The aim of the project is to ***use multiple regression to analyse a data set with multiple predictor variables.***

There are two components to this group project:

1. Presentation delivered in your computer lab in week 12.
2. Executive summary submitted through Canvas by 3pm Monday 29 October.

You will work with the data set assigned to you (see below) or you may use another data set of your choosing, **after having it approved by Garth**.

**Executive summary**

The executive summary (report) is restricted to two pages. You may optionally include a third appendix page for references, any large tables and/or additional figures that may not fit in the two pages.

In the two pages you should include:

1. **Abstract:** A one paragraph summary of what you set out to learn, and what you ended up finding. It should summarise the entire report.
2. **Introduction:** A discussion of what questions you are trying to answer.
3. **Data set:** Describe details about how the data set was collected (if known) and the variables in the data set.
4. **Analysis:** Describe how you used multiple regression to analyse the data set. Specifically, you should discuss how you carried out the steps in analysis discussed in class, i.e., exploration of data to ***find an initial reasonable model, checking the model and changes to the model based on your checking of the model.***
5. **Results:** Provide inferences about the questions of interest.
6. **Discussion and conclusion:** Describe any limitations of your analysis and how they might be overcome in future research and provide brief conclusions about the results of your study.

Your report does **not need** to be a R Markdown document. If you would like to make your report a fully reproducible R Markdown report, I recommend using one of the templates from the **[rticles](https://github.com/rstudio/rticles)** package (e.g. the PNAS template is a nice two column template that can fit lots of text, figures and tables into 2 pages). These require a TeX distribution installed. If you don’t already have an TeX distribution installed, you can get a basic working version with the **[tinytex](https://yihui.name/tinytex/)** package.

Your report **does not** need to show the code you used to generate the output and figures. However this should be available in a .R or .Rmd file in a group **git repository** and the link to the repository provided in the report.

**Presentation**

Your presentation should be **at most** 10 minutes - I suggest aiming for 7 minutes (roughly 7 content slides). It can be in **any format** that you’d like. R Markdown can generate various presentation formats, see [here](https://rmarkdown.rstudio.com/formats.html), or there are others such as [xaringan](https://github.com/yihui/xaringan) but this is more experimental.

All group members should have the opportunity to speak. The presentation should be an engaging overview of the 6 points covered in your report (outlined above). Take your audience on a journey with you - how did you approach the problem and what did you find?

|  |  |
| --- | --- |
| **Wine Quality Data Set**   Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical (relating to physics and chemistry of the wine) tests. | https://archive.ics.uci.edu/ml/assets/MLimages/Large186.jpg |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 4898 | **Area:** | Business |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 12 | **Date Donated** | 2009-10-07 |
| **Associated Tasks:** | Classification, Regression | **Missing Values?** | N/A | **Number of Web Hits:** | 741930 |

**Source:**

Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez>   
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal   
@2009

**Data Set Information:**

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [[Web Link]](http://www.vinhoverde.pt/en/) or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).   
  
These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

**Attribute Information:**

For more information, read [Cortez et al., 2009].   
Input variables (based on physicochemical tests):   
1 - fixed acidity   
2 - volatile acidity   
3 - citric acid   
4 - residual sugar   
5 - chlorides   
6 - free sulfur dioxide   
7 - total sulfur dioxide   
8 - density   
9 - pH   
10 - sulphates   
11 - alcohol   
Output variable (based on sensory data):   
12 - quality (score between 0 and 10)