Citi Bikes Analysis and Bike Demand Prediction

Julian DeGroot-Lutzner & Adi Salwan

12/14/2017

R Markdown

http://rmarkdown.rstudio.com. **Knit**

Background and Motivation

Our project is inspired by the Kaggle bike sharing demand competition. Bike sharing is a service that has become popular within the last few years in many cities across the United States. Bikes are checked in and out of a network of stations that keep track of each ride, its start and end time, its start/end location, as well as information about subscribers such as gender and age. The Kaggle competition challenged users to predict the hourly bike usage of a test dataset by building a model on a training set combing hourly weather data with hourly bike share usage in Washington, DC.

We planned to first visualize and understand the data and then try to find if there are any inconsistencies in bike stations traffic. Is there a bike station that people take bikes from but do not take bikes back to? If there was an bike station that is a net exporter - meaning it has an excess in demand of bikes - we wanted to make recomendations on how often CitiBikes should manually transport bikes back to the station.

Instead of using the Kaggle data, we wanted to wrangle bike sharing data from a different city's bike share dataset so that we could get familiar with the process of preparing a dataset for analysis. We decided on the Citibikes bike sharing dataset from New York City because there is data available on every ride from January 1st, 2013 to September 31, 2017. However, the dataset is very large, over 40 million observations, and the complete set cannot be ran in R. Our new goal became learning how to work with Big Data in R.

Working with Big Data in R

Big Data in R is when the data cannot fit in to memory. Instead, we stored the data in a SQL database and received the data using 'RMySQL' and 'dplyr.' We based our work process off of the Working with Big Data in R webinar.. The life cycle of a big data analysis project usually involves five parts. Subset (extract data to explore and work with), clarify (become familiar with the data and template a solution), develop (create a working model), productize (automate and integrate), and publish. In other words, first work with a smaller sample, then scale up the work to a larger dataset.

We worked on Jo Hardin's sever at the Pomona math department using MySQL. Since the server is within the Pomona network we needed to navigate both the Pomona ITS and Department firewall. Additonally, our limited permissions on our personal accounts made it difficult to update needed packages within R. These limitations forced us to learn, adapt, and change our project as needed.

The first difficulty in the project was uploading the Citibikes data to a SQL database. Citibikes publishes a zip file for each month of data so there were over 50 seperate files to download. We wrote scripts to download, alter, and write csv files from the the Citibikes server, to our server, and then onto the MySQL database. The difficulties served as a learning experience as we developed better skills in the command line as well as in writing Shell, Python, and SQL code. You can see some of this code in the setup-code folder. Unfortunately, as of now our complete process is not reproducible by a single code but we can come back and add more explanation on how to use this code later.

Dplyr allows integration with many different databases. We originally tried running code on the complete >40 million dataset. However, the SQL queries took too long for us to create a reasonable project within our

relatively short time frame. For example, some queries would take more than a day. In the future we could use other big data techniques like MapReduce to work more quickly. Instead we used the aforementioned big data work schedule of subsetting and scaling. First we took a random sample of 500,000 observations - approxiametly 1 % - from the original dataset.

