IST 652: Scripting for Data Analysis – Final Project Report

NYC Taxi Analysis

By:
Aditi Pala
Shubh Mody
Kartik Kaul
Pankaj Yadav

❖ Table of Contents:

❖ Introduction:

Introduced in 2023, Boro Taxis in "apple green" colour can be hailed only in the outer boroughs (except at the airports) and in the northern part of Manhattan, specifically above 96th street on the east side and above 110th street on the west side.

Medallion (yellow) cabs are concentrated in the borough of Manhattan, but can be hailed anywhere throughout the five boroughs of New York City and may be hailed with a raised hand or by standing at a taxi stand. New York City is arguably the taxi capital of America and home of the classic yellow taxicab.

❖ Dataset description:

Below is the data dictionary for the Green and Yellow Taxi datasets.

Field Name	Description
VendorID	A code indicating the LPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
lpep_pickup_datetime	The date and time when the meter was engaged.
lpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown

	6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the$0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed on hailed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount –This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.
Trip_type	A code indicating whether the tripwas a street-hail or a dispatchthat is automatically assigned based on the metered rate in use but can be altered by the driver. 1= Street-hail 2= Dispatch

❖ Below is the data dictionary for taxi zones dataset:

This data shows the NYC Taxi Zones, which correspond to the pickup and drop-off zones, or LocationIDs, included in the Yellow, Green, and FHV Trip Records published to Open Data.

Field Name	Description
Shape_Leng	Shape of the polygon
The_geom	Geometric of the polygon
Shape_Area	Area of the polygon
zone	Zones in NYC
LocationID	Location ID’s of respective boroughs
borough	Names of different boroughs

❖ Data Cleaning:

1. Performed initial data analysis to check the data types of each column

```
data_19.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6044050 entries, 0 to 6044049
Data columns (total 20 columns):
 #   Column                 Dtype  
---  ------                 -----  
 0   VendorID               float64
 1   lpep_pickup_datetime   object 
 2   lpep_dropoff_datetime  object 
 3   store_and_fwd_flag     object 
 4   RatecodeID             float64
```

2. Drop redundant columns

```
data_19.drop(['ehail_fee'],axis=1, inplace = True)
data_20.drop(['ehail_fee'],axis=1, inplace = True)
data_21.drop(['ehail_fee'],axis=1, inplace = True)
```

3. Rename columns as per data requirements

```
data_21 = data_21.rename(
    columns={
        "lpep_pickup_datetime": "pickup_time",
        "lpep_dropoff_datetime": "dropoff_time",
        "PULocationID":"pickup_location",
        "DOLocationID":"dropoff_location",
        "VendorID":"vendor_id",
        "RatecodeID":"ratecode_id"
    }
)
```

4. Merging all three datasets

```
data_df = pd.concat([data_21, data_20, data_19], axis=0)
```

5. Creating Functions to efficiently convert dates into datetime datatypes

```
def conv_datetime(dt):
    in_time = datetime.strptime(dt, "%m/%d/%Y %I:%M:%S %p")
    return datetime.strftime(in_time, "%m/%d/%Y %H:%M:%S")
```

6. Removing negative outliers by considering data where dropoff_time > pickup_time

```
data_df = data_df[data_df["dropoff_time"] > data_df['pickup_time']]
```

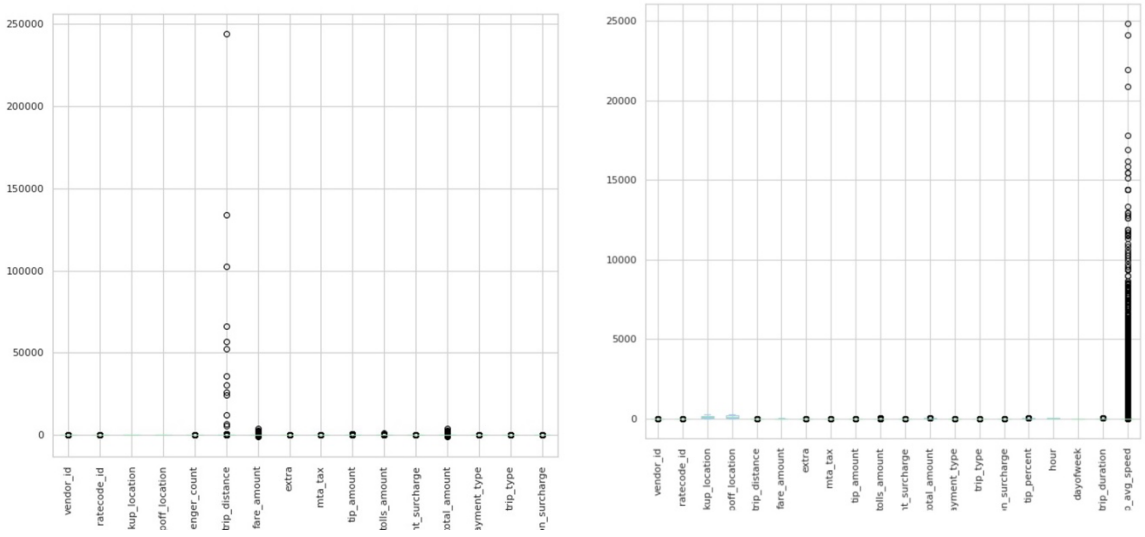
7. Fetching data between 2019 to 2021 and dropping NA values from payment type

```
data = data[pd.DatetimeIndex(data['pickup_time']).year >= 2019]
data = data[pd.DatetimeIndex(data['pickup_time']).year <= 2021]
data.dropna(subset=['payment_type'], inplace = True)
```

