Capstone 1 Milestone Report - Analyzing NYC Citibike Usage

Andy Pickering

July 21, 2017

Contents

1 Introduction

For my first capstone project, I chose to analyze data from the NYC Citibike bike-rental system. The problem is to determine which factors impact usage, and develop a model to predict the number of rides taken each day on the Citibike system. The main client is the Citibike management, who need to plan for anticipated demand and ensure that there are enough working bikes and staff on hand to meet the demand. Other clients could include city transportation departments, since bike usage likely affects ridership on other modes of transportation, as well as possibly introducing traffic issues. A more ambitious goal is to predict bike demand and availability at specific stations, to determine the most efficient distribution of bikes.

2 Data

2.1 Historical Data

The main dataset is historical records of trips taken on the Citibike system. These are made available by the Citibike program at `https://www.citibikenyc.com/system-data` in monthly files, spanning the time range July 2013 to March 2017. Each row in the dataset corresponds to one ride, and includes the following fields:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

Upon examining the rider age distribution, there appear to be a small number of incorrect values or typos. A very small fraction of riders have ages over 100, up to 159!. Since the oldest living person is 117 (`https://en.wikipedia.org/wiki/List_of_the_verified_oldest_people`) it seems likely that these are mistakes in the data. Birth year, not age, was recorded, so my best guess is that 18 was intered instead of 19 for the beginning of some years (ie 1895 instead of 1995). These age values will be converted to NaN in the dataset.

2.2 Weather

Historical daily weather data was downloaded from the Weather Underground website (`www.wunderground.com`) with the script `get_weather_data.py` . I use data from LaGuardia airport. The data contains the following fields:

- Max Temperature (F)
- Min Temperature (F)
- Mean Temperature (F)
- Precipitation (inches)
- Events (rain,snow,fog etc.)
- Cloud cover
- Max wind gust (mph)

2.3 Holidays

A list of US holidays was obtained through the holiday API. `https://holidayapi.com/`. This contains the name of each holiday and the date observed.

2.4 Limitations

Historical citibike data does not include any user ID, so I cannot determine the number of unique users or individual patterns.

Historical data does not include station status (number of bikes available etc). So I cannot predict station availability from this data. I began collecting streaming data in order to address this question in the future.

3 Data cleaning and wrangling

The combined historical citibike data was very large: more than 39 million rows (rides). Therefore I decided to store the data in a sqlite3 database and then query the database from python for analysis. A python script (`read_citibike_to_sql.`) was used to perform some cleaning and modification of the data and write it to the database. Redundant station-related variables (name, lat, long) are not saved in the table with ride data, in order to reduce the size. A separate table 'stations' is made with the info for unique stations, which can be joined to the 'rides' table using the station ID key if needed. Cleaning/modifications performed on the historical data include:

- Replacing spaces in variable names w/ underscores
- Converting date fields to appropriate data types
- Add variables such as year,month, and day to the dataframe
- Converting variables to categorical data type

4 Analysis

4.1 Citibike Usage

I began with an exploratory analysis of the data and patterns of ridership. The dataset contains 39,148,013 rides!

Findings:

- Ridership has increased over the 3 years (Figure 1).
- Part of this increase is due to expanasions of the system, increasing the number of stations and bikes. (Figure 4). I computed the number of stations each month as the number of unique station IDs in all the rides during that time. Note this only includes stations that were actually used; I assume every available station was used at least once, although the actual number of stations could be slightly different. I think it is close enough to account for the majority of this effect though.
- There is a strong seasonal cycle during each year, with more rides in the summer and early fall. (Figure 2). My assumption is that a lot of this is due to the seasonal cycle in weather; this will be tested further.
- Daily cycle: During the week, the majority of rides take place during the morning and evening rush hours (Figure 5). On the weekend, rides are more distributed throughout the day (Figure 6).

- A quick map examination (not shown) of the most used stations showed that during the week, there are several stations close to mass transit such as Grand Central Stations. On weekends, those stations are not most used; and several of the most used stations are now located near Central Park. This adds further evidence to the pattern of commuting during the week and recreational riding on weekends.
- The avearage duration of rides is longer on the weekend. A t-test found this difference is statistically significant.

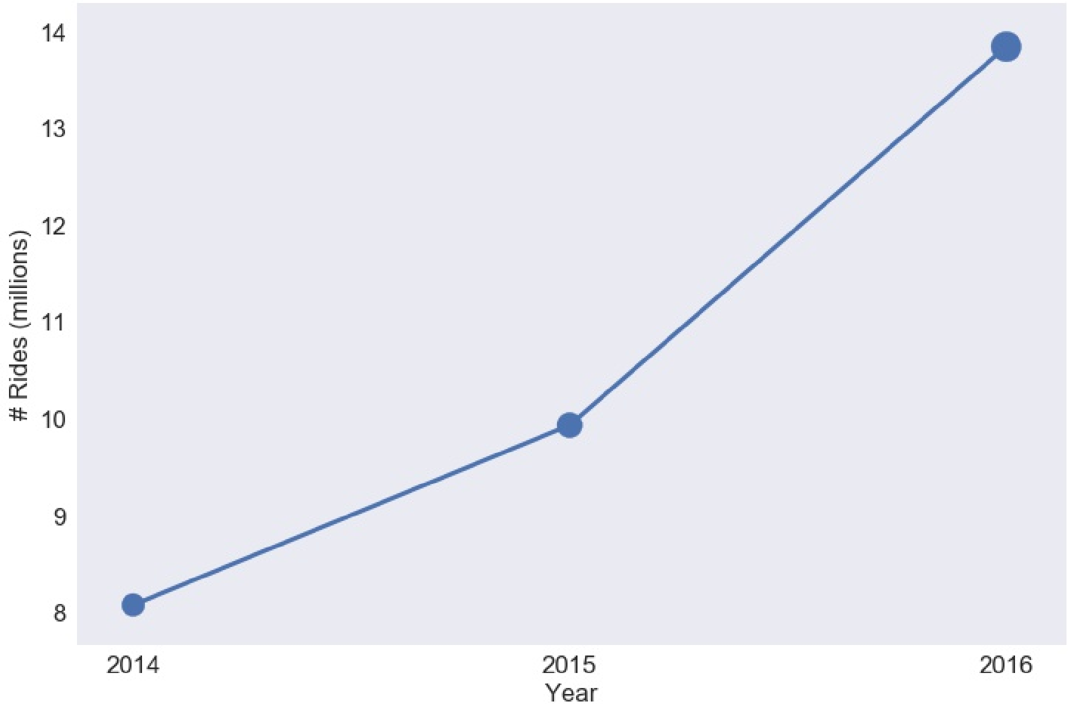

Figure 1: Total number of rides (in millions) taken each year. Only years for which complete data was available are plotted.

