NYC Taxi Dataset

I would like to thank CBA for the opportunity given to work on this dataset. It has certainly challenged me intellectually. Thanks for the great questions developed by the data science team in CBA.

Before we go deep into the answers, I would like to mention few important points as follow:

- I used Python (Jupyter notebook) and PowerBI for the technical answers to the questions. All EDA graphs found in this document were generated in PowerBI while Python (Jupyter) is used to process the dataset and generates the majority tables found in this document.
- Answers to the distribution questions on the 'Basic Questions' section, particularly question A, B, C, D and E, are using the two original datasets without any pre-processing steps. Limit was applied to some graphs due to a very long tail nature of the data.
- Answers to question G onwards are using the sample (10%) of the population datasets. Sample is pre-processed and few outliers were removed. A sample was taken due to limited computation I possess.
- This work was done using a limited computing I possess (16GB RAM and i5 processor). Thus, data processing and manipulation were done on a sample and few techniques that require higher computing power were not taken. Nevertheless, the steps followed and the given python code should run well with the entire population datasets should computing resources are made available.

Tools Used:

1. PowerBI for all EDA Graphs
2. Python (Jupyter) for data processing and manipulation

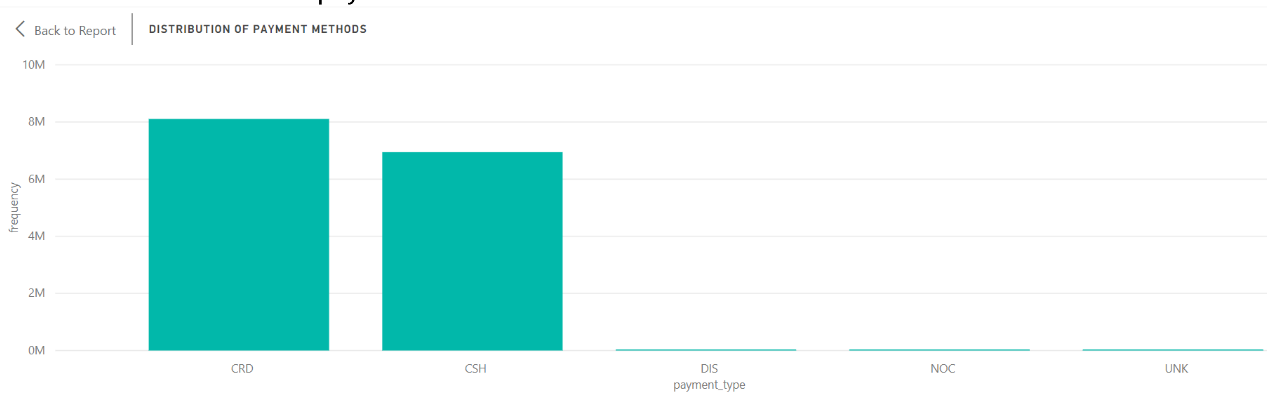
Basic Questions

a) What is the distribution of number of passengers per trip?

I am not sure if the 'per trip' is a typo error as number of passengers <u>per trip</u> can be obtained by looking up the passenger_count column of each record in the trip_data.csv. Plotting the 15 million points in a graph, e.g. scatter/line plot is incredibly huge and it is hard to see if any insights can be extracted. Therefore, since the next 5 questions asking for the frequency distribution, I am assuming that this question asks for the frequency distribution of number of passengers (not per trip) and created a plot as shown below.

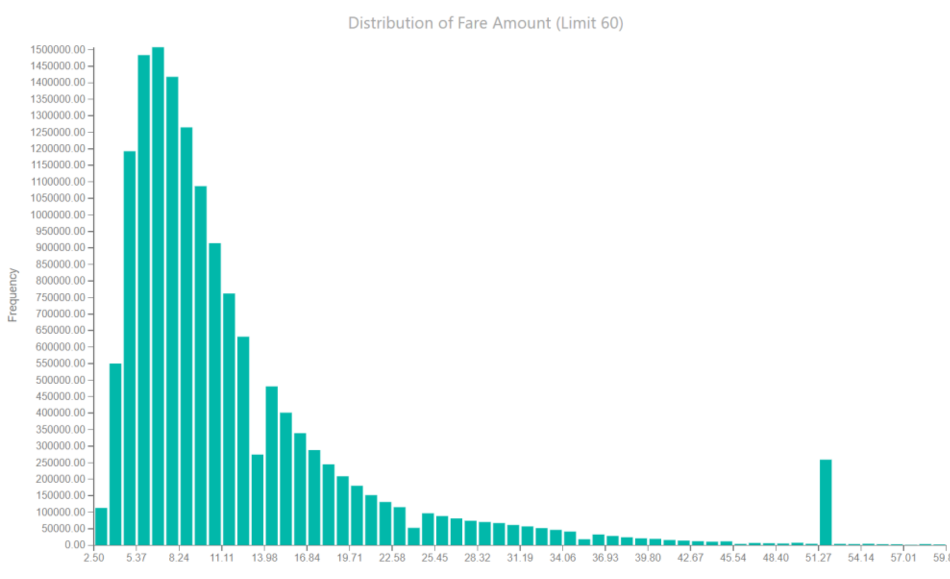
passenger_count	Count of medallion
0	229
1	10707072
2	1985742
3	609849
4	298146
5	890115
6	609313
8	1
9	1

b) What is the distribution of payment_type?
Below is the distribution of payment methods

