Assesment Report

on

“Problem Statement”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY

DEGREE

SESSION 2024-25

in

Name of discipline

By

Name: Kartikey Kumar

(Roll Number):202401100400106

Under the supervision of

“Mr. Abhishek Shukla”

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

📝 AI MSE Report: Data Visualization Based on Cancer Dataset

---

1. Introduction

The objective of this task is to visualize and interpret patterns in cancer diagnosis data. The dataset contains various medical measurements, and each entry is labeled as either benign or malignant. By creating visualizations, we aim to understand feature distributions and their relationship to diagnosis outcomes.

---

2. Methodology

1. **Data Loading**: Used Pandas to read a CSV file containing breast cancer data.
2. **Data Cleaning**: Removed unnecessary columns like `id` and `Unnamed: 32` which had no meaningful data.
3. **Visualization**:
    - Countplot to compare the number of malignant vs benign cases.
    - Histogram to check the distribution of `radius_mean`.
    - Boxplot to observe differences in `radius_mean` across diagnosis types.
    - Heatmap to analyze correlation between top features and diagnosis outcome.
4. **Libraries Used**: `pandas`, `numpy`, `seaborn`, and `matplotlib.pyplot`.

---

3. Code

python
CopyEdit

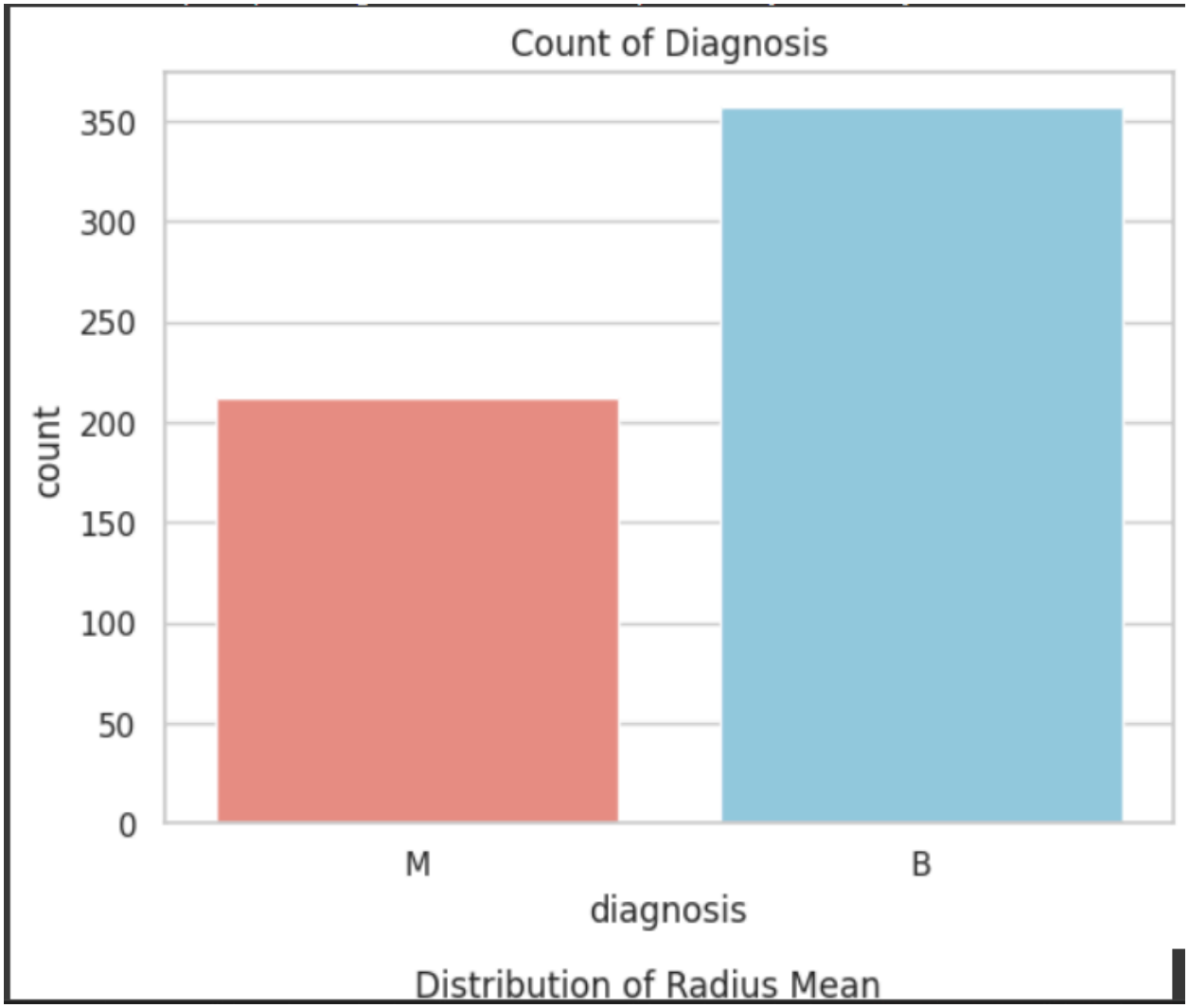
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
df = pd.read_csv("3. Predict Disease Outcome Based on Genetic and
Clinical Data.csv")

