To Fly or Not to Fly

MEET THE TEAM

JANET
Storyteller

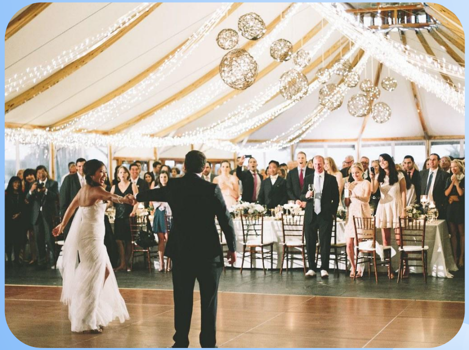
ERIC
Data Collector Guru

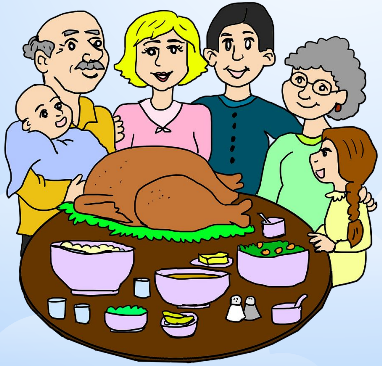
NEESHA
Data Cruncher

RYAN
Data Viz Whiz

The Story

Delayed or canceled flights can ruin business trips, vacations, family events, and so much more.





$28 Billion

FAA/Nextor estimated the annual costs of delays in 2018

Questions:

1. How does weather impact flight cancellations?
2. Are certain weather events impact the decision to cancel more than other?
3. Are certain airlines more prone to cancel flights based on weather?

Tools Used

Dashboard
Google Slides

Data Collection
Python, Kaggle, Kaggle API, PostgreSQL, SQLAlchemy, Quick DBD Psycopg2, pgAdmin, Jupyter Notebook

Data Viz
Matplotlib, D3, Leaflet.

Data Analysis
Machine Learning, Python 3.7, Scikit-learn

ERIC

Data Collector Guru

- Data Acquisition
- Preprocessing
- Database Storage
- Data Retrieval

Data Sources

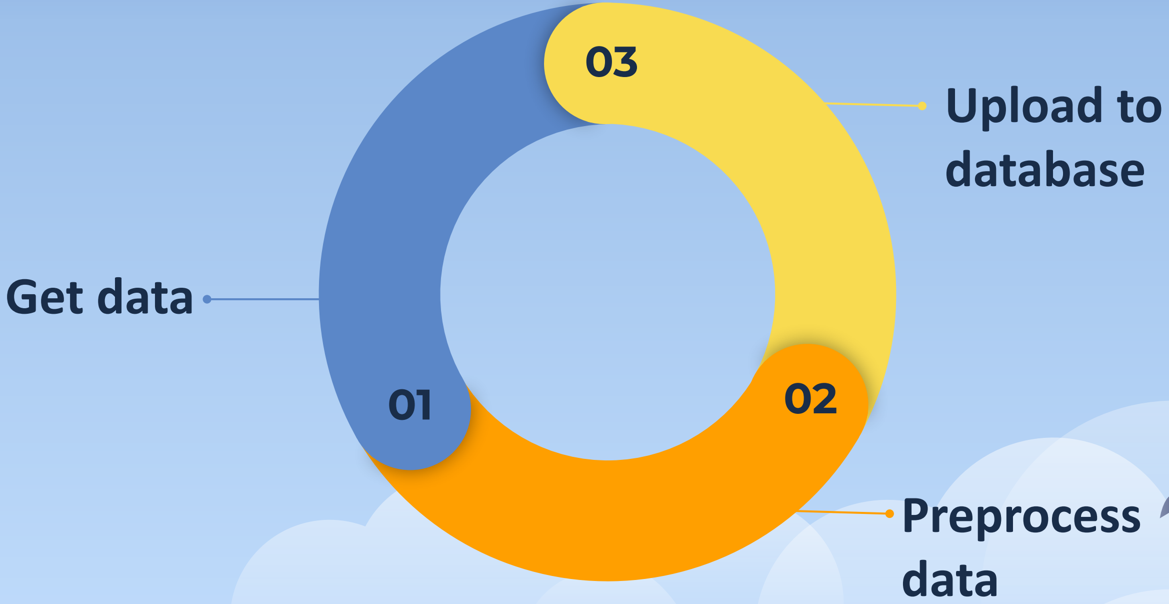
1. KAGGLE, Historical Flight Delay and Weather Data USA
   - United States Bureau of Transportation Statistics
   - National Oceanic and Atmospheric Administration
2. The Global Airport Database

