MIT 5032: Analytics Programming

Fall 2023

ASSIGNMENT 2

The due date for the assignment 2 is 11/12/2022 (11:59 pm). It must be submitted through the "Assignment 2" folder provided on Canvas. It is important that you specify your name in the assignment documents/attachments. The attachments name should follow the following convention:

MIT5032_ASSIGNMENT2_YOURNAME.DOCX (detailed project report including appropriate pictures)
MIT5032_ASSIGNMENT2_YOURNAME.R (R script file with comments and ready to execute code)

Problem 1 (1 Point)

A company has developed a new computer sound card whose average lifetime is unknown. In order to estimate this average, 200 sound cards are randomly selected from a large production line and tested; their average lifetime is found to be 5 years. The 200 sound cards represent a:

a) parameter

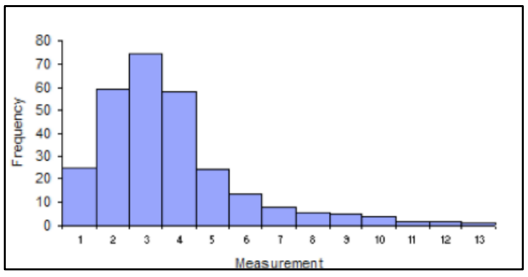
b) statistic

c) sample

d) population

Problem 2 (1 Point)

The histogram shown below is:



a) Symmetrical

b) Left Skewed

c) Right Skewed

d) None of the above

Problem 3 (1 Point)

If the distribution of data is right skewed, mean is smaller than median.

a) True b) False

Problem 4

You have been recently hired as a Data Scientist for the urgent care clinic and will spend the next two weeks reviewing all the patient records in this dataset. There is an ID number assigned to each patient, followed by patient age and gender. After this visit, the patients were mailed a survey to gather satisfaction scores. The results are represented in the opinion column. A score of 1 is very poor, 2 is poor, 3 is neutral, 4 is good, and 5 is very good. The clinic administrator has supplied the total time each patient spent in the clinic for this visit and also the number of prior visits to the clinic for each patient. The billing department has provided the payment type and the charges billed for the visit in charges column. The patient's admission blood pressure data come from the EHR.

ID	Age (years)	Gender (M/F)	Opinion	Visit Time (Min)	Charges ($)	Payment Type	Prior Visits	Admission Blood Pressure	
								Systolic	Diastolic
1	33.5	M	5	64.2	158	Self-Pay	4	136	71
2	21.2	F	2	69.4	159	Medicaid	3	112	65
3	56.4	F	1	81.1	178	Medicaid	0	156	88
4	53.9	M	3	31.6	124	Blue Cross	1	125	80
5	51.2	F	5	48.5	146	Aetna	8	133	62

The clinic administrator will ask you questions about these data and your answers will be used to set some performance goals and strategic objectives for the clinic. Considering the small portion of the dataset provided above, answer the following questions:

a. What type of data is found in each column (e.g., categorical [nominal]; interval [continuous], etc.)? (5 Points)

b. From the following numerical summary measures, which measures would you prefer to use to summarize the data in each column except for the column ID? (4.5 Points)

- Mean
- Median
- Count

c. From the following graphical methods, what type of methods would you prefer to use to represent the data in each column except for column ID? (4.5 Points)

- Bar Chart
- Histogram

Problem 5

Mini Case

1. MyHealthcare manufactures and sells blood pressure measurement and control products. Last year the company began selling its products online. Online sales have exceeded the company's expectations, and management is now considering strategies to increase sales even further. To learn more about the online shoppers, a sample of 50 transactions (refer to ShoppersData.csv on Canvas under Assignment 2) were selected from the previous month's sales. Data showing the day of the week each transaction was made, the type of browser used, the time spent on the website, the number of website pages viewed, and the amount spent by each of the 50 shoppers. MyHealthcare would like to understand the shoppers buying pattern in general. Specifically, the company uses the sample data to determine if online shoppers who spend more time and view more pages also spend more money during their visit to the website. The company would also like to investigate the effect that day of the week and the types of browsers have on sales.

Managerial Report

Use the methods of data analytics and Cran R to learn about the shoppers who visit the MyHealthcare site. Include the following in your report.

a. Appropriate graphical and numerical summaries for the length of time the shopper spends on the website, the number of pages viewed, and the amount spent per transaction. Interpret and discuss what you learn about MyHealthcare's online shoppers from these summaries. (5 points)

b. Report the frequency of transactions, the total dollars spent, and the mean amount spent for each day of week. What observations can you make about MyHealthcare's business based on the day of the week? Discuss and interpret your results. (5 points)

   Hint: Consider "aggregate" function available in the base R package to summarize data by group (day of week). Refer to R native help or the following URL for syntax related help: http://www.cookbook-r.com/Manipulating_data/Summarizing_data/

c. Report the frequency of transactions, the total dollars spent, and the mean amount spent for each type of browser. What observations can you make about MyHealthcare's business based on the type of browser? Discuss and interpret your results. (3 points)

d. Use appropriate graphical method and numerical summary statistics to explore the relationship between the time spent on the website and the dollar amount spent. Use the horizontal axis for the time spent on the website. Discuss and interpret your results. (3 points)

e. Use appropriate graphical method and numerical summary statistics to explore the relationship between the number of website pages viewed and the amount spent. Use the horizontal axis for the number of website pages viewed. Discuss and interpret your results. (2 points)

f. Use appropriate graphical method and numerical summary statistics to explore the relationship between the time spent on the website and the number of pages viewed. Use the horizontal axis to represent the number of pages viewed. Discuss and interpret your results. (2 points)

**********NOTE **********
Please feel free to consult your instructor by email or phone, if you have questions or need assistance!!

The following criteria will be used to grade the assignment:

- Data are explored and analyzed in many different ways.
- Calculations are detailed, accurate, and answers are correct.
- Graphs/pictures are labeled.
- Rigorous approach is used to solve a specific problem.
- Good coding practices are followed. For example, code are functional, readable, commented, and error free.
- Convincing conclusions are drawn.
- Writing style is understandable and organized in explaining investigation results and supporting conclusions.