Analyzing the Pricing of Airlines by Their Market Shares of Cities Around the United States

Jesse R. Carter

Western Governors University

Table of Contents

[**Overview** 3](#_Toc176856903)

[**A1. Research Question** 3](#_Toc176856904)

[**A2. Project Scope** 3](#_Toc176856905)

[**A3. Solution Overview** 3](#_Toc176856906)

[**Project Execution** 4](#_Toc176856907)

[**B1. project plan** 4](#_Toc176856908)

[**B2. Project planning methodology** 4](#_Toc176856909)

[**B3. project timeline and milestones** 6](#_Toc176856910)

[**C1. How data selection differed from the plan** 6](#_Toc176856912)

[**C2. How obstacles were handled** 7](#_Toc176856913)

[**C3. How unplanned data governance issues were handled** 7](#_Toc176856914)

[**Data extraction and preparation process** 7](#_Toc176856915)

[**Data analysis process** 8](#_Toc176856916)

[**E1. Methods used** 8](#_Toc176856917)

[**E2. Advantages and limitations of tools and techniques used** 8](#_Toc176856918)

[**E3. Explanation of analytical methods** 9](#_Toc176856919)

[**Results** 9](#_Toc176856920)

[**F1. Output of data solution** 9](#_Toc176856921)

[**F2. The practical significance of data solution** 10](#_Toc176856922)

[**F3. The overall success of the project** 10](#_Toc176856923)

[**G1. Conclusions drawn** 11](#_Toc176856925)

[**G2. How chosen tools support findings and effective storytelling** 11](#_Toc176856926)

[**G3. Recommendations of actions based on findings** 11](#_Toc176856927)

[**Panopto recording and Tableau public link** 12](#_Toc176856928)

[**Sources** 13](#_Toc176856929)

## **Overview**

## **A1. Research Question**

To what extent does an airline's market share of a city affect the overall price of flights? This information would be helpful for people who are seeking the best deal on flights from their local airports. While it is not the only aspect to consider when choosing their flight, I believe, however, it is something that will be helpful for people in finding the correct airline for their ongoing flights.

## **A2. Project Scope**

The price of everything is rising, and as such, the cost of air travel is rising as well. People who choose to fly need to get the best deal for their travels. There are multiple reasons to select any given airline such as the cost of flight, amenities, and route options to name just a few. There have been recent spikes in flight tickets from departing cities where certain airlines appear to have a major market share. Finding out if this is happening, with either one airline specifically, or with multiple airlines and to what extent it is or is not happening would benefit people and provide more knowledge of why airlines are charging the amount they are charging.

## **A3. Solution Overview**

For this project, I downloaded a dataset from Kaggle.com and used a Jupyter Notebook as the testing environment, as well as Python as the programming language within the environment. This combination allowed me to perform various calculations and data manipulations on the dataset accurately. Upon loading the dataset and any needed libraries for the project, I first went about looking into the data and what columns I believed I needed. I determined that I only needed a few columns such as the average fare, the average fare from the largest carrier, the market share from the largest carrier, and the origin airport for my analysis. I cleaned the data and went to work both visualizing the data and performing data manipulation to test and determine my hypothesis.

## **Project Execution**

## **B1. project plan**

By the end of this project, I managed to execute the project according to the intended plan. I managed to meet every objective that was previously discussed in task two. The objectives are listed below for further reference.

***Objective 1.0***

To obtain a clean dataset with which to use for the purpose of assisting with the project.

***Objective 2.0***

Use the cleaned dataset to perform the needed tasks such as finding the r2 as well as the p-value to resolve the goals question.

***Objective 3.0***

Create a Tableau dashboard with all the needed graphs and charts that support the final analysis of the project and then simplify the terms in the final analysis of the project for the end user.

***Deliverable 3.1***

A written report that will cover all the findings of the project and discuss how I came to that conclusion laid out in simple terms for the end user.

## **B2. Project planning methodology**

I used the ADDIE method for this project. I believe that this approach was the most logical choice as it allows for clear-cut steps to follow, and it did provide the desired outcome.

ANALYZE

I will analyze the information provided to me by the dataset and identify which data points I will use for the project. Some of the things I need to decide on will be as follows:

* What platform to host the data?
* How will I read and interact with said data?
* Will there be any tests I need to perform on the data?
* How will I represent the data when finished?

DESIGN

* I will come to a clear conclusion regarding this question by implementing the following steps.
* Collect the data from the dataset
* Transform and process the data into a set that will address the issue at hand
* Analyze the data to get a clear picture of the question
* Interpret the data and find an answer to the question posed
* Display the data results using Tableau

IMPLEMENTATION

I implemented the code for this project from steps 1 through 4 using Jupyter Notebook and Python. Then I used the data results from Step 5 using Tableau.

EVALUATION

* The project was evaluated through the length of each step to ensure that I was as thorough as possible before moving on to the next step.
* I evaluated the result to ensure that there was an appropriate result whether it be proving either the null or alternative hypothesis.

## **B3. project timeline and milestones**

## While the timeline’s overall project followed a similar trajectory, the start dates and task durations were slightly different. This was due to a variety of issues; however, the milestones did not alter from the original proposal. As noted below with the updated timeframe, this project took a bit longer due to environmental and other issues. In the end, the project took a little longer to get approval and as such, started later than planned. The collection of data and the cleaning/transforming of data stages were on schedule, however, the last four stages each took roughly one extra day. As such, it extended the overall project timeline to twelve days from the originally planned eight days.

## 

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Start Date** | **End Date** | **Task Duration** |
| Collection of Data | 08/26/2024 | 08/26/2024 | 1 Day |
| Cleaning and Transforming of Data | 08/27/2024 | 08/28/2024 | 2 Days |
| Training of Data Models | 08/29/2024 | 08/31/2024 | 3 Days |
| Testing and fine-tuning models | 09/01/2024 | 09/02/2024 | 2 Days |
| Analysis of Data | 09/03/2024 | 09/04/2024 | 2 Days |
| Display Data in Tableau | 09/03/2024 | 09/05/2024 | 2 Days |

**Methodology**

## **C1. How data selection differed from the plan**

The overall process of data selection stayed the same from the original proposed plan. However, when it came to the collection of the needed data, the original csv file from Kaggle went missing before I had a chance to download it. As such, it took a while for me to find it again. Eventually, it was re-uploaded by another user, and as such, I was able to finally start the project.

Once I had the overall data, I eventually determined that I only needed the information for the largest carrier, the origin airport, the largest carrier’s market share, the average fare, and the average fare for the largest carrier. I ended up not needing the smallest carrier’s information to prove or disprove the hypothesis.

## **C2. How obstacles were handled**

While I initially did not have any issues with collecting the data, my house did suffer from an extended power outage and as such, some of the needed files were lost. When I parsed through the files to determine what was missing, I determined that most of them were backed up. The only thing that was not backed up somehow was the original CSV file. Upon attempting to re-download, it from Kaggle, I discovered that the original publisher had closed their account. The next day, I tried to find it online through the original publisher's other listed works. However, it was soon clear to me that they had not published that file anywhere else. Just as I was about to reach out to discuss this new wrinkle, I rechecked Kaggle and found that a new author had published the same CSV under a different name. I proceeded to download that file and back it up immediately. I will update my references page with the new dataset link from Kaggle for this part of the project.

## **C3. How unplanned data governance issues were handled**

During the course of the project, I encountered no issues with unplanned data governance.

