Declaration on Plagiarism

Names:	Khilti Dedhia, Romil Sakariya (Pair K)
Student Numbers:	21264200, 21264095
Programme:	MSc in Computing
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	26th November 2021
Module Coordinator:	Dr Suzanne Little

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should we engage in plagiarism, collusion or copying. We have read and understood the Assignment Regulations. We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.

We have read and understood the referencing guidelines found at
http://www.dcu.ie/info/regulations/plagiarism.shtml,
https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Khilti Dedhia Date: 26th November 2021

Name: Romil Sakariya Date: 26th November 2021

A New York City Cabbie's guide to maximizing their income!

Abstract

As an NYC cab driver, incoming revenue from rides can vary substantially from one day to another. There are some locations that offer a higher rate of fares such as densely populated residential areas or office areas. But are there some specific timeframes that cab drivers can exploit in order to get quicker fares or larger tips? We successfully answer this very question by analysing the trends of fares and tips across 6 months, including the holiday season (i.e., December). Upon visually interpreting the data, we conclusively narrow down time frames where a cab driver should expect larger fares and tips. Unfortunately, we also uncover the fact that the holiday spirit does not extend towards higher tips for cab drivers in NYC. Whilst analysing this data we also come across some unexpected findings that we try to justify with some logical presumptions.

Dataset(s)

TLC (Taxi & Limousine Commission) publishes their yellow and green taxi trip records for each month of the year since 2009. To retrieve the trip records of our desired months (October 2020 to March 2021), we simply downloaded each month's trip records in a csv file from their website: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

The following table describes the size of each of the 6 batch files and one final combined dataset that we used for this project:

Filename	Number of trip records (rows)	File size
yellow_tripdata_2020-10.csv	1,681,131	147 MB
yellow_tripdata_2020-11.csv	1,508,985	132 MB
yellow_tripdata_2020-12.csv	1,461,897	128 MB
yellow_tripdata_2021-01.csv	1,369,765	120 MB
yellow_tripdata_2021-02.csv	1,371,708	120 MB
yellow_tripdata_2021-03.csv	1,925,152	169 MB
Final combined dataset	93,18,638	995 MB

The following table describes the attributes, data types and classification of data of each CSV file:

Attributes	Data Type	Classification of data
VendorID	float64	Miscellaneous data
tpep_pickup_datetime	object	Temporal Data
tpep_dropoff_datetime	object	
passenger_count	float64	Numerical count of passengers
trip_distance	float64	Distance based data
store_and_fwd_flag	object	Miscellaneous data
PULocationID	int64	Location based data
DOLocationID	int64	

Attributes	Data Type	Classification of data
RatecodeID	float64	Financial data
payment_type	float64	
fare_amount	float64	
extra	float64	
mta_tax	float64	
tip_amount	float64	
tolls_amount	float64	
improvement_surcharge	float64	
total_amount	float64	
congestion_surcharge	float64	

Our integrated dataset contains the 'Volume' one element of big data. For our analysis we would be using the large number of rows in the dataset while selectively using only the financial data mapped alongside the temporal data across all rides in our selected six months.

Data Integration, Cleaning, Processing, and Integration

Data Integration

Once we obtained all the 6 datasets, we integrated the files into one big final dataset that we used for further analysis using Pandas. We performed this data integration process on Google Colab where we first loaded all the 6 sub datasets and combined them into one final dataset. Once we obtained the integrated dataset, we moved on to clean and process it.

Data Cleaning

Immediately after integrating the datasets, we analysed the final dataset and interpreted that the data needed thorough cleaning. We carried out the following data cleaning operations:

1. Checked and removed NULL entries.
2. Checked and removed any duplicate entries.
3. Removed entries where trip distance was 0.
4. Removed entries where passenger counts were 0, and more than or equal to 6. (Since yellow taxis in New York has maximum 5 person's capacity)
5. Removed entries where the fare amount and total amount was 0.
6. Checked and removed entries from any other months that lie outside our 6 months range.

Before cleaning our data, the length of our dataset was 9,318,638. Once cleaned, it dropped down by approximately 10% to 8394145.