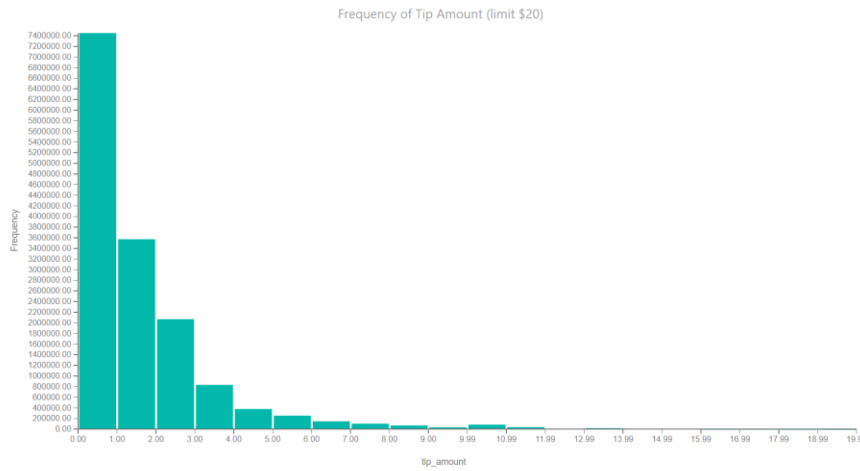


payment_type	Count of medallion
CRD	8105470
CSH	6943669
DIS	11552
NOC	33595
UNK	6182

c) What is the distribution of fare amount?
Below is the distribution of fare amount (right tail is limited to $60)

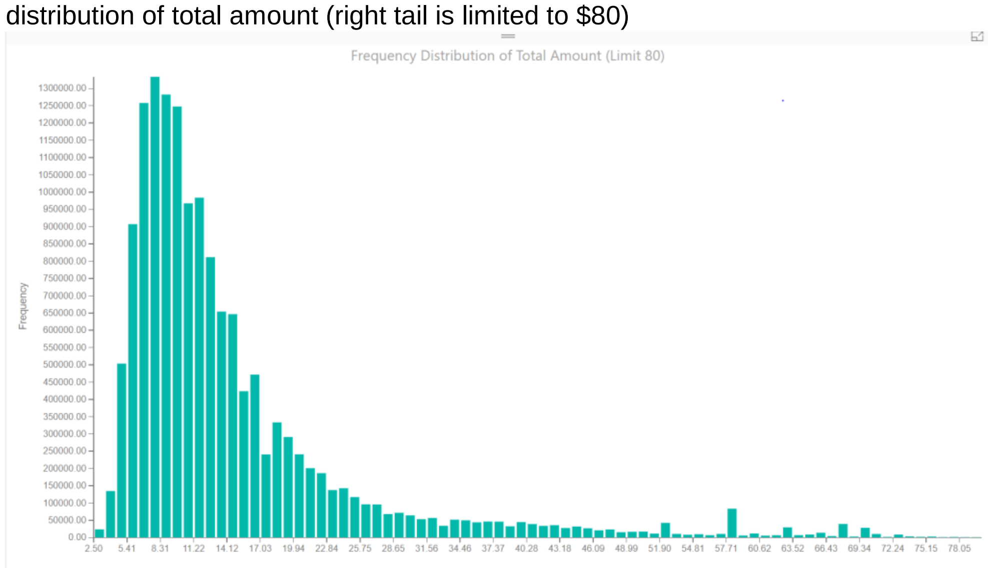


d) What is the distribution of tip amount?
Below is the distribution of tip amount (right tail is limited to $20)



e) What is the distribution of total amount?

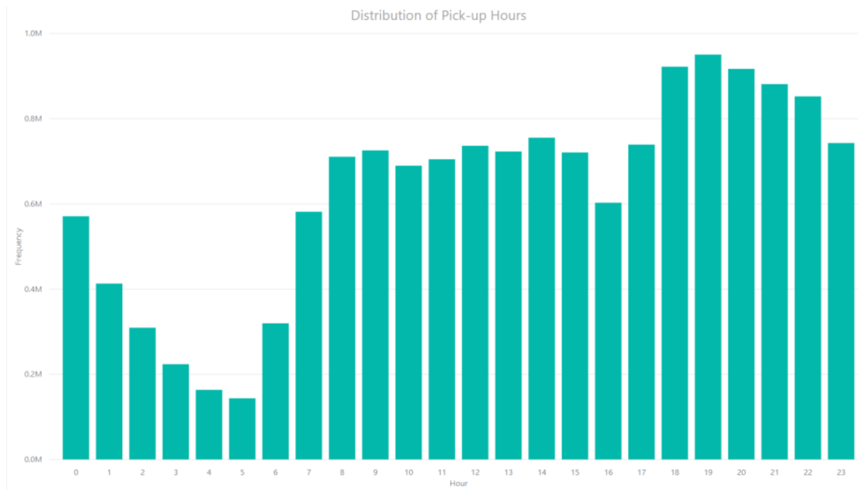
Below is the distribution of total amount (right tail is limited to $80)



f) What are top 5 busiest hours of the day?

As busiest hours were not specified to either pickup or dropoff, I used pickup hours to find the top 5 busiest hours below:

Hour	Count of medallion
19	950590
18	922177
20	917030
21	881281
22	852428



FROM HERE ONWARDS, A SAMPLE OF POPULATION IS TAKEN AND USED TO ANSWER THE QUESTIONS

A sample of the population is taken after data pre-processing due to a limited computing power I have (only a laptop with 16 GB memory). I took 10% sample from the population (1,472,600 records). Sample and population statistics are checked prior sample is used. These statistics are shown in the tables below. Data pre-processing step removes several outliers, such as pickup and dropoff longitude, latitude, trip_distance, trip_time_in_secs, etc.

Population stats

	count	mean	std	min	25%	50%	75%	max
rate_code	14726002.0	1.025439	0.295958	0.000000	1.000000	1.000000	1.000000	210.000000
passenger_count	14726002.0	1.713631	1.389747	1.000000	1.000000	1.000000	2.000000	9.000000
trip_time_in_secs	14726002.0	748.789875	545.889108	1.000000	363.000000	600.000000	960.000000	10800.000000
trip_distance	14726002.0	2.877417	3.312599	0.010000	1.070000	1.800000	3.200000	100.000000
pickup_longitude	14726002.0	-73.975442	0.034058	-74.240349	-73.992317	-73.982025	-73.967972	-73.708878
pickup_latitude	14726002.0	40.750990	0.026868	40.507389	40.736591	40.753448	40.767876	40.908543
dropoff_longitude	14726002.0	-73.974898	0.032766	-74.246284	-73.991615	-73.980499	-73.965118	-73.708870
dropoff_latitude	14726002.0	40.751381	0.030503	40.506039	40.735596	40.753956	40.768570	40.908535
fare_amount	14726002.0	12.161135	9.407520	2.500000	6.500000	9.500000	14.000000	500.000000
surcharge	14726002.0	0.327690	0.367364	0.000000	0.000000	0.000000	0.500000	15.000000
mta_tax	14726002.0	0.498882	0.023618	0.000000	0.500000	0.500000	0.500000	0.500000
tip_amount	14726002.0	1.332470	2.047839	0.000000	0.000000	1.000000	2.000000	200.000000
tolls_amount	14726002.0	0.239898	1.171710	0.000000	0.000000	0.000000	0.000000	20.000000
total_amount	14726002.0	14.560074	11.356654	2.500000	8.000000	11.000000	16.500000	500.000000

Sample stats

	count	mean	std	min	25%	50%	75%	max
rate_code	1472600.0	1.025081	0.279703	0.000000	1.000000	1.000000	1.000000	210.000000
passenger_count	1472600.0	1.713533	1.389870	1.000000	1.000000	1.000000	2.000000	6.000000
trip_time_in_secs	1472600.0	747.973048	545.058888	1.000000	362.000000	600.000000	960.000000	10076.000000
trip_distance	1472600.0	2.871971	3.303559	0.010000	1.070000	1.800000	3.200000	100.000000
pickup_longitude	1472600.0	-73.975472	0.033967	-74.234375	-73.992310	-73.982010	-73.967979	-73.709251
pickup_latitude	1472600.0	40.751010	0.026839	40.508209	40.736656	40.753452	40.767895	40.908157
dropoff_longitude	1472600.0	-73.974915	0.032692	-74.241127	-73.991615	-73.980499	-73.965141	-73.708923
dropoff_latitude	1472600.0	40.751400	0.030530	40.506039	40.735622	40.753941	40.768589	40.908260
fare_amount	1472600.0	12.141963	9.377005	2.500000	6.500000	9.500000	14.000000	300.000000
surcharge	1472600.0	0.327341	0.367248	0.000000	0.000000	0.000000	0.500000	3.000000
mta_tax	1472600.0	0.498914	0.023279	0.000000	0.500000	0.500000	0.500000	0.500000
tip_amount	1472600.0	1.327586	2.035957	0.000000	0.000000	1.000000	2.000000	134.250000
tolls_amount	1472600.0	0.238088	1.165869	0.000000	0.000000	0.000000	0.000000	20.000000
total_amount	1472600.0	14.533893	11.308021	3.000000	8.000000	11.000000	16.250000	330.000000

g) What are the top 10 busiest locations of the city?

