Abstract

The COVID-19 pandemic halted operations and normal activities in almost every sector but the travel industry arguably took one of the largest hits. In 2020, the airline industry had a record negative net income and a low number of monthly passengers and is slowly recovering from this economic downturn. Unfortunately, airlines are still facing obstacles to their success, as shown by their inability to perform as well as in previous years. Also, despite the number of monthly passengers returning to normal pre-pandemic levels, the airline industry is still not earning what they used to. In this project, we researched the causes behind this phenomenon and provided insights into how this can be combated.

Some key findings in our project include:

- Despite passenger numbers returning to pre-COVID levels, the airline industry is still not earning as much as they did pre-2020. The proportion of delays is also higher in the years following COVID.
- The number of total passengers is strongly correlated with overall delay times, carrier delays, and late aircraft delays. This suggests that airlines may be overbooking customers beyond operating capacity.
- Compared to other reasons like security, carrier, and so on, A flight is delayed due to the late arrival of the aircraft and has the highest total average delay minutes.
- The fuel costs in 2020 decreased significantly compared to the previous years, due to the impact of the COVID-19 pandemic on the aviation industry.
- United Airlines and American Airlines have consistently consumed the highest amount of fuel over the past 5 years.

Problem Statement

In this project, we wish to explore which factors truly impact flight delays and to provide insights to airlines on which areas to focus their optimization efforts. This will also help prove the case for optimizing flights to avoid delays to ultimately save the airline industry a massive amount of money and help provide a better service experience for consumers. Striking a balance between airline economy and passenger satisfaction is of main importance and supports customer retention. We will be using data from Chicago O'Hare Airport and Miami International Airport. Originally, we hoped to build a model that accurately predicts the effect of delays on the airline economy in the U.S. using historical data from years 2018-2022 to predict 2023 and use the actual values as a reference. We have instead generated a series of regression analyses to determine the impact of each factor in our collective datasets on the occurrence and length of delays.

In the future, we would hopefully have access to private datasets and have more adequate data to perform a well-rounded analysis that includes various aspects of airline economy and delays. We initially intended to only analyze data from O'Hare International Airport to add local personalization but quickly found that there was not enough data. Therefore, we also included

data from Miami International Airport to create a more well-rounded analysis as Florida is a state that has a very different climate compared to Illinois. Establishing a model that is able to predict delay times based on various inputs would be extremely valuable to airlines, as they would be able to prepare for delays ahead of time and hopefully minimize their negative impact.

Literature Review

Due to the ongoing climate crisis, extreme weather conditions have wreaked havoc on the airline industry [1]. Thunderstorms, smoky skies, and high temperatures make traveling less of a vacation and more of a potential danger. The Federal Aviation Administration has stated a shortage of workers as being another critical factor in determining flight performance. Although the total number of employees working in the U.S. airline industry has risen to 801,835 individuals, which is even higher than pre-pandemic times, there is a shortage of critical workers such as air traffic controllers. Although the airline industry is slowly rebuilding itself, profits are much lower than pre-pandemic times, even though the number of flights and employees are at a satisfactory level. We believe that delays are to blame for this extreme loss in airline profit margins and it is keeping the industry from making a full comeback.
Contrary to popular belief, weather is not the main cause of flight delays and this was also found in our summary analysis of the total number of delays and their delay types.

Interestingly, the internet has changed the way people plan and purchase flights as well as their expectations for flight prices [2]. A plethora of easily accessible sites allows consumers to now find the cheapest flights possible across airlines and purchasing platforms, down to the date and time someone should purchase their ticket in advance. Companies have noticed this surge of consumers choosing price over quality for flight options and have been attempting to reallocate their resources to finding flight routes that may not be the quickest– but rather the most cost-effective. Airlines have also resorted to laying off employees to save on operational costs in order to keep ticket prices low, which drives delays more. However, with how much delays cost airlines, this may be hurting them in the long run as a single flight delay can cause entire flight schedules to run off-track. Optimizing to decrease delays rather than optimizing for price may be the more fruitful option for both airlines and consumers.

This change can be seen in the 2022 airline financial data provided by the Bureau of Transportation. In 2022, 73.5% of airline revenue came from fares, an increase from the previous year, while 31.5% of operating expenses were attributed to labor, a decrease from the previous year [3]. Another way that airlines attempt to maximize profit is by overbooking [4]. This is because airlines are a service which means tickets are a perishable commodity, with a value that expires with the timeframe of the flight. In other words, each flight must be filled with as many passengers as possible to gain maximum profit and empty seats should be avoided. However, we anticipated that this would again drain profits as the total number of passengers can have influence delay times. We will explore these changes to the airline industry in our analysis and gauge their importance in order to provide future recommendations.

Data Processing - Pipeline details, data issues, assumptions/adjustments.

- **Filtering the Dataset:** based on specific years, airports, and airlines.
- **Cleaning**
    - Making it a Readable CSV
    - Various datasets/tables will be joined in R by Date or Unique Carrier or Airport Code as data keys.
    - Dealing with any missing values and replacing them with the mean value of the respective feature or deleting them.
    - Outliers will be detected using Excel MAX and MIN functions and further analyzed through boxplots if needed.
- **Get summary statistics for an overview of the data: mean, std dev, minimum, maximum**
- **Conducted data exploration and analysis**

After data processing, the next step is comprehensive exploration and analysis of the datasets.

- We will conduct various activities such as distributional analysis, summary statistics, visualization, time series analysis, and so on.
- **Descriptive Statistics:** Compute summary statistics (mean, median, standard deviation, etc.) for relevant variables. For example, weather conditions, and flight delay duration. These results can provide us with a baseline understanding of the data.
- **Visualization:** Use data visualization techniques such as histograms, and plots to visualize the distribution of data and then find relationships between variables.
- **Time Series Analysis:** Flight delays can have temporal patterns, perform time series analysis to identify trends and seasonality in delays. This can help us understand how delays vary over time, across months or seasons.
- **Correlation Analysis:** Compute correlation coefficients between different variables to determine if there are significant relationships between factors such as weather conditions, carrier, and security delays.
- **Software packages, applications, libraries, and associated tools, etc.**
  Softwares: RStudio, R
  Libraries/Packages: Tidyverse, Caret
    ggplot2:creating a wide variety of visualizations.
    dplyr: data filtering, transformation, aggregation, and more.
    readr: read CSV files.
  Associated tools: Excel, quick view of the dataset.

Firstly, we attempted to gather datasets that were relevant to our problem statement but quickly realized that a lot of datasets were incomplete or unavailable for public use. Therefore, we tried to find datasets that could relate to our master dataset from the BTS on flight delays.

Based on our initial research of the increasing flight delays caused by extreme weather and natural disasters, we attempted to use weather data which included monthly average precipitation and temperatures to model this correlation. However, little to no correlation was found and in hindsight, this could have been due to the fact that weather fluctuates through hours and months were too vast of a timeframe. Also, it was difficult to find datasets that specifically identified the occurrence of a wildfire or decreased levels of visibility due to smog.

Data Description

There are 6 datasets we used including weather data, airfare data, fuel data, O'Hare airport data, Miami airport data, and passenger data. We collected these datasets to analyze how these variables affect the response variable(delay time).

There are 144 observations and 5 variables in the weather dataset which include year, month, mean_in, mean_f, and airport variables. The variable types are numerical and categorical.
There are 69 observations and 9 variables in the airfares dataset which include Year, Quarter, Month, U.S. Average (Current $), U.S. Average (Inflation-Adjusted $), Chicago O'Hare - IL (Current $), Chicago O'Hare - IL (Inflation-Adjusted $), Miami - FL (Current $), Miami - FL (Inflation-Adjusted $) variables. The types are numerical.

There are 528 observations and 12 variables in the fuel dataset which include year, month, carrier, dom_consumption, dom_cost, dom_cpg, int_consumption, int_cost, int_cpg, tot_consumption, tot_cost, tot_cpg. The variable types are numerical. In the dataset, we can understand the change in domestic and international costs for fuel at that time.

There are 711 observations and 21 variables in the O'Hare airport dataset which include year, month, carrier, carrier_name, airport, airport_name, arr_flights, arr_del15, carrier_ct, weather_ct, nas_ct, security_ct, late_aircraft_ct, arr_cancelled, arr_diverted, arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay. The variable types are numerical and categorical.

There are 490 observations and 16 variables in the Miami airport dataset which include year, month, carrier, carrier_name, airport, airport_name, arr_flights, arr_del15, arr_cancelled, arr_diverted, arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay. The variable types are numerical and categorical.
There are 689 observations and 7 variables in the passenger dataset which include, year, month, carrier, airport, dom_passengers, intl_passengers, and total_passengers. The variable types are numerical.

R Join Example

```
Airport_Data = rbind(Miami,O'Hare)
df = inner_join(Airport_Data, Fuel_Data, by=c('year'='year',
'month'='month','airport'='airport'))
```

Data Analysis - Summary statistics, visualization, feature extraction.

In the predictive model, we have included several datasets as variables in the model to predict what kinds of factors would affect airline delays the most. We compare delay time in each type, passengers, and fuel costs to predict the pattern. In this section, we may base on the data visualization to get to understand dataset distribution.

