Analysis of Breast Cancer data using ggplots

Praveen Nagarajan

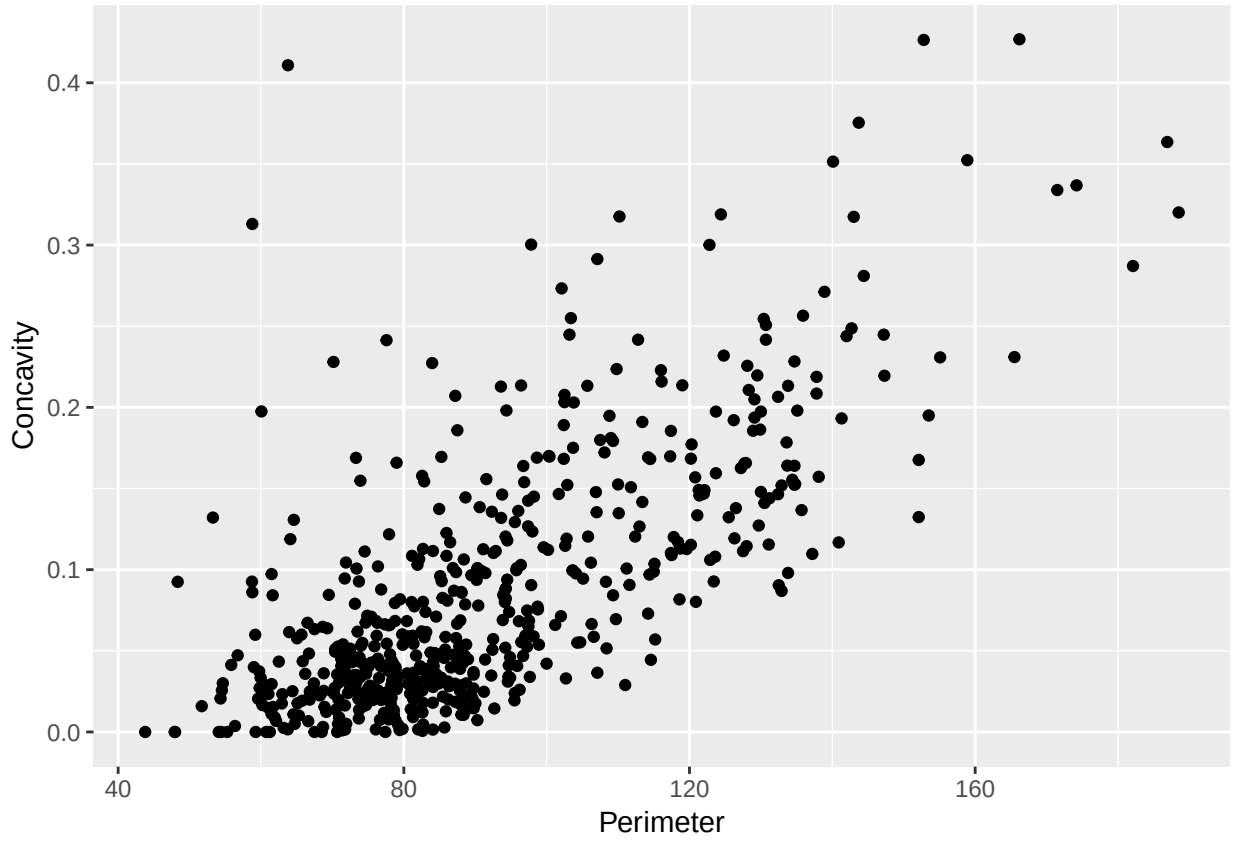
4/28/2019

Contents

1. geom_point

i. Initial Visualization

```
ggplot(data_set, aes(x = perimeter_mean, y = concavity_mean)) +geom_point(colour="black") +labs(x = "Pe
```