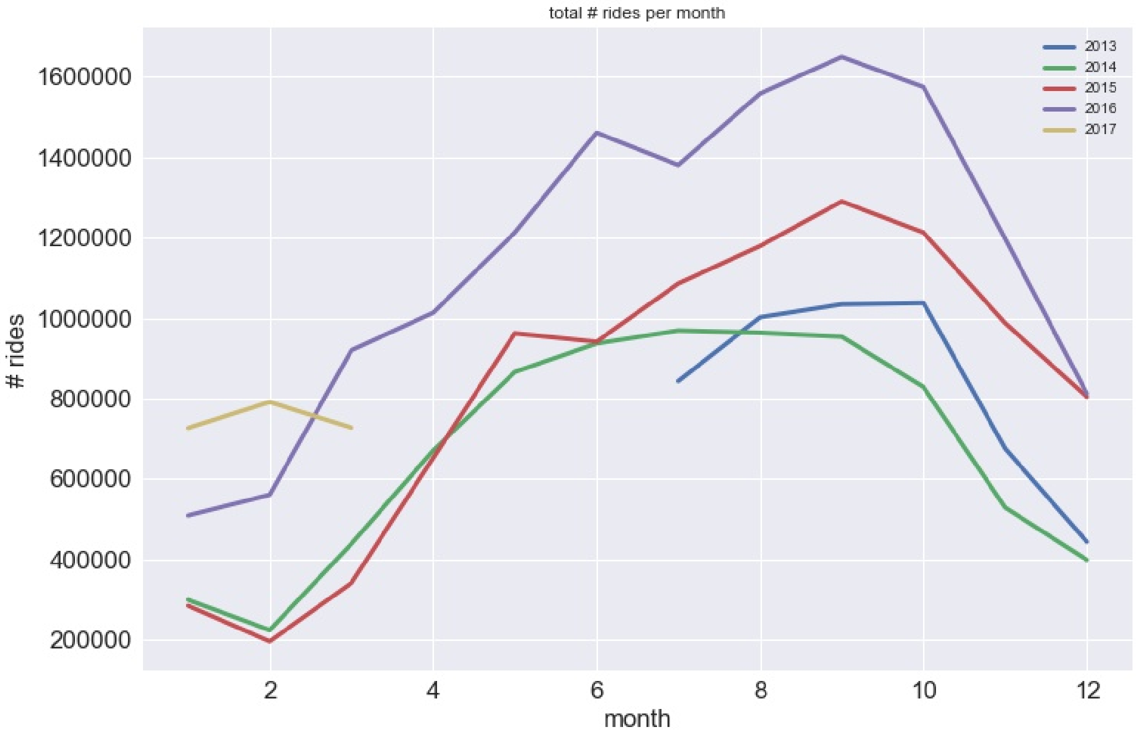

Figure 2: Total number of rides taken each month, for different years.

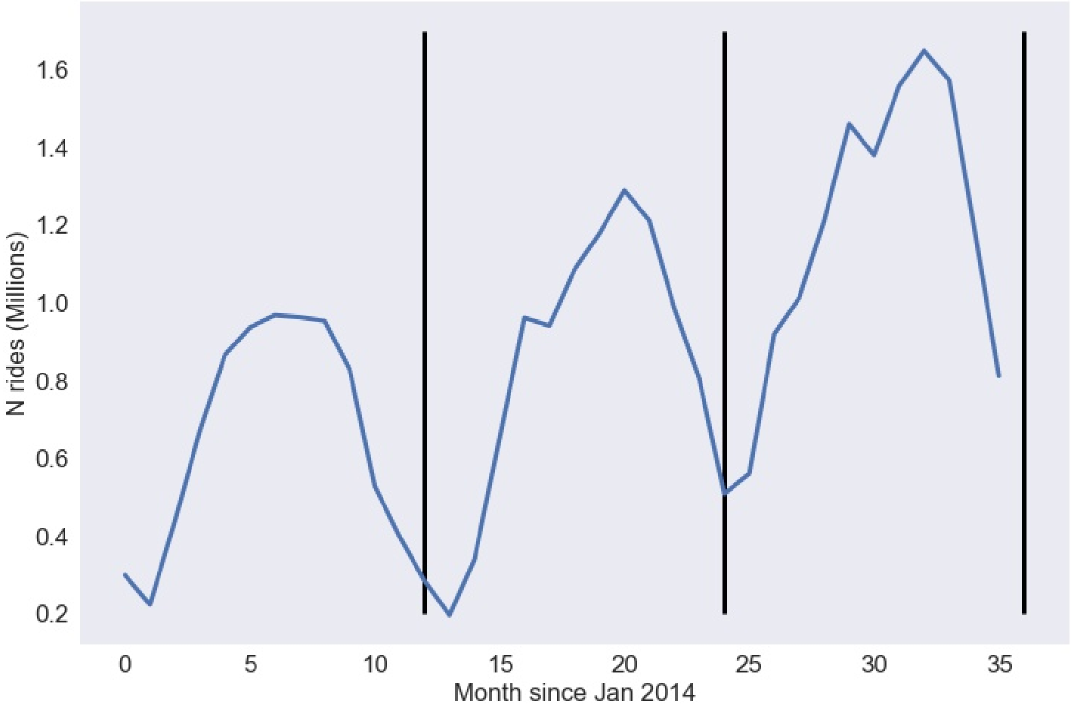

Figure 3: Total number of rides taken per month, plotted consecutively for 2014-2016.

