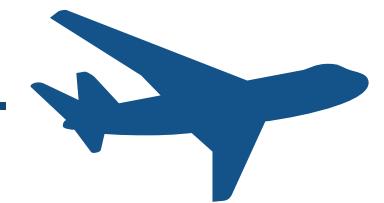
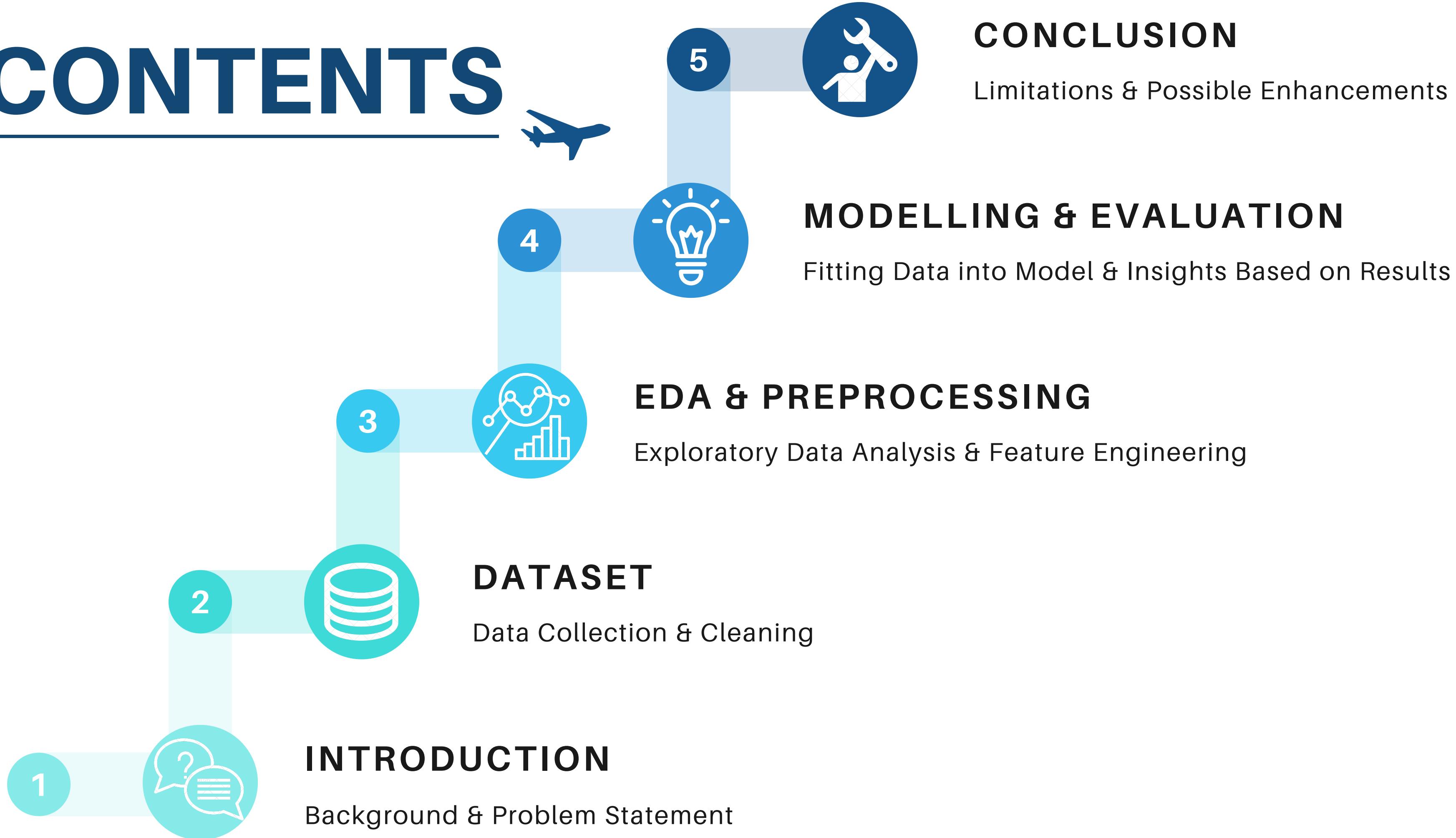


Flight Delay Prediction



LOOI JIA WEN | DSIF 3

CONTENTS



BACKGROUND



- Flight punctuality determines an airline's reliability & quality of service
- Airlines must constantly improve their flight punctuality to stay competitive
- Flight delay causes financial losses & decreases customers satisfaction



Average Cost of Flight Delay:

USD\$7,420

per minute



PROBLEM STATEMENT



- Accurately **predict** a flight delay
 - Airlines can prepare for delay contingency plans to cushion the loss
- Identify **features** that cause flights to delay (feature importance)
 - Plan better connecting flights and routes

