**Федеральное агентство по образованию Российской Федерации**

**Государственное образовательное учреждение**

**высшего профессионального образования**

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**Кафедра прикладной математики**

Преподаватель,

д.т.н. А.А. Халафян

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

**ОТЧЕТ**

**СТАТИСТИЧЕСКИЙ АНАЛИЗ АВТОМОБИЛЕЙ**

Работу выполнил студент 4 курса

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

Краснодар 2023

СОДЕРЖАНИЕ

[**1 Исходные данные 3**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138293)

[**2 Графический анализ 5**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138294)

[**2.1 2D Graphs 5**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138295)

[**2.1.1 2D Histogramms 5**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138296)

[**2.1.2 2D Scatterplots 6**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138297)

[**2.1.3 2D Box Plots 9**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138298)

[**2.2 Средство «закрашивание» 10**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138299)

[**2.3 3D SequentialGraphs 12**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138300)

[**2.3.1 Raw Data Plots 12**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138301)

[**2.3.2 Bivariate Histograms 13**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138302)

[**2.3.3 Box Plots 13**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138303)

[**3 Основные статистики 15**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138304)

[**3.1 Описательные статистики 15**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138305)

[**3.2 Корреляционная матрица 15**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138306)

[**3.3 Критерий Стьюдента сравнения средних 16**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138307)

[**3.3.1 t-test, independent, by groups 16**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138308)

[**3.3.2 t-test, independent, by variables 16**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138309)

[**3.3.3 t-test, dependent samples 16**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138310)

[**3.3.4 t-test, single samples 17**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138311)

[**3.4 Группировка и однофакторная ANOVA 17**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138312)

[**4 Частотный анализ 19**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138313)

[**4.1 Таблицы частот 19**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138314)

[**4.2 Таблицы кросстабуляции 21**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138315)

[**5 Корреляционный анализ 23**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138316)

[**6 Дисперсионный анализ 24**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138317)

[**7 Линейное многомерное моделирование взаимосвязей 27**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138318)

[**8 Нелинейное многомерное моделирование взаимосвязей 29**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138319)

[**8.1 Fixed Nonlinear Regression 29**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138320)

[**8.2 Логит регрессия 31**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138321)

[**8.3 Пробит регрессия 32**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138322)

[**8.4 Экспоненциальная регрессия 33**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138323)

[**8.5 Кусочно-линейная регрессия 34**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138324)

[**8.6 Определенная пользователем регрессия 35**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138325)

[**9 Канонический анализ 36**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138326)

[**10 Дискриминантный анализ 41**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138327)

[**11 Классификационный анализ без обучения 48**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138328)

[**11.1 Кластерный анализ 48**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138329)

[**11.1.1 Метод к-средних 48**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138330)

[**11.1.2 Двухвходовая кластеризация 50**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138331)

[**11.1.3 Древовидная кластеризация 52**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138332)

[**11.2 Деревья классификации 53**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138333)

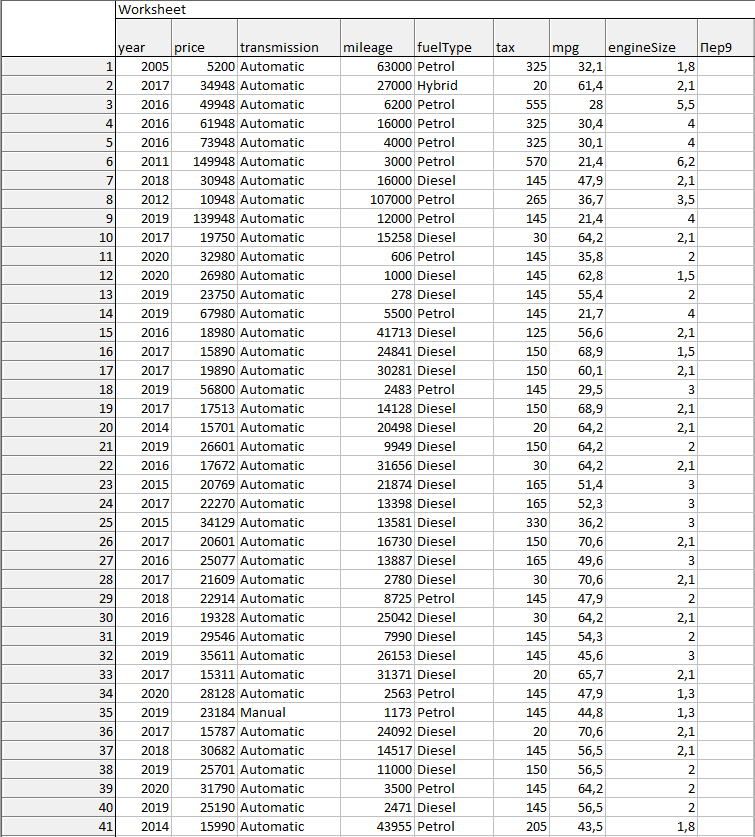
[**12 Методы редукции данных 57**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138334)

