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:

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
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)

EXPERIMENT 2

Aim: Python for data science
Objectives:

- Exploring data structures of python (list, tuple, dictionary, sets, array, dataframes)
- Exploring numpy and pandas libraries of python

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: 16th July 2020 by 9AM IST
- J2: 20th July 2020 by 9AM IST
- J3: 17th July 2020 by 9AM IST

Note:

- Students can choose python language

Solve the notebooks:

1. https://colab.research.google.com/drive/1sfhFtGOFWxukueoARioqVAicrFQpq3Z0
2. https://colab.research.google.com/drive/1v7o1c4NVO4NC8DgTvSkjou6XgFXAAL_J
3. https://colab.research.google.com/drive/1L4ccgm5nbTEMZh0QXAiYWZ0A5A62v9to
4. https://colab.research.google.com/drive/1dXXtQChNVkHIMG9043Px2Nj38yEjsDtB
5. https://colab.research.google.com/drive/1pPF7_IfcLk_Qb8oMLR-xuCH0SyGZARHX
6. https://colab.research.google.com/drive/177lHw7dLlRS-OPmw55VesOjCvuMshWw4
7. https://colab.research.google.com/drive/1SJrx0rjQS0NRMqBubjGjlZs4zpQNXrTh
8. https://colab.research.google.com/drive/1ajQEDmF-5eHpwOEhB4C0hJVx33DHEq2m
9. https://colab.research.google.com/drive/1P1JXqaMha7WEGoqjI4x3bEaMkGejfGWc
10. https://colab.research.google.com/drive/1Qd3gLh3IJODpp02e8U992Qvonhu6b91R
11. https://colab.research.google.com/drive/1C2Nn6bDJXxyGSW55lSJES9COrP5m2DFd
12. https://colab.research.google.com/drive/1bpycHoFbcJ21hg4ajlGGYfVe-GIgb1-Z
13. https://colab.research.google.com/drive/1RsvNZzK_wtWrtPw6DQs01VFcW9FlVCWp
14. https://colab.research.google.com/drive/1kQihatHNEvzISDUiPa5Pvvf1trAjPM9s
15. https://colab.research.google.com/drive/1WgQ54R5hYVq3cLxwRpz3lP2bl95nDBX2
16. https://colab.research.google.com/drive/1PvHPn6wBgD_ukJP8OYs-UakxmIBbvPNz
17. https://colab.research.google.com/drive/1UuRYs9QNQD9-LS4Q5-YUHMc1aQwl2Omi

Sample:
https://colab.research.google.com/drive/1T0Un0uIaQpjT4yRdJOzyiKFS0yN89V9V

Exercise 1

a. Introduction and Background.
    a. Import dataset store it as dataframe in python: `filename = "https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/auto.csv"`
    b. Add column headers to dataset
    c. Read the first 5 lines and display
    d. Are there missing data? Or illegal characters in the dataframe?
    e. Replace the character “?” with nan
    f. Count missing values in each column and print it with column name
        i. "normalized-losses": 41 missing data
        ii. "num-of-doors": 2 missing data
        iii. "bore": 4 missing data
        iv. "stroke" : 4 missing data
        v. "horsepower": 2 missing data
        vi. "peak-rpm": 2 missing data
        vii. "price": 4 missing data (Response)
    g. Delete price rows that have missing data
    h. Normalized losses,bore,stroke,horsepower,peak-rpm, - replace missing with mean of the column
    i. Num-of-doors replace missing with most frequent value in the column
    j. Reset the index of dataframe
    k. Check datatype of columns and convert numeric/quantitative variables to float or int
    l. Transform city-mpg and highway-mpg into liters/100km using conversion formula: L/100km = 235/mpg i.e. create two new column “city-L/100km” and “highway-L/100km”
    m. Normalize columns length, width, height so that their values range from 0 to 1. Hint: Replace original values with original_value/max_value
    m. Plot the histogram of horsepower to see its distribution
    n. Create three equal sized bins “low”, “medium”, “high” and organize values in column horsepower into new column “horsepower-binned”
    o. Plot distribution of “horsepower-binned”
    p. Convert “fuel-type” into one-hot-encoded variables. Repeat same for “aspiration” and then drop columns “fuel-type” and “aspiration”

Exercise -2: Exploring variables

1. Import the dataset:
   `path='https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/automobileEDA.csv'`

2. Import matplotlib, seaborn, numpy and pandas
3. See dimensions of data frame and its data types for each column
4. Calculate correlation between engine-size and price using corr function
5. Identify variables with positive or negative correlation with price
6. Identify datatype of “peak-rpm”
7. Using seaborn regplot() - plot relation between “engine-size” and “price”. Comment on your observation.
8. Identify using regplot() - which other variables can affect “price” and which do not affect it.
9. Use seaborn pairplot() to identify which variables can affect “price”
10. Draw a heatmap to plot the correlation in the dataframe
11. With seaborn boxplot() - compare “body-style” with “price”
12. Continue for other categorical variables in the dataset.
13. What do you infer from the boxplots about the relationship between the variables.
14. Use describe() to get descriptive statistics of numeric variables
15. Use describe() to get stats of categorical variables
16. Get unique values in each categorical variable along with their frequency. What do you understand by doing this?
17. Use groupby() to get the average price of “drive-wheels” wrt “price”. What do you understand by doing this?
18. Repeat step 17 for other categorical variables.
19. Use groupby() to find the average price for “drive-wheels”and”body-style” with price. Observation? Inference?
20. Use pivot() on the result of step 19 to get “drive-wheels” as index and “body-style” as columns. Observations? Inference?
21. Repeat step 19and20 for other combinations of independent variables wrt price. Observations? Inferences?
22. Draw heatmap for result of step20
23. Calculate the pearson correlation between “wheel-base” and “price”. What can you conclude from p-value (Hint: use stats from scipy which has pearsonr())
24. Perform one way ANOVA test using f_oneway() of stats to check if different groups of “drive-wheels” are correlated with “price”. What do you understand from F-test and p-value results?

END