Data Sources

1. KAGGLE, Historical Flight Delay and Weather Data USA

2. The Global Airport Database

Extract-Transform-Load



Download Datasets

Download Primary Dataset

```
import kaggle
from kaggle.api.kaggle_api_extended import KaggleApi
```

```
kag = KaggleApi()
kag.authenticate()
```

```
# Download primary dataset from Kaggle
kag.dataset_download_files(
    dataset=datasource_primary,
#     unzip=True,
    path=data_dir,
)

print('Download complete.')
```

```
Download complete.
```

Download Secondary Dataset

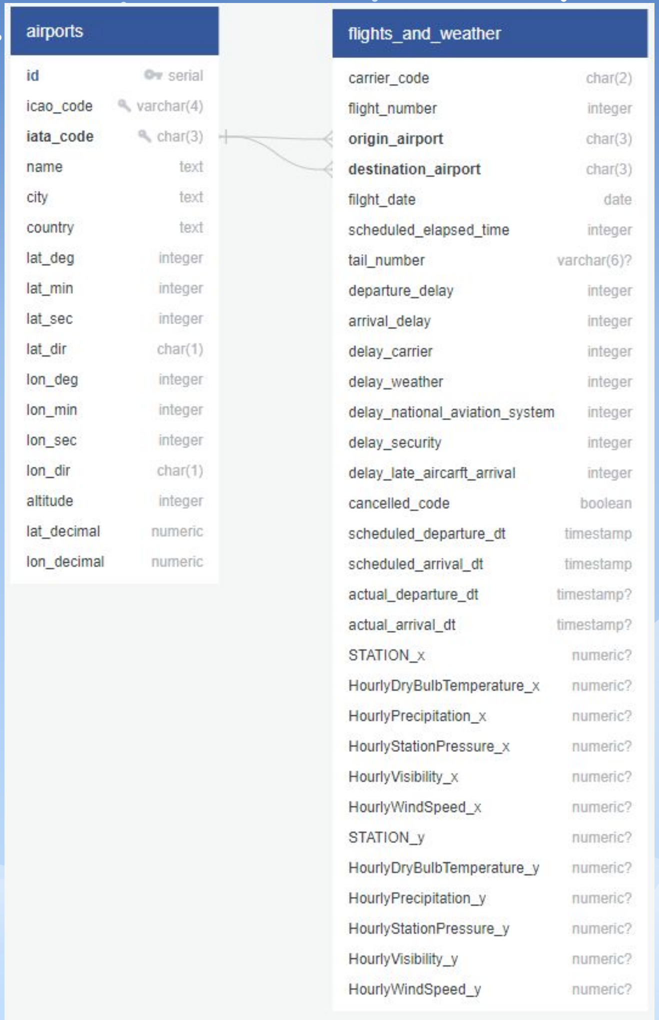
```
import requests
```

```
# Download the secondary dataset
response = requests.get(datasource_secondary)

try:
    with open(dataset_secondary, 'xb') as dl_file:
        for chunk in response.iter_content(chunk_size=128):
            dl_file.write(chunk)
        print('Download complete.')
except FileExistsError:
    print('Download complete. (File already exists.)')
```

```
Download complete.
```

