FLIGHT DELAY PREDICTION

Grace Pham, Heidi Tran, Jessica Jha, Khanh Tran

PROBLEM STATEMENT

People fly every day and delays happen everyday.

We hope to uncover certain trends from this data such

- as which airlines have the most delays
- which destinations are most impacted
- what are the causes of these delays
- how often delays occur
- build predictive classification models and compare which one can predict flight delays the best

RELEVANT WORK

Google is now using machine learning to predict flight delays

The company's Flights app will use historical data to warn users when it thinks their flight will be delayed

By James Vincent | Jan 31, 2018, 1:06pm EST

SHARE

Source: [The Verge 2018]

A statistical approach to predict flight delay using gradient boosted decision tree

Publisher: IEEE Cite This

Suvojit Manna ; Sanket Biswas ; Riyanka Kundu ; Somnath Rakshit ; Priti Gupta ; Subhas Barman

View Document

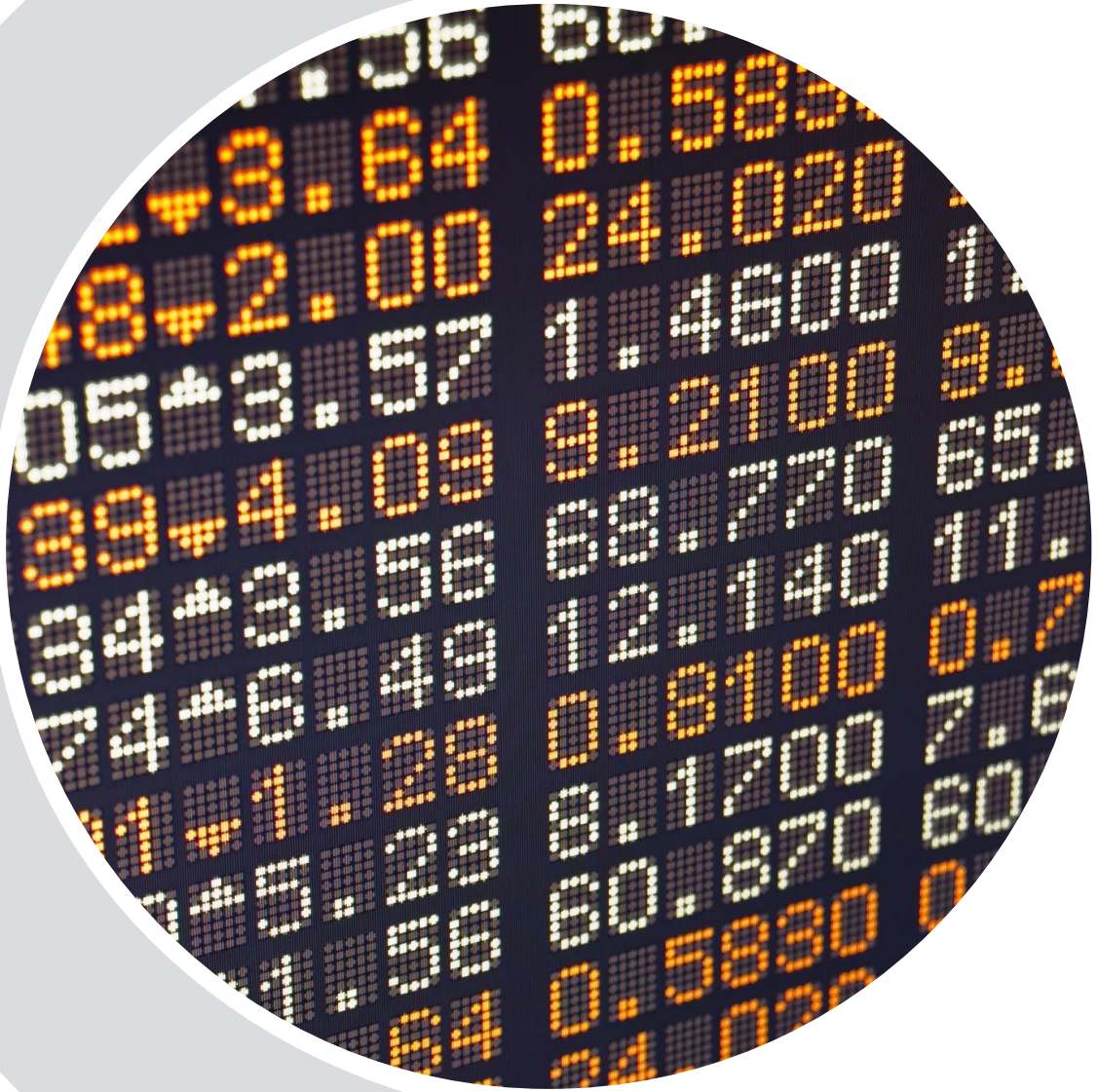
14 Paper Citations

1526 Full Text Views

Source: S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), 2017, pp. 1-5, doi: 10.1109/ICCIDS.2017.8272656.

DATA

- Source: **U.S. Department of Transportation Bureau of Transportation Statistics** through IBM Developer Website
- CSV format
- 194 million flights -> ~386,000+ records and 109 rows



DATA PRE-PROCESSING

Downsizing the data

- Full dataset: 7.2 GB → require a lot of computing power to process
- Data covers dated back to 1987 → stale data can harm the model
- Cut down the data with Spark
- New dataset
  - Data coverage: 2014 – 2019
  - Size: ~200MB

Pre-processing columns

- Removed columns
  - Redundant information
  - Unnecessary
- Added new columns that help future analysis
- Renamed columns for ease of use
- Checked for duplicated records
- Checked for missing values
  - Drop columns with > 95% missing values
  - Filled missing data for numeric columns
  - Dropped rows with missing categorical data
- Checked for outliers → keep outliers

MAP AIRLINE CODES TO THEIR NAMES

	Airline
0	9E
1	9E
2	9E
3	9E
4	9E
...	...
378952	FL
378953	FL
378954	FL
378955	FL
378956	FL

+

	Code	Airline_Name
0	02Q	Titan Airways
1	04Q	Tradewind Aviation
2	05Q	Comlux Aviation, AG
3	06Q	Master Top Linhas Aereas Ltd.
4	07Q	Flair Airlines Ltd.

Airline Mapping

=

	Airline	Airline_Name
0	9E	Endeavor Air Inc.
1	9E	Endeavor Air Inc.
2	9E	Endeavor Air Inc.
3	9E	Endeavor Air Inc.
4	9E	Endeavor Air Inc.
...	...	...
378952	FL	AirTran Airways Corporation
378953	FL	AirTran Airways Corporation
378954	FL	AirTran Airways Corporation
378955	FL	AirTran Airways Corporation
378956	FL	AirTran Airways Corporation

