ForProject_GrndWrk.r

vk0589

2023-07-23

```
setwd("C:/Users/vk0589/OneDrive - UNT System/Documents/UNT/Courses/ADTA5130")
#install.packages("tidyverse")
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.1
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
##
## 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 4.3.1
```

```
rawdata <- read_excel("AirlineMarch2023.xlsx")
# Quick Look at the Data
head(rawdata)
```

```
## # A tibble: 6 x 32
##    YEAR MONTH DAY_OF_MONTH DAY_OF_WEEK MKT_UNIQUE_CARRIER MKT_CARRIER_FL_NUM
##   <dbl> <dbl>        <dbl>       <dbl> <chr>                           <dbl>
## 1  2023     3            1           3 AA                                  1
## 2  2023     3            1           3 AA                                 10
## 3  2023     3            1           3 AA                               1002
## 4  2023     3            1           3 AA                               1004
## 5  2023     3            1           3 AA                               1006
## 6  2023     3            1           3 AA                               1007
```

```
## # i 26 more variables: 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 <chr>, DIVERTED <dbl>, CRS_ELAPSED_TIME <dbl>,
## #   ACTUAL_ELAPSED_TIME <dbl>, AIR_TIME <dbl>, DISTANCE <dbl>,
## #   CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>, NAS_DELAY <dbl>, ...
```

```r
#Selecting Data of Ohio City ######
ohio_data <- rawdata %>%
  filter(grepl(", OH", ORIGIN_CITY_NAME))
summary(as.data.frame(ohio_data))
```

```
##       YEAR          MONTH    DAY_OF_MONTH    DAY_OF_WEEK    MKT_UNIQUE_CARRIER
##  Min.   :2023   Min.   :3   Min.   : 1.00   Min.   :1.000   Length:11844
##  1st Qu.:2023   1st Qu.:3   1st Qu.: 9.00   1st Qu.:2.000   Class :character
##  Median :2023   Median :3   Median :16.00   Median :4.000   Mode  :character
##  Mean   :2023   Mean   :3   Mean   :16.19   Mean   :3.979
##  3rd Qu.:2023   3rd Qu.:3   3rd Qu.:24.00   3rd Qu.:5.000
##  Max.   :2023   Max.   :3   Max.   :31.00   Max.   :7.000
##
##  MKT_CARRIER_FL_NUM    ORIGIN          ORIGIN_CITY_NAME       DEST
##  Min.   :  34       Length:11844       Length:11844       Length:11844
##  1st Qu.:1165       Class :character   Class :character   Class :character
##  Median :2827       Mode  :character   Mode  :character   Mode  :character
##  Mean   :2935
##  3rd Qu.:4635
##  Max.   :9678
##
##  DEST_CITY_NAME      CRS_DEP_TIME     DEP_TIME      DEP_DELAY
##  Length:11844       Min.   : 505   Min.   :   1   Min.   : -34.000
##  Class :character   1st Qu.: 736   1st Qu.: 750   1st Qu.:  -7.000
##  Mode  :character   Median :1220   Median :1222   Median :  -3.000
##                     Mean   :1226   Mean   :1232   Mean   :   9.905
##                     3rd Qu.:1700   3rd Qu.:1659   3rd Qu.:   3.000
##                     Max.   :2241   Max.   :2348   Max.   :1931.000
##                                    NA's   :116    NA's   :116
##     TAXI_OUT       WHEELS_OFF     WHEELS_ON       TAXI_IN        CRS_ARR_TIME
##  Min.   :  2.0   Min.   :   1   Min.   :   1   Min.   : 1.000   Min.   :   5
##  1st Qu.: 11.0   1st Qu.: 806   1st Qu.: 917   1st Qu.: 6.000   1st Qu.: 930
##  Median : 13.0   Median :1235   Median :1325   Median : 8.000   Median :1339
##  Mean   : 15.8   Mean   :1259   Mean   :1353   Mean   : 9.618   Mean   :1374
##  3rd Qu.: 18.0   3rd Qu.:1714   3rd Qu.:1816   3rd Qu.:11.000   3rd Qu.:1829
##  Max.   :103.0   Max.   :2358   Max.   :2400   Max.   :94.000   Max.   :2359
##  NA's   :119     NA's   :119    NA's   :119    NA's   :119
##     ARR_TIME      ARR_DELAY        CANCELLED       CANCELLATION_CODE
##  Min.   :   1   Min.   : -45.0   Min.   :0.00000   Length:11844
##  1st Qu.: 924   1st Qu.: -16.0   1st Qu.:0.00000   Class :character
##  Median :1332   Median :  -6.0   Median :0.00000   Mode  :character
##  Mean   :1365   Mean   :   6.1   Mean   :0.01005
##  3rd Qu.:1825   3rd Qu.:   8.0   3rd Qu.:0.00000
##  Max.   :2400   Max.   :1928.0   Max.   :1.00000
##  NA's   :119    NA's   :138
##     DIVERTED        CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME    AIR_TIME
```

```
##  Min.   :0.000000   Min.   : 53.0    Min.   : 39.0       Min.   : 22.00
##  1st Qu.:0.000000   1st Qu.: 90.0    1st Qu.: 85.0       1st Qu.: 60.00
##  Median :0.000000   Median :110.0    Median :105.0       Median : 77.50
##  Mean   :0.001604   Mean   :127.1    Mean   :123.2       Mean   : 97.82
##  3rd Qu.:0.000000   3rd Qu.:150.0    3rd Qu.:146.0       3rd Qu.:121.00
##  Max.   :1.000000   Max.   :337.0    Max.   :378.0       Max.   :360.00
##                                      NA's   :138         NA's   :138
##     DISTANCE      CARRIER_DELAY     WEATHER_DELAY       NAS_DELAY
##  Min.   :  95.0   Min.   :   0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.: 347.0   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:  0.00
##  Median : 483.0   Median :   0.00   Median :  0.000   Median :  9.00
##  Mean   : 642.4   Mean   :  24.97   Mean   :  2.692   Mean   : 20.29
##  3rd Qu.: 869.0   3rd Qu.:  15.00   3rd Qu.:  0.000   3rd Qu.: 23.00
##  Max.   :2161.0   Max.   :1928.00   Max.   :814.000   Max.   :878.00
##                   NA's   :9590      NA's   :9590      NA's   :9590
##  SECURITY_DELAY   LATE_AIRCRAFT_DELAY
##  Min.   : 0.000   Min.   :   0.00
##  1st Qu.: 0.000   1st Qu.:   0.00
##  Median : 0.000   Median :   0.00
##  Mean   : 0.075   Mean   :  24.15
##  3rd Qu.: 0.000   3rd Qu.:  25.00
##  Max.   :75.000   Max.   :1197.00
##  NA's   :9590     NA's   :9590
```

```r
# ohio_data_bkp <- ohio_data #Backup fro Main Data

#Adding Date (Fixing the existing Date)
ohio_data$DATE = paste(ohio_data$YEAR,ohio_data$MONTH,ohio_data$DAY_OF_MONTH,
                       sep = "/")
ohio_data$DATE = as.Date(ohio_data$DATE, format = "%Y/%m/%d")

#Selecting Required Columns
ohdata2 <- ohio_data %>%
  select(DATE, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, MKT_CARRIER_FL_NUM, ORIGIN,
         ORIGIN_CITY_NAME, DEST, DEST_CITY_NAME, DEP_TIME, DEP_DELAY, TAXI_OUT,
         WHEELS_OFF, WHEELS_ON, TAXI_IN, ARR_TIME, ARR_DELAY, CANCELLED,
         CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME, AIR_TIME, DISTANCE)
# ohdata2 has required Columns that are needed for the Hypothesis
summary(ohdata2)
```

```
##       DATE             DAY_OF_WEEK    MKT_UNIQUE_CARRIER MKT_CARRIER_FL_NUM
##  Min.   :2023-03-01   Min.   :1.000   Length:11844       Min.   :  34
##  1st Qu.:2023-03-09   1st Qu.:2.000   Class :character   1st Qu.:1165
##  Median :2023-03-16   Median :4.000   Mode  :character   Median :2827
##  Mean   :2023-03-16   Mean   :3.979                      Mean   :2935
##  3rd Qu.:2023-03-24   3rd Qu.:5.000                      3rd Qu.:4635
##  Max.   :2023-03-31   Max.   :7.000                      Max.   :9678
##
##     ORIGIN          ORIGIN_CITY_NAME       DEST           DEST_CITY_NAME
##  Length:11844       Length:11844       Length:11844       Length:11844
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##     DEP_TIME      DEP_DELAY           TAXI_OUT       WHEELS_OFF
##  Min.   :   1   Min.   : -34.000   Min.   :  2.0   Min.   :   1
##  1st Qu.: 750   1st Qu.:  -7.000   1st Qu.: 11.0   1st Qu.: 806
##  Median :1222   Median :  -3.000   Median : 13.0   Median :1235
##  Mean   :1232   Mean   :   9.905   Mean   : 15.8   Mean   :1259
##  3rd Qu.:1659   3rd Qu.:   3.000   3rd Qu.: 18.0   3rd Qu.:1714
##  Max.   :2348   Max.   :1931.000   Max.   :103.0   Max.   :2358
##  NA's   :116    NA's   :116        NA's   :119     NA's   :119
##    WHEELS_ON       TAXI_IN          ARR_TIME      ARR_DELAY
##  Min.   :   1   Min.   : 1.000   Min.   :   1   Min.   : -45.0
##  1st Qu.: 917   1st Qu.: 6.000   1st Qu.: 924   1st Qu.: -16.0
##  Median :1325   Median : 8.000   Median :1332   Median :  -6.0
##  Mean   :1353   Mean   : 9.618   Mean   :1365   Mean   :   6.1
##  3rd Qu.:1816   3rd Qu.:11.000   3rd Qu.:1825   3rd Qu.:   8.0
##  Max.   :2400   Max.   :94.000   Max.   :2400   Max.   :1928.0
##  NA's   :119    NA's   :119      NA's   :119    NA's   :138
##    CANCELLED       CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME    AIR_TIME
##  Min.   :0.00000   Min.   : 53.0    Min.   : 39.0       Min.   : 22.00
##  1st Qu.:0.00000   1st Qu.: 90.0    1st Qu.: 85.0       1st Qu.: 60.00
##  Median :0.00000   Median :110.0    Median :105.0       Median : 77.50
##  Mean   :0.01005   Mean   :127.1    Mean   :123.2       Mean   : 97.82
##  3rd Qu.:0.00000   3rd Qu.:150.0    3rd Qu.:146.0       3rd Qu.:121.00
##  Max.   :1.00000   Max.   :337.0    Max.   :378.0       Max.   :360.00
##                                     NA's   :138         NA's   :138
##     DISTANCE
##  Min.   :  95.0
##  1st Qu.: 347.0
##  Median : 483.0
##  Mean   : 642.4
##  3rd Qu.: 869.0
##  Max.   :2161.0
##
```

