notebook_taxi101

February 3, 2022

1 Taxi 101

This report is divided into following sections. 1. Data Acquisition 2. Data Cleaning 3. Feature Engineering 4. Exploratory Data Analysis 5. Pricing engine w.r.t. market 6. Market analysis as a Use cases for a taxi launch - Experimental

1.1 Data Acquisition

Data is obtained from https://data.cityofchicago.org/Transportation/Taxi-Trips-2021/9kgb-ykyt. For simplicity I have downloaded data and put in data folder. This data acquisition can also be done via API (Not used here). There are more than 3 million rows, so it gonna take some time to download it.

For detailed data columns explanation, see the above link. In summary it has trip details, such as pickup time and location (area and coordinates), fare details, taxi company name, payment mode.

```
Time taken to load csv file :  13.630910873413086 secs
```

1.2 Data Cleaning

1. It starts with dropping the 'na' rows from those columns which will be crucial in an attempt to do valuable analysis or make informed decision. So dropping 'na' values from ['Taxi ID','Fare','Trip Seconds','Trip Miles','Dropoff Centroid Location','Pickup Centroid Location','Pickup Community Area','Dropoff Community Area'])
2. Second will be dropping the non-relevant columns like census tract. There are lot of na values, so dropping whole column makes sense.
3. Filtering Outliers using 3 sigma rule on Fare Values. See https://en.wikipedia.org/wiki/Standard_deviation
4. Finally Removing '0' fare values trips

As you can see below, we drop ~15% percent of the total available data

```
Percentage Drop = 14.91 %
```

1.3 Feature Engineering

1. Extract time feature like hour (0-23), minute(0-59), day number (1-31), month(1-12) , day of the week(0-6), week of the year from trip timing details.
2. Converting pickup and dropoff community area from float to int type.

3. Clustering Fare into three clusters. Why three you may ask? because of elbow mention for ‘kmeans’ and also it will be very clear in further EDA.
4. Adding a flag if the start trip time was midnight or not. Midnight here is 00:00 to 05:59.
5. Adding a flag if the start trip time was holiday or not as per # https://www.chicago.gov/city/en/narr/misc/city-holidays.html
6. Finally calculating distance between pickup and dropoff coordinated.(This may take time to process for all of the rows ~ 7-10 minutes on my machine)

```
100%|                    | 3359357/3359357 [04:30<00:00, 12402.51it/s]
```

1.4 Exploratory Data Analysis

This section is going to be a bit long. For EDA I have mostly referred to pie chart (https://en.wikipedia.org/wiki/Pie_chart), Kernel density estimation (https://en.wikipedia.org/wiki/Kernel_density_estimation), and Box Plot (https://en.wikipedia.org/wiki/Box_plot), along with GeoSpatial visualization.

1.4.1 Chicago Taxi Market Analysis

How big was the 2021 market in terms of revenue In total the taxi market for city of Chicago in 2021 was about ~$61 million

```
w.r.t. Fare only : $61,647,994.51
w.r.t. Fare and Tips : $67,637,960.95
For this report, wherever price is mentioned, it will refer to Fare only.
```

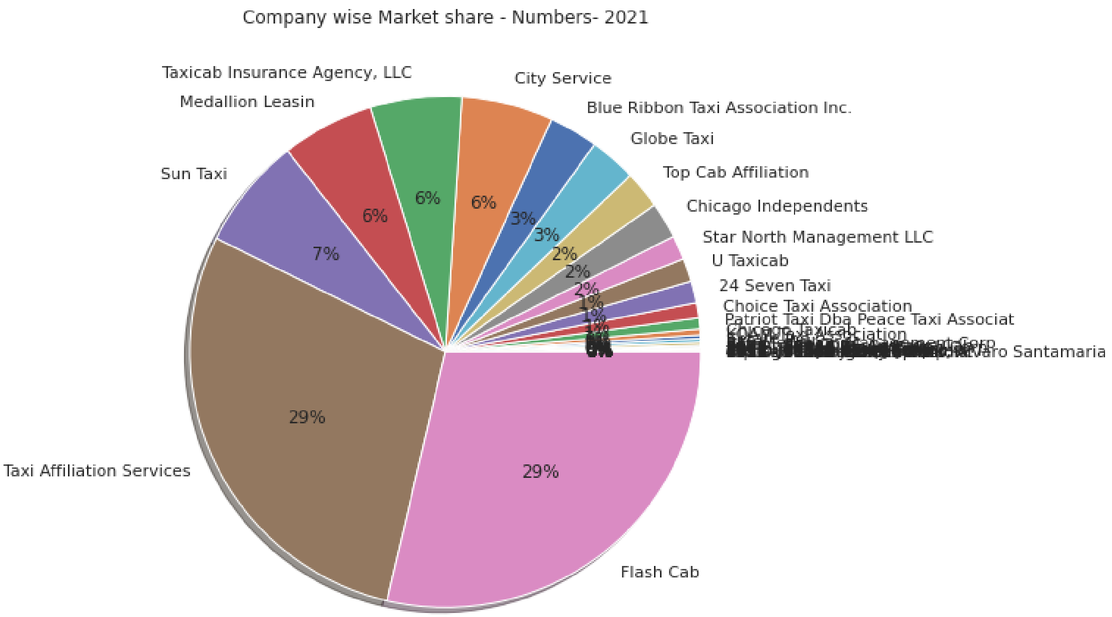
How many taxi trips were taken in 2021 In cleaned dataset there are 3,359,357 trip recorded

```
Total Taxi Trips : 3359357
```

1.4.2 Chicago Taxi Company’s Market Analysis

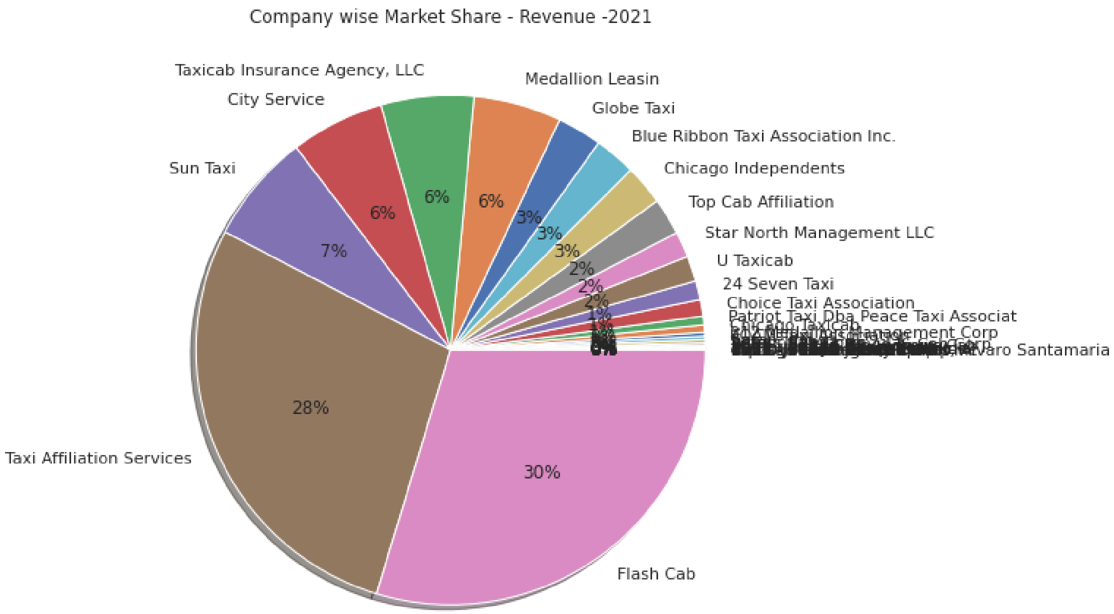
Company wise market share in terms of numbers for 2021 In terms of number of trips, ’Taxi Affiliation Services” and “Flash Cab” stand out as market leaders each with 29% share respectively.

Note: Please ignore the overlap labels on the pie chart, didn’t had time to do cosmetic changes the viz.



```
Top 5 Company wise market share in terms of numbers for 2021
Company
Flash Cab                          962804
Taxi Affiliation Services          962466
Sun Taxi                           237833
Medallion Leasin                   197068
Taxicab Insurance Agency, LLC      194535
dtype: int64
```

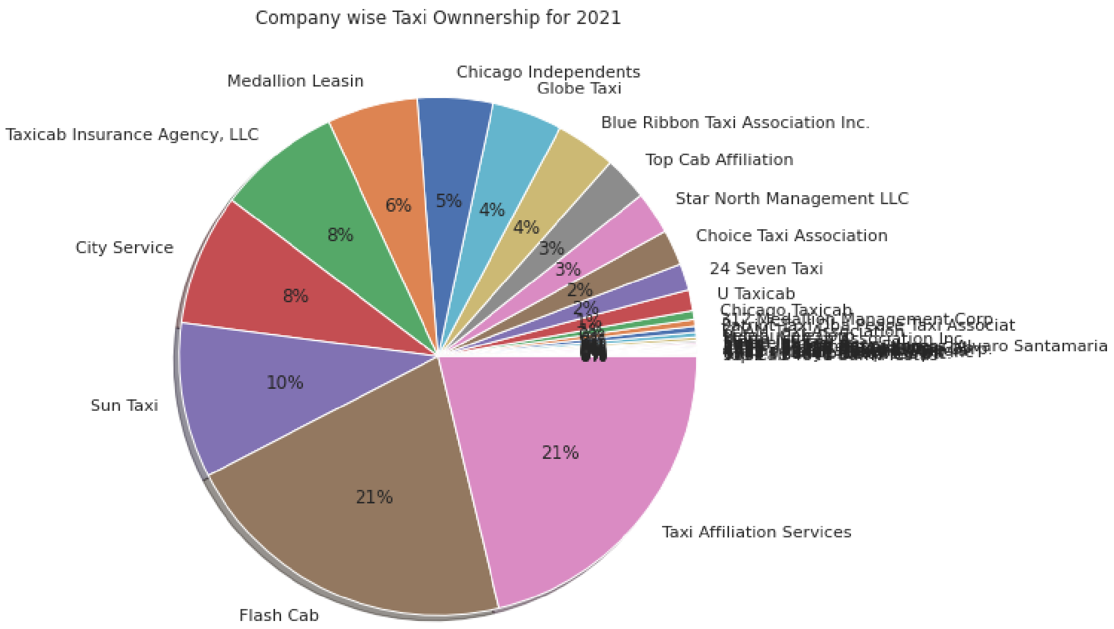
Company wise market share in terms of Revenue for 2021 In terms of total revenue, again 'Taxi Affiliation Services" and "Flash Cab" stand out as market leaders each with 28% and 30% share respectively.



```
Top 5 Company wise market share in terms of revenue for 2021.
Company
Flash Cab                        18307244.05
Taxi Affiliation Services        17243912.83
Sun Taxi                          4285506.01
City Service                      3669680.68
Taxicab Insurance Agency, LLC     3641440.22
Name: Fare, dtype: float64
```

Company wise Taxi Ownership for 2021 So what about number of taxi and company mapping. or how many taxi are with which companies.

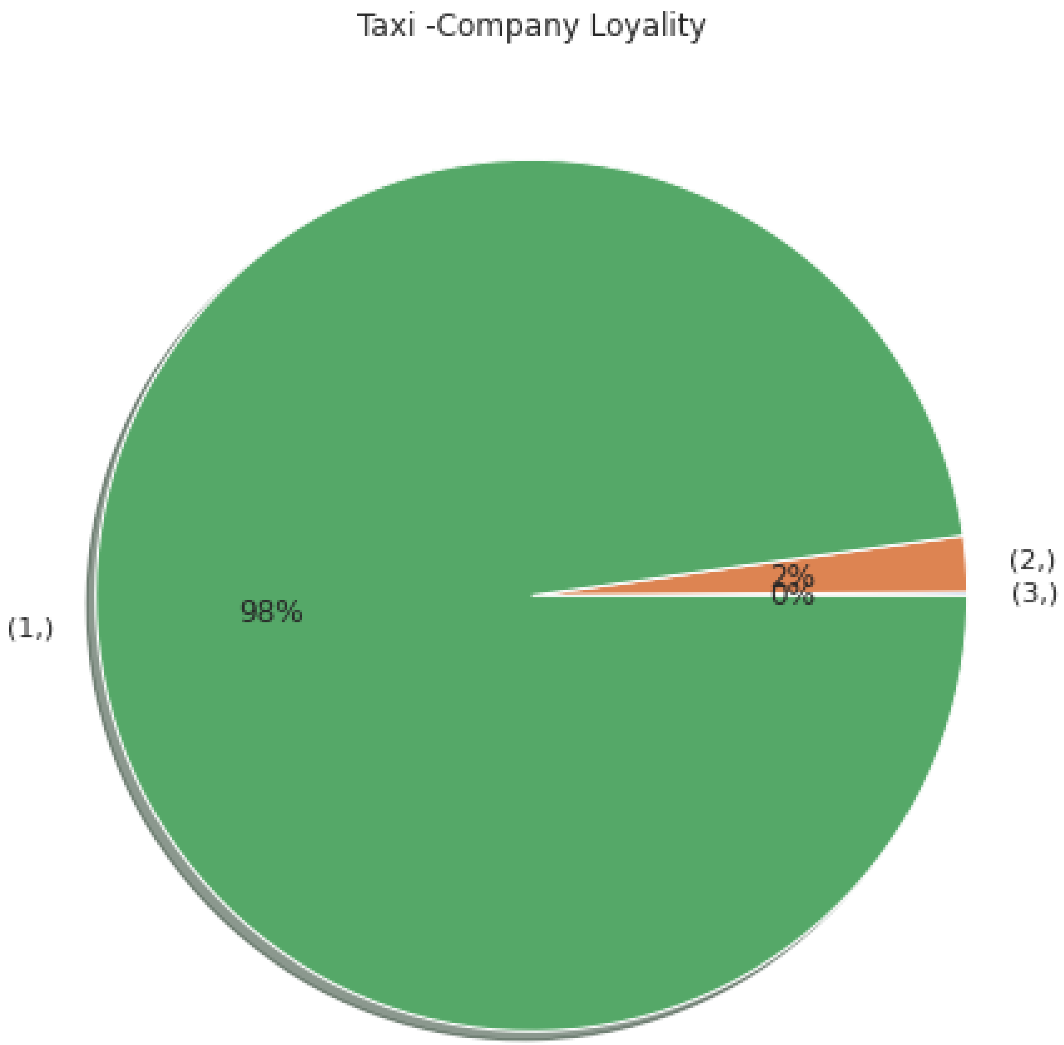
Again you can see, “Flash Cab” and “Taxi Affiliation Services” as market leaders, with both ~ 500 taxis, about ~20% for each of them.



```
Top 5 Company wise Taxi Ownership for 2021.
                               Taxi ID
Company
Taxi Affiliation Services          508
Flash Cab                          507
Sun Taxi                           232
City Service                       197
Taxicab Insurance Agency, LLC      185
```

Taxi to Company Loyality for 2021 This one is interesting. Since there are many companies,even though few of them are the market leaders. What does it say in terms of taxi to company loyalty. The assumption here, as per the database (https://data.cityofchicago.org/Transportation/Taxi-Trips-2021/9kgb-ykyt) is that the each Taxi will have the unique "Taxi ID" and will not change from company to company.