[**12.1 Факторный анализ 57**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138335)

[**12.2 Метод анализ главных компонент и классификация 61**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138336)

[**13 Многомерное шкалирование 65**](file:///K:\Халафян\primer%20otcheta.docx#_Toc233138337)

**1 Исходные данные**



В исходной таблице представлены данные по популярным автомобилям.

Показатели:

1. Year – год выпуска автомобиля.
2. Price – цена автомобиля при выпуске из завода.
3. Transmission – коробка передач.
4. Mileage – пробег.
5. FuelType – тип топлива.
6. Tax – налог на машину.
7. Mpg – ЛС.
8. EngineSize – размер двигателя.

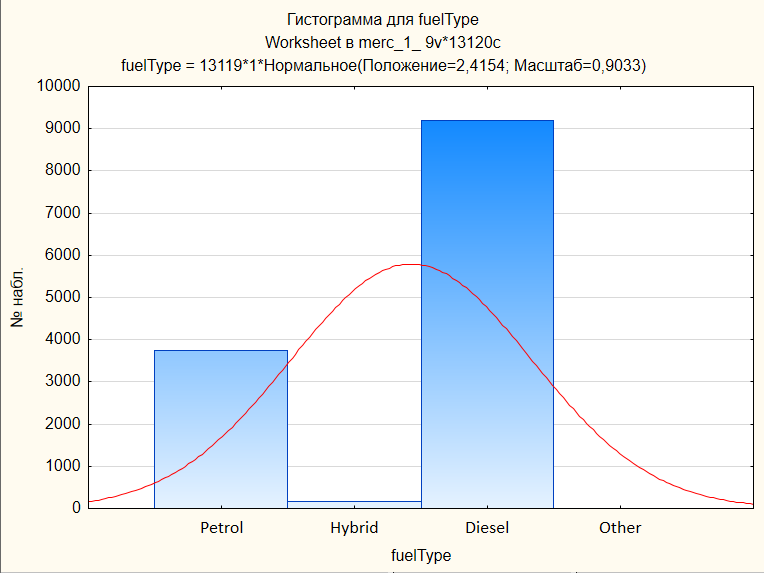
**2 Графический анализ**

**2.1 2D Graphs**

**2.1.1 2D Histogramms**

2D Histogramms являются графическими представлениями распределения частот выбранных переменных.

