School of Computer Science and Engineering

Fall Semester 2020-21

J Component Report

on

Flight Delay Prediction

submitted by

Varshitha Chennamsetti - 17BCE2208
A S Advaith - 17BCE0104
Sai Nikhil Prathipati - 17BCE2323
Rajesh K - 17BCI0147

submitted to
Prof. Ramani S

for

CSE3020: Data Visualisation

October 2020

1.Introduction

Delay is one of the most important performance metrics for any transportation system. Especially, commercial airlines perceive delay by the amount by which a flight is late or postponed. A delay can be shown as the difference between scheduled time and time of arrival/departure. Aviation controller authorities monitor all the thresholds for flight delay . Therefore flight delay is essential with respect to air transportation. Flight delays have bad impacts, especially for passengers, airlines, and airports. Given the uncertainty in their occurrence, passengers typically plan to tour many hours in advance, increasing their journey costs, to ensure their arrival on time. Also, airlines suffer through heavy fines, penalties and extra operation costs, including crew and aircrafts retention in airports. From the environmental sustainability point of view, delays might also cause environmental harm by increasing gasoline intake and gas emissions.

Delays additionally jeopardise airlines advertising and marketing strategies, considering companies depend upon customer's loyalty towards their frequent-flyer programs likewise customer's choice for travelling through the same airline will also be affected. The estimation of flight delay improves the operational decisions of airports, airport authorities and warns passengers to rearrange their plans accordingly. Their prediction is critical during the decision-making procedure for all commercial aviation airlines.

Commercial aviation is a complicated transportation system. Furthermore, a passenger follows their itinerary and plans their travel accordingly, whereas airlines have to plan numerous schedules in the background for the operation of an airline.

2.Identification and Exact Definition of the problem

Airlines have schedules that need to be followed by Pilots, flight attendants and aircrafts because of legal rests, duties, and maintenance plans for airplanes. So, any minor disruption within the system can affect the following flights. Moreover, disturbances may also create congestion at airspace, or different airports, creating queues in the taxiway and delaying other flights from different airlines. To develop the accurate prediction model can be difficult because of the complexity in the aviation system, and the overwhelming amount of data.
We can classify the delay by :

- **Delay propagation**
- **Root delay and Cancellation**

The delay propagation accounts for the interactions among flights by propagating any significant delays that are incurred at an airport to "downstream" flights and airports. In contrast , it is also important to predict further delays and understand their causes, keeping in mind that new problems might eventually occur and can be called a root delay problem. In some situations, delays can also lead to cancellations and forcing airlines and passengers to reschedule their travel.

2.1.Brief background of the topic

There are many things to consider when dealing with a flight transportation system. It has many complicated relations with the destinations, origins, the timings and the passengers. Many characteristics will be hence responsible for the flight's journey. Other than the above features, there are some other facts that can determine how the flight's journey would be like. Some of them are the weather conditions, how the air traffic is like, if there are any mechanical problems or if there is some problem with the amount of capacity that the airline has etc. These can determine how good or bad the flight journey will be like. Because as a flight customer these factors can count as a great deal when considering to fly again on the same airline again. Especially, when it comes to delays, the customer satisfaction is at stake. These small disturbances can also in counter disturb the schedules of other flights. It will taint the reputation of an airport too if there are many delays in that airport. Traditionally, the flight delay problem has been studied in two different ways. One way of studying the problem talks about the time the delay persists. While the other talks about the reasons for the delay of the airline and the cancellations of the airlines from the customer point of view. There have been many methods researched so far. Some of them are given below.

2.2.Literature Survey:

This Paper shows how delay propagation can be reduced by redistributing existing slack in the planning process, making minor modifications to the flight schedule while leaving the original fleeting and crew scheduling decisions unchanged. Computational results based on data from a major U.S. carrier are presented that show that significant improvements in operational performance can be achieved without increasing planned costs[1].

The aim of this paper is to analyze how selected data mining techniques can be integrated into entities such as airlines, airport retails sector and the airport itself is considered for this cause and the data mining techniques that can be applied to these entities to improve the current airport systems such as flight delay prediction, passenger profiling, segmentation, association rule mining are discussed to find better approaches for an intelligent airport system[2].

The aim of the paper is to provide a model that reduces the delay caused by the crew and to increase the efficiency. In this paper they use an integrated stochastic model for aircraft and crew scheduling. The model is solved by a heuristic solution approach based on column generation. Using this model delays propagated by crews and aircraft can be reduced significantly[3].

Statistical Analysis covers the use of regression models, correlation analysis, econometric models, parametric tests, non-parametric tests, and multivariate analysis. Delay multiplier and Recursive models help airlines predict delay and effects caused due to it through the network and helps us to estimate the cost of it[4].

Other economical models were created to analyse the efficiency of flight systems such as analysis of investment by the government agency or to predict the relationship between delays and passenger demand, fares, frequency and size of the plane[5].

Xiong et al built an econometric model based on pre-existing delays, potential delay savings, distance, characteristics of the destination airport and airline, frequency, aircraft size, occupancy rate and fare to understand which reasons lead airlines to cancel their flights[6].

The study was conducted about flight delay rate and the impact of aircraft turnaround time during maintenance check. This resulted in building a model which shows how the delays originated in certain places are affecting other airports[7].

Pathomsiri et al. used a non-parametric function to evaluate the efficiency of airports of the United States regarding delays. Reynolds et al. computed the correlation between levels of delays and capacities of the European airports[8,9].

They have calculated the daily average of delays to understand the principle causes of delays[10].

There have been usage of these models to forecast the delays and also give the cause of the delays. Expectation-maximization method was used in order to find the reasons for the departure delays. A prediction system model was built which considers many factors such as the type of season and the effects it has on the departure delay. Because of the obvious problem i.e finding the global extremes, they have adapted some of the genetic algorithms for the optimisation. They also consider the fact that all flights are connected and that one delay in one flight could very easily affect the other. So they consider the daily pattern of the schedules too. The model thus built shows robustness and fitness[11].

A probability function was also used to predict if there will be any cancellations due to the delays. This was done using a conditional probability by considering the assumption that the previous flight might have been delayed. The delays were predicted using a four parameter downstream model. They find the cost of it using the parameters of the customer[12].

Density functions such as Normal and Poisson distribution were also used. But the usage of them was concluded to be dependent on the type of data under consideration. They considered both the arrival as well as the departure delays. The Normal distribution fit better using the Normal distribution whereas the flight propagation period and the arrival delays were much better fitted using the Poisson distribution[13].

Flight delays in the U.S. domestic system is analyzed by estimating an econometric model of average daily delay that incorporates the effects of arrival queuing, convective weather, terminal weather conditions, seasonal effects, and secular effects (trends in delays not accounted for by other variables). From their estimation it was possible to quantify some sources of higher delays in late 2003 and early 2004 and track changes in delays that are not attributable to major causal factors.Their results suggest that when these factors are controlled for, delays decreased steadily from 2000 through early 2003 [14].

This paper examines the impact of hub and spoke congestion on flight times and departure delays in the US airline market. It measures the change in flight times resulting from infrastructure-constant changes in passenger demand. In addition, it quantifies the difference in flight times between airlines with large and small networks [15].

3. Data collection and Methodology

3.1Dataset

The dataset has been taken from a reliable online available government agency website that provides the air traffic delay statistics in the United States. The U.S. Department of Transportation's (DOT)

Attributes	Descriptions of Attributes
YEAR, MONTH, DAY, DAY_OF_WEEK	dates of the flight
AIRLINES	It is the IATA Code to identify unique airlines
ORIGIN_AIRPORT and DESTINATION_AIRPORT	Code attributed by IATA to identify the airports
SCHEDULED_DEPARTUR E and SCHEDULED_ARRIVAL	scheduled times of take-off and landing
DEPARTURE_TIME and ARRIVAL_TIME	real times at which take-off and landing took place
DEPARTURE_DELAY and ARRIVAL_DELAY	difference (in minutes) between planned and real times
DISTANCE	

Fig. 3.1 DESCRIPTION OF THE ATTRIBUTES INVOLVED IN THE DATASET.

Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. BTS compiles daily data for the benefit of the customers or for any data analysts. The dataset is of 2017 flight delays and cancellations.

3.2 Data Exploration

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways
5	OO	Skywest Airlines Inc.
6	AS	Alaska Airlines Inc.
7	NK	Spirit Air Lines
8	WN	Southwest Airlines Co.
9	DL	Delta Air Lines Inc.
10	EV	Atlantic Southeast Airlines
11	HA	Hawaiian Airlines Inc.
12	MQ	American Eagle Airlines Inc.
13	VX	Virgin America

Fig. 3.2 All the airlines in the dataset associated with particular IATA carrier codes.

Data cleaning is the critical initial step in evaluating the dataset for final analysis. With the enormous amount of data available, databases are prone to have noisy, missing and inconsistent data. The data in this project is obtained from BTS source, which has varying kinds of 31 variables involved, and may not be compatible with the format in which we require the data to use in Python. Data Cleaning helps in removing noisy data, and removing inconsistencies. Data cleaning is performed as follows: Dates and Times: The date format has been given in four variables format; it will be toned down to one particular format available in Python for ease of use. Filling Factor: In the data cleaning process, a missing value can be ignored, manually entered, given a constant value, or a mean value. In this case, it will be organizing and arranging the entire data frame to keep the relevant attributes and eliminate the ones which have missing values. This is done to increase the readability and feasibility of use. The fill factor gives us what percentage of space on each page to fill with data. The fill factor value obtained, in general, can be defined as a percentage from 1 to 100. Here, it has obtained a fill factor of >97%, which is quite satisfactory, that means 3% of the overall space can be used for future data growth. Further, we have established a statistical description of airlines, which involves classifying airlines on the basis of their punctuality; it is done using various statistical parameters.

3.3 Methodology Adopted : One Airline-All Airports

it is considered advantageous to make a single fit, which would take all the airports into account. Particularly, this would allow predicting delays on airports for which the number of data is low with a better accuracy. Here, to test, it has been chosen as the carrier=”AA”, that is, American Airlines, and in the data frame, a label has been assigned to each airport. The correspondence between the label and the original identifier has been saved in a list in Python. The next step involves incorporating the "One Hot Encoding" Method. In machine learning, to work with categorical variables, the categorical data is converted into numbers, which is required for both input and output data that are categorical. This method is applied in this case by creating a matrix where instead of the ORIGIN_AIRPORT variable that contains M labels, we build a matrix with M columns, filled with 1 and 0 depending on the correspondence with particular airports.

Linear Regression is first performed on this model, and extreme or large delays are underestimated and not taken into account, as explained. The upcoming figure Linear fit on Model depicts this.

```
MSE = 53.7430736542

'5.30%'
```

Fig. 3.3:MSE value and quality % of linear regression on Model obtained in Python

In practice, the quality of fit is also known by considering the number of predictions where the differences with data points (or real values) are greater than 15 minutes.

(No. of values >15min / No. of predictions (total)) *100

The value found here is 5.30%.

Further, Polynomial Regression is performed on the fit,

Upcoming figure Linear fit on the Model depicts it.

```
MSE = 49.5025438214
```

```
'4.81%'
```

Fig. 3.4:MSE score and quality % of Polynomial fit in Model 2 obtained in Python

The MSE score found is 49.502543. The quality of the fit is again judged by the above formula, and is found 4.81%.

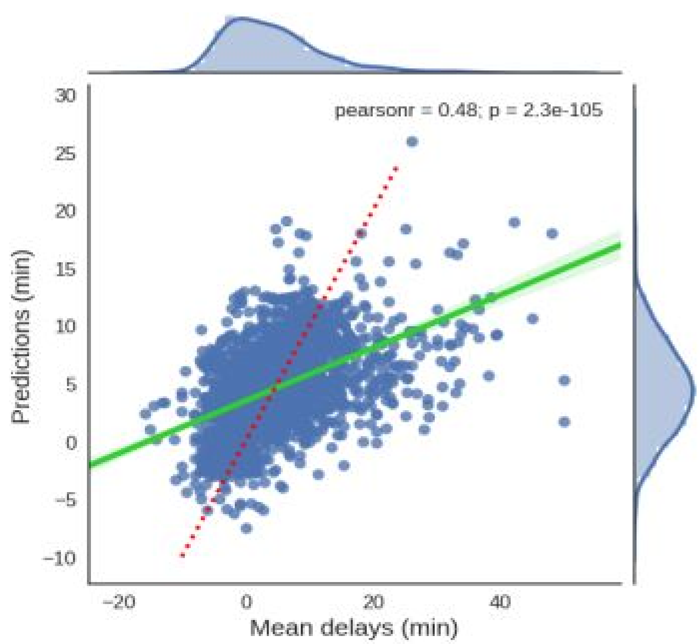

Fig. 3.5:Linear fit on Model .

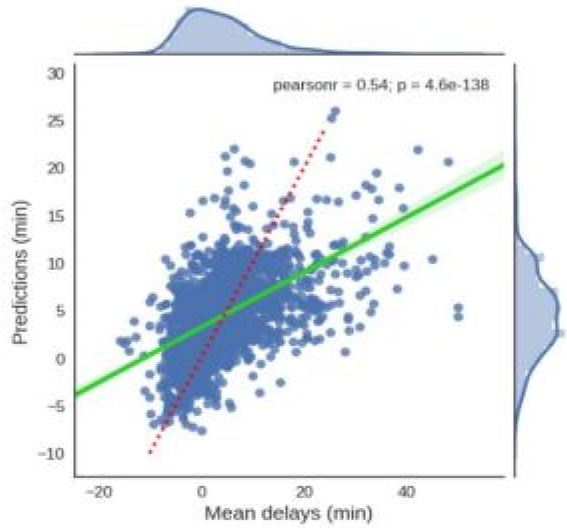

Fig. 3.6:Polynomial fit on Model .

Hence, it is evident that a polynomial fit improves the MSE score slightly, and is an efficient model. Testing the model against end-week data, using regularization to minimize the errors and over fitting:

```
'MSE = 59.36'

'Ecart = 7.70 min'
```

Fig. 3.7: MSE score of the final testing and quality percent obtained in Python.

The current MSE score is calculated on the basis of all the airports that are served by American Airlines, whereas previously it was calculated on the data of a single airport. The current model is therefore more generalized and efficient.

3.4 Experiments / Analytical Computations and tools used

Cross Validation Technique and K-Fold Technique Cross Validation is a very important technique for assessing the performance of machine learning models. It enables us in knowing how a machine learning model would generalize to an independent data set. The model dataset is divided into three sets: Training, test, and validation. The entire set is divided into K-folds or subsets, which is basically applying the K-fold technique, one of the ways of Cross Validation. Then, the K-1 folds are sent for training and the learning is done on it, then the model's generalization is checked on the test set, which contains just the remaining one fold; and this process goes on till the last fold. This method is used in the initial stages Model , for data splitting and increased efficiency.

MSE

The Mean Squared Error (MSE) is a measure of how close a fitted line is to the real data points. For every data point on the line, we take the distance vertically from the real point to the corresponding Y value on the curve fitted (which is the error), and square the value. The next step is to carry out the summation of all the squared error values corresponding to all the data points, and, in the case of a linear fit, the value we get is divided by the total number of observations minus 2. The squaring is to avoid negative values cancelling the positive values. The quality of the model is assessed by the Mean Squared Error score we get, the smaller the value, the closer the fit is to the real data and the accurate the machine learning model.

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

e_t = error value (predicted value-real value)
n=Total no. of attributes or points taken into account.
MSE value for Model , Linear Fit is (shown in Fig above), 53.7430 MSE
value for Model , Polynomial Fit (shown in Fig.above), 49.5025.

RMSE

Root Mean Squared Error (RMSE) is another quality that we calculate to measure the accuracy of a model. It is equal to the square root of the mean square error. It is considered as one of the most easily interpreted statistics, as it has the same units as the quantity plotted on the ordinate, which is the y-axis.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

et = error value(predicted value-real value)
n=Total no. of attributes or points taken into account.
The RMSE values are depicted using a variable Ecart, for both the models, in Fig of Linear regression and Fig of polynomial regression, Ecart value shown for Model is 4.81%. As the MSE value and RMSE value is lowest for the polynomial regression on Model , hence, it depicts that it shows the most accurate results, and the fitted line (the predicted results) is the closest to the real data points. Though, the final data set testing showed the Ecart value as **7.70 min**.

4.Results

A thorough data visualisation was done in order to find any new patterns or anomalies in the data. An interactive dashboard using Tableau was also constructed. The tableau dashboard created can be found online at:

https://public.tableau.com/profile/varshitha.chennamsetti#!/vizhome/Flights_16032703873650/Sheet1

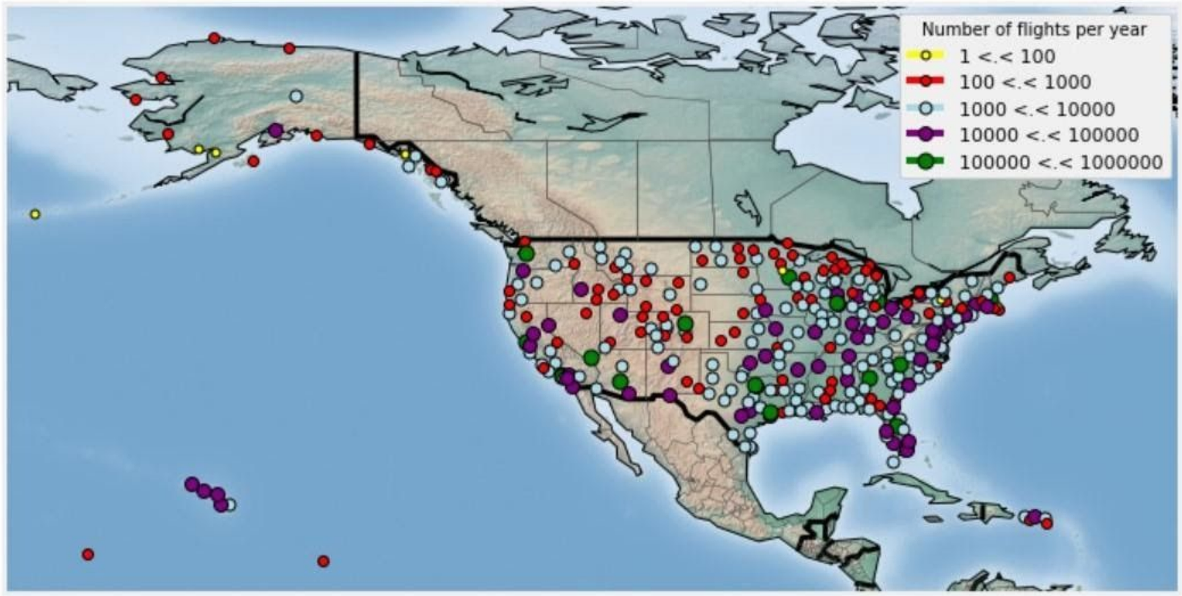

Fig.4.1: Origin airport and the flights flying from there per year

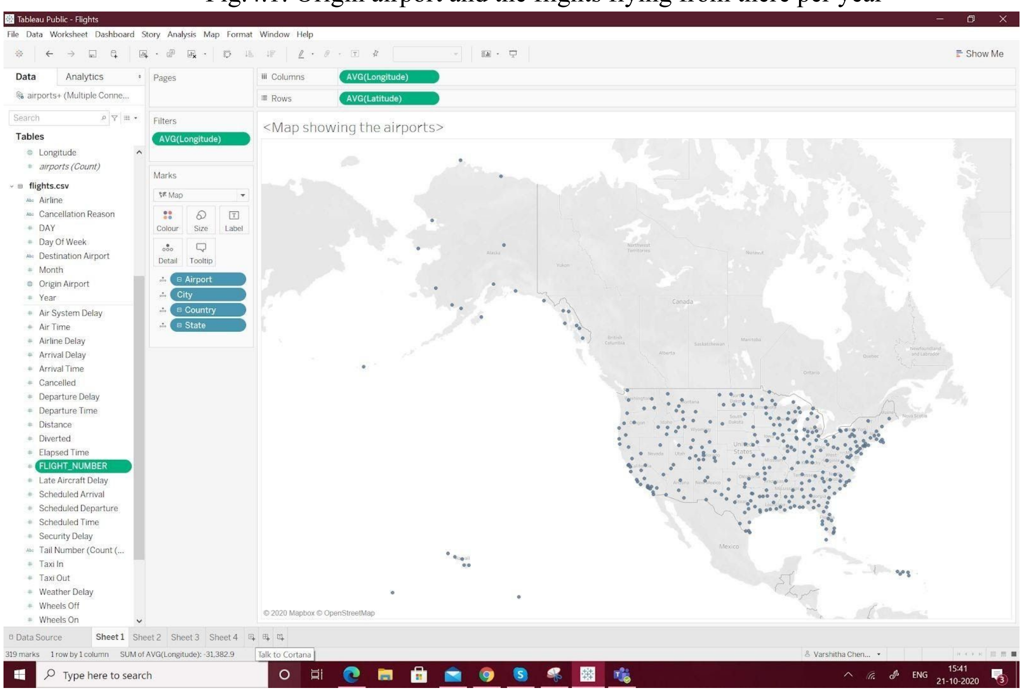

Fig 4.2: Map showing the origin airports using Tableau

The above graph shows the different places where airports of the flights in the data are flying through. The sizes and colours of it are based on the amount of flights that fly from that airport per year. From the graph it can be seen that most of the airports have the number of flights flying from there in the range 1000 to 10,000.

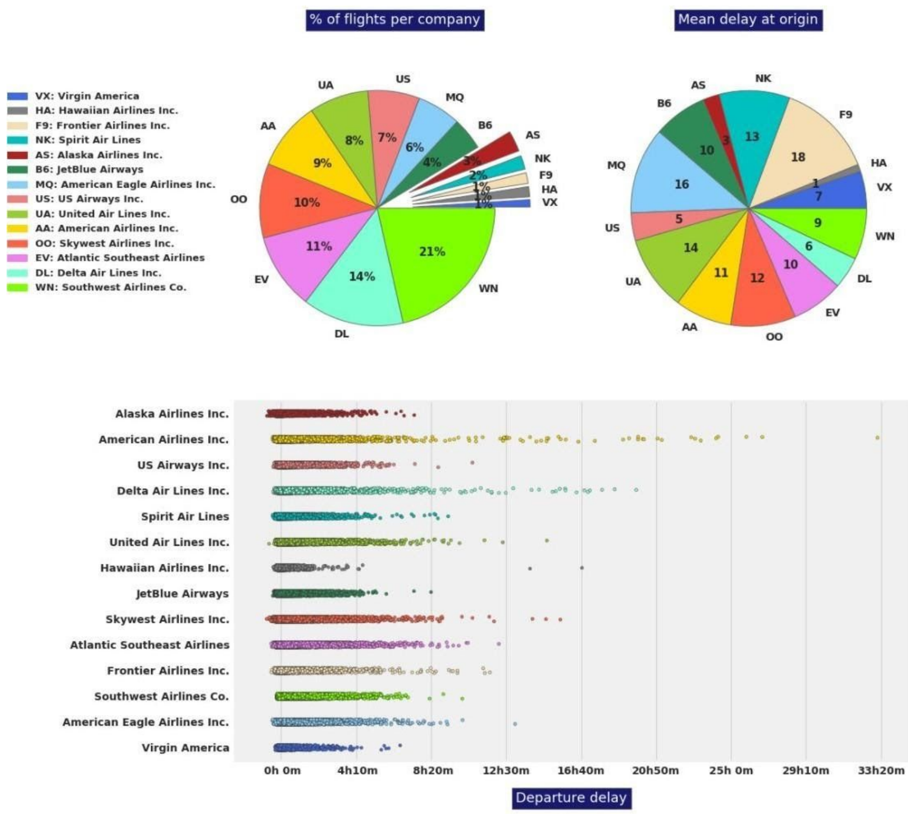

Fig 4.3: Flights and Mean delay per company

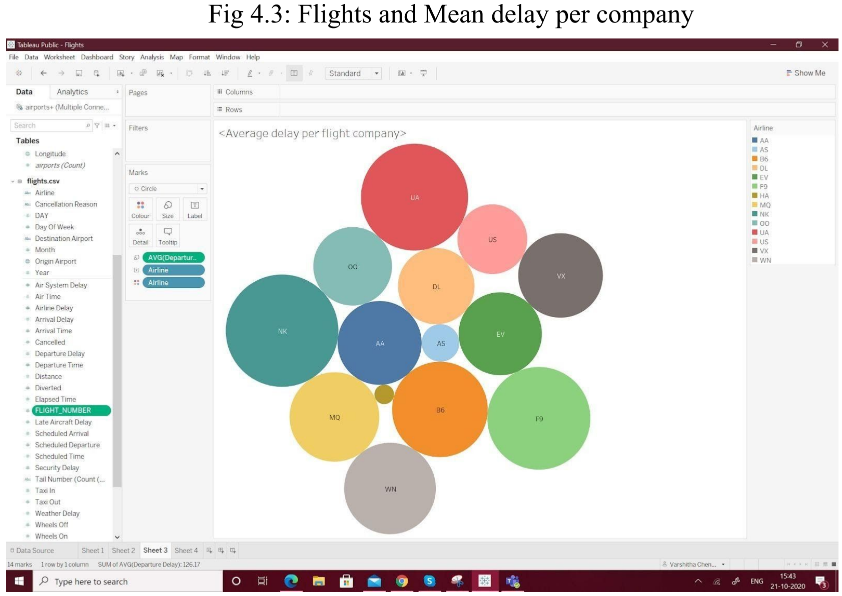

Fig 4.4: Mean delay per flight company using Tablea

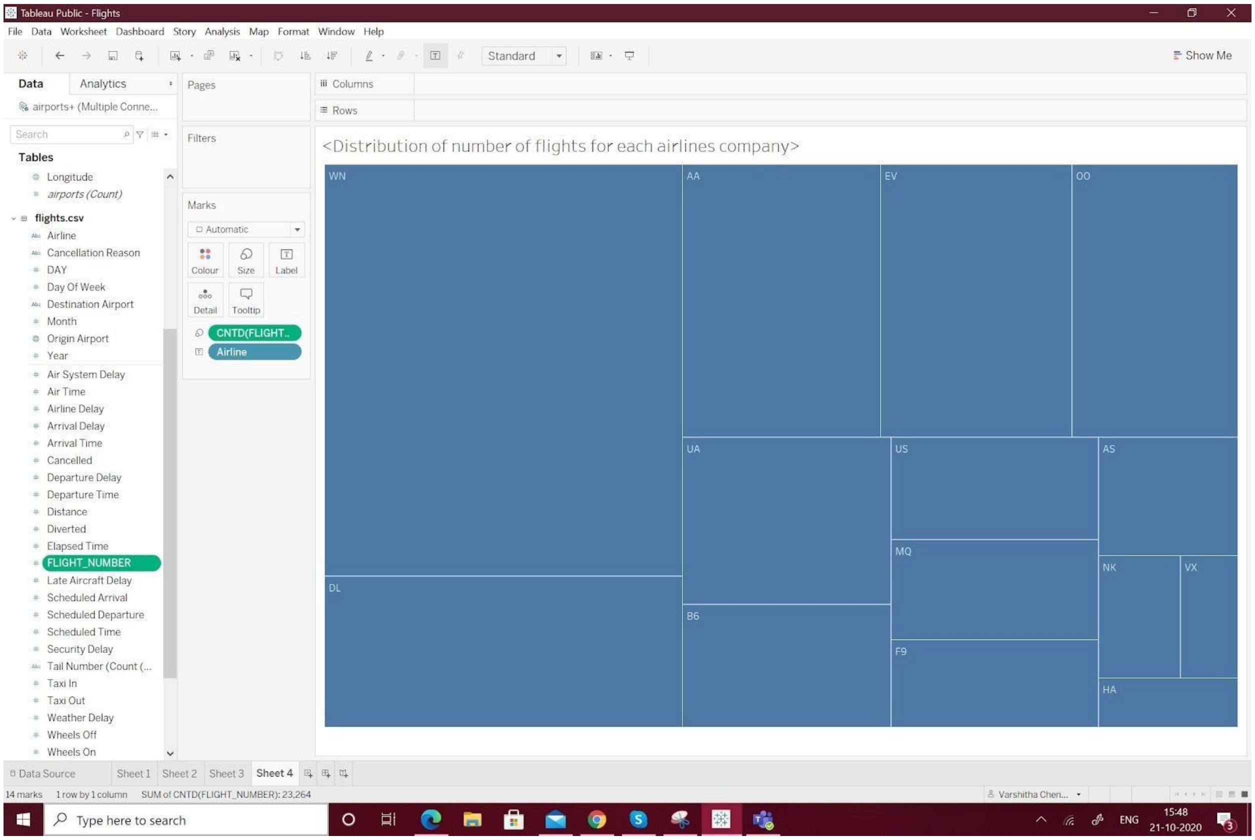

Fig 4.5: Tree map showing the Number of flights per each company

The above graph shows the number of graphs per company and the amount of delays at the origin. It can be seen that most of the airlines belong to the company called Southwest Airlines Inc. (WN) with 21%. There are also companies that have very few flights under them like the Virgin America airlines. The mean delay for each of the flights is around 11 minutes. This value is extremely low which means most of the flights do follow the schedule. But the distribution graph that is given below the two pie charts suggest that although most flights travel on their usual schedule, there are also few extremes that take upto almost a day of delay.

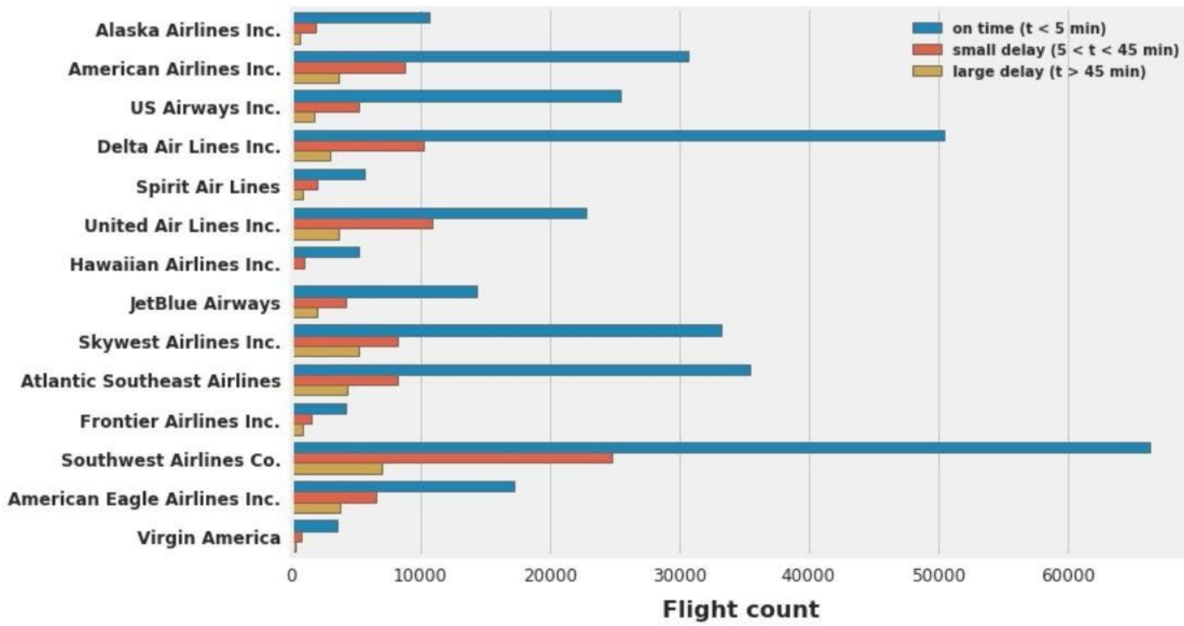

Fig 4.6: Time comparison for each flight company

The above distribution shows the number of flights of each company in terms of range of time of delay. It can be seen that most flights have a delay of less than 5 minutes but it also has a considerable amount of flights that have a delay in the range 5 minutes to 45 minutes. Like in the case of the Southwest Airlines Co., there are almost 25,000 flights that have a small delay and at least 7000 flights that have a large delay.

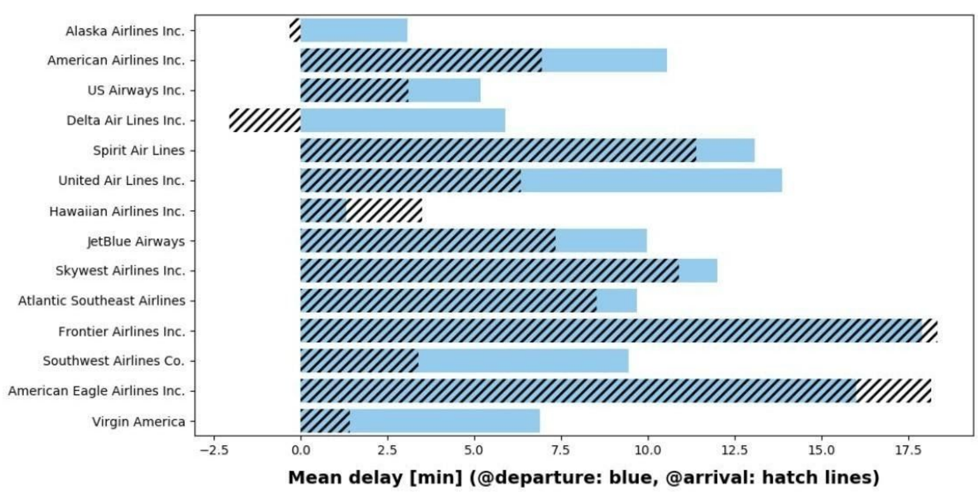

Fig 4.7: Comparison of departure and arrival delays

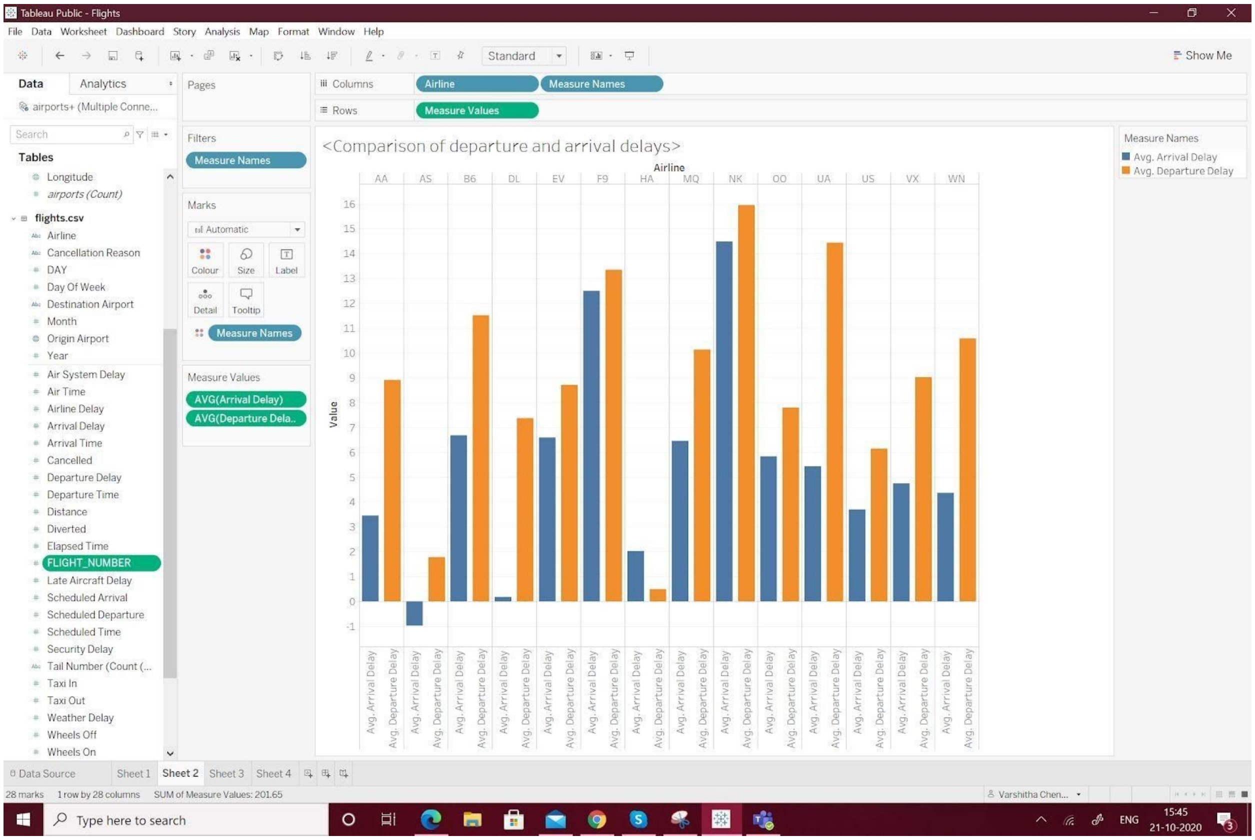

Fig 8: Comparison of arrival and departure delays on tableau

The above graph tells us whether there was an arrival or the departure delay. The hatched lines show the arrival delay whereas blue lines show the departure delay. It can be seen that there are almost no arrival delays. It is an indication that there are only departure delays as the arrival delays are mostly avoided because of the adjustment of speed of the airline based on the arrival time.

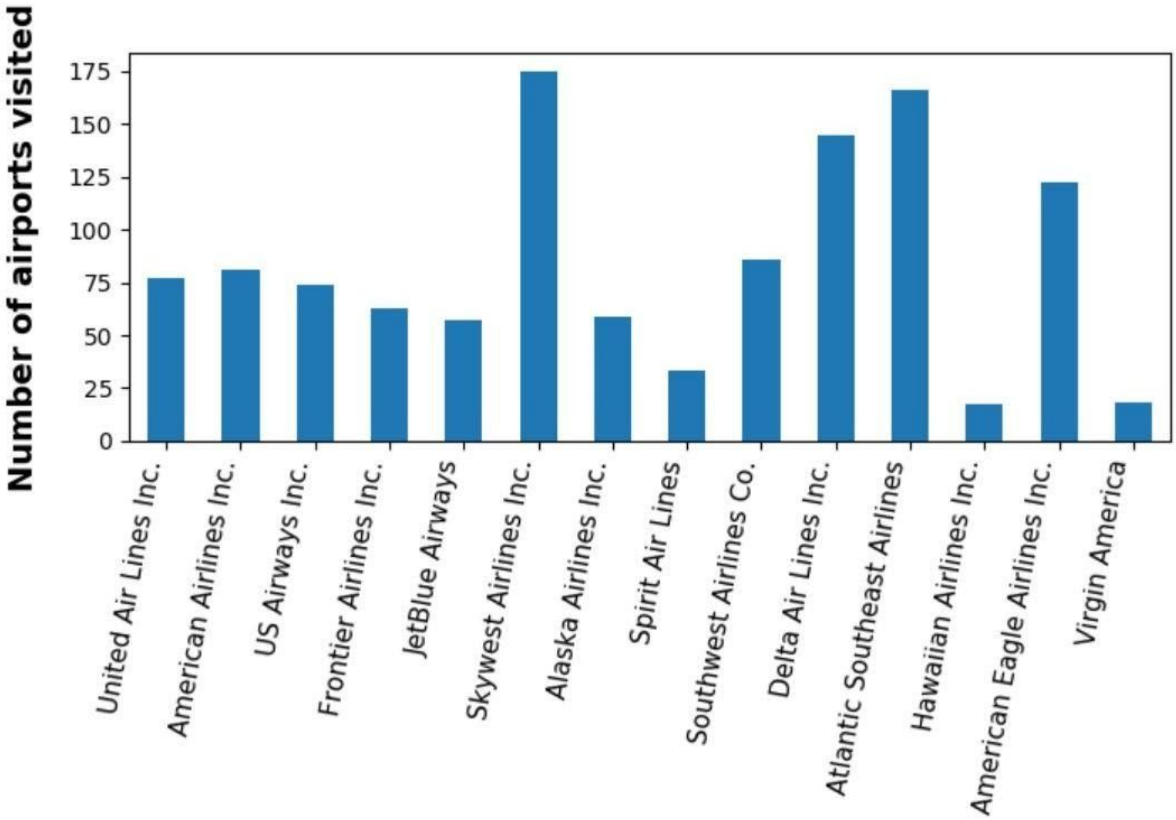

Fig 4.9: Number of airports visited by each airlines

The above graph shows the number of airports visited by each airline company. It can be seen that although Southwest Airlines Co. has the most number of airlines under it, it is the Skywest Airlines Inc. that has covered most of the airports.

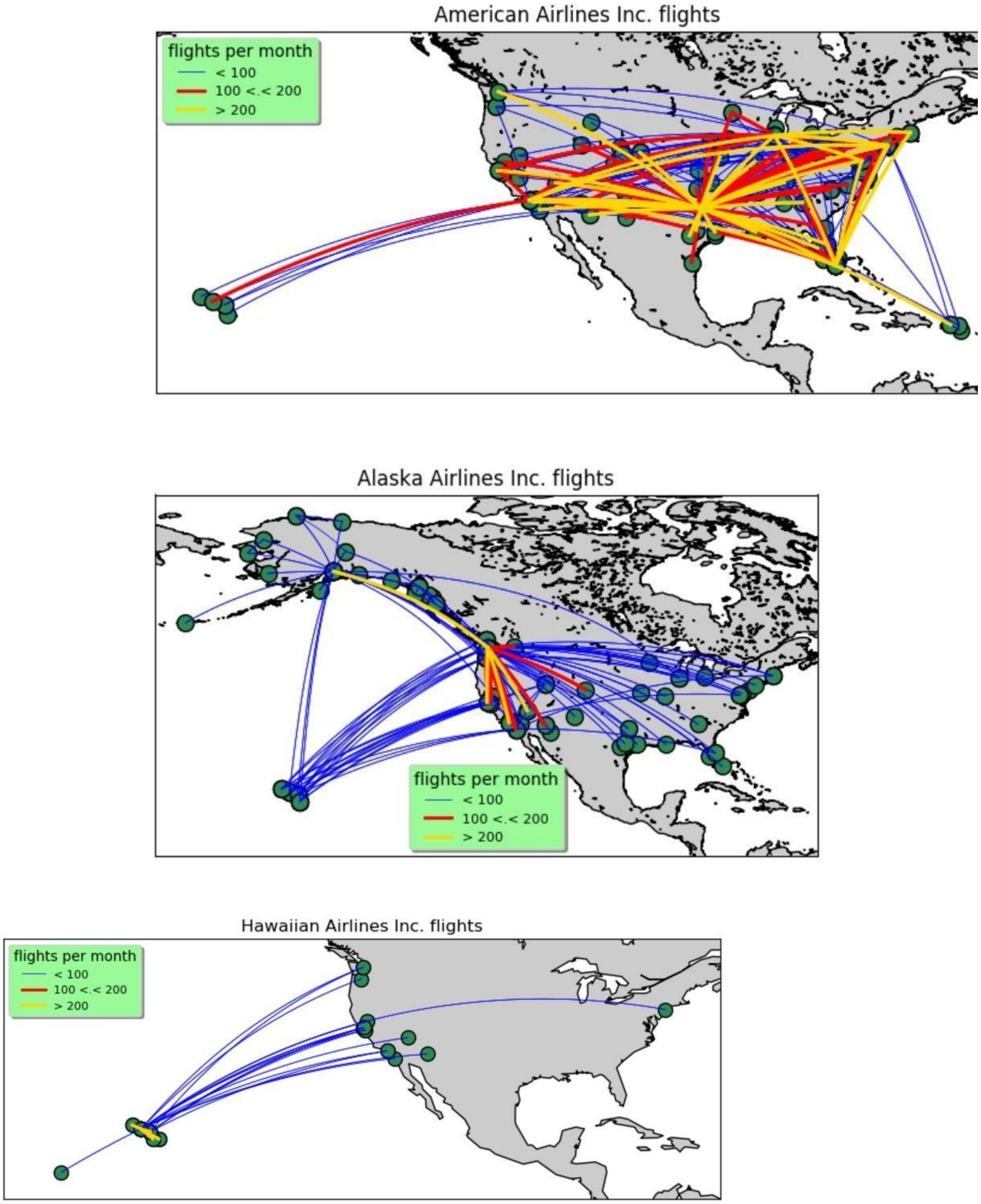

Fig 4.10: air traffic of each airline company

The above three maps show the route followed by the flights from three different companies per month. It can be seen that some companies have flights greater than 200 flying around each month while some have less than 100 and cover very less places in the US.

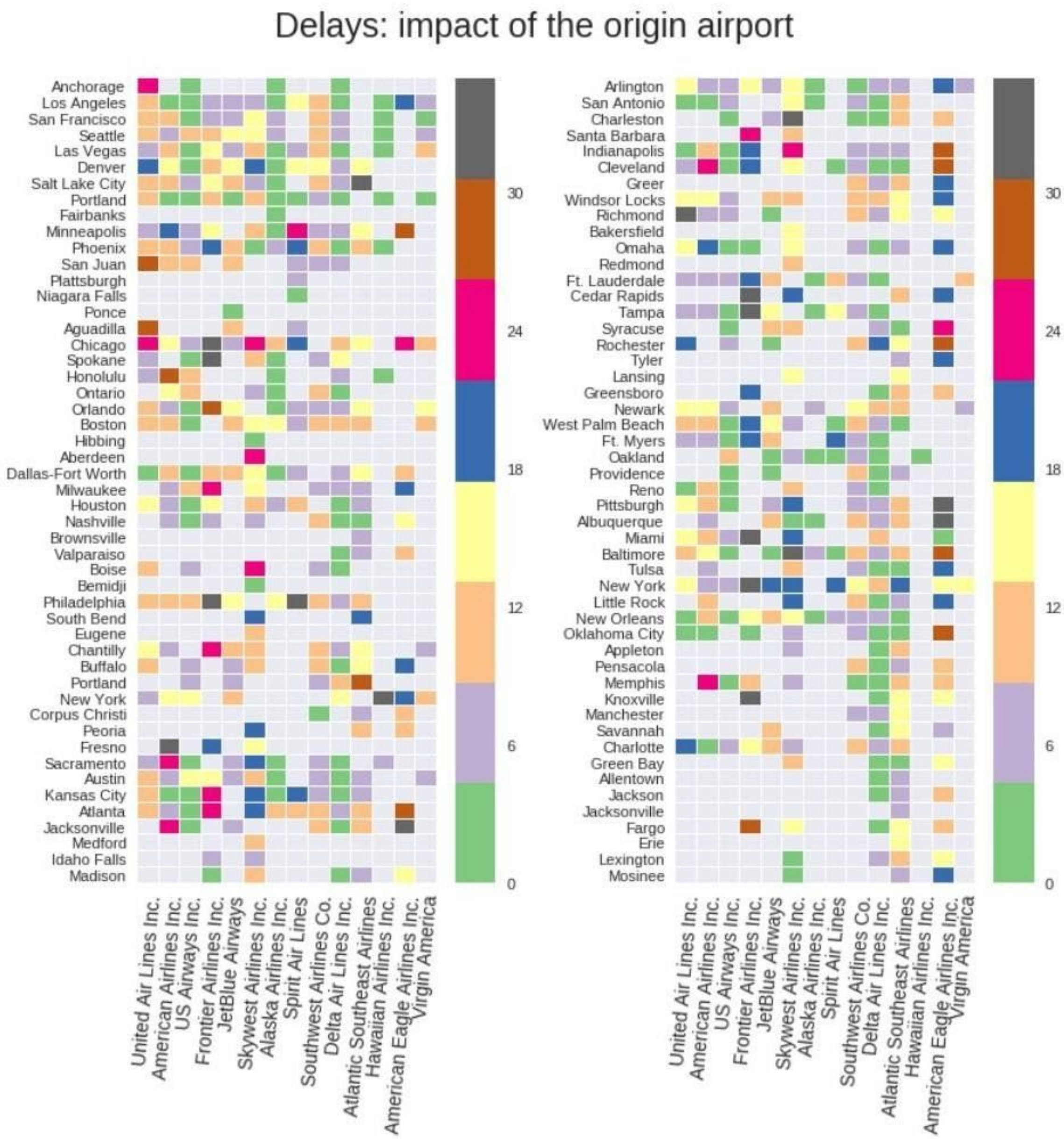

Fig 4.11: Impact of the origin airports on the delay

The above figure shows a heatmap concerning the delays based on the origin airport the flight travels from as we are only considering the arrival delay. It can be due to the weather or the traffic in that place. Like in the cases of metropolitan cities where most people travel to and from like New york. It can be seen that there is a delay of more than 30 minutes. It might be because there is more air traffic and bad weather conditions. This could also affect the other flights flying in that region. There is also a show of how many airports each airline company has delays of. It can be seen that delta airlines although having less than 5 mins delay have the most number of airports in which it delays in.

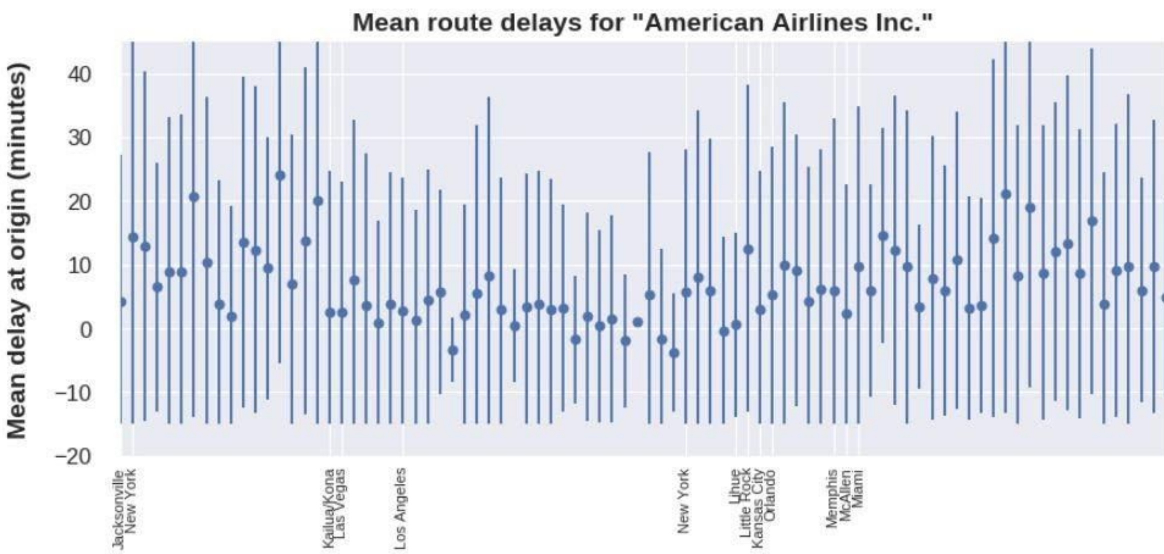

Fig 4.12: Mean route delays for the airline American airlines Inc.

