

# AirlineDelayAnalysis

2023-05-25

## Loading necessary libraries

```
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##       filter, lag

## The following objects are masked from 'package:base':
##       intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(mosaic)

## Warning: package 'mosaic' was built under R version 4.2.3

## Registered S3 method overwritten by 'mosaic':
##   method           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features. The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##       mean
```

```

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

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum

```

## Loading data and Summarization for Airlines Delay

```

airlines_data <- read.csv("Airline_Delay_Cause.csv")
summary(airlines_data)

```

```

##      year        month      carrier      carrier_name
##  Min.   :2023   Min.   :1    Length:4612      Length:4612
##  1st Qu.:2023  1st Qu.:1    Class :character  Class :character
##  Median :2023  Median :2    Mode  :character  Mode  :character
##  Mean   :2023  Mean   :2
##  3rd Qu.:2023 3rd Qu.:3
##  Max.   :2023  Max.   :3

##
##      airport        airport_name      arr_flights      arr_del15
##  Length:4612      Length:4612      Min.   : 1.0  Min.   : 0.00
##  Class :character  Class :character  1st Qu.: 47.0  1st Qu.: 8.00
##  Mode  :character  Mode  :character  Median : 93.0  Median : 20.00
##                           Mean   : 352.1  Mean   : 75.04
##                           3rd Qu.: 237.0  3rd Qu.: 52.00
##                           Max.   :18504.0  Max.   :3529.00
##                           NA's   :5      NA's   :5
##      carrier_ct      weather_ct      nas_ct      security_ct
##  Min.   : 0.00  Min.   : 0.000  Min.   : 0.00  Min.   : 0.0000
##  1st Qu.: 2.71  1st Qu.: 0.000  1st Qu.: 1.00  1st Qu.: 0.0000
##  Median : 7.67  Median : 0.640  Median : 4.01  Median : 0.0000
##  Mean   : 25.14  Mean   : 2.474  Mean   : 21.44  Mean   : 0.2244
##  3rd Qu.: 20.65  3rd Qu.: 2.000  3rd Qu.: 12.80  3rd Qu.: 0.0000
##  Max.   :904.87  Max.   :163.000  Max.   :999.84  Max.   :14.5500
##  NA's   :5       NA's   :5      NA's   :5      NA's   :5
##      late_aircraft_ct  arr_cancelled  arr_diverted  arr_delay
##  Min.   : 0.000  Min.   : 0.0  Min.   : 0.0000  Min.   : 0
##  1st Qu.: 1.705  1st Qu.: 0.0  1st Qu.: 0.0000  1st Qu.: 436
##  Median : 5.450  Median : 1.0  Median : 0.0000  Median : 1206
##  Mean   : 25.772  Mean   : 5.8  Mean   : 0.8055  Mean   : 4997

```

```

## 3rd Qu.: 16.655 3rd Qu.: 4.0 3rd Qu.: 1.0000 3rd Qu.: 3295
## Max. :1550.650 Max. :580.0 Max. :84.0000 Max. :337375
## NA's :5 NA's :5 NA's :5 NA's :5
## carrier_delay weather_delay nas_delay security_delay
## Min. : 0 Min. : 0.0 Min. : 0 Min. : 0.00
## 1st Qu.: 136 1st Qu.: 0.0 1st Qu.: 34 1st Qu.: 0.00
## Median : 437 Median : 30.0 Median : 159 Median : 0.00
## Mean : 1802 Mean : 280.6 Mean : 1002 Mean : 10.78
## 3rd Qu.: 1306 3rd Qu.: 182.0 3rd Qu.: 532 3rd Qu.: 0.00
## Max. :118554 Max. :20423.0 Max. :50792 Max. :1477.00
## NA's :5 NA's :5 NA's :5 NA's :5
## late_aircraft_delay
## Min. : 0
## 1st Qu.: 89
## Median : 369
## Mean : 1901
## 3rd Qu.: 1254
## Max. :155262
## NA's :5

```

```
print(length(unique(airlines_data$carrier)))
```

```
## [1] 15
```

```
print(length(unique(airlines_data$carrier_name)))
```

```
## [1] 15
```

```
print(length(unique(airlines_data$airport)))
```

```
## [1] 341
```

```
print(length(unique(airlines_data$airport_name)))
```

```
## [1] 341
```

```
print(nrow(airlines_data))
```

```
## [1] 4612
```

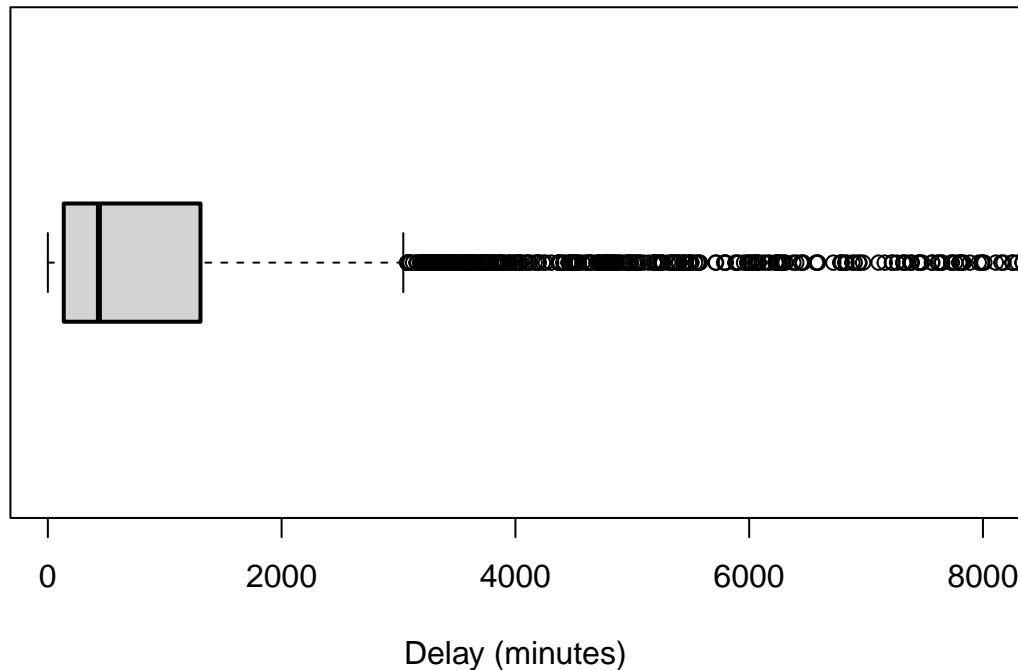
```
print(ncol(airlines_data))
```

```
## [1] 21
```