❖ Data Transformation:

```
data['vendor_id'] = data['vendor_id'].astype(np.float64)
data['ratecode_id'] = data['ratecode_id'].astype(np.float64)
data['trip_distance'] = data['trip_distance'].astype(np.float64)
data['fare_amount'] = data['fare_amount'].astype(np.float64)
data['extra'] = data['extra'].astype(np.float64)
data['mta_tax'] = data['mta_tax'].astype(np.float64)
data['tip_amount'] = data['tip_amount'].astype(np.float64)
data['tolls_amount'] = data['tolls_amount'].astype(np.float64)
data['improvement_surcharge'] = data['improvement_surcharge'].astype(np.float64)
data['total_amount'] = data['total_amount'].astype(np.float64)
data['payment_type'] = data['payment_type'].astype(int)
data['trip_type'] = data['trip_type'].astype(np.float64)
data['congestion_surcharge'] = data['congestion_surcharge'].astype(np.float64)
data['pickup_location'] = data['pickup_location'].astype(int)
data['dropoff_location'] = data['dropoff_location'].astype(int)
data['passenger_count'] = data['passenger_count'].astype(np.float64)
```

1. Box plot of dataset with outliers and without outliers:



2. For Trip distance, Tip amount:
    a. Checking where trip distance is more than 0
    b. Using InterQuartile Range concept to remove outliers

```
# trip distance > 0
data = data[data['trip_distance'] > 0]

trip_dist = data['trip_distance']
Q1 = trip_dist.quantile(0.25)
Q3 = trip_dist.quantile(0.75)
IQR = Q3 - Q1


data =  data[-((data['trip_distance'] < (Q1 - 1.5 * IQR)) | (data['trip_distance'] > (Q3 + 1.5 * IQR)))]
```

3. Total_amount, tip_amount:
    a. Checking where Total_amount is more than $2.5
    b. Using InterQuartile Range concept to remove outliers

Checking where total amount is more than or equal to 2.5
Note:Since the initial charge for NYC green taxi is $2.5, any transaction with a smaller total amount is invalid, thus it is to be dropped
Refer: https://www.nyc.gov/site/tlc/passengers/taxi-fare.page under "Standard Metered Fare"

```
# total amount >= 2.5
data = data[data['total_amount'] >= 2.5]

total_amt = data['total_amount']
Q1 = total_amt.quantile(0.25)
Q3 = total_amt.quantile(0.75)
IQR = Q3 - Q1

data =  data[-((data['total_amount'] < (Q1 - 1.5 * IQR)) | (data['total_amount'] > (Q3 + 1.5 * IQR)))]
```

4. For Passenger count:Consider passengers between 0 to 4:

Note: Why are we taking 0 passengers? Because Green taxis can also deliver packages, hence passengers are 0!

```
# passenger counter b/w 0 and 4
data = data[data["passenger_count"]>=0]
data = data[data["passenger_count"]<=4]
data["passenger_count"] = data["passenger_count"].fillna(data["passenger_count"].median())
```

5. Calculate tip percent based on tip_amount and total_amount:

```
data['tip_percent'] = 100*(data.tip_amount/data.total_amount)
```

6. Create new columns based on data requirements:

```
data["hour"] = pd.to_datetime(data['pickup_time']).dt.hour
data["dayofweek"] = pd.to_datetime(data['pickup_time']).dt.dayofweek
```

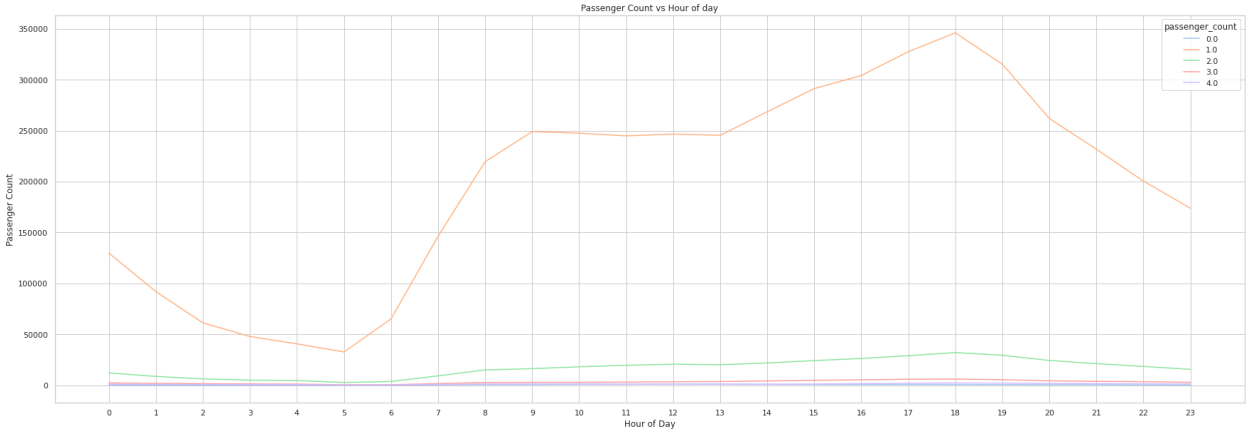
7. Calculate trip duration and average speed using existing columns:

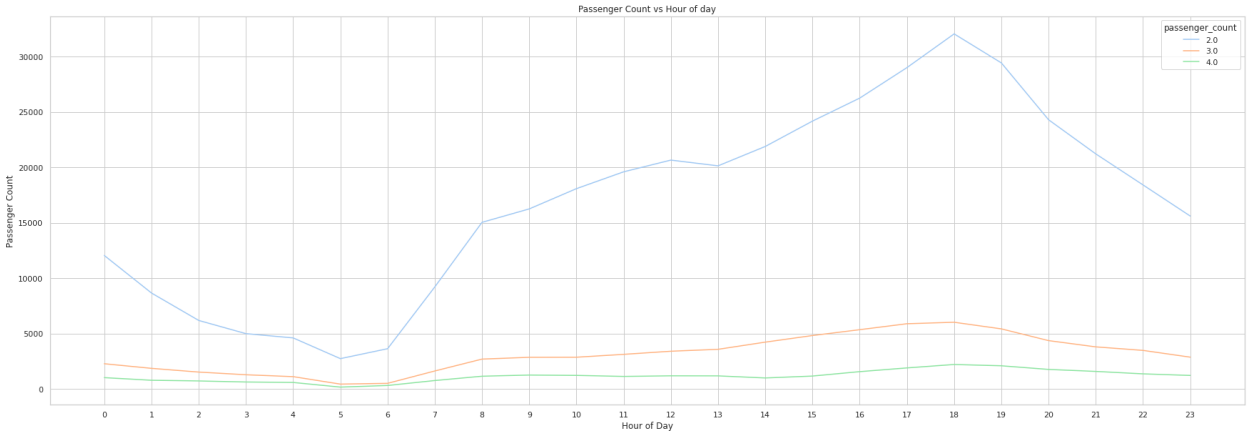
```
# Calc. Trip Duration - in Minutes
data['trip_duration'] = (data['dropoff_time'] - data['pickup_time']).dt.total_seconds()/60

duration = data['trip_duration']
Q1 = duration.quantile(0.25)
Q3 = duration.quantile(0.75)
IQR = Q3 - Q1

data =  data[-((data['trip_duration'] < (Q1 - 1.5 * IQR)) | (data['trip_duration'] > (Q3 + 1.5 * IQR)))]
```

```
# Avg Speed => speed = distance / time
data['trip_avg_speed'] = data['trip_distance'] / (data['trip_duration']/60) # miles/hour
data['trip_avg_speed']
```

❖ Preliminary Data Analysis

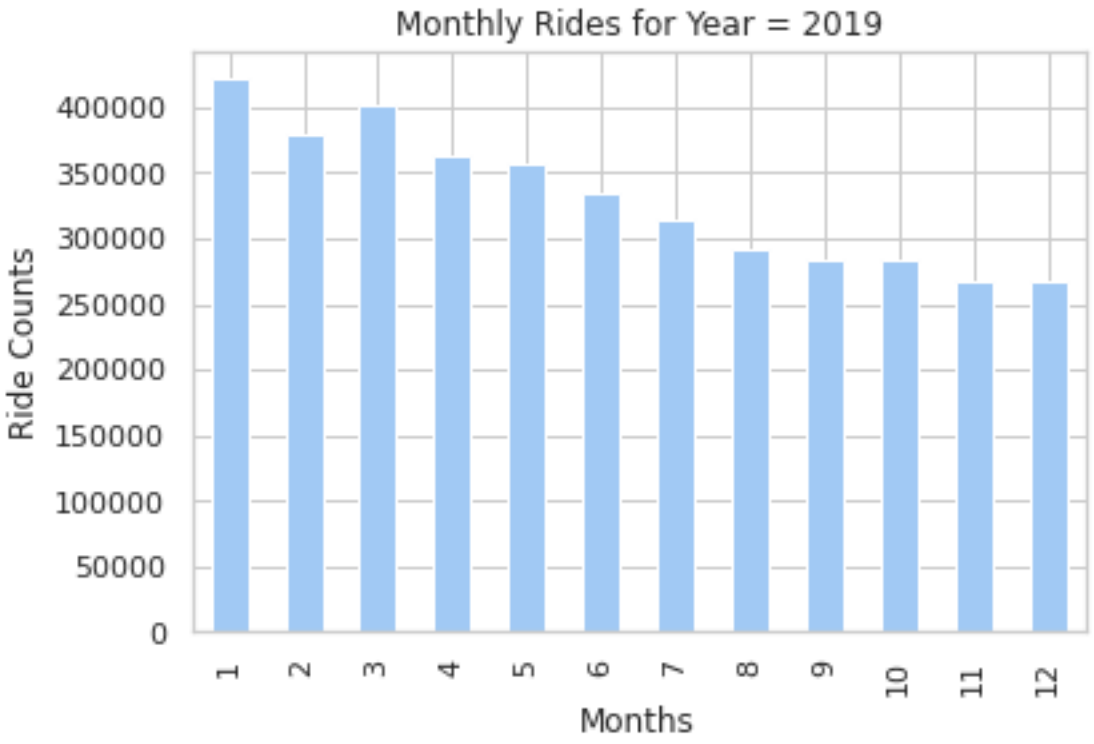
1. Trends of passenger count vs hour of the day:
    a. Green Taxi:



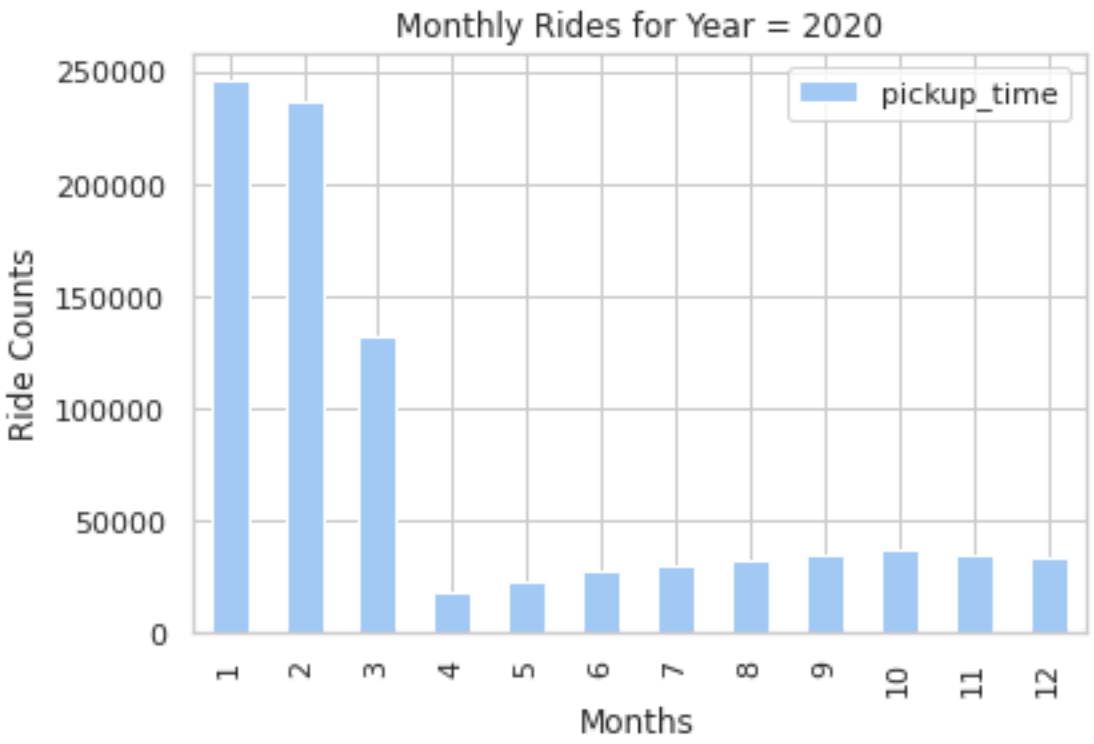


These graphs confirm that people prefer to travel with less passengers, or most likely to travel alone.

2. Monthly ride analysis for each year:



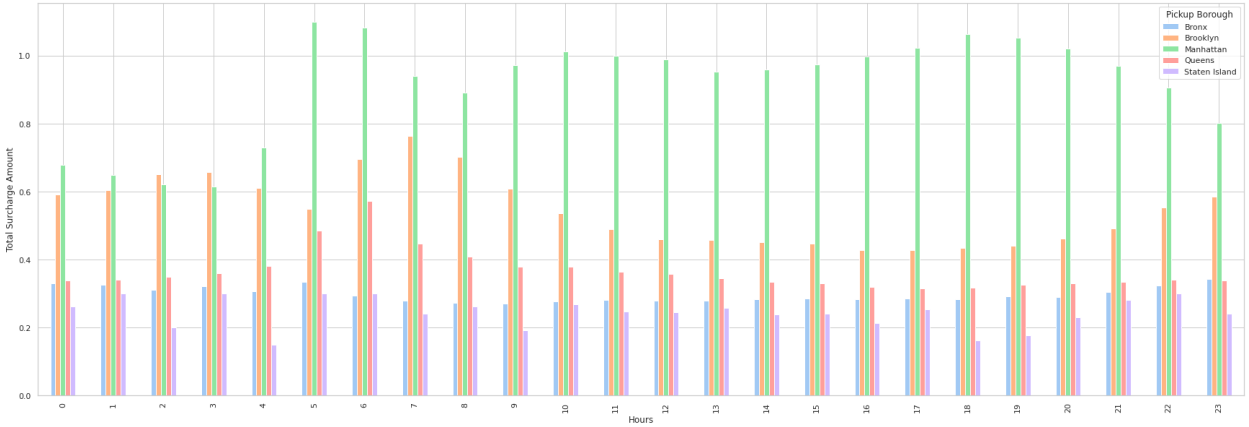
As the year was moving towards the end, the ride counts were steadily decreasing.



