Capstone 1 Milestone Report - Analyzing NYC Citibke Usage

Andy Pickering

July 17, 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 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

Station-related variables except id (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.

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 116 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 La-Guardia 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

A deeper dive into the data set: What important fields and information does the data set have? What are its limitations i.e. what are some questions that you cannot answer with this data set? What kind of cleaning and wrangling did you need to do? Are there other datasets you can find, use and combine with, to answer the questions that matter?

The combined historical citibike data was very big; 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. 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 x,x to categorical data type

4   Analysis

Any preliminary exploration youve performed and your initial findings. Test the hypotheses one at a time. Often, the data story emerges as a result of a sequence of testing hypothesis e.g. You first tested if X was true, and because it wasn't, you tried Y, which turned out to be true.

4.1   Citibike Usage

I began with an exploratory analysis of the data and patterns of ridership.
Findings:

- Ridership has increased over the 3 years (Figure 1).

- Part of this increase is due to expansions 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 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.
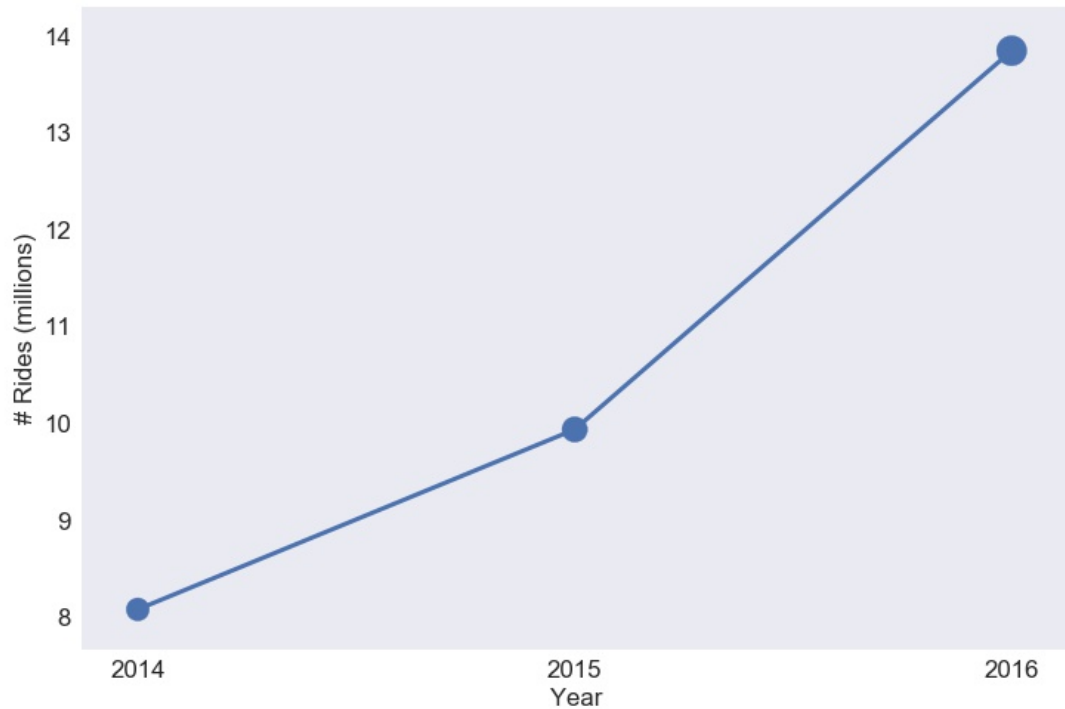
Figure 1: Total number of rides (in millions) taken each year. Only years for which complete data was available are plotted.
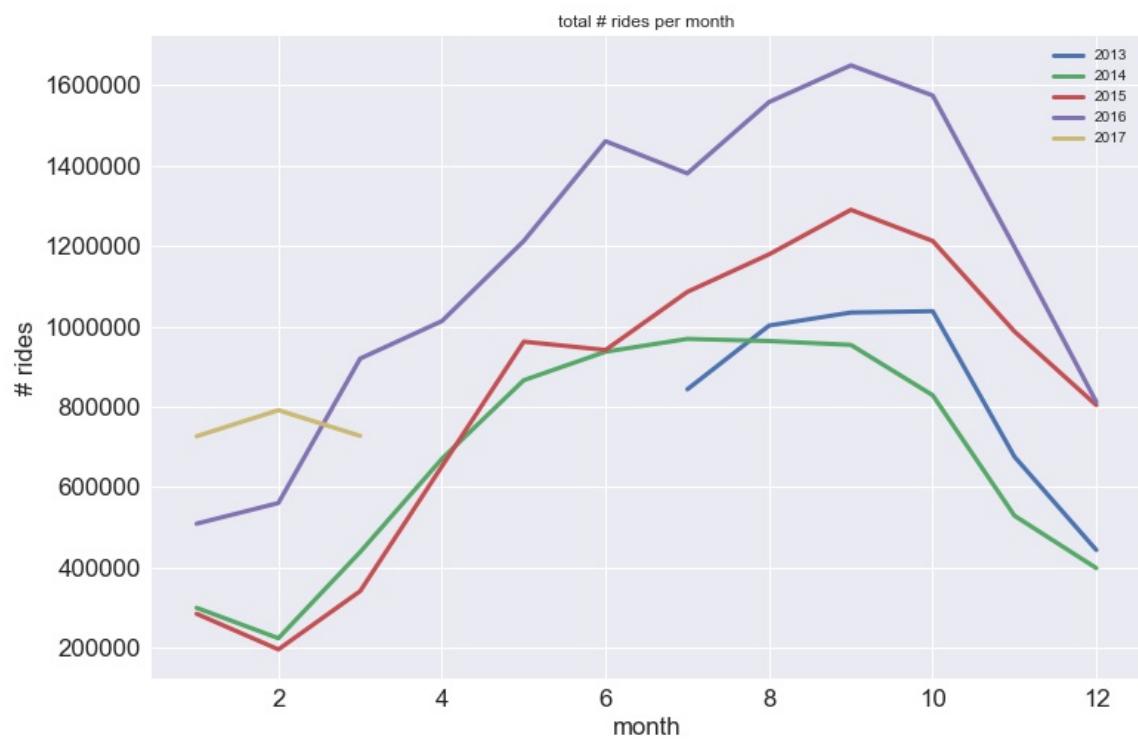
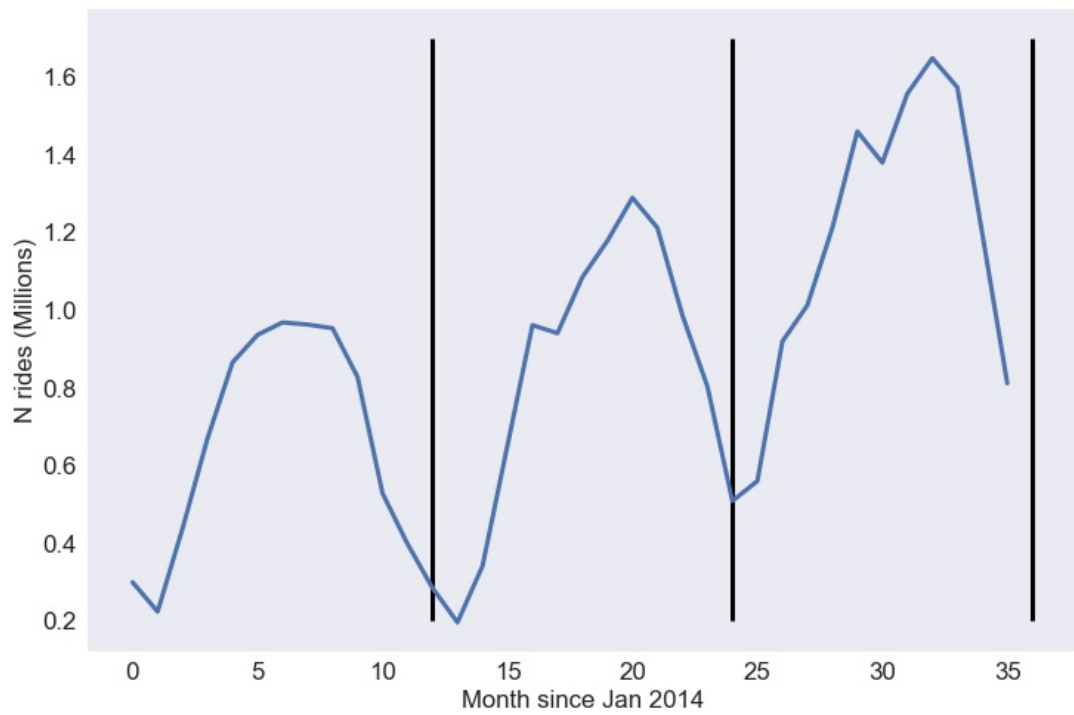Figure 2: Total number of rides taken each month, for different years.

