Flight Delay Prediction

By Group #3: Fatima, Travers, Matt, Yunkai

About the Project

✓ Selected topic and the Reason why we selected the topic

✓ Description of our source of data

✓ Questions we hope to answer with the data

✓ Description of the data exploration phase of the project

✓ Description of the analysis phase of the project

Flight Delays

In 2018, 18 % of all domestic flights were delayed.

As there were > 6 million flights in that year, that means 1,080,000 were delayed.

It is very frustrating even if you just experienced once.

What can we do?

We will use machine learning algorithms with the flight data, weather data, airport data to predict if a flight will be delayed or not. Let people have some sort of certainty with their travels.

Source of Data

Raw Data Source

https://www.transtats.bts.gov/databases.asp?Z1qr_VQ=E&Z1qr_Qr5p=N8vn6v10&f7owrp6_VQF=D

https://www.ncdc.noaa.gov/cdo-web/datasets

https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations?select=raw_data

Description of Data

We will use:
Airport_weather_2019.csv
Airport_list.csv
ONTIME_REPORTing_01.csv
…

- airport_weather_2019.csv
- airport_weather_2020.csv
- airports_list.csv
- ONTIME_REPORTING_2020_03.csv
- P10_EMPLOYEES.csv
- T3_AIR_CARRIER_SUMMARY_AIRPORT_ACTIVITY_2019.csv
- T3_AIR_CARRIER_SUMMARY_AIRPORT_ACTIVITY_2020.csv
- ONTIME_REPORTING_2020_01.csv
- ONTIME_REPORTING_2020_02.csv
- ONTIME_REPORTING_12.csv
- ONTIME_REPORTING_11.csv
- ONTIME_REPORTING_09.csv
- ONTIME_REPORTING_10.csv
- ONTIME_REPORTING_08.csv
- ONTIME_REPORTING_07.csv
- ONTIME_REPORTING_05.csv
- ONTIME_REPORTING_06.csv
- ONTIME_REPORTING_04.csv
- ONTIME_REPORTING_03.csv
- ONTIME_REPORTING_01.csv
- ONTIME_REPORTING_02.csv
- AIRPORT_COORDINATES.csv
- B43_AIRCRAFT_INVENTORY.csv
- CARRIER_DECODE.csv

Questions to Answer with the data

- Which carriers are most and least reliable for on-time departure?
- Which airports are best and worst for on-time departures?
- Which features in the data set are most correlated with a departure delay?
- Use the dataset to make predictions. Can you accurately predict a departure delay?
- Use the raw data files to re-tool the dataset and make our own prediction problem. Can we predict the reason for departure delay? Can we predict arrival delay?

Data Exploration

Airport_weather_2019.csv

NAME	STATION	DATE	AWND	PGTM	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	...	WT08	WT09	WESD	WT10	PSUN	TSUN	SN32	SX32	TOBS
ALBANY INTERNATIONAL AIRPORT, NY US	365	365	365	0	364	365	365	365	365	365	...	18	5	0	0	0	0	0	0	0
ALBUQUERQUE INTERNATIONAL AIRPORT, NM US	365	365	365	8	365	365	365	365	365	365	...	16	2	0	0	0	0	0	0	0
ANCHORAGE TED STEVENS INTERNATIONAL AIRPORT, AK US	365	365	365	361	365	365	365	365	365	365	...	25	0	0	0	0	0	0	0	0
ASHEVILLE AIRPORT, NC US	365	365	364	0	365	365	365	365	365	365	...	40	0	0	0	0	0	0	0	0
ASPEN PITKIN CO AIRPORT SARDY FIELD, CO US	365	365	365	31	365	0	0	0	365	365	...	80	1	0	0	0	0	0	0	0

