Predicting bike share rentals using weather and days of week in NYC

Cuny Data 621 Final Project

Group3: Sam Cohen-Devries, Alex Low, Neil Hwang, Joseph Elikishvili

Abstract: Bike-sharing programs have become popular in various metropolitan areas around the country. With over 100 major cities and university campuses having adopted the program, the program has emerged as not only a major policy tool to help commuters, but also a tourist attraction to help revive the lagging sector of the local economy. As the bike-sharing programs have evolved into more mature public infrastructure, attention has shifted to the economic viability of the program. In particular, the demand patterns for the bikes can be surprisingly hard to fathom, with seemingly uncountable number of potential covariates. In this paper, we attempt to describe and model daily bike rental volume using a set of predictors that previously have not been looked at closely, namely, the weather conditions and the day of the week. We show that each of these predictors is statistically significant at the bivariate level, with the wind and weekend binary having a significant negative effect, while temperature and weekday binary have a positive effect on the demand volume of bike rentals.

Introduction

In 2013 Mayor Bloomberg launched the Citi Bike bike share program in New York City. Modeled on similar initiatives in other cities across the world, the program is designed to provide bikes on demand to the public and promote alternative forms of transportation. Members of the public can visit stations across the city to rent a bike and then return it to the same or different station within the city.

Maintaining and operating this system requires significant data and analytics for the program to run smoothly. Citi Bike needs to be able to re-stock bike locations on a regular basis to ensure that bikes are readily available to customers. Usage patterns might vary on a wide set of variables such as location, time of day, day of week, weather conditions and many others.

In this study, our goal was to develop a data set and predictive algorithms to assist Citi Bike administrators in operating the system. In particular, we wanted to design an approach to predict usage based on location, weather, hour of day and day of the week.

Literature Review

Several studies have provided a broad background for the Citi Bike program, and others have examined in particular the current and future states of the program.

Kaufman and O'Connell at the Rudin Center for Transportation Policy at the NYU Wagner School released in 2017 a detailed report on the usage and capital statistics of the program. For instance, they reported that by the end of 2017, Cit Bike was estimated to have doubled in size since its inception in May 2013 with 12,000 bikes at 700 stations in Manhattan, Brooklyn, and Queens. As for the customer segmentation, the authors wrote that 73% are men and 25% are women, while a majority of them are "casual pass" holders. However, they observed that most trips are by annual members, who number 115,000 compared to over half a million casual passholders. As for accessibility, the authors noted that just 10 of the 700 stations generated over 7% of the trips, most notably those near Grand Central Station, Port Authority, and Penn Station.

Regarding this last point about the optimal location and number of stations, Garcia-Palomares et al used GIS-based methodology to examine the characterization of the optimization problem. A key finding from their study was that there exists a diminishing returns to the number of stations in terms of the target customer reach and accessibility for customers.

On the topic of predicting ridership of the bike sharing programs, Rixey has published a paper examining the significant predictors of the demand for bike rides by looking at multiple programs. In particular, the author found through a regression analysis that the following predictors were statistically significant: population density; retail job density; bike, walk, and transit commuters; median income; education; presence of bikeways; nonwhite population (negative association); days of precipitation (negative association); and proximity to a network of other bike-sharing stations.

In a related study, Daddio concluded in his masters thesis at the University of North Carolina at Chapel Hill that, based on a study of the Capital Bikeshare program in the D.C. area, the following predictors were the most statistically significant: population (aged 20-39), minority population, retail density (using alcohol licenses as a proxy), metrorail stations, and distance from the center of the bicycle sharing system.

On the technical aspect of our study involving a regularization method based on convex penalties, we note the technique published by Tibshirani et al for regression called "elastic net" that uses a penalty based in part on L1 (lasso) and L2 (the ridge regression) that helps develop a parsimonious model.

Methodology

For this study, we accessed three key data sets: Citi Bike usage data[1]; and two weather data sets - hourly precipitation and daily weather available from the National Climatic Data Center (NCDC), part of the National Oceanic and Atmospheric Administration (NOAA).[2]

From the Citi BIke data set, we collected trip date, trip start time, location ID and other key variables for exploration - which we grouped into hourly data. From the NCDC data sets, we collected hourly precipitation, daily minimum, maximum and "midpoint" temperature, and other daily weather variables.

We focused our research on June-December 2013, since the hourly precipitation data and Citi Bike usage data only overlapped for these dates. We then selected three months from different seasons - June, September and December - since there was too much data to combine the whole period and we wanted data from diverse seasons in case usage varied depending on month. We joined all these variables together based on hour of the day.

Other transformations included data cleaning such as removal of several records with missing data, converting some of the Na values to 0 and minor changes to categorical variable. This was followed with Box-cox transformation to alleviate some of the skew that was present in the data. Then, the predictors were standardized to minimize the effect of the different units of the predictors on the regression modeling and to enhance the interpretability of the model parameter estimates.

To test models, we created a number of models, beginning with the assumption that a generalized linear model with a Poisson distribution would be best, considering we are predicting a count variable. We first compared the null model (n ~ 1) against the full model (n ~ all predictors). Given that the full model scored higher, we went with an elimination process, as opposed to an additive process to determine the most efficient model. We compared models using the Akaike Information Criterion (AIC), as well as by analyzing plotted residuals vs. fitted values.

The histograms of transformed variables are depicted below in figure 1.

[1] Sourced from www.citibikenyc.com
[2] Sourced from www.ncdc.noaa.gov

Data transformation

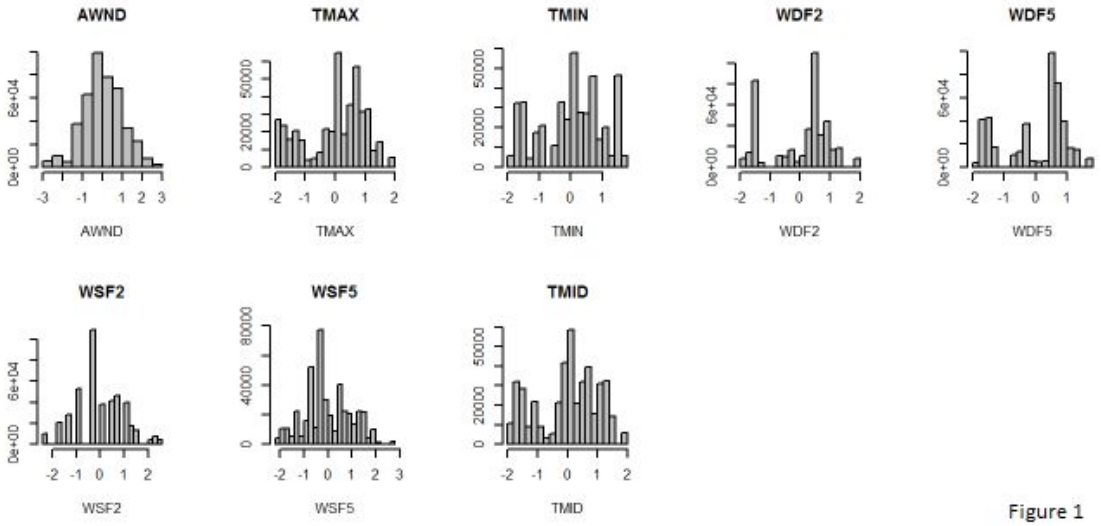


Figure 1

We have also analyzed predictor variables to identify any of the existing correlations as shown in the figure 2. below. Correlation plot suggested heavy correlation between several weather related variables that we took a note of and further used in model creation.

