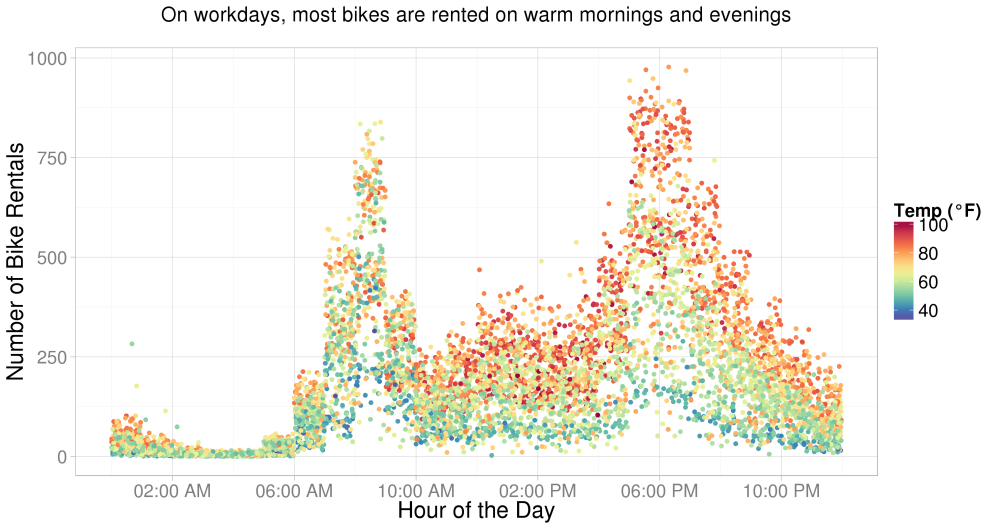
Benjamin Chen
a take-home interview exercise from DataRobot for Data Science intern

The Premise

http://www.kaggle.com/c/bike-sharing-demand



"Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

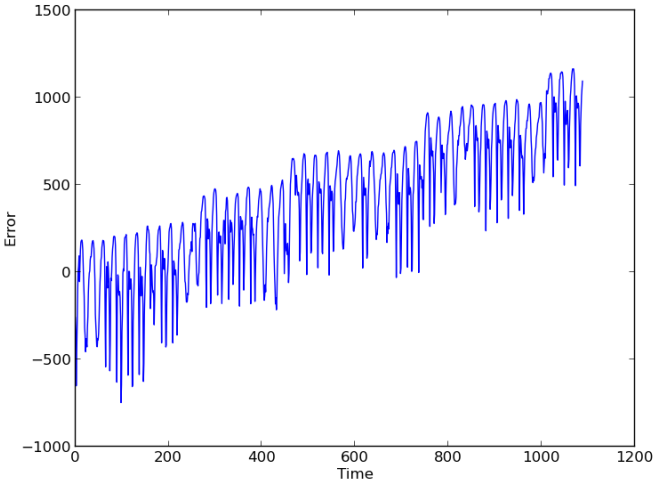
The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C."

The Raw DataSet

The data columns were as follows: datetime, season, holiday, weekday or not, weather, temperature, "feels like" temperature, humidity, and windspeed, number of non-registered user rentals initiated, number of registered user rentals initiated and the number of total rentals.

Data Transformation

The data is split for analysis into a training set of the first 90% of the data, instead of randomly because the application is forecasting. Conforming to Google Prediction API input guidelines, the columns that are not the output nor in the training set (number of non-registered user rentals initiated and number of registered user rentals initiated) are removed.



Prediction Results

After going through the Google Prediction API, the results are in results.csv, the results are as shown above. As expected, the error grows as we drift from the available data. It seems as if there is a cyclic component to the residuals that is not captured.

Complications in the Process

When going about this task, one of the big complications has been bumping up to the limitations of being an API: rate limiting, and the lack of transparency. Much time was spent figuring out how the undocumented pieces worked. Working with an API also creates somewhat of a different workflow and process, which de-emphasizes results because I have no control over the analytics process. Looking back, I should have prioritized analysis of results more than the results themselves.

Further Work

Any further progress would be made on residual analysis with visualizations to fine-tune my data transformation process, and refining the code to be easier to use and configure.