Drop Columns

	NAME	DATE	AWND	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN
0	ATLANTA HARTSFIELD JACKSON INTERNATIONAL AIRPO...	1/1/2019	4.70	0.14	0.0	0.0	64.0	66.0	57.0
1	ATLANTA HARTSFIELD JACKSON INTERNATIONAL AIRPO...	1/2/2019	4.92	0.57	0.0	0.0	56.0	59.0	49.0
2	ATLANTA HARTSFIELD JACKSON INTERNATIONAL AIRPO...	1/3/2019	5.37	0.15	0.0	0.0	52.0	55.0	51.0

airport_list

	ORIGIN_AIRPORT_ID	DISPLAY_AIRPORT_NAME	ORIGIN_CITY_NAME	NAME
0	12992	Adams Field	Little Rock, AR	NORTH LITTLE ROCK AIRPORT, AR US
1	10257	Albany International	Albany, NY	ALBANY INTERNATIONAL AIRPORT, NY US
2	10140	Albuquerque International Sunport	Albuquerque, NM	ALBUQUERQUE INTERNATIONAL AIRPORT, NM US
3	10299	Anchorage International	Anchorage, AK	ANCHORAGE TED STEVENS INTERNATIONAL AIRPORT, A...
4	10397	Atlanta Municipal	Atlanta, GA	ATLANTA HARTSFIELD JACKSON INTERNATIONAL AIRPO...

Link Weather with Airport

	DATE	AWND	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	DEST_AIRPORT_ID	DISPLAY_AIRPORT_NAME
0	2019-01-01	4.70	0.14	0.0	0.0	64.0	66.0	57.0	10397.0	Atlanta Municipal
1	2019-01-02	4.92	0.57	0.0	0.0	56.0	59.0	49.0	10397.0	Atlanta Municipal
2	2019-01-03	5.37	0.15	0.0	0.0	52.0	55.0	51.0	10397.0	Atlanta Municipal
3	2019-01-04	12.08	1.44	0.0	0.0	56.0	66.0	45.0	10397.0	Atlanta Municipal
4	2019-01-05	13.42	0.00	0.0	0.0	49.0	59.0	44.0	10397.0	Atlanta Municipal

ON-TIME-REPORTING _01.csv

	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	OP_UNIQUE_CARRIER	TAIL_NUM	OP_CARRIER_FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	ORIGIN_CITY_NAME	DEST_AIR
0	1	6	7	9E	N8694A	3280	10397	ATL	Atlanta, GA	
1	1	7	1	9E	N8970D	3280	10397	ATL	Atlanta, GA	
2	1	8	2	9E	N820AY	3280	10397	ATL	Atlanta, GA	
3	1	9	3	9E	N840AY	3280	10397	ATL	Atlanta, GA	
4	1	10	4	9E	N8969A	3280	10397	ATL	Atlanta, GA	
...	...	...	...	...	...	...	...	...	...	
583980	1	30	3	UA	N819UA	2024	14683	SAT	San Antonio, TX	
583981	1	30	3	UA	N37462	2022	14843	SJU	San Juan, PR	
583982	1	30	3	UA	N37462	2021	12264	IAD	Washington, DC	
583983	1	30	3	UA	N26967	2020	12266	IAH	Houston, TX	
583984	1	30	3	UA	N821UA	2017	12266	IAH	Houston, TX	

