Biostatistics 230 Statistical Graphs

Statistical Graphics Final Data Analysis Project:

Analysis of Delay of Commercial Flights in US in 2008

Ritesh Varyani

(UID: 904-406-389)

Source Of Dataset

For this project of Data Analysis, I analyze the US flight traffic during the year 2008. The dataset we consider is from ASA Sections on Statistical Computing and Statistical Graphics. This data is from a Data expo, 2009. The original source of the dataset is from united States Department of Transportation.

This is a dataset of flight arrival and departure details for all commercial flights within the USA for the year 2008. The dataset has a total of **7,009,728** rows and **29** columns. However, since there are 7 million rows in the dataset and 29 features ,we analyze the data only for 1st month of the year 2008. The resultant data has **605,765** rows and **29** columns.

Dataset Description And Preprocessing

The data we analyze is of one month and has features like DayOfWeek, DayOfMonth, DepTime, CRSDepTime(scheduled departure time), ArrTime, CRSArrivalTime(scheduled arrival time), UniqueCarrier, Origin, Destination, AirTime, ArrDelay, DepDelay, Distance and others. We ignore the flights which are cancelled or diverted because we do not have their final destinations in the dataset, hence analyzing the delay is not possible for them. Thus, after this data cleaning we are left with **587130** records with **24** features.
We then order the data by the date of the month and perform the analysis.

Since we have ~600,000 records to analyze. We begin with a little pre-processing where we find out all the different airports, different carriers, the number of flights departing or arriving at the airports throughout the period of one month. After all these preprocessing, we are able to start our analysis of delay.

The preprocessing result show that there are a total of **286** airports which have records of commercial flights. There are a total of **20 unique carriers** which are responsible for the commercial airline traffic.

Goals In Analysis Of Dataset

While performing analysis, we are indirectly figuring out which Carrier flights are usually late, which airports tend to be more congested and which flight-carriers lag in maintenance, thereby causing flight delays. Are there some delays for which the fight carriers and the airports, both are not responsible, like the weather.

We also identify whether there is correlation between the delays of flights related to their origins or destinations. Are these delays related to the distance of the flights? Are delayed flights for long-distances able to cover-up the delay by minimizing the airtime.

Analysis

After our preprocessing, we analyze the following aspects of the commercial airline records:

1. Is there a particular day when you must avoid boarding a flight which has higher chance of getting delayed?

We find out the percentage delay of flights across all airports during various days of the week, and see that there is are 21.50% to 29.50% of flights daily are delayed. The least percent of flight delayed are on Saturday and the flights the most delayed on a Thursday of a week. We are able to see these results in plot 1.
On normalizing this across all the days of the week, the probability of delay on any day of a flight is from 0.122(Saturday) to 0.168(Thursday). We are able to see these results in plot 2.

2. Which carrier to avoid when you book a flight in order to minimize your delay at the airport?

With this analysis, we find out which carrier has most delay and we must avoid while booking our flights. From our results in plot 3, we realize that WN(SouthWest) Airlines flights are more delayed as compared to other carriers. More than 20000 flights of SouthWest are delayed. The next carrier with second highest number of delayed flights is AA(American Airlines). About 16000 American Airlines flights are delayed. The least number of delayed flights are by HA(Hawaiian Airlines) and AQ(Aloha Airlines) which is about 250.
However, we also try to find out what percentage of total flights of a carrier are actually delayed as this might help us to better understand the delay rate of a carrier. From plot 4 we realize that UA(United Airlines) is the worst of the lot. About 35% of total flights by United Airlines are delayed. At the same time, the HA(Hawaiian Airlines) and the AQ(Aloha Airlines) have the least percentage delay(5-6%).

3. Identification of hubs based on the number of arriving and departing flights of different carriers

Next, we find out hubs of different carriers based on the number of arrivals and departures of flights throughout the month. The intention behind this is to measure delays at these hub airports and how they vary with non-hub airports. The table 1 shows the different hubs of different carriers.

4. Analyzing whether hubs airports are actually responsible for a major chunk of delays of flight-carriers

For this analysis, we calculate the total delay of all carrier across all airports. We end up with matrices of 286 airports as rows and 20 carriers as columns and we get and store the top 3 airport with maximum delay for each carrier. We compare these airports to our hubs found out in part 3 for respective carriers. Our conclusion is that there is no hub airport, which ranks in top 3 delays of any carriers. This is shown by table 2 which shows the top three departure delays of carriers across all airports and table 3 which shows the top three arrival delays of carriers across all airports. This actually shows that hub airports are good at handling their carriers and manage the take-offs and landings of these flights efficiently. Choosing a carrier whose hub is either your origin or your destination, gives us guarantee that our flight is least likely to get delayed.

5. In order to avoid delay, should you book from a airport which is smaller, or not a hub?

We measure number of arrivals and departures from all airports, and find out whether it is appropriate to book a flight from a big airport, which has a lot of flight arrivals and departures, or should we book from a smaller airport to minimize the chances of getting delayed. So we measure the departure delay against number of departures across all airports, and also the arrival delay against number of arrivals across all airports. The distributions obtained in plots 5 and 6 are both skewed and so, we take logarithms of these distributions along the number of departures or arrivals in both these plots to see it more clearly. The logarithms of these distributions are shown in plot 7(departure delay against logarithm of number of departures)) and plot 8(arrival delay against logarithm of number if arrivals).

