BIG DATA ANALYTIC APPLICATIONS SYMPOSIUM - FALL 2019

Analytics Project:

Relationship between Flight delays and Weather Data for different US airports

Team: Aishwarya Budhkar, Varada Hanumante

Abstract:

- Many factors influence the air travel delays
- The aviation areas generate huge amounts of data
- Weather, flight information and sensor data can be used to develop predictive model using big data analytics
- Influence of weather on the flight delays can lead to helpful insights for the airport and airline operations to enhance flight safety, improve customer satisfaction and reduce flight delays

Relationship between Flight delays and Weather Data for different US airports

Motivation:

- Citizens can use the application to find the status of their upcoming flight.
- Airline companies can use this analytic to manage their flight passengers.

Importance:

- Citizens can arrange for change in their plans after the flight in case the flight is getting delayed
- Airline companies can make proper arrangements for the passengers if the flight is getting delayed and make any schedule change for connecting flights

Relationship between Flight delays and Weather Data for different US airports

Goodness

We compared our analytic with existing researches/ findings in the domain and we then checked if our findings align with them and tried to analyze the causes of observed results

We used following paper for this purpose:

- Chained Predictions of Flight Delay Using Machine Learning

https://www.researchgate.net/publication/330185077_Chained_Predictions_of_Flight_Delay_Using_Machine_Learning

- Integrating Flight-related Information into a (Big) Data Lake

https://www.researchgate.net/publication/320968718_Integrating_flightrelated_information_into_a_Big_data_lake

Relationship between Flight delays and Weather Data for different US airports

Actuation/Remediation

1. Check status of upcoming flight

- Developed an application which takes user flight details as input and returns whether the flight will be delayed according to our model.
- User is required to provide flight date, time and airport within 5 days from that days date.
- Weather data forecast from NOAA API is collected.
- The features are used by the model to predict the delay chance.

2. An app which shows the flight delays for an airport for the past few years

- User can take into account the past information about delays to get an idea about delays while booking the flight

Relationship between Flight delays and Weather Data for different US airports

Data Sources:

Name: National Centers for Environmental Information, National Oceanic and Atmospheric Administration Link

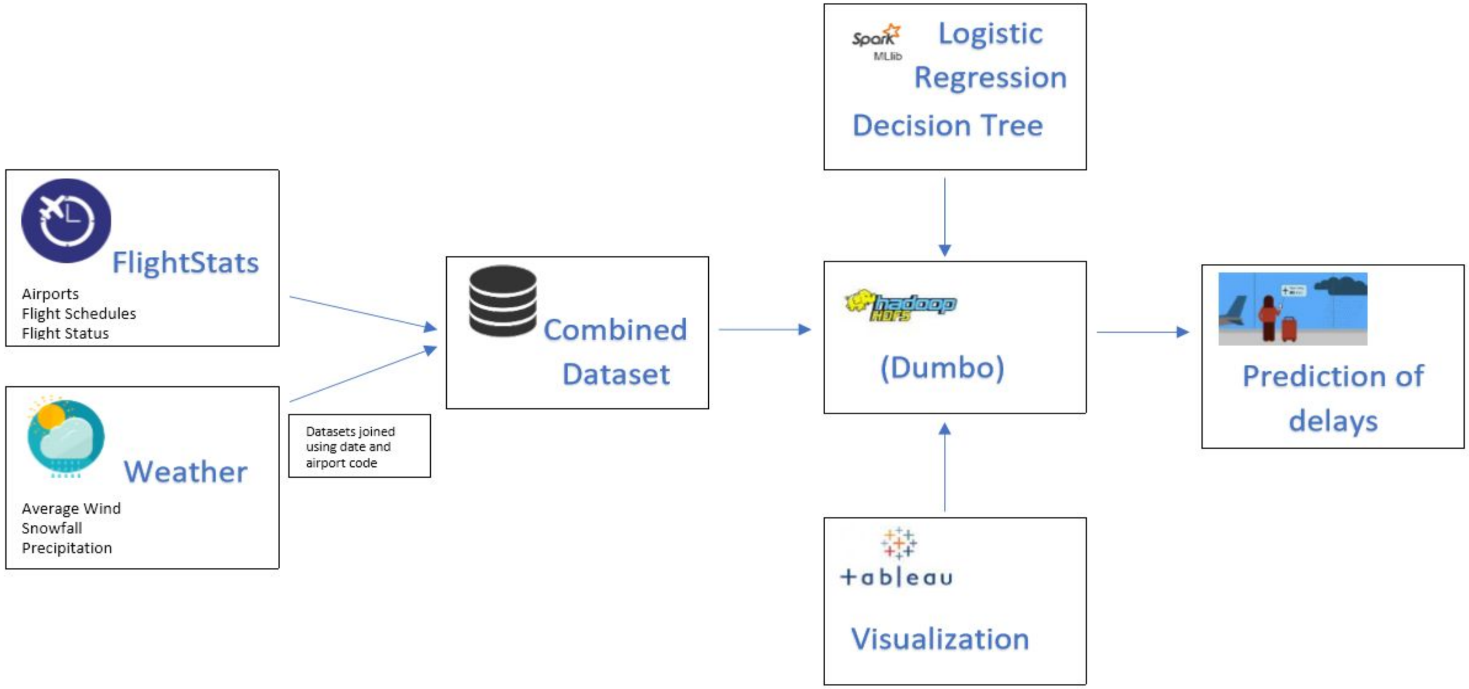
Description: This dataset includes data collected by various weather stations in US. It contains data fields like air temperature, precipitation, snow, wind, sun shine, etc.
Size of data: 0.5 GB

