Analiza klasycznego zbioru danych "iris"

Eksploracja danych w środowisku R

by Łukasz Stafiniak

1 Pochodzenie i zawartość zbioru danych

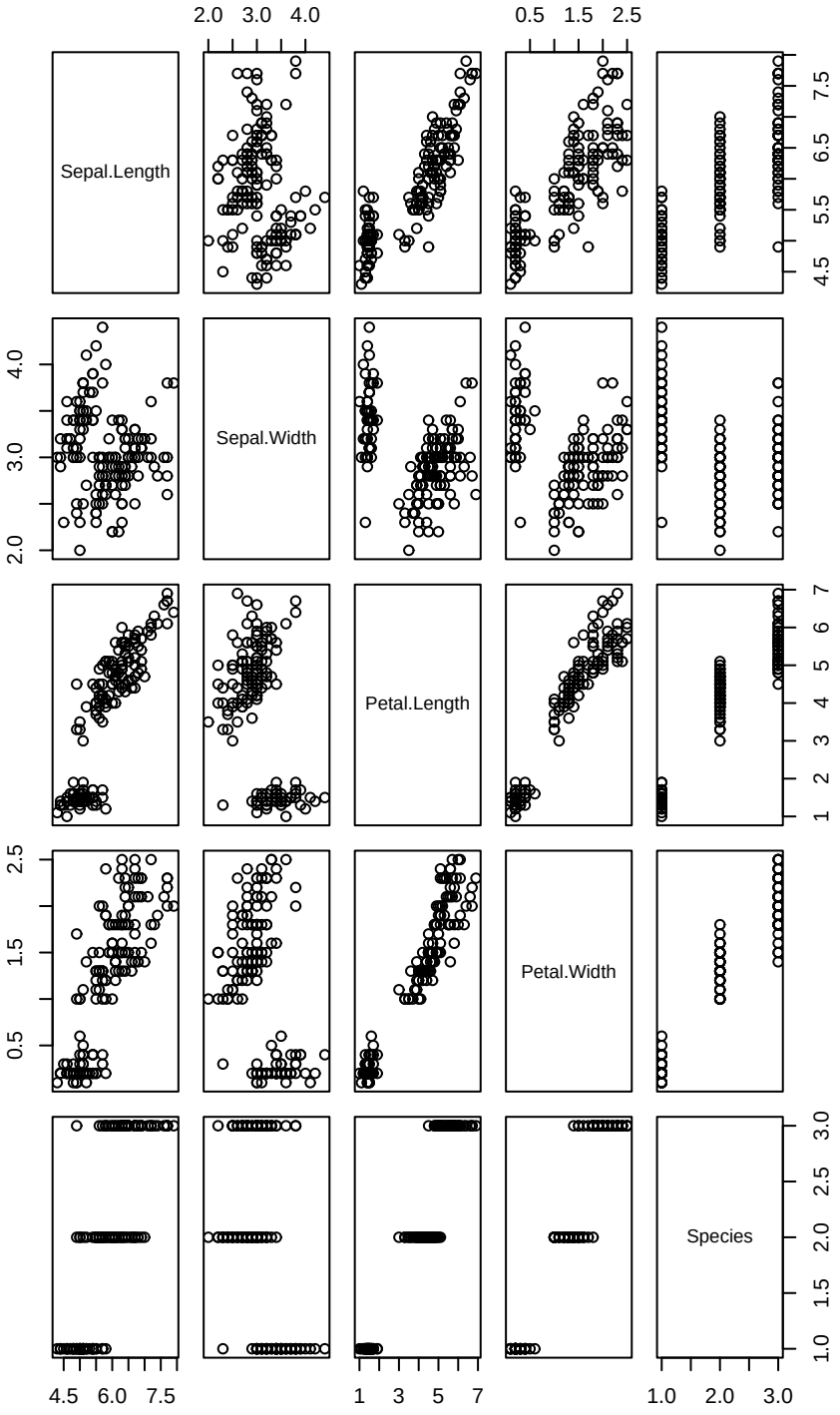
Analizujemy klasyczny zbiór danych zebranych przez E. Andersona (1935), spopularyzowany przez R. A. Fishera (1936) dotyczący gatunków irysów. Pomiary dotyczą długości i szerokości w centymetrach: petali czyli płatków kwiatu oraz sepali czyli listków kwiatu. Gatunkami są iris setosa, iris versicolor i iris virginica.

Mamy 150 obserwacji po 50 dla każdego gatunku. Zmienne w ramce zbioru danych nazywają się 'Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width', oraz 'Species'.

