

II. Analiza struktury – ćwiczenia

1. Wprowadzenie do SAS Enterprise Guide, wersja 7.1

1. Informacje ogólne. Organizacja pracy w EG. Przypisanie biblioteki. Podgląd wyników. Utworzenie nowego programu
2. Importowanie i eksportowanie danych
3. Atrybuty zbioru danych
4. Filtrowanie i sortowanie danych
5. Budowa zapytań SQL
6. Łączenie zbiorów danych
7. Formaty
8. Raportowanie
9. Próbkowanie
10. Transpozycja zbioru danych
11. Wykresy

2. Analiza struktury

1. Tabele liczebności i częstości
2. Miary położenia rozkładu, zróżnicowania, asymetrii oraz koncentracji
3. Obserwacje odstające



Wprowadzenie do SAS Enterprise Guide, wersja 7.1

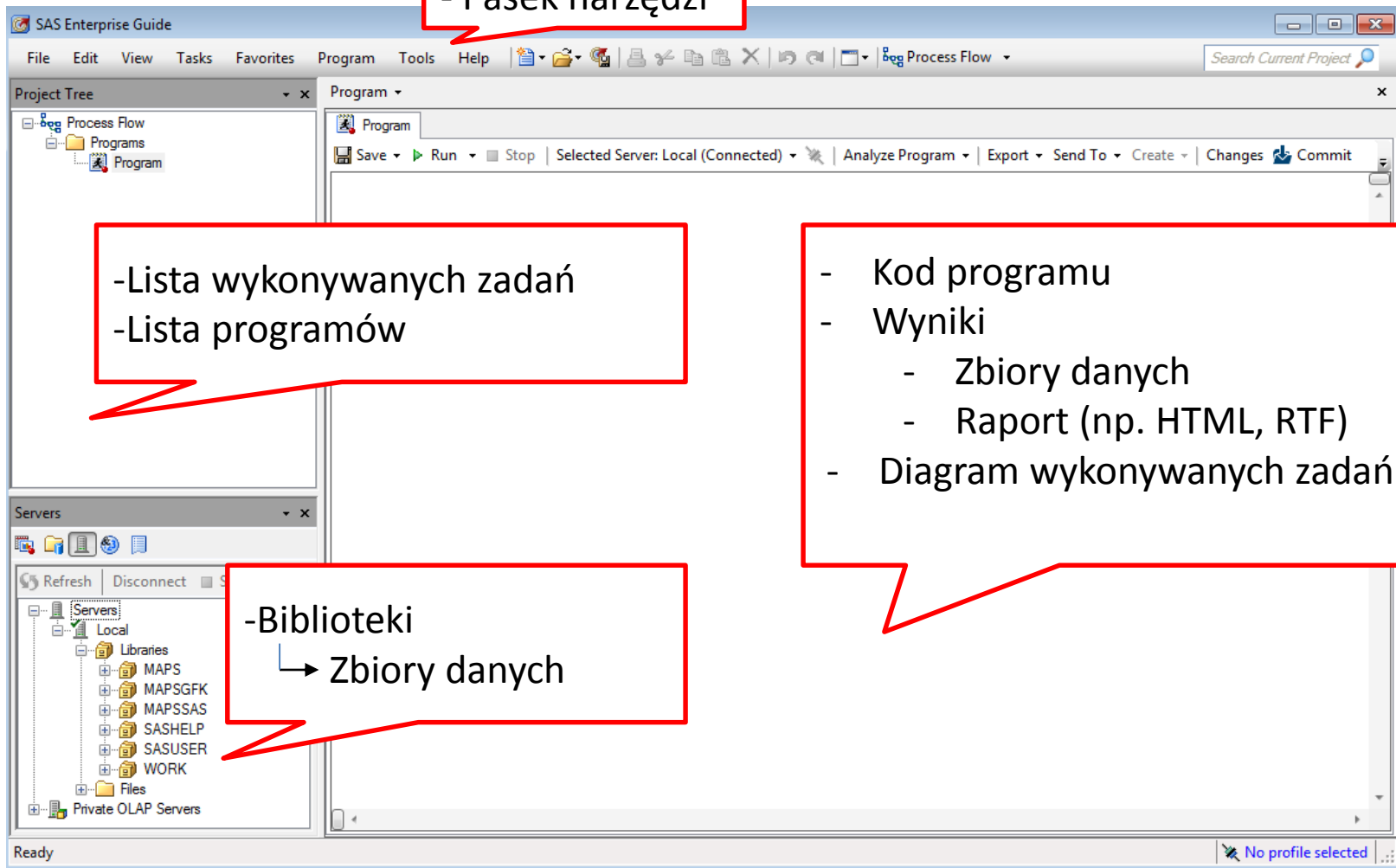
1. SAS Enterprise Guide, wersja 7.1

- Pasek narzędzi

-Lista wykonywanych zadań
-Lista programów

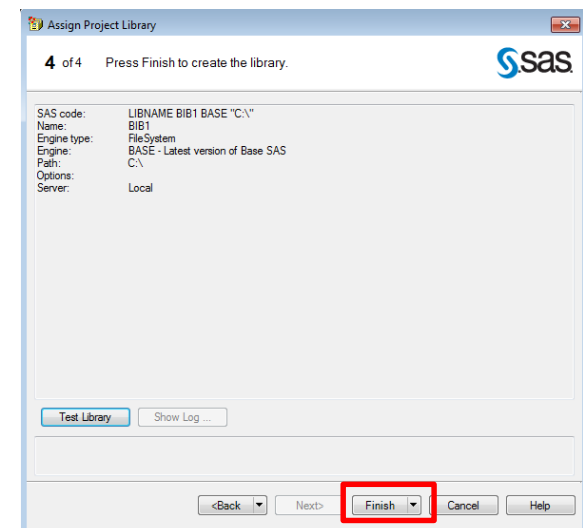
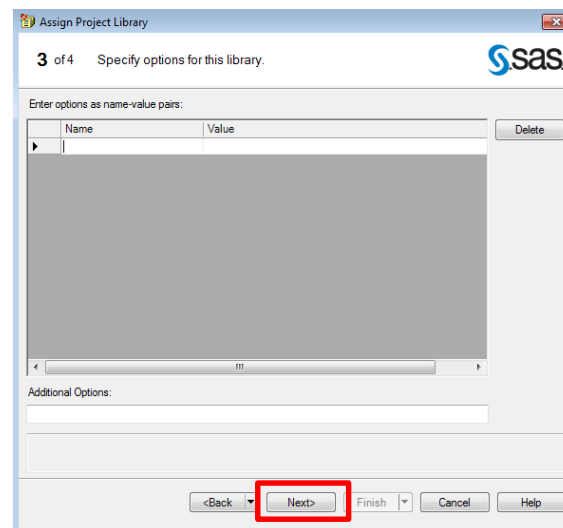
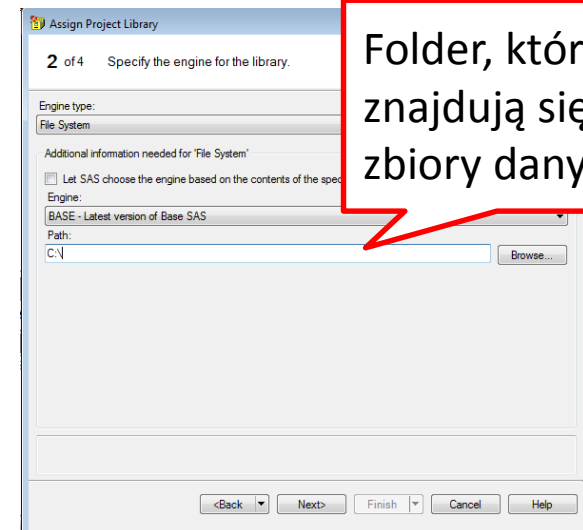
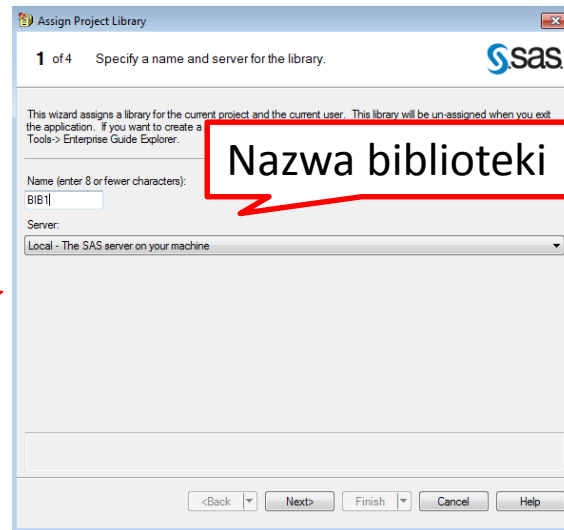
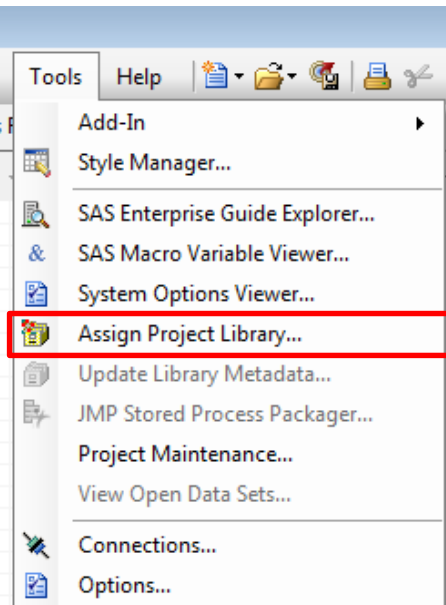
- Kod programu
- Wyniki
- Zbiory danych
- Raport (np. HTML, RTF)
- Diagram wykonywanych zadań

