Отчет о полученных результатах по проведенной семестровой работе

Данные, используемые для исследования, были получены с ресурса Kaggle. Их можно найти по ссылке - *https://www.kaggle.com/datasets/kapturovalexander/maang-share-prices-till-february-2024*

Наши данные представляют цены на акции технологических гиганта Alphabet Inc.

Рассмотрим столбцы наших данных и разберем, что они представляют из себя:

Date (Дата): Этот столбец содержит дату торговли акциями компании.

Open (Открытие): Это цена акций компании в начале торгового дня.

Close (Закрытие): Это цена акций компании в конце торгового дня.

High (Максимум): Это наивысшая цена, достигнутая акциями компании в течение торгового дня.

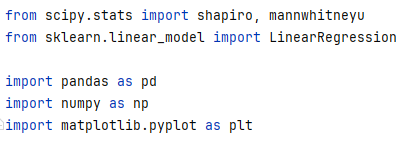
Low (Минимум): Это наименьшая цена, достигнутая акциями компании в течение торгового дня.

Adj Close (Скорректированное закрытие): Это цена закрытия акций компании, скорректированная на дробление или объединение акций, выплату дивидендов и другие корпоративные действия.

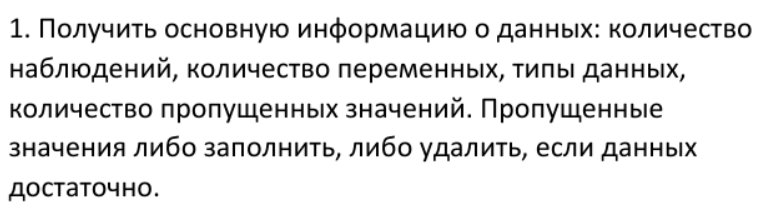
Volume (Объем): Этот столбец указывает на количество акций компании, которые были проданы или куплены в течение торгового дня.

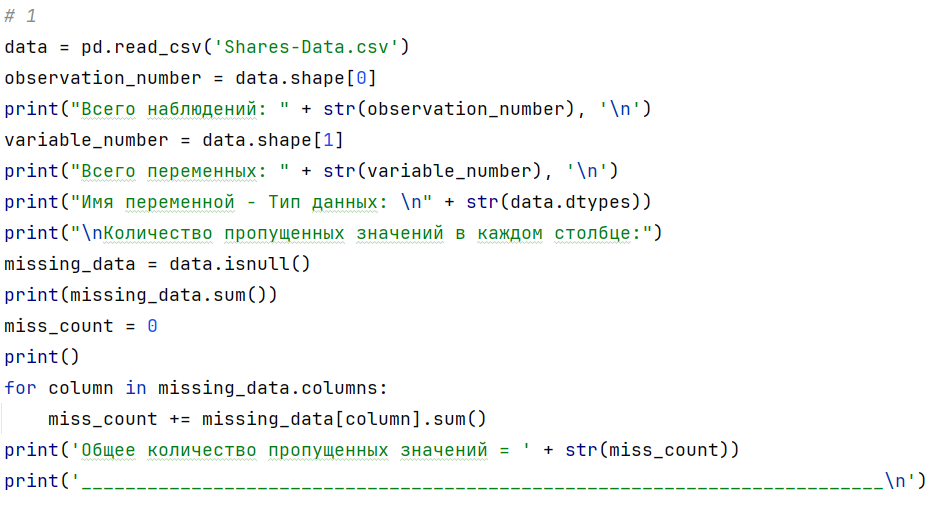
Теперь перейдем к непосредственной реализации нашей семестровой работы и коду:

В нашей работе мы будем использовать следующие базовые библиотеки python:



Пункт первый





Представленный код выполнит нужную работу. Как результат мы получим следующие данные.

Всего наблюдений: 4818

Всего переменных: 7

Имя переменной - Тип данных:

Date object

Open float64

High float64

Low float64

Close float64

Adj Close float64

Volume int64

dtype: object

Количество пропущенных значений в каждом столбце:

Date 0

Open 0

High 0

Low 0

Close 0

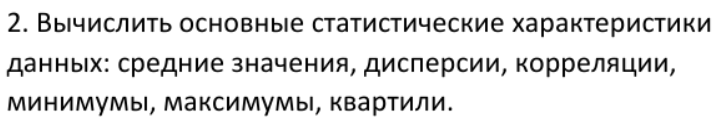
Adj Close 0

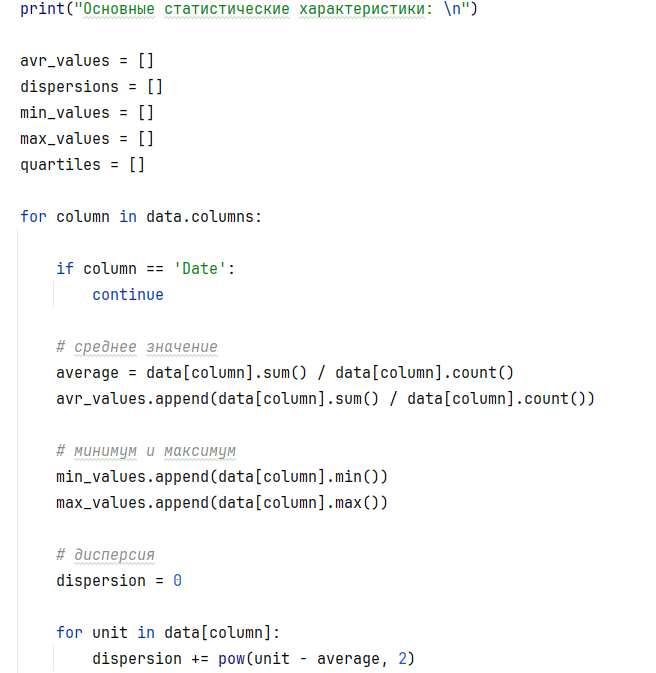
Volume 0

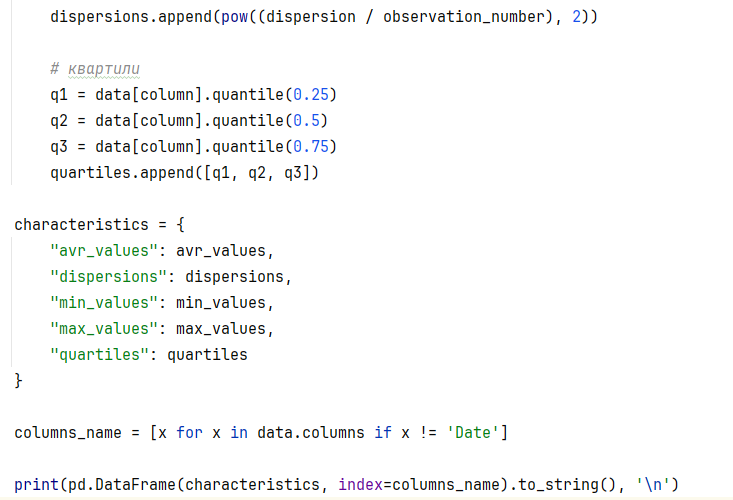
dtype: int64

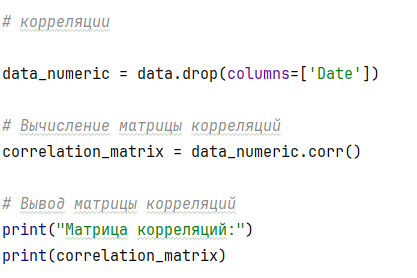
Общее количество пропущенных значений = 0

