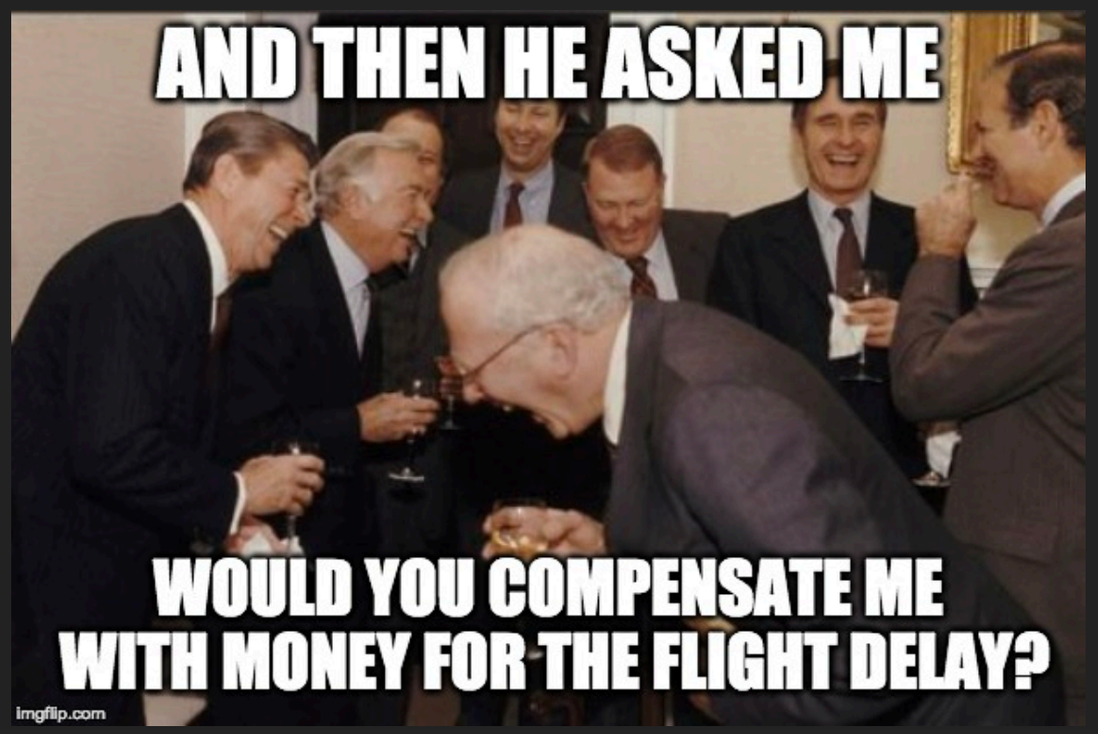
PREDICTING

FLIGHT DELAYS

Gus Eggers, Max Deyo, Micah Ima, Mike Rogers, Wiley Bui

OVERVIEW

- What causes flight delays?
- Over 1 billion travelers in 2018
- Southwest has the most passengers but also the worst on-time track record
- Hawaiian Airlines historically is not impacted by delays for various reasons
- Why do we care?
- Let's take a look!



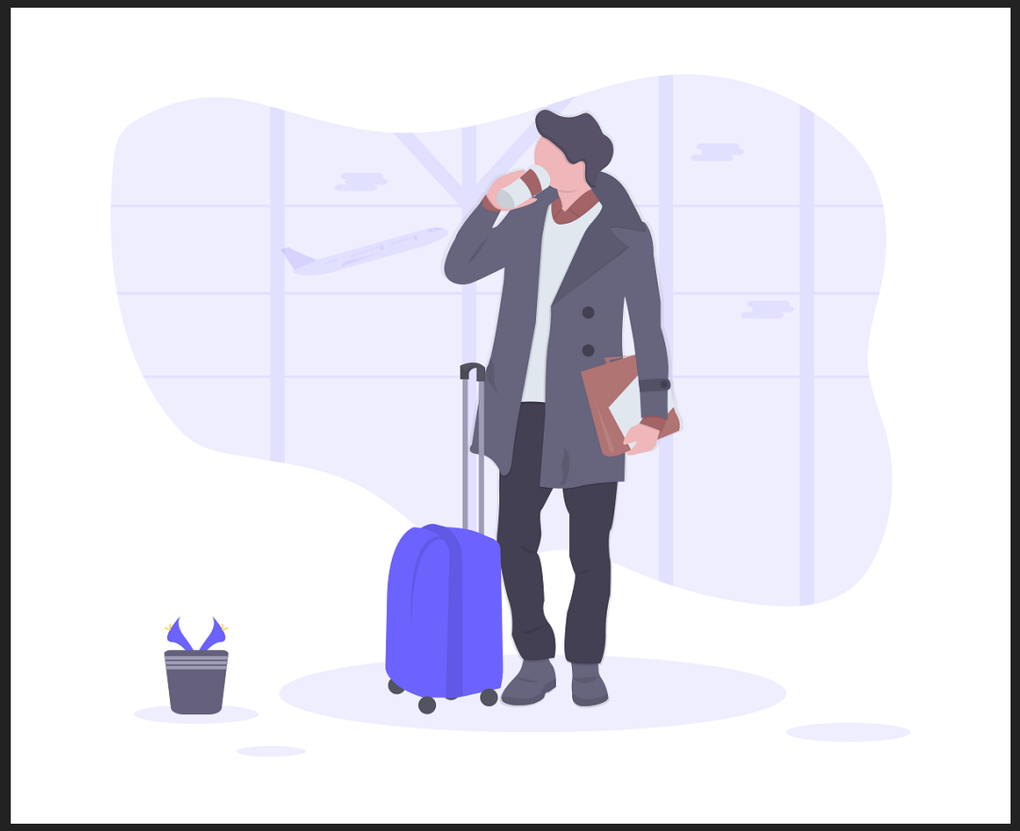
ISSUE

Have you ever needed to catch a flight on a timely matter and worried about being delayed? Not only do flight delays cause airlines to lose profit[8], but they can cause passengers to be late to family events, important business meetings, connecting flights, etc.

For the sake of passengers, our project aims to find the airline with the least likelihood of having a delayed flight based on your origin and destination.

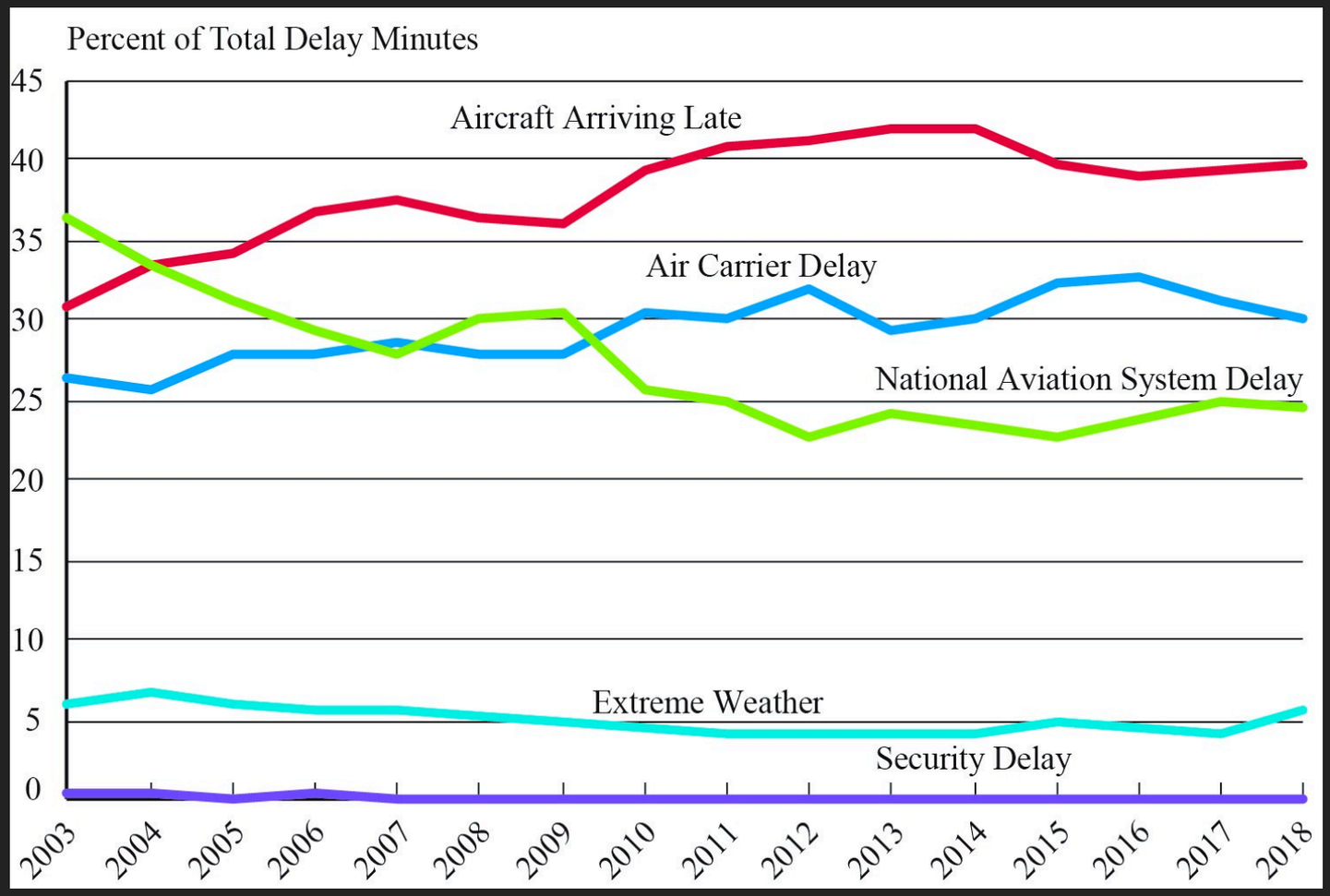
PROJECT GOALS

- predict future flights' statuses based on 2018 dataset
  - will I arrive early, on time, or delay by X minutes/hours?
  - examine past trends
    - time of the year
    - type of airlines
    - origin to destination
    - weather involved
    - air traffic control problems
    - other external factors
- based on the prediction, which type of flights are the best fit for
  - business travelers
  - high/middle/low-income travelers
  - anyone that needs to fly urgently



TOOLS

- Python Notebook
  - Anaconda
  - Google Colab
- Frameworks or libraries:
  - sklearn
  - matplotlib
  - pandas
  - numpy
  - seaborn
  - scipy
  - XGBoost
  - Catboost
  - LightGBM
- Dataset repositories:
  - https://opendata.socrata.com/
  - https://archive.ics.uci.edu/ml/datasets.php
  - https://www.kaggle.com/datasets
  - https://toolbox.google.com/datasetsearch
  - https://www.datazar.com/project/pb615a433-bac3-4483-b80b-717ff6118a61/files
  - https://www.data.gov/
  - https://data.worldbank.org/
  - https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236
  - https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-delays
  - https://www.bts.gov/topics/airlines-and-airports-0

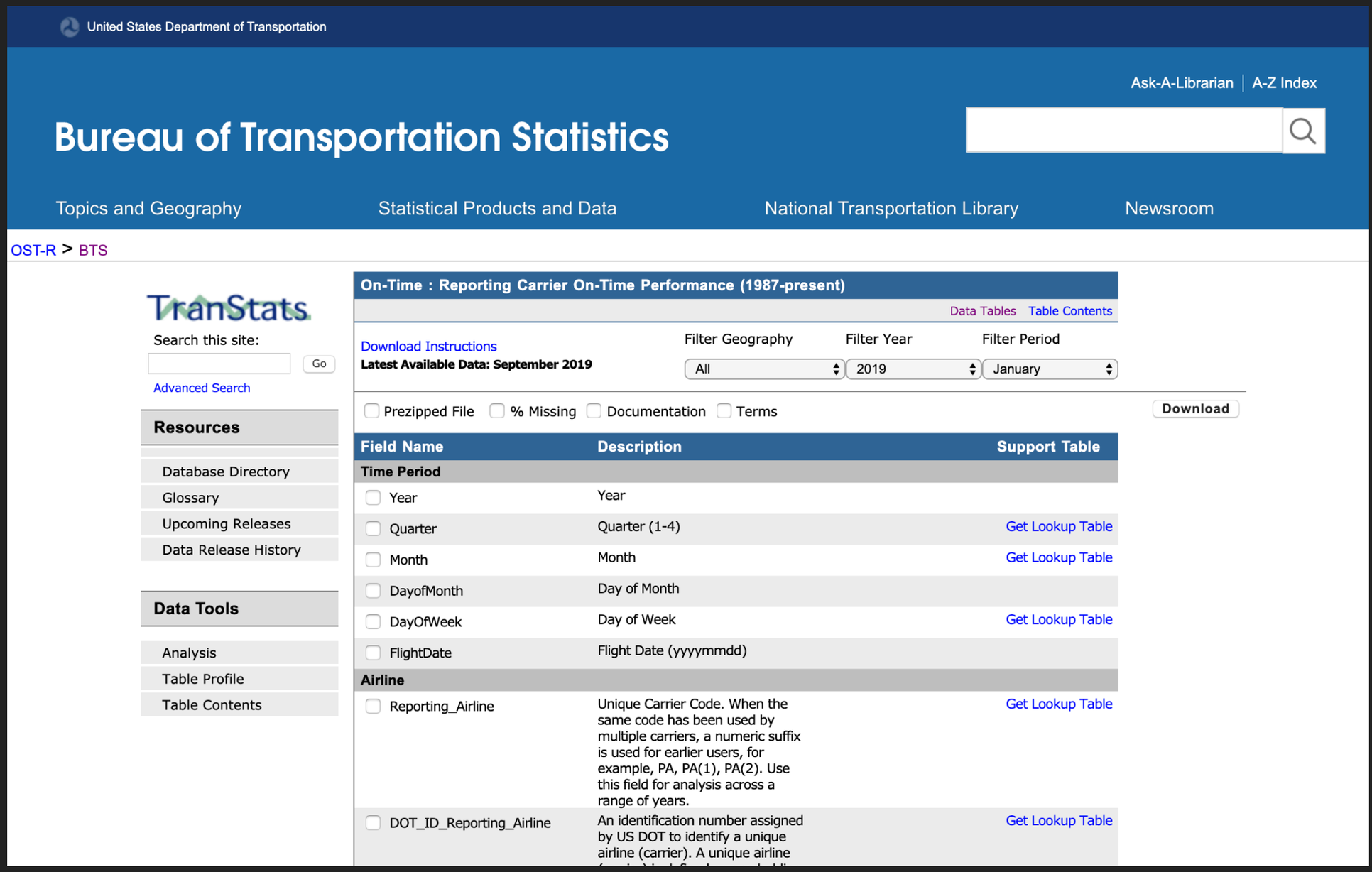


Source: **Bureau of Transportation Statistics**

Most of flight delays can be attributed to three main factors. The leading factor is waiting on an aircraft that simply arrives late. The next is air carrier delay, which can be maintenance, cleaning, baggage loading or fueling related. The third leading factor is a National Aviation System Delay, which can be due to many reasons, such as : non-extreme weather, heavy traffic volume, or air traffic control [5].

SOURCE OF DATASETS - US DEPARTMENT OF TRANSPORTATION

- https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time



Alternative dataset source: https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

SOURCE OF DATASETS - US DEPARTMENT OF TRANSPORTATION

- Choose Fields Based on Description and Filter by Date
- Main dataset: 2018.csv
- Lookup Table1: airports.csv
- Lookup Table2: airlines.csv

