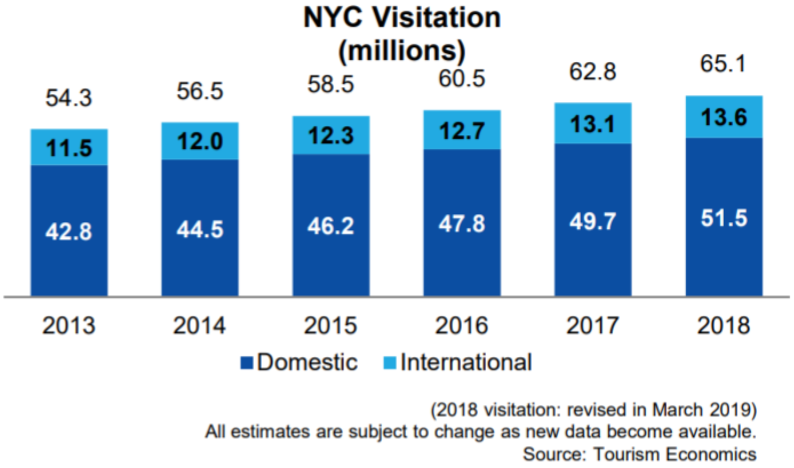
Commuting Solutions for Tourists Visiting NYC

by Vamsi Chinta, Alex Benitez, Faraz Mirza, Syed Kamil

Business Problem:

Over the time period of 2010-2018, NYC's travel volume has increased on average 4.2% per year, rising from 48.8 million to 65.1 million visits per year with an expected record-breaking forecast of 67.1 million visits for 2019[1].



(2018 visitation: revised in March 2019)
All estimates are subject to change as new data become available.
Source: Tourism Economics

FIGURE 1: NYC VISITATION PLOT

This allows for the 'travel and tourism' industry to be a 24/7/365 economic engine for the city[2].

The average minimum length of trip a tourist spends in NYC is 3 days and visits at least two attractions per day[3]; meanwhile, NYC's local population of 8.6 million people are doing their daily commutes. As you can see, this makes for a wide range of efficiency that a tourist can experience during their trip. In fact, NYC has an average one-way commute of 36 minutes[4]. With respect to cost, the average tourist can spend anywhere from 15 to 118$ per day on transportation.[5]

In order to make the most of their visit, it's in the tourist best interest to optimize their commutes around the city in both cost and time. Optimizing their commute time, allows tourists the luxury of enjoying the most attractions and allows them to afford more time to relax between activities. Optimizing their cost, makes their whole trip more affordable and allows tourists to spend more on food, souvenirs, and activities.

The scope of our project will be to help NYC tourists by advising them on how to optimally commute throughout the city with respect to both cost and time.

[1] https://indd.adobe.com/view/fcc4cd9f-7386-4b52-a39b-c401266a137f

[2] https://indd.adobe.com/view/fcc4cd9f-7386-4b52-a39b-c401266a137f

[3] https://indd.adobe.com/view/fcc4cd9f-7386-4b52-a39b-c401266a137f

[4] https://www.nytimes.com/2018/02/22/realestate/commuting-best-worst-cities.html

[5] https://www.budgetyourtrip.com/budgetreportadv.php?geonameid=5128581&countrysearch=&country_code=&categoryid=0&budgettype=1&triptype=0&startdate=&enddate=&travelerno=0

Analytical Framework

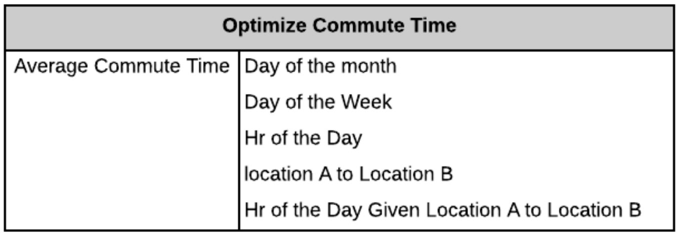
Metrics

Our project's analytical question is a two-part question and is as follows:

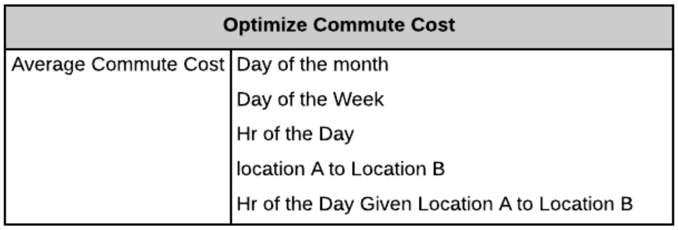
1. what are the possible ride-time's for commuting in NYC
2. what are the possible ride-cost's for commuting in NYC

The unit of measure for cost is average cost per taxi ride, we will use this measure to gage the market's patterns and trends. which is derived from subtracting the tip amounts from the total fare amounts. The time differences between pick-ups and drop-offs.

