HW_4_Geary

Marion Geary

```
library(tidyverse)
library(tidymodels)
library(skimr)
# setwd("/Users/Marion/Desktop/math386/hw/hw-4")
load('rad.Rdata')
```

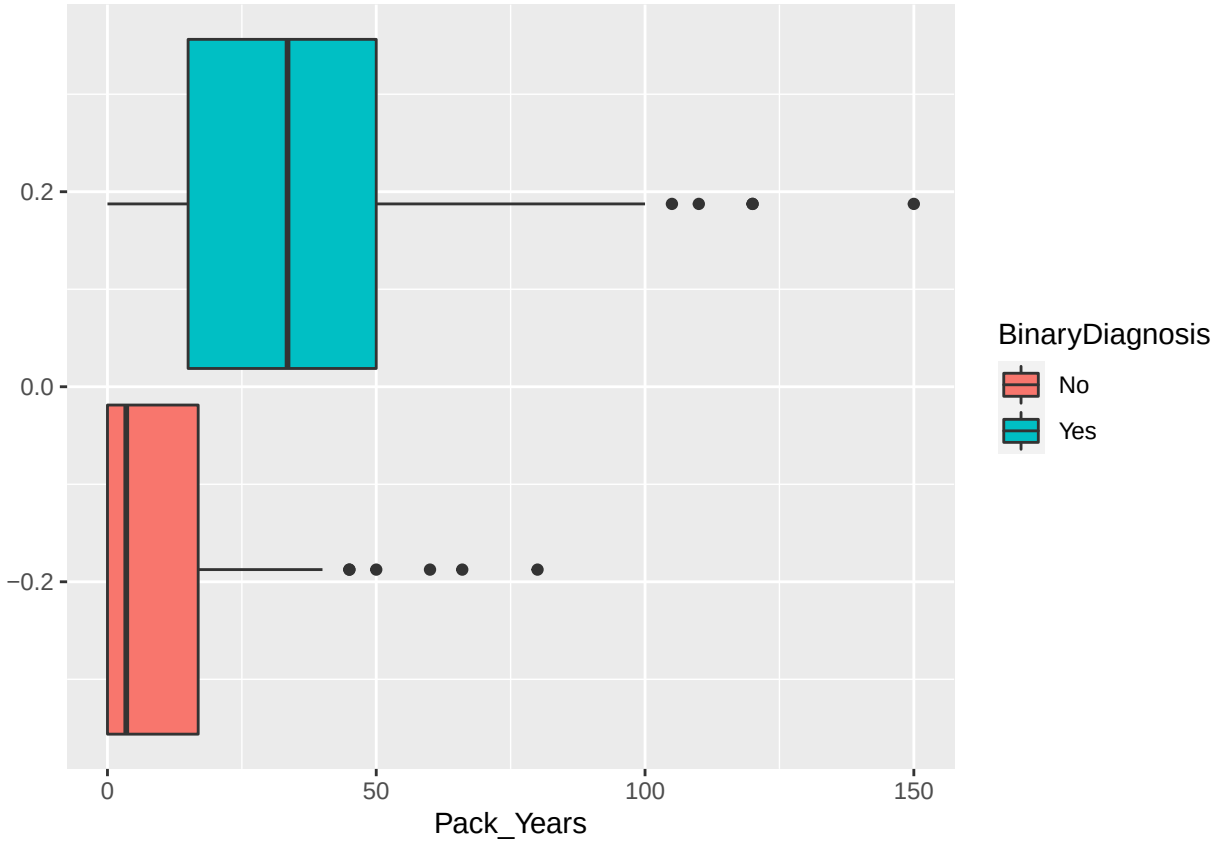
Exercise 1

```
rad <- rad %>% mutate(Sex = as.factor(Sex))
# skim(rad %>% select(BinaryDiagnosis:Sex))
```

