Przygotowanie danych do wizualizacji

A. M. Machno

Dane do wizualizacji.

W związku ze specjalną strukturą tworzenia wykresów w pakiecie `ggplot`, tzn. korzystanie z koncepcji gramatyki grafiki, dla każdego rodzaju wykresu może być potrzebna inna reprezentacja danych.

Przede wszystkim, funkcja `ggplot` w inny sposób traktuje zmienne różnego typu. Najczęstszym przypadkiem możliwych róznic w reprezentacji są zmienne liczbowe `integer` oraz zmienne kategoryczne.

Każda estetyka jest skalą związaną ze zmienną w danych. Jeżeli wielkość nas interesująca jest rozbita na dwie kolumny lub więcej niż jeden element występuje w jednej kolumnie, dane należy zmodyfikować.

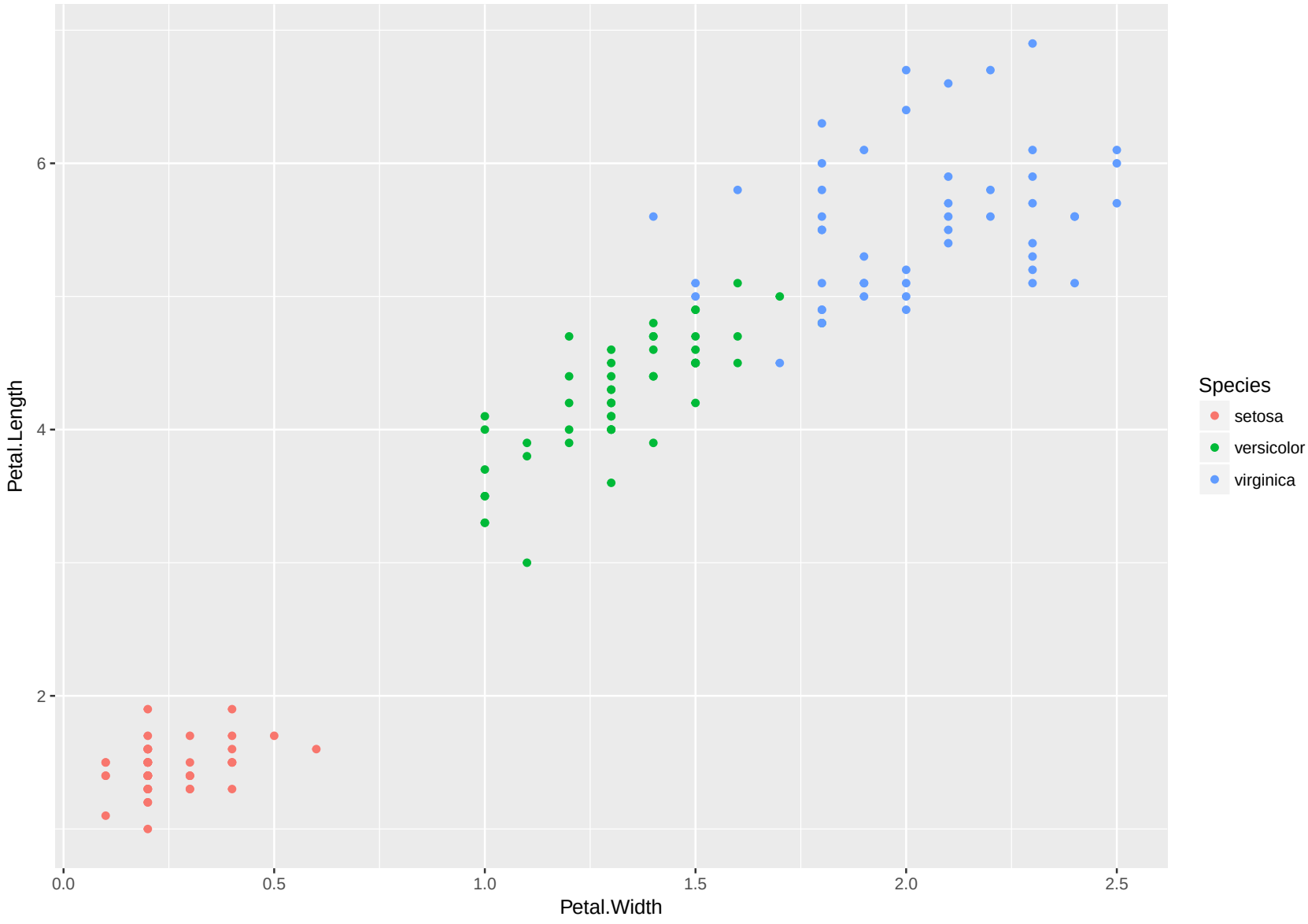
Dane iris

Dane iris są to dane 150 kwiatków trzech gatunków. Mierzonymi wielkościami są długości oraz szerokości płatków oraz działek kielicha. W kilku kolejenych slajdach zaprezentowane zostaną konkretne wykresy, a potem odtworzone poprzez modyfikację danych oraz wywołanie funkcji ggplot()

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.
##  $ Species     : Factor w/ 3 levels "setosa","versicolor
```

Wykres I


Petal.Length
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
Petal.Width
Species
setosa
versicolor
virginica

Dane do wykresu I

Oryginalne dane `iris` sa odpowiednie do tego wykresu, zmienna na osi x oraz osi y są w kolumnach, oraz zmienna określająca koloró również.

```
ggplot(iris, aes(x = Petal.Width, y = Petal.Length, col = S
  geom_point()
```

Czy te dane nadały by się aby stworzyć obok siebie wykresy długości od szerekości płatków oraz działek kielicha?

Wykres II


Length
Width
8
6
4
2
0
Value
setosa
versicolor
virginica
setosa
versicolor
virginica
Species
Part
Petal
Sepal

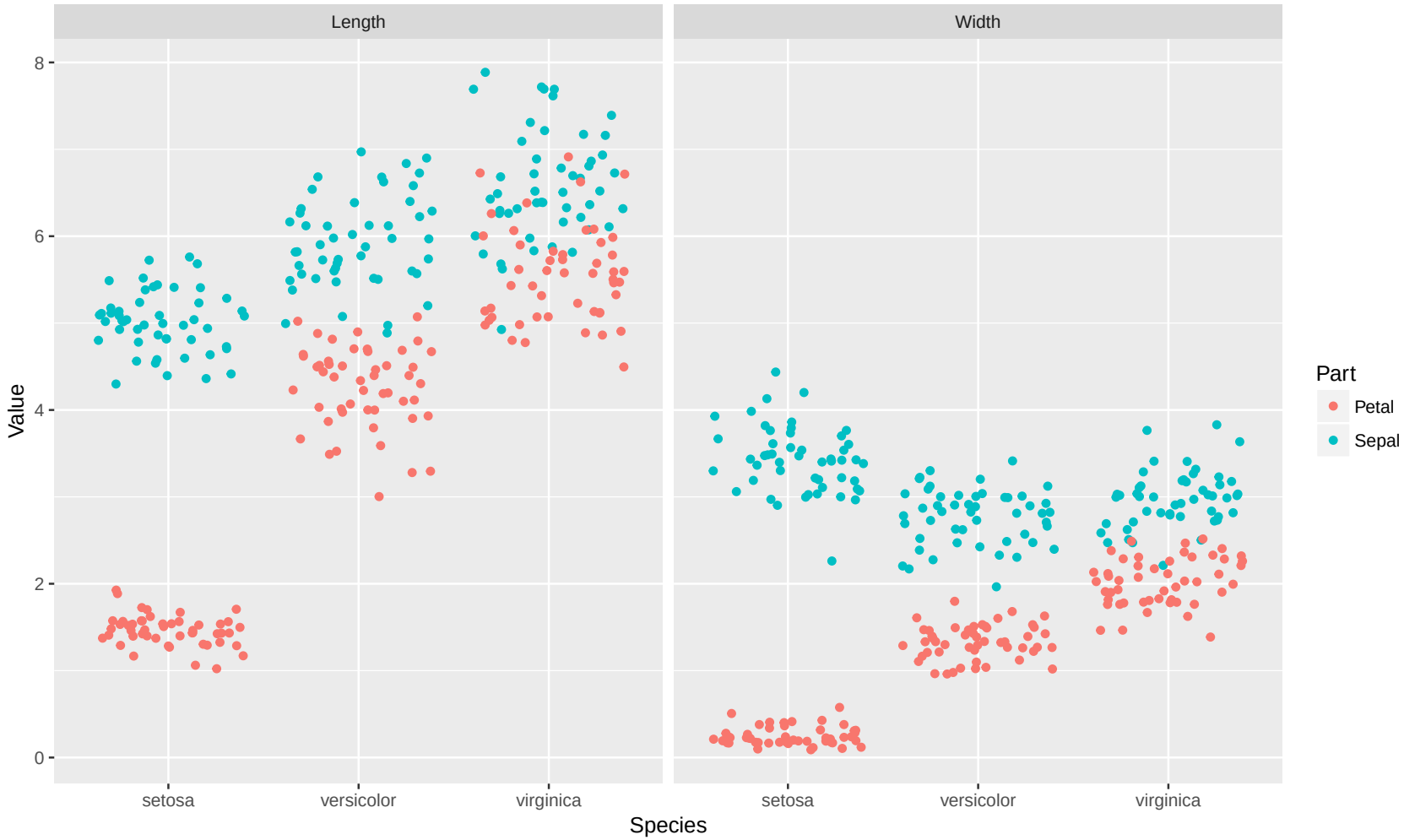
Dane do wykresu II