So about 98% percent of taxi's are with the single company. This also gives us the total number of taxis in Chicago for 2021 as "2340"

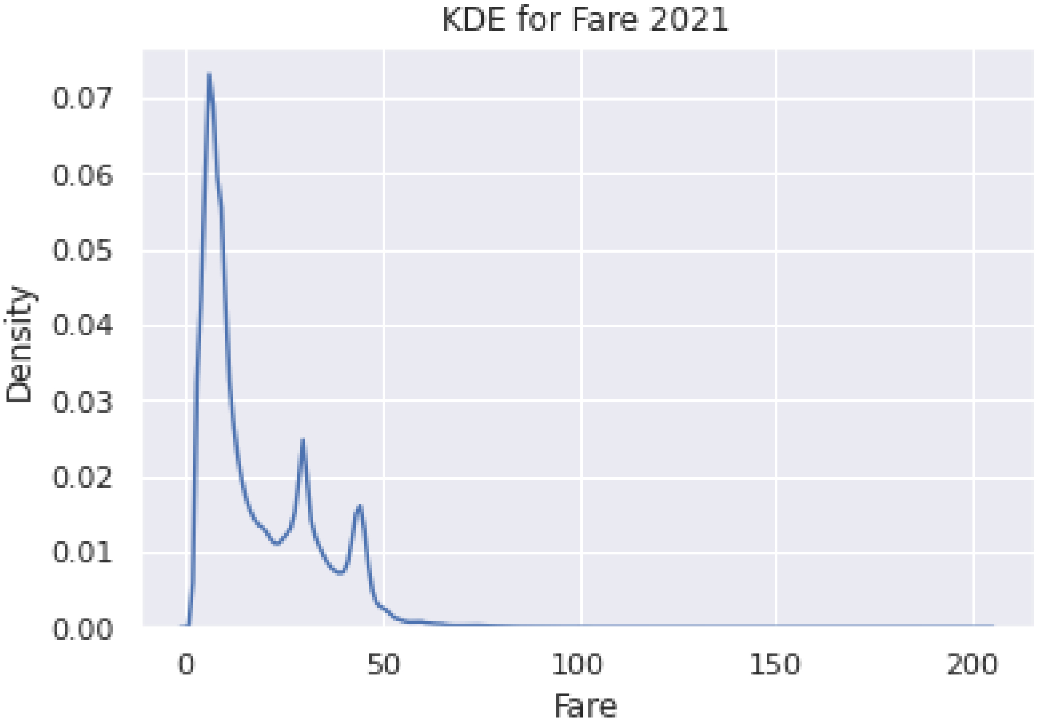


```
Taxi to Company loyality for 2021.
Company
1          2289
2            48
3             3
dtype: int64
Total number of Taxis = 2340
```

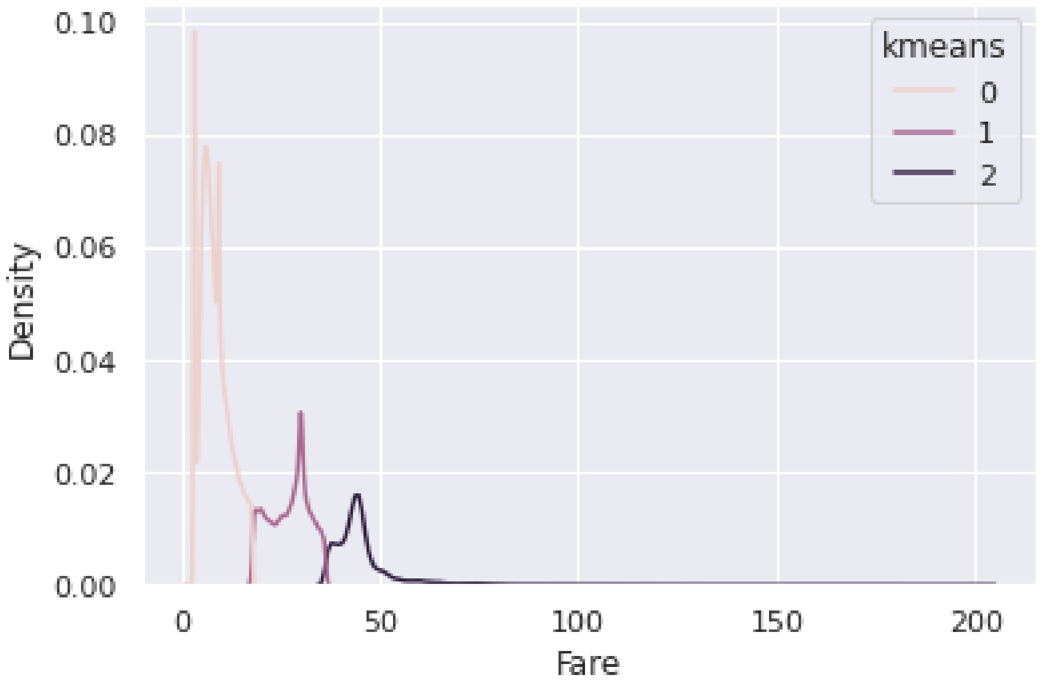
1.4.3 Fare Distribution and analysis

Next we are going to dive into ‘Fare’ analysis, its distribution, clustering and respective market size and revenue.

Kernel Density estimation for Fare As we can see, there clearly peaks at three places.(useful in k-means clustering)

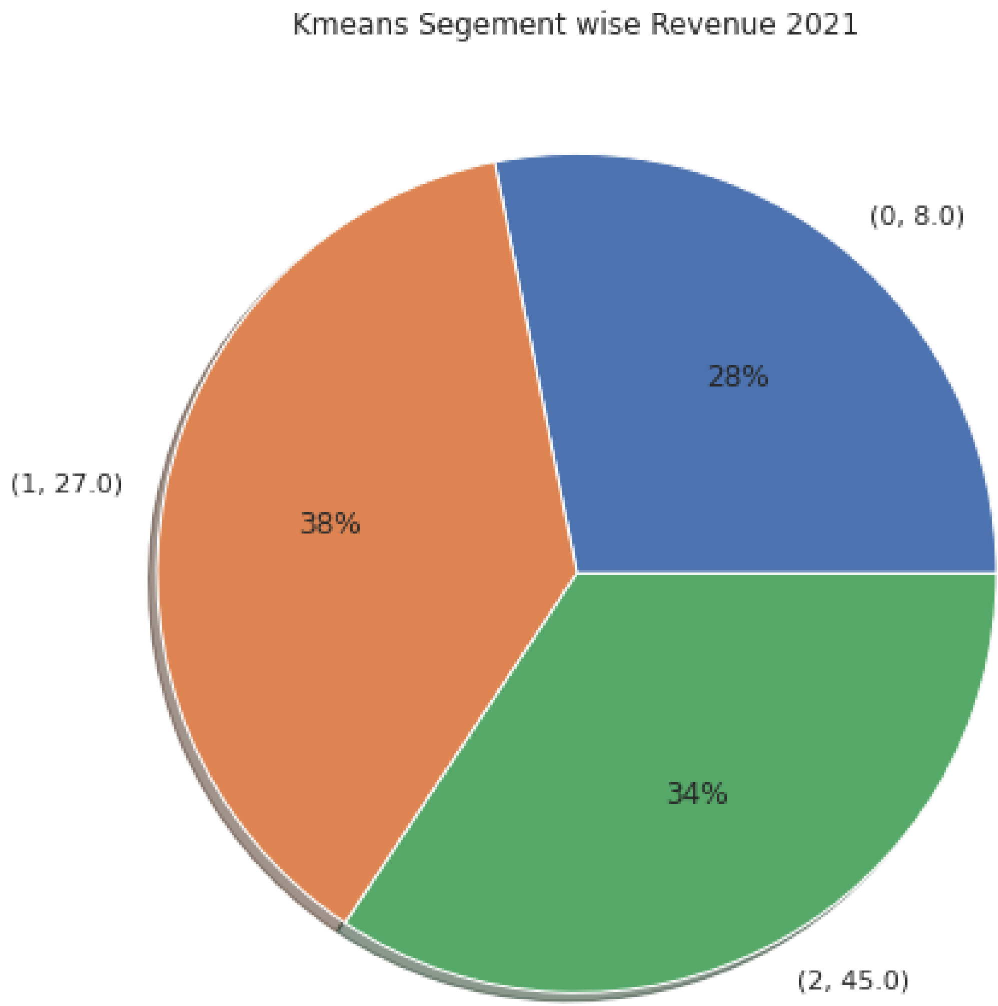


Kernel Density estimation for Fare with Kmeans feature As part of the feature engineering we also ran with 'kmeans' clustering on the Fare with cluster size of '3'.(Obtain from the above distribution and using elbow method https://en.wikipedia.org/wiki/Elbow_method_(clustering))



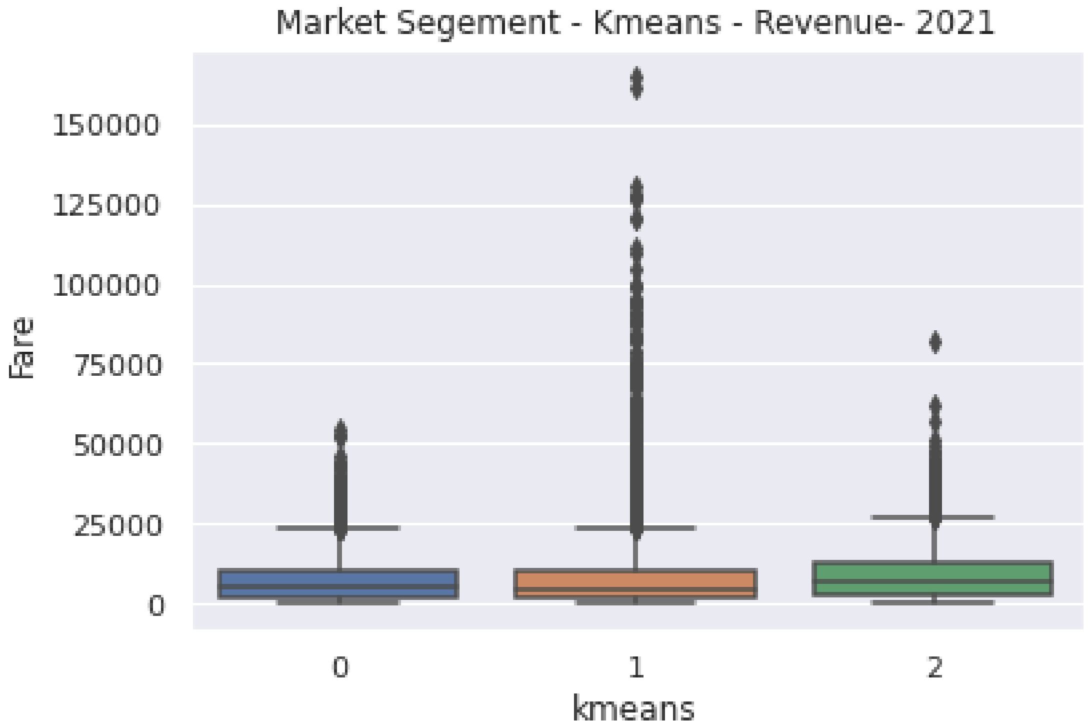
```
Kmeans Details (cluster_label, cluster_value) = [(0, 8.5), (1, 27.2), (2, 45.4)]
```

Market share w.r.t. Kmeans feature in terms revenue With respect to 'Kmeans' clustering, market can now be seen as comprising of the segment, low price ~ 8 dollars, mid price ~ 27 dollars and high price ~ 45 dollars. So what about revenue generated from this segments. As you can see in terms of revenue, mid-price segment generates about 38 percent of revenue

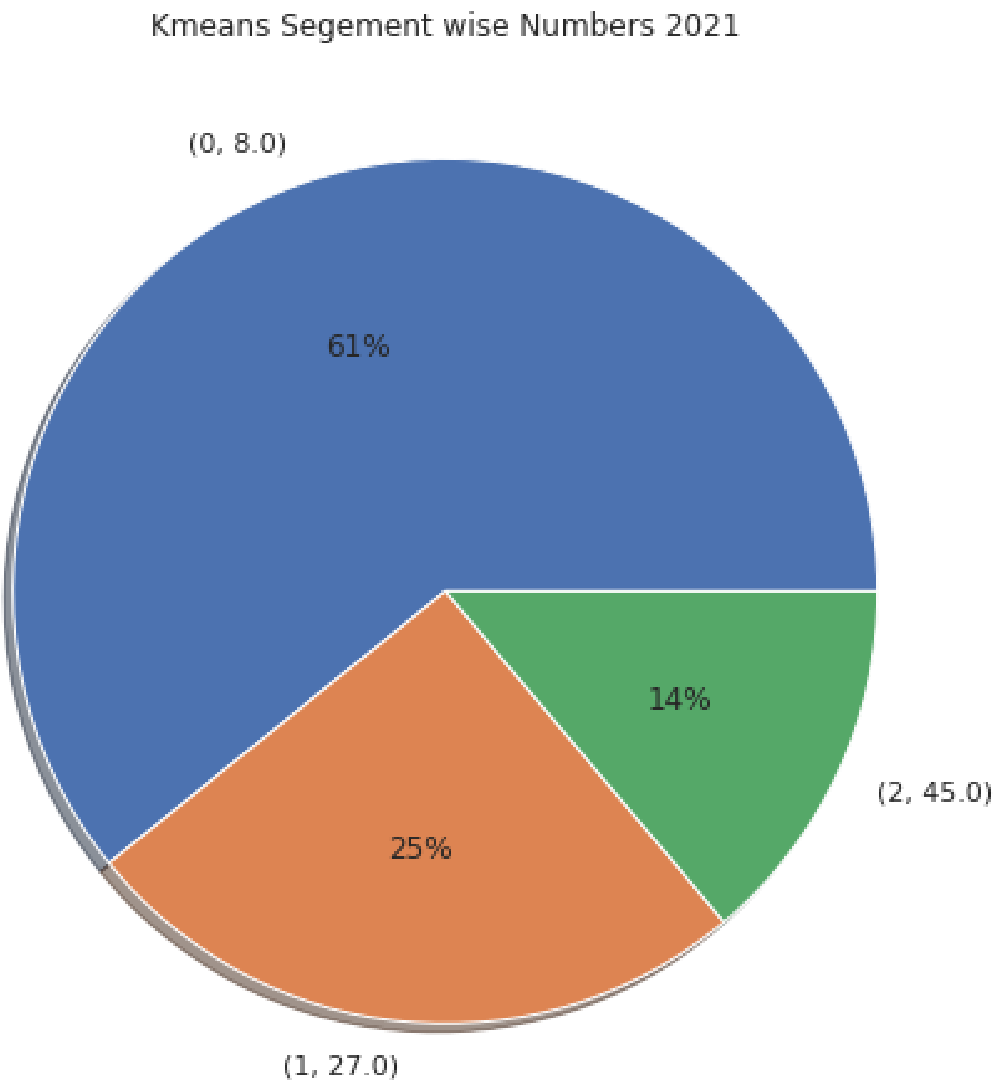


```
Market share w.r.t. Kmeans feature in terms revenue
kmeans
1    23152401.01
2    21147130.83
0    17348462.67
Name: Fare, dtype: float64
```

Market share w.r.t. Kmeans feature in terms revenue - Box Plot Box plot for price segment in terms of revenue

