HW11

Grace Okamoto

11/17/2021

Problem 1 is associated with your personal dataset project. Submit your work for all problems in a single .pdf compiled with knitr. The Homework Data folder has a starting .Rmd template with spaces for you to fill in your answers to the questions.

Problem 1

This question is related to your personal data set project. The goal for the end of this week is to put together an initial visualization of your data and to perform one piece of analysis on your data.

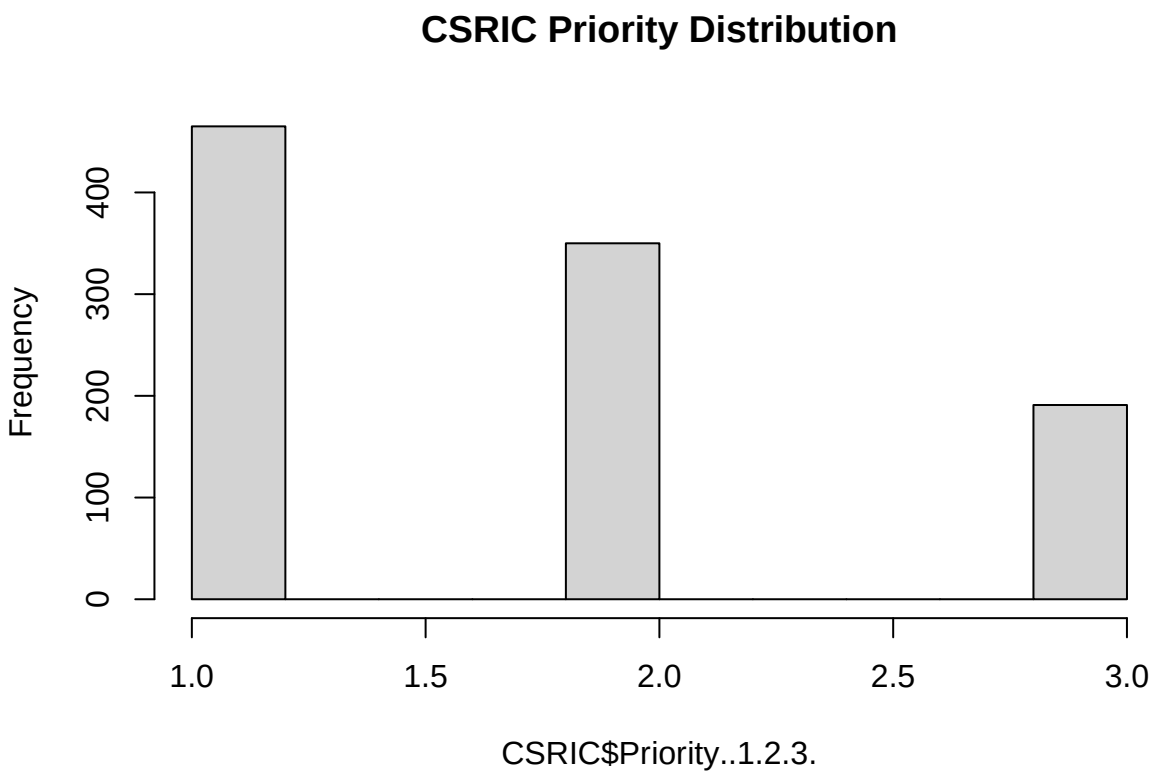
(A) Use your dataset (possibly extended from what you originally submitted week 9) to create a visualization that summarizes an important property of the data. You should follow the principles of good visualization design that we discussed in week 4 and make sure to include a title, axis labels, and any other necessary components.

```
CSRIC <- read.csv("CSRIC_Best_Practices.csv")
head(CSRIC)
```

```
##    BP.Number         Priority
## 1 12-10-0436 Highly Important
## 2 12-10-0437 Highly Important
## 3 12-10-0440 Highly Important
## 4 12-10-0447        Important
## 5 12-10-0448 Highly Important
## 6 12-10-0449         Critical
##
## 1                                                    Network Operators, Service Providers, and Public
## 2        Network Operators and Service Providers should aggregate routes where appropriate (e.g., sim
## 3                                    Network Operators and Service Providers should set and periodi
## 4                                          Network Operators and Service Providers should consider
## 5                                                    Equipment Suppliers should  where feasible, p
## 6 Network Operators, Service Providers and Public Safety should  where feasible, deploy fraudulent ti
##                                      Network.Type.s.
## 1 Cable; Internet/Data; Satellite; Wireless; Wireline;
## 2                                       Internet/Data;
## 3                                       Internet/Data;
## 4 Cable; Internet/Data; Satellite; Wireless; Wireline;
## 5 Cable; Internet/Data; Satellite; Wireless; Wireline;
## 6 Cable; Internet/Data; Satellite; Wireless; Wireline;
##                                    Industry.Role.s.
## 1 Service Provider; Network Operator; Public Safety;
## 2               Service Provider; Network Operator;
## 3               Service Provider; Network Operator;
## 4               Service Provider; Network Operator;
```