Airlines with Average Delay Minutes by Each Factor

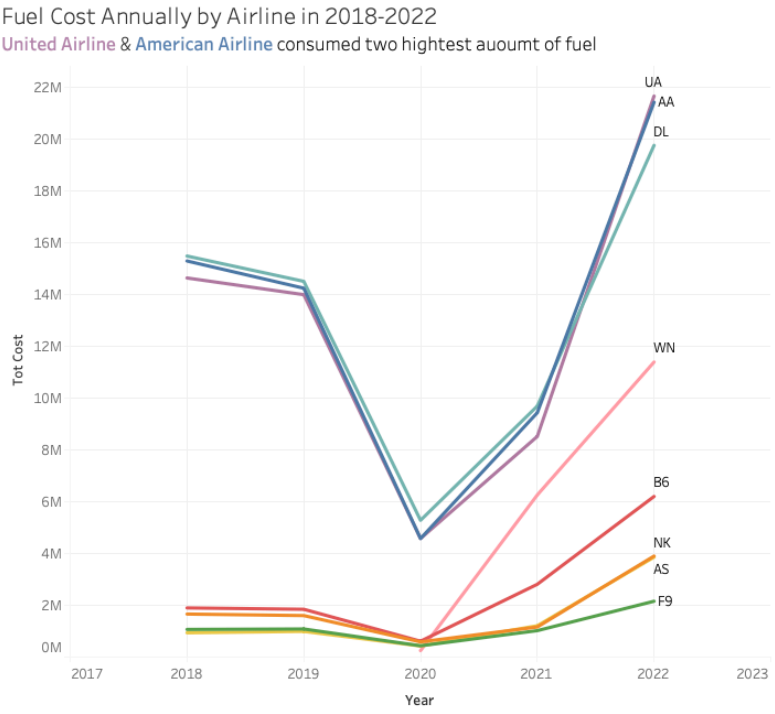
Airport	Carrier Name	Avg. Late Aircraft Delay	Avg. Nas Delay	Avg. Carrier Delay	Avg. Weather Delay	Avg. Security Delay
ORD	United Air Lines Inc.	24,754	21,761	16,403	4,222	8
	American Airlines Inc.	21,800	15,182	18,060	3,180	97
	Envoy Air	19,553	15,887	12,350	4,546	45
	Republic Airline	5,589	6,677	4,057	1,084	24
	Spirit Air Lines	2,551	5,231	2,006	541	57
	Delta Air Lines Inc.	2,055	4,929	3,564	738	11
	Frontier Airlines Inc.	1,610	982	991	63	0
	JetBlue Airways	769	1,220	1,267	102	6
	Endeavor Air Inc.	442	977	386	75	1
	Alaska Airlines Inc.	341	1,083	687	91	4
Grand Total		8,009	7,446	6,023	1,476	25

Based on the heat map for the main five delay types in each airline, we can observe that late arrival delay has the highest total average delay minutes of 8009 minutes in the past 5 years. Indicating a flight is delayed due to the late arrival of the aircraft assigned to operate that particular flight.

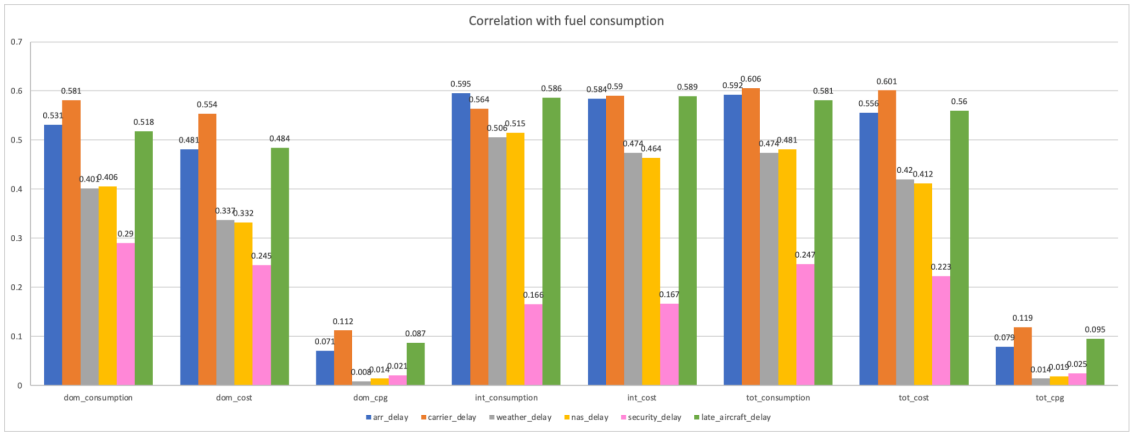
% of number of **Flight** and **Passengers** by Airlines

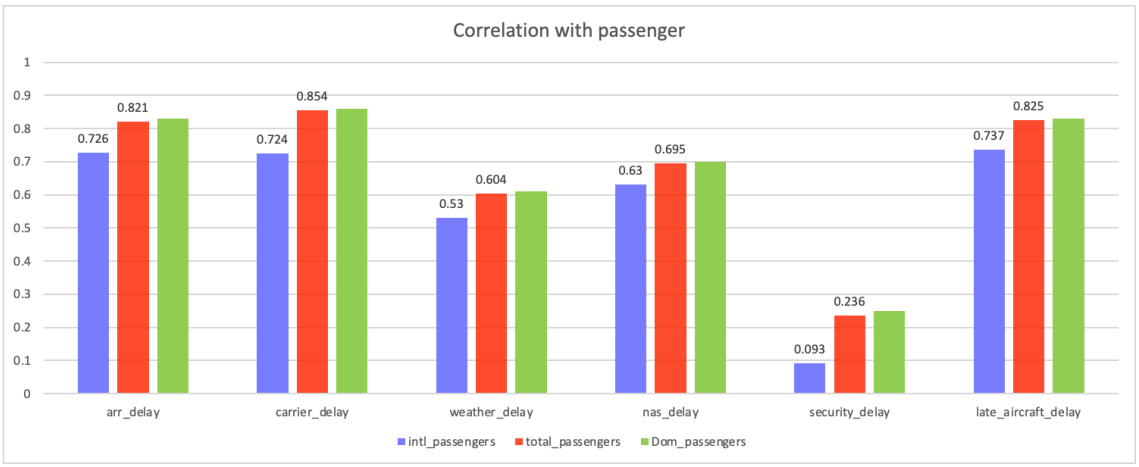
Carrier Name	% of Total Arr Flights along Table (Down)	% of Total Total Passengers along Table (Down)
United Air Lines Inc.	29.77%	40.48%
American Airlines Inc.	24.80%	31.53%
Envoy Air	23.61%	9.28%
Spirit Air Lines	3.78%	5.56%
Republic Airline	9.40%	4.73%
Delta Air Lines Inc.	4.47%	3.95%
Frontier Airlines Inc.	1.06%	1.85%
Alaska Airlines Inc.	1.33%	1.56%
JetBlue Airways	0.82%	0.64%
Endeavor Air Inc.	0.69%	0.31%
PSA Airlines Inc.	0.27%	0.11%
Grand Total	100.00%	100.00%