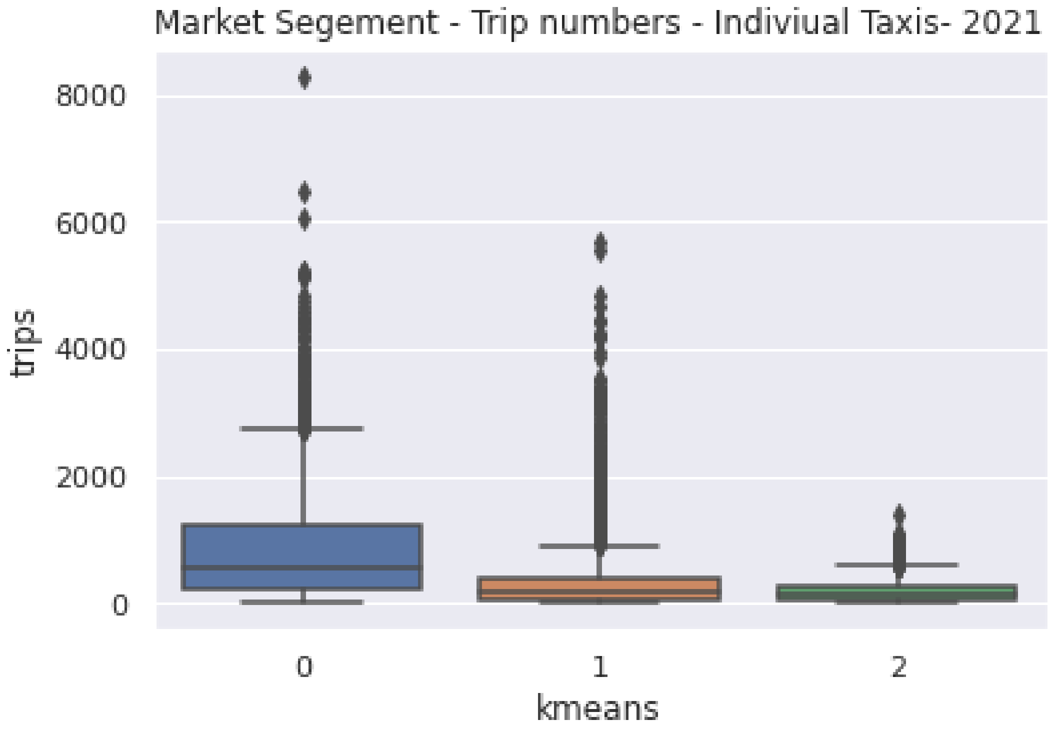


Market share w.r.t. Kmeans feature in terms trip numbers What about market share w.r.t. to 'kmeans' but in terms of trip numbers.As you can see in terms of revenue, low-price segment constitutes of about 61 percent of total number of trips.



```
Market share w.r.t. Kmeans feature in terms trip numbers
kmeans
0    2043154
1     850705
2     465498
dtype: int64
```

Market share w.r.t. Kmeans feature in terms trip numbers - Individual Taxi Wise - Box Plot This is Box plot for price segment in terms of trip numbers but grouped on individual Taxis. e.g. for low price range for each taxi, mean value of trips is about ~873 trips per annum.



```
Segment wise Mean number of Trips
                 trips
kmeans
0          873.516032
1          378.427491
2          207.348775
```

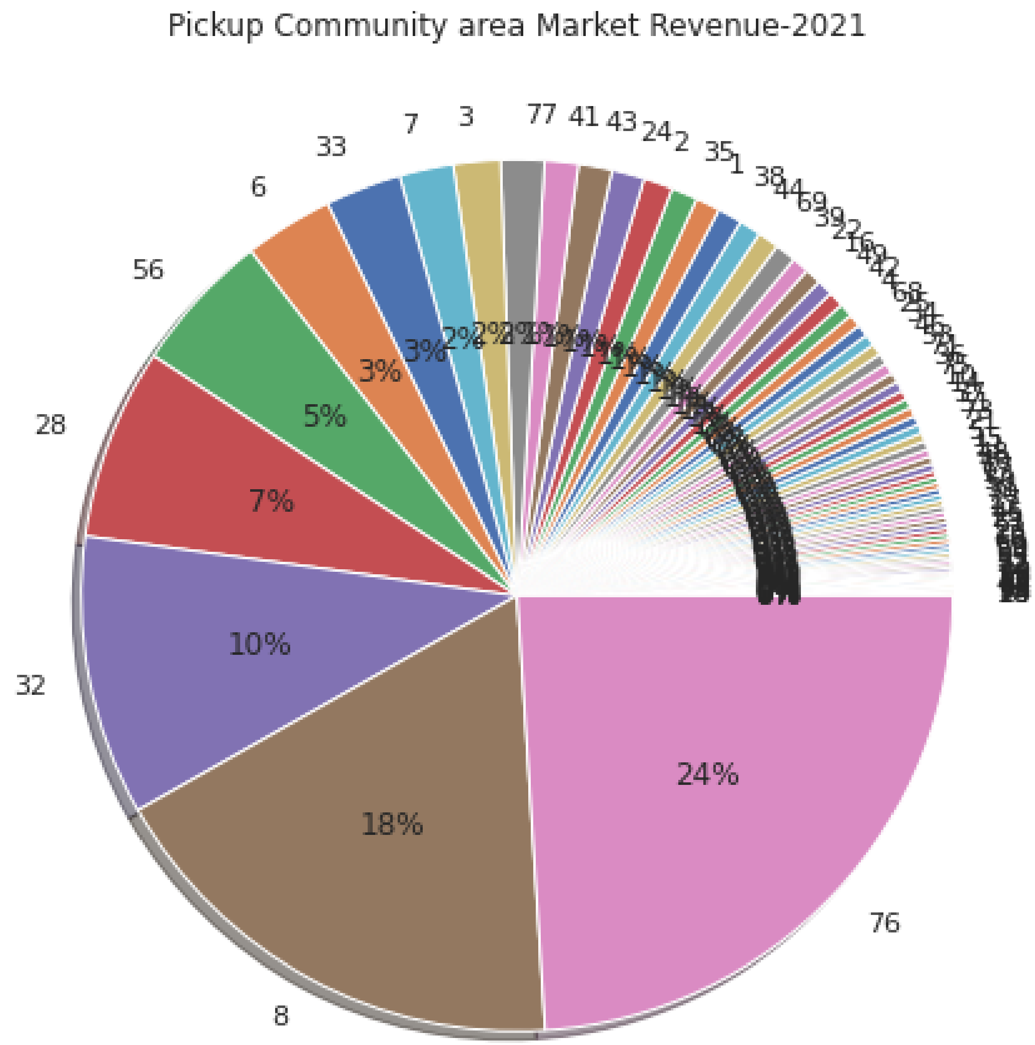
1.4.4 Location Distribution and analysis

This section is going to dive into 'Location' analysis, its distribution, clustering and respective market size and revenue.

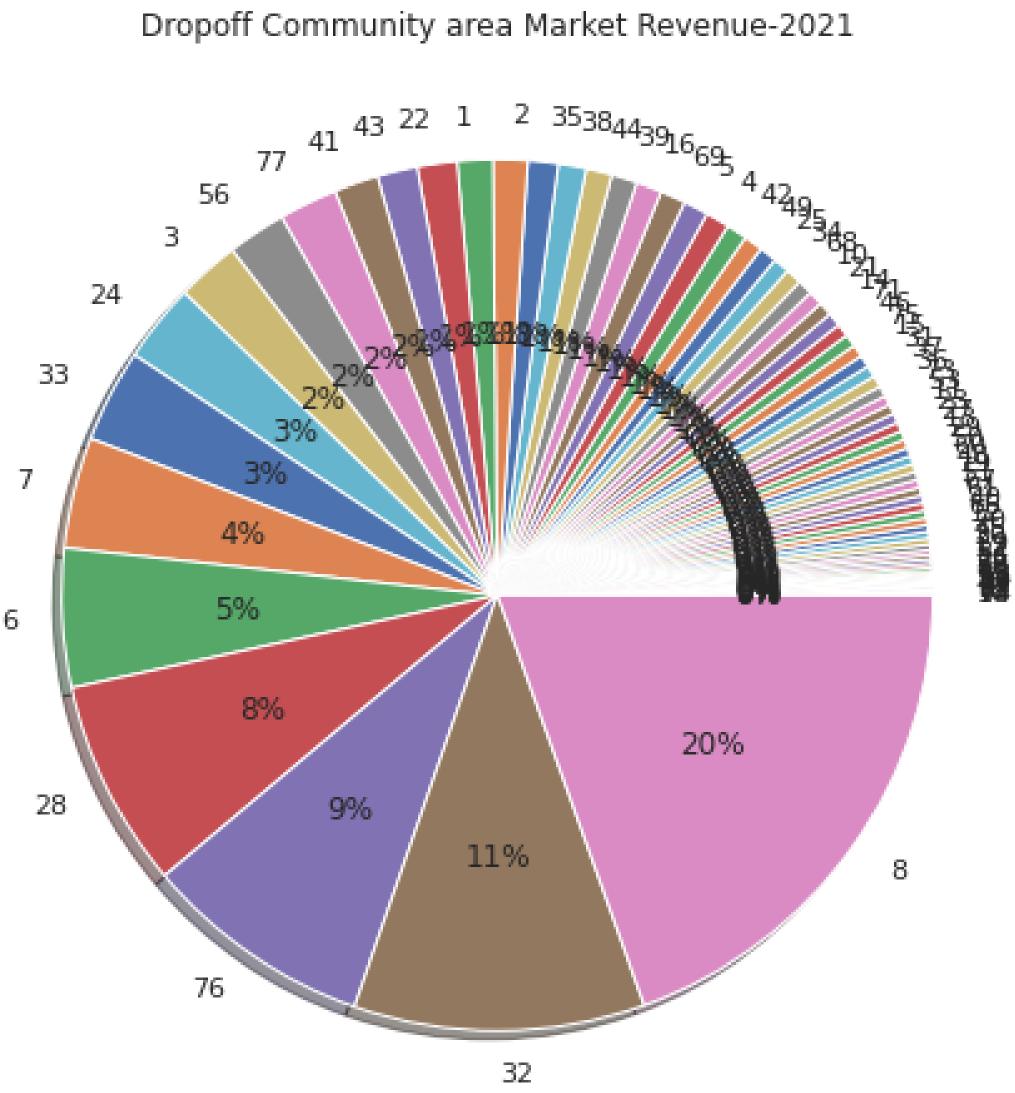
Market share w.r.t. Location feature in terms revenue With respect to Location, the data is given presented as number. I am going to use the same from here onwards. I am going to analyze for both pickup and dropoff location area.

As you can see in terms of revenue,for pickup, community area number "76" generates about 24% of revenue., followed by community area number "8" and "32" with 18% and 10% respectively.

Similarly for dropoff, community area number "8" generates about 20% of revenue., followed by community area number "32" and "76" with 11% and 9% respectively.


Pickup Community area Market Revenue-2021

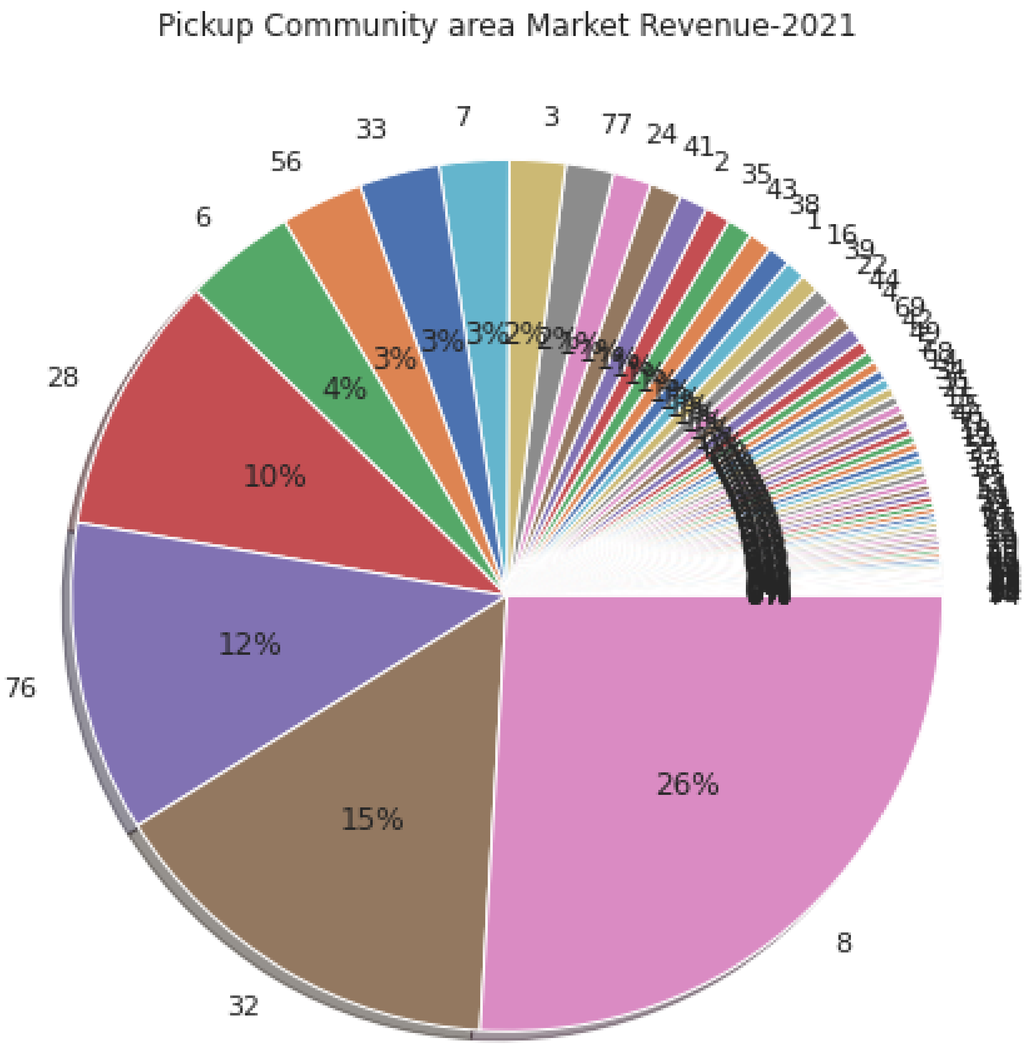
```
Market share for Pickup Community area in terms of Revenue-2021
Pickup Community Area
76    14781794.96
8     10986203.94
32     6394379.89
Name: Fare, dtype: float64
```



