Lab 8

AUTHOR
Dina Al Jibori

```
# load data and set "B" (benign) as the reference level
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.3     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.4     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
cells <- read_csv("https://www.dropbox.com/s/0rbzonyrzramdgl/ce
  mutate(diagnosis = factor(diagnosis, levels = c("B", "M")))
```

```
Rows: 569 Columns: 31
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (1): diagnosis
dbl (30): radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_m...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

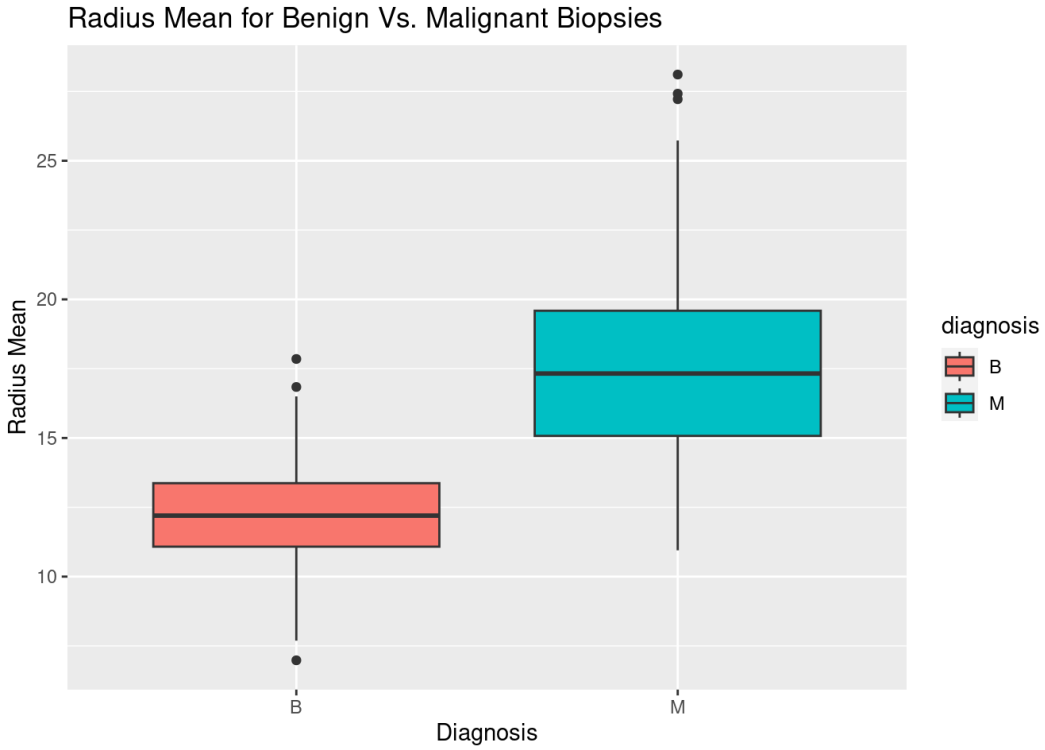
Question 1

The unit of observation for this would be an individual biopsy

Question 2

