Sta 325 Final Project

Calleigh Smith, Hannah Bogomilsky, Hugh Esterson, Maria Henriquez, Mariana Izon

11/22/2020

```r
library(readr)
library(dplyr)
library(tidyverse)
library(gridExtra)
library(mgcv)
library(patchwork)
```

```r
flights <- read_csv("data/flights.csv")

unique(flights$OP_CARRIER)
```

```
## [1] "AA" "DL" "B6" "AS"
```

```r
unique(flights$DEST)
```

```
##  [1] "LAX" "SFO" "SJC" "SAN" "PSP" "SMF" "OAK" "LGB" "ONT" "BUR"
```

```r
class(flights$CARRIER_DELAY)
```

```
## [1] "numeric"
```

```r
flights <- flights %>%
  mutate(CARRIER_DELAY = case_when(CARRIER_DELAY > 0 ~ 1,
                                   TRUE ~ 0),
         WEATHER_DELAY = case_when(WEATHER_DELAY > 0 ~ 1,
                                   TRUE ~ 0),
         NAS_DELAY = case_when(NAS_DELAY > 0 ~ 1,
                               TRUE ~ 0),
         SECURITY_DELAY = case_when(SECURITY_DELAY > 0 ~ 1,
                                    TRUE ~ 0),
         LATE_AIRCRAFT_DELAY = case_when(LATE_AIRCRAFT_DELAY > 0 ~ 1,
                                         TRUE ~ 0))
```

```r
flights
```

```
## # A tibble: 2,044 x 34
##     YEAR MONTH DAY_OF_MONTH DAY_OF_WEEK FL_DATE    OP_CARRIER TAIL_NUM
##    <dbl> <dbl>        <dbl>       <dbl> <date>     <chr>      <chr>
##  1  2020     1            1           3 2020-01-01 AA         N110AN
##  2  2020     1            2           4 2020-01-02 AA         N111ZM
##  3  2020     1            3           5 2020-01-03 AA         N108NN
##  4  2020     1            4           6 2020-01-04 AA         N102NN
##  5  2020     1            5           7 2020-01-05 AA         N113AN
##  6  2020     1            6           1 2020-01-06 AA         N103NN
##  7  2020     1            7           2 2020-01-07 AA         N113AN
```

```
## 8  2020     1           8           3 2020-01-08 AA         N106NN
## 9  2020     1           9           4 2020-01-09 AA         N102NN
## 10 2020     1          10           5 2020-01-10 AA         N117AN
## # ... with 2,034 more rows, and 27 more variables: OP_CARRIER_FL_NUM <dbl>,
## #   ORIGIN <chr>, ORIGIN_CITY_NAME <chr>, DEST <chr>, DEST_CITY_NAME <chr>,
## #   CRS_DEP_TIME <dbl>, DEP_TIME <dbl>, DEP_DELAY <dbl>, TAXI_OUT <dbl>,
## #   WHEELS_OFF <dbl>, WHEELS_ON <dbl>, TAXI_IN <dbl>, CRS_ARR_TIME <dbl>,
## #   ARR_TIME <dbl>, ARR_DELAY <dbl>, CANCELLED <dbl>, CANCELLATION_CODE <lgl>,
## #   DIVERTED <dbl>, CRS_ELAPSED_TIME <dbl>, ACTUAL_ELAPSED_TIME <dbl>,
## #   AIR_TIME <dbl>, DISTANCE <dbl>, CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>,
## #   NAS_DELAY <dbl>, SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>
```

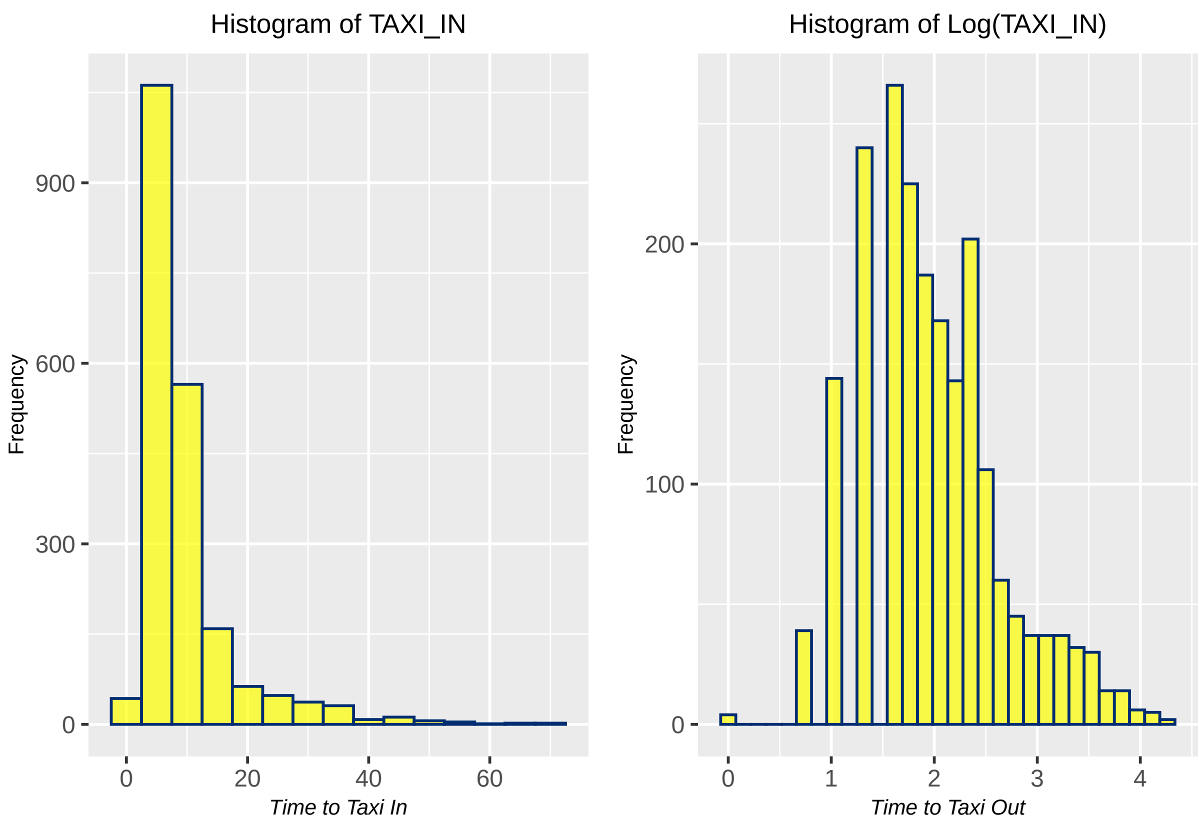
INDIVIDUAL PREDICTORS

Taxi Histograms

```r
pTAXI_IN <- ggplot(data = flights, aes(x = TAXI_IN)) +
  geom_histogram(binwidth = 5, fill = "#FFFF00", color = "#002D72", alpha = .7) +
  labs(x = "Time to Taxi In",
       y = "Frequency",
       title = "Histogram of TAXI_IN") +
  theme(plot.title = element_text(size = 10,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))



 # ggplot(train_data, mapping = aes(x = St2)) +
 #   geom_histogram(binwidth =2.5, fill = "#FFFF00", color = "#002D72", alpha = .7) +
 #   labs(x = xlab(bquote('St^2')),
 #       # xlab(bquote('Assimilation ('*mu~ 'mol' ~CO[2]~ m^-2~s^-1*')')),
 #        y = "Frequency",
 #        title = "Histogram of Stokes Number, Squared") +
 #   theme(plot.title = element_text(size = 10,hjust = 0.5),
 #         plot.subtitle = element_text(hjust = 0.5),
 #         axis.title.x.bottom = element_text(size = 8, face = "italic"),
 #         axis.title.y.left = element_text(size = 8))


pTAXI_OUT <- ggplot(data = flights, aes(x = TAXI_OUT)) +
  geom_histogram(binwidth = 5, fill = "#FFFF00", color = "#002D72", alpha = .7) +
  labs(x = "Time to Taxi Out",
       y = "Frequency",
       title = "Histogram of TAXI_OUT") +
  theme(plot.title = element_text(size = 10,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))


flights$log_TAXI_OUT <- log(flights$TAXI_OUT)
flights$log_TAXI_IN <- log(flights$TAXI_IN)
```

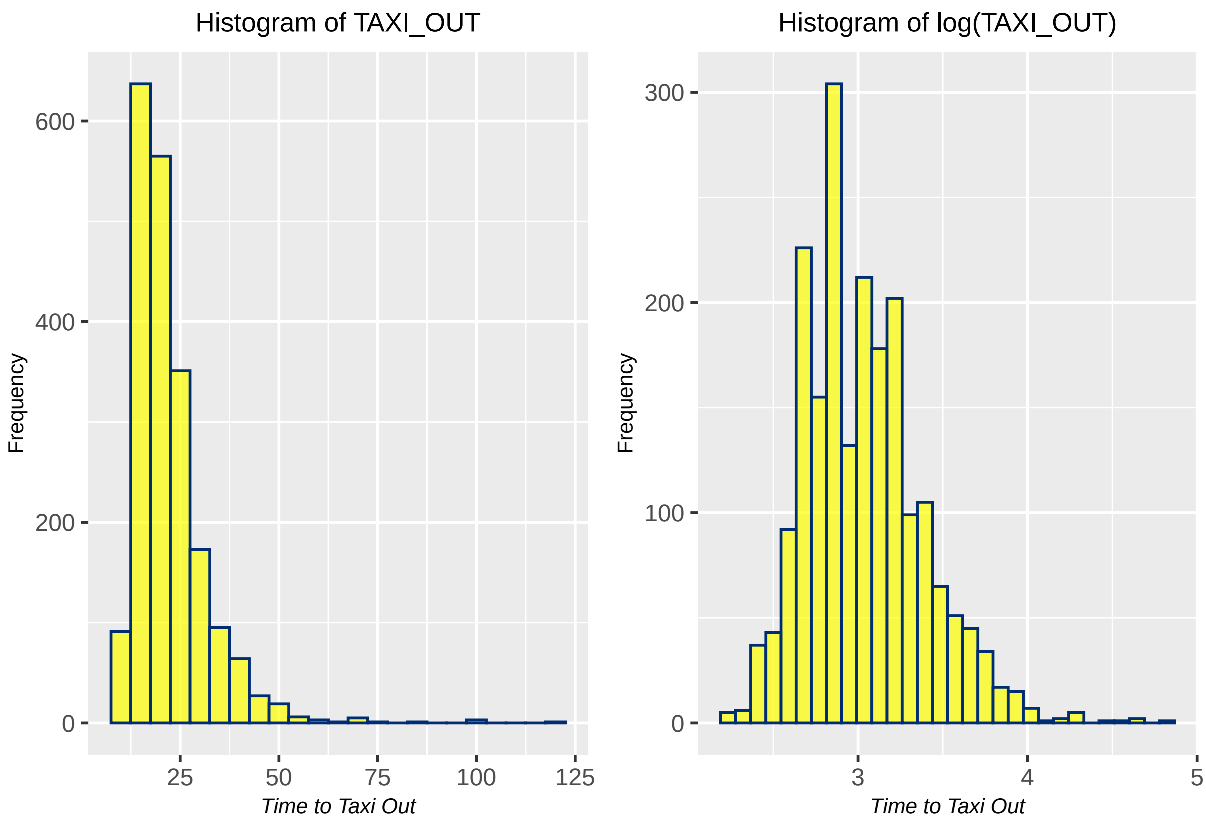
```r
plog_TAXI_OUT <- ggplot(data = flights, aes(x = log_TAXI_OUT)) +
  geom_histogram(fill = "#FFFF00", color = "#002D72", alpha = .7) +
  labs(x = "Time to Taxi Out",
       y = "Frequency",
       title = "Histogram of log(TAXI_OUT)") +
  theme(plot.title = element_text(size = 10,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))


plog_TAXI_IN <- ggplot(data = flights, aes(x = log_TAXI_IN)) +
  geom_histogram(fill = "#FFFF00", color = "#002D72", alpha = .7) +
  labs(x = "Time to Taxi Out",
       y = "Frequency",
       title = "Histogram of Log(TAXI_IN)") +
  theme(plot.title = element_text(size = 10,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))

pTAXI_IN + plog_TAXI_IN
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



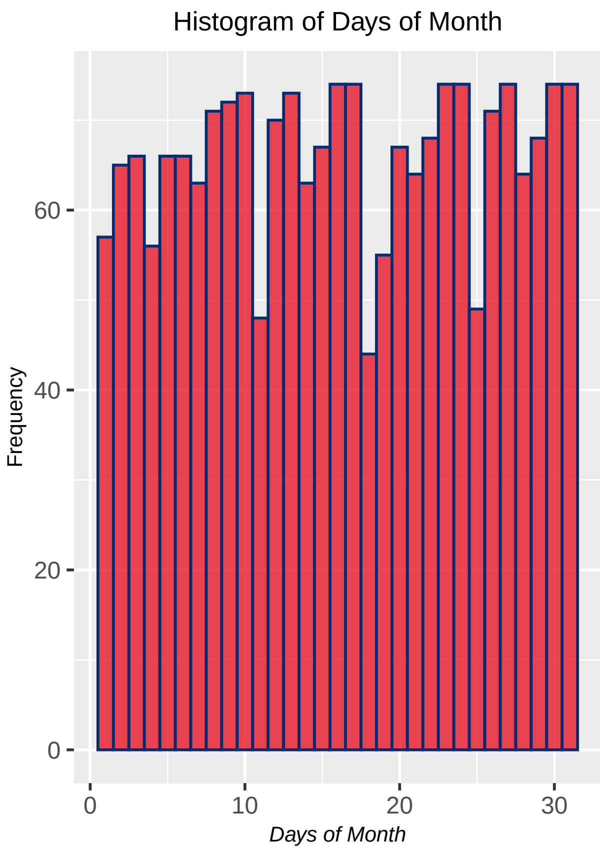