Name: Bureau of Transportation Statistics – flight data

Description: The airline trip records include fields capturing date, time, origin and destination airports, flight distance, flight departure time, flight departure scheduled time, flight departure delay, flight arrival time, flight arrival delay, flight arrival scheduled time, etc.
Size of data: 5 GB

Relationship between Flight delays and Weather Data for different US airports

Design Diagram:

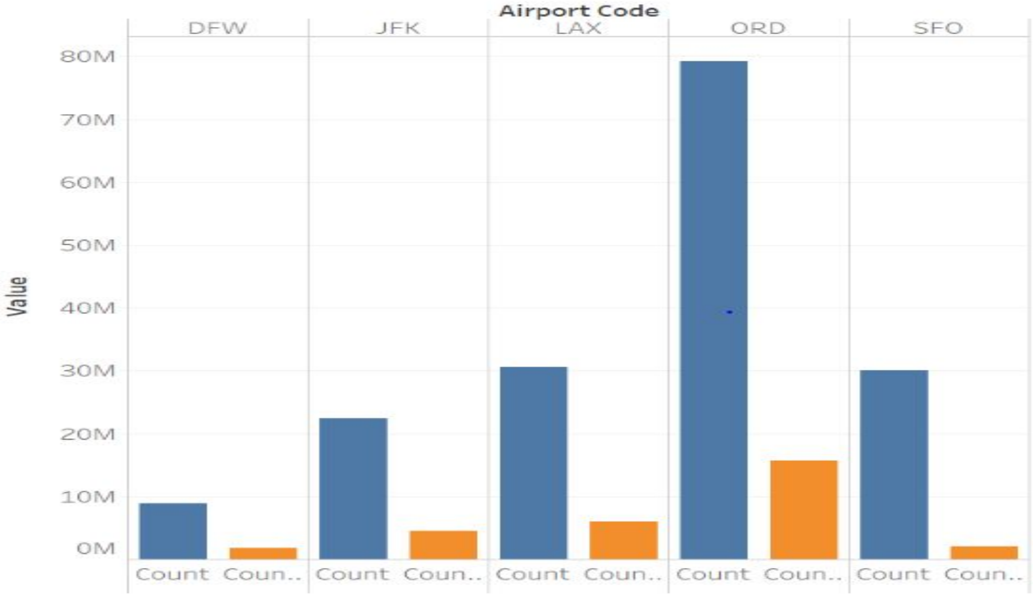


Platform on which the analytic ran:
NYU HPC Cluster

Relationship between Flight delays and Weather Data for different US airports

Code Challenge1:

1. Non-uniform data distribution



- Delayed flight records four times the number of non-delayed flight records.
- Synthetic Minority Over-Sampling techniques is used to generate synthetic samples to make the data distribution uniform

Relationship between Flight delays and Weather Data for different US airports

Code Challenge 2:

2. Choosing important features

- Number of factors regarding flight information and weather delays were available

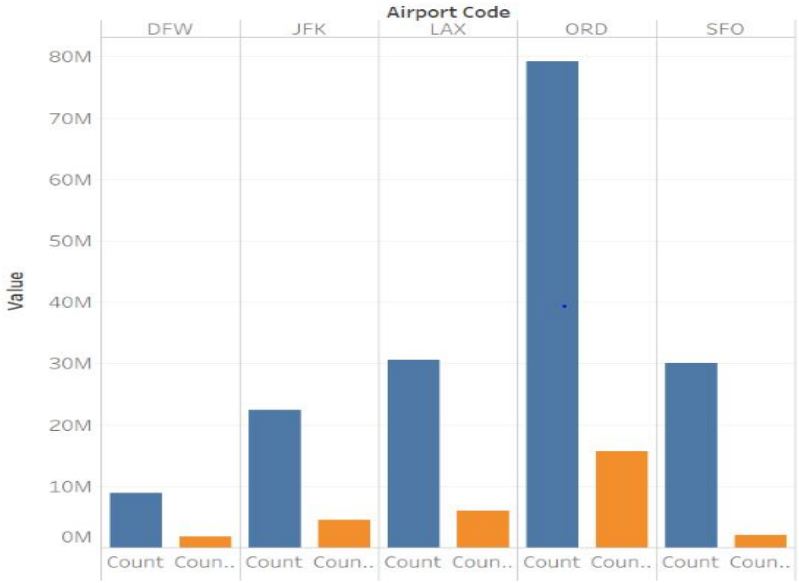
- As too many factors can confuse the classifier, it is important to choose to features affecting the analytic the most.

- Random forest algorithm was used to select the features according to the importance

Relationship between Flight delays and Weather Data for different US airports

Insight:

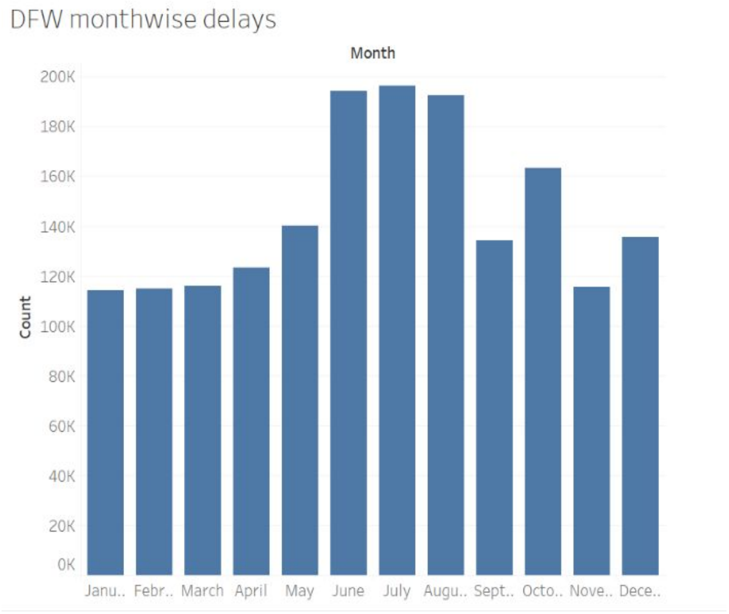
- ❖ SFO has least percentage of delays but other airports show almost equal percentage of delays

