64  649.65 1431.60
Seattle 2950.14 2814.77 2720.03 3271.40 2077.26
SF      3103.29 2929.07 2787.14 3054.27 2147.64  777.59
LA      2992.97 2800.55 2634.90 2728.86 2030.29 1207.14  488.18
Denver  1930.70 1750.84 1606.47 2015.18  969.35 1256.50 1180.68 1089.01
```

```
Coordinates:
        Dimension  1 Dimension  2
Boston      -1348.67      -462.40
NY          -1198.87      -306.55
DC          -1076.99      -136.43
Miami       -1226.94      1013.63
Chicago      -428.45      -174.60
Seattle      1596.16      -639.31
SF           1697.23       131.69
LA           1464.05       560.58
Denver        522.49        13.40

Stress: 0.02
```

Рис. 28

Вряд ли имеется смысл в таком анализе *MDS*, поскольку для любой одномерной метрики исходные расстояния на карте (*original distances*) и расстояния между точками на плоскости (*recovered distances*) можно считать эквивалентными. Но, по причине изогнутости поверхности земли или ошибки сходимости алгоритма, между значениями этих матриц дистанции всё же имеются 2% расхождения (*Stress* = 0.02). В окончательном итоге ординационная диаграмма на рис. 29 с нанесенными на нее названиями городов (Бостон и Лос-Анджелес, Сиэтл и Майями), построенная по координатам на *MDS*–шкалах, является полным аналогом географической карты США.

Рис. 29