## Box Plots

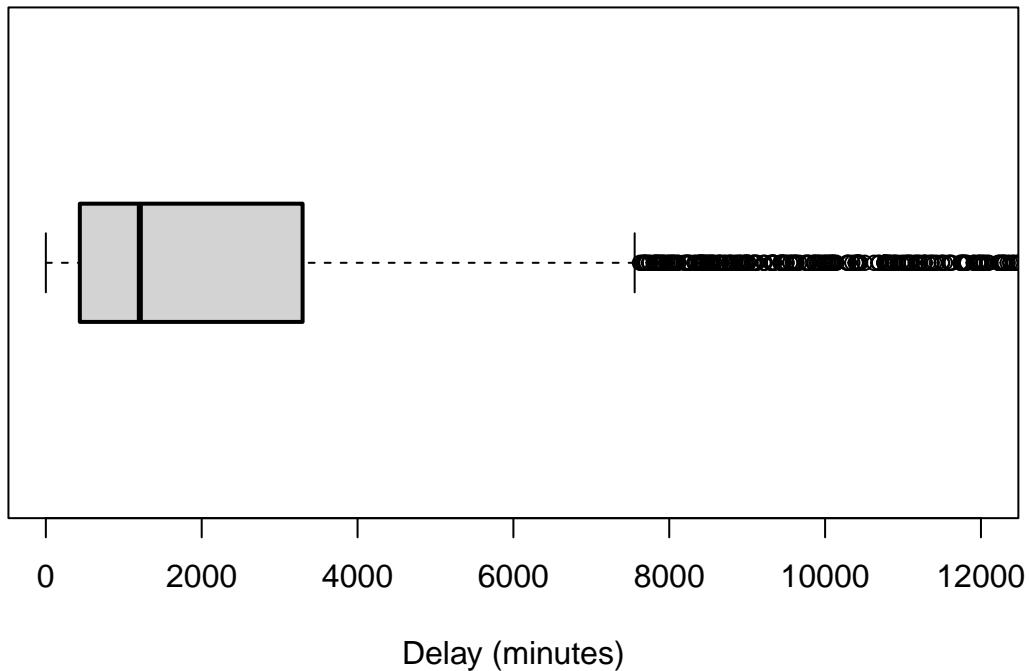
```
# Box Plot - carrier_delay  
boxplot(airlines_data$carrier_delay, main = "Distribution of Carrier Delays", xlab = "Delay (minutes)"  
       , boxlwd = 2, outwex = 0.5, boxwex = 0.5 , outline = TRUE , horizontal = TRUE ,  
       ylim=c(0,8000))
```

## Distribution of Carrier Delays



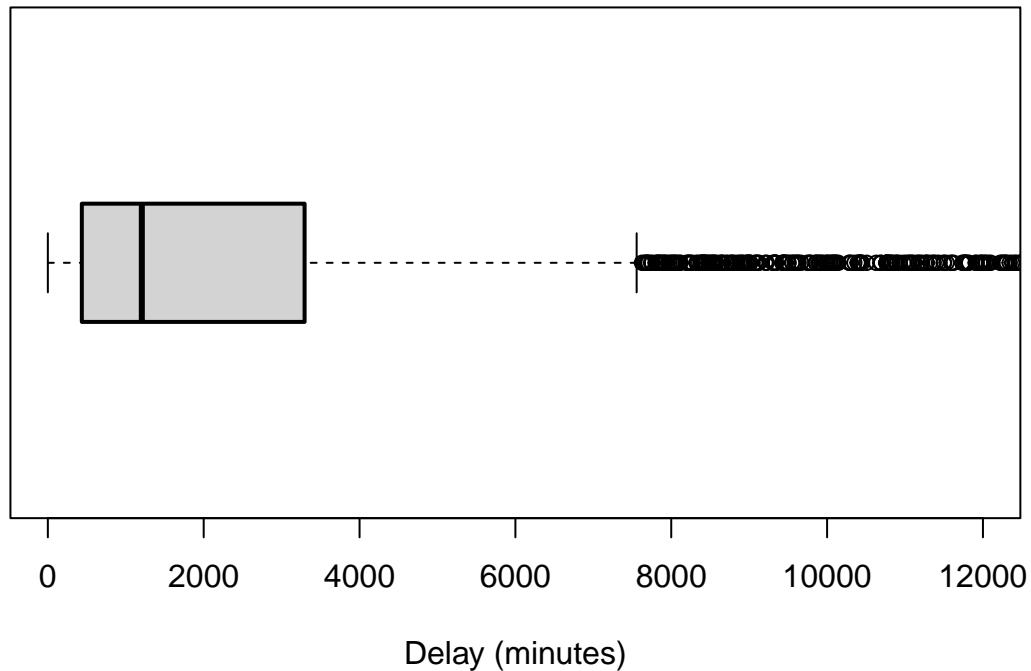
```
# Box Plot - arr_delay  
boxplot(airlines_data$arr_delay, main = "Distribution of Arrival Delays", xlab = "Delay (minutes)" ,  
       boxlwd = 2, outwex = 0.5, boxwex = 0.5 , outline = TRUE , horizontal = TRUE ,  
       ylim=c(0,12000))
```

