ST309 Project Instruction

1 The Project

(Maximum 10 A4 sides, excluding a cover page, figures, tables, a reference list and computing codes)

This project should be completed by a group with no more than 3 people, formed by your own choice, working on a non-trivial data project also chosen by you yourselves. The purpose of the project is two-fold:

- ♦ to demonstrate competency with data analysis techniques, and
- ♦ to show that you are capable of solving a practical problem from data analytic viewpoint.

Your final project report should include:

- Introduction of the problem tackled: what the problem is, why it is important for the subject matter, why the problem can be casted as a data analytic problem, what the goals of your study are. Don't worry too much if your analysis would not meet all your goals as you may be handicapped by the limitation of available data, time and etc. Good ideas are always valuable. You should also review the *status quo*: how the problem is handled in its subject area, and the previous attempts using data analysis.
- Description of data used in your analysis: source of data, appropriateness for your goal, any data cleansing conducted including dealing with missing values, outliers, transformation etc. Ideally data should be from your own subject areas, or something of interest to you or your subject, as you will have the relevant subject knowledge.
- Report of your data analysis, including a comprehensible description of the methods used, the reasons your approach, evaluation and interpretation of the results of your data analysis.
- Conclusion: What has been learned from your data analysis on the subject matter concerned? How helpful are the results from your data analysis in solving your practical problem? Are there any possible improvements on subject understanding, data collection, data analysis and etc?

A cautionary warning: make sure the data set which you choose for your project contain the relevant informaton, or at least partially.

Your report should be concise and informative. A dim view will be taken of barely adapted versions of existing descriptions (such as those in *Wikipedia*) or equations lifted from the literature without proper understanding.

2 Initial Project Outline

You should submit a one-page initial project outline via Moodle by **3 December 2020**, including the names of the group members, the tentative title of your project, the problems to be tackled, the data set to be analysed, and the statistical learning procedure(s) to be applied. Only one submission per group is required. Feedback will be given within one week after your submission.

3 Submission and Assessment

You may form a group of maximum 3 people to work together for this project. You should allocate tasks among the group members in an equitable fashion. You should submit a PDF version of your report (together with one or more cleaned-up R-script files) to Moodle by **4pm on Thursday 18th February 2021**. Again one submission per group only is required.

3.1 Project report

You should submit a word processed (by, for example, microsoft word or latex) printed report. The recommended length of the report is not more than 10 pages of A4 size, excluding a cover page, figures, tables, bibliography and computing codes. You should use an 11 point standard font (for example, times new roman) and 1.5 spacing. Your submission should contain the following sections:

1. A cover page
2. Main body, including, for example, introduction, data description, main results, conclusion (see Section 1 above).
3. Bibliography

The cover page should be marked clearly with "ST309 Group Project Report", should contain the title of project, the LSE examination number for each member of the group together with the individual contribution in percentages to the project; see Section 3.4 below. Please do *not* write your real names.

It is not advisable to include R codes or R outputs in your report, unless such an inclusion is essential for illustration of some key points. However you should submit to Moodle some cleaned-up R-markdown (`.Rmd`) files, or R-script (`.Rhistory`) files for the sessions that generated your results. The files should be in plain text. Remember that R code should be self-documenting – good use of comments is required.

You should make sure that the examiners can access the data used in your analysis, by either stating clearly the source of data, or submitting the data to Moodle. Split a large data file into several smaller files as Moodle only allows files not greater 100 Megabytes.

The presentation of results is very important. You should generate appropriate graphs or tables that summarize **succinctly** the results of your analysis. Your report and results should be intelligible to any reasonably well informed readers (such as those who have done ST107 or

ST108). Avoid using jargon as much as possible. You should show a critical awareness of any weaknesses in the approach you adopt and discuss possible extensions and improvements.

3.2 Plagiarism, references and bibliography

Plagiarism is taking someone else's work or ideas and passing them off as your own (adapted from Concise Oxford Dictionary definition). This arises in course work as sections of text lifted from books or internet sources and submitted as the student's own work. This is a very serious offense that is quite easy to detect. Plagiarism will result in instant failure (mark 0). You are strongly encouraged to read widely and assimilate ideas from as many sources as possible. However, when you use other people's work you must give a proper reference.

For **references** use the 'Harvard' or bracket system within the text. For example:

- "Lapointe and Legendre (1994) put forward . . ."
- "Several authors (Lapointe and Legendre, 1994; Silge and Robinson, 2018; Venables et al., 2018) make use of a . . ."
- " Silge and Robinson (2018, p.29) says, 'Sentiment analysis provides a way to understand the attitudes and opinions expressed in texts', but . . ."

