Analiza zestawu pomiarów kwiatów Irysa

304358, Piotr WITEK, czwartek 16^{15}
AGH, Wydział Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2020/2021

Kraków, January 24, 2021

Ja, niżej podpisany(na) własnoręcznym podpisem deklaruję, że przygotowałem(łam) przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

Piotr Witek

1 Streszczenie raportu

W ramach projektu na przedmiot Rachunek Prawdopodobieństwa i Statystyka dokonałem analizy danych na podstawie zestawu zawierającego pomiary kwiatów Irysa. Zestaw danych jest stosunkowo niewielki, zawiera 150 pomiarów, 3 rodzajów Irysów.

Korzystając z pakietu R wykonałem podstawową analizę danych. Badałem 4 cechy: Sepal Length, Sepal Width, Petal Length oraz Petal Width. Dokonałem analizy rodzaju rozkładu badanych danych wykorzystując test Shapiro-Wilka. Na podstawie wyników testu, wnioskuję, że rozkład cechy Sepal Width jest zbliżony do rozkładu normalnego. Operacje na pozostałych zmiennych wykluczyły podobieństwo do rozkładu normalnego.

Interesujące stało się badanie korelacji pomiędzy zmiennymi. Korzystając z operacji, które udostępnia środowisko do obliczeń statystycznych stworzyłem macierz korelacji, z której odczytałem, jakie zmienne korelują z innymi. Stworzyłem serię wykresów obrazujących te zależności z użyciem regresji, opisując jaka zmienna zależy od innej i w jakim stopniu.

Aby podkreślić czytelność i przejrzystość wykonywanej analizy, korzystałem z wielu tabel i wykresów.

Badany przeze mnie zestaw danych jest często wykorzystywany do nauki Machine Learningu, ze względu na przejrzystość danych, stosunkowo niewielką liczbę zmiennych, oraz ciekawe zależności między zmiennymi.

2 Opis danych

Dataset 'Iris spiecies' pochodzi ze strony `https://www.kaggle.com/uciml/iris` . Ten zbiór danych został opisany w 1936 roku i zawiera zmierzone wartości parametrów kwiatów trzech gatunków Irysów: Iris Setosa, Iris Versicolour i Iris Virginica.

Nazwy parametrów:

```
> names(data)
```

```
[1] "Id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm"
[5] "PetalWidthCm"  "Species"
```

Podsumowanie datasetu:

```
> summary(data)
```

```
       Id         SepalLengthCm    SepalWidthCm   PetalLengthCm
 Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000
 1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
 Median : 75.50   Median :5.800   Median :3.000   Median :4.350
 Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759
 3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
 Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900
  PetalWidthCm     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
```

3 Potrzebne biblioteki użyte w projekcie

```
> library(ggplot2)
> library(rstudioapi)
> library(gridExtra)
> library(grid)
> library(plyr)
> library(GGally)
> library(plotly)
> library(e1071)
```

4 Podstawowa analiza danych

W ramach podstawowej analizy danych dla każdej z czterech zmiennych (Sepal Length, Sepal Width, Petal Length, Petal Width) obliczono wartości takie jak: odchylenie standardowe, rozstęp, rozstęp kwartlowy, skośność, momenty oraz wykonano histogram.

4.1 Sepal Length

Podsumowanie:

```
> summary(data$SepalLength)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
```

Odchylenie standardowe:

```
> sd(data$SepalLength)

[1] 0.8280661
```

Rozstęp:

```
> range(data$SepalLength)

[1] 4.3 7.9
```

Rozstęp kwartylowy:

```
> IQR(data$SepalLength)

[1] 1.3
```

Skośność:

```
> skewness(data$SepalLength)

[1] 0.3086407
```

Momenty centralne:

```
> moment(data$SepalLength,0.25)

[1] 1.55188

> moment(data$SepalLength,0.5)

[1] 2.411318

> moment(data$SepalLength,0.75)
```

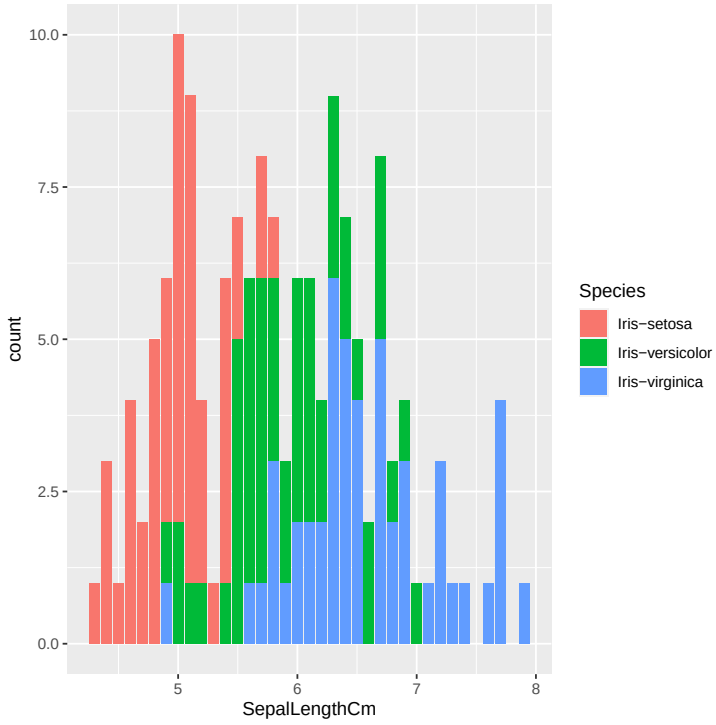
```
[1] 3.75136
```

```
> moment(data$SepalLength,1)
```

```
[1] 5.843333
```