-Biblioteki
↳ Zbiory danych



Przypisanie biblioteki

```
libname bib1 'C:\...\';
```



Podgląd wyników

Project Tree

- Process Flow
 - Assign Project Library (BIB1)
 - HPRACA
 - Summary Statistics

Summary Statistics

Input Data | Code | Log | Results

Filter and Sort | Query Builder | Data | Describe | Graph | Analyze | Export | Send To

	HOURS	AGE	EDU	WAGE	NKIDS	HINC	DF
1	15	34.91	1	16.42	1	20.53	1.00
2	35	55.85	2	3.41	0	17.33	0.00
3	43	30.66	2	8.38	0	58.38	0.00
4	49	31.17	3	8.04			
5	51	30.75	3	7.61			
6	42	33.44	3	4.89	1	42.64	0.00

Wejściowy zbiór danych

Wyniki (np. plik HTML)

Log – raport z wykonania programu

Kod programu

Utworzenie nowego programu

Project Tree

- Process Flow
 - Assign Project Library (BIB1)

Process Flow

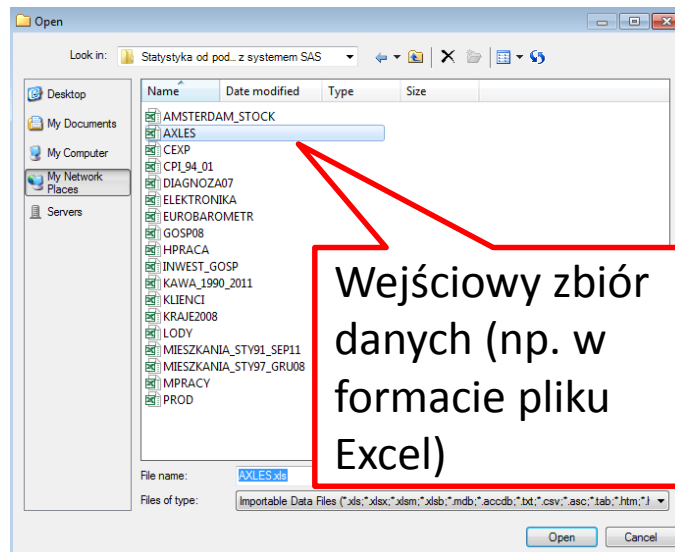
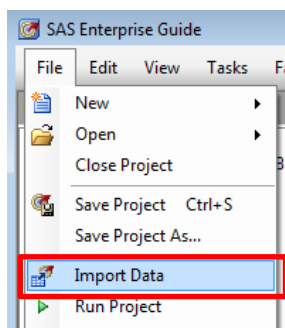
Run | Stop | Export