There are 7213446 rows and 27 columns

	0	1	2	3	4	5	6	7	8	9
DATE	2018-01-01	2018-01-01	2018-01-01	2018-01-01	2018-01-01	2018-01-01	2018-01-01	2018-01-01	2018-01-01	2018-01-01
AIRLINE	UA	UA	UA	UA	UA	UA	UA	UA	UA	UA
FLIGHT_NUMBER	2429	2427	2426	2425	2424	2422	2421	2420	2419	2418
ORIGIN	EWR	LAS	SNA	RSW	ORD	ORD	IAH	DEN	SMF	RIC
DESTINATION	DEN	SFO	DEN	ORD	ALB	OMA	LAS	CID	EWR	DEN
SCHEDULED_DEPARTURE	1517	1115	1335	1546	630	2241	750	1324	2224	1601
DEPARTURE_TIME	1512	1107	1330	1552	650	2244	747	1318	2237	1559
DEPARTURE_DELAY	-5	-8	-5	6	20	3	-3	-6	13	-2
TAXI_OUT	15	11	15	19	13	15	14	11	10	12
WHEELS_OFF	1527	1118	1345	1611	703	2259	801	1329	2247	1611
WHEELS_ON	1712	1223	1631	1748	926	1	854	1554	627	1748
TAXI_IN	10	7	5	6	10	2	6	6	9	8
SCHEDULED_ARRIVAL	1745	1254	1649	1756	922	14	916	1619	638	1813
ARRIVAL_TIME	1722	1230	1636	1754	936	3	900	1600	636	1756
ARRIVAL_DELAY	-23	-24	-13	-2	14	-11	-16	-19	-2	-17
CANCELLED	0	0	0	0	0	0	0	0	0	0
CANCELLATION_REASON	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DIVERTED	0	0	0	0	0	0	0	0	0	0
SCHEDULED_TIME	268	99	134	190	112	93	206	115	314	252
ELAPSED_TIME	250	83	126	182	106	79	193	102	299	237
AIR_TIME	225	65	106	157	83	62	173	85	280	217
DISTANCE	1605	414	846	1120	723	416	1222	692	2500	1482
AIRLINE_DELAY	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
WEATHER_DELAY	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
AIR_SYSTEM_DELAY	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
SECURITY_DELAY	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LATE_AIRCRAFT_DELAY	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

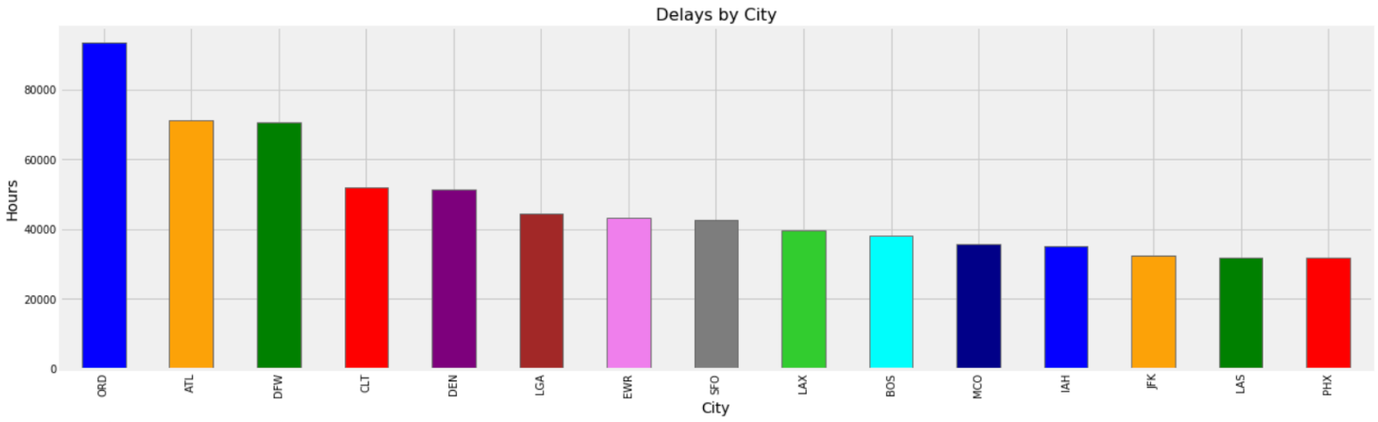
DATA PREPROCESSING

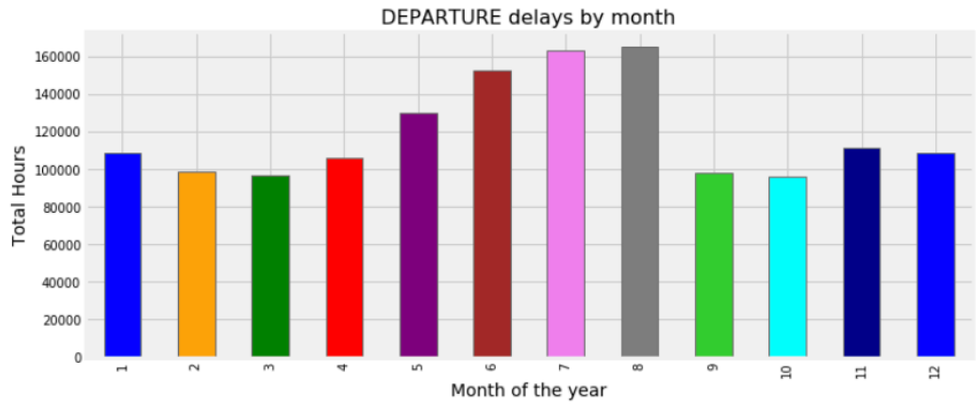
	column type	null values (nb)	null values (%)
DATE	object	0	0
AIRLINE	object	0	0
FLIGHT_NUMBER	int32	0	0
ORIGIN	object	0	0
DESTINATION	object	0	0
SCHEDULED_DEPARTURE	int64	0	0
DEPARTURE_TIME	float64	112317	1.55705
DEPARTURE_DELAY	float64	117234	1.62521
TAXI_OUT	float64	115830	1.60575
WHEELS_OFF	float64	115829	1.60574
WHEELS_ON	float64	119246	1.65311
TAXI_IN	float64	119246	1.65311
SCHEDULED_ARRIVAL	int64	0	0
ARRIVAL_TIME	float64	119245	1.65309
ARRIVAL_DELAY	float64	137040	1.89979
CANCELLED	float64	0	0
CANCELLATION_REASON	object	7096862	98.3838
DIVERTED	float64	0	0
SCHEDULED_TIME	float64	10	0.00013863
ELAPSED_TIME	float64	134442	1.86377
AIR_TIME	float64	134442	1.86377
DISTANCE	float64	0	0
AIRLINE_DELAY	float64	5860736	81.2474
WEATHER_DELAY	float64	5860736	81.2474
AIR_SYSTEM_DELAY	float64	5860736	81.2474
SECURITY_DELAY	float64	5860736	81.2474
LATE_AIRCRAFT_DELAY	float64	5860736	81.2474

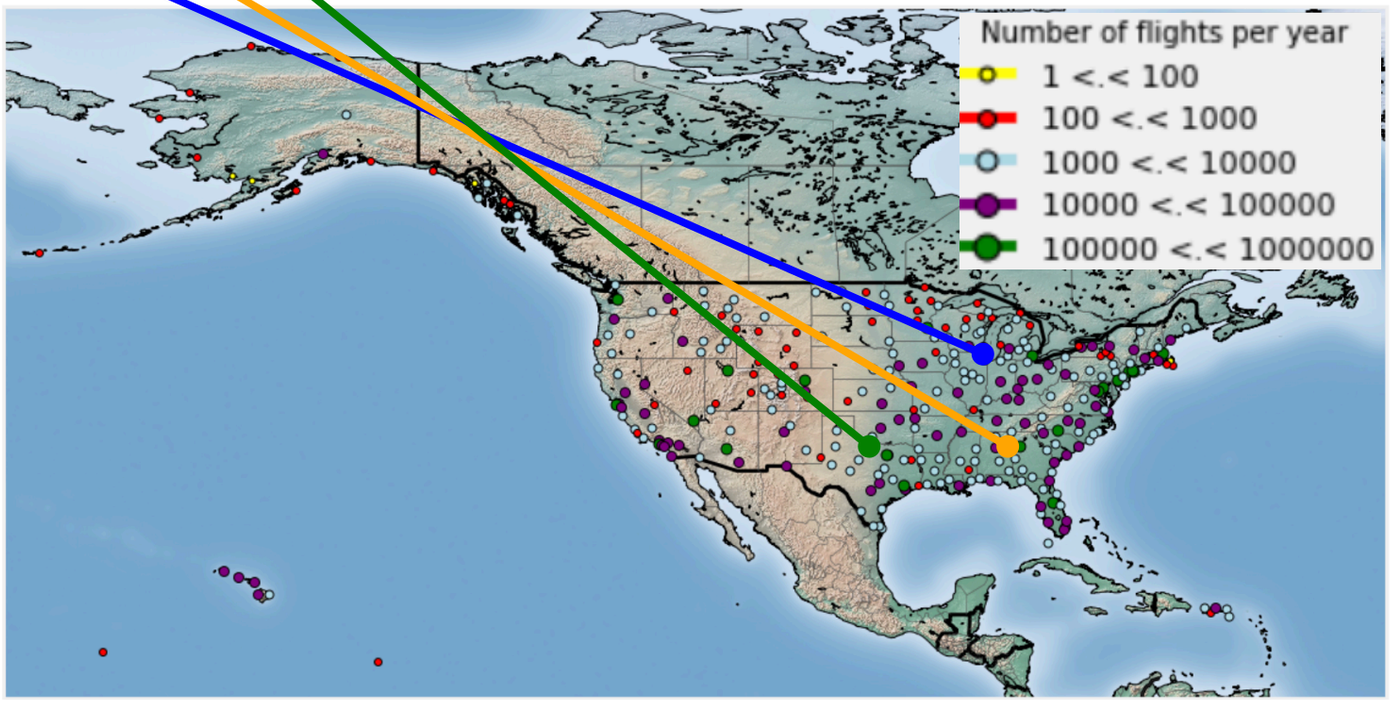
	AIRLINE	ORIGIN	DESTINATION	SCHEDULED_DEPARTURE	DEPARTURE_TIME
4	UA	ORD	ALB	2018-01-01 18:10:00	06:50:00
11	UA	ORD	CLE	2018-01-01 18:22:00	23:00:00
14	UA	ORD	BTV	2018-01-01 06:02:00	22:30:00
15	UA	MCO	LAX	2018-01-01 21:00:00	07:47:00
16	UA	EWR	SMF	2018-01-01 16:46:00	19:22:00

