Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной математики и кибернетики

Никонов Максим Викторович
316 группа

2020

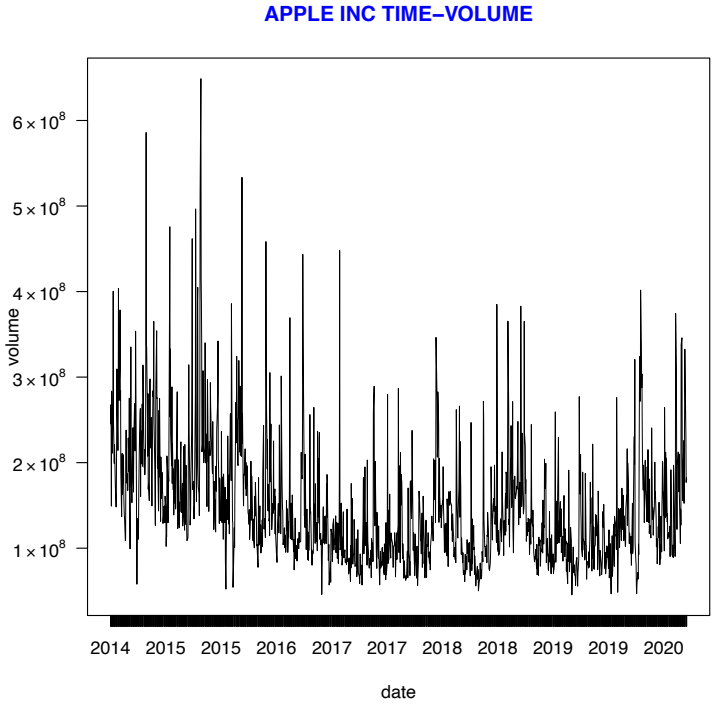
Для семестрового задания была выбрана выборка по акциям APPLE за 2014 по 2020 год по дням. Всего было взято 1510 наблюдений

```
Date,Open,High,Low,Close,Adj Close,Volume
2014-09-15,25.702499,25.762501,25.360001,25.407499,23.084803,245266000
2014-09-16,24.950001,25.315001,24.722500,25.215000,22.909901,267632400
2014-09-17,25.317499,25.450001,25.147499,25.395000,23.073446,243706000
2014-09-18,25.482500,25.587500,25.389999,25.447500,23.121145,149197600
2014-09-19,25.572500,25.587500,25.125000,25.240000,22.932617,283609600
2014-09-22,25.450001,25.535000,25.145000,25.264999,22.955326,211153600
2014-09-23,25.150000,25.735001,25.135000,25.660000,23.314219,253608800
2014-09-24,25.540001,25.712500,25.299999,25.437500,23.112061,240687200
2014-09-25,25.127501,25.177500,24.430000,24.467501,22.230740,400368000
2014-09-26,24.632500,25.187500,24.600000,25.187500,22.884920,249482000
2014-09-29,24.662500,25.110001,24.657499,25.027500,22.739540,199065200
2014-09-30,25.202499,25.385000,25.132500,25.187500,22.884920,221056400
2014-10-01,25.147499,25.172501,24.674999,24.795000,22.528292,205965200
2014-10-02,24.817499,25.055000,24.510000,24.975000,22.691843,191031200
2014-10-03,24.860001,25.052500,24.760000,24.905001,22.628239,173878400
2014-10-06,24.987499,25.162500,24.855000,24.905001,22.628239,148204800
2014-10-07,24.857500,25.030001,24.682501,24.687500,22.430620,168376800
2014-10-08,24.690001,25.277500,24.577499,25.200001,22.896267,229618800
2014-10-09,25.385000,25.594999,25.152500,25.254999,22.946241,309506000
2014-10-10,25.172501,25.507500,25.075001,25.182501,22.880367,265326400
2014-10-13,25.332500,25.445000,24.952499,24.952499,22.671396,214333600
2014-10-14,25.097500,25.129999,24.642500,24.687500,22.430620,254754400
2014-10-15,24.492500,24.787500,23.795000,24.385000,22.155775,403734400
2014-10-16,23.887501,24.430000,23.852501,24.065001,21.865028,28861800
2014-10-17,24.375000,24.750000,24.202499,24.417500,22.185303,272718800
2014-10-20,24.580000,24.990000,24.555000,24.940001,22.660042,310069200
2014-10-21,25.754999,25.754999,25.317499,25.617500,23.275604,378495600
2014-10-22,25.709999,26.027500,25.650000,25.747499,23.393719,273052400
2014-10-23,26.020000,26.262501,25.907499,26.207500,23.811670,284298800
2014-10-24,26.295000,26.372499,26.132500,26.305000,23.900255,188215600
2014-10-27,26.212500,26.370001,26.174999,26.277500,23.875267,136750800
2014-10-28,26.350000,26.684999,26.337500,26.684999,24.245510,192243600
2014-10-29,26.662500,26.842501,26.590000,26.834999,24.381798,210751600
2014-10-30,26.740000,26.837500,26.475000,26.745001,24.300032,162619200
2014-10-31,27.002501,27.010000,26.802500,27.000000,24.531713,178557200
2014-11-03,27.055000,27.575001,27.002501,27.350000,24.849722,209130400
2014-11-04,27.340000,27.372499,26.930000,27.150000,24.668011,166297600
2014-11-05,27.275000,27.325001,27.032499,27.215000,24.727062,149743600
```

