Sta 325 Final Project

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

11/22/2020

```
library(readr)
library(dplyr)
library(tidyverse)
library(gridExtra)
```

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

```
## Warning: 1 parsing failure.
##  row               col           expected actual               file
## 1143 CANCELLATION_CODE 1/0/T/F/TRUE/FALSE      A 'data/flights.csv'
```

```
unique(flights$OP_CARRIER)
```

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

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

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

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

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

```
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))
```

```
flights
```

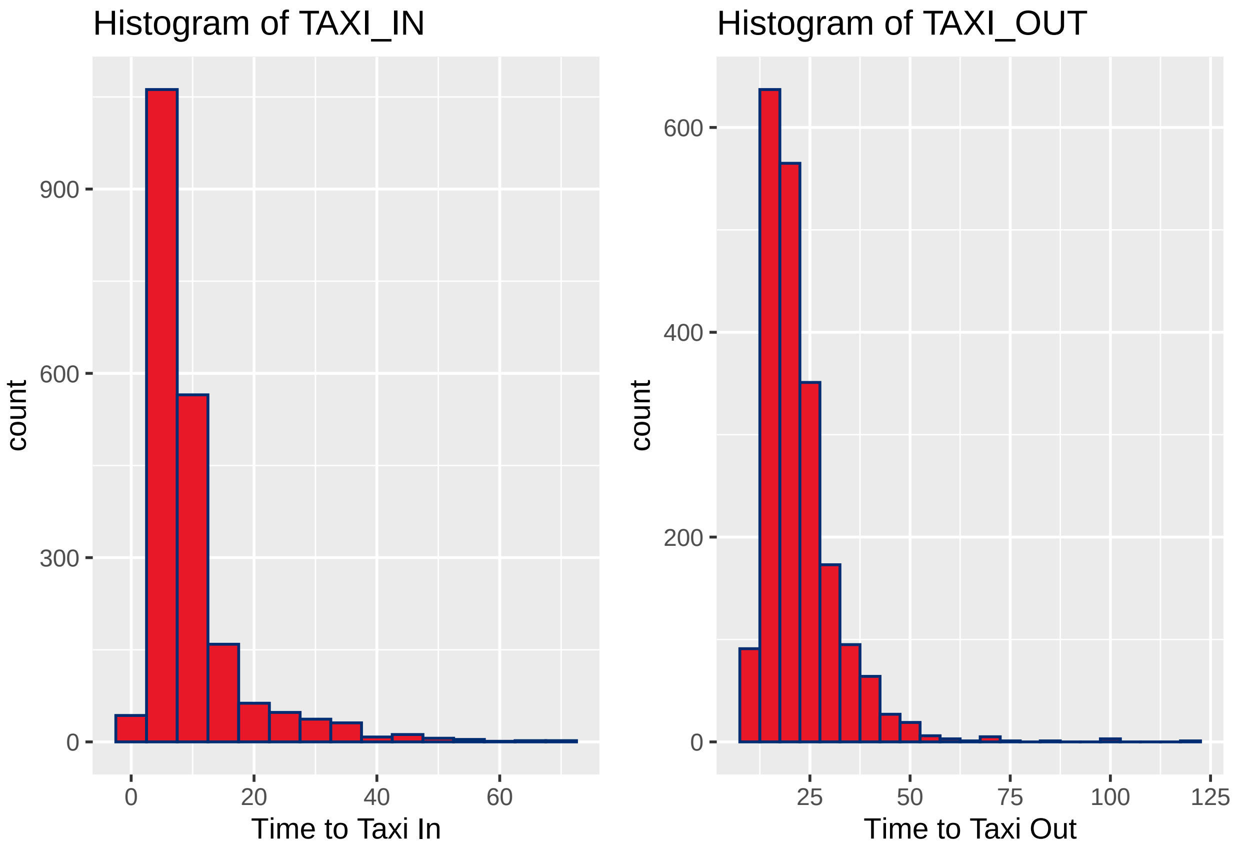
```
## # A tibble: 2,044 x 34
##     YEAR MONTH DAY_OF_MONTH DAY_OF_WEEK FL_DATE    OP_CARRIER TAIL_NUM
##    <dbl> <dbl>        <dbl>       <dbl> <date>     <chr>      <chr>
##  1  2020     1            1           3 2020-01-01 AA         N110AN
##  2  2020     1            2           4 2020-01-02 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
p00 <- ggplot(data = flights, aes(x = TAXI_IN)) +
  geom_histogram(binwidth = 5, fill = "#E81828", color = "#002D72") +
  labs(x = "Time to Taxi In",
       title = "Histogram of TAXI_IN")

p01 <- ggplot(data = flights, aes(x = TAXI_OUT)) +
  geom_histogram(binwidth = 5, fill = "#E81828", color = "#002D72") +
  labs(x = "Time to Taxi Out",
       title = "Histogram of TAXI_OUT")

grid.arrange(p00, p01, nrow = 1)
```

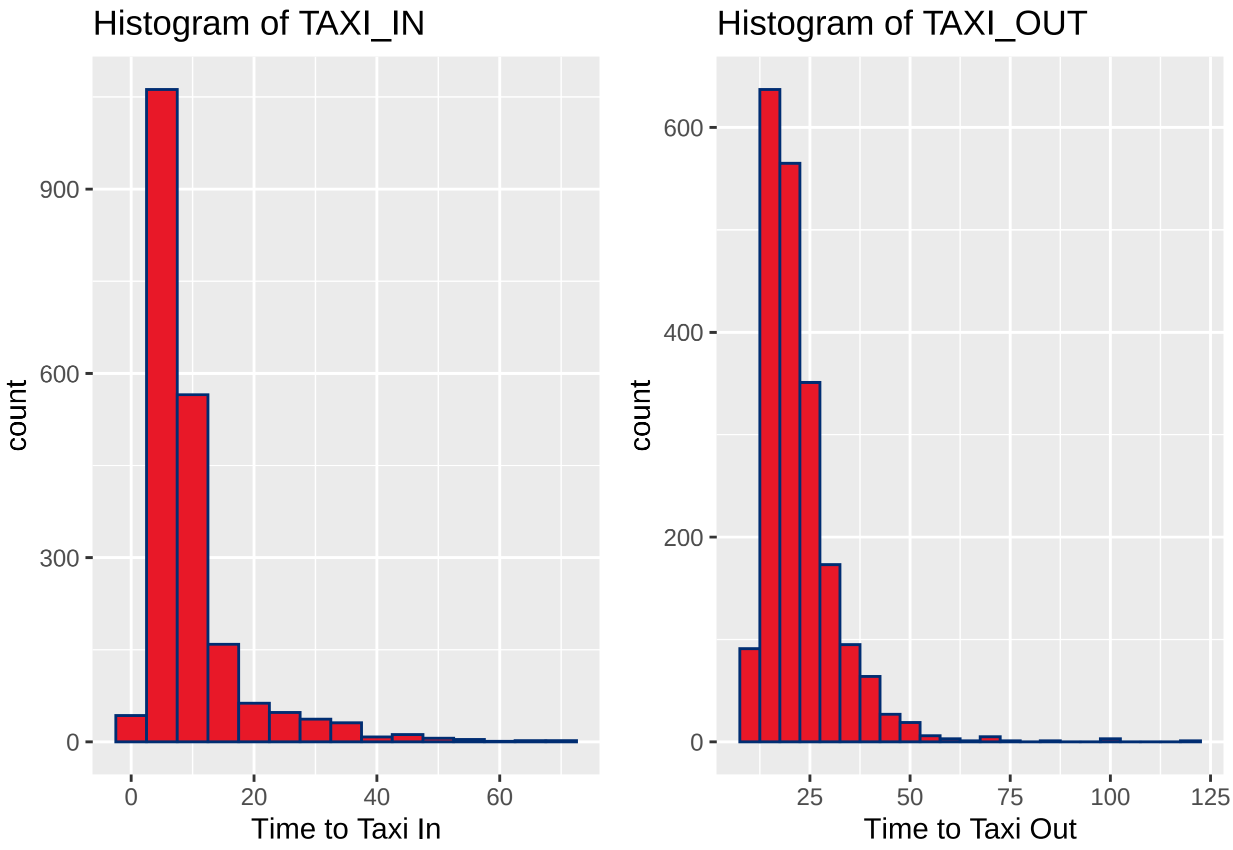
```
## Warning: Removed 1 rows containing non-finite values (stat_bin).

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Days of Month and Week

