Airline Data Analysis for Ohio State

ADTA 5130

Data Analysis I

Group 14

Chandra Sekhar Neelam – 11646448
Gopi Srinivas Yerramothu - 11654173
Namrata Sood - 11664403
Sreelekha Onteddu - 11661665
Venu Gopalan Krishnagiri Tuppal - 11613143

Under the guidance of

Dr. Amritha Thomas Ph.D.

Clinical Assistant Professor
Advanced Data Analytics

Contents

Research / Study Aim

The purpose of this research is to study the Airline Data in the state of Ohio and understand various aspects of it. The focus of this study is to find if Delays of the Flights are affected by factors such as Departing Airport or the Airline Service Provider. This study also covers the aspects such as relations between Delays, and other Time factors which are used to study the Travel time of the flight.

Introduction

Data Description

The Airline data contains all the data related to the flights that travelled in the Month of March 2023. The dataset has 33 Attributes with 616,234 Observations that describe the Journey of the Flight and Time factors related to it.

For the study, the data is focused on Ohio State. Ohio State data consists of 33 Attributes with 11,844 Observations. Following is complete summary of Ohio Data (with needed Attributes)

```
> summary(ohdata2)
      DATE             DAY_OF_WEEK MKT_UNIQUE_CARRIER MKT_CARRIER_FL_NUM ORIGIN           ORIGIN_CITY_NAME       DEST     
 Min.   :2023-03-01   1:1606      AA     :3401       1437   :   62      CAK: 304   Akron, OH      : 304   ORD    :1074  
 1st Qu.:2023-03-09   2:1460      DL     :2825       4685   :   62      CLE:3505   Cincinnati, OH:3544   ATL    : 798  
 Median :2023-03-16   3:1855      UA     :2161       5093   :   61      CMH:3575   Cleveland, OH :3505   LGA    : 791  
 Mean   :2023-03-16   4:2010      WN     :1663       419    :   60      CVG:3544   Columbus, OH  :3651   CLT    : 675  
 3rd Qu.:2023-03-24   5:2009      F9     : 613       1038   :   58      DAY: 784   Dayton, OH    : 784   DCA    : 630  
 Max.   :2023-03-31   6:1372      G4     : 549       1043   :   58      LCK:  76   Toledo, OH    :  56   EWR    : 605  
                      7:1532      (Other): 632       (Other):11483      TOL:  56                         (Other):7271  
        DEST_CITY_NAME     DEP_TIME       DEP_DELAY           TAXI_OUT      WHEELS_OFF     WHEELS_ON       TAXI_IN      
 Chicago, IL    :1476   Min.   :   1   Min.   : -34.000   Min.   :  2.0   Min.   :   1   Min.   :   1   Min.   : 1.000  
 New York, NY   :1287   1st Qu.: 750   1st Qu.:  -7.000   1st Qu.: 11.0   1st Qu.: 806   1st Qu.: 917   1st Qu.: 6.000  
 Washington, DC: 994    Median :1222   Median :  -3.000   Median : 13.0   Median :1235   Median :1325   Median : 8.000  
 Atlanta, GA    : 798   Mean   :1232   Mean   :   9.905   Mean   : 15.8   Mean   :1259   Mean   :1353   Mean   : 9.618  
 Charlotte, NC : 675    3rd Qu.:1659   3rd Qu.:   3.000   3rd Qu.: 18.0   3rd Qu.:1714   3rd Qu.:1816   3rd Qu.:11.000  
 Newark, NJ     : 605   Max.   :2348   Max.   :1931.000   Max.   :103.0   Max.   :2358   Max.   :2400   Max.   :94.000  
 (Other)        :6009   NA's   :116    NA's   :116        NA's   :119     NA's   :119    NA's   :119    NA's   :119     
    ARR_TIME       ARR_DELAY        CANCELLED CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME    AIR_TIME         DISTANCE     
 Min.   :   1   Min.   : -45.0   0:11725   Min.   : 53.0    Min.   : 39.0        Min.   : 22.00   Min.   :  95.0  
 1st Qu.: 924   1st Qu.: -16.0   1:  119   1st Qu.: 90.0    1st Qu.: 85.0        1st Qu.: 60.00   1st Qu.: 347.0  
 Median :1332   Median :  -6.0             Median :110.0    Median :105.0        Median : 77.50   Median : 483.0  
 Mean   :1365   Mean   :   6.1             Mean   :127.1    Mean   :123.2        Mean   : 97.82   Mean   : 642.4  
 3rd Qu.:1825   3rd Qu.:   8.0             3rd Qu.:150.0    3rd Qu.:146.0        3rd Qu.:121.00   3rd Qu.: 869.0  
 Max.   :2400   Max.   :1928.0             Max.   :337.0    Max.   :378.0        Max.   :360.00   Max.   :2161.0  
 NA's   :119    NA's   :138                                 NA's   :138          NA's   :138                      
```

Problem Statement

Following are the problem statements for the study conducted on the Airline Data of Ohio State:

1. ***Statement 1***: At the time of departure, does the departing airport affect the flight delays?
2. ***Statement 2***: Does an Airline Carriers (Service Providers) impact the delays that occur in the state of Ohio at the time of take-off?
3. ***Statement 3***: Are Reservation System's Elapsed Time and Actual Elapsed Time related?
4. ***Statement 4***: Does Departure Delay affect Arrival Delay with a Positive relation or No relation?
5. ***Statement 5***: Is Taxi In time (Travel Time between Landing and Arrival Airport Gate) related to Taxi Out Time (Travel Time between Gate and Take Off)?

By analyzing the above problem statements, we can understand the Relations among the attributes and Factors that cause or affect delays for the flights that depart from various airports in Ohio State

Hypothesis Testing

For the above provided problem statements, following are the Null and Alternative Hypotheses.

Hypothesis 1

To study the if Origin Airport affects the Delay at the Time of Departure.

We use Analysis of Variance (ANOVA) to find whether the average delay of all the flights from the Origin Airports are same or different. For ANOVA, we use Departure Delay (DEP_DELAY) as our Quantitative Variable (Measure of Time) and Origin Airport (ORIGIN) as Categorical Variable to study if Mean of the DEP_DELAY is same, under each ORIGIN variable or not.

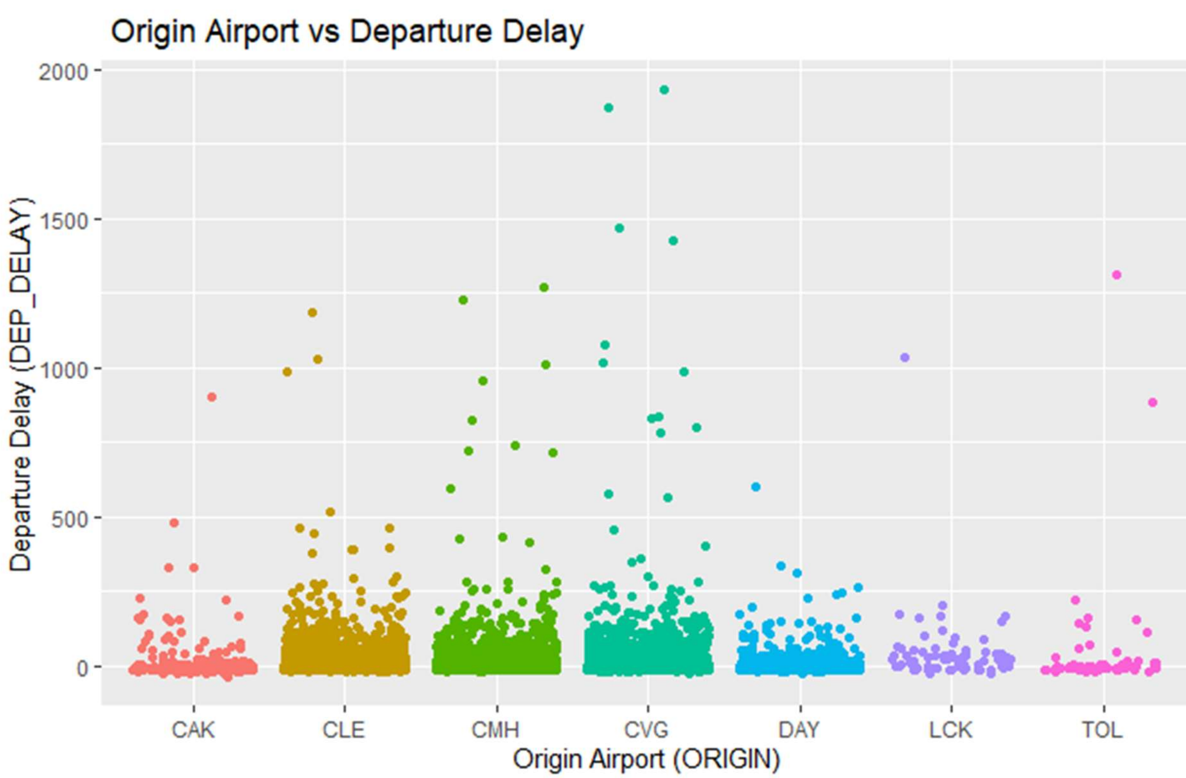
Following are the Null and Alternate Hypothesis for Hypothesis 1