Histogram:

```
> ggplot(data, aes(x=SepalLengthCm, fill=factor(Species)))+
+   geom_bar()+
+   scale_fill_discrete(name="Species")
```



4.2 SepalWidth

Podsumowanie:

```
> summary(data$SepalWidth)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.800   3.000   3.054   3.300   4.400
```

Odchylenie standardowe:

```
> sd(data$SepalWidth)
```

```
[1] 0.4335943
```

Rozstęp:

```
> range(data$SepalWidth)

[1] 2.0 4.4
```

Rozstęp kwartylowy:

```
> IQR(data$SepalWidth)

[1] 0.5
```

Skośność:

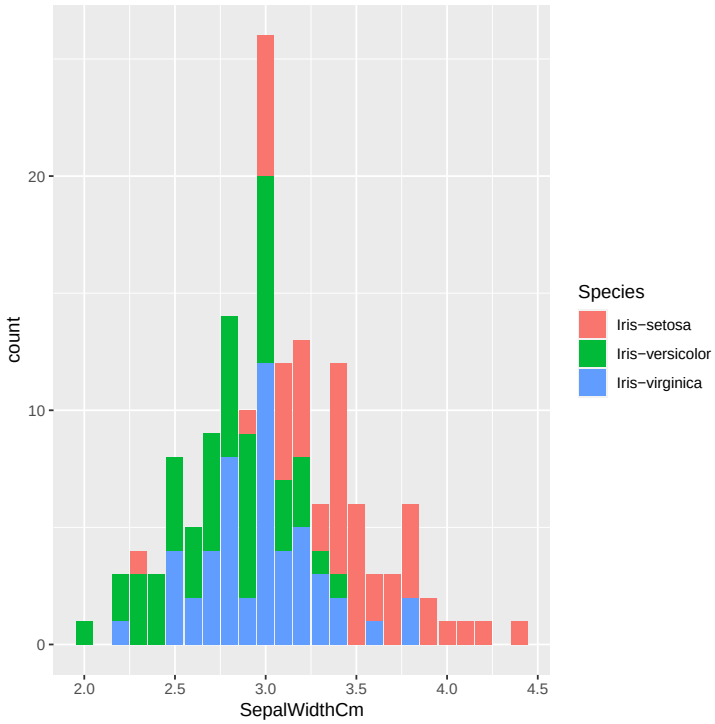
```
> skewness(data$SepalWidth)

[1] 0.3274013
```

Momenty centralne:

```
> moment(data$SepalWidth,0.25)

[1] 1.319481

> moment(data$SepalWidth,0.5)

[1] 1.743213

> moment(data$SepalWidth,0.75)

[1] 2.305896

> moment(data$SepalWidth,1)

[1] 3.054
```

Histogram:

```
> ggplot(data, aes(x=SepalWidthCm, fill=factor(Species)))+
+   geom_bar()+
+   scale_fill_discrete(name="Species")
```



4.3 Petal length

Podsumowanie:

```
> summary(data$PetalLength)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.600   4.350   3.759   5.100   6.900
```

Odchylenie standardowe:

```
> sd(data$PetalLength)
```

```
[1] 1.76442
```

Rozstęp:

```
> range(data$PetalLength)
```

```
[1] 1.0 6.9
```

Rozstęp kwartylowy:

```
> IQR(data$PetalLength)
```

```
[1] 3.5
```

Skośność:

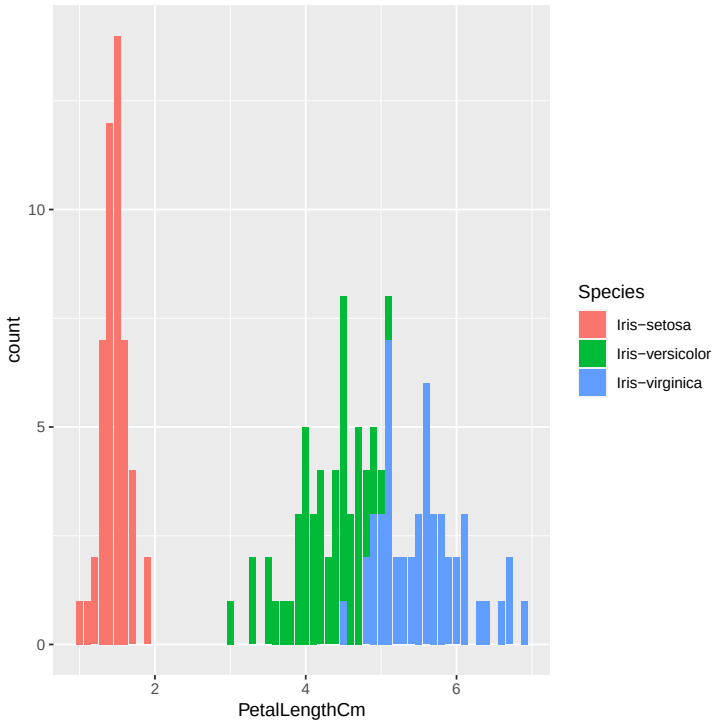
```
> skewness(data$PetalLength)

[1] -0.2689994
```

