*Politechnika Krakowska © Statystyka i Probabilistyka*

**PROJEKT NR 1 Sprawozdanie**

1. **Wizualizacja Danych. Miary położenia i miary rozproszenia.**
   1. **Opisz dane ‘painters’: “The Painter’s Data of the Piles”**

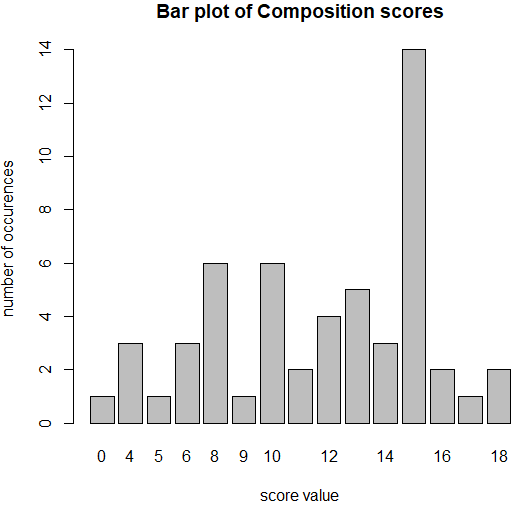
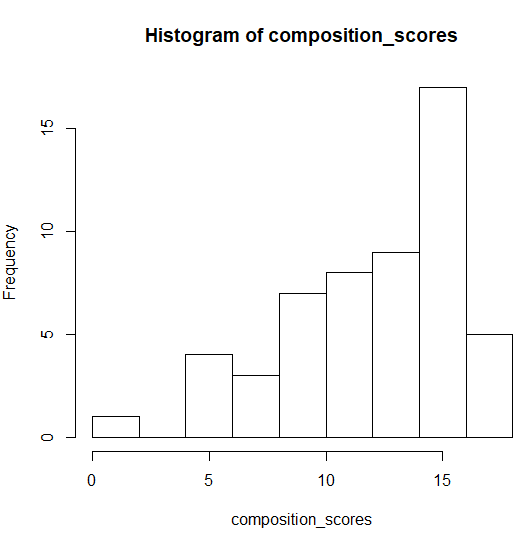
Zadany zbiór danych składa się z informacji stanowiących subiektywną ocenę warsztatu malarzy klasycznych wystawioną przez żyjącego w 18 w. krytyka sztuki de Pilesa’a. Każdy z artystów został oceniony w 20 stopniowej skali (0-20) w czterech różnych kategoriach: Composition (kompozycja), Drawing (rysunek), Colour (chromatyka) i Expression (ekspresja). Dodatkowo każdemu z malarzy przypisano nurt w sztuce (School), w myśl którego tworzył, a który można również utożsamić z okresem, w którym żył i malował.

Struktura rekordu w bazie danych:

* Nazwisko malarza (tekst)
* Punkty za kompozycję (liczba całkowita ∈ {0,1..20})
* Punkty za rysunek (liczba całkowita ∈ {0,1..20})
* Punkty za chromatykę (liczba całkowita ∈ {0,1..20})
* Punkty za ekspresję (liczba całkowita ∈ {0,1..20})
* Szkoła/nurt (pojedynczy literał), gdzie:
  + A : Renaissance - Renesans
  + B : Mannerist - Manieryzm (późny renesans)
  + C : Seicento – kontrreformacja oraz początki baroku
  + D : Venetian – styl wenecki
  + E : Lombard – styl lombardzki
  + F : Sixteenth Century – XVI w.
  + G : Seventeenth Century XVII w.
  + H : French – styl francuski
  1. **Pokaż szereg rozdzielczy dla ‘composition scores’ w ‘painters’.**



* 1. **Pokaż histogram danych kolumnowych ‘composition scores’ w ‘painters’.**

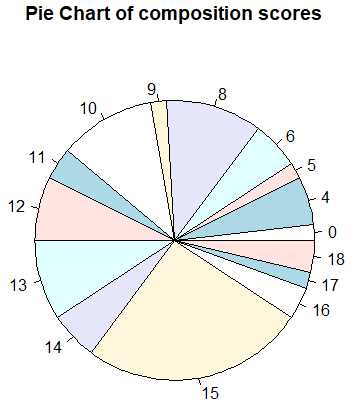


Rysunek 1. Bar plot Rysunek 2. Histogram

* 1. **Opisz różnicę pomiędzy histogramem, a bar chart.**

Histogram pokazuje częstość występowania wartości pomiarów w wyznaczonym przedziale wartości, natomiast wykres słupkowy obrazuje liczbę wystąpień danej wartości pomiaru w całej populacji (zbiorze).

