Flight Status Prediction Visualization

By,
Hansa Pradhan

Instructor,
Dr Reza Jafari

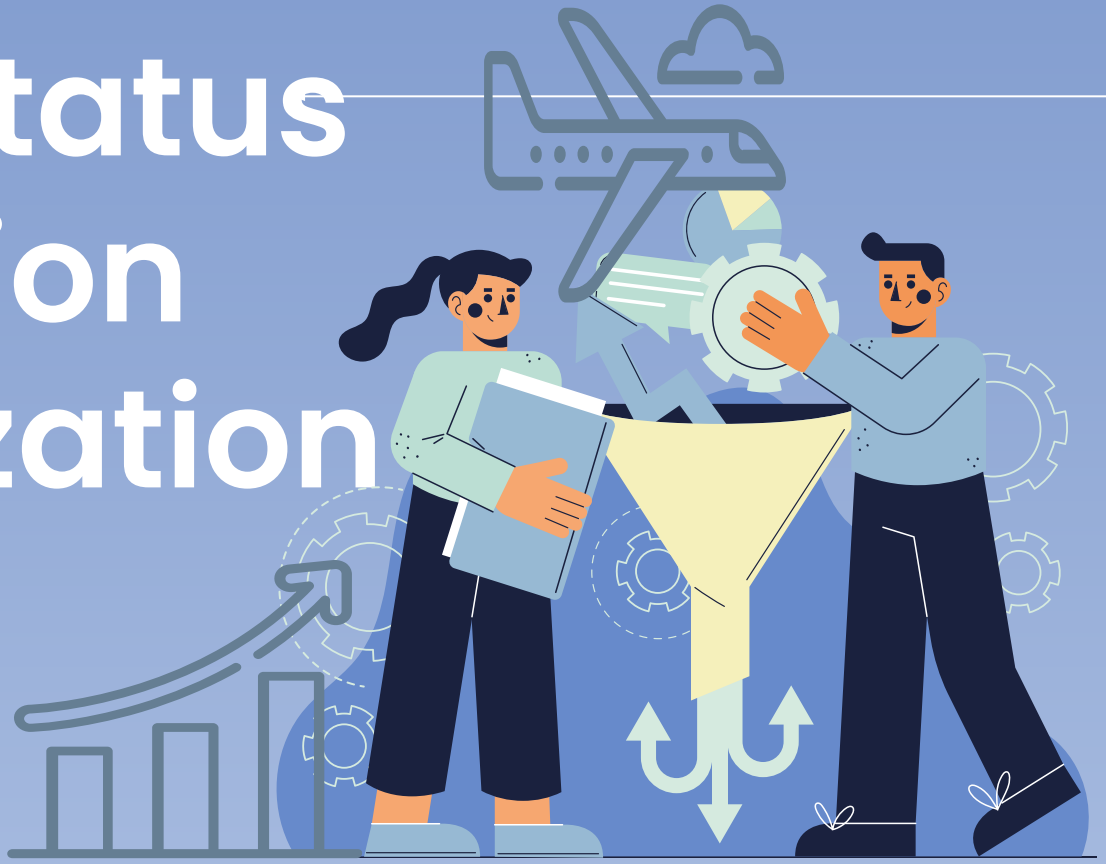


Table of *contents*

01

What *is it?*

Describing the dataset

02

Statistical Analysis

All the statistical analysis done on dataset

03

Visualizations

The plots I created for the term project

04

Lesson Learnt

Lessons from doing the project

01

What *is it?*

Lets learn about the dataset

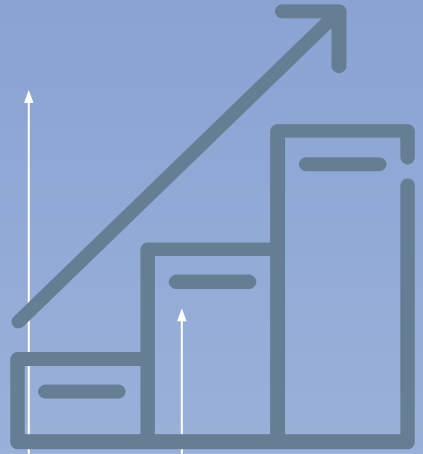
Flight Status Prediction Dataset

- Its a dataset hosted on Kaggle.
- The dataset source : TranStats data,
- Time period (January 2018 -2022)
- The file formats available (CSV and Parquet).
- 120 columns, 5 csv(approx 3M rows)

Why *data Visualization?*

- Clarity and Understanding
- Storytelling
- Decision Support
- Impactful Communication

