Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

Artur Skowroński

Nr albumu: 411423

Statystyczna analiza danych w zadaniach

Praca zaliczeniowa

przygotowana pod kierunkiem

dr hab. Piotra Wójcika

Warszawa, styczeń 2022

Spis treści

Zadanie 1

Przetestuj hipotezę, ze dla czterech losowo wybranych województw liczba osób zaliczanych do grona swoich przyjaciół przez niepalących mężczyzn jest w każdym z nich taka sama. Uwzględnij tylko osoby, które mają co najmniej jednego przyjaciela. Losowego wyboru także dokonaj za pomocą kodu 4GL.

Rozwiązanie:

Na początku użyłem funkcji proc survey, która pozwoliła mi losowo dobrać 4 województwa, dla których analizowałem to zadanie. W moim przypadku okazały się nimi być: Dolnośląskie, Lubelskie, Śląskie oraz Zachodniopomorskie.

W przypadku tych 4 województw postanowiłem sprawdzić normalność dla zmiennych hp39 oraz hp43. Ze względu na niskie pvalue'a, w obydwu przypadkach odrzucono hipotezy 0, mówiące o normalności obu rozkładów.

Procedura UNIVARIATE
Dopasowany rozkład Normalny dla HP39 (ILE OSOB ZALICZA PAN DO GRONA SWOICH PRZYJACIOL?)

Parametry dla rozkładu Normalny		
Parametr	Symbol	Ocena
Średnia	mi	6.143629
Odch. std.	sigma	6.196044

Testy dopasowania dla rozkładu Normalny				
Testowanie	Statystyka		Wartość p	
Kołmogorow-Smirnow	D	0.210831	Pr. > D	<0.010
Cramer-von Mises	W-kwadr.	31.162523	Pr. > W-kwadr.	<0.005
Anderson-Darling	A-kwadr.	177.617763	Pr. > A-kwadr.	<0.005

Procedura UNIVARIATE
Dopasowany rozkład Normalny dla HP43 (CZY PALI PAN PAPIEROSY?)

Parametry dla rozkładu Normalny		
Parametr	Symbol	Ocena
Średnia	mi	1.75664
Odch. std.	sigma	0.429187

Testy dopasowania dla rozkładu Normalny				
Testowanie	Statystyka		Wartość p	
Kołmogorow-Smirnow	D	0.471291	Pr. > D	<0.010
Cramer-von Mises	W-kwadr.	126.776949	Pr. > W-kwadr.	<0.005
Anderson-Darling	A-kwadr.	677.552380	Pr. > A-kwadr.	<0.005

Przeszedłem więc do wykonania nieparametrycznego testu Kruskalla-Willisa. H0 dla tego test głosi, iż mediana osób zaliczanych do grona swoich przyjaciół przez niepalących mężczyzn jest w każdym z wybranych województw taka sama. Poniżej przedstawiam wyniki:

Procedura NPAR1WAY

Test Kruskala-Wallisa		
Chi-kwadrat	DF	Pr. > chi-kw.
32.8406	3	<.0001

Przy założeniu poziomu istotności na poziomie 5%, należy odrzucić hipotezę 0. W związku z tym, należy uznać, że liczba osób zaliczanych do grona przyjaciół ze względu na województwo jest różna.

Zadanie 2

Przedstaw na mapie Polski podział województw (cylindryczne słupki + 4 kolory) wg udziału kobiet, które nie uprawiają aktywnie żadnej formy sportu czy ćwiczeń fizycznych, a reagując na kłopoty lub trudne sytuacje w ich życiu:

- zwracają się o radę i pomoc do innych ludzi LUB

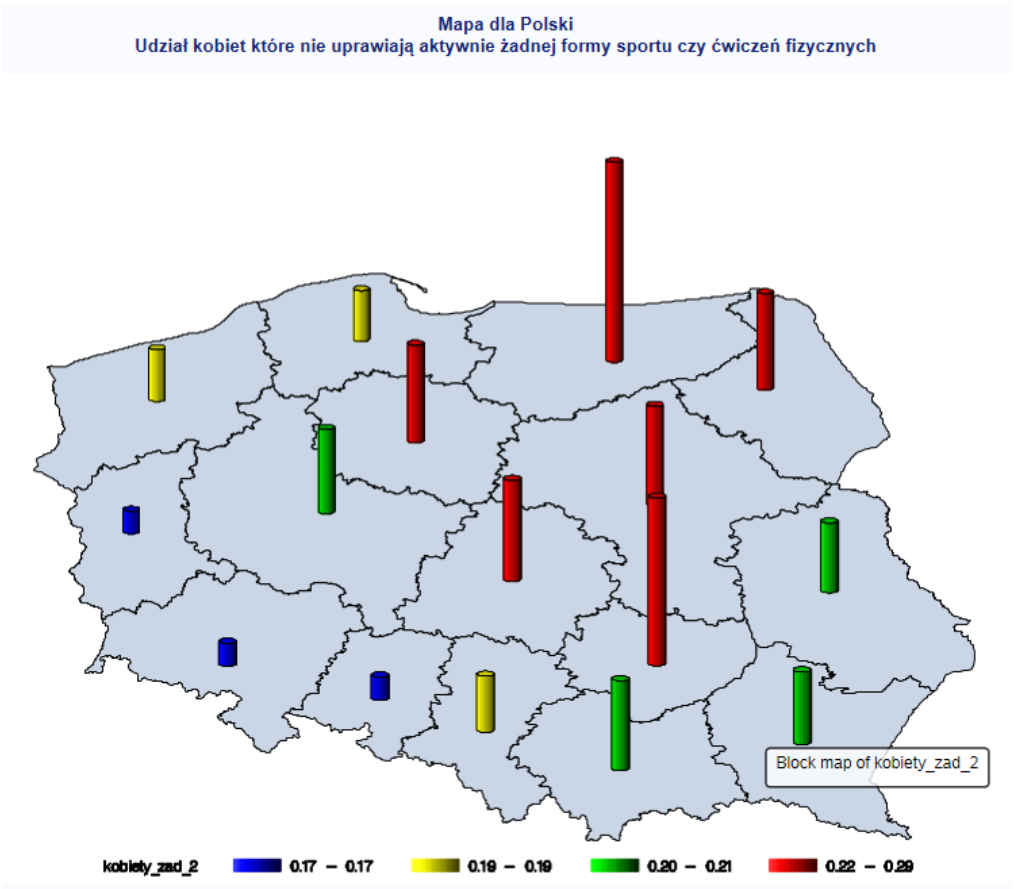
- mobilizują się i przystępują do działania LUB

- zajmują się innymi rzeczami, które odwracają uwagę i poprawiają nastrój

Wartości procentowe w legendzie mapy proszę zaokrąglić do 2 miejsc po przecinku.

Rozwiązanie:

Po uwzględnieniu warunków z zadania otrzymałem mapę Polski z podziałem na województwa, na której można odczytać udział niećwiczących kobiet, które różnie reagują na trudne sytuacje w życiu (warunki zgodne z powyższym opisem do zadania). Okazuje się, iż stosunek takich osób jest najwyższy w województwach m.in.: Warmińsko-Mazurskim oraz Mazowieckim. Natomiast najniższy wskaźnik występuje dla województw m.in. Pomorskiego i Zachodniopomorskiego.



