ALY 6110 Module 6 Assignment

John DiSessa, Lawrence Mrowka, and Rachel Su

ALY 6110 Data Management and Big Data

Northeastern University College of Professional Studies

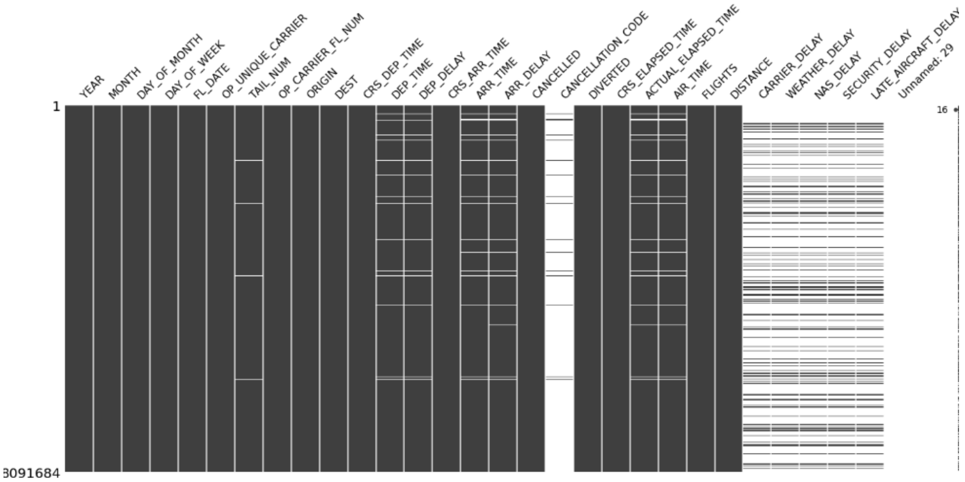
Introduction

Flying has changed the world. People and objects are able to be moved across large distances within hours. This fact has been largely taken for granted, and individuals often expect the complex task of air travel to be completed on time. However, that is not always the case. Flights can be canceled and delayed with little to no notice for the passengers and those connected to them. These travel plans can be wildly inconvenient for a whole host of reasons. Therefore, in a previous module we performed analysis in order to help travelers avoid the dreaded flight delay and cancellation. To provide useful insights to travelers we focused on identifying the best airlines to fly on, the best airports to travel through, and the best days to travel on. The results from each of these three focus areas could then be combined to inform air travelers about the best ways for them to travel — assuming the goal is to limit their risk of experiencing a flight delay or cancellation. However, we were unsure of how to weigh the results of each of the individual analyses. Therefore, we produced a correlation table to find which of our focus areas has the greatest impact on flight delays and cancellations.

EDA

<u>Missing value investigation</u>

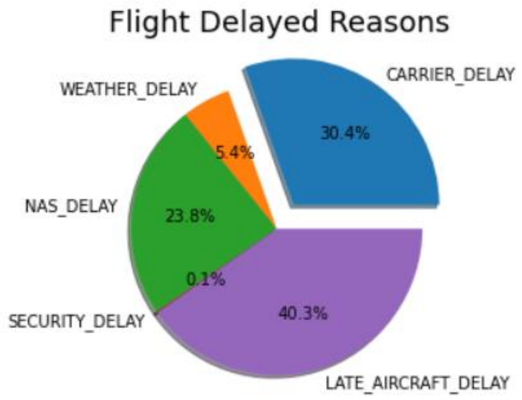
Using the missingno Python library to identify and visualize missing data to get a sense of our dataset. The columns



CANCELLATION_CODE , CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY, and LATE_AIRCRAFT_DELAY show large portions of missing data. which make sense due to the flights that are operated without delay and cancellation are still the majority.

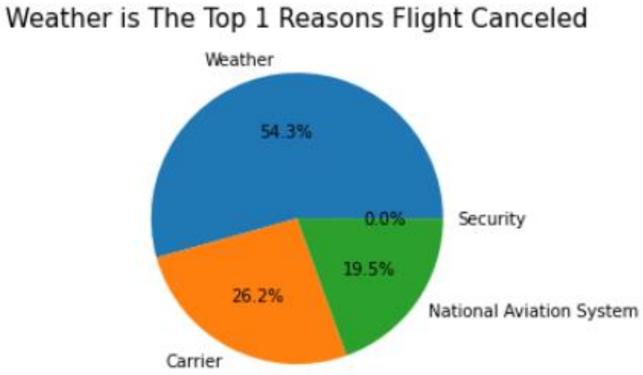
Overall Delay Causes

The main reason cause the flight delayed is “late aircraft delayed” .It means the previous flight with same aircraft arrived late, causing the present flight to depart late.



