Sunday, August 29, 2021 7:54 PM

Exploratory Data Analysis -

- 1) Device - Laptop
- 2) Target Feature - Classification (conversion)
    - a. Count plot - **Yes/No**
    - b. Value counts - check the distribution - skewed or not? - ****Insights**** **(we will tackle in the modelling Notes-2)**

    - a. Pair plot
    - b. Separate the categorical and numerical features
    - c. Categorical -
        - i. Univariate Analysis -
            - 1) Jaipur/Jaipure - hygiene checks in the data
                - a) Check the categories of the columns - more than expected categories like - adult (convert it into two categories)
                - b) Suppose city - 60% - Delhi, 1% Ahmedabad - try to merge for later purpose
            - 2) **Missing values -** Treat them with a method - Mode/Max freq/KNN imputer from sklearn
            - 3) Check that - "?"/special characters - value counts on each of the categorical - if you can run a loop
            - 4) Create a few count plots to show freq - run a loop to get all the plots in 1 go - ****Insights****
        - ii. Bi-variate Analysis -
            - 1) ~~Categorical to categorical (X1 v/s X2) - stack bar plot~~
            - 2) Categorical to numerical (X1_cat v/s X2_num) - bar plot/swarm/violin/bar - ****Insights****
            - 3) Categorical to Target Feature (X1_cat v/s Target_conversion) - stackbar - ****Insights****

    - d. Numerical -
        - i. Univariate Analysis -
            - 1) Hygiene checks on the data
            - 2) Missing values - Mean/Median/KNN imputer/simple imputer
            - 3) Distribution and box plots with a loop - ****Insights****
            - 4) Outliers - boxplot - IQR method/**percentile method (99%,95%)**
            - 5) Distribution and box plots with a loop - verify the outliers are removed - ****Insights****
            - 6) Skewness in the data - right skewed - ~~take a log else take a squareroot~~
        - ii. Bi-variate Analysis -
            - **1) Correlation -**
                - a) Correlation between (X1_num v/s X2_num) - heatmap - ****Insights****
                - b) Scatter plots (X1_num v/s X2_num) - regplot - ****Insights****
            - 2) Relation with target feature (X1_num v/s Target) - BOX/Swarm/violin - ****Insights****
            - 3) Relation with Categorical feature (X1_num v/s X1_cat) - BOX/Swarm/violin - ****Insights****
        - iii. Try to see the separation between the - creation the distribution plot with a hue of target - **Pair plot**

- 3) Device - Mobile
    - a. Pair plot
    - b. Separate the categorical and numerical features
    - c. Categorical -
        - i. Univariate Analysis -
            - 1) Jaipur/Jaipure - hygiene checks in the data
            - 2) **Missing values -** Treat them with a method - Mode/Max freq/KNN imputer from sklearn
            - 3) Check that - "?"/special characters - value counts on each of the categorical - if you can run a loop
            - 4) Create a few count plots to show freq - run a loop to get all the plots in 1 go - ****Insights****
        - ii. Bi-variate Analysis -
            - 1) ~~Categorical to categorical (X1 v/s X2) - stack bar plot~~
            - 2) Categorical to numerical (X1_cat v/s X2_num) - bar plot/swarm/violin/bar - ****Insights****
            - 3) Categorical to Target Feature (X1_cat v/s Target_conversion) - stackbar - ****Insights****

    - d. Numerical -
        - i. Univariate Analysis -
            - 1) Hygiene checks on the data
            - 2) Missing values - Mean/Median/KNN imputer/simple imputer
            - 3) Distribution and box plots with a loop - ****Insights****
            - 4) Outliers - boxplot - IQR method/**percentile method (99%,95%)**
            - 5) Distribution and box plots with a loop - verify the outliers are removed - ****Insights****

6) Skewness in the data - rigth skewed - ~~take a log else take a squareroot~~

ii. Bi-variate Analysis -

1) Correlation -

a) Correlation between (X1_num v/s X2_num) - heatmap - ****Insights****

b) Scatter plots (X1_num v/s X2_num) - regplot - ****Insights****

2) Relation with target feature (X1_num v/s Target) - BOX/Swarm/violin - ****Insights****

3) Relation with Categorical feature (X1_num v/s X1_cat) - BOX/Swarm/violin - ****Insights****

iii. Try to see the separation between the - creation the distribution plot with a hue of target - **Pair plot**

Optional - Github update
do it pythonic way - try to use as many functions and loops as possible