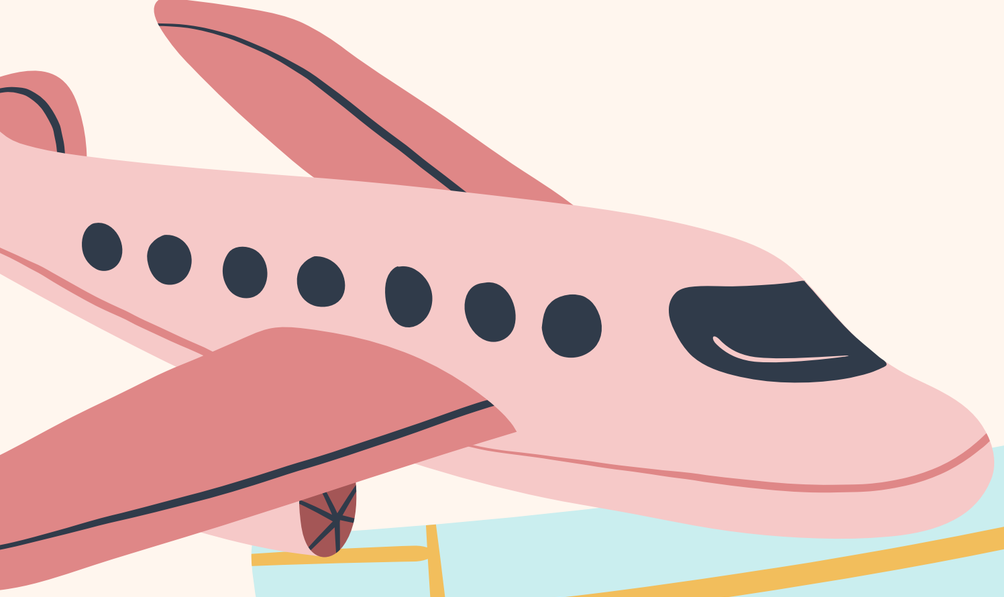


SC1015 MINI PROJECT

Presented by: Sai Harsha, Mustafa, Yifei

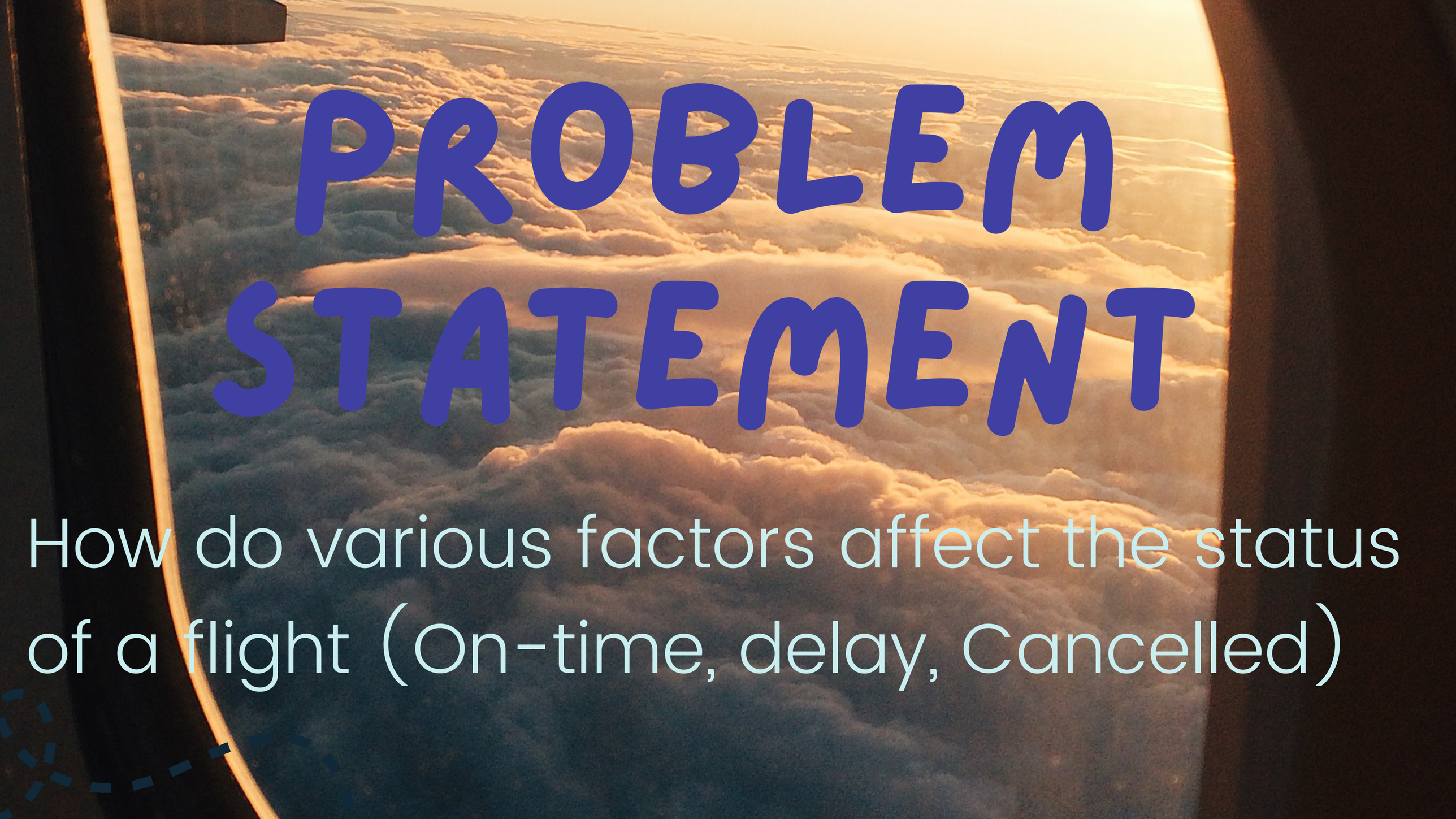
TOPIC OUTLINE

1)INTRODUCTION
2)DATA PREPARTION
3)MODEL & ANALYSIS
4)OUTCOMES

INTRODUCTION

Flight punctuality is a critical factor in the airline industry, as it impacts customer satisfaction, operational efficiency, and overall business performance. Analyzing early flight arrivals can provide valuable insights into the factors that contribute to on-time performance.

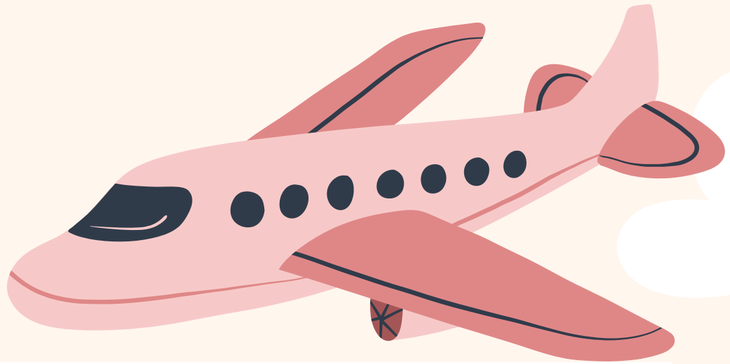
By predicting Departure delays, airlines and airports can better **allocate** resources. This can lead to more **efficient** operations and reduced delays.



PROBLEM STATEMENT

How do various factors affect the status of a flight (On-time, delay, Cancelled)

DATASET



PATRICK ZELAZKO · UPDATED 4 MONTHS AGO

43 | New Notebook

Flight Delay and Cancellation Dataset (2019-2023)

Data Analysis, Exploratory, Time-Series, Visualization (~29m total rows; 3m SRS)

Data Card | Code (6) | Discussion (1) | Suggestions (0)

