THE UNIVERSITY OF TEXAS ARLINGTON

INSY 5339 002 Principles of Business Data Mining

Final Project Report

Subscription Strategy Revamp for Citi Bike Application

Professor: Dr. Jayarajan Samuel

GROUP NUMBER- 08

Team Members:

Mohit Kumar Dundu	***1002011177***
Ayvee Nusreen Anika	***1002027982***
Suchit Surendra Kakirde	***1001837315***
Pratik Prashant Phirke	***1002021297***

1. Executive Summary:

Bike-sharing systems are a method of renting bicycles in which the process of getting a membership, renting, and returning the bicycle is automated through a network of kiosk stations located around a city. These systems let customers rent a bike from one location and return it to another as needed. There are now over 500 bike-sharing schemes in operation around the world. Citi Bike Rental Application is a service that allows users to rent bikes for a set length of time. Everything is computer-controlled, from bike booking to payment options. After providing all the necessary information to borrow a bike from a 'dock,' the system unlocks it, and the bike can be returned to another dock belonging to the same system.

Motivation/Background:

Citi Bike, a bike-rental system in the United States, recently reduced its subscription during the Covid19 pandemic. They are struggling to maintain the current market condition. As a result, Citi Bike intends to offer cheap annual memberships to their casual clients who have the potential to become registered members. Previously, it was based on managers' instincts and largely on the company's yearly or semiannual reports. As part of Citi Bike's Data Analytics Team, we are providing analysis for company management to determine where and when they may send their promotional team to promote cheap memberships to its casual users.

Loyalty Program:

As part of the loyalty program development, we will answer three W's inquiries for management WHEN, WHERE, and WHO. Our predictive algorithm is based on one year of customer rental data and is ready to identify potential consumers to whom the company may provide discounted subscriptions at specific times and locations as part of the establishment of a loyalty program.

When? We have focused on the optimal times to conduct the campaign for converting the non-subscribers to subscribers after running our model on the dataset that the customer provided. This choice will determine on what day and what time the promotions can be run to attract new members.

Where? We have riders’ pick and drop-off locations which helped us to prioritize locations based on certain timings.

Who? From our dataset, we know how many customers and subscribers are using our bikes, who are our frequent users, what time they usually rent and for how long they ride, and which age group or gender our riders belong- all this information helped us to find our potential target groups.

In our final model, we have applied prediction techniques to achieve our goals such as Ratio and Variance calculations to answer our 1st question “When”, Clustering for answering “Where” and Logistic Regression to answer our 3rd question i.e. "Who”, using other two answers in the final model

Some other techniques which we have used to support our final model are Correlation Matrix and Linear Regression (OLS). We used Correlation Matrix to determine our most relevant variables to the final model, whereas OLS to compare the result of our finalized model's (Logistic regression) accuracy.

2. Data Description:

The data we are providing from Citi Bike NYC is an immensely humongous dataset. We have picked 2019 data which will help us to predict the company's previous success factors while making new subscribers post the Covid scenario. We received this data from the official site of CitiBike (Ref: https://ride.citibikenyc.com/system-data) We will be mainly focusing on the demographic aspects of the dataset. Below are the different variables from our dataset.

1. **tripduration** - Duration in Seconds - Continuous
2. **starttime** - Start Time and Date - Continuous
3. **stoptime** - Stop Time and Date - Continuous
4. **start station id** - ID of Start Station - Continuous
5. **start station name** - Name of Start Station - STRING
6. **start station latitude** - Latitude of start station - Continuous
7. **start station longitude** - Longitude of start station - Continuous
8. **end station id** - ID of End Station - Continuous
9. **end station name** - Name of End Station - STRING
10. **end station latitude** - Latitude of End station - Continuous
11. **end station longitude** - Longitude of End station - Continuous
12. **Bike ID** - Bike ID - Continuous
13. **usertype** - Customer = 24-hour pass or 3-day pass user; Subscriber = -Annual Member - Binary
14. **gender** – 0= unknown,1=male; 2=female - Categorical
15. **birthyear** - Year of Birth – Continuous

3. Exploratory Data Analysis and Visualization:

Exploratory data analysis is a statistical procedure that evaluates data sets to identify and summarize their key characteristics, typically with the aid of visual tools. Even if a statistical model may be used, EDA is typically used to look at what the data may tell us besides what formal modeling or hypothesis testing can.

Tools Used:

We have Used Python and its libraries such as matplotlib to perform our visualization and answer our questions.

Visualization techniques:

We used Bar charts, Scatter Plots and Histograms for data visualization to support and analyze our problem.

4. Methodology:

Data Visualization and Exploratory Data Analysis (After initial cleaning): The process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. We tried to analyse whether there are any null values or outliers in the dataset and how to wrangle/handle them. Hence, we used Python to check data quality and pre-process dataset to deal with missing, null, or duplicate values. We also used Microsoft excel, and our focus is on **weekdays, weekend, distance, gender, age group, times, and locations**. This visualization helped us to understand the insights of our dataset and look further into few patterns this dataset already has.

Predictive Analysis: It is the collection and interpretation of data to uncover patterns and trends. We extracted date and time and categorized it into weekdays, weekends, and time intervals. We have also transformed user type 0-1 for our ease of work. Then we have used this information for ratio and variance calculation. Low ratio (Subscriber/ Customer) and low variance are required for our project. We did cluster analysis to identify the best locations as per our need, followed by logistic regression on transformed data to suggest to our management where and what time they can give discounted subscriptions.

```
# Extracting Hour

df['Start Hour'] = pd.to_datetime(df['Start Time']).dt.hour


df['End Hour'] = pd.to_datetime(df['End Time']).dt.hour
```

Fig 1: Python code for extraction of hour from 'Start Time' and 'End Time' avariables.

```
import pandas as pd
cust_out_file_name = r'D:\MS_Courses\3.Fall\Mining\Project\FINAL DATASET\All_Processed_Hour-NEW.csv'
df = pd.read_csv(cust_out_file_name)
df['usertype'].replace(['Customer', 'Subscriber'],
                        [0, 1], inplace=True)
df.to_csv(cust_out_file_name)
```

Fig 2: Python code to transform 0-1 user type to 'Customer' and 'Subscriber'.

