Predicting Bike Share Ridership based on Weather Data in Seattle

Joey Rodriguez and Daniel Bhatti

2024-11-22

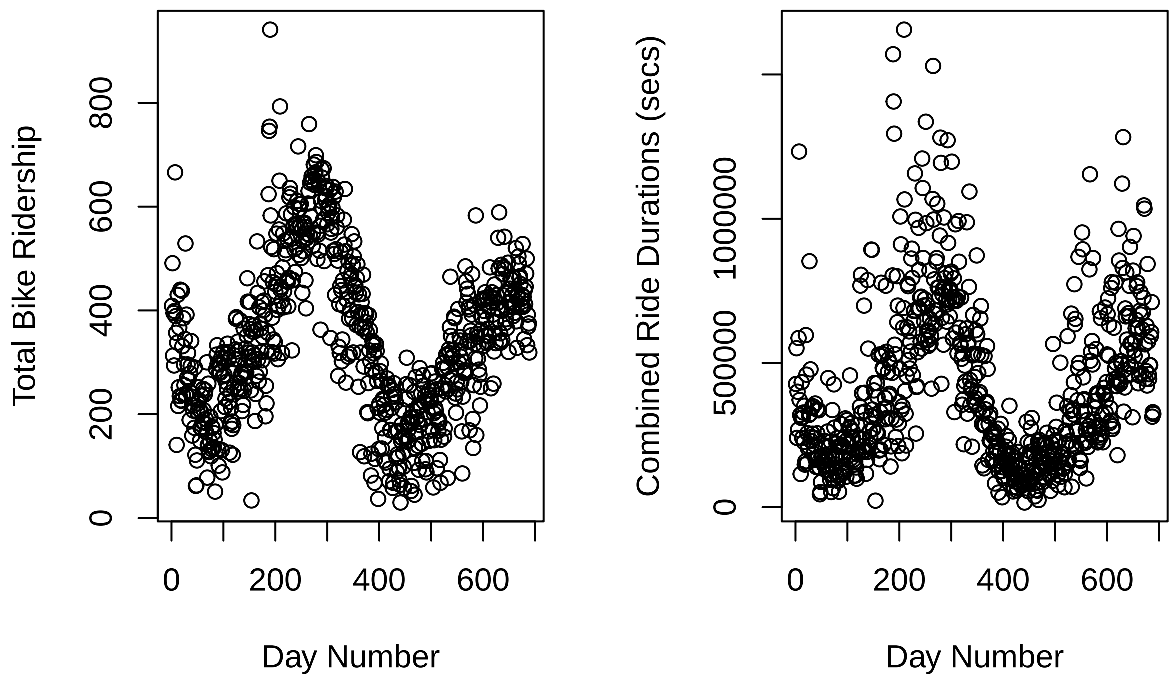
Introduction

A bike share system – or simply bike share – is a service available to residents and tourists of many North American cities. Bike share connects riders with bikes which they can rent and ride from their smartphone. People choose ride share for commuting and for leisure; along with other modes like private car, ride share, mass- and micro- transit, bike share is one option within a suite of transportation options, designed by planners and engineers to get people where they need to go.

The interaction between weather and ridership is intuitive. Favorable weather is marked by sunshine, warm temperatures, moderate humidity, and low wind speed. Humanity's proclivity for the outdoors is highest when the weather is uneventful. When the weather outside is frightful – think freezing temperatures, gusty, and rainy conditions – we prefer the indoors. Especially in the United States with its auto-centric development patterns, poor weather often justifies a "mode-shift'' for those who own a car. When the weather is poor and the infrastructure allows for it, why not drive!?