* 1. **Pokaż pie chart danych ‘composition scores’.**

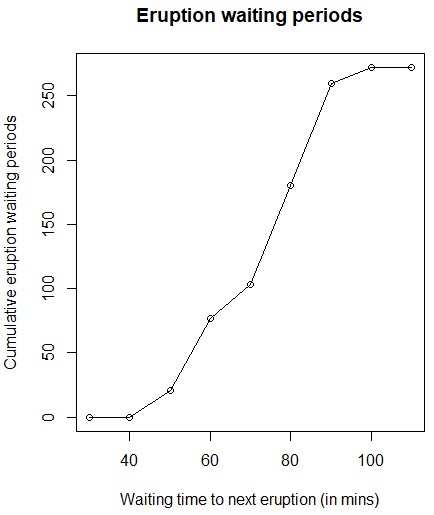


Rysunek 3. Pie chart

* 1. **Pokaż dystrybuantę danych ‘eruption waiting periods’ z danych ‘faithful’.**



* 1. **Pokaż graficznie dystrybuantę danych kolumnowych ‘eruption waiting periods’ in ‘faithful’.**



Rysunek 4. Cumulative Frequency Graph

* 1. **Dla wybranej próbki wylicz miary położenia oraz rozproszenia.**

Miary zostały wyliczone dla zbioru ‘faithful’ dla wartości długości trwania erupcji gejzerów (eruptions).

> erupt.mean

[1] 3.487783

> erupt.median

[1] 4

> erupt.mode

[1] 1.867

Miary położenia centrum

> erupt.min

[1] 1.6

> erupt.max

[1] 5.1

> erupt.range

[1] 3.5

> erupt.variance

[1] 1.302728

> erupt.stadard\_dev

[1] 1.141371

> erupt.quantile\_1

25%

2.16275

> erupt.quantile\_3

75%

4.45425

> erupt.iqr

75%

2.2915

> erupt.outlier\_left

25%

-1.2745

> erupt.outlier\_right

75%

1.017

> erupt.summary

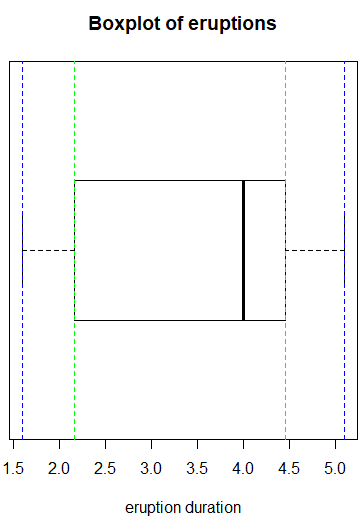
Min. 1st Qu. Median Mean 3rd Qu. Max.

1.600 2.163 4.000 3.488 4.454 5.100

Miary rozproszenia

* 1. **Dla wybranej próbki pokaż box plot i skomentuj czy istnieją obserwacje odstające.**

Obserwacje odstające nie występują w wybranej próbce.



Rysunek 7. Boxplot, blue – min, max c [Q1-1.5IQR, Q2+1.5IQR]

1. **Standaryzacja rozkładu normalnego**
   1. **Zdefiniuj z-scores oraz opisz do czego służy.**

z-scores jest miarą określającą odległość wartości danego pomiaru x względem innego pomiaru (najczęściej dla średniej) liczoną w liczbie odchyleń standardowych.

Dzięki z-scores możliwe jest wychwytywanie wartości znacznie wyróżniających się w danym zbiorze, przyjmuje się, iż jeśli z-scores jest pomiędzy -2s, 2s jest to normalnym przedziałem dla wartości w zbiorze, natomiast gdy wartość leży o więcej niż 3 odchylenia standardowe od średniej, wtedy mówimy o tzw. wartości odstającej.

