Guidelines For Data Science Case Study Project

Introduction to Data Science

Following is an overview of the guidelines for the data science case study project. The project must result in the production of at least three items:

1. Fully documented code
2. A project notebook (can be either an RStudio notebook or a jupyter notebook)
3. A written report following the prescribed format

The goals for the project assignment are

1. for students to gain confidence and foundational skills in implementing and applying the typical data science workflow: import, explore, format, transform, analyze, model, and draw conclusions from data.
2. for students to gain experience in effectively presenting and communicating results or findings based on an implementation or application of the typical data science workflow.
3. for students to make a connection between data science methodology and substantive expertise in a specific domain.

Important∗ Your work on this project must be original. Simply repeating information from another source or reproducing the work of someone else will result in a zero grade for this project since this constitutes plagiarism. Furthermore, you **must** cite sources (at least three reliable scholarly sources) and provide references. See further details below regarding appropriate sources.

The written report should be prepared using LaTeX(this can be done within RStudio) and should follow this format:

1. A section that provides a clear description of the problem and the objectives for the data analysis. Additionally, sufficient background on the specific domain of expertise for the project should be provided.
2. A description of the data. This should include but is not limited to a statement of where the data set(s) was obtained, the format of the original data set(s), what is included in the data set(s), number of and names of variables, number of observations, etc.
3. Description of the process of loading,formatting, and transforming the data.
4. All plots must be of high quality and should be included in both the written report and the project notebook. All plots should have appropriate and clear titles, labels, captions, etc. The appearance and effectiveness of plots is extremely important.
5. A description of the process and results from an exploratory data analysis.

6. A section on modeling the data to make predictions, learn something, etc. This should make a clear connection with description of the problem and the objectives for the data analysis set forth in the first section.

7. An evaluation of the model(s) used and an explanation of all results.

8. Conclusion/summary of the data science process and the obtained results.

You are allowed some flexibility in the order in which the required parts of the paper appear in your report. However, your paper should follow some sort of clear logical progression. See the case study book chapters posted on D2L for a examples.

The project notebook should follow roughly the same structure as the written report but requires less detail in written descriptions and more detail in terms of code and the use of code. You will be provided with a template notebook.

The various parts of the assignment together with their grade proportions and tentative due dates are:

1. selection and approval of a data set(s) - 5% (due $\approx$ September 14)

2. written data description - 5% (due $\approx$ September 21)

3. background, problem description, and objectives - 10% (due $\approx$ October 5)

4. programmatically loading, formatting, and transforming of the data set(s) - 12% (due $\approx$ October 12)

5. data visualization and summarization - 12% (due $\approx$ October 26)

6. modeling of the data - 12% (due $\approx$ November 16)

7. model evaluation - 12% (due $\approx$ November 28)

8. conclusions and report initial draft - 12% (due $\approx$ December 7)

9. report final version - 20% (due $\approx$ December 12)

Students will be allowed to revise earlier parts of the project assignment by the final draft submission in order to receive (some but not all) points back.

The way in which each of the separate components of your project grade will be determined is as follows:

1. (20%) The amount of clear effort that was put into the project

2. (20%) The amount of original thought that was required

3. (20%) The amount of independent learning

4. (20%) The overall appearance and presentation of the work, for example,

   (a) clarity of the paper, does the written paper read well?

   (b) is the grammar, punctuation, and spelling correct?

   (c) is the paper well organized?

(d) can the reader easily follow the flow of ideas?

5. (5%) Use of proper citations, you may use either APA or MLA format as long as you are consistent

6. (15%) The degree of sophistication demonstrated in use of data analysis techniques

As far as sources go, you should only use valid scientific publications such as textbooks, peer-reviewed journal articles, etc. I would definitely discourage the use of material on the internet that may not be reliable such as blogs or something posted on someone's personal website. Since this is meant to be a scientific endeavor you should be able to back up your claims in a reliable manner and using resources that are not professionally supported in some way is not the best way to do so.

This project should be a significant effort, you should spend **a lot** of time on it. It counts way more than an individual homework assignment, do not treat it as if it were a simple homework assignment. I am happy to provide assistance on any aspect of the project as long as you come to me well in advance of the due date.