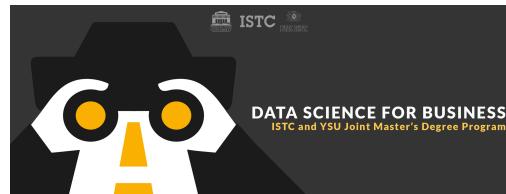


Yerevan State University
Faculty of Economics and Management
1002/M91, Data Mining, Spring 2020

Final Project Guidelines

Lusine Zilfimian

April 29, 2020



General Information

As a part of the assigned work for our course, students are to engage in a final project. The project will be carried out in groups of 2 (or 3 if the project is sophisticated enough, subject to instructor's approval). The final project must be *original* work and you should use the real-world data (try to use nontrivial data from Kaggle, [CRRC](#), [WDI](#), [SCRA](#), scraped data, or another unique data). Ideally, you should try to do something a bit interesting, like studying a data set that has not been thoroughly evaluated, or using different approaches. This is an application-based project but you should not worry about "the best" result. The goal of the final project to apply topics we learn during classes on the data and discover something interesting. You are free in choosing both the area of research (data, topic) and the algorithms to use.

Credits: 30% of the total grade.

Deadlines

- May 9th Group Creation
- May 15th Final Project Proposal due
- May 28th Final Project Report due
- June 1st Presentations

Proposal

You must submit a proposal for the instructor's approval. Your project one-page proposal should tell the instructor about the problem you aimed to solve, ideas about how you might begin: general approach and methods (decision trees, regressions etc.) you intend to perform, data you are going to use and approximate effort required. It must consist of a preliminary title, a high-level summary of your project and outline your main goals, a brief

description of what you plan to do. The purpose of the proposal is to make sure that you are on the right track and obtain the useful feedback from the instructor.

Final Report

The required element of the final report:

- Title
- Introduction: (the main scientific concept of the problem, summarizes the paper, and the goal of the project)
- Model: (the description of the data sets, the evaluation metrics, the data mining tools used, and any other details related to the main work)
- Results (the experiment result, the performance metrics should be given and be compared with theoretic results, etc.)
- Related Work: A brief description of related work, with citations to relevant papers.
- Conclusion: (the quality of your results, material on future work, etc.)

Final Project Presentation

The last lecture of classes will be dedicated to project presentations, each group will give short presentations on what they have accomplished (5-7 minutes each). The presentation should be done using **R Shiny**. You will receive a grade based on your final project and presentation. *Extensions* will be granted only in cases of extreme necessity. Late days will not apply to the final reports.

Software

You should use R programming language to do your final project, for the presentation you should use R Shiny.

Grading Policy

The final project grade will be based on the following components:

Criteria	Description	%
Curiosity	<ul style="list-style-type: none">Ability to creatively apply old tools in new waysThe data used is authenticThe goal of the research	30%
Skepticism	<ul style="list-style-type: none">Appropriate visualizationQuestions and answers	40%
Organization	<ul style="list-style-type: none">Clarity and quality of R codePresentation quality	30%

Software

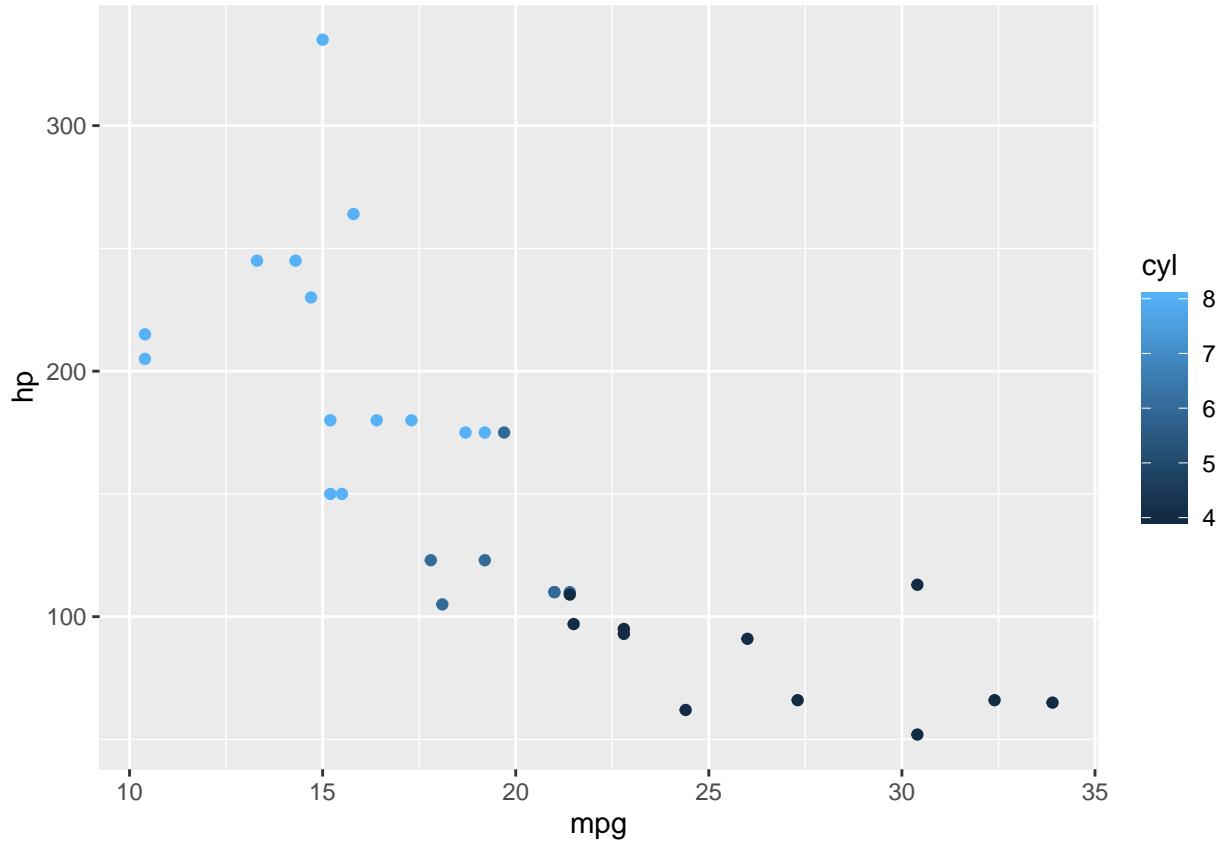
You should use **R programming language** to do your final project and for the presentation you should use **R Shiny**.