Data Processing

Once the dataset was cleaned, we noticed that we needed to perform some data processing steps in order to efficiently explore and visualize the data. We introduced the following attributes in order to work on the data effortlessly:

1. An 'id' attribute, so that every trip will have its unique identifier.
2. Converted the data type of pickup datetime attribute from 'object' to 'datetime'.

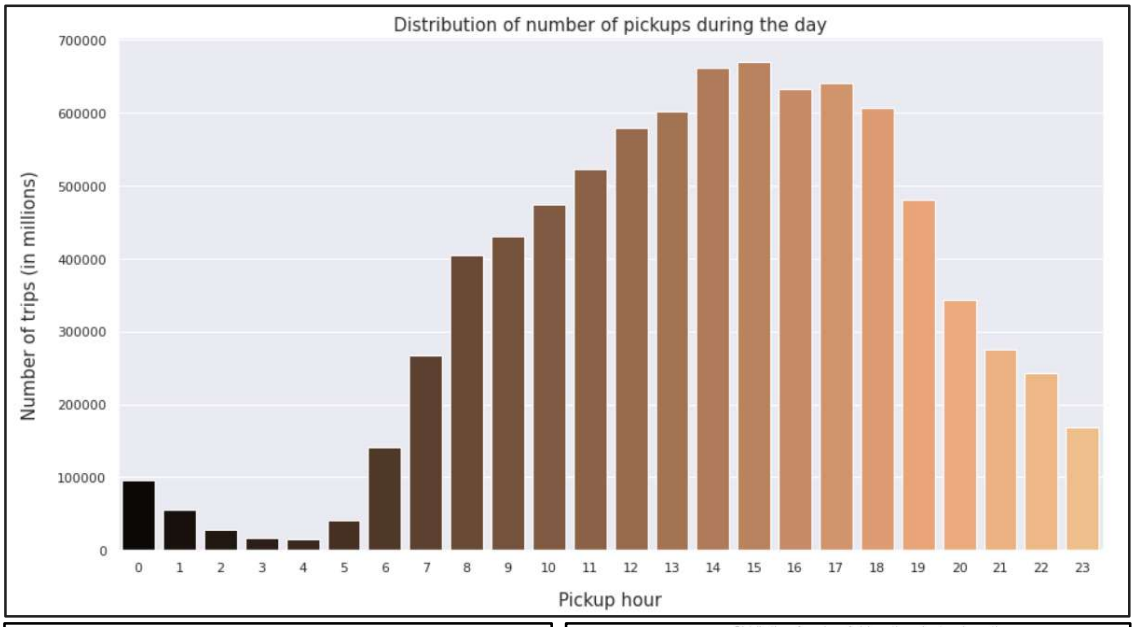
3. Extracted pickup month, pickup day, and pickup hour attributes from the pickup datetime attribute.
4. Calculated tip percentage - simply using tip amount would not be logical since it depends on the fare amount itself. In order to sufficiently normalize the comparison parameter across all fares in all months, we calculated tip percentage.

By performing these steps, we could explore data in a much better fashion.

