**Instructions**: **(Points will be deducted for failure to follow instructions)**

**Use the *UCDavisData* dataset to answer the questions below using PyCharm. Write a professional report of your analysis.**

1. You must turn in neatly formatted copies of both the write up, your python code as a single script, and your sample exported as a csv.
2. Your script must be thoroughly commented out, demonstrating that you understand what each block of code is doing.
3. Your script must run from top to bottom without any errors. I should only have to change the file path to import the data and click run.
4. You should save your files using the naming convention, lastname\_firstname\_pythonproject (e.g. carder\_nicole\_pythonproject.doc and carder\_nicole\_pythonproject.py)
5. You should include a title page with your name, course, project name, and data.
6. You should include page numbers in your report.
7. Your results must be written in clear, professional English, with proper grammar, spelling, and punctuation.
8. Your answers should be numbered according to the original question number.
9. You must include proper formatting and descriptive titles, axis labels, and data labels on all graphs and tables. You will likely need to clean up some of the category names in the data frame before creating your visualizations.
10. Copy and paste your tables and graphs neatly into your final document. You can clean up your tables in excel, if you wish, but the analysis must be performed in Python.
11. Graphs and tables should be number sequentially according to the way they appear in your write up (i.e. Figure 1: Hist…, Figure 2: Box…, Table 1: Data…, Table 2: Desc…, etc.). Be sure to reference the graphs and tables in your write up when reporting on the analysis.

**Data Summary:**

Data from n=173 college students at the University of California Davis in two different majors. Non liberal arts (n=148) and liberal arts (n=25). The data were collected in the Spring quarter of 2000.

**Variable Names in order from left to right:**

|  |  |  |
| --- | --- | --- |
| **Column** | **Name** | **Description** |
| A | **Sex** | What is your biological sex? (Male or Female) |
| B | **Seat** | What is your seat located in the classroom? (Front, Middle, Back) |
| C | **libarts** | Are you a liberal arts major or not? |
| D | **TV** | On average, how many hours of TV do you typically watch in a WEEK? |
| E | **computer** | On average, how many hours do you spend on the computer in a WEEK? |
| F | **Sleep** | On average, how many hours of sleep do you get at night? |
| G | **alcohol** | How many alcoholic beverages do you consume in a week? |
| H | **Height** | What is your height, in inches? |
| I | **momheight** | What is your mom’s height, in inches? |
| J | **dadheight** | What is your dad’s height, in inches? |
| K | **exercise** | On average, how many hours of exercise do you typically get in a week? |
| L | **GPA** | What is your current GPA? |

|  |  |
| --- | --- |
| **Univariate Quantitative Analysis:** |  |

1. Provide a table of descriptive statistics for the quantitative variables (including all measurements of central tendency and dispersion).
2. Create a histogram and boxplot for 3 quantitative variables of your choice.
3. For the 3 quantitative variables, explain why the mean or the median is the best representation of central tendency and the standard deviation or IQR of dispersion. This is decided by examining the histogram and boxplot. Identify the appropriate measures and state the associated values from the descriptive statistics table in your write up.

|  |  |
| --- | --- |
| **Univariate Categorical Analysis**: |  |

1. Create univariate frequency tables (frequency and percent) for the 3 categorical variables.
2. Create a bar chart and pie chart for the 3 categorical variables.
3. For each variable comment on the frequency distribution including the mode.

|  |  |
| --- | --- |
| **Bivariate Analysis 2 Categorical**: |  |

1. Using the Pivot Table tool, provide a contingency table for two categorical variables that you are interested in exploring the relationship between. Determine which variable you think may ‘explain’ the response of the other variables. The variable that explains should be on the rows and the variable that responds should be on the columns.
2. Report on the frequency and percent distribution of the two variables. Comment on anything you find particularly interesting about the relationship.
3. Create a grouped bar chart, a stacked bar chart, and a 100% stacked bar chart. The variable that ‘explains’ should be along the x-axis and the variable that ‘responds’ should be represented by the colors of the bars in the legend.
4. Describe the differences between these graphics and determine which graph(s) is a better visualization of the relationship of the two variables.

|  |  |
| --- | --- |
| **Bivariate Analysis 1 Cat 1 Quant**: |  |

1. Perform a stratified analysis of the mean of the quantitative variables “TV”, “computer”, “Sleep”, “alcohol”, “exercise”, and “GPA” by the “Seat” variable. Compare and contrast the mean of the quantitative variables between the categories.
2. Pick one of the quantitative variables that you feel has an interesting relationship with your categorical variable. Explore the relationship further by creating side-by-side boxplots of the quantitative variable stratified by the categorical variable you chose.
3. Report on the distributions of the quantitative variable across the different categories. Compare things like the median, IQR, range, quartiles, and outliers. Comment on anything you find particularly interesting about the relationship.

|  |  |
| --- | --- |
| **Bivariate Analysis 2 Quant**: |  |

1. Create a scatterplot of two quantitative variables that you are interested in exploring the relationship between. Determine which variable you think may ‘explain’ the response of the other variables. The variable that explains should be on the x-axis and the variable that responds should be on the y-axis. (For example, the number of square feet in a house explains the price of the house. Square feet is the explanatory variable and price is the response variable)
2. Describe the relationship you observe between these two variables, based on shape, strength, and direction.

|  |  |
| --- | --- |
| **Confidence Interval**: |  |

1. Construct and report 90%, 95% and 99% confidence intervals for a quantitative variable of your choice.
2. What do you notice as the confidence level increases? Interpret the 95% confidence interval in context.

|  |  |
| --- | --- |
| **Variable Creation**: |  |

1. Create a new categorical variable from one of the original quantitative variables.
2. Explain how and or why you chose the cut points for your categories.
3. Perform a stratified analysis of an original categorical variable of your choice by your newly created categorical variable. Report on the frequency and percent distributions between the categorical variable. Comment on anything you find particularly interesting about the relationships.

|  |  |
| --- | --- |
| **Sampling**: |  |

1. Generate a random sample of 30 observations called “ucdavis\_sample1”.
2. Export your new sample to a CSV and turn in your sample with the project deliverables. Use the naming convention lastname\_firstname\_gss\_sample1 (carder\_nicole\_ucdavis\_sample1.csv)