Teacher: Womble

Group Members:

John Saltanovitz; Tian Qiu; Zeqing Cai; Tyler Crumrin

30 points for organization/neatness and individual submission of individual effort.

1. **Data (5 points)**

|  |  |
| --- | --- |
| Variable | Variable Type |
| Arrival Delay (ArrDelay) | Discrete |
| Origin (Origin) | Categorical |

B. Decide on a Question (5 points)

What is the arrival delay relationship among the LAX JFK and BOS, these three different airports?

C. Inference Method 1 (50 points)

**a) Code has been Attached in appendix.**

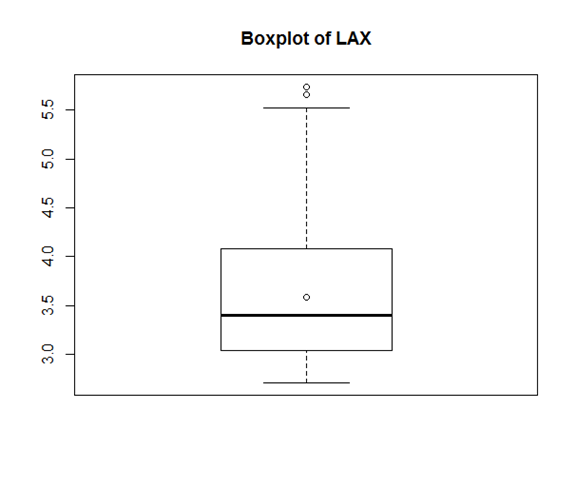
**b) (5 pts.) What procedure should be used and why? Remember, this needs to be determined**

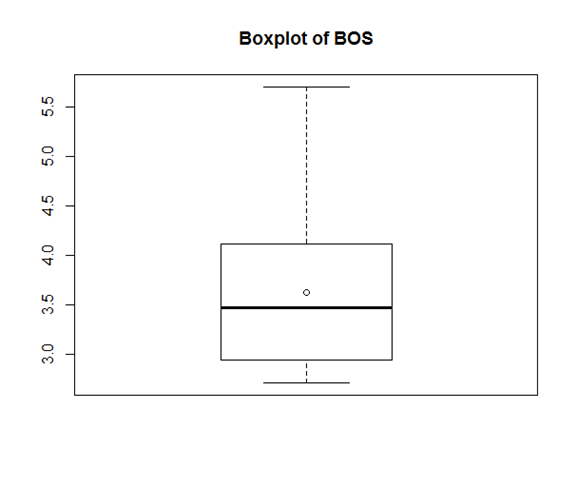
The two sample t-test hypothesis test procedure should be used because we want to compare the Arrival Delay (ArrDelay) for each of the 3 airports (LAX, JFK, and BOS). The two sample t hypothesis test allows allows us to compare the mean values of arrival delay for each airport to determine if there is a significant difference.

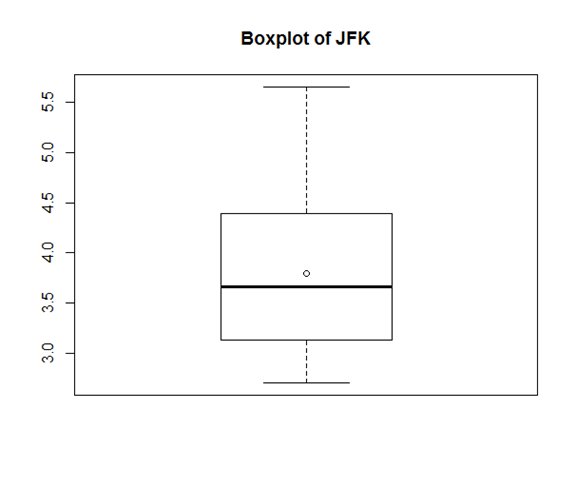
**BEFORE you analyze the data.**

**c) (5 pts.) Graphically display the data as appropriate for your answer in part b) with an interpretation of the output. If a transformation is needed, this should be done before this step.**

