REASON BEHIND NO. OF DELAYS OF AIRFLIGHT

Detecting reason for delays of airflight in all regions of world.

WANG YI MING
200548371

Contents

Introduction:

Delay is a central issue in air transportation action. As a presentation metric, it influences normal strategy concerns. Postpone impacts traveler fulfillment and forces costs. A new exact review by Mayer and Sinai (2003a) observes that air gridlock because of aircraft hubbing and over-planning of trips at air terminal offices are the essential drivers of flight delays.

The following aims and objected be achieved by this research project.

- Best time of day, day of week and time of year for minimum delays.
- Nostalgic planes are cause of delays.
- No. of people flying for different location is changing over time.
- Are delay in first airport reason behind delay in other airport as well?
- Model for predicting no. of delays.

Literature Review:

Allan et al. concentrated on a few deciding reasons for flight delay at the Newark International Airport (EWR) utilizing a thorough methodology. The outcomes show that antagonistic climate conditions, low roofs, and low perceivability conditions emphatically impact flight delays. Likewise, Asfe et al. examined the major causal variables of flight delays by positioning various elements utilizing the insightful progressive interaction. They tracked down specialized disappointment and postponed passages as two of the most persuasive variables In view of the ID of causal elements, further investigates investigated the quantitative impact of each component on flight delay. By dissecting the attributes of flight takeoff and appearance delays by developing likelihood thickness capacities, Mueller et al. investigated a few causal variables of postponements, for example, traffic volume, airplane type, airplane support, carrier tasks, climate conditions, change of methods on the way, limit requirements, client assistance issues, and late airplane or group appearance. The outcomes show that climate added to 69% of the postponements. Various outcomes can be accomplished by various strategy and factors; research aftereffects of Kwan and Hansen show that air terminal blockage added to roughly 32% of the normal deferrals, where a progression of econometric models was set up to recognize the vital causal variables of flight delays, including air terminal clog, absolute traffic, and on the way climate. As well as recognizing the causal elements and their quantitative impact on flight delay, more examinations center around the advancement of models to decide the likelihood of airplane delay. Wesonga et al. proposed and assessed a different parametric methodology, which incorporates the clearly critical meteorological and avionics boundaries, to foresee the likelihood of airplane delay.

Methodology:

To perform the research a secondary data is obtained from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7 for comparison of two years 2007 and 2008. It is mixed qualitative and quantitative data obtained for large number of respondents. To compare the performance of two years we have to merge data set. So, data is merged for study variables. The data is obtained for 23 different variables among which 15 of study variables are selected. So, for performing the analysis firstly we have to clean the data for missing values and replacing them by zero. Next is to examine for minimum no. of delays using minimum function in R and detect the corresponding days, year and month. Further for detecting older planes as cause of delay we would examine the average delays for different periods of time using t test. Further, change in no. of people flying is determined by time series plot for different regions of the world.

Later, delay in flight is detected by delay in flight on another airport is detected by correlation among them. At the end we have to propose a model that detect the reason behind delays by observing significant variables using linear regression model.

Data Analysis:

To achieve the required results we have firstly to read the data and examine for missing values which are than replaced by zero in R.

```
> delay=read.table(file.choose(),sep = ",",header = T)
> delay[1,]
  Year Month DayofMonth DayOfWeek DepTime ArrTime Distance WeatherDelay XOSDelay
1 2008     1          3         4    1343    1451      393            0        0
  SecurityDelay LateAircraftDelay
1             0                 0
> delay.df=as.data.frame(delay)
```

Now, we have to examine the month of year for which delays are minimum. Thus, data is examined for minimum delays and corresponding month is noted.

```
> delays=filter(delay.df, delay.df$Month == min(delay.df$Month))
> delays$Month
   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [81] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [161] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [201] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [241] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [281] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [321] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [361] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [401] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [441] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [481] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [521] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [561] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [601] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [641] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [681] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [721] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [761] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [801] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [841] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [881] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [921] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [961] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

The maximum no. of delays appeared in month of January.

Next is the detection of day of month for which minimum number of delays of flight appeared. So, similar technique would be used for examining the month having minimum delays.

