My analysis is going to be on the Bike Sharing Dataset from the UCI Machine Learning Repository. The data comes from two years' worth of records in the Capital Bikeshare program in Washington, D.C. This program is setup much like Zipcar in that a renter picks up a bike from a location in the city, uses it for a duration, and drops it at the same or another location.

The Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, Portugal contributed the data set after performing their own analysis. Their goal focused on detecting city events by the usage rate of the bike sharing program. From their own statement, "*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.*"

The data set is composed of 16 independent attributes. These attributes fall into two main categories: day-related and weather-related. My interest is determining if a correlation exists between the day of the week and ridership and/or between weather factors and ridership. For this reason, I created a dependent variable *ridersup* with labels of UP and NOTUP. To assign labels, I simply took the median of the daily total attribute and used the rule "If the total is greater than the median, assign UP. Otherwise, assign NOTUP." The method of creating the dependent variable may need to change over time as data is analyzed further.

Basic analysis shows some interesting features. From Figure 1, it is clear that the general trend in ridership is up, but consistent dips appear. Those dips are ripe for analysis on weather conditions and holidays.

The histogram of the dependent variable (Figure 2) shows slightly more days with increased ridership than not. This result is not so surprising since the variable was created by taking a median. Given that the data comes from the start of the program through year 2, it may be more interesting to designate some period as the "early adopter" period and use it as the basis for determining UP and NOTUP rather than the use of median.

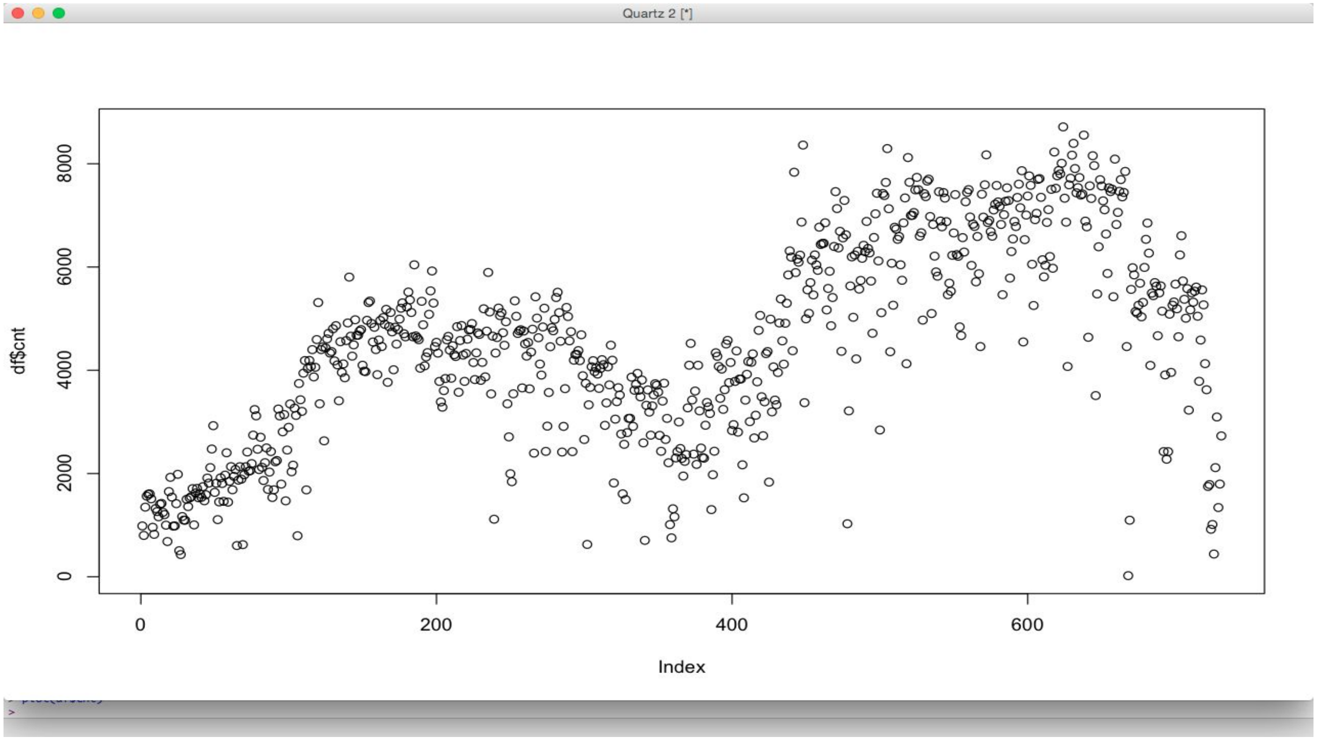
An interesting attribute is "weathersit". From the data description section of the data set page:

- *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*

Figure 3 show the plot for this particular attribute. The absence of points at the 4 level shows a climate more suited to bicycling than northern climates of the US Northeast and Midwest in Winter. Most points land in the 1-2 level. Interesting to note is that this attribute makes it easy to see the effects of snow and ice - conditions related to cold - but it does not help with conditions related to being too warm that may affect ridership. In order to gauge the effects of negative warm conditions - high temperature and humidity - the use of attributes for temperature and humidity need to be used.

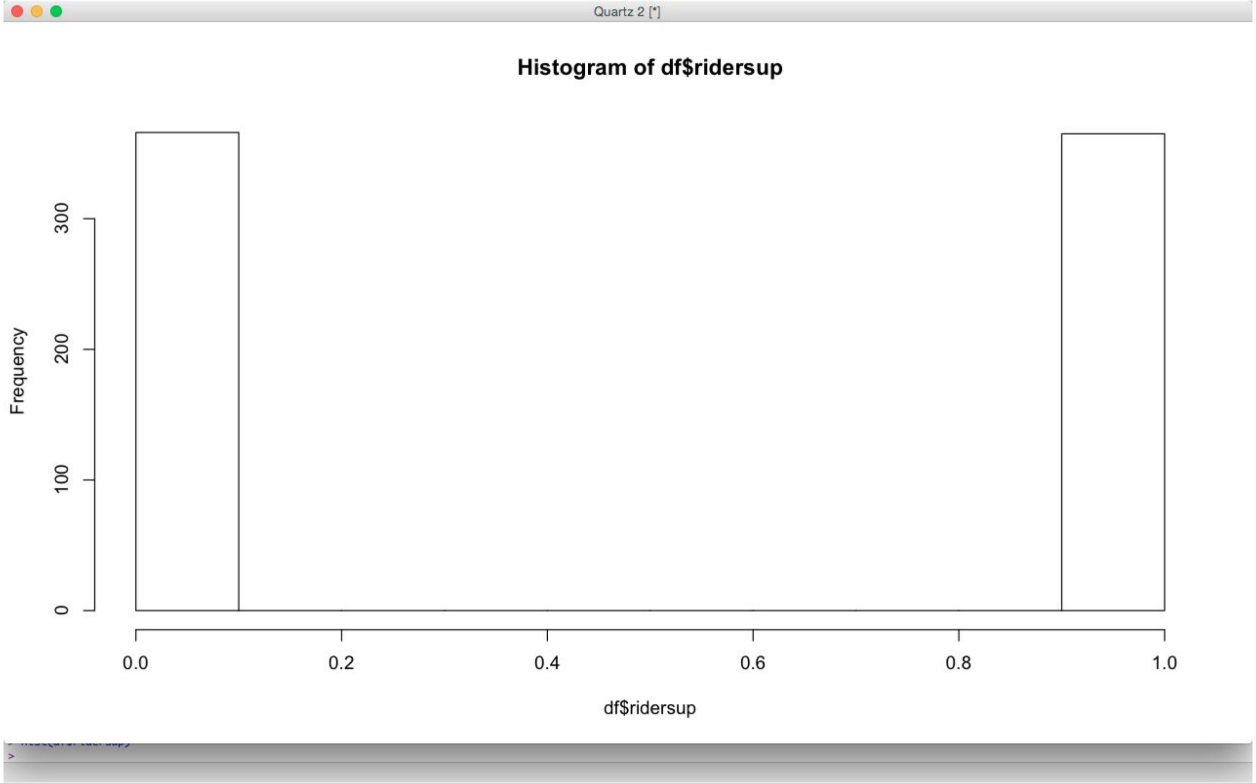
Weather is half of the analysis on ridership patterns. The other half is day of the week and holidays, both of which are independent attributes. An interesting hypothesis to test would be holidays landing on certain days of the week yield an increase in ridership, outside the trend for days around it, whereas, holidays landing on other days do not.