# Drop unnecessary columns
df = df.drop(columns=["id", "Unnamed: 32"])

# 1. Countplot
sns.countplot(x="diagnosis", data=df, palette={"B": "skyblue", "M":
"salmon"})
plt.title("Count of Diagnosis")
plt.show()

# 2. Histogram
sns.histplot(df["radius_mean"], bins=30, kde=True,
color="mediumseagreen")
plt.title("Distribution of Radius Mean")
plt.show()

# 3. Boxplot
sns.boxplot(x="diagnosis", y="radius_mean", data=df, palette={"B":
"lightblue", "M": "lightcoral"})
plt.title("Radius Mean by Diagnosis")
plt.show()

# 4. Heatmap
df["diagnosis"] = df["diagnosis"].map({"M": 1, "B": 0})
corr = df.corr()
top =
corr["diagnosis"].abs().sort_values(ascending=False).head(11).index
```

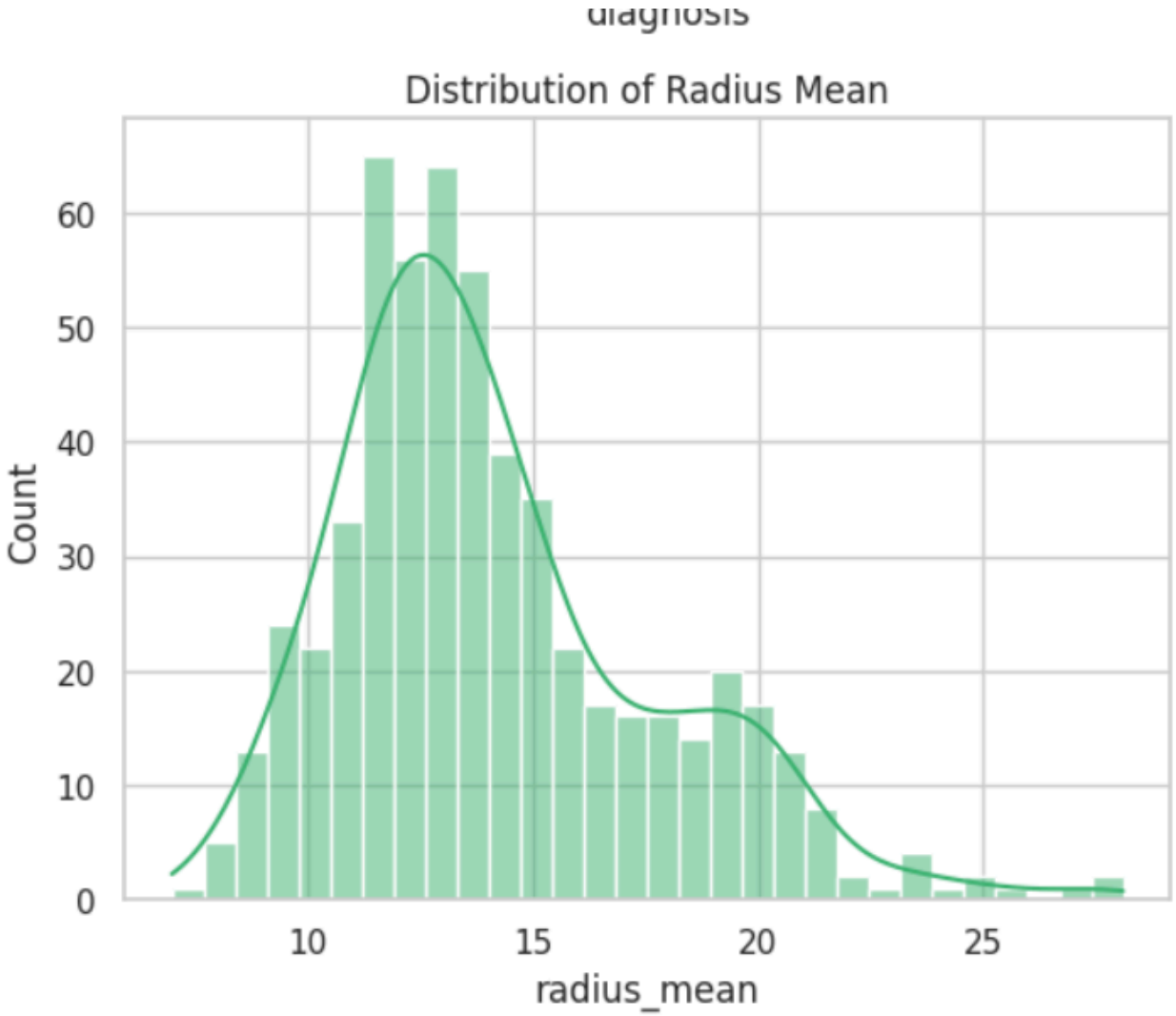
```
sns.heatmap(df[top].corr(), annot=True, cmap="coolwarm")
plt.title("Top Feature Correlations with Diagnosis")
plt.show()
```

