

Fall 2020 Math 4 Project 1

In groups of three or four, explore an existing data set, and carry out tests of statistical significance on selected variables. You will summarize your analyses in a written report.

You will use the dataset “mtcars” included with a basic installation of R. A description of this dataset can be found at the following website.

<https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html>

- The dataset has two binary categorical variables and several quantitative variables. You will choose two of the quantitative variables to analyze.
- If your group is ambitious and wants to propose an alternative data set, I will consider it, provided it has two binary categorical variables (with at least 10 successes and 10 failures in each category) and two quantitative variables.

Timeline

- By midnight of October 6, one person in the group must email me the names of your group members.
- By midnight of October 8, one person in the group must email me the names of the four variables you will analyze, and your hypotheses about them. [5 points]
- By midnight of October 15, one person in your group must email me graphs and statistical summaries of your data, and one person must email me results of tests of statistical significance tests. [30 points]
- By midnight of October 22, one person in your group must email me the final written report. [20 points]
- By midnight of October 23, each person must email me a peer evaluation of their group members. [5 points]

Deductions for missed deadlines or incomplete submissions will be assessed to the entire group. If your group has an issue with a non-participating group member that you are unable to resolve, please contact Professor de Moor.

Report

The report should be structured as follows.

- Cover page: containing a title for your study, the names of your group members, and a short (<100 words) paragraph that summarizes the topic, design, analysis, and conclusions of your study.
- Introduction: Introduce the topic, say why you chose it, and explain why the reader should be interested. This section should be less than one page.
- Design: Define the variables you are analyzing and explain how each was measured or observed. Include units of measurement where appropriate. Describe the observational units and how they were selected. Most of this information should be immediately available from the R dataset documentation website given above.
- Analysis: Compare and contrast the distribution of your quantitative variables across the two groups of each categorical variable. Analysis should be both **Visual** (e.g. dot plots or histograms or box plots) and **Numerical** (mean and SD broken down by categories of the categorical variables). Construct confidence intervals and/or conduct hypothesis tests where appropriate.
- Conclusions: What trends do you see in your study results? How broadly do you expect these trends hold? What are possible confounding variables? (Remember, confounding variables must be associated with **both** the explanatory and the response variables!)

Visual and Numerical Analysis Scoring (15 points)

- Dotplots or histograms of quantitative variables overall. [5 points]
- Box plots of quantitative variable broken down by category. [5 points]
- Mean and SD of quantitative variables overall and by category. [5 points]

Inference Scoring (15 points)

- Confidence intervals for quantitative variables [5 points]
- Significance tests for each categorical variable [5 points]
- Significance tests for each quantitative variable [5 points]

*Note: In order to conduct a significance test, you will have to formulate null and alternative hypotheses. Do this **before** you do any data analysis. The rationale for your hypotheses will be part of your report's originality score.*

Written Report Scoring (20 points)

- Interest, originality, initiative [5 points]
- Results summary and overall insight [10 points]
- Quality of writing and organization [5 points]

Peer Evaluation

Separately from the report, each student must evaluate their group members on a scale from 1 to 10 (with 10 being the highest) based on their involvement in the project. Each student will also write a sentence or two describing the level of involvement of the other group members.