Northeastern University
College of Professional Studies

ALY6000 Module 1 Project Assignment 2

Instructor: Dr. Dee Chiluiza, PhD

Overview and Rationale

This assignment is designed to provide you with hands-on experience in performing descriptive statistical methods on a data set. The data set is provided in an Excel workbook and contains a wide range to data types that you will need to work with.

This assignment also allows your instructor to make a general evaluation of your basic skills on data descriptive analysis and presentation.

Assignment Summary

Project Due Date is Sunday at 11:59 pm.

Using the data provided in the attached Excel workbook, apply the methods of graphical and numerical descriptive statistics.

Follow the instructions in the project document to analyze the data presented in the Excel workbook. Then complete a Power Point report summarizing your results.

Important note on the report: for this project, your report will be a Power Point file containing no more than 8 slides.

Make sure to give a good presentation format to your tables and graphs. Don't just copy your raw graphs from Excel or R, provide the proper X- and Y-axes labels, make sure the text is big enough to be easily read.

Both figures and tables should have good titles followed by a short description of the data presented.

Titles for tables go on the top. Titles for figures go on the bottom. Make sure your power point report looks professional.

Create a good introduction slide:

First, make a list of the topics you must cover in your report (this is a good practice for any report). In this case, for example:

1. Talk about US population and migration rates.
2. Talk about the importance of descriptive statistics,
3. Talk about the importance of good graphical display of data.
4. Explain the purpose of this assignment.
5. Briefly describe the dataset you will analyze.

What other topics you can think of? Make sure that all the topics are addressed in your introduction, use at least two academic references when needed. Be brief.

Northeastern University
College of Professional Studies

Use the data presented in the Excel File "Module 1 Data.xlsx" to develop the following tasks:

This data set contains the 2017 population for all counties in United States.

Important: Please develop this assignment using R. If you still don't feel ready to use R, please feel free to use Excel instead.

Q1. Basic descriptive statistics.

Use the data presented in the tab "**Three_States**" which contains the population data for all counties in Alabama, Mississippi and Louisiana.

Obtain basic descriptive data from each state. Include as many parameters as you consider important (mean, SD, median, etc).

Using a well-organized table, present your results in one slide.

This table should allow easy comparisons among the three states.

Explain any difference you observe between the three states.

Q2. Pareto Chart

Use the data in tab "**Population_State**."

Organize the table based on population size per state.

Create a Pareto chart of the top 25 states with the highest population.

Present the graph in one slide.

Write a short line describing the utility and purpose of Pareto charts.

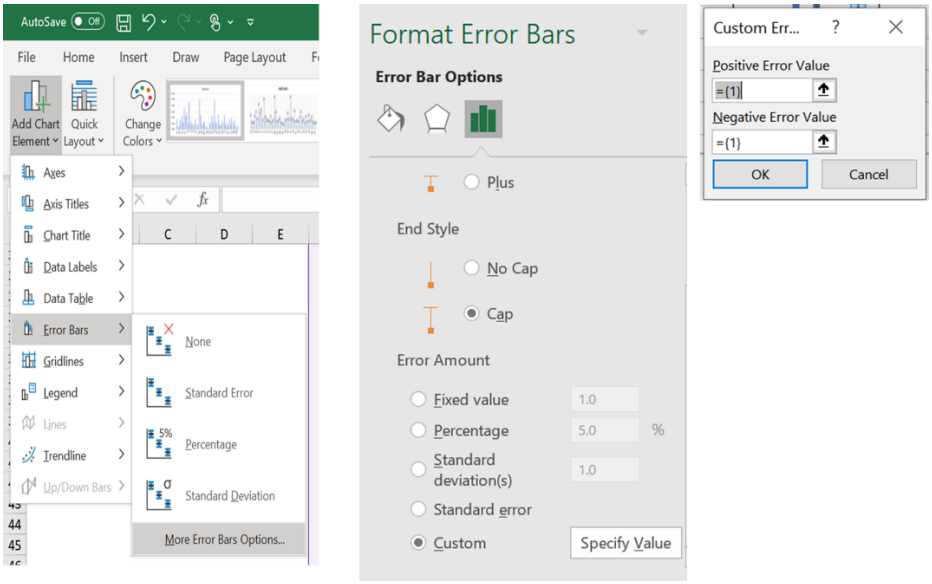
Write a short line describing, in your opinion, what is the most important observation on your graph.

Q3. Let's see the difference between bar graphs and box plots.

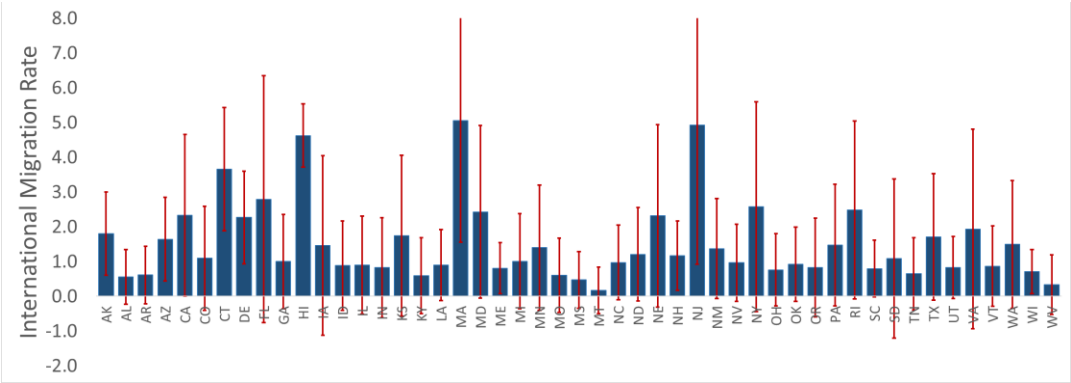
Use the date in the tab "**IMR_bargraph**"

(a) Create a bar graph to display the mean IMR for all the states in the data set. Add SD as error bars. Follow these steps:
Select the columns for State and Mean. Go to "Insert" in the top menu, then select "Insert column or bar chart."
Extend your graph to the right to make sure all states are displayed.
Then add standard deviation as error bars, follow these steps.
Click on your graph. Go to "Design" on the top menu, then click on "Add Chart element" **>** "Error Bars" **>** "More Error Bar Options."
Click "Custom" (at the bottom)
Click "Specify value."

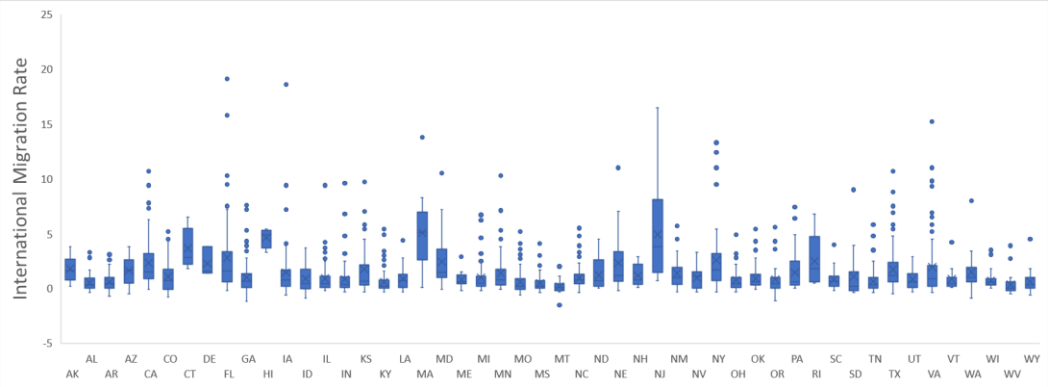
Activate the “Positive Error Value” and in the data table, select the all the points under SD (standard deviation).
Repeat the process for “Negative Error Value”



Your graph should look more or less like the graph below.
Add your colors, font type or size; develop your own style. Notice that I used minus 2 as the lower limit on the x-axis with the purpose to see the bottom limits of the error bars. If you did it right, the error bars should be different from each other to reflect their different nature.



(b) Create a box plot for all states using the data on tab “**IMR_boxplot”**.
Select the data for State and IMR.
Go to “Insert” on the top menu, and from the graphs, select “**Box and Whisker**.”
Your graph should look more or less like the graph below.
Add your colors, font type or size; develop your own style.



Present both graphs in one slide. Notice that for the bar graph we used only the mean and standard deviation, and for the box plot we used the whole data information for each state, which include the IMR values for individual counties.

Write a small paragraph to explain, in your opinion, which graph is better to present the data.

Write a short line describing, in your opinion, what is the most important result on your graphs.

Q4. Choose your graph

Population density is the relationship of the population per unit area.

Use the data in tab "**Density**."

Add the size of each state in **square kilometers** (column C). This data is not provided, you must find it in internet. Provide a reference for the source of your data.

Calculate the population density (Column D).

Create a graph of your choice that in your opinion, better display this data (population density of each state). There are many options in excel. You can also look at the internet for ideas.

Present this graph in one slide.

Write a short line to explain your choice of graph.

Make careful observations of the results.

Format & Guidelines

A report submitted in **Power Point** format, no longer than **8 Slides**.

Make sure your name is clearly described in the tile page.

Present all relevant graphs showing your work and results, making meaningful observations.

The power point file should follow the following format:

(i) *Title slide*.

(ii) *Introduction slide* (Develop ideas and concepts about US occupations and statistical tool to be utilized. Be substantial and include references).

(iii) *Analysis slides* (each slide must have a title. Include all figures and tables as indicated above).

(iv) *Conclusion and Reference slide* (One slide including all your conclusions and final observations about the statistical tools and graphs used, and about your results. Use references to support your conclusions. In this slide also include a minimum two academic references, proper format will be graded

Hint: Make sure that the text in your graphs (labels) are clearly visible, and that each graph or table has a short description of the data presented.

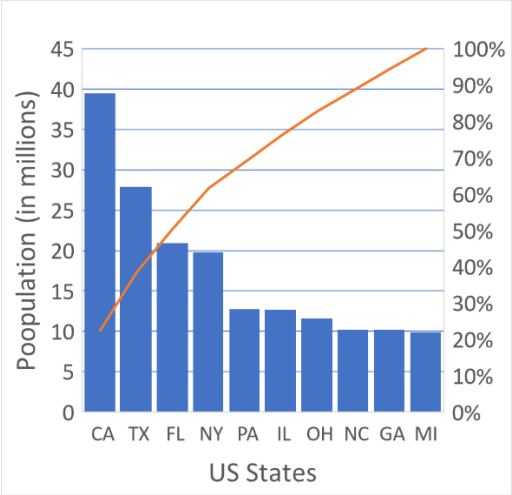


Figure 1. Pareto chart indicating the top 10 most populous US states during 2017. Bars indicate descending order of population in millions among those 10 states, line graph indicates cumulative percentage.

Please remember: your reports are very important, in class or at work. Make sure they look professional. Your reports should be as short as possible but containing all the relevant information. In general, use your reports to let me know you have learnt and understood the statistical tools you just used, and for extra points, use deep critical thinking to provide examples of practical applications in real life scenarios.
