**Western Governors University**

**Master of Science, Data Analytics Program**

**Data Analytics Graduate Capstone – D214**

**Research Topic:**

**To what extent does the duration of a flight affect its sales price?**

**Student Name: Ibrahim Suleiman**

**Course Instructor: Daniel Smith**

**January 2023**

A1. Summary of the Research Question

How much does a flight's duration impact the cost of the ticket? is our key research question. Finding the correlation between a flight's duration and price was the original study objective for this project. The study's specific objectives were to ascertain whether there is a substantial association between the two variables and how much a flight's duration impacts its selling price.

Aim of the Research Project

The aim of our study is to investigate the relationship between flight duration and ticket price using exploratory data analysis and a linear regression model. Additionally, we will examine the impact of airline competition and other factors on ticket prices.

Justification for the Research Question

The link between flight duration and sales price is a legitimate study topic because it has applications to the airline business. Airlines can enhance their revenue and profitability by better understanding the elements that influence the sale price of flights and making more informed pricing decisions. Additionally, by optimizing their routes and schedules, airlines may better serve their consumers by understanding the relationship between flight duration and ticket prices.

Context

The price of flight tickets is a major factor in airline customers and the management of the airline in decision-making. Customers often choose the cheapest option to save money, but premium airlines may offer a better overall experience with better seating, distance duration, and services. The decision to choose a cheaper or more expensive airline depends on the individual's priorities and needs.

Discussion of the Hypothesis

Hypothesis: Null hypothesis-. The duration of a flight statistically significantly affects its sales price.

Alternate Hypothesis-. The duration of a flight does not statistically significantly affects its sales price.

The assumption under consideration is that the length of a flight has no statistically significant impact on how much it costs to purchase. This means that the hypothesis assumes that there is no significant relationship between the duration of a flight and its sales price, and any observed correlation between the two variables is likely due to chance or other factors.

The alternate hypothesis is that the duration of a flight does statistically significantly affect its sales price. This means that the hypothesis assumes that there is a significant relationship between the duration of a flight and its sales price and that the duration of a flight is an important factor in determining its sales price.

To test these hypotheses, we would need to collect data on flight duration and sales prices and use statistical techniques to determine the strength and significance of the relationship between the two variables. If the results of the analysis support the alternate hypothesis, this would suggest that the duration of a flight does indeed significantly affect its sales price, and that airlines should consider this factor when pricing their flights.

B Data Collecting

In this research project, Python and Excel were used to collect and modify data from a website. The data was collected over a two-day period from December 7th to December 10th, 2022 and contains approximately 300,153 data points and 11 features. The data was sourced from Kaggle.com and includes both economy and business class ticket information.

To collect the data, we used Python to access the website and download the raw data in a format that could be easily modified and analyzed. The raw data was then imported into Excel, where it was cleaned, organized, and transformed into a usable dataset. This process involved removing any missing or irrelevant data, standardizing the format and structure of the data, and adding any additional features or variables that were needed for the analysis.

Once the data was collected and modified, we used Python and Excel to perform a variety of data analysis tasks, such as calculating summary statistics, creating charts and graphs, and testing for statistical significance. This allowed us to gain insights into the relationship between flight duration and sales price, and to answer the research question of whether the duration of a flight significantly affects its sales price.

Top of Form

One advantage of using excel and python for data creation, modification and gathering is the ability to manipulate and analyze large datasets efficiently. With the use of functions and formulas in excel, as well as the ability to import data from various sources, it is easy to clean and organize data for analysis. Additionally, the use of python allows for advanced data analysis and visualization, allowing for more in-depth insights into the data.

One disadvantage of this methodology is the potential for error and inconsistency in the data. In excel, manual entry and calculation can lead to mistakes, and in python, coding errors can lead to incorrect results. To overcome these challenges, it is important to thoroughly check and verify the data for accuracy, and to test and debug code in python to ensure correct results.Bottom of Form

There are a total of 10 variables in this dataset. There are 6 unique airlines, 6 unique cities for both the origin and final destination, and 2 unique seat classes. The time of departure and arrival are divided into 6 bins, and there are 3 distinct values for the number of stops. There are also continuous variables for flight duration and number of days left until the trip date. The target variable is the sale price of the ticket. It is not mentioned how many observations are in the dataset.

