Analyzing the predictability of flight cancellations for Clients' flight booking app

STEPS

DATASET IMPORT » LIBRARIES » DATA EXPLORATION » VISUALIZATION » ANALYSIS » ALGORITHM DATA » ALGORITHMS » INSIGHT

OUTCOME-DRIVEN + **SIMPLE** + **ACTIONABLE**

OUTCOME-DRIVEN

- Clients' needs are taken into consideration in each and every step

SIMPLE

- One notebook to browse various data angles

ACTIONABLE

- Clear insight ready to support Clients' decisions

DATA EXPLORATION

Since our dataset has more air-related categories than is recquired, we imported only relevant ones to ensure project clarity.

The most important metric is separately counted.

```
flights = pd.read_csv('flights.csv', usecols = ['MONTH','DAY','DAY_OF_WEEK','AIRLINE','ORIGIN_AIRPORT','DESTINATION_AIRPORT','DISTANCE','SCHEDULED_ARRI
flights
```

```
/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2882: DtypeWarning: Columns (7,8) have mixed types.Specify dtype option on impo
rt or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

	MONTH	DAY	DAY_OF_WEEK	AIRLINE	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DISTANCE	SCHEDULED_ARRIVAL	CANCELLED	CANCELLATION_REASON
0	1	1	4	AS	ANC	SEA	1448	430	0	NaN
1	1	1	4	AA	LAX	PBI	2330	750	0	NaN
2	1	1	4	US	SFO	CLT	2296	806	0	NaN
3	1	1	4	AA	LAX	MIA	2342	805	0	NaN
4	1	1	4	AS	SEA	ANC	1448	320	0	NaN
...	...	...	...	...	...	...	...	...	...	...
5819074	12	31	4	B6	LAX	BOS	2611	819	0	NaN
5819075	12	31	4	B6	JFK	PSE	1617	446	0	NaN
5819076	12	31	4	B6	JFK	SJU	1598	440	0	NaN
5819077	12	31	4	B6	MCO	SJU	1189	340	0	NaN
5819078	12	31	4	B6	JFK	BQN	1576	440	0	NaN

5819079 rows × 10 columns

