Assignment 3

Name – Sachin

Reg no – 21BAC10036

Campus – Vit Bhopal

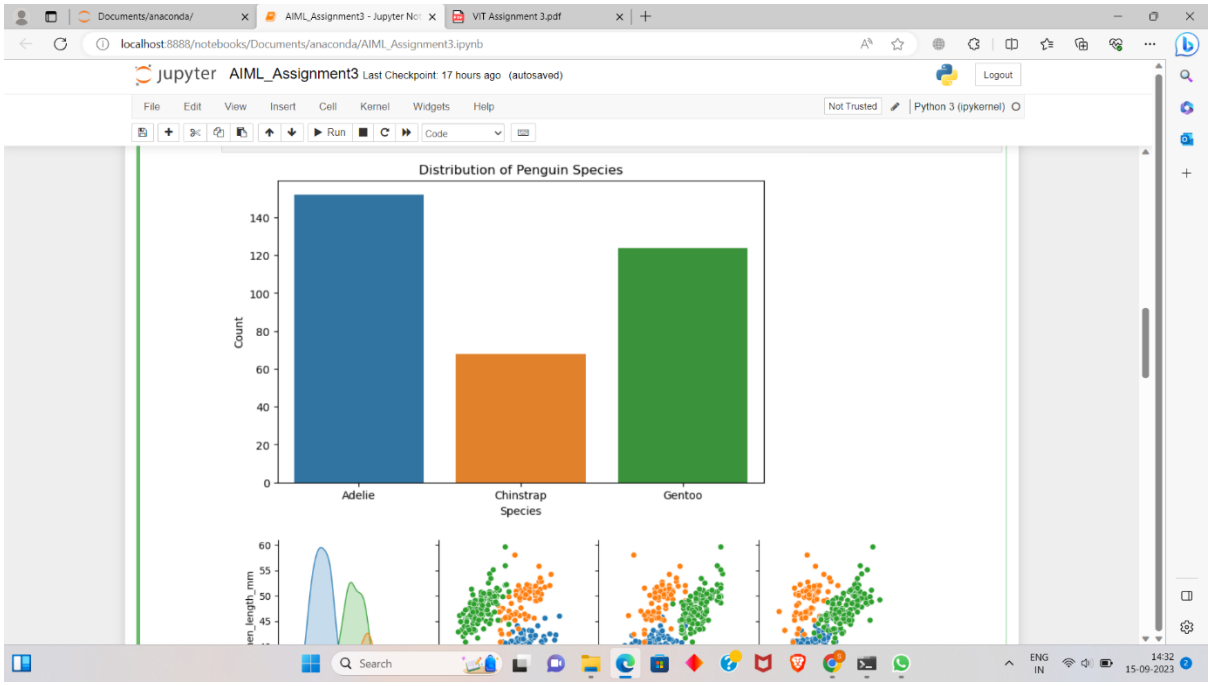
Code:

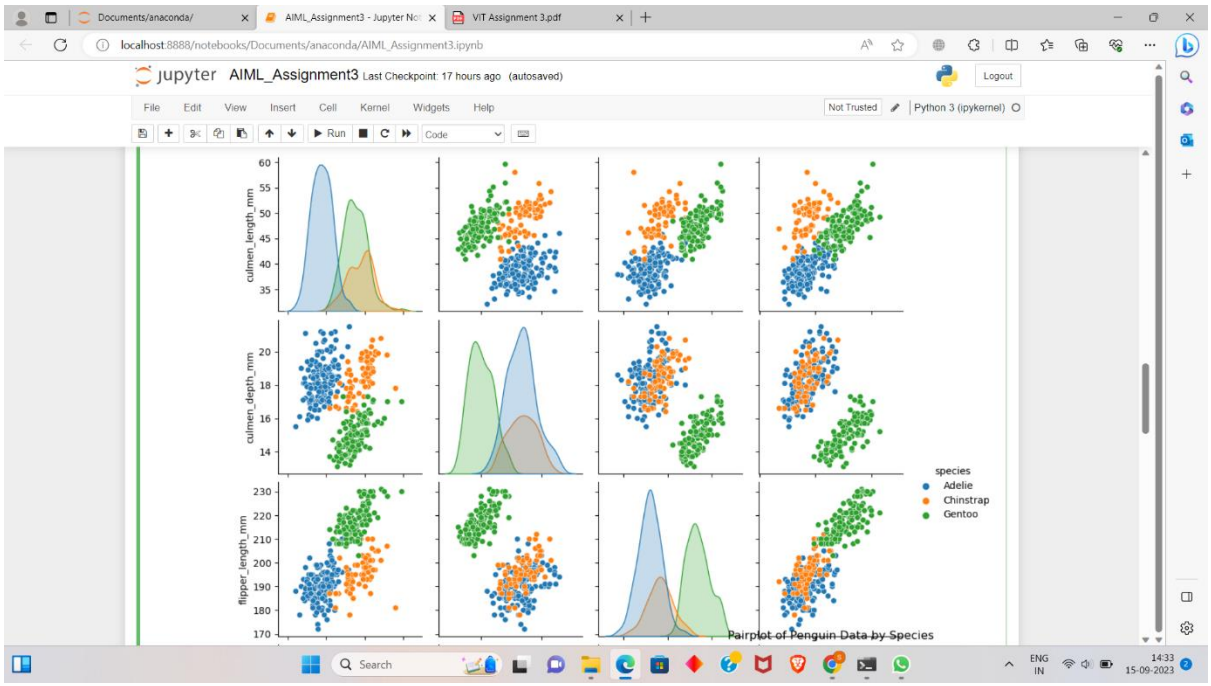
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.model_selection import train_test_split
from scipy import stats

# Load the dataset
data = pd.read_csv("penguins_size.csv")

# 1. Univariate Analysis
plt.figure(figsize=(8, 5))
sns.countplot(data=data, x='species')
plt.title('Distribution of Penguin Species')
plt.xlabel('Species')
plt.ylabel('Count')
plt.show()

# 2. Bivariate Analysis
sns.pairplot(data=data, hue='species')
plt.title('Pairplot of Penguin Data by Species')
```

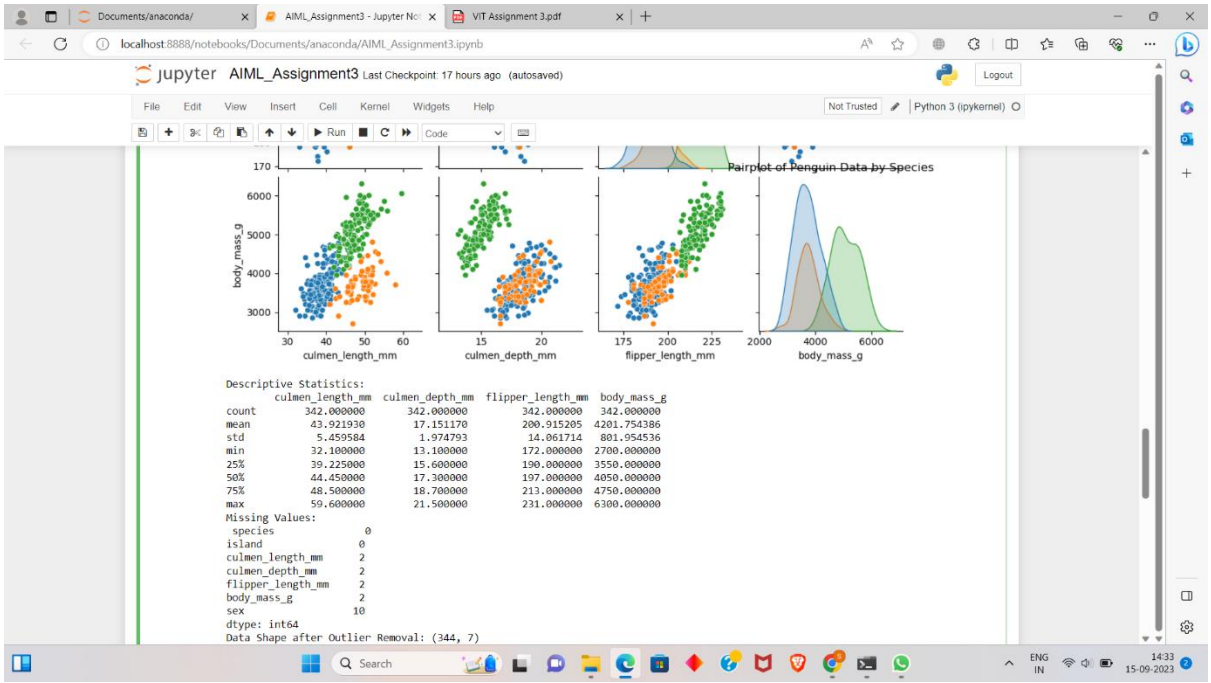
```
plt.show()

# 3. Descriptive Statistics
desc_stats = data.describe()
print("Descriptive Statistics:\n", desc_stats)

# 4. Handling Missing Values
missing_values = data.isnull().sum()
print("Missing Values:\n", missing_values)

# 5. Outlier Detection and Handling (Z-score method)
z_scores = np.abs(stats.zscore(data.select_dtypes(include=np.number)))
threshold = 3  # Adjust the threshold as needed
outlier_rows, outlier_cols = np.where(z_scores > threshold)
data_no_outliers = data.drop(data.index[outlier_rows])
print("Data Shape after Outlier Removal:", data_no_outliers.shape)

# 6. Correlation Analysis
numeric_data = data_no_outliers.select_dtypes(include=np.number)
correlation_matrix = numeric_data.corr()
print("Correlation Matrix:\n", correlation_matrix)

# 7. Check for Categorical Columns and Perform Encoding
categorical_columns = data_no_outliers.select_dtypes(include=[object]).columns
if 'species' in categorical_columns:
    categorical_columns = categorical_columns.drop('species')  # Exclude target column

label_encoder = LabelEncoder()
for column in categorical_columns:
    data_no_outliers[column] = label_encoder.fit_transform(data_no_outliers[column])
```

```
print("Data after Categorical Encoding:\n", data_no_outliers.head())

# 8. Split the data into dependent and independent variables
X = data_no_outliers.drop(columns=['species'])
y = data_no_outliers['species']
# Display the contents and shapes of X and y
print("Independent Variables (X):\n", X.head())  # Display the first few rows of X
print("Dependent Variable (y):\n", y.head())  # Display the first few rows of y
print("Shape of Independent Variables (X):", X.shape)
print("Shape of Dependent Variable (y):", y.shape)


