Predictive & Descriptive Analytics Code 7

Pedro Cuellar

Introduction

Utilizing data from airline flight delay records in 2018, I will use descriptive analytics to identify an independent variable and identify the correlation between the variable and 'ARR_DELAY' (arrival delays). Afterward, I will use predictive analytics to analyze arrival delays for future years from chosen airlines. I will support my findings with statistical code analysis and visuals.

Question

What predictor variables can most efficiently predict arrival delays for future years from three similar airlines?

Data Overview – Descriptive Analytics

Step 1: Identify potential independent variables for analysis

After reading and researching, column headers, departure delays, and distance travel were key indicators that could potentially predict arrival delays.

Input: Descriptive statistics for key indicators

```
#prints descriptive statistics on departure delays, grouping by airline name to determine predictor variable
print("Descriptive statistics on departure delays, grouping by airline name")
print (df.groupby(['OP_CARRIER_NAME'])['DEP_DELAY'].describe())

#prints descriptive statistics on distance, grouping by airline name to determine predictor variable
print("Descriptive statistics on departure delays, grouping by airline name")
print (df.groupby(['OP_CARRIER_NAME'])['DISTANCE'].describe())
```

Output: Descriptive statistics

Descriptive statistics on departure delays, grouping by airline name

OP_CARRIER_NAME	count	mean	std	min	25%	50%	75%	max
Alaska Airlines Inc.	924.0	42.188312	56.020458	-17.0	5.0	27.0	58.0	457.0
Allegiant Air	112.0	59.223214	68.866722	-4.0	25.0	40.0	67.5	513.0
American Airlines Inc.	56433.0	53.346127	70.725875	-15.0	13.0	36.0	70.0	1568.0
Comair Inc.	23009.0	54.816898	59.811674	-20.0	17.0	36.0	73.0	883.0
Delta Air Lines Inc.	44566.0	59.378988	80.586431	-21.0	19.0	37.0	70.0	1199.0
Endeavor Air Inc.	11786.0	66.849313	89.601392	-16.0	21.0	45.0	84.0	1206.0
Envoy Air	19749.0	42.209276	50.462858	-23.0	8.0	31.0	60.0	1086.0
ExpressJet Airlines Inc.	11389.0	65.663008	97.074197	-18.0	17.0	43.0	83.0	1426.0
Frontier Airlines Inc.	3048.0	73.173885	81.084777	-21.0	24.0	51.0	97.0	754.0
JetBlue Airways	3778.0	72.510323	76.230025	-23.0	26.0	50.0	94.0	699.0
Mesa Airlines Inc.	7612.0	65.085260	105.942931	-16.0	14.0	39.0	79.0	1487.0
Republic Airlines	16775.0	54.124888	73.678188	-20.0	9.0	37.0	74.0	1263.0
SkyWest Airlines Inc.	28559.0	66.382541	102.945044	-24.0	15.0	41.0	83.0	2710.0
Southwest Airlines Co.	13225.0	55.494669	53.462991	-10.0	24.0	40.0	68.0	564.0
Spirit Air Lines	5604.0	66.793005	100.678418	-15.0	14.0	39.0	87.0	1527.0
United Air Lines Inc.	33679.0	59.447252	80.450175	-18.0	13.0	40.0	78.0	1279.0
Virgin America	47.0	54.808511	65.223575	-13.0	8.0	34.0	73.0	240.0