Data Exploration



- Fetching data

```
flight_data.shape
```

```
(429036, 61)
```

- Handling Missing Values
  - Fillna and dropna
- Data Types Conversion
  - Datetime conversion, 0,1 to True false

Columns...

```
flight_data.columns
```

```
Index(['FlightDate', 'Airline', 'Origin', 'Dest', 'Cancelled', 'Diverted',
       'CRSDepTime', 'DepTime', 'DepDelayMinutes', 'DepDelay', 'ArrTime',
       'ArrDelayMinutes', 'AirTime', 'CRSElapsedTime', 'ActualElapsedTime',
       'Distance', 'Year', 'Quarter', 'Month', 'DayofMonth', 'DayOfWeek',
       'Marketing_Airline_Network', 'Operated_or_Branded_Code_Share_Partners',
       'DOT_ID_Marketing_Airline', 'IATA_Code_Marketing_Airline',
       'Flight_Number_Marketing_Airline', 'Operating_Airline',
       'DOT_ID_Operating_Airline', 'IATA_Code_Operating_Airline',
       'Tail_Number', 'Flight_Number_Operating_Airline', 'OriginAirportID',
       'OriginAirportSeqID', 'OriginCityMarketID', 'OriginCityName',
       'OriginState', 'OriginStateFips', 'OriginStateName', 'OriginWac',
       'DestAirportID', 'DestAirportSeqID', 'DestCityMarketID', 'DestCityName',
       'DestState', 'DestStateFips', 'DestStateName', 'DestWac', 'DepDel15',
       'DepartureDelayGroups', 'DepTimeBlk', 'TaxiOut', 'WheelsOff',
       'WheelsOn', 'TaxiIn', 'CRSArrTime', 'ArrDelay', 'ArrDel15',
       'ArrivalDelayGroups', 'ArrTimeBlk', 'DistanceGroup',
       'DivAirportLandings'],
      dtype='object')
```

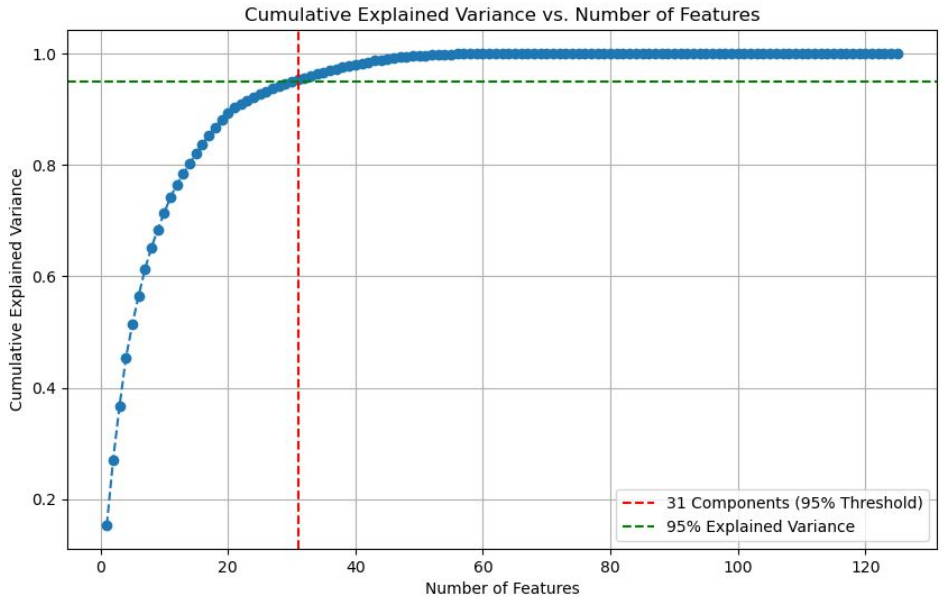
Statistical Analysis

- Normalization Test:
  Ks_test, shapiro test daugustino test.
  Implemented in dash

- Created the plots to show the normalization: Used Boxcox transformation to convert it to gaussian distribution