DEFINING FLIGHT DELAY

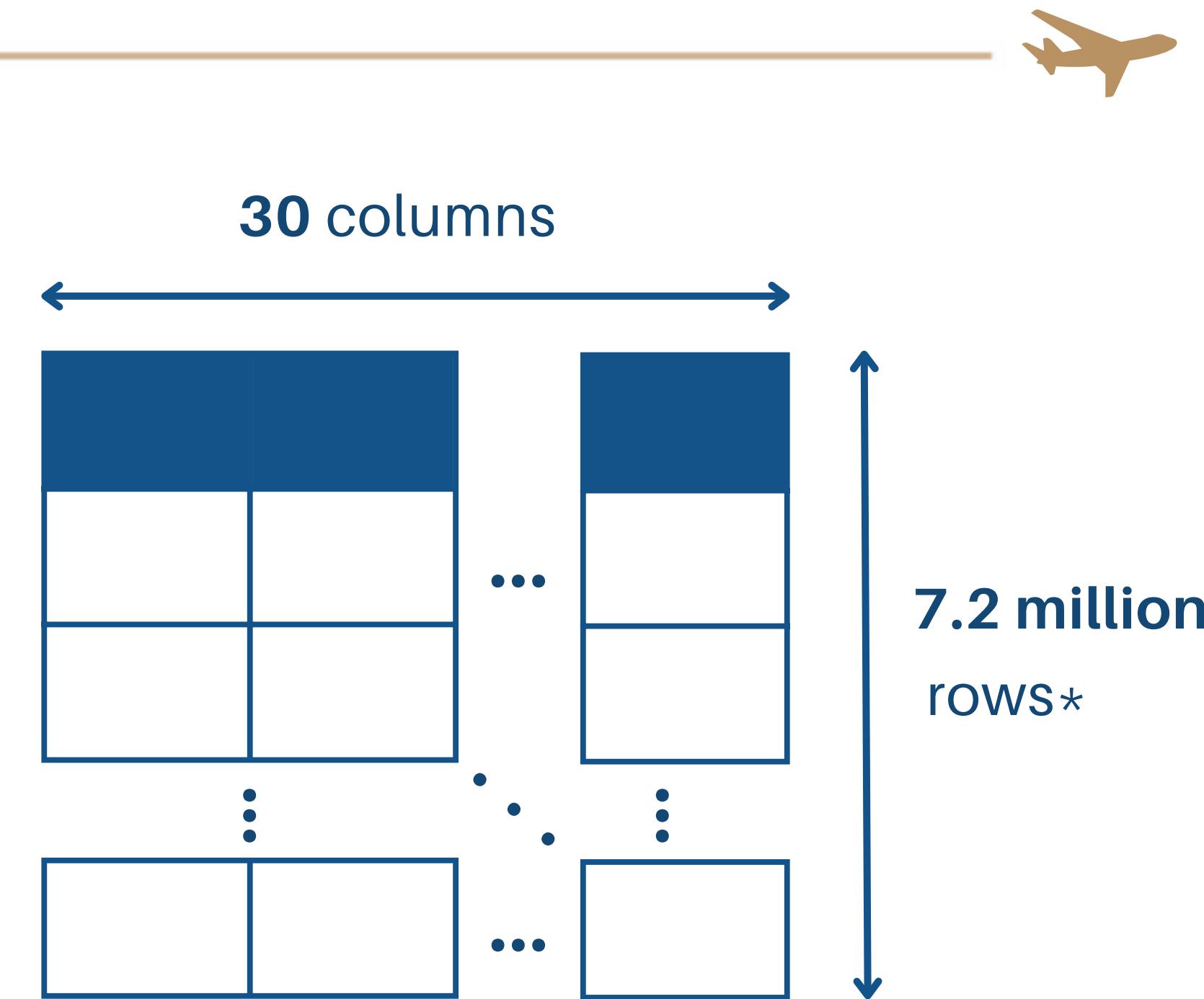


Flight takes off/lands **15 minutes**
later than its scheduled time

- *Federal Aviation Administration*

DATASET

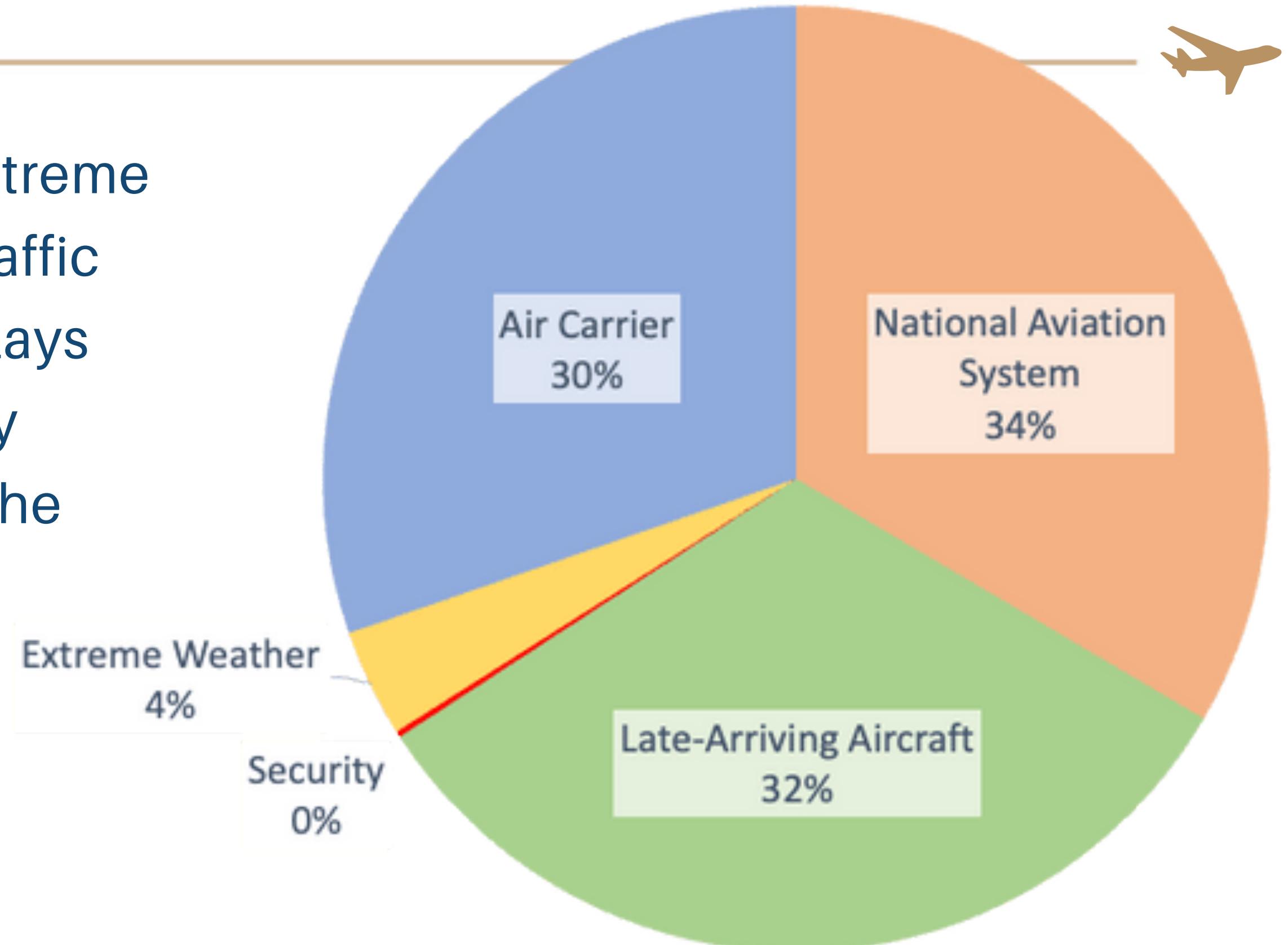
- Bureau of Transportation Statistics
(United States Department of Transportation)
- Flight Data from **2019**
(Pre-covid, peak airtravel)
- Features
 - Scheduled & Actual Flight Time
 - Origin & Destination
 - Airlines & Aircraft
 - Delay Types