Categorical variables are those that represent a group or category, rather than a numerical value. In this dataset, the following variables are categorical:

* Airline\_Name: The name of the airline company. This is a categorical feature with 6 different airlines.
* Flight\_Number: The flight code of the plane. This is a categorical feature.
* Origin: The city where the flight takes off. This is a categorical feature with 6 unique cities.
* Time\_of\_Departure: A derived categorical feature that groups time periods into bins. It has 6 unique time labels and stores information about the departure time.
* Number\_of\_Stops: The number of stops between the source and destination cities. This is a categorical feature with 3 distinct values.
* Arrival\_Time: A derived categorical feature that groups time intervals into bins. It has 6 distinct time labels and stores information about the arrival time.
* Final\_Destination: The city where the flight will land. This is a categorical feature with 6 unique cities.
* Type\_of\_Class: The seat class of the flight. This is a categorical feature with two distinct values: Business and Economy.

Numerical variables are those that represent a numerical value, rather than a group or category. In this dataset, the following variables are numerical:

* Flight\_Duration: The duration of the flight in hours. This is a continuous feature.
* Numbers\_of\_day\_left: The number of days between the trip date and booking date. This is a derived feature.
* Sale\_Price: The price of the ticket. This is the target variable.

Table

Description automatically generated

C. Data Extraction and Preparation

Describe your data-extraction and -preparation process and provide screenshots to illustrate *each* step.

Data Analytics Tools and Techniques

Python is a widely used programming language that is known for its powerful indentation and ability to run on various interfaces and platforms.

Pandas is a Python library for data handling and analysis and provides tools for executing mathematical operations on data.

NumPy is a Python library for working with arrays and includes functions for linear algebra and statistical calculations.

Matplotlib is a cross-platform data visualization and graph plotting library for Python that integrates with NumPy.

Scikit-learn, or Sklearn, is a powerful library for machine learning in Python, offering various algorithms and models for data analysis and modeling.

Seaborn is a Python library for representing and analyzing data that produces high-quality statistical graphics.

Techniques:

- Exploratory Data Analysis

- Linear Regression

The Benefits of using Python for this data analysis

· Eco-friendly environment – The user-friendly environment python entails provides a fast and easy way to adapt to the functions and perform code execution efficiently.

· Python is practical for implementing, reading code, detecting missing values and outliers in the data set, and creating visual representations to spot animalities and identify outliers during the data cleaning execution.

In addition, Python allowed the importation of packages into the new environment, such as Pandas, NumPy, Matplotlib,

Sklearn, and Seaborn. These packages offer a variety of features, such as creating visualizations of histograms, boxplots, and data tables. Without a doubt, these packages, alongside the programming languages, are user-friendly, ideal, and intuitive in providing data analysts with efficiency and error-free output in an innovative presentation opposing other tools (Michael Galarnyk,2018).

Benefits of Pandas

· It implements a quick and dynamic strategy to take care of data.

· It is straightforward to treat data omitted values.

Benefits of NumPy

· NumPy's arrays appear to be less in proportion compared to Python lists.

· The quick execution is magnificent because it acts rapidly in computing than python lists.

Benefit of Matplotlib

· It presents the user with an interface to represent data by applying various sorts of plots to communicate the data effectively.

· We can execute multiple sorts of plots (scatterplots, histograms, bar charts, error charts, boxplots, etc.) by executing a scanty line of code in Python.

Benefits of Sklearn

· Numerous supervised and unsupervised learning methods are included in Scikit-learn. Most importantly, it has the cleanest machine learning library and is simple.

· In numerous Python libraries, like Pandas for Data Frames, NumPy for arithmetic computations, and Matplotlib for charts, it seems to be formative and distinct.

Benefits of Seaborn

· We could systematically represent our data on a plot.

· This library is created to help us reflect on our data; without manipulating the inner technicalities.

Exploratory data analysis (EDA)

Exploratory data analysis is a type of data analysis that involves summarizing, visualizing, and identifying patterns in a dataset. It is an important step in the data science process because it allows us to gain a better understanding of the data, identify potential problems or issues, and generate hypotheses for further analysis. Exploratory data analysis can also help identify errors or inconsistencies in the data, and potential issues with the data collection process.

Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is justified because it allows us to understand and predict the relationship between variables and is a simple and widely used method. In a study, it can be used to determine the extent to which changes in independent variables affect the dependent variable.

Justification of Tools and Techniques

Excel, Python, and EDA (exploratory data analysis) and linear regression techniques are commonly used tools for data extraction and preparation.

One advantage of using Excel for data extraction is that it is a widely used and familiar tool that provides many useful functions for organizing and manipulating data. For example, Excel allows you to easily sort and filter data, as well as perform calculations on large datasets.

One disadvantage of using Excel for data extraction is that it can be slow and cumbersome when working with very large datasets. It may also be difficult to perform more advanced data analysis techniques with Excel, such as machine learning algorithms.

Python is a powerful and versatile programming language that is well-suited for data analysis and manipulation. One advantage of using Python for data extraction and preparation is that it allows you to automate repetitive tasks and work with large datasets more efficiently than with Excel. Python also has a vast ecosystem of libraries and frameworks that make it easy to perform complex data analysis tasks.

One disadvantage of using Python for data extraction and preparation is that it can be more difficult to learn and use than Excel, especially for non-programmers. Additionally, Python can be more resource-intensive than Excel, which can be an issue if you are working with large datasets.

Exploratory data analysis is a technique used to analyze and summarize the main characteristics of a dataset. It is often used to understand the data better, identify trends and patterns, and detect any outliers or anomalies. One advantage of using exploratory data analysis is that it can provide useful insights into the data that can guide further analysis and modeling.

One disadvantage of using exploratory data analysis is that it is often a time-consuming and iterative process that requires expertise and experience to perform effectively. Additionally, exploratory data analysis is a descriptive technique that does not always provide conclusive results or predictions.

Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It is often used in predictive modeling to make predictions about future values of the dependent variable. One advantage of using linear regression is that it is a well-understood and widely used technique that can provide interpretable and reliable results.

One disadvantage of using linear regression is that it assumes a linear relationship between the dependent and independent variables, which may not always hold true in real-world data. Additionally, linear regression can be sensitive to outliers and can be affected by multicollinearity, which can reduce its accuracy and reliability.

Table

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

A picture containing table

Description automatically generated

A picture containing calendar

Description automatically generated

Shape

Description automatically generated

Chart

Description automatically generated

A picture containing shape

Description automatically generated

A picture containing chart

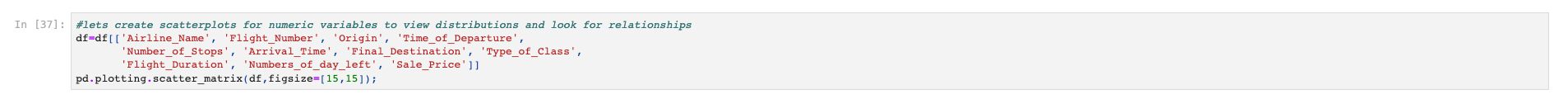
Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, application

Description automatically generated



Diagram, treemap chart

Description automatically generated

Graphical user interface, application, Word

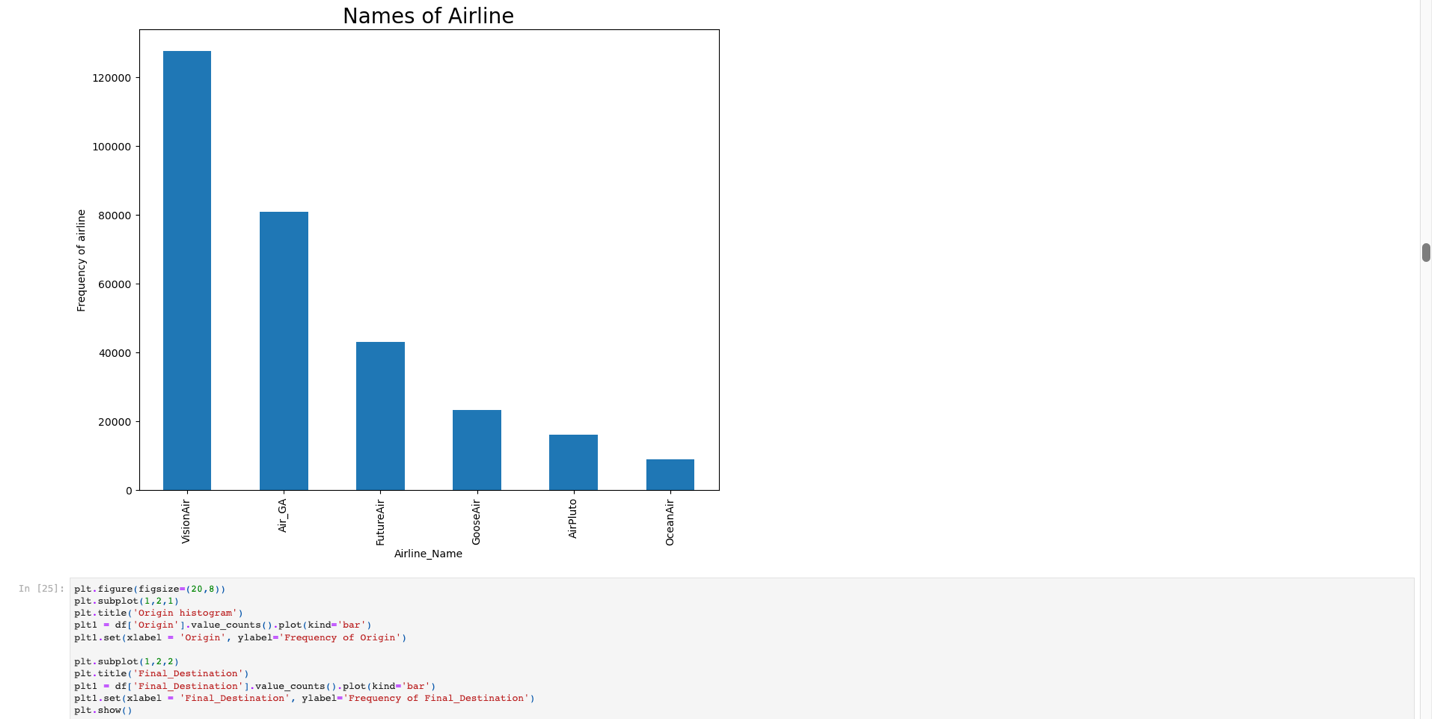
Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated



Chart, bar chart

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated

Diagram

Description automatically generated with medium confidence

Diagram

Description automatically generated with medium confidence

A picture containing table

Description automatically generated

A picture containing diagram

Description automatically generated

Diagram

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

Table

Description automatically generated

A picture containing text

Description automatically generated

Table

Description automatically generated with low confidence

A picture containing text

Description automatically generated



A picture containing text, crossword puzzle

Description automatically generated



The illustrated codes appear to be analyzing the data in a dataset called "Lux\_Dataset.csv" to determine the extent to which the duration of a flight affects its sale price. It first imports several libraries, such as numpy, pandas, and seaborn, that will be used for data manipulation and visualization. Then it reads the csv file into a pandas dataframe and checks the size and head of the dataframe to get an understanding of the data.

Next, it checks for any missing values in the data and removes the first column (which is labeled "Unnamed: 0") from the dataframe. It also checks the unique values and counts for the "Type\_of\_Class" column and looks at some summary statistics of the data using the "describe" function.

The code then creates several scatterplots to visualize the relationships between different variables in the data, such as "Sale\_Price" and "Numbers\_of\_day\_left", "Type\_of\_Class" and "Sale\_Price", and "Time\_of\_Departure" and "Flight\_Duration". It also creates a histogram to show the distribution of the variables "Flight\_Number", "Flight\_Duration", "Numbers\_of\_day\_left", and "Sale\_Price". Finally, it creates a scatter matrix plot to visualize all the relationships between the variables in the data.