Assign Project Library (BIB1)

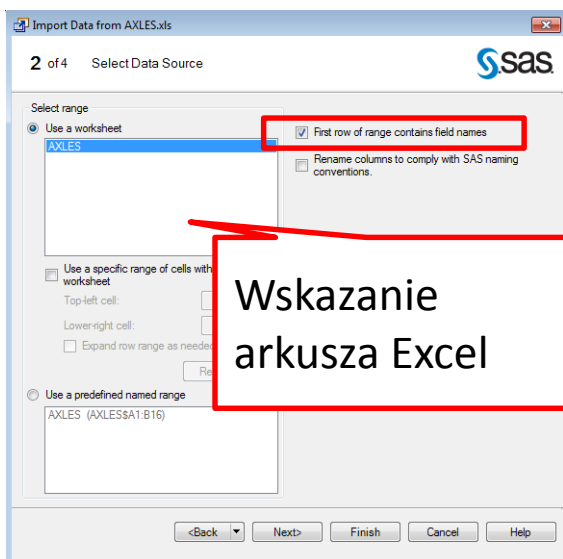
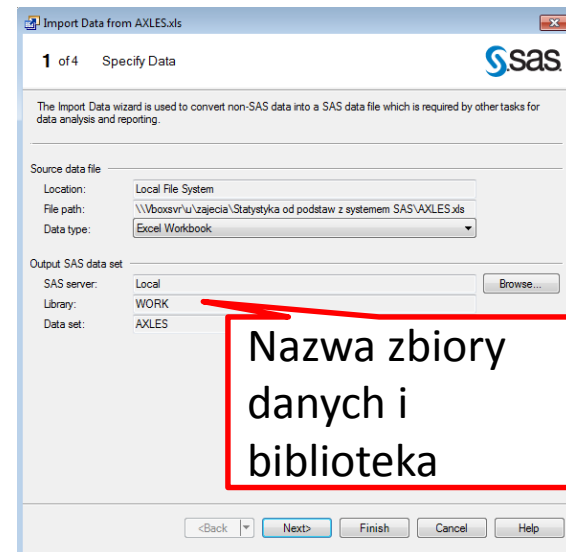
New

- Project
- Data
- Program**
- Task
- Report
- Stored Process
- Note
- Process Flow
- Ordered List

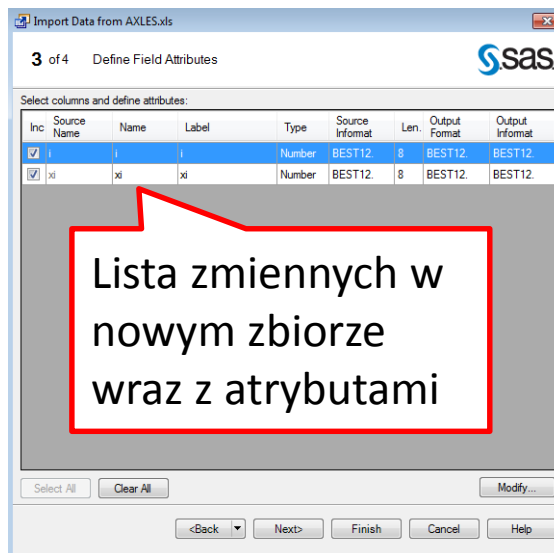
2. Importowanie danych (Plik zewnętrzny → Plik SAS) [1]