*Cancelled & Diverted flights are dropped

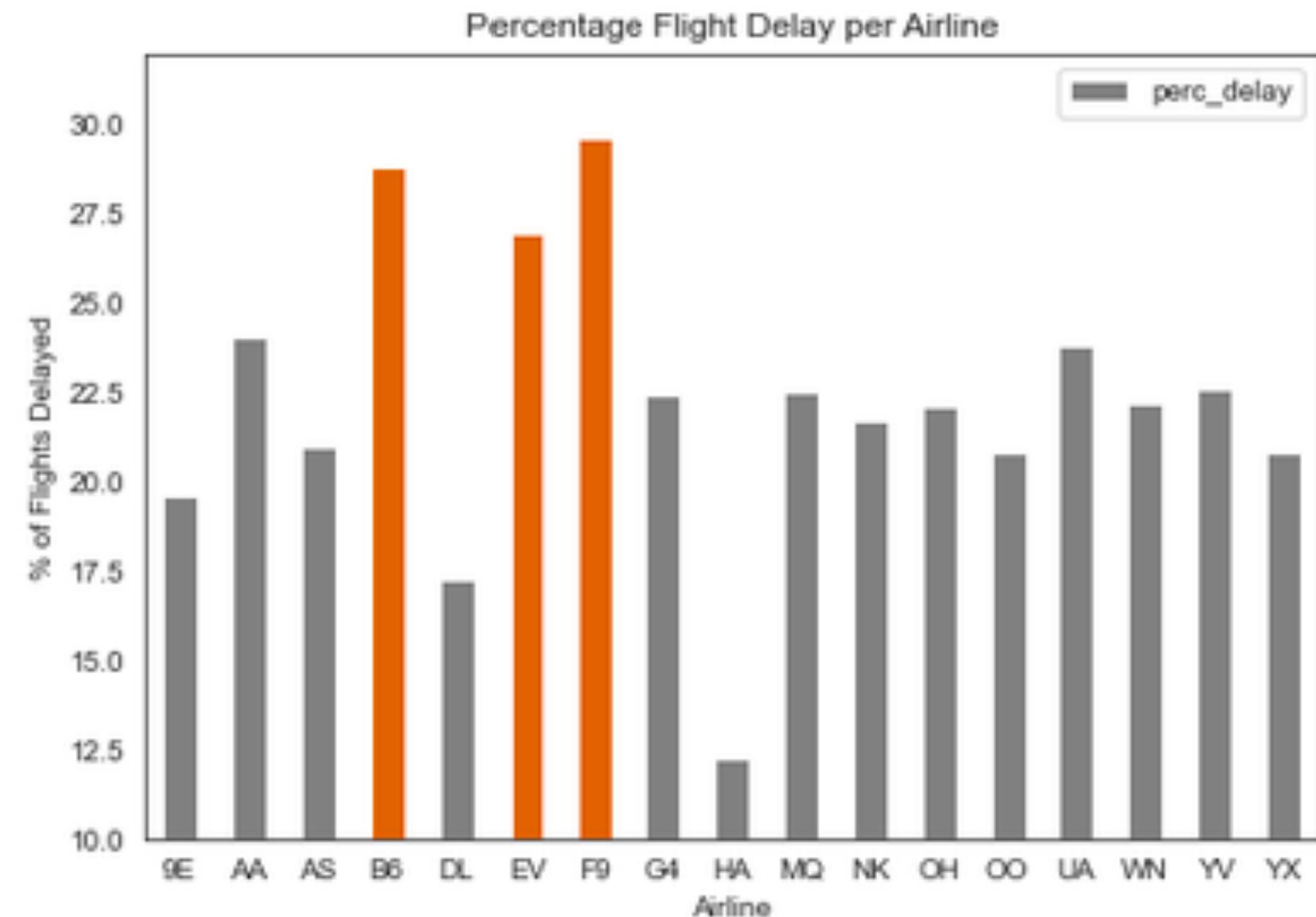
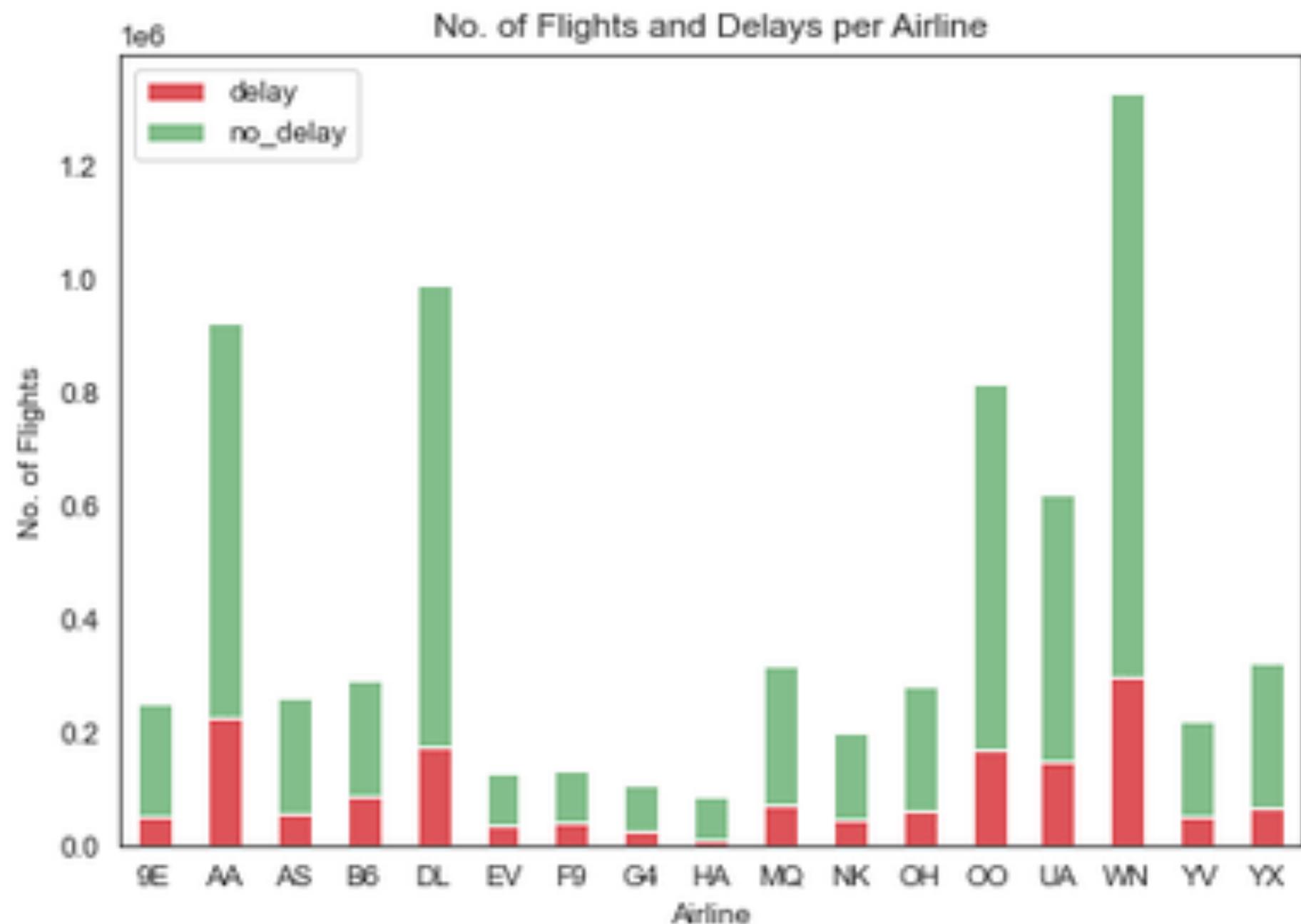
TYPES OF DELAY

1. **NAS:** Delays due to non-extreme weather conditions & air traffic
2. **Late-Arriving Aircraft:** Delays due to previous flight delay
3. **Air Carrier:** Delays within the airline's control



