Bversity
1st Ever Virtual College for Biotech

Name : snehal Ashok Kanase
Contact:snehalkanase156@gmail.com
Mobile:7397804023
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?

ANSWERS

question 1.

1. **Range:**

- The values range from 0.07259 to 0.1275.

2. **Distribution:**

- Benign cases (B) the values in the lower range (around 0.07259 to 0.08183).
- Malignant cases (M) have a wider range of values, extending up to 0.1275.

3. **Median (Middle Line in Box):**

- The median value for benign cases (B) appears to be lower, indicating a lower central tendency.
- The median value for malignant cases (M) appears to be higher, suggesting a higher central tendecy.

4.Box height (interquartile Range - IQR):

- The box for benign cases is smaller, showing a narrower range of values in the middle 50% of the data.
- The box for malignant cases is taller, showing a wider spread of values in the middle 50% of the data.

5. Whiskers (Vertical Lines Extending from Box):

- The whiskers for benign cases may be shorter, indicating a smaller range of values beyond the middle 50%.
- The whiskers for malignant cases may be longer, suggesting a larger range of values beyond the middle 50%.

6. Outliers (Individual Points Beyond Whiskers):

In Malignant cases the outliers beyond the whiskers, indicating potential extreme values.

The distribution of "fractal_dimension_worst" appears to differ between benign and malignant cases. Malignant cases exhibit a higher central tendency, a wider spread of values, and the presence of potential outliers compared to benign cases.

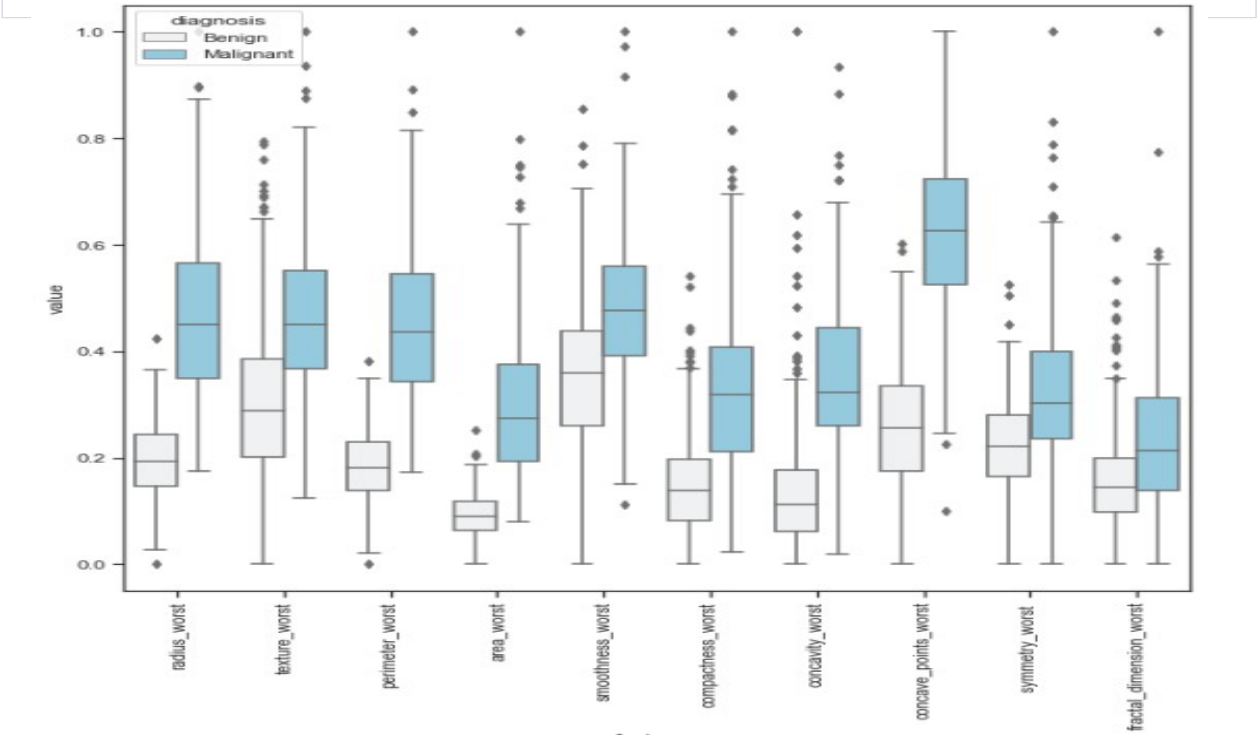
Question 2. Data:

count : 541.000000
mean : 14.175410
std : 3.527352
min :6.981000
25% :11.740000
50% :13.400000
75% :16.020000
max :28.110000

