Sprawozdanie z projektu

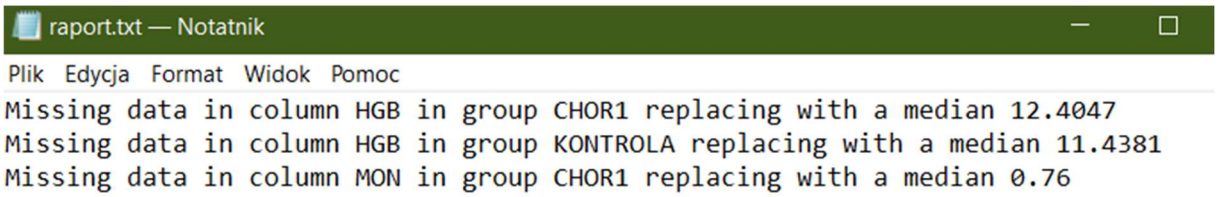
Weronika Koga

Plik README

Porusza ważne kwestie m.in jak uruchomić program, aby uzyskać prawidłowe wyniki, jak przygotować dane, co jest potrzebne do wykonania analizy i zwracane jako rezultat działania programu.

Braki danych

Funkcja Replace_blank_with_NA() odpowiada za zamianę ‘’ ‘’ występujących w pliku na NA. Następnie funkcja Remove_NA() usuwa NA w kolumnach nienumerycznych oraz zastępuje NA medianą z danej grupy w kolumnach numerycznych. Wszystkie zmiany zapisywane są do pliku raport.txt. W przypadku kolumn numerycznych - w jakiej kolumnie i grupie wykryto brak w wartości oraz ile wynosi mediana, która jest wstawiana w to miejsce. W przypadku kolumn nienumerycznych – numer usuniętego wiersza.



Raport istniejących grup

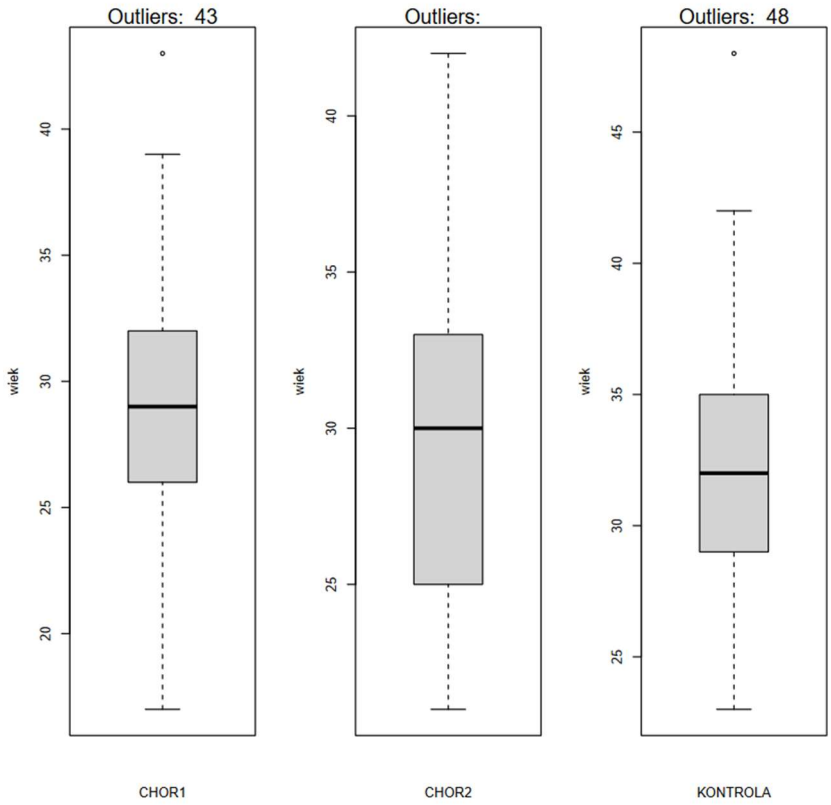
Funkcja count_groups() odpowiada za wypisanie wszystkich istniejących w pliku csv grup, wraz z ilością wierszy do niej przypisanych, do pliku raport.txt.

```
GROPUS AND THEIR SIZE_____________________

"CHOR1" 25
"CHOR2" 25
"KONTROLA" 25
```

Wartości odstające

Funkcja Outliers_detection() zapisuje do pliku raport.txt wartości odstające w każdej kolumnie numerycznej i dla każdej grupy. Tworzy również boxploty uwidaczniające wartości odstające lub ich brak zapisywany do pdf „Outliers”.
Przykład :

```
OUTLIERS_____________________


wiek CHOR1 43
wiek CHOR2 no outliners
wiek KONTROLA 48
```