## Distribution of Arrival Delays



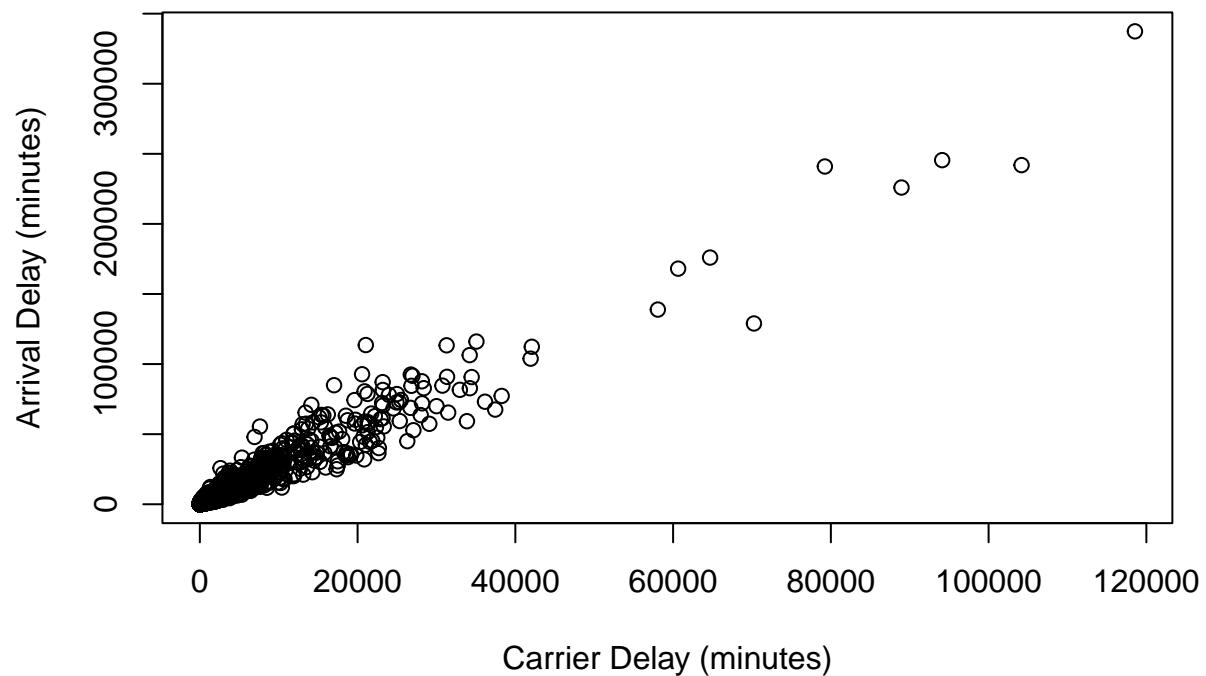
```
# Box Plot - arr_delay
boxplot(airlines_data$arr_delay, main = "Distribution of Weather Delays", xlab = "Delay (minutes)" ,
        boxlwd = 2, outwex = 0.5, boxwex = 0.5 , outline = TRUE , horizontal = TRUE ,
        ylim=c(0,12000))
```

## Distribution of Weather Delays



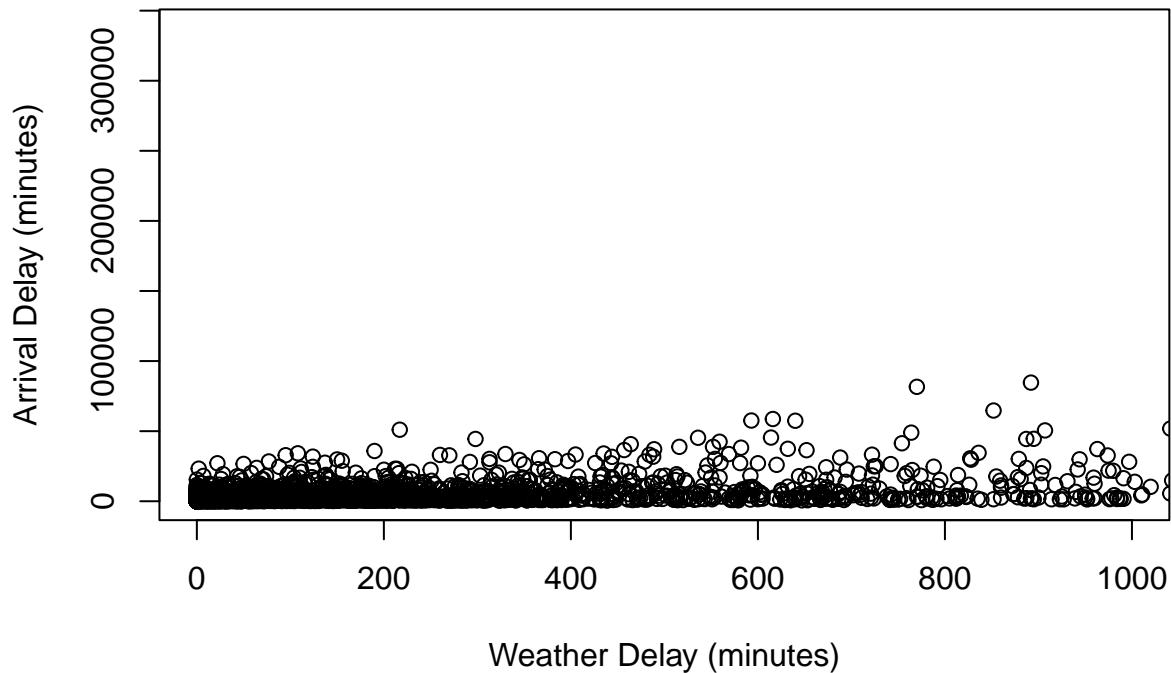
```
# Scatter Plot
plot(airlines_data$carrier_delay, airlines_data$arr_delay,
      main = "Arrival Delay vs Carrier Delay",
      xlab = "Carrier Delay (minutes)", ylab = "Arrival Delay (minutes)")
```

## Arrival Delay vs Carrier Delay



```
# Scatter Plot
plot(airlines_data$weather_delay, airlines_data$arr_delay,
      main = "Arrival Delay vs Weather Delay",
      xlab = "Weather Delay (minutes)", ylab = "Arrival Delay (minutes)",
      xlim = c(0,1000))
```

## Arrival Delay vs Weather Delay



## Linear Regression

```
# Select the relevant columns for the linear regression analysis
airlines_data_1m <- airlines_data %>%
  select (arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay)

# Fit a linear regression model
arr_carrier_model <- lm(arr_delay ~ carrier_delay , data = airlines_data_1m)
arr_weather_model <- lm(arr_delay ~ weather_delay , data = airlines_data_1m)
arr_nas_model <- lm(arr_delay ~ nas_delay , data = airlines_data_1m)
arr_security_model <- lm(arr_delay ~ security_delay , data = airlines_data_1m)
arr_aircraft_model <- lm(arr_delay ~ late_aircraft_delay , data = airlines_data_1m)

#Print the model summary
print(summary(arr_carrier_model))

## 
## Call:
## lm(formula = arr_delay ~ carrier_delay, data = airlines_data_1m)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -57616    -548    -191     223   57544 
## 
```

```

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 215.55778  54.86562  3.929 8.66e-05 ***
## carrier_delay   2.65278   0.00974 272.365 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3528 on 4605 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9415
## F-statistic: 7.418e+04 on 1 and 4605 DF, p-value: < 2.2e-16

```

```
print(summary (arr_weather_model))
```

```

##
## Call:
## lm(formula = arr_delay ~ weather_delay, data = airlines_data_1m)
##
## Residuals:
##      Min       1Q     Median       3Q       Max
## -110151    -2062    -1577     -206   172750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2095.2375   160.0484   13.09   <2e-16 ***
## weather_delay   10.3403    0.1562   66.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10450 on 4605 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.4875, Adjusted R-squared:  0.4874
## F-statistic:  4380 on 1 and 4605 DF, p-value: < 2.2e-16