EXPLORATORY DATA ANALYSIS

OVERVIEW

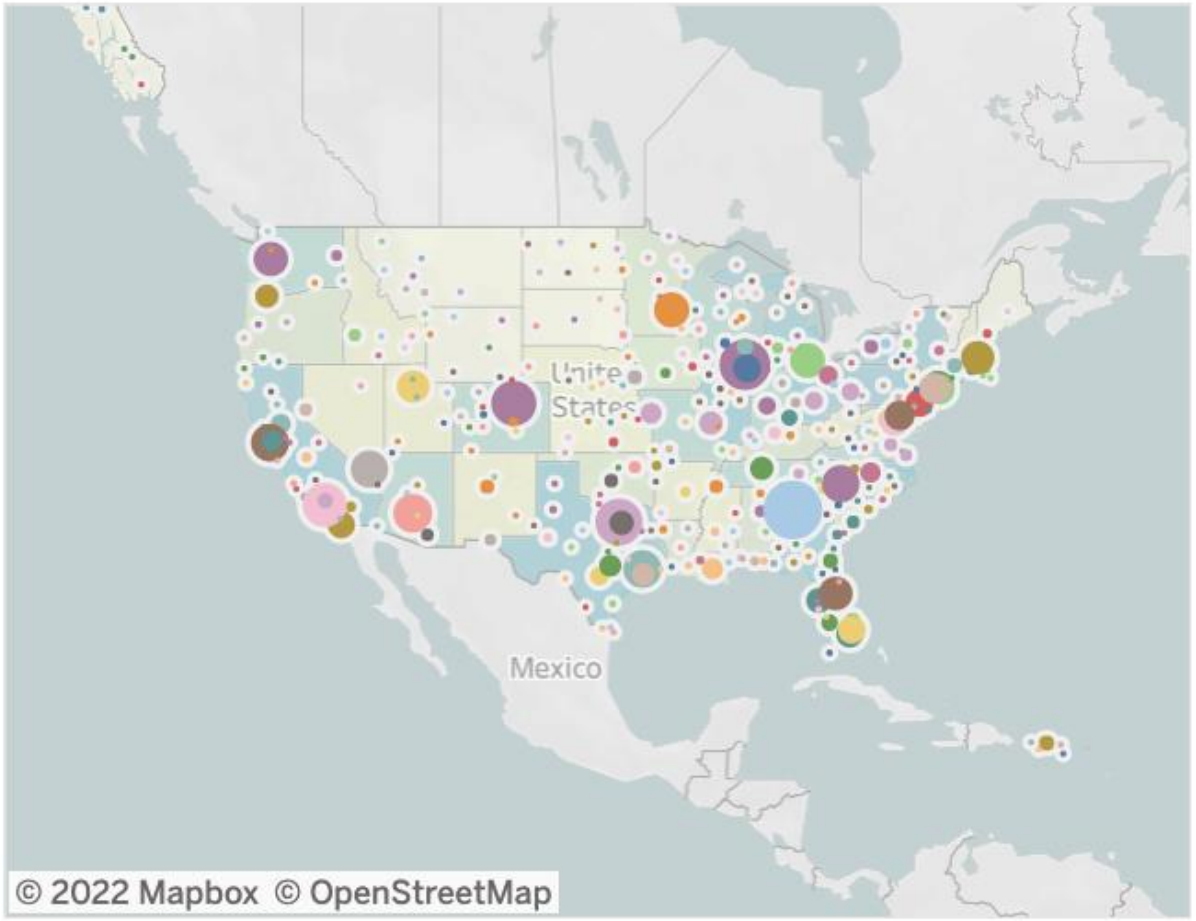
36.01% of flights delayed on departure

36.29% of flights delayed on arrival

Mean departure delay is 13 minutes

Max departure delay is 31.3 hour

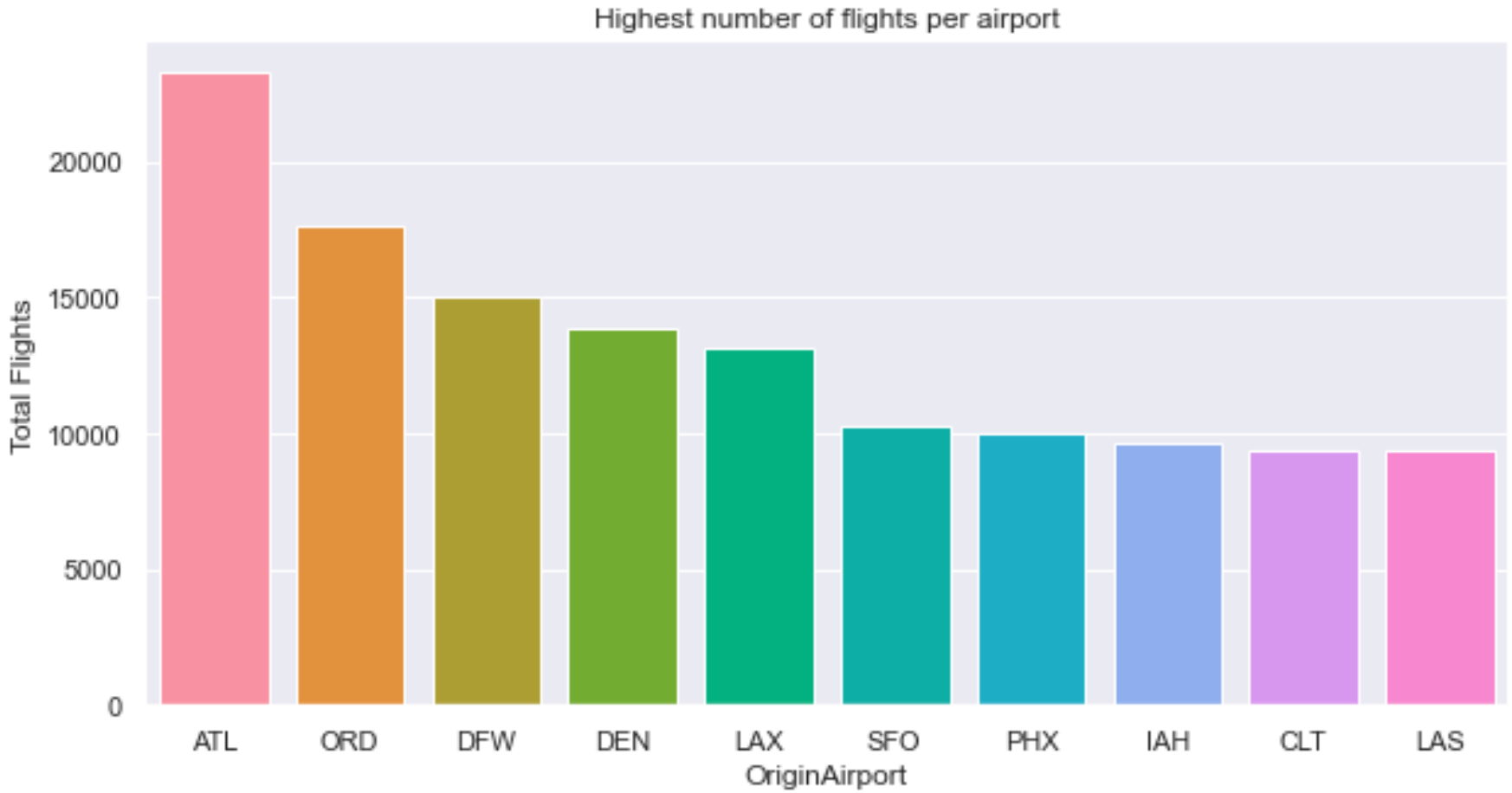
Airports with highest number of delays



- Condensed in major U.S. cities:
- Atlanta, Georgia
- Chicago, Illinois
- Denver, Colorado
- Phoenix, Arizona
- L.A. and S.F. California
- Las Vegas, Nevada
- Dallas-Fort Worth and Houston, Texas
- New York City, New York
- Newark, New Jersey