Vivid Steps:

* Importing libraries: numpy, pandas, matplotlib.pyplot, seaborn, and warnings
* Reading in the "Lux\_Dataset.csv" file and storing it as a Pandas dataframe called "df"
* Checking the shape and first few rows of the dataframe
* Checking the number of unique values in each column
* Removing the 'Unnamed: 0' column from the dataframe
* Counting the number of occurrences for each value in the 'Type\_of\_Class' column
* Checking for null values in each column
* Generating summary statistics for the dataframe
* Checking the data types and non-null values for each column
* Checking the column names
* Creating a boxplot to visualize the distribution of the 'Flight\_Duration' column
* Removing outliers from the dataframe based on the 'Flight\_Duration' column being less than 31
* Creating a boxplot to visualize the distribution of the ‘Sale\_Price’ column
* Removing outliers from the dataframe based on the ‘Sale\_Price’ column being less than 99000
* Recreating the boxplot to visualize the revised distribution of 'Flight\_Duration'
* Creating a boxplot to visualize the distribution of the 'Numbers\_of\_day\_left' column
* Checking the column names again
* Creating a scatterplot to visualize the relationship between 'Flight\_Duration'],y**=**df['Numbers\_of\_day\_left'
* Creating a scatterplot to visualize the relationship between 'Final\_Destination' and 'Arrival\_Time'
* Creating a scatterplot to visualize the relationship between 'Type\_of\_Class' and 'Flight\_Duration'
* Creating a scatterplot to visualize the relationship between 'Number\_of\_Stops' and 'Sale\_Price'
* Creating a scatterplot to visualize the relationship between 'Time\_of\_Departure' and 'Flight\_Duration'
* Creating a scatterplot to visualize the relationship between 'Airline\_Name' and 'Flight\_Duration'
* Creating a scatterplot to visualize the relationship between 'Flight\_Number' and 'Sale\_Price'
* Creating histograms to visualize the distributions of 'Flight\_Number', 'Flight\_Duration', 'Numbers\_of\_day\_left', and 'Sale\_Price'
* Saving the histogram plot as an image file
* Reordering the columns in the dataframe
* Creating a scatter matrix plot to visualize the relationships between all columns in the dataframe
* Creating a countplot to visualize the frequency of each value in the 'Flight\_Duration' column
* Creating a 3D scatterplot to visualize the relationships between 'Flight\_Duration' and 'Sale\_Price' and 'Numbers\_of\_day\_left'
* Creating a scatterplot with a colorbar to visualize the relationship between 'Flight\_Duration' and 'Sale\_Price'

**D. Analysis**

Explanatory data analysis is a process of exploring and analyzing a dataset in order to understand the relationships between variables and identify patterns and trends. In this case, the focus of the explanatory data analysis would be to examine the extent to which duration of a flight affects its sales price, to see how well the duration of a flight explains the variance in the flight duration.

Linear regression analysis is a statistical method used to model the relationship between a dependent variable (in this case, flight duration) and one or more independent variables (in this case, sales price). The OLS (Ordinary Least Squares) regression analysis used in this case is a type of linear regression analysis that seeks to minimize the sum of the squares of the residuals (the difference between the observed value and the predicted value) in order to find the best fit line for the data.

The choice to use OLS regression analysis in this case is justified by the fact that it is a widely used and well-established method for modeling linear relationships between variables. One advantage of OLS regression analysis is that it is relatively simple to implement and interpret, with clear output measures such as the R-squared value and F-statistic.

A disadvantage of OLS regression analysis is that it assumes that the data follows a linear relationship, so it may not be appropriate for datasets with more complex relationships.