Note: Trying to avoid similar projects/themes, compare your topics with topics of people who have already decided on a project topic.

My favorite Rules of Thumb ☺

Example of a bad graph

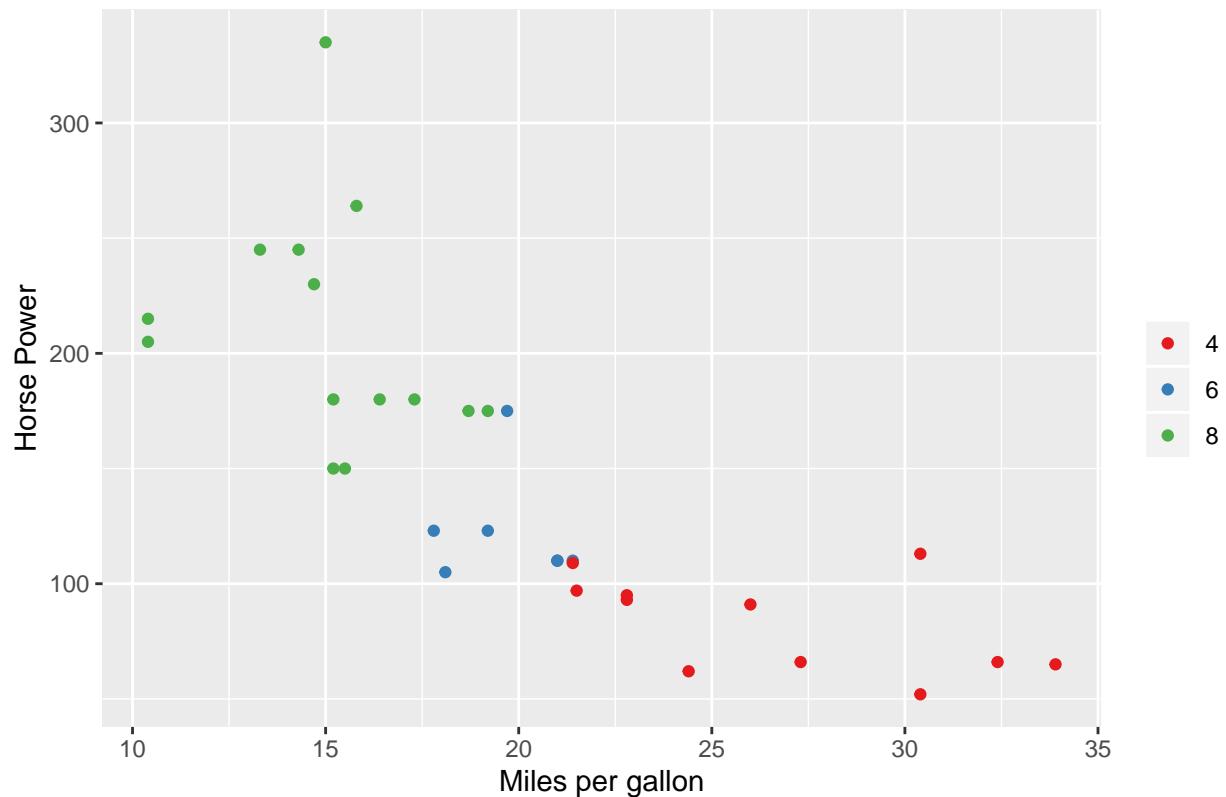
```
ggplot(mtcars, aes(x=mpg, y=hp, col = cyl)) + geom_point()
```



Example of a good graph

```
ggplot(mtcars, aes(x = mpg, y = hp, col = factor(cyl))) +
  geom_point() +
  labs(x = "Miles per gallon", y = "Horse Power",
       title = "Scatterplot for horsepower and miles per gallon", col = "") +
  scale_color_brewer(palette = "Set1") # this is not mandatory :)
```

Scatterplot for horsepower and miles per gallon



Example of a bad code

```
mt1 <- mtcars[mtcars$am==1,]  
mean(mt1$hp)
```

```
## [1] 126.8462
```

```
mt1 <- mtcars[mtcars$am==0,]  
mean(mt1$hp)
```

```
## [1] 160.2632
```

Example of a good code

```
library(dplyr)  
mtcars %>%  
  group_by(am) %>%  
  summarise(mean=mean(hp))
```

```
## # A tibble: 2 x 2  
##       am   mean  
##     <dbl> <dbl>  
## 1     0    160.  
## 2     1    127.
```

Useful packages

- ggVis
- plotly
- lubridate
- tidyr
- wordcloud
- tidytext
- sunburstR
- network3d
- scatter3d
- sentimentr
- formatR

Good Luck!