**Федеральное агентство по образованию Российской Федерации**

**Государственное образовательное учреждение**

**высшего профессионального образования**

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**Кафедра информационных технологий**

Преподаватель,

к.п.н. В.А. Акиньшина

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ОТЧЕТ**

**СТАТИСТИЧЕСКИЙ АНАЛИЗ ПОДДЕРЖАННЫХ**

**АВТОМОБИЛЕЙ МАРКИ AUDI В ВЕЛИКОБРИТАНИИ**

Работу выполнил студент 4 курса Дакукин Иван Евгениевич

факультета компьютерных технологий и прикладной математики  
спец. 01.03.02 – Прикладная математика и информатика

Краснодар 2023

СОДЕРЖАНИЕ

[**1 Исходные данные 4**](#_Toc157481815)

[**2 Графический анализ 6**](#_Toc157481816)

[**2.1 2D Graphs 6**](#_Toc157481817)

[**2.1.1 2D Histograms 6**](#_Toc157481818)

[**2.1.2 2D Scatterplots 8**](#_Toc157481819)

[**2.1.3 2D Box Plots 13**](#_Toc157481820)

[**2.2 Средство «закрашивание» 15**](#_Toc157481821)

[**2.3 3D SequentialGraphs 18**](#_Toc157481822)

[**2.3.1 Raw Data Plots 18**](#_Toc157481823)

[**2.3.2 Bivariate Histograms 18**](#_Toc157481824)

[**2.3.3 3D Box Plots 19**](#_Toc157481825)

[**3 Основные статистики 21**](#_Toc157481826)

[**3.1 Описательные статистики 21**](#_Toc157481827)

[**3.2 Корреляционная матрица 22**](#_Toc157481828)

[**3.3 Критерий Стьюдента сравнения средних 23**](#_Toc157481829)

[**3.3.1 t-test, independent, by groups 23**](#_Toc157481830)

[**3.3.2 t-test, independent, by variables 24**](#_Toc157481831)

[**3.3.3 t-test, dependent samples 24**](#_Toc157481832)

[**3.3.4 t-test, single samples 24**](#_Toc157481833)

[**3.4 Группировка и однофакторная ANOVA 25**](#_Toc157481834)

[**4 Частотный анализ 29**](#_Toc157481835)

[**4.1 Таблицы частот 29**](#_Toc157481836)

[**4.2 Таблицы кросстабуляции 32**](#_Toc157481837)

[**5 Корреляционный анализ 35**](#_Toc157481838)

[**6 Дисперсионный анализ 37**](#_Toc157481839)

[**7 Линейное многомерное моделирование взаимосвязей 43**](#_Toc157481840)

[**8 Нелинейное многомерное моделирование взаимосвязей 46**](#_Toc157481841)

[**8.1 Fixed Nonlinear Regression 46**](#_Toc157481842)

[**8.2 Логит регрессия 47**](#_Toc157481843)

[**8.3 Экспоненциальная регрессия 48**](#_Toc157481844)

[**8.4 Определенная пользователем регрессия 51**](#_Toc157481845)

[**9 Дискриминантный анализ 54**](#_Toc157481846)

[**10 Классификационный анализ без обучения 60**](#_Toc157481847)

[**10.1 Кластерный анализ 60**](#_Toc157481848)

[**10.1.1 Метод к-средних 60**](#_Toc157481849)

[**10.1.2 Двухвходовая кластеризация 63**](#_Toc157481850)

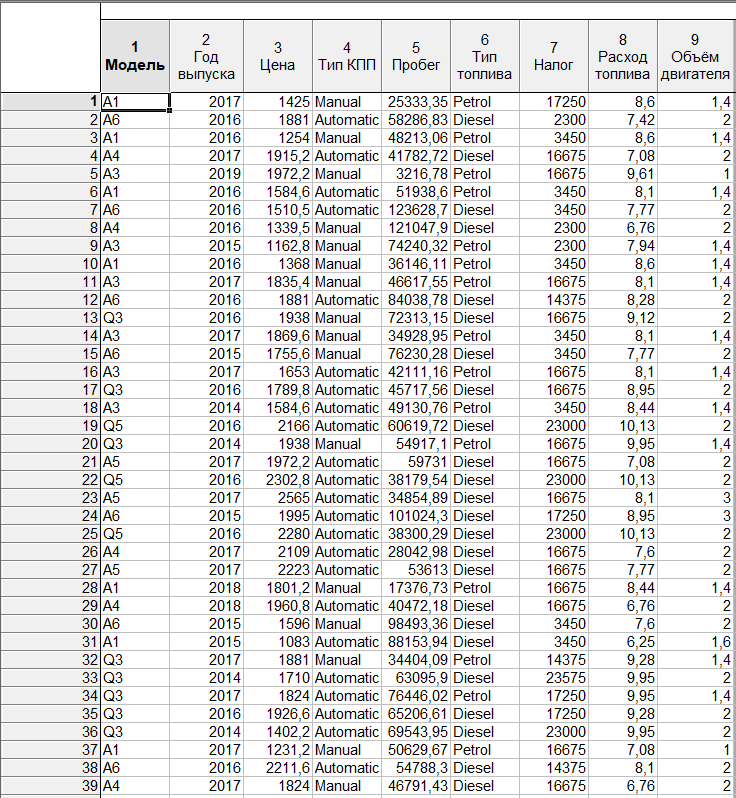
[**10.1.3 Древовидная кластеризация 65**](#_Toc157481851)

[**10.2 Деревья классификации 66**](#_Toc157481852)

[**11 Методы редукции данных 72**](#_Toc157481853)

[**11.1 Факторный анализ 72**](#_Toc157481854)

# Исходные данные



В исходной таблице представлены данные поддержанных автомобилей марки Audi в Великобритании.

Показатели:

1. Модель – модель автомобиля. Всего представлено 26 различных моделей автомобилей;
2. Год выпуска – год выпуска автомобиля;
3. Цена – цена автомобиля в тысячах рублей;
4. Тип КПП – тип коробки передач. Всего представлены 3 типа коробки передач: Manual – ручная, Automatic – автоматическая, Semi-auto – полуавтоматическая;
5. Пробег – пробег автомобиля в тысячах километров;
6. Тип топлива – тип топлива автомобиля. Всего представлено 2 типа: Diesel – дизель, Petrol – бензин;
7. Налог – налог на автомобиль в рублях;
8. Расход топлива – расход топлива автомобиля в литрах на 100 километров;
9. Объём двигателя – объём двигателя в литрах.

# Графический анализ

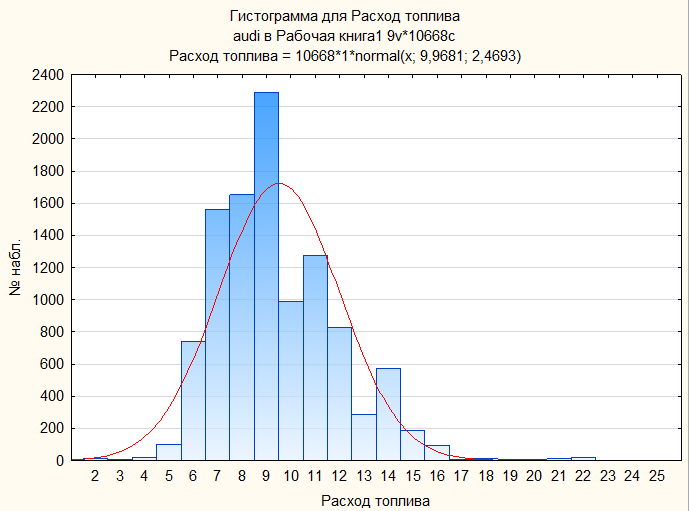
Графический анализ данных нужен для визуализации данных с помощью различных графических инструментов. Он используется для изучения распределения данных, выявления выбросов, аномалий и корреляции между переменными.

## 2D Graphs

### 2D Histograms

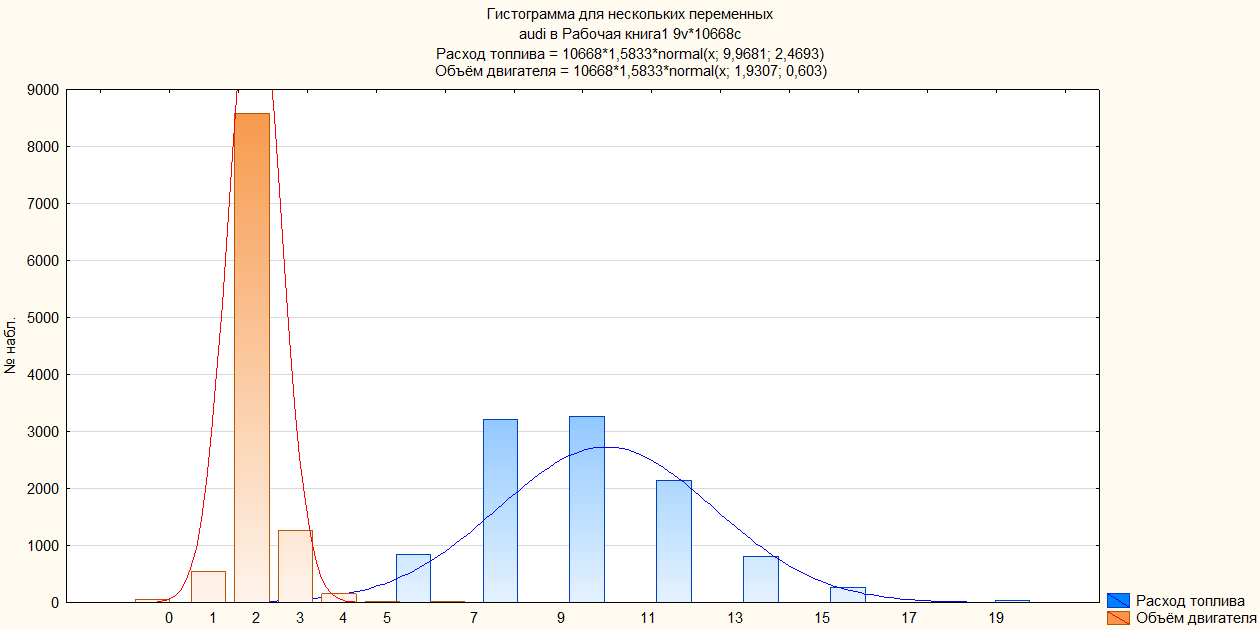
2D Histogramms являются графическими представлениями распределения частот выбранных переменных. Для каждого интервала (класса) рисуется столбец, высота которого пропорциональна частоте класса. Гистограмма наглядно показывает, какие значения или диапазоны значений исследуемой переменной являются наиболее частыми, насколько сильно они различаются, как сконцентрировано большинство наблюдений вокруг среднего, является распределение симметричным или нет, имеет ли оно моду или несколько мод. Различают несколько видов гистограмм.

2D Histogramms Regular (простые) – столбчатая диаграмма распределения частот для выбранной переменной.



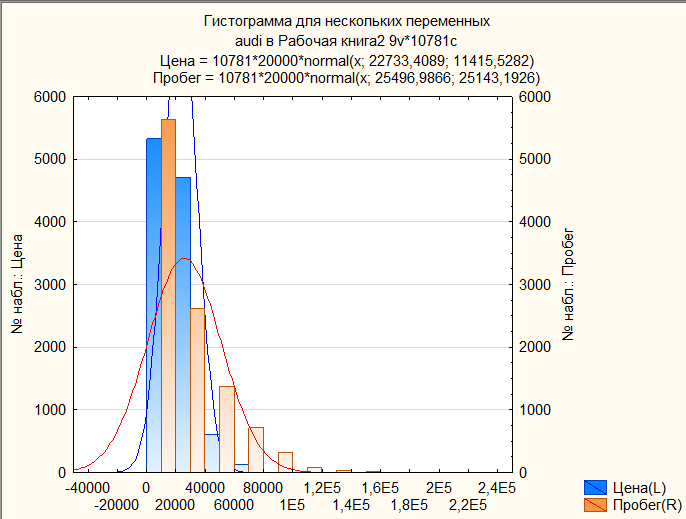
Данная 2D Histogramms Regular построена по столбцу «Расход топлива». Из гистограммы видно, что чаще всего встречаются значения от 9 до 10.

2D Histogramms Multiple (составные) – изображают распределение частот для нескольких переменных на одном графике.



В данной 2D Histogramms Multiple параметрами выступают объём двигателя автомобиля и расход топлива. Из гистограммы видно, что чаще всего можно встретить автомобили с объёмом двигателя 2 литра, а также можно увидеть, что чаще всего встречаются автомобили с расходом в диапазоне от 9 до 11 литров на 100 километров.

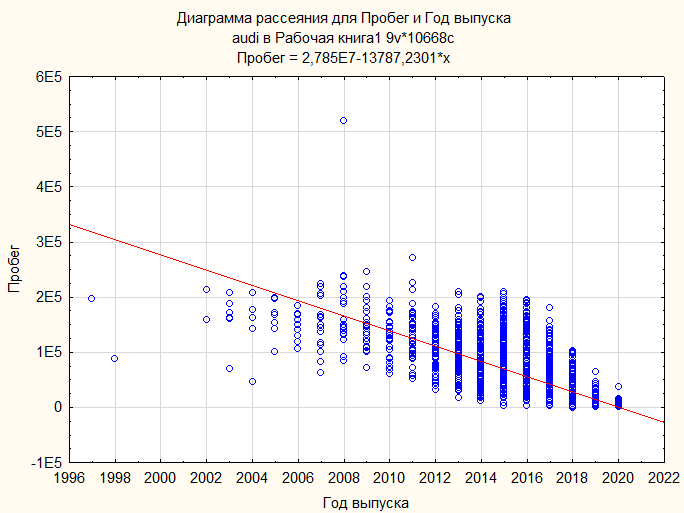
2D Histogramms Double-Y (с двойной осью Y) – комбинация двух по-разному масштабированных составных гистограмм.



