Experiment 02 - Exploratory Data Analysis and Visualization

Roll No.	19
Name	Manav Jawrani
Class	D15A
Subject	Business Intelligence Lab
LO Mapped	LO1: Identify sources of Data for mining and perform data exploration
Grade	

Aim: To perform exploratory data analysis and visualization on the dataset using python.

Introduction:

- **What is EDA?**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

- **Why is EDA important in data science?**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

- **What are EDA tools?**

Specific statistical functions and techniques you can perform with EDA tools :
Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables. Univariate visualization of each field in the raw dataset, with summary statistics. Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at. Multivariate visualizations, for mapping and understanding interactions between different fields in the data. K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be

clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression. Predictive models, such as linear regression, use statistics and data to predict outcomes.

Descriptive analysis - Central tendency

(explain and execute how central tendency can be identified. Explain the inference with respect to your dataset.)

- **What is Central Tendency?**

  Measures of central tendency are summary statistics that represent the center point or typical value of a dataset. Examples of these measures include the mean, median, and mode. These statistics indicate where most values in a distribution fall and are also referred to as the central location of a distribution. You can think of central tendency as the propensity for data points to cluster around a middle value.

  In statistics, the mean, median, and mode are the three most common measures of central tendency. Each one calculates the central point using a different method. Choosing the best measure of central tendency depends on the type of data you have.

  According to our dataset-

  In a healthcare dataset focused on stroke data, central tendency could be used to describe various aspects of the population affected by stroke. For example, the mean age of onset could give an average value for when people typically have a stroke, while the median could provide a more robust measure of central tendency if the age data is heavily skewed in one direction. Additionally, central tendency measures could be used to describe the average severity of symptoms, frequency of hospital admissions, length of stay, and other relevant variables. By understanding the central tendency of these variables, healthcare providers can get a general sense of the typical experience of people with stroke and use this information to guide their treatment and prevention efforts.

Descriptive analysis - Dispersion

Data dispersion refers to the degree of variation or spread in a set of data. It can be measured in several ways including:

- Range: The difference between the largest and smallest values in a dataset.

- Interquartile Range (IQR): The difference between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset.

- Variance: The average of the squared differences from the mean in a dataset.

- Standard Deviation: The square root of the variance, it is a measure of the average distance of each data point from the mean.

According to our dataset-

Dispersion refers to the spread or variation in a dataset. In the context of a healthcare dataset focused on stroke data, dispersion could be used to describe the range and variability of various variables, such as age of onset, severity of symptoms, frequency of hospital admissions, length of stay, and so on. Common measures of dispersion include the range, variance, and standard deviation. The range gives the difference between the highest and lowest values in the data, while variance and standard deviation describe the spread of the data around the mean. By understanding the dispersion of these variables, healthcare providers can get a sense of how much variation there is in the population affected by stroke and use this information to guide their treatment and prevention efforts.

Correlation

Identifying correlations between different attributes in a dataset is an important step in understanding the relationships between variables and can inform further analysis. There are several approaches to identify correlations:

- Visualization: Plotting the data in a scatterplot can quickly show any potential linear relationships between two variables. If there is a positive correlation, the data points will form an upward slope; if there is a negative correlation, the data points will form a downward slope.

- Correlation Coefficient: The Pearson correlation coefficient is a measure of the linear relationship between two variables. It ranges from -1 (a perfect negative correlation) to 1 (a perfect positive correlation) with values near 0 indicating no correlation.

- Covariance: Covariance measures the joint variability of two variables. Positive covariance indicates that the variables are positively related (increase together), while negative covariance indicates that they are inversely related (one increases while the other decreases).

According to our dataset-

Correlation refers to the relationship between two variables in a dataset. In the context of a healthcare dataset focused on stroke data, correlation could be used to examine the relationship between variables such as age of onset, severity of symptoms, frequency of hospital admissions, and length of stay. A positive correlation means that as one variable increases, the other variable also tends to increase, while a negative correlation means that as one variable increases, the other variable tends to decrease. Correlation does not necessarily imply causality, but it can provide important information about the relationship between variables and help healthcare providers identify potential risk factors for stroke. By understanding the correlation between different variables, healthcare providers can develop more targeted and effective treatment and prevention strategies.

Data Visualization

Data visualization is the process of representing data in a visual format, such as a graph or chart, to help understand and communicate insights and patterns in the data. Some common data visualization techniques include:

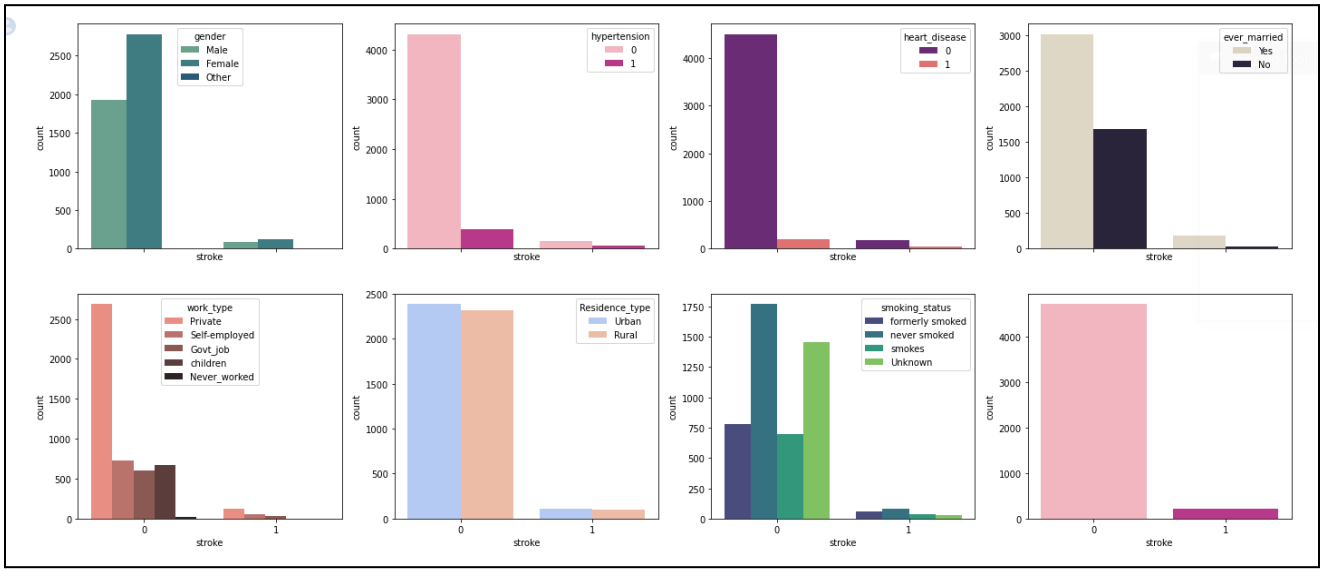
- Bar Charts: Used to compare categories or groups of data, such as the number of sales for each product.
- Line Charts: Used to show trends over time, such as the growth of a company's stock price.
- Scatter Plots: Used to show the relationship between two variables, such as the relationship between height and weight.
- Histograms: Used to show the distribution of data, such as the distribution of ages in a population.
- Pie Charts: Used to show proportions or percentages, such as the distribution of expenses in a budget.
- Box Plots: Used to show the spread and skewness of data, such as the spread of exam scores in a class.
- Heat Maps: Used to show the relationship between two variables, such as the relationship between temperature and humidity.

According to our dataset-

Data visualization refers to the process of creating graphical representations of data to help identify patterns, trends, and relationships. In the context of a healthcare dataset focused on stroke data, data visualization can be an important tool for understanding the distribution of variables such as age of onset, severity of symptoms, frequency of hospital admissions, and length of stay. Common types of data visualization include bar charts, line charts, scatter plots, histograms, and box plots. These visual representations can help healthcare providers identify outliers, trends, and patterns in the data that might not be immediately apparent from simply looking at the raw data. By using data visualization, healthcare providers can gain a deeper understanding of the population affected by stroke and use this information to guide their treatment and prevention efforts.