- The range of values is given by the difference between the maximum and minimum values.
- Range = Max - Min = 28.11 - 6.981 = 21.129.

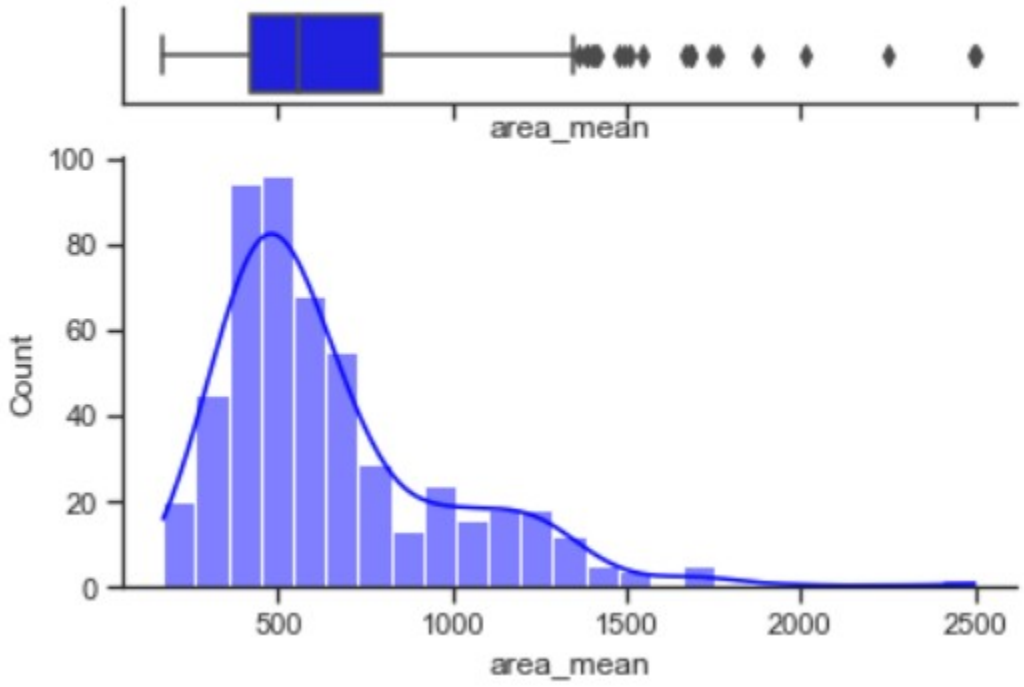
- **Skewness:**
  - The distribution skewness can be assessed based on the comparison of the mean, median (50%), and the quartiles (25%, 75%).
  - Mean = 14.175410
  - 50% (Median) = 13.4
  - 25% = 11.74
  - 75% = 16.02

From these values, I observe that the mean is slightly greater than the median, suggesting a slight rightward or positive skewness. The higher values (75th percentile) are more spread out than the lower values (25th percentile), contributing to the positive skewness.