Data Exploration

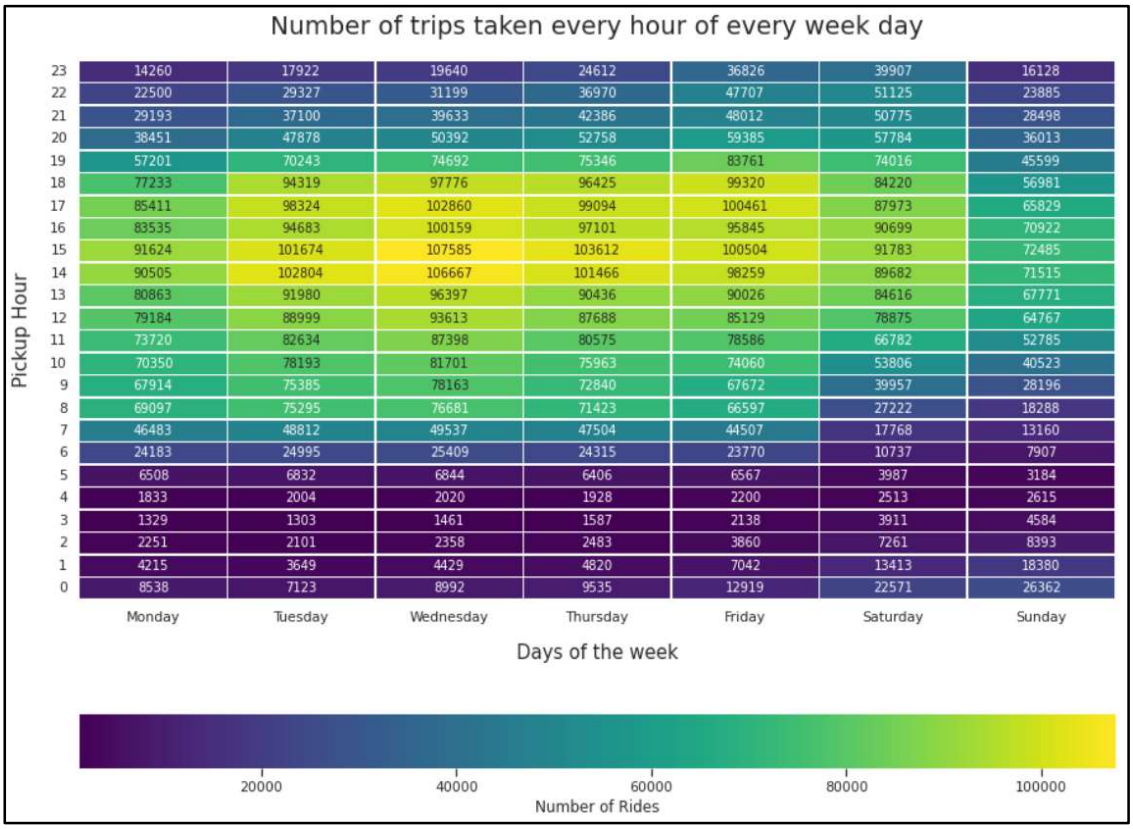
Once we had the data cleaned and processed, we began some basic exploration that sprung to mind for this dataset. For instance, we wanted to know when the maximum rides were taken during the 24 hours of the day. Another exploratory factor was the number of rides taken during weekdays versus weekends. We also explored some other features such as payment types and passenger count to elicit some interesting patterns.

We spend a significant amount of time exploring the data since this phase of the assignment would be crucial while drawing conclusions. This dataset has a variety of attributes varying from location based data to financial based data. It was during exploration when we discovered some surprising patterns from the dataset. One such unique finding was the difference in the number of rides taken on a weekday versus the weekend. In a city like New York, one would expect more trips being taken on weekends rather than weekdays, but the data reveals exactly the opposite (one justification could be the imposed weekend lockdown restrictions due to the COVID-19 pandemic).