# 9. Scaling the Data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
print("Scaled Data:\n", X_scaled)

# 10. Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42)

# 11. Check the shapes of training and testing data
print("Training Data Shape - X_train, y_train:", X_train.shape, y_train.shape)
print("Testing Data Shape - X_test, y_test:", X_test.shape, y_test.shape)
```

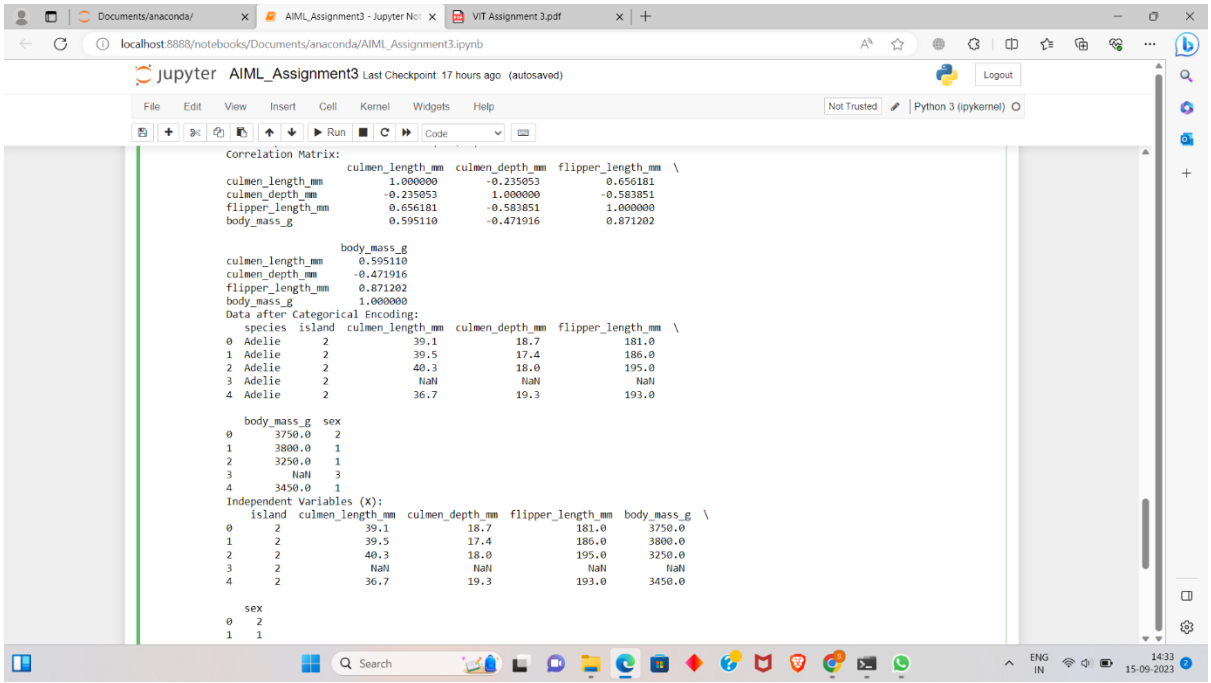
Output:


Distribution of Penguin Species
Count
Adelie
Chinstrap
Gentoo
Species

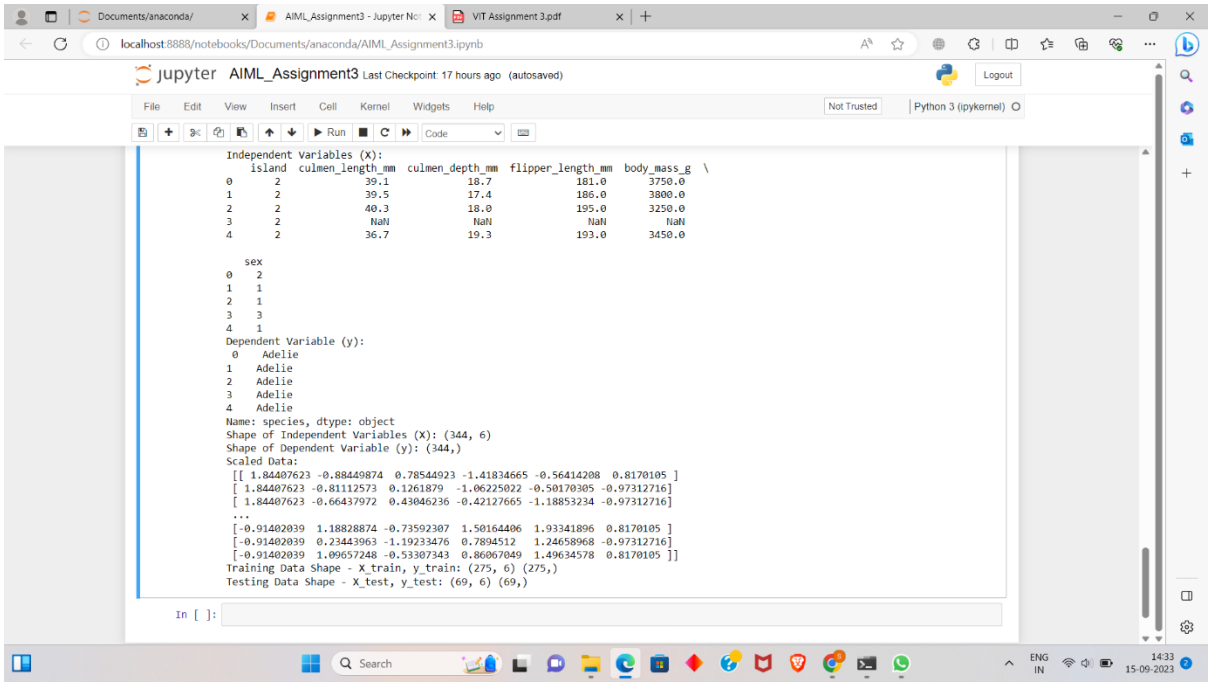

culmen_length_mm
culmen_depth_mm
flipper_length_mm
species
Adelie
Chinstrap
Gentoo
Pairplot of Penguin Data by Species