Understanding the Random Sample

Exploratory Graphs on Ride Use

We wrote a lot of code, but we decided to use as little code as possible in our write up so that it is easier to read. If you're interested in learning how to make cool graphs like these please download our Rmd file and take a look for yourself.
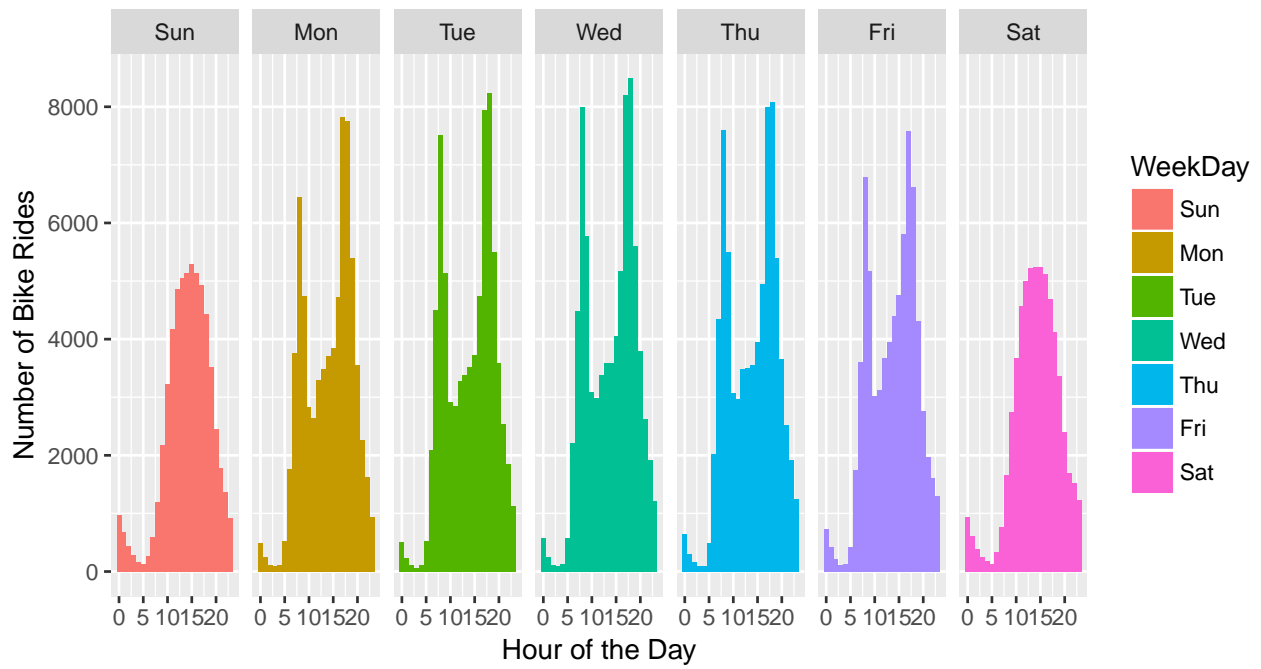
Here are some graphs that should help you get an understanding