https://public.tableau.com/app/profile/j.jha/viz/Info442-Maps/Story1

AIRPORTS WITH HIGHEST NUMBER OF FLIGHTS



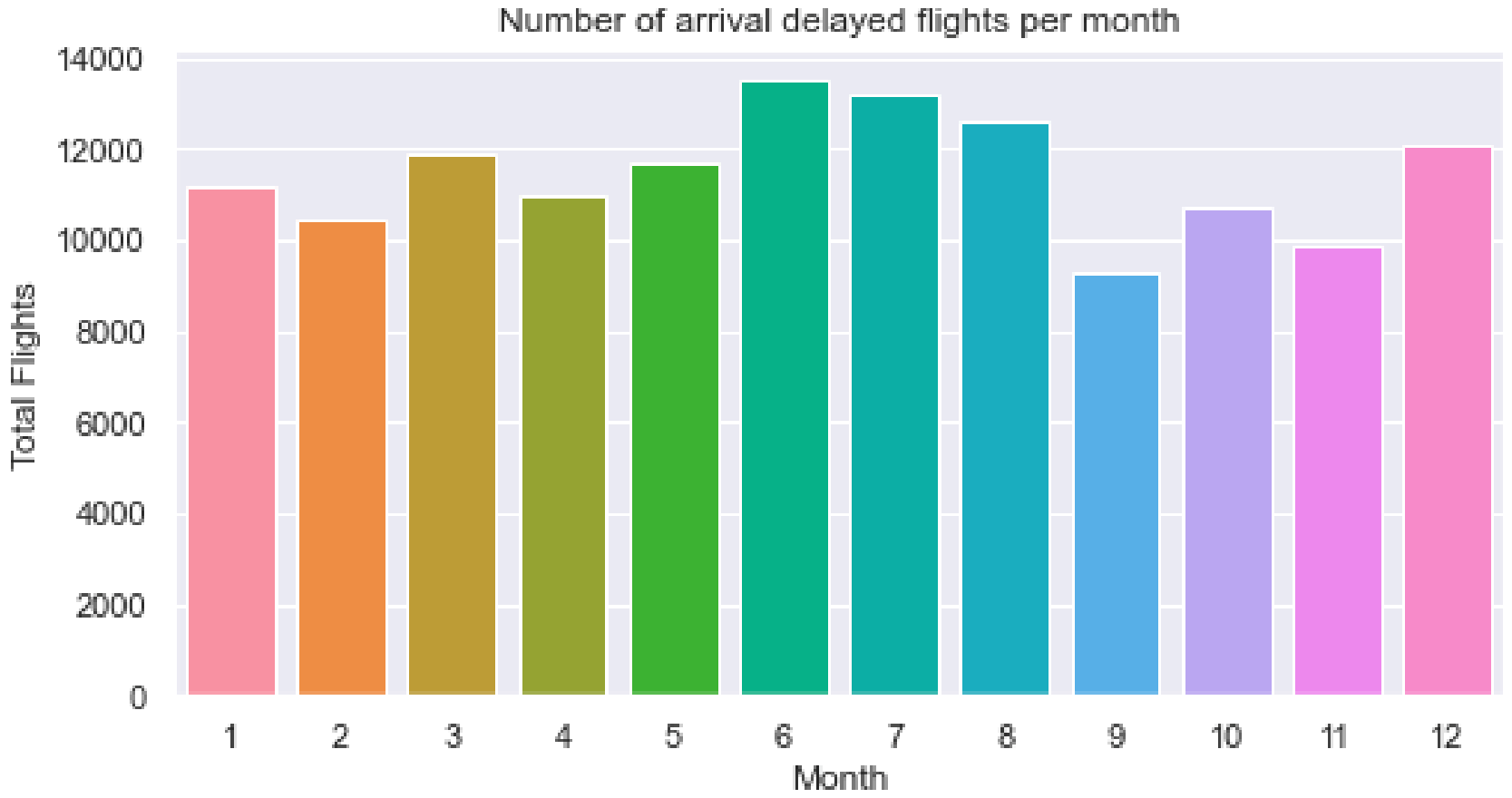
https://www.cnn.com/travel/article/worlds-busiest-airports-2021/index.html
https://usabynumbers.com/busiest-airports-in-the-us/

AIRPORTS WITH LEAST -> MOST FREQUENT DELAYS

	OriginAirport	Percentage_of_Flights_Delayed_on_Departure
336	IAH	34.905367669872 79%
41	ATL	36.4641121575710 66%
149	CLT	36.605621805792 17%
538	PHX	38.274556684598 53%
191	DFW	38.410375789823 74%
631	SFO	38.977294779739 96%
508	ORD	39.454235361000 57%
388	LAX	40.931653302787 325%
384	LAS	43.890675241157 56%

	OriginAirport	Percentage_of_Flights_Delayed_on_Arrival
41	ATL	32.542037586547 97%
336	IAH	36.094735753438 826%
538	PHX	37.876070930464 24%
384	LAS	38.413719185423 37%
388	LAX	38.564337533409 7%
632	SFO	40.122258878323 31%
151	CLT	40.353492333901 194%
191	DEN	40.374774774774 77%
193	DFW	41.769205187894 91%
508	ORD	41.841955656623 08%

