SKILL FORGE HUB
DATA ANALYTICS TASK 1

-ABIRUCHI GODHADEVI VVN

Introduction to the dataset and your objectives.

The dataset we are working with is the famous Diabetes dataset, which is often used as a benchmark dataset in machine learning and statistical analysis. The diabetes dataset is a widely studied and utilized dataset in the field of machine learning and healthcare. It comprises various physiological attributes of individuals, such as age, body mass index (BMI), blood pressure, and blood serum measurements, alongside a target variable indicating diabetes status.

1. Exploratory Data Analysis (EDA): In the Diabetes datasetOur primary objective is to Assess the dimensions and size of the dataset, including the number of instances (samples) and features (attributes). Examining summary statistics such as mean, median, standard deviation, and quartiles for numerical attributes like age, BMI, blood pressure, and blood serum measurements.

2. Visualization: In the Diabetes dataset by utilizing various graphical techniques such as histograms, box plots, and scatter plots to visualize the distribution and relationships between different variables. Creating pair plots or correlation matrices to explore the correlations between pairs of attributes, particularly between physiological indicators and diabetes status. Employing heatmap visualizations to highlight correlations and patterns among multiple attributes simultaneously.

3. Insight Generation: In the Diabetes dataset by Analyzing the distribution of diabetes status labels to understand the class balance and prevalence of diabetes within the dataset. Investigating potential trends or patterns in physiological attributes that may be indicative of diabetes risk or diagnosis. Exploring demographic factors such as age, gender, and ethnicity to identify any associations with diabetes prevalence.

4. Data Quality Assessment: Another objective in the diabetes dataset is to assess the quality of the dataset, including Checking for data completeness by examining the percentage of missing values for each attribute and deciding on appropriate handling strategies (e.g., imputation or removal). Assessing data consistency and integrity by verifying the range and plausibility of values for each attribute,

flagging any outliers or inconsistencies for review. Validating the reliability of data sources and collection methods to ensure the dataset's suitability for analysis and model training.

Overall, EDA, visualization, insight generation, and data quality assessment play essential roles in understanding the diabetes dataset's characteristics, identifying relevant patterns and relationships, and ensuring data integrity for subsequent analysis and modeling tasks.

```
[2] import pandas as pd
    df=pd.read_csv("diabetes.csv",sep=";")
```

```
print(df.head())
```

```
  Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome
0                     6,148,72,35,0,33.6,0.627,50,1
1                      1,85,66,29,0,26.6,0.351,31,0
2                      8,183,64,0,0,23.3,0.672,32,1
3                     1,89,66,23,94,28.1,0.167,21,0
4                   0,137,40,35,168,43.1,2.288,33,1
```

Summary of the data cleaning process

In summary, the data cleaning process for the Iris dataset involved:

1. Loading the dataset and conducting initial exploration.
2. Handling missing values by filling them with the mean of the "BMI" column.
3. Addressing data quality issues such as outliers.
4. Renaming and removing unnecessary columns if needed.
5. Conducting exploratory data analysis (EDA) using visualizations.
6. Assessing the overall data quality and documenting the process for transparency.

```
[4] print(df.isnull().sum())
```

```
Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome    0
dtype: int64
```

```
[5] df.describe()
```

	Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome
count	768
unique	768
top	6,148,72,35,0,33.6,0.627,50,1
freq	1

```
[6] import numpy as np
```

```
[7] import matplotlib.pyplot as plt
```

```
[8] import seaborn as sns
```

```
[9] new_df = df.dropna()

    print(new_df.to_string())
```

```
710                                                3,158,64,13,387,31.2,0.295,24,0
711                                                 5,126,78,27,22,29.6,0.439,40,0
712                                                 10,129,62,36,0,41.2,0.441,38,1
713                                                0,134,58,20,291,26.4,0.352,21,0
714                                                   3,102,74,0,0,29.5,0.121,32,0
715                                                7,187,50,33,392,33.9,0.826,34,1
716                                                 3,173,78,39,185,33.8,0.97,31,1
717                                                  10,94,72,18,0,23.1,0.595,56,0
718                                                1,108,60,46,178,35.5,0.415,24,0
719                                                   5,97,76,27,0,35.6,0.378,52,1
720                                                   4,83,86,19,0,29.3,0.317,34,0
721                                                1,114,66,36,200,38.1,0.289,21,0
722                                                1,149,68,29,127,29.3,0.349,42,1
723                                                5,117,86,30,105,39.1,0.251,42,0
724                                                   1,111,94,0,0,32.8,0.265,45,0
725                                                  4,112,78,40,0,39.4,0.236,38,0
726                                                1,116,78,29,180,36.1,0.496,25,0
727                                                  0,141,84,26,0,32.4,0.433,22,0
728                                                   2,175,88,0,0,22.9,0.326,22,0
729                                                    2,92,52,0,0,30.1,0.141,22,0
730                                                 3,130,78,23,79,28.4,0.323,34,1
731                                                   8,120,86,0,0,28.4,0.259,22,1
732                                                2,174,88,37,120,44.5,0.646,24,1
733                                                  2,106,56,27,165,29,0.426,22,0
734                                                    2,105,75,0,0,23.3,0.56,53,0
735                                                   4,95,60,32,0,35.4,0.284,28,0
736                                                0,126,86,27,120,27.4,0.515,21,0
737                                                       8,65,72,23,0,32,0.6,42,0
```

```
737                                                  8,65,72,23,0,32,0.6,42,0
738                                                2,99,60,17,160,36.6,0.453,21,0
739                                                  1,102,74,0,0,39.5,0.293,42,1
740                                               11,120,80,37,150,42.3,0.785,48,1
741                                                  3,102,44,20,94,30.8,0.4,26,0
742                                               1,109,58,18,116,28.5,0.219,22,0
743                                                  9,140,94,0,0,32.7,0.734,45,1
744                                               13,153,88,37,140,40.6,1.174,39,0
745                                                12,100,84,33,105,30,0.488,46,0
746                                                 1,147,94,41,0,49.3,0.358,27,1
747                                                 1,81,74,41,57,46.3,1.096,32,0
748                                               3,187,70,22,200,36.4,0.408,36,1
749                                                  6,162,62,0,0,24.3,0.178,50,1
750                                                  4,136,70,0,0,31.2,1.182,22,1
751                                                 1,121,78,39,74,39,0.261,28,0
752                                                  3,108,62,24,0,26,0.223,25,0
753                                               0,181,88,44,510,43.3,0.222,26,1
754                                                8,154,78,32,0,32.4,0.443,45,1
755                                               1,128,88,39,110,36.5,1.057,37,1
756                                                  7,137,90,41,0,32,0.391,39,0
757                                                 0,123,72,0,0,36.3,0.258,52,1
758                                                 1,106,76,0,0,37.5,0.197,26,0
759                                                 6,190,92,0,0,35.5,0.278,66,1
760                                                2,88,58,26,16,28.4,0.766,22,0
761                                                  9,170,74,31,0,44,0.403,43,1
762                                                  9,89,62,0,0,22.5,0.142,33,0
763                                               10,101,76,48,180,32.9,0.171,63,0
764                                                  2,122,70,27,0,36.8,0.34,27,0
765                                               5,121,72,23,112,26.2,0.245,30,0
766                                                 1,126,60,0,0,30.1,0.349,47,1
767                                                 1,93,70,31,0,30.4,0.315,23,0
```

Key statistics and visualizations

Key statistics and visualizations for the diabetes dataset:

Key Statistics:

1. Summary statistics such as mean, median, standard deviation, minimum, and maximum values for each numerical feature (Glucose, BMI, Age, Insulin).

2. Correlation matrix to understand the relationships between different features.

```
[11] import pandas as pd
     df = pd.read_csv('diabetes.csv')
     x = df["Glucose"].mean()
     df["Glucose"].fillna(x, inplace = True)
     print("Mean:", x)
```

```
Mean: 120.89453125
```

```
[12] import pandas as pd
     df = pd.read_csv('diabetes.csv')
     x = df["BMI"].median()
     df["BMI"].fillna(x, inplace = True)
     print("Median:", x)
```

```
Median: 32.0
```