```r
pTAXI_OUT + plog_TAXI_OUT
```

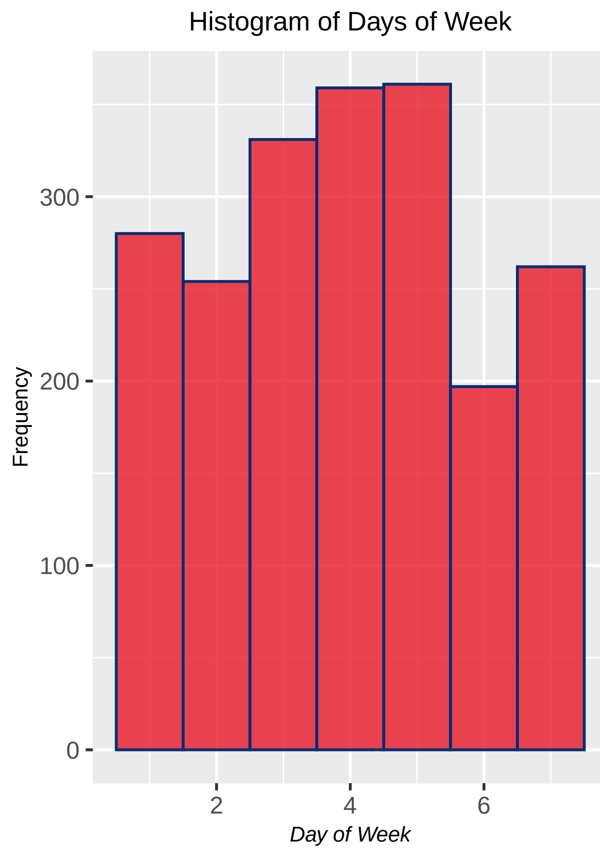
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Days of Month and Week

```
p02 <- ggplot(data = flights, aes(x = DAY_OF_MONTH)) +
  geom_histogram(binwidth = 1, fill = "#E81828", color = "#002D72", alpha = .8) +
  labs(x = "Days of Month",
       y = "Frequency",
       title = "Histogram of Days of Month") +
    theme(plot.title = element_text(size = 10,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))

p03 <- ggplot(data = flights, aes(x = DAY_OF_WEEK)) +
  geom_histogram(binwidth = 1, fill = "#E81828", color = "#002D72", alpha = .8) +
  labs(x = "Day of Week",
       y = "Frequency",
       title = "Histogram of Days of Week") +
    theme(plot.title = element_text(size = 10,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))

grid.arrange(p02, p03, nrow = 1)
```

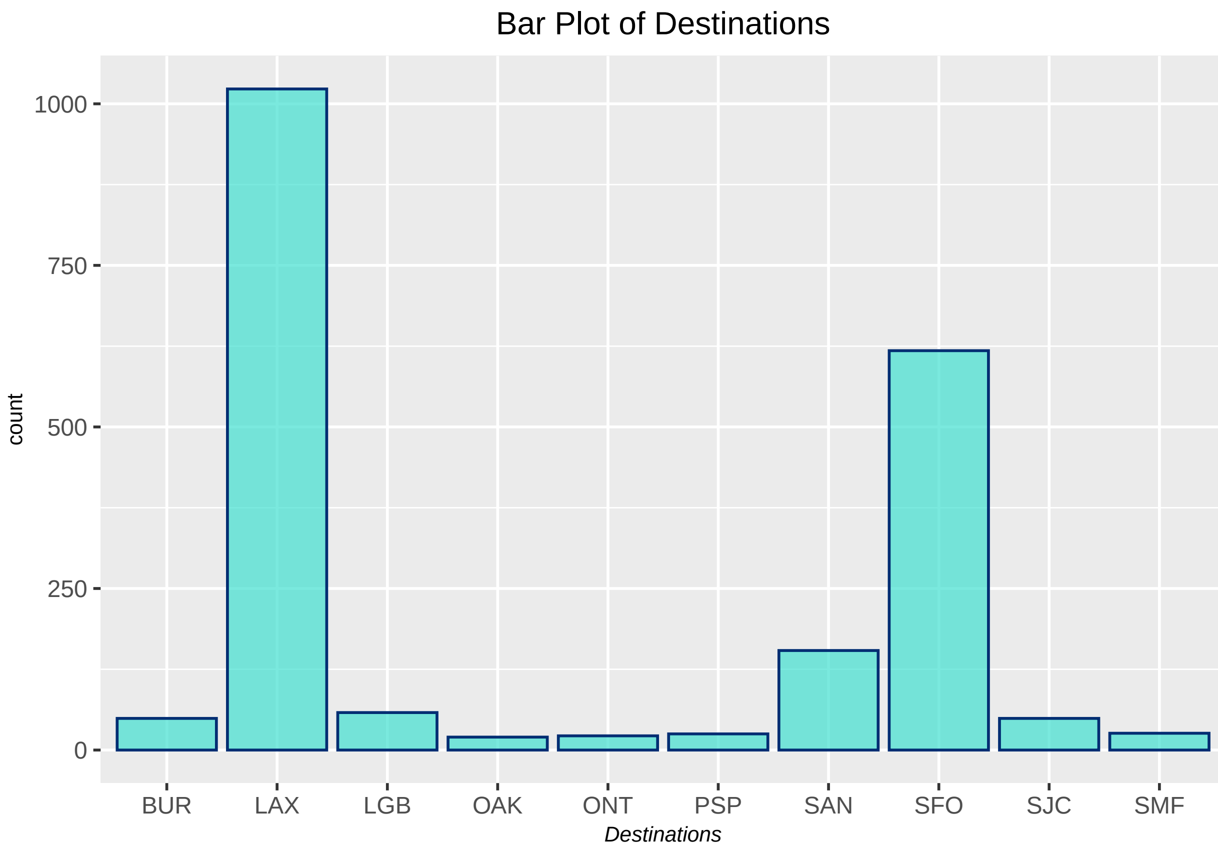




Destination Locations

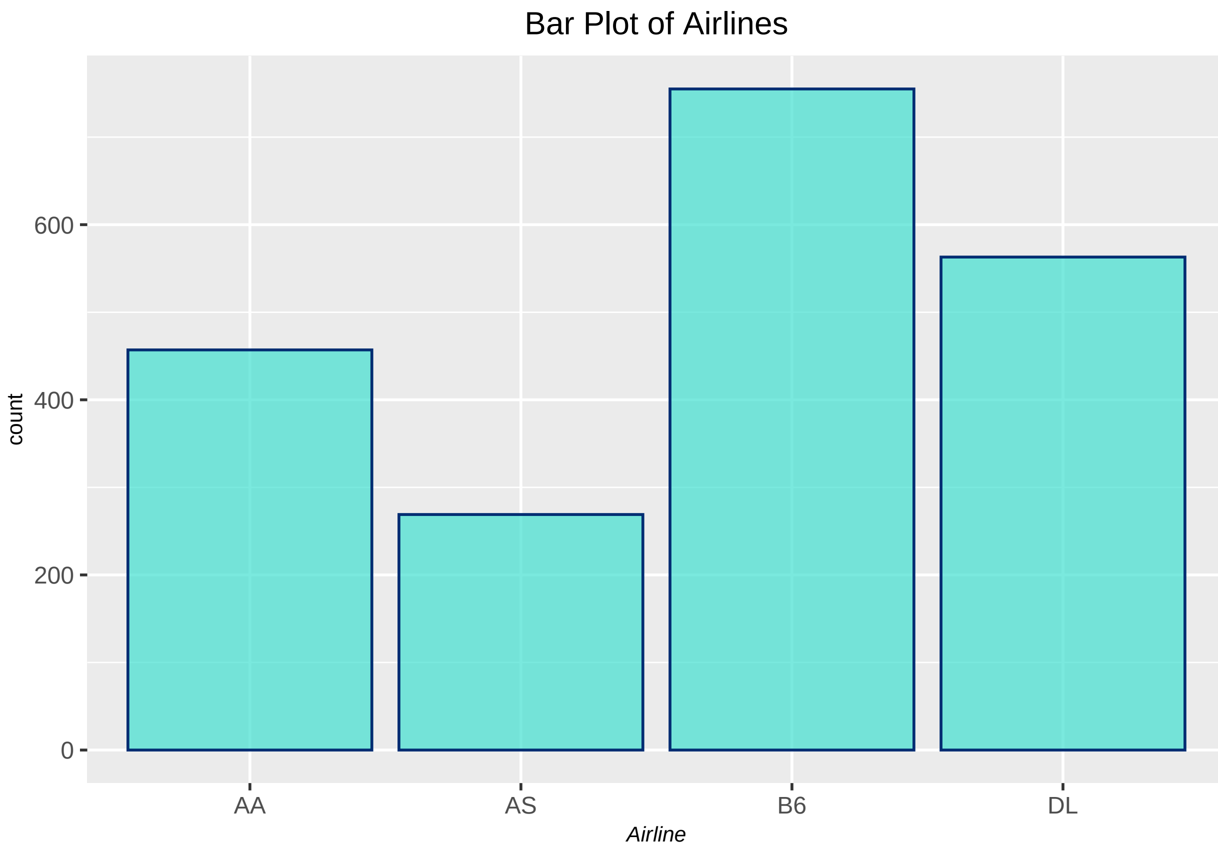
Origin is all JFK, but we could consider the different destination locations.

```
ggplot(data = flights, aes(x = DEST)) +
  geom_bar(fill = "#40E0D0", color = "#002D72", alpha = .7) +
  labs(x = "Destinations",
       title = "Bar Plot of Destinations") +
  theme(plot.title = element_text(size = 12,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))
```

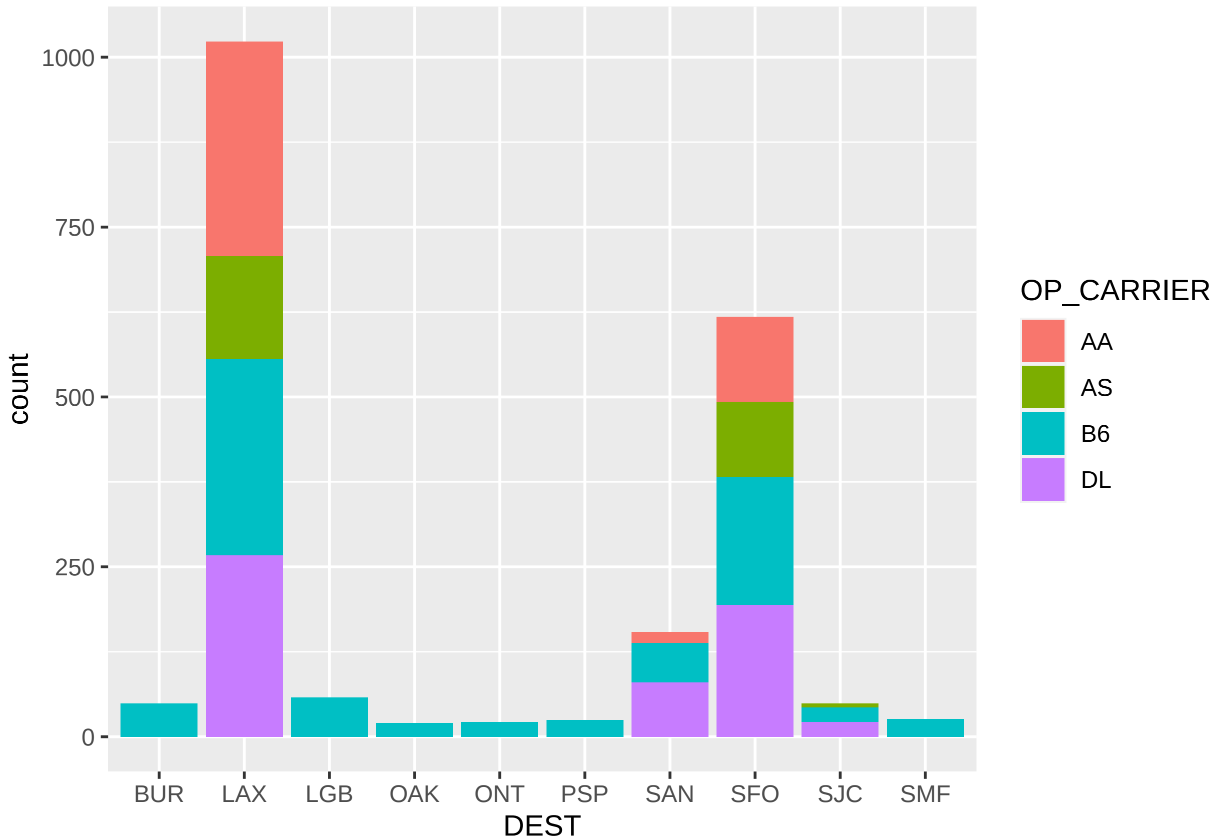


Airlines

```
ggplot(data = flights, aes(x = OP_CARRIER)) +
  geom_bar(fill = "#40E0D0", color = "#002D72", alpha = .7) +
  labs(x = "Airline",
       title = "Bar Plot of Airlines") +
  theme(plot.title = element_text(size = 12,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))
```

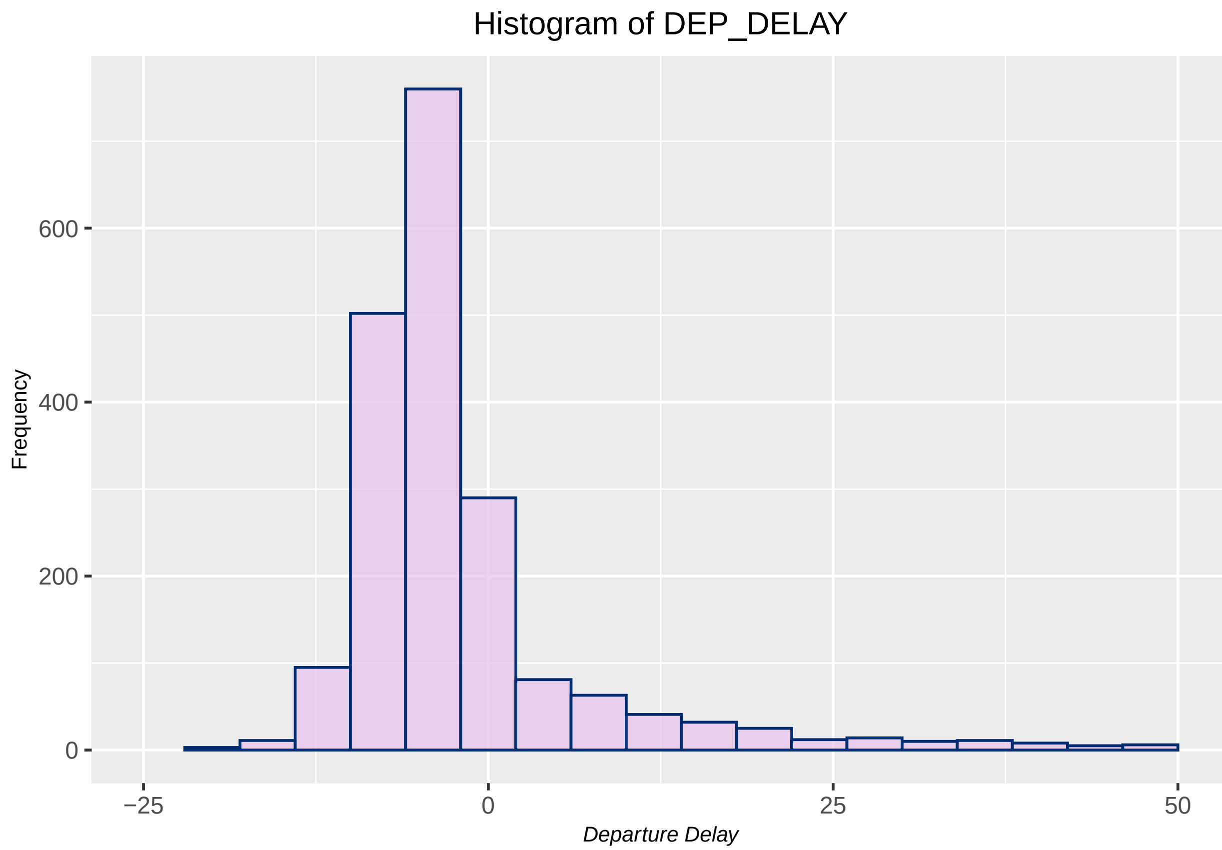



