МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Ходор Иван Андреевич

Обработка и структурный анализ данных использования общественного транспорта

Курсовой проект
студента 3 курса 3 группы

"Допустить к защите"

_____

"___"_____ 2020г

Научный руководитель
Вертинская А.Е.
ассистент кафедры ДМА

МИНСК 2020

Содержание

Введение

За последнее столетие популярность автомобиля как личного средства передвижения возросла просто колоссально. Каждая среднестатистическая семья имеет как минимум одну машину, но не редкость так же и 2, и 3 автомобиля на семью. Но содержание личного автомобиля - довольно хлопотное дело: необходимо иметь парковочное место, пополнять бак топливом, проводить регулярный технический осмотр. Гораздо полезнее для собственного здоровья и окружающей среды передвигаться пешком или с помощью велосипеда, но что же делать, если в конкретный момент важна скорость и комфорт? На помощь приходит общественный транспорт!

Больше всего нас интересует такси. Раз речь идёт о скорости, то пассажиру жизненно необходимо иметь представление о том, как долго придётся ожидать такси, как долго продлится сама поездка. Для решения этих задач не существует чёткого алгоритма, так как существует широкое множество факторов, оказывающих влияние на конечный результат. Именно потому на помощь приходят методы машинного обучения, способные находить и строить скрытые зависимости между, на первый взгляд, несвязанными вещами.

В данной работе попробуем предсказать время поездки в такси в городе Нью-Йорк на основе информации о поездках в жёлтых такси за периоды май 2019 и май 2020 годов. Также это хорошая возможность проследить некоторые изменения, наступившие из-за эпидемии COVID-19, связанные с карантином, социальным дистанцированием и подобным.

Глава 1
Подготовка данных

1.1 Постановка задачи

По имеющимся данным о поездках построить алгоритм, предсказывающий длительность поездки различными методами машинного обучения, описать недостатки и достоинства каждого используемого метода.

1.2 Описание данных

1.2.1 Исходные данные

Для решения поставленной задачи были выбраны данные с Национального ресурса хранения информации о передвижении такси «NYC(taxi limousine comission)». Было выбрано два периода: май 2019 и май 2020. Причинами послужило несколько факторов: во-первых, благоприятная погода для перемещений, что позволило сформировать большую выборку; во-вторых, после начала карантина прошло уже некоторое время, что может позволить проследить поведение жителей после некоторого времени нахождения на самоизоляции.

Посмотрим на предоставленные данные:

1. VendorID - номер одной из двух компаний, которая занималась перевозкой.

2. tpep pickup datetime, tpep dropoff datetime - дата начала и конца поездки.

3. passenger count - количество пассажиров.

4. trip distance - расстояние поездки в милях.

5. PULocationID, DOLocationID - номера районов начала и конца поездки(чуть конкретнее рассмотрим ниже).

6. store and fwd flag - информация о том, как была получена информация о поездке: передавалась во время поездки или была выгружена после конца смены водителя.

7. payment type - тип оплаты поездки(кредитной картой, наличными деньгами, неизвестно, поездка «в долг», за бонусные баллы по локальной программе).

8. extra - 0.5 или 1 доллар доплаты за ночную поездку.

9. tip amount - автоматически заполняется при оплате чаевых через карту.

10. total amount - общая сумма, оплаченная за поездку.

В данных 2020 года есть пропущенные значения в колонках о количестве пассажиров, получении системой данных, номере перевозчика и типе оплаты. Т.к. всё это категориальные данные, заполним пропуски самым частым значением, которое встречается в остальных данных.

1.2.2 Формирование новых данных

Очевидно, что из имеющихся данных можно получить дополнительную информацию. Получим следующие данные:

1. Число начала поездки.

2. Час начала и конца поездки.

3. День недели начала поездки(день конца игнорируем, потому что таких поездок очень мало).

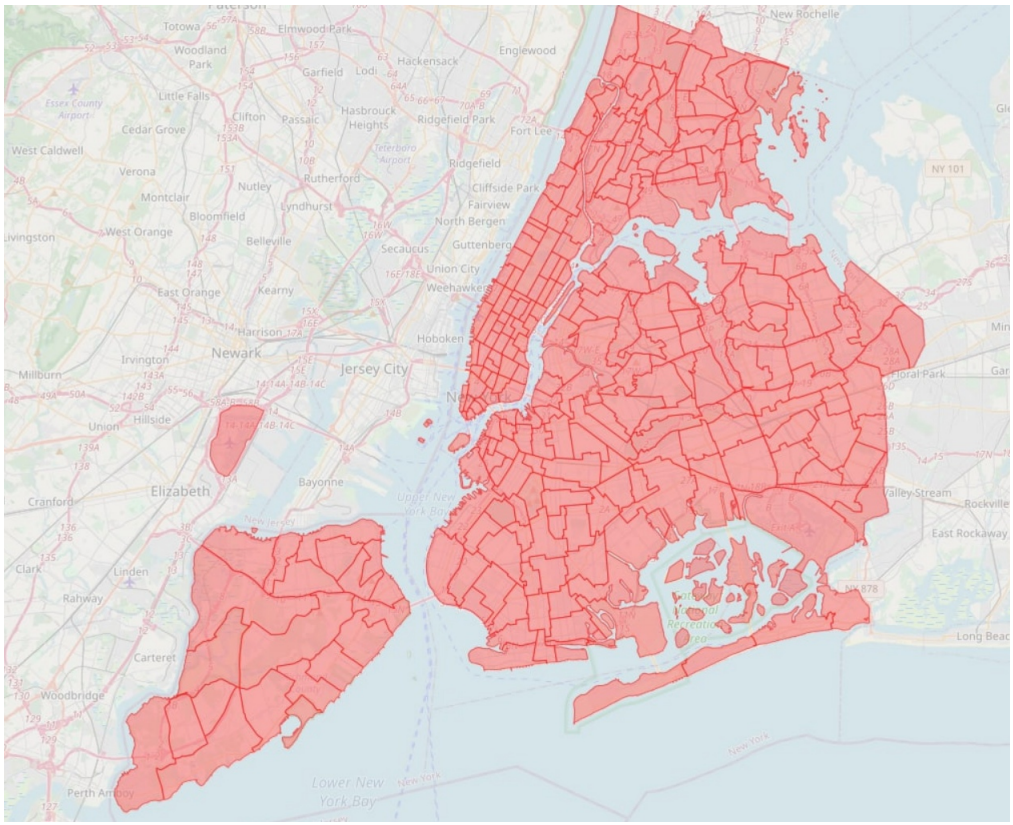
4. Время продолжительности поездки в секундах.