In order to obtain NYC neighbour locations for both pickup and dropoff coordinates, I downloaded NYC neighbours coordinates from this link: https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23. I then use the equirectangular distance formula for each pickup and dropoff coordinates to find the distance to the 299 neighbours. The neighbour location for pickup or dropoff is the one that has the shortest distance.

As the question does not mention whether it refers to pickup or dropoff locations as the busiest locations, I take both top 10 busiest locations for pickup and dropoff. Below are the top 10 busiest locations for pickup from sample data:

	pickup_neighbor	cnt_top_pickup_neighbor
138	Midtown	130154
219	Sutton Place	97402
119	Lincoln Square	96130
139	Midtown South	92491
77	Flatiron	82503
46	Clinton	70747
39	Chelsea	66913
149	Murray Hill	58854
118	Lenox Hill	58789
229	Upper West Side	51168

Below are the top 10 busiest locations for dropoff from sample data:

	dropoff_neighbor	cnt_top_dropoff_neighbor
164	Midtown	137893
258	Sutton Place	88370
143	Lincoln Square	88046
165	Midtown South	85703
92	Flatiron	71093
53	Clinton	66590
45	Chelsea	57916
141	Lenox Hill	56069
176	Murray Hill	55282
270	Upper West Side	53380

h) Which trip has the highest standard deviation of travel time?

The trip with highest standard deviation of travel time is from Battery Park City to Sunset Park with standard deviation (in secs) of 3356.917 within the sample.

	pickup_neighbor	dropoff_neighbor	stdev_travel_time
414	Battery Park City	Sunset Park	3356.917157
5157	Greenwich Village	Bayside	3224.406922
11346	Turtle Bay	East Tremont	2892.066735
832	Briarwood	Tudor City	2628.589403
1264	Bushwick	Maspeth	2388.606707
9938	South Ozone Park	Springfield Gardens	2288.185377
4455	Forest Hills	Midtown	2167.313650
9702	South Corona	Springfield Gardens	2163.746750
5889	Kew Gardens Hills	Clinton	2078.460969

i) Which trip has most consistent fares?

Data/values consistency is normally achieved by the lowest or zero standard deviation. Thus, consistent fares means trips that has zero or lowest standard deviation. There are numerous trips with zero (0) standard deviation in the data. However, this trip has low number of occurrences. To answer this question, I count the number of trips made between locations and obtain a trip with a zero standard deviation and the highest number of occurrences. This trip is from Lincoln Square (pickup location) to Springfield Gardens (dropoff location) with total occurrences of 252 trips and fare amount of $52 within the sample.

	pickup_neighbor	dropoff_neighbor	total_trips	avg_fare	stdev_fare_amount
6350	Lincoln Square	Springfield Gardens	252	52.0	0.0
8104	Murray Hill	Springfield Gardens	186	52.0	0.0
10218	Springfield Gardens	Manhattan Valley	146	52.0	0.0
11817	Upper West Side	Springfield Gardens	121	52.0	0.0
4996	Gramercy	South Ozone Park	113	52.0	0.0
1893	Chelsea	Springfield Gardens	100	52.0	0.0
7441	Midtown South	Brookville	99	52.0	0.0
11647	Upper East Side	South Ozone Park	89	52.0	0.0

Open Ended Questions

a) **In what trips can you confidently use respective means as measures of central tendency to estimate fare, time taken, etc.**

Means can be used as measures of central tendency for estimation purpose if the trip has zero or low standard deviation. The following trips from the sample can use means as measures of central tendency to estimate fare and time taken:

	pickup_neighbor	dropoff_neighbor	total_trips	stdev_fare_amount	stdev_travel_time
307	Battery Park City	Bensonhurst	2	0.0	0.0
558	Bensonhurst	Bath Beach	3	0.0	0.0
2442	Clinton	Schuylerville	2	0.0	0.0
5506	High Bridge	Riverdale	2	0.0	0.0

However, if other amount types, such as surcharge, tax, tip and toll are taken into account for estimation, only the following trips in the sample that can use means as measures of central tendency:

	pickup_neighbor	dropoff_neighbor	total_trips	stdev_fare_amount	stdev_travel_time	stdev_surcharge	stdev_tax	stdev_tip	stdev_toll
558	Bensonhurst	Bath Beach	3	0.0	0.0	0.0	0.0	0.0	0.0
2442	Clinton	Schuylerville	2	0.0	0.0	0.0	0.0	0.0	0.0

b) **Can we build a model to predict fare and tip amount given pick up and drop off coordinates, time of day and week?**

This question is a bit ambiguous in a sense that the data given for the model are just pick-up & drop off coordinates, time of day and week. If only three types of data given, it will be impossible to build a model as there is no target labels to train, i.e. past fare and trip amount.

