Assignment 1

Matthias
Bennett

January 30th, 2024

1 Statistical Distribution of Attributes

	radius_mean	smoothness_mean	texture_se
mean	14.221321	0.096536	1.200766
median	13.46000	0.09646	1.07300
Q1	11.760000	0.086770	0.828200
Q3	16.130000	0.105400	1.471000

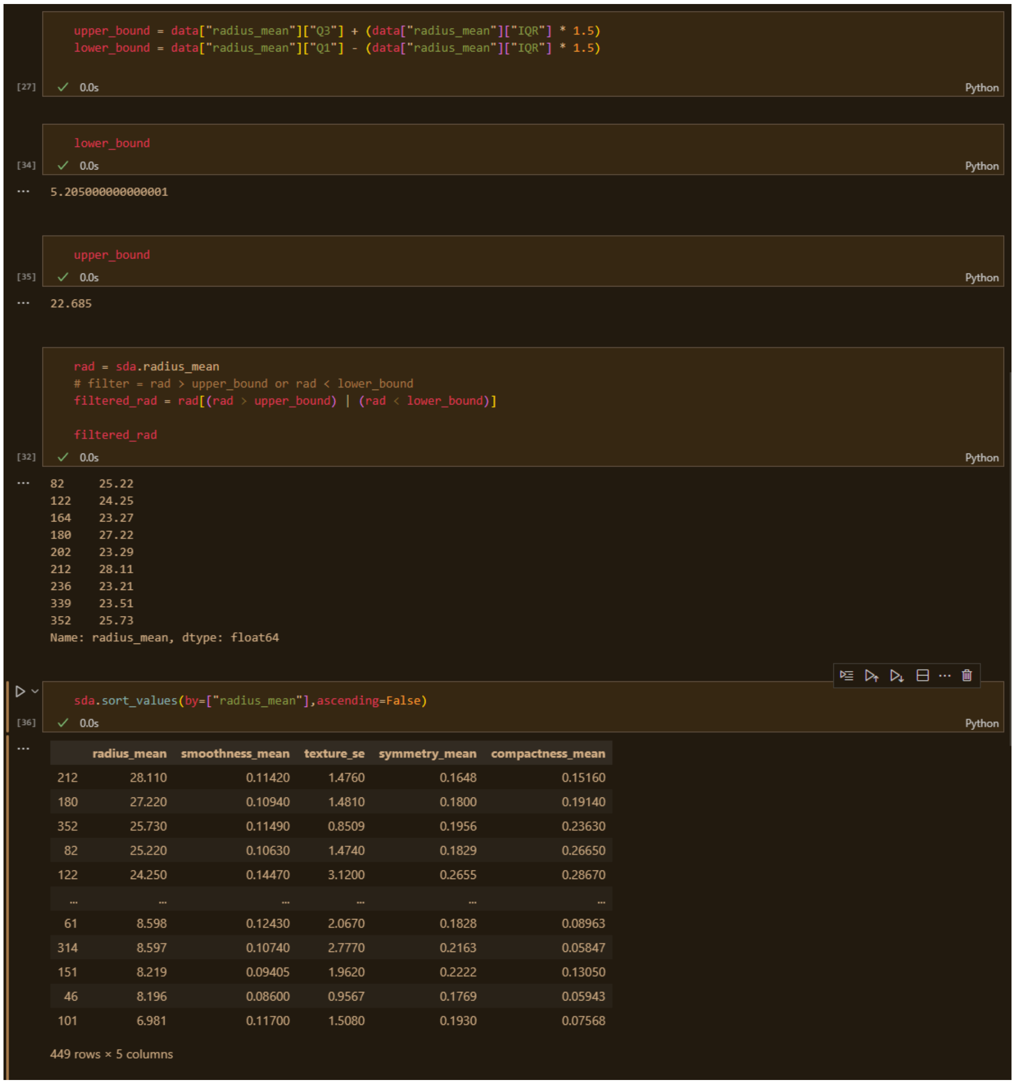
texture-se: The mean texture-se is slightly higher than the median which indicates a right skew in the distribution. The IQR for texture-se is 0.6428 which suggests a moderate spread in the data. This spread indicates some variability in among the samples, but not as pronounced as in the radius-mean.

smoothness-mean: The mean and median of smoothness are very close, suggesting a nearly symmetrical distribution. The proximity of these two measures indicates that the attribute does not have a pronounced skew. The IQR for smoothness is also relatively small, indicating that most values are concentrated around the median.

radius-mean: The mean radius is slightly higher than the median. This suggests a slight right skew in the distribution, indicating that there are a few larger values pulling the mean upwards compared to the median. The IRQ is 4.37. This relatively wide range suggests variability in the data, indicating that the radius-mean varies considerably across the dataset.

2 Outlier Identification

Here we calculate the upper bound and lower bound, then filter out all of the outliers, and sort them in descending order. The resulting table contains all of the outliers.



```python
upper_bound = data["radius_mean"]["Q3"] + (data["radius_mean"]["IQR"] * 1.5)
lower_bound = data["radius_mean"]["Q1"] - (data["radius_mean"]["IQR"] * 1.5)
```

```python
lower_bound
```

```
5.205000000000001
```

```python
upper_bound
```

```
22.685
```

```python
rad = sda.radius_mean
# filter = rad > upper_bound or rad < lower_bound
filtered_rad = rad[(rad > upper_bound) | (rad < lower_bound)]

filtered_rad
```

```
82     25.22
122    24.25
164    23.27
180    27.22
202    23.29
212    28.11
236    23.21
339    23.51
352    25.73
Name: radius_mean, dtype: float64
```

```python
sda.sort_values(by=["radius_mean"],ascending=False)
```