Figure 3: Total number of rides taken per month, plotted consecutively for 2014-2016.
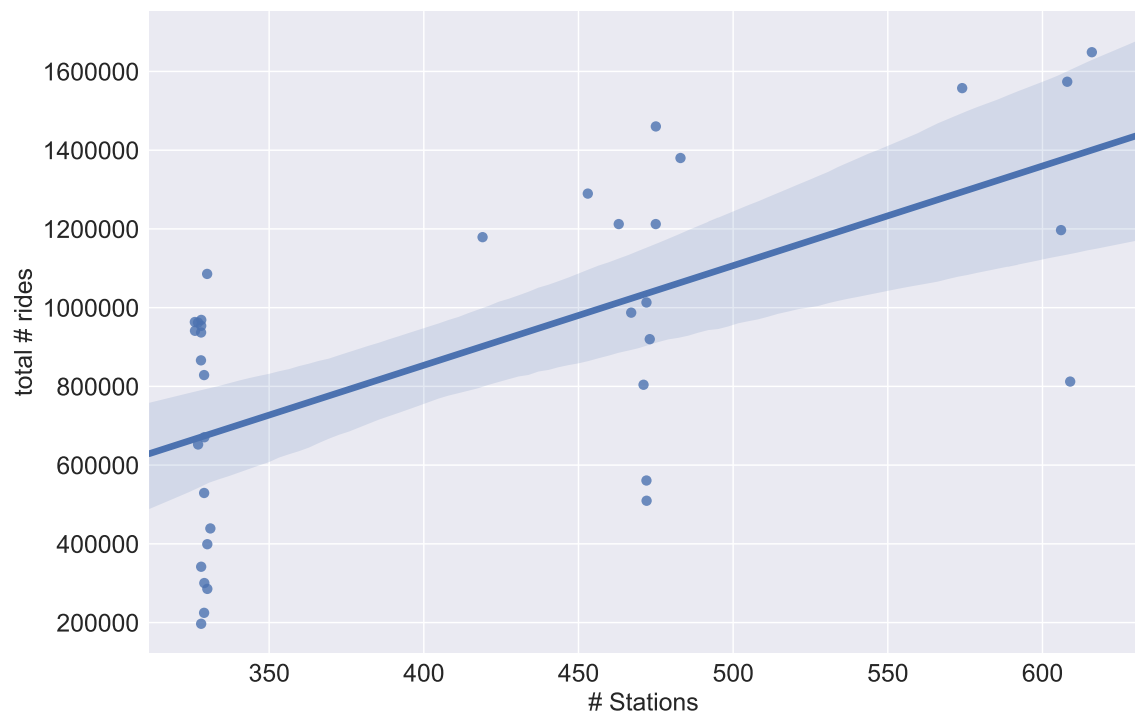
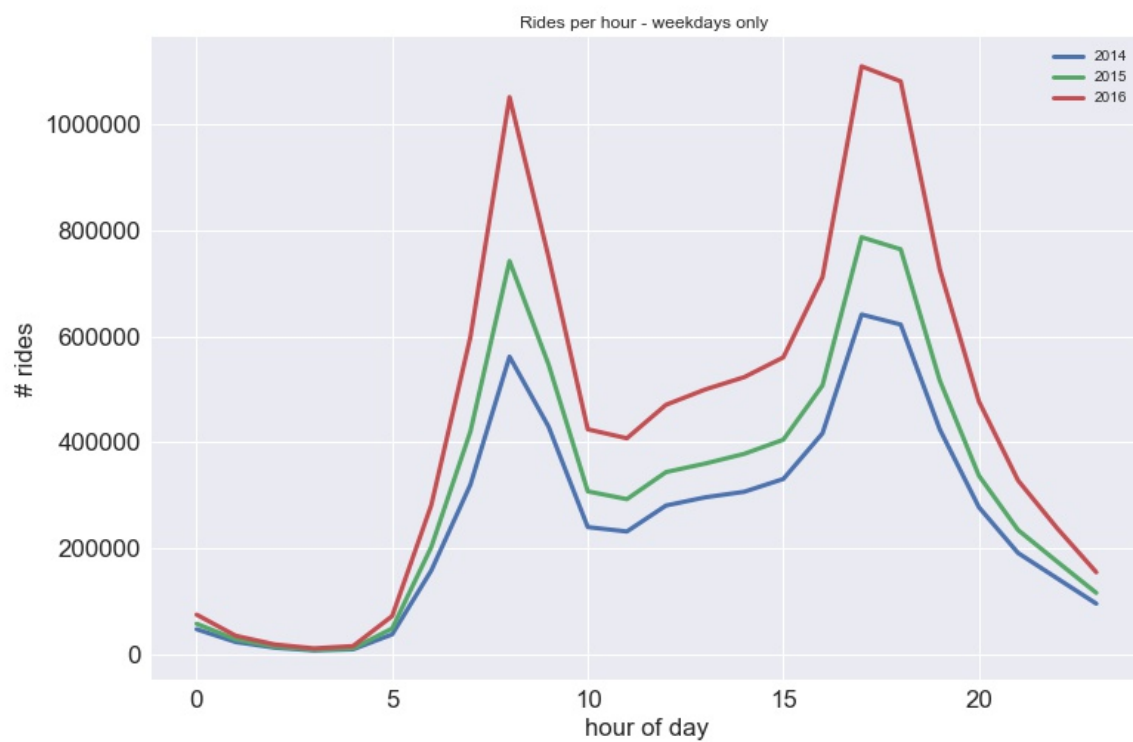Figure 4: Total number of rides vs the number of stations.