5. Средняя скорость в м/с.

Также заметим, что в последний понедельник мая(27.05.19 и 25.05.20) в Америке проводится федеральный праздник День Памяти, являющийся выходным.

1.2.3 Удаление невалидных данные

Для удаления нереальных данных нужно сначала разобраться с тем, что такое LocationID. Весь Нью-Йорк поделен на 263 района для более удобной организации различных процессов, связанных с такси:



Соответственно удалим все поездки, у которых номера районов выходят за границы отрезка $[1; 263]$. Расстояние в данных даны от точки начала до точки конца, но мы не имеем конкретных данных об этих пунктах(лишь о зонах), что немного затрудняет анализ и визуализацию, но тем не менее можно проследить популярные направления.

Очевидно, что стоит удалить поездки, у которых время окончания поездки раньше, чем начала. Также установим ограничения для средней скорости: от 1 до 50 м/с, что явно достаточно для такого города, как Нью-Йорк с его большим трафиком. Ещё нас не интересуют поездки с отрицательной стоимостью, и поездки, в которых было преодолено нулевое расстояние за ненулевое время, при этом стоимость положительна.

Глава 2
Предварительный анализ

2.1 Некоторые статистики, предоставляемые Python

Посмотрим на информацию, которую для pandas.DataFrame можно получить с помощью функции describe().

2019 год	Кол-во пас-ров	Расст.(м)	Длит-сть(сек)	Средняя скорость(м/с)
Среднее	1.57	4875	899.7	5
Мин.	0	16	2	1
Макс.	9	319196.5	51841	50

2020 год	Кол-во пас-ров	Расст.(м)	Длит-сть(сек)	Средняя скорость(м/с)
Среднее	1.27	5997	707	7.45
Мин.	0	16	2	1
Макс.	6	491009.6	42960	50

Сделаем несколько заключений:

1. Среднее количество пассажиров снизилось с 1.57 до 1.27 из-за социального дистанцирования. Также максимальное количество уменьшилось от 9 до 6 человек(среди жёлтых такси так же есть минивены, потому такое большое количество неудивительно).

2. Средняя дистанция в 20м году выросла с 5км до 6км. Можем сделать вывод, что люди стали чаще пользоваться такси для каких-то длительных поездок, ведь если расстояние небольшое, может быть рентабельнее просто пойти пешком, чтобы ни с кем зря не контактировать.

3. Несмотря на увеличение дистанции средняя продолжительность поездки уменьшилась(соответственно выросла средняя скорость) из-за меньшей загруженности дорог.

2.2 Краткое описание районов Нью-Йорка

В Нью-Йорке есть 5 основных районов: Manhattan, Queens, Brooklyn, Staten Island и Bronx.

Manhattan является главным туристическим местом, а также скоплением главных бизнес-центров Нью-Йорка. Тут очень развита инфраструктура (не)активного времяпрепровождения.

Queens является одним из наиболее "спальных"районов города. Однако тут расположен Национальный аэропорт имени Джона Кеннеди - аэропорт с самым большим трафиком пассажирских и грузовых авиаперевозок в мире, что в купе с тем фактом, что тут происходит множество спортивных мероприятий, добавляет посещаемости приезжими.

Brooklyn - самый населённый район Нью-Йорка. При этом тут очень развита инфраструктура локальных мероприятий, что позволяет жителям не ездить куда-то, а "жить на месте".

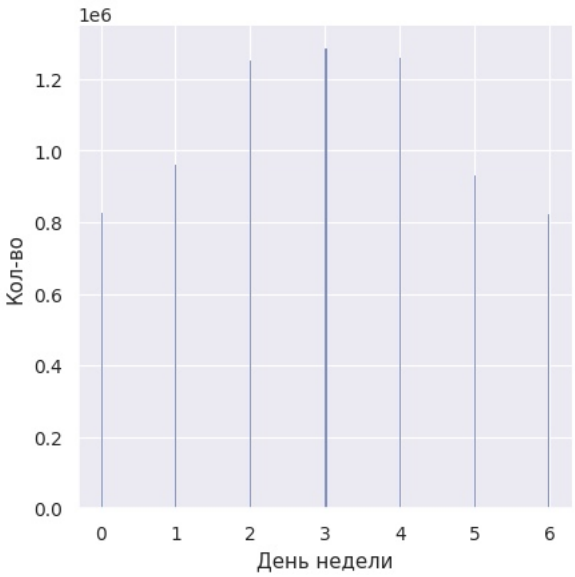
В Bronx'е находится печально известный своей неблагополучностью квартал Camp Apache. Однако сейчас там всё спокойно. Более того, среди туристов имеют популярность несколько иностранных кварталов

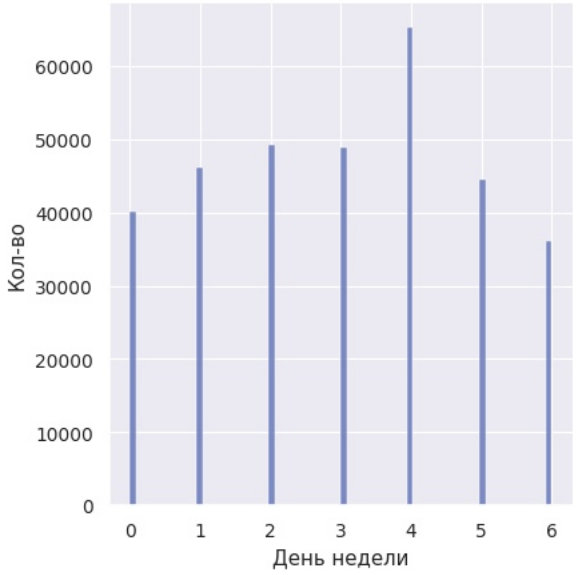
Staten Island - самый тихий район с наименьшей численностью населения. Туристов сюда не тянет в силу того, что тут "почти не на что смотреть". Однако с его замечательной природой можно отдохнуть от суеты других районов.

2.3 Анализ с помощью визуализации

Предварительно отметим, что количество поездок за год очень сильно сократилось: $7.3 \cdot 10^6$ против $3.3 \cdot 10^5$. Потому графики потчи всегда будем приводить отдельно друг от друга, чтобы можно было проследить лишь тенденции к использованию транспорта.

Распределение поездок по дням недели(0-понедельник, 6-воскресенье):

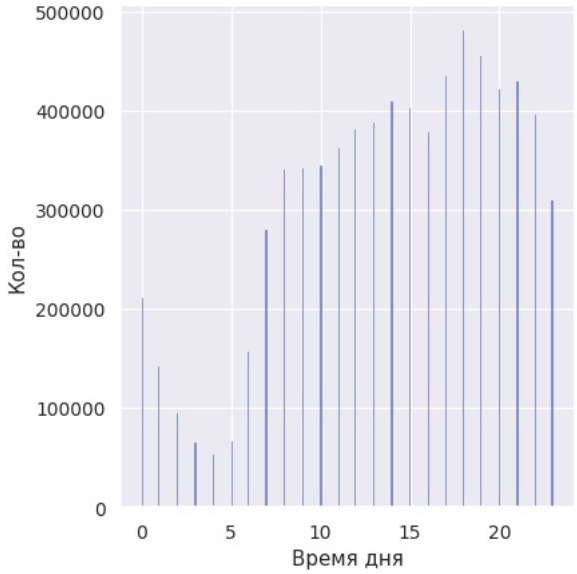


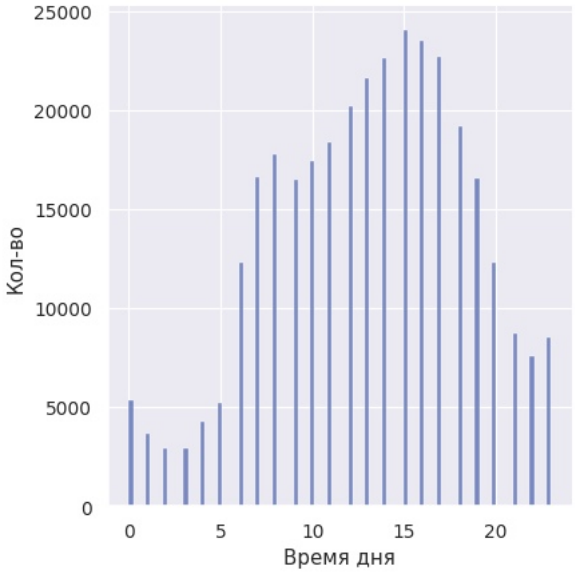


Выше видим выброс в пятницу в 2020м году. Возможно это объясняется режимом самоизоляции, который очень хочется нарушить в конце рабочей недели(либо в целом возвращение с работы домой, если это не является ежедневной рутиной).

В противоложность этому основная нагрузка в 19м году приходится на середину рабочей недели. Возможно, причина этому тот факт, что вторник является самым удачным днём для сложных задач, а дальше уже хочется немного поотлынивать, потому жители занимаются своими делами, в том числе разъездами.

Распределение поездок в течение суток:

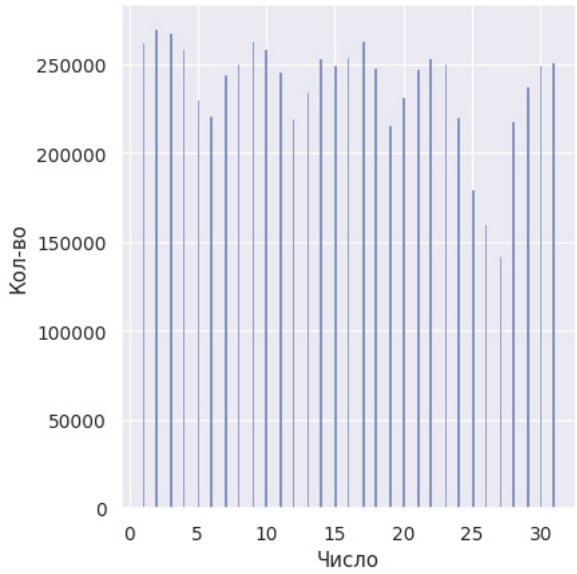


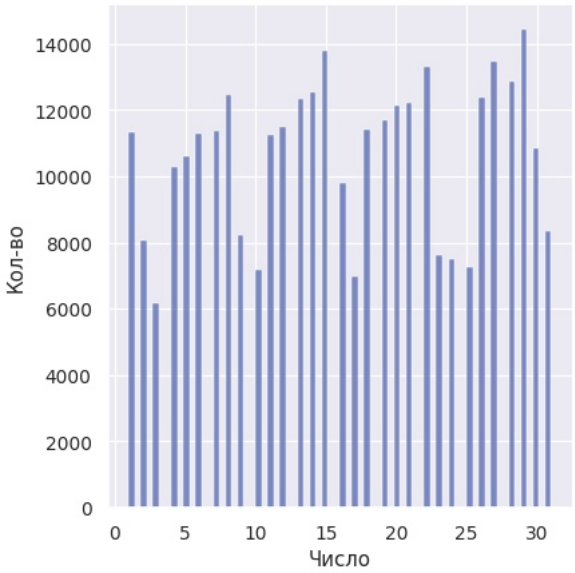


В 19 году видим основную нагрузку на 6-7 часов вечера, т.к. жители возвращаются с работы попутно заезжая по делам.

Однако в 20м году пик приходится на 3 часа дня. Возможно это потому что весь город подумал, что днём будет мало людей на улицах из-за работы и побежал делать свои дела.