```
p02 <- ggplot(data = flights, aes(x = DAY_OF_MONTH)) +
  geom_histogram(binwidth = 1, fill = "#E81828", color = "#002D72") +
  labs(x = "Days of Month",
       title = "Histogram of Days of Month")

p03 <- ggplot(data = flights, aes(x = DAY_OF_WEEK)) +
  geom_histogram(binwidth = 1, fill = "#E81828", color = "#002D72") +
  labs(x = "Day of Week",
       title = "Histogram of Days of Week")

grid.arrange(p00, p01, nrow = 1)
```

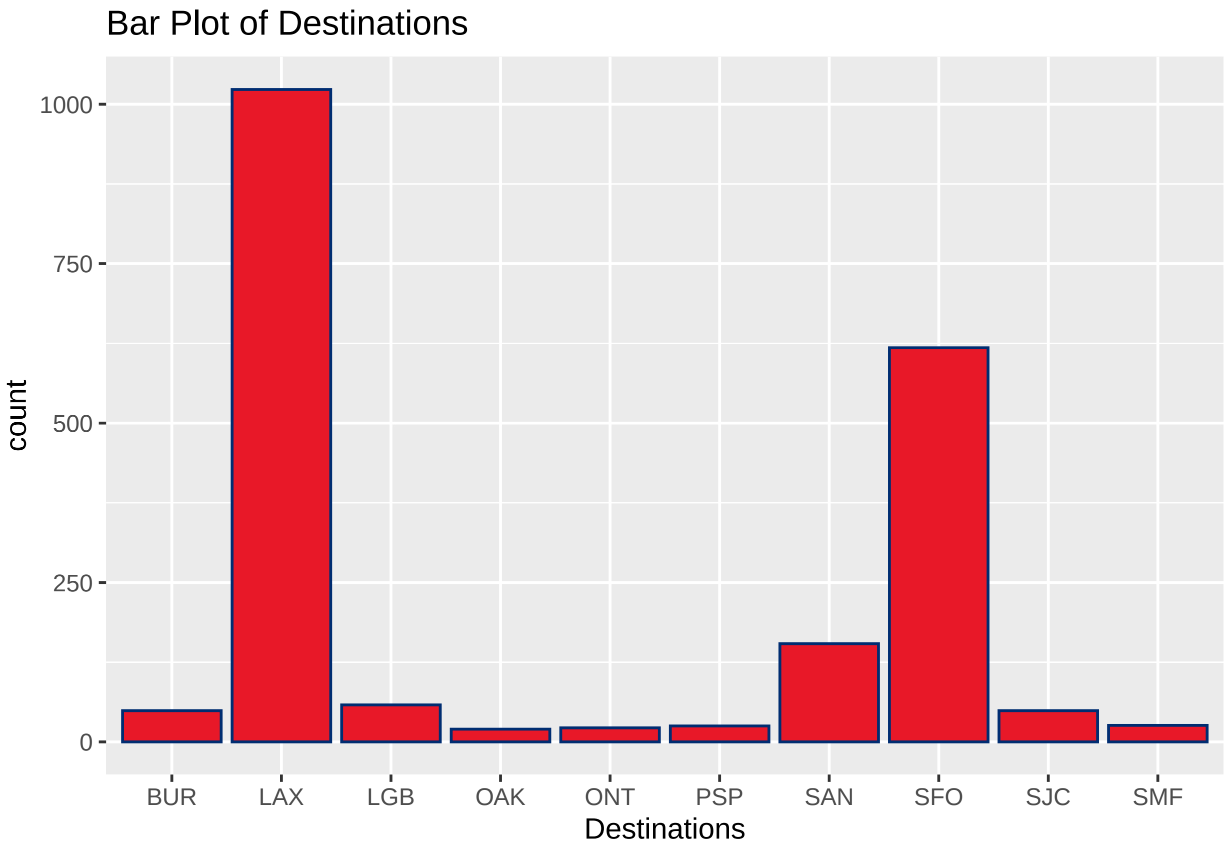
```
## Warning: Removed 1 rows containing non-finite values (stat_bin).

## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Destination Locations

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

```r
ggplot(data = flights, aes(x = DEST)) +
  geom_bar(fill = "#E81828", color = "#002D72") +
  labs(x = "Destinations",
       title = "Bar Plot of Destinations")
```

