BUDT704 Project Proposal

Bike Sharing Analysis

**Group 15 [Int Elligence]:** Kai-Hsiang Lin, Shubhangkar Girish Jain, Marco Sesay, Samrat Leuva, Shivakumar Vastrad, Yanjie Chen, Bradley R. Mascarenhas

# Introduction

Bike-sharing systems have been a go-to mode of transportation for all the major cities around the world. It provides a fun, interesting, and expedient way to explore the city, in addition to solving the last mile connectivity problem along with it being environmentally friendly. But there are situations when people are not able to find a bike and have to rely on more expensive and inconvenient options to perform their work. To solve this problem, we took a dataset of the Chicago bike-sharing system, **Divvy** to understand how different time periods and weather influence the demand for rides and trip duration for various user types and ensure adequate bikes are available across the city so that everyone has the option of reducing their carbon footprint by using Divvy.

## Questions of Interest

1. Analyze how temperature influences the count of trips and trip duration on subscribers and non-subscribers. Identify the range of temperature that creates the most demand.
2. Identify the top 10 most popular stations for both start and end stations, and pinpoint them on several maps to analyze the driving factors that make them the most popular ones.
3. Analyze the riding behaviors of subscribed and unsubscribed users during national holidays and non-holidays, and see whether the month/season of the holidays will influence the trip number and duration.
4. Predicting the demand of bikes based on the time (hours) in a day throughout a week - both weekdays and weekends

# Data Processing and Analysis

## Dataset Description

The data set for the project is acquired from Kaggle. The user who uploaded the data, pooled information from the “Divvy” portal and corresponding weather details from the “wunderground” website.

Source:

* Divvy Data: <https://www.divvybikes.com/system-data>
* Weather Data: <https://www.wunderground.com/>
* Kaggle: <https://www.kaggle.com/yingwurenjian/chicago-divvy-bicycle-sharing-data>

The data is taken for the period 2013 - 2017, for the purpose of this project we are making use of the raw data set presented by the user. The data in particular has two sections to it, the first half corresponds to the details of the bike sharing platform “Divvy.” It includes the following variables:

1. Categorical variables: trip\_id, usertype, gender, from\_station\_name, from\_station\_id, to\_station\_name, to\_station\_id.
2. Numeric variables: starttime, stoptime, tripduration, dpcapacity\_start (docking point), dpcapacity\_end (docking point), latitude\_start, longitude\_start, latitude\_end and longitude\_end.

The second half of the data, corresponding to the details of the weather includes:

1. Numerical variables: temperature, windchill, dewpoint, humidity, pressure, visibility, wind\_speed and precipitation
2. Categorical variables: events and conditions

Data Processing Tasks

* Select out a specific year, which is 2017 to try to reflect the situation now
* Extract year, month, week, date and hour from start\_time and end\_time and create new columns accordingly
* Handle missing values: replace them with feasible numbers (mean or median)
* Calculate the frequency of trips and the average trip duration based on the variables
* Group the temperatures by every 10 degrees for each bar
* Identify unique values of events (weather conditions) and user types
* Identify all federal holidays
* Segment a week into weekdays and weekends and create a column accordingly

## Data Analysis

1. The visualization for our analysis should be a dual-axis chart with bars and lines. Temperature should be the value on the x-axis, and trip count and trip duration should be on the two sides of the y-axis. The trip count should be in bars, and average trip duration should be in a line.
2. Create two separate maps to locate the top 10 starting and ending stations. Based on their actual locations, come up with the reasons and factors that lead them to become the most popular stations.
3. Select out the weeks with holidays and those without holidays, and create a dual-axis bar chart with two colors representing subscribers and non-subscribers.
4. Create a line chart. The lines themselves will visualize how the demand of rides changes within a day, and the colors of the lines will represent weekdays and weekends for comparison.

# Expected Findings

1. As the temperature becomes more ideal and comfortable, the number of trips and trip duration will increase. As the temperature becomes more extreme, the number of trips and trip durations will decrease. We also believe that the influence of temperature will be weaker on subscribers than on non-subscribers.
2. Top 10 starting stations should be close to the business/financial center or crowded areas. Since bikes serve as an intermediate medium of transportation, we believe that Top 10 ending stations should be located close to major transportation hubs like metro or bus stations.
3. We believe that during holidays, the number of trips for non-subscribers will be higher than that of subscribers because a large number of tourists travelling to Chicago for vacations will create a surge in demand. We also think that we will see a higher number of trips and trip duration during holidays in spring or fall than those in winter or summer (the weather tends to be more extreme) for both subscribers and non-subscribers.
4. We came up with a prediction that the time with the most demand during weekdays should be around 7-9am and 5-7pm when riders are typically riding to and off work. As for weekends, the time with the highest demand is supposed to be a little later, which is around 11am-1pm and 4-6pm when the temperature is the most comfortable.

# Project Timeline

|  |  |  |
| --- | --- | --- |
| **Task** | **Task Lead** | **Due Date** |
| 1. Extract year from starttime, select out records from 2017 | Shubhangkar | 10/24 |
| 1. Handle missing values | Sean | 10/29 |
| 1. Extract months, weeks, dates and hours and create new columns accordingly | Shivakumar | 11/5 |
| 1. Bin the temperature in steps of 10 degrees | Marco | 11/5 |
| 1. Identify unique values in events and user types | Marco | 11/6 |
| 1. Segment a week into weekdays and weekends and create a column accordingly | Yanjie | 11/6 |
| 1. Store all federal holidays in a list | Samrat | 11/10 |
| 1. Calculate the frequency of rides and average trip duration based on the variables that we are analyzing | Bradley | 11/10 |
| 1. Analyze and prepare reports for question of interest 1 and 2 with supporting visualizations | Samrat | 11/26 |
| 1. Analyze and prepare reports for question of interest 3 and 4 with supporting visualizations | Yanjie | 11/26 |