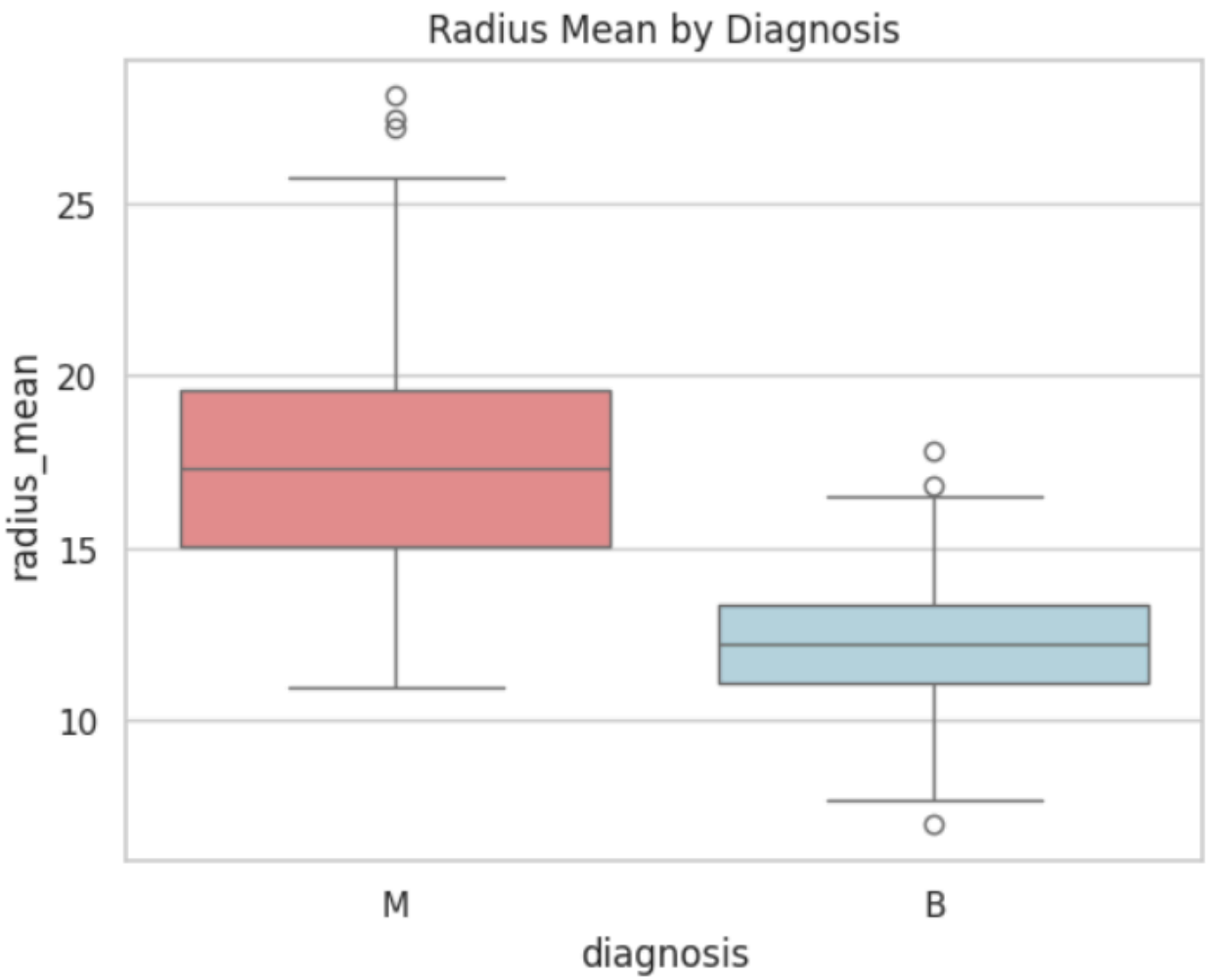
---

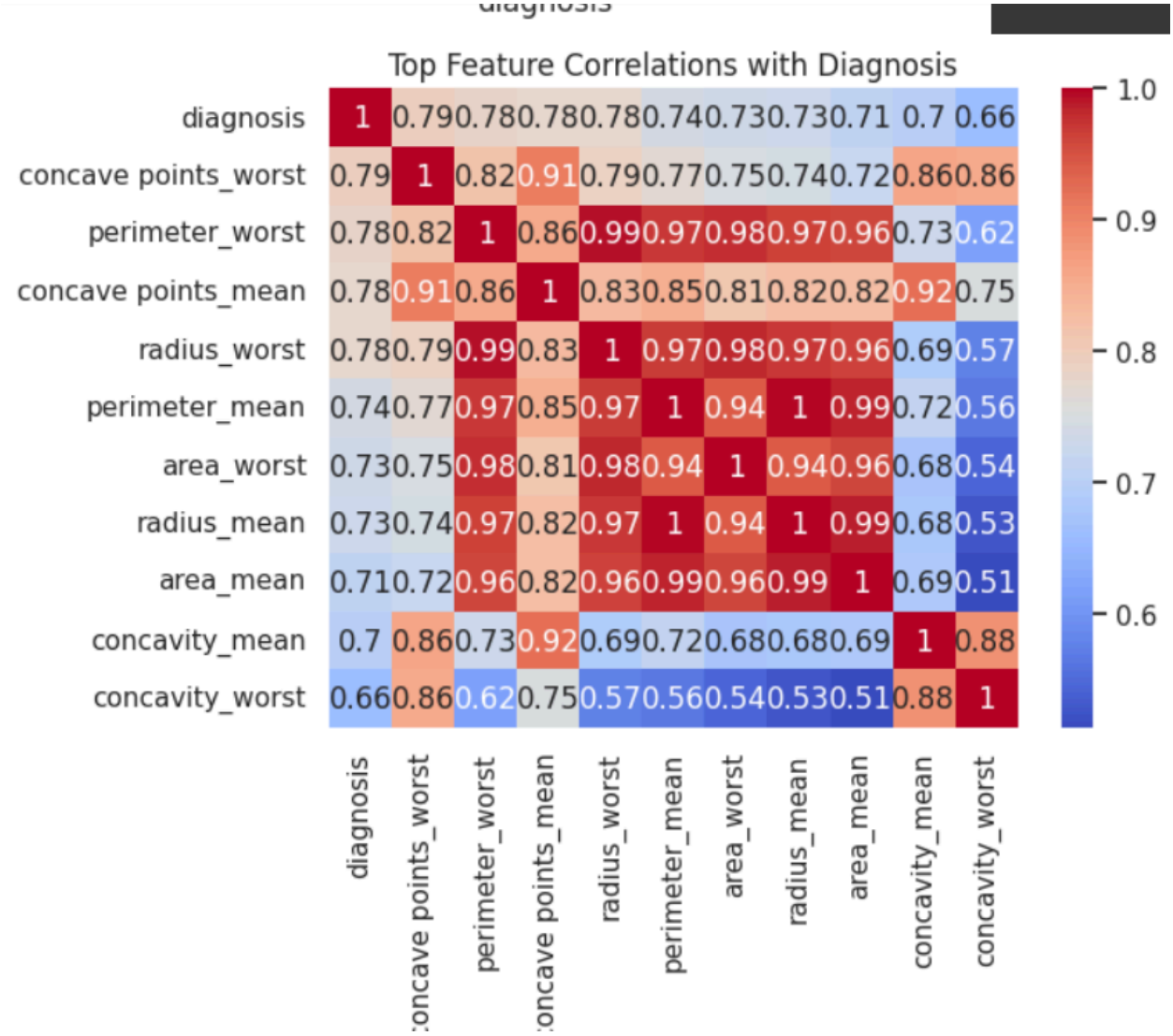
4. Output/Result

📌 Screenshots of graphs (Countplot, Histogram, Boxplot, Heatmap).




diagnosis
Distribution of Radius Mean
Count
0
10
20
30
40
50
60
10
15
20
25
radius_mean


Radius Mean by Diagnosis
radius_mean
25
20
15
10
M
B
diagnosis



---

5. References/Credits

- Dataset Source: Cancer Wisconsin Dataset
- Libraries: Pandas, NumPy, Matplotlib, Seaborn