**CHICAGO DIVVY BIKE SHARING ANALYSIS**

*By Ananth Nath, Vijaya Rani, Amratansh Sharma, Jiangtao Xie, Syed Almas Rizvi, Himanshu Premchandani*

**Problem Statement**

Divvy is Chicagoland’s bike share system with 6,000 bikes available at 570+ stations across Chicago and Evanston. Divvy provides residents and visitors with a convenient, fun and affordable transportation option to explore Chicago, commute to work or school and many more. Divvy is available 24\*7 and riders have access to all bikes and stations across the system.

Divvy has decided to shut down its operation in Chicago due to Federal regulation. A new company 'Ecoride' would be launching its bike sharing business & would be replacing Divvy in Chicago. Our problem statement is to **explore & analyze the customers trip behavior** under different scenarios using Divvy's past 4 years data. We can help the new company analyze various factors that shapes the decision of riding a bike by customers and we can help EcoRide establish their business and minimize the supply and demand gap.

**The Objective of this Study**

There are a few opportunity areas & potential customer segments that Divvy was unable to tap. We can help Ecoride identify these segments that they should focus on to increase their customer base and maximize the profits.

We are helping Ecorideto understand the bike sharing market and help them strategize a plan to minimize the gap between supply and demand of the bikes at popular stations. We also want to come up with an analysis of the past data of customers to make Ecoride understand the customer behavior and the important factors which have some major influences such as geographical conditions, customer demographics, etc.

**Data Background**

We are taking trip data from Divvy. This data is available online openly for 2013-2017. We performed basic cleaning of this dataset first (dropping null values etc.). This dataset was then loaded into Jupyter Notebook and later filtered for the year 2017 and stored in a dataframe (named "datan"). This data contains 2.9 million rows and 23 attributes.

The original Divvy study available online was focused on answering questions like:

* How far do the riders go? Where do riders go? Most frequent time when they ride?
* The busiest pick-up & drop-off stations?
* Which days of the week are the busiest one?

**Computational Steps**

* Data was loaded using pandas, cleaned (rows with null values were dropped) and filtered(columns that were not required were dropped) to prep it for analysis
* We conducted an Exploratory analysis of the dataset. In this analysis we examined the following
* Number of trips by hour, day, gender, weather, temperature, usertype
* Trip duration by hour, day
* Reducing the data size to 1% of original dataset by using Random Sampling
* Plotting pair-plot and correlation coefficient of independent variables and extracting the relevant attributes
* Built Single & Multi variable Linear regression models to understand the effect of attributes such as time, duration, temperature and day on number of trips & distance travelled
* Applied K-means clustering to group the customers based on similar characteristics from the information given in the dataset
* Calculated Distance covered for each trip using longitude & latitude of start & end points for 2.9 Million observations
* Plotted Heatmap of the pick-up stations during peak/rush hour on the map of Chicago using Folium
* Plotted all the unique pick-up stations on the map of Chicago using Folium
* Plotted route maps of Top 5 pick-up stations using Folium
* Monte Carlo Simulation of the survey conducted for onboarding more female riders

**Computational challenges that we faced & proposed solution**

1. **Huge Data Size**- We had around 2.9 Mn observations which made it difficult for us to run some complex analysis like making a pair-plot or clusters out of it. To overcome this, we sampled 1% of the data, i.e., 29,000 observations to perform our analyses.
2. **Memory Error**- To classify hours into rush & non-rush hours, we had to create dummy values to identify rush hour as 1 and non-rush hour as 0. When we tried to do this using a loop, it was extremely heavy on the processor and our machine ran out of memory. We then later tried this using 'map' function, and it turned out to be very smooth. Similarly to calculate the distance between stations, we used apply() to reduce the time required for the computation.
3. **Plotting route-map & heatmap**- Due to the large data size, we had issues plotting the route map and heatmap of our dataset. To tackle this, we plotted them for the top 5 busiest stations (most popular starting & end points)