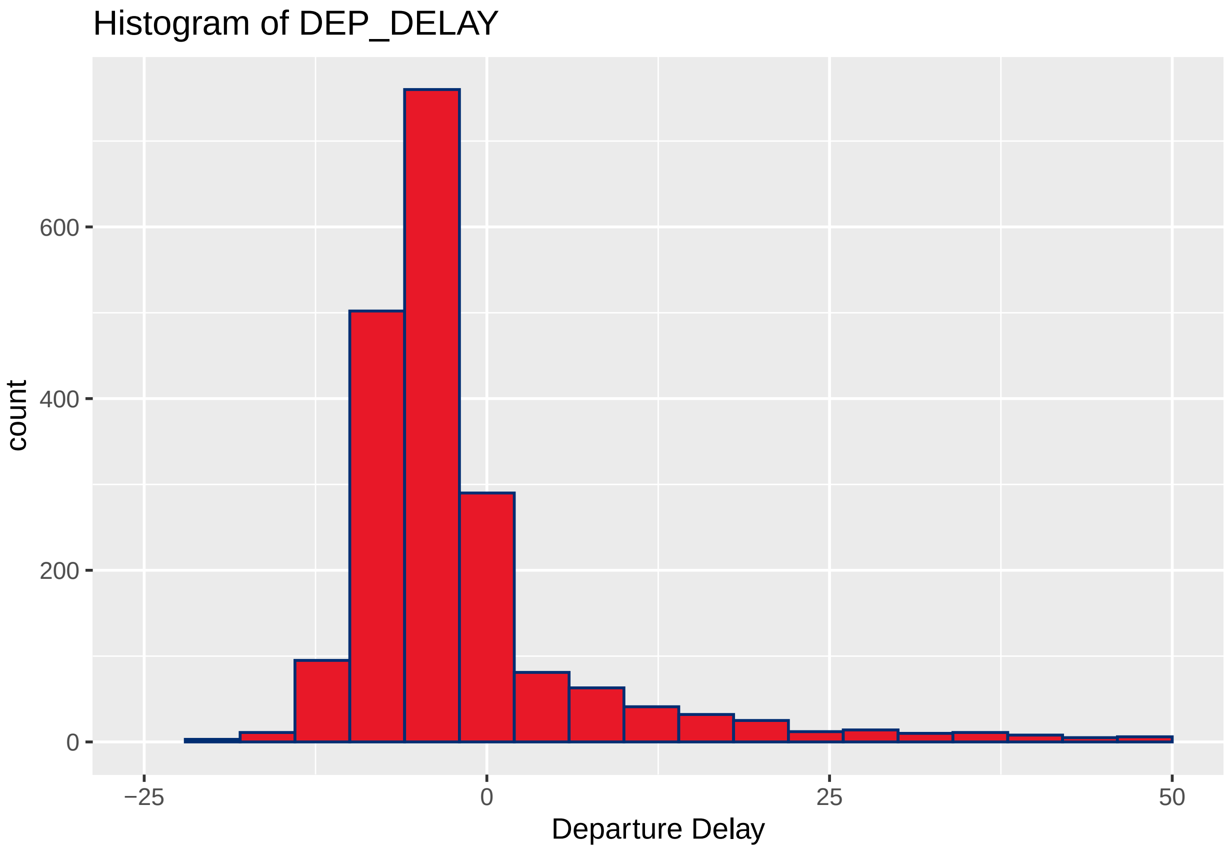


Depart Delay Histogram

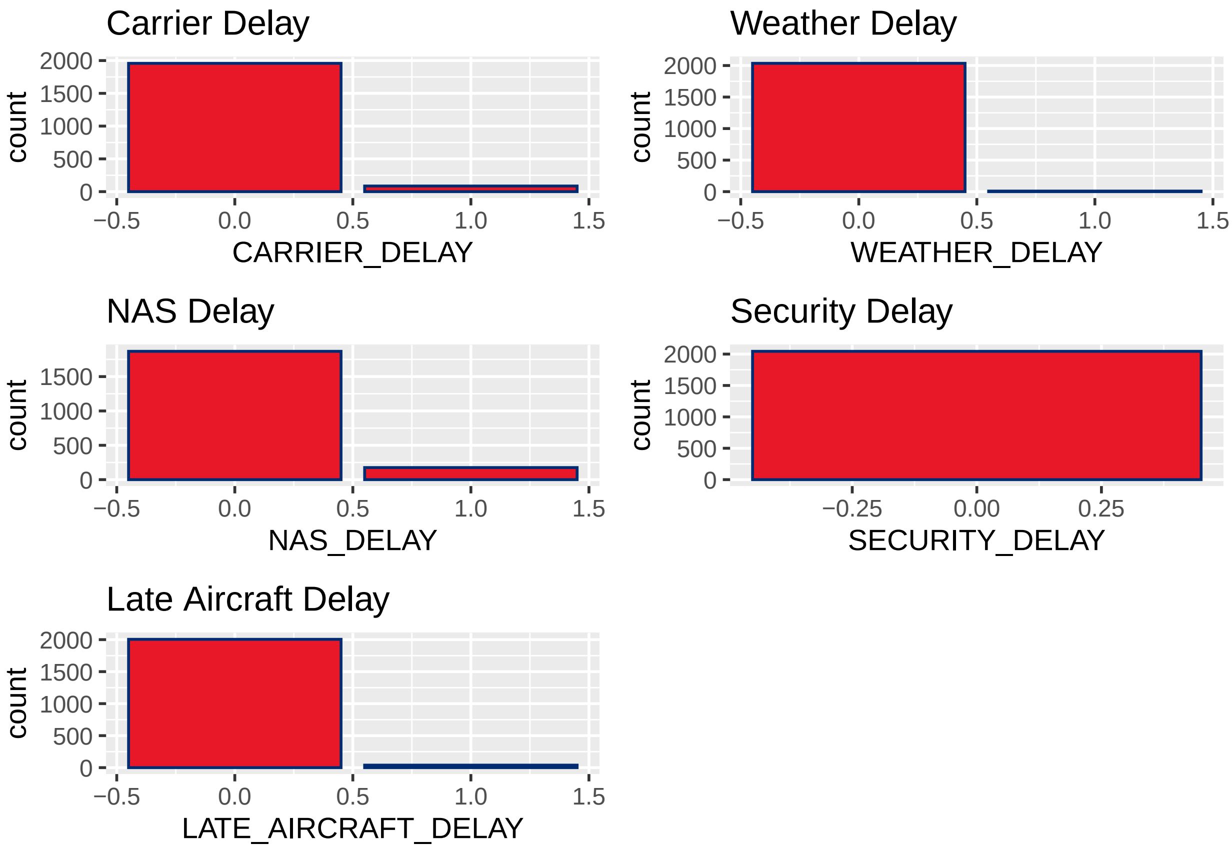
```r
ggplot(data = flights, aes(x = DEP_DELAY)) +
  geom_histogram(binwidth = 4, fill = "#E81828", color = "#002D72") +
  xlim(-25, 50) +
  labs(x = "Departure Delay",
       title = "Histogram of DEP_DELAY")
```

```
## Warning: Removed 75 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