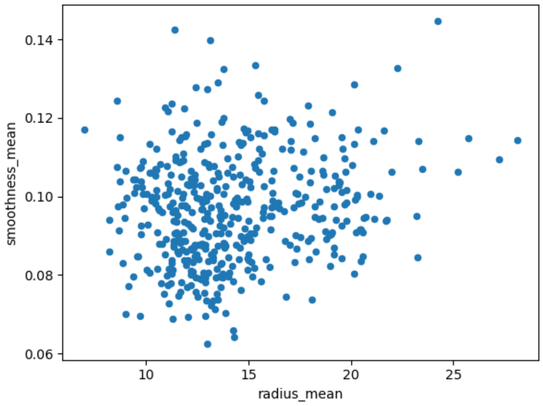
	radius_mean	smoothness_mean	texture_se	symmetry_mean	compactness_mean
212	28.110	0.11420	1.4760	0.1648	0.15160
180	27.220	0.10940	1.4810	0.1800	0.19140
352	25.730	0.11490	0.8509	0.1956	0.23630
82	25.220	0.10630	1.4740	0.1829	0.26650
122	24.250	0.14470	3.1200	0.2655	0.28670
...	...	...	...	...	...
61	8.598	0.12430	2.0670	0.1828	0.08963
314	8.597	0.10740	2.7770	0.2163	0.05847
151	8.219	0.09405	1.9620	0.2222	0.13050
46	8.196	0.08600	0.9567	0.1769	0.05943
101	6.981	0.11700	1.5080	0.1930	0.07568

449 rows × 5 columns

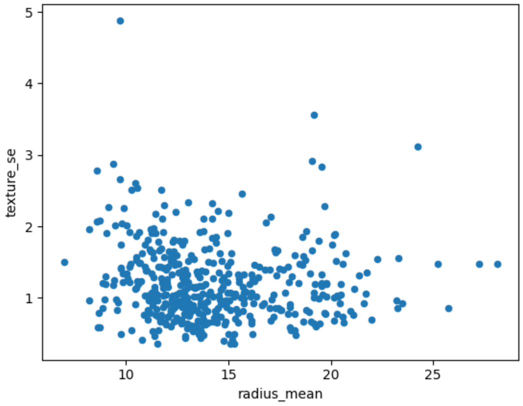
3 Correlations

In this section we will explore the correlations between different attributes

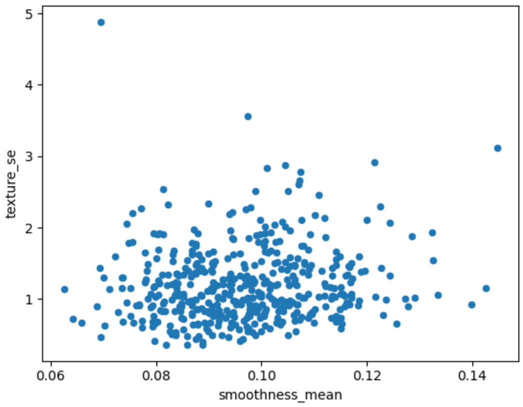
In these three scatter plots, we do not see any strong correlation between any of these attributes. If there was a correlation we would see some sort of pattern, whether positive, negative, neutral, etc.



No Correlation

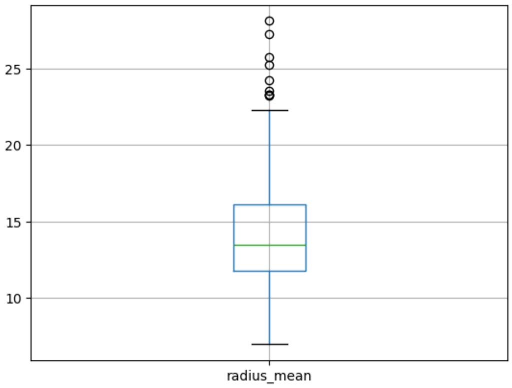


No Correlation

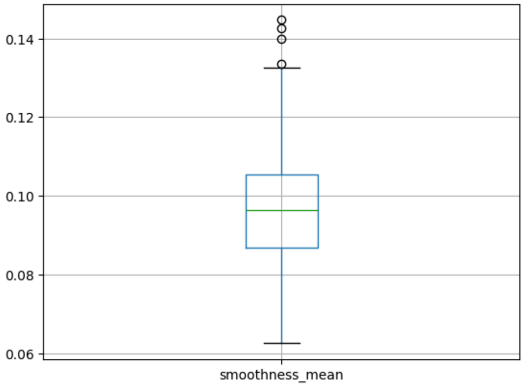


No Correlation

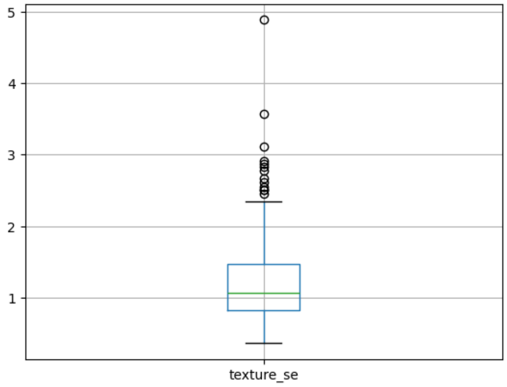
4 Boxplots



radius-mean: This boxplot shows a slight right skew since the median is slightly closer to the bottom of the IRQ box and the top or right whisker is slightly longer than the other. All of the outliers appear on the high side.



smoothness-mean: This boxplot depicts a fairly symmetrical distribution of data with few outliers.



texture-se: This boxplot shows texture-se has a very narrow interquartile range indicating that most of the data is clustered around the median. It also reveals that there are many outliers, some being very extreme.