

Assignment-2

Data Analysis on NETFLIX

Introduction(Part-I):

The large dataset contains several entries on the netflix TV shows and movies.

These are the 12 columns in the data:

'show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description'

The rows are 8807. Some columns have missing values, such as 'director', 'cast', 'country' etc.

The nature of the data is that its a mix of categorical and numerical data.

Dataset Overview:

Example entry :

show_id: s1 (show_id is the unique identifier to)

type: Movie (it is a binary data TV show or Movie)

title: Dick Johnson Is Dead

director: Kirsten Johnson

cast: Dakota Johnson, ryan gosling

country: United States

date_added: September 25, 2021

release_year: 2020

rating: PG-13 (Categorical data)


duration: 90 min

listed_in: Documentaries (genre)

description: As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable. (string)

we can analyze the content on Netflix like genre distribution, country-wise content, etc

Basic statistics:



	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

This is the total number of null values in each column :

```
show_id      0
type         0
title        0
director     2634
cast         825
country      831
date_added   10
release_year  0
rating       4
duration     3
listed_in    0
description  0
dtype: int64
```

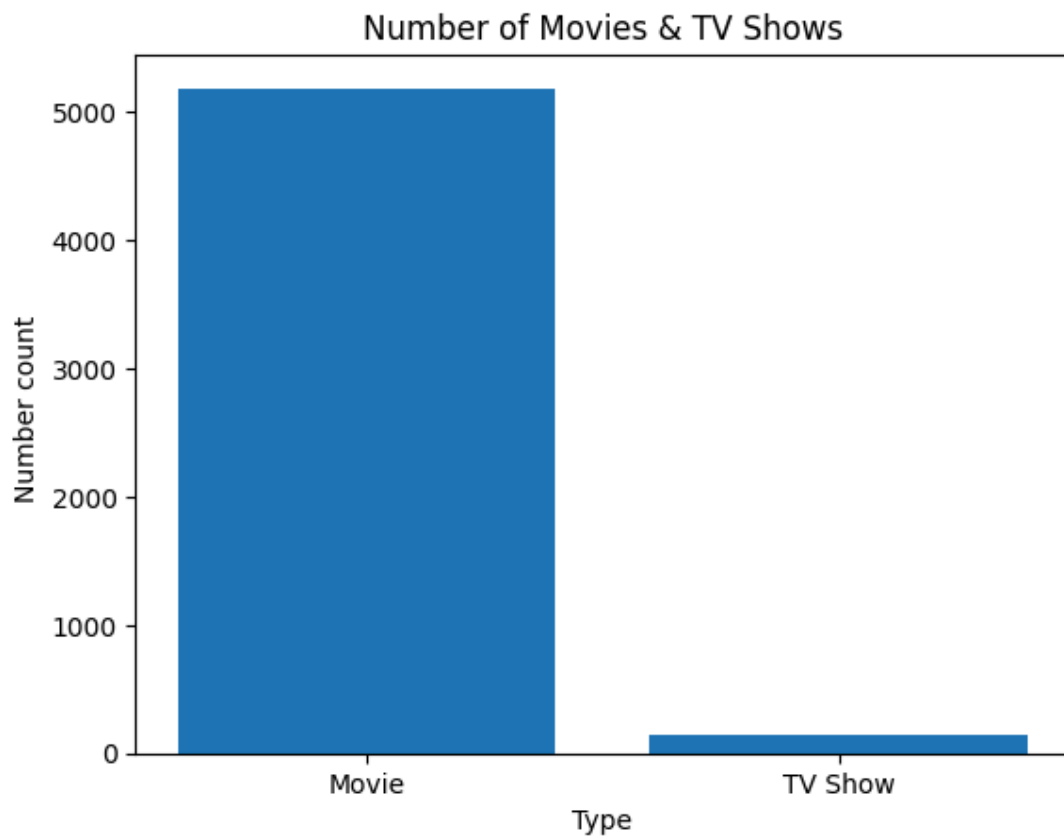
We are preprocessing and dealing with the null values in such way, for these columns
'director','cast','country','rating'

For these columns I removed the entire rows, For column 'duration' I am replacing the null value with the most common value in the column.

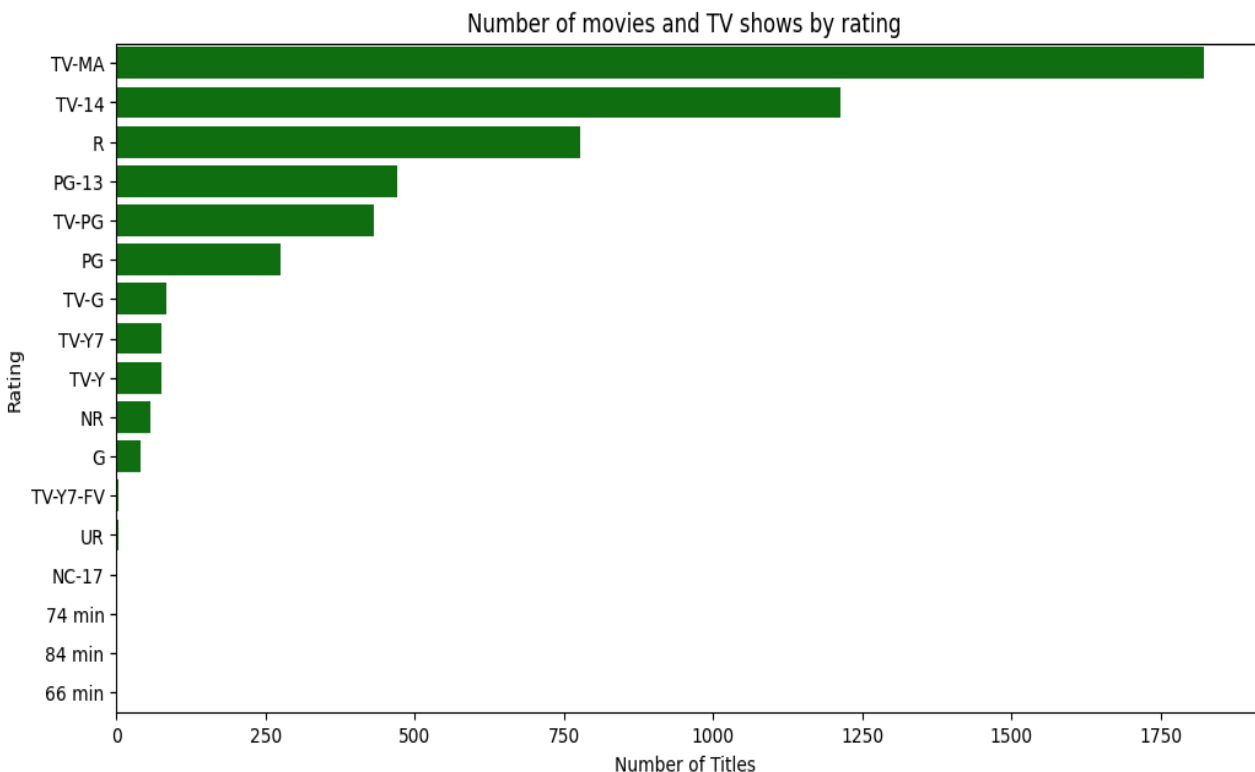
I am also converting a few columns to numeric categorical data.

