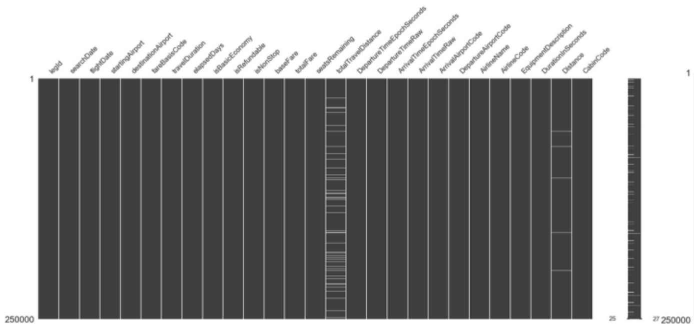
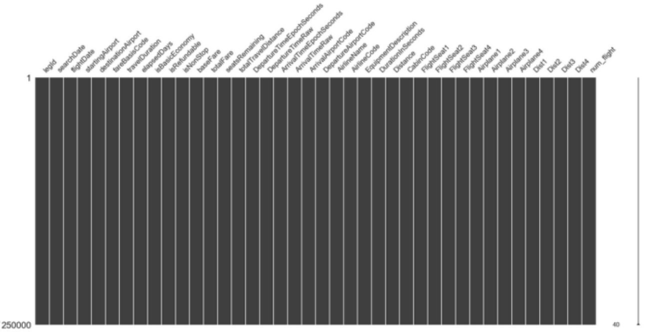
During the Covid-19 lockdown, air travel plummeted, and the airline industry lost a significant amount of revenue from the restrictions. In 2022, it was reported the industry lost around $220 billion worldwide. Only recently have the passenger count begun to recover to pre-pandemic levels, particularly the airlines in the United States are performing strongly. While the revenue is close to numbers achieved in 2019, we're not quite there yet; thus, we, as travel-data and analytics consulting company, were hired to analyze trends that could help increase airline revenue. Two competing potential strategies are to increase ticket prices or decrease ticket prices. The issue is if we increase ticket prices, we could increase revenue but harm passenger numbers; in contrast, we could decrease ticket prices to attract more passengers. The latter option seems more appealing to the airline stakeholders who would like to fill up as many seats as possible. Therefore, ticket prices decreasing in the upcoming weeks, and increasing the overall revenue in the US airline industry would be deemed a success. Although decreasing the ticket prices could bring in more customers, this could lead to the airlines spending more on amenities for flights, paying more for fuel given the higher weight loads, and risking the overall customer service declining. In situations where flights are delayed or canceled the airlines will deal with more complaints and this could potentially lead to a higher loss in profits.

Our stakeholder is the association, Airlines for America (A4A) – they represent major airlines in North America and are responsible for lobbying favorable regulations in taxation and competition. **Nicholas E. Calio** is the current CEO and President of the A4A, **Paul Archambeault** is the Chief Financial and Operating Officer, **Rebecca Spicer** is Senior Vice President of Communications. We will use our findings – a presentation at the conclusion – to advise to them the best course of actions, and they will in turn advise the airlines they represent in hopes of achieving their goals. Given the criteria, this will be a regression problem as the task will be to predict the prices of tickets.