```
cells %>%
```

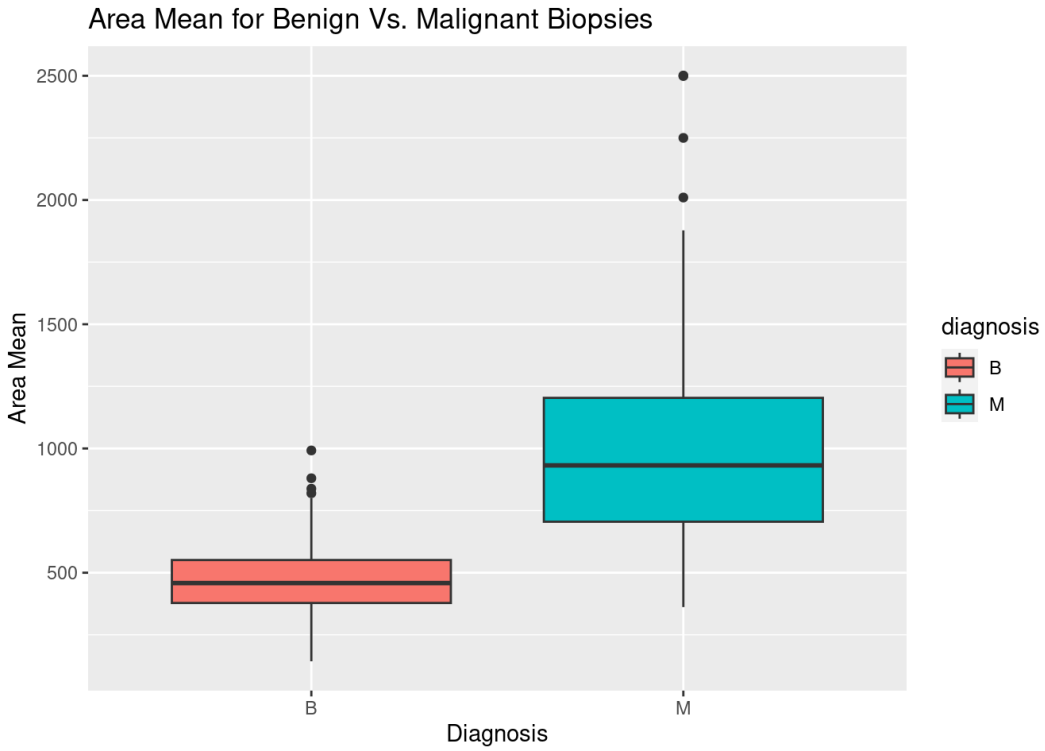
```
  ggplot(aes(x = diagnosis, y = radius_mean , fill = diagnosis))
  geom_boxplot()+ labs(y= "Radius Mean", x= "Diagnosis",
        title = "Radius Mean for Benign Vs. Malignant Biopsies")
```



The Key takeaway from this plot is that the average radius of maliginant tumors is much higher than the average rate radius for benign tumors. This information could help in diagnoaing maliginant and benign tumors based on the radius of the tumor.

Question 3

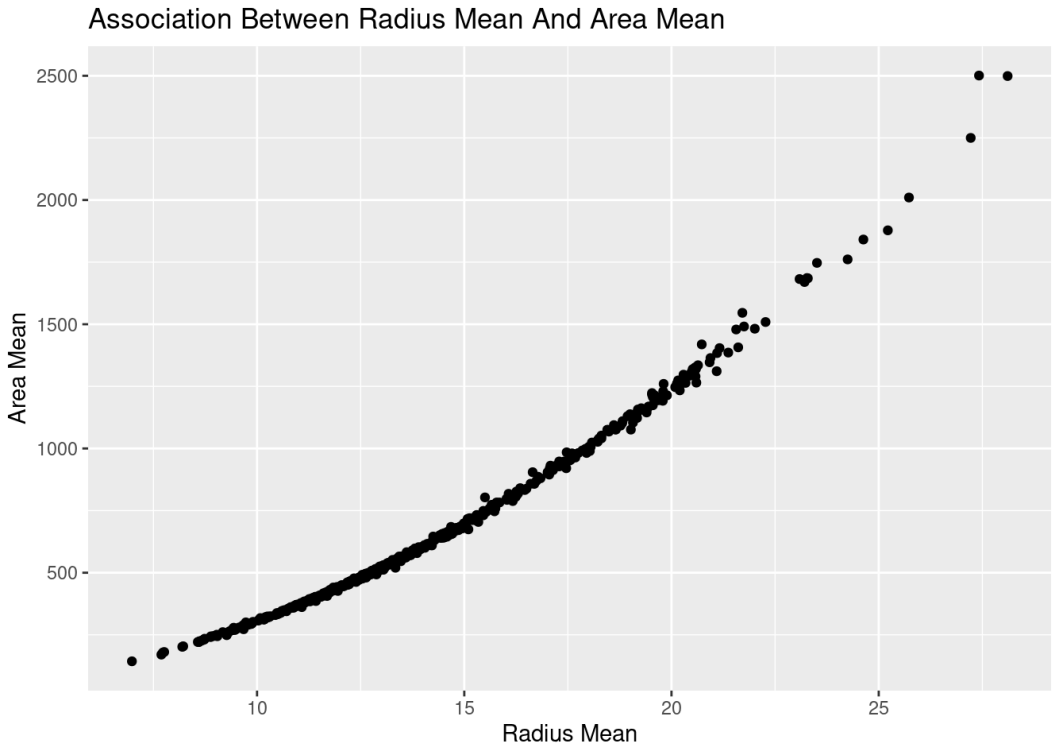
```
cells %>%
  ggplot(aes(x = diagnosis, y = area_mean , fill = diagnosis))
  geom_boxplot()+ labs(y= "Area Mean", x= "Diagnosis",
        title = "Area Mean for Benign Vs. Malignant Biopsies")
```



The Key takeaway from this plot is that the average area of maliginant tumors is much higher than the average area for benign tumors. This information could help in diagnoaing maliginant and benign tumors based on the radius of the tumor.

Question 4