Momenty centralne:

```
> moment(data$PetalLength,0.25)

[1] 1.355716

> moment(data$PetalLength,0.5)

[1] 1.874027

> moment(data$PetalLength,0.75)

[1] 2.634887

> moment(data$PetalLength,1)

[1] 3.758667
```

Histogram:

```
> ggplot(data, aes(x=PetalLengthCm, fill=factor(Species)))+
+   geom_bar()+
+   scale_fill_discrete(name="Species")
```



4.4 Petal Width

Podsumowanie:

```
> summary(data$PetalWidth)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   0.300   1.300   1.199   1.800   2.500
```

Odchylenie standardowe:

```
> sd(data$PetalWidth)
```

```
[1] 0.7631607
```

Rozstęp:

```
> range(data$PetalWidth)
```

```
[1] 0.1 2.5
```

Rozstęp kwartylowy:

```
> IQR(data$PetalWidth)
```

```
[1] 1.5
```

Skośność:

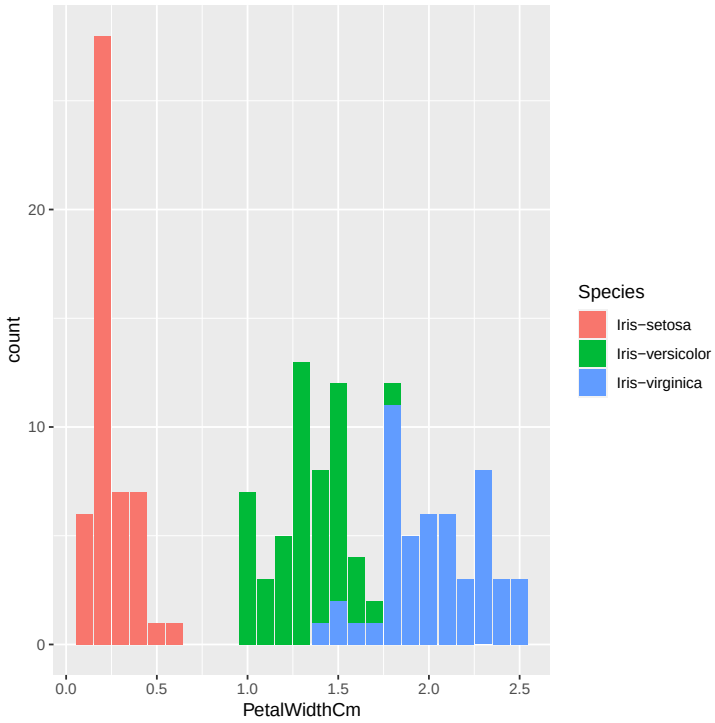
```
> skewness(data$PetalWidth)

[1] -0.102906
```

Momenty centralne:

```
> moment(data$PetalWidth,0.25)

[1] 0.9843897

> moment(data$PetalWidth,0.5)

[1] 1.017222

> moment(data$PetalWidth,0.75)

[1] 1.08996

> moment(data$PetalWidth,1)

[1] 1.198667
```

Histogram:

```
> ggplot(data, aes(x=PetalWidthCm, fill=factor(Species)))+
+   geom_bar()+
+   scale_fill_discrete(name="Species")
```

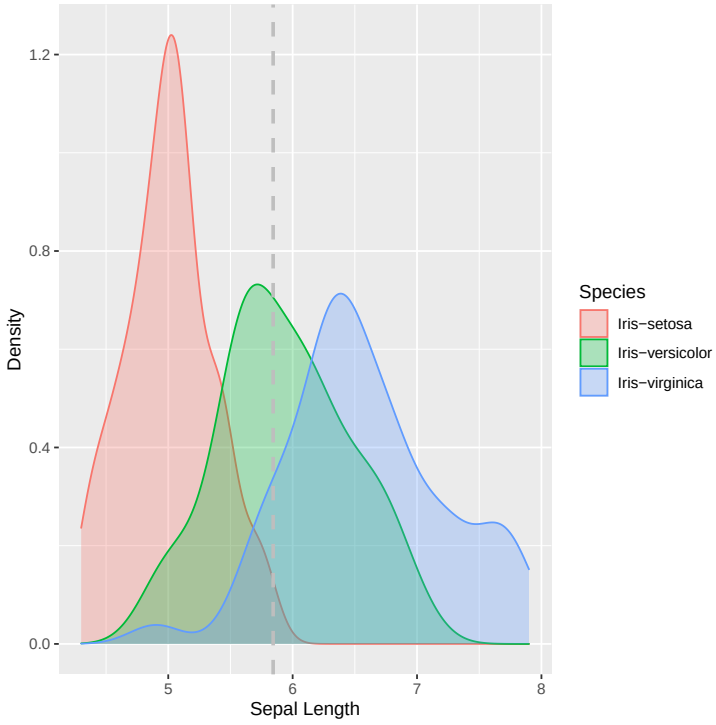

count
20
10
0
0.0
0.5
1.0
1.5
2.0
2.5
PetalWidthCm
Species
Iris−setosa
Iris−versicolor
Iris−virginica

5 Analiza danych przy użyciu histogramu gęstości

Uwtorzono histogramy obrazujące gęstości każdego atrybutu z zaznaczenie podziału na rodzaje irysów. Dzięki tym wykresom możemy przewidywać rozkład dla każdego parametru oraz jesteśmy w stanie zauważyć rozdział rodzajów irysów.

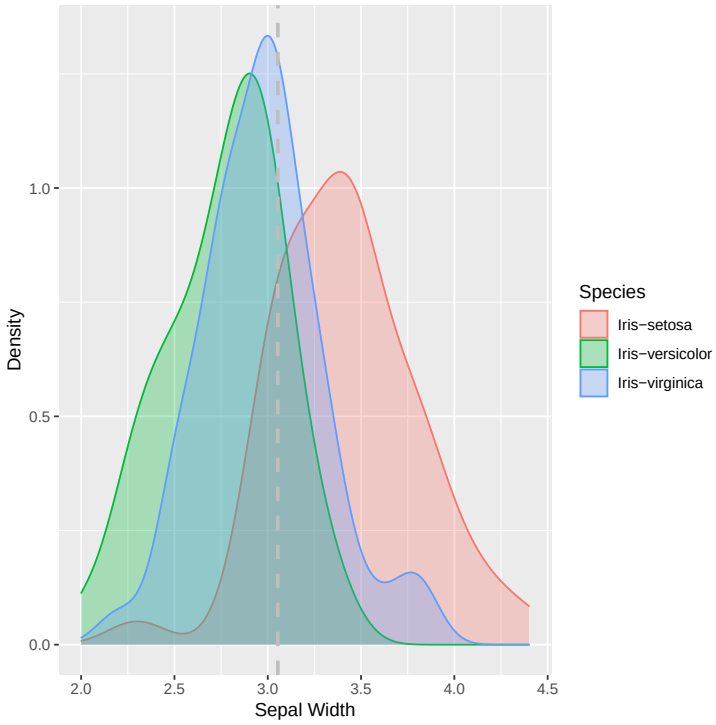
5.1 Sepal Length

```
> DhistSl <- ggplot(data, aes(x=SepalLengthCm, colour=Species, fill=Species)) +
+   geom_density(alpha=.3) +
+   geom_vline(aes(xintercept=mean(SepalLengthCm),  colour=Species),linetype="dashed", colo
+   xlab("Sepal Length") +
+   ylab("Density")
> DhistSl
```



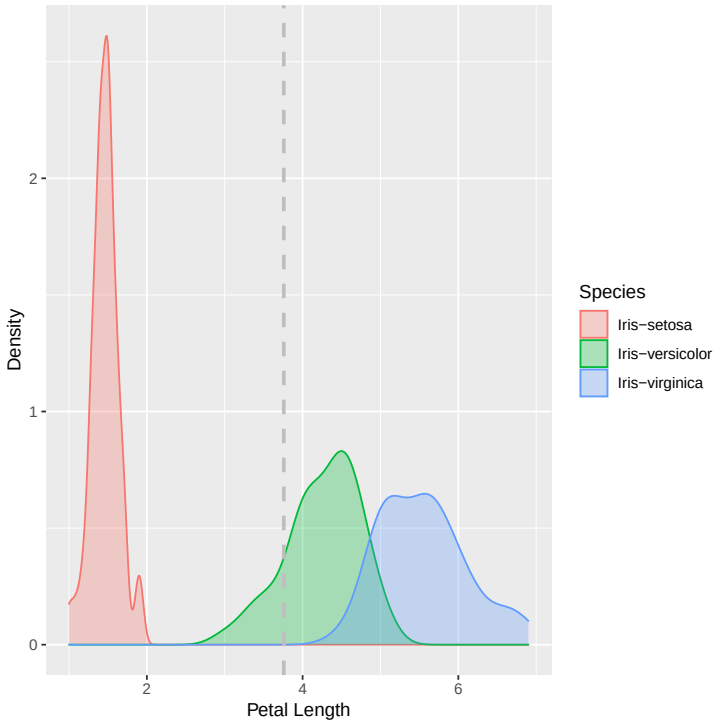
5.2 Sepal Width

```
> DhistSw <- ggplot(data, aes(x=SepalWidthCm, colour=Species, fill=Species)) +
+   geom_density(alpha=.3) +
+   geom_vline(aes(xintercept=mean(SepalWidthCm),  colour=Species),linetype="dashed", color=
+   xlab("Sepal Width") +
+   ylab("Density")
> DhistSw
```



