Iris Project

Annie Dang

10/16/2021

Initial Exploratory Data Analysis

Here we are loading in the iris dataset with "read.csv" and displaying the first few entries of this dataset using the head() function. This process is very similar to that of pandas.read_csv() and df.head(). There are 5 columns, 4 of which being quantitative (with the first one being and index column) and 1 being qualitative. The last column contains the labels of the iris species.

```
iris = read.csv("/Users/anniedang/Desktop/iris.csv")
head(iris)
```

```
##   X sepal_length sepal_width petal_length petal_width     species
## 1 0          5.1         3.5          1.4         0.2 Iris-setosa
## 2 1          4.9         3.0          1.4         0.2 Iris-setosa
## 3 2          4.7         3.2          1.3         0.2 Iris-setosa
## 4 3          4.6         3.1          1.5         0.2 Iris-setosa
## 5 4          5.0         3.6          1.4         0.2 Iris-setosa
## 6 5          5.4         3.9          1.7         0.4 Iris-setosa
```

We are now displaying the frequencies of each species present in the iris dataset. This process is appending a percentage column to a grouped table of iris species. This process can be imitated in pandas using the groupby function, combined with creating a new column of percentages. In this dataset, there is an equal distribution of each species. There are 150 entries in this dataset.

```
cbind(freq = table(iris$species), percentage=prop.table(table(iris$species)) * 100)
```

```
##                 freq percentage
## Iris-setosa       50   33.33333
## Iris-versicolor   50   33.33333
## Iris-virginica    50   33.33333
```

This displays six desciptive statistics relating to the different quantitative columns of this dataset. This process is very similar to that of df.describe in Python's pandas.

```
summary(iris)
```

```
##        X           sepal_length    sepal_width     petal_length
##  Min.   :  0.00   Min.   :4.300   Min.   :2.000   Min.   :1.000
##  1st Qu.: 37.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
##  Median : 74.50   Median :5.800   Median :3.000   Median :4.350
```

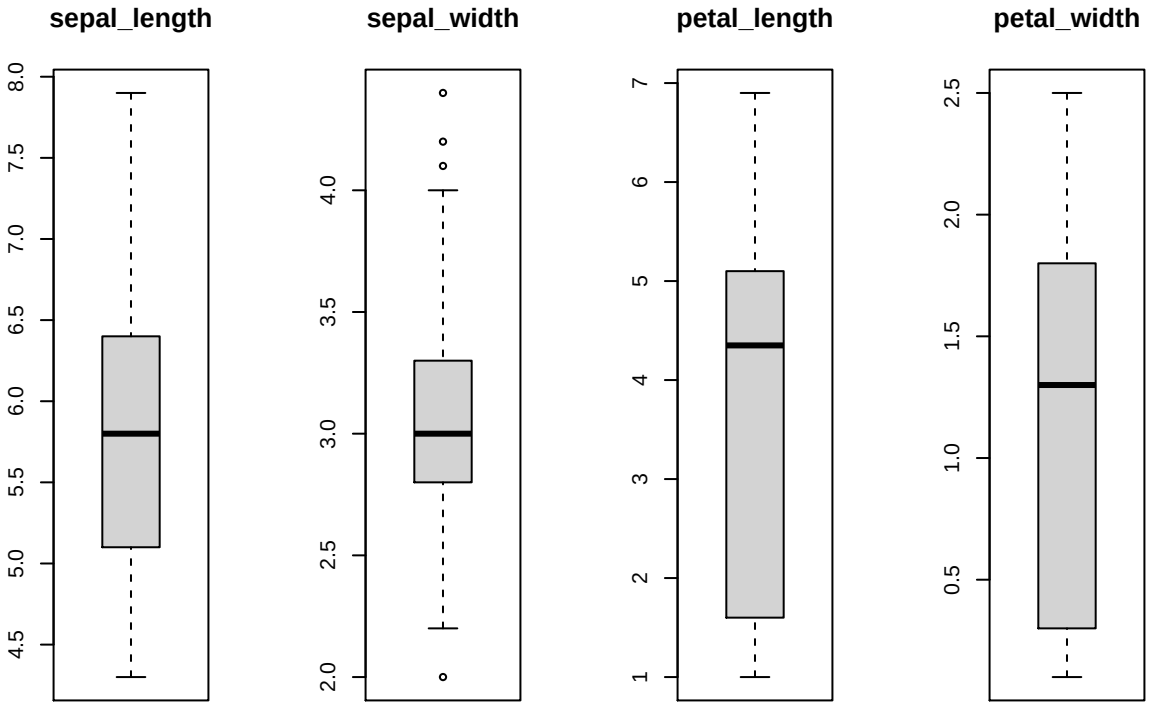
```
##  Mean   : 74.50   Mean   :5.843   Mean   :3.054   Mean   :3.759
##  3rd Qu.:111.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
##  Max.   :149.00   Max.   :7.900   Max.   :4.400   Max.   :6.900
##   petal_width       species
##  Min.   :0.100   Length:150
##  1st Qu.:0.300   Class :character
##  Median :1.300   Mode  :character
##  Mean   :1.199
##  3rd Qu.:1.800
##  Max.   :2.500
```