Months Most Delayed



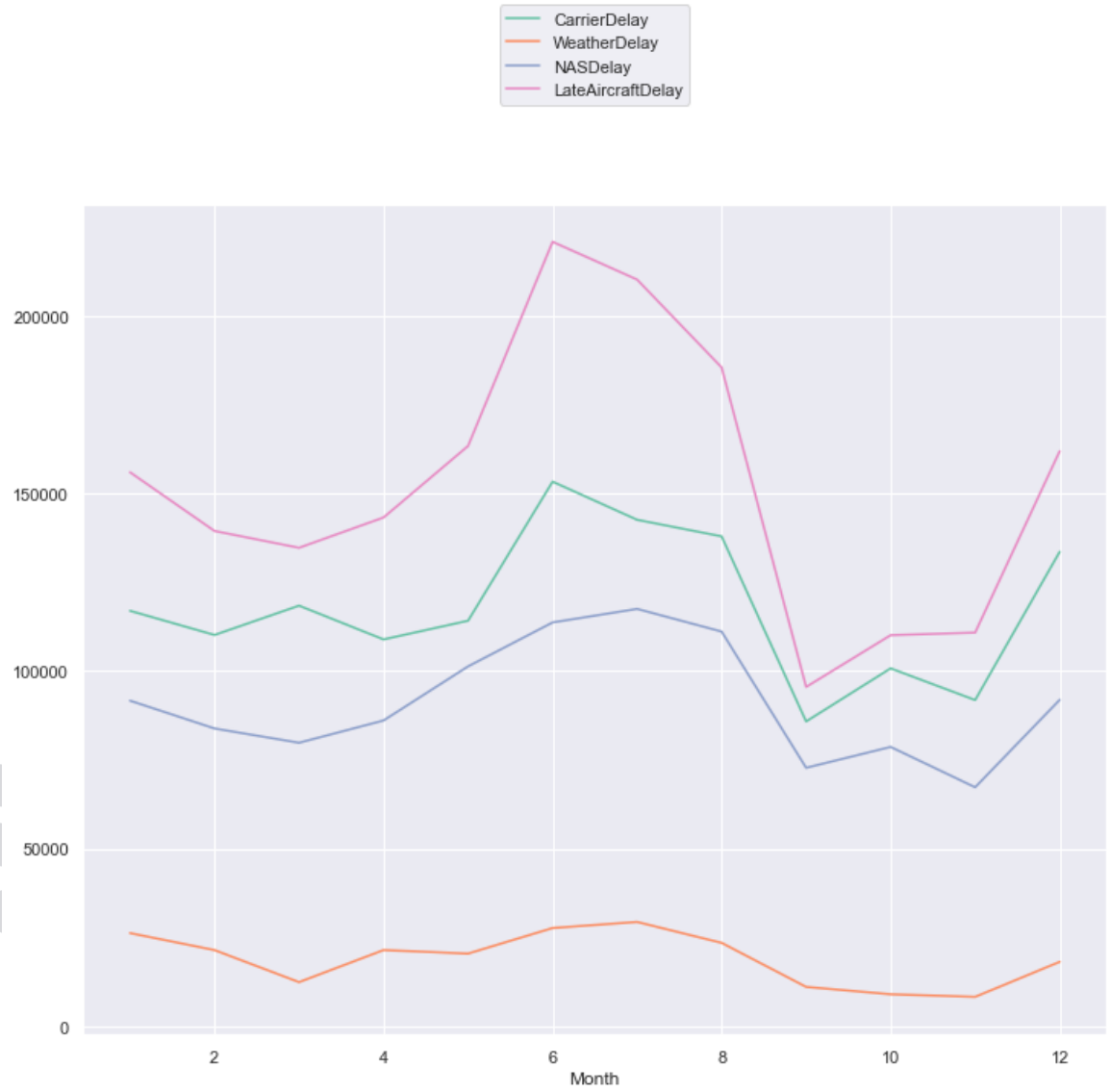
SPECIFIC DELAYED FACTORS

Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).

Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.

National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.

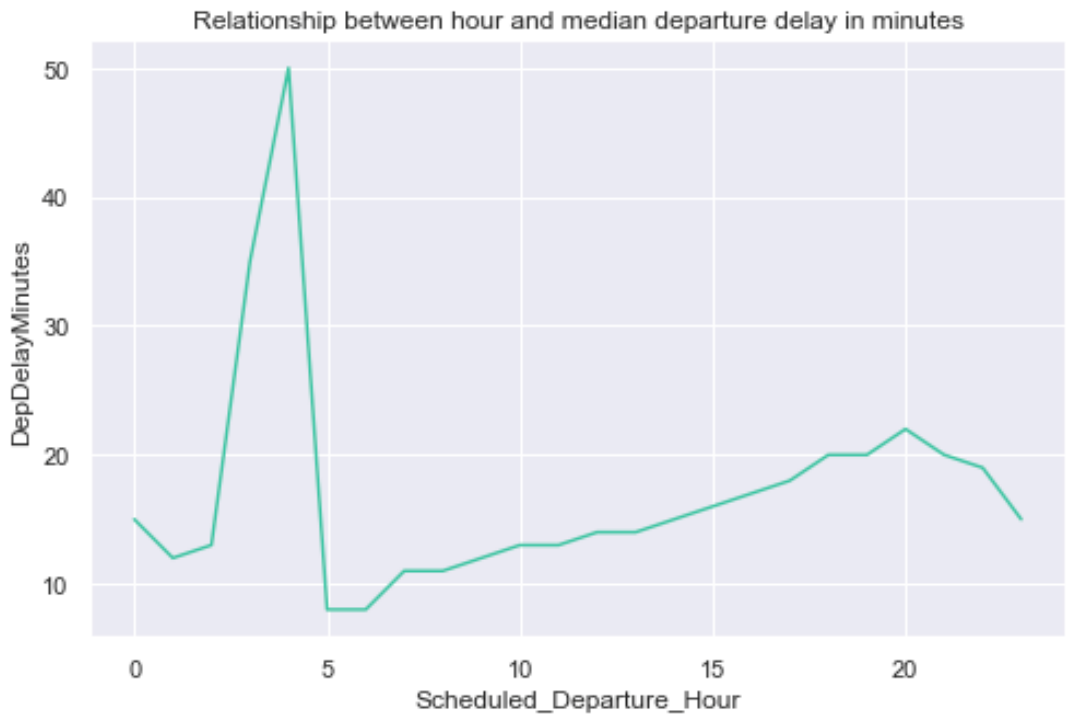
Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late – fun fact this is called delay propagation in the aviation world



Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics (bts.gov)