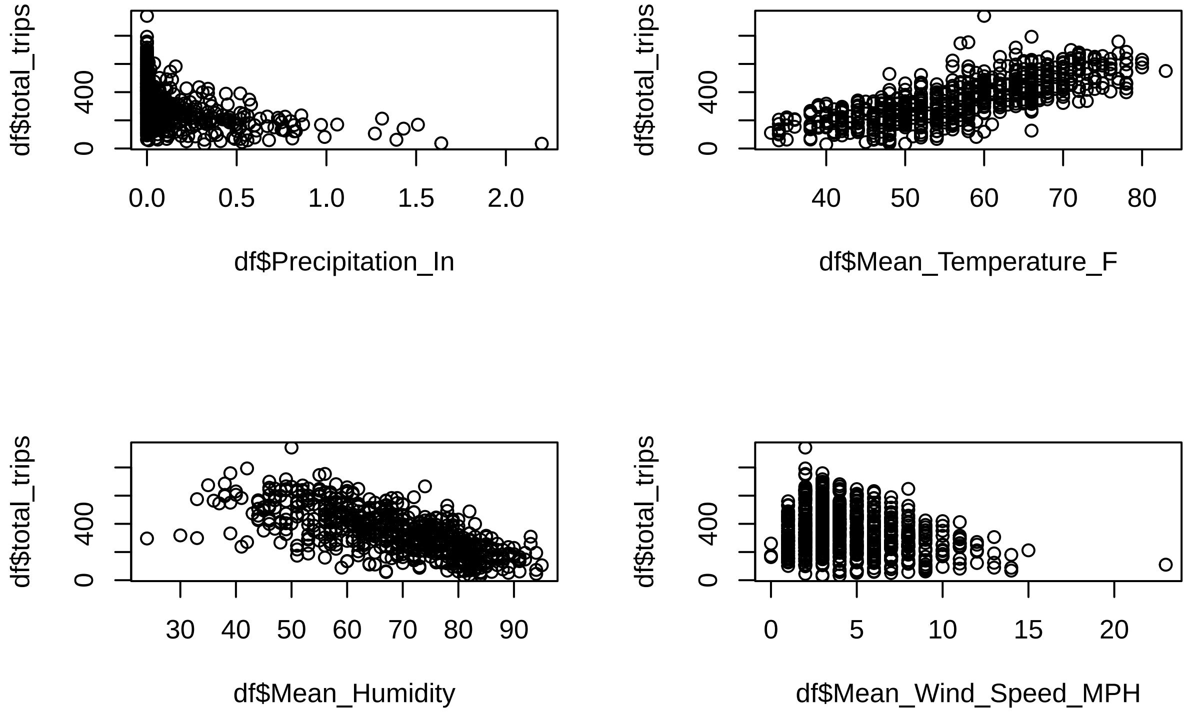
In this paper, we will fit a regression model to several variables describing the weather in order to predict daily bicycle ridership by trip. We will later consider trip duration as our dependent variable. Seattle has a reputation as a rainy city, and it's for good reason. There were 287 rainy days between 10/13/2014 and 08/30/2016, a total of 689 days! Yet bike share was active in Seattle over those 689 days, totaling 236,044 trips across the system.