```
flights['CANCELLED'].value_counts()
```

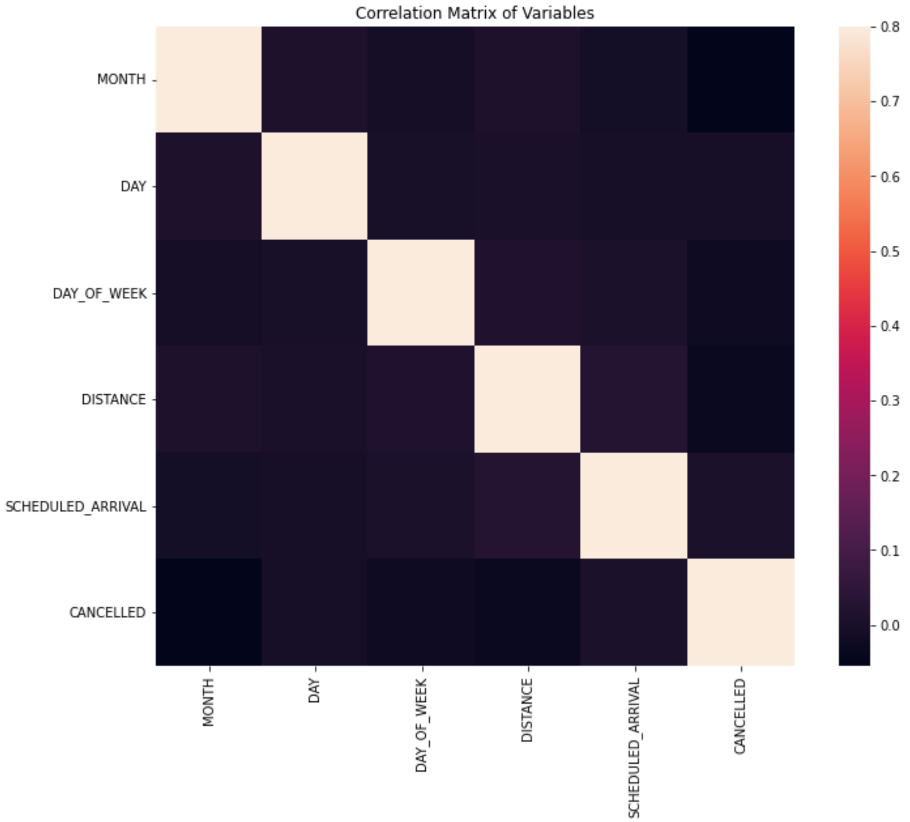
```
0    5729195
1      89884
Name: CANCELLED, dtype: int64
```

Cancelled flights are only 2% of all flights. While the number is miniscule, implementing cancel checks into the Clients app would be beneficial for user experience, if they lead to good predictability.

DATA EXPLORATION

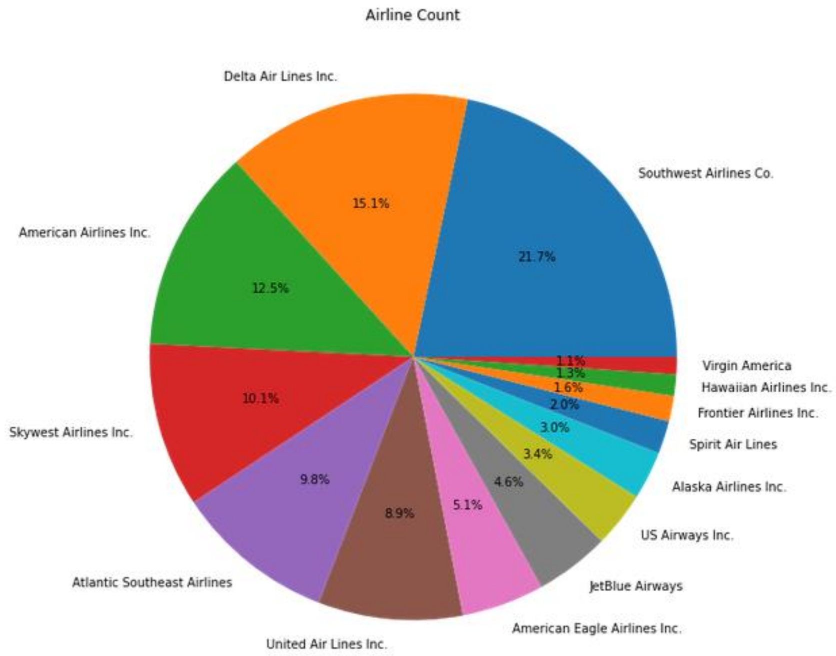
Correlation matrix shows no correlations, which does not bode well for our project.

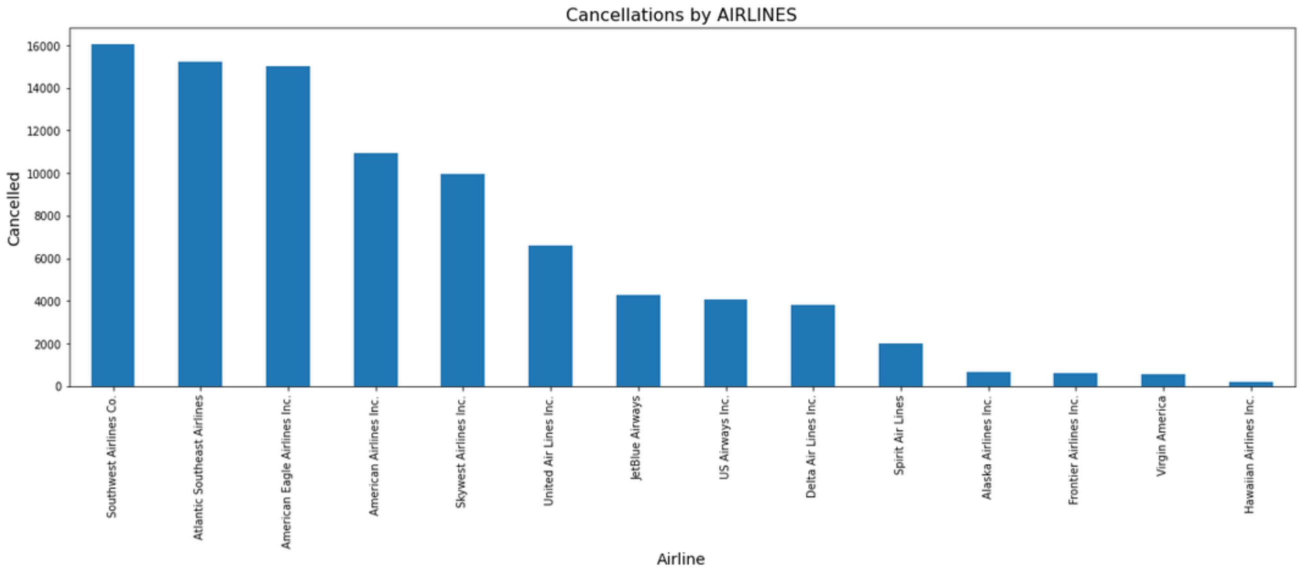
```python
corrmat = flights.corr()
f, ax = plt.subplots(figsize=(12, 9))
sns.heatmap(corrmat, vmax=.8, square=True);
plt.title("Correlation Matrix of Variables")
