Introduction

In recent years, New York City is facing an uncertain future where the consistent and unpredictable climate-changing than ever before. In particular, the increasing temperature and precipitation led by rising sea levels are all damageable for a coastal city on the world map. How to ease and fix these urgent problems become a challenge for all New Yorkers? According to the Mayor's Office of Climate and Sustainability report shows, over 1,000 deaths happen each year because of poor air pollution, and nearly 30% of the Vehicle's main engines are still highly relying on fossil fuels that are responsible for the city's greenhouse emission (GHG) which is still the city's second-largest source of GHG emissions, after buildings.

Electric vehicles (EVs) have been seen as one of the most crucial and effective ways of dealing with these climates issues and reaching carbon neutrality. The New York City Department of Transportation (NYC DOT) is also committed to reaching an ambitious carbon neutrality goal by reducing GHG emissions by 80% by 2050. Even though the city's EVs movements are already marching on the way, there are still lots of challenges ahead.

The Yellow Taxicabs, as the key role in New Yorkers' daily commuting, are the only vehicles that have the right to pick up street-hailing and prearranged passengers anywhere in New York City. By law, there are 13,587 taxis in New York City. The demand for these taxis at peak hours increases the delay time for commuters and passengers causing great inconvenience. Furthermore, according to the Environmental Protection Agency's carbon calculator, the average New York taxi emits more than 100,000 pounds of carbon dioxide each year. The total fleet emits nearly 580,000 metric tons. That is nearly 800 pounds for each resident of Manhattan and the equivalent of what 500,000 acres of pine forests would store in carbon for one year.

These taxis due to high demand and is the only way for transport most commuters are forced to travel using these taxis even for shorter distances. Therefore, an obvious remedy would be to have alternate means of transport guided by a demand prediction system, an ideal one would be a quick eco-friendly transport chain or to encourage carpooling for increased utilization of the cabs and to minimize single passenger trips.

The demand for taxis is affected by numerous factors such as pick-up and drop-off times, the precise location of the pick-ups and drop-offs, fares, weekends, holidays, time of the year, festivals, etc. Additionally, the weather could play an important role in crowd patterns and the likelihood of a commuter choosing a taxi during a specific weather condition.

As we intend to propose alternative methods of transport, we are extensively looking for sales data related to EV models in the industry market by analyzing their brands, affordability, range, efficiency, batteries, and overall appeal. With help of the location data, demand heatmaps can be created by clustering them into regions, further determining high and low demand regions. To understand the role of Fare, Payment Type, and weather in greater detail and further capture predictive patterns using these features.

Our project is aiming to find out those factors affecting the demand for yellow taxis in New York city thereby proposing more economical alternate transport ways of accelerating EVs adoption and achieving carbon neutrality in NYC.

Analytical Methodology

Description of the datasets:

Many websites such as Kaggle, data.world, nyc.gov were visited to find suitable datasets. After exploring several datasets, we've identified that the following 3 datasets are suitable for our study. The description of each dataset is provided below:

1. **NYC Taxi dataset:**

   The dataset contains all the trip records from Jan 2016 to Jun 2016 of New York Yellow Cabs. This data was originally published by NYC Taxi and Limousine Commission (TLC). The dataset contains approximately 1.45 million trip records spanning for over 6 months for the entire New York city.

   Data fields

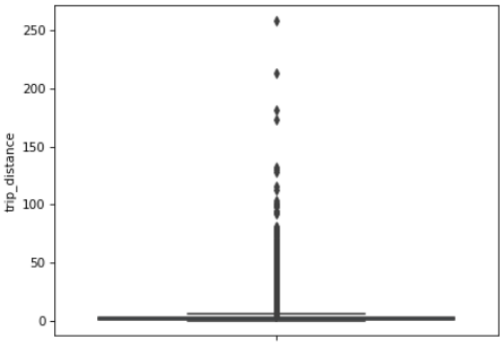
   id - a unique identifier for each trip
   vendor_id - a code indicating the provider associated with the trip record
   pickup_datetime - date and time when the meter was engaged
   dropoff_datetime - date and time when the meter was disengaged
   passenger_count - the number of passengers in the vehicle (driver entered value)
   pickup_longitude - the longitude where the meter was engaged
   pickup_latitude - the latitude where the meter was engaged
   dropoff_longitude - the longitude where the meter was disengaged
   dropoff_latitude - the latitude where the meter was disengaged
   store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
   trip_duration - duration of the trip in seconds

   Outlier Removal

   The dataset was filtered to exclude all the rows that contain missing values.

```
df = df.dropna()
df.isnull().sum()
```

   Finally, a box plot was plotted to inspect for outliers in trip durations and trip times.



Feature Engineering

Additional data fields were engineered from the existing data fields to improve the analysis and model training.

Day name – Name of the day when pickup happened

Weekday or weekend – whether the pickup day was a weekday or weekend

Pickup hour – hour of the pickup

Day period – based on the pickup hour, broadly classified as morning, evening, afternoon and night

Based on the longitude and latitude of the pickup and drop-off, with the help of Bing Map API we were able to get the address and zip code for all the pickups and drop-offs. Using this information, the pickups and drop-offs were segregated into 5 boroughs of the New York City. The information was further used to extract the travel distance and duration between pickup location and drop-off location.

Pickup_zipcode – Zip code of the pickup location

Dropoff_zipcode – Zip code of the drop-off location

Trip Distance – Distance of the most efficient route from source to destination

Trip Duration – Duration from source to destination during traffic

```
def get_zipcode(df, geolocator, lat_field, lon_field):
    location = geolocator.reverse((df[lat_field], df[lon_field]),timeout=200)
    try:
        df["zip code"] = location.raw['address']['postcode']

    except:
        df["zip code"] = np.NaN

    return df
```

```
def calculate_distance(df, pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude):
    sleep(0.3)
    routeUrl = "http://dev.virtualearth.net/REST/V1/Routes/Driving?wp.0=" + str(df[pickup_latitude]) + ","
               + str(df[pickup_longitude]) + "&wp.1="+ str(df[dropoff_latitude]) + "," + str(df[dropoff_longitude])
               + "&key=" + bingMapsKey
    try:
        request = urllib.request.Request(routeUrl)
        response = urllib.request.urlopen(request)
        jsonResponse = json.load(response)
        df["travelDistance"] = jsonResponse["resourceSets"][0]["resources"][0]["travelDistance"]
        df["travelDuration"] = jsonResponse["resourceSets"][0]["resources"][0]["travelDuration"]
        df["travelDurationTraffic"] = jsonResponse["resourceSets"][0]["resources"][0]["travelDurationTraffic"]
    except:
        df["travelDistance"] = np.NaN
        df["travelDuration"] = np.NaN
        df["travelDurationTraffic"] = np.NaN
    return df
```