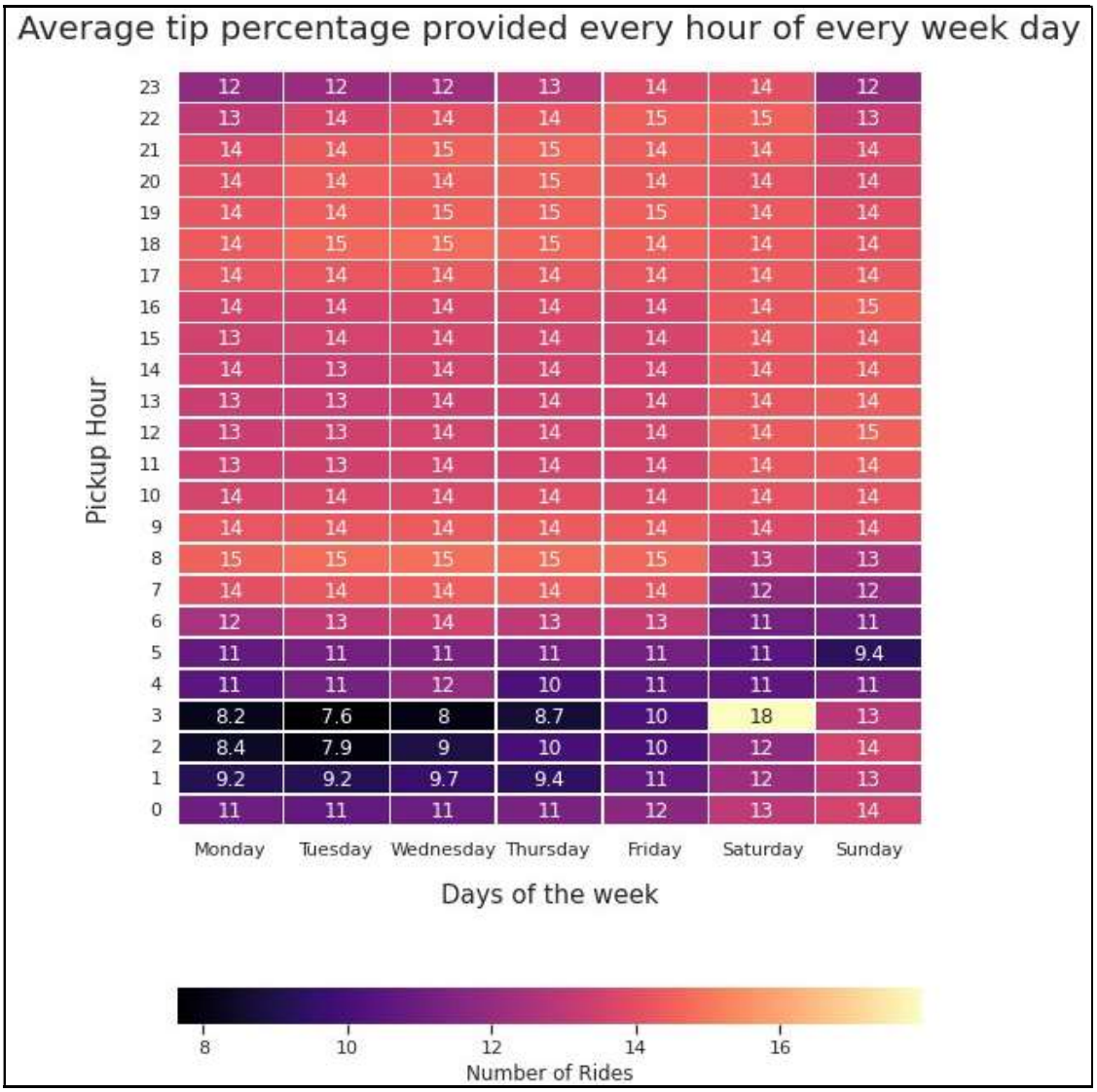
Visualisation

Upon exploring the data, it was evident that the factors such as number of rides and tips were the two factors that the NYC cab driver can modulate in order to maximize their income. Subsequently, we decided to focus on these attributes alongside our temporal data. As expected, the number of rides completed by a cab driver would have a substantial impact on their financial performance as well. So, we began by exploring the exact number of pickups done by all the yellow cabs in NYC across all the 7 days in a week and all the 24 hours of the day. To represent all these 3 attributes effectively, we chose to embed them into a heatmap. Heatmaps are a very efficient form of graphical representation that can handle representing 3-dimensional data in a 2-dimensional format.



