Study of Average Departure Delays of US Domestic flights by Airlines: Principal Component analysis

This study investigates the different airport’s and airline’s characteristics like number of departures & arrivals, arrival & departure delays, average taxi out & taxi in times, number of flight cancellations, number of diverted flights, variation in latitude & longitude and distance travelled, in determining which airlines are most likely to have a delay. The data is of domestic US airlines. Specifically, this study gives a closer look of all these airline related characteristics that helps determine most troubled airlines in terms of delays.

## Problem Statement

**Perform principal component analysis to predict which airlines are most likely to have the most average departure delays amongst US domestic flights/airlines based on an observed set of explanatory variables.**

## Constraints and Limitations

The biggest constraints on this observational study come from the fact that the data are only for the month of January from year 2015, so no causal inferences can be made about the relationship between explanatory and response variables. It is possible this data does not include all the explanatory variables that more appropriately describe the delays and all airport characteristics.

This model would be more accurate if we could use any time dependent features like weather API, daily max wind speeds, and precipitation but since weather and mechanical failure delays are unpredictable, we just analyze the model based on data of past flights.

## Data Set Description

The data was collected by the Department of Transportation Bureau of Transportation Statistics of all domestic flights during January 1st to March 10th, 2015. We acquired the data from Kaggle ( <https://www.kaggle.com/usdot/flight-delays>

The data set contained three separate csv files:

1. The airlines.csv detailed the abbreviations of IATA\_code used for airlines.
2. The airports.csv detailed name and descriptions of airports.
3. The flights.csv detailed arrival and departure delays, their time to taxi, wheels on and off times, cancellations and diversions.

The flights dataset contained 470K rows with individual entries for all flights, by reducing it to the month of January. We performed data cleaning to significantly reduced the dimensionality while maintaining much of the explained variance by taking averages and totals wherever needed as well as by working with grouped structure so that each week’s summaries are shown per airline. These data set contained variables that could be used to predict the Airlines most likely to suffer flight delays. Total variable set after merging the 3 files above are as below:

1. YEAR: Year of the flight
2. MONTH: Month of the flight
3. DAY: Day of the flight
4. DAY\_OF\_WEEK
5. AIRLINE : Airline the flight belongs to
6. FLIGHT\_NUMBER
7. TAIL\_NUMBER
8. ORIGIN\_AIRPORT
9. DESTINATION\_AIRPORT
10. ORG\_LAT: Origin airports latitude
11. ORG\_LONG: Origin airports longitude
12. DEST\_LAT: Destination airports latitude
13. DEST\_LONG: Destination airports longitude
14. SCHEDULED\_DEPARTURE: Scheduled departure in hours and mins (4 digit numeric)
15. DEPARTURE\_TIME: Actual departure in hours and mins (4 digit numeric)
16. DEPARTURE\_DELAY: Departure delay in mins
17. TAXI\_OUT: Amount of time spend taxying before taking off in mins
18. WHEELS\_OFF: Exact time of the flight taking off (wheels off the ground)
19. SCHEDULED\_TIME: Schedule time in mins of flight
20. ELAPSED\_TIME: Actual time in mins of flight
21. AIR\_TIME: Time spent in air in mins
22. DISTANCE: Distance between origin and destination airport
23. WHEELS\_ON: Exact time of the flight landing (wheels on the ground)
24. TAXI\_IN : Amount of time spend taxying after landing in mins
25. SCHEDULED\_ARRIVAL
26. ARRIVAL\_TIME
27. ARRIVAL\_DELAY
28. DIVERTED: 1 if flight diverted and 0 if not
29. CANCELLED: 1 if flight cancelled and 0 if not

Taking the airlines data set as reference for name and IATA code and merging the airports and flights data set for longitude and latitude information of airports and further taking averages and percentages of certain variables, we came up with our final data set that had a total 16 variables (excluding the ones we thought were not relevant) and was 70 rows.

**Final variables in the data set are:** Airline, Day(grouped by week per airline), Departures, Arrivals, D\_Ontime(Percentage of departures on time), A\_Ontime(Percentage of arrivals on time), Arrival\_Delay(Avg arrival delay), Departure\_Delay(Avg of departure delay), Taxi\_Out(avg of taxi out), Cancelled, Diverted, Taxi\_in, LAT(variation in latitudes for all flights run by an airline in that month), LONG(variation in latitudes for all flights run by an airline in that month), Wheels\_Off(avg wheels off), Distance.

## Exploratory Data Analysis

The data might seem straightforward as it has information on percentage of on time departures. However, there is more to this analysis. The frequency of having late flights for an airline and how late the flights are departing are being considered as well. These two factors individually might not make sense but when combined would have a Pareto affect (80/20 rule), where 20% of the flights can contribute to majority of the delays.

As the focus is on delays by airlines, it is important to understand what constitutes a delay. Any flight departing by >5 mins delay is considered late or to have a departure delay in our analysis. Same rule applies to arrival delays as well.

**Distribution of Departure Delay and Arrival Delay:** Average departure delay for flights is between 2-17 minutes (Table 1) and average arrival delay for flights is between 2-14 minutes (Table 2). It can be assumed that the arrival delay is caused due to the departure delay and probably increased air time or taxi\_in at arrival.

|  |  |
| --- | --- |
| **Table 1** | **Table 2** |
|  |  |

**Histogram 1**  **Histogram 2**

**Distribution of percent on-time departures and arrivals:** It is evident from the below histograms (Histograms 3 and 4) that most flights are arriving and departing on time.

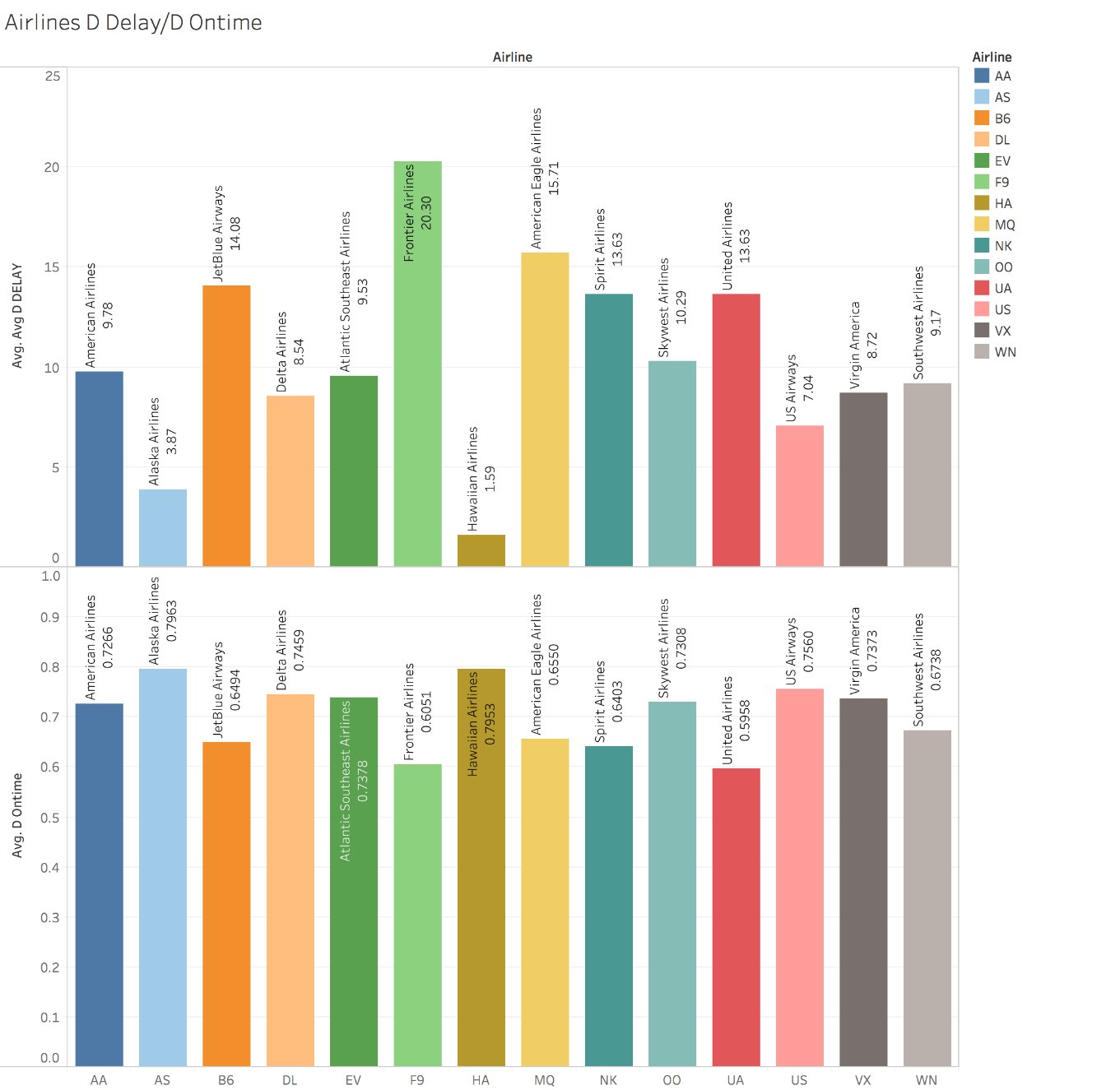
|  |  |
| --- | --- |
| Departure on time Around 74 percent | Arrival on time around 70 percent |
| **Histogram 3** | **Histogram 4** |

**Distribution of total arrivals and departures:** The arrival and departure data appear to be heavily right skewed. Most airlines seem to schedule less than 24K flights per month (Histograms 5 and 6).

|  |  |
| --- | --- |
| Most airlines have less than 5 K flights departing | Most airlines have less than 5 K flights arriving |
| **Hisogram 5** | **Histogram 6** |

**Percent Departures on time vs. Avg Delay times per airline:**

The below bar graph (Graph 1) clearly demonstrates that there is no clear relationship between the percentage of on time departures and average delay time for a given airline. F9, MQ, B6 seems to have the longest delay times. In contrast, HA and AS have the shortest average departure delay.



**Graph 1**

**Flight frequencies by airline:**

Even though F9, MQ and B6 have the longest delays they don’t seem to be the busiest in terms of operating number of flights in a month as shown below in Graph 2:



**Graph 2**

## Correlations between explanatory variables

Looking at the initial correlation matrix (Table 3) and scatter plot matrix (Table 4), we see a high degree of correlation between many explanatory variables. There is almost a perfect correlation between departures and arrivals and many of the characteristics are derivatives of each other, so this data set is a good candidate of PCA analysis. This can be perfect example for PCA to reduce the multicollinearity issues and reduce the number of predictors.

A screenshot of a cell phone

Description generated with high confidence

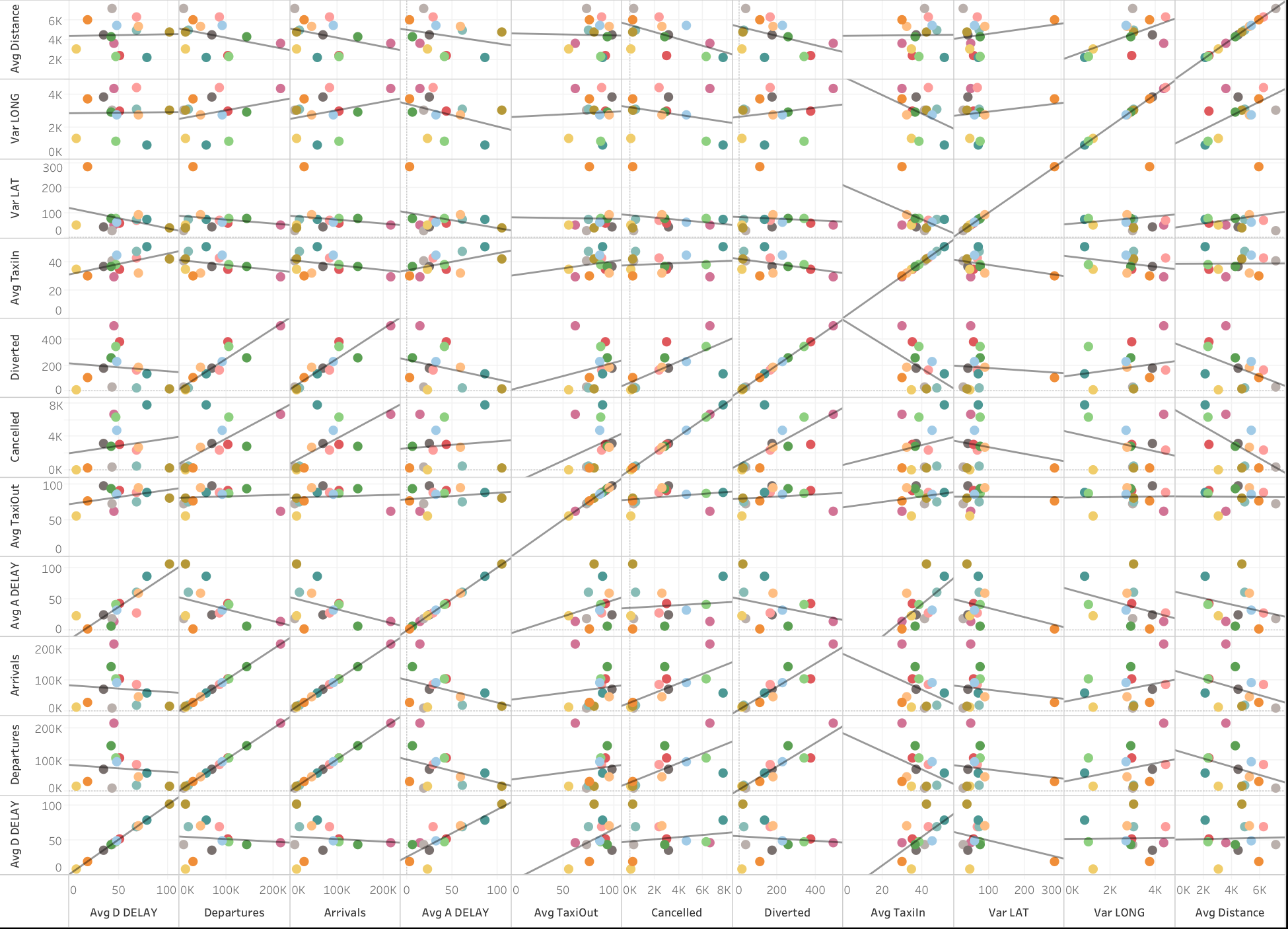
**Table 3**

A picture containing text

Description generated with high confidence

**Table 4**

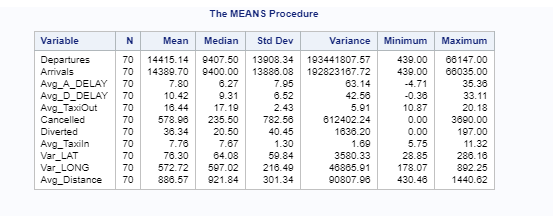
Looking at the matrix plot below (Graph 3), the average departure delays are strongly correlated with average arrival delays, as well as taxi out and taxi in. This makes sense, because if a plane is spending more time taxying before reaching the gate or after leaving the gate(departures) it effects the departure and arrival delay times. Interesting observation is that latitude is negatively correlated to the delays. We might assume that flights traveling longer distances in North – South directions have lesser delays.



**Graph 3**

## Principal Component Analysis

We are now able to move forward with the PCA analysis. PCA is all about determining which linear combinations of original factors also called as principal components explain the largest possible variance. So measuring the features on same scale is equally important. Upon examination of summary statistics (Table 6), Arrival and Departure variables show large variance. So, it would be wise to run PCA on correlation matrix because variables do not have same units and would be heavily influenced by variables with large variance.



**Table 6**

By examining the eigen value matrix (Table 7) and the Scree Plot (Graph 4), it has been decided to select the first three principal components as these account for almost 73% of the variance in data.

A screenshot of a cell phone

Description generated with very high confidence

**Table 7**

A close up of a map

Description generated with very high confidence

**Graph 4**

**Examining principle components and their associated original factors:**

PC1: The first principal component seems to be composed of the factors reflecting how busy an airport/airline activity is. It weights number of departures, arrivals, diversions, and cancellations relatively equally, as well as factors related to taxiing (Table 8). An airline high on PC1 likely has longer delays at any point in an airplane’s journey.

PC2: The second principal component negatively weights departures and arrivals and positively on average arrival delay (Table 8). An airline high on PC2 is more likely to have lesser number of departures and arrivals but longer delays.

PC3: The third principal component positively weights longitude, distance and departure delay (Table 8). An airline high in PC3 is more likely to cover airports over wider distances over longitudes. This component strongly accounts for longitude so an airline high in PC3 is more likely to be covering distances horizontally (east to west).

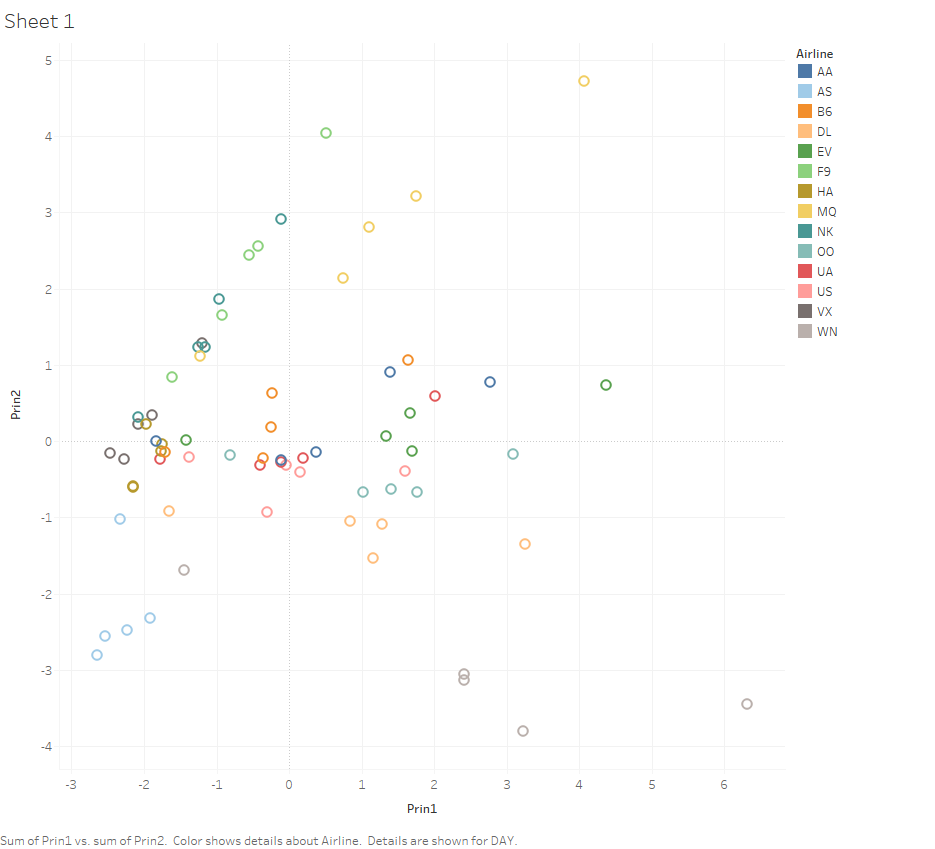
A picture containing text

Description generated with high confidence

**Table 8**

**Examining airlines against PC1 and PC2**

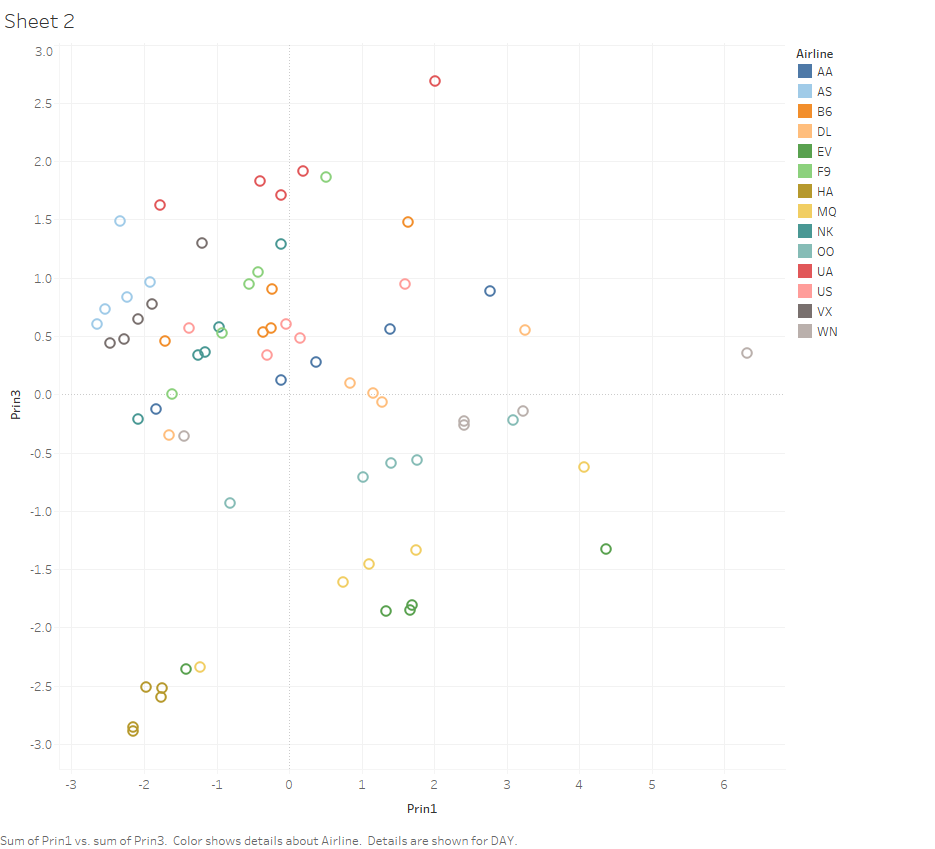
As seen in graph 4, WN (Southwest Airlines) has highest PC1 and lowest PC2. Which means Southwest airlines is most likely to have larger delays, however with highest frequency of flights due to negative correlation on PC2. It can also be observed in *graph 4* that southwest airlines in fact has the highest frequency of flights.



**Graph 4**

**Examining airlines against PC1 and PC3**

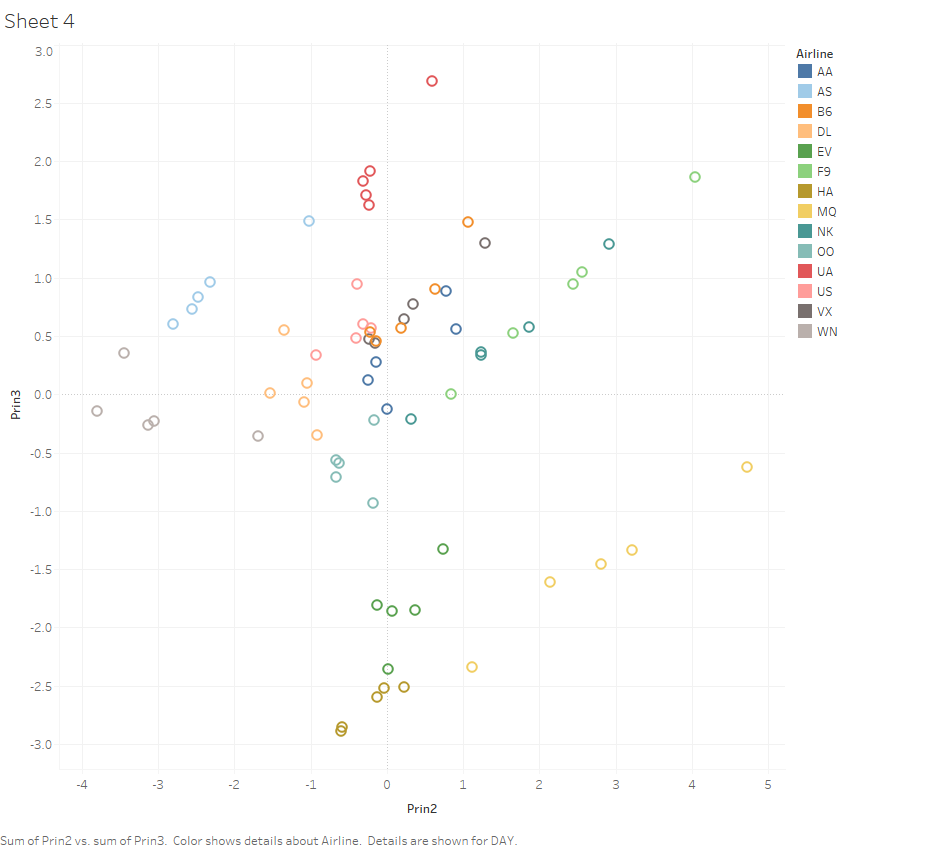
As seen in Graph 5*,* UA(United Airlines) has the highest PC3 and is moderate on PC1. Therefore, US airways is likely to have highest coverage of flights from east to west while the delays are moderate.



**Graph 5**

**Examining airlines against PC1 and PC3**

As seen in Graph 6UA(United airlines) and AS(Alaska Airlines) seems to have higher PC3. Therefore, they are likely to have more coverage from east to west. However, Alaska airlines is lower on PC2, therefore has higher frequency.



**Graph 6**

## Conclusions

We tried to analyze the primary components based on statistics of past flight and with the features that are available at any time such as airline, month, departure and arrival time, taxi time. There may be many other unknown and unpredictable factors like weather which causes delay, so our prediction is limited. Despite this, we are successful in performing principal component analysis to predict which airlines are most likely to have the most average departure delays. Since this is an observational study, the results apply to only those airports and airlines observed and does not indicate any level of Causality.

There were a couple of clear lessons from the study which can be included but further work on collecting other informative features about flights like size of the plane, direction of flights, no of passengers on the flights and work on sophisticated modelling techniques could be the possible next steps and could be a topic for future research.

APPENDIX:

FILENAME REFFILE '/home/skc00/Semester II/Statistics II,6372/Week10 project 2/Flights.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

Data flightdelay;

set work.import;

run;

proc print data= flightdelay;

run;

Basic PCA

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ods rtf ;

ods graphics on;

proc means data =flightdelay n mean median std var min max maxdec=2;

var Departures Arrivals Avg\_A\_DELAY Avg\_D\_DELAY Avg\_TaxiOut Cancelled Diverted Avg\_TaxiIn Var\_LAT Var\_LONG Avg\_Distance ;

run;

proc corr data=flightdelay plots=matrix(histogram);

var Avg\_D\_DELAY Departures Arrivals Avg\_A\_DELAY Avg\_TaxiOut Cancelled Diverted Avg\_TaxiIn Var\_LAT Var\_LONG Avg\_Distance;

run;

\*with avg dep delay as response variable ;

proc princomp plots=all data=flightdelay out=pca;

var Departures Arrivals Avg\_A\_DELAY Avg\_TaxiOut Cancelled Diverted Avg\_TaxiIn Var\_LAT Var\_LONG Avg\_Distance;

run;

proc gplot data=pca;

plot prin2\*prin1=Airline;

run;

proc gplot data=pca;

plot prin3\*prin2=Airline;

run;

proc gplot data=pca;

plot prin3\*prin1=Airline;

run;

ods graphics off;

ods rtf close;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

PCA Regression

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc corr data=pca plots=matrix(histogram);

var Avg\_D\_DELAY prin1 - prin3;

run;

\* without Cancelled Diverted;

run;

proc reg data=pca;

model Avg\_D\_DELAY = prin1-prin3;

run;