The above figure shows the mean delay for each kind of route for the airlines coming from the company of American Airlines Inc. Since there are many number of flights, we only visualised the flight routes for American Airlines Inc. There are especially many delays of about 20 minutes in the case of routes in New york and miami.

```
In [56]:
lm = linear_model.LinearRegression()
model = lm.fit(X,Y)
predictions = lm.predict(X)
print("MSE =", metrics.mean_squared_error(predictions, Y))
```

```
MSE = 53.7430736542
```

Fig 4.13: Mean square error of Linear regression model

The above shows the linear regression model fitting and the mean square error of the predictions. In this case, the mean square error is 53.74. Here only the ‘AA’ airline is taken for the prediction.

```
In [60]:
result = regr.predict(X_)
print("MSE =", metrics.mean_squared_error(result, Y))
```

```
MSE = 49.5025438214
```

Fig 4.14: Mean square error of Polynomial regression model

The above figure shows the fitting of polynomial regression of degree 2. It was seen as slightly better than the linear regression model with a mean square error of 49.5.

5.Summary

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system.

The purpose is not to obtain the best possible prediction but rather to emphasize on the various steps needed to build such a mode, the importance of the separation of the dataset during the training stage and how cross-validation helps in determining accurate model parameters. It shows how to build linear and polynomial models for univariate or multivariate regressions and also, It gives us some insight on the reason why regularisation helps us in developing models that generalize well.

5.1.Conclusion

With the continuous growing travel demand, limited capacity for airports, and fast growing aviation traffic, flight delay prediction has become an essential thing in the aviation industry. Accurate flight delay prediction has become indispensable to all airport congestion and improve relatively low on time performance for major commercial airports.

This timeline showed a dominance of delay propagation and root delayover cancellation analysis. This used to focus on statistical analysis and operational research approaches in the past. However, as the data volume grows, we noticed the use of machine learning and data management is increasing significantly. In this project we are using Linear regression and Polynomial regression and comparing their outputs to see which one gives an accurate output of prediction. We can observe that Polynomial regression gives the better result when it comes to difference between predictions and real delays which are greater than 15 minutes.

There is a risk of overfitting by proceeding that way and the free parameters of the model will be biased. Hence, the model will not allow a good generalization. In what follows, we therefore split the datas in order to train and then test the model. The first part dealt with an exploration of the dataset, with the aim of understanding some properties of the delays registered by flights. This exploration gave us the occasion of using various visualization tools offered by python. The second part of the notebook consisted in the elaboration of a model aimed at predicting flight delays. We used polynomial regressions and showed the importance of regularisation techniques.

5.2.Scope for Future Study

Future Research can be conducted to enhance the accuracy of the proposed model by using both domestic and international flights. In addition, cargo flights were avoided in the flight prediction due to lack of waiting data on cargo flights. Collecting reliable data from air traffic control authorities and embedding them in the proposed delay prediction model is another promising yet challenging future work.

Delays can be induced by different sources and airports, airlines, en route airspace or an ensemble of them. For analysis purposes, one may assume a simplified system where only one of these actors or any combination of them is considered.
It should be noted that any scope of application can be combined with any problem mentioned. Some work focused on airports to predict delays for all departs considered all airlines anden route airspace indifferently, Airports are also the focus when the objective is to investigate their efficiency based on delays of all carriers. On the other hand, only airlines are considered when comparing the performance of two airlines under the same conditions

An ensemble of airport and en route airspace were studied to understand the relationship between congestion and delays. Others considered airports and airlines as well to evaluate capacity problems and airlines decisions. There are many possibilities to ensemble scopes.This becomes important when studying the dynamics of air transportation systems, mainly when targeting root delay.

REFERENCES:

[1] S. Ahmadbeygi, A. Cohn, and M. Lapp. Decreasing airline delay propagation by re-allocating scheduled slack. *IIE Transactions (Institute of Industrial Engineers)*, 42(7):478–489, 2010.

[2] C. Ariyawansa and A. Aponso. Review on state of art data mining and machine learn- ing techniques for intelligent Airport systems. In *Proceedings of 2016 International Conference on Information Management, ICIM 2016*, pages 134–138, 2016.

[3] V. Du ¨ck, L. Ionescu, N. Kliewer, and L. Suhl. Increasing stability of crew and aircraft schedules. *Transportation Research Part C: Emerging Technologies*, 20(1):47–61, Feb. 2012.

[4] R. Beatty, R. Hsu, L. Berry, and J. Rome. Preliminary evaluation of flight delay propagation through an airline schedule. 2nd USA/Europe Air Traffic Management R&D Seminar, 7(4):259–270, 1998.

[5] B. Zou and M. Hansen. Flight delays, capacity investment and social welfare under air transport supply-demand equilibrium. Transportation Research Part A: Policy and Practice, 46(6):965–980, July 2012.

[6] J. Xiong and M. Hansen. Modelling airline flight cancellation decisions. Transportation Research Part E: Logistics and Transportation Review, 56(Supplement C):64–80, Sept. 2013.

[7] L. Hao, M. Hansen, Y. Zhang, and J. Post. New York, New York: Two ways of estimating the delay impact of New York airports. Transportation Research Part E: Logistics and Transportation Review, 70(Supplement C):245–260, Oct. 2014.

[8] S. Pathomsiri, A. Haghani, M. Dresner, and R. J. Windle. Impact of undesirable outputs on the productivity of US airports. Transportation Research Part E: Logistics and Transportation Review, 44(2):235–259, Mar. 2008.

[9] A. J. Reynolds-Feighan and K. J. Button. An assessment of the capacity and congestion levels at European airports. Journal of Air Transport Management, 5(3):113–134, July 1999

[10] M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak. Detecting periodic patterns of arrival delay. Journal of Air Transport Management, 13(6):355–361, Nov. 2007.

[11] Tu, Y., Ball, M. O., & Jank, W. S. (2008). Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. Journal of the American Statistical Association, 103(481), 112-125.

[12] Boswell, S. B., & Evans, J. E. (1997). Analysis of downstream impacts of air traffic delay. Lincoln Laboratory, Massachusetts Institute of Technology.

[13] Mueller, E., & Chatterji, G. (2002). Analysis of aircraft arrival and departure delay characteristics. In AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum (p. 5866).

[14]Hansen, M., and C. Y. Hsiao (2005), Going South? An Econometric Analysis of US Airline Flight Delays from 2000 to 2004, Presented at the 84th Annual Meeting of the Transportation Research Board (TRB), WashingtonD.C.'05.

[15] Rosen, A. (2002), Flight Delays on US Airlines: The Impact of Congestion Externalities in Hub and Spoke Networks, Department of Economics, Stanford University .