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



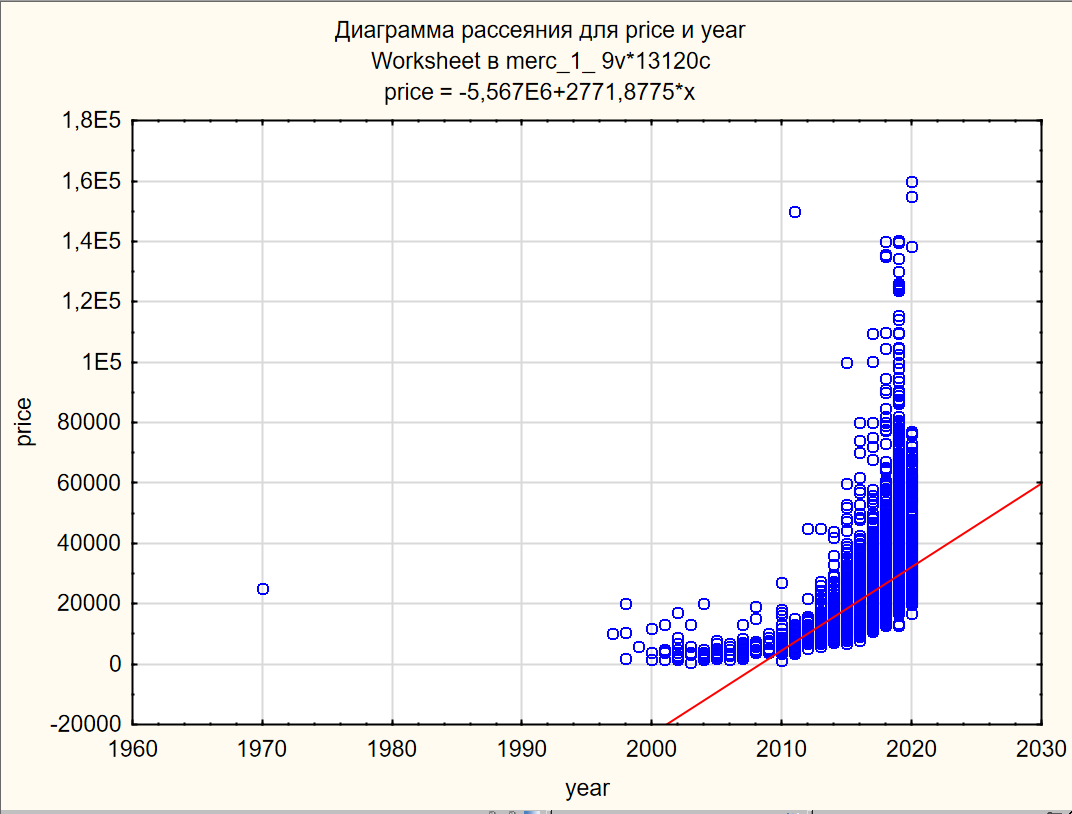
Данная 2D Histogramms Regular построена используя данные «fuelType»». Из гистограммы видно, что чаще всего встречаются автомобили, использующие бензин (9150 из 13000).

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

**2.1.2 2D Scatterplots**

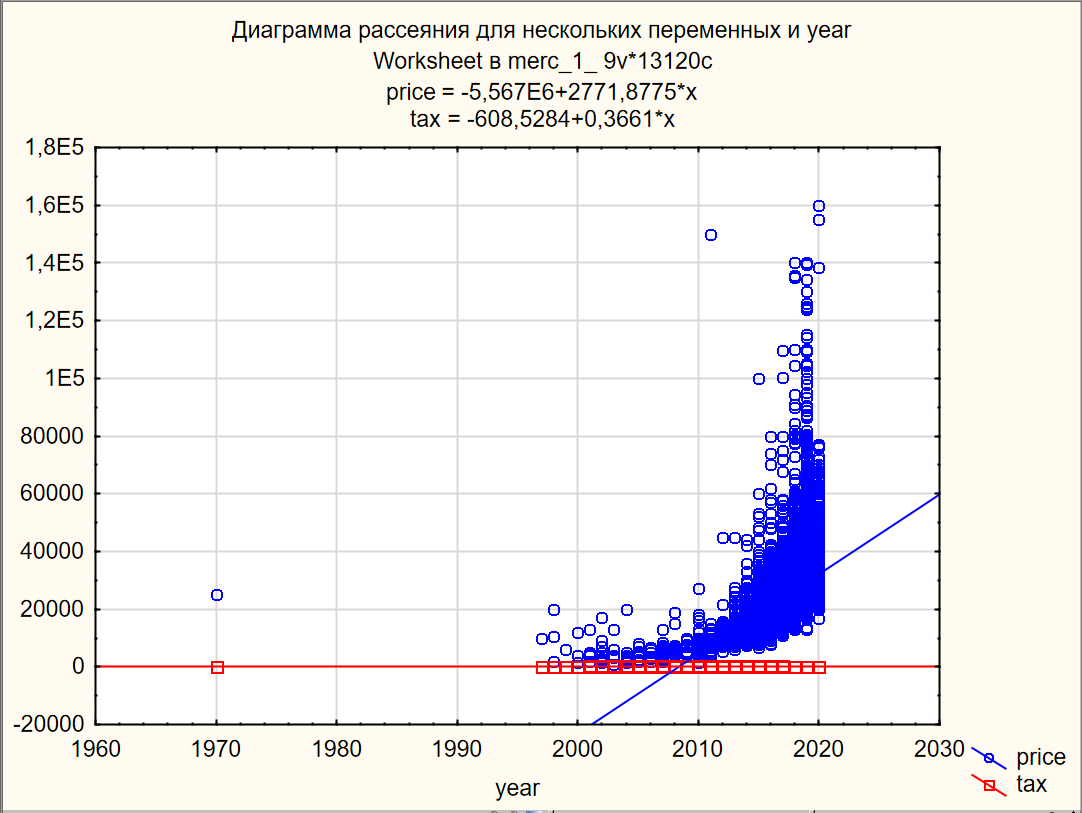
2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными.