Null Hypothesis (H$_0$): $\mu_{CAK} = \mu_{\mathrm{CLE}} = \mu_{\mathrm{CMH}} = \mu_{\mathrm{CVG}} = \mu_{\mathrm{DAY}} = \mu_{\mathrm{LCK}} = \mu_{\mathrm{TOL}}$ i.e., Average Flight Delays in all Airports of Ohio are not Different, where μ is Mean of DEP_DELAY

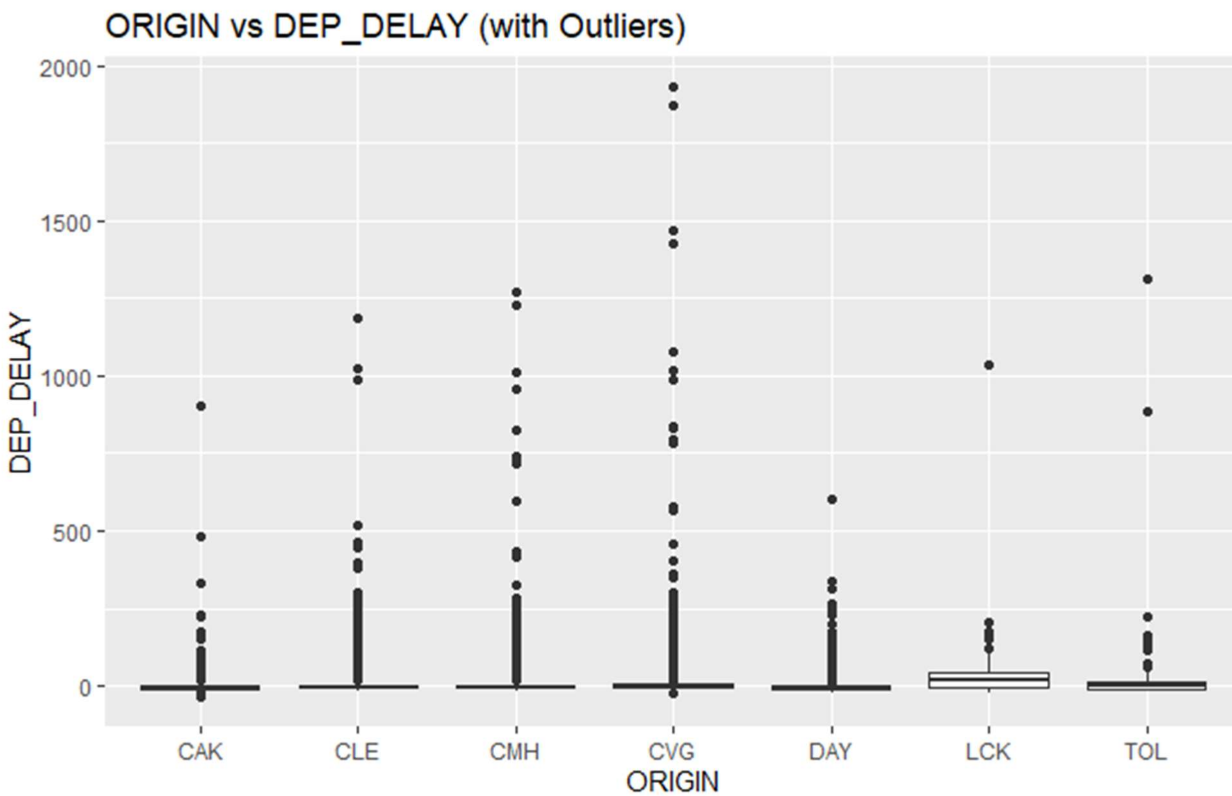
Alternate Hypothesis (***H$_A$***): At least 1 Mean is different i.e., $\mu_{CAK} \neq \mu_{\mathrm{CLE}}$ or $\mu_{CAK} \neq \mu_{\mathrm{TOL}}$ or so-on, 1 or more Means of the DEP_DELAY is different from each other under different ORIGIN.

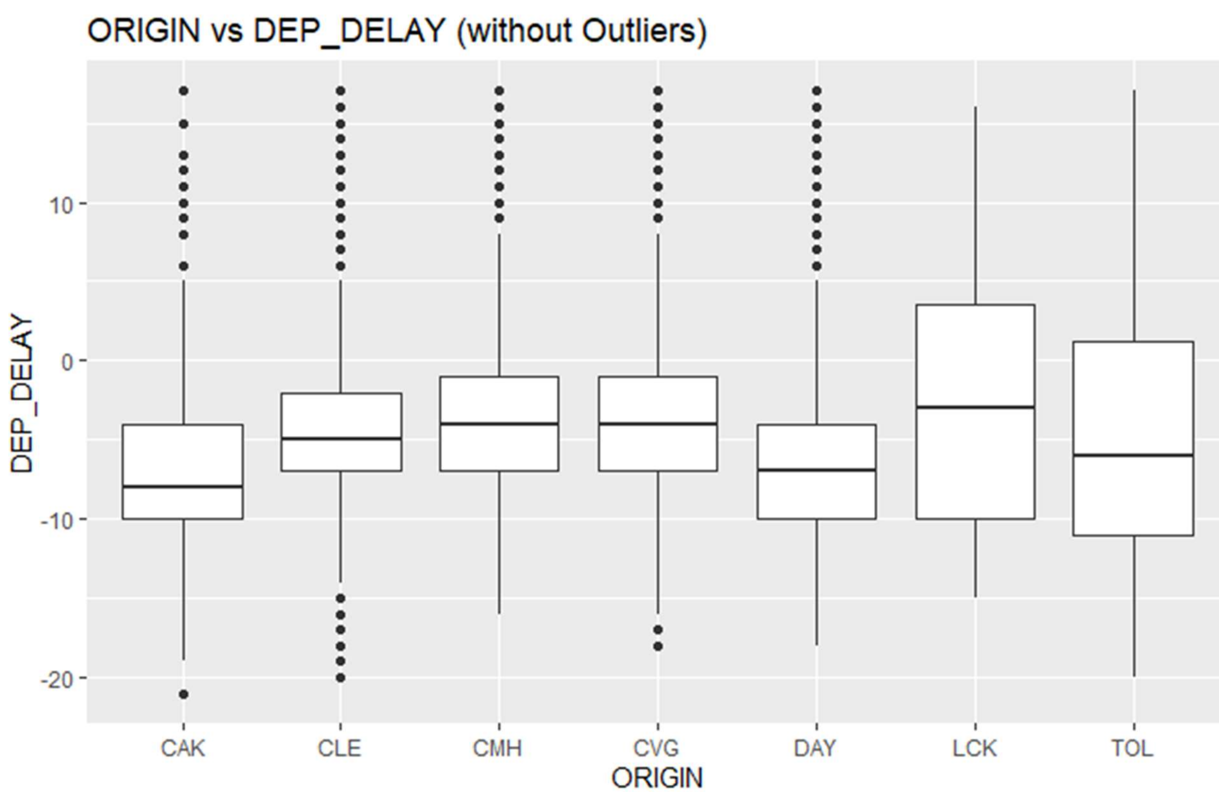
Procedure

Using R and RStudio, we select the columns that are needed for Hypothesis 1 (ORIGIN, ORIGIN_CITY_NAME, DEP_DELAY). Following Graph shows the general outline of Departure Delays based on Origin Airport



Null Values and Outliers are removed from the data as Outliers affect the ANOVA (Data Selection and Cleaning). The following Boxplots represent the data before and after cleaning.





After cleaning data, we perform ANOVA to determine whether our Null Hypothesis is accepted or not. Following is result obtained after performing ANOVA for clean data:

```
> #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
```

The above image shows the results obtained for the Analysis of Variance Test in R Programming using `aov( )` function and DEP_DELAY and ORIGIN as parameters.