About Dataset

Airline Flight Delay and Cancellation Data, August 2019 - August 2023

US Department of Transportation, Bureau of Transportation Statistics

https://www.transtats.bts.gov

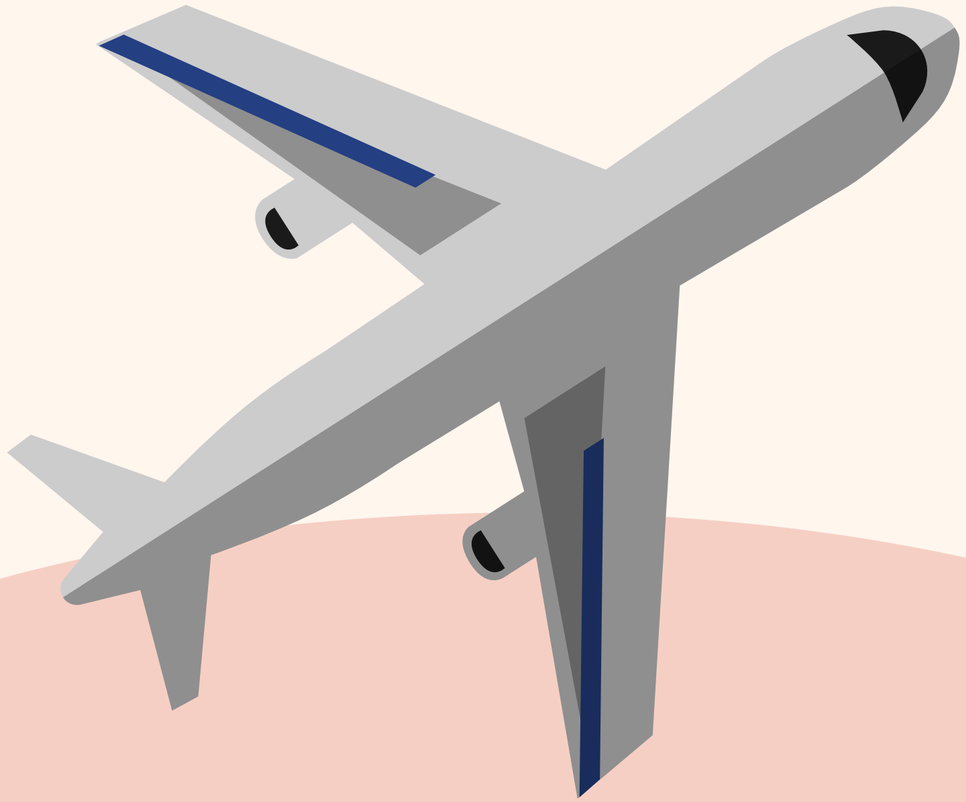
MOTIVATION

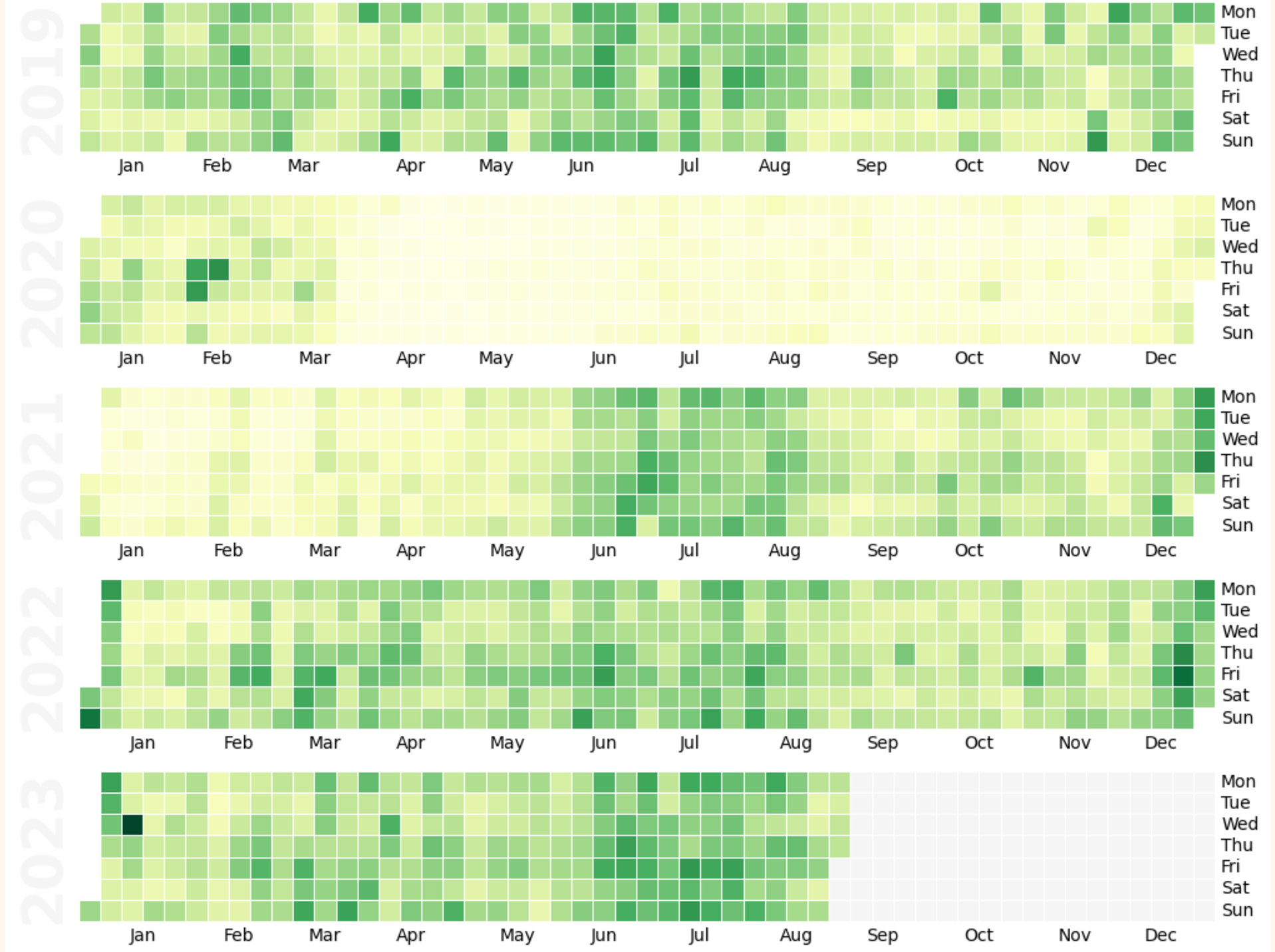
Every time a flight is scheduled to depart, unforeseen circumstances such as favorable weather conditions, or efficient operations can result in flights departing earlier than scheduled.

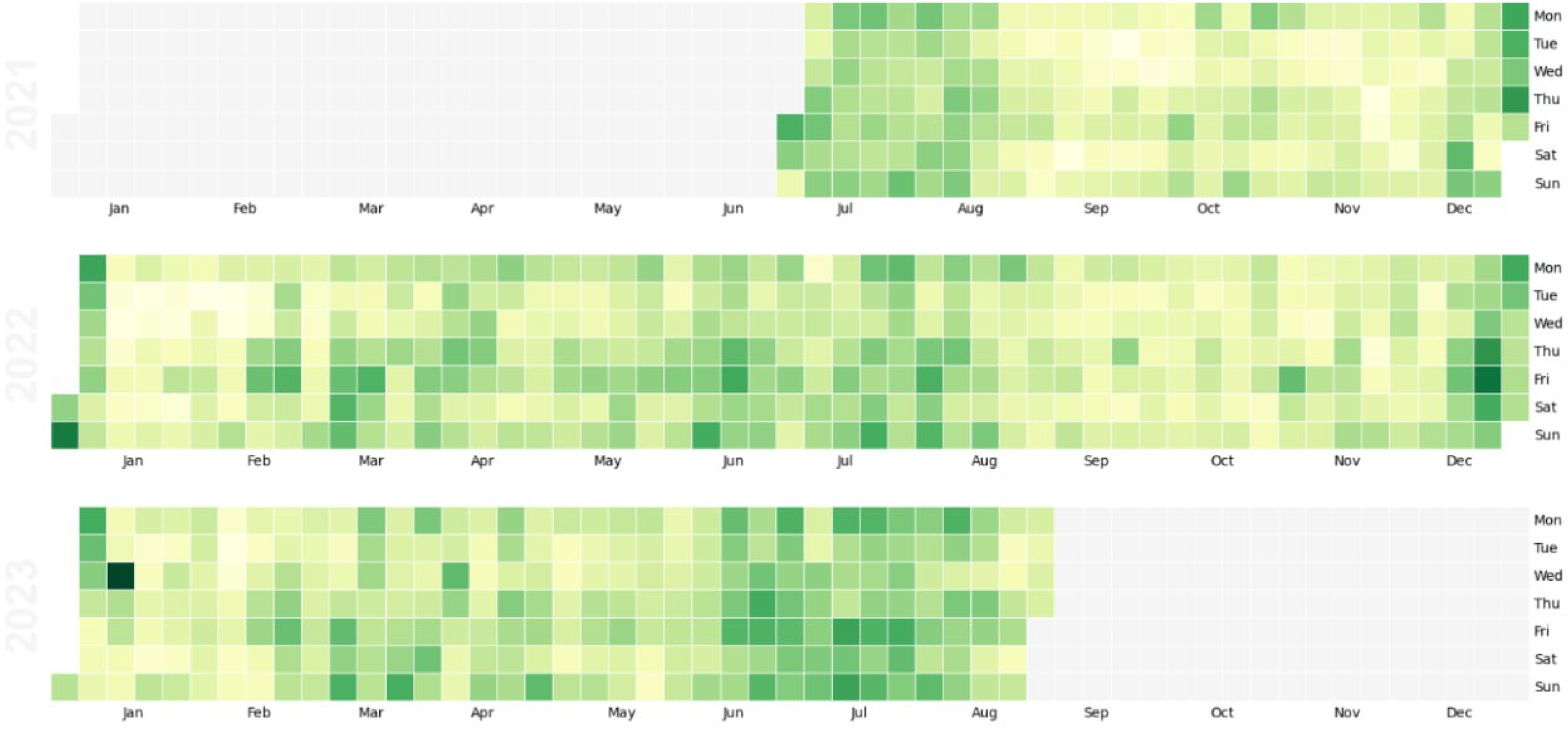
Passengers may find that their flights are ready for boarding sooner than anticipated

By accurately predicting the likelihood of a flight being early, our aim is to assist travelers in better planning their time and minimizing disruptions to their schedules.

DATA PREPARATION




2019
2020
2021
2022
2023
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


2021
2022
2023
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

DATA CLEANING

1) The initial dataset had over 32 columns, but we only selected the relevant 23 ones, that can be seen in the picture.

2) Cleaned the data to remove all flights from 2021-2023 as we chose not to consider flights from that year.

```
FlightDate                        object
Airline                           object
Airline_dot                       object
Airline_code                      object
DOT_Code                           int64
Flight_Number_Reporting_Airline    int64
Origin                            object
OriginCityName                    object
Dest                              object
DestCityName                      object
CRSDepTime                         int64
DepTime                          float64
DepDelay                         float64
TaxiOut                          float64
WheelsOff                        float64
WheelsOn                         float64
TaxiIn                           float64
CRSArrTime                         int64
ArrTime                          float64
ArrDelay                         float64
Cancelled                        float64
CancellationCode                  object
Diverted                         float64
CRSElapsedTime                   float64
ActualElapsedTime                float64
...
WeatherDelay                     float64
NASDelay                         float64
SecurityDelay                    float64
LateAircraftDelay                float64
dtype: object
```

MODIFICATIONS

OnTime: Flight departure within 15 minutes

Cancelled: Flight is cancelled

Delayed: Flight departure more then 15 minutes

Early: Flight departure before planned time

```
df['Cancelled'] = df['Cancelled'].astype(bool)

df['DelaySituation'] = 'OnTime'
df.loc[df['Cancelled'], 'DelaySituation'] = 'Cancelled'
df.loc[(df['DepDelay'] > 15) & (~df['Cancelled']), 'DelaySituation'] = 'Delayed'
df.loc[(df['DepDelay'] < 0) & (~df['Cancelled']), 'DelaySituation'] = 'Early'

display(df.head(20))
```

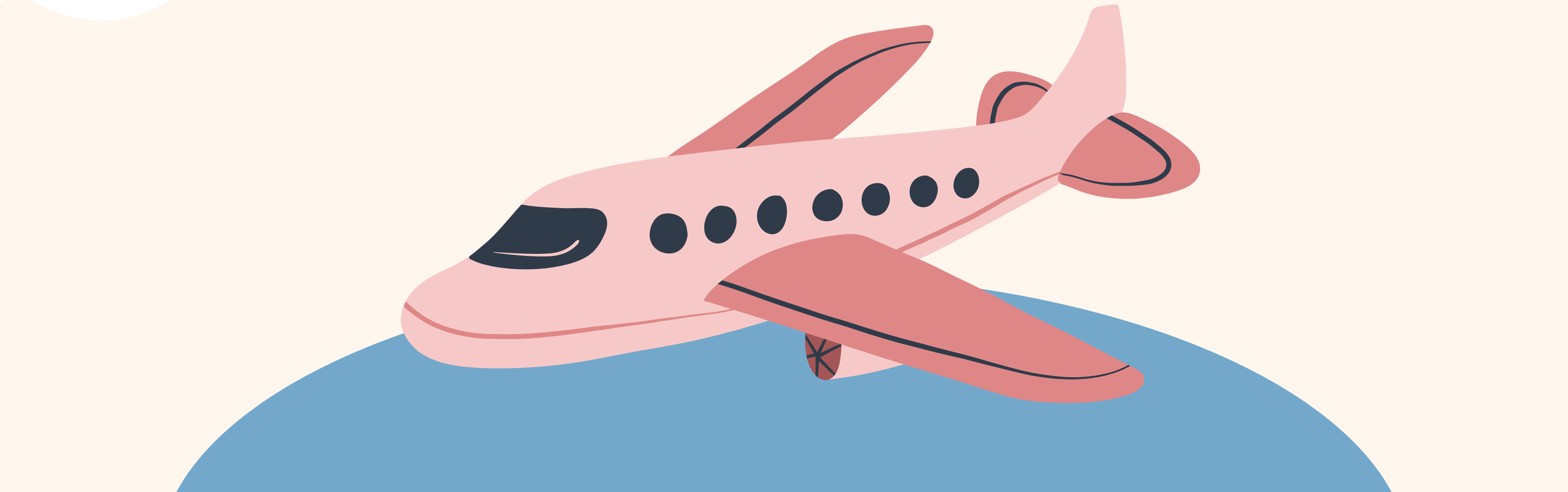
1. Adding Variables: DelaySituation

- Early (Before 0 minutes)
- On Time (0 - 15 minutes)
- Small Delays (15 - 30 minutes)
- Medium Delays (30 - 60 minutes)
- Large Delays (More than 60 minutes)