Furthermore, we compare the total percentage of passengers with the total number of flights for each airline. Over 70% of passengers took United Airlines and American Airlines in the past 5 years and we can reflect this in the number of flights. These two airlines have over 50% of the total flights. Based on these visualizations, we may indicate that increased passenger numbers typically align with a higher total count of flights. **We can consider assuming that the total number of passengers on an airline may influence the time, reflecting potential associations between passenger load and flight delays.**



Based on the dataset on fuel consumed and costs, we can simply understand the distribution of these 5 years. The COVID-19 pandemic affects the travel industry worldwide. It shows that fuel costs in 2020 decreased significantly compared to the previous years, likely due to the impact of the COVID-19 pandemic on the aviation industry. We also can understand that United Airlines and American Airlines have consistently consumed the highest amount of fuel over the past 5 years. Through visualization, **we can establish a connection between fuel consumption, the number of flights, and consequently, the passenger load for each airline.**





Delays vs. Fuel Consumption Bar Chart

- Arr_delay, Carrier_delay, and Late_aircraft_delay Relationships:

The analysis indicates substantial relationships between Arrival Delays, Carrier Delays, and Late Aircraft Delays with International Fuel Cost and Consumption. These delays exhibit a noteworthy correlation with the corresponding international fuel consumption and cost metrics.

- Comparison with Domestic Fuel Metrics:

Interestingly, the relationships between these delays and Domestic Fuel Cost and Consumption appear to be comparatively lower. Domestic fuel metrics show a less pronounced correlation with Arrival Delays, Carrier Delays, and Late Aircraft Delays when compared to their international counterparts.

Delays vs. Passenger Bar Chart

- Domestic Passengers Correlation with Delays:

The data pertaining to Domestic Passengers demonstrates a slightly higher correlation with delays when compared to International Passengers. This suggests that delays in the aviation system may have a marginally stronger association with domestic passenger metrics than with international passenger metrics.

Proposed Methodology (Initial/Final)

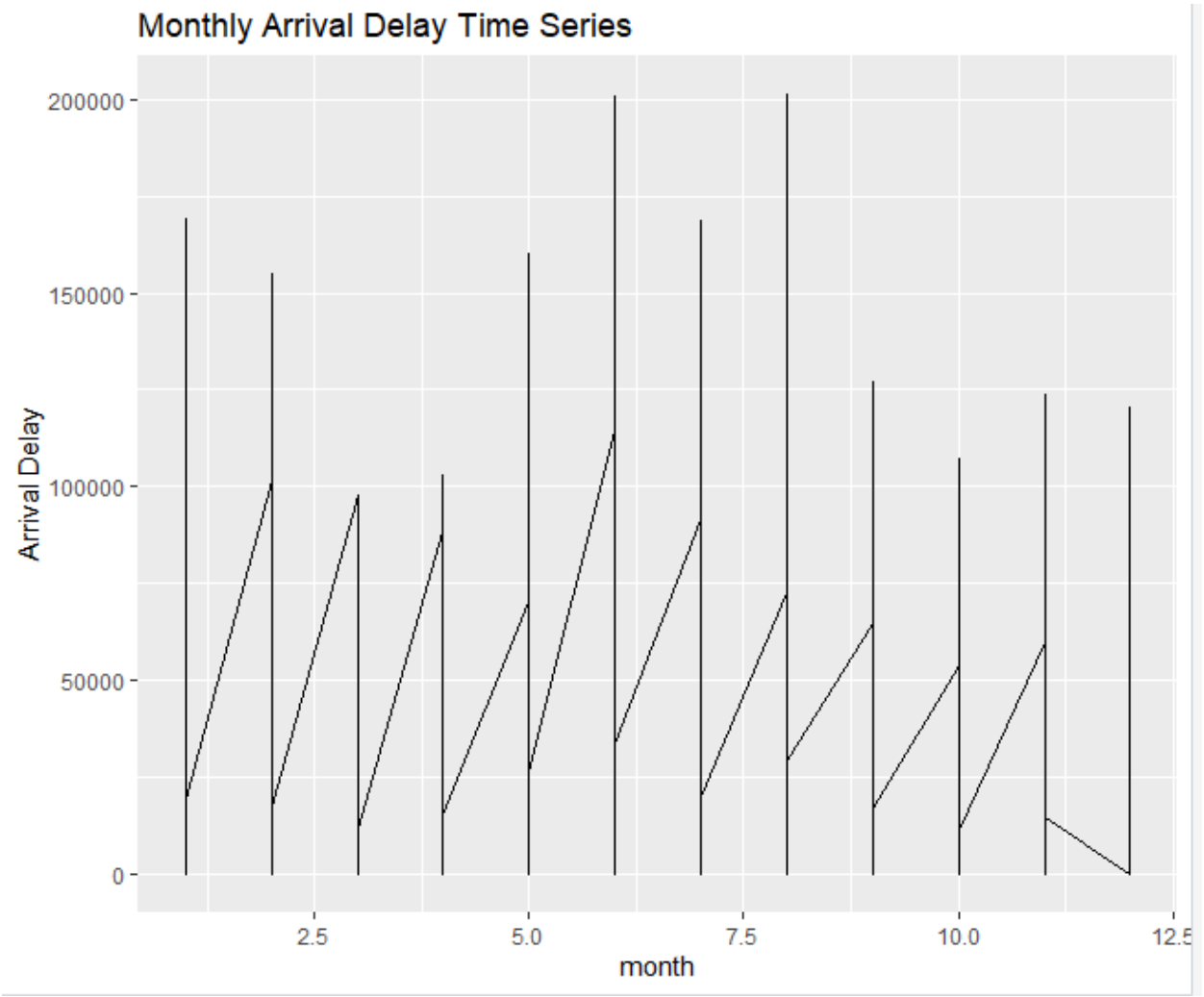
Initial: We will build a multiple linear regression model, embedded with Carrier, Weather (Average Precipitation and Average Temperature per month), Security, National Airspace System (NAS), and Late Aircraft as predictors. We will determine whether the predictors are significant or not after running the regression analysis and checking the coefficients and p-values. The response will be a dummy variable that takes on the value 1 if the flight is delayed or 0 if the flight is on schedule, at the most basic level model. We may use time-series regression, logistic regression, or lasso regression once we obtain more data or wish to expand on certain questions.