```
Descriptive Statistics:
       culmen_length_mm  culmen_depth_mm  flipper_length_mm  body_mass_g
count        342.000000       342.000000         342.000000   342.000000
mean          43.921930        17.151170         200.915205  4201.754386
std            5.459584         1.974793          14.061714   801.954536
min           32.100000        13.100000         172.000000  2700.000000
25%           39.225000        15.600000         190.000000  3550.000000
50%           44.450000        17.300000         197.000000  4050.000000
75%           48.500000        18.700000         213.000000  4750.000000
max           59.600000        21.500000         231.000000  6300.000000
Missing Values:
 species              0
island               0
culmen_length_mm     2
culmen_depth_mm      2
flipper_length_mm    2
body_mass_g          2
sex                 10
dtype: int64
Data Shape after Outlier Removal: (344, 7)
```



```
Correlation Matrix:
                   culmen_length_mm  culmen_depth_mm  flipper_length_mm  \
culmen_length_mm           1.000000        -0.235053           0.656181
culmen_depth_mm           -0.235053         1.000000          -0.583851
flipper_length_mm          0.656181        -0.583851           1.000000
body_mass_g                0.595110        -0.471916           0.871202

                   body_mass_g
culmen_length_mm      0.595110
culmen_depth_mm      -0.471916
flipper_length_mm     0.871202
body_mass_g           1.000000
Data after Categorical Encoding:
   species  island  culmen_length_mm  culmen_depth_mm  flipper_length_mm  \
0  Adelie       2              39.1             18.7              181.0
1  Adelie       2              39.5             17.4              186.0
2  Adelie       2              40.3             18.0              195.0
3  Adelie       2               NaN              NaN                NaN
4  Adelie       2              36.7             19.3              193.0

   body_mass_g  sex
0       3750.0    2
1       3800.0    1
2       3250.0    1
3          NaN    3
4       3450.0    1
Independent Variables (X):
   island  culmen_length_mm  culmen_depth_mm  flipper_length_mm  body_mass_g  \
0       2              39.1             18.7              181.0       3750.0
1       2              39.5             17.4              186.0       3800.0
2       2              40.3             18.0              195.0       3250.0
3       2               NaN              NaN                NaN          NaN
4       2              36.7             19.3              193.0       3450.0

   sex
0    2
1    1
```


Documents/anaconda/
AIML_Assignment3 - Jupyter No
ViT Assignment 3.pdf
localhost:8888/notebooks/Documents/anaconda/AIML_Assignment3.ipynb
jupyter AIML_Assignment3 Last Checkpoint: 17 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Independent Variables (X):
island culmen_length_mm culmen_depth_mm flipper_length_mm body_mass_g \
0 2 39.1 18.7 181.0 3750.0
1 2 39.5 17.4 186.0 3800.0
2 2 40.3 18.0 195.0 3250.0
3 2 NaN NaN NaN NaN
4 2 36.7 19.3 193.0 3450.0
sex
0 2
1 1
2 1
3 3
4 1
Dependent Variable (y):
0 Adelie
1 Adelie
2 Adelie
3 Adelie
4 Adelie
Name: species, dtype: object
Shape of Independent Variables (X): (344, 6)
Shape of Dependent Variable (y): (344,)
Scaled Data:
[[1.84407623 -0.88449874 0.78544923 -1.41834665 -0.56414208 0.8170105]
[1.84407623 -0.81112573 0.1261879 -1.06225022 -0.50170305 -0.97312716]
[1.84407623 -0.66437972 0.43046236 -0.42127665 -1.18853234 -0.97312716]
...
[-0.91402039 1.18828874 -0.73592307 1.50164406 1.93341896 0.8170105]
[-0.91402039 0.23443963 -1.19233476 0.7894512 1.24658968 -0.97312716]
[-0.91402039 1.09657248 -0.53307343 0.86067049 1.49634578 0.8170105]]
Training Data Shape - X_train, y_train: (275, 6) (275,)
Testing Data Shape - X_test, y_test: (69, 6) (69,)
In [ ]:
Search
ENG
IN
14:33
15-09-2023