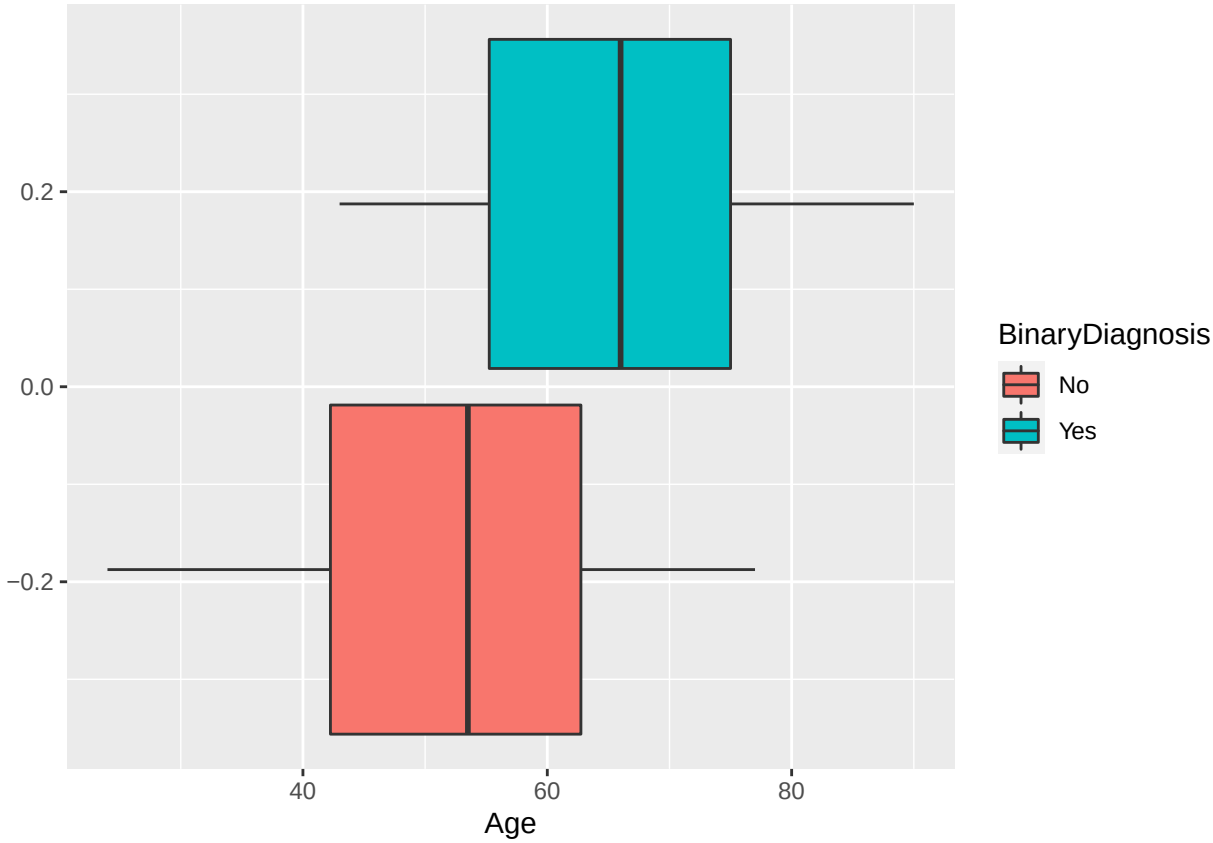
From `skim()`, I can see that the binary diagnosis has a fairly even split between yes and no. Sex also has a fairly even split, with slightly more females. The average age is 59.9 with a standard deviation of 13.7, so age varies pretty significantly. Pack years has an average of 26.2 with a standard deviation of 29.0, so the data is very right skewed, probably due to a mix of non-smokers with a few heavy smokers.

Exercise 2

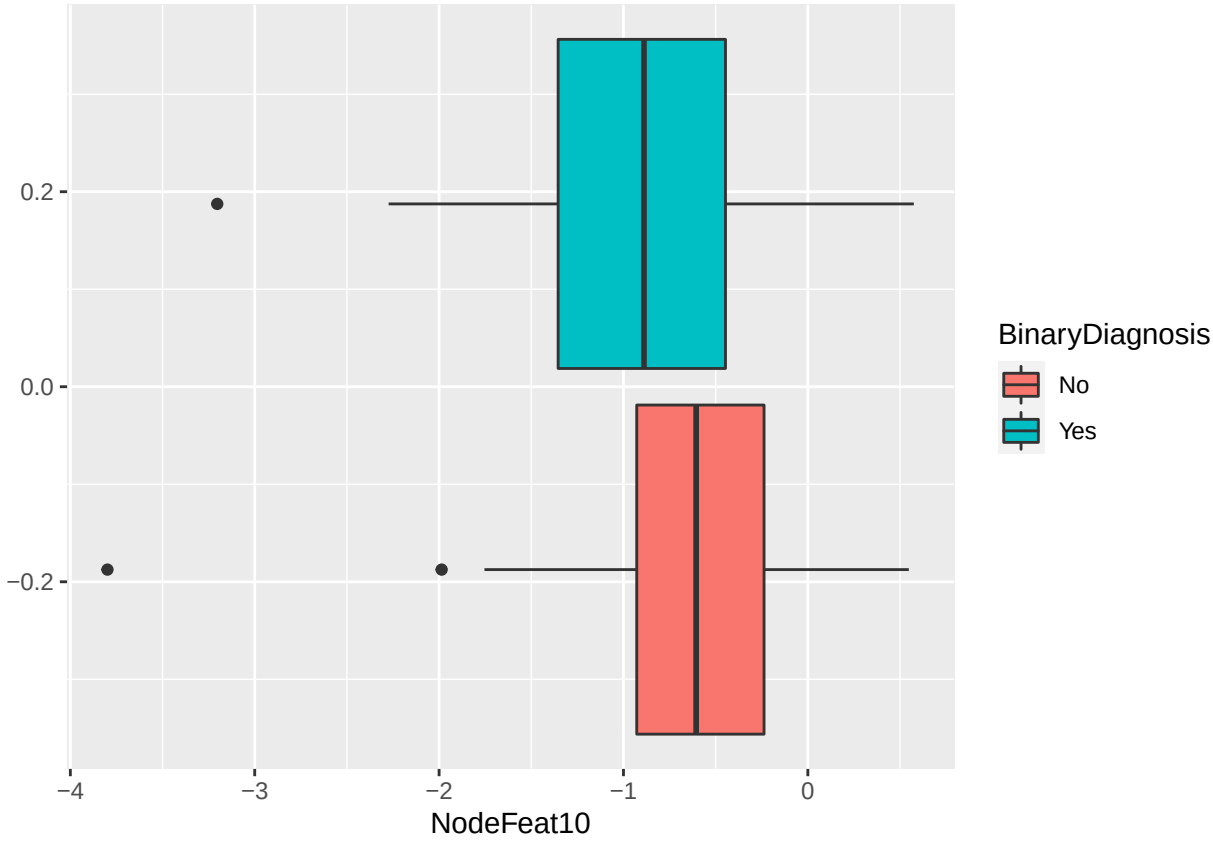
```
ggplot(data=rad) + geom_boxplot(aes(x = Pack_Years, fill=BinaryDiagnosis))
```



