Университет ИТМО

# Практическая работа №2-3

по дисциплине «Визуализация и моделирование»

**Автор:** Кумпан Виктор Викторович
**Поток:** 1.2
**Группа:** K3241
**Факультет:** ИКТ
**Преподаватель:** Чернышева А.В.

Санкт-Петербург, 2021 г.

## Аннотация

В данной работе начнем с формулировки 5 гипотез. Далее рассмотрим предобработку датасета: обработка пустых ячеек, нормализация дат, нормализация типов данных, балансирование значений ячеек. И попытаемся подтвердить или опровергнуть эти гипотезы при помощи анализа построенных графиков.

# 1 Сформулируем 5 гипотез по датасету

- Датасет взять за период с 22.02.2018-22.02.2020 именно в этот промежуток попадает пандемия (01.01.2020-01.09.2020) и можно выдвинуть гипотезу, что акции компаний значительно упадут в этот промежуток.
- Так как акции разных секторов экономики США присутствуют в датасете (IT, Oil, Fashion), гипотетически IT сектор более легко переживет всемирный кризис и акции компаний этого сектора начнут подниматься большими темпами.
- Акции Fashion сектора будут активно подниматься начиная с (01.06.2020 - 01.11.2020) в связи с всемирным трендом ЗОЖ и пандемия заставила людей задуматься о своем здоровье
- Американские корпорации в сегменте Oil гипотетически не смогут так быстро реабилитироваться после мирового кризиса, как компании других секторов, это связано с их низким уровнем приспособленности к быстрым изменениям на финансовом рынке.
- Гипотетически в послековидное время, капитализации компаний IT сектора будет очень близким к компаниям нефтегазового сектора.

# 2 Предобработка датасета

Первым делом проверим, есть ли пустые ячейки в датасете. Как оказалось их нет, были замечены только пропущенные дни, это связано с тем, что фондовый рынок не работает в праздничные и в выходные дни. Но это никак не сказалось, на качестве датасета.

```
[('Dell', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0}),
 ('Chevron', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0}),
 ('ExxonMobil', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0}),
 ('IBM', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0}),
 ('Nike', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0}),
 ('PVH', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0}),
 ('Apple', {'Дата': 0, 'Цена': 0, 'Откр.': 0, 'Макс.': 0, 'Мин.': 0, 'Объём': 0, 'Изм. %': 0})]
```

Рис. 1: Количество пустых ячеек

Следующая проблема с которой я столкнулся это некорректная структура даты **20.02.2021** . Pandas не смог преобразовать ее к типу datatime, необходимо было изменить порядок даты, после этого он скушал такую структуру **2021.02.20**.

Несмотря на то, что пустых ячеек в датасете не было, были ячейки в которых поставлен **'-'** и они принадлежали только одному столбцу **Объем** в данном случае мы можем считать, что это мусор. Но он мешает суммировать столбцы, поэтому нам нужно было с этим бороться. Решение было принять его за 0, исходя из логики столбца в котором встречался этот символ.

| Дата | Цена | Откр. | Макс. | Мин. | Объём | Изм. % |
|---|---|---|---|---|---|---|
| **14.11.2020** | 66,11 | 66,11 | 66,11 | 66,11 | - | 0,00 % |

Рис. 2: Пример встречаемой ошибки с тире

Более очевидная задача была балансирование ячеек в связи с необходимость преобразовать их к определенным типам. Здесь я столкнулся с тем, что в столбцах **'Цена','Откр.','Макс.','Мин.'** вместо положенной **'.'** стоит **','** это делает невозможным привести к численному типу в Pandas. Также в столбцах **'Объем','Изм.'** стоят **('M','K'),'%'** соответственно, их тоже необходимо удалить, для привидения к числовому типу.

| Дата | Цена | Откр. | Макс. | Мин. | Объём | Изм. % |
|---|---|---|---|---|---|---|
| **22.02.2021** | 127,53 | 127,94 | 129,67 | 125,62 | 65,81M | -1,80 % |

Рис. 3: Пример необработанного датасета

# 3 Визуализация данных и проверка гипотез

В данном разделе мы рассмотрим некоторые графики, с реализацией и полным их списком вы можете ознакомиться на Google colab , и проверим выдвинутые гипотезы.

Рассмотрим график стоимости акции с 22.02.2018 - 22.02.2021 акций компаний разных секторов, и можем заметить, что резкий обвал приходится на начало 2020 г. и продолжается до 04.2020, из этого следует, что компании всех секторов подверглись влиянию коронокризиса, что отражает одну из наших гипотез.