Results

The above results show that the main effect of ORIGIN on DEP_DELAY is statistically significant and small ($F(6, 9952) = 33.36$, $p < .001$; Eta2 = 0.02, 95% Confidence Interval [0.01, 1.00]) generated using R Library – 'report'

As *p-value* is less than 0.001 with is less that the Significance Level (α) = 0.05. We reject H_0 and conclude that the Mean (DEP_DELAY) of at least 1 Origin Airport is different and Origin Airport affects the Departure Delay for the Flights that take off from various Airports from Ohio State.

Hypothesis 2

To study if there is an impact on Flight Delays at Departure Time by the Airline Carriers (Service Providers).

We use Analysis of Variance (ANOVA) to find whether the average delays of all the flights of the Airline Carriers are same or different. For ANOVA, we use Departure Delay (DEP_DELAY) as our Quantitative Variable (Measure of Time) and Airline Carrier (MKT_UNIQUE_CARRIER) as Categorical Variable to study if Mean of the DEP_DELAY is same, under all MKT_UNIQUE_CARRIER variables or not.

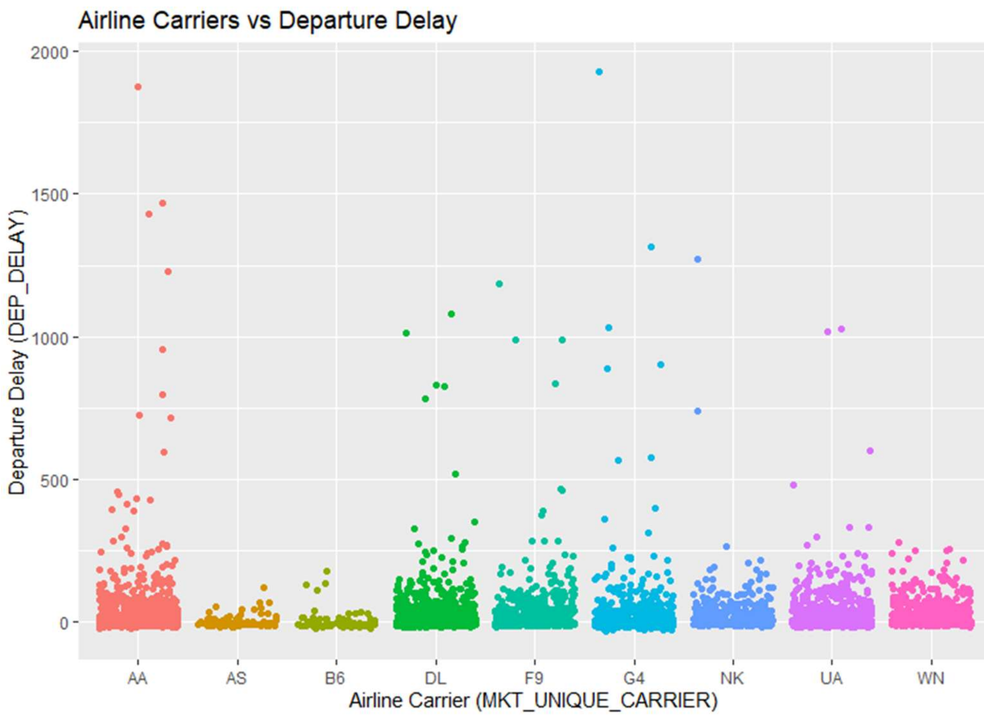
Following are the Null and Alternate Hypothesis for Hypothesis 2

Null Hypothesis (H_0): $\mu_{AA} = \mu_{\mathrm{AS}} = \mu_{\mathrm{B6}} = \mu_{\mathrm{DL}} = \mu_{\mathrm{F9}} = \mu_{\mathrm{G4}} = \mu_{\mathrm{NK}} = \mu_{\mathrm{UA}} = \mu_{\mathrm{WN}}$ i.e., Average Flight Delays for all Airline Carriers are Equal / Not Different, (where μ is Mean of DEP_DELAY)

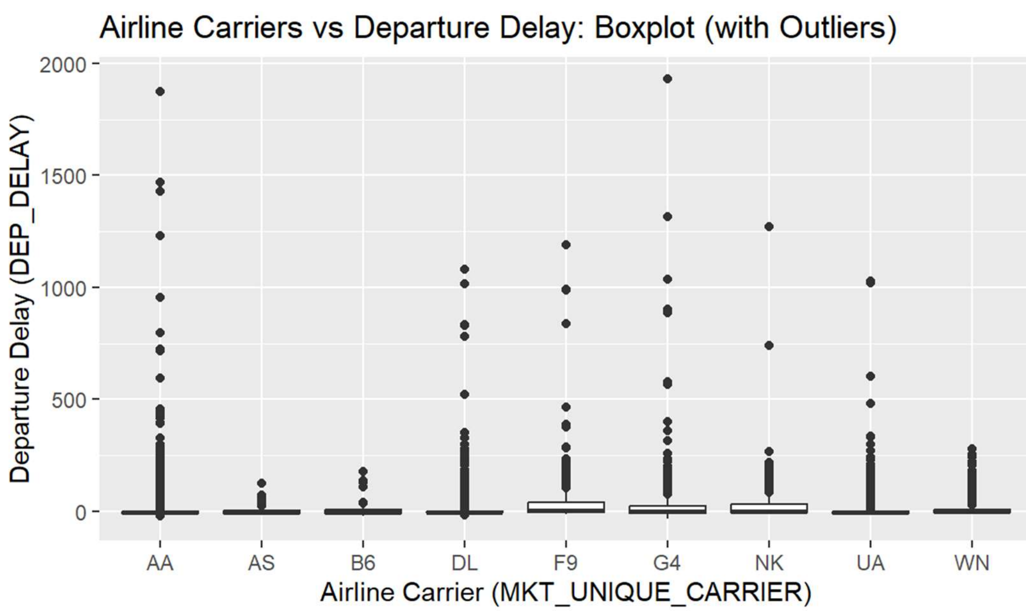
***Alternate Hypothesis* (H_A)**: At least 1 Mean is different i.e., $\mu_{AA} \neq \mu_{\mathrm{UA}}$ or $\mu_{WN} \neq \mu_{\mathrm{B6}}$ or so-on, 1 or more Means of the DEP_DELAY is different from each other for MKT_UNIQUE_CARRIER.

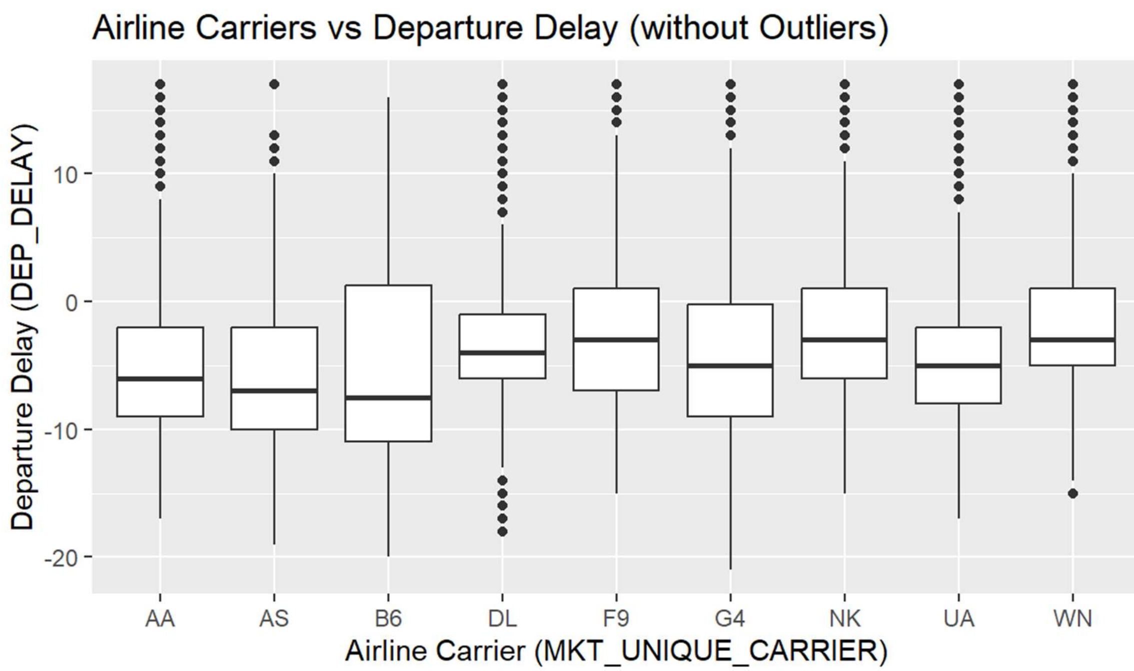
Procedure

Using R and RStudio, we select the columns that are needed for Hypothesis 2 (MKT_UNIQUE_CARRIER, MKT_CARRIER_FL_NUM, DEP_DELAY). Following Graph shows the general outline of Departure Delays for all Airline Carriers that Depart from Ohio State



We remove Null Values and Outliers (instead of replacing them as they cover only 9.1% of data) and use the cleaned data for Analysis of Variance. Following Graphs shows the comparison for data with and without outliers.