Data Visualization:

- It employs Matplotlib to create a bar plot in order to provide a visualization of the counts for each type.
- The x-axis, then, represents the types, and the y-axis the amount of movies included within these types.
- Chart Title: Number of Movies by Genre type .
- This graph clearly shows a quick sense of the number of movies in each type is more than TV Shows in the dataset.

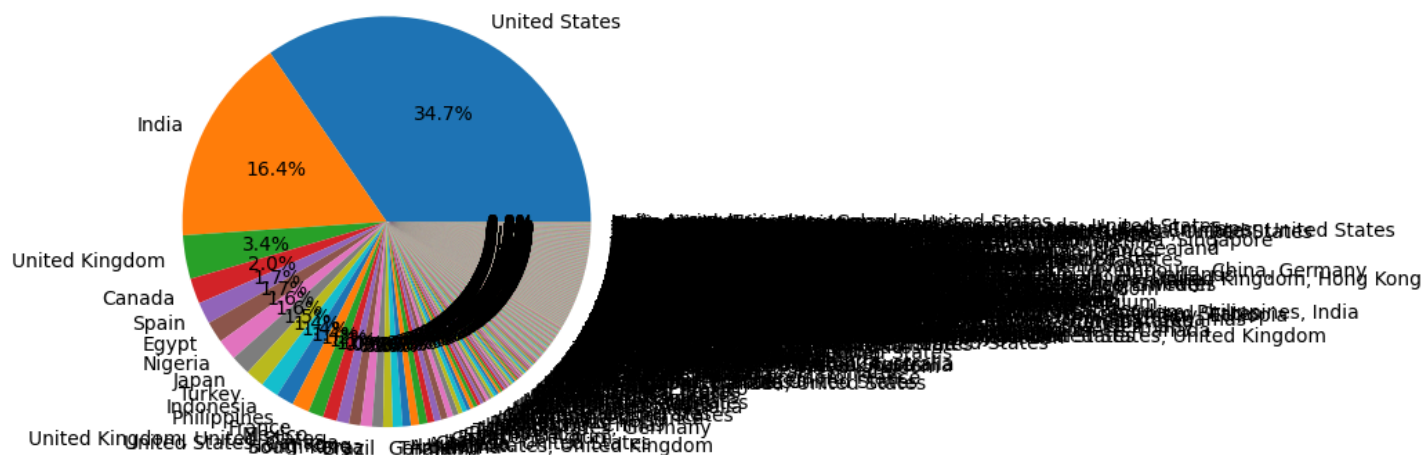


- From the bar plot of Figure, it can easily be seen that there are way more movies than TV shows. This is expected, since, in general, movies tend to be more popular and produced than TV shows.
- As for movies, the most frequent one is "Drama", followed by "Comedy" and "Action". All these demonstrate that these genres are the most popular among the people using Netflix.
- The least common, "Musical", indicates a general trend of musicals being much less popular than other categories of movies.
- Interestingly, "Documentaries" are more prevalent in the dataset than "Reality TV" material. This may be a reflection of how documentaries may be much more informative and educational in nature, keeping in mind reality TV shows can also be quite repetitive and sensationalist at times.
- Overall, the bar plot is pretty useful to get an idea of the various types of movies and shows on Netflix, wherein a positive recommendation to the viewers of what to watch depending on its availability on the platform could be made or even new content can be developed which can be popular.



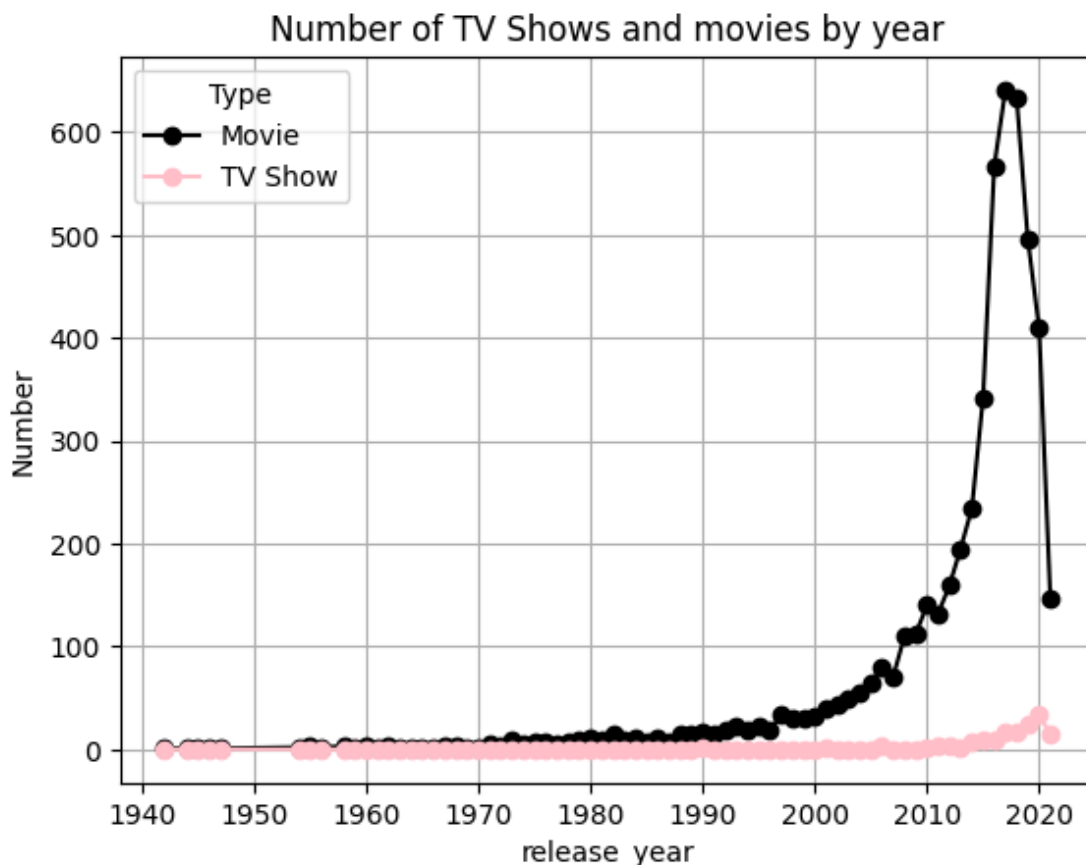
- It can be noticed from the pie chart that the majority of the movies and TV shows, 60.9%, are from the United States. That is not surprising because the United States is the biggest market for Netflix and, consequently, it should be expected that is where the most content would come from, but also the United States has a big tradition of making popular movies and TV shows.
- The next most common country for movies and TV shows is the United Kingdom, or 6.5%, India with 5.5%, and Canada with 4.5% of the content. These countries are all huge in the production of movies and TV shows and have a large proportion of the viewers who speak the English language.
- Other nations represented in the pie chart include France, Spain, Germany, Japan, and South Korea. All these countries are very famous for their high-quality movies and TV shows, and they are widely followed on Netflix.
- This generally means that the pie chart reveals the diversity in movies and TV shows from across the world that Netflix offers. This is very much expected to attract a large viewership and also make it one of the popular streaming services.