plt.show()
```



VISUALIZATION AND ANALYSIS

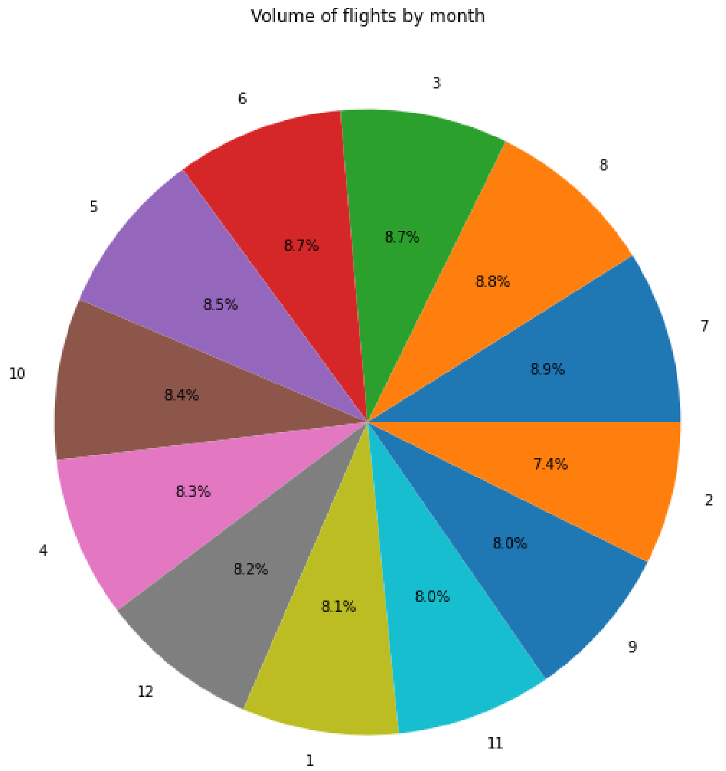
Considering total flights by airline, it is no surprise that No #1 in amount of flights is also the No #1 in cancellations. However, the ratio does not continue. Atlantic Southeast Airlines is second in cancellations amount while contributing to only 10% of all airline traffic. Surprisingly, Delta Air Lines has a remarkably low amount of cancellations compared to their flights volume.

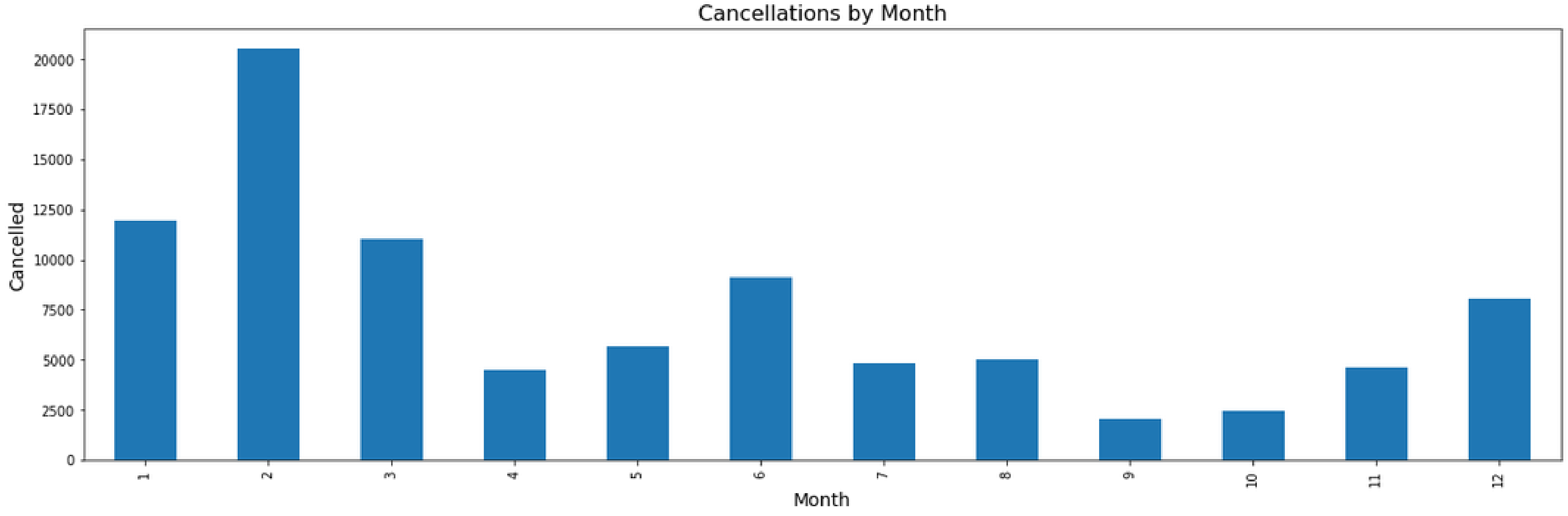




VISUALIZATION AND ANALYSIS

Significant increase in cancellations in February while the total amount of flights is the years' lowest. According to