	variable	missing values	filling factor (%)
0	ARRIVAL_DELAY	137040	98.100215
1	ELAPSED_TIME	134442	98.136231
2	ARRIVAL_TIME	119245	98.346907
3	DEPARTURE_DELAY	117234	98.374785
4	DEPARTURE_TIME	112317	98.442949
5	SCHEDULED_TIME	10	99.999861
6	AIRLINE	0	100.000000
7	ORIGIN	0	100.000000
8	DESTINATION	0	100.000000
9	SCHEDULED_DEPARTURE	0	100.000000
10	SCHEDULED_ARRIVAL	0	100.000000

PEEK AT DATASET

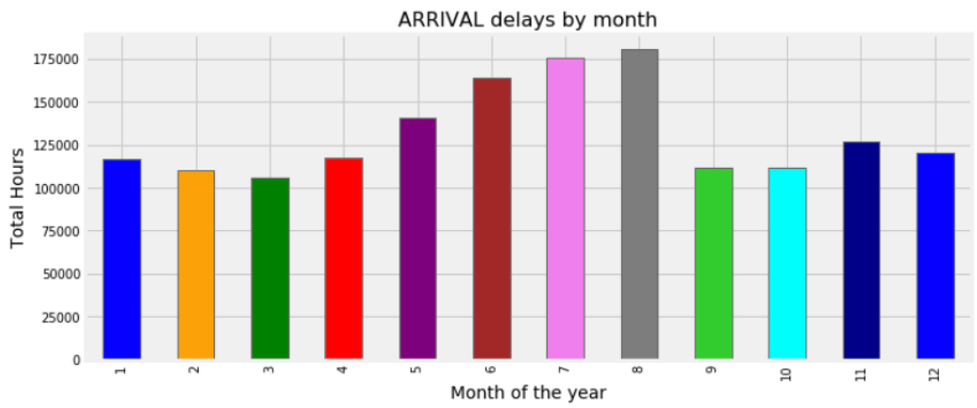






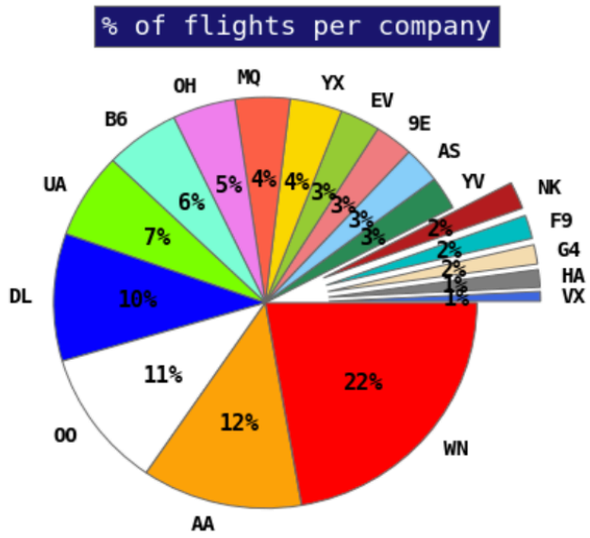
There tend to be more delays during summer months, this is primarily due to weather (harder to fly in the summer because of increased take off space for flights and higher than normal temperatures [9]) and higher volume of passengers.

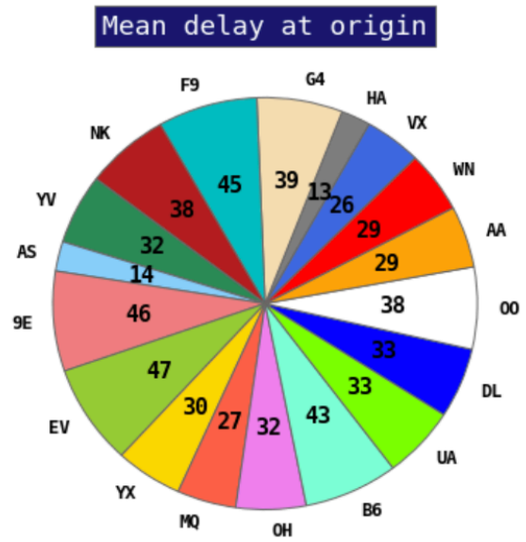
Additionally, larger airports warrant more delays, which makes sense. In general, more flights = more delays.

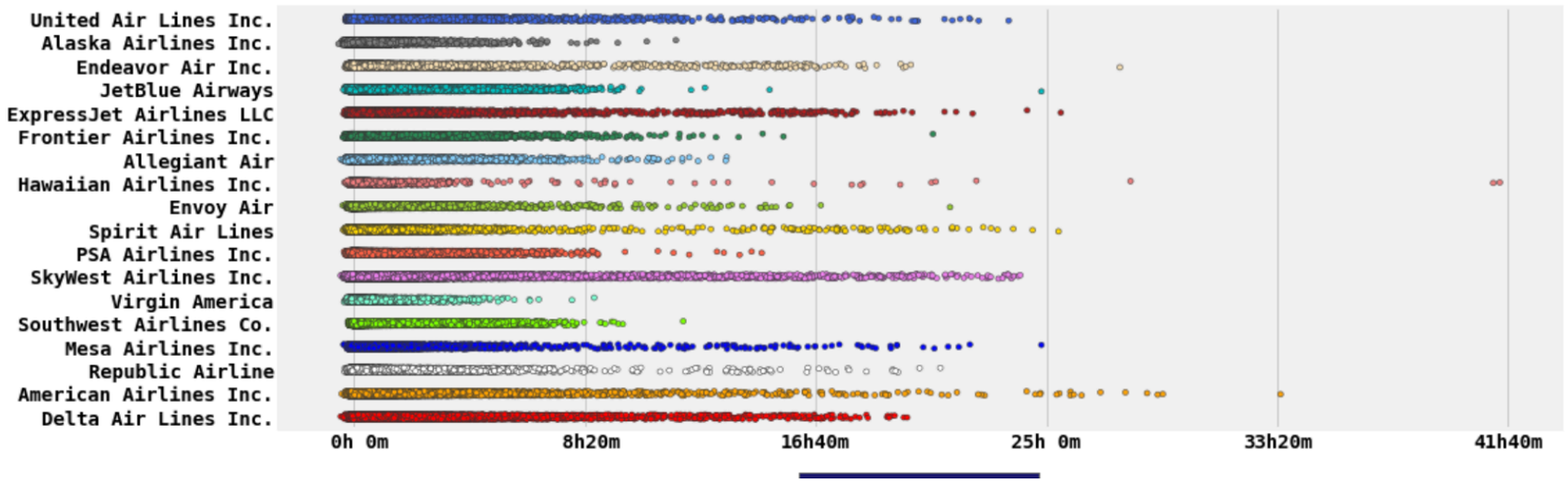


SOME STATISTICS FOR 2018

VX: Virgin America
HA: Hawaiian Airlines Inc.
G4: Allegiant Air
F9: Frontier Airlines Inc.
NK: Spirit Air Lines
YV: Mesa Airlines Inc.
AS: Alaska Airlines Inc.
9E: Endeavor Air Inc.
EV: ExpressJet Airlines LLC
YX: Republic Airline
MQ: Envoy Air
OH: PSA Airlines Inc.
B6: JetBlue Airways
UA: United Air Lines Inc.
DL: Delta Air Lines Inc.
OO: SkyWest Airlines Inc.
AA: American Airlines Inc.
WN: Southwest Airlines Co.