In the **bibliography**, list works cited in alphabetical order by author. Give the author's surname, initials, year of publication, and title of work. For books, give the place of publication and publisher; for journal articles give the name of the journal, volume and pages; for articles in published collections, give the name of the editor, title of book, place of publication and publisher; for online sources give the source organisation, the URL and the date accessed. The appropriate use of italics and bold is important. For example (in the order from collection, journal article, book, online source):

1. Lapointe, F.-J. and Legendre, P. (1994). A classification of pure malt scotch whiskies. *Journal of the Royal Statistical Society, Series C.* **43**, 237-257.
2. Silge, J. and Robinson, D. (2018). *Text Mining with R.* O'Reilly, Beijing.
3. Venables, W.N., Smith, D.M. and the R Core Team (2018). *An Introduction to R.* Available at `http://cran.r-project.org/doc/manuals/R-intro.pdf`

Note. Online sources are in abundance with diverse qualities. Be careful to use only reputable sources.

3.3 Mark scheme

Each project will be marked out of 100. The breakdown for different parts/aspects is as follows.

Description of problem tackled	10
Data description and data preparation	20
Data analysis	25
Conclusion	10
Readability and efficiency of R-codes	15
Presentation and use of language	20

3.4 Grades according to contribution

The grade for each group member will be a function of her/his contribution defined as follows:

$$\text{member grade} = \text{project grade} \times \frac{\text{member contribution}}{\text{maximum contribution}}$$

For example, for a group with 3 members contributing, respectively, 50%, 30% and 20%, and the project grade is 72 (out of 100), the individuals grades are:

$$72 \times \frac{50}{50} = 72, \quad 72 \times \frac{30}{50} = 43.2, \quad 72 \times \frac{20}{50} = 28.8.$$

Appendix I. How to identify a project topic

There is no uniquely correct way to identify a topic for your project. Nevertheless it might be helpful to following the steps listed below.

1. Identify a problem with objective(s) which you are interested, preferably related to your expertise.
2. Looking for data set(s) accordingly.
3. Identify Y (if any) and X_1, X_2, ..., and learning methods to be used.
4. Conducting some initial data analysis; you may start with a relative small data (sub)set at this stage.
5. You should then revisit 1, 2, 3 above again, to see if you should modify the problem, if data contains the relevant information to achieve your objectives or there is a need of a new data set, and if the chosen methods should be changed or more methods can be applied etc.

After **a few iterations** you should be happy with the finanalised problem together with relevant data set(s) and the methods to be used.

Appendix II. Useful data sites

To find adequate data set(s) is a key for this project. The data should not be too simple/small, ideally should be linked with your own subject(s).

- Google Dataset Search
  `https://toolbox.google.com/datasetsearch`
- UCI Machine Learning Repository
  `https://archive.ics.uci.edu/ml/index.php`
- Kaggle Datasets
  `https://www.kaggle.com/datasets`

- RDataMining Free Datasets
  `http://www.rdatamining.com/resources/data`

- Thomson Reuters Text Research Collection (of News stories)
  `https://trec.nist.gov/data/reuters/reuters.html`

  Note. Require an online application.

- Free eBooks - Project Gutenberg
  `https://www.gutenberg.org/`

- Yahoo!Finance
  `https://uk.finance.yahoo.com/`

- London Data Store
  `https://data.london.gov.uk/dataset/`

- US Bureau of Labor Statistics
  `https://www.bls.gov/`

- US Vote Data
  `http://k7moa.com/`

- Stock and Watson Econometrics data
  `http://fmwww.bc.edu/ec-p/data/stockwatson/datasets.list.html`

Appendix III. Sample reports

List below contains some good or not-so-good reports/papers on data analytic projects. Most of them are the results of substantial efforts, requiring more time and space allowed for this project. Nevertheless you may pick some useful ideas how to conduct a good data project.

- Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science.
  `https://www.liebertpub.com/doi/abs/10.1089/big.2018.0092`

  This was done as a nice example of data science projects.

- A Classification of Pure Malt Scotch Whiskies
  `http://www.dcs.ed.ac.uk/home/jhb/whisky/lapointe/text.html`

  Again an excellent and substantial data project. We will cover a part of the story in the course.

- There are 3 case study reports on mining text data in the last 3 chapters of book entitled "*Text Mining With R*" by J. Silge and D. Robinson, 2018, O'Reilly.
  `https://www.tidytextmining.com/`

  As the book focuses on data mining techniques, it pays a little attention on elaborating explicitly the goal of analysis and the interpretation of the results.

- Customer Puzzled Behavioral Analysis: a step towards valuing customer's interests.
  `http://www.iaeme.com/MasterAdmin/UploadFolder/IJMET_09_07_041/IJMET_09_07_041.pdf`

  Used market-basket analysis techniques which will be covered late in the course.

- Mining News Stories to Predict Stock Price Movement.
  Provost, F. and Fawcett, T. (2013). *Data Science for Business*. O'Reilly: Pages 268-277.
  Available online from LSE Library.

  Not enough details on how the data analysis results were produced.