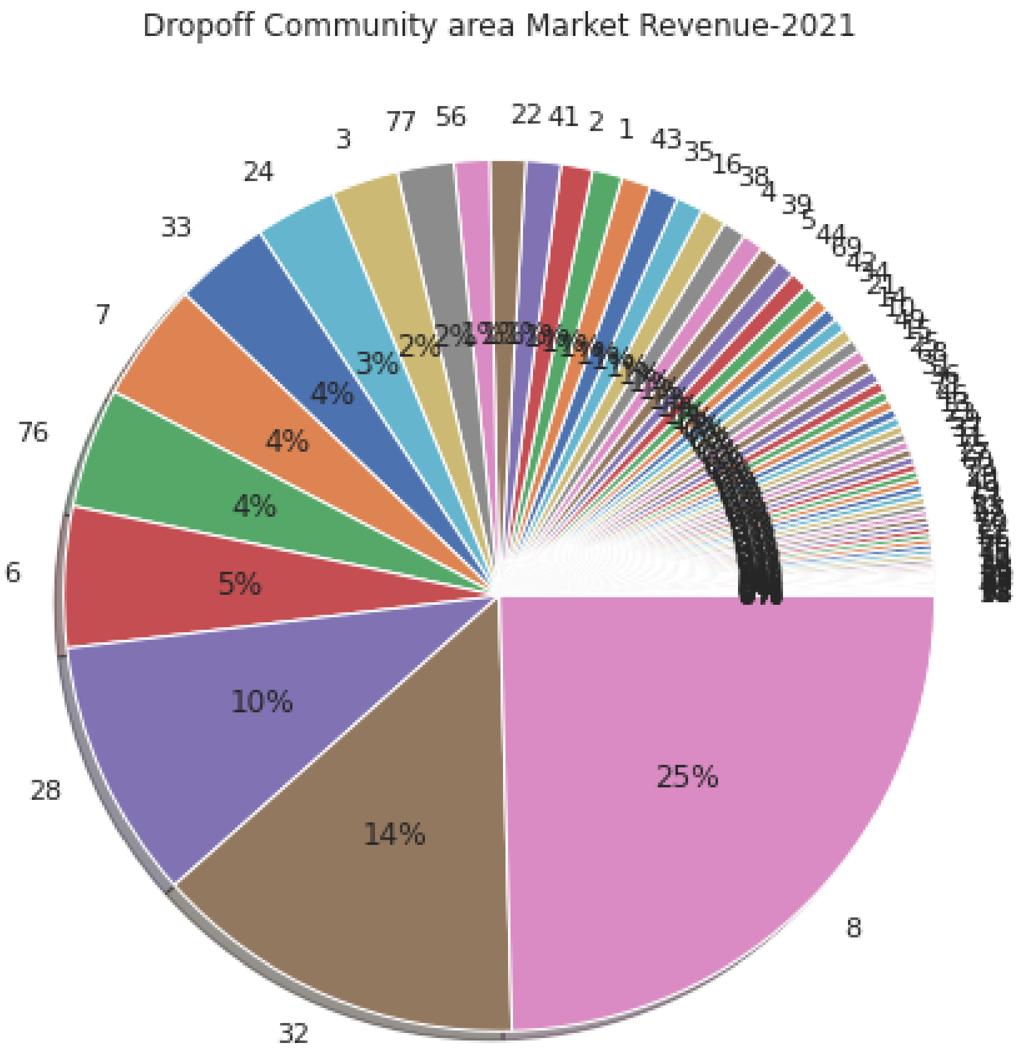
```
Market share for Dropoff Community area in terms of Revenue-2021
Dropoff Community Area
8      12035792.87
32      6655422.91
76      5259568.21
Name: Fare, dtype: float64
```

Market share w.r.t. Location feature in terms Numbers of trips As you can see below in terms of number of trip,for pickup, community area number “8” constitutes of about 26% of total number of trip, followed by community area number “32” and “76” with 15% and 12% respectively.

Similarly for dropoff, community area number “8” constitutes of about 25% of total number of trip, followed by community area number “32” and “28” with 15% and 10% respectively.

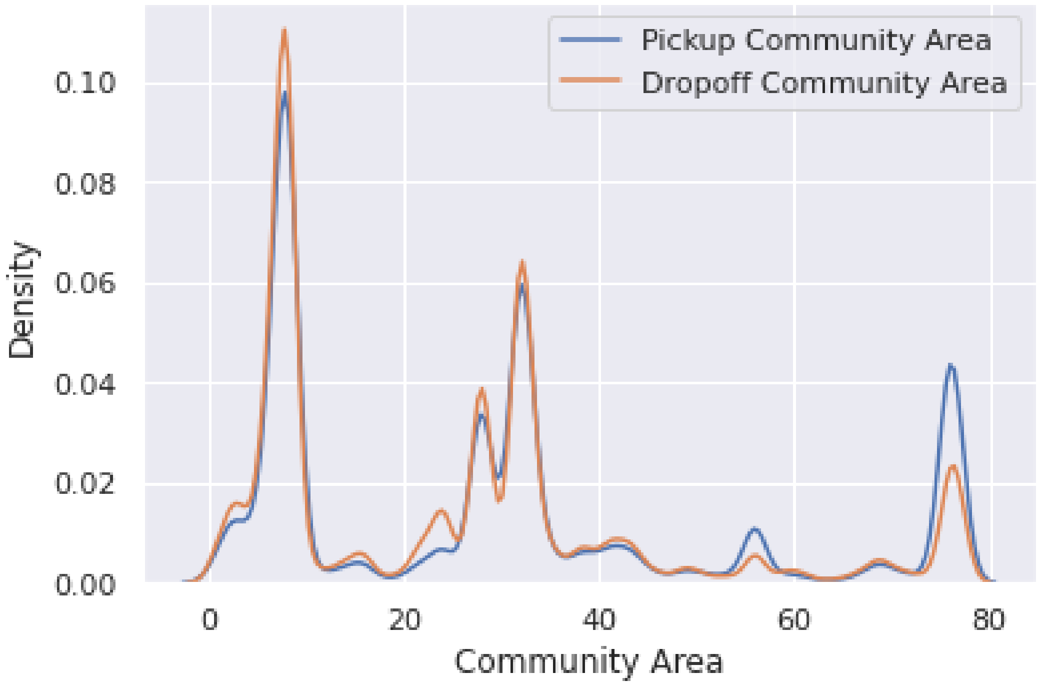


```
Market share for Pickup Community area in terms of Numbers-2021
Pickup Community Area
8     872794
32    509147
76    387525
dtype: int64
```

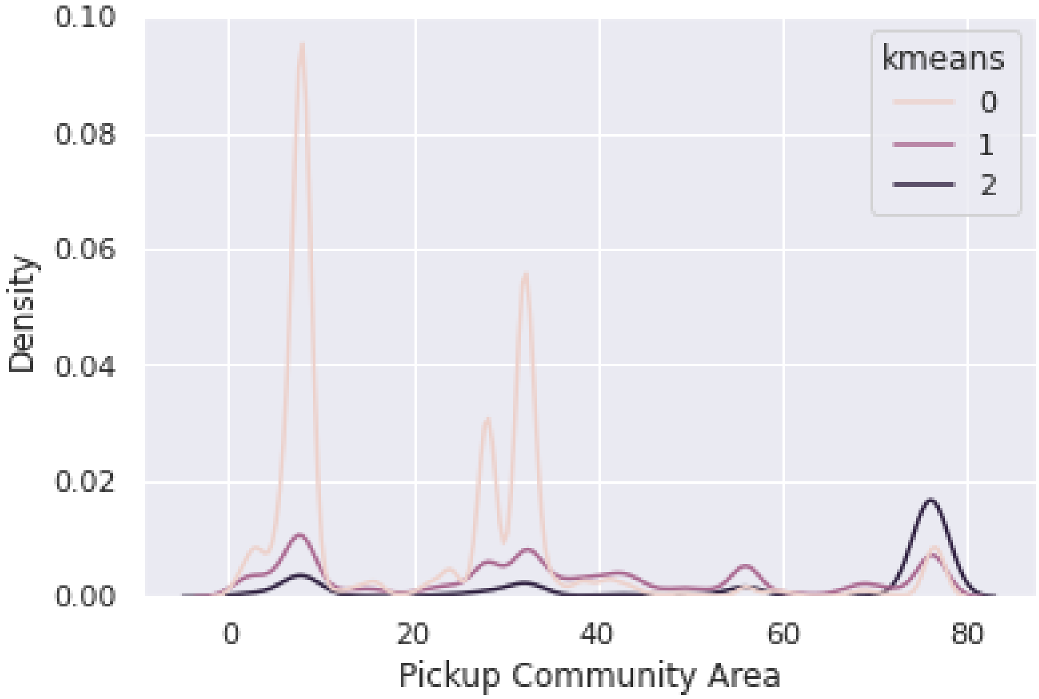


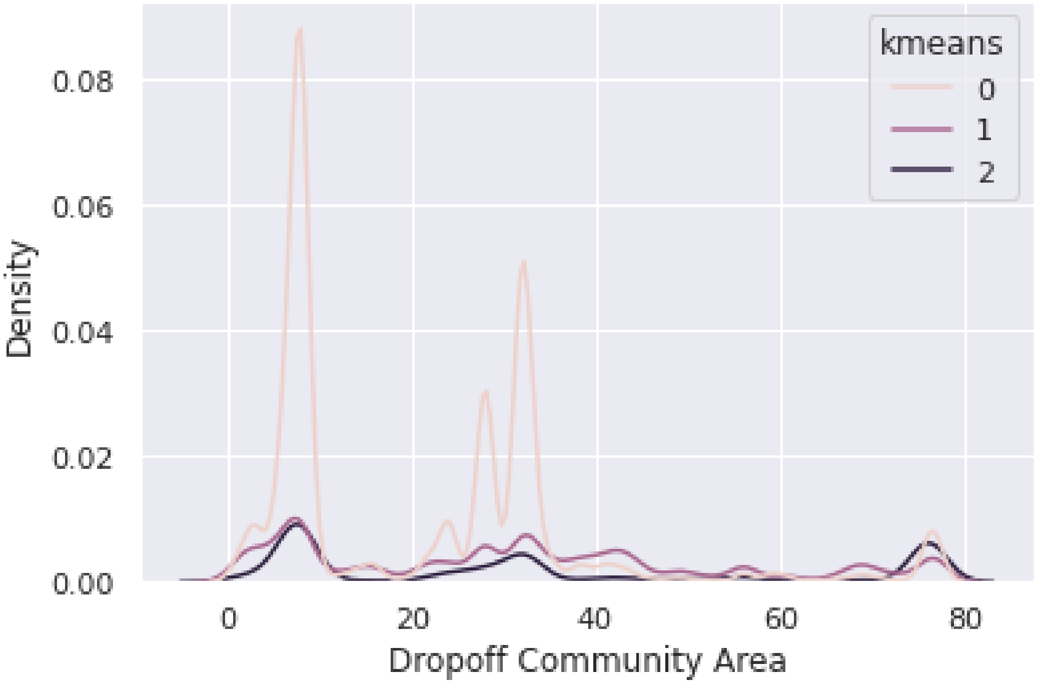
```
Market share for Dropoff Community area in terms of Numbers-2021
Dropoff Community Area
8     824587
32    465875
28    324399
dtype: int64
```

Market share w.r.t. Location feature in terms Numbers of trips - Kernel Density Estimation

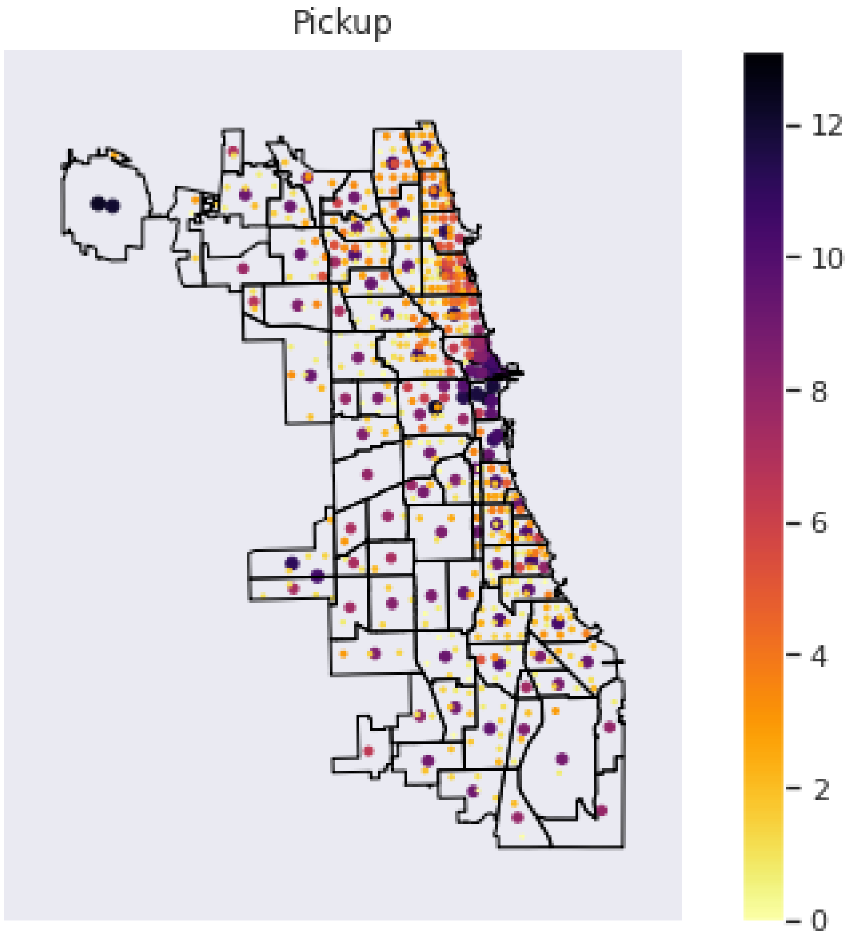


Market share w.r.t. Location feature in terms Numbers of trips and using kmeans information - Kernel Density Estimation

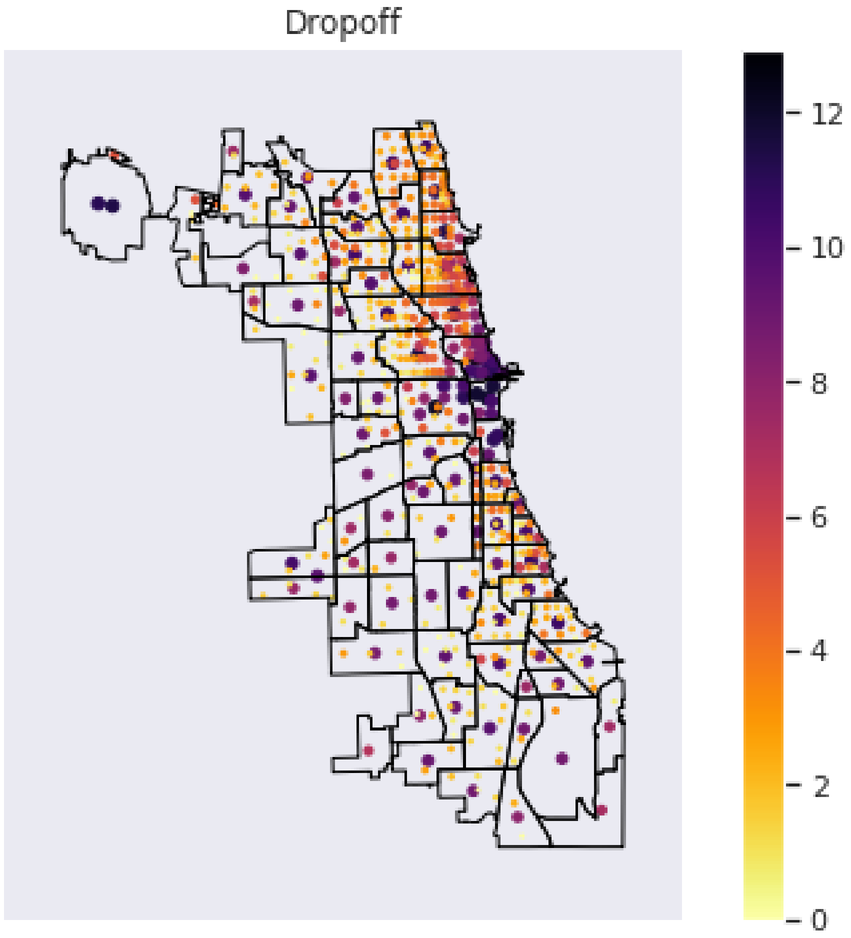




Market share w.r.t. Pickup Location feature in terms Numbers of trips - Geo Spatial Visualization Note: The numbers on the scale are in log range for better visibility



