DSBDA Assignment 7

Details 1. Author : Aditya Muthal 2. Roll Number : 33249 3. Batch : M10 4. Class : TE10

Problem Statement

Visualize the data using Python libraries matplotlib, seaborn by plotting the graphs for assignment no. 2 and 3

Implementation details

1. Dataset URLs
    1. Facebook metrics : https://archive.ics.uci.edu/ml/datasets/Facebook+metrics
    2. Heart Disease : https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. Python version : 3.7.4
3. Imports :
    1. pandas
    2. numpy
    3. matplotlib
    4. seaborn

Dataset details

1. Facebook Metrics :
    1. Given dataset is a representative of some of the Facebook metrics which are assosciated with the posts on social media.
    2. These metrics are indicative of the engagement of the users with the corresponding post.
    3. It includes various types of posts and their details

2. Heart Disease Dataset :
    1. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.
    2. The "goal" field refers to the presence of heart disease in the patient.
    3. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

4. The names and social security numbers of the patients were recently removed from

Importing required libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
%matplotlib inline
```

A) Visualization for Facebook metrics dataset

1) Loading the dataset

```
facebook_dataset = pd.read_csv("./dataset_Facebook.csv", sep=";")
facebook_dataset.head()
```

	Page total likes	Type	Category	Post Month	Post Weekday	Post Hour	Paid	Lifetime Post Total Reach	Lifetime Post Total Impressions	Lif En
0	139441	Photo	2	12	4	3	0.0	2752	5091	
1	139441	Status	2	12	3	10	0.0	10460	19057	

2) Distribution of data based on type of Post

```
# Acquiring unique post values
post_types = facebook_dataset.Type.unique()
post_types
```

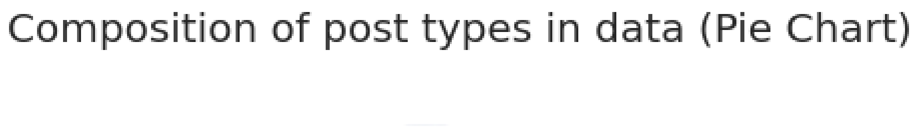
```
array(['Photo', 'Status', 'Link', 'Video'], dtype=object)
```

```
# Generating frequency data for each type of post

frequency_data = {}
```

```
for post in post_types:
    subset = facebook_dataset[facebook_dataset.Type == post]
    frequency_data[post] = subset.shape[0]

frequency_data
```

```
     {'Photo': 426, 'Status': 45, 'Link': 22, 'Video': 7}
```

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

# Adds subplot on position 1
ax = fig.add_subplot(111)

# Generating legend for pie chart
legend = [
    "Photo",
    "Status",
    "link",
    "Video"
]

# Defining explode values
explode = [0.1, 0.1, 0.1, 0.1]

# Generating and displaying piechart
plt.pie(
    x=frequency_data.values(),
    labels=legend,
    explode=explode,
)
plt.title("Composition of post types in data (Pie Chart)", fontsize=20)
plt.show()
```

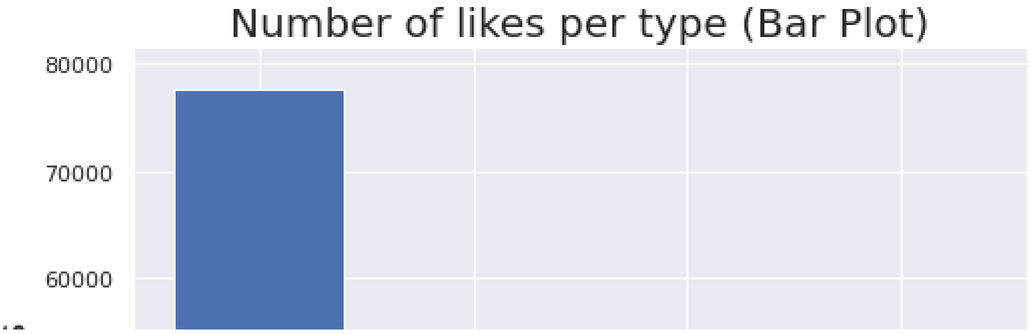


3) Likes per type of data

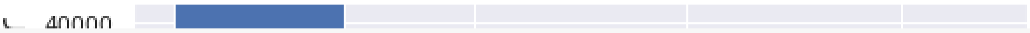
```
# Generating data for count of likes
likes_per_type = {}

for post in post_types:
    subset = facebook_dataset[facebook_dataset.Type == post]
    likes_per_type[post] = subset.like.sum()

likes_per_type
```

```
{'Photo': 77610.0, 'Status': 7952.0, 'Link': 1613.0, 'Video': 1620.0}
```

```
# Generating bar graph
fig = plt.figure(figsize=(8, 8))

# Adds subplot on position 1
ax = fig.add_subplot(111)

# Generating and displaying bar chart
plt.bar(
    x=likes_per_type.keys(),
    height=likes_per_type.values()
)
plt.xlabel("Type of Post", fontsize=20)
plt.ylabel("Number of Likes", fontsize=20)
plt.title("Number of likes per type (Bar Plot)", fontsize=20)
plt.show()
```



