US AIRLINE FLIGHTS DELAYS ANALYSIS AND PREDICTION

Group 10
Xinze Yu,
TzeHow Lee,
Aiwu Song

Content

Articulated problem

Dataset Description

Flight Delay Analysis

Prediction Model

Stakeholder Engagement

Conclusion

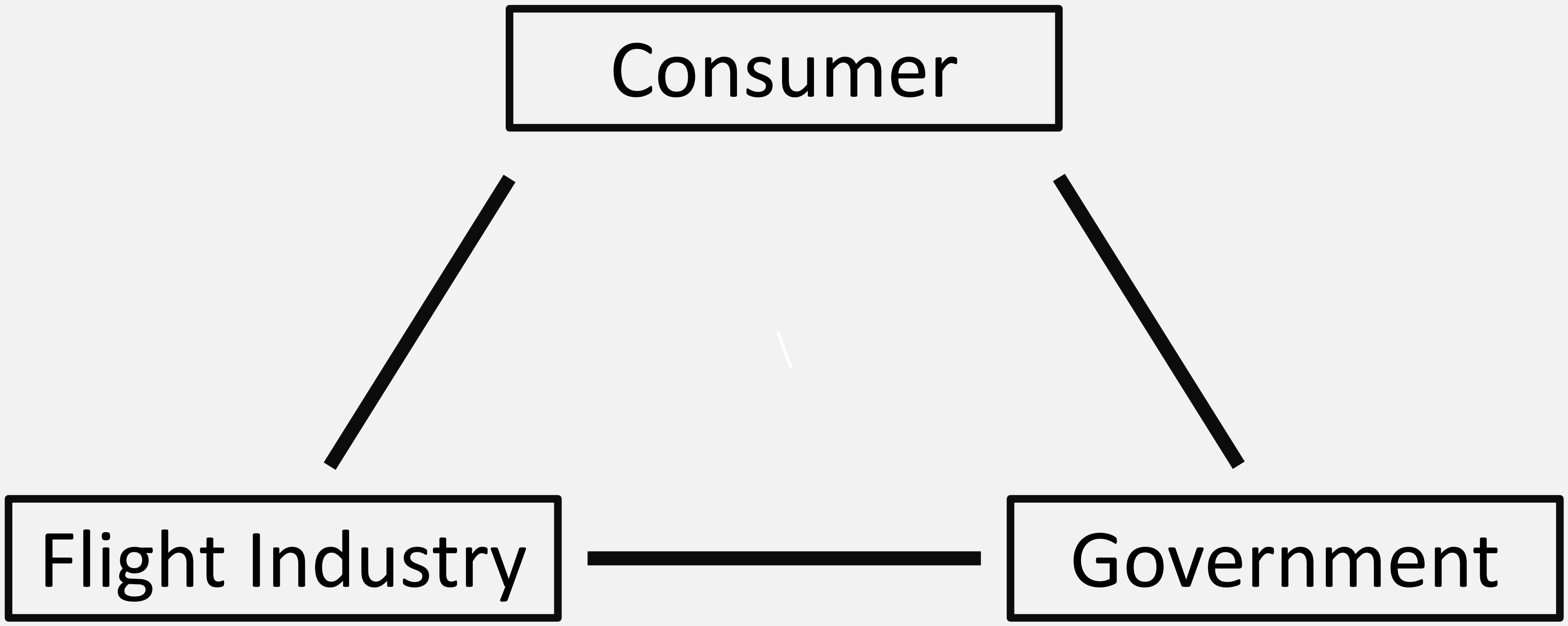
Articulated problem

US Airline Flight Delays Analysis and Prediction Model

- Finding patterns in causes of delays.
  Including weather data.

- Reduce substantial economic losses
  to passengers, airlines, and airport operators. [1]