With the dataset obtained after cleaning the outliers and null values, Analysis of Variance is performed on it to decide whether to accept or reject Null Hypothesis.

```
> 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
```

Above image shows the results of Analysis of Variance performed in R Programming by passing DEP_DELAY and MKT_UNIQUE_CARRIER as the parameters for the function `aov( )`.

Results

The results from the above performed analysis confirms that the main effect of MKT_UNIQUE_CARRIER on DEP_DELAY is statistically significant and small ($F(8, 9950) = 55.90$, $p < .001$; Eta2 = 0.04, 95% Confidence Interval [0.04, 1.00]), generated using R Library – ‘report’.

The results drawn show that the *p-value* is less than 0.001 which is less than Level of Significance (α) = 0.05. We reject H_0 and conclude that the Mean (DEP_DELAY) of at least 1 Airline Carrier (Service Provider) is different and Airline Carriers affects the Departure Delay for the Flights that take off from various Airports from Ohio State.

Hypothesis 3

To verify the relation between Reservation System's Elapsed Time and Actual Elapsed Time for all the flights that travel from Ohio to other states.

We use Linear Regression to verify whether the variable ACTAUL_ELAPSED_TIME (Actual Elapsed Time) is dependent on the factor / variable CRS_ELAPSED_TIME (Computer Reservation System's Elapsed Time) and find the Correlation Coefficient for ACTUAL_ELAPSED_TIME and CRS_ELAPSED_TIME

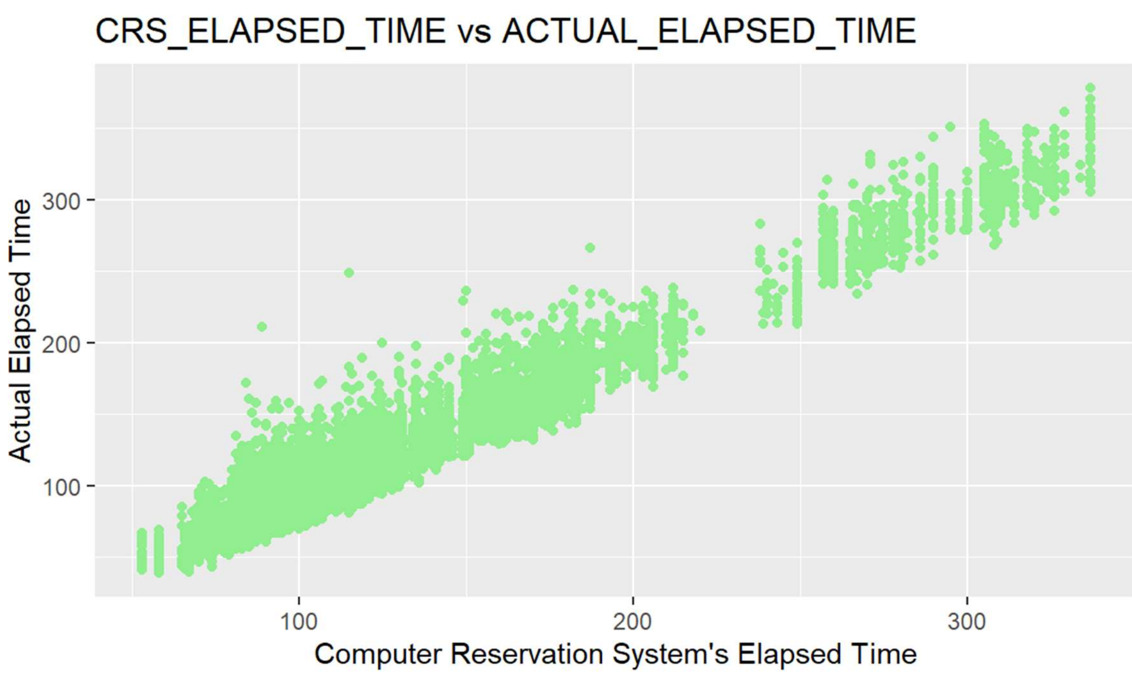
Following are the Null and Alternate Hypothesis for Hypothesis 3

Null Hypothesis (H₀): $\beta_1 = 0$, i.e., ACTUAL_ELAPSED_TIME and CRS_ELAPSED_TIME has no linear relation, (where AET = $\beta_0 + \beta_1$ CET + ε, β_0, β_1 – Coefficients, ε – Error Term)

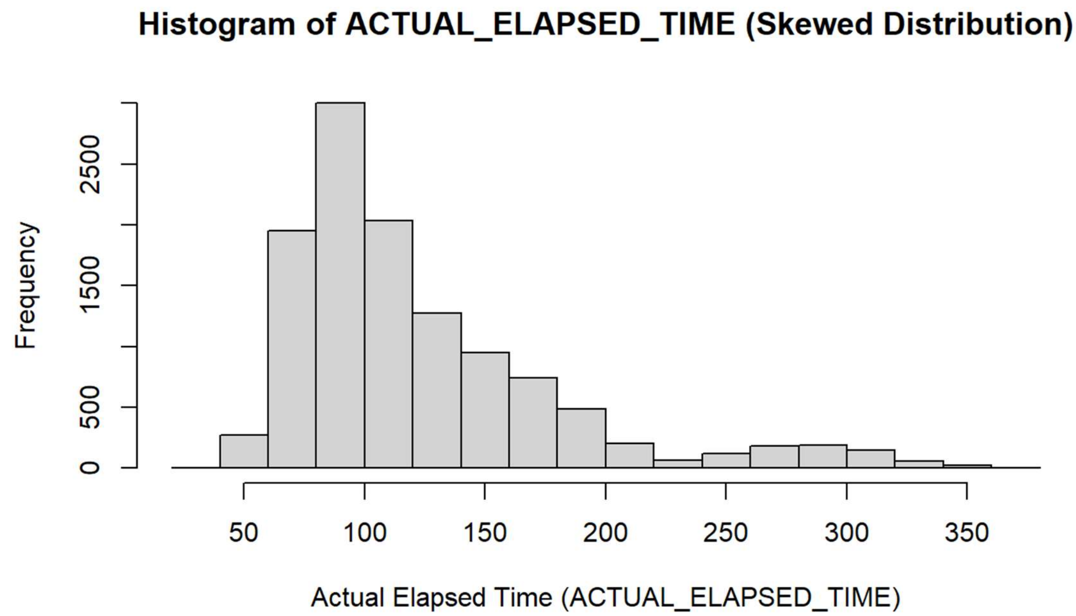
Alternate Hypothesis **(H_A)**: $\beta_1 \neq 0$ i.e., ACTUAL_ELAPSED_TIME depends on CRS_ELAPSED_TIME and both variables have either positive or negative linear relationship.

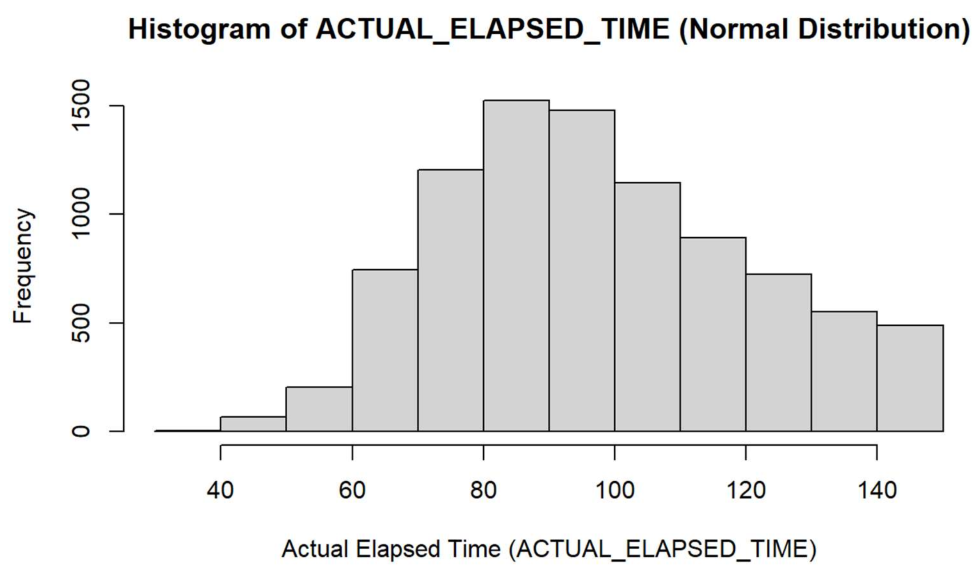
Procedure:

In R, we select the attributes CRS_ELAPSED_TIME and ACTUAL_ELAPSED_TIME that are explanatory and response variables respectively for Hypothesis 3. The following Graph gives the general idea of CRS_ELAPSED_TIME vs ACTUAL_ELAPSED_TIME.



