# EEMB 146 Final Report: Example with Iris Data

Tatum Katz

5/13/2020

**This is a merely a template or the "bare bones" of what your final report should look like. I reccomend copying and pasting this into a .Rmd in your .Rproj. Once pasted in your .Rmd, I reccomend filling in the sections with your writing and code. Pay attention to the document format and figure labels within {r}.**

## Abstract

Here, I would write ~4 sentences about why this research is relevant, what I did, and my findings.

## Introduction

Here, I would give a general overview of my dataset; the variables and how it was collected. I would also talk about my research question, and why it is interesting and important.

## Exploratory Data Analysis

Here, I would mostly state the findings that the petal variables are nonnormal, and references the figures I made. Check out the R code to see how I got these figures to show up where I wanted them to! (Fig. 1). Then, I would reference Figure 2, which is a nice example of a use of the function par. Hey, did you notice it automatically numbers my figures in the order they appear?

## Statistical Methods

Here, I would explain generally what I did, and then make subsections!

**Does petal length vary by species?** I would CLEARLY state my hypotheses, and the test I ran. I would describe what ANOVA is, what its assumptions are and how I checked them, You could even get mathy here if you wanted to with some *latex*, but it isn't required.

**What variables are good predictors of petal length?** Same as above, but you could get super fancy with some

$$latexonitsownlineusingdoubledollarsigns$$

.

## Results

Now, I would generally explain my results, and go for those subsections again bc i love me some subsections.

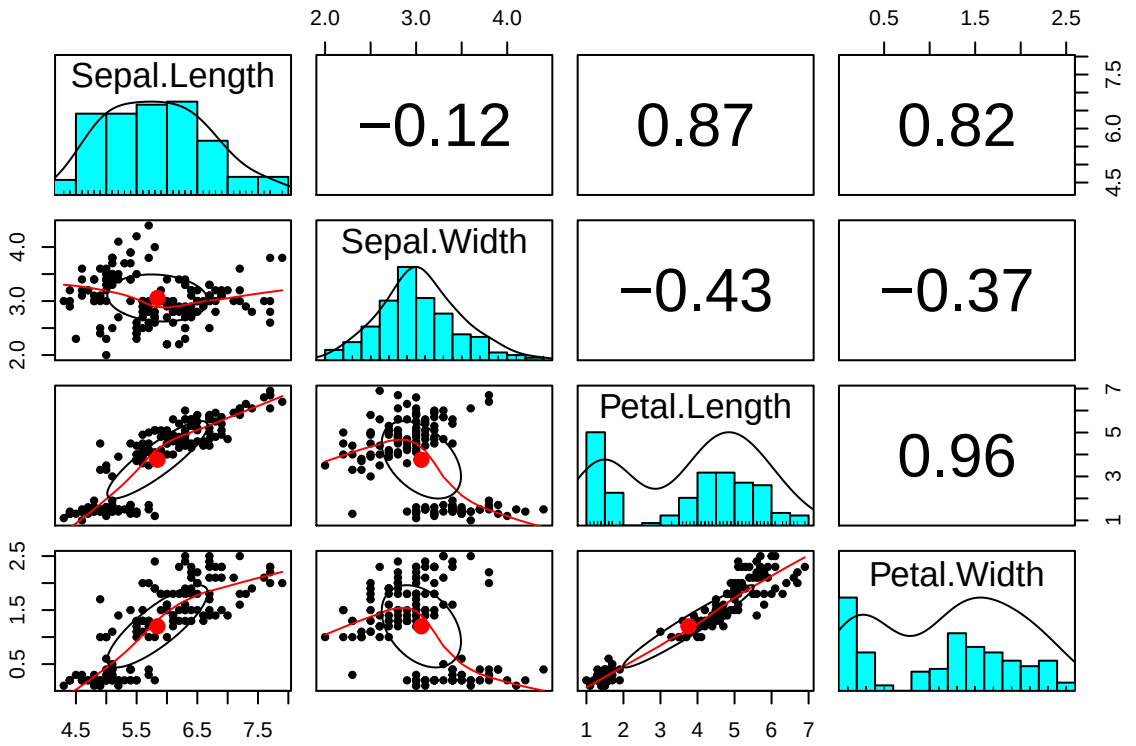



Figure 1: i would write a VERY informative figure caption here that explained EVERY THING about this plot!

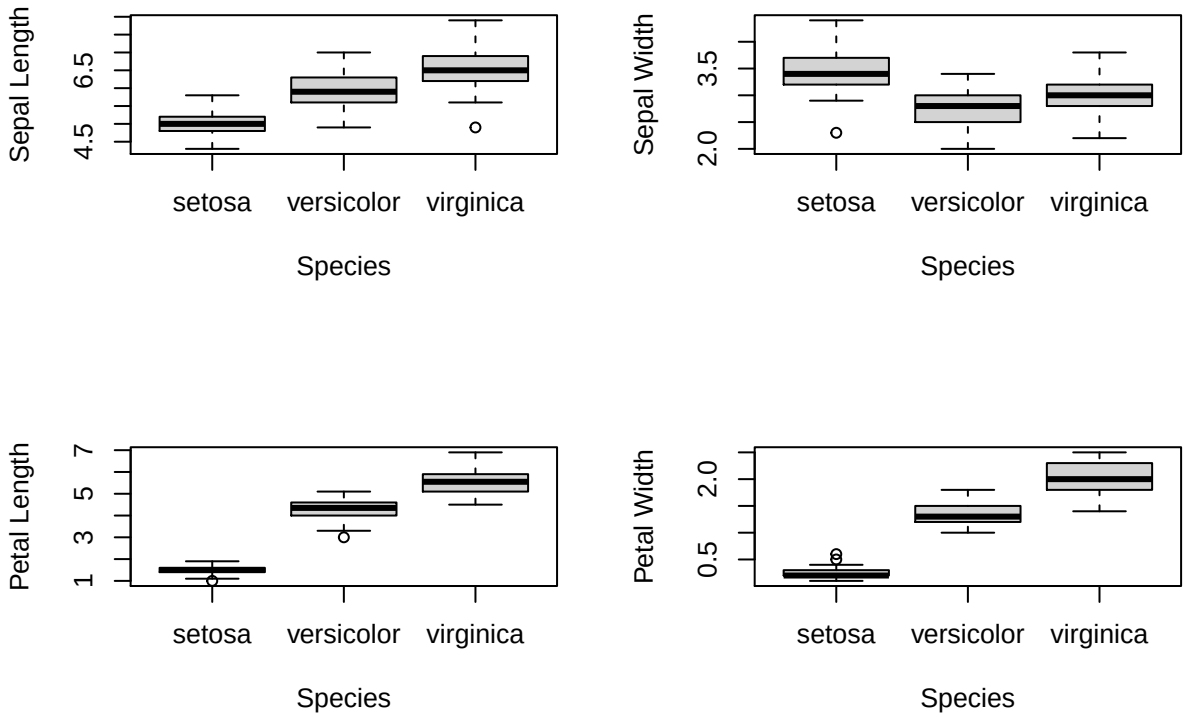



Figure 2: i would write a VERY informative figure caption here that explained EVERY THING about this plot!

**Does petal length vary by species?** I would report all my statistical results like this: I found a significant difference across species' mean petal lengths (one-way ANOVA; F-value = 1180, p-value $< 0.05$). And then reference Figure 3!

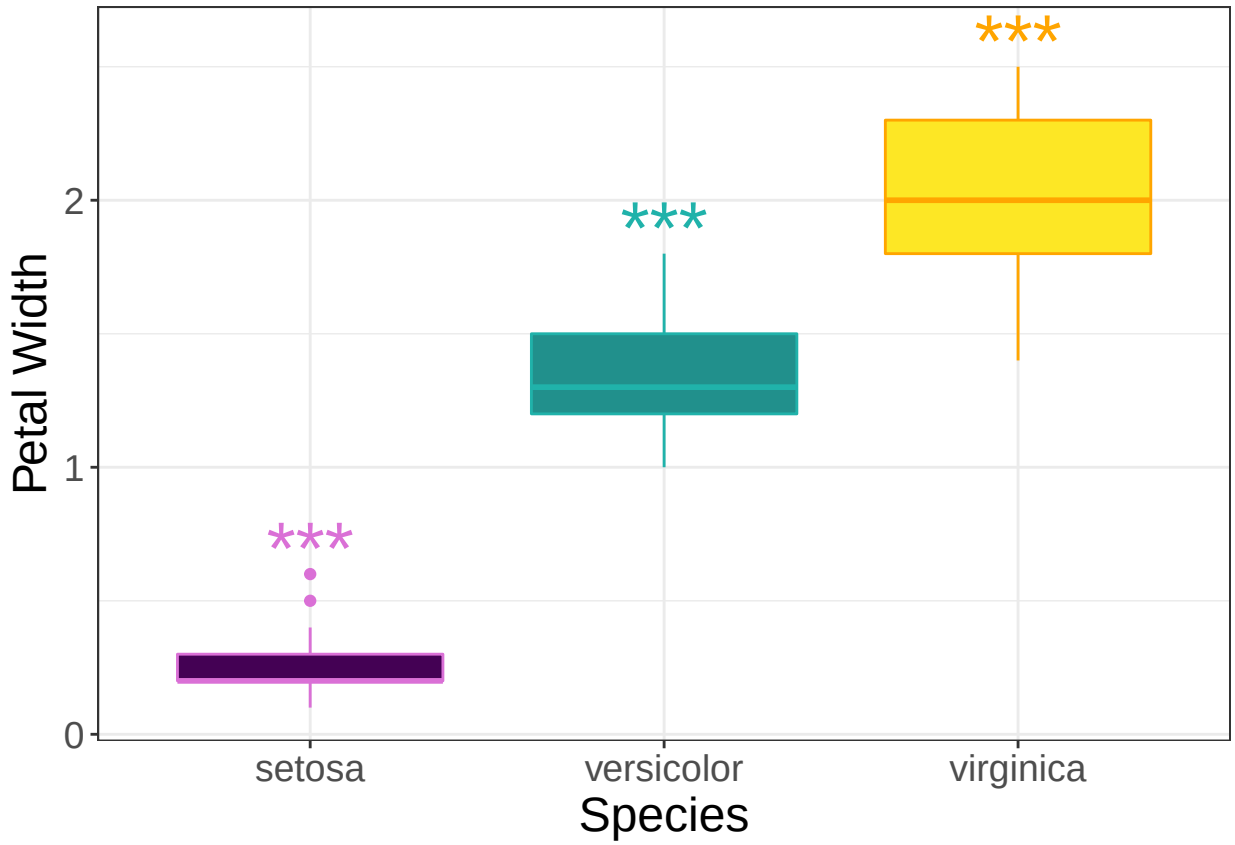



Figure 3: i literally did all my analyses in the appendix and then copied and pasted my favorite plots up here so they show up in the main body!

**What variables are good predictors of petal length?** Rinse, and repeat (Figure 4).

```
## `geom_smooth()` using formula 'y ~ x'
```

## Discussion

Here, I would wrap up everything and explain the biological signficiance. What didn't I do? What were the limitations? What would I do next?

## References

Site each paper and package here. Try using the citation function to get your cites for each package!