Business Model Example



Business Model Example

Consumer

Flight Tracking App: Flighty

Consumer Report from WSJ

Business Model Example

Flight Industry

- Airlines
- Airport Management Companies
- Businesses that rely on aviation

Business Model Example

Governor

U.S. Department of Transportation

- Aviation Consumer Protection
- Rules, Guidance, and Enforcement Orders

Dataset: Sources

U.S. Department of Transportation

NATIONAL WEATHER SERVICE

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

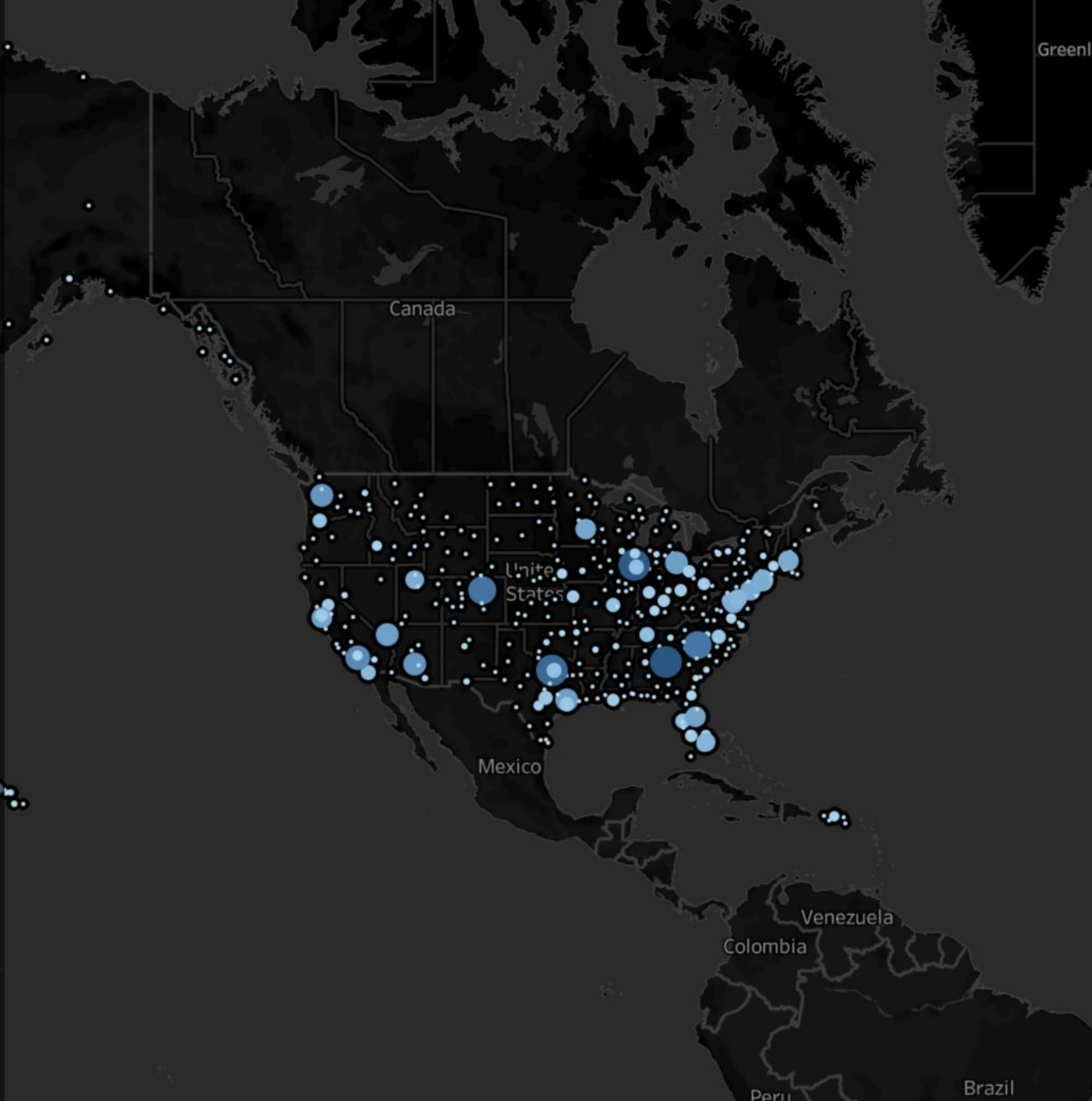
Flight Delay Analysis

Description of Dataset

	Carrier	Airport	Time	Delay	DOW	CarrierFact	NASFact	SecurityFact	Weather
0	AA	ATL	Night	Delay	3	0	0	0	Normal
1	UA	ATL	Noon	On Time	3	1	0	0	Normal
2	UA	ATL	Evening	On Time	3	0	0	0	Normal
3	UA	ATL	Night	On Time	3	0	0	0	Normal
4	UA	ATL	Night	On Time	3	0	0	0	Normal
5	UA	ATL	Evening	On Time	3	0	0	0	Normal
6	UA	ATL	Afternoon	Delay	3	0	0	0	Normal
7	AA	ATL	Afternoon	On Time	3	0	0	0	Normal
8	AA	ATL	Afternoon	On Time	3	0	0	0	Normal
9	AA	ATL	Evening	On Time	3	0	0	0	Normal

- The original dataset was downloaded from USDT, there are in total 35 columns data, for our analysis and prediction, we only select 'carrier', 'departure airports', 'actual departure time', 'Day of week' etc., in total 8 facts with corresponding delay types. (We divided time into 5 periods and delay into 2 types) [2]
- Also, we collected airport weather information from National Weather Service web and in conjunction with other data. (4 types of weather)

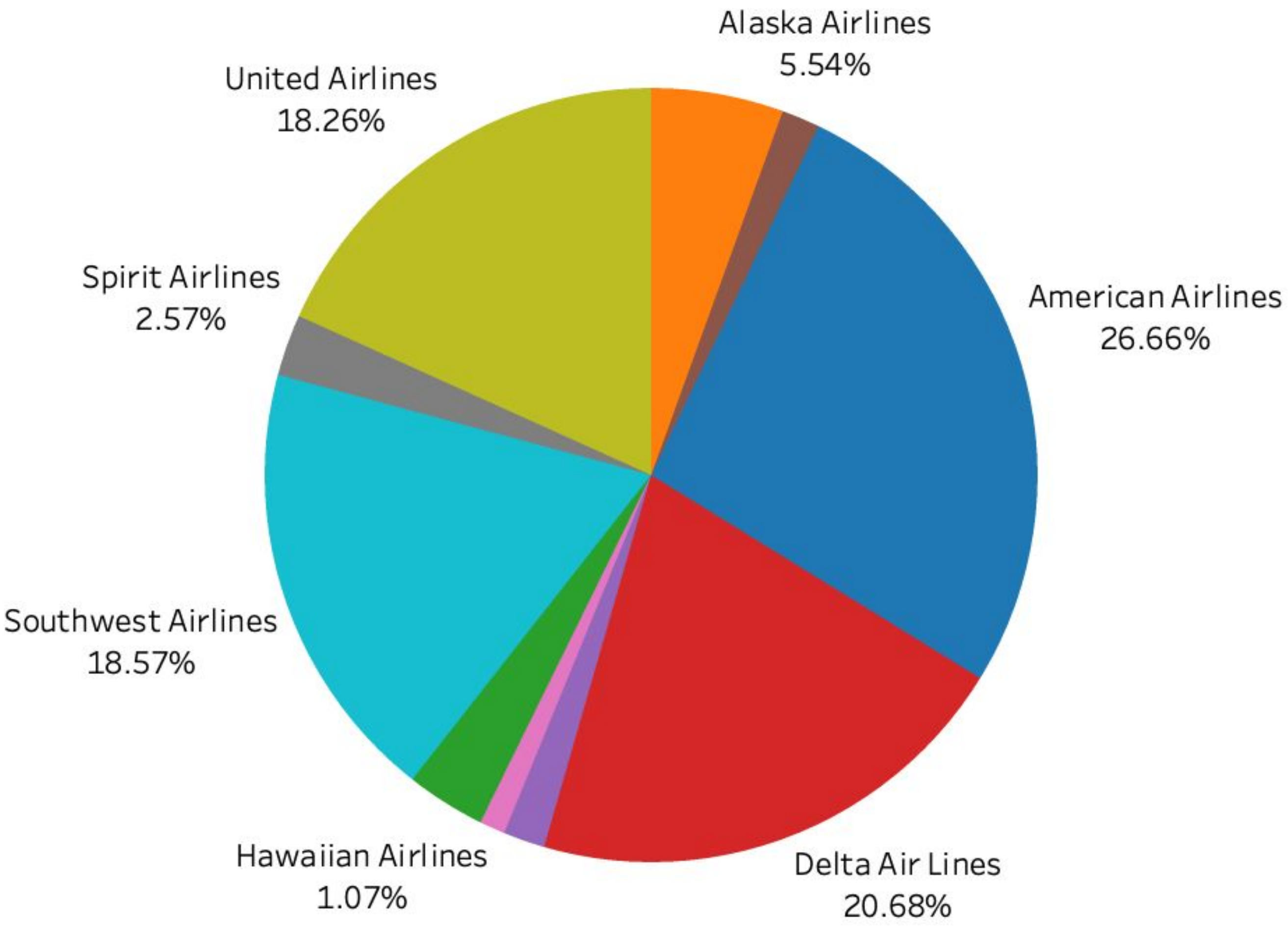
- Entire view of the airport map with its flight number, the large the circle the more flights it owns

