CAPSTONE PROJECT FINAL REPORT – GROUP 5

ALY6140: Analytics Systems Technology

Professor: Zhi Richard He

By

Jalajakshi Lali, Praharsha Pallempati, Vinodh Cherukupalli

Date: 18th August 2022

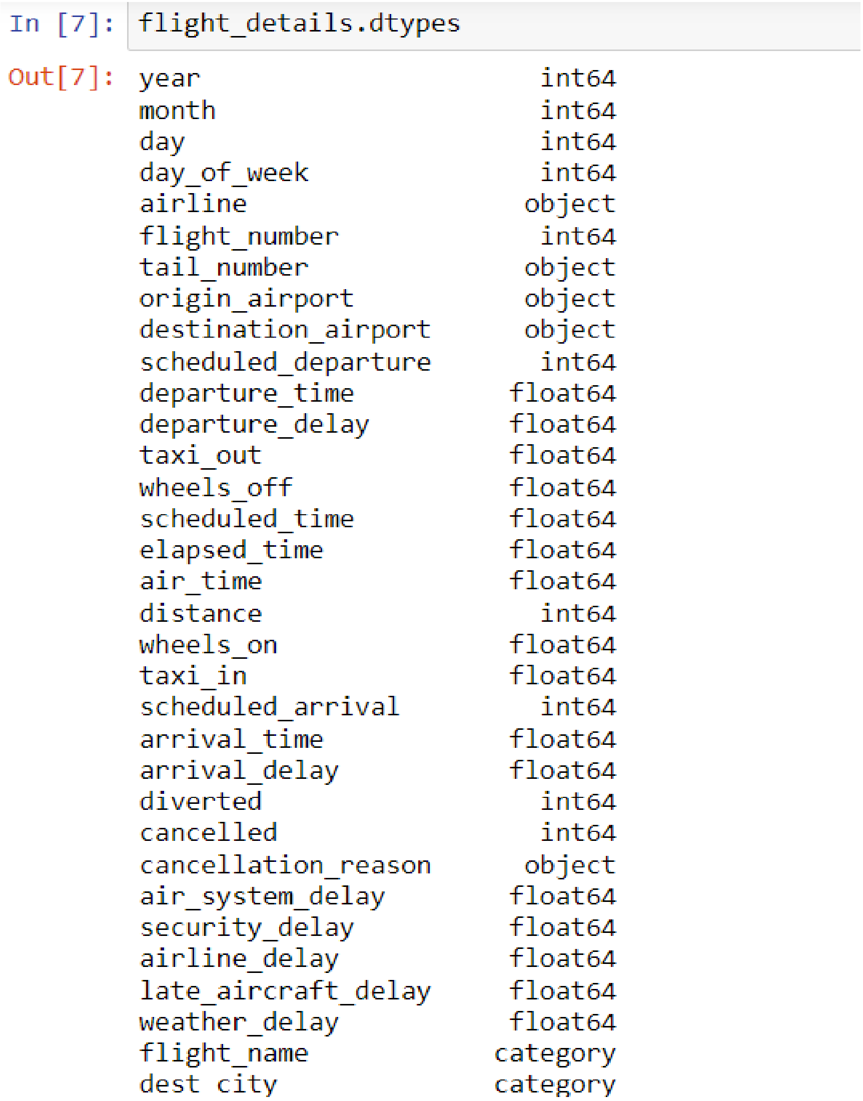
INTRODUCTION

Air Travel Consumer Report, produced by the DOT once a month, normally includes a summary of the number of on-time, delayed, cancelled, and diverted flights. BTS began its investigation into the causes of flight delays in June 2003. Through BTS, the public can access both the raw data and the summarized statistics. It's officially summer vacation time, and as usual, there will be a slew of airline cancellations and delays. Cancellations, which messes up everyone's travel arrangements and tests people's tolerance levels to the breaking point. To give just one example, this year's Memorial Day weekend saw the cancellation of over 2,000 flights by U.S. airlines (Jacobson, S. H., 2022). Reasons such as bad weather, heavy air traffic, insufficient crew members, etc., can cause flight cancellations and delays. We chose this data set to investigate the causes of flight delays and develop methods for making accurate, timely predictions that may be used to aid passengers. The "2015 Flight Delays and Cancellations" dataset from Kaggle was selected for our final assignment. The U.S. government keeps tabs on the percentage of domestic flights that depart and arrive on time, regardless of whose airline oversees the flight. The BTS is the statistics arm of the DOT (Department of Transport) (BTS). You can use this data set to analyze and forecast flight delays and cancellations, as well as learn more about the causes of these problems. There are a total of 31 records in the dataset. Details such as months, days of the week, scheduled arrivals, cancellation reasons, air system delays, security delays, etc. in separate columns or variables. About 582,000 records have missing data for various variables. Tail number, departure time, departure delay, arrival delay, etc. are all examples of variables for which we lack data. Consequently, the information must receive a cleaning.

This capstone aims to answer business issues and make predictions by using EDA on the raw dataset to provide patterns and insights. The project's overarching goal is to determine the best time of day, weekday, and season to fly to reduce the likelihood of delays. How reliably does the weather report predict flight delays? Which airlines provide the least amount of downtime due to delays and cancellations? Which of the many possible factors have a high degree of association? Which variables can be used to anticipate outcomes, and which can be utilized to make predictions? Three models would be developed to make forecasts regarding delays and cancellations. We'll construct a decision tree classifier model and a Random Forest regressor to examine flight delay trends. Logistic regression would be constructed to make cancellation forecasts.

EXPLORATORY DATA ANALYSIS:

One of the goals of this research is to develop a system that can foresee flight disruptions. Sorting the dependent and independent variables allows us to better understand what's causing the flight delays. Before diving into analysis, we need to extract the data and bring it into our Python environment. The flights.csv, airports.csv, and airlines.csv data were imported into pandas to facilitate this. After the data was imported, we identified the required columns for the analysis and deleted the rest. To remove the missing data, the averages are substituted. In order to employ it in the creation of prediction models, the mean is substituted for a large number of null values in the data, such as those found in the arrival and departure delays. The column names are transformed lower case and the data from all the three csv files are mapped with one another based on the common columns. After the data was mapped, four distinct datatypes emerged. Float64, int64, Object, and category.



```
In [7]:  flight_details.dtypes

Out[7]:  year                          int64
         month                         int64
         day                           int64
         day_of_week                   int64
         airline                      object
         flight_number                 int64
         tail_number                  object
         origin_airport               object
         destination_airport          object
         scheduled_departure           int64
         departure_time              float64
         departure_delay             float64
         taxi_out                    float64
         wheels_off                  float64
         scheduled_time              float64
         elapsed_time                float64
         air_time                    float64
         distance                      int64
         wheels_on                   float64
         taxi_in                     float64
         scheduled_arrival             int64
         arrival_time                float64
         arrival_delay               float64
         diverted                      int64
         cancelled                     int64
         cancellation_reason          object
         air_system_delay            float64
         security_delay              float64
         airline_delay               float64
         late_aircraft_delay         float64
         weather_delay               float64
         flight_name                category
         dest city                  category
```