Each of these objectives listed above can be broken further down into key metrics, each of the which, would allow us to relate the data to the objectives we wish to satisfy. Please see the tables below for the key metrics used.



Optimize Commute Time	
Average Commute Time	Day of the month
	Day of the Week
	Hr of the Day
	location A to Location B
	Hr of the Day Given Location A to Location B



Optimize Commute Cost	
Average Commute Cost	Day of the month
	Day of the Week
	Hr of the Day
	location A to Location B
	Hr of the Day Given Location A to Location B

The data used for this project is the commute data collected on NYC's yellow taxi cabs across time range of April to May 2014. In turn, the data may be biased and may have hidden seasonality trends within it on a yearly basis. It's worth mentioning also that given the dynamics of the emerging competitors in the market, the trends captured may have experienced a concept drift since 2014. However, with there being 29.4 million records within the two months of data, a strong assumption can be made that the insights gained are still apparent, but the degree of the trend may be contingent to the time frame under consideration.

With the data given, we can gauge insights on traffic trends in terms of time and location. And in turn, we'd be able to advise tourists on both when they should either consider or avoid commuting.

Our cost analysis, although biased to yellow taxi cabs, will be able to give us insights on cost trends for rides in terms of time and location. With this, we can also advise tourists on both when and where they should either consider or avoid commuting with respect to cost.

Technology Platform

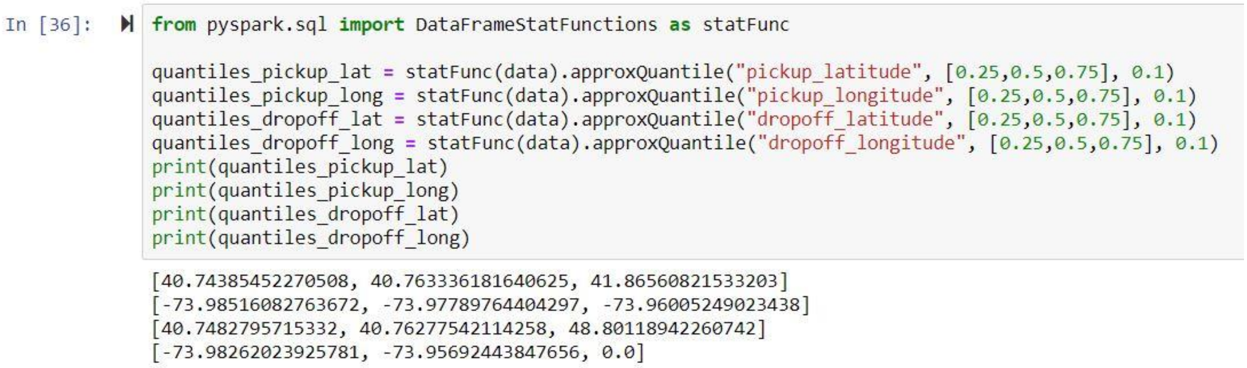
Our Dataset has a total of 29.4 million record and 18 attributes for each record.

Due to the high volume of data, the platform used for this project is Hadoop and PySpark. The GSU cluster runs PySpark 2.2, while quite a step above the original 1.6 version, still lacks some functionality added in later versions. This limited how we interacted with the data and how we were able to extract useful information.

Data Quality Report

feature	**passenger_count**	**trip_distance**	**pickup_longitude**	**pickup_latitude**	**dropoff_longitude**	**dropoff_latitude**	**fare_amount**
count	29.4 Million	29.4 Million	29.4 Million	29.4 Million	29.4 Million	29.4 Million	29.4 Million
mean	1.70	2.97	-72.3407	39.8499	-72.2885	39.8222	$12.74
Std-dev	1.36	3.50	10.8764	5.9957	11.0421	6.0874	$10.47
min	0.00	0.00	-873.9940	-180.0000	-736.4833	-180.0000	$2.50
max	9.00	100.00	112.4355	405.0167	121.8270	404.8833	$500.00

Looking at the Data Quality Report, we could tell there were outliers in latitude and longitude points. We were about to apply IQR outlier removal technique on the data but realized something interesting:



```
In [36]:  from pyspark.sql import DataFrameStatFunctions as statFunc

          quantiles_pickup_lat = statFunc(data).approxQuantile("pickup_latitude", [0.25,0.5,0.75], 0.1)
          quantiles_pickup_long = statFunc(data).approxQuantile("pickup_longitude", [0.25,0.5,0.75], 0.1)
          quantiles_dropoff_lat = statFunc(data).approxQuantile("dropoff_latitude", [0.25,0.5,0.75], 0.1)
          quantiles_dropoff_long = statFunc(data).approxQuantile("dropoff_longitude", [0.25,0.5,0.75], 0.1)
          print(quantiles_pickup_lat)
          print(quantiles_pickup_long)
          print(quantiles_dropoff_lat)
          print(quantiles_dropoff_long)

          [40.74385452270508, 40.763336181640625, 41.86560821533203]
          [-73.98516082763672, -73.97789764404297, -73.96005249023438]
          [40.7482795715332, 40.76277542114258, 48.80118942260742]
          [-73.98262023925781, -73.95692443847656, 0.0]
```

A lot of these points are way too far from New York city. Even further than the neighboring states. There's a lot of zero's as well. IQR technique would not consider them as outliers so we set the boundaries ourselves and filtered them out.

Latitude - Between 36 and 44
Longitude – Between -71 and -76

Our range includes the neighboring states as well since it is possible if people get pickup/drop-off from/to the neighboring states.
After filtering out,
Total Number of Records = 28,689,398 (28.7 Million)
Almost 700,000 invalid points (Noise)

K-Means Clustering

After filtering out the data, we applied k-means clustering on the data. Since the data was big and k-means is computationally heavy, we only ran it from k=10 to 16. There was a huge drop in mean distance on k=12 so we decided to go with that. We applied k-means on pickup points as well as drop off points separately and got 20 unique clusters in total. Each pickup and drop-off point in the dataset was assigned its respective cluster number.

Reverse Geocoding

We used OpenStreetMap API to get the neighborhood names and Zip codes of these cluster centers we got. We mapped them on our dataset according to their cluster number. Once we generalized all our datapoints in these neighborhood regions, we were able to dig deeper into our analysis that can be tailored according to the customer's need. The key metrics we used will be discussed later in the report.

Performance metrics

The performance metrics used are the following:

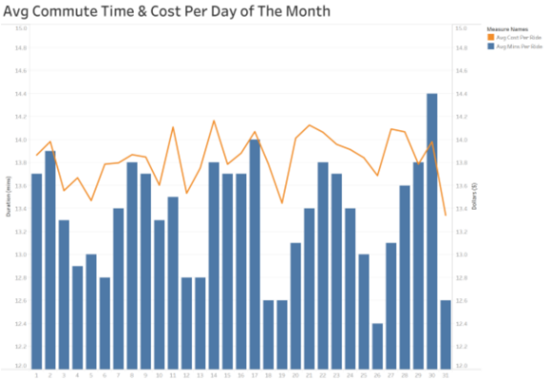
1. Mins/Ride
2. Cost in US$/Ride

Results

Part 1: Descriptive Analytics

To analyze the cost effectiveness, make riding Yellow Taxi by analyzing cost with respect to time. Unit of measure for cost in this case is average cost per ride. The unit of measures for time was day and hour resulting in *'average cost of a ride per day' and 'average cost of a ride per hr'* as our initial set of key metrics.

Metric 1: average cost and time per ride/day of the month

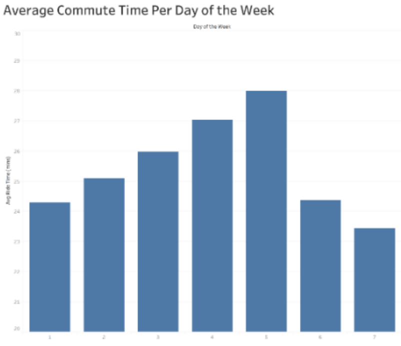


Using both average cost and duration of ride by the day of the month, we can show which days of the months would be optimal.

To get cost of each ride we subtract the tips from the total amount, as tips are voluntary and are subjective to each tourist. This gives us an unbiased fare amount. After which we merged it to our data frame. We then grouped the data by the day of the month.

We then fitted the averages to a histogram to visually view any trends in cost on a day-to-day basis.

Metric 2: average time per ride/day of the week

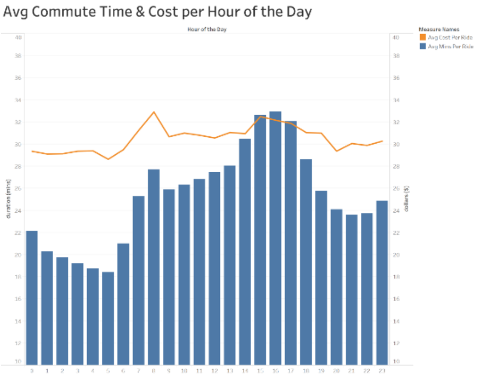


Using average cost and time of ride by day of the week, we can see which days of the week to recommend.

We then took the averages and plotted them on a line graph to visually view any significant differences between the days of the week.

Based on our analysis, we recommend tourists to consider commuting on the weekends because they'd save on average ~10mins per ride.

Metric 3: average cost and time per ride/Hour of the Day



Using average cost of each ride by the hour of the day, we can now gauge which part of the day would be optimal.

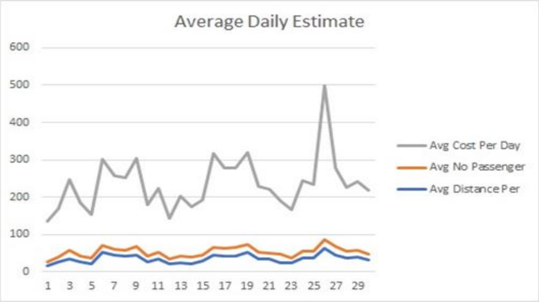
We then fitted the averages to a histogram to visually view any insights on what parts of the day to either consider or avoid traveling during.

We recommend

Part 2: Prescriptive Analytics

Our second part of the analysis focuses on how we perform data mining on the information retrieve in the Part 1 section. This section contains valuable insight that would help tourist to make good decisions with their commuting needs. Here we are providing the best time to travel from certain location we predict best place for Pickup location.

Metric 4: <u>Estimated Average cost and durations</u>



This is the scope of what we are trying to simplify for the tourist. Providing the breakdown of the commuting infrastructure in the city of New York, this detail is categorized by the day of the month and it shows the average cost, distance, and passenger travels on day to day basis. Some of the spike in the data relates to the busiest time when travelling, and in this case, it was related to the holiday that occurs every year in May.

Other aspects that are helpful for customers to identify cost depending on the commuting day, on average passenger from $54 to $65 per ride depending on the distance they covered. This insight also shows the traffic flow from within the city and shows the most cost and time effective day of the month to travel.

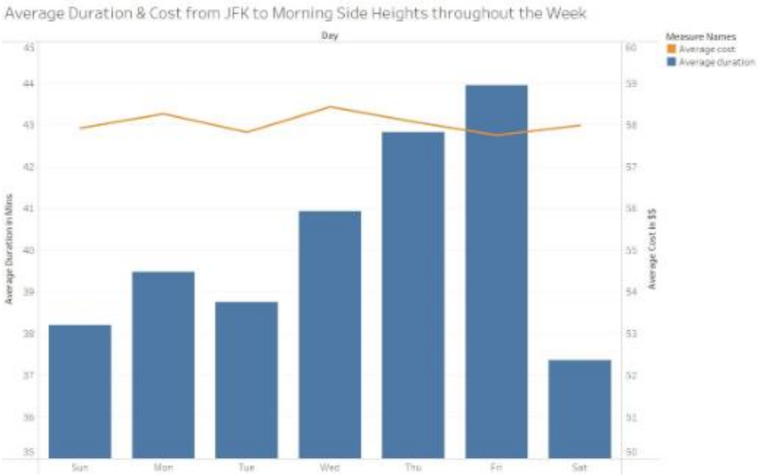
metric 5: <u>Avg cost and time of ride/ day of the week for respective target locations</u>
This metric represents the Average Cost of a Ride based on the pickup and drop off location. By location, we mean the neighborhood we were able to define through k-means, not latitude and longitude points. We have broken down that average cost by day of the week to see if there's a fluctuation of costs throughout the week.

This metric represents the Average Duration of a Ride based on the pickup and drop off location. By location, we mean the neighborhood we were able to define through k-means, not latitude and longitude points. We have broken down that average duration by day of the week to see if there's a fluctuation of average time per ride throughout the week.

metric 6: <u>Avg cost of ride/ the hour of the day for respective target locations</u>
This metric represents the Average Cost and time of a Ride based on the pickup and drop off neighborhood and hour of the day.