```r
ggplot(data = flights, aes(x = DEST, fill = OP_CARRIER)) +
  geom_bar()
```

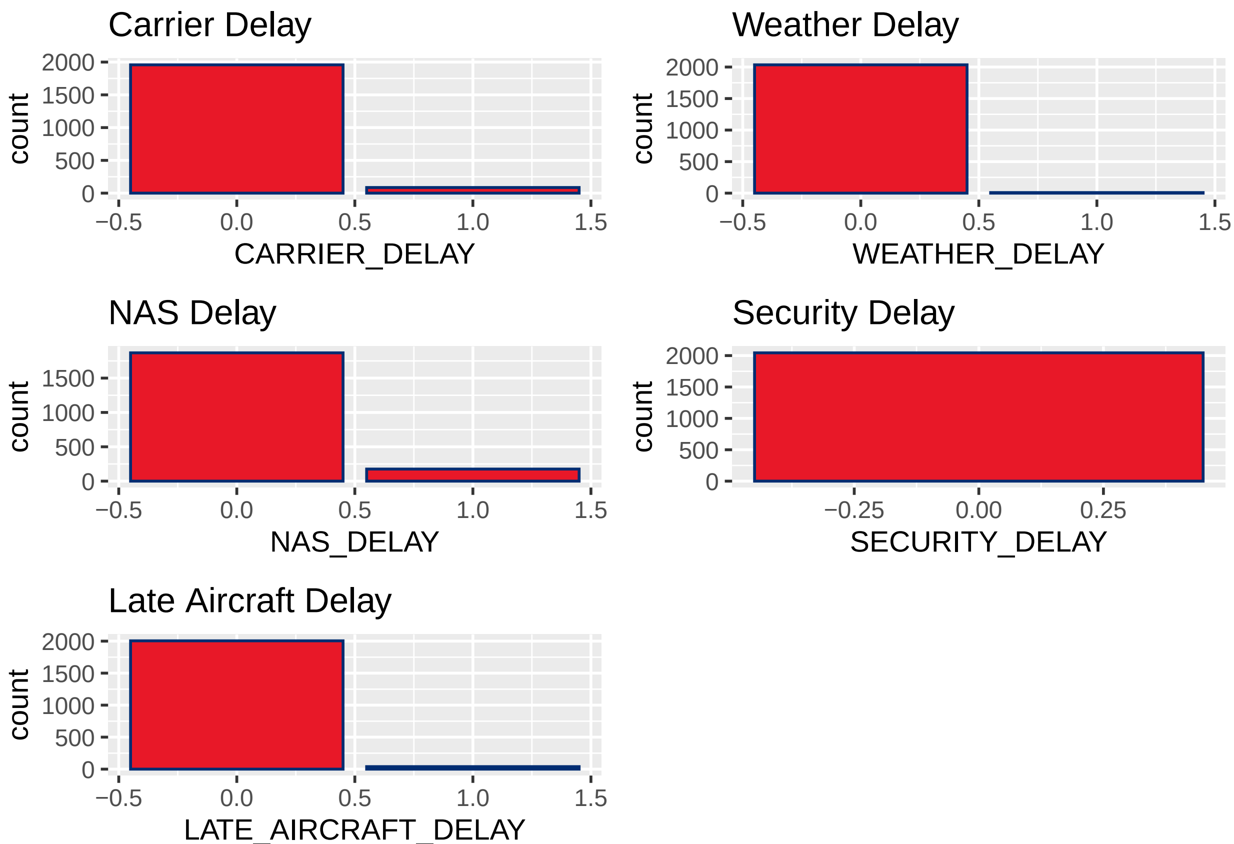


Depart Delay Histogram

```
ggplot(data = flights, aes(x = DEP_DELAY)) +
  geom_histogram(binwidth = 4, fill = "#e9c2ed", color = "#002D72", alpha = 0.7) +
  xlim(-25, 50) +
  labs(x = "Departure Delay",
       y = "Frequency",
       title = "Histogram of DEP_DELAY") +
  theme(plot.title = element_text(size = 12,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))
```




```
p1 <- ggplot(data = flights, aes(x = CARRIER_DELAY)) +
  geom_bar(fill = "#E81828", color = "#002D72") +
  labs(title = "Carrier Delay")

p2 <- ggplot(data = flights, aes(x = WEATHER_DELAY)) +
  geom_bar(fill = "#E81828", color = "#002D72") +
  labs(title = "Weather Delay")

p3 <- ggplot(data = flights, aes(x = NAS_DELAY)) +
  geom_bar(fill = "#E81828", color = "#002D72") +
  labs(title = "NAS Delay")

p4 <- ggplot(data = flights, aes(x = SECURITY_DELAY)) +
  geom_bar(fill = "#E81828", color = "#002D72") +
  labs(title = "Security Delay")

p5 <- ggplot(data = flights, aes(x = LATE_AIRCRAFT_DELAY)) +
  geom_bar(fill = "#E81828", color = "#002D72") +
  labs(title = "Late Aircraft Delay")

grid.arrange(p1,p2,p3,p4,p5, nrow = 3)
```



From this EDA of the categorical variables, we probably should not perform analysis with `SECURITY_DELAY` since all of them are classified as 0.

```r
flights %>%
  count(WEATHER_DELAY)
```

```
## # A tibble: 2 x 2
##   WEATHER_DELAY     n
##           <dbl> <int>
## 1             0  2035
## 2             1     9
```

Furthermore, only 9 flights are classified with a weather delay, so it may not be good for our model to include this as a variable for right now.

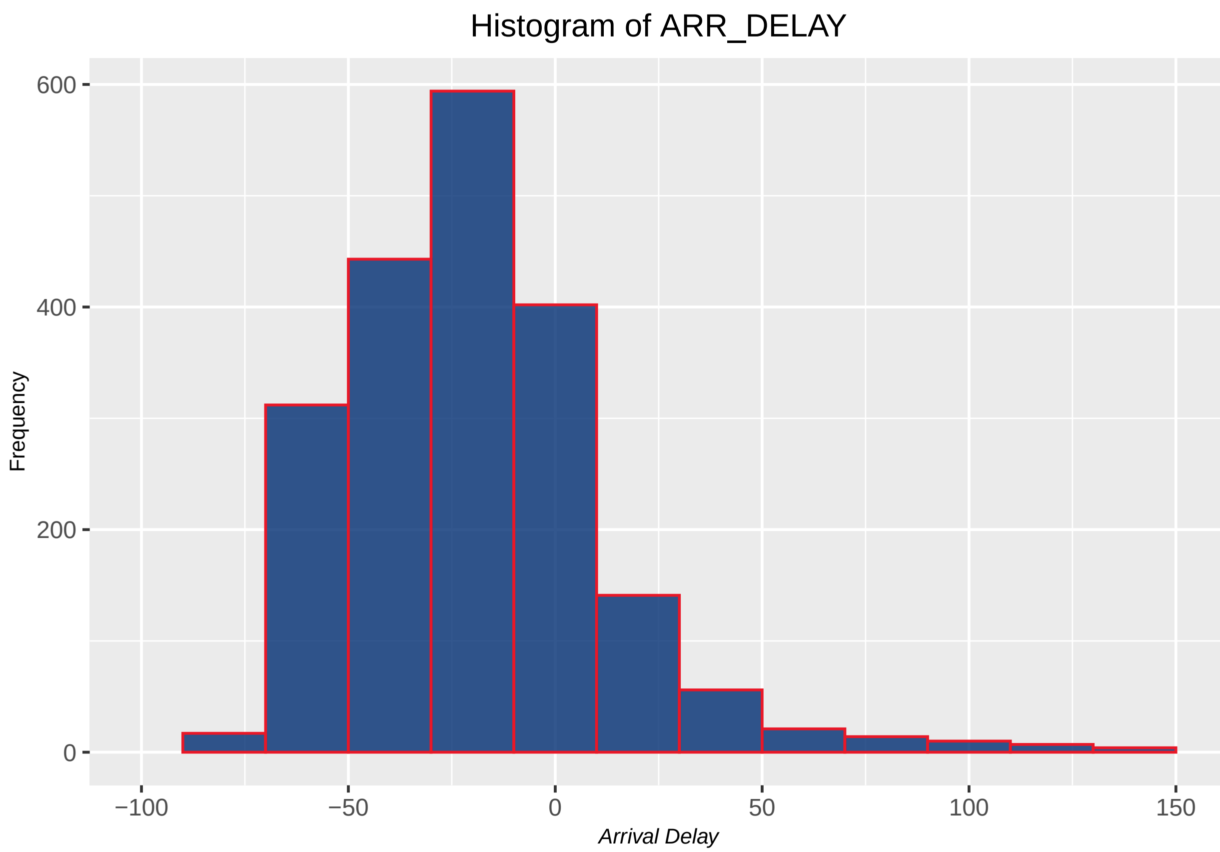
Overall, the categorical delay predictors I would think we could use are: Carrier Delay, NAS Delay, and Late Aircraft Delay

RESPONSE VARIABLE: ARRIVAL DELAY TIME

I just made it a different color so that when I scroll up to look at distributions I can easily tell the response from predictors (definitely can change at the end).

```r
ggplot(data = flights, aes(x = ARR_DELAY)) +
  geom_histogram(binwidth = 20, fill = "#002D72", color = "#E81828", alpha = 0.8) +
  xlim(-100, 150) +
  labs(x = "Arrival Delay",
       y = "Frequency",
       title = "Histogram of ARR_DELAY") +
```

```r
  theme(plot.title = element_text(size = 12,hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        axis.title.x.bottom = element_text(size = 8, face = "italic"),
        axis.title.y.left = element_text(size = 8))
```

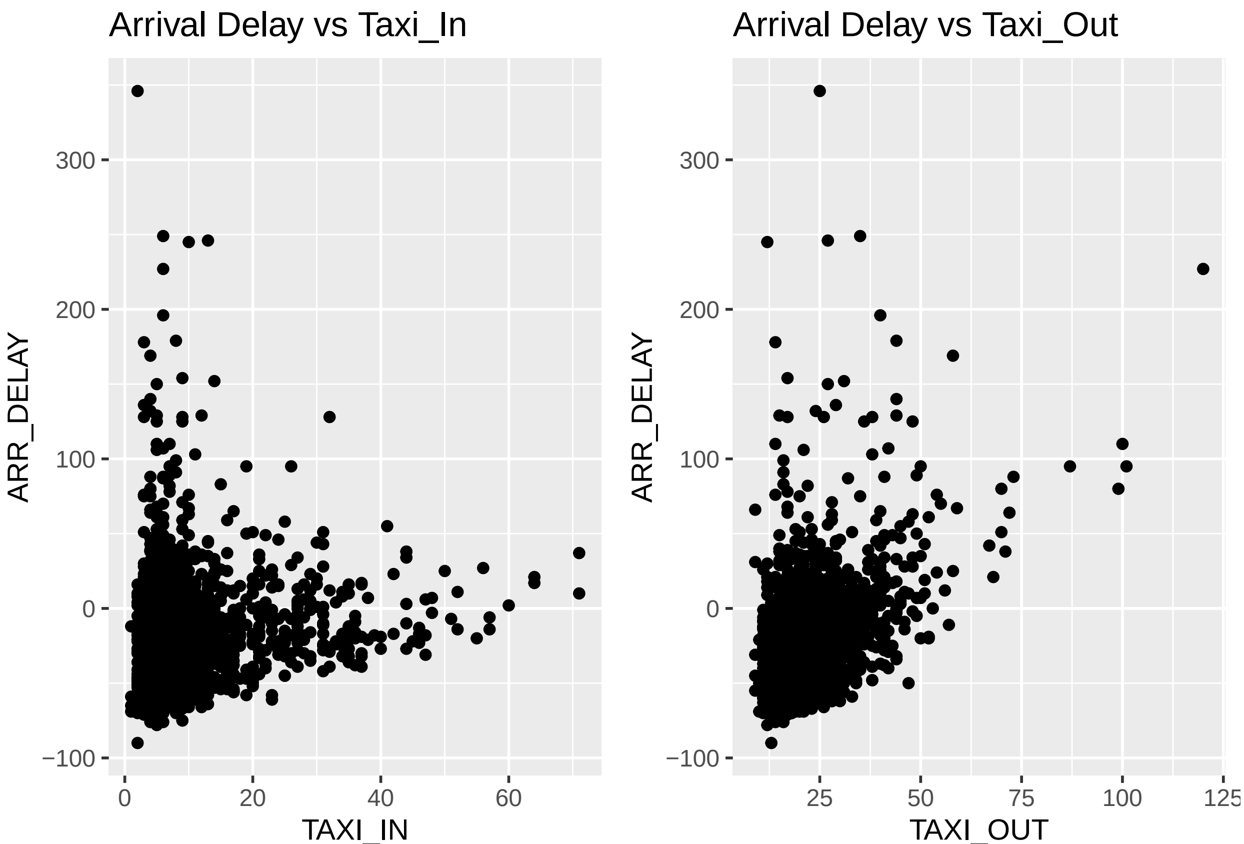


PREDICTORS VS RESPONSE

ARR_DELAY and TAXI_IN / TAXI_OUT

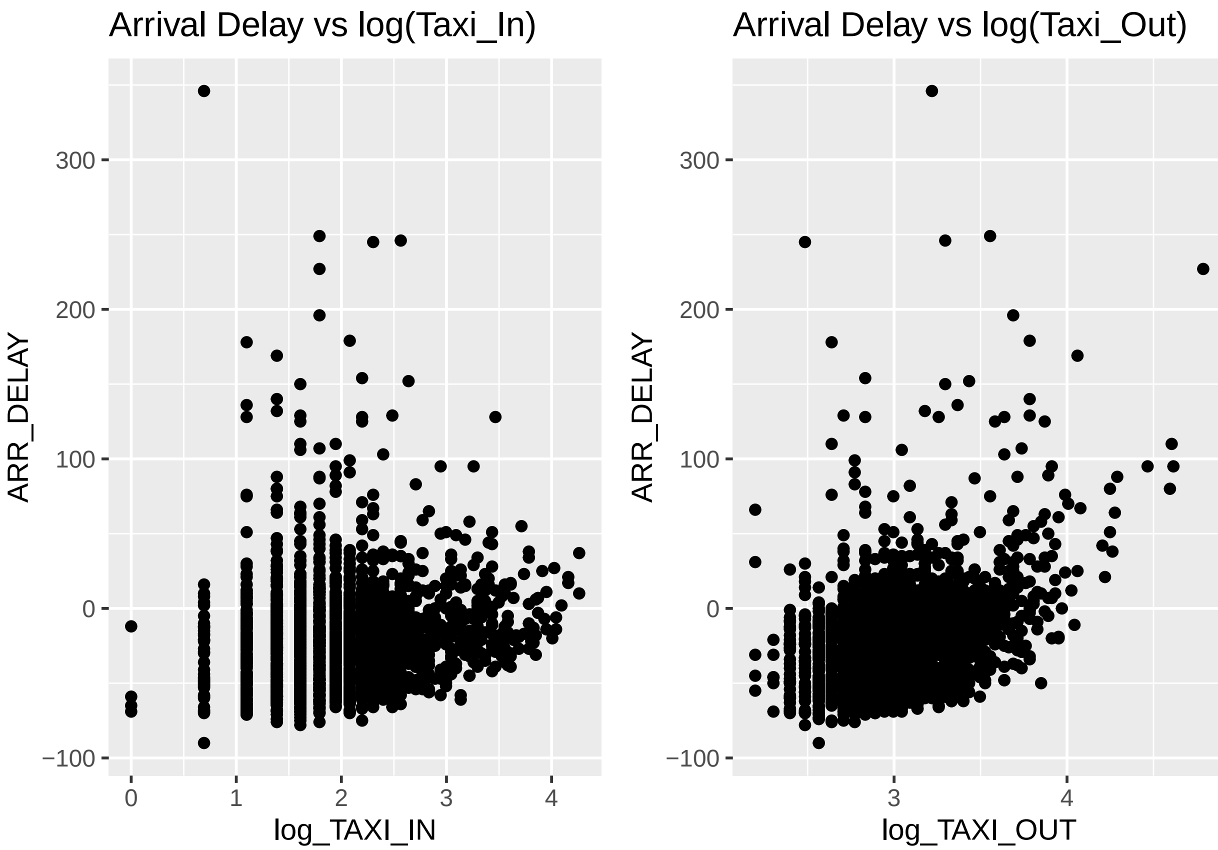
```r
p6 <- ggplot(data = flights, aes(y = ARR_DELAY, x = TAXI_IN)) +
  geom_point() +
  labs(title = "Arrival Delay vs Taxi_In")