Charakterystyka badanych grup

Funkcja Characteristics() dla każdej grupy w każdej kolumnie numerycznej przygotowuje raport minimalnej i maksymalnej wartości, średniej, mediany oraz 1 i 3 kwartylu.

```
CHARACTERISTICS_____________________

CHOR1
wiek
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   26.00   29.00   29.56   32.00   43.00
```

Analiza porównawcza pomiędzy grupami

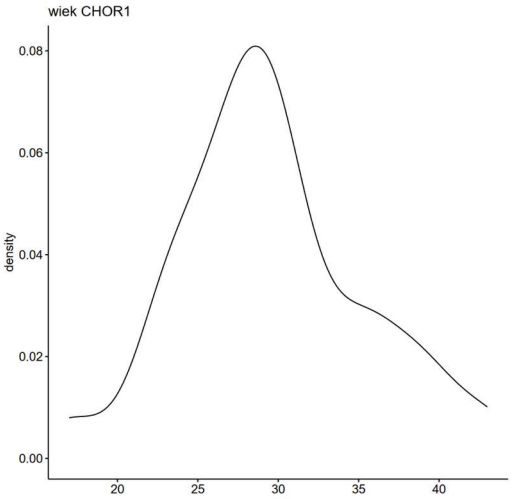
Dla każdego z testów przyjęłam wartość graniczną p.value=0.05

Homogeniczność i rozkład normalny

Funkcja Homogenity_of_variance_raport() sprawdza za pomocą testu Levene'a czy dla każdej kolumny numerycznej wariancja jest homogeniczna. Funkcja Normal_distribution_raport() przeprowadzając test Shapiro-Wilka sprawdza czy rozkład jest normalny. Wyniki tego sprawdzenia są raportowane do pliku raport.txt.
Obie te funkcje wywoływane są w funkcji Density_normal_and_homogenic_info() która dodatkowo zwraca listę z wektorami :
1.) nazw kolumn dla których istniała jakaś grupa która nie miała rozkładu normalnego
2.) nazw kolumn z homogeniczną wariancją.
Dodatkowo tworzony jest pdf „Density" z wykresami obrazującymi rozkłady w każdej grupie w kolumnie numerycznej.

Przykład:

```
 PLT
homogeneous variance
CHOR1 PLT p > 0.05 - normal distribution
CHOR2 PLT p < 0.05 - NOT normal distribution
KONTROLA PLT p < 0.05 - NOT normal distribution
```



Czy istnieją różnice pomiędzy grupami

Funkcja Statistics_test wywołuje funkcję Density_normal_and_homogenic_info() a następnie korzystając z tego co zwraca ta funkcja wywołuję funkcję Apply_test().
Apply_test() korzystając z ilości grup oraz informacji które grupy dla jakich kolumn mają rozkład normalny i które kolumny mają wariancje homogeniczną, stosuje odpowiedni test statystyczny, którego nazwa, rezultat oraz to dla jakich danych był przeprowadzony, jest zapisywany do pliku raport.txt.
Obecne testy:
-Anova_test (gdy grup jest więcej niż 2 a dane mają rozkład normalny i wariancję homogeniczną)
-Tukey_test (przeprowadzany automatycznie gdy test Anova wykaże p.value<0.05)
-Kruskal_test (gdy grup jest więcej niż 2 i warunek na Anova test nie jest spełniony)
-Dunn_test (przeprowadzany automatycznie gdy test Kruskala wykaże p.value<0.05)

-T_Student (gdy grupy są dwie a dane mają rozkład normalny i wariancję homogeniczną)
-Welch (gdy grupy są dwie a dane mają rozkład normalny i niehomogeniczną wariancję)
- Wilcoxon (gdy grupy są dwie a dane nie mają rozkładu normalnego)

Przykład:

```
 HGB
Test Kruskala 0.001 < 0.05 - there are differences between groups
Test Dunna 0.003 < 0.05 - there are differences between groups  CHOR1 - KONTROLA
Test Dunna 0.003 < 0.05 - there are differences between groups  CHOR2 - KONTROLA
```

Analiza korelacji

Funkcja Correlation_analysis() używa testu Spearmana i zapisuje do pliku raport.txt pomiędzy którymi kolumnami w obrębie jakiej grupy występuje korelacja oraz jaka jest jej siła i kierunek. Dodatkowo tworzony jest plik pdf „Correlations” z wykresami korelacji o ile takie istnieją. Regresja liniowa wskazuje kierunek korelacji a pole dookoła niej reprezentuje siłę korelacji pomiędzy danymi. Punkty dookoła reprezentują dane. Tworzony jest również plik pdf „Heatmaps” z heatmapą korelacji kolumn w obrębie każdej grupy (w funkcji Generate_heatmap()).

Przykład:

```
CORRELATION ANALYSIS_______________


 CHOR1
ERY  i  HGB :  Strong positive correlation
ERY  i  HCT :  Strong positive correlation
HGB  i  HCT :  Very strong positive correlation
HGB  i  MCHC :  Strong positive correlation
HCT  i  MCHC :  Positive correlation of medium intensity
```