capital bikeshare

Predicting daily patterns in rentals

1 Problem

UNDERSTAND PATTERNS OF BIKESHARE USE IN WASHINGTON, D.C.

To better understand use patterns in a bikeshare fleet, the goal is to describe frequency and duration of rentals for casual and registered users, define high and low demand times, and develop a model to predict rental frequency for a given day in a year.

2 Client

The client is Capital Bikeshare, and results from this analysis will allow a clearer understanding of how users interact with the bikeshare program. This information could be helpful for growing the business, for example, targeting advertising dollars on days or seasons with a high proportion of unregistered users might be the most efficient way to increase registration. Being able to predict high demand time will also assist with gauging appropriate number of available bicycles system wide.

3 Data Acquisition

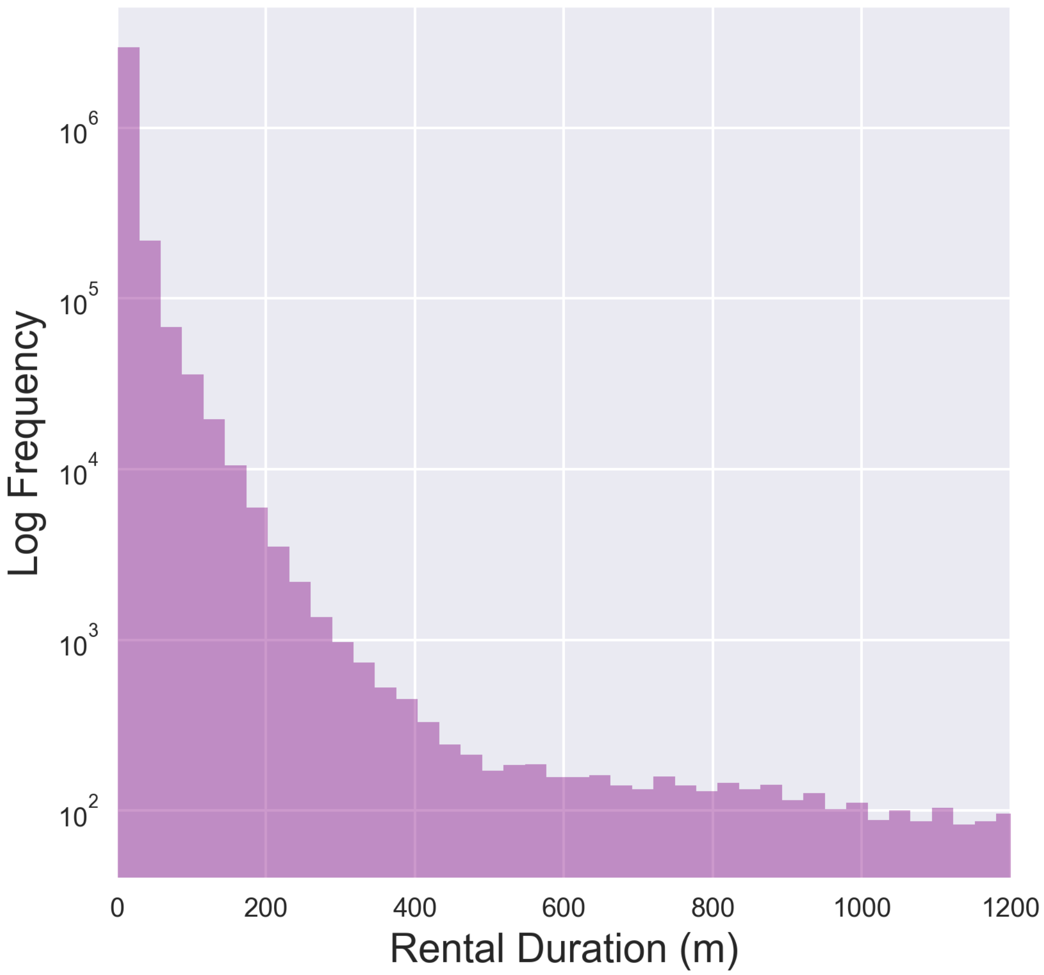
Capital Bikeshare posts quarterly data reports of bike trip times, start and end locations, and type of user (registered or casual). Each trip is on one line of data, and includes start time, end time, duration, start station ID and address, end station ID and address, and user type (registered or casual). These data are readily and publicly available at https://www.capitalbikeshare.com/system-data.