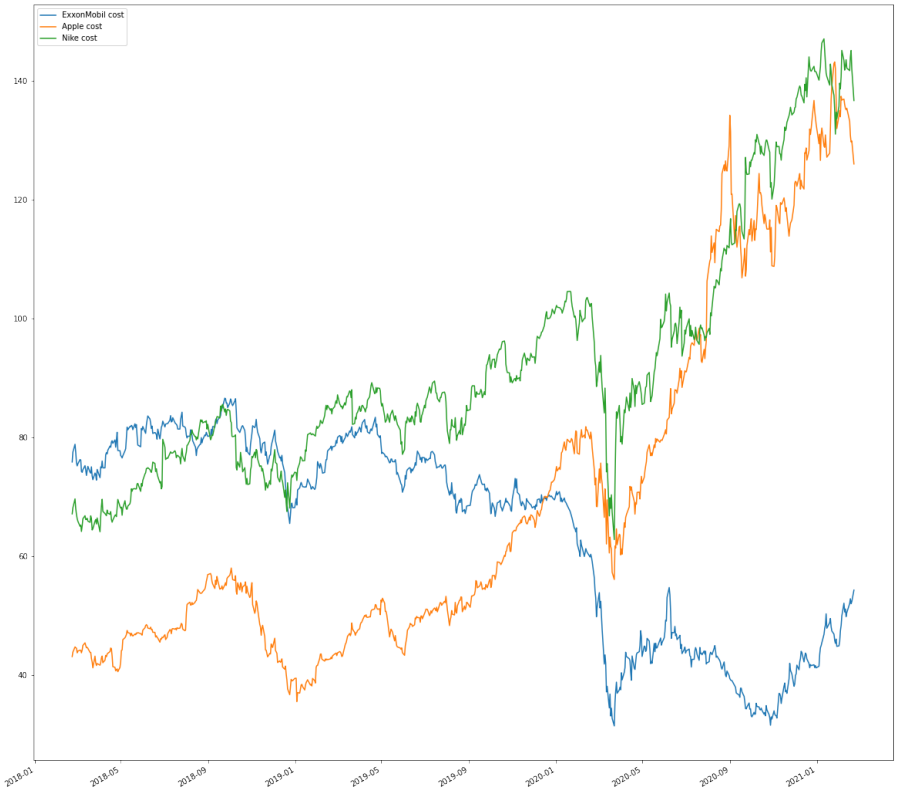


Рис. 4: Стоимости акций Apple, Nike, Exxon 22.02.2018 - 22.02.2021

Из предыдущего графика видим, что, действительно, акции Apple поднимаются после коронокризиса очень быстро, но все ли компании в данном секторе так стремительно выросли? Из графика ниже можем заметить, что акции IBM не так стремительно растут, поэтому однозначно ответить тут нельзя. В данном случае надо рассматривать компании в отдельности. Но даже при таком графике, мы можем убедиться, что IT сектор более легко пережил коронокризис, что подтверждает нашу 2-ю гипотезу.

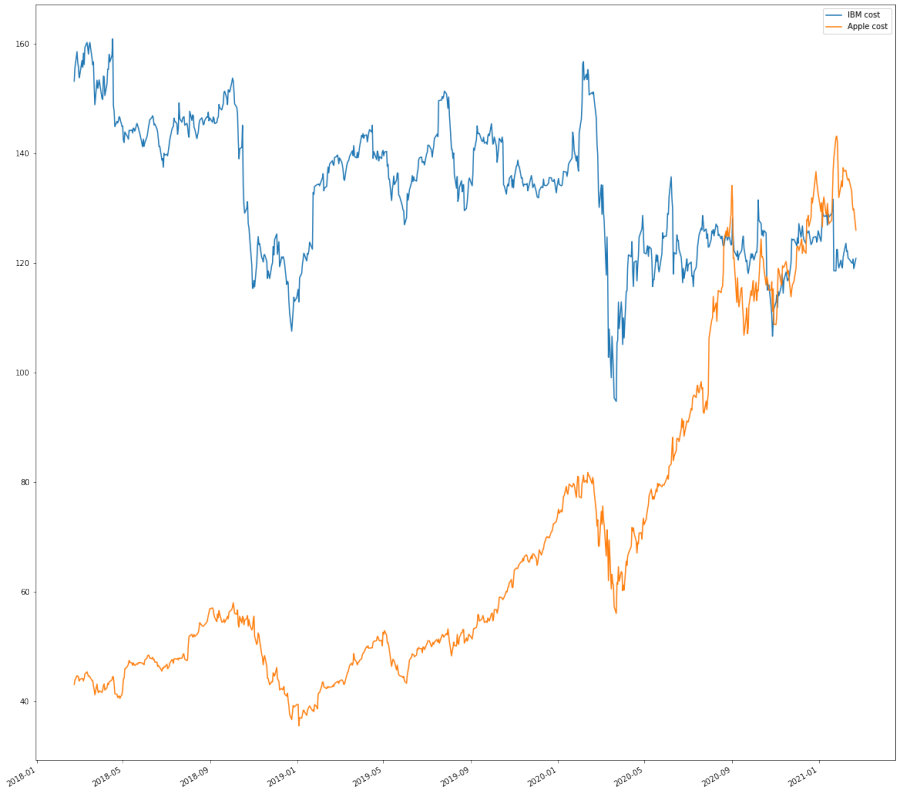


Рис. 5: Стоимости акций Apple, IBM 22.02.2018 - 22.02.2021

Рассмотри компанию Nike сектора Fashion и оценим влияния коронокризиса на стоимость акций. Из графика ниже, можем заметить, что данная компания растет на протяжении всего периода взятого для датасета. Но как это не странно, компания испытала резкий взлет в период коронокризиса, несмотря на то, что в этот период она достигла своего ценового минимума. Этому есть явная причина, люди во всем мире начали задумываться над своим здоровьем

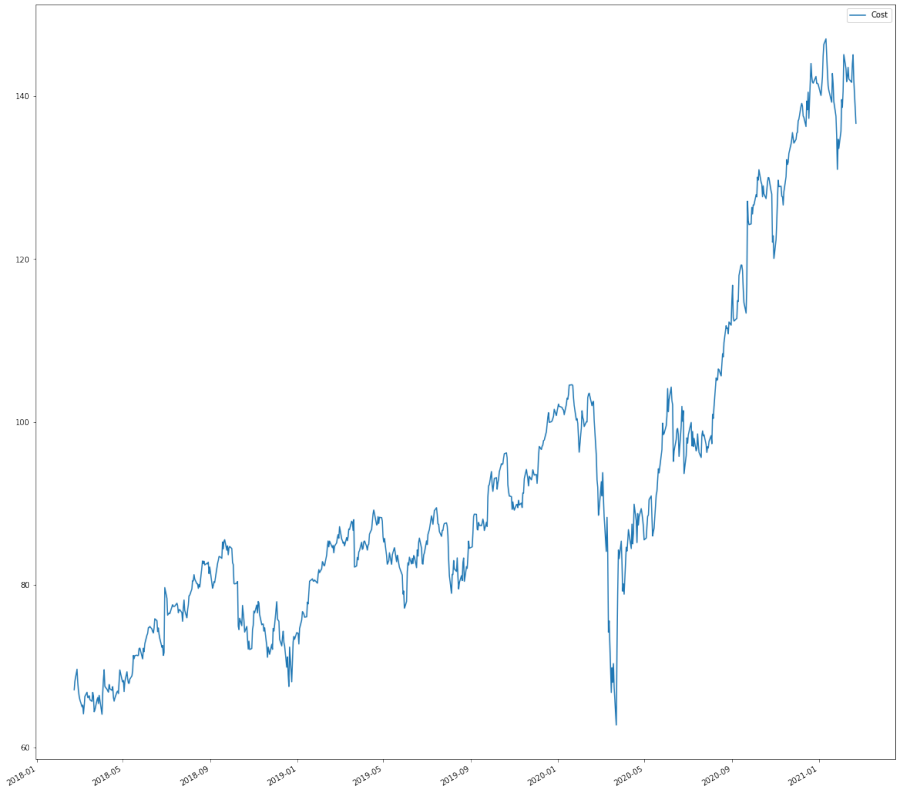


Рис. 6: Стоимости акций Nike 22.02.2018 - 22.02.2021

Рассмотрим нефтегазовый сектор, по графику 7 видим, что происходит падение цен на акции, причем в определенное время акции реагируют практически одинаково, за исключением некоторых ситуаций из этого можем сделать вывод, что они имеют общие факторы воздействия, одним из которых является договоры ОПЕК. И если мы хотим собирать диверсифицированный портфель, то нам нерелевантно брать все эти акции, достаточной одной из них. Также может заметить, что на протяжении всего периоды цены на акции испытывают падение. Если более детально проанализировать график в пероид коронокризиса, то темы выхода из него довольно проблематичные, что является следствием одной из наших гипотез.