- Noticed most columns were right skewed

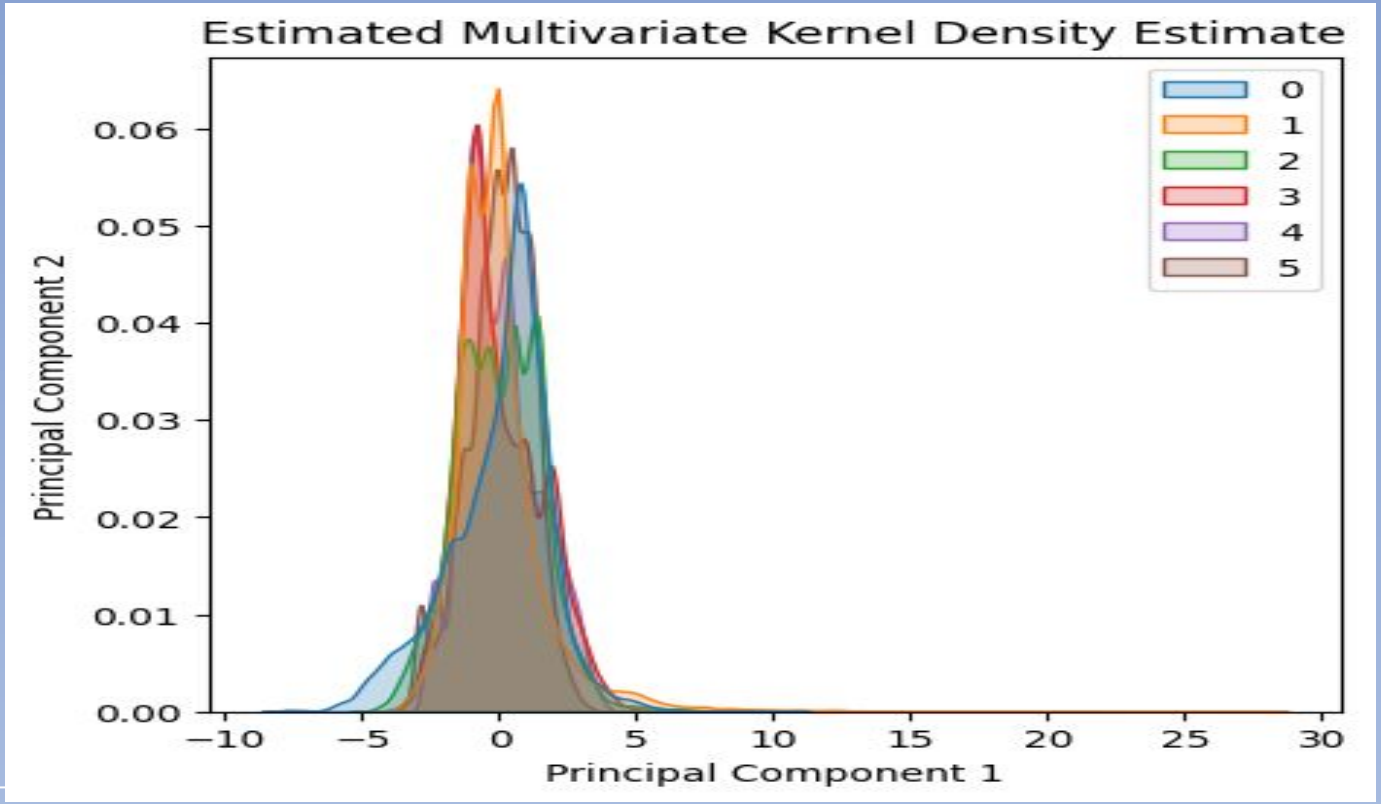
PCA



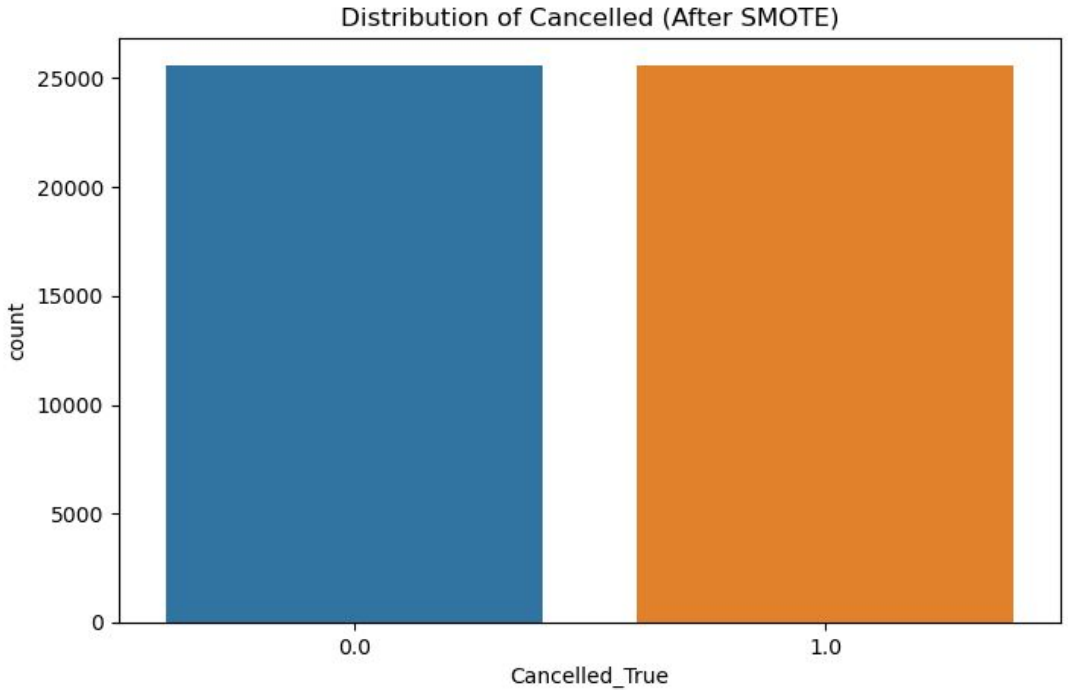
```
Number of features needed for 95% explained variance: 31
```

```
Condition Number: 2.1534166360430776e+16
```