However, I assume that past fare and trip amount are available. In this regard, Yes, we can build a model. These are the two approaches that I could think of and examples of both approaches are implemented in the Jupyter notebook:

1. Regression Method
   From pickup and dropoff coordinates, I can obtain their neighbour locations using the technique mentioned in Basic Question G. I have nine independent variables (i.e. pickup longitude, pickup latitude, dropoff longitude, dropoff latitude, pickup neighbour, dropoff neighbour, week, pickup hour, dropoff hour). LabelEncoder or One hot encoding can be applied to the categorical variables such as pickup/dropoff neighbors and week while StandardScaler can be used to normalise the numerical variables. The dependent variables are fare_amount and tip_amount. I created two regression models, one for predicting the fare_amount and the other for predicting the tip_amount. Dataset was split into training and testing sets where the training set was used to train regression algorithms and testing set was used to evaluate the models based on the selected performance metrics. Regression algorithms such as linear regression, SVR, Random Forest Regression, etc. can be experimented. Performance metrics such as MSE, R2, etc. can be used to evaluate the models.

2. Clustering Method
   I first clustered the pickup and dropoff coordinates, pickup and dropoff neighbors, time and week. K-means cluster algorithm was used for the clustering and cluster evaluation methods such as elbow method, silhouette index, AIC can be used to find the best k clusters. For each cluster, I found the highest and lowest price from all points belong to the cluster together with its mean, median and standard deviation. Cluster profiles the data points and gives an indicator of the mean and median price points that belong to the cluster. In prediction stage, each data point can belong to the cluster and the price can be estimated from the cluster mean/median price.

The implementation example of these methods can be found in the jupyter notebook.

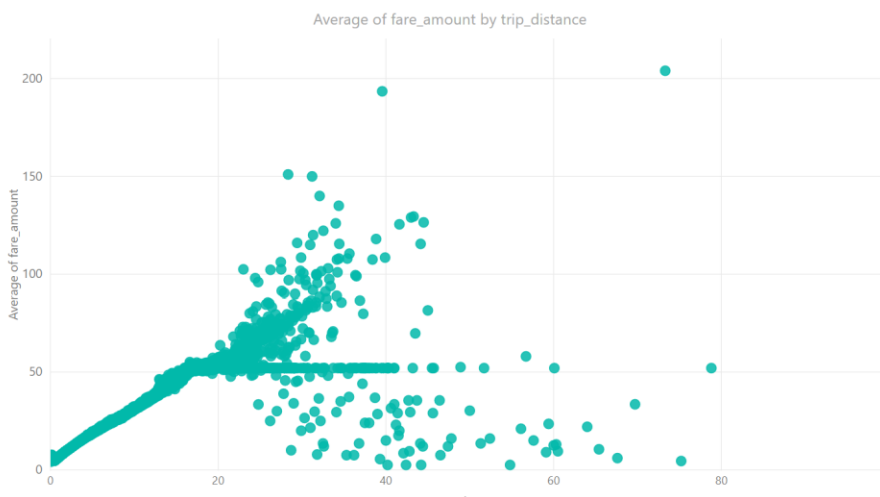
c) If you were a taxi owner, how would you maximize your earnings in a day?

Let's look at the definition of earning as follow:

Earning = Amount Received – Costs

The total amount received in a trip consists of various types of amount (i.e. tools, tip, surcharges, etc.). Some of these amount types can be claimed by the taxi owner (i.e. fare amount) while others (e.g. tax, toll charge, etc.) need to be paid to the providers. I will make a common assumption here that the toll charge and tax will be levied and paid by the customer(s) on each taxi trip and, therefore, are not considered as costs nor earnings to the taxi owner. Surcharge is a bit different in a sense that some surcharges such as credit card, airport, etc. will be paid to the service providers but others such as late night surcharge, etc. can be claimed by the taxi owner. However, without data about the type of surcharge made available or further information from the business department, it is hard to differentiate between the claimable and non-claimable surcharges. Therefore, I make another assumption that the surcharge is non-claimable by the taxi owner and will be charged to the customer(s). *This left us with fare amount and tax amount as claimable earnings*.

On the other hand, the cost of each trip is from petrol. Petrol consumption depends on the distance and time taken for each trip. My hunch and experience tell me that there are correlations between fare amount and trip distance and between fare amount and trip time. So, I plot these graphs from the sample data as seen below.