- During the COVID-19 pandemic, after March 2020, the ride counts took a hit with almost less than 50000 ride counts.

- As we move from 2020 to 2021, the ride counts steadily increase and come to some sense of stability that started in early 2019 i.e., pre-covid phase.

3. Analysis of surcharge amount during different hours of the day among different pickup hours:

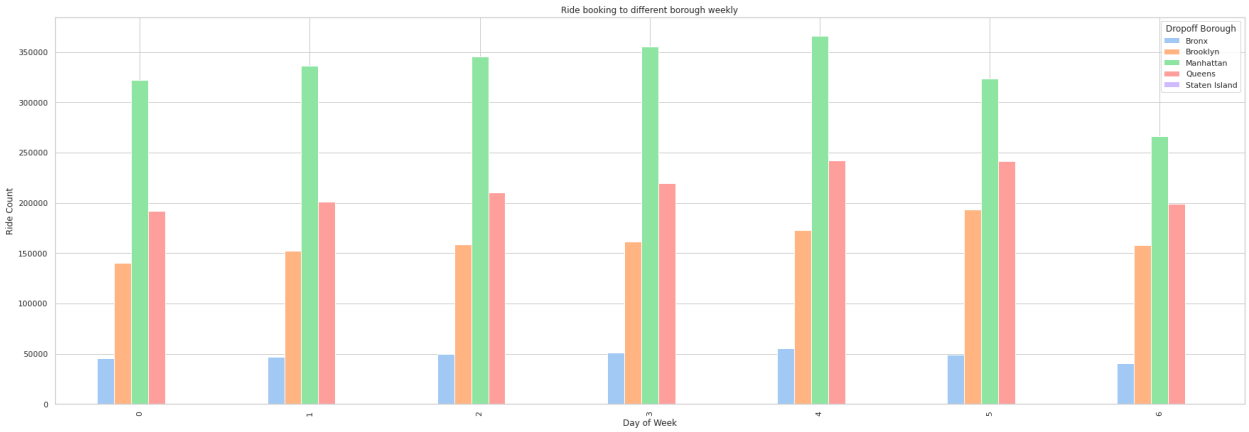


- Manhattan has one of the highest pickup surcharges throughout the day followed by Brooklyn. Least surcharge has always been in Staten Island.

- Throughout the day, there is surcharge among all the pickup boroughs.

- During the early hours (5am-6am), there is a high inflow of surcharge records, hence surcharge would be highest at those times. During the evening hours, highest surcharge is between (6pm -7pm)

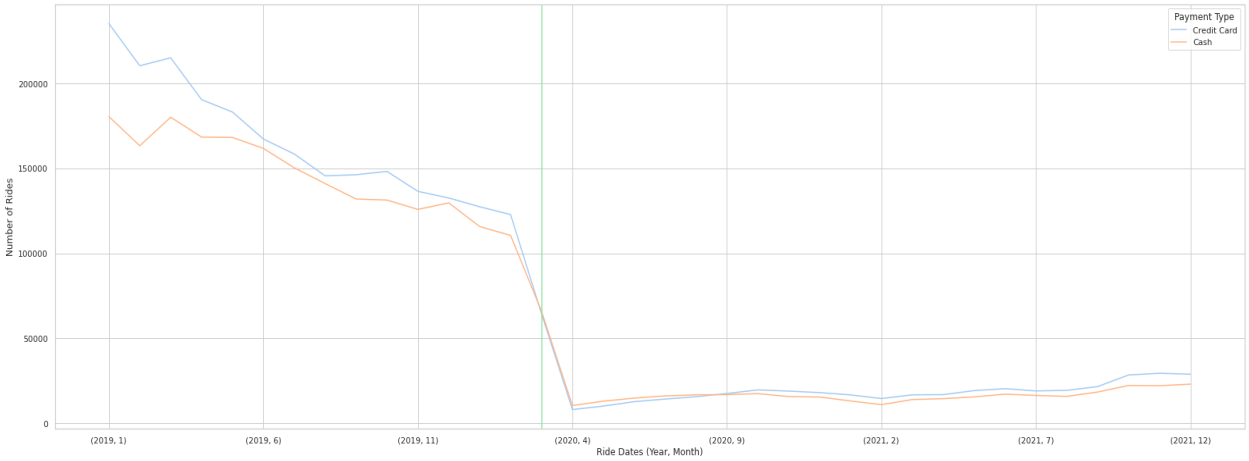
4. Analysis of surcharge amount during different days of the week among different drop-off hours:



- Highest surcharge throughout different drop off boroughs has been Manhattan, followed by Queens.
- Bronx has always had the least surcharge as compared to other boroughs.