Boxplots of Log Arrival Delay for each airport:

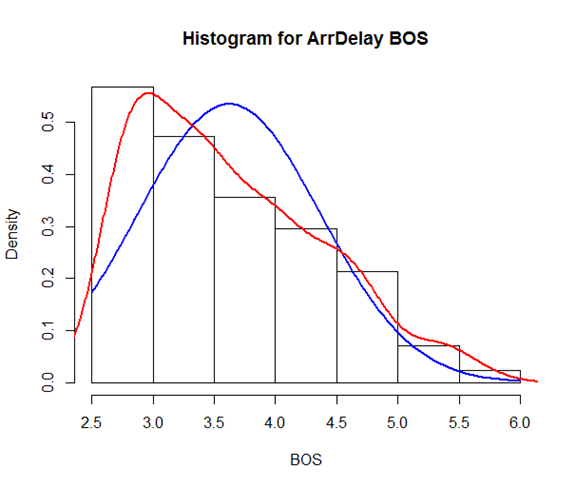


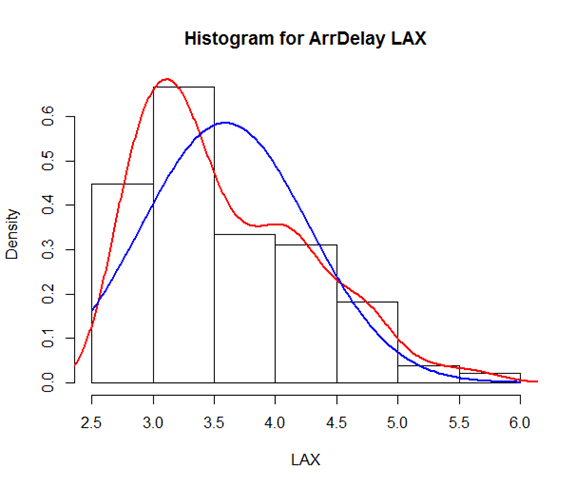
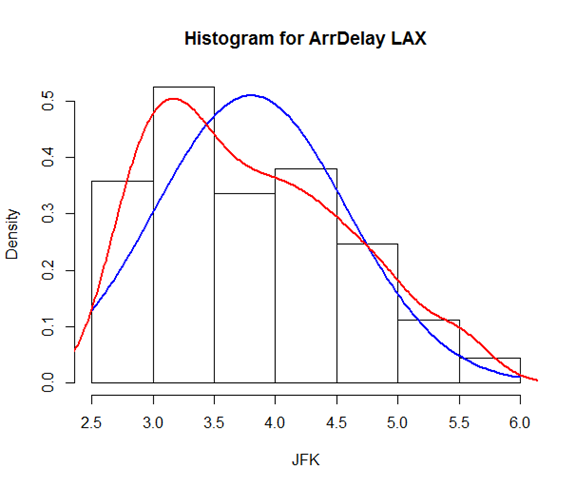


