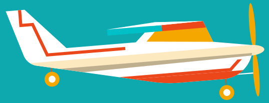
Predicting Airline Delay

By: Naufal Fadhilah

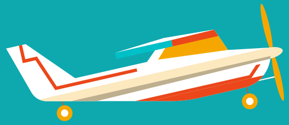


01
Data Understanding

02
Exploratory Data Analysis & Data Pre processing

03
Machine Learning Model

04
Conclusion and Recommendation

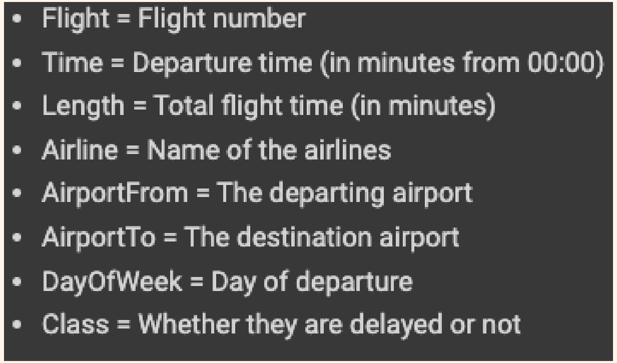


01

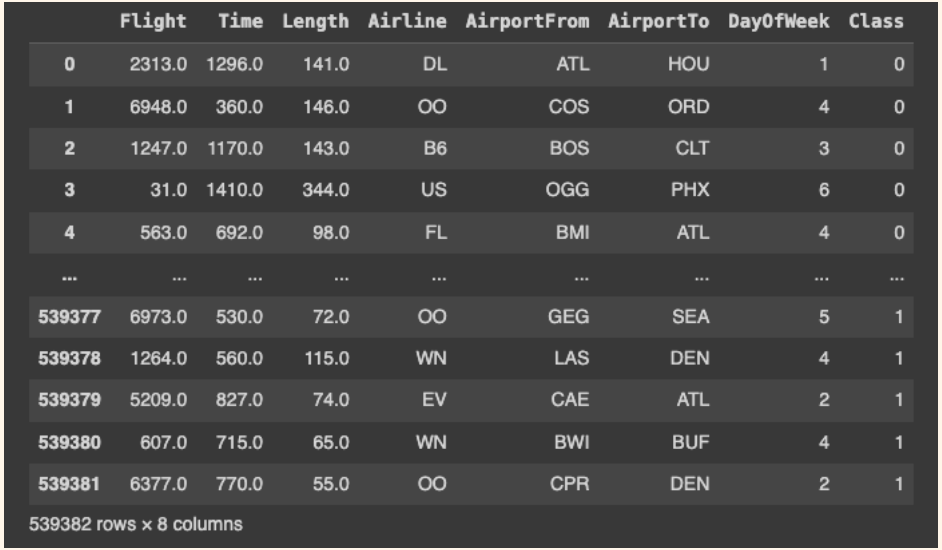
Data Understanding

Data Understanding

Obtained through Kaggle, this dataset is about predicting whether the airline will be delayed or not. Where it contains 539.382 data with 8 columns.

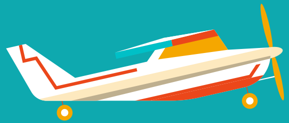


- Flight = Flight number
- Time = Departure time (in minutes from 00:00)
- Length = Total flight time (in minutes)
- Airline = Name of the airlines
- AirportFrom = The departing airport
- AirportTo = The destination airport
- DayOfWeek = Day of departure
- Class = Whether they are delayed or not



	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class
0	2313.0	1296.0	141.0	DL	ATL	HOU	1	0
1	6948.0	360.0	146.0	OO	COS	ORD	4	0
2	1247.0	1170.0	143.0	B6	BOS	CLT	3	0
3	31.0	1410.0	344.0	US	OGG	PHX	6	0
4	563.0	692.0	98.0	FL	BMI	ATL	4	0
...	...	...	...	...	...	...	...	...
539377	6973.0	530.0	72.0	OO	GEG	SEA	5	1
539378	1264.0	560.0	115.0	WN	LAS	DEN	4	1
539379	5209.0	827.0	74.0	EV	CAE	ATL	2	1
539380	607.0	715.0	65.0	WN	BWI	BUF	4	1
539381	6377.0	770.0	55.0	OO	CPR	DEN	2	1