583985 rows × 33 columns

Join the data

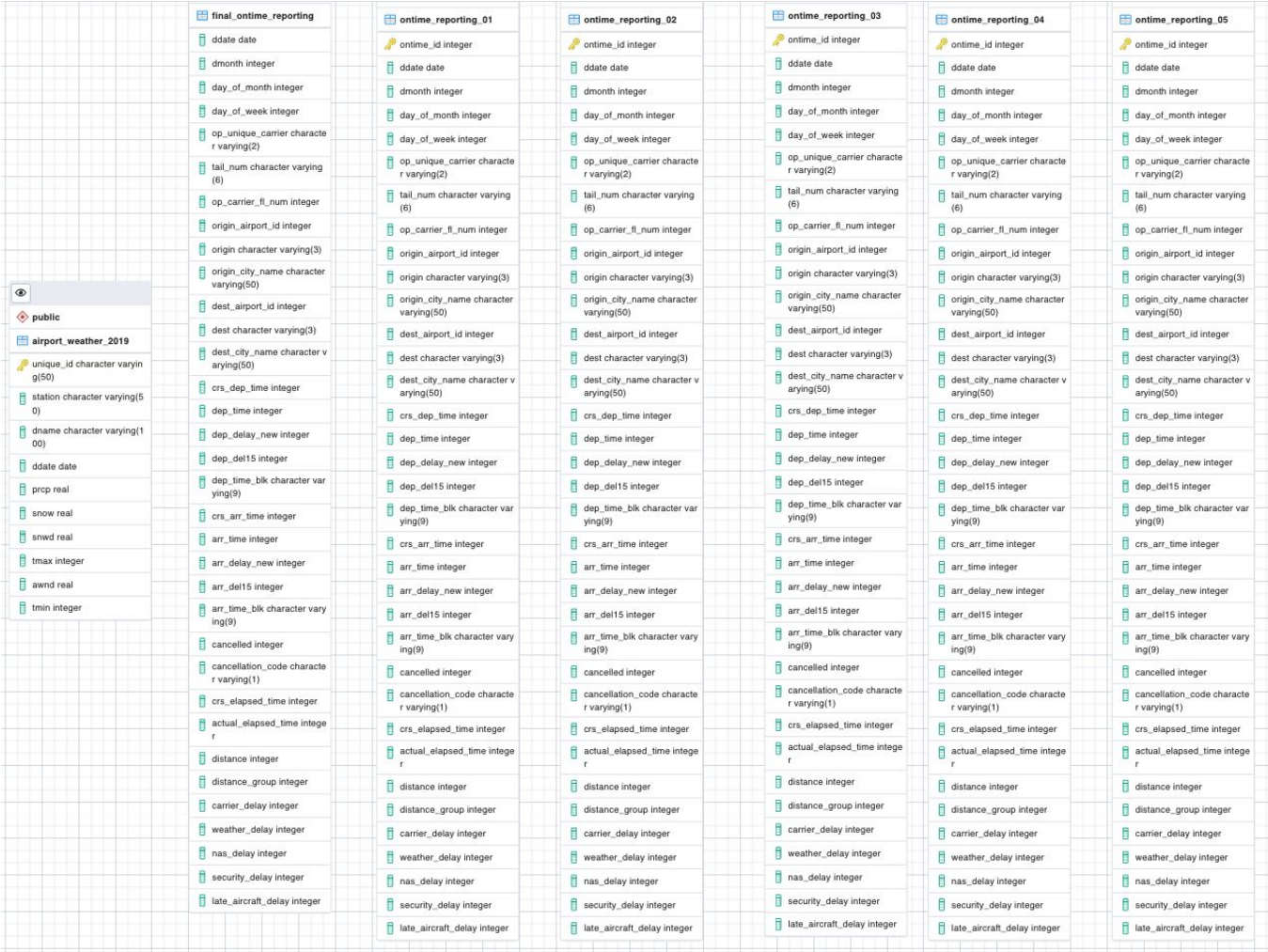
1	DATE	583985	non-null	datetime64[ns]
2	DAY_OF_WEEK	583985	non-null	int64
3	OP_UNIQUE_CARRIER	583985	non-null	object
4	TAIL_NUM	581442	non-null	object
5	OP_CARRIER_FL_NUM	583985	non-null	int64
6	ORIGIN_AIRPORT_ID	583985	non-null	int64
7	ORIGIN	583985	non-null	object
8	ORIGIN_CITY_NAME	583985	non-null	object
9	DEST_AIRPORT_ID	583985	non-null	int64
10	DEST	583985	non-null	object
11	DEST_CITY_NAME	583985	non-null	object
12	CRS_DEP_TIME	583985	non-null	int64
13	DEP_TIME	567633	non-null	float64
14	DEP_DELAY_NEW	567630	non-null	float64
15	DEP_DEL15	567630	non-null	float64
16	DEP_TIME_BLK	583985	non-null	object
17	CRS_ARR_TIME	583985	non-null	int64
18	ARR_TIME	566924	non-null	float64
19	ARR_DELAY_NEW	565963	non-null	float64
20	ARR_TIME_BLK	583985	non-null	object
21	CANCELLED	583985	non-null	float64
22	CANCELLATION_CODE	16726	non-null	object
23	CRS_ELAPSED_TIME	583851	non-null	float64
24	ACTUAL_ELAPSED_TIME	565963	non-null	float64
25	DISTANCE	583985	non-null	float64
26	DISTANCE_GROUP	583985	non-null	int64
27	CARRIER_DELAY	105222	non-null	float64
28	WEATHER_DELAY	105222	non-null	float64
29	NAS_DELAY	105222	non-null	float64
30	SECURITY_DELAY	105222	non-null	float64
31	LATE_AIRCRAFT_DELAY	105222	non-null	float64
32	AWND	451632	non-null	float64
33	PRCP	451632	non-null	float64
34	SNOW	300284	non-null	float64
35	SNWD	288421	non-null	float64
36	TAVG	375095	non-null	float64
37	TMAX	451632	non-null	float64
38	TMIN	451632	non-null	float64
39	DISPLAY_AIRPORT_NAME	451632	non-null	object