For our analysis, we selected Manhattan as the primary region of focus. Hence, we filtered the dataset to include pickups only from Manhattan. We did this by identifying all the zip codes that are part of Manhattan and filtered the dataset based on these zip codes.

```
import pandas as pd
import numpy as np

zipcode_df = pd.read_csv("manhattan_zipcodes.txt", sep="\t")
vehicle_df = pd.read_csv("ny_ev_registrations_public.csv")
sample_df = pd.read_csv("sample_train_withzip.csv")

print(type(zipcode_df["Zip Code"]))
zipcode_df["Zip Code"] = zipcode_df["Zip Code"].astype(int)
vehicle_df["Zip Code"] = vehicle_df["Zip Code"].astype(int)
sample_df["Zip Code"] = pd.to_numeric(sample_df["Zip Code"], errors='coerce')
sample_df = sample_df.dropna(subset=['Zip Code'])
sample_df["Zip Code"] = sample_df["Zip Code"].astype(int)

final_df1 = vehicle_df[vehicle_df["Zip Code"].isin(zipcode_df["Zip Code"])]
final_df2 = sample_df[sample_df["Zip Code"].isin(zipcode_df["Zip Code"])]
final_df1.to_csv("manhattan_evs.csv", index=False)
final_df2.to_csv("sample_train.csv", index=False)
```

Average speed per trip was calculated:

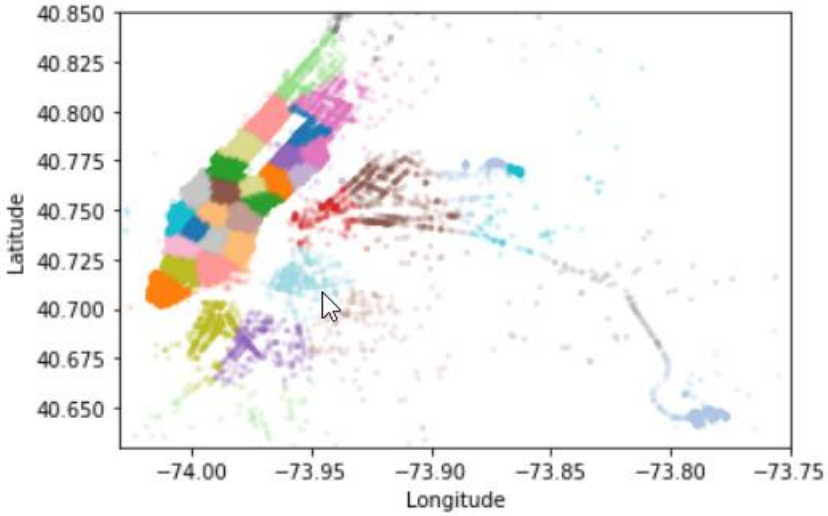
```
train.loc[:, 'avg_speed'] = 1000 * train['distance'] / train['trip_duration']
```

Clustering of Regions using location data with unsupervised learning:

The cleaned location data was introduced into K-means clustering to group locations based on the pick-up location. To make sure the clusters are of reasonable area, it was determined to form clusters based on a criterion in which the min inter cluster haversine distance between the cluster centers in 0.5 miles with maximum being 2 miles.

```
def min_distance(cluster_len,cluster_centers):
    lessth2 = []
    moreth2 = []
    good_points = 0
    bad_points = 0
    min_dist=1000
    for i in range(0, cluster_len):
        good_points = 0
        bad_points = 0
        for j in range(0, cluster_len):
            if j!=i:
                distance = gpxpy.geo.haversine_distance(cluster_centers[i][0], cluster_centers[i][1],cluster_centers[j][0],
                min_dist = min(min_dist,distance/(1.60934*1000))
                if (distance/(1.60934*1000)) <= 2:
                    good_points +=1
                else:
                    bad_points += 1
        lessth2.append(good_points)
        moreth2.append(bad_points)
    neighbours.append(lessth2)
    print (" for K ",cluster_len,"\n Avg. Number of Clusters, inter cluster dist < 2):", np.ceil(sum(lessth2)/len(lessth2)),
```

```
def find_clusters(i):
    kmeans = MiniBatchKMeans(n_clusters=i, batch_size=10000,random_state=22).fit(coords)
    data['pickup_cluster'] = kmeans.predict(data[['pickup_latitude', 'pickup_longitude']])
    cluster_centers = kmeans.cluster_centers_
    cluster_len = len(cluster_centers)
    return cluster_centers, cluster_len
```



Therefore, the clusters formed are intended to be used during the model building phase for better prediction of trip duration.

2. **NYC Weather dataset:**

Weather data for the New York city from Jan 2016 to Jun 2016 was download from [NWS New York Significant Weather Events Archive]. This data can be used to analyse the weather patterns in NYC and to analyse the effects of weather on taxi demand. Below is the description of the data:

Data fields:
Time – Date and time of the reading
Temp – Air Temperature
Wdsp – Wind speed
Dewpt – Dew point Temperature
Vism – Measured visibility
Pressure – Current pressure
Icon – Indicates current weather such as clear, cloudy, rain etc.
Rain – Rain in millimeters
Snow – Snowfall in millimeters
Tornado – True if there's an active tornado in the region

3. **NYC EV Registration Dataset**

The electric vehicle registration data was obtained from Atlas EV Hub ([State EV Registration Data – Atlas EV Hub]) for the New York city from 2010 to 2021. The main objective of using this dataset is to analyze people's opinion towards sustainable means of transportation and to analyze the trends of EV sales for certain vehicles. Dataset was later filtered to include data from Manhattan region only using the registration zip code as filter criteria. The description of the dataset is as follows:

Data fields:
Zip Code – Zip code of the registration
Registration Valid Date – Date the vehicle was registered on

VIN Prefix – Vehicle Index prefix
DMV ID – DMV ID provided by the NYC Vehicle registration department
DMV Snapshot – Snapshot of the DMV
VIN Model Year – Specifies the vehicle model and the year of manufacturing
Registration Expiration Date – Registration expiration date
State – State in which the vehicle was registered
Vehicle Name – Name of the vehicle provided by the manufacturer
Technology – Type of technology used by the vehicle (BEV, PHEV etc.)

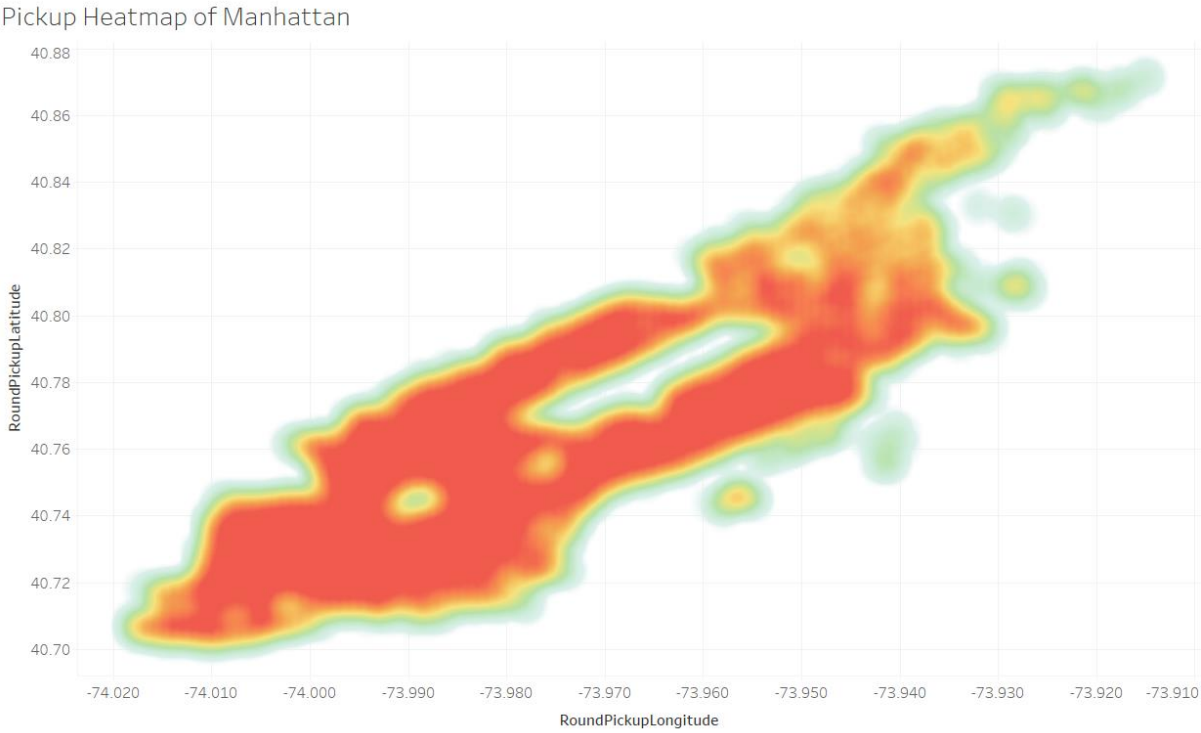
Tools and Technologies used

1. Tableau
2. Python

Findings of the Analysis:

Descriptive Analysis:

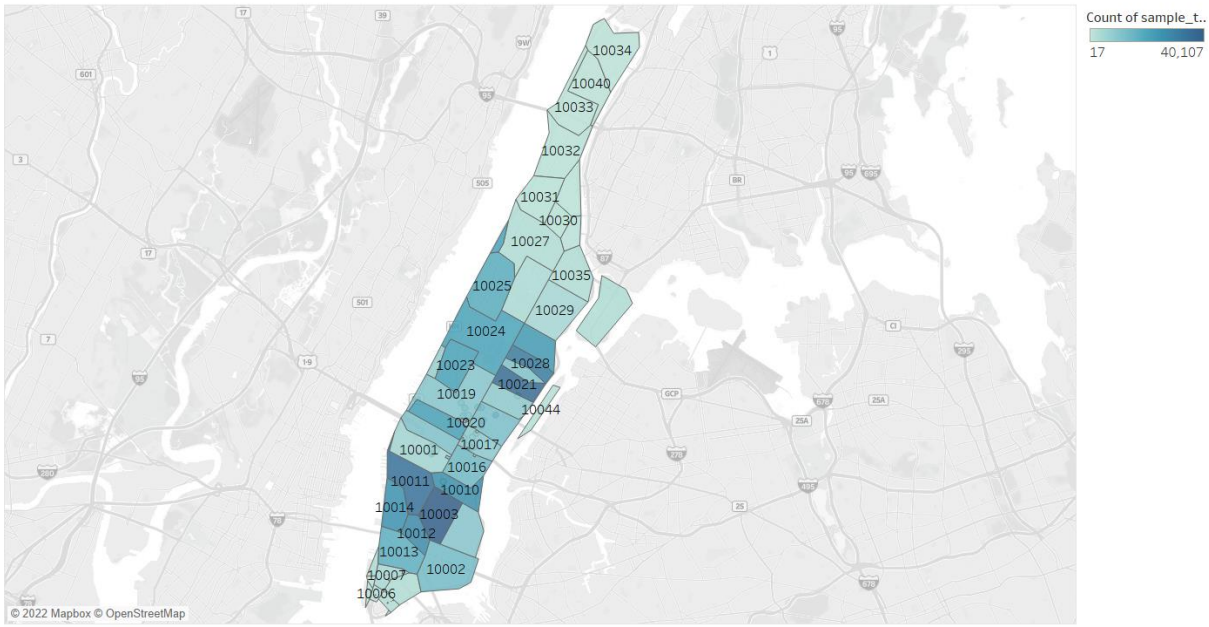
A correlation heatmap was generated using the pickup longitude and latitude as X and Y axes. From the graph we can observe that most of the trips originated from mid and lower Manhattan.



RoundPickupLongitude vs. RoundPickupLatitude. The data is filtered on Zip Code, Day Period and Month. The Zip Code filter keeps 100 members. The Day Period filter keeps 1. Morning, 2. Afternoon, 3. Evening and 4. Night. The Month filter keeps 6 of 6 members.

Another heatmap was generated based on the zip code of each pickup for Manhattan region. We can observe that most pickups have happened at the southern part of Manhattan, in particular, zip codes 10003 and 10011 with 40101 and 31200 pickups respectively. Zip codes 10021 and 10028 see fairly high number of pickups of 34000 each, where as most other southern zip codes see moderate pickups. The northern Manhattan sees minimal demand for taxis with zip code 10034 being the lowest with 76 pickups.

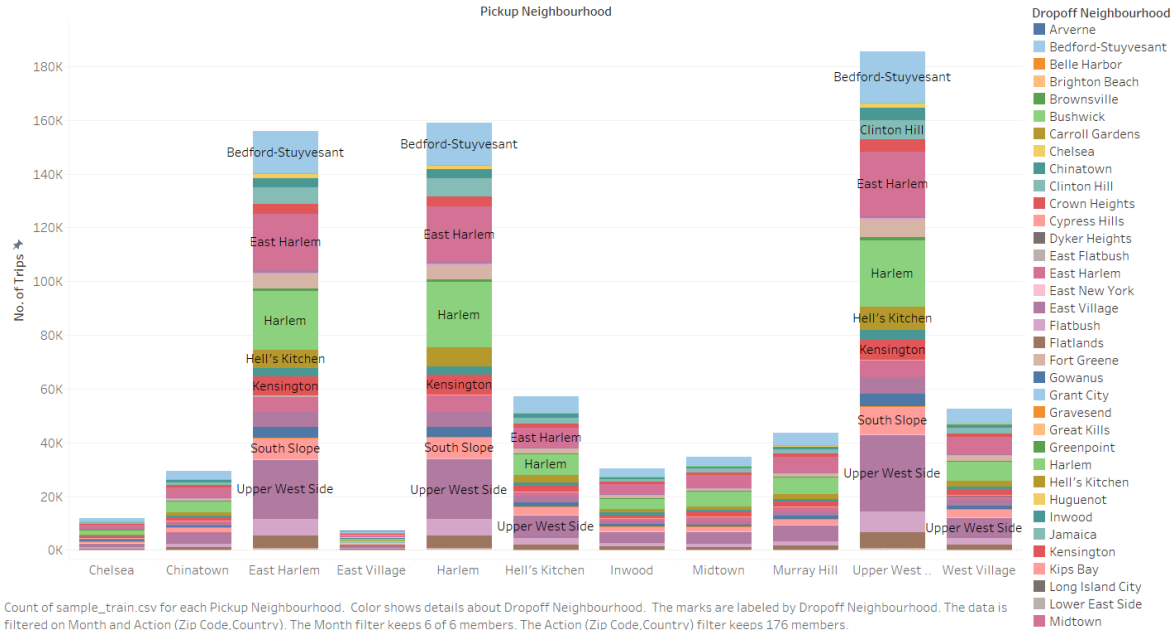
Pickup Heatmap



Map based on Longitude (generated) and Latitude (generated). Color shows count of sample_train.csv. The marks are labeled by Zip Code. Details are shown for Country. The data is filtered on Month, which keeps 6 of 6 members. The view is filtered on Zip Code, which keeps 100 members.

By using the neighborhood information which was feature engineered to analyze pickup and drop-off patterns. X-axis on the chart represents the pickup neighborhood, Y-axis represents the number of pickups, and the colors represent the destination neighborhood. Most pickups have originated from Upper Westside, Harlem, and East Harlem. It is surprising that these neighborhoods are also the most popular destinations along with Bedford and Kensington.

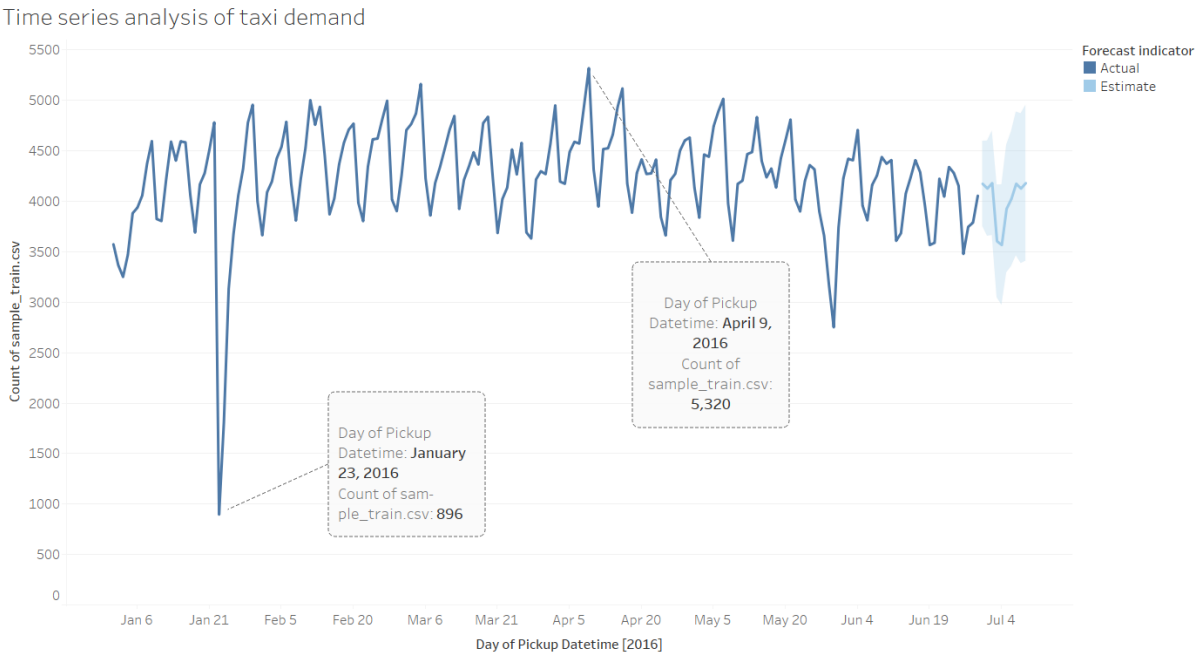
Pickup Dropoff



Count of sample_train.csv for each Pickup Neighbourhood. Color shows details about Dropoff Neighbourhood. The marks are labeled by Dropoff Neighbourhood. The data is filtered on Month and Action (Zip Code,Country). The Month filter keeps 6 of 6 members. The Action (Zip Code,Country) filter keeps 176 members.

Time series analysis was performed on the entire dataset to analyze the demand of taxi in Manhattan everyday for six months. From the time series chart, we understood that the taxi demand follows a specific pattern where Monday's see the least demand for taxis and the demand gradually grows as the week progresses, hitting the peak demand usually on a Friday or Saturday and falling on Sunday and Monday.

Another interesting fact that the overall demand for taxis grew from January to April, with highest number of pickups on April 9, 2016 and gently falling from April to June.



The trend of count of sample_train.csv (actual & forecast) for Pickup Datetime Day. Color shows details about Forecast indicator. The data is filtered on Month and Action (Zip Code,Country). The Month filter keeps 6 of 6 members. The Action (Zip Code,Country) filter keeps 176 members.

On Jan 23, 2016, significantly less pickups of just 896 were recorded. After analyzing the weather, it was identified that New York was buried in 3ft snow after a category 5 blizzard.

By using the 6-month data, time series prediction was performed to estimate the demand for the following 10 days. The parameters used for the prediction and the results are provided below.

