King Saud University
College of Computer and Information Sciences

Department of Information Technology

IT362: Principles of Data Science

Logbook

Prepared By

Section	Student Name	ID
80657	Afnan Alkharji	443200897
	Deema Alfarhoud	444200955
	Hussah Alotaibi	444201039
	Falwa alkhalifah	444200745
	Sana Alotaibi	444200426

Supervised by:

Dr. Abeer Aldayel

Phase	Task	Date	Description	Challenges
Phase 1: Data Collection, Research, and Assessment	Accessing GitHub API to collect data	September 18th	We started using python in google colab to apply the collecting code. We decided to collect 1000 observations for each year (250 per quarter) except for 2025 we ended up collecting almost 750 observations. The features we decided to collect for each repository were ["name", "owner", "url", "description", "stars", "forks", "created_at", "pushed_at", "primary_language", "all_languages_bytes","topics", "contributors_count"] to expand on the study.	API Rate Limits & Time-consuming collection, we solved this issues by dividing the task on the group members to get it done faster.
	Finalizing the raw data file	September 23rd	We gathered all the data collected by the group in one file and made sure they are compatible.	Merging all years' files together smoothly, we solved this issue by merging them via python then checking manually.
	Data checking and visualizing.	September 24th	We used python libraries such as numpy and matplotlib and seaborn to: -Check for any duplicate rows or missing values, there were no duplicates, but we found some missing values and we decided to only remove rows with missing "Primary programming language" value. -Transformed "stars", "forks" to numeric valus. -Use diagrams to showcase "Distribution of repository stars", "Top 10 programming	Hesitation regarding wither to remove all rows with any missing values or only the ones that misses "Primary programming language", eventually we decided to only remove rows with missing "Primary programming language" value only.

			languages used", "Repositories created over time", and "Stars vs Forks scatter plot".	
	Writing the 1st phase's report	September 22nd	We started writing the 1st draft for this phase's report.	No challenges faced.