Второй пункт

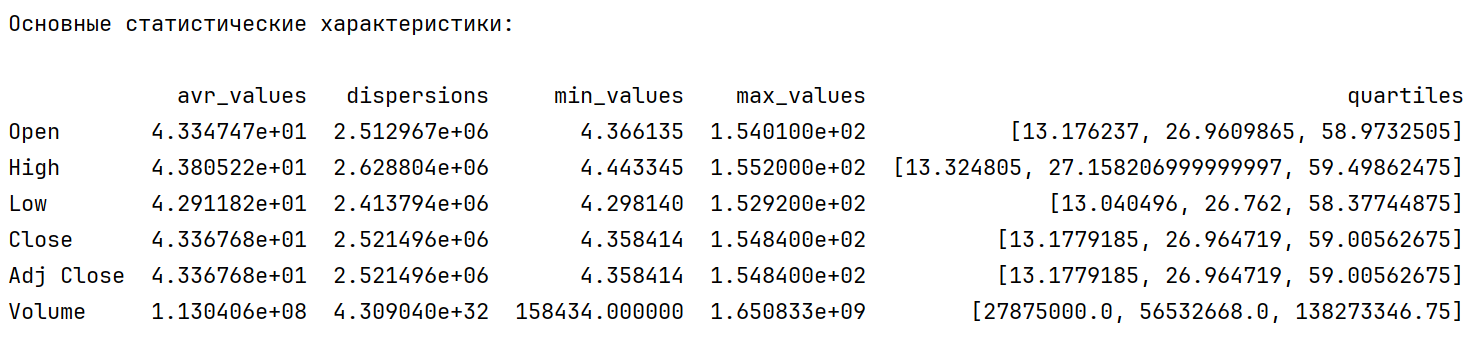




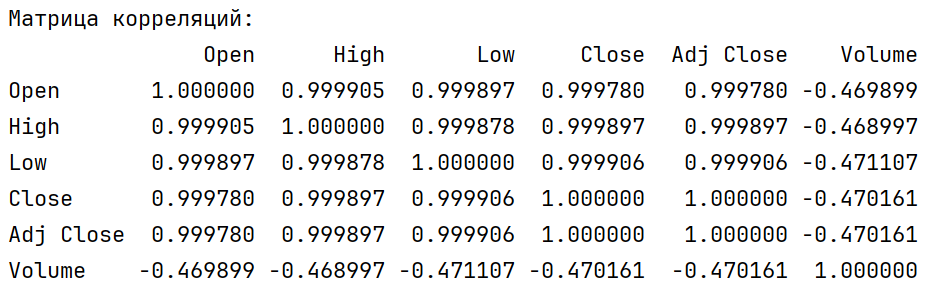




Данный код любезно конструирует для нас два DataFrame или проще сказать матрицы, благодаря котором мы можем удобно рассмотреть основные средние значения всех переменных на первом DataFrame.

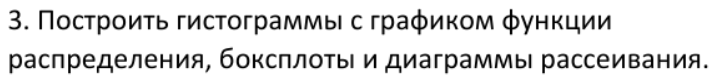


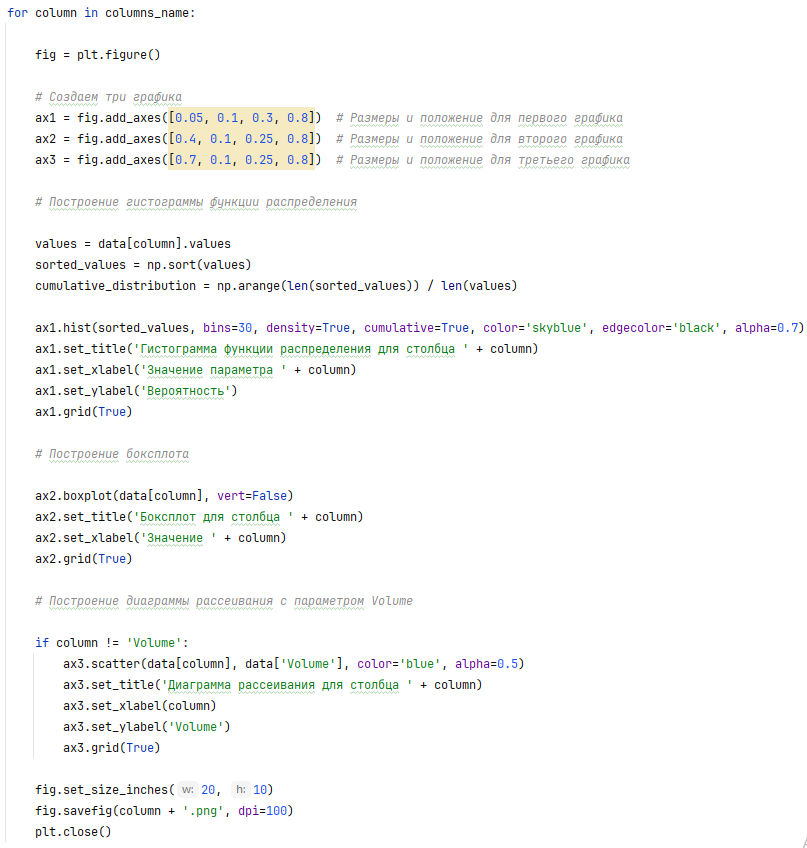
Отдельный DataFrame я сделал для корреляций. Это сделано для удобство представления данных, так ка в рамках биржевой торговли может быть важна корреляция каждой пары данных из возможных.