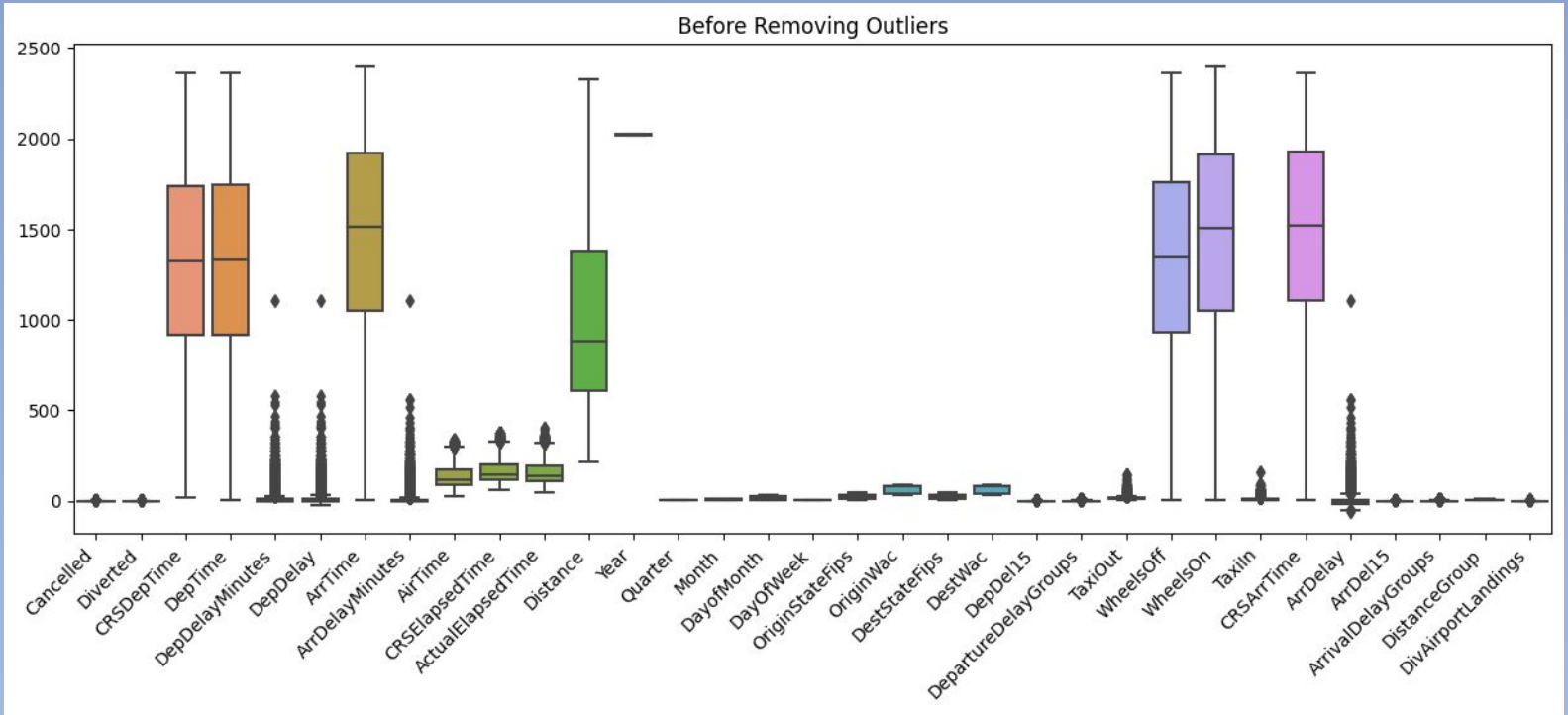
Multivariate KDE



SMOTE


Distribution of Cancelled (After SMOTE)
25000
20000
15000
10000
5000
0
count
0.0
1.0
Cancelled_True

