Practicum #3

Due: Friday, December 3 @ 6pm

In these exercises, we will be using the iris data built into R. Begin by loading the data:

```
data(iris)
```

You can read about it using `?iris`

Problem 1:

Explore the iris data:

a. What variables are present? Are they continuous or categorical?

Four continuous variables: Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width. One categorical variable: Species.

b. For the categorical variables, provide a table of how many samples there are in each category. Show the R code you used. (*hint: check out the* `table` *function*)

```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

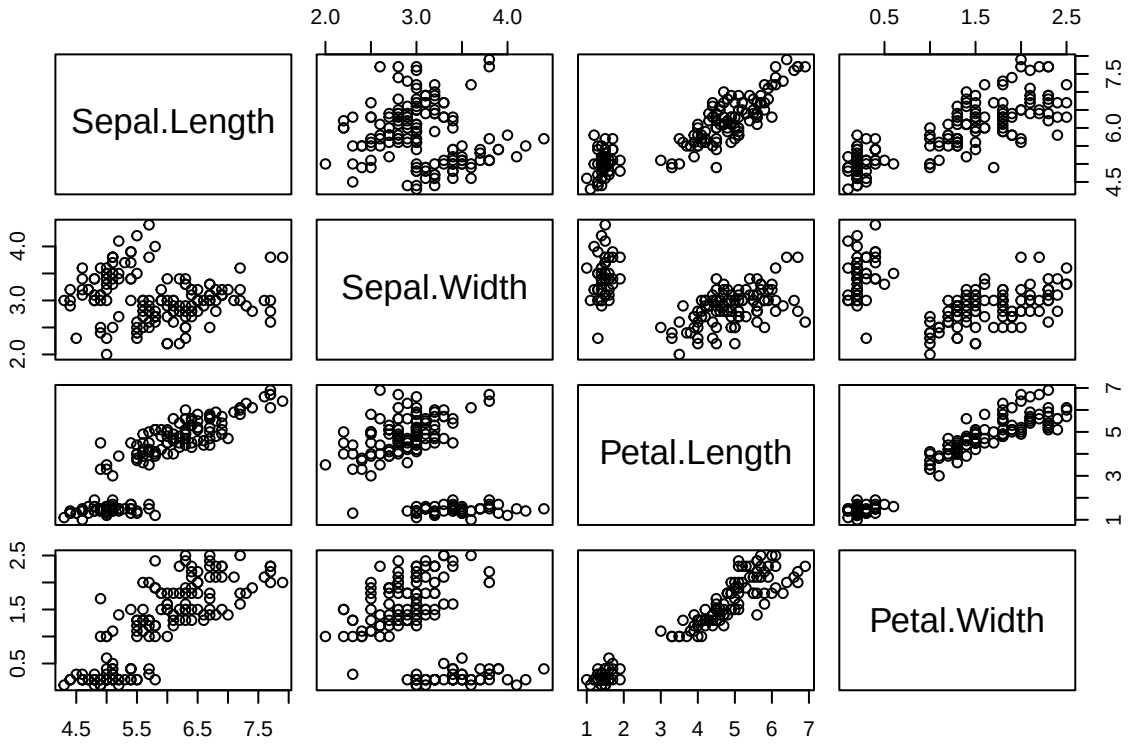
c. Provide the range (minimum, maximum) of the continuous variables.

minimum and maximum can be found with min() and max()

```
## # A tibble: 4 x 3
##   name           min   max
##   <chr>        <dbl> <dbl>
## 1 Petal.Length   1     6.9
## 2 Petal.Width    0.1   2.5
## 3 Sepal.Length   4.3   7.9
## 4 Sepal.Width    2     4.4
```

d. Another useful exploratory tool is the `pairs()` function, which plots the columns of a data frame against each other. Produce a scatterplot matrix (a pairs plot) of the continuous variables.

