Data Management and Exploratory Data Analysis CSC8631 Coursework (Semester 1, 2021) Report

Name: Sampath
Student id: 210449525

Introduction

Having completed some sort of data analysis, we often want to automate that process so that it will be executed at regular intervals. What that means is that code must to generated so that data acquisition, data cleaning, model development, document creation, and other components are fully executed from start to finish without any intervention from a human. At this point there is a need for Data pipeline.
A **data pipeline** is a set of actions that ingest raw data from disparate sources and move the data to a destination for storage and analysis. A pipeline also may include filtering and features that provide resiliency against failure.

To understand how a data pipeline works, think of any pipe that receives something from a source and carries it to a destination. What happens to the data along the way depends upon the business use case and the destination itself. A data pipeline may be a simple process of data extraction and loading, or, it may be designed to handle data in a more advanced manner, such as training datasets for machine learning.

Source: Data sources may include relational databases and data from SaaS applications. Most pipelines ingest raw data from multiple sources via a push mechanism, an API call, a replication engine that pulls data at regular intervals, or a webhook. Also, the data may be synchronized in real time or at scheduled intervals.

Destination: A destination may be a data store — such as an on-premises or cloud-based data warehouse, a data lake, or a data mart — or it may be a BI or analytics application.

Transformation: Transformation refers to operations that change data, which may include data standardization, sorting, deduplication, validation, and verification. The ultimate goal is to make it possible to analyze the data.

Processing: There are two data ingestion models: **batch processing**, in which source data is collected periodically and sent to the destination system, and **stream processing**, in which data is sourced, manipulated, and loaded as soon as it's created

Workflow: Workflow involves sequencing and dependency management of processes. Workflow dependencies can be technical or business-oriented.

About Source code

For this project I have used "cyber-security-1_question-response" dataset and I am applying most of the data pipeline fundamentals to extract some of the useful insights from the data.

I used dplyr package for the analysis. Then I imported the dataset "cyber-security-1_question-response". The dimensions of the dataset is
76601 Rows and 7 Columns. Then i analyzed some data and dropped some unnecessary columns. And dropped the rows which are having empty cells. I prefferred to drop the rows because the number of rows having empty cells are very less as compared to the totally filled rows. Then I find out how many distinct values are present in each and every column of the dataset. So here I completed the preprocessing part.

Now came the analysis part. So after seeing the columns I framed some questions.
1. Which question is mostly correct in between all quizzes?
For this question I took "question_number" and "correct" columns and after grouping. I concluded that question 1 is mostly correctly marked and incorrectly marked.

2. Total number of Corrected questions and Incorrected questions?
For this question I took "correct" column and took sum of true and false values so I concluded that

```
"Total number of corrected question: 41949"
"Total number of incorrected question: 34652"
```

3. Top 10 high scorers?
For this question I took "learning_id" and "correct" column and after grouping and sorting I concluded that "2f483763-596f-4ede-ba44-20937ce7eda9" is the top scorer with 24 correct answers.

4. Which quiz question is most correctly marked?
For this question I took “quiz_question” and “correct” columns and after grouping and sorting. I concluded that question “1.7.5” is mostly correctly marked for 3237 times.

5. Which quiz question is most incorrectly marked?
For this question I took “quiz_question” and “correct” columns and after grouping and sorting. I concluded that question “1.7.6” is mostly incorrectly marked for 4025 times.

6. Which step number got the most correctly marked?
For this question I took “step_number” and “correct” columns and after grouping and sorting. I concluded that step “7” is mostly correctly marked for 19002 times.

7. Which step number got the most incorrectly marked?
For this question I took “step_number” and “correct” columns and after grouping and sorting. I concluded that step “18” is mostly incorrectly marked for 13690 times.

These are the insights I can conclude from this dataset.