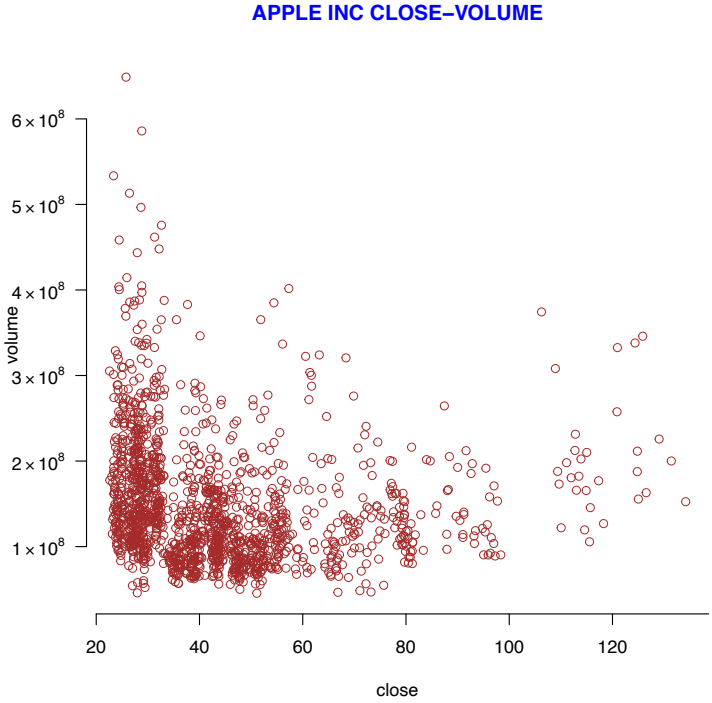
В ходе выполнений домашних заданий были выбраны столбцы Date, Close, Volume, отражающие дату, конечную цену и объем продаж. Стоит отметить, что дата имеет формат int, поэтому каждый раз приходилось переформатировать данные, иногда как Фактор, иногда как дату. Также в ходе выполнения добавились столбцы num и num2, которые помогли исследовать функции, которые были бы неприменимы к моим данным. Отметим, что столбцы генерируются каждый раз случайным способом и не отражены на рисунке.

Рассмотрим представленные данные

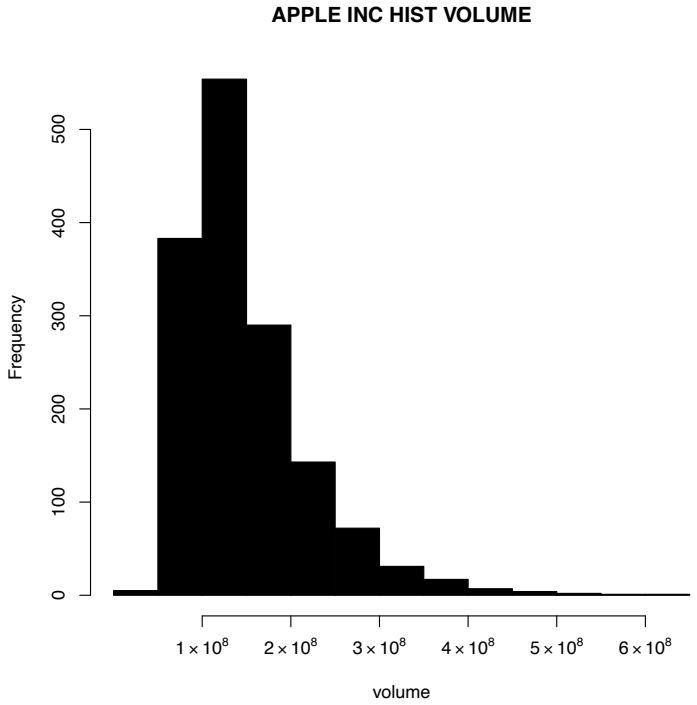
Связь между конечной ценой от времени, можно отметить, что в целом цена за единицу в течением времени растет

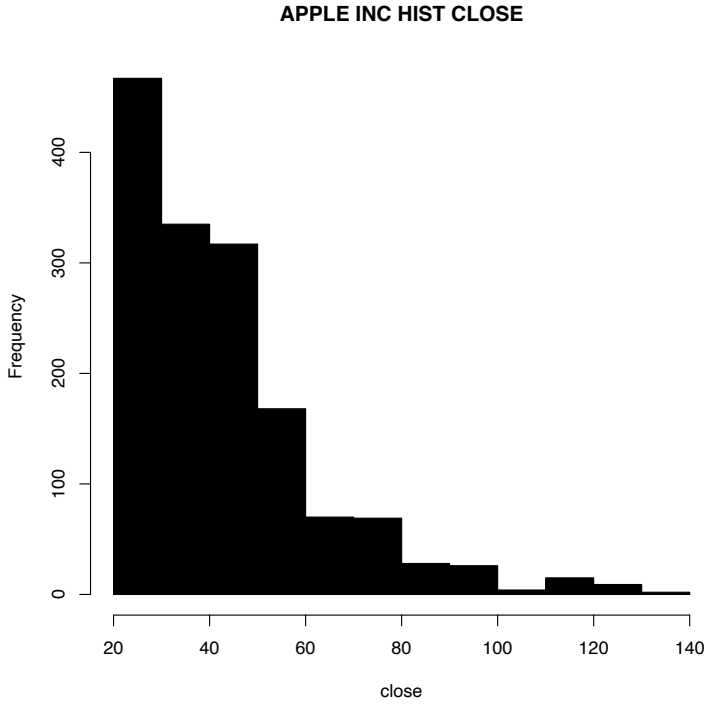


Связь между объемом продаж от времени, данные в целом сильно разнятся в течение года