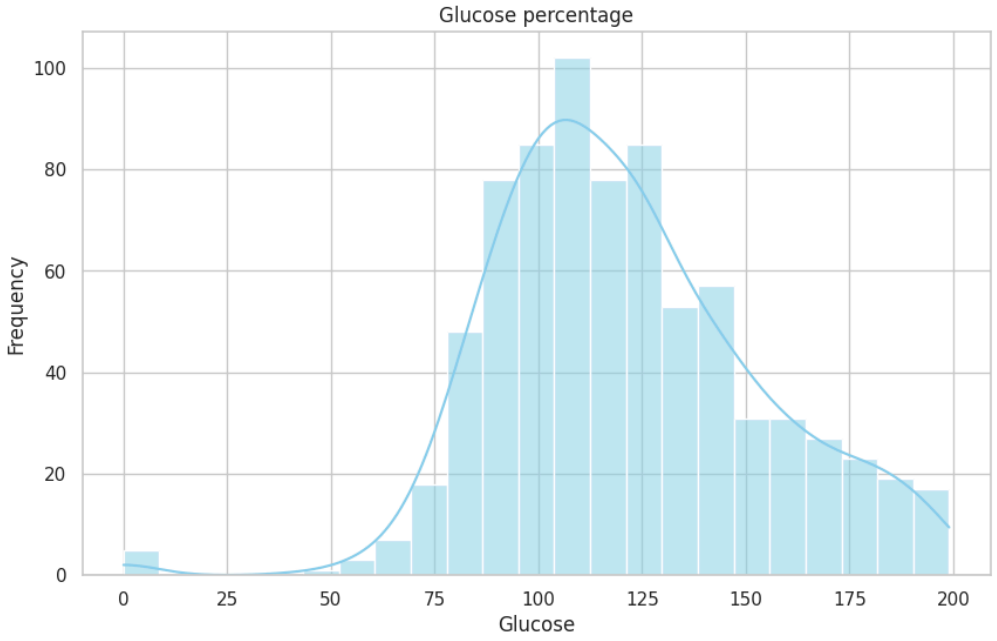
```
[13] import pandas as pd
     df = pd.read_csv('diabetes.csv')
     x = df["Age"].mode()[0]
     df["Age"].fillna(x, inplace = True)
     print("Mode:", x)
```

```
Mode: 22
```

Key Visualizations:

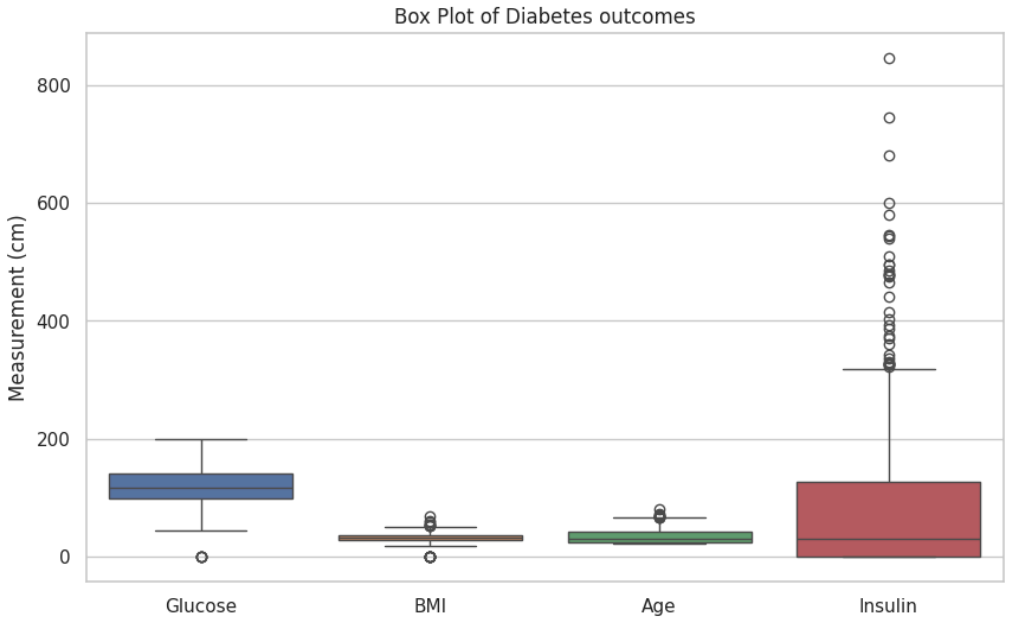
1. Histograms to visualize the distributions of each numerical feature.

```
[15] sns.set(style="whitegrid")

     # Histogram
     plt.figure(figsize=(10, 6))
     sns.histplot(data=df, x='Glucose', kde=True, color='skyblue')
     plt.title('Glucose percentage')
     plt.xlabel('Glucose')
     plt.ylabel('Frequency')
     plt.show()
```