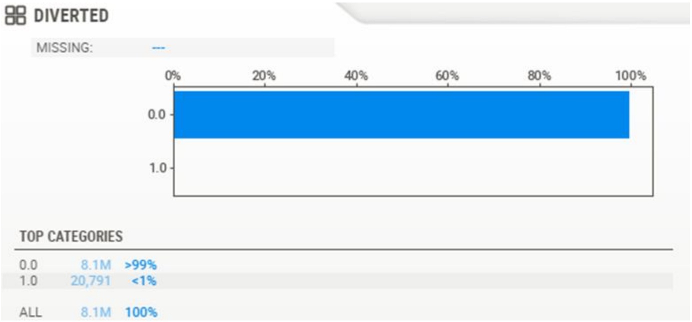
Cancellation Reasons

Weather is the Top reason to cause flight cancellations. Followed by Carrie, National Aviation System and we have less than 1% of flight cancellations were due to security.



Diversion Ratio

There are only less than 1 % of US domestic were diverted in 2019



MISSING: ---

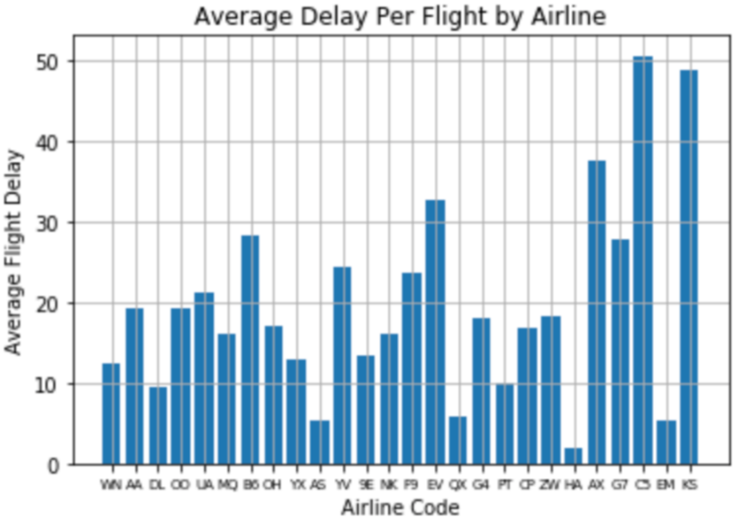
TOP CATEGORIES

0.0	8.1M	>99%
1.0	20,791	<1%
ALL	8.1M	100%

Analysis

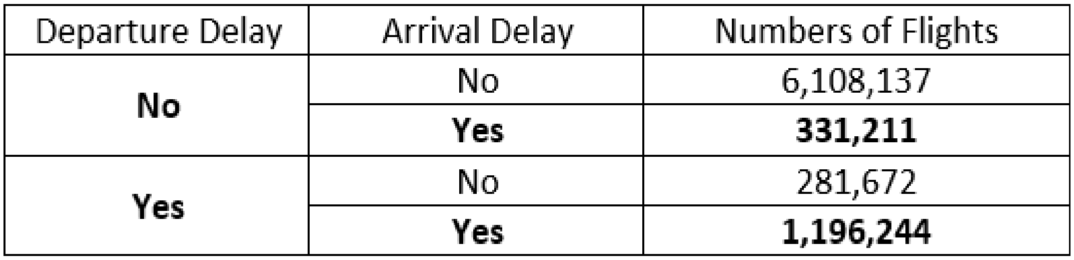
We used the Jupyter Notebook platform to perform our data analysis in Python. The first step in our analysis process was to prepare the data set. This process consisted of creating new variables, editing existing variable types, and investigating and handling outliers. With the necessary adjustments made to the flight data set we were then able to individually analyze three flying focus areas: airline selection, airport selection, and flight date selection. Upon drawing focus area specific conclusions we then combined the results to provide travelers with an optimal flying plan to avoid unexpected delays and cancellations.