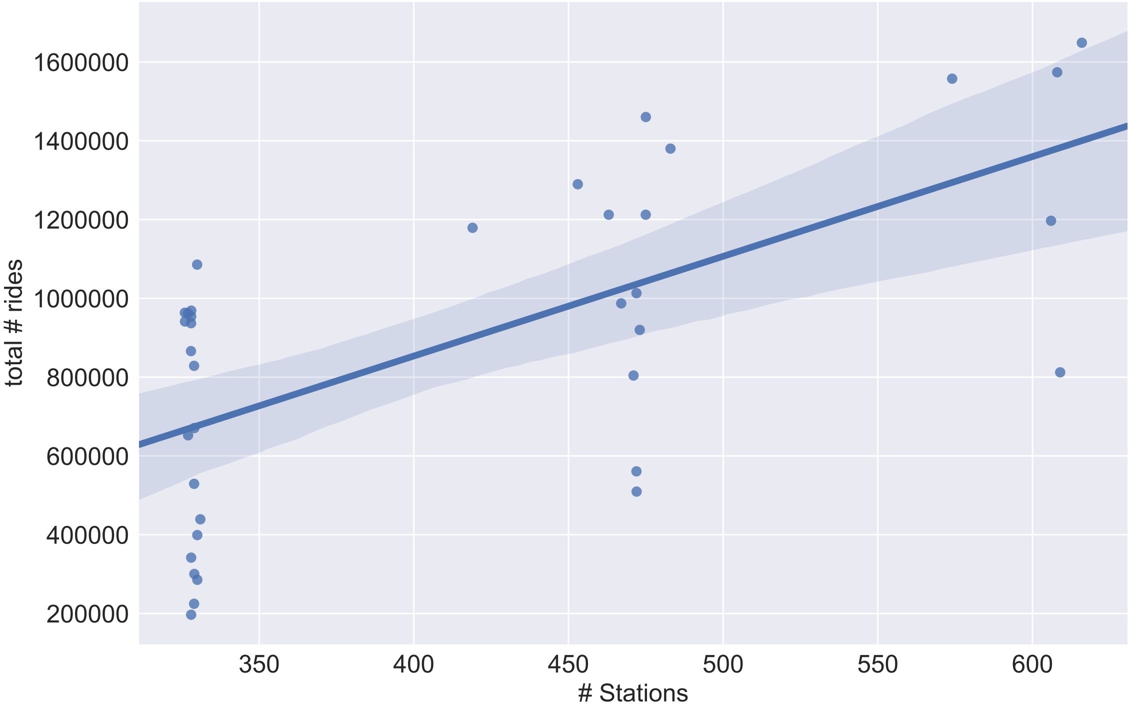

Figure 4: Total number of rides vs the number of stations.

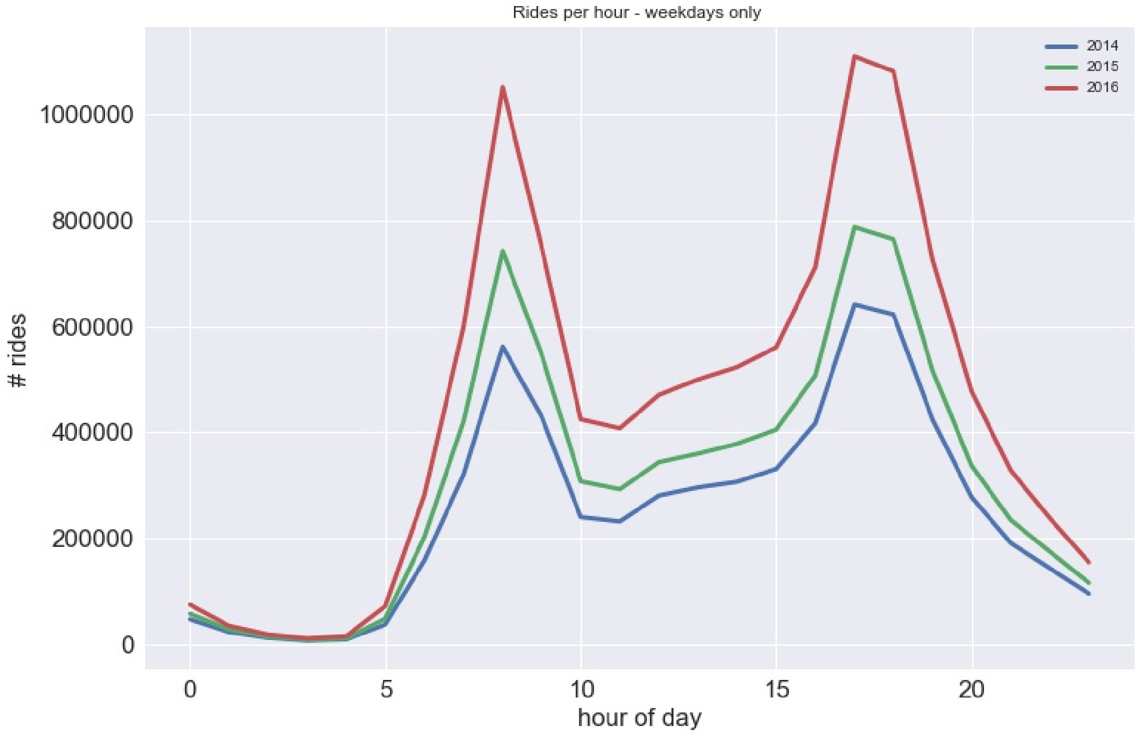

Figure 5: Total number of rides taken each hour of day, for weekdays (Monday-Friday).