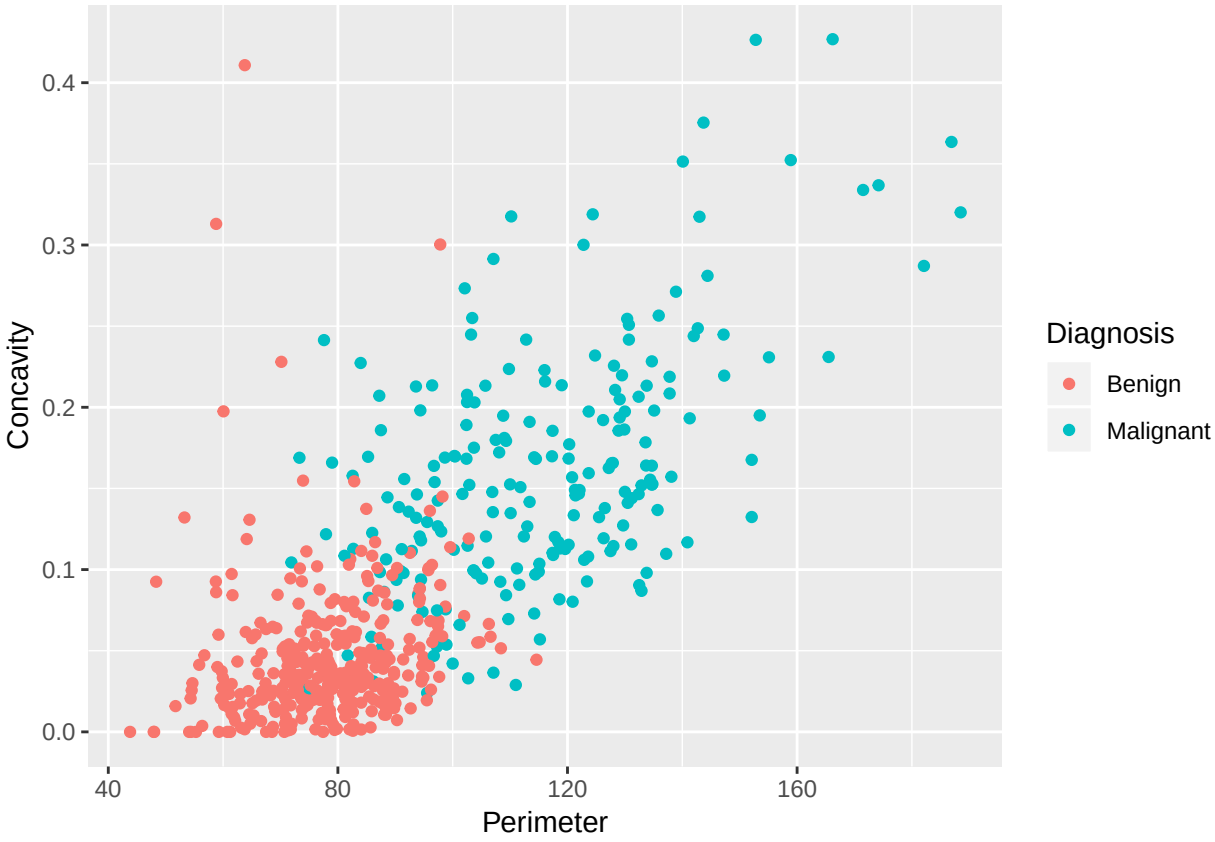
The first graph above shows a strong positive relationship between the Perimeter of a particular tumor sample and its Concavity.

The Perimeter of a particular tumor sample is the mean size of the tumor's core.

The Concavity expresses the mean severity of the concave portions of the tumor's contour.

We will be exploring this relationship further using the variations of the ggplot attributes.