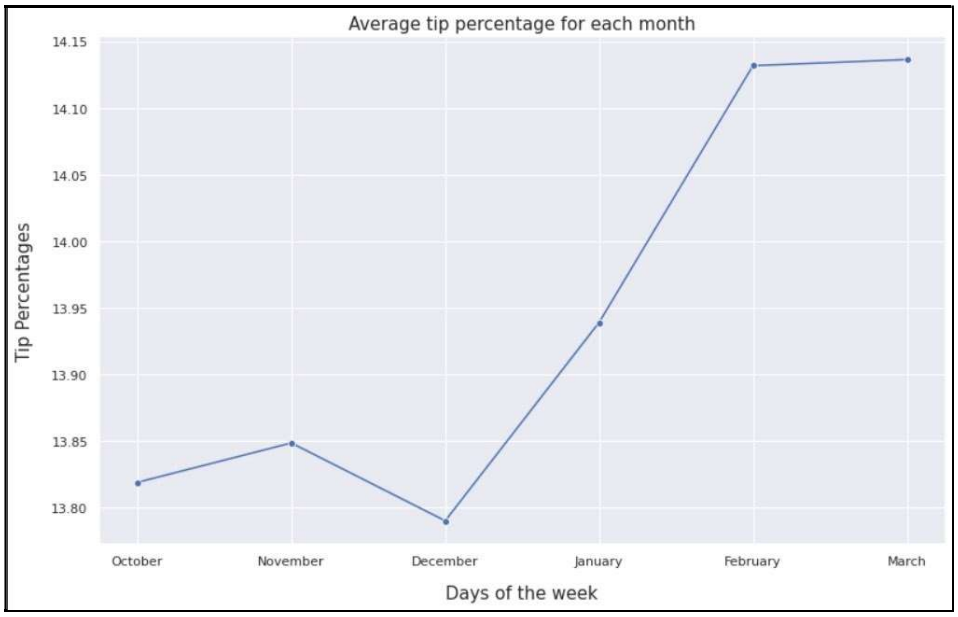
Upon plotting this heatmap graph, we can conclusively observe that there have been substantially more rides on a Tuesday, Wednesday, or Thursday evening. Unsurprisingly, this can easily be justified by many working class people taking a cab back home from work.

Another factor that significantly affects a cab drivers' income is the total amount in tips that they acquire. Logically, in order to maximize their income, a cab driver would want to know at what time during a particular day do the passengers feel particularly generous with their tips. To answer this question, we again plot a similar heatmap which this time maps the average tip paid by passengers across all the 7 days in a week and all the 24 hours of the day.

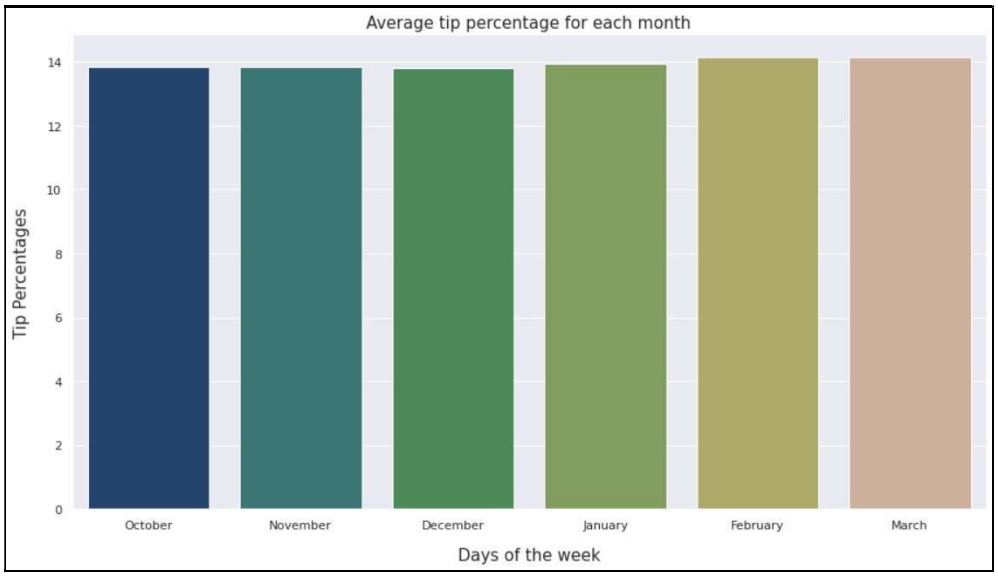


This heatmap does not reveal any substantial higher tipping period in a week. A couple of noteworthy observations that can be drawn from this heatmap are passengers tend to tip lower for night trips except passengers taking a trip on a Saturday night around 3:00 a.m. Strangely, cab drivers can expect an average tip of 18% from passengers that they pick up around 3:00 a.m. on a Saturday.