## **Data extraction and preparation process**

Once I had the dataset downloaded and loaded into Jupyter, I began by looking over the dataset, I determined that, while there was a large amount of data here, I only really needed five columns to perform my analysis and answer the question posed. I started by determining how many rows were in the dataset so I could determine the overall size. It was clear early on that this dataset was quite large and, as such, it might have some data issues. Once I selected the columns that I needed, I decided to rename them for a clearer description of the data. It was determined that there were some missing data in a few of the columns I needed, however, due to the sheer size of the data, I decided to just drop the NaN values, as there would still be more than enough data points to come to a logical conclusion.

Once I had everything figured out, I decided to start by creating a new data frame with just the columns I would be using and then dropped all the Nan values as described above. This gave me a clear, concise place to start with the data.

## **Data analysis process**

## **E1. Methods used**

I used a variety of charts and graphs to analyze the data during this stage, such as a scatter matrix for a better visualization of the chosen variables on a high level. I used a correlation matrix in the form of a heatmap to help visualize the relationship between the values such as the largest fare and average fare. I rounded the data visualization out with some bar charts that helped bring the top ten values together in a clear manner.

When it came to statistical analysis of the data, I decided to find the r-squared value of the largest market as it related to the average fare. I then used a z-score test on the average fare, largest fare, and largest market share fields. Then I finished the testing by determining the p-value of everything. Since the alpha is set to a threshold of 0.05, with the r-squared values, z\_scores, and p-values, we can then determine whether the null or alternative hypothesis would be accepted.

I finally proceeded with creating a simple dashboard in Tableau showing the market share of carriers by airport, the sum of the largest fare and average fare based on market size as well as the percentile of average fare by the count of market size of the carrier.

## **E2. Advantages and limitations of tools and techniques used**

I used a few primary tools for this project such as Jupyter Notebook and Python. There are some major advantages as well as some minor disadvantages with these tools.

**Advantages**

The combination of these two together allows a lightweight, flexible way to document, visualize, and process data in a reproducible method.

**Disadvantages**

Using this specific combination of tools can become cumbersome and lead to a lack of efficiency due to the many libraries that they are dependent on.

## **E3. Explanation of analytical methods**

I used a range of analytical methods during this project including the following:

* Processing and cleaning data to ensure no unintentional errors when it came time to test and evaluate the models.
* I used a range of visualizations to better understand the correlation between the data and to assist in finding the correct data to input into the testing models.
* Regression analysis was used to determine the r-squared value to determine the correlation between the datasets.
* I then used a Z-test to determine the normal distribution of the data and compared the p-values to determine if the differences were significant to prove or disprove the null hypothesis.

## **Results**

## **F1. Output of data solution**

I used Logistic regression as the base of my statistical model. I first reduced the dataset to only the columns I would need to conduct the test. From there, I conducted the logical regression testing and evaluated it with a z-test and p-value testing.

The results are as below

* R-squared value of.91
* Z-score of 2.08 on the average fare, 1.30 on the largest fare, and 6.21 on the market share
* The P-value mean of everything was found to be .49

From these tests, I can determine that the model is highly accurate due to the r-squared value being so high. After applying the z-test to every needed column, I found an interesting correlation. It appears that all three columns (AverageFare, LargestFare, and MarketShareLG) appear to have a strong positive z score. This would indicate that the Alternative Hypothesis would be correct and there is a direct correlation between an airline's market share in each market and its overall price. With a p-value at.49, I would conclude that the Alternative hypothesis is proven.

## **F2. The practical significance of data solution**

With the ever-rising price of travel, it is nice to understand why companies charge what they do. In doing so, one thing to consider for flight tickets is whether a company is charging their customers more based on whether they can just because there is no other option for the consumer. This could be perceived as a poor business tactic by some people, and they should know if the airlines are engaging in this practice. It will better inform customers on where to spend their limited travel funds and avoid supporting businesses with poor consumer practices.