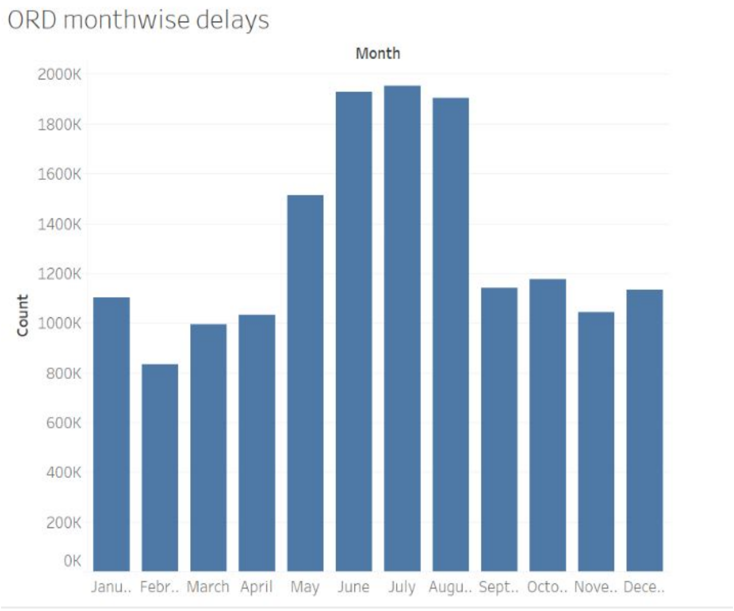


Relationship between Flight delays and Weather Data for different US airports

Insight:

- The flight departure delays were high at ORD and DFW showed high number of delays from June to August.

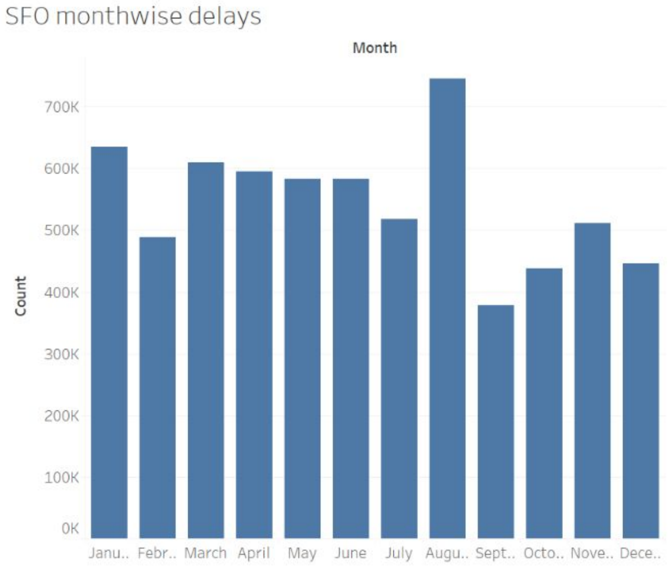


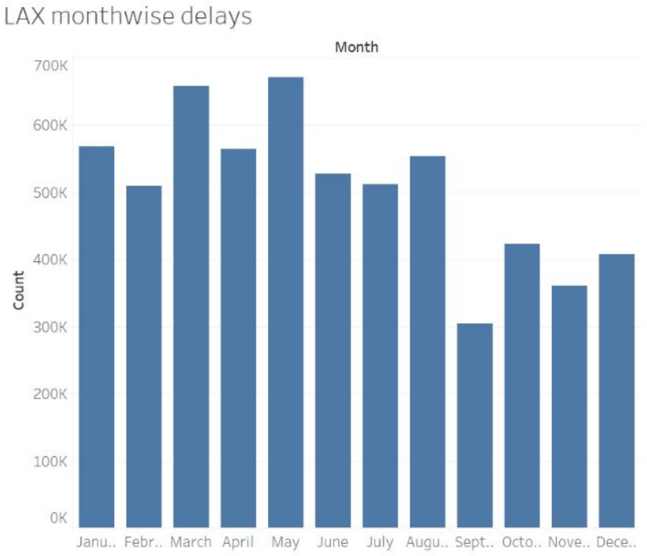


Relationship between Flight delays and Weather Data for different US airports

Insight:

- ❖ The delays distribution was spread evenly at LAX and SFO, less delays were observed from September to December.





