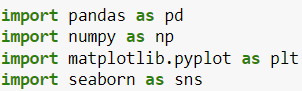
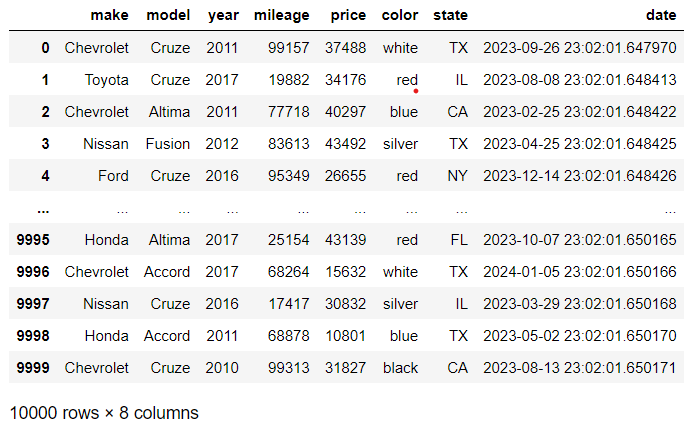
**Отчет по практикуму (1-6)**

В качестве датасета я взял цены автомобилей Chevrolet, Toyota, Nissan, Ford и Honda в США. Помимо цены, в таблице содержится информация о годе выпуска, пробеге, цвете машины, а также о дате продаже, и в каком штате происходила сделка. Вот ссылка на данные:

<https://www.kaggle.com/datasets/at3191/us-car-prices>



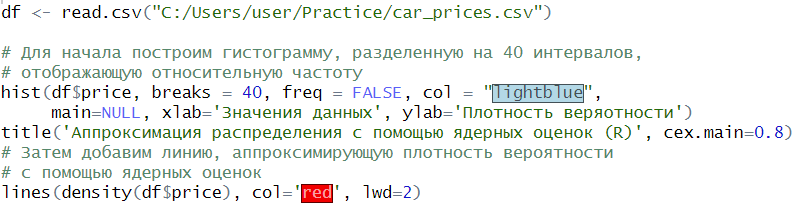




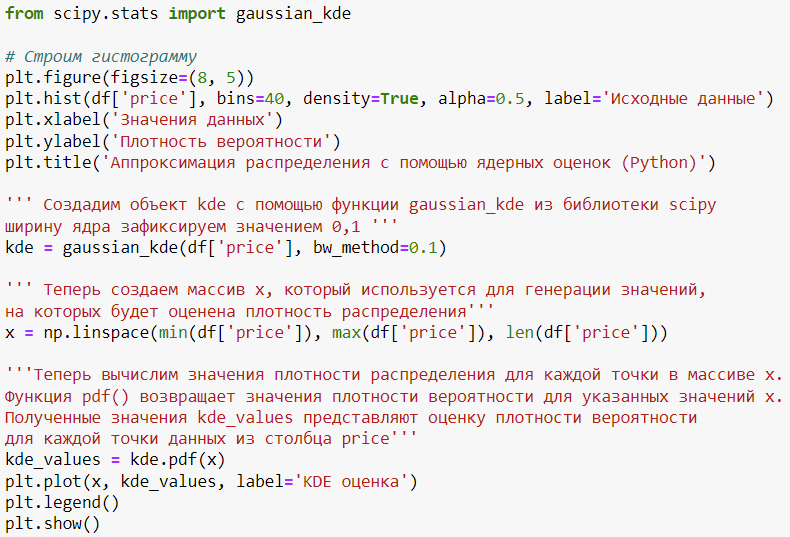
Для выполнения заданий на Python я использовал Jupyter Notebook. Для выполнения заданий на R использовал RStudio.

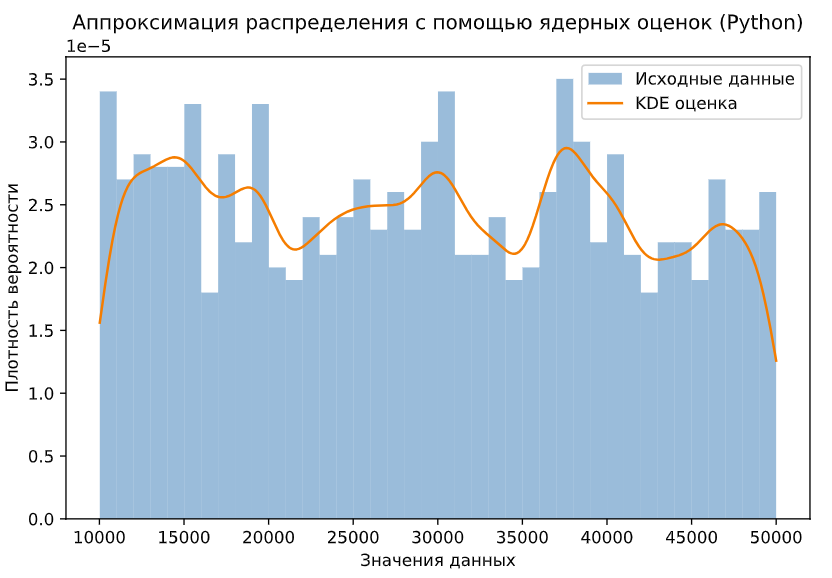
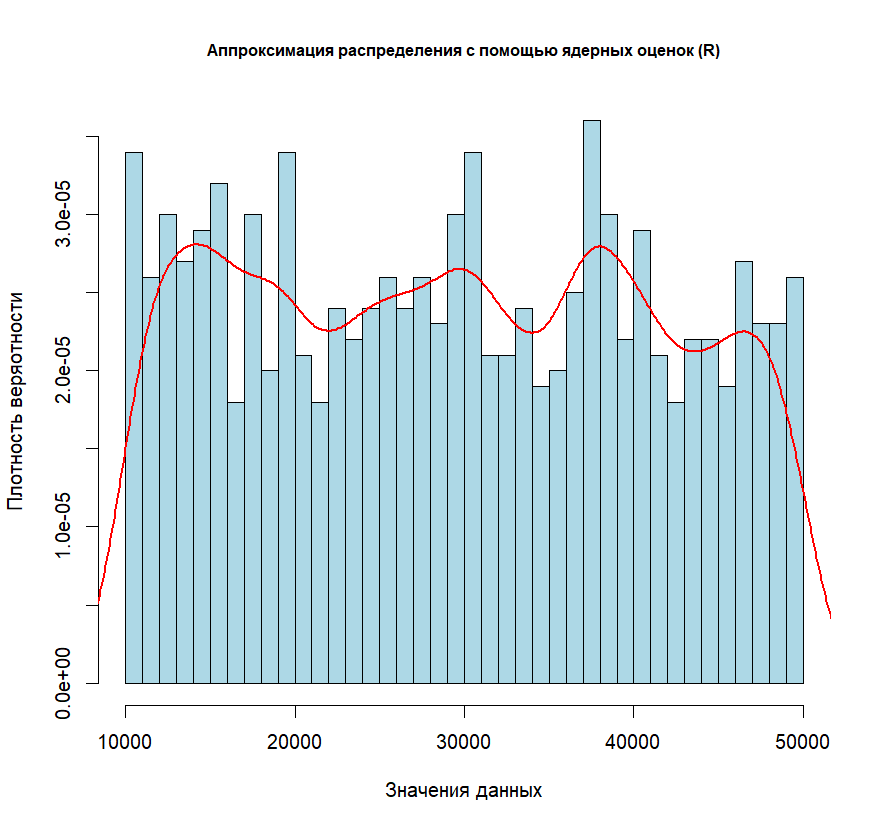
**1. Реализовать аппроксимацию распределений данных с помощью ядерных оценок.**

**R:**



**Python:**

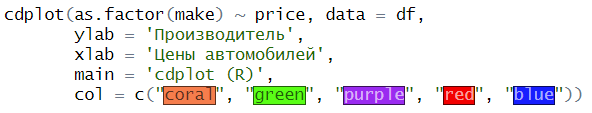




**2. Реализовать анализ данных с помощью cdplot, dotchart, boxplot и stripchart.**

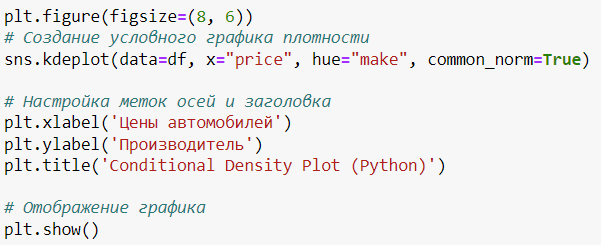
C помощью **cdplot (Conditional Density Plot)** построим график условных плотностей распределения цен автомобилей в зависимости от их производителя.

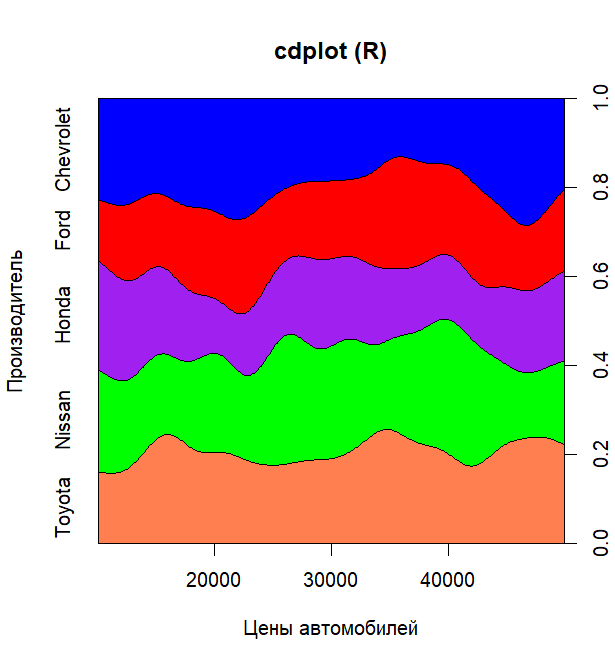
**R:**

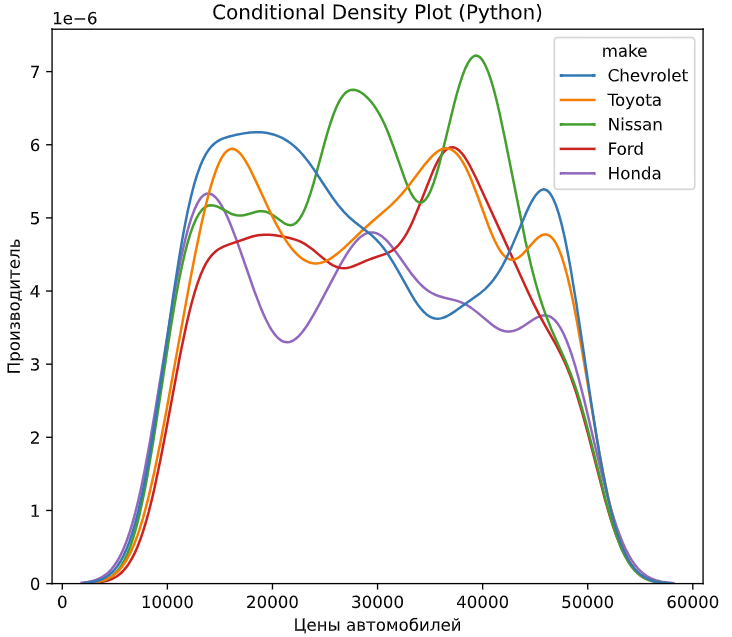


**Python:**

На Python нет функции cdplot, но для построения графика условных плотностей я использовал функцию **kdeplot** из библиотеки **Seaborn**.