Figure 5: Total number of rides taken each hour of day, for weekdays (Monday-Friday).
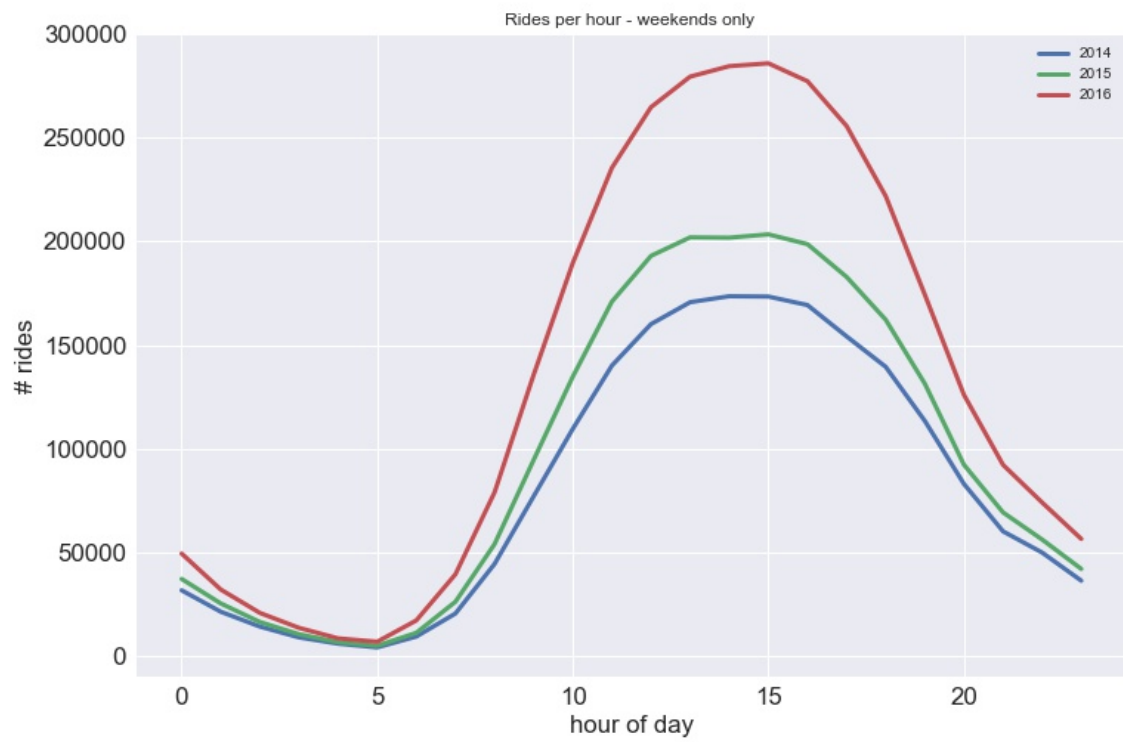
Figure 6: Total number of rides taken each hour