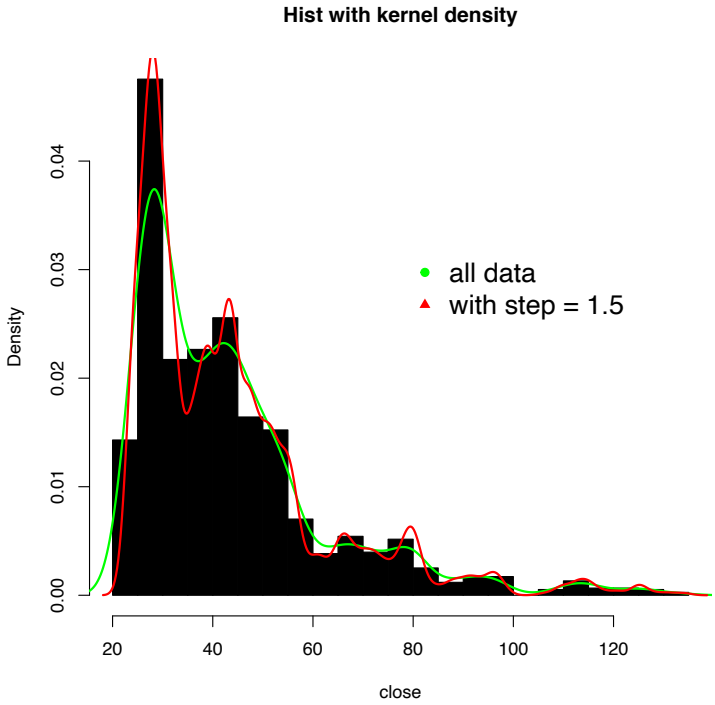


Связь между объемом и ценой, можно заметить, что плотность точек в начале координат больше, чем в конце, по мере увеличения цены, объемы сильно уменьшается, в то время как при увеличении объема, не ограничивая общности справедливо сказать, что за наибольшую заинтересованность в покупке составляются более дешевые акции, хоть и тенденция роста конечной цены акций большая, на предыдущем этапе все меньше людей заинтересованы покупать их



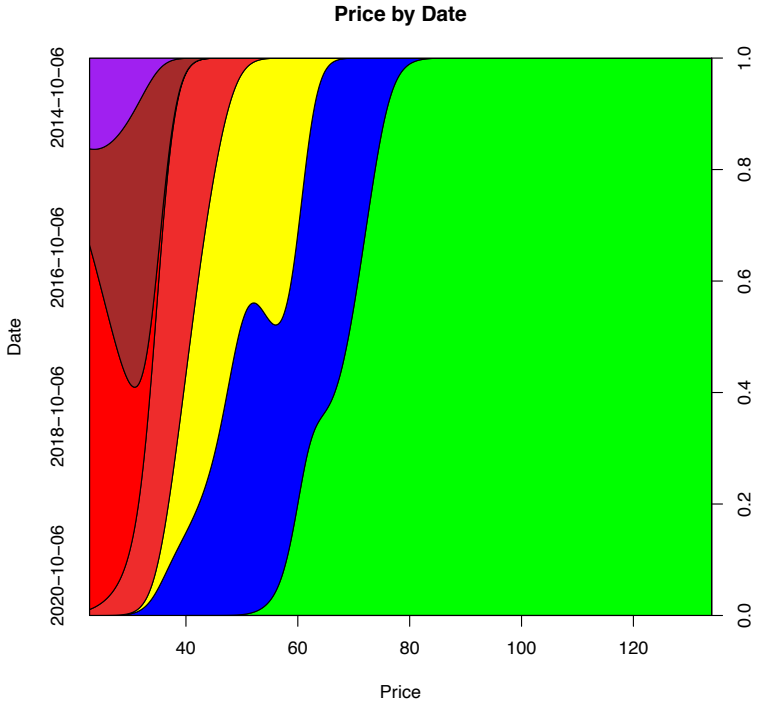


Построение гистограмм дает более общее понимание для исследования, не трудно заметить, что большая часть людей предпочитало покупать акции в размере от 1*10ˆ8 до 2*10ˆ8, в то время, как наиболее популярная цена наблюдалась на промежутке от 20 до 40 условных единиц (в настоящем в долларовом эквиваленте) за одну единицу

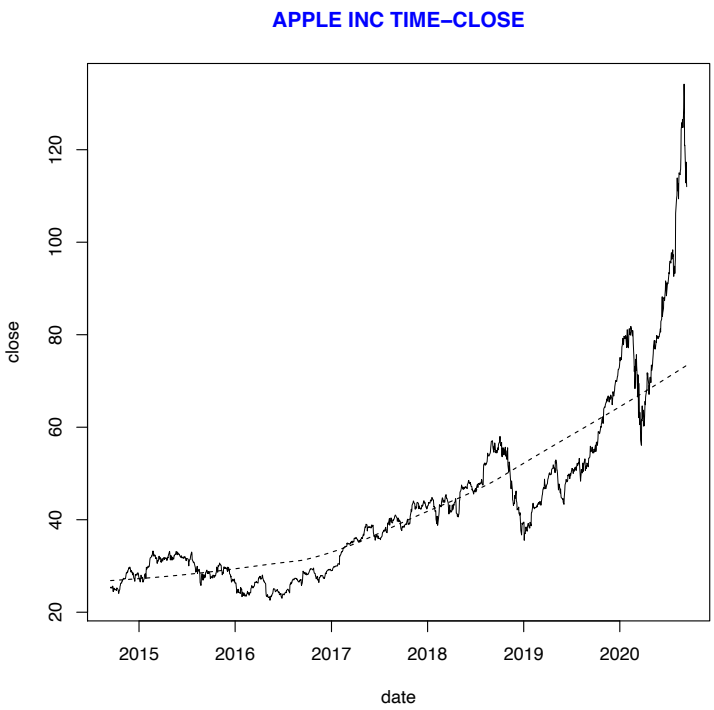


Это так же можно заметить при разной аппроксимации гистограммы, выбранный промежуток явно отражен

