DAO2702 Business Analytics

Team Python Pros

Project Report

Table of Contents

1. Business Problem and Objectives

1.1 Background

Trump Brothers is a travel agency that mainly deals with the domestic tourism scene within the US. As a travel agency, their main goal is to maximize profits by cutting down on the costs incurred from making travel arrangements for clients, and increasing revenue through increasing sales for attractive travel deasl. One major component affecting overall profits is Flight Fares. Trump Brothers recognizes that flight fares are affected by a multitude of factors, from macro considerations such as geopolitical relations and pandemics, to micro considerations such as purchase timing, route competition, seat availability, and even the specific day and time of flight (Younus, 2022). This makes it extremely difficult for them to predict flight fares and pinpoint the optimal deal to maximize profits. This problem is exacerbated by the Covid-19 pandemic, which made flight fares even more dynamic. In the United States (US), the pandemic brought about a decline in average airfares of $24 in the first quarter of 2020 (Zotova, 2021). Fast-forward to the post-pandemic era, a confluence of re-opening borders, inflation and other factors have contributed to US airfares shooting back up by 25%, the largest jump since 1989 (Holzhauer, 2024). After taking a hit during the pandemic, Trump Brother notices an opportunity with the trend of increased tourism post-Covid – a historic high of 4.7 billion people is forecasted to travel in 2024 (Loi, 2023). To effectively ride the wave and maximize their profits, Trump Brothers has sought our consulting company, PythonPros, to better understand factors that influence increasing flight fares in order to minimise operational expenditure and plan more cost effective tour packages for their customers.

1.2 Project Objectives

To achieve these goals, Python Pros will first provide Trump Brothers with a thorough analysis of what are the factors that affect flight fares in the US. We will then make use of data visualisations to help them better understand the relationships between each factor and flight fares. Finally, we will consolidate our findings by creating a robust prediction model based on linear regression to enable accurate predictions on flight ticket prices. Ultimately, we hope to advise Trump Brothers on how to make more informed decisions when purchasing flight tickets and how they can effectively adjust their pricing strategies when selling price packages so as to maximise profits.

2. Data Set and Handling

2.1 Data Source

To solve this problem, we will be using the "Flight Prices" dataset (Flight Prices (kaggle.com)) obtained from Kaggle. This dataset contains information about one-way, domestic flights scraped from Expedia between 2022-04-16 and 2022-10-05 in the US. This data source was selected for its comprehensiveness, offering over 27 columns, from departure time to airlines and even number of seats remaining for each listing, and over 100,000 records of data. Thus, it provides us with many potential variables to investigate and their effects on US domestic flight ticket prices, from which we will draw insights through our data visualisation.

2.2 Data Cleaning

Before drawing insights from the chosen data set, data cleaning was performed to remove irrelevant information. First, unnecessary columns, like the unique listing identifier "legID", which do not provide any information were dropped, while incomplete records which have null values were also removed. The scope of the dataset was then narrowed to fit the considerations of travel agencies which specialise in price and speed better. We focused only on direct flights which get customers to their destination in the shortest amount of time and considered the starting airport to only be from Los Angeles airport (LAX) to balance availability of data with model complexity. We have also selected to use total flight fare instead of base fare.

2.3 Assumptions

Some assumptions were made in interpreting the data. First, the dataset is assumed to be accurate and up to date. Secondly, we assume general uniformity in booking algorithms across all airlines, where industry trends are expected to be generalisable across all airlines for domestic flights within the US. Thirdly, we assume that the dataset used is representative of a diverse and comprehensive sample of fare prices across all dates, destinations and airlines, and is not skewed towards a particular factor.

3. Data Visualisations

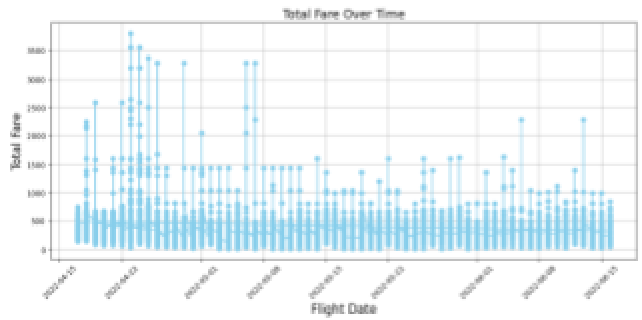
To understand the relationships between total flight fare and various numerical and categorical variables, in this section, we will conduct data visualisation, allowing us to identify meaningful trends and insights to help travel agencies obtain the cheapest flight fares based on their requirements. By performing the relevant calculations, we were also able to select the most suitable variables for creating the linear regression model (see Section 4) that will ultimately aid in predicting the total flight fare.

3.1 Time Series Variable

3.1.1 Variation of Total Flight Fare with Date of Flight

A line plot of flight fares over a 60-day period reveals fluctuating prices, which drop and increase within a few days' intervals. This proves the dynamism of flight fares and further highlights the importance of understanding the underlying factors that cause such variabilities. However, this variable will be left out of the prediction model as time series data is insignificant in linear regression models (Section 4.1).