```r
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.

```
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.

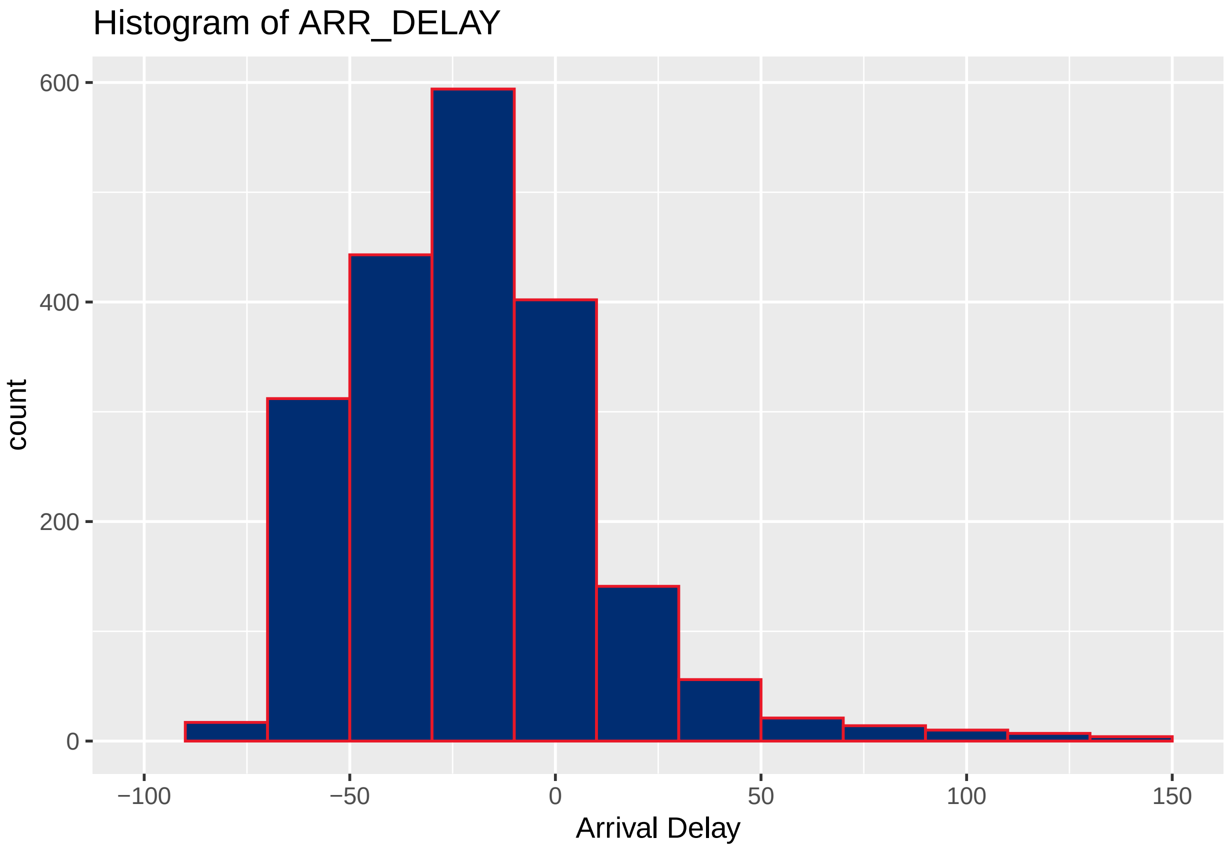
Overall, the categorical delay predictors I would think we could use are: Carrier Delay, NAS Delay, and Late Aircraft Delay

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).

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

```
## Warning: Removed 22 rows containing non-finite values (stat_bin).
## Warning: Removed 1 rows containing missing values (geom_bar).
```

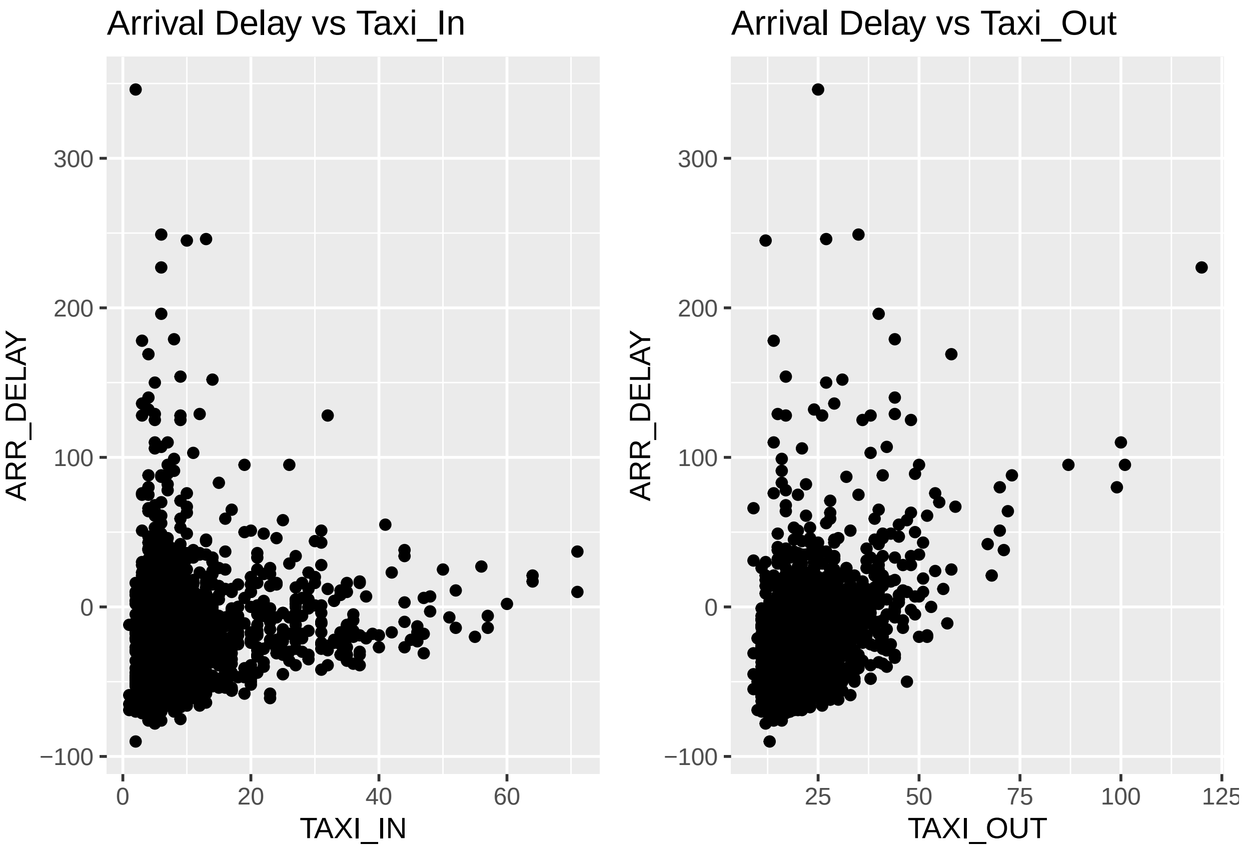


PREDICTORS VS RESPONSE

ARR_DELAY and TAXI_IN / TAXI_OUT