If we plot a Histogram for the ACTUAL_ELAPSED_TIME (the Response variable), we can see that ACTUAL_ELAPSED_TIME follows Skewed-Graph. For performing Regression, we need to obtain normally distributed variable, so we remove the outliers and check if ACTUAL_ELAPSED_TIME follows Normal Distribution. The following histograms show ACTUAL_ELAPSED_TIME before and after removing Outliers.





After obtaining normally distributed data, Linear Regression is performed with `lm( )` by passing ACTUAL_ELAPSED_TIME and CRS_ELAPSED_TIME as the parameters as shown below to find the relation values by getting the Coefficient of Correlation for the parameters.

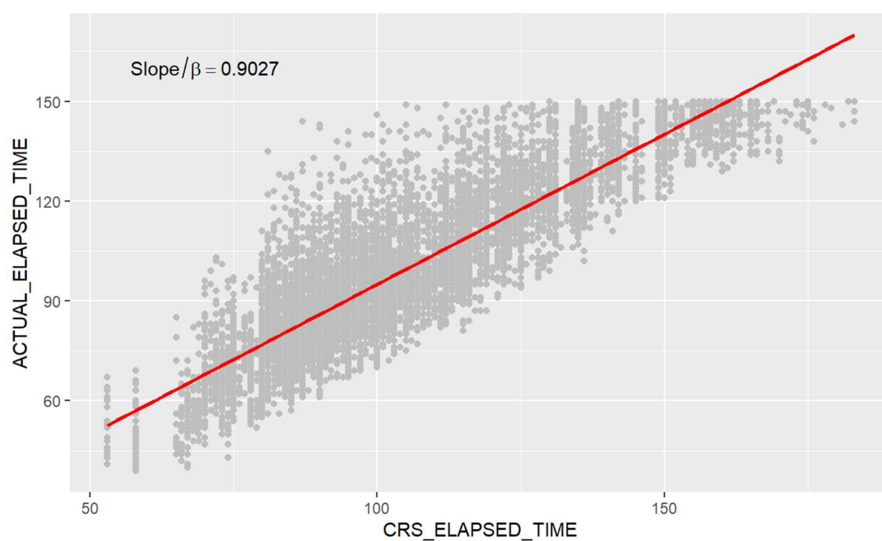
```
> result3_n <- lm(ACTUAL_ELAPSED_TIME ~ CRS_ELAPSED_TIME, data = h3t_n)
> summary(result3_n)

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
```

Results

Results from the above performed analysis says that the effect of CRS_ELAPSED_TIME on ACTUAL_ELAPSED_TIME is statistically significant and positive (beta = 0.90, 95% Confidence Interval [0.89, 0.91], t(9002) = 171.11, $p < .00$; Std. beta = 0.87, 95% Confidence Interval [0.86, 0.88]) – generated using R Library – 'report'

The value of Coefficient of Correlation $\beta_1 = 0.902$ which is not equal to zero ($\beta_1 \neq 0$). So, we reject the Null Hypothesis H_0 and conclude that ACTUAL_ELAPSED_TIME is dependent on the independent Variable CRS_ELAPSED_TIME. We can also say that with increase of each minute of CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME increase by 0.902 minute (54.1 Sec).

Following graphs shows the Linear Regression Model (Red Line / Curve) of CRS_ELAPSED_TIME vs ACTUAL_ELAPSED_TIME



Hypothesis 4

To study and find whether Departure Delay affects Arrival Delay positively or negatively for the flights that depart from Ohio State.

We use Linear Regression to verify whether the variable ARR_DELAY(Arrival Delay) is dependent on the factor / variable DEP_DELAY (Departure Delay) and find the is the Correlation Coefficient for ARR_DELAY and DEP_DELAY is positive, zero or negative.

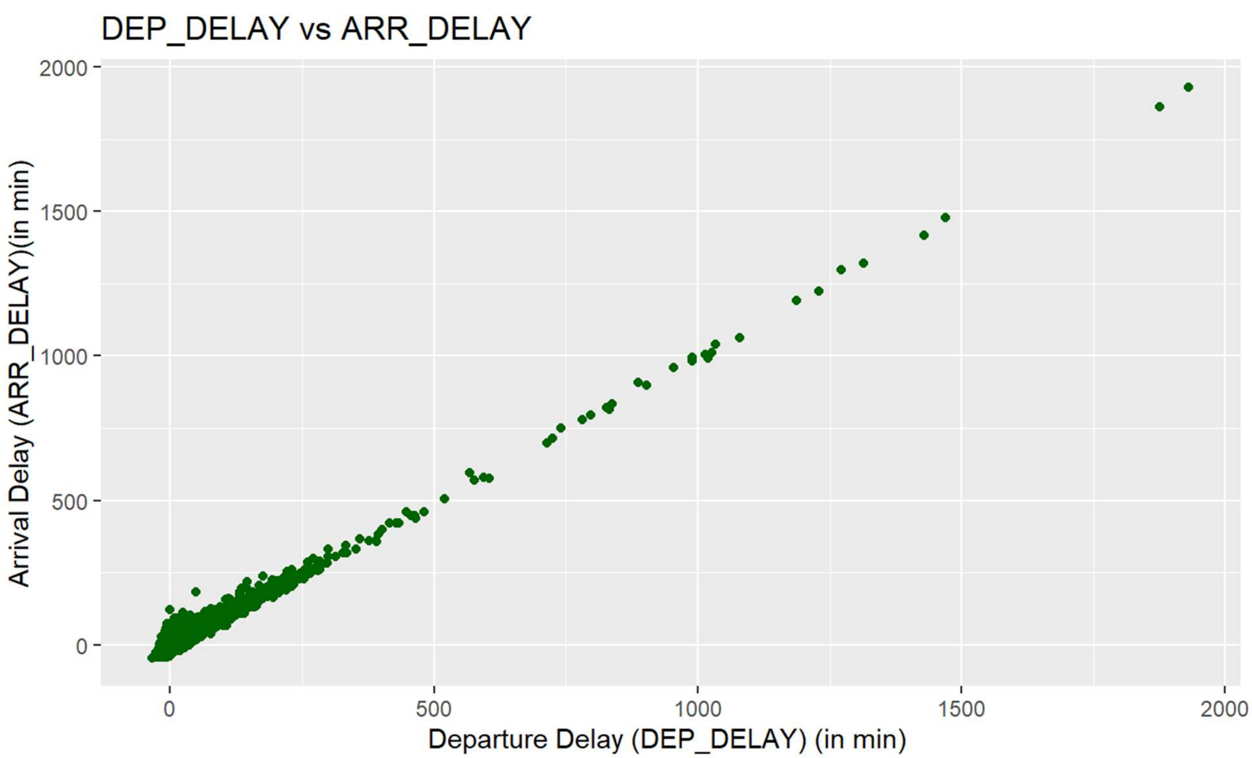
Following are the Null and Alternate Hypothesis for Hypothesis 3

***Null Hypothesis* (H_0)**: $\beta_1 \geq 0$, i.e., ARR_DELAY and DEP_DELAY has positive relation or no linear relation, (where ARR_DELAY = $\beta_0 + \beta_1$ DEP_DEALY + ε, β_0, β_1 – Coefficients, ε – Error Term)