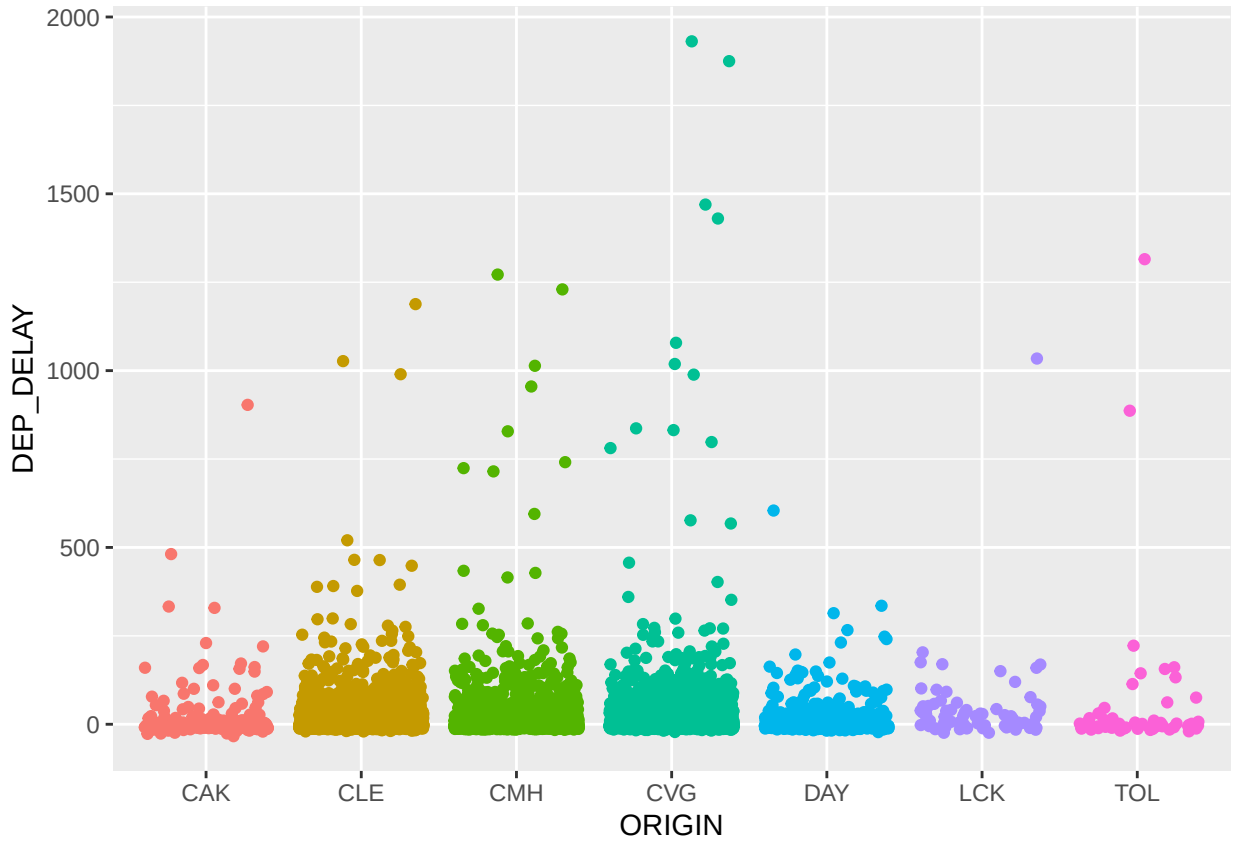
```r
#ohdata2_bkp <- ohdata2

#Changing data type to factors - For Categorical Variables
ohdata2$DAY_OF_WEEK <- as.factor(ohdata2$DAY_OF_WEEK)
ohdata2$MKT_UNIQUE_CARRIER <- as.factor(ohdata2$MKT_UNIQUE_CARRIER)
ohdata2$MKT_CARRIER_FL_NUM <- as.factor(ohdata2$MKT_CARRIER_FL_NUM)
ohdata2$ORIGIN <- as.factor(ohdata2$ORIGIN)
ohdata2$ORIGIN_CITY_NAME <- as.factor(ohdata2$ORIGIN_CITY_NAME)
ohdata2$DEST <- as.factor(ohdata2$DEST)
ohdata2$DEST_CITY_NAME <- as.factor(ohdata2$DEST_CITY_NAME)
ohdata2$CANCELLED <- as.factor(ohdata2$CANCELLED)
summary(ohdata2)
```

```
##       DATE            DAY_OF_WEEK MKT_UNIQUE_CARRIER MKT_CARRIER_FL_NUM
##  Min.   :2023-03-01   1:1606      AA     :3401       1437   :   62
##  1st Qu.:2023-03-09   2:1460      DL     :2825       4685   :   62
##  Median :2023-03-16   3:1855      UA     :2161       5093   :   61
##  Mean   :2023-03-16   4:2010      WN     :1663       419    :   60
##  3rd Qu.:2023-03-24   5:2009      F9     : 613       1038   :   58
```

```
##  Max.   :2023-03-31   6:1372      G4     : 549       1043   :   58
##                       7:1532      (Other): 632       (Other):11483
##  ORIGIN           ORIGIN_CITY_NAME      DEST            DEST_CITY_NAME
##  CAK: 304   Akron, OH     : 304    ORD    :1074   Chicago, IL   :1476
##  CLE:3505   Cincinnati, OH:3544    ATL    : 798   New York, NY  :1287
##  CMH:3575   Cleveland, OH :3505    LGA    : 791   Washington, DC: 994
##  CVG:3544   Columbus, OH  :3651    CLT    : 675   Atlanta, GA   : 798
##  DAY: 784   Dayton, OH    : 784    DCA    : 630   Charlotte, NC : 675
##  LCK:  76   Toledo, OH    :  56    EWR    : 605   Newark, NJ    : 605
##  TOL:  56                          (Other):7271   (Other)       :6009
##     DEP_TIME      DEP_DELAY          TAXI_OUT       WHEELS_OFF
##  Min.   :   1   Min.   : -34.000   Min.   :  2.0   Min.   :   1
##  1st Qu.: 750   1st Qu.:  -7.000   1st Qu.: 11.0   1st Qu.: 806
##  Median :1222   Median :  -3.000   Median : 13.0   Median :1235
##  Mean   :1232   Mean   :   9.905   Mean   : 15.8   Mean   :1259
##  3rd Qu.:1659   3rd Qu.:   3.000   3rd Qu.: 18.0   3rd Qu.:1714
##  Max.   :2348   Max.   :1931.000   Max.   :103.0   Max.   :2358
##  NA's   :116    NA's   :116        NA's   :119     NA's   :119
##    WHEELS_ON       TAXI_IN          ARR_TIME      ARR_DELAY      CANCELLED
##  Min.   :   1   Min.   : 1.000   Min.   :   1   Min.   : -45.0   0:11725
##  1st Qu.: 917   1st Qu.: 6.000   1st Qu.: 924   1st Qu.: -16.0   1:  119
##  Median :1325   Median : 8.000   Median :1332   Median :  -6.0
##  Mean   :1353   Mean   : 9.618   Mean   :1365   Mean   :   6.1
##  3rd Qu.:1816   3rd Qu.:11.000   3rd Qu.:1825   3rd Qu.:   8.0
##  Max.   :2400   Max.   :94.000   Max.   :2400   Max.   :1928.0
##  NA's   :119    NA's   :119      NA's   :119    NA's   :138
##  CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME    AIR_TIME         DISTANCE
##  Min.   : 53.0    Min.   : 39.0       Min.   : 22.00   Min.   :  95.0
##  1st Qu.: 90.0    1st Qu.: 85.0       1st Qu.: 60.00   1st Qu.: 347.0
##  Median :110.0    Median :105.0       Median : 77.50   Median : 483.0
##  Mean   :127.1    Mean   :123.2       Mean   : 97.82   Mean   : 642.4
##  3rd Qu.:150.0    3rd Qu.:146.0       3rd Qu.:121.00   3rd Qu.: 869.0
##  Max.   :337.0    Max.   :378.0       Max.   :360.00   Max.   :2161.0
##                   NA's   :138         NA's   :138
```

```r
### HYPOTHESIS TESTING ###
## Hypothesis 1 : ANOVA for DEP_DELAY vs ORIGIN #####
# Reference: https://statsandr.com/blog/anova-in-r/
# Graph for Ohio Data - Departure Delay vs Origin Airport
ggplot(ohdata2) +
  aes(x = ORIGIN, y = DEP_DELAY , color = ORIGIN) +
  geom_jitter() +
  theme(legend.position = "none")
```

```
## Warning: Removed 116 rows containing missing values (`geom_point()`).
```

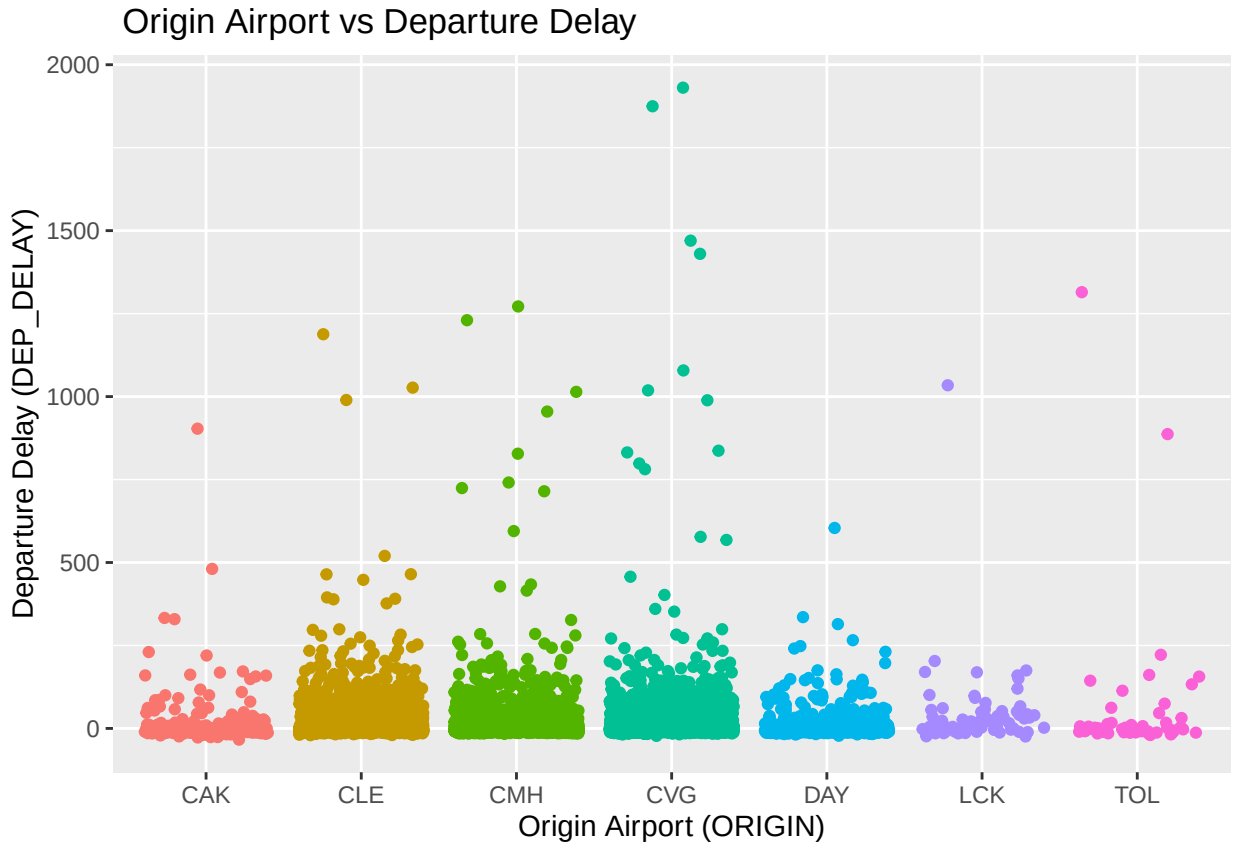



```
#ANOVA for Departure Delay with Origin Airport
#
#NULL and ALTERNATE Hypothesis:
# Ho : Average Flight Delays in all Airports of Ohio are not Different.
#      Airport of Origin doesn't affect Flight Delays
#      i.e.,  Mean(CAK)=Mean(CLE)=Mean(CMH)=Mean(CVG)=Mean(DAY)=Mean(LCK)=Mean(TOL)
# Ha: Average flight Delays is different for atleast 1 Airport in Ohio.
#     Origin Airport affects the Flight Delays
#     i.e., Mean Delay of 1 or more Airport(s) is different compared to others Airport

# ANOVA for Unclean Data (with Null Values and Outliers)
result1 <- aov(ohdata2$DEP_DELAY~ohdata2$ORIGIN)
summary(result1)
```

```
##                   Df   Sum Sq Mean Sq F value   Pr(>F)
## ohdata2$ORIGIN     6   251543   41924    10.1 3.66e-11 ***
## Residuals      11721 48662318    4152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 116 observations deleted due to missingness
```

```
ggplot(ohdata2) +
  aes(x = ORIGIN, y = DEP_DELAY , xlab = "Origin Airport (ORIGIN)", ylab = "Departure Delay (DEP_DELAY)
  geom_jitter() +
  theme(legend.position = "none") +
  labs(x = "Origin Airport (ORIGIN)", y = "Departure Delay (DEP_DELAY)", title = " Origin Airport vs Dep
```

```
## Warning: Removed 116 rows containing missing values (`geom_point()`).
```

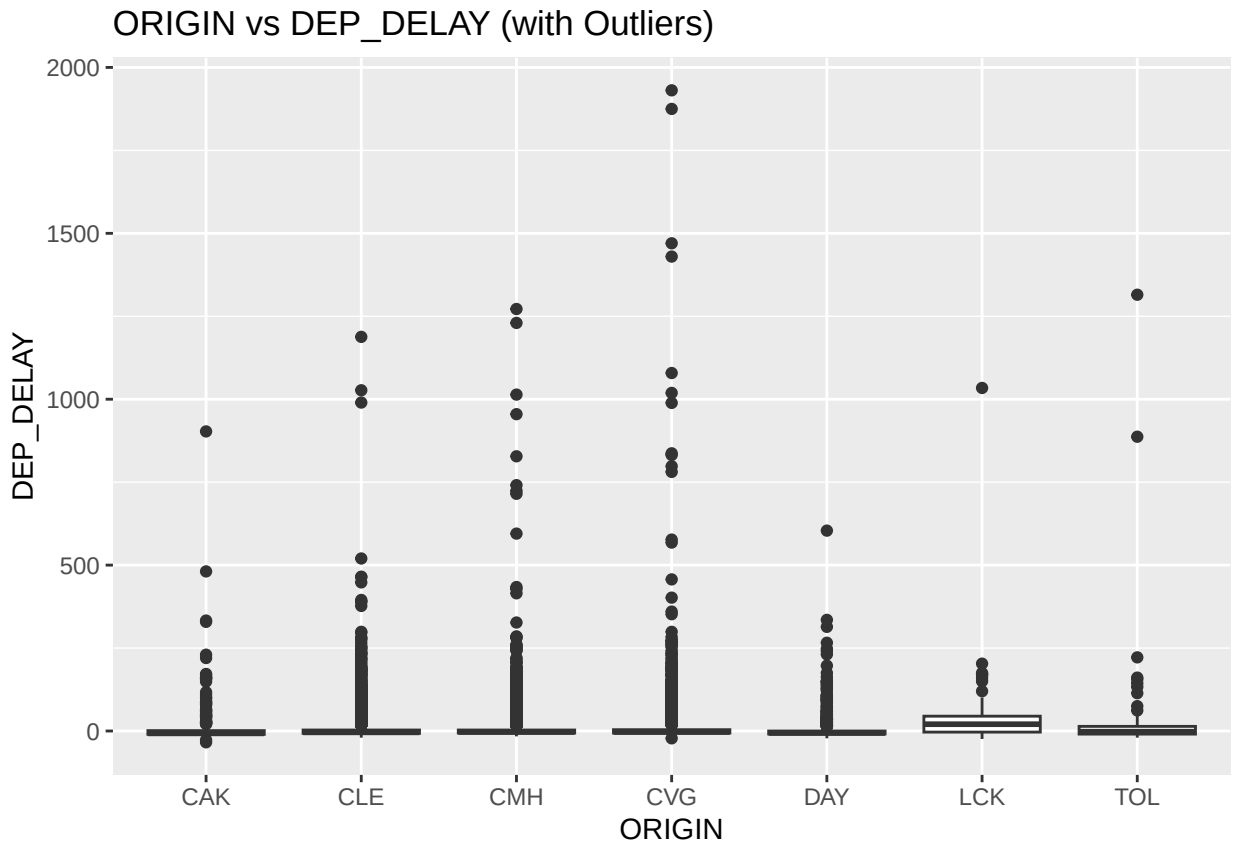



```
# Selecting Columns reguired for Hypothesis 1
h1 <- ohdata2 %>% select(ORIGIN, ORIGIN_CITY_NAME ,DEP_DELAY)
summary(h1) #h1
```

