Shared Car Locations
Project Report

Project Description:

Shared rental cars and services have become more prominent and easy to use because of recent technological advancements. The more logistically automated approach is to use the service where customers pick up and drop off the car at their convenience. The objective of this project is to improvise the car-share service by analyzing the pattern of commute which can be used to make changes to maximize profit and availability, and the number of rides of the car-sharing model. We use the location data (latitude and longitude) of shared car parking locations (updated every 2 minutes) and analyze the dataset for the above objective.

Motivation:

The car sharing companies such as the one discussed in our project - AutoTel have their aim to maintain geospatial balance while running their service which ensures that cars are readily available at the time where they are demanded. This problem is the biggest challenge for the car service companies since this task is difficult to achieve because the customers using the service do not have the same inclination towards this optimization problem. This reason motivates our group to use machine learning and work on this challenging task to attempt solving the issues such as geospatial availability of the cars, figuring out high- low demand and supply regions for car availability, how external factors affect the ride sharing service and so on.

About the Dataset:

AutoTel is a mobile application for a shared car project formed in the city of Tel Aviv. Here the users of this service would reserve a car using this application and then pay for the ride by the minutes it was reserved. This project started in October 2017 that was used by more than 7500 users and more than 3500 of the users were recorded to reserve cars one or more times each week.

1. Data Cleaning and Data Preprocessing:

Data Cleaning:

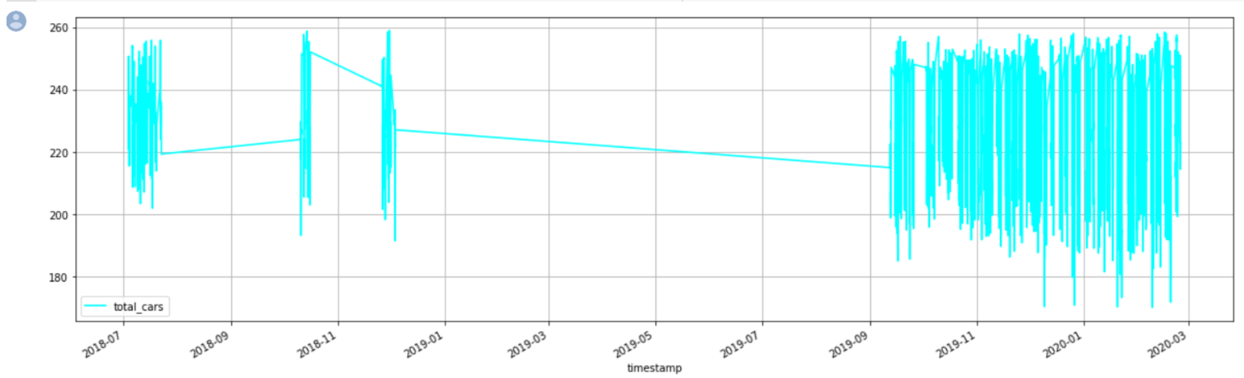
- The AutoTel dataset contains the information of the total number of cars at each parking location for every two minute intervals. It is possible that for some time-stamps there would be no car parked at a parking spot. Such rows from the table are of no use to predict the car's availability. We removed such data points to reduce the total computation cost and memory.

- Other than this the dataset we obtained was clean and 100% usable.

Data Preprocessing:

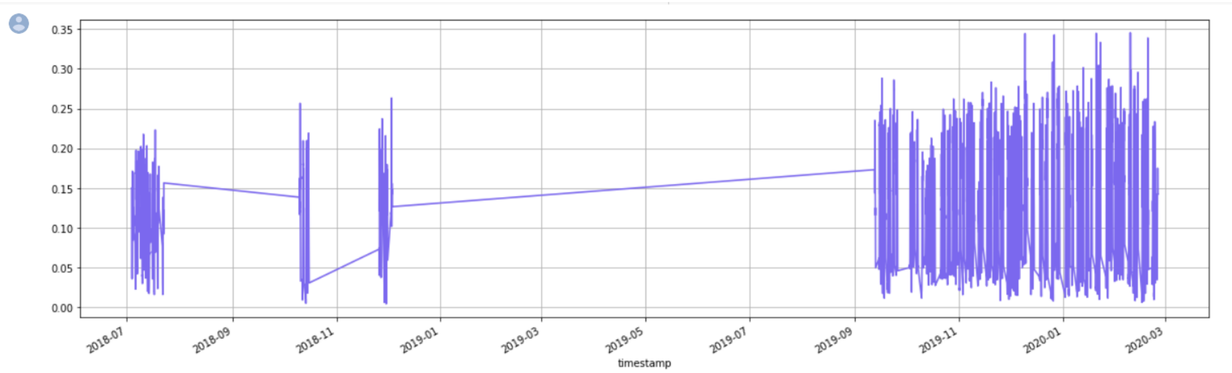
- Inorder to perform GIS operations and Geo Joins in the dataset, the pandas dataframe is converted to GeoPandas Dataframe with the geometry feature consisting of points of parking locations on the map.