Number of Movies & TV Shows by Country

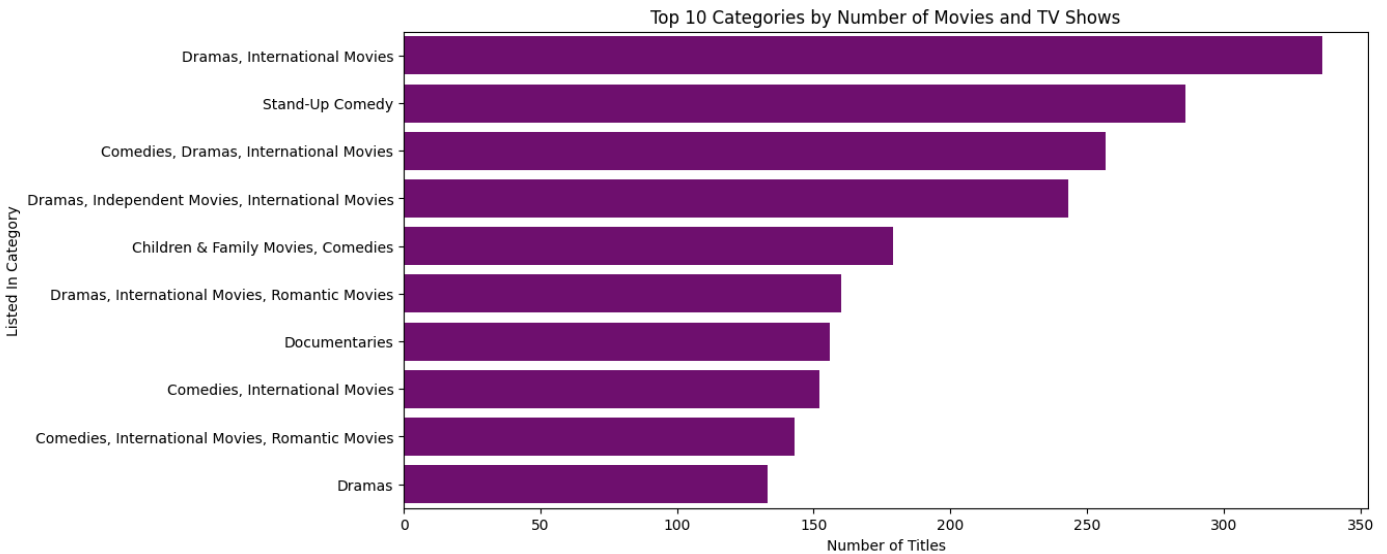


- This line plot displays the average length of movies and TV shows on Netflix over time. It has the year on the x-axis and average length in minutes on the y-axis.
- One notices the overall uptrend in the average length of movies and TV shows over time. Some reasons for this might include the increased popularity of streaming services, the growing demand for longer-form content or many other reasons.
- The increase in movie duration on average has been more significant. Much of the reason may be that movies, typically, are more costly to produce than television shows, and so there are more available resources.
- Therefore, the average length of movies and TV shows takes a nosedive in 2020. This is probably due to the pandemic that delayed and/or suspended most productions.

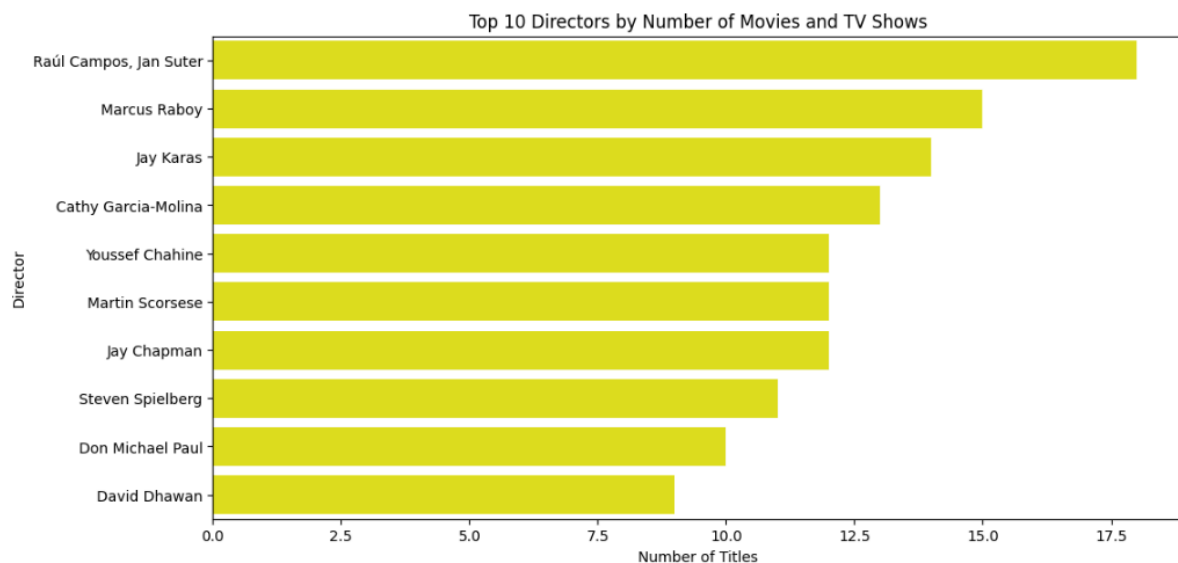
2
(4)



- This plot shows the number of movies and TV shows that are listed in each category. Thus, the most common categories are "International Movies", "Dramas", and "Comedies". This means that Netflix has a lot to offer its users in terms of variety and that there is something for everyone to watch.
- Also, the plot shows there are a pretty small number of movies and TV shows that are listed in the following categories: "Anime Features", "Classic Movies", and "Cult Movies". This can imply that these categories are not that popular amongst the Netflix customer base or, as another option, it may mean that Netflix just doesn't have as much content in these categories.



- This graph shows the most weighted directors to have movies and/or TV shows on Netflix with Jan Suter being the most with 39 titles. Raúl Campos comes second with 26 movies and Jan Kounen follows closely at third with 25.
- From the plot, it is also evident that most of the top 10 are male. In this case, the plot implies that the film and television industry has a gender disparity in terms of hiring where males are assigned the role of directors more than women.
- What is more, the plot shows that they are of different countries origin: the United States, Spain, France, and Germany. That means Netflix tries to do its best so that users would be able to watch the content of the great variety of cultures and points of view.



Analysis of Breast Cancer Wisconsin

Introduction:(Part-I)

The Breast Cancer Wisconsin is one of the dataset which is commonly used for the analysis and Prediction Purposes of breast Cancer Diagnosis. It contains a variety of features that describe the characteristics of cell nuclei present in the biopsy samples. The dataset is used mostly for the classification techniques between malignant and benign tumors.

Dataset Overview:

The dataset contains total entries of 569 and total columns of 33.