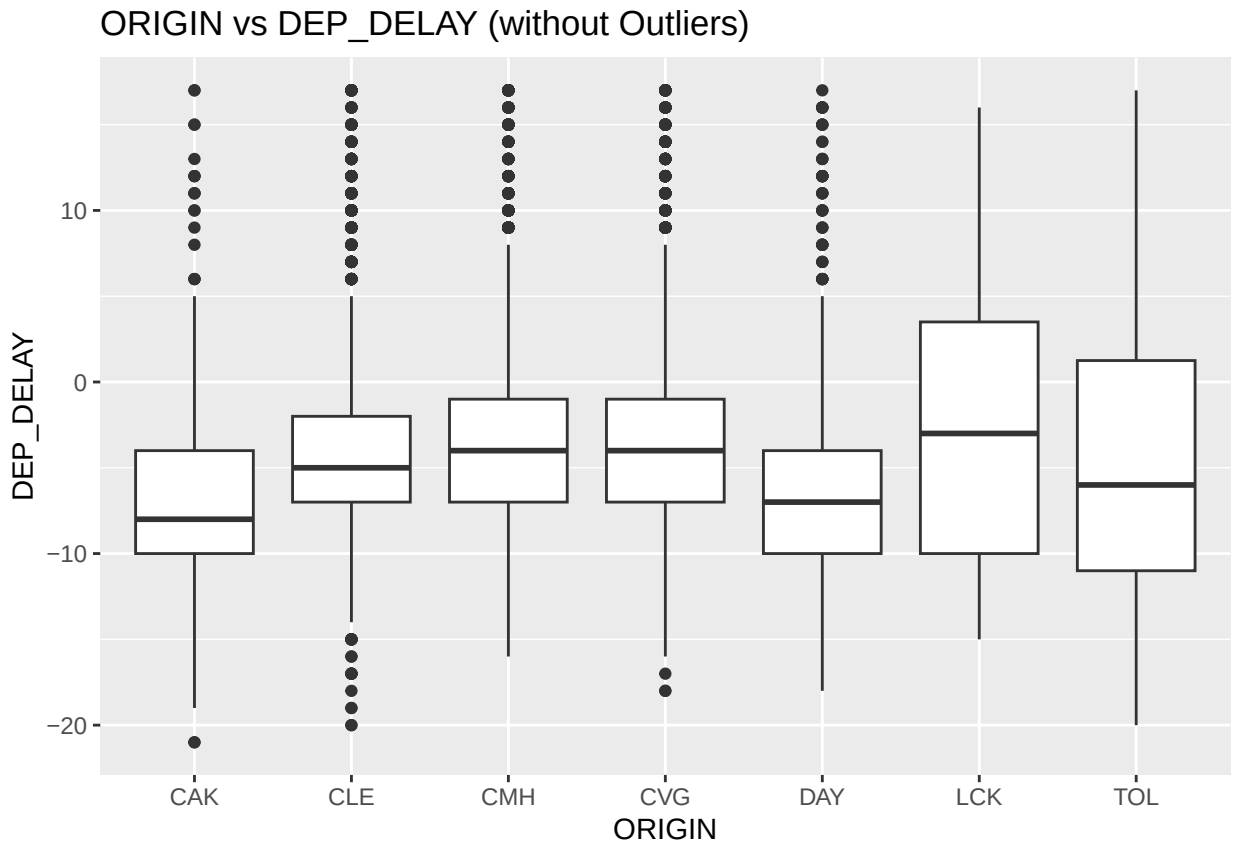
```
##  ORIGIN           ORIGIN_CITY_NAME   DEP_DELAY
##  CAK: 304   Akron, OH     : 304    Min.   : -34.000
##  CLE:3505   Cincinnati, OH:3544    1st Qu.:  -7.000
##  CMH:3575   Cleveland, OH :3505    Median :  -3.000
##  CVG:3544   Columbus, OH  :3651    Mean   :   9.905
##  DAY: 784   Dayton, OH    : 784    3rd Qu.:   3.000
##  LCK:  76   Toledo, OH    :  56    Max.   :1931.000
##  TOL:  56                          NA's   :116
```

```
ggplot(ohdata2) +
  aes(x = ORIGIN, y = DEP_DELAY) +
  geom_boxplot() +
  theme(legend.position = "none")+
  labs(title = "ORIGIN vs DEP_DELAY (with Outliers)")
```

```
## Warning: Removed 116 rows containing non-finite values (`stat_boxplot()`).
```




```
#Clean Data - Remove Outliers and Null Values for ANOVA
#Finding Mean and Standard Dev without Missing Value
#group_by(h1, ORIGIN) %>%
#  summarise(
#    mean = mean(DEP_DELAY, na.rm = TRUE),
#    sd = sd(DEP_DELAY, na.rm = TRUE))
na.omit(h1) -> h1t # Removed NAs; h1 with no Null Values - h1t
#Finding Q1, Q3 and IQR for Outliers
Q1_1 <- quantile(h1t$DEP_DELAY, 0.25)
Q3_1 <- quantile(h1t$DEP_DELAY, 0.75)
IQR_1 <- IQR(h1t$DEP_DELAY)
h1t_no <- subset(h1t,
                 h1t$DEP_DELAY > (Q1_1 - 1.5*IQR_1) &
                   h1t$DEP_DELAY < (Q3_1 + 1.5*IQR_1))
# h1t with No Outliers - h1t_no

# Graph for a quick look on Data without Outliers
ggplot(h1t_no) +
  aes(x = ORIGIN, y = DEP_DELAY) +
  geom_boxplot() +
  theme(legend.position = "none")+
  labs(title = "ORIGIN vs DEP_DELAY (without Outliers)")
```




```
#Applying ANOVA for No Outlier Data
result1_no <- aov(DEP_DELAY ~ ORIGIN, data = h1t_no)
# result1_no has the results for Hypothesis 1 with h1t_no (No Outliers and NAs)
summary(result1_no) # Stats of result1_no
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## ORIGIN         6   6946  1157.7   33.36 <2e-16 ***
## Residuals   9952 345334    34.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#install.packages("report")
library(report)
```

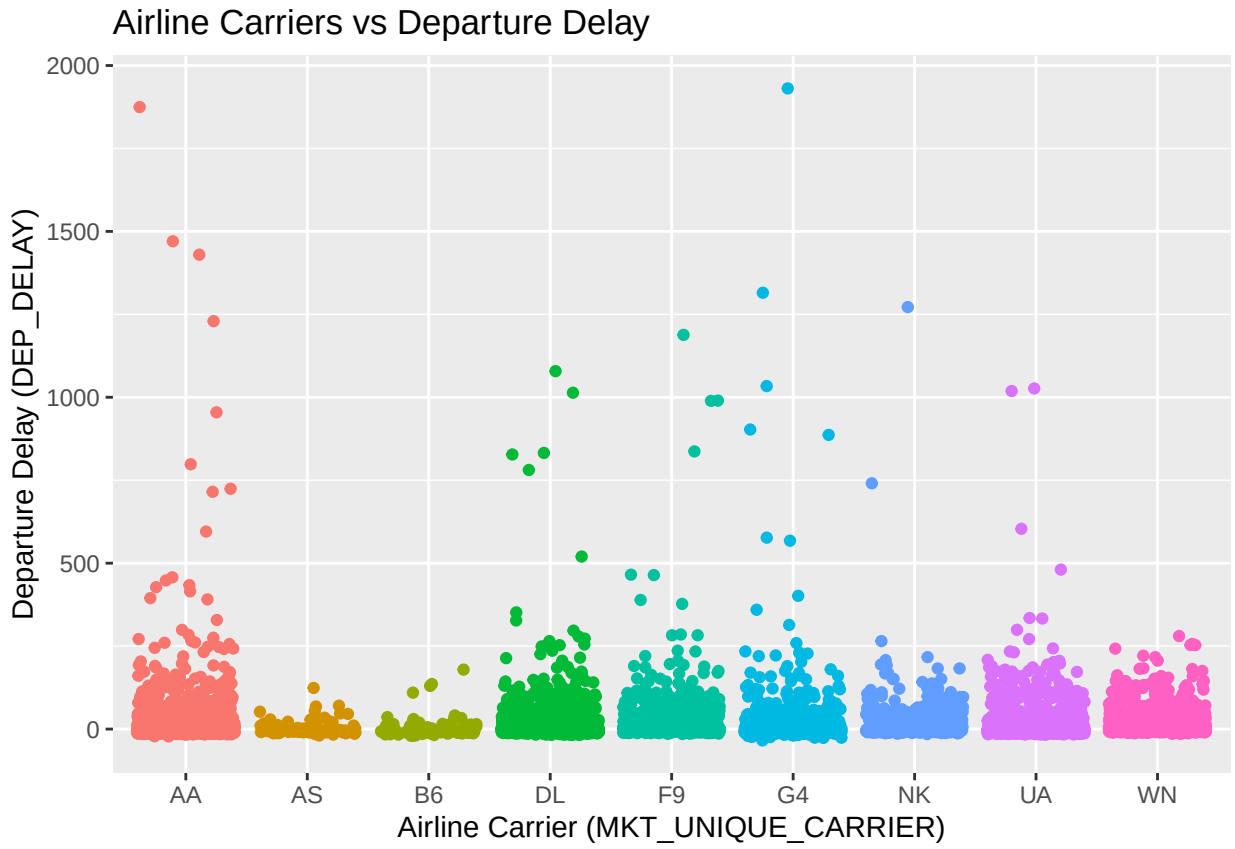
```
## Warning: package 'report' was built under R version 4.3.1
```

```
report(result1_no) # Final Report of our Results in Hypothesis 1
```

```
## The ANOVA (formula: DEP_DELAY ~ ORIGIN) suggests that:
##
##   - The main effect of ORIGIN is statistically significant and small (F(6, 9952)
## = 33.36, p < .001; Eta2 = 0.02, 95% CI [0.01, 1.00])
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
## Hypothesis 2 : ANOVA for TAXI_OUT vs MKT_UNIQUE_CARRIER #####
#Graph for Ohio Data - Departure Delay Time vs Fight Carrier
ggplot(ohdata2) +
  aes(x = MKT_UNIQUE_CARRIER, y = DEP_DELAY , color = MKT_UNIQUE_CARRIER) +
  geom_jitter() +
  theme(legend.position = "none") +
  xlab("Airline Carrier (MKT_UNIQUE_CARRIER)") +
  ylab("Departure Delay (DEP_DELAY)") +
  labs(title = "Airline Carriers vs Departure Delay")
```

```
## Warning: Removed 116 rows containing missing values (`geom_point()`).
```




```
#ANOVA for Departure Delay Time with Flight Carrier
# NULL and ALTERNATE HYPOTHESIS
##
# Ho : Mean Time out for all the Flight Carriers are not different (or) There is no
#     difference in the average Departure Delay time for all the Flight Carrier Companies
#      i.e.,  Mean(AA)=Mean(AS)=Mean(B6)=Mean(DL)=Mean(F9)=Mean(G4)=Mean(NK)=Mean(UA)
#               =Mean(WN)
# Ha: Average Departure Delay Time is not same for the Flight Carriers and atleast 1 Flight
#     Carrier's Departure Delay Mean is different
#     i.e., Mean Departure Delay Time of 1 or more Carrier(s) is different compared to others
##

# ANOVA for Unclean Data (with Null Values and Outliers)
```

```r
result2 <- aov(ohdata2$DEP_DELAY~ohdata2$MKT_UNIQUE_CARRIER)
summary(result2)
```

```
##                               Df   Sum Sq Mean Sq F value Pr(>F)
## ohdata2$MKT_UNIQUE_CARRIER     8   804751  100594    24.5 <2e-16 ***
## Residuals                  11719 48109110    4105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 116 observations deleted due to missingness
```

```r
#Clean Data - Remove Outliers and Null Values for ANOVA
h2 <- ohdata2 %>% select(MKT_UNIQUE_CARRIER, MKT_CARRIER_FL_NUM, ORIGIN_CITY_NAME, DEP_DELAY)
summary(h2)
```

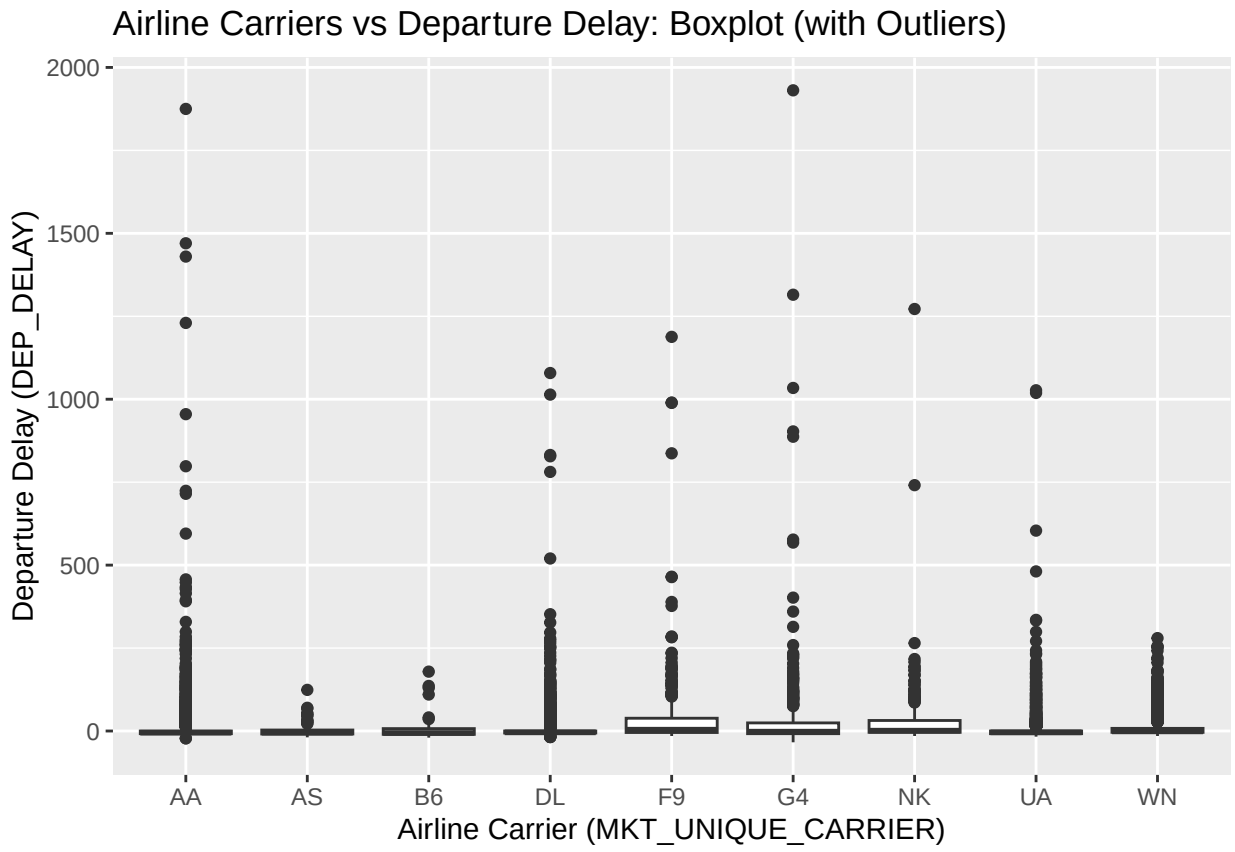
```
##  MKT_UNIQUE_CARRIER MKT_CARRIER_FL_NUM       ORIGIN_CITY_NAME
##  AA     :3401       1437   :   62      Akron, OH     : 304
##  DL     :2825       4685   :   62      Cincinnati, OH:3544
##  UA     :2161       5093   :   61      Cleveland, OH :3505
##  WN     :1663       419    :   60      Columbus, OH  :3651
##  F9     : 613       1038   :   58      Dayton, OH    : 784
##  G4     : 549       1043   :   58      Toledo, OH    :  56
##  (Other): 632       (Other):11483
##    DEP_DELAY
##  Min.   : -34.000
##  1st Qu.:  -7.000
##  Median :  -3.000
##  Mean   :   9.905
##  3rd Qu.:   3.000
##  Max.   :1931.000
##  NA's   :116
```

```r
dim(h2) # h2 Dimensions
```

```
## [1] 11844     4
```

```r
ggplot(ohdata2) +
  aes(x = MKT_UNIQUE_CARRIER, y = DEP_DELAY) +
  geom_boxplot() +
  theme(legend.position = "none") +
  labs(title = "Airline Carriers vs Departure Delay: Boxplot (with Outliers)") +
  xlab("Airline Carrier (MKT_UNIQUE_CARRIER)") + ylab("Departure Delay (DEP_DELAY)")
```

```
## Warning: Removed 116 rows containing non-finite values (`stat_boxplot()`).
```