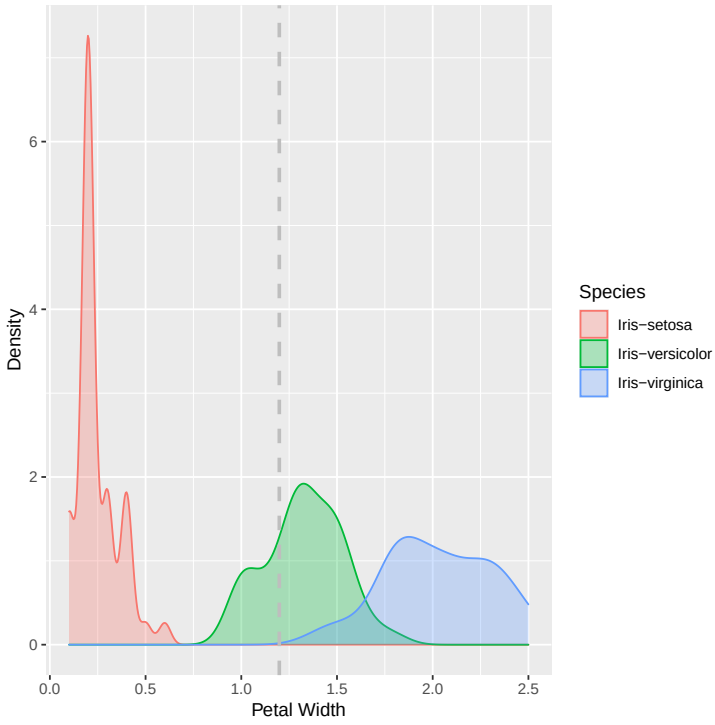
5.3 Petal Length

```
> DhistPl <- ggplot(data, aes(x=PetalLengthCm, colour=Species, fill=Species)) +
+   geom_density(alpha=.3) +
+   geom_vline(aes(xintercept=mean(PetalLengthCm),  colour=Species),linetype="dashed", colo
+   xlab("Petal Length") +
+   ylab("Density")
> DhistPl
```



5.4 Petal Width

```
> DhistPw <- ggplot(data, aes(x=PetalWidthCm, colour=Species, fill=Species)) +
+   geom_density(alpha=.3) +
+   geom_vline(aes(xintercept=mean(PetalWidthCm),  colour=Species),linetype="dashed", color=
+   xlab("Petal Width") +
+   ylab("Density")
> DhistPw
```



5.5 Wnioski

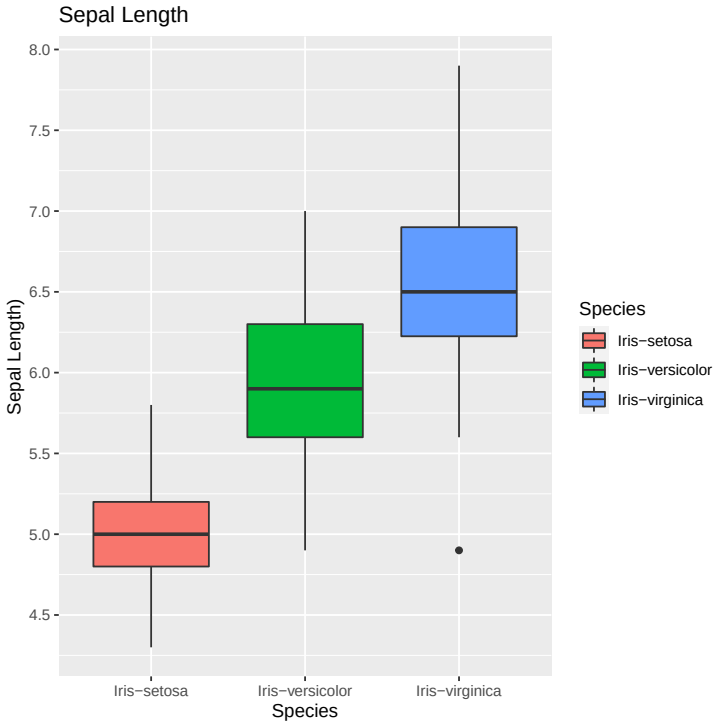
Na podstawie analizy wykresów gęstości można podejrzewać, że rozkład zmiennej Sepal Width jest zbliżony do rozkładu normalnego.

6 Analiza z użyciem wykresów pudełkowych

Analizę przeprowadzono w celu wykrycia wartości odstających(tzw. outliers) oraz w celu zobrazowania pokrycia się różnych klas irysów. Wykresy pudełkowe stanowią dobre narzedzie do ilustrowania różnic pomiędzy porównywanymi grupami.