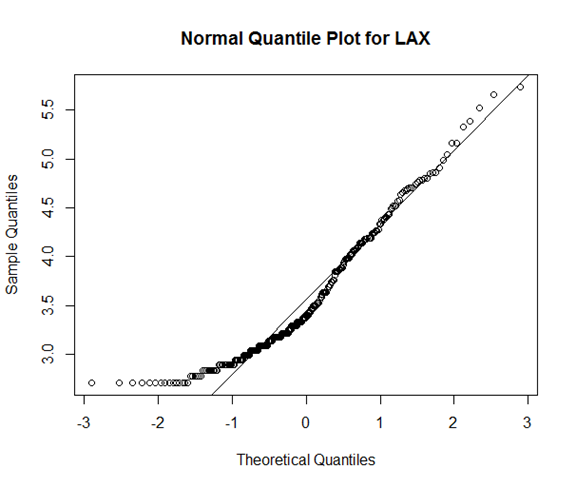


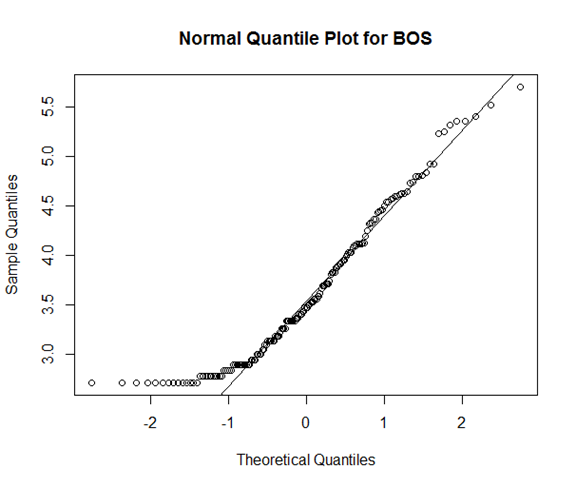
The distributions of arrival delay do not seem to vary significantly between each airport. Each distribution has very nearly the same values for means, inner fences, and outer fences, so we don’t expect to find significant differences in log arrival delay that correlate with these airports.

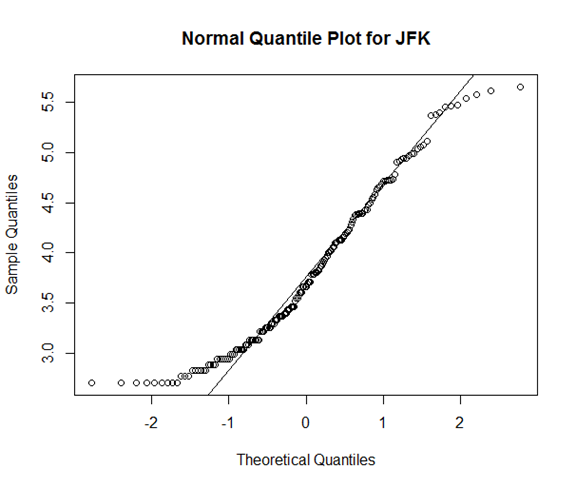
**d) (10 pts.) Determine if the appropriate assumptions are correct. Please provide all graphs and explain your decision. If the assumptions are not correct for your methodology and you perform the analysis, you will lose 25 points. You may assume that the data set is SRS as you have been assuming for the rest of the semester.**











We can see from the normal quantile plots that the distribution of arrival delay times for each airline is normal because the values in each sample follow the normal line with minimal skewing. We can assume that the airports are far enough away from each other that they do not affect each other's’ arrival delay times so they can be assumed to be independent. Lastly, we can assume that the variances in the values of arrival delay for each airport are the same because each flight will have to deal with the same factors that will cause flights to be delayed.

**e) (20 pts.) Perform the appropriate inference with a significance level of 0.05. This may consist of more than one step depending on the methodology in step b). The possible methodologies are**

**Step 1: Definition of the terms**

is the population mean of log Arrival Delay from LAX

is the population mean of log Arrival Delay from BOS

**Step 2: State the hypothesis**

H0:- = 0

Ha:

**Step 3: Find the Test Statistic, p-value, report DF**

t = -0.43638, df = 334.36, p-value = 0.6628

**Step 4: Conclusion**

alpha=.05

Since .6628>.05 we should fail to reject the H0, this means that and are the same.

**Step 1: Definition of the terms**

is the population mean of log Arrival Delay from LAX

is the population mean of log Arrival Delay from JFK

**Step 2: State the hypothesis**

H0:- = 0

Ha:

**Step 3: Find the Test Statistic, p-value, report DF**

t = -2.9209, df = 346.27, p-value = 0.003719

**Step 4: Conclusion**

alpha=.05

Since .003719<.05 we should reject the H0, this means that two of the mean log Arrival Delay times are different.