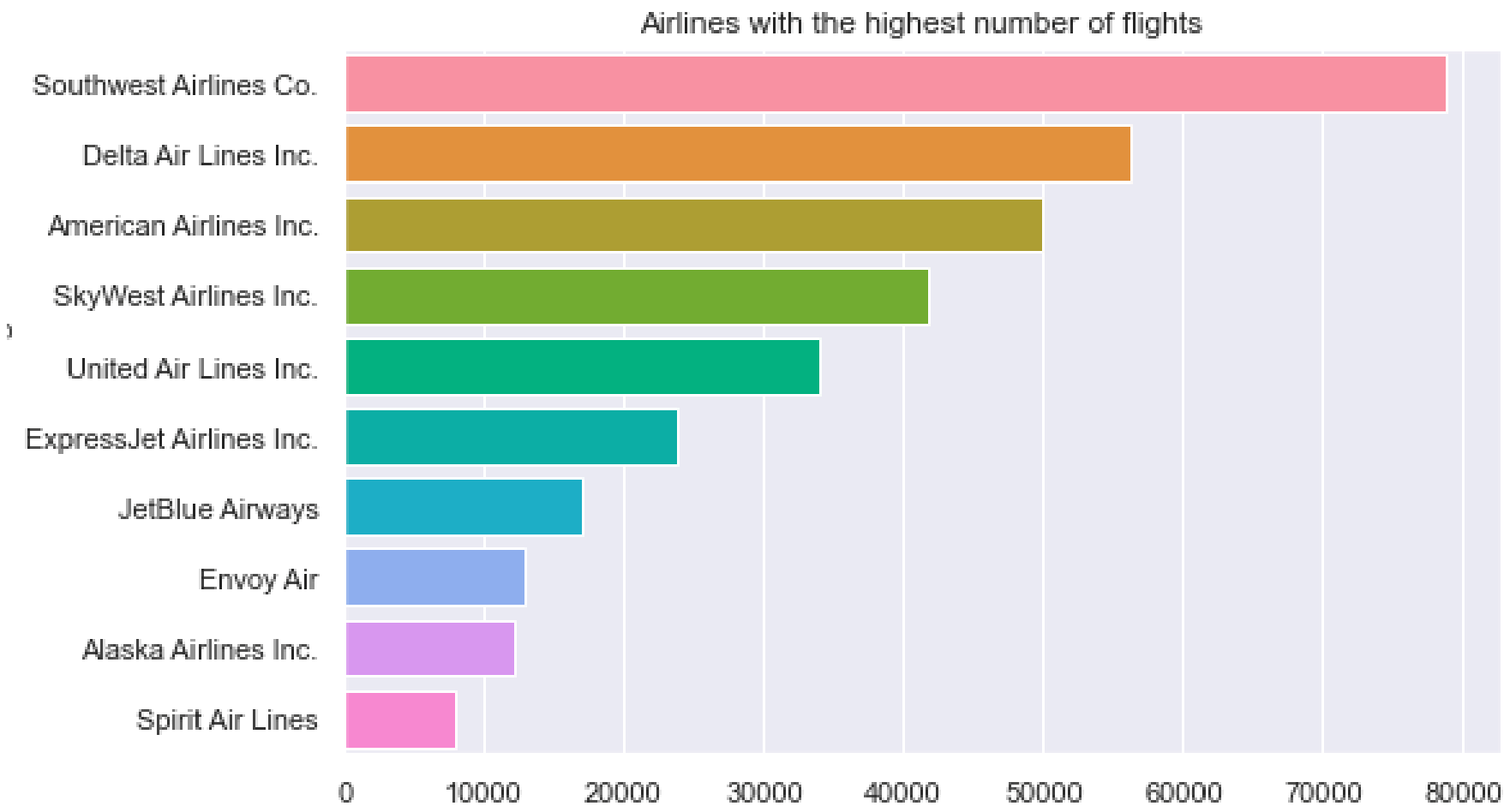
Time of Day most Delayed

Departure



Arrival

Airlines with highest number of flights



Airlines with least -> most frequent departure/arrival delay

Arrival: Envoy, JetBlue, Southwest

	Airline_Name	Percentage_of_Flights_Delayed_on_Arrival
9	Delta Air Lines Inc.	28.811143966653603%
37	United Air Lines Inc.	34.31723490405493%
3	Alaska Airlines Inc.	35.25262467191601%
29	SkyWest Airlines Inc.	36.04210627277959%
33	Spirit Air Lines	36.46702047005307%
7	American Airlines Inc.	37.55103674645745%
15	ExpressJet Airlines Inc.	37.97998408776852%
31	Southwest Airlines Co.	39.43540292970487%
21	JetBlue Airways	39.816124469589816%
13	Envoy Air	40.643046667707196%

Departure: Southwest, JetBlue, United

	Airline_Name	Percentage_of_Flights_Delayed_on_Departure
3	Alaska Airlines Inc.	27.903543307086615%
29	SkyWest Airlines Inc.	28.28985229234606%
9	Delta Air Lines Inc.	31.490256154476 47%
15	ExpressJet Airlines Inc.	31.82446296218751%
13	Envoy Air	33.088809115030436%
33	Spirit Air Lines	34.58428102097549%
7	American Airlines Inc.	34.700984708990475%
37	United Air Lines Inc.	38.304676955577726%
21	JetBlue Airways	40.21098538425271%
31	Southwest Airlines Co.	47.891654279579726%

MODELING

Feature Selection

7 Categorical features

- Airline
- Origin State
- Destination State
- Origin Airport
- Destination Airport
- Origin City
- Destination City

7 Numerical features

- Year
- Month
- Day
- Taxi In
- Taxi Out
- Wheels Off
- Wheels On

Encoding categorical data

- Origin and Destination data must have the same encoded values

  → use the same encoder for two columns

- Used LabelEncoder convert each categorical value into a number

- OneHotEncoder convert each value into a new column of 0s and 1s

Human-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

New dataframe for modeling

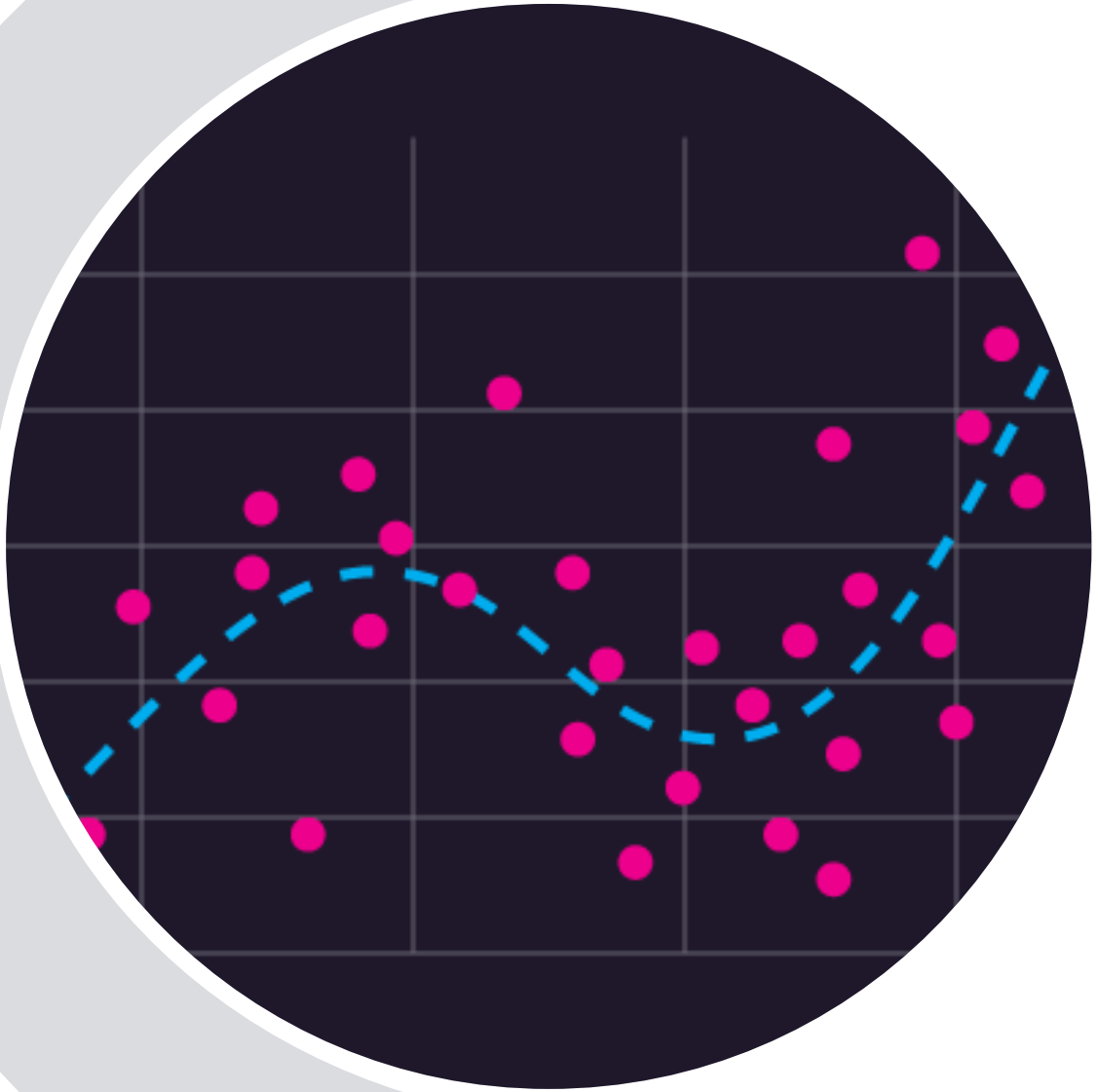
Combining the features columns and columns after encoded, we construct a new data frame to run models on.

► **airline_encoded_df**

```
numeric_cols = list(set(features).difference(set(cat_cols)))
otherCols = airline_df[numeric_cols].values
airline_encoded_df = np.hstack([otherCols,
                                encodedCityNameCols,
                                encoded_Origin_Dest_Cols,
                                encodedAirportsCols,
                                encodedAirlinesCols])
```

```
airline_encoded_df
```

```
matrix([[   6.,   19., 2019., ...,    0.,    0.,    0.],
        [   4.,   17., 2018., ...,    0.,    0.,    0.],
        [  10.,   20., 2019., ...,    0.,    0.,    0.],
        ...,
        [   3.,    9., 2014., ...,    0.,    0.,    0.],
        [   7.,   18., 2014., ...,    0.,    0.,    0.],
        [  11.,   11., 2014., ...,    0.,    0.,    0.]])
```



REGRESSION

LINEAR REGRESSION & DECISION TREE

Linear Regression

- Target variable: **DepDelayMinutes**
- Independent variables: 14 attributes mentioned above, including 7 encoded variables.
- Results:
  - Mean Squared Error (MSE): 1637.96
  - Coefficient of determination: 1.00
  - Root Mean Squared Error (RMSE): 40.47

Linear Regression

- The RMSE score (40.91%) shows that the average distance between the observed data values and the predicted data values is large. This model was not doing well in fitting the dataset

- With an R-squared score of 1.00, the model is indicated to be a perfect model where the fitting line fits perfectly with the data points. This contradicts the positive MSE.