❖ Research Questions Addressed:

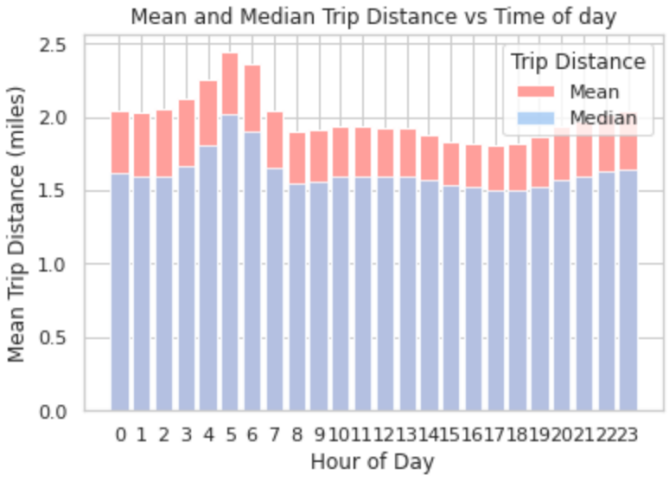
1. Trend comparison of top 2 payment types:

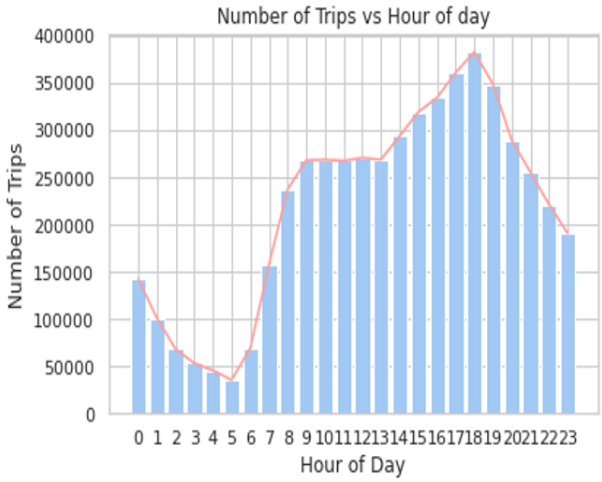


- Credit Cards and Cash were the two preferred methods of payment.
- There is a sharp decline in the graph which directly corresponds to the decline in the number of trips starting March 2020.

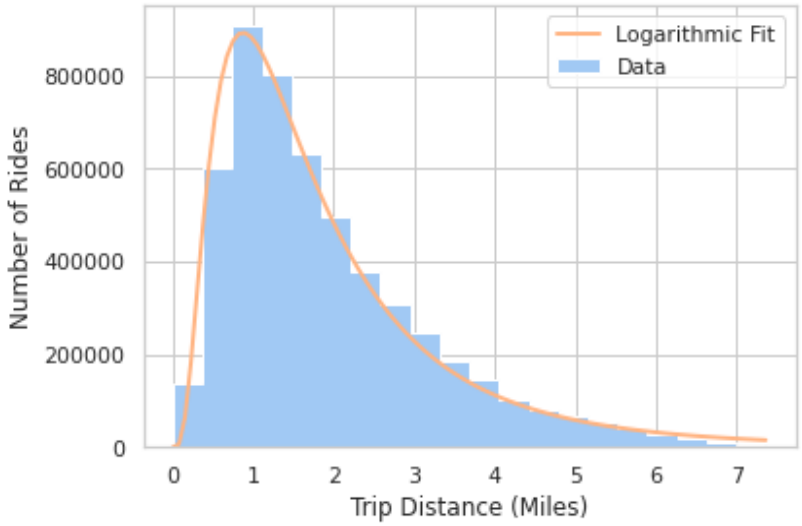
- The green vertical line signifies the official date of lockdown. This implies that COVID-19 was the cause of the down trend, and it started even before the official lockdown announcement was made.