To summarize, irrespective of location we will be predicting the best time first. After that, we will predict the best location for that best time to provide a discounted subscription to those potential casual customers.

5. **<u>Results of Data Visualization:</u>**

- Histogram of Duration.

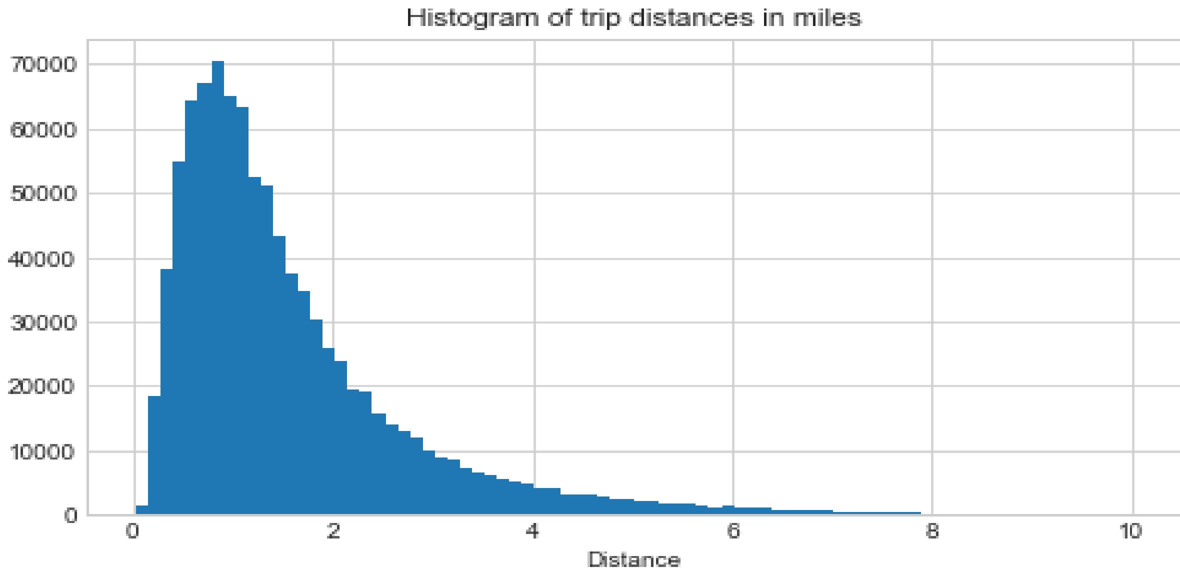


Fig 3: Histogram of Distance Vs Trips.

We can see from above histogram that the bike trips are primarily **short-distance trips** (less than 4 miles). There are few trips longer than 8 miles. We can also get more insights by grouping the trip distances by gender, age, and user type, whose histograms are shown below respectively.

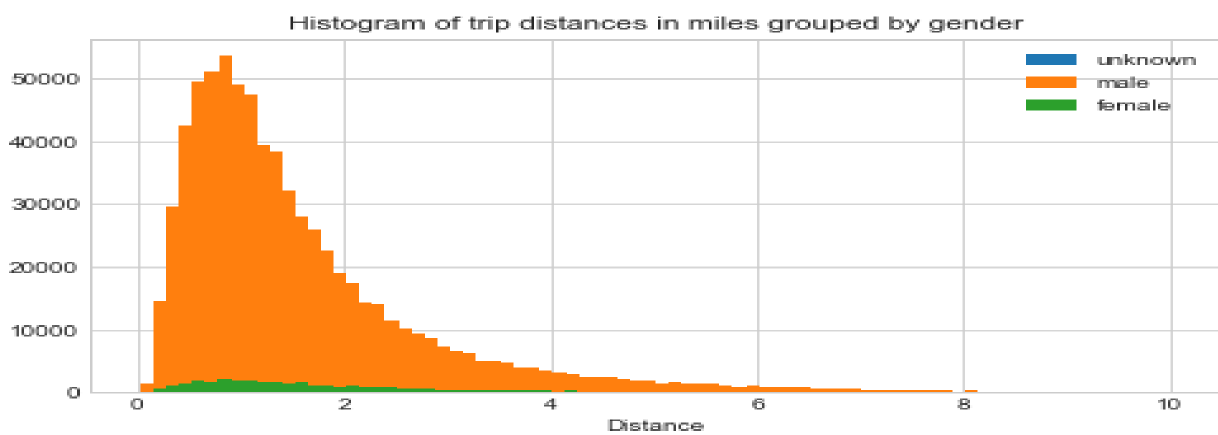


Fig 4: Histogram of Distance Vs Trip (Group Trip by Gender).

We can see from above histogram that Citi bike users are primarily **male**.