of CitiBike usage. After the graphs is a write up about some of the takeaways we found.
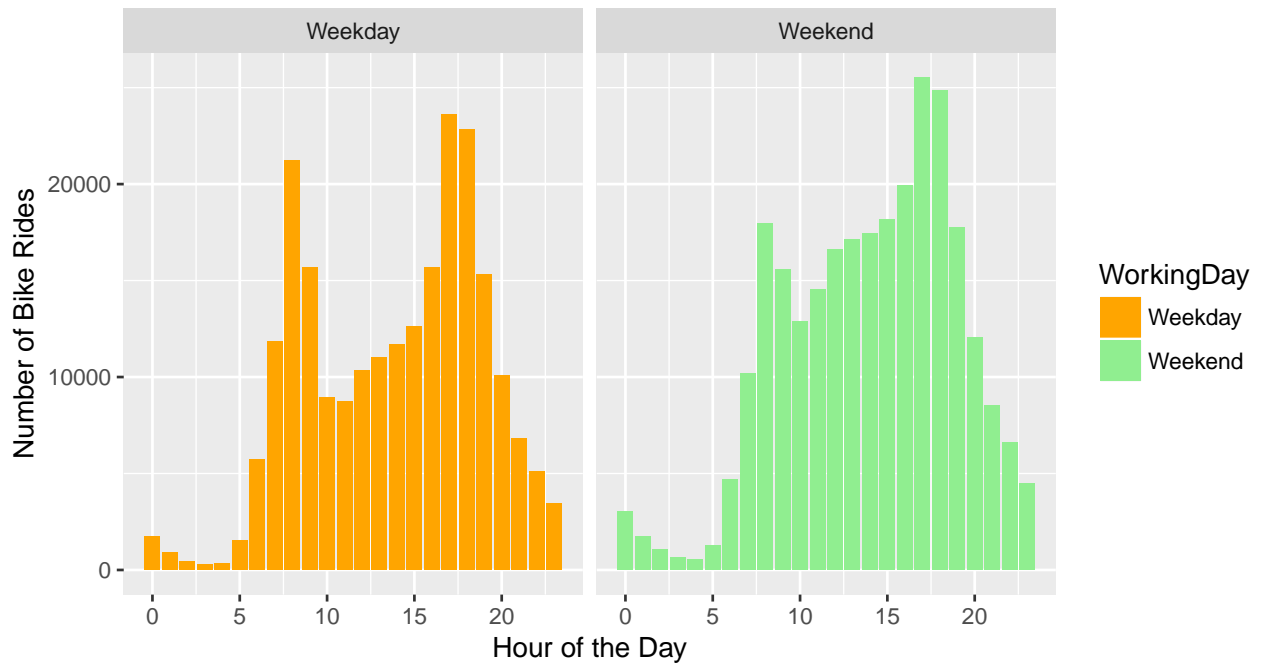
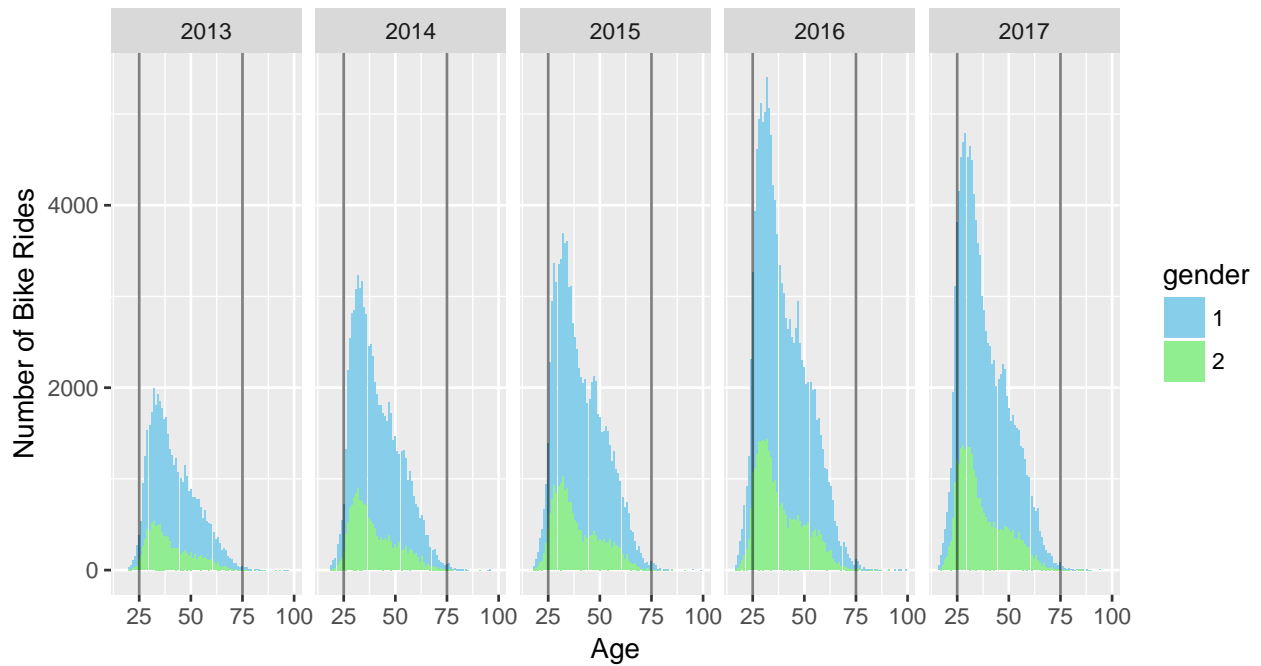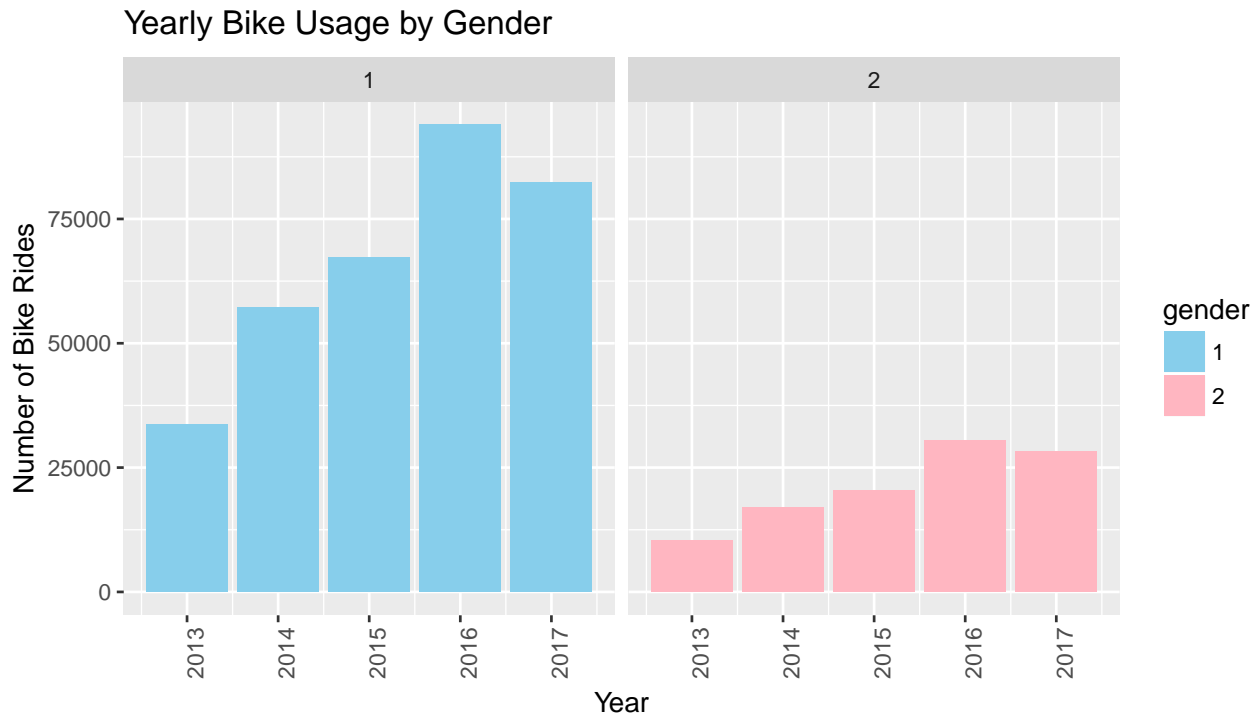### Bike