Assignment-3

W.Raja Vikram Bhatt
Sr.No.15289

October 31, 2017

1. In 2007.csv, 20 distinct carriers are present(with unique carrier name).

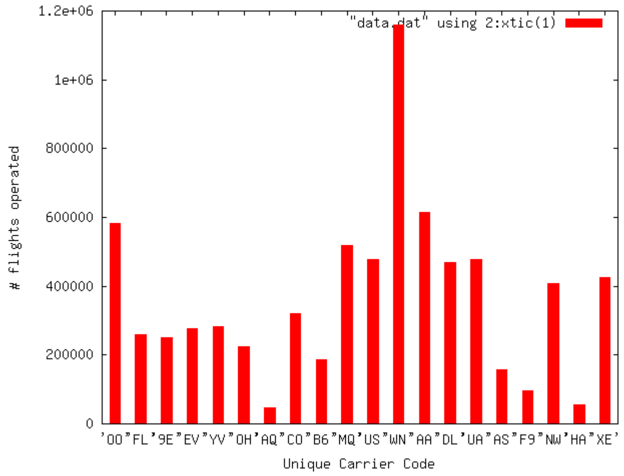
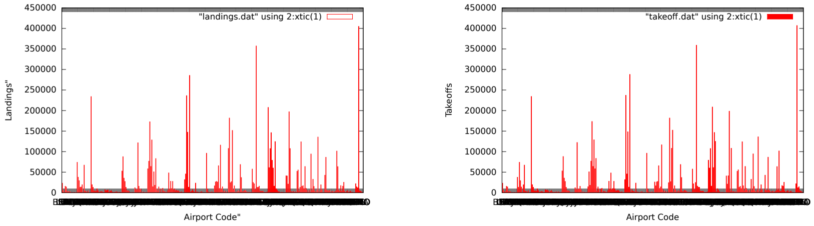


Figure 1: Flights operated in 2007

2. Fig.1 displays number of flights operated annually(in 2007) by each carriers.

3. 310 distinct airports(2007.csv)

4. a3_1_4.py generates two output files landings.dat and takeoffs.dat, which contains airport code and #flights operated.Figure.2 shows landings and takeoff at different airports in year 2007.



(a) Landings

(b) Takeoffs

Figure 2: Landings and Takeoffs at different airports in 2007

5. 14962 flights cancelled on 11-Sep-2011.

Table 1: Top 10 airports with most average delay

Average Delay(in minutes)	Airport Code
76.03	ACK
71.20	SOP
55.85	PIR
51.27	MCN
50.70	HHH
49.04	GNV
43.86	MEI
41.47	ACY
40.74	EWR
40.73	AGS

6. Above Table.1 shows top 10 airports with average delay.

7. Below Figure.3 shows delays variation between 1999 and 2002.The delays seem to be decreased between after 2001, but the delays are high in 2001 maybe due to the aftermath of 9/11 incident.

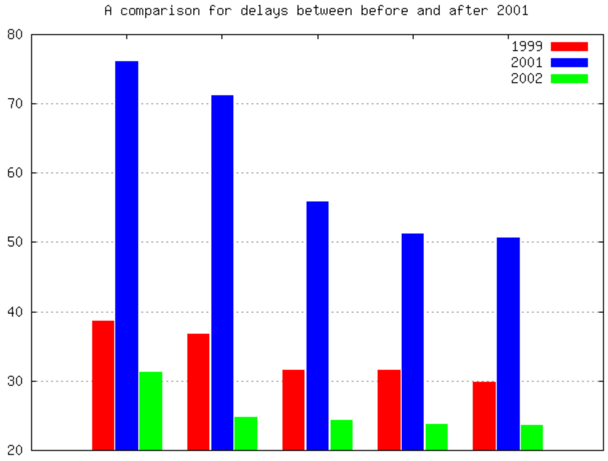


Figure 3: Top 5 delays in 1999,2001 and 2002

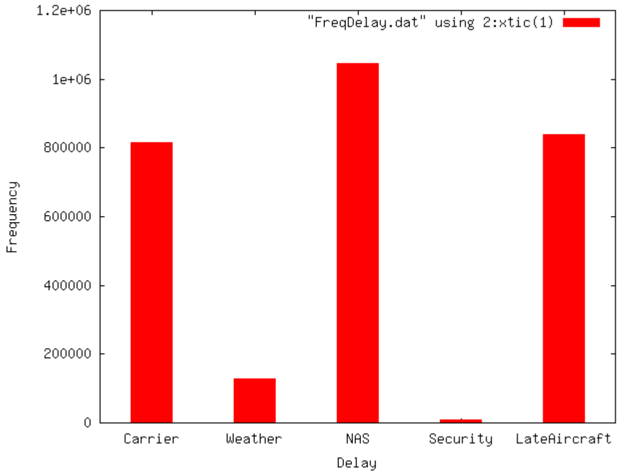


Figure 4: Frequency vs Reason for delay in 2007

8. NASdelay is the most frequent cause for delay.(Figure.4)

Table 2: Top 5 routes with longest average distance

Distance in miles	Route(Origin-Destination)
4962.0	EWR-HNL
4502.0	ATL-HNL
4433.0	CVG-HNL
4431.0	OGG-ATL
4243.0	HNL-ORD

