Descriptive Statistics with Python

Graded Assignment

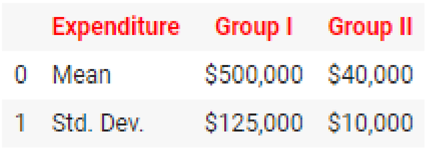
Section 1

1. A statistics test was conducted for 10 learners in a class. The mean of their score is 85 and the variance of the score is zero. What can you interpret about the score obtained by all learners?

2. In a residential locality, the mean size of the house is 2224 square feet and the median value of the house is 1500 square feet. What can you interpret about the skewness in the distribution of house size?

   Are there more bigger or smaller houses in the residential locality?

3. The following table shows the mean and variance of the expenditure for two groups of people. You want to compare the variability in expenditure for both groups with respect to their mean. Which statistical measure would you use to evaluate the variability in expenditure? Please provide an explanation for your answer.



	Expenditure	Group I	Group II
0	Mean	$500,000	$40,000
1	Std. Dev.	$125,000	$10,000

.

4. During the survey, the ages of 80 patients infected by COVID and admitted to one of the city hospitals were recorded and the collected data is represented in the less than cumulative frequency distribution table.

Age (in yrs.)	No. of Patients
5 - 15	6
15 - 25	11
25 - 35	21
35 - 45	23
45 - 55	14
55 - 65	5

   a. Which class interval has the highest frequency?
   b. Which age was affected the least?
   c. How many patients of age 45 years and above were admitted?
   d. Which is the modal class interval in the above dataset
   e. What is the median class interval of age?

5. Assume you are the trader and you have invested over the years, and you are worried about the average return on investment. What average method would you use to compute the average return for the data given below?

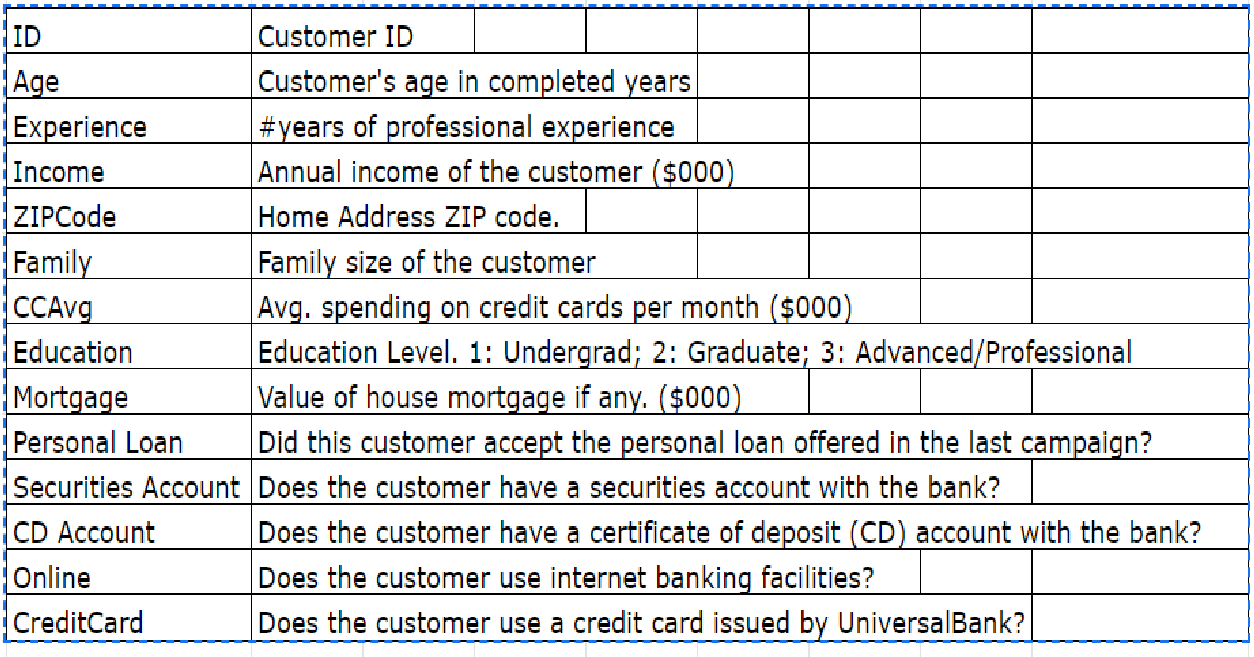
Year	Return	Asset Price
2015	36%	5000
2016	23%	6400
2017	-48%	7890
2018	-30%	9023
2019	15%	4567
2020	31%	3890

6. Suppose you have been told to measure the average height of all the males on the earth. What would be your strategy for the same? Would the average height be a parameter or a statistic? Justify your answer.

7. Calculate the z score of the following numbers:
   X = [4.5,6.2,7.3,9.1,10.4,11]

Section 2

You are expected to perform statistical analysis for the Bank Personal Loan Modelling dataset. Below is the data dictionary. For questions, 8 to 20 use the Bank Personal Loan Modelling dataset and answer the given questions.



ID	Customer ID
Age	Customer's age in completed years
Experience	#years of professional experience
Income	Annual income of the customer ($000)
ZIPCode	Home Address ZIP code.
Family	Family size of the customer
CCAvg	Avg. spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?

8. Give us the statistical summary for all the variables in the dataset.
9. Evaluate the measures of central tendency and measures of dispersion for all the quantitative variables in the dataset.
10. What statistical method will you use to examine the presence of a linear relationship between age and experience variables? Also, create a plot to illustrate this relationship.
11. What is the most frequent family size observed in this dataset?
12. What is the percentage of variation you can observe in the 'Income' variable?
13. The 'Mortgage' variable has a lot of zeroes. Impute with some business logical value that you feel fit for the data.
14. Plot a density curve of the CCAvg variable for the customers who possess credit cards and write an interpretation about its distribution.
15. Do you see any outliers in the dataset? If yes, what plot you would think will be suitable to showcase to the stakeholders?
16. Give us the decile values of the variable 'Income' in the dataset.
17. Give the IQR of all the variables which are quantitative and continuous.
18. Do the higher-income holders spend more on credit cards?
19. How many customers use online banking? Do customers using bank internet facilities have higher income?
20. Using the z-score of the income variable, find out the number of observations outside the +-3σ.

Deliverables for Solution and Grading Rubric
Required deliverables

- A Jupyter notebook detailing the steps.
- For questions 1 to 6, you can include your explanation and solution in the Jupyter Notebook.
- For questions 7 to 20, you would need the Jupyter Notebook to derive the solution.

Student-facing and faculty rubrics

- Total of points 100.

- 5 points per question.
- In question 4, each sub-question will carry 1 point.