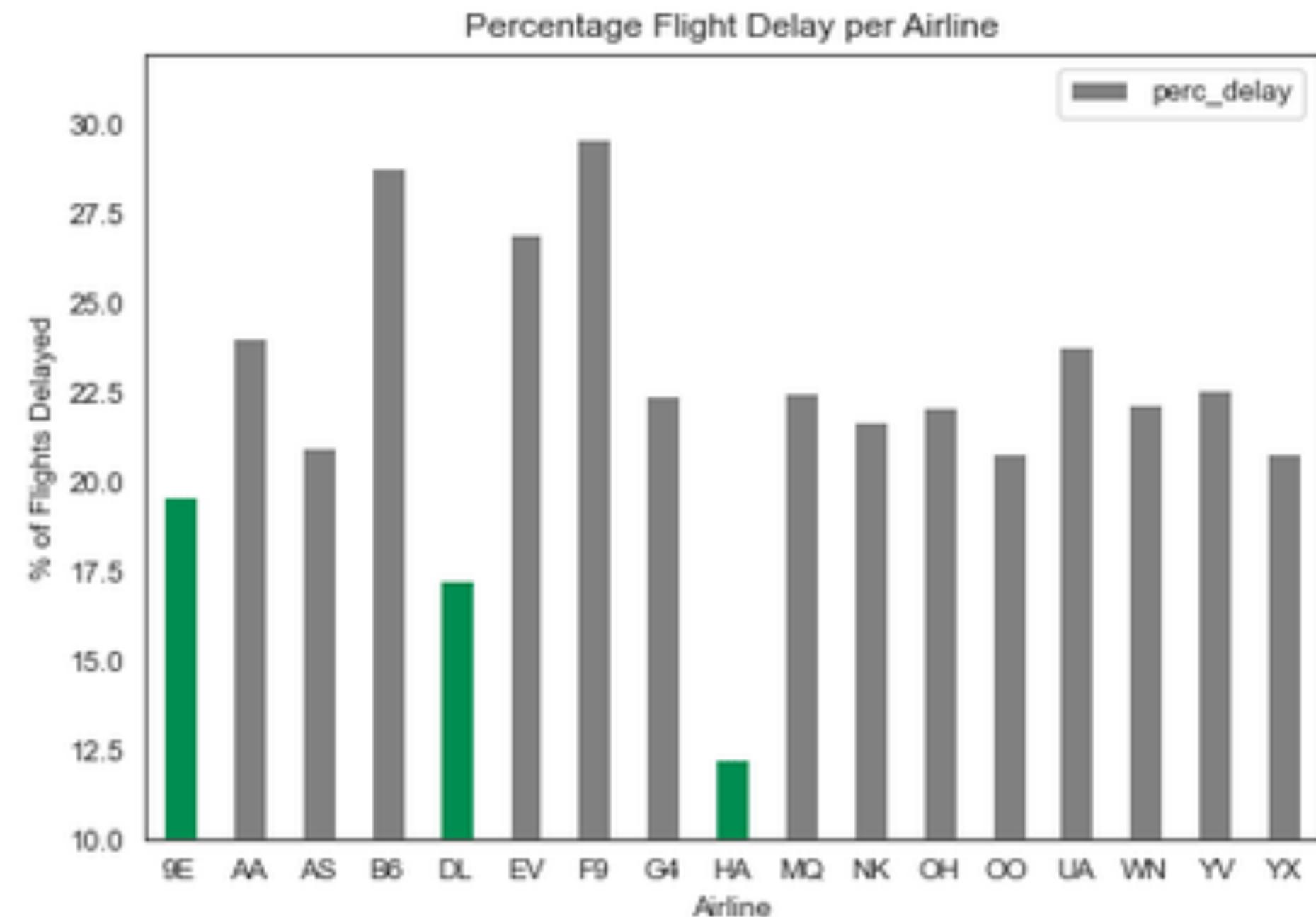
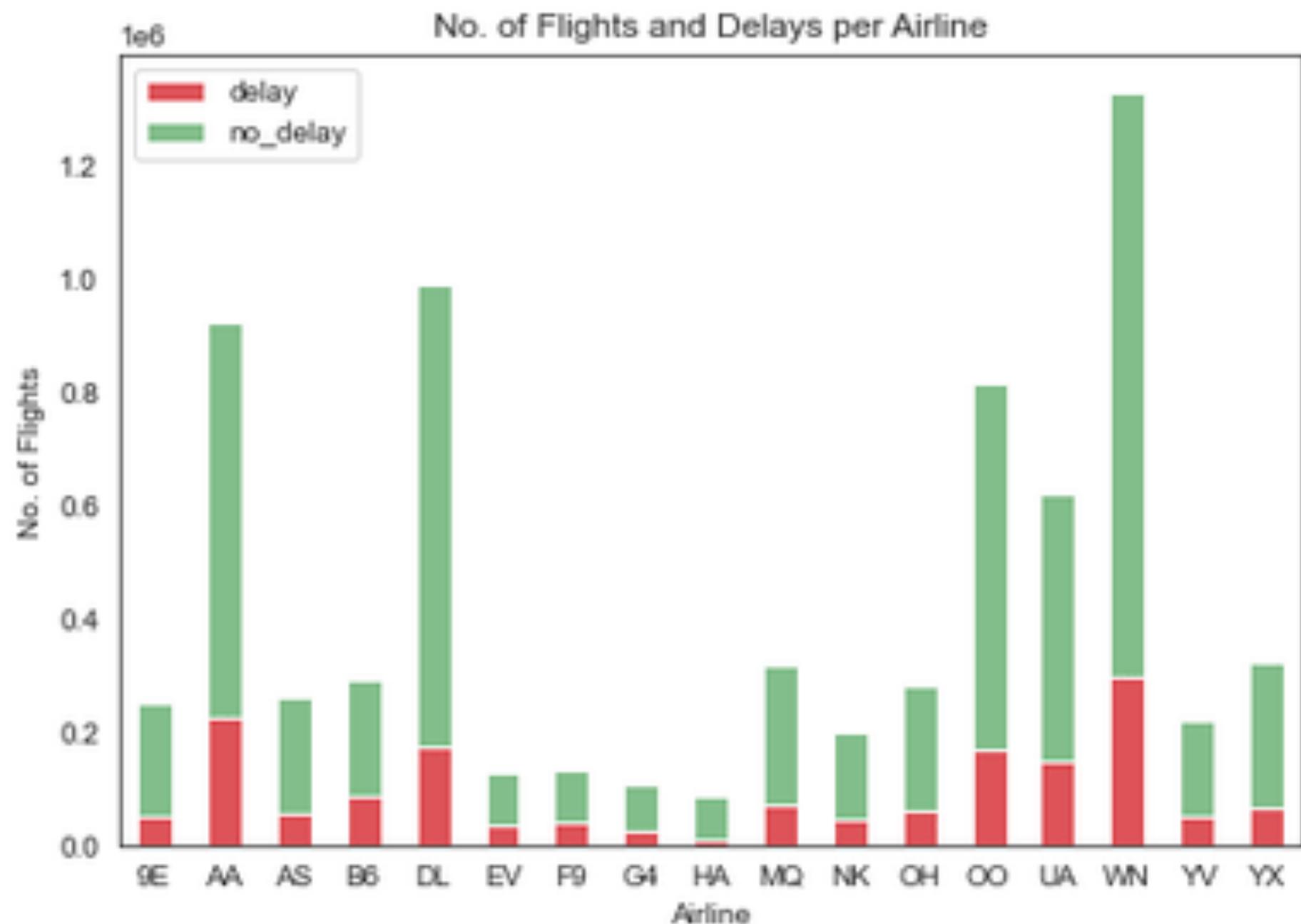
AIRLINES WITH THE HIGHEST FREQUENCY OF DELAYS

1. JetBlue (B6) - 29%
2. ExpressJet (EV) - 27%
3. Frontier (F9) - 30%

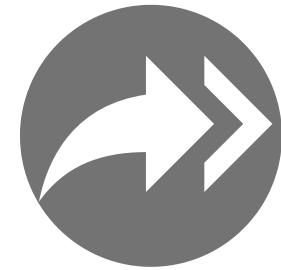
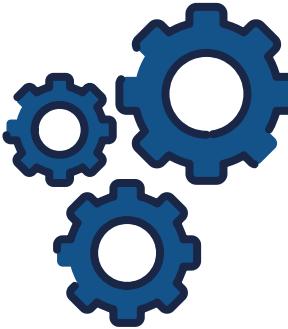


AIRLINES WITH THE LOWEST FREQUENCY OF DELAYS

1. Hawaiian (HA) - 12%
2. Delta (DL) - 17%
3. Endeavor (9E) - 19%



FEATURE ENGINEERING



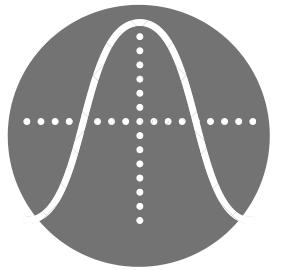
1 - Log Transform



2 - Sector per Day &
Previous Sector Delay



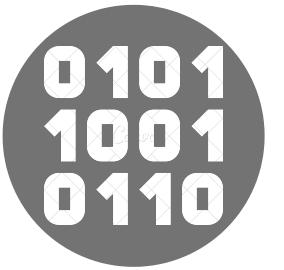
3 - Cyclical
Features Encoding



4 - Statistical Test



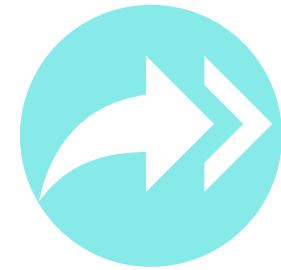
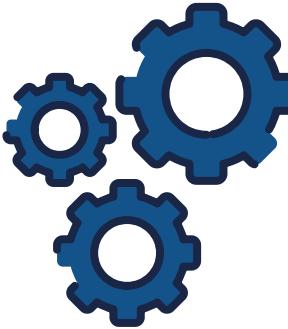
5 - Polynomial
Features