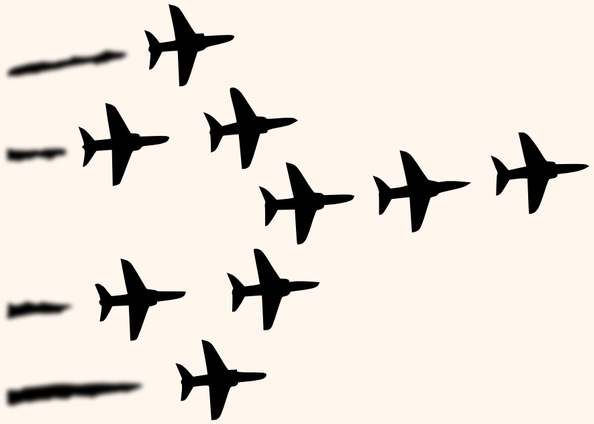
```
bins = [-np.inf ,0,15, 30, 60, np.inf]
labels = ['Early', 'OnTime','SmallDelays', 'MediumDelays', 'LargeDelays']
df['DelayCategory'] = pd.cut(df['DepDelay'], bins=bins, labels=labels, right=False)
```

2. Adding Variables: DelayCategory

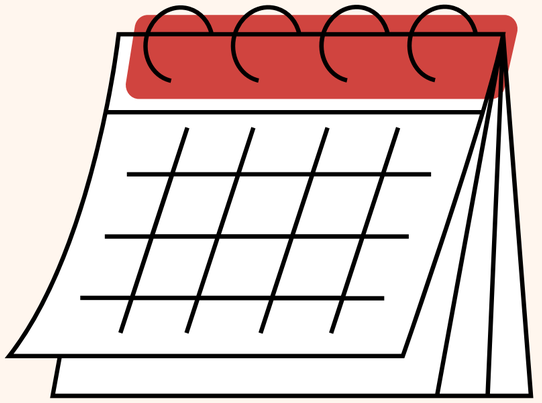
EXPLORATORY DATA ANALYSIS (EDA)



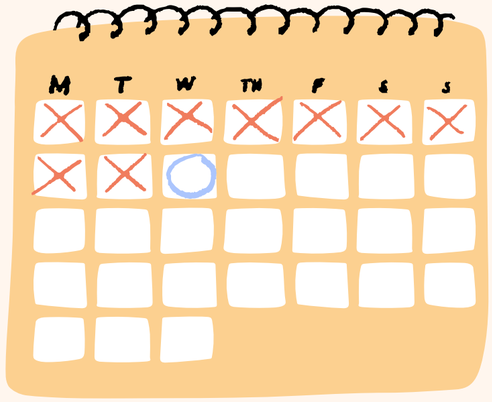
FACTORS



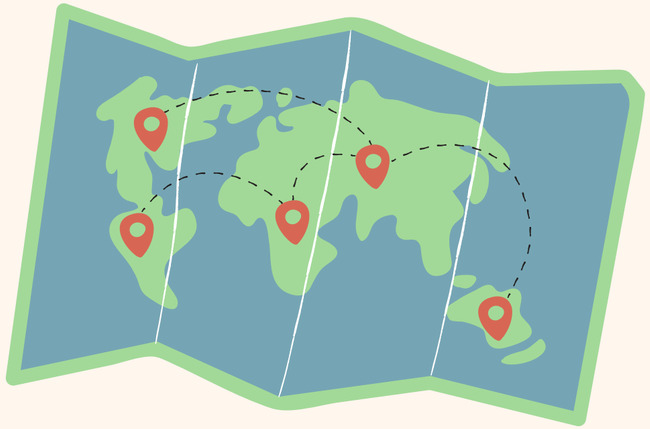
Flight Company



Month of the Year



Day of the Week



Origin City

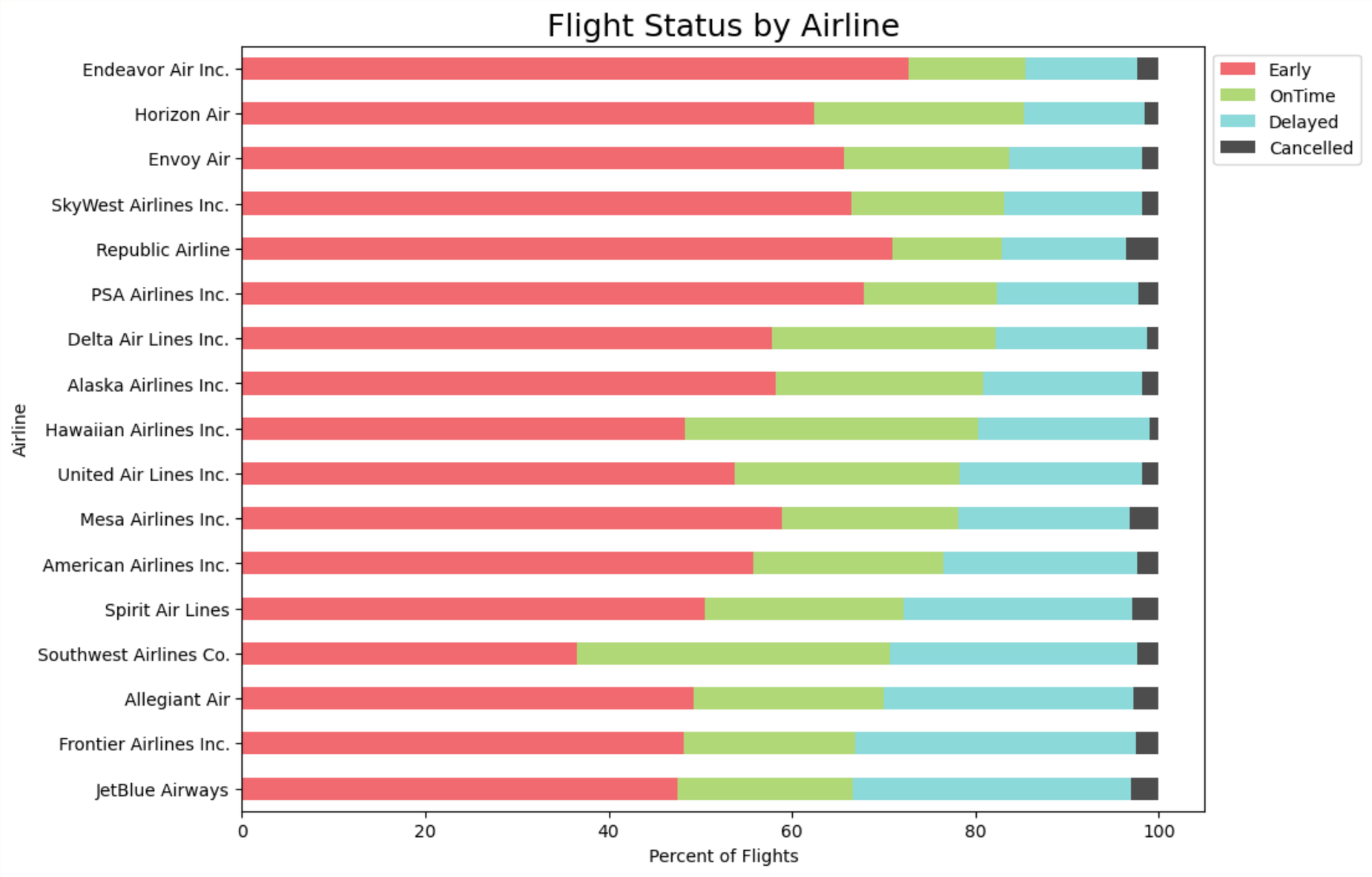
FLIGHT COMPANY

WHY?