* 1. **Wyniki egzaminu SAT Math dla studentów mają średnią 543 i standardowe odchylenie 110.** 
     1. **Wylicz z-scores dla: 300, 400, 500, 600, 700, 800**

> mean

[1] 543

> sd

[1] 110

> scores$scores

[1] 300 400 500 600 700 800

> results.difference

[1] -243 -143 -43 57 157 257

> results.zscores

[1] -2.2090909 -1.3000000 -0.3909091 0.5181818 1.4272727 2.33636

* + 1. **Oblicz wartości SAT Math dla poszczególnych z-scores: -2.09, -1.3, -0.39, 0.52, 1.43, 2.34**

> mean

[1] 543

> sd

[1] 110

> results.scores = scores$zscores\*sd+mean

> results.scores

[1] 313.1 400.0 500.1 600.2 700.3 800.4

Jak łatwo zauważyć wyniki różnią się, jest to skutek przyjętych zaokrągleń dla wartości z-scores, dla porównania wyniki uzyskane na podstawie obliczonych bez zaokrągleń z-scores w poprzednim podpunkcie.

> mean

[1] 543

> sd

[1] 110

> results.scores = results.zscores\*sd+mean

> results.scores

[1] 300 400 500 600 700 800

* + 1. **Porównaj inputs i outputs z poprzednich punktów**

Wyniki różnią się, w związku z przyjętymi w podpunkcie ii. zakrągleniami, więcej w podpukcjcie ii.

* 1. **Dla wybranej próbki:**

Wybrana próbka : painters$expression

> express.values

[1] 9 8 10 6 2 6 6 13 0 0 4

* + 1. **Wylicz wartość średnią.**

> express.mean

[1] 5.818182

* + 1. **Wylicz standardowe odchylenie.**

> express.sd

[1] 4.118694

* + 1. **Wylicz Z-score, korzystając ze wzoru oraz funkcji scale()**

> express.zscore

[1] 0.77253094 0.52973550 1.01532638 0.04414463 -0.92703713 0.04414463 0.04414463

[8] 1.74371269 -1.41262800 -1.41262800 -0.44144625

> express.scale\_sd

[,1]

[1,] 0.77253094

[2,] 0.52973550

[3,] 1.01532638

[4,] 0.04414463

[5,] -0.92703713

[6,] 0.04414463

[7,] 0.04414463

[8,] 1.74371269

[9,] -1.41262800

[10,] -1.41262800

[11,] -0.44144625

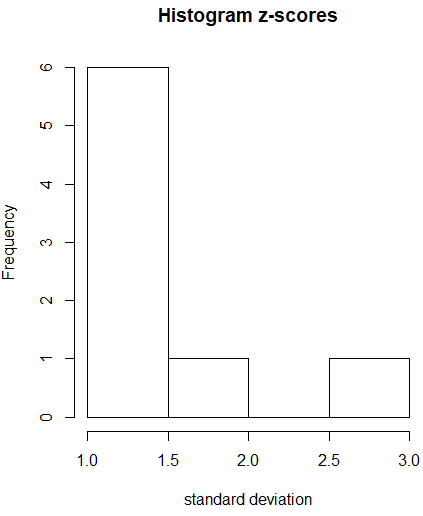
attr(,"scaled:center")

[1] 5.818182

attr(,"scaled:scale")

[1] 4.118694

* + 1. **Pokaż z-score graficznie. Opisz oś poziomą oraz pionową. Zinterpretuj wyniki.**



Rysunek 8. Z-score histogram

Na histogramie zobrazowano rozkład **wartości bezwzględnych** z-scores dla wybranej próbki ze zbioru painters$expression, jak można zauważyć żadna z wartości w próbce nie „odstaje” od średniej o więcej niż 3 odchylenia standardowe, co więcej jedynie jedna wartość odstaje o więcej, niż 2 odchylenia, co stanowi o stosunkowo spójnych danych w zbiorze (oceny za ekspresje w wybranej próbce, nie różnią się od siebie mocno).

