

Descriptive Stats Project

Due by midnight on Sunday, June 30

Project Overview

- Your goal is to describe the characteristics of the typical student taking an on-line statistics course. You are given the Student survey data from the past couple of semesters. A list of the actual survey questions is included with the project link in Blackboard. You might want to look over this list when you are interpreting your results.
- I have gone through the data and performed a preliminary “clean up” but I have left in a few errors and ambiguous responses. Your first step will be to look through the data and finish the “clean up”. (FYI the official term for cleaning up data is “data scrubbing”.) You will then summarize and analyze your cleaned-up data using the techniques from Weeks 1-4.
- Your report will consist of the answers to each of the **Project Task** questions. In some parts, only a table or graph is needed, in others an interpretation or explanation is needed, and in others both are required. The longest section will most likely be the justification of your data scrubbing.

Submission Guidelines

- Submit your report as a well-formatted Word document.
 - SPSS graphs and tables should be pasted into the appropriate location in your document.
 - Make sure to change the size of your graphs in Word so that they fit nicely into your report.
- Attach your Word document and Excel data file containing your clean data. Save your Word document as “DS Project” followed by your initials. Save your Excel document as “DS Project data” followed by your initials. They must be submitted as separate documents (not in a ZIP file).
 - Your documents must be saved in Windows format with the “.docx” and “.xlsx” extensions. This means that I can’t accept reports saved in Pages or Google docs.
- The due date is Sunday, June 30 and late submissions will be marked down 1 point for each day late. The last day to submit your project for grades is Friday, July 5.

Project Tasks

[2 pts] Format your report so that it looks professional. Try to look at your report from the eyes of an outsider. You might want to ask someone around you for an impartial and honest opinion because that is how I will look at your work. You might even insist that they be brutally honest in their critique – it will help you in the long run.

- All graphs should be created in SPSS and formatted so that the categories (represented by the bars or slices) can be distinguished when printed in black & white. This can be done by adding labels to bars or slices, or shading with patterns.
- You will need to export your SPSS output to a Word document so that you can paste your graphs and tables into your Project (without the irrelevant SPSS output code).

1. **[4 pts]** Read the document “Data Verification Guidelines” and follow these guidelines to “clean up” the data. You must “clean up” **all** variables, not just the ones you are analyzing in your report.
 - ❖ Since “cleaning up” is a subjective exercise, you will need to justify **all** changes you made to the data and the reasons why you made each change. Do not assume that a change you make is obvious even if it only involves changing units. ALL changes made to a data value must be documented. This will likely be the longest part of your report and must be included in your Word document; do not use the “Insert comment” feature in Excel to document your changes.
 - ❖ Some changes occur only once or twice and need a specific explanation. Other changes are made to multiple entries and can be explained as a group. Here are some examples of wording you might use to indicate these changes and your reasons for making the changes.
 - **The entry “New Yurk” in row 7 of the “State” variable was corrected to “New York”.**
 - **For consistency, all state abbreviations were changed to the full state name, for example “MA” becomes “Massachusetts”, “NY” becomes “New York”, and so on.**
 - ❖ It may be necessary for you to do a bit of simple research to validate certain types of data if you are not familiar with the variable. For example, names and/or spellings of U.S. states, other countries, or makes of vehicles.
 - ❖ When you import your Excel data into SPSS make sure to verify that your variables have the correct measures. You must also document any values that have been removed by SPSS when you do this. You may actually want to review these and decide if you want to handle these “lost values” differently in your clean-up process so that you don’t lose them.
 - ❖ Remember that you need to submit your Excel file showing your “cleaned up” data.
2. **[3 pts]** Select a categorical variable from the list below and analyze it.

#2	Eye color
#4	Birth location
#10	Make of first vehicle
#11	Cell phone provider

 - a) Make a frequency table for the variable you chose.
 - Your frequency table should include at least 4 different categories (rows) and no more than 8. If necessary, combine responses into an “Other” category (but give it a more descriptive name than “Other”).
 - If you use an “Other” category, it should not include more than 25% of the total data values.
 - b) Create either a bar or pie chart in SPSS that displays the relative frequency distribution of your chosen variable. Make sure that your chart displays the relative frequency for each data value.
 - Note that no single category can make up more than 60% of the data values.
 - Make sure to include a descriptive title on your graph. You should also add enhancements to your graph to improve its readability when printed in black & white. Refer to the Week 1 notes for the specific requirements for each type of graph.
 - c) Write a short paragraph interpreting the above results. What do they tell you about the typical on-line statistics student?

3. [7 pts] Select a quantitative variable from the list below and analyze it.

- | | |
|----|------------------|
| #1 | Height |
| #3 | Age |
| #5 | Texts |
| #6 | Surfing |
| #7 | TV |
| #8 | On-line shopping |
| #9 | Eating out |

- Use SPSS to find the mean, standard deviation and five-number summary of your variable. Include your SPSS output. Make sure to clearly state your five-number summary.
- Create a boxplot for your variable in SPSS and insert it in your document. Add a descriptive title.
- Calculate the upper and lower fences for your variable and use these to determine if there are any outliers. Make sure to show your calculations.
- Use the mean and standard deviation to determine whether there are any unusual and/or rare values. Make sure to justify your answer and show your calculations.
- Make a frequency **table** for your variable by hand (not using SPSS) by grouping responses into “bins”. Make sure your bins are the same width and that there aren’t any gaps between successive bins (but a bin can have a frequency of 0). Remember to use the left-endpoint rule.
 - You should have at least 4 bins and no more than 8 in your frequency table.
 - No single bin can contain more than 40% of the total data values.
- Write a short paragraph interpreting the above results. What do they tell you about the typical on-line statistics student?

4. [4 pts] Select two of the quantitative variables listed in Task 3 and analyze the relationship between them.

- Create a scatterplot in SPSS. Make sure to add a descriptive title.
 - Add the regression line to your scatterplot. Make sure the regression equation and R^2 -value are displayed.
 - There must be at least some correlation between the variables you choose. Specifically, the R^2 value must be **at least** 0.1.
- Describe the strength, form and direction of the relationship.
- Does your scatterplot have any outliers? Make sure to show or explain how you know.
- Interpret the slope and intercept of the regression equation in context. You must also indicate whether the intercept makes sense or whether it is just the starting point for the model. Make sure to refer to the Week 4 notes for what it means to interpret “in context”. Many students lose points because they don’t check.
- Write a short paragraph interpreting the above results. What do they tell you about the typical on-line statistics student?

Note that if you are having difficulties with this part of the Project, you should refer to your Week 4 notes and verify that each of your variables has the correct measure.