```
> delays$DayofMonth
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [41] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [81] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[121] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[161] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[201] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[241] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[281] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[321] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[361] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[401] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[441] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[481] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[521] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[561] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[601] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[641] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[681] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[721] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

The minimum no. of delays appeared at third day of every month. Now, we have to determine the day of week for which maximum number of delays appeared. Later we have to discuss the day of week for minimum delays.

```
> delays$DayOfWeek
  [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 [41] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 [81] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[121] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[161] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[201] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[241] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[281] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[321] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[361] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[401] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

The above output indicates minimum no. of delays appeared on fourth day of week such as Thursday.

Thus, from these three outputs we can suggest that minimum no. of delays appeared on 3rd January and on Thursdays for this particular data. So, if one tries to minimize no. of delays than he has to travel on Thursdays. Next is the effect of nostalgic plane on delay. So, average no. of delays would be examined for both periods of time. Thus, data is divided for both periods.

Null and Alternative hypothesis:

Ho: The average delays from older planes is less than or equal to that of new planes.

H1: The average delays from older planes is more than that of new planes.

Level of Significance:

Alpha=0.05

Test Statistics:

$$t = \frac{\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Critical Region:

We would reject null hypothesis if p value is less than level of significance.

Calculations:

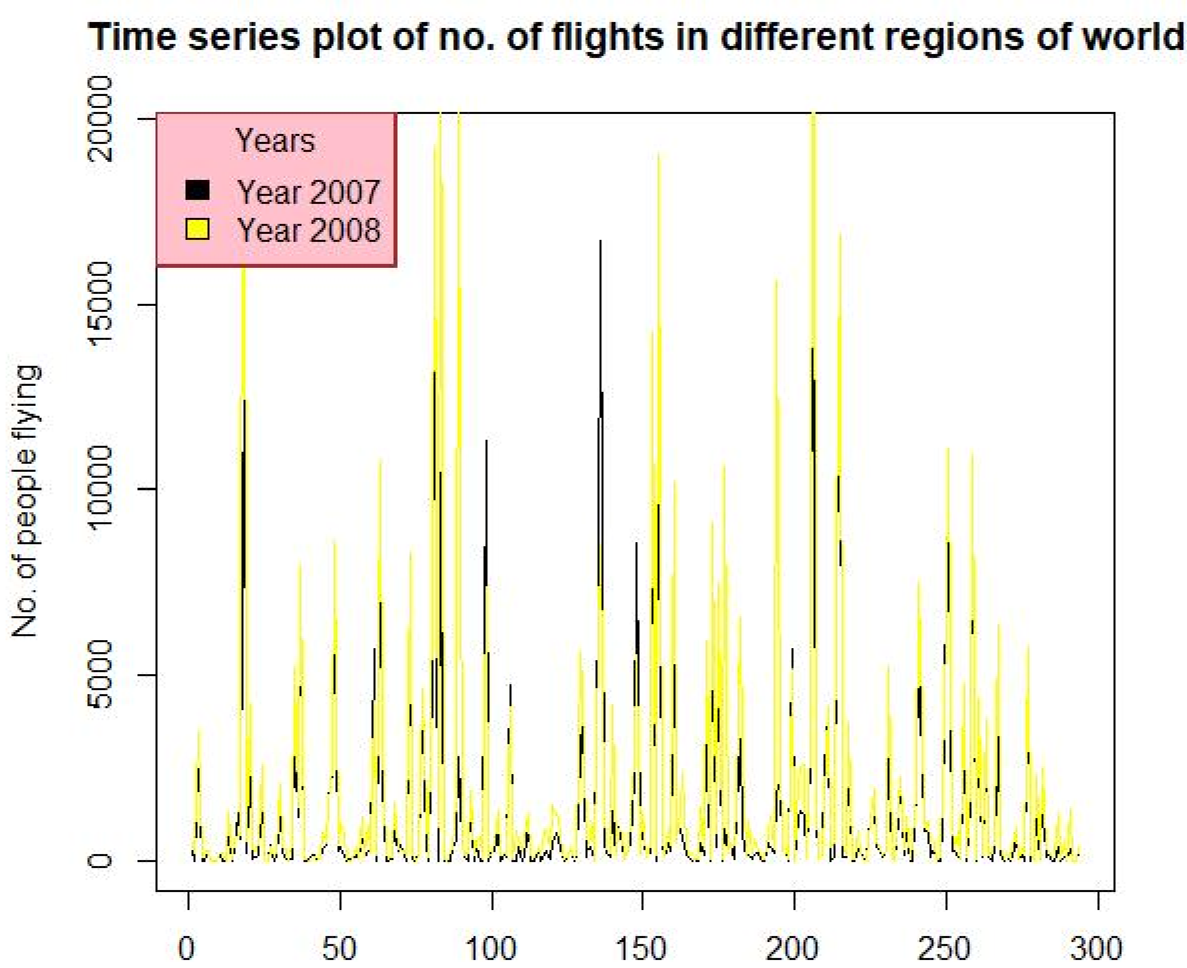
```
> no.of.delay=rowSums(delay.df[,8:11],na.rm = T)
> new.data=cbind(delay.df,no.of.delay)
> older=new.data[592929:1002463,]
> newer=new.data[1:592928,]
> new.data[1,]
  Year Month DayofMonth DayOfWeek DepTime ArrTime Distance WeatherDelay XOSDelay
1 2008     1          3         4    1343    1451      393            0        0
  SecurityDelay LateAircraftDelay no.of.delay
1             0                 0           0
```