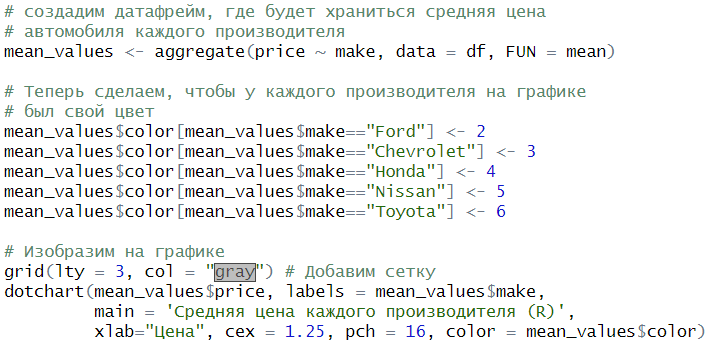


****

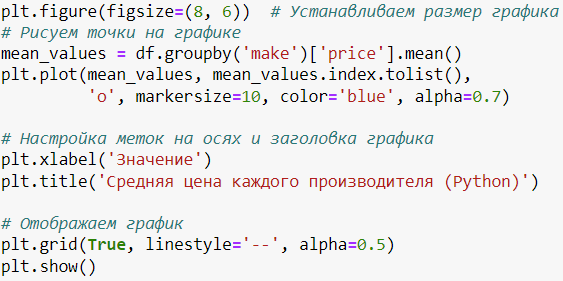
****

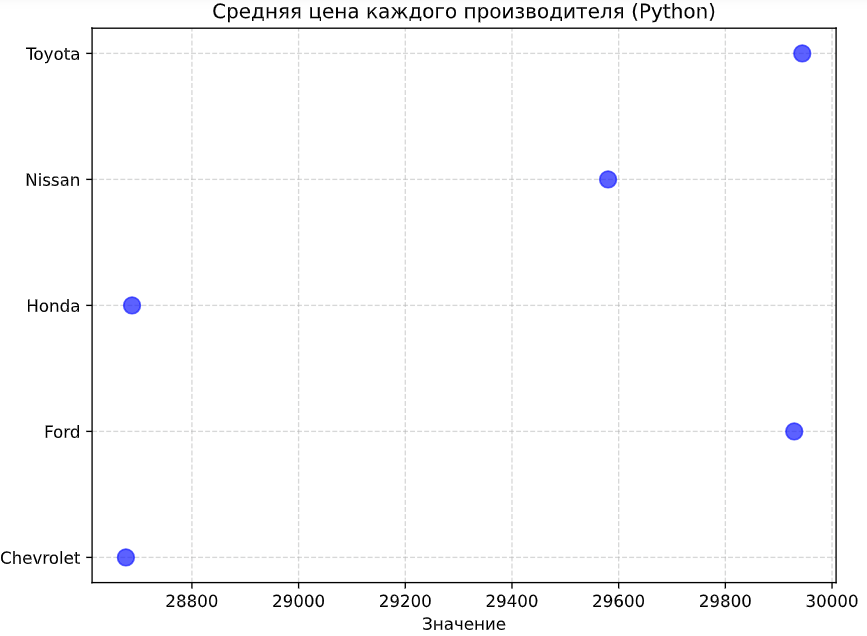
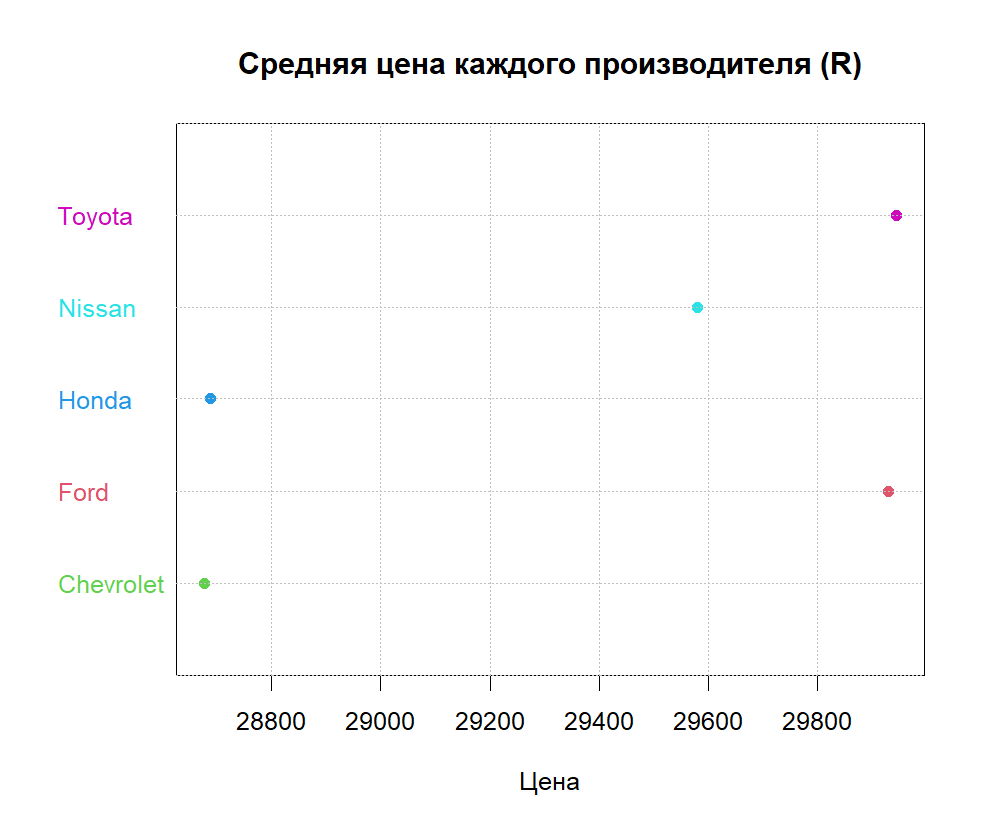
С помощью **dotchart** изобразим среднюю цену автомобилей каждого производителя

**R:**



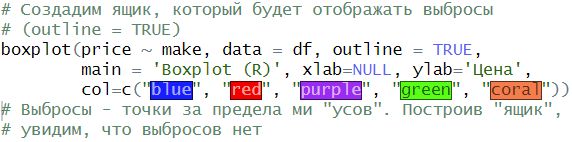
**Python:**



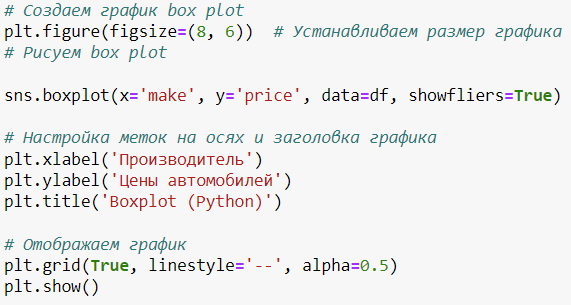


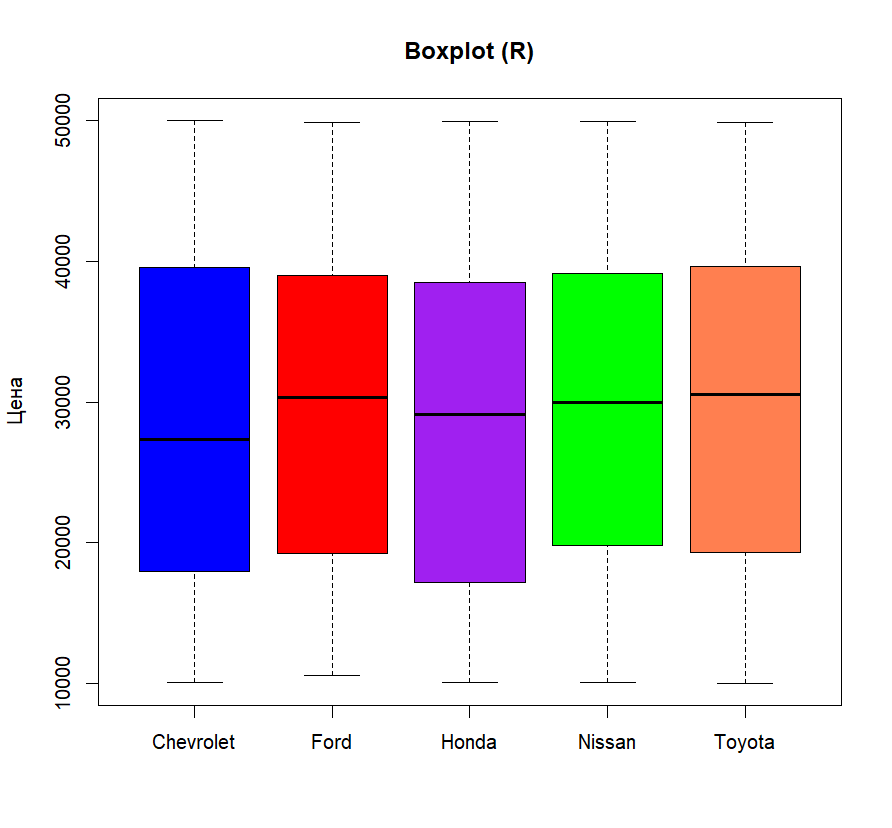
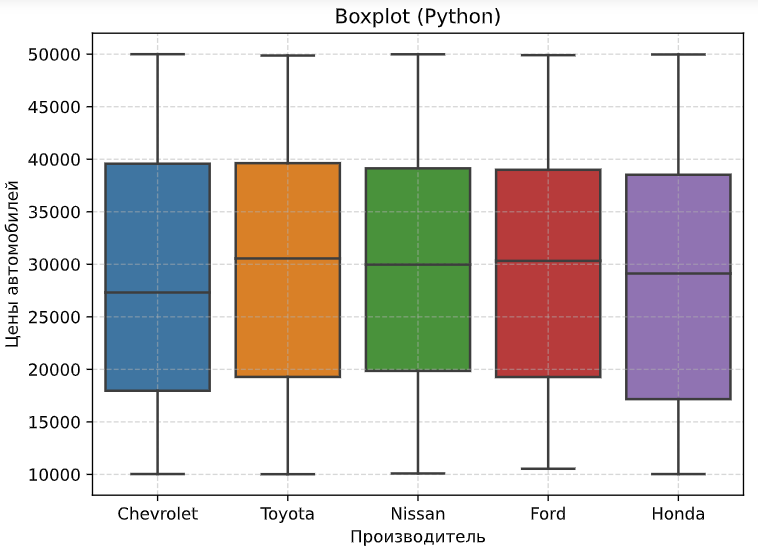
Теперь создадим **boxplot** («ящик с усами»), который отображает максимальное и минимальное значения, верхний и нижний квартили, медиану и выбросы, для цены автомобиля в зависимости от производителя. Выбросы – точки за пределами «усов». Построив ящик, увидим, что выбросов нет.

**R:**



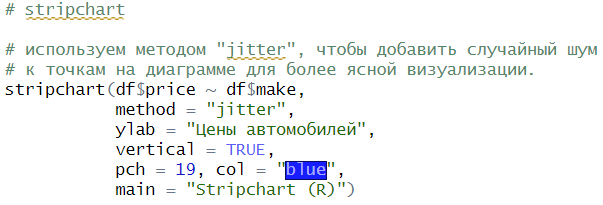
**Python:**



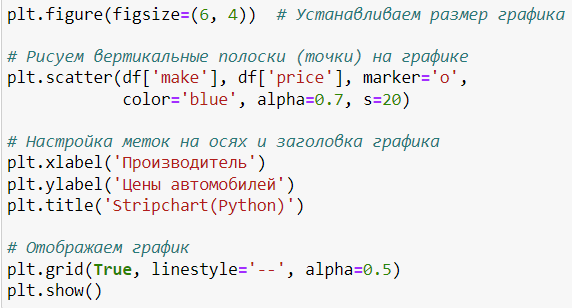
 

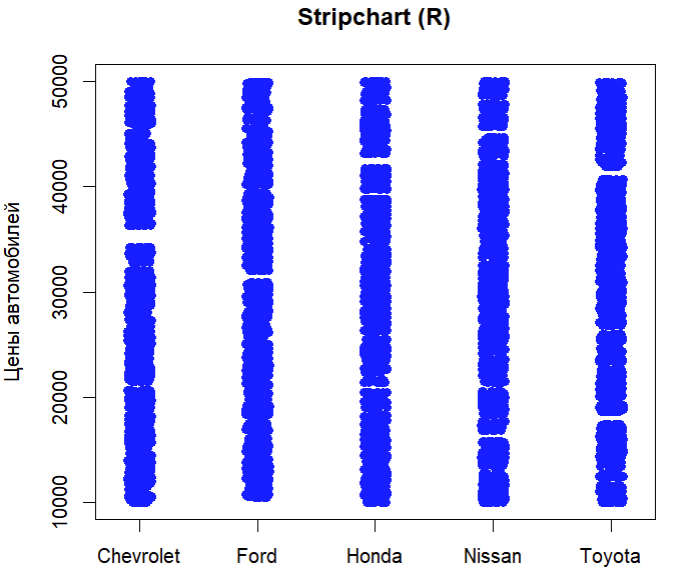
С помощью **stripchart** выясним, как распределены цены автомобилей каждого производителя.

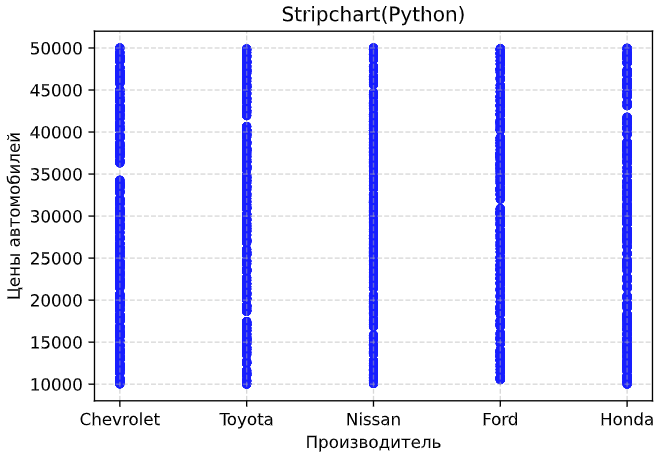
**R:**



**Python:**



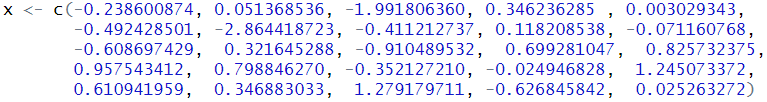




На графике видно, что цены каждого производителя имеют почти все значения от 10 000 до 50 000 за исключением некоторых точек. Можно заметить, что автомобили от производителя Chevrolet, к примеру, не имеют цену, равную 35 000 долларов.

**3. Проверить, являются ли наблюдения выбросами с точки зрения формальных статистических критериев Граббса и Q-теста Диксона. Визуализировать результаты.**

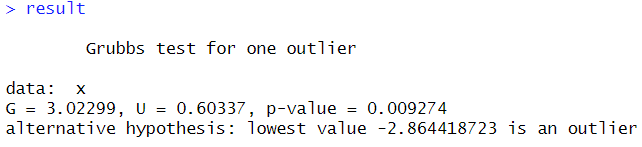
Критерий Граббса и Q-тест Диксона применяется к данным, имеющим нормальное распределение. Я взял выборку объемом 25 со следующими значениями:



Теперь применим эти два теста для выяснения того, является ли минимальное значение выбросом.

**Реализация на R:**

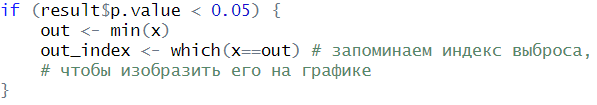




Основная гипотеза: минимальное значение не является выбросом.

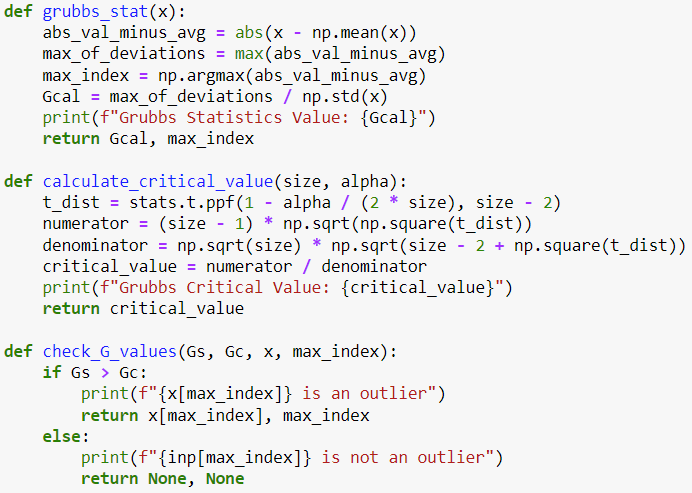
Альтернативная: минимальное значение является выбросом.

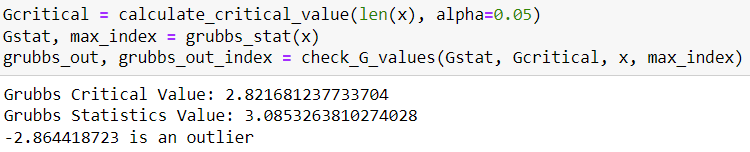
Если p-value > 0.05 (уровень значимости), то мы не отвергаем основную гипотезу. В противном случае принимаем альтернативную. В данном случае p-value = 0.009724 < 0.05, следовательно, мы принимаем альтернативную гипотезу и заключаем, что минимальное значение является выбросом.



**Python**:

В Python нет встроенной функции. При реализации я опирался на статью в википедии: <https://ru.wikipedia.org/wiki/%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%93%D1%80%D0%B0%D0%B1%D0%B1%D1%81%D0%B0>

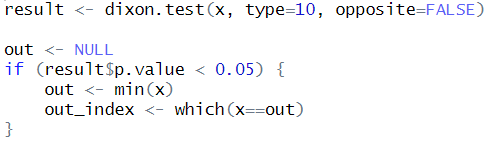


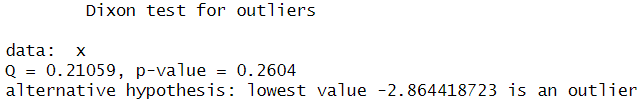


Видим, что значения критерия G в R и Python получились примерно одинаковыми.

**Теперь применим Q-тест Диксона:**

**R:**

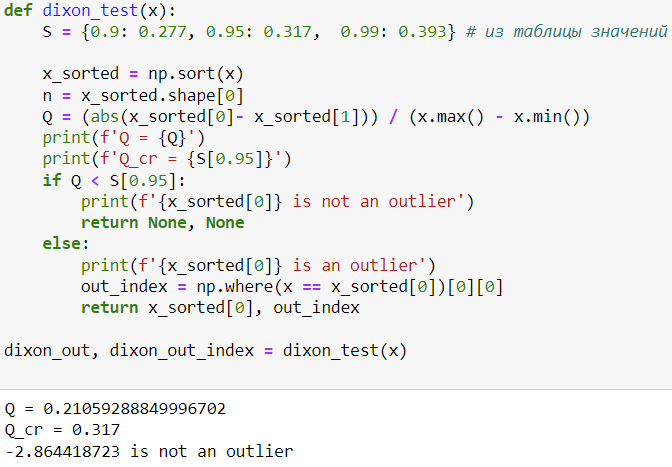
****



В данном случае p-value=0.2604 > 0.05, следовательно, мы не отвергаем основную гипотезу и заключаем, что минимальное наблюдание не является выбросом точки зрения Q-теста Диксона.

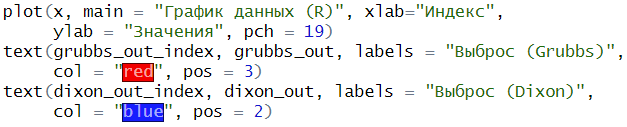
**Python:**

В Python нет встроенной функции. При реализации я опирался на статью: <https://wiki5.ru/wiki/Dixon%27s_Q_test>

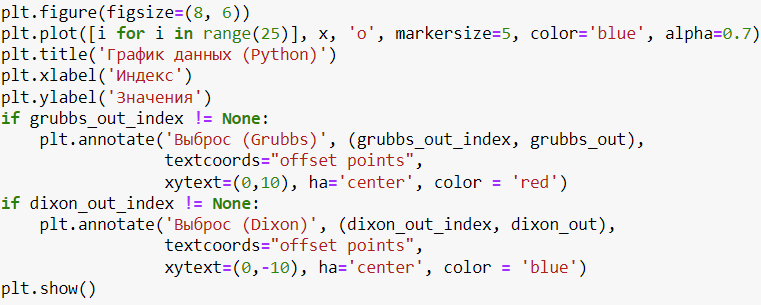


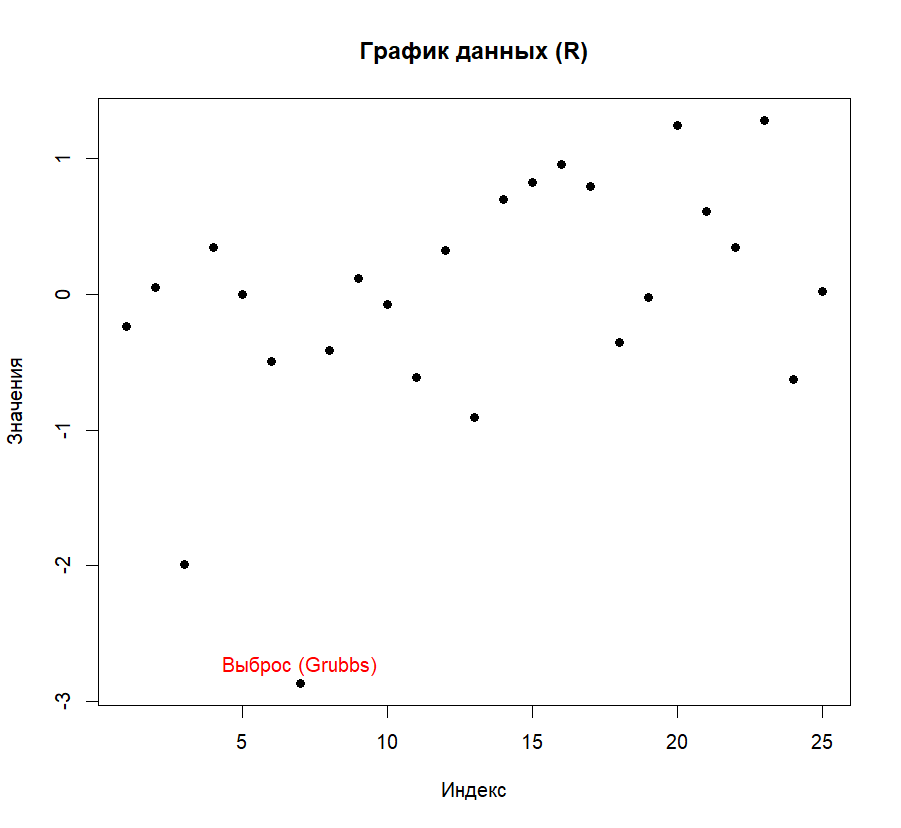
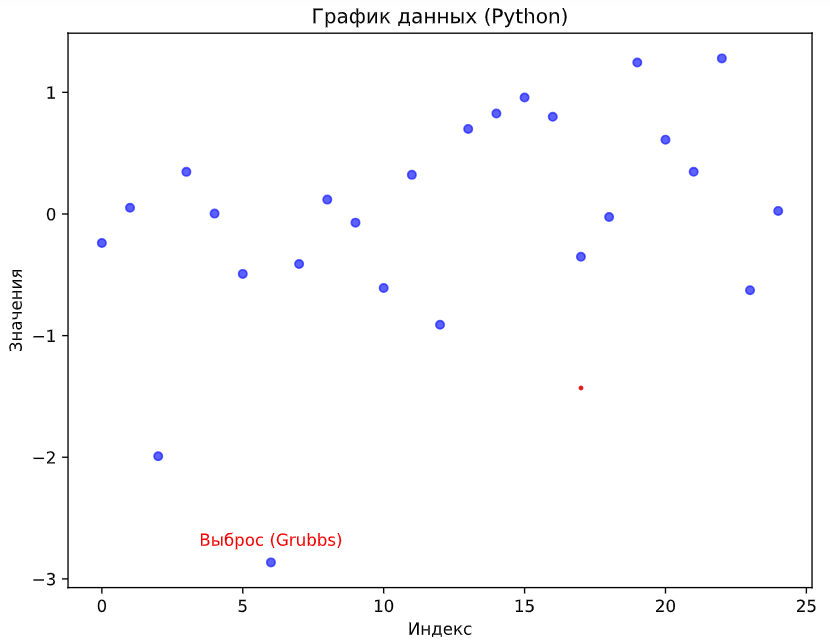
Мы видим, что в обеих реализациях значение Q одинаковое, равное 0.21059. Теперь визулизируем результаты.

