**Data Analytics Capstone Topic Approval Form**

The guidelines for the Data Analytics Graduate Capstone course require a student to demonstrate the "application of academic and professional abilities developed as a graduate student. The capstone challenges students to integrate skills and knowledge from several program domains into one project." As a result, it is highly recommended that your topic should resolve a current or perceived business problem. As stated, you want to exemplify what you have learned in your MSDA program to showcase your skills. Remember, your research topic exemplifies scholarship and research at the highest level and is significant and helpful to potential employers in identifying your abilities. Your research will always be something you can look back on with pride. Finally, it is recommended for students to use publicly available data sets for transparency and external validity.

The purpose of this document is to help you clearly state the research question you will be exploring in your capstone project, your project’s scope, and your timeline in order to ensure that these align with your degree emphasis. Without clearly defining each of these areas, you will not have a complete and realistic overview of your project, and it cannot be accurately assessed whether your project will be acceptable for the purposes of these courses. Of course, if this a project that you have already completed at work or elsewhere, this should be easy to fill in! Many students do use a project they have already completed in the past year or two. In that case, you will write the proposal as if the project had not been done yet, and when you report on your project, you will use your complete after-implementation report. If you have not yet done your project, this document can help make sure the scope is within the acceptable range for this capstone. A course instructor will approve this form before submitting this task for evaluation. The task will not be evaluated without a course instructor’s signature. The course instructor may ask for additional information before approving the form.

Before submitting this form for approval, please remove all italicized directions in the form.

Please only submit a Topic Approval Form that has been signed by a course instructor for evaluation.

**Student Name: Ibrahim Suleiman**

**Student ID: 001429984**

**Capstone Project Name:** "The Effect of Flight Duration on Sales Price: A Regression Analysis"

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

This project does not involve human subjects research and is exempt from WGU IRB review.

**Research Question:**

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. The research questions we aim to answer include: To what extent does the duration of a flight affect its sales price?

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.

**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.

**Data:**

The dataset contains information about flight prices options from Kaggle.com. The dataset has 300,153 data points and 11 features.

**Variables**

1. Airline\_Name: The name of the airline company. This is a categorical feature with 6 different airlines.
2. Flight\_Number: The flight code of the plane. This is a categorical feature.
3. Origin: The city where the flight takes off. This is a categorical feature with 6 unique cities.
4. 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.
5. Number\_of\_Stops: The number of stops between the source and destination cities. This is a categorical feature with 3 distinct values.
6. 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.
7. Final\_Destination: The city where the flight will land. This is a categorical feature with 6 unique cities.
8. Type\_of\_Class: The seat class of the flight. This is a categorical feature with two distinct values: Business and Economy.
9. Flight\_Duration: The duration of the flight in hours. This is a continuous feature.
10. Numbers\_of\_day\_left: The number of days between the trip date and booking date. This is a derived feature.
11. Sale\_Price: The price of the ticket. This is the target variable.

Text, table

Description automatically generated

**Data Gathering**:

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

**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

**Justification of Tools/Techniques**:

**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**

· Scikit-learn entails diversely supervised & unsupervised learning algorithms. Most significantly, its simplicities as well as the cleanest machine learning library.

· It appears to be formative and unify distinctively along several Python libraries, including Matplotlib for charts, Numpy for arithmetic calculations, as well as Pandas for DataFrames.

**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 (EDA) 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 researchers to gain a better understanding of the data, identify potential problems or issues, and generate hypotheses for further analysis. EDA 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 researchers 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.

**Project Outcomes:**

The expected outcome of this research project is to identify the extent to which the duration of a flight affects its sales price. This will be done by using exploratory data analysis and linear regression to analyze the relationship between the two variables. The final outcome will be an interpretation of the model's performance, as indicated by the R-squared value, and its implications for the relationship between flight duration and ticket prices.

**Projected Project End Date: 03/31/2023**

**Sources:**

Shubham, B.(2022). Flight Price Prediction; Predict Fllight Price, practise 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>

<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>

**Reference**

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

**Course Instructor Signature/Date**:

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.