Market share w.r.t. Dropoff Location feature in terms Numbers of trips - Geo Spatial Visualization Note: The numbers on the scale are in log range for better visibility

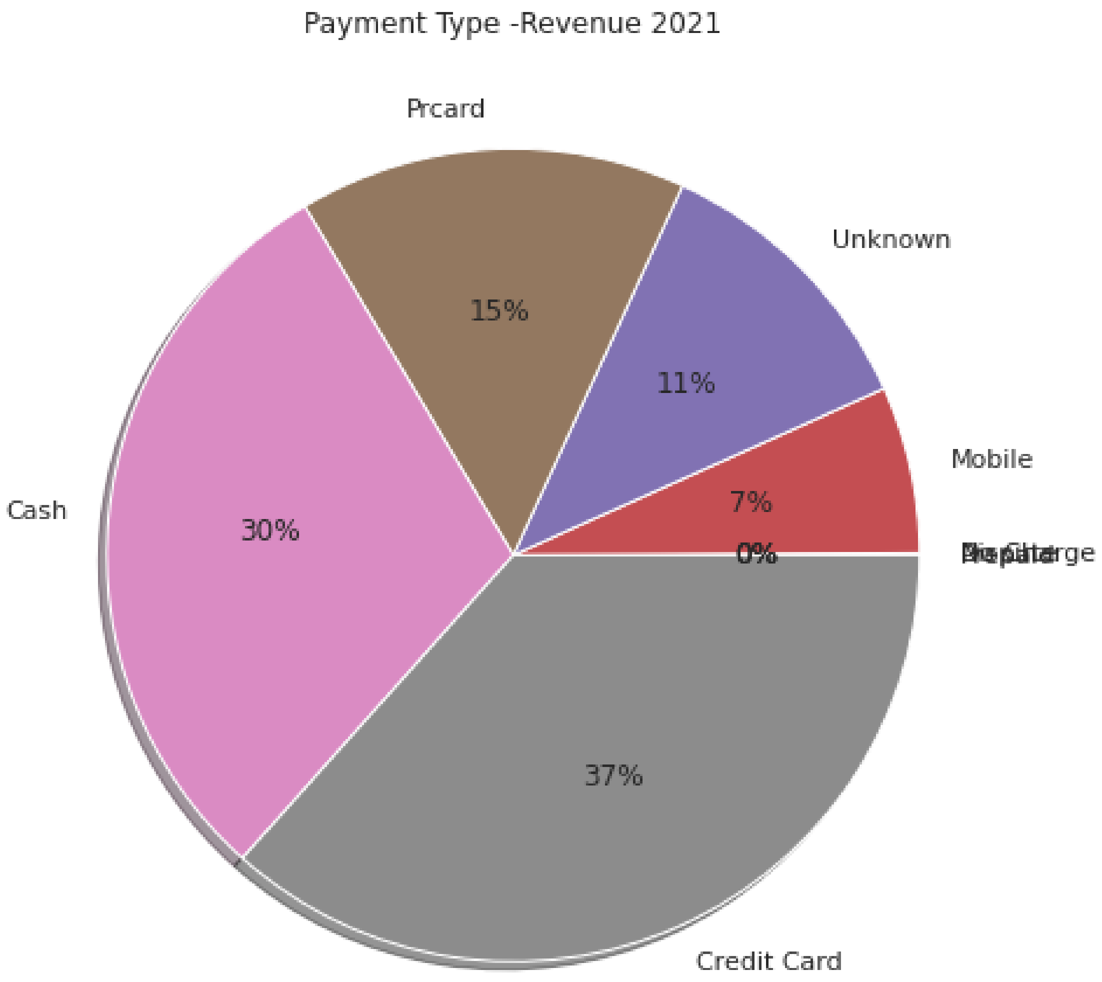


1.4.5 Payment Type Distribution and analysis

This section is going to dive into 'Payment Type' analysis, its distribution, clustering and respective market size and revenue.

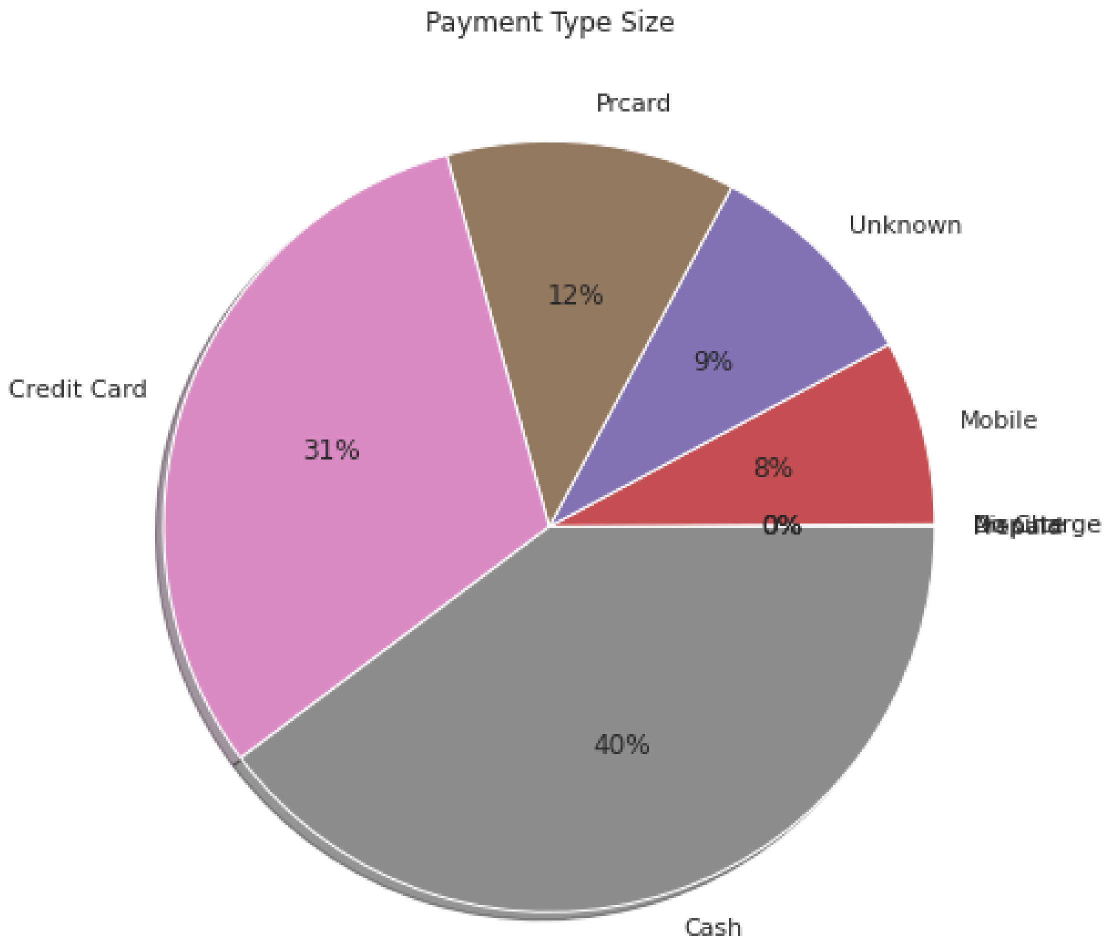
Market share w.r.t. Payment Type feature in terms revenue With respect to Payment Type, the data is available in the database.

As you can see in terms of revenue,for payment type, "Credit Card" generates about 37% of revenue, followed by "Cash" as 30%



```
Market share w.r.t. Payment Type feature in terms revenue-2021
Payment Type
Credit Card    22566969.07
Cash           18379902.23
Prcard          9504402.39
Name: Fare, dtype: float64
```

Market share w.r.t. Payment Type feature in terms of numbers As you can see in terms of revenue,for payment type, “Cash” constitutes of about 40% of total number of trips, followed by “Credit Card” as 31%



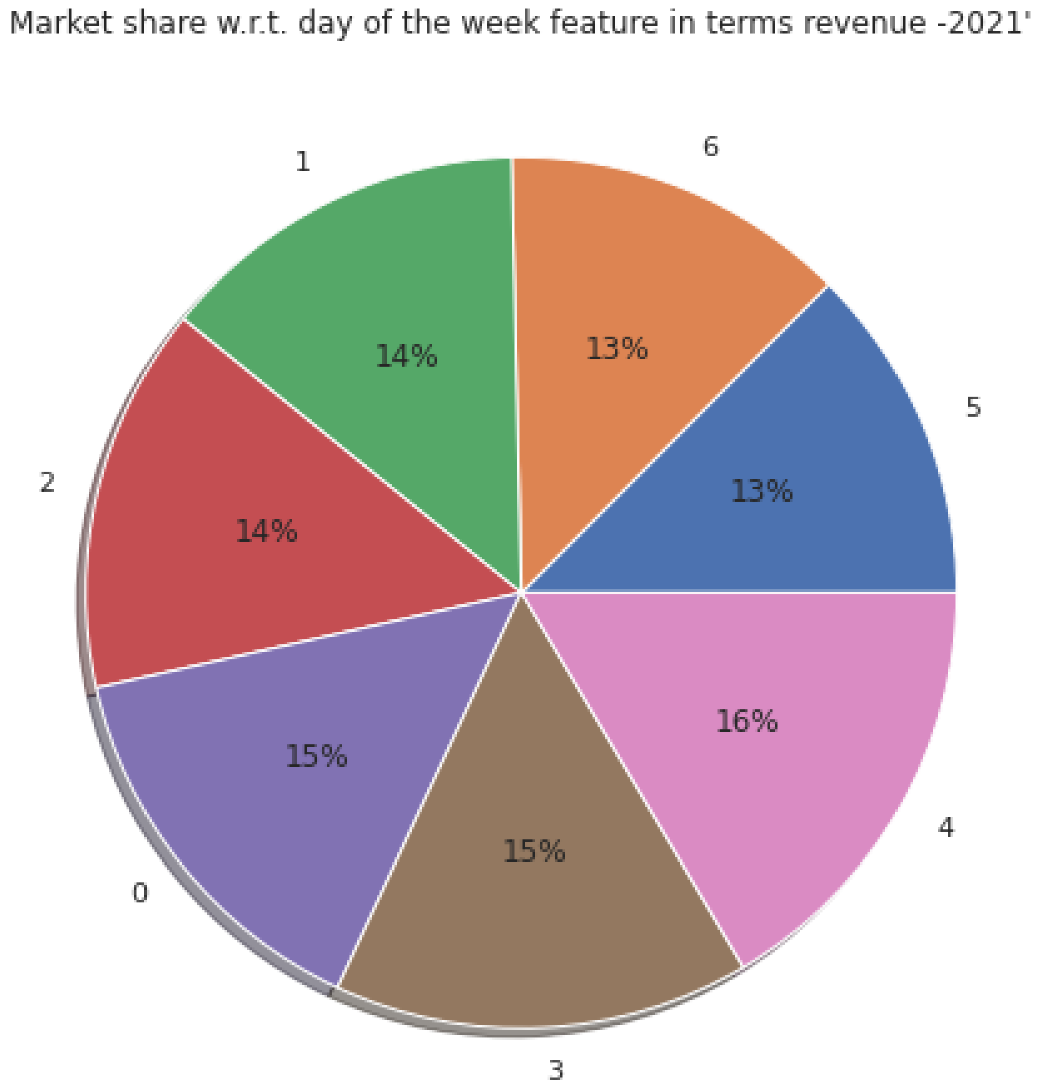
```
Market share w.r.t. Payment Type feature in terms Numbers-2021
Payment Type
Cash           1335583
Credit Card    1039903
Prcard          407671
dtype: int64
```

1.4.6 Trip time Distribution and analysis

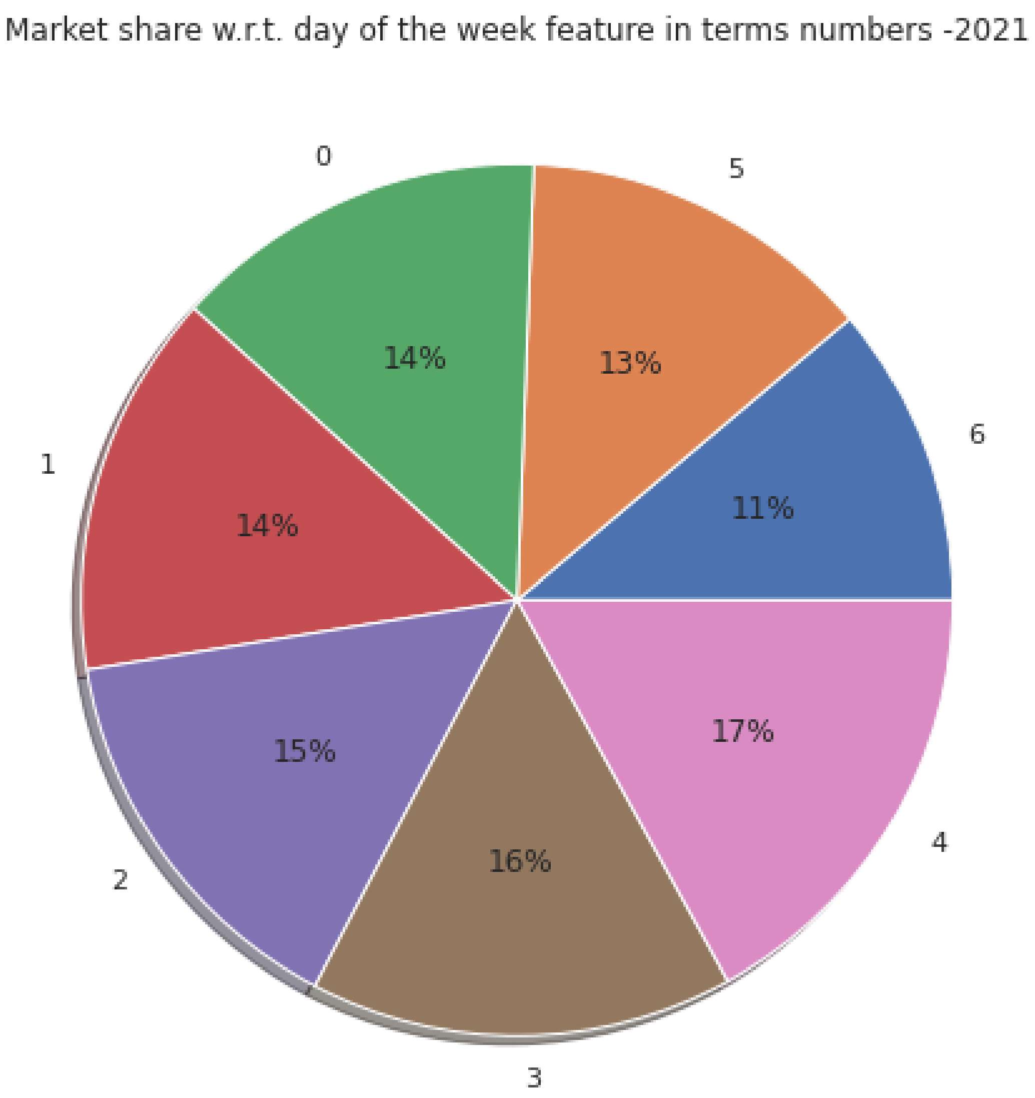
This section is going to analyze into 'Trip time', its distribution, clustering and respective market size and revenue. whether it was a weekday or, midnight or holiday.