Zadanie 3

3.1. Podpunkt a

Przeprowadź analizę głównych składowych dla:

- zmiennych opisujących różne przekonania i postawy (pytanie 57)

ALBO

- zmiennych opisujących zadowolenie z różnych aspektów życia (pytanie 63)

Opisz kolejne kroki, uzasadnij wybór liczby składowych, zinterpretuj je i nadaj im czytelne nazwy, a następnie zapisz nowe zmienne komponentowe w wynikowym zbiorze.

Rozwiązanie:

W celu przeprowadzenia analizy głównych składowych na początku przygotowano zestawienie zawierające macierz z wartościami własnymi (eigenvalues). Z kryterium Kaisera, wiemy, iż gdy dana wartość własna jest większa od 1, to może być ona uznana za komponent. W moim przypadku wybrano 7 takich komponentów.

Procedura FACTOR
Metoda początkowych czynników: Składowe główne

Oceny ładunku a priori: ONE

Wartości własne macierzy korelacji: łącznie = 20 średnio = 1				
	Wartość własna	Różnica	Udział	Skumulowany
1	2.86839534	0.74160333	0.1434	0.1434
2	2.12679201	0.63094515	0.1063	0.2498
3	1.49584687	0.18389161	0.0748	0.3246
4	1.31195525	0.18386896	0.0656	0.3901
5	1.12808629	0.03305592	0.0564	0.4466
6	1.09503037	0.07525111	0.0548	0.5013
7	1.01977927	0.04121515	0.0510	0.5523
8	0.97856411	0.10483812	0.0489	0.6012
9	0.87372599	0.03524346	0.0437	0.6449
10	0.83848253	0.05784432	0.0419	0.6868
11	0.78063821	0.01942445	0.0390	0.7259
12	0.76121375	0.02109507	0.0381	0.7639
13	0.74011868	0.03509073	0.0370	0.8009
14	0.70502795	0.05226856	0.0353	0.8362
15	0.65275939	0.04370576	0.0326	0.8688
16	0.60905363	0.05201706	0.0305	0.8993
17	0.55703657	0.02899528	0.0279	0.9271
18	0.52804129	0.03449995	0.0264	0.9535
19	0.49354134	0.05763018	0.0247	0.9782
20	0.43591116		0.0218	1.0000

Następnie usunąłem ładunki mniejsze (nieistotne) od wartości bezwzględnej z liczby 0.5. Dla pozostałych danych przeprowadziłem standaryzację zmiennych w zakresie [0,1], w celu dostosowania zmiennych do takiej samej skali. Ładunkom przypisano następujące nazwy:

F1. Przywiązanie do wartości materialnych
F2. Niedocenianie niektórych ludzi
F3. Pozytywne nastawienie do życia
F4. Konsumpcjonizm
F5. Legalizacja związków partnerskich
F6. Przywrócenie kary śmierci
F7. Patriotyzm

Natomiast poniżej, zdecydowałem się zamieścić podstawowe statystyki dla wszystkich siedmiu wybranych komponentów.

Procedura MEANS

Zmienna	N	Średnia	Odch. std.	Minimum	Maksimum
Factor1	9909	-3.72875E-17	1.0000000	-2.9685330	3.1892306
Factor2	9909	6.059225E-17	1.0000000	-3.6615768	2.8704830
Factor3	9909	-5.16289E-17	1.0000000	-3.1478245	5.0190570
Factor4	9909	4.589236E-17	1.0000000	-4.2836333	2.8521197
Factor5	9909	-2.3233E-16	1.0000000	-3.0136827	3.2195091
Factor6	9909	2.131485E-16	1.0000000	-3.3733099	4.3633145
Factor7	9909	-1.65284E-16	1.0000000	-3.3399467	4.5437666

3.2. Podpunkt b

Dla zmiennych komponentowych uzyskanych w podpunkcie a. wykonaj analizę skupień metoda k-średnich. Wybierz optymalną liczbę grup. Sprawdź czy rozwiązanie metody k-średnich z losowym wyborem początkowych środków ciężkości można poprawić używając metody k-średnich wspomaganej metodą hierarchiczną. Wybierz najlepsze Twoim zdaniem rozwiązanie, zapisz je w zbiorze danych i zinterpretuj uzyskane grupy w odniesieniu do zmiennych użytych do grupowania.

Rozwiązanie:

W tym podpunkcie wykorzystano zbiór wynikowy z poprzedniego podpunktu. W pierwszym kroku, na uzyskanym zbiorze z podpunktu 3a, przeprowadzono procedurę fastclus, z parametrem początkowym maxclus = 4, w celu sprawdzenia jak mniej więcej zachowują się dane stosując metodę k-średnich.

91 Observation(s) were omitted due to missing values.

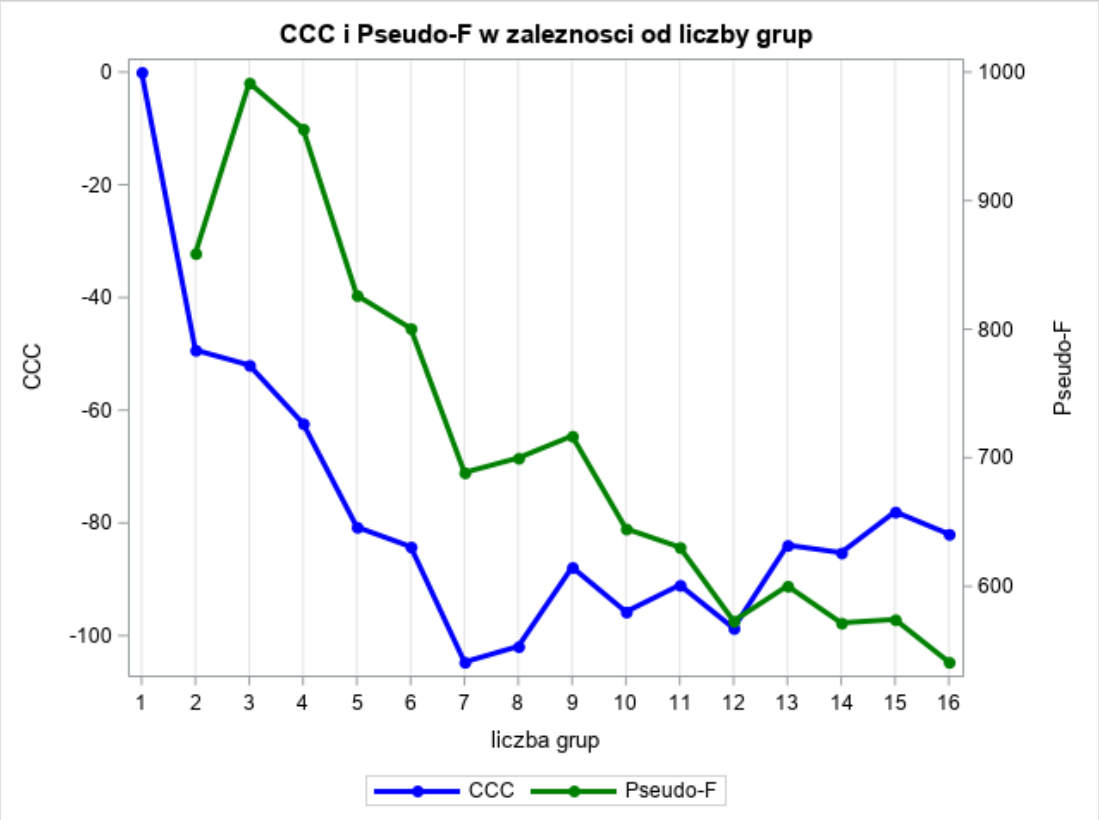
Statystyki dla zmiennych				
Zmienna	Całk. odch. std.	Wewn. odch. std.	R-kwadrat	RSQ/(1-RSQ)
Factor1	0.16240	0.11639	0.486498	0.947412
Factor2	0.15309	0.13641	0.206255	0.259850
Factor3	0.12245	0.12034	0.034366	0.035590
Factor4	0.14014	0.11160	0.366048	0.577406
Factor5	0.16043	0.14246	0.211740	0.268617
Factor6	0.12926	0.12758	0.025998	0.026692
Factor7	0.12684	0.12349	0.052544	0.055458
OVER-ALL	0.14292	0.12588	0.224498	0.289488

