FINAL PROJECT

DUE DEC 10

11:59 PM on MyCourses

GUIDELINES

The midterm project was an introduction to predictive analytics. It was a first dive into the basics of data science.

The final project, in a sense, is a replica of the midterm project. So, broadly speaking, the idea of the final project is the same as the midterm project: produce a statistical report where you use analytics. There are two key differences, however.

First, you can choose one of five datasets, and develop any idea that you want. At the end of this document, I have listed the five datasets. Accordingly, the final project will test your ability to come up with creative data-science ideas.

Second, the final project must rely on the techniques we learned during the second half of the course. This means that your project should either rely on:

(i) Classification techniques
(ii) Tree-based methods
(iii) Unsupervised (machine) learning techniques.

Of course, you may use statistical techniques from the first part of the semester. But I expect your statistical analysis to be heavily based on the second part of the course.

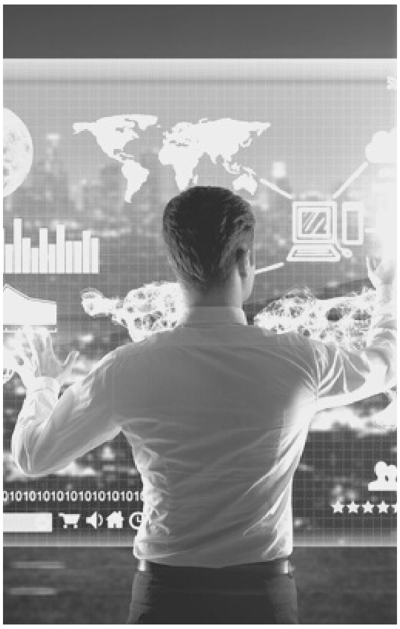
Note 1: This is a final examination. As such, we will not be offering advice via email or office hours. We will only be responding to clarification questions (e.g., an issue downloading a dataset).

Note 2: As an official final examination, the rules on time limits will be strict (as per McGill policy). we will not be granting extensions to anyone.

The same rules that apply to a three-hour exam apply to this take-home project.

PROJECT GOALS

Goal 1: Sharpening your data scientist skills

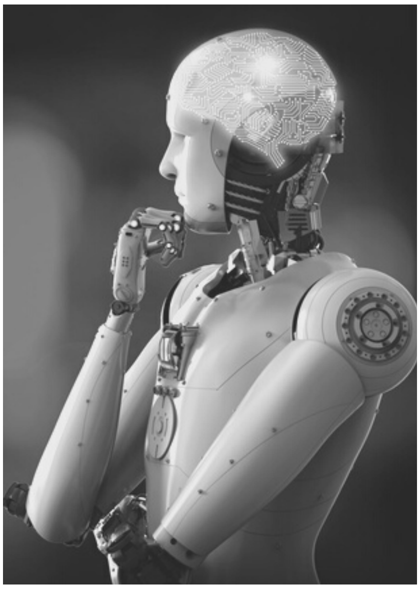


The midterm project provided you with the ability to apply data-science techniques to conduct predictive analysis. You also gained the skills to deploy a statistics project as a group. In the feedback sessions, I gave you specific pointers about the strengths and weaknesses of your midterm project, and also numerous tips on how to improve your final project report. This final project will give you a second chance to craft a refined version of your midterm project, and to improve on the weaknesses from your midterm project. I expect you to fully incorporate the feedback I gave you during your personalized feedback sessions, and to apply this feedback to improve your final report. I will have higher standards when it comes to the final project.

Goal 2: Applying new data techniques

After the midterm project, we learned many new techniques, including:

- Classification I: Logistic Regression
- Classification II: Linear Discriminant Analysis
- Tree-based methods: Regression Trees, Classification Trees, Bagging, and Random Forests, Boosting
- Unsupervised Learning I: Principal Component Analysis
- Unsupervised Learning II: Clustering Methods



I expect you to demonstrate that you have mastered some of the topics in the second half of the course.

