Memo

|  |  |
| --- | --- |
| To: | Dr. Landowski |
| From: | Yu Sheng Lu |
| Date: | Aug 12, 2022 |
| Re: | Project Proposal |

**Topic: Analyze public workstation usage**

**Data Description:**

The main dataset contains the reviews for the business on Yelp.

|  |  |  |
| --- | --- | --- |
| Data | Description | Example |
| date | The date for rating and review | 2017-04-07 |
| stars | The rating of the business | 5 |
| text | The reviews in text | You can't really find anything wrong with this place, ......... |
| business\_id | The ID for the business on Yelp | YvrylyuWgbP90RgMqZQVnQ |

I am going to use the Yelp business data as a supplementary dataset for the review dataset. This will help me to identify the business name and the review count for the rating.

**Research Questions:**

* What are the most used words for a good rating?
* How often the positive word being use have a high rating?
* What are the most used words for a bad rating?
* How often the negative word being use have a low rating?
* What location have the best review?

**Data Preparation Plan**

1. Load the data into Python
2. Transfer the json file to csv file
3. Concatenate the business and review data from each file into one data set
4. Filter the data set to only records in Groups we are interested in analyzing
5. Check for missing data and filter out the data that we don’t need