Statystyka pseudo-F	955.79
Przybliżone ogólne oczekiwane R-kwadrat	0.31934
Sześcienne kryterium skupień	-62.339

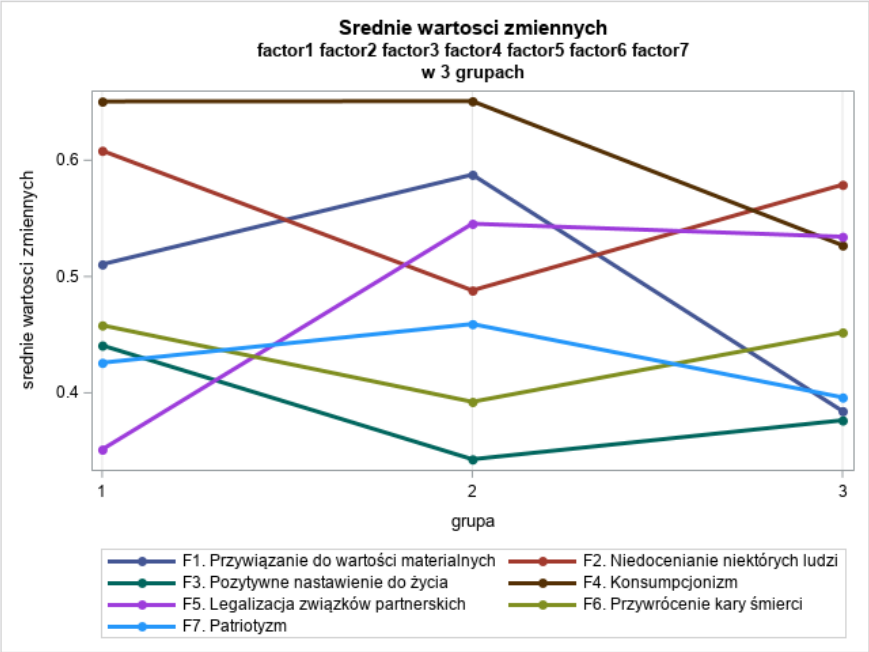
Początkowa wartość statystyki pseudo-F kształtuje się na poziomie 955.79, a dla CCC ta wartość wynosi niestety -62.339, co już może dawać sygnały, iż istnieje prawdopodobieństwo uzyskania lekko zniekształconych grup. W kolejnych krokach podjęto próbę poprawy tych wyników, a w szczególności statystyki CCC.

Skorzystano z makra „sgplotCCCk.sas", w celu sprawdzenia rozkładów wartości dla statystyk CCC oraz Pseudo-F. Parametr maxclus ustawiono na 16, tak aby odpowiadał liczbie województw w Polsce.

Na poniższym wykresie widać, iż ciężko znaleźć znaczące załamanie. Co prawda wartości dla Pseudo-F są dodatnie, jednakże tego samego nie można powiedzieć o CCC. Przypomnę, iż chcielibyśmy aby CCC oraz pseudo-F były jak najwyższe (i dodatnie). Z wykresu wynika, iż optymalna liczba segmentów wynosi 3. W następnych iteracjach wykorzystano tę informację.



W następnej iteracji postanowiono skupić się nad faktorami oraz wyeliminować te, które cechują się najniższym odchyleniem standardowym. Usunięcie mało zróżnicowanych faktorów, może pozytywnie wpłynąć na wzrost statystyki CCC oraz pseudo-F. Uruchomiłem więc makro "sgplotsegk.sas" i postanowiłem, iż z dalszej analizy należy usunąć factor6 factor7.

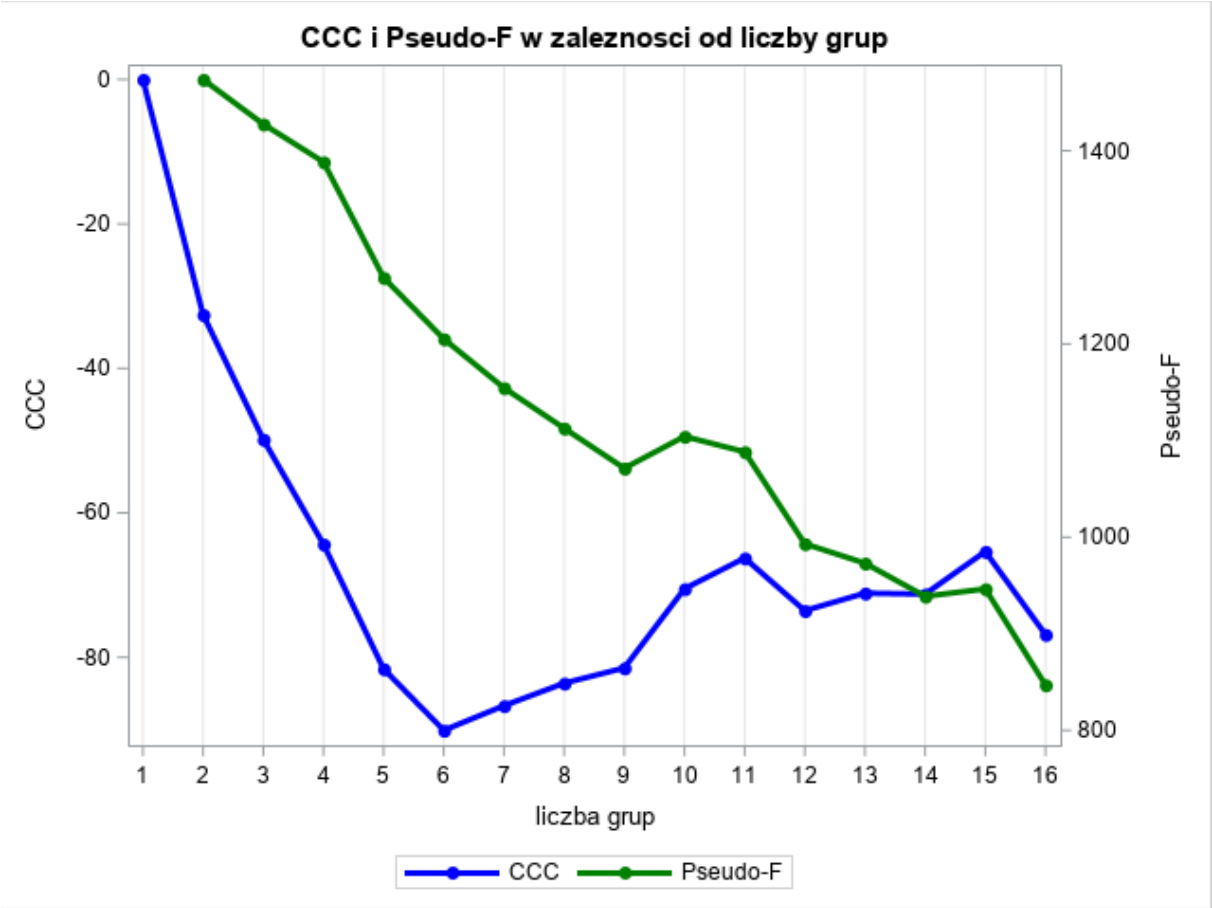


Srednie wartosci zmiennych
factor1 factor2 factor3 factor4 factor5 factor6 factor7
w 3 grupach

Procedura MEANS

Zmienna	Etykieta	Średnia	Odch. std.
Factor1	F1. Przywiązanie do wartości materialnych	0.4911145	0.0840116
Factor2	F2. Niedocenianie niektórych ludzi	0.5589417	0.0511653
Factor3	F3. Pozytywne nastawienie do życia	0.3863604	0.0406326
Factor4	F4. Konsumpcjonizm	0.6071368	0.0585402
Factor5	F5. Legalizacja związków partnerskich	0.4786115	0.0890542
Factor6	F6. Przywrócenie kary śmierci	0.4345863	0.0297052
Factor7	F7. Patriotyzm	0.4261902	0.0258267

Następnie, ponownie uruchomiłem makro „sgplotCCCk.sas” nie uwzględniając w nim dwóch ostatnich faktorów. Udało mi się pozytywnie wpłynąć zarówno na wartości Pseudo-F jak i CCC. W porównaniu do bazowej iteracji, co prawda utraciłem jednoznaczne załamanie, ale dla grup = 3 uzyskałem wyższe wartości statystyk CCC oraz Pseudo-F.



W ostatnim kroku postanowiłem sprawdzić czy poprzez metodę hierarchiczną można również wpłynąć pozytywnie na wyniki. Wykorzystałem do tego zbiór danych, w którym nie uwzględniałem faktorów 6-7. W moim przypadku okazuję się, że metoda hierarchiczna pomogła uzyskać lepsze wyniki. Skorzystałem z makra "kmeanssuph.sas" i widać, iż CCC wzrosło niewiele, bo z poziomu -50 do -34. Natomiast Pseudo-F wzrosło z poziomu około 1400 do 1675. Dla obydwu przypadków, zapisałem wyniki do jednego zbioru wynikowego.

91 Observation(s) were omitted due to missing values.

Statystyki dla zmiennych				
Zmienna	Całk. odch. std.	Wewn. odch. std.	R-kwadrat	RSQ/(1-RSQ)
Factor1	0.16240	0.11609	0.489077	0.957244
Factor2	0.15309	0.10629	0.518099	1.075114
Factor3	0.12245	0.12186	0.009761	0.009857
Factor4	0.14014	0.13581	0.061009	0.064973
Factor5	0.16043	0.15582	0.056884	0.060315
OVER-ALL	0.14845	0.12833	0.252764	0.338266

Statystyka pseudo-F	1675.43
Przybliżone ogólne oczekiwane R-kwadrat	0.31557
Sześcienne kryterium skupień	-34.745

3.3. Podpunkt c

Dla najlepszego rozwiązania analizy skupień z podpunktu b. dokonaj profilowania uzyskanych grup z wykorzystaniem zmiennych:

- *wiek/przedział wiekowy*
- *pleć*
- *klasa wielkości miejscowości zamieszkania*
- *poziom wykształcenia*
- *osobisty dochód miesięczny (na rękę) średnio z ostatnich 3 miesięcy*
- *korzystania z komputera*
- *korzystania z Internetu*
- *korzystania z usług bankowych*
- *liczba książek (jakichkolwiek) przeczytanych (wysłuchanych) w ciągu ostatnich 12 miesięcy*
- *oceny całego dotychczasowego życia*

Rozwiązanie:

W celu dokonania profilowania wykorzystałem procedurę proc freq dla najlepszego uzyskanego wyniku z zadania b. W moim przypadku wykorzystałem zbiór wynikowy wyliczony poprzez przeprowadzenie metody hierarchicznej z 5 faktorami. Poniżej przedstawiam charakterystyki grup dla wszystkich 3 podziałów.

1 grupa.

- wiek/przedział wiekowy: 45-59 lat
- pleć: kobieta
- klasa wielkości miejscowości zamieszkania: miasto
- poziom wykształcenia: ZASADNICZE ZAWODOWE
- osobisty dochód miesięczny (na rękę) średnio z ostatnich 3 miesiecy:2000
- korzystania z komputera: TAK
- korzystania z Internetu: TAK
- korzystania z usług bankowych: TAK
- liczba książek (jakichkolwiek) przeczytanych (wysłuchanych) w ciągu ostatnich 12 miesięcy: 0
- oceny całego dotychczasowego życia: UDANE

2 grupa.

- wiek/przedział wiekowy: 45-59 lat
- pleć: kobieta
- klasa wielkości miejscowości zamieszkania: wieś
- poziom wykształcenia: ZASADNICZE ZAWODOWE
- osobisty dochód miesięczny (na rękę) średnio z ostatnich 3 miesięcy: 2000
- korzystania z komputera: TAK
- korzystania z Internetu: TAK
- korzystania z usług bankowych: TAK
- liczba książek (jakichkolwiek) przeczytanych (wysłuchanych) w ciągu ostatnich 12 miesięcy: 0
- oceny całego dotychczasowego życia: UDANE

3 grupa.

- wiek/przedział wiekowy: 65+ lat
- płeć: kobieta
- klasa wielkości miejscowości zamieszkania: miasto
- poziom wykształcenia: ZASADNICZE ZAWODOWE
- osobisty dochód miesięczny (na rękę) średnio z ostatnich 3 miesięcy: 2000
- korzystania z komputera: TAK
- korzystania z Internetu: TAK
- korzystania z usług bankowych: TAK

- liczba książek (jakichkolwiek) przeczytanych (wysłuchanych) w ciągu ostatnich 12 miesięcy: 0
- oceny całego dotychczasowego życia: UDANE

Zadanie 4

4.1. Podpunkt a

Dokonaj hierarchicznej analizy skupień NA POZIOMIE WOJEWÓDZTW analizując ich podobieństwo pod względem:

- pierwszego kwartyla wieku respondenta

- przeciętnego pragnienia do życia (pytanie "Jak silne w tych dniach jest Pana pragnienie życia?")