```
> t.test(older$no.of.delay,newer$no.of.delay,alternative = "greater")

        Welch Two Sample t-test

data:  older$no.of.delay and newer$no.of.delay
t = 4.2077, df = 875623, p-value = 1.29e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.158341      Inf
sample estimates:
mean of x mean of y 
 10.30998  10.05002 
```

Conclusion:

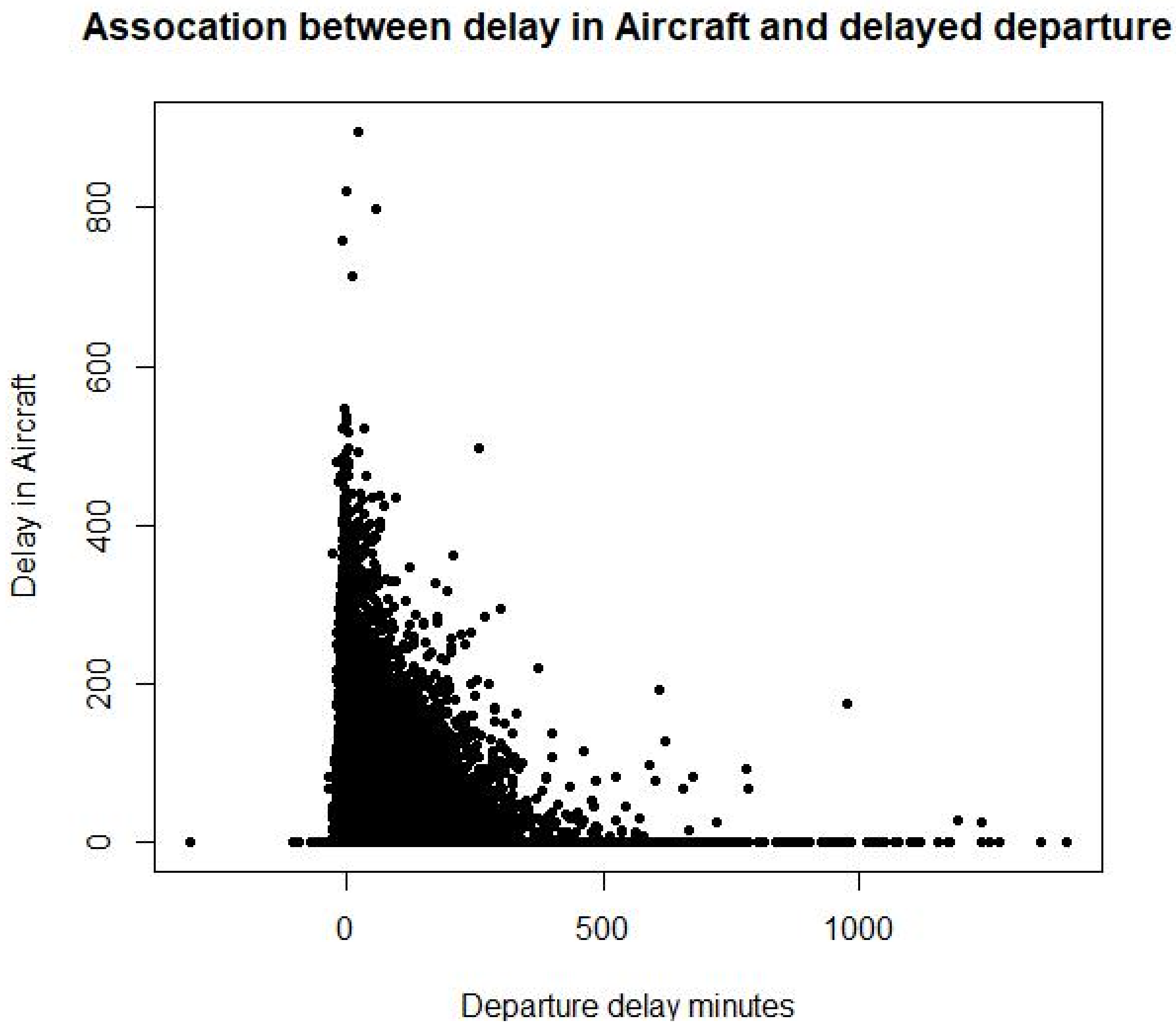
For examining reason behind no. of delays we have to observe either delays are because of older planes or not. So, average number of delays for two periods would be tested using t test. P value for test is very small which indicates enough evidence for possible rejection of null hypothesis concluding the no. of delays for older planes would be more than those of newer planes. So, nostalgic planes are more harmful than new ones in terms of delays.

Next we have to determine whether the no. of people travelling for different regions due to delays is decreased by time or not. So, time series plot is used for two different time periods.



This indicated that by time no. of persons flying in regions of world changed. As, yellow lines in above graph are larger than black lines which is for previous year. So, we can say that no. of people flying by these delays is not decreased by time.

Next we have to determine whether delays in one airport effects the delays of flights on other airport as well or not. So, association among both variables would be best choice. Here, delays in minutes is quantitative variable and delay in aircraft for another airport is also quantitative variable. So, for both of quantitative variables correlation is discussed by scatter plot which would predict stronger association if values follow any pattern and they are closer to each other. Thus, scatter plot among both variables is shown below;



The scatter plot indicated that values follow a pattern and they are closely related with each other which predict high relationship among them. The scatter plot predicts negative relationship among them as increasing delay in departure time of one airport decreases delay in aircrafts at another airport so, we can't say that delays in airports are associated or in other words by controlling no. of delays of one airport might not affect the delays of other airport as well.

Later we have to detect the reason behind no. of delays in all regions of world for two time periods. As, departure delay in minutes is quantitative variable so, linear regression would be best choice which is detected by weather delay, security delay and aircraft delay.