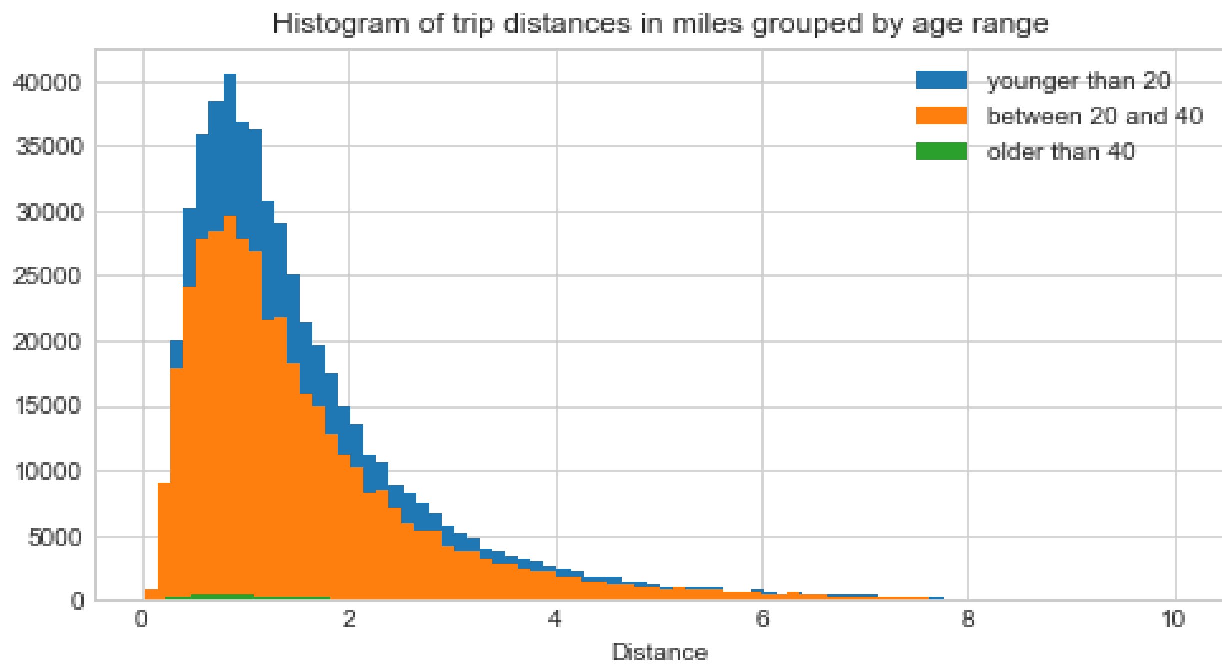


Fig 5: Histogram of Distance Vs Trips (Group by Age).

The histogram above shows that users are primarily **young and middle-aged** people.

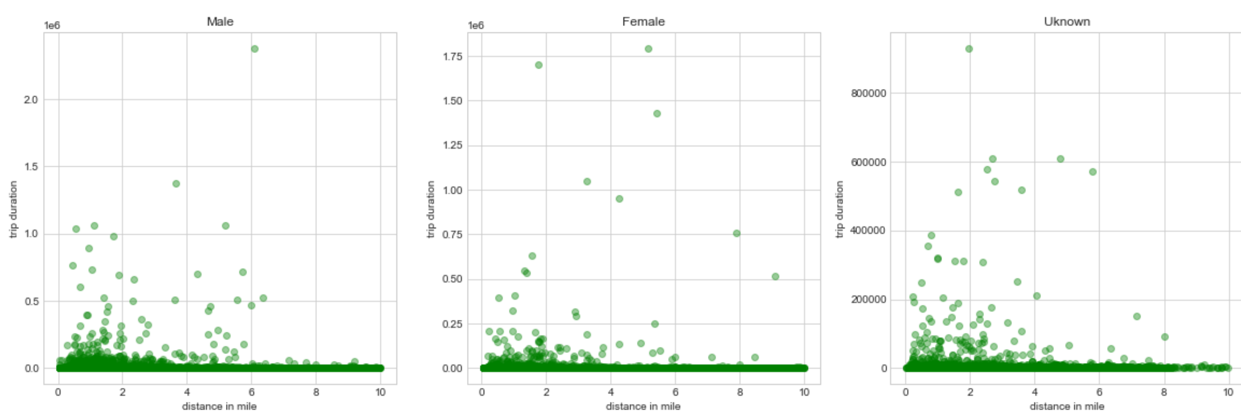


Fig 6: Scatter plot of Distance Vs Trip (Male and Female).

The scatter plot shows that male users are more likely to ride relatively longer distance than female users.

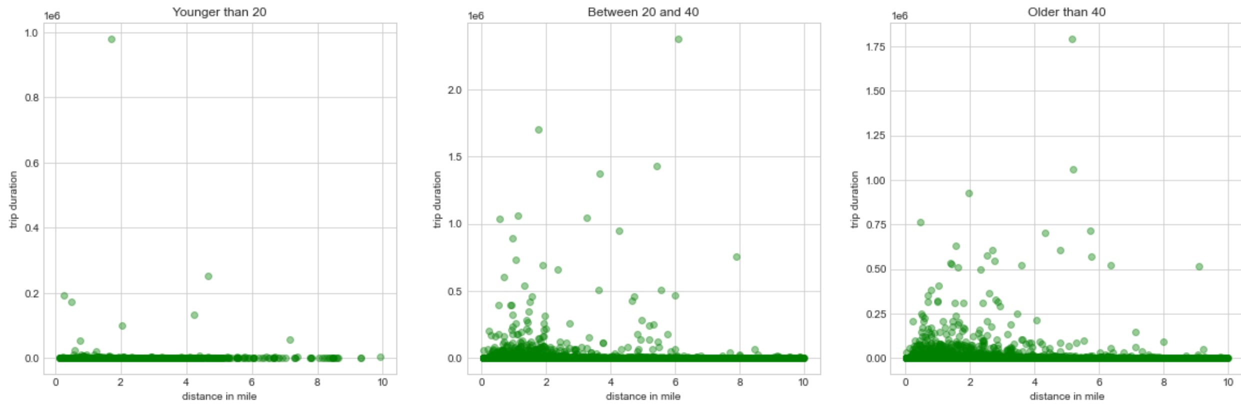


Fig 7: Scatter plot of Distance Vs Trip (Age).

The scatter plot shows that young users use Citi bikes primarily for short distance ride, while middle and old age people are more likely to ride a relatively longer distance.