Распределение в течение месяца:



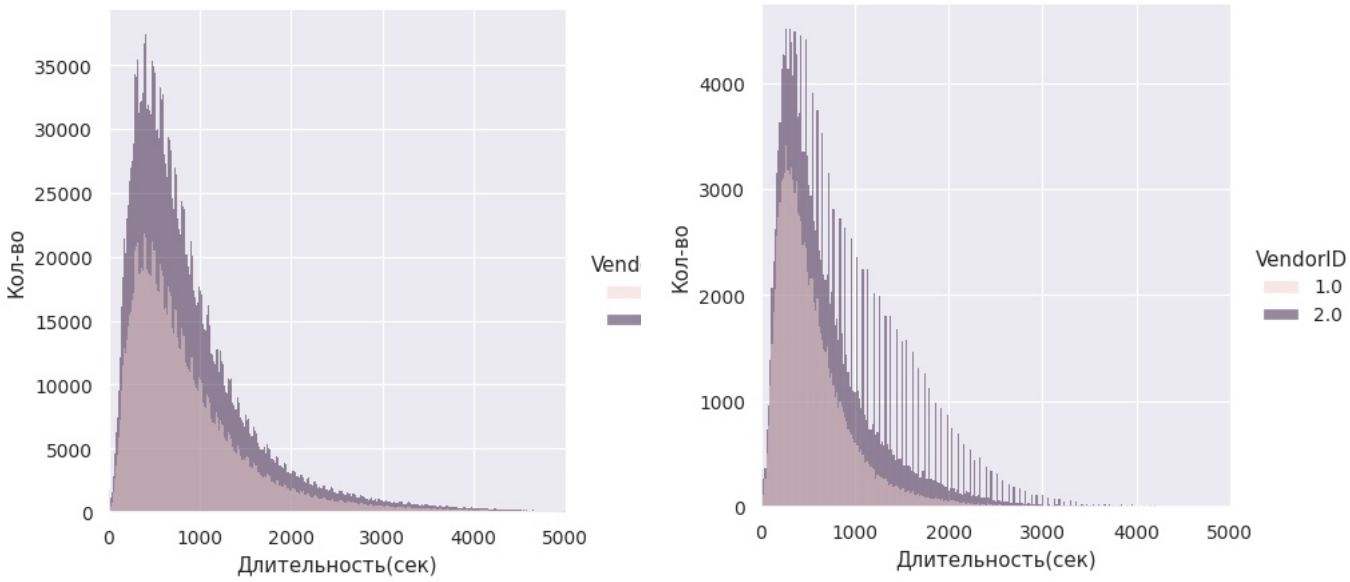


В оба года видим снижения использования средств передвижение во время выходных. Однако в 19м году имеется тенденция к постепенному снижению использования такси в течение рабочих дней(люди начинают готовится

к отпуску и собственно отдыхать), тогда как в 20м году правило обратно: использование повышается, - тут это обуславливается тем, что люди уже устали сидеть некоторое время на карантине и начинают пренебрегать рекомендациями о самоизоляции.

Также отметим резкое(даже относительно выходных) снижение поездок 27.05.2019 и 25.05.2020, т.к., как уже упоминалось, в этот день федеральный праздник День Памяти, являющийся выходным.

Распределение длительностей поездки:

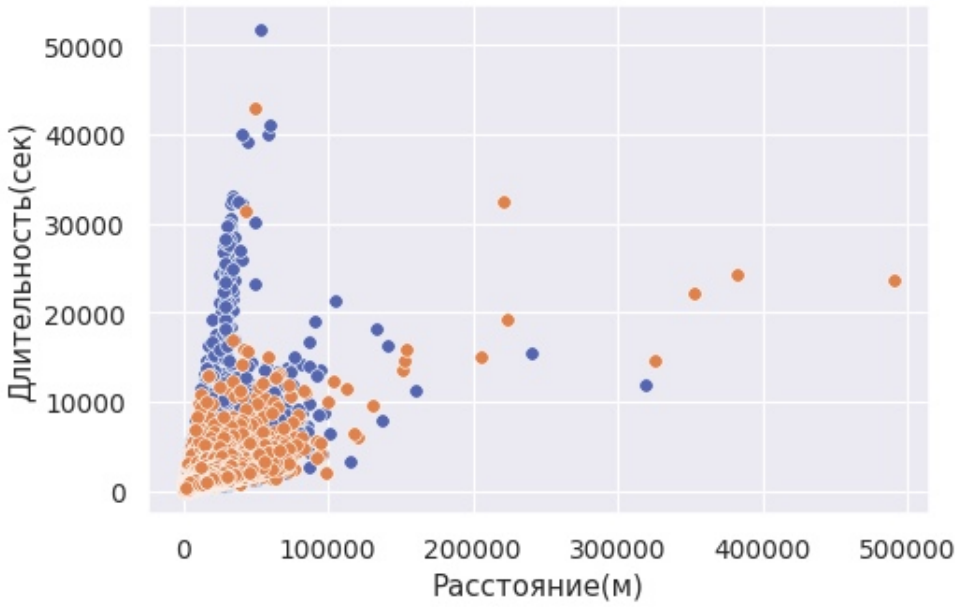


На 2м графике существуют видимые выбросы. Возможно это из-за частых поездок по более популярным маршрутам, которых, как мы видим, существует очень много(например из района Manhattan до аэропорта имени Джона Кеннеди).

Легко увидеть, что пользователи совершают очень много коротких поездок(до 30 минут) и очень мало длинных(более часа). На ум приходят две причины: немалая стоимость длинной поездки + средний пассажир ездит ради скорости и независимости от большинства других факторов(расписание общественного транспорта, движение от точки до точки).

Также видим, что в целом за год появилась тенденция ещё больше использовать в качестве перевозчика компанию №1. Возможно это связано с какими-то локальными улучшениями.