As mentioned above, the first focus area of the analysis was to identify the most reliable airline for travelers to fly on. The basic analysis performed in module 4 informed us that travelers should avoid flying Peninsula Airways (KS) and Commuteair (C5) because greater than 40% of the flights featured delays and averaged a delay of at least 25 minutes per flight. We also learned that Empire Airlines (EM),



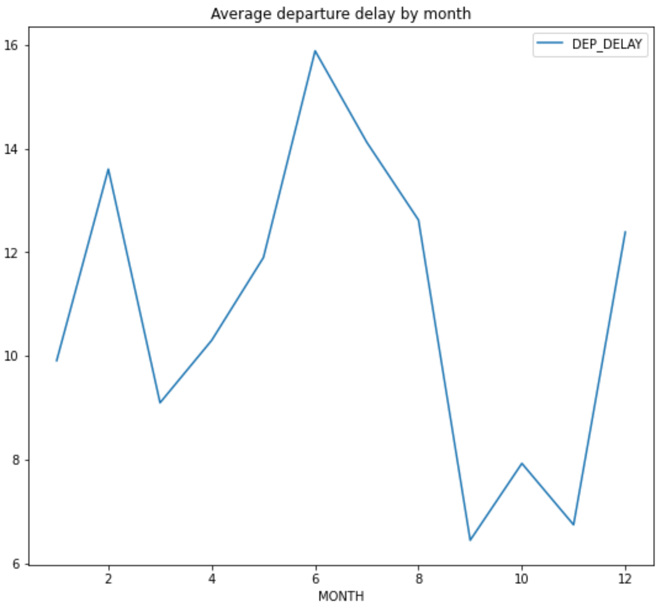
Alaska Airlines (AS), and Hawaiian Airlines (HA) sported the lowest mean flight delay values out of the airlines in the dataset. The plot below shows the distribution of mean flight delays for each airline. Finally, we learned that Endeavor Airlines (9E), Air Wisconsin (ZW), and Delta Airlines (DL) had the lowest flight delay rates — all less than 30%. Based on these results, among others from the prior analysis, we were able to make decisive recommendations on what airlines should be trusted and which airlines should not. We felt that the insights produced could be useful, but also might not deserve complete trust.

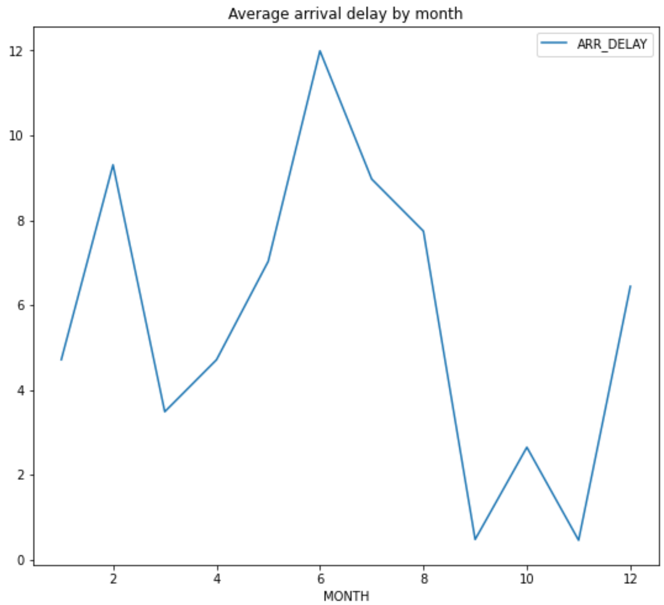
The second focus area from our basic analysis centered around travel locations. Our purpose is to find the delay rate among the airports. We considered that the departure or arrival time from the flight's original destination is delayed for more than 15 minutes. As expected, arrivals are more likely to have delays if there is a delay on departure.