```
## 5                                Equipment Supplier;
## 6 Service Provider; Network Operator; Public Safety;
##                                                    Keywords
## 1                           Network Operations; Procedures;
## 2   Cyber Security; Network Operations; Network Provisioning;
## 3                  Industry Cooperation; Network Operations;
## 4                               Liaison; Network Operations;
## 5 Hardware; Network Elements; Network Provisioning; Software;
## 6                        Cyber Security; Network Operations;
##   Public.Safety.and.Disaster
## 1                       TRUE
## 2                      FALSE
## 3                      FALSE
## 4                      FALSE
## 5                      FALSE
## 6                       TRUE
##                                                Reference cable internet.Data
## 1                                                   TRUE          TRUE
## 2                                                  FALSE          TRUE
## 3                                                  FALSE          TRUE
## 4 Note: This Best practice could impact 9-1-1 operations.  TRUE          TRUE
## 5                                                   TRUE          TRUE
## 6 Note: This Best practice could impact 9-1-1 operations.  TRUE          TRUE
##   satellite wireless wireline Service.Provider Network.Operator
## 1      TRUE     TRUE     TRUE             TRUE             TRUE
## 2     FALSE    FALSE    FALSE             TRUE             TRUE
## 3     FALSE    FALSE    FALSE             TRUE             TRUE
## 4      TRUE     TRUE     TRUE             TRUE             TRUE
## 5      TRUE     TRUE     TRUE            FALSE            FALSE
## 6      TRUE     TRUE     TRUE             TRUE             TRUE
##   Priority..1.2.3. Equipment.Supplier Property.Manager Government Public.Safety
## 1                2              FALSE            FALSE      FALSE          TRUE
## 2                2              FALSE            FALSE      FALSE         FALSE
## 3                2              FALSE            FALSE      FALSE         FALSE
## 4                1              FALSE            FALSE      FALSE         FALSE
## 5                2               TRUE            FALSE      FALSE         FALSE
## 6                3              FALSE            FALSE      FALSE          TRUE
```

```
hist(CSRIC$Priority..1.2.3., main = "CSRIC Priority Distribution")
```



