[2:15 PM, 7/6/2020] Soham Data Science Nmims: 2:13 PM

https://upscfever.com/upsc-fever/en/data/en-exercises-1.html

2:14 PM

Experiment No. 1

Aim: Exploring variable in a dataset

Objectives:

Exploring Variables in a Dataset

Learn how to open and examine a dataset.

Practice classifying variables by their type: quantitative or categorical.

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:

Create github repo and Upload results to repo and paste repo link in common excel sheet

Link to excel- https://docs.google.com/spreadsheets/d/1G0HMQkWOuAb6mk89BRauMMhhjga0OaXjTbUC-IwSu8M/edit?usp=sharing

Submission dates

J1: 9th July 2020 by 9AM IST

J2: 13th July 2020 by 9AM IST

J3: 10th July 2020 by 9AM IST

Note:

Students can choose any programming language

Questions:

What are the categorical variables in this dataset?

What are the quantitative variables in this dataset?

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

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))

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))

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))

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)

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)

[2:16 PM, 7/6/2020] Soham Data Science Nmims: Q1,2 - depression

Q3- friends

Q4-actor\_age.csv

Q5,6-actors\_age.csv

Q7 - grad\_data.csv

Q8-ratings.csv