For the convenience of analysis, We have eliminated several columns that were not all that necessary for our EDA and the development of future models. The columns include 'year','flight_number','airline','distance','tail_number','taxi_out','scheduled_time','departure_time','wheels_off','elapsed_time','air_time','wheels_on','day_of_week','taxi_in','origin_airport', 'destination_airport','flight_name','dest_city','orig_city','orig_state','dest_state','orig_country',' dest_country','orig_lat','orig_lon','dest_lat','dest_lon'

We have laid the foundation for our prediction model from the exploratory data analysis of this project.

CANCELLATIONS

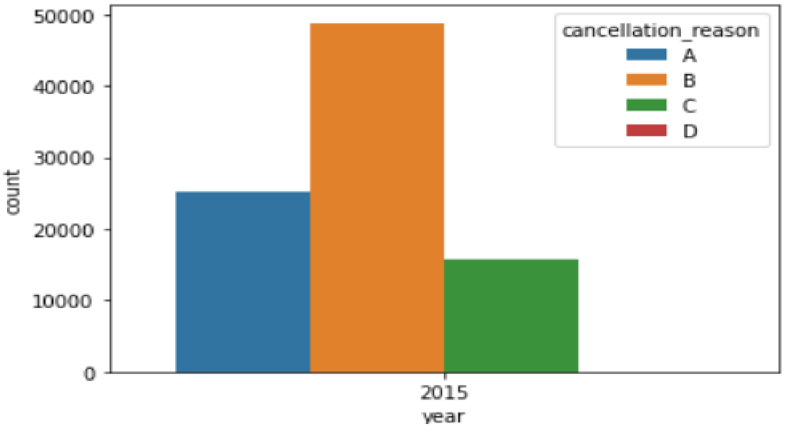
On observations, the cancellations of flights are happening due to various reasons in a year.

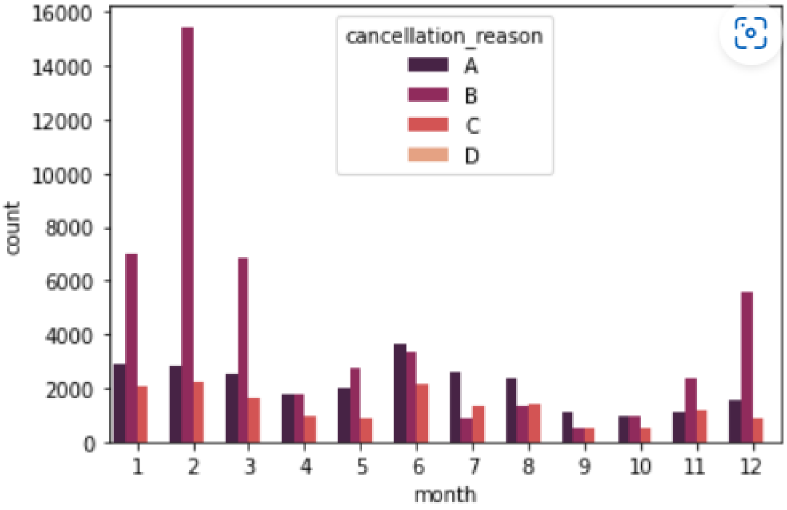
Cancellation Reason A represents Airline/Carrier

Cancellation Reason B represents Weather

Cancellation Reason C represents National Air System

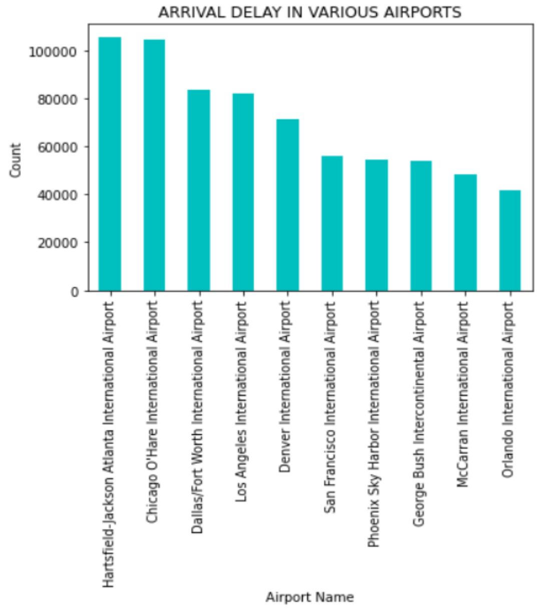
Cancellation Reason D represents Security

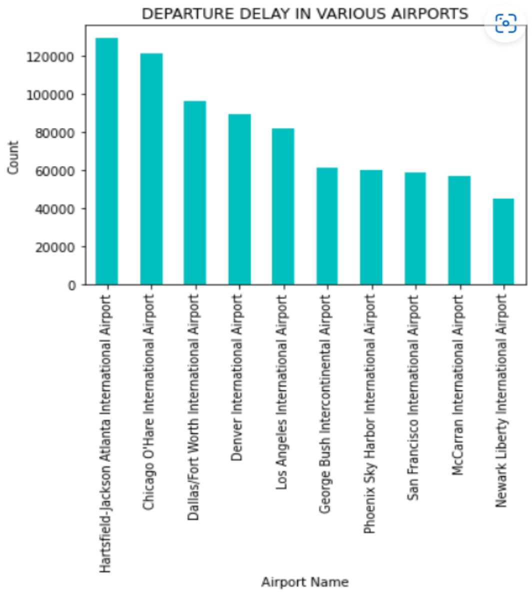




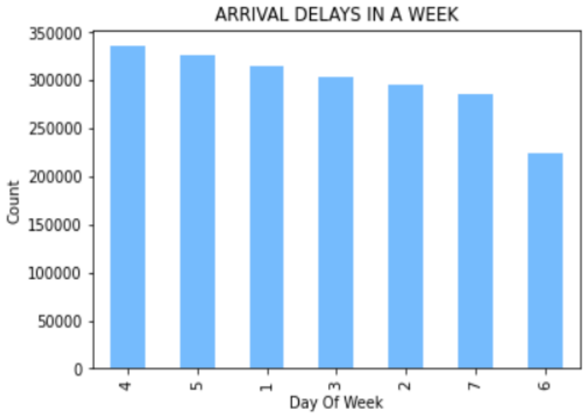
From these graphs we can observe that most of the cancellations are happening due to unsupported weather conditions. There is spike in cancellations in the months of December, January, February and March as these months fall under Winter season. Hence, from this we can understand that the cancellations of flights are more in Winter than any other.