- Plotting is not possible due to the difference between the shape of the independent variables and dependent variable. The lack of visualization is considered a big downsize of this approach.

- **This is not an ideal approach.**

Decision Tree

- Target variable: **DepDelayMinutes**
- Independent variables: 14 attributes mentioned above, including 7 encoded variables.
- Max depths: 2 and 5
- Results:

	Max depth = 2	Max depth = 5
MSE	1578.20	1578.20
R^2	1.00	1.00

Decision Tree

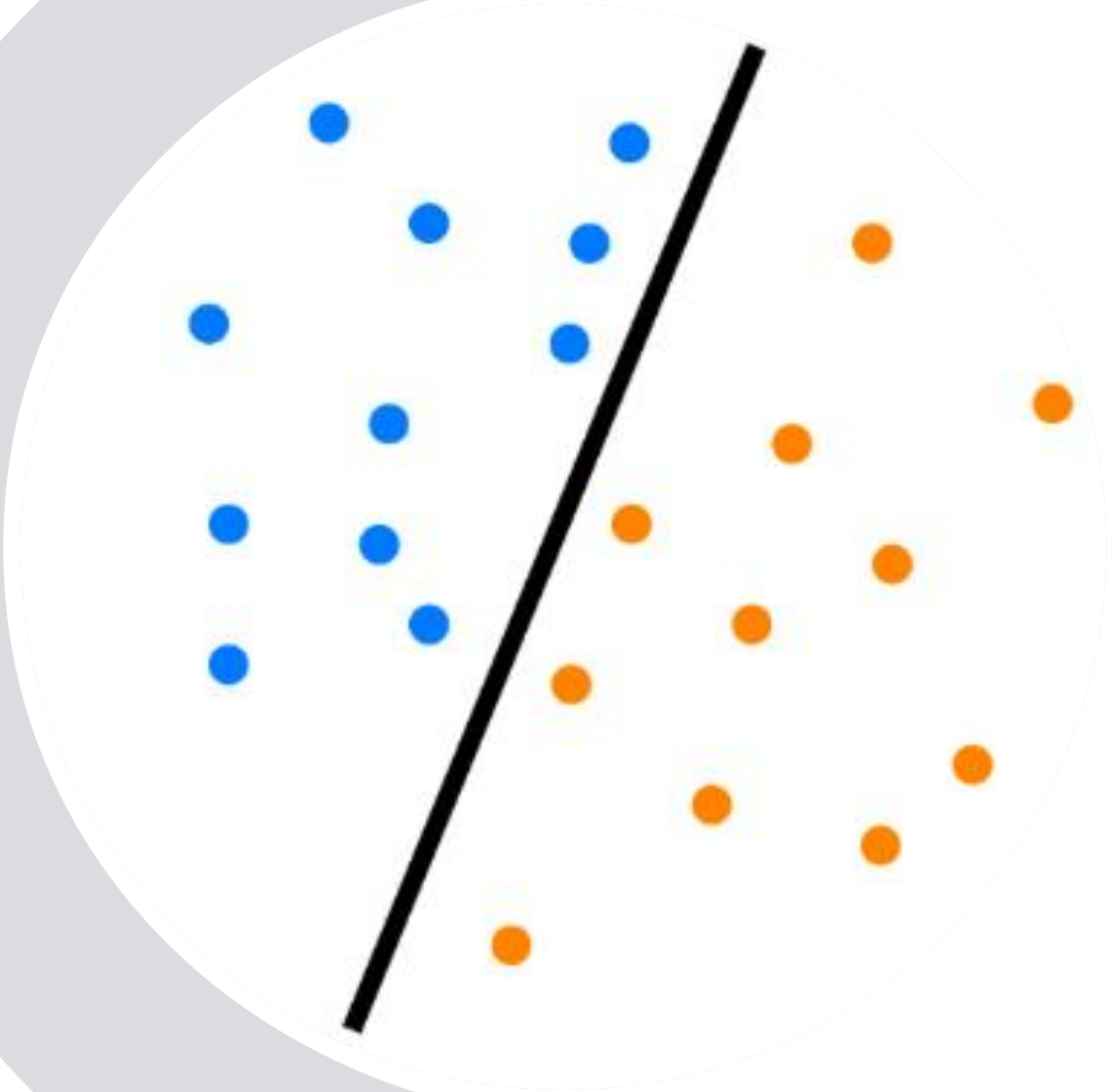
- Again, we witness the same issue with the MSE and R^2 score pair as the one in the linear regression analysis above.

➤ **This is not an ideal approach.**

CHALLENGE WITH REGRESSION MODELS

- One-hot encoding explodes the 7 categorical decision variables into a massive matrix of numerical values, which prevents regression models from fitting the training data efficiently.
- Visualization is not viable.

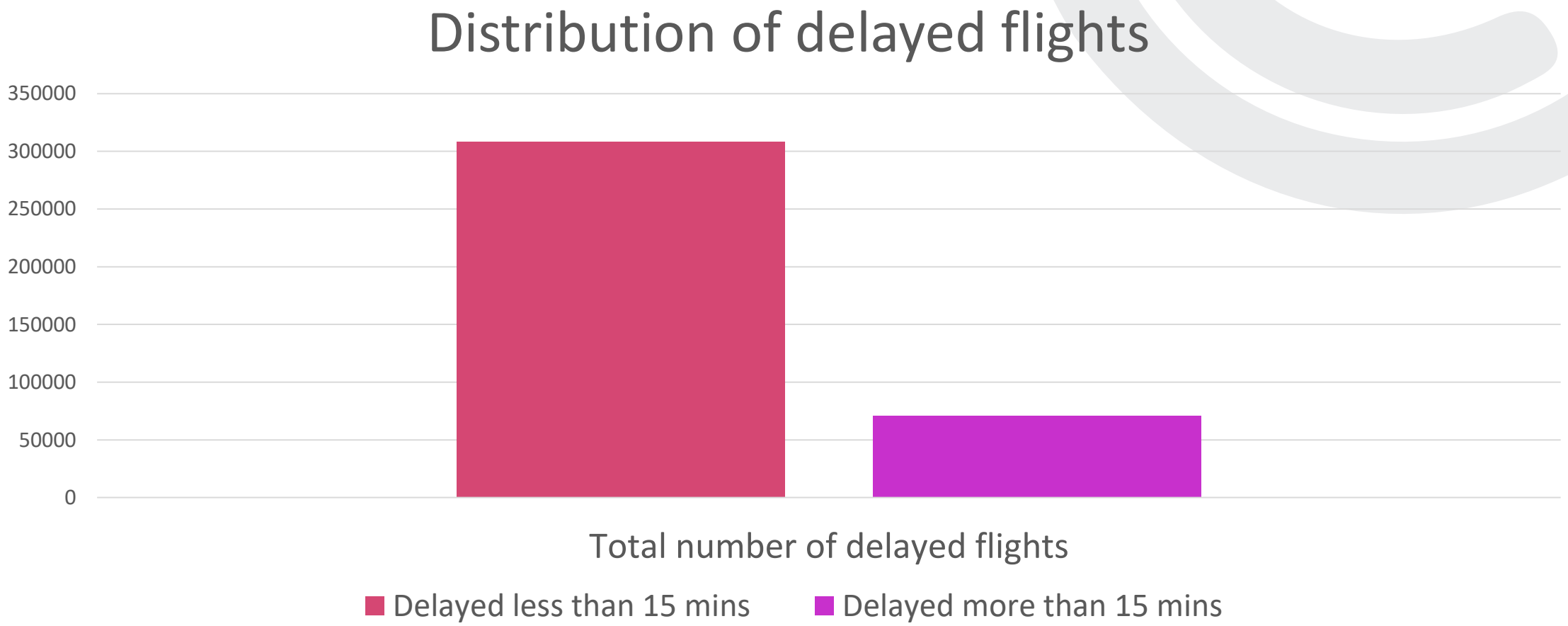
➤ **With half of the attributes being categorical, it is more reasonable to go for classification models.**