The below image has been plotted number of cars vs the day frames.As per the below graph the maximum number of cars available in a day is 260.

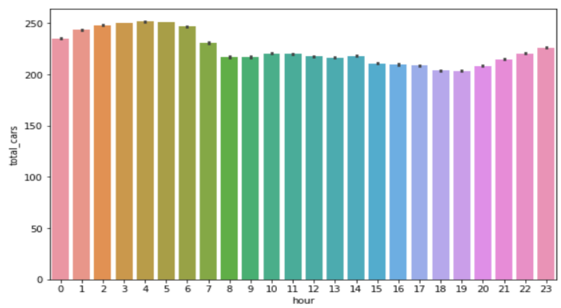


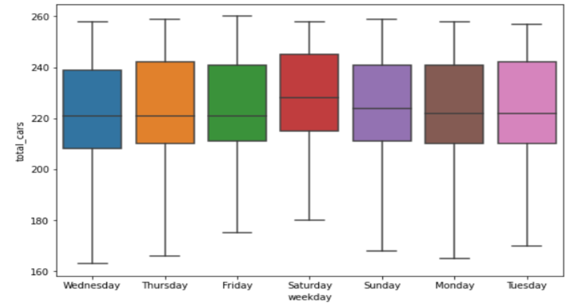
Using the above assumption we analyzed the usage rate at a particular timestamp.
The usage rate is given by : (total number of cars - available cars) / total number of cars.

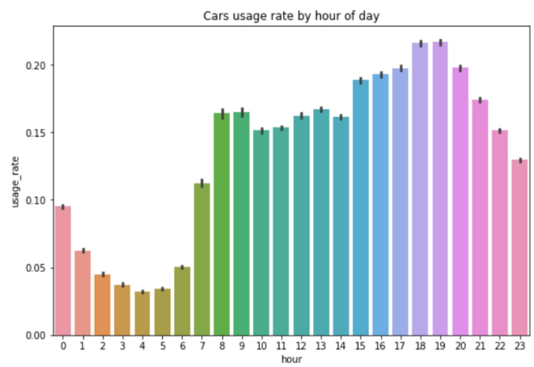
As per the analysis the average usage is 25% and can go up to 35% on some peak business days and can be as low as 5%-0%.

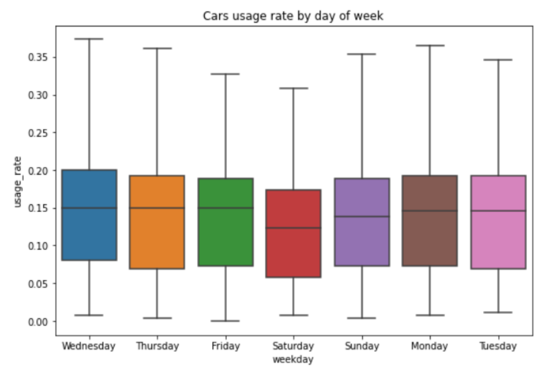


Additionally, analysis has been done on the total number of cars used in an hour of the day and also the day which has maximum car usage. The figure below says that the cars are used most during the afternoon and evening, while the cars are utilized least from midnight to very early morning.
We also found that Saturday is the most demanding day of the week. We then analyzed the usage rate by the hour of day and usage rate by day of the week using the above given usage rate formula.







