Untitled

none

12/6/2021

Data summary

This data consist of 150 rows with 5 columns. Column summary are given below.

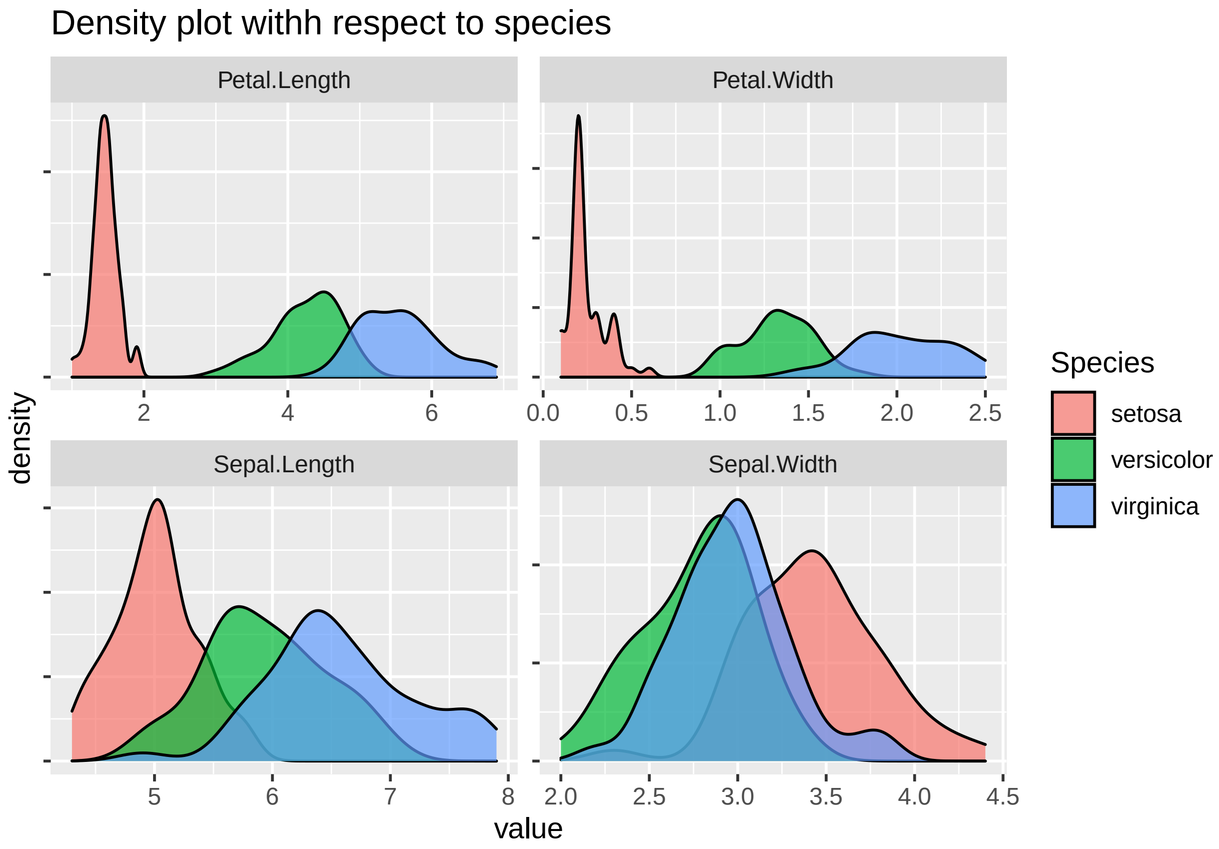
- Sepal.Length: Length of the sepal (in cm)
- Sepal.Width: Width of the sepal (in cm)
- Petal.Length: Length of the petal (in cm)
- Petal.Width: Width of the petal (in cm)
- Species: Species name

```
iris %>%
    summary()
```

```
##   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  
##                 
##                 
## 
```