(B) Upload your figure to your github repository for the project and incorporate it into the readme file, along with a description (caption) of the plot that describes its contents.

(C) Choose one of the analytical techniques that we have discussed in class so far (either from the EDA section or the (un)supervised learning techniques) and apply it to your data. Explain in a brief paragraph why you chose this technique and what you learned about your data by performing this analysis (you don't have to upload this to github). This can be anything from a simple EDA technique like identifying outliers, making a box plot of a column, evaluating the summary statistics (Tukey's 5 numbers), to one of the regression models from supervised learning, or one of the clustering methods from unsupervised learning. This is not an exhaustive list of possibilities and as with the previous components doesn't need to be long or drawn out.

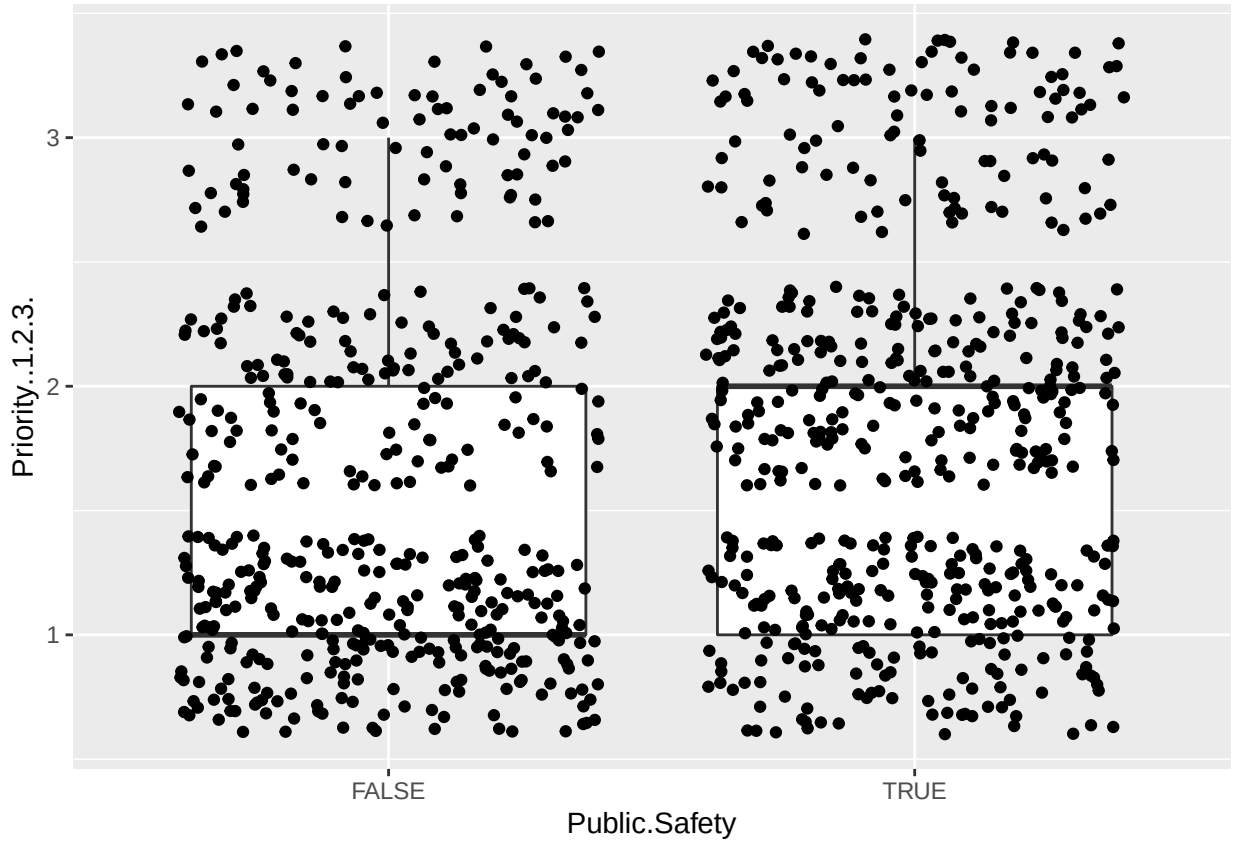
```
library(ggplot2)
summary(CSRIC$Priority..1.2.3.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   1.000   2.000   1.728   2.000   3.000      86
```

```
ggplot(data = CSRIC, mapping = aes(x=Public.Safety, y=Priority..1.2.3.)) + geom_boxplot() + geom_jitter
```

```
## Warning: Removed 86 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 86 rows containing missing values (geom_point).
```



Through this strange and convoluted bar chart that makes me despise everything I know in life, we can see that there is a slight tendency for events with "False" as their Public Safety value to be 1, compared with the "True" box plot, which has comparatively fewer events with a priority value of 1, and more events with the 2 and 3 priority values. ## Problem 2

In your own words, please write brief answers to the following:

(A) What is the curse of dimensionality?

The curse of dimensionality refers to the problems that can arise in higher-dimensional spaces when adding additional variables.

(B) What is the difference between MDS and PCA?

Multi-Dimensional Scaling is a way of converting distances to lower dimensions, whereas Principal Component Analysis converts distance to lower dimensions with the largest amount of variance.

(C) Give an example of a dataset for which dimension reduction would be a useful first step.

This could be useful in a graph of cities and their distances from each other in order to visualize the distances.

(D) What is a dendrogram and what type of clustering method is it used to represent?

A dendrogram is a representation of hierarchial clustering, where the branches of the tree conglomerate together to represent the clusters.

(E) What is the difference between supervised and unsupervised learning?