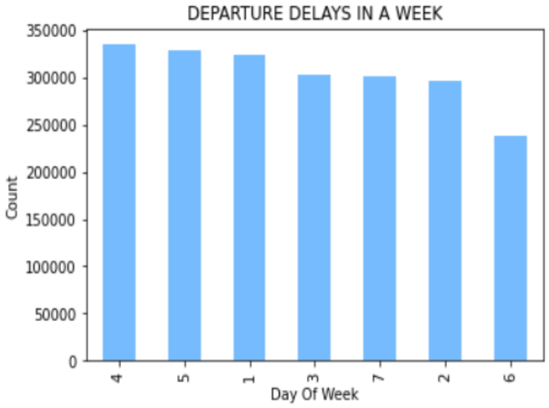
ARRIVAL AND DEPARTURE DELAYS



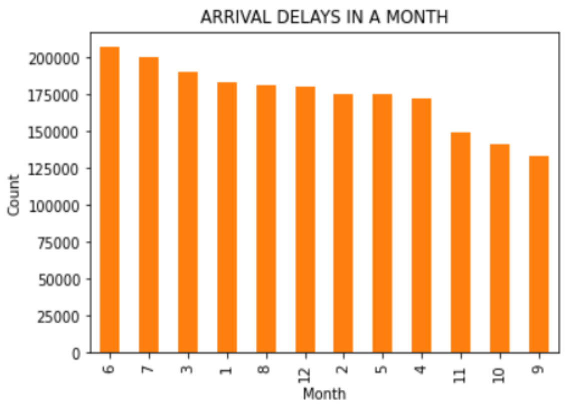


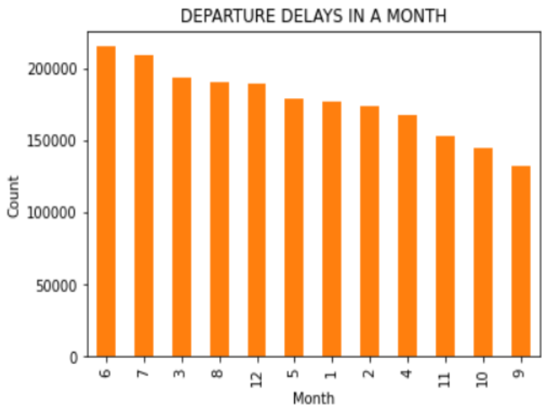
The above charts show that Atlanta, Chicago, and Dallas airports experience the most delays in arrivals and departures. These three airports are often recognized as the busiest in the United States. Delays in getting into an airport are shown to correspond with those in getting out.



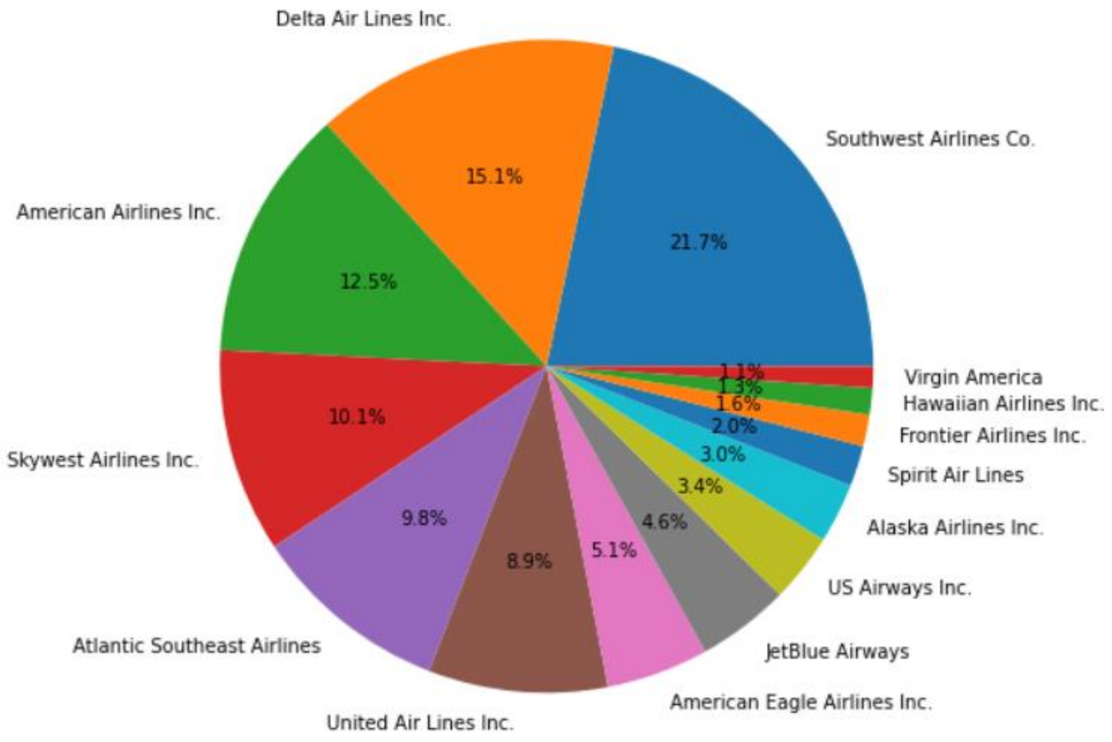


We note that delays within a week tend to occur on the fourth and fifth days or Wednesday and Thursday. There are typically more flights on Wednesdays and Thursdays than on any other days of the week. Reason being, avoiding peak demand and prices by taking off on a Wednesday or Thursday makes air travel more affordable.