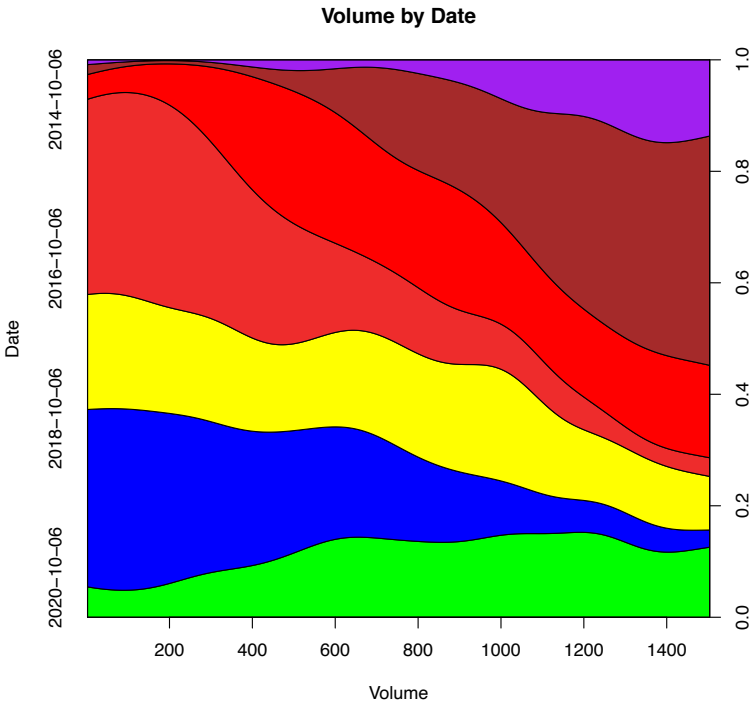
Можно разбить выборку на подвыборку по годам, что нетрудно сделать, из этого



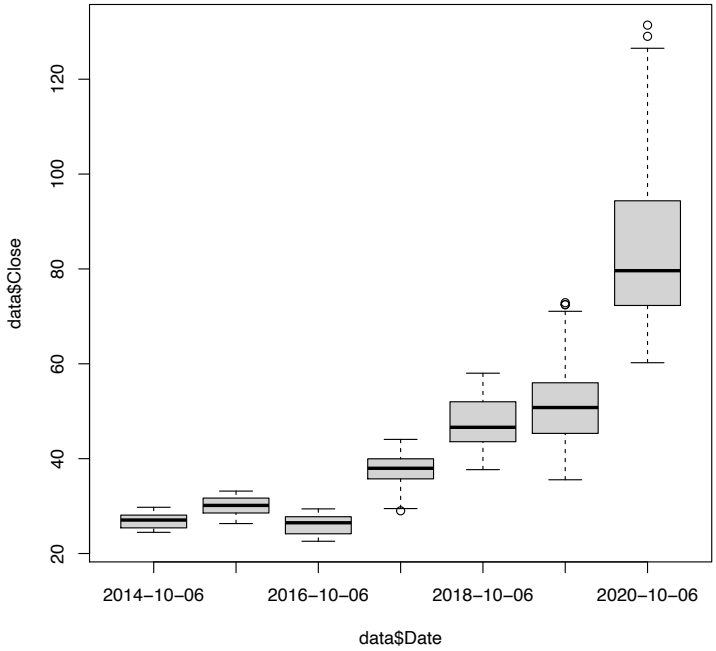
Можно наглядным образом показать в какое время можно было купить акцию по определенной цене, каждый цвет в свою очередь символизирует конкретный год, так например Зеленый цвет иллюстрирует собой 2020 год



Да, не трудно убедиться, что если бы я хотел купить акцию по цене 120 условных единиц (в настоящем в долларом эквиваленте), то я бы смог сделать это только в 2020 году

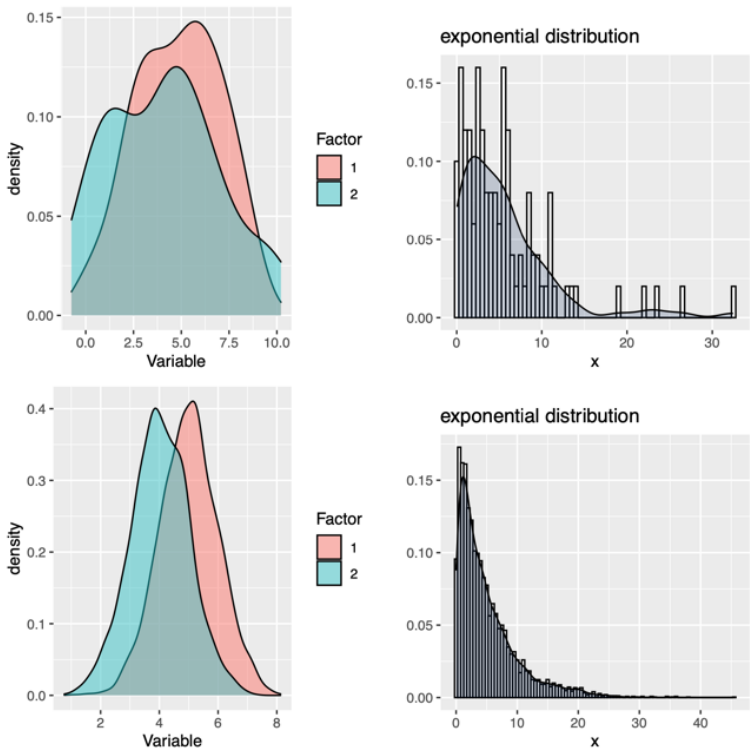


Проводя аналогичные действия и для объема, можно наглядно понять, в каком году и по какой цене можно было купить единицу акций



Черная черта — это медиана (среднее значение) то, что в прямоугольнике вверх и в низ серым цветом это 25% данных Пунктир это 75% данных. Все остальное это выбросы, которые обозначаться кругами