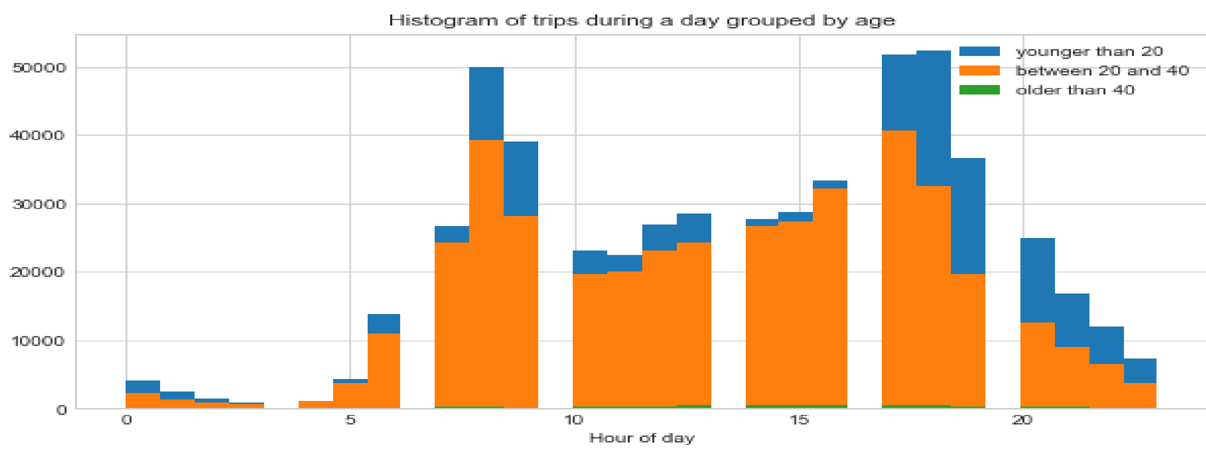


Fig 8: Histogram of Hour of day Vs Trip (with respect to age).

As the histogram shows above, there are more trips during rush (or office) hours (7am-9am and 16pm-19pm) in one day. We expect to see more people under the age of 20 using Citi bike during rush hours since the probability of having a car for them is low and the roads have more traffic.

Trips by hour on weekdays and weekends:

From the below two bar charts about pick-up times on weekday and weekend, we can see that weekday riders mainly use Citi Bikes to commute to and from work (we suppose as these are mostly office hours), with peak hours from 8–9 AM and 5–6 PM. On the other hand, weekend riders prefer a more leisurely schedule, with most rides occurring in mid-afternoon.

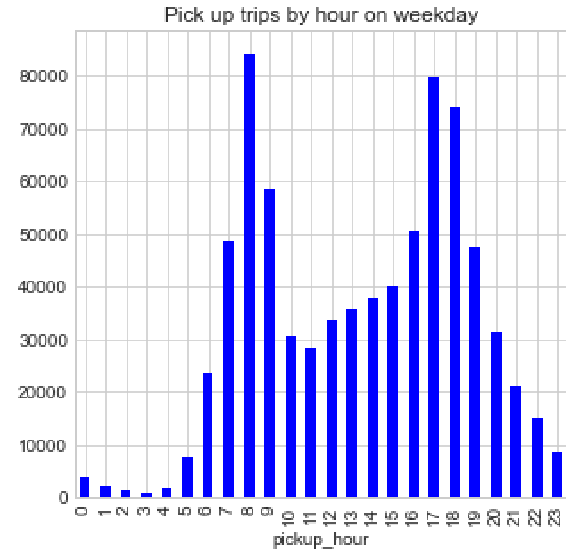


Fig 9: Pickup hour Vs Trip (Weekday)

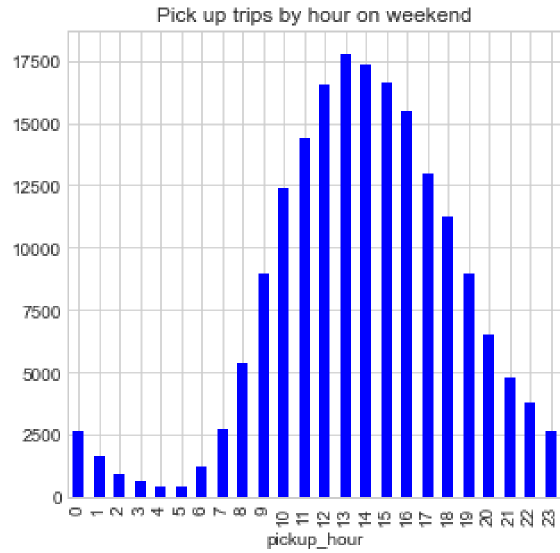
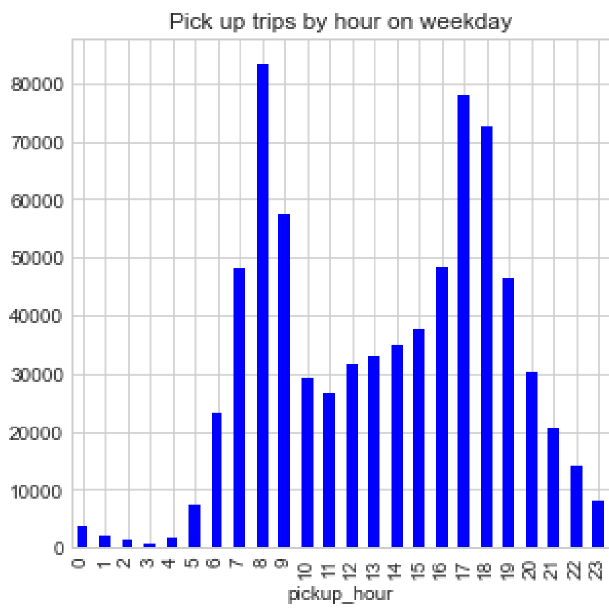
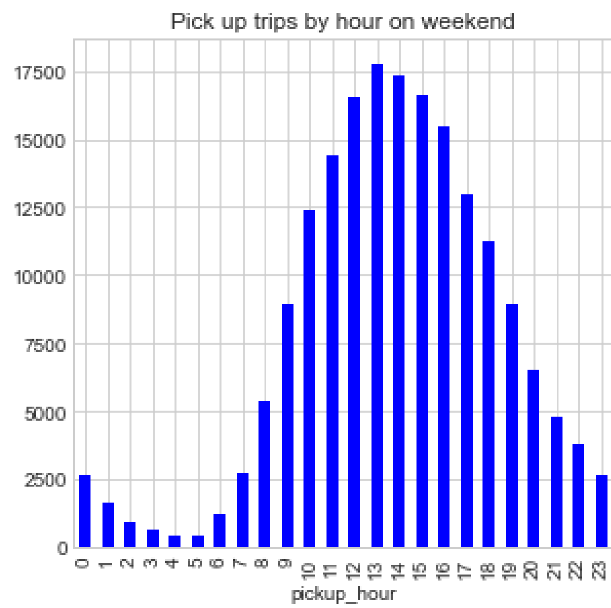


Fig 10: Pickup hour vs Trip (Weekend)

From the below graph, we can conclude annual subscribers mainly use Citi Bikes to commute to and from work, with peak hours from 8–9 AM and 5–6 PM. And on weekend, they mainly use it for entertainment, with peak in mid-afternoon.