On-time-data with weather data

	Unnamed: 32	DATE	DAY_OF_WEEK	OP_UNIQUE_CARRIER	TAIL_NUM	OP_CARRIER_FL_NUM	ORIGIN_AIRPORT_ID	ORIGIN	ORIGIN_CITY_NAME	DEST_AIRPORT_ID	...
0	NaN	2019-01-01	2	WN	N739GB	2147	11259	DAL	Dallas, TX	10140	...
1	NaN	2019-01-01	2	OO	N943SW	5045	13184	MBS	Saginaw/Bay City/Midland, MI	13930	...
2	NaN	2019-01-01	2	OO	N679SA	5046	14543	RKS	Rock Springs, WY	11292	...
3	NaN	2019-01-01	2	YV	N83329	6016	10713	BOI	Boise, ID	12266	...
4	NaN	2019-01-01	2	AA	N102NN	1140	10721	BOS	Boston, MA	12478	...

Data Analysis

Database ERD



Database Output

Data output | Messages | Notifications

	unique_id character varying (50)	station character varying (50)	dname character varying (100)	ddate date	prcp real	snow real	snwd real	tmax integer	awnd real	tmin integer
1	USW0001387443471	USW00013874	ATLANTA HARTSFIELD...	2019-01-06	0	0	0	69	6.49	43
2	USW0001387443526	USW00013874	ATLANTA HARTSFIELD...	2019-03-02	0	0	0	66	7.16	51
3	USW0001387443560	USW00013874	ATLANTA HARTSFIELD...	2019-04-05	0.07	0	0	76	6.93	56
4	USW0001387443602	USW00013874	ATLANTA HARTSFIELD...	2019-05-17	0	0	0	88	5.82	66
5	USW0001387443620	USW00013874	ATLANTA HARTSFIELD...	2019-06-04	0	0	0	86	8.5	70
6	USW0001387443654	USW00013874	ATLANTA HARTSFIELD...	2019-07-08	0.09	0	0	93	6.71	74
7	USW0001387443702	USW00013874	ATLANTA HARTSFIELD...	2019-08-25	0	0	0	78	11.86	67
8	USW0001387443711	USW00013874	ATLANTA HARTSFIELD...	2019-09-03	0	0	0	95	6.26	73
9	USW0001387443774	USW00013874	ATLANTA HARTSFIELD...	2019-11-05	0	0	0	69	7.83	54
10	USW0001387443807	USW00013874	ATLANTA HARTSFIELD...	2019-12-08	0	0	0	57	11.86	47
11	USW0001387443526	USW00013874	ATLANTA HARTSFIELD...	2019-03-02	0	0	0	66	7.16	51
12	USW0001387443561	USW00013874	ATLANTA HARTSFIELD...	2019-04-06	0	0	0	73	7.83	59
13	USW0001387443602	USW00013874	ATLANTA HARTSFIELD...	2019-05-17	0	0	0	88	5.82	66
14	USW0001387443620	USW00013874	ATLANTA HARTSFIELD...	2019-06-04	0	0	0	86	8.5	70
15	USW0001387443654	USW00013874	ATLANTA HARTSFIELD...	2019-07-08	0.09	0	0	93	6.71	74
16	USW0001387443702	USW00013874	ATLANTA HARTSFIELD...	2019-08-25	0	0	0	78	11.86	67
17	USW0001387443712	USW00013874	ATLANTA HARTSFIELD...	2019-09-04	0	0	0	97	8.28	75
18	USW0001387443745	USW00013874	ATLANTA HARTSFIELD...	2019-10-07	0	0	0	81	6.93	67