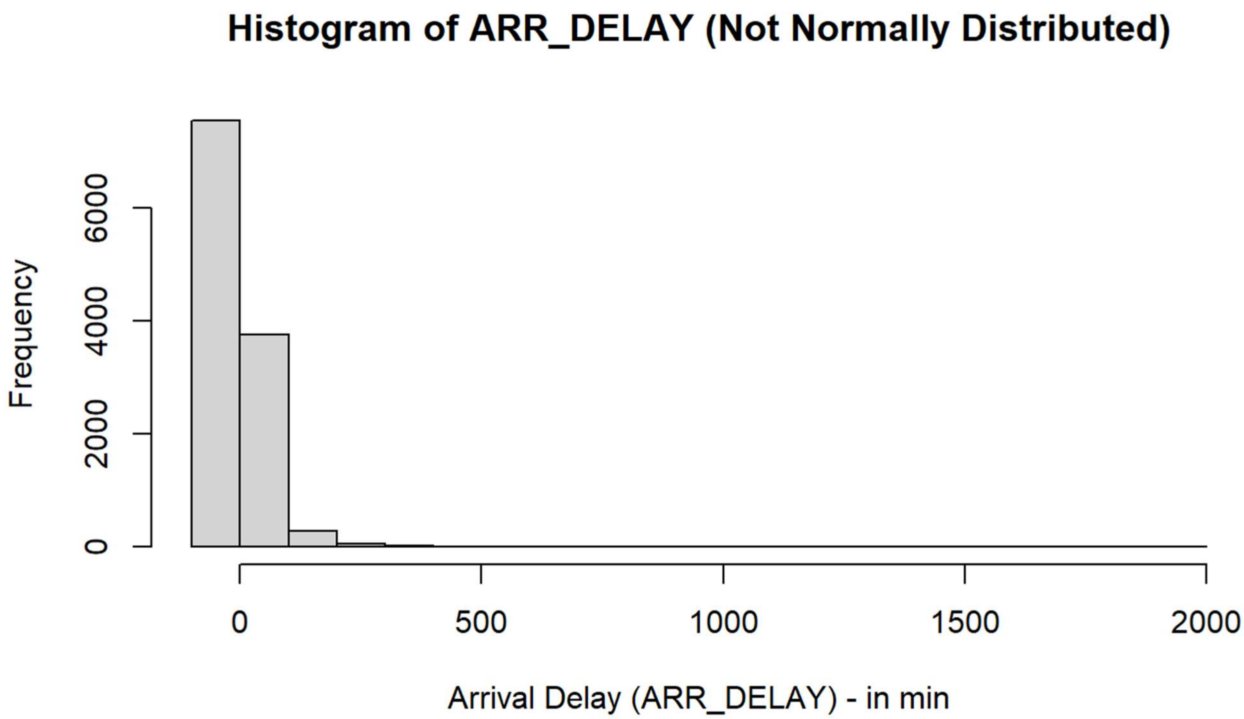
***Alternate Hypothesis* (H_A)**: $\beta_1 < 0$ i.e., ARR_DELAY and both variables have negative linear relationship.

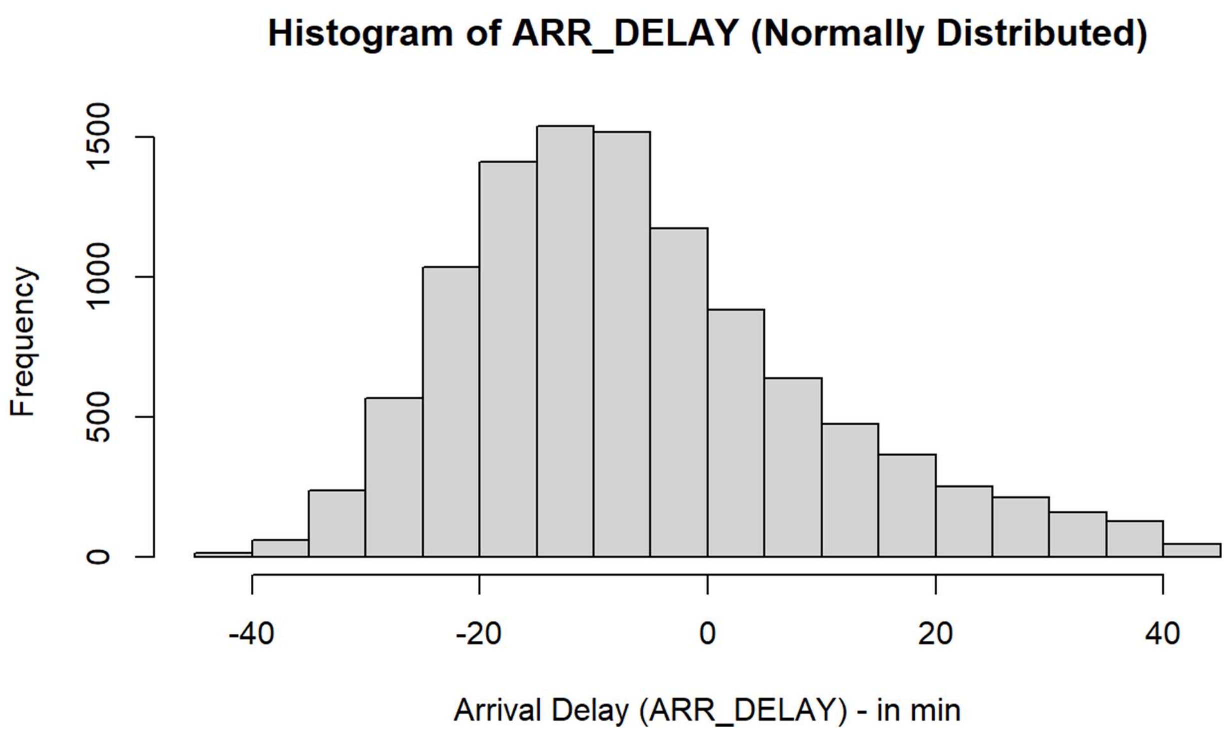
Procedure

Using R, we select the columns DEP_DELAY (exploratory variable) and ARR_DELAY (Response Variable) to find if the Correlation Coefficient of ARR_DEALY with respect to DEP_DELAY is greater than or equal to zero, or not for Hypothesis 4. The following graph gives overview of DEP_DELAY vs ARR_DELAY



When a histogram for the Response Variable is plotted, it is seen that ARR_DELAY doesn't follow normal distribution. So, we clean the data by removing the null values and outliers and see that the response variable is normally distributed for performing linear regression. The following histograms shows the structure of ARR_DELAY with and without Outliers:





After obtaining Normally Distributed data, we perform Linear Regression analysis with DEP_DELAY and ARR_DELAY using `lm( )` function in R Programming. Following is the result obtained after performing analysis:

```
> 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
```

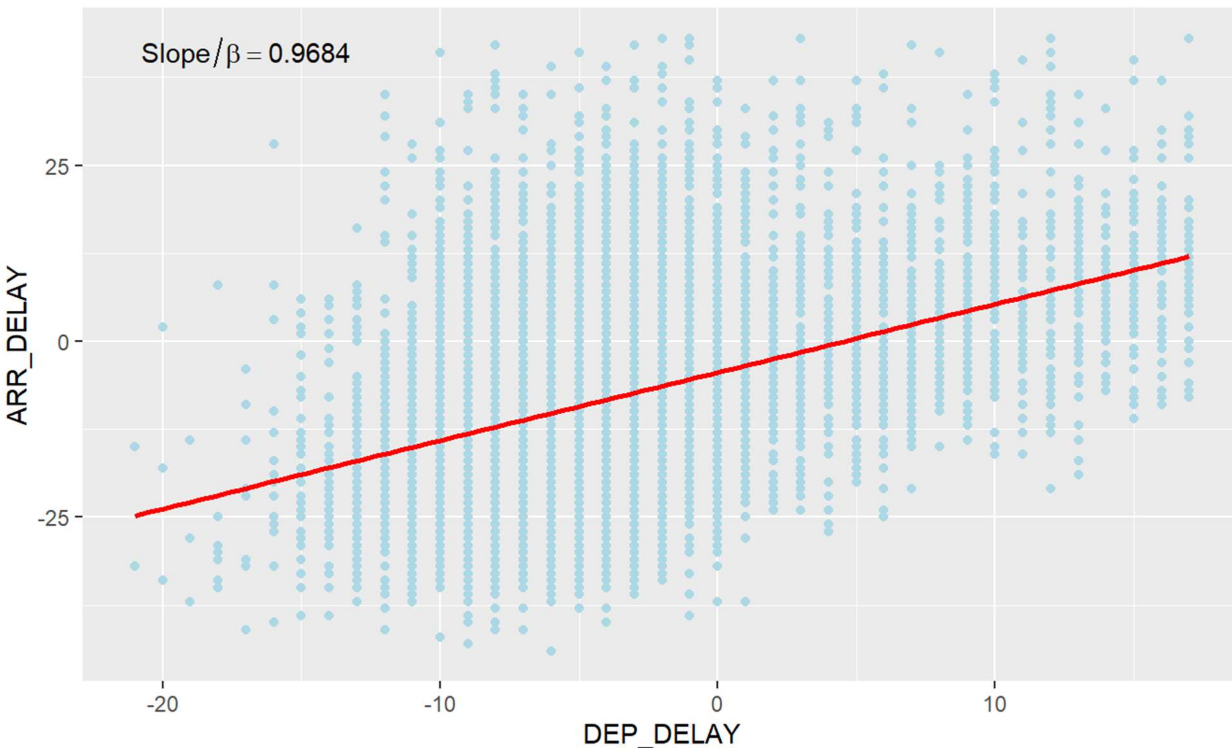
We get the above results by removing outliers from the exploratory variable DEP_DELAY

Results

The Results obtained from states that the effect of DEP DELAY on ARR_DELAY is statistically significant and positive (beta = 0.97, 95% Confidence Interval [0.93, 1.01], t(9890) = 46.90, $p < .001$; Std. beta = 0.43, 95% Confidence Interval [0.41, 0.44]), generated with R Library – ‘report’

The value of Coefficient of Correlation $\beta_1 = 0.968$ which is greater than or equal to zero ($\beta_1 \geq 0$). So, we do not reject the Null Hypothesis H_0 and conclude that ARR_DELAY is positively dependent on the independent variable DEP_DELAY. We can also say that with increase of each minute of CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME increase by 0.968 minute (58.1 Sec).