Preliminary Entity Relationship Diagram



Airports Data

Flights and Weather Data

Airports Data

- General text conversion

Flights and Weather Data

Airports Data

- General text conversion
- Column definitions: data types and unique values

Flights and Weather Data

Airports Data

- General text conversion
- Column definitions: data types and unique values

Flights and Weather Data

- Redundant data: date vs year-month-day

Airports Data

- General text conversion
- Column definitions: data types and unique values

Flights and Weather Data

- Redundant data: date vs year-month-day
- Irrelevant data: cancellation types

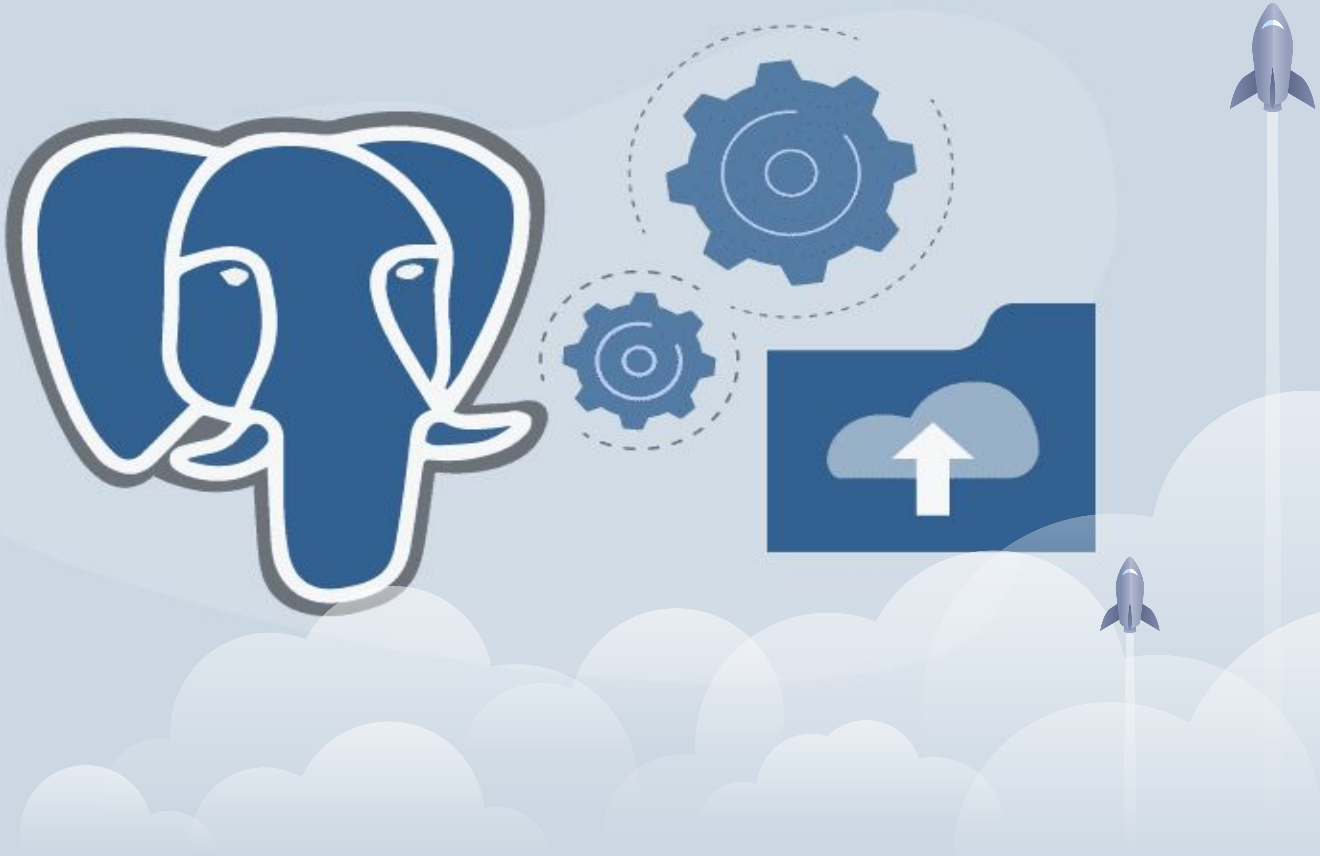
Airports Data

- General text conversion
- Column definitions: data types and unique values

Flights and Weather Data

- Redundant data: date vs year-month-day
- Irrelevant data: cancellation types
- PostgreSQL complication: FOREIGN KEY CONSTRAINT

PostgreSQL Database Integration



5,468,069

Rows….Whoa! That’s a lot of data!