Стоит отметить, что в данной аналитике и далее мы не используем столбец Date, так как это не имеет смысла, этот столбец нужен только для указания даты для подсчитанных основных характеристик.

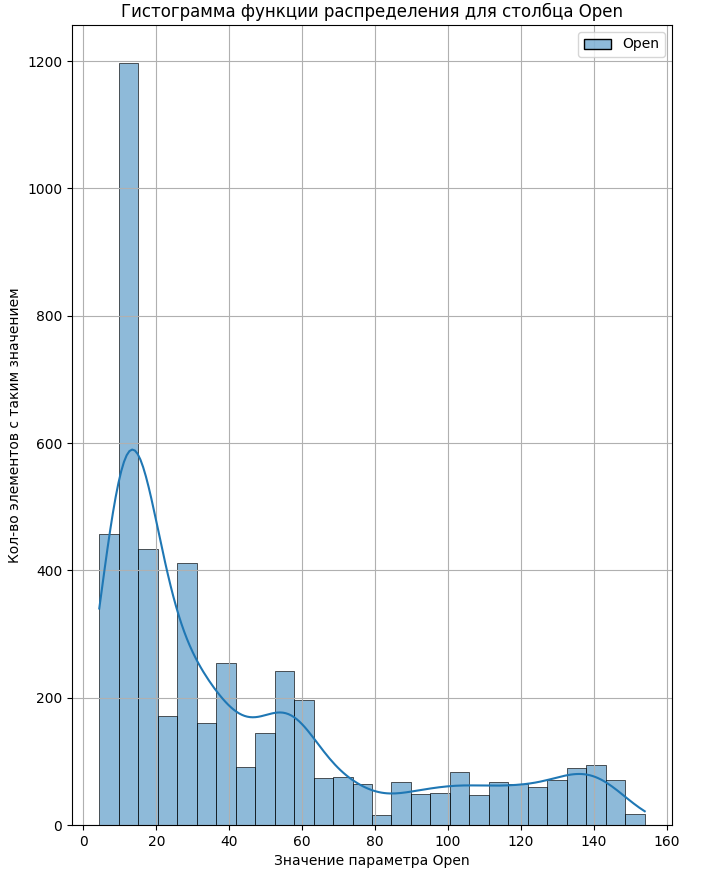
Третий пункт

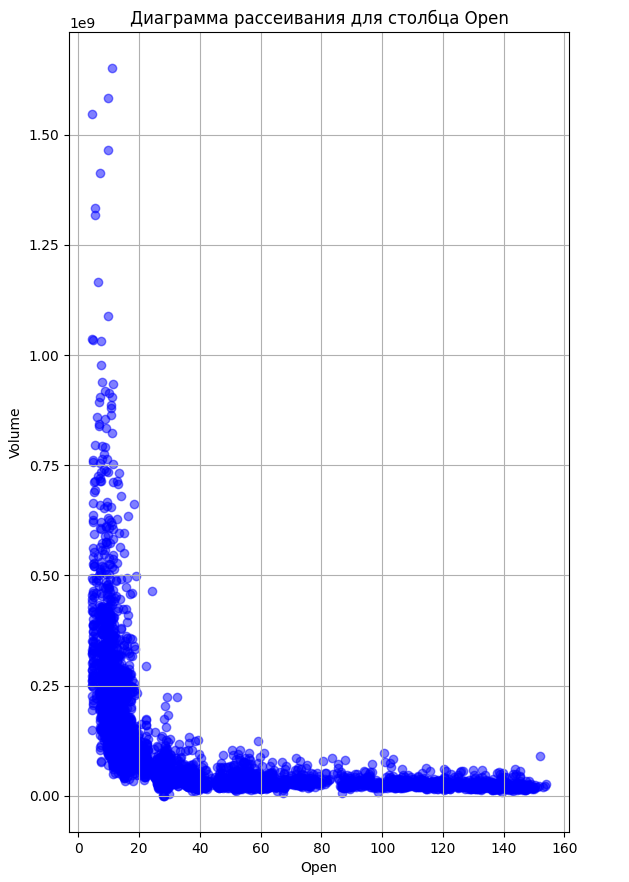
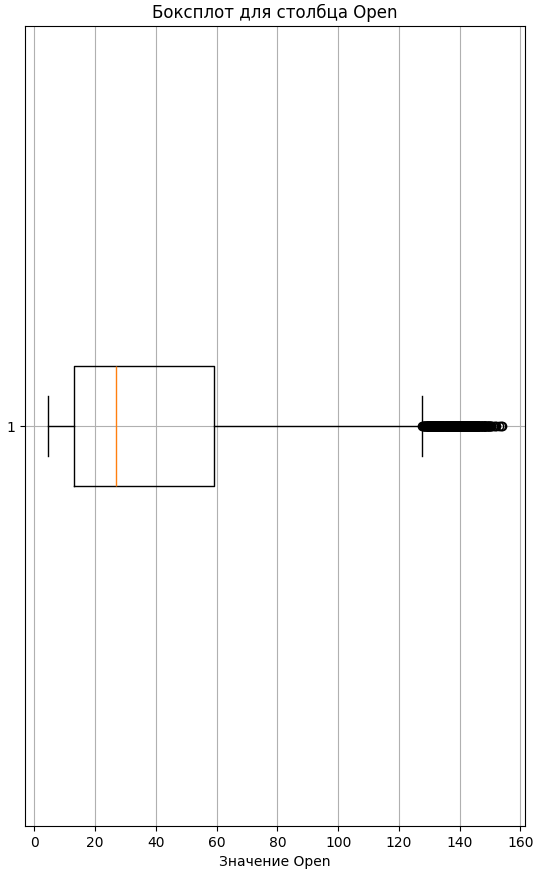