Usage in Different Months (July 2013 to September 2017)
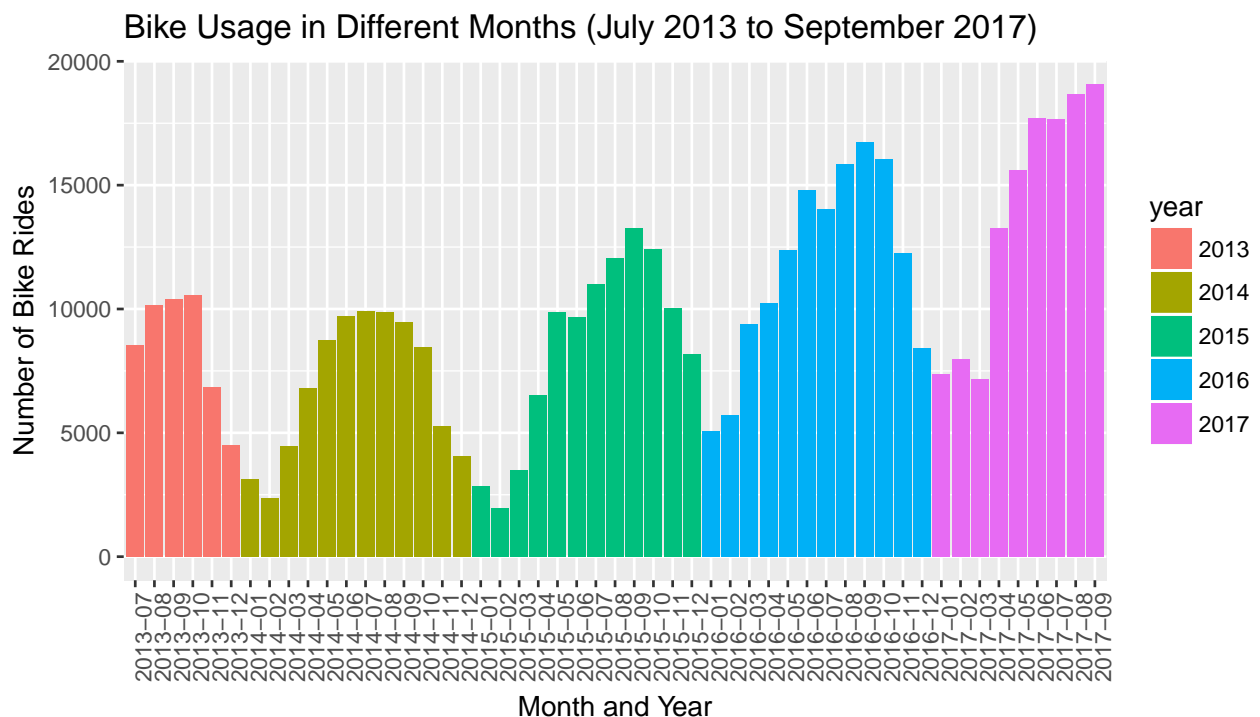
Weekly Bike Rides in 2016
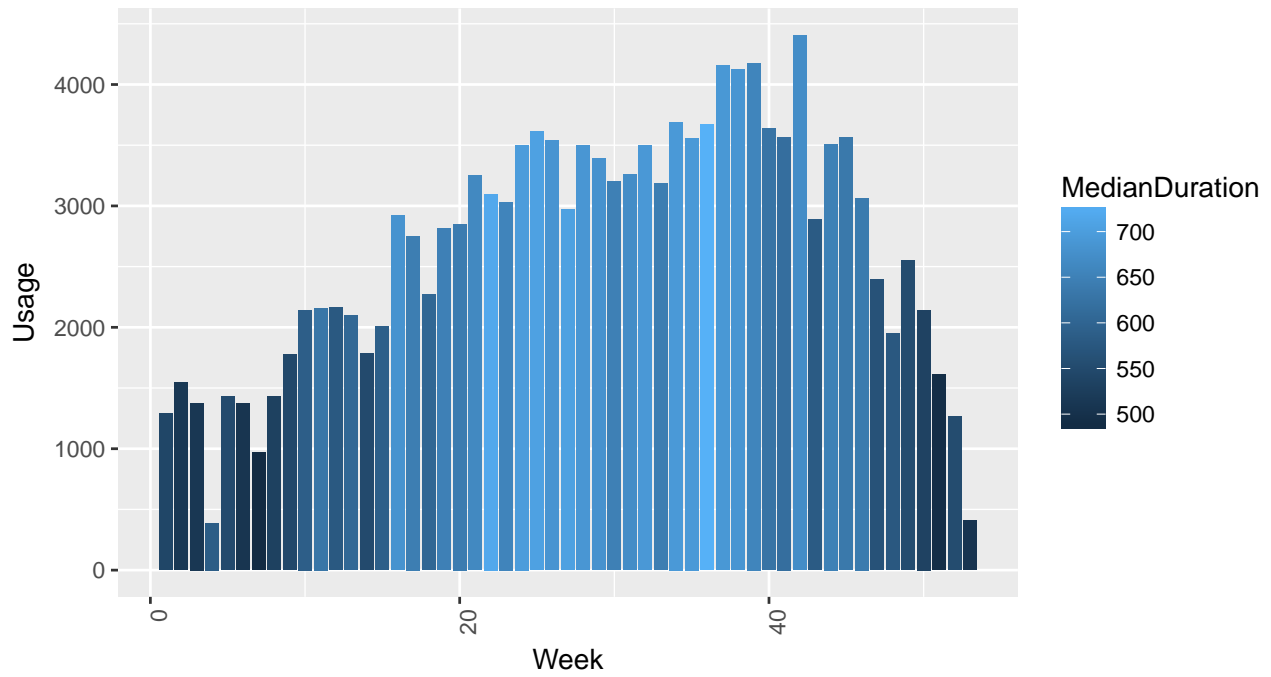
Bike Usage on Different Days