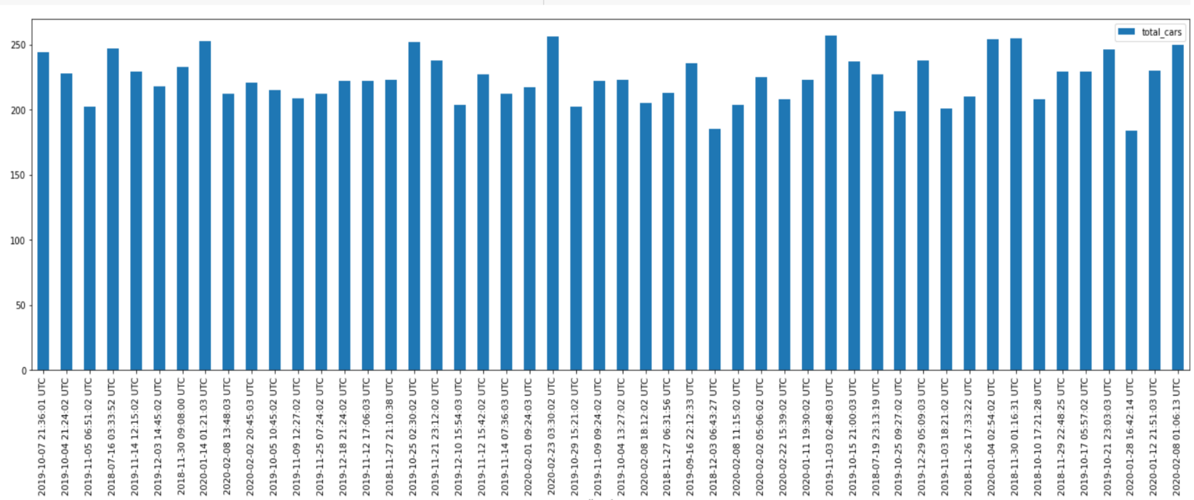


2. Outcomes and Visualization of the project :

A. Output1 : Predict the availability of cars at a certain parking location

Below is an image wherein we plotted a graph of timestamps vs the total number of cars available in parking lots during that time period.This was done directly from the dataset we had and plotting it in a form of bar graph using the 2 columns, timestamp and total_cars from our dataset.

```
timeVsCars = pd.DataFrame(dataOfCarsByTimestamp, columns=["timestamp", "total_cars"])
timeVsCars.plot(x="timestamp", y="total_cars", kind="bar", figsize=(25, 7))
plt.show()
```



Then we have created data polygons which enabled us to group the data by the parking lots and hence we were able to predict the number of cars available per parking lot! This was implemented using Geo-joins.

Below is an image showing how we created polygons from latitude and longitude of the data.

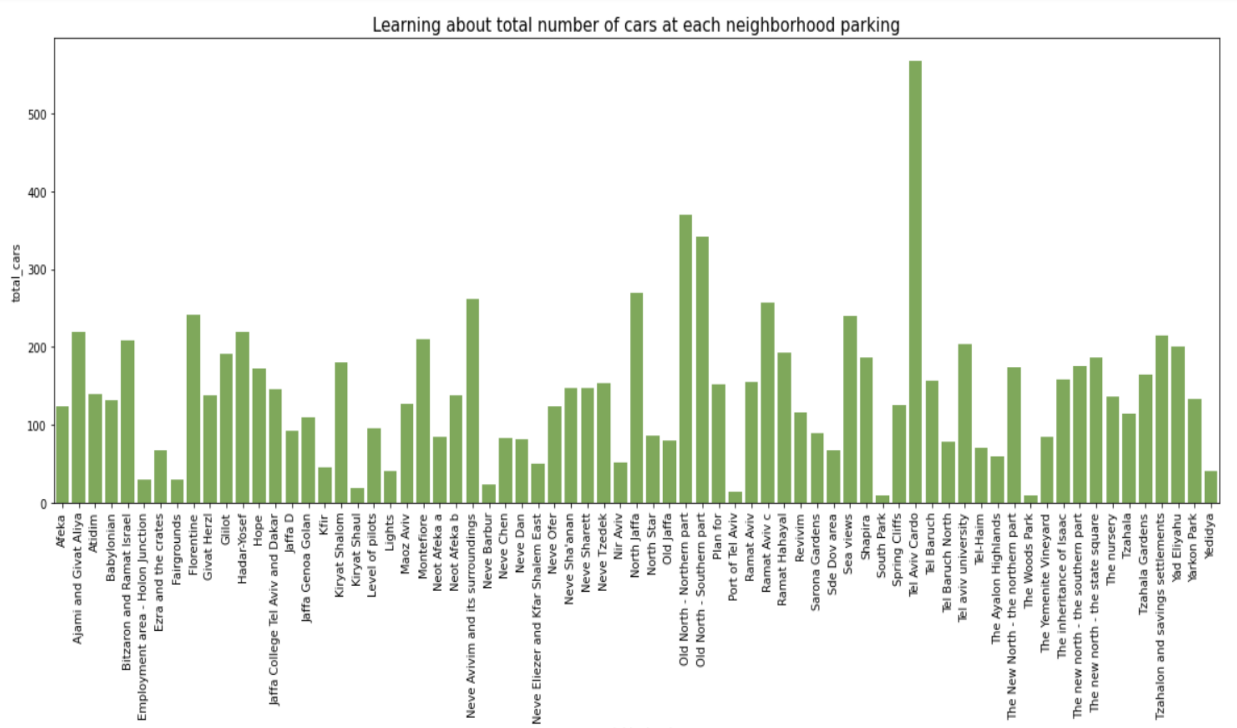
Creating polygon objects and saving to pandas dataframe

```python
def CreatePolygon(input):
    #print(input)
    polygonCreation = wkt.loads(input)
    polygonPoints = polygonCreation.exterior.coords[:]
    polygonPoints.append(polygonPoints[0])
    #print(polygonPoints)
    return Polygon(polygonPoints)
```

```python
dataByLocation['polygon'] = dataByLocation['area_polygon'].apply(CreatePolygon)
neighborhood_map = dataByLocation.set_index('neighborhood_name')['polygon'].to_dict()
dataByLocation
```

	oid_shchuna	neighborhood_id	neighborhood_name	date_import	Shape_Area	area_polygon	polygon
0	1	3	Sde Dov area	2015-10-06 08:36:43 UTC	1.307955e+06	POLYGON ((34.788389 32.123656, 34.788383 32.12...	POLYGON ((34.788389 32.123656, 34.788383 32.12...
1	2	2	Spring Cliffs	2015-10-06 08:36:43 UTC	1.288925e+06	POLYGON ((34.796243 32.130388, 34.795859 32.12...	POLYGON ((34.796243 32.130388, 34.795859 32.12...
2	3	7	Ramat Aviv c	2015-10-06 08:36:43 UTC	1.101582e+06	POLYGON ((34.808554 32.128712, 34.808641 32.12...	POLYGON ((34.808554 32.128712, 34.808641 32.12...
3	4	11	Tel aviv university	2015-10-06 08:36:43 UTC	1.459851e+06	POLYGON ((34.802354 32.118062, 34.80239 32.118...	POLYGON ((34.802354 32.118062, 34.80239 32.118...
4	5	6	North Star	2015-10-06 08:36:43 UTC	4.212298e+05	POLYGON ((34.783522 32.098636, 34.783493 32.09...	POLYGON ((34.783522 32.098636, 34.783493 32.09...

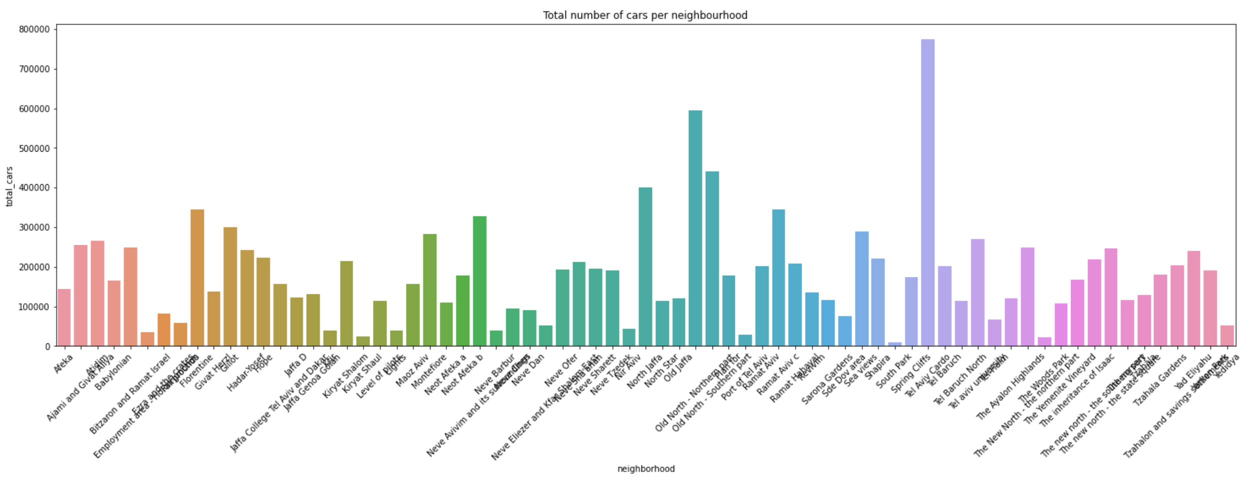
Then using geo joins we used 2 datasets(neighborhood parking spots dataset and car locations dataset) for a single analysis to find out the total number of cars available at a parking location and plotted it as below in the form of a bar graph.