Goal 3: Test Your Creativity

In the midterm project, you were given a task: predict the IMDB ratings of twelve movies. In this project, it is your responsibility to come up with a topic. This will allow you to explore different directions, and to work on something you feel passionate about. At the same time, it will test your creativity, i.e, your ability to find meaning and explore interesting ideas related to data science.

Dataset list

You may choose one of the following five datasets, and work on any question/problem you wish to develop, as long as you are following the guidelines above. The datasets are provided in myCourses. Data dictionaries for each dataset are provided in the links below.

Note: You may not work on any other dataset

Dataset 1: Gun violence dataset

Comprehensive record of over 260K US gun violence incidents from 2013-2018.
(https://www.kaggle.com/jameslko/gun-violence-data)

Final Project

Dataset 2: Olympic events data

Basic bio data on athletes and medal results from Athens 1896 to Rio 2016
(https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)

Dataset 3: Shark tank pitches

Dataset containing the various pitches for funding on popular TV Show Shark Tank
(https://www.kaggle.com/rahulsathyajit/shark-tank-pitches)

Dataset 4: Chocolate bar ratings

Expert ratings of over 1,700 chocolate bars
(https://www.kaggle.com/rtatman/chocolate-bar-ratings)

Dataset 5: Automobile dataset

Dataset comprising the characteristics and selling price of most auto models
(https://www.kaggle.com/toramky/automobile-dataset/data)

DELIVERABLE

The format of the final project is, in essence, very similar to the midterm project. But, of course, the standards for the final project are much higher (now that you know how to write a report!) The report should be typed, clear, and aesthetically pleasing. The report should be between **4 and 7 pages**; I'm not too demanding on page limits, so this is a rough guideline. You shouldn't feel constrained to write more or less if you feel the need to say something (or you feel you've said enough). Exhibits (e.g., extra tables, figures, etc.): up to 10 pages may be appended. Please use the following framework to organize your paper (again, just a rough suggestion—feel free to deviate).

1

Introduction (0.5 pages)

Here, you provide a summary of the project, the goals, etc.

2

Data Description (1 - 1.5 pages)

Here, you describe the distributions of the different variables, and the relationship between these variables.

3

Model Selection & Methodology (1 - 1.5 pages)

Here, you will tell us which methodology you used to build your model.

4

Results (1 - 2 pages)

Present the results of your final model.

5

Classification/predictions and conclusions (1 - 2 pages)

Summarize your findings and present managerial conclusions.

6 Appendices (max 10 pages)

All tables and exhibits should be after the conclusions. All tables should be labeled and named.