The following graph shows the positive slope for DEP_DELAY vs ARR_DELAY and it is generated by performing Linear Regression Model



Hypothesis 5

To study the relation between Taxi Out Time and Taxi In Time for the Flights that take off from Ohio State. (Taxi In time is the Time taken for the Flight to reach Arrival Airport Gate after it lands. Taxi Out time is the Time taken for the flight to travel from Gate at Departing airport till it takes off)

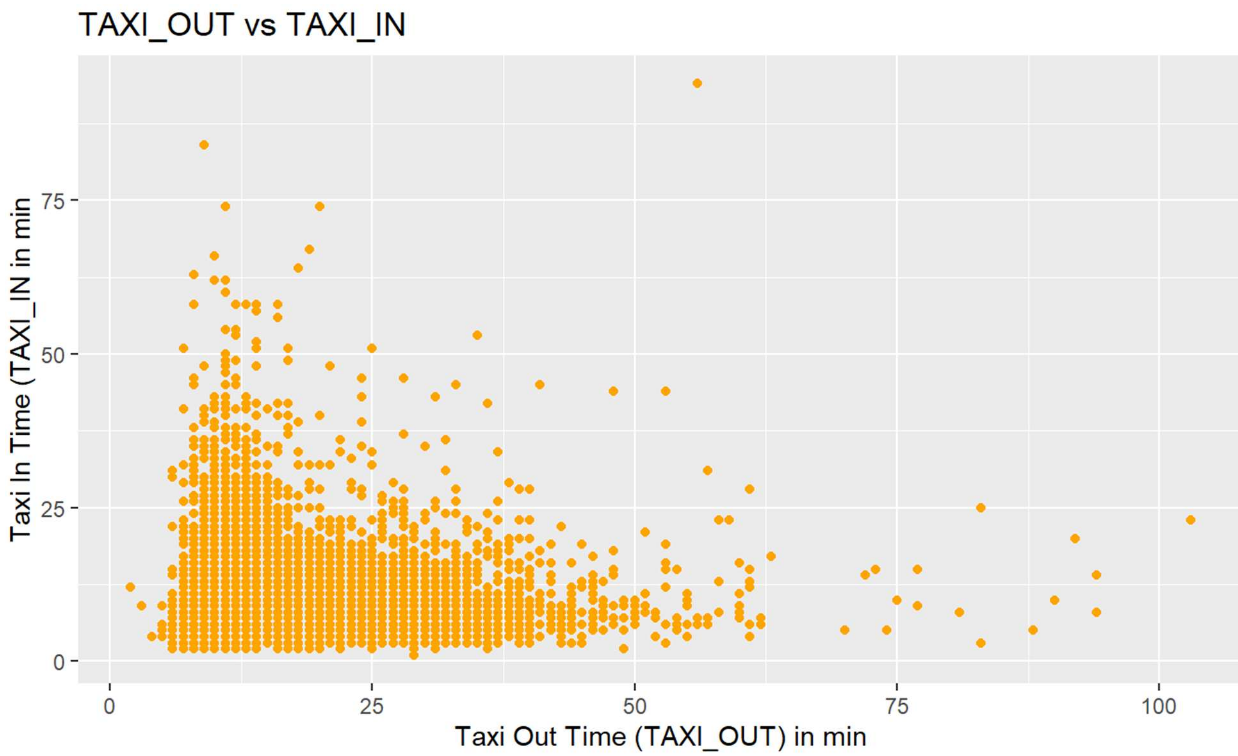
Using Linear Regression Model, we can find whether our response variable TAXI_IN has a linear relation with our exploratory variable TAXI_OUT. We perform a Two-tailed test to accept or reject our Null Hypothesis and Alternate Hypothesis which is given as follows:

Null Hypothesis (H_0): $\beta_1 = 0$, i.e., TAXI_IN and TAXI_OUT has no linear relation, (where TAXI_IN = $\beta_0 + \beta_1$ TAXI_IN + ε, β_0, β_1 – Coefficients, ε – Error Term)

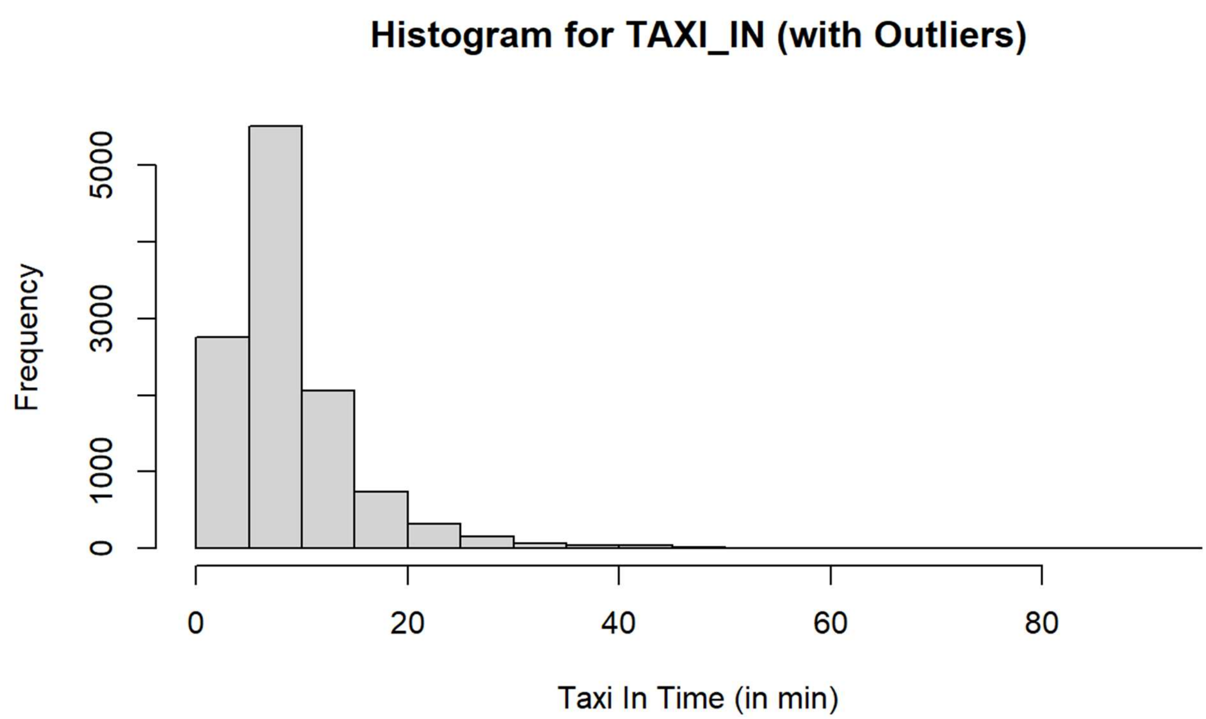
Alternate Hypothesis (***H_A***): $\beta_1 \neq 0$ i.e., TAXI_IN depends on TAXI_OUT and both variables have either positive or negative linear relationship.