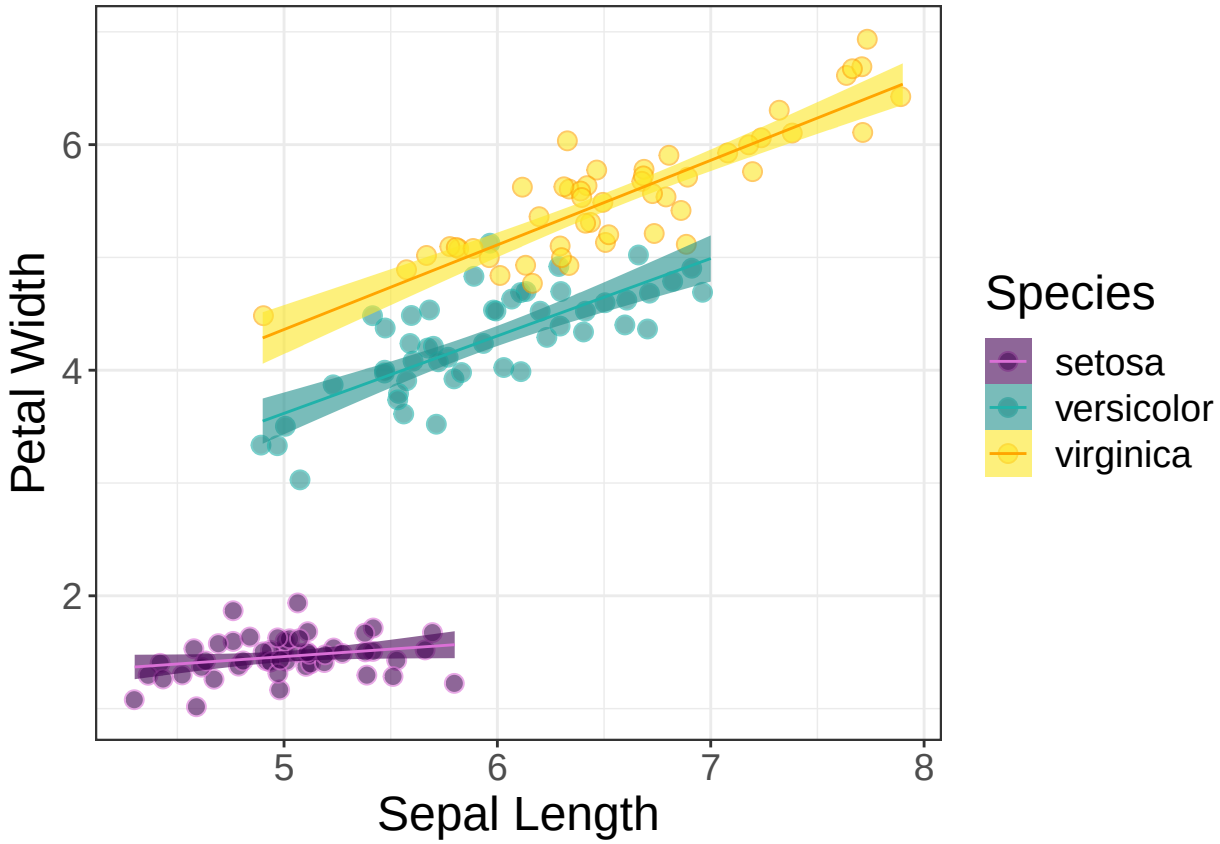



Figure 4: i literally did all my analyses in the appendix and then copied and pasted my favorite plots up here so they show up in the main body!

## Appendix

### Exploratory data analysis

```
#general data checking
head(dat)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

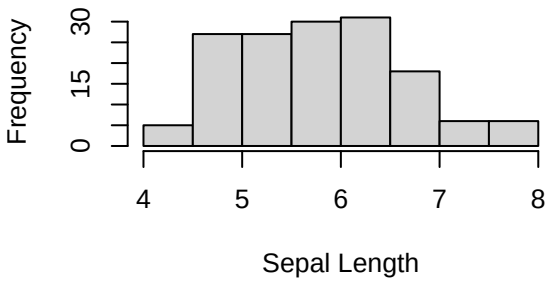
```
levels(dat$Species)
```

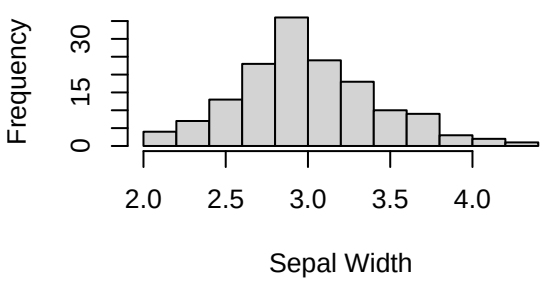
```
## [1] "setosa"     "versicolor" "virginica"
```

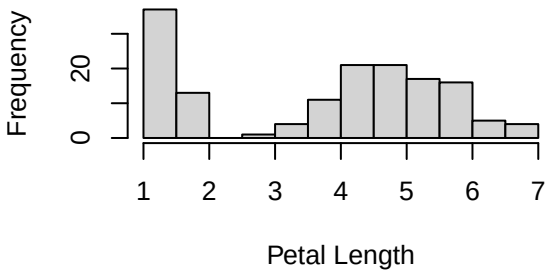
```
summary(dat)
```

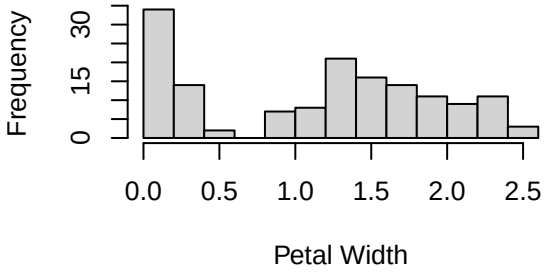
```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
#check numeric vars for distributions and normality
par(mfrow=c(2,2))
hist(dat$Sepal.Length, main="", xlab="Sepal Length")
hist(dat$Sepal.Width, main="", xlab="Sepal Width")
hist(dat$Petal.Length, main="", xlab="Petal Length")
hist(dat$Petal.Width, main="", xlab="Petal Width")
```




```
par(mfrow=c(1,1))

shapiro.test(dat$Sepal.Length) #not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$Sepal.Length
## W = 0.97609, p-value = 0.01018
```

```
shapiro.test(dat$Sepal.Width) #normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$Sepal.Width
## W = 0.98492, p-value = 0.1012
```

```
shapiro.test(dat$Petal.Length) #not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$Petal.Length
## W = 0.87627, p-value = 7.412e-10
```

```
shapiro.test(dat$Petal.Width) #not normal
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$Petal.Width
## W = 0.90183, p-value = 1.68e-08
```

```
par(mfrow=c(2,2))
qqPlot(dat$Sepal.Length) #looks normal
```

```
## [1] 132 118
```

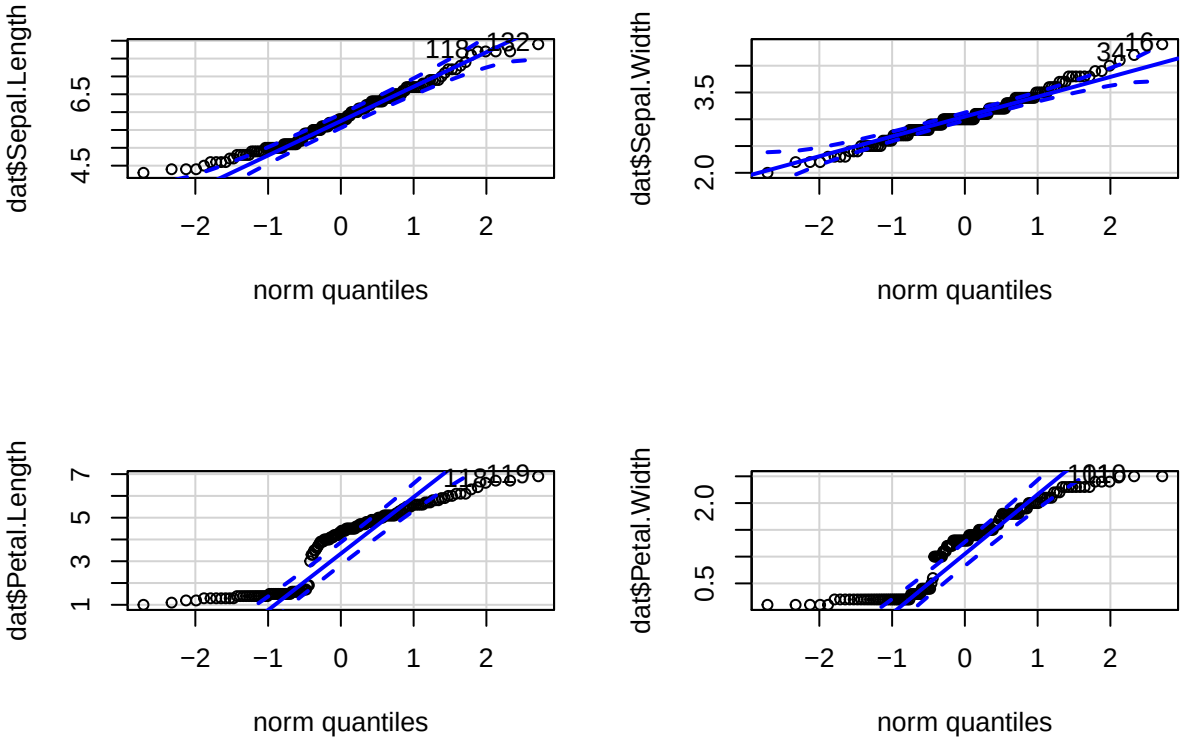
```
qqPlot(dat$Sepal.Width) #looks normal
```