Market share w.r.t. day of the week feature in terms revenue and number of trips. With respect to day of the week of the trip, the data is feature engineered as day of the week with Monday=0, Sunday=6.

As you can see in terms of revenue, the day of the week is in few percent difference. Same is the case with for number of trips.

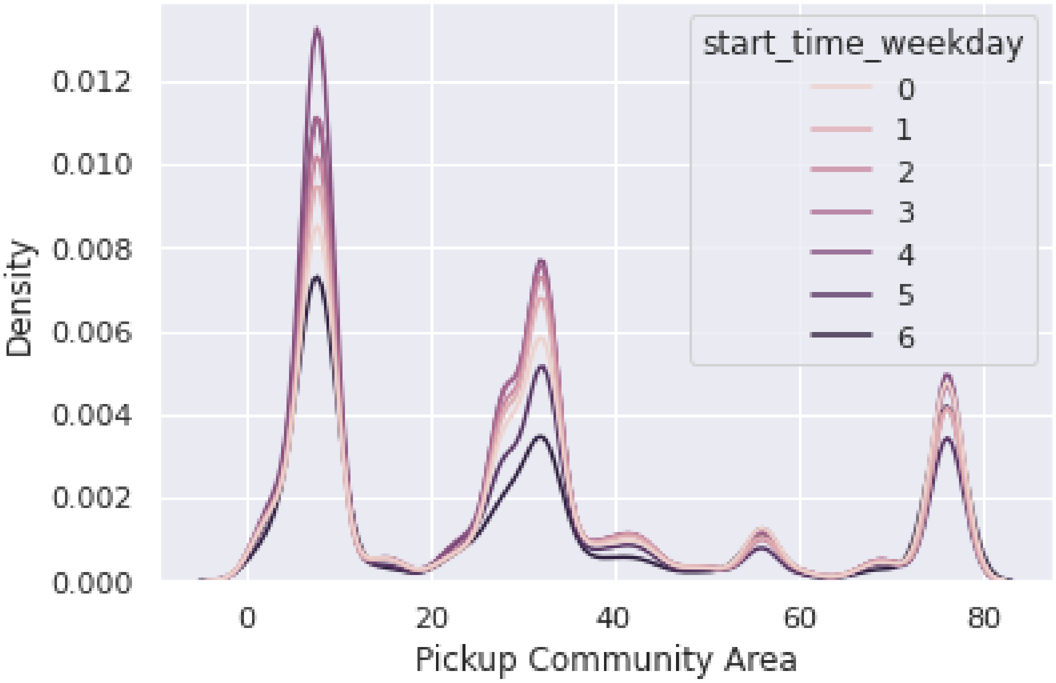


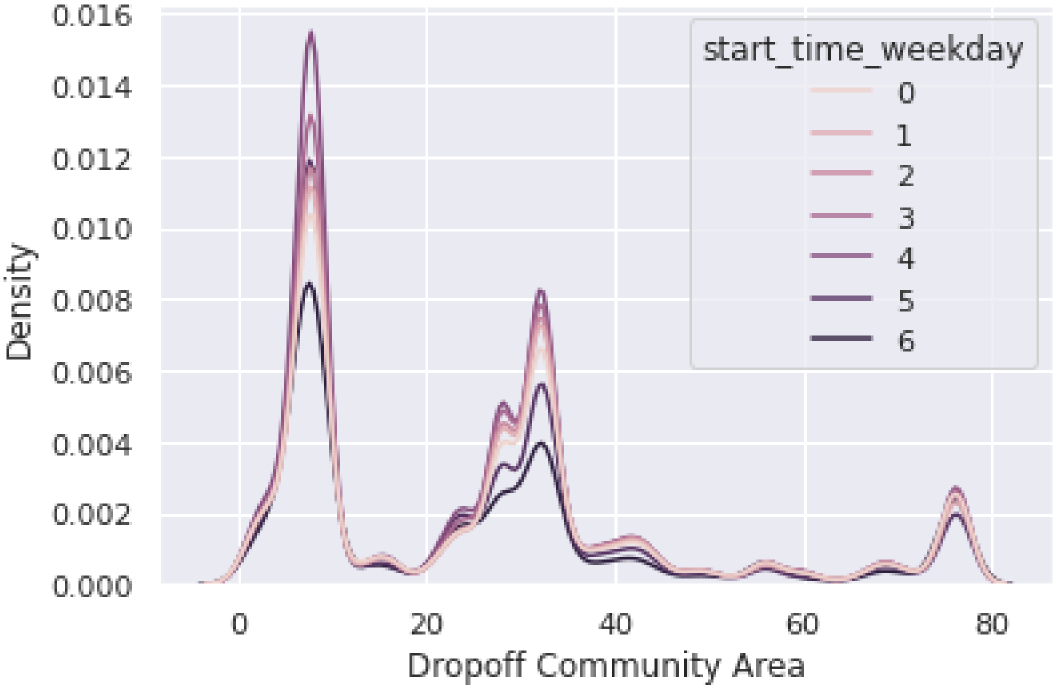
```
Market share w.r.t. day of the week feature in terms revenue -2021
start_time_weekday
4    10158567.05
3     9528429.80
0     8971967.31
2     8845283.41
1     8518208.79
6     7906542.99
5     7718995.16
Name: Fare, dtype: float64
```



```
Market share w.r.t. day of the week feature in terms numbers -2021
start_time_weekday
4    569662
3    528030
2    496048
1    478205
0    467831
5    444516
6    375065
dtype: int64
```

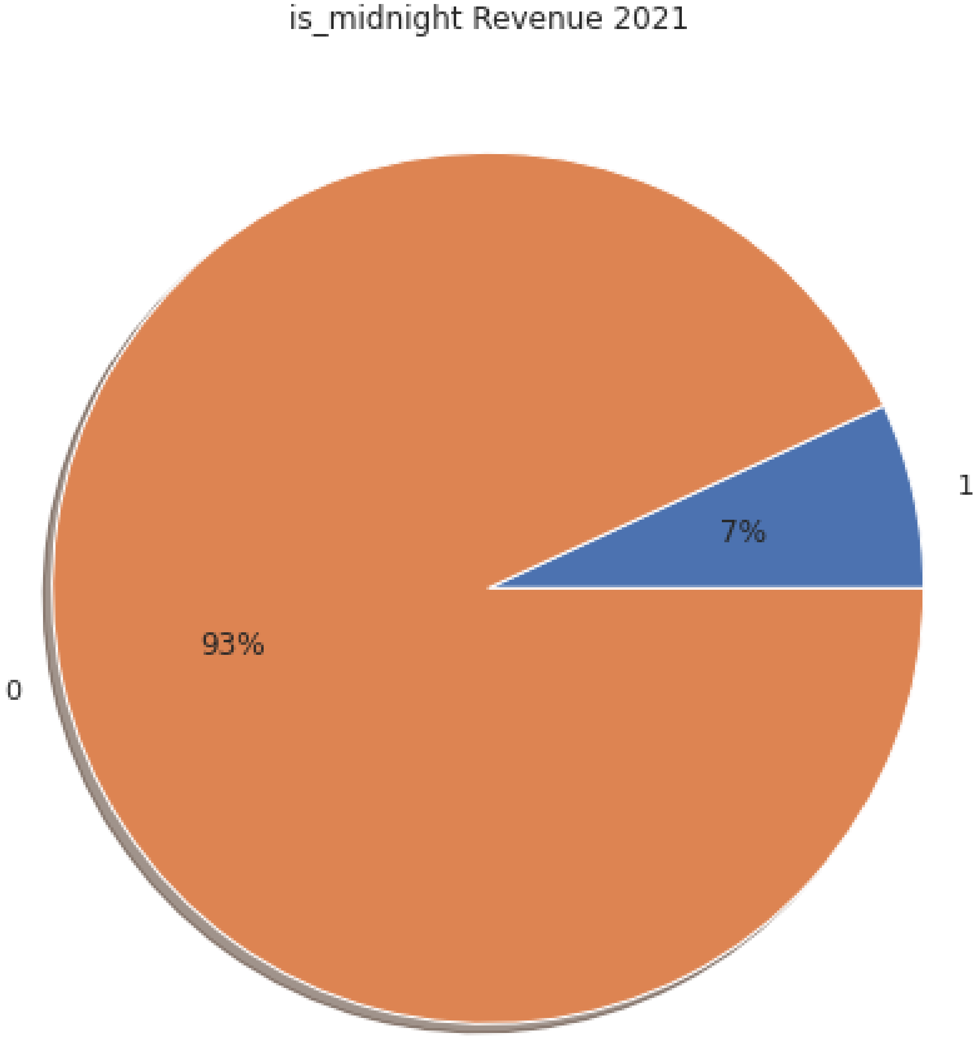
Market share w.r.t. day of the week, with location data - Kernal Density Estimator
With respect to day of the week of the trip, and the location area is plotted below as KDE.

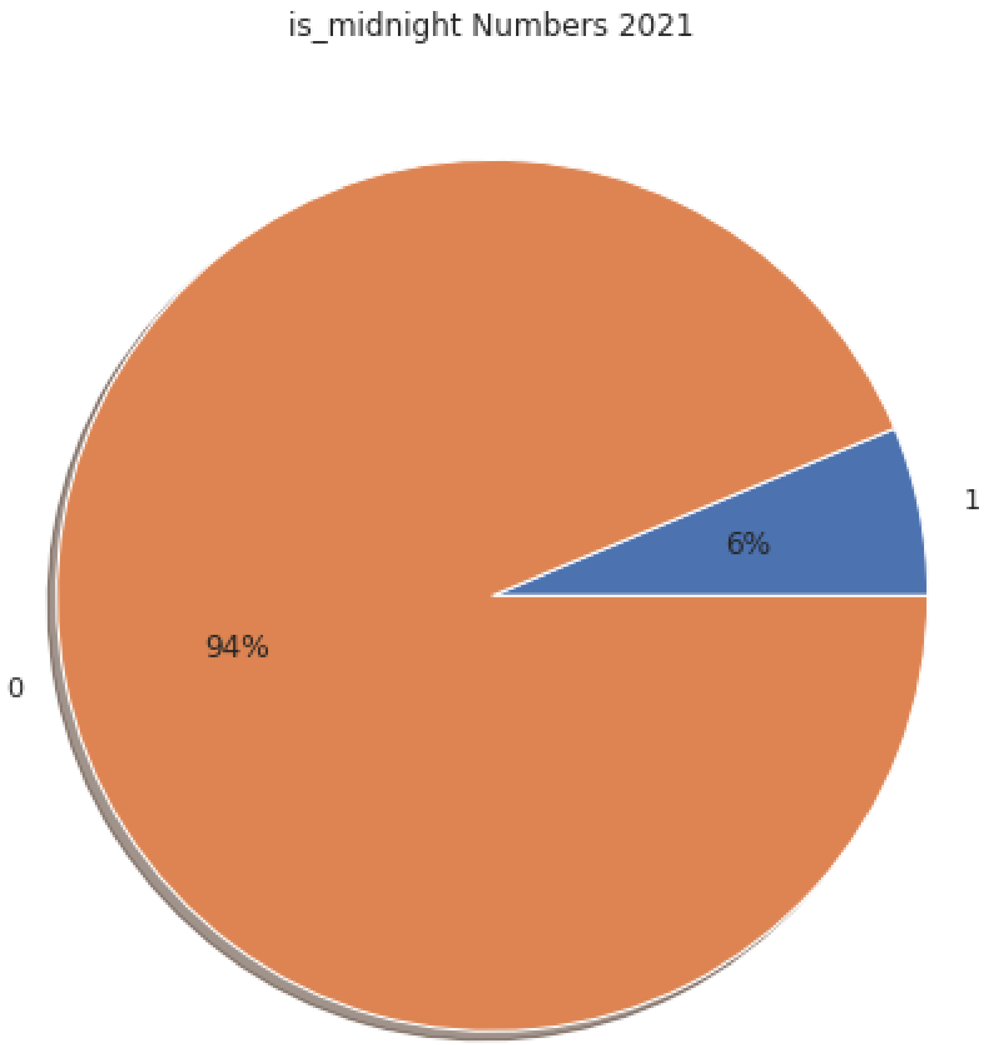




Market share w.r.t. midnight or not in terms revenue and number of trips. With respect to midnight or not the data is feature engineered as midnight if start time of the trip is in range 00:00 to 05:59. 1 is for midnight and 0 is for non-midnight

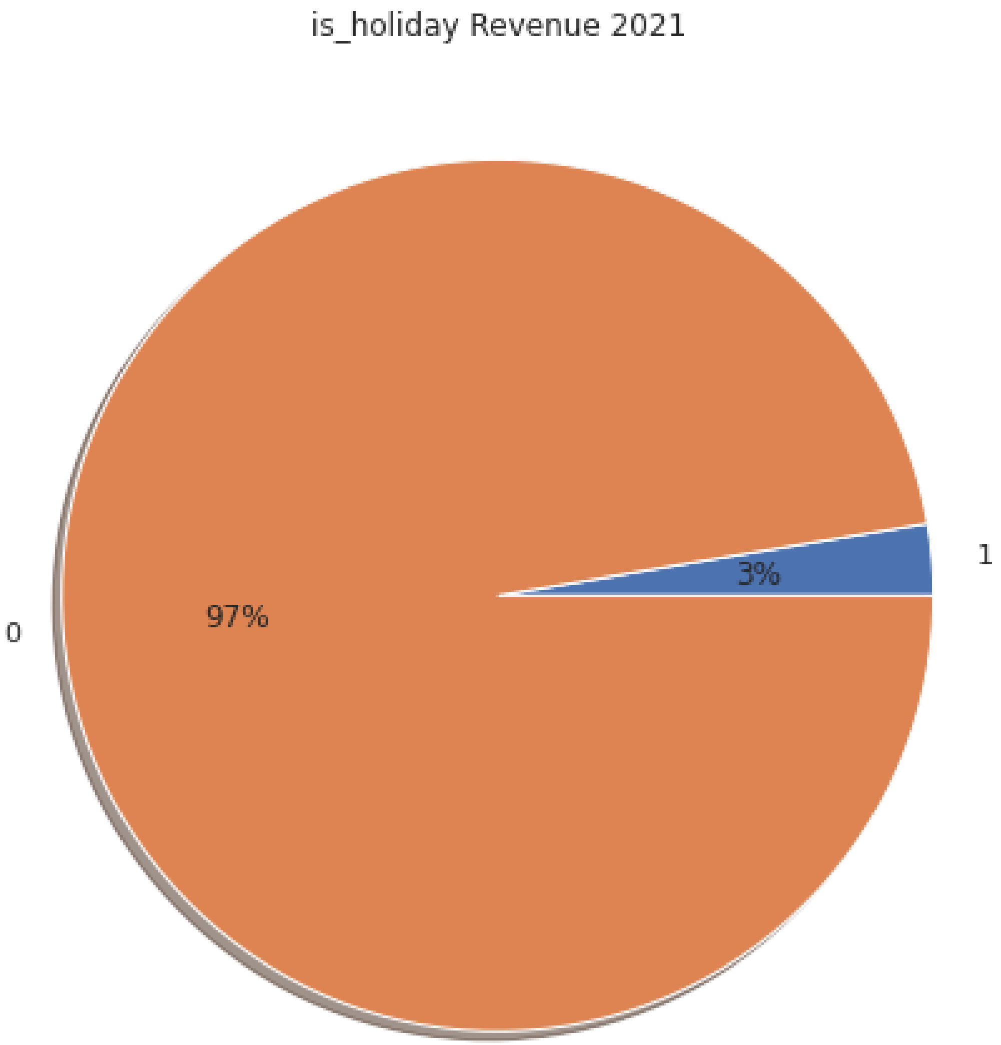
As you can see, midnight revenue and number of trips generates and constitutes of about 7% and 6% on the market.