p7 <- ggplot(data = flights, aes(y = ARR_DELAY, x = TAXI_OUT)) +
  geom_point() +
  labs(title = "Arrival Delay vs Taxi_Out")

grid.arrange(p6,p7, nrow = 1)
```

```
## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).
```




```
plog6 <- ggplot(data = flights, aes(y = ARR_DELAY, x = log_TAXI_IN)) +
  geom_point() +
  labs(title = "Arrival Delay vs log(Taxi_In)")

plog7 <- ggplot(data = flights, aes(y = ARR_DELAY, x = log_TAXI_OUT)) +
  geom_point() +
  labs(title = "Arrival Delay vs log(Taxi_Out)")
grid.arrange(plog6,plog7, nrow = 1)
```

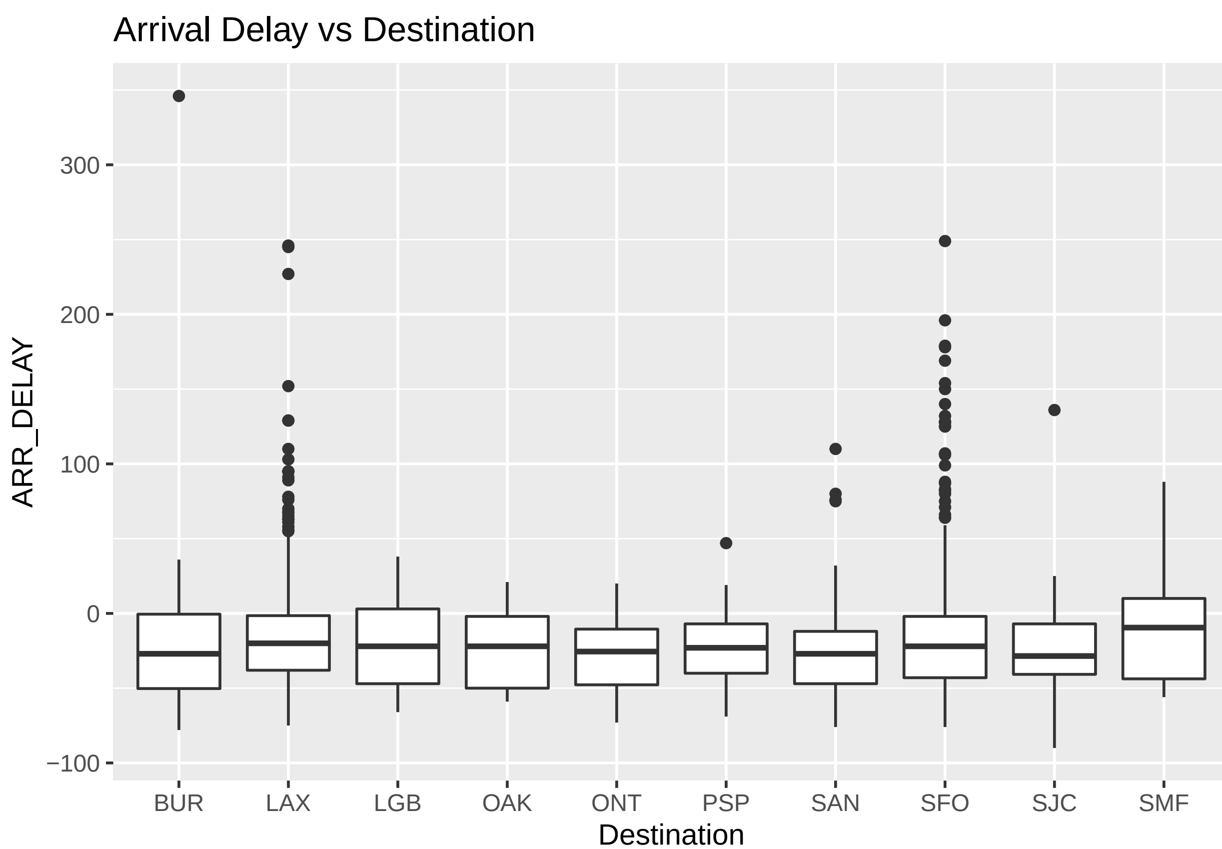
```
## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing missing values (geom_point).
```



These plots above suggest that we may want to transform the variables at some point.

```
ggplot(data = flights, aes(y = ARR_DELAY, x = DEST)) +
  geom_boxplot() +
  labs(x = "Destination",
       title = "Arrival Delay vs Destination")
```

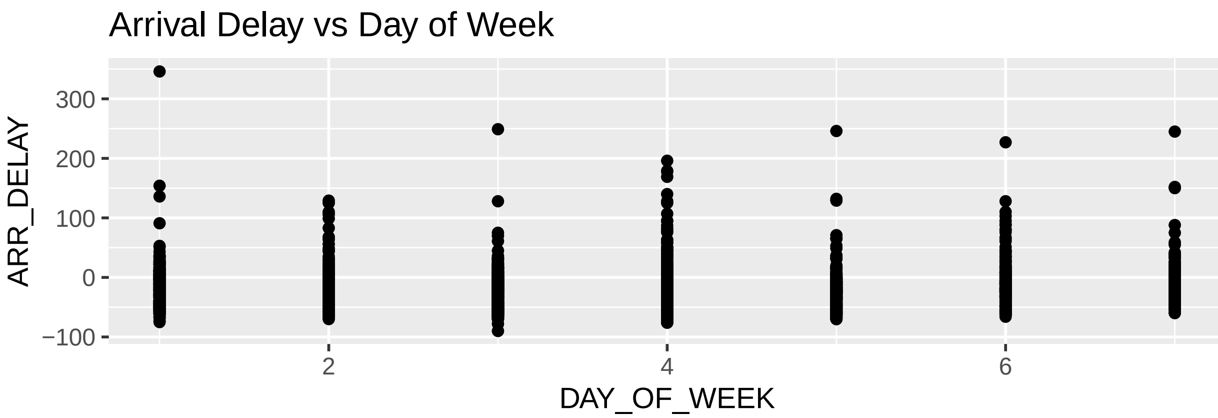
```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```

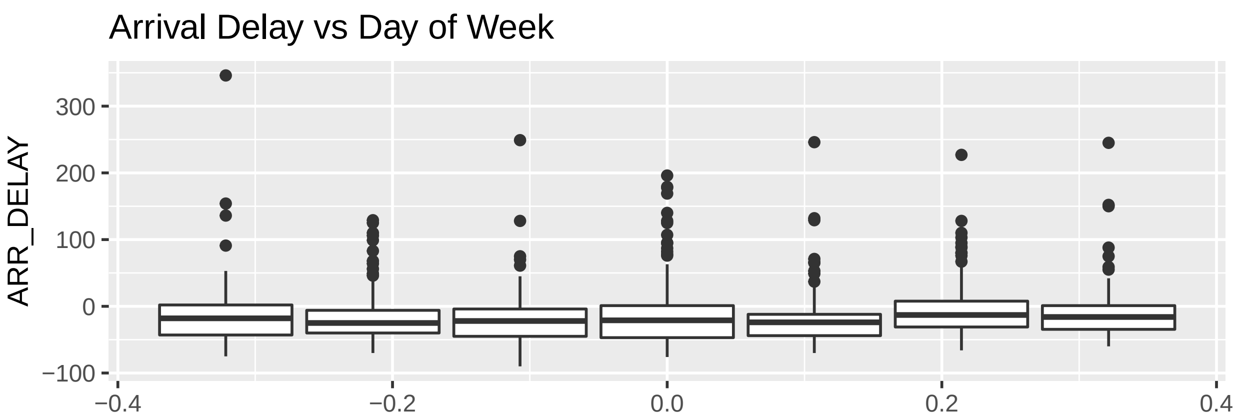


ARR_DELAY and DAY_OF_WEEK

```r
p8 <- ggplot(data = flights, aes(y = ARR_DELAY, x = DAY_OF_WEEK)) +
  geom_point() +
  labs(title = "Arrival Delay vs Day of Week")

p9 <- ggplot(data = flights, aes(y = ARR_DELAY, group = DAY_OF_WEEK)) +
  geom_boxplot() +
  labs(title = "Arrival Delay vs Day of Week")

grid.arrange(p8,p9, nrow = 2)
```

```
## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```

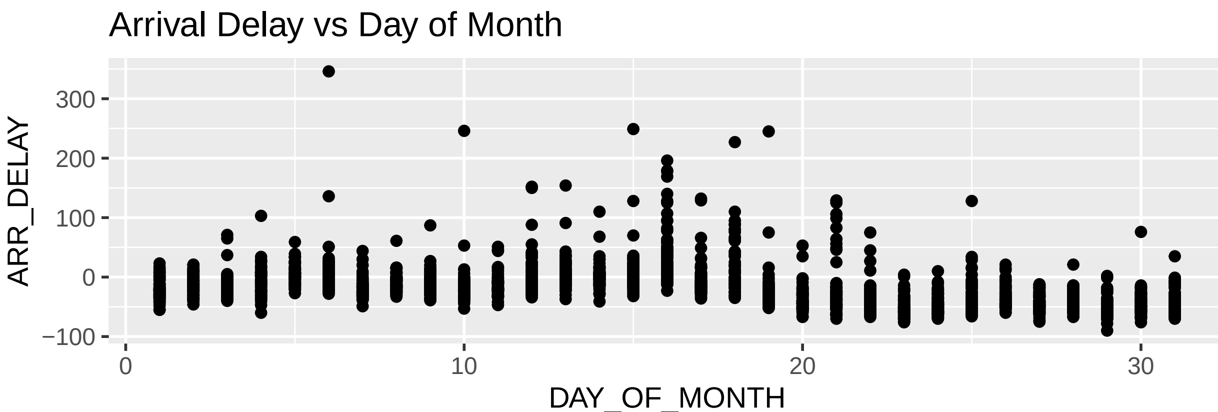


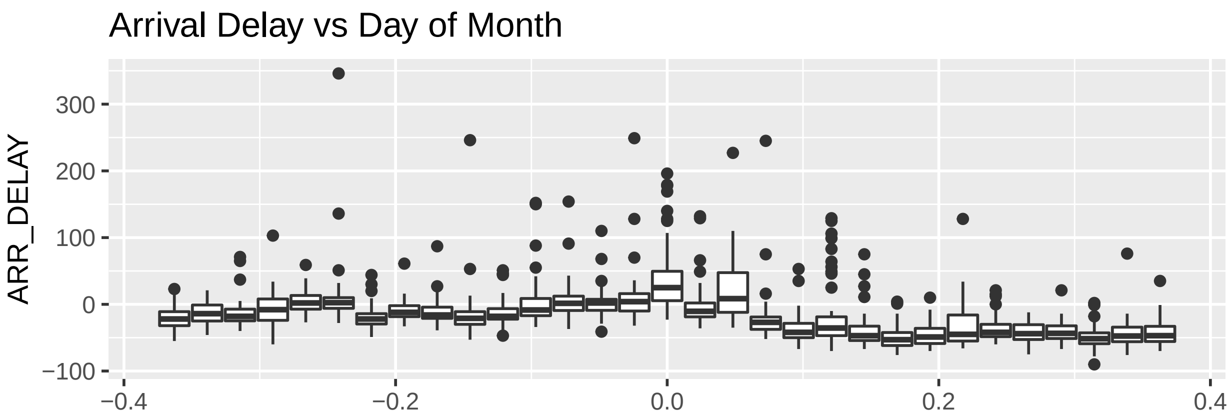


ARR_DELAY and DAY_OF_MONTH