SkyWest has an 11% delay rate, consistently long delay time.

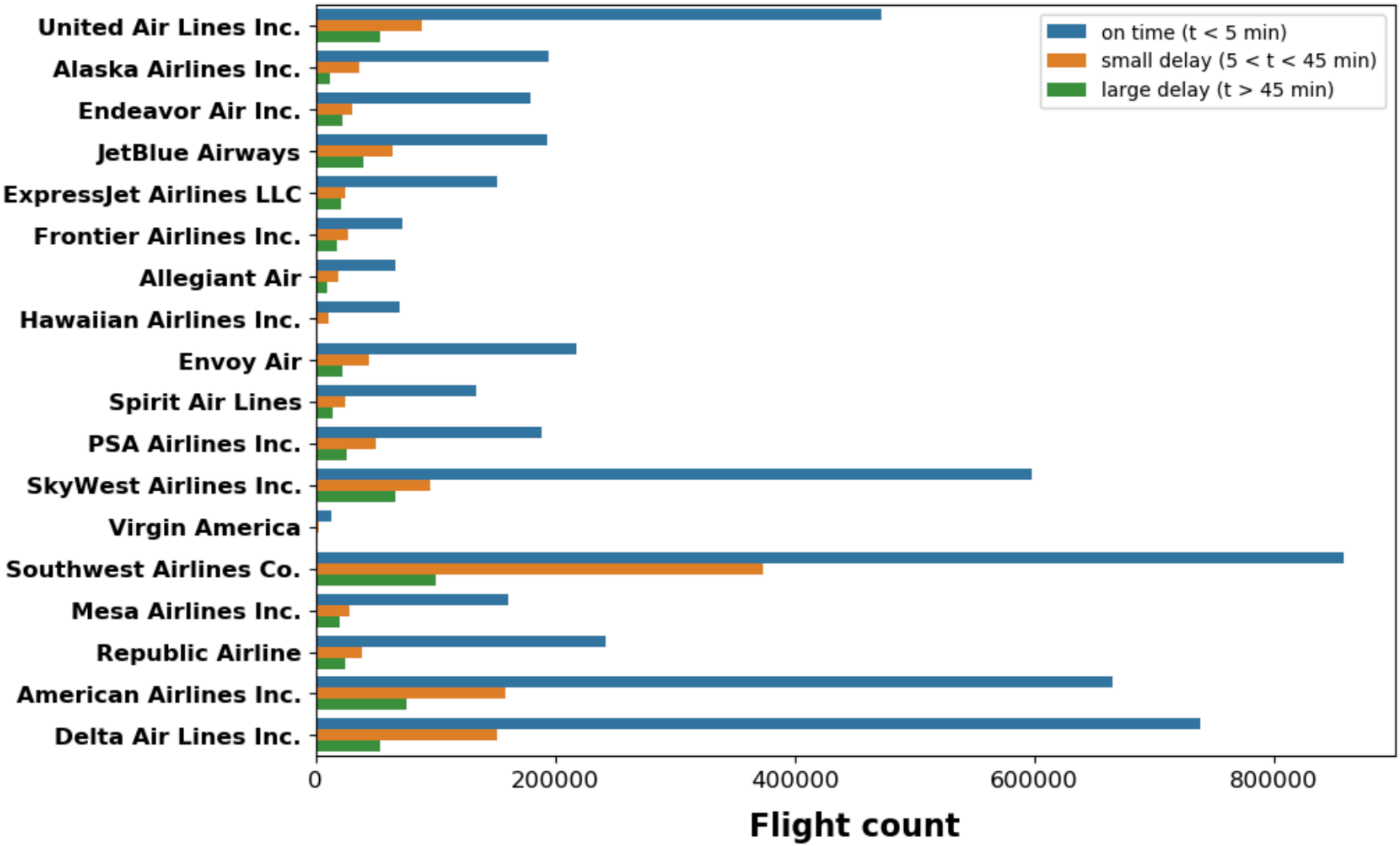
Southwest has a 22% delay rate, however not as long of a delay time as SkyWest.

Delta has a 10% delay rate, with delay times being an average of SkyWest and Southwest's time.

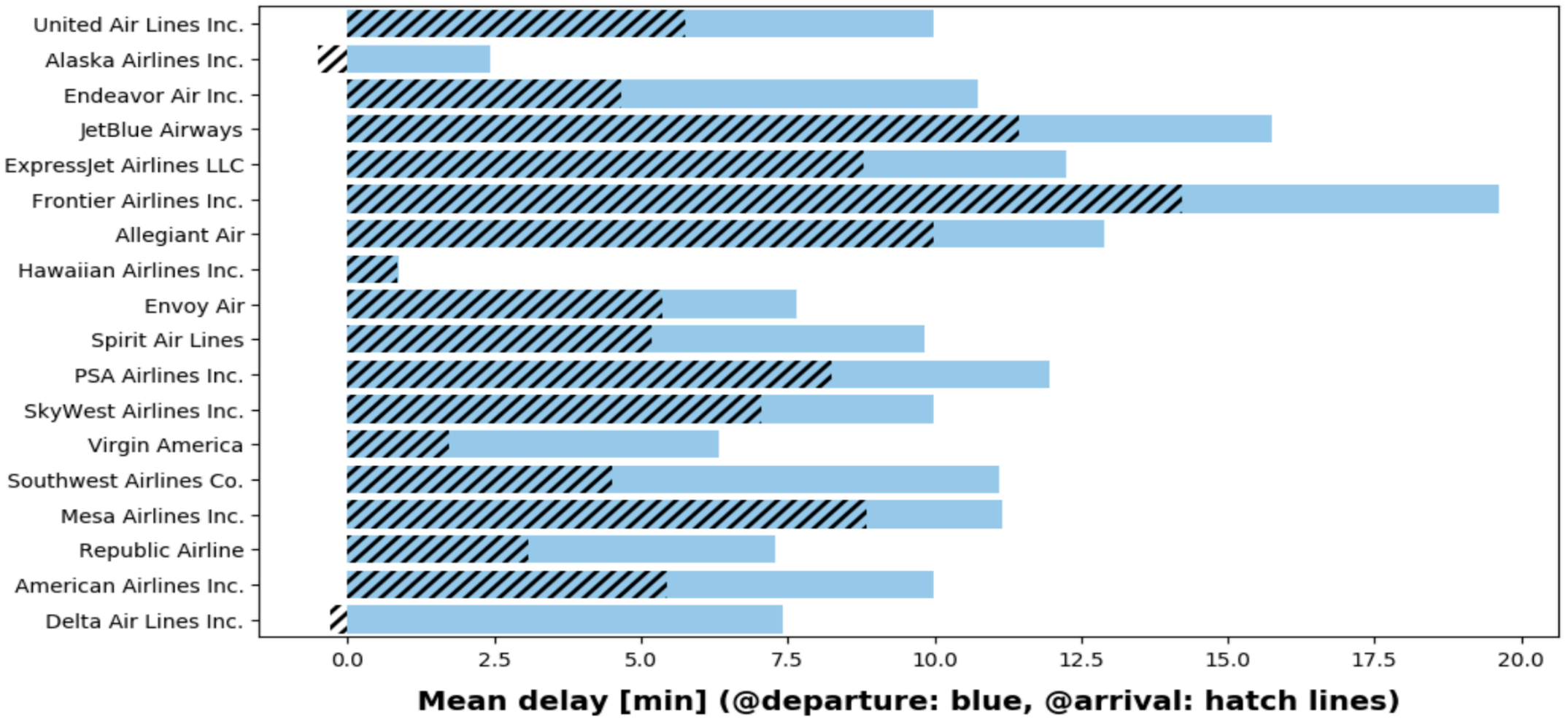
Based on the data, the most likely airlines to have a delayed flight are (in order) : Southwest, American, SkyWest, Delta, United, and JetBlue.