4) Counting number of paid and unpaid posts



```
# Generating bar graph
fig = plt.figure(figsize=(8, 8))

# Adds subplot on position 1
ax = fig.add_subplot(111)

sns.countplot(x=facebook_dataset.Paid)

plt.xlabel("Paid posts (O : unpaid, 1: paid)", fontsize=20)
plt.ylabel("Count", fontsize=20)
plt.title("Count of paid and unpaid posts (Count plot)", fontsize=20)

plt.show()
```

Count of paid and unpaid posts (Count plot)

B) Heart Disease dataset

1) Loading the dataset

```
heart_dataset = pd.read_csv("./processed.cleveland.csv", header=None)
heart_dataset.head()
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

2) Renaming columns

```
heart_dataset.columns = [
    "age",
    "sex",
    "chest_pain",
    "trestbps",
    "cholestrol",
    "fbs",
    "restecg",
    "thalach",
    "exang",
    "oldpeak",
    "slope",
    "ca",
    "thal",
    "num"
]
```

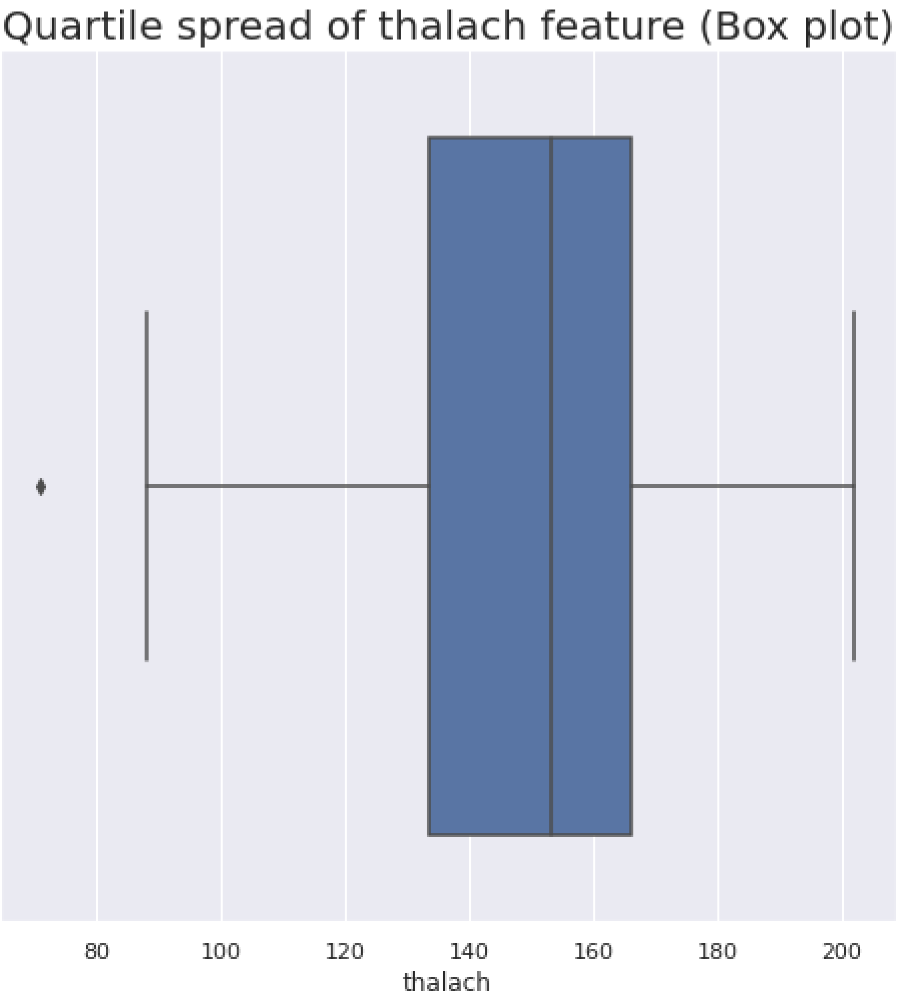
```
heart_dataset.head()
```

	age	sex	chest_pain	trestbps	cholestrol	fbs	restecg	thalach	exang	oldpe
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	

3) Quartile spread of thalach feature

4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0