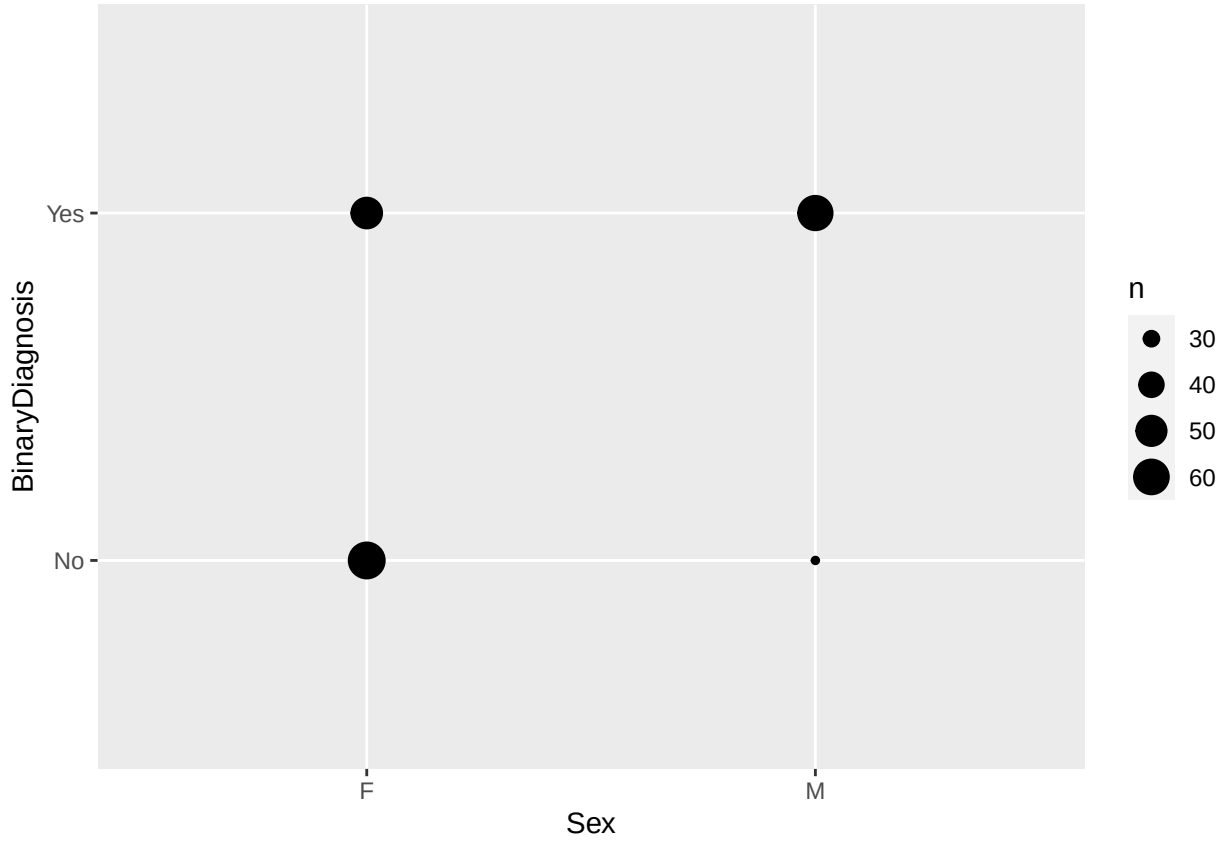

```
ggplot(data = rad, aes(x = Age)) + geom_boxplot(aes(fill = BinaryDiagnosis))
```



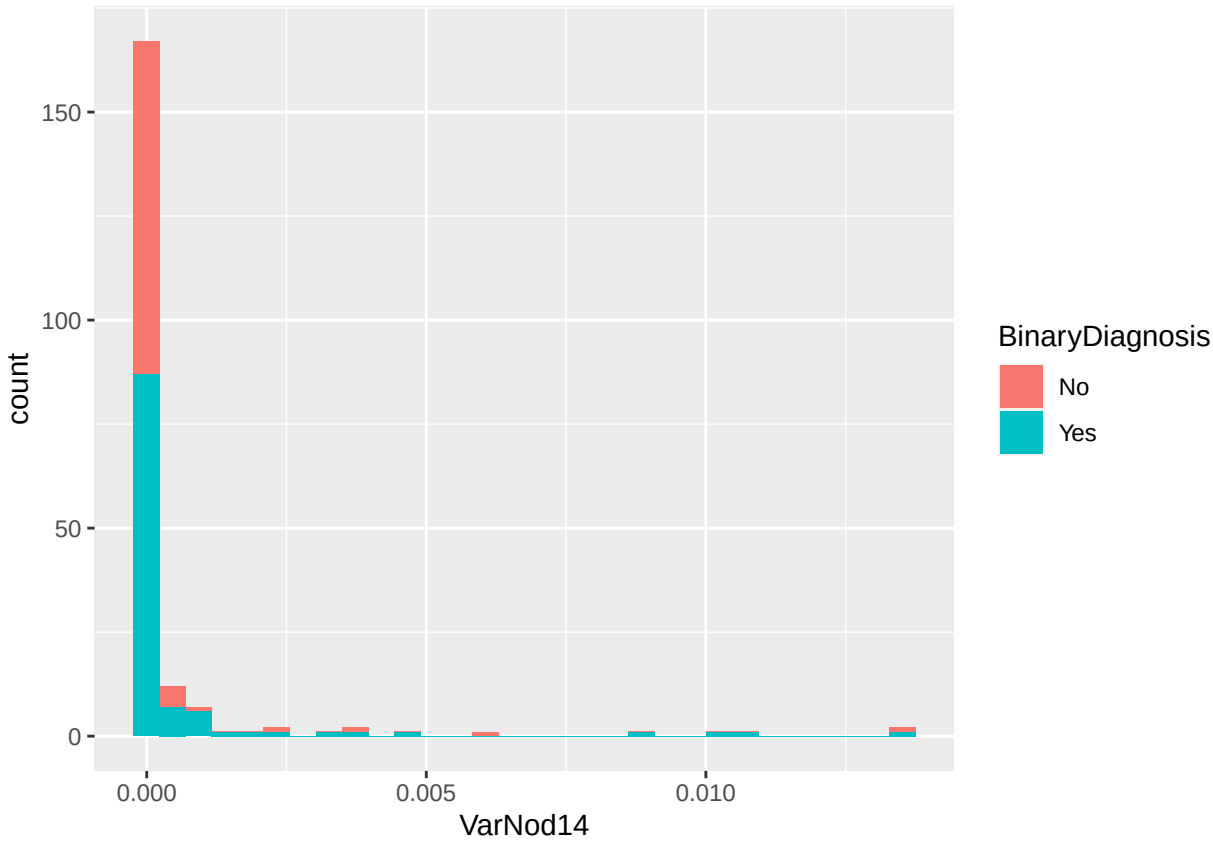

```
ggplot(data = rad, aes(x = NodeFeat10)) + geom_boxplot(aes(fill = BinaryDiagnosis))
```