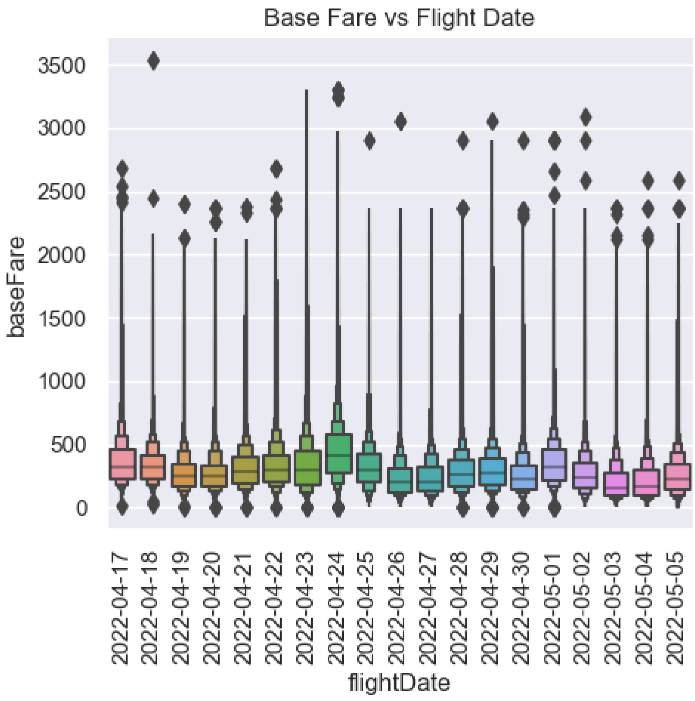
First step of the project is data wrangling, the step where we collect, organize, and clean our data. The dataset provided to us contains 250,000 rows of entries by 27 columns worth of features. The data has flight departures spanning from April 17th, 2022, to May 5th, 2022, with travel distance, and the fare for the tickets. Flights sometimes have layovers, meaning a passenger takes a flight to another airport and take another flight to get to their destination instead of flying to their destination directly. The data collected contains layover flights; however, the resulting data has certain columns with their data clumped up together as a result, so these will need to be separated into their own individual columns. Furthermore, imputation of missing data is also required on three of the columns. We will begin with the imputation step, which is filling any missing data. The three columns with missing data are: "EquipmentDescription", "totalTravelDistance", and "Distance". For the "EquipmentDescription" column, we took the most flown aircraft type used that for the missing data. The other two columns used the average distance of their respective columns for imputation. A visualization of the before and after imputation can be seen below, with the white horizontal lines representing missing data:

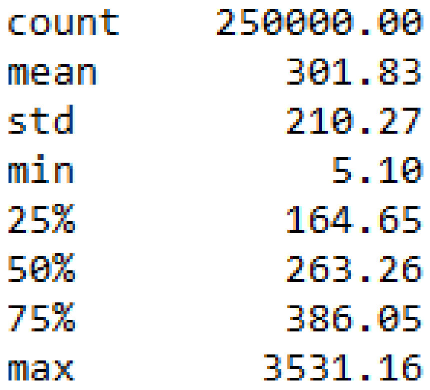




Once the imputation is completed and no columns is missing any data, we deal with the columns that has multiple entries grouped together. We simply iterate through all the entries and split on the "||" and the split data are given their own columns.

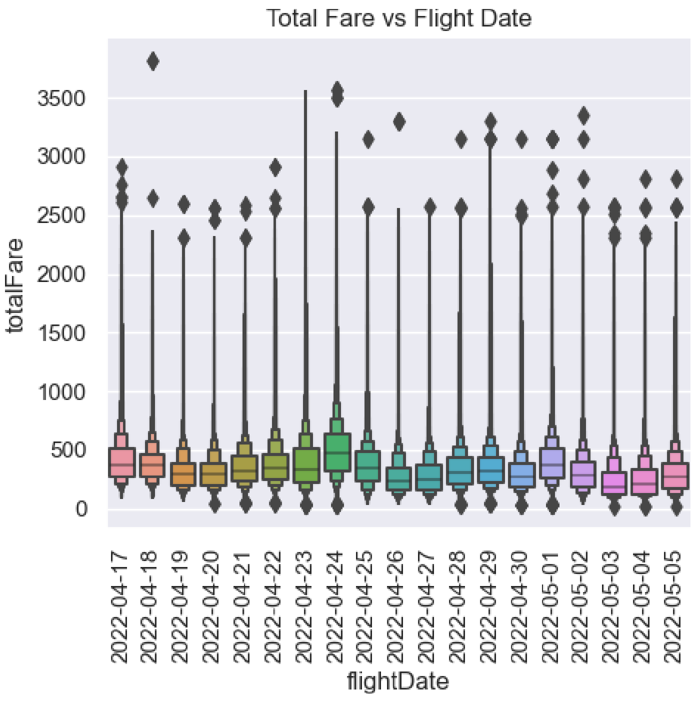
The data exploration stage will give us key insights about the underlying patterns of our data. The analysis mainly pertained to the fare of the tickets and the fluctuations between April 17th, 2022, to May 5th, 2022. An investigation on the base fare was carried out and then compared to the total fare, the total fare being the base fare plus taxes and any additional addon costs, to mark the differences between the two fares. The visualization of the base fares and total fares plotted against the date of the flights can be seen below:

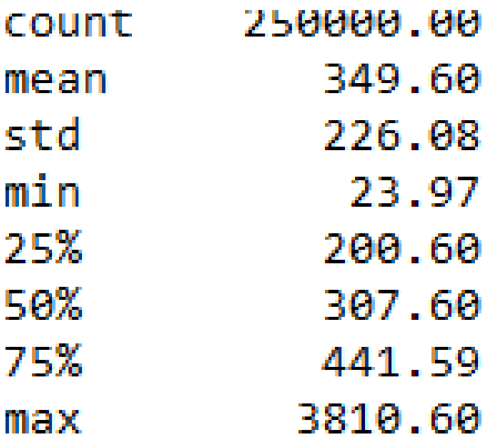




```
count    250000.00
mean        301.83
std         210.27
min           5.10
25%         164.65
50%         263.26
75%         386.05
max        3531.16
```

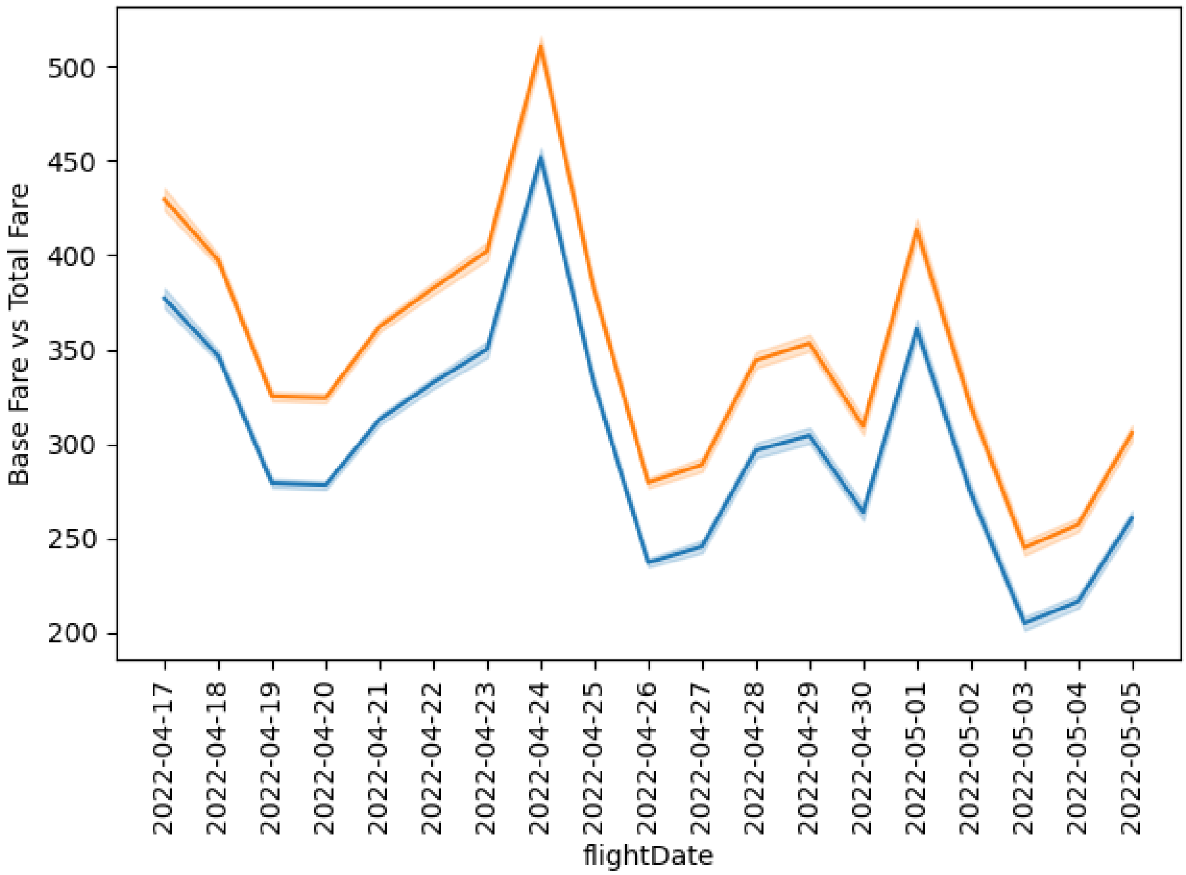
According to the analysis, the average base fare is $301.83, the lowest ticket cost being just $5.10 and the highest ticket cost being $3531. These are quite some outliars, but we don't want to remove any of them since we're dealing with various seating accommodations. Lets compare this to the total fare:



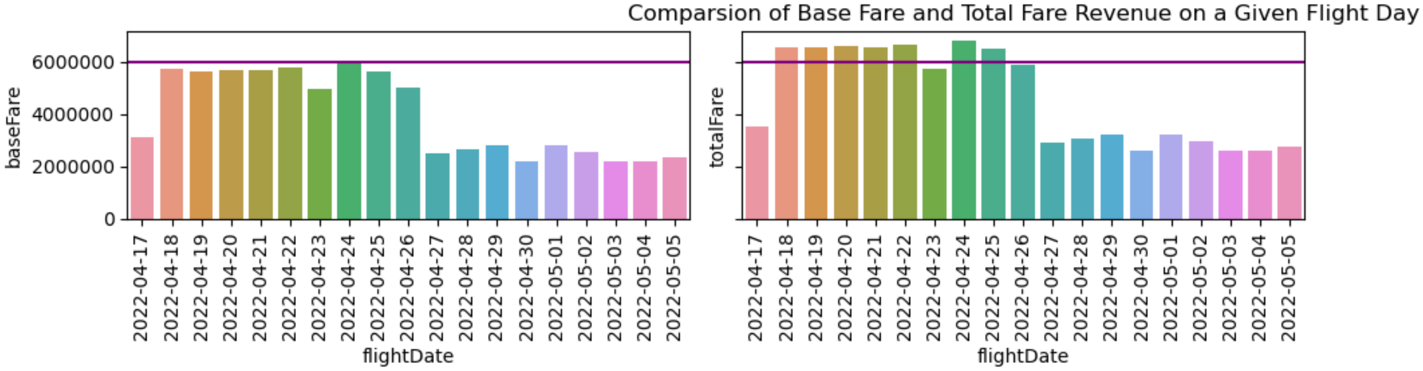


```
count    250000.00
mean        349.60
std         226.08
min          23.97
25%         200.60
50%         307.60
75%         441.59
max        3810.60
```

The math works out an average difference of $47.77 between the average base fare and total fare. The there is quite a difference between the min fare and max fare after the addon costs. The plot below shows the average base fare in the blue line, and the average total fare in the orange line. A trend does seem to show with the time series.

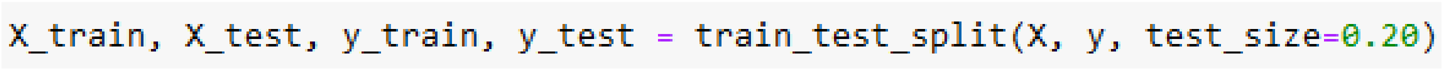


Here we’re conducting a hypothesis test to determine if the fare prices correlates with the days of the week, specifically if the prices are higher during the weekends, this will be our null hypothesis. The alternative hypothesis will simply be if the fare prices are not higher during the weekends. We already have the visualization above to see the date with the highest peak (highest fares) was on April 24th, 2024, with the next highest on May 01st, 2022. The two dates were a Saturday and Sunday respectively, does that does give initial validation that the null hypothesis is correct.


Comparsion of Base Fare and Total Fare Revenue on a Given Flight Day

The plot above shows the revenue generated on a given flight date, and the purple line indicates the $6 million revenue point. Without the addition revenue from the addon costs; the total revenue of the base fare never crosses the $6 million threshold; however, with the additional costs from total fare, it does seem the airlines crosses this threshold regularly.