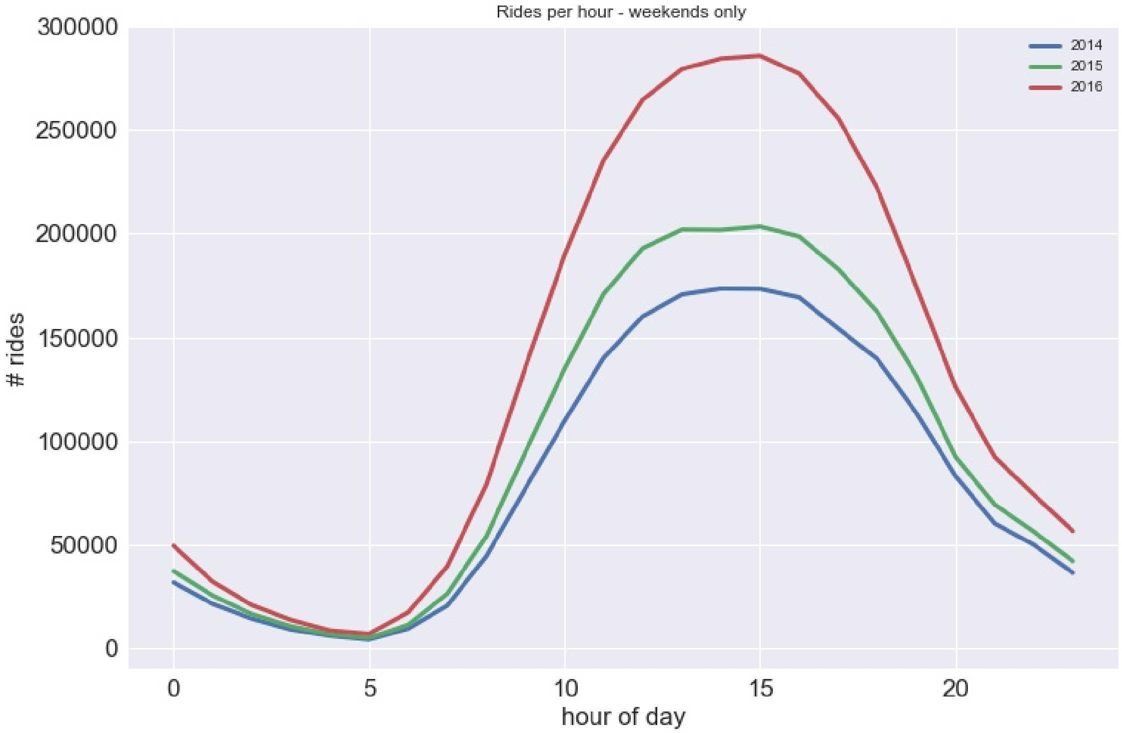

Figure 6: Total number of rides taken each hour of day, for weekend days (Saturday and Sunday).

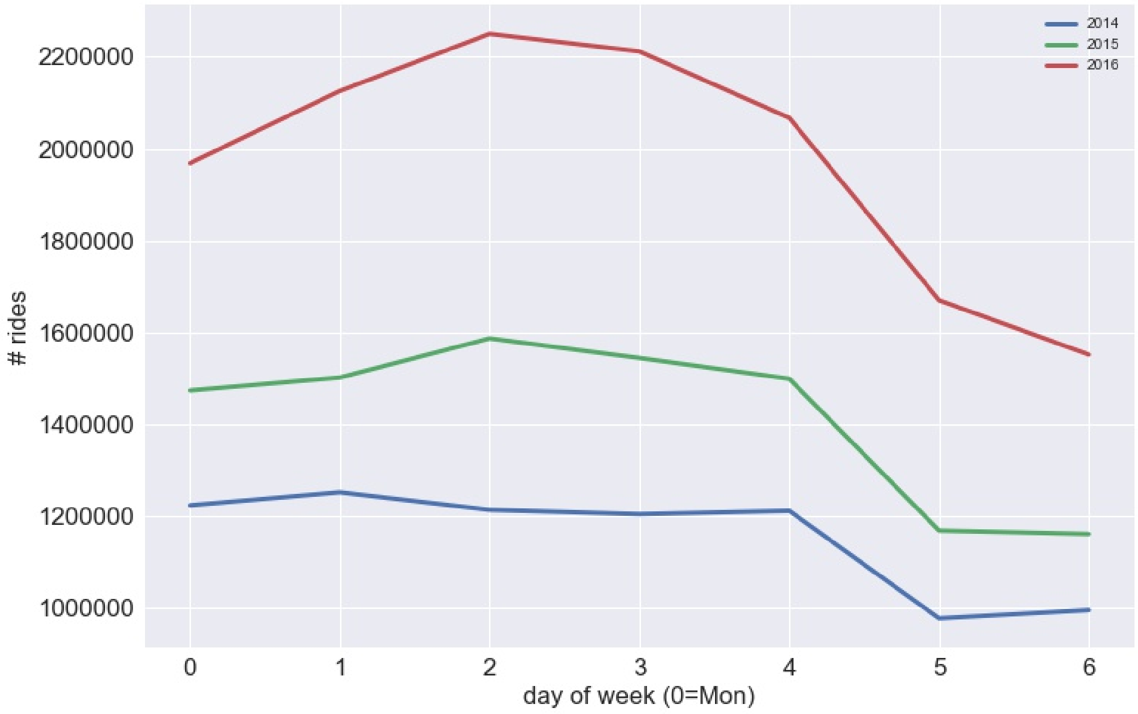

Figure 7: Total number of rides taken for each day of the week. Monday is day 0.

4.2 Weather

4.2.1 Temperature

The daily number of rides is postitively correlated with temperature (Figure 9). This makes sense; anyone who has ridden a bike knows it is much more enjoyable when it's not freezing! A linear regression for 2015 data gives an R^2 of 0.62. It appears that there is a lot more variance for temperatures above 50^o; this is something to look into further. I think this may be partly due to added stations in mid 2016.