From these two plots, we see that, the delays are spread when the number of arrivals or departures is less and for the airports which are big enough to handle more airline departures and arrivals, this departure or arrivals delay narrows down as compared to these smaller airports. It is proper to book your flights from airports which handle more flights or from hub airports.

Although, there is one interesting point, for airport which only handle less than 250 flights, there are some airports with negative delays. These might be the airports which are in remote areas which is why there might not be much flight traffic.

6. If your flight is delayed, is there a chance you might still reach on time or make-up for some time if it is a long distance flight?

In this part, we are finding out whether arrival delay is influenced by the distance of the flight. If it is a long distance flight, is there a chance that some delay is reduced and the time is made up during the flight. From the results in the plots 9 and 10, where we show the arrival delay as a function of distance of flights, we have plotted these plots for two of the twenty carriers-AA(American Airlines)(plot 9) and DL(Delta Airlines)(plot 10).

From these plots, we are able to see that the regression line fitted on both these plots has slope of ~0(almost 0). This indicates that there is not much correlation between the arrival delay of a flight and the distance which it has travelled. Hence, if you are on a long-distance flight which is delayed, more often than not, you will not reach on -time at your destination or cover-up some of the time lost.

7. Does weather play a part in the delay of flights? How much percent of flights delayed are delayed because of weather?

We find out what percent of all delay is due to weather. We do this to understand, which carriers' flights were actually affected by weather and were delayed not because of airport administration or the carrier's lack of efficiency.

From the plot 11, we are able to see, that a majority of carrier(17 from 20) have 0.7% to 13% percent of all delays caused have been contributed to by weather. There is only one carrier OH, for which weather is responsible for 32.50% of delays.

There are two carriers HA(Hawaiian Airlines) and AQ(Aloha Airlines), which have a negative weather delay, because there overall arrival delay is negative. Throughout the month, their cumulative arrival delay is negative, which implies that most of their fights often reached their destinations before the scheduled arrival time.

Conclusion

Thus, from our analysis, we are able to see that if we are to plan our domestic journey in the US, we must book the flight of a carrier which has a hub at either or source or the destination, should avoid booking a flight on a Thursday. Also, we must try to avoid United Airlines as 35% of its total flights are delayed. We must try to book our flights from airports which handle larger traffic, because they usually have very less delays, as found out in part 5 of the analysis. If either HA(Hawaiian Airlines) or AQ(Aloha Airlines) carriers fly from our origin to the destination, we must take them, as they have been found to reach their destinations before the scheduled arrival time, resulting in negative delays. However, these airlines have very few flights in a month(about 250) and between limited airports and so are not always viable option.

Appendix

Tables:

Carrier	Hub Airport Found
9E	DTW
AA	DFW
AQ	HNL
AS	SEA
B6	JFK
CO	IAH
DL	ATL
EV	ATL
F9	DEN
FL	ATL
HA	HNL
MQ	DFW
NW	DTW
OH	CVG
OO	SLC
UA	ORD
US	CLT
WN	LAS
XE	IAH

Table 1: Hub airports for carriers.

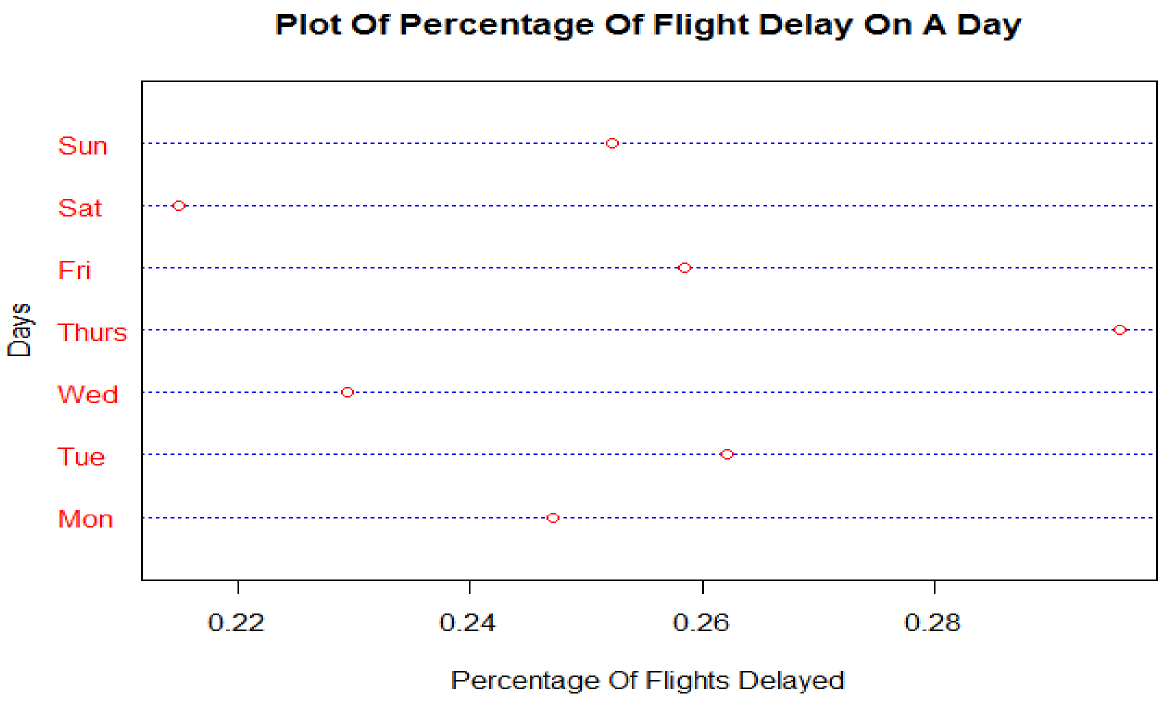
Hub Airport Found	**Highest Departure Delay**		
	First Airport	Second Airport	Third Airport
DTW	PLN	SLC	FWA
DFW	GUC	RNO	SFO
HNL	RNO	SNA	SMF
SEA	SFO	ORD	MIA
JFK	ORD	SFO	FLL
IAH	ORD	BDL	SFO
ATL	LIT	OAK	TYS
ATL	DFW	RDU	SAV
DEN	SFO	GEG	LAX
ATL	SFO	MSY	FLL
HNL	SMF	SFO	SJC
DFW	MQT	ROC	GRB
DTW	STT	SJU	SFO
CVG	ROA	TPA	FWA
SLC	ROA	IND	CEC
ORD	GSO	JAC	RIC
CLT	ORD	ANC	SJU
LAS	SFO	MDW	LAX
IAH	BGR	SFO	ORD
DTW	AVP	TVC	BUF

Table 2: Comparison of highest departure delay with the hub airports found. There is not a single match which indicates hub airports do not have significant departure delays.

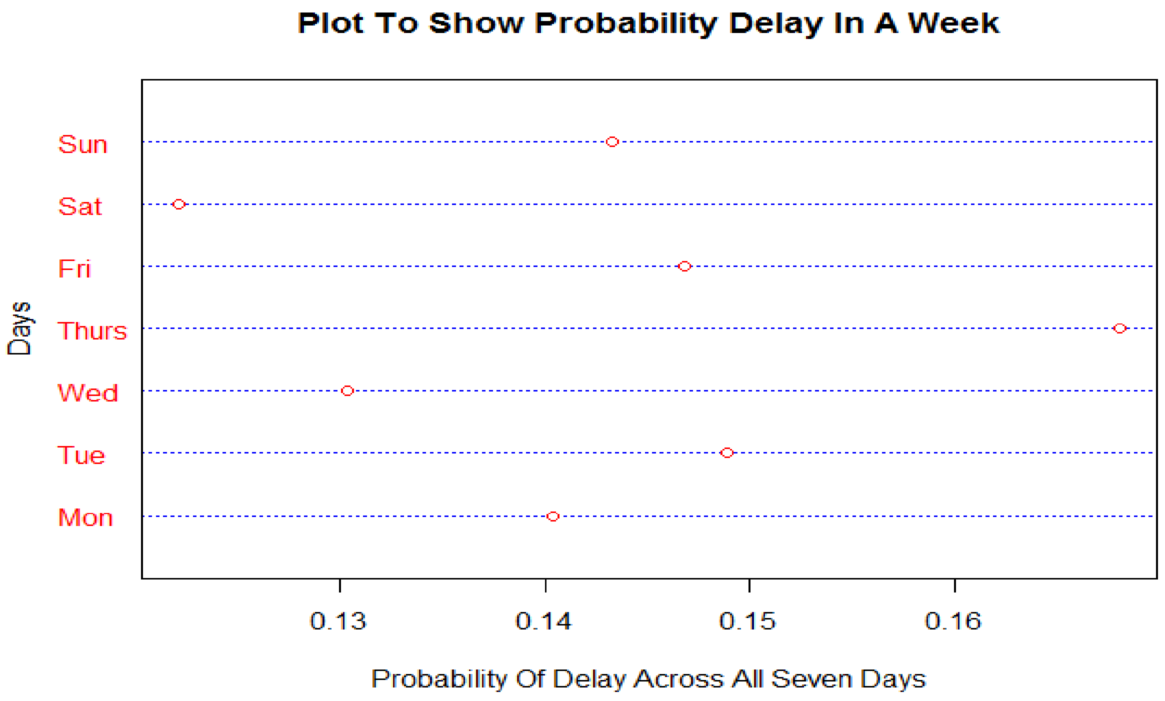
Hub Airport Found	**Highest Arrival Delay**		
	First Airport	Second Airport	Third Airport
DFW	SFO	RNO	PSP
HNL	OAK	SAN	SNA
SEA	CDV	YAK	ORD
JFK	SFO	ORD	LGA
IAH	ORD	SFO	BDL
ATL	OAK	CAE	MKE
ATL	SAV	RDU	MCI
DEN	SFO	ABQ	GEG
ATL	SFO	LAX	DFW
HNL	LAS	SMF	SFO
DFW	MQT	CMI	TVC
DTW	SFO	STT	CID
CVG	DEN	TPA	HPN
SLC	ORF	SAV	SFO
ORD	GSO	MDT	PBI
CLT	SFO	ORD	KOA
LAS	SFO	TUS	LAX
IAH	ORD	SFO	LGA
DTW	CAK	CWA	MBS

Table 3: Comparison of highest arrival delay with the hub airports found. There is not a single match which indicates hub airports do not have significant arrival delays.