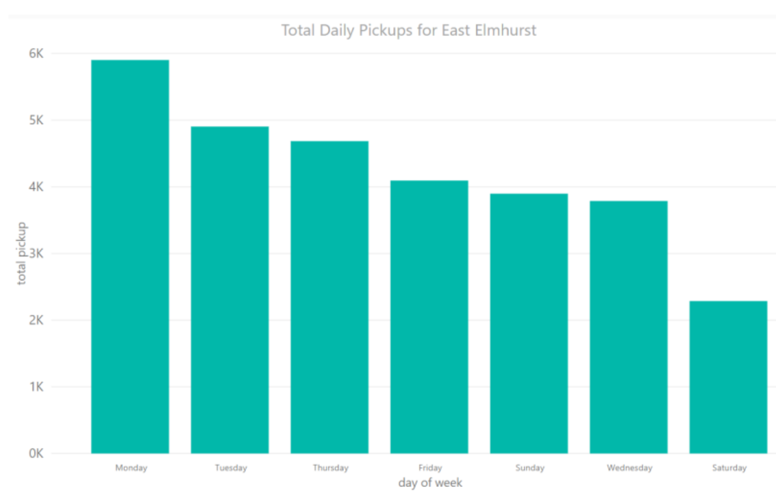


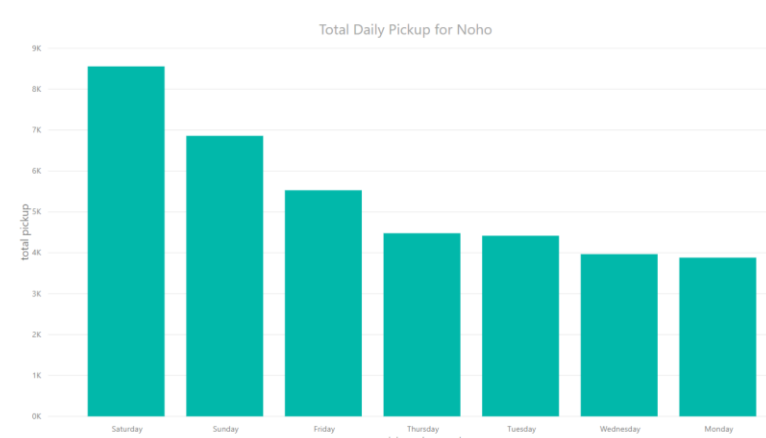
We can clearly see that the correlations exist which means fare considers both trip distance and trip time. This implies that fare has already taken into account the cost of petrol consumption no matter how long or short or how far or short the trip is. *As the trip duration and length are considered in fare amount, we should not worry about their effect on the petrol cost*. Also, note that as data for daily fuel price is not available, we won't be able to calculate the petrol cost. Therefore, I make another assumption that daily petrol cost remains static and that the higher the fare the more earnings the taxi owner will get after subtracting petrol cost from the fare.

Considering the facts and assumptions above, our goal must be *obtaining the highest fare and tip amount*. Here are few strategies that I can follow to maximise my earnings as a taxi owner in a day:

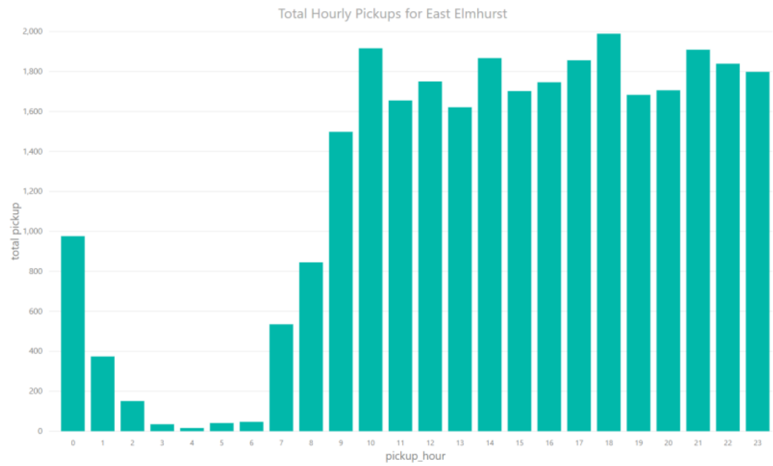
1. I will first try to go to the busiest pickup locations shown in Basic Question G. Midtown, for example, is one of the highest pickups of all days. This ensure that I always get passengers and not sitting idle. However, busiest locations are normally high in supply too as most taxi owners will go there, which means more competitions to get customers. Unfortunately, taxi supply data is not made available. In case the supply in such locations is too high, I will go to less busiest locations while considering the demand for such location during day of the week and time of the day as discussed in the next points.