```

```
print (summary (arr_nas_model))
```

```

##
## Call:
## lm(formula = arr_delay ~ nas_delay, data = airlines_data_1m)
##
## Residuals:
##      Min       1Q     Median       3Q       Max
## -70239    -1025    -718     176  145597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 967.63751  107.92052   8.966   <2e-16 ***
## nas_delay      4.02231    0.03233 124.428   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6988 on 4605 degrees of freedom
##   (5 observations deleted due to missingness)

```

```

## Multiple R-squared:  0.7707, Adjusted R-squared:  0.7707
## F-statistic: 1.548e+04 on 1 and 4605 DF,  p-value: < 2.2e-16

print (summary (arr_security_model))

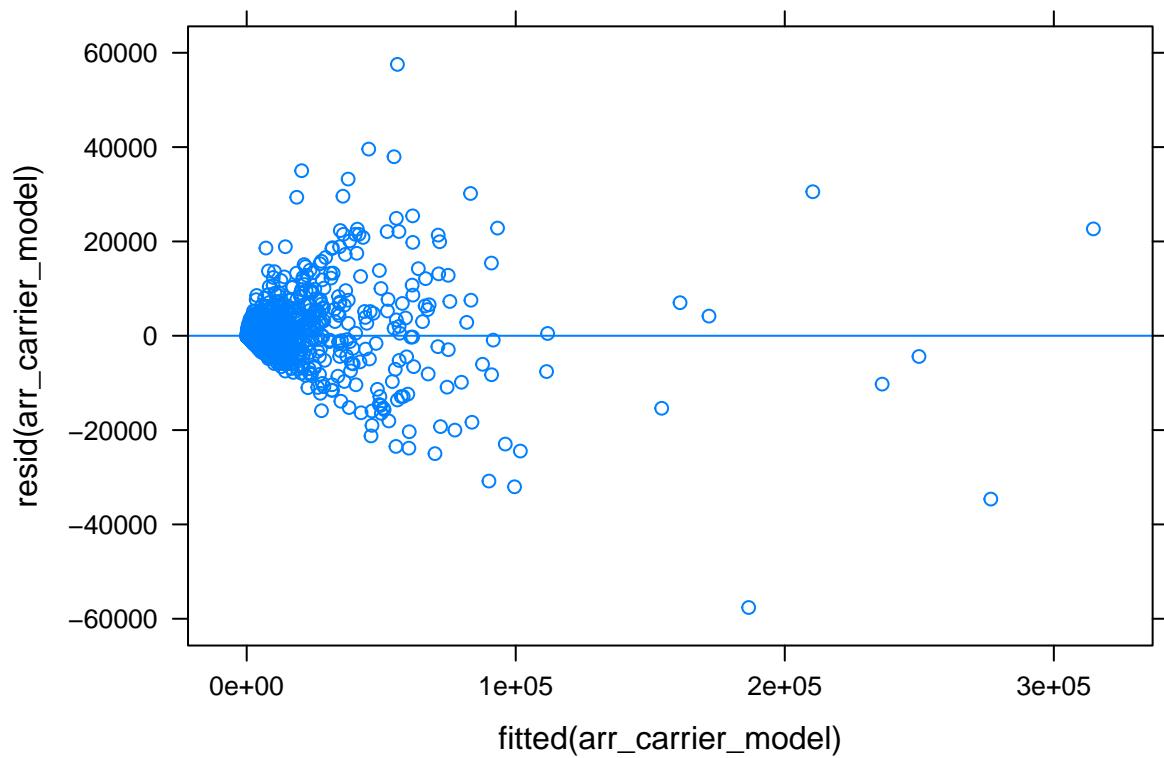
## 
## Call:
## lm(formula = arr_delay ~ security_delay, data = airlines_data_1m)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -188983    -3170    -2502     -806   276897 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3471.691   195.605   17.75 <2e-16 ***
## security_delay 141.454     4.009   35.28 <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12950 on 4605 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.2128, Adjusted R-squared:  0.2126
## F-statistic:  1245 on 1 and 4605 DF,  p-value: < 2.2e-16

print (summary (arr_aircraft_model))

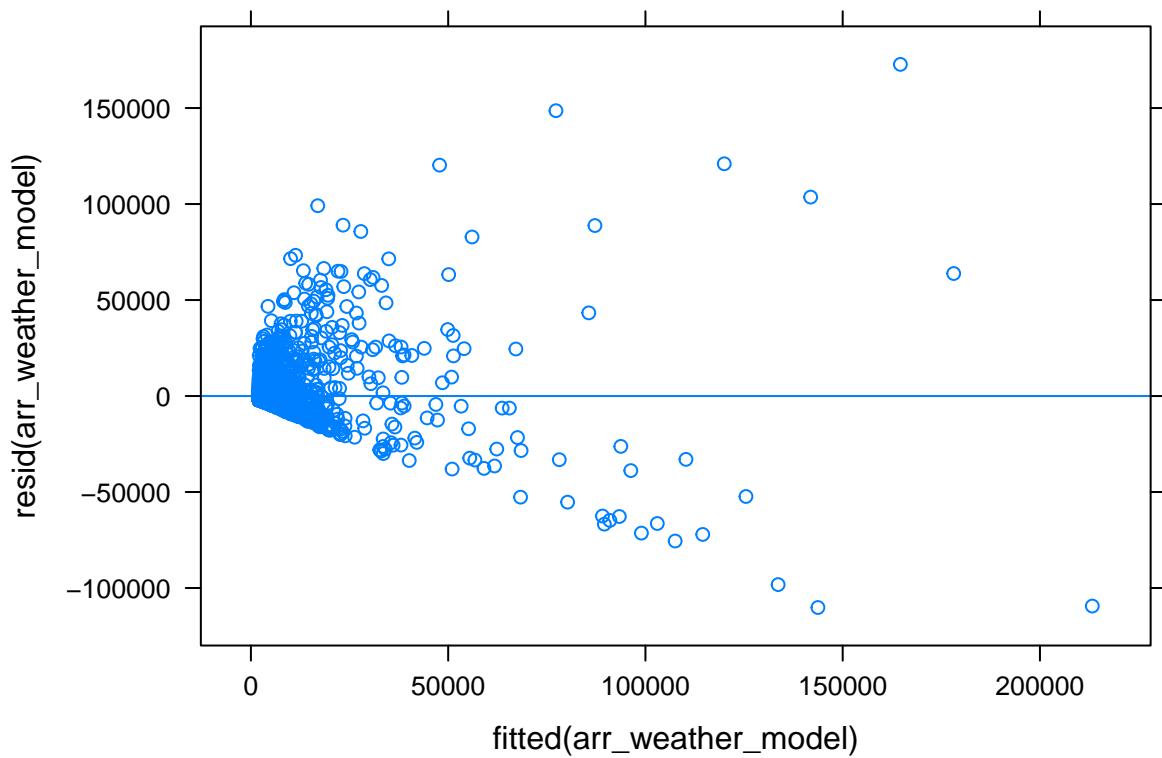
## 
## Call:
## lm(formula = arr_delay ~ late_aircraft_delay, data = airlines_data_1m)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -26529    -672    -416     113   50317 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 547.24086   51.70081   10.59 <2e-16 *** 
## late_aircraft_delay 2.34028    0.00813   287.85 <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3349 on 4605 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.9473, Adjusted R-squared:  0.9473
## F-statistic: 8.286e+04 on 1 and 4605 DF,  p-value: < 2.2e-16

xyplot(resid(arr_carrier_model) ~ fitted(arr_carrier_model), data=airlines_data_1m, type=c("p", "r"))

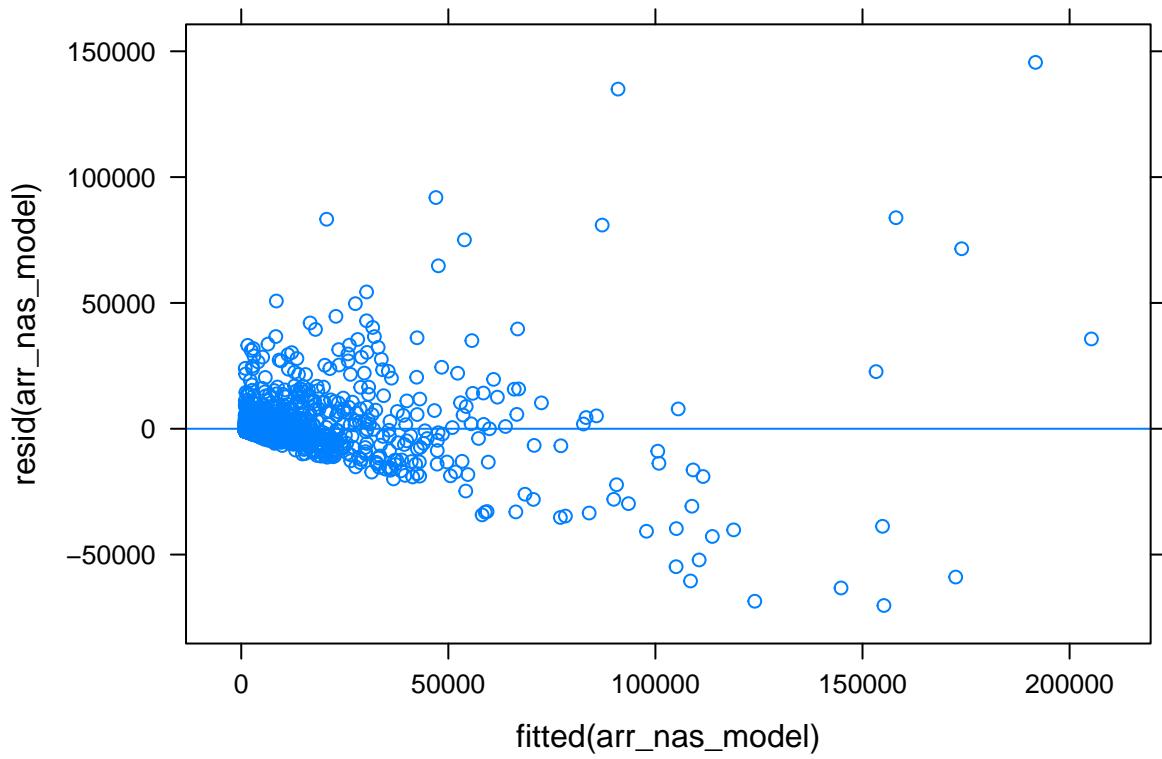
```