6 - One-Hot
Encoding

Dummifying our
categorical features

FEATURE ENGINEERING



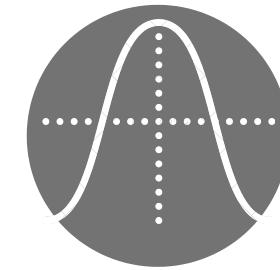
1 - Log Transform



2 - Sector per Day &
Previous Sector Delay



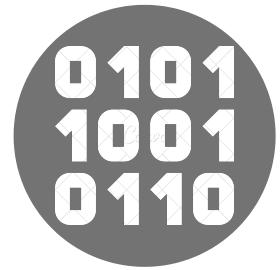
3 - Cyclical
Features Encoding



4 - Statistical Test

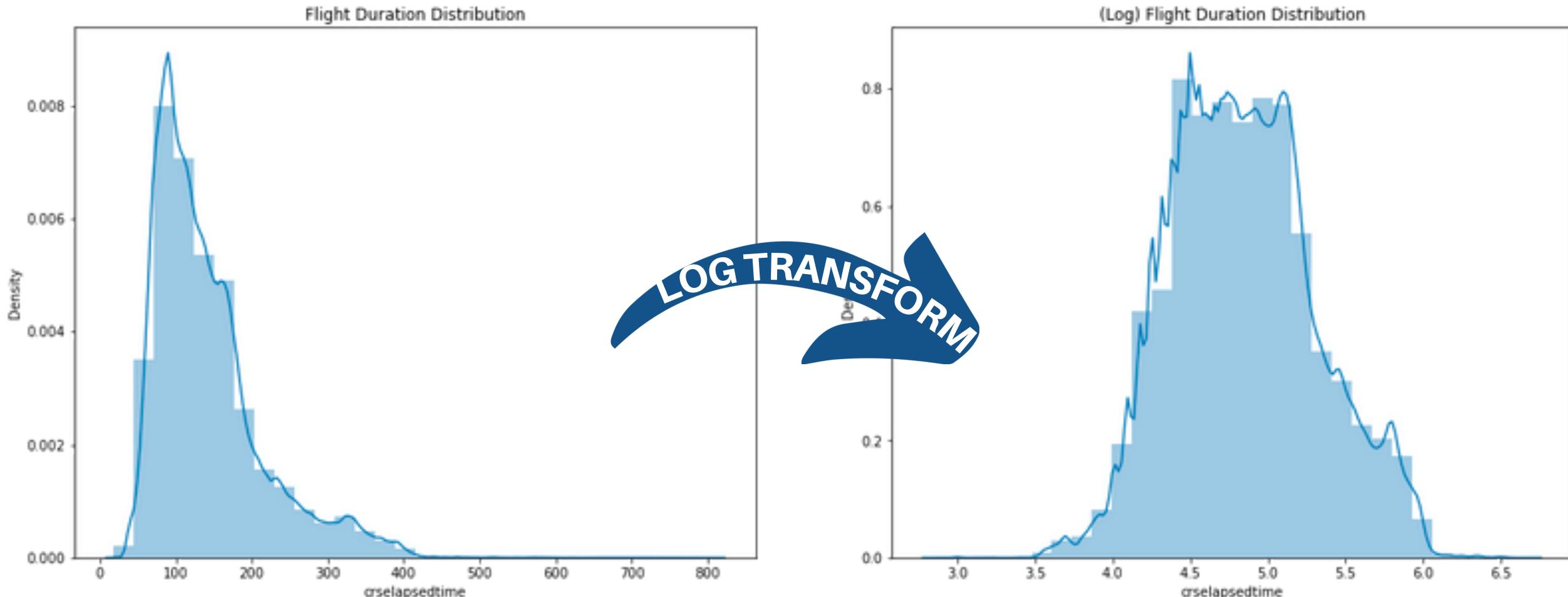


5 - Polynomial
Features

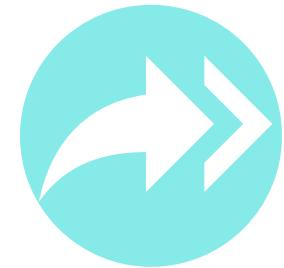
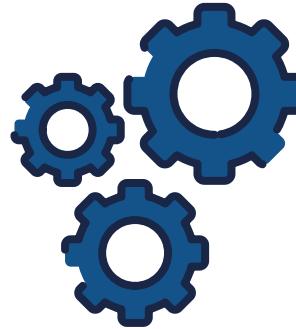


6 - One-Hot
Encoding

- Flight duration & flight distance distributions are **positively skewed**
- More normally distributed after **Log-Transformation**



FEATURE ENGINEERING

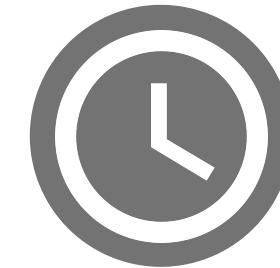


.....>

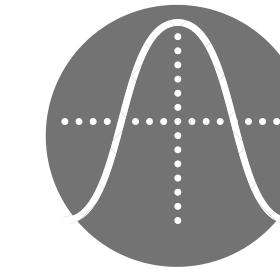


1 - Log Transform

2 - Sector per Day &
Previous Sector Delay



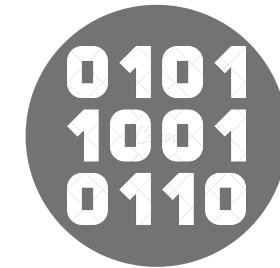
3 - Cyclical
Features Encoding



4 - Statistical Test



5 - Polynomial
Features

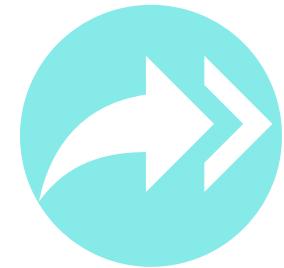
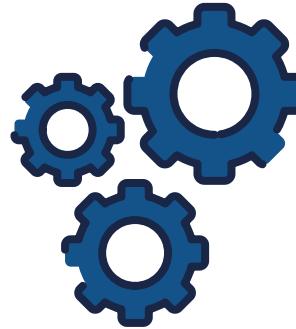


6 - One-Hot
Encoding

- ***nth_sector*** covered by an aircraft in a day, in **chronological order**
- e.g., **10 Sectors** covered by flight no. N489HA on 01 Jan 2019

nth_sector	crsdeptime	origin	crsarrtime	dest
1	0503	HNL	0543	OGG
2	0611	OGG	0649	HNL
3	0725	HNL	0806	LIH
4	0834	LIH	0910	HNL
5	0950	HNL	1031	LIH
6	1059	LIH	1145	OGG
7	1215	OGG	1255	HNL
8	1405	HNL	1450	OGG
9	1518	OGG	1555	KOA
10	1625	KOA	1713	HNL

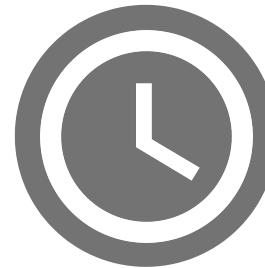
FEATURE ENGINEERING



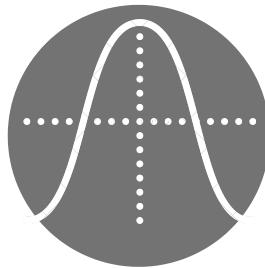
.....>



1 - Log Transform



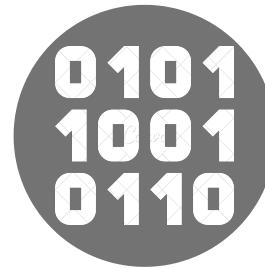
2 - Sector per Day &
Previous Sector Delay