ii. Indicate points as Benign/Malignant

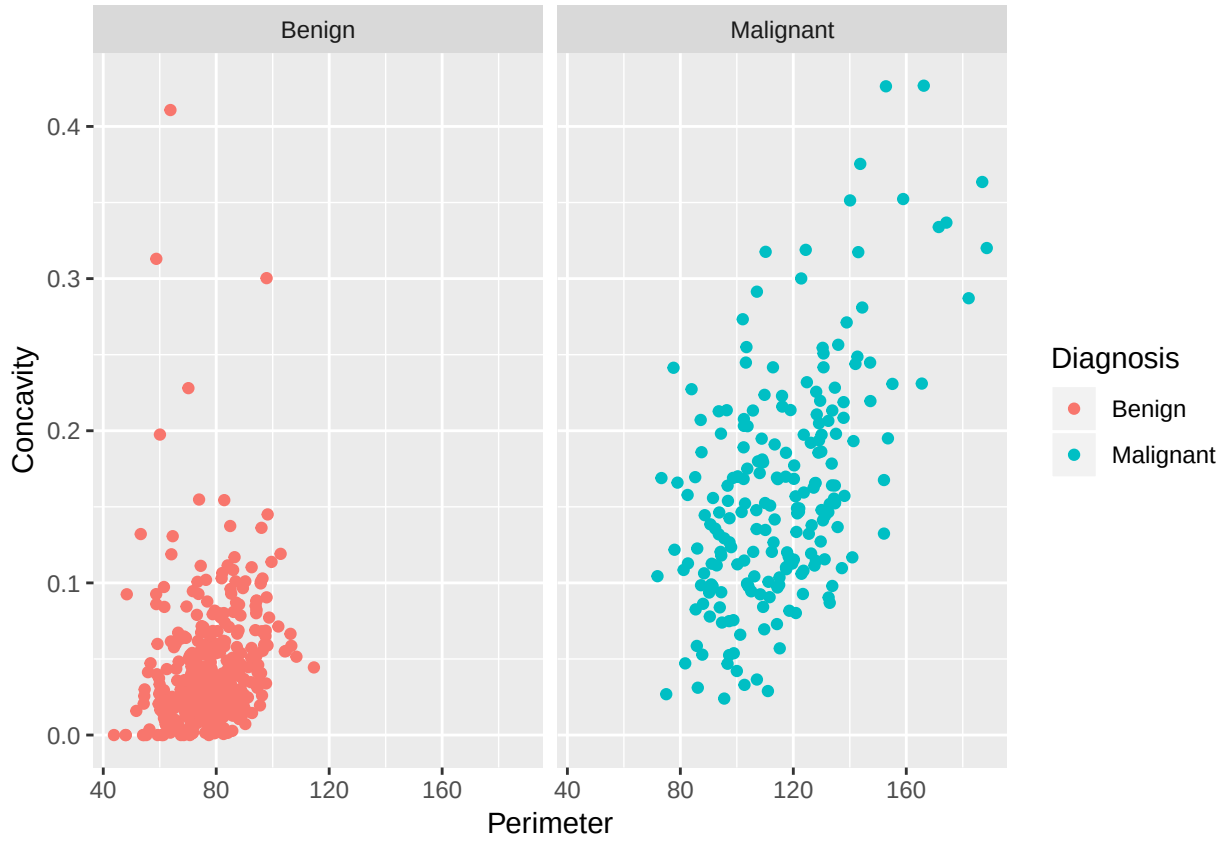


In this above graph, we add different colors for the Diagnosis levels (Benign, Malignant).

We can visually see a relationship between Perimeter, Concavity, and the Diagnosis.

Note that Benign is colored Red, and Malignant is colored Teal.

iii. Seperate the points labeled Benign/Malignant



This above graph, shows the tumor samples seperated by their classification as Benign/Malignant.

We can see a few trends:

Benign examples are clustered with low perimeter and concavity.

Malignant examples have high perimeter (which feels intuitive).

Numerical observations:

Benign cases have Perimeter less than 120.

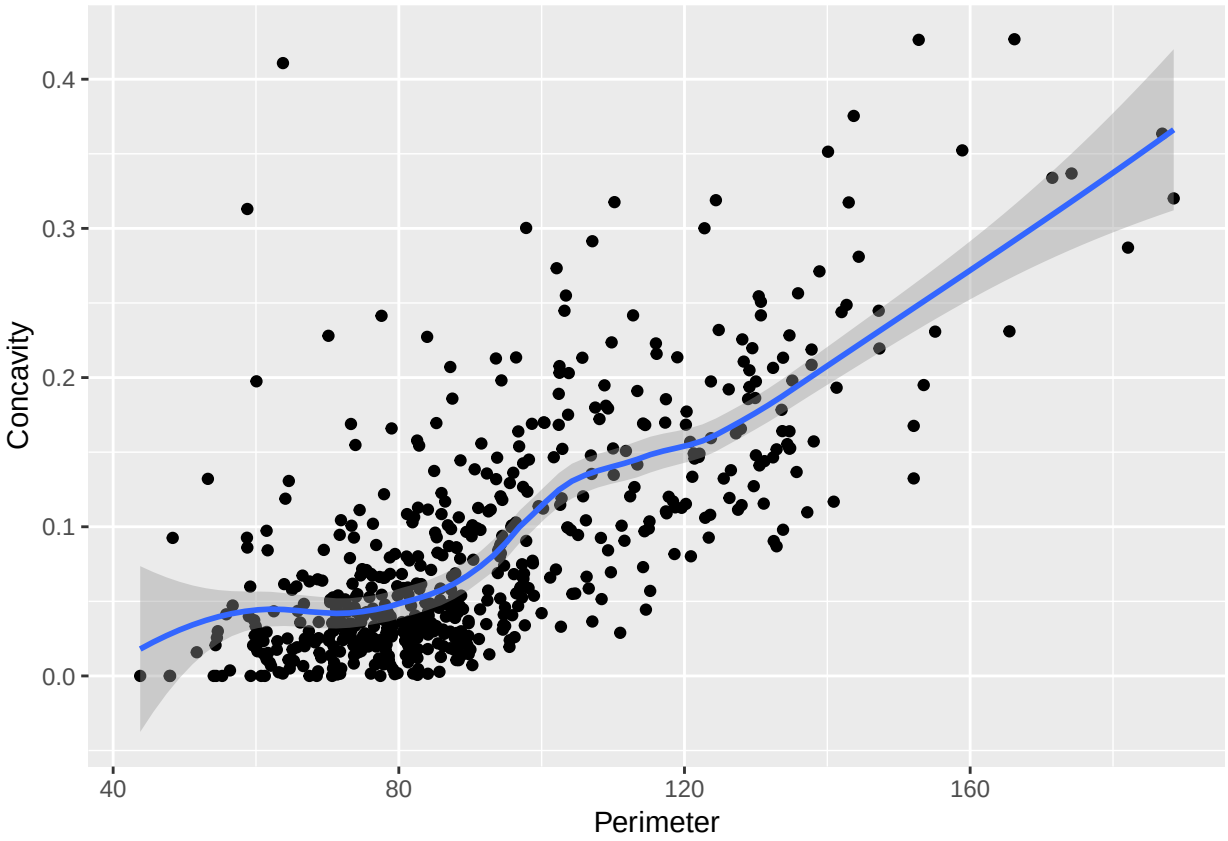
Malignant cases have Perimeter greater than 60.

2. geom_smooth

i. loess - "Locally Estimated Scatterplot Smoothing" visualization

```
ggplot(data_set, aes(x = perimeter_mean, y = concavity_mean)) +geom_point() +geom_smooth(span=.5) +labs
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



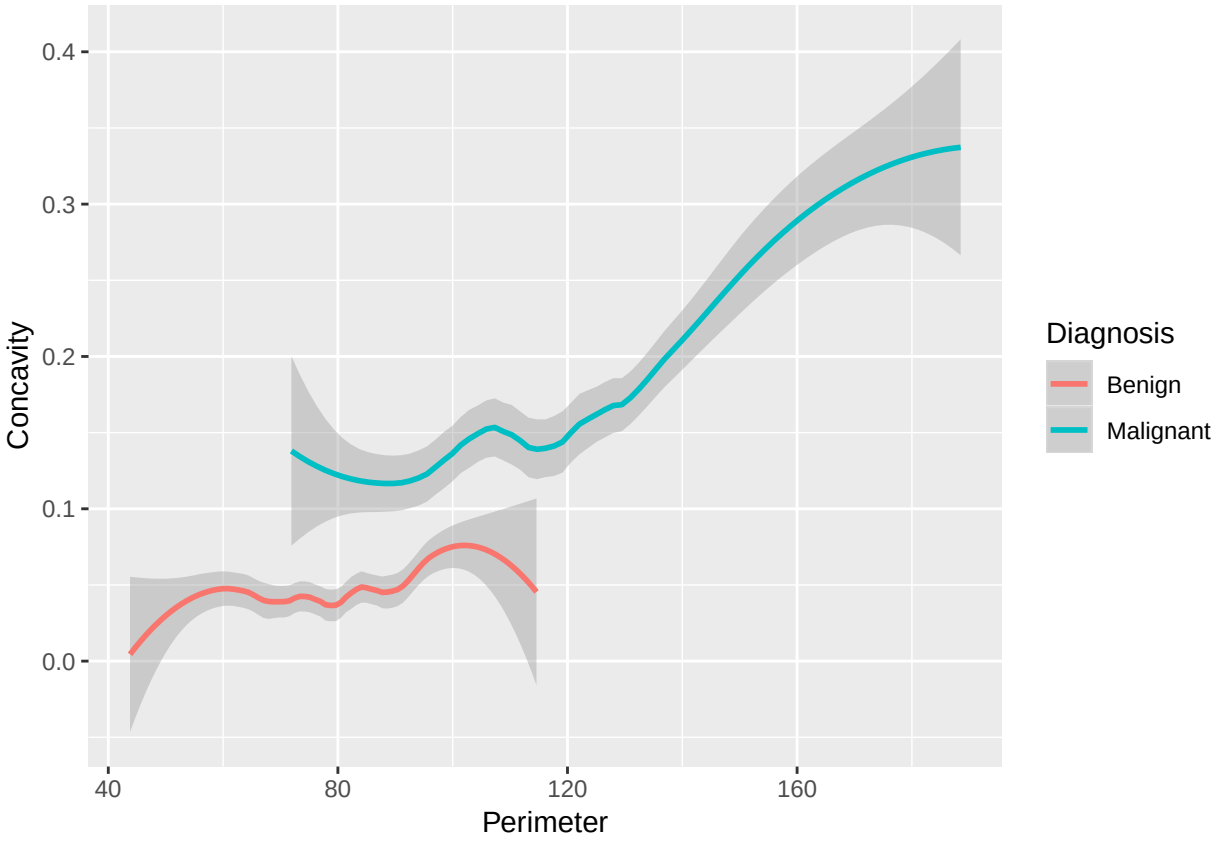
This plot overlays a loess-fitted curve to the scatterplot.

We can see the fitted line and the estimated confidence.

ii. loess visualization seperated by Benign/Malignant

```
ggplot(data_set, aes(x = perimeter_mean, y = concavity_mean, colour=Diagnosis)) +geom_smooth(span=.5) +
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



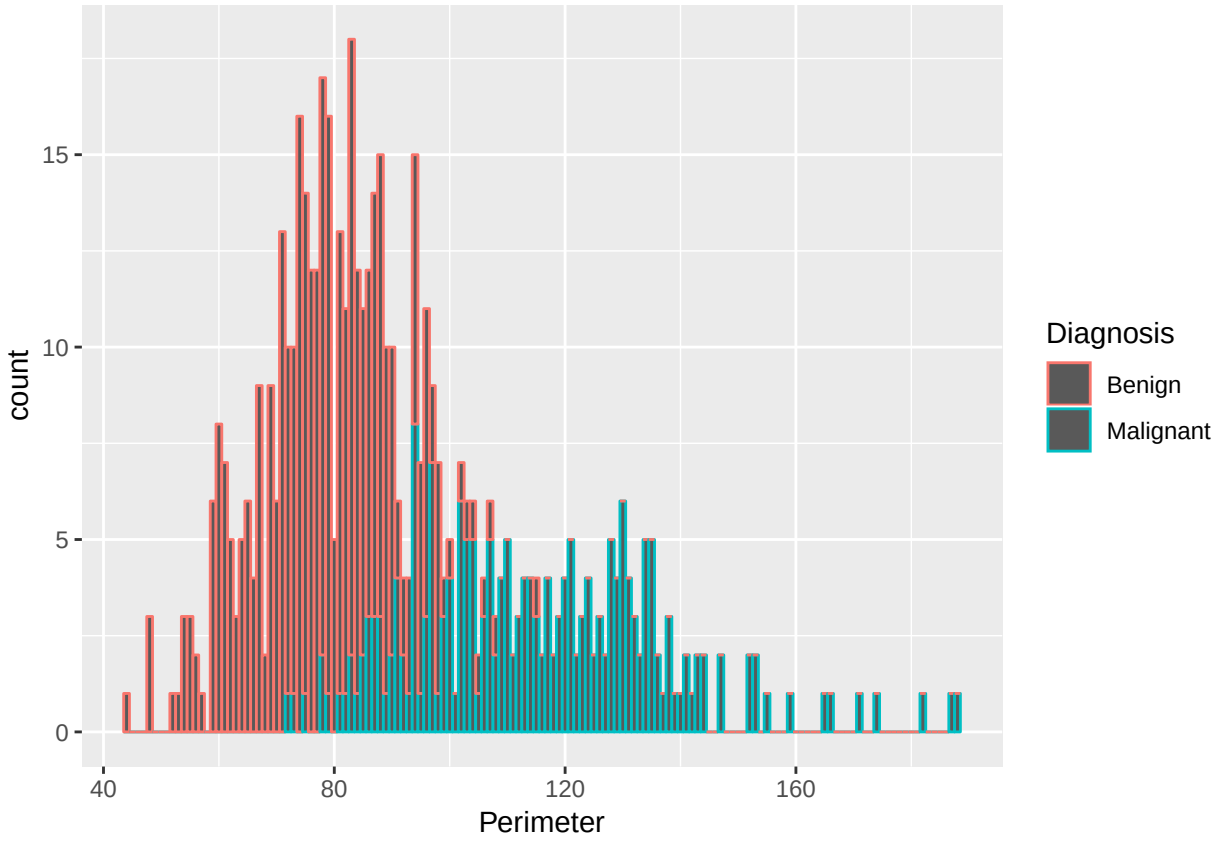
We can see a fitted line generated for both Benign cases and Malignant cases.