Relationship between Flight delays and Weather Data for different US airports

Insight:

- Validation accuracy was 78 % when model was trained using data collected from sources

- To remove bias during training SMOTE was used to generate data with 50% delayed flights and 50% non-delayed and cancelled flights

- Validation accuracy of 87 % was obtained

- Addition of features like arrival delay and flight distance led to increased accuracy to 92% with Random Forest Algorithm

- Amongst the 4 algorithms, Logistic Regression, SVM, Decision Trees, Random Forest, highest accuracy was obtained with random forests

Relationship between Flight delays and Weather Data for different US airports

Obstacles:

1. Cleaning NOAA Weather data

- Difficult to parse due to more number of columns in some rows
- There were many missing values
- Some dates were having inconsistent formats

2. Cleaning Bureau of Transportation Statistics - Flight Data

- There were extra columns in few records
- Few values of dates were inconsistent
- Some origin and destination airport codes were wrong
- There were negative values for flight distance
- Many iterations for cleaning as data was huge

3. Non-uniform data distribution

- As the non delayed flight records outnumber the delayed records by over 4 times, it was necessary to remove the bias while training the model.

Relationship between Flight delays and Weather Data for different US airports

Next Steps

- Adding more flight related information to improve the analysis
- Use Random Forest algorithm to select important features
- Develop a Chained delay prediction model for connected flights
- Use other Machine Learning models and compare the performance
- Addition of data for more airports and airlines can lead to more insights

Relationship between Flight delays and Weather Data for different US airports

Summary

Flight delays are positively corelated with weather conditions though there are number of other factors which have an impact on taxi pickups like arrival delay, propagation delay, etc. There is chained delay due to late arrival and late departure for connecting flights. All factors together can be used to predict the flight delays for a particular date and airport.

Acknowledgements

- Thanks to NYU HPC for helping with Tableau and Dumbo.
- Thank you Professor Suzanne Mcintosh for guiding us, helping us with every blocker we encountered and constantly encouraging us.
- Thanks to course grader Lakshmi for helping us with obstacles during this project.

Relationship between Flight delays and Weather Data for different US airports

References

1. U.S. Department of Transportation. Research and Innovative Technology Administration. Bureau of Transportation Statistics Scheduled Intercity Transportation: Rural Service Areas in the United States. June 2005. Washington, DC: 2005
2. Chang-Geun. Oh. “Application of Big Data Systems to Aviation and Aerospace Fields; Pertinent Human Factors Considerations,”, International Symposium on Aviation Psychology 2017, May 2017..
3. Alice Sternberg. Jorge Soares. Diego Carvalho. and Eduardo Ogasawara. A Review on Flight Delay Prediction, arXiv:1703.06118 http://arxiv. org/abs/1703.06118, March 2017.
4. Navoneel Chakrabarty. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines, https://arxiv.org/abs/1903.06740, March 2019.Deri, Joya & Moura, Jose. (2015). Taxi data in New York city: A network perspective. 1829-1833. 10.1109/ACSSC.2015.7421468.
5. Jun Chen. Meng Li. Chained Predictions of Flight Delay Using Machine Learning, AIAA Science and Technology Forum and Exposition 2019, CA, January 2019.
6. Kebing Li. Colby College (2019) Investigating the effect crime has on Uber and Yellow Taxi pickups in NYC
7. National Centers for Environmental Information. National Oceanic and Atmospheric Administration.