**R:**



**Python:**



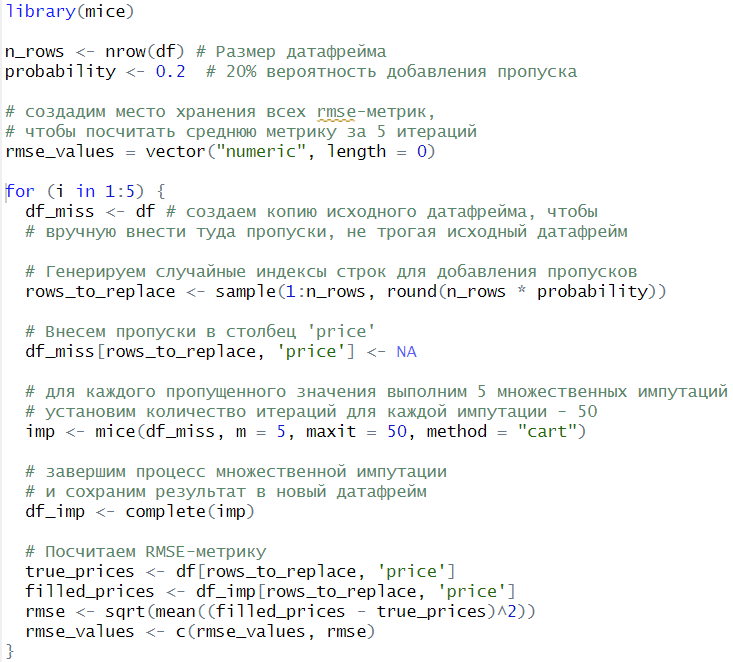
 

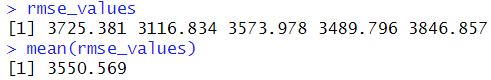
**4. Воспользоваться инструментами для заполнения пропусков в данных. Пропуски внести вручную и сравнить результаты заполнения с истинными значениями.**

**R:**

Я вручную внес пропуски в столбец с ценами на автомобили случайным образом 5 раз, чтобы посчитать среднюю RMSE-метрику за 5 итераций.

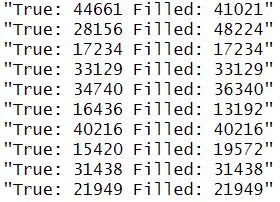
Для заполнения пропусков я пробовал различные методы: случайные леса, линейную регрессию и так далее, но большинство методов давали большое значение RMSE-метрики (около 16 тыс.). Наименьшую метрику показывал метод *cart (Classification and regression trees)*, равную 3550 долларов.



****

Выведем первые 10 истинных и заполненных значений.





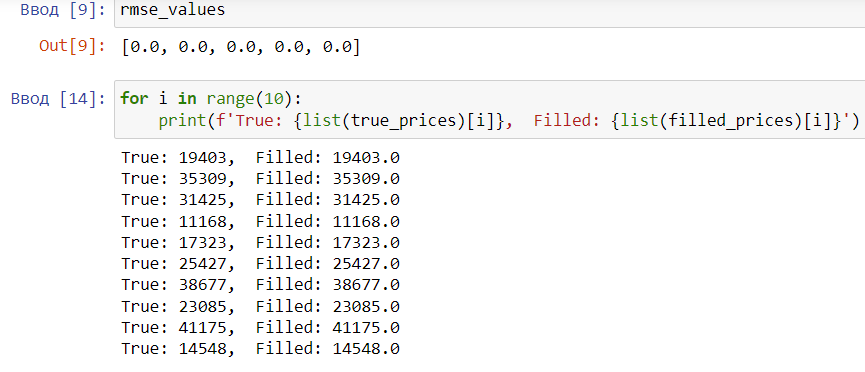
Видим, что некоторые значения в точности совпадают.

**Python:**

Для заполнения пропущенных значений на Python я использовал библиотеку **scikit-learn** с методом **IterativeImputer,** который базируется на модели **регрессионного дерева принятия решений (DecisionTreeRegressor).**

Это метод всегда давал нулевое значение RMSE-метрики, и все заполненные значения в точности совпадали с истинными.



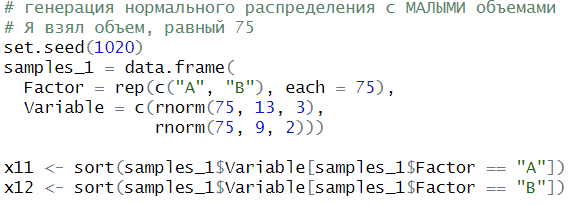


**5. Сгенерировать данные из нормального распределения с различными параметрами и провести анализ с помощью графиков эмпирических функций распределений, квантилей, метода огибающих, а также стандартных процедур проверки гипотез о нормальности (критерии Колмогорова-Смирнова, Шапиро-Уилка, Андерсона-Дарлинга, Крамера фон Мизеса, Колмогорова-Смирнова в модификации Лиллиефорса и Шапиро-Франсия). Рассмотреть выборки малого (не более 50-100 элементов) и умеренного (1000-5000 наблюдений) объемов.**

Для малого объема я сгенерировал 2 выборки с объемом 75 со следующими параметрами

1) матожидание = 13, стандартное отклонение = 3

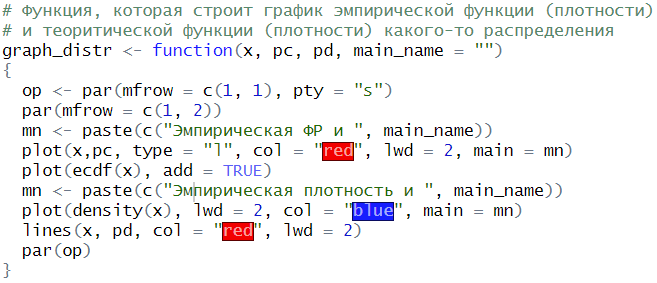
2) матожидание = 9, стандартное отклонение = 2

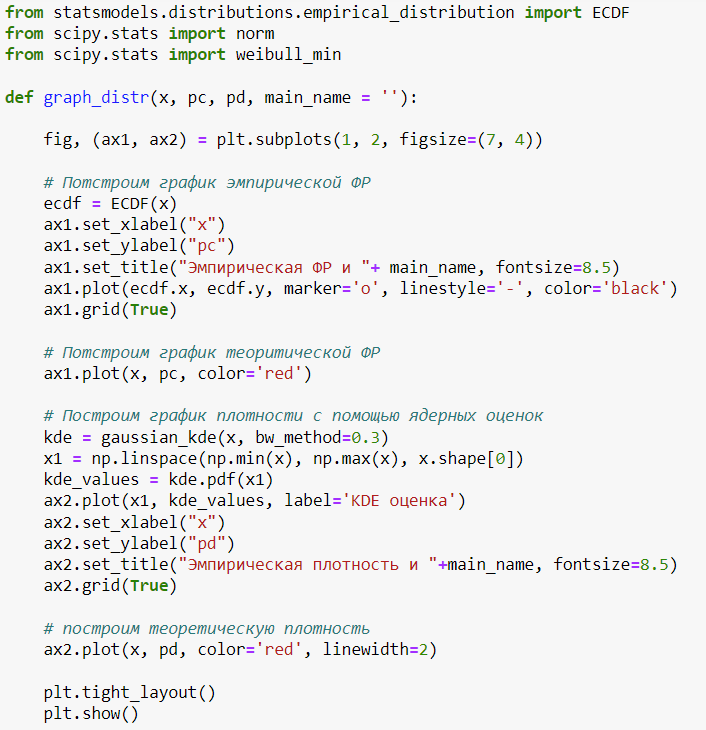
****



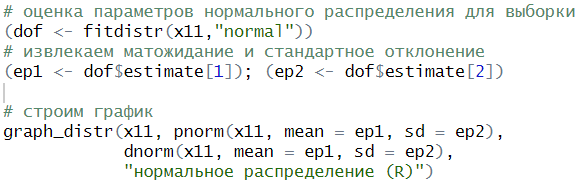
Проведем анализ с помощью графиков эмпирических фукнций распределения и теоритических функций **нормального распределения и распределения Вейбулла.**

Для начала определим функцию, которая строит график эмпирической функции (плотности) и теоритической функции (плотности) какого-то распределения.

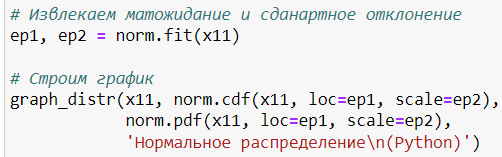


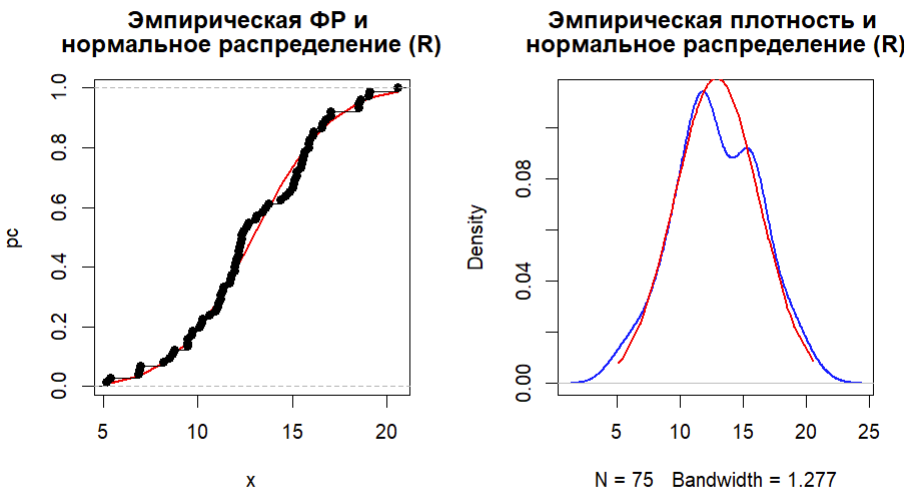


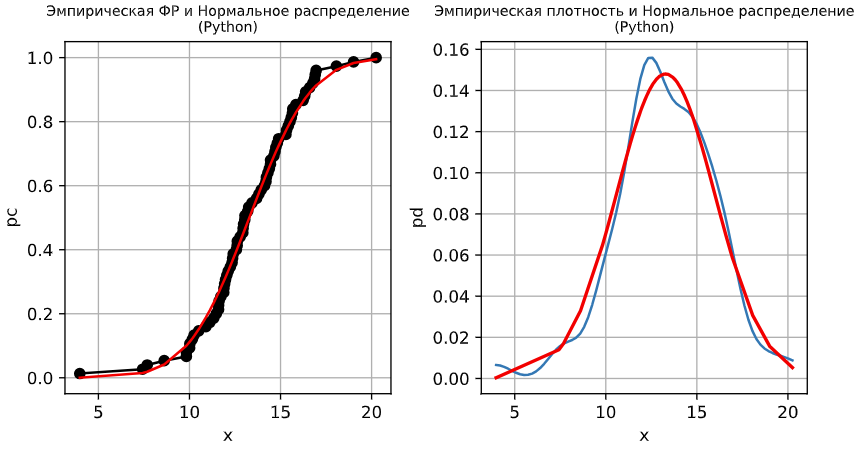
Исследуем первую выборку с малым объемом с параметрами (13, 3) на соотвествие **нормальному распределению**



