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.

Questions:

1. What are the categorical variables in this dataset?

The dataset used is depression. The categorical variables are as follows:’

* + Gender
  + Treat
  + Hospt
  + Outcome

1. What are the quantitative variables in this dataset?

The dataset used is depression. The quantitative variables are as follows:

* + Time
  + AcuteT
  + Age

1. Describe the distribution of the variable "friends" in dataset - Survey that asked 1,200 U.S. college students about their body perception

friends <- read.csv("/Users/nihaltodankar/Desktop/friends.csv")

friends

View(friends)

friends$Friends

t = table(friends$Friends)

t

prop = prop.table(t)

prop

p

percent = prop\*100

percent

* + From the above code it seen that 50.1667% of students find it same to make friends with same sex as well as the opposite sex
  + While 36.1667% find it easy becoming friends with the opposite sex
  + And 13.667 % find it easy becoming friends with the same sex.

1. Describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, centre, spread and outliers (Dataset - Best Actor Oscar winners (1970-2013))
   * Shape : Right skewed meaning the actors recived an award at a younger age
   * Centre : It is centered around 40-45
   * Spread : Entire age group i.e. from 30to75 is covered.
   * Outlier : There is an outlier around the age of 76
2. 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))
   * 1. There are 44 observations
     2. Mean age = 44.97727
     3. min = 29

lower quartile = 38.0

median = 43.5

upper quartile = 50.25

max = 76

1. 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))
   * 1. Before the age of 43
     2. The range of age is 76 - 29 = 47
     3. This range is = 50.25 – 8.0 = 12.25
2. 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)
   * For Class.I we have std = 1.56892
   * For Class.II we have std = 4.00
   * For Class III we have std = 2.63117
3. 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)
   * Class.II has the largest standard deviation while Class.III has the smallest standard deviation