* + 1. **Oblicz średnią oraz standardowe odchylenia otrzymanych z-scores.**

> express.mean

[1] 5.818182

> express.zscore\_sd

[1] 1

* + 1. **Oblicz min i max z-score, co oznaczają te wartości?**

> ekspress.zscore\_min

[1] -1.412628

> ekspress.zscore\_max

[1] 1.743713

Wartość najmniejsza oraz najwieksza dla z-score pozwala określić, czy w zbiorze znajdują się wartości odbiegające od średniej, dla wybranej próbki dane są „typowe” i mieszczą się w przedziale [-2s, 2s].

* + 1. **Dla wybranych 3-ech z powyższych wartości użyj funkcji pnorm, zinterpretuj wynik.**

> pnorm(express.zscore[0:3])

[1] 0.7801000 0.7018523 0.8450249

Otrzymane wyniki są wartością dystrybuanty (percentyl określający ile procent pomiarów z próbie/populacji jest <= od wybranej wartości) w wybranym punkcie.

* 1. **Podaj definicje:**
     1. **Percentyl** – wskazuje położenie pomiaru w całej próbie, określa procentowo ile pomiarów ze zbioru jest mniejszych (i analogicznie ile jest większych) od wybranego pomiaru.
     2. **Kwartyl** – Percentyl wskazujący położenie punktu, dla którego 25% z pomiarów jest mniejszych (Q1) lub punkt, dla którego 75% pomiarów jest mniejszych (Q2).
     3. **Rozstęp kwartylowy** – jest to różnica Q3-Q1 wyznaczająca przedział wartości, w którym powinno się mieścić 50% pomiarów z badanego zbioru.
  2. **Używając dowolnych danych ilościowych, wylicz:**

Wybrane dane : painters$colour

> colour.values

[1] 0 4 4 4 4 5 6 6 6 6 6 6 7 7 8 8 8 8 8 9 9 9 10 10 10 10 10 10 10 12 12 12 13 13 14 14 14 15 15 16 16 16 16 16 16

[46] 16 16 17 17 17 17 17 18 18

* + 1. **Kwartyle**

> colour.Q1

25%

7.25

> colour.Q3

75%

16

* + 1. **Percentyle: .32, .48, .86**

> colour.P32

32%

8

> colour.P48

48%

10

> colour.P86

86%

16

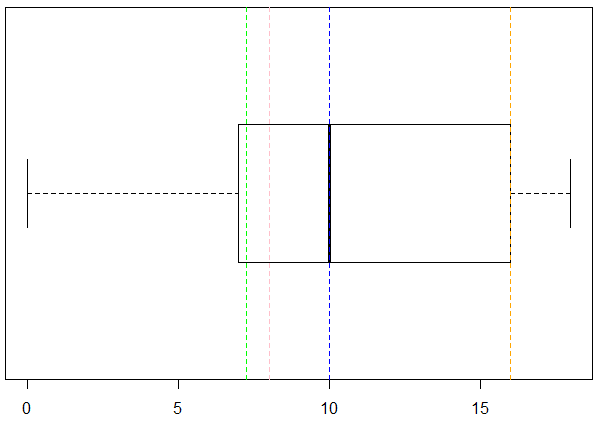
* + 1. **Rozstęp ćwiartkowy**

> colour.Q3-colour.Q1

75%

8.75

W celu lepszego zobrazowania uzyskanych wyników, utworzony został boxplot:



Rysunek 9. Boxplot dla danych colour zbioru painters, zielony – Q1, Q3 (pokrywa się z P86), różowy P32, niebieski P48 – niemal pokrywa się ze medianą, pomarańczowy P86. Wszystkie 3 percentyle mieszczą się w rozstępie ćwiartkowym.

Kod źródłowy projektu znajduje się na zdalnym repozytorium (github) :

<https://github.com/daterka/statistics>

lub bezpośrednio pod linkiem:

[kod źródłowy](https://github.com/daterka/statistics/blob/master/%5BSiP%5D%5Blab%5D%5BProjekt_01%5D%5BDaniel_Terkala%5D%5B121122%5D%5BData_Science%5D.R)