I.Area of Operation

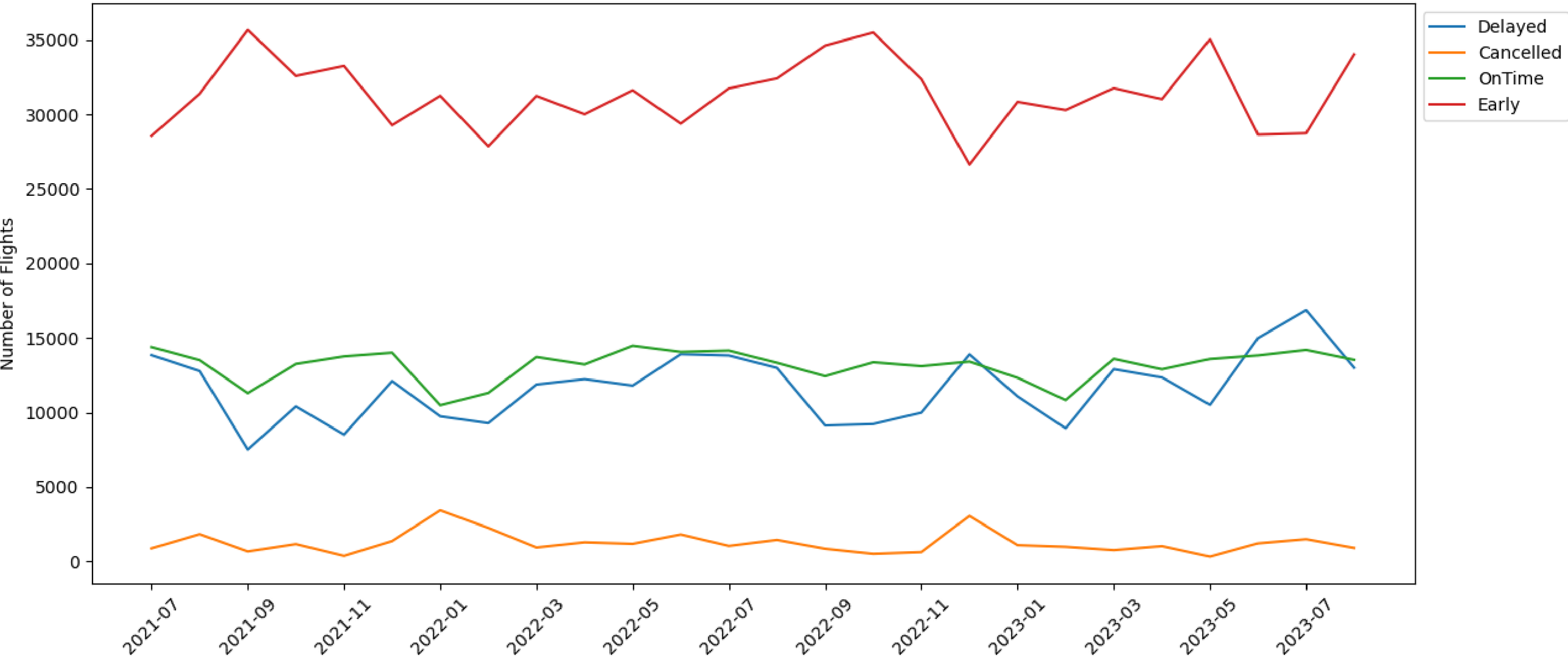
II.Fleet Size

III.Operational Issues



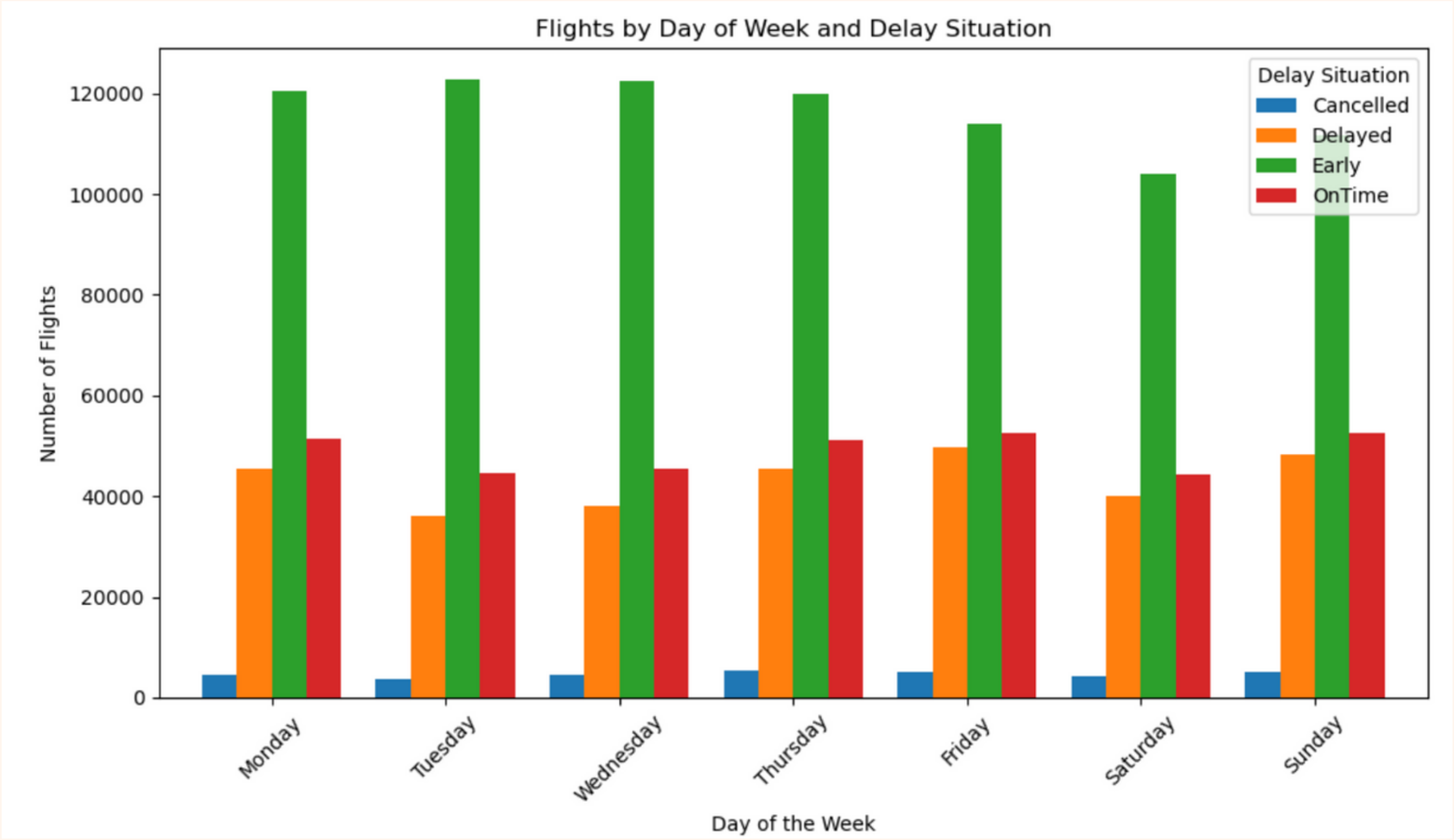
MONTH OF THE YEAR

Ho do various factors affects the status of a flight (On-time, Early, Delay and Cancelled)



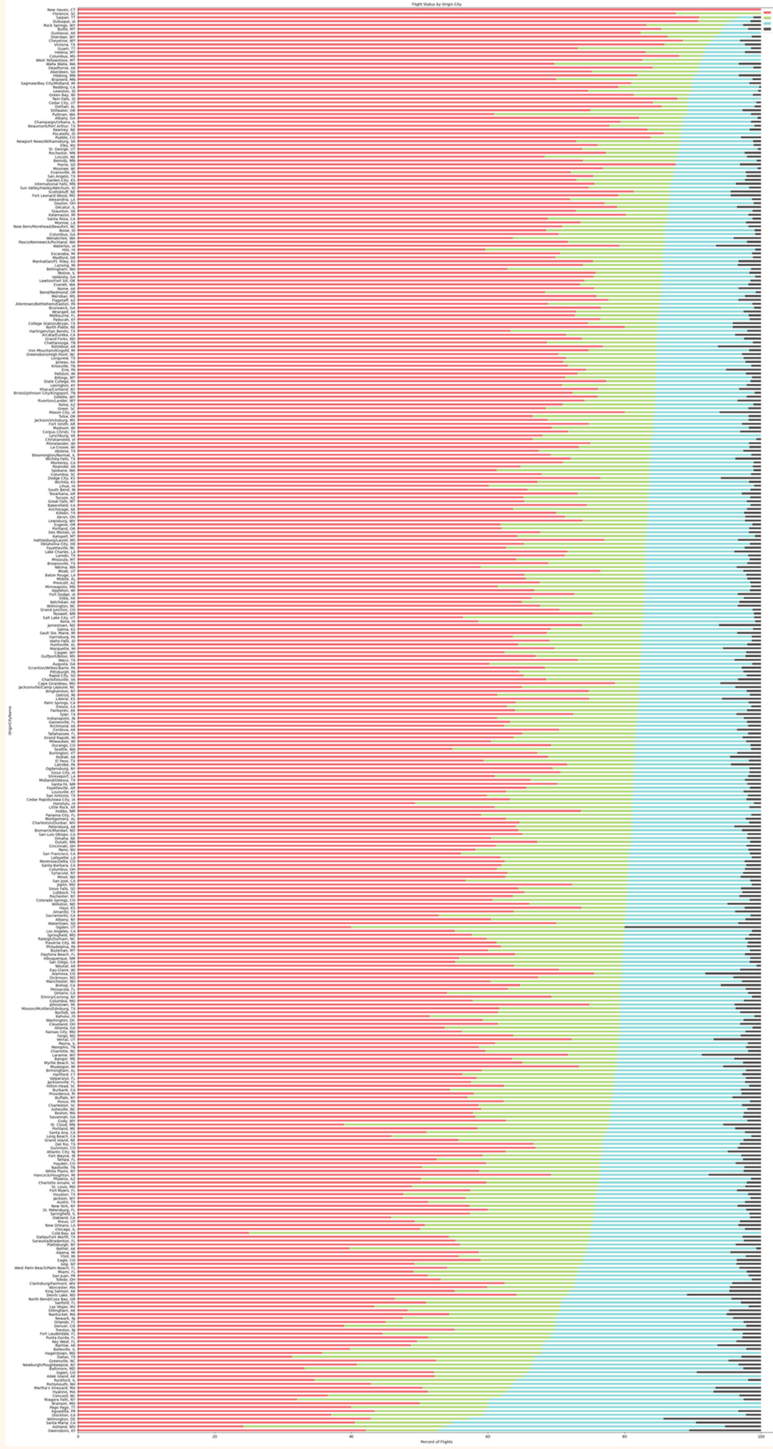
DAY OF THE WEEK

Ho do various factors affects the status of a flight (On-time and Delayed)

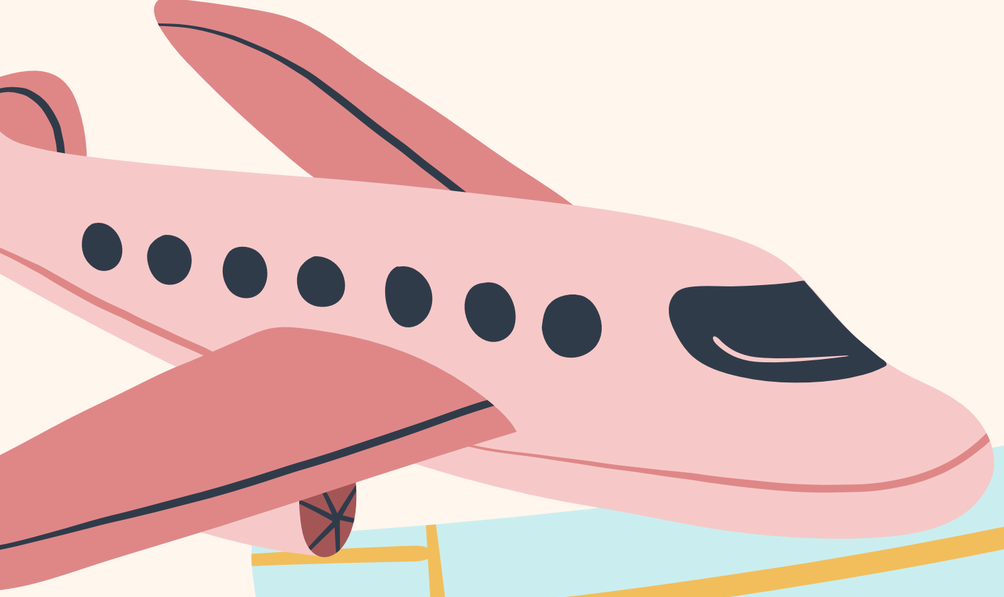


ORIGIN CITY

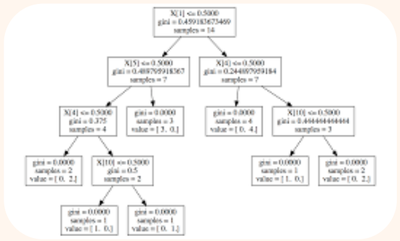
How do various factors affects the status of a flight (On-time, Early, Delay and Cancelled)