**Step 1: Definition of the terms**

is the population mean of log Arrival Delay from JFK

is the population mean of log Arrival Delay from BOS

**Step 2: State the hypothesis**

H0:- = 0

Ha:

**Step 3: Find the Test Statistic, p-value, report DF**

t = 2.1863, df = 345.96, p-value = 0.02946

**Step 4: Conclusion**

alpha=.05

Since .02946<.05 we should reject the H0, this means that two of the mean log Arrival Delay times are different.

Additionally we chose a Tukey Test because we want to compare all of the means in a pairwise fashion.

**f) (5 pts.) A conclusion in words that relates to the context of the question. This should be a short paragraph in length and should be understandable to someone who has not taken a course in statistics which explains your conclusions of the part. This should answer part of the question that you posed in Part B.**

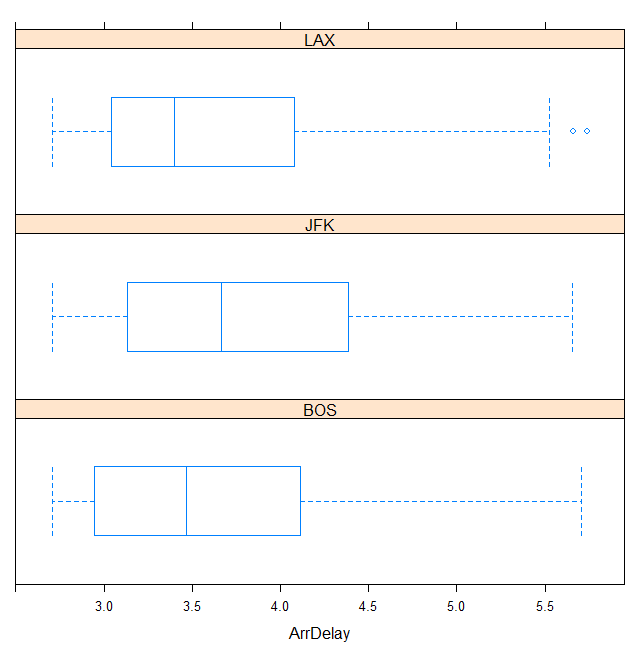
In conclusion, the t hypothesis test showed that we had to reject the null hypothesis that the population mean of log arrival delay from BOS and JFK, JFK and LAX are the same. We can say that the mean of log arrival delay from BOS and LAX is the same. So in one sentence is that in these three airports, only the BOS and LAX airport has the same population mean of log arrival delay time. You would expect the same delay time in these two airports.

D. Inference Method 2 (50 points)

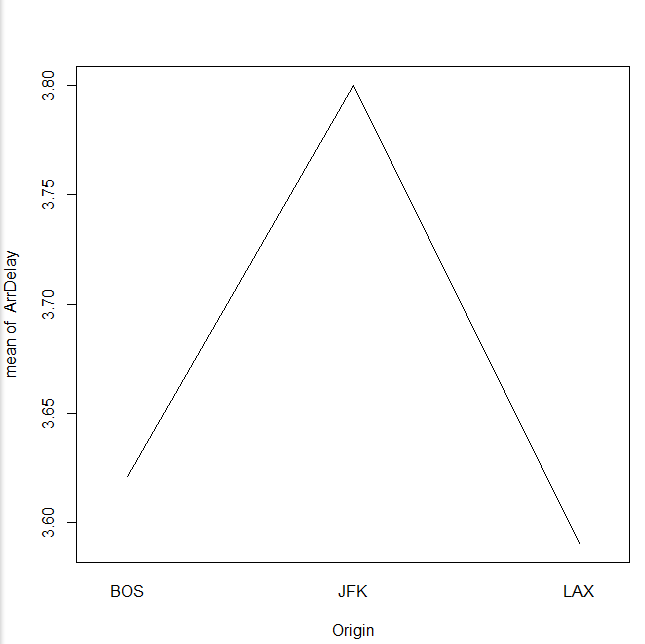
1. **Code is attached in appendix**
2. **What procedure should be used and why?**

The ANOVA procedure should be used because we want to compare the Arrival Delay (ArrDelay) from 3 different airports (LAX, JFK, and BOS). The ANOVA and the Tukey Method allow us to compare if there is a difference (significant) between the three airports. This can be down as long as all of the requirements are met for the ANOVA test.