One last factor that we thought would contribute to higher than usual tips was the generosity of spirit during the holiday season. It is well known that people in general become more giving during holiday period but what we wanted to statistically calculate was: does the generosity also extend towards cab drivers in terms of higher tips? To analyse this, we mapped the average tip percentage in a month against said month. We chose to represent this analysis in a line chart expecting a spike in the December month but instead we noticed almost a slight dip in December and a gradual growth from January through March.



Although we see some dips and spikes in this line chart, we observed that the overall range of these dips and spikes is within 0.5%. Since we were dealing with temporal data, using a line chart for this analysis was our natural tendency, but after plotting a line chart for this data we found it to be misleading. Hence, we decided to use a bar chart since it emphasizes the point that there is not much variance in tip percentages inside or outside the holiday period.



Design Choices

The colours used in our charts were chosen to emphatically support the point that we intend on making through that chart. For instance, we used the 'viridis' colour scheme in our first heatmap since the yellow colour range around the maximum number of trips taken during the week represents the peak significantly in the entire chart.

Tools and libraries used

We decided to perform all the actions from integrating to visualizing our dataset in a Google Colab notebook. Google Colab provided us the flexibility to run individual code blocks one at a time instead of having to wait for the entire code to execute. This feature is also handy while collaborating among each other since we were able to upload the unintegrated data files to a shared folder on Google Drive and load it into pandas dataframes unilaterally. We used the following python libraries throughout this assignment:

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Calendar

Conclusion

After integrating, cleaning, processing, analysing, interpreting, and visualising our dataset, we came up with the following conclusions that may assist an NYC cab driver to maximize their income:

1. The best time to maximize the number of trips would be Tuesday, Wednesday, or Thursday evening around 2:00 p.m. to 4:00 p.m.
2. A bad time to find a fare would be Monday, Tuesday, or Wednesday late night around 3:00 a.m.
3. Number of fares on a weekend may not be as high as that on a weekday.
4. Expect the number of rides to progressively increase as the day moves along from morning to evening and gradually decrease as the day moves along from evening to night.
5. Average tips of around 14% are generally expected throughout the day.
6. Although on a weeknight, this average tip percentage falls to about 7.5 to 8%.
7. An unexpectedly high tip can be expected from a fare picked up on a Saturday around 3:00 a.m. (if you're lucky enough).
8. The holiday season does not majorly affect the tips provided by passengers.

Distribution of work as a pair

As a pair, we collaborated effectively by each taking vivid interest in the entire assignment. Though we did divide up duties and responsibilities eventually we reviewed, critiqued, and enhanced each other's work. The following list briefly describes the tasks performed by both of us:

- **Khilti Dedhia**
    - Gathered all the datasets, combined them, integrated, cleaned, and explored the dataset.
    - Implemented the heatmap to study the average tip percentages provided each hour of every week.

- **Romil Sakariya**
    - Planned out the process of cleaning, performed data processing and data exploration.
    - Implemented the heatmap to study the number of trips taken each hour of every week.
    - Implemented the line chart to study the average tip percentages for each month.

Future Works

Given time, this dataset can further be worked upon more carefully to analyse much more information other than just the temporal distribution of financial data. One such implementation could be the spread of financial data across the location-based data that is available in this dataset. Some interesting spatial graphs can be plotted using python's 'geoplot' library. Another interesting case study (that we briefly discussed taking up) was to analyse the effect of the COVID pandemic lockdown restrictions on the NYC taxi sector.

As a final addition to our visualization, we wanted to implement some interactive elements to our charts such as hover effect tooltips but due to deadline constraints we were not able to carve out sufficient time to work on this part of the project.

References

- Dataset: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Gallery of graphs to choose from: https://seaborn.pydata.org/examples/index.html
- Implementing a heatmap: https://stackabuse.com/ultimate-guide-to-heatmaps-in-seaborn-with-python/