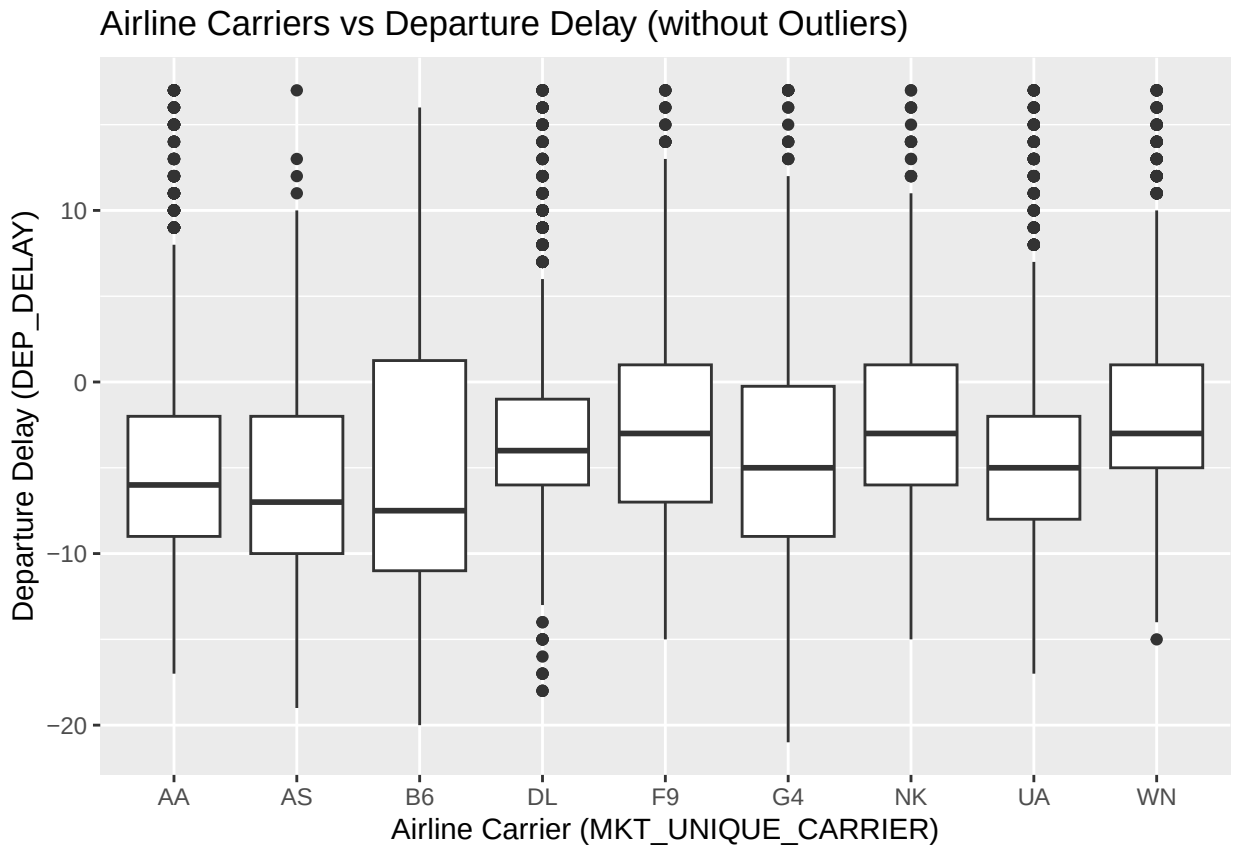

```
#Removing Outliers and Missing Values
na.omit(h2) -> h2t # Removed NAs; h2 with no Null Values - h2t
dim(h2t) #h2t Dimensions (After Removing NAs)
```

```
## [1] 11728     4
```

```
#Finding Q1, Q3 and IQR
Q1_2 <- quantile(h2t$DEP_DELAY, 0.25)
Q3_2 <- quantile(h2t$DEP_DELAY, 0.75)
IQR_2 <- IQR(h2t$DEP_DELAY)
h2t_no <- subset(h2t,
                 h2t$DEP_DELAY > (Q1_2 - 1.5*IQR_2) & h2t$DEP_DELAY < (Q3_2 + 1.5*IQR_2))
dim(h2t_no)
```

```
## [1] 9959    4
```

```
# h2t with No Outliers - h2t_no
ggplot(h2t_no) +
  aes(x = MKT_UNIQUE_CARRIER, y = DEP_DELAY) +
  geom_boxplot() +
  theme(legend.position = "none")+
  labs(title = "Airline Carriers vs Departure Delay (without Outliers)") +
  xlab("Airline Carrier (MKT_UNIQUE_CARRIER)") + ylab("Departure Delay (DEP_DELAY)")
```




```
#Applying ANOVA for Hypothesis 2 with No Outlier Data
result2_no <- aov(DEP_DELAY ~ MKT_UNIQUE_CARRIER, data = h2t_no)
summary(result2_no)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## MKT_UNIQUE_CARRIER     8  15153  1894.1    55.9 <2e-16 ***
## Residuals           9950 337128    33.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

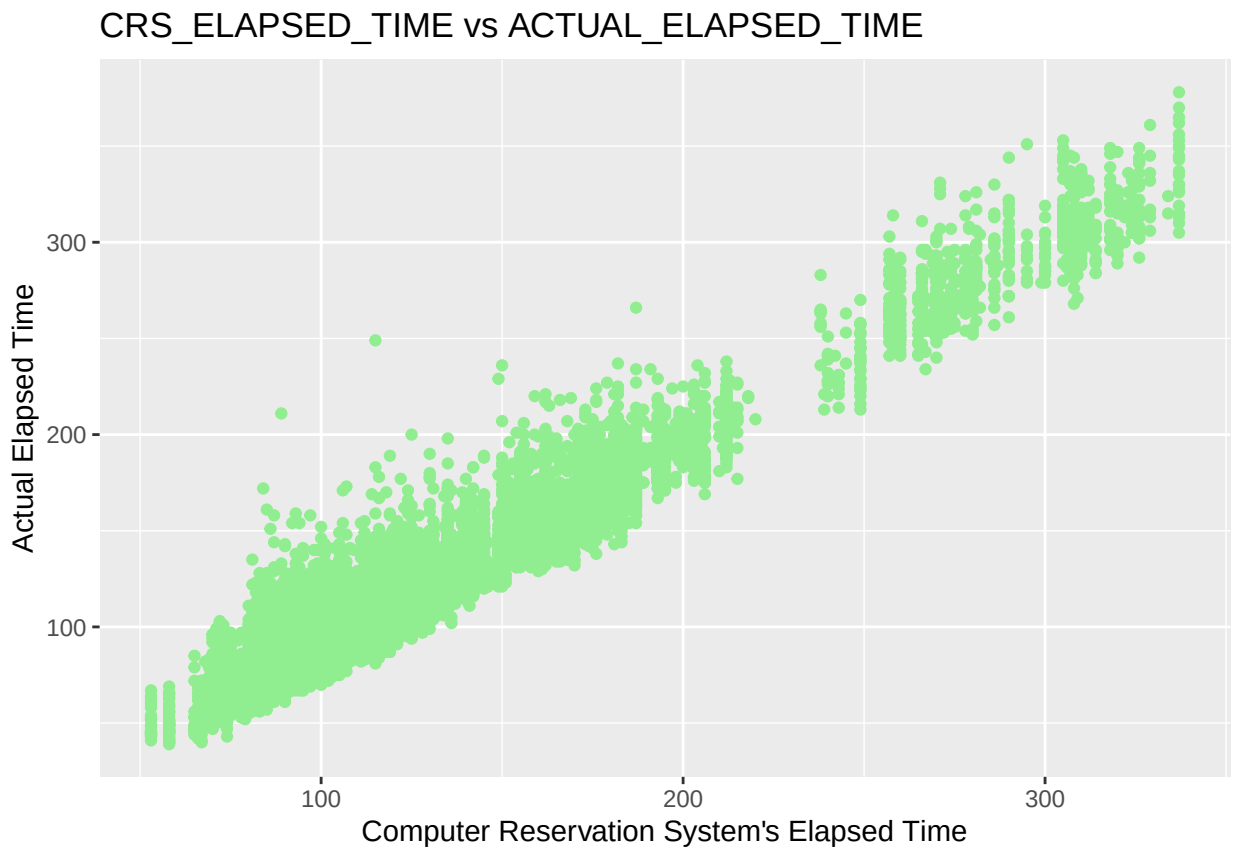
```
library(report)
report(result2_no)
```

```
## The ANOVA (formula: DEP_DELAY ~ MKT_UNIQUE_CARRIER) suggests that:
##
##   - The main effect of MKT_UNIQUE_CARRIER is statistically significant and small
## (F(8, 9950) = 55.90, p < .001; Eta2 = 0.04, 95% CI [0.04, 1.00])
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
## Hypothesis 3 : Regression for CRS_ELAPSED_TIME vs ACTUAL_ELAPSED_TIME ####
#Graph for Ohio Data - Reservation System's Elapsed Time vs Actual Elapsed Time
ggplot(ohdata2) +
  aes(x = CRS_ELAPSED_TIME, y = ACTUAL_ELAPSED_TIME) +
```

```
  geom_point(color = "Light Green")+
  labs(title = "CRS_ELAPSED_TIME vs ACTUAL_ELAPSED_TIME") +
  xlab("Computer Reservation System's Elapsed Time") + ylab("Actual Elapsed Time")
```

```
## Warning: Removed 138 rows containing missing values (‘geom_point()‘).
```

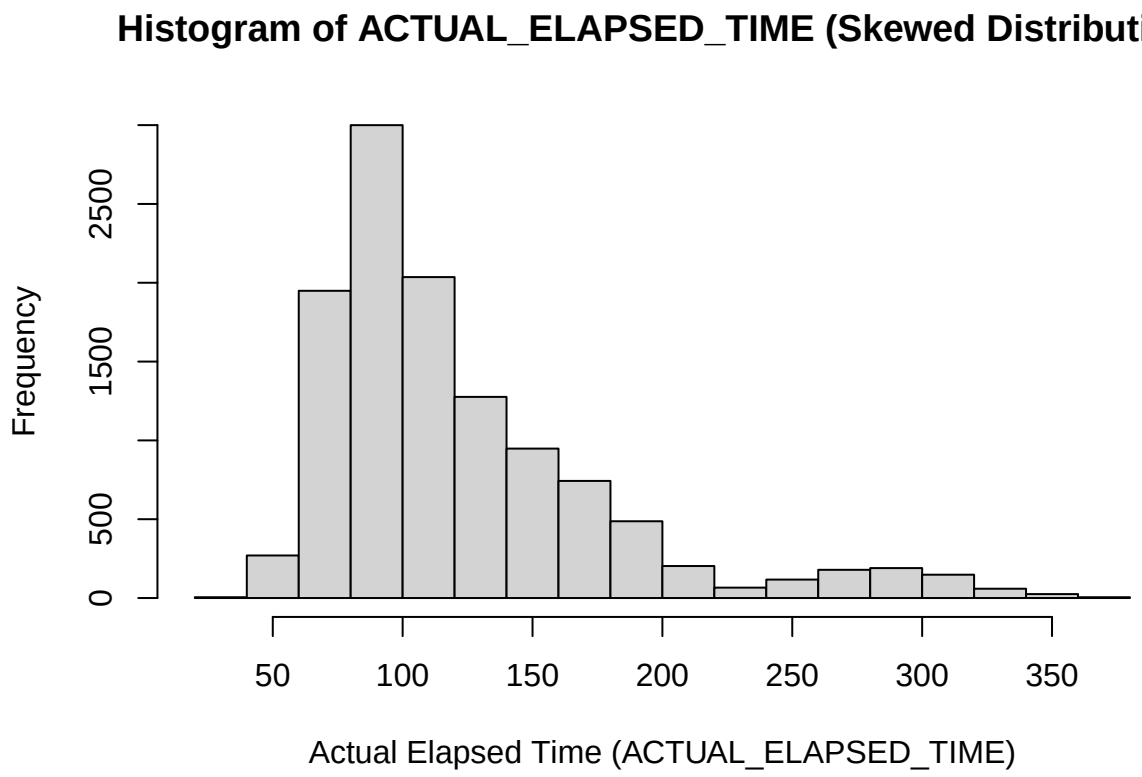



```
#Regression for Reservation System's Elapsed Time vs Actual Elapsed Time
# Elapsed Time = Diff b/w Departure Time and Arrival Time
#
# Ho: There is No Relation between CRS_ELAPSED_TIME and ACTUAL_ELAPSED_TIME
#     i.e., B1 = 0 (B is Beta)
# Ha: CRS_ELAPSED_TIME and ACTUAL_ELAPSED_TIME are related to each other
#     (positively or negatively)
#     i.e., B1 != 0
result3 <- lm(ohdata2$ACTUAL_ELAPSED_TIME ~ ohdata2$CRS_ELAPSED_TIME)
summary(result3)
```

```
##
## Call:
## lm(formula = ohdata2$ACTUAL_ELAPSED_TIME ~ ohdata2$CRS_ELAPSED_TIME)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.976  -8.698  -1.820   6.662 137.952
```

```
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -5.697745   0.315657  -18.05   <2e-16 ***
## ohdata2$CRS_ELAPSED_TIME  1.015176   0.002287  443.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.38 on 11704 degrees of freedom
##   (138 observations deleted due to missingness)
## Multiple R-squared:  0.944,  Adjusted R-squared:  0.9439
## F-statistic: 1.971e+05 on 1 and 11704 DF,  p-value: < 2.2e-16
```

```
hist(ohdata2$ACTUAL_ELAPSED_TIME, xlab = "Actual Elapsed Time (ACTUAL_ELAPSED_TIME)",
     main = "Histogram of ACTUAL_ELAPSED_TIME (Skewed Distribution)")
```




```
# Cleaning / Correcting and Selecting appropriate data
h3 <- ohdata2 %>% select(ORIGIN, CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME)
dim(h3)
```

