STAT 435: Project Requirements/Rubric

The course STAT 435 requires that all students participate in project. This project can be done individually or in groups of *at most three (3)* students.

Purpose: The project is meant for students to develop an independent thinking towards all aspects of statistical data analysis. This includes the preliminary stages: 1. Gathering/cleaning data, 2. Framing an objective statistically, i.e., to frame a subjectively interpretable problem as a objective statistical model and related statistical question(s). Then the implementation stage: 3. Choosing the required method that addresses the questions of interest and finally, 4. Concluding your analysis by describing the interpretable conclusions from the analysis performed.

Guidelines:

1. All students are free to choose a problem of their own. The only requirement is to make sure that you use one or more methods or related methods that we have learnt in this course.
2. A first step is to identify a particular dataset that is of interest to the student. One can find an innumerable number of datasets online, for e.g., at the repository *https://archive.ics.uci.edu/ml/index.php*.
3. While or after choosing a dataset one needs to frame a question of interest statistically. For this purpose, think of the following questions,
   - *What statistical model can I use in my data with respect to my problem of interest.*
   - *Am I trying to perform a the prediction of a continuous variable (i.e, a linear regression type problem)*
   - *Am I tyring to perform a classification type problem (i.e., logistic regression/LDA/QDA etc.).*
   - *Can I perform one or related hypothesis tests towards my problem of interest or to obtain more statistical evidence towards it.*
4. If required or in order to test the precision of your statistical model, you can consider splitting your data into a *testing* and *training* data. All model building is done on the *training* data and precision is measured by how well predictions are made on the *testing* data.
5. After setting up a statistical model one needs to go back to our class materials and see how to implement it in *R*. One can always consider implementing more than one model and then analyze which one does better. For e.g., different classifiers (logistic regression/LDA/QDA/KNN) can be implemented if one is trying to perform a classification exercise and then compared with each other with appropriate tools.
6. After all analysis has been conducted. All results should be illustrated in a properly formatted document. This includes presenting any technical output, for e.g. p-values, classification error, cross validation error, roc curves etc. and what you interpret from them.
7. All Data and Rcode utilized for your project should also be supplied as part of the submitted document/folder.
8. As a final step make sure to provide a conclusion about your findings from the project. This should be your interpretation in context of the specific problem that choose.

Rubric:

The project will be graded with the following rubric:

1. Problem setup (choosing a data, defining a problem and writing the associated statistical question) (20%),
2. Implementation (50%): have appropriate tools been utilized towards (1),
3. Depth (10%): for e.g., more than one approach to answering a question or has sufficient statistical evidence been illustrated (in keeping with what has been taught in the course),
4. Conclusion (10%): is your conclusion in coherence with the statistical analysis.
5. Presentation (10%): have you provided your results in a clear and readable format.

Additional Comments:

1. If you are unable to find a problem on your own, please email the instructor and a problem will be assigned to you.
2. If you are working in groups then you must include a section in your report indicating which participant did exactly what as part of the project.