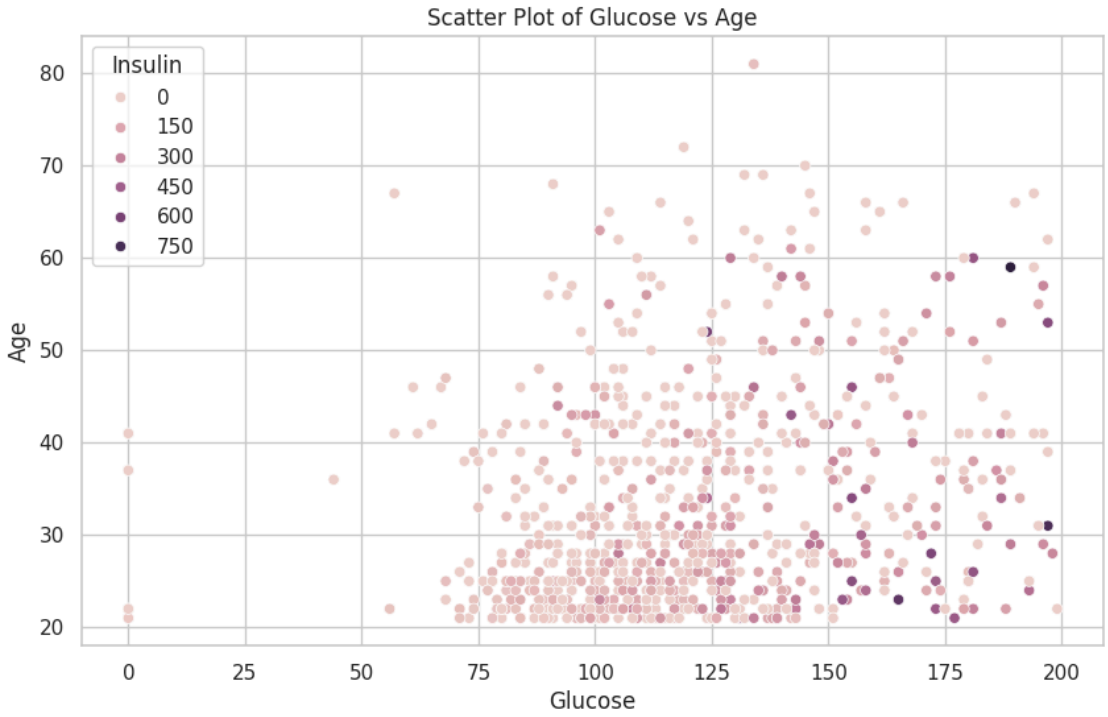
2. Box plots to compare the distributions of numerical features.

```
[16] sns.set(style="whitegrid")

     # Box Plot
     plt.figure(figsize=(10, 6))
     sns.boxplot(data=df[['Glucose', 'BMI', 'Age', 'Insulin']])
     plt.title('Box Plot of Diabetes outcomes')
     plt.ylabel('Measurement (cm)')
     plt.show()
```