```
p10 <- ggplot(data = flights, aes(y = ARR_DELAY, x = DAY_OF_MONTH)) +
  geom_point() +
  labs(title = "Arrival Delay vs Day of Month")

p11 <- ggplot(data = flights, aes(y = ARR_DELAY, group = DAY_OF_MONTH)) +
  geom_boxplot() +
  labs(title = "Arrival Delay vs Day of Month")

grid.arrange(p10, p11, nrow = 2)
```

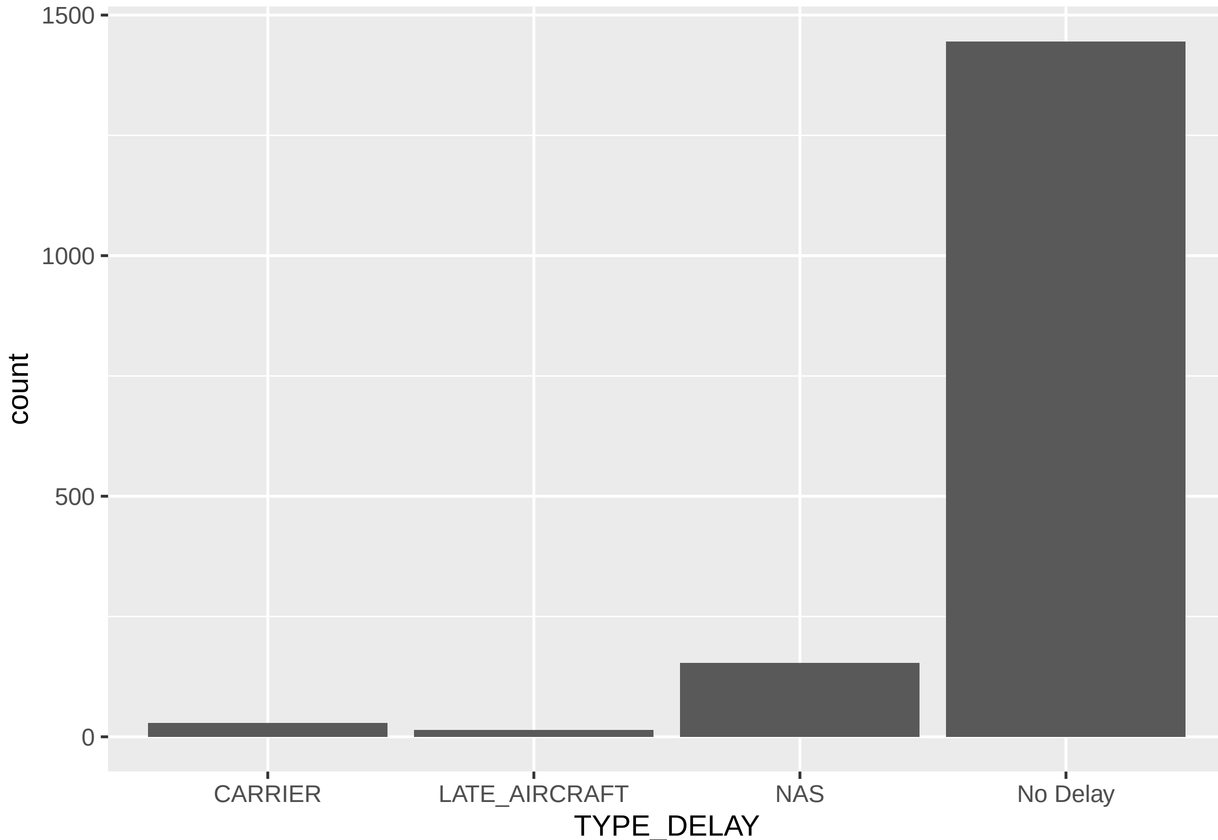
```
## Warning: Removed 11 rows containing missing values (geom_point).

## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```





Further Data Cleaning

```r
# take only SFO/LAX since all 4 carriers fly there
flights <- flights %>%
  filter(DEST == "SFO" | DEST == "LAX") %>%
  mutate(TYPE_DELAY = case_when(NAS_DELAY == 1 ~ "NAS",
                                CARRIER_DELAY == 1 ~ "CARRIER",
                                LATE_AIRCRAFT_DELAY == 1 ~ "LATE_AIRCRAFT",
                                TRUE ~ "No Delay"))

ggplot(data = flights, aes(x = TYPE_DELAY)) +
  geom_bar()
```




```r
unique(flights$TYPE_DELAY)
```

```
## [1] "No Delay"      "NAS"           "LATE_AIRCRAFT" "CARRIER"
```

SPLITTING DATA

```r
set.seed(1234)
flights <- flights %>%
  mutate(id = row_number())
train <- flights %>%
  sample_frac(0.8)
test <- anti_join(flights, train, by = "id")
```

LINEAR MODELS

Variables that I think we could explore: department delay time, days of month, days of week, taxi-in, taxi-out, destination, Carrier Delay, NAS Delay, and Late Aircraft Delay.

Full Model

```r
lm.01 <- lm(ARR_DELAY ~ DEP_DELAY + DAY_OF_WEEK + OP_CARRIER + DEST + CRS_DEP_TIME + CRS_ARR_TIME + log

#plot(lm.01)
#summary(lm.01)
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:patchwork':
##
##     area

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
step_model <- stepAIC(lm.01, direction = "backward", trace = FALSE)
#summary(step_model)

lm.02 <- lm(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME  + log_TAXI_OUT + log_TAXI_IN + TYI
#summary(lm.02)
#anova(step_model, lm.02)

lm.03 <- lm(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME  + log_TAXI_OUT + log_TAXI_IN + TYI

#anova(lm.02, lm.03)


final_model <- lm(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME  + log_TAXI_OUT + log_TAXI_II

anova(lm.03, final_model)
```

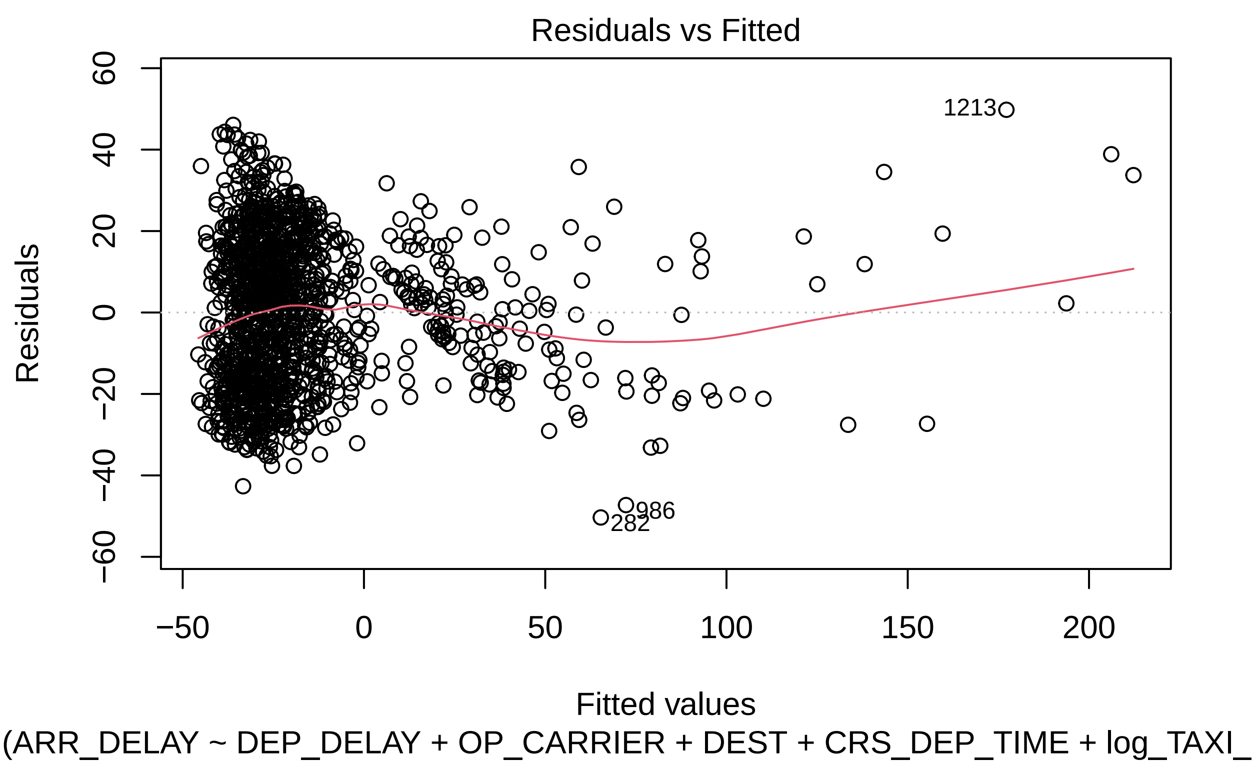
```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME + log_TAXI_OUT +
##     log_TAXI_IN + TYPE_DELAY + OP_CARRIER:DEST + DEST:log_TAXI_IN
## Model 2: ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME + log_TAXI_OUT +
##     log_TAXI_IN + TYPE_DELAY + OP_CARRIER:DEST + DEST:log_TAXI_IN +
##     log_TAXI_OUT:DEP_DELAY
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   1295 417435
## 2   1294 415006  1    2429.3 7.5745 0.006003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

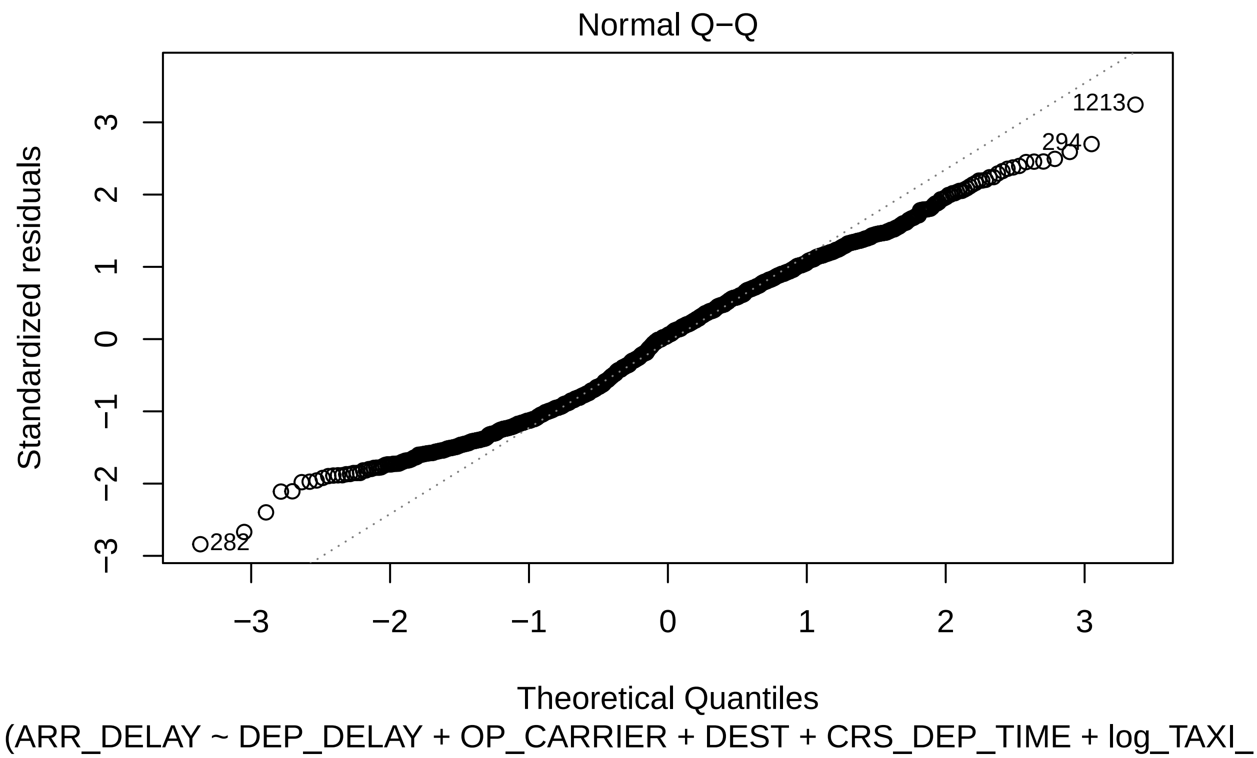
```r
summary(final_model)
```

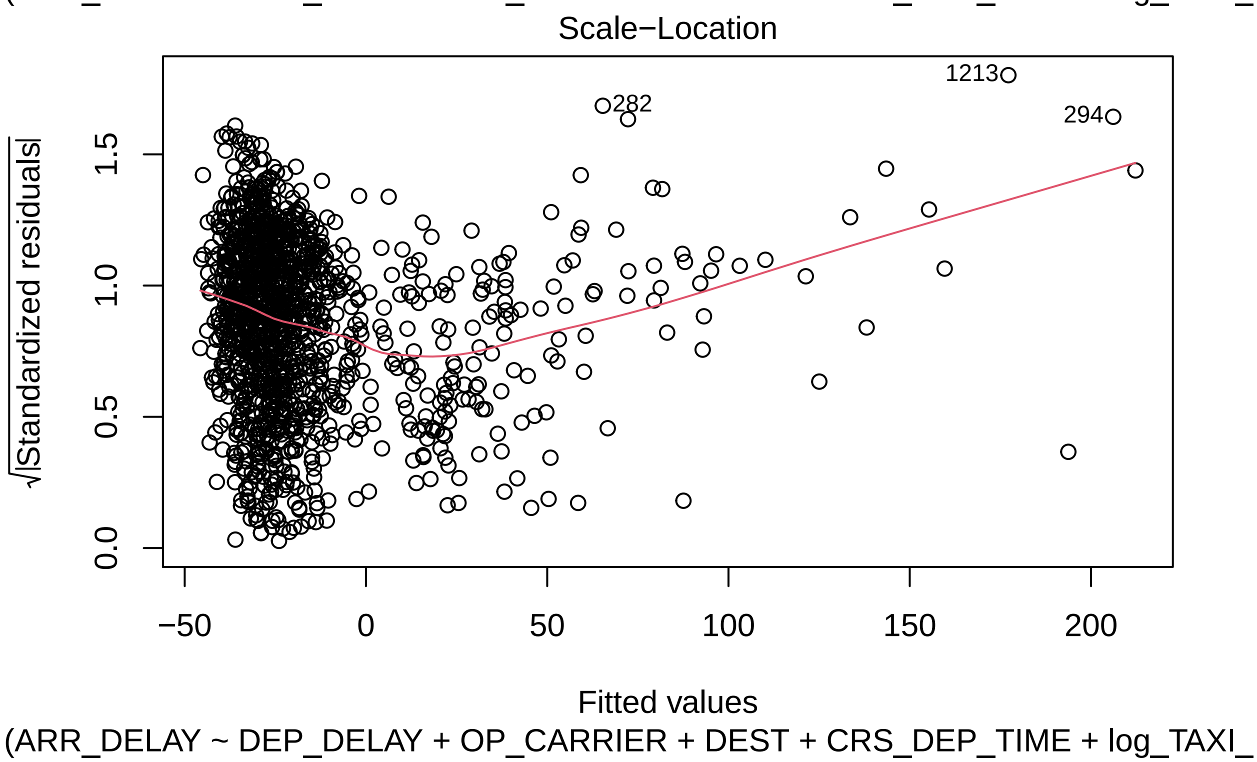
```
##
## Call:
## lm(formula = ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME +
##     log_TAXI_OUT + log_TAXI_IN + TYPE_DELAY + OP_CARRIER:DEST +
##     DEST:log_TAXI_IN + log_TAXI_OUT:DEP_DELAY, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.329 -14.930   0.958  13.682  49.785
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
```

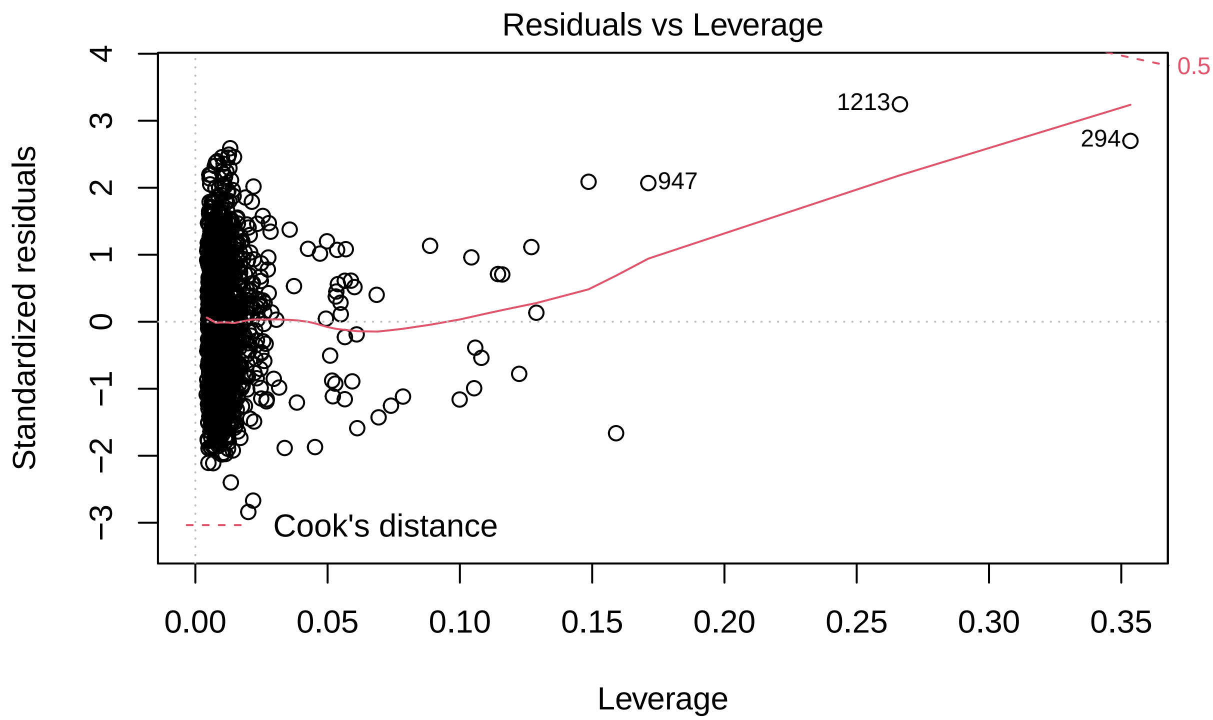
```
## (Intercept)              -79.875033   7.154618 -11.164  < 2e-16 ***
## DEP_DELAY                  0.473110   0.147155   3.215 0.001336 **
## OP_CARRIERAS              -4.960311   2.061338  -2.406 0.016252 *
## OP_CARRIERB6               5.663369   1.662176   3.407 0.000676 ***
## OP_CARRIERDL              -2.355823   1.689398  -1.394 0.163414
## DESTSFO                    7.590819   4.318505   1.758 0.079028 .
## CRS_DEP_TIME              -0.003717   0.001056  -3.522 0.000444 ***
## log_TAXI_OUT              20.280646   1.592782  12.733  < 2e-16 ***
## log_TAXI_IN                7.910495   1.054399   7.502 1.16e-13 ***
## TYPE_DELAYLATE_AIRCRAFT   -4.298799   6.682903  -0.643 0.520174
## TYPE_DELAYNAS             23.936600   4.539529   5.273 1.57e-07 ***
## TYPE_DELAYNo Delay       -16.409905   4.460083  -3.679 0.000243 ***
## OP_CARRIERAS:DESTSFO       5.527479   3.284028   1.683 0.092589 .
## OP_CARRIERB6:DESTSFO      -4.714933   2.847131  -1.656 0.097958 .
## OP_CARRIERDL:DESTSFO       0.359385   2.861596   0.126 0.900077
## DESTSFO:log_TAXI_IN       -4.397237   1.926591  -2.282 0.022628 *
## DEP_DELAY:log_TAXI_OUT     0.125352   0.045546   2.752 0.006003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.91 on 1294 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7213, Adjusted R-squared:  0.7178
## F-statistic: 209.3 on 16 and 1294 DF,  p-value: < 2.2e-16
```