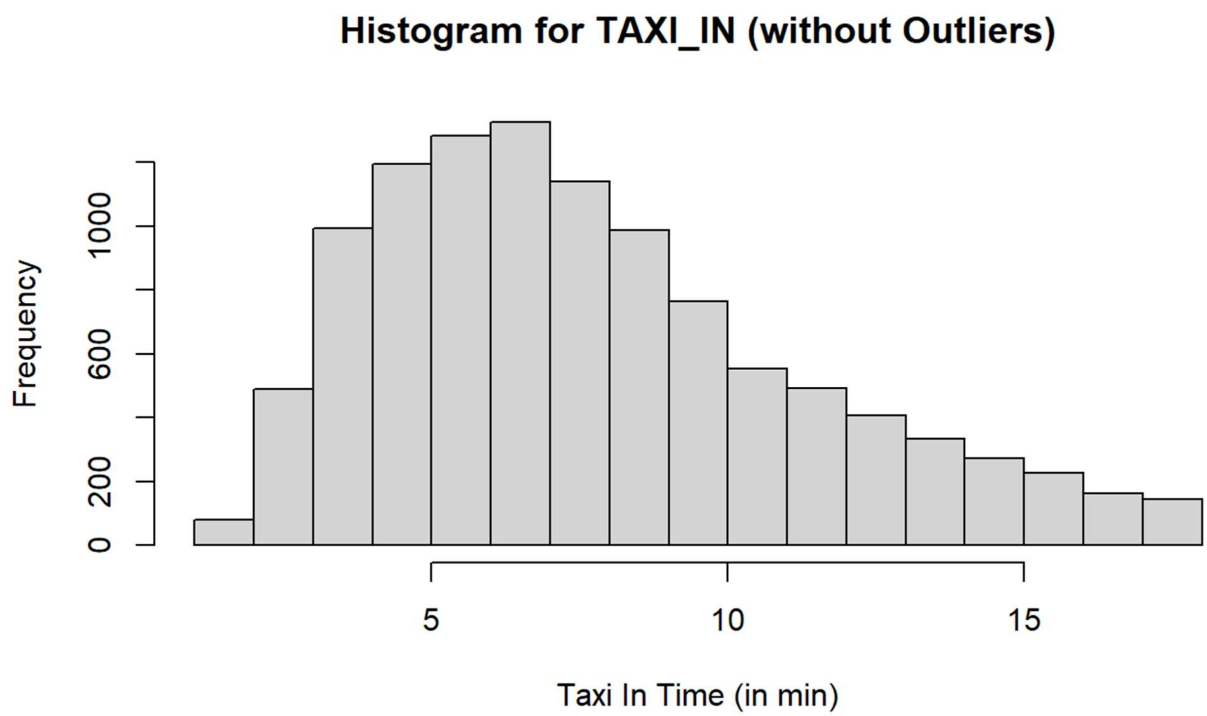
Procedure

We consider the variables TAXI_OUT (exploratory variable) and TAXI_IN (respondent variable) to find the type of linear relationship that exists between TAXI_IN and TAXI_OUT. Following graph is plotted TAXI_OUT against TAXI_IN and it gives the glimpse of the data:



As the TAXI_IN and TAXI_OUT don't follow normal distribution, we remove the outliers and null values and see that the variables follow normal distribution to perform linear regression analysis. The following histograms show the Distribution of Respondent variable TAXI_IN with and without Outliers:





After cleaning the data (removing Null Values and Outliers), we perform Linear Regression analysis with function `lm( )` and parameters TAXI_IN and TAXI_OUT in R Programming. Following image describes the Test Stats that we get after performing linear regression on TAXI_IN and TAXI_OUT:

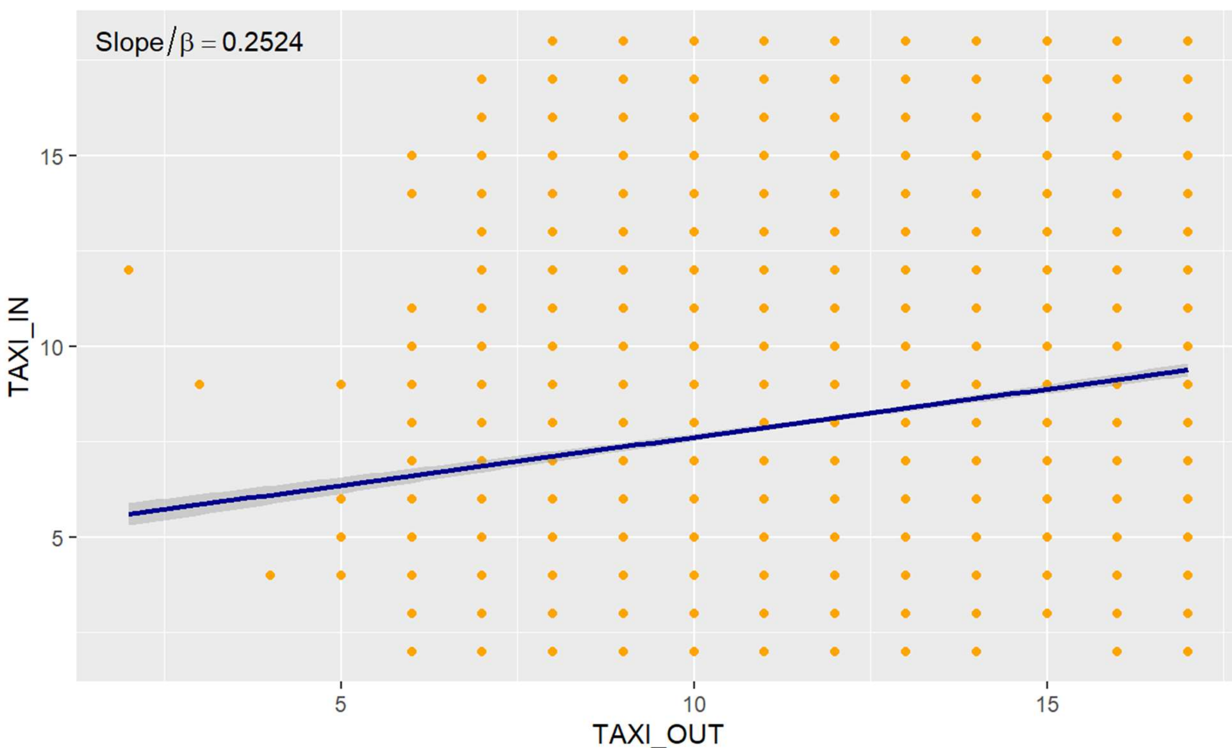
```
> result5_no2 <- lm(TAXI_IN ~ TAXI_OUT, data = h5t_no2)
> summary(result5_no2)

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
```

Results

From the results obtained through function `lm( )`, we can say that the effect of TAXI OUT on respondent variable TAXI_IN is statistically significant and positive (beta = 0.25, 95% Confidence Interval [0.22, 0.28], t(7949) = 16.82, $p < .001$; Std. beta = 0.19, 95% Confidence Interval [0.16, 0.21]), generated using the R Library - ‘report’

With the value of Coefficient of Correlation $\beta_1 = 0.252$, which means it is not equal to zero ($\beta_1 \neq 0$), we can reject the Null Hypothesis H_0 and conclude that respondent variable TAXI_IN and exploratory variable TAXI_OUT have positive linear relationship. We can also say that with increase of each minute of CRS_ELAPSED_TIME, ACTUAL_ELAPSED_TIME increase by 0.252 minute (15.1 Sec).

The following graph shows the linear model that is generated with TAXI_OUT vs TAXI_IN:



Conclusion

From the above drawn results, we can conclude that the Flight Delays depends on Origin Airport (Airport from with Flight departs) and Airline Carrier (Service Provider) for the flights that take off from the Ohio State. Along with this, we found that the respondent variables ARR_DELAY, ACTUAL_ELAPSED_TIME and TAXI_IN are in a positive linear relationship with the exploratory variables DEP_DEALY, CRS_ELAPSED_TIME and TAXI_OUT respectively.

Also, if we focus on the fifth hypothesis, TAXI_IN and TAXI_OUT’s correlation coefficient is 0.25 whereas other two had 0.9, thus confirming that the relation between TAXI_IN and TAXI_OUT turning out to be weak. If we look at description of the attributes TAXI_OUT and

TAXI_IN, we can see that they are the time taken to fly off the ground from the Origin Gate and reach the Destination Gate after landing. If we consider this, we shouldn't be having any kind of relation between either as they depend on Departing Airport and Arrival Airport Runway size. Yet, instead of obtaining $\beta_1 = 0$, we got 0.25 stating that these two attributes are in linear relation. So, we can say that either there are other factors that affect these attributes, or this can be a Type-I Error where we are rejecting the Null Hypothesis even though it is correct.

Recommendations

We could improve the method of obtaining results by properly handling Null Values and Outliers. In this case, we removed the Null values and Outliers (10% of the data). Instead, we could say that these Outliers are present not because of incorrect entry but they are valid data, and they should be considered. So, instead of removing them, if they were handled properly, maybe we could have got a better understanding of the data and obtain better results that approve or reject the hypotheses.