We had to deviate from our original project plan of optimizing flight arrival times through the prediction of delay types and delay times because we were unable to find usable datasets that were available publicly. We attempted to reach out to OAG Aviation Group and FlightAware but were unfortunately immediately redirected to the sales team to make a purchase. We tried out best to piece together data found from the BTS Airline Datasets, such as fuel expenditure and employee numbers.

Model Training Process - Feature Engineering, Model Selection, and Testing Results

In the model training process, we would like to elaborate on our data model individually. First, we will define our main topic **Passenger data,** and then describe the additional approach to **Fuel Expenditure**. The following will include our modification of input variables to improve the model's performance, capture patterns, and list the interpretation from our data model.

Times series plot



We generate the result by (ggplot). It helps us to visualize how the arrival delay changes over months. We may see from the chart that January, July and August have the highest arrival delays among all over months.

Predictive modeling

```
> cat("Predicted future delay:", future_prediction, "\n")
Predicted future delay: 54488.45
>
```

We are trying to predict the future delay of airlines. We chose an “American Airlines” flight from “Miami” airport with flight number “3995”. The methodology includes data splitting (training and testing dataset), model training, prediction, and evaluation. The delay time of this specific airline in the future will be 54488.45 minutes.

Passenger Data

In the Passenger Data, we consider applied **correlation analysis** and a **linear regression** model. These methodological choices were based on data processing and data visualization. In

correlation analysis, it provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis calculates parameters in a linear equation that can be used to predict the values of one variable based on the other.

Correlation Analysis

Based on the feature of the correlation analysis, we can estimate the strength of the linear relationship between two variables and compute their relations. We selected passenger and delay variables and calculated correlation coefficients between passenger and delay variables.

The higher the coefficient close to 1, the stronger the correlation is.

Variable	arr_delay	carrier_delay	weather_delay	nas_delay	security_delay	late_aircraft_delay
intl_passengers	0.726	0.724	0.530	0.630	0.093	0.737
total_passengers	0.821	0.854	0.604	0.695	0.236	0.825
Dom_passengers	0.83	0.86	0.61	0.70	0.25	0.83

Linear Regression

Second, we loaded the passenger data and ran linear regression. We defined predictor variables (X: domestic and international passengers) and target variables (y: arrival delay). Fitted a linear regression model (lm) to predict arrival delay based on passenger counts.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.833e+03  9.536e+02   2.971 0.003087 **
dom_passengers   1.293e-01  7.219e-03  17.906  < 2e-16 ***
intl_passengers -1.763e-01  4.979e-02  -3.541 0.000428 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18370 on 619 degrees of freedom
Multiple R-squared:  0.6885,    Adjusted R-squared:  0.6875
F-statistic:   684 on 2 and 619 DF,  p-value: < 2.2e-16
```

OBJ

Briefly, interpretation based on the result:

- R-squared: 0.688 - The model explains about 68.8% of the variance in 'arr_delay'.
- Intercept: Approximately 2832.8 minutes predicted delay when both passenger counts are zero.
- dom_passengers': Coefficient of 0.1293, suggesting a positive relationship with 'arr_delay'.

- 'intl_passengers': Coefficient of -0.1763, indicating a negative relationship with 'arr_delay'.
- P-value less than 0.05, indicating all coefficients are statistically significant.

Multicollinearity

By reviewing our result, we noticed it might cause potential concerns related to multicollinearity since our model predictor variables are highly correlated. To test the multicollinearity, we calculated Variance Inflation Factors (VIF).

dom_passengers	intl_passengers
6.309168	6.309168

We may consider a VIF above 5 or 10 as an indication of multicollinearity. Here VIF is over 6.3. In our case, the correlation between dom_passengers and intl_passengers is approximately 0.917. It explains the multicollinearity problem.

Tested domestic and international passengers separately

To reduce multicollinearity and did further observation to discover potential interaction effects. We chose to test the dom_passenger and intl_passenger separately with arr_delay respectively. The decision to separate the analysis could affect the stability and interpretability of the coefficients in a combined model.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.038e+03  8.990e+02   4.492 8.42e-06 ***
dom_passengers 1.058e-01  2.901e-03  36.479  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18540 on 620 degrees of freedom
Multiple R-squared:  0.6822,    Adjusted R-squared:  0.6817
F-statistic:  1331 on 1 and 620 DF,  p-value: < 2.2e-16
```

Briefly, interpretation based on the result for domestic passengers:

- R-squared: 0.682. This suggests that approximately 68.2% of the variability in arr_delay is explained by the number of domestic passengers. It indicates a strong relationship.
- dom_passengers: The coefficient is 0.1058. This means for each additional domestic passenger, the arr_delay is expected to increase by approximately 0.1058 minutes, holding all else constant.