### 2D Scatterplots

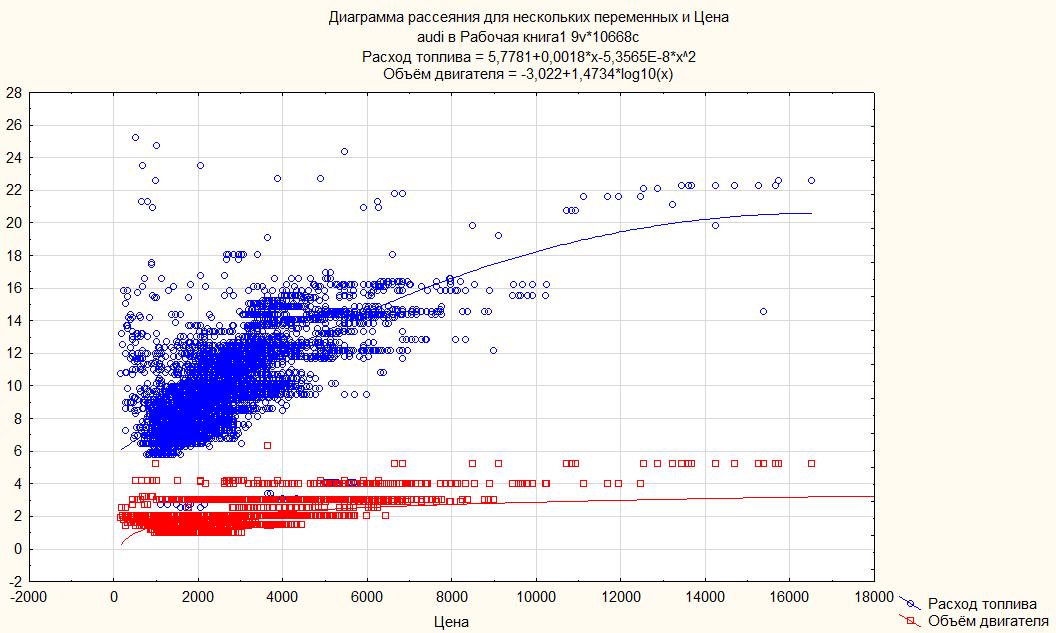
2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными.

2D Scatterplots Regular (простые) – визуализируют зависимость между двумя переменными X и Y.



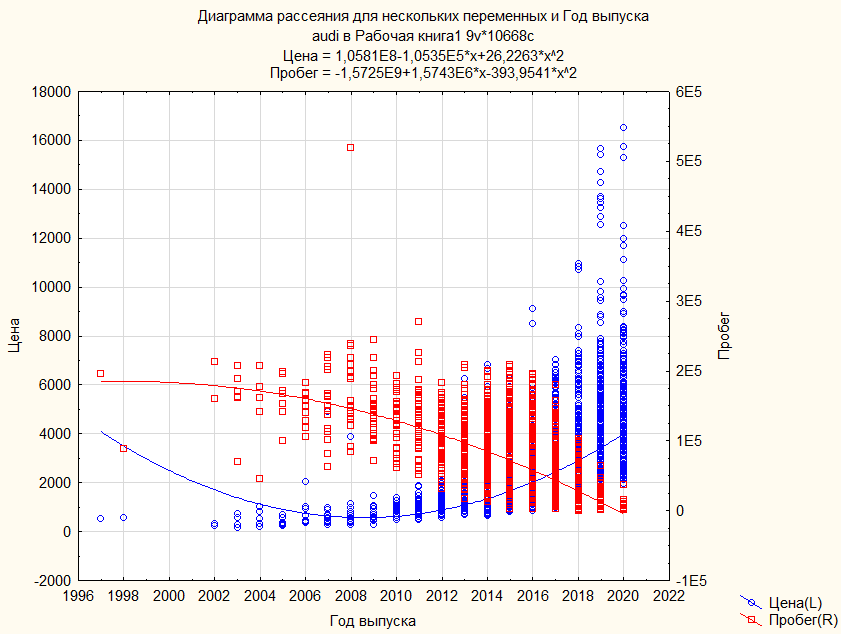
Данная 2D Scatterplots Regular диаграмма отображает зависимости между годом выпуска автомобиля и его пробегом с линейной подгонкой.

2D Scatterplots Multiple (составные) – состоит из нескольких зависимостей и изображает несколько корреляций.



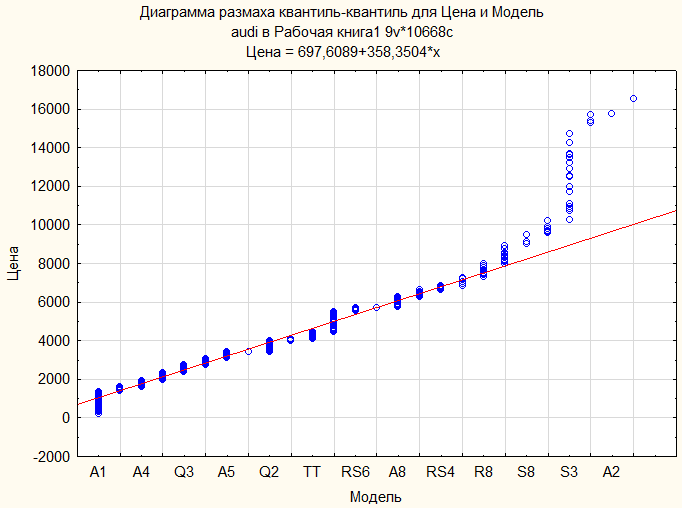
Данная 2D Scatterplots Multiple диаграмма отображает зависимости между количеством расходуемого топлива, объёмом двигателя и ценой с полиномиальными подгонками.

2D Scatterplots Double-Y (с двойной осью Y) – комбинация двух составных диаграмм рассеяния для одной переменной X и двух различных наборов переменных Y.

****

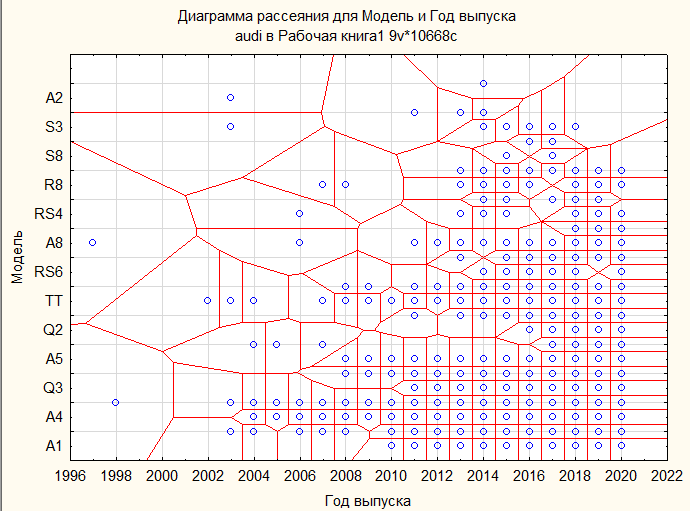
Данная 2D Scatterplots Double-Y диаграмма отображает комбинацию двух составных диаграмм: первая зависимость между годом выпуска автомобиля и его пробега, вторая зависимость – цена автомобиля и год выпуска. Отсюда видно, что чем новее машина, тем больше её стоимость и меньше её пробег, при этом стоимость возрастает быстрее.

2D Scatterplots Quantile (квантилей) – показывает зависимость между квантилями двух переменных, позволяющая оценить сходство эмпирических распределений.



Данная 2D Scatterplots Quartile отображает зависимость между квантилями переменных модель автомобиля и его цена.

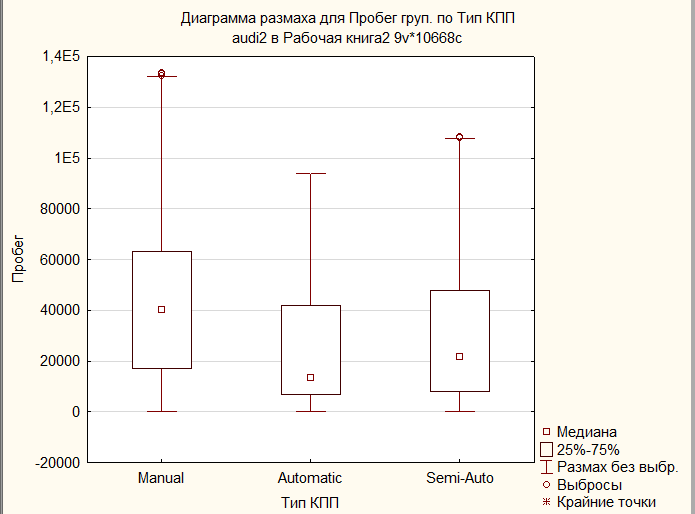
2D Scatterplots Voronoi (Вороного) – показывает разделение пространства между точками данных. Пространство между отдельными точками делится границами на такие области, каждая точка которых находится ближе к заключенной внутри точке, чем к любой другой соседней точке.



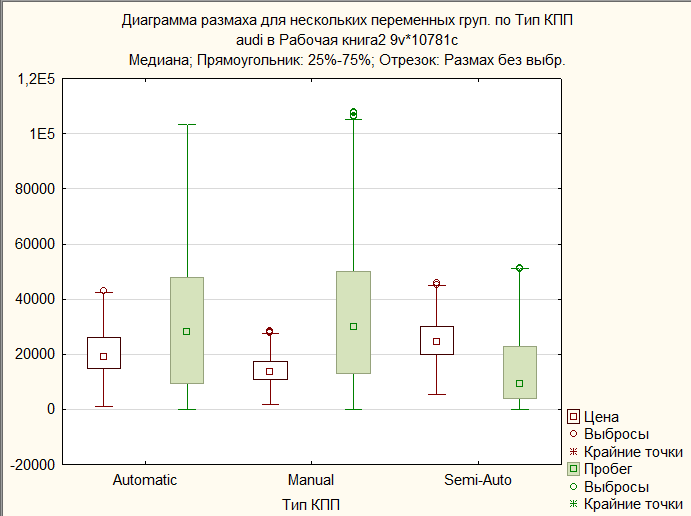
Данная 2D Scatterplots Voronoi диаграмма построена для переменных год выпуска и модель.

### 2D Box Plots

2D Box Plots (графики ящика – диаграммы размаха) – на этих диаграммах изображаются диапазоны или характеристики распределения значений выбранной переменной отдельно по группам, заданным категориальной переменной.



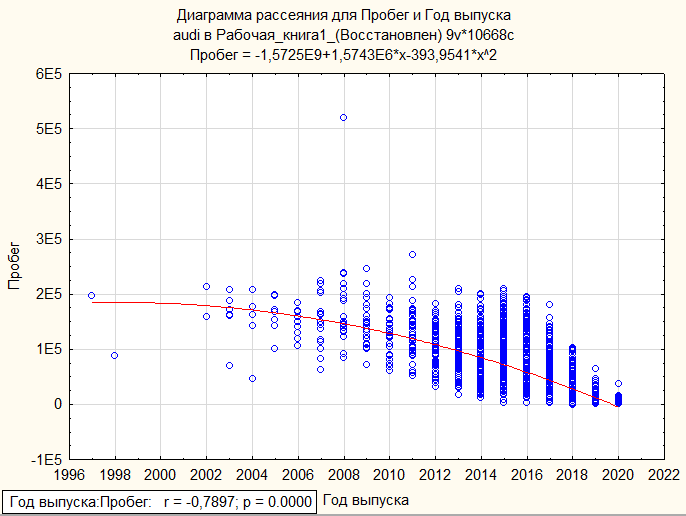
На этой 2D Box Plots Regular диаграмме представлены характеристики распределения значений пробега автомобилей по типу коробки передач. Медианами для Manual, Automatic, Semi-Auto будут 40182, 13427 и 21784 соответственно.



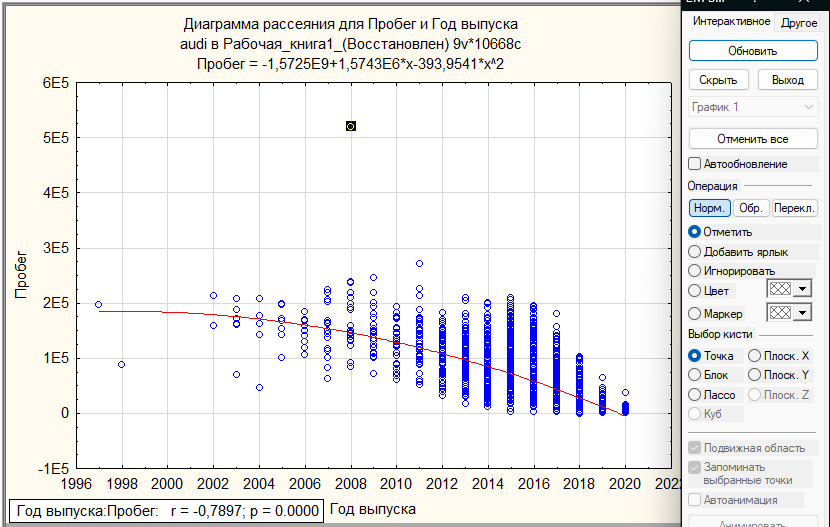
На 2D Box Plots Multiple графике изображены параметры распределения данных о пробеге и цене для групп автомобилей с различными типами коробки передач.

## Средство «закрашивание»

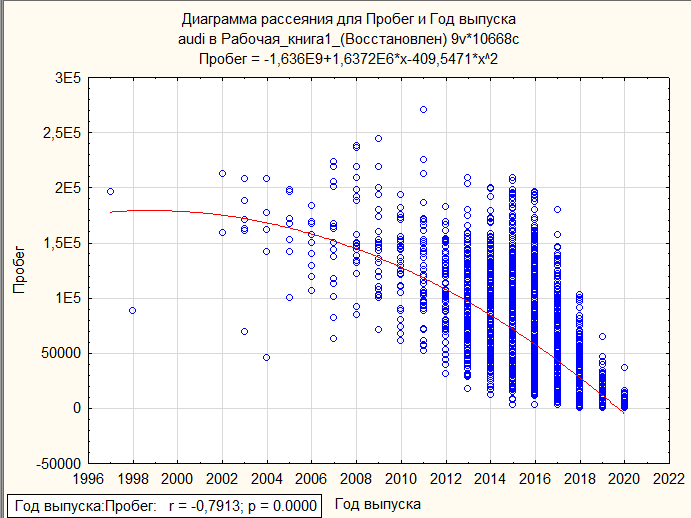
Построим диаграмму рассеивания для переменных пробег и год выпуска.



Из диаграммы видно, что есть точки, которые располагаются далеко от линии регрессии, поэтому аппроксимация может быть некачественной. Для исключения этой ситуации воспользуемся средством «закрашивание», чтобы исключить некоторые точки.



Теперь с помощью тех же средств закрашивания можно исключить эту точку:

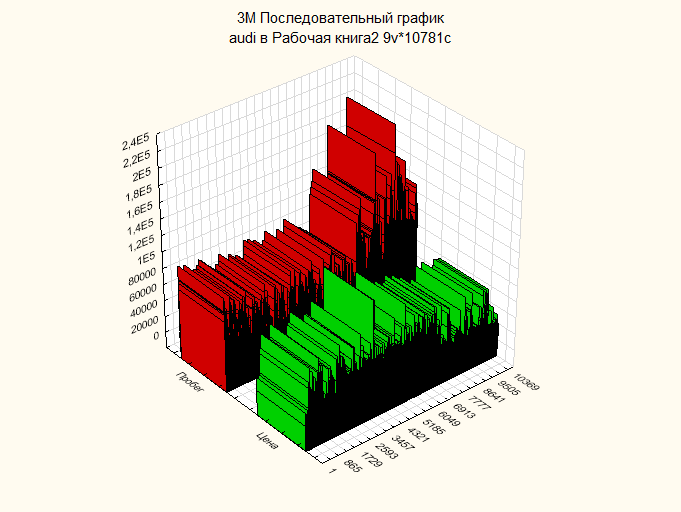


Сравнивая r до и после исключения точки, можно сказать, что качество аппроксимации изменилось.

