Group Contract

Group 16
Group Members:

Name	Student Number
Loay Al Abri	8483229
Roy Oh	57552671
Sachit Sharma	15141393
Anne Zhou	34844902

(1) Goals

What are our team goals for this project?

- We aim to provide a high quality output, where every group member contributes effectively.
- Complete the project within the timeline.
- Create a work environment where every group member enjoys and learns something new.

What do we want to accomplish?

- We want to accomplish an insightful project that can lead to productive outcomes that could be applied to a wider context.
- Build a strong relationship with other team members.

What skills do we want to develop or refine?

- Develop writing reproducible and readable codes.
- Apply statistical modelling methods to the GitHub Repositories dataset to extract useful insights from the data.
- Build a better understanding of what we've studied in the classroom.

Expectations

- We expect each member to keep on track and communicate frequently to prevent gaps in understanding.
- Instagram is the main way of communication between group members.
- Meetings will be scheduled through the Instagram group chat. We should have a meeting before submitting every group deliverable.
- We expect everyone to be familiar with the material in class and put a good amount of effort into the project.

(2) Policies & Procedures

We will give each other weekly feedback if a member tends to fall behind. In addition, if a group member fails to fulfil their part in the contract, it will be clearly stated in the peer evaluation stage.

(3) Consequences:

If a team member shows any misconduct, we can talk to them first about the issue and come up with a collaborative solution.
If a group member displays a free-rider(piggybacking) behaviour, it will be discussed with the TAs and other group members, and we will decide accordingly the consequences of their actions.

Data:

We would like to explore the following about the data.

- What variables are crucial in predicting how many starts a GitHub repository will get?
  - We can analyse different models created to predict GitHub repository stars to determine which variables are useful in predicting the response.
- Explore the correlation between variables.
  - We will achieve this by visualizing scatterplots of the variables and fitting linear regression models accordingly. Also, we shall compute the correlation coefficient between numeric variables.
- Explore the most popular topics using the description column.
  - We can achieve this by using some Natural Language processing techniques to group repositories with the same topic, and then rank them based on the mean number of stars.
- What is the relationship between the number of starts a Github and the number of issues and forks it has? Is a linear model appropriate for the data?
  - We can create a linear model for the data and estimate the coefficients of the model. We can analyze if the model assumptions are met and whether a linear model is appropriate.

Description of the dataset:
The dataset includes a list of top GitHub projects by the number of starts. It includes a brief description of the repository, URL, created date, updated date, homepage URL, size, starts, forks, issues, watchers, language, license, topics, has issues (boolean), has projects (boolean), has downloads (boolean), has wiki (boolean), has pages (boolean), has discussions (boolean), is fork (boolean), is archived (boolean), is template (boolean), and default branch for every repository in the dataset.

The dataset can be accessed here.

Signatures:

Loay Al Abri #8483229
Sachit Sharma #15141393

Roy Oh #57552671
Anne Zhou #34844902