breast_cancer_da

2025-08-06

Read data

```
data <- read.csv("~/Desktop/breast_cancer_da/Breast_cancer_dataset.csv")
```

Manual Correlations

Get the correaltions between diagnosis and the relevant values of radius of tumor

```
data_summary <- data |>
  group_by(diagnosis) |>
  summarise(
    max_radius = max(radius_mean, na.rm = TRUE),
    min_radius = min(radius_mean, na.rm = TRUE),
    avg_radius = mean(radius_mean, na.rm = TRUE)
  )

print(data_summary)
```

```
## # A tibble: 2 x 4
##   diagnosis max_radius min_radius avg_radius
##   <chr>          <dbl>      <dbl>      <dbl>
## 1 B               17.8       6.98       12.1
## 2 M               28.1      11.0        17.5
```

Correlation beween raidus_mean and area_mean

```
cor(data$radius_mean, data$area_mean, use = "complete.obs", method = "pearson")
```

```
## [1] 0.9873572
```

Full Correlation Matrix for All Numeric Variables

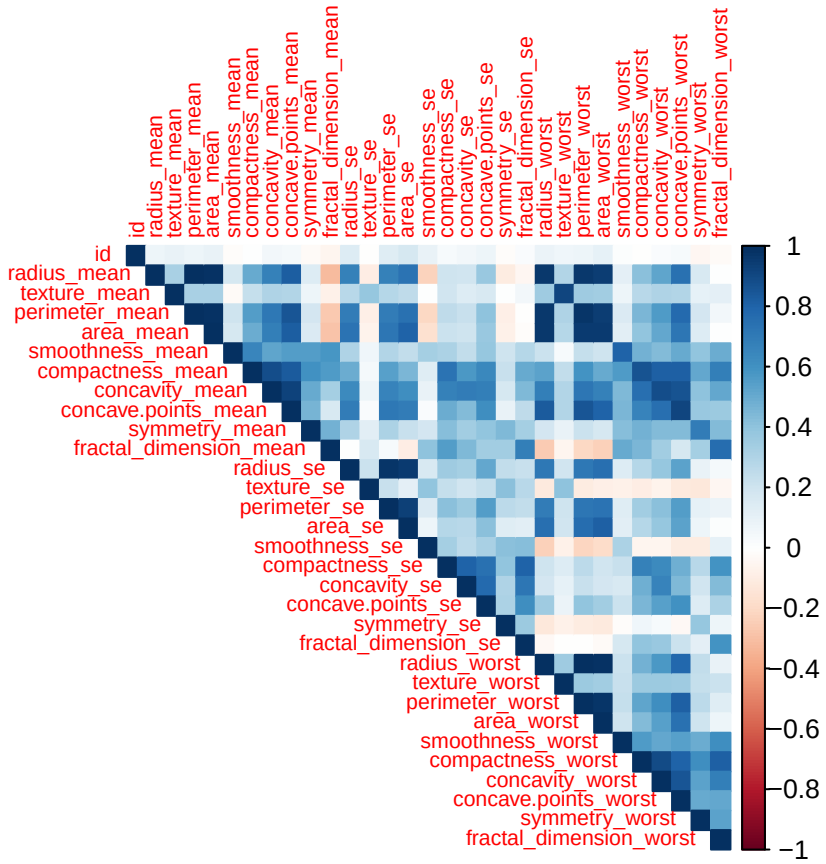
```
numeric_data <- data %>%
  select(where(is.numeric))

cor_matrix <-
  cor(numeric_data, use = "complete.obs", method = "pearson")
print(head(round(cor_matrix, 2)))
```

```
##                    id radius_mean texture_mean perimeter_mean area_mean
## id               1.00        0.07         0.10           0.07      0.10
## radius_mean      0.07        1.00         0.32           1.00      0.99
## texture_mean     0.10        0.32         1.00           0.33      0.32
## perimeter_mean   0.07        1.00         0.33           1.00      0.99
## area_mean        0.10        0.99         0.32           0.99      1.00
```

```
## smoothness_mean -0.01        0.17        -0.02           0.21      0.18
##                 smoothness_mean compactness_mean concavity_mean
## id                        -0.01            0.00           0.05
## radius_mean                0.17            0.51           0.68
## texture_mean              -0.02            0.24           0.30
## perimeter_mean             0.21            0.56           0.72
## area_mean                  0.18            0.50           0.69
## smoothness_mean            1.00            0.66           0.52
##                 concave.points_mean symmetry_mean fractal_dimension_mean
## id                             0.04         -0.02                  -0.05
## radius_mean                    0.82          0.15                  -0.31
## texture_mean                   0.29          0.07                  -0.08
## perimeter_mean                 0.85          0.18                  -0.26
## area_mean                      0.82          0.15                  -0.28
## smoothness_mean                0.55          0.56                   0.58
##                 radius_se texture_se perimeter_se