9. Table.2 shows the result for top 5 routes with average distance(2007).
10. In this problem, I have used average speed as metric for fastest route.

$$\text{Average Speed} = \frac{\text{Distance}}{\text{Airtime}}$$

I have taken average over which has same origin-destination(aggregateByKey) and the result shown in Table.3.

Table 3: Top 5 routes aircraft fly fastest on average

Average Speed miles/minute	Route(Origin-Destination)
30.28	TYS-JAN
18.68	AUS-TUS
18.66	IAD-LGB
18.41	FLL-LGB
18.37	JFK-TUS

11. Edge list file has been stored in a3_4_1.out file.

12. Edge list file generated in 4_1, has been used to generate .dot file. I have used **pygraphviz library** on my local machine to convert the edge list file to directed graph data structure.I have also assumed A− >B and B− >A are different routes.Below Figure.5 shows the network graph generated by using 2007.csv, with the following options:

    ```
    twopi -Goverlap=scale -Ncolor=green -Nfontname=courier -Nfontsize=18
    -Estyle=dotted -Ecolor=grey -Efontcolor=blue -Teps graph.dot -o fig.eps
    ```

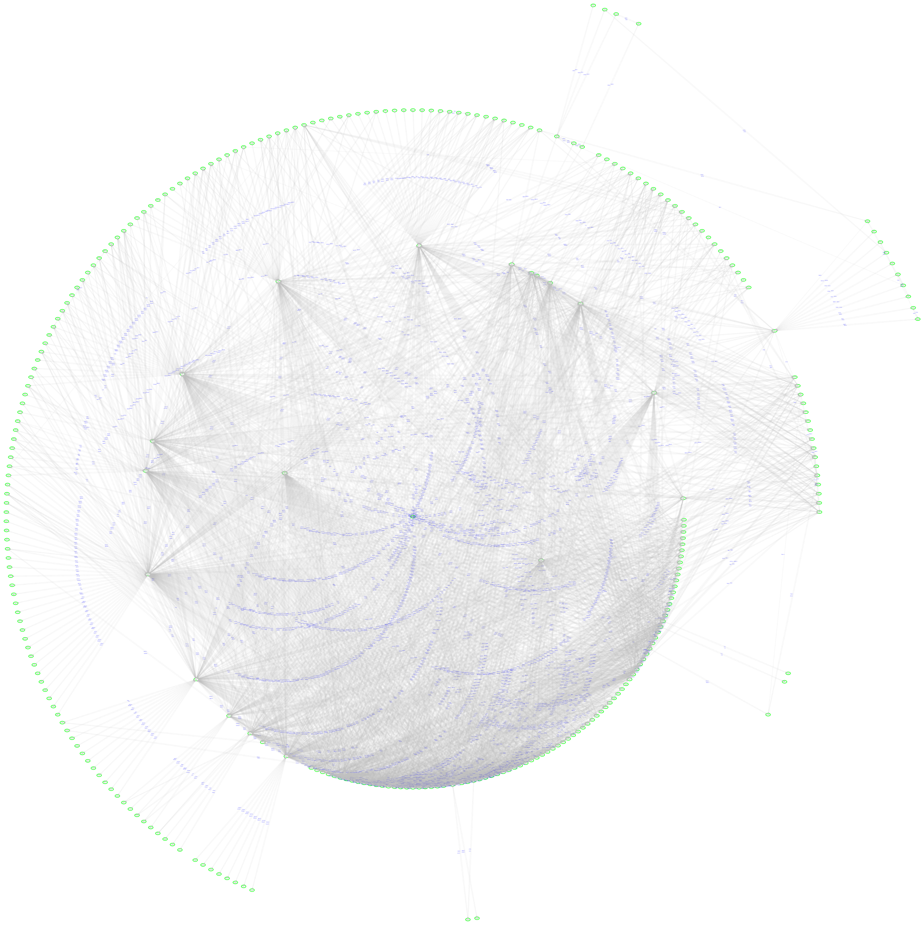


Figure 5: Network graph generated for 2007.csv

Bibliography

[1] Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia-Learning Spark,**Lightning-Fast Big Data Analysis** O'Reilly Media (2015)

[2] Tomasz Drabas, Denny Lee **Learning PySpark**,Packt Publishing (2017)