```
# first four columns
pairs(iris[,1:4])
```



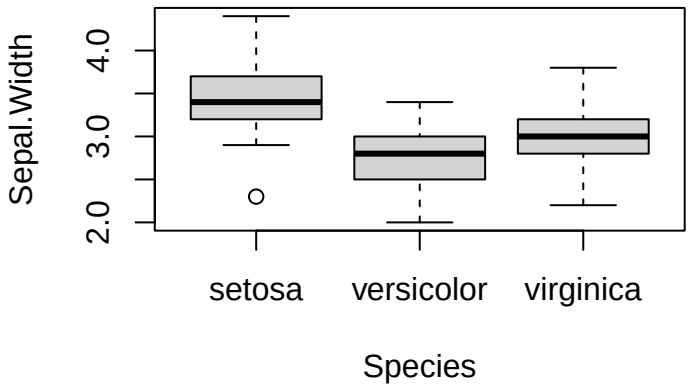

```
# or using dplyr
pairs(iris %>% dplyr::select(-Species))
```

Problem 2:

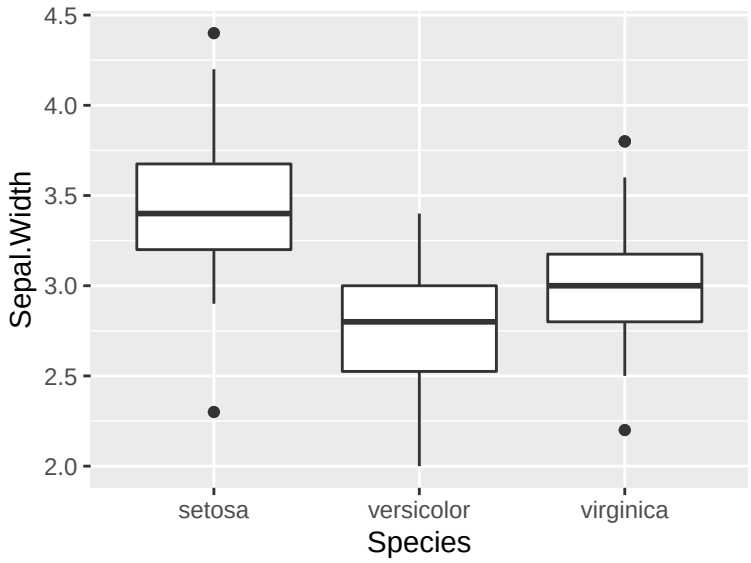
Do iris flower sizes depend on species?

a. In class, we learned about the formula syntax in R, which uses the tilde character `~` to mean "as a function of." Many functions use this syntax, including the `boxplot()` function. For example, you can generate a boxplot for two variables with `boxplot(y ~ x)`. Using this syntax (or `ggplot2` if you prefer), Produce boxplots of sepal width broken down by iris species.

```
boxplot(Sepal.Width ~ Species, data = iris)
```



```
# ggplot alternative
ggplot2::ggplot(iris) +
  ggplot2::aes(x = Species, y = Sepal.Width) +
  ggplot2::geom_boxplot()
```



b. For each iris species, compute the mean sepal width (Show your work.)

```
# several ways to do this
# cleanest with tidyverse:
iris %>%
  # group by a factor - in this case species
  dplyr::group_by(Species) %>%
  # then apply a function to all the data in each factor - in this case mean
  dplyr::summarize(mean = mean(Sepal.Width))
```

```
## # A tibble: 3 x 2
##   Species     mean
##   <fct>      <dbl>
## 1 setosa      3.43
## 2 versicolor  2.77
## 3 virginica   2.97
```

```
# other ways: using base R subsetting
mean(iris$Sepal.Width[iris$Species == "setosa"])
```

```
## [1] 3.428
```

```
mean(iris$Sepal.Width[iris$Species == "versicolor"])
```

```
## [1] 2.77
```

```r
mean(iris$Sepal.Width[iris$Species == "virginica"])
```

```
## [1] 2.974
```

```r
# using filter instead
mean(dplyr::filter(iris, Species == "setosa")$Sepal.Width)
```

```
## [1] 3.428
```

```r
# or using filter and pull
mean(dplyr::filter(iris, Species == "setosa") %>% dplyr::pull(Sepal.Width))
```

```
## [1] 3.428
```

```r
# other ways too!
```

c. For each iris species, compute the sample SD of the sepal width (Show your work.)

```r
# can calculate in any of the same ways as above ^
# for simplicity, I will use group_by() and summarize()

iris %>%
  dplyr::group_by(Species) %>%
  dplyr::summarize(sd = sd(Sepal.Width))
```

```
## # A tibble: 3 x 2
##   Species       sd
##   <fct>      <dbl>
## 1 setosa     0.379
## 2 versicolor 0.314
## 3 virginica  0.322
```

d. Using your answers above, fill out a complete ANOVA table for sepal width vs. species. Be sure to also calculate the F statistic. (Show your work.)

For the ANOVA table, we need the SS(between) and SS(within), df(between) and df(within), MS(between) and MS(within). And finally the F statistic.