Density plot of different variables

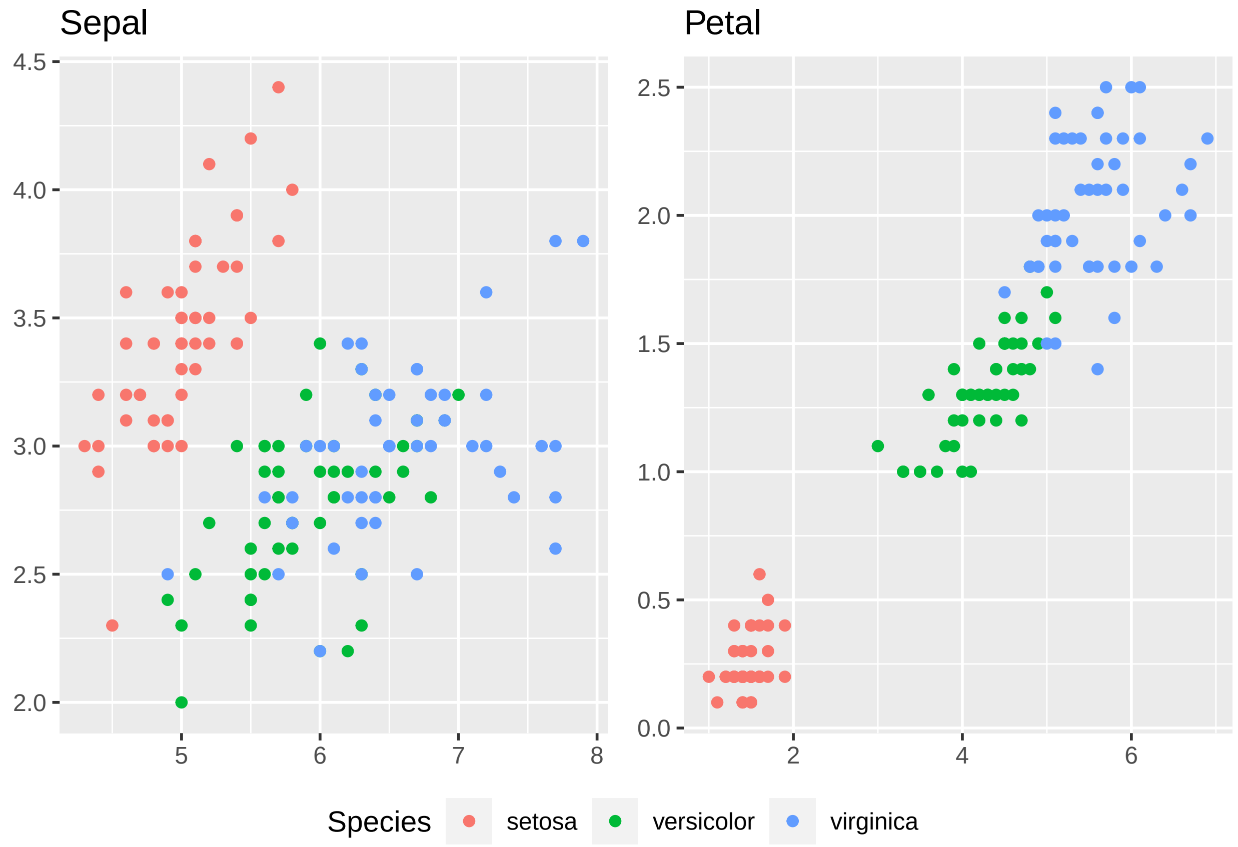
```
iris %>%
    pivot_longer(-Species) %>%
    ggplot(aes(value, fill = Species)) +
    geom_density(alpha = .7) +
    facet_wrap(~name, scales = "free") +
    scale_y_continuous(labels = NULL) +
    labs(title = "Density plot withh respect to species")
```



So without the sepal width other 3 variables are very sensitive to the species class.

Scatter plot of the Sepal and Petal lengh with respect to the Species

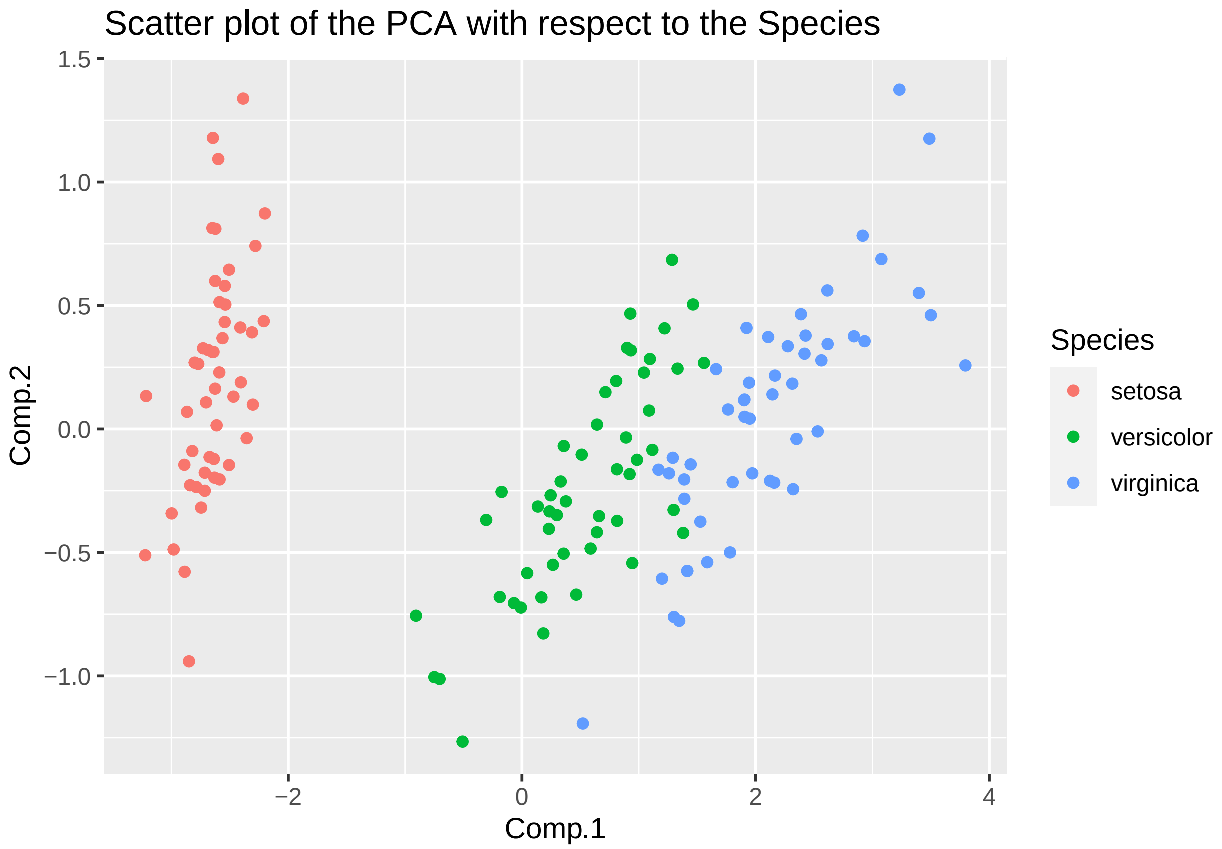
```r
ggarrange(
    iris %>%
        ggplot(aes(Sepal.Length, Sepal.Width, col = Species)) +
        geom_point() +
        labs(title = "Sepal",x = NULL, y = NULL),
    iris %>%
        ggplot(aes(Petal.Length, Petal.Width, col = Species)) +
        geom_point() +
        labs(title = "Petal",x = NULL, y = NULL), common.legend = T, legend = "bottom"
)
```



So we can see that Petal length and Petal width can separate the species and Sepal length and sepal width can separate the species partially since the species "versicolor" and "verginica" are mixed up in that plot.