```
## [1] 16 34
```

```
qqPlot(dat$Petal.Length) #looks not normal
```

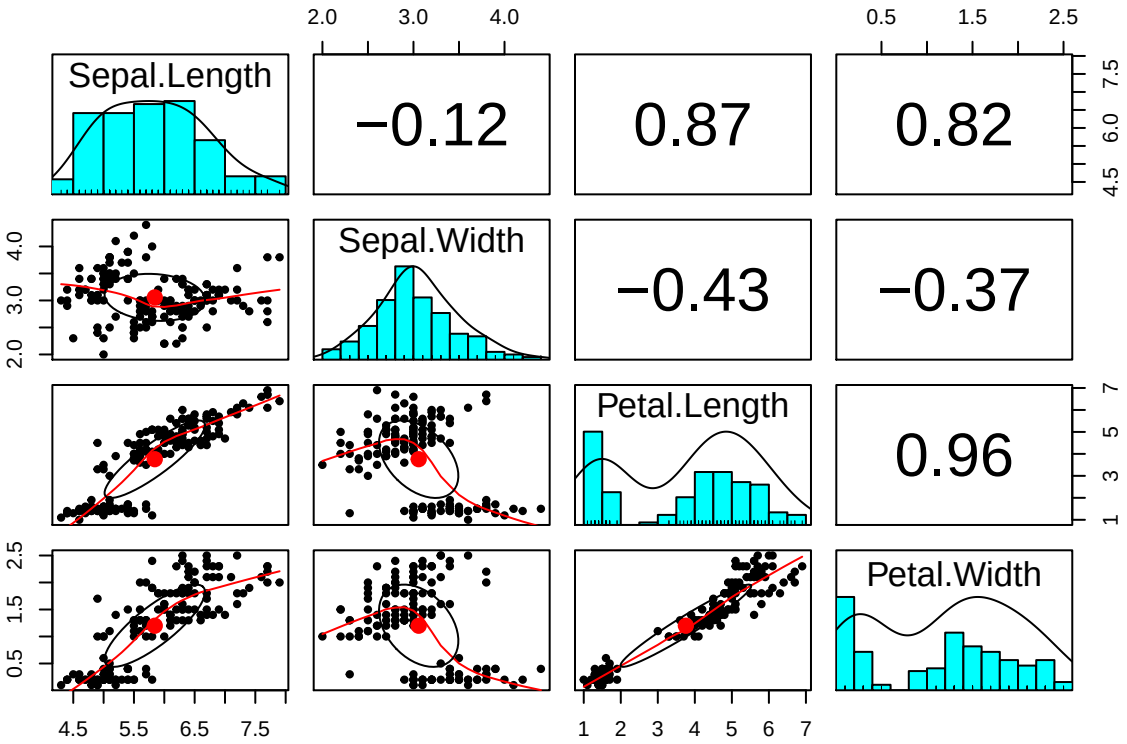
```
## [1] 119 118
```

```
qqPlot(dat$Petal.Width) #looks not normal
```

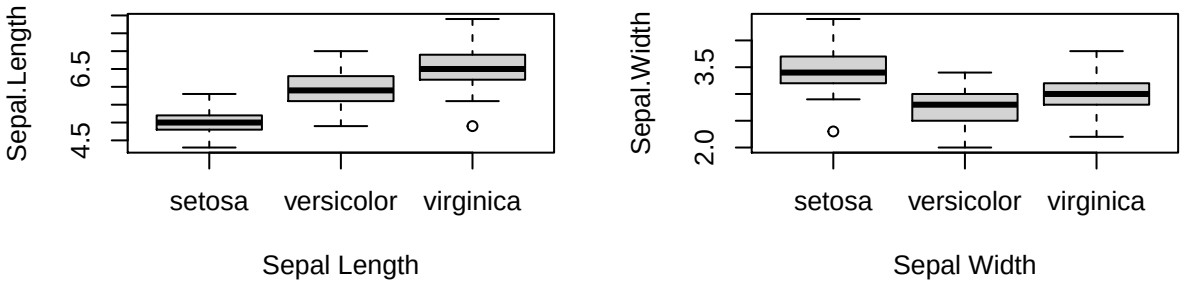


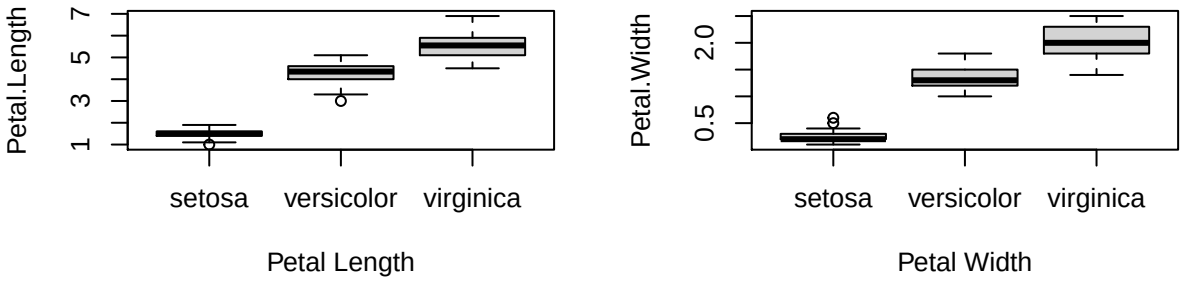