Multivariate > Maps > Attributes – осуществляется анализ главных компонент (*Principle component analysis – PCA*), в котором в качестве матрицы сходства между объектами используется корреляционная матрица Пирсона. В примере на рис. 30 таблица наблюдений включает 7 переменных (в терминологии разработчиков – *Attributes*) и 6 объектов (*Brand*), поименованных в поле `retailer`.

В *Summary* приводится стандартный набор таблиц *PCA*-анализа. В таблице *Fit measures* ("Показатели подгонки") дается информация о том, какая доля вариации исходных данных объясняется каждой главной компонентой (фактором). Первый фактор объясняет 56,4%, второй фактор – 42% , а оба вместе – 98,4% совокупной дисперсии, то есть потеря информации от уменьшения размерности данных с 7 переменных до двух факторов составляет всего 1,6%. Таблица *Attributes communalities* показывает, какая доля дисперсии каждой из исходных переменных объясняется обеими факторами. В таблице *factor scores* приводятся факторные оценки для объектов, показывающие уровень их связи с латентными переменными. Таблица *factors loadings* показывает величины факторных нагрузок, т.е. уровень корреляции между исходными и латентными переменными.

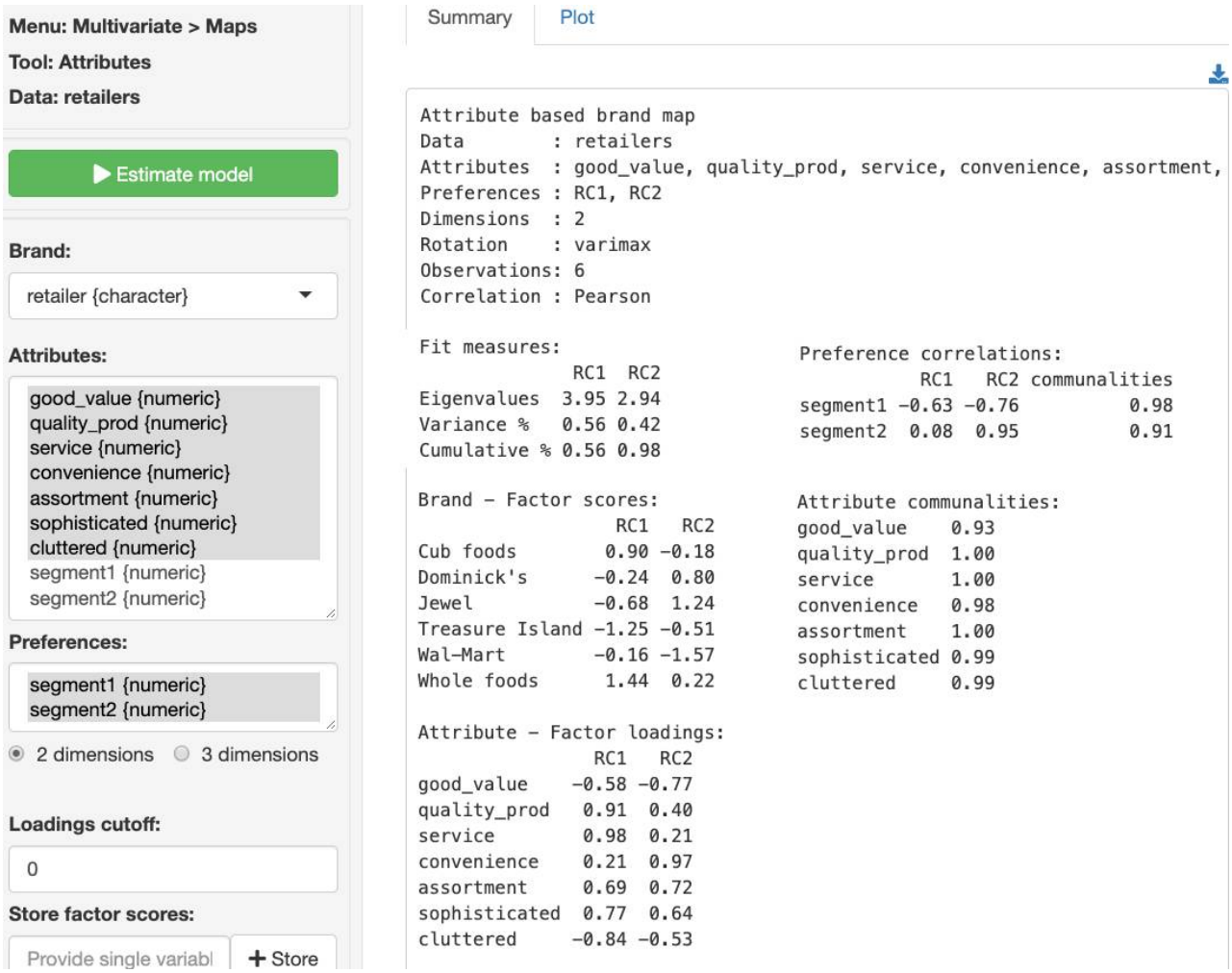


Рис. 30

Некоторым расширением стандартного алгоритма *PCA* является включение в анализ дополнительных переменных, называемых "предпочтительными" (*preferences*). Для этих переменных вычисляются корреляции с каждым из латентных факторов (и обеими вместе).

Ординационная диаграмма переменных и объектов (биплот) приведена на рис. 31.

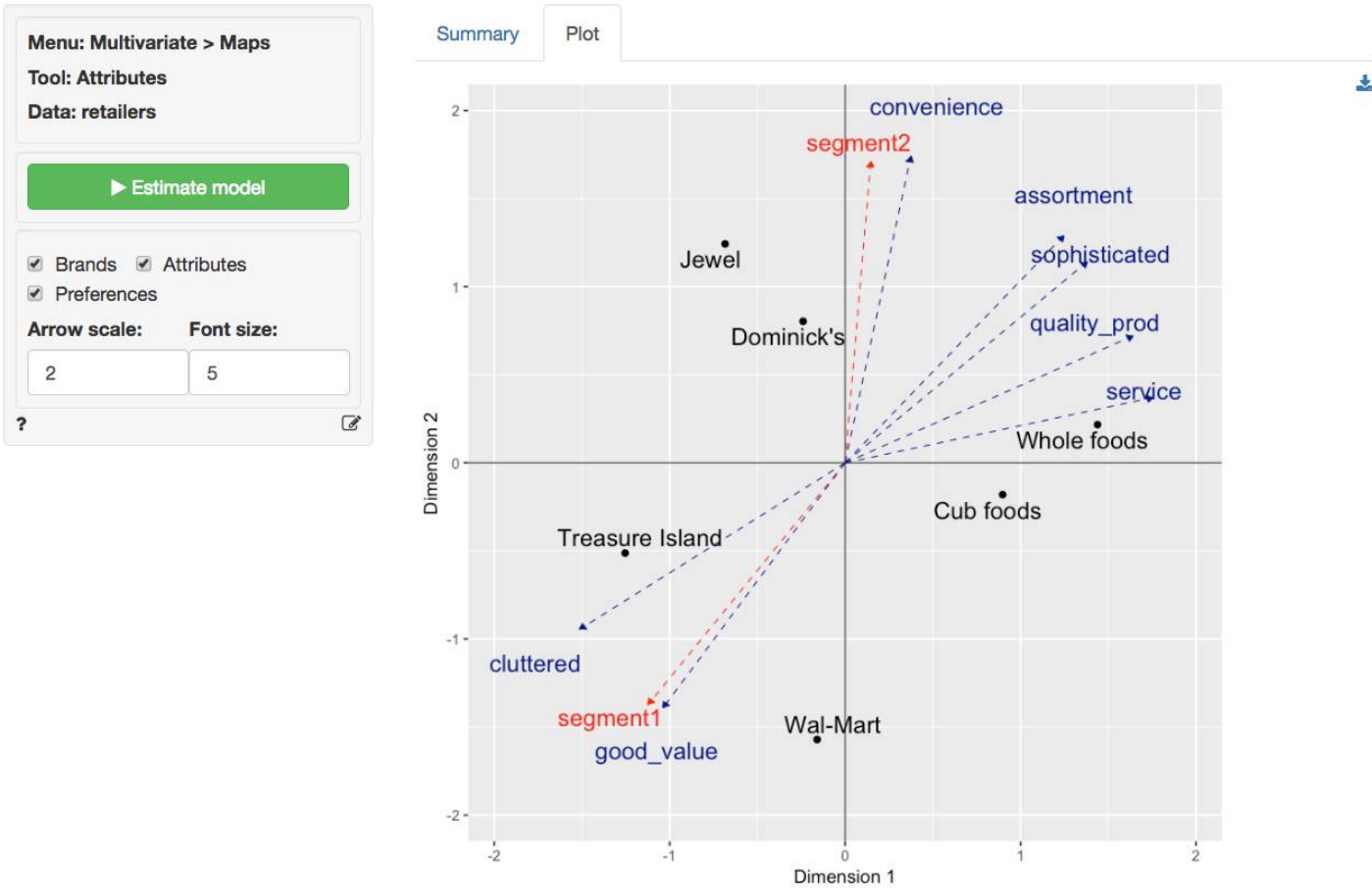


Рис. 31

Multivariate > Factor > Pre-factor **-** выполняется несколько тестов, связанных с начальными предпосылками анализа главных компонент и его сходимостью к удовлетворительному решению. В частности, поскольку в один фактор объединяются только сильно коррелирующие между собой переменные, данные должны быть коллинеарными.

Тесты Бартлетта и Кайзера-Мейера-Олкина (*KMO*) оценивают степень взаимной зависимости исходных переменных: уровень значимости для теста Бартлетта ниже 0,05 и значение *KMO* более 0,5 предполагают наличие существенной корреляции в данных. В тесте *Collinearity* ("Коллинеарность") для каждой исходной переменной значения выше 0,4 считаются приемлемыми.

Другой проблемой является определение количества факторов, необходимых для описания структуры, лежащей в основе данных. Для этого используется график "каменной осыпи", т.е. график собственных значений по отношению к числу факторов в порядке их извлечения. Иногда в этом сюжете наблюдается перелом или локтевой сгиб. Наблюдательный человек всегда найдет его там, где ему это предпочтительно.

Multivariate > Factor > Factor **-** этот пункт меню лишь в следующих деталях отличается от ***Multivariate > Maps > Attributes*** **:**

- кроме классического PCA, анализ может проводиться с использованием метода максимального правдоподобия (*Maximum Likelihood*);
- больше не упоминаются "предпочтительные" переменные (*preferences*);
- добавлена возможность поворота (*Rotate*) ординационых диаграмм, например, варимаксное вращение (*Varimax*).