historical data, at that time USA was experiencing a polar vortex which is a plausible cause for beforementioned trends in data.

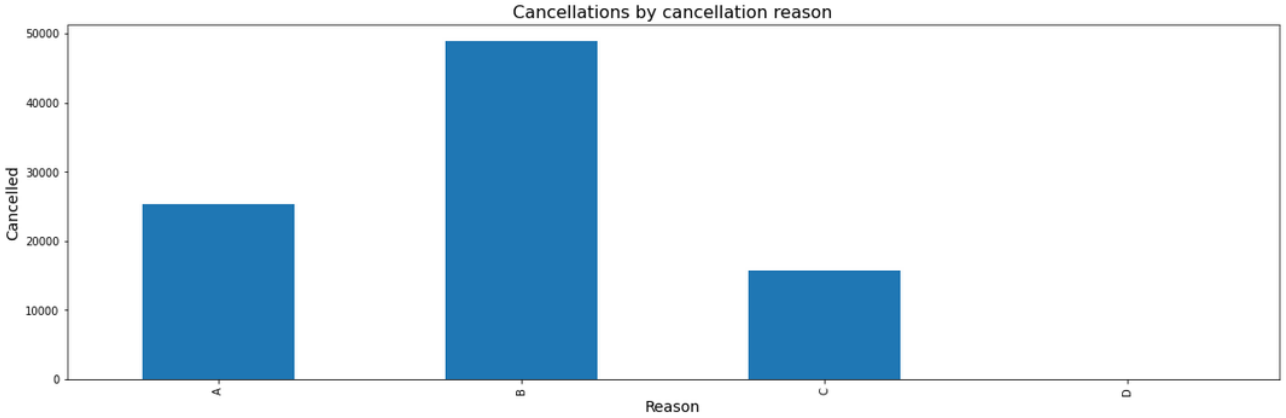




VISUALIZATION AND ANALYSIS

Most cancellations were due to weather, which is somewhat predictable. Carrier and NAS reasons are impossible to predict, and Security reasons can be omitted.

In conclusion, the data is too scarce to create any correlations or predictions - just 2% of a dataset is not enough to form any remarks. Predicting the weather seems to have the biggest impact on flight cancellations, however that cannot be done with given dataset about flights.



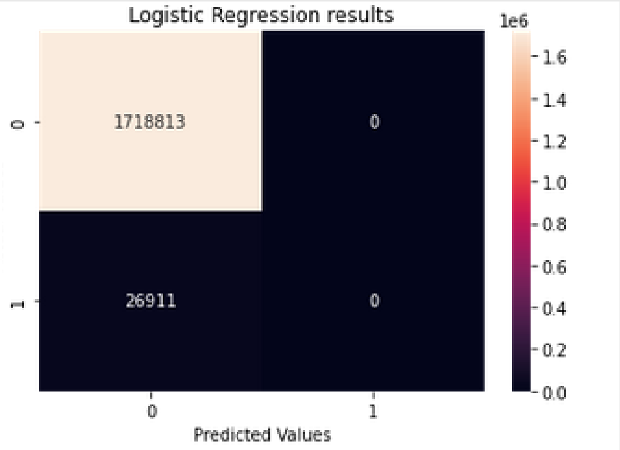