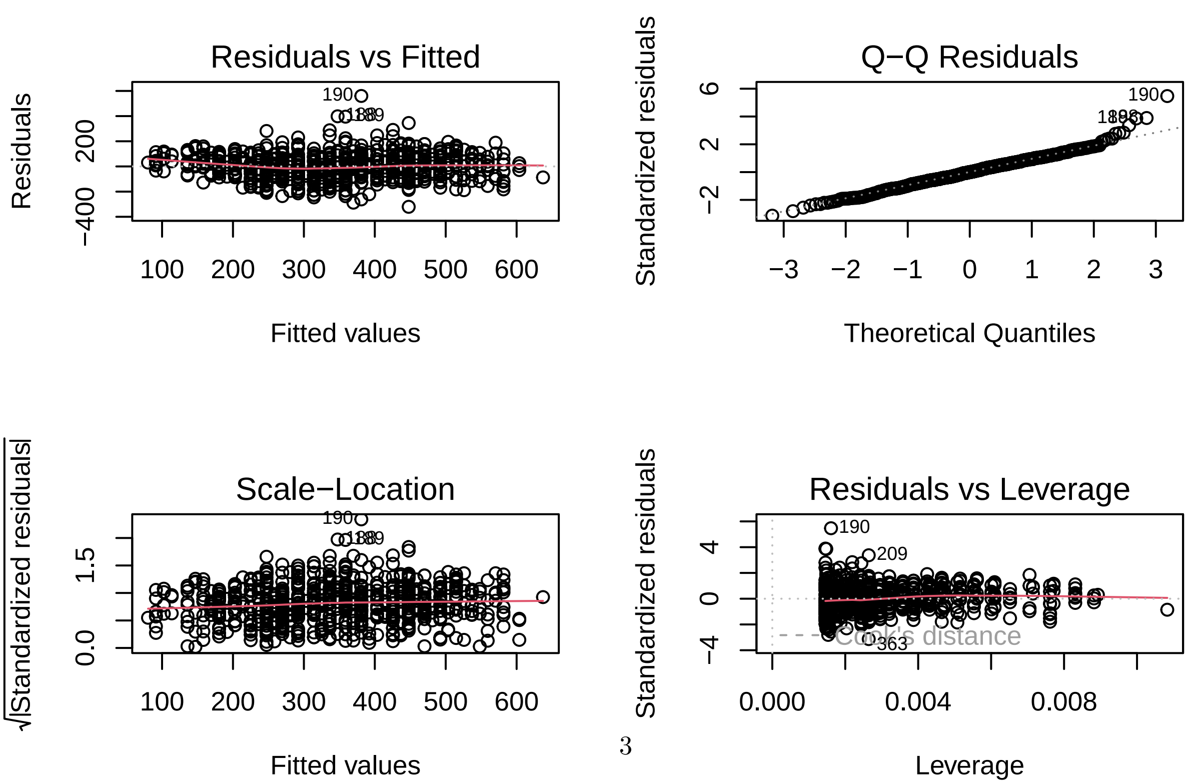
Bike Ridership



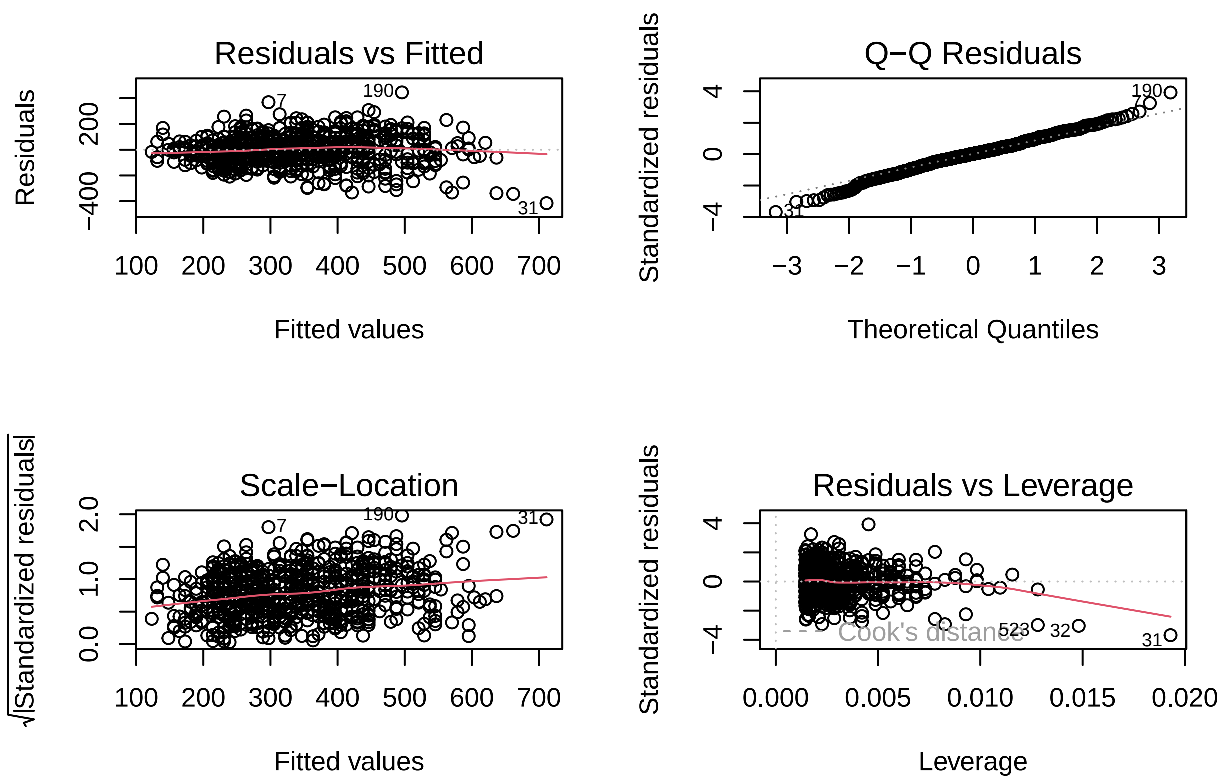
Methods/ Analysis

Best Predictors



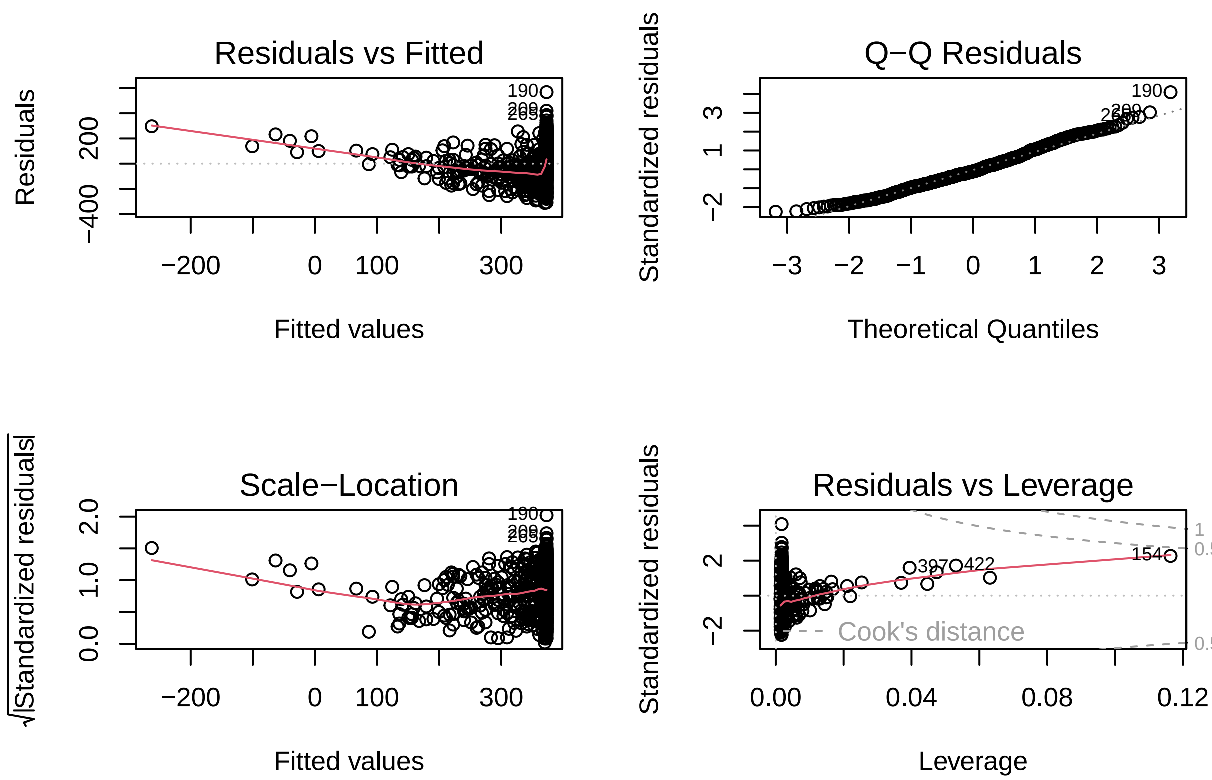


```
##
## Call:
## lm(formula = total_trips ~ Mean_Temperature_F, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -320.82  -64.18   -0.20   66.33  560.02
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -287.4165    21.6101  -13.30   <2e-16 ***
## Mean_Temperature_F   11.1399     0.3756   29.66   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 102.5 on 686 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5618, Adjusted R-squared:  0.5612
## F-statistic: 879.6 on 1 and 686 DF,  p-value: < 2.2e-16
```