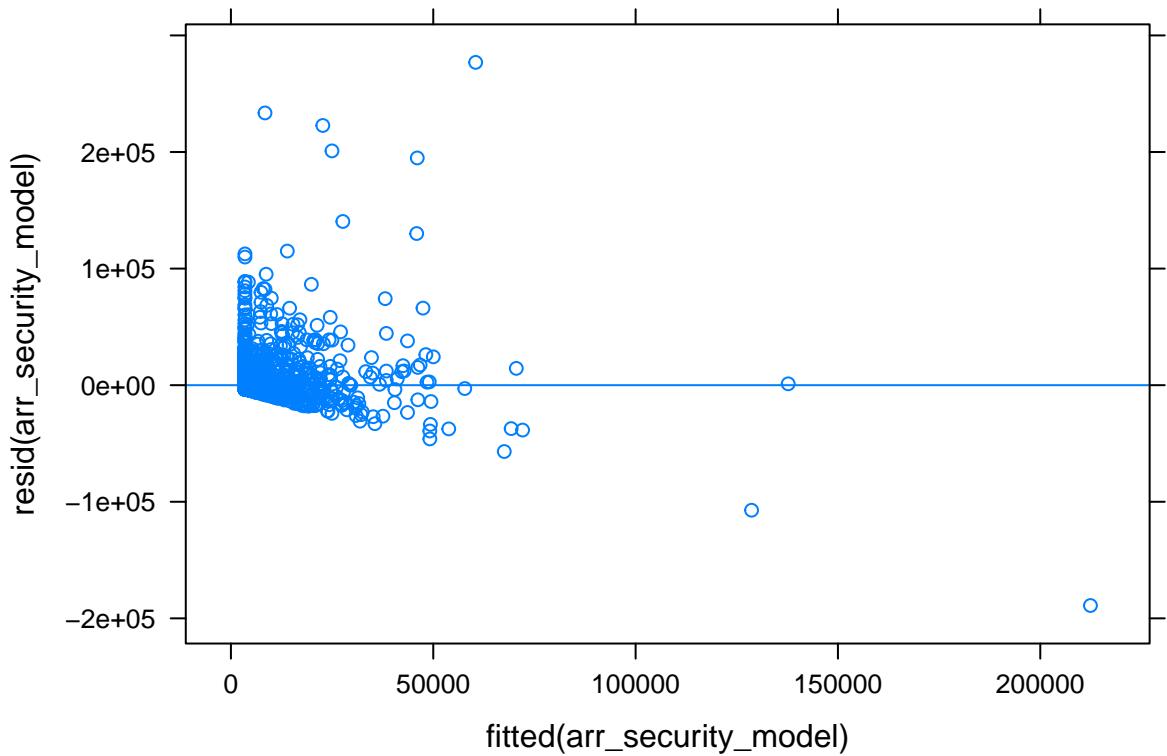
```
xyplot(resid(arr_weather_model) ~ fitted(arr_weather_model), data=airlines_data_1m, type=c("p", "r"))
```