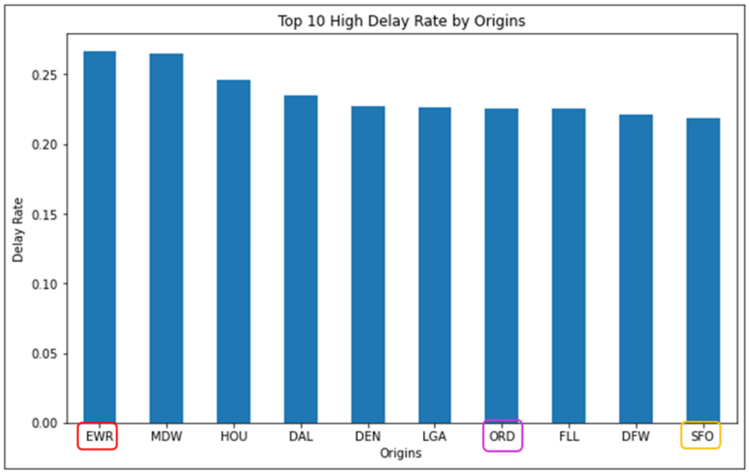


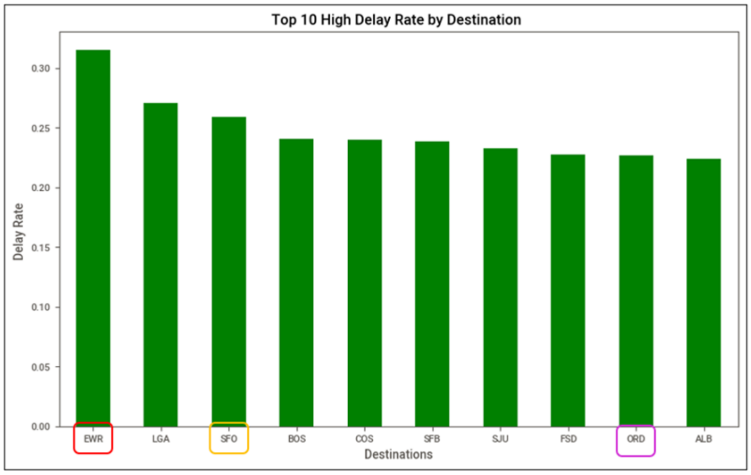
Departure Delay	Arrival Delay	Numbers of Flights
No	No	6,108,137
	Yes	**331,211**
Yes	No	281,672
	Yes	**1,196,244**

Now, we know there is a correlation between departure and arrival delayed. Furthermore, we check if there is a certain monthly pattern on delayed flights during the year. Both departure and arrival delays show a similar pattern. Delays focused on February, June and December, with a peak on average delays in July 2019.