- mediany dochodu miesięcznego netto (na rękę) spodziewanego za 2 lata

- średniego indeksu BMI (waga w kg / (wzrost w metrach)^2)

- proporcji osób, które w minionym tygodniu oglądały telewizję mniej niż przez 2 godziny dziennie

- proporcji osób z wykształceniem co najmniej średnim

- odstępu międzykwartylowego dla liczby osób zaliczanych do przyjaciół

Wybierz metodę pozwalającą uzyskać skupienia o minimalnej wewnątrzgrupowej wariancji. Dendrogram wyświetl poziomo.

Rozwiązanie:

W celu wykonania analizy skupień na poziomie województw na początku należało wykonać 7 warunków z powyższego polecenia. W tym celu niezwykle pomocne okazały się być funkcje proc freq, które pomogła mi sprawdzić jak wygląda rozkład niektórych zmiennych oraz proc mean, dzięki której mogłem wyznaczyć różne statystyki takie jak np. mediana.

Uzyskane wartości zostały przyrównane do zmiennych z przedziału faktor1-faktor7. Poniżej zamieszczam dokładny opis dla przypisanych labelek.

F1. Pierwszy kwartyl wieku respondenta
F2. Przeciętne pragnienie do życia
F3. Mediana dochodu miesięcznego netto

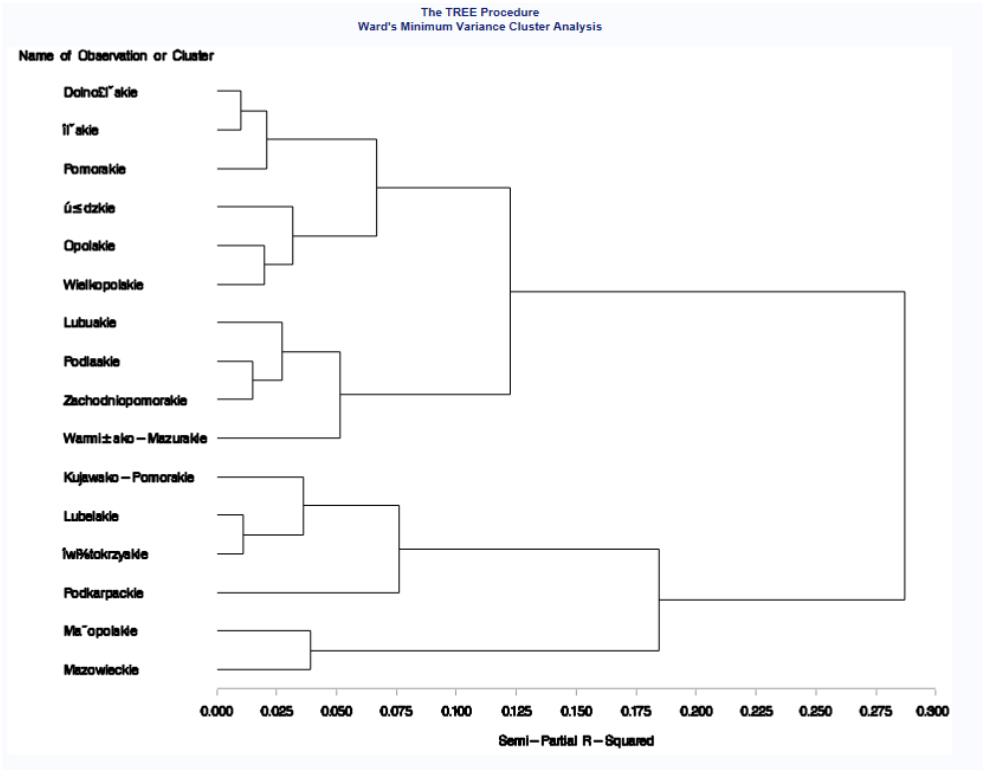
F4. Średni indeks BMI

F5. Proporcja osób, które w minionym tygodniu oglądały telewizję mniej niż przez 2 godziny dziennie

F6. Proporcja osób z wykształceniem co najmniej średnim

F7. odstęp międzykwartylowy dla liczby osób zaliczanych do przyjaciół

Następnie poprzez funkcję proc cluster dokonałem hierarchicznej analizy skupień. W celu uzyskania skupienia o minimalnej wewnątrzgrupowej wariancji, skorzystałem z metody Warda.



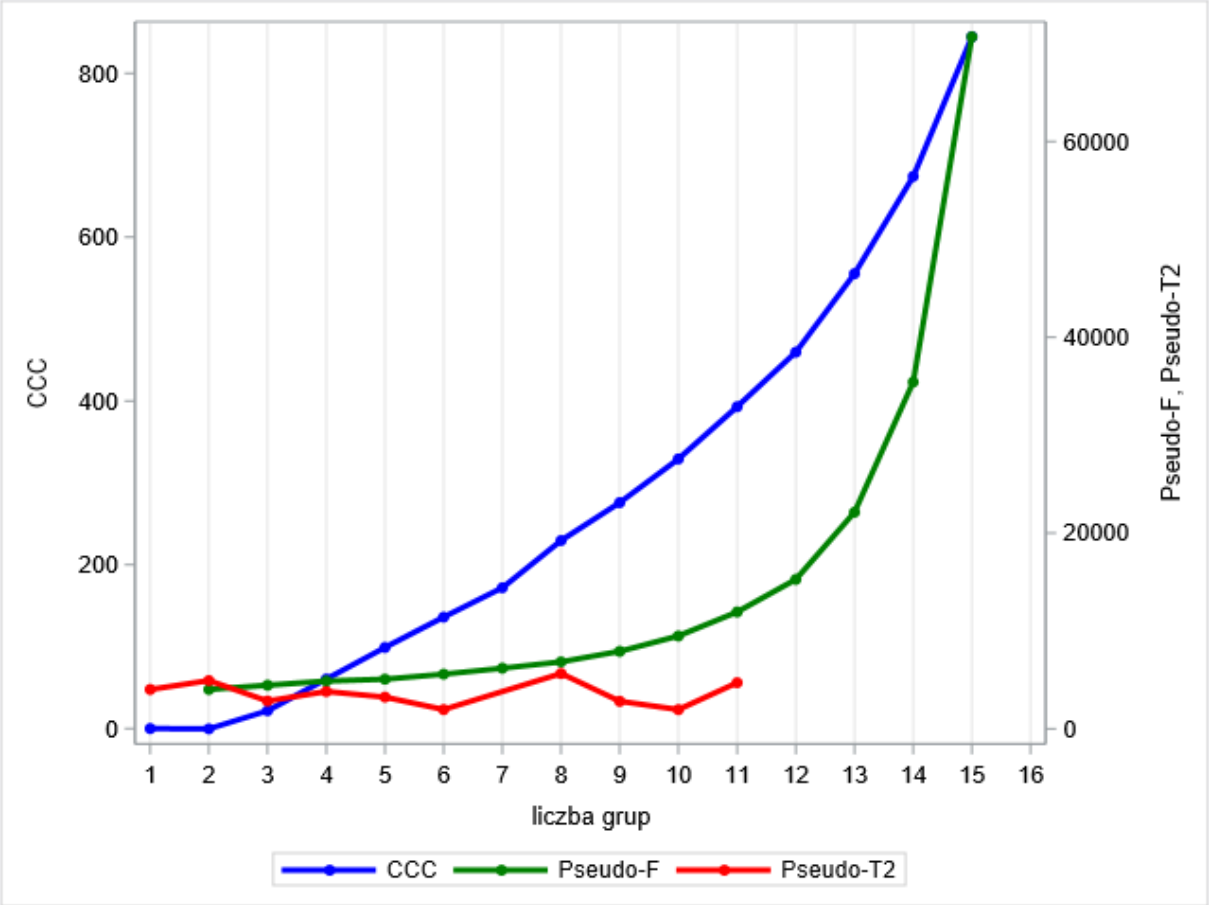
Na podstawie powyższego wykresu, ciężko jednoznacznie stwierdzić jaka liczba skupień wydaję się być najlepszą. Bezpieczniejszym pomysłem będzie podanie potencjalnego przedziału, dla którego liczba skupień oscyluje pomiędzy 2-6.

4.2. Podpunkt b

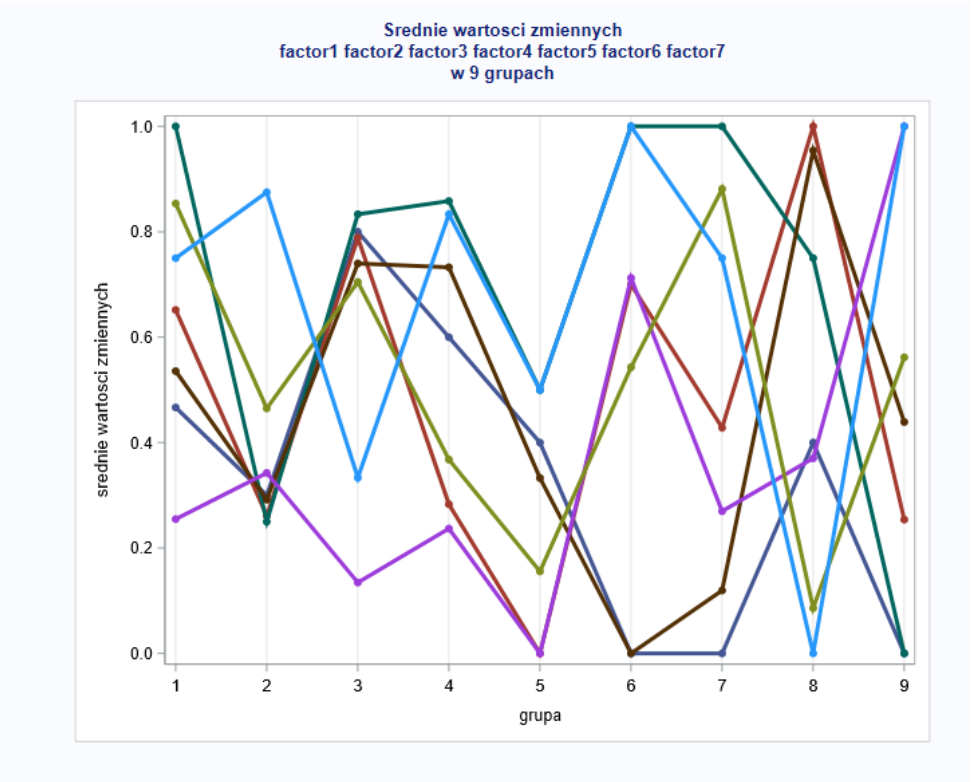
Wybierz optymalne rozwiązanie. Wyświetl średnie wartości charakterystyk wykorzystanych do grupowania dla poszczególnych skupień i zinterpretuj uzyskane grupy.

Rozwiązanie:

W celu doboru odpowiedniej liczby skupień wykorzystałem plik sgplotCCCh, w którym argument maxncl ustawiłem na 16, co ma odpowiadać liczbie wszystkich polskich województw.



Wraz ze wzrostem liczby grup, można wręcz pokusić o stwierdzenie, iż zarówno CCC jaki i Pseudo-F zaczynają wzrastać nawet w wykładniczym tempie. Zdecydowanie inaczej wygląda wykres dla statystyki Pseudo-T2. Dla grupy 8mej zaliczono niewielki wzrost wartości, a już dla następnej grupy zauważono spadek o taką samą wartość. Ze względu na ten skok, uważam, iż optymalna liczba grup powinna być równa aż 9. W następnym kroku postanowiłem sprawdzić zróżnicowanie faktorów. Niemniej ze względu na ich wysokie odchylenia standardowe, w ostatecznym rozrachunku postanowiłem niczego nie usuwać. Moją konkluzję potwierdza poniższy wykres dla średnich wartości zmiennych, na którym widać iż rozkład zmiennych zdecydowanie nie jest „płaski”.



Srednie wartosci zmiennych
factor1 factor2 factor3 factor4 factor5 factor6 factor7
w 9 grupach

The MEANS Procedure

Variable	Label	Mean	Std Dev
factor1	F1. Pierwszy kwartyl wieku respondenta	0.3404167	0.2708729
factor2	F2. Przeciętne pragnienie do życia	0.4871441	0.3008314
factor3	F3. Mediana dochodu miesięcznego netto	0.6930729	0.3413850
factor4	F4. Średni indeks BMI	0.4673897	0.2942118
factor5	F5. Proporcja osób, które w minionym tygodniu oglądały telewizję mniej niż przez 2 godziny dziennie	0.3628436	0.2889902
factor6	F6. Proporcja osób z wykształceniem co najmniej średnim	0.5177656	0.2639564
factor7	F7. odstęp międzykwartylowy dla liczby osób zaliczanych do przyjaciół	0.6713542	0.3145424

Tak więc w moim ostatecznym zbiorze wynikowym, liczba skupień wynosi 9, natomiast liczba faktorów jest równa 7.

Zadanie 5

5.1. Podpunkt a

Zweryfikuj hipotezę, ze rozkład pragnienia życia (w tych dniach) dla osób, które nie były w minionym miesiącu:

- w kinie, teatrze lub na koncercie, ani
- na spotkaniu towarzyskim
nie zależy od klasy wielkości miejscowości zamieszkania

Rozwiązanie:

W celu zweryfikowania tej hipotezy na początku sprawdziłem, czy występuje normalność rozkładu HP40. Pvalue wyszło mniejsze od 0.05, tak więc odrzuciłem hipotezę zerową mówiącej o normalności zmiennej HP40. W związku z tym wykonałem nieparametryczny test Kruskala-Willisa. Pvalue dla tego testu wyniosło 0.4323, tak więc nie ma podstaw do odrzucenia hipotezy 0. Rozkład pragnienia życia nie zależy od klasy wielkości miejscowości zamieszkania.