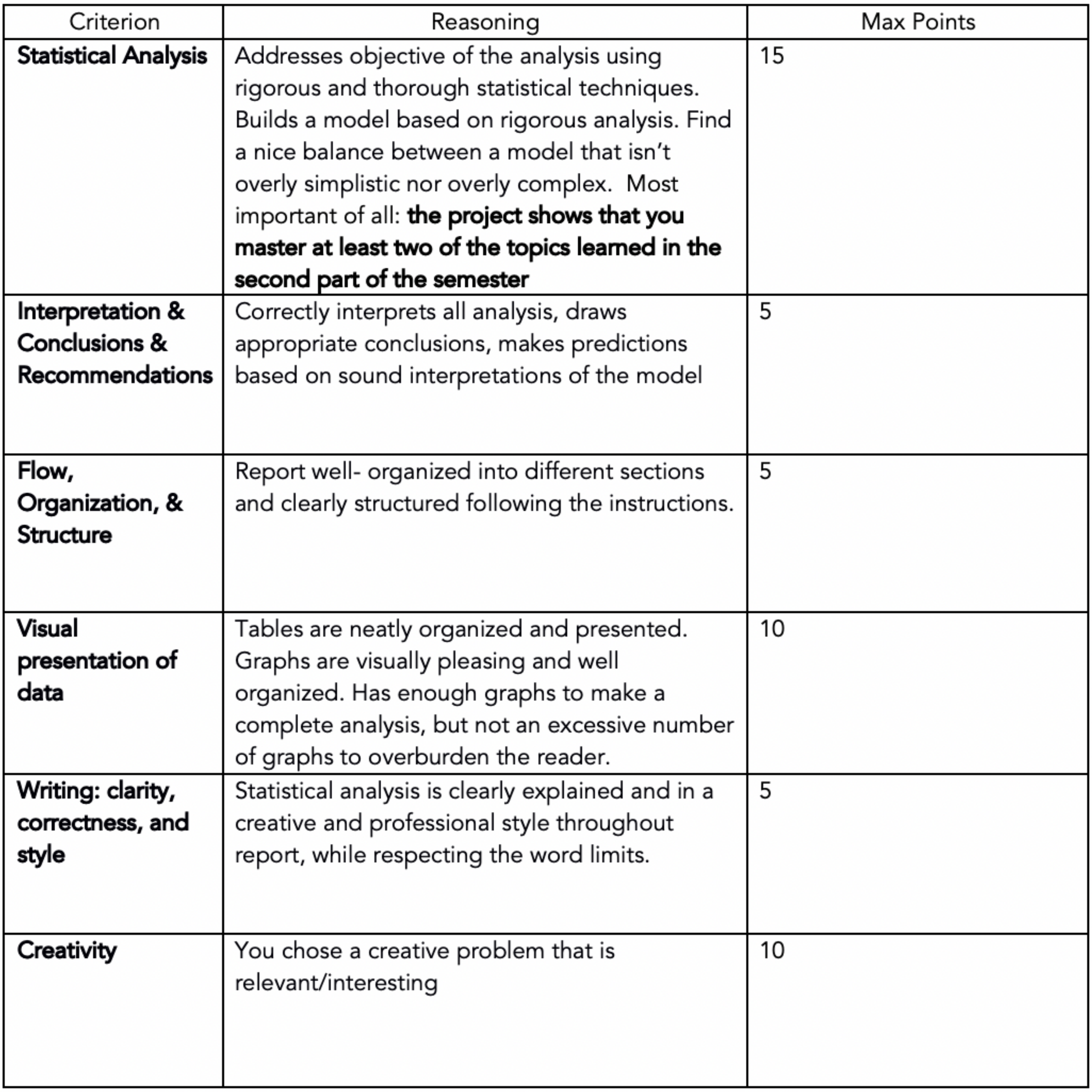
7 Code

Please attach your R code at the end.

Note: I recommend you use LaTeX, since it ultimately makes your life easier and creates beautiful formatting for sections. But again, no points will be taken off, nor added, if you decide to use Microsoft word, or another word processor.

GRADING

Your grade will be out of 50, and you will be graded on the following criteria:



Criterion	Reasoning	Max Points
Statistical Analysis	Addresses objective of the analysis using rigorous and thorough statistical techniques. Builds a model based on rigorous analysis. Find a nice balance between a model that isn't overly simplistic nor overly complex. Most important of all: **the project shows that you master at least two of the topics learned in the second part of the semester**	15
Interpretation & Conclusions & Recommendations	Correctly interprets all analysis, draws appropriate conclusions, makes predictions based on sound interpretations of the model	5
Flow, Organization, & Structure	Report well- organized into different sections and clearly structured following the instructions.	5
Visual presentation of data	Tables are neatly organized and presented. Graphs are visually pleasing and well organized. Has enough graphs to make a complete analysis, but not an excessive number of graphs to overburden the reader.	10
Writing: clarity, correctness, and style	Statistical analysis is clearly explained and in a creative and professional style throughout report, while respecting the word limits.	5
Creativity	You chose a creative problem that is relevant/interesting	10