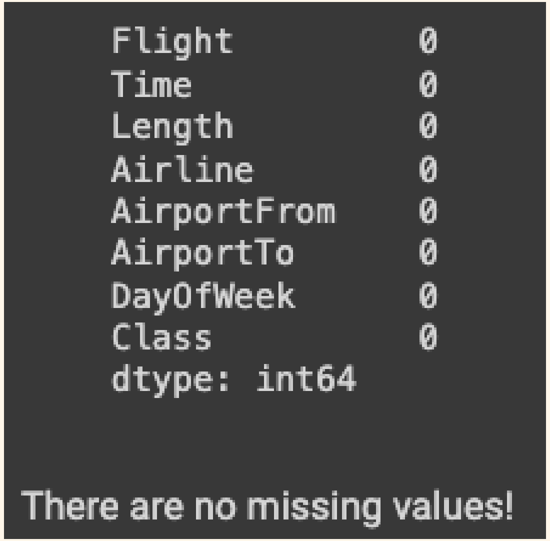
539382 rows × 8 columns



02

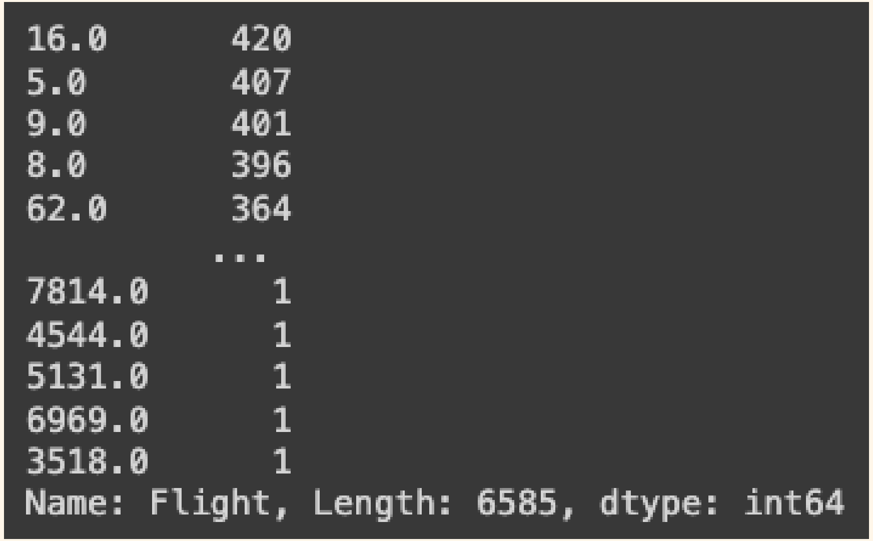
Exploratory Data Analysis & Data Pre processing

Exploratory Data Analysis & Data Preprocessing



```
Flight         0
Time           0
Length         0
Airline        0
AirportFrom    0
AirportTo      0
DayOfWeek      0
Class          0
dtype: int64
```