These data were provided in a very manageable format. When importing, I created a time index based on the start time of each rental. These data were also combined with weather data (daily maximum and minimum temperature) from a weather station in the National Arboretum in Washington, D.C.

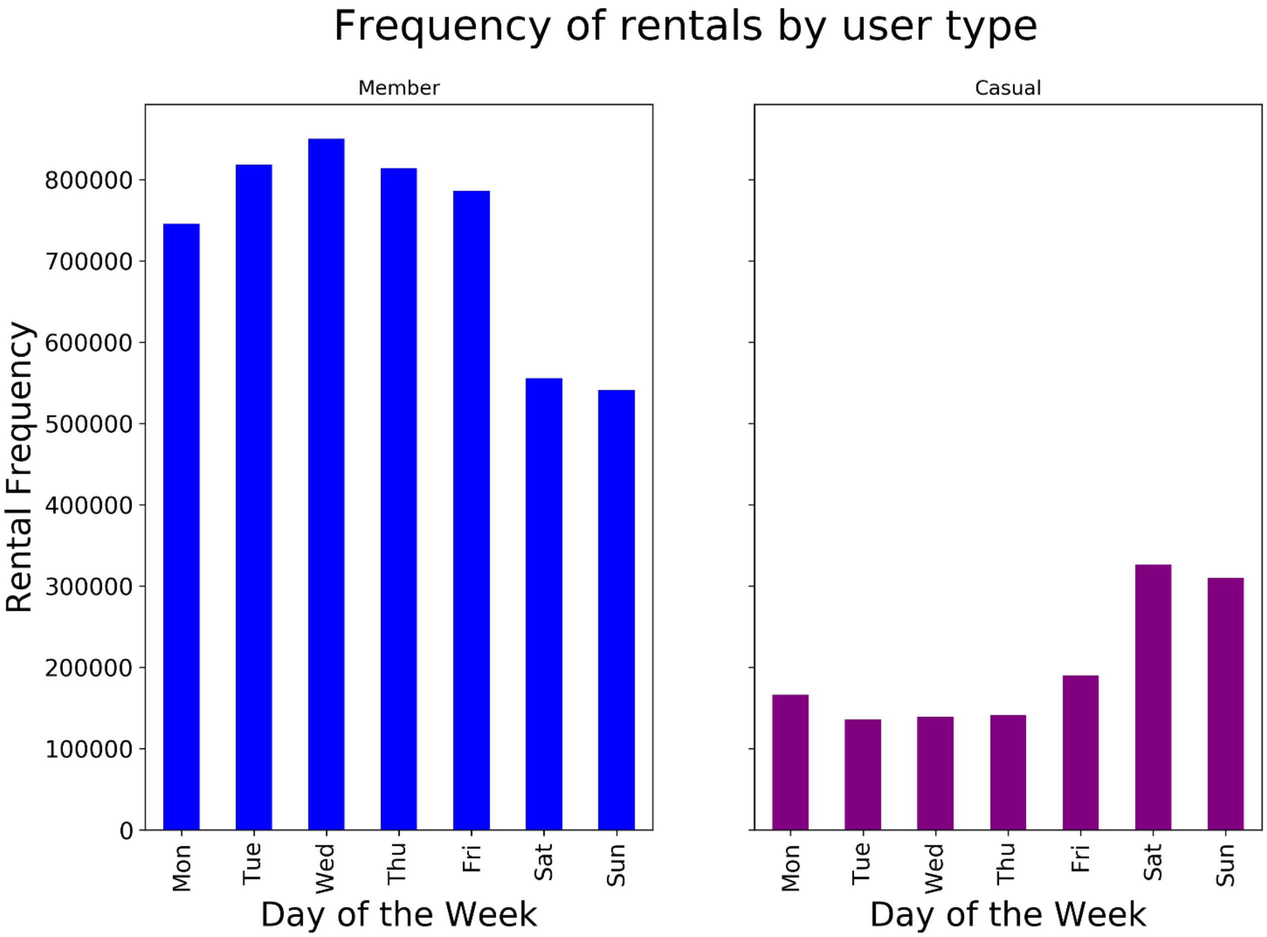
To explore the data and build the predictive model, I used data from 2015 and 2016. The model was tested on data from 2017.

4 Describing patterns in the data (EDA)

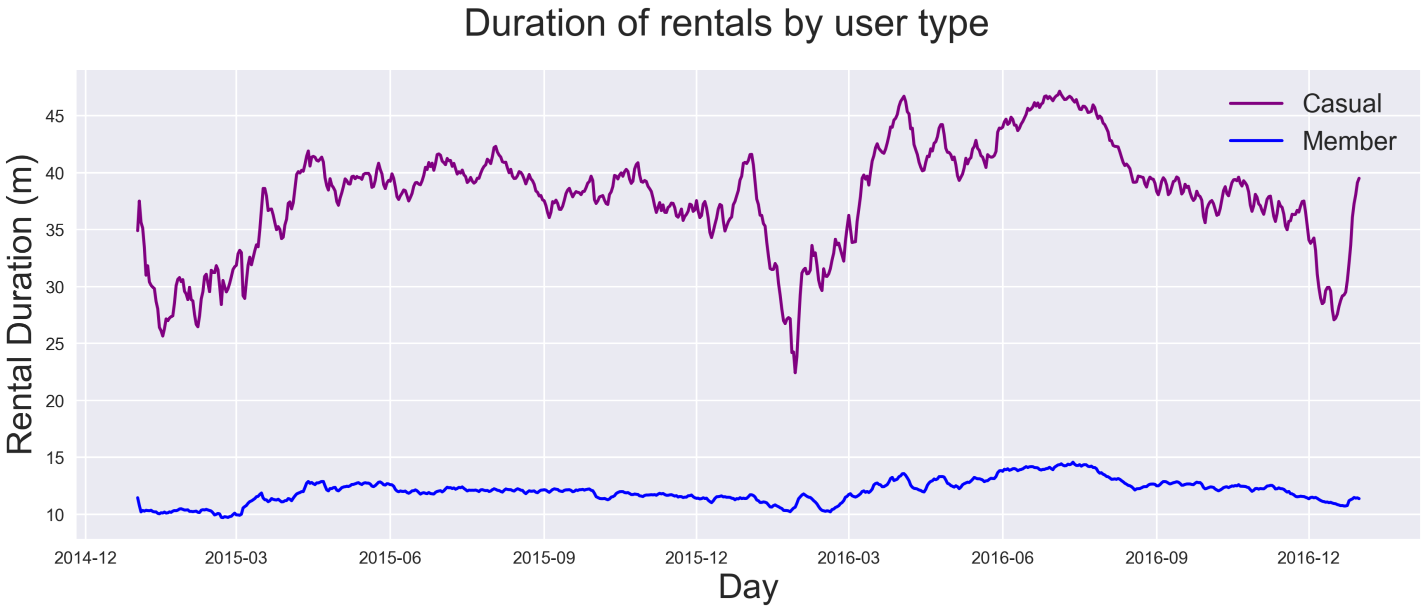
One interesting characteristic of the test data is that the duration of rentals has a non-normal distribution. Although most rentals range from approximately 5-25 minutes, there is a long tail of much longer durations.



