Configuration Manual

MSc Research Project
Programme Name

Yash Rajesh Suryawanshi
x22227431

School of Computing
National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Yash Rajesh Suryawanshi		
Student ID:	x22227431		
Programme:	MSc in Data Analytics	**Year:**	2023-24
Module:	Research Project		
Lecturer:	Abdul Qayum		
Submission Due Date:	12th August 2024		
Project Title:	Leveraging Weather Data for Improved Flight Delay Prediction: A Comparative Analysis of Decision Trees and Random Forests		
Word Count:	368	**Page Count:**	6

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Yash Rajesh Suryawanshi

Date: 12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Yash Rajesh Suryawanshi
x22227431

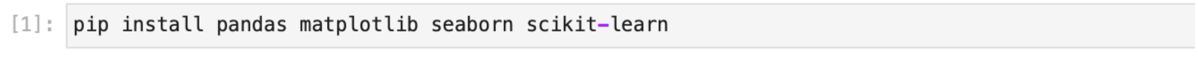
1 Introduction

A step-by-step guidelines is provided in this document to configure and run the Flight Delay Prediction project. The project uses two machine learning models that is Gradient bosting and Random forest to predict delays in flight using a publically available dataset containing various features related to flights.

2 System Requirements

- Python Version: 3.11 or above
- Libraries:
    - pandas
    - matplotlib
    - seaborn
    - scikit-learn

Ensure above mentioned libraries are installed. If they are not installed you can install them by using



```
[1]: pip install pandas matplotlib seaborn scikit-learn
```

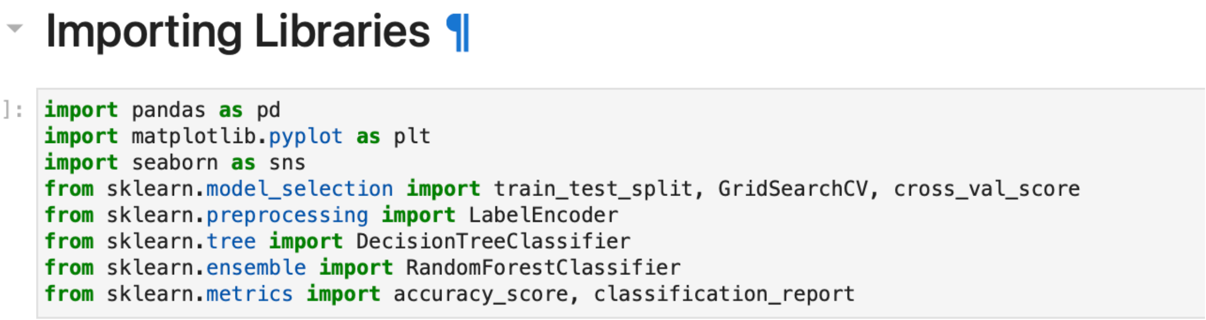
3 Project Structure

- Dataset: archive/full_data_flightdelay.csv
- Main Script: This script runs to process the data, train models, and make predictions is contained within the provided code.

4 Configuration Steps

4.1 Importing Libraries

- At the beginning of your script, import the necessary libraries:

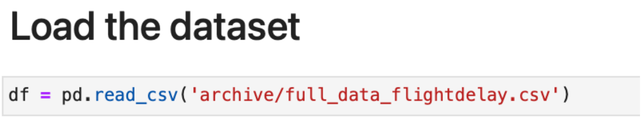


Importing Libraries ¶

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
```

4.2 Loading the Dataset

Load Dataset using Pandas



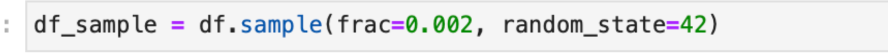
Load the dataset

```
df = pd.read_csv('archive/full_data_flightdelay.csv')
```

5 Data Preprocessing

5.1 Sampling the Data

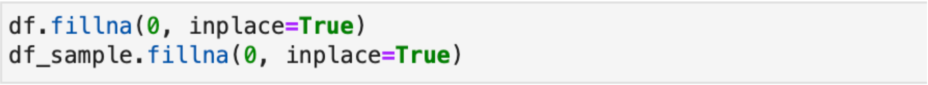
Take Small Sample of data for initial experimentation