The NPAR1WAY Procedure

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
0.6167	1	0.4323

5.2. Podpunkt b

Zweryfikuj hipotezę, ze częstość udziału w nabożeństwach osób mieszkających w największych miastach nie zależy od płci.

Rozwiązanie:

W celu zweryfikowania tej hipotezy na początku sprawdziłem, czy występuje normalność rozkładu HP38. Pvalue wyszło mniejsze od 0.05, tak więc odrzuciłem hipotezę zerową mówiącą o normalności zmiennej HP38. W związku z tym wykonałem nieparametryczny test Kruskala-Willisa. Pvalue dla tego testu wyniosło <.0001, tak więc należy odrzucić hipotezę 0. Częstość udziału w nabożeństwach osób mieszkających w największych miastach zależy od płci.

The NPAR1WAY Procedure

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
16.1647	1	<.0001

5.3. Podpunkt c

Zweryfikuj hipotezę, że średnie zadowolenie z pracy i ze swego wykształcenia dla osób, które przynajmniej 1 godzinę tygodniowo poświęcają na czytanie prasy, są sobie równe.

Rozwiązanie:

W celu zweryfikowania poprawności tej hipotezy wykorzystałem procedurę paired t-test, w którym powiązanymi ze sobą zmiennymi były HP63_11 z HP63_13. Pvalue dla tego testu wyniosło <0.0001, tak więc należy odrzucić hipotezę 0. Średnie zadowolenie z pracy i ze swojego wykształcenia dla osób, które przynajmniej 1 godzinę tygodniowo poświęcają na czytanie pracy, nie są sobie równe.

The TTEST Procedure

Difference: HP63_11 - HP63_13

N	Mean	Std Dev	Std Err	Minimum	Maximum
4179	-0.0775	1.1268	0.0174	-5.0000	5.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-0.0775	-0.1117	-0.0434	1.1268	1.1032	1.1515

DF	t Value	Pr > |t|
4178	-4.45	<.0001

5.4. Podpunkt d

Osobno dla kobiet i dla mężczyzn zweryfikuj hipotezę, ze taki sam procent osób uważa, że ogólnie rzecz biorąc większości ludzi można ufać oraz, że demokracja ma przewagę nad wszelkimi innymi formami rządów

Rozwiązanie:

Na początku postanowiłem podzielić zbiór ze względu na płeć. Zgodnie z warunkami zadania, za pomocą procedury proc sql i case when, wyznaczyłem osoby, które sądzą, iż ludziom można ufać, a demokracja ma przewagę nad innymi formami rządów. Następnie, do porównania rozkładów procentowych wykorzystałem procedurę proc freq. Zarówno dla mężczyzn jak i dla kobiet, statystyka pvalue dla tego testu wyniosła <0.001, tak więc należy odrzucić hipotezę 0. Zarówno dla grupy mężczyzn jak i kobiet procent osób, które uważają że ogólnie rzecz biorąc większości ludzi można ufać oraz, że demokracja ma przewagę nad wszelkimi innymi formami rządów jest różny.

Mężczyźni:

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of UFAC by DEMO		
	DEMO		
UFAC	0	1	Total
0	2782 63.81 74.81 88.35	937 21.49 25.19 77.37	3719 85.30
1	367 8.42 57.25 11.65	274 6.28 42.75 22.63	641 14.70
Total	3149 72.22	1211 27.78	4360 100.00

Statistics for Table of UFAC by DEMO

McNemar's Test		
Chi-Square	DF	Pr > ChiSq
249.1564	1	<.0001

Simple Kappa Coefficient			
Estimate	Standard Error	95% Confidence Limits	
0.1283	0.0154	0.0981	0.1585

Sample Size = 4360

Kobiety:

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of UFAC by DEMO		
	DEMO		
UFAC	0	1	Total
0	3716 65.89 77.76 87.66	1063 18.85 22.24 75.87	4779 84.73
1	523 9.27 60.74 12.34	338 5.99 39.26 24.13	861 15.27
Total	4239 75.16	1401 24.84	5640 100.00

Statistics for Table of UFAC by DEMO

McNemar's Test		
Chi-Square	DF	Pr > ChiSq
183.8588	1	<.0001

Simple Kappa Coefficient			
Estimate	Standard Error	95% Confidence Limits	
0.1353	0.0142	0.1076	0.1631

Sample Size = 5640

5.5. Podpunkt e

Używając stosownych testów statystycznych odpowiedz na pytanie, czy zadowolenie ze sposobu spędzania wolnego czasu zależy od statusu społeczno-zawodowego respondenta.

Rozwiązanie:

W celu sprawdzenia powyższej hipotezy, na początku zweryfikowałem normalność zmiennej HP63_12. Pvalue wyszło mniejsze niż 0.05, tak więc w następnym kroku wykorzystałem nieparametryczny test Kruskala Wallisa. Również i w tym teście, ze względu na równie niską wartość pvalue, musiałem odrzucić hipotezę 0. Zadowolenie ze sposobu spędzania wolnego czasu nie zależy od statusu społeczno-zawodowego respondenta

The NPAR1WAY Procedure

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
176.5431	8	<.0001

.

5.6. Podpunkt f

Policz i podaj interpretacje miary/miar współzależności oceny całego swojego dotychczasowego życia i miesięcznego dochodu (na rękę).

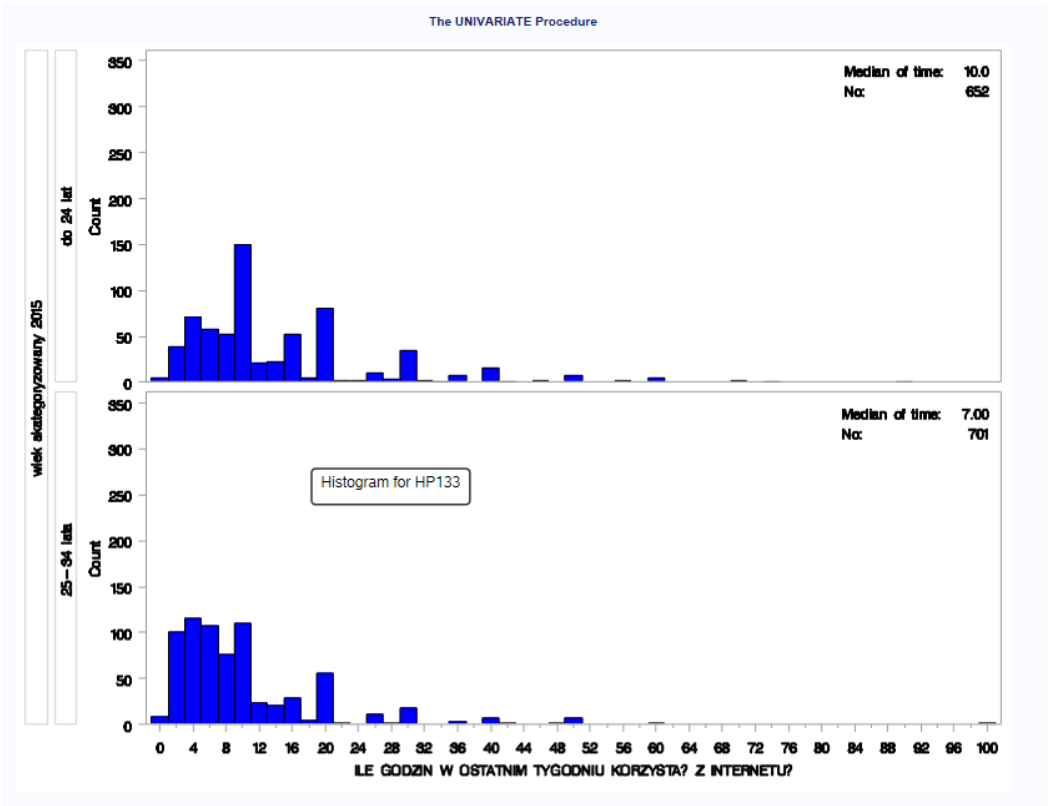
Rozwiązanie:

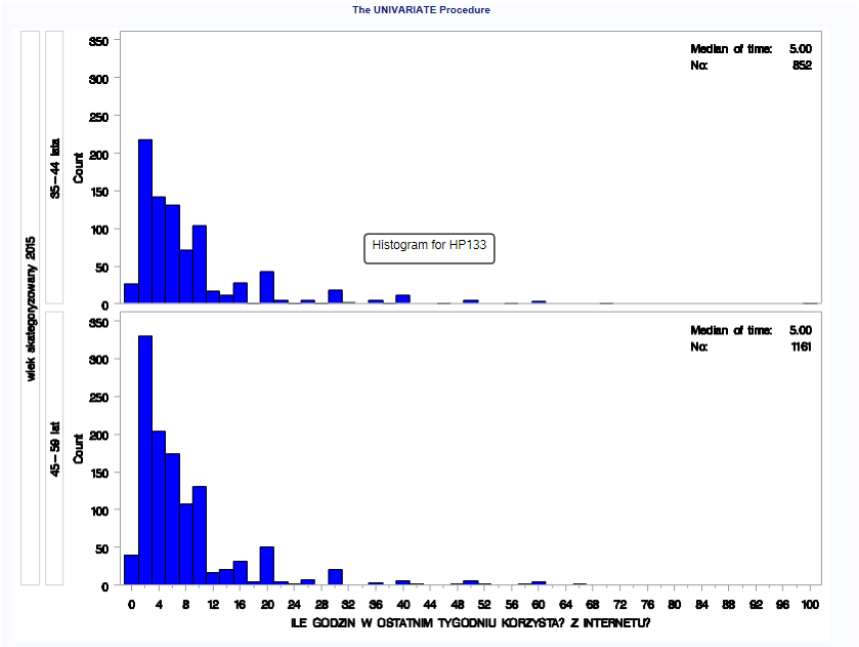
W celu zdecydowania, którą miarę współzależności należy wykorzystać (Pearson lub Spearman), początkowo sprawdziłem normalność zmiennej HP3 oraz HP65. W obu przypadkach pvalue było mniejsze od przedziału ufności 0.05, tak więc nie spełniłem założenia o normalności rozkładu zmiennych. W takim wypadku, musiałem posłużyć się współczynnikiem korelacji Spearmana. Wartość tego współczynnika wynosi -0,23952. W związku z tym zależność pomiędzy oceną całego swojego dotychczasowego życia a miesięcznego dochodu jest ujemnie skorelowania.

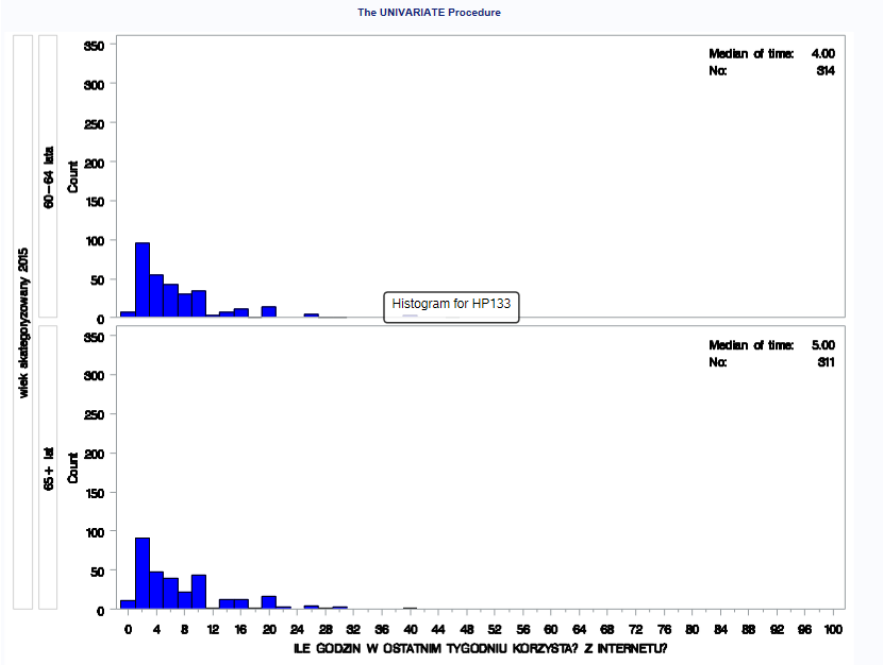
Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	HP3	HP65
HP3 JAK OCENIA PAN SWOJE CALE DOTYCHCZASOWE ZYCIE	1.00000 9984	-0.23952 <.0001 7543
HP65 PANA WLASNY (OSOBISTY) DOCHOD MIESIECZNY NETTO (NA REKE)	-0.23952 <.0001 7543	1.00000 7556

5.7. Podpunkt g

Narysuj i zinterpretuj wykres podstawowych parametrów rozkładu liczby godzin korzystania z Internetu w ostatnim tygodniu w zależności od grupy wiekowej. Ogranicz się do osób, które nigdy nie płaciły za treści dostępne w Internecie.







Pierwszą rzecz, która zdecydowanie rzuca się w oczy jest fakt, iż liczba godzin korzystania z Internetu dla wszystkich grup wiekowych, oscyluje w okolicach od 0 do 20 godzin, co może świadczyć o prawostronnej asymetrii rozkładów. Po drugie, można zauważyć, iż największa liczba osób, które nie płaciły nigdy za treści dostępne w Internecie znajduje się w przedziale wiekowym 45-55 lat. Niemniej jednak, mediana liczby godzin spędzanych w Internecie jest zdecydowanie najwyższa dla grupy ankietowanych z przedziału wiekowego do 24 lat. Co ciekawe, można zauważyć, iż respondenci w wieku 60+ pomimo niewykupywania nigdy treści dostępnych w Internecie, w większości nie są osobami negatywnie nastawionymi do świata wirtualnego.