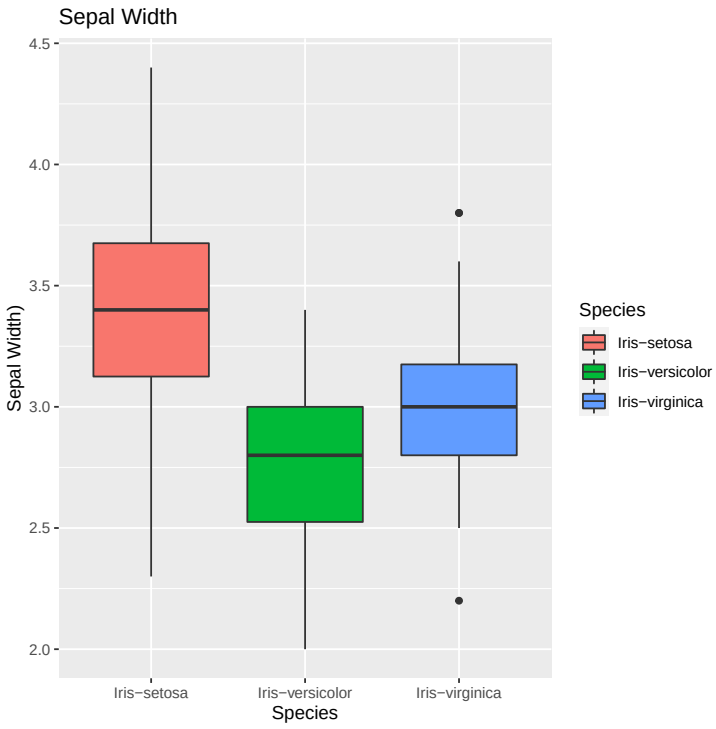
6.1 Sepal Length

```
> ggplot(data, aes(Species, SepalLengthCm, fill=Species)) +
+   geom_boxplot()+
+   scale_y_continuous("Sepal Length)", breaks= seq(0,30, by=.5))+
+   labs(title = "Sepal Length", x = "Species")
```



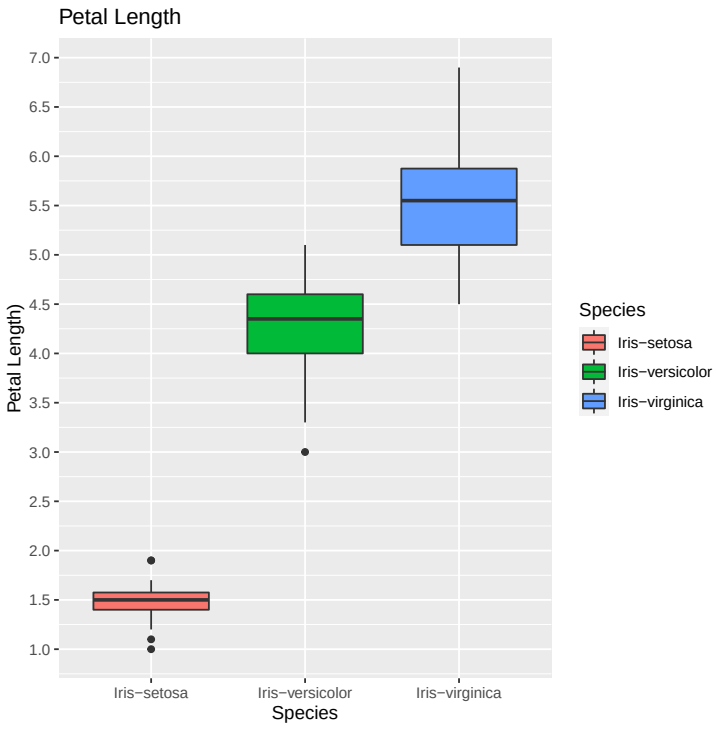
6.2 Sepal Width

```
> ggplot(data, aes(Species, SepalWidthCm, fill=Species)) +
+   geom_boxplot()+
+   scale_y_continuous("Sepal Width)", breaks= seq(0,30, by=.5))+
+   labs(title = "Sepal Width", x = "Species")
```



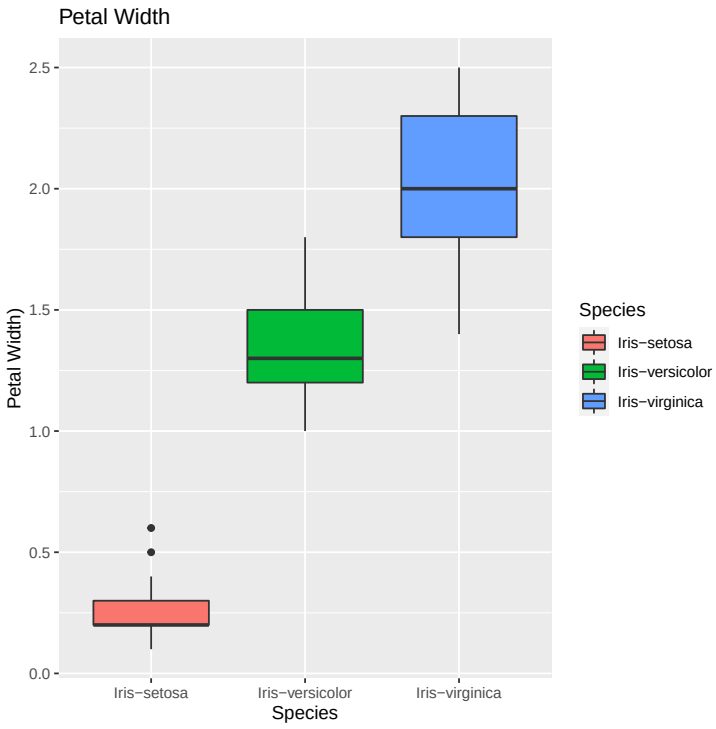
6.3 Petal Length

```
> ggplot(data, aes(Species, PetalLengthCm, fill=Species)) +
+   geom_boxplot()+
+   scale_y_continuous("Petal Length)", breaks= seq(0,30, by=.5))+
+   labs(title = "Petal Length", x = "Species")
```



