MISM3515 Final Report
Airfare Pricing Exploration

Team 5
Jack Krolik, Daniel Han, Lidong He, Quyen Dao
12/5/2021

Business Question:

Air travel is an essential form of transportation for anyone who wishes to travel between large distances in a quick and efficient manner. This has been demonstrated by the sheer number of individuals who have traveled using this mode of transportation. For instance, in 2019, more than 4.5 billion passengers were reported boarding planes by the global airline industry (Mazareanu). Of those traveling in 2019, 57% of individuals over the age of 70 and between the ages of 30 and 49 reported that they had been on a plane more than once during the year (Kunst). While 55% of individuals between the ages of 18 and 29 reported traveling by plane more than once (Kunst). These age groups make up the largest share of air travel mainly due to either leisure, business, or school-related purposes. While more people have used air travel over recent years to get from place to place. It is interesting to note that air travel costs have in fact decreased over time (McCartney) due to its highly regulated industry by the government and by states. For example, in 2000 the average airfare was about $524 adjusted for inflation while in 2018 that number has fallen to $371, a 26% decrease (Annual US Domestic Average Itinerary Fare in Current and Constant Dollars). This also happened because of new aircraft technology making planes more efficient. Although airfares have decreased in price, there is still no easy way to compare prices in a relatively easy and trustworthy manner. Sites like Kayak.com use data from airlines, but their incentives are completely misaligned. This is because the airlines pay to have their fares listed which is how they make money which presents a conflict of interest between the consumers and airlines. Airlines therefore will charge consumers a higher price due to the fact that they have to pay extra to aggregator sites like Kayak.com. Therefore, with our project, we wanted to answer the following questions to help consumers gauge their airfares.

How are airfares priced, and is there an optimal time and location to purchase them?

As there are many different variables to be addressed in the calculation of airfares, we want to make the data transparent in order for us and others to understand it to the best of their ability. The airfare in this analysis is based on the fare per person which represents the dollar amount that the ticket was purchased given an itinerary of the airfare cost.

How do different airports and airlines differ in price and what is their travel demographic?

Due to the fact that many students and business professionals travel during the year, we want to understand what places are the busiest and how their airfares differ in different locations. This portion of analysis will be useful when looking for the airport or airline that not only offers the best service, but price as well. We want to understand whether there is certain geography that is more optimal for airfares than others and how they are different.

We hope that by using our analysis, business and regular personnel will have a better understanding of what variables impact the price of their airfare tickets and whether there are better options available in order to make a more prudent decision.

Dataset:

The dataset is from the Bureau of Transportation Statistics (BTS) which is part of the US Department of Transportation that organizes, analyzes, and makes information available on domestic transportation systems. They also improve the quality of reported statistical data through research, updated guidelines, and the promotion of data acquisition strategies. Due to the fact that it is part of a governmental organization, we expect that it is fully reliable and there is little chance of bias or statistical corruption. The data we gathered, in particular, comes from a BTS survey called the *Airline Origin and Destination Survey* which is a sample of airline ticket data from different airlines collected each quarter by the Office of Airline Information of the Bureau of Transportation Statistics. The survey consists of three different tables: Coupon, Market, and Ticket survey. These surveys are quite similar yet differ in some of the variables they include. For instance, the Coupon survey does not include the airfare of the ticket but does include a variable for Fare Class, although it was not recommended to use this for data analysis. In addition, the Market survey includes variables about the market fare of the ticket, but when digging deeper into the dataset, it was found that the variables had an extremely poor correlation to each other, and the models were predicting results that were far too inaccurate given a testing sample. Finally, there was the Ticket survey which included the airfare and all relevant variables that were needed for our analysis. The variables included were the FPP, which is the price of the ticket per person. This would be our Y variable for our analysis. For all other variables, please see the appendix.

Data Formatting:

There was a lot that went into the formatting of our dataset to get it ready for analysis. First, once the files were all downloaded, the columns such as 'Quarter' had to be adjusted so that they were qualitative variables rather than quantitative variables. To do this, the quarters were each named 'Quarter 1/2/3/4' depending on which quarter the ticket was from. Each file was a different quarter and therefore had to be aggregated to create one large data frame consisting of all the quarters. Being that the dataset was so large, each year was roughly 2.5 million rows long, there was an inability to do analysis for more than one year. However, it would have been helpful to train the model in different years as we could how prices on a yearly basis. Although, we were able to use these quarters (1-4) as a representation of time/seasonality which was sufficient. The next step once the data was in one file was to drop all unnecessary rows/columns. To do this, we dropped the 'Bulk_Fare' column as it only had one unique value and all rows where the 'Dollar_Cred' column was 0. The Dollar_Cred column signified whether the airfare was in dollars and if it was credible or not. Since we wanted our airfares to be in dollars and the data to be accurate, we dropped these respective rows. In addition, because the dataset was so large, to make it more efficient we wanted to drop all rows where the airport and airlines were not significant enough to our analysis. As such, we found out what airlines were most popular and filtered them by the number of times they occurred in the original dataset. If they occurred less than 100,000 times, they were dropped resulting in the top 12 airlines (See Appendix for more

details). We did the same thing for the airport locations in which if it occurred more than 220,000 times resulting in the top 26 airlines (See Appendix for more details). To further narrow down the data, we looked at histograms of the FPP, Passenger, and Distance columns via pyplot and calculated the mean and standard deviations to determine what rows to remove based on whether they were outliers. We defined outliers as being more than three standard deviations from the mean of the respective column data. Therefore, by removing the outliers from these columns, we were able to see via histograms that the data was much less skewed (see the Appendix for details).

Key Findings:

As our dataset was so large, Python was the best option to format our data in an efficient manner. In terms of our analysis, we used various regression models through Python's pandas libraries. This includes Linear Regression, Cross-Validation K-Fold, SelectKBest (f-regression), KNN, Decision Trees (regression), and Random Forest (regressor). These were the main regression models that we covered during the semester and therefore were the ones we used. Unfortunately, due to the size of the data, we are unable to run grid search, yet we used SelectKBest as a proxy for selecting important variables as discussed below. We used FPP as our y variable while the rest of the variables were used as x variables (See Appendix for details). We also used a test size of .3 and a random state of 1. Further, due to the fact that the data was not equal in scale, one passenger was not equivalent to one dollar, for instance, we had to scale the data by running MinMaxScaler before completing any modeling. In addition to our models, we used the matplotlib libraries in order to illustrate visualizations of our analysis in a concise manner. This helps our audience receive the information faster and in a simpler fashion.

Linear Regression

We first conducted a linear regression model to find out exactly the relationship between our Y variable (FPP, or fare per person), and our X variables. As linear regression is used to "predict" the value of the Y variable by basing it off the value of another, we used linear regression to find connections and exemplify the relationship between FPP and the other variables. After conducting the model, we received outputs of an RMSE of 141.526 and an R^2 score of 0.687. While this score isn't exceptionally high, we were able to pull from the model the conclusion that there is some linear relationship between FPP and our x variables.

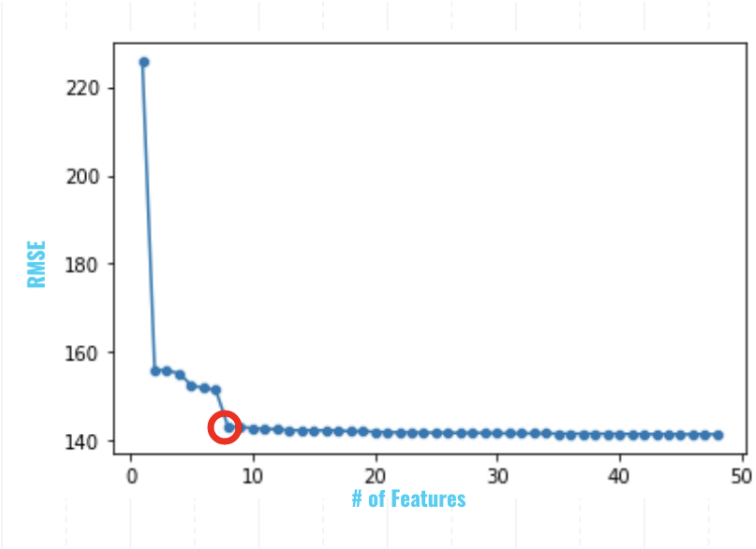
K-Fold Cross-Validation

Our cross-validation analysis was useful for determining how accurate our linear regression model was. This is because it is possible that based on the randomization of the testing sample and training sample, the model could be either more accurate or less accurate than in aggregate. Therefore, by using K-fold cross-validation we were able to feel confident that our model was consistent with its outputs regarding the r^2 and rmse. To do this, we used cross-validation on linear regression using ten splits as the dataset was very large and we did not have the bandwidth to run more but felt confident that ten was sufficient once we saw the results. As such, the mean

of the r^2 was .683, while the mean of the rmse was 141.84. This is fairly strong given that the data was heavily skewed in the beginning and while we did condense it, it is still somewhat tumultuous. In addition, when considering the standard deviations of both the r^2 (3.57) and rmse (.01795), neither output lies further than 3 standard deviations away from the mean (see Appendix for graphs). With this optimal linear regression model, we hope that our audience (leisure travelers, business personnel, and students) will be able to take advantage of the transparency in our pricing model and compare it to what they see online.

SelectKBest

For our SelectKBest model, we created a for-loop over a linear regression model to determine the best variables to select. In total there were forty-eight features (including dummy variables) and the model ran on every single one while creating a list of the root mean squared error each time. Next, we graphed the outputs using a scatter plot, as shown below and determined that nine features were the most optimal. By running the BestKSelect model again on nine features, we were able to analyze that the features selected were *Coupons, Round_Trip, Itin_Yield, Distance, Miles_Flown, Total_Fare, Passengers, Reporting_Carrier_NK* (Frontier Airlines)*, Reporting_Carrier_F9* (Spirit Airlines). Based on this analysis, a few interesting points are important to note. First, no 'Quarter' column was selected even though in our original hypothesis time was an extremely important variable to the outcome of the airfare given the seasonality of the airline industry. Secondly, the two airlines that were selected ('Reporting_Carrier') are both discount airlines and are the 8th & 9th most popular airlines by frequency in the dataset. On the survey, this is quite confusing given the fact that there are airlines that occur much more frequently in the dataset and therefore should be important when it comes to airfares. Thirdly, no airport code variable ('ORIGIN') was included in the feature analysis. This means that the origin of a flight is mostly irrelevant to its airfare which seems confusing on the surface given that more urban markets should seem to have higher prices. However, in hindsight, the price of an airfare should not depend on the origin, rather what is more important is the distance of the flight given that there are so many possibilities in which the airfare could be depending on its origin. In addition, upon further analysis, we were able to see just how exactly BestKSelect determines which features are most important given a dataset. We did this by analyzing the correlation of the x variables to the y variable, FPP. As shown below, the most correlated variables were selected from the dataset no matter if they were positive or negative. As the reason for the feature selection of both discount airlines were confusing at first, by digging deeper we were able to see that the airfares of these airlines were heavily negatively correlated. This would make sense given that they are discount airlines, yet it is interesting to note their relationship.



	FPP
COUPONS	0.273356
ROUNDTRIP	0.327553
ITIN_YIELD	0.451897
PASSENGERS	-0.17221
DISTANCE	0.357673
MILES_FLOWN	0.356688
TOTAL_FARE	0.305443

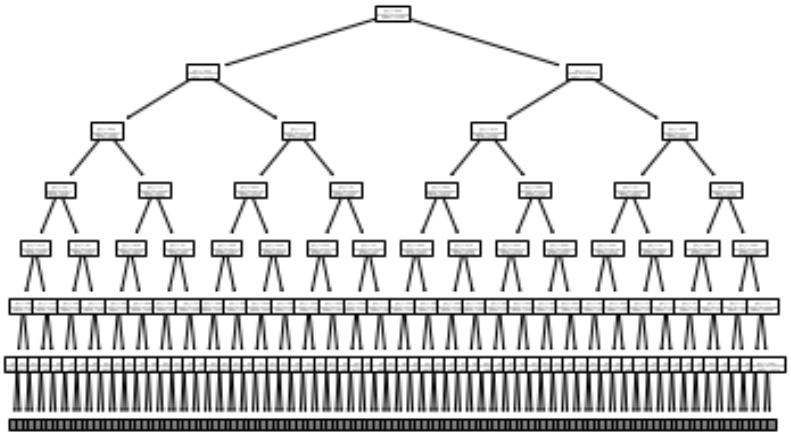
REPORTING_CARRIER_AA	0.043127
REPORTING_CARRIER_AS	0.00936422
REPORTING_CARRIER_B6	-0.00599862
REPORTING_CARRIER_DL	0.124258
REPORTING_CARRIER_F9	-0.159457
REPORTING_CARRIER_NK	-0.204842
REPORTING_CARRIER_OO	-0.00463334
REPORTING_CARRIER_UA	0.0978129
REPORTING_CARRIER_WN	-0.108823
REPORTING_CARRIER_YV	0.0151098
REPORTING_CARRIER_YX	-0.00266671

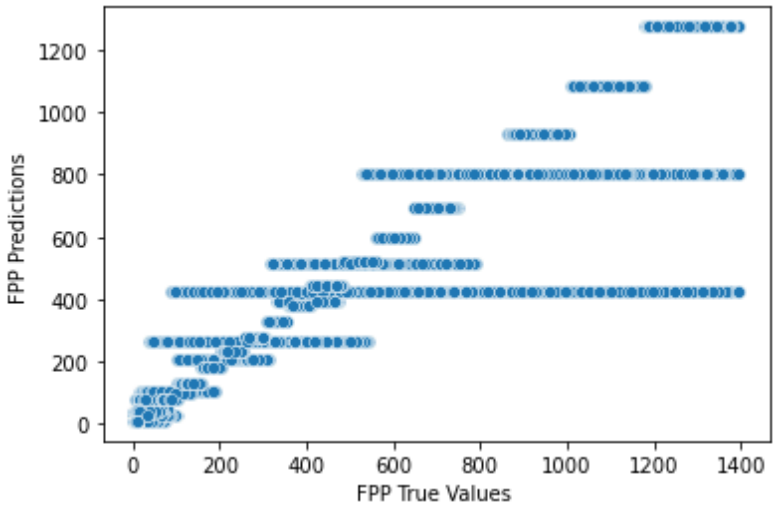
Decision Tree

- y-variable: FPP (Fare Per Person)
- x-variables: DISTANCE, MILES_FLOWN, TOTAL_FARE, PASSENGERS, ITIN_YIELD

Because we are working on continuous variables, we decided to drop all the binary variables as well as string variables and use the decision tree regressor, and split the data into subsets for having better outcome predictions. We also did not scale the data for the decision tree method as we thought that it would not be necessary.

Without determining the depth of the decision tree, the outcome was a highly positively correlated relationship between the x and y variables. The rmse and the r_square are so high that the model was overfitting. We decided to reduce the number of depths to 5 from 11 so that it resulted in the rmse of 61.305 and the r_square of 94.13%, not too high rmse over a thousand and a fairly good prediction of the model.



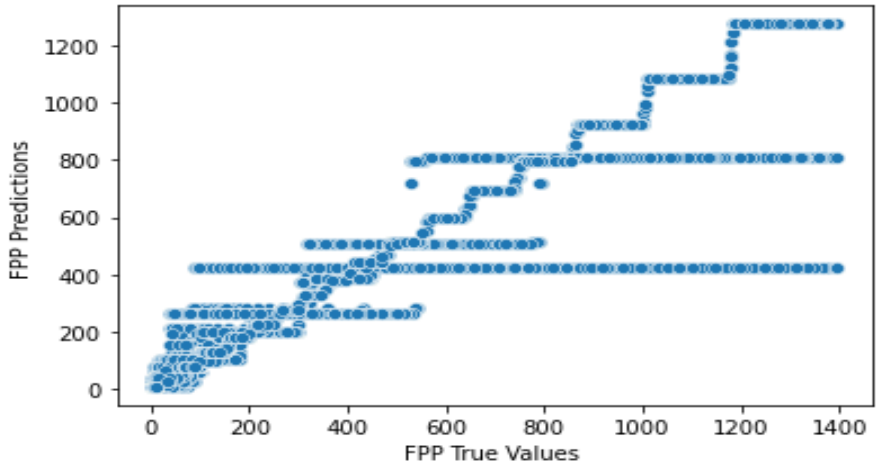


Random Forest

- y-variable: FPP (Fare Per Person)
- x-variables: DISTANCE, MILES_FLOWN, TOTAaL_FARE, PASSENGERS, ITIN_YIELD

We used the same variables for the random forest, dropping all the binary as well as string variables to have a better prediction of the training data. We decided to use the random forest method because it could help predict the fare price per person better using a group of decision trees.

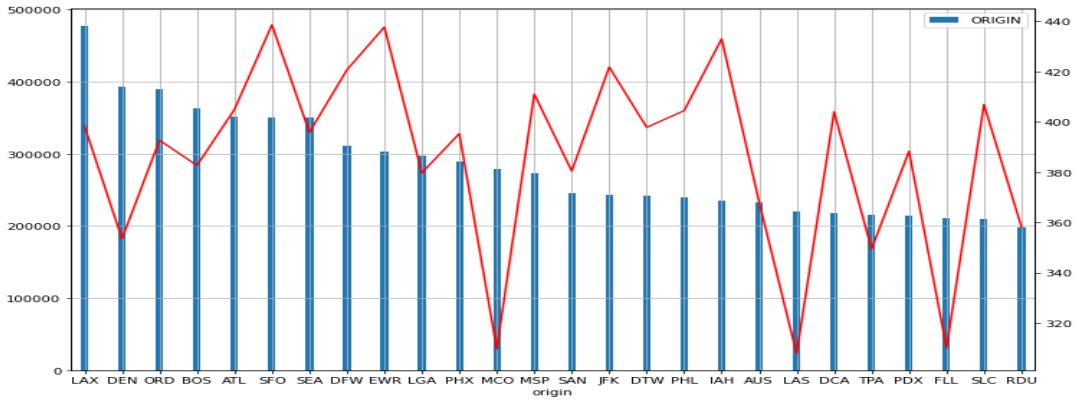
Keeping the same maximum depth of 5, the result was slightly better than the decision tree method. The model's rmse was 59.97 and the r_square was 94.38% fitting the data. Overall, the random forest gave us good metrics of model performance.



Visualizations:

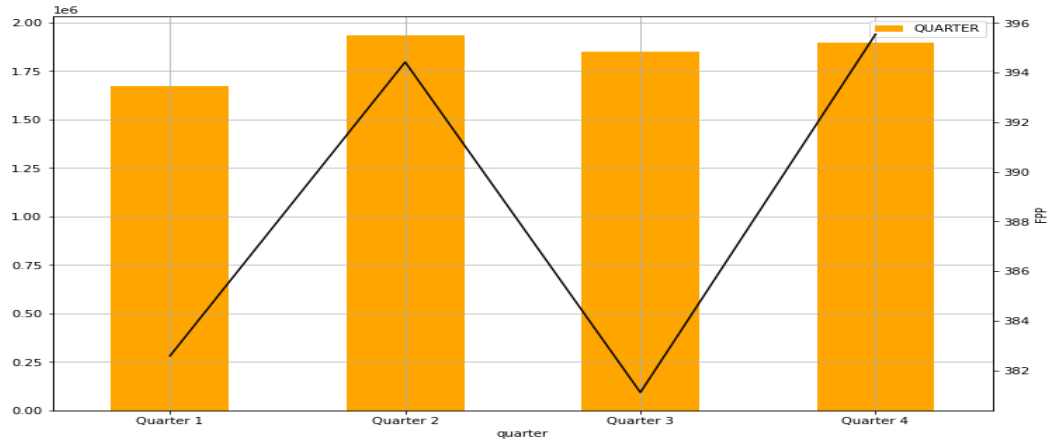
FPP and originals:

In the beginning, our group believed that more frequent flights in original places would have a massive impact on the ticket prices in the following two aspects. First, the most frequently original site will probably have many different tickets to the same places. Thus, those tickets will form a competition situation, and the prices will decrease a lot. Second, we believe that more frequent original sites will lower the fixed cost to create a scale of the economy by attracting more customers to go to other places from more frequent places. Thus, reducing the prices to attract customers is the most effective way. However, as the visualization shows, ticket prices in the most frequent places do not have the least prices. Thus, we conclude that the average FPP and the average number of flights from the original place in 2019 do not have any relationship.



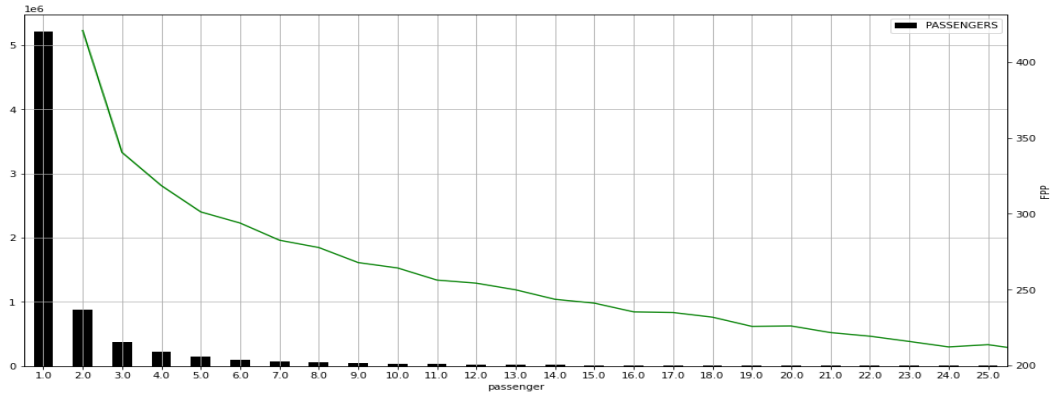
FPP and Quarters:

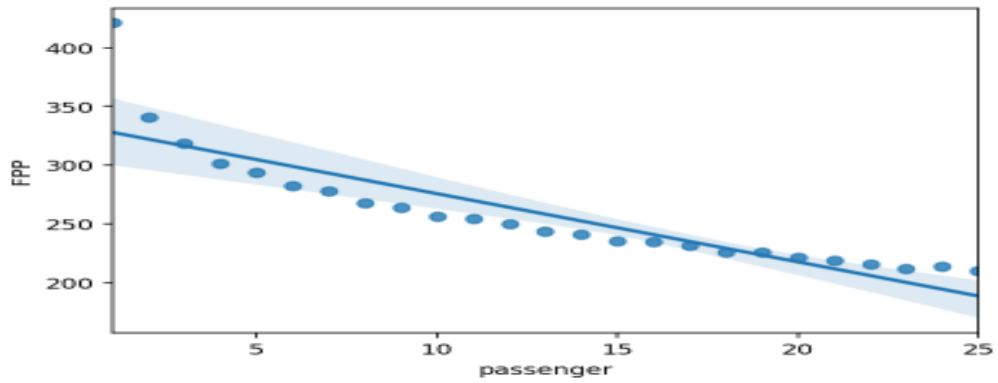
The Average FPP will be influenced by the quarters significantly. This situation is because there are summer and winter vacations in quarters 2 and 4. During these two quarters, many employees or students will have time to travel around the world or return home from working places or universities. Therefore, the demand for these two quarters will increase significantly, but the supply of flights will not change. Thus, quarters 2 and 4 will be more expensive than quarters 1 and 3.



FPP and the Average number of passengers:

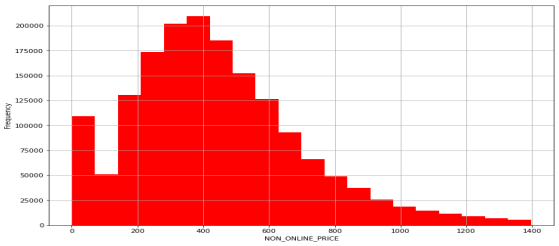
Company team travel or family travel is a common situation that there are many people buy the ticket together. However, flights alone are another common situation for some businessmen. Thus, our group wants to figure out whether more people buy the same flight ticket, the ticket prices will decrease a lot. Through the appendix graph, we find that the number of people taking the same flight has negatively correlated with FPP, and except the passenger travel along(which is the outline of the regression line), every one passenger adds to the same flights, the FPP will decrease. Our group believes that when people travel alone, he is probably a businessman. Thus, he always pays attention to the time rather than ticket prices. And they can afford the higher prices ticket like business class. But for the family or company team pays attention to the cost and always booked the ticket in advance.





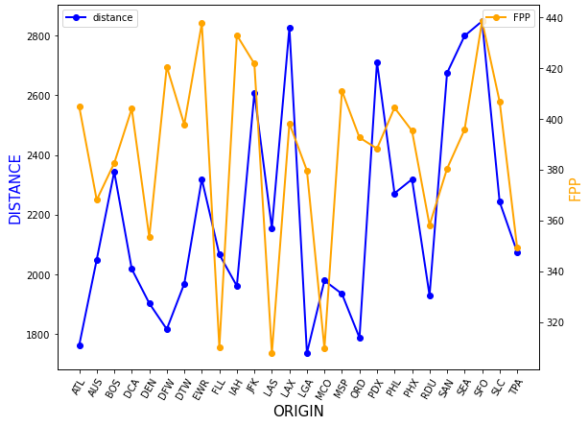
FPP and online orders:

Right now, ordering a flight ticket online is a very convenient way. Thus, our group is trying to find whether ordering the tickets online will have lower prices than not ordering online. Through the appendix graph, we can find that ordering online can not make sure you have a cheaper ticket than not_online order every time. However, we find that the online order ticket will have lower prices than non-online orders most of the time. As the graph shows, the most frequently online order ticket prices are between 140-340, but for not-online orders, the ticker prices are between 270-470.



FPP and distance in a different area:

Finally, we find that using the average FPP and average distance in different areas will help us find the most cost-effective original prices. Because we find that even the similar distance flights, the ticket prices will change dramatically by flying from different original places. For example, when you try to fly to 2800 miles, the prices in SFO are much higher than LAX. Thus, if people want to save money, they can consider flying from LAX rather than SFO.



Recommendations:

To summarize our analysis, we were able to identify a few key relationships that affect airfare prices. Firstly, the relationship between the airfare and the number of passengers. This is because individuals who purchase in bulk receive discounts or are buying cheaper seats given that they are families. On the other hand, business personnel who mostly purchase one ticket are buying more expensive seats. Additionally, it is buying online can drastically reduce the airfare because in-person tickets are generally last minute and the airfares are therefore higher. Furthermore, we learned that while the 'Quarter' column was not selected in our model for feature selection, it has a very strong correlation demonstrated by our visualizations. Due to the fact that tickets are much more expensive during the summer and winter, given that many people travel during the summer for vacation and the winter for holidays. As demonstrated in our analysis, there is extremely high variability in airfare prices and it is very difficult to predict them. This is mainly in regards to the fact that airfare prices are mostly idiosyncratic, meaning they change dynamically as the flight date approaches and it can be very difficult to predict given we were only able to receive the quarter as an input for time. However, our best model, Random Forest, did return strong results and we hope that our audience will be able to use it as a proxy for their analysis. Although, we recommend that it be used more as a benchmark rather than an accurate representation of what the data should be like. Otherwise, we hope that our analysis will aid in the effort to make airfare prices more transparent to all.

Challenges:

One of the main challenges we faced was that the size of the dataset was too large, with millions of rows, which caused some issues when we were trying to run the models. For example, when we tried to run Grid Search to look for the best parameters as well as tried to tune the model, it was unsuccessful because the data took endless amounts of time and our laptops could not run them. The other challenge associated with the data size was that at the first step of formatting the data, it took lots of time to download and figure out ways to run the models in an efficient manner. We also struggled with the numerous airlines and airports from the dataset so we compressed the data to only using the top airlines and airports. It would have been useful for our analysis if we could use all the airlines and airports listed. The other challenge that we encountered was the esoteric nature of the dataset. Having gathered the various surveys made it more difficult to know the best options and we potentially missed the important data points. We were clueless about the information gathered and whether or not it only skewed toward certain groups of people or social class. There were a lot more variables and it was difficult to spot

which ones would be helpful in the analysis to select and download. Their definitions are complex and confusing as well. Moreover, the dataset is highly skewed. We had to figure out which variables contained outliers to get rid of them that would have heavily affected airfare prediction. There is ample reason to believe that we were able to mitigate this problem to the best of our ability. Some data was invalid such as the Dollar Cred variable. The Bulk_Fare variable was useless for our analysis, yet it might have been important in other circumstances and other business questions. Although, aggregating the data together would have created repetitions in the dataset making it harder to analyze and most variables that were present in the other surveys were also present in the one we used for analysis. Based on our ability to address the issues that came with our dataset, we believe that it was suitable to answer our business question and gave us all the relevant information we needed from our analysis. If we were to complete the analysis again, we would use multiple years, given that only 2019 was used this time. We would choose a less complex and unique enough the analysis would be merited. Lastly, we would ensure that the dataset would not be too large to handle and run our analysis.

