Assignment 2 (20%)
STQD6324 Data Management
SEMESTER 2 2024/2025

Airline on-time performance

Have you ever found yourself stranded at an airport due to a delayed or cancelled flight, wondering if the situation could have been anticipated? This project offers you the opportunity to explore airline performance data to identify patterns and factors affecting flight delays and cancellations.

You will be working with the "Airline On-Time Performance" dataset from Kaggle: https://tinyurl.com/u8rzvdsx, which contains airline performance data from 1987 to 2008. Please refer to the appendix to find the specific year assigned to you for analysis.

The challenge:

Using the dataset for your selected year, analyze the data and answer the following questions:

1. **Delay Patterns:**
   - What times of day (morning/afternoon/evening) have the lowest average delays?
   - Which days of the week show better on-time performance?
   - During which months or seasons are flights most likely to be on time?

2. **Delay Factors:**
   - Identify and rank the top 3-5 factors contributing to flight delays, based on the delay categories provided in the dataset.
   - Quantify the impact of each factor (in minutes of delay and percentage of total delays).

3. **Cancellation Analysis:**
   - Identify the primary reasons for flight cancellations as categorized in the dataset.
   - Determine if cancellations correlate with specific airlines, airports, or time periods.

4. **Problematic Routes:**
   - Identify specific routes (origin-destination pairs), carriers, or flight numbers that show consistently poor performance.
   - Analyse the reasons these particular flights are prone to delays or cancellations.

For each question, use either **Pig** or **Hive** to extract insights from the dataset, and generate figures to explain your findings using **Python** or **R**. Use your creativity to answer these questions, and you are encouraged to search online for ideas on how others have approached similar challenges, such as on this link: https://tinyurl.com/bdejna9e.

Please ensure that the scripts and codes used to generate your findings are included in the main report. You can use **Jupyter Notebook** with markdown for this purpose. The submission deadline is **2025-06-08**, and please share your work through **GitHub**.

Name	Year
LING WEI JIET	1989
SITI NURNIERA BASMA	1990
MAO JINLIN	1991
TAN YIDAN	1992
GAO YURU	1993
ZHAO WANPENG	1994
YANG FANGJIN	1995
ZHANG ZHUORUI	1996
TENG SIYAN	1997
LI, TAORUI	1998
JIANG ZITIAN	1999
NG CHIN WEN	2000
WANG RONGCHENG	2001
MIAO YAOWEI	2002
FARAH SYAHIRAH	2003
TAO HUANXIN	2004
ADAM SUHAIL	2005
AHMAD HATHIM	2006
AZRUL ZULHILMI	2007
NUR YASMIN NADHIRAH	2008

Criteria	Marks		
Reproducibility	**5** The notebook is 100% reproducible	3 The notebook is reproducible with a few missing steps	1 The notebook is not reproducible
Plots	**10** All the plots are i. suitable, ii. easy to understand iii. observations are properly explained	5 Some of the plots are i. suitable, ii. easy to understand iii. observations are properly explained	3 The plots are i. not suitable, ii. hard to understand iii. observations are poorly explained
Overall GitHub presentation	**5** The overall GitHub is i. properly structured, ii.each section neatly organized, iii. easy to follow	3 Part of the GitHub is i. properly structured, ii.each section neatly organized, iii. easy to follow	1 The GitHub is i. poorly structured, ii. each section is not organized, iii. hard to follow