NEESHA

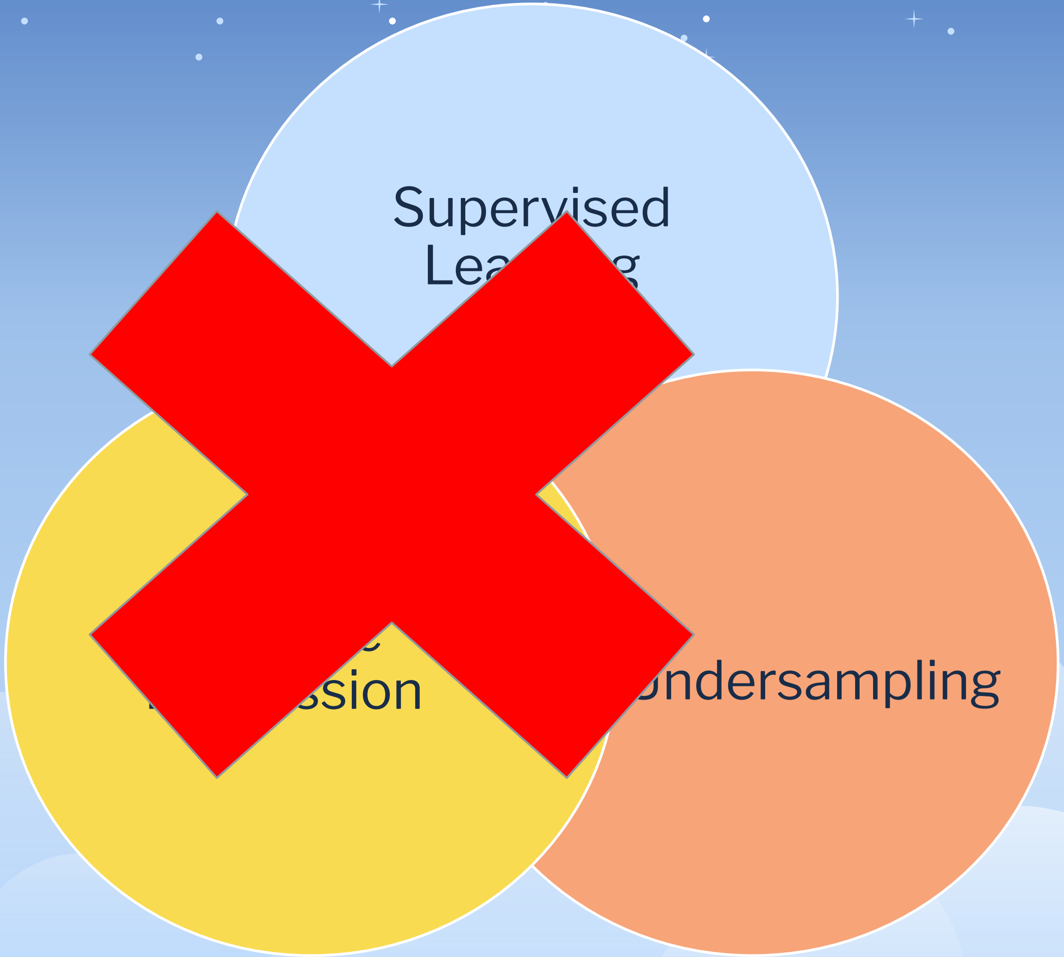
Data Cruncher

Description of the analysis phase of the project.

Machine Learning Model.

Analysis results.

Machine Learning Model



Prep for Machine Learning Model

Merging the tables: After merging the flight_weather data with the airport data, we had 5468069 rows × 36 columns to analyze the data.

5468067	5512901	DL	2436	ATL
5468068	5512902	DL	3826	ATL

5468069 rows × 36 columns

Datasets and Airline Codes

Airports for which weather data was available.