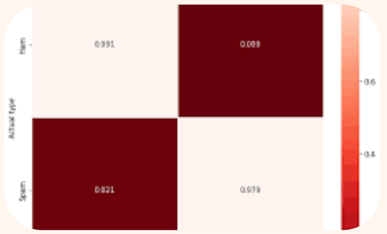
MODEL & ANALYSIS



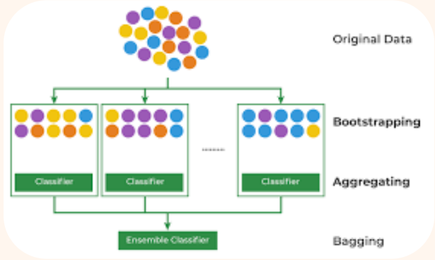
MODELS USED



DecisionTreeClassifier

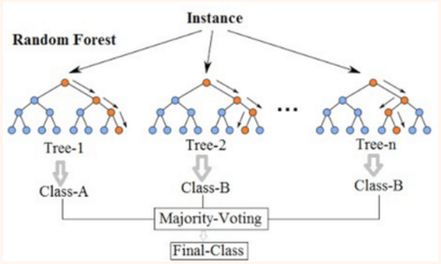


BernouliNB

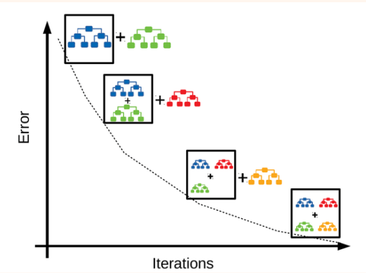


BaggingClassifier

RandomForestClassifier

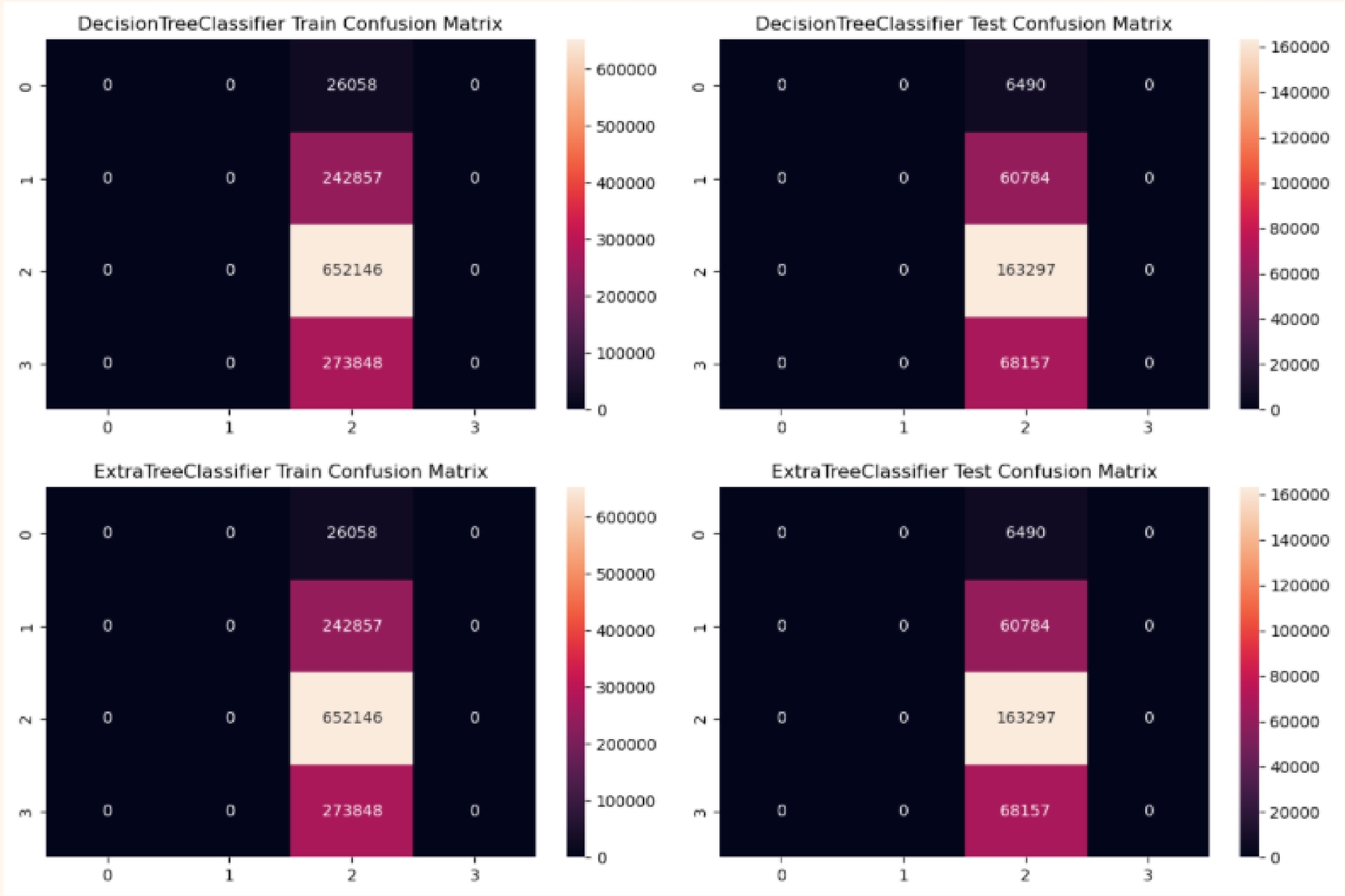


ExtraTreeClassifier

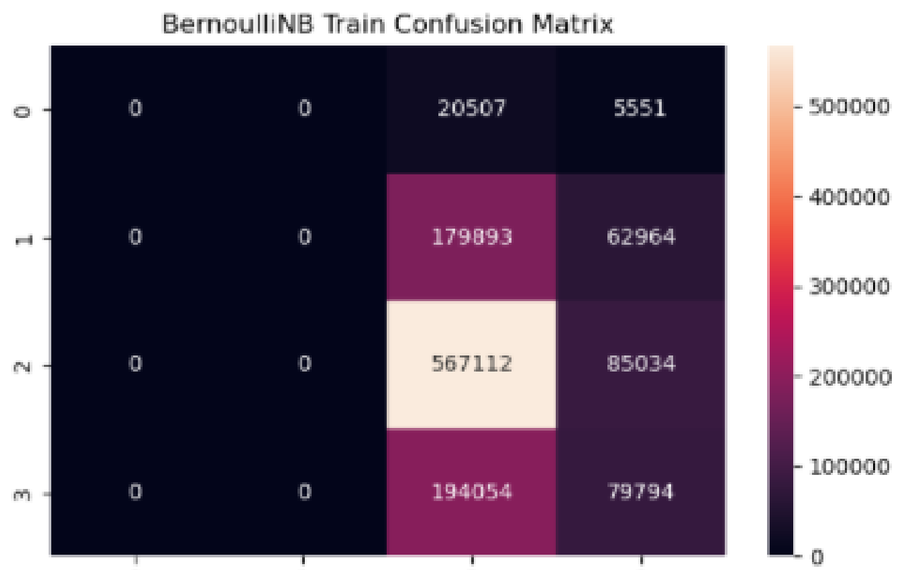


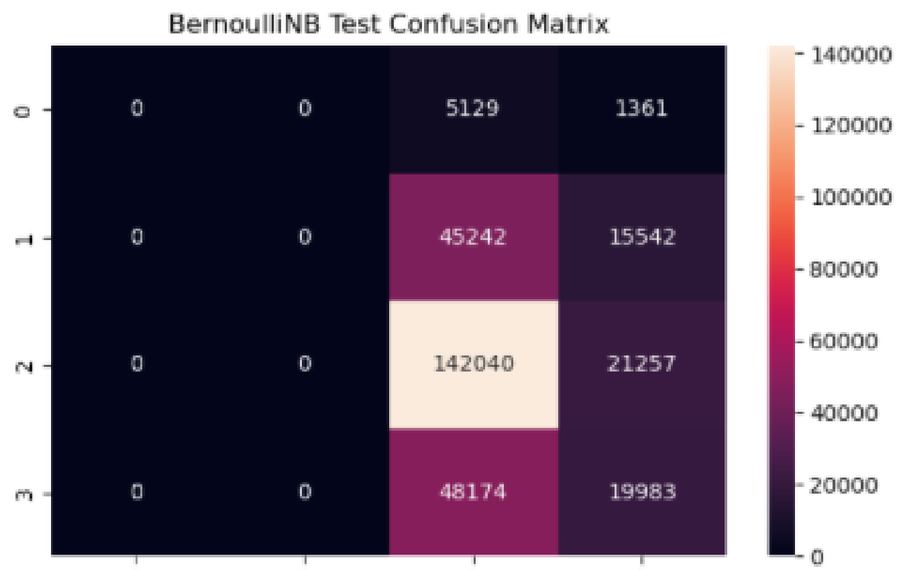
GradientBoostingClassifier

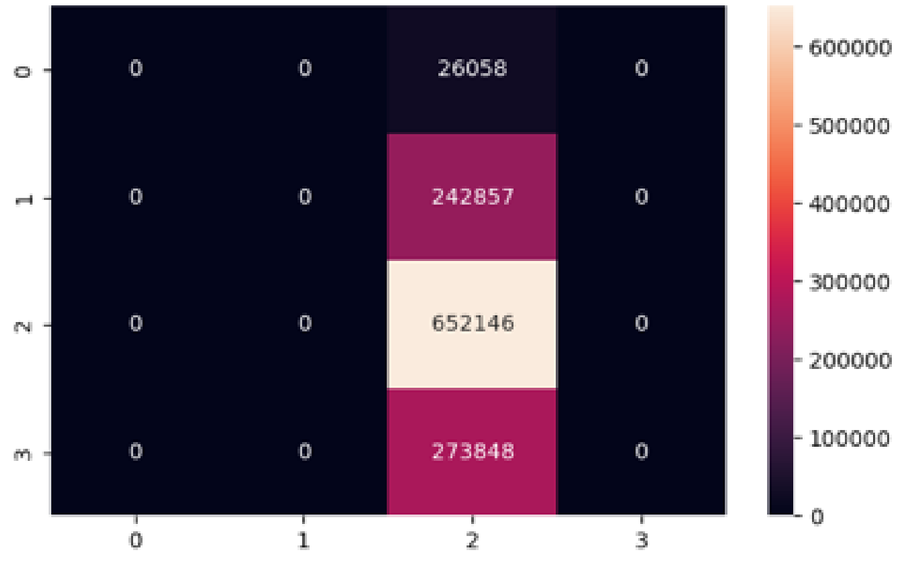
ROUND 1: ONE VARIABLE

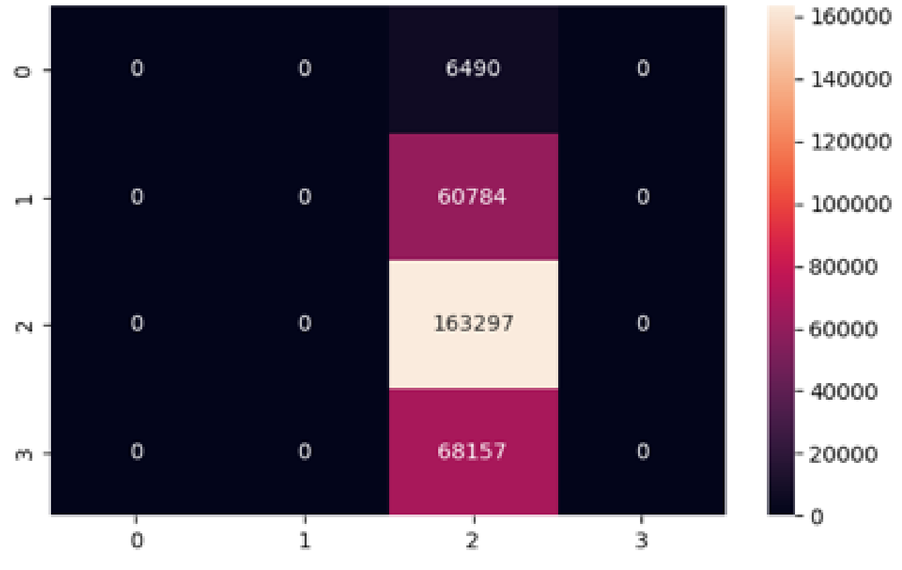


ROUND 1: ONE VARIABLE