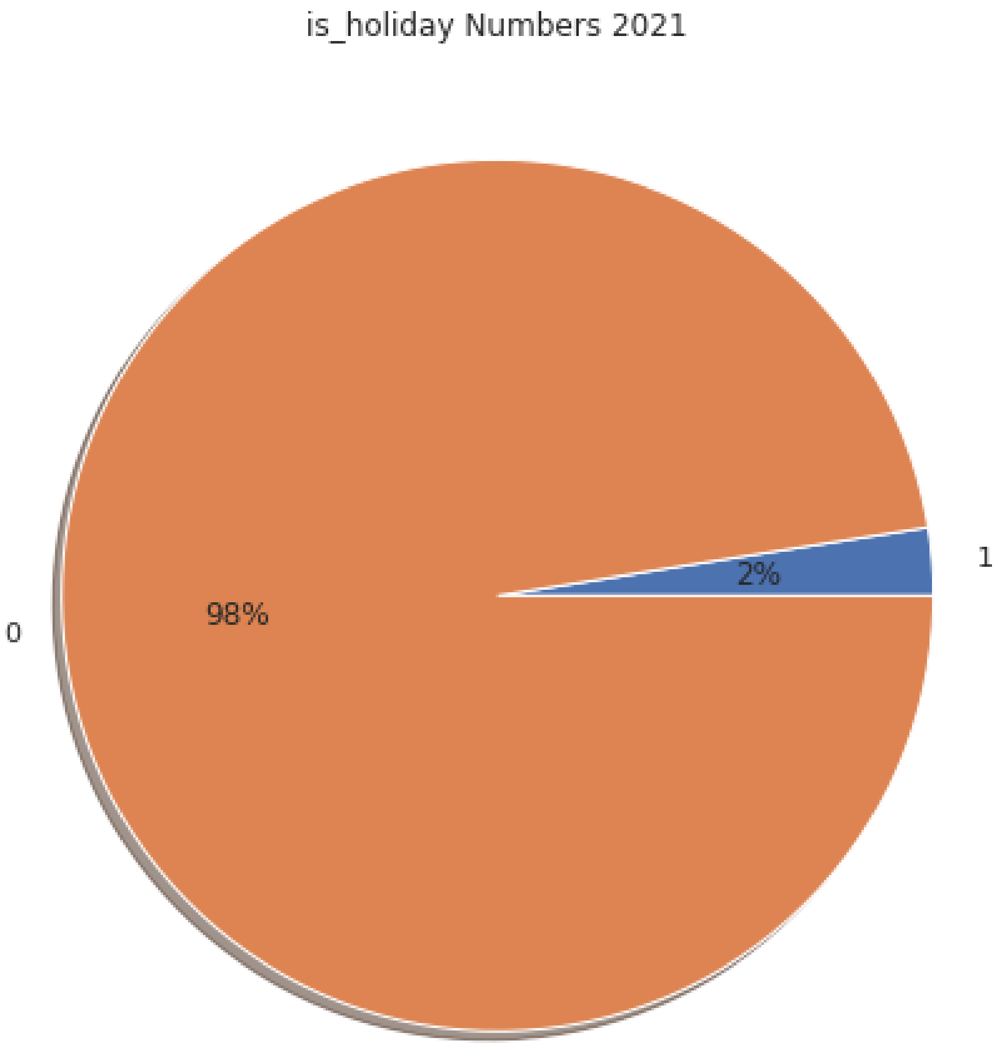




Market share w.r.t. holiday or not in terms revenue and number of trips. With respect to holiday or not, the data for feature engineering is obtained from https://www.chicago.gov/city/en/narr/misc/city-holidays.html. 1 is for holiday and 0 is for non-holiday

As you can see, holiday revenue and number of trips generates and constitutes of about 3% and 2% on the market.


is_holiday Revenue 2021
1
3%
0
97%



1.5 Pricing engine w.r.t. market

For pricing, I have approached the problem with the market distribution. For pricing I am building a regression model (using XGBoost) with input features as trip time, pickup location and dropoff location. If developed as an API, rest of the feature engine functions can be written separately. But for this demo, I am sampling data from the already existed data not used for training as it is already feature engineered.

1.5.1 Coordinate to Area

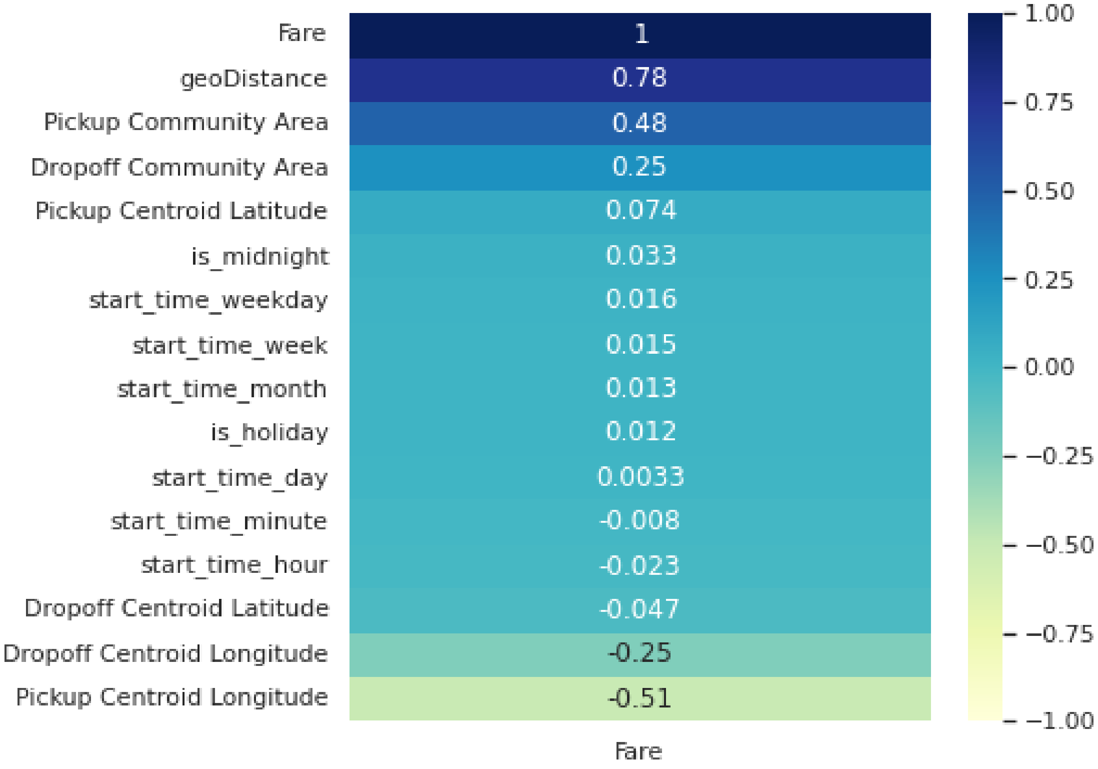
This helper function is written keeping is mind that in practice that the user will just input the location, there will be API running in the background to get the coordinates and based on the coordinates, we will obtain community area for pickup and drop off location. For this I am using ‘KNeighborsClassifier’.

So may ask why?. The answer is because the area number is used as one of the input feature for regression model.

```
As as an example for Corrdinates [-87.633308, 41.899602] the area obtained is 8
```

1.5.2 Analyzing the Input Feature - Correlation

As you can see the input feature used for regression and it correlation. Distance from pickup and dropoff has the higher correlation, followed by pickup area and etc. Since this data is available when booking a taxi, same is used as a input features for regression model.



1.5.3 Price Engine Training

For price engine training, XGBoost is selected for training regression model.

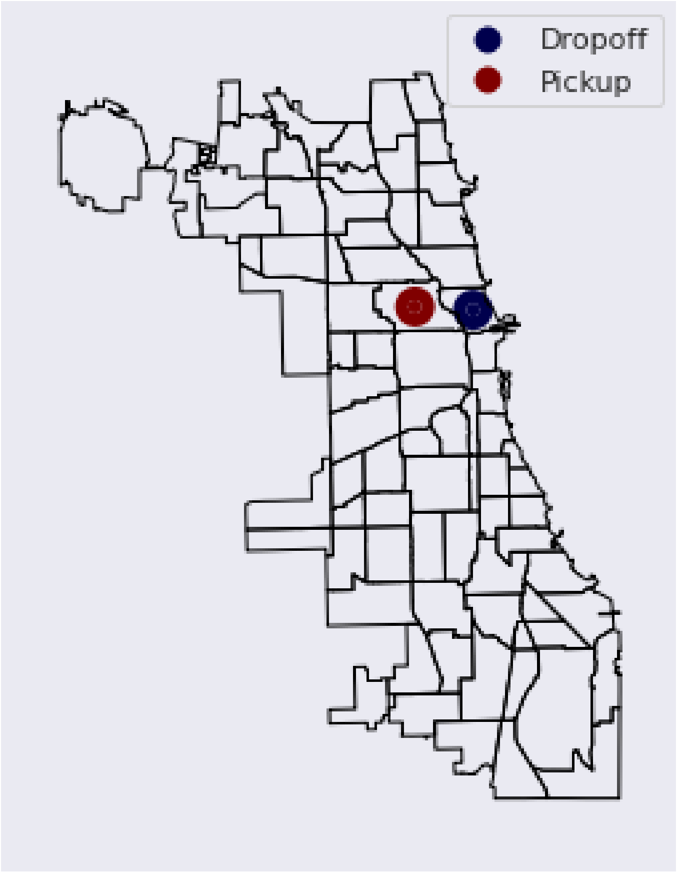
```
[0]     validation_0-rmse:20.84183
[100]   validation_0-rmse:6.18009
[200]   validation_0-rmse:6.15588
[300]   validation_0-rmse:6.14658
[400]   validation_0-rmse:6.14378
[500]   validation_0-rmse:6.14320
[562]   validation_0-rmse:6.14385
```

```
rmse = 6.196053332822338
mae = 3.138344272920899
```

1.5.4 Price is right - Running Price Engine on Sample data from Test Set

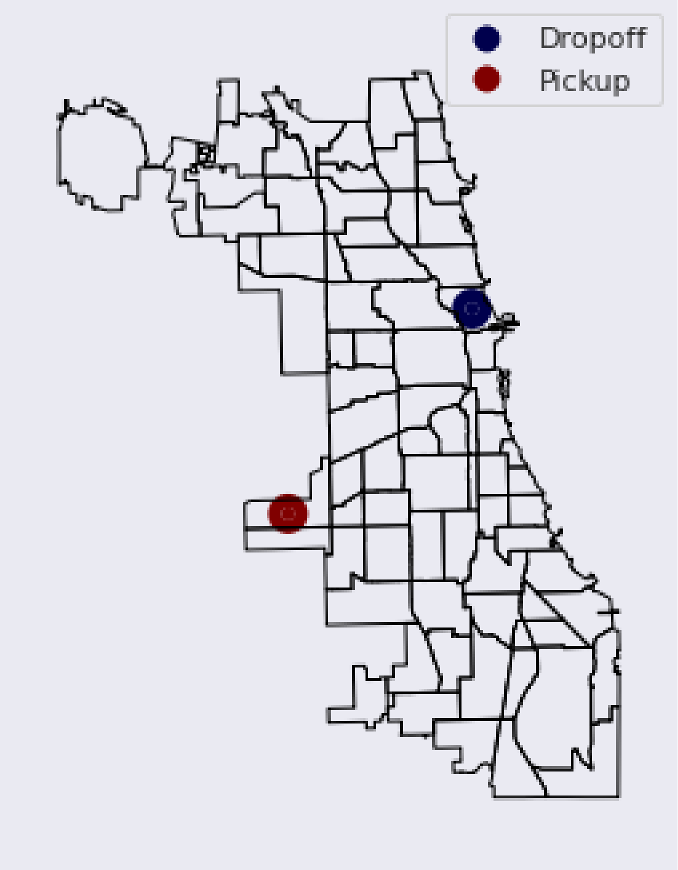
As mentioned above, after user select location and time, there will some back-end process to run and obtain all feature engineered input. For this demo I am sampling data from test set which is already cleaned and processed. Below I am running few samples with different multiplier. multiplier in this case is for either upsurge demand or downtrend demand.

```
Sample Trip #1
```



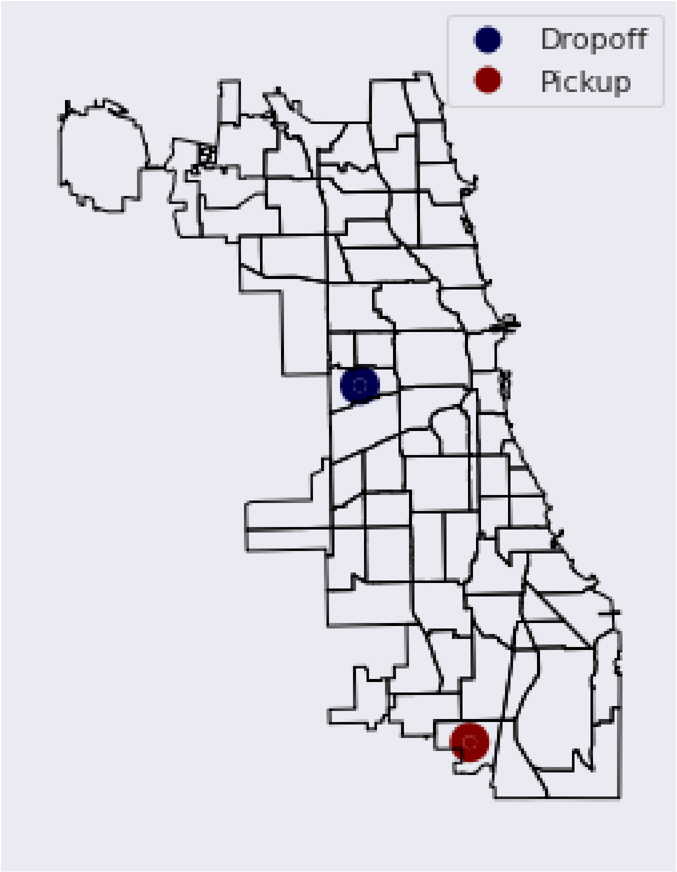
```
Actual Price = 10.12
Trip Fare as per price Engine  = 11.6
```

```
Sample Trip #2
```



```
Actual Price = 34.25
The Multiplier for the current time is :0.8
Trip Fare as per price Engine  = 26.160000000000004
```

```
Sample Trip #3
```



```
Actual Price = 47.75
The Multiplier for the current time is :1.2
Trip Fare as per price Engine  = 56.879999999999995
```

1.6 Market analysis as a Use cases for a taxi launch - Experimental

Note: This is still experimental.

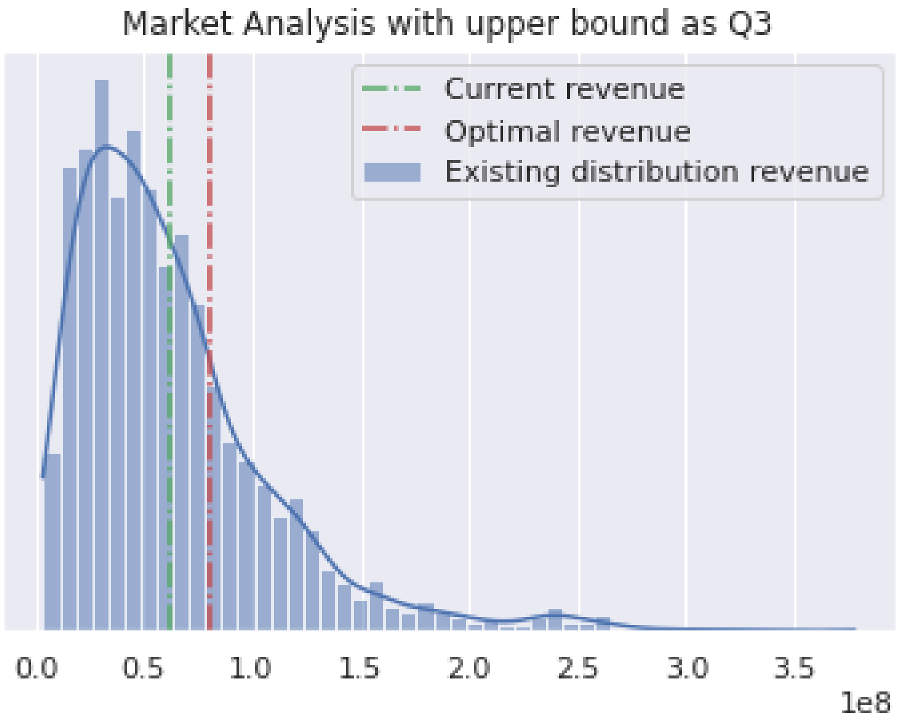
For Taxi Launch, I am trying to see the optimized revenue of the market in terms number of trips per price segment i.e. low-price, medium-price and high-price per taxi.

As a linear programming formulation, its running as the 3 variable which a number of trip per segment per taxi with a lower bound values as Q1-quantile number for price-segment. The objective is to maximize the revenue i.e. number of trips per price segment multiplied by kmeans cluster centroid values (8, 27, 45). Upper bound of the variables can be selected as ‘Q3 Quantile, Mean, and Median’ values. Please see the above box plot on number of trips per price segment for reference. In addition to the upper bound, taxi-market share percentage can also be analysed.

The same methodology can be extended to trip per community area as well with minor changes in problem formulation.

1.6.1 With upper bound as Q3 Quantile

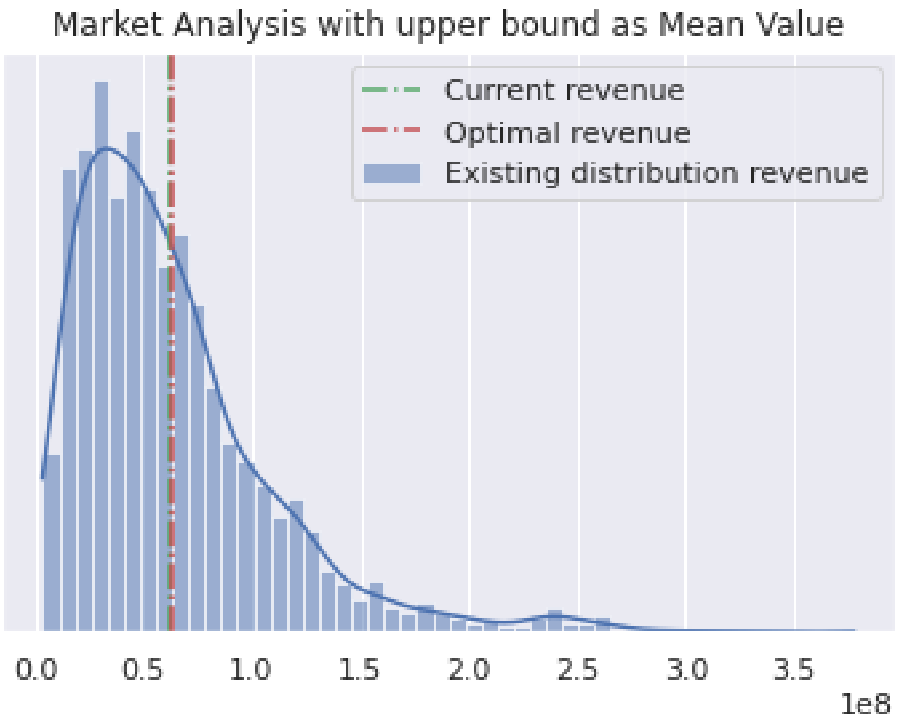
Let's run the market analysis with upper bound as Q3 quantile. The plot comprise of three things, revenue distribution with the existing distribution of the trips per segment, the current revenue and the max optimized value with upper bound as Q3 Quantile.



```
Current Revenue $61,647,994.51
Optimal Revenue with upper bound as Q3 Revenue $80,287,038.00
```

1.6.2 With upper bound as Mean Value

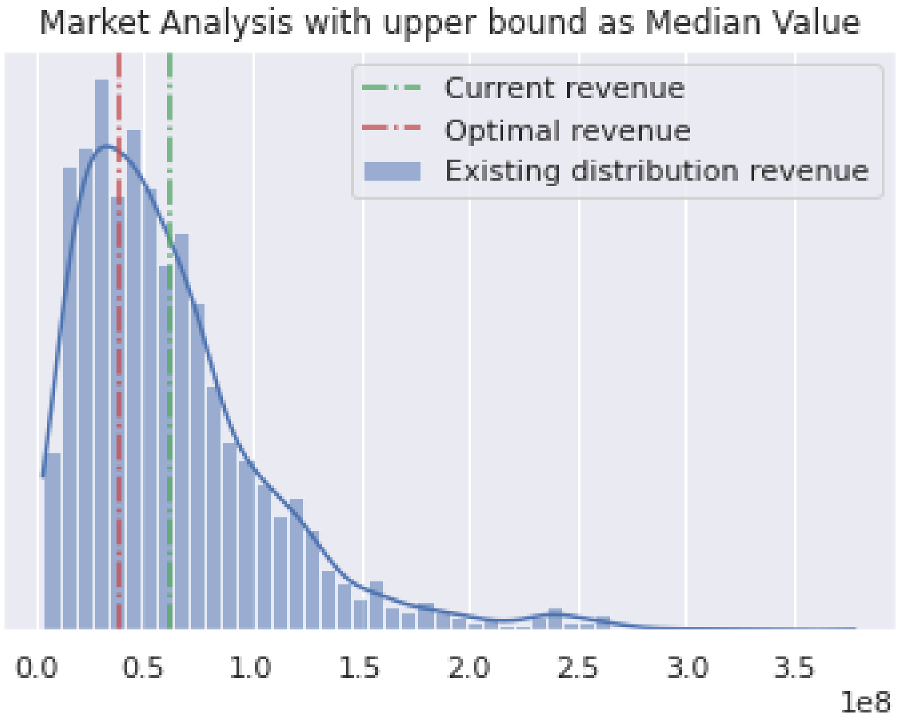
Market analysis with upper bound as Mean Value. The plot comprise of three things, revenue distribution with the existing distribution of the trips per segment, the current revenue and the max optimized value with upper bound as Mean Value.



```
Current Revenue $61,647,994.51
Optimal Revenue with upper bound as Mean Revenue $62,825,022.00
```

1.6.3 With upper bound as Median Value

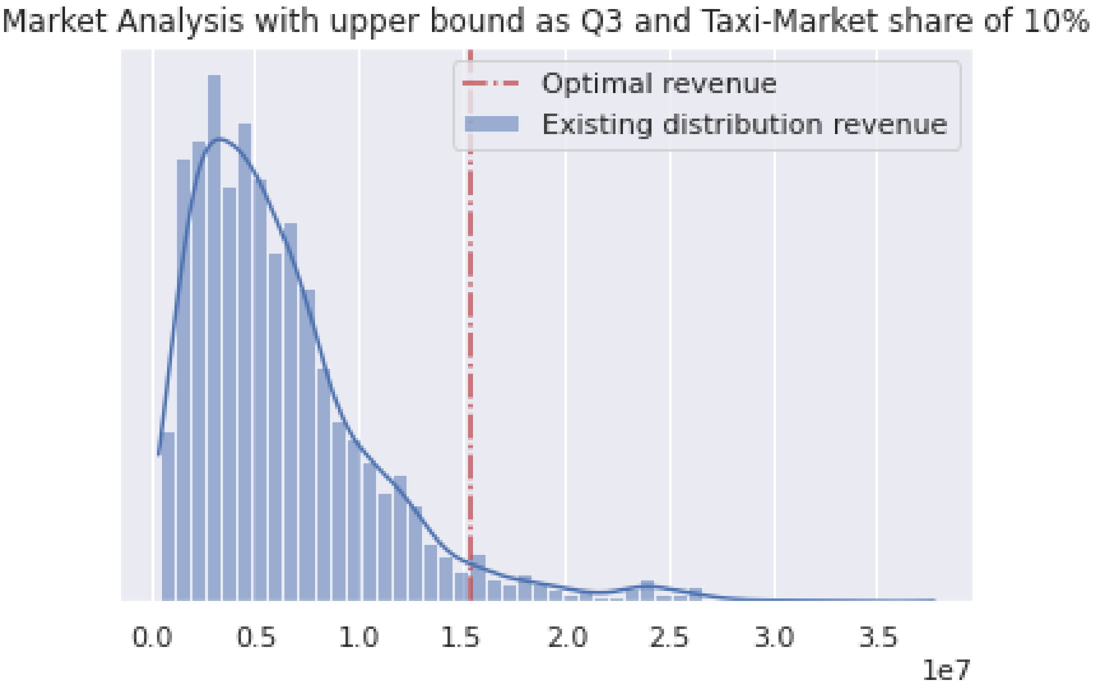
Market analysis with upper bound as Median Value. The plot comprise of three things, revenue distribution with the existing distribution of the trips per segment, the current revenue and the max optimized value with upper bound as Median Value.



```
Current Revenue $61,647,994.51
Optimal Revenue with upper bound as Median Revenue $38,931,516.00
```

1.6.4 With Taxi Market share of 10% and upper bound as Q3 Quantile

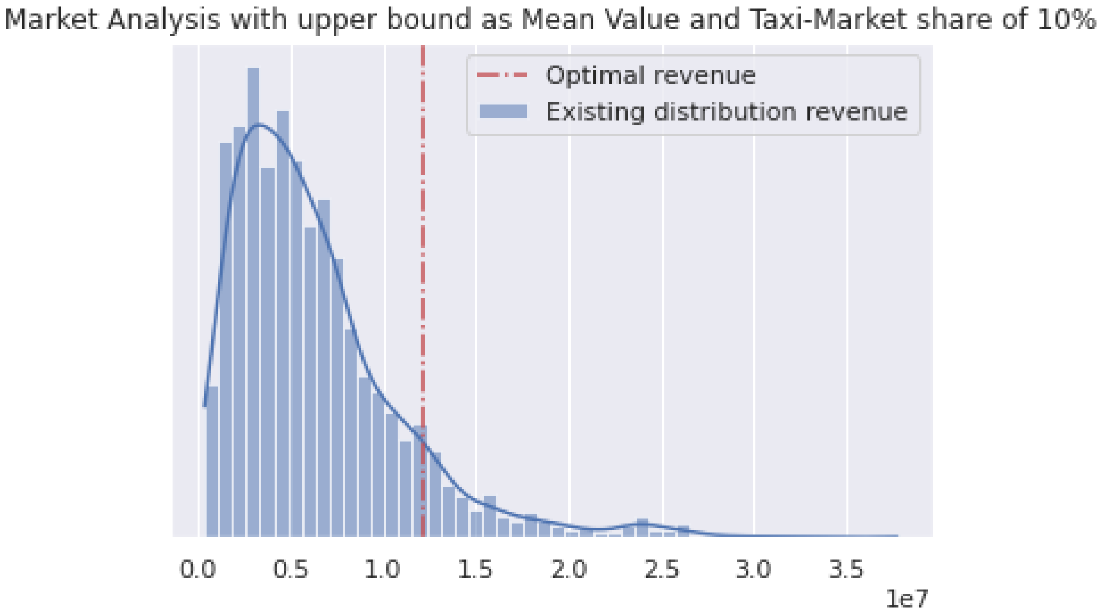
Market analysis with Taxi Market share of 10% and upper bound as Q3 Quantile. The plot comprise of three things, revenue distribution with the existing distribution of the trips per segment and the max optimized value with market share of 10% along with upper bound as Q3 Quantile Value.



```
Optimal Revenue with upper bound as Q3 Revenue and Taxi-Market share of 10% :
$15,439,815.00
```

1.6.5 With Taxi Market share of 10% and upper bound as Mean

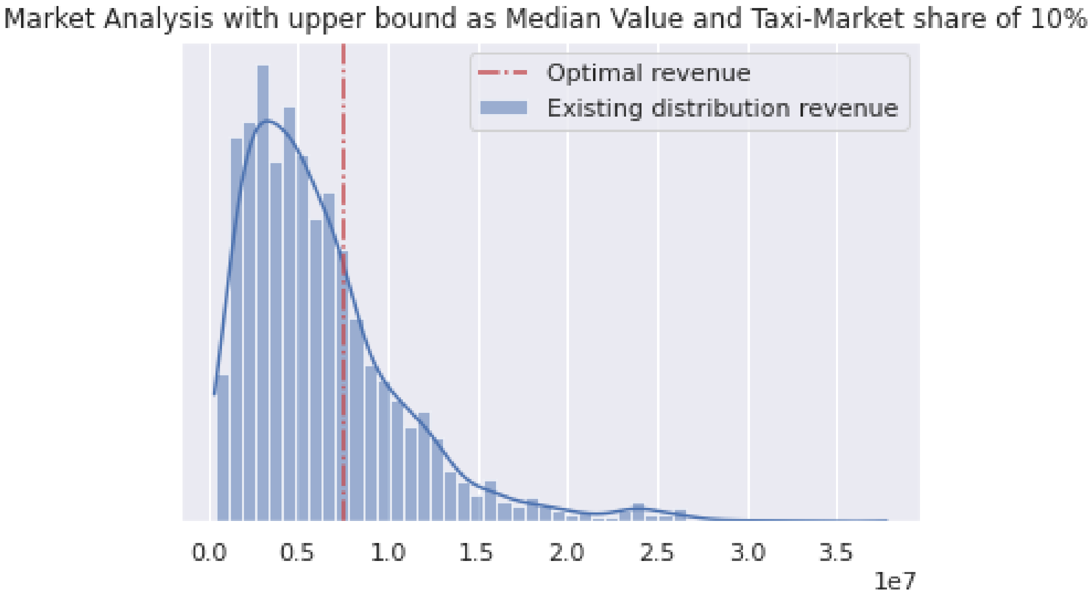
Market analysis with Taxi Market share of 10% and upper bound as Mean value. The plot comprise of three things, revenue distribution with the existing distribution of the trips per segment and the max optimized value with market share of 10% along with upper bound as Mean Value.



```
Optimal Revenue with upper bound as Mean Revenue and Taxi-Market share of 10% :
$12,081,735.00
```

1.6.6 With Taxi Market share of 10% and upper bound as Median

Market analysis with Taxi Market share of 10% and upper bound as Median Quantile. The plot comprise of three things, revenue distribution with the existing distribution of the trips per segment and the max optimized value with market share of 10% along with upper bound as Median Value.



```
Optimal Revenue with upper bound as Median Revenue and Taxi-Market share of 10%
: $7,486,830.00
```