3 - Cyclical
Features Encoding



4 - Statistical Test



5 - Polynomial
Features

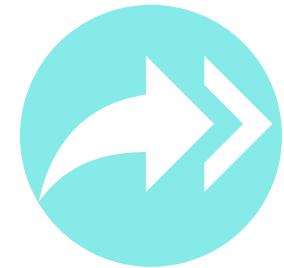
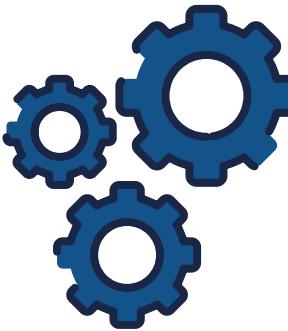
6 - One-Hot
Encoding

• Previous Sector Information

e.g. Origin Airport, Arrival Hour, Flight Duration, Flight Distance, Time Delayed

nth_sector	origin	previous_origin	arrdelay	previous_arrdelay
1	SFB	NaN	46	NaN
2	XNA	SFB	35	46.0
3	SFB	XNA	27	35.0
4	FWA	SFB	27	27.0

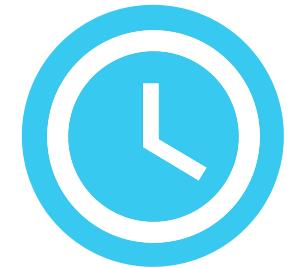
FEATURE ENGINEERING



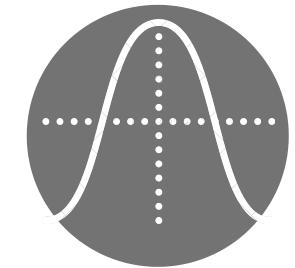
1 - Log Transform



2 - Sector per Day &
Previous Sector Delay



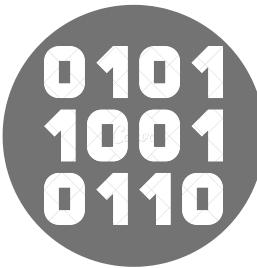
3 - Cyclical
Features Encoding



4 - Statistical Test

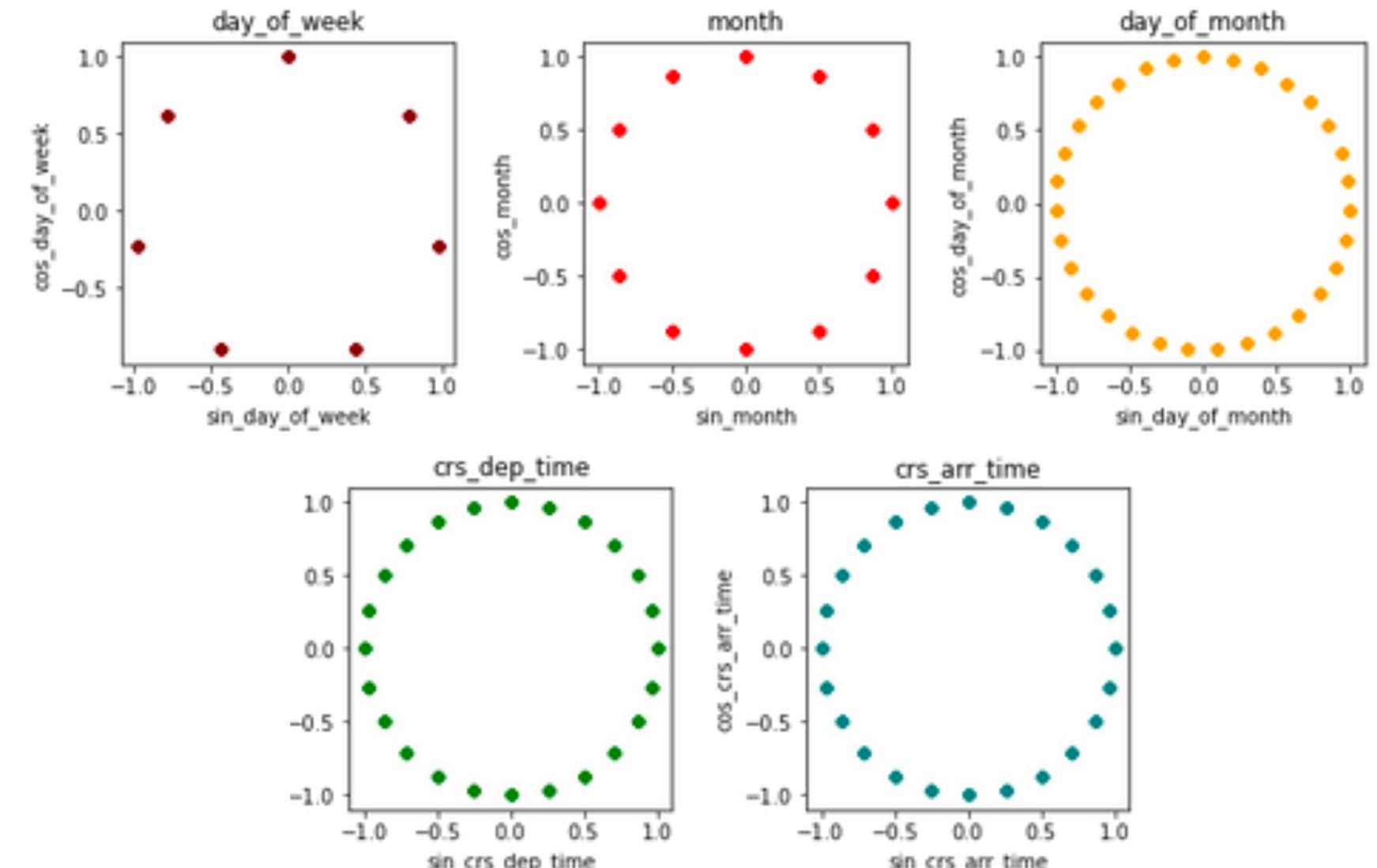


5 - Polynomial
Features

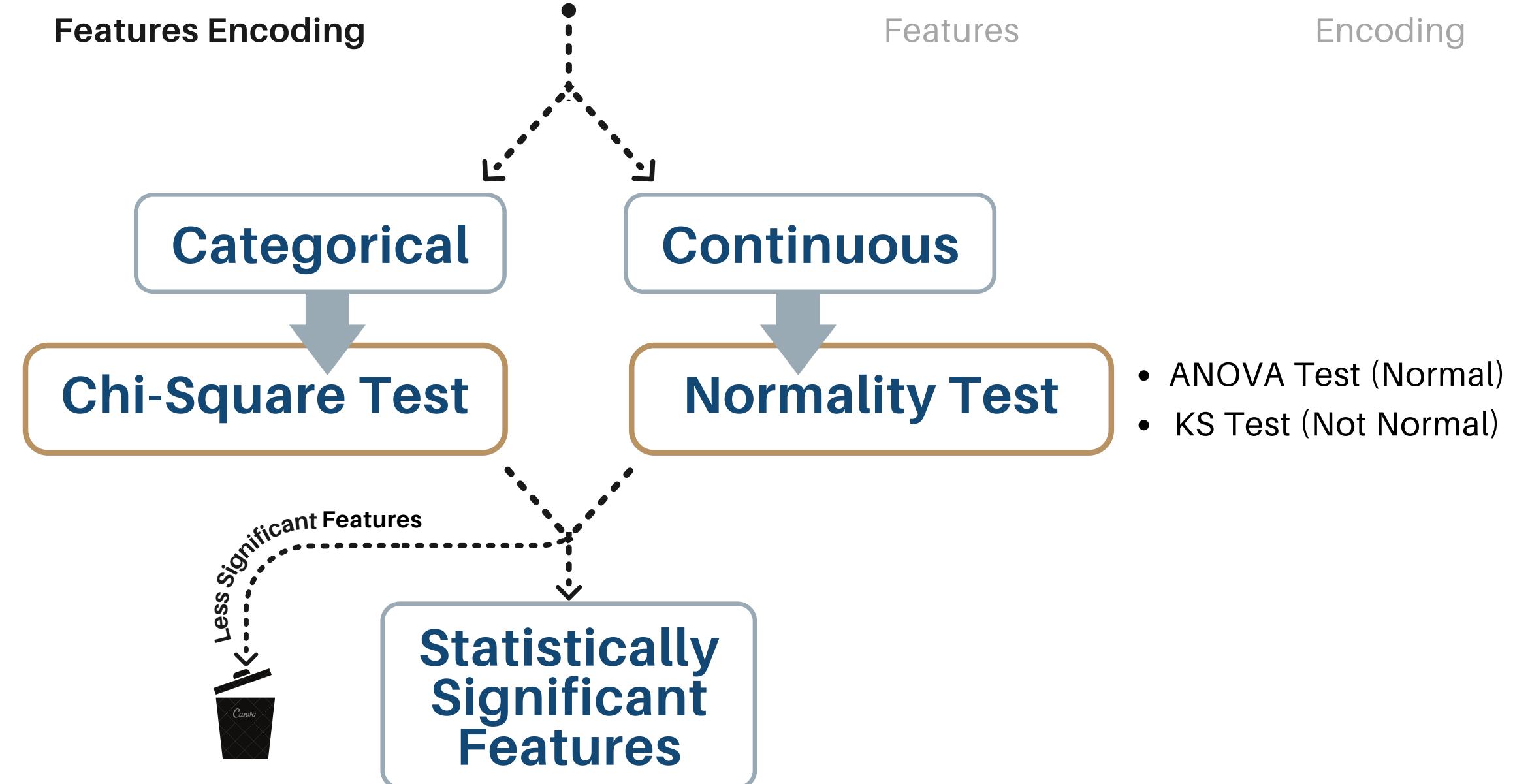
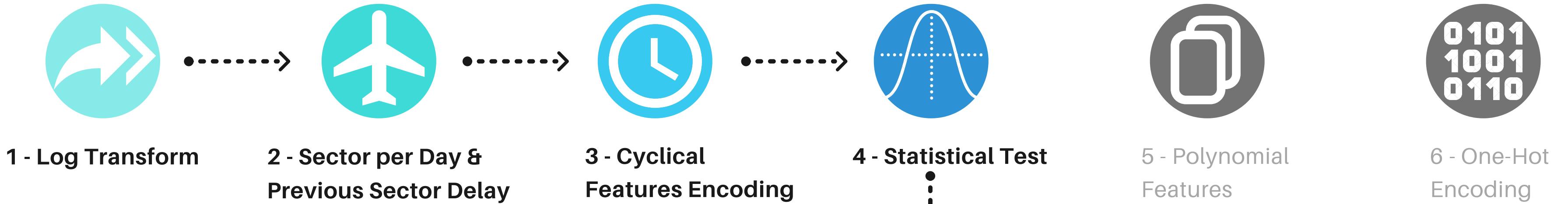