- Both the intercept and the coefficient for dom_passengers are statistically significant (P < 0.05).

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.199e+04  9.907e+02   12.11   <2e-16 ***
intl_passengers 6.416e-01  2.441e-02   26.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22620 on 620 degrees of freedom
Multiple R-squared:  0.5271,    Adjusted R-squared:  0.5264
F-statistic: 691.1 on 1 and 620 DF,  p-value: < 2.2e-16
```

Briefly, interpretation based on the result for international passengers:

- R-squared: 0.527. This suggests that approximately 52.7% of the variability in arr_delay is explained by the number of international passengers.
- intl_passengers: The coefficient is 0.6416. This means for each additional international passenger, the arr_delay is expected to increase by approximately 0.6416 minutes, holding all else constant.
- Both the intercept and the coefficient for intl_passengers are statistically significant (P < 0.05).
- The model indicates a significant positive relationship between the number of international passengers (intl_passengers) and the arrival delay (arr_delay). The impact of international passengers on arrival delay appears to be stronger than that of domestic passengers (based on the higher coefficient). This result suggests that increases in international passengers are associated with greater increases in arrival delays, compared to domestic passengers.

In the analysis of Passenger Data, we may observe correlation analysis and linear regression were statistically significant in understanding the relationship between passenger variables and arrival delays. The linear regression model explained a significant portion of the variability in delays (R-squared: 0.688). Furthermore, both domestic and international passenger counts showed statistically significant impacts on delays, with domestic passengers positively influencing delays and international passengers indicating a negative effect. By addressing this through VIF calculations, we separate tests for domestic and international passengers for multicollinearity. The model has shown a stronger positive relationship between international passengers and delays compared to domestic passengers. Overall, we can base this comprehensive analysis provides insights into the factors influencing arrival delays in passenger data.

Fuel Expenditure

Next, we focus on our additional approach to fuel expenditure. In the data set, we conducted the data with various regression, including, **multiple linear regression**, **random forest**, **lasso regression**, and **ridge regression** for predicting arrival delays based on fuel expenditure variables.

Multiple Linear Regression

In this step, We make our assumptions on our data visualization, that there is a linear relationship between the dependent variable and each of the independent variables. By defining predictor variables (consumption, cost, cost per gallon for both domestic and international) and target variable (arr_delay), we can predict arrival delay based on fuel-related factors.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6.907e+03  4.215e+03  -1.639 0.101660
dom_consumption  2.458e-01  7.293e-02   3.371 0.000788 ***
dom_cost        -9.001e-02  3.182e-02  -2.829 0.004796 **
dom_cpg         -1.026e+04  5.471e+03  -1.876 0.061032 .
int_consumption -2.847e-02  9.385e-02  -0.303 0.761724
int_cost         1.428e-01  3.994e-02   3.576 0.000371 ***
int_cpg          1.336e+04  5.402e+03   2.473 0.013609 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23210 on 762 degrees of freedom
Multiple R-squared:  0.3782,    Adjusted R-squared:  0.3733
F-statistic: 77.23 on 6 and 762 DF,  p-value: < 2.2e-16
```

Briefly, interpretation based on the result:

- R-squared: 0.378, indicating that approximately 37.8% of the variability in arr_delay can be explained by the model.
- F-statistic: 77.20, with a p-value of 2.77e-75, suggesting that the model is statistically significant.

Coefficients**:**

- dom_consumption: Coefficient of 0.2570, meaning for each unit increase in domestic consumption, arr_delay increases by 0.2570 minutes, significantly.
- dom_cost: Coefficient of -0.0931, indicating that an increase in domestic cost is associated with a decrease in arr_delay, significantly.
- dom_cpg: Coefficient of -8891.3793, not significant at the 5% level (p-value: 0.075)
- int_consumption: Coefficient of -0.0319, not significant (p-value: 0.735)
- int_cost: Coefficient of 0.1427, indicating a significant positive relationship with arr_delay.
- int_cpg: Coefficient of 12400, significantly associated with arr_delay.

Random Forest

We also applied the random forest approach, It is primarily used for predictive modeling and is powerful for multicollinearity in some areas. Since random forest provides high accuracy, flexibility, and ease of use by applying random forest, we implement this approach to understand what the result would be.

```
 randomForest(formula = arr_delay ~ dom_consumption + dom_cost +      dom_cpg
on + int_cost + int_cpg, data = data,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 680024338
                    % Var explained: 20.77
> |
```

Briefly, interpretation based on the result:

- The model explains 20.76% of the variance in arr_delay. And based on the result, we consider the explanatory power might be low. We are trying Lasso and Ridge regression.