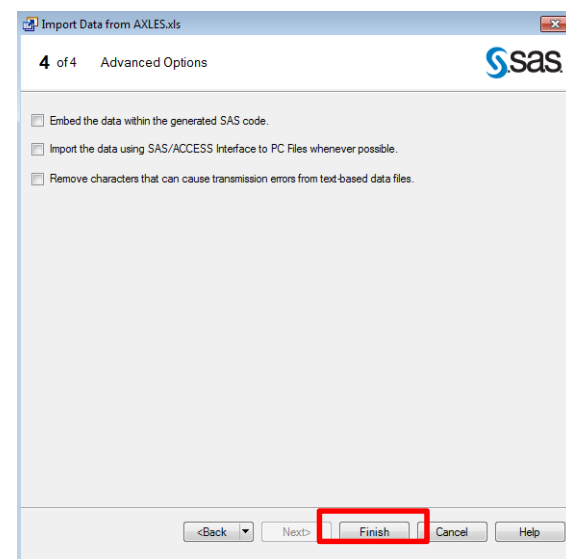
Wejściowy zbiór danych (np. w formacie pliku Excel)



Wskazanie arkusza Excel



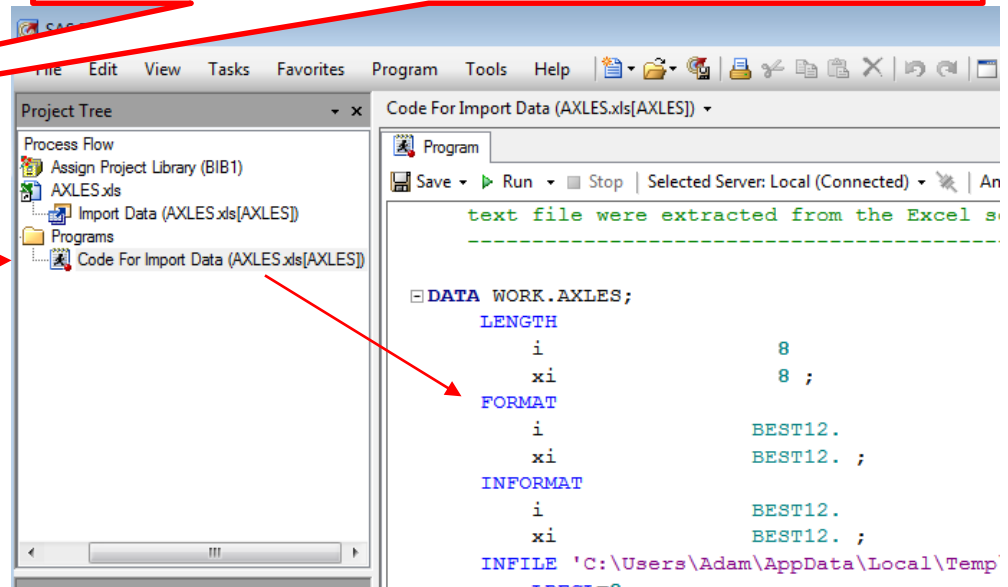
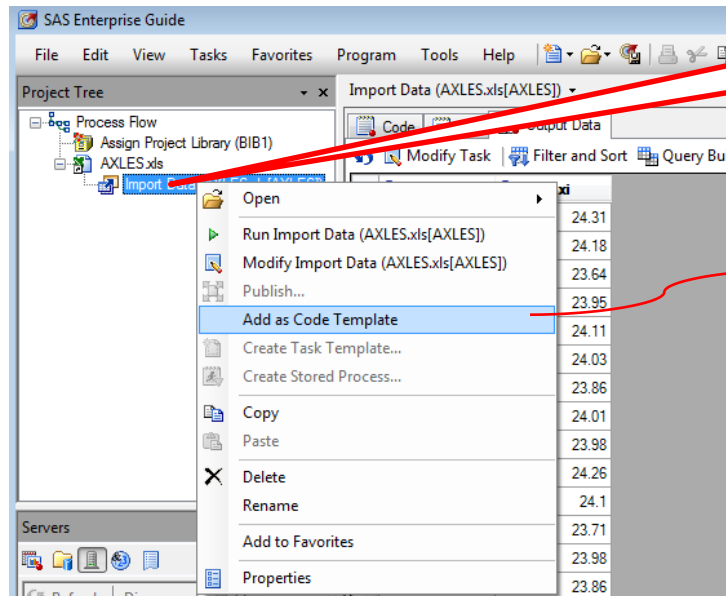
Lista zmiennych w nowym zbiorze wraz z atrybutami