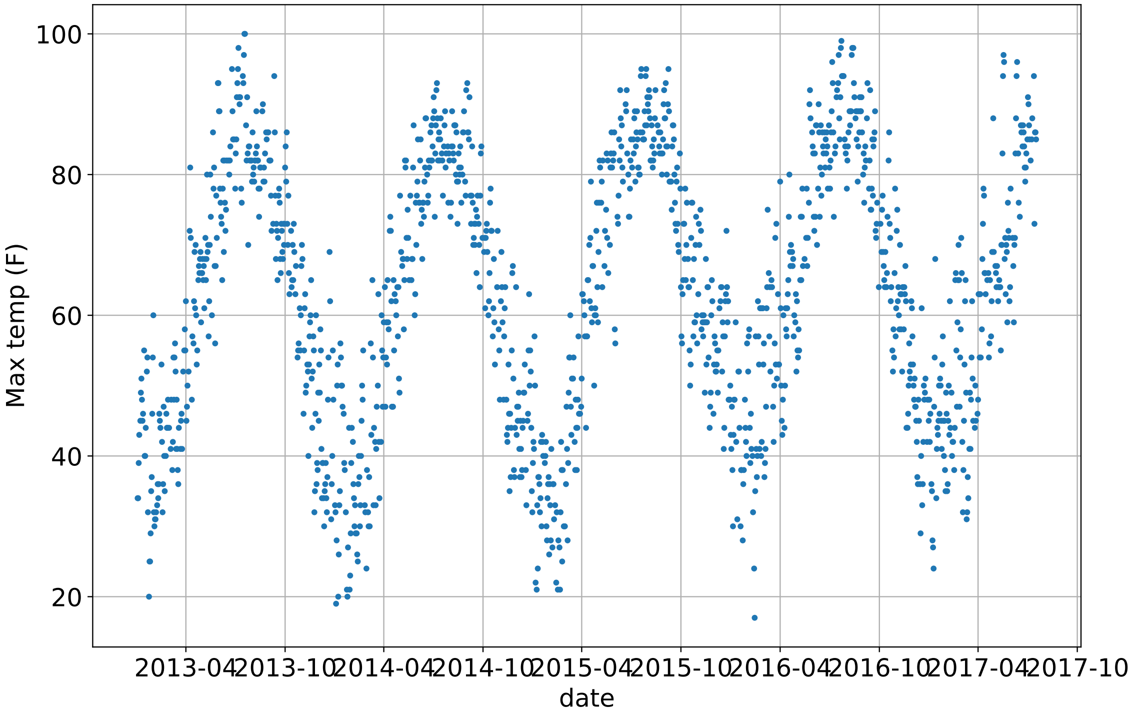

Figure 8: Daily mean temperature (F).

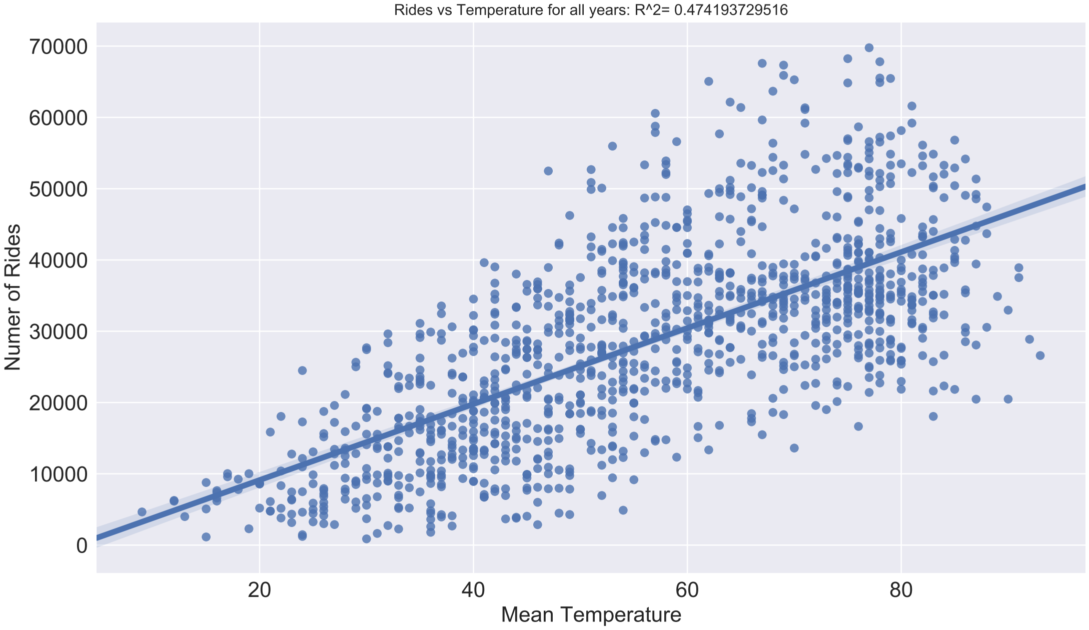

Figure 9: Rides vs Temp

4.2.2 Precipitation

Number of rides is negatively corelated w/ precipitation (Figure 11).

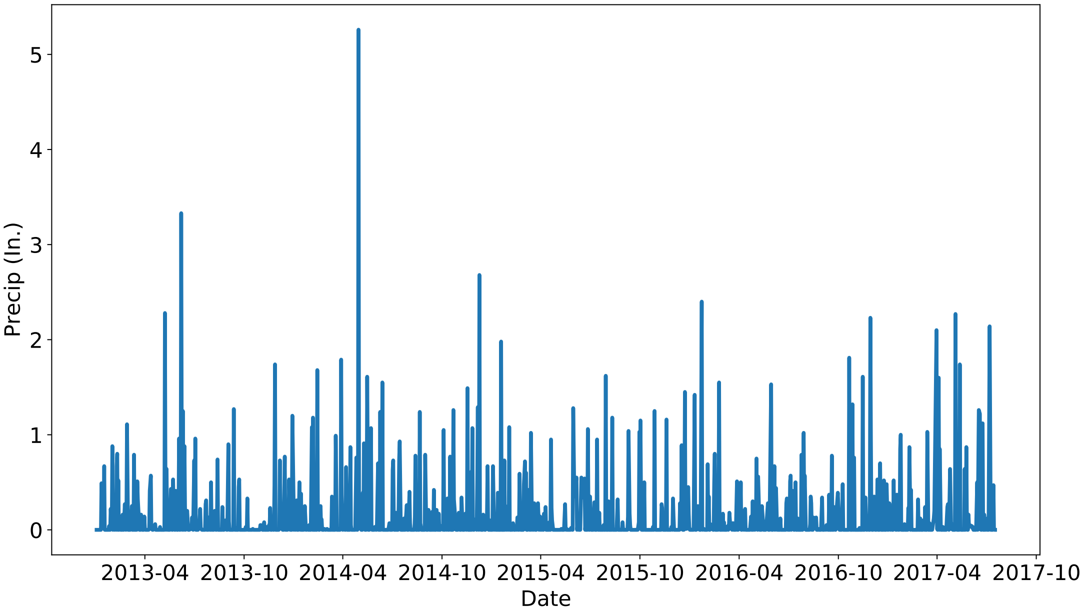

Figure 10: Precipitation (inches).

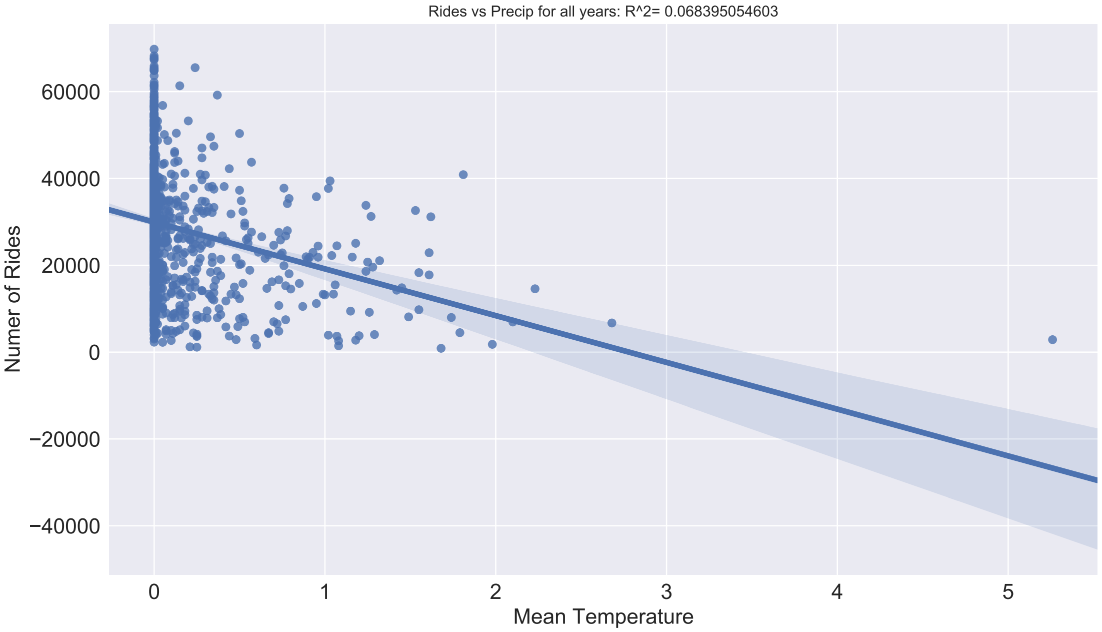

Figure 11: Rides vs precip

4.2.3 Wind Gust

Number of rides is negatively corelated w/ wind (Figure 13).

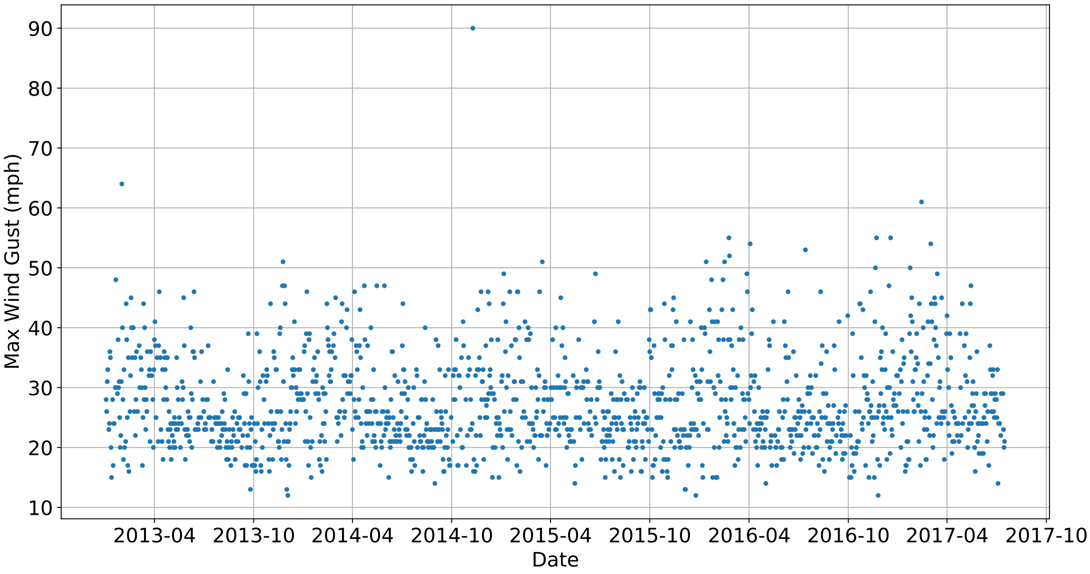

Figure 12: Max wind gust (mph).

