Data Management and Data Analytics Capstone Topic Approval Form

Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Student Name: Paul Smith

Student ID: 000412833

Capstone Project Name: Analysis of Airline Delay and Cancellation Data, 2009 - 2018

Project Topic: This project will examine airline delay and cancellation data from 2009 to 2018 to determine the causes of flight delays.

Research Question: What causes airline delays?

Hypothesis: The overall reason for airline flight delays is Late Aircraft Delay.

Context: A 2010 report sponsored by the Federal Aviation Administration (FAA) analyzed a variety of cost components caused by flight delays. This included the cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The report concluded that the total cost of all US air transportation delays in 2007 was $32.9 billion. Clearly, flight delays are a serious and widespread problem in the US. An exploration of multi-year (2009 – 2018) airline delay and cancellation data will be made to determine what are the causes of airline delays.

Data: I will collect Airline Delay and Cancellation Data datasets from www.kaggle.com. This data has been combined from multiple US Government (Bureau of Transportation Statistics) datasets.

The source of the US Government data:
https://www.bts.dot.gov/browse-statistical-products-and-data/bts-publications/airline-service-quality-performance-234-time

These Kaggle datasets will be combined to form multi-year data in an effort to offer additional insights:

- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2009.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2010.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2011.csv

- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2012.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2013.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2014.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2015.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2016.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2017.csv
- https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-2018/data?select=2018.csv

The datasets are owned by the United States Department of Transportation's Bureau of Transportation Statistics. This department was founded in 1966 and its mandate was to collect and disseminate transportation statistics (https://www.bts.dot.gov/learn-about-bts-and-our-work/history-bts). It provides various publicly downloadable datasets (https://www.bts.dot.gov/airline-data-downloads). I will not use any restricted, private, or propriety data.

Data Gathering: I will download tabular data from the Kaggle website as detailed earlier. I will cleanse and combine the data as needed.

Data Analytics Tools and Techniques: I will perform exploratory data analysis and produce a correlation matrix to show whether various delays can be predicted from the others. Delayed flight data and cancelled flight data will also be examined to determine if these factors influence airline delays, otherwise those data will be removed from the dataset to simplify the analysis. I will use Python with a Jupyter Notebook to assist in the data acquisition, cleansing, manipulation, and analysis.

Justification of Tools/Techniques: To explore the potential correlations between the airline delays I will need to perform bivariate and potentially multivariate analysis. This will involve producing a correlation matrix and/or scatter plots as needed to refine the results.

Application Type, if applicable (select one):

- ☐ Mobile
- ☐ Web
- ☒ Stand-alone

Programming/Development Language(s), if applicable: Python, Jupyter Notebook.

Operating System(s)/Platform(s), if applicable: N/A (cross-platform).

Database Management System, if applicable: N/A.

Project Outcomes: The project will produce a clean dataset from the Kaggle datasets (derived from the US Government source data). Code used in the data analysis will be included. A report will be produced that summarizes airline delays, and any correlations between the various delay causes.

Projected Project End Date: 8/31/2022

Sources: Total Delay Impact Study, October 2010, https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

---

Human Subjects or Proprietary Information
Does your project involve the potential use of human subjects? (Y/N): N
Does your project involve the potential use of proprietary company information? (Y/N): N

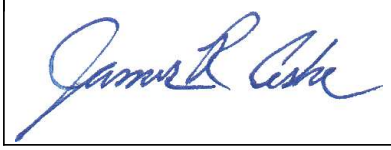
STUDENT SIGNATURE

P. Smith_____

By signing and submitting this form, you acknowledge that any cost associated with the development and execution of your data analytics solution will be your (the student) responsibility.

TO BE FILLED BY A COURSE INSTRUCTOR

The capstone topic is approved by a course instructor.



Monday, August 29, 2022

Project Compliance with IRB (Y/N): Y