Appendix:

Variables and definitions:

FPP [Y VARIABLE]	Price of ticket per person
TOTAL_FARE	Total ticket price
ORIGIN	Airport Code of the origin airport
COUPONS	Number of coupons in the itinerary
ONLINE	Whether the ticket was sold online
ITIN_YIELD	Fare per mile (miles flown)
REPORTING_CARRIER	Airline Code
PASSENGERS	Number of passengers
DISTANCE	Distance of itinerary (includes ground transport)
MILES_FLOWN	Itinerary Miles Flown (track miles)
QUARTER	Quarter of the year

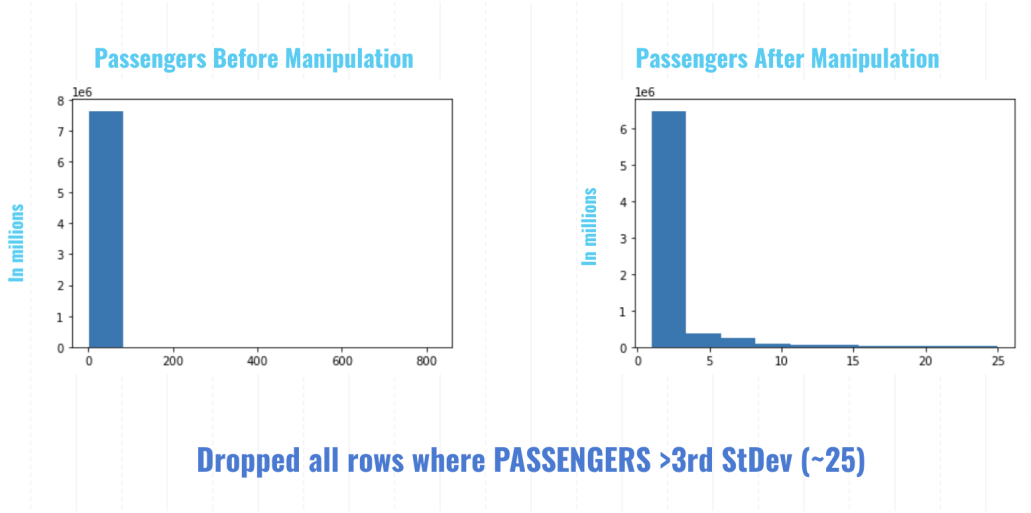
Airlines used for analysis (airline codes)

Index	airline	REPORTING_CARRIER
0	AA	1545865
1	DL	1536245
2	UA	1176456
3	WN	1014920
4	AS	408844
5	B6	345760
6	OO	344494
7	NK	332590
8	YX	245673
9	F9	174482
10	9E	128300
11	YV	98681

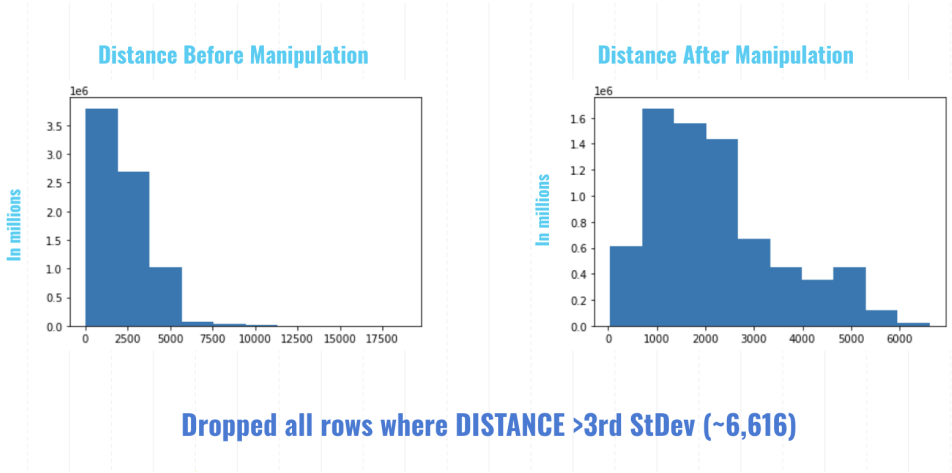
Airports used for analysis (airport codes)

Index	airport	ORIGIN
0	LAX	477846
1	DEN	393455
2	ORD	389773
3	BOS	362832
4	ATL	351699
5	SFO	350957
6	SEA	350018
7	DFW	311439
8	EWR	303286
9	LGA	297014
10	PHX	289845
11	MCO	278975
12	MSP	272735
13	SAN	245173
14	JFK	242870
15	DTW	241752
16	PHL	239426
17	IAH	234712
18	AUS	233218
19	LAS	220709
20	DCA	217896
21	TPA	215509
22	PDX	213879
23	FLL	210376
24	SLC	209259
25	RDU	197657