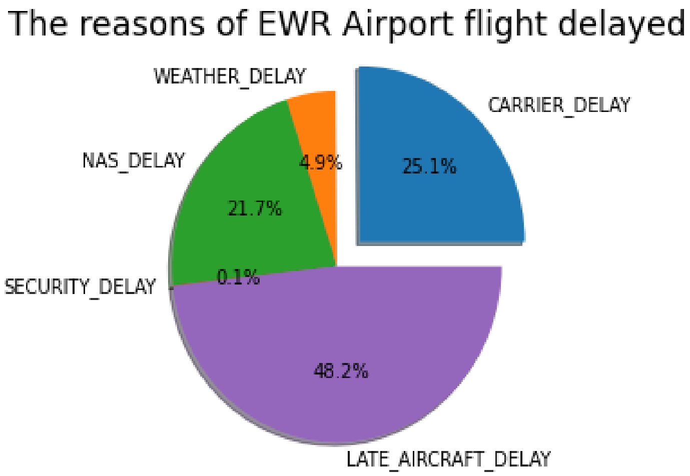


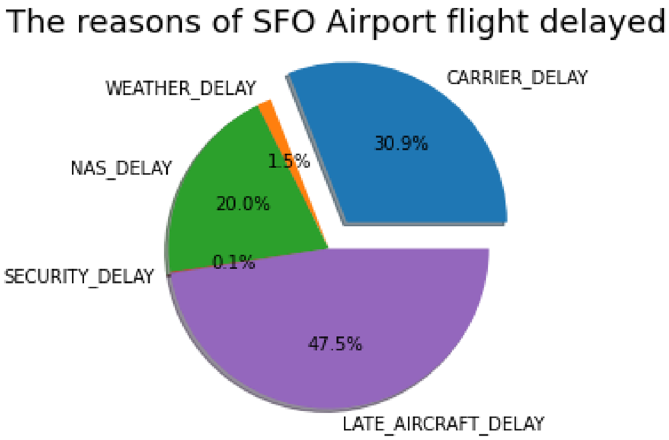


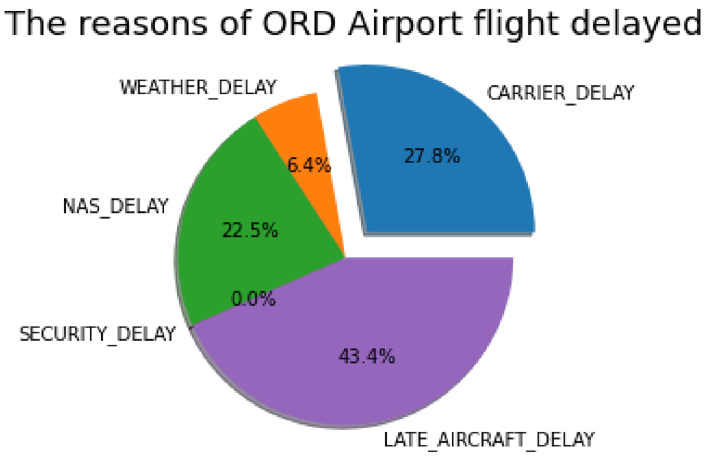


According to our analysis, EWR, SFO and ORD airports are on both of the top 10 high delay rate lists. EWR is the top highest on both lists.

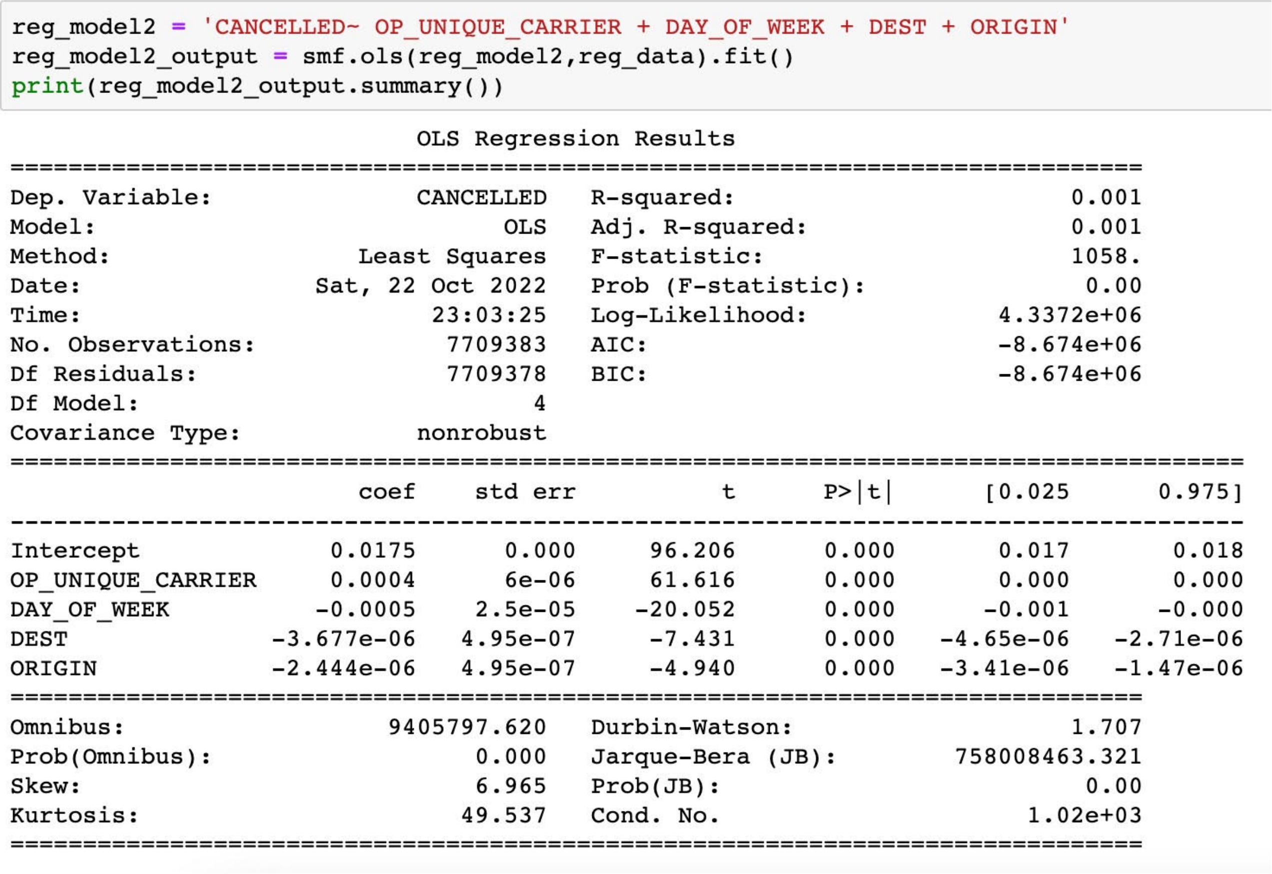
If we look into the delayed reasons of the above three airports which are on both top delayed lists. We found that the top 3 reasons that cause delays are all the same. The first reason that accounts for around 45% of delays is late aircraft, followed by carrier delay and weather delay. This insight might be helpful to these three airports understanding their main causes of flight delay and taking reactions to improve it.