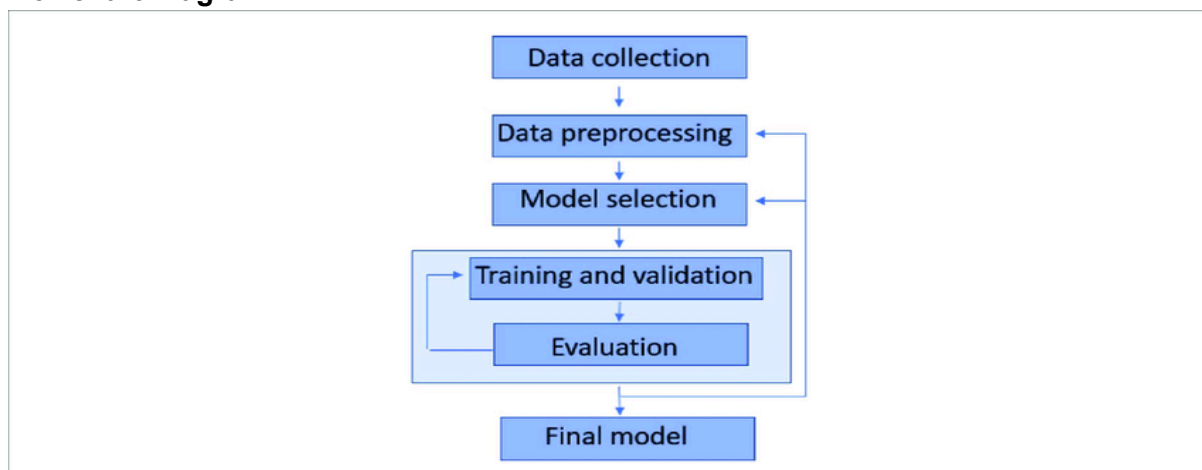
Key columns:

- Id: This is the unique identifier for the each column
- Diagnosis: There is a Diagnosis sample with M as Malignant and B stands for benign
- Features: There are overall 30 digitized images of fine needle aspirate of a breast mass. These features describe the characteristics of the cell nuclei present in the image.
- Some columns are:
 - radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean.
- Unnamed:32: A column with all values missing, which can be dropped from the analysis.

Purpose:

The dataset is to predict the classification and Prediction models and to determine the statistical Analysis of the data.

Flowchart Diagram:



Import Libraries:(Part 1 and Part 2):

```
# Import all the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
```

Data PreProcessing:

1. Identifying the Missing Values:

- Using the breast_cancer dataset the code checks columns for missing values. `isnull().sum()`. This function returns the number of missing values in each column. Telling you where you have missing values is useful to you to know where to impute data into.

2. Dropping Unnamed Column:

- A column (below an unnamed extra column, probably added because of an extra comma at the end of the CSV) is dropped from the dataset. This is achieved with the `breast_cancer` command. (`columns = ['Unnamed: 32']`) As no non-null values are present in this column(That i remember initial data review), dropping such columns is safe here.

3. Dropping Rows with Missing Data:

- Fix whatever needs to be fixed in the problematic column using all the methods available, and once we do that we go and drop by row anything that remains. The aim of this command is to drop any rows that have an NA somewhere, that way only complete rows go through the model training process.

4. Diagnosis Mapping to Numeric Values

- The numerical data in the diagnosis column stored as strings ('M' stands for malignant, 'B' for benign) are converted to the binary format. The transformation is essential for machine learning models that require numerical input. The mapping is carried out using `breast_cancer_cleaned['diagnosis']=breast_cancer_cleaned['diagnosis'].map(RENAMING).map({'M': 1, 'B': 0}).astype(float)`. Convert 'M' to 1, 'B' to 0 and then change the dtype to float to save time on preprocessing of data with changing dtypes in some columns after generating csv files.

Statistics:

The `.describe()` function in pandas calculates the description statistics of variables that hold numeric data values in a dataframe. it also ignores nan values by default. Typically the result will contain:

- count: It obtains the count of non-null records in each column
- mean: Calculate the mean for request columns.
- std standard deviation: It measures the amount of variation or dispersion of a set of values.
- Min: The minimum value of each column in a data frame.
- 25% first quartile: The value below which 25 percent of each column of data lies.
- 50% median : The middle value of each column of data.
- 75% third quartile: The value below which 75 percent of each column of data lies.
- Max: maximum The maximum value in each column.

The below are some of the outputs.

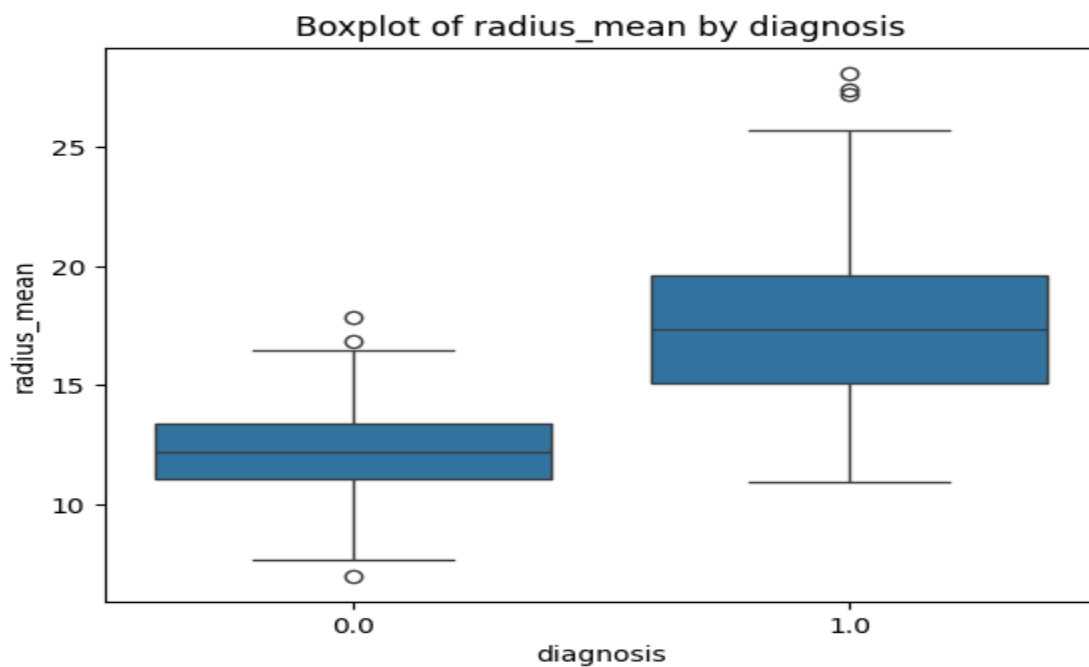
id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
count	5.69E+02	569	569	569	569	569
mean	3.04E+07	0.372583	14.127292	19.289649	91.969033	654.889104
std	1.25E+08	0.483918	3.524049	4.301036	24.298981	351.914129
min	8.67E+03	0	6.981	9.71	43.79	143.5
25%	8.69E+05	0	11.7	16.17	75.17	420.3
50%	9.06E+05	0	13.37	18.84	86.24	551.1

75%	8.81E+06	1	15.78	21.8	104.1	782.7
max	9.11E+08	1	28.11	39.28	188.5	2501

Data Visualization:

1. Boxplot of radius_mean by diagnosis:

- It can be seen quite clearly from the boxplot that the average size of the radii for malignant tumors is generally larger than those for benign ones. The median and the spread IQR, are larger in the malignant cases.
- The presence of outliers, throughout the malignant group in particular, suggests that some malignant tumors have much larger values of radius_mean than is typical, perhaps indicating a more aggressive or advanced tumor state.
- This is a key visualization in the support of diagnosis in a medical sense, due to the fact that, in this plot, there is clear visual differentiation between radii of benign and malignant breast tumors and hence help in strategies towards predictive modeling and diagnosis.



