IDS-Final Project

Avisek Choudhury, Aldo Adriazola, Kait Arnold

3/08/2020

Contents

Dataset

The physicians have identified a data set that consists of over 500 measurements from Fine Needle Aspiration (FNA) of breast tissue masses. In an FNA, a small needle is used to extract a sample of cells from a tissue mass. The cells are then photographed under a microscope. The resulting photographs are entered into graphical imaging software. A trained technician uses a mouse pointer to draw the boundary of the nuclei. The software then calculates each of ten characteristics for the nuclei. This process is repeated for most or all of the nuclei in the sample.

The data consists of measurements of the cell nuclei for the following characteristics:

1. radius
2. texture
3. perimete r
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeterˆ2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" - 1)

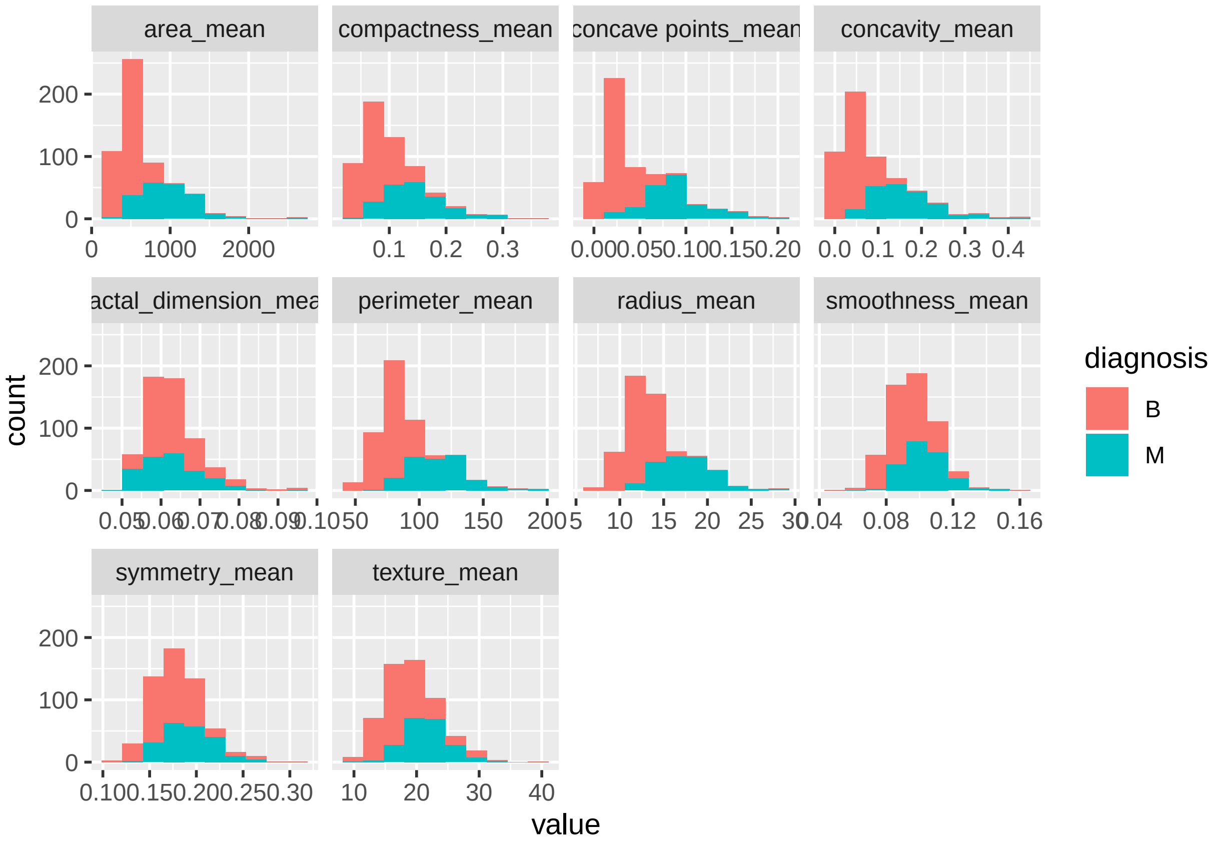
Measurements of these ten characteristics are summarized for all cells in the sample. The dataset consists of the mean, standard error of the mean, and maximum of the 10 characteristics, for a total of 30 observations for each. Additionally, the data set includes an identification number and a variable that indicates if the tissue mass is malignant (M) or benign (B).

```r
library(tidyverse)

#Load the dataset
cancer_df <- read_csv('C:/MSDS/Spring 2020/IDS/Project/FNA_cancer.csv')

#Print the dataset
head(cancer_df)
```