```
## [1] 11844     3
```

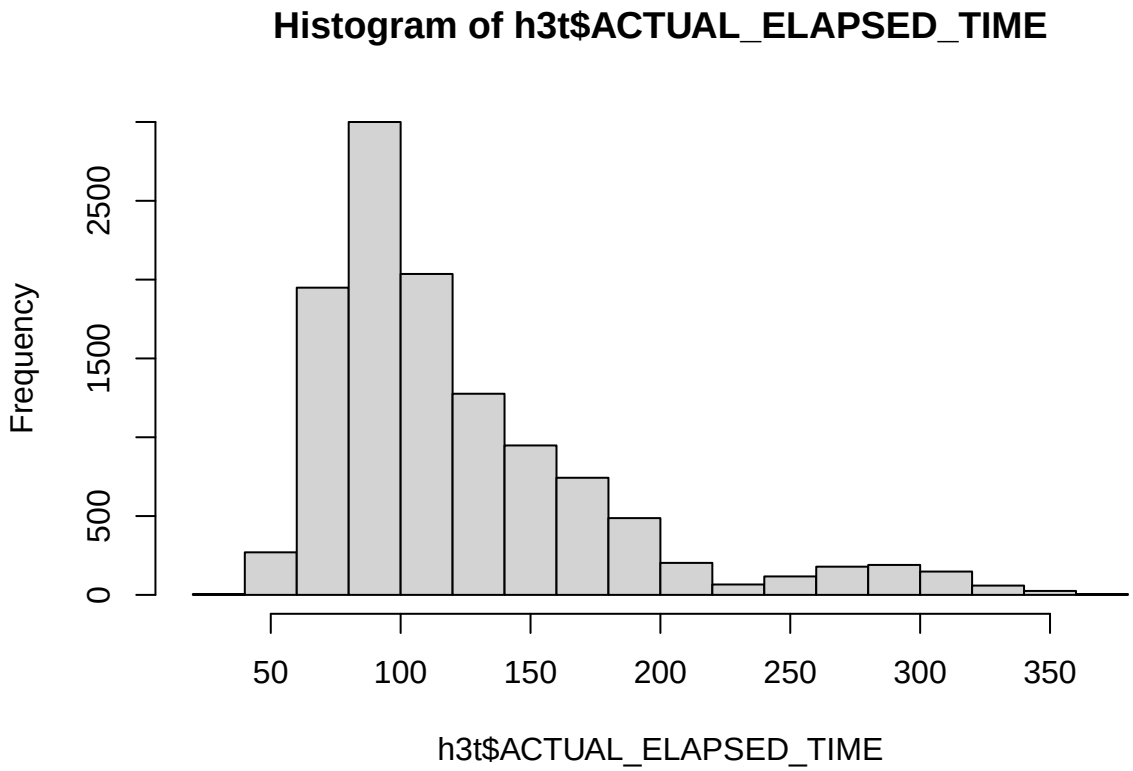
```
summary(h3)
```

```
##  ORIGIN     CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
##  CAK: 304   Min.   : 53.0    Min.   : 39.0
##  CLE:3505   1st Qu.: 90.0    1st Qu.: 85.0
##  CMH:3575   Median :110.0    Median :105.0
##  CVG:3544   Mean   :127.1    Mean   :123.2
##  DAY: 784   3rd Qu.:150.0    3rd Qu.:146.0
##  LCK:  76   Max.   :337.0    Max.   :378.0
##  TOL:  56                    NA's   :138
```

```
#Removing Null Values
h3t <-  na.omit(h3) #h3t is h3 after removing NAs
dim(h3t)
```

```
## [1] 11706     3
```

```
hist(h3t$ACTUAL_ELAPSED_TIME)
```



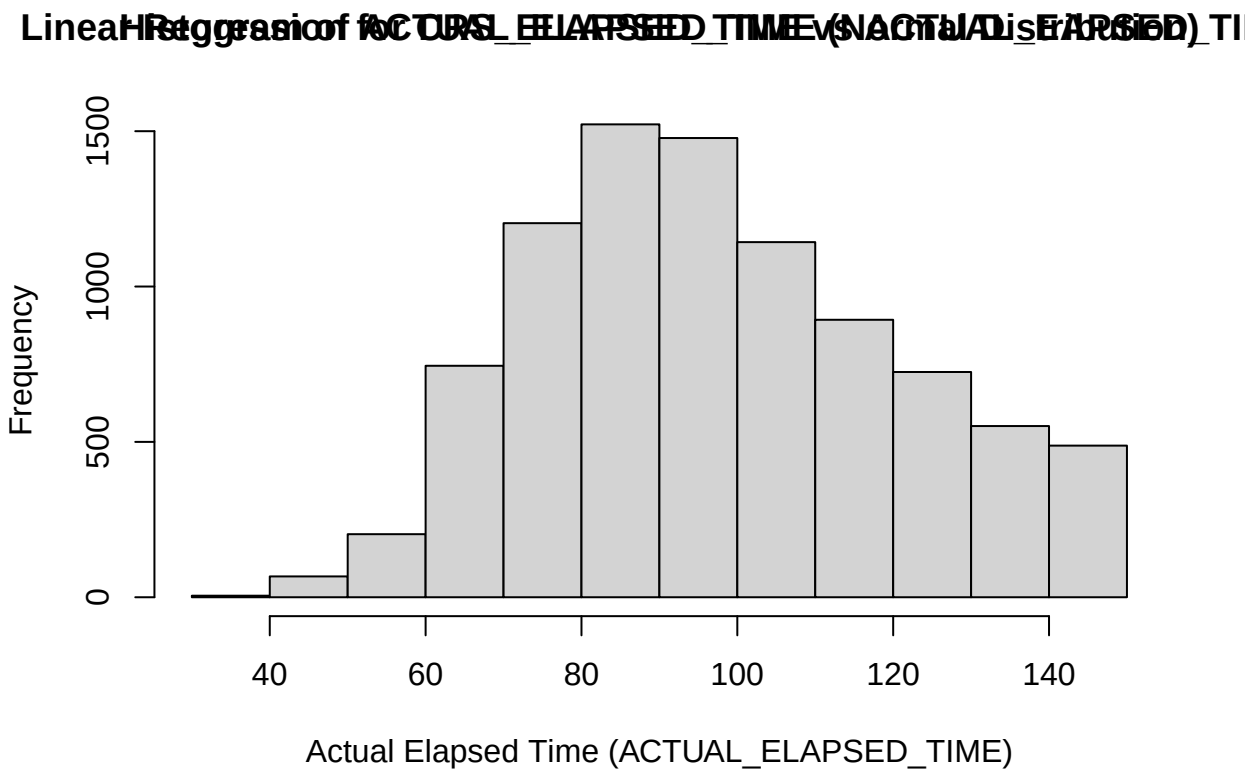

```
# Doesn't follow Normal Distribution. Deleting values after 180 to get NormDist
h3t_n <- h3t %>% filter(ACTUAL_ELAPSED_TIME <= 150) # h3t_n for NormDist Data
dim(h3t_n)
```

```
## [1] 9024    3
```

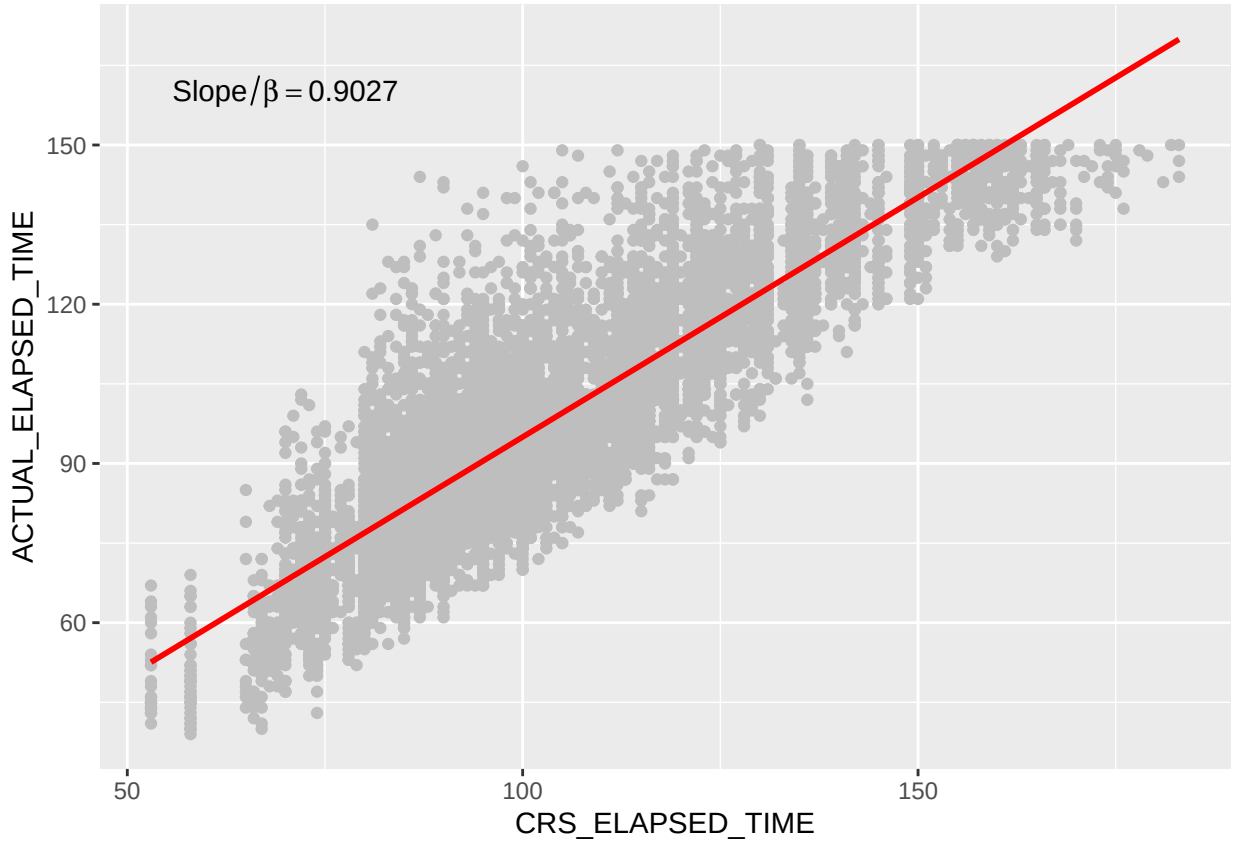
```r
hist(h3t_n$ACTUAL_ELAPSED_TIME, xlab = "Actual Elapsed Time (ACTUAL_ELAPSED_TIME)",main = "Histogram of
# h3t_n looks somewhat Normal Compared to h3t, So we will use h3t_n for Regression

result3_n <- lm(ACTUAL_ELAPSED_TIME ~ CRS_ELAPSED_TIME, data = h3t_n)
summary(result3_n)
```

```
##
## Call:
## lm(formula = ACTUAL_ELAPSED_TIME ~ CRS_ELAPSED_TIME, data = h3t_n)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.516  -7.819  -1.570   6.181  60.749
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.716930   0.559263   8.434   <2e-16 ***
## CRS_ELAPSED_TIME 0.902694   0.005275 171.114   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.35 on 9022 degrees of freedom
## Multiple R-squared:  0.7645, Adjusted R-squared:  0.7644
## F-statistic: 2.928e+04 on 1 and 9022 DF,  p-value: < 2.2e-16
```

```r
ggplot(h3t_n, mapping = aes(CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME)) +
  geom_point(color = "Grey")+
  geom_smooth(method = "lm",se =  FALSE ,color = "Red") +
  title(main = "Linear Regression for CRS_ELAPSED_TIME vs ACTUAL_EAPSED_TIME") +
  annotate("text",x=70,y=160,label=(paste0("Slope / beta ==",round(coef(result3_n)[2],4))),parse=TRUE)
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



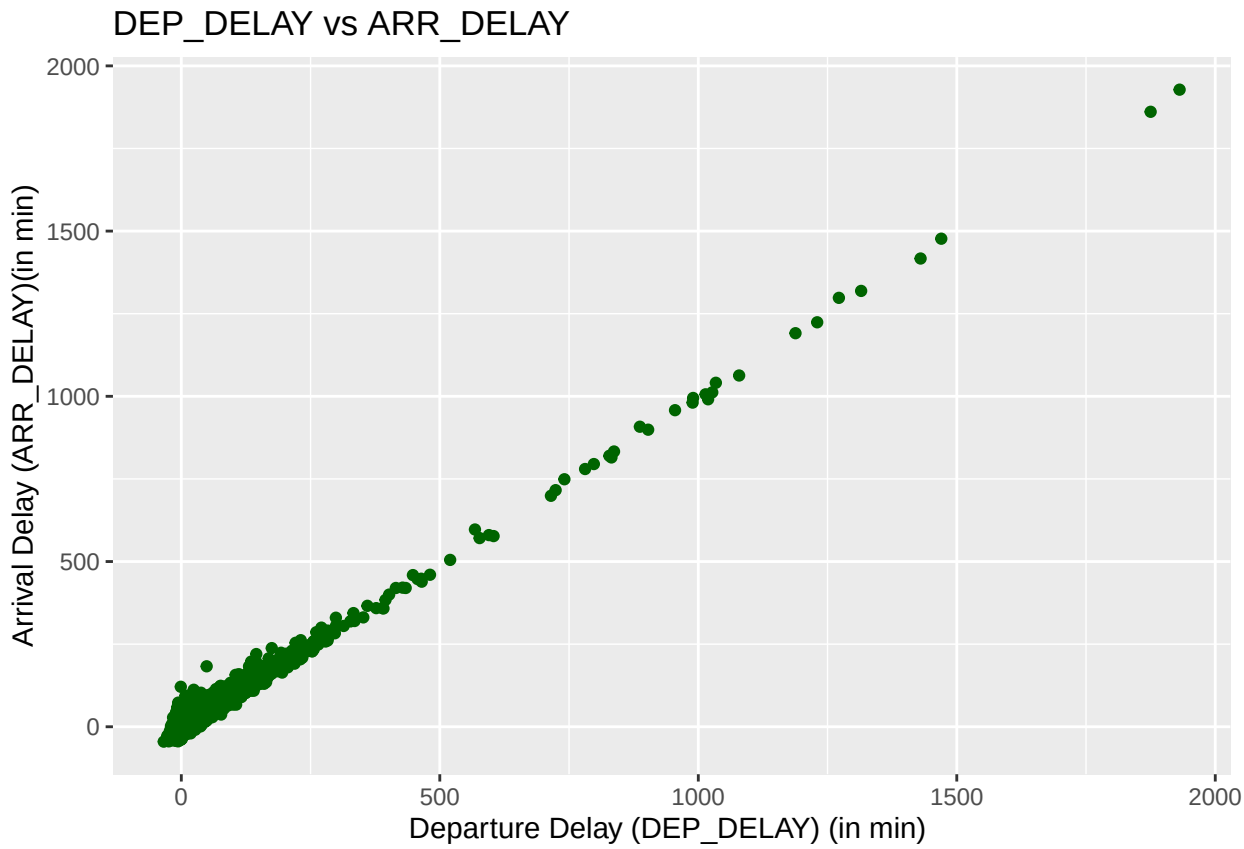

```
#Can add Equation on Graph here

library(report)
report(result3_n)
```

```
## We fitted a linear model (estimated using OLS) to predict ACTUAL_ELAPSED_TIME
## with CRS_ELAPSED_TIME (formula: ACTUAL_ELAPSED_TIME ~ CRS_ELAPSED_TIME). The
## model explains a statistically significant and substantial proportion of
## variance (R2 = 0.76, F(1, 9022) = 29280.00, p < .001, adj. R2 = 0.76). The
## model's intercept, corresponding to CRS_ELAPSED_TIME = 0, is at 4.72 (95% CI
## [3.62, 5.81], t(9022) = 8.43, p < .001). Within this model:
##
##   - The effect of CRS ELAPSED TIME is statistically significant and positive
## (beta = 0.90, 95% CI [0.89, 0.91], t(9022) = 171.11, p < .001; Std. beta =
## 0.87, 95% CI [0.86, 0.88])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```
## Hypothesis 4 : Regression for DEP_DELAY vs ARR_DELAY ####
#Graph for Ohio Data - Departure Delay vs Arrival Delay
ggplot(ohdata2) +
  aes(x = DEP_DELAY, y = ARR_DELAY) +
  geom_point(color = "Dark Green")+
```

```
  labs(title = "DEP_DELAY vs ARR_DELAY") +
  xlab("Departure Delay (DEP_DELAY) (in min) ") + ylab("Arrival Delay (ARR_DELAY)(in min)")
```

```
## Warning: Removed 138 rows containing missing values (`geom_point()`).
```




```
#Regression for Departure Delay vs Arrival Delay
#
# Ho: There is Positive or No Relation between DEP_DELAY and ARR_DELAY
#     i.e., B1 >= 0
# Ha: DEP_DELAY and ARR_DELAY are related to each other(positively or negatively)
# DEP_DELAY doesn't affect ARR_DELAY
#     i.e., B1 < 0
result4 <- lm(ohdata2$ARR_DELAY ~ ohdata2$DEP_DELAY)
summary(result4)
```

```
##
## Call:
## lm(formula = ohdata2$ARR_DELAY ~ ohdata2$DEP_DELAY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.465  -9.174  -2.187   6.803 137.629
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.792909   0.125513  -30.22   <2e-16 ***
## ohdata2$DEP_DELAY  1.003351   0.001922  522.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 11704 degrees of freedom
##   (138 observations deleted due to missingness)
## Multiple R-squared:  0.9588, Adjusted R-squared:  0.9588
## F-statistic: 2.726e+05 on 1 and 11704 DF,  p-value: < 2.2e-16
```

