# 

# COMP 5070 Statistical Programming for Data Science

## Assignment 2

* This assignment is worth 25% of your overall grade.
* You have to submit R-code file and a word (or PDF) document with a nice-looking report. No archived files are allowed, that is, no zip, rar, tar, 7z, etc...
* You do NOT need to include the data files provided to you, as it can be assumed I have them too and all data files are stored in the working directory on my computer.
* The assignment is out of 100 marks. Marks for each individual questions are listed on the following pages. Below are possible penalties on the top of the total mark:

1. Upto 10 marks for none-optimised code. It is not always possible to avoid loops, however you should aim to use vectorisation where possible.
2. Upto 10 marks for not clear coding style. Code clarity is an important part of your submission. Thus, you should choose meaningful variable names and adopt the use of comments. You don't need to comment every single line, as this will affect readability. However, you should aim to comment at least each section of code.
3. Upto 10 marks if the code does not run successfully.
4. Upto 10 marks for poor presentation of the written report.

• Assignments submitted late, without an extension being granted, will attract a penalty of 10 marks per each day or any part thereof beyond the due date and time.

• *Plagiarism is a specific form of academic misconduct.* Although the University encourages discussing work with others and the Social Forum will support this, ultimately this assignment is to represent your individual work. If plagiarism is found, all parties will be penalised. You should retain copies of all assignment computer files used during development of the solution to Assignment 2. These files must remain unchanged after submission, for the purpose of checking if required.

# Yelp Reviews

***Summarising Online Reviews via Sentiment Mining***

|  |  |
| --- | --- |
| **source: yelp.com** | ***Rate Us On Yelp!***  Yelp is an online business that accepts and publishes reviews by anyone and everyone about local businesses, ranging from dentists to hairstylists, hotels and shopping.  A particularly popular business type to review on yelp are restaurants, cafes and other eating establishments. In this question, we are going to analyse a number of yelp reviews for restaurants, cafes, etc., and provide our statistical insights. |

## Data and Background Information

Yelpers have written hundreds of million reviews to date. Due to Yelp’s astounding popularity as a go-to site for recommendations (or warnings!) Yelp has become a very important site, particularly for small businesses who can achieve success or close down, based on their online reviews.

The data for this question is yelp\_reviews.csv provided on the assignment page. The variables are:

* user\_id: anonymised Yelp reviewer id.
* business\_id: anonymised Yelp business id.
* date: date of review (dd/mm/yyyy format).
* stars: star rating. 1 star is the lowest rating, 5 stars is the best rating.
* review\_length: the review length in characters.
* votes\_funny: the number of votes received indicating the review was funny.
* votes\_useful: the number of votes received indicating the review was useful.
* votes\_cool: the number of votes received indicating the review was cool.
* votes\_total: the total number of votes.
* pos\_words: the total number of ‘positive’ words used in the review.
* neg\_words: the total number of ‘negative’ words used in the review.
* net\_sentiment: the overall sentiment of the review – the difference between pos\_words and neg\_words. A positive number implies a positive review (more positive words than negative) and a negative number indicates the opposite.

For this assignment, you need to produce a report summarising a collection of requested statistical analyses and visualisations of the data. See the next page for details.

You will need to submit a proper written report and R-script file. As a guideline, excluding tables/figures, 2-3 pages of writing will be sufficient for the report. I won’t strictly count words so if you go over/under by a bit that’s fine, but this is a good ballpark to aim for.

The report should contain:

1. An introduction outlining the analysis to follow/background information. The introduction can be up to 2 paragraphs. For the purposes of this assignment a paragraph is 6-8 sentences. (5 marks)
2. A statistical summary of *stars, review\_length, pos\_words, neg\_words* and *net\_sentiment*. Discuss general impressions from this statistical summary. (15 marks)
3. Create tables of the counts of numbers of positive words per review and the counts of numbers of negative words per review. From these tables, produce a plot to display the first 20 entries. You may either produce one plot per table, or one plot overall – this is entirely your choice. Discuss the trends you see in this data. (12 marks)
4. Now repeat step 3 for the data in *net\_sentiment* and share any insights you might have about the output. (12 marks)
5. Present and discuss the average review length per star category. Produce an appropriate visualisation of your choice of the average. Explain your choice of average and discuss the general behaviour of the data you have presented. (12 marks)
6. Analyse reviews voted as useful (variable *votes\_useful*): are there any relationship between usefulness of reviews and (separately) star-rating provided and length of the review. (12 marks)
7. Study the number of reviews per day and how does it change over time. (12 marks)
8. Select the best business and the best user in the dataset. You must create and explain your own criteria of being “the best”. (12 marks)
9. Conclusions. (8 marks)

The general rule is that every question in your report should have graph/s, numbers, and discussion. There might be situations when it is impossible or impractical to have all three elements of the reporting. For example, conclusions don’t have data visualisations; some questions might not have meaningful numerical summary. However, you should always aim to present all three elements.

Don’t include programming code in your report. Don’t include discussions like “I used function xyz() to calculate total value” or “Using package ggplot and created a graph”. These are examples of a poor presentation. Your readers are managers that need results of your analysis to get better understanding of the business and to support their decision-making process. They are not interested in the programming code or functions.