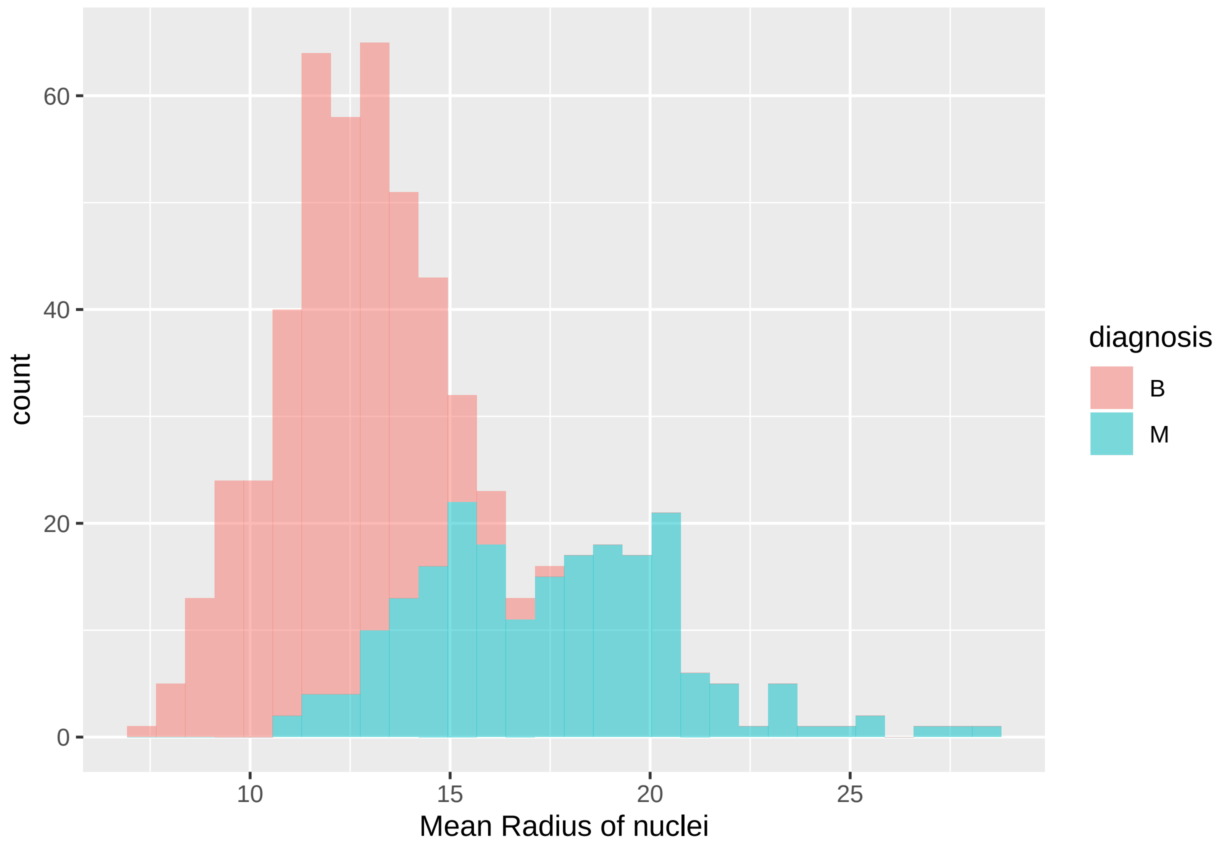
```
## # A tibble: 6 x 32
##       id diagnosis radius_mean texture_mean perimeter_mean area_mean
##    <dbl> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
## 1 8.42e5 M                18.0         10.4           123.      1001
## 2 8.43e5 M                20.6         17.8           133.      1326
## 3 8.43e7 M                19.7         21.2          130       1203
## 4 8.43e7 M                11.4         20.4           77.6      386.
## 5 8.44e7 M                20.3         14.3          135.      1297
## 6 8.44e5 M                12.4         15.7           82.6      477.
## # ... with 26 more variables: smoothness_mean <dbl>, compactness_mean <dbl>,
## #   concavity_mean <dbl>, `concave points_mean` <dbl>, symmetry_mean <dbl>,
## #   fractal_dimension_mean <dbl>, radius_se <dbl>, texture_se <dbl>,
## #   perimeter_se <dbl>, area_se <dbl>, smoothness_se <dbl>,
## #   compactness_se <dbl>, concavity_se <dbl>, `concave points_se` <dbl>,
## #   symmetry_se <dbl>, fractal_dimension_se <dbl>, radius_worst <dbl>,
## #   texture_worst <dbl>, perimeter_worst <dbl>, area_worst <dbl>,
## #   smoothness_worst <dbl>, compactness_worst <dbl>, concavity_worst <dbl>,
## #   `concave points_worst` <dbl>, symmetry_worst <dbl>,
## #   fractal_dimension_worst <dbl>
```

Exploratory data analysis (EDA)

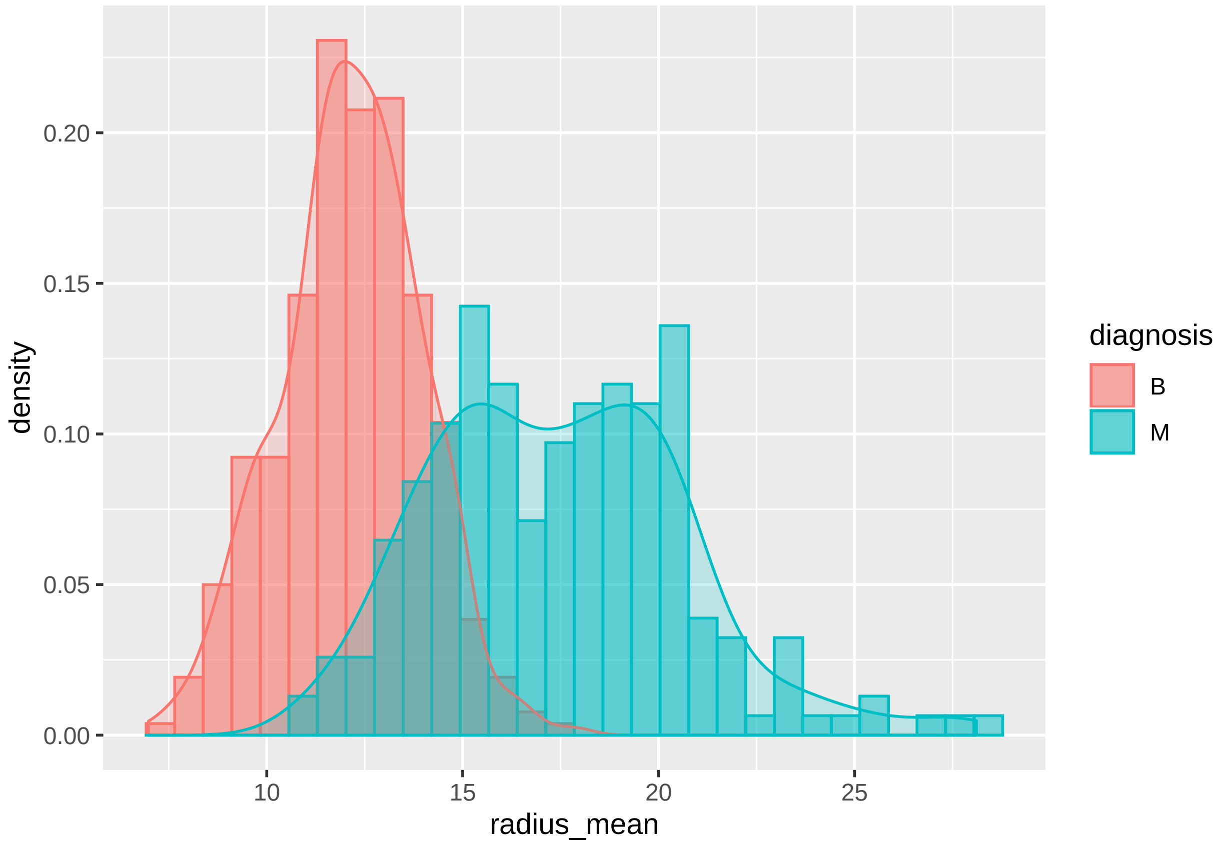
```
# ggplot(cancer_df[ , c(3:12)] %>% gather(), aes(value)) +
#     geom_histogram(bins = 10) +
#     facet_wrap(~key, scales = 'free_x')

ggplot(cancer_df[ , c(2:12)] %>%
         pivot_longer(cols = radius_mean:fractal_dimension_mean),
       aes(value, fill = diagnosis)) +
    geom_histogram(bins = 10) +
    facet_wrap(~name, scales = 'free_x')
```