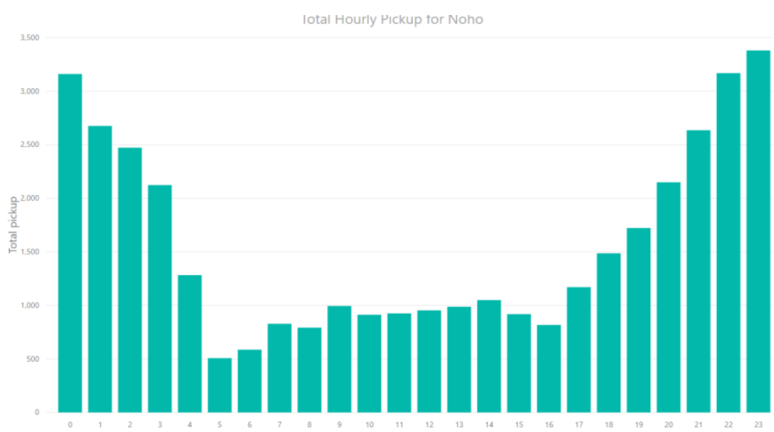
2. Pickup demand for each location differs based on day of the week. I will go to the places with high pickups on specific days. For example, I will prefer to go East Elmhurst on Tuesday and Monday rather than on Saturday and Sunday as pickup demand for Saturday and Sunday for this location are much lower. On the other hand, I will go to Noho for Saturday and Sunday as these days have the highest pickups.





3. The pickup demand for each location also differs based on time. I will go to places with high pickup demands on specific hour. For example, pickup demand in Noho peaks during morning and afternoon (05:00-16:00) and lower in midnight (20:00 – 03:00). Equipped with this information, I will bring my taxi to Noho at midnight and early morning while I will go to East Elmhurst in the morning till afternoon. This ensures I always go to the busiest hours for each location and receive orders.





4. The other strategy I can follow is by aiming to pick up passengers in the locations where my earnings can be maximised. As earning consists of fare and tip amount, I can find the average earning amount per trip for each pickup neighbour. I will also add a condition that the pickup locations must have at least 10000 total trips monthly to increase my probability of getting the passengers. The table below shows the top 10 pick up locations, total trips and average earnings per trip. The location examples we used above, i.e. East Elmhurst and Noho, are in the top 10 earnings per trip. I can also consider other locations listed in the table below and complement this information with strategy # 2 and 3 above to maximise my earnings in a day.

Table: Top 10 locations for highest earning per trip (fare + tip)

	pickup_neighbor	trip_count	sum_fare_tip_amount	fare_tip_per_trip
209	Springfield Gardens	10837	541200.42	49.940059
62	East Elmhurst	29555	1014848.84	34.337636
75	Financial District	24085	431884.98	17.931699
7	Battery Park City	12664	221031.99	17.453568
223	Tribeca	14342	205284.06	14.313489
43	Civic Center	18436	261319.76	14.174428
142	Morningside Heights	13417	189081.80	14.092703
200	Soho	18470	245084.57	13.269332
121	Little Italy	13878	183149.88	13.197138
156	Noho	37690	496292.31	13.167745

d) If you were a taxi owner, how would you minimize your work time while retaining the average wages earned by a typical taxi in the dataset?

To minimize my work time while retaining the average wages, these are several strategies I could take:

1. Go to the most pickup and dropoff locations. The sample dataset shows that all top 10 locations for pickup and dropoff are same neighbors. By going to these neighbors, it increases the chances of getting another pickup after dropping off passengers. For example, the most popular pickup neighbour, Midtown, has 9 of the top dropoff locations in the top 10 pickup locations (see table below). This ensures that I keep getting passengers after dropoff and not idling around.

Table: Top 10 popular dropoff locations for Midtown and their pickup counts

	dropoff_neighbor	dropoff_count	pickup_count
0	Midtown	14912	130154
3	Sutton Place	9066	97402
2	Lincoln Square	9921	96130
1	Midtown South	12688	92491
5	Flatiron	6657	82503
4	Clinton	8103	70747
8	Chelsea	4615	66913
6	Murray Hill	4928	58854
7	Lenox Hill	4690	58789
10	Upper West Side	4161	51168

2. Go to the pickup locations that normally have the shortest average trip time and healthy total pickup numbers (>10000 per month). Locations such as Murray Hill & Tudor City have healthy total pickup numbers and their average time taken on the trip is around 10 mins with average fare and tip of $11.58 and $12.34 respectively. This allows me to quickly go back to the pickup locations and get on another trip.

Table: Pickup neighbors with their average fare+tip and time taken
sorted by average time taken ascending

	pickup_neighbor	total_trips	avg_fare_tip	stdev_fare_tip	avg_time_in_secs	stdev_time_in_secs
91	Gramercy	40124	11.452796	7.760670	638.525446	423.104239
149	Murray Hill	58854	11.588127	8.555972	646.006474	430.090663
228	Upper East Side	47283	10.935765	7.217209	652.017469	466.682847
219	Sutton Place	97402	11.327060	7.823062	658.755652	445.557007
118	Lenox Hill	58789	11.294900	7.403346	658.912943	451.940626
34	Carnegie Hill	46663	11.541641	7.757064	661.236033	475.673903
224	Tudor City	14698	12.340395	9.265064	661.851136	420.156576
229	Upper West Side	51168	11.669159	8.182454	662.529902	492.994910
130	Manhattan Valley	23498	12.183437	8.768816	664.815857	500.486192

Note that this strategy can be combined with Strategy # 2 (days of the week) and 3 (time hour) of the Open Question C above so that I could go to these neighbors on the right day and time.