2D Scatterplots Regular (протые) - визуализируют зависимость между двумя переменными X и Y.



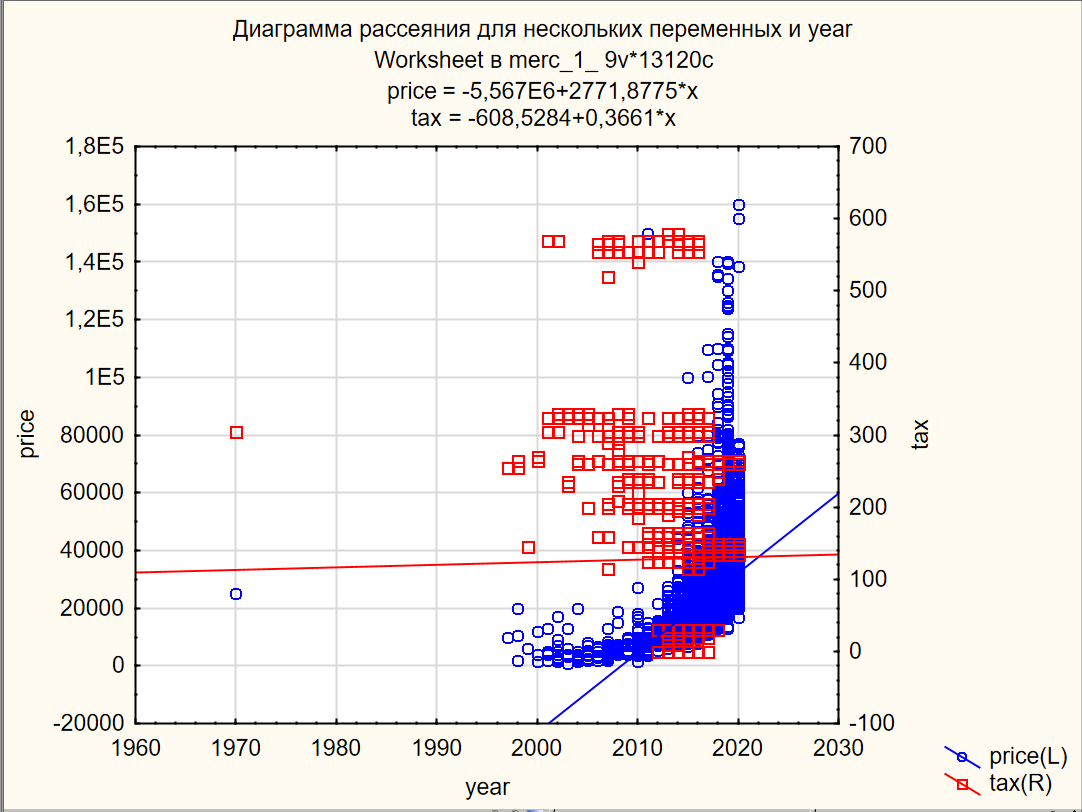
Данная 2D Scatterplots Regular диаграмма отображает зависимости между годом выпуска и ценой машины.

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



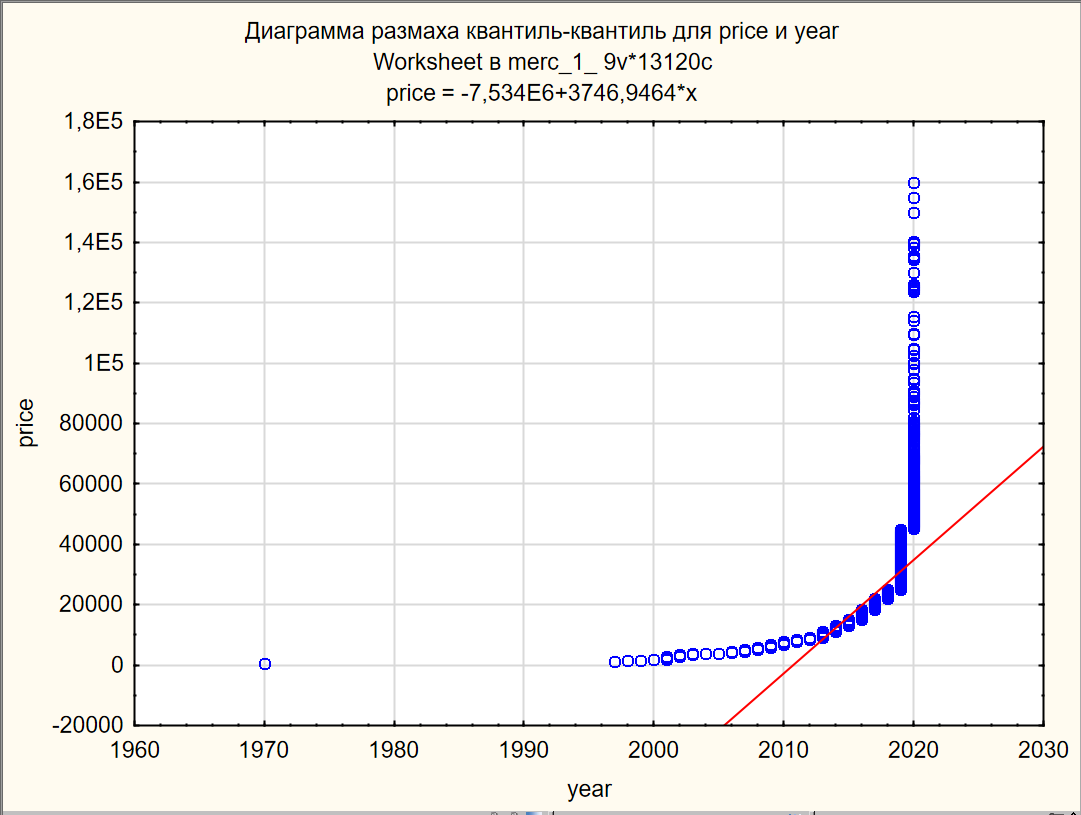
Данная 2D Scatterplots Multiple диаграмма отображает зависимости между годом выпуска и ценой машины и между годом выпуска и налогом на машину.

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 и ценой машины, вторая зависимость между годом выпуска и налогом на машину.

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



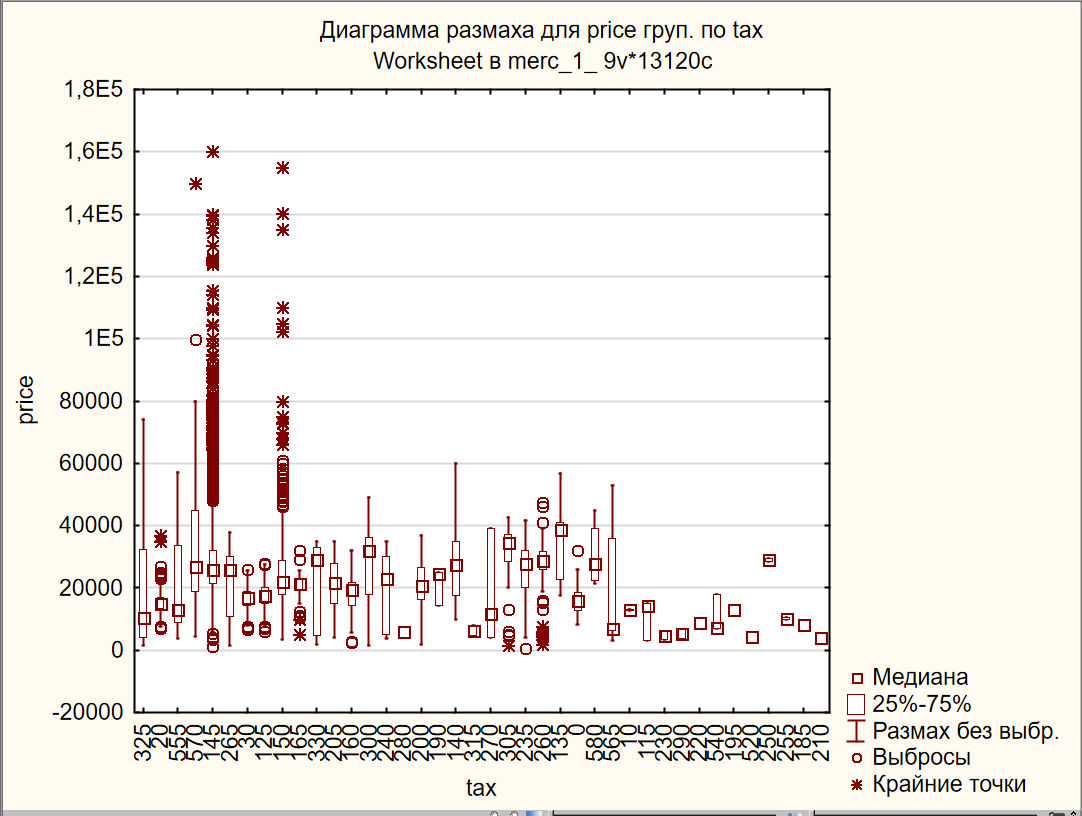
Данная 2D Scatterplots Quartile отображает зависимость между годом выпуска и ценой машины.