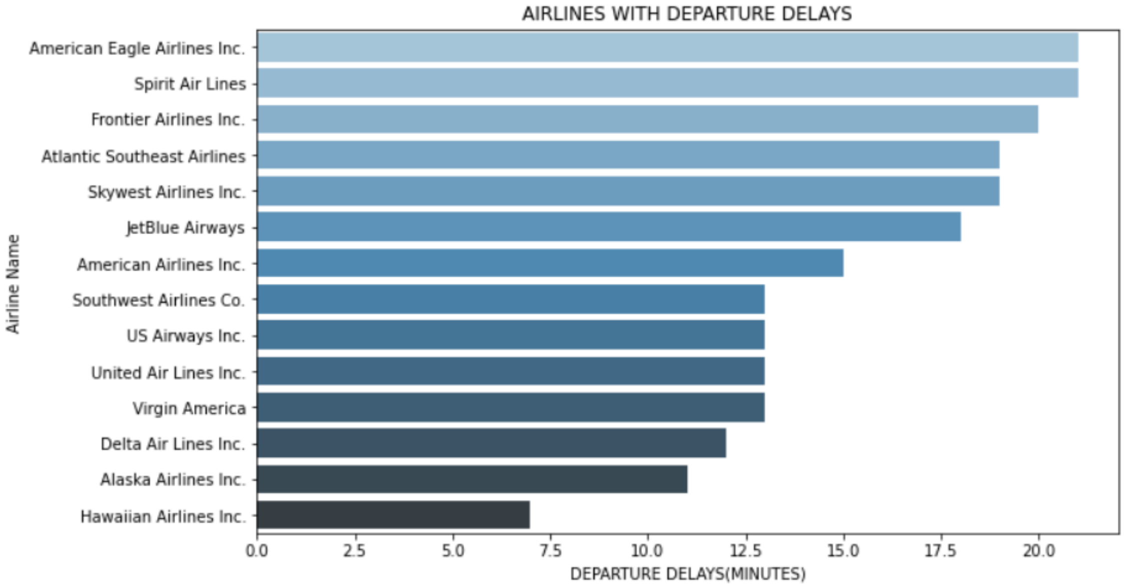


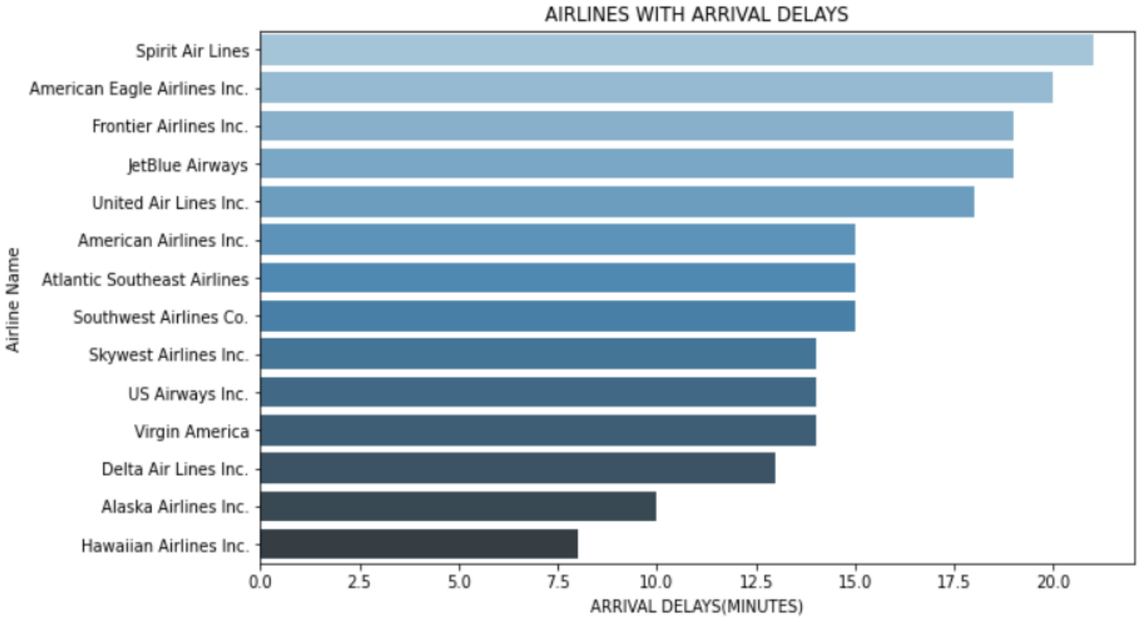
When looking at the number of delays inside a single month, we can see that June, July, and August have the most delays.



The above graphs depict the percentage of flights travelled with respect to airline company as of 2015. It is observed that Southwest Airlines, Delta Airlines, American Airlines are the top 3 airline companies that have highest flights scheduled.

Further, we will be observing the scheduled flight delays that are occurring with respect to each airline company,

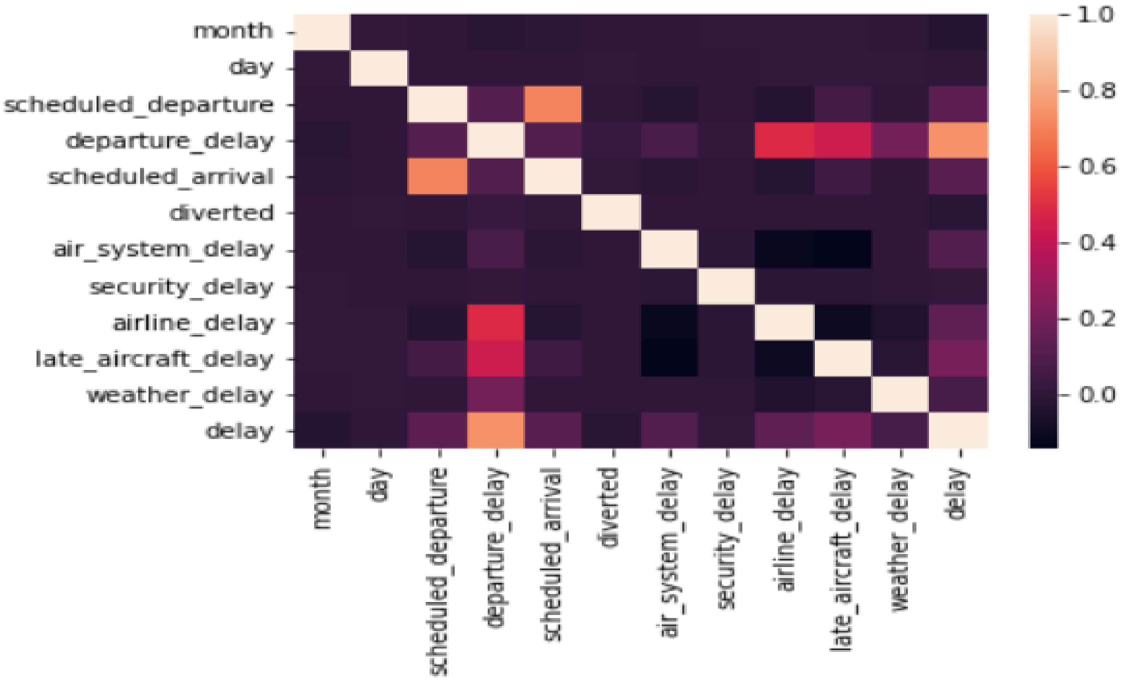




The American Eagle Airlines, Spirit Airlines, Frontier Airlines are the most delayed airlines when compared to others. Though they hold very less percentage of flights scheduled i.e., 5.1%, 2%,1.6% respectively the flights are getting delayed.