Outlier Analysis

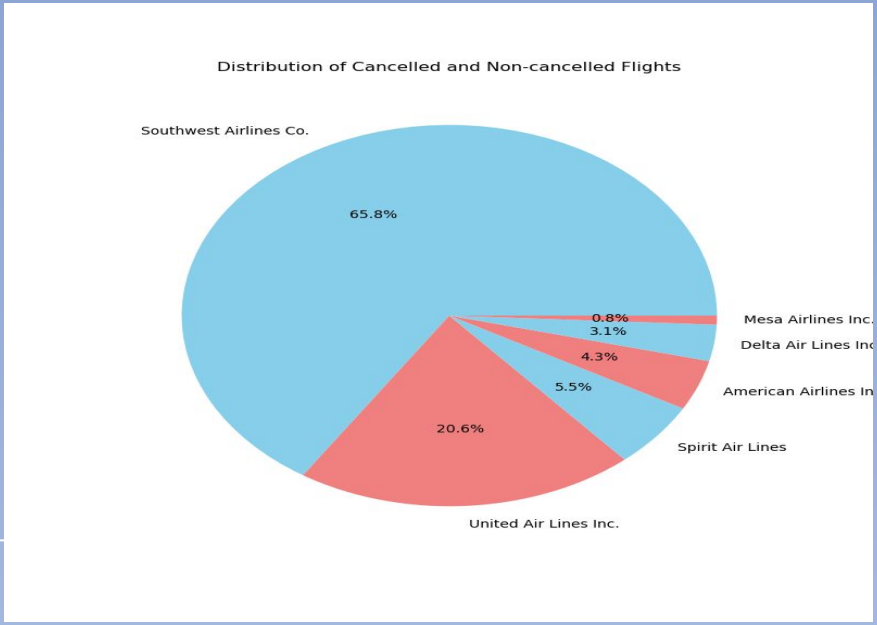


02

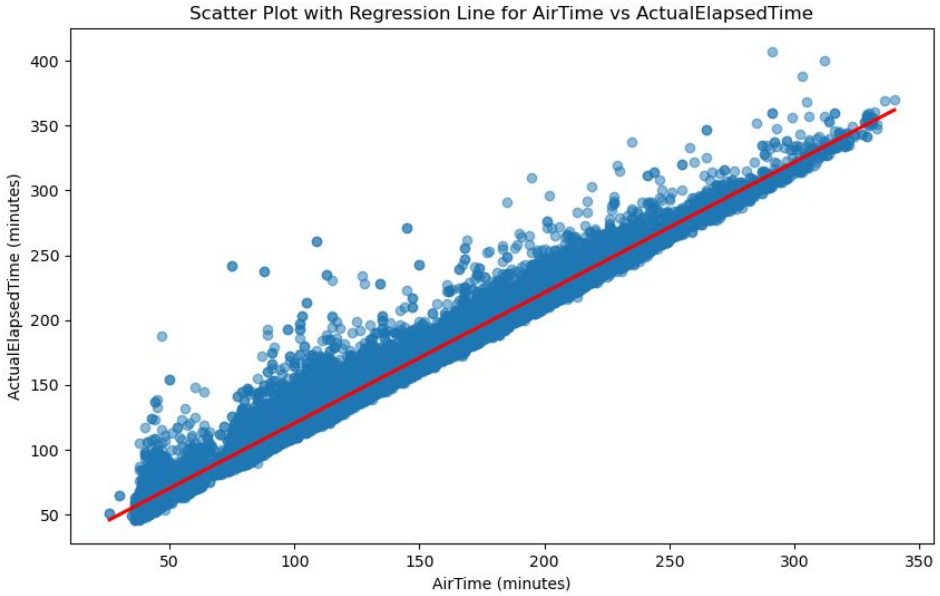
Project *Visualization*

Lets see the plots generated....

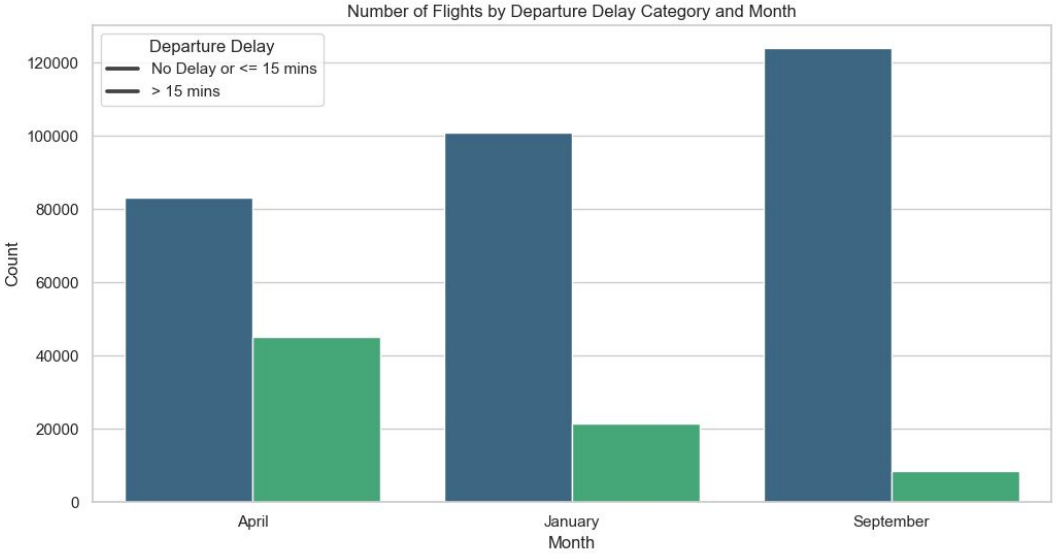
Pie Plot...


Distribution of Cancelled and Non-cancelled Flights
Southwest Airlines Co.
65.8%
0.8%
Mesa Airlines Inc.
3.1%
Delta Air Lines Inc
4.3%
American Airlines In
5.5%
Spirit Air Lines
20.6%
United Air Lines Inc.

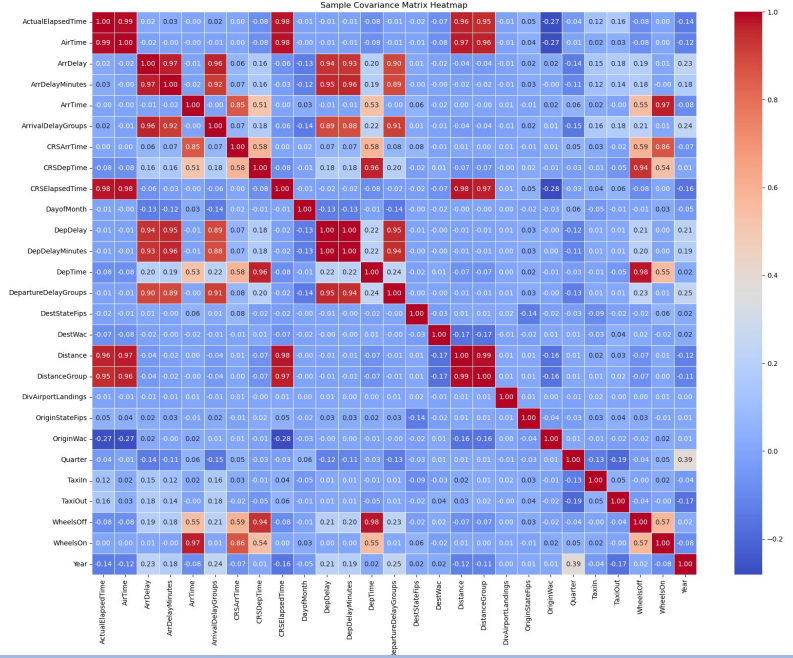
Scatter Plot...



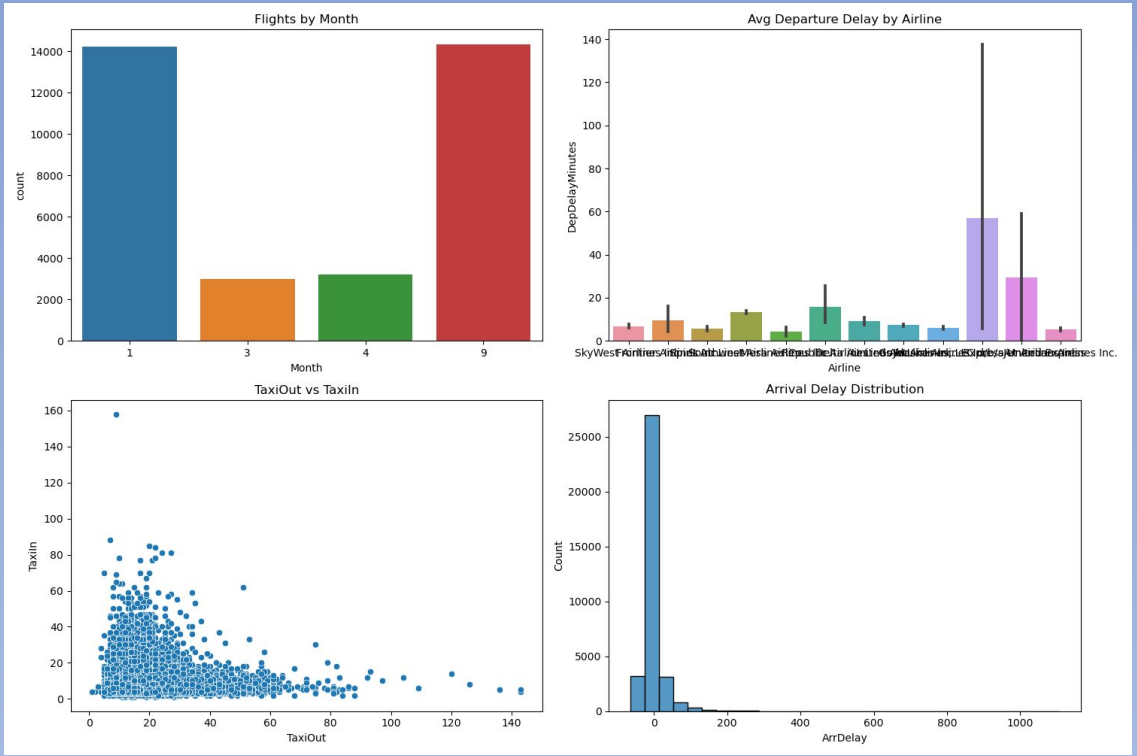
Barplot...


Number of Flights by Departure Delay Category and Month
Departure Delay
No Delay or <= 15 mins
> 15 mins
120000
100000
80000
60000
40000
20000
0
Count
April
January
September
Month

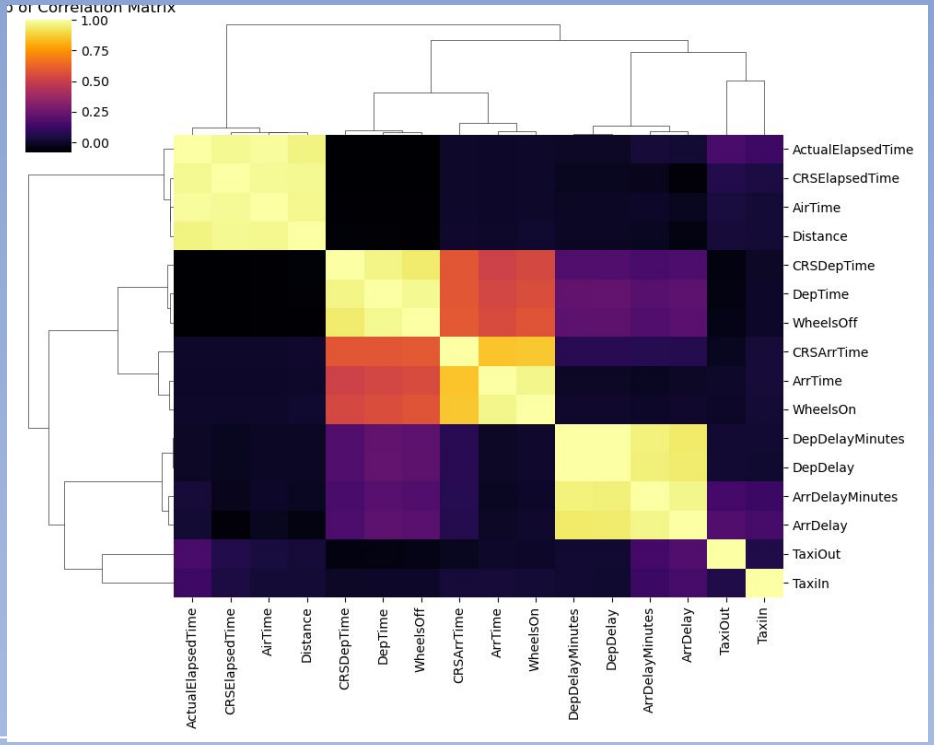
Heatmap..


