Module 6 Final project

Course: ALY6000 Introduction to Statistics and Data Analytics

Prof. Dr. Dee Chiluiza, PhD

Data Analysis

Overview and Rationale

Being able to ask appropriate questions of data is an important part of the work of data analytics. It is also critical to be able to interpret the results of the analysis. This assignment is intended to familiarize you with the data sets and to get you thinking about key business questions you can answer from this data.

Course Outcomes

This assignment directly links to the following learning outcomes from the course syllabus:

CO2: Identify basic statistics measures of central tendency and variance

CO3: Utilize the "R" as a tool set for processing and analyzing basic data

CO4: Demonstrate how the analysis of data impacts operational and strategic decision making.

CO5: Visualize data in a compelling way to enable data driven storytelling.

Assignment Summary and Instructions

Choose one of the data sets from the list presented in this assignment.

The assignment has three parts. In the Appendix of this assignment, you are provided an example of how the questions in Part I should be answered.

Initial analysis of the data set

Do not add these observations on your report

TO DO:

Open the Folder you created on your computer for this class.

This data set contains information about Global Sales in 2016 of a Super Store Company. The variables names are self-explanatory.

Copy and paste inside that folder, the data set you choose (there are three sets).

Open the RStudio project for this class and create a new R Script file named:

M6Final_Mylastname_Myfirstname.R

Replace Mylastname_Myfirstname with your last name and first name, respectively.

Import the CSV data set into your R Studio, it will appear in your environment window.

When you are importing the data set into R, remove (skip) the following fields you will not use:

Row ID • Order ID • Ship Date • Postal Code • Customer ID • Customer Name • ProductID • ProductName.

In order to understand the data, we first need to run some descriptive statistics on the data sets.

---

*Remember my recommendation: "**Take your data set on a Date**."*
Spend time understanding the whole set, specially, how many variables (columns) does it have? What type of variables they are (categorical or numerical)?
If they are categorical, how many categories do they have?
How many observations there are.
What type of information numerical variables contain?
(Is it price, weight, distance, speed?)
Are there missing values or errors in the set?
This is basically the process of descriptive statistics presented to yourself.

Notice that some categorical variables contain hundreds of categories (product ID), or about 100 (Country), while others are under 20 categories (Region, Market, Segment, etc.). Keep

this in mind when choosing the categories you plan to manipulate. Make sure you do not overwhelm yourself with big tables, do only if you feel comfortable, this is your challenge.

My recommendation: Create graphical or numerical description for every category, these are your "couch figures," I also call them your "living room" figures, your "coffee time" figures, etc. Remember I mentioned this in class. These tables and figures **are for you only! To know and to understand your data set.** Do not present them to your audience, ONLY what is interesting.

Part 1. Cover page and introduction section

Present the following on your report

Following the recommendation you have received during the class:

1. Present a good Title page.

2. Present a well informative introduction section, this will measure your understanding of the topic and analytical processes for data analysis:

Your introduction needs good information and good organization. This applies for any report you make. Try to separate each topic in a paragraph, and use references in all of them to support the information you are presenting:

- **General topic:** Show your understanding of the topic related to the data set, in this case, sales and anything related to the business: corporations, global market, importance of analytics for this industry, etc. You choose the aspect you want to present to your audience. As a guide, write a paragraph of 5 to 8 lines for this topic.

- **Data set description**: Briefly mention the nature of the data set you are about to use.

- **Problem identification**: Make sure the problem is real and explain why it is important to address it. Imagine that you work for this company and you are given this data set. What questions would you ask to improve the company performance? Would you focus on profits, market size, sales, shipping cost? You choose the aspects that are very interesting to you. That is why I recommended you to "*take your data set on a date*" and learn very well every aspect of the data.

- **Plan**: Briefly describe your plan to address the problem, in this case, the analytical and visualization tools you plan to use.

- Use references to support each aspect/information presented.

Part 2. Analysis section (2 tasks)

TASK 1

TO DO: Present on your report:

- Start by providing the following for each appropriate variable in the dataset.
- When choosing the variables you want to process, choose a minimum of 5 categorical and 5 numerical variables, then use them to create logical/business questions.
- First briefly explain the whole data set and then and explanation why you chose those 10 variables (5 categorical, 5 numerical).

Present a summary of the data you choose using tables and graphs.