```
## [1] 101 110
```

```
par(mfrow=c(1,1))

#more numeric var investigation
pairs.panels(dat[,-5]) #include as a fig
```




```
#numeric vars vs. species
par(mfrow=c(2,2)) #include as a fig
boxplot(Sepal.Length ~ Species, data=dat, xlab="Sepal Length")
boxplot(Sepal.Width ~ Species, data=dat, xlab="Sepal Width")
boxplot(Petal.Length ~ Species, data=dat, xlab="Petal Length")
boxplot(Petal.Width ~ Species, data=dat, xlab="Petal Width")
```

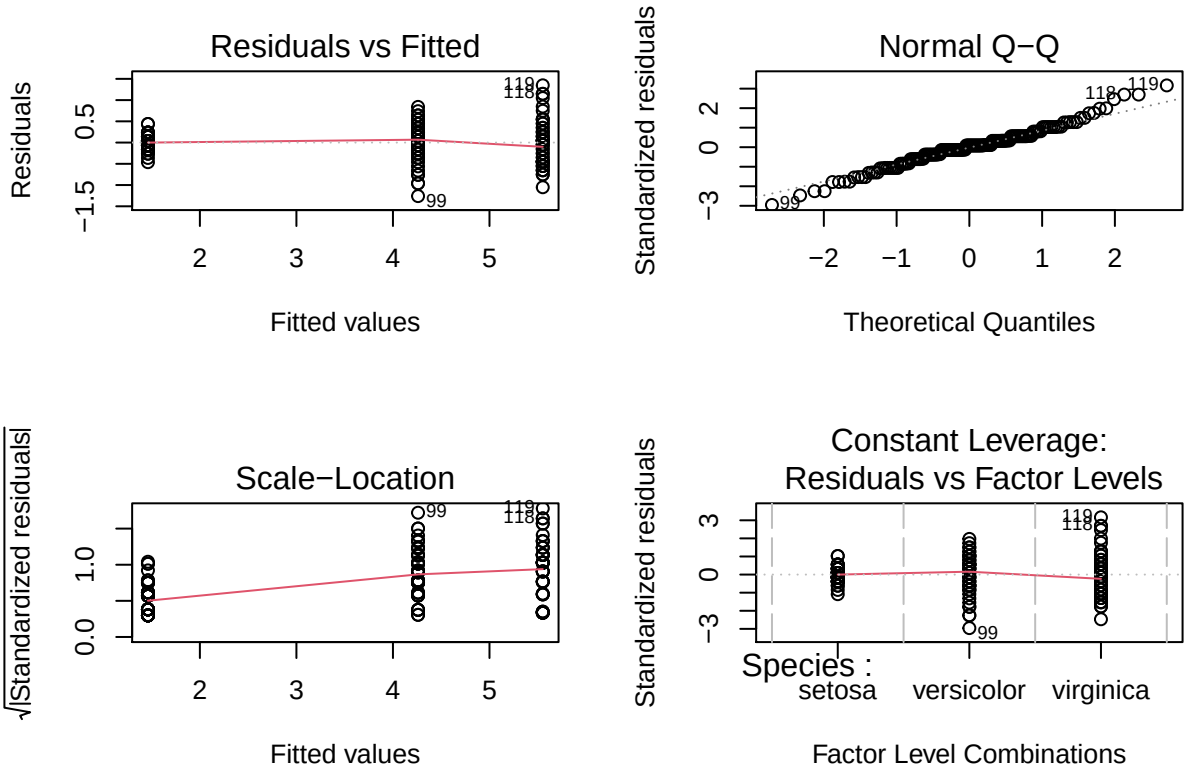


```r
par(mfrow=c(1,1))

#based on all this, here are my findings:

#petal vars not normal, may need to be transformed
#species seems like a good predictor of each numeric var, perhaps an ANOVA, or multivariate linear mode
#sepal length and petal length look correlated
#sepal length and petal width look correlated
#petal length and petal width look correlated

#so, i think i will run an anova on petal length ~ species for my "comparing means" test
#and run a linear model to see what variables are most predictive of petal length for my "predictive" t
```

### Statistical methods

```r

#goal: one way anova of petal length by species

#step 1: check assumptions of anova
#indep data? probs
#normality
#equal variance

fit1 <- lm(Petal.Length ~ Species, data=dat)
par(mfrow=c(2,2))
plot(fit1) #oof looks a little cone shaped
```

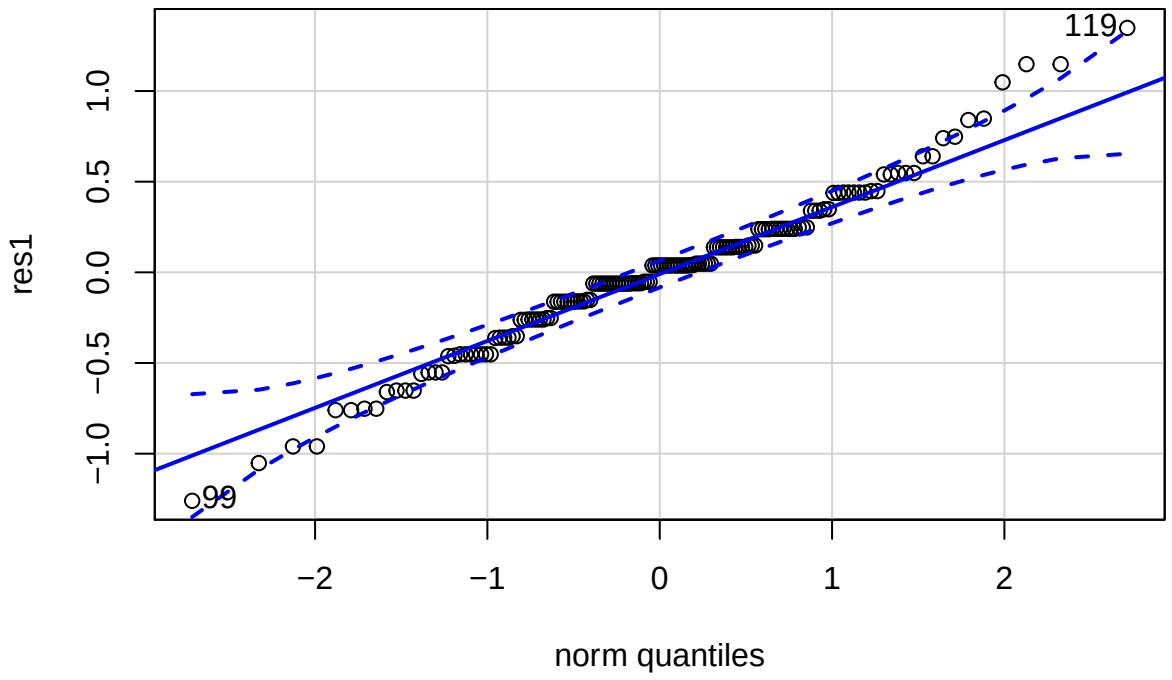



### ANOVA