```
df_sample = df.sample(frac=0.002, random_state=42)
```

5.2 handling Missing Values

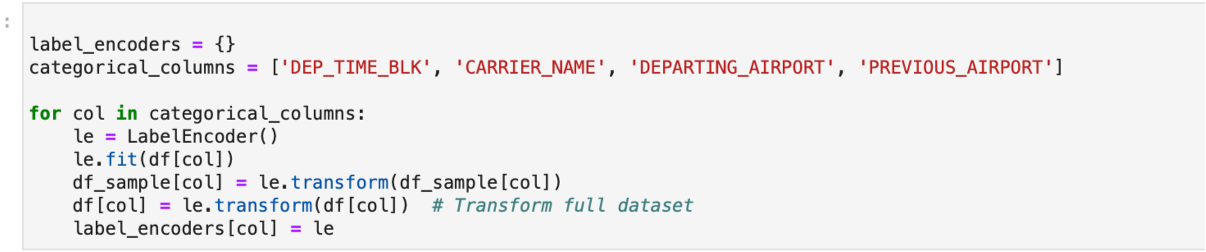
Fill Any missing values in dataset



```
df.fillna(0, inplace=True)
df_sample.fillna(0, inplace=True)
```

5.3 Encode Categorial Variables

Change non-numerical variables into numerical



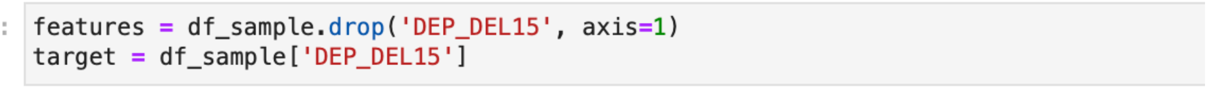
```
label_encoders = {}
categorical_columns = ['DEP_TIME_BLK', 'CARRIER_NAME', 'DEPARTING_AIRPORT', 'PREVIOUS_AIRPORT']

for col in categorical_columns:
    le = LabelEncoder()
    le.fit(df[col])
    df_sample[col] = le.transform(df_sample[col])
    df[col] = le.transform(df[col])  # Transform full dataset
    label_encoders[col] = le
```

6 Model Training

6.1 Defining features and target variables

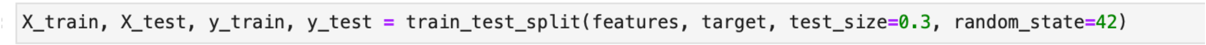
Separate the features and the target variable



```
features = df_sample.drop('DEP_DEL15', axis=1)
target = df_sample['DEP_DEL15']
```

6.2 Splitting the dataset

Split the dataset into training and testing sets

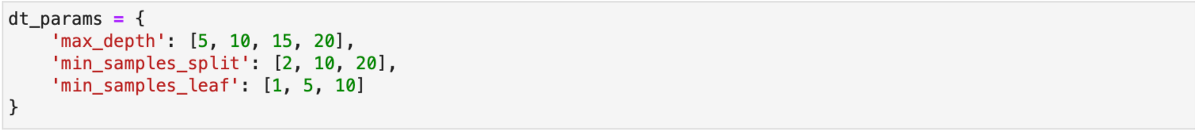


```
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.3, random_state=42)
```

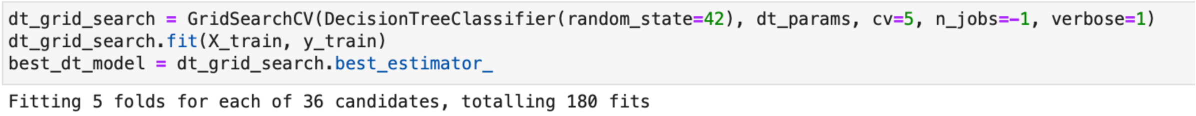
7 Training and hyperparameter tuning

7.1 Decision tree classifiers

Define hyperparameters and perform Grid Search