Multivariate > Cluster > Hierarchical – выполняет последовательное разбиение исходных объектов на группы с использованием иерархической кластерной процедуры. Для выполнения анализа необходимо указать список включаемых переменных и меру для подсчета дистанции между каждой парой объектов (*Distance measures*). Можно использовать 8 различных формул сходства/расстояния: Эвклида, Хемминга, городских кварталов, Гувера, Минковского, суммы минимумов и др. Необходимо также указать один из семи возможных алгоритмов кластеризации (Уорда, центроидный, средней связи, полной связи и т.д.). Для визуализации результатов есть две возможности: дендрограмма (см. рис. 32) и графики для оценки оптимального числа классов.

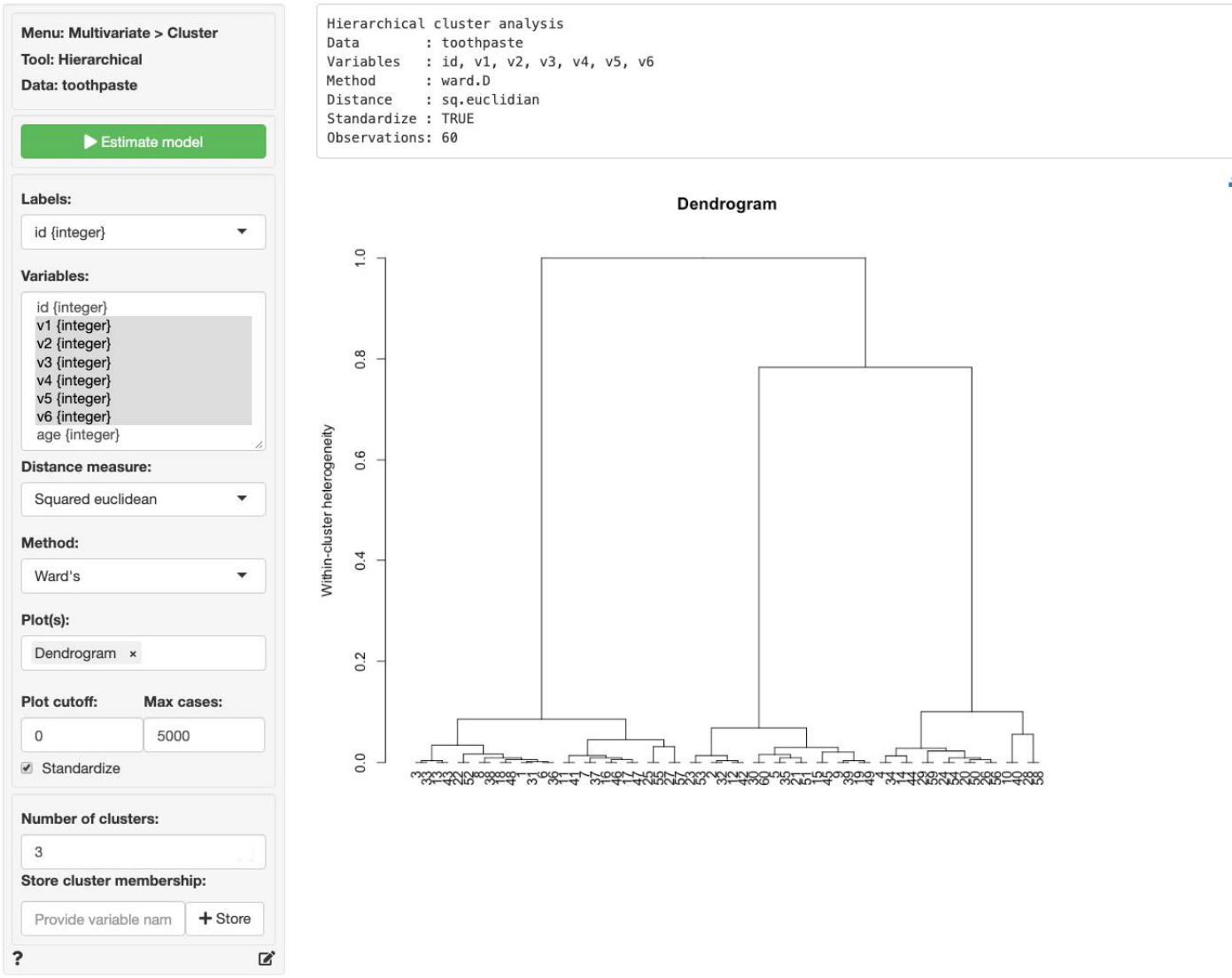


Рис. 32

Во второй группе графиков оценивается изменение показателя внутриклассовой дисперсии по мере разбиения исходной совокупности на группы. На каком-то этапе можно усмотреть, что темпы снижения этой дисперсии резко уменьшаются – на графике *Scree* (родственник "каменистой осыпи" для анализа главных компонент) это происходит в точке, называемой "локтем". Если отсечка графика установлена в 0, то мы видим результаты для всех возможных кластерных решений. Другой график *Change* показывает динамику изменения внутриклассового разброса в виде столбчатой диаграммы.