```r
par(mfrow=c(1,1))
res1 <- fit1$residuals
shapiro.test(res1) #just barely non-normal
```

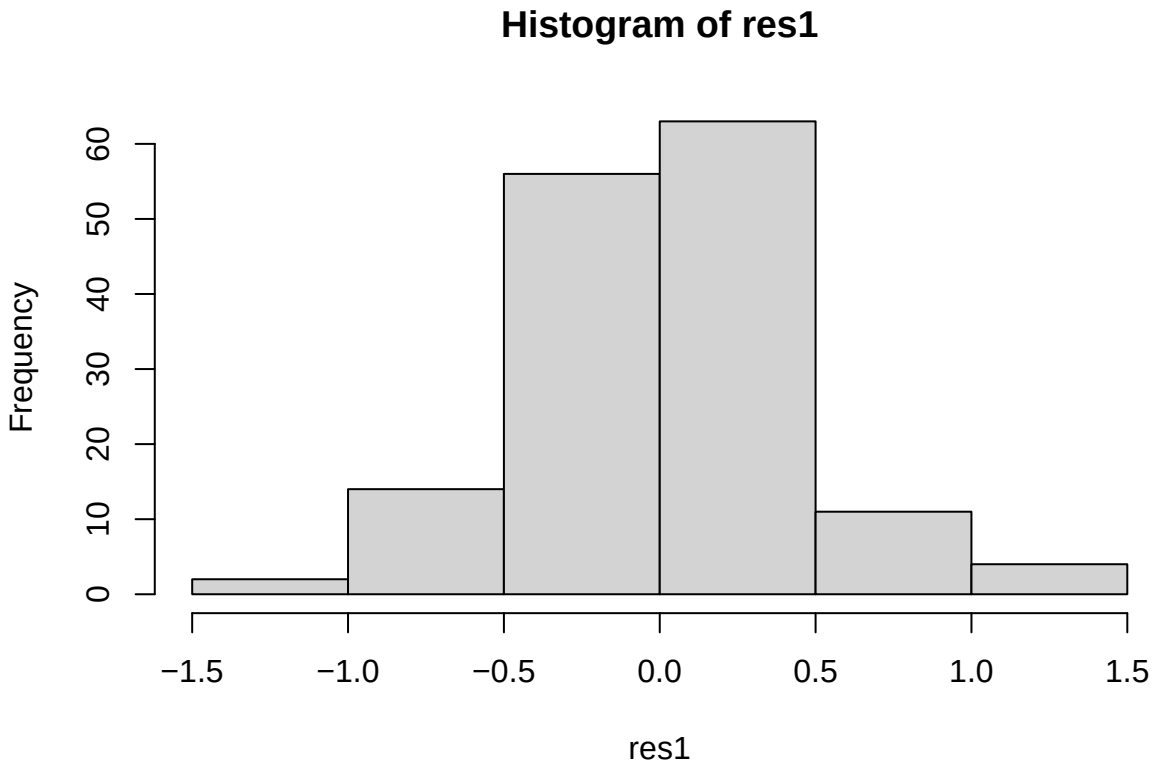
```
##
##  Shapiro-Wilk normality test
##
## data:  res1
## W = 0.98108, p-value = 0.03676
```

```r
qqPlot(res1) #actually looks normal
```




```
## [1] 119  99
```

```r
hist(res1) #hmm, not obviously skewed
```




```
#so the variances aren't great but the data are normal enough so maybe i can ignore that violation!

#step 2: run the anova

irisaov <- aov(Petal.Length ~ Species, data=dat)
summary(irisaov) #species highly significant, as expected from EDA. only added digits arg bc otherwise
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

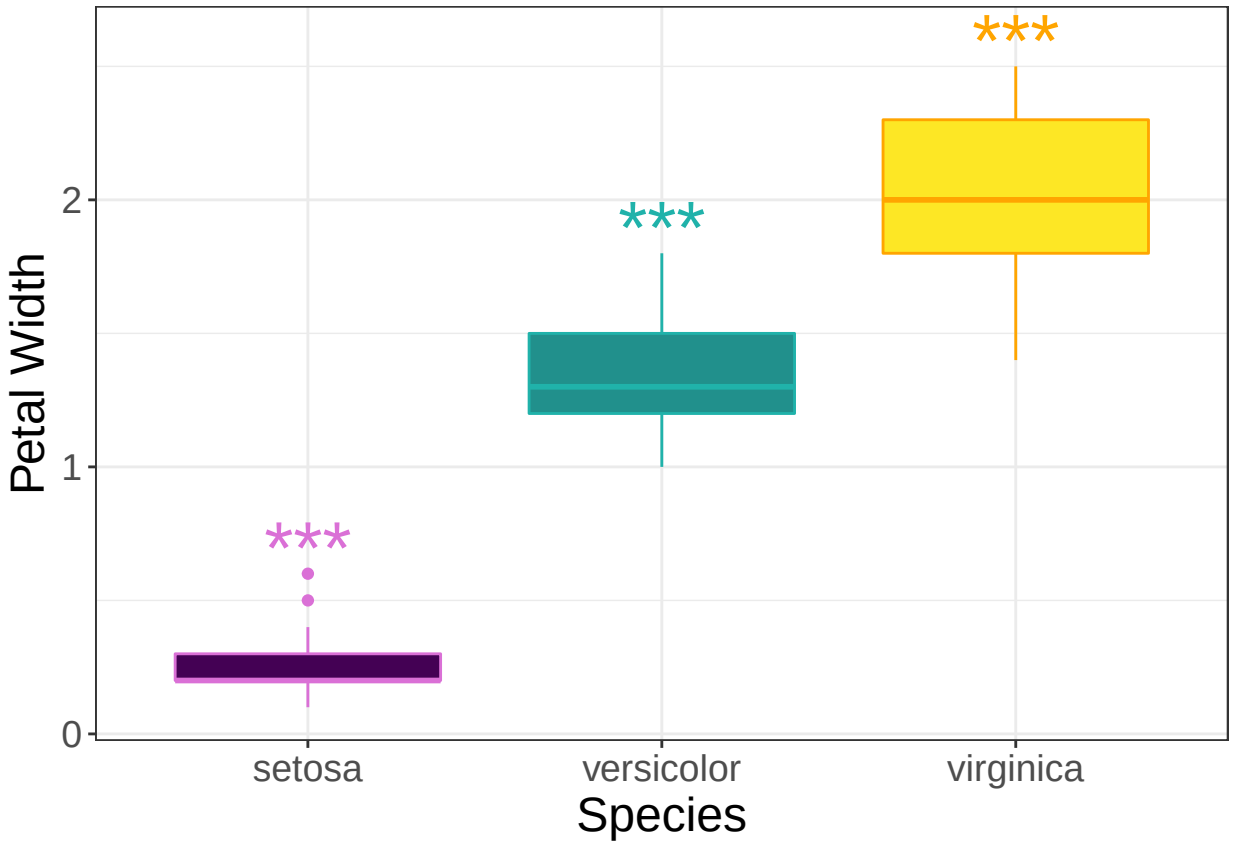
```
#step 3: post-hoc analyses

TukeyHSD(irisaov) #whoa p-values should never be 0. this is a rounding error. lets extract with more pr
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Petal.Length ~ Species, data = dat)
##
## $Species
##                       diff     lwr     upr p adj
## versicolor-setosa    2.798 2.59422 3.00178     0
## virginica-setosa     4.090 3.88622 4.29378     0
## virginica-versicolor 1.292 1.08822 1.49578     0
```

