

Descriptive Statistics Using R

AMPBA 2023

Assignment I

Total Points: 100

This deliverable has 40% weightage in the Consolidated Total Score.

Note: The Assignment submission form should be submitted as well. Your submission will not be considered without the Assignment Submission form being submitted.

1. Think of a business question, related to your job or your interests. State the statistical question and briefly describe the data you would collect to answer this question. Write a few lines (2-3) that explain why this data is appropriate. [5 points]
2. Which of the following are statistical questions?
 - a. How many days are in July?
 - b. How many Google Chrome users in Hyderabad use its incognito mode feature?
 - c. How often do Netflix users in India watch movies suggested by the platform?
 - d. How much is Elon Musk's salary?
 - e. Are temperatures higher in Texas, USA or Andhra Pradesh, India during summer months?[3 points]
3. Which of the following statements about the difference between a population and sample are correct?
 - a. Using a sample gives more accurate results than using a population
 - b. Both terms, population, and sample, can be used interchangeably in Statistics.
 - c. The population consists of all possible elements we are interested in measuring. A sample is a subset of the population.

- d. Drawing samples is too complex, so researchers typically collect data on all population elements.

[3 points]

4. A company has 50 employees. The minimum, maximum and median annual salaries are Rs. 3,20,000, Rs. 54,00,000 and Rs. 21,42,000 respectively. Four more employees have just joined the company with annual salaries Rs. 2,40,000, Rs. 54,12,000, Rs.2,84,000 and Rs.54,20,000. The updated median salary:
- a. Is less than Rs. 21,42,000
 - b. Is exactly equal to Rs.21,42,000
 - c. Is greater than to Rs.21,42,000
 - d. Cannot be determined from the given data

[3 points]

5. If a histogram representing the data points in a data is skewed to the right, then:
- a. Mean is less than median
 - b. Mean is equal to median
 - c. Mean is greater than median
 - d. No categorical comparative statement regarding mean and median is possible

[3 points]

6. Consider data of maximum temperatures (measured in degree centigrade), for 50 consecutive days in a town. The data are measured on which scale?
- a. Nominal
 - b. Ordinal
 - c. Interval
 - d. Ratio

[3 points]

Questions 7 & 8 are based on the following data table.

GMAT scores of applicants to a Business Analytics course						
Interval	501-550	551-600	601-650	651-700	701-750	751-800
Frequency	25	40	42	31	26	21

7. What is the median class of the above GMAT scores?

- a. 501-550
- b. 93
- c. 601-650
- d. 95

[3 points]

8. After the second round of applications, this data table was updated. 13 students were added to the Interval 551-600 and 22 students were added to 701-750. What is the modal class of the new data?

- a. 501-550
- b. 551-600
- c. 601-650
- d. 701-750

[3 points]