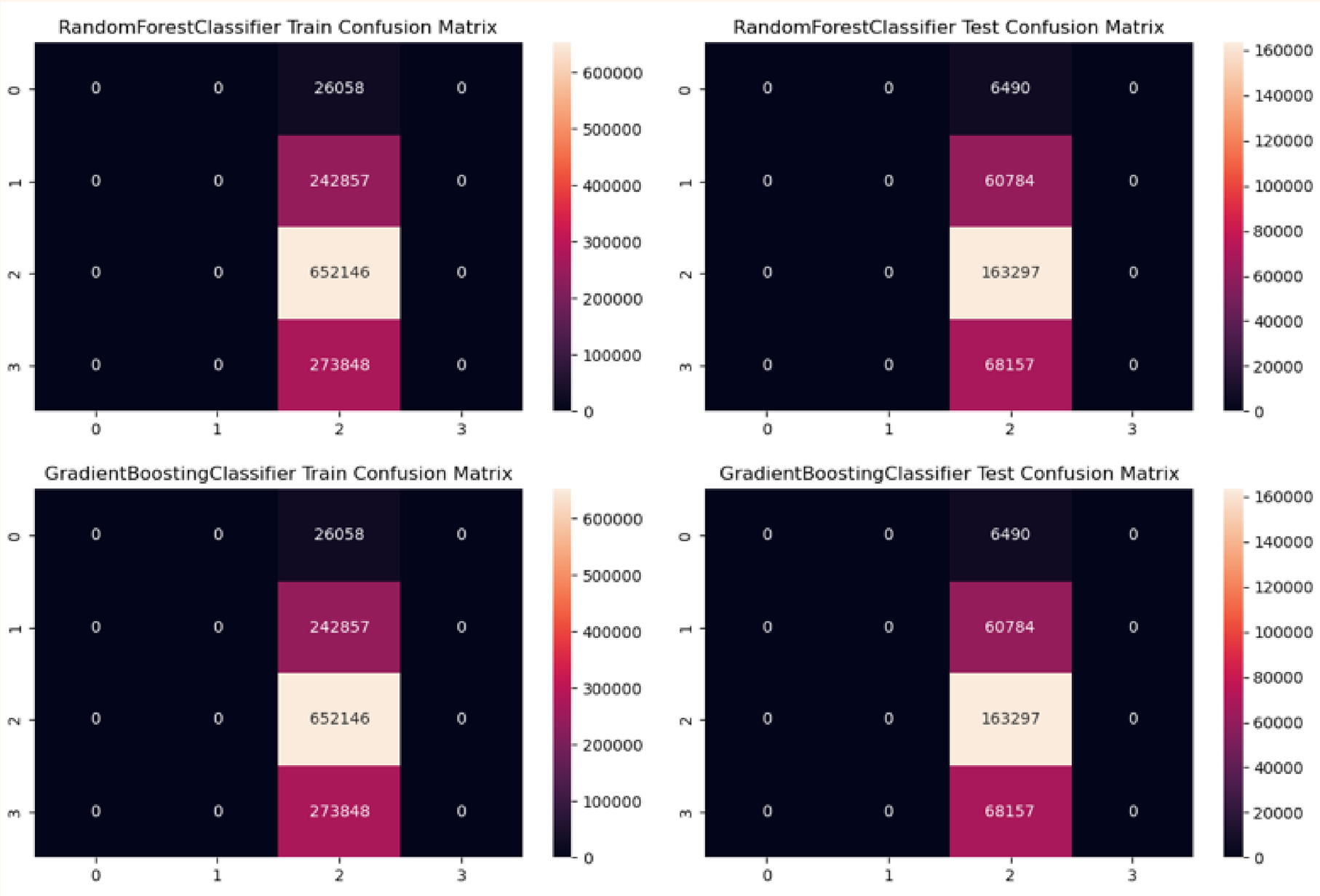








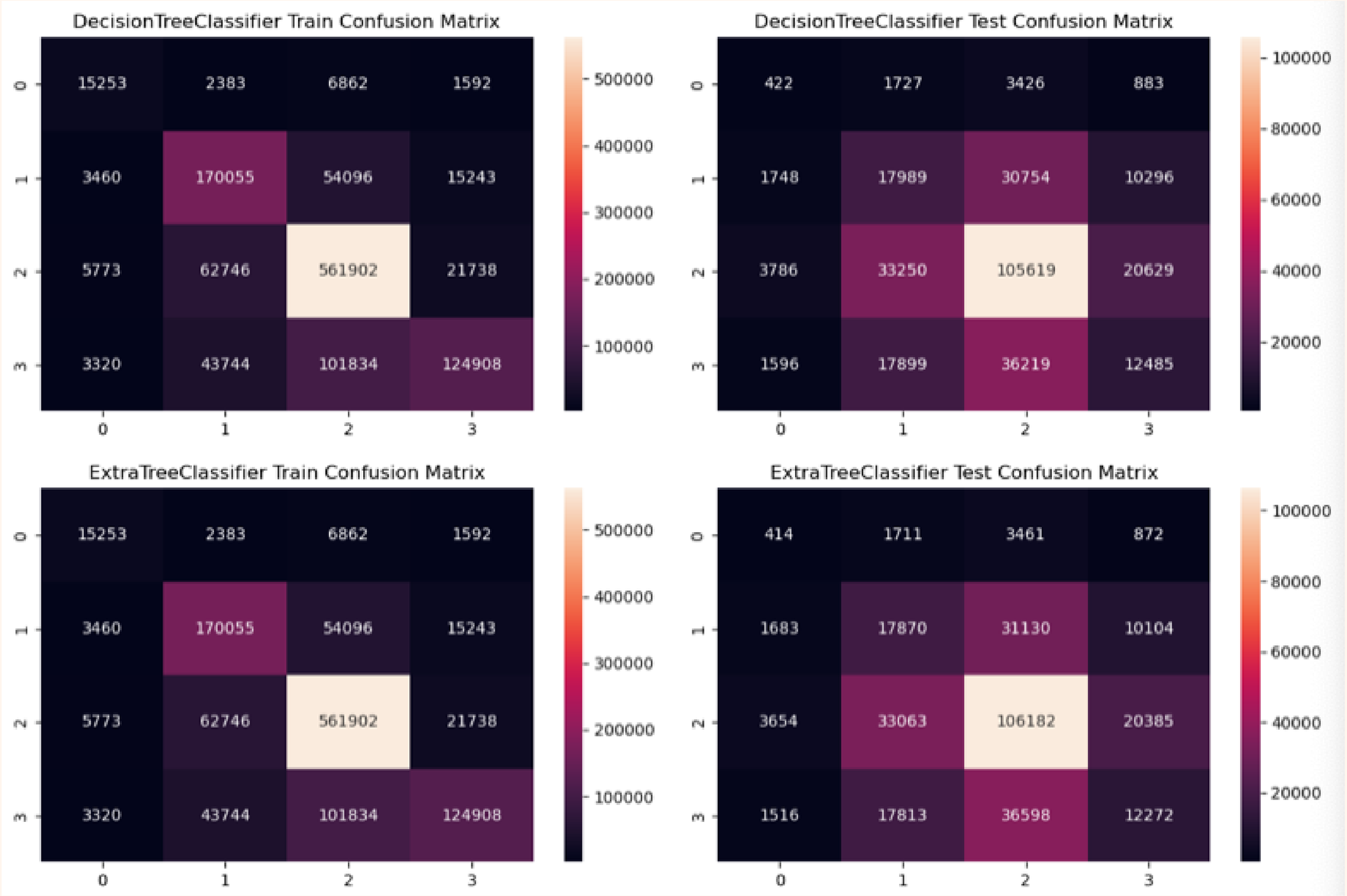
ROUND 1: ONE VARIABLE



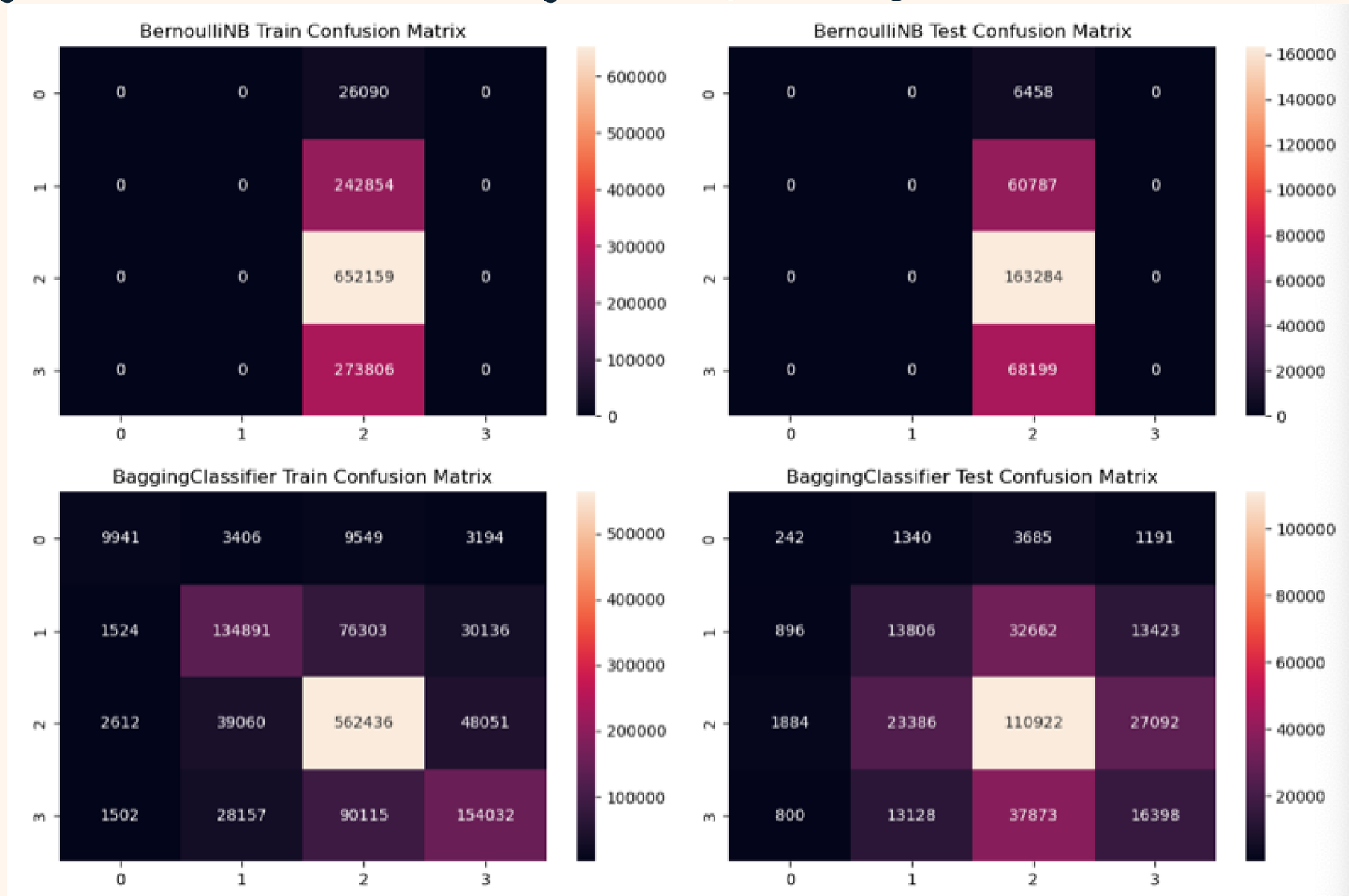
ROUND 1: ONE VARIABLE

name	train_accuracy	train_f1_score	test_accuracy	test_f1_score
DecisionTreeClassifier	0.545770	0.385394	0.546641	0.386407
ExtraTreeClassifier	0.545770	0.385394	0.546641	0.386407
BernoulliNB	0.541385	0.455715	0.542376	0.456690
BaggingClassifier	0.545770	0.385394	0.546641	0.386407
RandomForestClassifier	0.545770	0.385394	0.546641	0.386407
GradientBoostingClassifier	0.545770	0.385394	0.546641	0.386407