Multivariate > Cluster > K-clustering. Алгоритмы неиерахического разделения (*partitioning algorithms*) осуществляют декомпозицию набора данных, состоящего из *n* наблюдений, на заранее известное число *k* групп (кластеров). При этом выполняется поиск *центроидов* – максимально удаленных друг от друга центров сгущений точек C_k с минимальным разбросом внутри каждого кластера. *Radiant* использует разделяющие алгоритмы *k*-средних Мак-Кина (*k-means*) и *k-prototypes*. Исходные данные и параметры задаются по той же схеме, что и при иерархической кластеризации, но указание числа классов *k* (*Number of clasters*) является обязательным.

На вкладке *Summary* показаны основные результаты при разделении на 3 класса: заселенность классов, средние значения и доля внутриклассовой дисперсии исходных переменных в каждом классе. Кривые плотности распределения исходных переменных для каждого класса можно получить на вкладке *Plot* – см. рис. 33.

```
K-means cluster analysis
Data         : toothpaste
Variables    : v1, v2, v3, v4, v5, v6
Clustering by: K-means
Standardize  : TRUE
Observations : 60
Generated    : 3 clusters of sizes 18 | 26 | 16

Cluster means:
             v1   v2   v3   v4   v5   v6
 Cluster 1 1.67 3.00 1.89 3.44 5.56 3.22
 Cluster 2 5.85 3.38 6.15 3.38 1.92 3.77
 Cluster 3 3.38 5.75 3.25 6.00 3.75 5.88

Percentage of within cluster heterogeneity accounted for by each cluster:

Cluster 1 24.94%
Cluster 2 39.54%
Cluster 3 35.52%

Between cluster heterogeneity accounts for 73.04% of the
total heterogeneity in the data (higher is better)
```

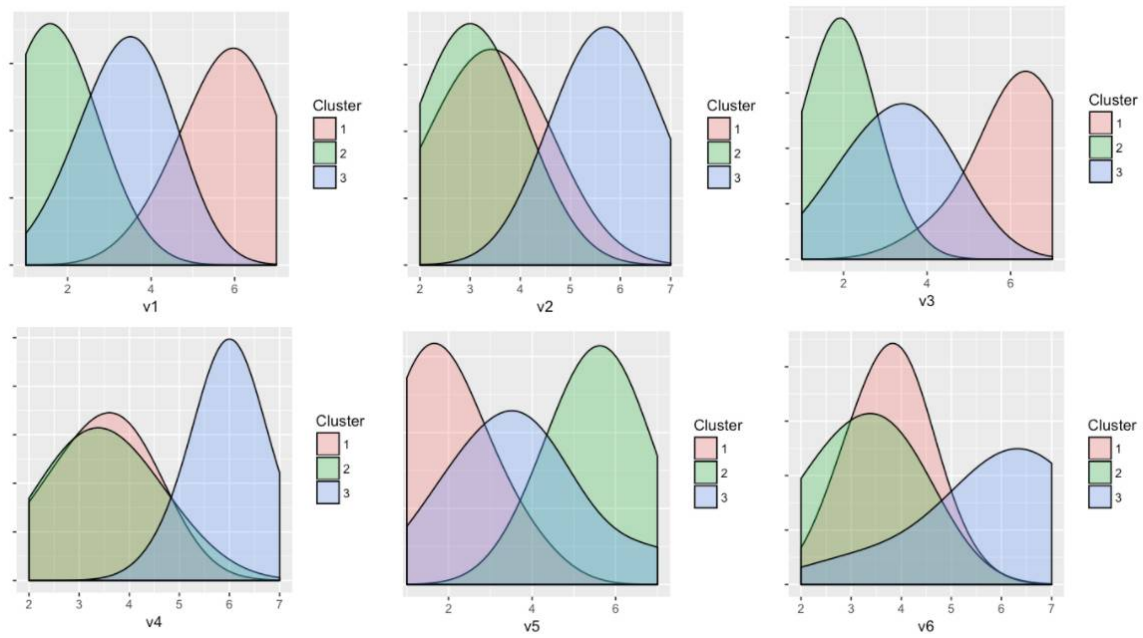


Рис. 33

Multivariate > Conjoint > Conjoint. В *conjoint*-анализе (т.е. совместном анализе – ***con****sider* ***joint****ly*) объекты условно описываются набором факторов (например, для товара это размер упаковки, цвет, масса, цена и т.д.), а каждый фактор имеет несколько уровней. Максимальное число объектов со всеми возможными сочетаниями уровней факторов равно произведению числа уровней и их количество может получиться слишком большим. Поэтому используются различные экспериментальные планы ***Design > Design of Experiments*** из разных уровней факторов, позволяющие оптимизировать количество комбинаций и сформировать набор прототипов, представляющих собой различные возможные варианты свойств объектов.

Респонденту предоставляется набор объектов с разными характеристиками и предлагается оценить, насколько вероятно, что он воспользуется каждым из представленных предложений. С помощью регрессионного анализа рассчитываются полезности уровней фактора, показывающие, насколько наличие того или иного уровня влияет на общую оценку объекта. В итоге формируется рейтинг объектов, оптимальных с точки зрения общей полезности по совокупности всех их свойств и одновременно не слишком дорогих и конкурентоспособных.