```
## 'data.frame':    600 obs. of  4 variables:
##  $ Species: Factor w/ 3 levels "setosa","versicolor",..
##  $ Part   : chr  "Sepal" "Sepal" "Sepal" "Sepal" ...
##  $ Measure: chr  "Length" "Length" "Length" "Length" ..
##  $ Value  : num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ..
```

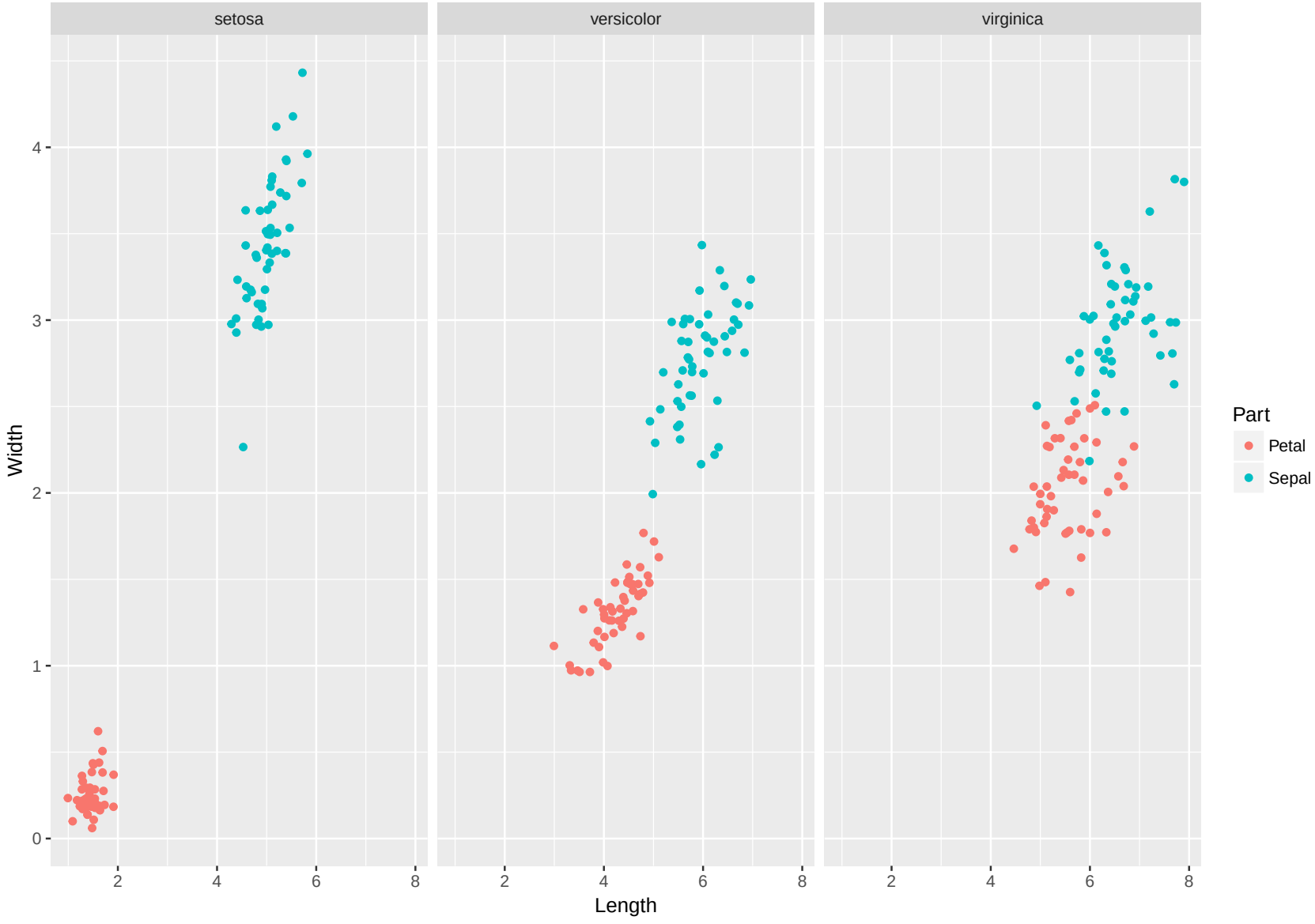
```
iris.tidy <- iris %>%
  gather(key, Value, - Species) %>%
  separate(key, c("Part", "Measure"), "\\.")
```

Stworzenie wykresu II