2. Violin Plot by Texture Mean by Diagnosis:

- Distribution differences: Malignant tumors will on the average have a wider distribution of the means of the texture features, shifted towards higher values. That can point to tumor texture being more heterogeneous in malignant cases than in benign ones.
- Texture Mean as a Diagnosis Indicator: Texture Mean as a Diagnostic Indicator The plot implies that the mean of the texture can be used as a diagnostic indicator also, where one can easily find the difference in its texture properties for benign versus malignant tumor.
- Implications for Diagnosis: Through this scatter plot, we realize that it is not just size - as captured in the variable `radius_mean`, but even the texture of tumors holds importance in classifying a case as benign or malignant. Such information, while to be exploited by clinicians, but more so by procedures of automated diagnosis, which would enable a more accurate prediction of the nature of a tumor.



Correlation Matrix and Heat Map:

1. Color Coding and Correlation Strength

- The color scale on the right side of the heatmap runs from dark red to white with dark red representing the strongest positive correlation, if not 1.0, and white for the most negative or closest to -1.0, and that intermediate shades of pink to light red for more or less high positive correlations.
- A value near 0, represented by a neutral cream color, shows that the paired features have no linear relationship.

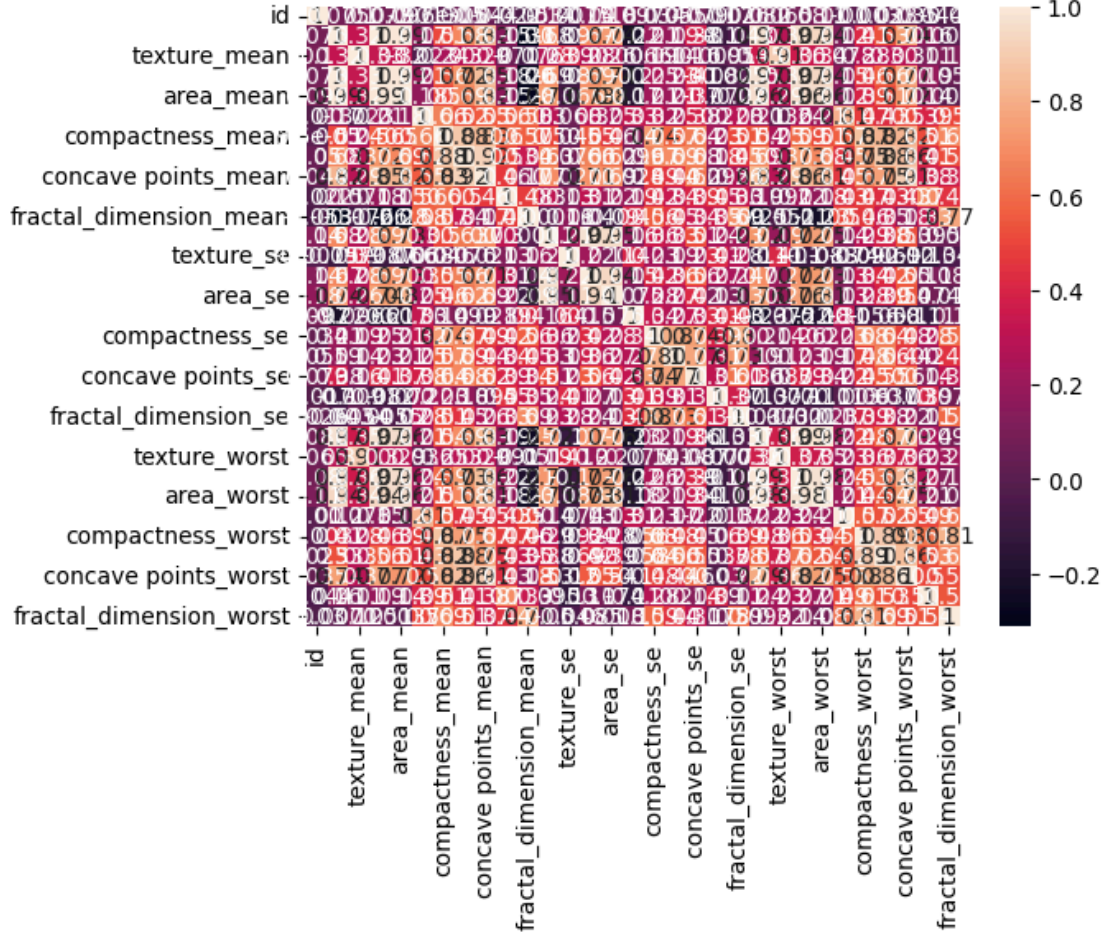
2. High Correlation Clusters

- There are several big dark red squares on the heat map; these are in clusters between those features all measuring the same aspects of the tumors but on different scales, namely: mean, worst and SE, standard error.
- radius_mean, perimeter_mean, area_mean are highly correlated all of these are describing size.
- radius_worst, perimeter_worst, and area_worst all show a very similar high-correlation pattern, indicating the various measurement concepts of tumour severity or progression have very consistent relationships to each other.
- Finally, compactness_mean, concavity_mean, and concave points_mean all show similar highs in their correlations with each other. This indicates that these textural features vary together as the physical characteristics of the tumour change or develop.

3. Possible Existence of Multicollinearity in Predictive Model Making

- High correlations among multiple feature sets imply that one set is multicollinear with respect to another if used together in some predictive model like linear regression. This can result in skewed results from this model or in inaccurate coefficient estimates.
- Then it would be careful about choosing the features or feature reduction techniques PCA - Principal Component Analysis, to deal with multicollinearity so that one has more robust and interpretable models.