```
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)
```

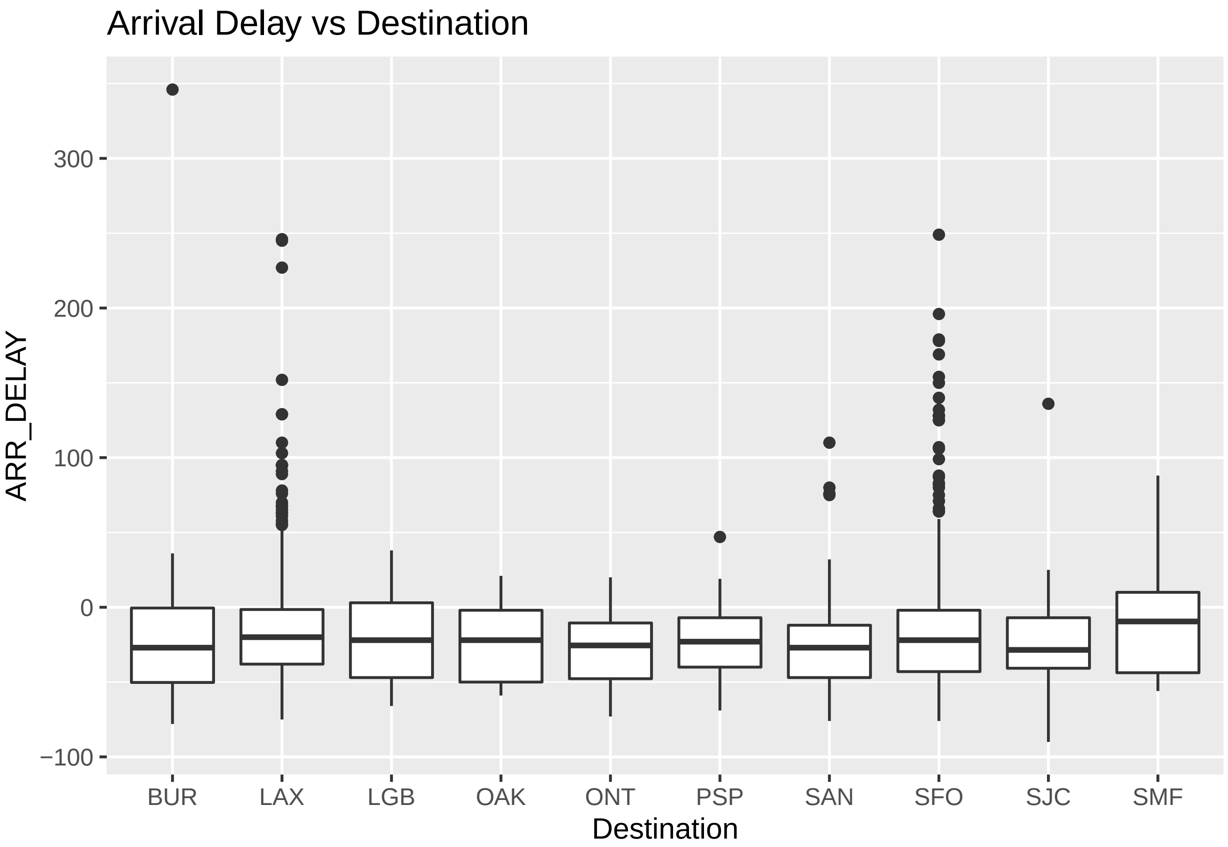
```
## Warning: Removed 11 rows containing missing values (geom_point).

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



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).
```

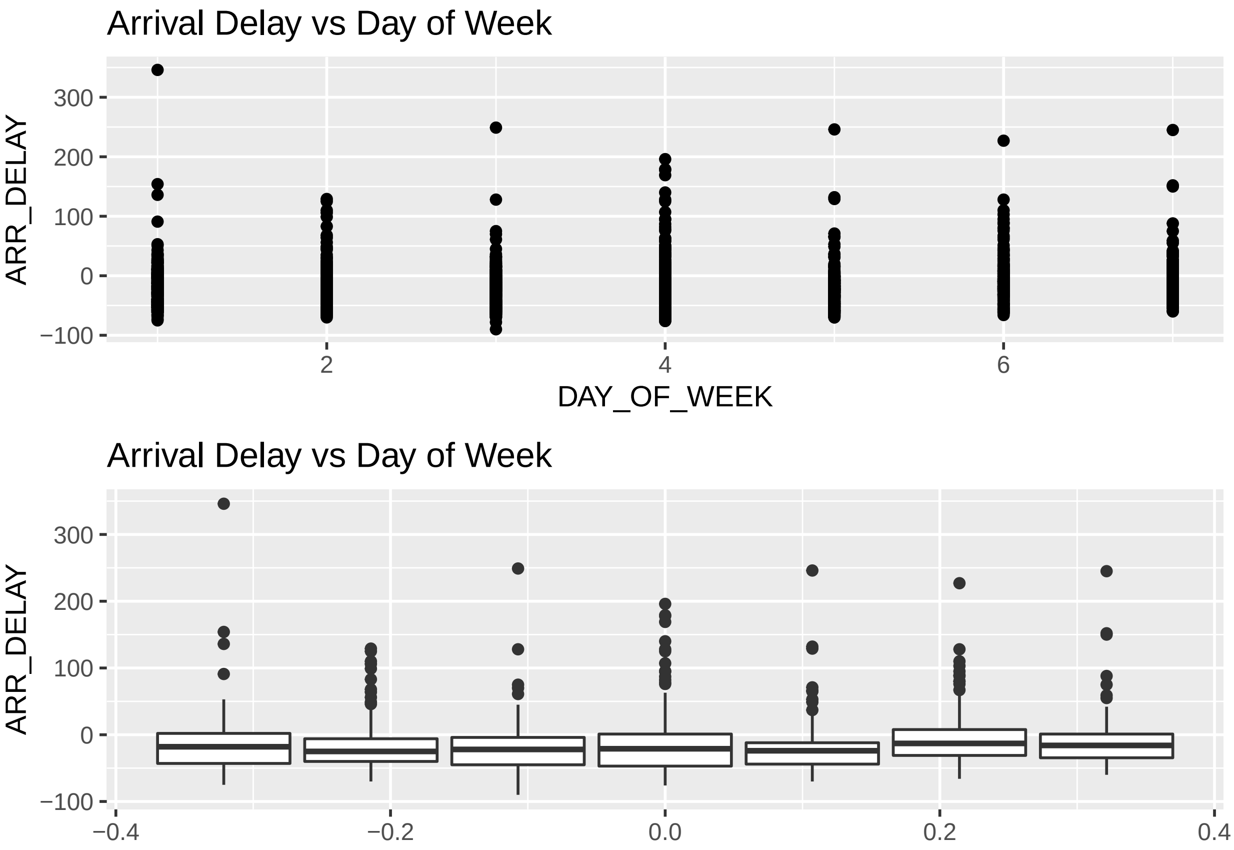


ARR_DELAY and DAY_OF_WEEK