2 Podstawowe charakterystyki zmiennych

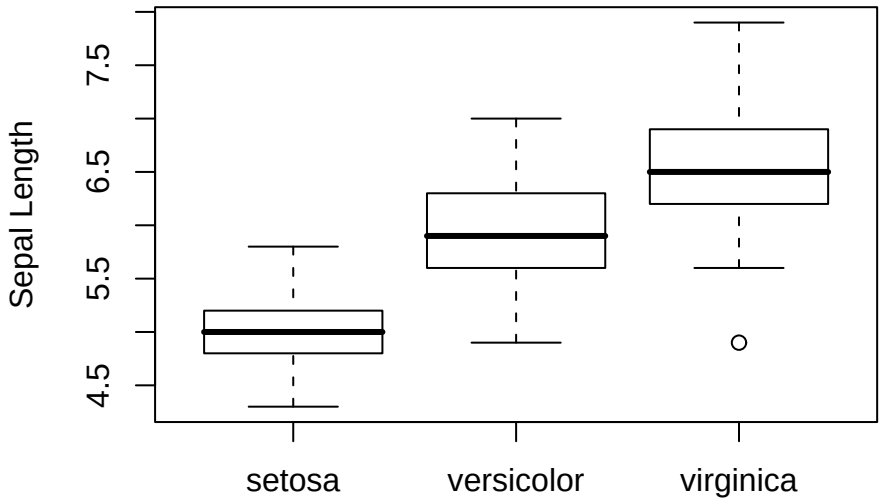
2.1 Porównanie zmiennych pod względem dyskryminacji gatunków

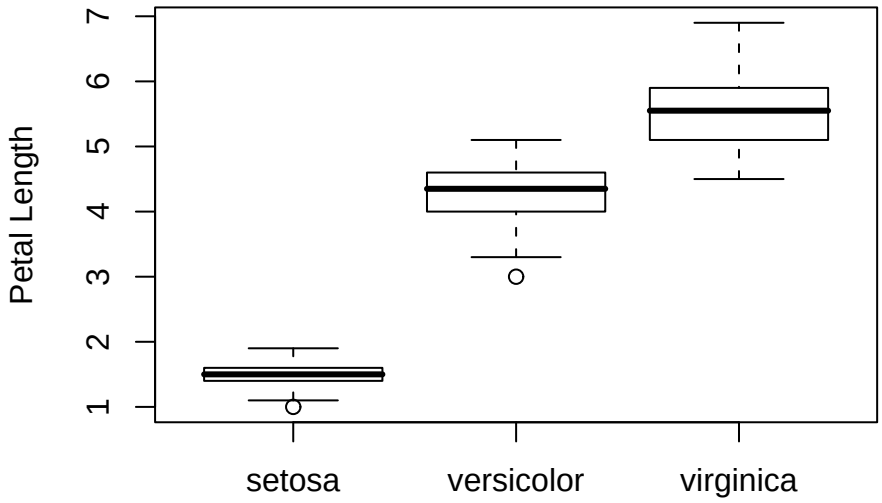
Przyjrzyjmy się obserwacjom na wykresach rozrzutu:



Wyraźnie rozdzielone są dwie grupy obserwacji, z kolumny “Species” odczytujemy, że mniejsza grupa odpowiada gatunkowi setosa, a większa gatunkom versicolor i virginica.

Zobaczmy na wykresie pudełka-wąsy, jak własności petali i sepali rozróżniają gatunki.