6. <u>Predictive Analysis Details:</u>

Ratio calculation: We calculated ratio of total subscribers to non-subscribers throughout the year for all time intervals**.**

```
ratios={}
for i in range(0,24):
    ratios[i]=(len(df[(df['usertype']==1) & (df['Start Hour']==i)])/len(df[(df['usertype']==0) & (df['Start Hour']==i)]))
print(ratios)
#plt.bar(rartios.keys(), ratios.values(), width, color='g')
```

```
{0: 22.318181818181817, 1: 17.190697674418605, 2: 12.134831460674157, 3: 17.38372093023256, 4: 36.41379310344828, 5: 89.6179775
2808988, 6: 90.73431734317343, 7: 103.63967611336032, 8: 84.66698292220114, 9: 44.616778523489934, 10: 19.620982986767487, 11:
13.796271186440679, 12: 12.844330729868032, 13: 11.735602704593145, 14: 11.678305924757828, 15: 12.797902764537655, 16: 17.3157
03380588875, 17: 33.036271309394266, 18: 38.218735788995, 19: 33.080712166172106, 20: 25.03728813559322, 21: 23.64438254410399
4, 22: 22.14909090909091, 23: 19.463587921847246}
```

Fig 11: Python code for the ratio of subscriber to non-subscriber

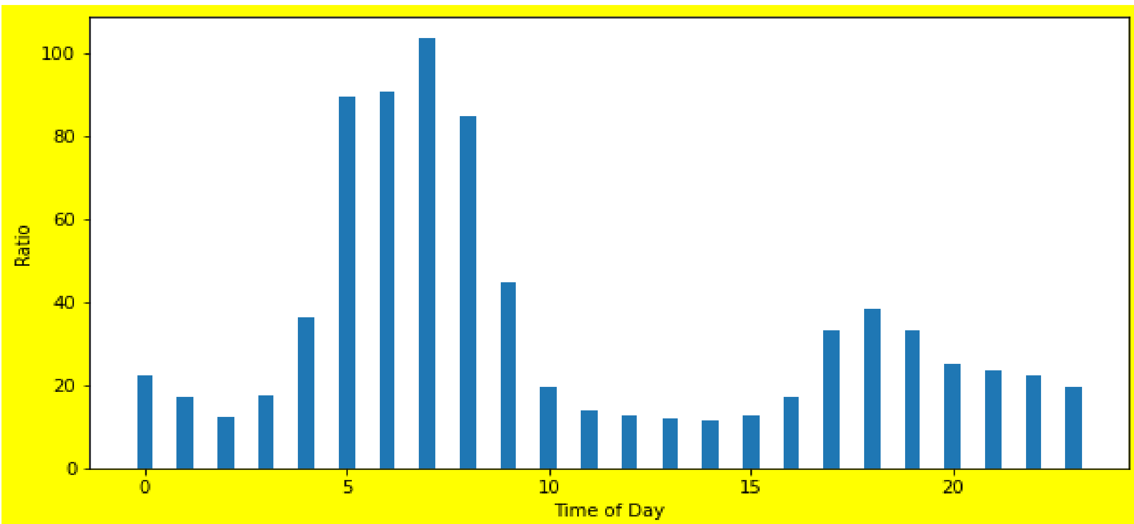


Fig 12: Time of Day vs Ratio.

We found ratio between subscribers to non-subscribers to understand the spread of users across the days and times. **Low ratio means more subscribers compared to nonsubscribers**.

We plotted above bar chart to analyze the ratio of total subscribers to non-subscribers throughout the year, at that time interval.

Example, 0 represents ratio between 12am to 12:59am.

Variance calculation: We need to find best time and location to give our subscription. That is why we need to find the variance of ratio. Variance represents how constant the ratio is throughout the span of time.

Firstly, after calculating the ratio of subscribers to non-subscribers for a particular hour on a given day. Example, we first find ratio of subs to non-subs between 12am to 12:59am on Jan 1st. This ratio is one data point. Likewise, we get 365 ratios for throughout the year for every hour.

So, variance of 0: represents variance of all the ratios for 365 days between 12am and 12:59am.

Variance for all the 365 ratios for each hour:

```
# CALCULATE VARIANCE FOR EACH OF THESE HOURS FOR ALL THE DAYS
cust_out_file_name = r'D:\MS_Courses\3.Fall\Mining\Project\FINAL DATASET\All_Processed_Hour-NEW.csv'
df = pd.read_csv(cust_out_file_name)
dates = df['Start Date'].unique()
var_ratio = {}
ratio_hour = {}
nr = []
dr = []
# i = 1
for i in range(24):
    lst = []
    for j in dates:
        try:
            rat = (len(df[(df['usertype']==1) & (df['Start Hour']==i) & (df['Start Date']==j)])/
                    len(df[(df['usertype']==0) & (df['Start Hour']==i) & (df['Start Date']==j)]))
        except ZeroDivisionError:
            rat = 0
        lst.append(rat)
    ratio_hour[i] = lst
```

Fig 13: Python code to calculate variance for each hour for all the days.