Visualizations

First, we split the dataset into independent and dependent variables using indexing, very similar to python indexing.

```
x <- iris[ , 2:5]
y <- iris[ , 6]
```

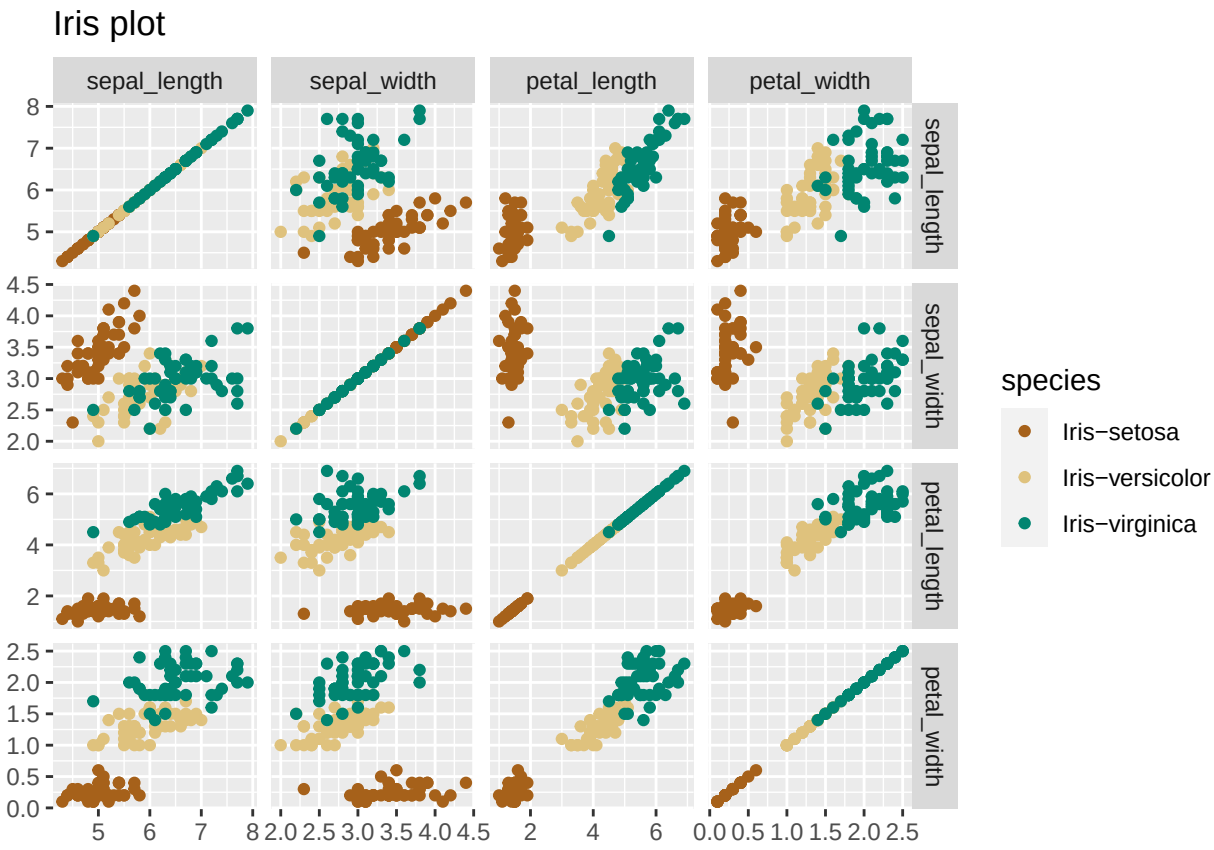
Here, we diplay barplots to describe the distribution of the first four columns of the iris dataset; we can do the same in python using matplotlib's bar function. We see here that petal length and petal width has the largest variance in its values and sepal width has the least variance in its values. Sepal width also has the most out-



liers.
This is a pairplot of the iris dataset. As we can see, the species have distinct clustering especially when compairing petal length and sepal length, petal length and petal width, petal width and sepal width, or sepal width and petal length. Iris-setosa can be linearly separated from the other two species, while there is

some overlap between iris-versicolor and iris-virginica. This implies that these iris attributes may be good features to use for modeling. This process of a pairplot is very similar to that of Python's seaborn pairplot.

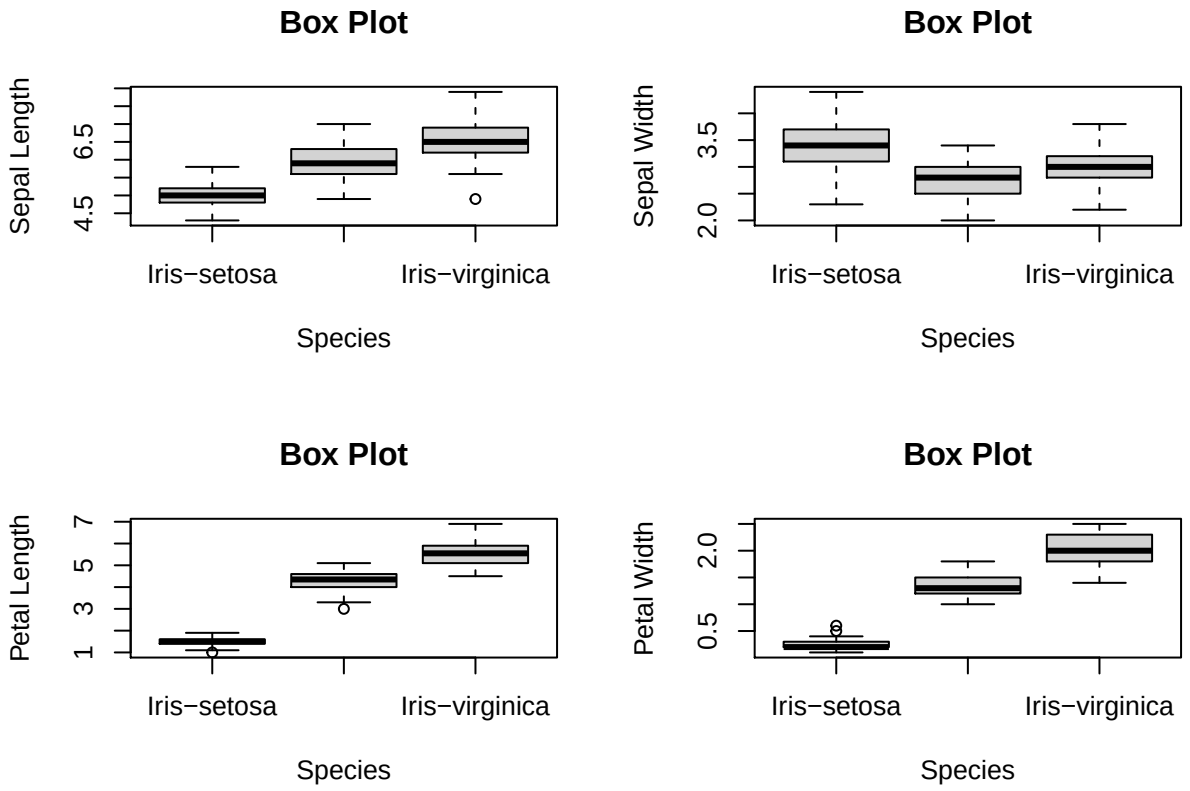
```
colormap = c('#a6611a', '#dfc27d', '#018571')
PairPlot(iris, colnames(iris)[2:5], "Iris plot",
         group_var = "species", palette=NULL) +
         ggplot2::scale_color_manual(values=colormap)
```