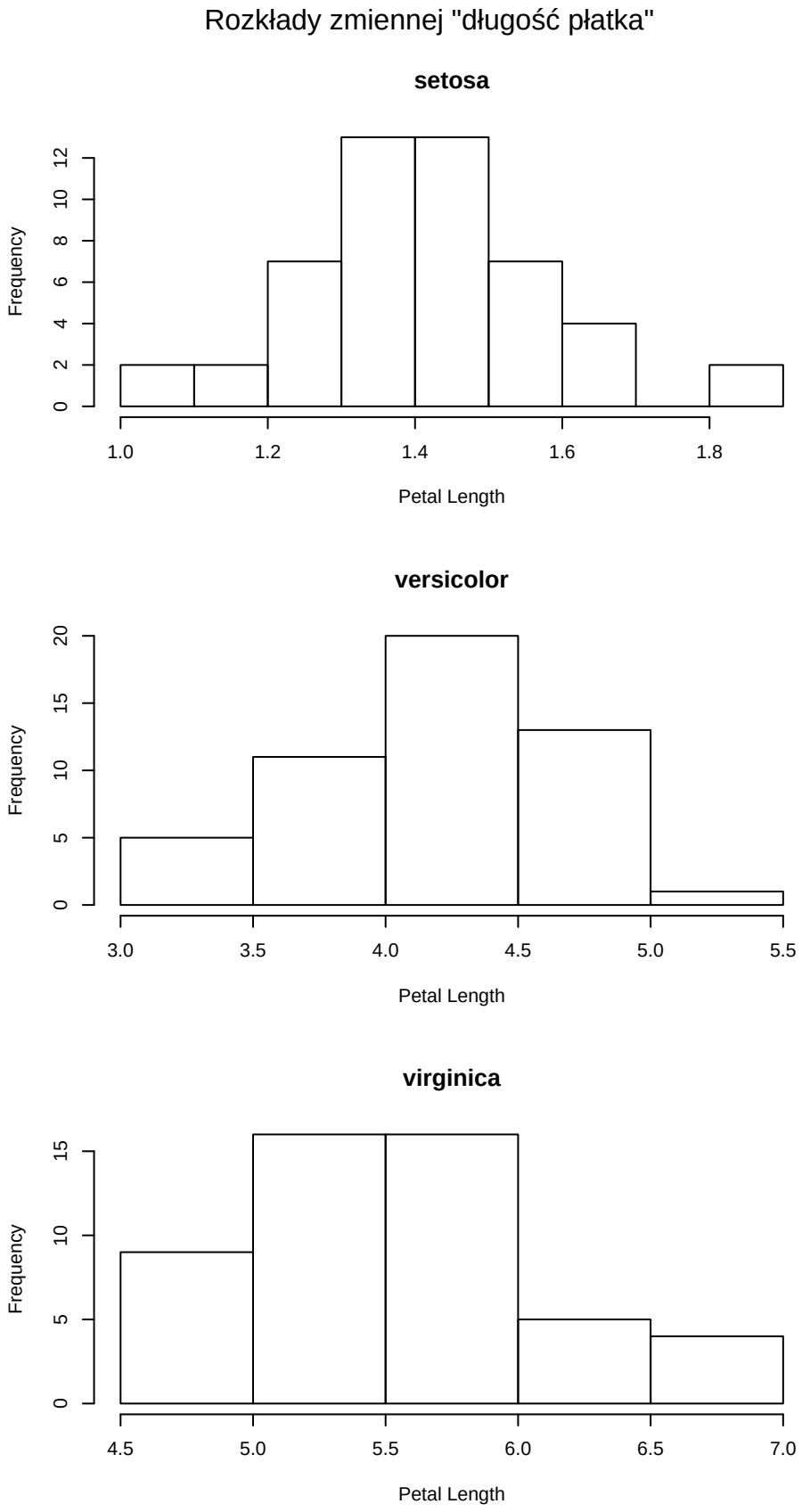




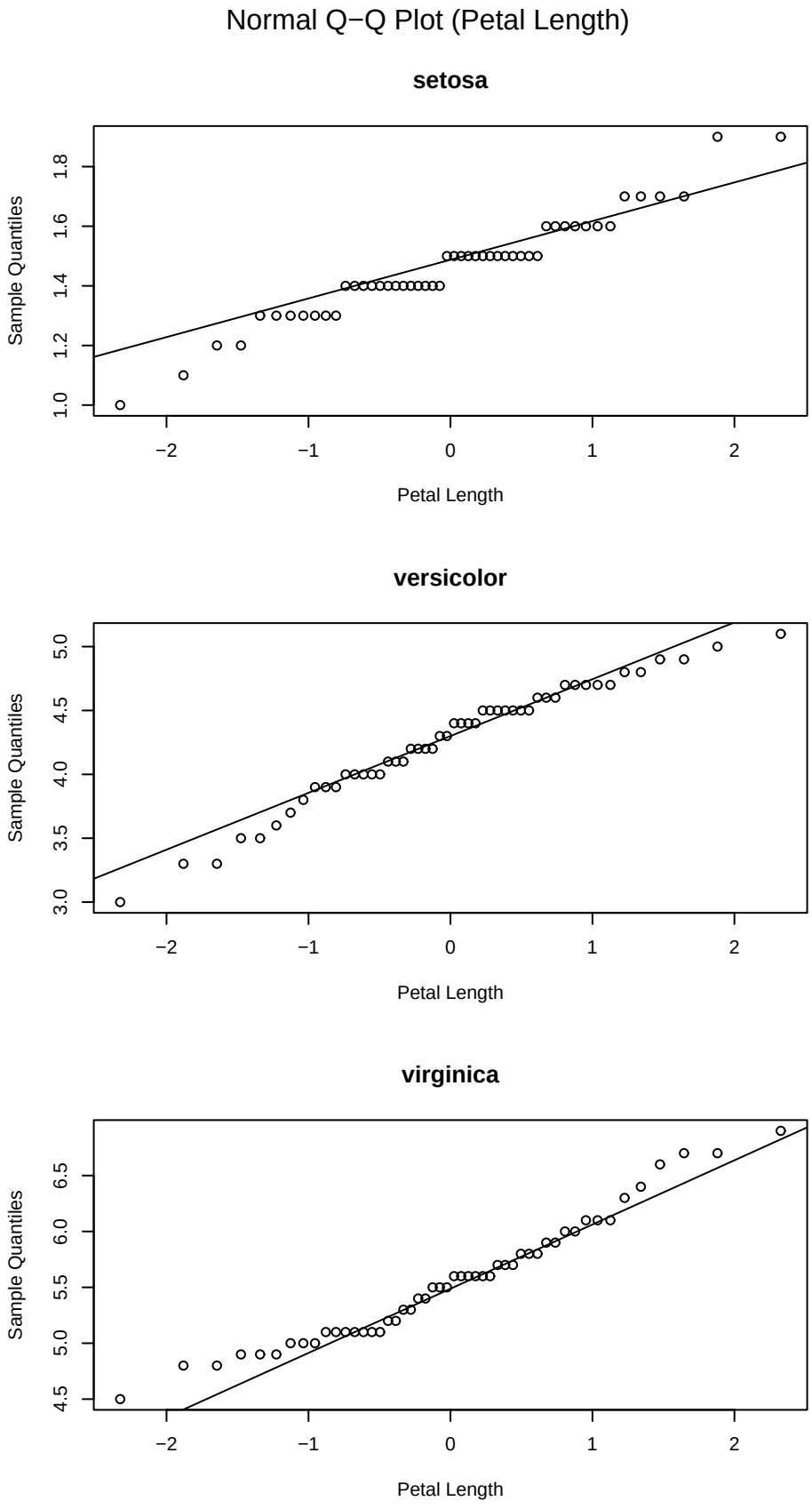
Widzimy, że długość płatków (petali) lepiej różnicuje gatunki kwiatów, w szczególności pozwala dokładnie oddzielić gatunek setosa, a versicolor i virginica pozwala rozróżnić z błędem mniejszym niż 1/4 przypadków.