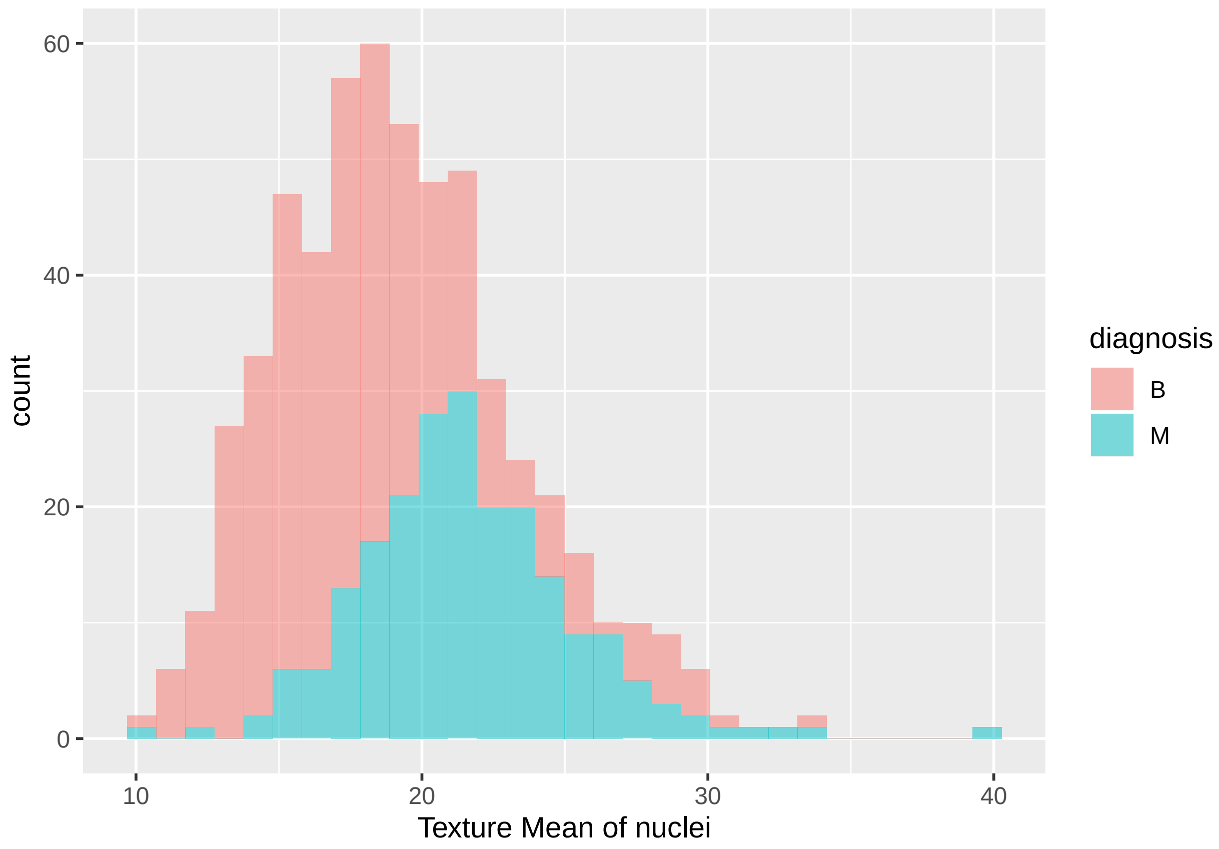
```
#Histogram of the Mean Radius of nuclei
ggplot(data = cancer_df, aes(x = radius_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Mean Radius of nuclei')
```



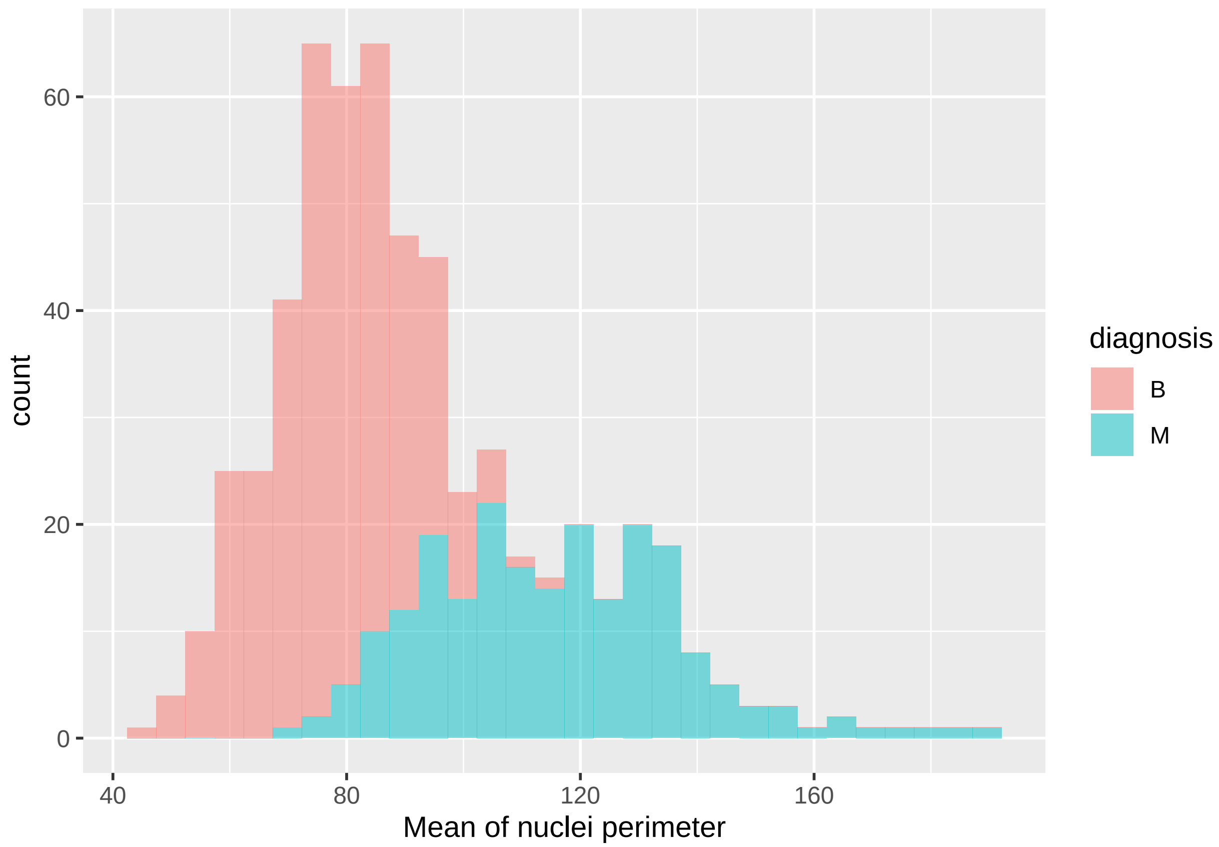
```
ggplot(cancer_df, aes(x = radius_mean, color = diagnosis, fill = diagnosis)) +
 geom_histogram(aes(y=..density..), alpha=0.5,
                position="identity")+
 geom_density(alpha=.2)
```