Figure 1: Total Flight Fare Over Time



3.2 Numerical Variables

3.2.1 Correlation Matrix Analysis

From the correlation matrix heatmap, total travel distance (totalTravelDistance) and total travel duration (duration_h) displayed the highest correlation coefficients with total fare, at 0.48 and 0.47 respectively. This suggests a strong relationship between total fare and these variables.

However, while other variables like days left to booking or departure time, exhibit lower correlation coefficients with flight fare, it does not necessarily indicate a lack of strong relationship. In cases where the relationship between a variable and flight fare is strong but non-linear, the correlation coefficient may still be low (Investopedia, 2023). Therefore, we will visualise the data for these variables to investigate the relationship between them.

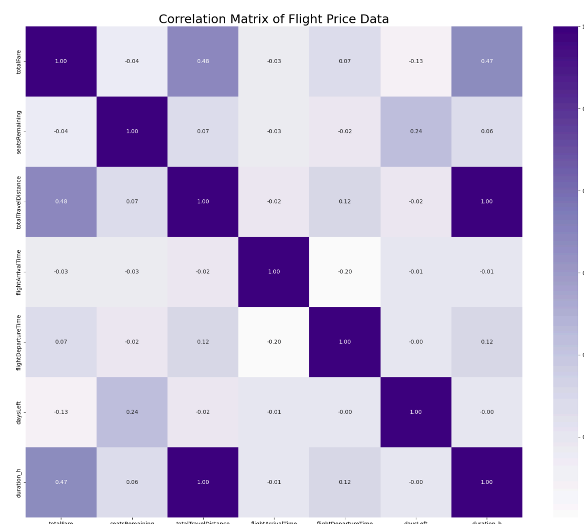


Figure 2: Correlation Matrix of Flight Price data

3.2.2 Variation of Total Fare with Total Flight Duration

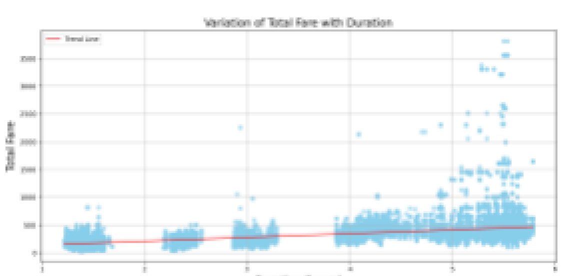


Figure 3: Scatter Plot of Total Flight Fare against Travel Duration

The linear relationship is shown by the red trend line and indicates a proportional increase in total flight fare with longer flight durations. This variable can be utilised to fit a linear regression model.

The total fare exhibits greater variability as the duration of the flight increases, indicating a wider range of fares for longer flights. The total fare remains stable within the range of \$0 to \$500 between 1 to 4 hours, so Trump Brothers can focus on flight options within this duration range to get better estimates of the flight fare and make more informed decisions.

3.2.3 Variation of Total Fare with Total Travel Distance

The linear relationship between flight fare and travel distance is shown by the red trend line, hence the variable is suitable for use in a linear regression model.

At shorter distances between 0 and 2000km, the total fare remains relatively stable, falling within the range of $0 to $750. However at travel distances upwards of 2000km, there is greater variation in total fare evident from the wider spread of data points above and below the trend line (Figure 4), indicating higher volatility in pricing for longer-distance flights. This insight is valuable as it highlights the potential for offering cost-effective flight options to farther destinations within this stable travel distance range.

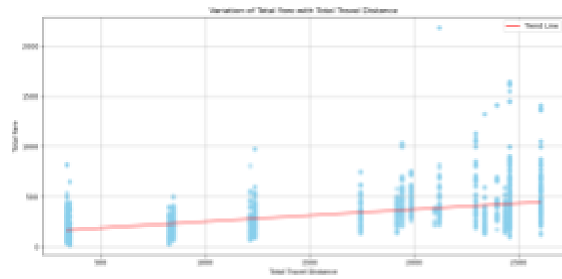


Figure 4: Scatter Plot of Total Flight Fare against Travel Distance

3.2.4 Variation of Mean Total Fare with Days Left

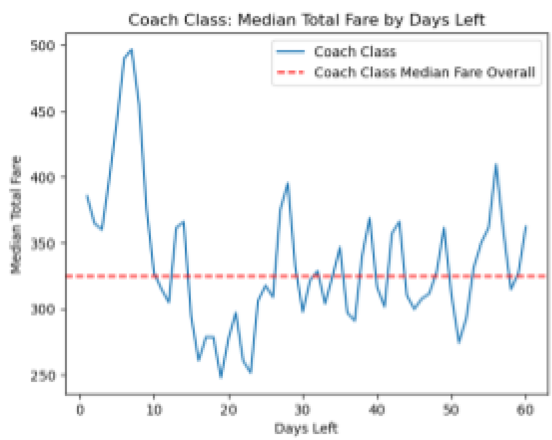


Figure 5: Median Total Fare by Days Left for Coach Class

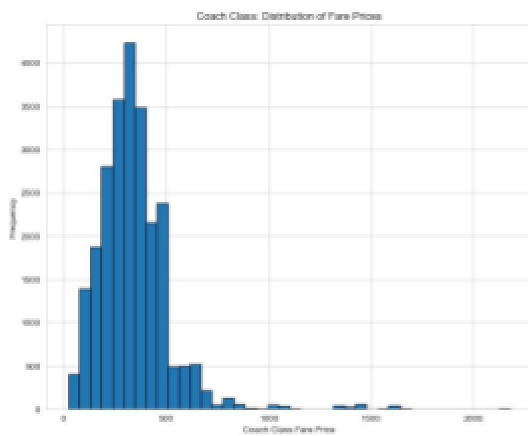


Figure 6: Distribution fo Fare Prices for Coach Class

For Coach Class, prices between Day 10 to Day 28 are consistently low, lower than the Median Total Fare Overall with exception to a small spike that happens around Day 12. The price then peaks in the last 10 days before the flight. Median price is used as the distribution of coach prices follows a right skewed distribution spread, since the median is less sensitive to outliers. It can be derived that airlines raise their coach prices significantly in the last 10 days before the actual flight and drop their prices for around 20 days before that time period.

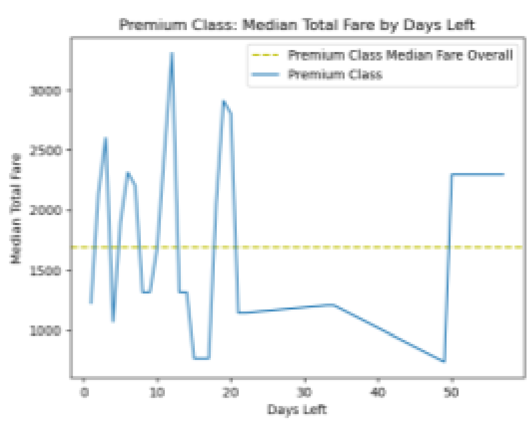


Figure 7: Median Total Fare by Days Left for Premium Class

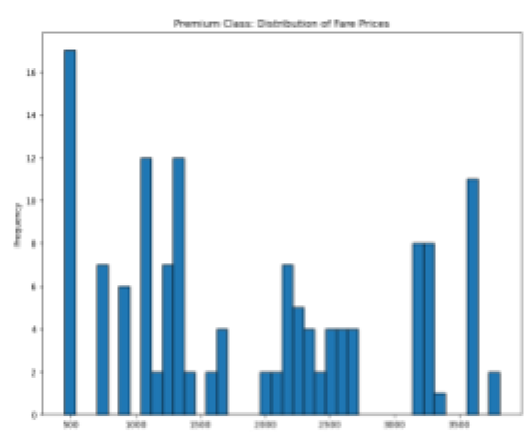


Figure 8: Distribution of Fare Prices for Premium Class

For Premium Class, between Day 20 to Day 50, prices are consistently low, significantly below the Median Total Fare Overall. Median Total Fare Overall is used instead of mean as fares in Premium Class do not show a clear distribution of prices. Hence, the median represents a more stable estimate of the central tendency of Premium Class prices. There is a high guarantee that from Day 20 to 50, the travel agencies can confidently book their Premium Class tickets for a lower price.

3.3 Categorical Variables

3.3.1 Analysis of Flight Departure Days on Total Flight Fare

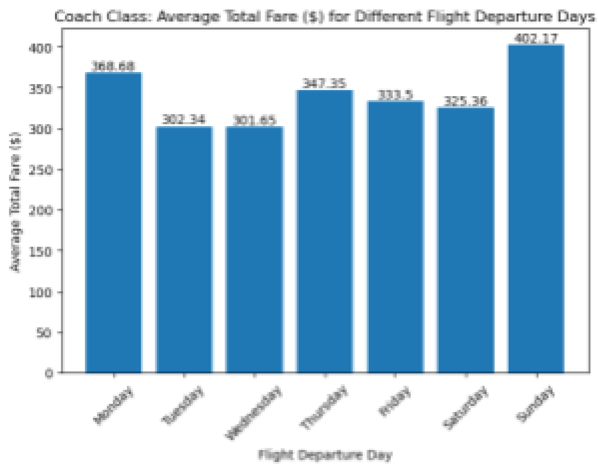


Figure 9: Average Total Fare across Different Flight Departure Days for Coach Class