1. **Graphically display the data as appropriate for your answer in part b) with an interpretation of the output.**



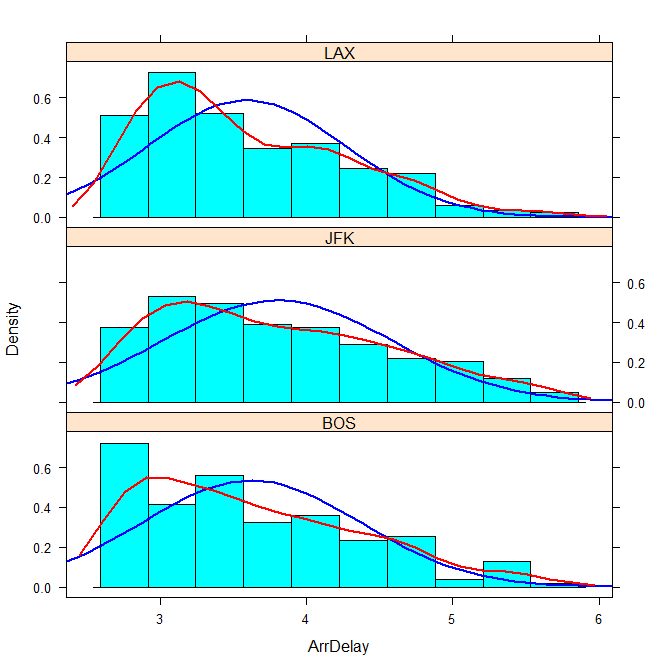
Even though the means are different, they are not significantly different since the boxplots overlap each other.



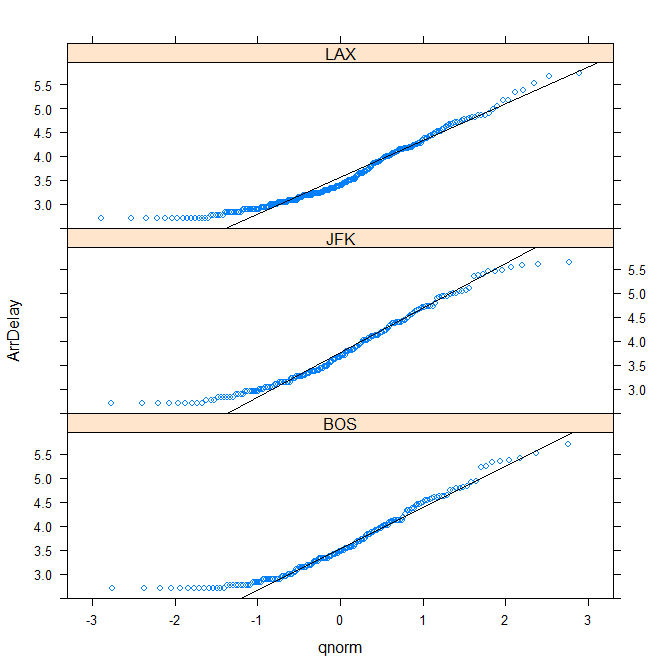
The Interaction plot shows that the mean ArrDelay is the highest for JFK but LAX and BOS are close together (similar).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Origin | n | sample mean transformed | sample mean | sample standard deviation transformed | sample standard deviation |
| LAX | 264 | 3.59 | 47.37 | 0.68 | 43.37 |
| BOS | 169 | 3.62 | 51.17 | 0.75 | 49.2 |
| JFK | 179 | 3.8 | 62.16 | 0.78 | 57.75 |

1. **Determine if the appropriate assumptions are correct. Please provide all graphs and explain your decision.**



The Histogram has a large sample size and the lines reflect data that is normally distributed. According to the histogram plots the data is appropriate for ANOVA analysis.