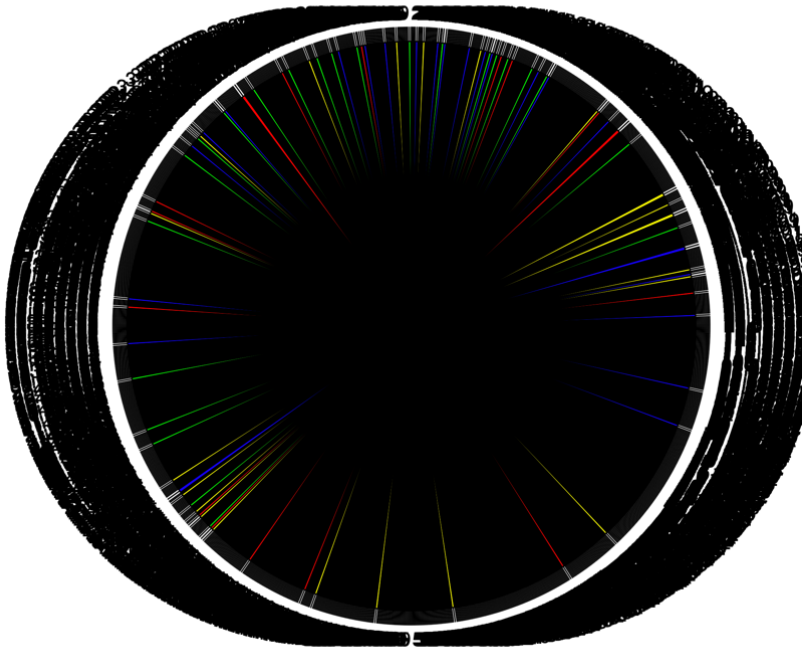
Построение boxplot дает явное понимание, что выбросов не так много, но все же они есть, подчеркнем, что в реальных данных это наблюдается почти повсеместно, однако в представленных данных это наблюдается не так часто в сравнении с другими из личного опыта. Отметим, что максимально значение цены у нас совпадает с выбросом по нескольким методам ! Это достаточно удобно исследовать. На практике построение и анализ boxplot крайне необходим, поскольку, например, если бы за этим следил бот, то он смог бы купить выбросы снизу и продать их к цене, близкой к нормальной для данного участка, именно такой алгоритм использует примерно 70% ботов на фондовых рынках, в свою очередь боты составляют примерно 44% от общего числа участников торгов, то есть это достаточно развито и даже в России, из личного опыта такие алгоритмы используются на торговой площадке Steam



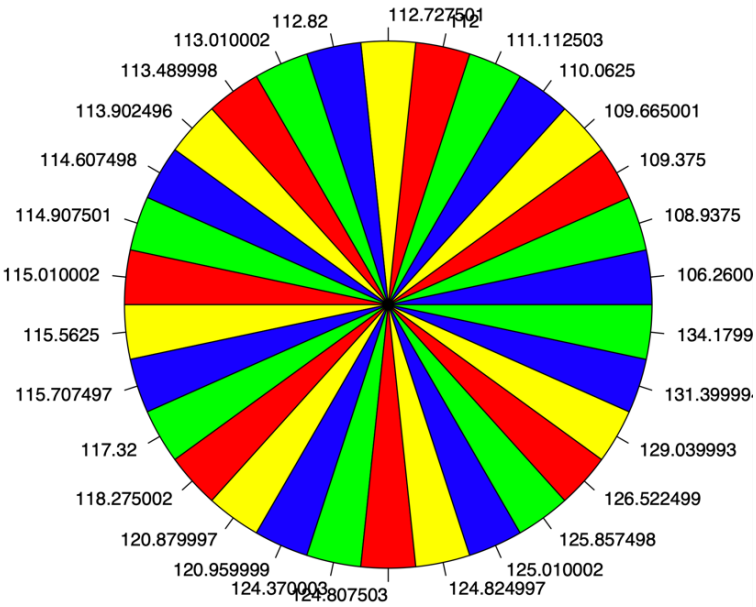
Были сгенерированы плотности с большим количеством наблюдений и с малым

Из этого было наглядно продемонстрирован факт того, что при увеличении выборки, данные стремятся к своему нормальному распределению, что достаточно очевидно, но теперь это и наглядно понятно

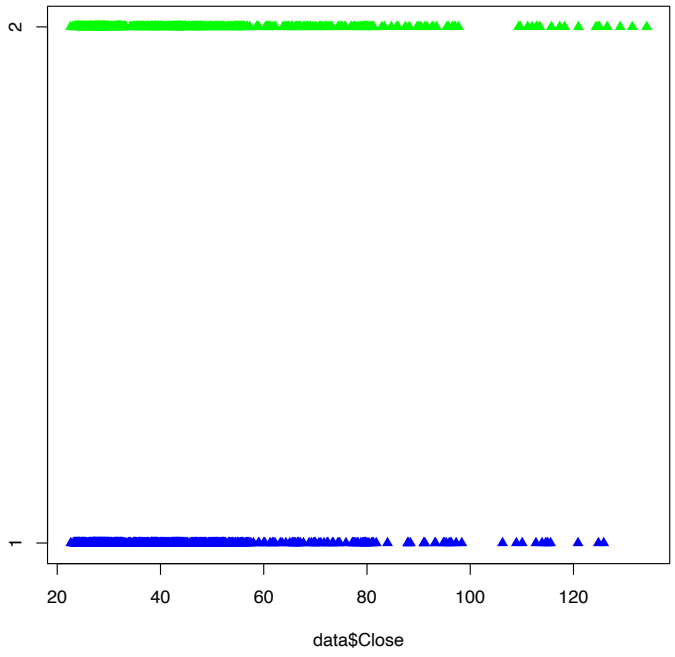
Соответственно при большом количестве наблюдений можно понять, какое это было распределение, однако при малом количестве об этом можно лишь догадываться, тк например нормальное и биномиальное распределения похожи

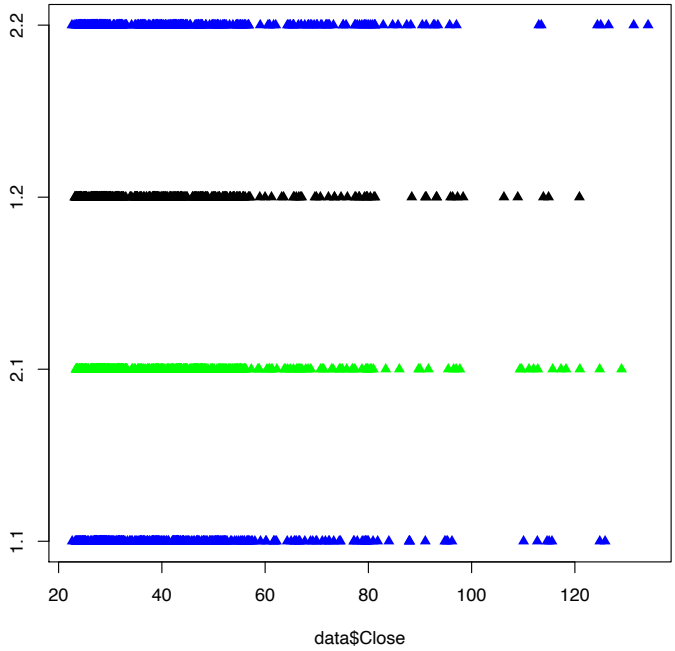


Рассматривая pie для цены сложно делать какие-то выводы в виду полного отсутствия информативности, это происходит из-за большой выборки.