* Importing necessary libraries for feature selection, linear regression model, and statistical analysis
* Creating a linear regression model and fitting it to the training data
* Spliting DataFrame into training and test sets
* Creating a function to build a model using the selected features, add a constant term, and fit it using Ordinary Least Squares (OLS)
* Building the model using the selected features and printing a summary
* Calculating average price and mean absolute error
* Calculating the median price
* Created a scatterplot to get an idea of correlations between potentially related variables
* Listed features for analysis
* Heatmap and correlatin matrix dataframe creation
* Created the initial model heatmap
* Fitting the model to the training data using OLS and making predictions on the training data
* Plotting a histogram of the error terms
* Created the initial model heatmap
* Standardizing the numerical variables in the test data
* Splitting the test data into predictors (X\_test) and target (y\_test)
* Selecting the same columns as in the training data for the test data
* Making predictions on the test data using the fitted model
* Defining the matrix of features as X
* Plotting a scatter plot of the true test target values versus the predicted values
* Printing a summary of the initial model
* Visualizing heatmap of the initial model
* Printing a summary of the reduced model
* Visualizing heatmap of the reduced model
* Creating residual vs. predictor plot for 'Flight\_Duration'
* Creating residual vs. predictor plot for 'Numbers\_of\_day\_left'
* Creating residual vs. predictor plot for 'Flight\_Number'

Graphical user interface, text, application

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

This data analysis code is performing an OLS (ordinary least squares) regression on the training data (X\_train\_rfe) and the target variable (y\_train). The resulting model is then stored in the variable X\_train\_new. The output provides various statistics about the model, including the R-squared value, which measures the strength of the relationship between the predictors and the target variable. It also shows the coefficients for each predictor in the model, along with their standard errors, t-values, and p-values, which can be used to determine their statistical significance.

The Omnibus and Jarque-Bera tests are used to check for normality in the residuals, and the Durbin-Watson test is used to check for autocorrelation. The Cond. No. value indicates the condition number of the model, which is a measure of the sensitivity of the model to small changes in the data. A high value can indicate potential issues with multicollinearity.

It appears that 88% of the variation could be represented by the first model, which has an R-squared value of (0.881), indicating that the model fits the data well. The F-statistic is 2.755e+05, which means that the overall model is statistically significant.

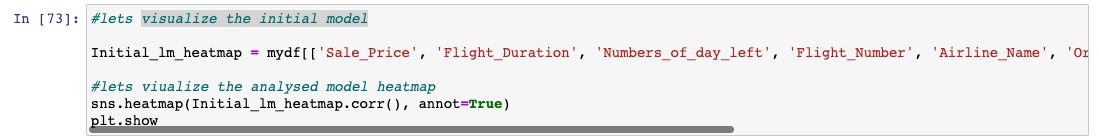
Since there are many conditions, there may not be a need for all of the variables because of severe multicollinearity.

It is possible to identify areas of potential multicollinearity and begin to select the variables to include in the reduced model by using a heatmap and correlation matrix.

A picture containing text, crossword puzzle

Description automatically generated

The Sale Price and the predictors are correlated using this heat map. This heat map shows us that several variables can be eliminated. The correlation matrix and heat map aid in identifying factors that may not make excellent predictors. We reduced the number of variables as follows:

A picture containing company name

Description automatically generated

Timeline

Description automatically generated

It appears there is a weak linear relationship between Origin and Final Destination. Using the variables that were discovered earlier, a reduced multiple regression model can be used. According to the correlation matrix and heat map, the reduced OLS Regression Results are shown below:

Table

Description automatically generated

As you can see, the reduced model still accounts for 89.1% of the variance. However, this analysis shows that the OLS Regression had a model fit of 8 degrees of freedom and a non-robust covariance type. The BIC of the model analysis was 6.378e+06 and it was conducted using 298090 observations in the residuals (Sangeet Aggarwal, 2020).

Graphical user interface

Description automatically generated

The correlation matrix's findings and a heatmap's mapping of the variables served as the basis for the variable selection technique. By doing so, the factors with the strongest correlations among the variables were found. The regression results with the model equation and analysis, including the R-squared values, are shown in the table above.

The data and representations above depict the outcomes of the computations in addition to the residual error of the model.