2.2 Rozkłady i zależności zmiennych

Przyjrzyjmy się rozkładowi zmiennej "długość płatka". Najmniejszy rozrzut ma ona dla gatunku setosa (odchylenie standardowe = 0.17cm, rozstęp międzykwartylowy = 0.18cm), a największy dla gatunku virginica (odchylenie standardowe = 0.55cm, rozstęp międzykwartylowy = 0.78cm), jednak rozrzut względny wszystkie gatunki mają bardzo zbliżony (odchylenie standardowe wynosi ok. 10% średniej).



Histogramy pokazują rozkład dzwonowy, porównajmy z rozkładem normalnym (Q-Q plot):


Normal Q−Q Plot (Petal Length)
setosa
Sample Quantiles
Petal Length
versicolor
Sample Quantiles
Petal Length
virginica
Sample Quantiles
Petal Length

Długość płatków ma rozkład bardzo bliski normalnemu ("schody" na wykresach oznaczają małą rozdzielczość pomiaru), versicolor ma odrobinę cięższy lewy ogon (punkty pod prostą), a virginica prawy (punkty nad prostą). (Oznacza to, że obserwacje czasami będą mniejsze, odpowiednio większe, niż "byśmy się spodziewali".)

Zbadajmy teraz zależności zmiennych od siebie – macierz korelacji:

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Najbardziej zależne są od siebie parametry płatków, następnie długości listków i płatków. Szerokość listków jest mało związana z pozostałymi parametrami. Największe zdolności predykcji pozostałych zmiennych ma długość płatków (Petal.Length), wyznaczmy więc odpowiednie formuły przy pomocy regresji liniowej:

```
Petal.Width = -0.3631 + 0.4158 * Petal.Length
Sepal.Length = 4.3066 + 0.4089 * Petal.Length
Sepal.Width = 3.4549 - 0.1058 * Petal.Length (?)
```

Dla przykładu, listki (na ogół) "startują" od długości 4.3cm, i potem rozciągają się jeszcze na 0.4 długości płatka.

Warto teraz jeszcze raz spojrzeć na wykresy rozrzutu. Z odpowiedniego wykresu widać, że kuszący wniosek "im płatki dłuższe i szersze, tym listki węższe", jest błędny. Analizę potrzebujemy przeprowadzić dla poszczególnych gatunków osobno. (SW–Sepal.Width, PL–Petal.Length)

	setosa	versicolor	virginica
korelacja SW $\sim$PL	0.18	0.56	0.40
model	$\mathrm{SW} = 2.86 + 0.38\,\mathrm{PL}$	$\mathrm{SW} = 1.18 + 0.37\,\mathrm{PL}$	$\mathrm{SW} = 1.67 + 0.23\,\mathrm{PL}$