- **EDA Lab1:-**

- Let us assume df repesents a dataframe.
- df.dtypes shows us the data types.
- If we want to change a datatype in the data set column, then we can use the function. For example:-

df-adult['sex'] = df-adult['sex'].astype('object')

- Quantile function is used to check the values under a specified Quantile.
- Value_Counts() function counts the types of values in each column or attribute.
- Z Score is used for standard deviation normalization.
- threshold = 3 means, after 3 observations, we will take any observation as an outlier.
- abs means absolute value.
- Proportiontocut:- It cuts a part or fragment of data from both the ends (both the tails).

proportionto cut = 0.01 means it will cut 1% of the observations from both the tails.

- It is a good sign when we are trimming the mean.
- Most of the observations in a standard Normal Distribution fall between -3σ to $+3\sigma$. It is a good practice to take threshold as 3.
- .Var() function gives the variability.
- If the standard deviation is near to zero, then there is no variation in the range of data.
- Coefficient of variance $= \dfrac{\text{standard Deviation}}{\text{np.abs(mu)}}$

  where mu is mean μ.

- what is the range of pearson's correlation coefficient.

→ -1 to +1 (-1 is strong negatively correlated and +1 is strong positively correlated).

- who is considered as the father of EDA :-

→ John Tukey.

- The difference between box plot and violin plot:-

→ In violin plot we can also see the distribution of the data. but in the box ~~plo~~ plot we can only see the quantile (or) the five point summary.

- What is the diffeunce between Binomial and Bernoulli Distubution?

→ Bernoulli is the limiting case of binomial distribution.
In Bernoulli distribution we only have one trial but in binomial distribution we have multiple (or) n number of trials.

EDA Lab 2 :-

- We have normalization and ~~standardization~~ tranformation techniques in EDA.
- Sometimes data doesn't follow a normal distribution, there tranfomation techniques can help us make the data follow normal distribution.

- Extreme observation means an outlier.

- hist_kws = dict (Cummulative = True)

  ~~These~~ This is histogram argument, it says hist_kws argument will show the histograms (The histograms which are cummutatively added).

- KDE stands for Kernel Density Estimation. It gives us the rough idea on what is the trend present in the histogram.
- Variation is a term that is synonymous with the standard deviation. std is the measure of the dispersion of data. Dispersion means how much varied the data is.

- Does standard Deviation have a unit?

→ It has a unit (the unit of what std we are going to ~~cat~~ find.), (the unit of the feature we are analyzing).

- If std is not unitless, is there a measure of dispersion which is unitless.

→ Coefficient of variation is a unitless quantity.

$$\boxed{\text{Coefficient of variation} = \frac{\sigma}{\mu}}$$.

- Statisticians like to say dummy variable and ML people like to say one hot encoded variables.

- get-dummies is a sparse matrix, means there are a lot of zeros and very less ones.
- standard normal distribution has mean = 0 and variance = 1.
- StandardScaler will do that above one.
- the two general techniques are standard scaler and the min-max scaler.
- Inverse of logarithmic transformation is exponential transformation.
- ~~the exp~~

- A given feature is negatively skewed, what can be done to transform it into a symmetrical distribution?

→ Use the logarithmic transformation

- Why do we use cross tabulation?

→ It is a form of creating tables from extracted entries from the data. Cross tabular means we will have values at the rows and the columns also. We take some features and put it into the rows and we take some features and put it into the rows.

- Can we have a "None" as a key in the Dictionary.

→ Yes, we can.

- What is the appropriate data type for Date time column or a feature that contains time stamps.

→ datetime, pd.-- datetime

- which plot can display the line of trend between two variables scattered in a plot.

→ IM Plot (or) LM plot.

- How can we set the universal plot size and dimension for a notebook.

→ rc params (params is a parameter which is in matplotlib).

- How can we only display upper triangle in the heat map.

→ we can do zero masking (use mask function).

- How to create a random sparse matrix in python?

→ we can use the random function. In sparse matrix it will have only random zeros and ones.

- Suppose we implemented Label Encoder for some feature. How can we get the original categorical labels back.

→ fit_transform for transformation.

inverse_transform

- KDE → Kernel Density Estimation.