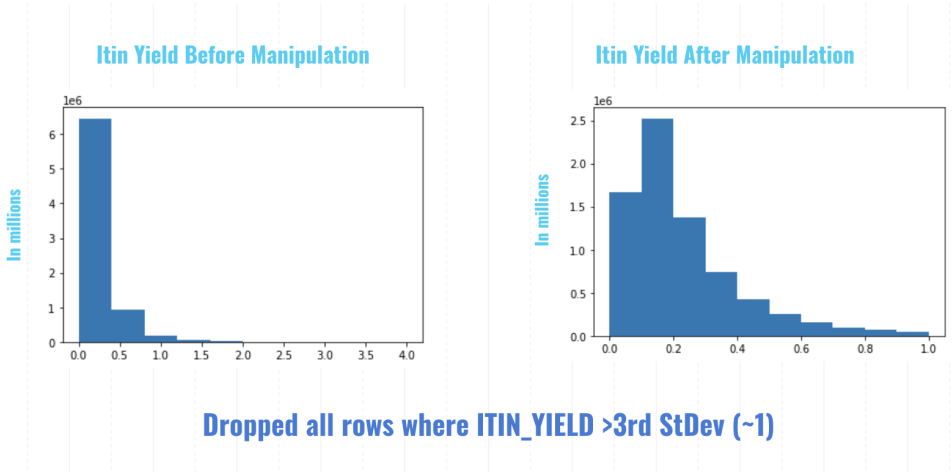
Passenger outlier manipulation



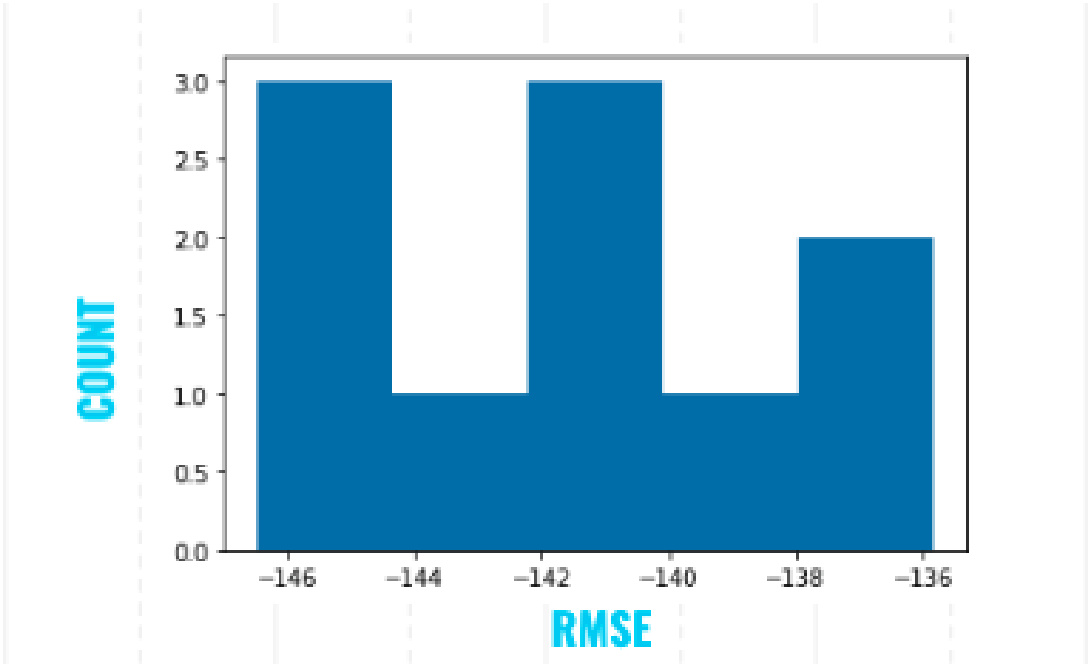
Distance outlier manipulation



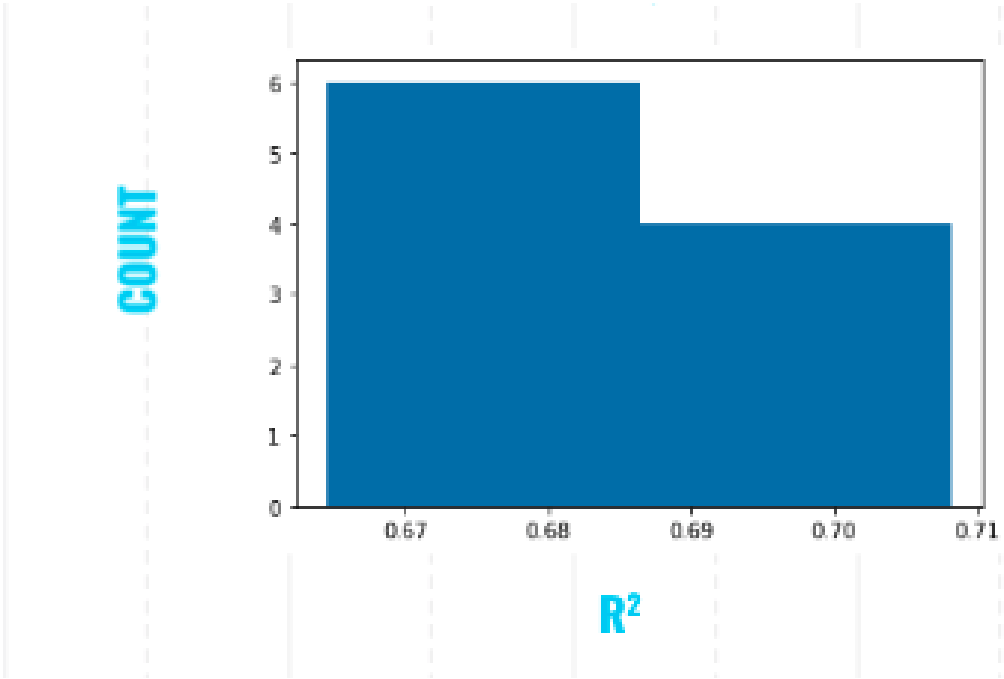
Itin Yield outlier manipulation



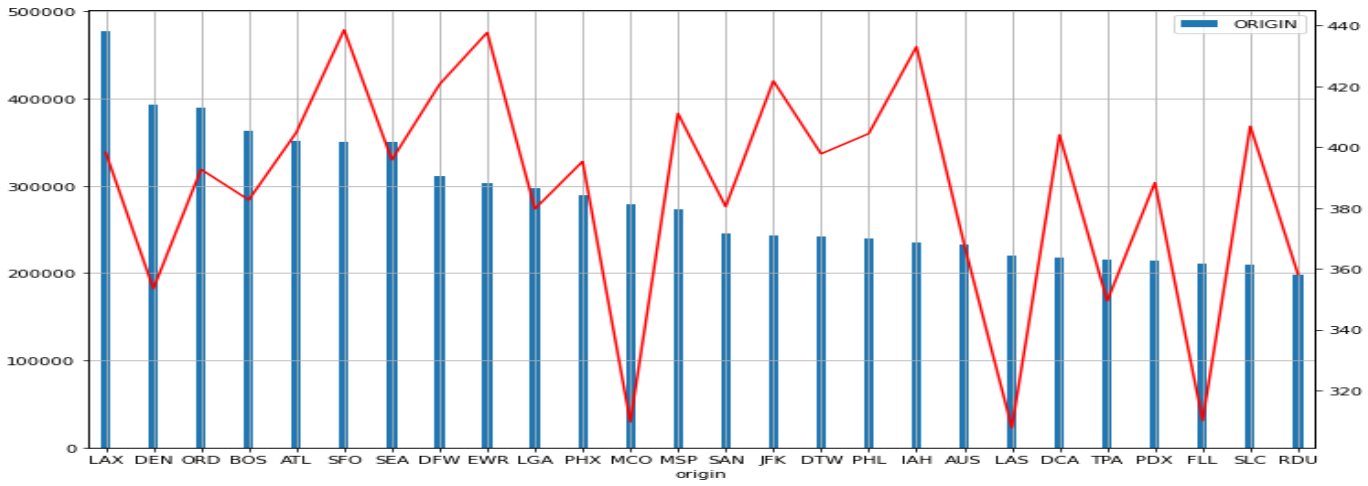
Outputs of rmse using cross-validation



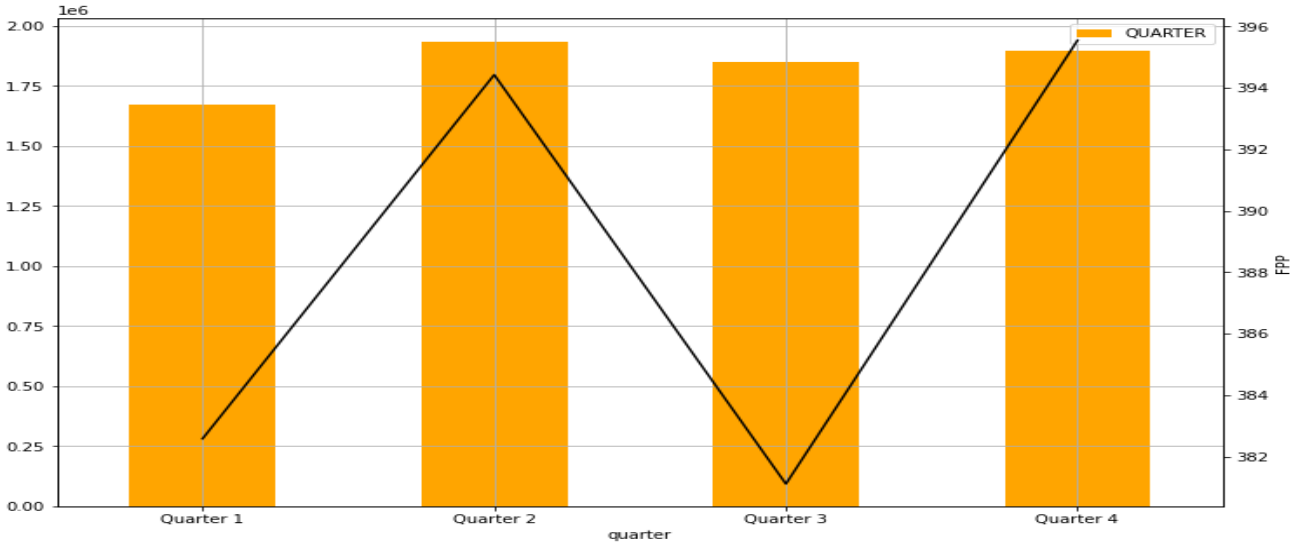
Outputs of r^2 using cross validation



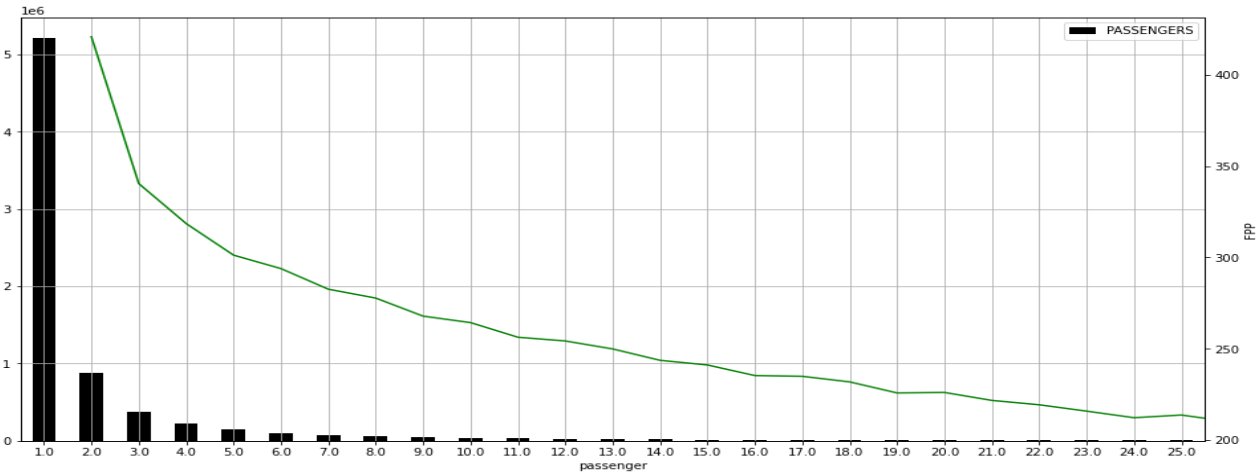