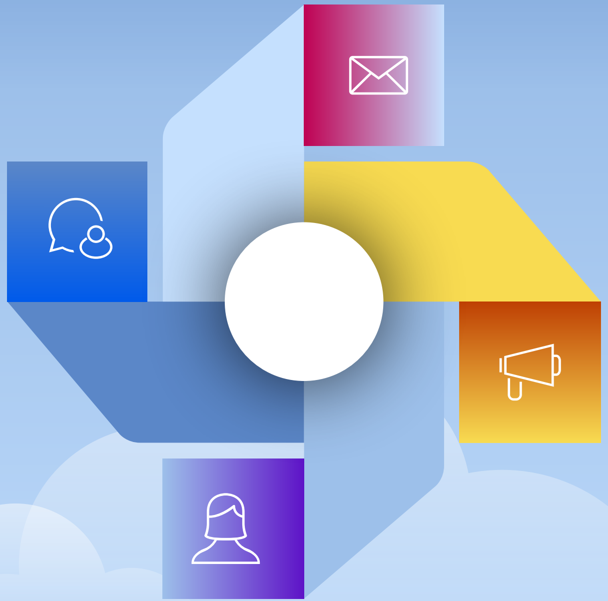
AA	American Airlines	G4	Allegiant Air
AS	Alaska Airlines Inc	HA	Hawaiia
B6	JetBlue Airways	NK	Spirit Air Lines
DL	Delta Air Lines Inc	UA	United Air
F9	Frontier Airlines Inc	WN	Southwest A

Cancellation Codes

	Carrier Caused
	Weather
C	National Aviation System
	Security

Note: N is not on the list and represents "None" or "Not cancelled".

Features Selections



01 **Missing values**

02 **Label Encoding**

03 **Normalize values and scale**

04 **Feature creation**

Features Selection:

Flight and Delayed reasons other than weather conditions

Kept	Dropped
Origin_airport, origin_lat, origin_lon,	Id, carrier code
'Destination_airport, destination_lat, destination_lon	delay_national_aviation_system', 'delay_security
'departure_delay	delay_security
'arrival_delay'	delay_late_aircarft_arrival'

Weather parameters:

Kept	
Hourlydrybulbtemperature_x hourlyprecipitation_x hourlystationpressure_x hourlyvisibility_x hourlywindspeed_x	hourlydrybulbtemperature_y hourlyprecipitation_y hourlystationpressure_y hourlyvisibility_y hourlywindspeed_y

Data clean up

Dataset after removing the unwanted columns

5468069 rows × 21 columns

Removal of Missing Values

hourlydrybulbtemperature_x	2073
hourlyprecipitation_x	9881
hourlystationpressure_x	2073
hourlyvisibility_x	2073
hourlywindspeed_x	2073
station_y	2078
hourlydrybulbtemperature_y	2078
hourlyprecipitation_y	9896
hourlystationpressure_y	2078
hourlyvisibility_y	2078
hourlywindspeed_y	2078
origin_lat	382438
origin_lon	382438
destination_lat	382775
destination_lon	382775
dtype: int64	

hourlydrybulbtemperature_x	0
hourlyprecipitation_x	0
hourlystationpressure_x	0
hourlyvisibility_x	0
hourlywindspeed_x	0
station_y	0
hourlydrybulbtemperature_y	0
hourlyprecipitation_y	0
hourlystationpressure_y	0
hourlyvisibility_y	0
hourlywindspeed_y	0
origin_lat	0
origin_lon	0
destination_lat	0
destination_lon	0
dtype: int64	

Prep for Machine Learning Model

```
df_new['cancelled'].value_counts()
```

```
f    4674943
t      33957
Name: cancelled, dtype: int64
```

After Undersampling

```
df_new_f.shape
```

```
(33957, 21)
```

```
33957 * 2
```

```
67914
```

```
df_new_t.shape
```

```
(33957, 21)
```

```
df_final = pd.concat
```

```
df_final.shape
```

```
(67914, 21)
```

Splitting into training and test datasets

```
# Use the train_test_split function to create training and testing subsets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,
    y,  random_state=1, stratify=y, test_size=0.1)
X_train.shape
```

```
(61122, 20)
```

Label Encoding:

```
from sklearn.preprocessing imp
```

```
Le = LabelEncoder()
```

```
y_train[:5]
```

```
22488    f
47918    t
14565    t
11452    f
3805     f
Name: cancelled, dtype: object
```

```
y_train_cln
```

```
array([0, 1, 1, ..., 1, 1, 0])
```