Но даже урезание выборки не дает свои плоды, поскольку секторы демонстрируются одинаковыми, не дает представления о своих промежутках, для анализа необходимо группировка диапазонов

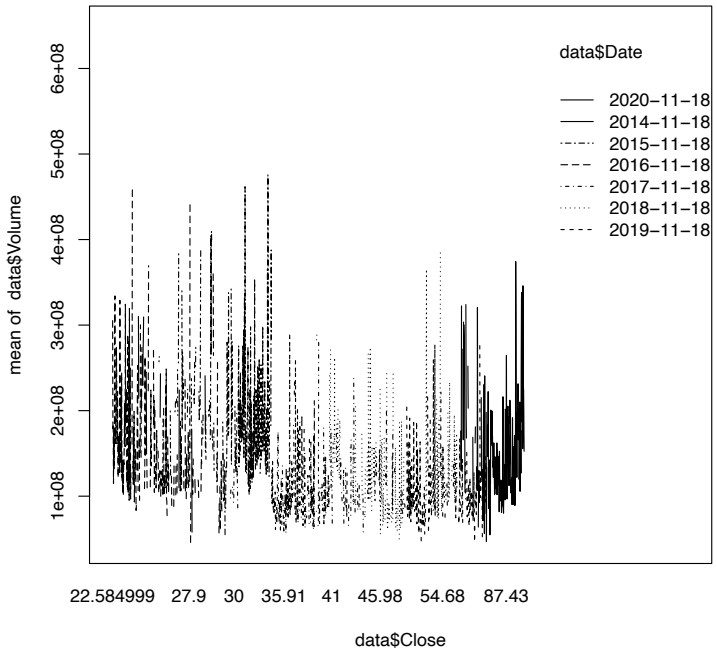




По

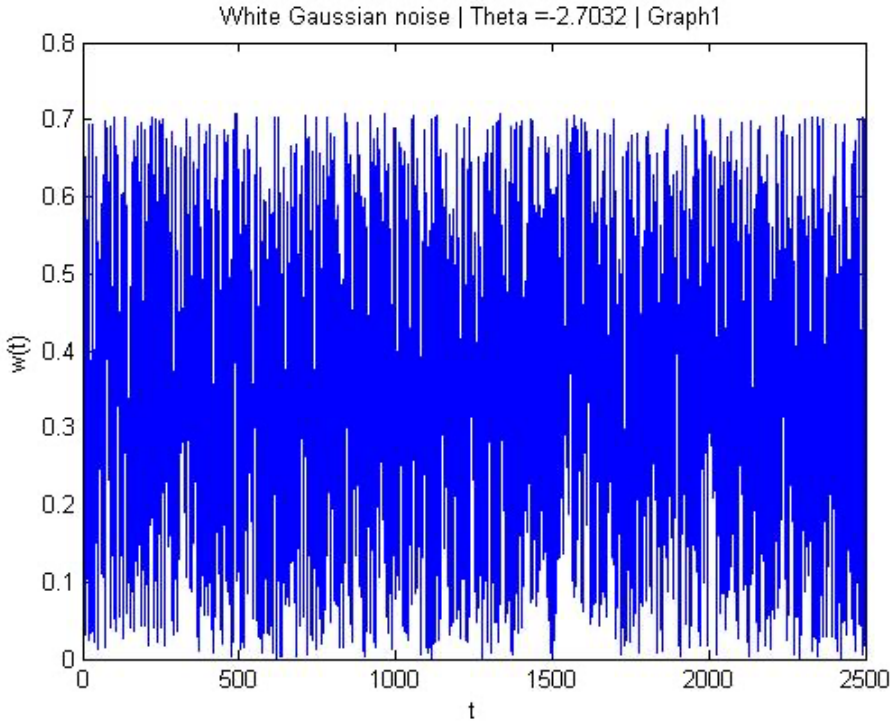
По представленным картинам сложно делать какие-то выводы, однако можно смело говорить о том, что выбросы не сильно влияют на картину в целом и их можно убрать, однако полученная выборка является в корне другой, так например мы знаем, что максимальное значение это выброс, поэтому точность предсказаний будет неточной

Выборка оказалась удобной для применения метода хи-квадрат, но совершенно неудобно для теста Фишера, МакНемара. Проверялось влияние на цену сезонности и объем, однако как выяснилось, что влияния как такого нет (p-value < 2.2e-16), тесты Фишера и др проводились со сгенерированными даными, поскольку не были применимы для исследования.



Дисперсионный анализ – это форма линейной регрессии, поэтому в идеале существует линейная связь между независимыми переменными и зависимой переменной. Одним из источников нелинейной связи является взаимодействие между двумя независимыми переменными: когда одна переменная меняет значение, другая меняет свою связь с зависимой переменной. Проверка взаимодействия между независимыми переменными является базовой диагностикой.