Correlation plot

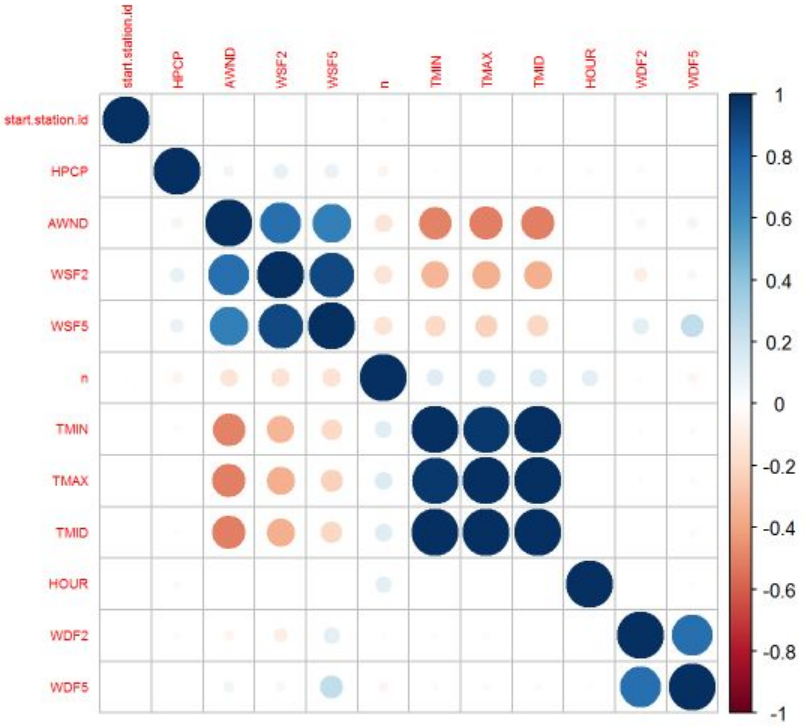
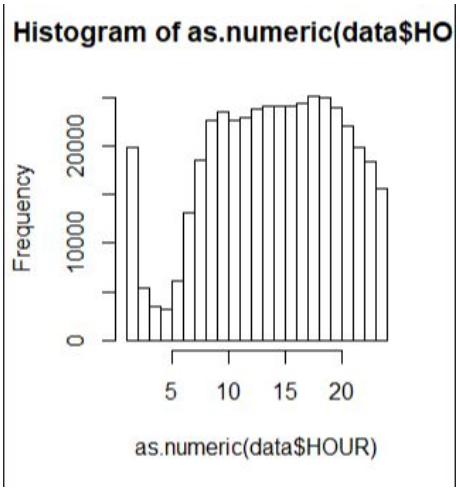


Figure 2

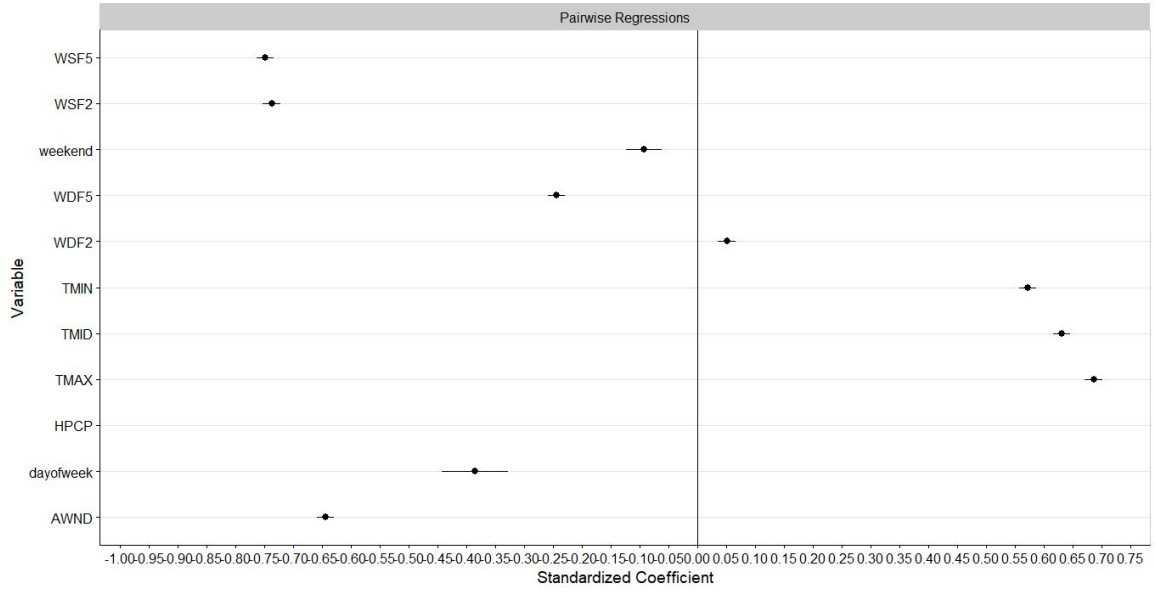
In addition to the variables in the source datasets, we derived several variables that we thought would provide interesting insights for us. One is the day of the week for each observation. And another is an indicator variable for whether a date in question falls on a weekend (i.e., Friday, Saturday, Sunday). This last point was especially important for us to consider since our literature review revealed that a majority of the rides are by customers that work in the city, while an overwhelming majority of the customers are “casual” passholders that primarily ride the bikes on weekends. The third derived predictor variable was whether the time of the rental falls on a “peak” period. As is shown below, the rental volume distribution is uniform from about 7am to about 8pm, with a noticeable drop thereafter:



Results

To get a sense of the relative importance of each of the individual predictors in predicting the response, we first conducted bivariate correlations between the response variable and each of the predictors, and estimated the slope coefficients for the resulting simple linear regression lines. Below is the pairwise bivariate plots that show the coefficients and the

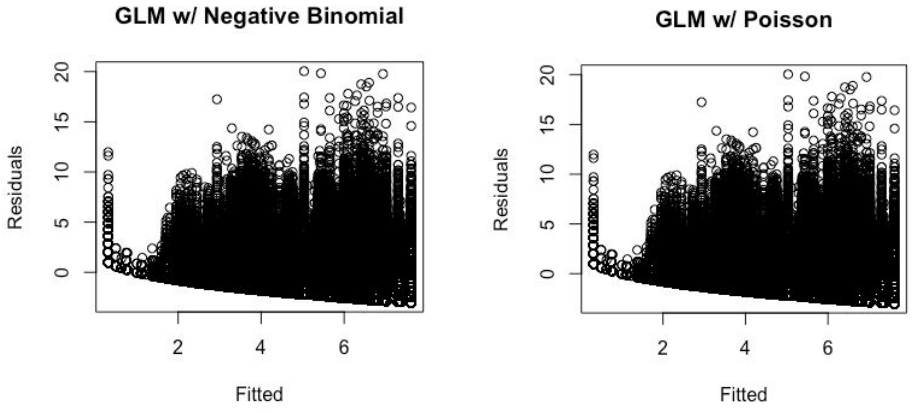
standard errors for all the predictors:

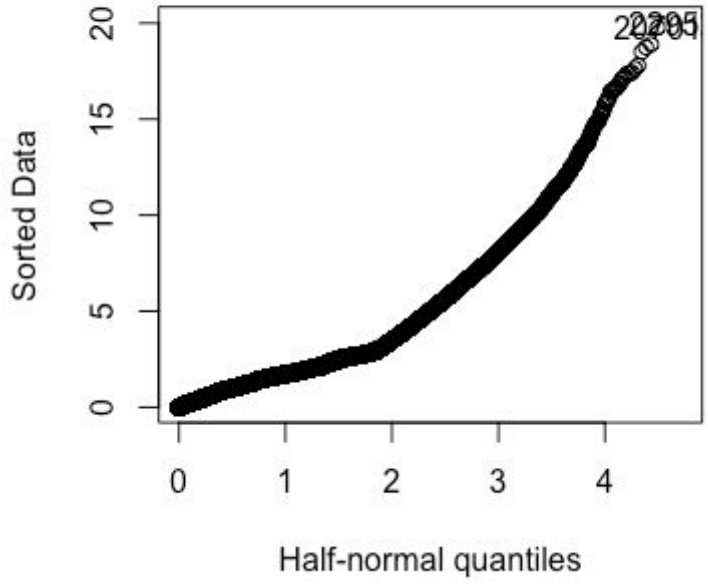


The above plot shows that all of the predictors are significant at the individual predictor level, since none of the predictors' standard errors crosses the y-intercept. The interpretation of the plot would be in terms of the number of standard deviations in each predictor. For instance, a one standard deviation increase in AWND (average wind speed) is associated with about a 0.65 unit decrease in hourly rentals. Not surprisingly, all of the wind speed predictors are negatively associated with the response, while temperature is positively related.

One interesting observation is that a weekend dummy variable is negatively associated with the rentals. This suggests that a significant number of the riders work in the city, and hence the tourist demand for the weekend rental volume might be overshadowed by the decline in the commuters' demand for the bikes.

Although we began with the assumption that a Poisson distribution would be ideal for our predicted value, when compared with a generalized linear model using a negative binomial distribution, the latter scored better using the AIC score, although both models showed a nearly identical distribution of residuals. We were also fortunate that the model scored best without transforming the response variable, which makes interpretation more straight-forward.





Discussion

We were not surprised to discover that that the most significant factor in our model was precipitation. We would not expect to see a high usage rate for bicycles if it is raining outside. As mentioned before, the results did not show higher usage rate on the weekends, which we found to be counter-intuitive.

Although we were able to settle upon a model for this case, it was not ideal in terms of standard statistical analysis. It does not necessarily appear to be due to shortcomings of the model, but instead more likely that some additional data points might be useful.

One limitation of our data set was that many of the weather variables were only available at a daily level, whereas hourly weather variables could make our prediction more precise.

Another significant limitation was that our data set was collected for months during the first 6 month of the launch of Citi Bike. It is very possible that usage at the outset of the program might vary from usage once the program got into a more steady state.

We would recommend that Citi Bike administrators further investigate the impact of weather on usage rates to inform how they restock bike stations. We recommend that they fund future research that looks more closely at usage rates based on a broader set of hourly conditions as that data becomes available and over a longer time period to ensure that the patterns we noted were not specific to the launch phase of the program.

References

Daddio, D.W. (2012). *Maximizing Bicycle Sharing: An Empirical Analysis of Capital Bikeshare Usage*. (Unpublished masters thesis). University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33(1)*, 1-22.

Garcia-Palomares, J., Gutierrez, J., and Latorre, M. (2012). Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography, 35(1-2),* 235-246.

Kaufman, S. and O'Connell, J. *Citi Bike: What Current Use and Activity Suggest for the Future of the Program*. New York, NY: NYU Wagner Rudin Center for Transportation Policy and Management.

Rixey, R. (2014). Station-Level Forecasting of Bikesharing Ridership Station Network Effects in Three U.S. Systems. *Journal of the Transportation Research Board, 2387*, DOI: 10.3141/2387-06.