2. Analysis of Taxi Trips throughout the Day:



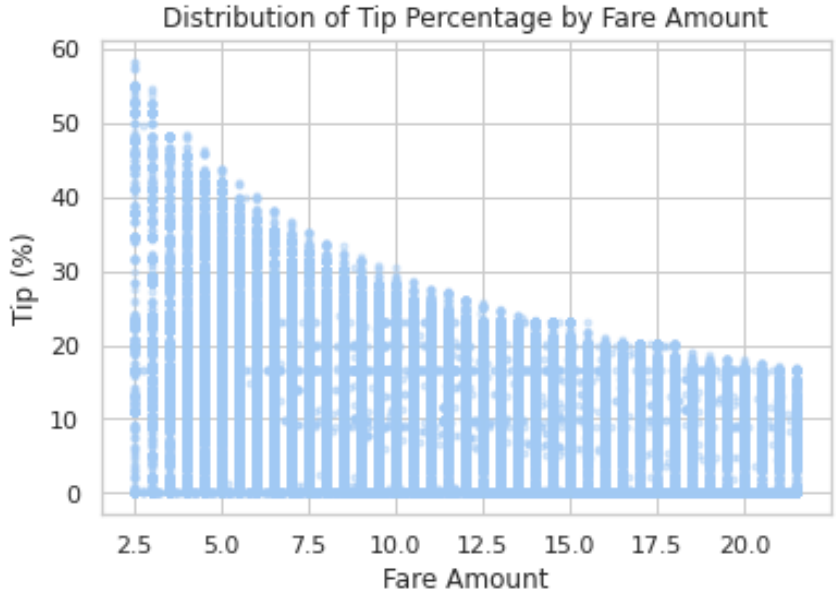


- Trips that occur in early morning hour relatively cover longer distances than those happening in the rest of the day.
- In contrast, we see that the number of trips gradually increases as the day progresses.



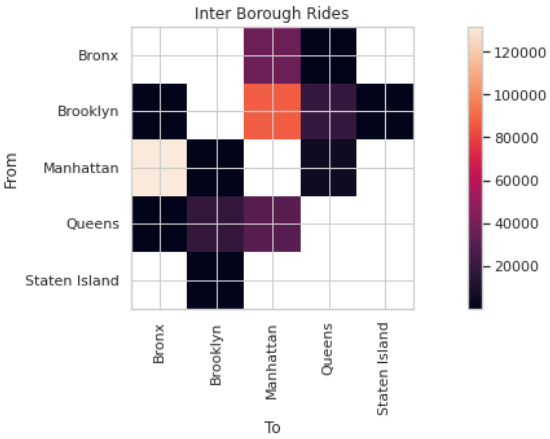
- We can confirm our analysis as this graph also depicts that the most rides occur between 0 to 2 miles.

3. Analysis of Tip amount based on Fare Prices:



- The percentage of the tip amount goes on decreasing as the Fare amount of the trip ride increases.

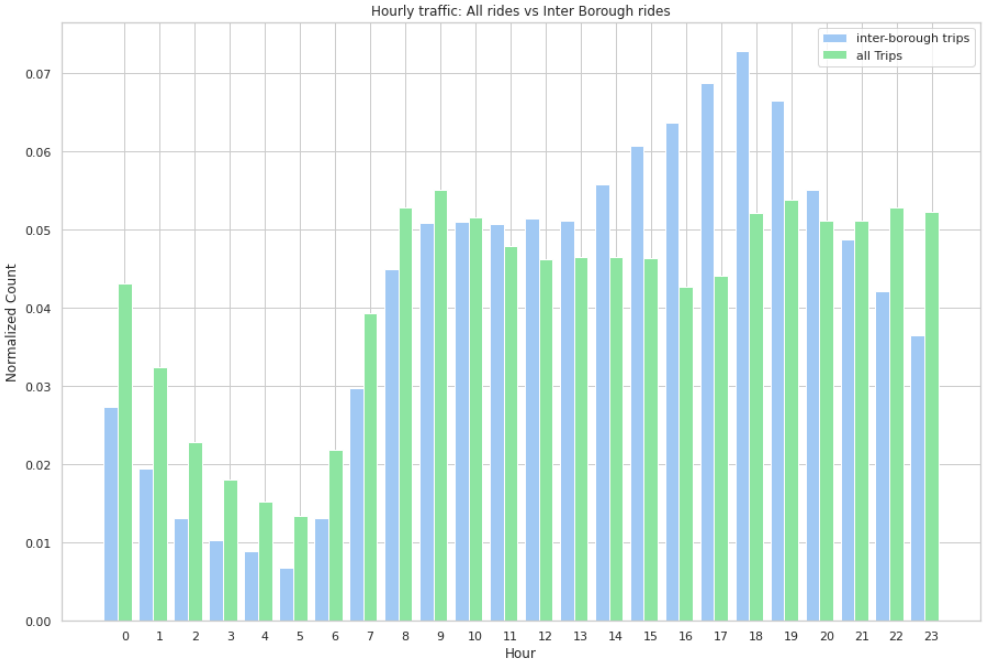
4. Analysis of Trips based on Location:



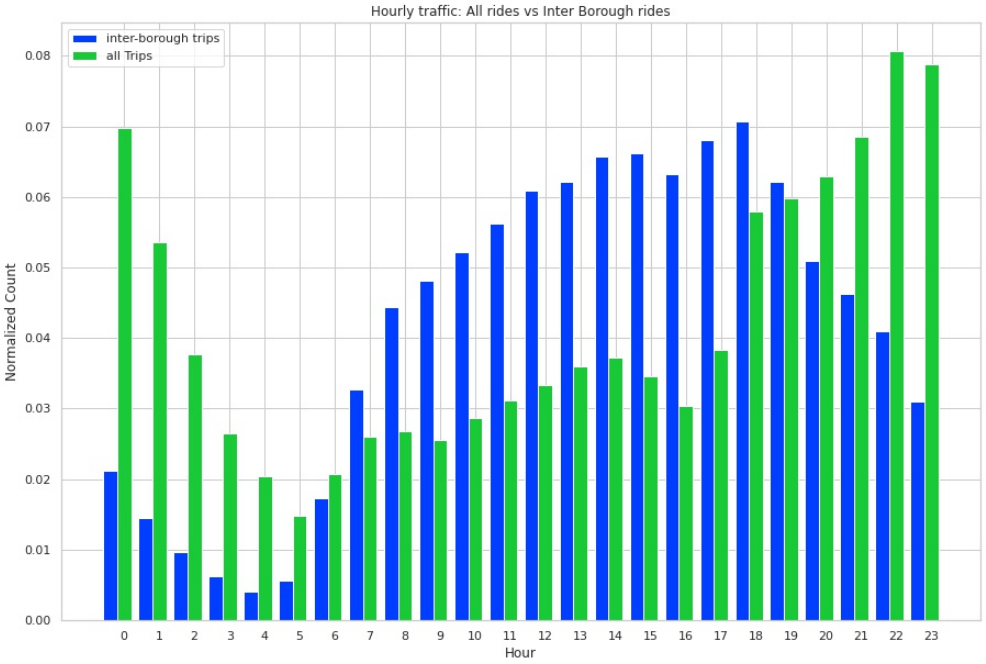
dropoff_borough / pickup_borough	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Bronx	NaN	NaN	36548.0	243.0	NaN
Brooklyn	2.0	NaN	87191.0	18805.0	9.0
Manhattan	131728.0	119.0	NaN	5557.0	NaN
Queens	372.0	18483.0	30269.0	NaN	NaN
Staten Island	NaN	8.0	NaN	NaN	NaN

- The 2D raster image represents the pivot table shown below. It signifies the
- number trips from and to a particular borough.
- The highest number of inter-borough trips are happening from Manhattan to Bronx, followed by Brooklyn to Manhattan and then Bronx & Queens to Manhattan.
- There is a gap in data for Staten Island because people usually prefer travelling to Staten Island via the ferries.

5. Hourly Analysis of Inter-Borough Trips Vs All Trips:
    a. Green taxi:

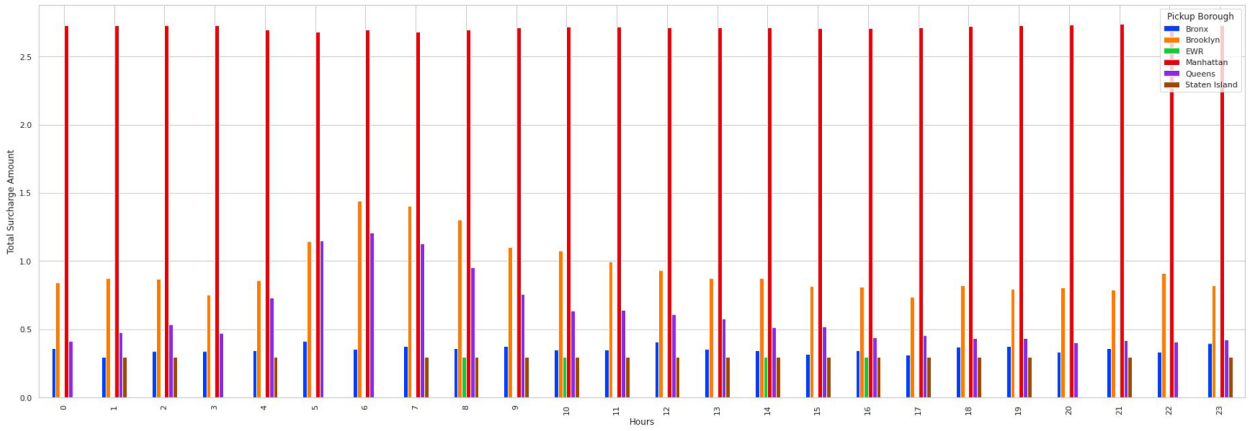


    b. Yellow Taxi:



- While the taxi trips to different boroughs increases during the evening hours, we can see that there is decrease in the intra-borough trips.

❖ **Comparative Analysis with green taxi:**



- The total surcharge amount for Manhattan is about $2.75 which is greater in comparison from green taxis.

❖ Conclusion

- Credit Cards and Cash were the two preferred methods of payment where people with credit cards pay higher tips more frequently.
- Green Taxis, cover very a smaller number of fares during late night hours, which indicates people prefer to pre book their cabs with regards to safety concern.
- There is a sharp decline in the number of trips starting February 2020 even though the official lockdown date was mid-March 2020 which is represented by green vertical line. This implies that COVID-19 was the cause of the down trend, and it started even before the official lockdown announcement was made.
- Trips that occur in early morning hours relatively cover longer distances than those happening in the rest of the day.
- Majority of the bookings begin to start early morning during office hours and there is a spike in number of cabs booked around 5pm and 6pm as the offices close and people rush towards home.
- In contrast, we see that the number of trips gradually increases as the day progresses.
- Most rides occur between 0 to 2 miles.

❖ Reference: Link