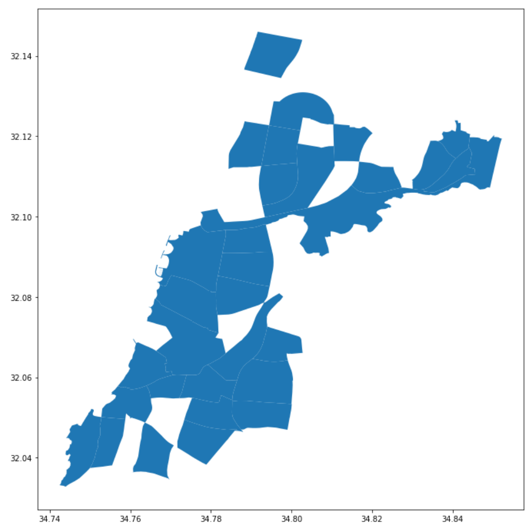
B. **Output 3: Analyzing profitable locations to enhance the usage of the service based on the dataset.**

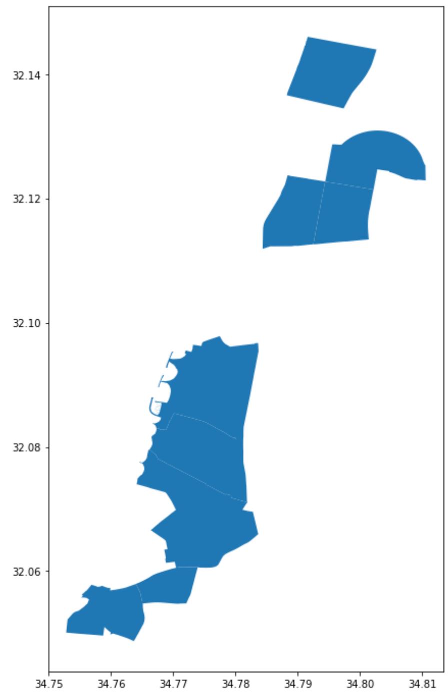
On gaining more experience with the geopandas gis functions, we learned to plot the city map outlines with the help of Polygon geometry using geo boundary data in maps. The Polygon object consists of the vertices of the polygons which practically is the boundary vertices on the map.

In order to print the areas on the map with 50% of usage and 25% of usage, we combined the parking location (primary) data and the map data of polygons. The sum of the total number of cars present all together at a parking location has been used as the metric to decide the usage of that parking location.



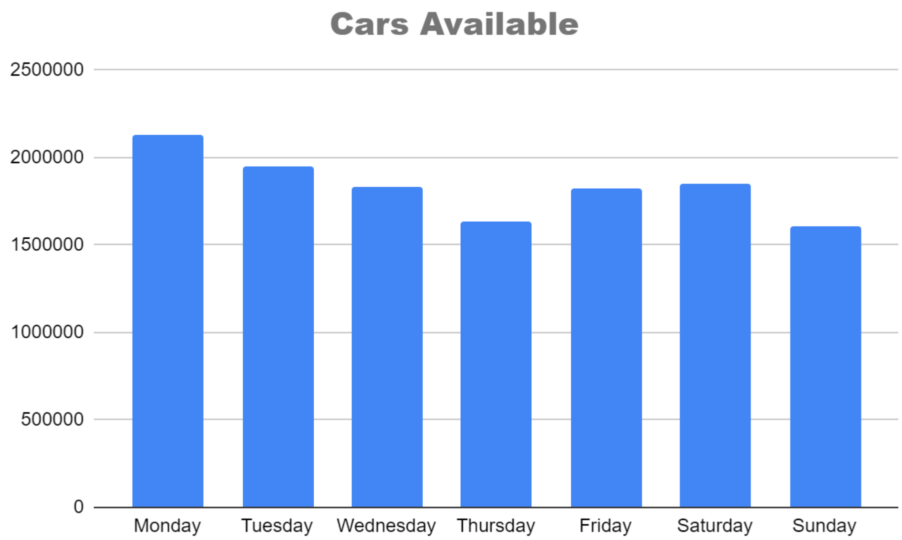
Here are the map locations of the neighborhoods with 50% and 25% of the usage.





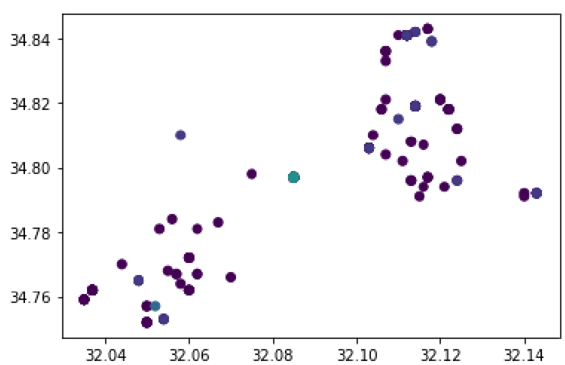
C. Output 4: Analyzing factors affecting the usage of service and Availability of cars.

Adding **weekday** data as a separate column in the dataframe.
Used datetime function to add the weekday column in the dataframe for each entry of the data. This helps us to know the availability of cars in a given parking location based on the day of the week information.