Bar and Line chart:relationship between Average FPP and average number of flight from the origin place in 2019

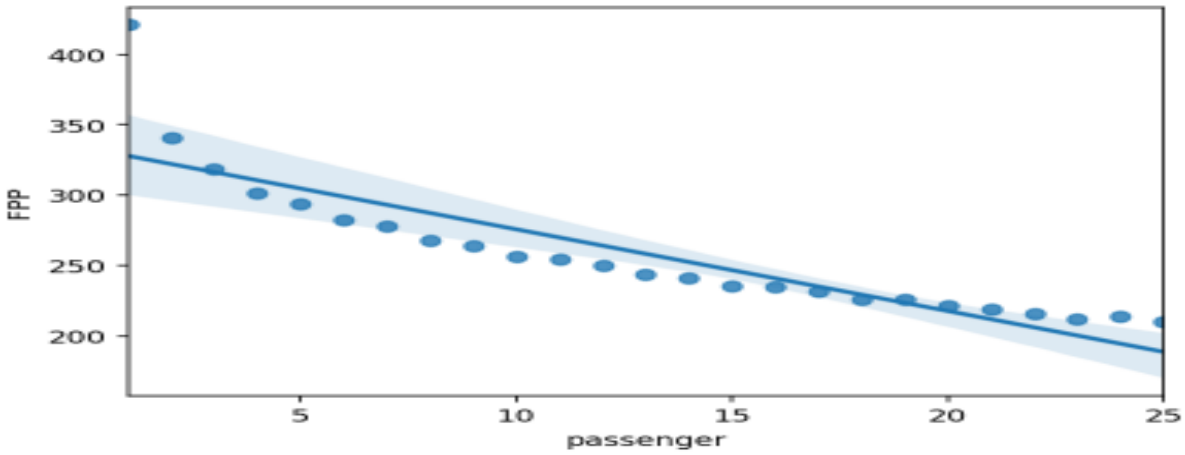


Bar and Line chart:relationship between Average FPP and average number of flight in each quarter of 2019



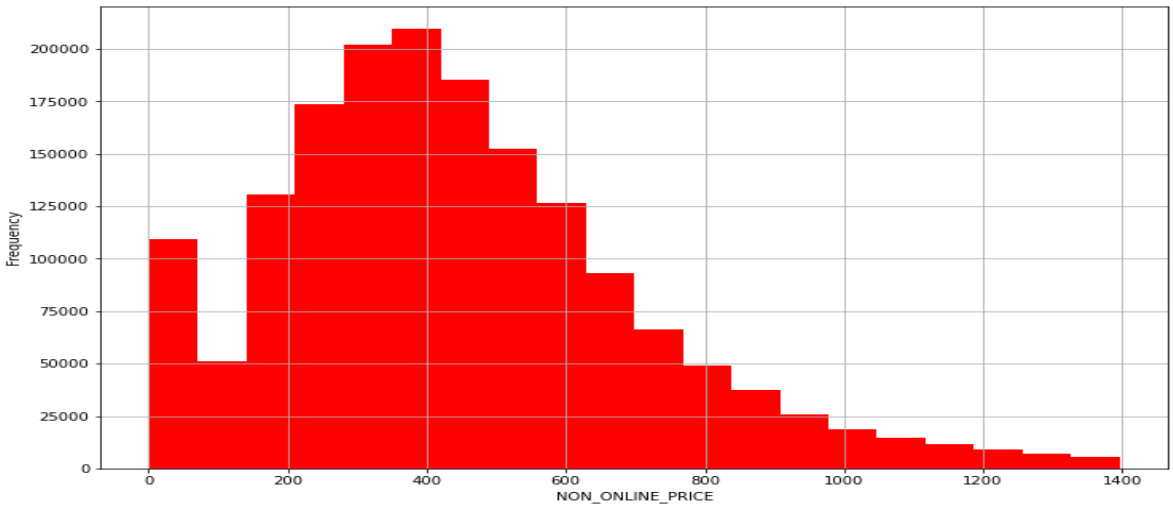
Bar and Line chart/Regplot plot :Relationship between the Average number of passage in one flight and average FPP in 2019