Unsupervised learning is where the data is not modeled with a specific reactionary variable in mind. It is more meant to explore the nature of the data, instead of seeking a specific answer to a particular question. However, knowing when success has been achieved is much more difficult, as success in this scenario is quite complicated and is difficult to show with numerical metrics. On the other hand, supervised learning generally has specific questions which the analyst seeks to answer, and success is much more relative, and can be shown numerically.

Problem 3

This problem checks your understanding of PCA and K means clustering.

(A) Load the iris data as a dataframe. This is a classic dataset (originally published in 1936) that is frequently used as an initial set of test data. It consists of four numerical columns reporting the length and width of the sepal and petals of 150 iris plants and a final column reporting the specific subspecies. The version in the week 11 data folder also has an additional column with numerical column representing the subspecies.

(A) Make a scatterplot of petal width vs. petal length colored by the subspecies.

(A) Use R to perform PCA on the data with the four numerical columns as inputs and make a scatter plot of the top two principal components colored by subspecies.

(A) What proportion of the variance is explained by these two components?

$$0.7659 + 0.1843 = 0.9496$$

(or 95% of the variance is explained by these two components)

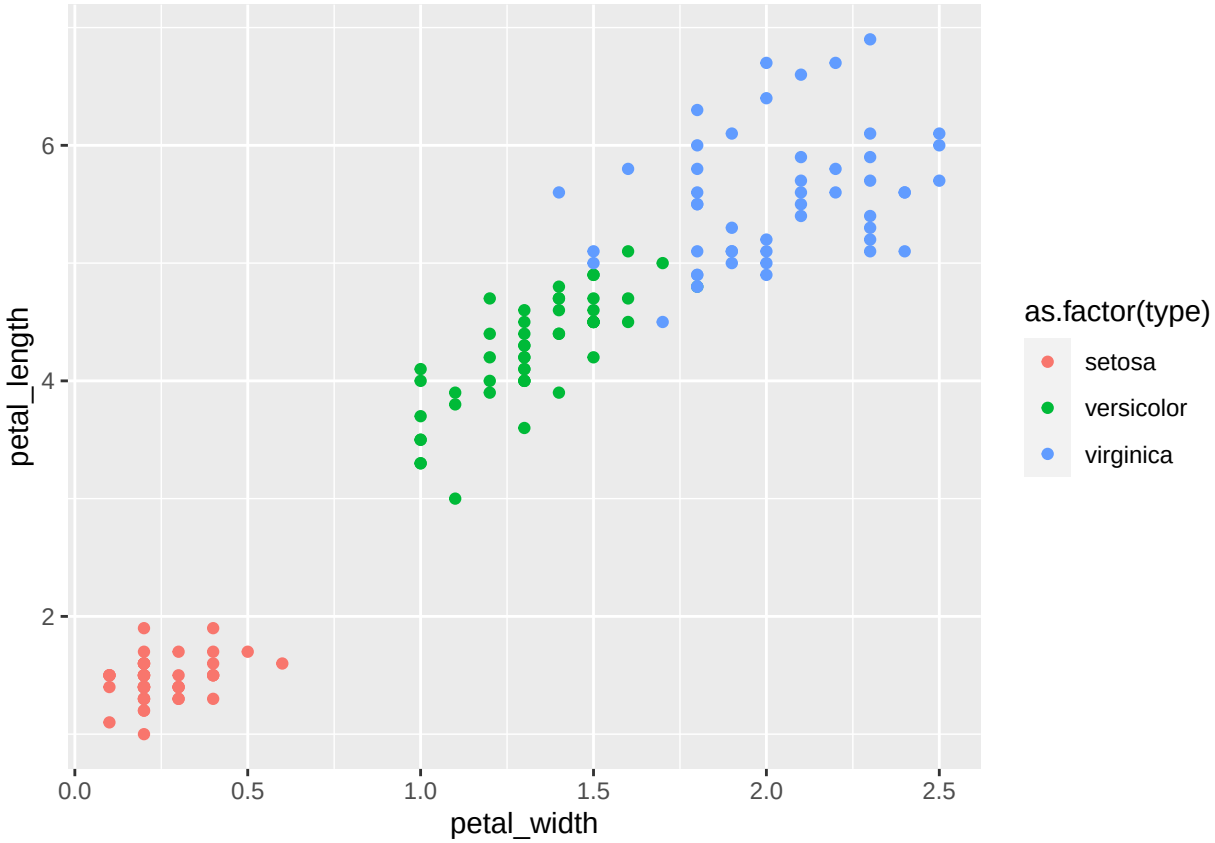
(A) What are the loadings for each of the original numerical columns?