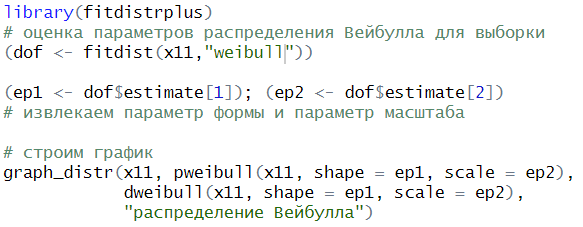


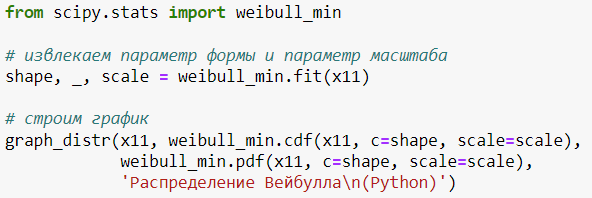


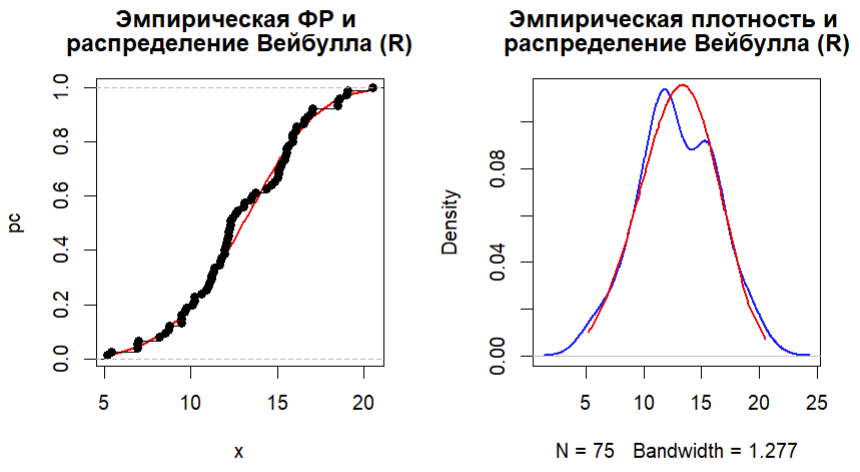


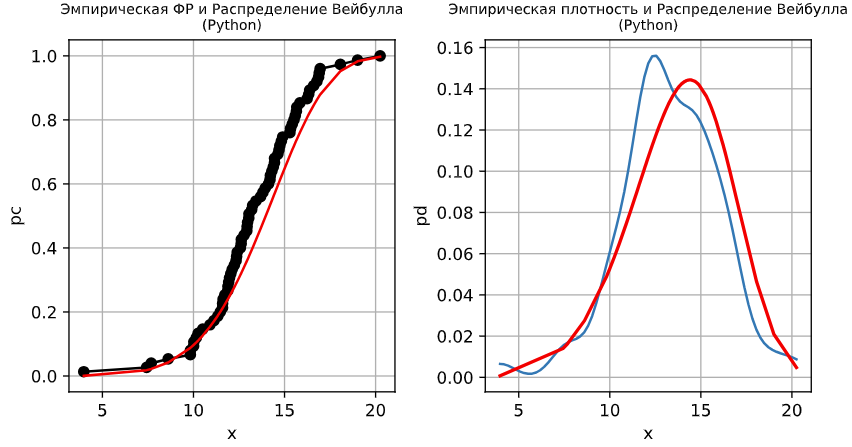


Теперь исследуем на соотвествие распределению **Вейбулла.**

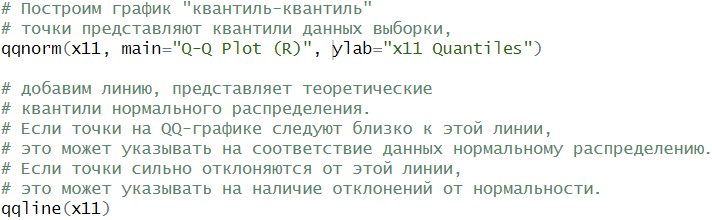


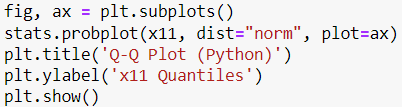


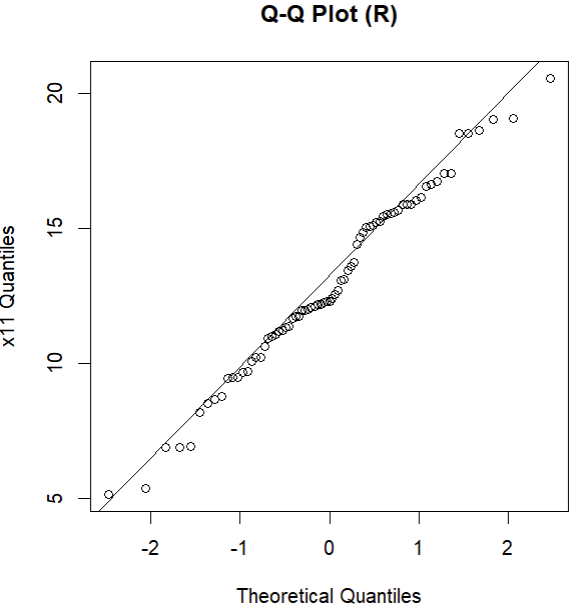


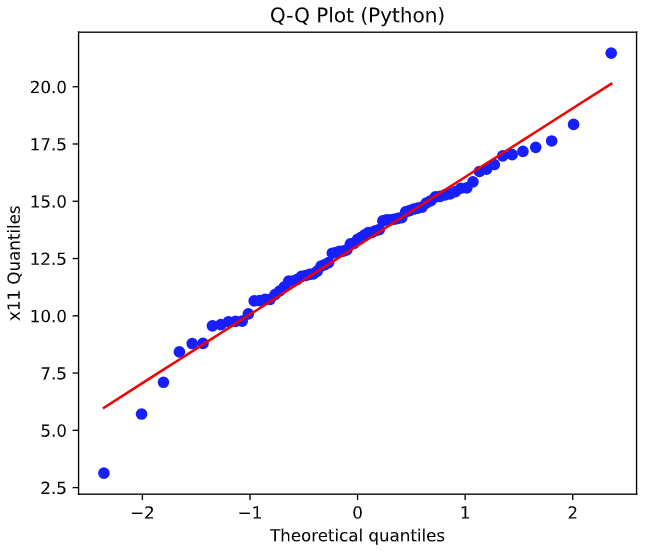


Теперь проведем анализ с помощью **графиков квантилей**.

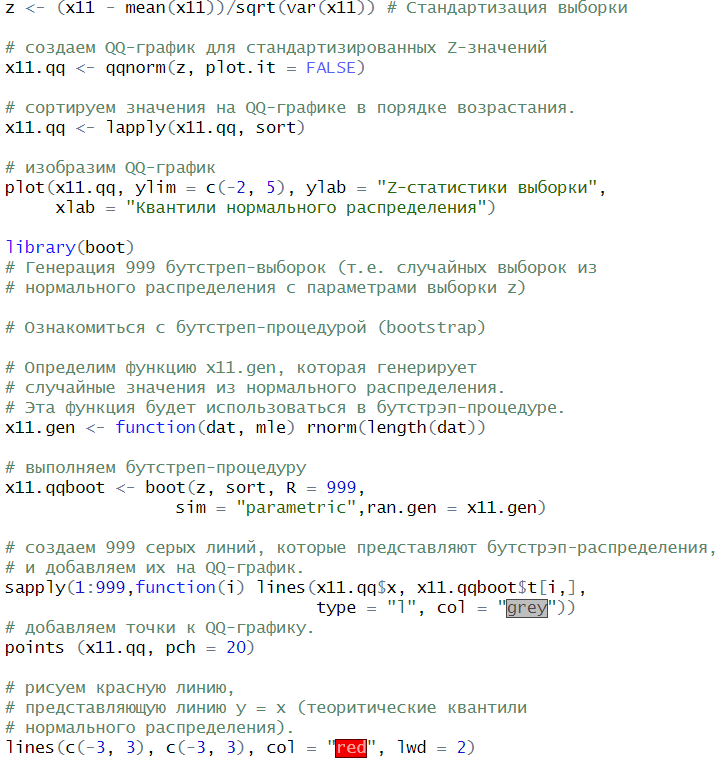


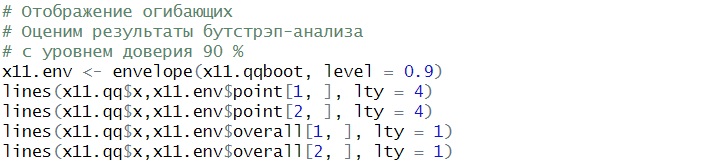


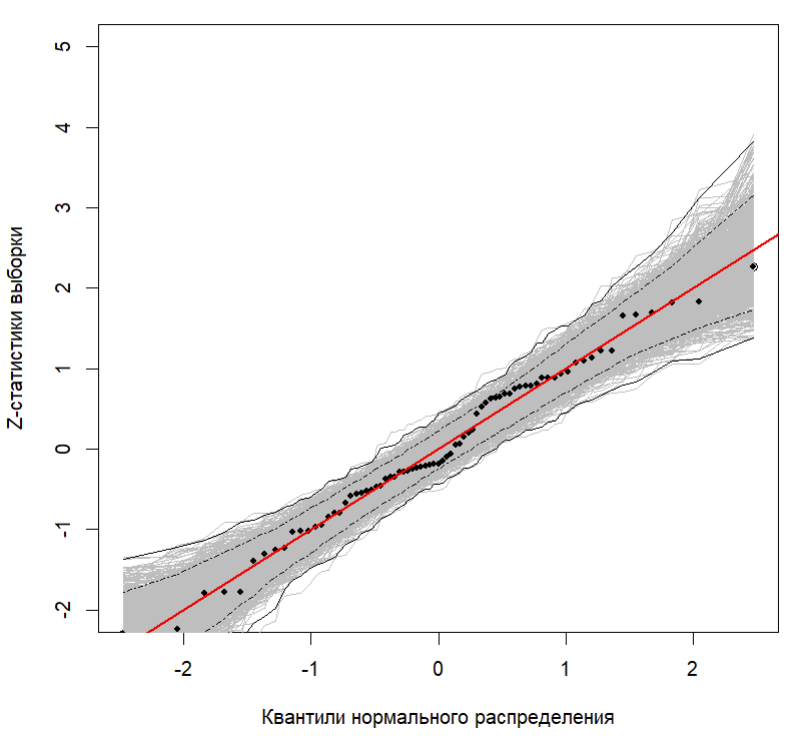




Используем **метод огибающих:**

****





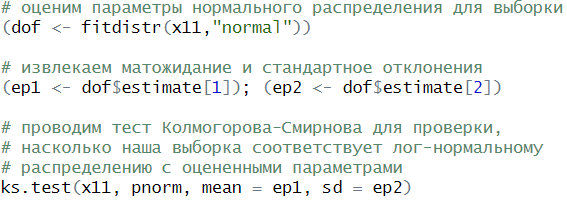
Теперь проверим **гипотезы о нормальности**