**Slowest part of our code:**

* Plotting pairplots of Independent variables
* Plotting route maps & heatmaps- While plotting the route & heatmap of bike trips, it took longer to map it as folium is using latitudes and longitudes of nearly 3 million points.
* Calculating the distance travelled for each trip using latitude & longitude of trip\_start location & trip\_end location- This code is taking around 90 seconds to run

**Time it takes for your code to run if the size of the data were to double, triple, quadruple**

Our dataset consists of ~2.9 Mn observations, **it's taking around 10-11 mins to run the whole analysis** that we have performed. Upon taking a sample of our dataset and running the entire analysis with twice, thrice & four times the sample size, we found that our code run time increases by around 2.5% each time we double the data-size. This time distribution is plotted in our code.

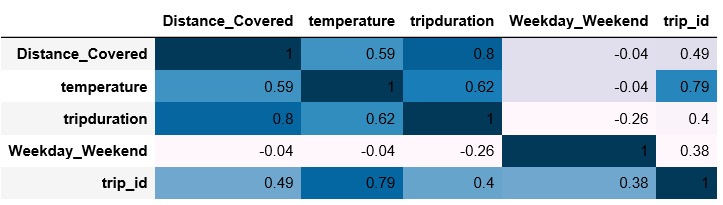
**ANAYSIS SUMMARY & RESULTS**

As per the initial analysis that we conducted, we can summarize our findings as follows:

* The number of rides have been catching on in the city through the years. Number of trips has increased by 86% from 2013 to 2017.
* Most of the trips are taken when people are getting to work/school (8 am and 5 pm) during Weekdays.
* We also know that the best day to get a ride hustle free are Saturdays and Sundays, the rest of the week is fairly busy with Tuesdays being the busiest.
* Summer is the busiest season for the bike renting business with the number of rides peaking in August.
* Canal St and Adams St is the busiest station – maximum no of trips start and end at this station.

1. **Correlation matrix**

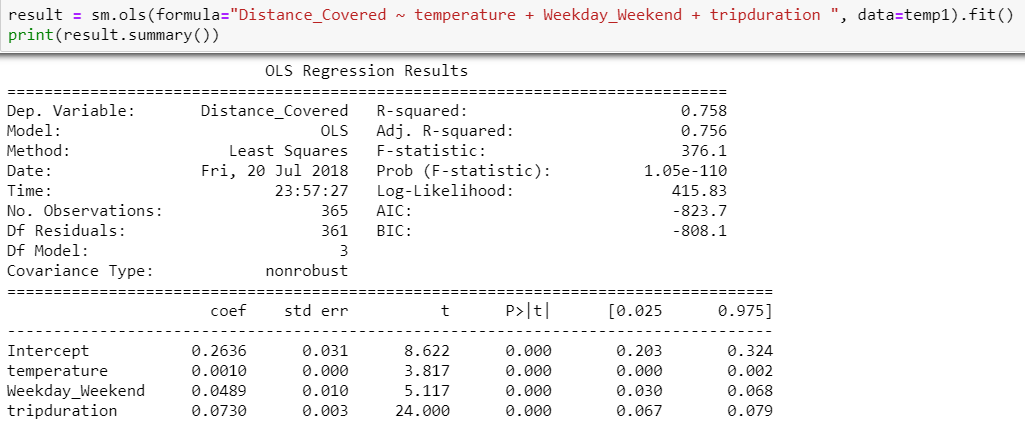
To get an overview of correlation between different independent variables, we plotted the correlation matrix. These results were used to pick the variables for linear regression model