(A) Apply K means clustering to the four numeric columns with three clusters.

(A) Apply K means clustering to the two principal components with three clusters.

(A) Which of the two K means clusterings is more accurate at predicting the subspecies correctly?

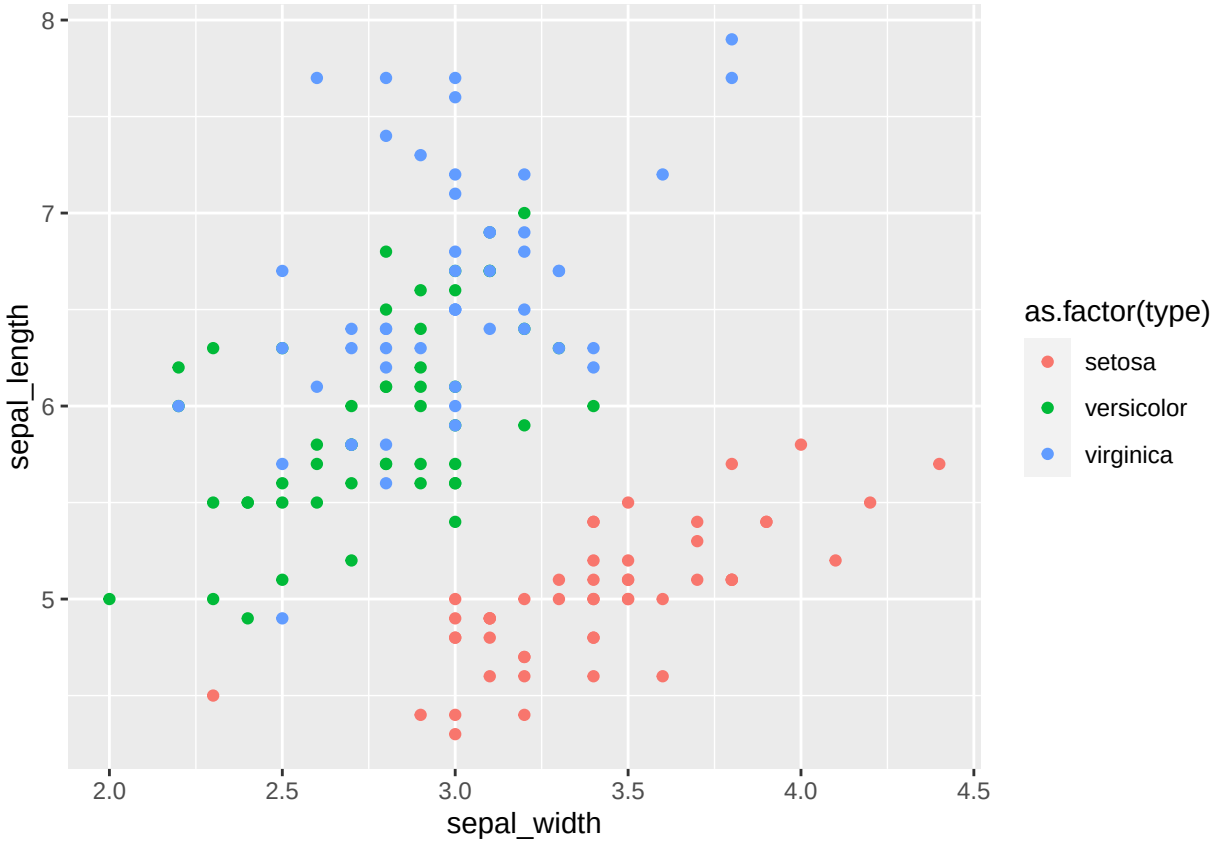
```
iris <- read.csv("iris_data.csv")
ggplot(iris,aes(x=petal_width,y=petal_length,color=as.factor(type))) + geom_point()
```




```
irisPCA <- prcomp(iris[,c(1:4,6)],center=TRUE,scale=TRUE)
summary(irisPCA)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.9569 0.9599 0.4319 0.20521 0.14297
## Proportion of Variance 0.7659 0.1843 0.0373 0.00842 0.00409
## Cumulative Proportion  0.7659 0.9502 0.9875 0.99591 1.00000
```

