Analysis of Washington, DC Bike Share Data from 2011 and 2012

Robert Brasso

Abstract— **Bike sharing has grown significantly over the last 10 years and most major cities around the world implement some form of bike sharing. Due to innovations in bike sharing technology we now have a plethora of transaction data at our fingertips. This paper seeks to find trends in bike sharing by analyzing the bike sharing data from Capital Bikeshare in Washington, DC between 2011 and 2012.**

I. Introduction

OVER the past several years, bike sharing programs have started to pop up in major cities all over the world and existing programs have been noticing tremendous growth. Bike sharing programs intend to solve the problem of traffic congestion in urban centers, while providing a healthier and environmentally friendly alternative to traditional means of commuting like car and buses.

Capital Bikeshare was launched in Washington, DC in 2010. Across the city, there are 440 stations with up to 3700 bikes. In order to rent a bike, patrons go to one of these 440 locations, either sign up for a membership or purchase a one-time pass, ride the bike, and return it to any of the other available stations when they have finished their session with the bike. Because of the transactional nature of renting the bike, data can be collected about the specifics of the bike rentals. We can use this data to gain a better understanding of how a bikeshare program works in a major city.

The purpose of this paper is to perform data preprocessing, exploratory statistics, and create a regression model to answer the central question of whether or not weather conditions are a reliable predictor of the amount of bike rentals for a bike share program in a major US city.

II. Dataset Description

The dataset analyzed in this paper was taken from the UCI machine learning repository and contains data from a bike share program in Washington, DC from the beginning of 2011 to the end of 2012. The dataset can be found at the following link: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.

The dataset is broken up into two similar, but different tables. The first represents only the data for patrons that rented bikes for an entire day and the other data set represents patrons that rented bikes on an hourly basis. The 'days' dataset contains 731 records, while the 'hours' dataset contains 17379 records.

The categories for both datasets are the same, except there is no 'hr' category in the days dataset. The attribute names and descriptions are listed below:

- **Instant** – record index.
- **Dteday** – Date.
- **Season** – '1' for winter, '2' for spring, '3' for summer, and '4' for fall.
- **Year** – '0' represents 2011, '1' represents 2012.
- **Mnth** – Months represented from 1 – 12, with January equaling '1' and December equaling '12'.
- **Hr** – hours from 0-23 with '0' representing 12am and '23' representing 11pm.\
- **Holiday** – Whether the day of instance was a holiday or not.
- **Weekday** – The day of the week from 0-6 with '0' representing Sunday and '6' representing Saturday.
- **Workingday** – If the day is neither a weekend or a holiday it is represented as '1', otherwise it is '0'.
- **Weathersit** – a coded scale of the weather conditions on the day of instance based on a cross reference of weather information from http://www.freemeteo.com.
    - '1' represents ideal conditions. Clear, few clouds, or partly cloudy.
    - '2' represents mist and cloudy, mist and broken clouds, mist and few clouds, or just mist.
    - '3' represents light snow, light rain and thunderstorms, or light rain and scattered clouds.
    - '4' represents worst conditions. Heavy rain, ice pallets, thunderstorm and mist, or snow and fog.
- **Temp** – represents the temperature in Celsius.
- **Atemp** – represents the feel temperature in Celsius.
- **Hum** – represents humidity.
- **Windspeed** – represents windspeed.
- **Casual** – The total amount of casual (non-registered) users for a given instance.
- **Registered** – The total amount of registered users for a given instance.
- **Cnt** – Sum of Casual + Registered

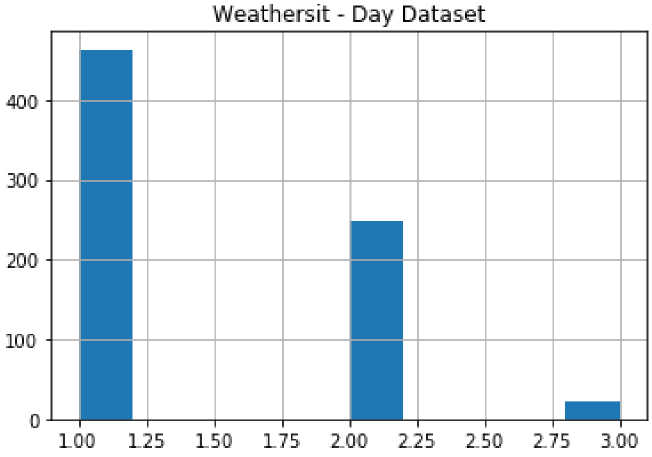
III. Exploratory Data Analysis

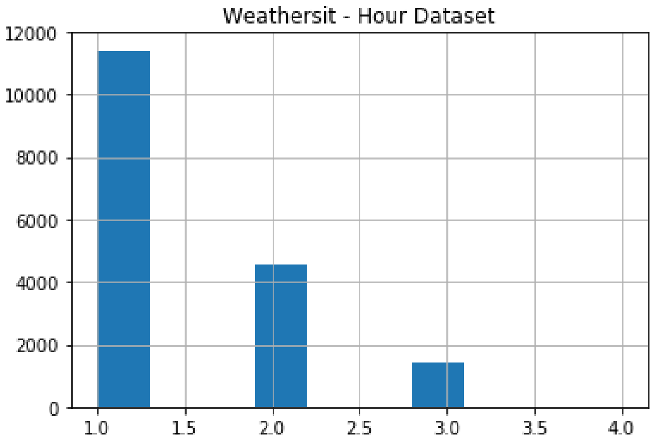
During the process of running descriptive statistics on the dataset I primarily used histograms and scatterplots. After creating histograms on all of the attributes for both of the datasets, I observed there was very little difference between the day and hour datasets. The majority of results held true between both datasets.

Of all of the attributes, there appeared to be no significant difference between the days of the week for bike rentals. Each day of the week had almost an identical number of rentals in both of the datasets. In both datasets, about 18.83% of the patrons were casual users and 81.17% were registered users. Likewise, contradicting my own prejudice about the data, the season of the rental also does not appear to be a significant factor as the data was nearly evenly distributed amongst the seasons.

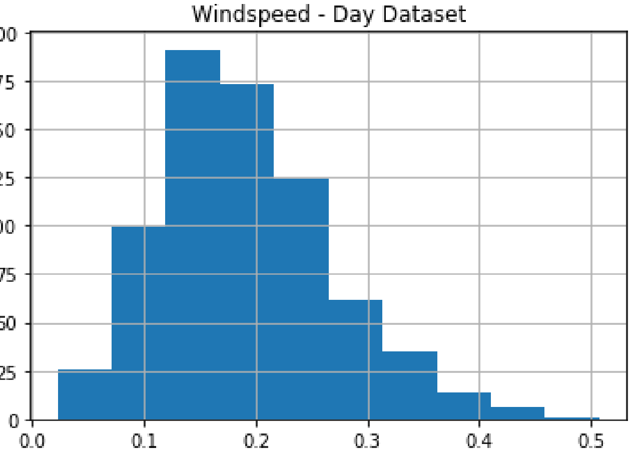
Temperature statistics showed that the highest frequency of rentals occurred while the temperature was between 8.2 and 32.8 degrees Celsius which roughly translates to 47 to 91 degrees Fahrenheit, which falls in line with the average temperatures for Washington, DC. Similarly, the highest frequency of rentals occurred when the humidity was between 40% and 90% which is also average for Washington, DC. Additionally, research is needed to determine if humidity and temperature play a role in the total rental counts.

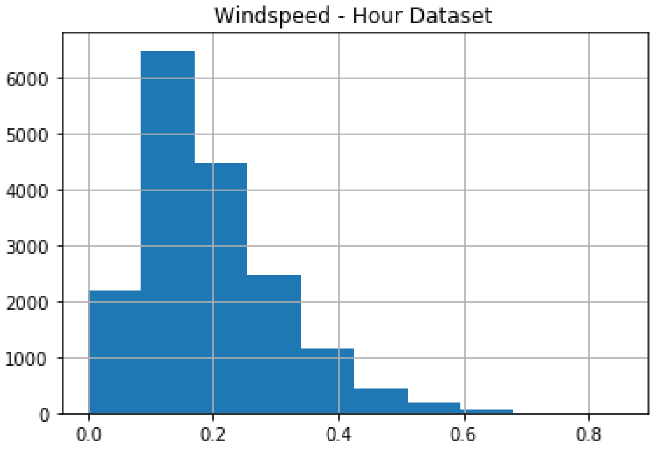
Two attributes revealed what appear to may be some correlative results. Weather situations appears to have a direct relationship the amount of bike rentals. The lower the number for weather situations (lower indicating better conditions), the higher the number of rentals. As the number increased, the number of rentals appeared to decrease.



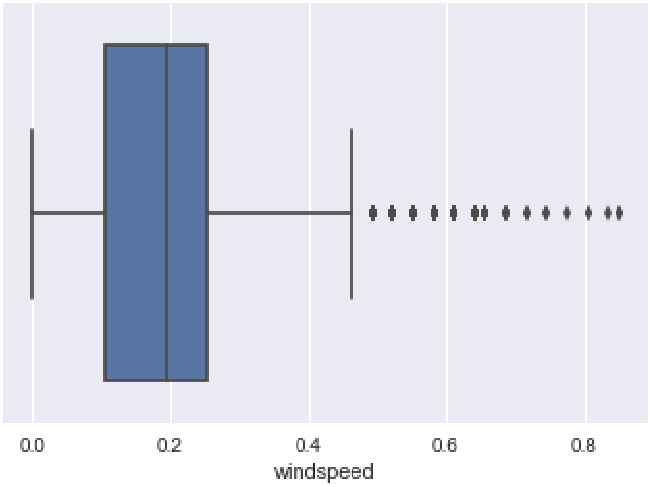


Similarly, it appeared that as windspeed increased beyond 13.4 mph (.2 *67), the amount of rentals sharply decreased.





The following boxplot also indicates that the highest frequency of instances occurred when the windspeeds were around 13.4mph.

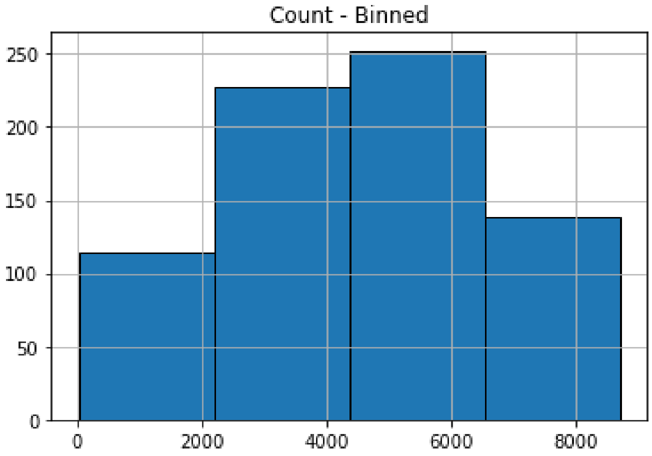


Since the highest frequency occurs during what appears to be the average windspeed for Washington,DC we do not need to investigate this attribute further. Due to the results of the exploratory data analysis, it appears the best attribute to investigate further are weathersit.

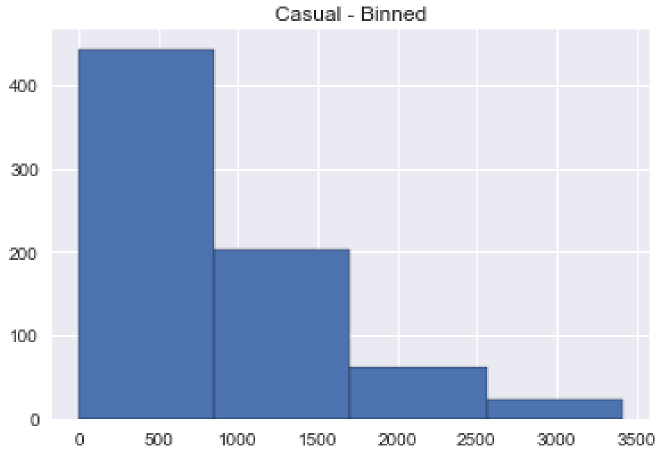
IV. Data Preprocessing

The dataset provided for this paper had already handled much of the data preprocessing. Prior to working with this dataset, I checked the entire dataset for missing and null values and there were none. Furthermore, certain attributes that would have needed normalization had already been normalized. Temp had been normalized on a scale from 0 to 1 and divided to a max of 41 degrees Celsius. Atemp was normalized and divided to a max of 50 degrees Celsius. Hum was normalized and divided to a max of 100 (representing 100% humidity). Lastly windspeed was normalized and divided to a max of 67 miles per hour.

The only continuous variables in the dataset that could benefit from binning were the casual, registered, and cnt attributes. From binning the cnt attribute we are able to determine the vast majority of instances had between 2000 and 6000 total users.



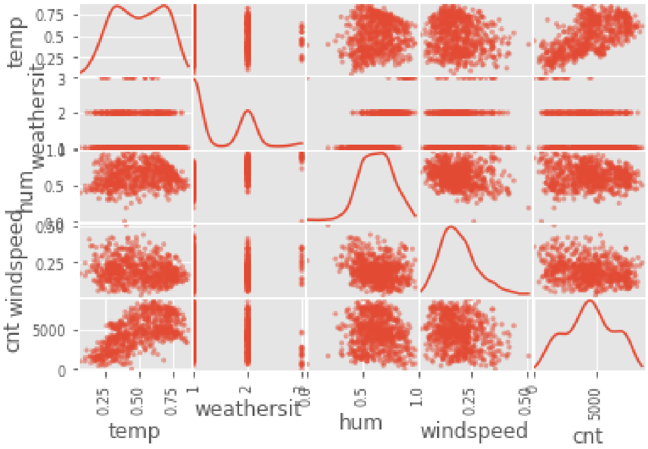
Additionally, using binning we can identify that nearly all of the instances had under 1000 casual users.



V. Regression Model

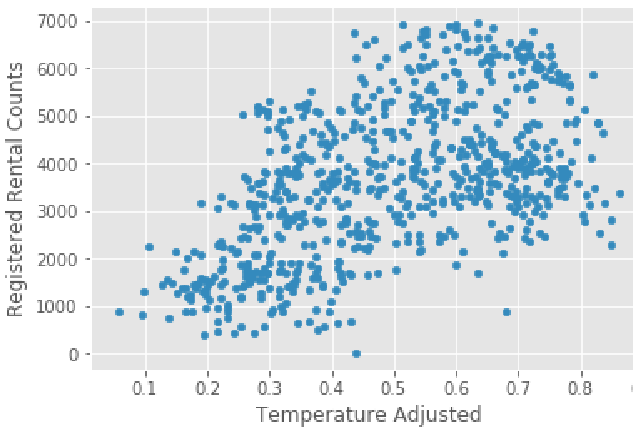
From the exploratory data analysis, it appeared the two categories of most interest were temperature, humidity, weathersit, and windspeed. Specifically, I wanted to analyze how these attributes effected the count of bike rentals.

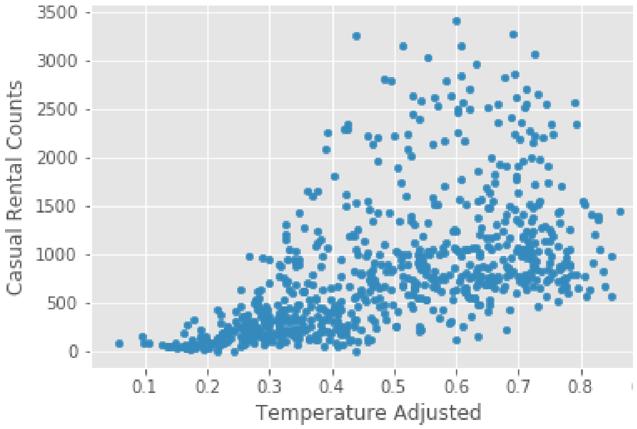
Before performing linear regression on these variables, I created a scatterplot matrix to ensure the variables were worth running further analysis on.



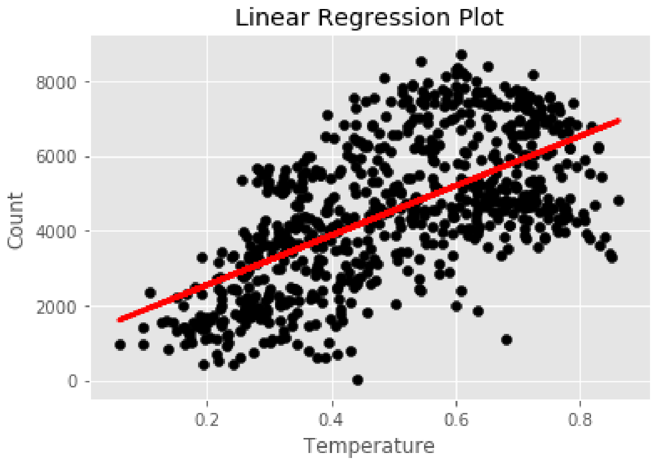
From this matrix, I determined that analyzing the relationship between temperature and count was worth further exploration. Additionally, I wanted more information on windspeed, humidity, and weathersit so I ran the Pearson Coefficient to see if they had any correlation. The correlation between humidity and cnt was weak negative (-.1006), windspeed and cnt was weak negative (2345), and lastly weathersit and cnt was weak negative (-.2974). After these results, I decided that none of these were likely substantial attributes to determine bike rental counts.