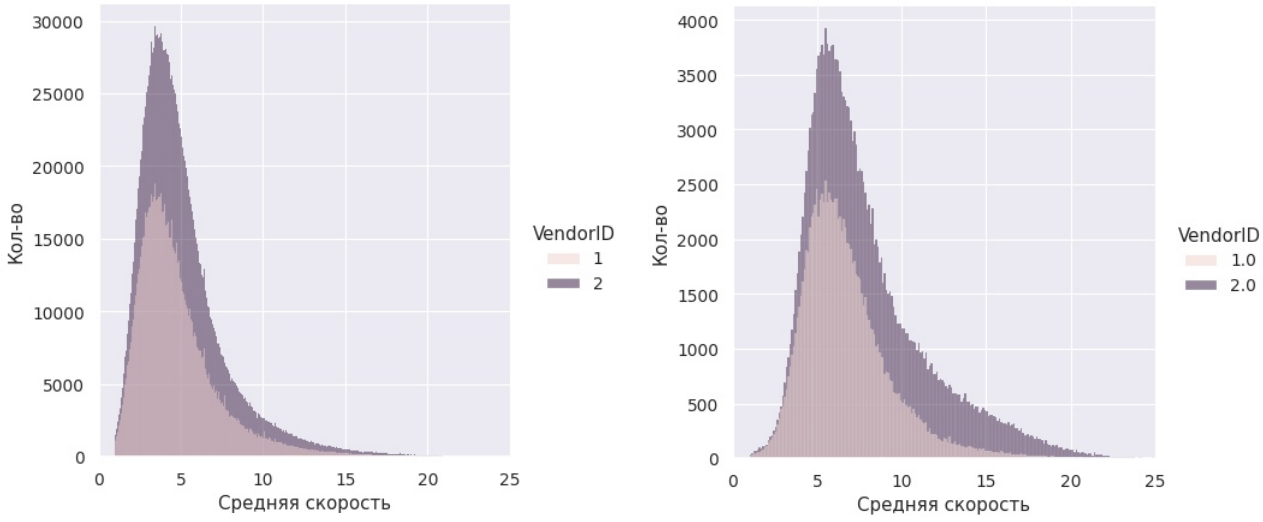
Распределение длительности поездок от времени(синий - 19 год, оранжевый - 20):



Тут видим, что в целом они распределены примерно одинаково относительно друг друга(этому виной в том числе и то, как мы убирали выбросы), т.е. эпидемия повлияла в данном случае только на количество поездок и уменьшение числа длительных поездок.

Но есть очень много поездок с короткими дистанциями, но значительным временем. Это может объясняться большим трафиком в загруженные части дня. Систематически такой проблемы в 20 году не наблюдалось в силу того, что в целом загруженность узлов была меньше, следовательно меньше пробок.

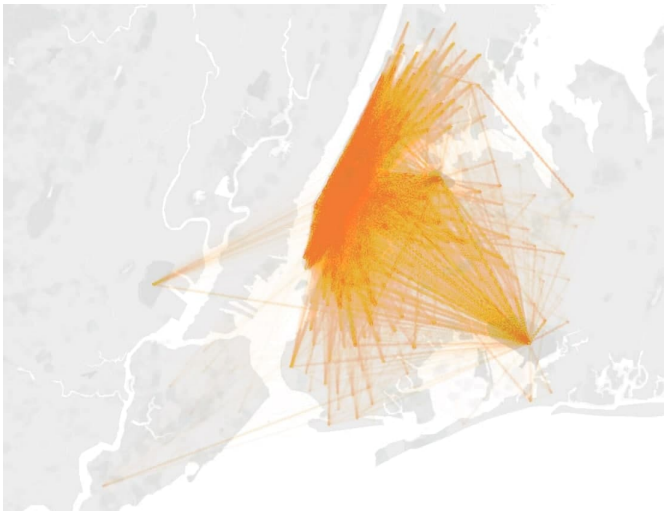
Распределение средней скорости:



Пики расположены около средних скоростей, которые составляют соответственно 18 и 27 км/ч, что вполне нормально для Нью-Йорка с его трафиком и пробками. Для передвижения со скоростью выше 48 км/ч(13.3 м/с) требуется либо быть каким-нибудь служебным автомобилем(скорая, полиция или пожарные), либо нарушать скоростной режим(чем, видимо, не пренебрегает некоторое количество водителей).

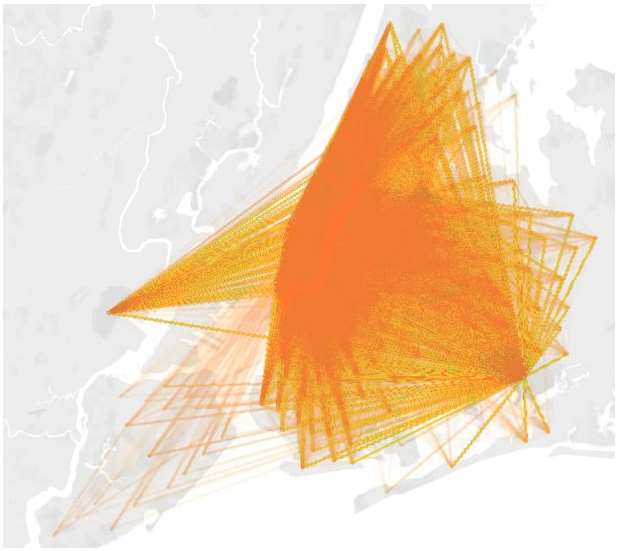
Тут же можем понять, что в 2020 году для быстрых переездов чаще использовался перевозчик №2. Вкупе с данными с графиков о времени переезда можем попробовать сделать вывод, что водители данной компании чаще нарушают: они быстрее ездят, но пассажиры немного опасаются использовать этого перевозчика.

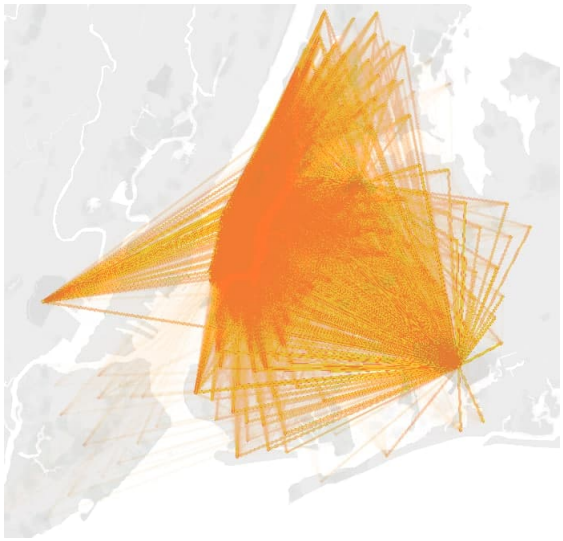
Посмотрим на поездки на картах. В силу того, что в данных есть информация только о зонах, картинка будет специфична.