```
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).
```

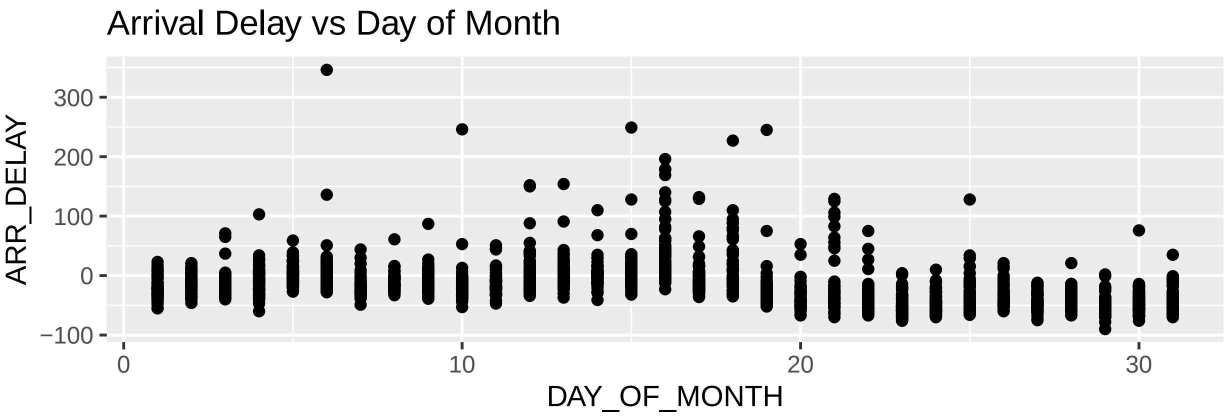
```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```

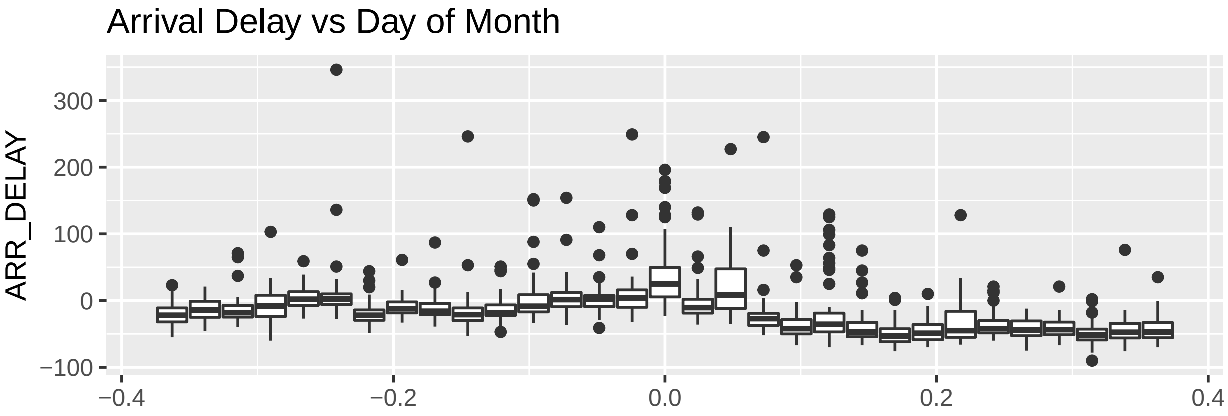


ARR_DELAY and DAY_OF_MONTH

```
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).
```





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

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 = flights)

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 = flights)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -45.272  -9.855  -1.160   9.038  47.085 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         -23.06202    2.38159  -9.683   <2e-16 ***
## DAY_OF_MONTH         -1.34816    0.03581 -37.652   <2e-16 ***
## DAY_OF_WEEK          -0.10808    0.16812  -0.643   0.5204    
## TAXI_IN               0.58755    0.04106  14.309   <2e-16 ***
## TAXI_OUT              0.74925    0.03657  20.487   <2e-16 ***
## DESTLAX               0.69458    2.12940   0.326   0.7443    
## DESTLGB               2.86719    2.80487   1.022   0.3068    
## DESTOAK               0.69036    3.87894   0.178   0.8588    
## DESTONT              -2.90918    3.68260  -0.790   0.4296    
## DESTPSP              -3.04613    3.52845  -0.863   0.3881    
## DESTSAN              -2.05384    2.36601  -0.868   0.3855    
## DESTSFO               0.66839    2.14716   0.311   0.7556    
## DESTSJC              -6.75916    2.98926  -2.261   0.0239 *  
## DESTSMF               5.22127    3.48381   1.499   0.1341    
## DEP_DELAY             0.92577    0.01658  55.847   <2e-16 ***
## CARRIER_DELAY         2.84882    1.87193   1.522   0.1282    
## NAS_DELAY            32.57154    1.31617  24.747   <2e-16 ***
## LATE_AIRCRAFT_DELAY  -1.36546    2.58424  -0.528   0.5973    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14.29 on 2015 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8398, Adjusted R-squared:  0.8384 
## F-statistic: 621.3 on 17 and 2015 DF,  p-value: < 2.2e-16
```

Select Model with AIC

```
library(MASS)
```

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

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

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