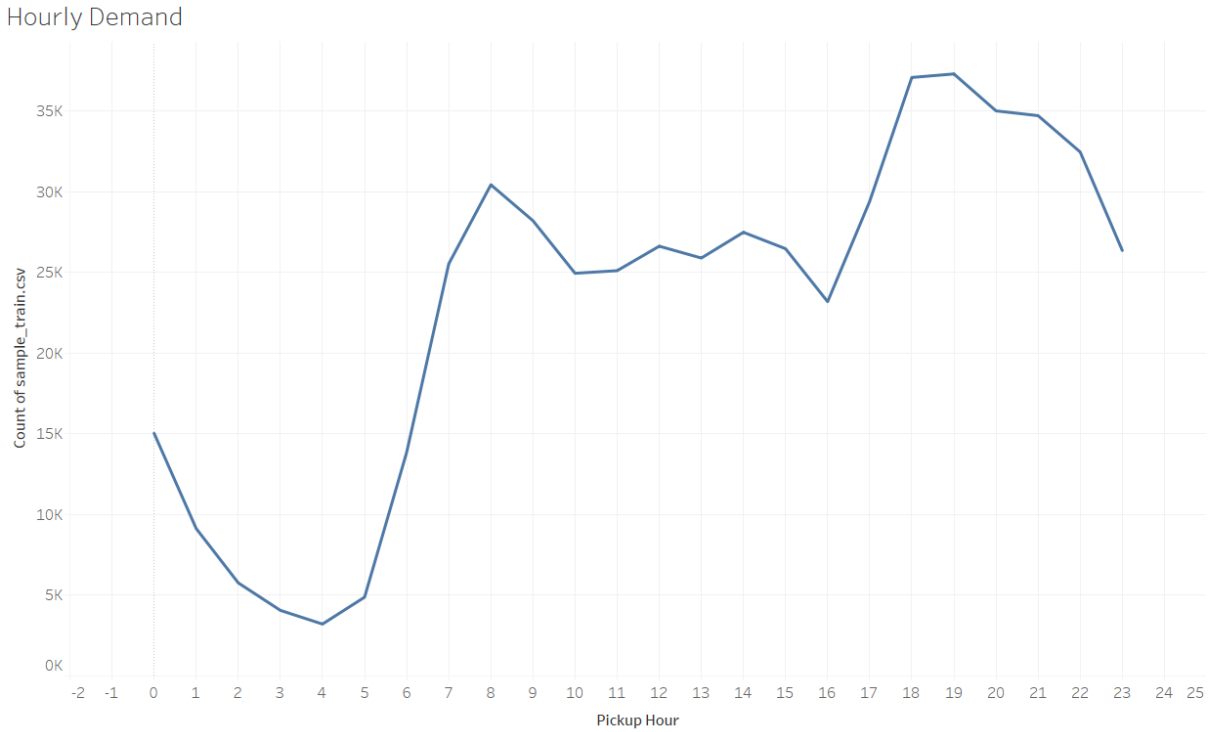
Count of sample_train.csv

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	None	Additive	215	159	0.57	3.9%	1,309	0.500	0.000	0.175

Count of sample_train.csv

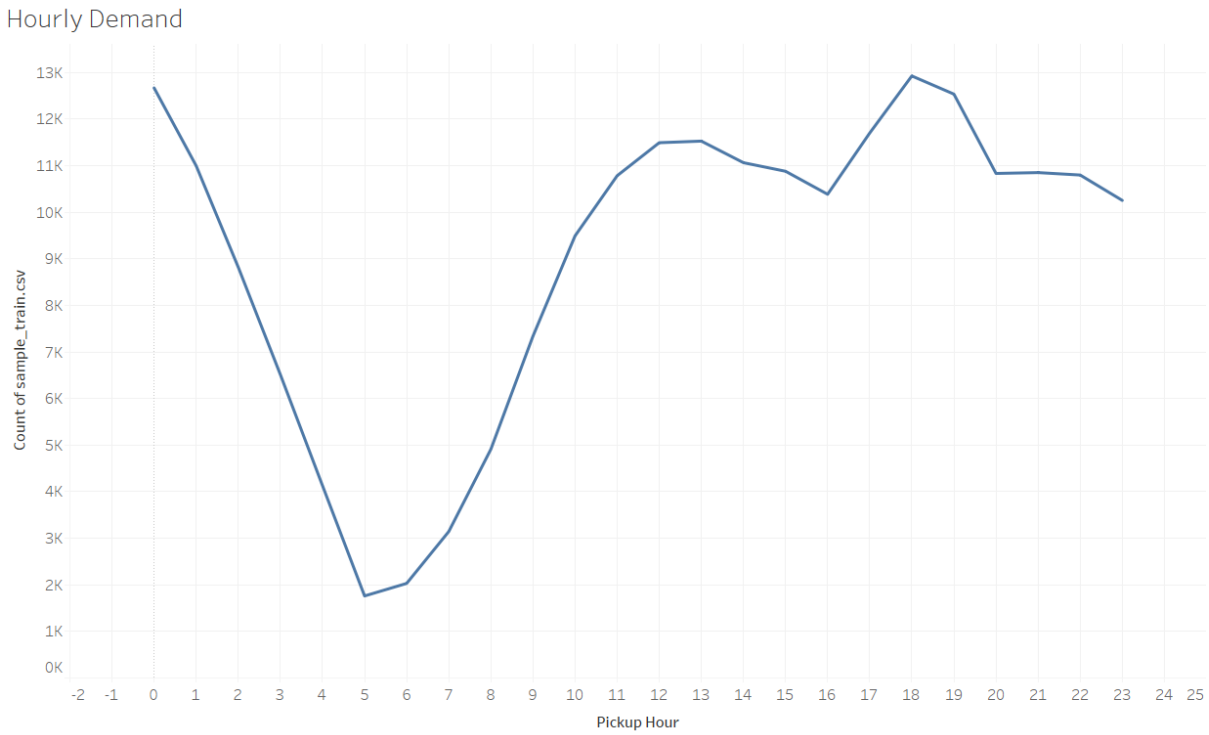
Initial	Change From Initial	Seasonal Effect				Contribution		
June 30, 2016	June 30, 2016 – July 9, 2016	High		Low		Trend	Season	Quality
4,174 ± 422	6	July 9, 2016	209	July 4, 2016	-405	0.0%	100.0%	Ok

Taxi demand based on the hour of the day for weekdays and weekends was also visualized. During the weekdays, the demand is at its lowest at 4am and rises sharply during the breakfast time reaching the local peak at 8am. The pickup rate falls a little and stays consistent through out the day reaching the peak at 7pm in the evening and slowly falling as the night progresses.