CLASSIFICATION

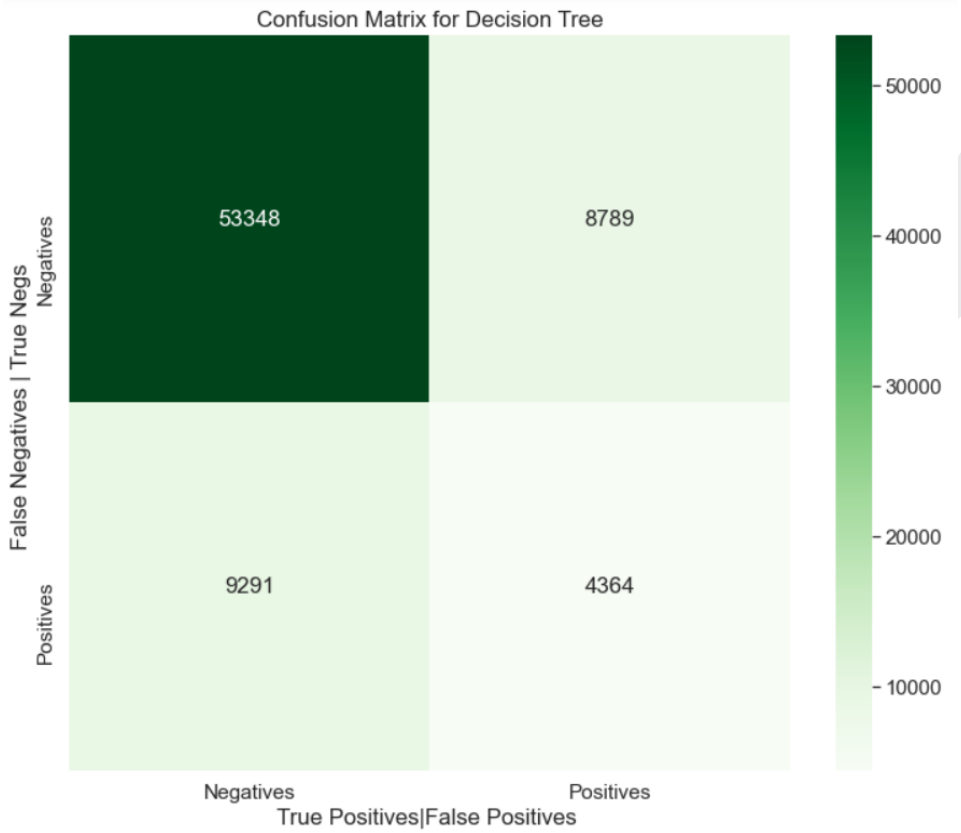
DECISION TREE & RANDOM FOREST

Create a binary label columns for classification



Decision Tree

	precision	recall	f1-score	support
0	0.85	0.86	0.86	62137
1	0.33	0.32	0.33	13655
accuracy			0.76	75792
macro avg	0.59	0.59	0.59	75792
weighted avg	0.76	0.76	0.76	75792



- Model accuracy is 76%.
- F-1 score for the positive label is around 33% which is very low

→ High occurrence of False Negative in prediction: Lots of flights with ***less than 15 minutes delay*** were classified as having ***more than 15 minutes delay***

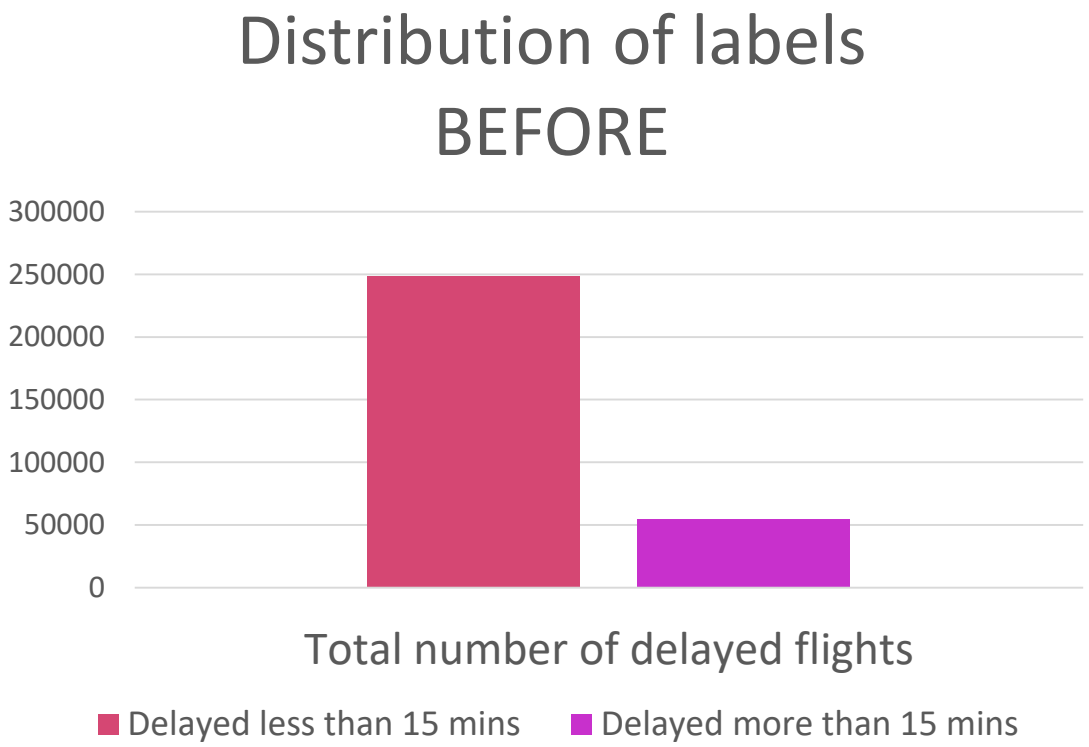
Reason: Highly skewed data (more flights with shorter delay time than flights with long delay time)

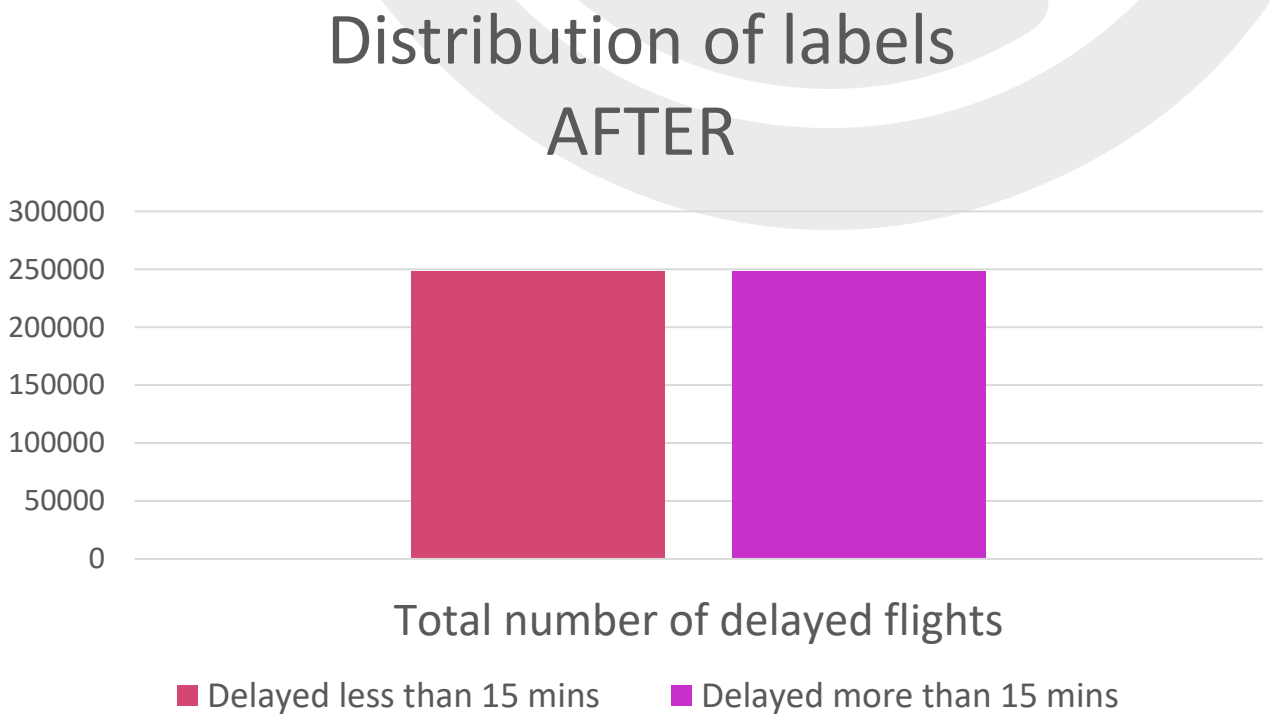
→ **Our approach**: Use oversampling:

- Randomly selected examples from the minority class (high delay time) and added them to the training dataset
- Dataset becomes more balanced

Over sampling with SMOTE

Data is skewed → Over-sampled the positive class in training data





Decision Tree with Over Sampling

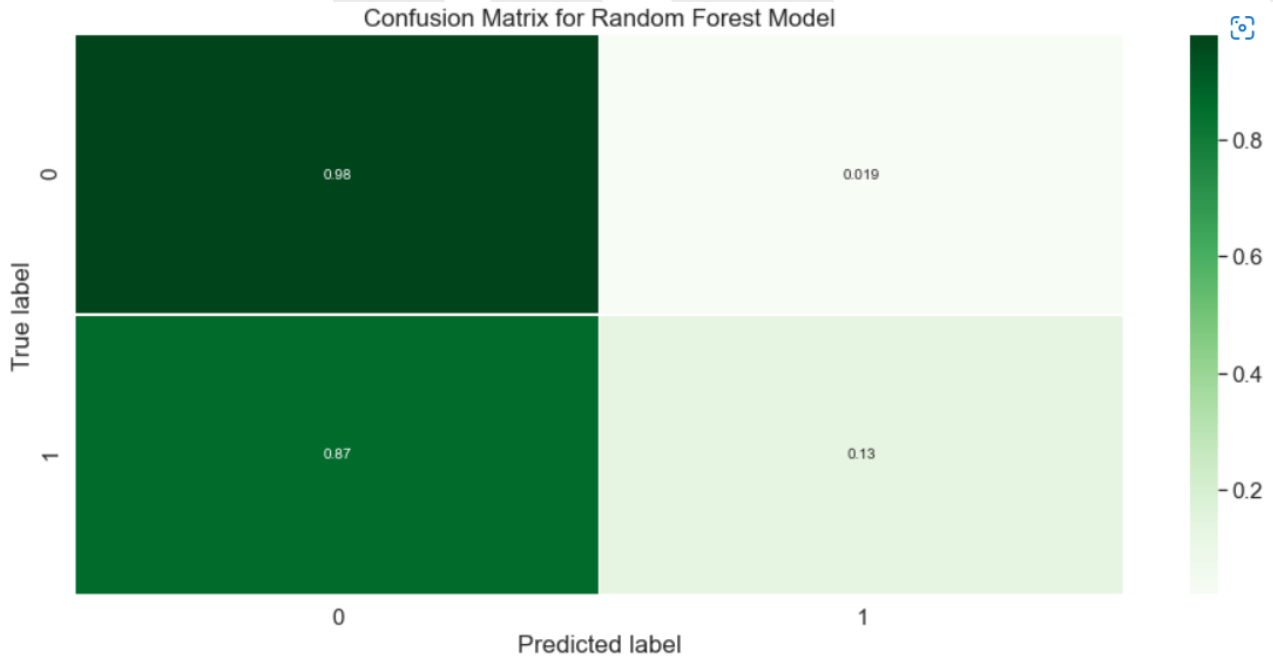
	precision	recall	f1-score	support
0	0.85	0.86	0.86	62137
1	0.35	0.34	0.34	13655
accuracy			0.77	75792
macro avg	0.60	0.60	0.60	75792
weighted avg	0.76	0.77	0.76	75792

- Model accuracy is 77% (an 1% increase)
- F-1 score for the negative label is around 86% (unchanged)
- F-1 score for the positive label is around 34% (an 1% increase)

→ This is **not** a considerable increase from F-1 score before oversampling
→ This model's still not doing well after oversampling approach

Random Forest with Over Sampling

	precision	recall	f1-score	support
0	0.84	0.98	0.90	62137
1	0.60	0.13	0.22	13655
accuracy			0.83	75792
macro avg	0.72	0.56	0.56	75792
weighted avg	0.79	0.83	0.78	75792



- Decent accuracy score: 82.8%
- F-1 score for the positive label is around 22%
- F-1 score for the negative label is around 90%
- Decent Weighted F-1 score: 78%

COMPARISON

1 DECISION TREE

- Decent accuracy score: 77%
- F-1 score for the negative label is around 86%
- F-1 score for the positive label is around 34%
- Average Weighted F-1 score: 60%

2 RANDOM FOREST

- Decent accuracy score: 82.8%
- F-1 score for the negative label is around 90%
- F-1 score for the negative label is around 90%
- Decent Weighted F-1 score: 78%

Performances are similar: Random Forest is better with Accuracy Score, and Decision Tree has higher F-1 score

OUTCOMES & IMPACTS

Steps to increase model performance

Penalizing the model - imposes an additional cost on the model for making classification mistakes on the minority class during training

Hyper-parameter tuning using gridsearch, randomSearch, Optuna, XGBoost, or Bayesian Optimization; Using Undersampling together with Oversampling

Re-analyzing data download/processing methods & examine feature importance

Utilizing cloud clusters to work with more complete datasets

https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Conclusions

EDA:

- Southwest and JetBlue has the most frequent delays
- Las Vegas and Chicago has the highest delay rate
- average of 36% of all flights are delayed

Modeling:

- Classification preferred over regression
- Utilizing SMOTE negatively impacted our models
- Decision Trees performed (comparatively) well to Random Forest

THANK YOU

ANY QUESTIONS/COMMENTS/CONCERNS?