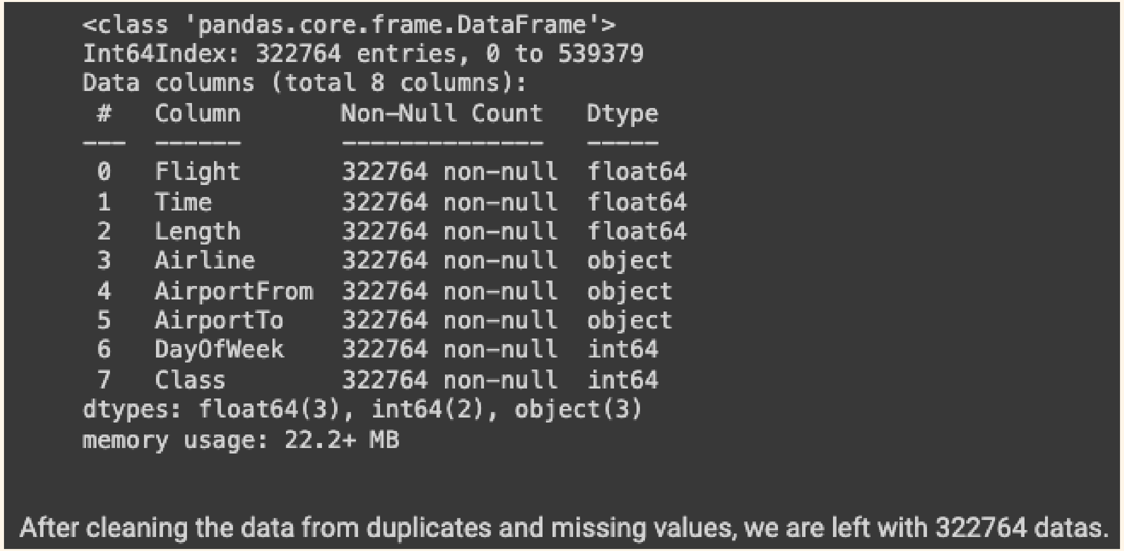
There are no missing values!



```
16.0      420
5.0       407
9.0       401
8.0       396
62.0      364
         ...
7814.0      1
4544.0      1
5131.0      1
6969.0      1
3518.0      1
Name: Flight, Length: 6585, dtype: int64
```

It is found that there are 216618 duplicated data, they all are coming from the flight column.

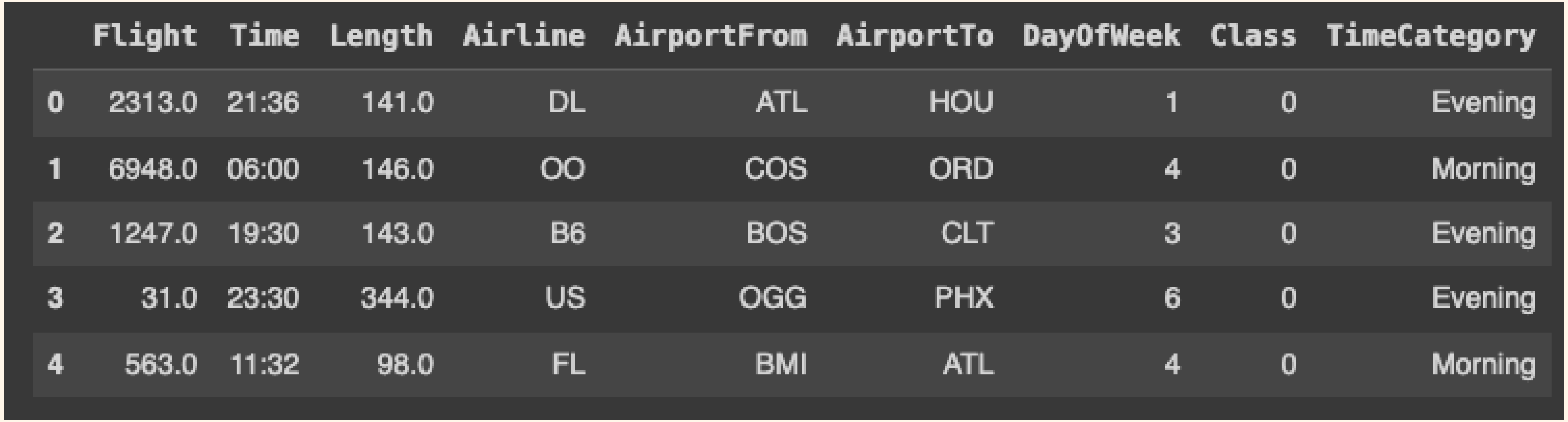
Exploratory Data Analysis & Data Preprocessing