However, it seems that whether a user is registered can have a big influence on usage behavior. For example, whether rentals occur on the weekend or on weekdays:



Or how long rentals last (duration):



My initial hypotheses for these differences based on user type is that users of each category have predictable characteristics. For example, residents may tend to register for the program, bike shorter distances, and rent for shorter times. We could speculate that trip purpose might be personal errands, lunch on a workday, or transport to work meetings. Non-residents (i.e. tourists), on the other hand, may rent a bike to facilitate longer distance sight-seeing agendas, or even circular trips (no destination). Perhaps even unregistered locals use the bikeshare program only on weekends for purposes like those postulated for tourists.

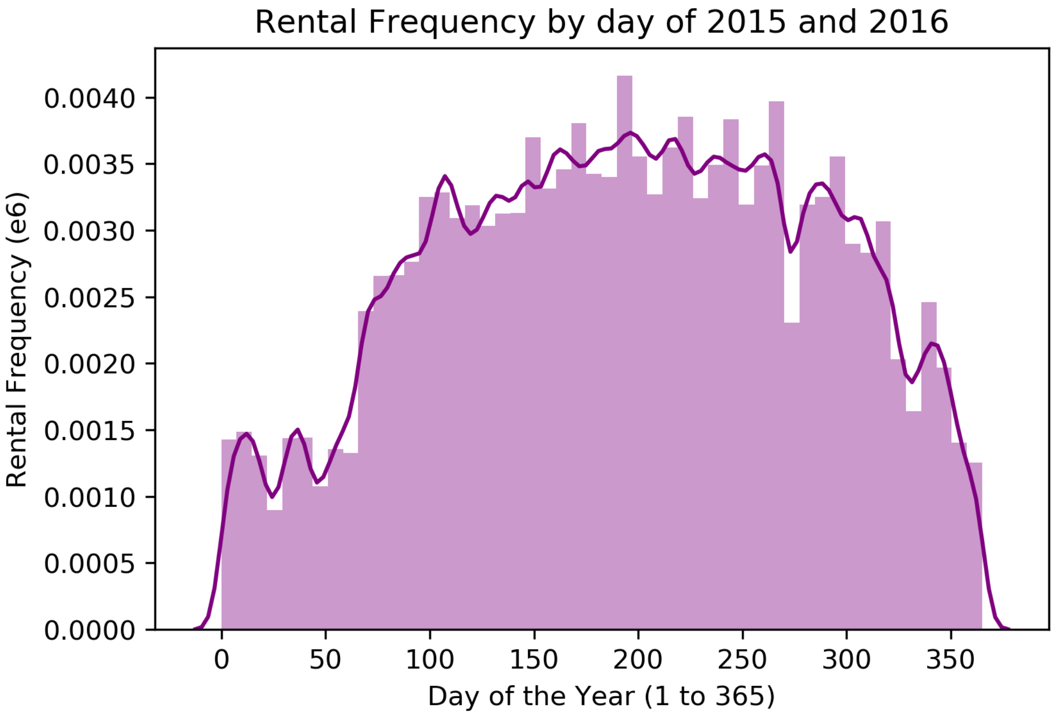
5 Model Development

My goal was to predict demand (number of rentals) on a given day (e.g. in 2017). To do this, I tested two subsets of models, a linear model with varying numbers of features and a regularized linear model (Ridge regression).

5.1 Features

- Time (Days in the year, 1-365)
- $(\text{Time})^{2}$
- Daily maximum temperature
- Daily minimum temperature
- $(\text{Daily maximum temperature})^{2}$
- $(\text{Daily minimum temperature})^{2}$
- Weekday dummy variables ('day_0', … 'day_5': 0 or 1)
- Holiday indicator variable (0 or 1)

These features were designed to capture the both the progression of time and the cyclical pattern in time. This histogram shows that bike rentals do seem to follow a parabola-like pattern through a given year, with the greatest number of rentals from ~March through November. Temperature is also expected to follow a similar pattern, so I included both minimum and maximum daily temperature, in addition to the square of those values. Finally, to incorporate social patterns in bike rentals, I designed dummy variables to represent each weekday and holidays.



