

DATA 115: Final Project

This document provides details and guidelines for the final project, which accounts for 40% of your final grade. As this is the largest component of your grade, you should read the sections below carefully in order to be successful. The purpose of this project is to give you an opportunity synthesize all of the material that we have covered in the course and showcase your R and data analytics skills by performing a case study on real data.

1 Project Outline

You will select a dataset from one of the resources below and analyze it in R using the techniques we have discussed in class. Your main goal will be to provide an answer to a related ‘Big Question’ chosen from among those associated to your selected dataset. You have to understand the data you chose and frame your big question. Along the way, you will need to process the data, conduct exploratory data analysis, perform supervised and/or unsupervised learning techniques that you deem fit and justify the conclusions that you draw while constructing a satisfactory answer to your question.

2 Main Components

You will summarize and describe your work in a final report, submitted as a .pdf by 11:59Pm on 16th June, 2021. The report will be evaluated on the following criteria:

- **‘Big Question’:** Is the question interesting, clearly stated, and specific? Is the chosen dataset a reasonable option for addressing this question?
- **Visualizations:** Are the visualizations used to represent the analysis effective and complete?
- **Analysis:** Are the methods that we used to analyze the data appropriate and carried out correctly? Is the analysis thorough and logically conducted?
- **Conclusions:** Do the final conclusions provide a satisfactory answer to the ‘Big Question’? Are the conclusions supported by the analysis that was performed and presented?
- **Reproducibility:** Are the data processing and cleaning methods well-documented? Is the code used to analyze the data correct and easily interpretable?
- **Presentation:** Is the final report well organized and neat? Are the plots designed with care, including color choices, appropriate labelling, and other aspects of good visualization design?

2.1 Report

Your final report should thoroughly describe your experiences and analysis, as though you were reporting the results of this project to a manager or supervisor. The document should be submitted as .pdf by the beginning of the finals period for this class. The following information should be contained in the report, as well as appropriate figures from the analysis incorporated into the text of your markdown file:

- Describe the dataset and why you selected it for this project.
- Describe any processing problems you identified with the data and how you overcame those issues.

- Describe your ‘Big Question’ and why the data is a good choice to answer it.
- Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.
- Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered.
- Describe your final conclusions based on your analysis and support them with analytics on your dataset.
- Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.

3 Data Sources

You should select the underlying dataset from one of the following sources. Note that these resources provide additional information about the data beyond the specific values and you should incorporate a discussion of the relevant pieces of this information (how the data was gathered, what was the original purpose of the data, what cleaning steps were performed before the data was uploaded, etc.) into your report.

- Wine Quality
- Community and Crime
- Video game sales
- Baseball Stat
- COVID-19

4 Scheduling

We will progress through the project each week. A part of the project questions will be included in the homework. However, the final report will be due on 16th June, EOD.

- **Week 3:** Selecting and understanding the dataset and deciding your “Big Question” you want to answer through your analysis
- **Week 4:** Exploratory Data Analysis. Take a deep look at all columns and data properties. Address ‘Big Question’, determine conclusions, decide on follow up questions to ask and answer. What does your project objective demand-Regression or Classification?
- **Week 5 :** Analyze your data based on what all you have learnt and you deem fit to answer your big question
- **Week 6:** Compile final report, generate attractive figures.