Correlation matrix



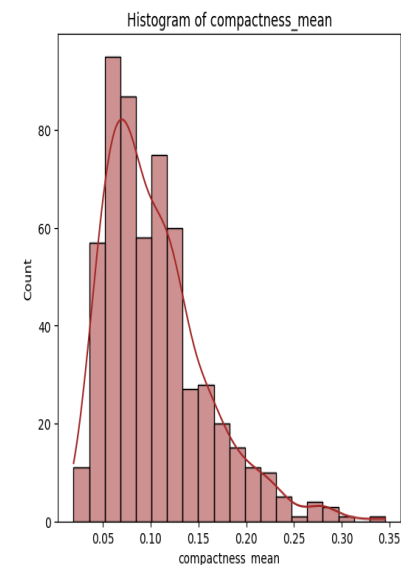
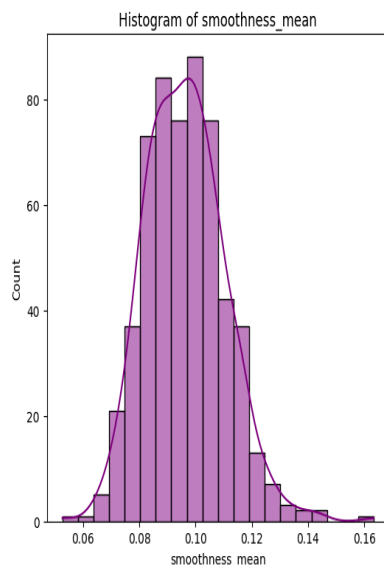
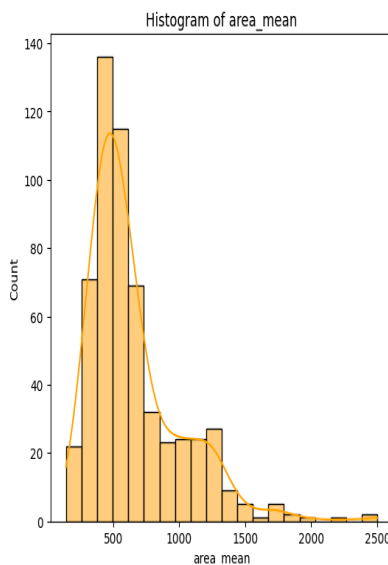
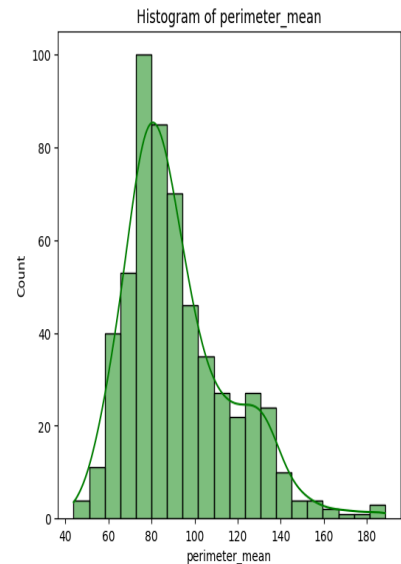
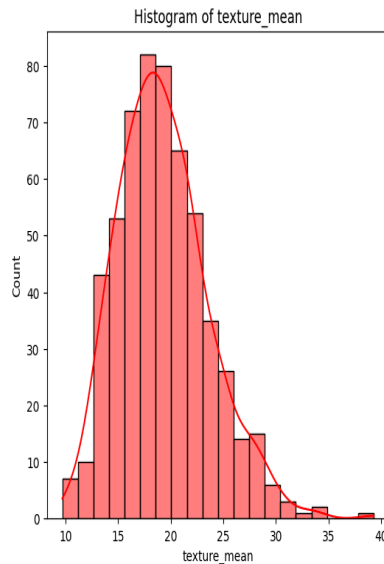
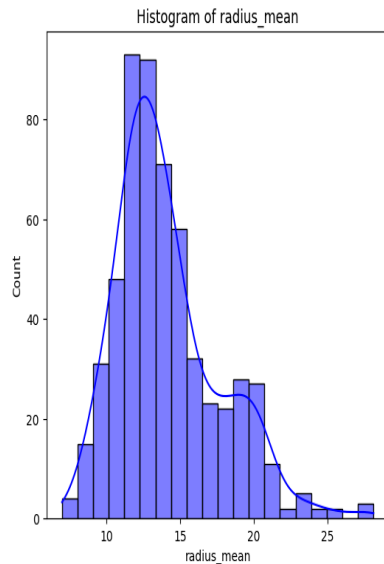
Histograms:

General Insights:

Histograms above all give a touch of a picture about how these measures are distributed in the dataset of samples of breast cancer.

Bimodality in radius mean and perimeter mean are especially interesting findings, suggesting that really there are two subgroups present in this data set probably benign and malignant corresponding.

Distributions are thus helpful in indicating the trends that the data possesses, and from the trends, proper statistical study and model formulation can be carried out.



BoxPlots:

1. Overall Box Plot Comparisons Across

Median: In most of the variables, it is easy to note that the median for value 1.0 in the malignant category is way higher compared to zero in the benign category. This means most of these measurements tend to be bigger in malignant tumors.

Spread and IQR: Most variables have a larger spread and IQR in the malignant group. At the same time, benign shows a smaller spread, indicating the divergence of tumor characteristics in the malignant cases.

Outliers: In both the categories, for all the features there exists notable outliers that show that there are extremely high or low values present than the usual range of the data

2. Individual Feature Analysis

Radius Mean : Boxplot-Blue

Benign: Closer clustering and lower median implies that the tumor size is smaller.

Malignant: Higher median as well as spread shows that the tumor sizes are larger in size which is vital in knowing about the aggressiveness of the cancer

Texture Mean : Boxplot-Green :

Benign: Smaller median and fewer high outliers

Malignant: High variance and more outliers could be due to more inhomogeneous and irregular texture of the tumor

Perimeter Mean Boxplot in red

Benign: Uniformly smaller with compact IQR complement the observation for radius mean

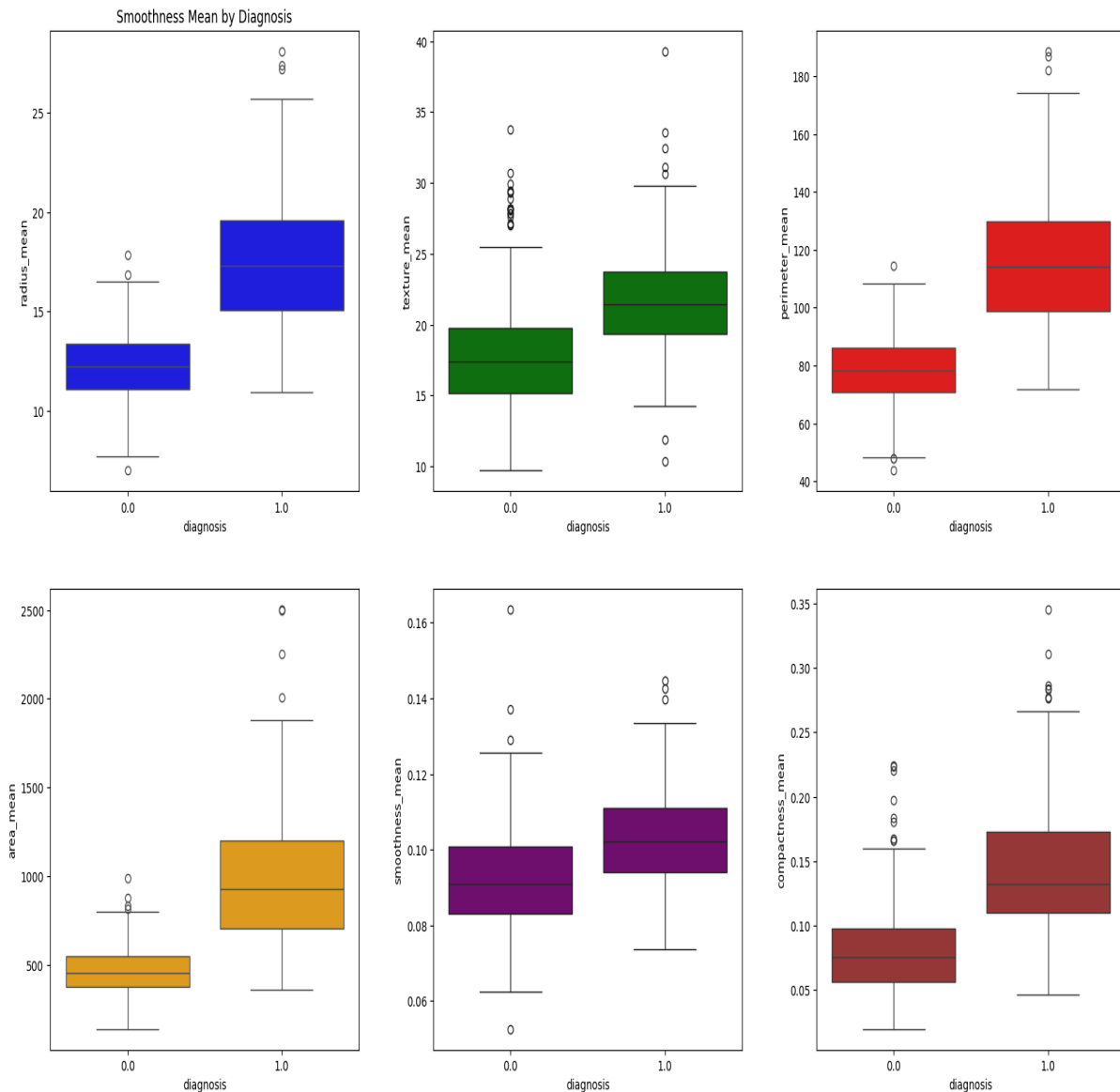
Malign An: Much larger perimeter values with extensive range supports the trend observed both by radius and area

3. Implications of the Findings

Diagnostic Value: This difference in medians and distributions between benign and malignant tumors along features like radius, texture and perimeter proves their role in diagnosis.

Treatment Planning: The extent and nature of variation, in the case of malignant tumors, might be useful in planning treatments since, from the exploratory phase, the features were noted to have higher values indicating advanced disease.

Research and Analysis Outliers in general, most importantly the presence of outliers in itself, are crucial to point out further research into possible causes that lead to extreme values and on general trends in distribution of its effects on patients.



Part 2: Linear Regression Analysis:

Linear Regression:

Linear regression is a method in statistics used to model and analyze the relationship that exists between a dependent variable and one or more independent variables. It tries to establish a linear relationship between variables by fitting a linear equation to observed data.

1. Loss Value and Weight Vector:

Weight Vector:

The weights are computed on the basis of the `ols_regression` which is called the Ordinary Least Square method using the formula $w = (X^T X)^{-1} X^T y$. The resulting vector weights include both the intercept and the slope coefficient for the radius mean variable after normalization.

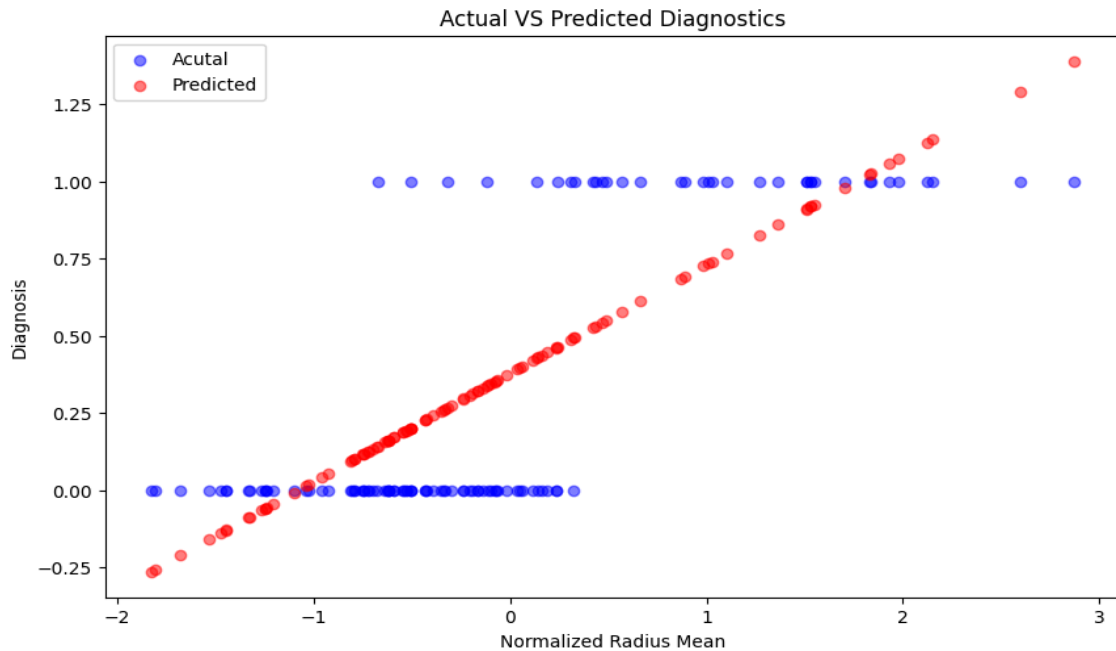
Output: `Calculated Weights: [0.37897371 0.35176842]`

Loss Value(RSME): The Loss Value here is calculated using the root mean square error formula which is a common way to measure the difference between value predicted by a model and the values observed. The RMSE for the test set is computed in the code and displayed at the end.

Output: `rmse: 0.30314124083531985`

2. Plot Comparing the Predictions VS The Actual Test Data:

The following plot is generated for visual comparison between actual and predicted values based on diagnosis with values predicted by the radius mean normalized. This plot will let us know how well our model is performing in regards to preciseness of the predictions. This has actual data points in blue and predicted points in red overlaid in a scatter plot to highlight discrepancies and agreements.



Discussion of the Advantages/Disadvantages of Use of OLS Estimate:

Advantages:

Efficiency: Computationally, OLS is very simple and hence efficient for small to moderate size data sets.

Interpretability: The coefficients of the model are easily interpretable in terms of its effect on the response variable.

Best Linear Unbiased Estimator: Under the classical assumptions of the linear model, the OLS estimates are BLUE, that is, the most efficient among all linear unbiased estimators.

Disadvantages

Sensitivity to Outliers: OLS can be easily influenced by outliers since it seeks to minimize the sum of squared errors. These kinds of errors are significantly raised by outliers.

Assumption Dependent: Performance and validity are the essence of OLS that relies on several assumptions- linearity, independence, homoscedasticity, and normality of errors. It's self-explained that the estimates might be wrong and inconsistent if these assumptions are breached.

Binary Outcome Limitation: From your code, it appears you have binary target variable 'diagnosis' which justifies the absence of any breast cancer disease in that patient. Now OLS are not a good choice for binary outcomes. You will get predicted values all over the

place, greater than a 1 or less than a 0, which in a binary context, this is not interpretable. Here, logistic regression would be the best option.

Part 3: Ridge Regression Analysis:

Ridge Regression:

Ridge regression, like Tikhonov regularization, is multiple regression data analysis in the case of suffering from multicollinearity. If the independent variables were highly correlated, then the least squares estimate becomes very sensitive to random errors in the observed response; this leads to overfitting and hence poor predictive performance.