Temperatures in DC (at the National Arboretum) ranged from 11-99°F in 2015 and 2016. The average high temperature was 69°F and average low temperature was 50°F. Average daily precipitation was 0.11", with 2.85" in one day being the maximum for the 2015-2016 period.

5.2 Model development process

I used older data to train each model (2015-2016) and the most recent entire year of available data to test accuracy (2017). Models 1-3 used a linear model (sklearn Linear Regression) and fit an intercept.

Model 1: time and time^2

> This model including only the day of the year and the day of the year squared explains 33% of the variation in daily rentals.

Model 2: Time, time^2, daily max temp, and daily min temp, and min/max temp squared (*6 features*)

> This model also includes daily maximum and minimum temperature, and seems to only explain 42% of the variance.

Model 3: Time, time^2, daily max temp, daily min temp, square of daily min/max temp, day of the week (7 dummy variables), holiday indicator (*14 features*)

> This final model that includes the day of the week and holiday indicator seems to be the worst predictor of daily totals only explaining 42% of the variance.

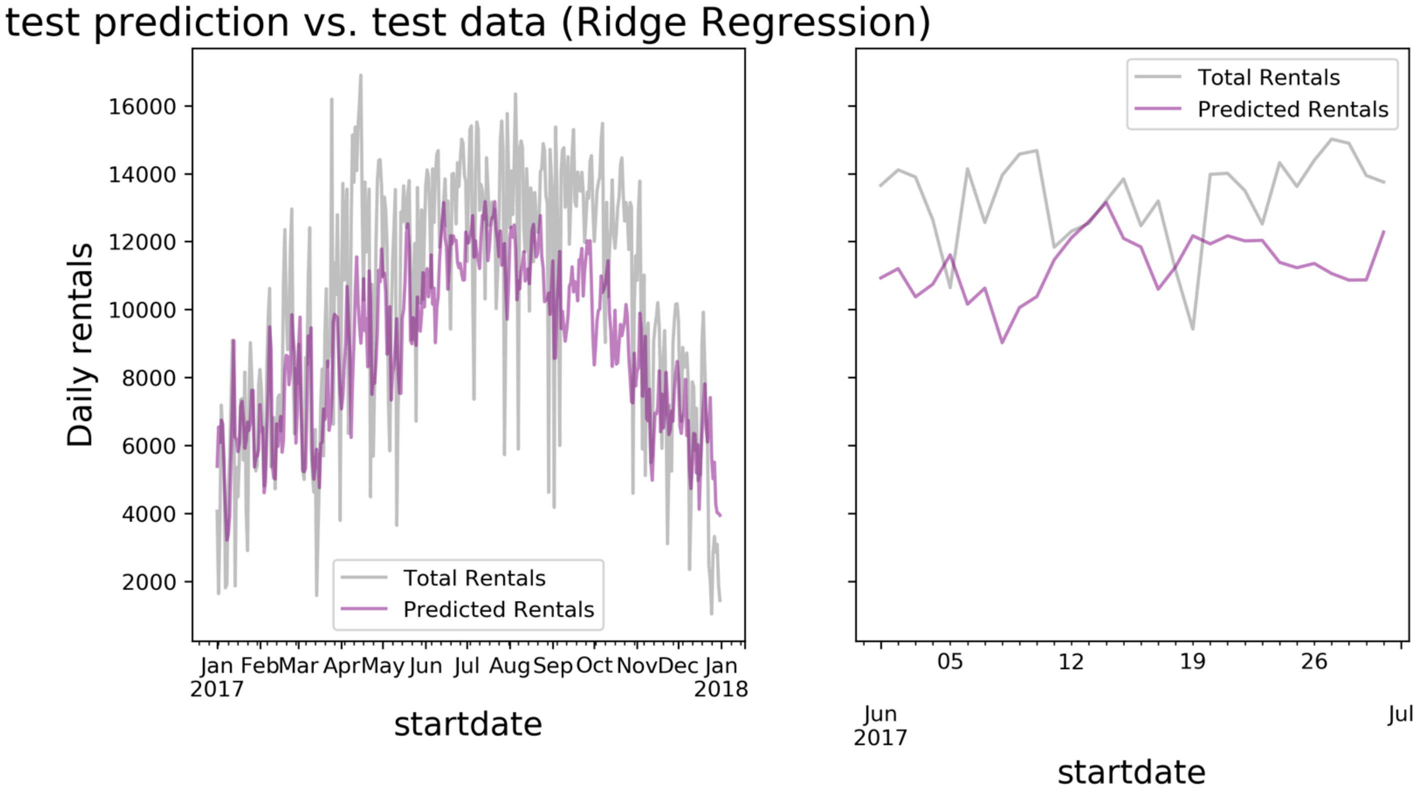
Model 4: Model 3 features, Ridge Regression – **BEST MODEL**