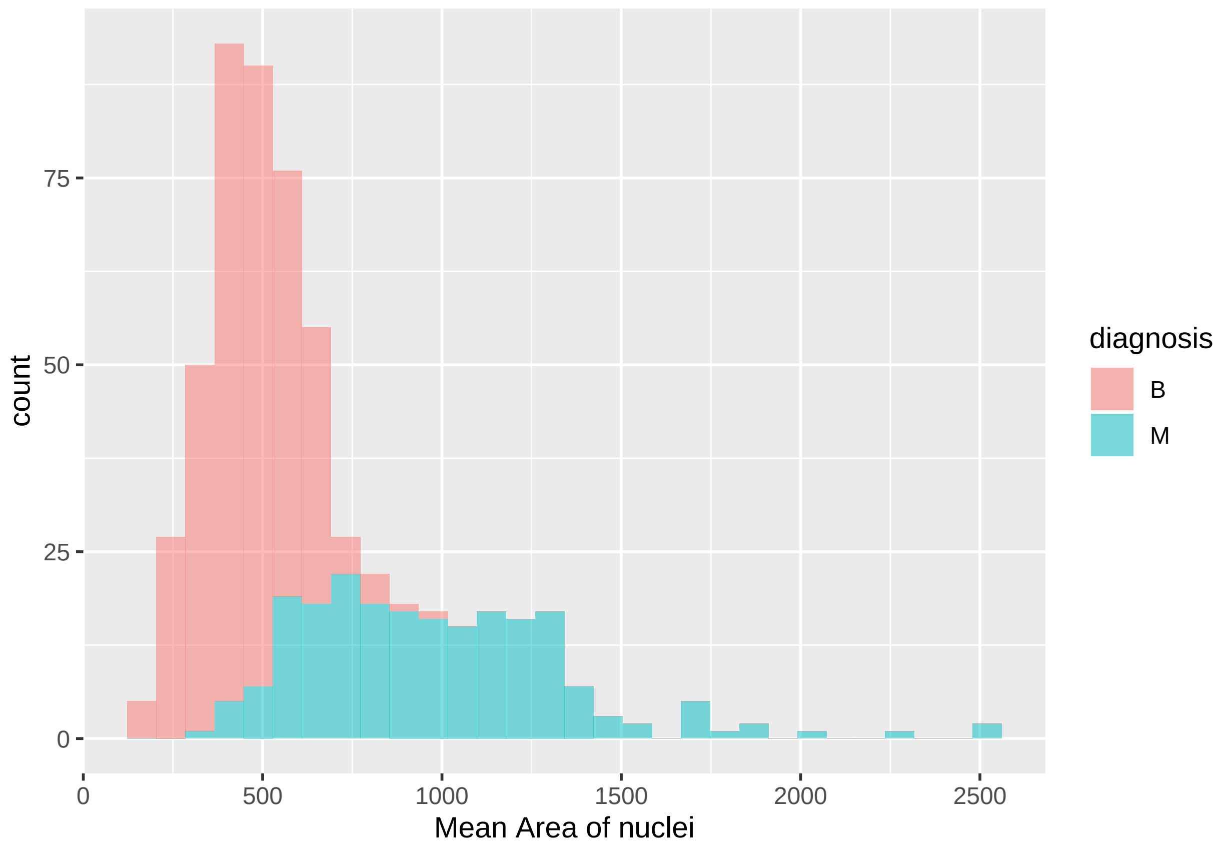
```
#Histogram of the texture_mean
ggplot(data = cancer_df, aes(x = texture_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Texture Mean of nuclei')
```