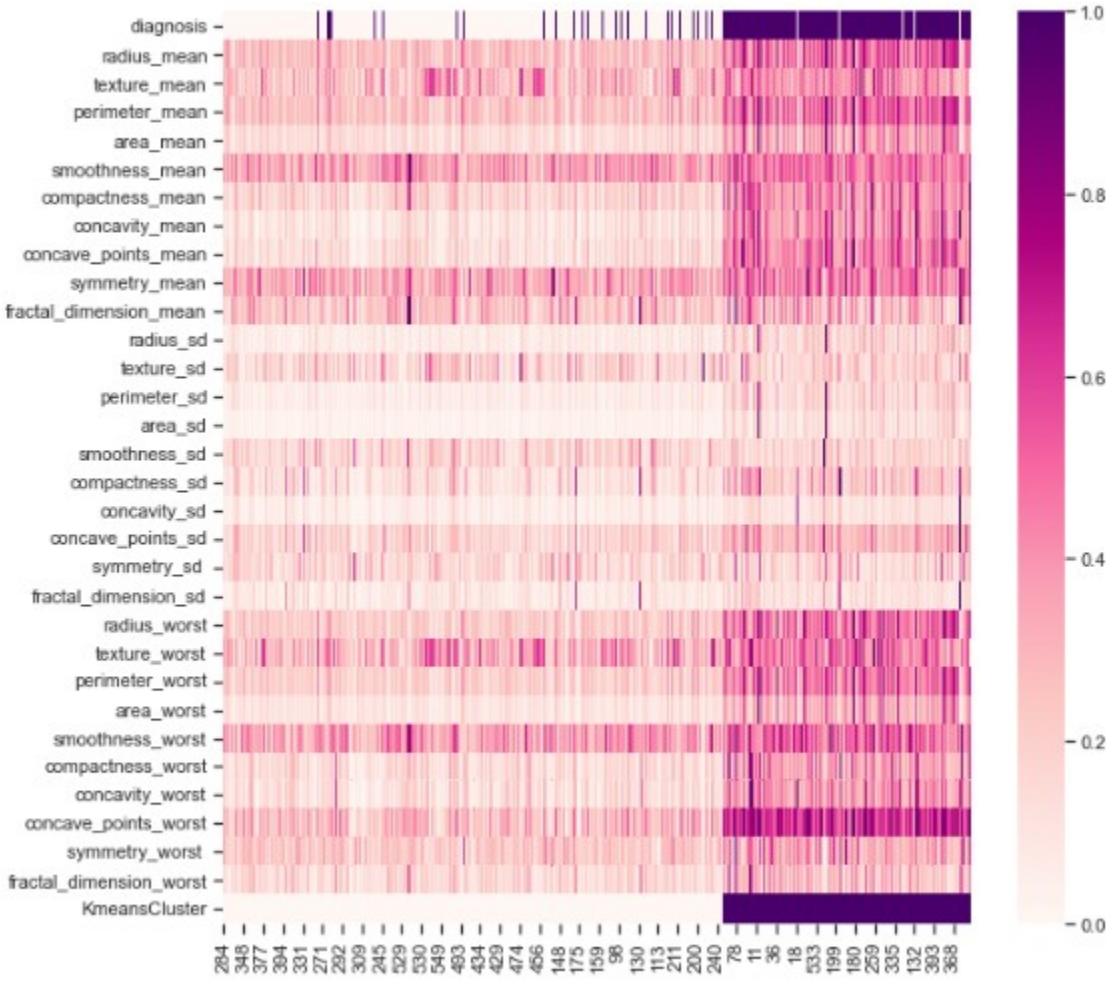


question 3.

- Yes, The box plot for the 'area_mean' feature indicates the presence of outliers, the range of outliers shown in box plot is approximately ranges after 1500 to 2500 and visible as individual points beyond the whiskers.
- Outliers in 'area_mean' may significantly affect the analysis by skewing summary statistics and potentially influencing the overall interpretation of the data.
- These extreme values should be carefully considered, as they might represent abnormal instances or errors in the measurement process, impacting the generalizability of findings.
- Outliers can distort the shape and spread of data distributions, affecting the interpretation of patterns and relationships.
- Also affect the mean (average) of a dataset. The mean is sensitive to extreme values.
-



Question 4.



- The heatmap is created using Seaborn, where rows represent features, columns represent data points, and the color intensity represents the feature values.
- Hotter colors indicate small distance so corelation is more , while cooler colors indicate large distance so corelation is less.
- 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

Question 5.

The limitations present in data and their impact on Conclusions are following :

A. Missing Data:

Incomplete or missing data points can lead to biased analyses and incomplete understanding. So the conclusions may be skewed or limited if missing data are not handled appropriately. The analysis might not accurately represent the true characteristics of the population.

B. outliers in “area_mean”:

There are presence of outliers in data set which shows potential date anomalies.outliers may influence the statistical measures.