DANA 4820
Winter 2024
TERM PROJECT INFORMATION

The objective of the term project is to provide you with the opportunity to gain experience in the application probability distribution for categorical data, maximum likelihood estimation, frequentist approach and Bayesian approach*, chi-square tests for two-way and multi-way contingency tables, Generalized linear model, logistic regression, Poisson regression, Multinomial Logistic regression, and loglinear models for contingency tables and counts* to large records of data.

Each student is required to complete a term project. It will be worth 25 percent of your final grade. Those who fail to complete a satisfactory term project will receive a final grade for the course no higher than "D".

Students work in a group of four students whenever possible (if the number of students in a section is divisible by 4). Each team member in the project group is expected to contribute equally. The powerpoint file to be submitted should allocate one slide to include the names of team members and their estimated contribution in percentages. Members of any group who do not contribute to the term project in a satisfactory manner will receive a final grade for the course no higher than "D".

The requirements for successful completion of the term project are summarized as follows:

- Decide among team members on suitable topics. You are encouraged to discuss your ideas with your instructor.

- A project proposal (one page) outlining the objectives of your study, data source, and a description of variables will be due on **16th February, 2024.** You are welcome to hand in your proposal earlier if you wish to start the project sooner.

- You are to search for an (must be one) existing public dataset from reputable sources (one such source is www.kaggle.com). The dataset must have at least 150 rows of observations. You are required to provide the data files as part of your completed project.

- The **Presentation** will be in class on **1st and 3rd April, 2024**. The powerpoint file should be uploaded to Brightspace by on March 31, 2024 by 11:59 pm. You should include an introduction, objectives of the study, literature review, methods, results (tabular and graphical summaries), summary of results, limitations of the study, conclusions and recommendations. The grading rubric for your final project is found on the last page of this document.

DANA 4820 – ***Winter 2024***
Term project

The following are the items you should have in your final project.

I. Proposal [20 marks]

A project proposal (1 page, pdf format only) is due on ***February 16*** (the proposal must be uploaded to Brightspace group folder (see Assessment->Assignment) before 11:59 pm on the date due)**.** Each team submits ONE proposal. The proposal needs to outline the study objectives of your project, population of interest, information on how and where you obtain your dataset, sample size and a description of variables.

Your proposal must be

- Typed
- **Exactly** 1 page in length (any longer than 1 page may result in loss of marks)
- Your data collection involves using an existing publicly available dataset from a reputable source. Your proposal must identify the source of the dataset, including identifying which variables you will use. You may get some ideas on suitable datasets from these links:
  https://github.com/awesomedata/awesome-public-datasets
  https://www.kaggle.com/datasets
  https://archive.ics.uci.edu/ml/datasets.php

(you are not limited to what is listed here and you may seek your own datasets based on your team's interest).

In your proposal, you are to include

- Group members' names (student ID is optional, as a student has a right to keep his/her student ID private)
- Population
- Clear statements of the study objectives (can be phrased as questions) – e.g., How many principal components describe variables in the data set? What are the underlying factors associated with variables in the study? How best do we select variables for regression analysis.? There should be exactly 3 statements (questions of interest), for a group of 4 and 4 statements for a group of 5.

- The purpose of identifying these questions of interest is so that you can provide answers to them in your project, providing a clear direction of what you are supposed to do in this project. Providing answers to these questions is just a part of the project, and you are expected to do more than that (including drawing suitable graphs and charts, and applying as much of what you learnt in class to this project).
- Variables (you should have significant number of variables, ideally more than 5, for your analysis) Sample size (the original dataset must have at least 150 rows.)
- Source of dataset (where did you get this dataset).

The proposal is worth 20 marks (20% of the project) and marks are allocated as described below.

1. (4 marks) Define the questions of interest (that is, statements of study objectives).

2. (8 mark) Define the population of interest, size, source of data
You must be specific. E.g., "students" is not specific enough for a definition of a population. Use the format as taught in class: "The population is made up of all ....." is a good definition.

3. (8 marks) Define the variables you use in your project and provide the source of your dataset.
Identify variables. Try to select variables that have a good chance of being strongly related to the question you are trying to answer.

Please read the rest of this document before working on the proposal and make sure you understand the full project requirements in order to come up with a good proposal.

Details on the project:

Your powerpoint slides and data file must be uploaded to D2L/Brightspace. The powerpoint slides should cover the following:

Data Collection

1. Describe your data set.

2. How you obtained your data.
The data file should be provided (include units of measurement). You should upload the dataset.

3. Discuss possible bias.
Discuss the type(s) of bias you may expect in your study (even though you did not perform data collection but used a publicly available dataset).

Literature Review

You are to review 1 paper that uses one of your intended statistical methods in analyzing similar data. Review should be a brief description of what was done to answer research questions in the paper.

Analysis

1. Apply at least 3 of the methods discussed in class.

2. Apply to further statistical analysis

It is my expectation that you apply as much of what is taught in class as possible. Go through the lecture notes and check that you did not miss anything that could be added.

Conclusion

What's the conclusion to the questions you posed? Be sure to back up your answer by referring to your analysis. Summarize the results of your study**.** You also need to explain how you would improve this project if you were to do it again.

• Your completed project is in the form of a powerpoint presentation and must strictly be no longer than 25 minutes.

• This is a **group project**. Your group should consist of a maximum of 5 students

• Each team member in the project is expected to contribute equally and you should list down in one slide the names of the team members and their estimated contribution in percentages (e.g., Alex 25%; Bob 25%; Claire 25%; Donnie 25%)

• **Your project is graded as follows:**

Proposal	**20 marks**
Powerpoint content Part I Sampling and Data collection	**10 marks**
Literature Review	**10 marks**
Analysis	**40 marks**
Conclusion	**10 marks**
Multimedia (relevant photos, videos, etc.) and powerpoint overall quality (max length 10 min)	**10 marks**
	Total: 100 marks