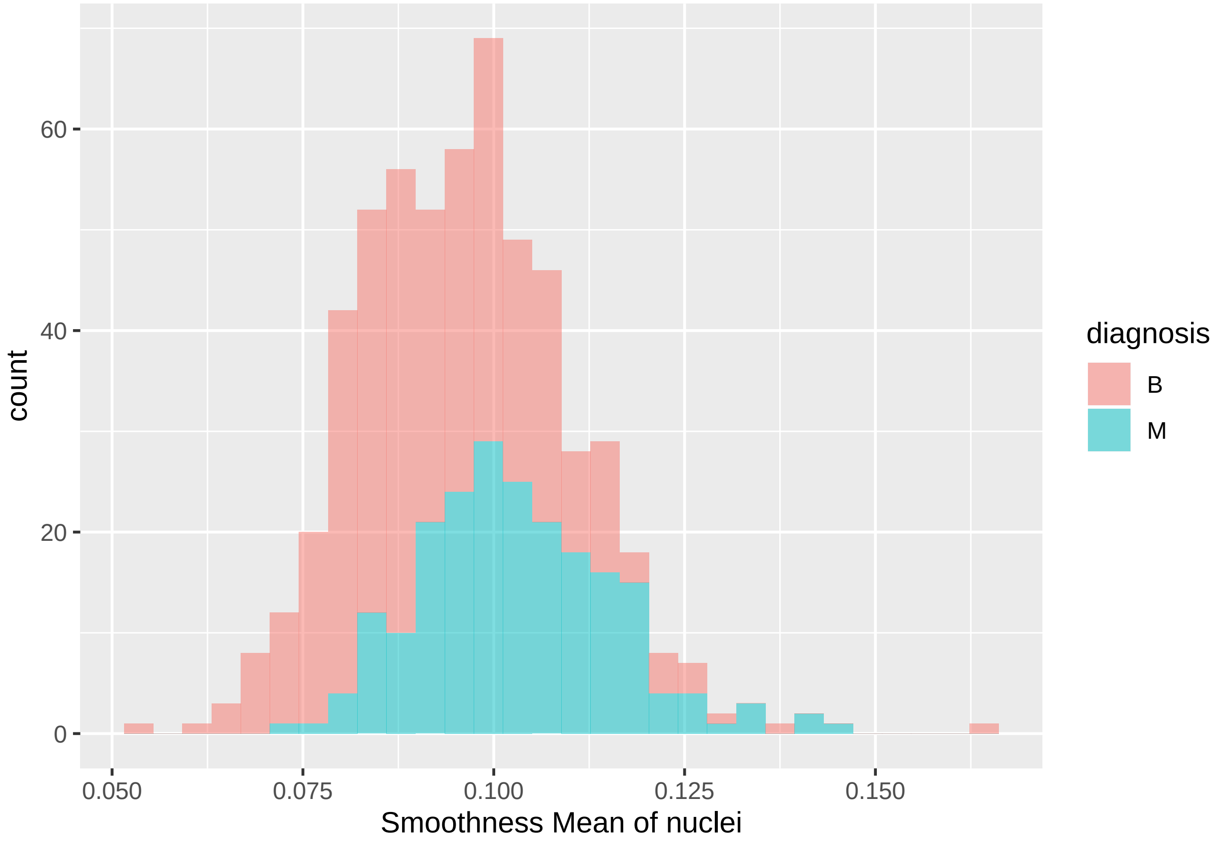
```
#Histogram of the perimeter_mean
ggplot(data = cancer_df, aes(x = perimeter_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Mean of nuclei perimeter')
```



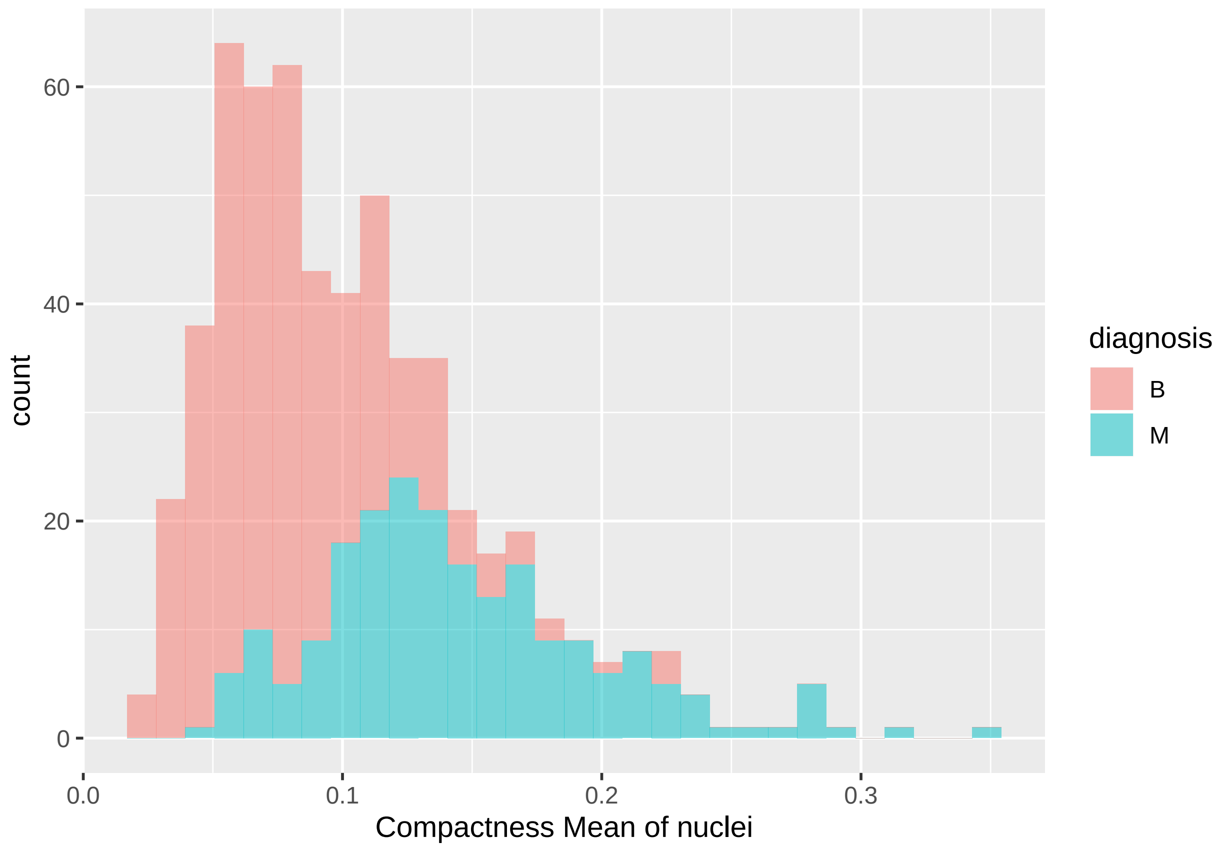
```
#Histogram of the area_mean
ggplot(data = cancer_df, aes(x = area_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Mean Area of nuclei')
```