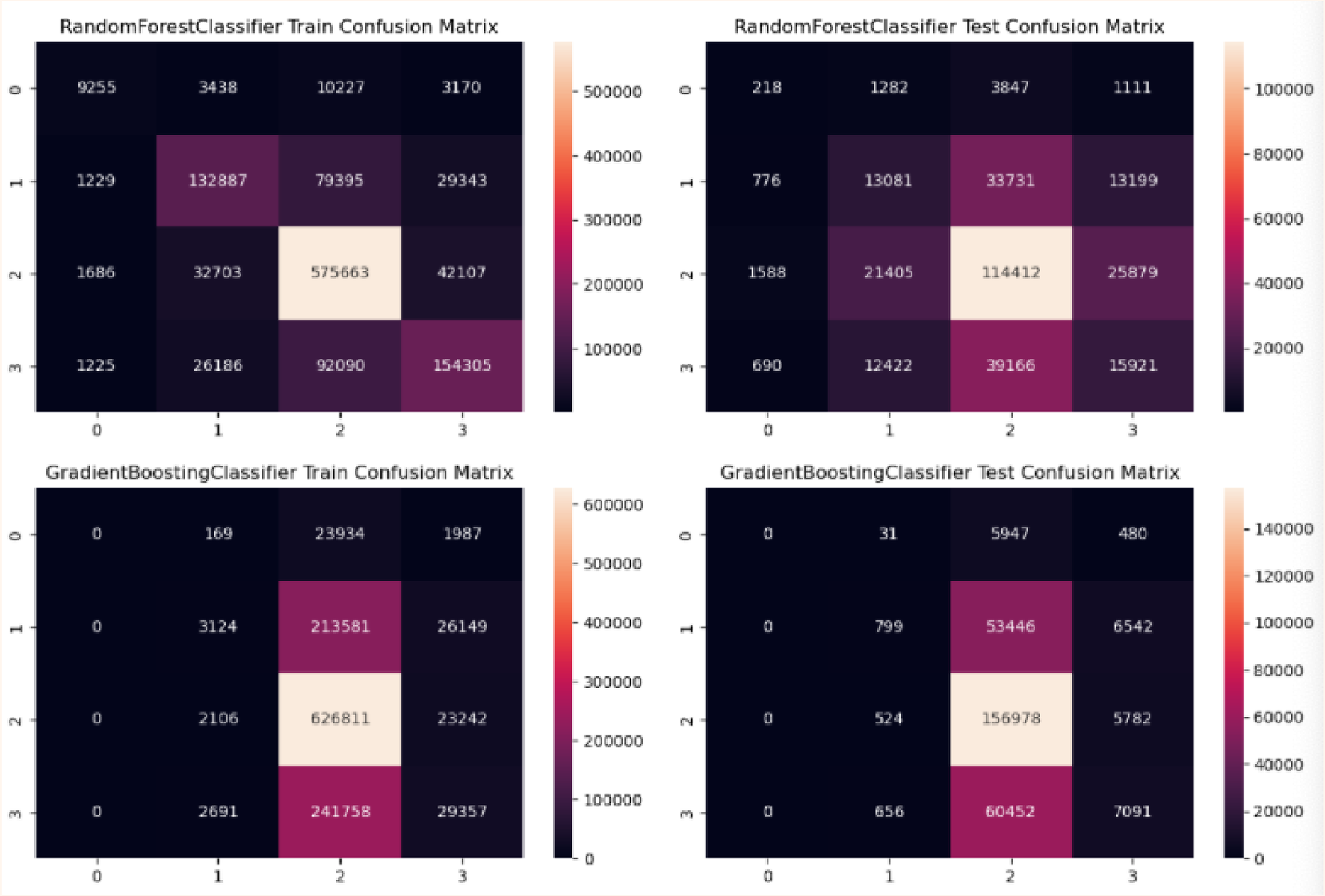
ROUND 2: ADDING MORE FEATURES



ROUND 2: ADDING MORE FEATURES



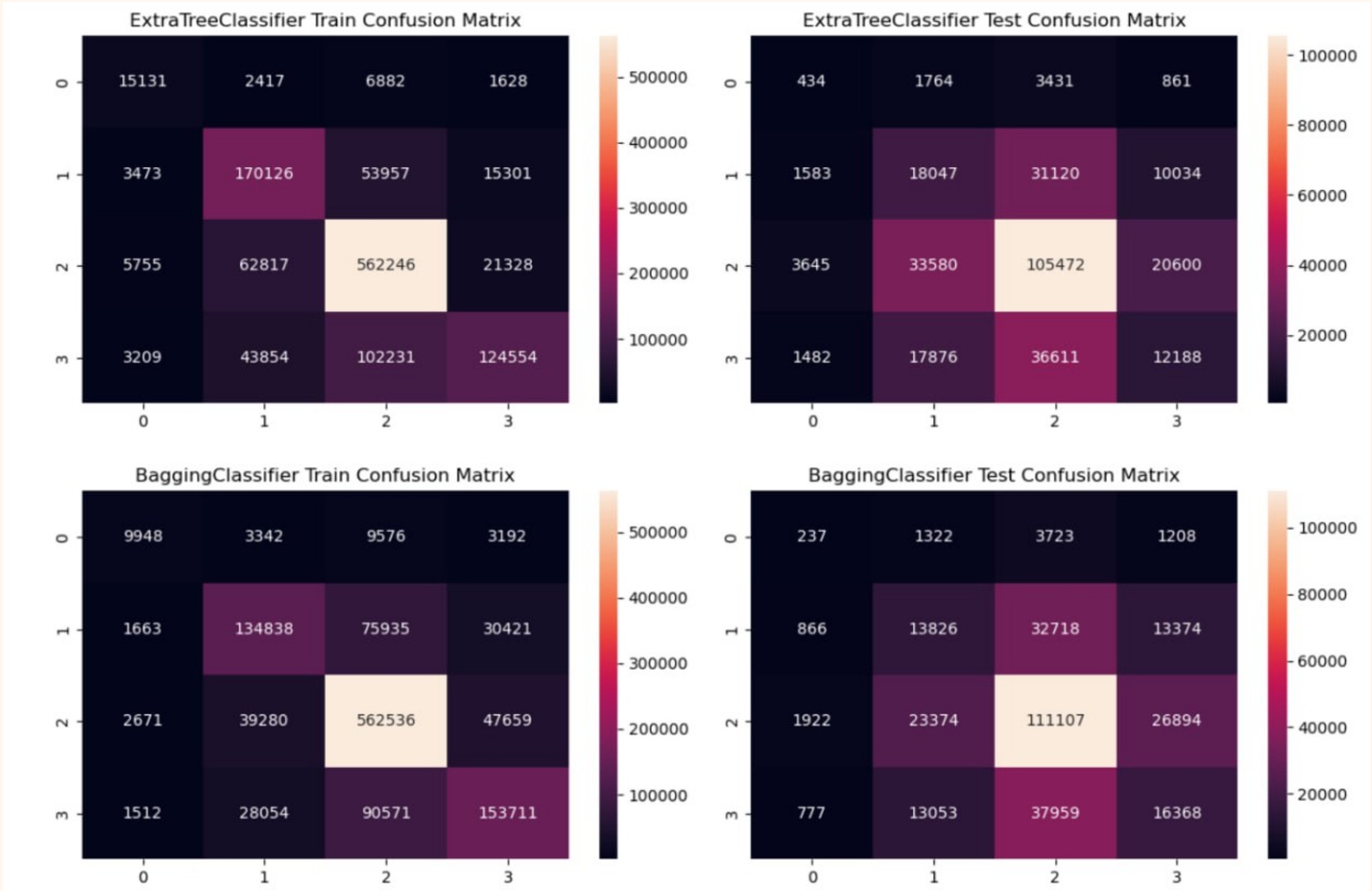
ROUND 2: ADDING MORE FEATURES



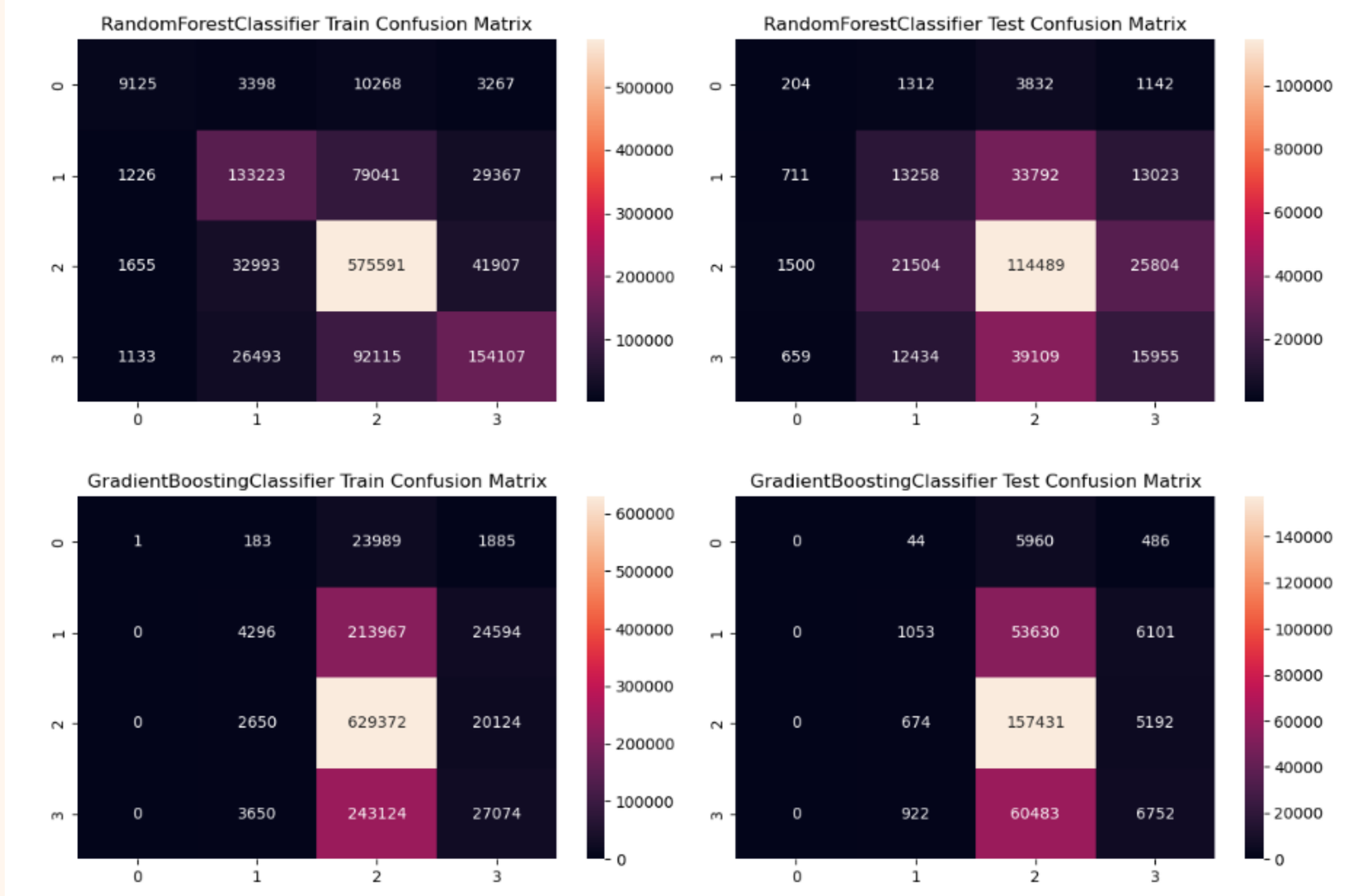
ROUND 2: ADDING MORE FEATURES

name	train_accuracy	train_f1_score	test_accuracy	test_f1_score
DecisionTreeClassifier	0.729861	0.721216	0.456988	0.447880
ExtraTreeClassifier	0.729861	0.721216	0.457734	0.447570
BernoulliNB	0.545781	0.385407	0.546598	0.386356
BaggingClassifier	0.720808	0.712834	0.473233	0.458286
RandomForestClassifier	0.729855	0.720006	0.480812	0.461178
GradientBoostingClassifier	0.551751	0.432148	0.551900	0.431854

ROUND 3:
ADDING POLYNOMIAL FEATURES



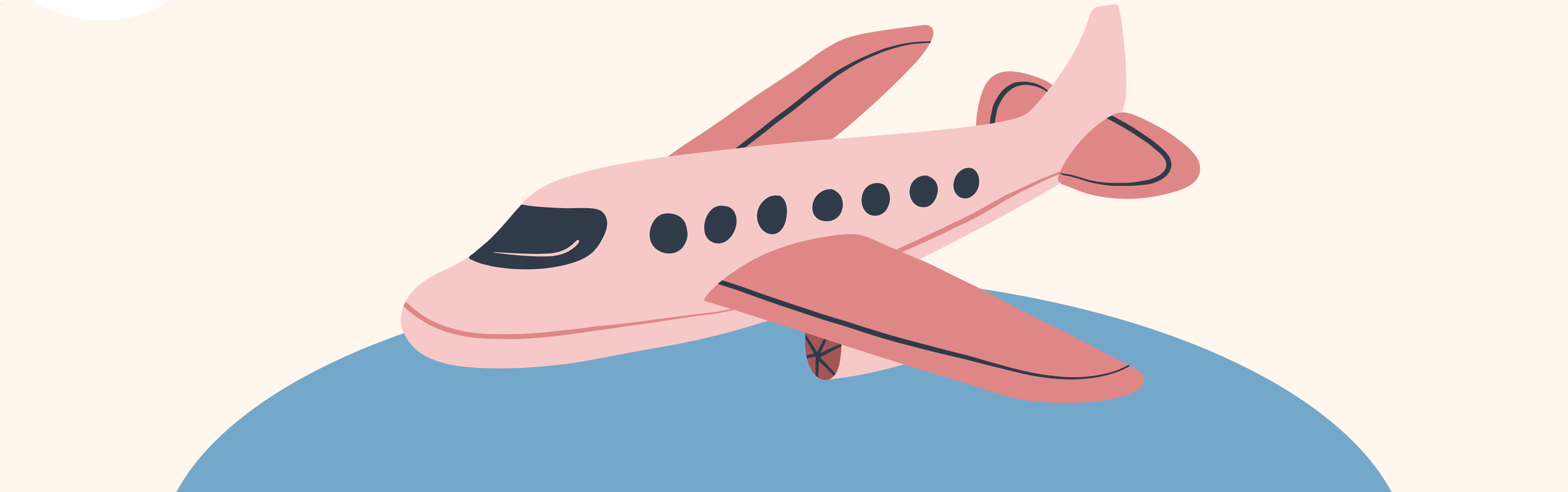
ROUND 3:
ADDING POLYNOMIAL FEATURES



ROUND 3

	name	train_accuracy	train_f1_score	test_accuracy	test_f1_score
0	ExtraTreeClassifier	0.729810	0.721051	0.455736	0.445946
1	BaggingClassifier	0.720585	0.712579	0.473802	0.458605
2	RandomForestClassifier	0.729801	0.719945	0.481729	0.462035
3	GradientBoostingClassifier	0.552965	0.432352	0.553132	0.432756