## **F3. The overall success of the project**

I would say that overall, the analysis was successful. Through the testing, I found a correlation between the average market share and the highest charged fare. Whether or not this can be directly related to the market share alone, however, was determined to be false. It appears based on my test that there is no direct correlation between an airline's market share in a given market and its overall price. There could be other deciding factors in why a company charges what they do, such as more flights more aircraft maintenance at the airport, more staff needed, or more fuel. The fact that the company has a larger market share could just be that they determined that the city is a good place to have more of a central hub. Whatever the case, it is undetermined that companies are charging more just because they can.

## 

## **Takeaways**

## **G1. Conclusions drawn**

The major takeaway from this project is that there appears to be a correlation between the data, and it is strong enough to fully conclude that there is a direct relationship between an airline's market share and the price they charge for a ticket. There is, however, no method to test with the given data whether the price increase is solely due to the market share. There are a wide range of factors that do play a role in the pricing of tickets; however, we cannot fully factor in all the possible reasons, nor the weight of each factor to a satisfactory degree.

## **G2. How chosen tools support findings and effective storytelling**

The choice of renaming the columns that were used, as well as creating a new data frame based on only those columns assisted greatly in cleaning up the data for a more manageable project. The data visualizations generated for this project demonstrate a more detailed understanding of the needed data as well as how it relates to the question at hand. The scatter matrix was a good starting point for data visualization, where the individual bar charts based on the top ten frequencies of the important values assisted greatly in making the data more accessible. The pie chart helped determine the market share percentages of the largest carrier as well.

The combination of the r-squared value, the z scores of each major column, and the mean of the p values helped prove the test results and their accuracy. Overall, the model was highly accurate based on the r-squared test, and the p values determined that the results were unlikely to be a result of random chance. The conclusions from the z-tests show that while there is a correlation.

## **G3. Recommendations of actions based on findings**

Since this project is based on an article, I found about my hometown airport, it was more of a hypothetical test to see if there was a correlation between market share and flight prices. It is wise for any traveler to think about what goes into the price of things, especially their flights. If they are okay with a company, knowing that they might get charged more then that is up to them. However, if they realize what is going on and they can make more informed decisions as it regards flights out of their home airport, then this project has accomplished its goal. An informed traveler is a better traveler, and it might make their experience better overall.

## **Panopto recording and Tableau public link**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1ea37f62-4ec4-4228-9d93-b1e8010882a9>

<https://public.tableau.com/app/profile/jesse.carter/viz/AirlineMarketShareProject/Dashboard1?publish=yes>

## 

## **Sources**

# References

Beers, B. (2024, 08 01). *P-Value: What It Is, How to Calculate It, and Why It Matters*. Retrieved from investopedia.com: https://www.investopedia.com/terms/p/p-value.asp

Bobbitt, Z. (2023, 02 28). *Pandas: How to Create Bar Chart to Visualize Top 10 Values*. Retrieved from statology.org: https://www.statology.org/pandas-bar-chart-top-10/

isitapol2002. (2024, 08 06). *How to find a P-Value from a Z-Score in Python?* Retrieved from geeksforgeeks.org: https://www.geeksforgeeks.org/how-to-find-a-p-value-from-a-z-score-in-python/

Jikadara, B. (2024, 08 01). *US Airline Flight Routes and Fares 1993-2024*. Retrieved from kaggle.com: https://www.kaggle.com/datasets/bhavikjikadara/us-airline-flight-routes-and-fares-1993-2024

reddit.com. (2023). *Finding max of the difference between two columns in dataframe*. Retrieved from reddit.com: https://www.reddit.com/r/learnpython/comments/176roka/finding\_max\_of\_the\_difference\_between\_two\_columns/

stackoverflow. (2023). *Pandas - Compute z-score for all columns*. Retrieved from stackoverflow.com: https://stackoverflow.com/questions/24761998/pandas-compute-z-score-for-all-columns

Waskom, M. (2024). *seaborn.heatmap*. Retrieved from https://seaborn.pydata.org: https://seaborn.pydata.org/generated/seaborn.heatmap.html