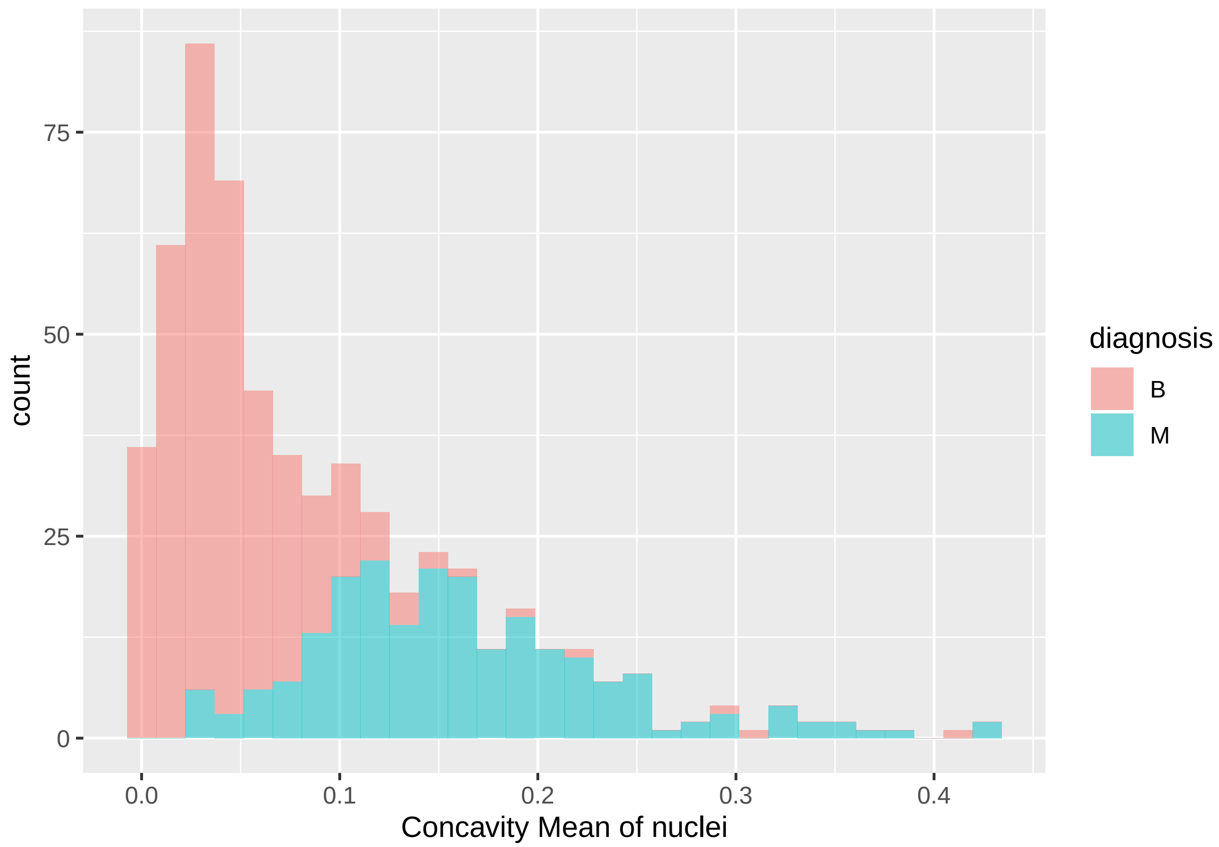
```r
#Histogram of the smoothness_mean
ggplot(data = cancer_df, aes(x = smoothness_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Smoothness Mean of nuclei')
```



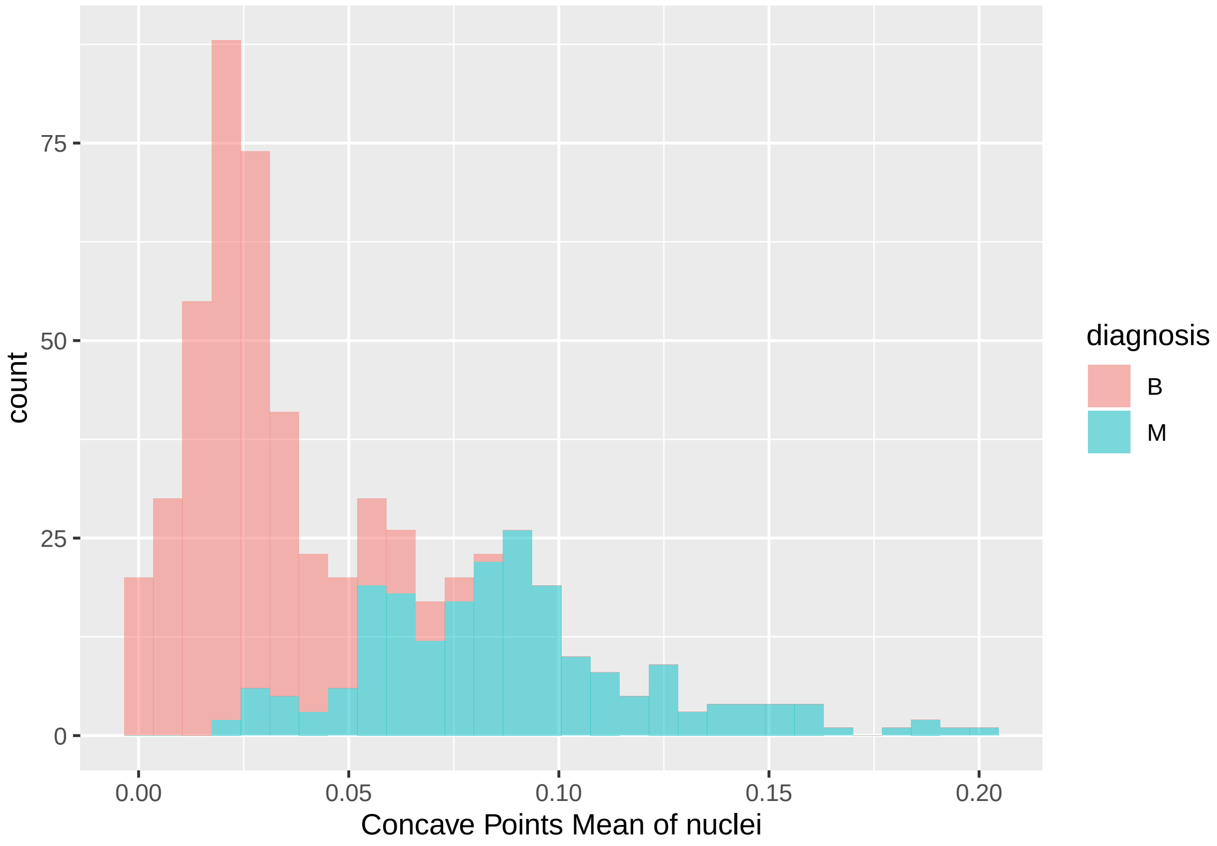
```
#Histogram of the compactness_mean
ggplot(data = cancer_df, aes(x = compactness_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Compactness Mean of nuclei')
```