6.4 Petal Width

```
> ggplot(data, aes(Species, PetalWidthCm, fill=Species)) +
+   geom_boxplot()+
+   scale_y_continuous("Petal Width)", breaks= seq(0,30, by=.5))+
+   labs(title = "Petal Width", x = "Species")
```



6.5 Wnioski

Przy niektórych ”pudełkach” można zauważyć pojedyncze wartości odstające. Dla zmiennych PetalWidth i PetalLength boxy nie pokrywają się, a dla zmiennym SepalWidth i SepalLength pokrywają się.

7 Analiza rodzaju rozkładu badanych cech

Do analizy użyto testu Shapiro-Wilk z uwagi na stosunkowo niewielki rozmiar danych. W ramach testu rozkładu wykorzystano paramentr jakim jest kurtoza. Kurtoza rozkładu normalnego wynosi 0.

7.1 Sepal Length

```
> kurtosis(data$SepalLengthCm)

[1] -0.6058125

> shapiro.test(data$SepalLengthCm)

        Shapiro-Wilk normality test

data:  data$SepalLengthCm
W = 0.97609, p-value = 0.01018
```

7.1.1 Wnioski

Kurtoza jest ujemna, wartości cechy mniej skoncentrowane niż przy rozkładzie normalnym. p-value < 0.05, rozkład nie jest normalny.

7.2 Sepal Width

```
> kurtosis(data$SepalWidthCm)

[1] 0.1983681

> shapiro.test(data$SepalWidthCm)

        Shapiro-Wilk normality test

data:  data$SepalWidthCm
W = 0.98379, p-value = 0.07518
```

7.2.1 Wnioski

Kurtoza jest dodatnia, lecz bliska 0, wartości cechy bardziej skoncentrowane niż przy rozkładzie normalnym. p-value > 0.05, przyjmujemy hipotezę, że rozkład jest zbliżony do normalnego.

7.3 Petal Length

```
> kurtosis(data$PetalLengthCm)

[1] -1.416683

> shapiro.test(data$PetalLengthCm)

        Shapiro-Wilk normality test

data:  data$PetalLengthCm
W = 0.87642, p-value = 7.545e-10
```

7.3.1 Wnioski

Kurtoza jest ujemna, wartości cechy mniej skoncentrowane niż przy rozkładzie normalnym. p-value < 0.05, rozkład nie jest normalny.

7.4 Petal Width

```
> kurtosis(data$PetalWidthCm)

[1] -1.357368

> shapiro.test(data$PetalWidthCm)

        Shapiro-Wilk normality test

data:  data$PetalWidthCm
W = 0.90262, p-value = 1.865e-08
```

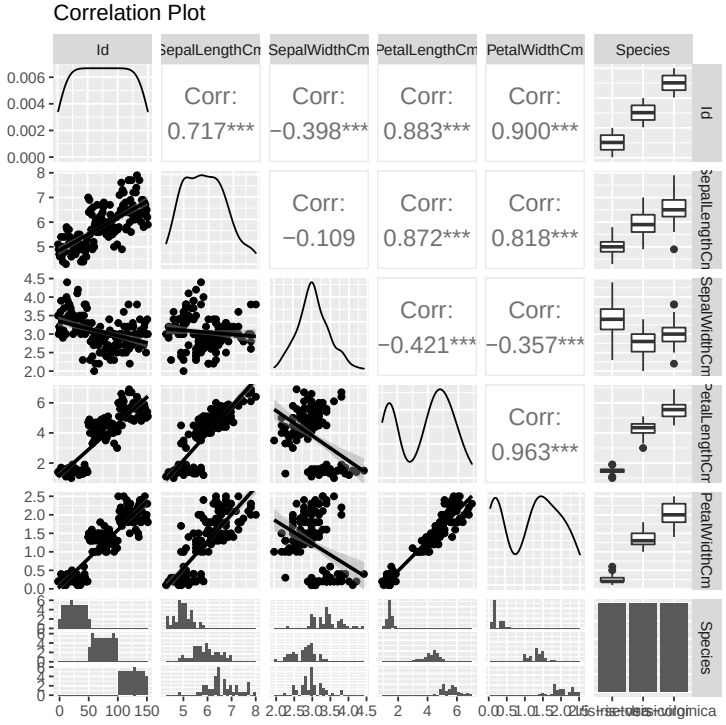
7.4.1 Wnioski

Kurtoza jest ujemna, wartości cechy mniej skoncentrowane niż przy rozkładzie normalnym. p-value < 0.05, rozkład nie jest normalny.