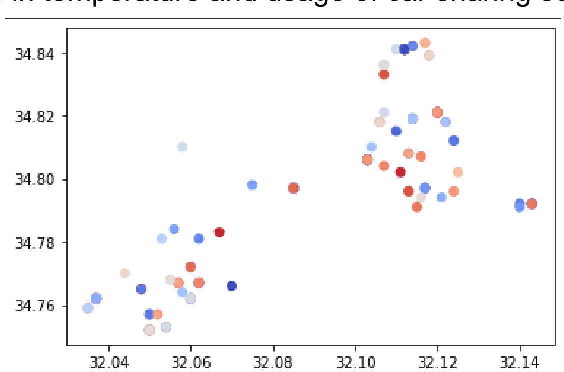


The given chart shows the total number of cars that were parked in the parking locations in the given day from all of the data collected. It can be inferred from the chart that cars are available mostly on Monday. And the least available for sharing on Thursday.

Adding **Temperature** information in the data
Installed meteostat library to add the weather data (temperature) as a separate column based on the parking location to see the similarities in temperature and usage of car sharing services



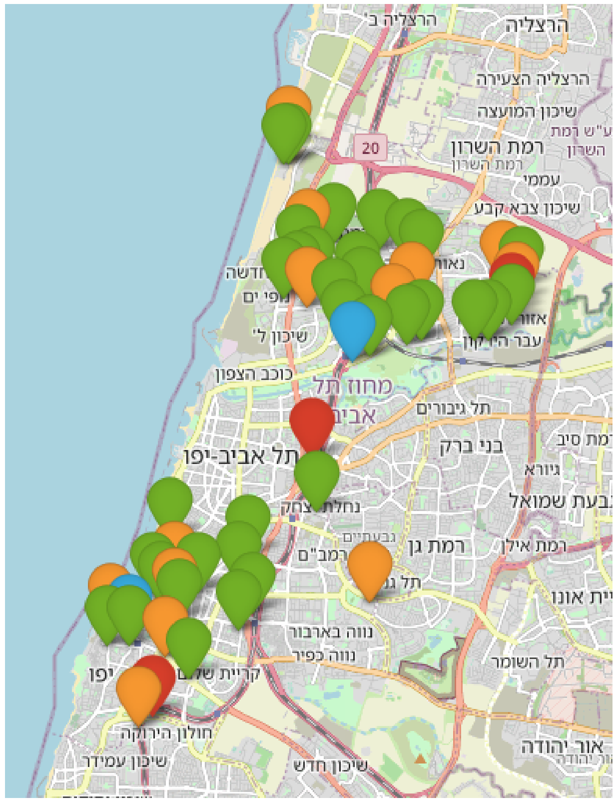
Number of cars



Temperature

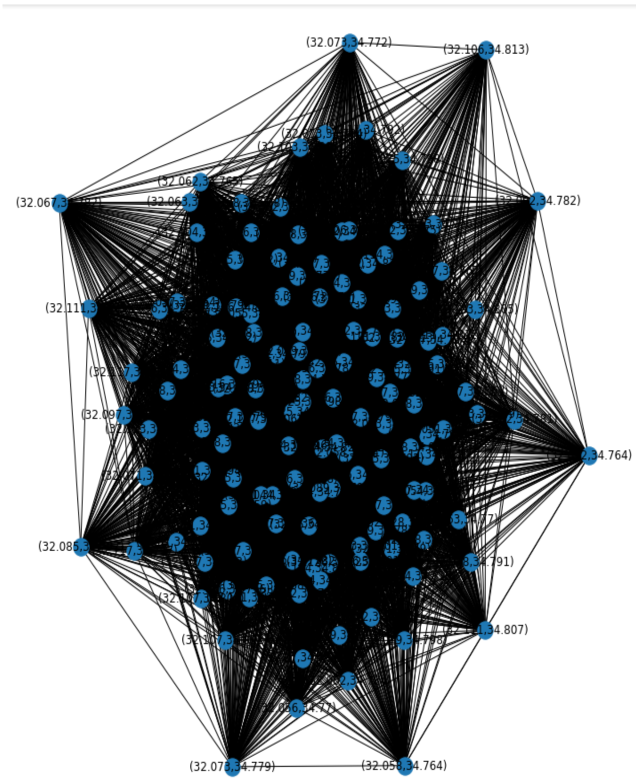
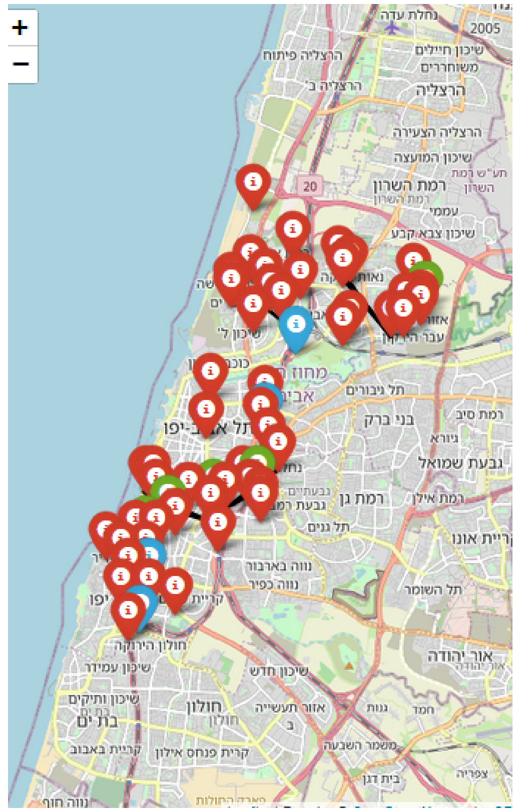
The above plot shows the relation between temperature and availability of cars at the given location. Higher temperature is represented by red shade and higher number of car availability is shown by lighter shade of blue.

Based on Location -
The availability of cars also depends on the location of the city of tel aviv. The graph shown below using Folium shows the availability of different numbers of cars. Blue Indicates That 6 or more cars are available at that location, green indicates 4 cars, orange indicates 5 cars are located at that location, and the rest are red.



D. **Output 2 and 5: At a given timeframe, predict the location where a car would go from a selected location; Identifying prime locations/areas for ride-sharing service for computing revenue models**

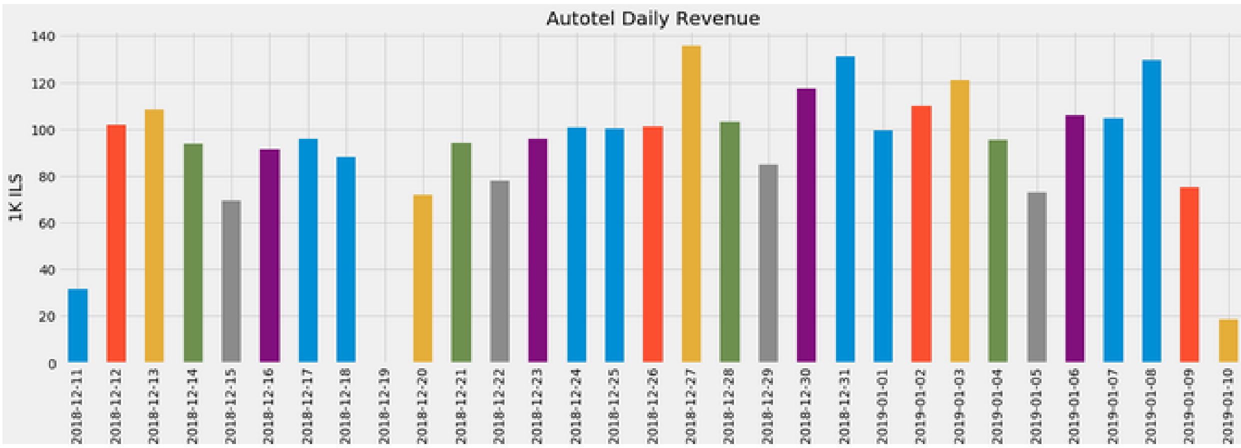
We have attempted to perform network analysis on the locations where cars are parked to identify significant locations and important pairs of locations. To implement this, we first determined all car parking locations. Further, some locations of interest are determined. This includes locations where 4 or more cars can be parked. Further, the next step is we collected all the trips that the cars have made. What this does is that it makes a count of all the trips made by any car between 2 parking locations. This is particularly important because we wish to know which car parks are more important. Then using that trip information, it is possible to build directed and undirected graphs.

