Debbie McMurrick

DSC530 Data Exploration and Analysis

Final Project

November 16, 2024

My statistical/hypothetical question is “What variables impact finding the lowest fare for an airline ticket?” I found a dataset for US Airline flights and fare from 1993 – 2024. This dataset covers airline flight routes, fares, and passenger volumes within the United States. There are several questions this dataset could answer but I chose this one because I like to travel so I wanted to see what my chances are of finding a good fare, and what impacts those prices.

During the EDA process, I found a lot of outliers in the data for some of the fields I chose to analyze. However, I decided not to remove a lot of them because I felt that it would erase pertinent data relating to fare\_low. Instead, I did the first part of EDA without removing the outlies and then from the CDF portion to the end, I did do some manipulation of the data (like passenger counts greater than zero and removing NaN’s). This helped to provide a more accurate analysis at different phases of the EDA.

These were some of my findings during the EDA process:

1. A Kernel Density Estimate was used to see the analytical distribution, which showed most low fares were skewed to the left with the media low fare being ~$180.
2. Distance between Airports (nsmiles) had a moderate, positive linear correlation, with a Pearson’s Correlation Coefficient of 0.417 (significant).
3. Number of passengers (passengers) indicated a weak, negative linear relationship, with a Pearson’s Correlation Coefficient of -0.211 (not as significant).
4. Hypothesis Analysis showed we can confidently reject the null hypothesis on both scenarios.
5. Regression analysis used multiple explanatory variables, with a. R-squared value of 0.75, which suggests a strong relationship between the predictors and the response.
6. All P values were 0.00, which indicates that each independent variable is statistically significant at the common significance level of 0.05.

I don’t believe any of the other variables would have helped in the analysis. However, there may be an opportunity to create new variables with the data that may provide additional insight. I do think this dataset could answer several hypothetical questions with additional time for analysis. An example of this would be a way to analyze the airline carriers within the data by grouping them by airline and finding if there are certain airlines that had more low fares than others.

One assumption I made was to leave the outliers in the data because I wanted to first see what the results looked like with the outliers. It appears that some of the data may have been inaccurately entered and I felt it needed a deeper analysis to feel confident to remove those lines. Some of the fares were very high, even in comparison to the same route. There were also some passenger counts that were less than zero, which really shouldn’t happen. I did end up removing some NaNs and looking at passengers greater than zero for more accuracy in the PMF, CDF, KDE, Hypothesis Testing, and Regression Analysis.

The main challenge for me was picking a dataset and a statistical question. It took me a long time to zero in on one. Additionally, the PMF and CDF were a bit challenging in just understanding what it was showing me. Aside from that, I enjoyed the project, and it taught me a lot on EDA, different techniques, and regression analysis.

Dataset: <https://www.kaggle.com/datasets/bhavikjikadara/us-airline-flight-routes-and-fares-1993-2024>