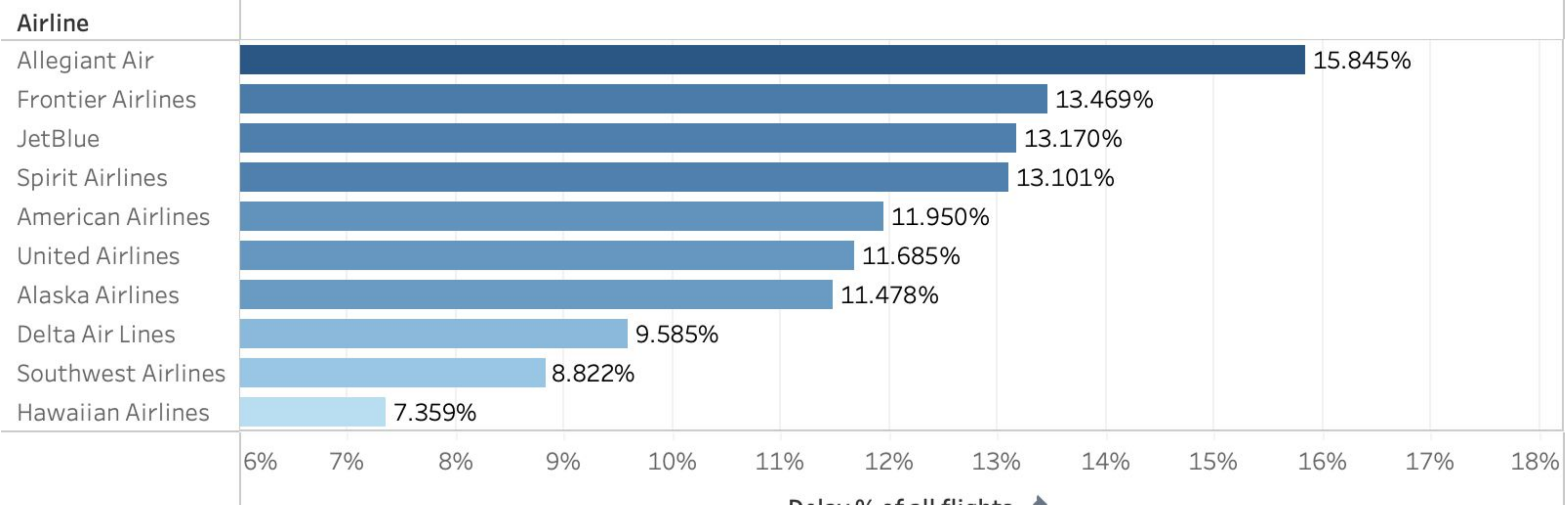


Descriptive Data Analysis – Market share for main airlines in US

- From the pie chart we can see that American Airlines, Delta Air Lines, Southwest Airlines and United Airlines are the top airlines in US since they have almost total 73% market shares.

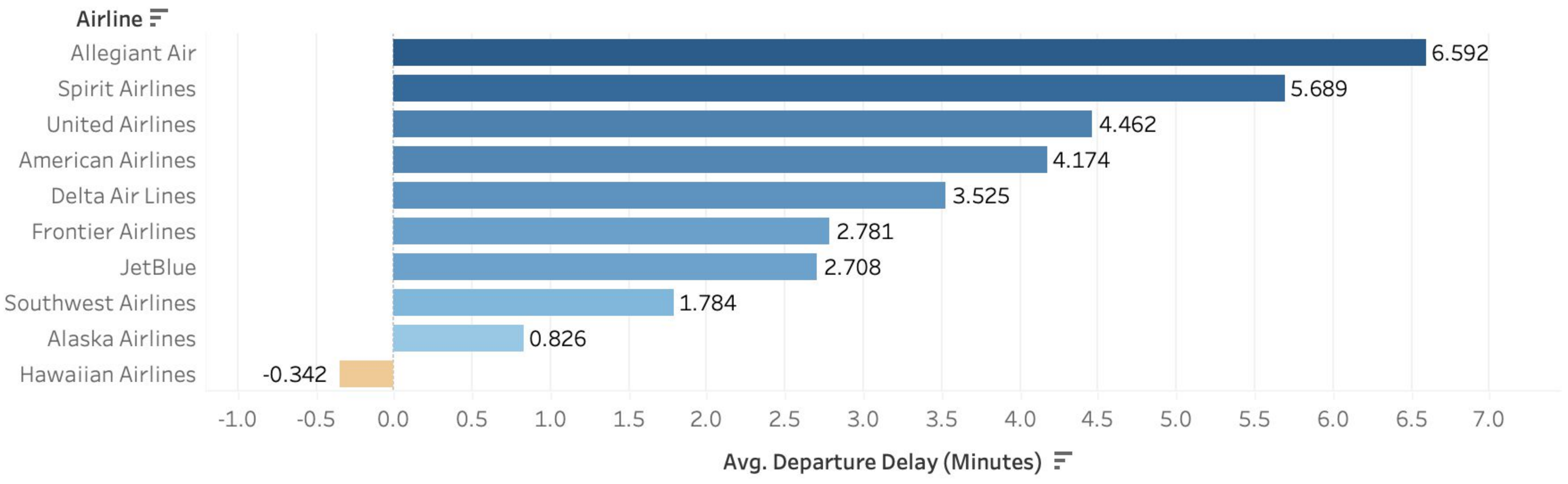
Airline	
American Airlines	26.66%
Delta Air Lines	20.68%
Southwest Airlines	18.57%
United Airlines	18.26%
Alaska Airlines	5.54%
JetBlue	3.32%
Spirit Airlines	2.57%
Frontier Airlines	1.74%
Allegiant Air	1.59%
Hawaiian Airlines	1.07%





Descriptive Data Analysis – Delay ratio rank for each airlines

As we can see above, Hawaiian Airlines has the best performance for only having a 7.359% of delay, while Allegiant Air is the worst.

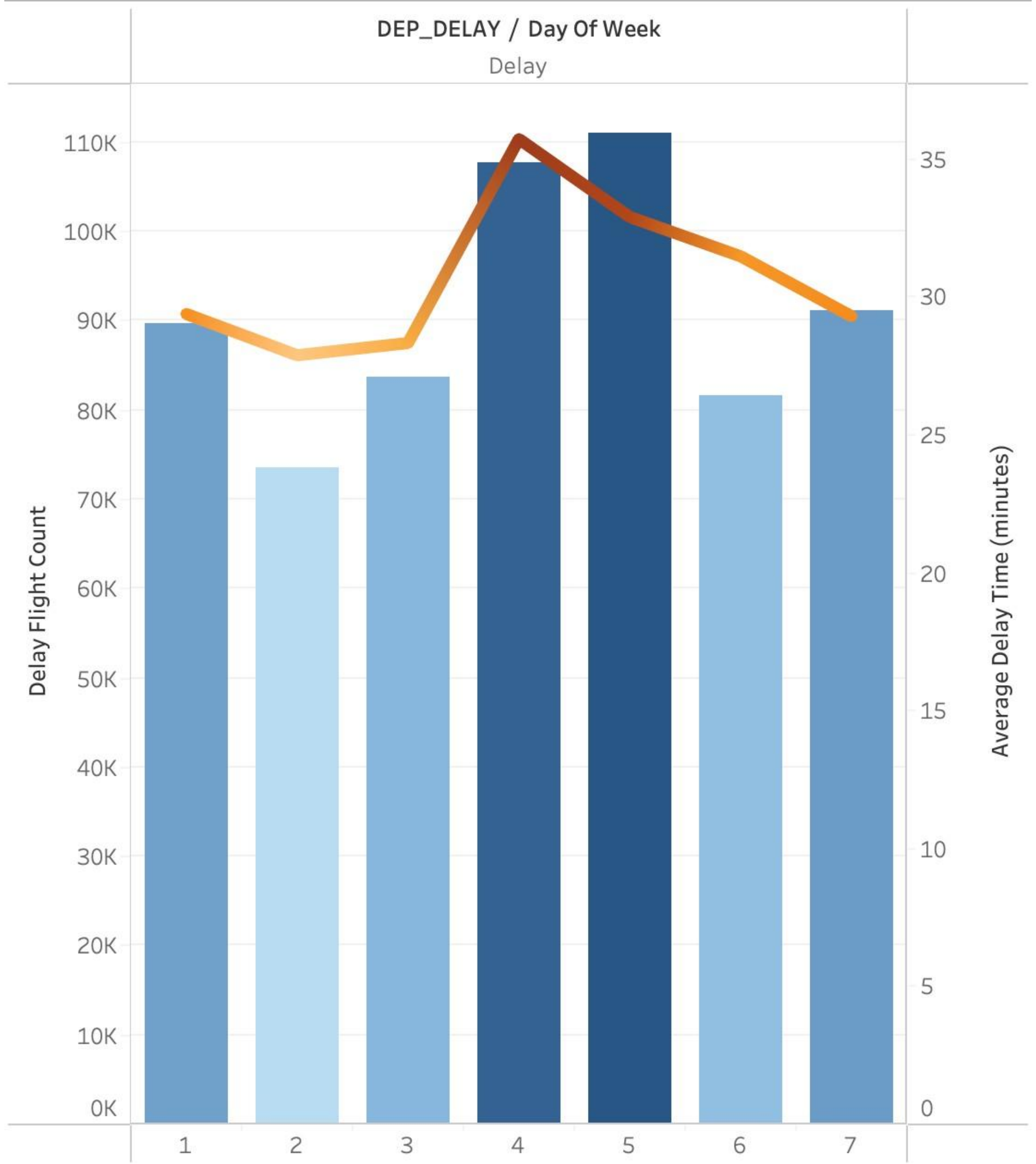


Descriptive Data Analysis – Avg delay minutes rank for each airlines

From the chart, we can see Hawaiian airlines has the best performance the average delay minutes is -0.342 (they even departure early!), but while Allegiant Air is still the worst.

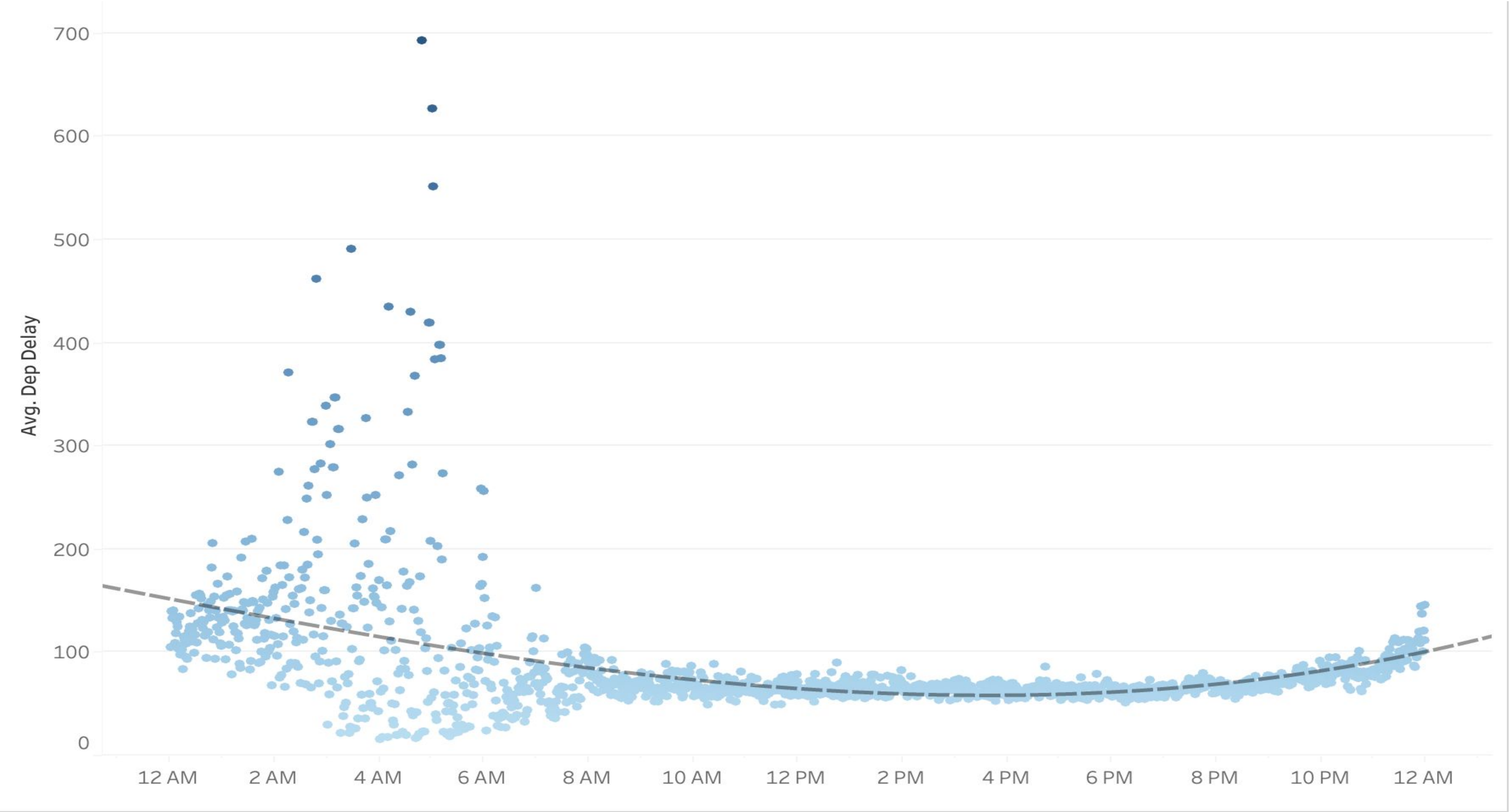
Descriptive Data Analysis – Relationship with delays and weekday

- We can see that avg delays time are more on Thursday and Friday. But variation is not too much.
- Also, the delay counts are more on Thursday and Friday. But still no significant difference.



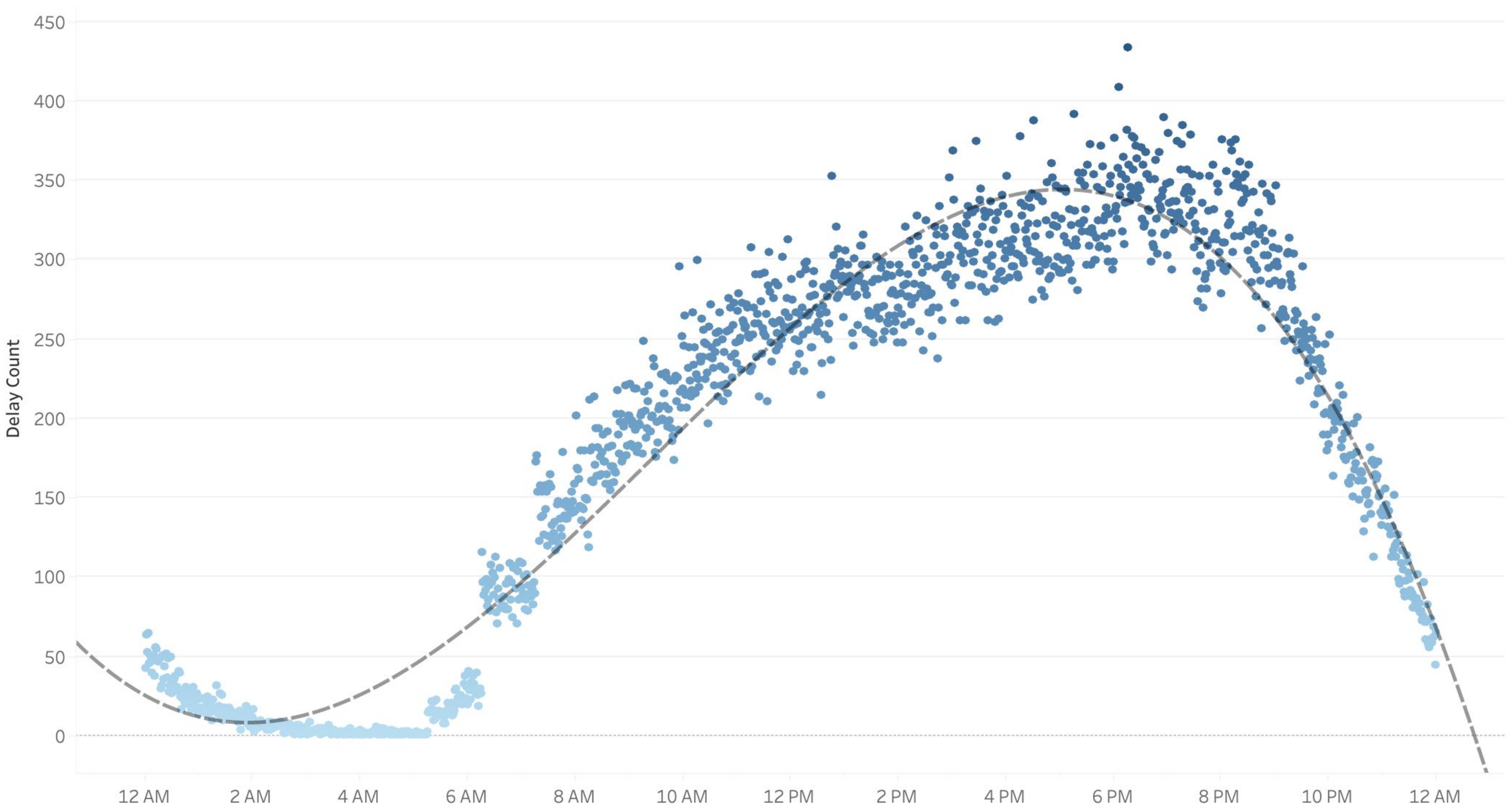
Descriptive Data Analysis – Temporal variability of delays

For average departure delay, it is more likely to face an extreme delay experience at early morning (2am – 6am), and the rest of the day tends to be more stable.



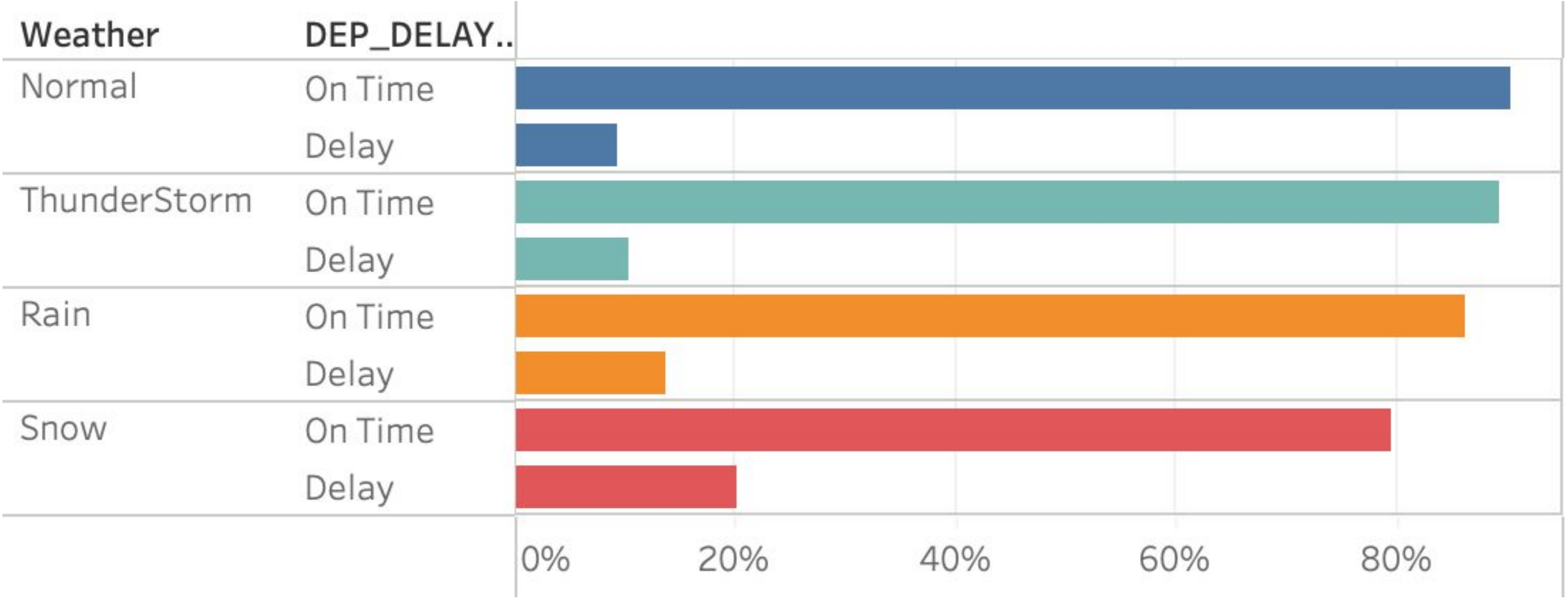
Descriptive Data Analysis – Temporal variability of delays

For delay frequency, most delay happens at 10am to 9pm, peak is around 7pm.



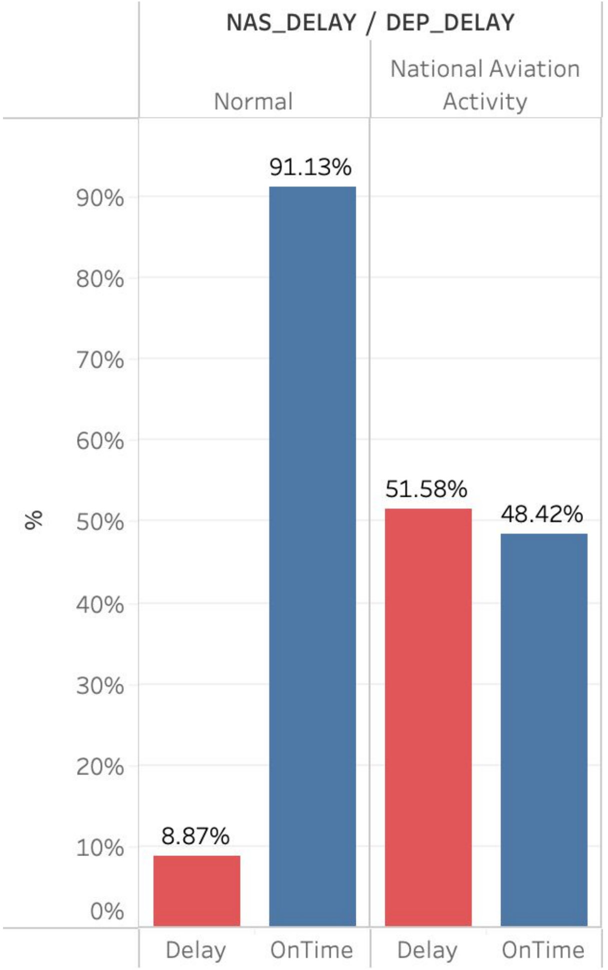
Descriptive Data Analysis – Delay with different weather

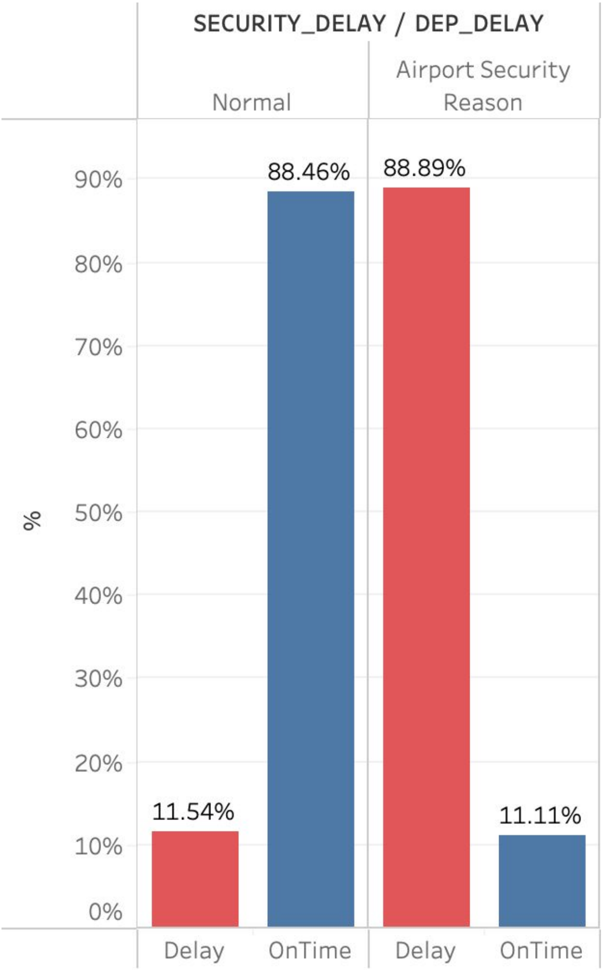
With different weather conditions, the overall On-time rate still above 80%

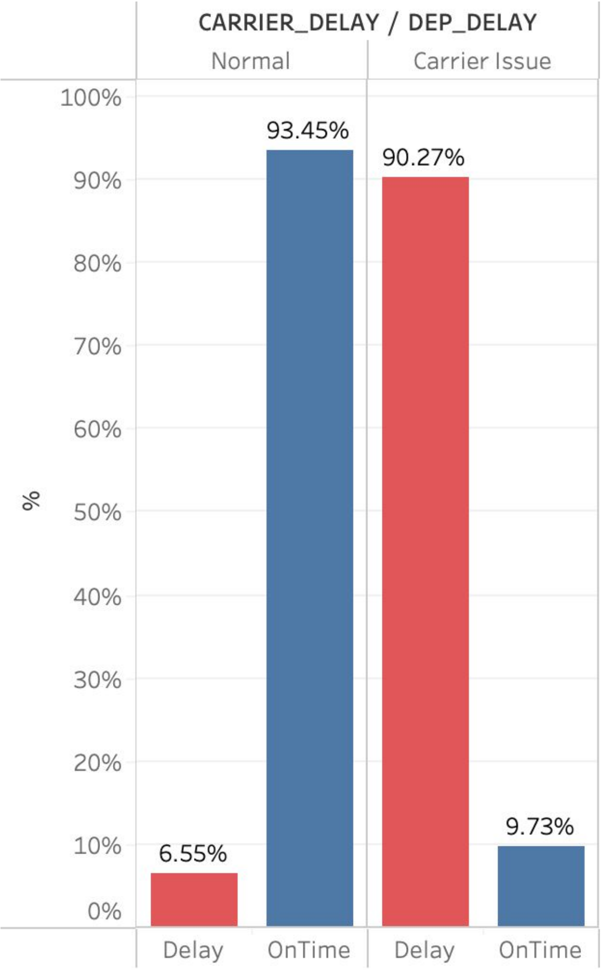


Descriptive Data Analysis – Other Major Delay Reasons

With different major delay reasons , National Aviation Activity cause around 50% delay rate. Airport Security Reason and Carrier cause nearly 90% delay rate







Prediction Model

Appropriate data – Create dummy variables

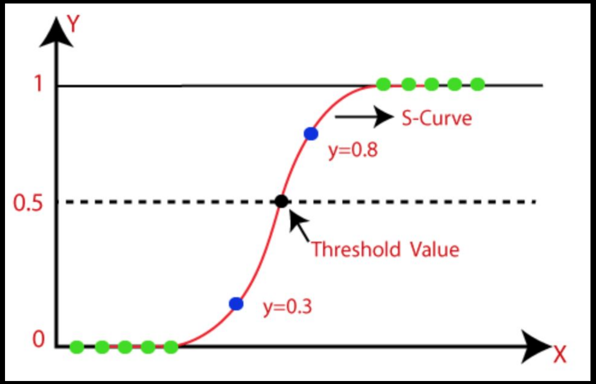
- Since most of our inputs are discrete variables, so we need to convert them into dummy variables for further regression.
- Also, we have 1 target variable (whether the flight will delay or not) [3]

Carrier_UA	Carrier_WN	...	Dow_6	Dow_7	Weather_Normal	Weather_Rain	Weather_Snow	Weather_ThunderStorm	CarrierFact	NASFact	SecurityFact	Delay
0	0	...	0	0	1	0	0	0	0	0	0	Delay
1	0	...	0	0	1	0	0	0	1	0	0	On Time
1	0	...	0	0	1	0	0	0	0	0	0	On Time
1	0	...	0	0	1	0	0	0	0	0	0	On Time
1	0	...	0	0	1	0	0	0	0	0	0	On Time

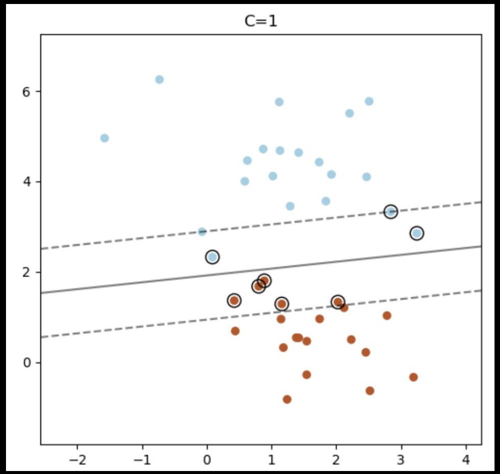
Approaches for analyzing the data - Methods

- Binary Classification:
- For binary classification, we are interested in classifying data into one of two binary groups - these are usually represented as 0 and 1, but in our data, they are ‘On Time’ and ’Delay’.
- We have tried 4 binary classifier, Naïve Bayes, Logistic Regression and Linear SVC(Support Vector Classifier) and Random Forest Classifier. [4]

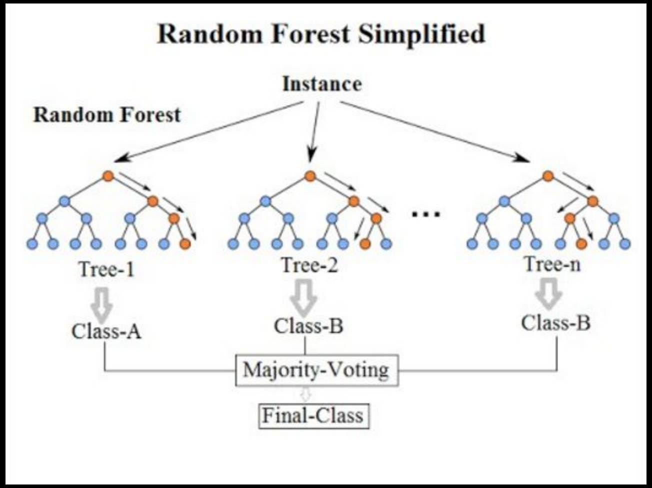
Naïve Bayes



Logistic Regression



Linear SVC



Random Forest

Approaches for analyzing the data - Codes

```
prediction_table = []
prediction_eval_table = []
NB_pipeline = Pipeline([
                ('clf', MultinomialNB(fit_prior=True, class_prior=None)
                        )])

print('... Processing {}'.format('Delay'))
# train the model using X & y
NB_pipeline.fit(X_train, train['Delay'])
# compute the testing accuracy
prediction = NB_pipeline.predict(X_eval)
```

Naïve Bayes

```
prediction_table = []
prediction_eval_table = []
SVC_pipeline = Pipeline([
                ('clf', LinearSVC()
                        )])

print('... Processing {}'.format('Delay'))
# train the model using X & y
SVC_pipeline.fit(X_train, train['Delay'])
# compute the testing accuracy
prediction = SVC_pipeline.predict(X_eval)
```

Linear SVC

```
prediction_table = []
prediction_eval_table = []
LR_pipeline = Pipeline([
                ('clf', LogisticRegression(solver='sag')),
            ])

print('... Processing {}'.format('Delay'))
# train the model using X_dtm & y
LR_pipeline.fit(X_train, train['Delay'])
# compute the testing accuracy
prediction = LR_pipeline.predict(X_eval)
```

Logistic Regression

```
prediction_table = []
prediction_eval_table = []
RFC_pipeline = Pipeline([
                ('clf', RandomForestClassifier(n_estimators=500, max_depth=20)
                        )])

print('... Processing {}'.format('Delay'))
# train the model using X & y
RFC_pipeline.fit(X_train, train['Delay'])
# compute the testing accuracy
prediction = RFC_pipeline.predict(X_eval)
```

Random Forest

Accuracy of result - Metrics

- We will consider the accuracy, precision, recall and F1-Score as our metrics.
- Since it's an unbalanced classification problem and we care more about the recall.

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy of result - Metrics

	precision	recall	f1-score	support
Delay	0.83	0.50	0.62	10627
On Time	0.94	0.99	0.96	81119
accuracy			0.93	91746
macro avg	0.89	0.74	0.79	91746
weighted avg	0.93	0.93	0.92	91746

Naïve Bayes

	precision	recall	f1-score	support
Delay	0.90	0.46	0.61	10627
On Time	0.93	0.99	0.96	81119
accuracy			0.93	91746
macro avg	0.91	0.73	0.79	91746
weighted avg	0.93	0.93	0.92	91746

Logistic Regression

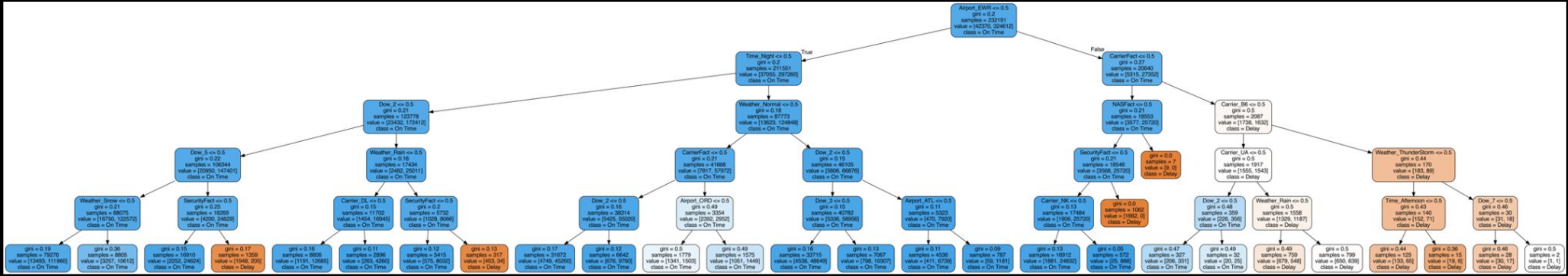
	precision	recall	f1-score	support
Delay	0.90	0.46	0.61	10627
On Time	0.93	0.99	0.96	81119
accuracy			0.93	91746
macro avg	0.91	0.73	0.79	91746
weighted avg	0.93	0.93	0.92	91746

Linear SVC

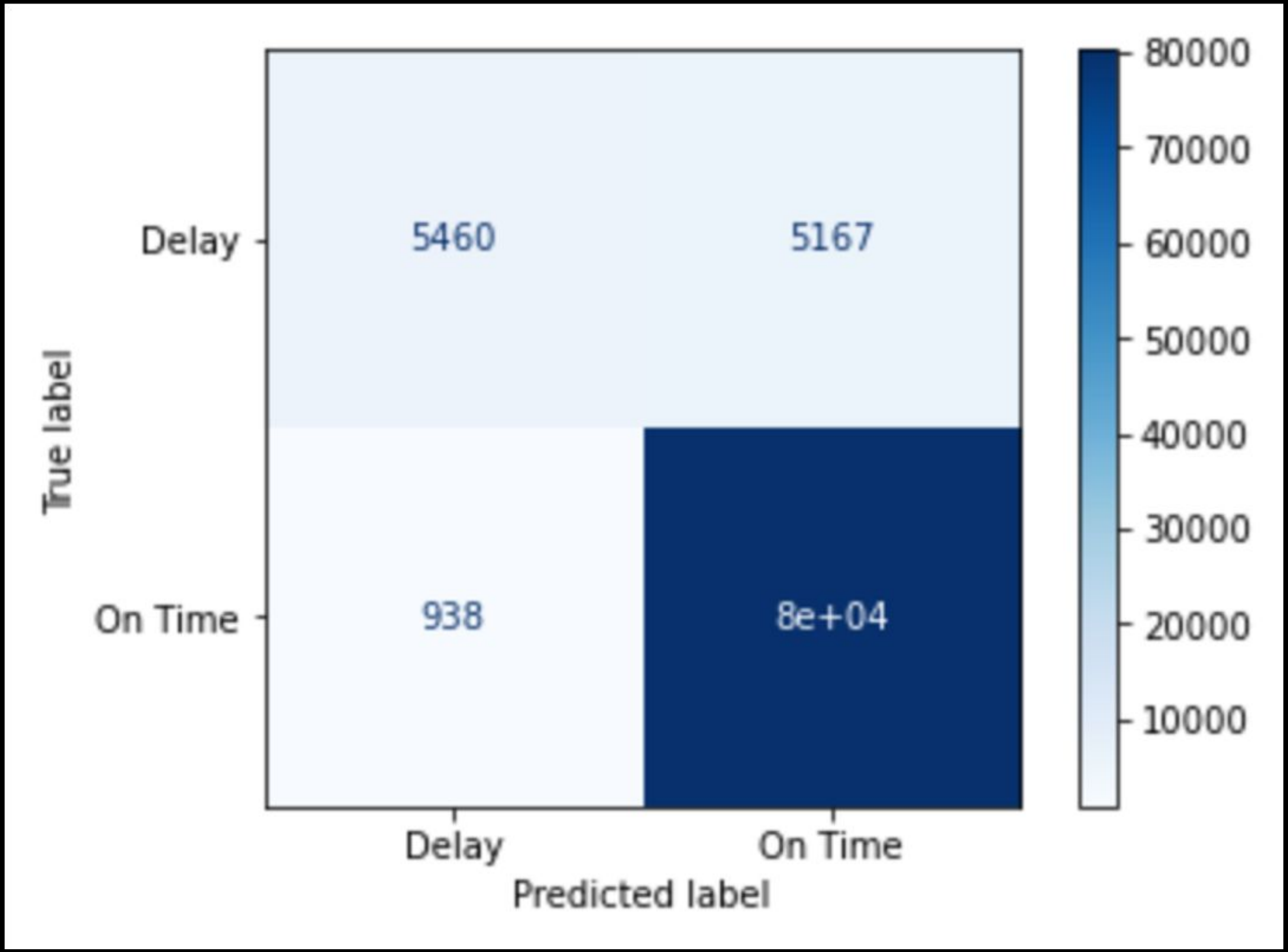
	precision	recall	f1-score	support
Delay	0.85	0.51	0.64	10627
On Time	0.94	0.99	0.96	81119
accuracy			0.93	91746
macro avg	0.90	0.75	0.80	91746
weighted avg	0.93	0.93	0.93	91746

Random Forest

Random Forest Visualization



Random Forest Confusion Matrix

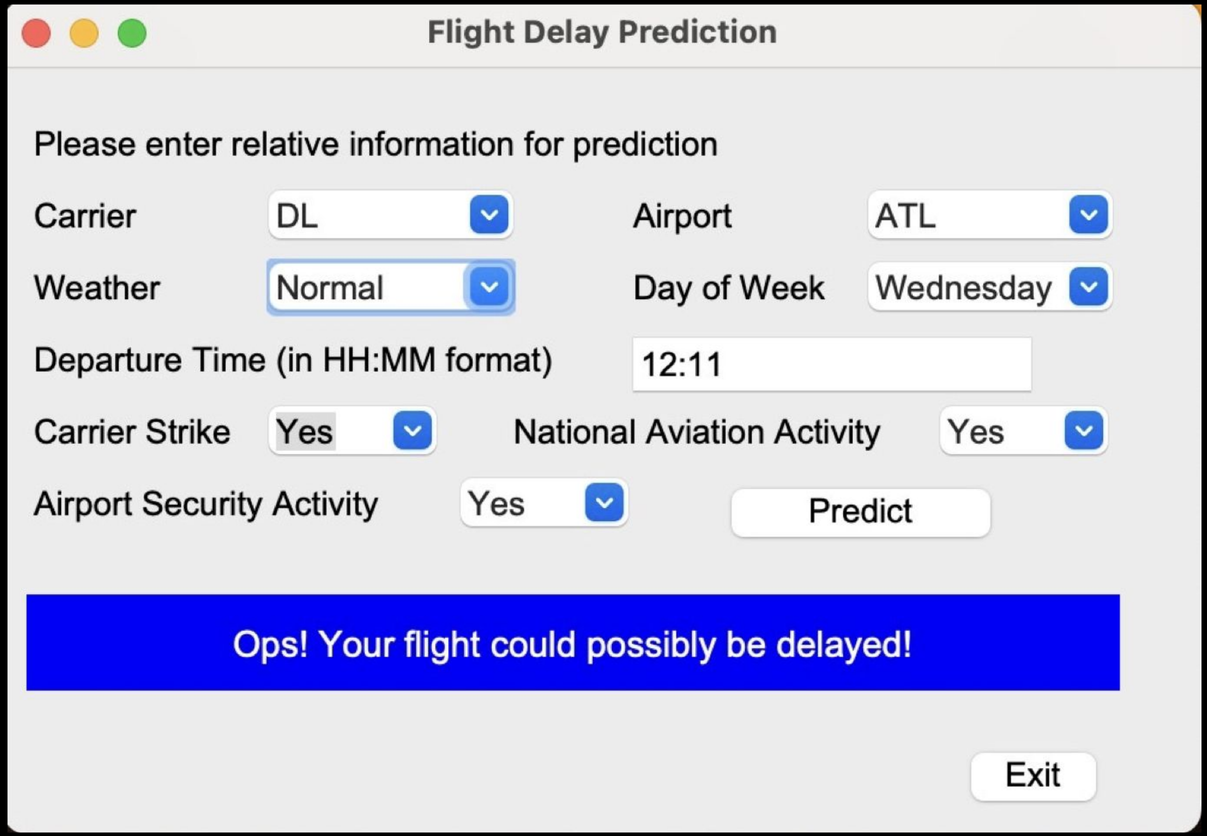


Stakeholder Engagement

Stakeholder Engagement

- We designed a UI for users to predict their flights delay or not.

- As we can see in the picture, we select delta airline, from Atlanta airport, the weather is normal and its Wednesday noon, the prediction is delay! [5]



Conclusion

Analysis Conclusion

- All airlines have a more than 80% on time rate.
- Hawaiian Airlines has the best performance in punctuality, meanwhile the Allegiant Air has the worst on time rate.
- In different weather conditions, the overall flight on time rate are still more than 80%.
- While most delays occur between 10 a.m. and 9 p.m. during the day, the peak occurs around 7 p.m. There was little difference in delay performance between days of the week

Model Conclusion

- Performed data preprocessing
- Created dummy variables
- Different Methods and Metrics
- Compared different model approaches performance: Random Forest performs the best

Project Conclusion

- Real-life business question
- Data from United States Department of Transportation and National Weather Service
- No Privacy Concern
- Business Insights and application from analysis and prediction model
- For Stakeholder Engagement, Delivered a user interface for customers

References

1. Understanding the Reporting of Causes of Flight Delays and Cancellations
2. Bureau of Transportation Statistics (BTS)
3. S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016, pp. 1-6, doi: 10.1109/DASC.2016.7777956.
4. Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, Gang Chen, “Flight delay prediction for commercial air transport: A deep learning approach,” Transportation Research Part E: Logistics and Transportation Review, Volume 125, 2019, Pages 203-221, ISSN 1366-5545.
5. G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140-150, Jan. 2020, doi: 10.1109/TVT.2019.2954094.