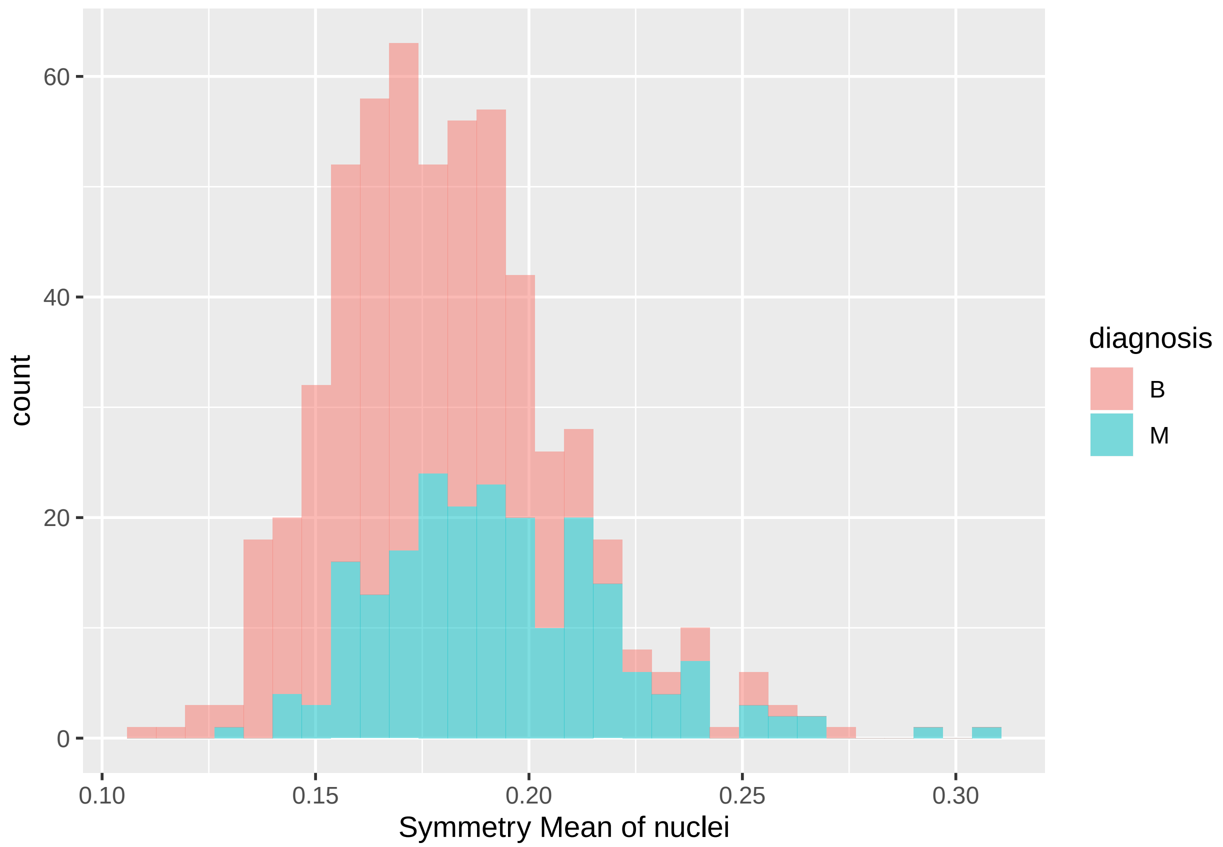
```
#Histogram of the concavity_mean
ggplot(data = cancer_df, aes(x = concavity_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Concavity Mean of nuclei')
```



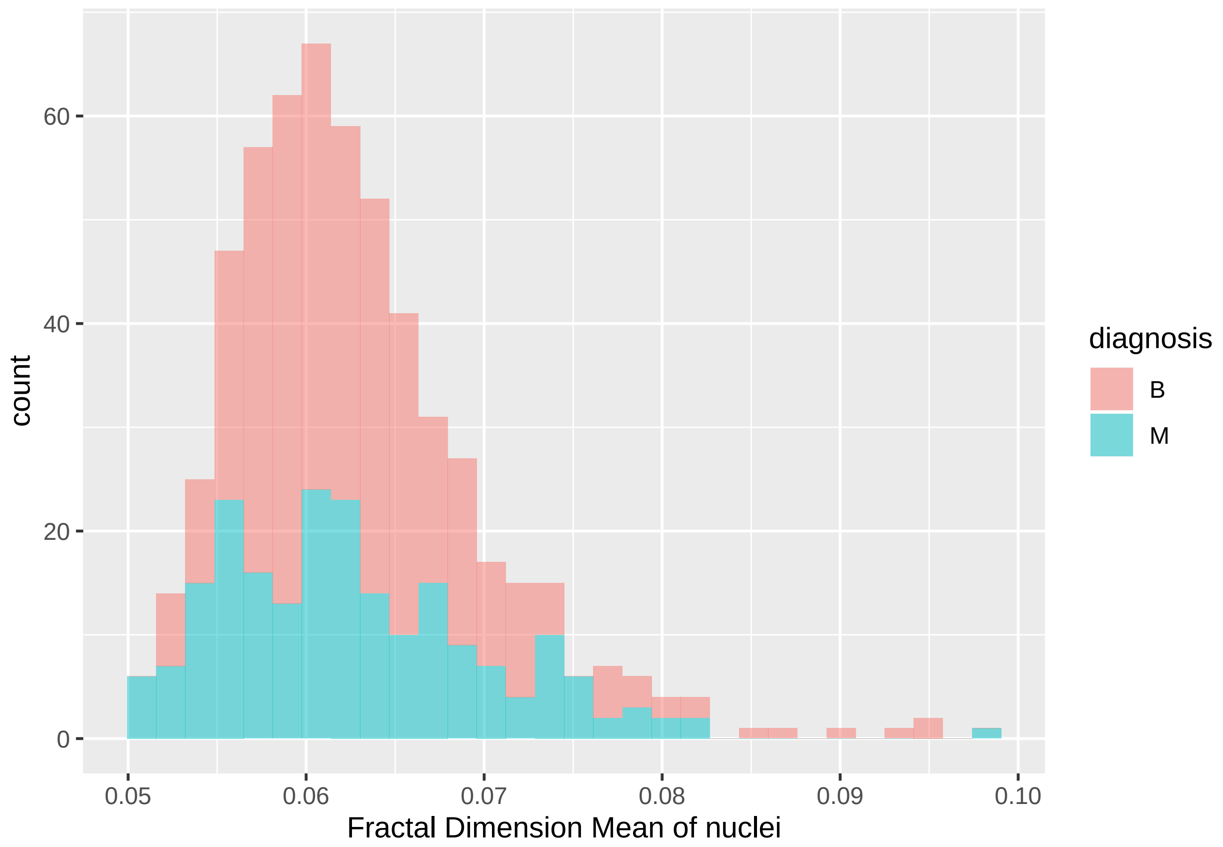
```
#Histogram of the concave points_mean
ggplot(data = cancer_df, aes(x = `concave points_mean`)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Concave Points Mean of nuclei')
```



