Paweł Galewicz 234053
Karol Podlewski 234106

Rok akademicki 2019/20
Poniedziałek, 12:30

JĘZYKI PROGRAMOWANIA W ANALIZIE DANYCH – LABORATORIUM

Zadanie 1

Opis implementacji

Stworzone rozwiązanie to programy konsolowe, zaimplementowane w języku Python 3.7.5. W projekcie wykorzystano następujące biblioteki: NumPy, Pandas, Matplotlib, Seaborn.

Część na ocenę dostateczną

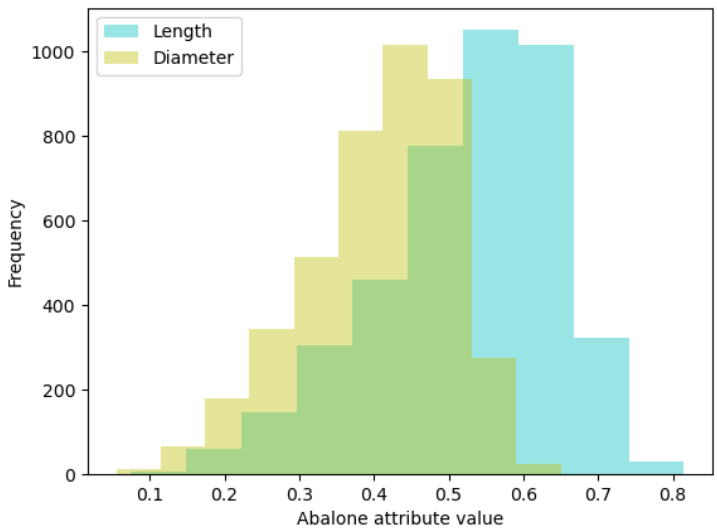
Z dostępnych zbiorów danych wybrano zbiór *abalone*. Dla cech jakościowych uzyskano następujące wyniki:

Cecha	Dominanta
Sex	M

Dla cech ilościowych uzyskano następujące wyniki:

Cecha	Mediana	Wartość minimalna	Wartość maksymalna
Length	0,545	0,075	0,815
Diameter	0,425	0,055	0,65
Height	0,14	0	1,13
Whole weight	0,7995	0,002	2,8255
Shucked weight	0,336	0,001	1,488
Viscera weight	0,171	0,0005	0,76
Shell weight	0,234	0,0015	1,005
Rings	9	1	29

Dla najbardziej skorelowanych ze sobą cech ilościowych (Length oraz Diameter) uzyskano następujący histogram:



Część na ocenę dobrą

Do tego eksperymentu wybrany został zbiór *births* dostarczający informacji na temat średniej dziennej liczbie urodzeń. Dane zebrane zostały w latach 1969 – 1988.

Hipotezę H_0, którą należało zbadać było stwierdzenie, że średnio dziennie rodzi się 10000 dzieci. Za hipotezę alternatywną przyjęto zaprzeczenie H_0.

Ze względu na specyfikę zbioru (istotne dane są ilościowe), do przetestowania hipotezy wybrano test *t-Studenta*, zaś za poziom istotności statystycznej p przyjęto 5%.

Test dał następujący rezultat:

Statystyka t	**Wartość p**
-26.616	$4.73*10^{-149}$

Otrzymana wartość p jest zdecydowanie mniejsza od dopuszczalnej granicy 0.05, więc hipoteza H_0 została odrzucona.

Następujący histogram prezentuje rozkład danych wraz z zaznaczoną hipotezą H_0: