Capital Bikeshare Ride Prediction: Data Wrangling

Author: Matthew Cassi
Date: September 25, 2017

I. Project Description

The Capital Bikeshare Prediction project will involve predicting how long a ride will take between two Bikeshare locations in Washington D.C.. This analysis will take into account the day of the week, weather, and whether the ride falls on a holiday or not.

II. Datasets

The data for the prediction comes from two different sources: UCI Machine Learning Repository for the weather data and the Capital Bikeshare website for the time and start/end point data. The data will need to be joined together to perform the analysis needed for prediction. The UCI dataset contains dates and weather information (wind, temperature, humidity, number of bikes per day) for each day between January 2011 and December 2012. The bikeshare dataset contains start times, end time, start locations, end locations, and member type from the second quarter of 2011 through the first quarter of 2012. The first quarter of data for 2011 was not included because Capital Bikeshare was fairly new and the number of rides were lower compared to the other quarters. The first quarter of 2012 provided a more accurate picture of the rides from January through March.

The UCI data can be found here: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset (http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset)
The Bikeshare data can be found here: https://s3.amazonaws.com/capitalbikeshare-data/index.html (https://s3.amazonaws.com/capitalbikeshare-data/index.html)

III. Loading and Cleaning the Datasets

All five datasets contain date columns that needed to be loaded in as datetime objects to make it easier for filtering and merging of the datasets. To handle the date coolumns, the partse_dates attributes were used on in the read_csv functions. This created all date columns as datetime objects in Python.

When inspecting the first few rows of each dataframe, each datasets have columns that are not necessary for the analysis and prediction aspects of the project. The UCI weather dataset contains extra columns like month, year, and instant (an index column). The bikeshare dataset contains additional columns like duration (text field) and bike serial number. All of these columns were removed as they would not impact the analysis to be performed. In addition to removing columns, some columns had to be renamed so all columns match, which is needed for the the joining of the datasets.

There were not many missing values in each dataset. The bikeshare data contained 11 rows that did not have a final end point. These rows were removed from the bikeshare dataset. The UCI data did not have any null or missing values.

One of the Bikeshare datasets contained negative values for the time difference column. After doing some research, this occurred because of daylight savings time. 60 minutes were added to the negative values to make adjust for the 1 hour difference.

Prior to joining the datasets together, the date columns from all datasets needed to be in the same format. The UCI weather dataset date format is YYYY-MM-DD, while the Bikeshare datasets used a YYYY-MM-DD HH:MM:SS format. Two additional columns were created in the Bikeshare datasets with the YYYY-MM-DD

format. The original columns with the hours, minutes, and seconds were kept. In addition to creating those two columns, another columns was created for the time difference, in minutes, between the start and end times of the bikeshare dataset.

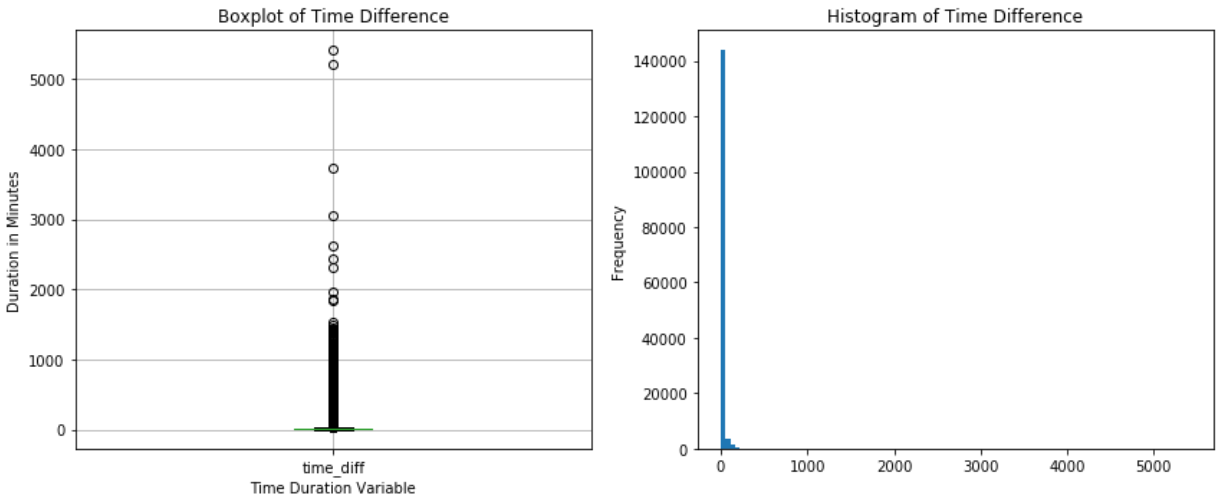
IV. Joining and Concatenating the Datasets

Each dataset needed to be joined to have a complete dataset needed for prediction purposes. The datasets were left joined with each bikeshare dataset as the left dataset and the UCI data as the right dataset. This was done so that all of the values from the bikeshare dataset remained in the dataset and only the dates that matched from the UCI dataset were included.

Once the datasets were all joined together, they needed to be concatenated so all the quarterly data would be combined.

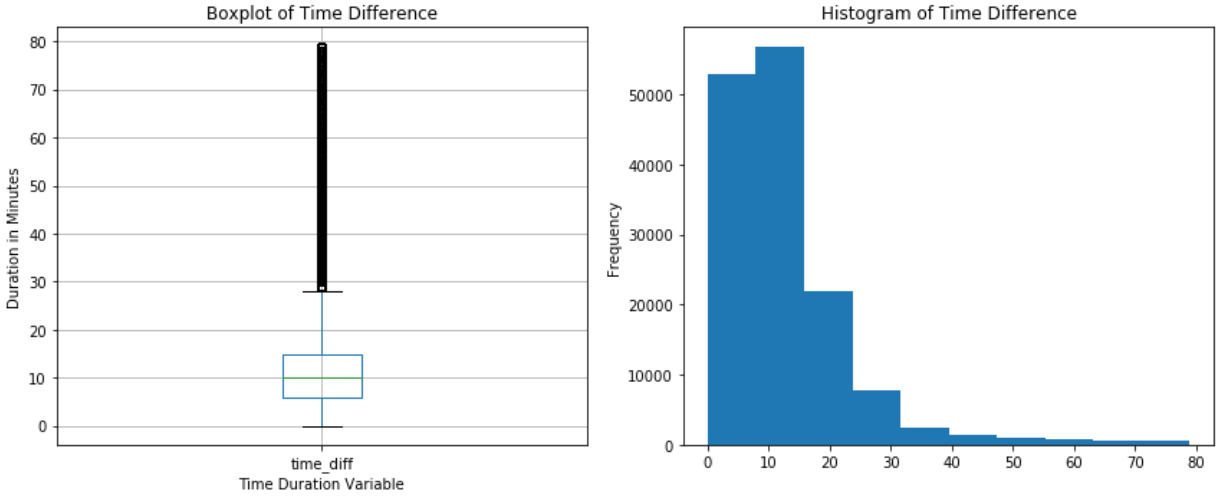
V. Outliers

After cleaning and joining the datasets together, the data was inspected for outliers. The main column to check was the duration column of the data. The duration column contained many values that were very large with the max being well over 5000 minutes.



The box plot above shows the time difference column and the shape of the distribution of the column is difficult to discern. The shape of the distribution cannot be determined because the values of the time difference are so spread out. The box of the box plot cannot be displayed because of these outliers. These outliers would have to be removed in order to show the distribution of the data.

An arbitrary value of 80 minutes was selected for the cutoff of outliers. While 80 minutes (also, anything greater than 32 minutes) falls out of the Interquartile Range, I felt that was a good value for the a bike ride in the Washington D.C. area based on traffic, weather, and distance between some of the stations.



After removing the outliers of this dataset, the box of the box plot became visible and the distribution of the time difference looked like it was right skewed.