Descriptive statistics on distance traveled, grouping by airline name

OP_CARRIER_NAME	count	mean	std	min	25%	50%	75%	max
Alaska Airlines Inc.	924.0	1889.089827	297.201143	1721.0	1721.0	1744.0	1874.0	2846.0
Allegiant Air	112.0	737.098214	308.739254	583.0	594.0	594.0	668.0	1514.0
American Airlines Inc.	56433.0	939.867631	548.577126	83.0	575.0	802.0	1182.0	4243.0
Comair Inc.	23009.0	363.657221	202.605894	75.0	203.0	335.0	500.0	1322.0
Delta Air Lines Inc.	44566.0	725.800274	491.554921	106.0	404.0	606.0	821.0	4502.0
Endeavor Air Inc.	11786.0	507.975479	276.196467	83.0	292.0	461.0	692.0	1416.0
Envoy Air	19749.0	457.753962	247.223824	83.0	286.0	431.0	594.0	1437.0
ExpressJet Airlines Inc.	11389.0	485.890245	227.432688	67.0	304.0	463.0	617.0	1195.0
Frontier Airlines Inc.	3048.0	1047.285761	421.491160	373.0	762.0	1005.0	1337.0	2174.0
JetBlue Airways	3778.0	822.749074	240.870487	184.0	740.0	867.0	950.0	1182.0
Mesa Airlines Inc.	7612.0	640.601550	334.433495	140.0	347.0	643.0	912.0	1530.0
Republic Airlines	16775.0	642.293651	359.313720	67.0	335.0	606.0	844.0	1501.0
SkyWest Airlines Inc.	28559.0	503.322595	323.452385	67.0	235.0	448.0	717.0	1730.0
Southwest Airlines Co.	13225.0	725.854140	390.079909	153.0	481.0	594.0	764.0	2149.0
Spirit Air Lines	5604.0	931.504461	390.600909	305.0	606.0	854.0	1120.0	1947.0
United Air Lines Inc.	33679.0	1051.494670	552.261372	108.0	719.0	925.0	1400.0	4243.0
Virgin America	47.0	1791.212766	51.971537	1739.0	1744.0	1744.0	1846.0	1846.0

Step 2: Reduce the scope of analysis

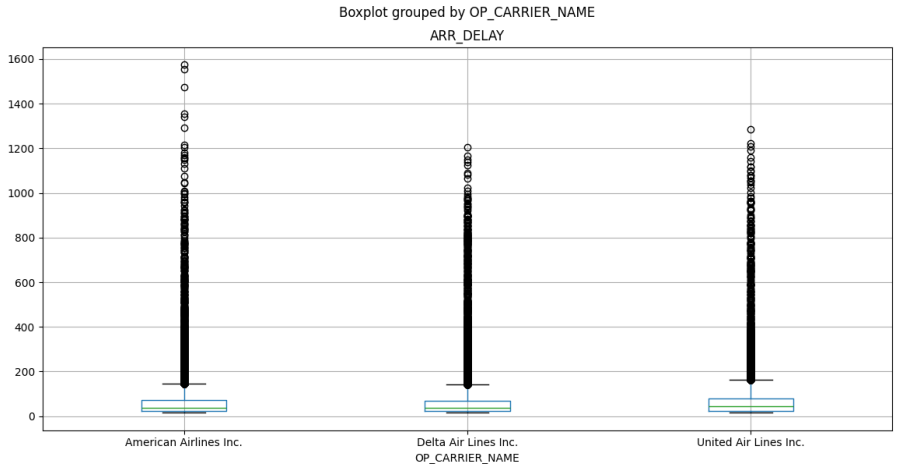
After my analysis, I chose to go with departure delays as my predictor variable. To further reduce the scope of my analysis, I chose American Airlines Inc, Delta Air Lines Inc, and United Air Lines Inc due to their similar amount of airline “count” and “mean” values under departure delays.

Step 3: Visualize the data from arrival delays

Input: Use boxplots to analyze arrival delays from three chosen airlines

```
#VISUAL REPRESENTATION
filter_query= "OP_CARRIER_NAME == 'Delta Air Lines Inc.' or OP_CARRIER_NAME == 'American Airlines Inc.' or OP_CARRIER_NAME == 'United Air Lines I
smaller_df= df.query(filter_query)
smaller_df.boxplot(column='ARR_DELAY', by='OP_CARRIER_NAME')
plt.show()
```