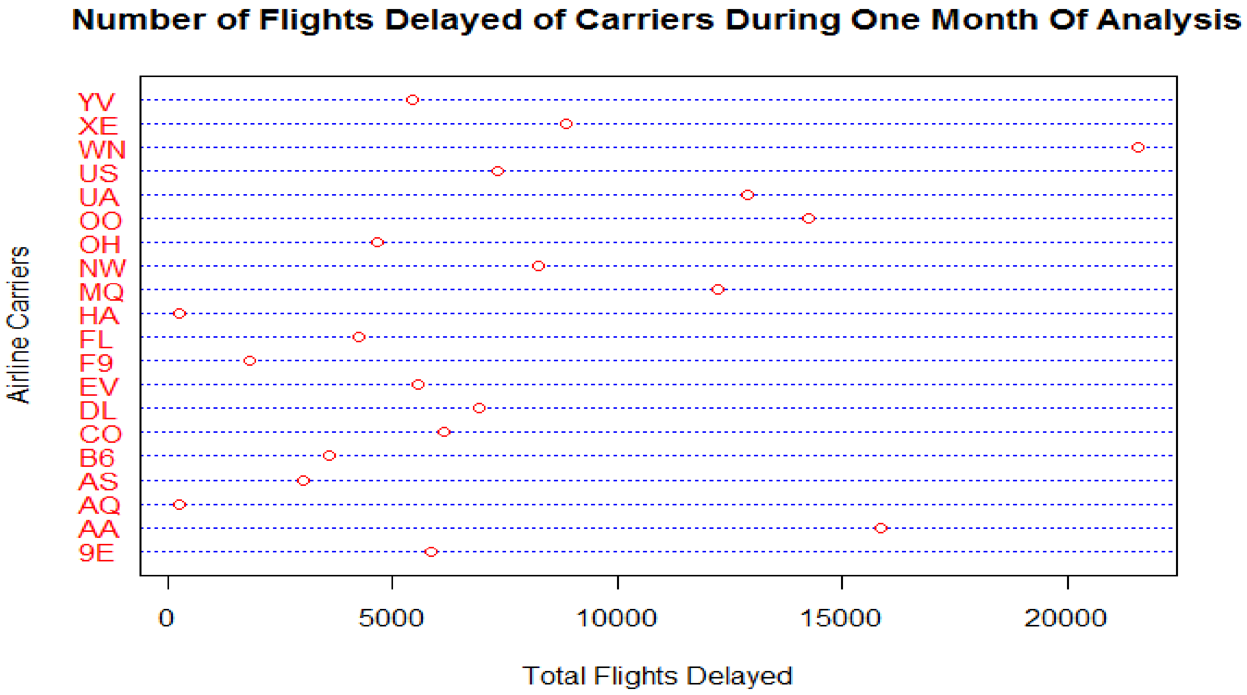
Plots:



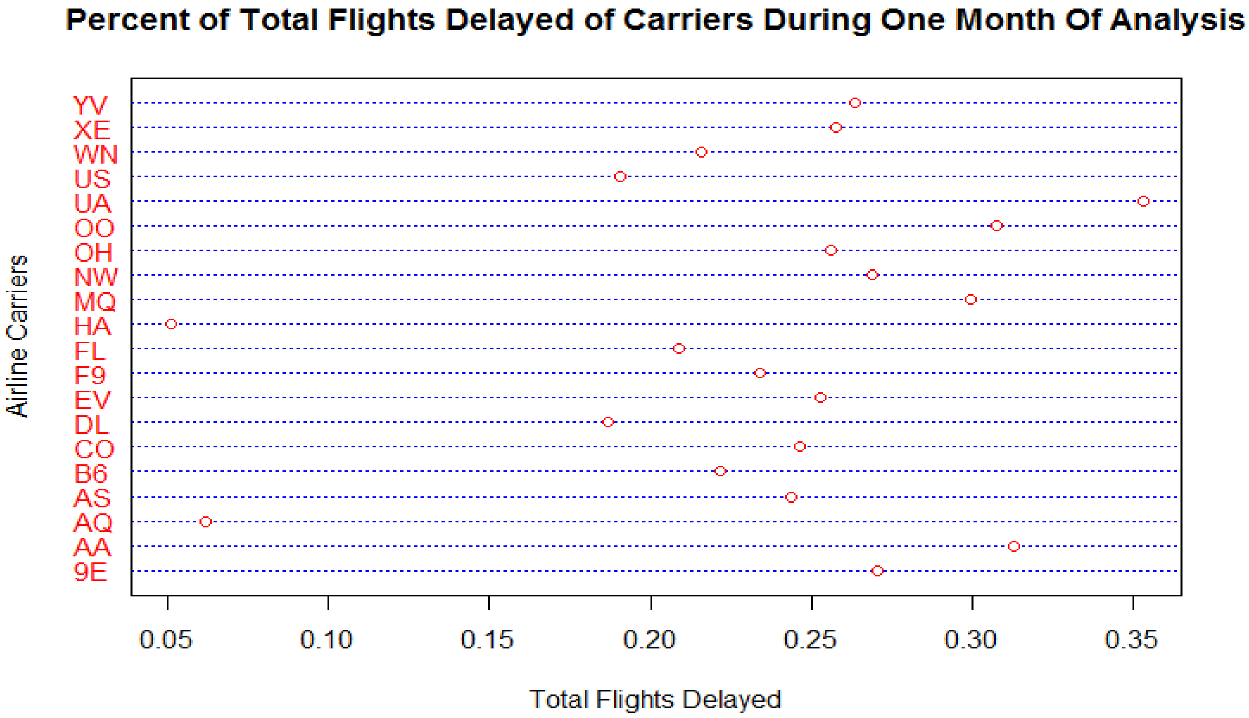
Plot 1: Percentage of flights delayed daily throughout the month.