Importowanie danych [2]

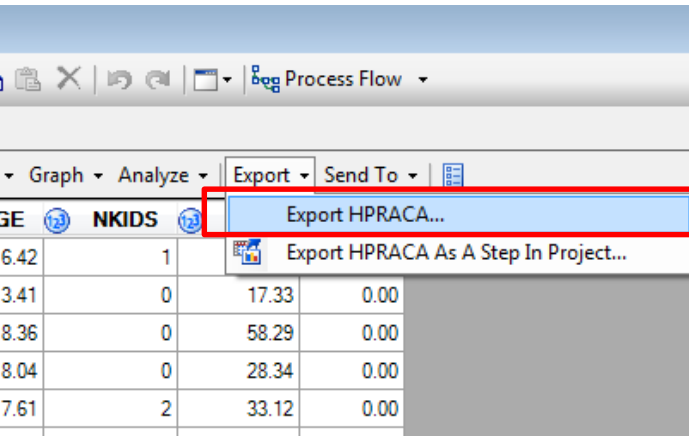
```
proc import datafile=„C:\...\axles.xls"
  out=out.axles
  dbms=xls
  replace;
  getnames=yes;
run;
```

Wyświetlenie kodu programu wykorzystanego do wykonania danego zadania (w tym przypadku do importowania danych)



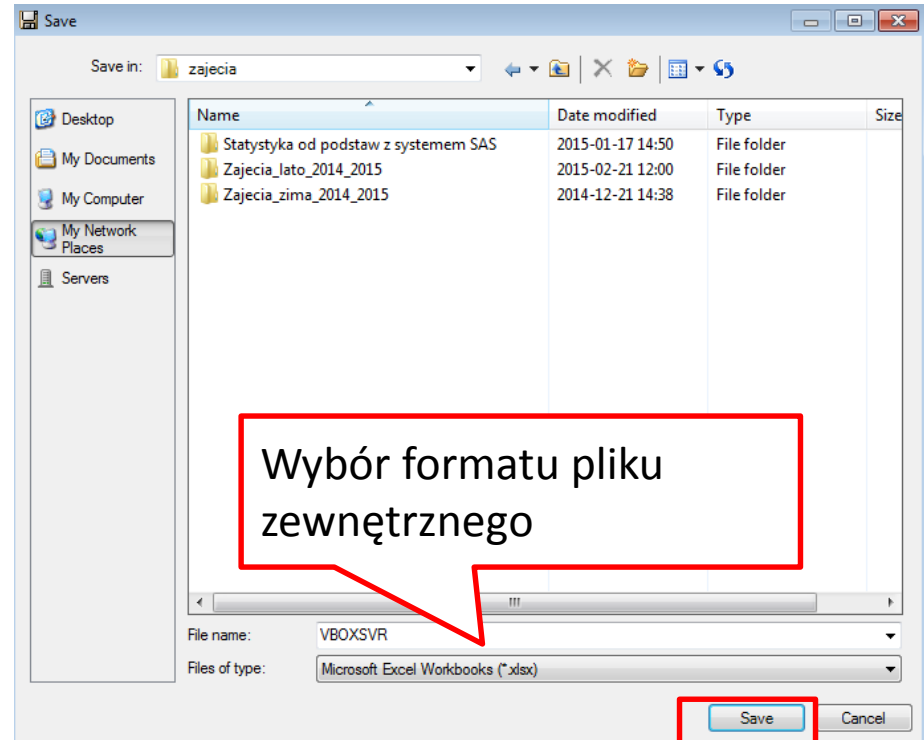
Eksportowanie danych (Plik SAS → Plik zewnętrzny)

```
proc export data=bib1.axles
  outfile="\\Vboxsvr\u\zajecia\nowy_zbior.xls"
  dbms=xls
  replace;
run;
```