Standard Scaler

	origin_airport	destination_airport	departure_delay	arrival_delay	station_x	hourlydrybulbtemperature_x
22488	SAN	SAT	2	-6	7.229002e+10	75.0
47918	DFW	other	0	0	7.225900e+10	73.0

```
X_train_cln
```

```
array([[ 0.        ,  0.        ,  0.        , ..., -1.47694024,
        -1.26551277, -0.37827086],
       [ 0.        ,  0.        ,  0.        , ..., -0.28838617,
         1.2004544 , -1.43207941],
```

Using Binary Classification and Running Logistic Regression

```
LogisticRegression(random_state=1)
```

```python
# Make predictions using the test data
y_pred = classifier.predict(X_test_cln)
results = pd.DataFrame({
     "Prediction": y_pred,
     "Cancelled": y_test_cln
}).reset_index(drop=True)
results.head()
```

	Prediction	Cancelled
0	0	1
1	1	0
2	0	0
3	1	1
4	1	1

```python
# Validate using test data
from sklearn.metrics import accuracy_score
accuracy_score(y_test_cln, y_pred)
```

```
0.71849234393404
```

LOGISTIC CONFUSION MATRIX

		predicted	
		cancelled	not cancelled
actual	cancelled	[illegible]	959
	not cancelled	985	2411

mlm_t01_logistic_confusion_matrix.png

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.71	[illegible]	0.71	3396
1	[illegible]	[illegible]	0.71	3396
accuracy			0.71	6792
macro avg	0.71	0.71	0.71	6792
weighted avg	0.71	0.71	0.71	6792

mlm_t01_logistic_classification_report.png

Logistic Confusion Matrix

		predicted	
		cancelled	not cancelled
actual	cancelled	2437	959
	not cancelled	985	2411

Classification Report

	precision	recall	f1-score	support
0	0.71	0.72	0.71	3396
1	0.72	0.71	0.71	3396
accuracy			0.71	6792
macro avg	0.71	0.71	0.71	6792
weighted avg	0.71	0.71	0.71	6792