```r
# SSB = sum for all categories: category_n(category_mean - grand_mean)^2

# grand mean = mean of all sepal.length...
mean(iris$Sepal.Width)
```

```
## [1] 3.057333
```

```r
# setosa:
dim(dplyr::filter(iris, Species == "setosa"))
```

```
## [1] 50  5
```

```
# there are 50 observations for each species...
SSB = 50*(3.43 - 3.0573)^2 + 50*(2.77-3.0573)^2 + 50*(2.97-3.0573)^2
SSB
```

```
## [1] 11.45339
```

```
# SSW = sum for all categories: (category_n - 1)(var_category)
SSW = (50-1)*0.379^2 + (50-1)*0.314^2 + (50-1)*0.322^2
SSW
```

```
## [1] 16.95013
```

MS = SS / df

df(between) = k-1 = 3-1 = 2

df(within) = N - k = 150 - 3 = 147

F = MS(between) / MS(within)

Putting it all together...

	SS	df	MS	F
Between	11.45	2	5.725	49.78
Within	16.95	147	0.115	

e. Based on your table, what is the standard deviation of sepal width for *all* species? How does it compare to `sd(iris$Sepal.Width)`?

```
# standard deviation of all samples = sqrt(MS(within))
sqrt(0.115)
```

```
## [1] 0.3391165
```

```
# compare to sd(population)
sd(iris$Sepal.Width)
```

```
## [1] 0.4358663
```

Pretty similar! But not exact.

f. Based on your table, and using the `p*` family of functions (such as `pnorm` and `pt`), compute a p value for the ANOVA. Also provide, in words, an interpretation of the result.

```
# for anova we are going to use the F distribution
pf(49.78, 2, 147, lower.tail = F)
```

```
## [1] 3.101069e-17
```

```
# note: answers may vary for (d) and (e) depending on rounding
```

The p-value is very small, leading us to reject the null hypothesis that the variance between the mean sepal lengths of the tree species does not contribute significantly to the overall variance observed in the sepal lengths. Alternatively, there is a difference between one or more of the three species sepal length.

g. In class, we learned about using `anova(lm(Y~X,data=myData))` to obtain an ANOVA analysis of `Y` versus the categorical variable `X` using the dataframe `myData`. Use this syntax to perform the ANOVA analysis you did by hand above. Do the results agree with yours?

```
anova(lm(Sepal.Width ~ Species, data = iris))
```

```
## Analysis of Variance Table
##
## Response: Sepal.Width
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 11.345  5.6725   49.16 < 2.2e-16 ***
## Residuals 147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes everything agrees! I think I was off a bit because of rounding... but either way the p-value is highly significant. Note: R does not return values $< 2.2e\text{-}16$ because it might be imprecise.

h. Remember, the ANOVA is a single test which tells us if the three species's distribution of sepal widths came from the same overall distribution or from two or more distributions. However, if we reject the null hypothesis, it does not tell us which species differ. Use a TukeyHSD test to perform pairwise t-tests between species to see which, if any, comparisons are significantly different.

```
# remember, must use aov() with TukeyHSD instead of anova(lm())
TukeyHSD(aov(Sepal.Width ~ Species, data = iris))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sepal.Width ~ Species, data = iris)
##
## $Species
##                       diff         lwr        upr     p adj
## versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
## virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
## virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

The Tukey HSD test performs pairwise comparisons of differences in means and adjusts the p-value to take into account the multiple testing. For this example, it looks like all pairwise comparisons are significant at an experiment wide significance level of 0.05 (p-value < 0.009)

i. Compare a regular t-test for sepal width of versicolor vs. virginica. How do the p-values differ? Was this what you expected?