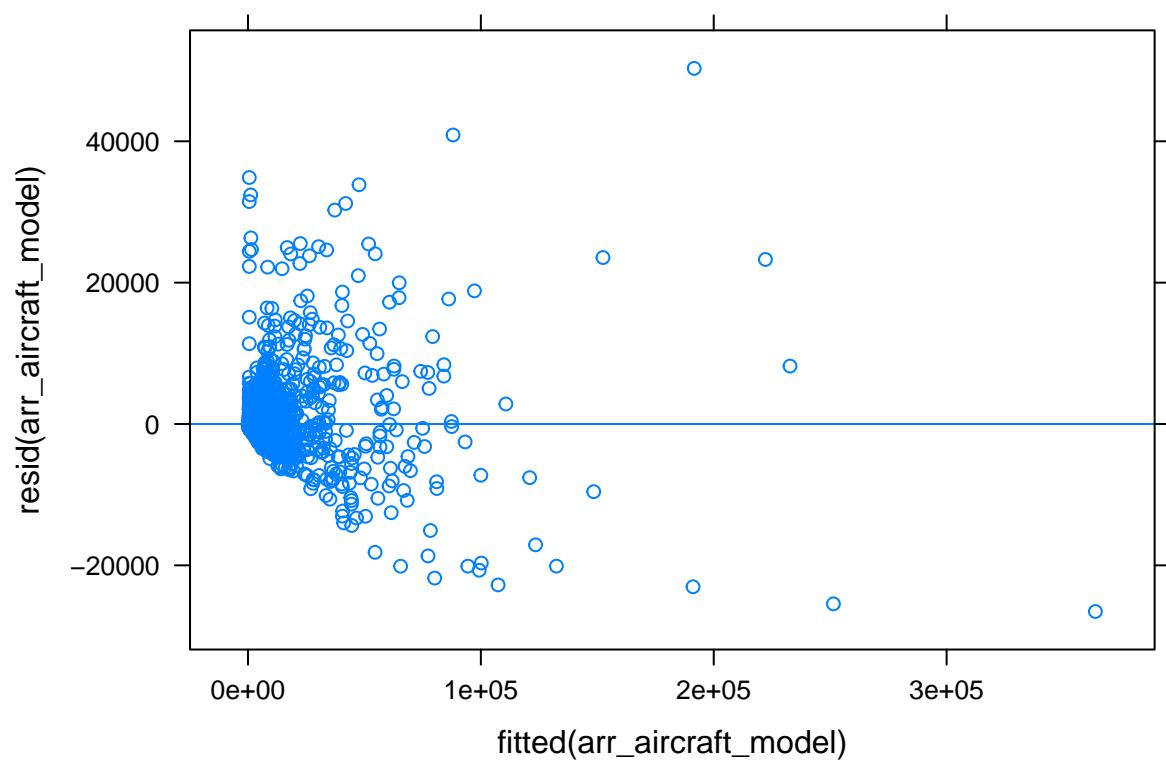
```
xyplot(resid(arr_nas_model) ~ fitted(arr_nas_model), data=airlines_data_1m, type=c("p", "r"))
```



```
xyplot(resid(arr_security_model) ~ fitted(arr_security_model), data=airlines_data_1m, type=c("p", "r"))
```

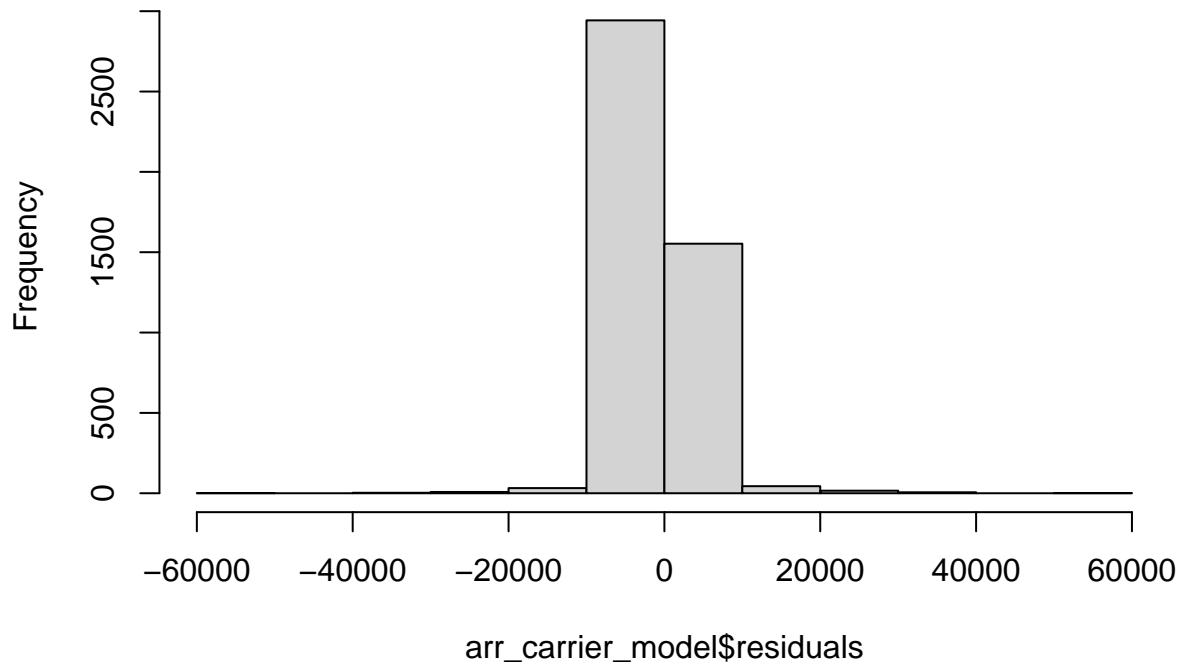


```
xyplot(resid(arr_aircraft_model) ~ fitted(arr_aircraft_model), data=airlines_data_1m, type=c("p", "r"))
```