1. **Regression (Single & Multiple)**

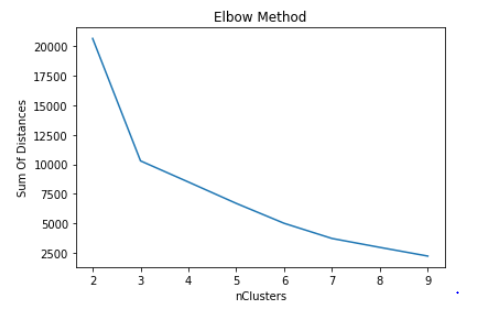
* We built single & multi variable regression models to study the impact of few relevant attributes (using correlation matrix) such as trip duration, temperature, gender, weather, gender on number of trips & distance travelled by the customers
* By running single variable regressions, we found significant positive linear relationship between **number of trips** and **temperature** with an R-sq value of 29%. All other models had R-sq values of less than 10%.
* Multi variable Regression results

1. **Number of trips** is highly correlated to **temperature** & **Weekend/Weekday** with an R-sq value of ~77.5%. Both temperature & Weekend/Weekday have significant positive linear relationship with the number of trips that the customers take
2. Another significant multi variable model is **Distance covered** is highly correlated to **temperature**, **Weekday/Weekend** & **trip duration** with an R-sq value of 75.8%.



1. **Clustering (K-means)**

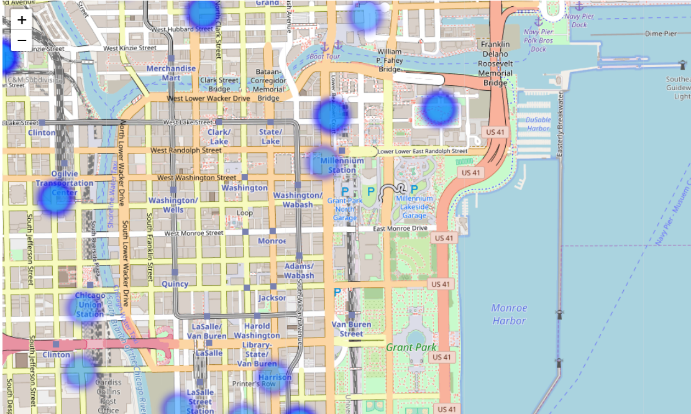
* We conducted K Means clustering to identify group of customers with similar characteristics
* We transformed distance, gender, hour and weekend/weekday variables into categorical variables by introducing some dummy variables.
* The number of Clusters is approximately 8 if we take gender, hour & Weekend/Weekday variables while running K-means clustering. On further adding another variable that was the Distance covered by the rider (Whether higher than or lower than 2 miles)we observed the following
  + We get approximately 3 clusters which implies that distance plays an important role in characterizing the type of users.



1. **Route-map/Heatmap**

We plotted a heatmap of rides during rush hours. It helped us to identify the stations where the company should try to maximize their supply to meet customer demands.

* Canal St & Adams St
* Clinton St & Washington Blvd
* Clinton St & Madison St
* Kingsbury St & Kinzie St
* Franklin St & Monroe St



**CONCLUSION**

Based on the above analysis, we come up with several recommendations:

1. Busy stations like 'Canal St & Adams St' need proper inventory so that high demand can be met. We can shuttle bikes from the stations that are the most frequent target stations from Canal Street to meet the demand.
2. We should exploit the significant amount of difference between number of male and female customers. Incentives can be provided to females for onboarding them.
3. The number of rides on weekdays are significantly higher than weekends. Hence, we should try different pricing models to increase profits over weekends
4. **Distance covered** is highly correlated to **temperature**, **Weekday/Weekend** & **trip duration** with an R-sq value of 78%.

**Interesting hypotheses that would be interesting to investigate further**

1. Dynamic surge pricing based on rush hours is an interesting model which can be implemented. Higher the demand, higher should be the price for the ride
2. Investigating demographic variables like age, profession, address, family size etc. might help us in estimating trip & rider behavior
3. If we can classify the stations as "business", "shopping center", "tourist", "educational institute", etc, we can use this information to train our model to give a more precise clustering result.
4. To target potential female riders. As per the current data, males are 3 times more likely to ride a bike than female. Female segment is an untapped market which we must further explore to increase company's customer base