Experiment No. 1

Aim: Exploring variable in a dataset

Objectives:

1. Exploring Variables in a Dataset
2. Learn how to open and examine a dataset.
3. Practice classifying variables by their type: quantitative or categorical.
4. Learn how to handle categorical variables whose values are numerically coded.

Link to experiment: <https://upscfever.com/upsc-fever/en/data/en-exercises-1.html>

Submission:

1. Create github repo and Upload results to repo and paste repo link in common excel sheet
2. Link to excel- <https://docs.google.com/spreadsheets/d/1G0HMQkWOuAb6mk89BRauMMhhjga0OaXjTbUC-IwSu8M/edit?usp=sharing>

Submission dates

1. J1: 9th July 2020 by 9AM IST
2. J2: 13th July 2020 by 9AM IST
3. J3: 10th July 2020 by 9AM IST

Note:

1. Students can choose any programming language

Questions:

1. What are the categorical variables in this dataset?
2. What are the quantitative variables in this dataset?
3. Describe the distribution of the variable "friends" in dataset - Survey that asked 1,200 U.S. college students about their body perception n
4. Describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, center, spread and outliers (Dataset - Best Actor Oscar winners (1970-2013))
5. Getting information from the output: a. How many observations are in this data set? b. What is the mean age of the actors who won the Oscar? c. What is the five-number summary of the distribution? (Dataset - Best Actor Oscar winners (1970-2013))
6. Get information from the five-number summary: a. Half of the actors won the Oscar before what age? b. What is the range covered by all the actors' ages? c. What is the range covered by the middle 50% of the ages? (Dataset - Best Actor Oscar winners (1970-2013))
7. What are the standard deviations of the three rating distributions? Was your intuition correct? (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)
8. Assume that the average rating in each of the three classes is 5 (which should be visually reasonably clear from the histograms), and recall the interpretation of the SD as a "typical" or "average" distance between the data points and their mean. Judging from the table and the histograms, which class would have the largest standard deviation, and which one would have the smallest standard deviation? Explain your reasoning (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)

Ans 1-The categorical variables are 1) Hostp because the numbers represent codes, which are used to identify individual hospitals and place them into categories.As such, the numbers used for the codes (1, 2, 3, 5, and 6) have no arithmetic meaning; 2) Treat because the treatment received by the patients is in the form of categories (Lithium, Imipramine, or Placebo); 3) Outcome since recurrence is in the form of two categories (Recurrence or No Recurrence) and 4) Gender because the numbers represent two distinct categories: Female and Male.Thus, the numbers used to represent gender (1 = Female; 2 = Male) have no arithmetic meaning.

Ans 2-The quantitative variables are 1) Time since it can take on multiple numerical values, which have arithmetic meaning (i.e., it makes sense to add, subtract, multiply, divide, or compare the magnitude of such values).2) Age since it can take on multiple numerical values, which represent a characteristic of the patient; and 3) AcuteT because it can take on multiple numerical values to represent a characteristic of the patient.

Ans 3- The students are NOT divided equally among the three categories. About 50% of the students find it as easy to make friends with the opposite sex as with the same sex. Among the remaining 50% of the students, the majority (36.2%) find it easier to make friends with people of the opposite sex, and the remainder (13.7%) find it easier to make friends with people of their own sex.

Ans 4-Shape: the distribution is skewed left. This means that most actors receive the best acting Oscar at a relatively younger age (before age 48), and fewer at an older age.

Center: The distribution seems to be centered at around 42-43. This means that about half the actors are 42 or younger when they receive the Oscar, and about half are older.

Spread: The age distribution ranges from about 30 to about 75. The entire dataset is covered, then, by a range of 45 years. It should be noted, though, that there is one high outlier at around age 75, and the rest of the data ranges only from 30 to 60.

Outliers: As mentioned above, there is one high outlier at around age 75.

Ans 5 -a. There are n = 44 observations in the data set (representing the age of the Best Actor Oscar winners of the 44 years from 1970 through 2013).

b. Mean = 44.98

c. The five-number summary is: min = 29, Q1 = 38, M = 43.5, Q3 = 50.5, Max = 76

Ans 6-a. Half the actors won the Oscar before age 43.5 (the median). b. The range covered by all the ages is: Range = Max - min = 76 - 29 = 47. c. The range covered by the middle 50% of the ages is: IQR = Q3 - Q1 = 50.5 - 38 = 12.5.

Ans 7 -Here are the three standard deviations :Class I: 1.6Class II: 4.0Class III: 2.6Note that through this example, we also learn that the number of distinct values represented in a histogram does not necessarily indicate greater variability.

Ans 8 - In class I, almost all the ratings are 5, which is also the mean. The average distance between the observations and the mean, then, would be very small. In class II most of the observations are far from the mean (at 1 or 9). The average distance between the observations and the mean in this case would be larger. Class III is the case where some of the observations are close to the mean, and some are far, so the average distance between the observations and the mean would be somewhere in between class I and II. This observation would lead me to conclude that the standard deviation would be ranked (from smallest to largest): Class I, Class III, Class II.

Name – Anoop Mishra

Roll No- J063

BTech Data Science

3rd Year