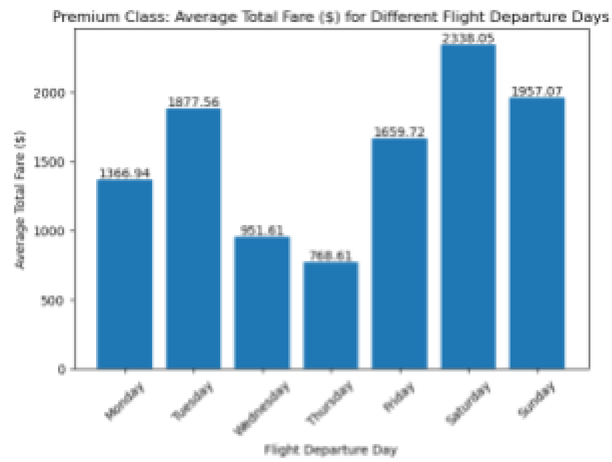


Figure 10: Average Total Fare across Different Flight Departure Days for Premium Class

Across the days of the week for Coach Class, Wednesday produced the lowest average fares of $301.65, closely followed by Tuesday at $302.34, while Sunday had a clear peak with the highest average fare of $402.17 (Figure 9). Meanwhile, price appears to fluctuate more greatly across the week for Premium Class tickets, with the cheapest fares still being in the midweek on Thursdays ($768.61) and Wednesdays ($951.61), peaking on Saturday ($2338.05) at thrice Thursday's fares. (Figure 10). This suggests that flying specifically on Tuesdays and Wednesdays for Coach Class, or Wednesdays and Thursdays for Premium Class allows one to get the lowest average total fares, with savings of up to 68% compared to peak fares. This is consistent with secondary research which recommends flying during the midweek for the cheapest tier (Coyle, 2024). In fact, simply choosing to fly on any day other than the day in which fares peak can achieve cost savings of at least 9.3% in the most modest comparison between Monday and Sunday fares for Coach Class .

3.3.2 Analysis of Airline and Destination Airport on Total Flight Fare

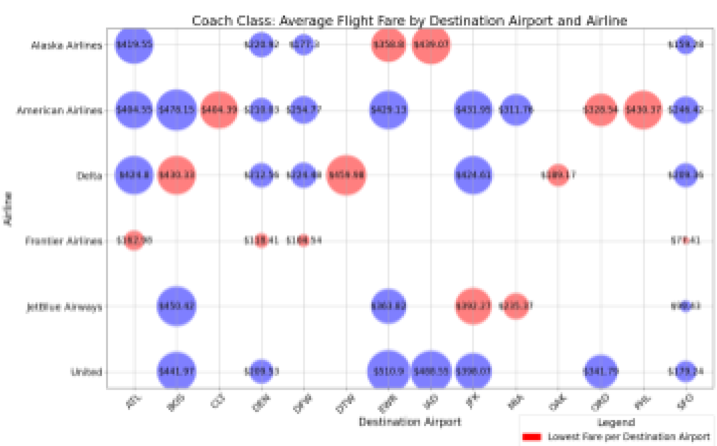


Figure 11: Bubble Plot of Average Flight Fare by Destination and Airline for Coach Class

From the bubble plots, we can examine each column to identify the cheapest airline for each airport (highlighted red) and examine each row for destinations each airline supports.

For coach flights, Frontier Airlines provides the cheapest flights. Prices are more than 3x cheaper on average compared to other airlines to San Francisco (SFO). This is in line with it being a budget airline (BudgetAir, 2024). However, it only serves flights to 4 out of 14 of the total destination airports listed. American Airlines is the most

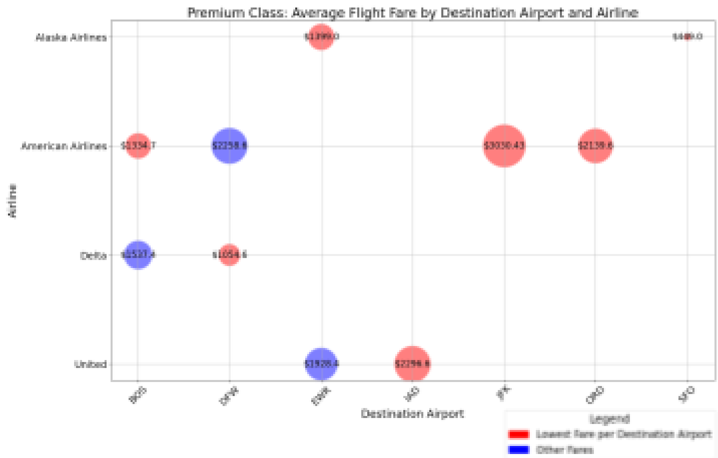


Figure 12: Bubble Plot of Average Flight Fare by Destination and Airline for Premium Class

versatile in terms of destinations, as the only choice that flies to Charlotte Douglas (CLT) and Philadelphia (PHL) and serves a total of 11 different destinations.

There are far fewer destination airports and airlines which offer Premium Class tickets (Figure 11). Even amongst the 7 destinations with premium flights, a maximum of two airlines per destination provide such service, with most airports only having one airline that provides premium flights. American Airlines is the most versatile airline as it offers Premium Class tickets to 4 destinations. Comparing coach and premium flight tickets for the same airline and destination, premium tickets are between 300% to 702% more expensive.