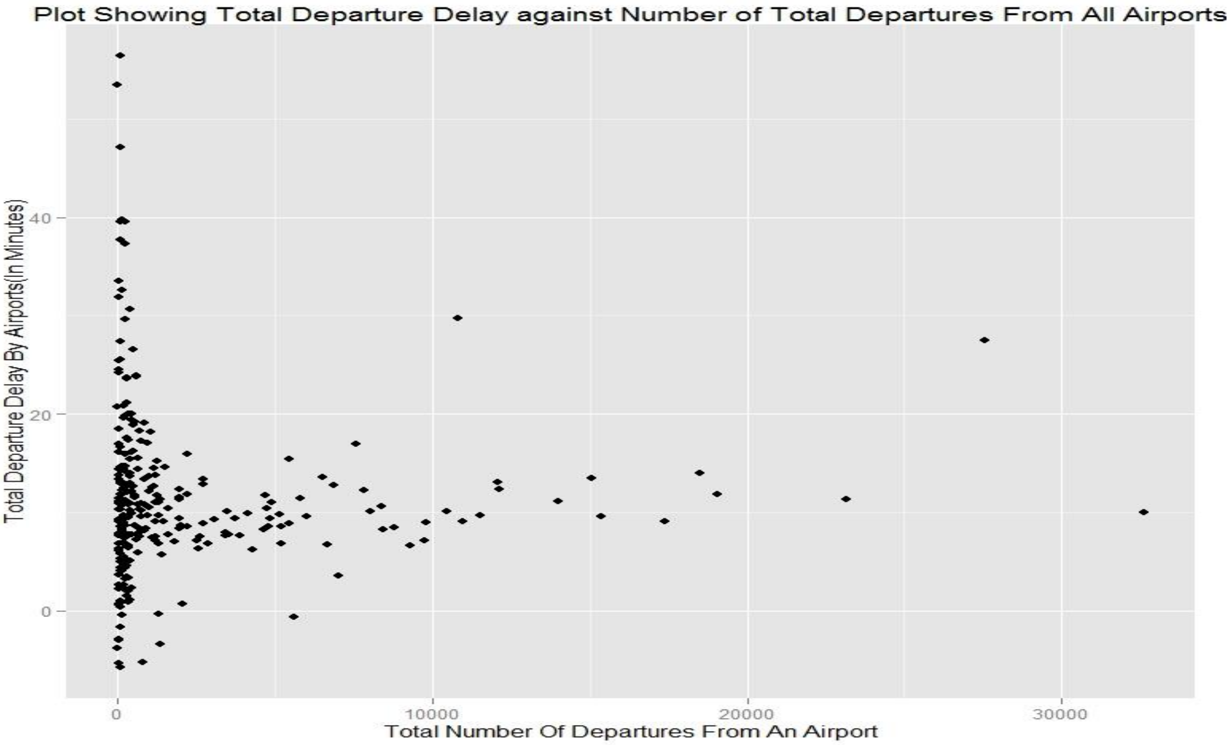
Plot 2: Probability of delay of flights on a day of a week.



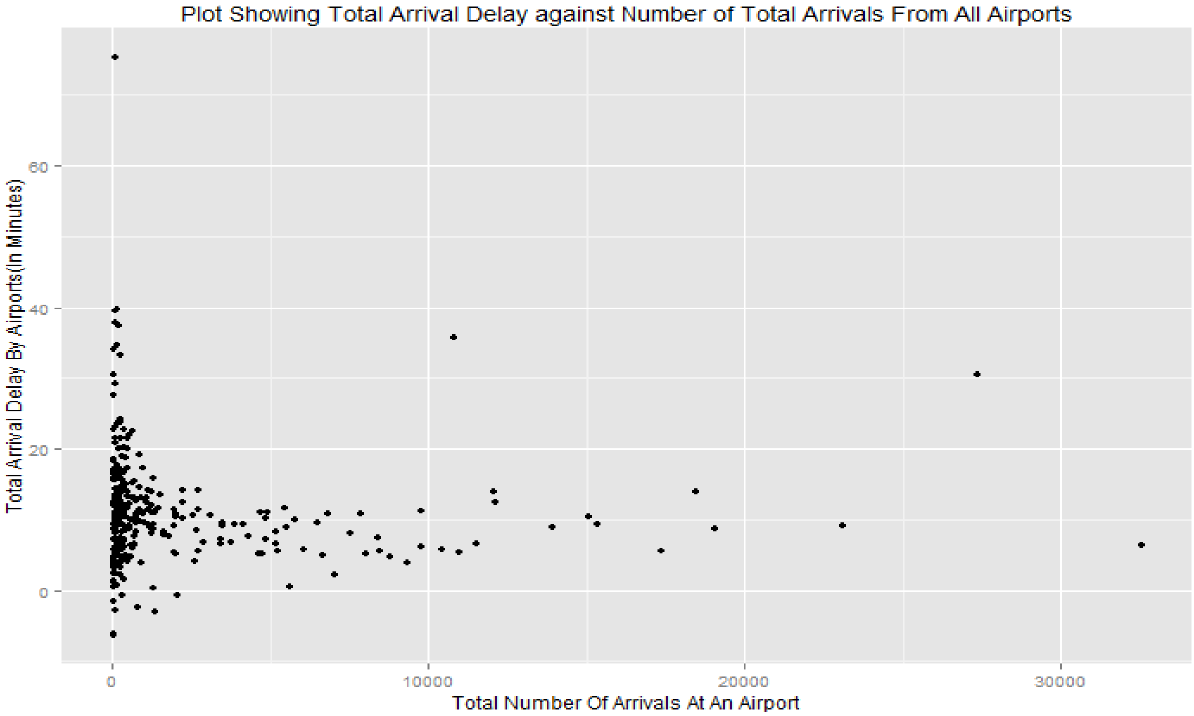
Plot 3:Plot of number of flights delayed by carrier.



