COMP30770 – Programming for Big Data
Group Project Plan
Lindsay Atkinson (25205120), Ethan Epperson (25201495)

Question 1 – Introduction

The dataset that will be used for this project comes from Open Psychometrics. It includes over a million answers to an online questionnaire designed to assess five different behavioral traits and assign personality types. The data was gathered between 2016 and 2018.

In terms of volume, challenges in handling this dataset could arise due to the large number of responses. Processing over 1 million responses requires significant computational power and time, thus frequent analysis is challenging and resource-intensive. Accordingly, it will be important to thoroughly review the code before execution to prevent unnecessary reruns. Additionally, using big data frameworks to split and process data across multiple servers will allow for improved speed and efficiency.

This dataset is highly structured, with the majority of data being numerical values ranging from 1 to 5, based on how closely the respondent aligns with the statement. While the consistency makes the data easier to work with, it also limits data diversity, as it lacks quantitative or categorical insights that could offer a deeper analysis. The dataset does offer some external comparison points, such as timestamps, geographical locations, and screen size; however, additional factors like socioeconomic status, cultural background, etc. would allow for a more interesting analysis of how personality types differ with life experiences. Incorporating additional datasets to analyze these correlations would be an exciting next step for this project.

Link to dataset: https://www.kaggle.com/datasets/tunguz/big-five-personality-test

Question 2 – Objective

The main objective of analyzing this dataset is to cluster the individual responses into different personalities, and discover which are the most common. Then, exploring how the common personalities correlate with data on the respondents' interaction with the questionnaire. By examining the time of response, time taken to complete the survey and time taken on the finalization page, the study aims to identify if there are patterns in terms of how different personality types engage with the questionnaire.

Question 3 – Small Data Analysis Tasks

To achieve the required objective of recognizing prominent personality types and how they interact with the questionnaire are as follows, several small data analysis tasks will be completed. First, responses will be clustered based on similar answers and organized into tables with proper classification of personality types/traits. From these clusterings, the most prevalent personality types will be identified to determine the dominant associations. Then, the dominant personalities will be compared with the behavioural data such as time elapsed during the survey, screen size, and time required to finalize their answers to see if there are any correlations. Finally, the personalities and their behaviours will be compared to their geographical locations to see if geography plays a systematic role in respondents behaviour.

Question 4 – Big Data Analysis Tasks

To improve the efficiency of the small data analysis tasks, MapReduce can be employed to enable optimized aggregation of the responses, allowing for rapid identification of popular personality types and associations. Additionally, Hadoop can be used to partition data into countries

- Q4: How to use Big Data Technologies (e.g., MapReduce on Hadoop or Spark, etc.) to improve those computation/storage-intensive tasks (answers can be tentative)?