Breast Cancer Wisconsin Dataset Analysis & Dashboard Summary

Understand the Dataset

This dataset contains medical measurements from breast cancer biopsies. There are 569 rows and 32 columns, including an ID, a diagnosis column (target), and 30 numeric features that describe the shape and structure of cell nuclei. Each feature has three types: mean, standard error (SE), and worst (maximum).

1. What does each column mean?

1) ID number : A unique number assigned to each patient (not useful for analysis).
2) Diagnosis (M = malignant, B = benign) - Target variable since the final motto is understand patterns in malignant vs benign tumors
3)Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Class distribution: 357 benign, 212 malignant

2. What are the target classes? (diagnosis: M = Malignant, B = Benign)

The target column (Diagnosis) has two possible values:

- M (Malignant) → (212 cases).
- B (Benign) → (357 cases).

Data Cleaning & Preprocessing

1. Check for missing values -

   No missing values as given in the description. Also verified it using
   df.isnull().sum()

2. Remove ID column -

   ```
   (df = df.drop(columns=["id"])
   ```

3. Encode target values if needed (e.g., M = 1, B = 0)

   ```
   df['diagnosis'] = df['diagnosis'].map({'M': 1, 'B': 0})
   ```

Exploratory Data Analysis (EDA):

Perform visual and statistical analysis to answer:

1. What are the most important features that differentiate malignant from benign?

   a. Radius (mean of distances from center to points on the perimeter)

   b. Area

   c. Perimeter

   d. Concavity (severity of concave portions of the contour)

   e. Texture (standard deviation of gray-scale values)

2. How are features like radius_mean, texture_mean, and area_mean distributed for each diagnosis?

   Radius_mean, **Texture_mean**, and **Area_mean** are generally **higher for malignant tumors**, indicating that tumors with **larger sizes** and **rougher textures** are more likely to be malignant.

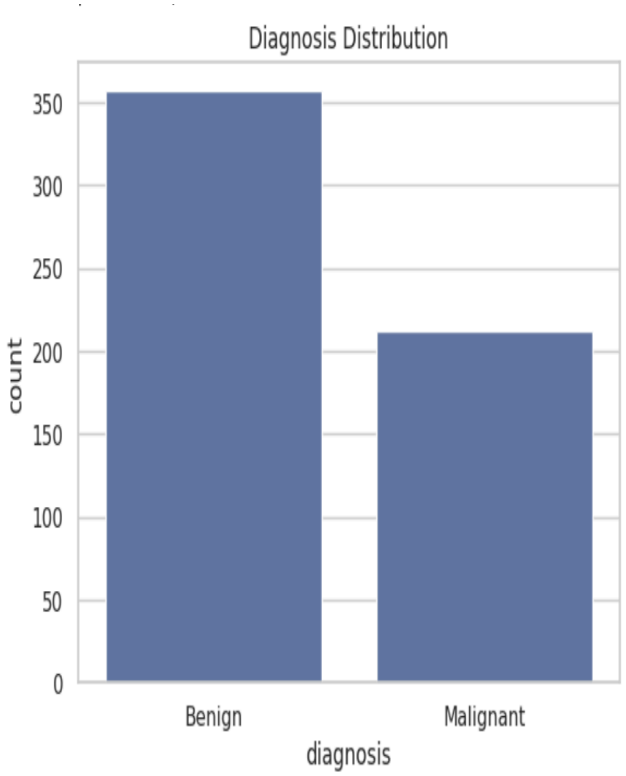
3. Are there any strong correlations between features?

- There are **strong correlations** between features like **radius_mean**, **perimeter_mean**, and **area_mean**, as well as their corresponding **worst features**.
- These strong correlations suggest that these features might be measuring similar aspects of tumor size and shape.

Plots utilized for Analysis

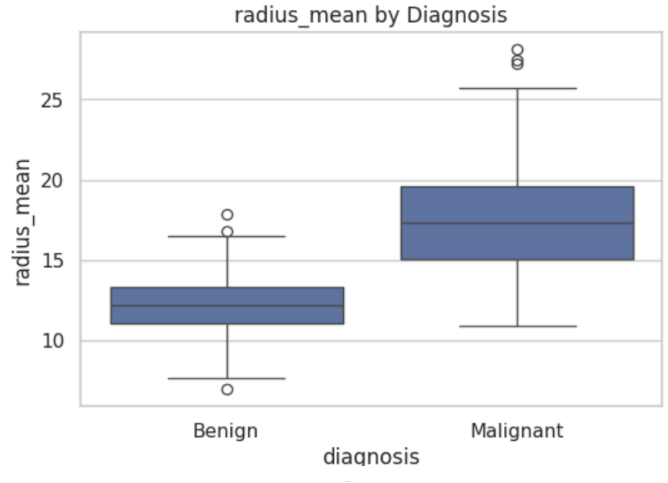
Bar chart

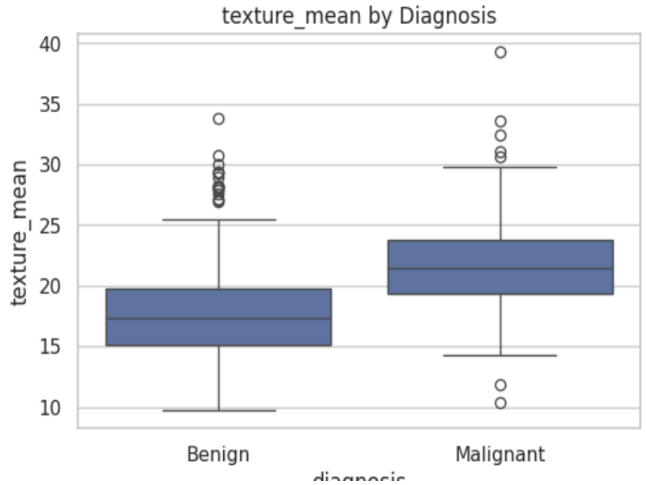
Showing the number of benign and malignant cases.

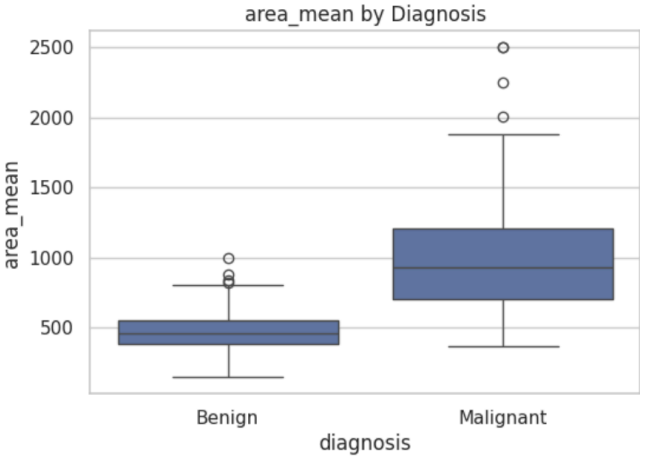


Boxplots

to **compare distributions of the key features.**

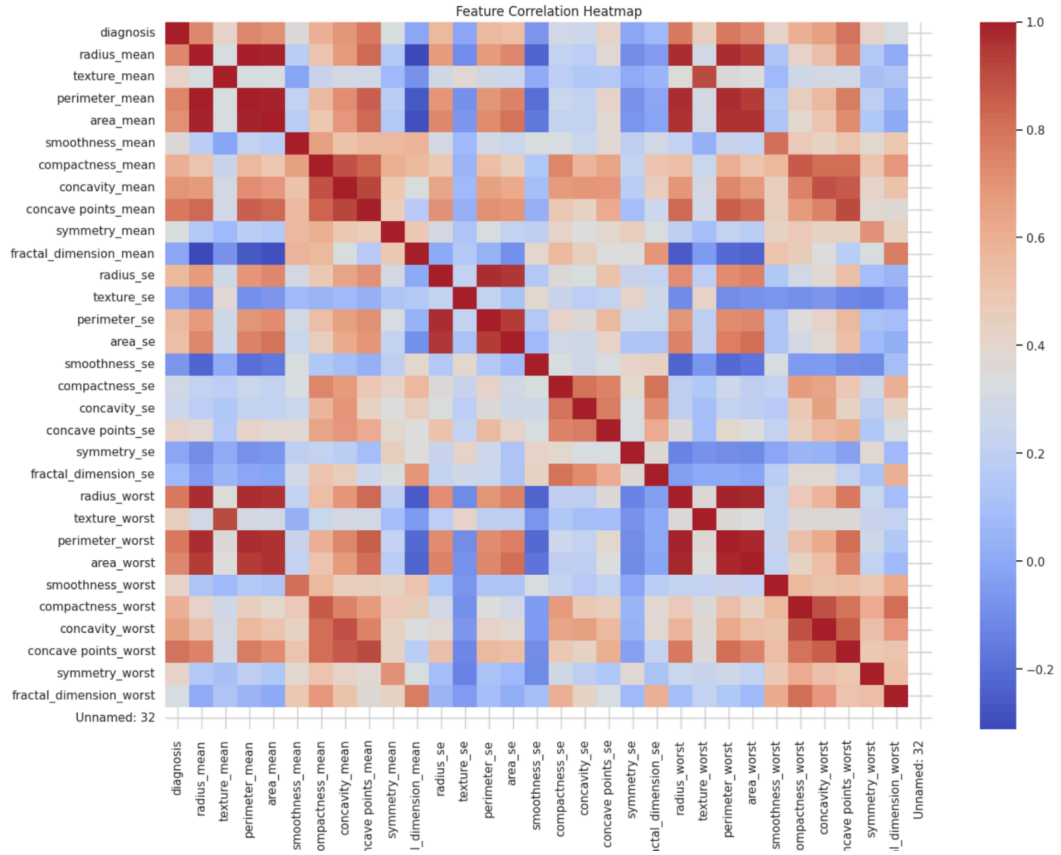




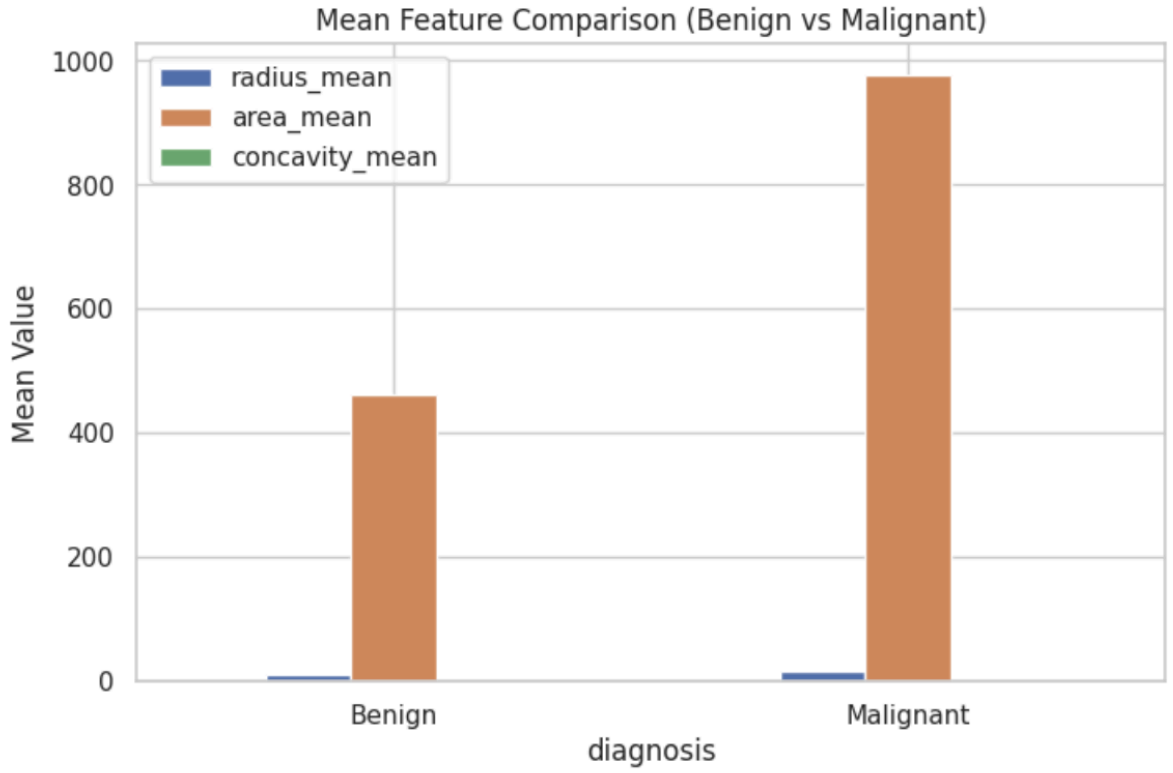


Heatmap

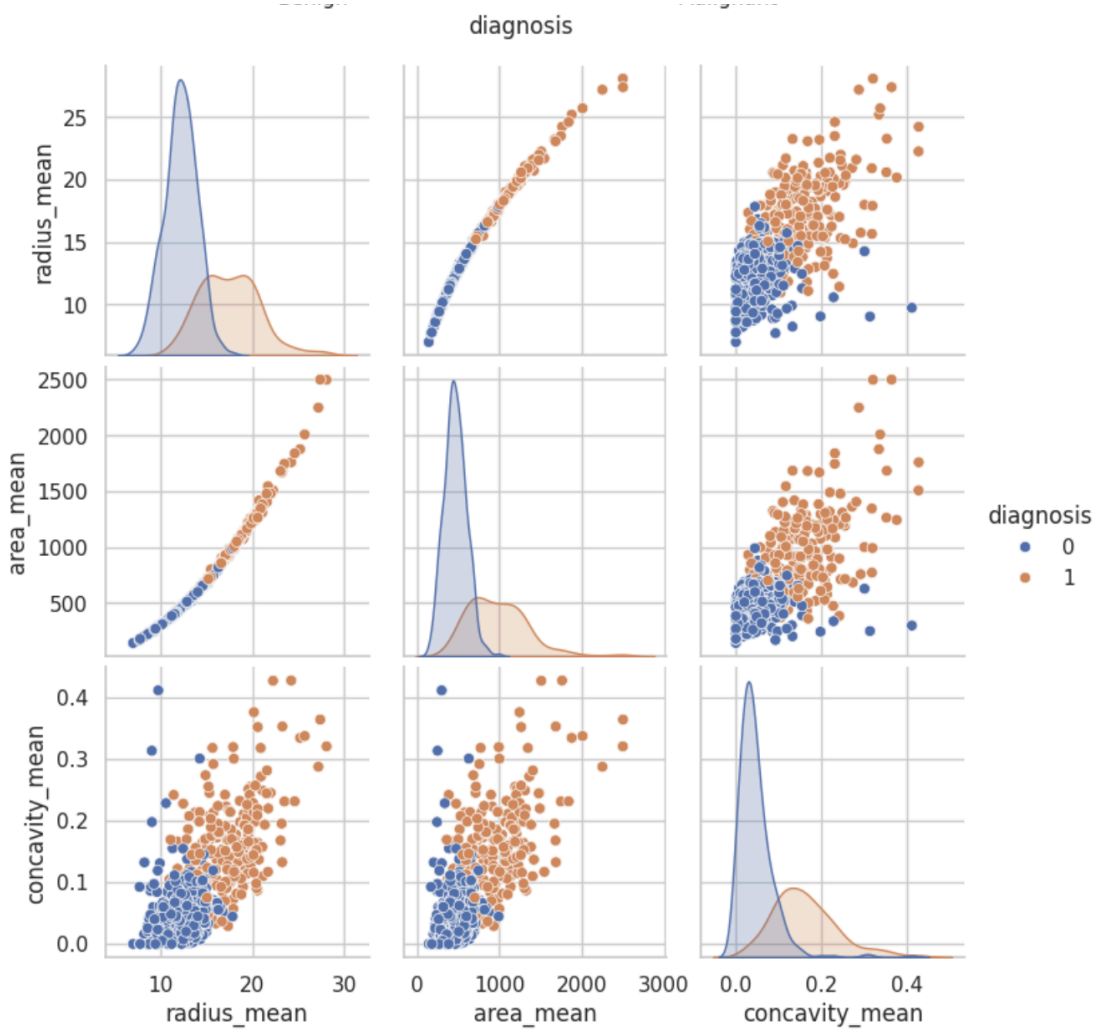
to findout the correlations between the features



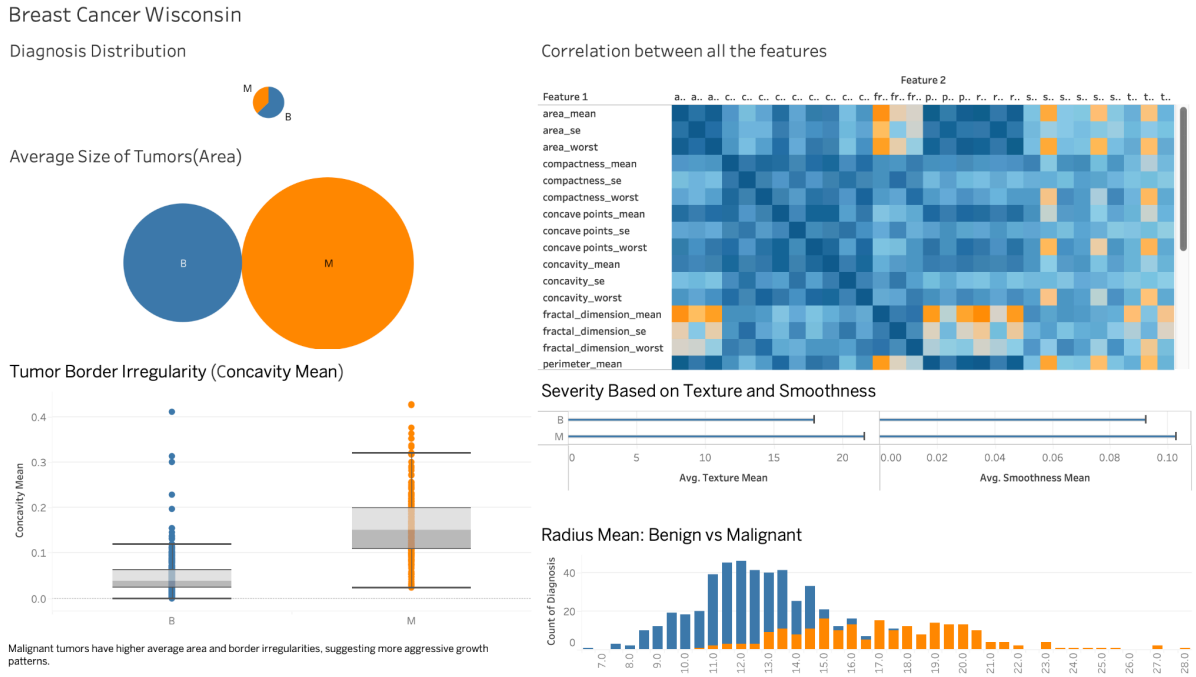
4. Bar chart



Pairplot


diagnosis
radius_mean
area_mean
concavity_mean
diagnosis
0
1
10
20
30
0
1000
2000
3000
0.0
0.2
0.4

Dashboard



This chart shows how the dataset is split:

🟦 **Benign (B)**: Non-cancerous tumors
🟧 **Malignant (M)**: Cancerous tumors

- **Malignant tumors are generally larger**, with higher radius and area.
- They have **more irregular, concave borders**.
- Features like **texture and smoothness** also tend to differ in subtle but consistent ways.
- The heatmap reveals that **many features are highly related**, especially size-based ones.
- These patterns can help doctors or machine learning models **predict malignancy** earlier.

Insight Summary – Breast Cancer Wisconsin Dataset Analysis

This analysis explored the Breast Cancer Wisconsin dataset to understand the differences between **benign** and **malignant** tumors. Using visualizations and statistical comparisons, several patterns were identified that may help support **early detection and diagnosis**.

Key Findings:

1. **Tumor Size Is a Strong Indicator**
   Malignant tumors tend to be **larger in radius, area, and perimeter** compared to benign ones. This was clearly shown in the **bubble chart (area)** and **radius histogram**, where most malignant tumors had higher values.

2. **Border Irregularity Matters**
   Features like **concavity_mean** showed that **malignant tumors have more irregular and bumpy borders**. This was demonstrated in the **box plot**, where malignant cases had significantly higher concavity values.

3. **Surface Characteristics Vary**
   Malignant tumors also showed **slightly higher average texture and smoothness**, suggesting **greater structural complexity or abnormalities** on the tumor surface.

4. **Strong Correlation Between Features**
   The heatmap revealed that many features are **highly correlated**, especially those related to size (e.g., radius, area, and perimeter) and shape (e.g., concavity and concave points).

Recommended Focus Features for Early Detection:

Based on this analysis, the following features are most useful for identifying potentially malignant tumors:

- **Radius Mean** – strong size-based separator
- **Area Mean** – reinforces the importance of tumor size
- **Concavity Mean** – highlights border irregularities associated with malignancy

Conclusion:

By examining key size, shape, and surface features, this analysis shows clear differences between benign and malignant tumors. These patterns can help inform medical professionals and may support the development of predictive models for **early breast cancer detection**