**CA682 Data Management And Visualisation - Assignment**

|  |  |
| --- | --- |
| Student’s name(s): | Sarath Mukundan Adavakkat,  Suhara Fathima Shahul Hameed |
| Student Number(s): | 23267908, 23266257 |
| Program: | Master of Science in Computing (Data Analytics) |
| Project title: | Airlines with the highest departure and arrival delays from 2021 to 2023 |
| Module Code: | CA682 |
| Lecturer: | Dr Suzanne Little |
| Assignment Submission Date: | 1 December 2023 |
| GitLab Repository: | <https://gitlab.computing.dcu.ie/shahuls2/data-visualisation-assignment> |
| Presentation Link: | <https://drive.google.com/file/d/1ZARnGwZBUdz4gUV2SFiDstXbarYeQ87d/view?usp=drivelink> |

Table of Contents

[Declaration 3](#_Toc151938839)

[Abstract 3](#_Toc151938840)

[Dataset 4](#_Toc151938841)

[Data Exploration, Processing, Cleaning and/or Integration 4](#_Toc151938842)

[Data Visualisation 5](#_Toc151938843)

[Conclusion 7](#_Toc151938844)

[References 7](#_Toc151938845)

# **Declaration**

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious. I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy. I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. Any use of generative AI or search will be described in a one-page appendix including prompt queries.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work. By signing this form or by submitting this material online I confirm that this assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy.

Name: Sarath Mukundan Adavakkat Date:01/12/2023

Name: Suhara Fathima Shahul Hameed Date: 01/12/2023

# 

# **Abstract**

Airline delays, which involve both departure and arrival delays, are a critical feature of air travel. Departure delays are when flights do not take off on time, and arrival delays are when flights land later than expected. Passengers, airlines, and airport operations are all affected by these delays, which have an influence on schedules, expenses, and overall consumer satisfaction. Analyzing delay data allows stakeholders to discover trends, allowing them to implement initiatives that increase timeliness, optimize operations, and improve the travel experience for everyone involved in the aviation business.

The visualization provides a detailed examination of airline punctuality from 2021 to 2023, including departure and arrival delays. This visualization highlights the top airlines most affected by delays, providing insights on their performance throughout the selected timeframe.

# **Dataset**

The data was collected from the Kaggle website.

Link: <https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023>

Firstly, the Volume for the dataset is 5.04 GB which consists of 29,380,334 rows total and after cleaning 16,867,390 rows. It contains a large amount of flight-related data gathered over a number of years (2019-2023), potentially collecting significant records. Secondly, the dataset contains a variety of structured data items (such as flight details, dates, and airlines). So, the Variety property is also satisfied.

Therefore, we have considered 2V’s- Volume and Variety properties of Big Data.

# **Data Exploration, Processing, Cleaning and/or Integration**

The original dataset consists of data from 2019 to 2023, recorded in separate CSV files for each year. By focusing the analysis on the years 2021-2023, the volume of data considered for analysis was narrowed down. Pandas were utilized to store the data in data frames, which were then merged into a single data frame. Python was utilised for cleaning the datasets.

The dataset initially comprised 33 columns, which most likely held various types of information for delay and cancellation. The cancellation information column was eliminated as the focus was mainly on delay. After removing 19 columns judged unsuitable for analysis, the dataset was reduced to 14 columns. This reduction could imply concentrating on select features of flight data essential to the investigation while ignoring others.

The airline codes were replaced with respective names from a dictionary generated from the AIRLINE\_CODE\_DICTIONARY.csv file. The flight date(FL\_DATE) column was converted to timestamp format from string object which was then sorted according to dates and omitted the rows with missing values demonstrating data manipulation and preparation. However, it was also critical to ensure that the exclusion of these rows does not have a major impact on the overall integrity of the dataset.

This data preparation process, which includes filtering columns, dealing with missing values, converting formats, and mapping codes to names, is critical before undertaking in-depth analysis. It's an important step in ensuring that the data utilized for subsequent analysis is consistent, relevant, and reliable.

# **Data Visualisation**

Design process including a sketch of planned visualisation:

A graph and line drawing on a white paper

Description automatically generated

Each airline's departure and arrival delays are concentrated for visualisation. The difference in minutes between the scheduled take-off time and the true take-off time is known as the departure delay of the flight. The arrival delay is the time difference between the scheduled arrival time and the flight's actual departure time, measured in minutes.

Positive delays show that the flight was delayed, while negative delays indicate that the flight departed/arrived early.

The distribution of delays across different airlines, specifically departure (Dep\_Delay) and arrival (Arr\_Delay) delays, provides a thorough insight of how these delays vary within each airline. The identification of variations in airline delay patterns was explored. This investigation aids in understanding typical delay ranges and delay frequency. Bar charts are useful for comparing and displaying the number of delays experienced by various airlines.

Tableau Desktop was chosen as the visualization tool and the cleaned data was loaded for further analysis.

**Which Airlines show the most delay with respect to counts of departure and arrival delays from 2021 to 2023?**

Screenshot or image of visualisation:

A graph with a line going up

Description automatically generated

Choice of chart or graph type and design:

1. A graph was plotted that demonstrates the comparison of airlines over time on departure and arrival delays.
2. The chart used for visualization was a bar chart and a line chart for departure delay and arrival delay respectively.
3. The airline is a nominal variable which is represented in the x-axis, and the count of departure and arrival delays is represented by two y-axes.
4. The combination of the bar and line charts allows us to see the relationship between the two variables throughout the same time period.
5. Because the bar graph takes up the majority of the area and the foreground of the graph, we chose a gentle purple colour that relaxes the viewer's eye.
6. While the line graph occupies a relatively little surface area and is in the background, we have coloured it teal to contrast with it and catch the viewer's attention.
7. The title is short and axes were labeled with appropriate labels and prominent fonts so that users could readily comprehend them.
8. The grid lines have been eliminated to keep the graph's design as simple as possible.
9. Because tableau graphs are responsive, the graph can be hovered to obtain the exact values.

Interactivity or animation and communication message from the graph

1. It was observed that, just after the Covid pandemic the number of flights that were flown was less which showed a less delay pattern in the first quarter of 2021.
2. Travelling during the holidays will have a significant impact on the number of delays experienced by any airline.
3. A gradual increase pattern was observed in the delay of Southwest Airlines, American Airlines, Delta Airlines, United Airlines and SkyWest Airlines during the second and third quarters in 2021, 2022 and 2023. They were accountable for most of the delays, as a result of the huge amount of flights they provided.
4. This data collection and accompanying visualizations show a consistent pattern in flight delays, but this does not imply that they can be extrapolated to the rest of the year or future years.

# **Conclusion**

List of tools or libraries used to create the visualisation

All of the cleaning and processing of the dataset was done in Python, and Tableau Desktop was chosen as the visualization tool.

Outcome of visualisation

Several airlines were compared with respect to the count of flight departure and arrival delays. Southwest Airlines, American Airlines, Delta Airlines, United Airlines and SkyWest Airlines showed high delays in the selected time period of 2021-23. There was a significant shift in airline delays over three years.

There was a functionality that we were technically unable to plot the graph with two separate filters for the bar and line graph individually in Tableau. But this was accomplished with logical conditions when rows were selected to plot.

There was an equal contribution on our side from the initial stage of dataset searching, data cleaning and processing, data visualization, making report, and presentation. At first, even though we found many datasets of our interests, they didn’t satisfy the 3V’s of big data properties.

# **References**

[1]www.kaggle.com. (n.d.). Flight Delay and Cancellation Dataset (2019-2023). [online] Available at: <https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023> [Accessed 29 Nov. 2023].

[2]help.tableau.com. (n.d.). Tutorial: Get Started with Tableau Desktop. [online] Available at: [https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial home.htm](https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial%20home.htm).

[3]help.tableau.com. (n.d.). Visual Best Practices. [online] Available at: <https://help.tableau.com/current/pro/desktop/en-us/visual_best_practices.htm>.