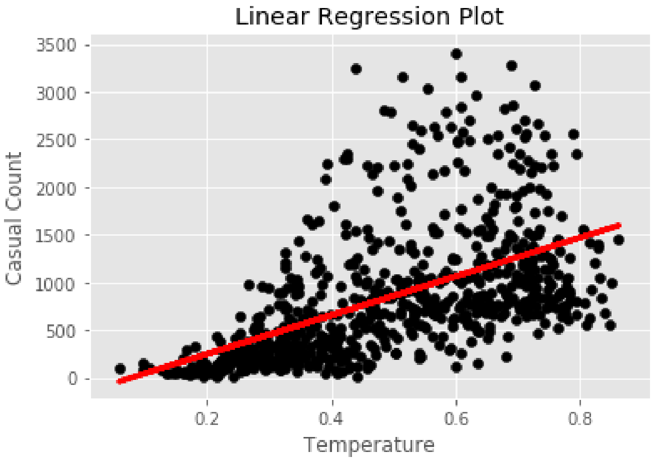
Continuing in my analysis of temperature to count, I decided to run scatterplots on both temperature to casual and temperature to registered.

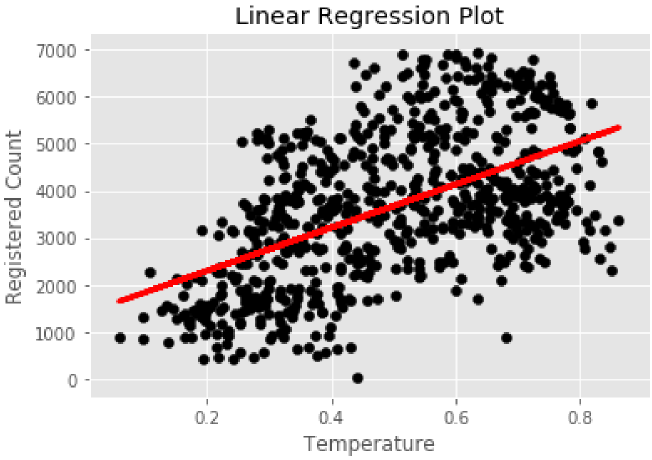




From these scatterplots, I ran my linear regression models and found lines of bet for each of the scatterplots. Below are the linear regression plots for temp to cnt, casual, and registered and the linear regression plot for windspeed to casual.







We can clearly see that there is a positive correlation between temperature and bike rental counts.

Running the Pearson coefficient matrix verified these results. The correlation between temperature and counts were positive and much stronger than all of the other attributes in the dataset. The Pearson scores between temp and casual were .5432, between temp and registered were .5400, and between temp and cnt were .6275.

Lastly, I ran the r-squared and p-value for temperature on count. The r-squared value is .6275, which indicates that the data somewhat fits the model. The p-value came back at nearly zero (2.81062239759e-81), which indicates that we must reject the null hypothesis and there is strong relationship between these two attributes.

VI. Conclusion

While the majority of statistics included in this dataset have little to no effect on bike rentals, there is a strong and clear correlation between temperature and bike rentals. As temperatures get hotter, people are much more likely to rent bikes from a bike share. While this seems fairly intuitive, it is important to point out that this data comes from the Washington DC bike share. It is possible that the warmer temperatures coincide with tourism season and perhaps tourism is another major factor in bike share rentals. If further research is done on this subject, I believe investigating that effect of tourism on bike share numbers would be important as well.

One potential application of this finding is that bike shares could try to fluctuate pricing models between summer and winter or they could try to increase access to bikes during summer months. Using this data, it might make sense to find some sort of temporary summer option for bike sharing such as partnering with hotels, coffee shops, and other local businesses. By using their space and having them manage the transaction, the bike share does not need to worry about leasing property and handling transactions. The money saved could be kicked back to the businesses who in turn benefit from having additional potential customers use their facilities and perhaps purchase their items or services.

References

[1] 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.

[2] https://www.capitalbikeshare.com/how-it-works