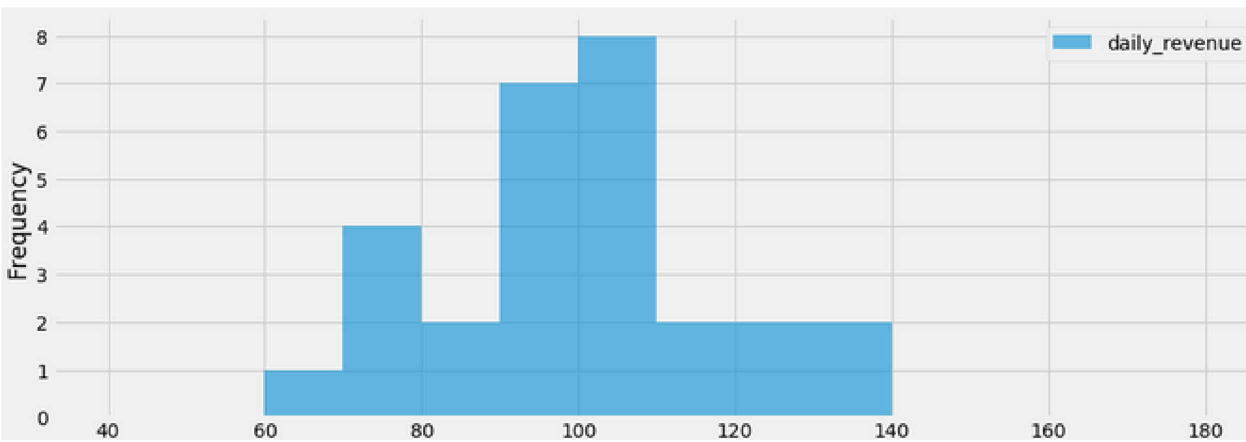
 

These graphs yield different information about the network. The number of trips made will be used as the edge weight between nodes during analysis. Then using that trip information, it is possible to build directed and undirected graphs. Below, a typical network type graph is shown. The outlier can quickly be identified and analyzed.

Trip	Count
('(32.114,34.842)', '(32.118,34.839)')	11611
('(32.063,34.782)', '(32.07,34.764)')	9611
('(32.063,34.782)', '(32.067,34.767)')	8876
('(32.107,34.833)', '(32.122,34.818)')	6988
('(32.067,34.767)', '(32.07,34.764)')	6879
('(32.111,34.802)', '(32.113,34.799)')	6634
('(32.066,34.786)', '(32.075,34.801)')	6531
('(32.107,34.833)', '(32.12,34.821)')	6363
('(32.066,34.786)', '(32.07,34.795)')	6279
('(32.066,34.776)', '(32.07,34.766)')	6174

The highly connected routes can be shown on a map to help figure out why these locations have such high connectivity. The top 10 routes of high connectivity are shown. We assume all rides cost 1.3 ILS per minute. The average daily revenue is estimated to be around 100K ILS (ILS refers to Israel currency). Since we have top parking locations and most used routes between pairs of mostly used parking locations, pricing models for these locations can be updated by the ride-sharing company for these routes rather than having standard rates for all car sharing rentals.





3. Conclusion :
In order for this service to be reliable and profitable, the service provider needs a clear picture of the supply and demand of the cars at respective parking locations. Moreover, for such a business the revenue is generated per mile a car is driven. If a car sits ideally in some remote parking area and there is deficiency of cars at busier locations, then the business is losing on potential profits.

During our project implementation and data analysis we successfully predicted the below outcomes:

1. Predict the availability of cars at a certain parking location in a specific time frame
2. At a given timeframe, predict the location where a car would go from a selected location
3. Find profitable locations by analyzing the dataset to enhance the usage of the service based on the dataset
4. Determine and Analyze the External factors that are affecting the usage of the service
5. Separate price model for separate areas

These outcomes can be helpful for taking good business decisions and making it more profitable. Some visualizations were also plotted using matplotlib in python. There could be more such use cases and factors affecting the business which need to be analyzed to propose better strategies for improving the service.
From the data visualization it has been noticed that at many locations, more cars are currently located at many parking locations which are not being used. This can be improved by relocating the said cars to the locations which are almost always out of cars.

From the Network information of the trips that cars make, cars can be smartly located at locations which are most traveled and future location prediction of cars can be made with higher confidence of further future.

The timeline information about the availability of cars helps to understand the usage of car sharing service throughout the year and day in the city of Tel-Aviv.

4. Future Work :

One of the challenges in this project is that the dataset has many records, moreover it is a geospatial dataset. In order to process this dataset big data techniques are required.

Another challenge would be to combine different datasets based on geospatial features which require spatial join algorithms to generate a clear picture of the data. In order to perform clear analysis on this dataset, we have to figure the combination logic for different outcomes.

Parallel computing in the cloud can be used to face this challenge and the calculation can be done faster. Using recent and everyday data from the government website of Tel-Aviv city will give us better understanding and better predictions about the availability of the car at specific locations.

Using neural networks or other machine learning algorithms for better prediction of locations and availability can be made, which is nearly impossible right now since we have billions of data points and computation would be extremely expensive.

5. References :

Related Work along with Taxonomy has been submitted in detail in a separate word document.