Areas for future study could be correlations between drop off location and social establishments like pubs or concert halls. Also, a low density of pick ups for a particular location could lead to a study of security or population density affecting pick ups. However, neither attribute is part of the provided data set. A search on the *capitalbikeshare.com* program site shows a link provided to get firsthand data from the organization.



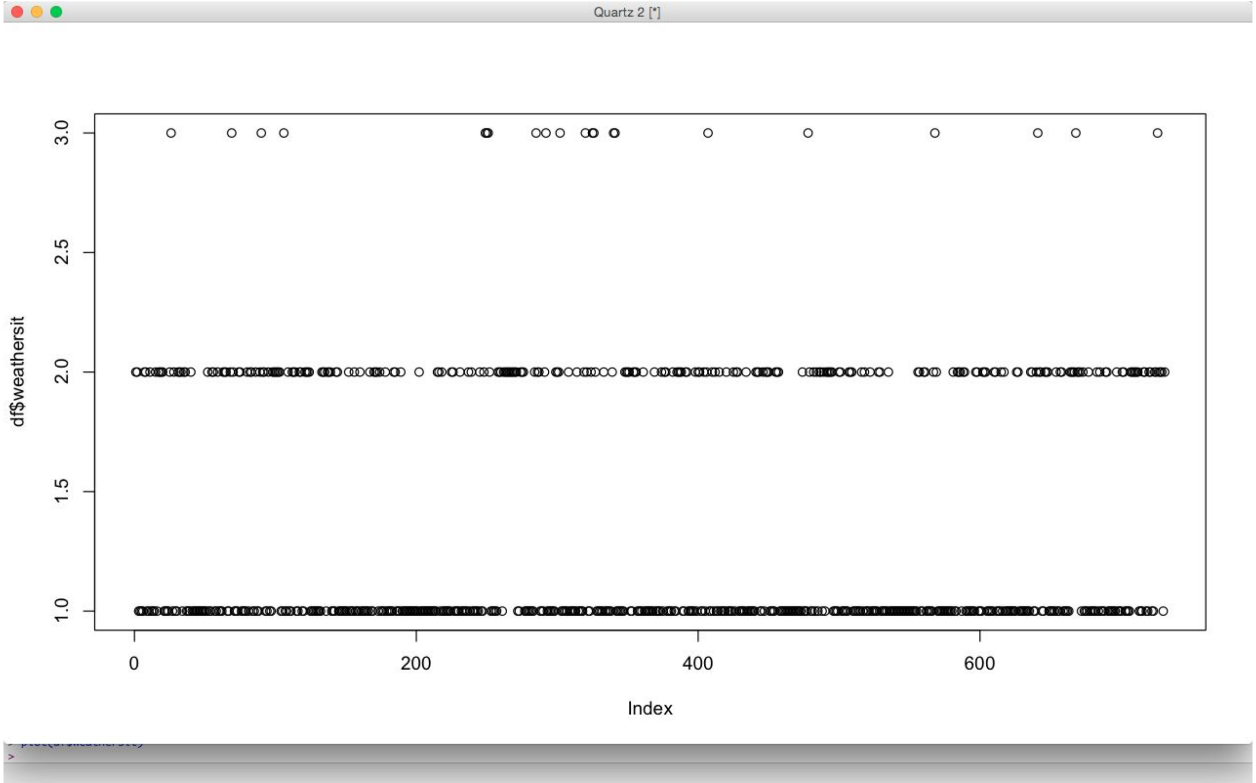
```
> plot(df$cnt)
>
```

Figure 1. Daily Ridership Plot



```
> hist(df$ridersup)
>
```

Figure 2. Riders Up/Not Up Histogram



```
> plot(df$weathersit)
>
```

Figure 3. Weather Situation Plot