> Because the number of features has grown to 14 in the third model, using a regularization is appropriate to penalize overfitted models. Using RidgeCV from sklearn.linear_model, I fit a ridge regression to the 2015/2016 training data using 5-fold cross validation to evaluate the parameter α (0.1). The R^2 for the ridge regression model with all features is 0.59, or 59% of the variation, which is a substantial improvement over the R^2 for the full model without the regularization component.

6 Results

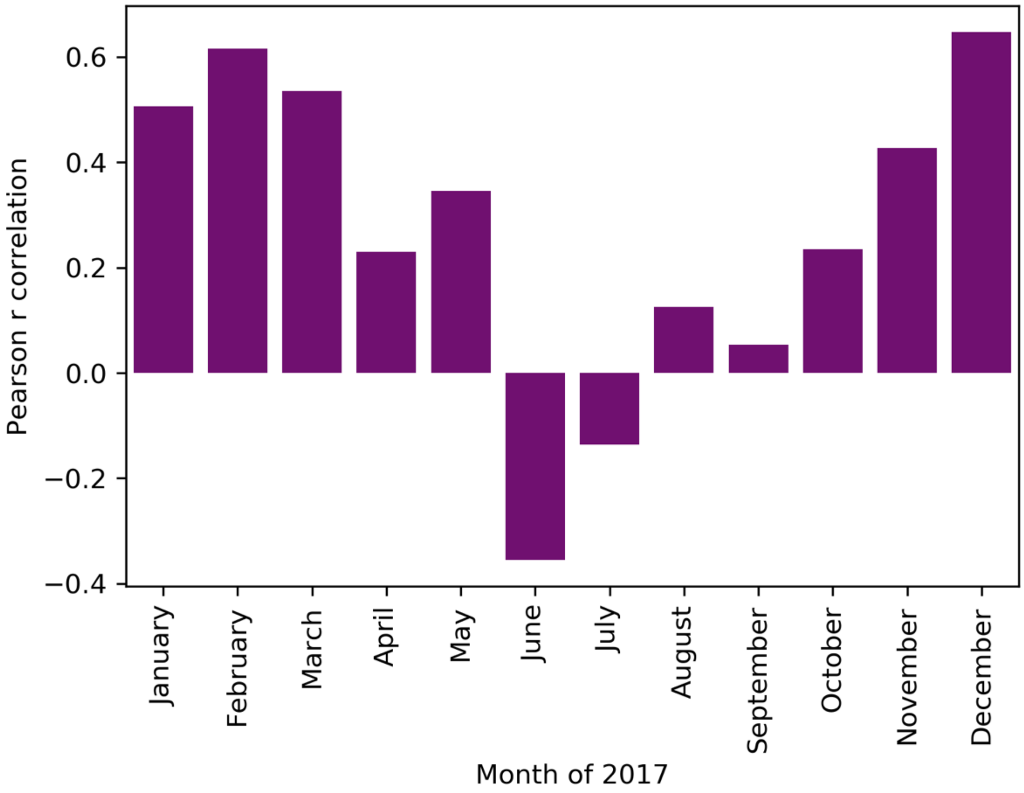
How well does the Ridge Regression model predict the test data?