Assessing the boxplots of each feature stratrified by species, we see that there is less of an overlapping distribution between each species for sepal length, petal length, and petal width.

```
par(mfrow=c(2,2))
boxplot(sepal_length ~ species, data=iris,
     main="Box Plot",
     xlab="Species",
     ylab="Sepal Length")
boxplot(sepal_width ~ species, data=iris,
     main="Box Plot",
     xlab="Species",
     ylab="Sepal Width")
boxplot(petal_length ~ species, data=iris,
     main="Box Plot",
     xlab="Species",
     ylab="Petal Length")
boxplot(petal_width ~ species, data=iris,
     main="Box Plot",
```

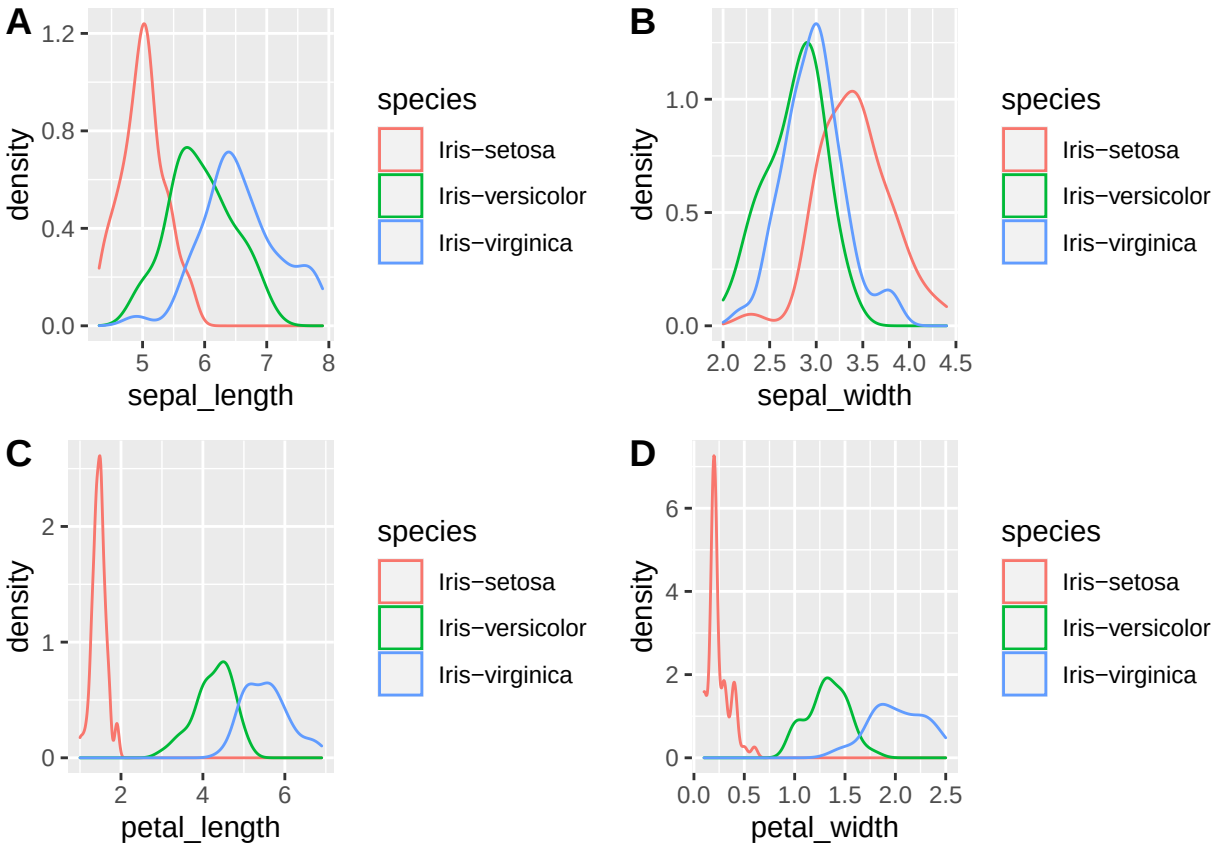
```
     xlab="Species",
     ylab="Petal Width")
```



A diffirent view of the different iris attributes stratified by species. In this view, we can see that the distributions are in fact a little more overlapping than the boxplots would suggest.

```
iris1 = ggplot(iris, aes(x = sepal_length, color = species)) + geom_density()
iris2 =ggplot(iris, aes(x = sepal_width, color = species)) + geom_density()
iris3 =ggplot(iris, aes(x = petal_length, color = species)) + geom_density()
iris4 =ggplot(iris, aes(x = petal_width, color = species)) + geom_density()

plot_grid(iris1, iris2, iris3, iris4, labels = "AUTO");
```



