**Big-data-analysis-on-flight departure delay and the speed of the Airplane to cover the distance-using-Revolutionary-R-open**

The arrival and departure time of the flight can change according to the delay, weather, any technical difficulties and many other factors. In this project we try to analyze at what speed the airplane travels if the flight’s departure time is delayed for few minutes. NOTE-The weather can also change the arrival time of an airplane. The below summary gives detailed analysis on the flight dataset with functions followed with the output for each function

rxOptions(reportProgress = 0)

File name: C:\Users\Chinmay\Desktop\KaggleProjects\GroupoBimbo\2007.xdf

Number of observations: 7453215

Number of variables: 29

Number of blocks: 15

Compression type: zlib

Variable information:

Var 1: ArrDelay, Type: integer, Low/High: (-312, 2598)

Var 2: DepDelay, Type: integer, Low/High: (-305, 2601)

Var 3: Distance, Type: integer, Low/High: (11, 4962)

> rxSummary(~DepDelay + Distance,data = flightXdf)

Call:

rxSummary(formula = ~DepDelay + Distance, data = flightXdf)

Summary Statistics Results for: ~DepDelay + Distance

Data: flightXdf (RxXdfData Data Source)

File name: C:\Users\Chinmay\Desktop\KaggleProjects\GroupoBimbo/2007.xdf

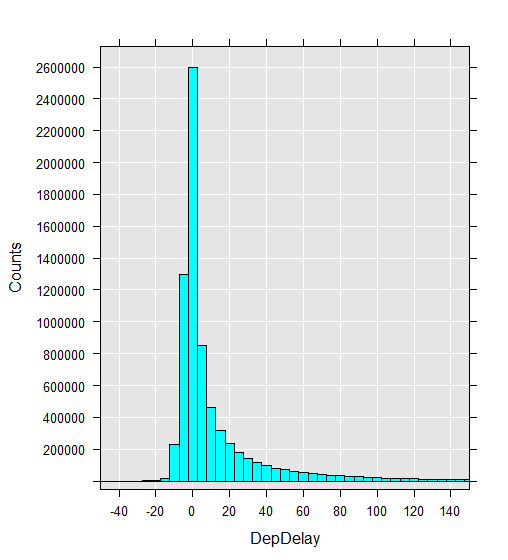
Number of valid observations: 7453215

Name Mean StdDev Min Max ValidObs MissingObs

DepDelay 11.39914 36.14189 -305 2601 7292467 160748

Distance 719.80579 562.30512 11 4962 7453215 0

> rxHistogram(~DepDelay,data = flightXdf,xAxisMinMax = c(-50,150),numBreaks = 500,xNumTicks = 10)



New XDS file to calculate speed and used only AirTime, Distance,DepDelay,ArrDelay,airspeed

> rxGetInfo(data = NewflightXdf,getVarInfo = TRUE)

File name: C:\Users\Chinmay\Desktop\KaggleProjects\GroupoBimbo\2007\_airSpeed.xdf

Number of observations: 7453215

Number of variables: 5

Number of blocks: 15

Compression type: zlib

Variable information:

Var 1: AirTime, Type: integer, Low/High: (0, 1257)

Var 2: Distance, Type: integer, Low/High: (11, 4962)

Var 3: DepDelay, Type: integer, Low/High: (-305, 2601)

Var 4: ArrDelay, Type: integer, Low/High: (-312, 2598)

Var 5: airSpeed, Type: numeric, Low/High: (0.2053, 1074.0000)

> rxSummary(~airSpeed,data = NewflightXdf)

Call:

rxSummary(formula = ~airSpeed, data = NewflightXdf)

Summary Statistics Results for: ~airSpeed

Data: NewflightXdf (RxXdfData Data Source)

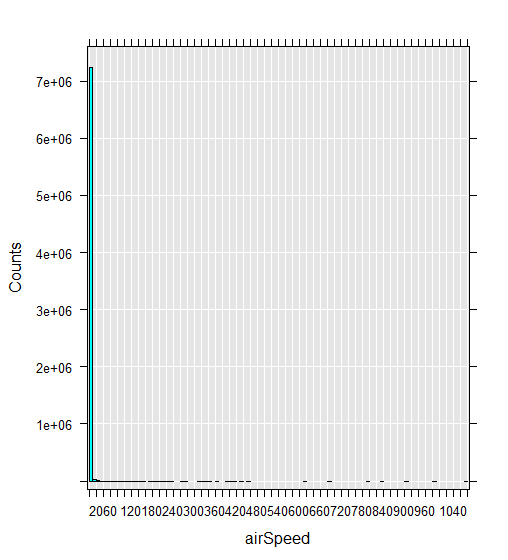
File name: C:\Users\Chinmay\Desktop\KaggleProjects\GroupoBimbo/2007\_airSpeed.xdf

Number of valid observations: 7453215

Name Mean StdDev Min Max ValidObs MissingObs

airSpeed 6.559545 1.986091 0.2052506 1074 7275191 178024

> rxHistogram(~airSpeed,data = NewflightXdf)



> rxSummary(~airSpeed,data = NewflightXdf)

Call:

rxSummary(formula = ~airSpeed, data = NewflightXdf)

Summary Statistics Results for: ~airSpeed

Data: NewflightXdf (RxXdfData Data Source)