```
cells %>%
  ggplot(aes(x = radius_mean, y = area_mean)) +
  geom_point()+ labs(y= "Area Mean", x= "Radius Mean",
    title = "Association Between Radius Mean And Area Mean")
```



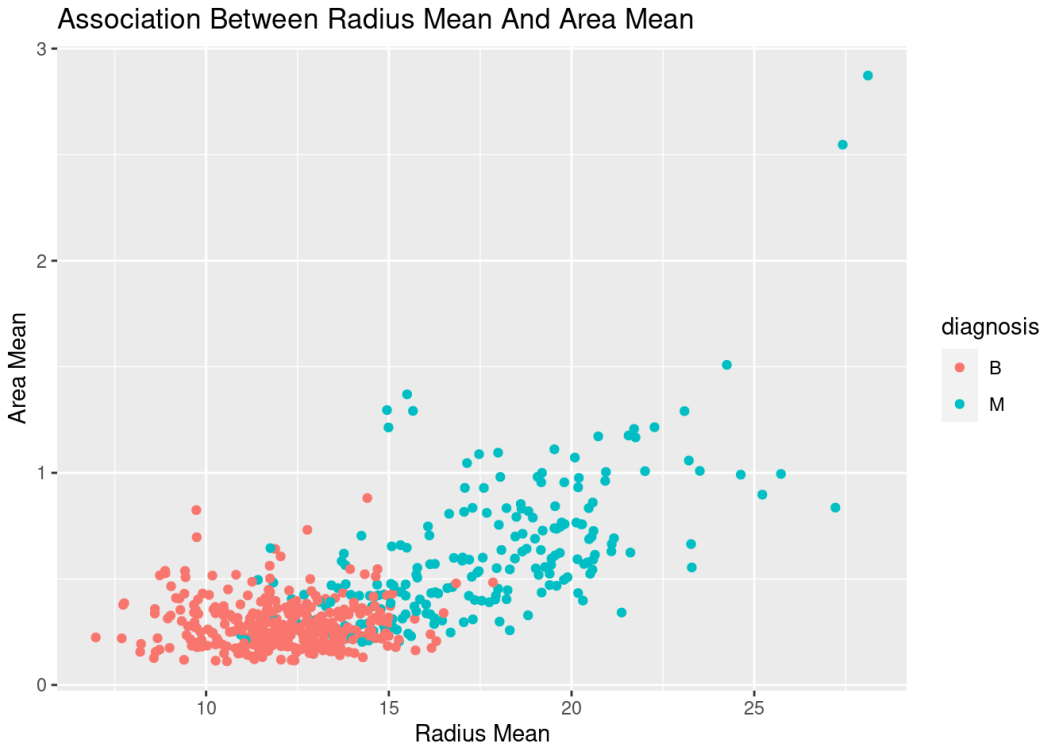

```
print(cor(cells$radius_mean, cells$area_mean))
```

```
[1] 0.9873572
```

The plot shows a strong, nonlinear, exponential, and positive relationship between Radius Mean and Area Mean. The correlation coefficiant also confirms it with it being 0.987. This shape most likely occurs because of the the radius is part of the equation to calculate area for circular objects so therefore as radius increases area will in turn also increase

Question 5

```
cells %>%
  ggplot(aes(x = radius_mean, y = radius_sd , color = diagnosis
  geom_point()+ labs(y= "Area Mean", x= "Radius Mean",
            title = "Association Between Radius Mean And Area M
```




```
cells %>%
  group_by(diagnosis) %>%
  summarise(correlation = cor(radius_mean, radius_sd)) %>%
  print()
```

```
# A tibble: 2 × 2
  diagnosis correlation
  <fct>           <dbl>
1 B             -0.0278
2 M              0.639
```

The correlation coefficiant for Benign Biopsies is around -0.28 which means theres a weak negative relationship between the radius mean and radius standard deviation. This means its most likely that the a smaller radius has nore variability. While the correlation coefficiant for Benign Biopsies is around 0.64 which means theres a medium positive relationship between the radius mean and radius standard deviation. This means its most likely that the a larger radius has more variability. The relationship between radius_mean and radius_sd are different for benign biopsies vs. malignant biopsies as seen with their differing coorelations. An explination for these differences can be explained by how benign tumors have more variability in sizes while Malignant tumors have a tendency to be larger sizes.

Question 6

```
set.seed(123)
set_type <- sample(x = c('train', 'test'),
                   size = 569,
                   replace = TRUE,
                   prob = c(0.8, 0.2))
cells <- cells %>%
  mutate(set_type = set_type)

cells_train <-cells %>%
  filter(set_type == 'train')

cells_test <- cells %>%
  filter(set_type == 'test')
```

There are 110 observation in the teams set while there are 459 observations of the training data set.

Question 7

```
model1<- glm(diagnosis~texture_mean, data=cells_train, family= '
model1
```

```
Call:  glm(formula = diagnosis ~ texture_mean, family =
"binomial",
    data = cells_train)

Coefficients:
 (Intercept)  texture_mean
     -5.5245        0.2547

Degrees of Freedom: 458 Total (i.e. Null);  457 Residual
Null Deviance:      602.9
Residual Deviance: 511  AIC: 515
```

```
probability <- exp(5.5245 + 0.2547 * 15) / (1 + exp(5.5245 + 0.2
probability
```