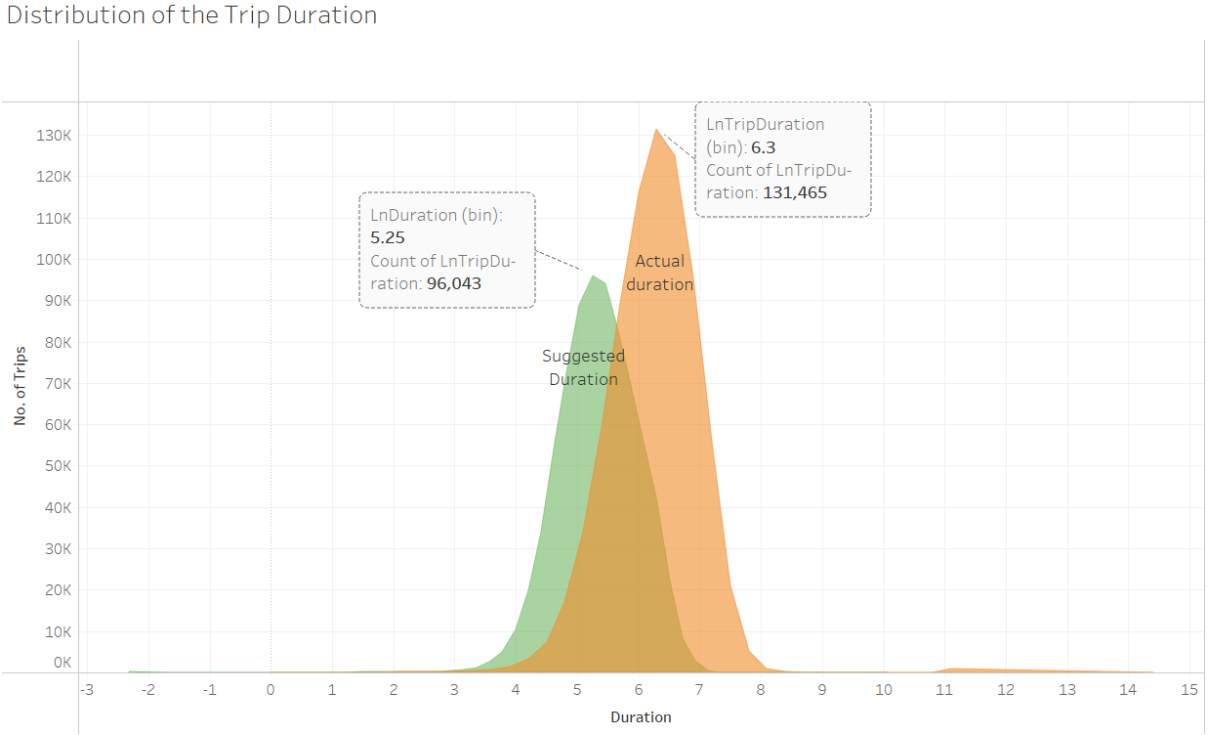
The trend of count of sample_train.csv for Pickup Hour. The data is filtered on Weekday Or Weekend and Month. The Weekday Or Weekend filter keeps Weekday. The Month filter keeps 6 of 6 members.

On the weekends, there is a high demand at 12AM which steeply falls by 5am to the minimum. However, the pickup rates increase in the morning till 11AM and stays consistent reaching the peak at 6pm in the evening.

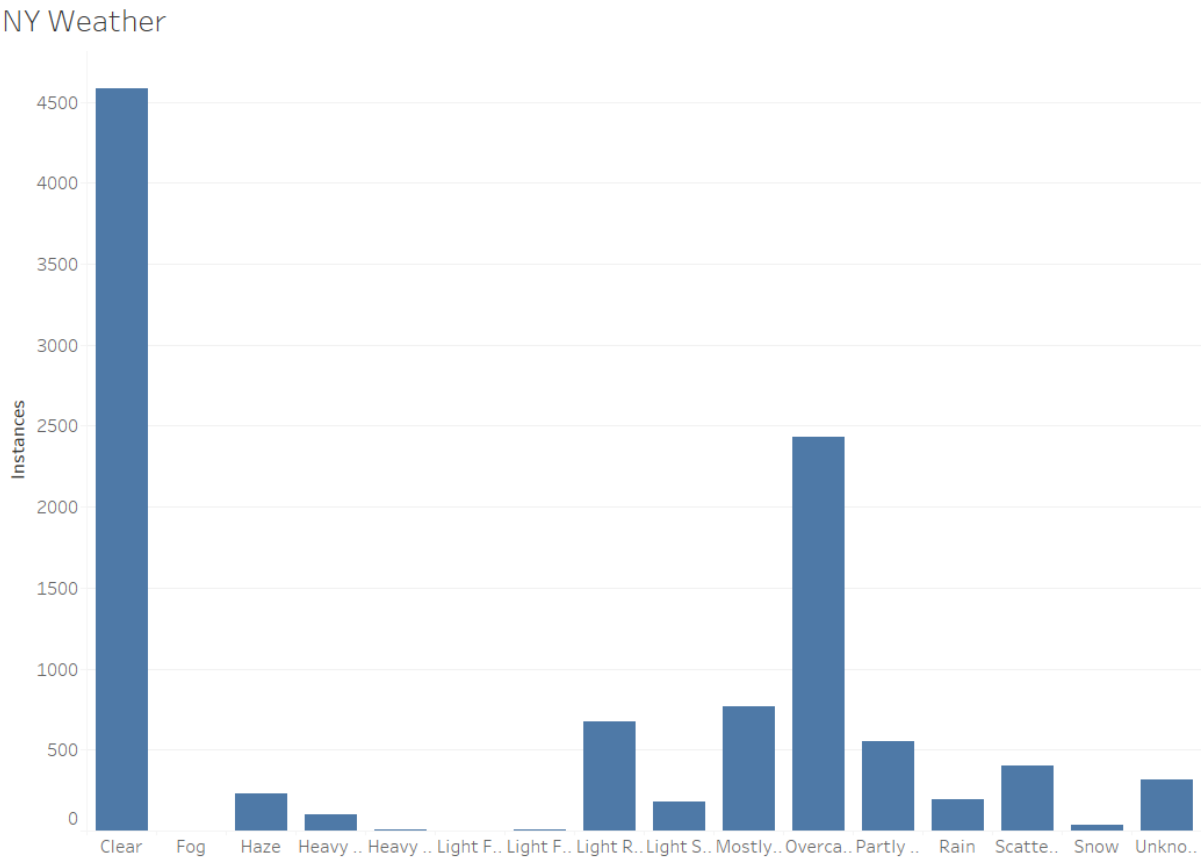


The trend of count of sample_train.csv for Pickup Hour. The data is filtered on Weekday Or Weekend and Month. The Weekday Or Weekend filter keeps Weekend. The Month filter keeps 6 of 6 members.