Among the top 3 airlines companies, The Delta Airlines have very less delays when compared with Southwest and American Airlines.

From the filtered columns we have plotted a heat map to find the correlation between the variables.



PREDICTION MODELS

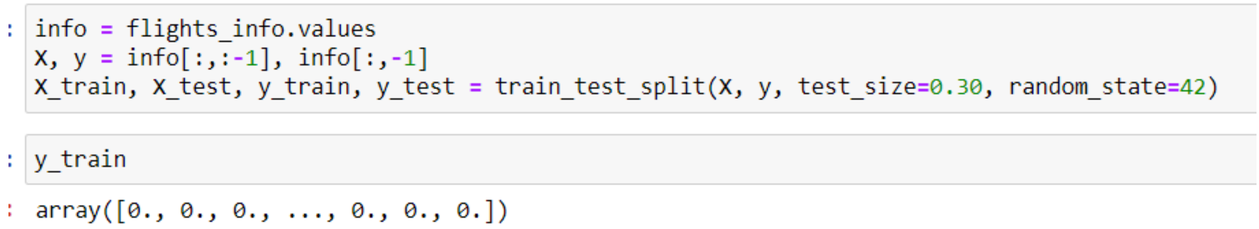
To predict the delay time and the cancellation patterns for the test data from train data we used 3 different predictive models, they are:

1. Decision Tree Classifier Model – Arrival Delays
2. Logistic Regression Model - Cancellations
3. Random Forest Regressor Model – Departure Delays

DECISION TREE CLASSIFIER MODEL:

A parameter-free approach to unsupervised learning that can be utilized for classification and regression. It is employed to build a model that can predict the value of a target variable by the identification and application of decision rules inferred from the attributes of the data. By doing this, we ensure that the model properly represents the data.

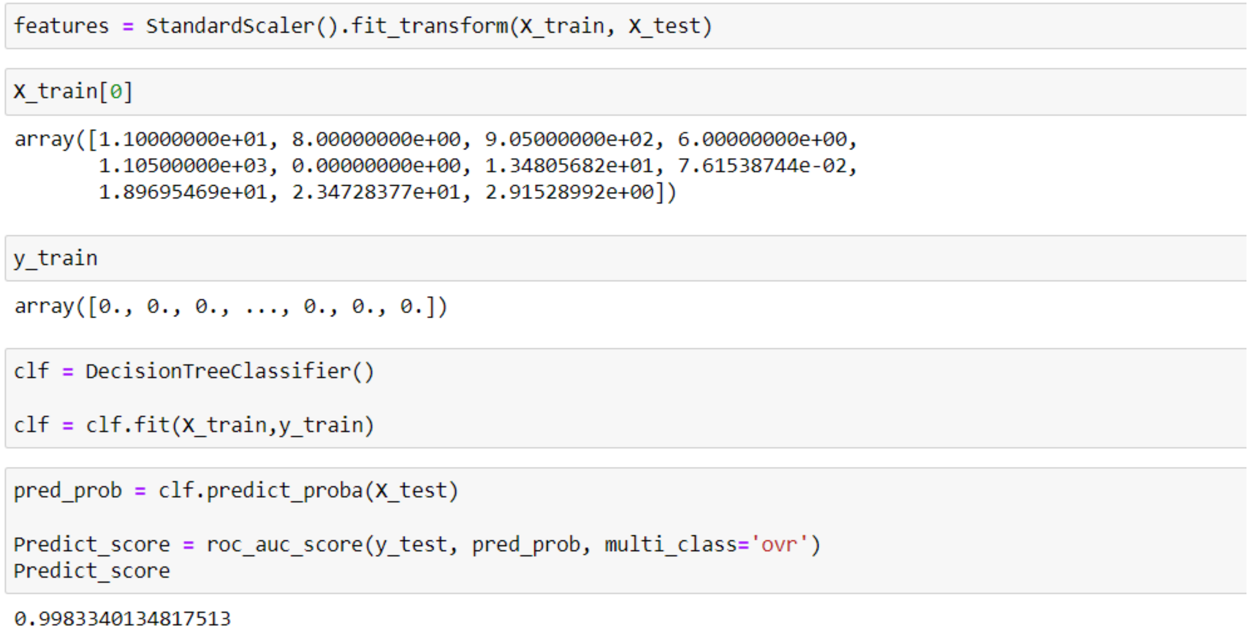
To better understand the causes of airline delays, such as weather, air system, and security issues, we will develop a decision tree classifier model to apply to our dataset.



```
info = flights_info.values
X, y = info[:,:-1], info[:,-1]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
```

```
y_train
```

```
array([0., 0., 0., ..., 0., 0., 0.])
```



```
features = StandardScaler().fit_transform(X_train, X_test)
```

```
X_train[0]
```

```
array([1.10000000e+01, 8.00000000e+00, 9.05000000e+02, 6.00000000e+00,
       1.10500000e+03, 0.00000000e+00, 1.34805682e+01, 7.61538744e-02,
       1.89695469e+01, 2.34728377e+01, 2.91528992e+00])
```

```
y_train
```

```
array([0., 0., 0., ..., 0., 0., 0.])
```

```
clf = DecisionTreeClassifier()

clf = clf.fit(X_train,y_train)
```

```
pred_prob = clf.predict_proba(X_test)

Predict_score = roc_auc_score(y_test, pred_prob, multi_class='ovr')
Predict_score
```

```
0.9983340134817513
```

We have separated the information into training data and test data using the train test split function. Both X train and X test data are formatted using the usual scalar function. DecisionTreeClassifier() is used to create a model, and then the model is fit to the X train and y train datasets. Using the predict proba function, we were able to determine that the model has a 99.8 percent success rate.

The model evaluation is performed using the RUC and AUC score which is 0.998 which is almost equivalent to 1. Results with an area under the ROC curve (AUC) between 0.9 and 1 were deemed to be very good.

