Question 4.2

The *iris* data set `iris.txt` contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Iris). *The response values are only given to see how well a specific method performed and should not be used to build the model.*

Use the R function `kmeans` to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

Below is the code for questions 4.2.

I. In general the approach was:

1. Load the data
2. Using ggplot verify how the data looks for each variable to make a visual comparison of how the attributes compare with each other.
3. Use all the data to create the bet k-means model to determine the best combination of predictor which is:

The predictor table is

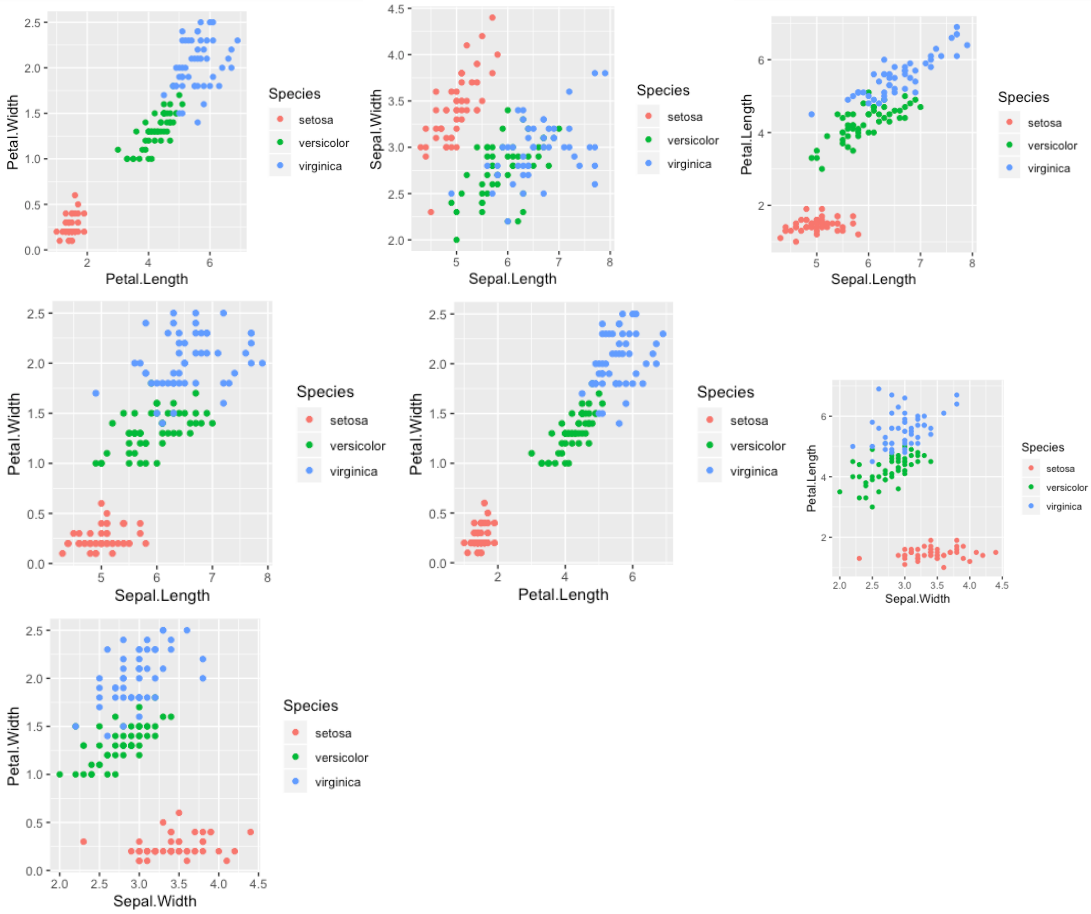
	1	2	3
setosa	**0**	**0**	**50**
versicolor	**48**	**2**	**0**
virginica	**14**	**36**	**0**

4. Plot the cluster to visualize
5. Plot the elbow diagram to determine the K value which is 3
6. Determine how well it predicted which is 89%

```
#Import the data
data <- read.table("~/Documents/Georgia Tech Masters/EdX/FA20Introduction to Analytics Modelin/HW2/iris.txt", header = TRUE, row.names = 1)

#set random Generator seed
set.seed(123)
head(data, 3)

ggplot(data, aes(Petal.Length, Petal.Width, color= Species)) + geom_point()
ggplot(data, aes(Sepal.Length, Sepal.Width, color= Species)) + geom_point()
ggplot(data, aes(Sepal.Length, Petal.Length, color= Species)) + geom_point()
ggplot(data, aes(Sepal.Length, Petal.Width, color= Species)) + geom_point()
ggplot(data, aes(Sepal.Width, Petal.Width, color= Species)) + geom_point()
ggplot(data, aes(Sepal.Width, Petal.Length, color= Species)) + geom_point()
```



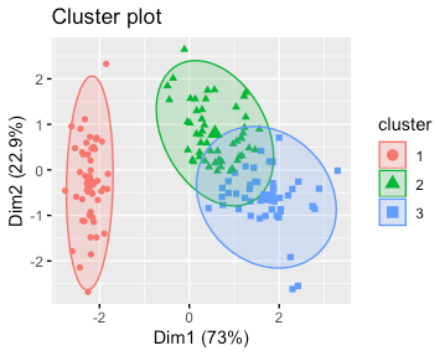
read the best model with k means with all the data

```
best_model <-kmeans(data[,1:4],3,nstart=5,iter.max=15)
table(data[,5], best_model$cluster)
```

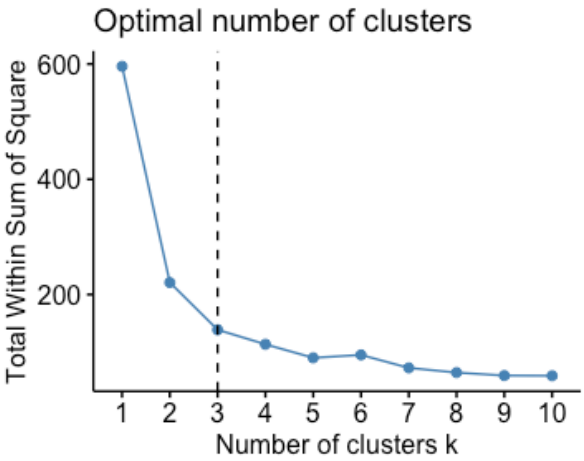
The predictor table is

	1	2	3
setosa	**0**	**0**	**50**
versicolor	**48**	**2**	**0**
virginica	**14**	**36**	**0**

```
data_filter <- scale(data[,1:4])
fviz_cluster(best_model, data = data_filter, geom = "point",
        stand = FALSE, ellipse.type = "norm")
```



fviz_nbclust(data_filter, kmeans, method = "wss") + geom_vline(xintercept = 3, linetype = 2)



predicted_species<- best_model$cluster
accuracy<- sum(predicted_species == data[,5]) / nrow(data)
accuracy
89%