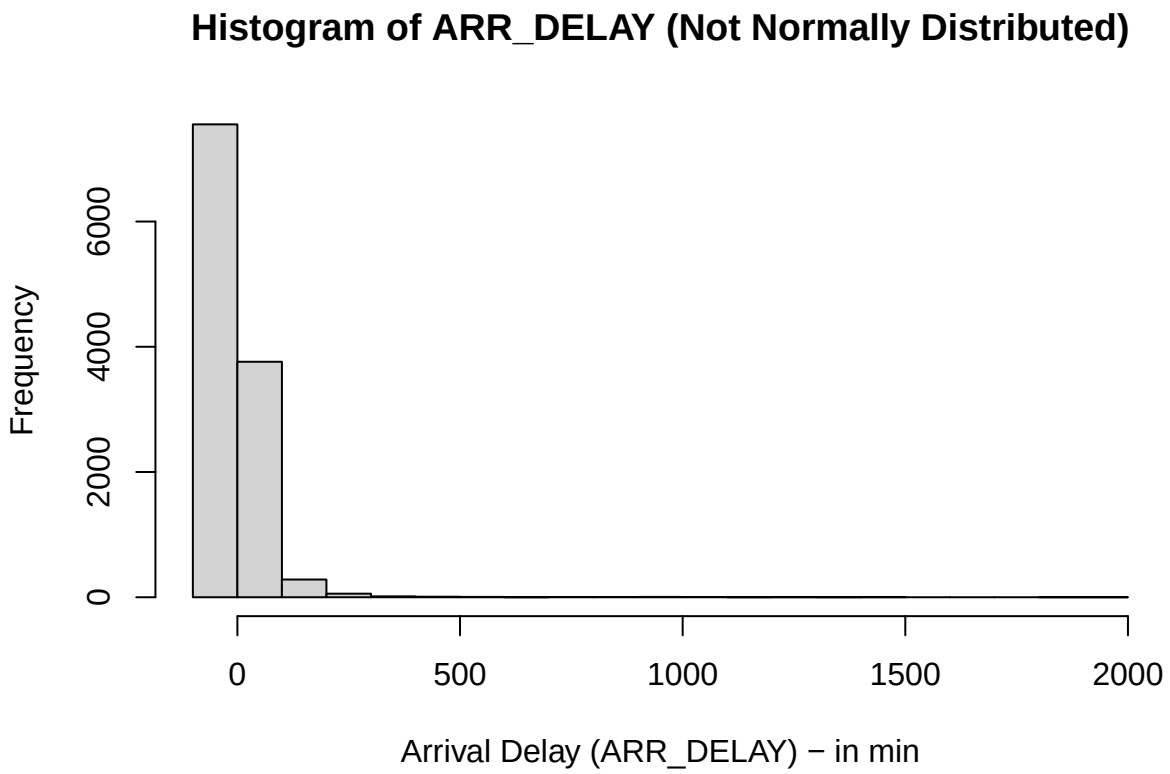
```r
#Selecting Required Data
h4 <- ohdata2 %>% select(ORIGIN, DEP_DELAY, ARR_DELAY)
summary(h4)
```

```
##  ORIGIN       DEP_DELAY          ARR_DELAY
##  CAK: 304   Min.   : -34.000   Min.   : -45.0
##  CLE:3505   1st Qu.:  -7.000   1st Qu.: -16.0
##  CMH:3575   Median :  -3.000   Median :  -6.0
##  CVG:3544   Mean   :   9.905   Mean   :   6.1
##  DAY: 784   3rd Qu.:   3.000   3rd Qu.:   8.0
##  LCK:  76   Max.   :1931.000   Max.   :1928.0
##  TOL:  56   NA's   :116        NA's   :138
```

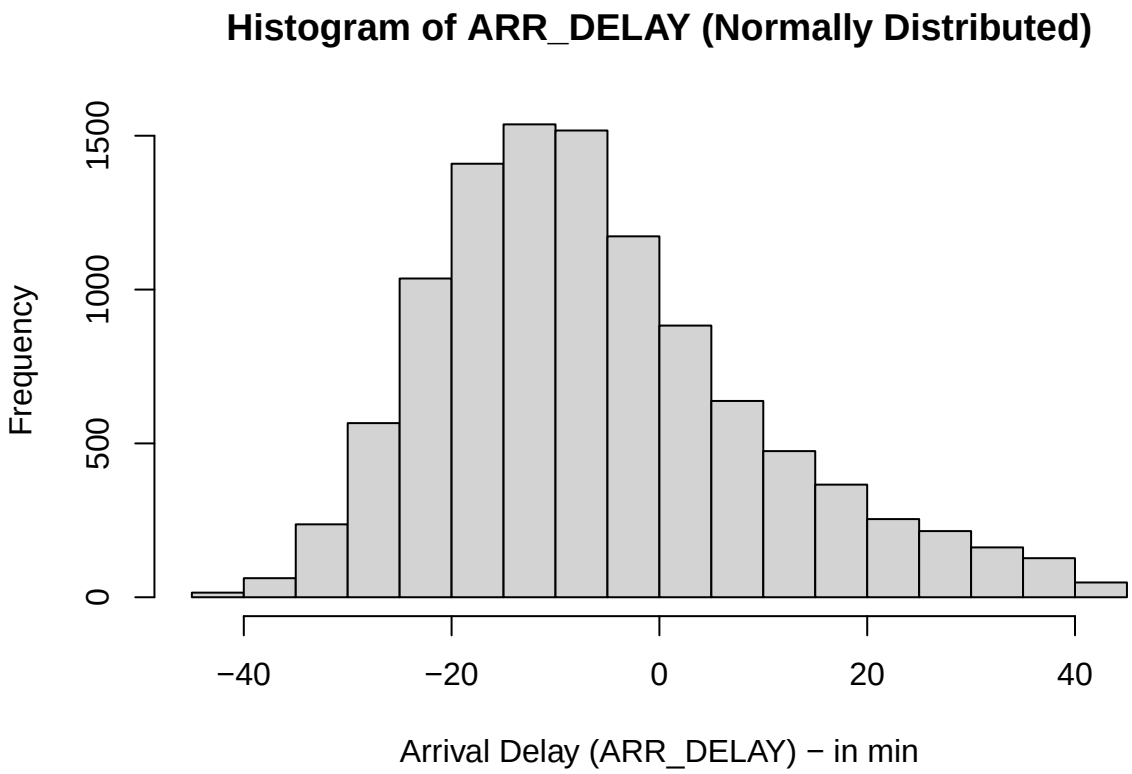
```r
dim(h4)
```

```
## [1] 11844     3
```

```r
#Removing Null Values
h4t <- na.omit(h4)
hist(h4t$ARR_DELAY, xlab = "Arrival Delay (ARR_DELAY) - in min", main = "Histogram of ARR_DELAY (Not No
```



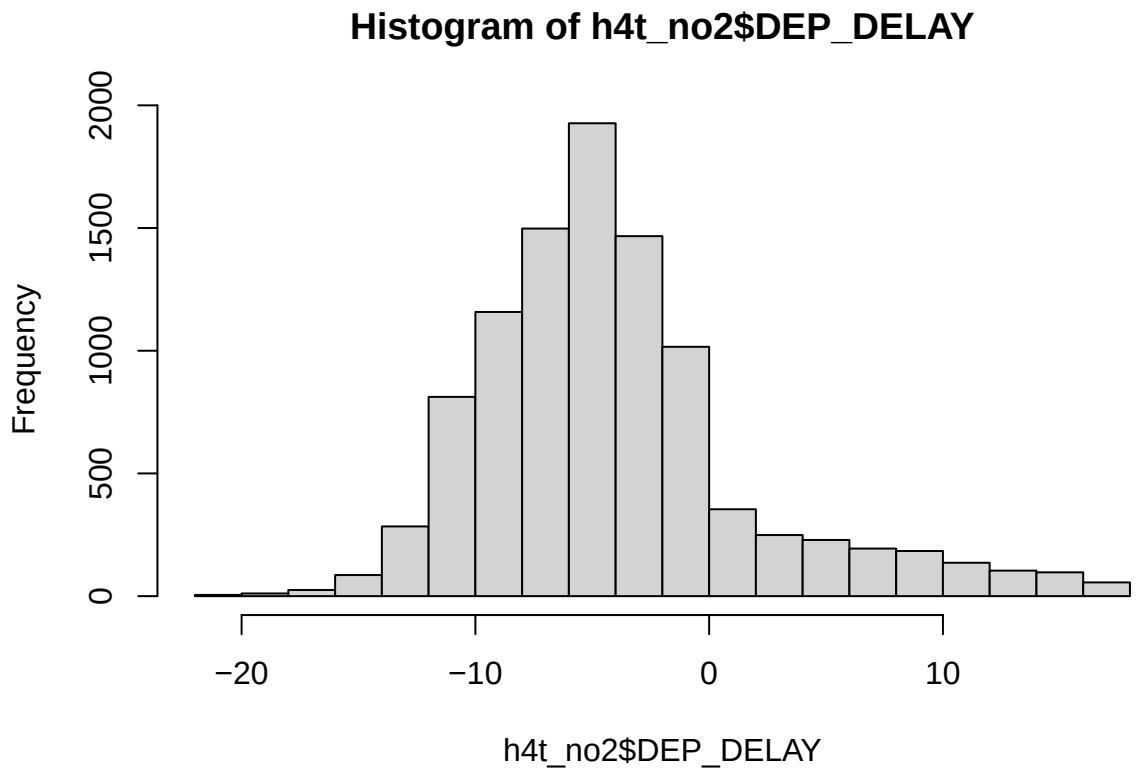

```r
#Doesn't look NormDist; Removing Outliers to change it to NormDist
Q1_4 <- quantile(h4t$ARR_DELAY, 0.25)
Q3_4 <- quantile(h4t$ARR_DELAY, 0.75)
IQR_4 <- IQR(h4t$ARR_DELAY)
h4t_no <- subset(h4t,
                 h4t$ARR_DELAY > (Q1_4 - 1.5*IQR_4) &
                   h4t$ARR_DELAY < (Q3_4 + 1.5*IQR_4))

hist(h4t_no$ARR_DELAY, xlab = "Arrival Delay (ARR_DELAY) - in min", main = "Histogram of ARR_DELAY (Nor
```




```
#Now it looks like Normal Distribution

#Applying Regression Analysis
result4_no <- lm(ARR_DELAY ~ DEP_DELAY, data = h4t_no)
summary(result4_no)
```

```
##
## Call:
## lm(formula = ARR_DELAY ~ DEP_DELAY, data = h4t_no)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.926  -8.247  -1.461   6.716  54.646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.71606    0.11643  -40.51   <2e-16 ***
## DEP_DELAY    0.89300    0.01033   86.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.01 on 10718 degrees of freedom
## Multiple R-squared:  0.4108, Adjusted R-squared:  0.4108
## F-statistic:  7473 on 1 and 10718 DF,  p-value: < 2.2e-16
```

```r
#Still getting -ve coeff even though Graph is +ve Linear
# Working on DEP_DELAY Outliers
# Using Same data from Hypothesis 1
h4t_no2 <- subset(h4t_no,
                  h4t_no$DEP_DELAY > (Q1_1 - 1.5*IQR_1) &
                    h4t_no$DEP_DELAY < (Q3_1 + 1.5*IQR_1))

hist(h4t_no2$DEP_DELAY)
```