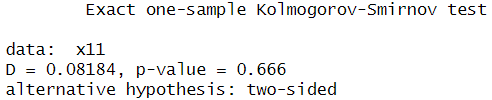
*Основная гипотеза – распределение может быть нормальным*

*Альтернативная – распределение не является нормальным*

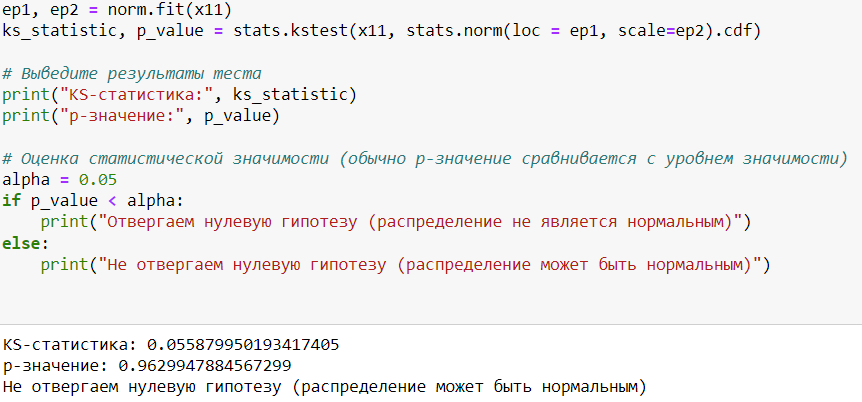
Если в следующих критериях значения p-value < 0.05 (уровень значимости), то мы отвергаем основную гипотезу и заключаем, что распределение не является нормальным. В противном случае – распределение может быть нормальным.

**R:**

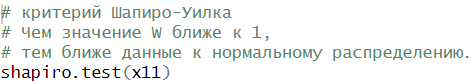
****

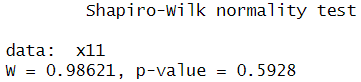
****

**Python**

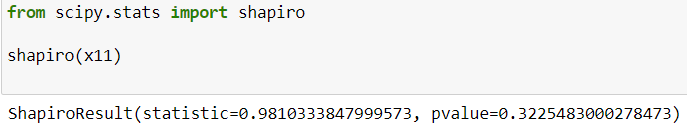
****

**R:**

****

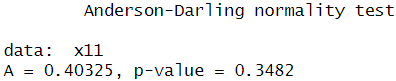
****

**Python:**

****

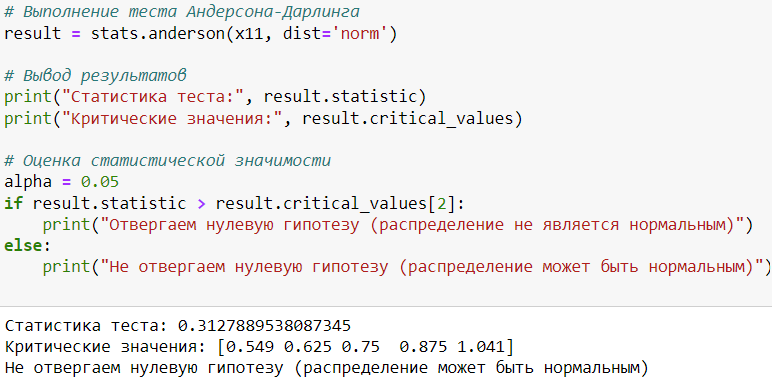
**R:**

****

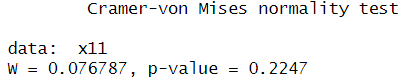
****

**Python:**

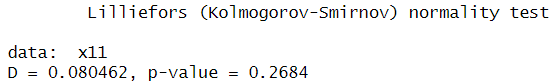
Если значение тестовой статистики не превышает никакое критическое значение (в нашем случае *result.critical\_value[2] -* значения при уровне значимости, равным 0.05), то результат не является значимым, и мы не отвергаем нулевую гипотезу.

****

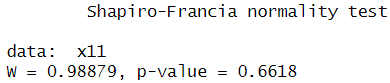
****

****

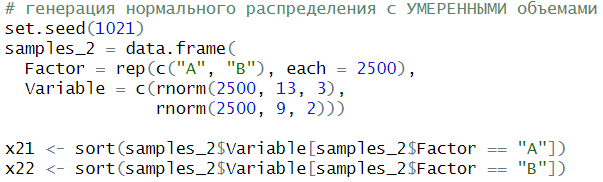
****

****

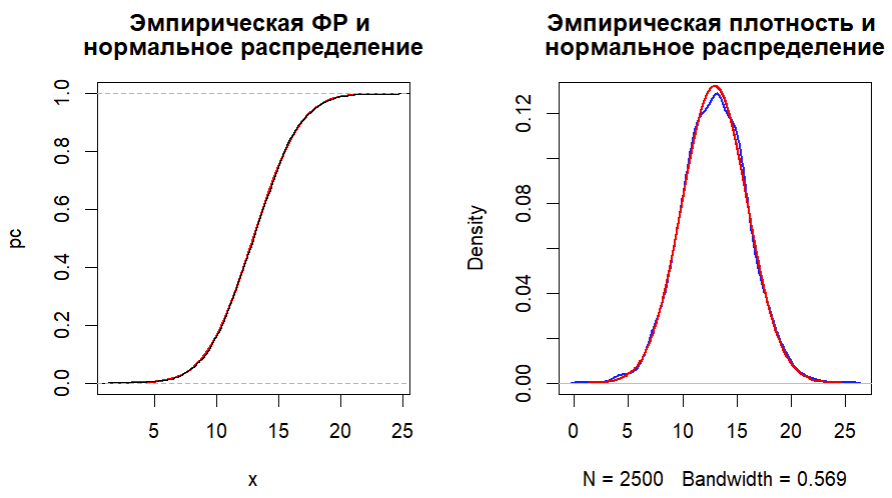
****

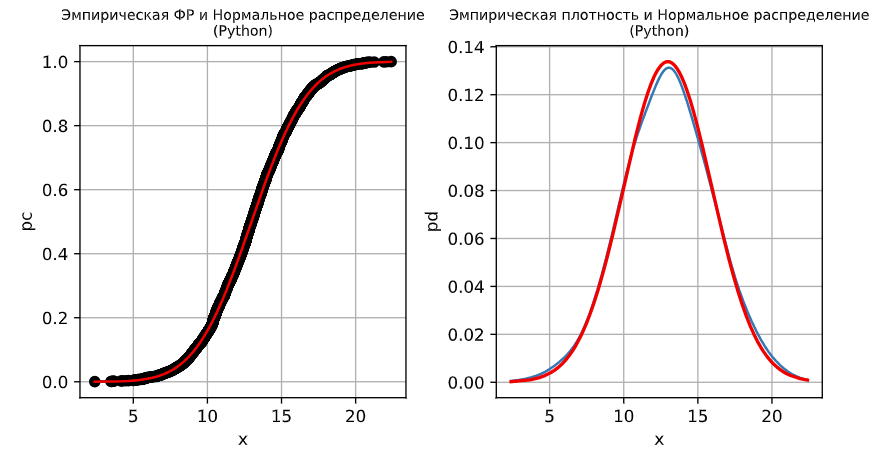
****

Я исследовал вторую выборку с малым объемом с параметрами (9, 2), построил графики и проверил гипотезы, но позвольте мне не предоставлять отчет по исследованию этой выборки, а сразу перейти к выборке с умеренным объемом, равным 2500, с теми же параметрами, что и выборка x11. Проделаем те же действия и посмотрим как увеличение объема выборки повлияет на результат.