6.1 Time



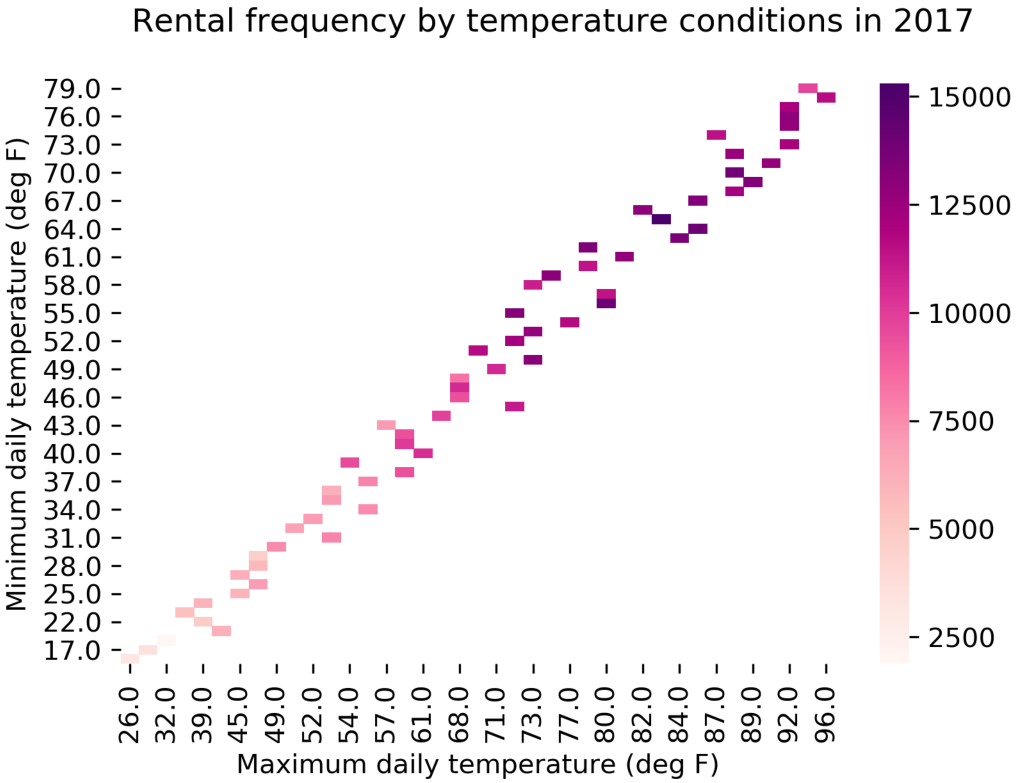
This figure shows the model-predicted rentals for each day in 2017 (purple) over the actual rentals in 2017 (grey). A quick visual assessment leads to the conclusion that the model performs well at capturing the rise and fall of rentals through the year. The right-hand panel shows a magnification of just the month of June 2017, where you can see daily fluctuations are only mildly mirrored by the model.

There is some noticeable departure from the actual data in late summer in early fall, when the model is mostly under-predicting demand. This next figure shows a breakdown by month in 2017 of how well the model correlates with the actual data, which shows very low or negative correlation June-September.



6.2 Temperature

It may also be useful to use a weather forecast to improve predictions on a short timescale. There is a clear indication of lower rentals in cooler weather (cool max and min temperature, lower left quadrant):



7 Conclusions and Next Steps

Overall, the chosen best model did reasonable job of predicting demand on the Capital Bikeshare system. Some improvements could be made in looking for factors that might explain the departure of the model in June-September. For example, perhaps rentals are increasing year over year, and this model does not take that aspect of growth into account. Another aspect to consider might be other aspects of seasonality associated with being the national capital, such as congress recess or large events like inauguration.

8 Addendum

Please find relevant code notebooks at https://github.com/mskaerthomason/Springboard-Capstone-1.