Tingyu Yang
Mako Gedi
Final Project
Data Science

Proving the Dependence of Bike-Renting Behaviors on Environmental Conditions

Our final project set out to study the correlation between bike-renting behaviors and environmental factors including normalized temperature, humidity, wind speed and various weather conditions ranging from clear skies clear skies (denoted by a value of 1) to ice pellets and thunderstorms (4). With environmental conciousness on the rise, world interest in sustainable transportation is growing exponentially. In fact, global research indicates that as of the year 2013, there were more than 500 bike-sharing systems around the world (Fanaee-T et al., 2013) and the number has nearly doubled today (DeMaio, 2016). The rise in communal bike-renting behavior can be in part attributed to convenience, or rising interest in sustainability but another unique facet to consider is that contrary to public transportation systems like buses, trains or the subway, bike-sharing systems are a less standard form of transportation, and thus rely less on the rentor needing to get from point A to point B and more on the renter feeling like renting a communal bike. But is communal bike-renting likely to become a standard means of transportation in the near future? We predict that this really depends on environmental factors. That is, we hypothesize that exposure to harsh environmental conditions plays a major role in decreases in bike rentals.

The data set we used to study this correlation is from the *University of California Irvine Machine Learning Repository*. The particular file we used is a daily log for biking-sharing system that encapsulates rental activity data (among other variables) for every single day between the year 2011 and 2012 in Washington, D.C.. In total, the file contains 731 datapoints per variable. With this data, we anticipate proving our hypothesis that although this new environmentally-friendly form of transportation is convenient and growing, it is also easily undermined by the environment itself.

Statistics and Inference

We decided that the first step to take in working with our data was to use python to calculate statistical parameters including median, mean, and standard deviation. We did this by referring back to the Statistics and Inference assigment we had in Data Science.To be able to obtain the parameters we set out to obtain, we had to first choose which of the 17 variables (columns) in the file were relevant to our experimental question: Does extreme temperature effect on bike-renting behavior? We decided the most relevant of the variables were temperature, feeling temperature, weather conditions, humidity, windspeed and total renters.

Next, we set out to find the statistical parameters of the total number of renters in relationship to “hot”, “cold”, and “pleasant” temperatures. To do this we first had to go through the temperature data and see which ranges of temperatures cold be classifed into each of the three categories. Ofcourse the data was normalized (and in celcius) so we had to do a little backtracking to figure out what we had to do to the temperature data we had to get it into a form that we could understand. We looked in the supplementary text file that accompanied the data and found that the data was normalized by a factor of 41. We then added a column that multiplied all the given values by 41 and from the data in this column we were able to get a better idea of the total range of temperatures and select from there what range of values correlated to our classifications: “cold”: < 10℃, 10℃ < “pleasant” < 20℃, “hot”: > 20℃. From here we noted the normalized temperature values correlating to the degrees celcius parameters we set, and used a for loop to set the three ranges we wanted to look at. After setting the ranges, we obtained the following statistical parameters from the dataset:

Number of Renters	Total	Temperature “cold”	Temperature “pleasant”	Temperature “hot”
Mean	4504	1678	3692	5613
Median	4548	1530	3649	5339
Standard Deviation	1937.2	733.5	1608.9	1470.9

Table 1. Statistical parameters show trend of bike rentals increasing with temperature.

Data Visualization

From looking at ranges in termperature alone we were able to conclude that the total number of renters nearly doubles during pleasant temperature as compared to cold, and are even higher during hotter temperatures. Next we dived into generating visualizations of the data. First we looked at the relationship between bike rentals and specific weather conditions ranging in hazardousness from clear skies to ice pellets and thunderstorms (Figure 1, Table 2).

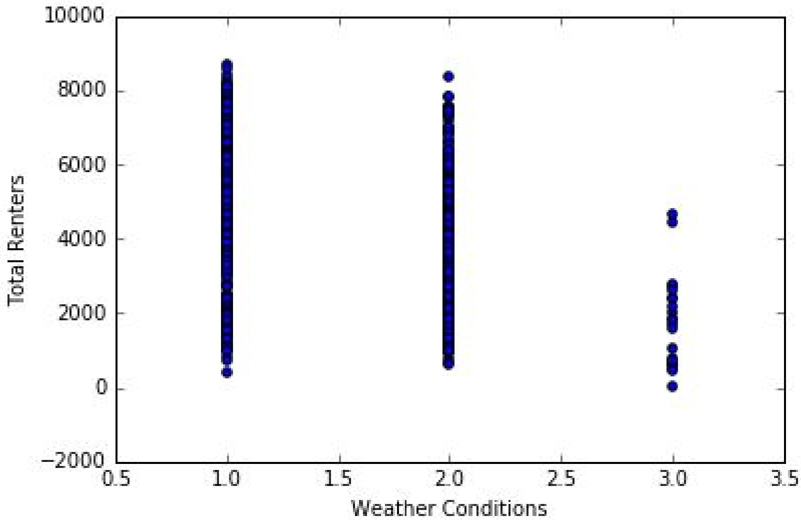

Figure 1. Bike rentals decrease with increasing hazardousness of weather conditions. Correlation of weather condition and total bike rentals is -0.29739. No weather conditions recorded to be level 4 in Washington D.C. between the years 2011 and 2012.

1	Clear, Few clouds, 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

Table 2. Key for weather type denoted by number.

Our next step was to obtain correlation plots to better visualize the relationship between bike rentals and temperature. For this plot, we decided that feeling temperature would be more effective of a measure than actual temperature (Figure 2). However, to be as clear as possible, we also re-generated the correlation plot showing both feeling temperature (yellow) and actual temperature (blue) plotted against total bike rentals (Figure 3).

Tingyu Yang
Mako Gedi
Final Project
Data Science

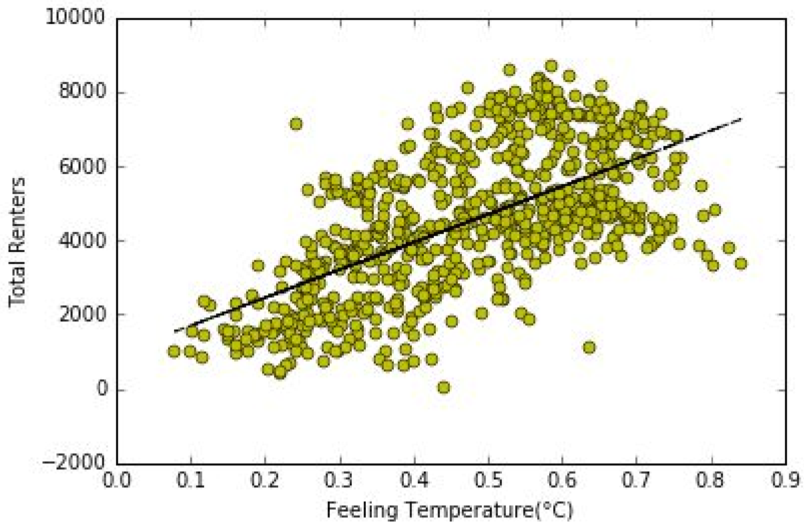

Figure 2. Bike rentals increase with increasing feeling temperature. Correlation of feeling temperature and total bike rentals is 0.63107.

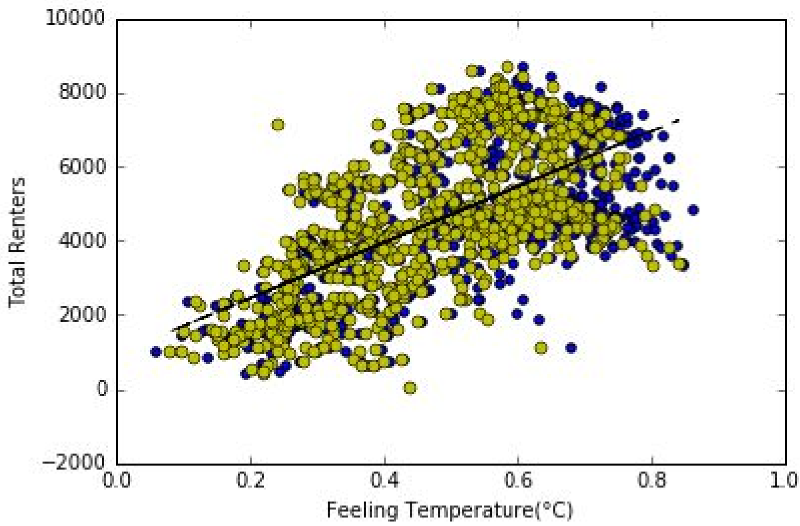

Figure 3. Bike rentals increase with increasing feeling temperature and actual temperature. Correlation of feeling temperature and total bike rentals is 0.63107.

Additionally we generated correlation plots to investigate the relationship between humidity and bike rentals (Figure 4), and windspeed and bikerentals (Figure 5). Both variables were very weakly negatively correlated with the total number of bike rentals.