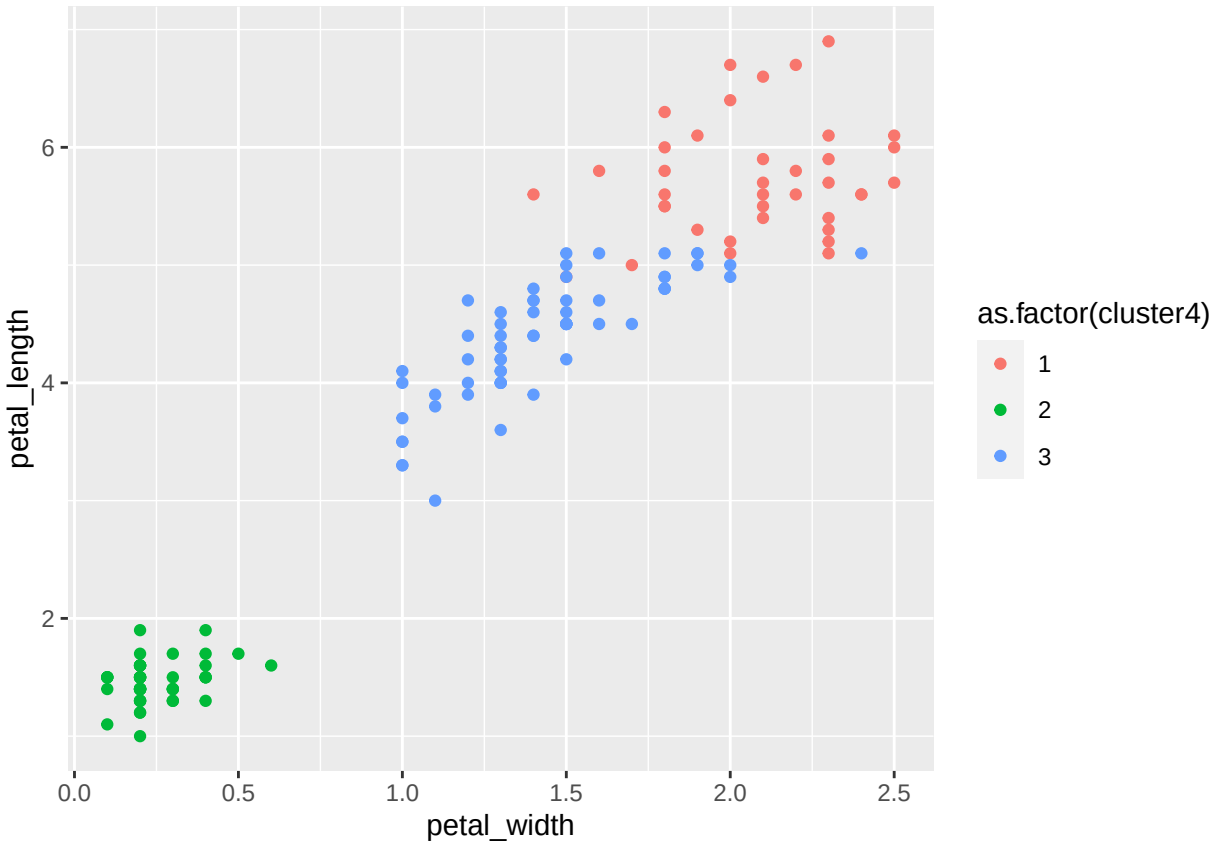
```
ggplot(iris,aes(x=sepal_width,y=sepal_length,color=as.factor(type))) + geom_point()
```