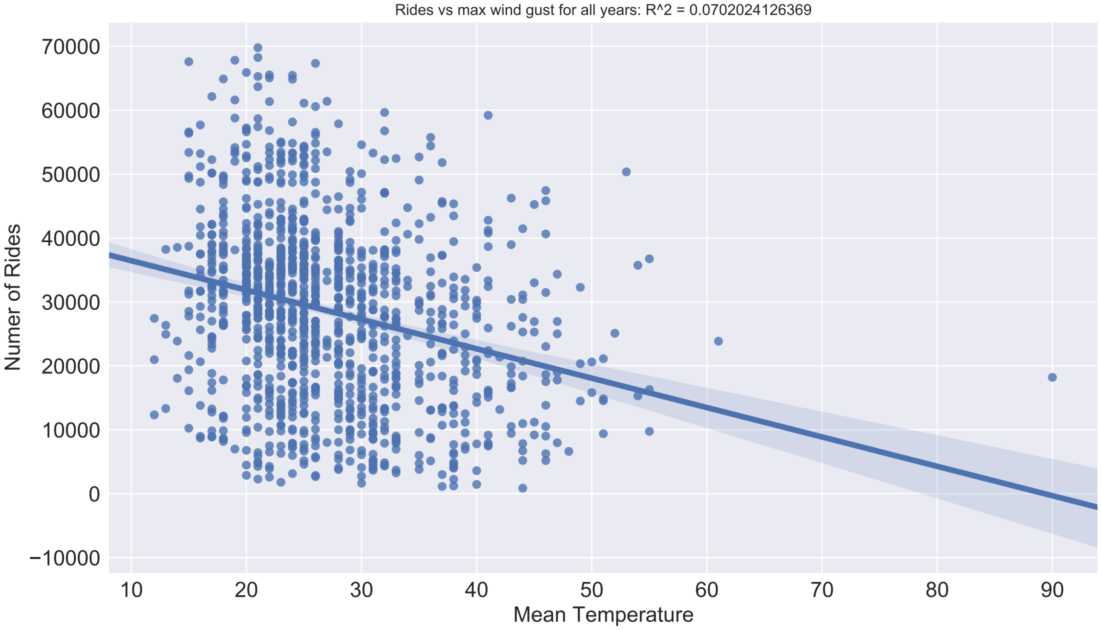

Figure 13: Rides vs max wind gust

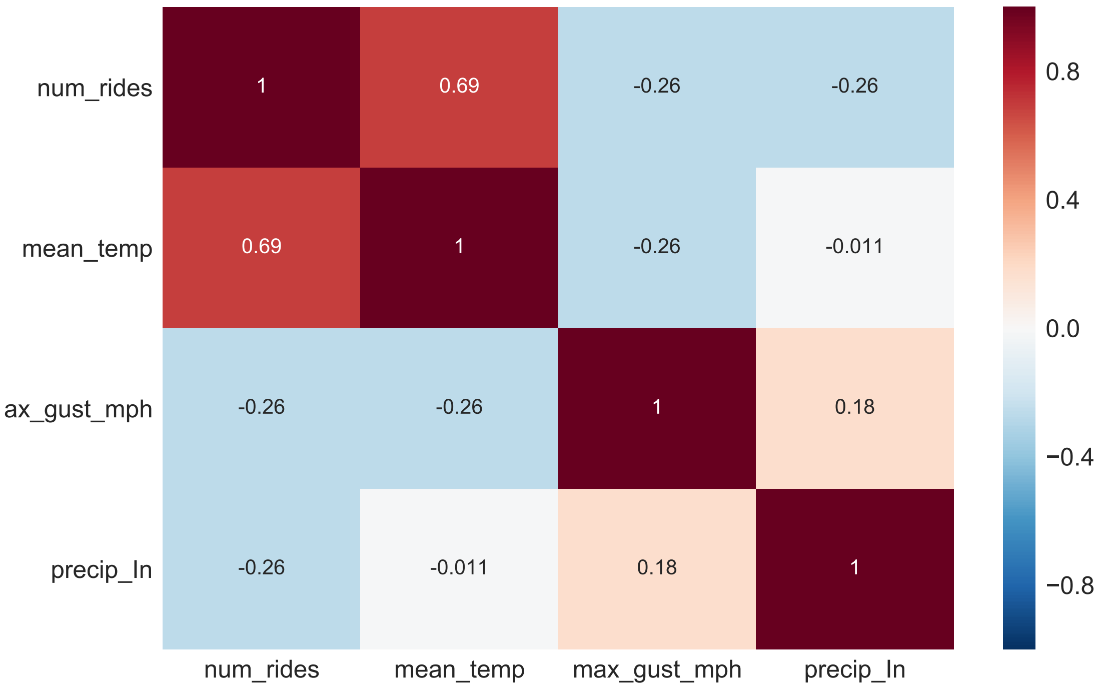

Figure 14: Correlations between daily number of rides and weather variables.

5 Regression Model

I have begun building a regression model with the goal of predicting the daily number of rides taken. Data from 2016 were randomly split into training (70%) and test (30%) sets. I first tried a linear regression with the following features: day of week, number of stations, mean temperature, precipitation, max wind gust, and cloud cover. Day of week was converted to dummy variables. This gave an $R^2 = 0.78$ on the test set.

Examination of the residuals from this model indicated that some of the larger residuals were likely associated with holidays such as Thanksgiving and Christmas (Figure 16). Adding holidays (yes/no) to the linear regression did not improve the model performance much.

I then tried a random forest regressor, optimized w/ GridsearchCV. This achieved an improved $R^2 = 0.87$, and improved the residuals (Figure 17).

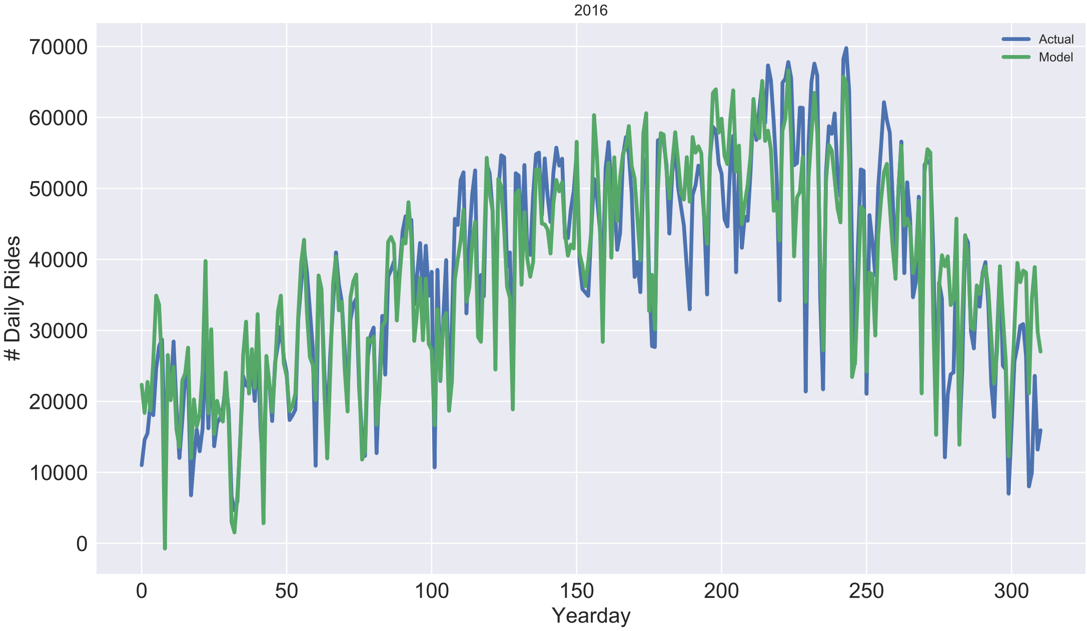

Figure 15: Predictions from linear regression (no holidays) and actual values for 2016.