Lasso Regression

We understand that Lasso and Ridge are regularization techniques that address overfitting and multicollinearity, but they differ in the type of penalty applied to the coefficients and the impact on variable selection. Lasso tends to provide sparse models with exact zero coefficients, while Ridge retains all predictors with reduced impact. In this part, we conduct these two regressions to expand our further research.

Ridge Regression

We keep building our model with ridge regression. Since ridge regression is particularly useful when there is multicollinearity in the data, effectively prevents the model from becoming overly sensitive to the training data.

```
> coef(ridge_model, s = best_lambda_ridge)
7 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)     -1.086175e-17
dom_consumption  1.736100e-01
dom_cost        -2.765799e-02
dom_cpg         -4.857935e-02
int_consumption  2.506453e-01
int_cost         2.295330e-01
int_cpg          5.444143e-02
```

Briefly, interpretation based on the result:

- dom_consumption 1.736100e-01: For each standard deviation increase in domestic consumption, arr_delay is expected to increase by 0.1736 standard deviations, assuming other variables are held constant. This indicates a positive relationship with arr_delay.
- dom_cost -2.765799e-02: This coefficient suggests a small negative relationship with arr_delay. As the domestic cost increases by one standard deviation, arr_delay decreases by 0.0277 standard deviations.
- dom_cpg -4.857935e-02: The negative coefficient for domestic cost per gallon implies that as it increases, arr_delay slightly decreases, but the effect is relatively small.
- int_consumption 2.506453e-01: A positive coefficient indicates that an increase in international consumption is associated with an increase in arr_delay. The magnitude of the effect is somewhat more substantial than domestic consumption.
- int_cost 2.295330e-01: A positive coefficient here shows a significant positive relationship with arr_delay, suggesting that increases in international cost led to increases in arr_delay.
- int_cpg 5.444143e-02: The coefficient for international cost per gallon is positive, indicating a smaller yet positive effect on arr_delay compared to other factors.

Lasso Regression

```
> coef(lasso_model, s = best_lambda_lasso)
7 x 1 sparse Matrix of class "dgCMatrix"
                         s1
(Intercept)     -5.930008e-17
dom_consumption  4.678569e-01
dom_cost        -4.335997e-01
dom_cpg         -2.311335e-01
int_consumption  7.530255e-03
int_cost         5.684052e-01
int_cpg          3.225778e-01
```

Briefly, interpretation based on the result:

- dom_consumption and int_cost have much higher coefficients in the Lasso model, suggesting a more significant impact on arr_delay.
- The sign and general direction of the coefficients are consistent across both models.

Predictions and Mean Squared Error (MSE)

We include the MSE for both models, providing a measure of the model's accuracy. We may consider a lower MSE indicating a better fit to the data.

```
> print(paste("Ridge Regression MSE:", ridge_mse))
[1] "Ridge Regression MSE: 0.628688774418688"
> print(paste("Lasso Regression MSE:", lasso_mse))
[1] "Lasso Regression MSE: 0.621302260567014"
```

Briefly, interpret based on the results:

- Since the Lasso MSE (0.6213) is slightly lower than the Ridge MSE (0.6287), it suggests that the Lasso model might be fitting the data slightly better than the Ridge model.

The multiple linear regression model explained 37.8% of the variability in delays, emphasizing the significance of variables such as domestic consumption and cost. The random forest approach, known for its predictive accuracy, showed an explanatory power of 20.76%. Also, The performance of Lasso regression gives out higher coefficients for influential variables like domestic consumption and international cost. Predictive performance, measured by Mean Squared Error (MSE), favored lasso slightly over Ridge (MSE: Lasso 0.6213, Ridge 0.6287). These findings contribute valuable understanding of the relationship between fuel expenditure and arrival delays, guiding predictive modeling and understanding key contributing factors.

Conclusion - Positive/Negative results, recommendations, caveats/cautions.

- Impact of passenger numbers: Flight delays are largely influenced by both domestic and international passenger numbers. The linear regression models show a greater influence of international passengers on delays than that of domestic passengers. Under such a scenario, international flights are more influential when it comes to causing delays and thus airlines need to optimize this. It includes improving the boarding process and efficient use of resources.

-Fuel expenditure: The analysis shows a complicated relation between fuel usage, expenses, and flight delay. As far as we are concerned, airlines need to consider different fuel management approaches that can be used as replacements for delay prevention. For example, efficient use of fuel may result in less time spent on ground operations, and subsequently fewer delays.

-Economic factors: Airfare year and quarter reports reflect the pandemic and seasonal pressures. Under such circumstances, Airlines should reschedule capacity in the seasonal pattern, expecting demand fluctuations and reacting to limit delays.

-Airport-Specific Insights: data from Chicago O'Hare and Miami International Airports show insights into how local factors contribute to delays. As the situation cannot be thoroughly solved, airports should consider local factors such as airport layout, traffic flow, and operational challenges and develop tailored strategies.

The pandemic has reshaped the airline industry, demanding a reevaluation of operational strategies. By focusing on the key areas above aligned with advanced predictive modeling, airlines may effectively reduce delays.

Data Sources - Links, downloads, access information.