19	USW0001387443807	USW00013874	ATLANTA HARTSFIELD...	2019-12-08	0	0	0	57	11.86	47
20	USW0001387443526	USW00013874	ATLANTA HARTSFIELD...	2019-03-02	0	0	0	66	7.16	51
21	USW0001387443562	USW00013874	ATLANTA HARTSFIELD...	2019-04-07	1.08	0	0	80	5.59	62
22	USW0001387443602	USW00013874	ATLANTA HARTSFIELD...	2019-05-17	0	0	0	88	5.82	66
23	USW0001387443620	USW00013874	ATLANTA HARTSFIELD...	2019-06-04	0	0	0	86	8.5	70
24	USW0001387443654	USW00013874	ATLANTA HARTSFIELD...	2019-07-08	0.09	0	0	93	6.71	74
25	USW0001387443702	USW00013874	ATLANTA HARTSFIELD...	2019-08-25	0	0	0	78	11.86	67
26	USW0001387443713	USW00013874	ATLANTA HARTSFIELD...	2019-09-05	0	0	0	95	12.75	76
27	USW0001387443746	USW00013874	ATLANTA HARTSFIELD...	2019-10-08	0	0	0	83	7.83	66
28	USW0001387443776	USW00013874	ATLANTA HARTSFIELD...	2019-11-07	0.67	0	0	67	7.61	55

Total rows: 1000 of 5692785 | Query complete 00:05:29.204

Machine Learning Model

Logistic Regression to Predict Flight Departure Delay

Logistic Regression is a statistical method for predicting binary outcomes from data.

- We will use this model to take the DEP_DEL15 dimension data and split flights it into classes.
- The DEP_DEL15 dimension data is "0" for ontime departure and "1" for a delay.
- The DEP_DEL15 dimension data is boolean confirming if the minute count in DEP_DELAY_NEW is greater than 15 min. Is is industry knowledge that fl counted as late unless the delay is > 15.
- The model may need multiple algorithms to make an accurate prediction.
- We can calculate logistic regression flight delay by adding an activation function as the final step to our linear model.
- This converts the linear regression output to a probability.

Questions

- What type of Logistical regression algorithms will offer the best prediction of flight departure delay?
- Can we make a prediction of planes that could need service based on CARRIER_DELAY and unique values in TAIL_NUM?

```
import matplotlib.pyplot as plt
import pandas as pd
from pathlib import Path
import matplotlib.pyplot as plt
```

Sample data from data source

```
df = pd.read_csv(Path('/Users/traverslaville/Desktop/Classwork/Final_Project_Group_3/Resources/ONTIME_REPORTING_2020_03.csv'))
df.head()
```