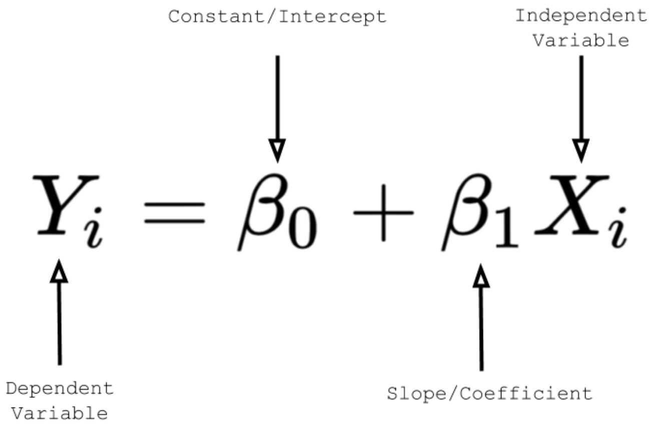
The third step in the data science pipeline, once we’ve concluded our data exploration, is the preprocessing data – preparing the data to feed into machine learning models. The main task here is converting any categorical data into numerical data since machine learning algorithms only accepts numerical values. The encoder utilized here are the dummy encoder and the label encoder; the dummy encoder was used to convert features with low category varieties such as the “isBasicEconomy”, “isRefundable” and “isNonStop”, basically features that can be represented through Boolean. The label encoder was used for categorical features with high cardinality, which avoids generating too many new columns which increases the dimensionality, by give each unique category a value to represent them. After all the categorical features have been encoded, the data can then be split into independent “X” and dependent “y” variables. The independent variable will include everything but the “totalFare” column since this is the feature the model will attempt to predict, so the “totalFare” column is stored on the “y” variable. The variables will be farther be split into 80% of the data being used for training and the remaining 20% used for testing.



```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)
```

The final preprocessing task is to scale the “X_train” and “X_test” variables to reduce the impact of outliers by scaling the data into a standard normal distribution. The data is finally ready to be used for machine learning.

Given the task is a regression problem to predict ticket prices, the first model we will utilize is the Linear Regression model. The model works by predicting the target variable through using the values of another variable. It creates a line of best fit that minimizing the difference between the predicted value and actual values. The metric used to measure the success of our model will be R-square which measures the variance for the dependent variable. We will also use root-mean-squared-error (RMSE) which is the sum of the difference between the predicted and actual values, squared. Then the mean is taken, and we take the square root. Starting, we fitted the “X_training” and “y_training” data into our linear regression model. The fitted linear regression model is used to predict our “y” value (total fare) by passing it our “X_testing” data. After our predicted “y” values is returned, the predicted values are compared to the actual “y” values for R-square and RMSE. The R-square value for the base linear regression model was 0.9999850897515424 and the RMSE value was 0.8755734132742768. The highest score achievable with R-square is 1, and the score from the linear regression model is as close of a score possible to the limit. Perhaps the model overfitted, if this is the situation then we’ll need to add penalty to the linear regression model. Regularization is a technique that penalties large coefficients (the slope) since large coefficients leads to overfitting.



$$Y_i = \beta_0 + \beta_1 X_i$$

Two regularization methods are Ridge Regression and Lasso Regression – Ridge takes the squared value of the coefficient, while Lasso takes absolute value of the coefficient. A third method, called the Elastic Net, is a hybrid of Ridge and Lasso which takes both the absolute value and squared value of the coefficient. Shown below is the R-square and RMSE values for the three methods.

Ridge Regression

Iteration	RMSE	R-Square
1	0.00218557	1.00000000
2	0.00219641	1.00000000
3	0.00214804	1.00000000
4	0.0021553	1.00000000
5	0.00216584	1.00000000
Average	0.002170232	1.000000000

Lasso

Iteration	RMSE	R-Square
1	1.07676125	0.9999769
2	1.06427845	0.99997762
3	1.07141032	0.99997708
4	1.06376864	0.99997723
5	1.11169273	0.99997738
Average	1.077582278	0. 999977242

Elastic Net

Iteration	RMSE	R-Squared
1	66.22040541	0.91691644
2	66.93095756	0.91657176
3	64.78928854	0.91478787
4	65.71959101	0.91600541
5	64.35291631	0.91419513
Average	65.602631766	0.915695322

The tables above show the results of the three models performed on cross validation of five folds. We see that the Ridge Regression model most likely overfitted given its high R-Square values, similarly to Lasso which has a score very close to 1 and both models have low RMSE values as well. The RMSE is essentially tell us there is about $1 or less the airline stakeholders could change about their ticket fares. Elastic Net looked more promising; it still has a high R-Square value, but the RMSE tells is there is about a $65.60 in the average price of a ticket fare that could be adjusted. Overall, we would go for the Elastic Net Model given the fear of overfitting with the other two models.