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 322764 entries, 0 to 539379
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Flight       322764 non-null  float64
 1   Time         322764 non-null  float64
 2   Length       322764 non-null  float64
 3   Airline      322764 non-null  object
 4   AirportFrom  322764 non-null  object
 5   AirportTo    322764 non-null  object
 6   DayOfWeek    322764 non-null  int64
 7   Class        322764 non-null  int64
dtypes: float64(3), int64(2), object(3)
memory usage: 22.2+ MB
```

After cleaning the data from duplicates and missing values, we are left with 322764 datas.

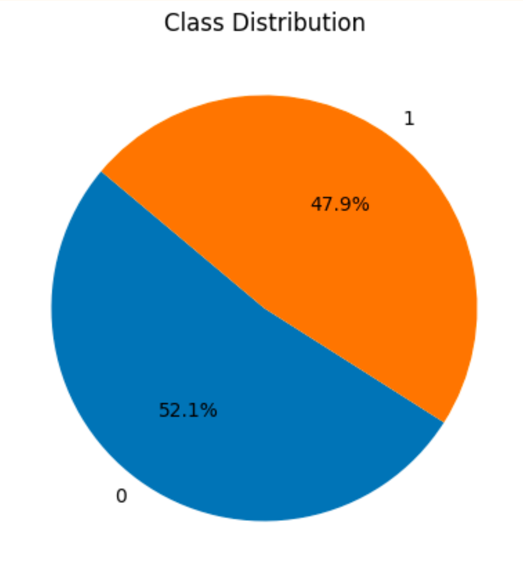
Exploratory Data Analysis & Data Preprocessing

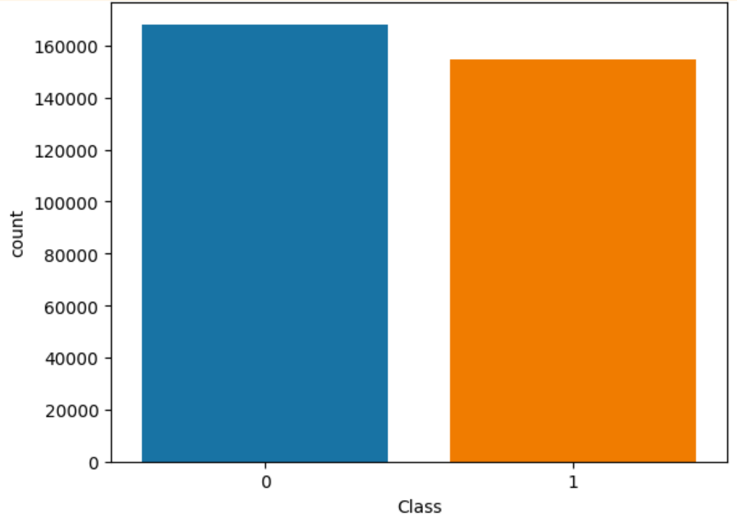


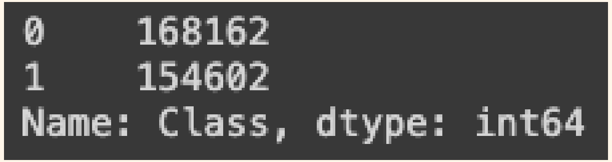
	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class	TimeCategory
0	2313.0	21:36	141.0	DL	ATL	HOU	1	0	Evening
1	6948.0	06:00	146.0	OO	COS	ORD	4	0	Morning
2	1247.0	19:30	143.0	B6	BOS	CLT	3	0	Evening
3	31.0	23:30	344.0	US	OGG	PHX	6	0	Evening
4	563.0	11:32	98.0	FL	BMI	ATL	4	0	Morning

Here Time format from minutes have been turned into HH:MM format. Also a new column called "TimeCategory" has been added.

Exploratory Data Analysis & Data Preprocessing

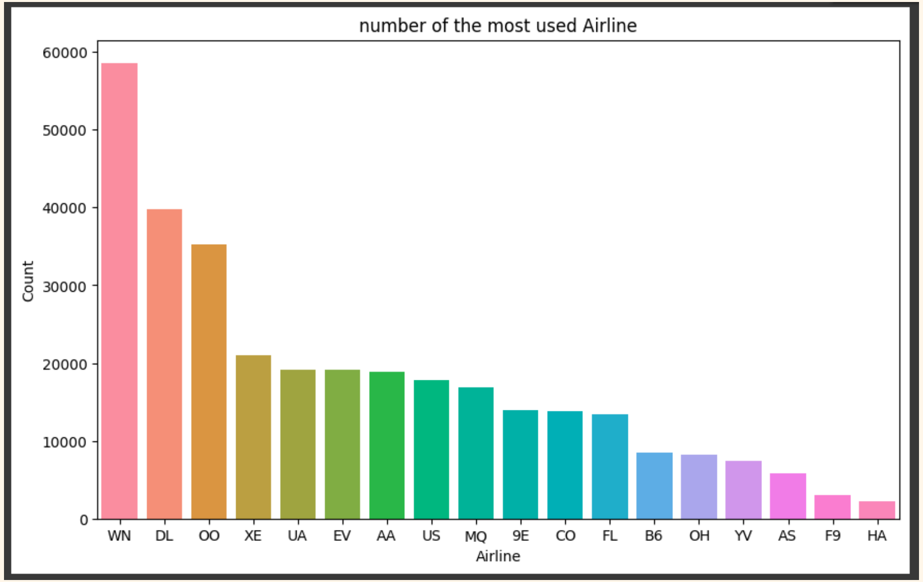


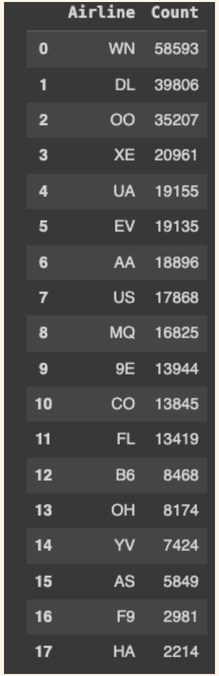