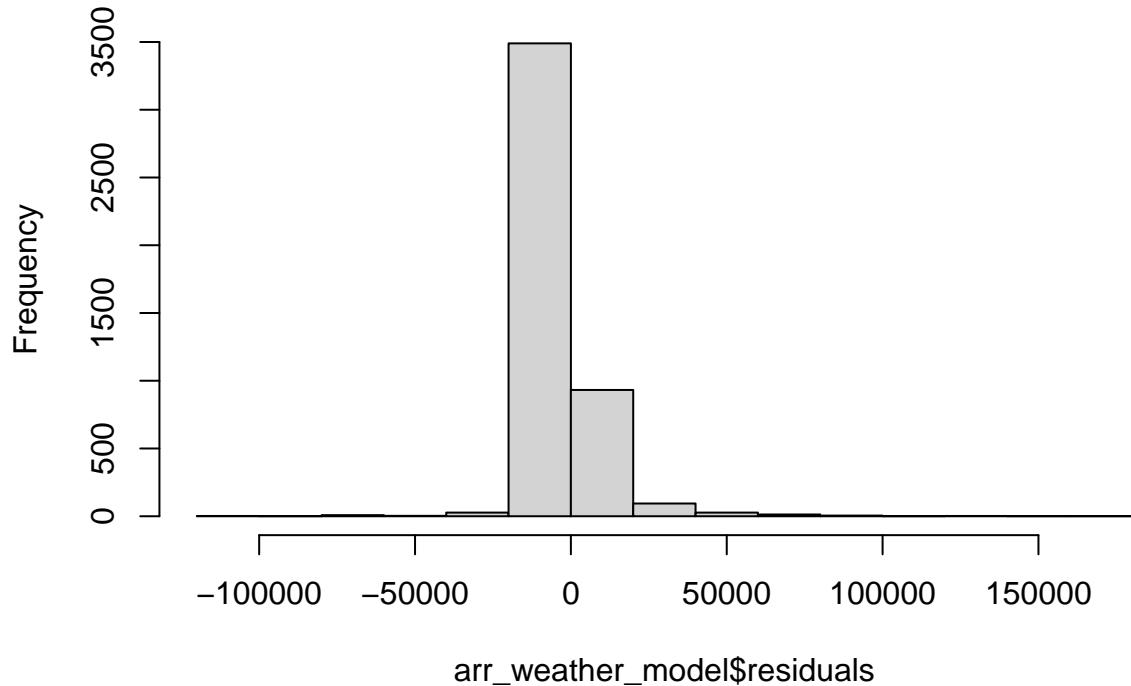
```
hist(arr_carrier_model$residuals)
```

**Histogram of arr\_carrier\_model\$residuals**



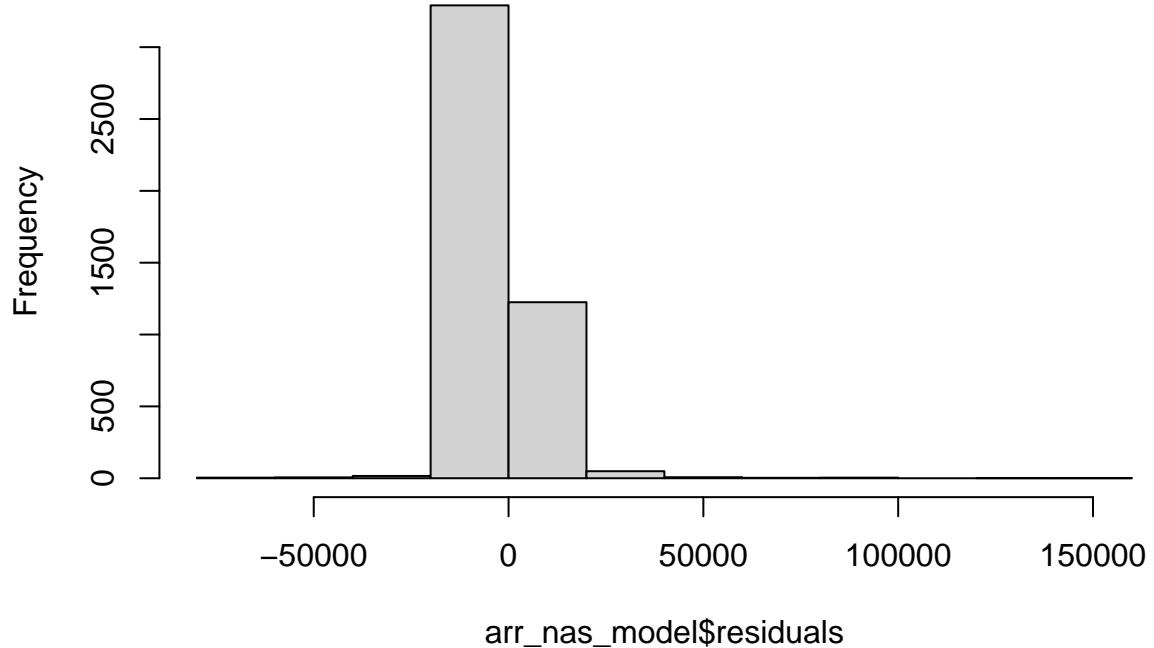
```
hist(arr_weather_model$residuals)
```