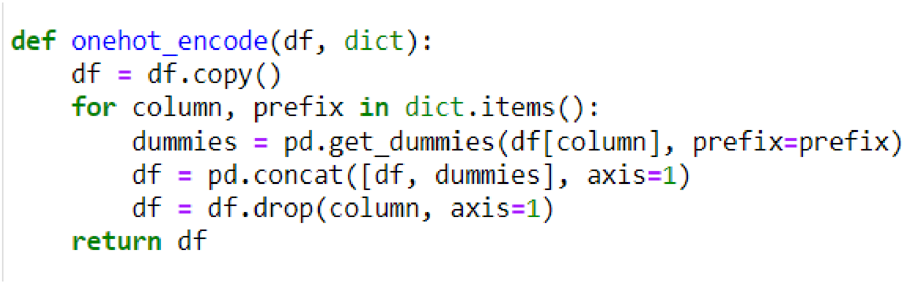
LOGISTIC REGRESSION MODEL

Logistic Regression is a useful tool for resolving classification problems grounded on probability theory. It is employed in cases where the dependent (or output) variable is of a categorical kind. The most often canceled flights between the origin and destination will be identified by splitting the data into test and training sets for logistic regression.

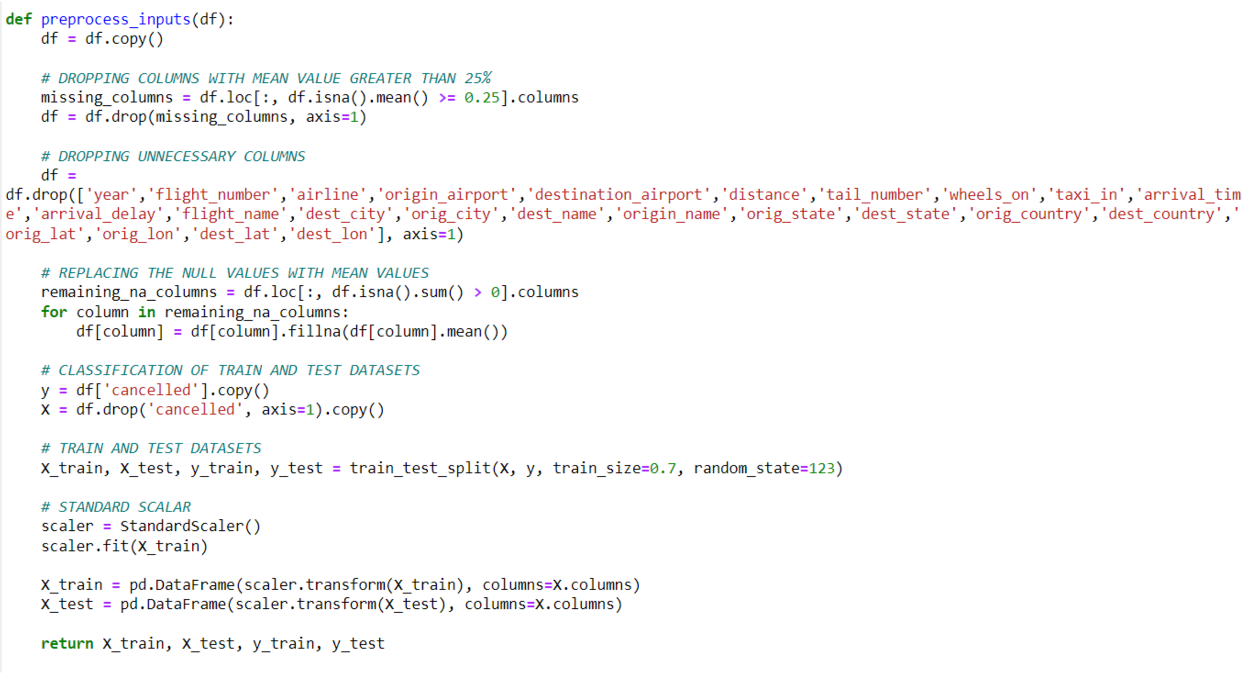
With the help of onehot encode and reprocess inputs, we were able to locate and replace any missing values in the inputs to this regression model. The preprocess inputs function is where the dataset is divided. We have separated the information into training data and test data using the train test split function. The X train and y train datasets are fitted using the LogisticRegression() algorithm.

We have built a function, evaluate model(), that figures out the test accuracy, shows a confusion matrix, and generates a classification report.

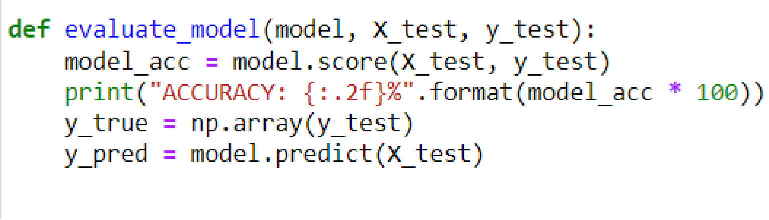
Py file:



```python
def onehot_encode(df, dict):
    df = df.copy()
    for column, prefix in dict.items():
        dummies = pd.get_dummies(df[column], prefix=prefix)
        df = pd.concat([df, dummies], axis=1)
        df = df.drop(column, axis=1)
    return df
```

