DATA VISUALISATION

Matplotlib Library

SeaBorn

can't create Pie plot.

- ~~Matplot Lib~~

- **Matplot lib** : Matplotlib is a library used for Data visualisation for creating highly customized plot.

- **SeaBorn** : SeaBorn is a python library built on top of matplotlib and provides additional function-ality, making certain task easier.

★ SeaBorn is easy to use than Matplotlib.

★ We will use SeaBorn for Data Visualis-ation

SOME BASICS OF MATPLOTLIB YOU NEED TO KNOW

(I) import matplotlib.pyplot as plt.

(II) Change figure size

```
plt.figure(figsize=(8,5))
```

width → 8, height → 5

(III) Create subplots.

```
fig, ax = plt.subplot(1, 2, figsize=(12,5))
```

1 row, 2 column

```
ax[0].set_title("First Plot")
ax[1].set_title("2nd Plot")
```

(IV) Set Title & axis of Plot

```
plt.title("My Plot")
plt.xlabel("X-axis")
plt.ylabel("Y-Axis")
```

(V) Show Plot or Print Plot.

plt.show()

(VI) save Plot.

plt.savefig("myplot.png", dpi=300)

INTEGRATED WITH SEABORN

GENERAL

(I) SET FIGURE SIZE
(II) SNS => Barchart/Boxplot/...
(III) Title
(IV) x-label
(V) y label
(VI) plt.grid (True or False)
(VII) plt.show()

SEABORN (SNS)

Types of Plots in Seaborn

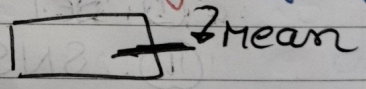
(I) Lineplot : Simplest one used to analyze trend

(II) Barplot : Only can be plotted where there distinct categories like Beer, Wine, Shorts.

count no. of entries in each category : used to show performance of Different Group

★ Calculate ⇒ Mean (avg) by default

(III) Histogram :



Histogram is a type of plot that shows distribution of numerical value variable by dividing data into intervals called bins & counting no. of data points in bins.

Histogram	Barplot
(I) Numerical Data	(I) Categorical data
(II) Show the distribution of Data	(II) Compare the value across categories

(III) Preq
with

(IV) Sc

(V) positively related y =

negatively related y = -x

(V) HEA
wh
m

- Use
- Rou

(III) Freq. of data within intervals | (III) Specific or aggreg-ated matric

(IV) **ScatterPlot**: used to visualize relationship b/w two numerical values

(a) **Positive Correlation**: Dots move upward from left to right.

positively related $y = x$

Negative Correlation: Dots moves downward from left to right.

negatively related $y = -x$

No Correl-ation: Dots are scattered randomly.

(V) **HEATMAP Plot**: Create 2D matrix where the intensity of color represents magnitude of value

- Used to find correlation in numerical variables in dataset.
- Rows & cols represent categories

↳ Simply count the number of ~~occurence~~ occurance

(VI) Count Plot: frequency distribution of categorical data.
↳ count the no. of entries in each category.

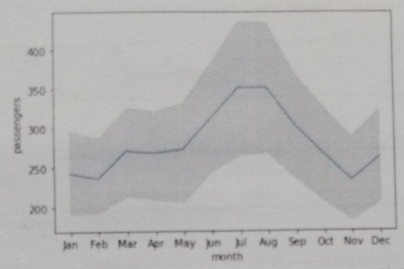
(VII) Voith Violin Plot:

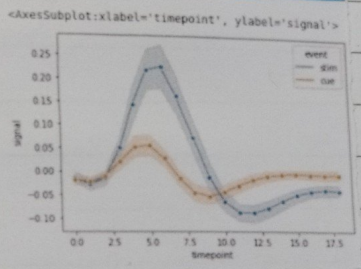
(VIII) BoxPlot: way of Displaying Dist-ribution of dataset based on 5 Things:

i) Minimum: Smallest datapoint (excluding outliers)
ii) First Quartile (Q_1): 25th 1·ile of data
iii) Median (Q_2): Middle value
iv) Third Quartile (Q_3): 75th 1·ile of data
v) Maximum: Largest Datapoint (excluding outliers

★ Used to identify outliers

Line Plot:

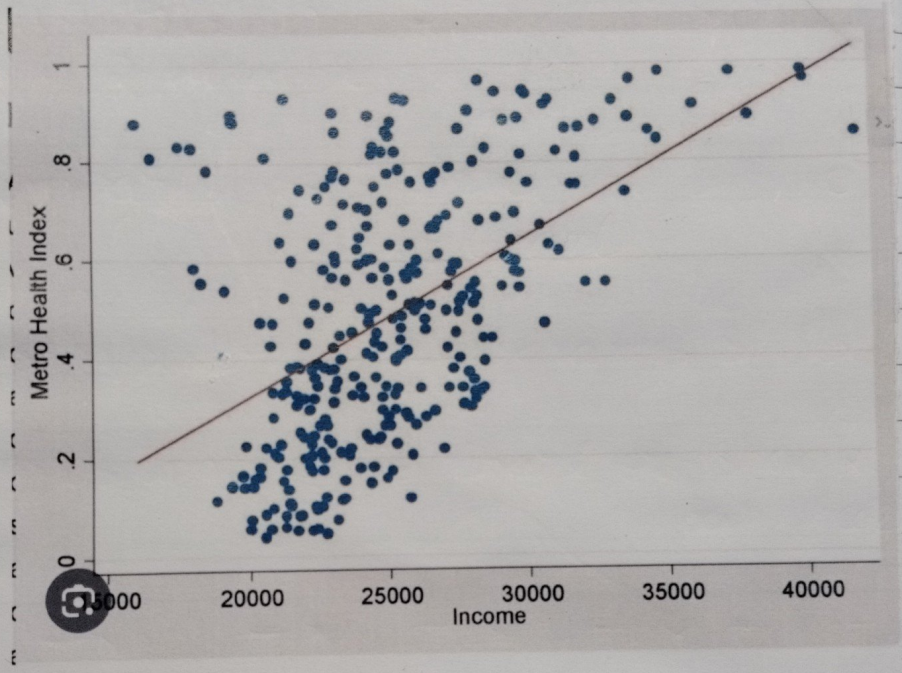




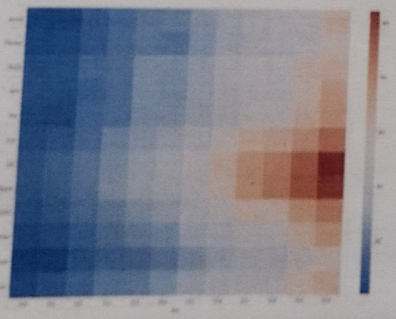
(excluding outliers)

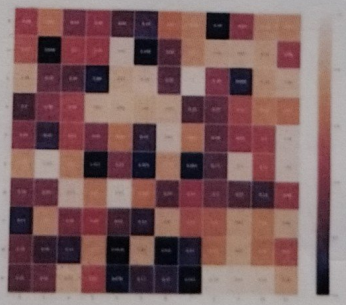
Bar Plot:

SCATTER PLOT

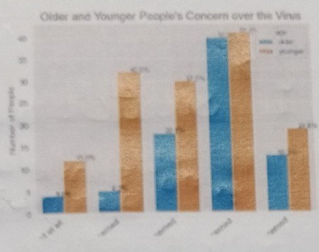


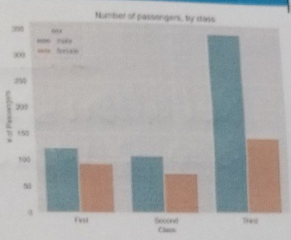
Heatmap Plot:



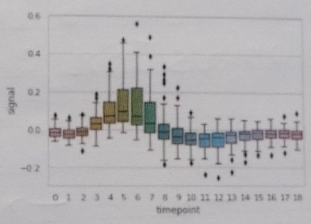


Count Plot:

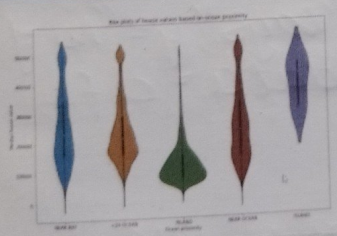


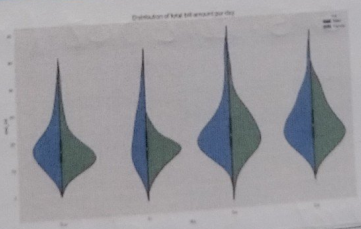


Box Plot:



Violin Plot:





- Relationship Blw 2 variable : Scatterplot
  not same as Bar plot
- Comparisons : Barchart & Line Chart
- Distribution : Histograms & Box Plot
- Composition : Pie Chart or Stacked Bar Chart

— x — x — x — x — x — x — x

① LINE PLOT

sns.lineplot(x="Days", y="No of People", data=df)

(IV)

Q What is the function of hue

→ hue allows you to group data by a specific category & visually distinguish using different colors.

→ sns.linep

Q What

↳ Pa

sche

→ sns.lineplo

(II) BAR

sns.ba

Q How

↳ Ju

as

→ sns.ba

→ sns.lineplot(x='x', y='y', hue="sex")
↳ it will group by sex.