Sample Covariance Matrix Heatmap
ActualElapsedTime
AirTime
ArrDelay
ArrDelayMinutes
ArrTime
ArrivalDelayGroups
CRSArrTime
CRSDepTime
CRSElapsedTime
DayofMonth
DepDelay
DepDelayMinutes
DepTime
DepartureDelayGroups
DestStateFips
DestWac
Distance
DistanceGroup
DivAirportLandings
OriginStateFips
OriginWac
Quarter
TaxiIn
TaxiOut
WheelsOff
WheelsOn
Year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

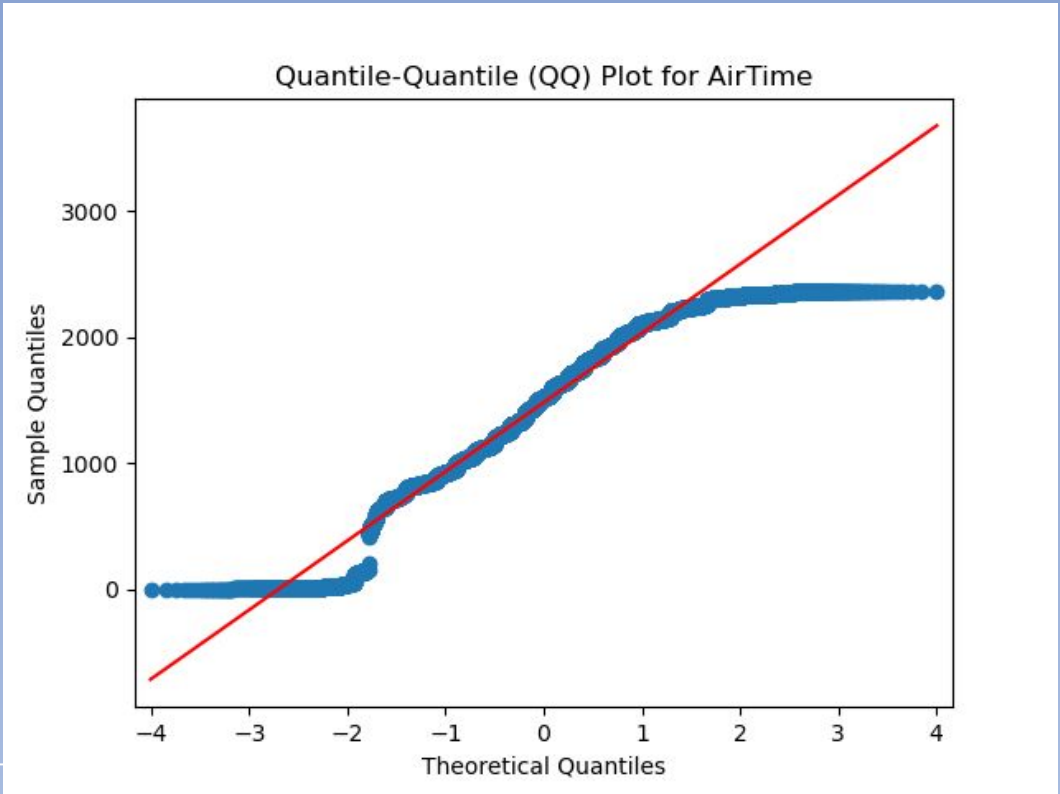
Subplots..


Flights by Month
count
Month
Avg Departure Delay by Airline
DepDelayMinutes
Airline
TaxiOut vs TaxiIn
TaxiIn
TaxiOut
Arrival Delay Distribution
Count
ArrDelay

ClusterMap..


1.00
0.75
0.50
0.25
0.00
ActualElapsedTime
CRSElapsedTime
AirTime
Distance
CRSDepTime
DepTime
WheelsOff
CRSArrTime
ArrTime
WheelsOn
DepDelayMinutes
DepDelay
ArrDelayMinutes
ArrDelay
TaxiOut
TaxiIn

QQ Plot..


Quantile-Quantile (QQ) Plot for AirTime
Sample Quantiles
3000
2000
1000
0
−4
−3
−2
−1
0
1
2
3
4
Theoretical Quantiles

Dash *Demo*

URL:

https://dashapp-jadd4w76ha-nn.a.run.app/

Lessons Learnt

- Data Consistency Matters
- Robust Data Preprocessing
- Error Handling is Crucial
- Impactful Communication

Thanks!

Do you have any questions?

hansapradhan@vt.edu