**Histogram of arr\_weather\_model\$residuals**



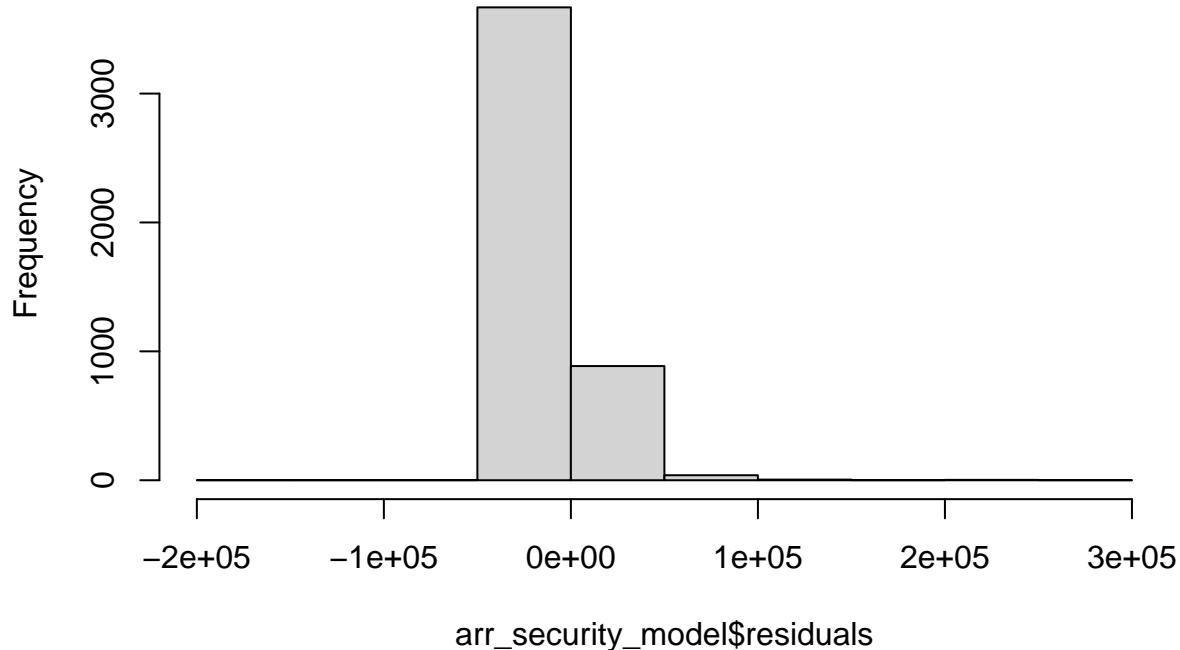
```
hist(arr_nas_model$residuals)
```

**Histogram of arr\_nas\_model\$residuals**



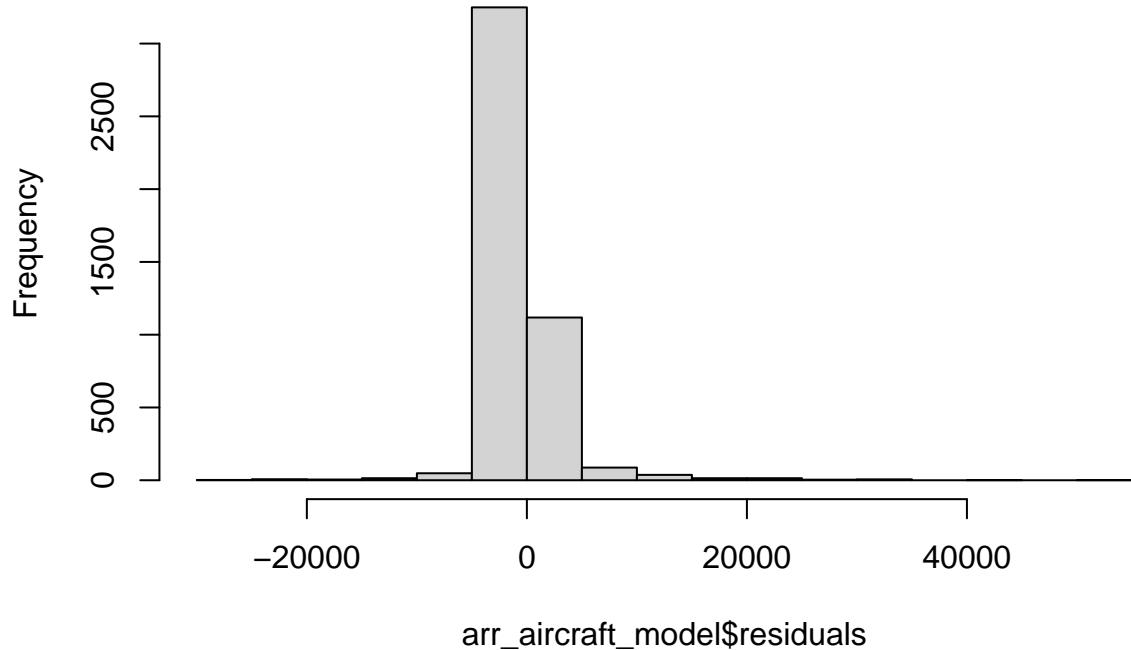
```
hist(arr_security_model$residuals)
```

**Histogram of arr\_security\_model\$residuals**



```
hist(arr_aircraft_model$residuals)
```

## Histogram of arr\_aircraft\_model\$residuals



## Chi-Square Test

```
delta <- subset(airlines_data, carrier_name == "Delta Air Lines Inc.")
american <- subset(airlines_data, carrier_name == "American Airlines Inc.")

# carrier_ct vs weather_ct
delta_table<- table(delta$carrier_ct, delta$weather_ct)
american_table <- table(american$carrier_ct, american$weather_ct)
chi_sq <- chisq.test(delta_table, american_table)

## Warning in chisq.test(delta_table, american_table): Chi-squared approximation
## may be incorrect

print(chi_sq)

##
## Pearson's Chi-squared test
##
## data: delta_table
## X-squared = 90954, df = 85360, p-value < 2.2e-16
```

```

#carrier_ct vs nas_ct
delta_table<- table (delta$carrier_ct, delta$nas_ct)
american_table <- table(american$carrier_ct, american$nas_ct)
chi_sq <- chisq.test(delta_table, american_table)

## Warning in chisq.test(delta_table, american_table): Chi-squared approximation
## may be incorrect

print(chi_sq)

##
## Pearson's Chi-squared test
##
## data: delta_table
## X-squared = 146650, df = 138904, p-value < 2.2e-16

#carrier_ct vs security_ct
delta_table <- table(delta$carrier_ct, delta$security_ct)
american_table <- table(american$carrier_ct, american$security_ct)
chi_sq <- chisq.test(delta_table, american_table)

## Warning in chisq.test(delta_table, american_table): Chi-squared approximation
## may be incorrect

print(chi_sq)

##
## Pearson's Chi-squared test
##
## data: delta_table
## X-squared = 7303.2, df = 6596, p-value = 1.293e-09

#carrier_ct vs late_aircraft_ct
delta_table <- table (delta$carrier_ct, delta$late_aircraft_ct)
american_table <- table (american$carrier_ct, american$late_aircraft_ct)
chi_sq <- chisq.test(delta_table, american_table)

## Warning in chisq.test(delta_table, american_table): Chi-squared approximation
## may be incorrect

print(chi_sq)

##
## Pearson's Chi-squared test
##
## data: delta_table
## X-squared = 147352, df = 137740, p-value < 2.2e-16

```

## Correlation Matrix

```

data <- airlines_data %>%
  select(carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay)

correlation_matrix <- cor(data, use = "complete.obs")
print(correlation_matrix)

##                                     carrier_delay weather_delay nas_delay security_delay
## carrier_delay                 1.0000000   0.7406985  0.7743065   0.4501237
## weather_delay                0.7406985   1.0000000  0.5202790   0.4013624
## nas_delay                     0.7743065   0.5202790  1.0000000   0.3883256
## security_delay                0.4501237   0.4013624  0.3883256   1.0000000
## late_aircraft_delay          0.9234710   0.5888050  0.8175952   0.4364896
##                                     late_aircraft_delay
## carrier_delay                  0.9234710
## weather_delay                 0.5888050
## nas_delay                      0.8175952
## security_delay                 0.4364896
## late_aircraft_delay           1.0000000

set.seed(1)
selected_columns_for_matrix <- c("carrier_name", "airport", "arr_flights", "arr_del15", "carrier_ct", "weather_ct", "nas_ct", "security_ct", "late_aircraft_ct", "airline_cancelled", "airline_diverted")
subset_data <- airlines_data[, selected_columns_for_matrix]
plot(subset_data, col="blue")

```