Bike Usage on Weekdays versus Weekends

Yearly Bike Usage by Age and Gender

Yearly Bike Usage by Gender

Based on a preliminary exploratory analysis the following insights were found which could be useful for Citi Bikes:

1) It was found that bike usage for both females and males has increased each year since the initiation of this bike sharing project in May 2013 and since the data became available in July 2013. After the missing observations for gender were removed, it was found that bikes are used predominantly by males. We believe this trend is unexpected and needs to be corrected. We posit that female prefer walking, which is why the bike usage for females is not very high as compared to bike usage for males. The 2017 data is only available till September, which is why when compared to 2016 it shows a decline.

2) It was found that bike usage has increased the most for the 25-50 years age group. The bike usage for the age group of 50-75 has increased significantly too, while the usage for the age group less than 25 increased too. The bike usage for the age group of 75 years or older has not increased much. Based on this information, we believe that the target market for Citi Bikes is 25 to 50 year olds.

3) In the year 2016 bike usage peaked from week 21 to week 44. This corresponds to the dates from May 16 to October 31. This is because bike usage is generally higher during the summer and fall months. It was also found that the median bike ride is longer during these months as the weather does not act as impediment for bikers. The highest median bike ride was in week 36 (12 minutes and 9 seconds), while the lowest was in week 7 (8 minutes and 8 seconds). Hence weather does play a part in biker's decision whether to bike or not. However, it remains to be seen which of the six weather predictors play the most important role in a user's decision to bike or not.

4) Based on the first graph, it can be seen that the overall popularity of this bike sharing system has surged significantly each year. It has increased each year since its inception. The number of bike rides continue to grow, but the number of bike rides in the last 4 months of the years are higher in 2013 than in 2014, albeit not by a big amount. This may have been because of the weather conditions. The average temperature in September 2013 was 67.2 degree Farenheit, 2 degrees higher than the average temperature in September 2014, which could have translated in the marginally lower number of bike rides in in September 2014 than in September 2013. Similarly, the average temperature in December 2013 was 2 degrees lower than the average temperature in December 2014. We hypothesize that weather conditions could be the reason for this slight decrease in bike rentals. However, it is still not quite clear about why the number of bike rides rented in October and November 2013 were marginally lower than

in October and November 2014.

5) The bike usage is significantly higher during the weekdays than during weekends. This could be because the bikes are predominantly used by commuters commuting to work on weekdays. This drop is the highest during the commuting hours from 7am to 9am and from 4pm to 7pm, which makes us believe that commuters account for a majority of the bike rides in New York City. Based on this information, Citi Bikes could shift its major focus to commuters instead of any other group. This could be undertaking or promoting campaigns to bike to work, which would increase their annual memberships.

It is important to note that the exploratory analysis is based on a sample of 500,000 observations, which is 1% of the total population. This sample is representative of the population because of its large size.

Station Analysis

Rationale for Choosing the weather data variables:

The variables relating to the weather data chosen are air temperature in Fahrenheit(tmpf), Dew Point Temperature in Fahrenheit (dwpf), Relative Humidity in percentage (relh), Wind Speed in knots (sknt), Pressure altimeter in inches (alti), and visibility in miles (vsby). We chose these variables because each of these variables can have an impact on a biker's decision to choose or not to choose to bike and can significantly affect bike usage. We chose these variables to build a preliminary model and make predictions based off the model. The other variables like wind direction in degrees from north, sky level coverages, sky level altitudes, and wind gust had some missing observations. We believe that these variables don't affect bike usage much and not including them would not affect the model.