area_se smoothness_se
## id                   0.14      -0.01         0.14    0.18          0.10
## radius_mean          0.68      -0.10         0.67    0.74         -0.22
## texture_mean         0.28       0.39         0.28    0.26          0.01
## perimeter_mean       0.69      -0.09         0.69    0.74         -0.20
## area_mean            0.73      -0.07         0.73    0.80         -0.17
## smoothness_mean      0.30       0.07         0.30    0.25          0.33
##                 compactness_se concavity_se concave.points_se symmetry_se
## id                        0.03         0.06              0.08       -0.02
## radius_mean               0.21         0.19              0.38       -0.10
## texture_mean              0.19         0.14              0.16        0.01
## perimeter_mean            0.25         0.23              0.41       -0.08
## area_mean                 0.21         0.21              0.37       -0.07
## smoothness_mean           0.32         0.25              0.38        0.20
##                 fractal_dimension_se radius_worst texture_worst perimeter_worst
## id                              0.03         0.08          0.06            0.08
## radius_mean                    -0.04         0.97          0.30            0.97
## texture_mean                    0.05         0.35          0.91            0.36
## perimeter_mean                 -0.01         0.97          0.30            0.97
## area_mean                      -0.02         0.96          0.29            0.96
## smoothness_mean                 0.28         0.21          0.04            0.24
##                 area_worst smoothness_worst compactness_worst concavity_worst
## id                    0.11             0.01              0.00            0.02
## radius_mean           0.94             0.12              0.41            0.53
## texture_mean          0.34             0.08              0.28            0.30
## perimeter_mean        0.94             0.15              0.46            0.56
## area_mean             0.96             0.12              0.39            0.51
## smoothness_mean       0.21             0.81              0.47            0.43
##                 concave.points_worst symmetry_worst fractal_dimension_worst
## id                              0.04         -0.04                   -0.03
## radius_mean                     0.74          0.16                    0.01
## texture_mean                    0.30          0.11                    0.12
## perimeter_mean                  0.77          0.19                    0.05
## area_mean                       0.72          0.14                    0.00
## smoothness_mean                 0.50          0.39                    0.50
```