```
ggplot(iris.tidy, aes(x = Species, y = Value, col = Part))
  geom_jitter() +
  facet_grid(. ~ Measure)
```



Wykres III


setosa
versicolor
virginica
Width
Length
0
1
2
3
4
2
4
6
8
Part
Petal
Sepal

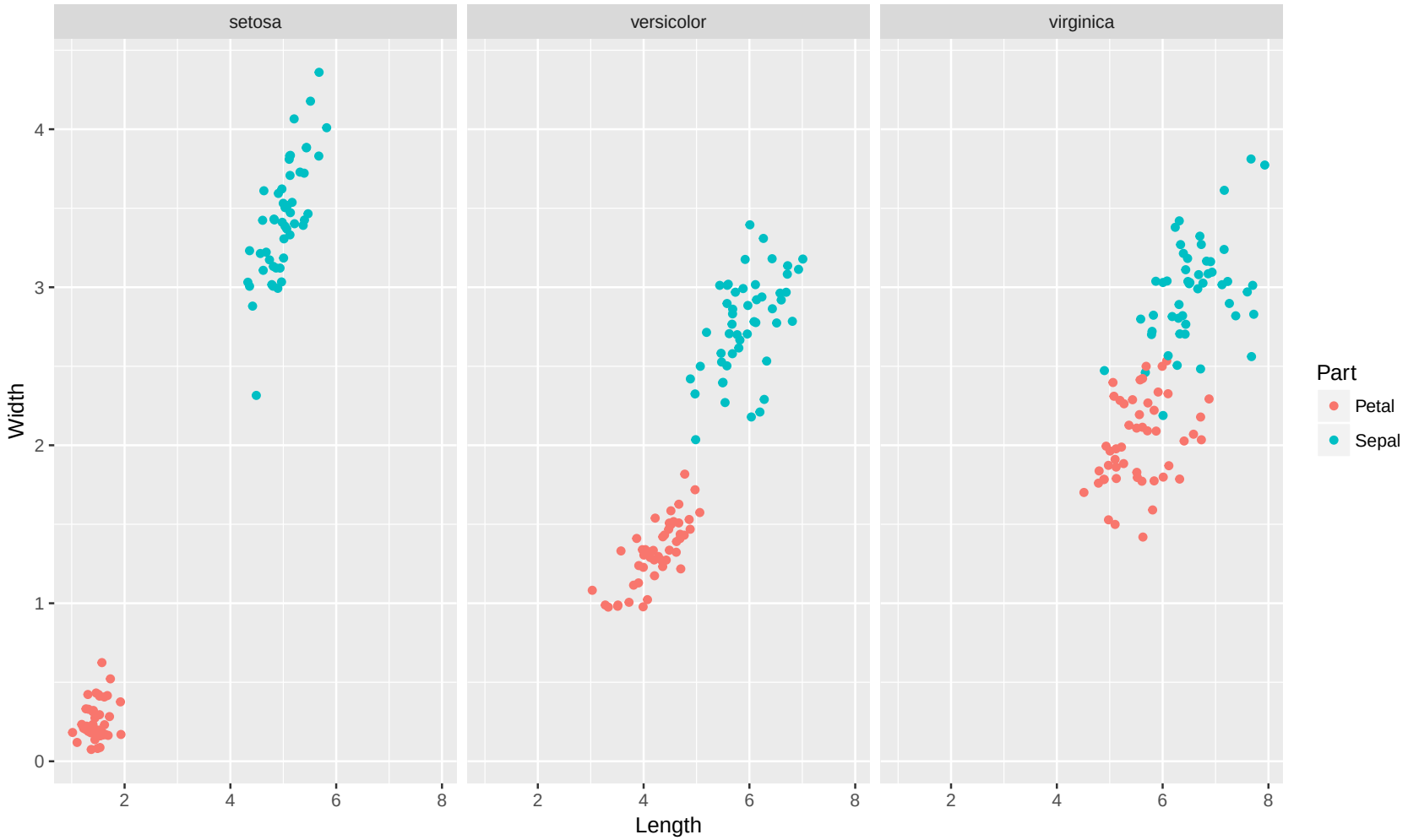
Dane do wykresu III