9. A survey by an electric company contains questions on the following:

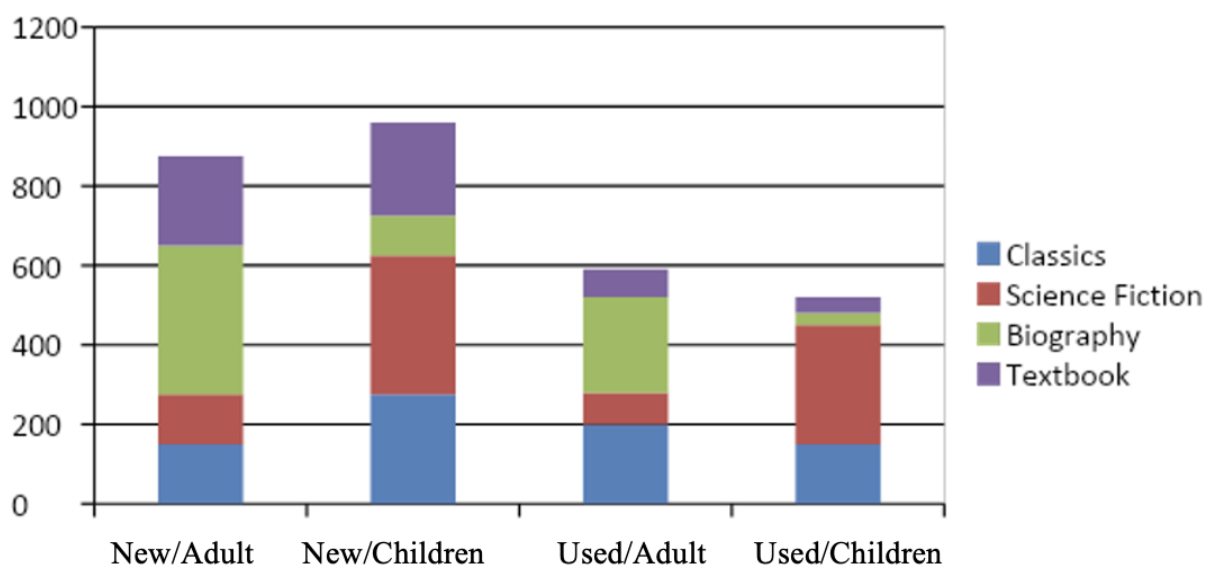
- a. Age of household head.
- b. Sex of household head.
- c. Number of people in household.
- d. Use of electric heating (yes or no).
- e. Number of large appliances used daily.
- f. Thermostat setting in winter.
- g. Average number of hours heating is on.
- h. Average number of heating days.
- i. Household income.
- j. Average monthly electric bill.
- k. Ranking of this electric company as compared with two previous electricity suppliers.

Describe the variables in these 11 items as quantitative or qualitative and describe the scales of measurement.

[5 points]

Questions 10 - 12 are based on the following information:

Matt's Bookstore sells both new and used books. Within each category it further sells both adult and children's books. The following chart gives the sales for each category of books for the year 2021.



10. In which category was sales of used books greater than new books?

- a. Children's Science Fiction
- b. Adult Classics
- c. Children's Classics
- d. Adult Biography

[5 points]

11. Which of the following statements can be inferred from the above chart?

- a. The category of children's books (new + used) that had the largest sales is also the category that had the least sales in the adult book (new + used) category.
- b. The category of adult book (new + used) that sold the most, is the category of children's book (new + used) that sold the least.

- c. The number of adult textbooks that are sold as new is twice that of the number of children's textbooks sold as used.
- d. Science fiction was the most sold category, followed by Classics, then Textbook and lastly, Biography.

[5 points]

12. Which type of book contributes the most to the bookstore's earnings?

- a. Children's book
- b. Adult book
- c. None of the Above
- d. Cannot be inferred

[4 points]

13. A real estate agent is trying to understand the nature of housing availability and home prices in and around a town in the United States. She has collected data from a random sample of 1047 homes sold in the last 12 months. Data was collected on the following variables and is available in the dataset named **assignment_1.csv** file.

- Price – the sale price of the house in \$
 - Living Area – in Sq. ft.
 - Bathrooms – number of bathrooms in the house (rooms with no tub or shower area are considered 0.5 baths)
 - Bedrooms – the number of bedrooms
 - Lot Size – size of the property on which the house sits (in acres).
 - Age – of the house in years
 - Fireplace – whether the house has a fireplace (Yes = 1, No = 0)
- a. Prepare a brief report summarizing the home values (prices) in and around this town. Use both graphical and numerical summaries. Your report should briefly describe what those summaries tell you, and anything of particular interest.
 - b. Create a histogram and boxplot for the *Living Area* variable. What does the histogram tell you that the boxplot does not, and vice-versa? Is the distribution symmetric? Comment.

[a: 45 Points, b:10]

Note: Questions 1-12 are based only on sessions 1-2. You can add ideas you learn in sessions 3-5 when answering question 13.

General Instructions:

1. This is an Individual assignment.
2. Do NOT submit .zip files otherwise the submission will not be considered.
3. Please include your name and PGIDs in all files in the submission.
4. Any late submission will attract a penalty as mentioned in the course outline.
5. The honor code for this submission is **2N-b**.
6. **Please look through the honor code restrictions carefully before attempting the assignment as there will be strong consequences for breaking them.**
7. **Please adhere to the given instructions, otherwise your submission will not be accepted, or a severe penalty will be applied**

Deadline: 21st January 2023, 11:55 pm