A normal distribution curve was plotted for the actual trip duration and the trip duration for the route suggested by Bing Maps API. On average the actual duration is significantly higher than the suggested duration. Actual duration has the mean of 6.3 which is around 9 min (EXP (6.3)) and the mean of suggested duration curve is 5.25 (4 mins).

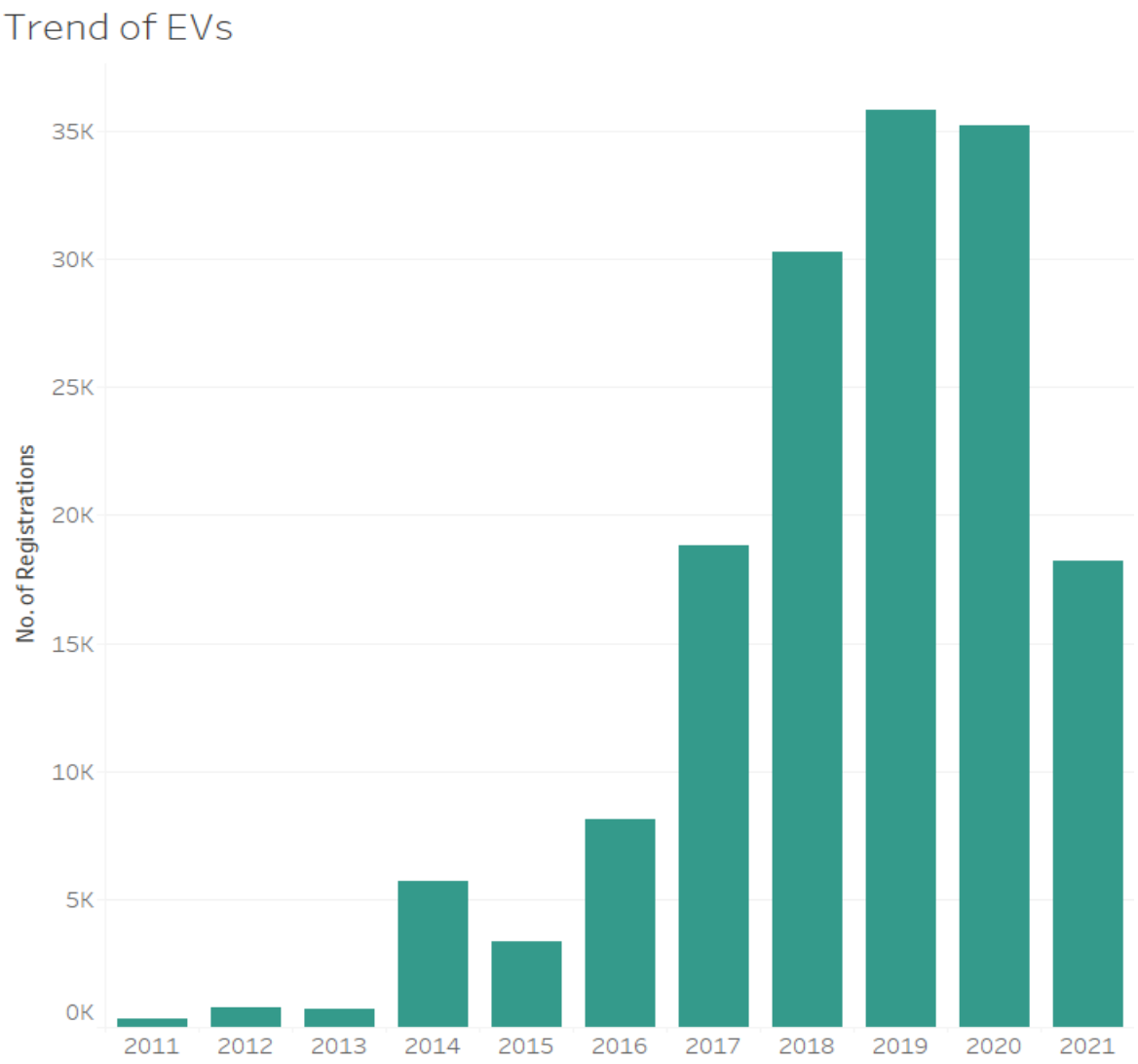


Below bar chart indicates the weather of New York during various instances of measurement. Majority of the time the weather is clear and sunny, sometimes its partly cloudy. We can further observe that New York sees very less snow or rainy days with very less instances of Heavy Rain or Snow.

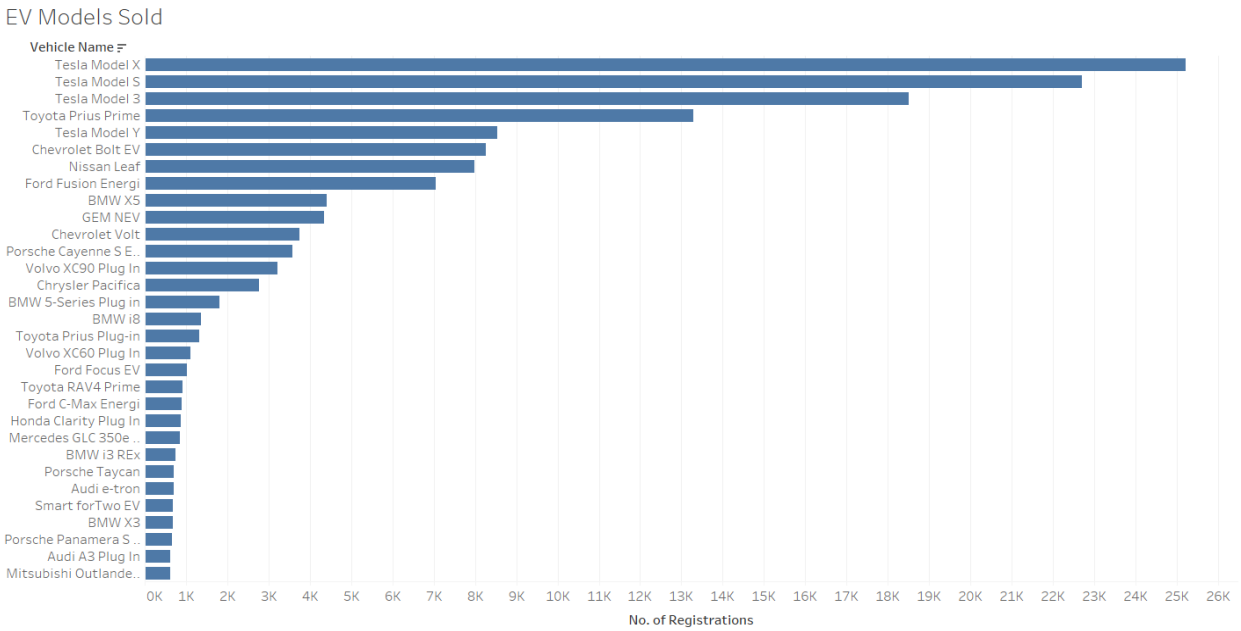


The registrations of Electric vehicles were analyzed to understand the trends of EV sales and customers opinion towards EVs and sustainability. The number registration from 2011 – 2013 stay low and see a small rise in 2014. From 2016 the registrations rise on steep curve, peaking at 2019 and falling off in the next couple of years.

