|  |  |
| --- | --- |
| **INFO 4120 Midterm Project**  **Professor: Langehennig** | 1 |

**PROJECT PURPOSE**

This project is designed for you to share with me some of what you have learned with Python at this point in the quarter (programming, data management, and statistical analysis modules). Your project will be to collect and summarize some data, do some basic checks, recoding, and summaries of the data. Then summarize what you find. You should define at least 2 research questions to explore in the data.

The topic for your data is of your own choice. But note the requirements for data below.

**Basic Requirements:**

Note: In some cases, there is a REQUIRED and also an ADVANCED option you can use to move your grade from average to above average.

|  |  |
| --- | --- |
| **REQUIRED** | **ADVANCED** |
| **PART ONE** | |
| You need to use at least 2 datasets that should be in .csv or .txt format.  Include the file(s) in your submission. | You use more than 2 datasets. (Note you can break a single dataset you find into 2 to meet this requirement) |
| The datasets when combined must have at least 1 ‘Categorical’ variable and 1 numerical variable and a total of at least 4 variables. |  |
| Create a Python Jupyter Notebook to read in your dataset(s), perform the analysis needed for your report, and includes the text of your Report. | Well documented/commented Python code. |
| **Written Report:** Discuss what you have discovered about your data. The report should have the following components: |  |
| EXECUTIVE SUMMARY   * (include introduction and basic conclusions) |  |
| RESEARCH QUESTIONS   * List at least 2 research questions you are exploring in your data | Exploring more than 2 research questions in your data. |
| THE DATA   * Use a pandas dataframe to store your data * Include your data sources * Include "word" descriptions of each of the data columns in your final dataframe and an example of the values stored in each of the columns. For each column, look for missing values (fill them if appropriate or use code to report there are no missing values). * Also recode at least one variable and create bins for one continuous variable. Make sure to describe what you did. * Save this data as a .csv “final version file” for your input data for PART TWO. | Program uses any package that we did not cover in the class notes. |
| DATA EXPLORATION   * Create basic summaries for your variables. * For continuous variables this would include basic min, max, median, mean, and standard deviation. Include appropriate distribution graphs. * For categorical variables, this would include counts for each category. Include appropriate graphs. * Create at least two summaries that compare multiple variables at the same time. * Discussion of your data exploration. What did you learn about the data from these explorations? * Explore each of your questions   + Include the appropriate analysis to answer the question   + Include a conclusion for each research question |  |
| CONCLUSION   * Include an overall conclusion from your data exploration and analysis of your research questions |  |
| Save as an .html file |  |
| **PART TWO** | |
| Create a second Python Jupyter Notebook that, when run, will present the user with menu choices that will allow a user to interactively explore the .csv final dataset you created from your Data Processing in Part One. | Script avoids repetitive code with loops, conditional code, and no ‘hard coding’ of values to be more flexible, etc. |
| Present the user with at least 4 choices of getting summaries from your data.  At least two should include a graph.  Include labels, legends, and descriptive summaries where appropriate. | Have more than 4 choices. |
| The menu should run from a function. |  |
| At the end of your code, call the function **multiple times trying all** of your menu options so I can see examples of your output. |  |
| Save as an .html file |  |

* **Submit your original .csv/.txt datasets.**
* **Submit your final .csv merged data for Part Two**
* **Submit your .html and .ipynb for Part One (Report)**
* **Submit your .html and .ipynb for Part Two (Menu Script)**