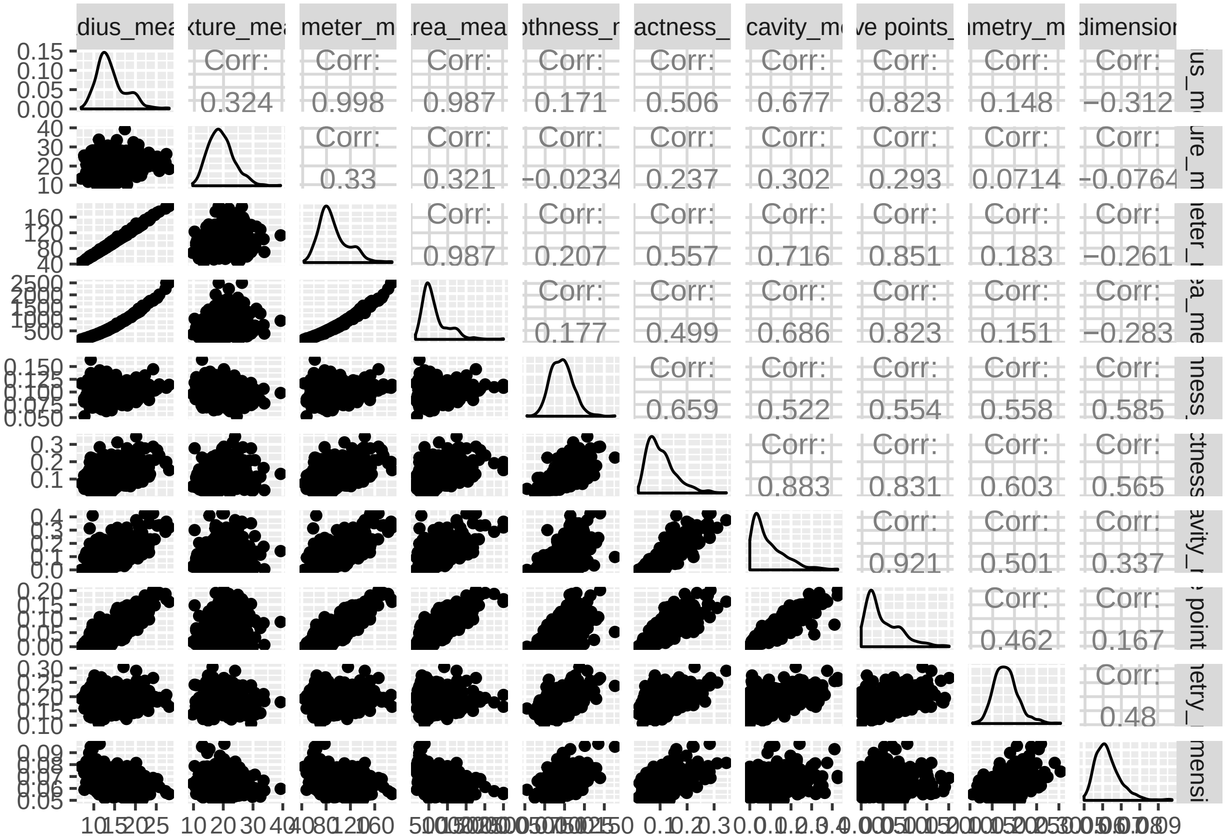
```
#Histogram of the symmetry_mean
ggplot(data = cancer_df, aes(x = symmetry_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Symmetry Mean of nuclei')
```



```
#Histogram of the symmetry_mean
ggplot(data = cancer_df, aes(x = fractal_dimension_mean)) +
  geom_histogram(aes(fill = diagnosis), alpha = 0.5) +
  xlab('Fractal Dimension Mean of nuclei')
```



```r
library(GGally)

#Pair plot between variables
ggpairs(cancer_df[ , c(3:12)])
```


dius_mea
xture_mea
meter_m
rea_mea
othness_r
actness_
cavity_me
ve points_
metry_m
dimensior
us_m
ure_m
eter_
a_me
ness_
tness
avity_
point
etry_
mensi
Corr:
0.324
0.998
0.987
0.171
0.506
0.677
0.823
0.148
−0.312
0.33
0.321
−0.0234
0.237
0.302
0.293
0.0714
−0.0764
0.987
0.207
0.557
0.716
0.851
0.183
−0.261
0.177
0.499
0.686
0.823
0.151
−0.283
0.659
0.522
0.554
0.558
0.585
0.883
0.831
0.603
0.565
0.921
0.501
0.337
0.462
0.167
0.48
0.15
0.10
0.05
0.00
40
30
20
10
160
120
80
40
2500
2000
1500
1000
500
0.150
0.125
0.100
0.075
0.050
0.3
0.2
0.1
0.4
0.3
0.2
0.1
0.0
0.20
0.15
0.10
0.05
0.00
0.30
0.25
0.20
0.15
0.10
0.09
0.08
0.07
0.06
0.05
10 15 20 25
10 20 30 40
40 80 120 160
500 1000 1500 2000 2500
0.050 0.075 0.100 0.125 0.150
0.1 0.2 0.3
0.0 0.1 0.2 0.3 0.4
0.00 0.05 0.10 0.15 0.20
0.10 0.15 0.20 0.25 0.30
0.05 0.06 0.07 0.08 0.09