Appendix

Data Preparation

```
library(stringr)
library(dplyr)
library(plyr)

citibike_data <- read.csv("C:/Data/201306-citibike-tripdata.csv", stringsAsFactors = FALSE)
weather_0613 <- read.csv("C:/Data/1351207.csv", stringsAsFactors = FALSE)
citibike_data$starttime <- substr(citibike_data$starttime,1,nchar(citibike_data$starttime)-3)
citibike_data$starttime <- substr(citibike_data$starttime,1,nchar(citibike_data$starttime)-3)

citi_counts <- citibike_data %>% dplyr::group_by(starttime) %>% count(start.station.id)

citi_counts$starttime <- str_replace_all(pattern = "-",replacement = "",citi_counts$starttime)
citi_counts$starttime <- substr(citi_counts$starttime,1,nchar(citi_counts$starttime)-1)
```

```
citi_counts$starttime <- paste(citi_counts$starttime,":00", sep="")

colnames(citi_counts)[1] <- "DATE"

citi_counts2 <- join(citi_counts,weather_0613,by="DATE",type="left")

citi_counts2$DATE2 <- substr(citi_counts2$DATE,1,nchar(citi_counts2$DATE)-6)

dailyweather_0613 <- read.csv("C:/Data/1351305.csv", stringsAsFactors = FALSE)
dailyweather_0613$DATE <- str_replace_all(pattern = "-",replacement = "",dailyweather_0613$DATE)

colnames(dailyweather_0613)[6] <- "DATE2"

dailyweather_0613 <- dailyweather_0613[,-c(1:5)]

citi_counts2 <- citi_counts2[,-c(4:8)]

citi_june <- merge(citi_counts2,dailyweather_0613,by="DATE2")

citi_june <- citi_june[,-c(19:24)]
citi_june <- citi_june[,-c(7,11,14)]

citi_june$TMID <- (citi_june$TMIN+citi_june$TMAX)/2

citi_june$HOUR <- str_sub(citi_june$DATE,10,14)

citi_june <- citi_june[,-2]

citi_june$start.station.id <- as.character(citi_june$start.station.id)

citi_june <- citi_june[,-c(6:8)]

#######

####Data EDA and Initial Transformation

#Read the data from 2 files
data1 <-
read.csv("C:/Users/joseph/Documents/GitHub/Cuny/Spring2018/621/FinalProject/citi_total1.csv",
header = TRUE)
data2 <-
read.csv("C:/Users/joseph/Documents/GitHub/Cuny/Spring2018/621/FinalProject/citi_total2.csv",
header = TRUE)

#Combine data into 1 file
data = rbind(data1, data2)

#Remove index
data$X = NULL
```

```
dim(data)

#Replacing
data$HPCP[is.na(data$HPCP)] <- 0

#Check for missing values
colSums(is.na(data))


#remove last line
tail(data)
data <- data[-nrow(data),]


library(Amelia)
missmap(data, main = "Missing values")

#Check for missing values
colSums(is.na(data))

#Check data types
str(data)

data$HOUR = as.numeric(factor(data$HOUR))

#Data Table
table.desc <- describe(data)
table.prep <- as.matrix(table.desc)
table.round <- round((table.prep), 2)
kable(table.round)

#Boxplots and Histograms
data1 = data
data1$X = NULL
data1$DATE2 = NULL
par(mfrow = c(3,5), cex = .5)
for(i in colnames(data1)){
  boxplot(data1[,i], xlab = names(data1[i]),
          main = names(data1[i]), col="grey", ylab="")
}

par(mfrow = c(3,5), cex = .5)
for(i in colnames(data1)){
  hist(data1[,i], xlab = names(data1[i]),
       main = names(data1[i]), col="grey", ylab="")
}

data_trans = data
data_trans$DATE2 = NULL
data_trans$start.station.id = NULL
data_trans$n = NULL
```

```
data_trans$HPCP = NULL
data_trans$HOUR = NULL

#Boxcox

library(caret)
dataB = preProcess(data_trans, c("BoxCox", "center", "scale"))
datanorm = data.frame(dataB = predict(dataB, data_trans))