Visualization of correlations

```
corrplot(cor_matrix, method = "color", type = "upper", tl.cex = 0.7)
```

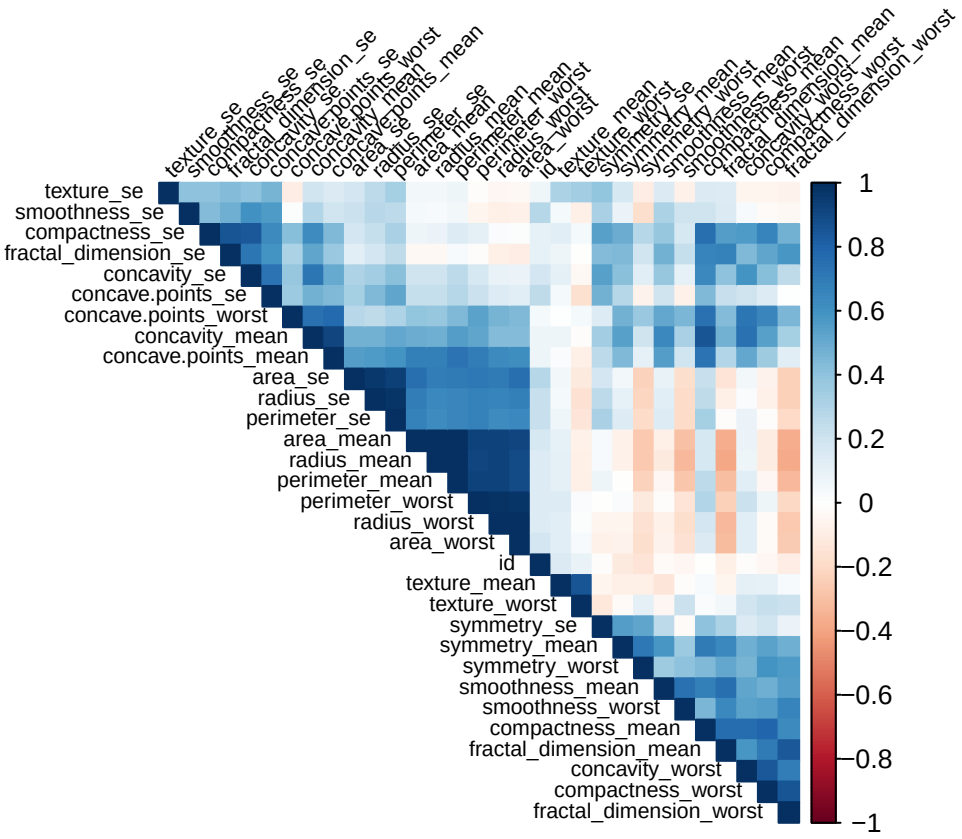


Correlation by Diagnosis Group

```
malignant <- data |>
  filter(diagnosis == "M") |>
  select(where(is.numeric))

benign <- data |>
  filter(diagnosis == "B") |>
  select(where(is.numeric))

cor_m <- round(cor(malignant, use = "complete.obs"), 2)
cor_b <- round(cor(benign, use = "complete.obs"), 2)
```

```
corrplot(
  cor_m,
  method = "color",
  type = "upper",
  order = "hclust",
  tl.cex = 0.7,
  tl.col = "black",
  tl.srt = 45
)
```

```
title("Malignant Tumors", outer = TRUE, line = -1, cex.main = 1.5)
```