OUTCOMES



CONCLUSION

More Features had More accurate test results.

Final accuracy still low despite modifciations done to the model

Polynomial Features added had no significant effect on test accuracy

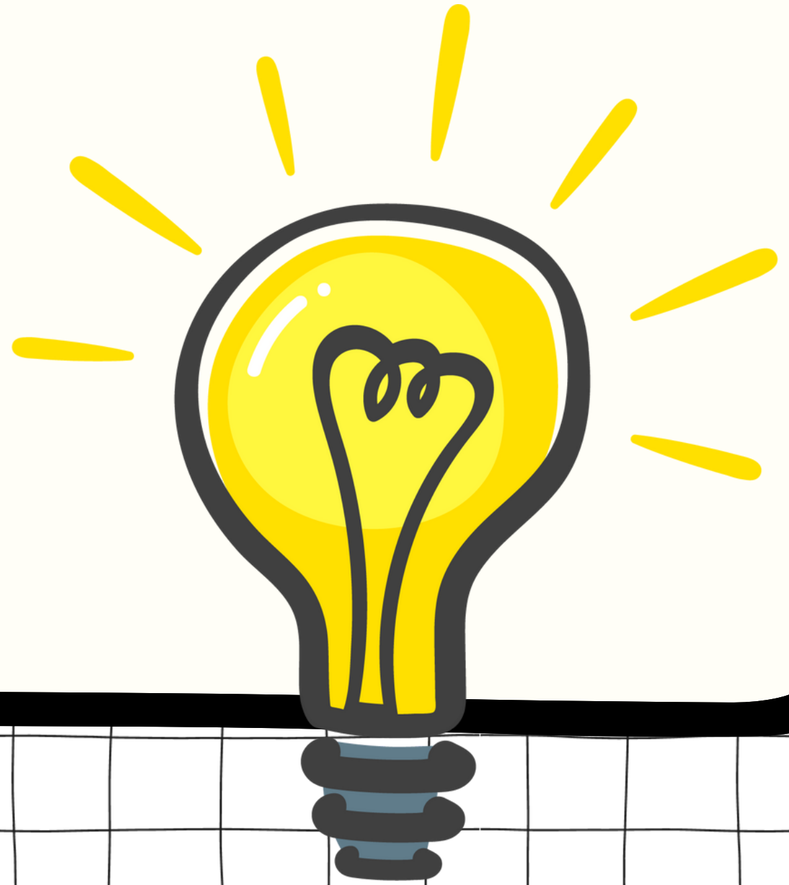
INSIGHTS

Still need to improve accuracy: trying newer models

Add new features from other data sets such as flight duration and number of on board passengers.

we could have further improved our project by also designing a model to predict the delay time of a flight.

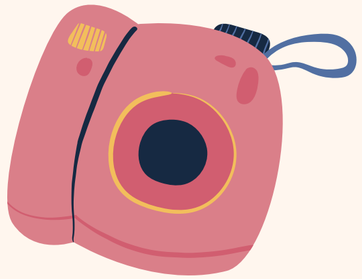
Thank You!



Presented by: Mustafa, Sai, Yifei

TRIP ESSENTIALS

Passport



Camera

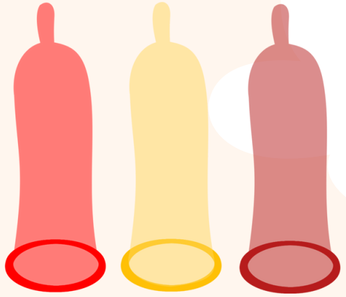
Headphones

Tickets

Luggage

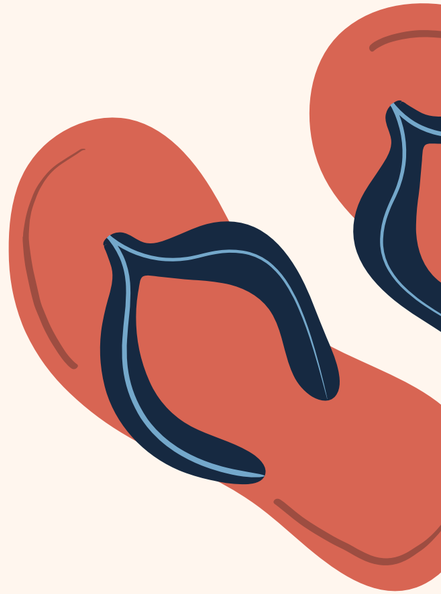
Clothes

Phone



Condoms

BOOK YOUR NEXT VACATION WITH US!



CONTACT US

@REALLYGREATSITE

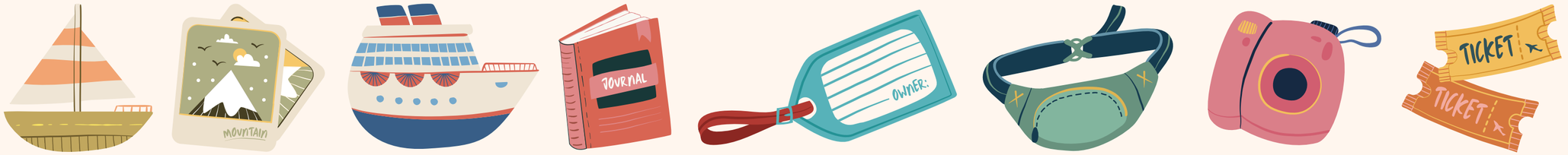
123-456-7890

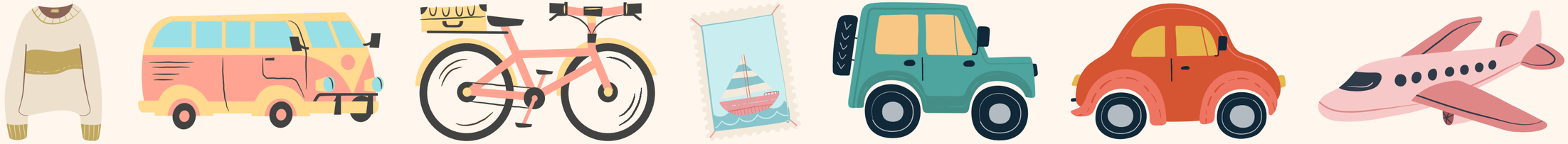
HELLO@REALLYGREATSITE.COM

WWW.REALLYGREATSITE.COM

123 ANYWHERE ST., ANY CITY

FREE RESOURCE PAGE





FREE RESOURCE PAGE