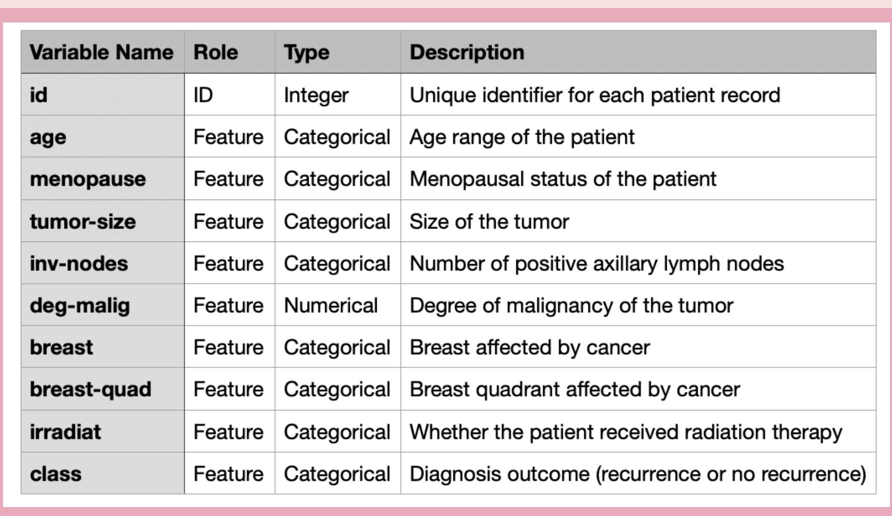
BREAST CANCER

Objectives:

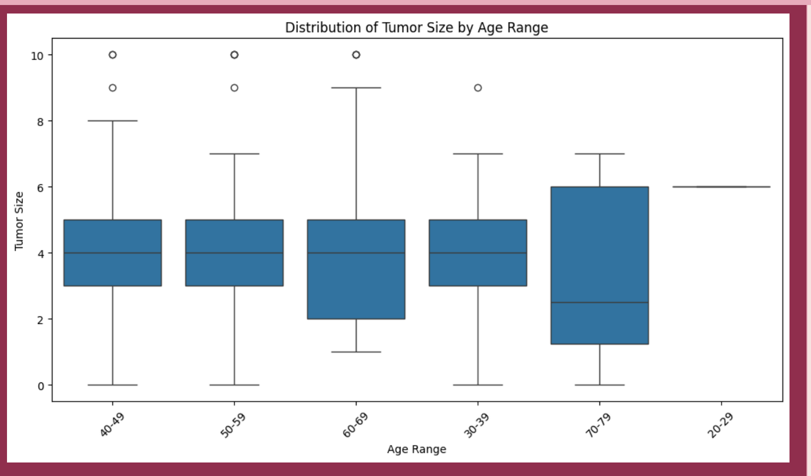
- Analyze Age Distribution: Explore the distribution of breast cancer cases across different age ranges to identify high-risk groups.
- Investigate Tumor Size: Examine how tumor size varies with age and its implications on malignancy.
- Visualize Correlations: Identify key correlations between clinical features and outcomes to guide future research.

What I found during MetaData

1. Entries and Features: The dataset contains 272 entries and 10 features, including age, tumor size, and degree of malignancy.
2. Basic Statistics: The degree of malignancy has a mean of 2.06, with a range from 1 to 3.
3. Visualization: I plotted histogram and box plot illustrate the distribution of age and tumor size across different demographics, providing insights into breast cancer patterns.



Variable Name	Role	Type	Description
id	ID	Integer	Unique identifier for each patient record
age	Feature	Categorical	Age range of the patient
menopause	Feature	Categorical	Menopausal status of the patient
tumor-size	Feature	Categorical	Size of the tumor
inv-nodes	Feature	Categorical	Number of positive axillary lymph nodes
deg-malig	Feature	Numerical	Degree of malignancy of the tumor
breast	Feature	Categorical	Breast affected by cancer
breast-quad	Feature	Categorical	Breast quadrant affected by cancer
irradiat	Feature	Categorical	Whether the patient received radiation therapy
class	Feature	Categorical	Diagnosis outcome (recurrence or no recurrence)



The box plot reveals that older age groups, particularly those aged 70-79, tend to have larger tumor sizes. This suggests that age plays a significant role in tumor development and highlights the necessity for age-specific treatment strategies. Focusing on early detection and personalized care for older patients can lead to improved outcomes and more effective management of breast cancer.

SUMMARY

By analyzing the breast cancer dataset, we identified significant patterns that are not immediately obvious. The average degree of malignancy is relatively consistent across age ranges, suggesting that malignancy is not heavily influenced by age. This insight helps focus attention on other risk factors. In contrast, the tumor size plot shows that older age groups tend to have larger tumors, highlighting the need for targeted treatment approaches based on age.

In my breast cancer data analysis project, I began by exploring the dataset using a combination of shell scripting and Python. I first used awk and sed commands in the terminal to extract key columns and analyze the structure of the dataset. This initial step helped me understand the data's dimensions and allowed me to isolate specific entries for further analysis, such as focusing on particular age groups. In Google Colab, I leveraged Python libraries like Pandas, Matplotlib, and Seaborn to dive deeper into the data. By encoding categorical variables into numerical values using LabelEncoder, I prepared the data for a comprehensive analysis. This transformation was crucial because it enabled me to perform correlation analyses and visualize data trends effectively.

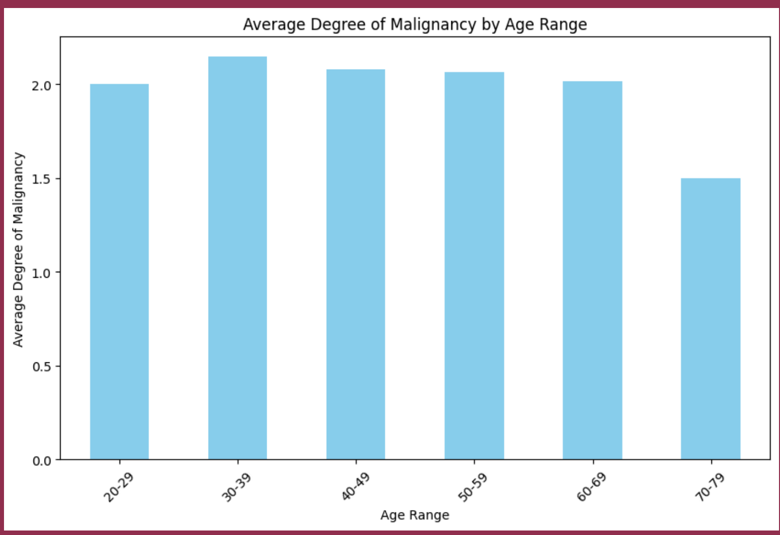
Tools:

First, I used Data Extraction Utilizing awk and sed commands to filter and inspect data columns, providing an initial understanding of dataset dimensions.
Second Data Transformation: Categorical variables were converted to numerical values using Python's LabelEncoder, enabling strong analysis.
Third, Employed Python libraries like Matplotlib and Seaborn in Google Colab to create meaningful visualizations that reveal hidden patterns.

python + colab

BASH & Shell Scripts

Statistics



The visualizations revealed significant insights into breast cancer patterns across different age groups. The histogram of age distribution highlighted that most cases occur in the 40-69 age range, emphasizing the importance of targeted screening for these groups. Additionally, the box plot of tumor size by age range showed that older age groups tend to have larger tumors, which could influence treatment decisions. Another point I want to focus on is the consistent degree of malignancy across age ranges, suggesting that age-related risk factors might impact tumor size more than malignancy degree. Overall, this project provided valuable insights into demographic and clinical factors associated with breast cancer, underscoring the importance of early detection and age-targeted interventions.

My future focus

Focus screening efforts on individuals aged 40-69, as this group shows the highest incidence of breast cancer. Early detection in this age range can lead to better outcomes.

Encourage continued research into the underlying causes of variations in tumor characteristics and malignancy. Investigating these factors can help identify new prevention and treatment approaches.