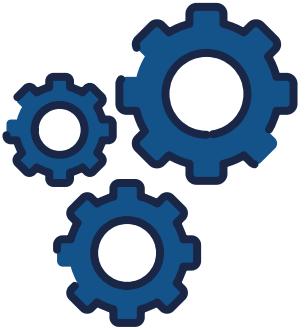


6 - One-Hot
Encoding

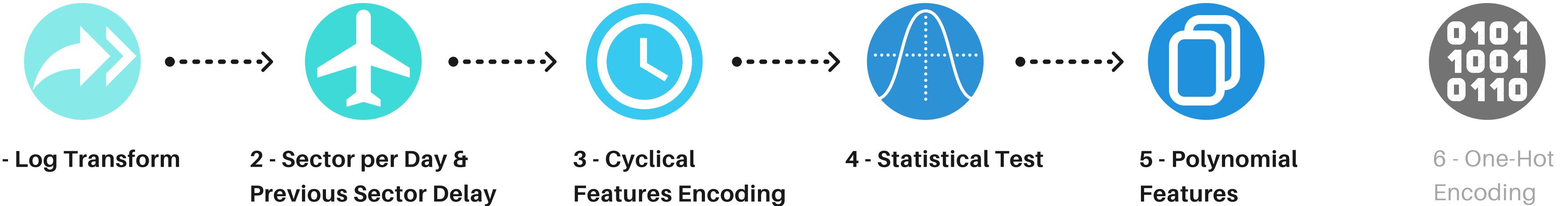
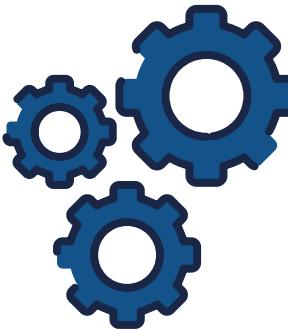
- **Cyclical Features Encoding**
by trigonometric data conversion
- Convert features into a representation that can preserve the cyclical properties of time data



FEATURE ENGINEERING



FEATURE ENGINEERING

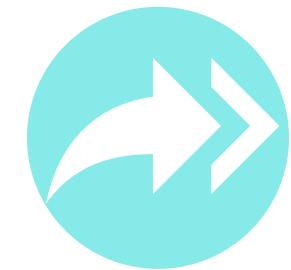
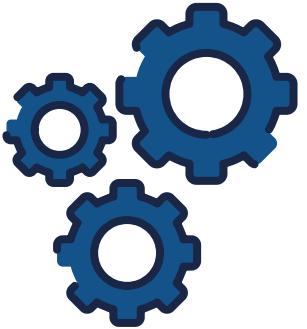


taxi out duration * log(estimated flight duration)
nth number of sector * taxi out duration
taxi out duration * taxi out duration

	delay	1
taxiout	0.24	
taxiout_X_log_crslapsedtime	0.24	
nth_sector_X_taxiout	0.23	
taxiout_X_taxiout	0.23	
cos_crs_arr_time	0.13	
nth_sector	0.11	

- Our original features have a relatively weak correlation to our target variable
- Polynomial features were manually chosen to **avoid multicollinearity** issues