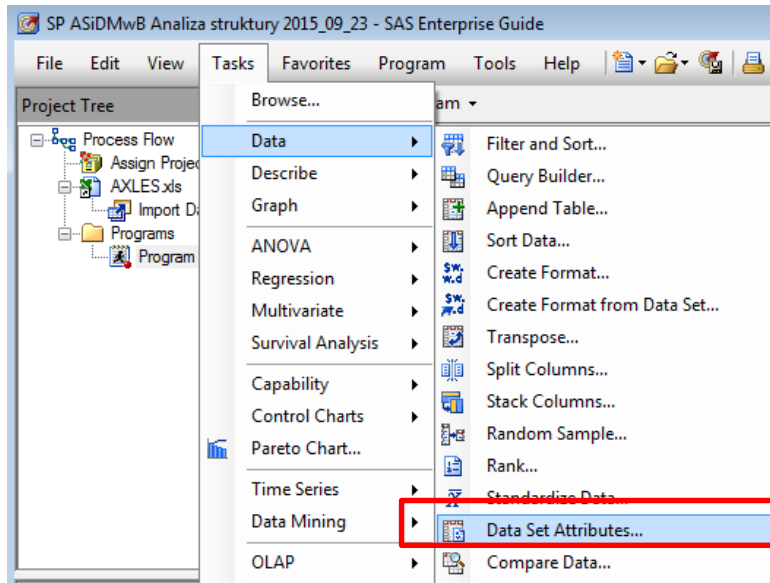
The screenshot shows the SAS software interface. The 'Export' menu option is highlighted with a red box. Below the menu, a table of data is visible.

GE	NKIDS		
6.42	1		
3.41	0	17.33	0.00
8.36	0	58.29	0.00
8.04	0	28.34	0.00
7.61	2	33.12	0.00



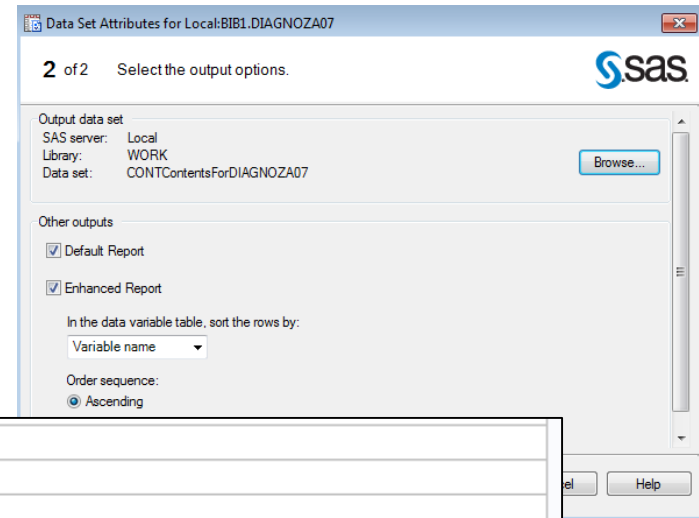
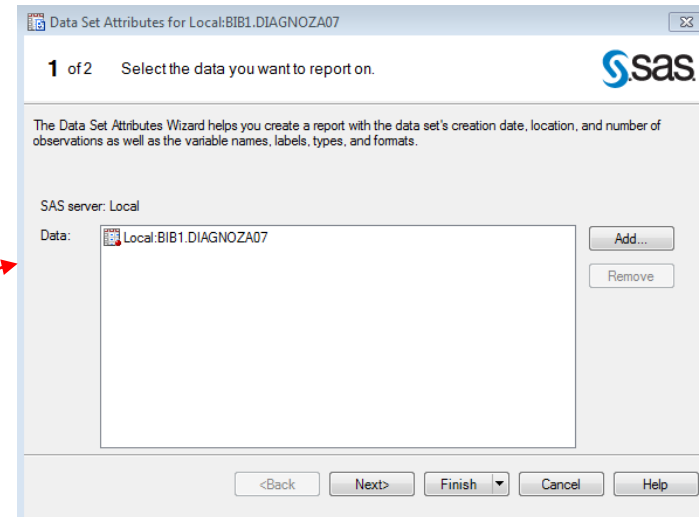
3. Atrybuty zbioru danych