3.3.3 Analysis of Departure Time Period on Total Flight Fare

Total travel duration was converted from a string format to total flight hours; and departure time was classified into 6 segments, “Early Morning” (4am-8am), “Morning” (8am-12pm), “Afternoon”(12m-4pm), “Evening”(4pm-8pm), “Night”(8pm-12pm) and ”Late Night” (12am-4am) . The outliers and data points are not symmetrically distributed so using median values will be a more robust measure of central tendency.

Coach Seat Class

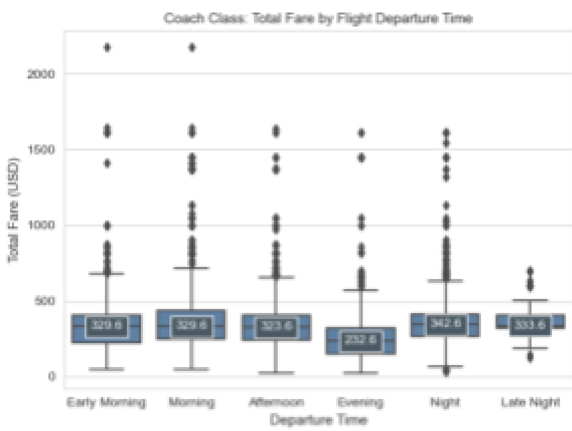


Figure 13: Box Plot of Total Fare by Departure Time for Coach Class

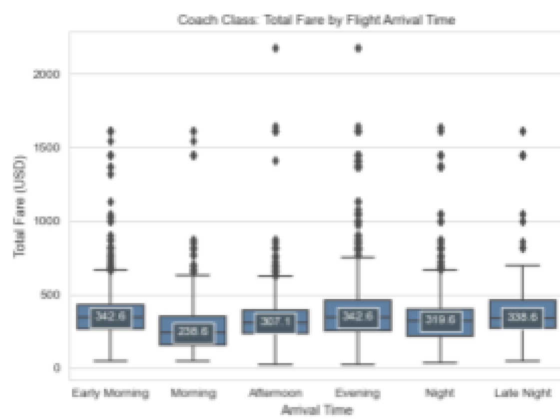


Figure 14: Box Plot of Total fare by Flight Arrival time for Coach Class

Total Flight Fare and Flight Departure and Arrival Time Periods are analysed for Coach Class. Median Departure Flight Prices are significantly lowest in the Evening (4pm - 8pm) at USD$232.6 with prices peaking at Night Time at a median price of USD$342.6 (Figure 12). Flight prices are more or less consistent for Early Morning, Morning, Afternoon and Late Night time flights with median total fare being at USD$329.6 for both Early Morning and Morning Time Periods (Figure 13). For arrival flight prices, median prices are the significantly lowest in the morning (8am-12pm) at USD$238.6 and highest in the Early Morning and Evening at USD$342.6.

Premium Seat Class

For Departure, flights that depart at night time have the lowest median fare at USD$1240.6 together with the afternoon time period. Prices peak at evening time with the highest median total fare at USD$2483.6 (Figure 15). For Arrival Flights, flights that arrive at Morning time of (8am-12pm) have the lowest median fare at USD$1240.6.

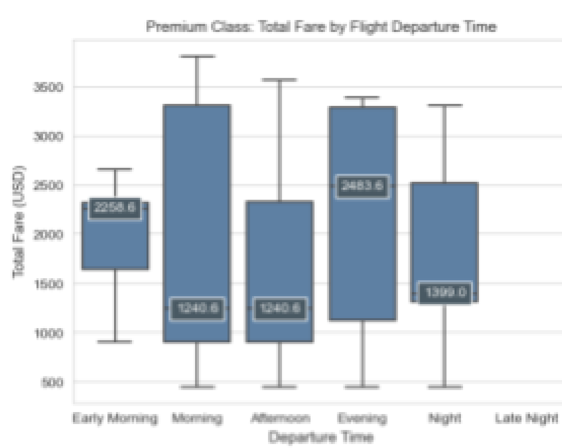


Figure 15: Box Plot of Total Fare by Flight Departure time for Premium Class

Late Night has a significantly highest median price at USD$3241.6 (Figure 16).

There are insufficient data points for Late Night Departure Flights and we will not consider Afternoon Arrival flights due to the low number of data points hence flight prices might be unrepresentative.

Overall, for Coach Class, evening departure and morning arrival will be the most ideal while Night Time departures and morning arrival will be the most ideal. For Premium Class, Morning or Afternoon Time departures and Morning Arrivals will be the most ideal while Evening Departures and Late Night Arrivals are the least ideal.