#Fix the colnames
colnames(datanorm) = colnames(data_trans)

#Transformation Review

par(mfrow = c(3,5), cex = .5)
for(i in colnames(datanorm)){
  hist(datanorm[,i], xlab = names(datanorm[i]),
       main = names(datanorm[i]), col="grey", ylab="")
}

datanorm$DATE2 = data$DATE2
datanorm$start.station.id = data$start.station.id
datanorm$n = data$n
datanorm$HPCP = data$HPCP
datanorm$HOUR = data$HOUR

## Correlation plot

library(corrplot)
#install.packages("corrplot")
correlations <- cor(data1)
corrplot(correlations, order = "hclust", tl.cex = 0.55)

######

citibike_data_sept <- read.csv("C:/Data/201309-citibike-tripdata.csv", stringsAsFactors = FALSE)
rain_0913 <- read.csv("C:/Data/1351594.csv", stringsAsFactors = FALSE)
citibike_data_sept$starttime <-
substr(citibike_data_sept$starttime,1,nchar(citibike_data_sept$starttime)-3)
citibike_data_sept$starttime <-
substr(citibike_data_sept$starttime,1,nchar(citibike_data_sept$starttime)-3)

citi_counts_sept <- citibike_data_sept %>% dplyr::group_by(starttime) %>% count(start.station.id)
duration_sept <- citibike_data_sept %>% dplyr::group_by(starttime,start.station.id) %>%
summarize(median.duration = median(tripduration))
#citi_counts_sept <- merge(citi_counts_sept,duration_sept, by=c("starttime","start.station.id"))

citi_counts_sept$starttime <- str_replace_all(pattern = "-",replacement = "",citi_counts_sept$starttime)

citi_counts_sept$starttime <- paste(citi_counts_sept$starttime,":00", sep="")

colnames(citi_counts_sept)[1] <- "DATE"
```

```
citi_counts_sept2 <- join(citi_counts_sept,rain_0913,by="DATE",type="left")

citi_counts_sept2$DATE2 <- substr(citi_counts_sept2$DATE,1,nchar(citi_counts_sept2$DATE)-6)

dailyweather_0913 <- read.csv("C:/Data/1351595.csv", stringsAsFactors = FALSE)
dailyweather_0913$DATE <- str_replace_all(pattern = "-",replacement = "",dailyweather_0913$DATE)

colnames(dailyweather_0913)[3] <- "DATE2"

dailyweather_0913 <- dailyweather_0913[,-c(1:2)]

citi_counts_sept2 <- citi_counts_sept2[,-c(4:5)]

citi_sept <- merge(citi_counts_sept2,dailyweather_0913,by="DATE2")

citi_sept<- citi_sept[,-c(16:21)]
citi_sept <- citi_sept[,-c(7,8,11)]

citi_sept$TMID <- (citi_sept$TMIN+citi_sept$TMAX)/2

citi_sept$HOUR <- str_sub(citi_sept$DATE,10,14)

citi_sept <- citi_sept[,-2]

citi_sept$start.station.id <- as.character(citi_sept$start.station.id)

#######

citibike_data_dec <- read.csv("C:/Data/201312-citibike-tripdata.csv", stringsAsFactors = FALSE)
rain_1213 <- read.csv("C:/Data/1351616.csv", stringsAsFactors = FALSE)
citibike_data_dec$starttime <-
substr(citibike_data_dec$starttime,1,nchar(citibike_data_dec$starttime)-3)
citibike_data_dec$starttime <-
substr(citibike_data_dec$starttime,1,nchar(citibike_data_dec$starttime)-3)

citi_counts_dec <- citibike_data_dec %>% dplyr::group_by(starttime) %>% count(start.station.id)
#duration_sept <- citibike_data_sept %>% dplyr::group_by(starttime,start.station.id) %>%
summarize(median.duration = median(tripduration))
#citi_counts_sept <- merge(citi_counts_sept,duration_sept, by=c("starttime","start.station.id"))

citi_counts_dec$starttime <- str_replace_all(pattern = "-",replacement = "",citi_counts_dec$starttime)

citi_counts_dec$starttime <- paste(citi_counts_dec$starttime,":00", sep="")