1. Graphs that help visualize the data. These can be bar charts, histograms, pie charts, box plot, etc. Be sure the chosen graph best represents the information you want to highlight
2. Explain the story the data is telling you, apply critical thinking.
3. What business question do your descriptive analyses answer? Provide a brief discussion of the findings.
4. If there are any unusual values, discuss them.
5. Present figures comparing variables, for example: sales per country, sales per region, total sales per segment, total profits per region, etc. Use your imagination. This is where you need to be careful in choosing the 10 variables for your project, you must understand them and be ready to ask interesting questions that you know you will be able to answer using R.
6. Identify additional questions that the data is leading you to ask. What new attributes are needed to answer those questions? This is important because you build expectative in your audience and you will answer it in Task 2 by creating new calculated fields.

Important: When selecting your 10 variables, use R codes to create the new set, DO NOT create new data sets (CSV files) outside R to import them. Again, select the variables into a new data set using the codes you learnt in class (check previous reports).

To provide a good presentation format to your tables, you can export the tables you created in R using the command write.table(x, file =, row.names = F, sep=",,"), where x is the name of the object you created in R, file= is the name you want to provide to your csv file.
This will save the csv file inside the Project working directory (same folder in your computer). Then go to that folder, open the file on Excel, and provide the table with a good presentation format before pasting the image on your report.

For example, you remember the inchBio.csv data set we used in class.

If I create a data frame with species using: as.data.frame(bio$species) I will obtain a long list of observations.

However, if I first create a table, and then I use the same code, I obtain a table listing the observations per category. These codes follow a logical sequence:

```
Table1 = table(bio$species)
Table2 = as.data.frame(Table1)
Table3 = rename(Table2, Species = Var1, Frequency = Freq)
write.table(Table3, file = "TableName.csv", row.names = F,
sep=",")
```

Try those codes with the inchBio set you already have inside your ALY6000 R project. Observe the results you obtain, and if you prefer, use that strategy to move data to excel for table formatting. You can try other strategies; I am just providing a simple solution. **Do not** submit any excel file you created this way; I will check them using your R codes.

TASK 2

Create at least one new calculated field, also called a new attribute or new variable, based on the data and the questions you identified in Part 1.

For your data set, compute calculations between appropriate variable values and create a new variables, remember the **mutate()** code you used on a previous project. As a reminder:

```
FinalTable = mutate(Frequencies_df,
                    CumFrequencies = cumsum(Frequencies),
                    Percentage = round((Frequencies)/676, 4),
                    CumPercentage = cumsum(Percentage)
```

An example: If the data shows yearly sales for different years, by month, calculate the increase or decrease in sales from month to month.

Another idea: Calculate the sales per country, region, or per segment. For this, you can:

(1) Create a new object containing only Country, Sales and Quantity.
(2) Create a new calculated field using (Sales*Quantity).
(3) Sort the new data using Country.
(4) Create a table to calculate the total sale per country.
(5) Present a bar graph with Country on the X-axis and Total sale on the Y-axis.

Another idea: Calculate the percentage that shipping cost represent per product price and see where it was more significant per country or per region. You can conclude that the company is spending too much money on shipping on a particular country and a new strategy should be developed.

Another idea: If you observe the price variable and multiply it by the quantity that was sold, keep in mind the discount given to some products: ((Price-(Price*Discount))*Quantity), and then remove the shipping cost also multiplied by the amount sold (ShippingCost*Quantity), you will see that the value you obtain is higher than the profits. This represents the cost associated with running the company. Can you calculate that value by creating a new category and then observe in which country or region the cost of running the company is the highest?

Another idea: Calculate the price of each product after the discount.

Use your imagination to ask simple logical questions to the data, questions that in your opinion, can provide meaningful information about how the company activity.

Please think about simple questions, do not complicate yourself. The purpose of task 2 is to measure your skills to create a new calculated field (a new variable) and to put it on perspective regarding categorical variables by asking simple logical business questions.

Compute basic descriptive statistics of the new variables (including the mean and median).

Present the data using tables and graphs and make significant observations about your results.

Notice that this task is basically an **open task**, where you need to use your imagination to create new calculated fields and use the new information mixed with one or more categorical variables to present interesting information to your audience.

Part 3. Conclusions, Reference and Appendix sections

Now that you have worked with the data, what is the data saying to you? Make a global analysis of the results you observed.

What have you learned about the attributes? What are some follow-up questions you would like to have answered? Identify 3-5 observations or follow-up questions that you have.

What did you learn from this project? What new skills you have added to your learning curve?

Properly cite all sources in the reference section using proper APA citation rules.

Complete all data management tasks in R and submit this file together with your report. Mention this file in the Appendix section at the end of your report.

What to Submit

A maximum 5 to 8 pages executive report (DOC or PDF format) with all your findings, figures, tables, observations, critical thinking applications, etc.

Due date

Saturday October 24th at 11:59 PM.