INFOSYS 750 Research Methods -Quantitative

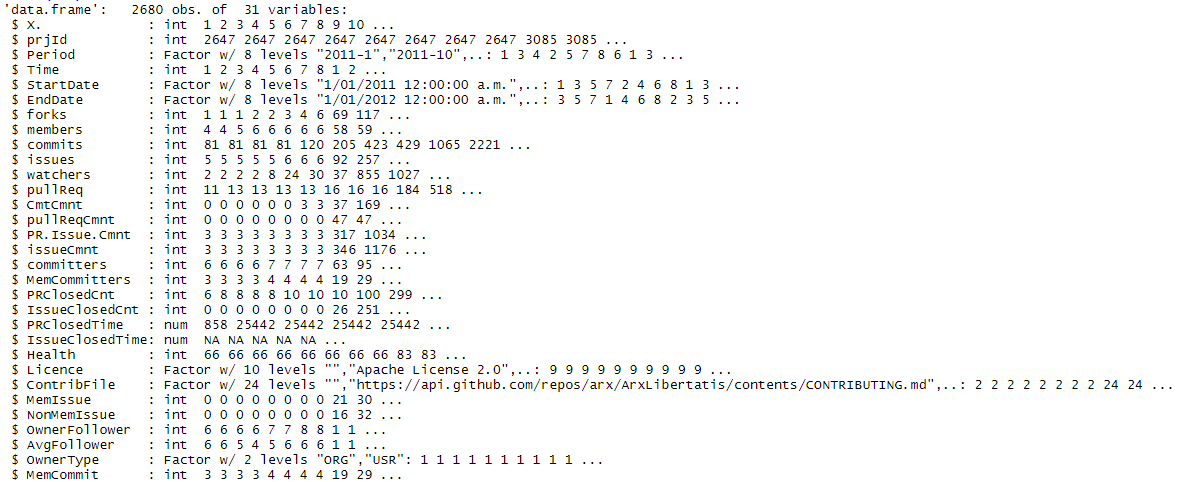
**Summer Ren (Jren791)**

1. Introduction
2. Literature Review and Research Question
3. Method
   1. Data Collection

The data used in this study was a panel dataset of project hosted on GitHub (https://github.com) where contains two years (8 quarters) of data for each project. GitHub is the largest open source repository with millions of projects. This allows up to analysis of how projects change over time. Furthermore, by using data from GitHub, we controlled other effects because all the projects are all saved in the same repository. In other word, the impact that from the different open source repository could have is removed.

Github supports pull requests with code review and comments features, which allow project owner or even users to maintain and update the open source project. It also has issue tracking (including feature requests) which can manage bus or some new requirements from users. Watchers means how many users interested in this open source project. On Github some projects just alive short years and then disappear but some can get chance to invest fund and become to a big company. Thus, more watcher means more chance. We get a data set of about 2680 objects consisting of watchers from 0 to 5153, start date from 01/01/2011 to 01/10/2012. Summary structure of the data are reported in Table 1.

**Table 1 Overall Summary Structure of the Data Set**



**Table 2 Dataset details**

|  |  |
| --- | --- |
| Variable | Definition |
| PrjID | A unique id number for each project |
| Period | Represents the current record contains data for which period of year |
| Time | A sequence for time of observations |
| StartDate | Beginning of observation |
| EndDate | End of observation |
| Forks | Number of times a project is forked |
| Members | Number of members |
| Commits | Number of coding activities |
| Issues | Number of problem/bugs raised or requests for new features |
| Watches | Number of people interested in project |
| PullReq | Number of code changes request for review. |
| CommitCmnt | Number of discussions on commits |

* 1. Definition of main variables, Visual Exploration

After overview the dataset, we can definition the main variables of this dataset. Base on Mr. Thompson’s requirements we picked following variable to analysis how projects change over time:

PrjID: A unique id number for each project;

Time: A sequence for time of observations;

Forks: Number of times a project is forked;

Members: It might relate with paying members or developers who want to maintain and support this OOS project;

Commits: This can represent the activity of the project;

Issues: Except developers to commit code, users also will find issue when they use the project. So, it also represents the activity of the project;

Watchers: It represents how many people interested in this project;

PullReq: Activity of the developer team representative;

CommitCmnt: Activity of the project representative.

Visual Exploration: (Visualize the growth plots for first ten projects)

This is the empirical growth plots for first 10 projects in watchers changing use study.

The graph shows that Watchers and Time appears linear between 0 to 9. And watchers tend to increase gradually from time 0 to 9. Identifying a suitable linear regression for level 1 model.



Between time 0-9, for project 3085, 3671, 5378,5495 and 5791, the watchers notable increase (10%-50%) then rest of project but for 2647 and 5624, the watchers stay constant over the time.

* 1. Data cleaning and preparation (any modification, transformation and sampling techniques that you have applied)

1. Analysis and Results