1. Introduction

   Breast Cancer is one of the fatal diseases which needs an early diagnosis and treatment. Analyzing the data for better identification of the tumor is need of the hour.

   Cancer cells come into existence if the body becomes a home for more cells that how many must be present. Benign and Malignant are the two types of tumors. Benign tumors are the non-cancerous ones with specific borders for spreading in the body while the malignant ones are cancerous ones that spread throughout the body.

   The aim of this report is to distinguish as well as get an understanding of the breast cancer in benign and malignant masses taking the other measurement characteristics into consideration.

   Dataset brief: Collection data of 443 benign sample tissues and 239 malignant sample tissues. 9 categorical attributes of these tissues are given with 1 as the least and 10 as the highest. Some of the data is missing.

2. Methods

   As mentioned in the introduction that there is missing data, assessed the missing data. Created a data frame without including the Sample ID number from the dataset. Created new variables for analyzing different points in the data.

   This report will have histograms, boxplots and correlation plots to describe the breast cancer data. These plots are used to give a better visualization for our data. As we have several observations, Boxplots are easier as they take up less space but give more components of data like percentiles, outliers, median which is necessary to investigate in this dataset. On the other hand, the histograms are used to here to give a broad overview of the data. Coming to the correlation plots, these are the best ones for this dataset to identify the linearity relationship between different characteristics of tissues in benign and malignant masses.

   Further, there will be a summary of the breast cancer data in various ways using the mean and standard deviation. Mean and Standard deviation come up as it is easy to analyze which are in the same units of the original data.

3. Results

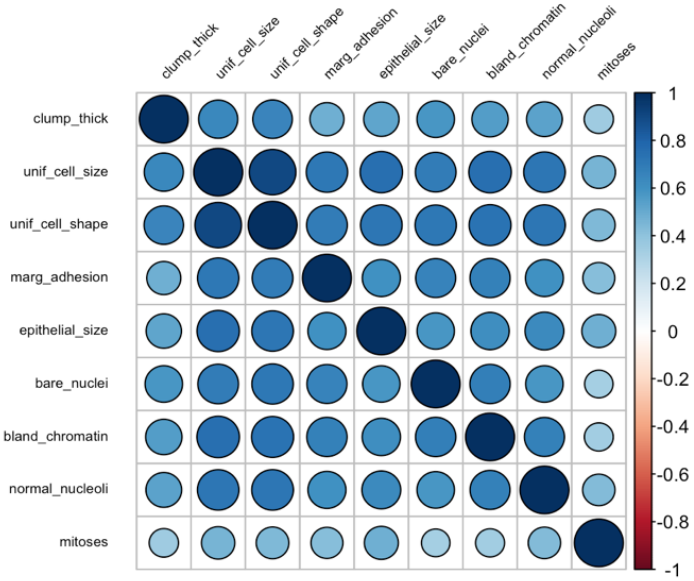

Fig 1

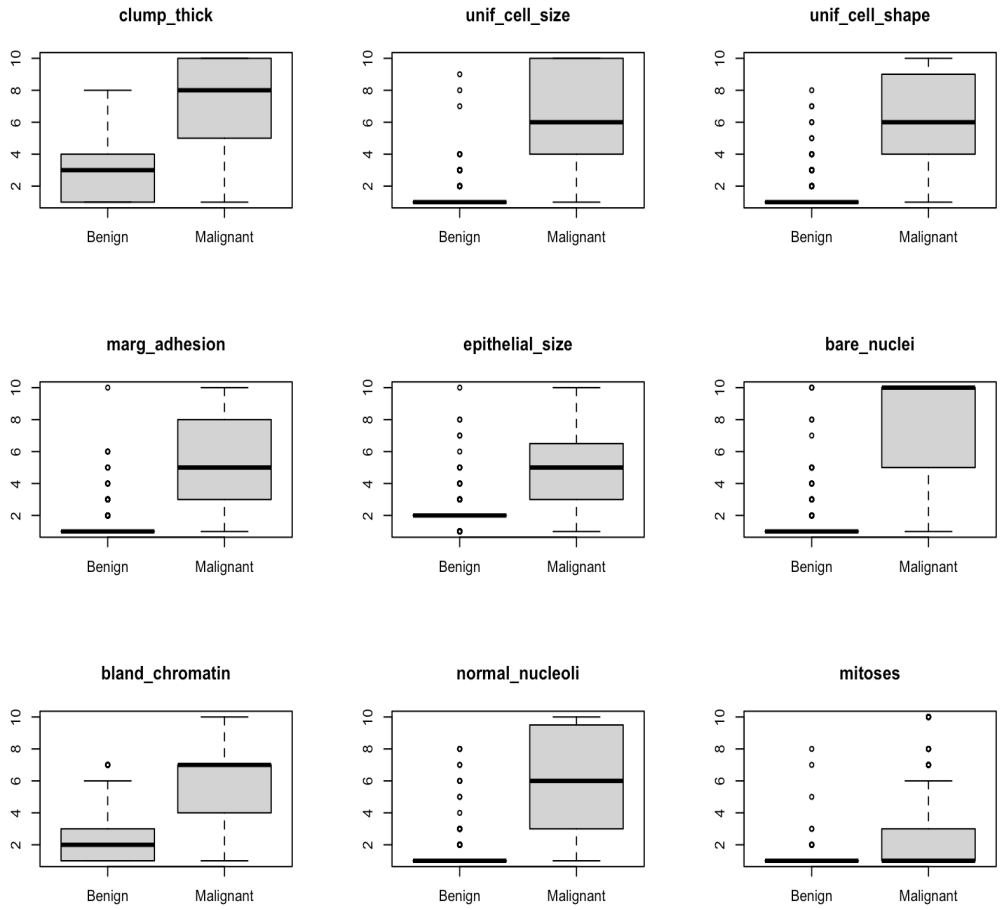

Fig 2

From the figure 1, we can see the relationsnip between the different attributes of the benign and malignant masses. We can take that most of the relationship of characteristics with mitoses are less related than compared with the others
From the figure 2, For non-cancerous breast cancer, most of the sample tissues have the characteristic measurements below 2 and rarely above 2. Moving to Malignant masses, all other measurements except mitoses lie in the quartile ranges and are opposite to that of benign mass even though they have a median around halfway. Normal_nucleoli in malignant mass are spread evenly than any other attribute. The median measurement for mitoses is nearly the same for both the tumors.
The average clump thickness for a breast cancer detection regardless of the type of tumor is around 4.4 being the highest and 1.6 for mitoses being the least for the given dataset.
The deviation from the mean is slightly higher for the bare_nucleoli with rounding to 4 while the least deviation is with the mitoses again.

4. Conclusion

There are few missing details in the bare_nucleoli attribute of the data. More investigation is necessary there.
From the results section, we can conclude that benign tumor has lower measurements in almost all the characteristics except mitoses while the malignant tumor has higher measurements.

References:

1. Aisha patel, Benign vs Malignant Tumors, July 30, 2020