```
t.test(dplyr::filter(iris, Species == "versicolor")$Sepal.Width,
       dplyr::filter(iris, Species == "virginica")$Sepal.Width)
```

```
##
##  Welch Two Sample t-test
##
## data:  dplyr::filter(iris, Species == "versicolor")$Sepal.Width and dplyr::filter(iris, Species == "
## t = -3.2058, df = 97.927, p-value = 0.001819
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.33028364 -0.07771636
## sample estimates:
## mean of x mean of y
##     2.770     2.974
```

The p-value from this t-test is 0.0018, which is smaller than the p-value obtained from the TukeyHSD test. This makes sense because the TukeyHSD test is accounting for multiple testing so it is likely resulting in a higher (less significant) adjusted p-value.

Problem 3

ANOVA assumes that the groups have equal variances. Let's check that assumption.

a. Consider the boxplots of sepal **length** broken down by iris species that you produced in part (a) of the previous problem. By eye, do you believe the variances are equal?

The setosa box looks narrower than the others, but it is hard to tell if it is significant by eye...

b. Although we learned about using the F statistic to compare MS(between) and MS(within), we can use an F test for a variety of things. For example, if we wanted to know if the variance of the setosa and virginica sepal lengths are equal, we can use an F test for equality of variance:

$$F = \frac{Var(Y)}{Var(X)}$$

- Perform an F test to compare two variances using the `var.test()` function. What is your interpretation of these results? (*hint: supply* ***`var.test()`*** *with the Sepal.Length for one species as* ***`x`*** *and the other species as* ***`y`***)

```
var.test(dplyr::filter(iris, Species == "setosa")$Sepal.Length, dplyr::filter(iris, Species == "virgini
```

```
##
##  F test to compare two variances
##
## data:  dplyr::filter(iris, Species == "setosa")$Sepal.Length and dplyr::filter(iris, Species == "virg
## F = 0.30729, num df = 49, denom df = 49, p-value = 6.366e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1743776 0.5414962
## sample estimates:
## ratio of variances
##          0.3072862
```

The small p-value leads us to reject the null hypothesis that the variances are equal and accept the alternative hypothesis that the variances are not equal.

c. If the equal variance or normality assumptions of ANOVA are violated, an alternative is to use the Kruskal-Wallis test, which is a generalization of the rank-sum test to > 2 groups. You can carry out a Kruskal-Wallis test in R using `kruskal.test()`. Read the help page and apply it here (*hint: you will not need `lm()`!*) – what do you conclude?

```
kruskal.test(Sepal.Length ~ Species, data = iris)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Sepal.Length by Species
## Kruskal-Wallis chi-squared = 96.937, df = 2, p-value < 2.2e-16
```

The p value is extremely small, leading us to reject the null hypothesis that species does not contribute to the **ranks of the sepal lengths.**

d. If (c) produced a significant result, do you think a TukeyHSD test would be applicable here to see which species are significantly different? Why or why not? If not, suggest an alternative. (You don't need to do the test)

Because TukeyHSD is performing t-tests, it has the assumption for normal data and is therefore a parametric test. Any non-parametric t-test with multiple hypothesis correction would be appropriate here. (Grading: don't have to name an actual test, just know that it needs to be non-parametric).

Problem 4:

Next we'll consider the relationship between sepal length and sepal width

a. Refer to the scatterplot matrix you produced in problem 1(d). Do you expect sepal length and sepal width to be positively correlated, negatively correlated, or uncorrelated?

To me, they seem mostly uncorrelated. Maybe a very weak negative correlation - definitely not a positive correlation

b. Compute the correlation between sepal length and sepal width. Are your expectations confirmed?

```
cor(iris$Sepal.Length, iris$Sepal.Width)
```

```
## [1] -0.1175698
```

Very weak negative correlation – almost no correlation

c. Now compute the correlation between sepal length and sepal width separately for each of the iris species. How do these compare to the overall correlation you observed in part (b)?

```r
# again, several ways to do this...
# I am tired of writing out all the subsetting, lets just make 3 dfs
setosa <- dplyr::filter(iris, Species == "setosa")
virginica <- dplyr::filter(iris, Species == "virginica")
versicolor <- dplyr::filter(iris, Species == "versicolor")

cor(setosa$Sepal.Length, setosa$Sepal.Width)
```

```
## [1] 0.7425467
```

```r
cor(virginica$Sepal.Length, virginica$Sepal.Width)
```

```
## [1] 0.4572278
```

```r
cor(versicolor$Sepal.Length, versicolor$Sepal.Width)
```

```
## [1] 0.5259107
```

They appear to all have moderate to strong positive correlations!!!?! So strange. This is an example of Simpson's Paradox.

d. Using R's `lm()` function, fit a linear model that predicts sepal length as a function only of sepal width and assign it to `fit1`.

```r
fit1 <- lm(Sepal.Length ~ Sepal.Width, data = iris)
```

e. Repeat part (d), but now include an additional term to model the difference in average sepal length between species, assigning it to `fit2`.