8 Korelacja

Macierz korelacji w datasecie:

```
> ggpairs(data = data,
+         title = "Correlation Plot",
+         upper = list(continuous = wrap("cor", size = 5)),
+         lower = list(continuous = "smooth")
+ )
```

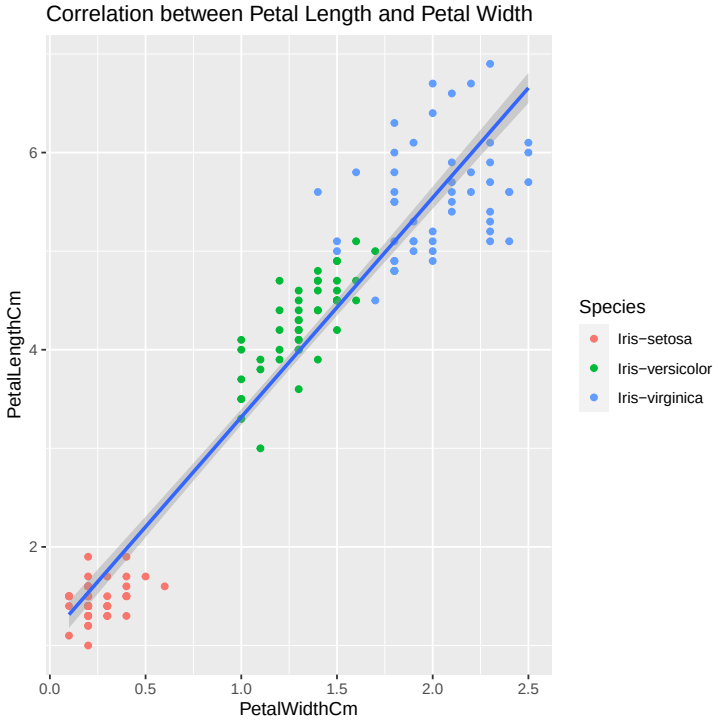



```
> cor(data[, 2:5])
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
SepalLengthCm	1.0000000	-0.1093692	0.8717542	0.8179536
SepalWidthCm	-0.1093692	1.0000000	-0.4205161	-0.3565441
PetalLengthCm	0.8717542	-0.4205161	1.0000000	0.9627571
PetalWidthCm	0.8179536	-0.3565441	0.9627571	1.0000000

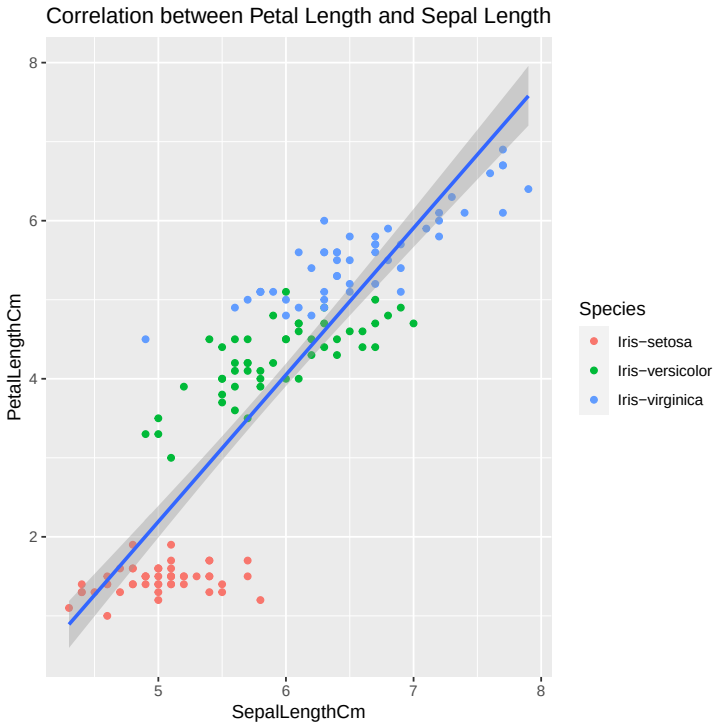
Zauważam, że istnieje wysoka korelacja pomiędzy PetalWidth i PetalLength - 96%

```
> ggplot(data, aes(x=PetalWidthCm, y=PetalLengthCm))+
+   geom_point(aes(colour=Species))+
+   geom_smooth(method='lm')+
+   ggtitle("Correlation between Petal Length and Petal Width")
```



Zauważam również silną korelację pomiędzy SepalLength i PetalLength - 87%

```
> ggplot(data, aes(x=SepalLengthCm, y=PetalLengthCm))+
+   geom_point(aes(colour=Species))+
+   geom_smooth(method='lm')+
+   ggtitle("Correlation between Petal Length and Sepal Length")
```



Korelacja pomiędzy SepalLength i PetalWidth wynosi 82%

```
> ggplot(data, aes(x=SepalLengthCm, y=PetalWidthCm))+
+   geom_point(aes(colour=Species))+
+   geom_smooth(method='lm')+
+   ggtitle("Correlation between Petal Width and Sepal Length")
```

Korelacja pomiędzy wartościami SepalLength i SepalWidth jest ujemna i wynosi -0.1. Ujemna korelacja oznacza, że wraz ze wzrostem/ spadkiem jednej zmiennej, druga zmnienna zachowuje się odwrotnie.

```
> ggplot(data, aes(x=SepalWidthCm, y=SepalLengthCm))+
+   geom_point(aes(colour=Species))+
+   geom_smooth(method='lm')+
+   ggtitle("Correlation between Sepal Length and Sepal Width")
```