The numbers have gone down in the last couple of years mainly because of the covid pandemic and the global chip shortage. Since most of the restrictions are being lifted and with the chip shortage coming to an end, the sales of EVs are predicted to go up.



The data was also analyzed to understand which companies and models are being sold more. This could be advantageous as the EV require specific types of charge technology and charging connectors. By looking at the bar chart, we can see that Tesla was the most popular automobile manufacturer with 3 of their models at the top (Model X, Model S, Model Y) followed by Toyota Prius Prime.



Predictive Analysis:

As the initial phase of Predictive analytics, we have built a Support Vector Regressor model and used root mean square log error as our model evaluation metric. The Steps are as follows:

The features that would benefit the prediction of trip duration were chosen to build a regression model.

```
do_not_use_for_training = ['id', 'pickup_datetime', 'dropoff_datetime',
                           'trip_duration', 'check_trip_duration',
                           'pickup_date', 'avg_speed',
                           'pickup_lat_bin', 'pickup_long_bin',
                           'center_lat_bin', 'center_long_bin',
                           'pickup_dt_bin', 'pickup_datetime_group',
                           'store_and_fwd_flag']
feature_names = [f for f in train.columns if f not in do_not_use_for_training]
X, y = train[feature_names], train["trip_duration"]
Xtest = test[feature_names]
```

```
Index(['vendor_id', 'passenger_count', 'pickup_longitude', 'pickup_latitude',
       'dropoff_longitude', 'dropoff_latitude', 'distance', 'pickup_weekday',
       'pickup_weekofyear', 'pickup_hour', 'pickup_minute', 'pickup_dt',
       'pickup_week_hour'],
      dtype='object')
```

The dataset split into train and test where 80% of data being train and remaining 20% being the test data.

```
from sklearn import model_selection
X_train, X_val, y_train, y_val = model_selection.train_test_split(X,y, test_size=0.2)
```

The features were standardized using a standard scalar in the model pipeline for the consistency of different scales of the features.

Hyper parameter tuning was performed using RandomSearchCV with various values of C (penalty parameter) such as {0.1, 1, 10}

```
from sklearn import linear_model, model_selection, metrics, pipeline, preprocessing, svm, compose
rmsle = metrics.make_scorer(lambda yt, yp: np.sqrt(metrics.mean_squared_log_error(yt, yp)),
                            greater_is_better=False)
params = {
    "regressor__linearsvr__C": [0.1, 1, 10]
}
model = model_selection.RandomizedSearchCV(
    compose.TransformedTargetRegressor(regressor=pipeline.make_pipeline(
        preprocessing.StandardScaler(),
        svm.LinearSVR()), func=np.log, inverse_func=np.exp),
    params, scoring=rmsle, n_jobs=None)
model.get_params()
```

After the model execution we obtained the optimal root mean square log error of -0.74

```
Model score: -0.7475479299118015
Wall time: 1h 26min 38s
```

As future steps we intend to optimize the model and improve its performance using the clustering results and by experimenting with various other machine learning algorithms.

Discussion

It is important to analyse demand to understand the customer interest for a product or service in a target market. Demand analysis techniques can be used to determine when and where a service can be introduced and how it can generate expected results or profit. It also gives a better understanding of the high-demand markets for the company's offerings, using which businesses can determine the viability of investing into such services.

From the trends of EV registrations it is apparent that interest in electric vehicles is growing every year among the people of New York City. This indicates that people are concerned about climate change and the carbon footprint of their fossil fuel vehicles. In addition, according to US Energy Information Administration the gas prices are on the rise recently and there are no signs of these prices coming down soon. This could serve as a motivation for people who are not eco-conscious to look at electric vehicles as an alternative. Therefore, by choosing electric vehicles carbon footprint can be reduced and help fight the climate change.

From the EV Sales, we understand that Tesla is the most popular automobile manufacturer. With all 4 models in the top 5, Tesla is undoubtedly the leader in EV market. Coming to specific models, Model X is at the top with overall sales followed by model S, 3 and Y at 2nd, 3rd, and 5th places respectively. By considering the launch dates, Model Y takes the lead in terms of sales as it is less expensive compared to other models. In conclusion, it is better to choose Tesla Model Y as taxis for the business.

New York City has pleasant weather most of the days with clear sunny days and very few snowy and rainy days. This is a perfect opportunity to introduce another alternative to taxis which is bike sharing. Bike sharing is a popular service which could be helpful for customers who are in a hurry and want to travel quickly. NYC Department of Transportation has already started an initiative of shared scooters

in few neighborhoods. Since these scooters can make use of the dedicated bike lanes, they are safer and faster when compared to conventional taxis and are economically feasible for customers.

From the analysis of taxi trip data, it is evident that 50 percent of all the taxi pickups in New York City happen in Manhattan. Additionally, from the pickup heatmap it is confirmed that nearly 90% of these trips are concentrated to mid and southern areas Manhattan. Further looking into the pickup-drop off chart, it is observed that most of these pickups originate from Upper Westside, Harlem, and East Harlem. People of these neighborhoods where demand is so high would be looking at alternatives for the traditional yellow cabs and with such high userbase it will be easier to launch new services. As a business, attracting the customers and gaining the market share is crucial to be successful. Hence, we would suggest on focusing these neighborhoods as an initial target area for the solution.

Finally, replacing traditional cars with electric vehicles and introducing shared bikes which could be rented by customers in Upper Westside, Harlem, and East Harlem can be helpful to understand the pros and cons of these services. This could be further analyzed to improve the services continuously and reduce the carbon footprint.

References

https://www1.nyc.gov/html/dot/downloads/pdf/electrifying-new-york-report.pdf

Gasoline and Diesel Fuel Update - U.S. Energy Information Administration (EIA)

Shared E-Scooter Pilot | Shared E-Scooter Pilot (nycdotscootershare.info)