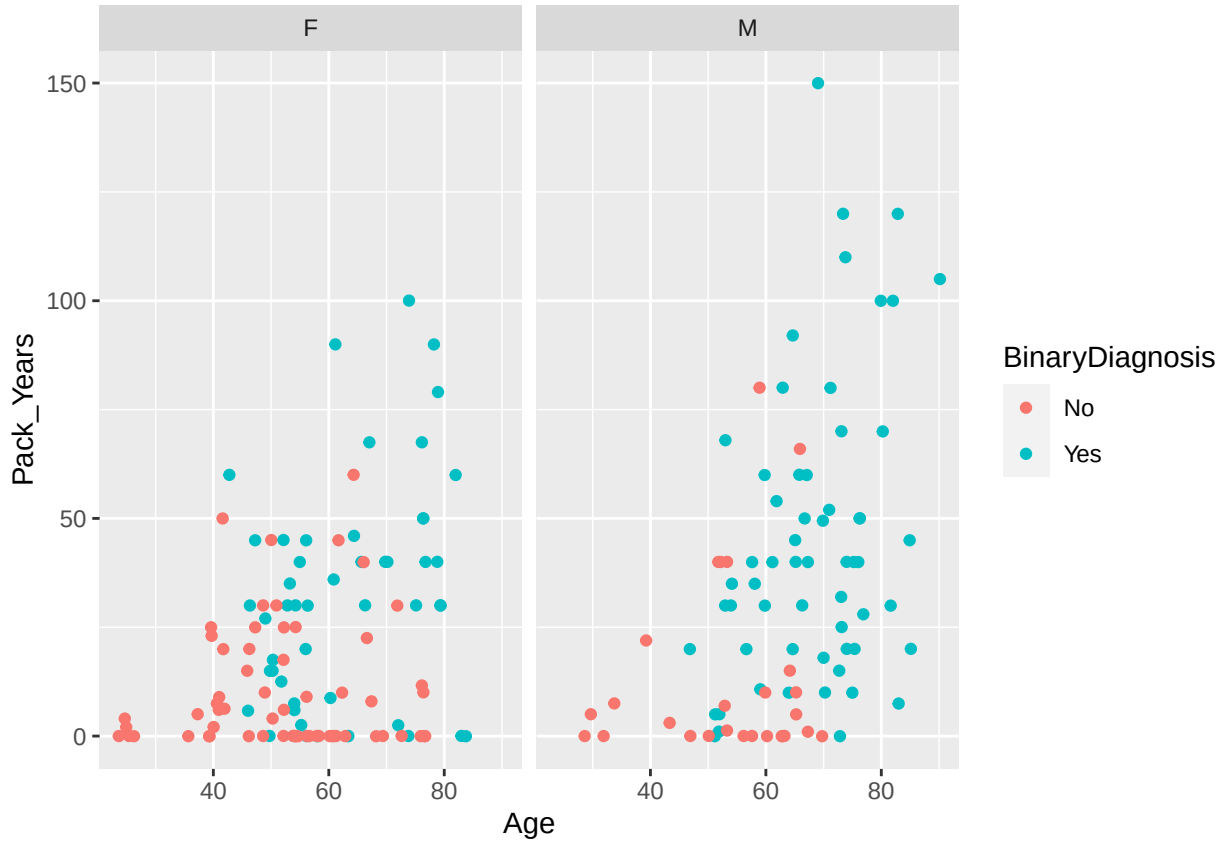

```
ggplot(data = rad) + geom_count(aes(x = Sex, y = BinaryDiagnosis))
```




```
ggplot(data = rad, aes(x = VarNod14)) + geom_histogram(aes(fill = BinaryDiagnosis))
```




```
ggplot(data = rad, aes(x = Age, y = Pack_Years)) + geom_jitter(aes(color = BinaryDiagnosis)) + facet_wra
```



The first plot shows that `Pack_Years` is a significant predictor for `BinaryDiagnosis`, with the median `Pack_Years` for `Yes` being over 30 `Pack_Years` over `No`. The next plot shows that `Age` is also a significant predictor of `BinaryDiagnosis`, with `Yes` having a median about 20 years higher than `No`. `NodeFeat10` may be a predictor for `BinaryDiagnosis`, with the box-plot showing that `Yes` has a lower median `NodeFeat10` than `No`. The count plot shows that `Sex` is likely a significant predictor of `BinaryDiagnosis`. It shows that most men in the study tested positive, while the split was pretty even between women. The histogram of `VarNod14` shows that `VarNod14` is not a good predictor of `BinaryDiagnosis`. The final plot shows all 4 non-radiomic variables. We see that `Pack_Years` and `Age` are both good predictors of `BinaryDiagnosis`, and these variables seem to have a stronger impact on males than females.

Exercise 3

```
knn_model <- nearest_neighbor(neighbors = 5, weight_func = "epanechnikov", dist_power = 2, mode = "class
```

Exercise 4

```
set.seed(987)
rad_split <- rad %>%
  initial_split(prop = .8)
rad_test <- testing(rad_split)
rad_train <- training(rad_split)
```

Exercise 5

```
rad_recipe <- recipe(BinaryDiagnosis ~ ., data = rad) %>%
  step_dummy(Sex) %>%
  step_normalize(all_predictors())

temp <- bake(prep(rad_recipe), new_data = NULL)
# glimpse(temp)
# commented out because of large output
```

Exercise 6

```
rad_wkflow <- workflow() %>%
  add_model(knn_model) %>%
  add_recipe(rad_recipe)

rad_fit <- fit(rad_wkflow, data = rad_train)

rad_res <- augment(rad_fit, new_data = rad_test)
```

Exercise 7

```
my_metrics <- metric_set(sens, spec, accuracy)
my_metrics(rad_res, truth = BinaryDiagnosis, estimate = .pred_class)
```

```
## # A tibble: 3 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 sens     binary         0.375
## 2 spec     binary         0.542
## 3 accuracy binary         0.475
```

Exercise 8

Sensitivity = $\frac{TP}{TP+FN}$ Sensitivity is the number of correctly predicted positives (in this case, `BinaryDiagnosis == Yes`) out of the real number of positives. It shows us what percentage of the patients with cancer are correctly being detected as positive out of those given a positive diagnosis.

Specificity = $\frac{TN}{TN+FP}$ Specificity is the number of correctly predicted negatives (`BinaryDiagnosis == No`) out of the real number of negatives. It shows what percentage of patients given a negative test result are actually cancer free. In this model, we care more about specificity as only want to give a negative result if we are absolutely sure a patient does not have cancer. The specificity here is low at 0.542, so the model needs refining.

Accuracy = $\frac{TN+TP}{TN+FP+TP+FN}$ Accuracy is the number of correct predictions out of the total number of predictions. In this model, we see that only 0.475 of cancer diagnoses are accurate. This is another sign that our model is not working well, as the model is giving an accurate diagnosis less than 50% of the time, making the model about the same as random chance.