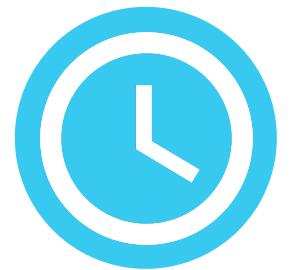
FEATURE ENGINEERING



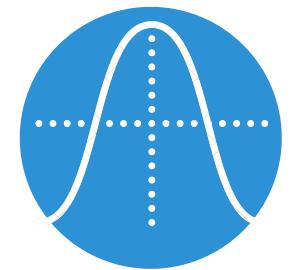
1 - Log Transform



2 - Sector per Day &
Previous Sector Delay



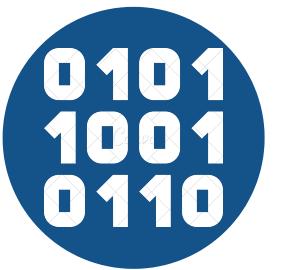
3 - Cyclical
Features Encoding



4 - Statistical Test

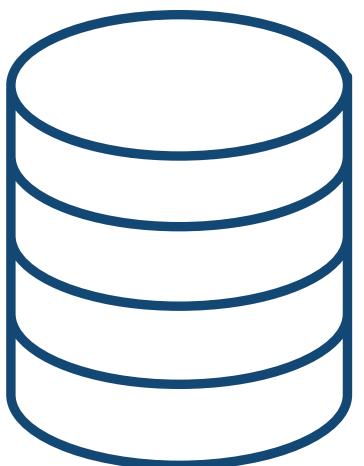


5 - Polynomial
Features



6 - One-Hot
Encoding

Dummifying our
categorical features



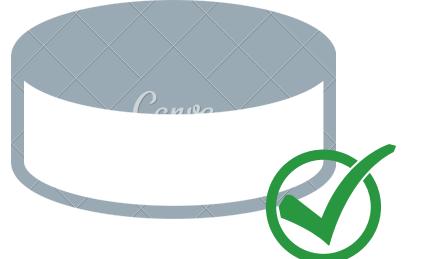
After Dummifying

~7,200K X 1,120

Choosing
Dummified
Features

based on r -value

Randomly
Selecting 10%
of the Data



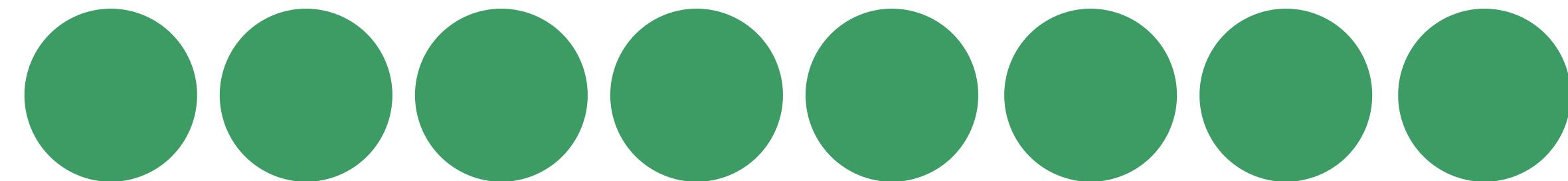
Ready for Modelling

~720K X 50

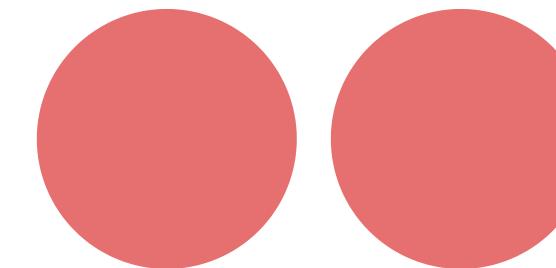
IMBALANCED DATASET



ON-TIME (0)



DELAYED (1)



MODEL EVALUATION

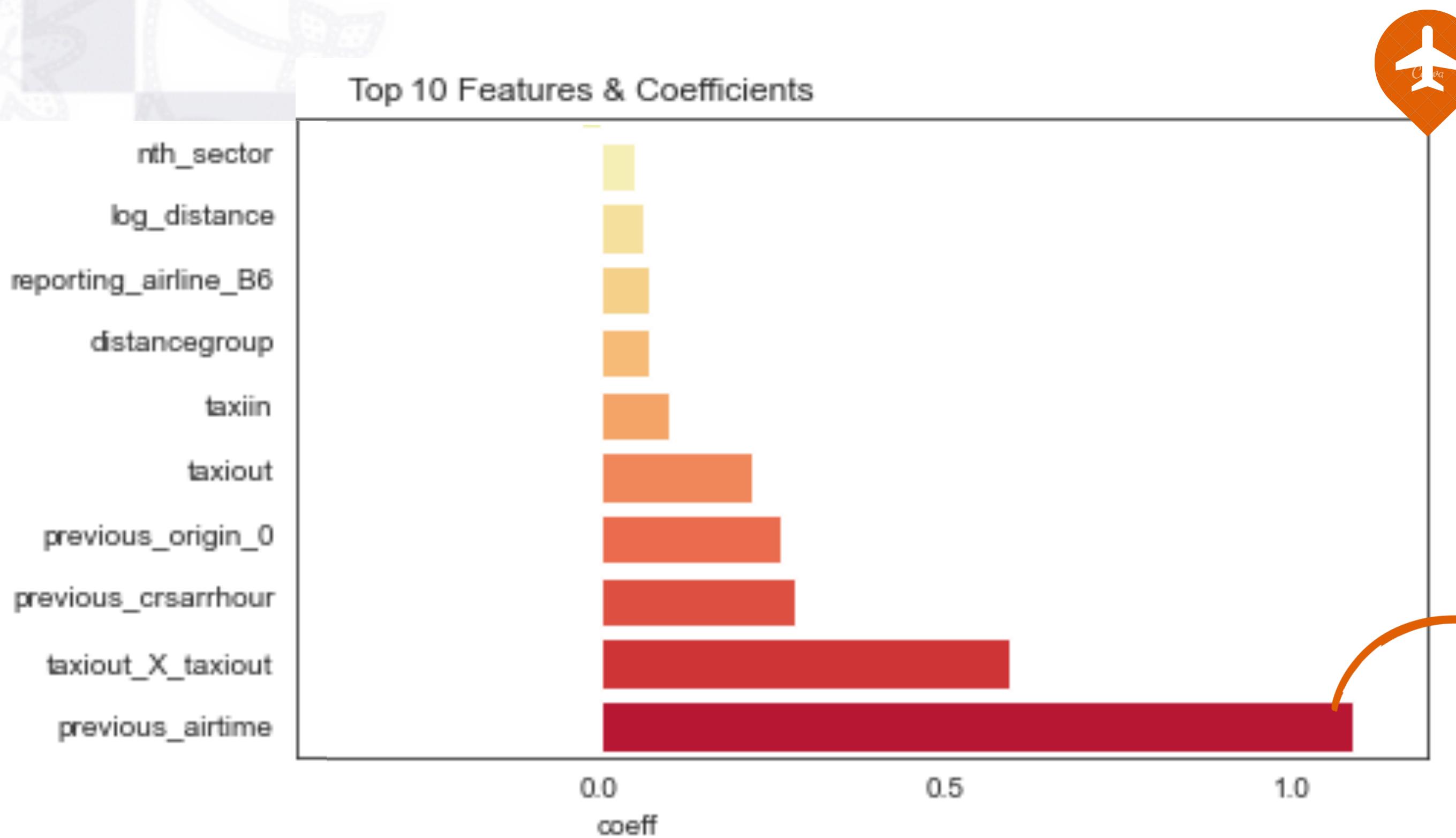


Success Metrics: 1. F1-Score
2. Recall } Best Model:
Logistic Regression

Model	Train	Test	Recall	F1 Score
Logistic Regression	0.6467	0.6464	0.6211	0.4342
Random Forest	0.9859	0.8011	0.1748	0.2195
AdaBoost	0.7998	0.799	0.1381	0.2195



FEATURE IMPORTANCE



Longer previous flight's airtime
⇒ higher chance of flight delay

RECOMMENDATIONS



MAINTAINING CUSTOMER SATISFACTION

Unavoidable Delays

- Notifying in advance
- Deploy extra staff
- Passenger welfare



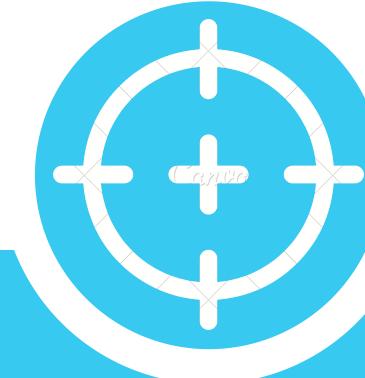
BETTER FLIGHT PLANNING

Avoiding Delays

- ↑ Buffer time for aircraft turnaround
- ↓ Connecting flights
- ↑ Point-to-point flights



LIMITATIONS



**1.
Poor F1
& Precision
Score**



**2.
No
International
Flight Data**



**3.
Change in
Future Flight
Patterns
(Post-COVID)**

POSSIBLE ENHANCEMENTS



- 01** ***Handle Imbalanced Data*** 
- 02** ***Analyse Airports*** 
- 03** ***Predict Extent of Delay*** 
- 04** ***Weather Forecast*** 

Methods: SMOTE, Threshold Moving, Random Undersampling/Oversampling, Class Weights

Analyse airport delays and traffic

Regression Analysis for Prediction

Complementing weather forecast to do real-time delay prediction



Thank You!