```
## 
## Call:
## lm(formula = ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + 
##     DEST + DEP_DELAY + CARRIER_DELAY + NAS_DELAY, data = flights)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -46.708  -9.916  -1.206   9.012  47.046 
## 
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -23.46377    2.29646 -10.217   <2e-16 ***
## DAY_OF_MONTH    -1.34746    0.03575 -37.691   <2e-16 ***
## TAXI_IN          0.58586    0.04100  14.290   <2e-16 ***
## TAXI_OUT         0.74758    0.03646  20.502   <2e-16 ***
## DESTLAX          0.69652    2.12839   0.327    0.744
## DESTLGB          2.85107    2.80358   1.017    0.309
## DESTOAK          0.65335    3.87739   0.169    0.866
## DESTONT         -2.92653    3.68135  -0.795    0.427
## DESTPSP         -3.07497    3.52673  -0.872    0.383
## DESTSAN         -2.07008    2.36509  -0.875    0.382
## DESTSFO          0.67879    2.14637   0.316    0.752
## DESTSJC         -6.74937    2.98824  -2.259    0.024 *
## DESTSMF          5.13431    3.47975   1.475    0.140
## DEP_DELAY        0.92217    0.01521  60.643   <2e-16 ***
## CARRIER_DELAY    2.86554    1.87104   1.532    0.126
## NAS_DELAY       32.56899    1.31525  24.763   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.28 on 2017 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8397, Adjusted R-squared:  0.8385
## F-statistic: 704.5 on 15 and 2017 DF,  p-value: < 2.2e-16
```

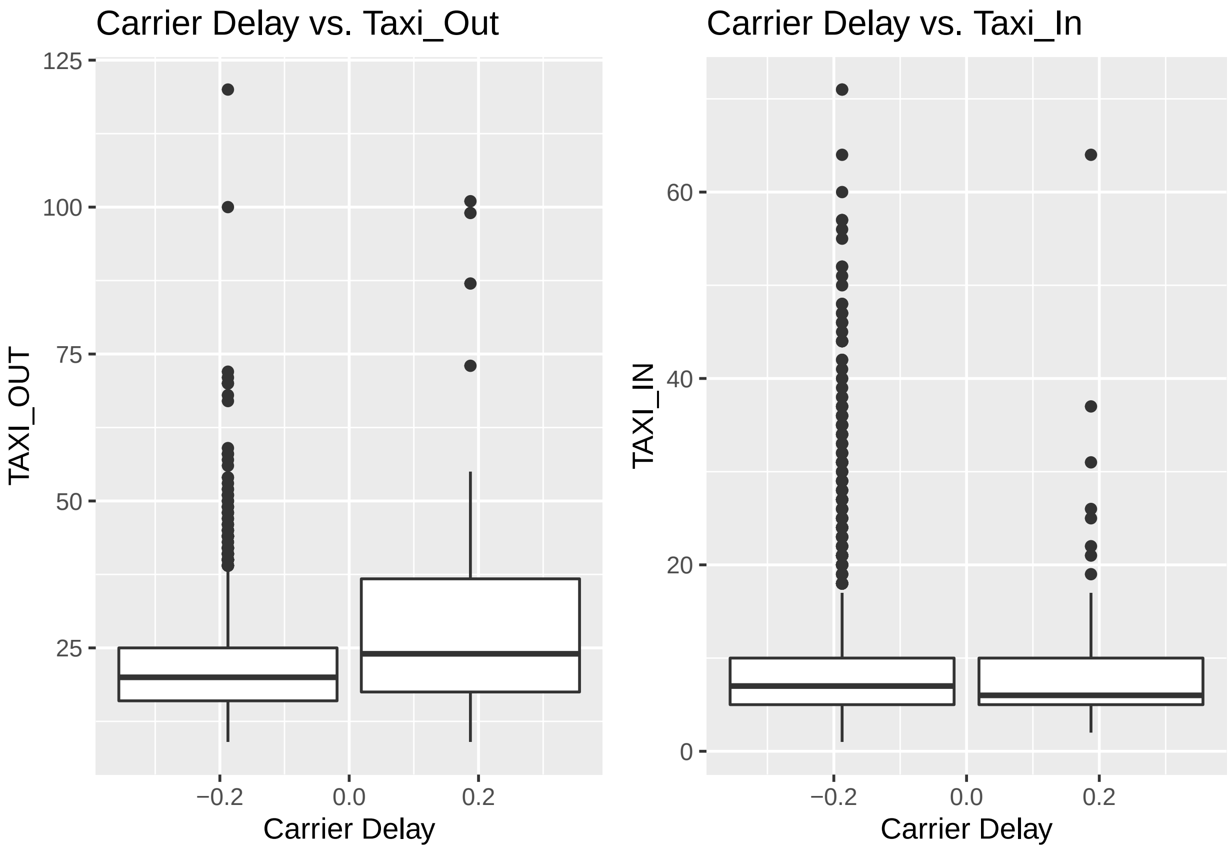
The only variables that were removed were DAY_OF_WEEK and LATE_AIRCRAFT_DELAY. Let's continue using the step_model then.

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.

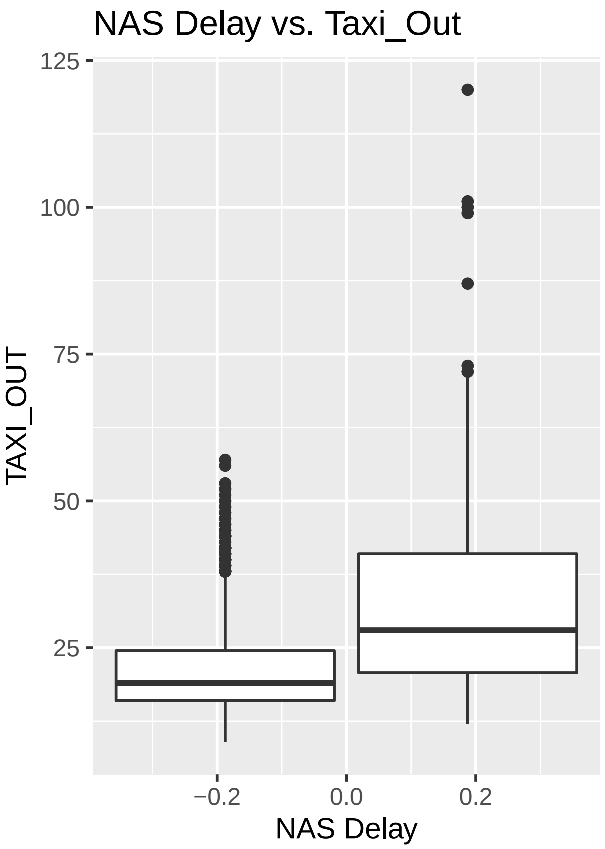
```r
p12 <- ggplot(data = flights, aes(group = CARRIER_DELAY, y = TAXI_OUT)) +
  geom_boxplot() +
  labs(title = "Carrier Delay vs. Taxi_Out",
       x = "Carrier Delay")

p13 <- ggplot(data = flights, 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)
```

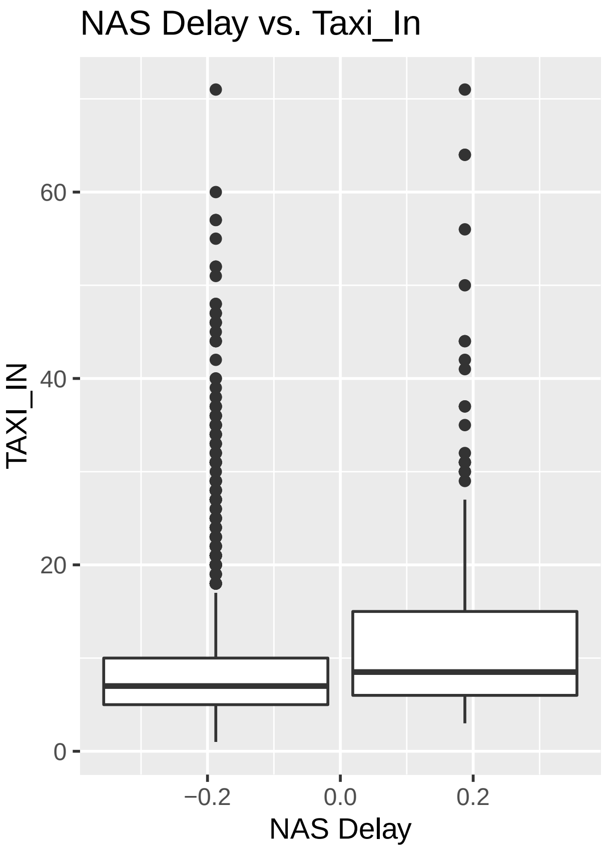
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



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

p15 <- ggplot(data = flights, 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)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```





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

# nas vs taxi in
interaction3 <- lm(ARR_DELAY ~ DAY_OF_MONTH +
                   DAY_OF_WEEK +
                   TAXI_IN +
                   TAXI_OUT +
                   DEST +
                   DEP_DELAY +
                   CARRIER_DELAY +
                   NAS_DELAY +
                  NAS_DELAY*TAXI_IN, data = flights)
```