The third part of our analysis consisted of trying to create a model from linear regression in order to predict the likelihood of a flight being canceled. We used Airline Carrier, Day of Week, Destination, and Origin as our independent variables and had Cancelled as our dependent variable. Unfortunately, we were unable to build a model since each independent variable had essentially 0 for coefficients. Each independent variable was technically able to be put into the model since each one was statistically significant, however the model as a whole was not a good predictor of cancelled flights. The adjusted R square value of .001 meant that our model was only able to account for .1% of the variation in cancelled flights. Rather than just assuming flights are cancelled at random, it is very likely that factors outside the database cause cancelled flights such as weather and airline staffing issues.

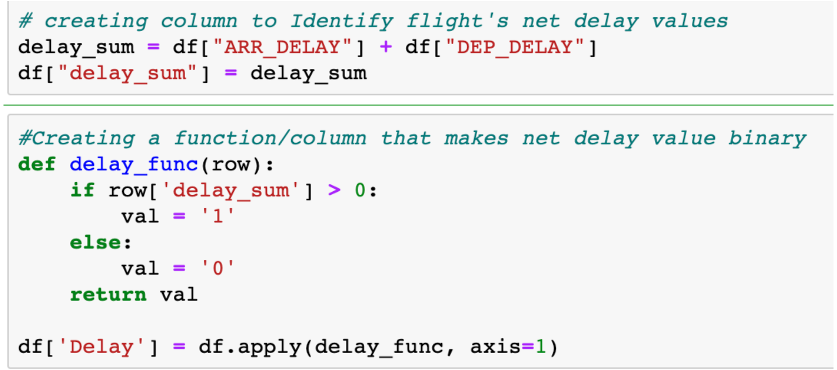


```
reg_model2 = 'CANCELLED~ OP_UNIQUE_CARRIER + DAY_OF_WEEK + DEST + ORIGIN'
reg_model2_output = smf.ols(reg_model2,reg_data).fit()
print(reg_model2_output.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              CANCELLED   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1058.
Date:                Sat, 22 Oct 2022   Prob (F-statistic):               0.00
Time:                        23:03:25   Log-Likelihood:             4.3372e+06
No. Observations:             7709383   AIC:                        -8.674e+06
Df Residuals:                 7709378   BIC:                        -8.674e+06
Df Model:                           4                                         
Covariance Type:            nonrobust                                         
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             0.0175      0.000     96.206      0.000       0.017       0.018
OP_UNIQUE_CARRIER     0.0004      6e-06     61.616      0.000       0.000       0.000
DAY_OF_WEEK          -0.0005    2.5e-05    -20.052      0.000      -0.001      -0.000
DEST              -3.677e-06   4.95e-07     -7.431      0.000   -4.65e-06   -2.71e-06
ORIGIN            -2.444e-06   4.95e-07     -4.940      0.000   -3.41e-06   -1.47e-06
==============================================================================
Omnibus:                  9405797.620   Durbin-Watson:                   1.707
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        758008463.321
Skew:                           6.965   Prob(JB):                         0.00
Kurtosis:                      49.537   Cond. No.                     1.02e+03
==============================================================================
```

After taking in the insights from the previous module’s analysis we wanted to investigate how we should value the insights. In order to do so we then sought out to identify which variables had the strongest correlations with flight delays and flight cancellations. To gain these insights we performed correlation analysis on the binary “Delayed” and “Canceled” variables — both displayed a 1 if the flight was delayed or canceled and a 0 if it was not. The binary “Delayed” variable was not part of the original data set and was created using the code below.