Petal length and petal width have the strongest positive correlation. However, petal length and sepal length as well as petal width and sepal length also have very strong postive correlations.

```
cor(x)
```

```
##              sepal_length sepal_width petal_length petal_width
## sepal_length    1.0000000  -0.1093692    0.8717542   0.8179536
## sepal_width    -0.1093692   1.0000000   -0.4205161  -0.3565441
## petal_length    0.8717542  -0.4205161    1.0000000   0.9627571
## petal_width     0.8179536  -0.3565441    0.9627571   1.0000000
```

Insights from EDA

From our EDA we can see that:

1. Setosa has the smallest petal length and petal width, while virginica has the largest petal length and width. Versicolor has avergae petal length and width.
2. Setosa is distringuishable and have very different characteristics (linearly separable) from the other species.
3. Setosa has smaller attributes and less outliers, versicolor has average attributes, and virginica has the longest attributes.
4. The strongest correlation is between petal width and petal length.
5. Versicolor and virginica are hard to separate for most characteristics

Modeling

Since all features (petal length, petal width, sepal length, sepal width) seem to be good indicators for species, all features will be used for modeling. Additionally, since the pairplot shows goos separability between the species for different characteristics, random forst and support vector machines (SVM) may be the best choice for modeling.

```
training_idx <- createDataPartition(iris$species, p=0.74, list = FALSE)
testing <- iris[-training_idx, 2:6]
iris_training <- iris[training_idx, 2:6]
```

```
# 10 fold cross validation and accuracy as our performance metric
cv <- trainControl(method="cv", number=10)
metric <- "Accuracy"

#SVM
fit.svm <- train(species~., data=iris_training, method="svmRadial", metric=metric, trControl=cv)

#Random Forest
fit.rf <- train(species~., data = iris_training, method="rf", metric = metric, trControl = cv)
```

Random Forest does seem to have better performance. Althoug, random forests are known to overfit to the training set by virtue of the algorithm's splitting nature.

```
summary(resamples(list(svm=fit.svm, rf=fit.rf)))
```

```
##
## Call:
## summary.resamples(object = resamples(list(svm = fit.svm, rf = fit.rf)))
##
## Models: svm, rf
## Number of resamples: 10
##
## Accuracy
##          Min.   1st Qu.    Median      Mean 3rd Qu. Max. NA's
## svm 0.8181818 0.9109848 0.9583333 0.9469697       1    1    0
## rf  0.8333333 0.9375000 1.0000000 0.9659091       1    1    0
##
## Kappa
##          Min.  1st Qu. Median      Mean 3rd Qu. Max. NA's
## svm 0.7317073 0.865625 0.9375 0.9204967       1    1    0
## rf  0.7500000 0.906250 1.0000 0.9487500       1    1    0
```

Surprisingly, we get 100% accuracy on the testing data.

```
predictions <- predict(fit.rf, testing)
confusionMatrix(as.factor(testing$species), as.factor(predictions))
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Iris-setosa Iris-versicolor Iris-virginica
```

```
##   Iris-setosa              13               0              0
##   Iris-versicolor           0              12              1
##   Iris-virginica            0               1             12
##
## Overall Statistics
##
##                Accuracy : 0.9487
##                  95% CI : (0.8268, 0.9937)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : 7.509e-16
##
##                   Kappa : 0.9231
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Iris-setosa Class: Iris-versicolor
## Sensitivity                      1.0000                 0.9231
## Specificity                      1.0000                 0.9615
## Pos Pred Value                   1.0000                 0.9231
## Neg Pred Value                   1.0000                 0.9615
## Prevalence                       0.3333                 0.3333
## Detection Rate                   0.3333                 0.3077
## Detection Prevalence             0.3333                 0.3333
## Balanced Accuracy                1.0000                 0.9423
##                      Class: Iris-virginica
## Sensitivity                         0.9231
## Specificity                         0.9615
## Pos Pred Value                      0.9231
## Neg Pred Value                      0.9615
## Prevalence                          0.3333
## Detection Rate                      0.3077
## Detection Prevalence                0.3333
## Balanced Accuracy                   0.9423
```