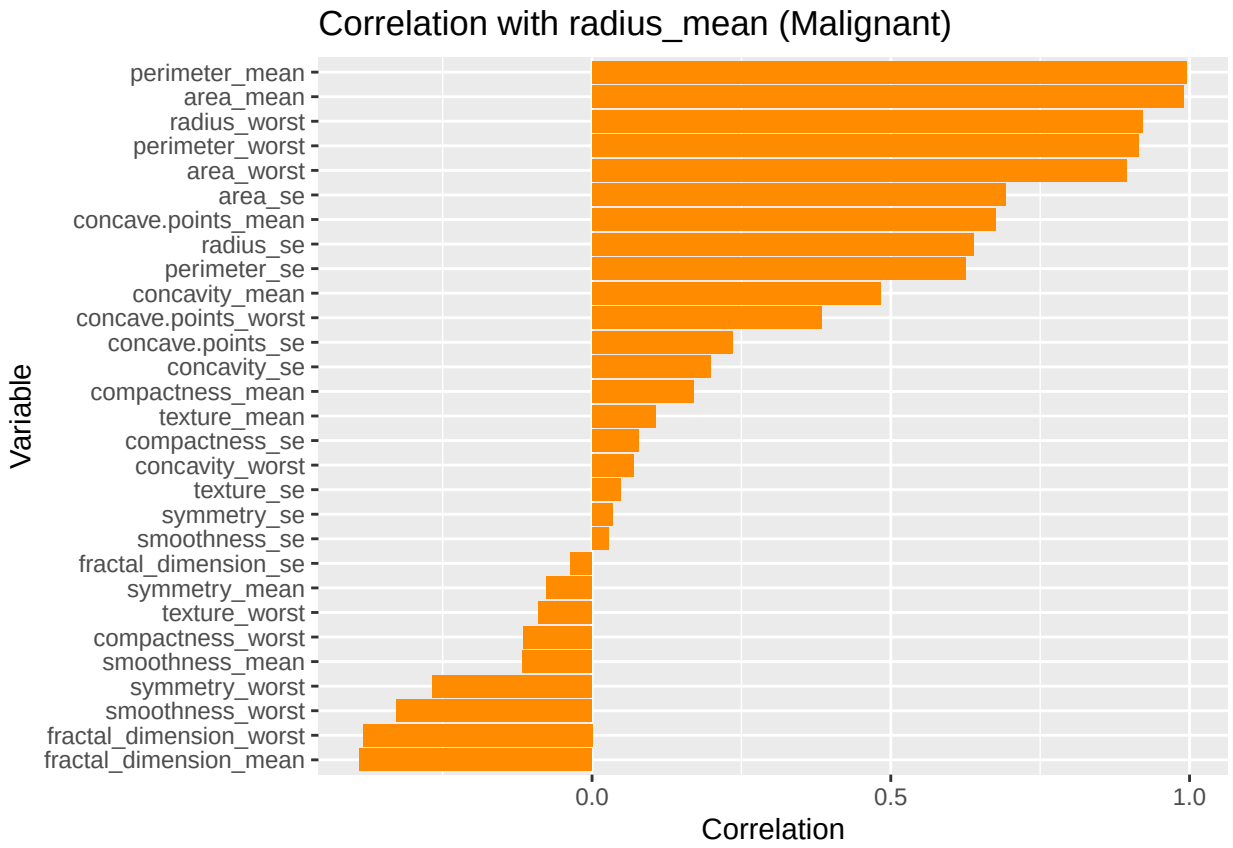
Correlations with radius_mean(Malgnant)

```
malignant_numeric <- malignant |>
  select(where(is.numeric)) |>
  select(- id)

cor_vec <- sapply(malignant_numeric,
                  function(x) cor(x,
                                  malignant_numeric$radius_mean,
                                  use = "complete.obs"))
cor_vec <- cor_vec[names(cor_vec) != "radius_mean"]
cor_df <- data.frame(
  variable = names(cor_vec),
  correlation = cor_vec
)
cor_df
```

```
##                                    variable correlation
## texture_mean                   texture_mean  0.10651642
```

```
## perimeter_mean                  perimeter_mean  0.99528148
## area_mean                            area_mean  0.99007845
## smoothness_mean                smoothness_mean -0.11603615
## compactness_mean              compactness_mean  0.16916808
## concavity_mean                  concavity_mean  0.48275037
## concave.points_mean        concave.points_mean  0.67551912
## symmetry_mean                    symmetry_mean -0.07644230
## fractal_dimension_mean  fractal_dimension_mean -0.38867852
## radius_se                              radius_se  0.63926965
## texture_se                            texture_se  0.04687304
## perimeter_se                        perimeter_se  0.62478537
## area_se                                  area_se  0.69235898
## smoothness_se                      smoothness_se  0.02803740
## compactness_se                    compactness_se  0.07825081
## concavity_se                        concavity_se  0.19757070
## concave.points_se              concave.points_se  0.23470949
## symmetry_se                          symmetry_se  0.03372713
## fractal_dimension_se        fractal_dimension_se -0.03676151
## radius_worst                        radius_worst  0.92165330
## texture_worst                      texture_worst -0.08888970
## perimeter_worst                  perimeter_worst  0.91487709
## area_worst                            area_worst  0.89393498
## smoothness_worst                smoothness_worst -0.32664876
## compactness_worst              compactness_worst -0.11482397
## concavity_worst                  concavity_worst  0.06962069
## concave.points_worst        concave.points_worst  0.38346081
## symmetry_worst                    symmetry_worst -0.26637550
## fractal_dimension_worst fractal_dimension_worst -0.38336710
```

```r
top_cor <- cor_df %>%
  arrange(desc(abs(correlation)))

ggplot(top_cor, aes(x = reorder(variable, correlation), y = correlation)) +
  geom_col(fill = "darkorange") +
  coord_flip() +
  labs(title = "Correlation with radius_mean (Malignant)",
       x = "Variable", y = "Correlation")
```



Correlation between diagnosis and tumor size

```
data$diagnosis_num <- ifelse(data$diagnosis == "M", 1, 0)

cor(data$diagnosis_num,
    data$radius_mean,
    use = "complete.obs", method = "pearson")
```

```
## [1] 0.7300285
```

Compare Means(t-test)