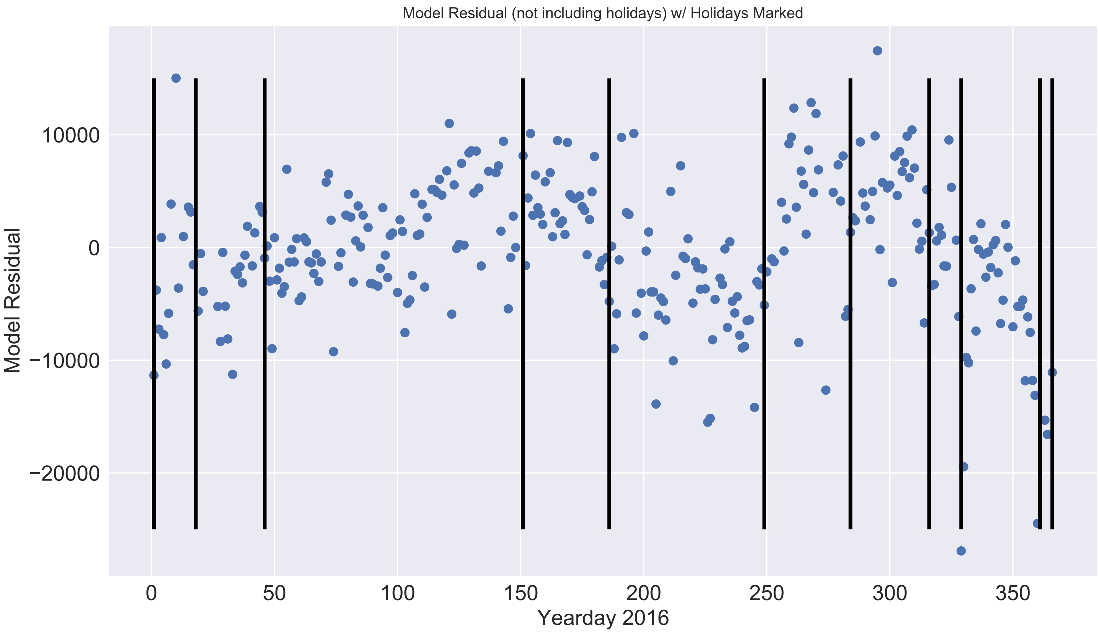

Figure 16: Residuals from linear regression (no holidays) for 2016. Vertical lines indicate holidays.

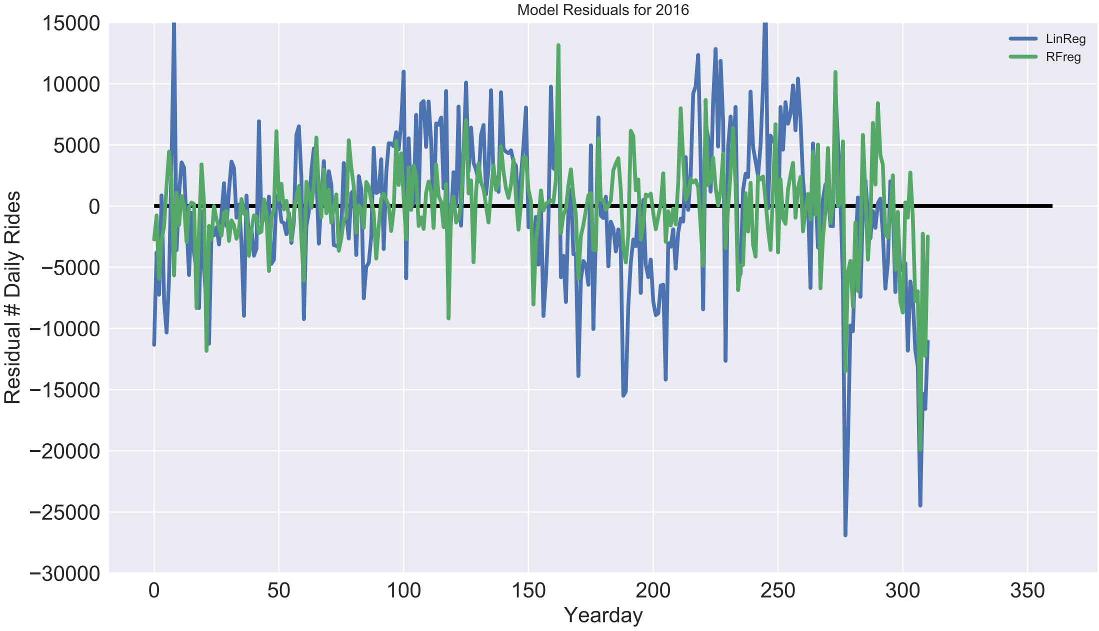

Figure 17: Residuals from linear model and random forest regressor for 2016.

6 Conclusions and Plan

At this point in the project, I have the following conclusions:

- There is a strong weekday/weekend pattern to the number of rides. More rides are taken during the week than on weekends. On weekdays, the majority of rides are taken during rush hours, suggesting heavy use of the system for commuting. On weekends, rides are more distributed throughout the day, suggesting more recreational use of the system.
- An important factor which I did not anticipate before doing the analysis is the changing number of stations and bikes, which essentially raises the baseline level of rides. This needs to be accounted for when trying to diagnose the effect of other variables such as temperature.
- The number of rides is positvely correlated with the daily mean temperature.
- Number of rides is negatively correlated with precipitation and max wind gust.
- A random forest regressor looks promising, with a $R^2 = 0.87$ on the test set for 2016.

Plan:

- Continue working on regression model to try to determine what features affect the number of rides, and predict the number of rides based on these features.
- Use all data in regression. Try splitting based on time, not just randomly.
- Create slide deck/presentation of results.
- Clean and organize final code to reproduce analysis.