7. Использование Radiant в статистической среде R

*7.1. Инсталляция и запуск **Radiant** в **R***

Как отмечалось в начале сообщения, *Radiant* можно использовать он-лайн без установки пакета и самой статистической среды R на своем компьютере по ссылке:

https://vnijs.shinyapps.io/radiant

Для этого достаточно современного интернет-браузера (например, *Internet Explorer* вер. 11 или выше, *FireFox*, *Chrome* или *Safari*).

Однако есть много соображений, чтобы развернуть модули *Radiant* на своем компьютере (или в локальной сети своей организации):

- При работе он-лайн вы помещаете свои, возможно, конфиденциальные данные на общедоступный сервер (объем загружаемых данных ограничен 10 МБ по соображениям безопасности).
- Скорость обработки данных на вашем компьютере может оказаться выше, а локализация ошибок или непредвиденных остановок становится более удобной (это особенно важно при выполнении трудоемких вычислений).
- Вы получаете доступ к исходным модулям текстов и будете понимать, как все это работает. Можно откорректировать некоторые скрипты, чтобы выполнять, например, построение графиков по вашему вкусу или нужной палитре цветов.
- Наконец, можно дополнить список функций *Radiant* необходимыми вам модулями или собственными разработками, т.е. по сути создать собственную версию *Radiant*.

Для установки `Radiant` на своем компьютере необходимо развернуть статистическую среду R вер. 3.3.0 или выше. Пакет целиком разработан на базе Shiny и весьма желательно иметь начальные представления об этом средстве разработки веб-приложений (см. наше сообщение https://stok1946.blogspot.com/2021/01/shiny.html).

Пакет `Radiant` доступен в CRAN и может быть установлен обычным порядком, например, подачей команд:

```
options(repos = c(RSM = "https://radiant-rstats.github.io/minicran",
                  CRAN = "https://cloud.r-project.org"))
install.packages("radiant")
```

В ходе инсталляции проверяется комплектность среды R и устанавливаются или обновляются все зависимые пакеты. В нашем случае установка потребовалась для 36 новых пакетов: `pillar, cpp11, lifecycle, tidyselect, vctrs, downloader, influenceR, visNetwork, tibble, tidyr, dplyr, shinyAce, writexl, shinyFiles, randomizr, patchwork, AlgDesign, polycor, NeuralNetTools, data.tree, DiagrammeR, broom, xgboost, pdp, GPArotation, clustMixType, radiant.data, radiant.design, radiant.basics, radiant.model, radiant.multivariate, import.`

Используйте команду `update.packages("radiant")` для обновления пакета.