```
A - Airline/Carrier
B - Weather
C - National Air System
D - Security
```

ALGORITHMS

Logistic regression has better scores, and no False Positives, therefore it is a slightly better model than Bayes one.

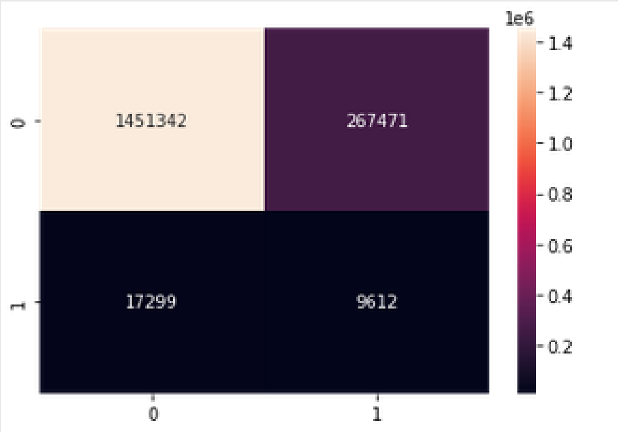
Logistic Regression

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1718813
1	0.00	0.00	0.00	26911
accuracy			0.98	1745724
macro avg	0.49	0.50	0.50	1745724
weighted avg	0.97	0.98	0.98	1745724



Naive Bayes

	precision	recall	f1-score	support
0	0.99	0.84	0.91	1718813
1	0.03	0.36	0.06	26911
accuracy			0.84	1745724
macro avg	0.51	0.60	0.49	1745724
weighted avg	0.97	0.84	0.90	1745724



INSIGHT

Looking at the data objectively, presented dataset does not allow for precise prediction of cancelled flights. On top of being rare, consisting of only 2% of all flights, cancellations rely heavily on unfavorable weather conditions.

Further modelling would be possible with the inclusion of weather data, however from the business side it is more cost efficient to just connect to a real-time weather provider through API and display in-app warnings about extreme weather situations in users' specified locations based on their travel plans.

THANK YOU