```r
plot(final_model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME + log_TAXI_
1213
294
282
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME + log_TAXI_
1213
282
294
0.0
0.5
1.0
1.5
−50
0
50
100
150
200



(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST + CRS_DEP_TIME + log_TAXI_

```
## SIGNIFICANT INTERACTIONS
#OP_CARRIER:DEST
#DEST:log_TAXI_IN
#CRS_DEP_TIME:DEST (***** makes zero intuitive sense - might not wanna do this)
#CRS_ARR_TIME:log_TAXI_IN
#log_TAXI_OUT:DEP_DELAY


#log_TAXI_OUT:CRS_DEP_TIME (verrrrrry close to 0.05)
library(broom)
final_linear_preds <- predict(final_model, train)
linear_MSE <- sum((final_linear_preds-train$ARR_DELAY)^2, na.rm=T)/328
linear_MSE
```

```
## [1] 1265.262
```

First, let's just fit a full linear model with all the variables we would like to explore.

```
full_model <- lm(ARR_DELAY ~ DAY_OF_MONTH +
                   DAY_OF_WEEK +
                   TAXI_IN +
                   TAXI_OUT +
                   DEST +
                   DEP_DELAY +
                   CARRIER_DELAY +
                   NAS_DELAY +
                   LATE_AIRCRAFT_DELAY, data = train)

summary(full_model)
```

```
##
```

```
## Call:
## lm(formula = ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + TAXI_IN +
##     TAXI_OUT + DEST + DEP_DELAY + CARRIER_DELAY + NAS_DELAY +
##     LATE_AIRCRAFT_DELAY, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.659  -9.913  -1.229   9.243  46.780
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -22.18852    1.58821 -13.971   <2e-16 ***
## DAY_OF_MONTH         -1.28951    0.04418 -29.187   <2e-16 ***
## DAY_OF_WEEK          -0.28103    0.20758  -1.354   0.1760
## TAXI_IN               0.55575    0.04785  11.615   <2e-16 ***
## TAXI_OUT              0.73768    0.04368  16.887   <2e-16 ***
## DESTSFO              -0.33517    0.82901  -0.404   0.6861
## DEP_DELAY             0.89165    0.02221  40.145   <2e-16 ***
## CARRIER_DELAY         2.30229    2.30029   1.001   0.3171
## NAS_DELAY            32.68992    1.54500  21.159   <2e-16 ***
## LATE_AIRCRAFT_DELAY   5.54853    3.24643   1.709   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.11 on 1301 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8259, Adjusted R-squared:  0.8247
## F-statistic: 685.8 on 9 and 1301 DF,  p-value: < 2.2e-16
```

```r
full_model_preds <- predict(full_model, train)
linear_MSE <- sum((full_model_preds-train$ARR_DELAY)^2, na.rm=T)/328
linear_MSE
```

```
## [1] 790.2284
```

Select Model with AIC

```r
library(MASS)
step_model <- stepAIC(full_model, trace = FALSE)
summary(step_model)
```

```
##
## Call:
## lm(formula = ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT +
##     DEP_DELAY + NAS_DELAY + LATE_AIRCRAFT_DELAY, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.702 -10.034  -1.314   9.034  46.852
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -23.31594    1.34840  -17.29   <2e-16 ***
## DAY_OF_MONTH         -1.28947    0.04400  -29.30   <2e-16 ***
## TAXI_IN               0.55710    0.04637   12.01   <2e-16 ***
```

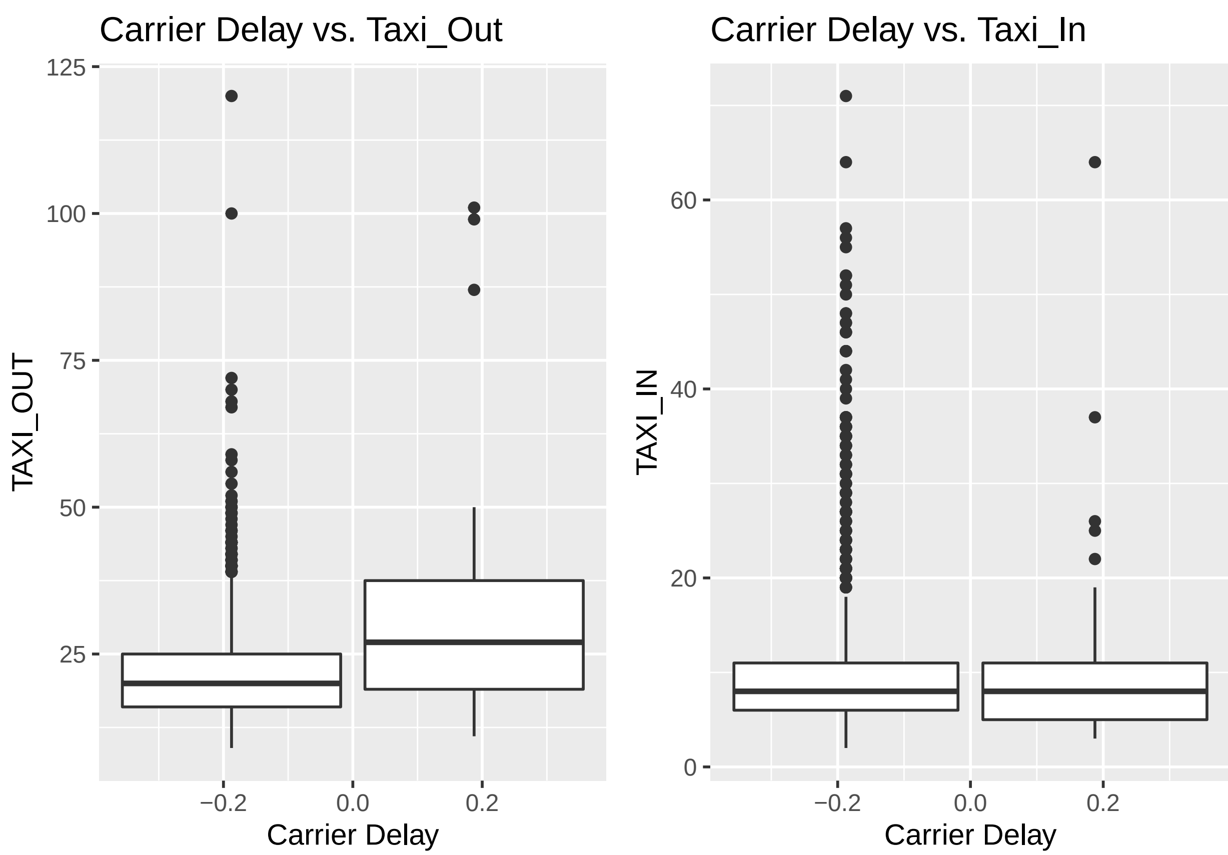
```
## TAXI_OUT              0.73506    0.04347   16.91   <2e-16 ***
## DEP_DELAY             0.89777    0.02100   42.76   <2e-16 ***
## NAS_DELAY            33.03098    1.50853   21.90   <2e-16 ***
## LATE_AIRCRAFT_DELAY   5.44580    3.24199    1.68   0.0932 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.12 on 1304 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8255, Adjusted R-squared:  0.8247
## F-statistic:  1028 on 6 and 1304 DF,  p-value: < 2.2e-16
```

The only variables that were removed were DAY_OF_WEEK and LATE_AIRCRAFT_DELAY. Let's continue using the step_model then.