Tingyu Yang
Mako Gedi
Final Project
Data Science

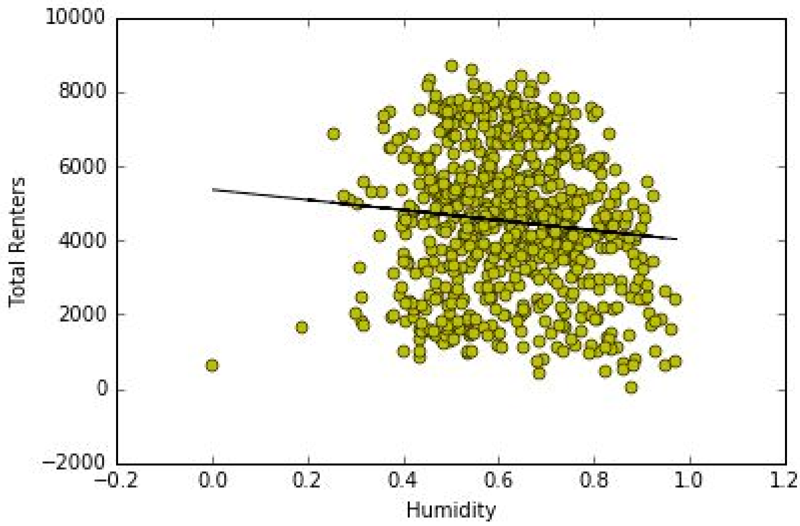

Figure 4. Bike rentals slightly decrease with increasing humidity. Correlation of humidity and total bike rentals is -0.10066.

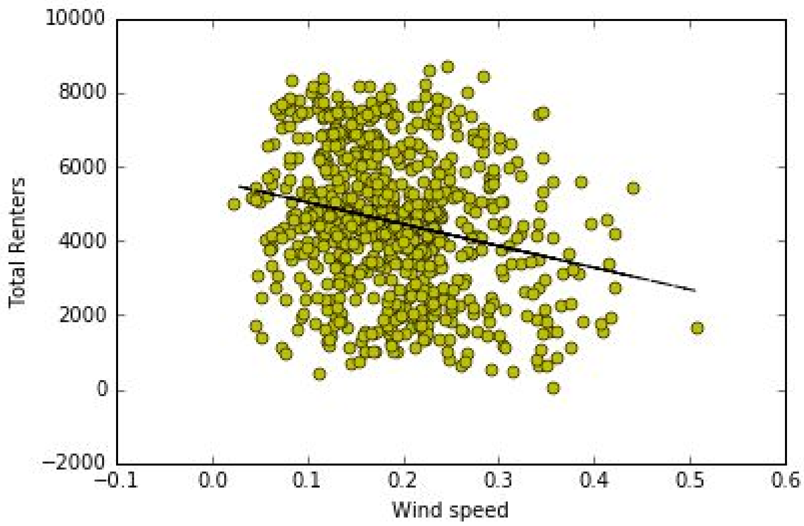

Figure 4. Bike rentals slightly decrease with increasing wind speeds. Correlation of wind speeds and total bike rentals is -0.23454.

Further Analysis: Multiple Regression Model

Lastly, we tried to use the multiple regression model to generate predictions of total bike rentals in considerion of these variables: weather, feeling temperature, humidity, and wind speed. The functions we used generated the following model:

Tingyu Yang
Mako Gedi
Final Project
Data Science

*Total renters = 0.8444218515250481+ Weather Situation*0.7579544029403025 + Feeling temperature *0.420571580830845 + Humidity* 0.25891675029296335 + Wind Speed *0.5112747213686085*

In order to make the function to be more accurate, we used the first 600 datapoints to train, making sure to use the random function so that the data will not be influenced by its given order (date). We saved the last 131 datapoints so that we would have some data to compare the predictions our model made with in order to assess the accuracy of our regression model.

Unfortunately, we found the predictions made by the model to be completely inaccurate when compared to the 131 datapoints we saved. We hypothesize that this was because the model of the functions and dataset we used were too big. We are not planning using the flawd predictions to support our case in this paper. However, we include a discussion of the model to put forth all of the strategies we utilized in our investigation.

In conclusion, based on our study, we can confidently state that the feeling temperature/actual temperture is positively correlated with the number of bike-rentals per day. More specifically, we conclude that most bike rentals occur when the weather is anywhere ranging from pleasant (10-20℃) to relatively hot (>20℃). Since the correlation between the temperature and total rentals is fairly close to 1, we can agree that the influence of temperture on renters is relatively strong. On the other hand, the relationship between wind speed or humidity with total rentals is negatively correlated. Thus, the more humid it is or the higher the wind speed, the less people tend to rent bikes. However, one thing to consider is that the (negative) influence of wind speed on renters is relatively bigger than the influence of humidity. Additionally, by examining the data points between weather conditions and total renters, we can speculate when the weather condition is at 1, which means the day is clear, there are likely to be more bike rentals. Moreover, the calculation of correlation between the weather condition and number of bike rentals indicates that harsher weather conditions yield a decrease in bike rentals.

All in all, it seems to us that despite not having been able to effectively predict the number of bike rentals using the multiple regression model, we have done what we set out to do, which was to prove that bike rentals depend on environmental factors. We believe that this is the reason bike-sharing systems are not likely to become a main form of transportation. For future directions, we would have liked to obtain larger and more current sets of data.

Tingyu Yang
Mako Gedi
Final Project
Data Science

References

DeMaio, Paul “The Need for Innovation” [Web log post]. (2016, February 14). Retrieved from http://bike-sharing.blogspot.com/

Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

Grus, J. (2015). *Data Science from Scratch:*. USA: O'Reilly.