Loss Value and Weight Vector:

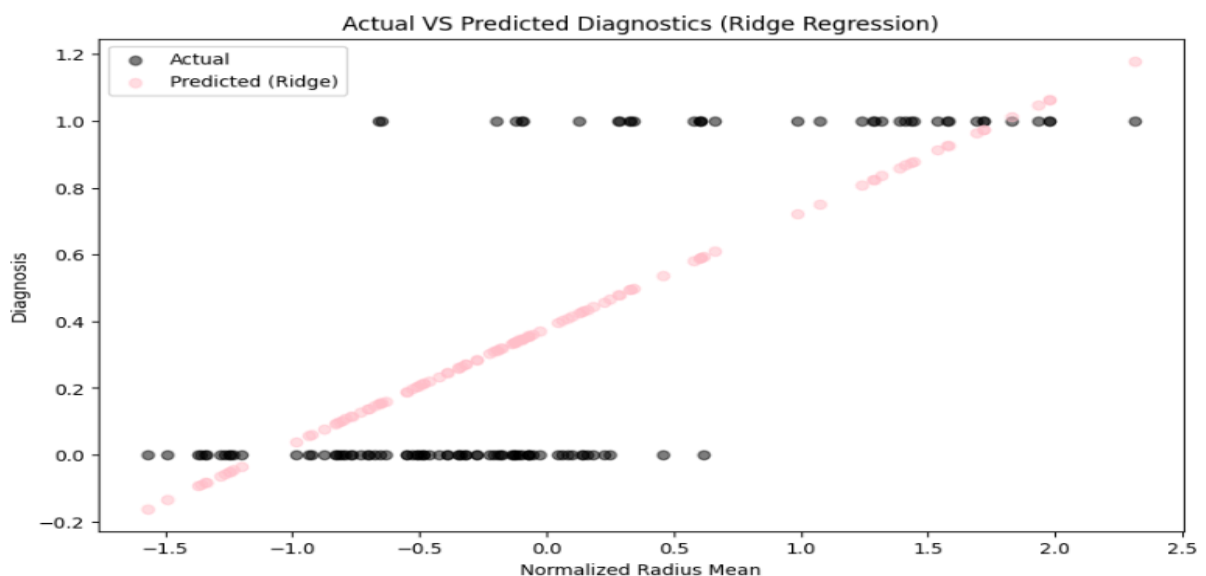
Loss Value:

```
RMSE (Ridge Regression): 0.32093312591789425
```

Weight Vector:

```
Calculated Ridge Regression Weights: [0.38060891 0.34577013]
```

plot comparing the predictions vs the actual test data



The plot shows that, by and large, the fitted values are very close to the observed values. However, there are some cases in which the difference between the predicted and actual values is quite far apart.

This could be most probably due to the fact that ridge regression is a linear model and the data may not exhibit perfect linearity.

The plot suggests that this will give a reasonable way to predict diagnosis based on mean radius using ridge regression.

Discuss the difference between Linear and Ridge regressions. What is the main motivation of using L2 regularization?

- Linear regression assumes data is linearly separable and linearity in the relationship between the features and the target variable; that is, the straight-line of the model.
- Ridge regression assumes a lack of linearity in the data separation process and nonlinear relations among features and a target variable. This thus means that the model cannot be represented as a straight line.
- L2 regularization will help prevent overfitting of the data. Overfitting is when a model has then become too complex, and it starts learning from the noise in the data. This may result in bad performance on predictions for new data. This will avoid overfitting by penalizing large weights by the model with the help of L2 regularization.
- This will bias the model toward a simpler solution, less likely to overfit the data.
- In general, as a methodology, ridge regression is more robust than linear regression. It's less prone to overfitting and able to handle many nonlinear relationships between features and a target variable.

Bonus Point: Gradient Descent:

Gradient Descent:

Gradient Descent is an optimization algorithm used for the minimization of a function by iteratively moving toward the steepest descent as defined by the negative of the gradient. In machine learning, it is normally used to optimize loss functions of models for linear and logistic regression and for neural networks. The idea is to find those values of parameters or weights that turn this cost function into a minimum, which measures how well predictions fit the real data.

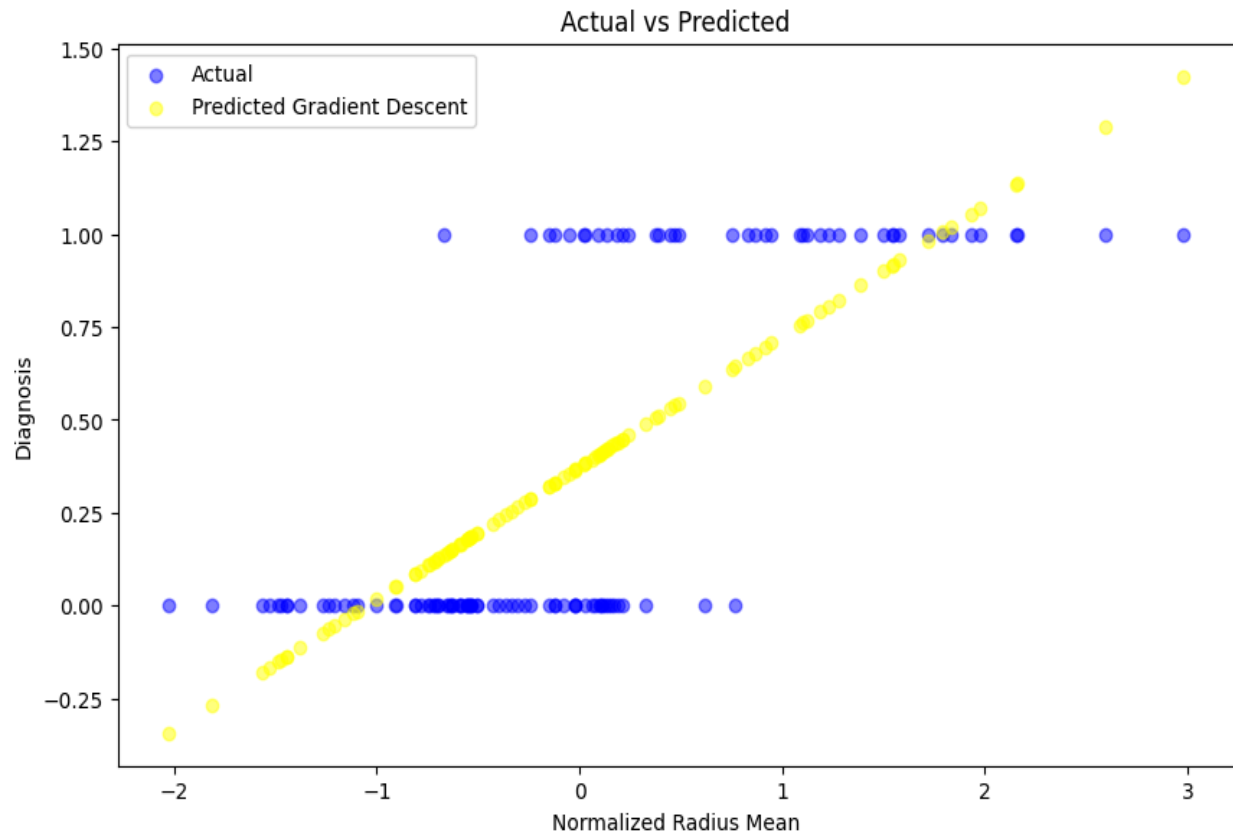
Calculated Weights:

```
Calculated Weights of Gradient Descent: [0.37293813 0.3530316 ]
```

Weight Vector:

RMSE (Gradient Descent): 0.3318704710282012

Analysis of the Graph:



- A model that picks up the trend of linearity between features and targets in that it systematically under- or over-estimates on different segments of data in a consistent manner, hence proving possible systematic bias or underfitting
- Most data points are concentrated in the middle range, where the model has densely predicted. Quantitative metrics like RMSE have to be considered to be very accurate in measuring and probably improving accuracy.

Contribution:

Team Member	Assignment Part	Contribution
Bhavitha Rapolu	Part 1	50%
	Part. 2	50%
	Part 3	50%
	Bonus	50%
Aditya Srivatsav Lolla	Part 1	50%
	Part. 2	50%
	Part 3	50%
	Bonus	50%

Note: There is an equal contribution from both of the teammates.