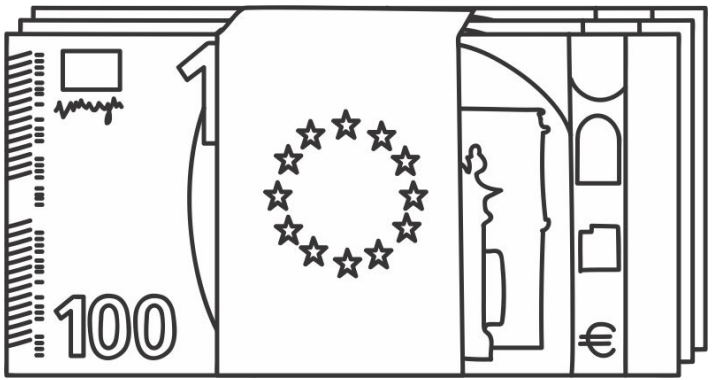
An Investigation into

Flight Delays

Rachael Alexandroff, Sofia Pignataro, Racquel Fygenson, Ruxin Shen
Group 13

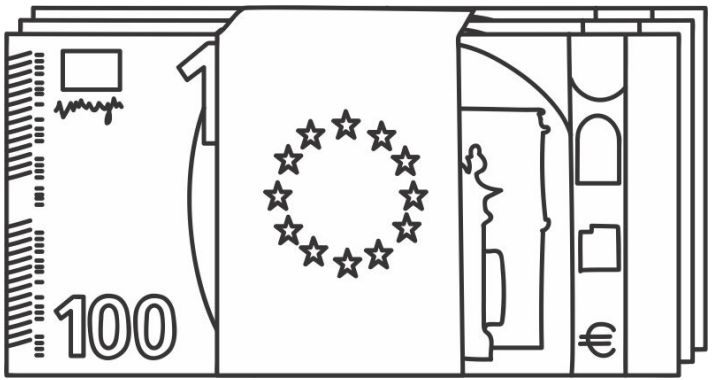
Flight delays are a hassle.

Flight delays are a hassle



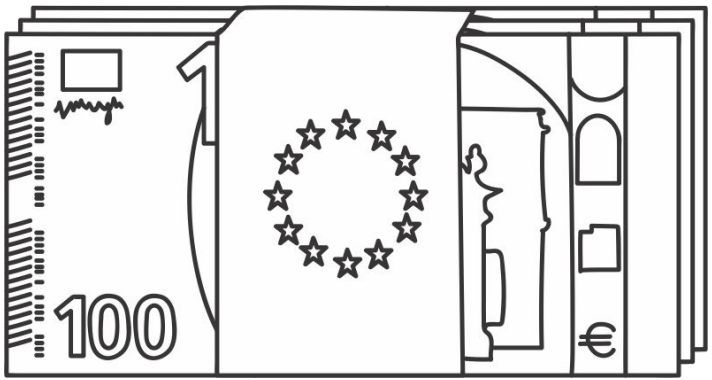
EU regulation 261/2004 requires
airlines to give you money
if your flight is
delayed > 3 hours!

Flight delays are a hassle

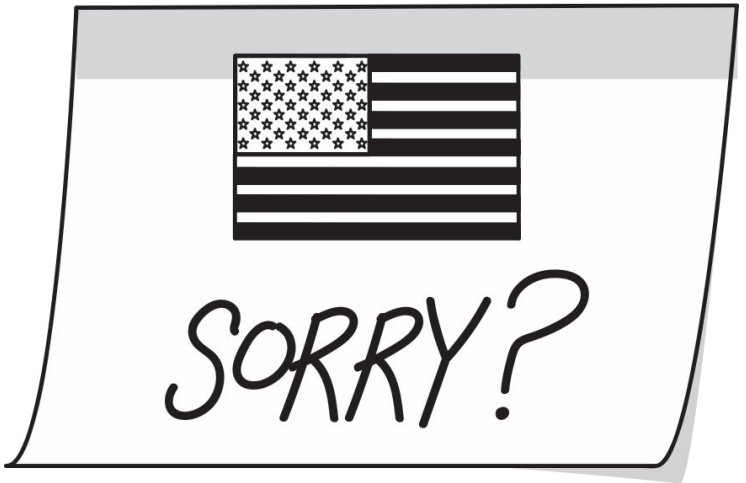


What about in the US?

Flight delays are a hassle



VS

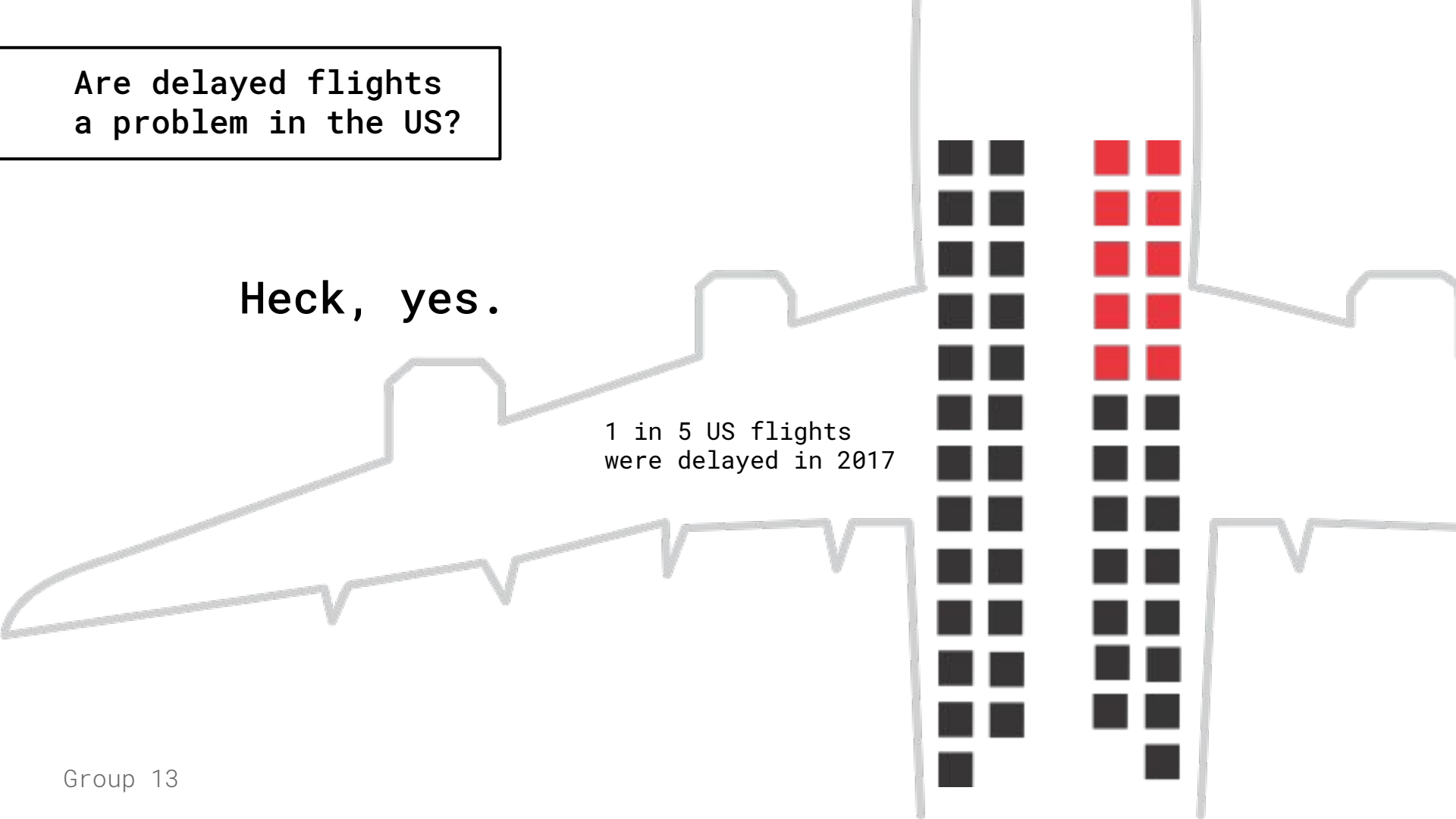


Flight delays are a hassle

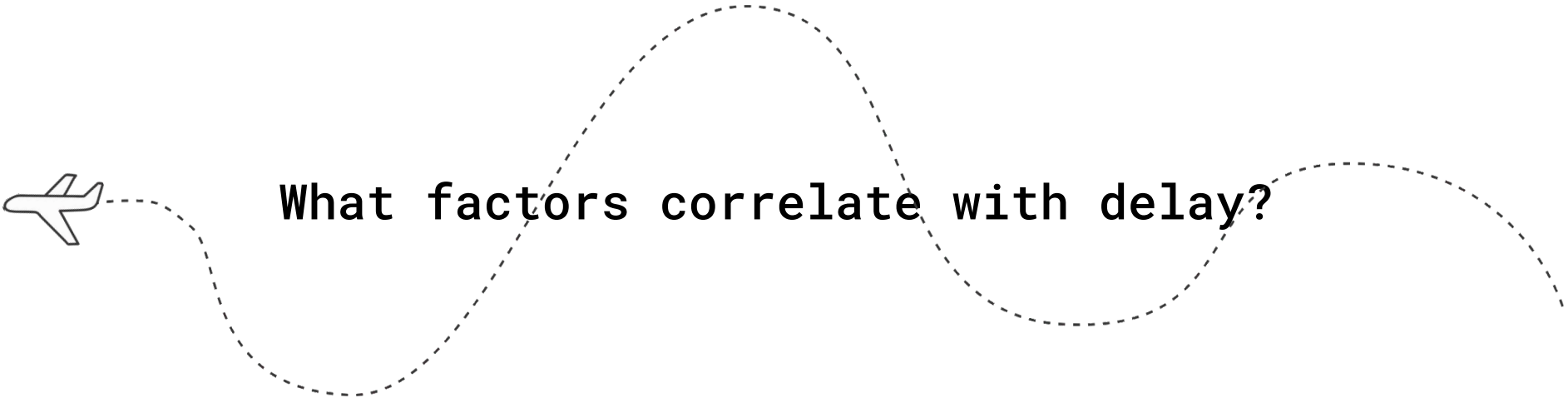
Are delayed flights a problem in the US?

Are delayed flights a problem in the US?

Heck, yes.

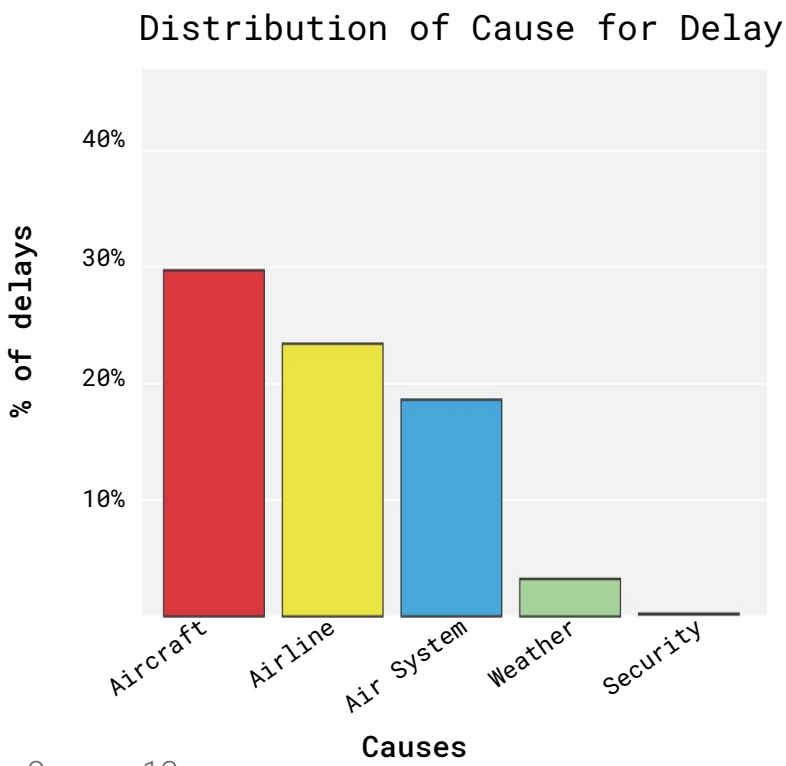


1 in 5 US flights were delayed in 2017

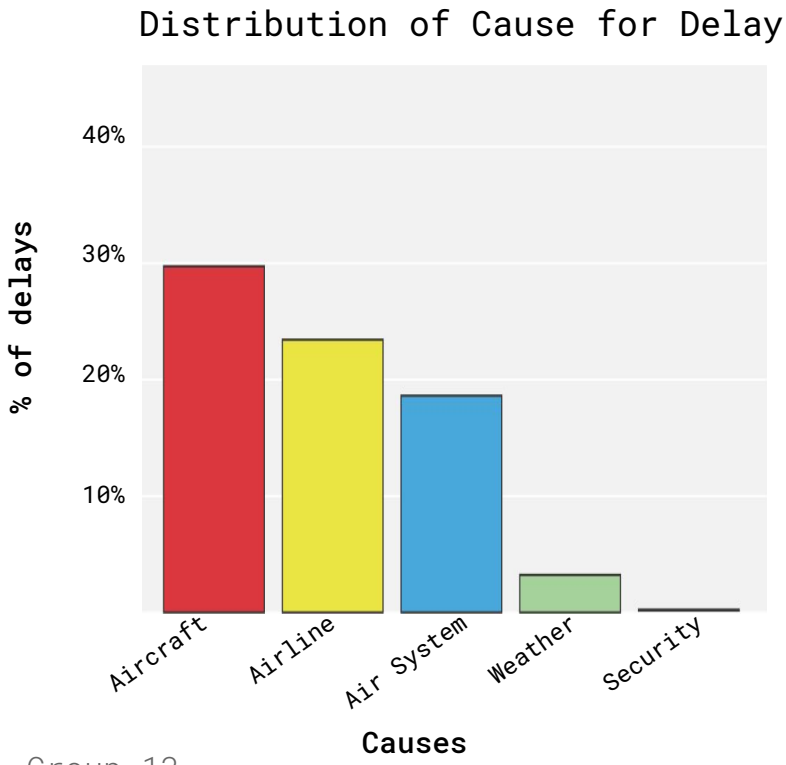


What factors correlate with delay?

What factors affect delay?

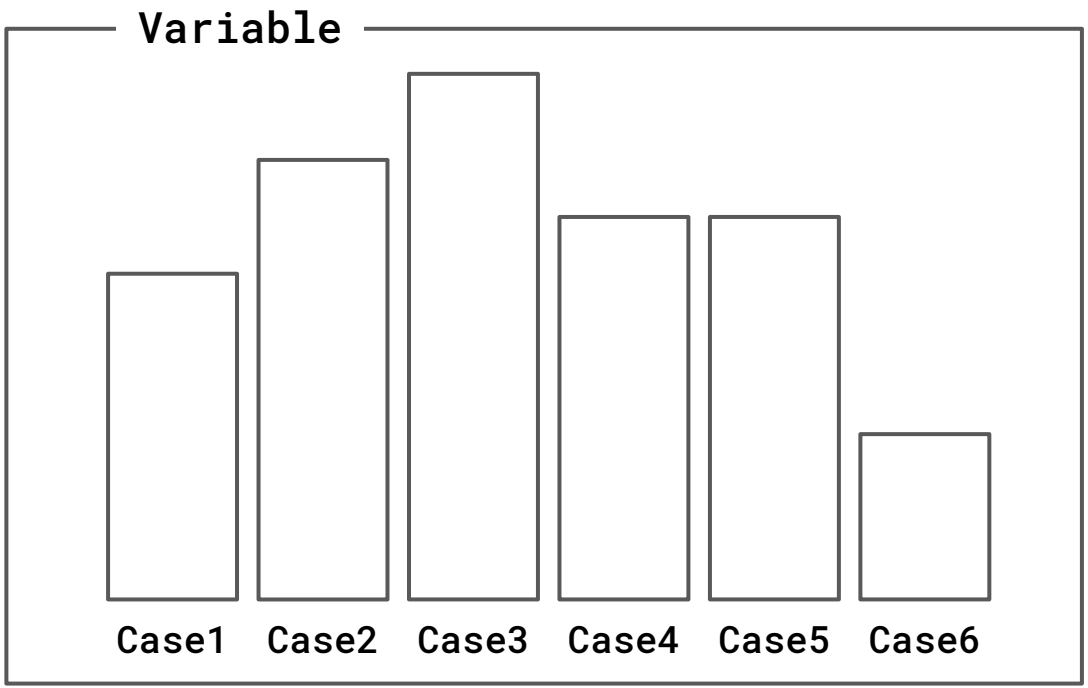


What factors affect delay?



What factors affect delay?

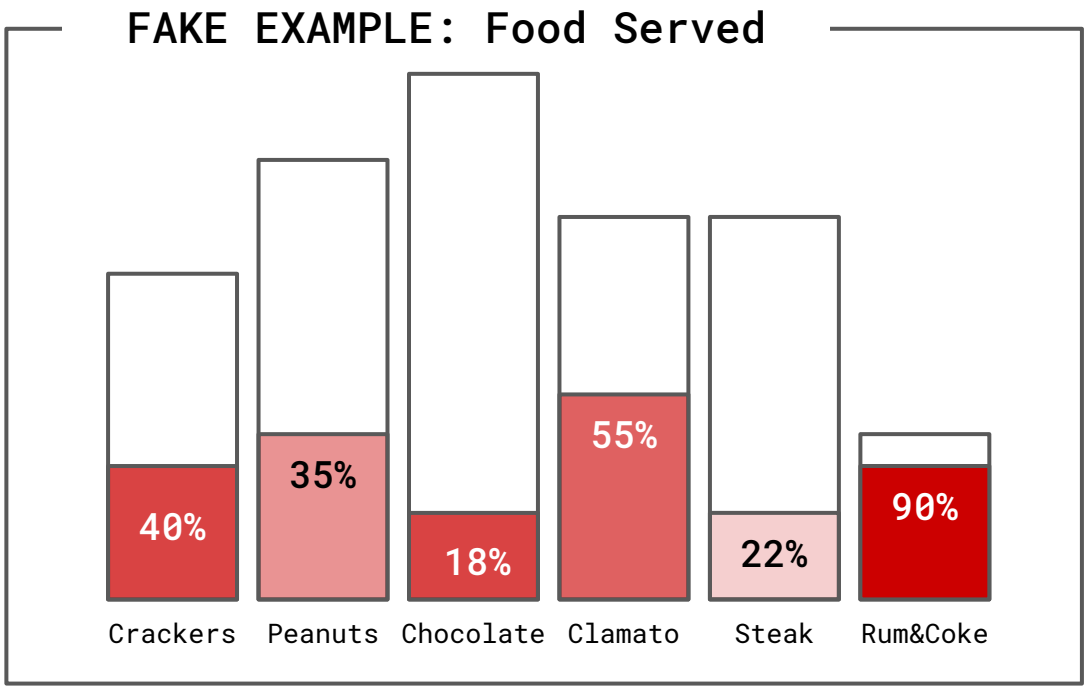
Proportional Summary*



Let's **separate by variables** we think might be relevant

What factors affect delay?

Proportional Summary*

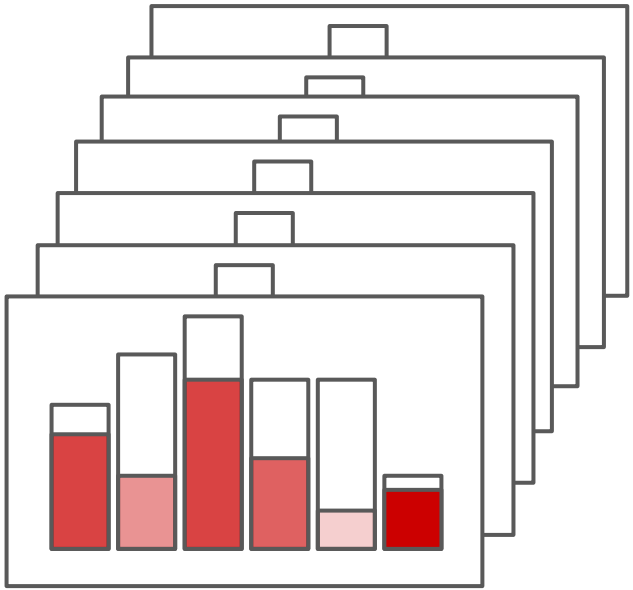


Let's **separate by variables** we think might be relevant

and look at the **proportion** of flights that are delayed

What factors affect delay?

Proportional Summary*

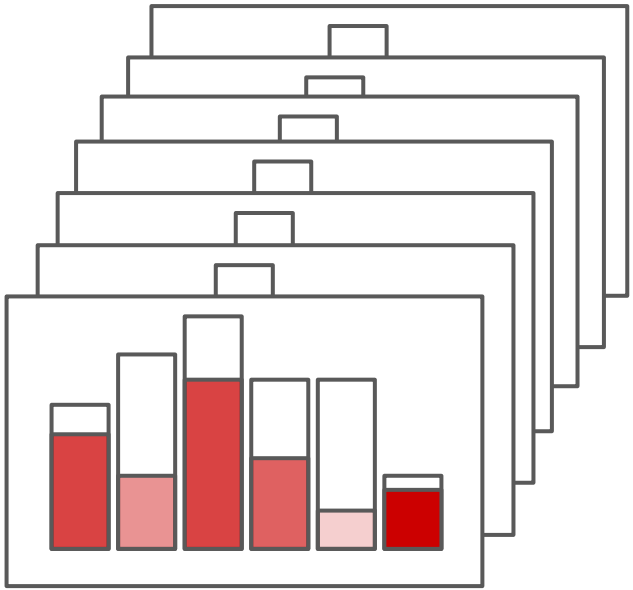


Variables we evaluated

- ➔ Day of the week
- ➔ Month of the year
- ➔ Time of day (4 buckets)
- ➔ Elapsed flight time
- ➔ Distance of flight
- ➔ Airlines
- ➔ Season

What factors affect delay?

Proportional Summary*

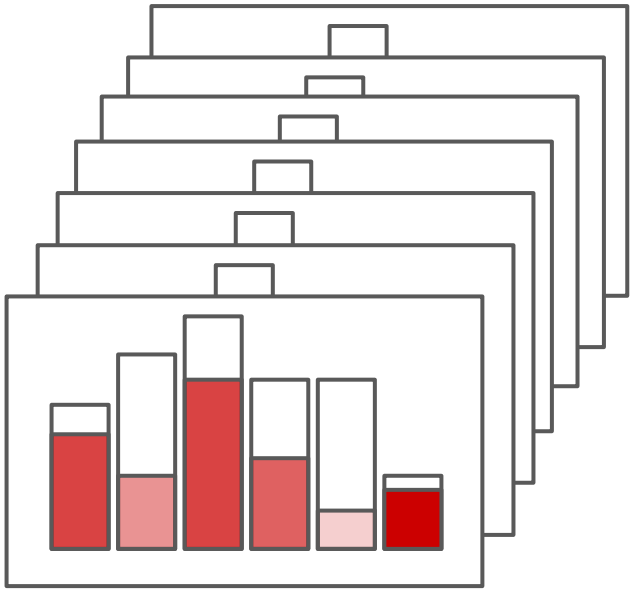


Variables we evaluated

- ➔ Day of the week
- ➔ Month of the year
- ➔ Time of day (4 buckets)
- ➔ Elapsed flight time
- ➔ Distance of flight
- ➔ Airlines (2 buckets)
- ➔ Season

What factors affect delay?

Proportional Summary*



Data Used

Flight Traffic

Data Not Used

Outside scope of question →

- Weather
- Fare
- Event

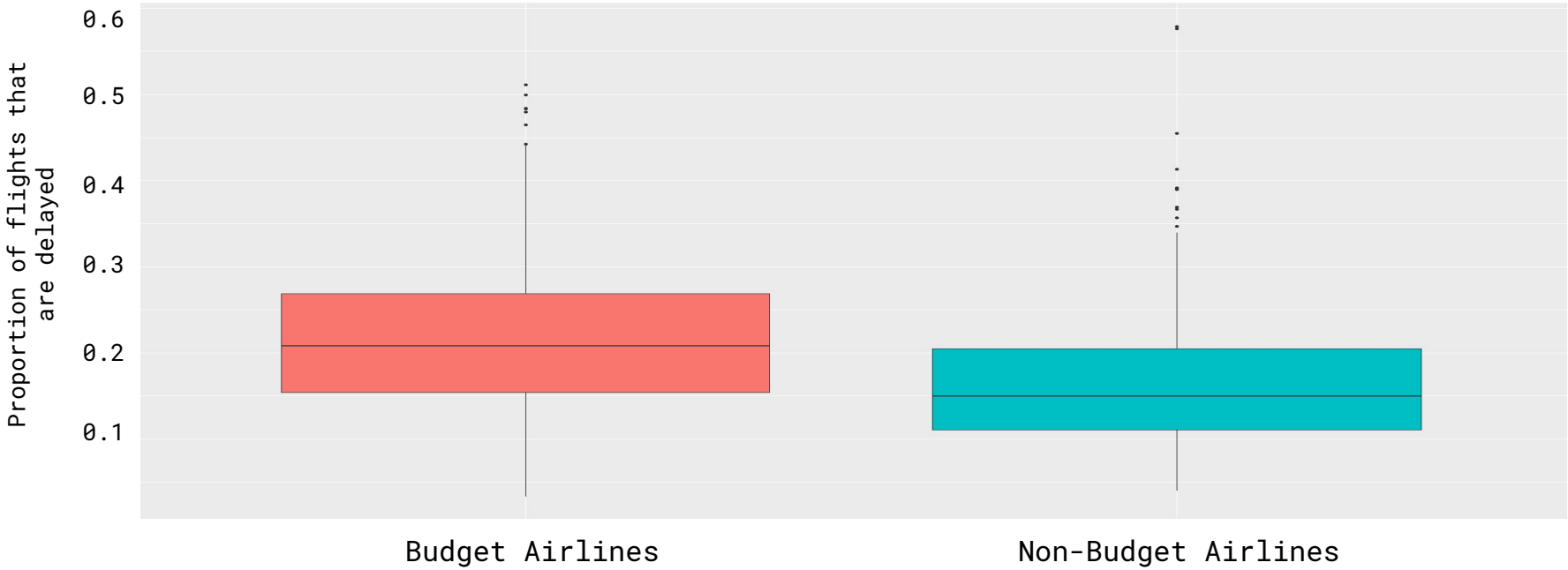
← Do not contain important data points (like ORD or DFW) that are represented in our main dataset

Variables we evaluated

- ➔ Day of the week
- ➔ Month of the year
- ➔ Time of day (4 buckets)
- ➔ Elapsed flight time
- ➔ Distance of flight
- ➔ Airlines (2 buckets)
- ➔ Season

What factors affect delay?

Hypothesis Testing



What factors affect delay?

Hypothesis Testing

Airline Type

	Budget
Not Budget	S

S	Significant
NS	Not Significant

H_0 = Proportion of delayed flights are equal

H_a = Proportion of delayed flights are not equal

"Budget" Airlines:

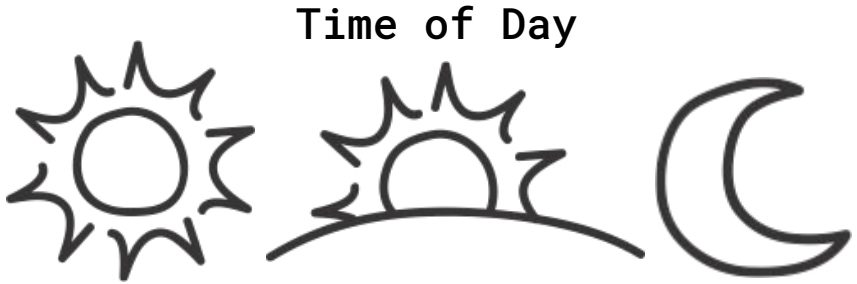
- ➔ Spirit
- ➔ JetBlue
- ➔ ExpressJet
- ➔ Frontier
- ➔ SkyWest
- ➔ Southwest
- ➔ Virgin

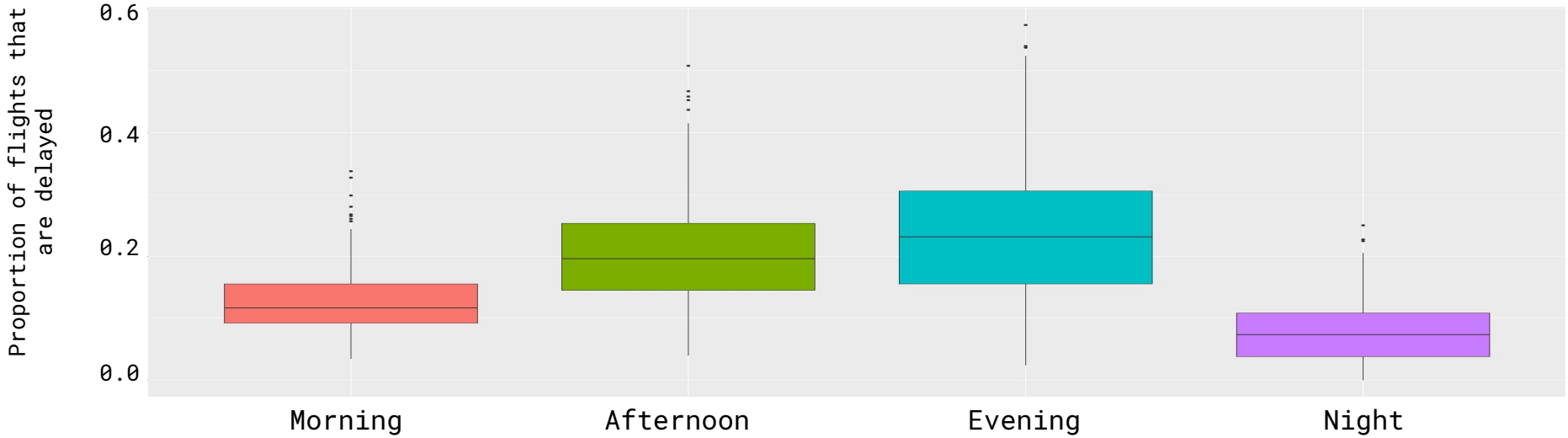
"Non-Budget" Airlines:

- ➔ American
- ➔ Delta
- ➔ Hawaiian Air
- ➔ United
- ➔ Alaska Air

What factors affect delay?

Hypothesis Testing





What factors affect delay?

Hypothesis Testing

Time of Day

	Morning	Afternoon	Evening
Afternoon	S		
Evening	S	S	
Night	S	S	S

H_0 = Proportion of delayed flights are equal

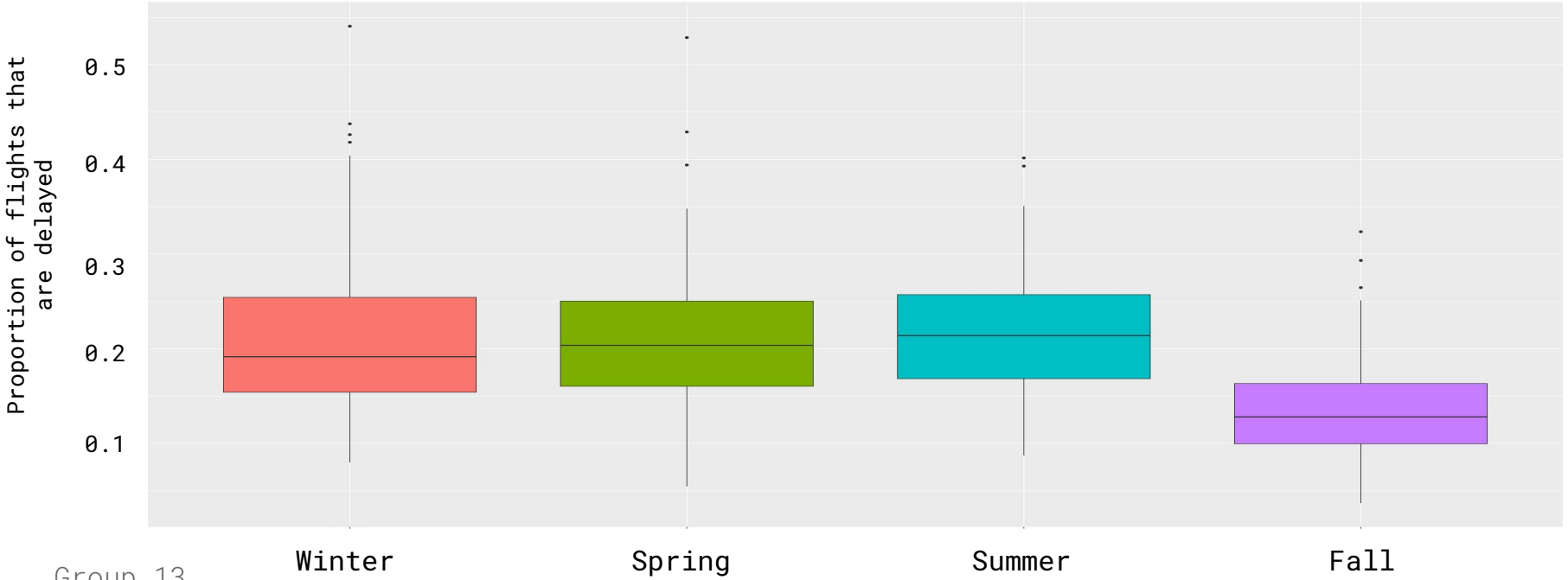
H_a = Proportion of delayed flights are not equal

S	Significant
NS	Not Significant

What factors affect delay?

Hypothesis Testing

Season



What factors affect delay?

Hypothesis Testing

Season

	Spring	Summer	Autumn
Summer	NS		
Autumn	S	S	
Winter	NS	NS	S

H_0 = Proportion of delayed flights are equal

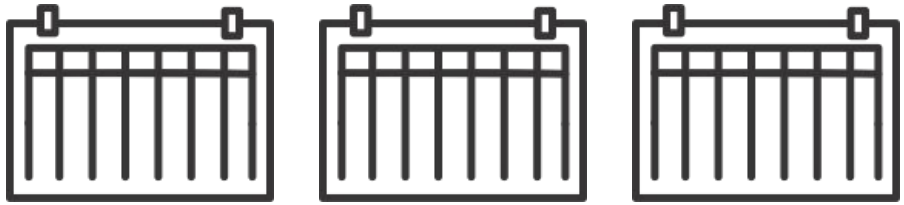
H_a = Proportion of delayed flights are not equal

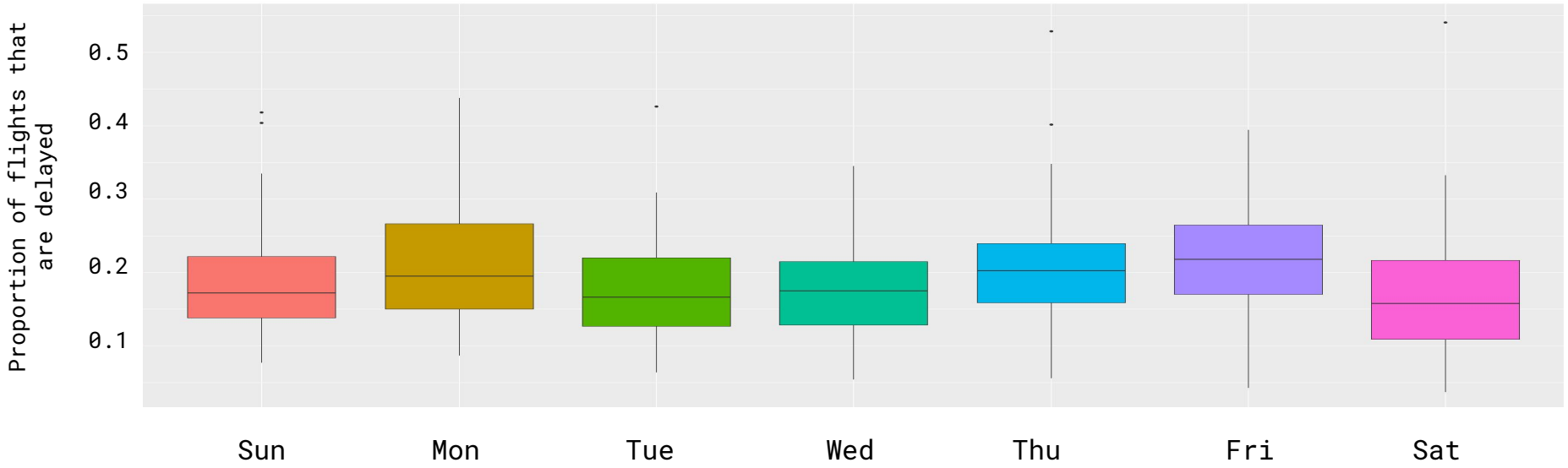
S	Significant
NS	Not Significant

What factors affect delay?

Hypothesis Testing

Day of the Week

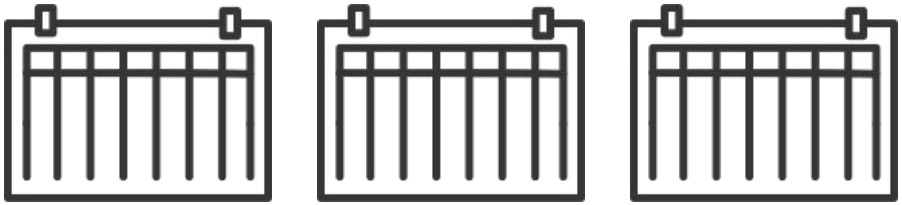




What factors affect delay?

Hypothesis Testing

Day of the Week



	Sun	Mon	Tue	Wed	Thu	Fri
Mon	S					
Tue	NS	S				
Wed	NS	S	NS			
Thu	S	NS	S	S		
Fri	S	NS	S	S	S	
Sat	S	S	S	S	S	S

H_0 = Proportion of delayed flights are equal

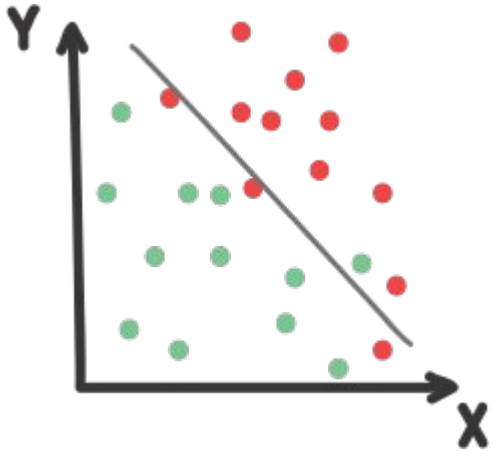
H_a = Proportion of delayed flights are not equal

S	Significant
NS	Not Significant

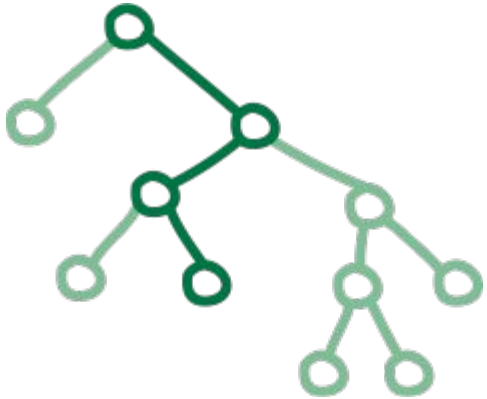
Can we predict delay?

Machine Learning

Benchmark:
Logistic Regression



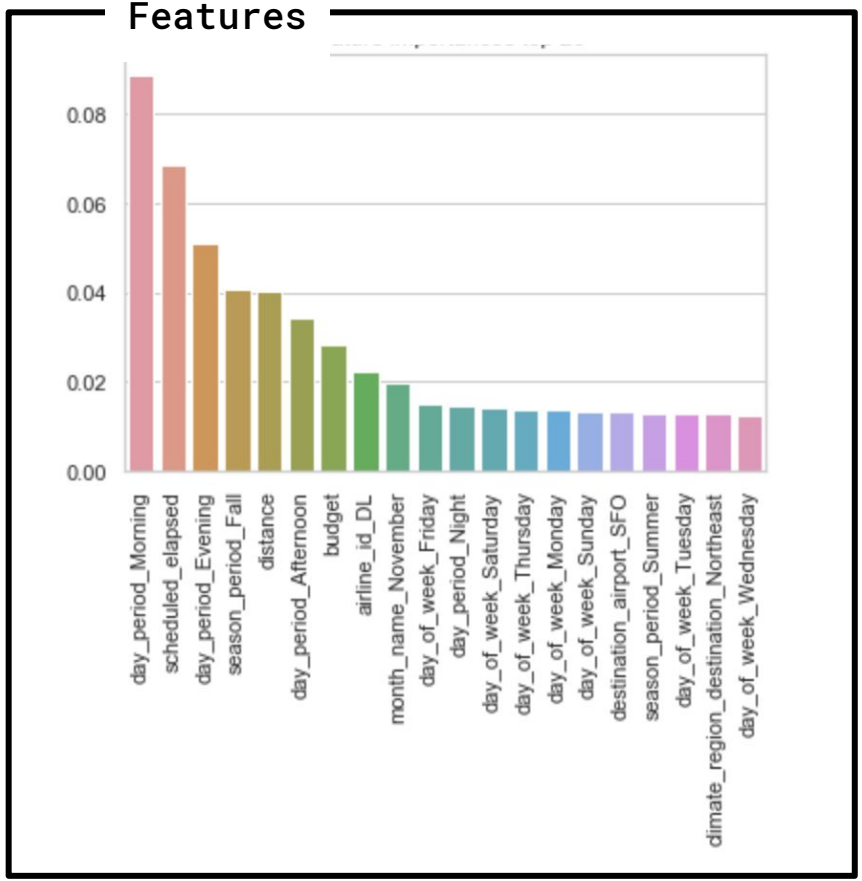
Modeling:
Random Forest



Can we predict instance of delay?

Random Forest

Delay: Y/N?

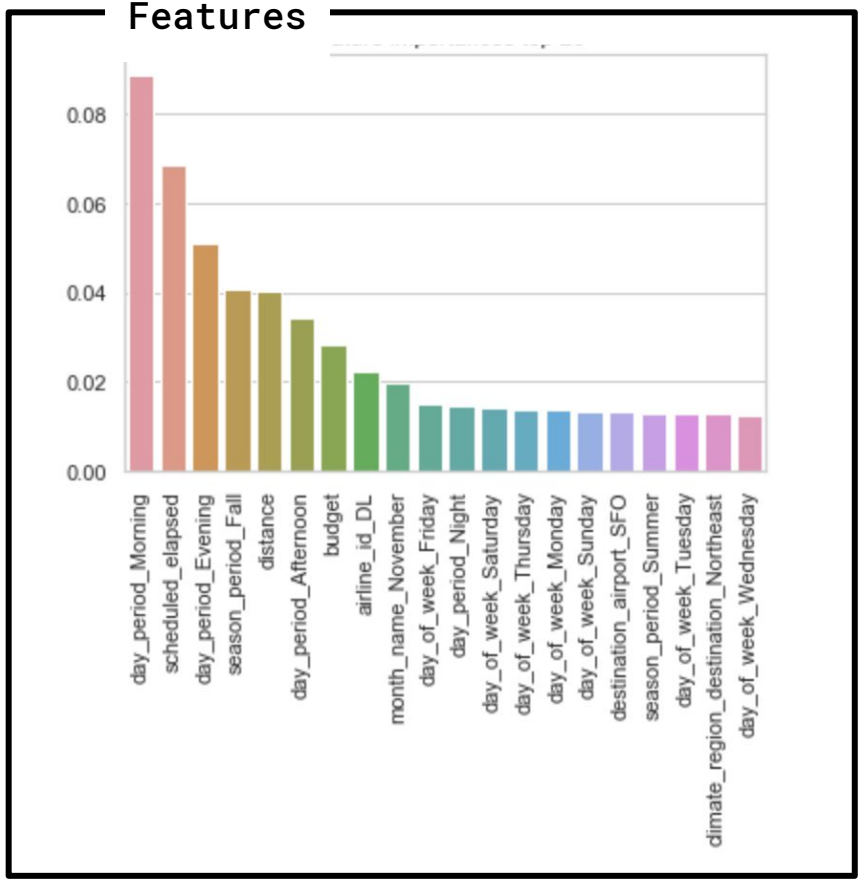


Can we predict instance of delay?

Random Forest

Delay: Y/N?

Time of day is an important feature



Can we predict instance of delay?

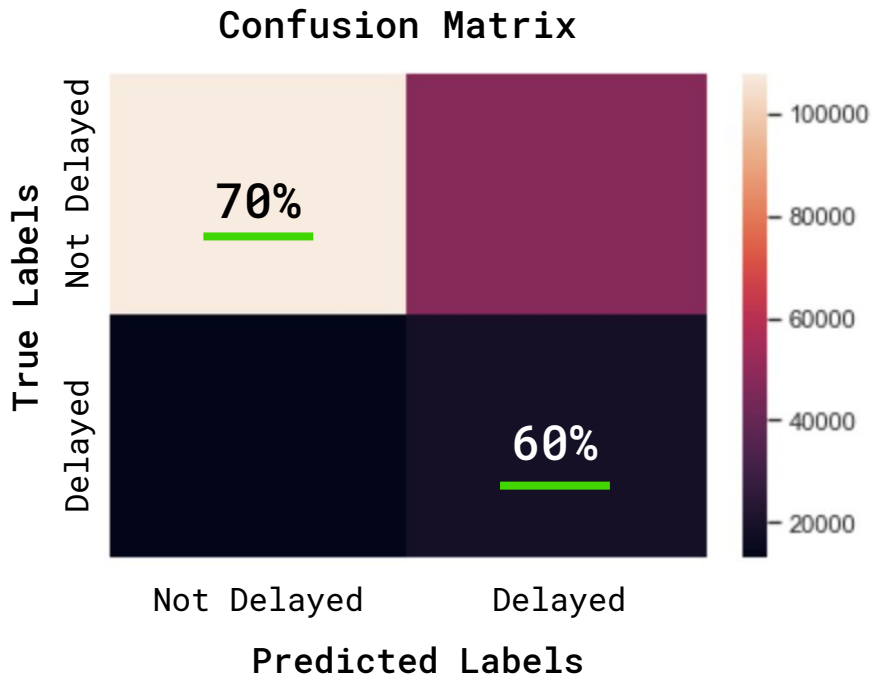
Random Forest

Delay: Y/N?

Right now, our model has a lot of false positives

Overall Accuracy: 68%

Recall



Can we predict instance of delay?

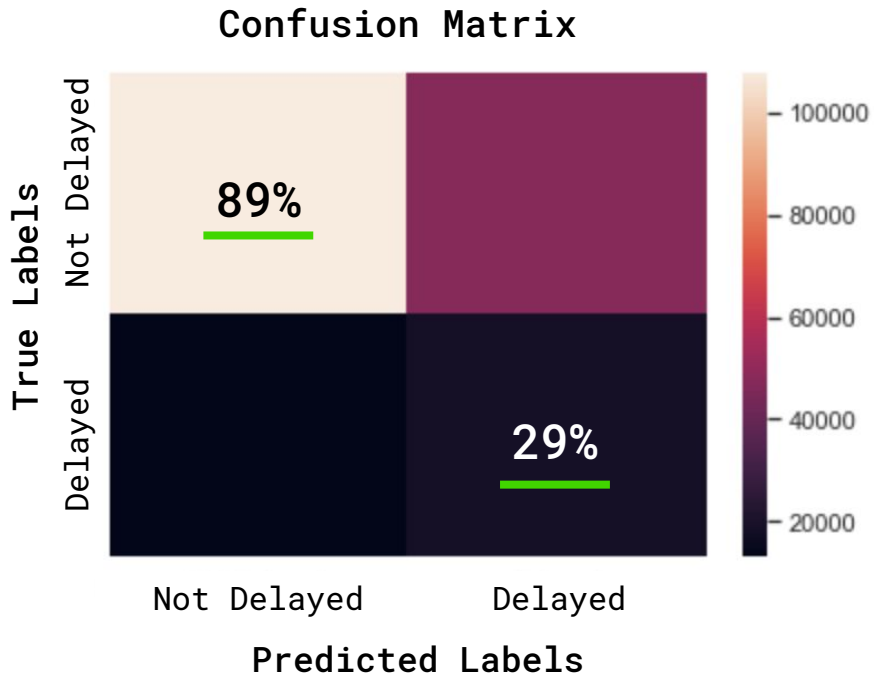
Random Forest

Delay: Y/N?

Right now, our model has a lot of false positives

Overall Accuracy: 68%

Precision

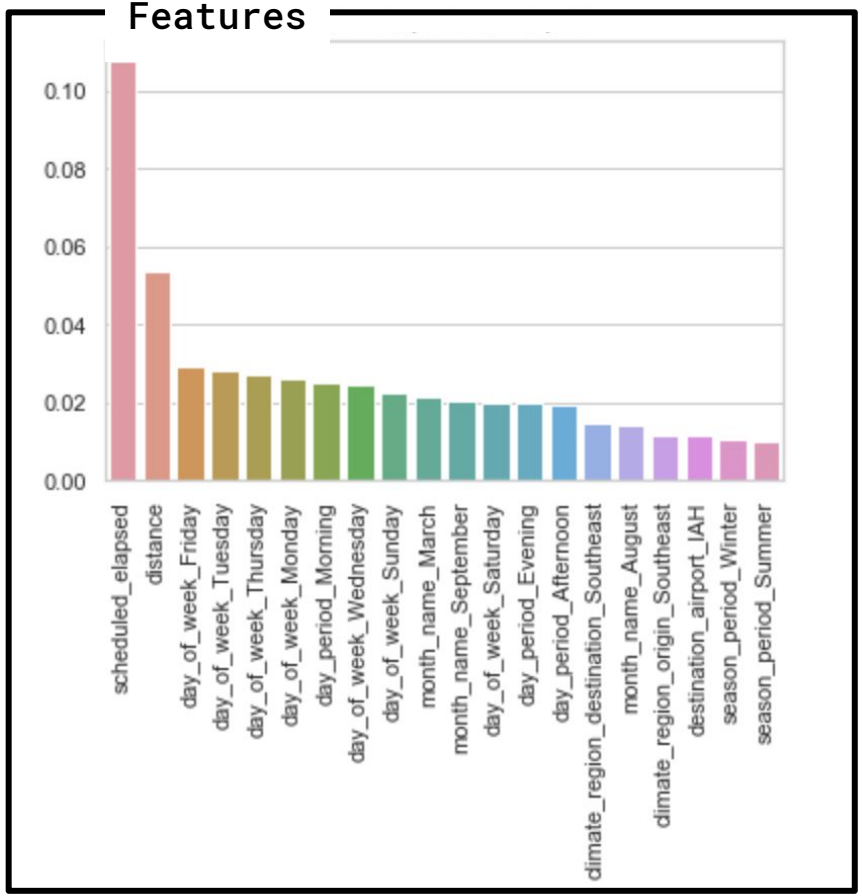


Can we predict length of delay?

Random Forest

Length of Delay?

Flight duration,
flight distance,
day of week,
and month are
important features

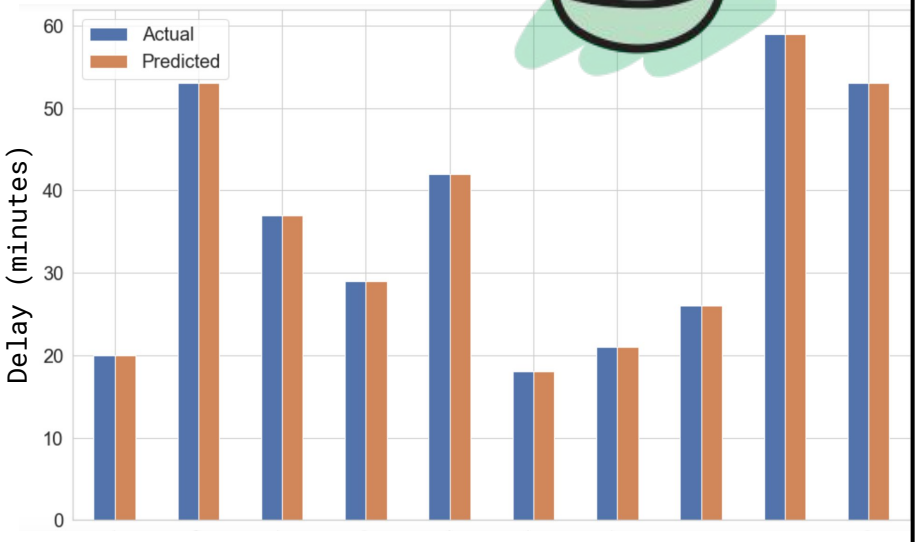


Can we predict length of delay?

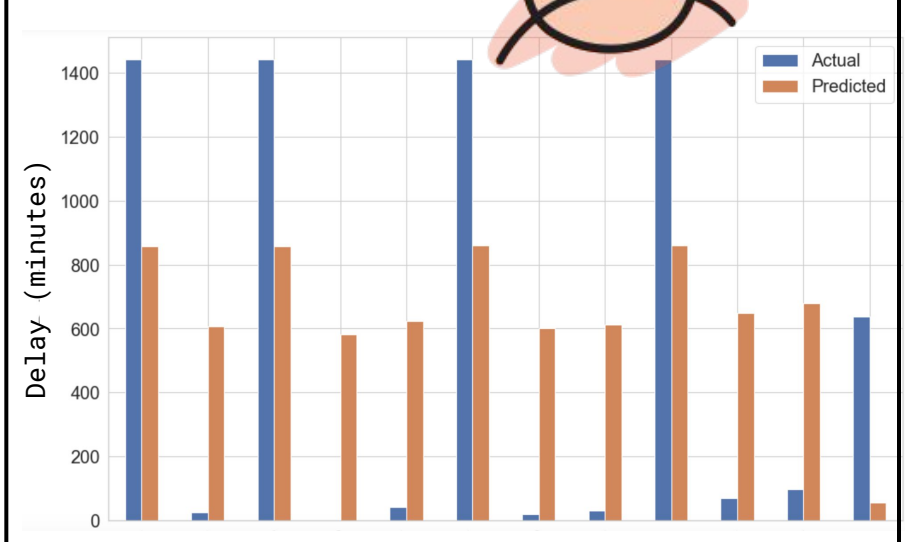
Short Delays

Can we predict length of delay?

Short Delays



Long Delays



Practical Application

(So What?)

Alert shoppers when a flight is at risk of being delayed.

Practical Application

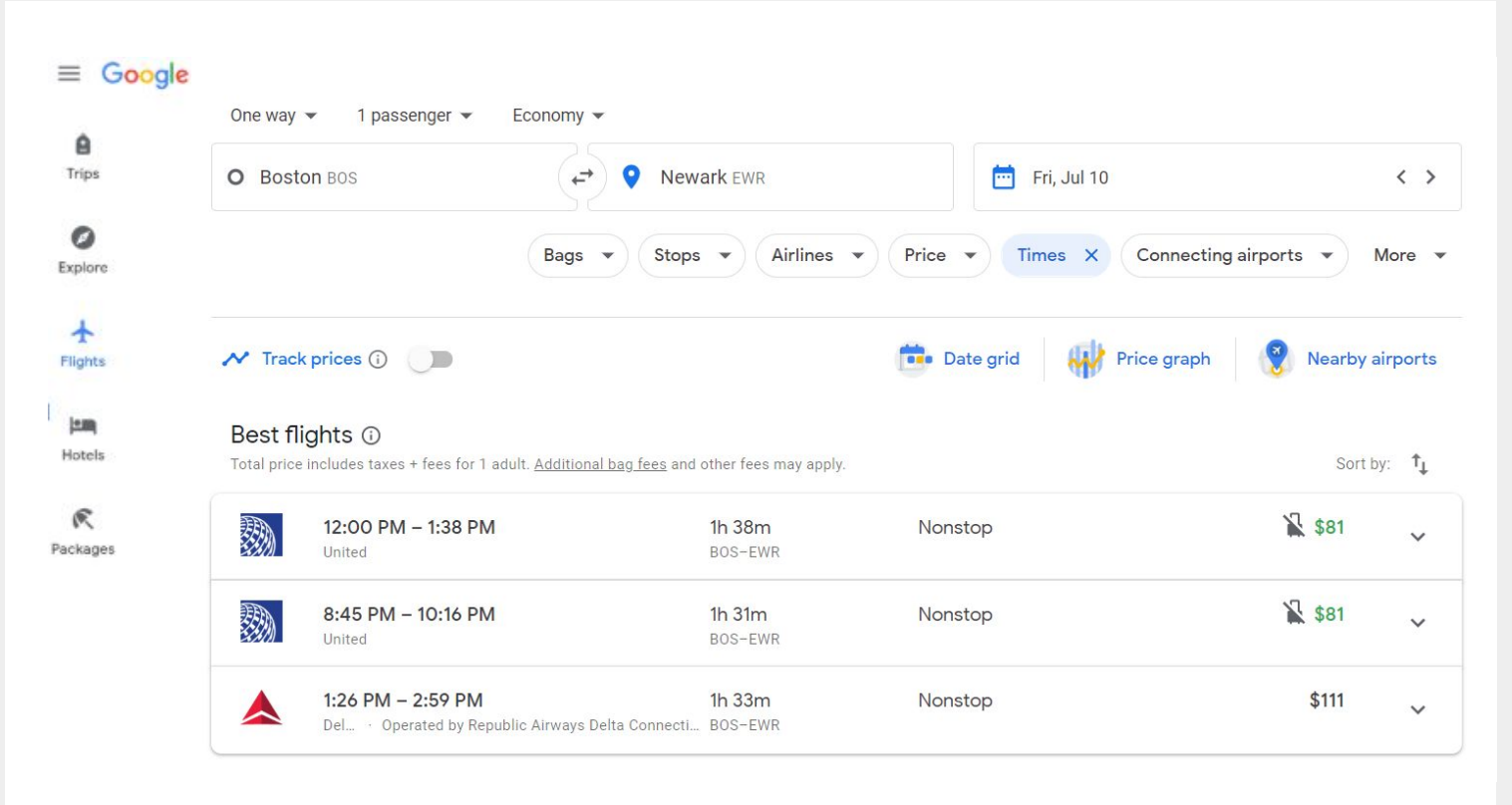
(So What?)

Alert shoppers when a flight is at risk of being delayed.

Practical Application

(So What?)

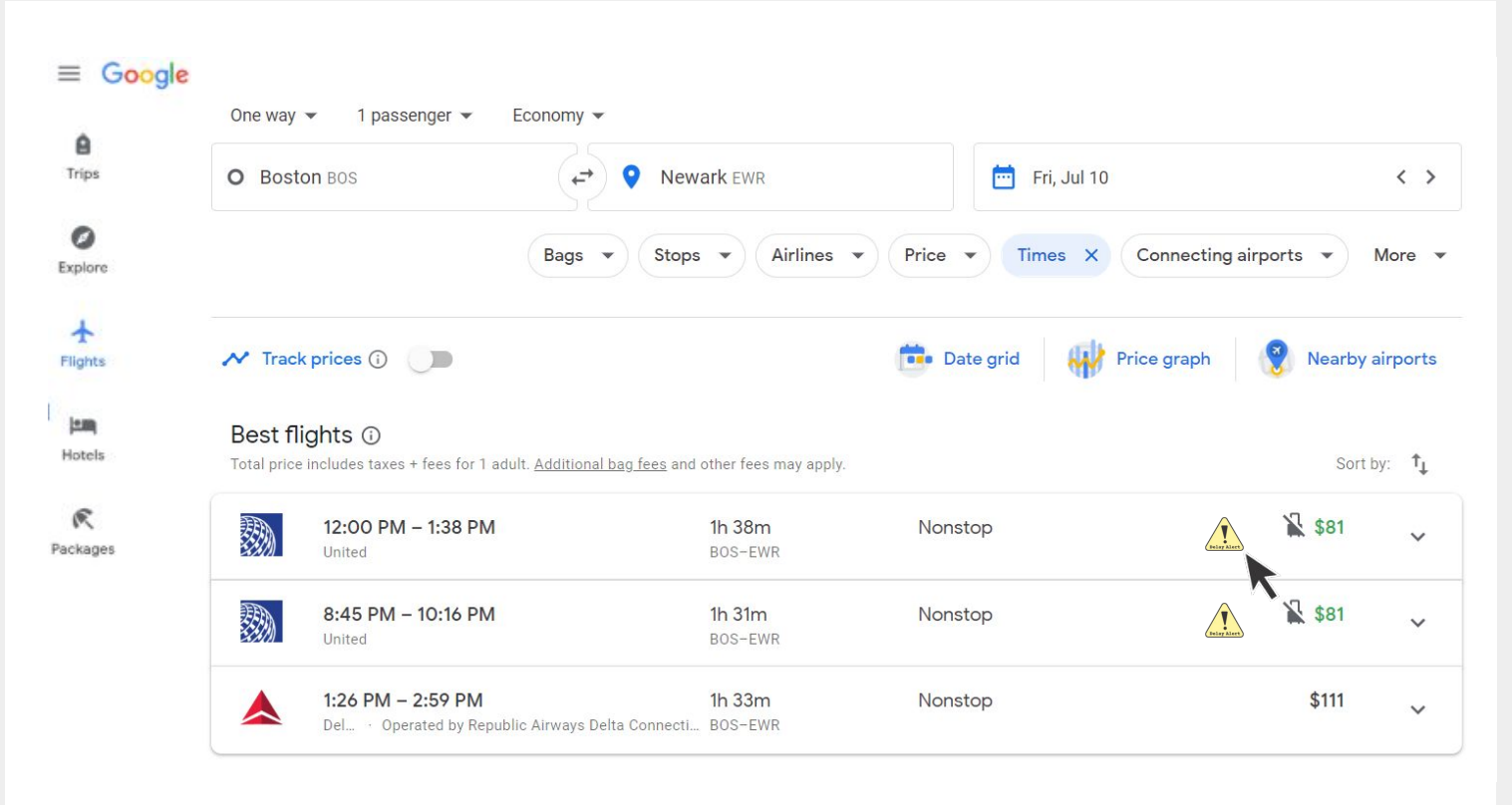
Alert shoppers when a flight is at risk of being delayed.



Practical Application

(So What?)

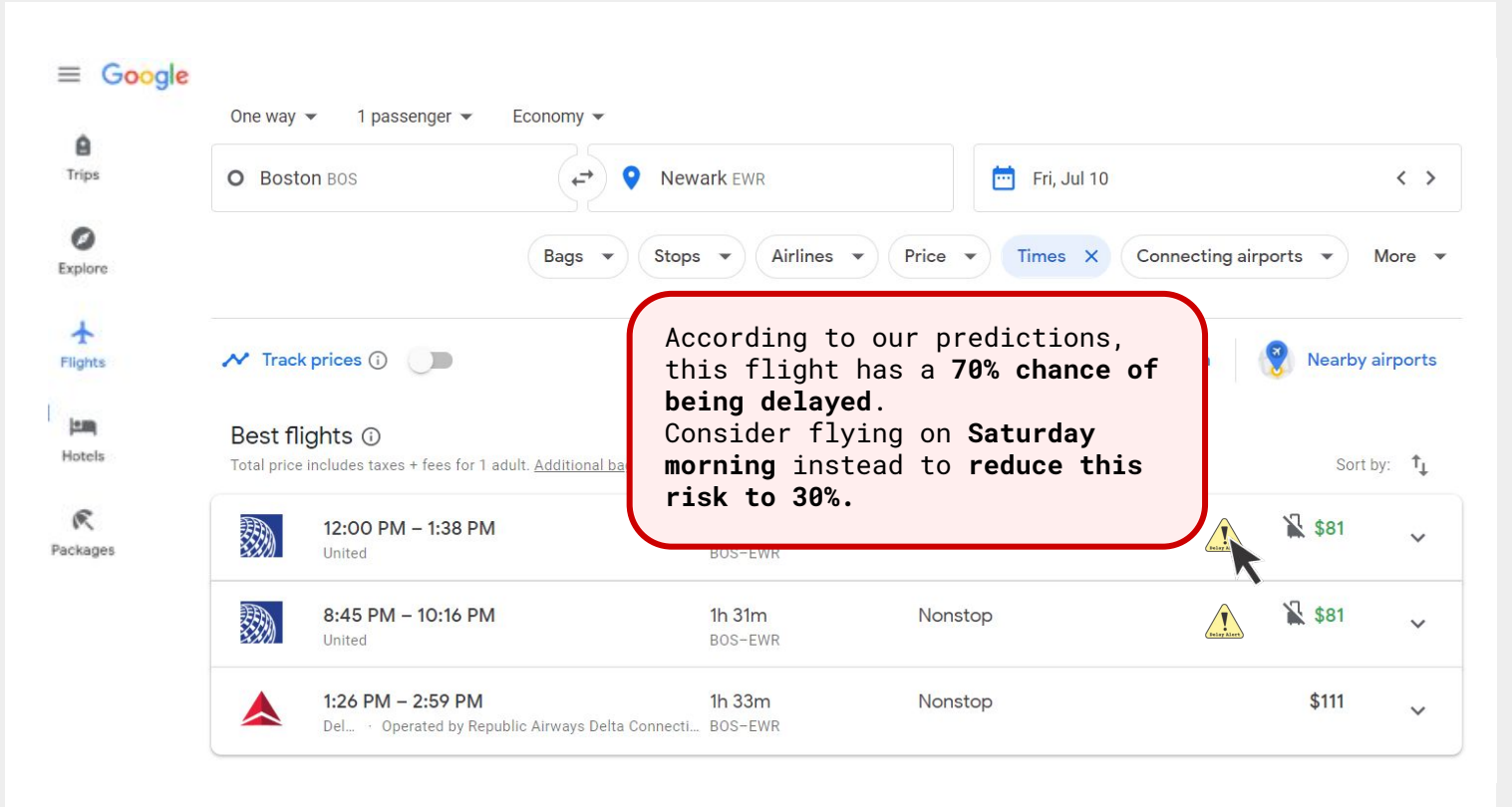
Alert shoppers when a flight is at risk of being delayed.



Practical Application

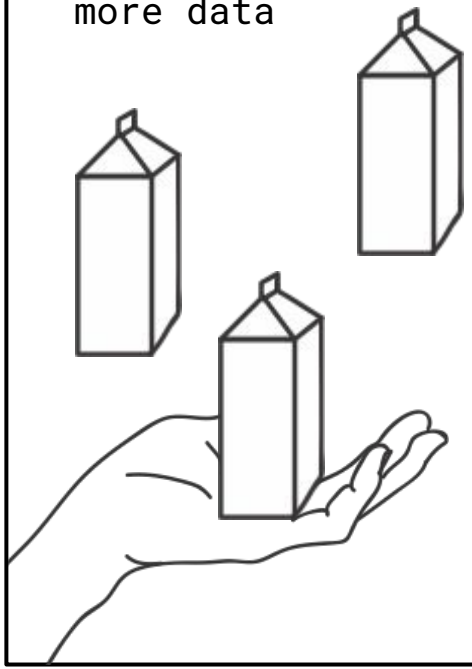
(So What?)

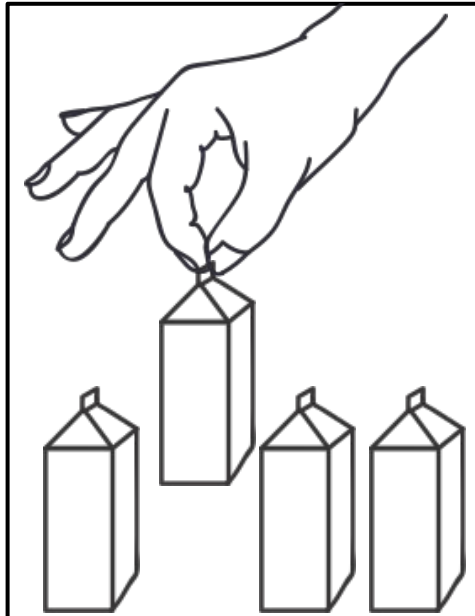
Alert shoppers when a flight is at risk of being delayed.



Things we could do better

Get our hands on more data





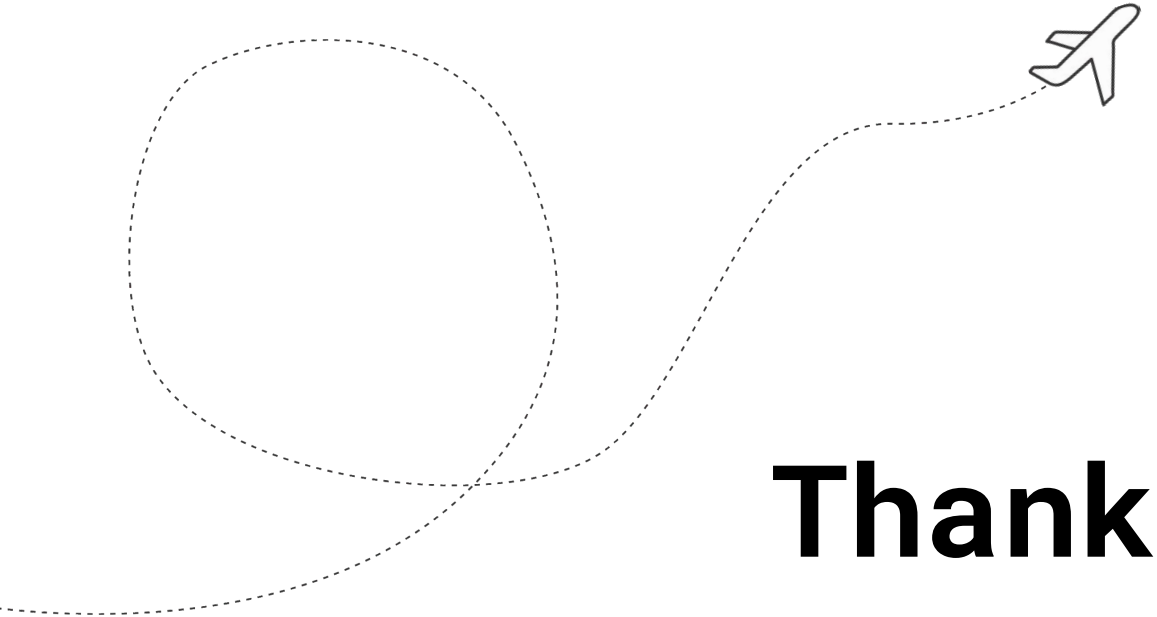
Optimize feature selection

Bucket delays by type:

- ➔ >3hrs (“Catastrophic”)
- ➔ <3hrs (“Not Catastrophic”)
- ➔ Cancelled

Bucket delays by cause:

- ➔ Airline
- ➔ Weather
- ➔ Air System
- ➔ Aircraft
- ➔ Security



Thank you

A quick overview...

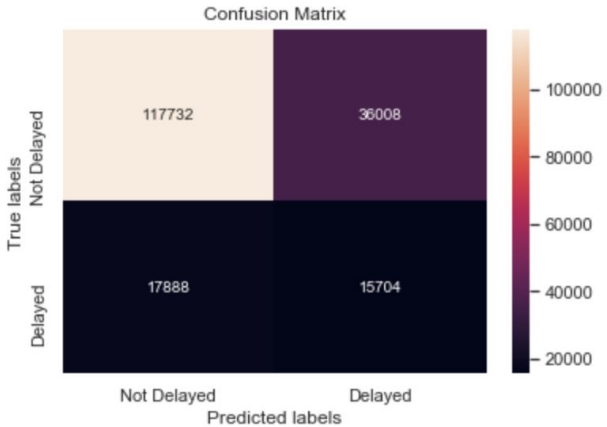
Time of day and **length of flight** most affect whether a flight will be delayed.

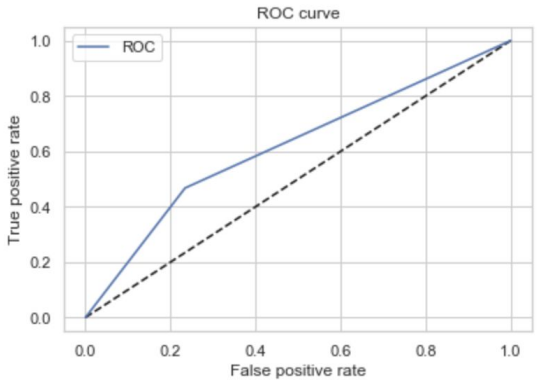
Our model recall is **60% for delayed** flights and **70% for non-delayed** flights.

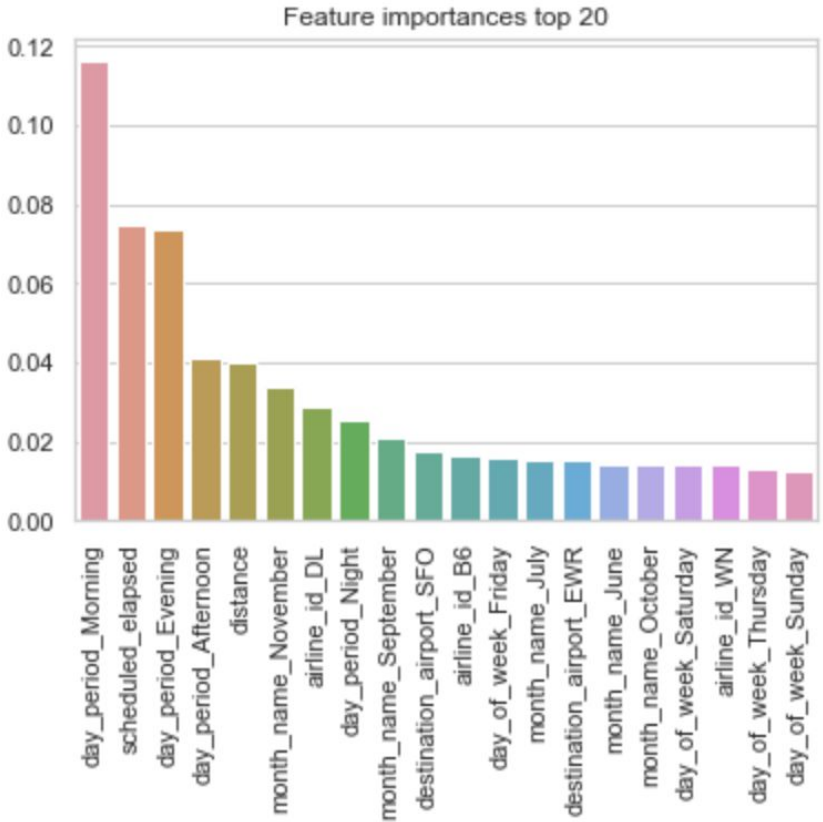
To improve we would use **more data**, optimize **feature selection** and investigate differences between **types of delays.**

... to open for questions :)

Outputs :\


Confusion Matrix
117732
36008
17888
15704
True labels
Not Delayed
Delayed
Not Delayed
Delayed
Predicted labels
100000
80000
60000
40000
20000


ROC curve
ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

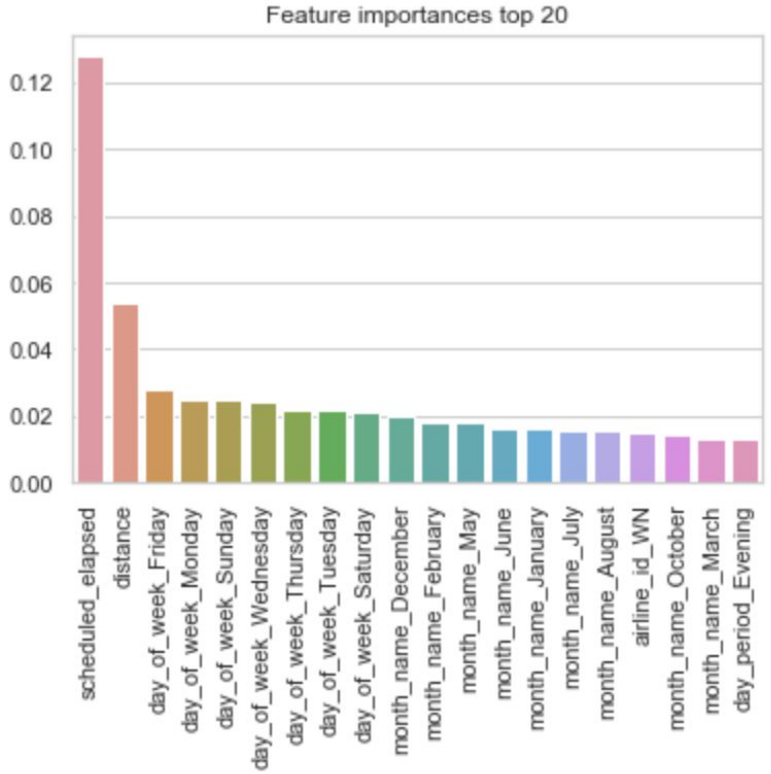

Feature importances top 20
0.12
0.10
0.08
0.06
0.04
0.02
0.00
day_period_Morning
scheduled_elapsed
day_period_Evening
day_period_Afternoon
distance
month_name_November
airline_id_DL
day_period_Night
month_name_September
destination_airport_SFO
airline_id_B6
day_of_week_Friday
month_name_July
destination_airport_EWR
month_name_June
month_name_October
day_of_week_Saturday
airline_id_WN
day_of_week_Thursday
day_of_week_Sunday

How Long would the Delay be ?

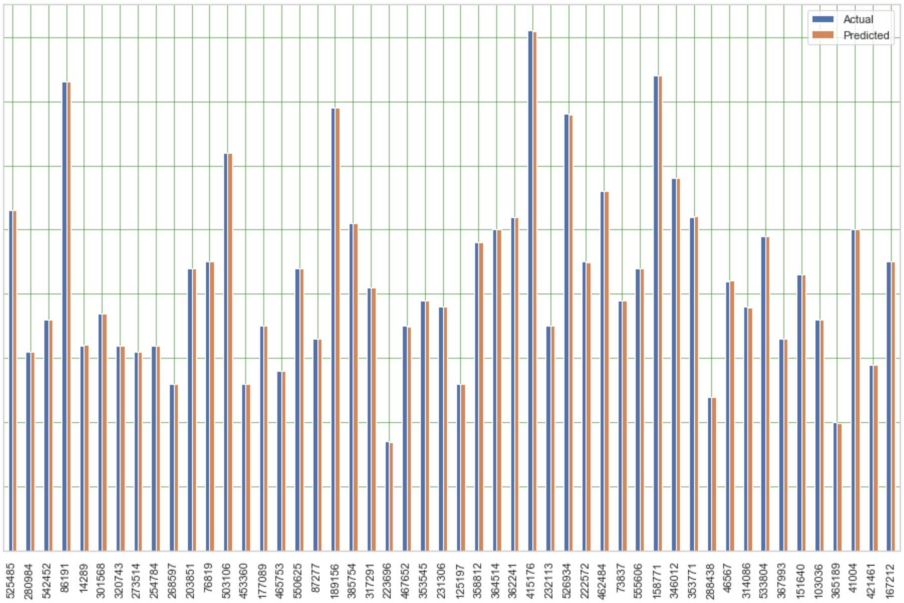
scheduled_elapsed
distance
day_of_week_Friday
day_of_week_Monday
day_of_week_Sunday
day_of_week_Wednesday
day_of_week_Thursday
day_of_week_Tuesday
day_of_week_Saturday
month_name_December
month_name_February
month_name_May
month_name_June
month_name_January
month_name_July
month_name_August
airline_id_WN
month_name_October
month_name_March
day_period_Evening

Important Factors

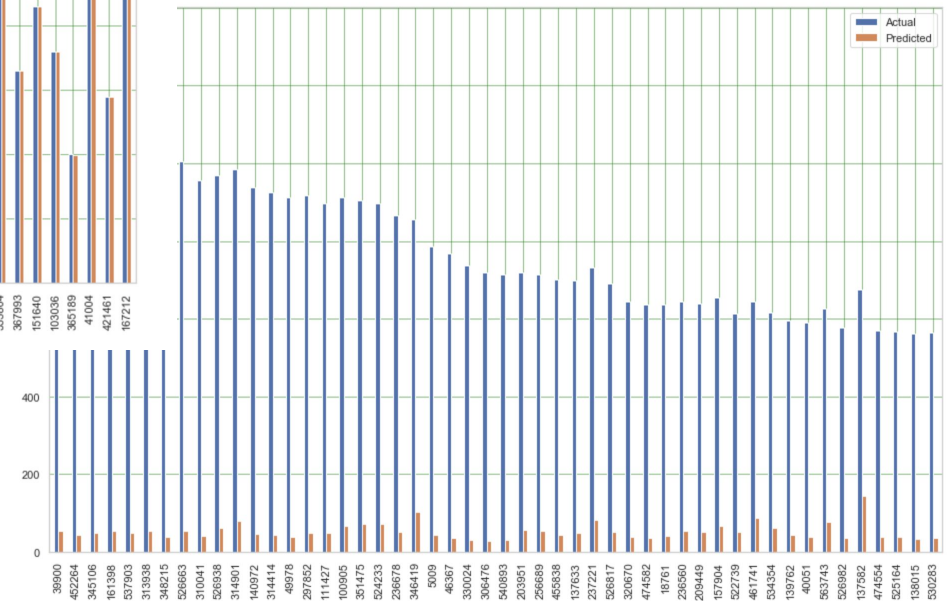
Scheduled duration
Distance
Day of week
Month



Good At predicting short-term delay



Not so well when the delay is long

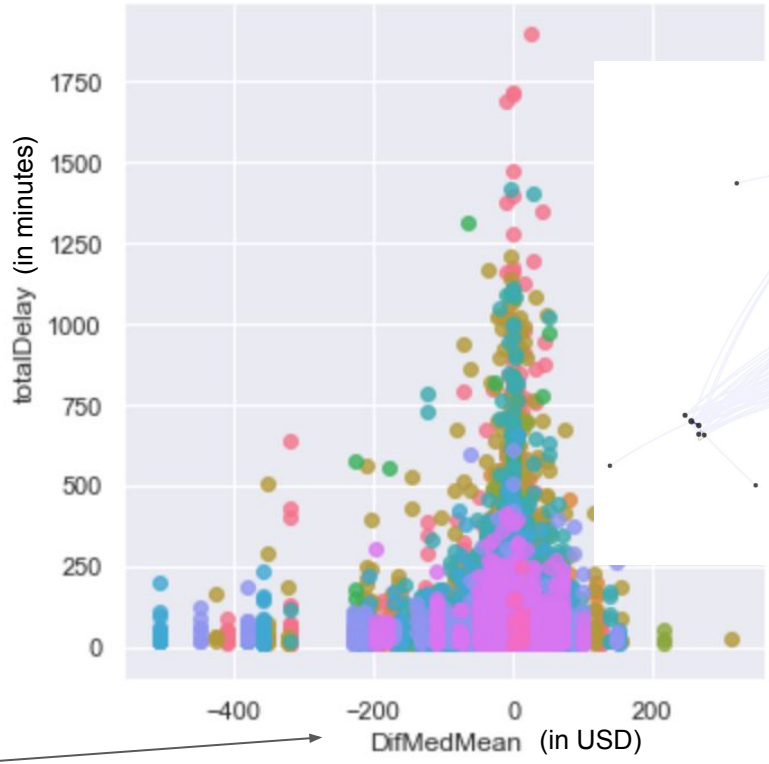


Recommendation: Alternative Airports?

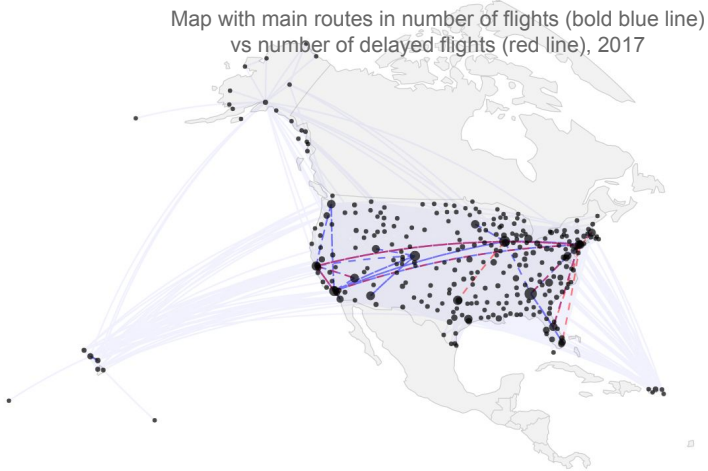
Run same on airports within a given distance-- still working on this code

Route Comparison: Fare vs. Delay

Do some airlines tend to have more delays? What about more



Map with main routes in number of flights (bold blue line) vs number of delayed flights (red line), 2017



We came up with a pricing metric that calculates the median price for a single flight and then compares it against the average median price for that flight path across airlines

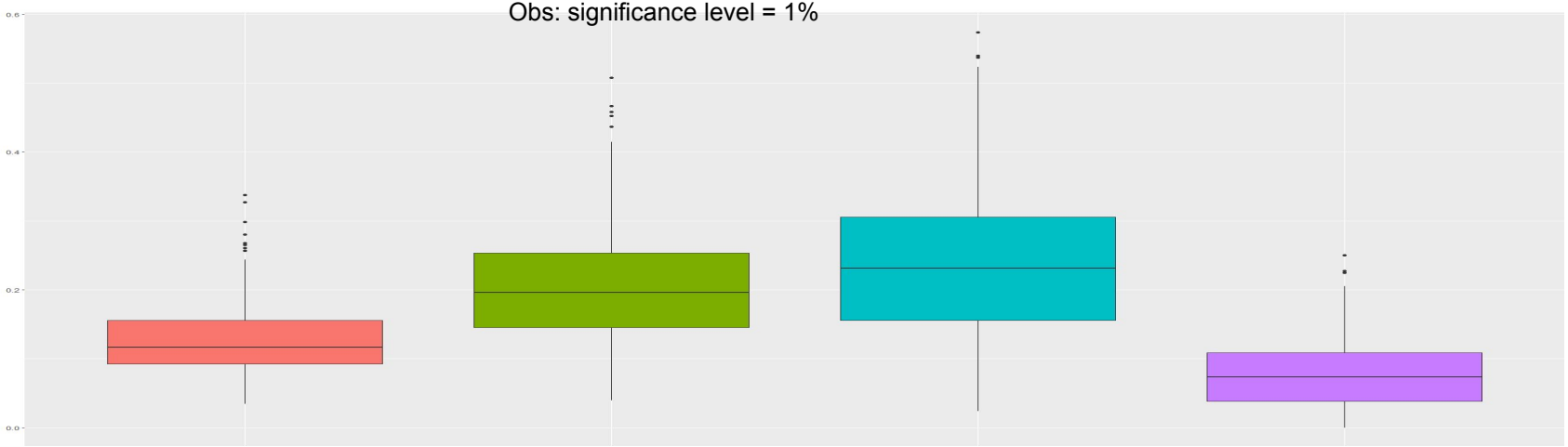
What factors affect delay?

Hypothesis Testing

	Morning	Afternoon	Evening
Afternoon	S		
Evening	S	S	
Night	S	S	S

In hypothesis testing with pairwise comparison procedures, we found out that all the periods of the day presented statistically significant difference in their proportions of delayed/cancelled flights.

Obs: significance level = 1%



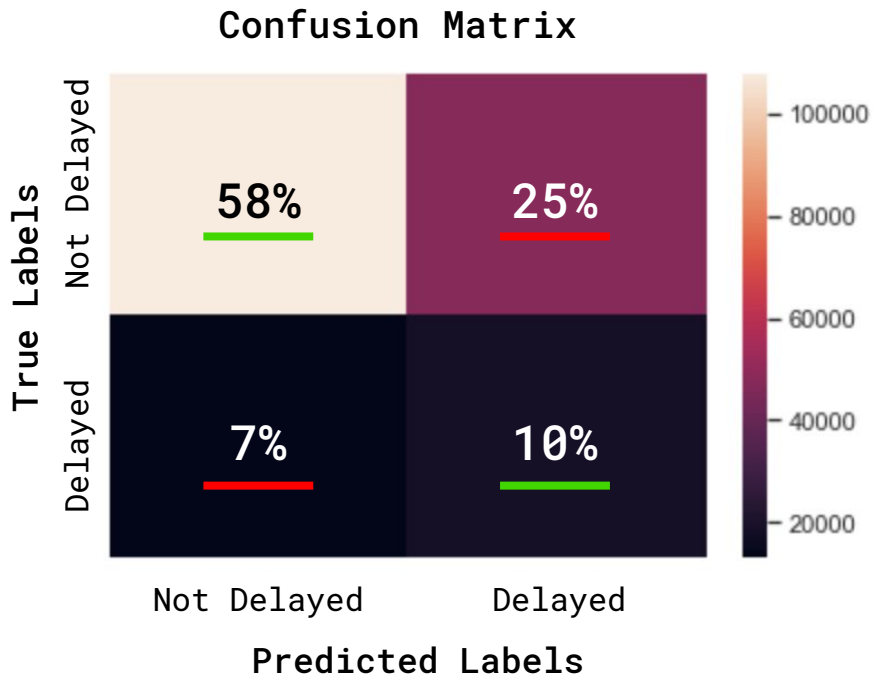
Can we predict instance of delay?

Random Forest

Delay: Y/N?

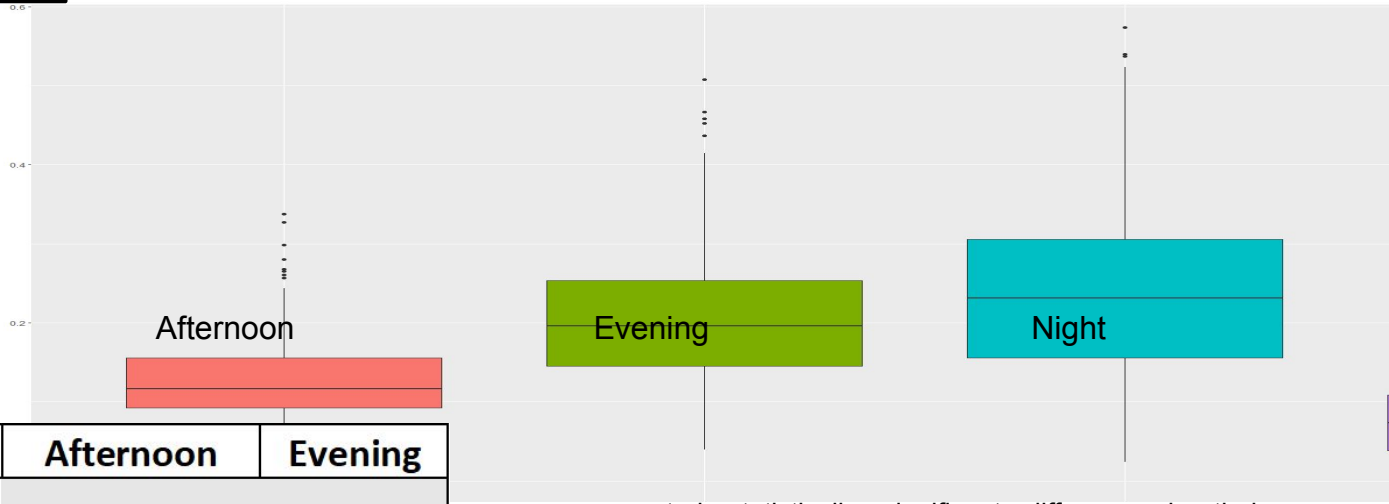
Right now, our model has a lot of false positives :/

Accuracy



What factors affect delay?

Hypothesis Testing



	Morning	Afternoon	Evening
Afternoon	S		
Evening	S	S	
Night	S	S	S

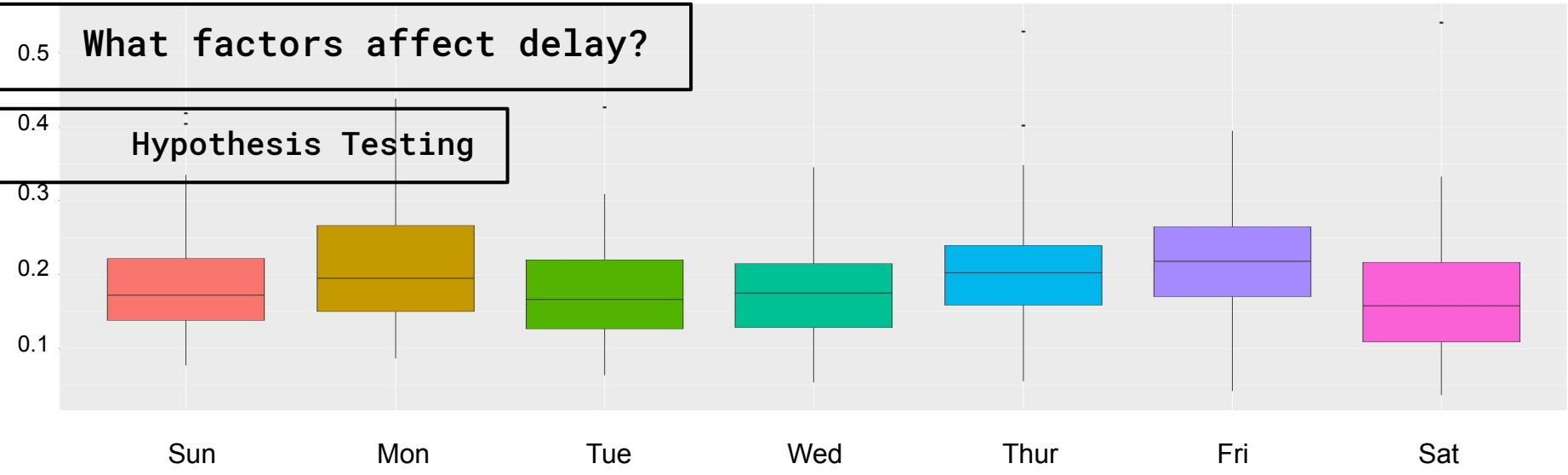
S: significant
NS: not significant

presented statistically significant difference in their proportions of delayed/cancelled flights.

Obs: significance level = 1%

What factors affect delay?

Hypothesis Testing



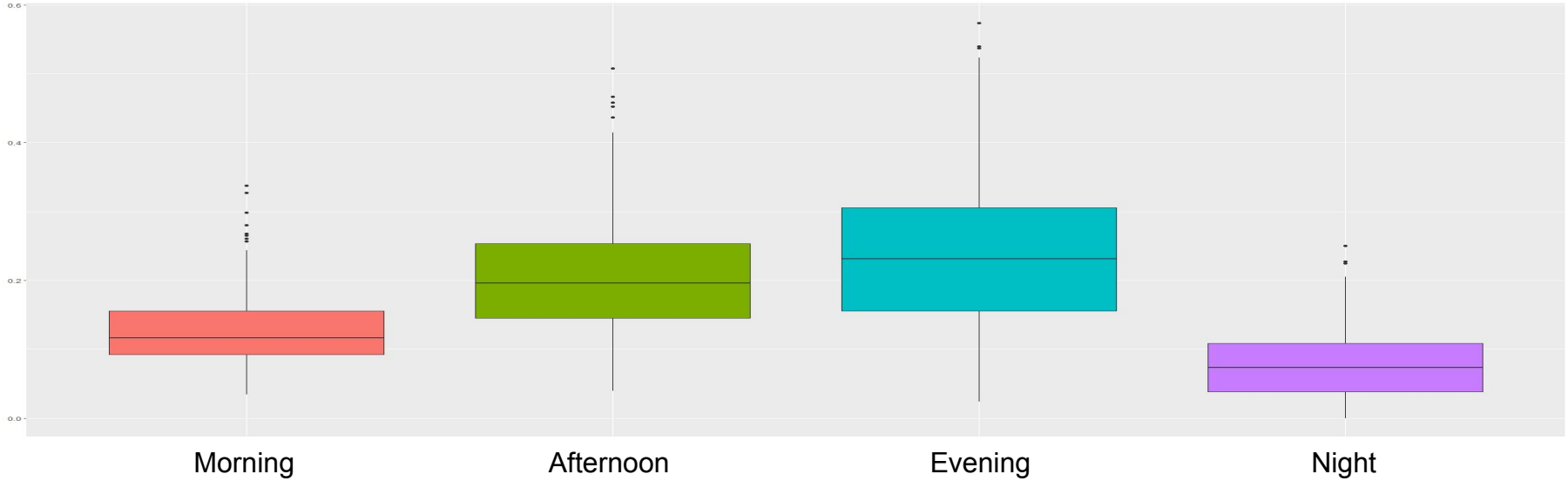
	Sun	Mon	Tue	Wed	Thur	Fri
Mon	S					
Tue	NS	S				
Wed	NS	S	NS			
Thur	S	NS	S	S		
Fri	S	NS	S	S	S	
Sat	S	S	S	S	S	S

S: significant
NS: not significant

In hypothesis testing with pairwise comparison procedures, we found out that the days of the week presented statistically significant difference in their proportions of delayed/cancelled flights, except for:

- Sunday versus both Tuesday and Wednesday
- Monday versus Thursday and Friday
- Tuesday versus Wednesday

Obs: significance level = 1%

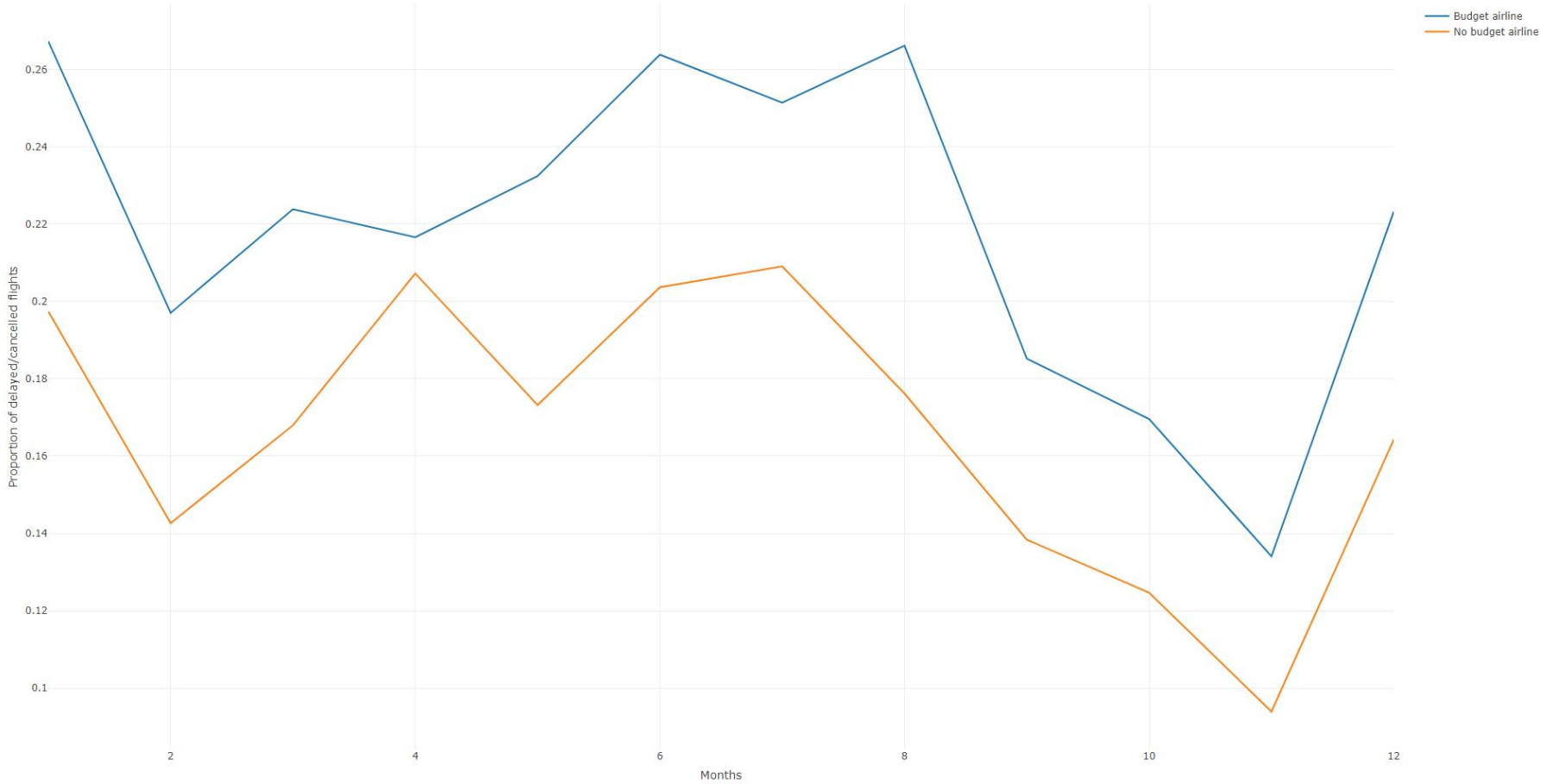


	Morning	Afternoon	Evening
Afternoon	S		
Evening	S	S	
Night	S	S	S

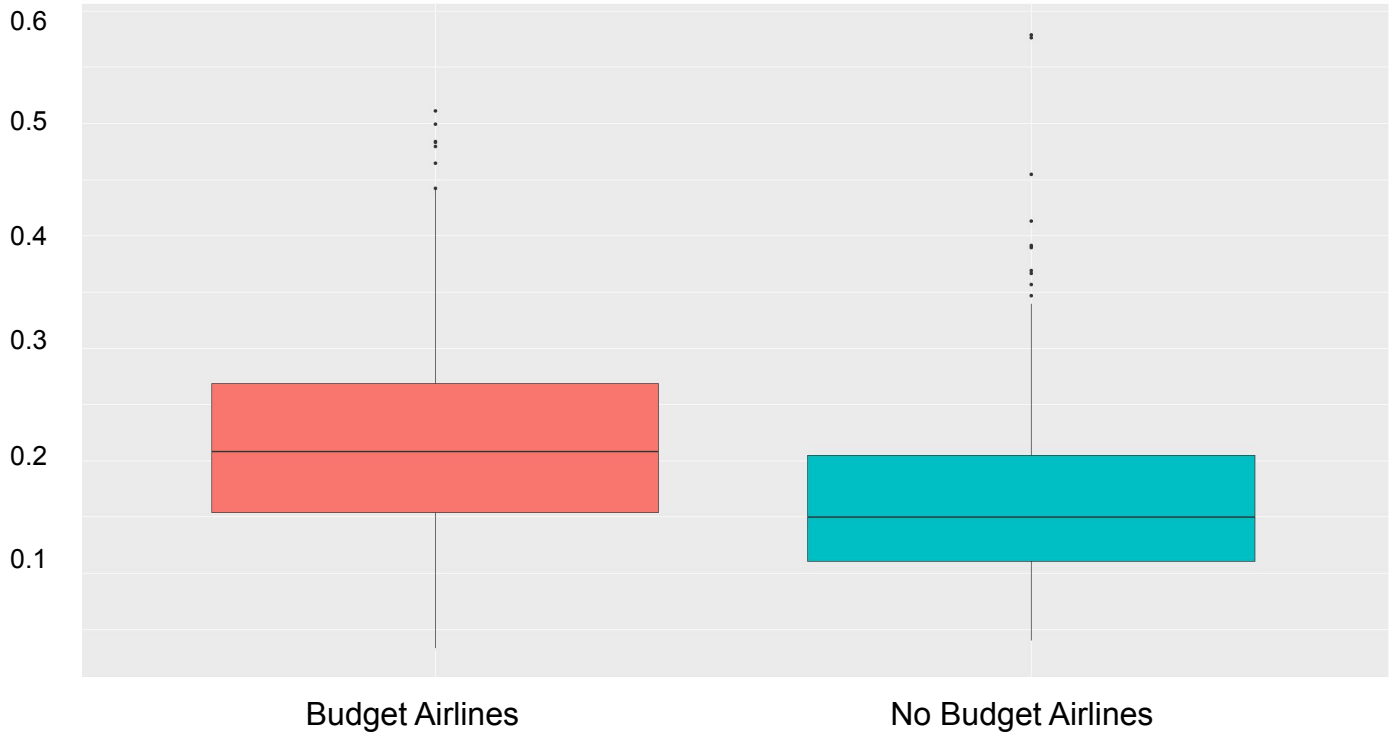
S: significant
NS: not significant

In hypothesis testing with pairwise comparison procedures, we found out that all the periods of the day presented statistically significant difference in their proportions of delayed/cancelled flights.

Obs: significance level = 1%

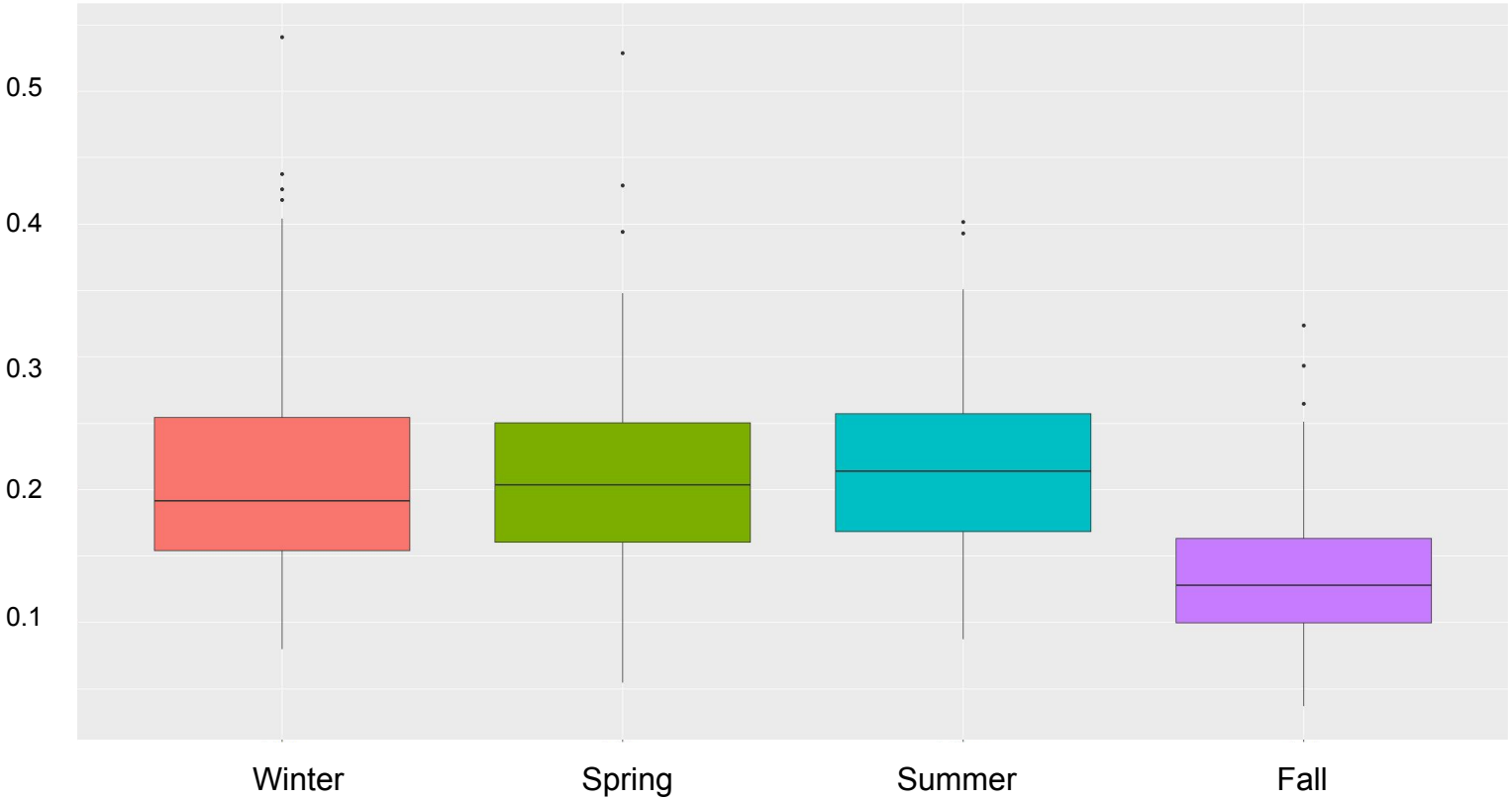

Budget airline
No budget airline
Proportion of delayed/cancelled flights
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
2
4
6
8
10
12
Months

Group 13



In hypothesis testing, we found out that budget airlines presented statistically significant difference in their proportions of delayed/cancelled flights.

Obs: significance level = 1%



In hypothesis testing with pairwise comparison procedures, we found out that Fall season flights presented statistically significant difference in their proportions of delayed/cancelled flights.
Obs: significance level = 1%

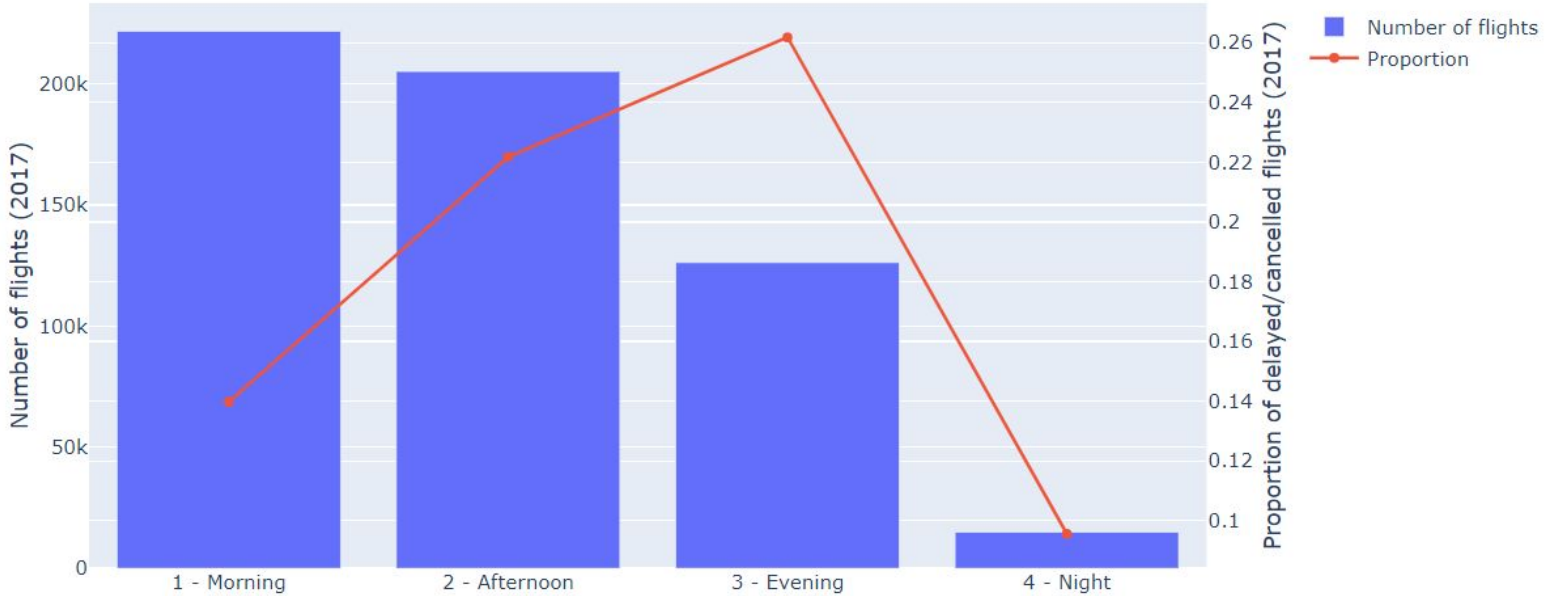
Create your own delay-adjusted fare

Some model for how much an individual is willing to pay to avoid/minute of delay

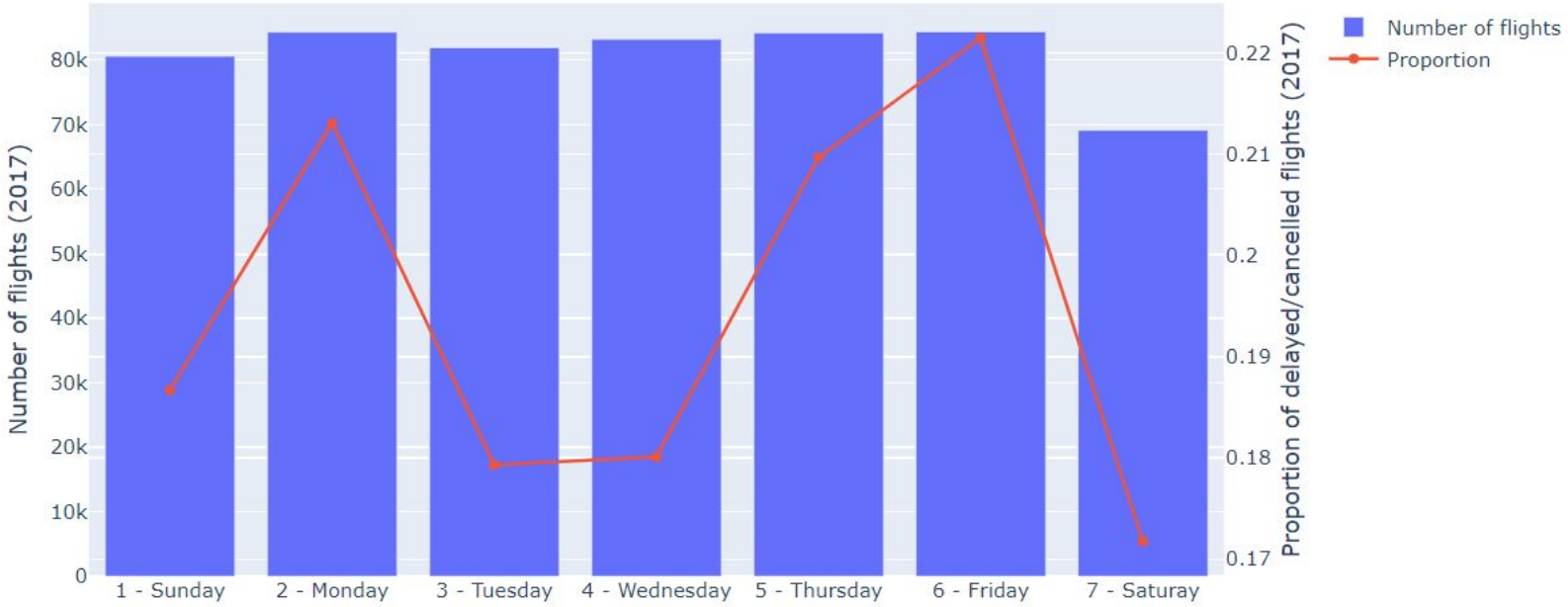
Model as a polynomial or exponential with some assumptions, user input

Exploratory Analysis

Proportional Summary*



Delay Comparison: What factors most affect delay?



Can we predict instance of delay?

Random Forest

Delay: Y/N?

Right now, our model has a lot of false positives :/

Accuracy