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 количество выплат и выплаты.

**2.1.3 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.

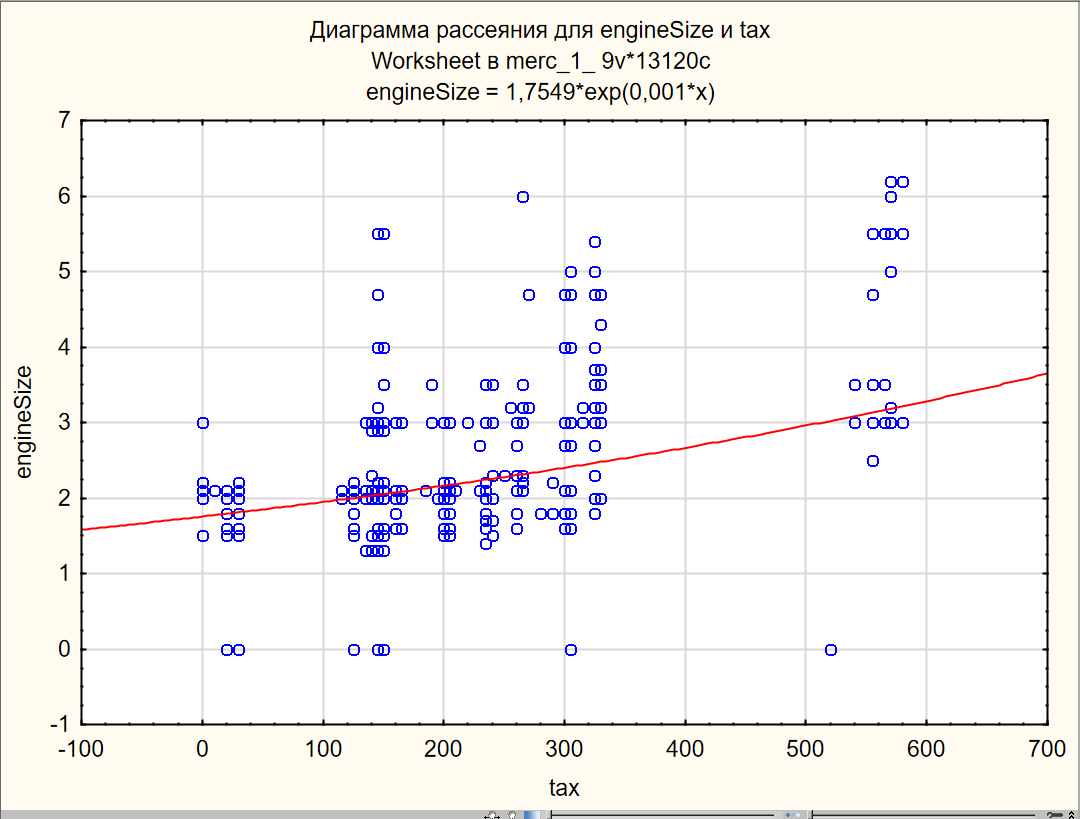


На этой 2D Box Plots Regular диаграмме представлены характеристики распределения значений переменной изменение за год по группам прибыльных или убыточных страховщиков.

На этой 2D Box Plots Multiple диаграмме представлены характеристики распределения значений переменных изменение за год и доля юрлиц по группам прибыльных или убыточных страховщиков.