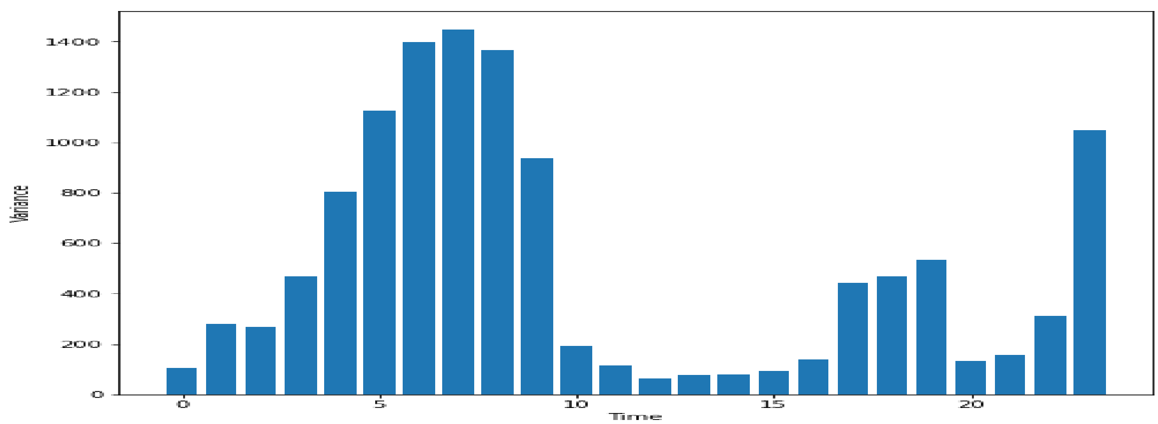


Fig 14: Time (each hour for all day) Vs variance.

Below mentioned score are the various calculated variances throughout the year using python programming.

```
Variance is:
{0: '103.291', 23: '1047.079', 11: '113.420', 20: '133.553', 16: '138.183', 21: '156.801', 10: '192.187', 5: '19505.010', 2: '2
67.022', 1: '279.027', 22: '311.129', 8: '3368.250', 6: '3397.273', 17: '444.652', 18: '467.021', 3: '468.326', 19: '533.769',
13: '61.958', 7: '6448.237', 12: '75.785', 14: '80.527', 4: '804.332', 15: '93.483', 9: '937.095'}
************************************************************************
```

<u>Clustering Analysis (k- means):</u>

The goal of cluster analysis or clustering is to organize a collection of objects into groups that are more similar (in some ways) to one another than to objects in other groups (clusters). In many domains, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics, and machine learning, it is a frequent statistical data analysis technique and a key goal of exploratory data analysis.

Cluster analysis is a general problem to be solved, not a particular algorithm. Different algorithms that have quite different ideas of what clusters are and how to find them effectively can accomplish it. Popular definitions of clusters include collections of individuals with proximity to one another, crowded regions of the data space, intervals, or specific statistical distributions.

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

To find the K i.e., no of clusters we used Elbow method.

```
K_clusters = range(1,10)
kmeans = [KMeans(n_clusters=i) for i in K_clusters]
Y_axis = df1_loc[['start station latitude']]
X_axis = df1_loc[['start station longitude']]
score = [kmeans[i].fit(Y_axis).score(Y_axis) for i in range(len(kmeans))]
# Visualize
plt.plot(K_clusters, score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

Fig 15: Python code to calculate the number of clusters i.e., ‘K’ using Elbow method.

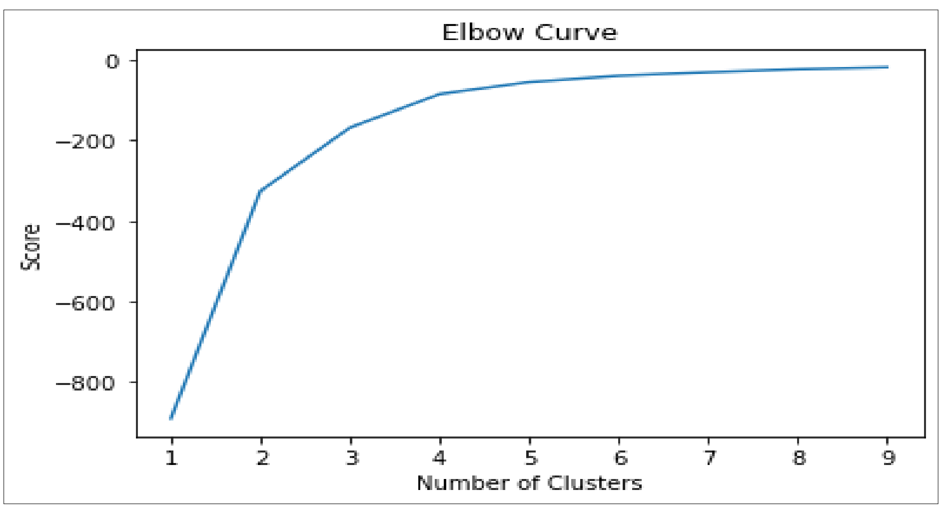


Fig 16: Number of Clusters Vs Score.

As shown below, using elbow method for k means clustering we successfully categorized various locations, where we can run our loyalty program into 3 clusters i.e., Clusters 0,1,2. Which further used in logistic regression model to predict which cluster is the best for running the loyalty program.

Here, k = 3 clusters. K means cluster classification with K = 3:

```
kmeans = KMeans(n_clusters = 3, init ='k-means++')
kmeans.fit(X) # Compute k-means clustering.
```

	start station latitude	start station longitude	cluster_label
0	40.778968	-73.973747	1
1	40.751873	-73.977706	1
2	40.785247	-73.976673	1
3	40.732219	-73.981656	1
4	40.727434	-73.993790	0
5	40.803865	-73.955931	2
6	40.784597	-73.949685	2
7	40.734546	-73.990741	0
8	40.676999	-74.006471	0
9	40.726218	-73.983799	1

After running the clustering algorithm, we can see below displayed results:

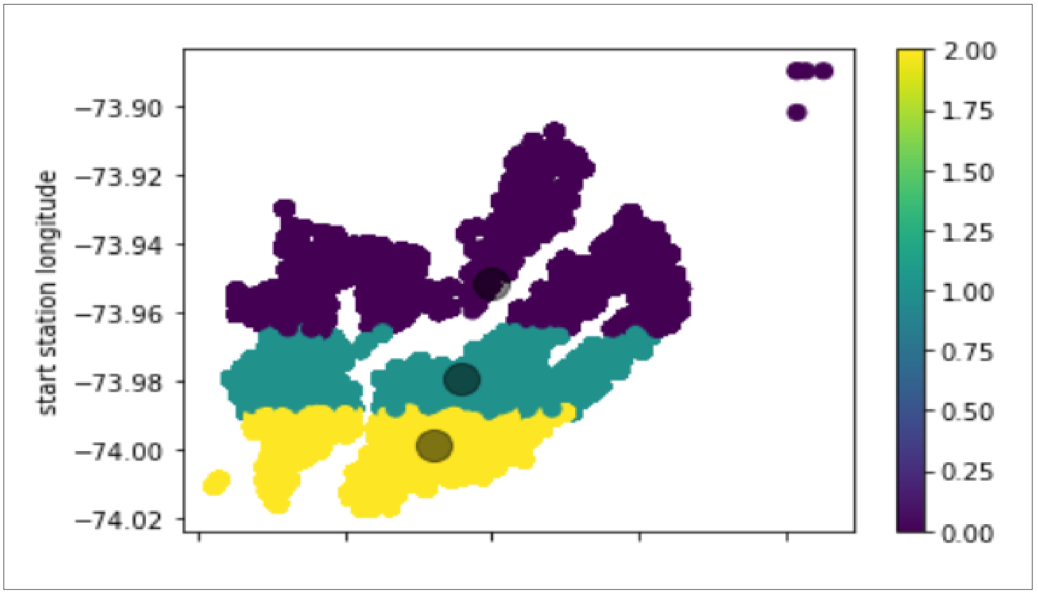


Fig 17: Categorized in three different locations as per the cluster 3.

Logistic Regression:

Logistic regression analysis is valuable for predicting the likelihood of an event. It helps determine the probabilities between any two classes. In a nutshell, by looking at historical data, logistic regression can help us to predict insights about or model.

We have used Logistic Regression to understand the relation between the Dependent variable and the Independent Variable and also to predict the significance of each variable in our predictive model.

Correlation analysis:

The main purpose of correlation is to allow experimenters to know the association or the absence of a relationship between two variables. When these variables are correlated, you'll be able to measure the strength of their association.

Overall, the objective of correlation analysis is to find the numerical value that shows the relationship between the two variables and how they move together.

Thus, to run the logistic regression successfully to see what variables are correlated with each other and can be dropped, we performed correlation analysis and the results can be seen below:

Fig 18: Correlation Matrix.

To improve the efficiency of the model, we dropped some variables including trip duration, latitude, longitude, birth year (age is present), other string columns and some calculated columns like distance, and speed. Finally, the variables we have decided to use in our model are:

	gender	age	speed	pickup_hour	day_of_week	cluster_label
0	1	48	0.200030	0	2	1
1	1	55	0.109729	0	2	1
2	1	32	0.206564	0	2	1
3	1	29	0.030978	0	2	0
4	1	40	0.260690	0	2	0

Fig 19: Variables to be used in model that were decided with help of correlation matrix.

Results on training dataset:

We have split our dataset into 70:30 (Training and Test accordingly).

```
Optimization terminated successfully.
         Current function value: 0.122735
         Iterations 9
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.261
Dependent Variable: usertype         AIC:              163486.5109
Date:               2022-11-26 14:59 BIC:              163554.9649
No. Observations:   665966           Log-Likelihood:   -81737.
Df Model:           5                LL-Null:          -1.1065e+05
Df Residuals:       665960           LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
------------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
gender             2.2582   0.0150 150.3136 0.0000  2.2288  2.2877
age                0.0091   0.0005  20.0081 0.0000  0.0083  0.0100
speed             16.5202   0.1217 135.7597 0.0000 16.2817 16.7587
pickup_hour       -0.0522   0.0013 -38.8372 0.0000 -0.0548 -0.0496
day_of_week       -0.1930   0.0035 -54.8555 0.0000 -0.1999 -0.1862
cluster_label     -0.0955   0.0090 -10.6606 0.0000 -0.1130 -0.0779
==================================================================
```

Fig 20: Logistic Regression Model results.

Based on the data related to user, our model predicts whether a customer can be converted to a subscriber or not. Since the output is a binary variable, prediction using Logistic Regression was best in our case. The R-squared value is 0.261 and **P-value** for all variables are most significant (P-value = 0).

Results on test dataset/ Sample data set to test our model:

```
yhat = result.predict(x)
prediction = list(map(round, yhat))
predicted_customer = pd.DataFrame(prediction)
# # comparing original and predicted values of y
# # print('A:', list(y_test.values))
# # print('P:', prediction)
```

```
final_file = r"D:\MS_Courses\3.Fall\Mining\Project\FINAL DATASET\1.REGRESSION_RESULT-FINAL.csv"
df['predicted_customer'] = predicted_customer
df.to_csv(final_file)
```

usertype	gender	age	pickup_hour	day_of_week	cluster_label	predicted_customer
0	1	34	13	7	0	0
0	1	25	14	2	0	1
0	1	29	15	2	1	1
0	2	50	6	1	2	0
0	2	69	0	4	2	0
0	1	22	14	2	1	0

Fig 21: Result after testing the model on test data set.

From above insights we can see that results are indeed surprising because the factors like gender and age are playing a dominant role in predicting whether our customer will be converted or not.

Also, results are surprising based on the timing as it completely overlaps with our ratio and variance model.

Model Validation:

While trying Linear regression (OLS) for our dataset, the output variable for user type was coming beyond 1 (30s, 40s, and even 100s). Upon putting a cap limit of 0 & 1. we see in Linear Regression (OLS) the error rate is >10%.