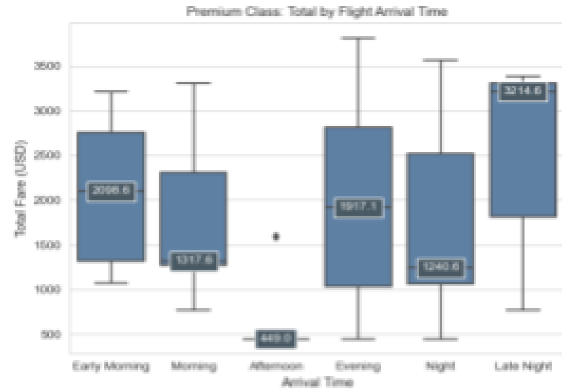


Figure 16: Box Plot of Total Fare by Arrival Time by Premium Class

4. Linear Regression Model

Our data analysis has given insight into how our travel agency can manage their operational costs by reducing flight fares. We can then use the identified variables as parameter inputs to build a predictive linear regression model to predict and optimise flight prices. This aims to help our travel agency purchase flight tickets when they are cheaper, plan more cost-effective itineraries, and better manage their pricing strategies. This was done using the Sci-kit Learn Machine Library.

4.1 Building the Model

Building the model using the Pipeline and LinearRegression function requires the selection of relevant variables to be input parameters. Firstly, we will exclude the time series data from our training data set as linear regression models do not consider time dependency and assume that observations are independent (D.Aman, 2023). Next, numerical variables which have shown to have some correlations with total flight fare have been kept (Figure 2). However, we noticed that total duration and total distance travelled have a similar correlation with flight fares. These variables are highly and positively correlated with each other (C. Lago et. al, 2021) because as distance travelled increases, duration of travel increases. Hence, it suffices to keep only one of the variables - totalTravelDistance - to observe its effects on the predictions. Lastly, categorical variables identified in our analysis were converted into dummy variables in the training data, using the OneHotEncoder function.

4.2 Feature Importance

In further analysing our linear regression model, we identified that the top few factors contributing to higher flight fares are total travel distance, flight arrival time, flight departure time and days left until departure (Figure 17). Feature importance is derived based on the coefficients of the linear model and indicates how much each variable feature contributes to increasing flight fares (Z. Enes, 2024), allowing us to unravel insights that observing correlation values in Section 3.2.1 did not reveal due to non-linear relationships between the numerical variables and light fare. This would allow the travel agency to give more importance to these factors when planning itineraries.

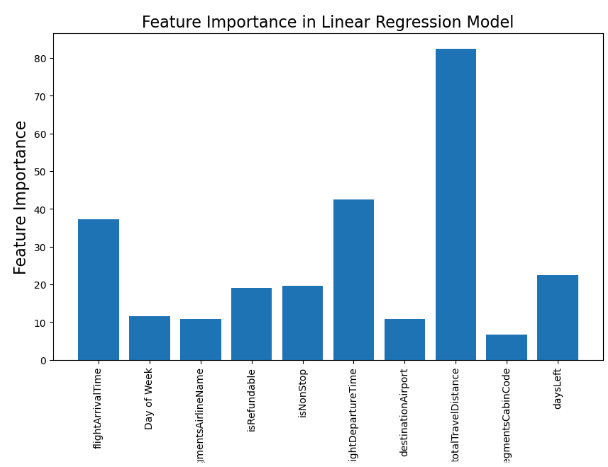


Figure 17: Feature Importance drive from model's coefficients

4.3 Model Performance based on Cross - Validation (CV) Scores

Using the cross_val_score and cross_val_predict functions, the Linear Regression model has an average accuracy of approximately 58.88% based on 10-fold CV. Based on the comparison between the predicted and actual flight prices (Figure 17), we can see that the model is effective in predicting flight fares when their price range is lower. However, when it increases beyond the $1000 range, the predictions start to fluctuate and

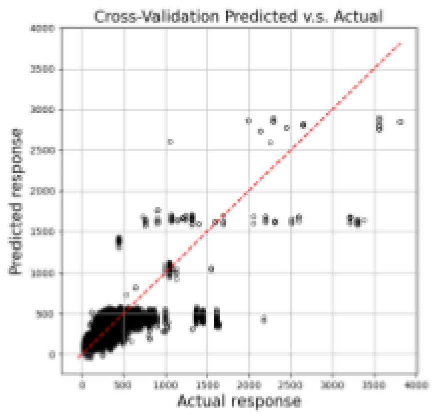


Figure 18: Cross-validation vs actual values of prediction model

become less accurate. This implies that the model can be applied to help our travel agency predict lower flight prices. However, for higher flight prices, we would consider fine-tuning the model, which will be discussed in Section 6.

4.4 Application of Model

Deviations of model accuracy between each CV is also relatively small, as seen from the violin plot's symmetric shape with the median sitting near the middle of the plot (Figure 18), which implies a relatively balanced distribution of CV scores. This suggests that the model is quite robust and performs well across different subsets of data, which is a good indication of the model's generalisability. This is important as the model was trained on a subset of the flight data, which includes direct flights departing from LAX only, and thus it can be generalised across the other subsets of the data.