Out of these airlines, SkyWest, Delta, United, and American have the longest delays. Therefore, they tend to be riskier airlines to choose a flight with if you're on a time crunch.

MORE STATISTICS FOR 2018



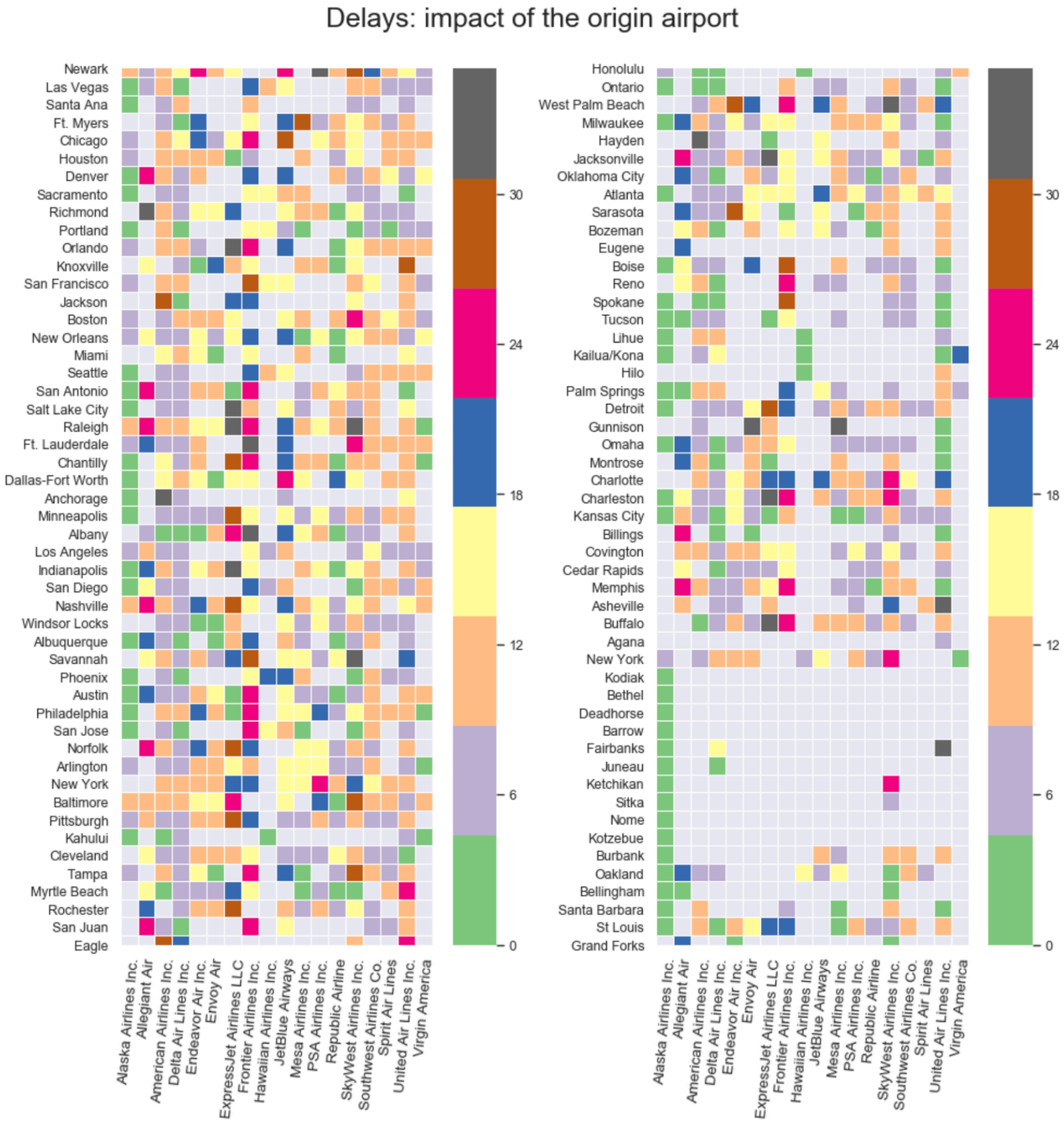
MORE STATISTICS FOR 2018



This figure shows that nearly all airlines have longer departure delays than arrival delays, with the notable exception of Hawaiian, Alaska, and Delta Airlines. Generally, the departure delays are longer because of the previous flight arriving late, flight needing maintenance, fueling, or bagging related delays. Some airlines, like Delta and Alaska next to no arrival delays.

In response to Hawaiian Airlines having the same departure and arrival delays, it's most likely due to there being less flights with this airline compared to the others. In Delta and Alaska's case, it's possible there are less/no arrival delays because of increased number of aircrafts, meaning they don't need to clean/refuel/unbag as often.

MORE STATISTICS



To Use This Graph:

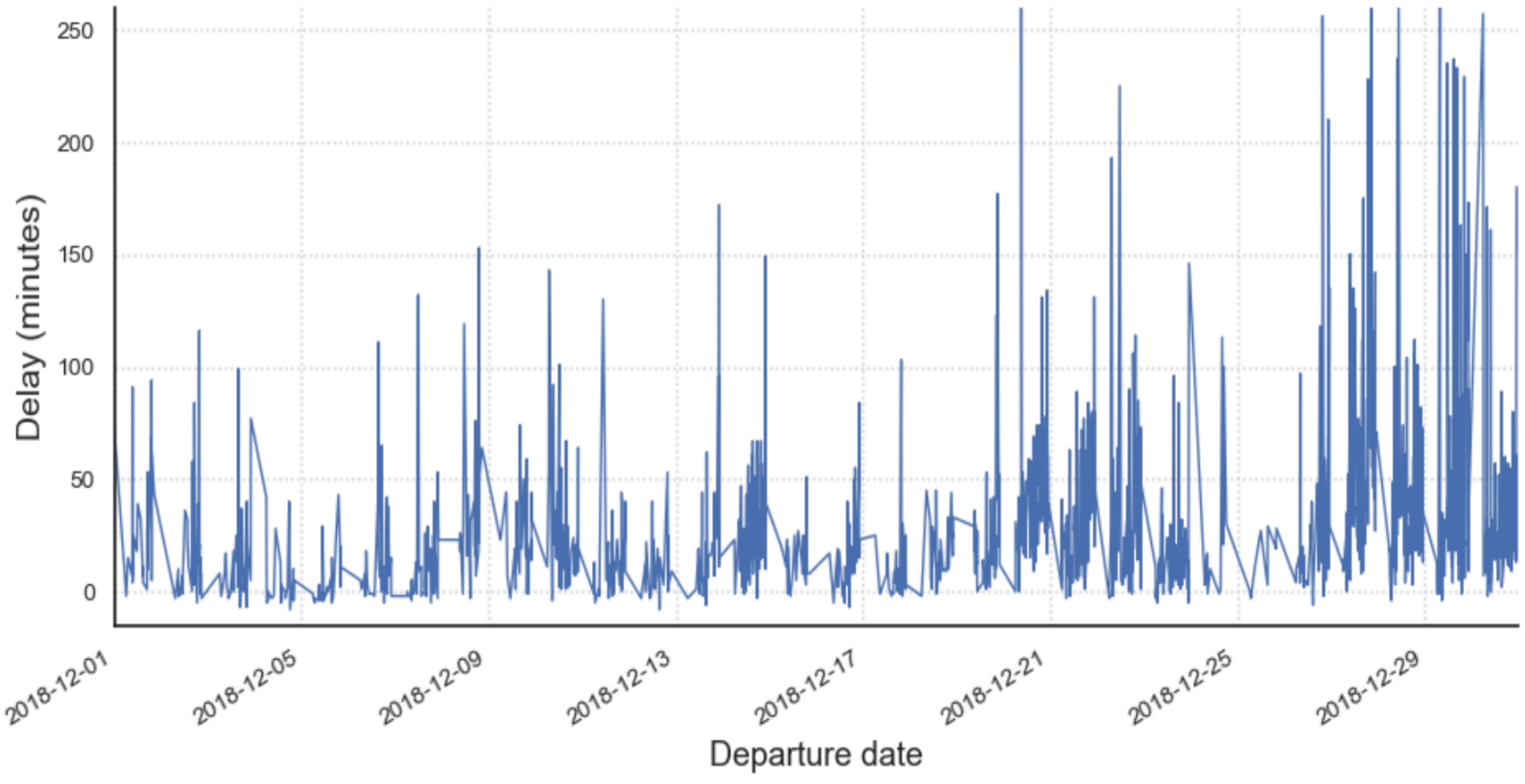
First, choose your location (column).

