Zbigniew Nowacki 234102 234102@edu.p.lodz.pl
Bartosz Jurczewski 234067 234067@edu.p.lodz.pl

Zadanie 1: Analiza statystyczna

1. Cel

Celem zadania było przeprowadzenie analizy statystycznej dla wybranych zbiorach danych.

2. Wprowadzenie

Plik *main3.py* zawiera skrypt realizujący następujące wymagania:
1. Dla poszczególnych atrybutów wyznaczyć medianę, minimum i maximum dla cech ilościowych oraz dominantę dla cech jakościowych.
2. Narysować histogramy dla dwóch cech ilościowych najbardziej ze sobą skorelowanych.
3. Zadbać o czytelność rezultatów oraz staranny i atrakcyjny wygląd histogramów.

Plik *main4.py* zawiera skrypt realizujący następujące wymagania:
1. Zbadać jedną z następujących hipotez:
   a) Dla danych Births zbadać hipotezę, że dzienna średnia liczba urodzeń dzieci wynosi: 10000.
   b) Dla danych manaus zbadać hipotezę, że średnia wysokość rzeki w manaus jest na wysokości punktu arbitralnego(wynosi 0).
   c) Dla danych quakes zbadać hipotezę, że średnia głębokość występowania trzęsienia ziemi wynosi 300 metrów.
2. Zwizualizować rozkłady na histogramie.
3. Zaznaczyć na wykresie punkt dotyczący badanej hipotezy.

3. Opis implementacji

Zadanie zostało zrealizowane przy użyciu języka **Python** w wersji 3.7, z wykorzystaniem bibliotek: *pandas*, *matplotlib* i *scipy*.

4. Wyniki

4.1. Cześć 1 - ocena dostateczna

Do spełnienia wymagań na ocenę dostateczną wybraliśmy zbiór *iris.data*. *Sepal length* to długość działki kielicha, a *petal width* to szerokość płatku irisa.

Wyniki dla poszczególnych atrybutów (cech ilościowych):

Mediana	
sepal length	5.80
sepal width	3.00
petal length	4.35
petal width	1.30

Tabela 1. Mediany dla poszczególnych atrybutów ze zbioru *iris.data*

Minimum	
sepal length	4.3
sepal width	2.0
petal length	1.0
petal width	0.1

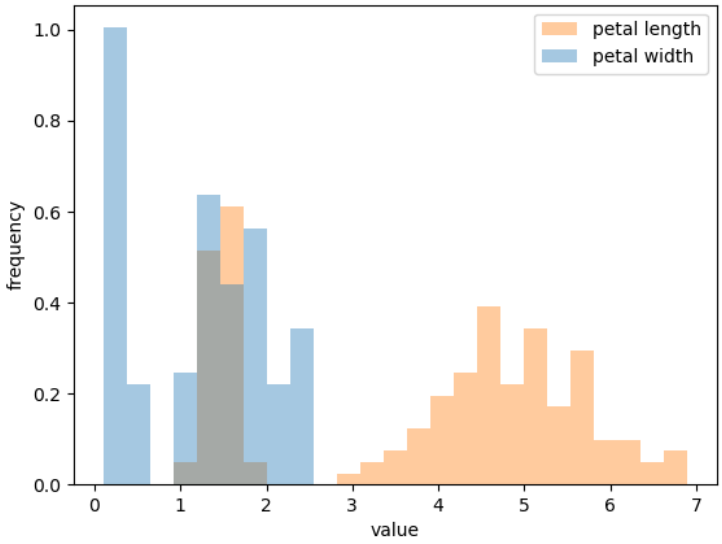
Tabela 2. Minima dla poszczególnych atrybutów ze zbioru *iris.data*

Maksimum	
sepal length	7.9
sepal width	4.4
petal length	6.9
petal width	2.5

Tabela 3. Maksima dla poszczególnych atrybutów ze zbioru

Dominantą dla cechy jakościowej czyli gatunkiem są *Iris-setosa*, *Iris-versicolor* i *Iris-virginica*. Wynika to z specyfiki danych - w pliku znajduję się po 50 irysów z każdego gatunku.

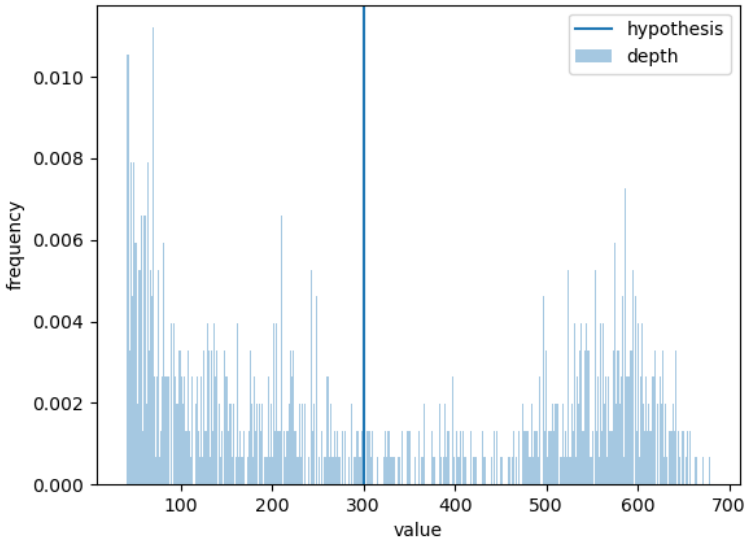
Dwie cechy ilościowe najbardziej skorelowane ze sobą to *petal length* i *petal width*. Ich relacja została pokazana na rysunku 1.



Rysunek 1. Histogram dla danych *iris*

4.2. Cześć 2 - ocena dobra

Zbiór danych: *quakes*.
Hipoteza: "średnia głębokość występowania trzęsienia ziemi wynosi 300 metrów".



Rysunek 2. Histogram dla danych *quakes*

Próbka danych na której badano hipotezę nie pochodzi z rozkładu normalnego. Wnioski wyciągnięte na podstawie takiej próbki należy odrzucić.

5. Wnioski

— Korzystanie z gotowych bibliotek do statystki znacznie przyspieszyło i ułatwiło pracę z danymi.
— Program oblicza poprawnie podstawowe funkcje statystyczne.
— Przed rozpoczęciem opracowania wyników należy upewnić się że próbka pochodzi z rozkładu normalnego.

Literatura

[1] *An Omnibus Test of Normality for Moderate and Large Size Samples*, 1971, Ralph B. D'Agostino, https://pdfs.semanticscholar.org/1493/2aec5e2128d2373d57f6dede6c3c7ed71f07.pdf

[2] *WnioskowanieStatystyczne/ Testowanie hipotez,* https://brain.fuw.edu.pl/edu/index.php/WnioskowanieStatystyczne/_Testowanie_hipotez#Przyk.C5.82ad_.28zastosowanie_r.C3.B3.C5.BCnych_test.C3.B3w_do_tych_samych_danych.29:_karma

[3] *A. Wosiak – Języki programowania w analizie danych,* https://ftims.edu.p.lodz.pl/pluginfile.php/136679/mod_resource/content/1/Python_wyklad_02_Statystyka.pdf

[4] https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-test/