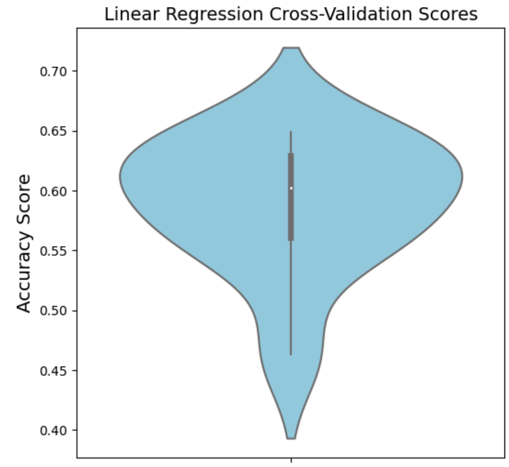


Figure 19: Violin plot of cross-validation scores

4.5 Limitations of Model

As seen from the evaluation of our model, it is relatively effective in predicting lower range flight fares. However, for higher range flight fares for more expensive travel itineraries the predictions tend to fluctuate and deviate away from the actual prices. This may be because the continuous variables that affect flight prices do not follow a normal distribution, as seen in Section 3.2.3 as well, which is one of the most important factors when building a linear regression model. It shows that data near the mean are more frequent in occurrence than data far from the mean (R. Sandeep, 2020), which is not applicable to flight fares which are affected by a multitude of factors and vary a lot.

Furthermore, from observing our dataset, it is extremely right-skewed because there is a significantly large amount of data on Coach Class flights, as compared to the Premium Class flights which are much cheaper than the Premium Class. This limits the prediction accuracy of our model as there aren't sufficient data points for the Premium Class flights to train the model on.

5. Recommendations

Having a comprehensive understanding of how the various factors affect flight prices, we have now come up with a three prong solution for Trump Brothers to strategically adopt for their current as well as upcoming series of tour packages in order to minimise costs on their end regarding flight tickets.

5.1 Design Low Cost Itinerary Packages

5.1.1 Deciding the destination and airlines of travel

For distance of travel, 250 km - 1500 km is a better range to consider since flight fares remain relatively stable between $0-$600 regardless of the distance, going beyond this range would result in more fluctuating costs. Additionally, Trump Brothers should maximise on distance to increase value for money and travel to farther destinations, applying this in their pricing strategy.

In choosing airlines, Frontier Airlines should be chosen if the package includes flight to ATL, DEN, DFW, SFO and the priority is cheap travel, whereas American Airlines should be picked if flights are to uncommon locations like CLT, PHL.

5.1.2 Deciding the date and time of the itinerary

We recommend flying on Tuesdays or Wednesdays and avoiding Sundays for Coach Class, and flying on Wednesday and Thursdays and avoiding Saturdays for Premium.

Overall, Flight Fare does not vary too much between different time periods for Coach Class but purchasing flights that depart at evening time and flights that arrive in the morning would be recommended For Premium Class, seems that Night Time Departures and Morning or Afternoon Arrivals would be more ideal

5.1.3 Deciding when to purchase the flight tickets

It would be ideal to book a flight between 10 to 50 days before takeoff for coach and 20 to 50 days before flight takeoff for premium as prices are more stable and within a cheaper range. Booking less than 10 days before the flight's departure might result in steep hikes in price.

5.2 Reduce Flight Fare Costs for Existing Travel Packages

Recognising that it is not always possible to design travel packages solely around minimising air fare cost, Trump Brothers can also use the insights from our data visualisations to reduce total flight fares for existing packages. Based on the fixed objectives in their existing packages, they can then optimise the relevant variables to get the cheapest fares. Two realistic scenarios are given below.

Case 1: If they only want to plan **budget itineraries** for customers who are looking for cheaper travel options, they can leverage the optimal flight departure/ arrival times, travel days, and choose the cheapest airlines (Frontier) and input these as parameters to predict the flight fare. They can then package their travel packages to go to the destinations that Frontier airlines offers

Case 2: If they want to plan a more **luxurious travel package**, they can fix the airlines to (insert more premium airlines) or focus on the more expensive destination, then experiment with the other variables to optimise the flight fare costs.

5.3 Optimise Pricing Strategy and Budgets

Finally, Trump Brothers can also utilise our flight fare prediction model to predict the precise cost of their flight once itineraries have been confirmed. Using this information, they can then adjust their pricing strategy of their travel packages and more effectively plan their budget. For instance, if predicted flight fares are lower than initially budgeted for, Trump Brothers can offer a discount for customers or increase the value of . Conversely, if predicted fares are higher than expected, they can adjust other components of the travel package like accommodation or food to manage costs.