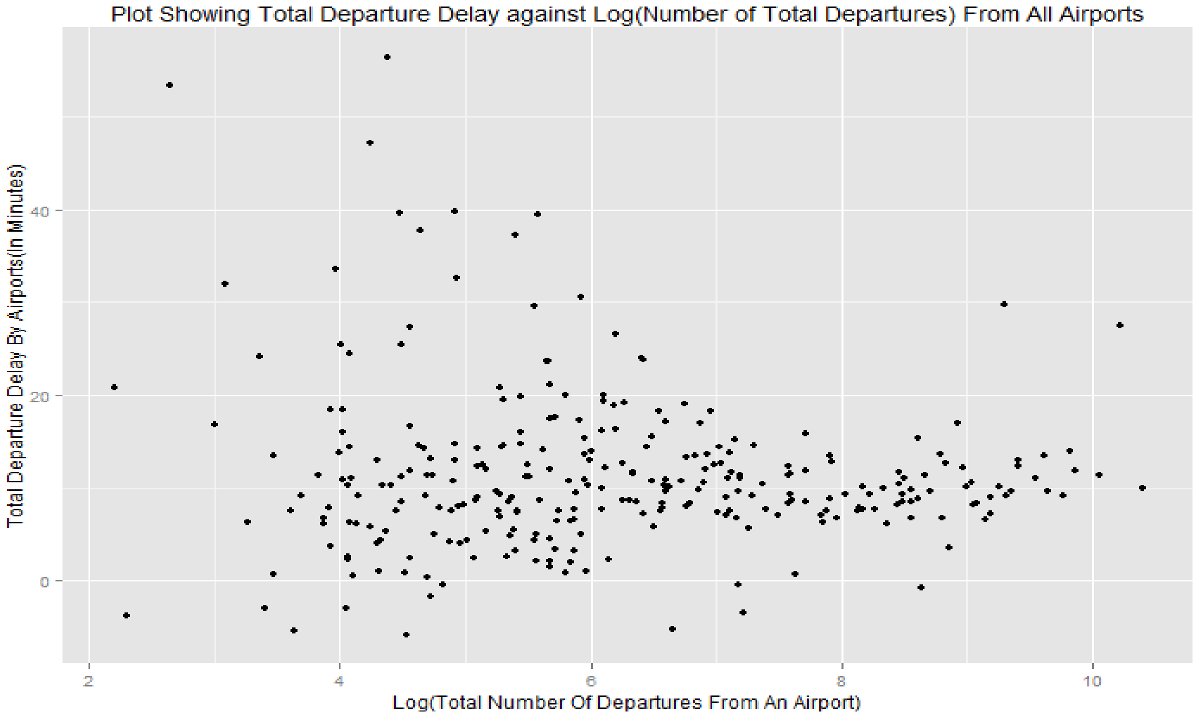
Plot 4: Percent of total flights which are delayed shown by carrier.