Данный код сохранит 6 фотографий в корневой директории проекта, каждая из которых содержит в себе три графика: гистограмма с графиком функции распределения, боксплоты и диаграммы рассеивания. Для примера приведем все 3 графика для столбца Open.

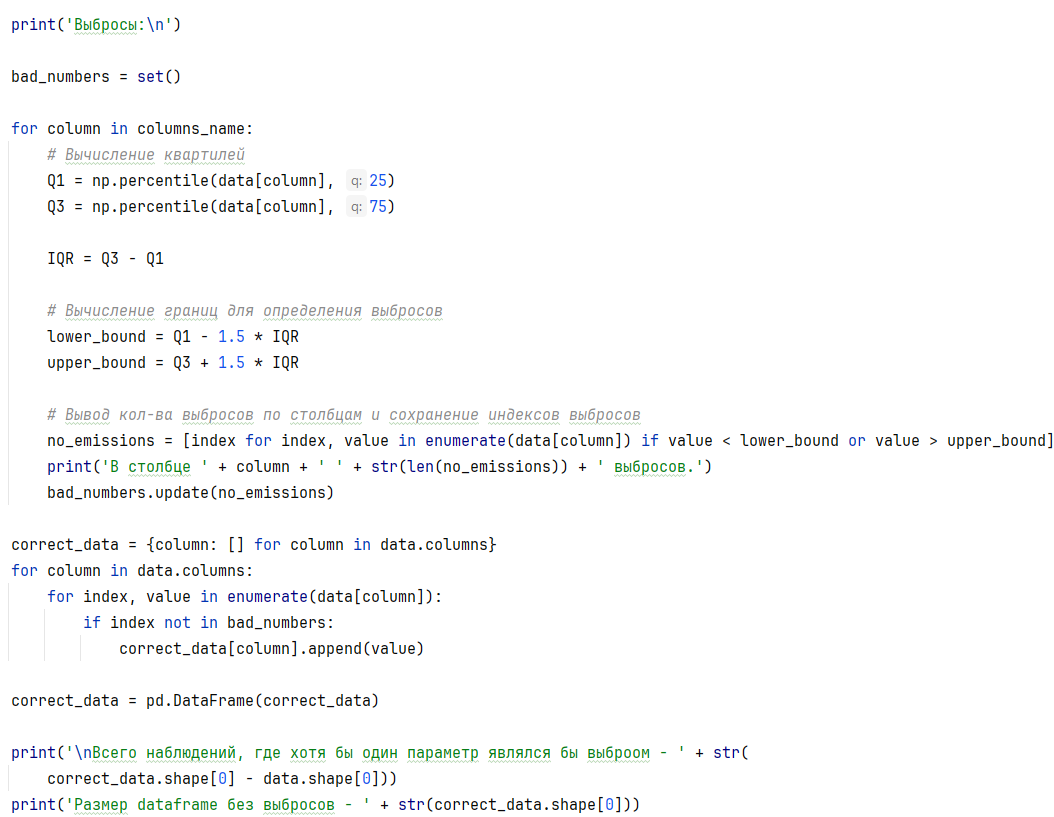
Все остальные графики также будут прикреплены к этому документы в виде png фотографий.