The Q norm shows that all of the data for each of the 3 airports lies close to the line. This is characteristic of data that is normally distributed. It is ok to continue with ANOVA analysis.

Constant standard Deviation

.78/.68=1.41<2

Therefore, it is ok to perform the ANOVA based on the SD analysis of the transformed data

1. **Perform the appropriate inference with a significance level of .05. (SEE PROJECT GUIDELINE FOR MORE INFORMATION)**

Df Sum Sq Mean Sq F value Pr(>F)   
Origin 2 5.1 2.5337 4.753 0.00895 \*\*  
Residuals 609 324.6 0.5331   
---  
Signif. codes:   
0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Step 1: Definition of the terms**

M1 is the population mean of log Arrival Delay from LAX

M2 is the population mean of log Arrival Delay from JFK

M3 is the population mean of log Arrival Delay from BOS

**Step 2: State the hypothesis**

Ho:M1=M2=M3

Ha:at least two M’s are different

**Step 3: Find the Test Statistic, p-value, report DF**

Fts=4.735

DF1=2

DF2=609

P-Value=.00895

**Step 4: Conclusion**

alpha=.05

Since .00895<.05 we should reject the Ho, this means that at least two of the mean log Arrival Delay times are different.

Additionally we chose a Tukey Test because we want to compare all of the means in a pairwise fashion.

Tukey multiple comparisons of means  
 95% family-wise confidence level  
  
Fit: aov(formula = ArrDelay ~ Origin)  
  
$Origin  
 diff lwr upr p adj  
JFK-BOS 0.17914636 -0.004834109 0.36312683 0.0582761  
LAX-BOS -0.03102562 -0.200011573 0.13796033 0.9025533  
LAX-JFK -0.21017198 -0.376255047 -0.04408892 0.0085963

This data shows that LAX-BOS and LAX-JFK are significantly different since the bounds do not contain 0.

1. **Conclusion in words that relates to the context of the question.**

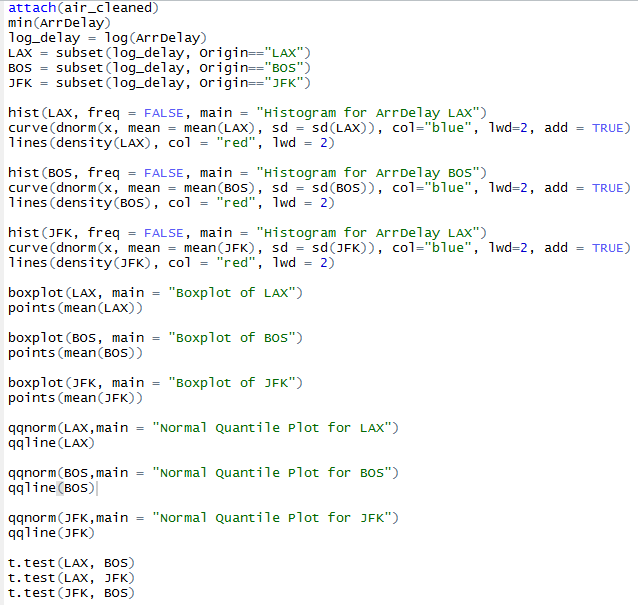
In conclusion, I would summarize by saying that the Los Angeles Airport tends to have a longer Arrival Delay Time at other airports compared to Boston and New York. The ANOVA showed that we had to reject the null hypothesis that all three arrival delay times were the same. Additionally, the Tukey test showed that LAX-BOS and LAX-JFK were significantly different but JFK-BOS were not. This shows again that LAX is not the same as the other two airports. When looking at the previous graphs it can be noted that LAX had higher means in all of the visual displays so this is not a surprise. My final conclusion from this portion would be that people flying from LAX or expecting flights from LAX should expect more delays when compared to flights leaving from Boston or New York.