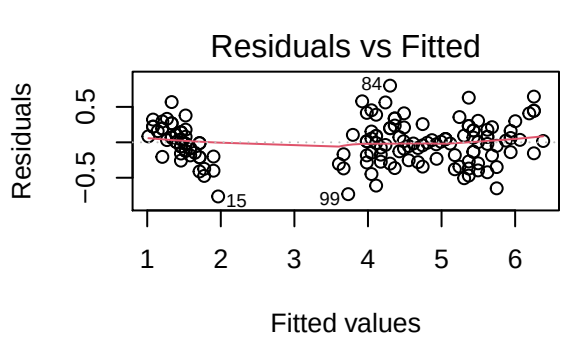
```r
thsd <- TukeyHSD(irisaov)
print(thsd, digits=15) #wow so super significant, they are all sig different from each other
```

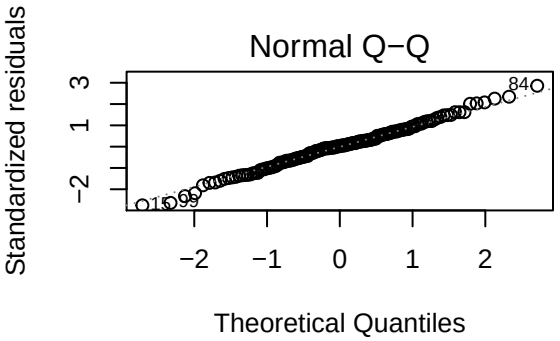
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Petal.Length ~ Species, data = dat)
##
## $Species
##                        diff             lwr              upr p adj
## versicolor-setosa    2.798 2.5942199870557 3.00178001294430 3e-15
## virginica-setosa     4.090 3.8862199870557 4.29378001294431 3e-15
## virginica-versicolor 1.292 1.0882199870557 1.49578001294431 3e-15
```

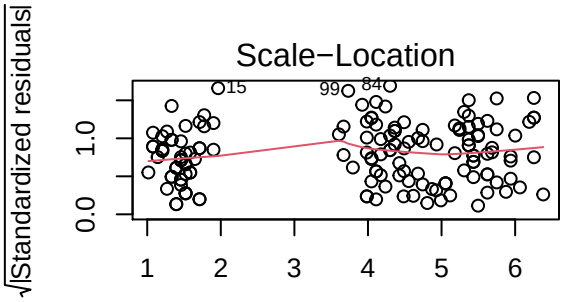
```r
#step 4: make a nice ggplot figure to summarize these results, include as a fig

label.df <- data.frame(Species = levels(dat$Species), #make significance stars
                       Petal.Width = c(.7, 1.9, 2.6)) #these are the y-axis locs of the stars

ggplot(dat, aes(x=Species, y=Petal.Width, fill=Species, color=Species))+ #base plot
  geom_boxplot(show.legend=F)+ #do a boxplot, but hide legend cause its just colors
  scale_fill_viridis_d()+ #set the fill colors
  scale_color_manual(values=c("orchid", "lightseagreen", "orange"))+ #set line colors
  ylab("Petal Width")+
  geom_text(data = label.df, label = "***", size=10, show.legend=F)+ #add those ***
  theme_bw()+ #cute theme
  theme(text = element_text(size=18)) #cause my eyes is tired and old
```

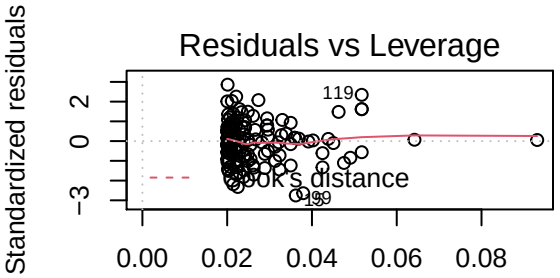



```
#goal: predict petal length by all, see what is important

#step 1: check assumptions of linear regression
#indep? i hope so
#normal residuals?
#equal variances?