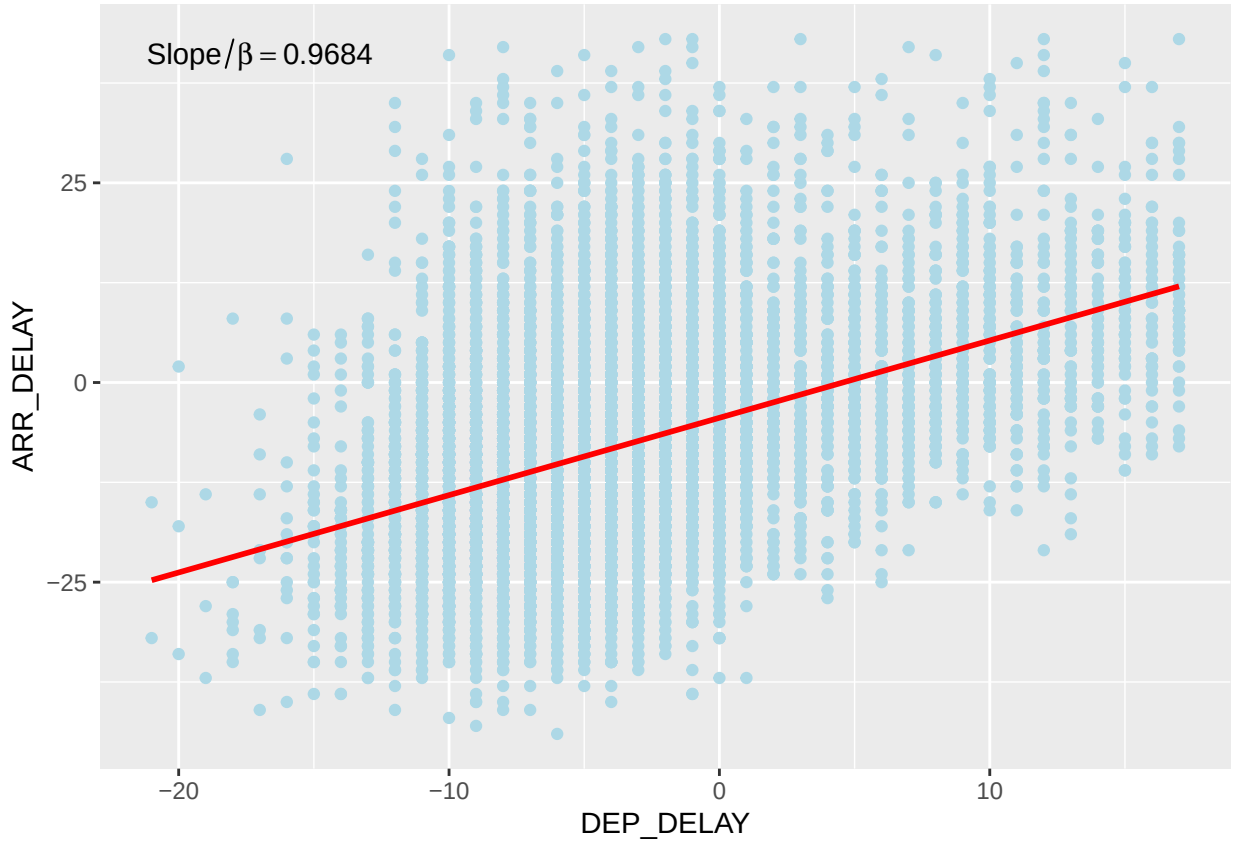

```r
result4_no2 <- lm(ARR_DELAY ~ DEP_DELAY, data = h4t_no2)
summary(result4_no2)
```

```
##
## Call:
## lm(formula = ARR_DELAY ~ DEP_DELAY, data = h4t_no2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.769  -8.358  -1.675   6.736  55.104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.42028    0.14270  -30.98   <2e-16 ***
## DEP_DELAY    0.96839    0.02065   46.90   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.17 on 9890 degrees of freedom
## Multiple R-squared:  0.1819, Adjusted R-squared:  0.1819
## F-statistic:  2199 on 1 and 9890 DF,  p-value: < 2.2e-16
```

```
ggplot(h4t_no2, mapping = aes(DEP_DELAY, ARR_DELAY)) +
  geom_point(color = "Light Blue")+
  geom_smooth(method = "lm",se = FALSE ,color = "Red") +
  annotate("text",x = -17 ,y = 41,
           label=(paste0("Slope / beta ==",round(coef(result4_no2)[2],4))),
           parse = T)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



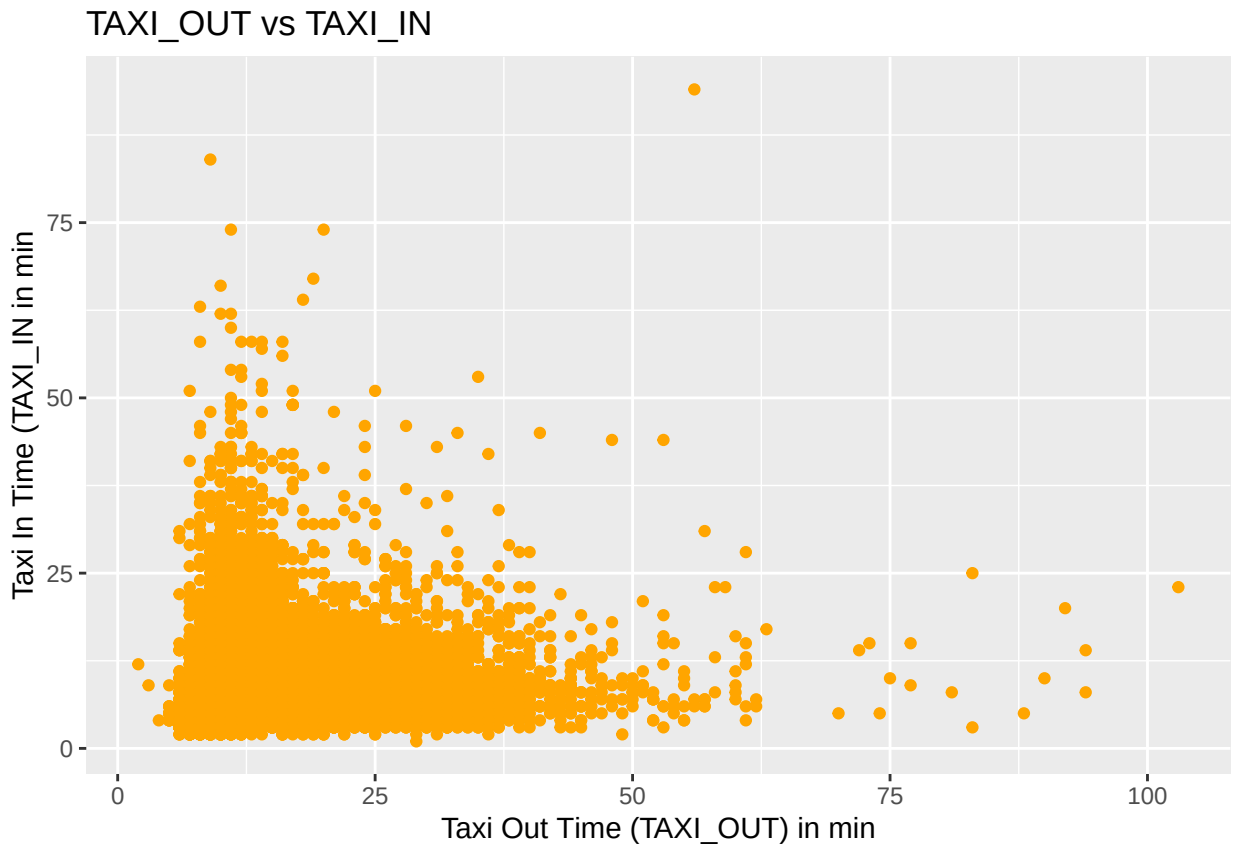

```
#Can add Equation on Graph here

library(report)
report(result4_no2)
```

```
## We fitted a linear model (estimated using OLS) to predict ARR_DELAY with
## DEP_DELAY (formula: ARR_DELAY ~ DEP_DELAY). The model explains a statistically
## significant and moderate proportion of variance (R2 = 0.18, F(1, 9890) =
## 2199.49, p < .001, adj. R2 = 0.18). The model's intercept, corresponding to
```

```
## DEP_DELAY = 0, is at -4.42 (95% CI [-4.70, -4.14], t(9890) = -30.98, p < .001).
## Within this model:
##
##   - The effect of DEP DELAY is statistically significant and positive (beta =
## 0.97, 95% CI [0.93, 1.01], t(9890) = 46.90, p < .001; Std. beta = 0.43, 95% CI
## [0.41, 0.44])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```
## Hypothesis 5 : Regression for TAXI_OUT vs TAXI_IN ####
#Graph for Ohio Data - Taxi Out Time vs Taxi In Time
ggplot(ohdata2) +
  aes(x = TAXI_OUT, y = TAXI_IN) +
  geom_point(color = "Orange") + xlab("Taxi Out Time (TAXI_OUT) in min") + ylab("Taxi In Time (TAXI_IN
  labs(title ="TAXI_OUT vs TAXI_IN")
```

```
## Warning: Removed 119 rows containing missing values (`geom_point()`).
```




```
# Graph doesn't have a Linear Relation
# Regression for Taxi Out Time vs Taxi In Time
# Taxi Out Time: Time Diff b/w Departure Time from  Airport Gate till Wheel Off
# at DEP Airport
# Taxi In Time: Time Diff b/w Wheel On Ground and Arrival at Airport Gate at
```

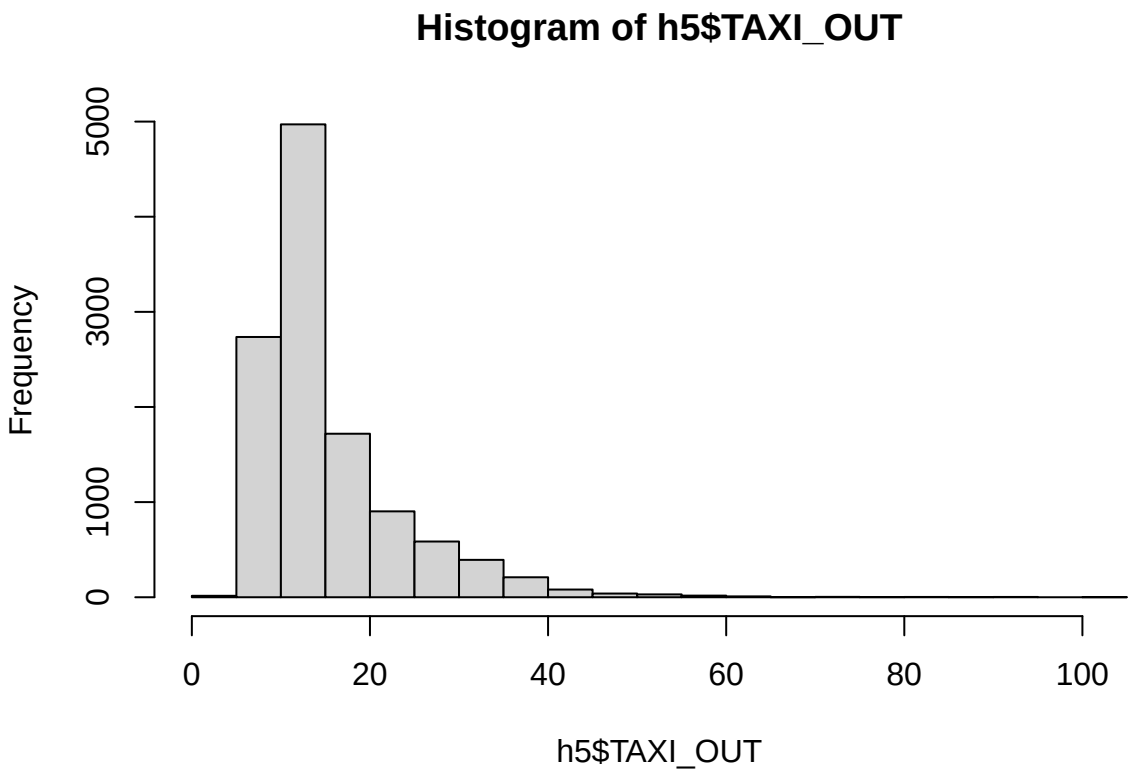
```r
# AVL Airport
# Ho: There is No Relation between TAXI_OUT and TAXI_IN Time
#     i.e., B1 = 0 (B is Beta)
# Ha: TAXI_OUT and TAXI_IN Time are related to each other
#     (positively or negatively)
#     i.e., B1 != 0

#Selecting Required Data:
h5 <- ohdata2 %>% select(ORIGIN,TAXI_OUT, TAXI_IN)
dim(h5)
```

```
## [1] 11844     3
```

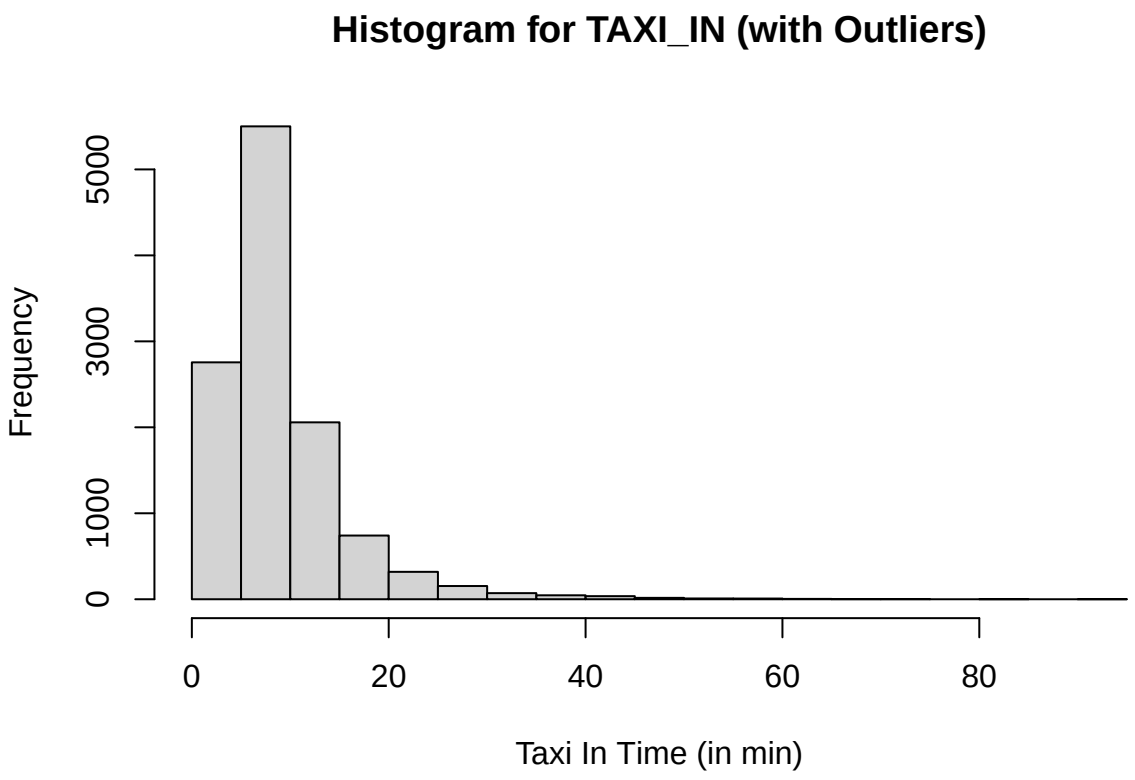
```r
result5 <- lm(TAXI_IN ~ TAXI_OUT, data = h5)
summary(result5)
```

```
##
## Call:
## lm(formula = TAXI_IN ~ TAXI_OUT, data = h5)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.164 -3.891 -1.588  1.639 82.859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.019824   0.129566  69.616  < 2e-16 ***
## TAXI_OUT    0.037883   0.007258   5.219 1.83e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.527 on 11723 degrees of freedom
##   (119 observations deleted due to missingness)
## Multiple R-squared:  0.002318,   Adjusted R-squared:  0.002233
## F-statistic: 27.24 on 1 and 11723 DF,  p-value: 1.828e-07
```