Если установка прошла успешно, то запустить `Radiant` с отображением формы такой же веб-страницы, что и в он-лайн, можно штатным способом командами:

```
library(radiant)
radiant()
```

*7.2. Некоторые пояснения по работе с **Radiant***

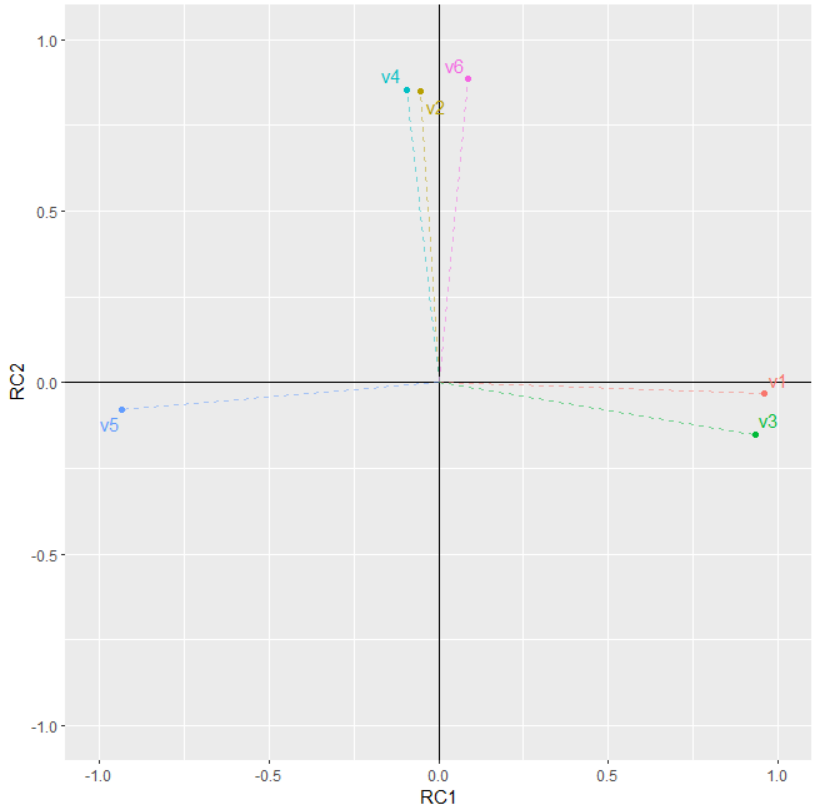
Сохранение и загрузка текущего состояния. При практической работе важно использовать файлы текущего состояния *Radiant*, функции сохранения и загрузки которых доступны в главном меню за значком 🖬 или во вкладке ***Data > Manage***. Чтобы сохранить текущие результаты анализа и настройки приложения в файл, нажмите на кнопку *Save radiant state file*. Сохранять файл состояния при работе он-лайн можно непосредственно на сервере без указания имени и тогда при повторном запуске приложения вы сразу оказываетесь в той точке выполнения расчетов и с теми настройками, которые были на момент сохранения.

Если сохранить файл состояния на локальном компьютере, то тогда можно открыть этот файл позже или на другом компьютере, используя пункт меню *Load radiant state file*. Вы тогда получите сразу доступ ко всем сохраненным объектам или настройкам и сможете продолжить работу с того места, где в свое время остановили вычисления. Файл состояния состоит из четырех разделов (списков): *input*, *r_data*, *r_info* и *r_state*, со структурой которых можно познакомиться, выполнив пункт меню *View radiant state*. Можно также поделиться этим файлом с другими пользователями, которые могут захотеть повторить ваши расчеты.

Доступ к функциям. Веб-интерфейс, поддерживаемый средствами *Shiny*-приложения *Radiant*, дает возможность решать широкий набор задач анализа данных, не прибегая фактически к использованию каких-либо кодов команд. Однако установка пакетов `Radiant` одновременно дает доступ ко всем его функциям и примерам, и вы можете предпочесть обратиться к ним напрямую, запустив соответствующий скрипт, в частности, непосредственно в интерактивной среде R. Например, анализ главных компонент можно осуществить как через пункты меню ***Multivariate > Factor > Factor*** (см. раздел 6), так и подав команды

```
library(radiant)
data(toothpaste)
result <- full_factor(toothpaste", nr_fact = 2,
           vars = c("v1", "v2", "v3", "v4", "v5", "v6") )
summary(result, cutoff = 0.1)
Factor loadings:
     RC1   RC2
v1  0.96
v2        0.85
v3  0.93 -0.15
v4        0.85
v5 -0.93
v6        0.88
```

```
Fit measures:
               RC1  RC2
Eigenvalues   2.69 2.26
Variance %    0.45 0.38
Cumulative % 0.45 0.82
plot(result, custom = FALSE)
```



Список некоторых функций и формат обращения к ним представлен на https://radiant-rstats.github.io/radiant.data/reference/index.html или в файлах помощи, а примеры обращения к ним – на https://radiant-rstats.github.io/docs/programming.html

Radiant обеспечивает мост к программированию в R путем экспорта текущего списка вызовов функций, используемых для анализа. Например, можно запустить веб-приложение *Radiant* и увидеть, что большинство страниц имеют значок 🖉 в левом нижнем углу экрана, при щелчке на который создается отчет по проведенному анализу на вкладке ***Report > R***. Эти команды можно скопировать и вставить в командную консоль, чтобы получить тот же результат, что и на веб-странице браузера.

Доступ к исходным кодам. *Radiant* , разумеется, не является в значительной мере полной и совершенной программой. Каждый статистический аналитик будет разочарован, если не увидит в составе пакета какого-нибудь привычного критерия или любимого метода обработки данных. Нам, в частности, недостает функций моделирования временных рядов, метода опорных векторов или прокрустова анализа. Ужасно раздражает эстетика графиков *ggplot* с мутно-серыми панелями и павлиньими цветами точек – рисунки с таким оформлением не примет ни один научный журнал. Но упаси нас бог упрекать в чем-то "пианистов" – они сыграли нам прекрасную пьесу. И, если вы заинтересованы в том, чтобы внести свой вклад в *Radiant* или расширить его под свои вкусы, взгляните на исходные коды функций пакета на *GitHub* по адресу https://github.com/radiant-rstats, который в свою очередь состоит из нескольких разделов *radiant.data*, *radiant.design*, *radiant.basics* и других, частично знакомых нам по предыдущему изложению. Изучив, как это все работает, пробуйте внести свои усовершенствования.

Выход на просторы Интернета. Если вы внесли некоторые изменения в функции пакета или хотите создать свой клон веб-приложения, можно запустить свой собственный экземпляр *Radiant* на *shinyapps.io*. Скопируйте репозитарий https://github.com/radiant-rstats/radiant и запустите *radiant/inst/app/for.shinyapps.io.R*. После этого откройте *radiant/inst/app/ui.R* и разверните приложение. Почти аналогично можно разместить приложение на любом другом сервере.

Радиант может работать теперь в облаке (например, *AWS*), для чего используется настроенный контейнер *Docker*. См. *https://github.com/radiant-rstats/docker* для подробностей.

Документация. Кроме уже упоминавшихся в этом разделе ссылок, дополнительные сведения можно получить также, изучив заметки и документацию на

- https://cran.r-project.org/web/packages/radiant/index.html
- http://radiant-rstats.github.io/radiant
- http://radiant-rstats.github.io/radiant.data
- http://radiant-rstats.github.io/radiant.design
- http://radiant-rstats.github.io/radiant.basics
- http://radiant-rstats.github.io/radiant.model
- http://radiant-rstats.github.io/radiant.multivariate
- https://github.com/radiant-rstats/radiant/issues
- https://shiny.rstudio.com/gallery/radiant.html

Наша роль состояла лишь в том, чтобы ознакомиться с ними, переосмыслить с точки зрения пользователя, мало знакомого с программированием в R, и адаптировать описание приложения в русском переводе.