За весь период мы ничего конкретного не вынесем. Особенно из прошлого года. Однако в этом году можно заметить, что популярным осталось направления Manhattan(и собственно весь он сам) и 2 аэропорта в Queens, а также в 19 году аэропорт Newark.

Поездки за рабочие дни и за выходные в мае 19 года:

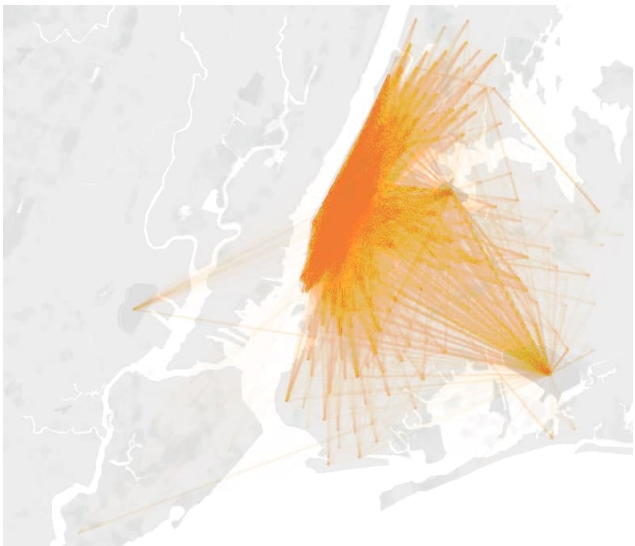


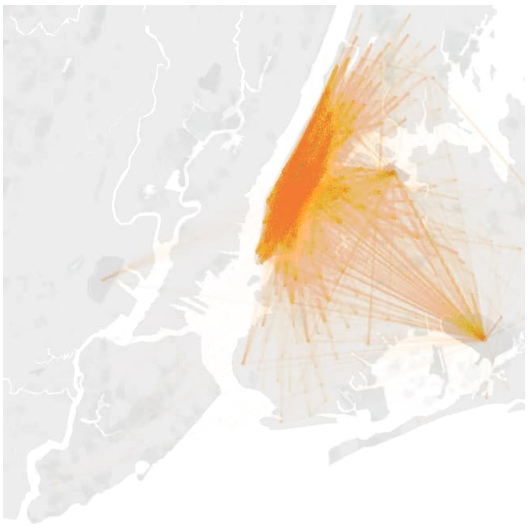


На выходных видим значительное уменьшение поездок в район Staten Island. Это, возможно, обусловлено тем, что много жителей едут в другие районы на работу, а вся инфраструктура для отдыха на выходных присутствует на месте. Более того, вспоминая, что этот район является довольно тихим, можно сделать вывод, что жители не хотят суеты ещё и в выходные.

Также можно заметить небольшое уменьшение трафика из аэропортов на выходных(вероятно потому Нью-Йорк популярный город даже больше для рабочих визитов, чем для отдыха).

Однако в целом плотность в центральных районах не изменилась.





В противоположность этому на выходных 20 года видно значительное снижение использования такси. Популярным остались Manhattan и аэропорты,

но очевидно снижение переездов между дальними частями города. Также можем заметить, что из аэропорта Newark почти полностью перестали ездить.

В целом можно понять, что Manhattan всегда имел большой трафик из-за множества туристов и большого количества бизнес-центров.

Глава 3
Применение различных моделей машинного обучения

3.1 Формирование репрезентативной выборки

В течение обучения я часто встречался с проблемой нехватки ресурсов(а именно оперативной памяти). Это объясняется тем, что используемые модели в sklearn реализованы на C и требуют сразу все данные(что затрудняет копирование выборки большого объёма). В противоположность такому подходу существуют другие модели, которые используют генераторы, что позволяет получать данные блоками фиксированного размера для уменьшения потребления памяти. Однако такие модели не рассматриваются в рамках текущего исследования.

Для решения данной проблемы было решено сформировать случайную выборку из имеющихся данных. Задача сформировать такую выборку, которая будет являться репрезентативной, сама по себе является хорошей темой для исследования, но мы не будем уделять этому много времени. Использовались следующие критерии репрезентативности:

- Урезанная выборка визуально повторяет все основные распределения основной(по времени суток, по дню месяца и т.д.).
- Качество предсказания на моделях для тестовой выборки уменьшается не очень сильно(на самом деле, оно даже немного улучшилось).

На основе этих критериев выборки были урезаны до размеров $9.7 \cdot 10^4$ и $7.9 \cdot 10^4$ соответственно.

3.2 Изменение представления категориальных данных

В имеющейся выборке имеется очень много категориальных данных. Для большинства моделей нельзя оставить их в имеющемся виде. Применим известный приём под названием OneHotEncoding: размножим колонку столько раз, сколько различных значений имеется в ней, после чего установим значение True, если поездка имеет такую категорию, False во всех остальных случаях.

Возникает проблема с колонками PULocationID, DOLocationID, т.к. мы имеем 263 различных значения, из-за чего мы будем использовать очень много памяти. Вместо этого мы сделаем по 5 колонок на пункты начала и конца поездки, где будет указано, какому конкретно району Нью-Йорка принадлежит данная точка. Это значительно ухудшит качество, но лучше так, чем никак.

3.3 Линейная регрессия

Линейная регрессия — используемая в статистике регрессионная модель зависимости одной переменной y от другой или нескольких других переменных x_i с линейной функцией зависимости.

Чётких правил по выбору размеров train- и test-выборок нет. Рекомендуется брать в каком-то из соотношений $90/10, 80/20, 70/30$.

Гораздо большим значением обладает размер самой выборки. Имеем размеры в $9.7 \cdot 10^4$ и $7.9 \cdot 10^4$. Это выборки со средним размером. Потому искусственно повышать размер train-выборки из-за недостаточности данных необходимости нет.

По принципу Парето возьмём 80/20.

Обучим линейную регрессию. Посмотрим, на качество предсказания для train-выборки, чтобы понять, не переобучилась ли она. Основным показателем качества будем использовать корень из MSE(mean squared error - среднее

квадратичное отклонение):

$$SMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Интерпретировать эту величину можно как среднее отклонение от ответа в секундах, т.е. если SMSE равна 300, то в среднем наша модель «врёт» на 5 минут.

