Exercises 1: Background –

https://colab.research.google.com/drive/1gM8vwiXAt4Ao-H-LlOIazB3z2Yqqf4gz

1. Introduction and Background.
2. Import dataset store it as dataframe in python: filename = "https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DA0101EN/auto.csv"
3. Add column headers to dataset
4. Read the first 5 lines and display
5. Are there missing data? Or illegal characters in the dataframe?
6. Replace the character “?” with nan
7. Count missing values in each column and print it with column name
8. "normalized-losses": 41 missing data
9. "num-of-doors": 2 missing data
10. "bore": 4 missing data
11. "stroke" : 4 missing data
12. "horsepower": 2 missing data
13. "peak-rpm": 2 missing data
14. "price": 4 missing data (Response)
15. Delete price rows that have missing data
16. Normalized losses,bore,stroke,horsepower,peak-rpm, - replace missing with mean of the column
17. Num-of-doors replace missing with most frequent value in the column
18. Reset the index of dataframe
19. Check datatype of columns and convert numeric/quantitative variables to float or int
20. 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.         [Text Wrapping Break]                        Hint:  Replace original values with original\_value/max\_value

1. Plot the histogram of horsepower to see its distribution
2. Create three equal sized bins “low”, “medium”, “high” and organize values in column horsepower into new column “horsepower-binned”
3. Plot distribution of “horsepower-binned”
4. Convert “fuel-type” into one-hot-encoded variables. Repeat same for “aspiration” and then drop columns “fuel-type” and “aspiration”

END - 10 mins

Exercise -2: Exploring variables

<https://colab.research.google.com/drive/1gtU7GZEuUspZMHdhuh7rO7iWxm7YZAcU>

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?