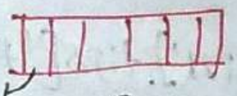
EDA

Univariate Analysis



Single column → If you calculate (Analysis) on Single column it's call ~~Single variate~~ univariate Analysis.

If Analysis on two columns → Biv Variate Analysis

If more then two columns → Multy Variate Analysis.

Mostly focuse on: which type of Data in

Data → Numerical Data ↓ height, weight, marks,

Data → categorical data: country, colure, Sex, city.

Step (I) ~~focuse~~ work on Categorical Data.

(i) countplot, PieChart

(ii) df['Pclass'].value_counts().plot(~~kide~~ kind = 'pie', autopct = "%.2f")

Step (II) Numerical Data:

(a) → Histogram → Histogram is a tools for checking the Range of the data

Eg ⇒ 10 – 20 Age → 15 peoples
20 – 30 Age → 40 P.S
30 – 40 Age → 30 peoples.
40 – 50 Age → 20 people.
↳ Range (Bins)

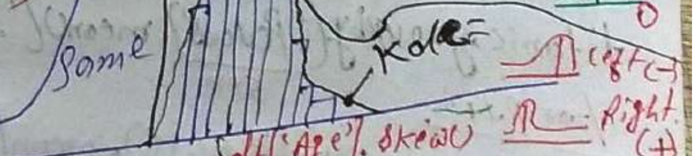
matplotlib

plt.hist(df['Age'])

You can know about Skewness from this plot.

(b) Distplot (seaborn)

sns.distplot(df['Age'])



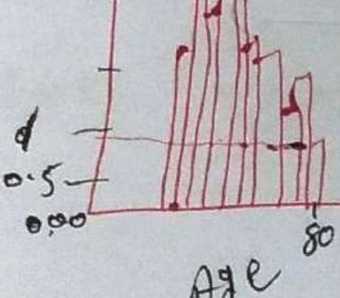
df['Age'].skew()

Kde → kernal density estimation.

Es plot se hum probability k bare me pta chlta hai

Q ⇒ Age 80 hone ka probabidity kya hai

Ans ⇒ 1 % chanse hai Age 80. hone ka.



(c) Boxplot

sns.boxplot(df['Age'])

~~EAD~~ EDA Bivariate

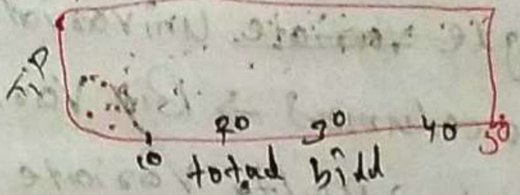
Work on 2 columns

(1) Scatter plot (Numerical - Numerical Data)

check Relation

(tips dataset)

sns.scatterplot(tips['total-bill'].tips['tips']), hue=df['sex']



Jais-jaise bill badh rha hai waise-waise tip v badh rha hai,

Multivariate

sns.satterplot(tips['total-bill'].tips['tip'], hue=df['sex'], style=df['smoker'], size=df['size'])
(1) (1) (3) (4) (5)

there are 5 column in single scatter plot

we can only do this with Scatter plot.

(2) Bar plot (Numerical - Categorical data)

(i) (ii)

Sns.barplot(titanic['Pclass'], titanic['Age'], hue=titanic['Sex'])

(3) Boxplot. (Numerical - categorical)

Sns.boxplot(titanic['Sex'], titanic['Age'], hue=titanic['Survived'])

(4) Distplot (Numerical - categorical)

sns.distplot(titanic[titanic['Surrived']==0]['Age'], hist=False)

sns. " ——— " ——— " 1 ——— " ——— ")

(5) Heatmap (Categorical - Categorical)

sns.heatmap(pd.crosstab(titanic['PClass'], titanic['Survived']))

for know the percentage

titanic.groupby('Pclass').mean()['Survived']*100 → p% show
sex

for plot

(titanic.groupby('PClass').mean['Survived']*100).plot(kind='bar')
sex

(6) ClusterMap (Categorical - Categorical)

pd.crosstab(titanic['Sibsp'], titanic['Surrived'])

for plot sns.clustermap(——— " ———)

⑦ Pair plot

sns. pairplot(iris, hue ='spacies')

↳ it will self decide and plot with same column, (Cate → Cate
name → nume)

⑧ Lineplot (Numerical - Numerical)

↳ it is basicly use for trend date and time ~~data~~ dataset

[Share market graph]

new = flights.groupby('year').sum().reset_index()

sns. line plot(new['year'], new['passengers']

for pivot table

flight.pivot_table(values='passengers', index='month', columns='year')

sns.heatmap(________ " ________)

sns.clustermap(________ " ________)