```r
fit2 <- lm(Sepal.Length ~ Sepal.Width + Species, data = iris)
```

f. As above, but now include the interaction between sepal width and species. Assign this to `fit3`.

```r
fit3 <- lm(Sepal.Length ~ Sepal.Width * Species, data = iris)
```

h. Print out the summary of `fit3` and interpret the results. What are the results telling you?

```r
summary(fit3)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width * Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26067 -0.25861 -0.03305  0.18929  1.44917
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.6390     0.5715   4.618 8.53e-06 ***
## Sepal.Width                      0.6905     0.1657   4.166 5.31e-05 ***
## Speciesversicolor                0.9007     0.7988   1.128    0.261
## Speciesvirginica                 1.2678     0.8162   1.553    0.123
## Sepal.Width:Speciesversicolor    0.1746     0.2599   0.672    0.503
## Sepal.Width:Speciesvirginica     0.2110     0.2558   0.825    0.411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4397 on 144 degrees of freedom
## Multiple R-squared:  0.7274, Adjusted R-squared:  0.718
## F-statistic: 76.87 on 5 and 144 DF,  p-value: < 2.2e-16
```

The model shows β_1's for each covariable (sepal width, species, and interaction). However, the only variable that is significantly different from 0 is Sepal.Width.

i. Look at the summaries for `fit1`, `fit2`, and `fit3`. The "Multiple R-squared" value gives the square of the correlation between the observed dependent variable y_i (in this case, sepal length) and the estimated $\hat{y}_i$ from your model. We can use the squared correlation between them as a measure of how well our model fits. Which gives the best R^2?

```
summary(fit1)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width, data = iris)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5561 -0.6333 -0.1120  0.5579  2.2226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.5262     0.4789   13.63   <2e-16 ***
## Sepal.Width  -0.2234     0.1551   -1.44    0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8251 on 148 degrees of freedom
## Multiple R-squared:  0.01382,    Adjusted R-squared:  0.007159
## F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30711 -0.25713 -0.05325  0.19542  1.41253
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.2514     0.3698   6.089 9.57e-09 ***
## Sepal.Width          0.8036     0.1063   7.557 4.19e-12 ***
## Speciesversicolor    1.4587     0.1121  13.012  < 2e-16 ***
## Speciesvirginica     1.9468     0.1000  19.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.438 on 146 degrees of freedom
## Multiple R-squared:  0.7259, Adjusted R-squared:  0.7203
## F-statistic: 128.9 on 3 and 146 DF,  p-value: < 2.2e-16
```

fit3 (the one with the most covariates) gives the highest R^2 value. Note: either fit2 or fit3 are appropriate answers here, depending on which R^2 you used. I will use a more clear example in the future...

j. Generally, adding more covariates will tend to produce a higher R^2. That doesn't necessarily mean it's a better model, however – you may be overfitting the data! A better way to select a model is to test if the variance explained by the larger model compensates for the degrees of freedom you introduce by adding covariates. Try `anova(fit1,fit2)`. Does `fit2` account for significantly more variance than `fit1`? What about `fit3`? Based on the ANOVA comparisons, which model would you choose?

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: Sepal.Length ~ Sepal.Width
## Model 2: Sepal.Length ~ Sepal.Width + Species
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    148 100.756
## 2    146  28.004  2    72.752 189.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value for the additional terms in fit2 – namely, the indicator of species – is significant; that is, the variance between the species contributes significantly to the overall variance in sepal length.

```
anova(fit1, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: Sepal.Length ~ Sepal.Width
## Model 2: Sepal.Length ~ Sepal.Width * Species
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    148 100.756
## 2    144  27.846  4     72.91 94.258 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This suggests that the species and interaction terms in fit3 are an improvement over fit1 (which has neither), in that those terms collectively account for significant variance in sepal length. However, we still need to check if fit3 does better than fit2, or whether fit2 is enough:

```
anova(fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: Sepal.Length ~ Sepal.Width + Species
## Model 2: Sepal.Length ~ Sepal.Width * Species
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    146 28.004
## 2    144 27.846  2   0.15719 0.4064 0.6668
```

This shows that the additional terms in fit3 vs fit2 – namely, the interaction between species and sepal width – does NOT account for a significant amount of the remaining variance after fit2! We would thus choose fit2 as a minimal predictive model.