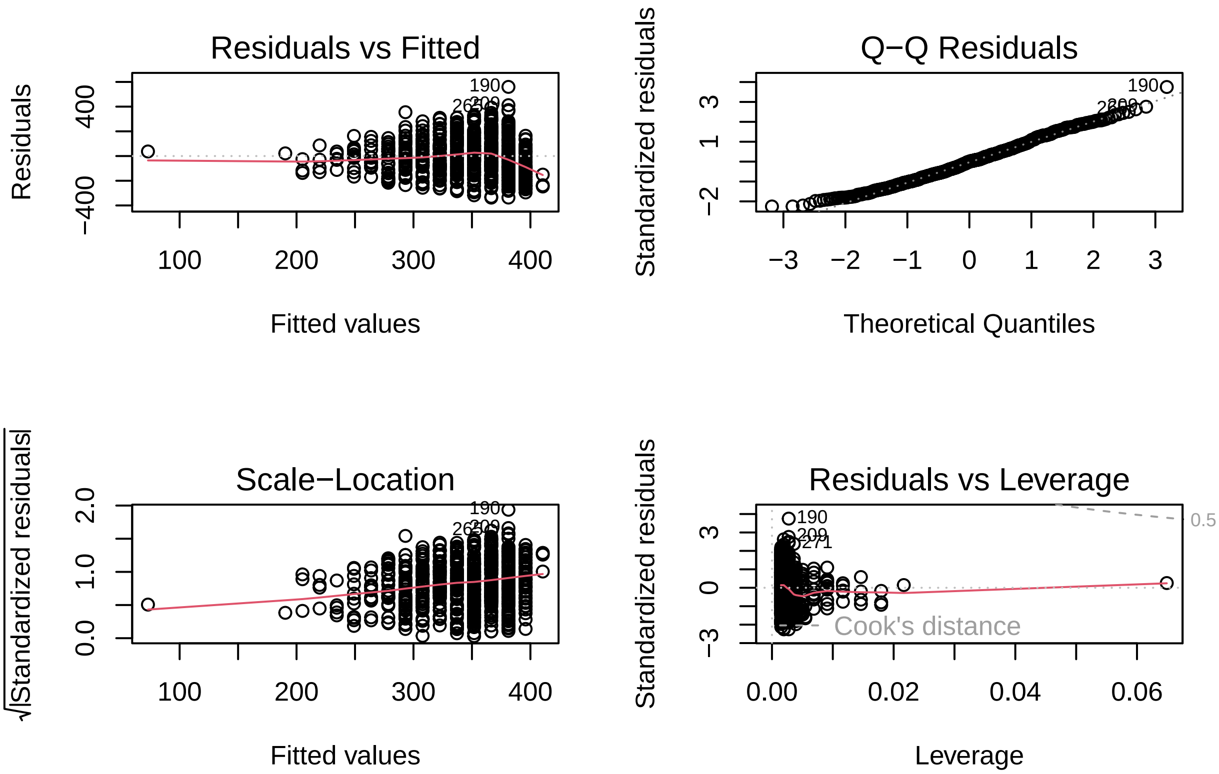

```
##
## Call:
## lm(formula = total_trips ~ Mean_Humidity, data = df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -415.28  -63.52    1.48   67.50  445.10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   910.0991    23.7690   38.29   <2e-16 ***
## Mean_Humidity  -8.2840     0.3412  -24.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 113.7 on 687 degrees of freedom
## Multiple R-squared:  0.4619, Adjusted R-squared:  0.4611
## F-statistic: 589.6 on 1 and 687 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = total_trips ~ Precipitation_In, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -312.06  -93.95  -14.95   86.05  568.05
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       372.947      5.805   64.24   <2e-16 ***
```

```
## Precipitation_In -288.941     22.514  -12.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 139.2 on 687 degrees of freedom
## Multiple R-squared:  0.1934, Adjusted R-squared:  0.1922
## F-statistic: 164.7 on 1 and 687 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = total_trips ~ Mean_Wind_Speed_MPH, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -336.54 -108.14    0.14   98.14  559.78
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          410.581     11.069  37.094  < 2e-16 ***
## Mean_Wind_Speed_MPH  -14.681      2.049  -7.163 2.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.5 on 687 degrees of freedom
## Multiple R-squared:  0.0695, Adjusted R-squared:  0.06815
## F-statistic: 51.31 on 1 and 687 DF,  p-value: 2.034e-12
```

```r
rainResiduals <- resid(outRain)
windResiduals <- resid(outWind)

variance_rain = lm(abs(windResiduals) ~ Mean_Wind_Speed_MPH, data = df)

rainpredictedvar = predict(variance_rain)

rainweights = 1/(rainpredictedvar^2)

RainWLS <- lm(total_trips ~ Precipitation_In, data = df, weights = rainweights)

#RainWLS did not produce meaningful improvements.

outDew = lm(total_trips ~ MeanDew_Point_F, data = df)

#Great plots R^2 is only 0.2

outWind = lm(total_trips ~ Mean_Wind_Speed_MPH, data = df)

#Good plots R^2 is a paltry 0.07

windweights <- 1/lm(abs(outWind$residuals)~outWind$fitted.values)$fitted.values^2

WindWLS <- lm(total_trips ~ Mean_Wind_Speed_MPH, data = df, weights = windweights)

#Using WLS on windspeed significantly improves it. The R^2 is 0.7

outRange = lm(total_trips ~ temp_range, data = df)

#Plots have 1 extremely influential point, the R^2 is 0.315
#If we are to use mean temp I think we can ignore this variable (as a basic regressor)

outVisible = lm(total_trips ~ Mean_Visibility_Miles, data = df)

#Plots are mid, R^2 is 0.1313

visresid =  resid(outVisible)

variance_vis = lm(abs(visresid)~Mean_Visibility_Miles, data=df)

vispredictvar = predict(variance_vis)

visweights = 1/(vispredictvar^2)

VisWLS <- lm(total_trips ~ Mean_Visibility_Miles, data = df, weights = visweights)

#Very unimpressive.

outSea = lm(total_trips ~ Mean_Sea_Level_Pressure_In, data = df)

#Criteria for being in full model is that it had a *** significance by itself

fullmodel = lm(total_trips ~ Mean_Temperature_F + Mean_Humidity+MeanDew_Point_F+Precipitation_In+Mean_Wi
```

```r
#Very interestingly Mean temp is not significant. Additionally visibility is far from significant

#Diagnostic plots look very good. R^2 is 0.71

partialmodel = lm(total_trips ~ Mean_Temperature_F + Mean_Humidity+MeanDew_Point_F+Precipitation_In+Mean

partialmodel2 = lm(total_trips ~ Mean_Humidity+MeanDew_Point_F+Precipitation_In+Mean_Wind_Speed_MPH, da

#Good diagnostics, R^2 is 0.71, all regressors are significant.

#I think that this is the best model.

testmodel = lm(total_trips ~ Mean_Temperature_F + MeanDew_Point_F, data=df)

#R^2 0.63, good diagnostics.

testmodel2 = lm(total_trips ~ Mean_Temperature_F + Mean_Humidity, data=df)

#R^2 0.6487, good diagnostics

testmodel3 = lm(total_trips ~ Mean_Temperature_F + Mean_Humidity+MeanDew_Point_F, data=df)

#Suddenly temperature is not significant


#powerTransform(cbind(df$total_trips,df$Mean_Temperature_F,df$#Mean_Humidity,df$MeanDew_Point_F,df$Prec

#This fails as powerTransform needs arguments to be strictly positive and the min
#of Precipitation and Wind speed are 0

powerTransform(cbind(df$total_trips,df$Mean_Temperature_F,df$Mean_Humidity,df$MeanDew_Point_F,df$Mean_Vi
```

```
## Estimated transformation parameters
##         Y1         Y2         Y3         Y4         Y5
##  0.7608250  0.7602633  0.6698861  1.1004605 10.6680860
```

```r
#trips: 0.761, Temperature: 0.760, Humidity: 0.67, Dew point 1.1, Visibility 10

#Therefore trips 3/4, temperate 3/4, humidity 2/3, Dew point no change visibility, visibility^2

df$total_trips_trans <- df$total_trips^(3/4)
df$Mean_Temperature_F_trans <- df$Mean_Temperature_F^(3/4)
df$Mean_Humidity_trans <- df$Mean_Humidity^(2/3)
df$Mean_Visibility_Miles_trans <- df$Mean_Visibility_Miles^2

df$total_trips_trans <- df$total_trips^(3/4)
df$Mean_Temperature_F_trans <- df$Mean_Temperature_F^(3/4)
df$Mean_Humidity_trans <- df$Mean_Humidity^(2/3)
df$Mean_Visibility_Miles_trans <- df$Mean_Visibility_Miles^10

transform_out <- lm(total_trips_trans ~ Mean_Temperature_F_trans + Mean_Humidity_trans + MeanDew_Point_

#The transformed model is not very impressive. Good diagnostics, R^2 of 0.6484
```

Conclusion/Discussion

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.