Sayra Moore & Caroline Sklaver
DATS6103 Group Proposal
10/30/2019

For the final project, we will be performing Exploratory Data Analysis, or EDA for London bike share use in 2015. We hope to determine the influence various factors have on city-shared bike use in London.

The dataset was originally retrieved from the UK government site 'Transport For London', and was merged with weather data from freemeteo.com/i-weather.com by timestamp, and was all made available as one dataset on kaggle.com. The dataset is made up of 10 columns including: timestamp, cnt (number), t1 (temperature), t2 (temperature feels), hum (humidity), wind_speed (km/h), weather_code, is_holiday (boolean), is_weekend (boolean), and season (code). From what we have observed so far, the data does not need to be cleaned since there are very few missing values.

Although we have not yet covered all of the possible data mining algorithms in our lectures, as of right now we believe we could use decision trees to determine the most influential variables when people decide to use bike share or not. On the other hand, we could use regression analysis to find which features predict the quantity of bike share usage and to what magnitude. In order to predict future bike share use based on weather, season, etc., we could potentially use K Nearest Neighbors (KNN). To do this, we will likely be implementing all of the Python packages we covered in class including: numpy, pandas, matplotlib, seaborn, pydotplus and sklearn, specifically to visualize and analyze the data.

Our reference materials that we will be applying will include lecture slides/notes, course textbooks, and information on the web. To judge the performance of our results we hope to perform statistical tests with which we can formulate null hypotheses and either accept or reject those hypotheses based on p-values, or alternatively, gini/entropy values and other measures of accuracy for any other algorithms we apply. Other metrics of evaluation may included graphical visualizations or correlations.

A tentative schedule for this project includes all cleaning and descriptive statistics completed by the end of this week (11/03), then we will spend the next two/three weeks on applying algorithms. Afterwards, we will complete the remainder of coding/statistical testing the week of 11/18. Lastly, the week of 11/25 we should be compiling all of our information and results into presentable format to then prepare for the final presentation on 12/04.