Second, choose your airline (row). Third, find the intersecting square. If the color is light gray, the airline isn't available at the selected location. Otherwise, compare the color to the key on the right.

Example :
Let's say we're in Detroit. We want to fly with United. According to the color key, the average delay time (if we're delayed) will be ~12 minutes.

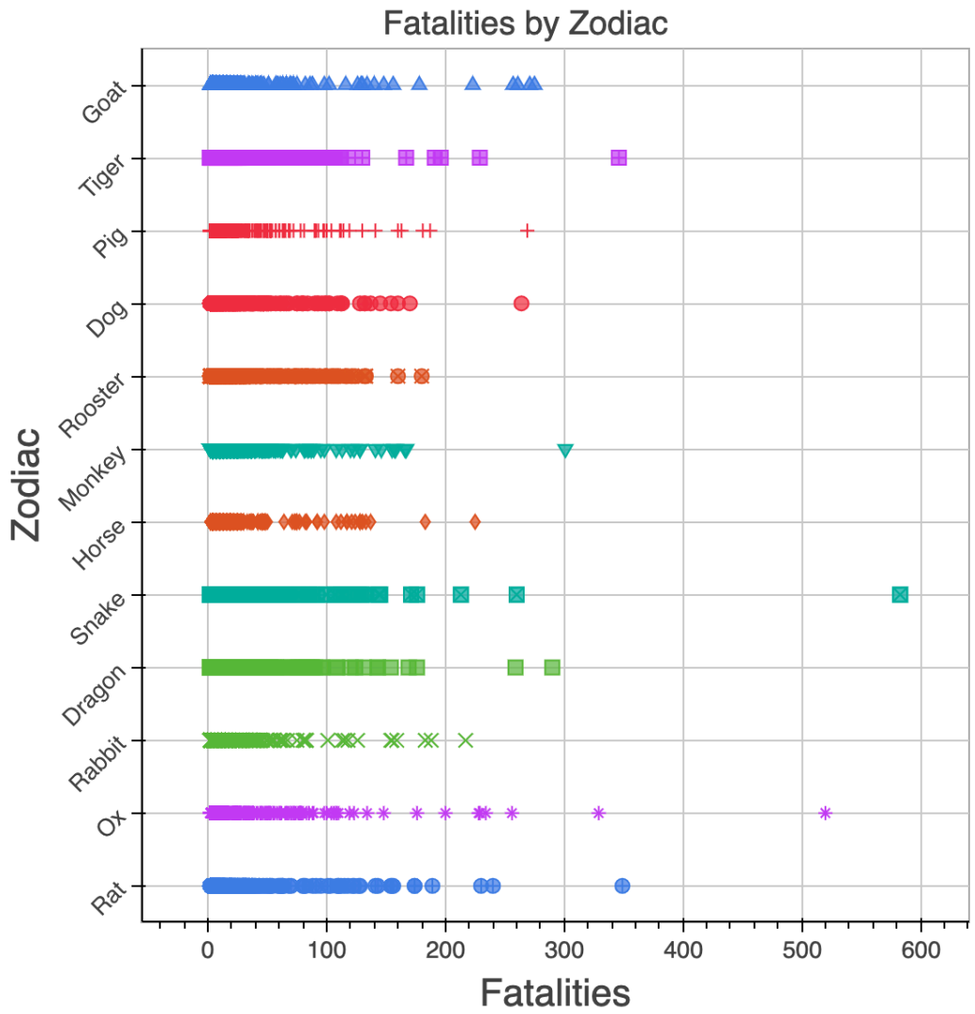
This graph can also be used to find the average delays based on individual location or airlines. For example, Chicago tends to have longer delays (as shown earlier) while places like Tucson have shorter delays.

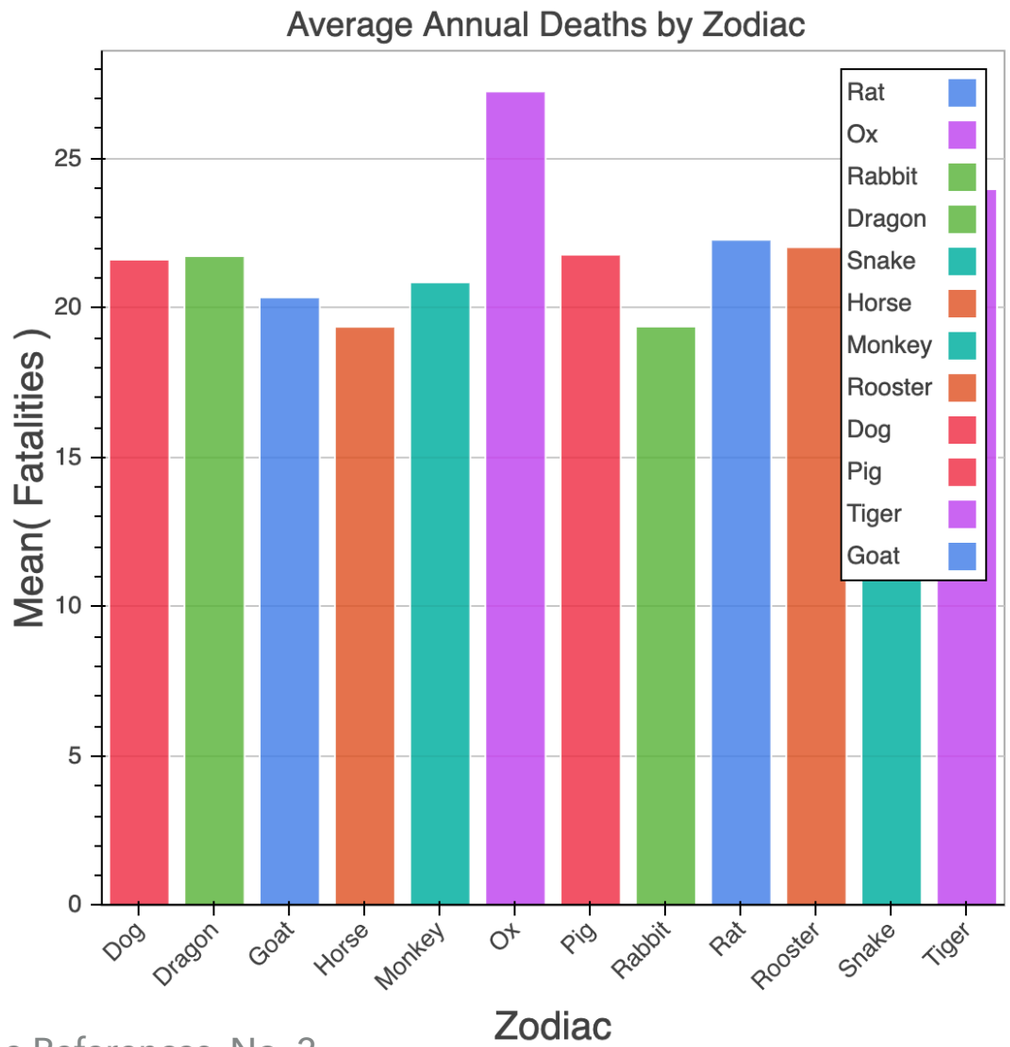
Heavy Traffic Leads to Increased Delays



OUR CRAZY IDEA – SORT OF

FLIGHTS NUMBERS BEGINNING WITH A NUMBER 4, ARE 2.58% LIKELY TO BE DELAYED ON DEPARTURE, AND ONLY 2.95% LIKELY TO BE DELAYED ON ARRIVAL, SECOND LEAST LIKELY BEHIND FLIGHT NUMBERS BEGINNING WITH A SIX.





Source for Zodiac Graphs: See References, No. 2

RANDOM FOREST CLASSIFIER

We chose a random forest classifier for our model but also tested a number of other classification models, including Catboost, LightGBM, XGBoots, LinearRegression + RandomForestClassifier combined.