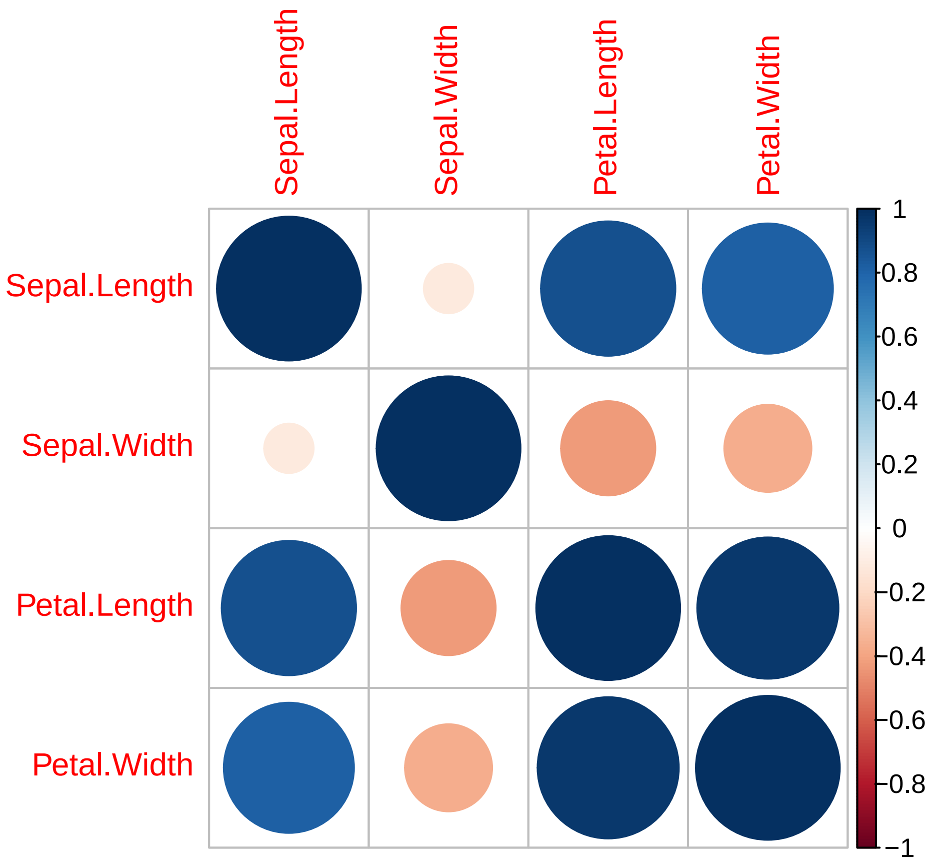
Scatter plot of the PCA with respect to the Species

```
(iris %>%
    select(-Species) %>%
    princomp())$scores %>%
    as.data.frame() %>%
    select(1:2) %>%
    bind_cols(Species = iris$Species) %>%
    ggplot(aes(Comp.1, Comp.2, col = Species)) +
    geom_point() +
    labs(title = "Scatter plot of the PCA with respect to the Species")
```



Correlation among the predictor variables

```
cor(select(iris, -Species)) %>%
corrplot::corrplot()
```



We can see a high correlation the explanatory variables. So this multi-collinearity may cause an performance drop of the model.

Spliting the test and train set

```
set.seed(1234)
iris_train <-
    iris %>%
    mutate(id = row_number(), .before = everything()) %>%
    group_by(Species) %>%
    slice_sample(n = 35)

iris_test <-
    iris %>%
    mutate(id = row_number(), .before = everything()) %>%
    anti_join(iris_train, by = "id") %>%
    select(-id)

iris_train <- select(iris_train, -id)
```

Fitting model

```
model <- multinom(formula = Species ~ ., data = iris_train)
```

```
## # weights:  18 (10 variable)
## initial  value 115.354290
## iter  10 value 10.451001
## iter  20 value 0.395906
## iter  30 value 0.063643
## iter  40 value 0.032814
## iter  50 value 0.023376
## iter  60 value 0.021200
## iter  70 value 0.019040
## iter  80 value 0.017101
## iter  90 value 0.012218
## iter 100 value 0.011383
## final  value 0.011383
## stopped after 100 iterations
```

```
table(predicted = predict(model, iris_test), true = iris_test$Species)
```

```
##               true
## predicted    setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         13         1
##   virginica       0          2        14
```

The accuracy of the multinomial logistic regression is (1 - 3/45) = 0.933 or 93%. Lets see whether the PCA can give us better performance or not

Fitting model with PCA

```
model_pca <-
    (iris_train %>%
         ungroup() %>%
         select(-Species) %>%
         princomp())$score %>%
    as.data.frame() %>%
    bind_cols(Species = iris_train$Species) %>%
    multinom(formula = Species ~ .)
```

```
## # weights:  18 (10 variable)
## initial  value 115.354290
## iter  10 value 4.343745
## iter  20 value 0.020857
## iter  30 value 0.002420
## iter  40 value 0.002053
## iter  50 value 0.001994
## iter  60 value 0.001105
## iter  70 value 0.000666
## iter  80 value 0.000543
## iter  90 value 0.000537
## iter 100 value 0.000447
## final  value 0.000447
## stopped after 100 iterations
```

```
table(predicted = predict(model_pca,
        newdata = (iris_test %>%
                       ungroup() %>%
                       select(-Species) %>%
                       princomp())$score), true = iris_test$Species)
```

```
##               true
## predicted    setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         13         1
##   virginica       0          2        14
```

The accuracy of the multinomial PCA logistic regression is (1 - 3/45) = 0.933 or 93%. which is equal to the the previous model. So both of the model actually performing the same way.

So the conclusion is we don't have enough evidence that whether the PCA logistic regression perform well or not.