**DATS 6103 Data Mining Fall 2019 Project Proposal**

**Presented by:** Ziya Zhao, Kaiqi Yu, Zichu Chen

1. **Introduction**

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back to another position. Currently, there are about over 500 bike-sharing programs around the world which are composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a **virtual sensor network** that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

**2. Data description**

There are two datasets we will use:

a. London: The CSV contains over 17,000 records with timestamp, number of bike, temperature, somatosensory temperature, humidity and wind speed of the hour, weather of the day, if the day is holiday, if the day is weekend and season of the day. All of the data is related to shared bicycles and weather collected in London from January 4, 2015 to December 31, 2016.

The data from cycling dataset is grouped by "Start time", this represent the count of new bike shares grouped by hour. The long duration shares are not taken in the count.

Variables:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Timestamp | Record Time |
| Cnt | Number of bike |
| T1 | Temperature |
| T2 | Somatosensory temperature |
| Hum | Humidity(%) |
| Wind\_speed | Wind speed(km/h) |
| Weather\_code | Weather(1-26) |
| Is\_holiday | The day of record is holiday(0/1) |
| Is\_weekend | The day of record is weekend(0/1) |
| Season | Season(0/1/2/3) |

b. Washington, D.C.: Bike Sharing in Washington D.C. Dataset has similar variables. Please use the link below:

<https://www.kaggle.com/marklvl/bike-sharing-dataset#hour.csv>

**3. Goals of analysis**

Ideally, we want to investigate how does Bike sharing usage in metropolitan area relates to weather factors and potentially holiday and weekend effects. Also, with bike sharing usage dataset from two capitols of the US and the UK. We can compare the usage pattern difference. Specifically, we want to test the following hypothesis:

1. Bike sharing usage has a non-linear relationship with city temperature. There might be a threshold effect where they have a positive correlationship below a certain temperature but a negative correlation after that.
2. There are holiday and weekend effects on the usage of bike sharing. Usage during holidays or weekend is larger compared to normal weekdays.
3. There should not be significant difference on the above results between the Washington D.C. dataset and London dataset.

**4. Analysis methods and tools**

We will use VSstudio and present statistical test, performance analysis and data visualization: scatter plot, bar plot, etc. Also, we will do Regression analysis: regression model, threshold model and in-sample prediction