```
# Generating bar graph
fig = plt.figure(figsize=(8, 8))

# Adds subplot on position 1
ax = fig.add_subplot(111)

sns.boxplot(x=heart_dataset.thalach)
plt.title("Quartile spread of thalach feature (Box plot)", fontsize=20)

plt.show()
```

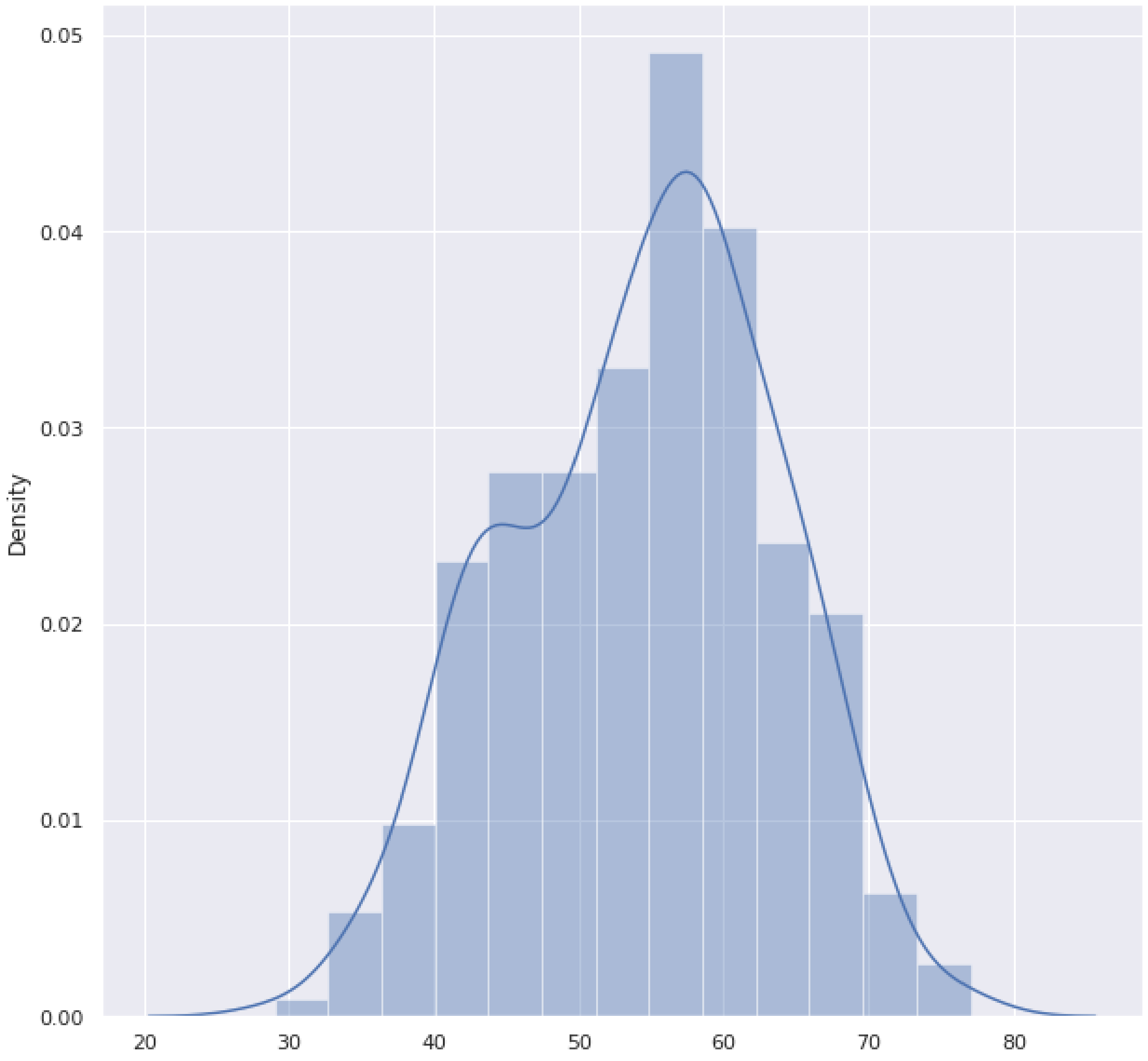


4) Distribution of age in entire dataset

```
# Generating bar graph
```

```
fig = plt.figure(figsize=(10, 10))