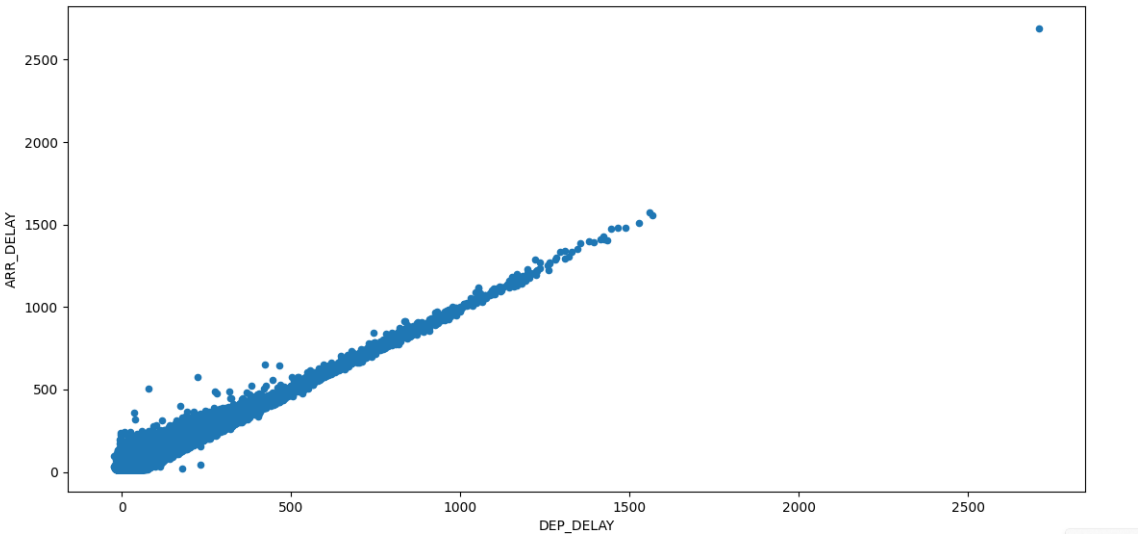
Output: Visual



Input: Use scatter plots to analyze arrival delays and departure delays correlation

```
df.plot.scatter(x='DEP_DELAY', y='ARR_DELAY')
plt.show()
```

Output: Visual



Descriptive Analytics Summary

By utilizing the box plot, the three airlines I chose have a similar box and outlier amount. The upper and lower lines are also like one another, ensuring an accurate comparison for the predictive analytics phase of the report. Furthermore, the strong correlation between my chosen predictor variable, departure delays, and arrival delays strengthens the accuracy of the predictive analytics.

Data Overview – Predictive Analytics

Step 1: Run linear regression model OLS

Input: OLS Code

```
#PREDICTIVE STATISTICS
#Run OLS
#define predictor and dependent variables
y = df['ARR_DELAY']
x = df['DEP_DELAY']
#add constant to predictor variables
x = sm.add_constant(x)
#fit linear regression model
model = sm.OLS(y, x).fit()
#view model summary
print(model.summary())
```

Output: OLS Analysis

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              ARR_DELAY   R-squared:                       0.927
Model:                            OLS   Adj. R-squared:                  0.927
Method:                 Least Squares   F-statistic:                 3.577e+06
Date:                Sun, 06 Apr 2025   Prob (F-statistic):               0.00
Time:                        12:30:03   Log-Likelihood:            -1.2441e+06
No. Observations:              280295   AIC:                         2.488e+06
Df Residuals:                  280293   BIC:                         2.488e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.0360      0.048    209.279      0.000       9.942      10.130
DEP_DELAY      0.9245      0.000   1891.330      0.000       0.924       0.925
==============================================================================
Omnibus:                   102928.006   Durbin-Watson:                   1.285
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           634414.579
Skew:                           1.643   Prob(JB):                         0.00
Kurtosis:                       9.597   Cond. No.                         122.
==============================================================================
```

Predictive Analytics Summary

As seen in the previous scatter plot, the relationship between arrival delays and departure delays is strong. Indicating that the use of my chosen predictor variable is a good comparison. Using the OLS's statistics, we can determine that an airline can still have an

arrival delay of around 10 minutes, even without a departure delay. As seen through the ‘coef’ value. Furthermore, for each minute a flight waits to depart, the arrival delay minutes increase by 0.9245.

Conclusion

By utilizing descriptive and predictive analytics, I found a strong correlation between arrival delays and departure delays. By utilizing the equation: Predicted Arrival Delay = 10.00360 + 0.9245*(departure delay minute). Airlines can potentially predict arrival delays to advise customers in advance and increase their statistical analysis of airline times.