```
dt_params = {
    'max_depth': [5, 10, 15, 20],
    'min_samples_split': [2, 10, 20],
    'min_samples_leaf': [1, 5, 10]
}
```

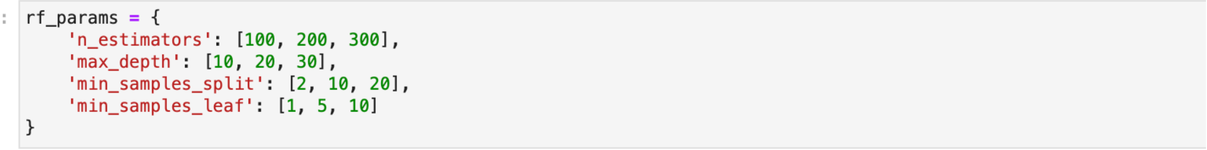


```
dt_grid_search = GridSearchCV(DecisionTreeClassifier(random_state=42), dt_params, cv=5, n_jobs=-1, verbose=1)
dt_grid_search.fit(X_train, y_train)
best_dt_model = dt_grid_search.best_estimator_
```

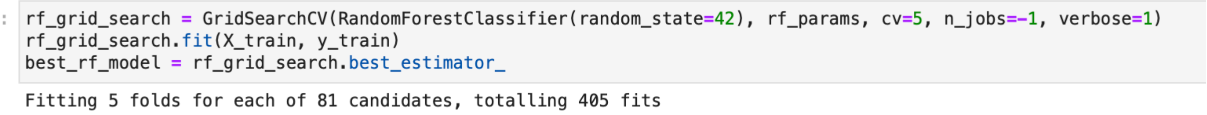
```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
```

7.2 Random Forest classifiers

Similarly, perform Grid Search for Random Forest Model



```
rf_params = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 20, 30],
    'min_samples_split': [2, 10, 20],
    'min_samples_leaf': [1, 5, 10]
}
```

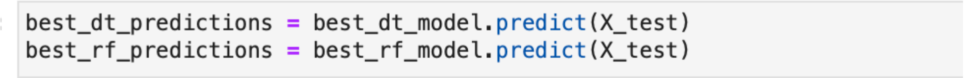


```
rf_grid_search = GridSearchCV(RandomForestClassifier(random_state=42), rf_params, cv=5, n_jobs=-1, verbose=1)
rf_grid_search.fit(X_train, y_train)
best_rf_model = rf_grid_search.best_estimator_
```

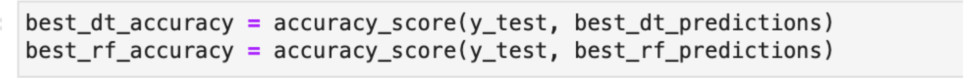
```
Fitting 5 folds for each of 81 candidates, totalling 405 fits
```

8 Feature Importance

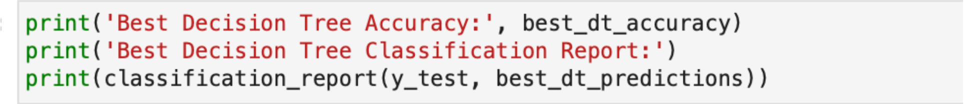
Evaluate the models using the test set



```
best_dt_predictions = best_dt_model.predict(X_test)
best_rf_predictions = best_rf_model.predict(X_test)
```



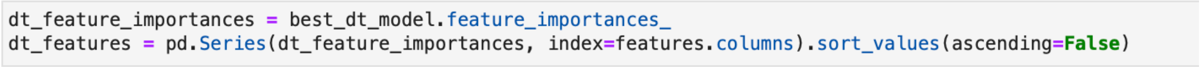
```
best_dt_accuracy = accuracy_score(y_test, best_dt_predictions)
best_rf_accuracy = accuracy_score(y_test, best_rf_predictions)
```



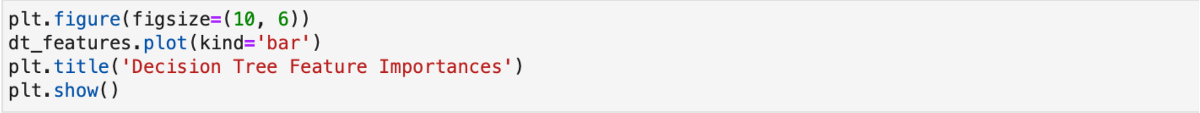
```
print('Best Decision Tree Accuracy:', best_dt_accuracy)
print('Best Decision Tree Classification Report:')
print(classification_report(y_test, best_dt_predictions))
```

9 Project Structure

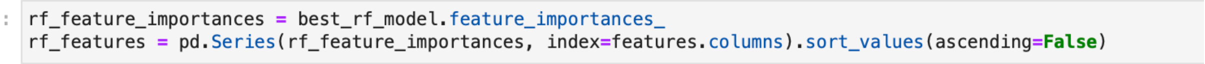
Analyse feature importance for both models



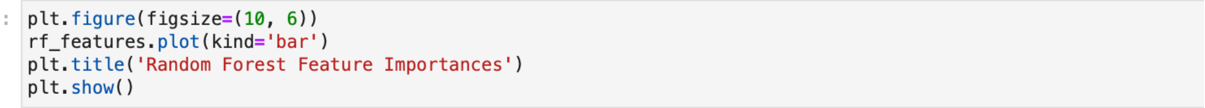
```
dt_feature_importances = best_dt_model.feature_importances_
dt_features = pd.Series(dt_feature_importances, index=features.columns).sort_values(ascending=False)
```