```
0    168162
1    154602
Name: Class, dtype: int64
```

Exploratory Data Analysis & Data Preprocessing

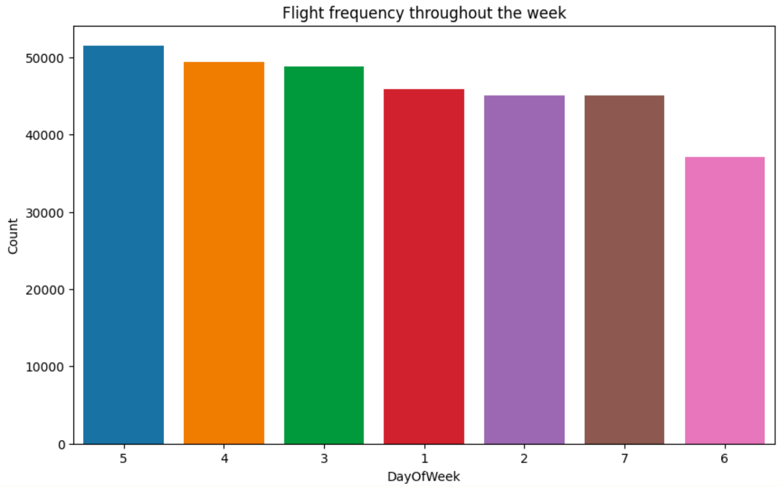


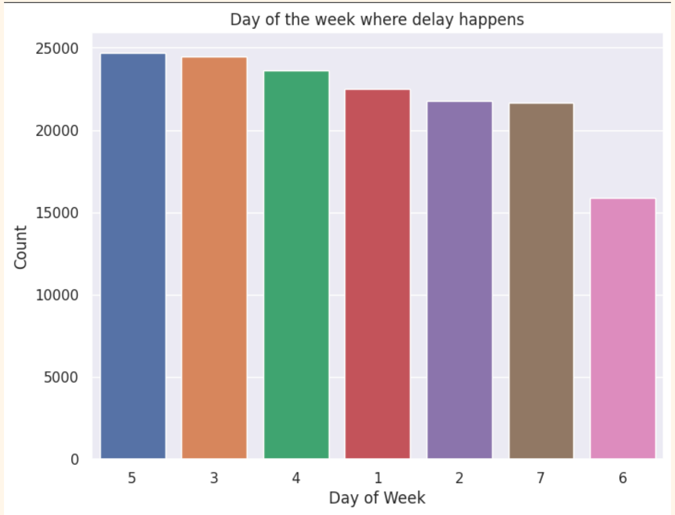


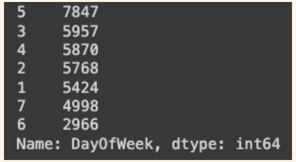
	Airline	Count
0	WN	58593
1	DL	39806
2	OO	35207
3	XE	20961
4	UA	19155
5	EV	19135
6	AA	18896
7	US	17868
8	MQ	16825
9	9E	13944
10	CO	13845
11	FL	13419
12	B6	8468
13	OH	8174
14	YV	7424
15	AS	5849
16	F9	2981
17	HA	2214

The graph shows the distribution of most used airline, WN being the most and HA being the least.

Exploratory Data Analysis & Data Preprocessing





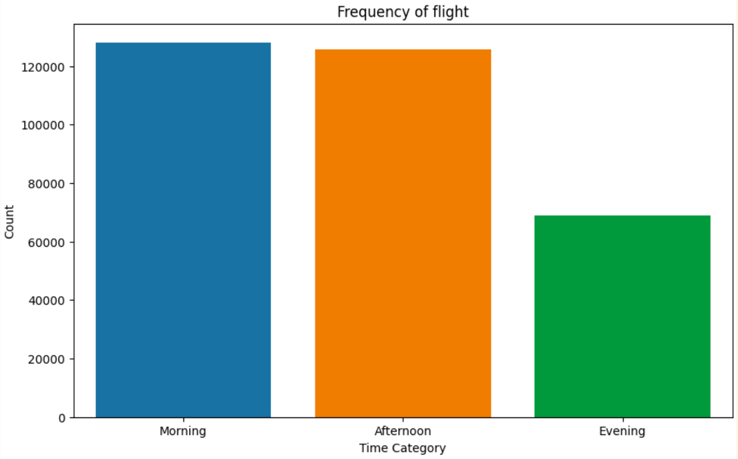


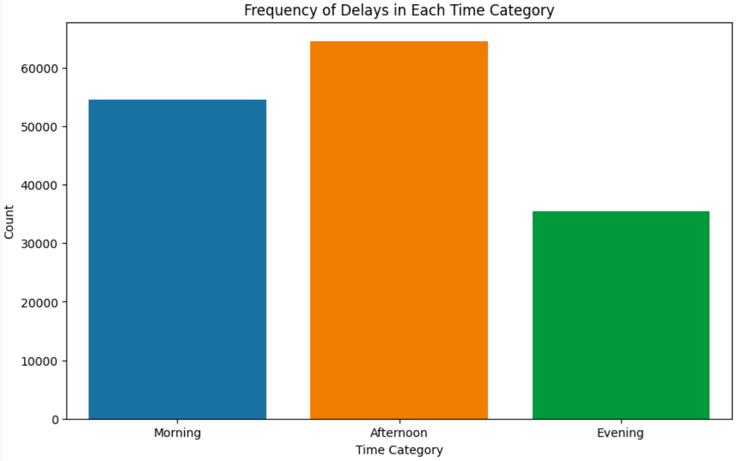
```
5    7847
3    5957
4    5870
2    5768
1    5424
7    4998
6    2966
Name: DayOfWeek, dtype: int64
```