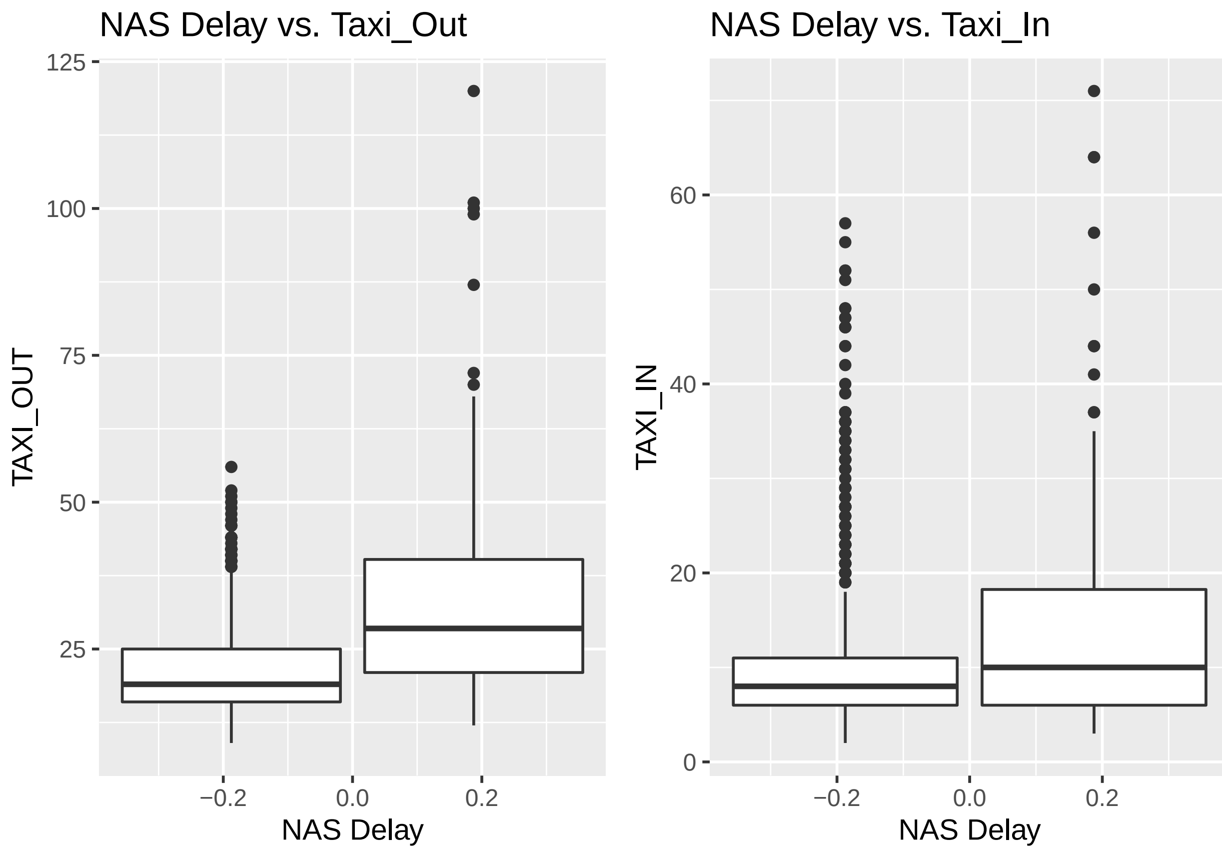
Interactions

Because there are so many levels to Destination, I don't know if we should necessarily include an interaction with this categorical variable. My suggestion would be to find interactions with carrier_delay and nas_delay.

```
p12 <- ggplot(data = train, aes(group = CARRIER_DELAY, y = TAXI_OUT)) +
  geom_boxplot() +
  labs(title = "Carrier Delay vs. Taxi_Out",
       x = "Carrier Delay")

p13 <- ggplot(data = train, aes(group = CARRIER_DELAY, y = TAXI_IN)) +
  geom_boxplot() +
  labs(title = "Carrier Delay vs. Taxi_In",
       x = "Carrier Delay")

grid.arrange(p12, p13, nrow = 1)
```




```
p14 <- ggplot(data = train, aes(group = NAS_DELAY, y = TAXI_OUT)) +
  geom_boxplot() +
  labs(title = "NAS Delay vs. Taxi_Out",
       x = "NAS Delay")

p15 <- ggplot(data = train, aes(group = NAS_DELAY, y = TAXI_IN)) +
  geom_boxplot() +
  labs(title = "NAS Delay vs. Taxi_In",
       x = "NAS Delay")

grid.arrange(p14, p15, nrow = 1)
```



From what I'm seeing in the plots above, there could be an interaction between taxi_out and carrier_delay. There also seems to be an interaction between NAS delay and taxi_out as well as a possible one between NAS delay and taxi_in. Let's test these three interactions below.

```
# carrier vs taxi out
interaction1 <- lm(ARR_DELAY ~ DAY_OF_MONTH +
                     TAXI_IN +
                     TAXI_OUT +
                     DEST +
                     DEP_DELAY +
                     CARRIER_DELAY +
                     NAS_DELAY +
                   CARRIER_DELAY*TAXI_OUT, data = train)
# nas vs taxi out
interaction2 <- lm(ARR_DELAY ~ DAY_OF_MONTH +
                     TAXI_IN +
                     TAXI_OUT +
                     DEST +
                     DEP_DELAY +
                     CARRIER_DELAY +
                     NAS_DELAY +
                   NAS_DELAY*TAXI_OUT, data = train)

# nas vs taxi in
interaction3 <- lm(ARR_DELAY ~ DAY_OF_MONTH +
                     TAXI_IN +
                     TAXI_OUT +
```

```
                   DEST +
                   DEP_DELAY +
                   CARRIER_DELAY +
                   NAS_DELAY +
                  NAS_DELAY*TAXI_IN, data = train)
```

```
anova(step_model, interaction1)
```

```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEP_DELAY + NAS_DELAY +
##     LATE_AIRCRAFT_DELAY
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEST + DEP_DELAY +
##     CARRIER_DELAY + NAS_DELAY + CARRIER_DELAY * TAXI_OUT
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   1304 259813
## 2   1302 260081  2   -268.59
```

```
anova(step_model, interaction2)
```

```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEP_DELAY + NAS_DELAY +
##     LATE_AIRCRAFT_DELAY
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEST + DEP_DELAY +
##     CARRIER_DELAY + NAS_DELAY + NAS_DELAY * TAXI_OUT
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   1304 259813
## 2   1302 260101  2   -288.77
```

```
anova(step_model, interaction3)
```

```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEP_DELAY + NAS_DELAY +
##     LATE_AIRCRAFT_DELAY
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEST + DEP_DELAY +
##     CARRIER_DELAY + NAS_DELAY + NAS_DELAY * TAXI_IN
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1304 259813
## 2   1302 258380  2    1432.9 3.6103 0.02732 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It actually seems that interaction3: NAS_DELAY and TAXI_IN is the only interaction that is statistically significant in predicting ARR_DELAY. Let's make this model our current model:

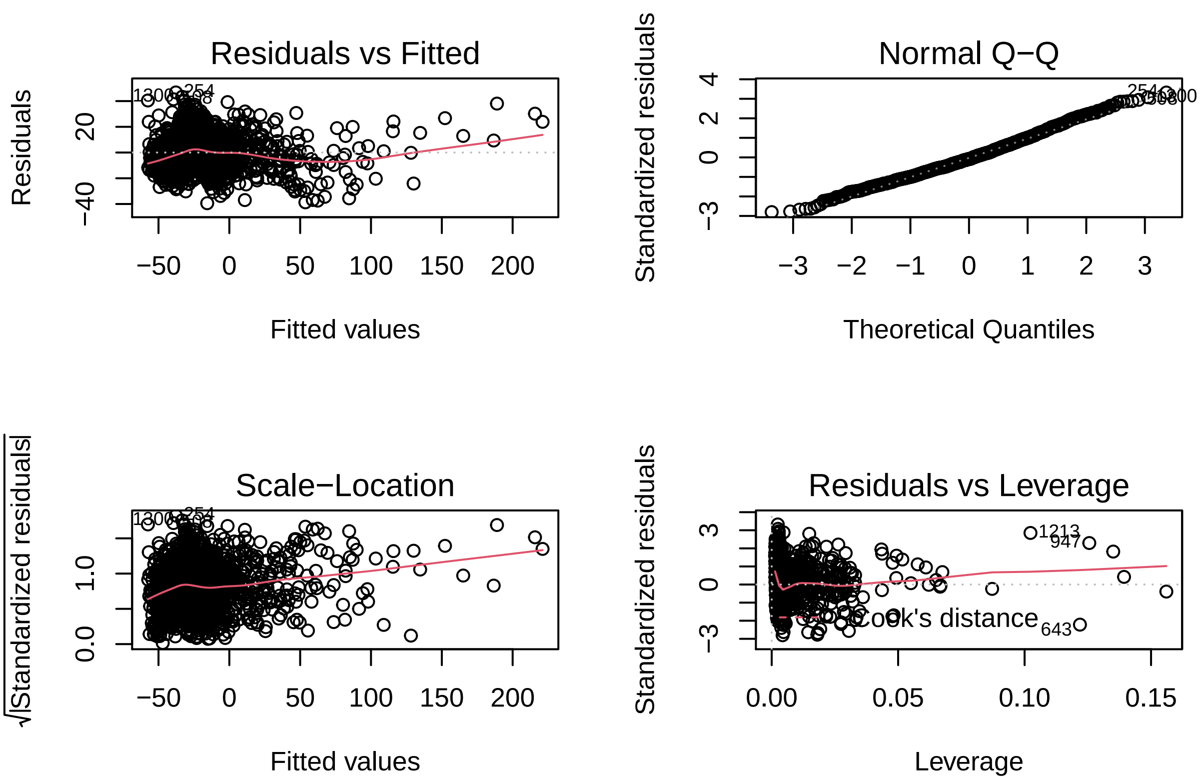
Final Linear Model

```
current_model <- interaction3

summary(current_model)
```

```
##
## Call:
## lm(formula = ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT +
```

```
##     DEST + DEP_DELAY + CARRIER_DELAY + NAS_DELAY + NAS_DELAY *
##     TAXI_IN, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.388  -9.698  -1.216   8.983  46.729
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -23.48832    1.41476 -16.602  < 2e-16 ***
## DAY_OF_MONTH      -1.29353    0.04411 -29.322  < 2e-16 ***
## TAXI_IN            0.62691    0.05364  11.688  < 2e-16 ***
## TAXI_OUT           0.72130    0.04348  16.587  < 2e-16 ***
## DESTSFO           -0.39829    0.82704  -0.482  0.63019
## DEP_DELAY          0.90193    0.01975  45.666  < 2e-16 ***
## CARRIER_DELAY      2.63404    2.29653   1.147  0.25161
## NAS_DELAY         37.29871    2.17272  17.167  < 2e-16 ***
## TAXI_IN:NAS_DELAY  -0.32214    0.10933  -2.946  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.09 on 1302 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8265, Adjusted R-squared:  0.8254
## F-statistic: 775.1 on 8 and 1302 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(current_model)
```



The diagnostic plots above suggest that this model decently satisfies the necessary conditions to assume a linear regression.

Response (Box-Cox) Transformation

```r
## AFTER SELECTED MODEL
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'

## The following object is masked from 'package:MASS':
##
##     boxcox

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

## The following object is masked from 'package:base':
##
##     print.default
```

```r
# bc_model <- boxcox(final_model, optimize = TRUE)
# bc_lambda <- bc_model$lambda
# bc_lambda
# plot(bc_model)
```

```r
# add Box-Cox transform to data
# train_data <- train_data %>%
#   mutate(bc_R_moment_1 = ((R_moment_1^bc_lambda) - 1)/bc_lambda)
#
# hist(train_data$bc_R_moment_1)
```

Test Error

```r
lm_preds <- predict(current_model, test)
sum((test$ARR_DELAY - lm_preds)^2, na.rm=T)/328
```

```
## [1] 220.1752
```

GAM MODEL

Initial Model

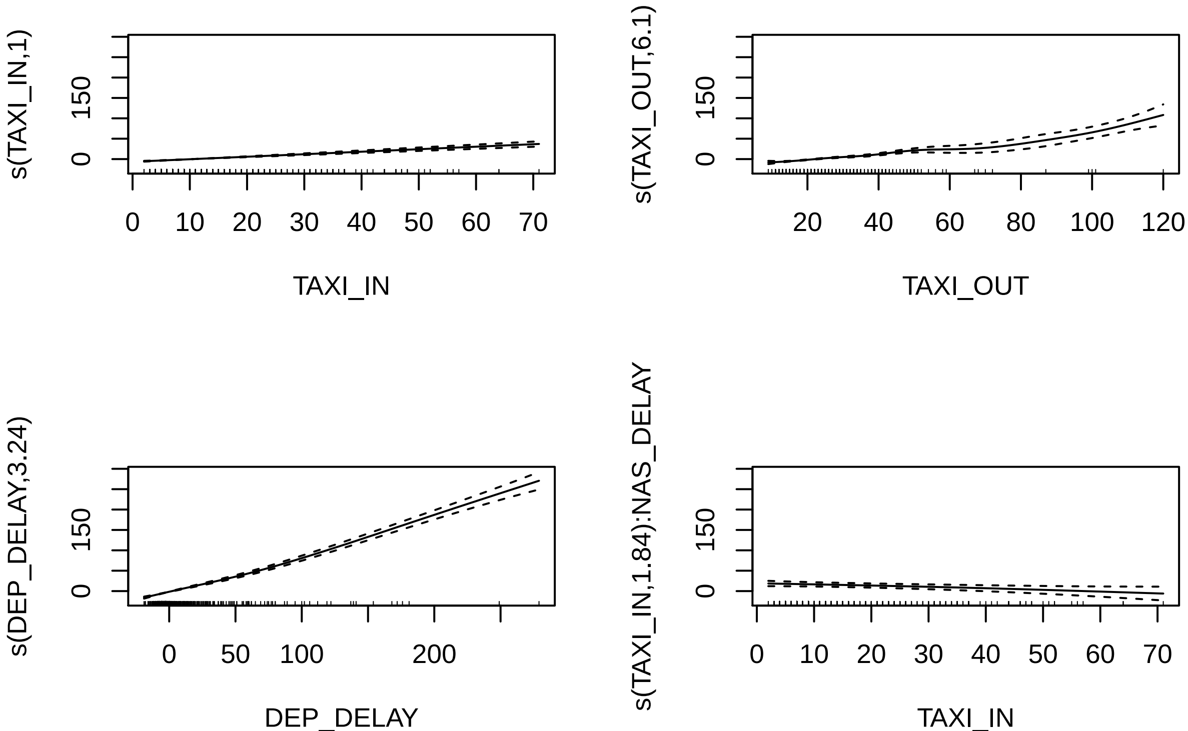
fit a gam model with numerical variables on a smoothing spline and including the interaction between NAS_DELAY and TAXI_IN

```r
gam00 <- gam(ARR_DELAY ~ DAY_OF_MONTH +
                   DAY_OF_WEEK +
                   s(TAXI_IN) +
                   s(TAXI_OUT) +
                   DEST +
                   s(DEP_DELAY) +
                   CARRIER_DELAY +
                   NAS_DELAY +
                   LATE_AIRCRAFT_DELAY +
```

