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. The welcome public service which allows people to borrow a bike from a dock and return it in another place. Until 2017, there’re over 500 bike sharing systems with 500,000 bikes around the world. 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. Different from other transport services, the duration of trip, start and end time arrival positions are explicitly recorded in these systems. These features make bike sharing system a virtual sensor network which can be used for sensing mobility of a city. Through this research, we expect to find any factors that influence people’s choice of using shared bicycles or not and predict how many times the bike sharing system would be used in a time period in Washington, DC.

1. Data Description
   1. Data Source

There’re 3 data resources we would use:

1. 2011-2012 Bike Sharing in Washington D.C. Dataset from Kaggle: This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in [Capital bikeshare system](https://www.capitalbikeshare.com/system-data) in Washington, DC with the corresponding weather and seasonal information. There’re 17379 rows in this dataset and each row is a record for an hour in year 2011 and 2012. The variables are shown below:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Instant | Record index |
| dteday | Date |
| Season | Season (1:springer, 2:summer, 3:fall, 4:winter) |
| yr | Year (0: 2011, 1:2012) |
| mnth | Month (1 to 12) |
| hr | Hour (0 to 23) |
| holiday | weather day is holiday or not (extracted from [Holiday Schedule](http://dchr.dc.gov/page/holiday-schedule)) |
| weekday | Day of the week |
| Workingday | If day is neither weekend nor holiday is 1, otherwise is 0 |
| weathersit | * 1: Clear, Few clouds, Partly cloudy, Partly cloudy * 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist * 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds * 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| temp | Normalized temperature in Celsius. The values are derived via:  (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) |
| atemp | Normalized feeling temperature in Celsius. The values are derived via: (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale) |
| hum | Normalized humidity. The values are divided to 100 (max). |
| windspeed | Normalized wind speed. The values are divided to 67 (max). |
| casual | Count of casual users |
| registered | Count of registered users |
| cnt | Count of total rental bikes including both casual and registered |

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

1. 2017 Bike Sharing in Washington D.C Dataset from capital bikeshare: With year 2011-2012’s sharing bike dataset as training set, we would use year 2017’s data as test set. This dataset contains trip history data in year 2017 from capital bike share. There’re 8738 rows in this dataset and each row is a record for bike sharing. The variables are shown below:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Duration | Duration of the trip |
| Start Date | Includes start date and time |
| End Date | Includes end date and time |
| Start Station | Includes start station name and number |
| End Station | Includes end station name and number |
| Bike Number | Includes ID number of bike used for the trip |
| Member Type | Indicates whether user was a "registered" member or a "casual" rider |

Link: <https://www.capitalbikeshare.com/system-data>

1. 2017 Daily Observations of Arlington County, VA Weather History from weather underground: The second dataset lack corresponding weather information so we use web scraping tools to build the third dataset as a supplement. This dataset contains weather information of each hour in 2017 and have these variables:

|  |  |
| --- | --- |
| **Name** | **Description** |
| Date | Date |
| Hour | Hour |
| TF | Feeling temperature |
| Humidity | Humidity |
| WindSpeed | Windspeed |
| weather | Description of the weather |

Link: <https://www.wunderground.com/history/daily/us/va/arlington-county/KDCA/date/2017-1-1>

The second and third dataset were combined to build a dataset with same important variables as that of the first dataset.

1. Data pre-processing

The 2011-2012 bike sharing in Washington DC dataset was provided and cleaned by Kaggle contributor Mark Kaghazgarian while the second data was a raw dataset with no weather information. Also, each row in the second dataset is a record of a specific bike sharing while in the first dataset, all the bike sharing records were counted in hours. So we grouped the data in the second dataset according to the variable Start Date. We also dropped the variables which are not contained in the first dataset.

Then we used chrome driver and beautiful soup to scrape the observed weather information in 2017 from the website weather underground. It took about 6 hours to scrape all the information because loading the website takes a longtime. The normal observe time of this observation station was on the 52nd minute in each hour so we chose all the normal observations and drop others to avoid repetition.

The second and third datasets were combined and converted into a new dataset named as “hour2017.csv”. This dataset has a structure same as the first dataset and has only time, weather information and count of total rental bikes. Also we normalized the feeling temperature, humidity and windspeed in the same way as the description of the first dataset.