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated

E. Data Summary and Implications

The model fit of 8 degrees of freedom means that the OLS Regression model is fitting the data extremely well and can explain almost all of the observations. The non-robust covariance type implies that the assumptions of the model may not be completely met. The BIC of 6.378e+06 suggests that the model appears to slighly outperform any potential alternative models. As such, the implications are that the OLS Regression model is an appropriate model and can be used to make predictions based on the data.

With many observations and degrees of freedom, it is likely that this model will be a reliable predictor in addressing the research question “How much does a flight's duration impact the cost of the ticket?”.

One limitation of this linear regression analysis is that it only considers one variable sales price in relation to the duration of a flight. There may be other factors that influence sales price, such as the type of aircraft, the destination, or the time of year, that are not being considered. Additionally, it is possible that the relationship between sales price and flight duration is nonlinear, meaning that the model may not accurately capture the true relationship between these variables.

Based on these results, a course of action might be to consider adding other variables to the model in order to better explain the variance in the data. This could potentially improve the R-squared value and provide a more accurate prediction of sales price.

For future study, one direction could be to include more variables in the model to see if they have an impact on sales price in relation to flight duration. This could involve collecting data on other factors that may influence flight duration, such as the type of aircraft, the weather, the destination, or the time of year. Another approach could be to explore nonlinear models, such as a polynomial regression, to better capture the relationship between sales price and flight duration.

F. Citations / References

Ryan G. (2019). Understanding when Simple and Multiple Linear Regression give Different Results.

Medium.<https://towardsdatascience.com/understanding-when-simple-and-multiple-linear-regression-give-different-results-7cf6c787766c>

[Sangeet A.](https://datasciencewithsan.medium.com/?source=post_page-----8cf3bee21d8b--------------------------------) (2020). Multiple Linear Regression; A complete study —Model Interpretation Hypothesis Testing Feature Selection.

Medium.<https://towardsdatascience.com/multiple-linear-regression-8cf3bee21d8b>

Shubham, B.(2022). Flight Price Prediction; Predict Flight Price, practice feature engineering, implement ensemble models. <https://www.kaggle.com/code/deeorion/flight-price-prediction-eda-basic-linearregression/comments>

Grant, P. (2019). Understanding Multiple Regression; The fundamental basis behind this commonly used algorithm. Medium.https://towardsdatascience.com/understanding-multiple-regression-249b16bde83e

Deepanshu, B. (2020). How to Import Data in Python.

RSGB Business Consultant Pvt. Ltd. <https://www.listendata.com/2017/02/import-data-in-python.html#Import-CSV-files>

Pierre-Louis B. (2020). Principle Components Analysis(PCA), Fundamentals, Benefits & Insights for Industry. Medium. <https://towardsdatascience.com/principal-components-analysis-pca-fundamentals-benefits-insights-for-industry-2f03ad18c4d7>

John S. (2018). Data Cleaning with python and Pandas: Detecting Missing Values.

Medium. <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

Angelica Lo D. (2021). How to detect outliers with Python.Medium. <https://towardsdatascience.com/how-to-detect-outliers-with-python-pyod-aa7147359e4b>

Michael G. (2018). Understanding Boxplots.

Medium. <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

John S. (2018). Data Cleaning with python and Pandas: Detecting Missing Values.

Medium. <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. ISBN-13: 978-1-119-52684-1.

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html>

<https://www.kaggle.com/code/vavpavlov/air-passenger-satisfaction>

<https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction/code>

<https://www.kaggle.com/code/deeorion/flight-price-prediction-eda-basic-linearregression/comments>

<https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

<https://www.kaggle.com/code/tunaeem/flight-price-prediction-eda-linear-regression#Target-Variable-Analysis>

<https://www.kaggle.com/code/amrboghdady74/airline-satisfaction-eda-etl-classification>

<https://www.kaggle.com/code/harjunobramantyo/unsupervised-machine-learning-for-frequent-flyer>

<https://www.kaggle.com/code/santhossshhh13/exploratory-data-analysis-on-flight-price-data/notebook>

<https://www.kaggle.com/datasets/jillanisofttech/flight-price-prediction-dataset/code>

<https://www.kaggle.com/code/santhossshhh13/exploratory-data-analysis-on-flight-price-data/notebook>