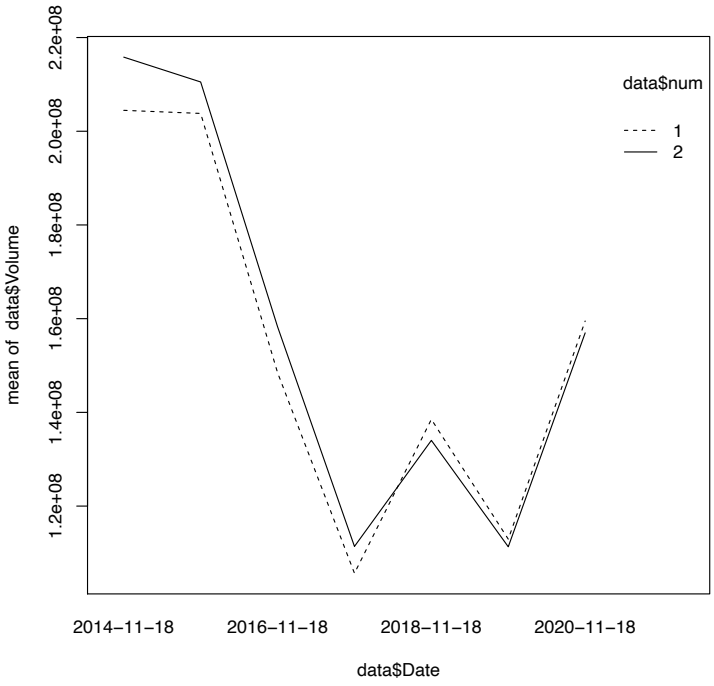
Для демонстрации зависимости дата выступает в качестве предикта, нетрудно заметить, что иллюстрация представляет собой Гауссовский шум, который продемонстрирован ниже.



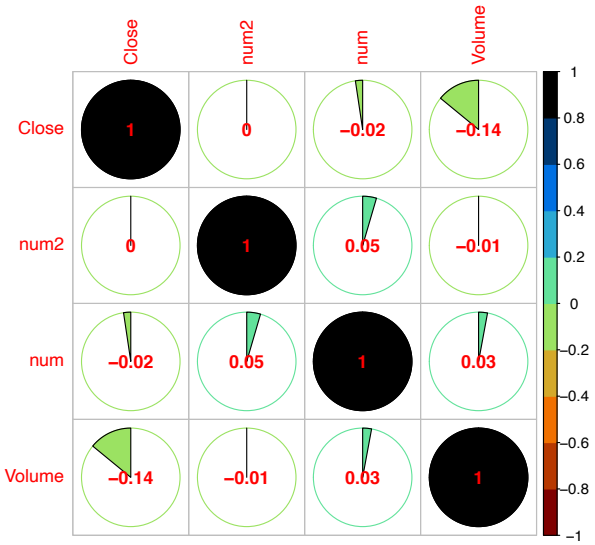
Для построения линейной регрессионной модели такие данные не подходят, однако в целом возможность построение регрессионной модели сохраняется, так для следующего шага построения предсказания необходимы некоторые дополнительные действия для построения например интегральной модели авторегрессии, называемая ARIMA, которая используется сейчас достаточно часто. И это возможно, но нельзя забывает о таких факторах, как цикличность, сезонность, их влияния здесь нет, но это не все факторы

Обычный дисперсионный анализ предполагает, что ваши данные имеют нормальное распределение. Он может допускать некоторое отклонение от нормальности, но крайние отклонения приведут к бессмысленным p-значениям. Критерий Краскела–Уоллиса представляет собой непараметрическую версию ANOVA. Это означает, что он не предполагает нормальности. Тем не менее он предполагает распределения одинаковой формы. Критерий Краскела–Уоллиса следует использовать как раз для данной выборки

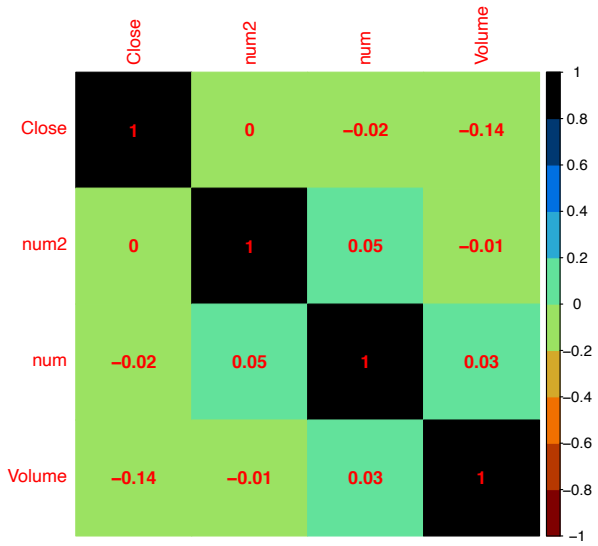
Для демонстрации регрессионной модели были взяты сгенерированные столбцы

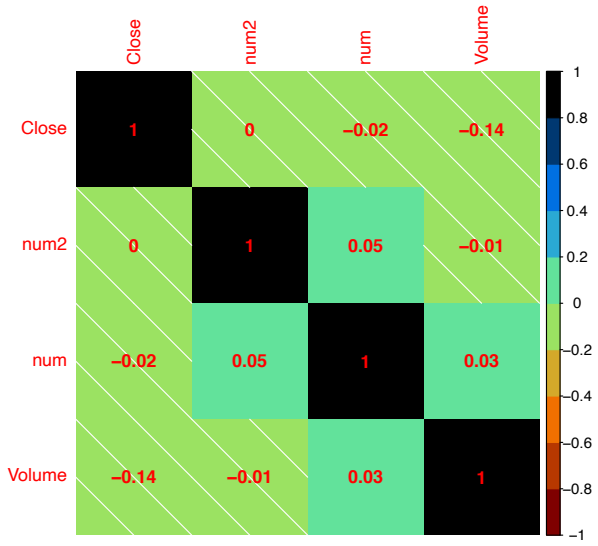


На тех участках, где прямые были параллельны, очевидно, есть линейная зависимость, на остальных участках зависимости не наблюдается