```
plt.figure(figsize=(10, 6))
dt_features.plot(kind='bar')
plt.title('Decision Tree Feature Importances')
plt.show()
```



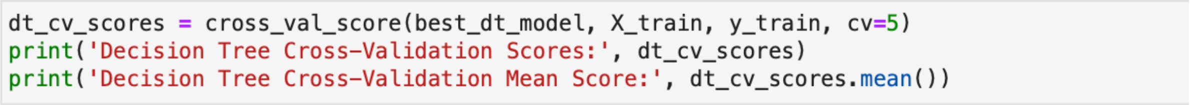
```
rf_feature_importances = best_rf_model.feature_importances_
rf_features = pd.Series(rf_feature_importances, index=features.columns).sort_values(ascending=False)
```



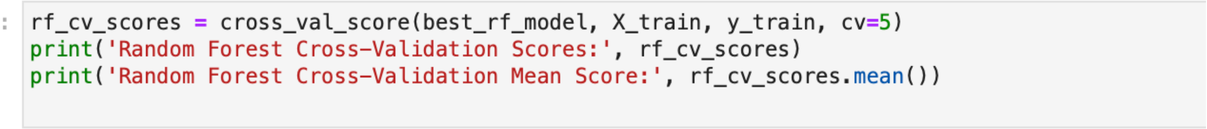
```
plt.figure(figsize=(10, 6))
rf_features.plot(kind='bar')
plt.title('Random Forest Feature Importances')
plt.show()
```

10 Cross Validation

Perform Cross Validation to ensure model robustness



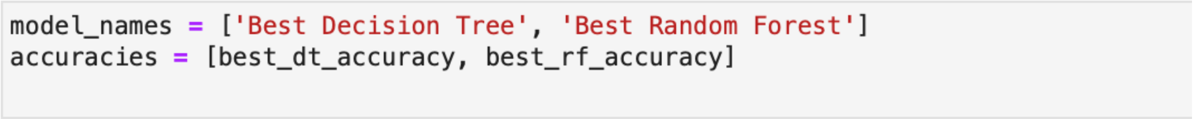
```
dt_cv_scores = cross_val_score(best_dt_model, X_train, y_train, cv=5)
print('Decision Tree Cross-Validation Scores:', dt_cv_scores)
print('Decision Tree Cross-Validation Mean Score:', dt_cv_scores.mean())
```



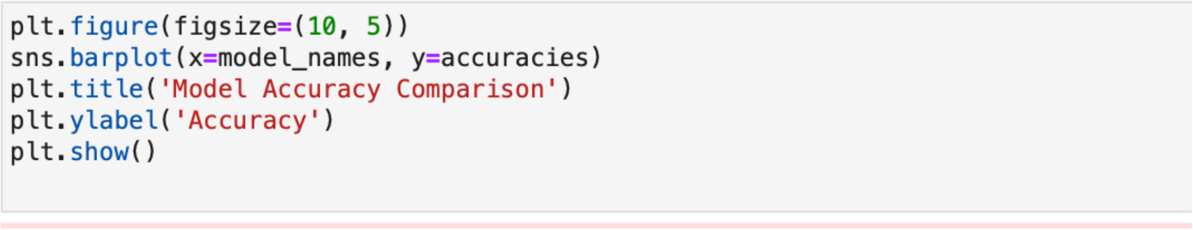
```
rf_cv_scores = cross_val_score(best_rf_model, X_train, y_train, cv=5)
print('Random Forest Cross-Validation Scores:', rf_cv_scores)
print('Random Forest Cross-Validation Mean Score:', rf_cv_scores.mean())
```

11 Model Selection

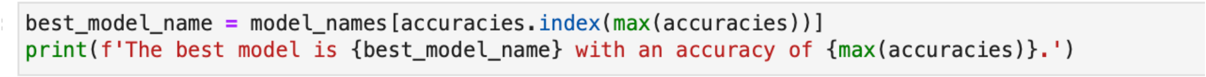
Finally determine which model performs best