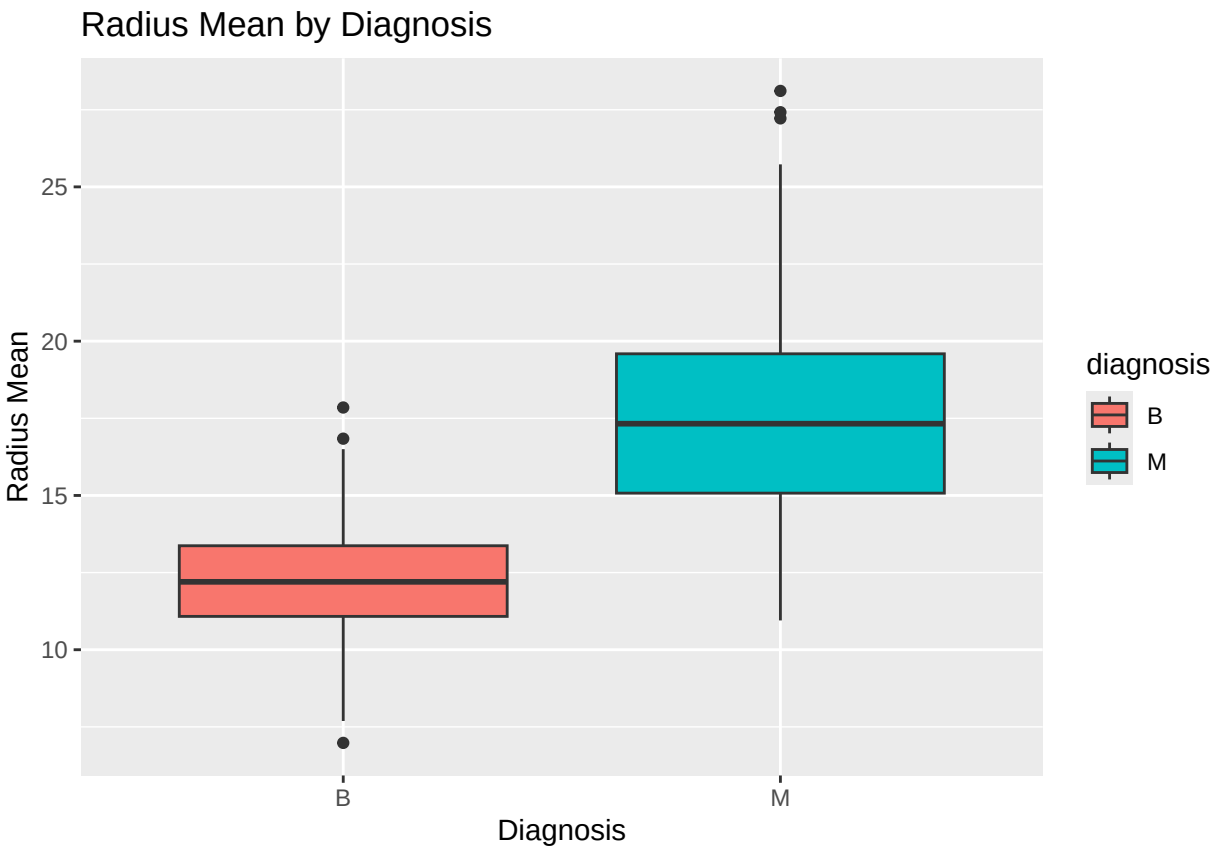
```
t.test(radius_mean ~ diagnosis, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  radius_mean by diagnosis
## t = -22.209, df = 289.71, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group B and group M is not equal to 0
## 95 percent confidence interval:
##  -5.787448 -4.845165
## sample estimates:
## mean in group B mean in group M
##        12.14652        17.46283
```

The p-value of such one-sided t-test is small enough. Thus, we are 95% confident to conclude that the tumor

size matters in diagnosis of breast cancor. Also, we can see that the all value inside the 95% confidence interval is negative, which intensifies the conclusion we made before that tumor size is an important factor in diagnosis the breast cancer.

```
ggplot(data, aes(x = diagnosis, y = radius_mean, fill = diagnosis)) +
  geom_boxplot() +
  labs(title = "Radius Mean by Diagnosis", y = "Radius Mean", x = "Diagnosis")
```



From the plot above, we can easily see the differece of mean tumor size between maglinant group and benign group.