RYAN,

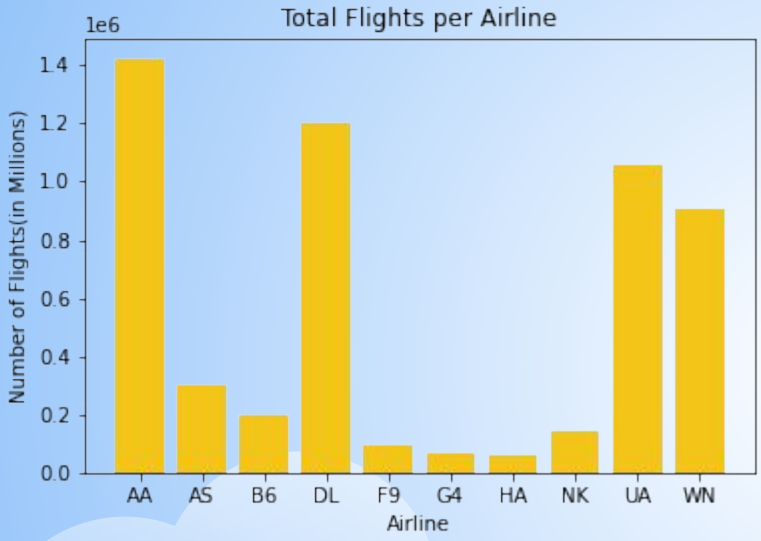
Data Viz Whiz

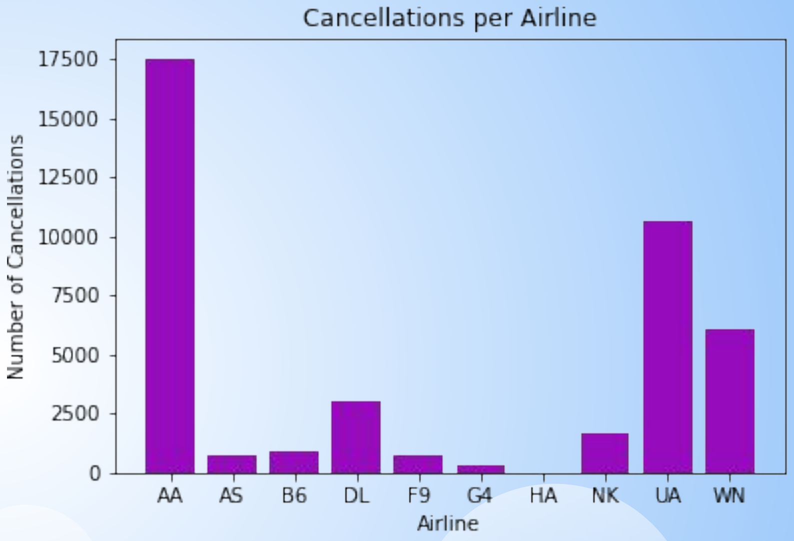
Visualization of analysis

Recommendation for future analysis

Anything the team would have done differently

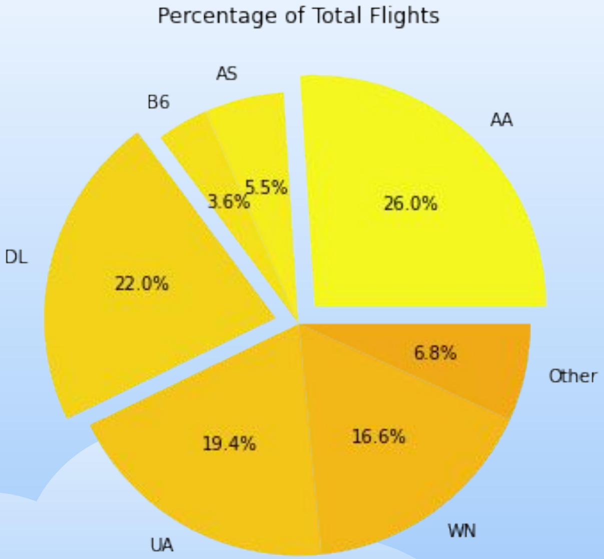
Airline Observations

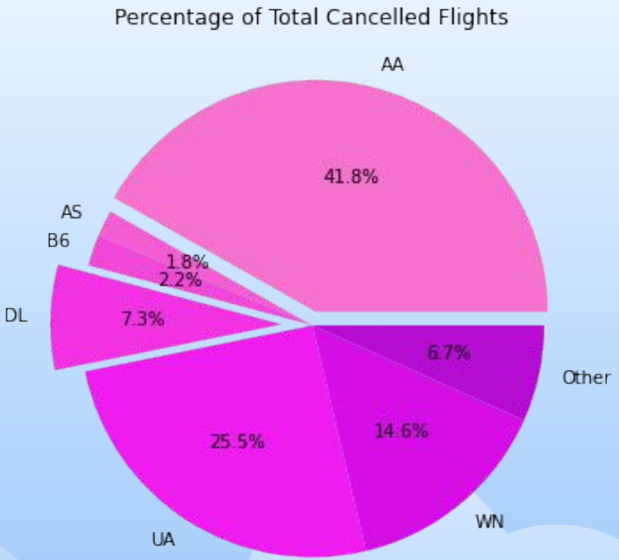




We noticed that although Delta Airlines (DL) is the second highest carrier by volume of total flights, they are fourth in number of cancellations. We wanted to look into this further, so we looked at percentages of total flights vs. percentages of cancellations (next slide).

Airline Observations





American Airlines (AA) accounts for about 26% of total flight volume, but about 42% of cancellations due to weather, whereas Delta (DL) accounts for about 22% of total flights but only about 7% of cancelled flights due to weather.

Further Visualizations

In the next segment we will create:

1. Line chart by month
2. Possibly an interactive map of airports around the country
3. Possibly a chart of cancellations by flight route
4. Possibly any visualizations from interesting findings when we run the ML model

FINAL THOUGHTS

Recommendations for future analysis

Things the team would have done differently

QUESTIONS ?


BUSINESS
Economy of the European Union
25 great jobs for people who love to travel
Business Company

Thank You!