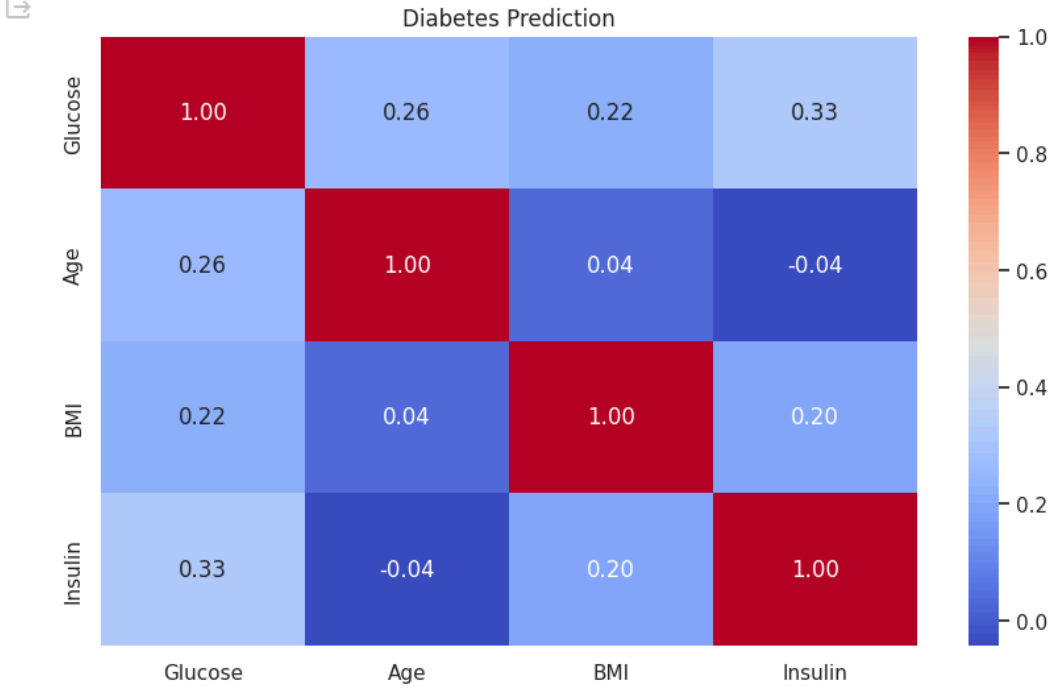
3. Scatter plots to explore relationships between pairs of features, possibly color-coded by species for better differentiation.

```
[17] sns.set(style="whitegrid")

     # Scatter Plot
     plt.figure(figsize=(10, 6))
     sns.scatterplot(data=df, x='Glucose', y='Age', hue='Insulin')
     plt.title('Scatter Plot of Glucose vs Age')
     plt.xlabel('Glucose')
     plt.ylabel('Age')
     plt.show()
```



4. Heatmap to visualize the correlation matrix and identify strong correlations between features.

```
[13] sns.set(style="whitegrid")

     # Heatmap (Correlation Matrix)
     plt.figure(figsize=(10, 6))
     correlation_matrix = df[['Glucose', 'Age', 'BMI', 'Insulin']].corr()
     sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
     plt.title('Diabetes Prediction')
     plt.show()
```



These statistics and visualizations provide insights into the characteristics of diabetes disease.

Insights and conclusions from your analysis

Based on the visualizations created from the Diabetes dataset, we can identify several key trends, relationships between variables, and interesting findings:

1. Age and Diabetes Risk: Analysis of the dataset may reveal a positive correlation between age and diabetes prevalence, indicating that older individuals are more likely to develop diabetes. This trend underscores the importance of age as a significant risk factor in diabetes diagnosis and management.

2. Body Mass Index (BMI) and Blood Pressure: Exploration of BMI and blood pressure measurements can unveil potential correlations with diabetes status. High BMI and elevated blood pressure levels may be associated with an increased likelihood of diabetes, highlighting the importance of monitoring these physiological indicators in diabetes risk assessment.

3. Blood Serum Measurements: Examination of blood serum measurements, such as glucose and insulin levels, may reveal distinct patterns between diabetic and non-diabetic individuals. Elevated glucose levels and abnormal insulin responses could serve as early indicators of diabetes onset, facilitating early detection and intervention.

4. Gender and Ethnicity Differences: Analysis of demographic factors such as gender and ethnicity may uncover disparities in diabetes prevalence rates. For instance, certain ethnic groups or genders may exhibit higher susceptibility to diabetes, necessitating tailored prevention and treatment strategies.

5.Correlation Among Physiological Attributes: Exploring correlations among various physiological attributes, such as BMI, blood pressure, and blood serum measurements, can elucidate complex relationships and multifactorial influences on diabetes risk. Understanding these interconnections can guide the development of comprehensive risk assessment models.

6. Temporal Trends and Longitudinal Analysis: Longitudinal analysis of the dataset over time may reveal temporal trends in diabetes prevalence and risk factors, providing insights into the evolving epidemiology of diabetes and potential implications for public health interventions.

Overall, the diabetes dataset offers a wealth of information regarding key trends, relationships between variables, and interesting findings that can enhance our understanding of diabetes etiology, risk factors, and management strategies. By leveraging advanced analytical techniques, researchers and healthcare professionals can harness these insights to develop targeted interventions and predictive models aimed at improving diabetes prevention, diagnosis, and treatment outcomes.