of day, for weekend days (Saturday and Sunday).
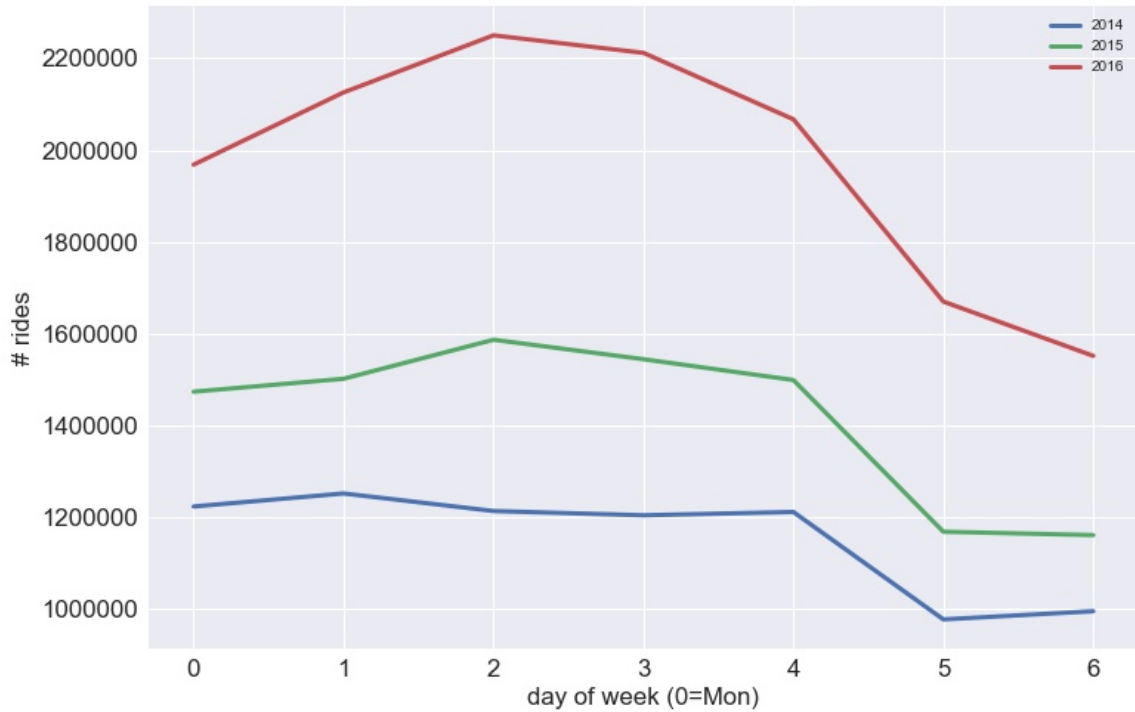
Figure 7: Total number of rides taken for each day of the week. Monday is day 0.