It is interesting to note that with a span of just 0.5 the loess curves for Benign and Malignant (including the areas for confidence) may be linearly seperable.

3. geom_histogram

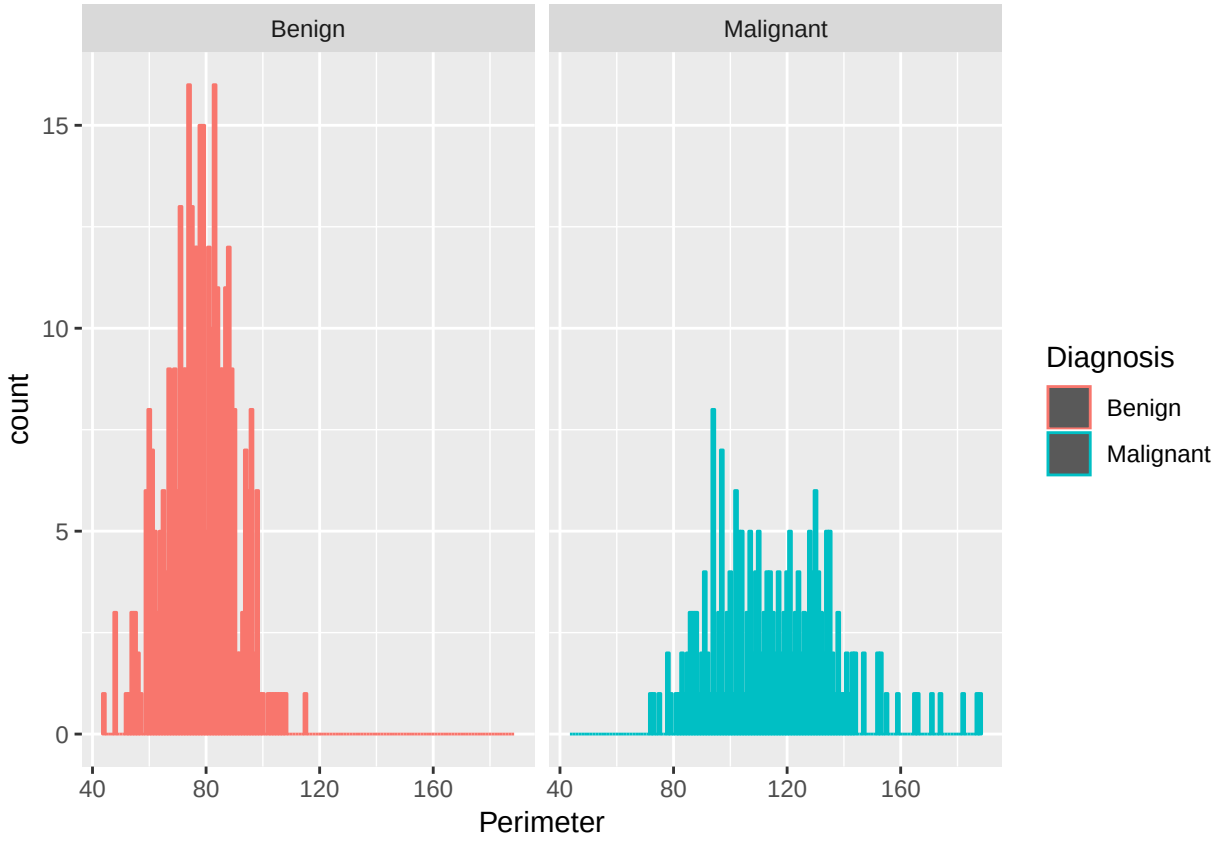
i. Histogram of Perimeter with Benign/Malignant classifications

```
ggplot(data_set, aes(perimeter_mean, colour=Diagnosis)) +geom_histogram(binwidth=1) +labs(x="Perimeter"
```



ii. Seperate Histograms of Perimeter with Benign/Malignant classifications

```
ggplot(data_set, aes(perimeter_mean, colour=Diagnosis)) +geom_histogram(binwidth=1) +facet_wrap(data_se
```


Benign
Malignant
15
10
5
0
count
40
80
120
160
Perimeter
Diagnosis
Benign
Malignant