Q What is Pallete?
↳ Pallete is used to set the color scheme of a plot.

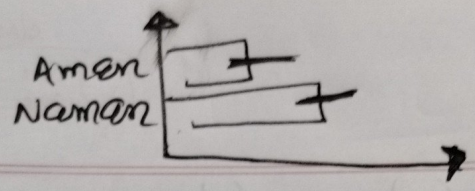
→ sns.lineplot(x='x', y='y', hue="sex", pallete="flare")

⑪ BAR PLOT

sns.barplot(x='who', y='fare', data=df)

Q How to Draw Horizontal Bar Plot.
↳ Just make categorical variable as y-axis.

→ sns.barplot(x='fare', y='who', data=df)

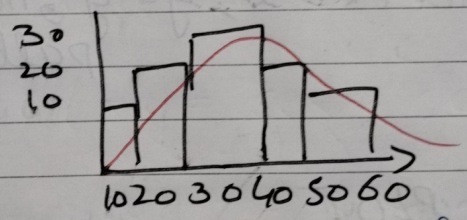


(III) HISTOGRAM (need only x)

→ sns.histplot (x="total_bill", data=df

★ hue, pallatte can be used everywhere (every plot)

→ Q how to get the density line



→ sns.histplot (x="total_bill", data=df, kde=True)

↓ Kernel Density estimation

(IV) Scatter Plot

→ sns.scatterplot (x="time", y="pulse", dataframe=df)

Q change the way of representing datapoints?

→ marker = "size" => Based on size
"s" => square
"*" => * represent

sns.scatterplot(x="time", y="pulse", data=df
, marker="s")

Q Wh You want to make datapoints transparent

0 < alpha < 1
↓
Transparent ↳ full visible

alpha = 0.5 => half Transparent (light sa hojayega datapoints)

⑦ HEATMAP

sns.heatmap(data="datas")
↓
should be a 2D matrix

[[1, 2, 3, 4, 5]
[9, 8, 7, 6, 5]
[⋮]
[⋮]]

color

→ sns.heatmap(data="dataset", cmap="Gnb
, alpha=0.5)

(VI) COUNT PLOT

→ sns.countplot (x="sex", data=df)

Q how to get countplot in horizontal.

→ sns.countplot (y="sex", data=df)

★ here also can use hue & palette

(VII) BOX PLOT

→ sns.boxplot (x=df["tips"], data=df)

Hmm....

→ sns.boxplot (x="sex", y="tip"
, data=df)

Q Plot Multiple lineplot using seaborn

```
x = [1, 2, 3, 4, 5]
y1 = [50, 40, 70, 80, 20]
y2 = [80, 20, 20, 50, 60]
y3 = [70, 20, 60, 40, 60]

sns.lineplot(x=x, y=y1, label='Enfield')
sns.lineplot(x=x, y=y2, label='Honda')
sns  ,,  (x=x, y=y3, label='Yamaha')
sns. ,,  (  ,,  , y=y4
```

Q Create Pie Plot or Pie Chart?

```
import matplotlib.pyplot as plt.
plt.pie(grade count, labels=grades,
        autopct='%1.1f%%')
plt.show()
```

column

V. Imp

EDA STEPS

2→ You need to make prediction

Steps

(I) Drop the column if there are many columns that you think is not important for your prediction

→ df = df.drop(['Engine fuel Type', 'vehicle style',], , axis = 1)

(II) To improve readibility, you can also rename columns.

→ df = df.rename(columns = {"Engine HP : HP", ... })

(III) Remove Duplicate rows

df.shape => (11914, 13)
df = df.drop_duplicates()
df.shape => (11139, 13)

(iv) Drop the missing or null values

i) print(df.isnull().sum())

↳ Print cols. with no. of null values.

Make	0
Model	0
HP	69
Cylinders	30

ii) df = df.dropna()

↳ drop all the rows with any of the col. with Null value.

(v) Detecting outliers.

• <u>Outlier</u>: Data point that deviate signif-icantly from the rest of datapoints.

sns.boxplot(~~df~~ x = df['HP'])

sns.boxplot(x = df['Price'])

(VI) Draw DIFFERENT PLOTS : (i) Histogram
variable & Prediction variable ← (ii) Scatterplot
(iii) Heat Map
↳ To find correlation

Heat Map

↳ c = df.corr(numeric_only=True)
sns.heatmap(c, annot=True)

☆ Univariate Analysis : Histogram, Boxplot, BarPlot Count Plot
☆ Bivariate Analysis : Scatterplot,
(Need Two Variables) Heatmap,
Boxplot
(x=x, y=y)