```r
                   s(TAXI_IN, by = NAS_DELAY), data = train)

summary(gam00)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + s(TAXI_IN) + s(TAXI_OUT) +
##     DEST + s(DEP_DELAY) + CARRIER_DELAY + NAS_DELAY + LATE_AIRCRAFT_DELAY +
##     s(TAXI_IN, by = NAS_DELAY)
##
## Parametric coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.73197    1.21228   1.429   0.1533
## DAY_OF_MONTH        -1.30430    0.04386 -29.736  < 2e-16 ***
## DAY_OF_WEEK         -0.25535    0.20508  -1.245   0.2133
## DESTSFO             -0.28562    0.82069  -0.348   0.7279
## CARRIER_DELAY        4.96143    2.33845   2.122   0.0341 *
## NAS_DELAY           18.45820    2.61969   7.046 2.99e-12 ***
## LATE_AIRCRAFT_DELAY  7.50690    3.25263   2.308   0.0212 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                        edf Ref.df      F  p-value
## s(TAXI_IN)           1.000  1.000 134.59  < 2e-16 ***
## s(TAXI_OUT)          6.103  7.128  42.97  < 2e-16 ***
## s(DEP_DELAY)         3.240  4.019 393.06  < 2e-16 ***
## s(TAXI_IN):NAS_DELAY 1.839  2.099  18.12 5.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 43/44
## R-sq.(adj) =  0.831   Deviance explained = 83.3%
## GCV =  195.4  Scale est. = 192.61    n = 1311
```

```r
par(mfrow = c(2,2))
plot.gam(gam00, se=TRUE)
```



Checking Lineartiy

TAXI_IN and the interaction between NAS_DELAY and TAXI_IN may be linear

```r
gam01 <- gam(ARR_DELAY ~ DAY_OF_MONTH +
                   DAY_OF_WEEK +
                   TAXI_IN +
                   s(TAXI_OUT) +
                   DEST +
                   s(DEP_DELAY) +
                   CARRIER_DELAY +
                   NAS_DELAY +
                   LATE_AIRCRAFT_DELAY +
                   TAXI_IN*NAS_DELAY, data = train)

anova(gam00, gam01, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + s(TAXI_IN) + s(TAXI_OUT) +
##     DEST + s(DEP_DELAY) + CARRIER_DELAY + NAS_DELAY + LATE_AIRCRAFT_DELAY +
##     s(TAXI_IN, by = NAS_DELAY)
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + TAXI_IN + s(TAXI_OUT) +
##     DEST + s(DEP_DELAY) + CARRIER_DELAY + NAS_DELAY + LATE_AIRCRAFT_DELAY +
##     TAXI_IN * NAS_DELAY
##   Resid. Df Resid. Dev       Df Deviance     F Pr(>F)
## 1    1290.3     248917
## 2    1290.8     249038 -0.51093  -120.95 1.229 0.2267
```

based on anova test, the model with smoothing splines on TAXI_IN and the interaction term is a better fit

More Anova

DAY_OF_WEEK and DEST have very high p-values, so let's try an anova test without including them

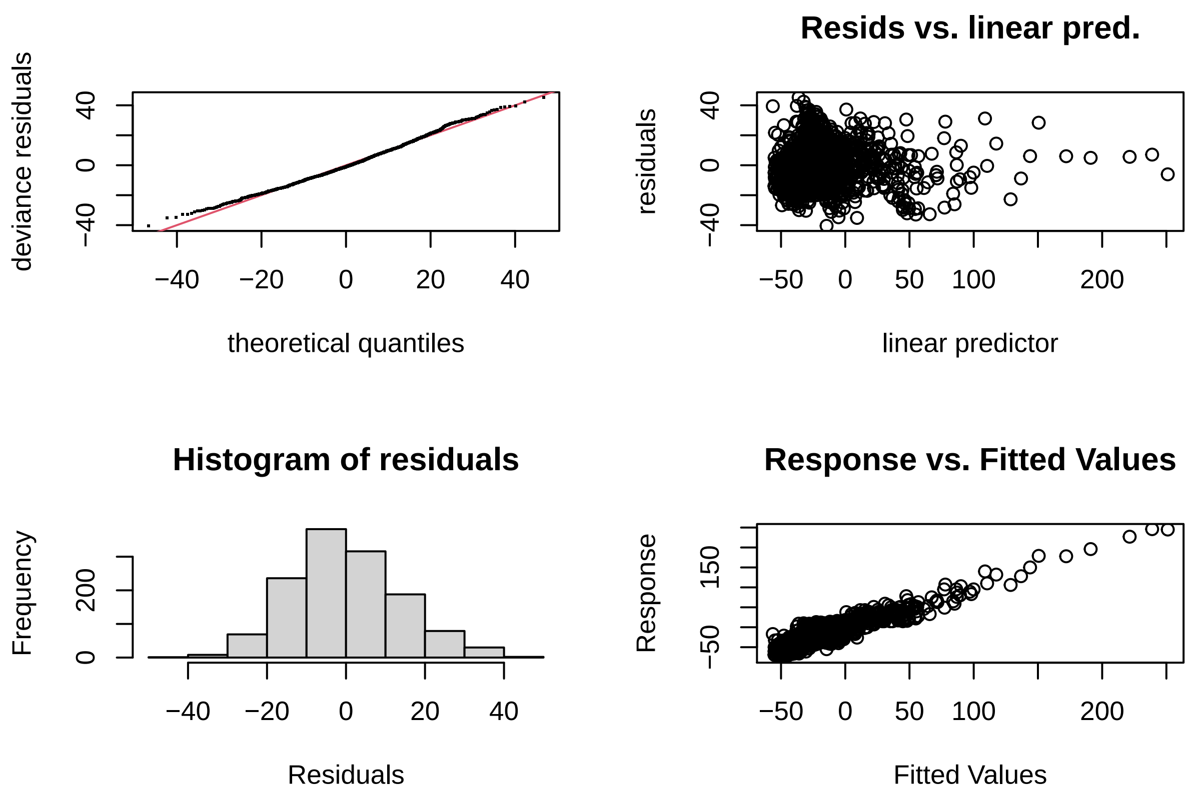
```
gam02 <- gam(ARR_DELAY ~ DAY_OF_MONTH +
                   s(TAXI_IN) +
                   s(TAXI_OUT) +
                   s(DEP_DELAY) +
                   CARRIER_DELAY +
                   NAS_DELAY +
                   LATE_AIRCRAFT_DELAY +
                   s(TAXI_IN, by = NAS_DELAY), data = train)

anova(gam00, gam02, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + s(TAXI_IN) + s(TAXI_OUT) +
##     DEST + s(DEP_DELAY) + CARRIER_DELAY + NAS_DELAY + LATE_AIRCRAFT_DELAY +
##     s(TAXI_IN, by = NAS_DELAY)
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + s(TAXI_IN) + s(TAXI_OUT) + s(DEP_DELAY) +
##     CARRIER_DELAY + NAS_DELAY + LATE_AIRCRAFT_DELAY + s(TAXI_IN,
##     by = NAS_DELAY)
##   Resid. Df Resid. Dev      Df Deviance      F Pr(>F)
## 1    1290.3     248917
## 2    1292.2     249218 -1.9182  -300.41 0.8131 0.4393
```

based on the anova test, the model including DAY_OF_WEEK and DEST is a better fit

Model Diagnostics

```
par(mfrow = c(2,2))
gam.check(gam00)
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 14 iterations.
## The RMS GCV score gradient at convergence was 6.788545e-06 .
## The Hessian was positive definite.
## Model rank =  43 / 44
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                         k'   edf k-index p-value
## s(TAXI_IN)            9.00  1.00    0.97   0.090 .
## s(TAXI_OUT)           9.00  6.10    1.01   0.665
## s(DEP_DELAY)          9.00  3.24    0.96   0.035 *
## s(TAXI_IN):NAS_DELAY 10.00  1.84    0.97   0.130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test Error

```
gam_preds <- predict.gam(gam00, newdata = test)
sum((test$ARR_DELAY - gam_preds)^2, na.rm=T)/328
```

```
## [1] 230.8748
```

TREES

Random Forests

```
library(tree)
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

By default, `randomForest()` uses $p/3$ variables when building a random forest of regression trees.

```
set.seed(1)
rf.delay <- randomForest(ARR_DELAY ~ DEP_DELAY + OP_CARRIER + DEST +
                           CRS_DEP_TIME + CRS_ARR_TIME + log_TAXI_OUT +
                           log_TAXI_IN + TYPE_DELAY + OP_CARRIER:DEST +
                           DEST:log_TAXI_IN + CRS_ARR_TIME:log_TAXI_IN +
                           log_TAXI_OUT:DEP_DELAY,
                   data = train, na.action = na.omit, importance = TRUE,
                   ntree=10000)
yhat.rf <- predict(rf.delay, newdata = test)
rf.MSE <- sum((test$ARR_DELAY - yhat.rf)^2, na.rm=T)/328
rf.MSE
```

```
## [1] 468.0524
```

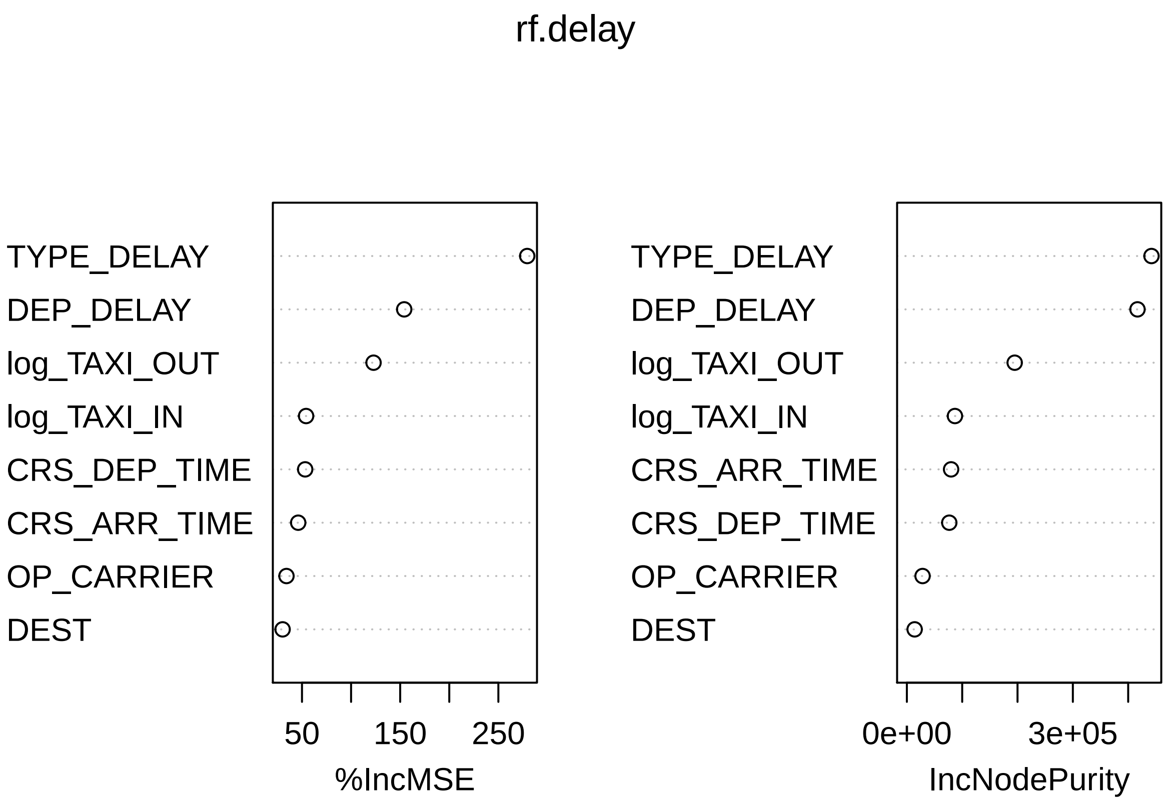
Using the `importance()` function, we can view the importance of each variable.

```
importance(rf.delay)
```

```
##                %IncMSE IncNodePurity
## DEP_DELAY    154.09231     416846.60
## OP_CARRIER    34.18926      28395.69
## DEST          30.25325      14047.33
## CRS_DEP_TIME  53.22776      76644.29
## CRS_ARR_TIME  46.13102      79938.18
## log_TAXI_OUT 122.86877     194906.75
## log_TAXI_IN   54.16845      86854.94
## TYPE_DELAY   279.36573     442114.64
```

Two measures of variable importance are reported. The former is based on the mean decrease in accuracy in predictions on the out of bag samples when a given variable is excluded from the model. The latter is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees (this was plotted in Figure 8.9 in the text). In the case of regression trees, the node impurity is measured by the training RSS and for classification trees by the deviance. Plots of these importance measures can be produced using the `varImpPlot()` function.

```
varImpPlot(rf.delay)
```



4. Boosting

Here we use the `gbm()` package, and within it the `gbm()` function, to fit boosted regression trees to the `train` data set. We run `gbm()` with the option `distribution = "gaussian"` since this is a regression problem. The argument `n.trees = 5000` indicates that we want 5000 trees, and the option `interaction.depth = 4` limits the depth of each tree.

```
# library(gbm)
# set.seed(1)
# boost.boston <- gbm(ARR_DELAY ~ DAY_OF_MONTH +
#                     TAXI_IN +
#                     TAXI_OUT +
#                     DEP_DELAY +
#                     CARRIER_DELAY +
#                     NAS_DELAY +
#                     LATE_AIRCRAFT_DELAY,
#                     data = train, distribution = "gaussian",
#                     n.trees=1000, interaction.depth=1, shrinkage=0.001, cv.folds=10)
```

The `summary()` function also provides a relative influence plot and also outputs the relative influence statistics.

```r
#summary(boost.boston)
```

We see that `lstat` and `rm` are by far the most important variables. We can also produce *partial dependence plots* for these two variables. These plots illustrate the marginal effect of the selected variables on the response after `integrating` out the other variables. In this case, as we might expect, median house prices are increasing with `rm` and decreasing with `lstat`.

```r
# par(mfrow = c(1,2))
# plot(boost.boston, i = "rm")
# plot(boost.boston, i = "lstat")
```

We now use the boosted model to predict `medv` on the test set:

```r
# yhat.boost <- predict(boost.boston, newdata = Boston[-train,],
#                        n.trees = 5000)
# mean((yhat.boost - boston.test)^2)
```

The test MSE obtained is 11.8; similar to the test MSE for random forests and superior to that for bagging. If we want to, we can perform boosting with a different value of the shrinkage parameter λ in Equation 8.10. The default value is 0.001, but this is easily modified. Here, we take $\lambda = 0.2$.

```r
# boost.boston <- gbm(medv~., data = Boston[train,],
#                     distribution = "gaussian", n.trees = 5000,
#                     interaction.depth = 4,
#                     shrinkage = 0.2,
#                     verbose =FALSE)
# yhat.boost <- predict(boost.boston, newdata = Boston[-train,],
#                        n.trees = 5000)
# mean((yhat.boost - boston.test)^2)
```