**E. Write a Final Conclusion (10 points)**

Using our first inference method, a two sample t hypothesis test, we determined that the mean of logarithm of arrival delay was not significantly different between the LAX and BOS airports, but was significantly different between JFK and both of these airports. Using our second inference method, an ANOVA test, we determined that the mean of logarithm of arrival delay was not significantly different between JFK and BOS, but was significantly different between LAX and both of these airports. When comparing the results of two tests, given that the results are contradictory to each other, it’s impossible to say for sure if there are differences in the true means of arrival delay. Based on the results from the t-test and the mean values of arrival delay for LAX, BOS, and JFK, I would concluded it’s much more likely that LAX and BOS have the same true mean arrival delay time, but it’s impossible to know for sure.

**Appendix:**

1. Method 1 code:



2. Method 2 code

**Organizing the data/Transformation**

>airline.new <- subset(airline\_calc, airline\_calc$Origin=="LAX"|

>airline\_calc$Origin=="BOS"|

>airline\_calc$Originr=="JFK")

>airline.new <- droplevels(airline.new)  
> library(lattice)  
> airline.new$ArrDelay<-log(airline.new$ArrDelay)  
> attach(airline.new)

**Creating the Boxplot**  
> windows()  
> bwplot(~ArrDelay|Origin,layout=c(1,3),pch="|")  
> windows()

**Interaction Plot**  
> trace<-rep(1,length(Origin))  
> interaction.plot(Origin,trace,ArrDelay,fun=mean,legend=F)

**Adding Data to table**

> tapply(ArrDelay,Origin,mean)  
 BOS JFK LAX   
3.621212 3.800358 3.590186   
> tapply(ArrDelay,Origin,length)  
BOS JFK LAX   
169 179 264   
> tapply(ArrDelay,Origin,sd)  
 BOS JFK LAX   
0.7463028 0.7822403 0.6814769

> tapply(ArrDelay,Origin,mean)  
 BOS JFK LAX   
51.17160 62.16201 47.37879   
> tapply(ArrDelay,Origin,sd)  
 BOS JFK LAX   
49.20221 57.74714 43.37554

**Histogram**  
> histogram(~ArrDelay | Origin, layout=c(1,3),type="density",  
+ panel=function(x,...)  
+ {panel.histogram(x,...)  
+ panel.mathdensity(dmath=dnorm,col="blue",lwd=2,  
+ args=list(mean=mean(x, na.rm=T), sd = sd(x,na.rm=T)),...)  
+ panel.densityplot(x,col="red",lwd=2,...)  
+ })

**QNORM line**

> qqmath(~ArrDelay| Origin, layout=c(1,3),panel = function(x){  
+ panel.qqmath(x)  
+ panel.qqmathline(x)  
+ })

**ANOVA**  
> fit<-aov(ArrDelay~Origin)  
> summary(fit)  
 Df Sum Sq Mean Sq F value Pr(>F)   
Origin 2 5.1 2.5337 4.753 0.00895 \*\*  
Residuals 609 324.6 0.5331   
---  
Signif. codes:   
0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**TUKEY Method**

> test.Tukey<-TukeyHSD(fit,conf.level=.95)  
> test.Tukey  
 Tukey multiple comparisons of means  
 95% family-wise confidence level  
  
Fit: aov(formula = ArrDelay ~ Origin)  
  
$Origin  
 diff lwr upr p adj  
JFK-BOS 0.17914636 -0.004834109 0.36312683 0.0582761  
LAX-BOS -0.03102562 -0.200011573 0.13796033 0.9025533  
LAX-JFK -0.21017198 -0.376255047 -0.04408892 0.0085963