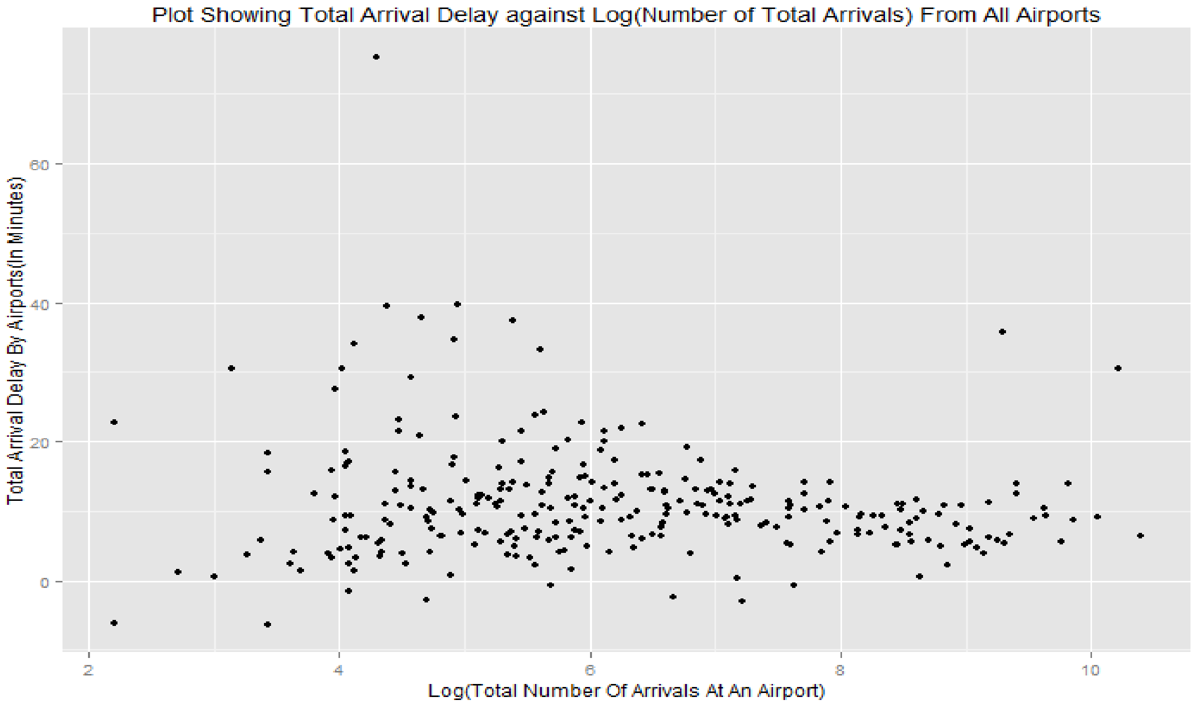
Plot 5: Qplot showing total departure delay against total number of departures across all airports.



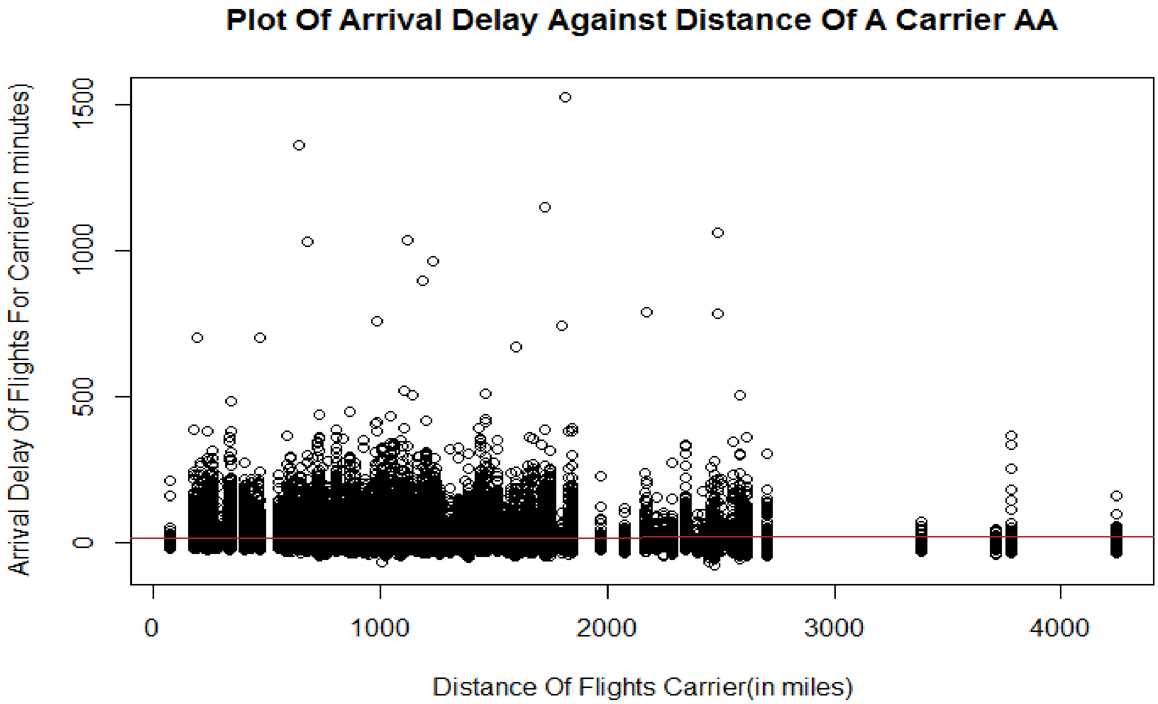
Plot 6: Qplot showing total arrival delay against total number of arrivals across all airports.



