Citi Bikes Analysis and Bike Demand Prediction

Julian DeGroot-Lutzner & Adi Salwan

12/14/2017

R Markdown

[ee]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.

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 because there is data available on every ride from January 1st, 2013 to September 31, 2017. The dataset is very large, over 40 million observations, and the complete set cannot be ran in R. Our new goal was too learn how to work with Big Data in R.

The first difficulty in the project was uploading the Citibikes data to a SQL database. We worked on Hardin's Outlier sever using MySQL. Outlier is within the Pomona network so 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.

Citibikes publishes a zip file for each month of data so there were over 50 seperate files to download.

```
#!/bin/bash
# A script to automate the download of all the Citibikes data

IFS=$'\n'       # make newlines the only separator
set -f          # disable globbing
for link in $(cat < "$1"); do
    wget "$link"
done
```

Working with Big Data in R

Understanding the Random Sample

Exploratory Analyis

Station Analysis