```
model_names = ['Best Decision Tree', 'Best Random Forest']
accuracies = [best_dt_accuracy, best_rf_accuracy]
```



```
plt.figure(figsize=(10, 5))
sns.barplot(x=model_names, y=accuracies)
plt.title('Model Accuracy Comparison')
plt.ylabel('Accuracy')
plt.show()
```



```
best_model_name = model_names[accuracies.index(max(accuracies))]
print(f'The best model is {best_model_name} with an accuracy of {max(accuracies)}.')
```

12 Power BI Visualization

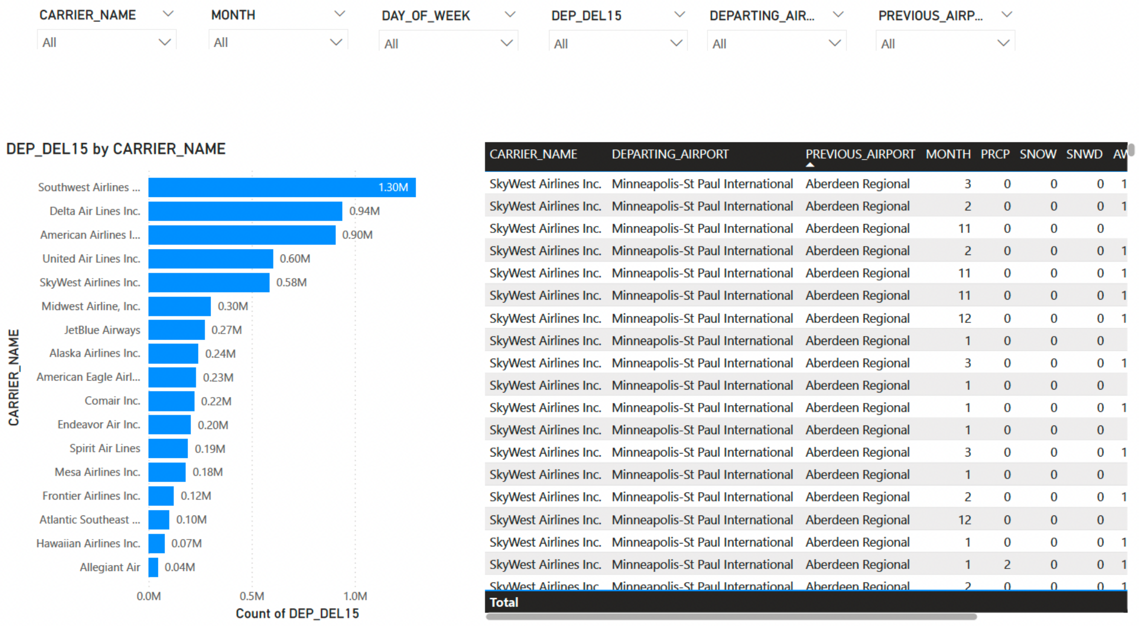
This project includes Power BI visualizations to better understand the distribution and impact of flight delays across different airlines and airports

12.1 Loading the dataset in power BI

- □ Import the dataset
- □ Use power BI's drag and drop feature to create visualizations

12.2 Key Visualization Created

- □ DEP_DEL15 by CARRIER_NAME: Displays the count of delayed flights by airline.
- □ Flight Delay Percentage: Shows the percentage of flights delayed versus those on time.
- □ Top 5 Delaying Airports: Lists the airports with the highest number of delayed flights.
- □ Delayed Flights by Airline: A bar chart showing the number of delayed flights for each airline.



CARRIER_NAME: All

MONTH: All

DAY_OF_WEEK: All

DEP_DEL15: All

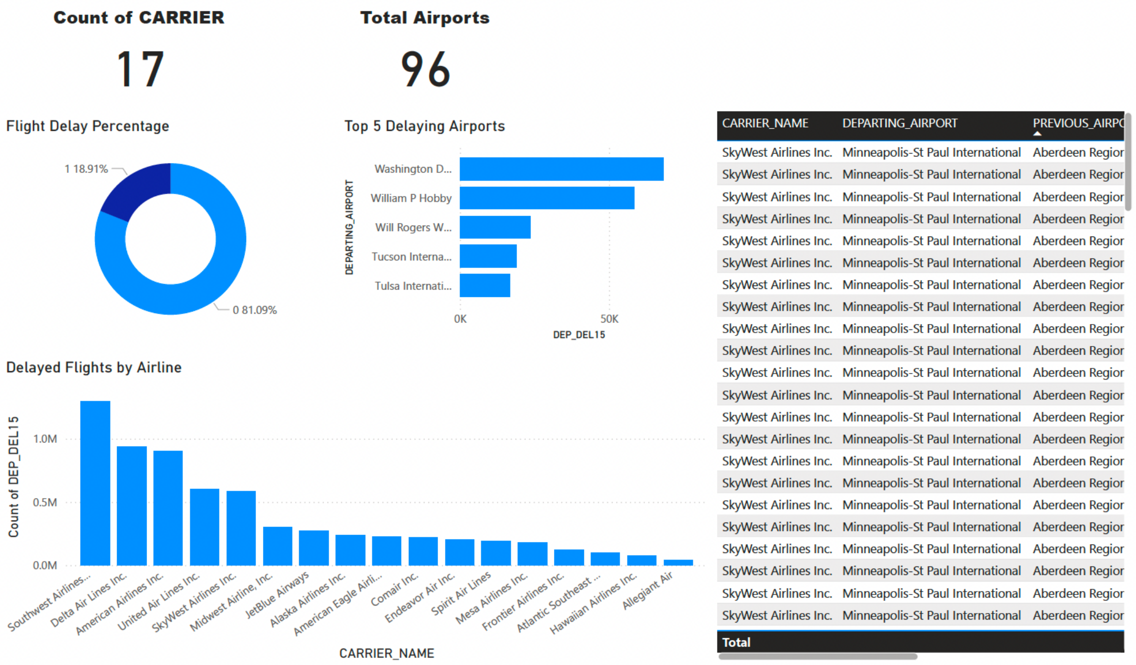
DEPARTING_AIR...: All

PREVIOUS_AIRP...: All

DEP_DEL15 by CARRIER_NAME

CARRIER_NAME
Southwest Airlines ... 1.30M
Delta Air Lines Inc. 0.94M
American Airlines I... 0.90M
United Air Lines Inc. 0.60M
SkyWest Airlines Inc. 0.58M
Midwest Airline, Inc. 0.30M
JetBlue Airways 0.27M
Alaska Airlines Inc. 0.24M
American Eagle Airl... 0.23M
Comair Inc. 0.22M
Endeavor Air Inc. 0.20M
Spirit Air Lines 0.19M
Mesa Airlines Inc. 0.18M
Frontier Airlines Inc. 0.12M
Atlantic Southeast ... 0.10M
Hawaiian Airlines Inc. 0.07M
Allegiant Air 0.04M
0.0M
0.5M
1.0M
Count of DEP_DEL15

CARRIER_NAME	DEPARTING_AIRPORT	PREVIOUS_AIRPORT	MONTH	PRCP	SNOW	SNWD	AW
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	3	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	11	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	11	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	11	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	12	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	3	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	3	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	12	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	2	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
Total							



Count of CARRIER

17

Total Airports

96

Flight Delay Percentage

Top 5 Delaying Airports

Delayed Flights by Airline

CARRIER_NAME	DEPARTING_AIRPORT	PREVIOUS_AIRPO
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regior
Total		