Solving Stats problem using R

## Problem statement for Stats:

Statistical Computing Org published Flight records in and out from US between 1987 – 2008. Its more than 12 GB of data that cannot be loaded into memory for processing. Main idea is how to play with data to get valuable insights.

<http://stat-computing.org/dataexpo/2009/the-data.html>

Following information needs to be analyzed using R:

1. Standard deviation: Check the standard deviation of distance travelled by American Airlines (AA)

2. Plot using R: Draw a boxplot of unique carrier distance with distance.

3. Plot using R: Direction of relationship between arrDelay and DepDelay by drawing a scattered plot.

4. Probability: What is the probability that a flight which is landing/taking off is “WN” Airlines

5. Probability: What is the joint probability of a flight getting cancelled and is supposed to travel less than 2000 miles given that the flight is “AA” Airlines.

6. Prediction: Suppose arrival delays of flights belonging to “AA” are normally distributed with mean 15 minutes and standard deviation 3 minutes. If the “AA” plans to announce a scheme where it will give 50% cash back if their flights are delayed by 20 minutes, how much percentage of the trips “AA” is supposed to lose this money.

7. Prediction: Assume that 65% of flights are diverted due to bad weather through the Weather System. What is the probability that in a random sample of 10 flights, 6 are diverted through the Weather System?

8. Linear regression: Do linear regression between the Arrival Delay and Departure Delay of the flights.

9. Multiple liner regression: Perform a multiple linear regression between the Arrival Delay along with the Departure Delay and Distance travelled by flights.

## What we want to achieve via this problem statement

Please note that each problem statement has 3 parts to be addressed, so even though solution of each one of them will be of a single line, there will be lot of thought process that must be given before reaching to that solution.

1. Mathematics itself

2. R language

3. Data is quite big in size, so cant be loaded and then can generate reports, sample data needs to be extracted out to ensure that P and Cs are as near to exact data as possible.

## Solutions

Standard

1. Standard deviation: Check the standard deviation of distance travelled by American Airlines (AA)

DF <- read.csv("/Users/anubandhans/Documents/R/2008.csv")

> View(DF)

> hist(DF$Distance,breaks = "fd")

> subset\_aa = subset(DF, UniqueCarrier = "AA")

> summary(subset\_aa$Distance)

Min. 1st Qu. Median Mean 3rd Qu. Max.

11.0 325.0 581.0 726.4 954.0 4962.0

> sd(subset\_aa$Distance)

[1] 562.1018

2. Plot using R: Draw a boxplot of unique carrier distance with distance.

DF <- read.csv("/Users/anubandhans/Documents/R/2008.csv")

boxplot(DF$Distance)

> boxplot(DF$Distance ~ DF$UniqueCarrier)

3. Plot using R: Direction of relationship between arrDelay and DepDelay by drawing a scattered plot.

DF <- read.csv("/Users/anubandhans/Documents/R/2008.csv")

plot(DF$DepDelay, DF$ArrDelay)

4. Probability: What is the probability that a flight which is landing/taking off is “WN” Airlines

> DF <- read.csv("/Users/anubandhans/Documents/R/2008.csv")

> dim(DF)[1]

[1] 7009728

> x <- subset(DF, uniqueCarrier = "WN")

> dim(x)[1]/nrow(DF)

[1] 1

5. Probability: What is the joint probability of a flight getting cancelled and is supposed to travel less than 2000 miles given that the flight is “AA” Airlines.

x <- subset(DF, uniqueCarrier = "AA")

> dim(x)

[1] 7009728 29

> y <- subset(DF, Distance < 2000)

> dim(y)[1]/ nrow(x)

6. Prediction: Suppose arrival delays of flights belonging to “AA” are normally distributed with mean 15 minutes and standard deviation 3 minutes. If the “AA” plans to announce a scheme where it will give 50% cash back if their flights are delayed by 20 minutes, how much percentage of the trips “AA” is supposed to lose this money.

DF <- read.csv("/Users/anubandhans/Documents/R/2008.csv")

> pnorm(20, mean = 15, sd=3)

[1] 0.9522096

7. Prediction: Assume that 65% of flights are diverted due to bad weather through the Weather System. What is the probability that in a random sample of 10 flights, 6 are diverted through the Weather System.

> dbinom(6, size =10, p=0.65)

[1] 0.2376685

8. Linear regression: Do linear regression between the Arrival Delay and Departure Delay of the flights.

> ArrDelay <- subset(DF, ArrDelay !="NA" & DepDelay!="NA")

> dim(ArrDelay)

[1] 6855029 29

> M <-lm(ArrDelay$ArrDelay ~ ArrDelay$DepDelay)

> M

Call:

lm(formula = ArrDelay$ArrDelay ~ ArrDelay$DepDelay)

Coefficients:

(Intercept) ArrDelay$DepDelay

-1.940 1.019

Summary(M)

confint(M)

2.5 % 97.5 %

(Intercept) -1.951354 -1.929553

ArrDelay$DepDelay 1.018350 1.018946

9. Multiple liner regression: Perform a multiple linear regression between the Arrival Delay along with the Departure Delay and Distance travelled by flights.

> ArrDelay <- subset(DF, ArrDelay !="NA" & DepDelay!="NA")

> M <-lm(ArrDelay$ArrDelay ~ ArrDelay$DepDelay + ArrDelay$Distance)

> M

Call:

lm(formula = ArrDelay$ArrDelay ~ ArrDelay$DepDelay + ArrDelay$Distance)

Coefficients:

(Intercept) ArrDelay$DepDelay ArrDelay$Distance

-1.061369 1.019154 -0.001213

Summary(M)

confint(M)

> confint(M)

2.5 % 97.5 %

(Intercept) -1.078702090 -1.044036839

ArrDelay$DepDelay 1.018856481 1.019452038

ArrDelay$Distance -0.001231805 -0.001194581