fit2 <- lm(Petal.Length ~ Sepal.Length + Species, data=dat)
par(mfrow=c(2,2))
plot(fit2) #looks homoskedastic to me
```



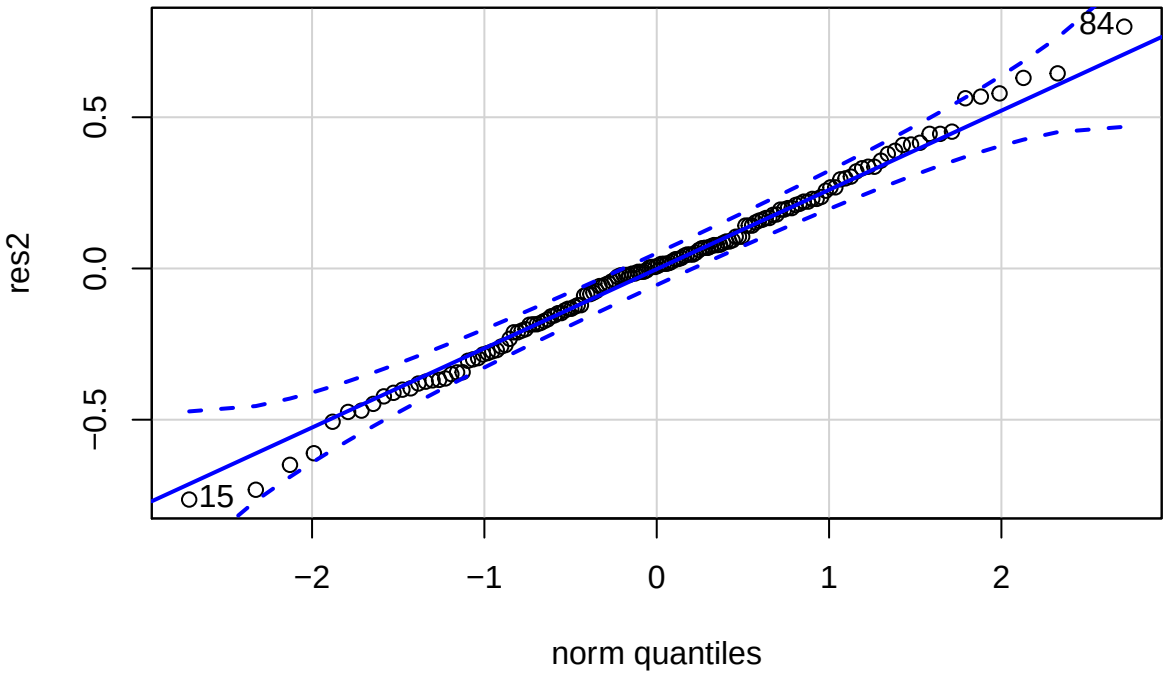




**Linear regression**

```
par(mfrow=c(1,1))
res2 <- fit2$residuals
qqPlot(res2) #she is normal
```




```
## [1] 84 15
```

```r
shapiro.test(res2) #yep its normo
```

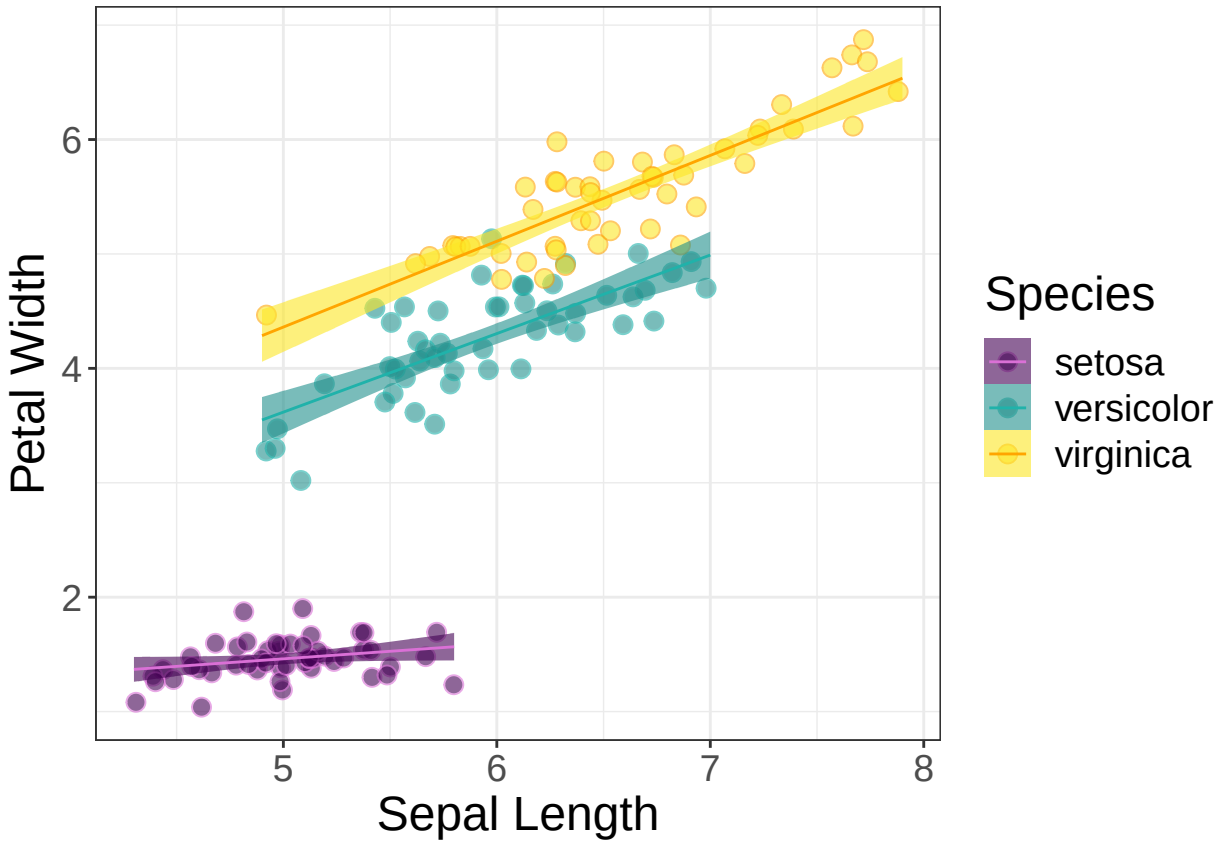
```
##
##  Shapiro-Wilk normality test
##
## data:  res2
## W = 0.99473, p-value = 0.8658
```

```r
#step 2: run the lm

summary(fit2) #EVERYTHING! is a significant predictor, which isn't shocking based on EDA
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length + Species, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76390 -0.17875  0.00716  0.17461  0.79954
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.70234    0.23013  -7.397 1.01e-11 ***
## Sepal.Length       0.63211    0.04527  13.962  < 2e-16 ***
## Speciesversicolor  2.21014    0.07047  31.362  < 2e-16 ***
## Speciesvirginica   3.09000    0.09123  33.870  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2826 on 146 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9744
## F-statistic:  1890 on 3 and 146 DF,  p-value: < 2.2e-16
```

```r
#step 3: make a pretty graph to summarize findings

ggplot(dat, aes(x=Sepal.Length, y=Petal.Length, color=Species, fill=Species))+
  geom_jitter(size=3, alpha=0.6, shape=21)+
  geom_smooth(method="lm", alpha=0.6, size=0.5)+
  scale_color_manual(values=c("orchid", "lightseagreen", "orange"))+
  scale_fill_viridis_d()+ #set the fill colors
  ylab("Petal Width")+
  xlab("Sepal Length")+
  theme_bw()+ #cute theme
  theme(text = element_text(size=18)) #cause my eyes is tired and old
```

```
## `geom_smooth()` using formula 'y ~ x'
```


Petal Width
6
4
2
5
6
7
8
Sepal Length
Species
setosa
versicolor
virginica