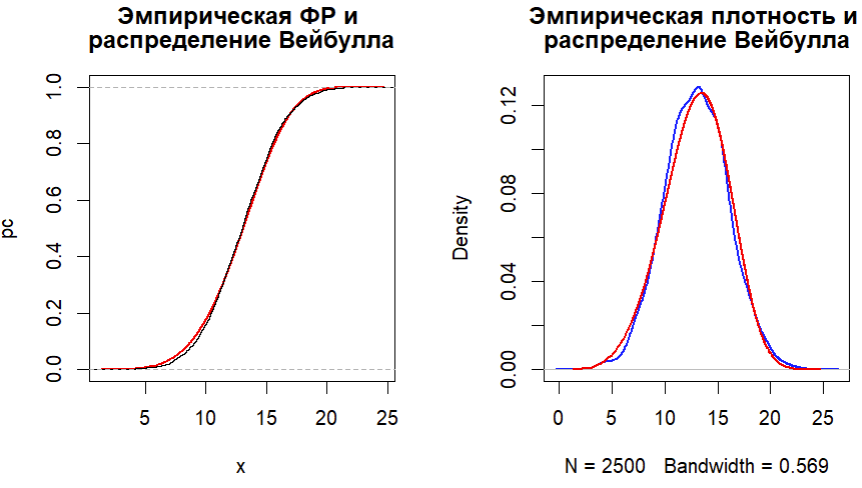


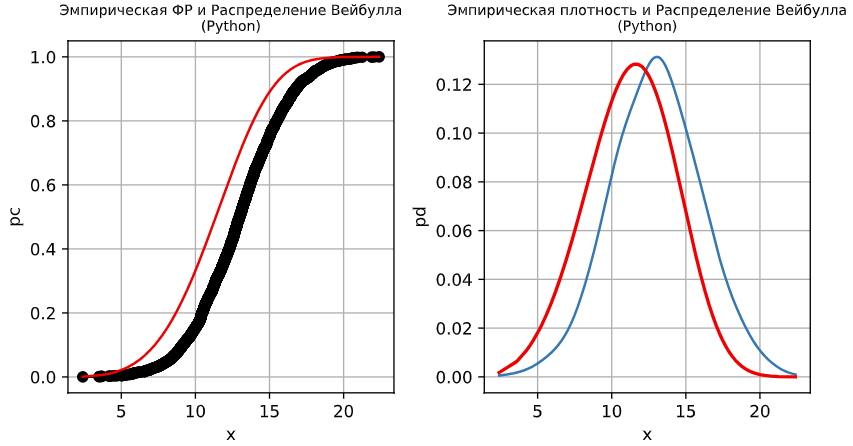


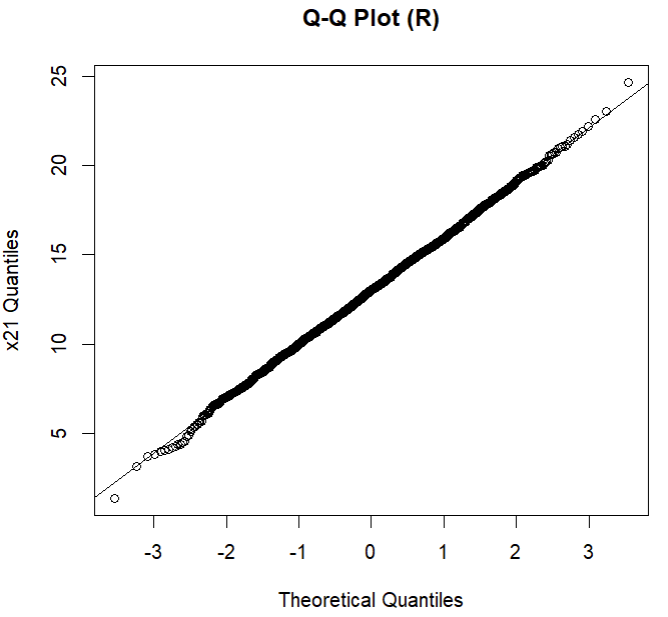


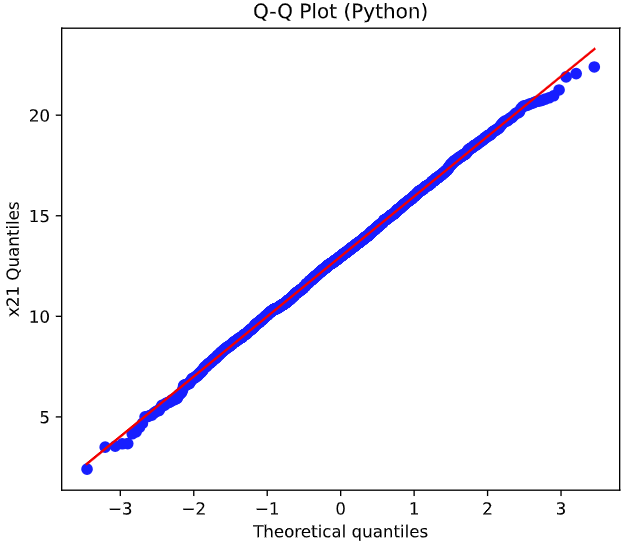


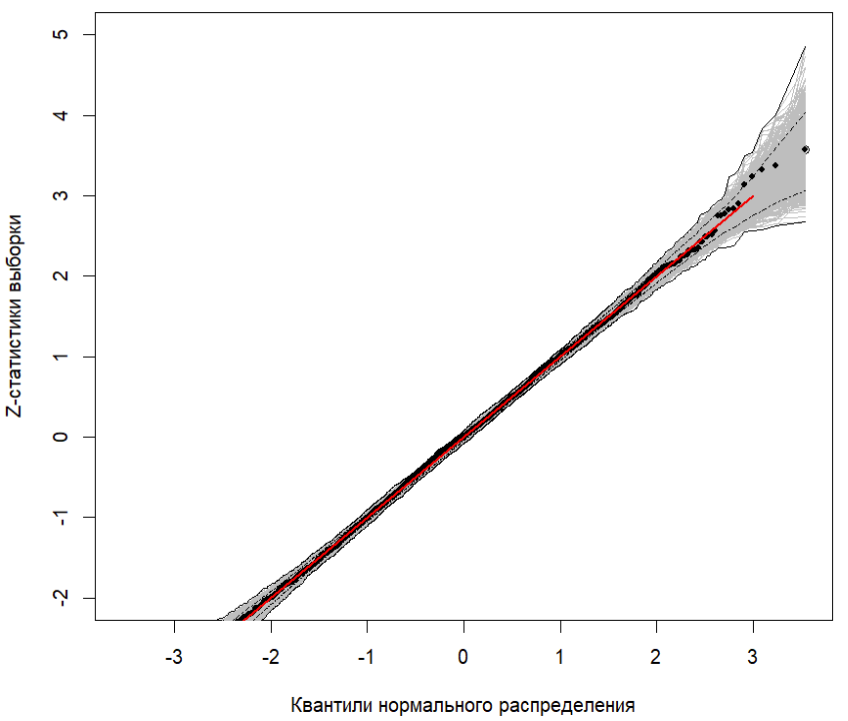
Видим, что при умеренном объеме выборки эмирическая функция распределения приблизилась к теоретической и практически совпадает с ней.







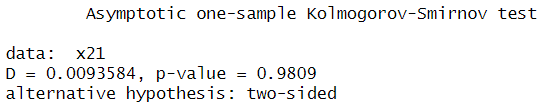




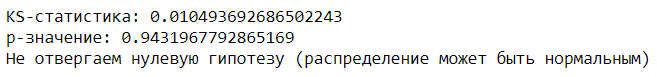
Видим, что при умеренном объеме выборки гораздо меньше точек лежат вне линии, представляющей теоретические квантили. Почти все точки лежат на этой линии.

Проверим **гипотезы о нормальности**.

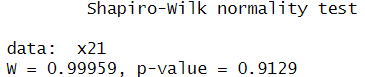
**R:**



**Python:**



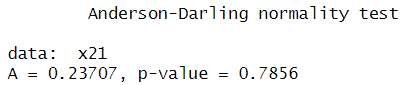
**R:**



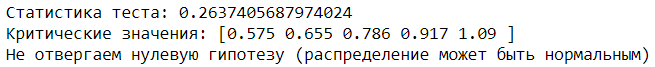
**Python:**

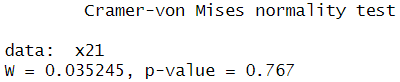


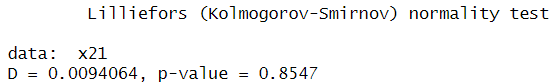
**R:**

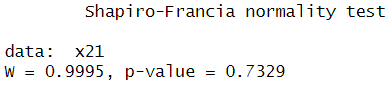


**Python:**

****

****

****

****

**Можно сделать вывод о том, что при увеличении размера выборки, распределение близится к теоретическому.**

**6. Продемонстрировать пример анализа данных с помощью графиков квантилей, метода огибающих, а также стандартных процедур проверки гипотез о нормальности. Рассмотреть выборки малого и умеренного объемов.**

Исследуем распределение величины пробега у автомобилей. Рассмотрим выборку малого объема. Я удалил из выборки все дубликаты, потому что тест Колмогорова-Смирнова проводится для данных без повторяющихся значений.

**R:**



**Python:**

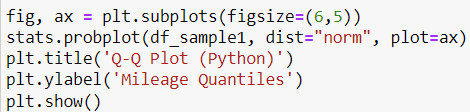
****

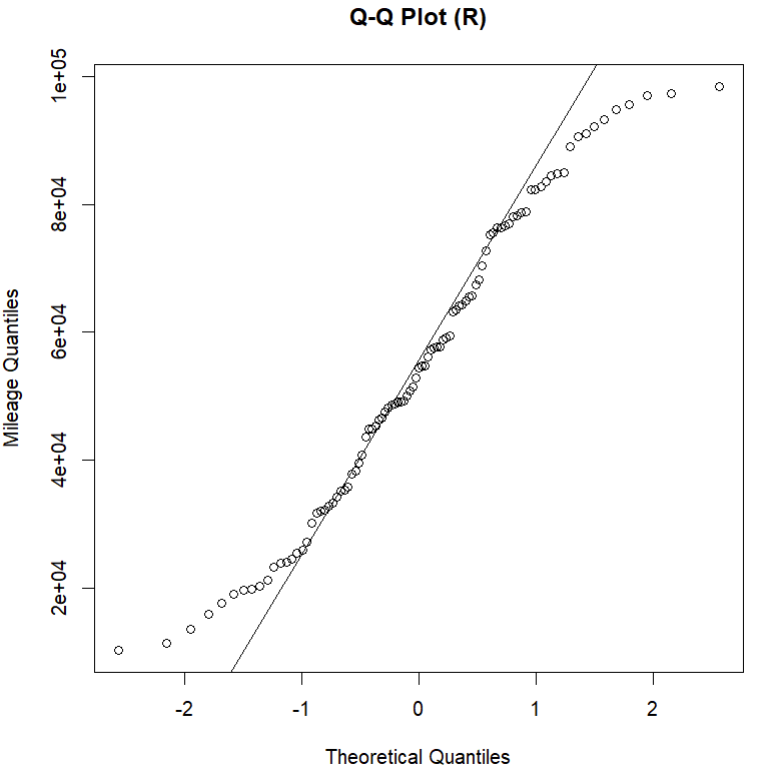
Проанализируем с помощью **графиков квантилей**:

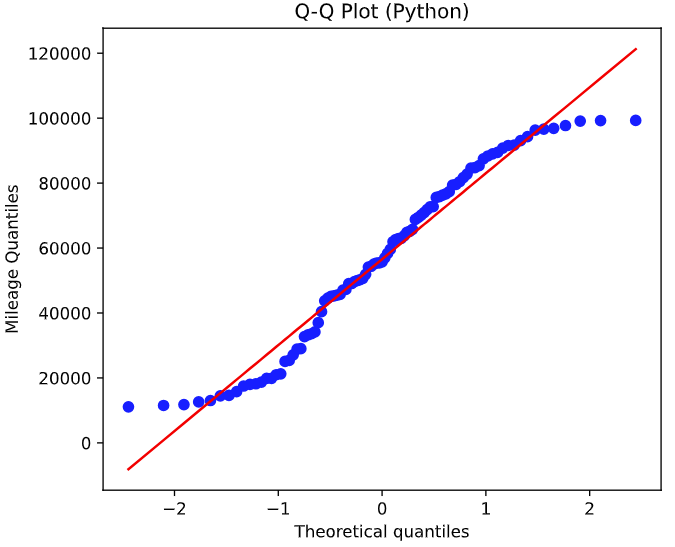
**R:**

****

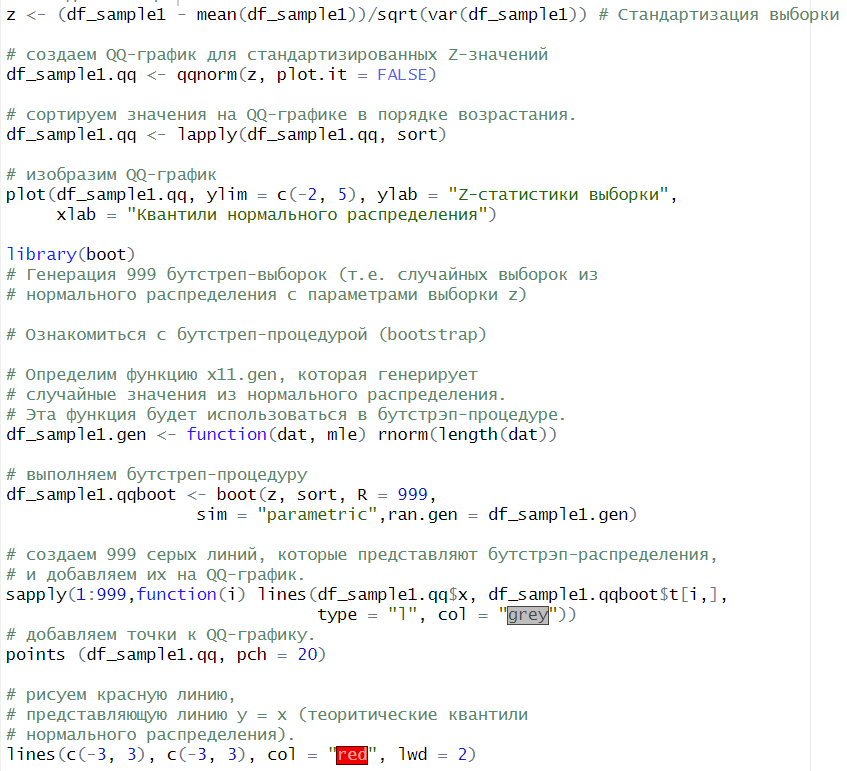
**Python:**

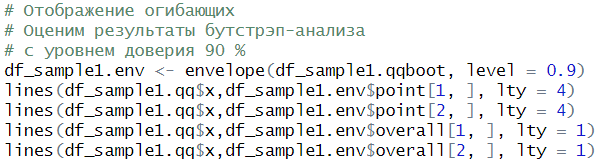


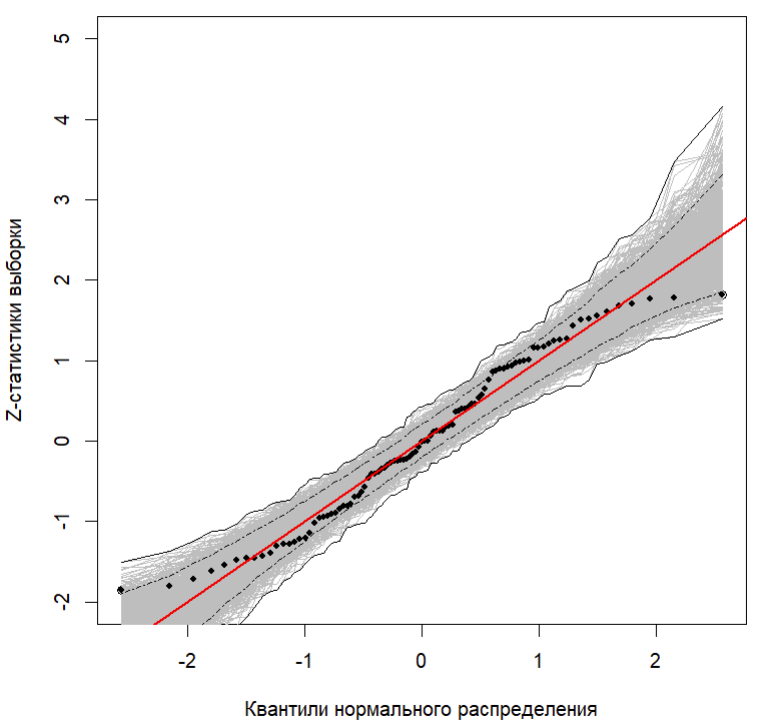




Используем **метод огибающих:**

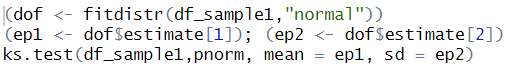
****

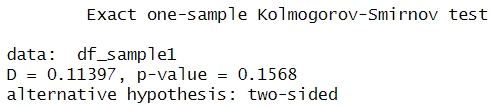
****

****

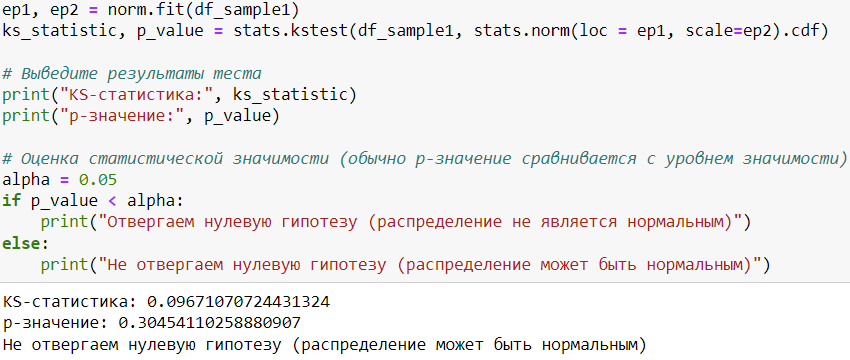
Проверим **гипотезы о нормальности:**

**R:**





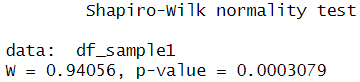
**Python:**

****

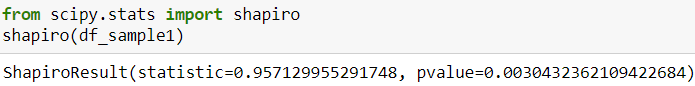
Видим, что в тесте **Колмогорова-Смирнова** p-value > 0.05, то есть у нас нет достаточных доказательств того, что распределение не является нормальным.

**R:**





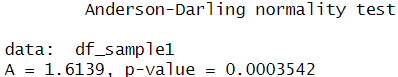
**Python:**

****

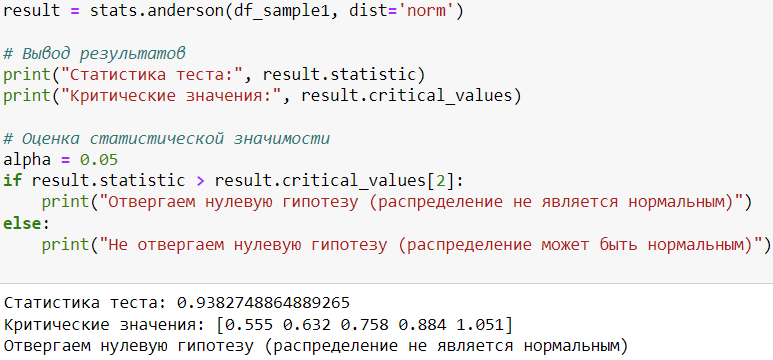
Видим, что в тесте **Шапиро-Уилка** p-value < 0.05, поэтому мы заключаем, что распределение не является нормальным.

**R:**





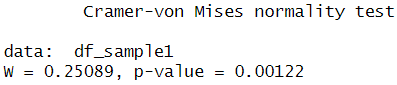
**Python:**

****

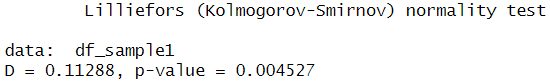
Видим, что в тесте **Андерсона-Дарлинга** значение тестовой статистики больше критического значения при уровне значимости, равным 0.05, и поэтому заключаем, что распределение не является нормальным.

**R:**

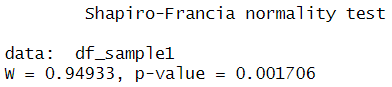
****

****

****

****

****

****

По итогам критериев **Крамера фон Мизеса, Колмогорова-Смирнова в модификации Лиллиефорса и Шапиро-Франсия**, заключаем, что распределение не является нормальным.

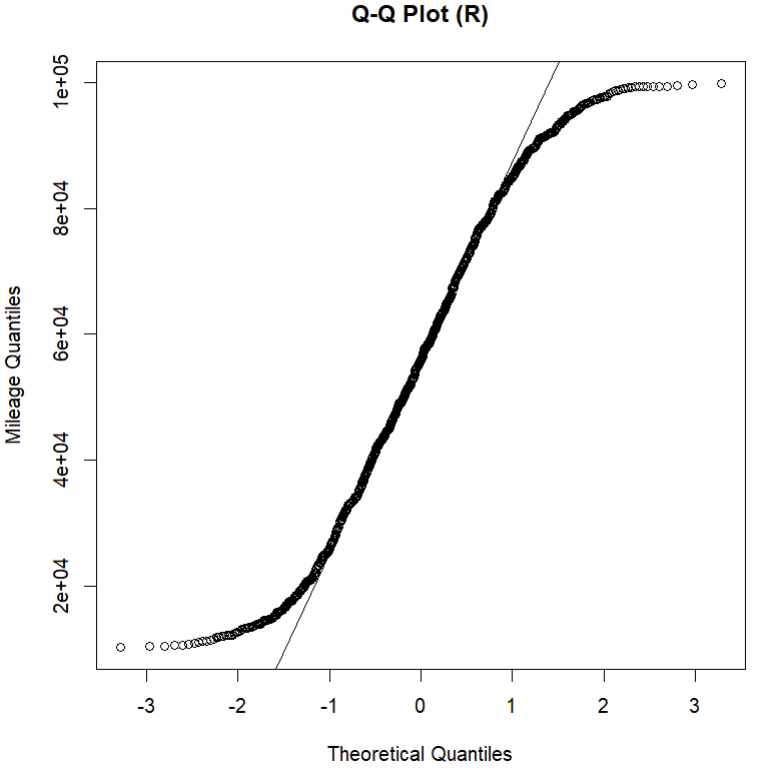
**Теперь исследуем выборку умеренного объема и проделаем те же действия.**

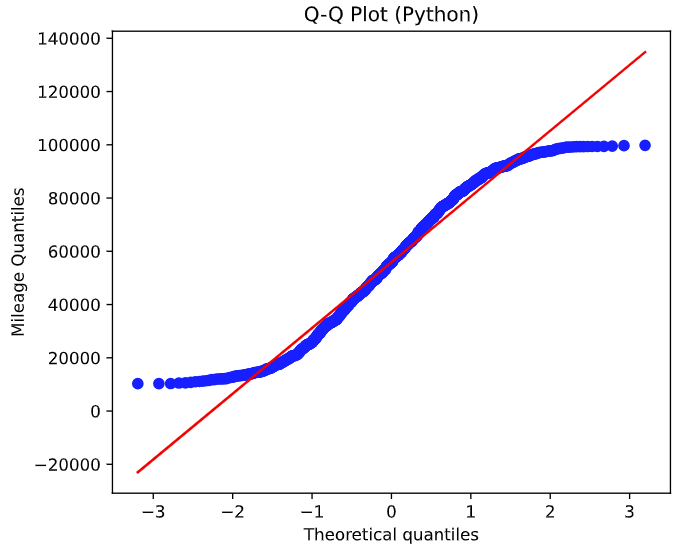
**R:**

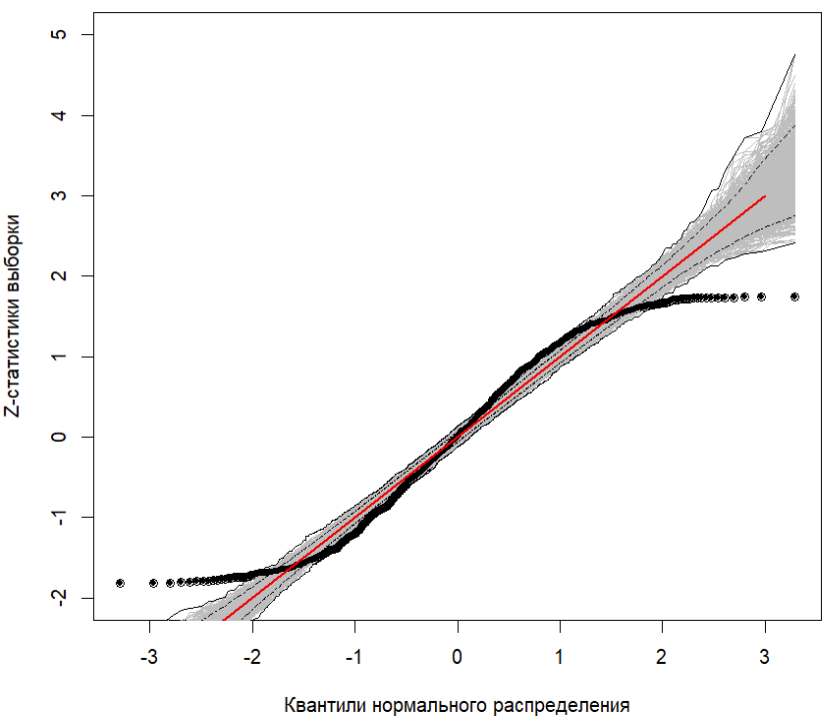


**Python:**

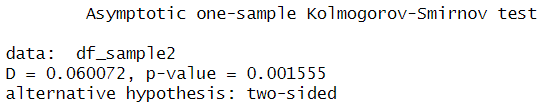
****

****

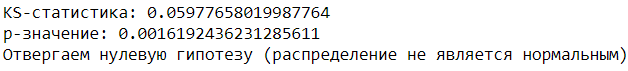


****

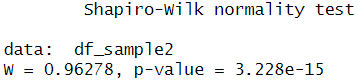
**R:**

****

**Python:**



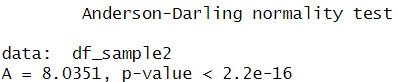
**R:**



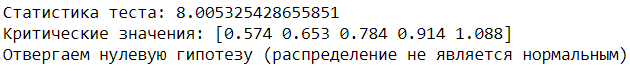
**Python:**

****

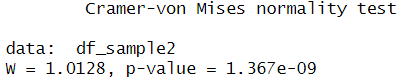
**R:**

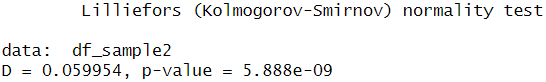
****

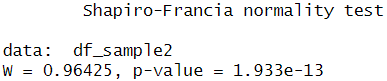
**Python:**

****

**R:**

****

****

****

При умеренном объеме выборки, все критерии показали, что **распределение не является нормальным.**