```
## 'data.frame':    300 obs. of  5 variables:
##  $ Species: Factor w/ 3 levels "setosa","versicolor",..
##  $ Flower : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ Part   : chr  "Petal" "Sepal" "Petal" "Sepal" ...
##  $ Length : num  1.4 5.1 1.4 4.9 1.3 4.7 1.5 4.6 1.4 5
##  $ Width  : num  0.2 3.5 0.2 3 0.2 3.2 0.2 3.1 0.2 3.6
```

```r
iris$Flower <- 1:nrow(iris)
iris.wide <- iris %>%
  gather(key, value, -Species, -Flower) %>%
  separate(key, c("Part", "Measure"), "\\.") %>%
  spread(Measure, value)
```

Stworzenie wykresu III

```
ggplot(iris.wide, aes(x = Length, y = Width, color = Part))
  geom_jitter() +
  facet_grid(. ~ Species)
```



Ćwiczenia

1. Dla danych `hflights` stworzyć wykres opóźnienia (oba typy: `DepDelay` oaz `ArrDelay`) od czasu podróży. Niech kształt punktów będzie zależny d rodzaju opóźnienia (wylot i przylot), a wielkość punktów od odległości lotu. Dotakowo niech obramowanie punktów będzie kolorowane w zależności od przewoźnika.
2. Co można zmodyfikować w wykresie 1. aby był bardziej czytelny?
3. Dla danych `hflights` stworzyć wykres czasu jazdy taksówką (oba typy: `TaxiIn` oaz `TaxiOut`) od daty. Niech kształt punktów określa typ jazdy taksówką, a wielkość punków odległość lotu. Dodatkowo storzyc krzywe opisujące zależność w grupach przewoźników.