**2.2 Средство «закрашивание»**

Построим диаграмму рассеивания для переменных количество заключённых договоров и количество выплат:



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. Предварительно можно узнать каким страховщикам соответствуют такие точки:



Теперь с помощью тех же средств закрашивания можно исключить эти точки:



Сравнивая r до и после исключения точки, можно сказать, что качество аппроксимации увеличилось (доля объясненной дисперсии).

**2.3 3D SequentialGraphs**

**2.3.1 Raw Data Plots**

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

соотношения между значениями переменных.



На данном графике показано соотношения значений между количеством заключённых договоров и количество выплат.

**2.3.2 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 страховая сумма и доля юрлиц.

**2.3.3 Box Plots**

Box Plots - это диаграммы размаха.

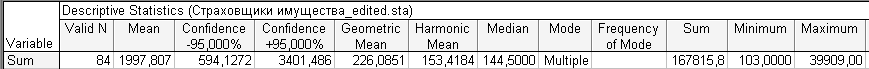
Построим диаграмму размаха для переменных изменение за год и доля юрлиц по группам прибыльных или убыточных страховщиков:



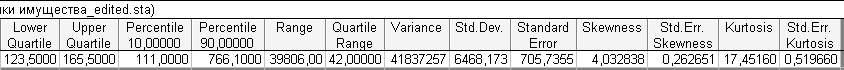
3 **Основные статистики**

**3.1 Описательные статистики**

Таблица описательных статистик для итогов по страхованию имущества:



*продолжение*



В таблице представлены следующие статистики: число наблюдений 84, среднее 1997.807, сумма 167815.8, медиана 144.5, геометрическое среднее 226.0851, гармоническое среднее 153.4184, стандартное отклонение 6468.173, дисперсия 41737257, стандартная ошибка среднего 705.7355, доверительные пределы для среднего 594.1272 и 3401.486, ассиметрия 4.032838, стандартная ошибка ассиметрии 0,262651, эксцесс 17.45160, стандартная ошибка эксцесса 0.519660, минимум 103, максимум 39909, нижний квартиль 123.5, верхний квартиль 165.5, 10-я процентиль (квантиль 0,1) 111, 90- процентиль (квантиль 0,9) 766.1, размах 39806, квартильный размах 42.

**3.2 Корреляционная матрица**

Из таблицы видно, что сильная корреляция только между количеством заключённых договоров и количеством выплат; средняя между кол. дог. и sum, между кол. вып. и sum, sum и финриски, sum и с-м риски и т.д., а слабая везде, где корреляция <=0.25. Красным отмечаны значения, где уровень значимости <0.05.

**t-test, independent**, by groups применяется если сравниваются средние количественной переменной значения которой разбиваются группирующей переменной на 2 группы!

**t-test, independent**, by variablesи применяется, если сравниваются 2 столбца одной и той же величины, но записанные для разных объектов! У вас в данных нет таких столбцов, поэтому этот анализ либо не проводим, либо придумываем 2 столбец, например цена автомобиля, но оговариваем,, что это цена другой группы авто!

**t-test, dependent** samples применяем, если сравнивается одна и та же величина при 2 повторных измерениях! Такой столбец надо также придумать!

**t-test, single samples** применяем, если сравниваем среднее группы с средним генеральной совокупности, которую тоже надо придумать!

**Группировка и однофакторная ANOVA** применяем, если сравниваем более 2 групп!!!!!!!!

**Дисперсионный анализ,** если сравниваем подгруппы по двум  группирующим переменным

**Ранговый ДА и конкордация Кендалла** применяем, если сравниваем более 2 повторных измерений!!!!

**критериями знаков и Вилкоксона** применяем, сравнивается одна и та же величина при 2 повторных измерениях

**Критерий манна-Уитни, Вальда-Вольфовица, Колмогорова-Смирнова** (у вас их нет) если сравниваются средние количественной переменной значения которой разбиваются группирующей переменной на 2 группы!

Если групп более 2, то применяем **критерий Краскела-Уоллиса**

**Дискриминантный анализ** проводим по группирующей переменной, полученной в результате кластеризации не менее, чем на 3 группы!

**3.3 Критерий Стьюдента сравнения средних**

**3.3.1 t-test, independent, by groups**

Проверим равенство средних для всех переменных по группам прибыльных и убыточных страховщиков на основе t-критерия (t-test, independent, by groups).