```r
hist(h5$TAXI_OUT) # Not in NormDist
```



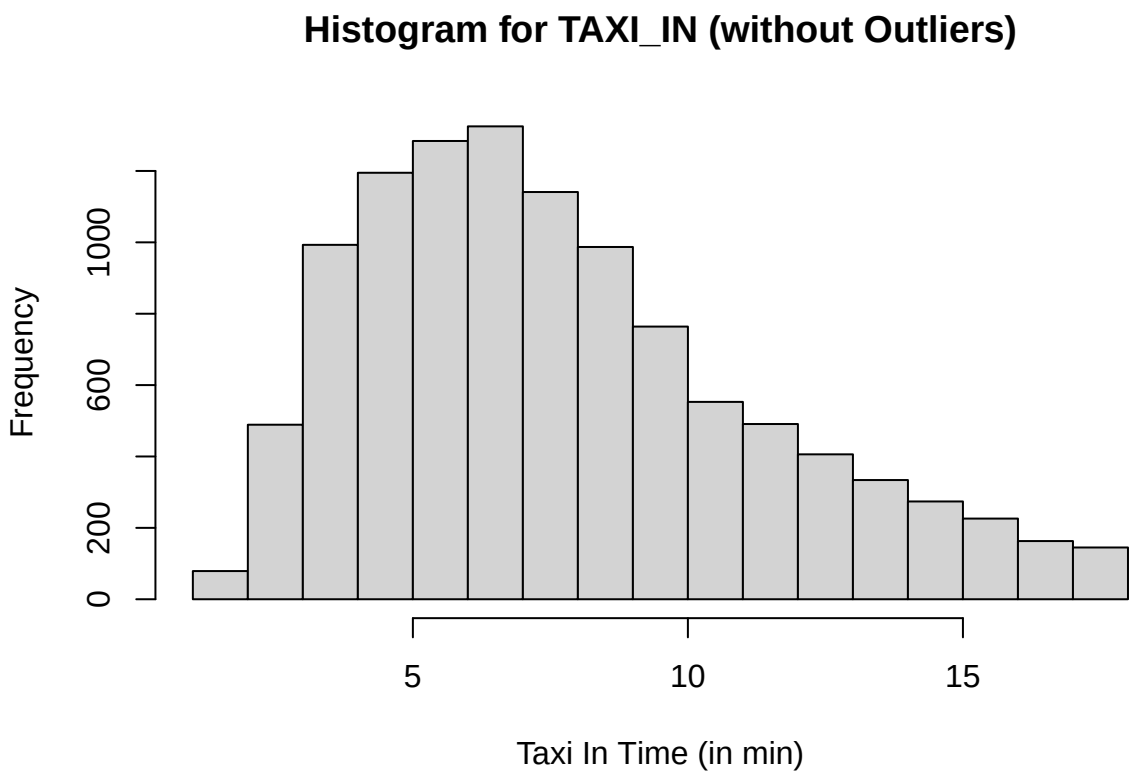

```
hist(h5$TAXI_IN, xlab = "Taxi In Time (in min)", main = "Histogram for TAXI_IN (with Outliers)") # Not
```




```
#Removing Null Values:
h5t <- na.omit(h5) #h5t has no NA values
dim(h5t)
```

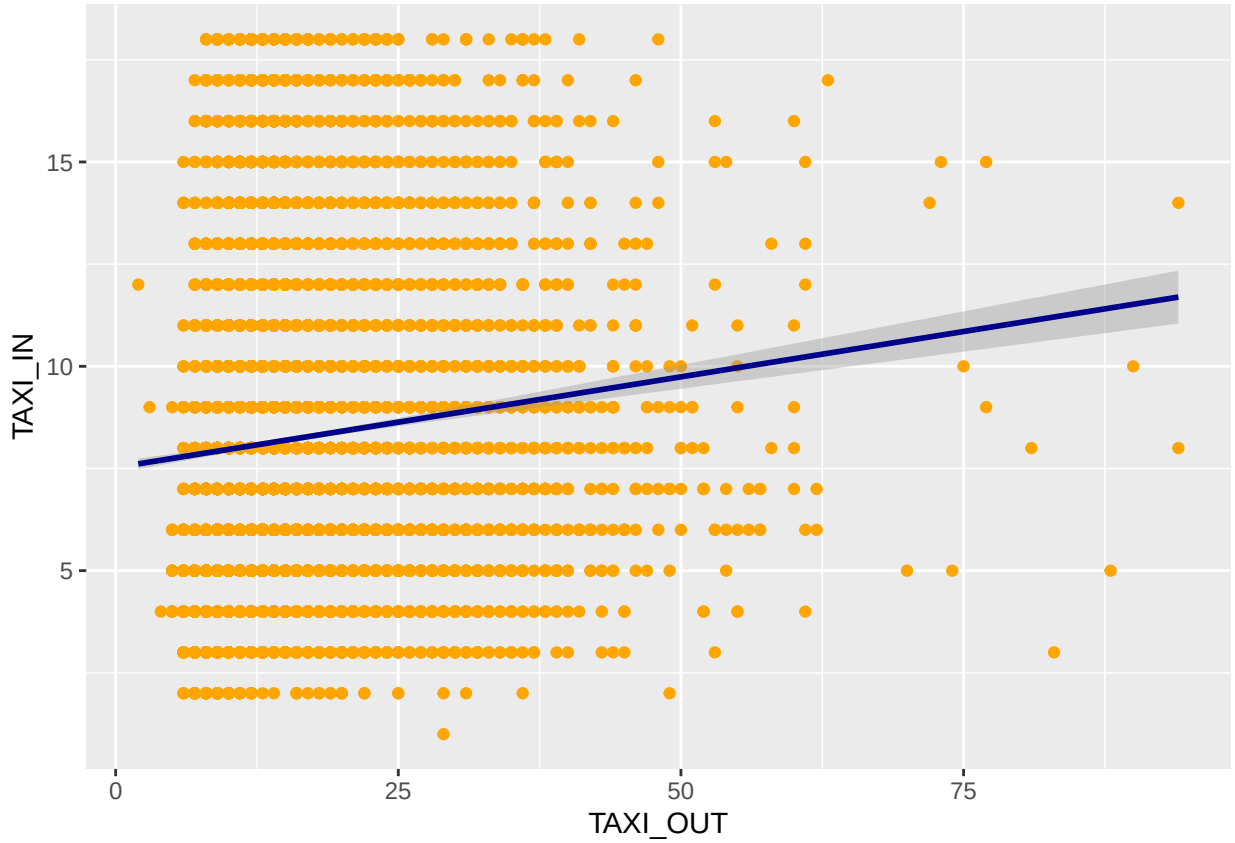
```
## [1] 11725     3
```

```
#Removing Outliers for TAXI_IN
Q1_5 <- quantile(h5t$TAXI_IN, 0.25)
Q3_5 <- quantile(h5t$TAXI_IN, 0.75)
IQR_5 <- IQR(h5t$TAXI_IN)
h5t_no <- subset(h5t,
                 h5t$TAXI_IN > (Q1_5 - 1.5*IQR_5) &
                   h5t$TAXI_IN < (Q3_5 + 1.5*IQR_5))

hist(h5t_no$TAXI_IN, xlab = "Taxi In Time (in min)", main = "Histogram for TAXI_IN (without Outliers)")
```

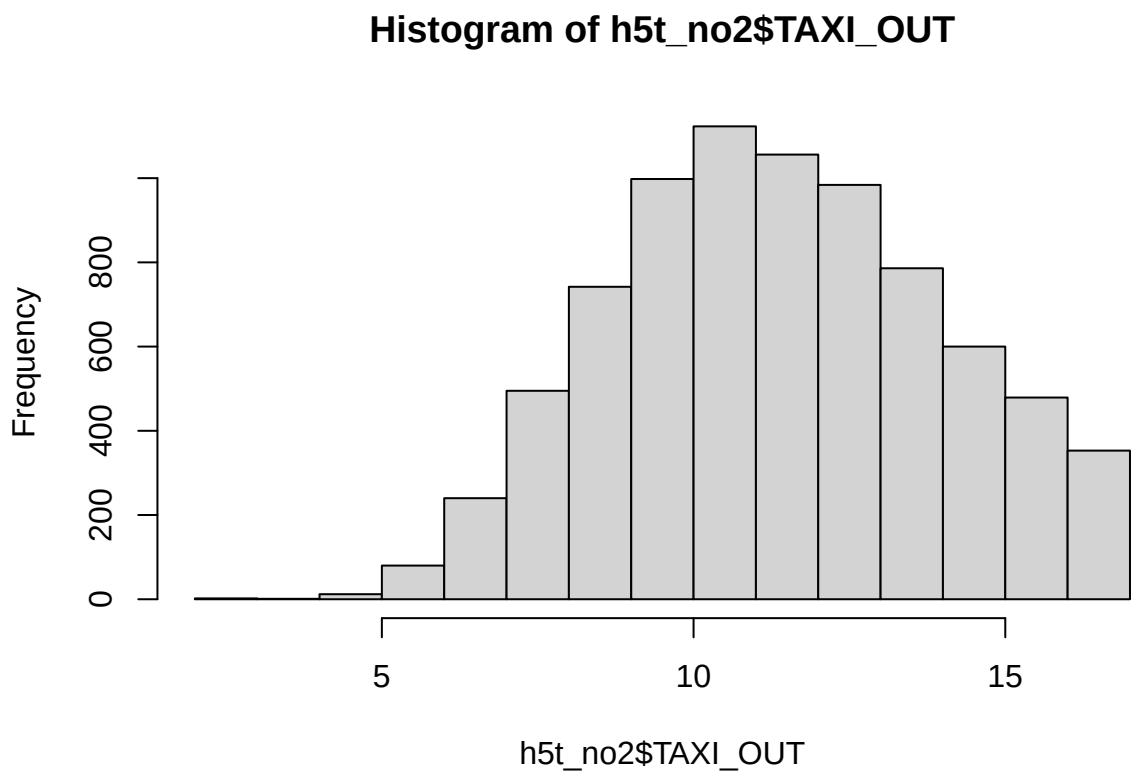



```
# After Removing Outliers from Taxi In
ggplot(data = h5t_no, mapping = aes(x = TAXI_OUT, y = TAXI_IN)) +
  geom_point(color = "Orange") +
  geom_smooth(method = "lm", color = "Dark Blue")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Using Data from h2 for TAXI_OUT Outliers
h5t_no2 <- subset(h5t_no,
                  h5t_no$TAXI_OUT > (Q1_1 - 1.5*IQR_1) &
                    h5t_no$TAXI_OUT < (Q3_1 + 1.5*IQR_1))

hist(h5t_no2$TAXI_OUT)
```



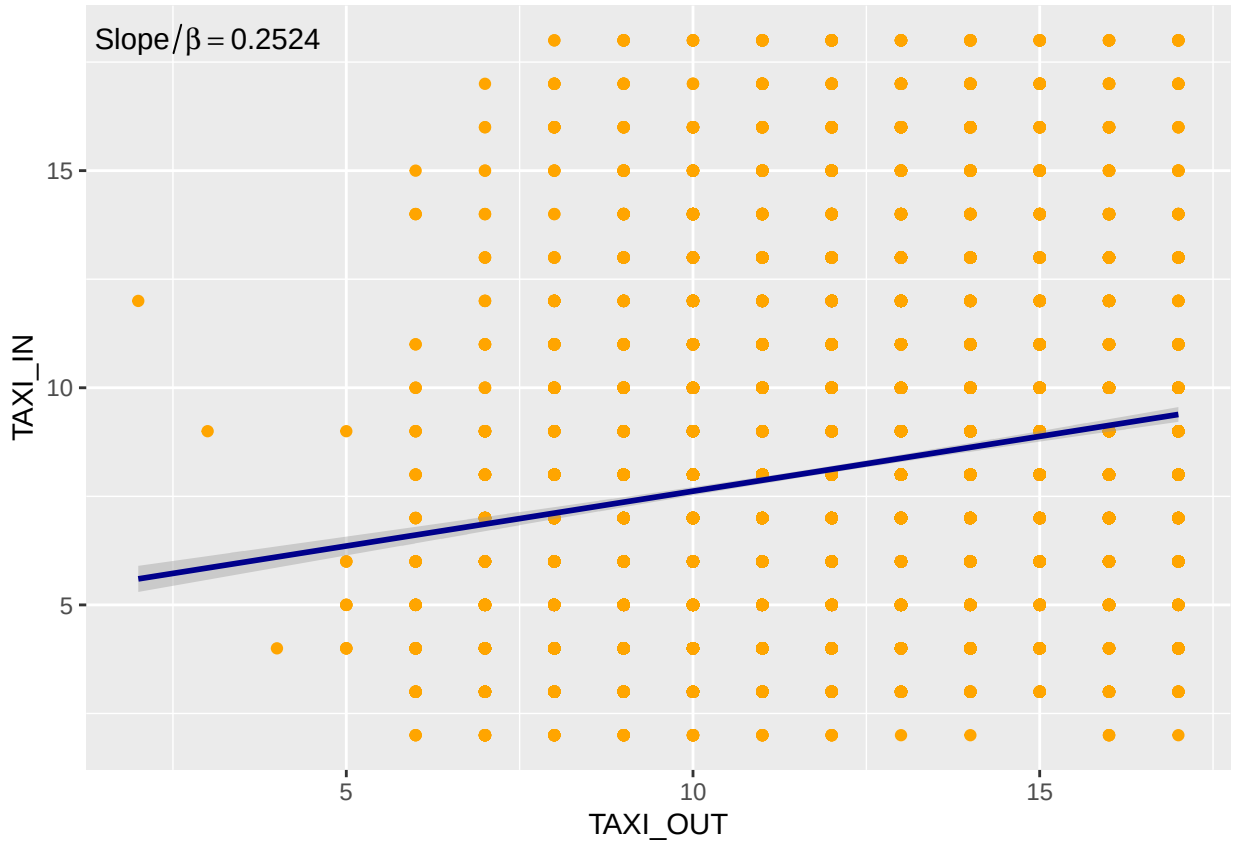

```
result5_no2 <- lm(TAXI_IN ~ TAXI_OUT, data = h5t_no2)
summary(result5_no2)
```

```
##
## Call:
## lm(formula = TAXI_IN ~ TAXI_OUT, data = h5t_no2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3855 -2.6282 -0.6184  2.1194 10.8865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.09402    0.18243   27.92   <2e-16 ***
## TAXI_OUT     0.25244    0.01501   16.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.53 on 7949 degrees of freedom
## Multiple R-squared:  0.03436,    Adjusted R-squared:  0.03424
## F-statistic: 282.8 on 1 and 7949 DF,  p-value: < 2.2e-16
```

```
#After removing Outliers from Taxi Out and Using lm()
ggplot(data = h5t_no2, mapping = aes(x = TAXI_OUT, y = TAXI_IN)) +
  geom_point(color = "Orange") +
```

```
  geom_smooth(method = "lm", color = "Dark Blue") +
  annotate("text",x = 3 ,y = 18,
           label=(paste0("Slope / beta ==",round(coef(result5_no2)[2],4))),
           parse = T)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
library(report)
report(result5_no2)
```

```
## We fitted a linear model (estimated using OLS) to predict TAXI_IN with TAXI_OUT
## (formula: TAXI_IN ~ TAXI_OUT). The model explains a statistically significant
## and weak proportion of variance (R2 = 0.03, F(1, 7949) = 282.85, p < .001, adj.
## R2 = 0.03). The model's intercept, corresponding to TAXI_OUT = 0, is at 5.09
## (95% CI [4.74, 5.45], t(7949) = 27.92, p < .001). Within this model:
##
##   - The effect of TAXI OUT is statistically significant and positive (beta =
## 0.25, 95% CI [0.22, 0.28], t(7949) = 16.82, p < .001; Std. beta = 0.19, 95% CI
## [0.16, 0.21])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

```
##
# Considering all the Results, we can say that all the Null Hypothesis are
#  Rejected except for Hypothesis 4
```