## 3D SequentialGraphs

### Raw Data Plots

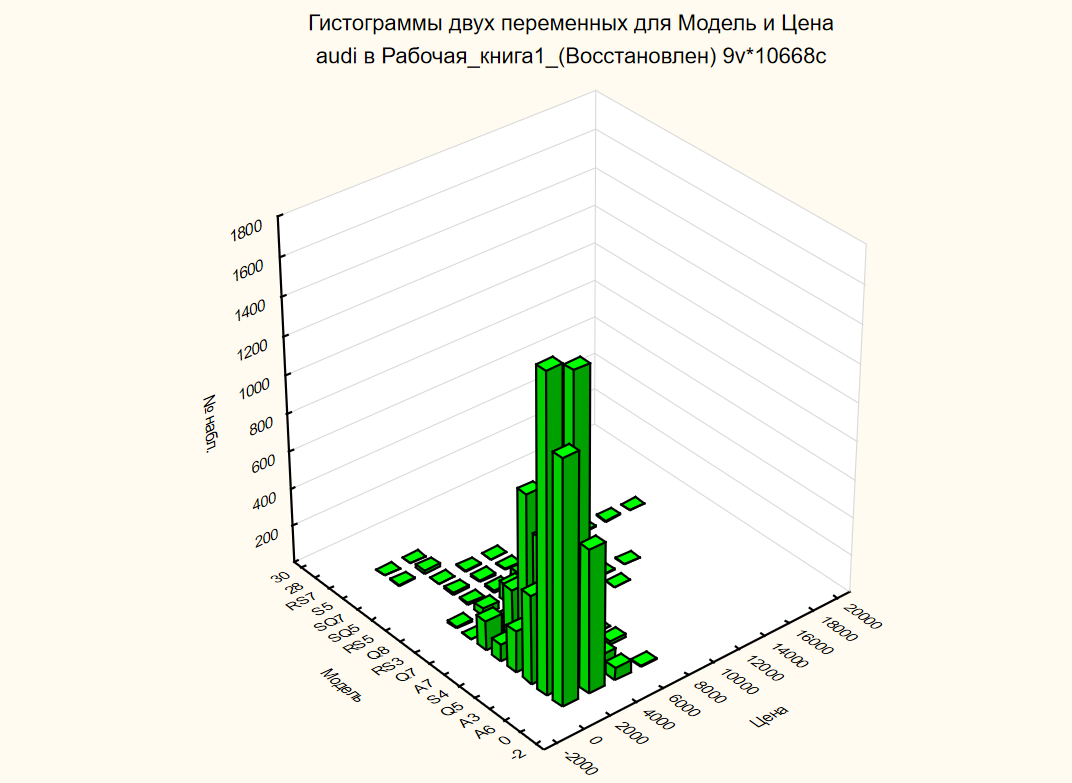
Raw Data Plots (графики исходных данных) – иллюстрируют соотношения между значениями переменных



На данном графике показано соотношения значений между ценой и пробегом автомобилей.

### Bivariate Histograms

Эти гистограммы можно рассматривать как сочетание двух простых гистограмм, соединенных так, чтобы можно было исследовать частоты совместного появления значений двух переменных.

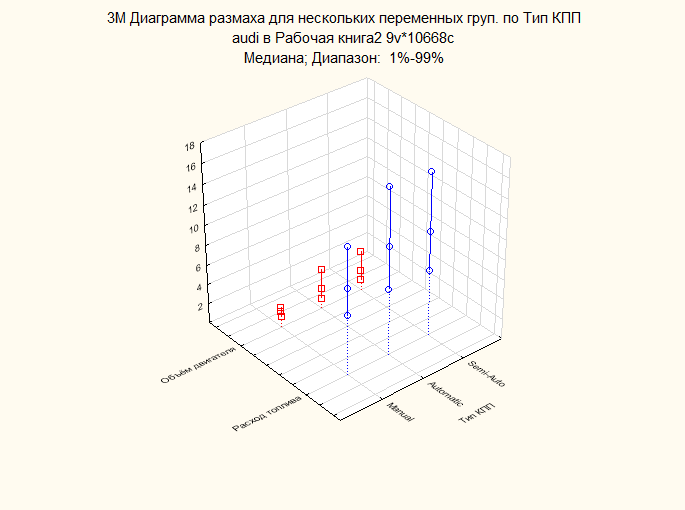


На данной гистограмме показано совместное распределение частот переменных цены и модели автомобиля.

### 3D Box Plots

3D Box Plots – это диаграммы размаха.

Построим диаграмму размаха для переменных расхода топлива и объема двигателя по группам потребляемого автомобилем топлива:

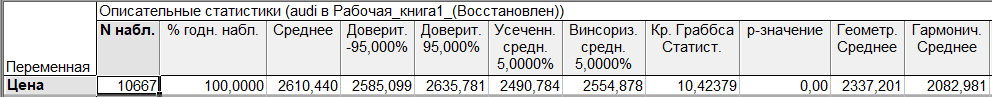


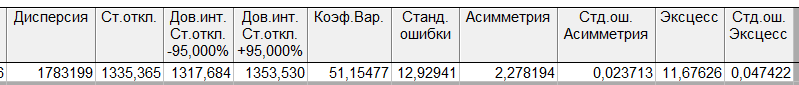
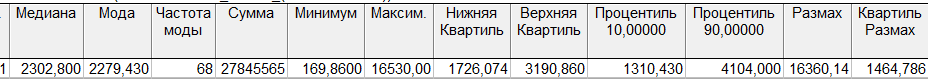
# Основные статистики

## Описательные статистики

Анализ данных желательно начинать с проверки данных, пытаясь выявить выбросы – нетипичные данные, которые могут быть ошибками сбора данных, их форматирования, передачи и т.д. Для этой цели можно использовать модуль Описательные статистики. Также они нужны для общей характеристики данных, по которым проводится анализ.

Таблица описательных статистик для цены автомобилей:





В таблице представлены следующие статистики: число наблюдений 10667, среднее 2610,440, сумма 27845565, медиана 2302,800, мода 2279,430, частота моды 68, геометрическое среднее 2337,201, гармоническое среднее 2082,981, стандартное отклонение 1335,365, дисперсия 1783199, стандартная ошибка среднего 12,92941, доверительные пределы для среднего 2585,099 и 2635,781, ассиметрия 2,78194, стандартная ошибка ассиметрии 0,023713, эксцесс 11,67626, стандартная ошибка эксцесса 0,047422, минимум 169,8600, максимум 16530, нижний квартиль 1726,074, верхний квартиль 3190,86, 10-я процентиль (квантиль 0,1) 1310,430, 90-я процентиль (квантиль 0,9) 4104, размах 16360,14, квартильный размах 1464,786.

## Корреляционная матрица

Между переменными (случайными величинами) может существовать функциональная связь, проявляющаяся в том, что одна из них определяется как функция от другой. В качестве меры зависимости между переменными используется коэффициент корреляции (r), который изменяется в пределах от –1 до +1. Если коэффициент корреляции отрицательный, это означает, что с увеличением значений одной переменной значения другой убывают. Если коэффициент корреляции не равен 0 (переменные называются некоррелированными), то это значит, что между переменными существует зависимость. Чем ближе значение модуля r к 1, тем зависимость сильнее.



Из таблицы видно, что год выпуска автомобиля и его пробег демонстрируют сильную отрицательную корреляцию со значением -0,79. Это означает, что для автомобилем с более поздним годом выпуска присущ меньший пробег. Также сильная положительная корреляция обнаруживается между расходом топлива автомобиля и его ценой. Это означает, что автомобили с более высокой ценой, как правило, обладают большим расходом топлива.

Умеренная корреляция присутствует, например, между переменными цены и года выпуска, цены и пробега, объема двигателя и цены, расхода топлива и налога.

Слабая корреляция присутствует, например, между переменными года выпуска и объема двигателя, объема двигателя и пробега, года выпуска и налога.

Красным отмечены значения, где уровень значимости <0.05.

Для более наглядного отображение корреляции между переменными, мы можем отобразить карту цветов корреляции:



## Критерий Стьюдента сравнения средних

t-критерий Стьюдента – общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента.

### t-test, independent, by groups

t-test, independent, by groups (t-критерий для независимых выборок с группирующей переменной) используется, если надо сравнить средние случайных величин двух независимых групп, полученных из одной выборки при помощи группирующей переменной. Другими словами, сравниваются средние двух подвыборок, записанные в одном столбце.

Проверим равенство средних для всех переменных по группам потребляемого автомобилями топлива на основе t-критерия (t-test, independent, by groups).



По данным таблицы можно сделать вывод, что отличия средних значимы, т.к. р<0,05, следовательно гипотезу о равенстве дисперсии для всех наших переменных отвергаем.

### t-test, independent, by variables

t-test, independent, by variables (t-критерий для независимых выборок) применяется, если надо сравнить средние случайных величин, полученных по двум разным (независимым) выборкам, но записанных в разных столбцах и имеющих одно и ту же единицу измерения.

В наборе данных отсутствуют независимые переменные имеющие одинаковые единицы измерения, поэтому проверить равенство средних на основе t-критерия (t-test, independent, by variables) будет невозможно.

### t-test, dependent samples

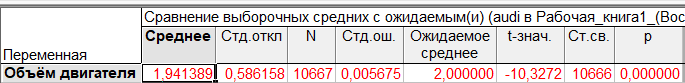
t-test, dependent samples (t-критерий для зависимых выборок) применяется, если надо сравнить средние случайных величин двух групп, одна из которых представляет повторные измерения другой.

В наборе данных отсутствуют зависимые переменные, поэтому проверить равенство средних на основе t-критерия (t-test, dependent samples) будет невозможно.

### t-test, single samples

t-test, single samples (простые выборки). Применяется, если надо сравнить среднее выборки со средней генеральной совокупности.

Проверим гипотезу, что объем двигателя у автомобилей не отличается статистически значимо от указанного среднего (2л).

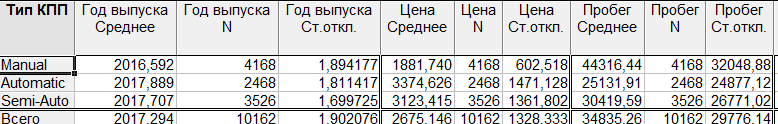


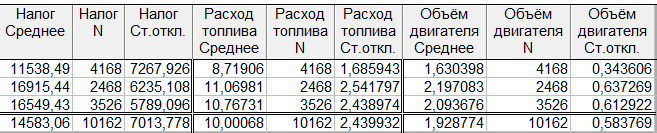
По данным таблицы можно сделать вывод, что среднее отличается существенно от указанного среднего, т.к. значение p<0.05

## Группировка и однофакторная ANOVA

Модуль дисперсионного анализа ANOVA используется для сравнения средних в более чем двух группах.

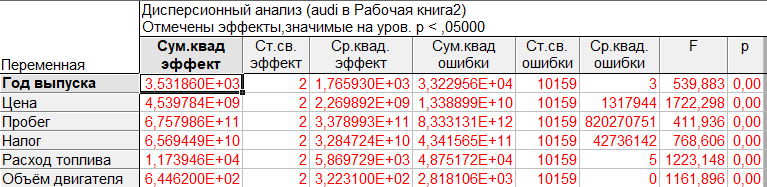
Будем использовать переменную Тип коробки передач в качестве группирующей. Построим таблицу разбиения на группы:

****



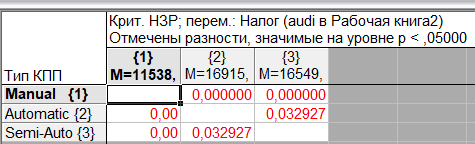
Данная таблица содержит ряд описательных статистик для указанных групп.

Теперь, для проверки значимости различий в средних указанных групп, используем процедуру анализа дисперсий, получим таблицу:



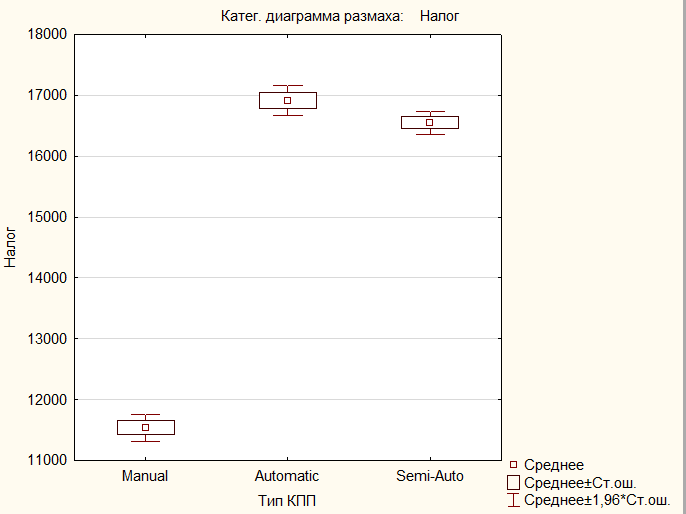
Из таблицы видно, что гипотеза для данных параметров неверна, т.к. p < 0.05.

Проведем LSD-test для переменной налог:

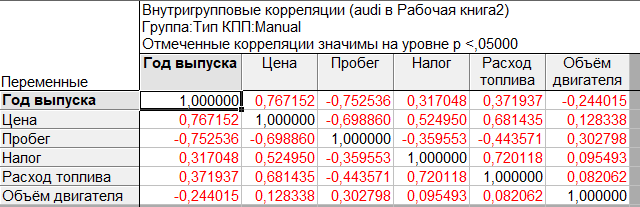


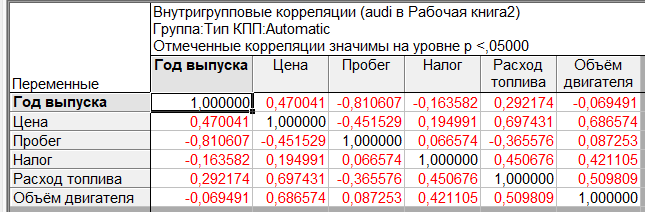
Из таблицы видно, что гипотеза для данных параметров неверна ни для какой из подгрупп, т.к. p < 0.05.

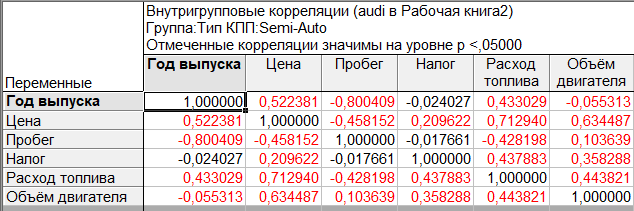
Для наглядности полученных результатов можно использовать инструмент Categorized box & whisker plot:



Внутригрупповые корреляции:







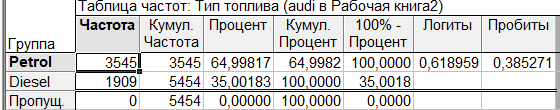
Из таблиц видно, что в разных группах зависимости между анализируемыми величинами проявляются по-разному. Так же видно, что значимые корреляции наблюдаются между годом выпуска и пробегом автомобиля, между расходом топлива и ценой.

# Частотный анализ

Частотный анализ — это метод статистического анализа, который используется для определения частоты появления определенных значений или категорий в наборе данных. Этот метод позволяет оценить распределение данных и выявить наиболее часто встречающиеся значения или категории. Частотный анализ может использоваться для исследования структуры данных, выявления зависимостей между переменными и построения моделей прогнозирования.

## Таблицы частот

Таблицы частот, или одновходовые таблицы, представляют собой простейший метод анализа категориальных (номинальных) переменных. Часто их используют как одну из процедур разведочного анализа, чтобы просмотреть, каким образом различные группы данных распределены в выборке.

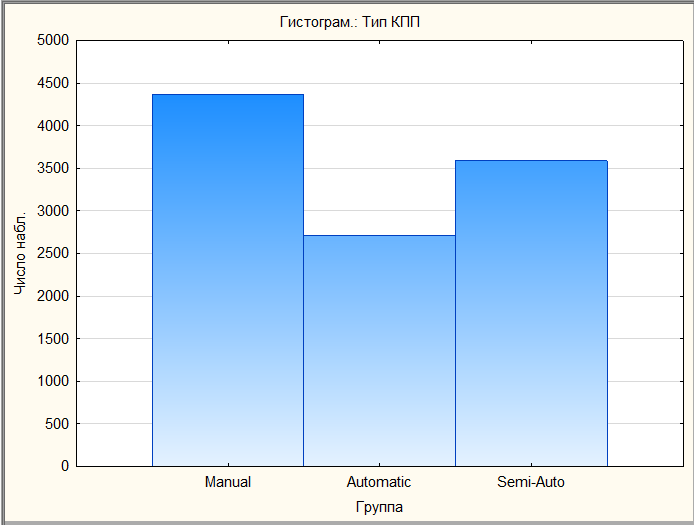
****

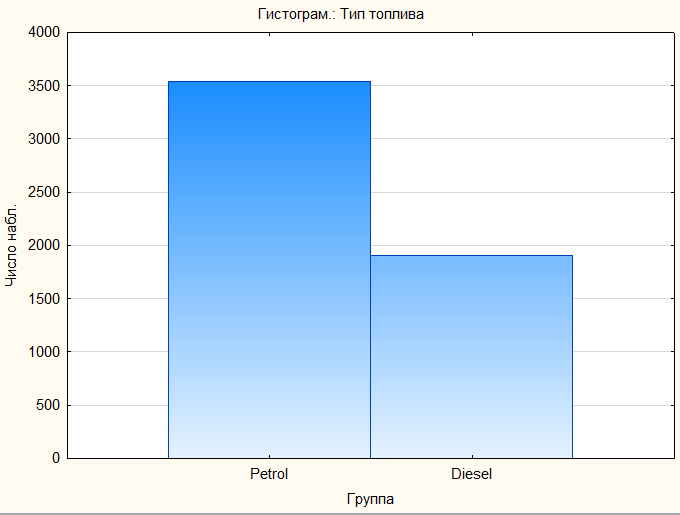
Данная таблица частот построена для категориальной переменной Тип топлива. В таблице для категорий Дизель и Бензин приведены следующие анализы: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты, 100%-кумулятивные проценты, логит преобразование частот, пробит преобразование частот.



Данная таблица частот построена для категориальной переменной тип коробки передач. В таблице для категорий Механическая, Автоматическая, Полуавтоматическая указаны те же показатели, что и в предыдущей таблице.

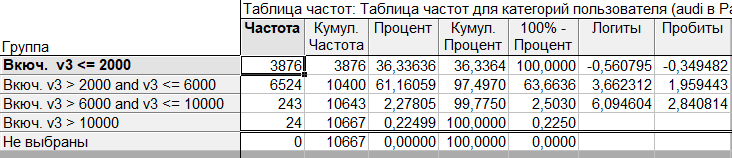
Гистограммы соответствующих категориальных переменных:





Из гистограмм видно, что больше всего автомобилей на бензине и с механической коробкой передач.

Построим таблицу частот по пользовательским категориям.

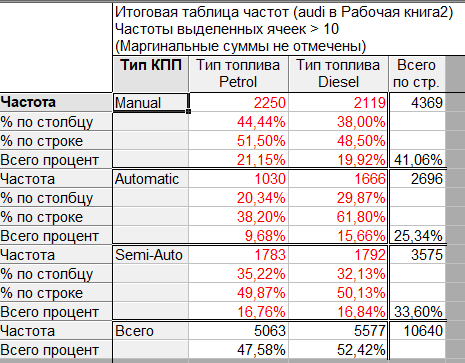


Каждая из этих категорий определяются интервалами: v3<=2000, v3 > 2000 and v3 <= 6000, v3 > 6000 and v3 <= 10000, v3 > 10000 где v3 – это переменная цены.

## Таблицы кросстабуляции

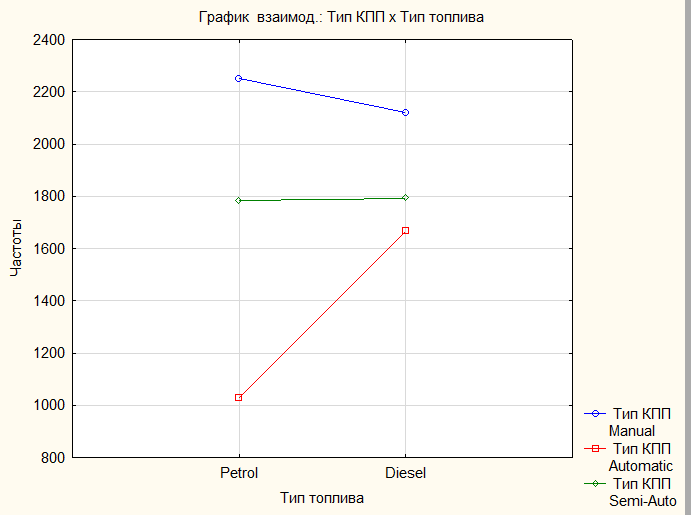
Кросстабуляция (сопряжение) – процесс объединения двух (или нескольких) таблиц частот так, что каждая ячейка (клетка) в построенной таблице представляется единственной комбинацией значений или уровней табулированных переменных. Таким образом, кросстабуляция позволяет совместить частоты появления наблюдений на разных уровнях рассматриваемых факторов. Исследуя эти частоты, можно определить связи между табулированными переменными.

Построим таблицу кросстабуляции на основании таблицы частот переменной Тип коробки передачи и таблицы частот переменной Тип топлива.



В данной таблице представлены частоты по типам топлива и типам коробки передач.

Графики взаимодействия частот:



Подробная двухвходовая таблица:

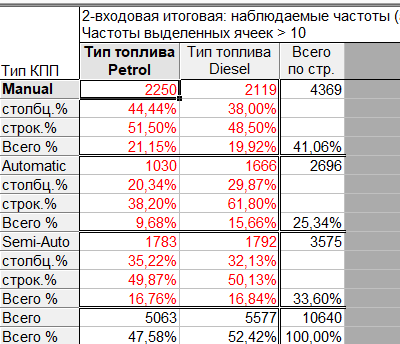


Таблица статистик для двухвходовой таблицы:



Критерий Пирсона – простой критерий проверки значимости связи между двумя категоризованными переменными. Нулевая гипотеза – “между переменными нет зависимости”. Таким образом, между исследуемыми переменными есть зависимость (нулевая гипотеза не принимается, так как p<0.05).

Из таблицы Statistics: следует, что между типом коробки передач и типом топлива существует статистически значимая взаимосвязь – уровни значимости p критериев Пирсона Хи-квадрат меньше 0,05.

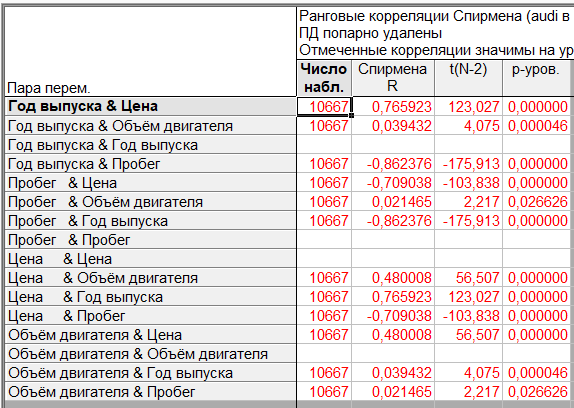
# Корреляционный анализ

Корреляционный анализ — это метод статистического анализа, который используется для изучения связи между двумя или более переменными. Он позволяет определить, насколько сильно и в каком направлении связь между переменными. Корреляционный анализ может быть полезен для выявления зависимостей между различными факторами и предсказания будущих значений переменных на основе известных данных.

Нулевая гипотеза – «между переменными корреляция статистически не значима».

Коэффициент Спирмена используется, если закон распределения переменных неизвестен или не является нормальным.

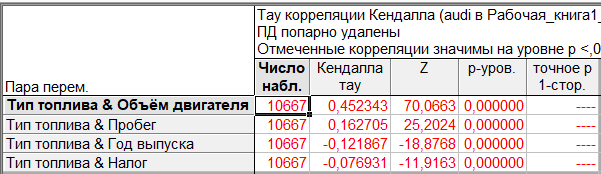
Вычислим данный коэффициент, чтобы проверить корреляцию между переменными Цена, Год выпуска, Объём двигателя и Пробег.



Так как p<0.05, то нулевая гипотеза отвергается, то есть корреляция между переменными статистически значима.

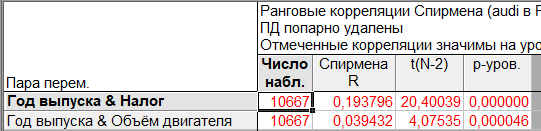
Коэффициент Кендалла используется, если хотя бы одна переменная категориальная.

Вычислим данный коэффициент, чтобы проверить корреляцию между переменными Налог, Объём двигателя, Пробег, Год выпуска и переменной Тип топлива.



Так как p<0.05, то нулевая гипотеза принимается, то есть корреляция между переменными статистически значима.

Коэффициент Гамма используют, если переменные содержат много повторяющихся значений.

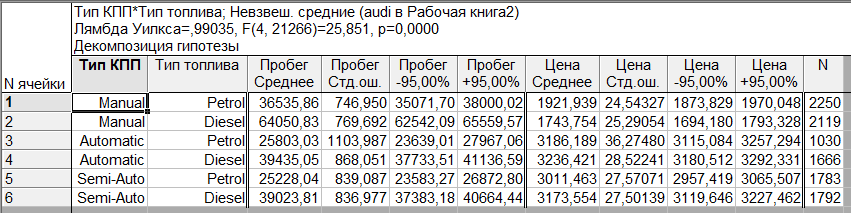


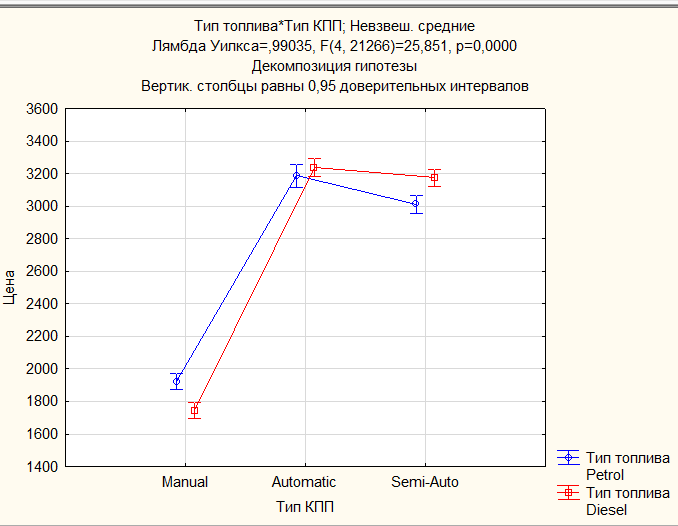
Так как p<0.05, то нулевая гипотеза принимается, то есть корреляция между переменными статистически значима.

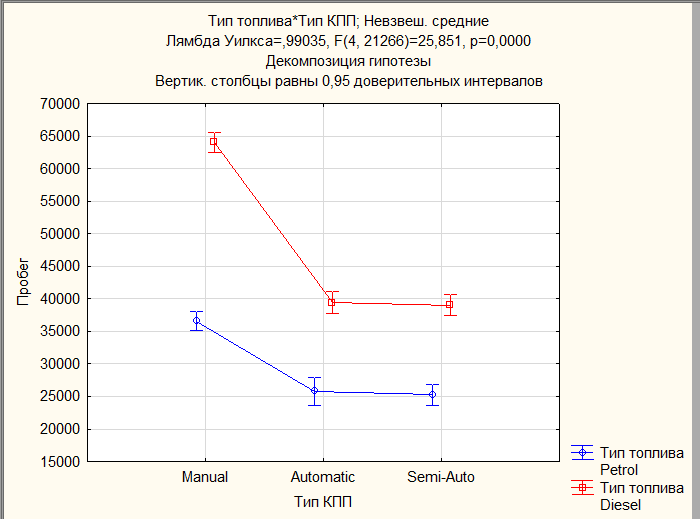
# Дисперсионный анализ

Дисперсионный анализ является наиболее общим методом сравнения средних. В дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от одного или нескольких качественных признаков (факторов).

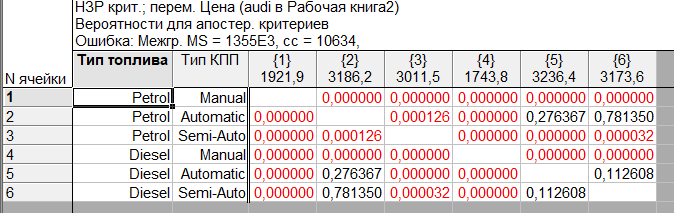
В качестве зависимых переменных выбраны пробег и цена, в качестве категориальных: тип топлива(дизель/бензин) и тип КПП (механика/автомат/полуавтомат).



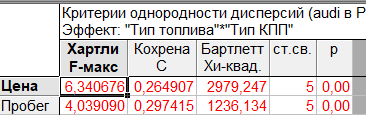




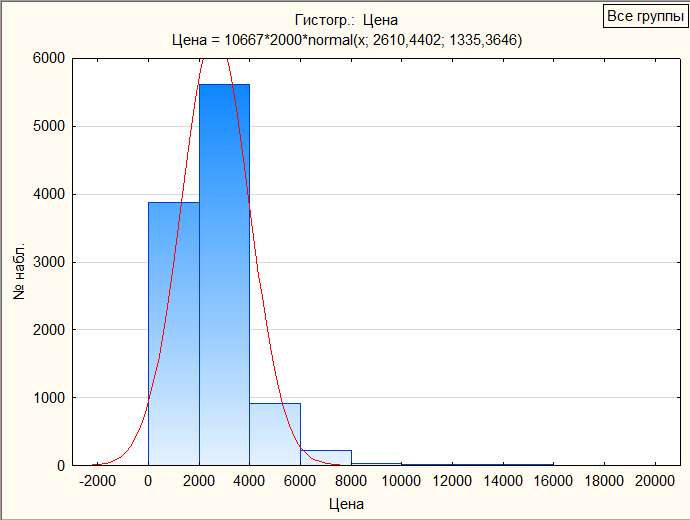
Из графиков видно, что среди автомобилей с механической коробкой передач (Manual) среднее значение переменной Цена больше у автомобилей с типом топлива Petrol, а средний Пробег больше у автомобилей с типом топлива Diesel; среди автомобилей с автоматической коробкой передачи (Automatic) среднее значение переменных Цена и Пробег больше у автомобилей с типом топлива Diesel; среди автомобилей с полуавтоматической коробкой передачи (Semi-Auto) среднее значение переменных Цена и Пробег больше у автомобилей с типом топлива Diesel;

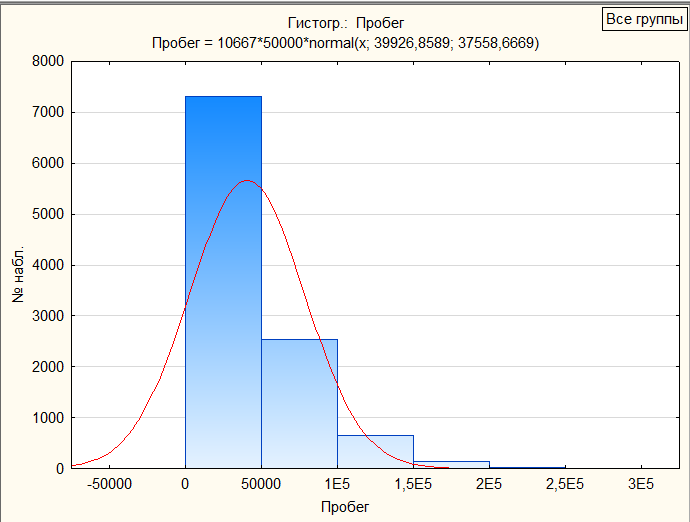


Из таблицы видно, что гипотеза о равенстве средних верна не во всех группах, то есть средние не равны в этих группах (значения должны быть больше 0.05, чтобы гипотеза была верна).

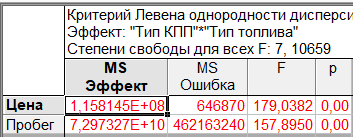


Из таблицы видно, что по критериям Кохрана, Хартли и Бартлетта гипотеза однородности дисперсии отклоняется для всех переменных из таблицы (отклоняется, если p<0.05).

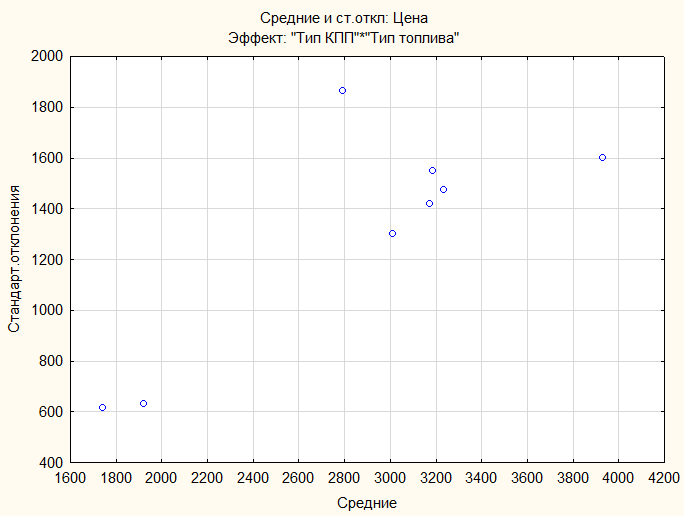




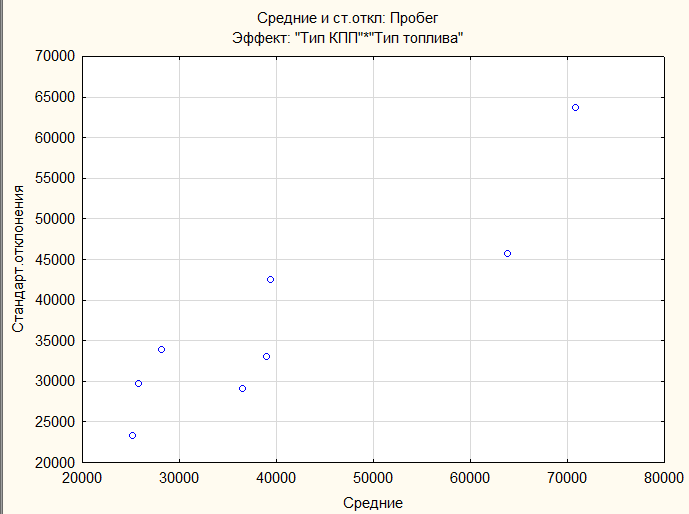
Из гистограмм видно, что общее распределение не соответствует нормальному распределению.



Из таблицы видно, что по критерию Левена гипотеза однородности дисперсии отклоняется только для переменных цена и пробег.



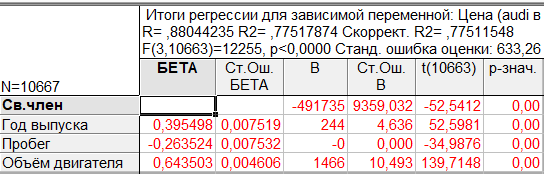
На данной диаграмме наблюдается скопление точек в диапазоне средних от 3000 до 3300 и в диапазоне стандартного отклонения от 1300 до 1600.



На данной диаграмме наблюдается скопление точек в диапазоне средних от 20000 до 40000 и в диапазоне стандартного отклонения от 20000 до 40000.

# Линейное многомерное моделирование взаимосвязей

Построим зависимость между переменными пробег, цена. Линейное многомерное моделирование представляет собой статистический метод анализа данных, используемый для изучения взаимосвязей между несколькими переменными. Он также известен как множественная линейная регрессия в случае, когда исследуются связи между зависимой переменной и двумя или более независимыми переменными.



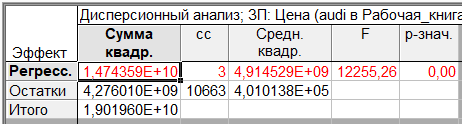
Из таблицы видно, что коэффициенты регрессии при переменных пробег и объём двигателя статистически значимы (p<0.1). Зависимость между откликом и предикторами достаточная (R2=0.775).

Построим таблицу частной корреляции по тем же параметрам.

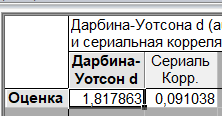


По данной таблице можно ранжировать влияния предикторов на отклик. Больше всего влияет Объём двигателя. Предиктор Объём двигателя имеет наибольшую частную и получастную корреляции (0,8 и 0,64 соответственно). Tolerance=1-R-square. p-value говорит о значимости частных коэффициентов корреляции переменных.

Проведем дисперсионный анализ для проверки гипотезы о значимости уравнения регрессии при помощи критерия Фишера.

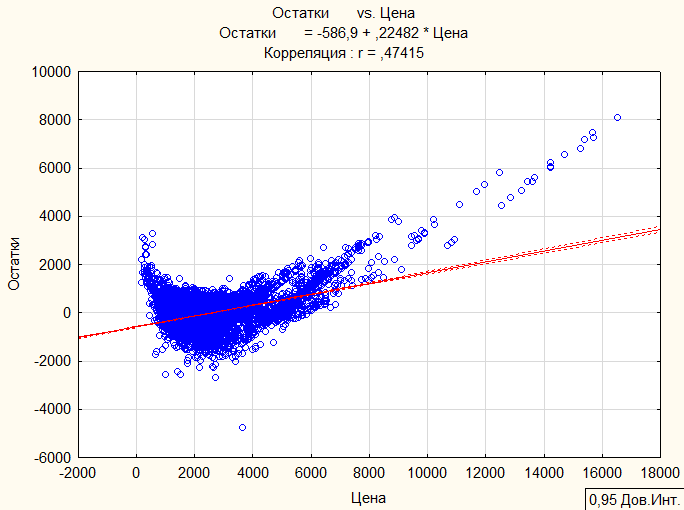


Так как р<0.05, то уравнение регрессии статистически значимо.

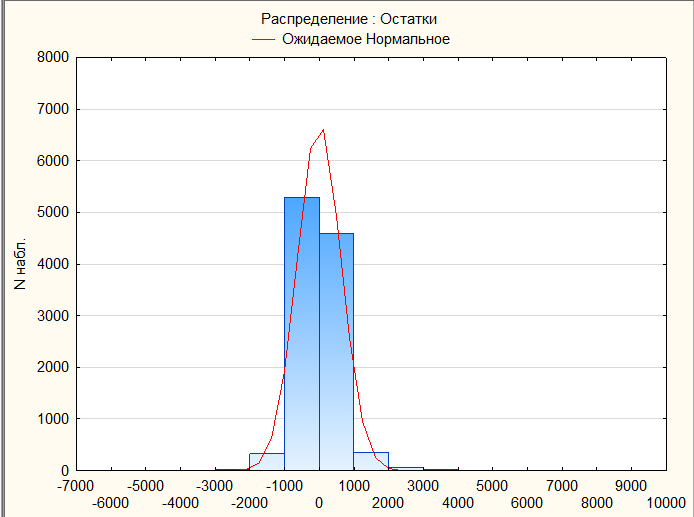


Из таблицы видно, что статистика Дарбина-Уотсона имеет небольшо значение ~1,818 при низкой сериальной корреляции ~0,09. Это свидетельствует о некоторой зависимости наблюдений, следовательно, можно говорить о недостаточной устойчивости некоторых значений коэффициентов регрессии, а значит о невысокой адекватности модели изучаемому процессу.

Графическое сравнение предсказанных и наблюдаемых значений отклика приведено ниже:



Построим гистограмму остатков.



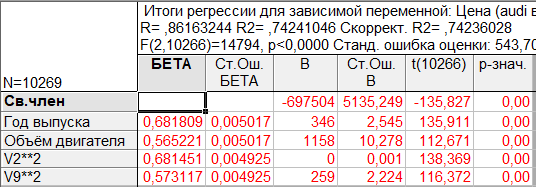
По данной гистограмме можно судить о соответствии закона распределения остатков нормальному закону.

# Нелинейное многомерное моделирование взаимосвязей

## Fixed Nonlinear Regression

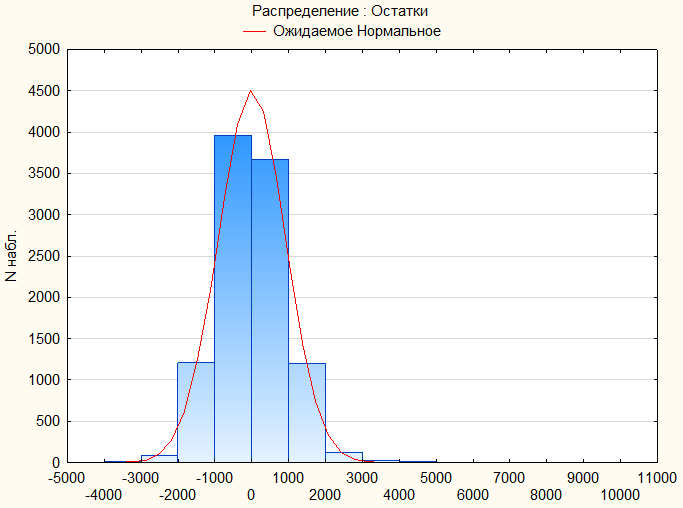
Этот модуль реализует множественный линейный регрессионный анализ с линеаризованной моделью.

Построим квадратичную зависимость цены, года и объёма двигателя.



Из таблицы видно, что Цена зависит от всех переменных. Как видно, скорректированный квадратный коэффициент детерминации высокий (R2=0.742).

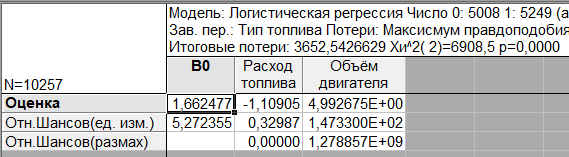
Для проверки адекватности модели построим гистограмму остатков.



Видно, что распределение остатков соответствует нормальному.

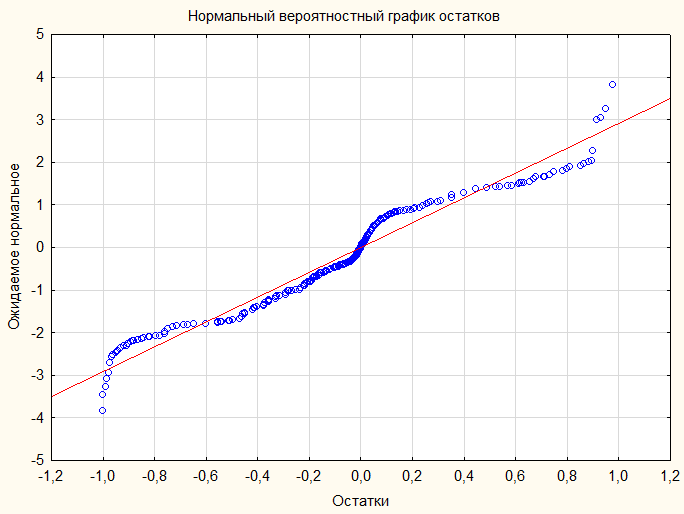
## Логит регрессия

Построим Логит модель зависимости бинарной переменной Тип топлива от Расхода топлива и Объёма двигателя.

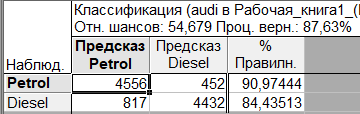


Регрессия значима, так как p<0.05 (отклоняется гипотеза об отсутствии связи).

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.



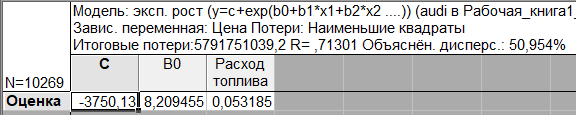
Построим таблицу Наблюдаемые и предсказанные.



Видно, что модель предсказывает автомобили на бензине (Petrol) с вероятностью 90,91%, а автомобили на дизеле (Diesel) с вероятностью 84,4%.

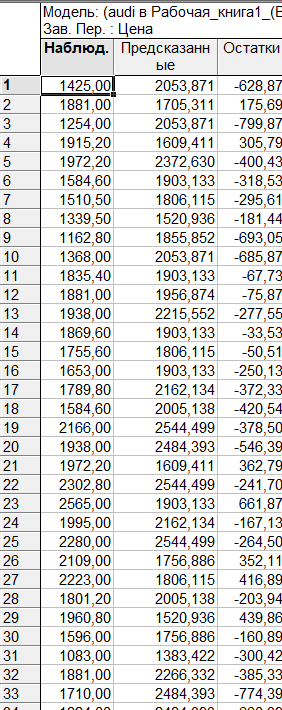
## Экспоненциальная регрессия

Построим экспоненциальную модель зависимости цены автомобиля от расхода топлива:

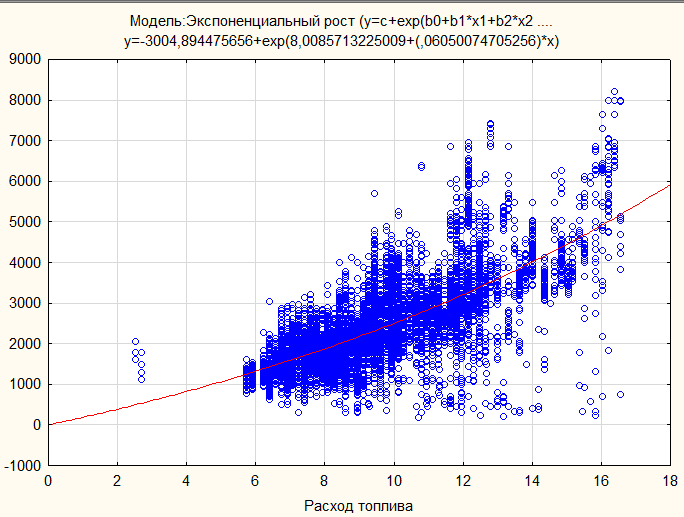
****

Как видно, коэффициент детерминации высокий (R2=0,71) и доля объясненной дисперсии средняя (50,95%)

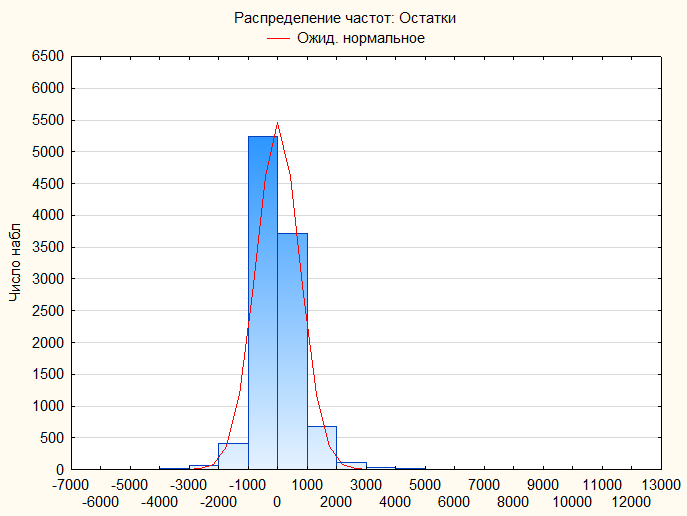
Часть таблицы предсказанных и наблюдаемых значений:



Построим подогнанную функцию.



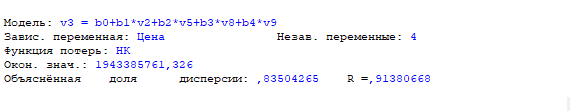
Гистограмма остатков:



Из гистограммы видно, что распределение остатков примерно соответствует нормальному, что свидетельствует об адекватности модели.

## Определенная пользователем регрессия

Построим следующую модель: Стоимость = b0 + b1\*(Год регистрации) + b2\*(Пробег) + b3\*(Расход топлива) + b4\*(Объём двигателя)

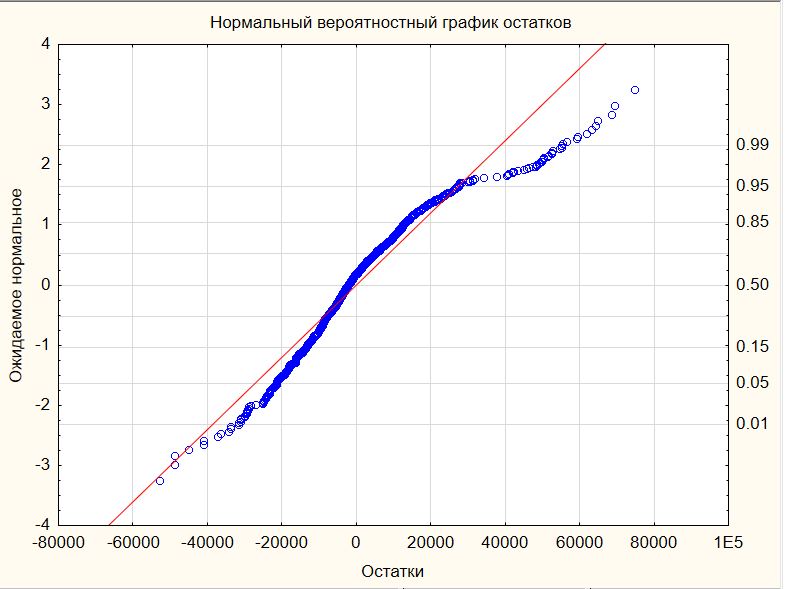
****

Из таблицы видно, что все параметры значимы.

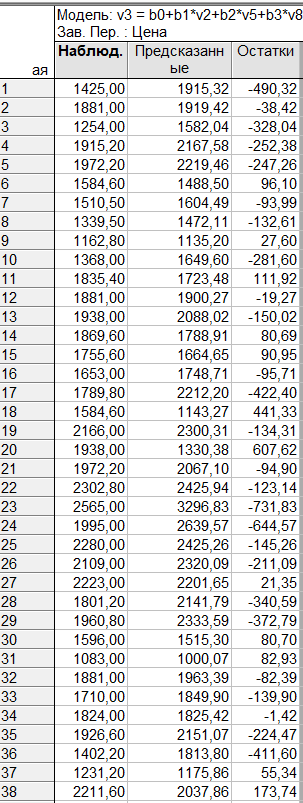
Таким образом регрессия имеет вид:

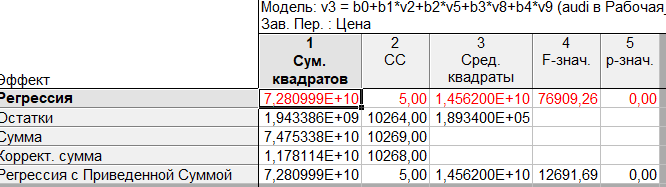
'цена' = -414377 + 250 \* 'год' + 0 \* 'пробег'+ 145 \* 'расход топлива' + 942 \* 'объём двигателя'. Коэффициент детерминации высокий (R2=0,91).

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.



Часть таблицы предсказанных и наблюдаемых значений:



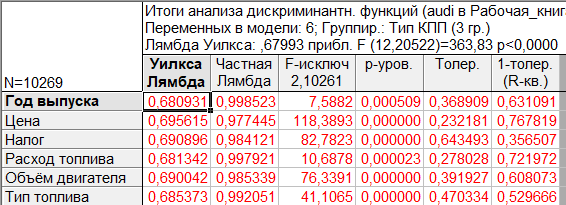
****

Из таблицы выше видно, что в целом регрессия значима.

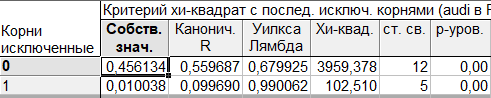
# Дискриминантный анализ

Дискриминантный анализ — это метод многомерного статистического анализа, который используется для определения, какие переменные наилучшим образом разделяют две или более группы объектов. Он позволяет выявить различия между группами объектов и определить, какие переменные наиболее важны для разделения групп. Основной идеей дискриминантного анализа является поиск линейных комбинаций переменных, которые максимально разделяют группы объектов. Эти комбинации называются дискриминантными функциями.

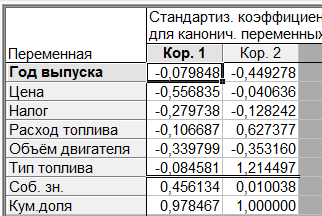
Задача состоит в том, чтобы по результатам рассмотрения всех показателей автомобилей определить тип коробки передач.



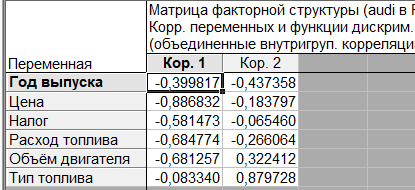
Значение лямбды Уилкса, лежащее около 0, говорит о хорошей дискриминации, около 1 - плохой. Чем ближе значение показателя предиктора к 1, тем более значим предиктор в модели.



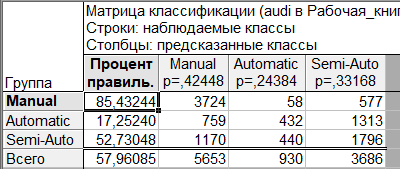
Из таблицы видно, что канонические корни статистически значимы(p<0,05).



В таблице выше приведены стандартизированные переменные дискриминантных функций.

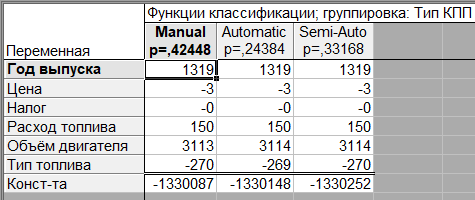


В этой таблице приведены объединенные внутригрупповые корреляции переменных (структурные коэффициенты). У переменной цена наибольшая корреляция с дискриминантной функцией 1, у типа топлива – с дискриминантной функцией 2.

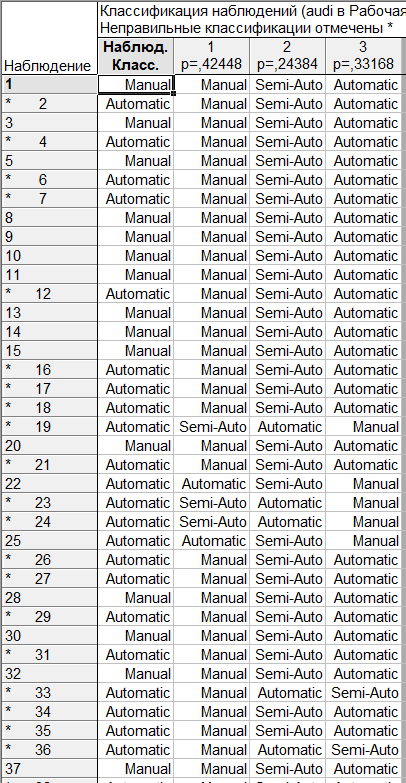


В матрице выше приведена информация о количестве и проценте корректно классифицированных наблюдений в каждой группе.

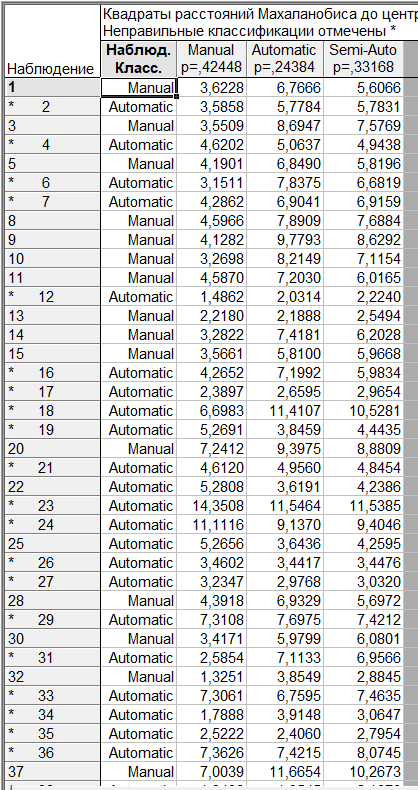
В следующей таблице приведены коэффициенты и свободные члены при переменных линейных функций:



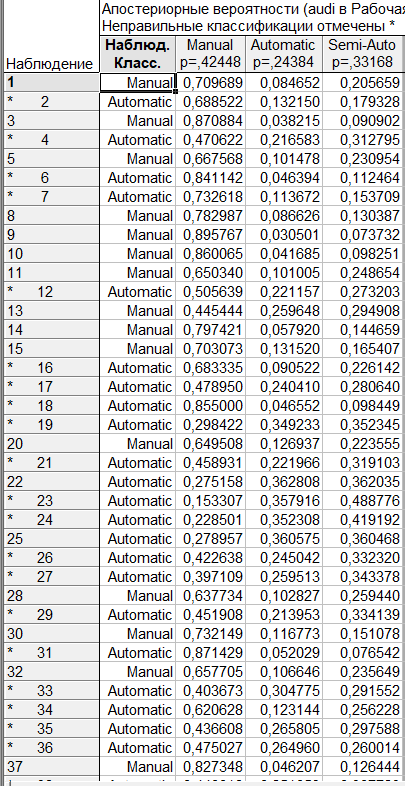
Следующая таблица содержит классификации для каждого наблюдения (часть таблицы). Наблюдения, которые не удалось классифицировать помечены знаком «\*»:



В следующей таблице представлены квадраты расстояний Махаланобиса каждого наблюдения от центроида группы, расстояния учитывают корреляции в модели (часть таблицы):



В таблице ниже каждому наблюдению поставлена в соответствие вероятность принадлежности к группе (часть таблицы):



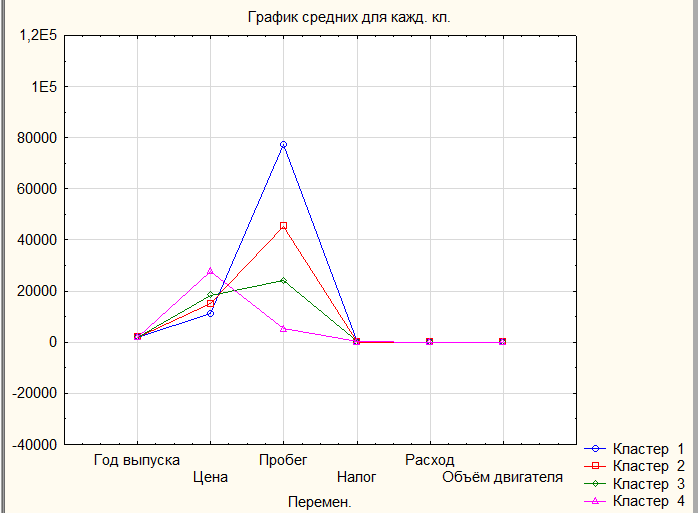
# Классификационный анализ без обучения

## Кластерный анализ

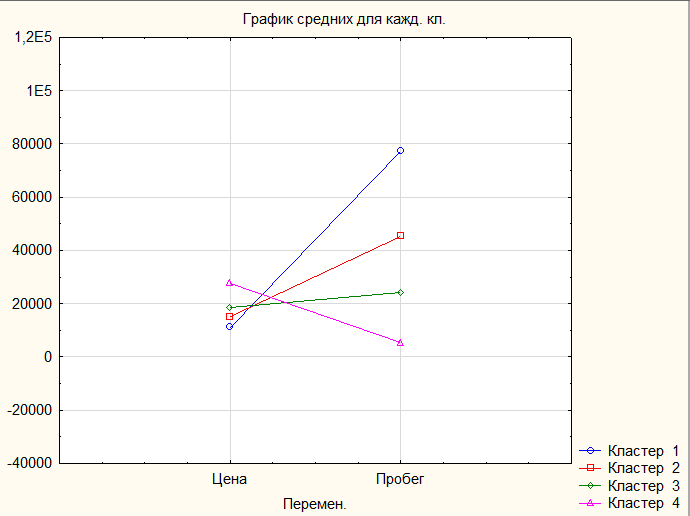
### Метод к-средних

Кластерный анализ — это метод многомерного статистического анализа, который используется для выявления групп объектов, которые более похожи друг на друга, чем на объекты из других групп. Он позволяет выделить группы объектов на основе их сходства и различия. Основной идеей кластерного анализа является поиск оптимального количества групп объектов, которые максимально отличаются друг от друга. Для этого используется различные метрики сходства и алгоритмы кластеризации.

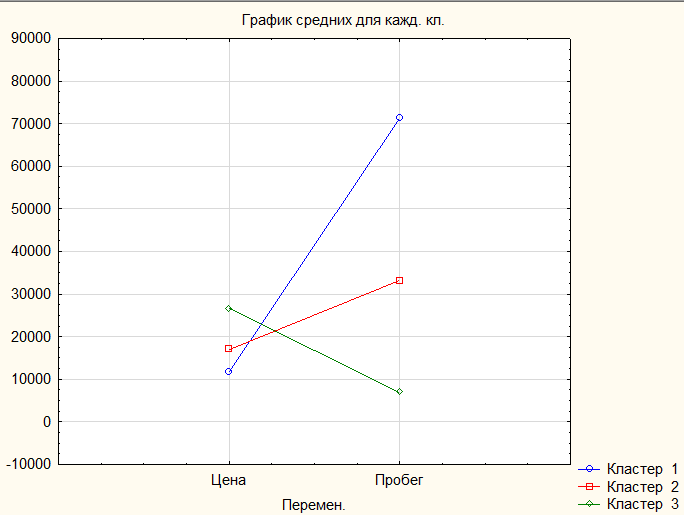
Попробуем разбить автомобили на кластеры, но до этого надо стандартизировать переменные. Варьируя количеством кластеров и исключая переменные, получили следующие результаты:



Из графика видно, что в кластерах средние значения всех параметров кроме пробег и цена незначительно отличаются друг от друга. Попробуем их исключить:



По графику средних значений для 4 кластеров по всем переменным видно, что разбиение на 4 кластеров не качественное. Разобьем объекты на 3 кластера:



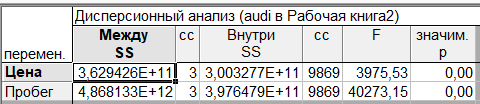
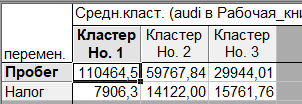
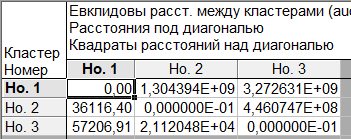


Таблица и график свидетельствуют об успешной классификации автомобилей. В таблицы приведены значения межгрупповых и внутригрупповых дисперсий признаков. Чем меньшее значение внутригрупповой и больше межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру. Так как p<0.05 для всех признаков, то вклады всех признаков в разделение объектов на группы существенные.



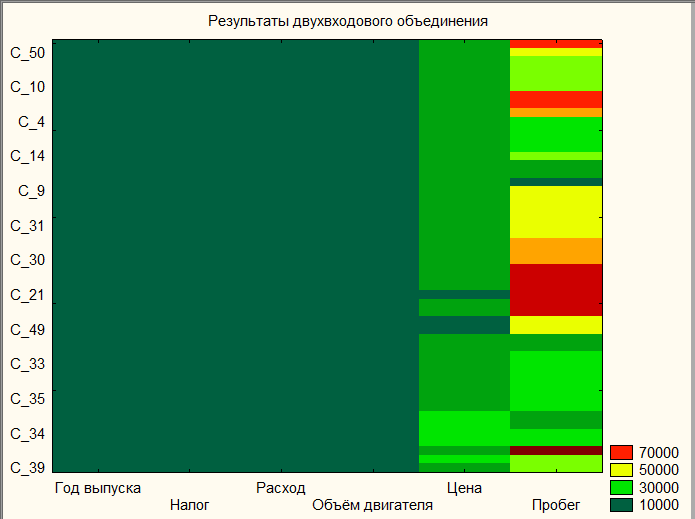


В первой из данных таблиц приведены средние для каждого кластера, во второй – евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

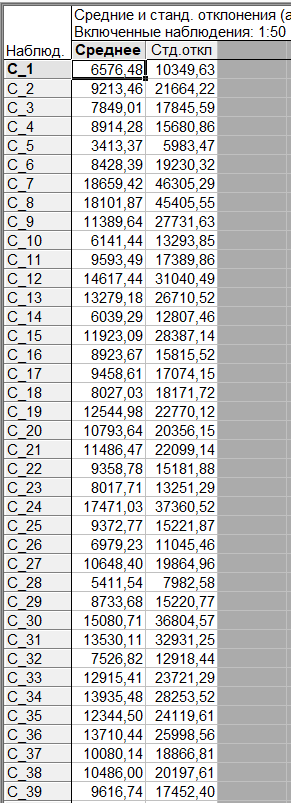
### Двухвходовая кластеризация

Проведем двухвходовую кластеризацию (одновременно и по переменным, и по строкам) для всех количественных признаков первых 50 автомобилям.

Ниже приведен цветной график результата кластеризации:

****

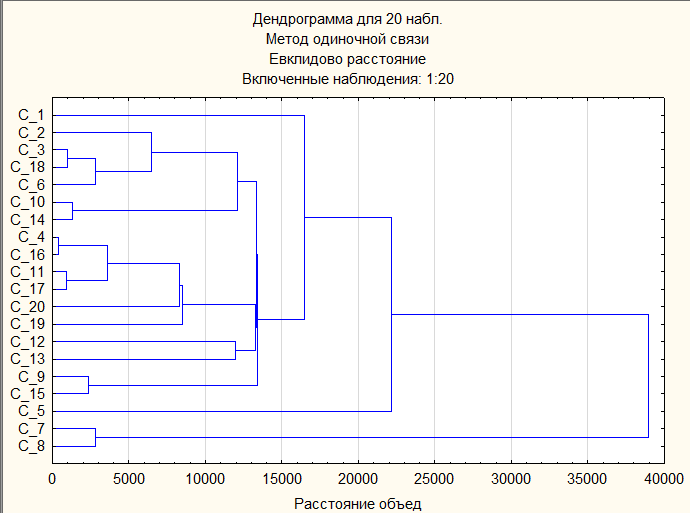
Из графика видно, что кластер уровня 10000 включает в себя большую часть из 50 наблюдений по переменным Год выпуска, Налог, Расход топлива, Объём двигателя.

****

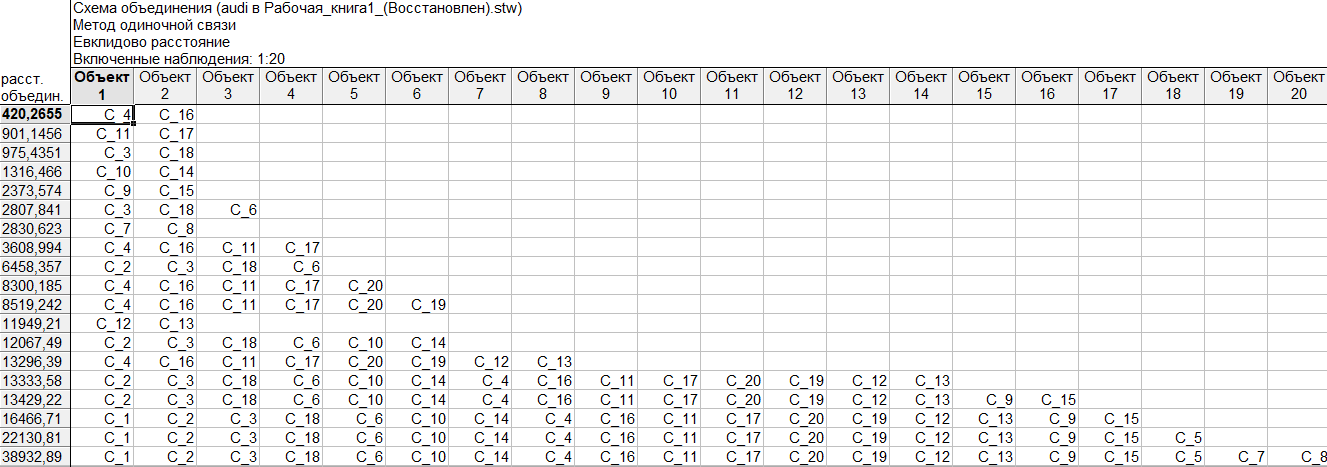
В данной таблице приведены значения средних и стандартных отклонений для строк.

### Древовидная кластеризация

Горизонтальная древовидная диаграмма имеет вид:



Правило объединения в кластеры показано в следующей таблице:



Построим график, на котором изображен порядок объединения строк.

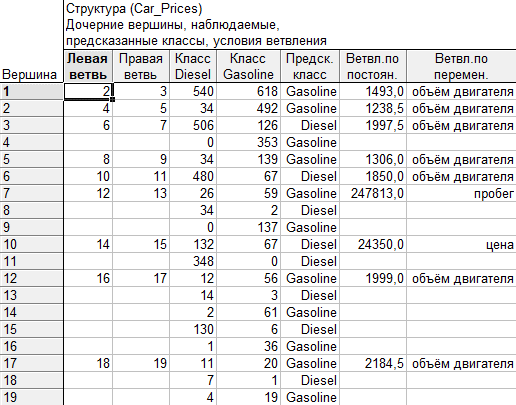
****

Заметим, что начиная с 18 шага расстояние объединения начало значительно увеличиваться.

## Деревья классификации

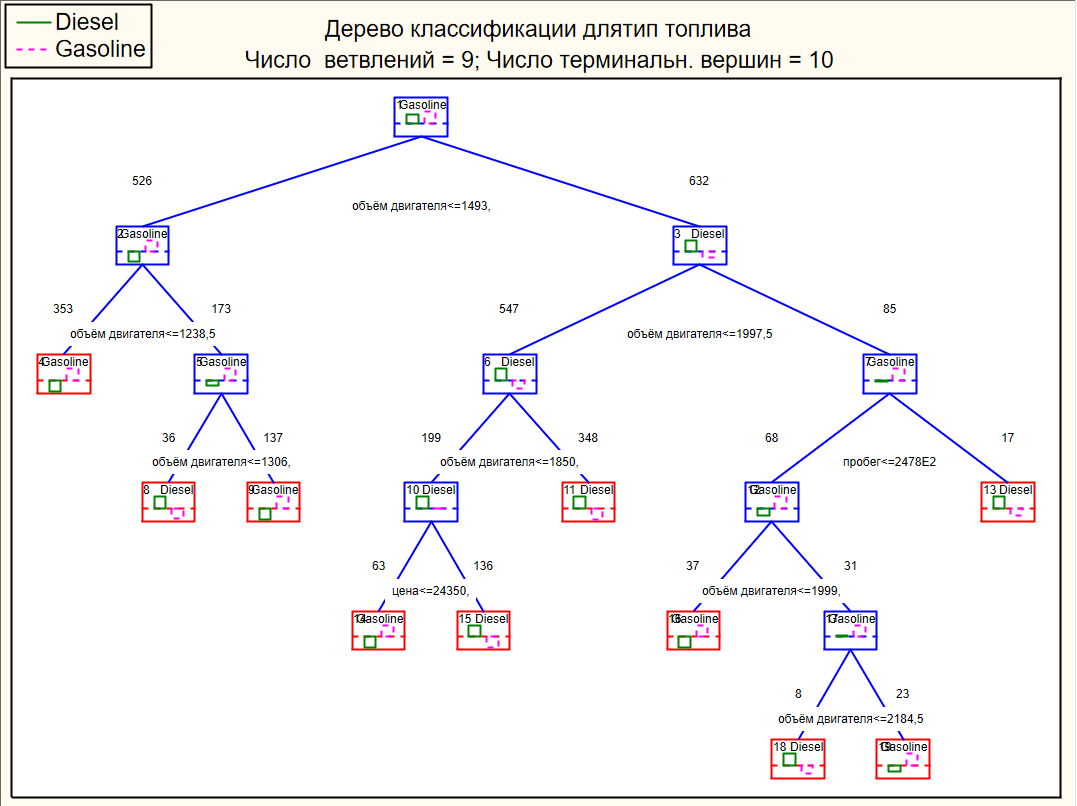
Деревья классификации – это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких независимых переменных.

Будим классифицировать по типу топлива.