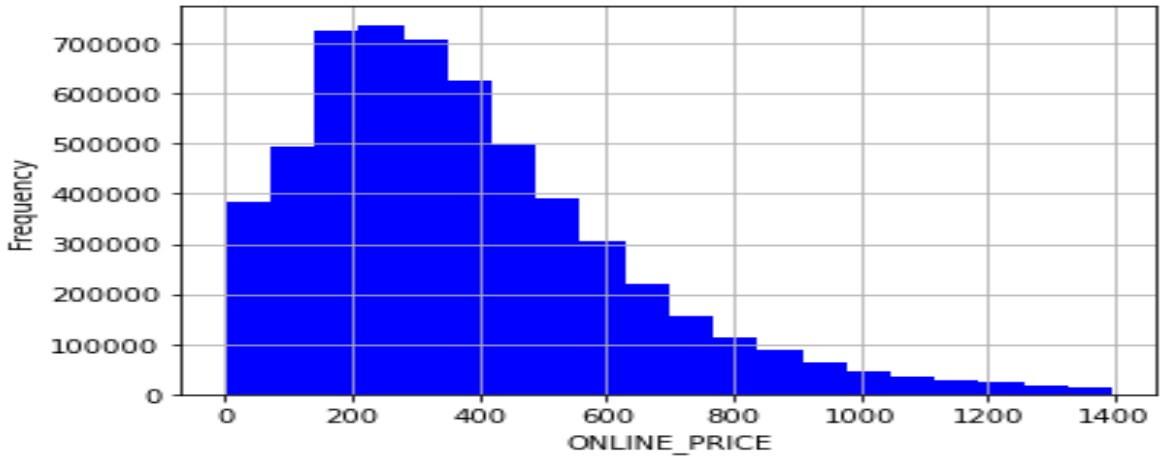


Histogram: Relationship between the whether is online order and Average price in FPP

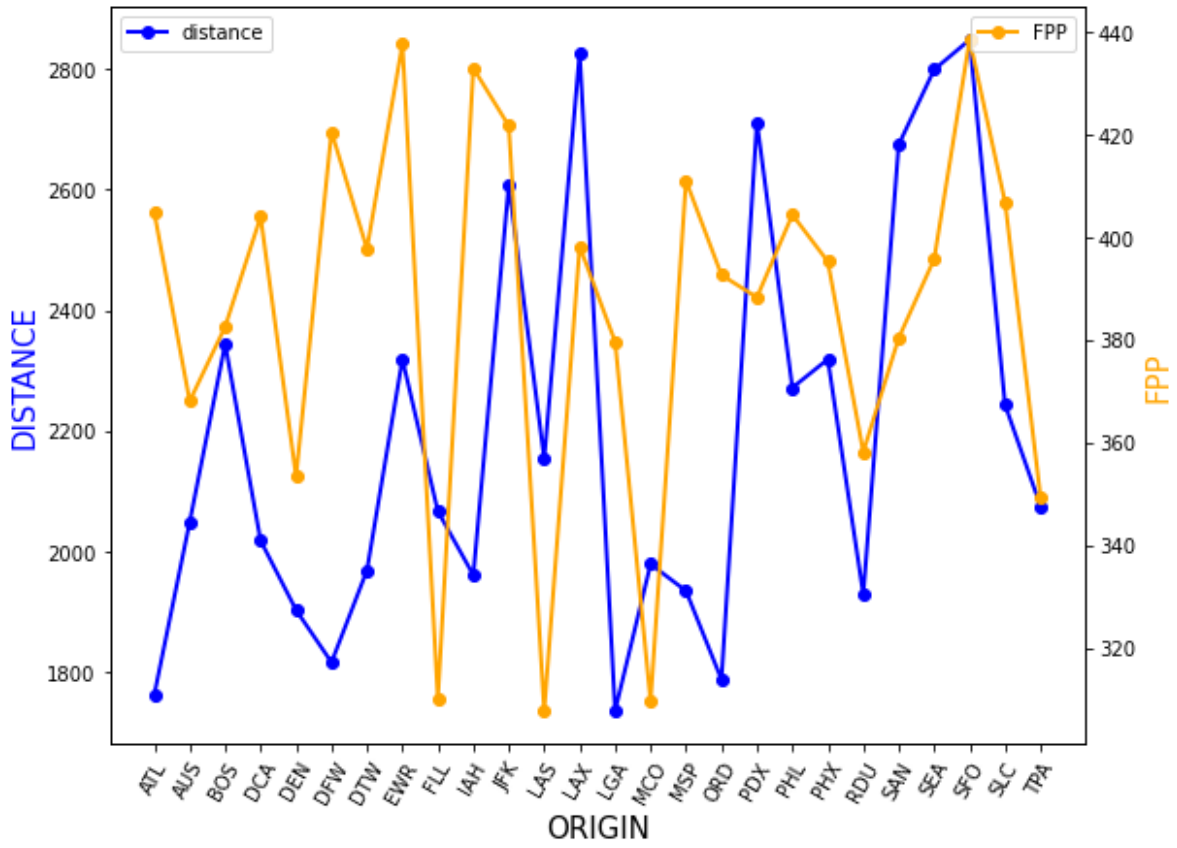
Non-online order:



Online order:



Line Graph: Relationship between the average distance and average FPP in difference origin in 2019



Works Cited

"Annual US Domestic Average Itinerary Fare in Current and Constant Dollars." *Bureau of Transportation Statistics*, https://www.bts.gov/content/annual-us-domestic-average-itinerary-fare-current-and-constant-dollars. Accessed 26 November 2021.

Kunst, Alexander. "• Leisure air travel frequency by age in the US 2019." *Statista*, https://www.statista.com/statistics/316365/air-travel-frequency-us-by-age/. Accessed 26 November 2021.

Lucas, Patrick, et al. "Demography, geography, and airport traffic - ACI Insights." *ACI Insights*, 3 October 2019, https://blog.aci.aero/demography-geography-and-airport-traffic/. Accessed 26 November 2021.

Mazareanu, E. "• Passenger air traffic each year." *Statista*, 5 October 2021, https://www.statista.com/statistics/564717/airline-industry-passenger-traffic-globally/. Accessed 26 November 2021.

McCartney, Scott. "Air Travel Prices Have Barely Budged in 25 Years. (It's True.)." *Wall Street Journal*, 10 August 2021, https://www.wsj.com/articles/why-airfare-has-barely-budged-in-25-years-11628600401. Accessed 26 November 2021.