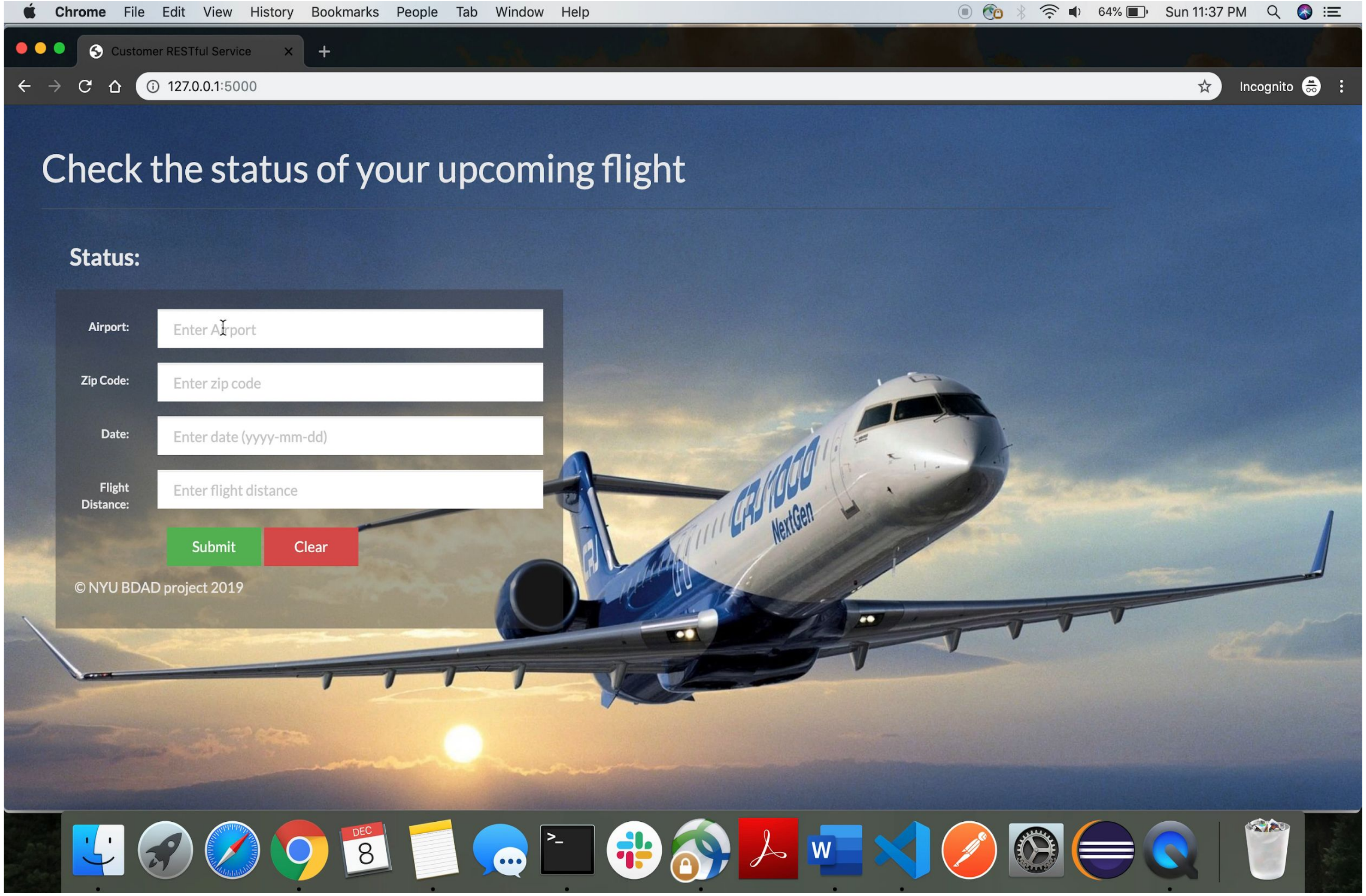
Relationship between Flight delays and Weather Data for different US airports

References

8. Veershetty Dagade. Mahesh Lagali. Supriya Avadhani. Priya Kalekar. Big Data Weather Analytics Using Hadoop, International Journal of Advanced Research in Computer and Communication Engineering, April 2015.
9. Sun Choi. Young Jin Kim. Simon Briceno. Dimitri Mavris. Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms, 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), September 2016.
10. Ning Yang. Lewis Westfall. Ms. Preeti Dalvi. A weather prediction model with Big Data, Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 2018.
11. Chawla, N.V. Bowyer, K.W. Hall, L.O. Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 2002, 16, 321–357.
12. Yao Wang. Deepak Kulkarni. Modeling weather impact on ground delay problems, SAE International Journal of Aerospace 4 (2), Nov. 2011, pp. 1207–1215.
13. Priyanka Chouksey. Abhishek Singh Chauhan. Weather Data Analytics using MapReduce and Spark, International Journal of Advanced Research in Computer and Communication Engineering, February 2017.

Relationship between Flight delays and Weather Data for different US airports

Demo



Check the status of your upcoming flight

Status:

Airport:	Enter Airport
Zip Code:	Enter zip code
Date:	Enter date (yyyy-mm-dd)
Flight Distance:	Enter flight distance

© NYU BDAD project 2019

Thank You