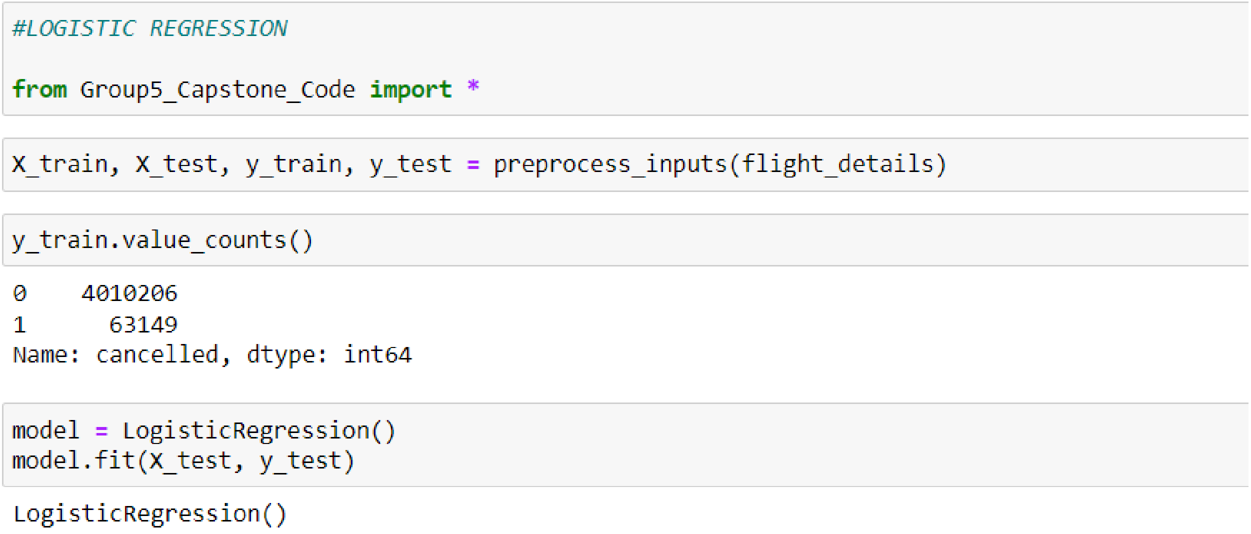


```python
def preprocess_inputs(df):
    df = df.copy()

    # DROPPING COLUMNS WITH MEAN VALUE GREATER THAN 25%
    missing_columns = df.loc[:, df.isna().mean() >= 0.25].columns
    df = df.drop(missing_columns, axis=1)

    # DROPPING UNNECESSARY COLUMNS
    df =
df.drop(['year','flight_number','airline','origin_airport','destination_airport','distance','tail_number','wheels_on','taxi_in','arrival_tim
e','arrival_delay','flight_name','dest_city','orig_city','dest_name','origin_name','orig_state','dest_state','orig_country','dest_country','
orig_lat','orig_lon','dest_lat','dest_lon'], axis=1)

    # REPLACING THE NULL VALUES WITH MEAN VALUES
    remaining_na_columns = df.loc[:, df.isna().sum() > 0].columns
    for column in remaining_na_columns:
        df[column] = df[column].fillna(df[column].mean())

    # CLASSIFICATION OF TRAIN AND TEST DATASETS
    y = df['cancelled'].copy()
    X = df.drop('cancelled', axis=1).copy()

    # TRAIN AND TEST DATASETS
    X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, random_state=123)

    # STANDARD SCALAR
    scaler = StandardScaler()
    scaler.fit(X_train)

    X_train = pd.DataFrame(scaler.transform(X_train), columns=X.columns)
    X_test = pd.DataFrame(scaler.transform(X_test), columns=X.columns)

    return X_train, X_test, y_train, y_test
```



```python
def evaluate_model(model, X_test, y_test):
    model_acc = model.score(X_test, y_test)
    print("ACCURACY: {:.2f}%".format(model_acc * 100))
    y_true = np.array(y_test)
    y_pred = model.predict(X_test)
```

IPYNB File:



```
#LOGISTIC REGRESSION

from Group5_Capstone_Code import *
```

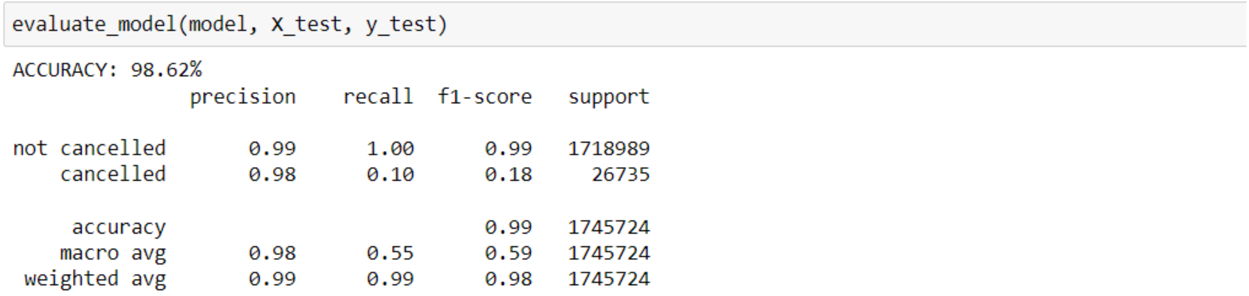
```
X_train, X_test, y_train, y_test = preprocess_inputs(flight_details)
```

```
y_train.value_counts()
```

```
0    4010206
1      63149
Name: cancelled, dtype: int64
```

```
model = LogisticRegression()
model.fit(X_test, y_test)
```

```
LogisticRegression()
```



```
evaluate_model(model, X_test, y_test)
```

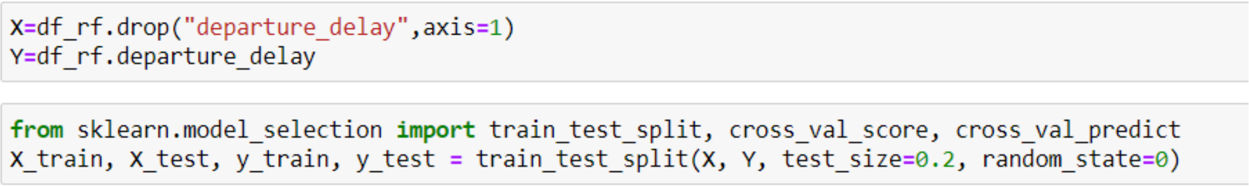
```
ACCURACY: 98.62%
              precision    recall  f1-score   support

not cancelled       0.99      1.00      0.99   1718989
    cancelled       0.98      0.10      0.18     26735

     accuracy                           0.99   1745724
    macro avg       0.98      0.55      0.59   1745724
 weighted avg       0.99      0.99      0.98   1745724
```

Thought the accuracy of this model built is 98.62% the recall value seems to be very less for the cancelled status values i.e., The model is not able to detect a specific category. Due to which we can conclude that this model is not a best fit, as we know Recall values play a major role in evaluating the model.

RANDOM FOREST REGRESSOR MODEL

Random forests employ averaging to build several decision tree classifiers on separate sub-samples of the dataset, which improves prediction accuracy and prevents over-fitting. Flight delay predictions at the monthly, annual, and daily levels will be made using Random Forest Regressor.

We will use the train test split function to separate the information into training and test sets. With the departure delay being the focus of our analysis, we will separate it out into its own column (Y), while the remaining columns will form X.