```
> model=lm(DepDel0y~Distance+WeatherDelay+XOSDelay+SecurityDelay+LateAircraftDelay ,data
 = delay)
> summary(model)

Call:
lm(formula = DepDel0y ~ Distance + WeatherDelay + XOSDelay +
    SecurityDelay + LateAircraftDelay, data = delay)

Residuals:
    Min      1Q  Median      3Q     Max
-317.09  -16.23  -12.37   -0.59 1393.78

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.197e+01  6.123e-02 195.447  < 2e-16 ***
Distance           4.765e-04  6.822e-05   6.985 2.86e-12 ***
WeatherDelay       1.478e-04  3.455e-03   0.043 0.965870
XOSDelay           8.012e-03  2.286e-03   3.505 0.000456 ***
SecurityDelay     -2.242e-03  1.683e-02  -0.133 0.894003
LateAircraftDelay  7.947e-03  1.693e-03   4.694 2.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.14 on 1002454 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  8.53e-05,  Adjusted R-squared:  8.031e-05
F-statistic:  17.1 on 5 and 1002454 DF,  p-value: < 2.2e-16
```

The regression model indicates distance and aircraft delay are significant reason behind no. of delays. The R square of model is small which indicates that smaller variations in delays are explained by these variables or in other words delay is dependent upon some other variables as well.

Summary and Conclusion:

The report is conducted to examine the reason behind number of delays. So, data is obtained from secondary source for large number of respondents thus to reduce the effect of bias produced by small samples. The t test for average no. of delays produced by older planes indicates more delays are caused by nostalgic machinery. Further, no. of people travelling by planes after delays is not reduced by time. Also, no. of delays is best predicted by distance travelled, NOS delay and Late air craft delay.