colnames(citi_counts_dec)[1] <- "DATE"

citi_counts_dec2 <- join(citi_counts_dec,rain_1213,by="DATE",type="left")

citi_counts_dec2$DATE2 <- substr(citi_counts_dec2$DATE,1,nchar(citi_counts_dec2$DATE)-6)
dailyweather_1213 <- read.csv("C:/Data/1351618.csv", stringsAsFactors = FALSE)
```

```
dailyweather_1213$DATE <- str_replace_all(pattern = "-",replacement = "",dailyweather_1213$DATE)

colnames(dailyweather_1213)[3] <- "DATE2"

dailyweather_1213 <- dailyweather_1213[,-c(1:2)]

citi_counts_dec2 <- citi_counts_dec2[,-c(4:5)]

citi_dec <- merge(citi_counts_dec2,dailyweather_1213,by="DATE2")

citi_dec <- citi_dec[,-c(16:21)]
citi_dec <- citi_dec[,-c(7,8,11)]

citi_dec$TMID <- (citi_dec$TMIN+citi_dec$TMAX)/2

citi_dec$HOUR <- str_sub(citi_dec$DATE,10,14)

citi_dec <- citi_dec[,-2]

citi_dec$start.station.id <- as.character(citi_dec$start.station.id)

citi_total <- rbind(citi_june,citi_sept,citi_dec)

## Data Transformation and Results

library(MASS)
library(faraway)

data$dayofweek <- weekdays(as.Date(as.character(data$DATE2),"%Y%m%d"))
data$isPeak <- ifelse(as.numeric(data$HOUR) > 7 & as.numeric(data$HOUR) < 20,1,0)
data$isWeekend <- ifelse(data$dayofweek %in% c("Saturday","Sunday"),1,0)

datanorm$dayofweek <- weekdays(as.Date(as.character(data$DATE2),"%Y%m%d"))
datanorm$isPeak <- ifelse(as.numeric(data$HOUR) > 7 & as.numeric(data$HOUR) < 20,1,0)
datanorm$isWeekend <- ifelse(data$dayofweek %in% c("Saturday","Sunday"),1,0)

modelFull <- lm(n ~ ., data = data)
modelNull <- lm(n ~ 1,data = data)

g.modelFull <- glm(n ~ ., data = data, family=poisson())
g.modelNull <- glm(n ~ 1, data = data, family=poisson())

#anova suggests using full model and eliminating
anova(modelFull,modelNull)

#as does comparing AIC of glm
g.modelFull$aic
g.modelNull$aic

#model1.colChk <- lm(n+10*rnorm(nrow(data)) ~ ., data = data)
#summary(model1.colChk)
```

```
#only take significant predictors (p>0.05)
sig.vars <- c("isPeak","HPCP","TMAX","TMIN","WDF2","WDF5","WSF5","TMID","isWeekend","n")

data2 <- subset(data,select=sig.vars)

model2 <- lm(n ~ .,data=data2)
summary(model2)

#check for colinearity
vif(model2)

#pare down variables to reduce collinearity
data3 <- subset(data2,select=-c(TMAX,TMIN,WDF2))

model3 <- lm(n ~ ., data=data3)
summary(model3)

#use glm
g.model2 <- glm(n ~ .,data=data3,family=poisson())
summary(g.model2)

#compare models
g.modelFull$aic
g.model2$aic

#compare plots
plot(residuals(model3) ~
     fitted(model3),xlab="Fitted",ylab="Residuals",main="Standard Linear Model")

plot(residuals(g.model2) ~
     fitted(g.model2),xlab="Fitted",ylab="Residuals",main="GLM w/ Poisson")

halfnorm(residuals(g.model2))

#check for dispersion
(dp <- sum(residuals(g.model2,type="pearson")^2)/g.model2$df.res)

summary(g.model2,dispersion=dp)

#negative binomial
g.model3 <- glm(n ~ .,negative.binomial(1),data=data3)
AIC(g.model3)

plot(residuals(g.model3) ~
     fitted(g.model3),xlab="Fitted",ylab="Residuals",main="GLM w/ Negative Binomial")

#use standardized variables
datanorm2 <- subset(datanorm,select=colnames(data3))
g.norm.model <- glm(n ~ .,negative.binomial(1),data=datanorm2)
summary(g.norm.model)
#use glm model with negative binomial distribution
```