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

```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEST + DEP_DELAY +
##     CARRIER_DELAY + NAS_DELAY
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + 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   2017 411411
## 2   2015 411194  2    216.96 0.5316 0.5878
```

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

```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEST + DEP_DELAY +
##     CARRIER_DELAY + NAS_DELAY
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + 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   2017 411411
## 2   2015 411252  2    159.04 0.3896 0.6774
```

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

```
## Analysis of Variance Table
##
## Model 1: ARR_DELAY ~ DAY_OF_MONTH + TAXI_IN + TAXI_OUT + DEST + DEP_DELAY +
##     CARRIER_DELAY + NAS_DELAY
## Model 2: ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + 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   2017 411411
## 2   2015 408708  2    2703.1 6.6634 0.001305 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

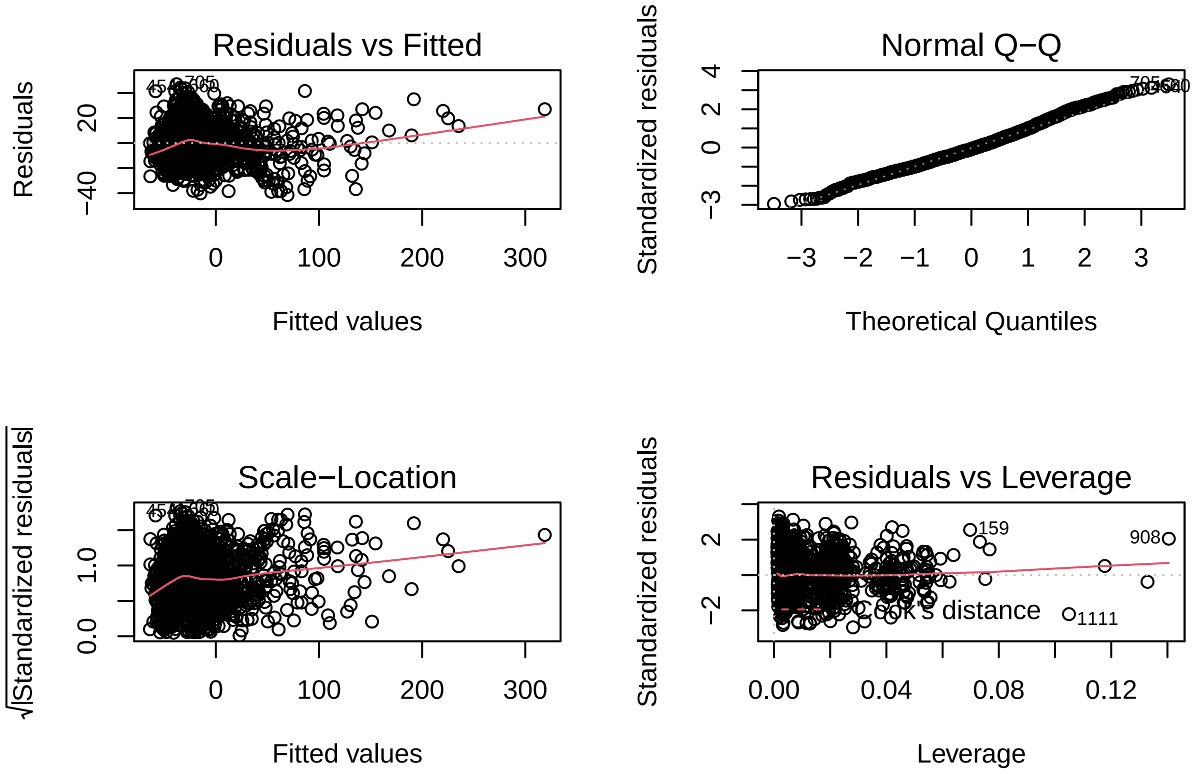
It actually seems that interaction3: NAS_DELAY and TAXI_IN is the only interaction that is statistically significant in predicting ARR_DELAY. Let's make this model our current model:

Final Linear Model

```
current_model <- interaction3

summary(current_model)
```

```
##
## Call:
## lm(formula = ARR_DELAY ~ DAY_OF_MONTH + DAY_OF_WEEK + TAXI_IN +
##     TAXI_OUT + DEST + DEP_DELAY + CARRIER_DELAY + NAS_DELAY +
##     NAS_DELAY * TAXI_IN, data = flights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.499  -9.698  -1.100   8.842  47.047
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -23.16948    2.37418  -9.759  < 2e-16 ***
## DAY_OF_MONTH      -1.35728    0.03577 -37.945  < 2e-16 ***
## DAY_OF_WEEK       -0.08743    0.16753  -0.522 0.601810
## TAXI_IN            0.65452    0.04503  14.536  < 2e-16 ***
## TAXI_OUT           0.73937    0.03656  20.221  < 2e-16 ***
## DESTLAX            0.42359    2.12379   0.199 0.841929
## DESTLGB            2.58596    2.79668   0.925 0.355257
## DESTOAK            0.56652    3.86702   0.147 0.883541
## DESTONT           -2.83203    3.67121  -0.771 0.440551
## DESTPSP           -3.21114    3.51775  -0.913 0.361436
## DESTSAN           -2.08223    2.35848  -0.883 0.377412
## DESTSFO            0.37092    2.14204   0.173 0.862541
## DESTSJC           -7.02861    2.98091  -2.358 0.018475 *
## DESTSMF            4.87675    3.47087   1.405 0.160160
## DEP_DELAY          0.91791    0.01521  60.332  < 2e-16 ***
## CARRIER_DELAY      2.82439    1.86585   1.514 0.130252
## NAS_DELAY         37.14992    1.83308  20.266  < 2e-16 ***
## TAXI_IN:NAS_DELAY  -0.35216    0.09814  -3.588 0.000341 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.24 on 2015 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.8408, Adjusted R-squared:  0.8394
## F-statistic: 625.9 on 17 and 2015 DF,  p-value: < 2.2e-16
```

```
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.