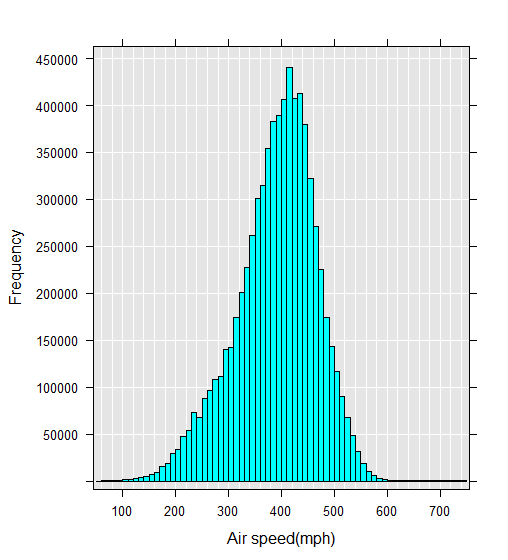
File name: C:\Users\Chinmay\Desktop\KaggleProjects\GroupoBimbo/2007\_airSpeed.xdf

Number of valid observations: 7453215

Name Mean StdDev Min Max ValidObs MissingObs

airSpeed 393.5727 119.1655 12.31504 64440 7275191 178024

> rxHistogram(~airSpeed,data = NewflightXdf,rowSelection = (airSpeed>50)&(airSpeed<750),scales = list(x=list(at=seq(100,700,by=100))),xlab=list(label="Air speed(mph)"),ylab=list(label="Frequency"),numBreaks = 5000,xNumTicks = 40)



> rxCor(formula = ~DepDelay + ArrDelay + airSpeed,data = NewflightXdf,

+ rowSelection = (airSpeed >50) & (airSpeed <800))

DepDelay ArrDelay airSpeed

DepDelay 1.00000000 0.93192674 0.02083635

ArrDelay 0.93192674 1.00000000 -0.06705657

airSpeed 0.02083635 -0.06705657 1.00000000

Correlation matrix

reg1 <-rxLinMod(formula = airSpeed ~DepDelay,data = NewflightXdf,rowSelection = (airSpeed>50)&(airSpeed <800))

> names(reg1)

[1] "coefficients" "residual.squares" "condition.number" "rank" "aliased"

[6] "coef.std.error" "coef.t.value" "coef.p.value" "total.squares" "y.var"

[11] "sigma" "residual.variance" "r.squared" "f.pvalue" "df"

[16] "y.names" "deviance" "aic" "params" "formula"

[21] "call" "fstatistics" "adj.r.squared" "nValidObs" "nMissingObs"

[26] "coefLabelStyle"

> print(summary(reg1),header =FALSE)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.904e+02 2.956e-02 13209.28 2.22e-16 \*\*\*

DepDelay 4.389e-02 7.822e-04 56.12 2.22e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.9 on 7250010 degrees of freedom

Multiple R-squared: 0.0004342

Adjusted R-squared: 0.000434

F-statistic: 3149 on 1 and 7250010 DF, p-value: < 2.2e-16

Condition number: 1

#next step is done to check the delay sequence wise for example 0 – 10 mins delay and speed accordingly.

rxDataStep(inData=NewflightXdf,outFile=NewflightXdf,varsToKeep = "DepDelay",transformVars = "DepDelay",transforms =

list(F\_DepDelay=cut(DepDelay,breaks = seq(from=-10, to=100,by=10))),append = "cols",

overwrite=TRUE)

print(rxSummary(~F\_DepDelay,data = NewflightXdf),header = FALSE)

Category Counts for F\_DepDelay

Number of categories: 11

Number of valid observations: 6819077

Number of missing observations: 634138

F\_DepDelay Counts

(-10,0] 3894806

(0,10] 1316316

(10,20] 554103

(20,30] 321440

(30,40] 211002

(40,50] 150920

(50,60] 114696

(60,70] 87504

(70,80] 68825

(80,90] 55093

(90,100] 44372

reg2 <-rxLinMod(formula = airSpeed ~F\_DepDelay,data = NewflightXdf,rowSelection = (airSpeed>50)&(airSpeed <800),dropFirst = TRUE)

class(reg2)

print(summary(reg2),header =FALSE)

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 387.89666 0.03821 10151.854 2.22e-16 \*\*\*

F\_DepDelay=(-10,0] Dropped Dropped Dropped Dropped

F\_DepDelay=(0,10] 10.97080 0.07603 144.291 2.22e-16 \*\*\*

F\_DepDelay=(10,20] 9.72038 0.10829 89.758 2.22e-16 \*\*\*

F\_DepDelay=(20,30] 8.74882 0.13843 63.199 2.22e-16 \*\*\*

F\_DepDelay=(30,40] 7.56250 0.16862 44.850 2.22e-16 \*\*\*

F\_DepDelay=(40,50] 6.25256 0.19792 31.591 2.22e-16 \*\*\*

F\_DepDelay=(50,60] 5.52251 0.22610 24.425 2.22e-16 \*\*\*

F\_DepDelay=(60,70] 4.47101 0.25797 17.332 2.22e-16 \*\*\*

F\_DepDelay=(70,80] 3.76816 0.29028 12.981 2.22e-16 \*\*\*

F\_DepDelay=(80,90] 3.32997 0.32385 10.283 2.22e-16 \*\*\*

F\_DepDelay=(90,100] 2.50516 0.36051 6.949 3.68e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.2 on 6780242 degrees of freedom

Multiple R-squared: 0.004016

Adjusted R-squared: 0.004014

F-statistic: 2734 on 10 and 6780242 DF, p-value: < 2.2e-16

Condition number: 3.3165

From the above final analysis we can say for 0-10 minutes delay the speed increased to 10 and it goes on decreasing as the delay time increases