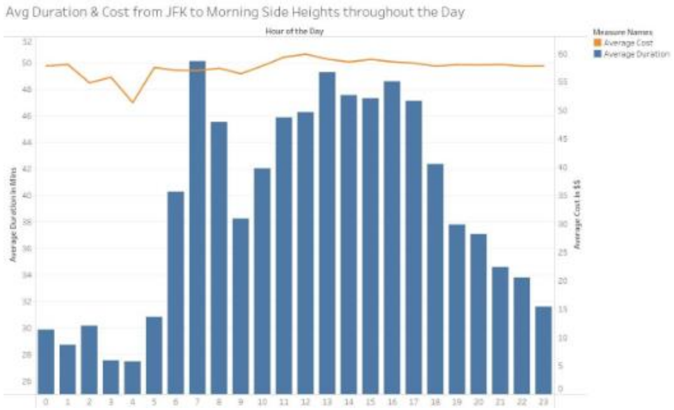
Conclusion

In our conclusion we have provided a case study where it shows how our optimization is going to make a difference in tourist commute and how it is saves them in cost if they decide to take our recommendations.



Case Study 1:

If the tourist decides to travel from JFK to one of the cities in Manhattan. According to our above diagram we have shown if a person decides to take a ride over the weekend (Saturday or Sunday), it will reduce the commute time significantly lower as compare to any other days. Although, if they cannot travel during weekend than one other day is Tuesday where traffic is much more favorable than other weekdays.

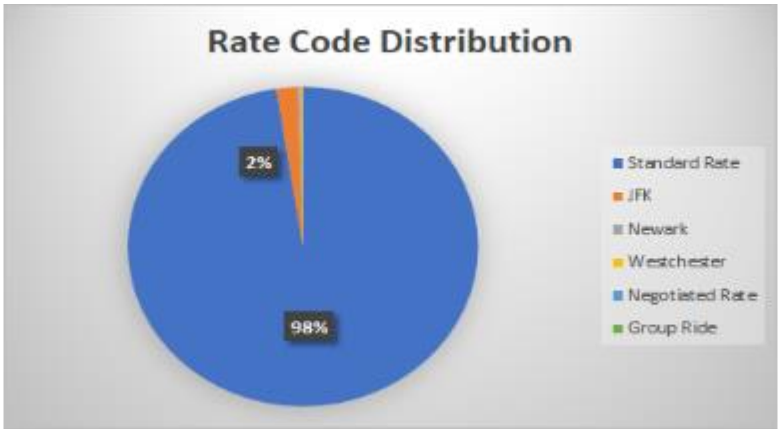


Case Study 2:

We have even taken a step further and broke it down by the hour in the day. Providing a diagram shown above it shows what are the optimal hour for a person to commute from JFK to one of the cities in Manhattan, if they take our recommendations than it will save 24 minutes and almost $4 per commute.

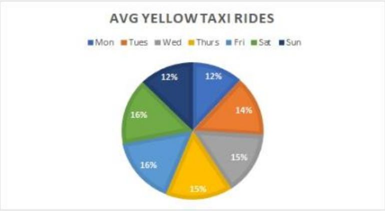
Interesting Facts:

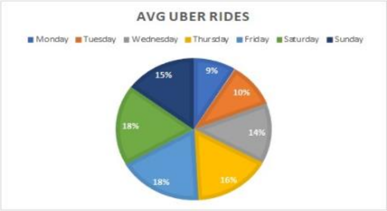
What we learn from the Yellow taxi rides that there are various types of rate codes which determines how the customer pays for the ride. These codes also show that the costs per ride is not constant and it can change depends on the area and the distance between destination. Mentioned below are the commonly used rate codes, and by distribution it seems the most used rate is the Negotiated Rate which shows the 98% of the rides utilize that rate code.



Other Interesting Fact:

Other interesting fact that we have identified how the rides are distributed between the two transportation companies. One is Yellow Taxi and the other is Uber, mentioned below is how the rides are distributed between two companies:





According to the analysis on the left, it shows that Yellow taxi is most popular during the weekdays, but Uber pickup more rides during the weekend.

Concluding Remarks

As you can see, given the data, our analysis can be very impactful for optimizing a trip with regards to commuting efficiently. As well as other insights that could help gauge how to commute. All in all, this project was a huge success and would recommend this business problem to future students.