Statistical graphics Final Data Analysis Project Proposal:

Analysis of Delay of Commercial Flights in US in 2008

Ritesh Varyani

UID:904-406-389

SOURCE OF DATASET

For the 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 is the link for the website. This data is from a Data expo, 2009. The original source of the dataset is from united States Department of Transportation, found on this link. For downloading the dataset, this is the direct link.

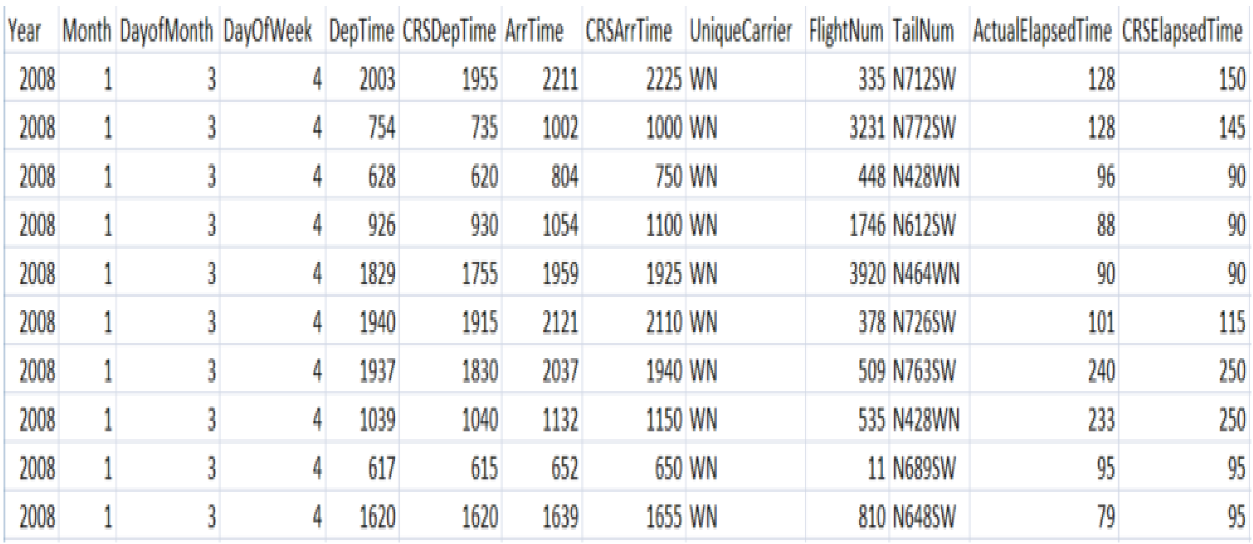
This is a dataset of flight arrival and departure details for all commercial flights within the USA for the year 2008. The dataset has a total of **7,009,728** rows and **29** columns. However, since there are 7 million rows in the dataset and 29 features ,we analyze the data only for 1st month of the year 2008. The resultant data has **229,292** rows and **29** columns.

GOALS OF ANALYSIS

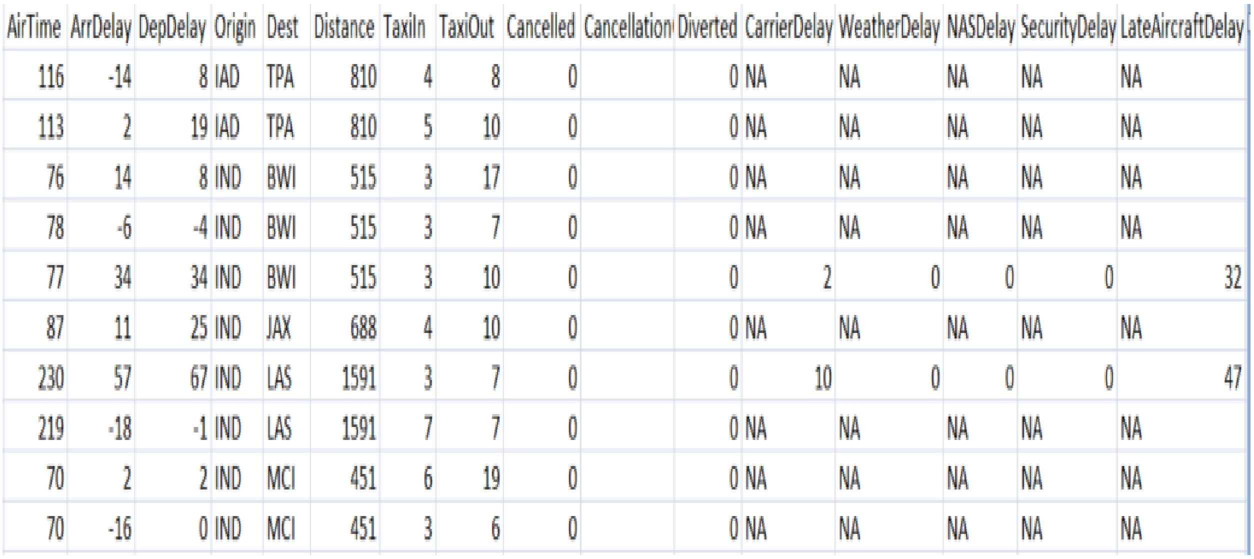
The main aim for this analysis is to identify which Carrier flights are usually late, which airports tend to be more congested and which flight-carriers lag in maintenance, thereby causing flight delays.

During this analysis, the aim lies to identify whether there is co-relation between the delays of flights related to their origins or destinations. We try to find the peak hours which signify too much amount of traffic at airports and how the delay of airlines vary during these peak hours.

DETAILS OF THE DATASET



Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime
2008	1	3	4	2003	1955	2211	2225	WN	335	N712SW	128	150
2008	1	3	4	754	735	1002	1000	WN	3231	N772SW	128	145
2008	1	3	4	628	620	804	750	WN	448	N428WN	96	90
2008	1	3	4	926	930	1054	1100	WN	1746	N612SW	88	90
2008	1	3	4	1829	1755	1959	1925	WN	3920	N464WN	90	90
2008	1	3	4	1940	1915	2121	2110	WN	378	N726SW	101	115
2008	1	3	4	1937	1830	2037	1940	WN	509	N763SW	240	250
2008	1	3	4	1039	1040	1132	1150	WN	535	N428WN	233	250
2008	1	3	4	617	615	652	650	WN	11	N689SW	95	95
2008	1	3	4	1620	1620	1639	1655	WN	810	N648SW	79	95



AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	Cancellation	Diverted	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay
116	-14	8	IAD	TPA	810	4	8	0		0	NA	NA	NA	NA	NA
113	2	19	IAD	TPA	810	5	10	0		0	NA	NA	NA	NA	NA
76	14	8	IND	BWI	515	3	17	0		0	NA	NA	NA	NA	NA
78	-6	-4	IND	BWI	515	3	7	0		0	NA	NA	NA	NA	NA
77	34	34	IND	BWI	515	3	10	0		0	2	0	0	0	32
87	11	25	IND	JAX	688	4	10	0		0	NA	NA	NA	NA	NA
230	57	67	IND	LAS	1591	3	7	0		0	10	0	0	0	47
219	-18	-1	IND	LAS	1591	7	7	0		0	NA	NA	NA	NA	NA
70	2	2	IND	MCI	451	6	19	0		0	NA	NA	NA	NA	NA
70	-16	0	IND	MCI	451	3	6	0		0	NA	NA	NA	NA	NA

All the airport codes and Carrier data will be replaced with the new numbers, preserving the nature of data, but masking the original names or IDs.

The **legend** for the 29 features is:

	Name	**Description**
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes

16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes