Bversity

1st Ever Virtual College for Biotech

Name : Aman Bashar

Contact: Amanbashar96@gmail.com

Mobile: 9085429899

Title: Python Exploratory Data analysis

---

**Answer the following questions below and upload them in the google form.**

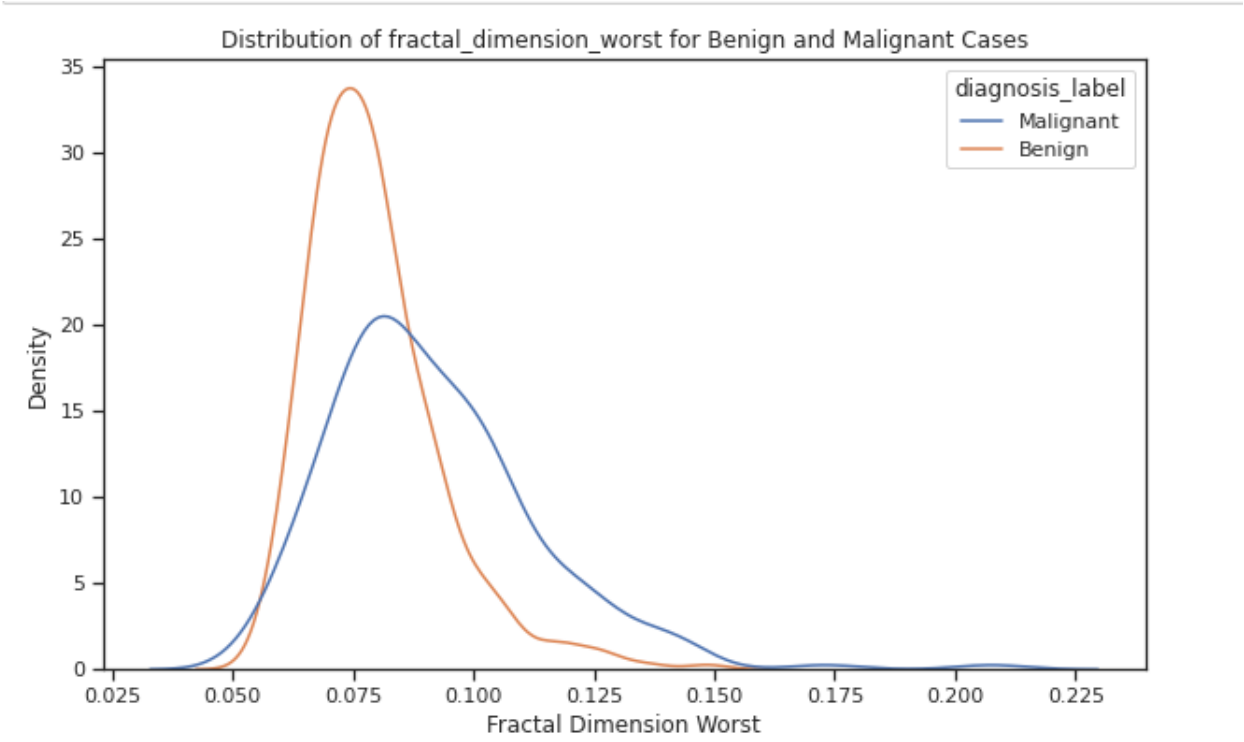
1. How does the distribution of feature “fractal_dimension_worst” differ between benign and malignant cases?
2. What is the range of values for the feature “radius_mean” and how skewed is its distribution?
3. Are there any outliers in feature “area_mean” and how might they affect analysis?
4. Based on the EDA, what factors seem to be most relevant to predicting breast cancer diagnosis?
5. What limitations are there in the data, and how might they affect our conclusions?

**Answer 1.**

**By Investigating the distribution of the feature 'fractal_dimension_worst' between benign and malignant cases reveals these insights:-**

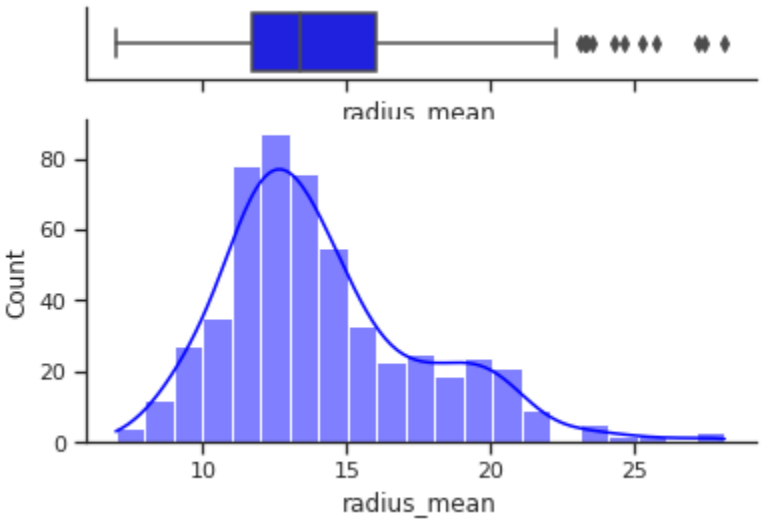
**Benign cases tend to concentrate around a certain range of fractal dimension values, while malignant cases exhibit a broader distribution with higher density in some regions.**

**This suggests that 'fractal_dimension_worst' may be a discriminative feature for distinguishing between benign and malignant tumors, that is it can be a potential indicator for identifying type of tumor.**

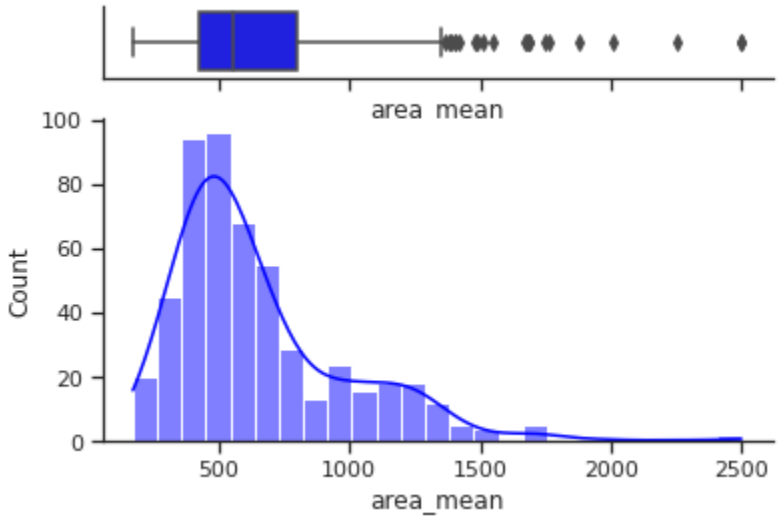


**Answer 2:**

**The range of values of the feature ' radius_mean' is 6.981 to 28.11 and Skewness of radius_mean is 0.998492, which is positively skewed.**



**Answer 3:**




**Yes, there are outliers present in the feature 'area_mean' as we can see it in the box plot. We also calculated that there are 39 outliers present in 'area_mean'.**

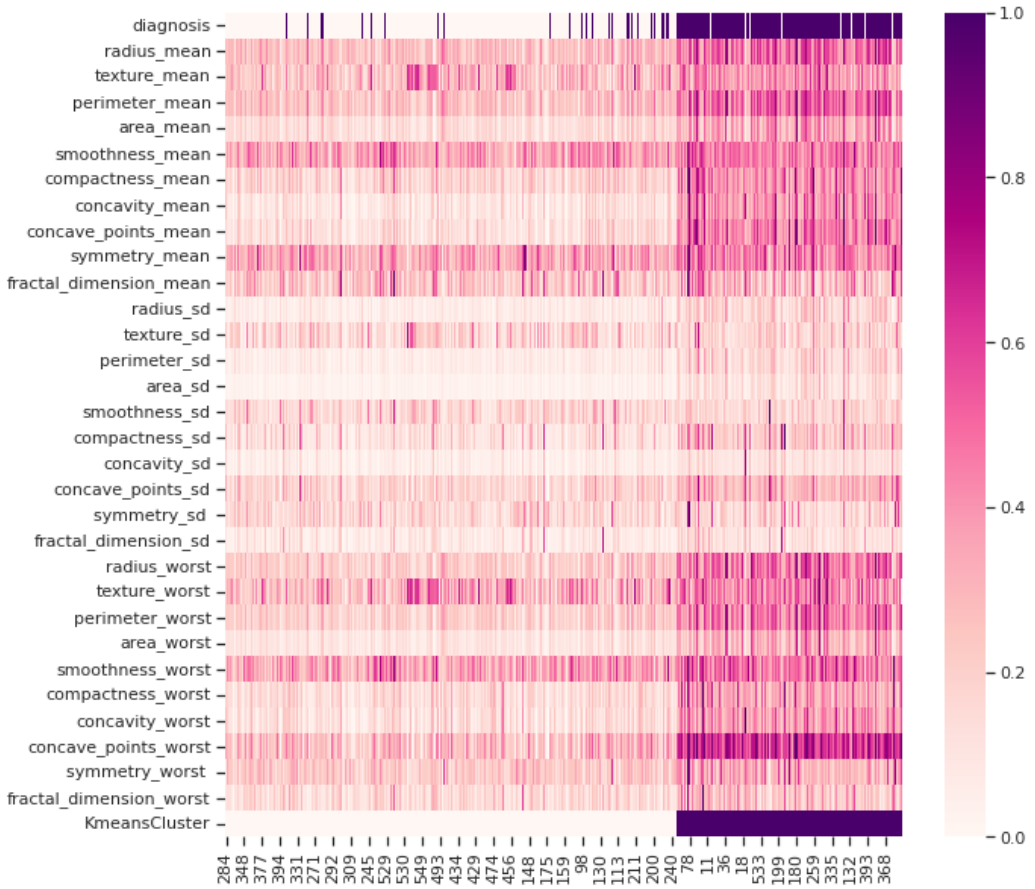
**Presence of outliers in dataset can have several impacts on the analysis, they are :-**

- **Outliers can impact summary statistics such as mean and Standard deviation.**
- **Outliers may violate the assumptions of certain statistical tests, leading to inaccurate results.**
- **Outliers can distort the shape and spread of data distributions, affecting the interpretation of patterns and relationships.**

**Answer 4:**

**Based on the correlation analysis of the data-set through heat-map,**
**we can conclude that following factors seem to be most relevant to predicting breast cancer diagnosis :-**

- **radius_mean**
- **perimeter_mean**
- **smoothness_mean**
- **symmetry_mean**
- **fractal_dimension_mean**
- **radius_worst**
- **texture_worst**
- **Smoothness_worst**
- **Concave_points_worst**



**Answer 5 :**

**limitation in the data and their impact on Conclusions:-**

**1. Imbalanced Data:-**

**we need to check if the dataset has an imbalance in the distribution of benign and malignant cases.**

**imbalance data can bias the model towards the majority class, potentially affecting the accuracy of predictions.**

**2. Missing Values:**

**We need to check if there's any missing values in the dataset.**

**This missing values can lead to biased or incomplete analysis.**

**3. Data Source and Bias-ness:-**

**Biases in data collection may introduce inaccuracies or limit the generalization of findings.**