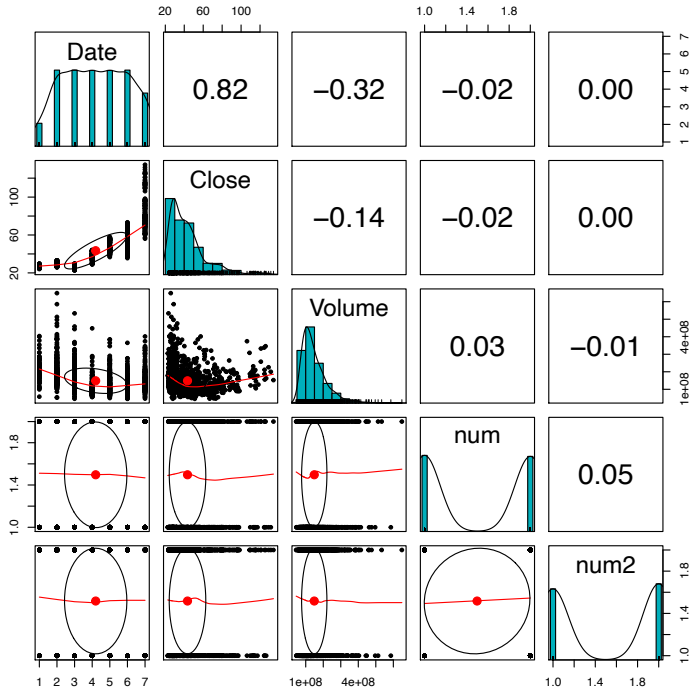
Исследуя график автокорреляции, нетрудно заметить, что данные практически не коллериуют между собой, что естественно плохо, но ожидаемо для реальных данных. Очевидный факт, что данные полностью коррелируют сами с собой, это и ожидалось, ведь данные одни и те же





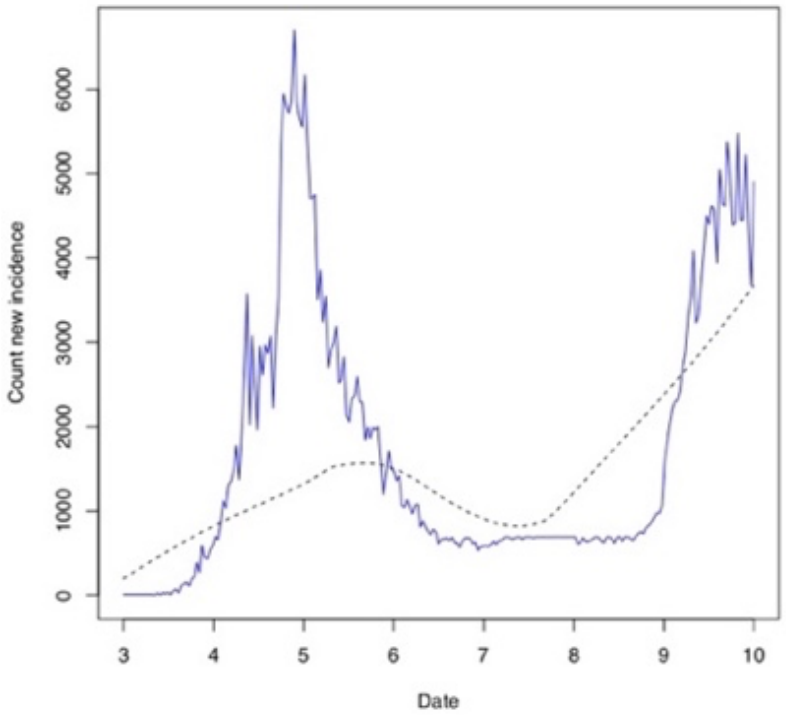
Наглядно продемонстрированно это и в других реализациях

В итоге о данных можно судить исходя из следующего рисунка, отметим, что все находится в коридоре



Так в верхнем треугольнике видны коэффициенты корреляции, по диагонали наблюдаем сами графики, а в нижнем треугольнике наглядно показаны средние приближения для каждой связи (красным цветом)

Подводя некоторые итоги, можно сказать, что цена акций на компанию APPLE стабильно растет и не поддается сезонности, цикличности, на общую тенденцию не сильно влияет влияет и кризис в мире, связанный с пандемией COVID-19 однако можно использовать это для анализа и применения методов машинного обучения, используя смеси распределений, например с помощью пакетов mixtools или RandomForest, используя EM алгоритмы, ниже приведена статистика заболеваний COVID-19



EM-алгоритм приводится в качестве одного из способов кластеризации. Другими словами, это метод машинного обучения без учителя, когда нам заранее не известны истинные ответы. EM-алгоритм (англ. expectation-maximization) — это общий метод нахождения оценок функции правдоподобия в моделях со скрытыми переменными, который из смеси распределений позволяет строить (приближать) сложные вероятностные распределения.

EM-алгоритм в задачах кластеризации используется как итеративный алгоритм, который на каждой итерации осуществляет два шага:

E-шаг. На первом E-шаге мы каким-либо образом, например, случайным, выбираем скрытые переменные, в нашем случае это будут математическое ожидание и стандартное отклонение . Используя выбранные переменные, рассчитываем вероятность отнесения каждого объекта к тому или иному кластеру. При последующих E-шагах используются скрытые переменные, определенные на M-шагах.

M-шаг. На M-шаге мы, в соответствии с полученными на E-шаге значениями вероятностей отнесения каждого объекта к тому или иному кластеру, пересчитываем скрытые переменные. Итерации повторяются до тех пор, пока не наступит сходимость

Таким образом можно прогнозировать данные, однако очевидно, что при каждом следующем дне прогноза вероятность будет падать

APPLE INC является стабильной на рынке акций и показывает не только устойчивость, но и рост своих акций, в современном мире это очень редко наблюдается, в действительности многие компании в мире сильно просели за аналогичным период

Компания демонстирует именно тот подход, который необходим многим компаниям. Однако как этого достичь остается нерассмотренной задачей, требующей еще долгого глубокого анализа