4.2 Relationship with Weather

The daily number of rides is postitively correlated with temperature (Figure 8). 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°; this is something to look into further. I think this may be partly due to added stations in mid 2016.

5 Regression Model

6 Conclusions and Plan

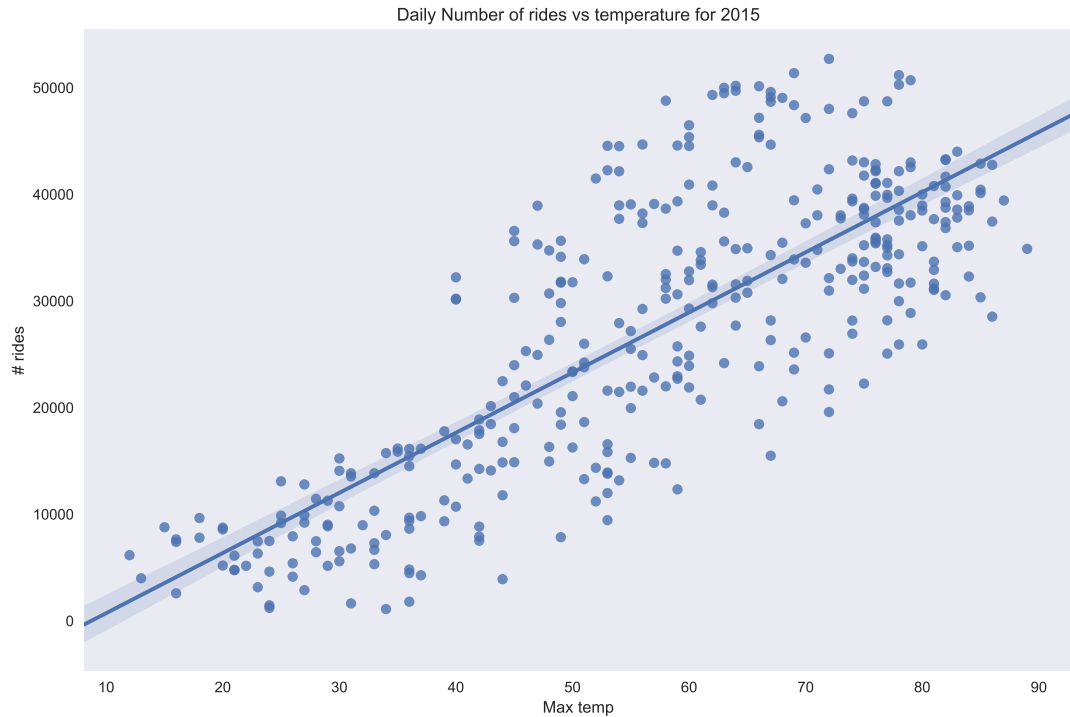At this point in the project, I have the following conclusions:

Figure 8: Total number of rides vs temperature in 2015.

- 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. A linear regression

Plan:

- Look at other weather variables such as rain, clouds, etc;

- Build regression model to try to predict number of rides.

-