```
[1] 0.9999126
```

The predicted probability for a biopsy with a mean texture of 15 is approximately 0.9999126. This means that the model assigns a very high probability for this biopsy being malignant.

Question 8

```r
p_hat_train <- predict(model1, cells_train, type = "response")

cells_train %>%
  mutate(p_hat = p_hat_train,
         y_hat = ifelse(p_hat > 0.5, "M", "B")) %>%
  summarize(misclass_train = mean(diagnosis != y_hat))
```

```
# A tibble: 1 × 1
  misclass_train
           <dbl>
1          0.281
```

```r
p_hat_test <- predict(model1, cells_test, type = "response")

cells_test %>%
  mutate(p_hat = p_hat_test,
         y_hat = ifelse(p_hat > 0.5, "M", "B")) %>%
  summarize(misclass_test = mean(diagnosis != y_hat))
```

```
# A tibble: 1 × 1
  misclass_test
           <dbl>
1         0.336
```

There is evidence that my model is overfitting due to the misclassification rate for testing being higher than the misclassification rate for training.

Question 9

```r
model2<- glm(diagnosis~texture_mean + compactness_mean + concavi
        concave.points_mean + fractal_dimension_mean, data=cells
model2
```

```
Call:  glm(formula = diagnosis ~ texture_mean +
compactness_mean + concavity_mean +
    concave.points_mean + fractal_dimension_mean, family =
"binomial",
    data = cells)

Coefficients:
           (Intercept)            texture_mean
compactness_mean
               -2.9343                  0.3051
1.1341
```

```
       concavity_mean     concave.points_mean
fractal_dimension_mean
                5.9363                129.6060
-169.3931

Degrees of Freedom: 568 Total (i.e. Null);  563 Residual
Null Deviance:      751.4
Residual Deviance: 176.2    AIC: 188.2
```

```
misclass_train <- cells_train %>%
  mutate(predictions_train = predict(model2, newdata = ., type =
         predicted_labels_train = ifelse(predictions_train >= 0.
  summarize(misclass_train = mean(predicted_labels_train != diag
misclass_train
```

```
# A tibble: 1 × 1
  misclass_train
           <dbl>
1         0.0501
```

```
misclass_test <- cells_test %>%
  mutate(predictions_test = predict(model2, newdata = ., type =
         predicted_labels_test = ifelse(predictions_test >= 0.5,
  summarize(misclass_test = mean(predicted_labels_test != diagno
misclass_test
```

```
# A tibble: 1 × 1
  misclass_test
           <dbl>
1        0.0909
```

There is evidence that my model is overfitting due to the misclassification rate for testing being higher than the misclassification rate for training but the difference is too small to make a definitive conclusion.

Question 10

A type II error is much worse because a false negative causes a delay to treatments if any which ultimately can cause the death of a patient.

Question 11

```
predictions2 <- cells_test %>%
  mutate(y_hat1 = predict(model1, newdata = ., type = "response"
  mutate(y_hat_label = ifelse(y_hat1 >= 0.5, "M", "B")) %>%
  mutate(falseNeg = ifelse(y_hat_label == "B" & diagnosis == "M"
```

```
summarise(predictions2, FalseNegatives = sum(falseNeg == "yes"))
```

```
# A tibble: 1 × 1
  FalseNegatives
           <int>
1             23
```

Question 12

```
predictions3 <- cells_test %>%
  mutate(y_hat1 = predict(model1, newdata = ., type = "response"
  mutate(y_hat_label = ifelse(y_hat1 >= 0.3, "M", "B")) %>%
  summarize(FalseNegatives = sum(ifelse(y_hat_label == "B" & dia
predictions3
```

```
# A tibble: 1 × 1
  FalseNegatives
           <dbl>
1              8
```

To lower the number of false negatives, I adjusted the classification threshold from 0.5 to 0.3.

Question 13

```
misclassification_rate <- cells_test %>%
  mutate(y_hat1 = predict(model1, newdata = ., type = "response"
         y_hat_label = ifelse(y_hat1 >= 0.3, "M", "B")) %>%
  summarize(MisclassificationRate = mean(y_hat_label != diagnos
misclassification_rate
```

```
# A tibble: 1 × 1
  MisclassificationRate
                  <dbl>
1                 0.327
```

The misclassification rate went down slightly 32.7% from the original 33.6%. This is due to the threshhold highly affecting how specific the classification rate is for diagnosis

Question 14

Many things are gained by shifting to algorithmic diagnoses like it can identify patterns of illnesses quicker, and most importantly is how efficient

it is since it’s able to diagnose much quicker. its also consistent in diagnoses. However theres many things that lost by shifting to algorithmic diagnoses like ethical concerns of algorithmic biases, the lack of understanding and empathy a machince can have and the lack of human interaction.