Datasets

(1) Bureau of Transportation Statistics (BTS): *Historical Airline Data, On-Time : Marketing Carrier On-Time Performance (Beginning January 2018)*
https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr

(2) Bureau of Transportation Statistics (BTS): *Passengers All U.S. Carriers - All Airports (Beginning January 2018)*
https://www.transtats.bts.gov/Data_Elements.aspx?Data=1

→ This dataset contains information on the performance of various marketing airlines and their flights for a given time period. This dataset allows you to select from a variety of numerical and descriptive features to download as a .csv file for further analysis.

(3) National Weather Service (NWS) & National Oceanic and Atmospheric Administration (NOAA): *Historical Weather Data, Climatological Data for CHICAGO O'Hare INTL AP, IL and MIAMI INTL AP, FL- Jan 2018 - Jan 2023*
https://www.weather.gov/wrh/Climate?wfo=lot

→ These datasets provide historical weather data for Chicago O'Hare International Airport and Miami International Airport. We have chosen to use precipitation data and temperature data in our analysis.

(4) Bureau of Transportation Statistics (BTS): *Airline Fuel Cost and Consumption (U.S. Carriers - Scheduled)*
https://www.transtats.bts.gov/fuel.asp?20=E

→ This dataset provides monthly historical fuel expenditure and consumption data for a variety of airline companies, beginning in January 2000.

Reference Data Tables

Unique Carrier Code	**Airline Carrier Name**
9E	Endeavor Air Inc.
AA	American Airlines Inc.

AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines LLC
F9	Frontier Airlines Inc.
MQ	Envoy Air
NK	Spirit Air Lines
OH	PSA Airlines Inc.
QX	Horizon Air
UA	United Air Lines Inc.
VX	Virgin America
WN	Southwest Airlines Co.
YV	Mesa Airlines Inc.
YX	Republic Airline

Main Dataframe

Column Name	Description	Data Type	Other Notes
year	Year of the data	Date	
month	The month of the data	Date	
carrier	Unique carrier code	String	
carrier_name	Name of the airline carrier	String	
airport	Airport code	String	
airport_name	Name and location of the airport	String	
arr_flights	Number of flights arriving	Numeric	
arr_del15	Number of flights delayed by 15 minutes or more	Numeric	
arr_cancelled	Number of flights cancelled	Numeric	
arr_diverted	Number of flights diverted	Numeric	

arr_delay	Total delay in minutes for arriving flights	Numeric	
carrier_delay	Delay attributed to the airline carrier in minutes	Numeric	
weather_delay	Delay attributed to weather conditions in minutes	Numeric	
nas_delay	Delay attributed to National Airspace System in mins	Numeric	NAS: National Airspace System
security_delay	Delay attributed to security issues in minutes	Numeric	
Late_aircraft_delay	Delay attributed to late aircraft arrivals in minutes	Date	
dummy_delayed	Binary variable indicating flight delay (1 or 0)	Numeric	Create a dummy variable for flight delay or not.(Delay=1,On time=0)
dom_passengers	Number of domestic passengers for a given airline	Numeric	
intl_passengers	Number of international passengers for a given airline	Numeric	
total_passengers	Total number of international passengers for a given airline	Numeric	

Fuel Dataframe

Column Name	Description	Data Type	Other Notes
dom_consumption	The amount of fuel consumed domestically, measured in million gallons.	Numeric	
dom_cost	The total cost of domestic fuel consumption, measured in million dollars.	Numeric	
dom_cpg	The cost per gallon of domestically consumed fuel, measured in dollars.	Numeric	
int_consumption	The amount of fuel consumed internationally, measured in million gallons.	Numeric	

int_cost	The total cost of internationally consumed fuel, measured in million dollars.	Numeric	
int_cpg	The cost per gallon of internationally consumed fuel, measured in dollars.	Numeric	
tot_consumption	The total amount of fuel consumed domestically and internationally combined, measured in million gallons.	Numeric	
tot_cost	The overall cost of both domestic and international fuel consumption, measured in million dollars.	Numeric	
tot_cpg	The overall cost per gallon of fuel consumption, considering both domestic and international consumption, measured in dollars.	Numeric	

Source code link:

https://github.com/DGyyyyy/Predictive-Analtics-Report/tree/main（In PA final folder)

Citations

[1] Kim, W. (2023, June 29). Why this summer is already shaping up to be a drag for travelers. Vox. Vox Media. https://www.vox.com/travel/23777457/summer-air-travel-delays-weather-canceled-delayed-flights. Accessed 1 December 2023

[2] Allon, G. (2023, July 23). How the internet screwed up air travel. Business Insider. Insider Inc.https://www.businessinsider.com/flight-delays-airline-fees-bad-service-online-booking-internet-travel-2023-7. Accessed 1 December 2023

[3] Bureau of Transportation Statistics. (2023, May 1). 2022 annual and 4th Quarter U.S. Airline Financial Data. 2022 Annual and 4th Quarter U.S. Airline Financial Data . U.S. DEPARTMENT OF TRANSPORTATION. https://www.bts.gov/newsroom/2022-annual-and-4th-quarter-us-airline-financial-data#:~:text=See%20the%20tables%20that%20accompany,7%2D12. Accessed 1 December 2023

[4] Freedman, J. (2023, June 9). Why do airlines overbook? the truth about overbooking. GetGoing. BCD Travel .

https://www.getgoing.com/blog/why-do-airlines-overbook/#:~:text=The%20point%20of%20all%20this,most%20flights%20are%20undoubtedly%20overbooked. Accessed 1 December 2023