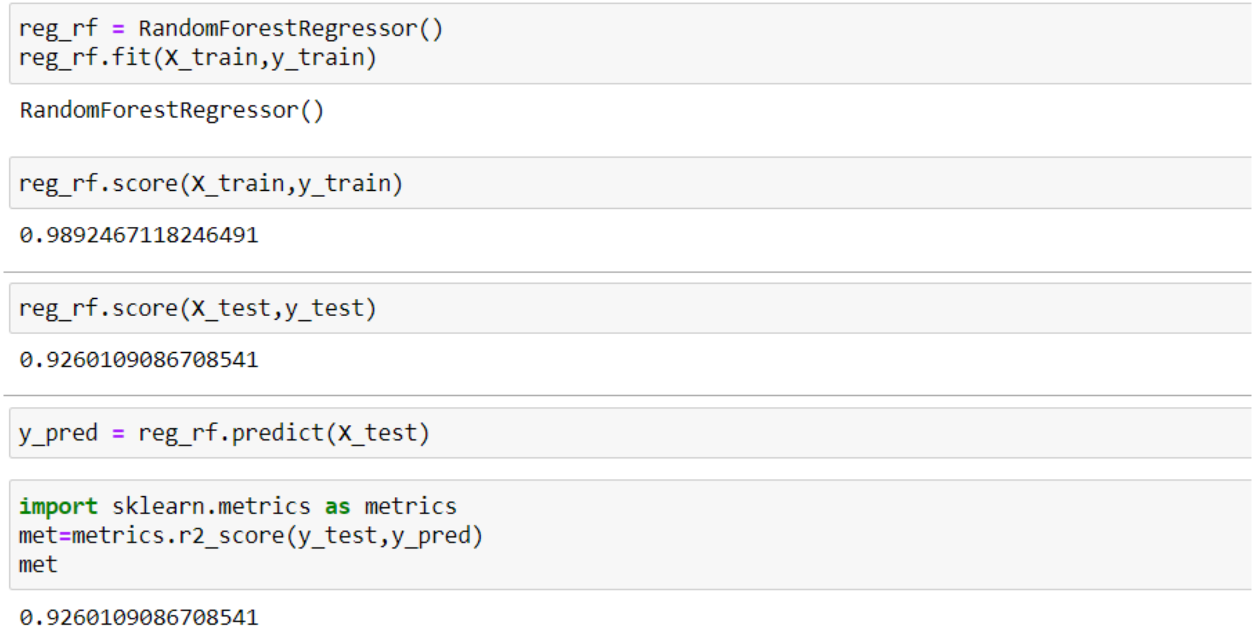


```
X=df_rf.drop("departure_delay",axis=1)
Y=df_rf.departure_delay
```

```
from sklearn.model_selection import train_test_split, cross_val_score, cross_val_predict
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=0)
```

Since the departure delays are predicted using this model we have bifurcated the train and test based on the departure delays with test size of 20% and train size of 80%.

On fitting the random forest model to the dataset, we obtained an accuracy score of 0.92 which is same for the auc roc score.



```
reg_rf = RandomForestRegressor()
reg_rf.fit(X_train,y_train)
```

```
RandomForestRegressor()
```

```
reg_rf.score(X_train,y_train)
```

```
0.9892467118246491
```

```
reg_rf.score(X_test,y_test)
```

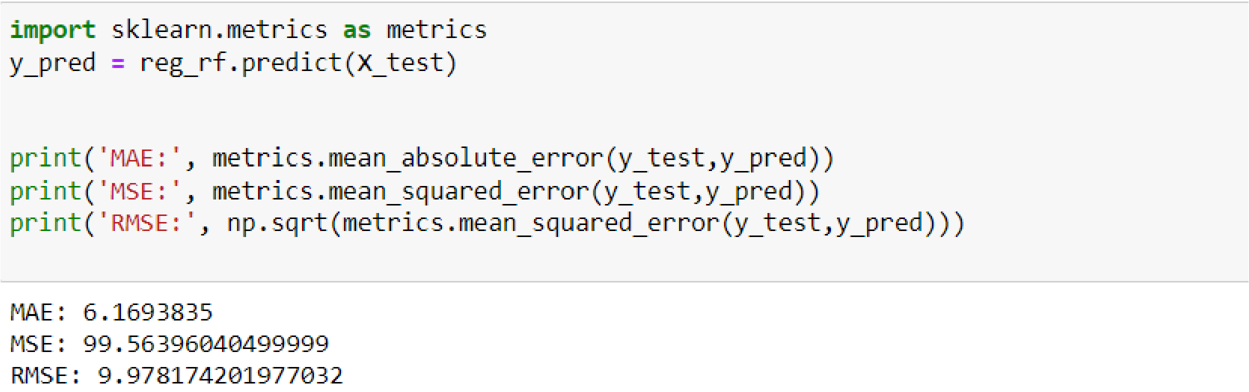
```
0.9260109086708541
```

```
y_pred = reg_rf.predict(X_test)
```

```
import sklearn.metrics as metrics
met=metrics.r2_score(y_test,y_pred)
met
```

```
0.9260109086708541
```

Using evaluation metric, we identified the below



```
import sklearn.metrics as metrics
y_pred = reg_rf.predict(X_test)


print('MAE:', metrics.mean_absolute_error(y_test,y_pred))
print('MSE:', metrics.mean_squared_error(y_test,y_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test,y_pred)))
```

```
MAE: 6.1693835
MSE: 99.56396040499999
RMSE: 9.978174201977032
```

For this model, since the R2 score is 0.92 and is at the acceptable limit we can say that the model is a good fit. Furthermore, we can improve the efficiency of the test on basis of RMSE. On tuning the data model further there is scope of improving the accuracy and evaluation metrics of this model.

CONCLUSION:

We first analyzed the dataset to determine which qualities were unevenly distributed and which numerical features were irrelevant to the modeling efforts. Based on our findings, the months of October, November, and December are often the least crowded and most reliable for air travel. As a rule, Fridays are the most convenient days of the week to take a trip. The weather is rarely responsible for flight delays.

We have also developed three models for making cancellation and delay forecasts. We have built a decision tree classifier model to analyze aircraft delays and identify patterns. We developed a logistic regression model to foresee dropouts. To determine the cause of the departure delay, we have constructed a third model using a random forest regressor, and then evaluated the models to determine which one yields the most accurate predictions.

MODEL	PARAMETER	ACCURACY	DECISION
DECISION TREE CLASSIFIER	ARRIVAL DELAYS	0.99	BEST FIT
LOGISTIC REGRESSION	CANCELLATION	0.98	NOT A BEST FIT
RANDOMFOREST REGRESSOR	DEPARTURE DELAYS	0.92	BEST FIT

REFERENCES:

- 2015 Flight Delays and Cancellations. (2017, February 9). Kaggle. 2015 Flight Delays and Cancellations | Kaggle

- Jacobson, S. H. (2022, July 11). Sheldon Jacobson: Frustrated by summer flight delays and cancellations? Here's why they're happening. Chicago Tribune. Reasons behind those frustrating flight delays, cancellations (chicagotribune.com)

- Shin, T. (2021, December 13). All Machine Learning Models Explained in 6 Minutes - Towards Data Science. Medium. All Machine Learning Models Explained in 6 Minutes | by Terence Shin | Towards Data Science

- Wikipedia contributors. (2022, June 23). List of the busiest airports in the United States. Wikipedia. List of the busiest airports in the United States - Wikipedia