3. Tip contributes to a big proportion of earning. If I would like to earn average wages while minimising my work time, I will need to find pickup locations where passengers offer big tip and often giving tips. Location such as East Elmhurst has a healthy tip ratio of 0.61 and its average tip amount of $6.596 is amongst the highest. Battery Park City is another location with a healthy tip ratio (0.63) and average high tip amount ($3.088).

Table: Pickup neighbors with their tip ratio and average tip amount above 10k trips sorted by average tip ratio ascending

	pickup_neighbor	pickup_count	sum_tips	tip_count	tip_ratio	avg_tip	stdev_tip
201	Tribeca	14342	23408.51	9293	0.647957	2.518940	1.963471
6	Battery Park City	12664	24523.49	7942	0.627132	3.087823	2.400272
57	East Elmhurst	29555	118066.74	17899	0.605617	6.596276	3.818021
213	West Village	42548	55122.23	24944	0.586256	2.209839	1.752996
69	Financial District	24085	45145.15	14032	0.582603	3.217300	2.384353
71	Flatiron	82503	102589.62	47858	0.580076	2.143625	1.642663
145	Noho	37690	51792.21	21809	0.578642	2.374809	1.747565

e) If you run a taxi company with 10 taxis, how would you maximize your earnings?

If I am running a taxi company with 10 taxis, I will spread out the 10 taxis to various busy neighbors following strategies listed in Open Question C and D. The distribution will follow the daily and hourly demand of each neighbour. During the busy peak hours period between 6-11pm, multiple taxis will be allocated to the downtown and financial district areas where the demand are the highest.

In non-peak hours, I will allocate my taxis to several healthy pickup locations that have high tips ratio and amount. For weekend and midnight shift, high demand locations such as Noho are the best to allocate the taxis as the pickup demand for these locations is at their highest.

Part 2: Open Showcase

Beside working full time as a Senior (Lead) Data Scientist at SAP, I am also holding an adjunct lecturer position (non-paid & part-time/casual position) at the University of Newcastle where I am currently supervising two PhD students in Data Science/Machine Learning. I have been active in the academic community publishing and presenting my ideas through academic papers that relate to the areas. This shows a continued contribution to the area of data science and machine learning.

Below are selected papers that I published recently with my students and co-contributors:

1. Predict Polarity of Review Text using NLP and Machine Learning
   Paper reference: Satia-Budhi, G., Chiong R., Pranata I., and Hu, Z., "Predicting rating polarity through automatic classification of review texts", IEEE Conference of Big Data and Analytics (ICBDA 2017), 2017.

   This research looked into automatic polarity classification of reviews using several machine learning techniques. Several Natural Language Processing (NLP) techniques such as bag of words, n-gram, TF-IDF were used in conjunction with the solution proposed. Millions of review text of Yelp! Were used and experimented to validate the proposed solutions. The objective is to train machine to understand the review text and provide polarity estimation of such review.

2. Measuring the Trustworthiness of Most Popular Users with Data Science.
   Paper Reference: I. Pranata, W. Susilo, "Are the most popular users always trustworthy? The case of Yelp", Electronic Commerce Research and Applications, vol. 20, pp. 30-41, 2016.

   This research investigated the trustworthiness of the most popular users in a popular review website, Yelp! I look into a cluster of the most popular users in Yelp! And use data science and statistic methods to review their trustworthiness in giving credible reviews

3. Automatic identification of malicious web domains through their online credibility and performance data.
   Paper reference: Hu, Z., Pranata, I. and Chiong R., "Identifying malicious web domains using machine learning techniques with online credibility and performance data", IEEE Congress on Evolutionary Computing (IEEE CEC 2016), Vancouver Canada 2016.

   This research looks into providing automatic credibility score of web domains through open and accessible web data. Open data of hundreds of malicious and valid web domains were scourged and pre-processed into several machine learning algorithms. The objective is to provide automatic identification and credibility score of malicious web domains. Open data includes SEO, Alexa, search rankings, load performance and many more

4. Experiment research on the effectiveness of targeted e-health program for weight loss
   Paper reference: MJ Hutchesson, PJ Morgan, R Callister, I Pranata, G Skinner, CE Collins, "Be positive Be health e: development and implementation of a targeted e-health weight loss program for young women" Journal of Telemedicine and e-Health, 2016

   This is a health related research where we developed a targeted e-health programme and test our hypothesis through a 6 months real experiment with participants who are young Australian women. Control and target groups were created in the 6 months research and several statistics and data science methods were used to validate results and hypothesis from the control and target groups.

There are few other research papers in the area that are either still under consideration for publications or waiting to publish. All these publications and work that I and my students are currently working on show the strong contribution that I have in advancing the data science and machine learning research.