Пункт четвертый





Данный код покажет нам основные характеристика выбросов, а именно:

В столбце Open 340 выбросов.

В столбце High 346 выбросов.

В столбце Low 341 выбросов.

В столбце Close 338 выбросов.

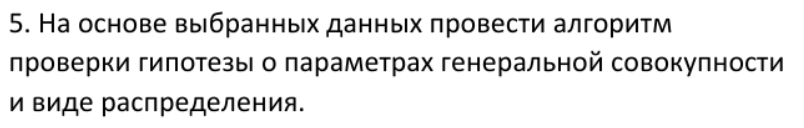
В столбце Adj Close 338 выбросов.

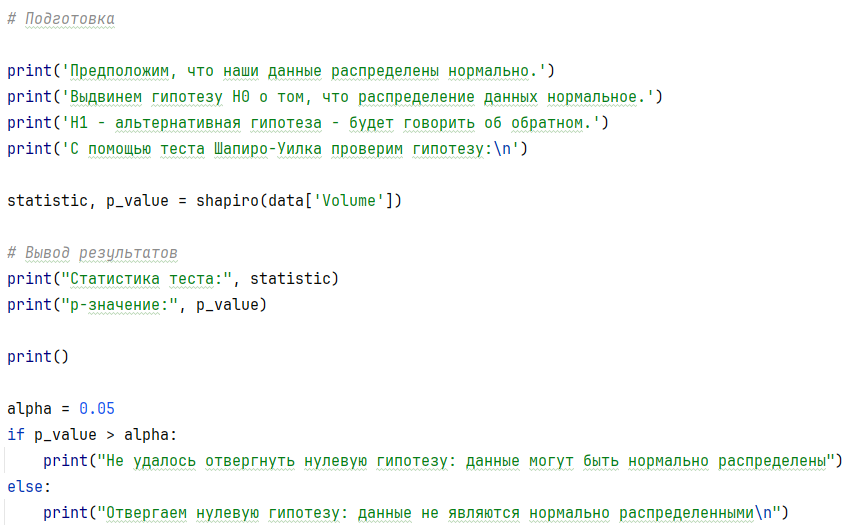
В столбце Volume 400 выбросов.

Всего наблюдений, где хотя бы один параметр являлся бы выброом - -749

Размер dataframe без выбросов – 4069

Пятый пункт



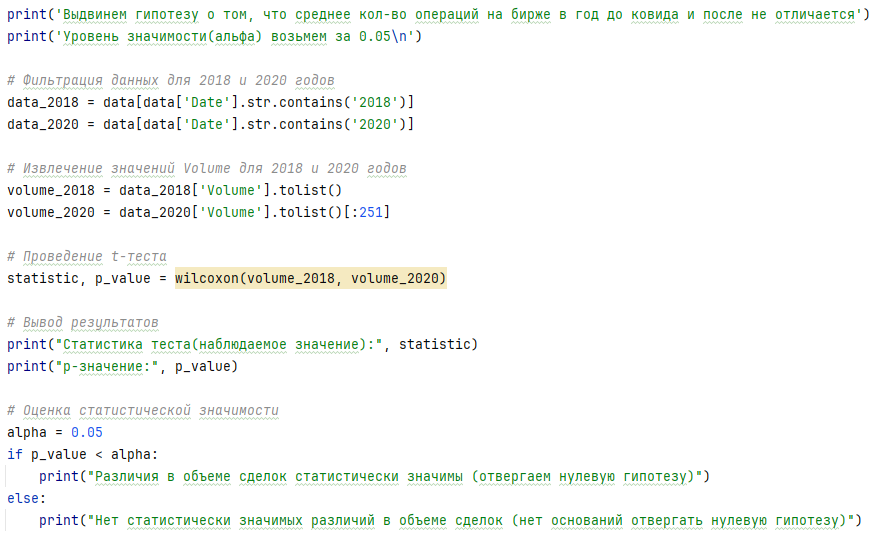
  
Для начала предположим, что наши данные распределены нормально и выдвинем гипотезу H-0 о том, что данные распределение данных нормальное. Проверим эту гипотезу с помощью теста Шапиро-Уилка и получим следующие данные:

Статистика теста: 0.6503599140057313

p-значение: 9.10845108899811e-72

Отвергаем нулевую гипотезу: данные не являются нормально распределенными

Теперь можем и переходить к основной гипотезе:



Выдвинем гипотезу о том, что среднее кол-во операций на бирже в год до ковида и после не отличается. Так как вспышка ковида началась в Китае, в 2019, возьмем 2018 год в качестве года до ковида и 2020 год в качестве года во время/после ковида. В этот период (ковидный) многие вынуждены были уйти на удаленку, бизнес потерпел изменения, и, конечно, эта ситуация коснулась и рынка ценных бумаг.

Теперь нужно определится с тестом. Мы будем проводить исследование на основании теста Уилкоксона.

Как мы уже выяснили, наши данные распределены не нормально, поэтому мы будем использовать непараметрический тест. Также стоит заметить, что мы будем исследовать две независимые выборки, однако акции будем учитывать только одной компании, но в разных условиях. Поэтому мы и выбираем критерий Уилкоксона.

Итак, наш код вернул нам необходимые данные о результатах проведения теста:

Выдвинем гипотезу о том, что среднее кол-во операций на бирже в год до ковида и после не отличается

Уровень значимости(альфа) возьмем за 0.05

Статистика теста (наблюдаемое значение): 12179.5

p-значение: 0.0016005773207351751

Различия в объеме сделок статистически значимы (отвергаем нулевую гипотезу)

На основании полученных результатов делаем вывод о том, что Ковид действительно повлиял на компанию Alphabet Inc.

Шестой пункт





Построим две линейные регрессии. Начнем с первой

Проведем построение линейной регресси по переменной Open как независимой переменной и Close как зависимой

Intercept: -0.006964719522152052

Coefficient: [1.00062685]

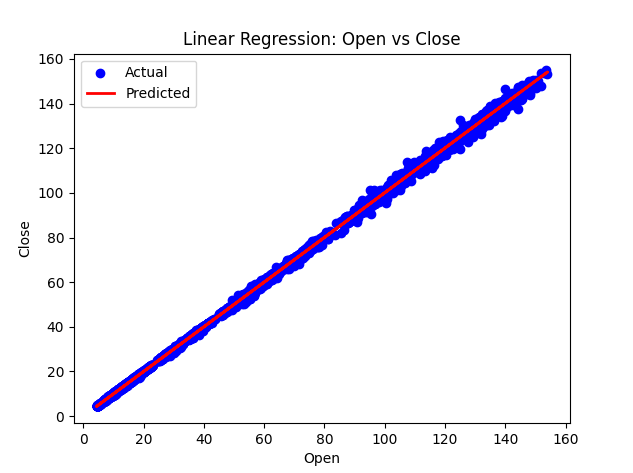
Это означает, что у нас есть следующая линейная зависимость: Close = [1.00062685] \* Open + -0.006964719522152052

Таким образом, каждый дополнительный единичный прирост в переменной Open приведет к увеличению переменной Close на примерно 1.00062685 единиц, при условии, что все остальные факторы остаются константными

Коээфициент корреляции - 0.999779670516443

С точки зрения связи мы получаем сильную положительную связь между группами

Также представим график, который отрисует наш код



Вторая регрессия



Данный код выдаст нам следующие результаты:

Проведем построение линейной регресси по разнице между наибольшей и наименьшой ценой продажи акции как независимой переменной и переменной Volume как зависимой

Intercept: 149594787.03671998

Coefficient: [-40915630.44188645]

Это означает, что у нас есть следующая линейная зависимость: Volume = [-40915630.44188645] \* Difference + 149594787.03671998

Таким образом, каждый дополнительный единичный прирост в переменной Difference приведет к увеличению переменной Volume на примерно -40915630.44188645 единиц, при условии, что все остальные факторы остаются константными

Коээфициент корреляции - 0.8060339735827999

С точки зрения связи мы получаем сильную положительную связь между группами

Также представим график и для этой линейной регрессии