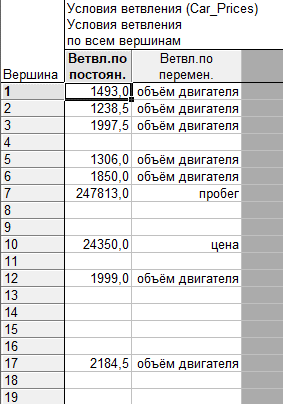


Из таблицы видно, что левая ветвь содержит 9 узлов с номерами: 2, 4, 6, 8, 10, 12, 14, 16 и 18, правая ветвь содержит тоже 9 узлов под номерами: 3, 5, 7, 9, 11, 13, 15, 17, 19. Из строки 1 видно, что в первой вершине классифицируется автомобиль с объёмом двигателя <=1493 бензинового типа. Условие разделения автомобилей по вершинам 2 и 3 следующее: если автомобиль у автомобиля объем двигателя <=1238,5 то это бензиновый иначе если объём двигателя <= 1997,5 то это дизельный.

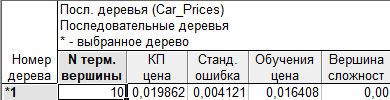
Граф дерева классификации изображен на следующем рисунке:



Условия ветвления:

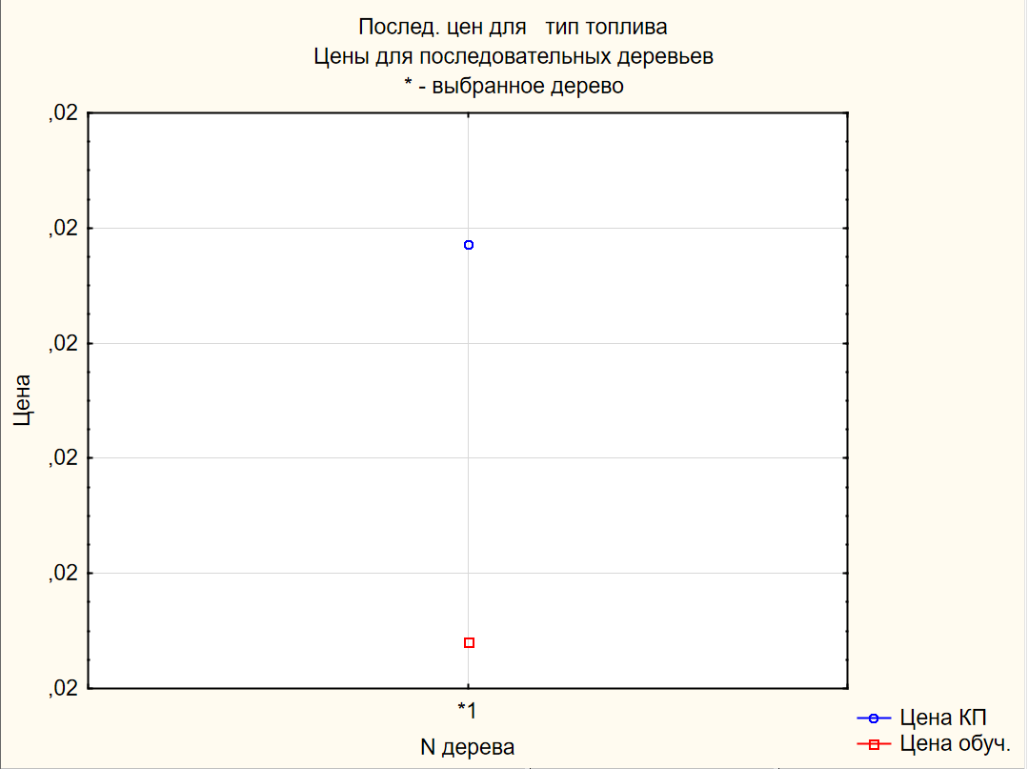


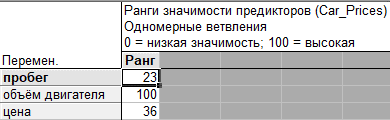
Последовательность деревьев:



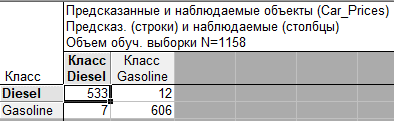
В таблице указаны терминальные вершины, цена кросс-проверки, ее стандартная ошибка, цена обучения и сложность каждого из усеченных деревьев. Звездочкой помечено дерево, которое было признано деревом «подходящего размера».

График последовательности цен:



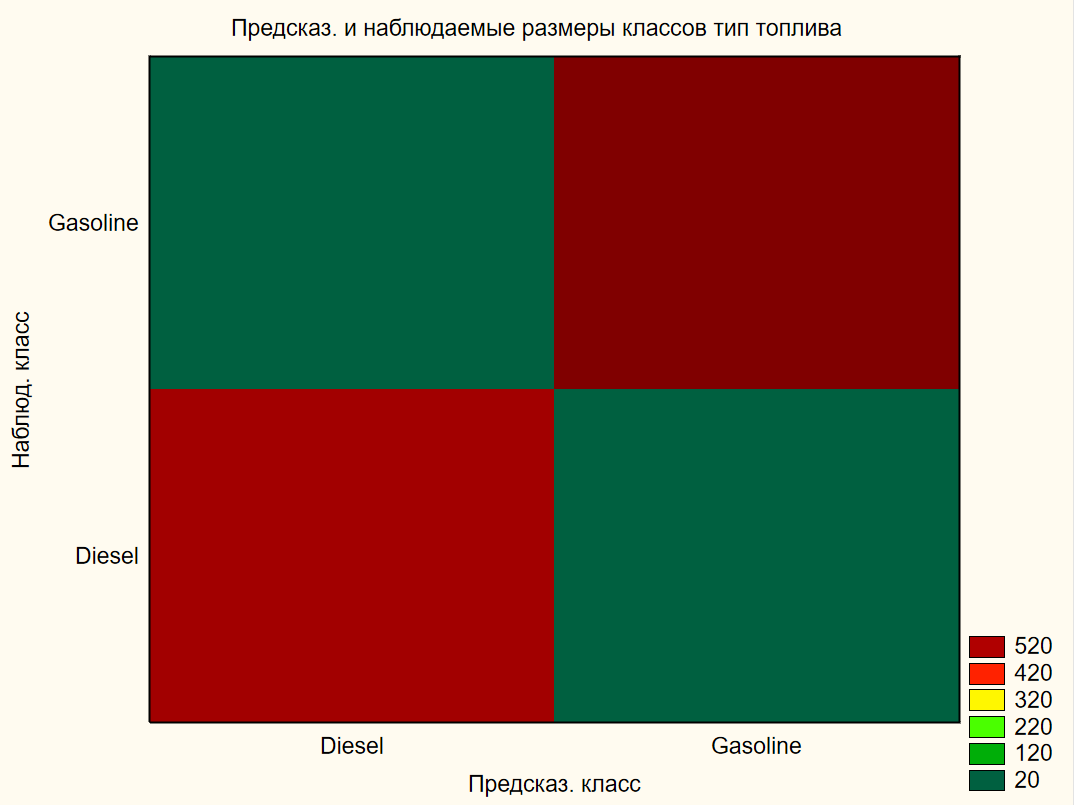


В данной таблице показана значимость предиктора

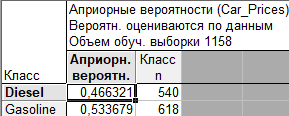


В данной таблице показана предсказанные и наблюдаемые объекты в классах.

На следующем рисунке показана дискретная карта линий уровня:

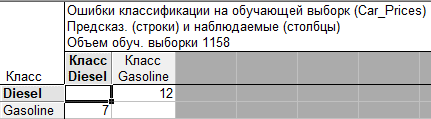


Априорные вероятности:



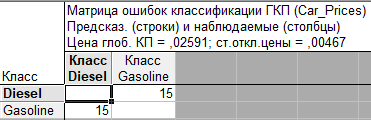
В этой таблице показаны априорные вероятности для каждого класса зависимой переменной и число элементов в каждом классе.

Ошибки классификации:



В этой таблице показано, сколько объектов каждого класса было ошибочно отнесено к другому классу.

В следующей таблице представлена матрица ошибок классификации глобальной кросс-проверки:



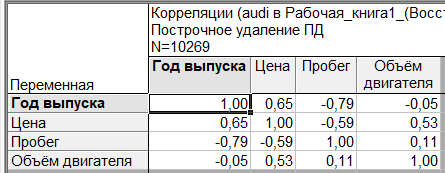
В данной таблице видно, что 15 автомобилей с Дизельным двигателем неверно классифицированы как автомобили с бензиновым, 15 машин бензинового типа неверно классифицированы как машины с дизельным. Кроме того, в таблице показана цена глобальной кросс-проверки и ее стандартное отклонение: 0,2591 и 0,00467 соответственно.

Таким образом, получено решающее правило, состоящее из одного этапа, которое позволит произвольный автомобиль классифицировать как дизельный или бензиновый.

# Методы редукции данных

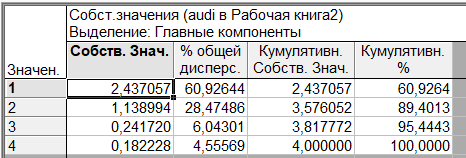
## Факторный анализ

Главная цель факторного анализа — это сокращение числа переменных и определение структуры взаимосвязей между переменными, т.е. классификация переменных



Из таблицы видно, что между переменными год выпуска и цена, пробег и цена, а также пробег и объём двигателя наблюдается заметная корреляция. Между переменными год выпуска и пробег наблюдается сильная корреляция.

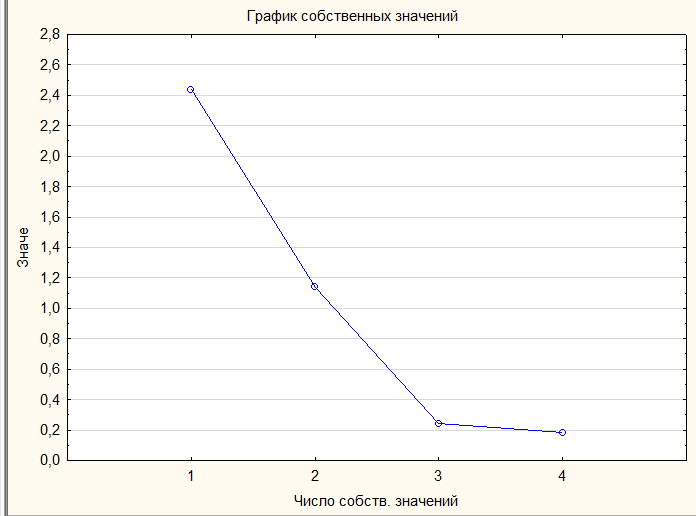
Просмотрим собственные значения факторов:



Во первом столбце приведены дисперсии выделенных факторов – собственные числа, во втором – процент от общей дисперсии. Как видно, первый фактор объясняет 59,53568% общей дисперсии, второй – 31,95491%. Необходимо понять, сколько факторов следует оставить.

По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются два фактора.

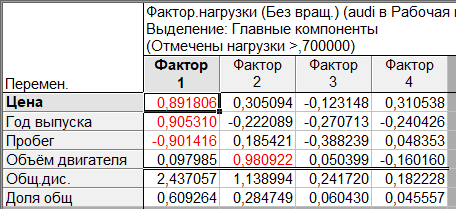
Критерий каменистой осыпи является графическим:

****

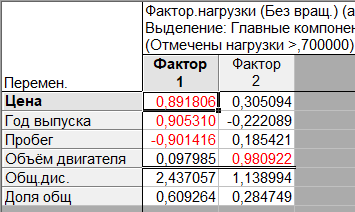
Надо выбрать такое место на графике, где убывание собственных значений слева направо максимально замедляется. Из графика видно, что в соответствии с этим критерием можно попытаться выделить 3 фактора.

Предположим, что число факторов неизвестно.

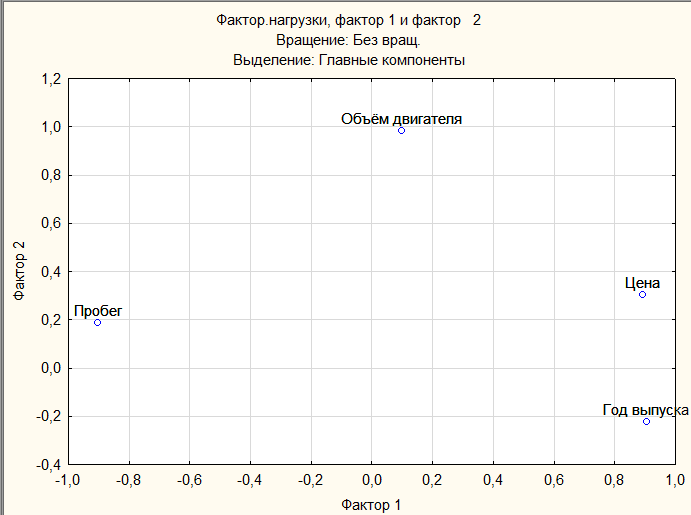
Корреляции между переменными и выделенными факторами:

****

Из таблицы видно, что первому и второму факторам соответствуют большие коэффициенты корреляции, чем остальным. Назначим число факторов 2.

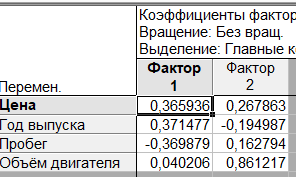


Из таблицы видно, что Фактор 1 имеет высокие факторные нагрузки по переменным год, пробег и цена и низкие по переменной объём двигателя. А Фактор 2 наоборот. Это означает, что выделенные два фактора наилучшим образом характеризуют данные.

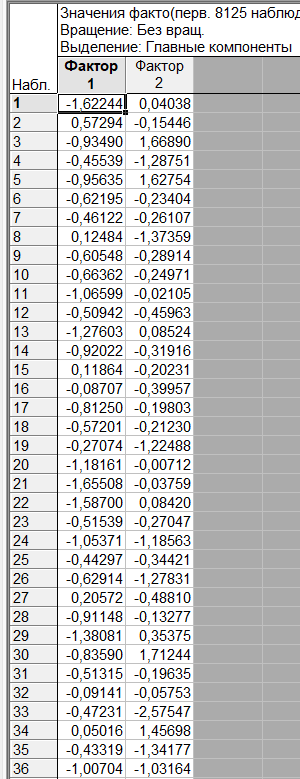
****

Этот график иллюстрирует соотношение между факторами и группами переменных. Видно, что группа переменной Объём Двигателя занимает на плоскости крайнее верхнее положение, группа переменной Пробег крайнее левое верхнее положение, группа переменной Год и Цена крайнее правое положение ниже.

Коэффициенты уравнений регрессий, по которым программа посчитает значения факторов для каждого наблюдения, представлены в таблице:



Значения факторов для каждого автомобиля представлены в таблице:



По данным из этой таблицы можно судить об отношении автомобилей к соответствующему фактору. Положительное значение фактора соответствует позитивному отношению автомобиля, а отрицательное – негативному.