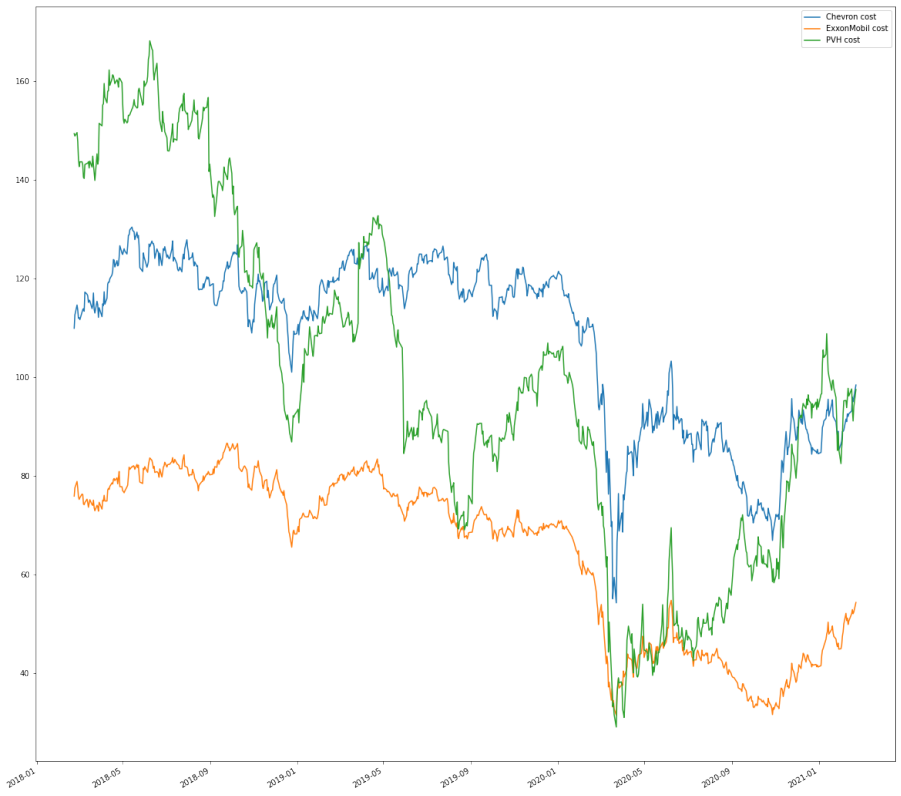


Рис. 7: Стоимости акций Apple, IBM 22.02.2018 - 22.02.2021

Проанализируем как коронокризис повлиял экономическое положение трех секторов (Fashion, OIL, IT). По диаграме 8 справа мы видим совокупность стоимости всех продавамых акций за период (01.01.2019-01.01.2020) явным лидером выступает нефтегазовый сектор с контрольной долей в 78% рынка, далее идет IT (18.4%) и Fashion(3.7%) сектор. На левой диаграмме все кардинально отличается,она приходится на период начала коронокризиса и выхода из него (01.01.2020-22.02.2021). Доля капитализации IT сектора выросла в 2.5 раза, Fashion в 1.5, а нефтегаз упал в 1.6 р. . Это свидетельствует о том, что именно в этот период рынок более благоприятный для этих секторов и они испытывают глобальный рост, даже по сравнению с фундаментальным сектором - нефтегазом.

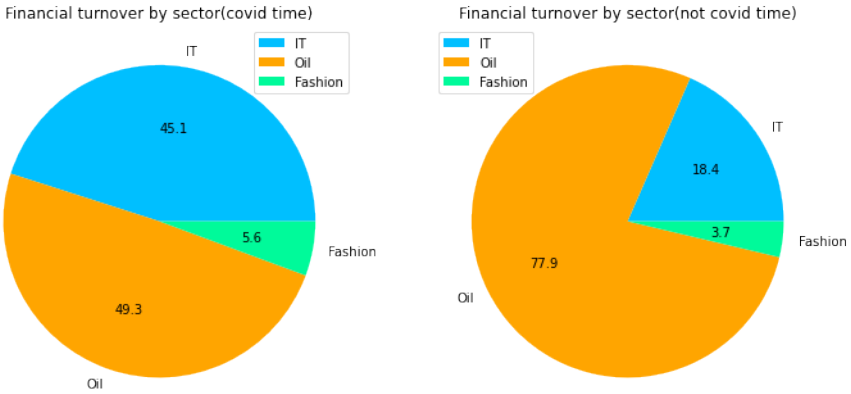


Рис. 8: Финансовый оборот секторов в различное время

# 4 Вывод

В ходе работы была осуществлена предобработка текста, по выявленным критериям, которые не позволяли работать с данными при помощи Pandas. После чего сформулировали 5 гипотез и при анализе построенных графиков удалось в какой-то мере подтвердить их. Вся визуализация данных выполнялась при помощи библиотеки matplotlib. С большим числом графиков по данному разделу вы можете найти в Google colab.