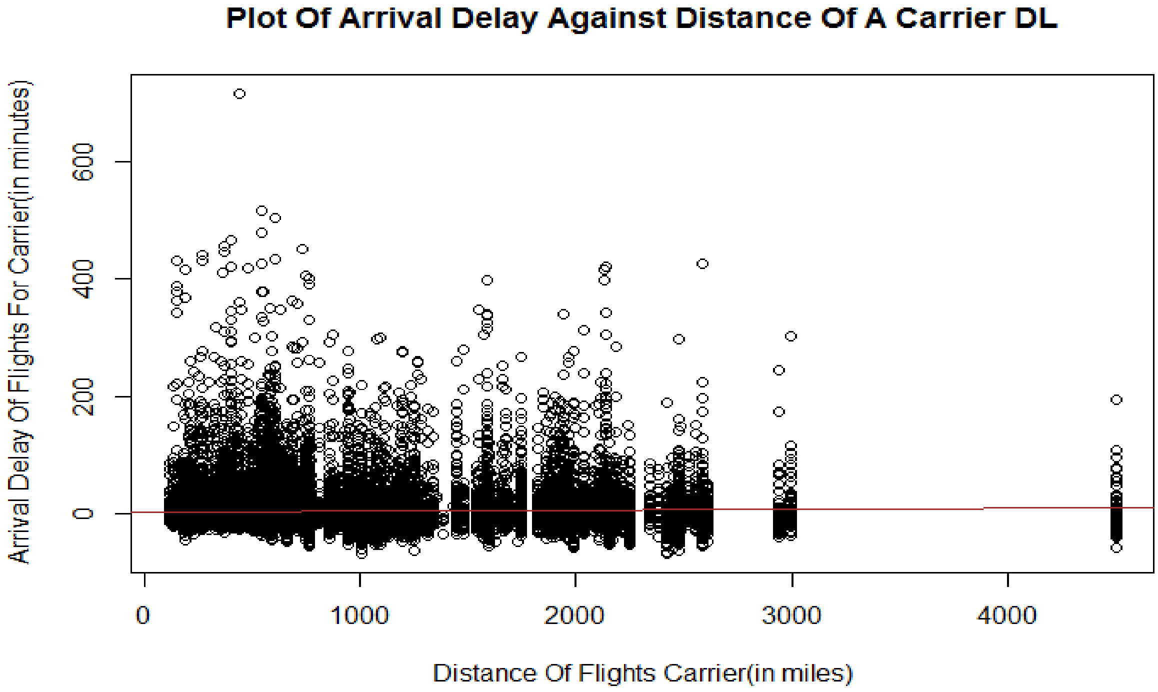
Plot 7: Qplot showing logarithm of total departure delay against logarithm of total number of departures across all airports.



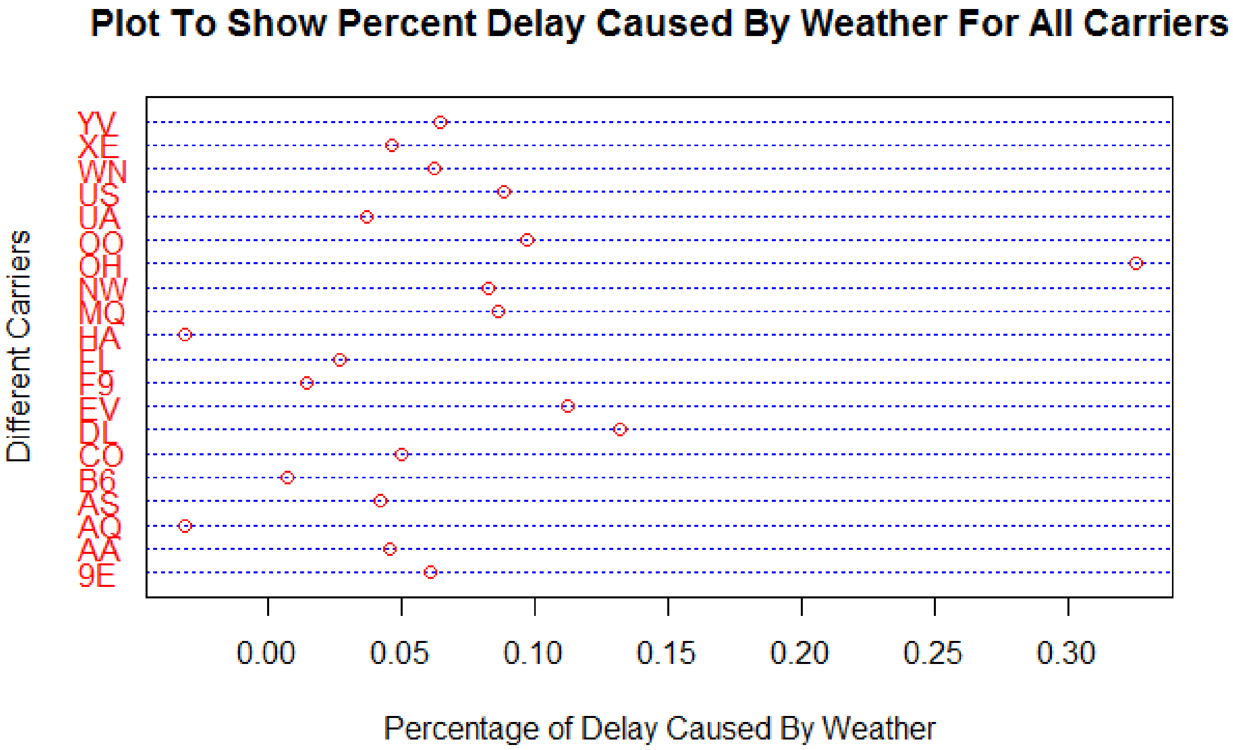
Plot 8: Qplot showing logarithm of total arrival delay against logarithm of total number of arrivals across all airports.



Plot 9: Plot of arrival delay of individual flights against distance of each of the flights for carrier AA(American Airlines)



Plot 10: Plot of arrival delay of individual flights against distance of each of the flights for carrier DL(Delta Airlines)



Plot 11: Plot of percentage of delay caused by weather shown across different carriers.