Для линейной регрессии для train- и test-выборок из данных 2019 года соответственно имеем следующие значения:

$$SMSE_{train} = 259;\ SMSE_{test} = 219$$

Аналогично обучим для 2020 года:

$$SMSE_{train} = 195;\ SMSE_{test} = 186$$

В 2020 году получили результаты немного лучше(хотя и совсем чуть-чуть). Причиной может быть бóльшая однородность данных(отсутствие большого количества явных выбросов, в то время как в 2019 году мы наблюдали, например, много коротких поездок).

Интересно также, что качество для тестовой выборки немного лучше, чем у train-выборки. Это может объясняться, например, спецификой данных, непригодностью их для данной задачи, или просто случайностью.

3.4 Обработаем данные перед обучением

Для этого перед обучением обработаем их одним из трёх способов: StandardScaler, MinMaxScaler, MaxAbsScaler.

Подробно останавливаться не будем, т.к. ни один из способов преобразования перед обучением линейной регрессии ничего не дал(действительно были заметны изменения, но порядка одной десятой от ошибки в случайную сторону).

3.5 Полиномиальная регрессия

Т.к. зависимость вряд ли линейная, обучим полиномальную регрессию - это обычная модель линейной регрессии, которая на вход кроме первой степени каждой колонки получает ещё все её степени до некоторой.

После обучения при второй степени всех данных получим следующие результаты:

$$SMSE_{train_{19}} = 133;\ SMSE_{test_{19}} = 104$$

$$SMSE_{train_{20}} = 65;\ SMSE_{test_{20}} = 108$$

Получили заметное повышение качества, однако за эти кроется значительный недостаток такого подхода: при большом количестве данных с увеличеним степени линейно растёт и потребляемая память, что делает затруднительным обучение на больших данных при ограниченных ресурсах. В данном случае, выше второй степени обучить не получилось из-за недостатка оперативной памяти.

3.6 Случайный лес

Random forest(с англ.— «случайный лес») — алгоритм машинного обучения, заключающийся в использовании ансамбля решающих деревьев. Основная идея заключается в использовании большого количества решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим.

Основными параметрами в данной модели являются количество деревьев в лесу и максимальная глубина каждого дерева. Переберём эти параметры в границах $[100; 1000]$ для количества деревьев и $[3; 15]$ для глубины. После долго ожидания(это самый большой минус данной модели) получаем, что наилучшее качество показывает модель с количеством деревьев 500 и глубиной 8.

Результаты данной модели:

$$SMSE_{train_{19}} = 288;\ SMSE_{test_{19}} = 351$$

$$SMSE_{train_{20}} = 281;\ SMSE_{test_{20}} = 414$$

3.7 Композиция моделей

Также хорошим вариантом является попробовать скомбинировать ответы нескольких моделей с какими-то коэффициентами. Однако в нашем случае это не даст никакого улучшения, т.к. качество предсказания случайным лесом в несколько раз хуже, чем у полиномиальной, и даже линейной регрессии.

Чтобы удостовериться в этом, давайте посмотрим на распределения отклонений от ответов для полиномиальной регрессии и случайного леса на тестовой выборке для 2019 и 2020 годов:

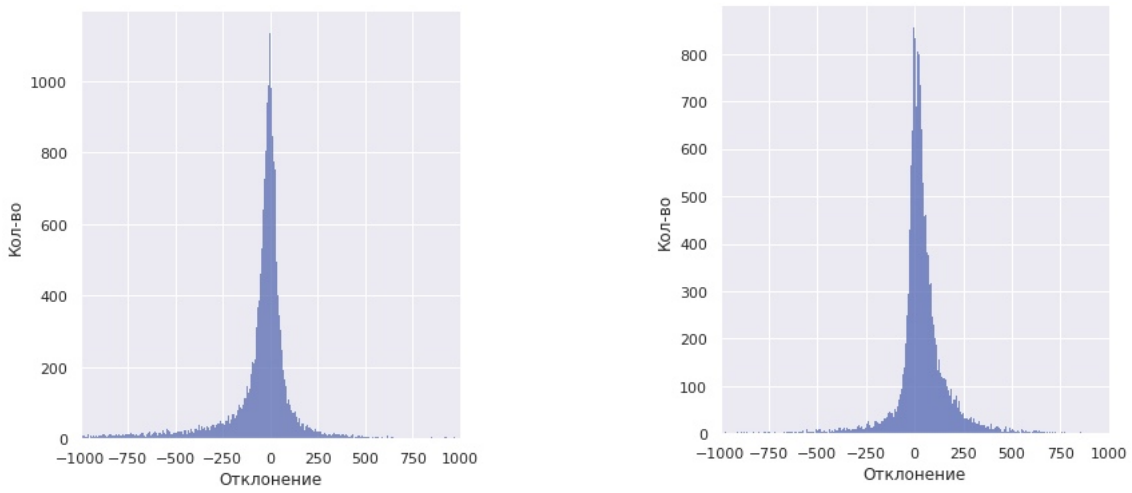


Рисунок 3.1 — Отклонения от ответа для случайного леса: 2019 и 2020 года

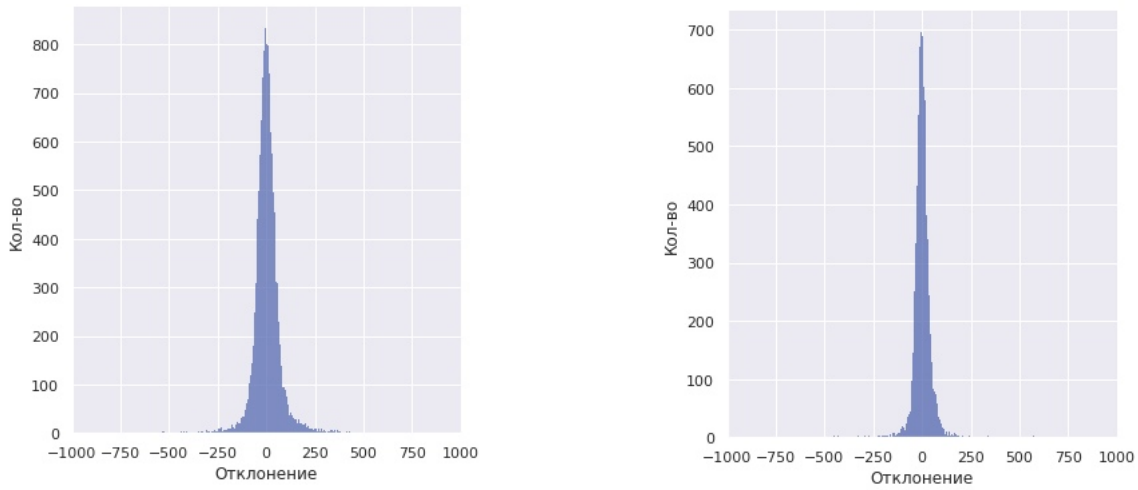
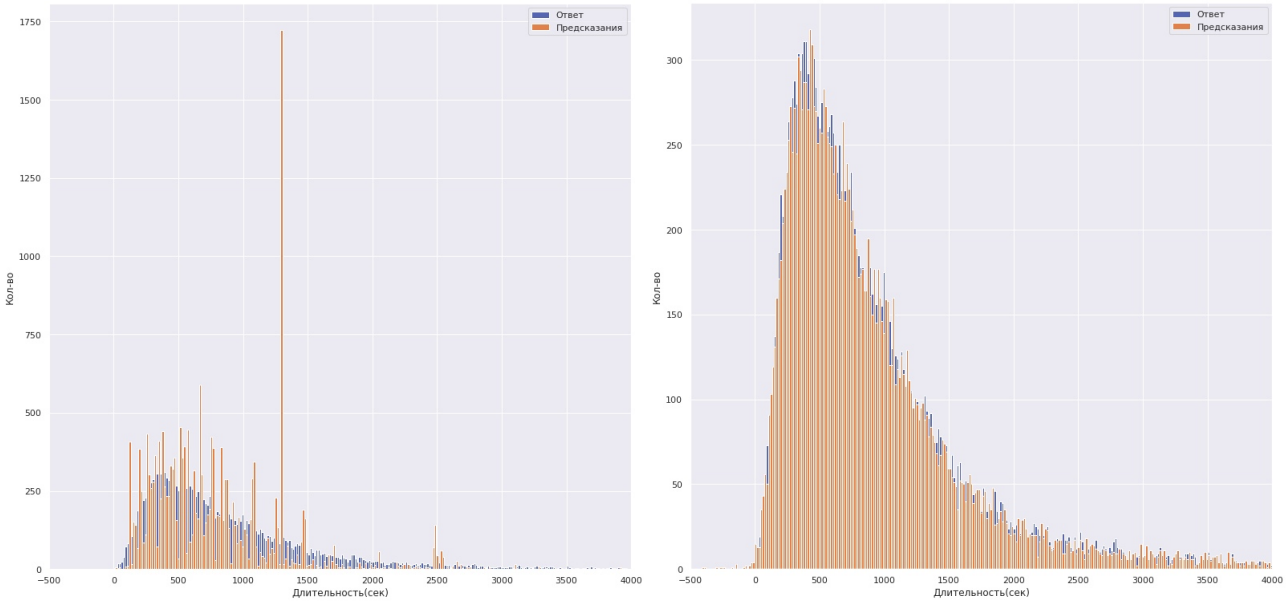


Рисунок 3.2 — Отклонения от ответа для полиномиальной регрессии: 2019 и 2020 года

Как видим, полиномиальная регрессия реже давала абсолютно верный ответ, но все её ошибки являются незначительными, в то время, как у случайного леса существует сильное отклонение.

И в конце посмотрим на распределение самих ответов, например для 2019 года, для случайного леса и полиномиальной регрессии:



Регрессия(справа) очень хорошо «угадала» верное распределение, в то время как случайный лес с этим не справился(и правда случайный).

3.8 Краткие выводы

Как мы видим, у каждой модели существуют свои достоинства и недостатки.

Для линейной регрессии это низкая точность при сложных зависимостях, однако низкое потребление памяти и времени.

Для полиномальной регрессии - способность работать с более сложными зависимостями(требуется лишь подобрать подходящую степень), быстродействие, но высокое потребление памяти.

Для случайного леса плюсов в данной работе выделить не удалось(что, конечно же не означает, что их нет). Минусами является большое время ожидания(до нескольких минут), высокое потребление памяти, если никак не ограничивать деревья.

Заключение

Мы познакомились с базовыми моделями машинного обучения: линейной и полиномиальной регрессиями, случайным лесом. Применили их на задаче предсказания длительности поездки в такси в зависимости от различных факторов. Оценили плюсы и минусы каждой модели. Добились средней ошибки в полторы минуты, что является вполне приемлемым отклонением для пользователя, но, конечно, плохим результатом относительно средней поездки.

Однако в области машинного обучения существует множество ещё неопробованных методов: от базовых(градиентный бустинг) до сложных(различные нейронные сети). Не рассматривались нами также методы регуляризации для отбора признаков, понимания того, как кореллируют признаки друг с другом. Всё это открывает широкое поле для деятельности в будущем.

Список использованных источников

1. Национальный ресурс хранения информации о передвижении такси NYC(taxi limousine comission) [Электронный ресурс] / – Данные, собранные с помощью таксометра во время поездок Taxi, Cabify и Uber. - Режим доступа: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page . – Дата доступа: 16.09.2020.

2. Хенрик Бринк, Джозеф Ричардс, Марк Феверолф. «Машинное обучение» - Издательство Питер, 2017. - С. 69-88.

3. Эндрю Траск. «Грокаем глубокое обучение» - Издательство Питер, 2019. - С. 27-38.

4. Официальная документация по модулю sklearn [Электронный ресурс] / - https://scikit-learn.org/ . - Режим доступа: 18.09.2020-13.12.2020.

5. Интернет-журнал о бизнесе и о России [Электронный ресурс] / - В какой день недели лучше работается и почему. - Режим доступа: https://secretmag.ru/opinions/ponedelnik-kak-socialnyj-konstrukt-chto-ne-tak-s-dnyami-nedeli.htm. - Дата доступа: 22.10.2020.

6. Репозиторий с исходным кодом: https://github.com/dasfex/taxi-trips.