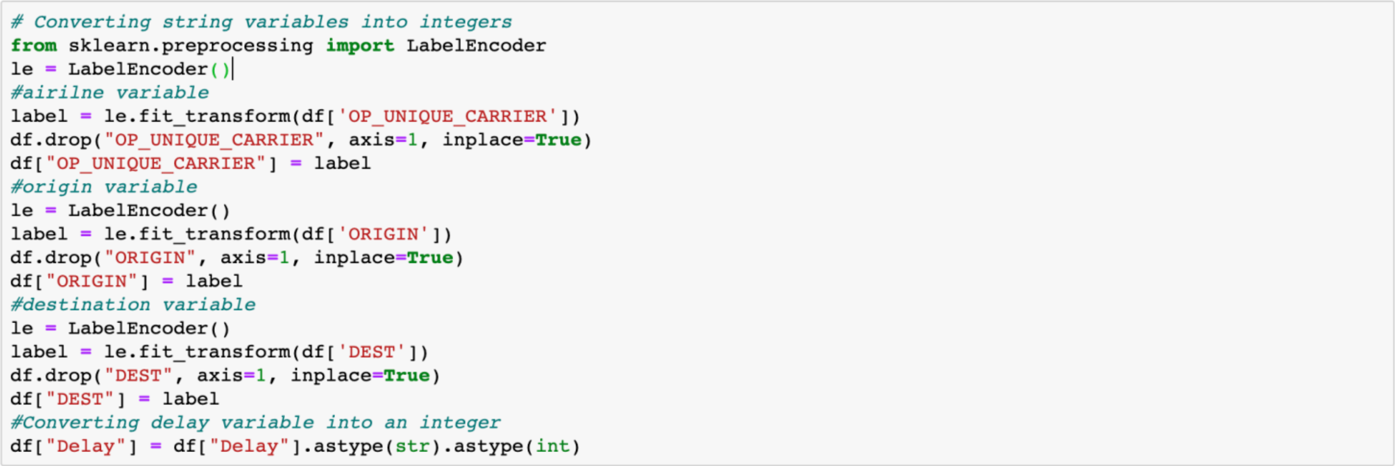
We chose to perform correlation analysis in this case because the coefficients would give us insight into the strength of the relationship between the independent variables — in this case airline, day of the week, destination, and origin. To perform the analysis we first needed to adjust the type of some of the variables. Airline, origin, and destination needed to be converted from string variables into numeric variables, and the “Delay” variable needed to be converted into a numeric. The code used to produce these data transformations is shown below.



```
# creating column to Identify flight's net delay values
delay_sum = df["ARR_DELAY"] + df["DEP_DELAY"]
df["delay_sum"] = delay_sum
```

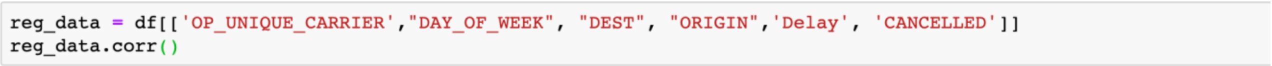
```
#Creating a function/column that makes net delay value binary
def delay_func(row):
    if row['delay_sum'] > 0:
        val = '1'
    else:
        val = '0'
    return val

df['Delay'] = df.apply(delay_func, axis=1)
```



```
# Converting string variables into integers
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
#airilne variable
label = le.fit_transform(df['OP_UNIQUE_CARRIER'])
df.drop("OP_UNIQUE_CARRIER", axis=1, inplace=True)
df["OP_UNIQUE_CARRIER"] = label
#origin variable
le = LabelEncoder()
label = le.fit_transform(df['ORIGIN'])
df.drop("ORIGIN", axis=1, inplace=True)
df["ORIGIN"] = label
#destination variable
le = LabelEncoder()
label = le.fit_transform(df['DEST'])
df.drop("DEST", axis=1, inplace=True)
df["DEST"] = label
#Converting delay variable into an integer
df["Delay"] = df["Delay"].astype(str).astype(int)
```