```
                         Results: Ordinary least squares
==================================================================================
Model:                   OLS              Adj. R-squared (uncentered): 0.955
Dependent Variable:      usertype         AIC:                         -197362.4708
Date:                    2022-12-03 09:56 BIC:                         -197294.0169
No. Observations:        665966           Log-Likelihood:              98687.
Df Model:                6                F-statistic:                 2.338e+06
Df Residuals:            665960           Prob (F-statistic):          0.00
R-squared (uncentered): 0.955            Scale:                       0.043533
--------------------------------------------------------------------------------------
                     Coef.      Std.Err.        t          P>|t|      [0.025     0.975]
--------------------------------------------------------------------------------------
gender               0.1879      0.0005     378.1814      0.0000     0.1870     0.1889
age                  0.0055      0.0000     321.8730      0.0000     0.0055     0.0056
speed                2.0194      0.0043     474.8365      0.0000     2.0110     2.0277
pickup_hour          0.0097      0.0000     207.3410      0.0000     0.0097     0.0098
day_of_week          0.0119      0.0001      89.1504      0.0000     0.0116     0.0121
cluster_label        0.0145      0.0003      43.5929      0.0000     0.0139     0.0152
----------------------------------------------------------------------------------
Omnibus:                 294221.761          Durbin-Watson:              1.995
Prob(Omnibus):           0.000               Jarque-Bera (JB):           1820639.367
Skew:                    -2.048              Prob(JB):                   0.000
Kurtosis:                9.988               Condition No.:              734
==================================================================================
```

usertype	gender	age	pickup_hour	day_of_week	cluster_label	predicted_customer
0	1	34	13	7	0	27
0	1	25	14	2	0	29
0	1	29	15	2	1	31
0	2	50	6	1	2	13
0	2	69	0	4	2	1
0	1	22	14	2	1	29

Whereas, Logistic Regression, we can see the Error Rate of 0.172 %, where we can see that 1's are getting converted to 0's.

usertype	gender	age	pickup_hour	day_of_week	cluster_label	predicted_customer
0	1	34	13	7	0	0
0	1	25	14	2	0	1
0	1	29	15	2	1	1
0	2	50	6	1	2	0
0	2	69	0	4	2	0
0	1	22	14	2	1	0

Thus, we have selected **Logistic Regression**, to achieve better accuracy in the final model.

Managerial Insights for providing promotion:

From both of our explanatory and predictive analysis we have seen that during office hours, there will be a greater number of subscribers who are daily users and who need the bike rides daily, but they may not may or may not be customers. Also, in the non-commuting hours there might be many customers who need a bike ride.

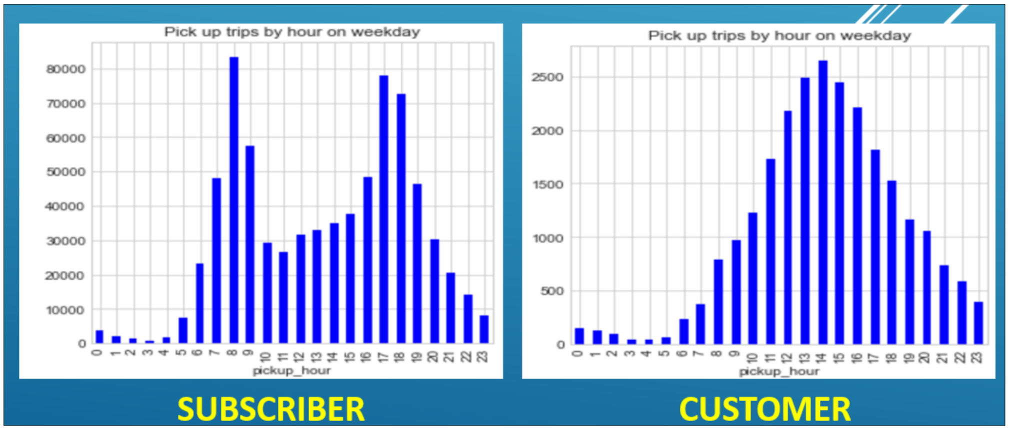


Fig 21: Subscriber Vs Trips and Customers Vs Trips

7. Conclusion:

So based on our model, we have come up with few best days, times, and locations to promote our loyalty program.

Best Day and Best Time: Tuesday and Wednesday have the highest conversion. Our target time would be 1.00pm to 3.00pm based on the conversions.

Day Count	Count	Daya
Count of 1	2592	MON
Count of 2	6779	TUES
Count of 3	5037	WED
Count of 4	3603	THURS
Count of 5	4221	FRI
Count of 6	4217	SAT
Count of 7	5456	SUN

Fig 22: Days with Highest Conversion.

Hour	Coversion
0	259
1	195
2	151
3	72
4	54
5	78
6	257
7	469
8	956
9	47
10	1731
11	2355
12	2917
13	3339
14	3371
15	3098
16	2714
17	2218
18	1815
19	1392
20	1191
21	874
22	637
23	424

Fig 23: Target time based on the conversion.

Cluster Segregation: Cluster zero has the higher conversion rate and that is why the top three locations in cluster zero are our targeted promotional locations. Later, we can expand our campaigns to other locations as well. And another insight from our analysis is there are many locations where bikes are not being picked up daily. They are occasional and even on those occasional days, only one or two bikes are being picked, so we can suggest our company that they could close off these locations as lower maintenance will lower their rental cost.

Count of 0	16478
Count of 1	9516
Count of 2	5910

Top 3 promotion Locations are:

1. 514-12 Ave & W 40 St
2. 3641-Broadway & W 25 St
3. 3255-8 Ave & W 31 St

Total conversion and Gender Segregation:

Total Customers	37560
Total subscr	913821
Total Conversions	31905

Male	22704
Female	9201

Also, we have noticed that 80% of customers who are born before 1980, and after 2002 are not getting converted. So, we tend to target age groups of relatively young people between 25-40 but not beyond as they have a higher chance of getting converted and male around this age has a higher chance of getting converted.