6. Conclusion

Looking ahead, we can improve our model's prediction accuracy across all ranges of flight by applying a logarithmic transformation on the variables, to obtain a linear function that is as close as possible to a straight line and remove the impact of the many outliers in our dataset (M. Yonchev, 2023). This would fix the skewness of the data and better predict flight fares by fine tuning our model. Furthermore, a random forest model can also be considered because it may be more suitable for this business problem's context, as the variables need not necessarily follow a normal distribution or have a linear relationship with the dependent variable. In addition, improvements to focus on regarding the data set itself include using data from a larger time frame instead of only 60 days, in order to capture seasonal trends and fluctuations that are also important factors to flight prices. Also, using a more generalisable dataset including flights outside of USA domestic would allow the model to be applied to broader business problems.

In conclusion, our data-driven analysis highlights key factors influencing US domestic flight fares: flight arrival and departure times, days remaining until ticket purchase, and total travel distance and duration. Leveraging these insights, coupled with our predictive model, we can empower our client, Trump Brothers, to make more informed decisions when crafting travel packages and setting prices. This strategic approach positions them to capitalise on the post-pandemic tourism surge and maximise profits.

7. Appendix

Features	Count/ Range
Search Date	2022-04-16 to 2022-04-18 (3 days)
Flight Date	2022-04-17 to 2022-06-05 (50 days)
Starting & Destination Airport	ATL, BOS, CLT, DEN, DFW, DTW, EWR, IAD, JFK, LAX, LGA, MIA, OAK,ORD, PHL, SFO
Airline Name	Alaska Airline, American Airline, Boutique Air, Cape Air, Contour Airlines, Delta, Frontier Airlines, JetBlue Airways, Key Lime Air, Southern Airways Express, Sprint Airlines, Sun Country Airlines, United
Travel Duration (h)	0 to 9.98
Total Travel Distance (miles)	282 to 3686
Ticket Class	First, Coach, Business
Fare of the Ticket (USD)	totalFare

Appendix A: Breakdown of data source used

8. References

AI, S. (2022, October 2). Flight fare prediction — time series ML project - Skillcate AI. Medium. https://medium.com/@skillcate/flight-fare-prediction-machine-learning-project-bc7363e6d9eb

Aman, D. (2023, November 24). Time series forecasting vs. regression - Danayt Aman. Medium. https://medium.com/@danaytaman/time-series-forecasting-vs-regression-cf89d0d0f3bd

Bhandari, A. (2020, March 19). Multicollinearity : Definition, causes and detection using VIF. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Coyle, M. (2019, August 29). The best days to book a flight and when to fly. NerdWallet. https://www.nerdwallet.com/article/travel/best-days-book-flight-fly

F. Siegel, A. (2016). Multicollinearity Problem - An overview. ScienceDirect Topics. https://www.sciencedirect.com/topics/mathematics/multicollinearity-problem

Fontanet-Pérez, P., Vázquez, X. H., & Carou, D. (2022). The impact of the COVID-19 crisis on the US airline market: Are current business models equipped for upcoming changes in the air transport sector? Case Studies on Transport Policy, 10(1), 647–656. https://doi.org/10.1016/j.cstp.2022.01.025

Frontier Airlines flights. (n.d.). Budgetair.CA. Retrieved April 19, 2024, from https://www.budgetair.com/en_ca/airlines/frontier-airlines

Holzhauer, B. (2022, May 14). Airline ticket prices are up 25%, outpacing inflation — here are the ways you can still save. CNBC.https://www.cnbc.com/select/airline-ticket-prices-are-up-25-percent-why-and-how-to-save/

Lago, C., Garzo, E., Moreno, A., Barrios, L., Martí-Campoy, A., Rodríguez-Ballester, F., & Fereres, A. (2021). Flight performance and the factors affecting the flight behaviour of Philaenus spumarius the main vector of Xylella fastidiosa in Europe. Scientific Reports, 11(1), 17608. https://doi.org/10.1038/s41598-021-96904-5

Loi, E. (2023, December 6). 4.7b people expected to travel by air in 2024, about 4% more than pre-pandemic levels: Iata. The Straits Times. https://www.straitstimes.com/world/47b-people-expected-to-travel-by-air-in-2024-about-4-more-than-pre-pandemic-levels-iata

Nickolas, S. (2015, March 25). Correlation coefficients: Positive, negative, and zero. Investopedia. https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Ram, S. (2020, October 16). Improve linear regression using statistics. Towards Data Science. https://towardsdatascience.com/statistics-supporting-linear-models-bfc24fb9781f

Wong, D. (2022). Flight prices. Kaggle. https://www.kaggle.com/datasets/dilwong/flightprices

Yonchev, M. (2023, March 26). Log Transformations and their Implications for Linear Regression. Medium. https://medium.com/@myonchev99/log-transformations-and-its-implications-for-linear-regression-f4570de0d69d#2640

Younus, J. (2022, October 27). Why do Flight Prices Keep Changing? How Airlines Price Their Tickets. https://www.alternativeairlines.com/blog/how-airlines-price-flights

Zotova, I. (2021). Impact of COVID-19 on airline pricing. Micronomics Economic Research and Consulting.

Zvornicanin, E. (2022, August 17). What is feature importance in machine learning? Baeldung on Computer Science. https://www.baeldung.com/cs/ml-feature-importance