- Flight happens mostly during day 5 and least at day 3

- Delay also happens the most at day 5

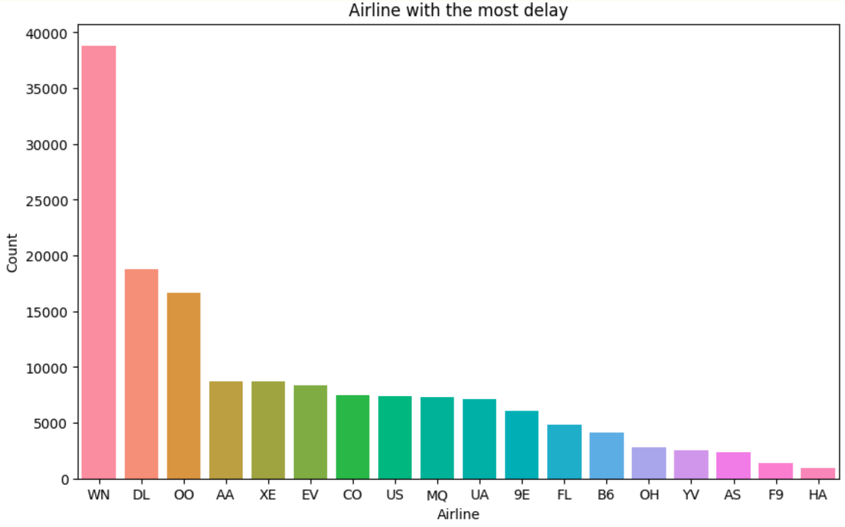
Exploratory Data Analysis & Data Preprocessing





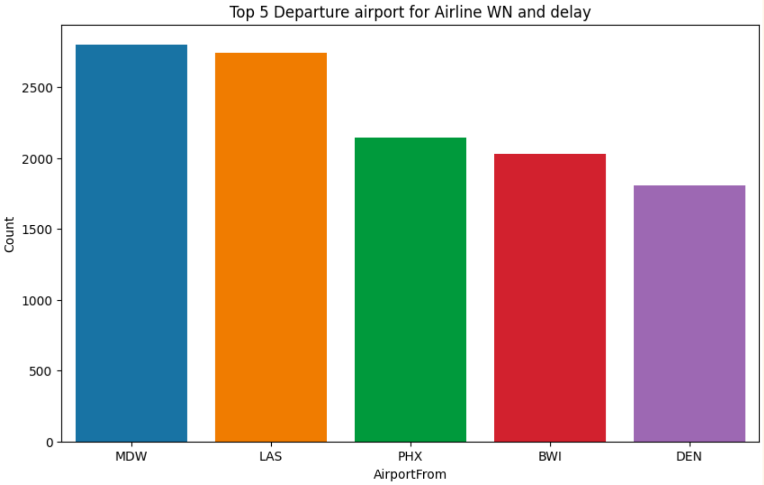
Most flights happens during morning time. However, the most delayed are in the afternoon.

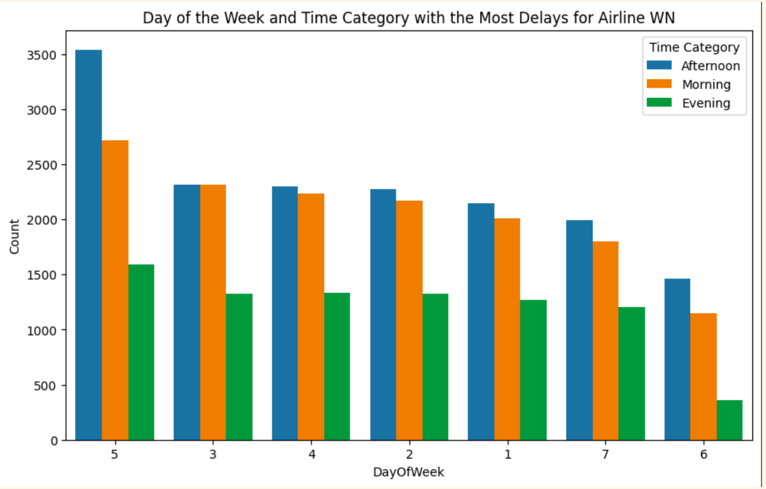
Exploratory Data Analysis & Data Preprocessing



WN is the airline with the most delay, coming in second is DL and the least delay airline being HA.

Exploratory Data Analysis & Data Preprocessing

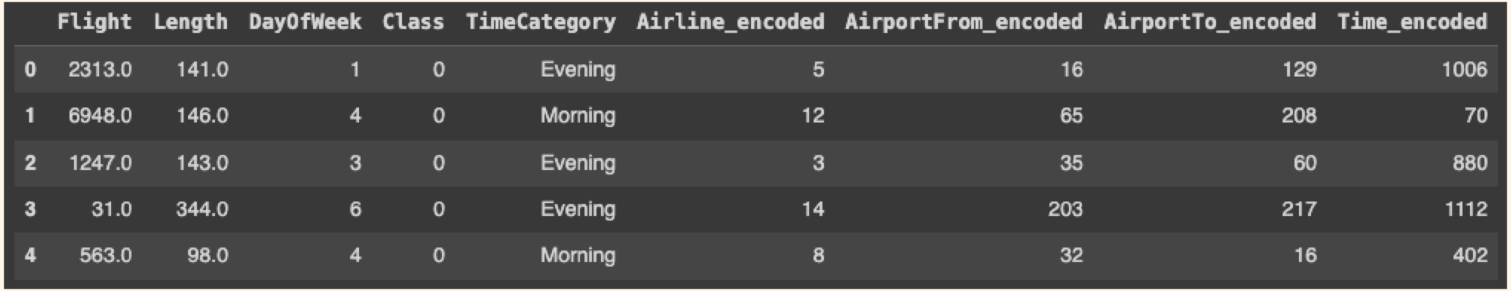




-WN mostly delayed from MDW airport

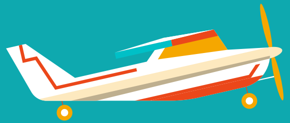
- Delay occurs mainly on day 5 in the Afternoon while the least being in the evening.

Exploratory Data Analysis & Data Preprocessing