	YEAR	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	CRS_DEP_TIME
0	2018	12	6	4	645
1	2018	12	6	4	700
2	2018	12	6	4	1133
3	2018	12	6	4	727
4	2018	12	6	4	1039

5 rows × 698 columns

```
print(train_x.shape,test_x.shape)
```

```
(475073, 697) (118769, 697)
```

```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(random_state=42, n_jobs=-1, n_estimators=100)
model.fit(train_x, train_y)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=None, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=-1, oob_score=False, random_state=42, verbose=0,
                       warm_start=False)
```

RANDOM FOREST CLASSIFIER

```
from sklearn.metrics import precision_score

train_predictions = model.predict(train_x)
precision_score(train_y, train_predictions)
```

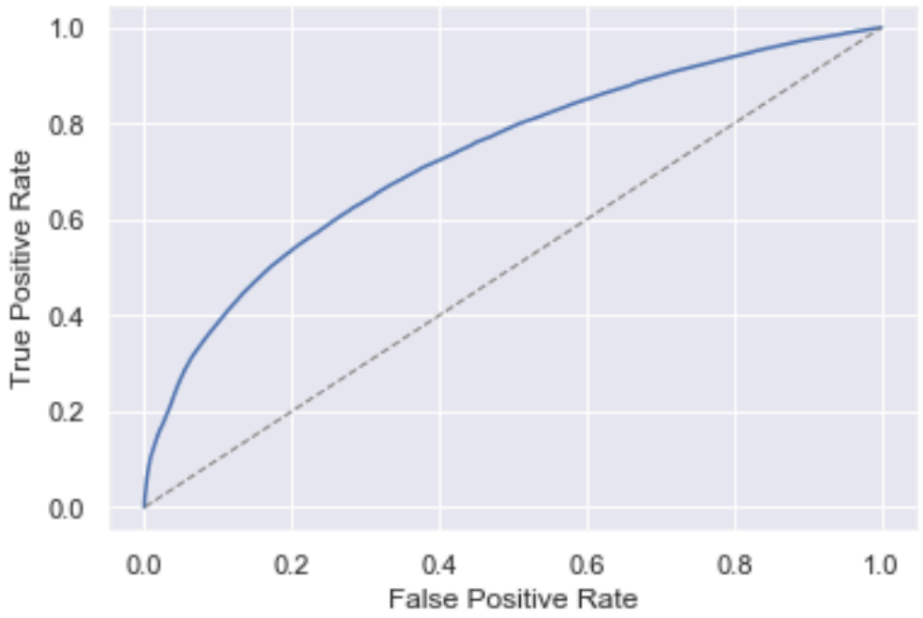
```
0.9762488197473135
```

```
from sklearn.metrics import recall_score

recall_score(train_y, train_predictions)
```

```
0.929832450082972
```

```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns

sns.set()
```

AUC: ~ 74%



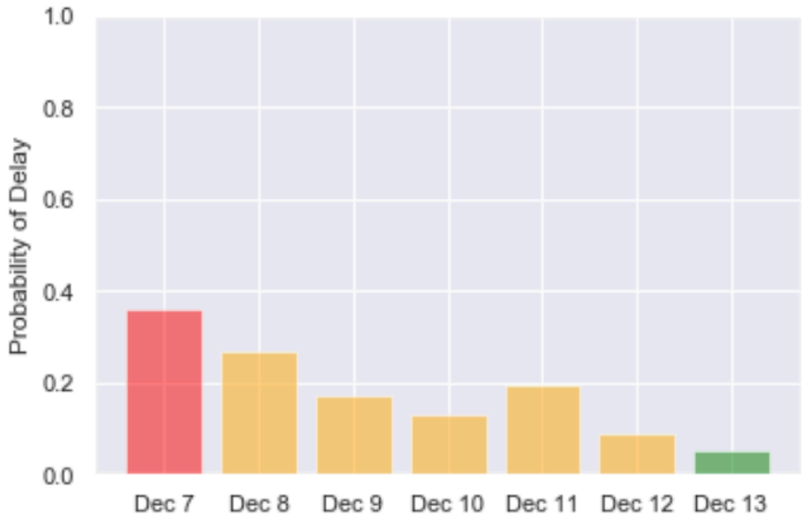
Individual Predictions

```
predict_delay('12/24/2019 13:40:00', 'ASE', 'IAH')
```

```
0.875
```

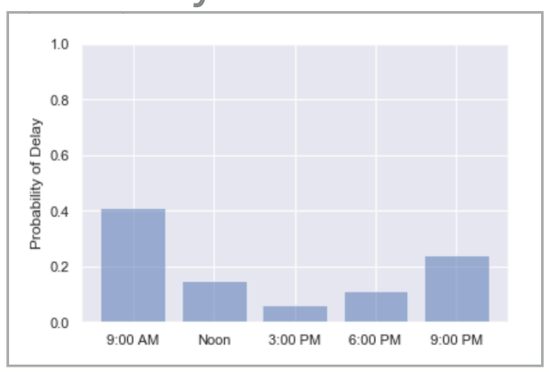
Aspen, CO to Houston, TX, Departure 1:40PM

```
import numpy as np

labels = ('Dec 7', 'Dec 8', 'Dec 9', 'Dec 10', 'Dec 11', 'Dec 12', 'Dec 13')
values = (predict_delay('12/1/2019 21:45:00', 'HNL', 'PDX'),
          predict_delay('12/2/2019 21:45:00', 'HNL', 'PDX'),
          predict_delay('12/3/2019 21:45:00', 'HNL', 'PDX'),
          predict_delay('12/4/2019 21:45:00', 'HNL', 'PDX'),
          predict_delay('12/5/2019 21:45:00', 'HNL', 'PDX'),
          predict_delay('12/6/2019 21:45:00', 'HNL', 'PDX'),
          predict_delay('12/7/2019 21:45:00', 'HNL', 'PDX'))
alabels = np.arange(len(labels))

plt.bar(alabels, values, align='center', alpha=0.5, color=['red','orange','o
plt.xticks(alabels, labels)
plt.ylabel('Probability of Delay')
plt.ylim((0.0, 1.0))
```



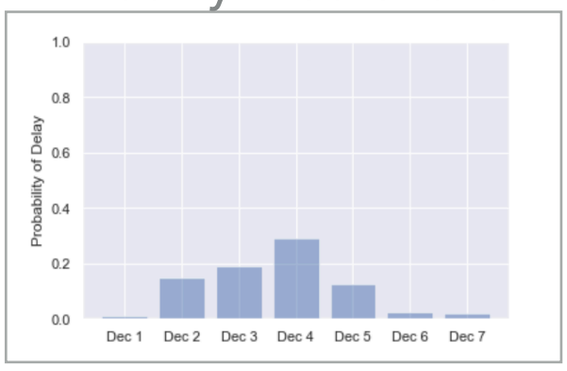
DEC 7-13, 2019 HONOLULU - PORTLAND

Hourly Predictions



DEC 1, 2019 SEATTLE TO ATLANTA

Weekly Predictions



DEC 1-7, 2019 HNL-SFO

REFERENCES

- 1. Gary Stoller (2019). “Most Flight Delays? In The Evening And Unrelated To Bad Weather” [online] Available at: https://www.forbes.com/sites/garystoller/2019/06/03/most-flight-delays-in-the-evening-and-unrelated-to-bad-weather/#695fcd384502
- 2. Do the Zodiacs Influence Aircraft Accidents? [online] Available at: https://www.kaggle.com/jeffd23/chinese-zodiac-and-aircraft-deaths
- 3. Cailey Rizzo | Travel + Leisure (2018). There's a secret code behind flight numbers — here's how to tell what it means [online] Available at: https://www.businessinsider.com/what-does-flight-number-mean-2018-3.
- 4. Yan, L., Dodier, R., Mozer, M., Wolniewicz, R. (2003) Optimizing Classifier Performance via an Approximation to the Wilcoxon-Mann-Whitney Statistic. Available at: https://pdfs.semanticscholar.org/df27/dde10589455d290eeee6d0ae6ceeb83d0c6b.pdf
- 5. Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics. [online] Available at: https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations.
- 7. Jake VanderPlas (2016). Python Data Science Handbook [online book] Available at: https://jakevdp.github.io/PythonDataScienceHandbook/
- 8. Fabien Daniel (2017). Predicting flight delays [online] Available at: https://www.kaggle.com/
- 9. James Vincent (2019). “Google is now using machine learning to predict flight delays” | The Verge [online] Available at: https://www.theverge.com/2018/1/31/16955580/google-flights-app-delays-machine-learning-economy