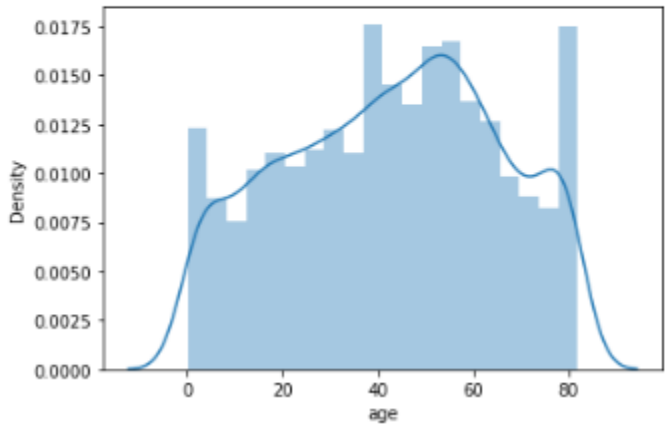
Implementation:

```
[ ]  fig, axes = plt.subplots(2, 4, sharex=True, figsize=(24,10))
     fig.suptitle('Count of all categorical variables')
     sns.countplot(ax=axes[0, 0], data=dataset, x='stroke',hue = 'gender',palette='crest')
     sns.countplot(ax=axes[0, 1], data=dataset, x='stroke',hue='hypertension',palette='RdPu')
     sns.countplot(ax=axes[0, 2], data=dataset, x='stroke',hue = 'heart_disease',palette='magma')
     sns.countplot(ax=axes[0, 3], data=dataset, x='stroke',hue = 'ever_married',palette="ch:s=-.2,r=.6")
     sns.countplot(ax=axes[1,0],data = dataset, x ='stroke' ,hue = 'work_type',palette="dark:salmon_r")
     sns.countplot(ax=axes[1,1],data = dataset, x ='stroke',hue='Residence_type',palette="coolwarm" )
     sns.countplot(ax=axes[1,2],data = dataset, x ='stroke',hue='smoking_status',palette='viridis' )
     sns.countplot(ax=axes[1,3],data = dataset, x ='stroke',palette='RdPu' )
     plt.show()
```



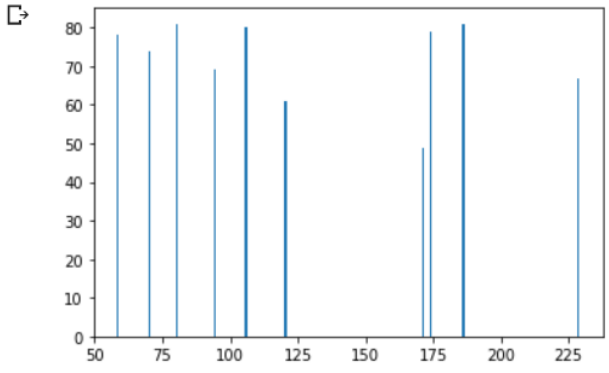

```
sns.distplot(dataset['age'])
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/distributio
  warnings.warn(msg, FutureWarning)
```



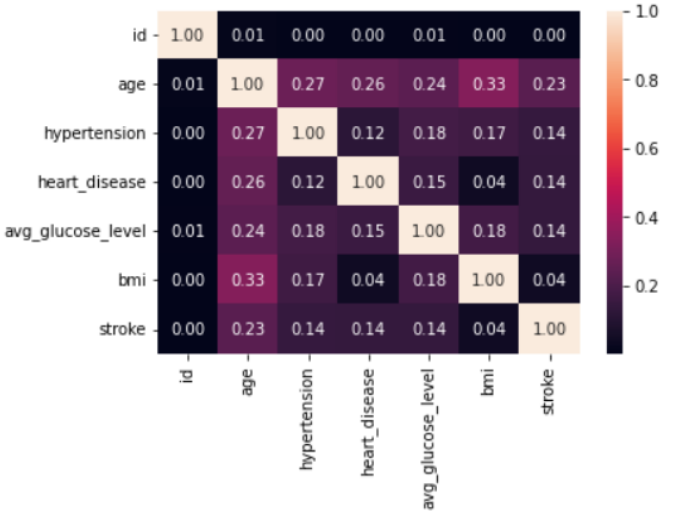
```
# Horizontal Bar Plot
plt.bar(name[0:10], price[0:10])

# Show Plot
plt.show()
```

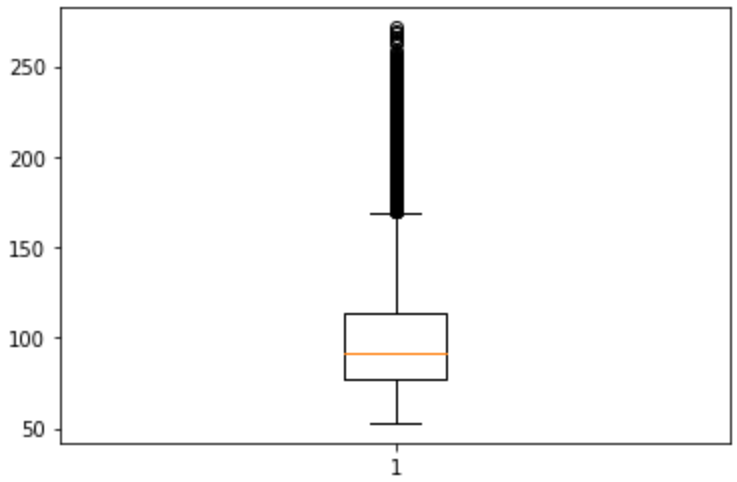



```
sns.heatmap(dataset.corr(), annot=True, fmt='.2f')
```

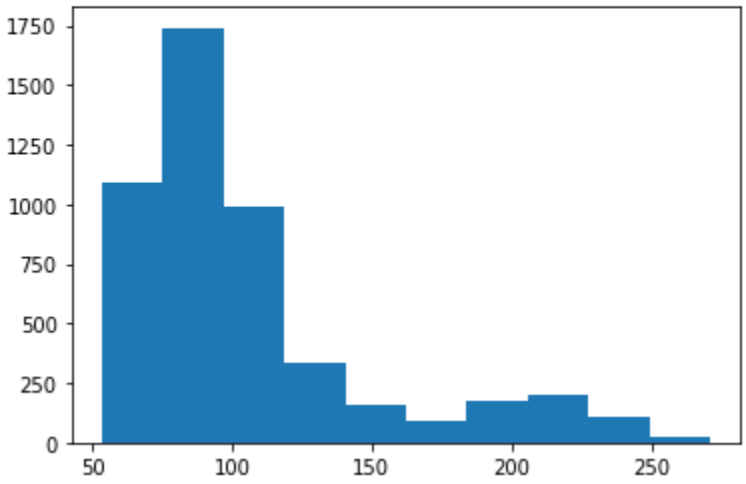
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f49c6320190>
```



```
[43] y = np.random.normal(df['avg_glucose_level'])
     plt.boxplot(y)
     plt.show()
```




```
x = np.random.normal(df['avg_glucose_level'])
plt.hist(x)
plt.show()
```



```
[26] df['avg_glucose_level'].mean()

     105.3051497249949

     df['avg_glucose_level'].std()

     44.42434066091561

[29] import statistics
     print(statistics.median(df['avg_glucose_level']))

     91.68

[30] df['avg_glucose_level'].var()

     1973.5220431570804

[43] y = np.random.normal(df['avg_glucose_level'])
     plt.boxplot(y)
     plt.show()
```

Conclusion:

Hence we have successfully performed Data Analysis and Visualisation on the Dataset of "Stroke Prediction".