# Adds subplot on position 1
ax = fig.add_subplot(111)

sns.distplot(x=heart_dataset.age)
plt.show()
```

```
/home/varadmash/anaconda3/envs/python3.7_TF2.0/lib/python3.7/site-packages/seaborn
  warnings.warn(msg, FutureWarning)
```

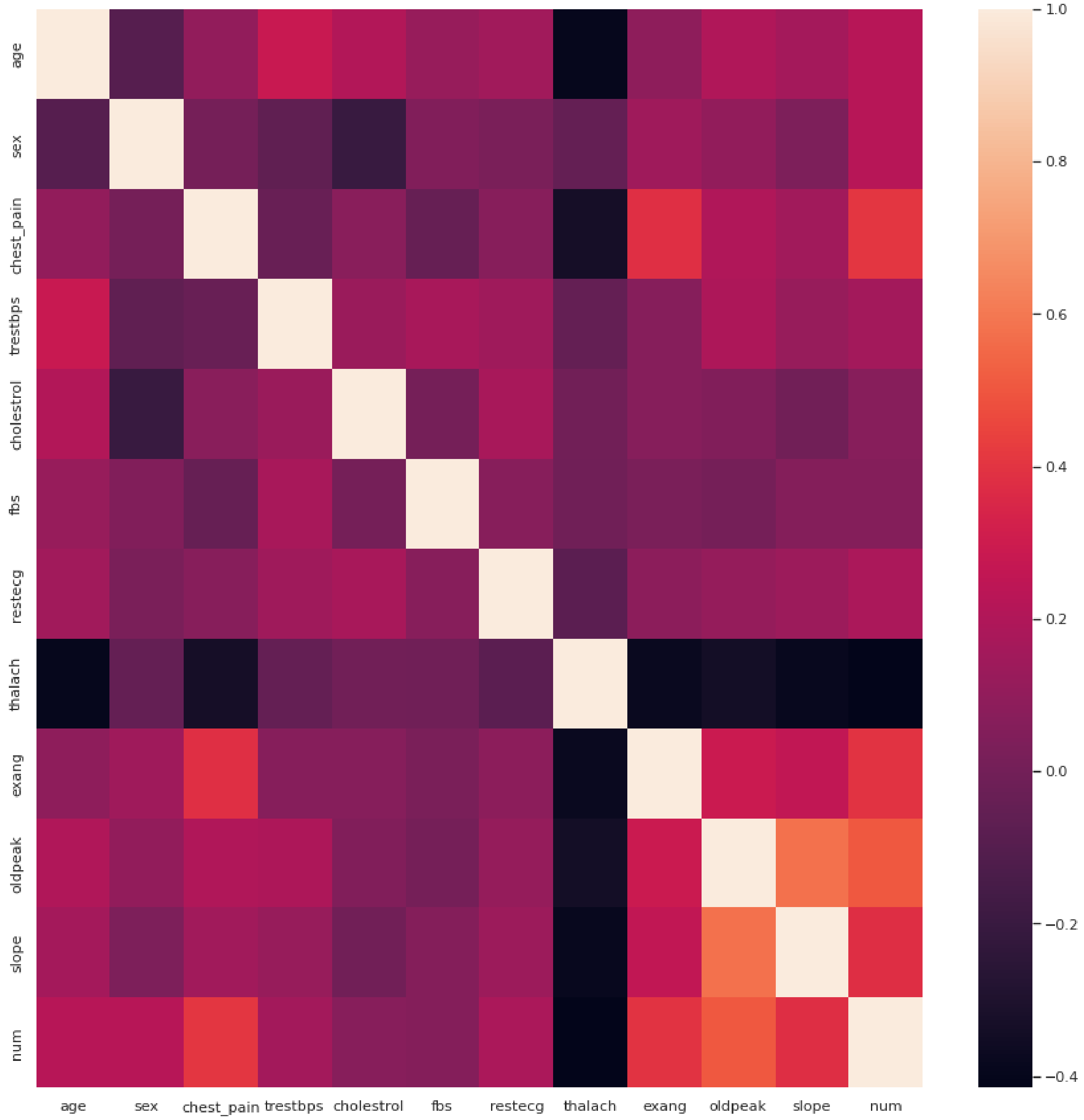


5) Checking correlation using heatmap

```
# Generating bar graph
fig = plt.figure(figsize=(15, 15))

# Adds subplot on position 1
ax = fig.add_subplot(111)

sns.heatmap(heart_dataset.corr())
```

```
plt.plot()
```

```
[]
```



Conclusion

1. Implemented following visualization methods :
    1. Pie chart
    2. Bar chart
    3. Count plot
    4. Box plot
    5. Distribution plot (Histogram)
    6. Heatmap

End of Notebook