Final Project

Statement

Datasets:

The data that will be used for the project is the US Behavioral Risk Factor Surveillance System (BRFSS) dataset. This dataset is composed of 330 variables from different topics with a total of 491,775 samples. A description of the variables and how the dataset was created can be found here.

Deliverables:

You will have to submit two files through **Moodle**:

- **A report in PDF format** that contains the main points of the project, justified answers to the proposed questions, and analyses of the results. The report does not need to be long but should demonstrate that you worked through the whole statement. Do not include figures or code without a comment about it. Remember to include a conclusion section.
- **A compressed folder** with all your code files and any additional file that you might want to attach (for example, a model which takes too long to train).
- **The organization and the quality of the code will be assessed and may penalize the final grade of the assignment.**
- **The format of the report will be assessed and may penalize the final grade of the assignment.**

Questions:

The objective of this part of the final project is to use the programming and statistics techniques learnt during the course. You may use R or Python to complete the tasks. For this project, four main sections should be developed:

1. **Research questions**
   In this section, you should come up with research questions that can be answered with the provided dataset. Try to make them interesting. If you are a group of 2 people, come up with 2 research questions. If you are a group of 3 people, come up with 3 research questions.
2. **Data**
   In this section, sampling techniques shall be used on the data to obtain a sample of the population from which we can draw conclusions on the whole dataset/US population. Here are some ideas for questions that may be answered (you can add more):
   - Describe the sampling method, which technique you have used and why.
   - May the results of your EDA be generalized to the whole population from the sample you have chosen?
   - Discuss potential sources of bias based on the dataset description, how can these biases affect your conclusions?
3. **EDA**
   In this section, perform an explorative analysis to answer the proposed questions in the previous section. Justify the results and draw conclusions based on this analysis.
4. **Inference statistics**
   In this section, perform a statistical analysis on the data to respond to the research questions proposed in section 1. You have to answer **at least one question via confidence interval** and **at least one question via hypothesis testing**. Alternatively, you can answer one of the questions using a Bayesian model.