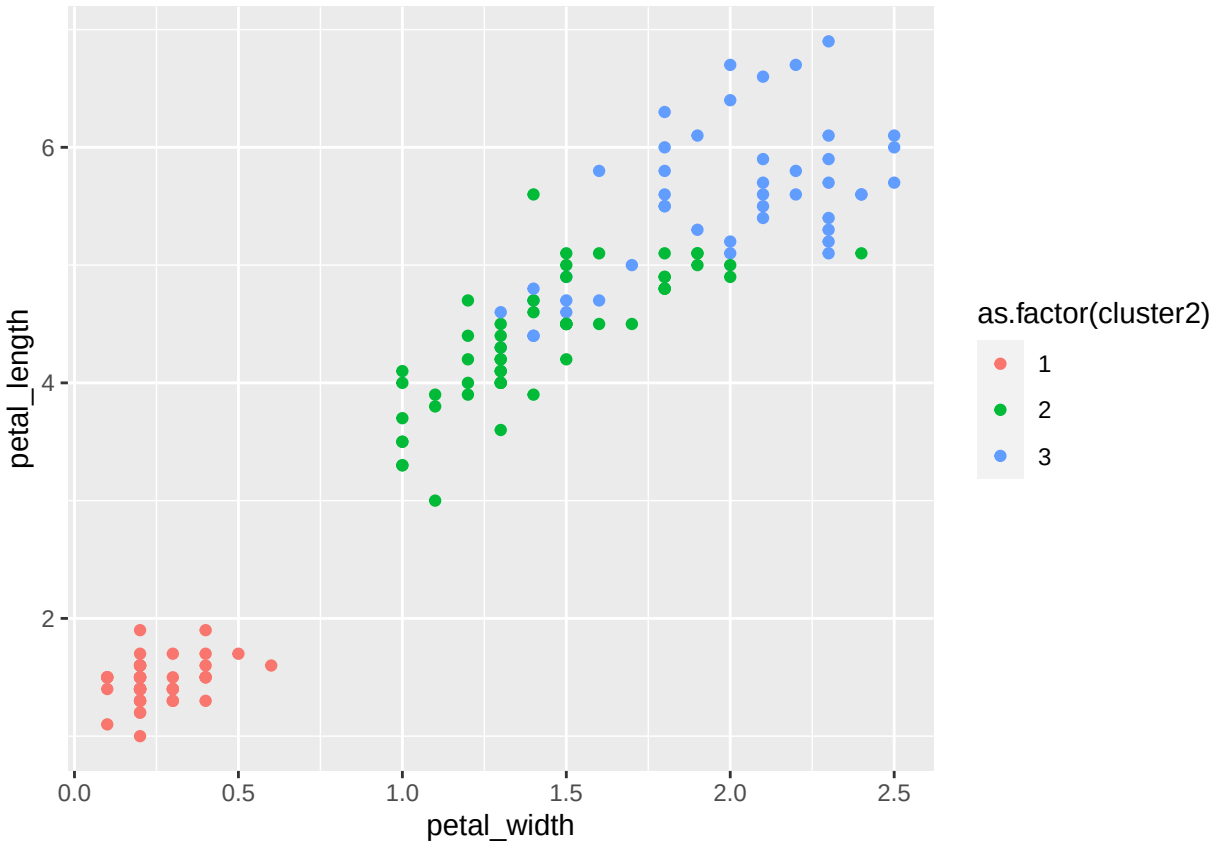

```
head(irisPCA$rotation)
```

```
##                     PC1         PC2         PC3          PC4         PC5
## sepal_length  0.4456261 -0.37804172  0.75213041 -0.140905632  0.27008753
## sepal_width  -0.2285825 -0.92268061 -0.28533914 -0.005005529 -0.12233999
## petal_length  0.5065716 -0.02641607  0.02912116  0.246522893 -0.82526712
## petal_width   0.4973627 -0.07007419 -0.38662225  0.609640302  0.47600623
## type_numeric  0.4951596  0.01169435 -0.45006290 -0.740057912  0.06661758
```

```
irisk4 <- kmeans(iris[,1:4, 5],3)
iris$cluster4 <- irisk4$cluster
ggplot(iris,aes(x=petal_width,y=petal_length,color=as.factor(cluster4))) + geom_point()
```




```
irisk2 <- kmeans(iris[,1:2],3)
iris$cluster2 <- irisk2$cluster
ggplot(iris,aes(x=petal_width,y=petal_length,color=as.factor(cluster2))) + geom_point()
```




```
irisk4$tot.withinss
```

```
## [1] 78.94084
```

```
irisk2$tot.withinss
```

```
## [1] 37.1237
```

Judging from the graphs that I may or may not have done incorrectly, the clustering using all of the columns is quite a bit off and looks weird, whereas the clustering using only the two principal components is much closer to that of the first graph. To check this, I used tot.withinss and observed that the value for the clustering with only the two principal components was much lower than that of the clustering for all values, meaning that it was more accurate.