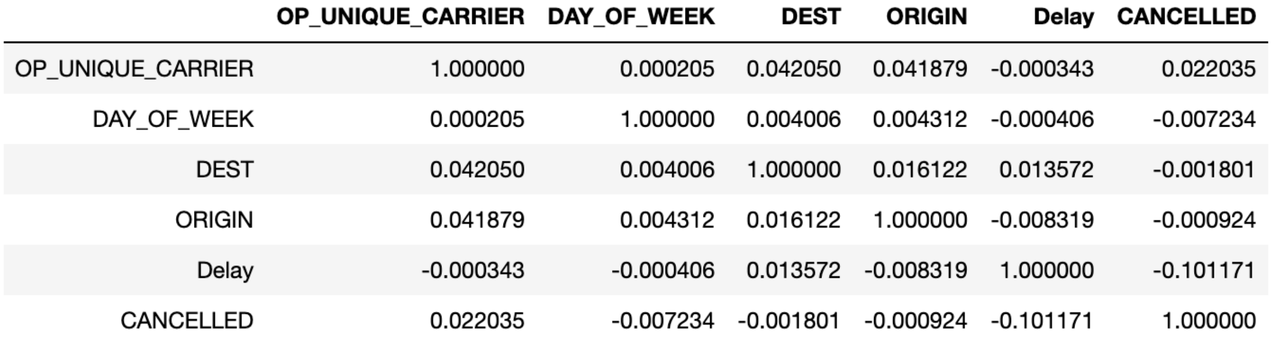
With the variables in appropriate forms we were now able to begin building the correlation table. The screenshot below shows the code used to build the correlation table.



```
reg_data = df[['OP_UNIQUE_CARRIER',"DAY_OF_WEEK", "DEST", "ORIGIN",'Delay', 'CANCELLED']]
reg_data.corr()
```

Results

The correlation output is displayed below. We focused on two columns: Delay and CANCELLED.



	OP_UNIQUE_CARRIER	DAY_OF_WEEK	DEST	ORIGIN	Delay	CANCELLED
OP_UNIQUE_CARRIER	1.000000	0.000205	0.042050	0.041879	-0.000343	0.022035
DAY_OF_WEEK	0.000205	1.000000	0.004006	0.004312	-0.000406	-0.007234
DEST	0.042050	0.004006	1.000000	0.016122	0.013572	-0.001801
ORIGIN	0.041879	0.004312	0.016122	1.000000	-0.008319	-0.000924
Delay	-0.000343	-0.000406	0.013572	-0.008319	1.000000	-0.101171
CANCELLED	0.022035	-0.007234	-0.001801	-0.000924	-0.101171	1.000000

The “Delay” column of the correlation table shows us which of our focus variables is most, and least, correlated with a flight being delayed. The destination variable was shown to have the strongest correlation with delayed flights (0.0135), and the airline variable, “OP_UNIQUE_CARRIER”, was shown to have the weakest correlation (-.000343) — the negative number means that the two variables are inversely correlated. These details *could* provide travelers focused on avoiding flight delays to focus on the rather than the airline insights presented in the previous module’s analysis. Further, the “CANCELLED” column sported the airline variable, “OP_UNIQUE_CARRIER”, as the most highly correlated (.022035). On the other hand, the flight's origin was the weakest correlated with flight cancellations (-0.000924). Therefore, travelers most concerned with the risk of flight cancellations should shift their focus to the analysis done on airlines in last week's module.

However, it is also relevant to note that all of the focus correlations displayed in this plot are extremely weak and should merely be used to compare importance *between* variables. Unfortunately, this result informs us that there is randomness in the delays and cancellations of flights that we are unable to protect against. Users could, however, use the insights provided to give themselves the greatest, albeit marginal, improvement of odds.

Conclusions

- Arrivals airport are more likely to have delays if there is a delay on departure airports.
- EWR airports has the highest delayed rate in 2019. For the travellers who fly from or to this airport should plan ahead to accommodate the high chances of flight delayed.
- The peak months of delayed flights are February, June and December.
- Travelers should avoid Peninsula Airways (KS) and Commuteair (C5) because of their high delay risk
- To avoid flight delays travelers should focus on destination
- To avoid flight cancellation travelers should focus on airline selection
- Weekday flights are more likely to be cancelled than weekend flights

References

Bobbitt, Z. (2022, March 15).Pandas: How to Create Bar Plot from GroupBy. Statology. Retrieved October 15, 2022, from https://www.statology.org/pandas-groupby-bar-plot/

Creating a new column based on other columns in pandas DataFrame. (n.d.). Retrieved October 15, 2022, from https://www.skytowner.com/explore/creating_a_new_column_based_on_other_columns_in_pandas

Sklearn.preprocessing.LabelEncoder. scikit. (n.d.). Retrieved October 22, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

How to explode a pie chart using Matplotlib in Python. (n.d.). Educative: Interactive Courses for Software Developers. Retrieved October 22, 2022, from https://www.educative.io/answers/how-to-explode-a-pie-chart-using-matplotlib-in-python