	Flight	Length	DayOfWeek	Class	TimeCategory	Airline_encoded	AirportFrom_encoded	AirportTo_encoded	Time_encoded
0	2313.0	141.0	1	0	Evening	5	16	129	1006
1	6948.0	146.0	4	0	Morning	12	65	208	70
2	1247.0	143.0	3	0	Evening	3	35	60	880
3	31.0	344.0	6	0	Evening	14	203	217	1112
4	563.0	98.0	4	0	Morning	8	32	16	402

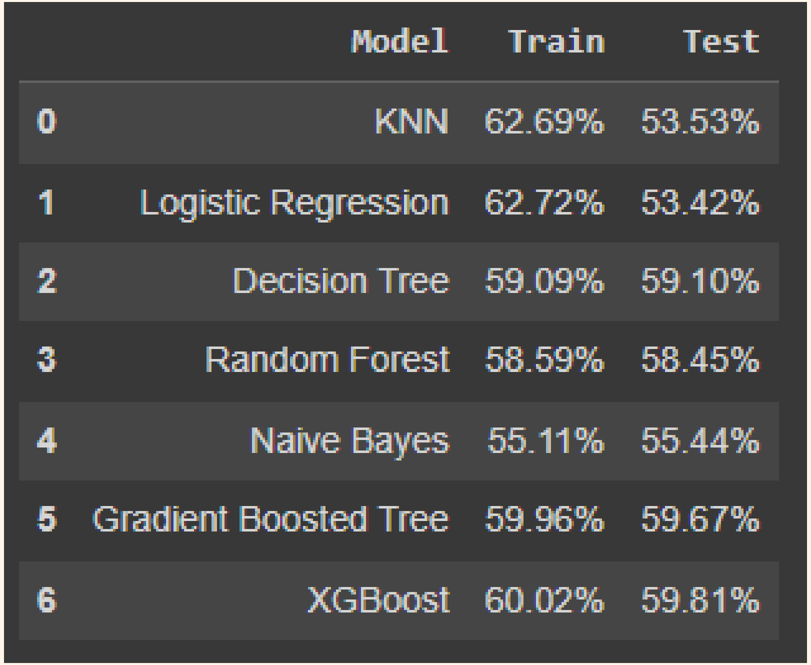
Encoding the classification column (Airline, AirportFrom, AirportTo and Time.



03

Machine Learning Model

Machine Learning Model

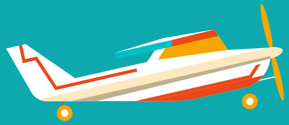


	Model	Train	Test
0	KNN	62.69%	53.53%
1	Logistic Regression	62.72%	53.42%
2	Decision Tree	59.09%	59.10%
3	Random Forest	58.59%	58.45%
4	Naive Bayes	55.11%	55.44%
5	Gradient Boosted Tree	59.96%	59.67%
6	XGBoost	60.02%	59.81%

- KNN and Logistic Regression seems toi be overfitting

- Gradient Boosted Tree has the second highest accuracy with 59.81%

- The best model to be used is XGBoost with accuracy of 59.81%



04

Conclusion and Recommendaiton

Conclusion and Recommendation

With this dataset, I ran through numerous different models, thus the model with the highest accuracy score is XGBoost with 59.81% to predict delay. While the lowest being Logistic Regression with an accuracy of 53.42%. However the recall for XGBoost is only 44.62% which is quite low. Thius suggest the model struggles to correctly identify the positive cases. For example in 100 delay it will only identify 39 delays.

Recommendation

- WN needs to add its flights on the evening more

- It is recommended for WN airline to increase its fleet

- Airline need to prepare and compensate the delayed passenger in advance to keep customer satisfaction