Zbiór danych



```
proc contents  
data=bib1.diagnoza07;  
run;
```

Fragment raportu



4	POW	Num	8	Powiat
5	ROK	Num	8	dc8-Rok urodzenia
7	SC	Num	8	dc10-stan cywilny

4. Filtrowanie i sortowanie danych (1)

SP ASiDMwB Analiza struktury 2015_09_23 - SAS Enterprise Guide

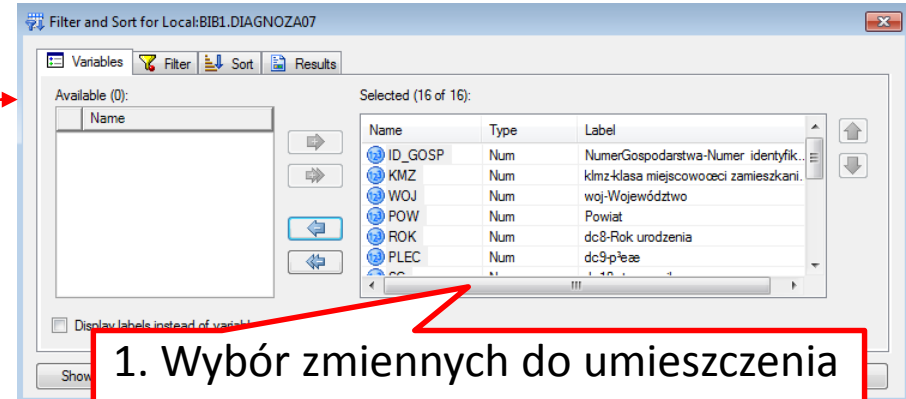
File Edit View Tasks Favorites Program Tools Help

Project Tree

- Process Flow
 - Assign Project Library (BIB1)
 - AXLES.xls
 - Import Data (AXLES.xls[AXLES])
 - DIAGNOZA07
 - Data Set Attributes
 - Programs
 - Program

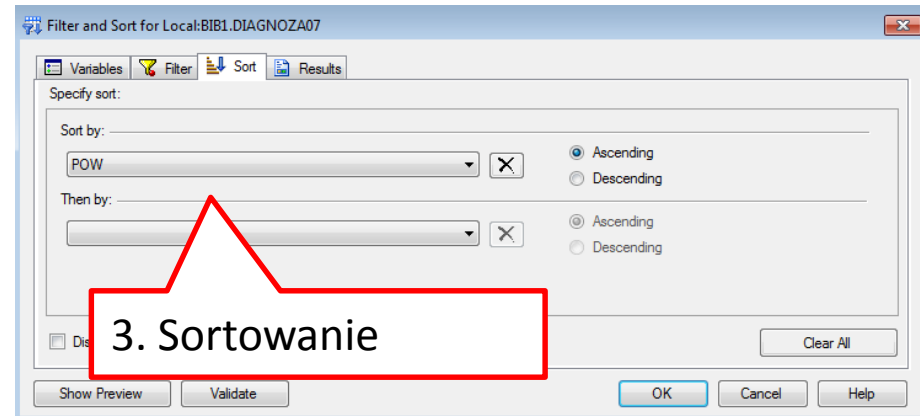
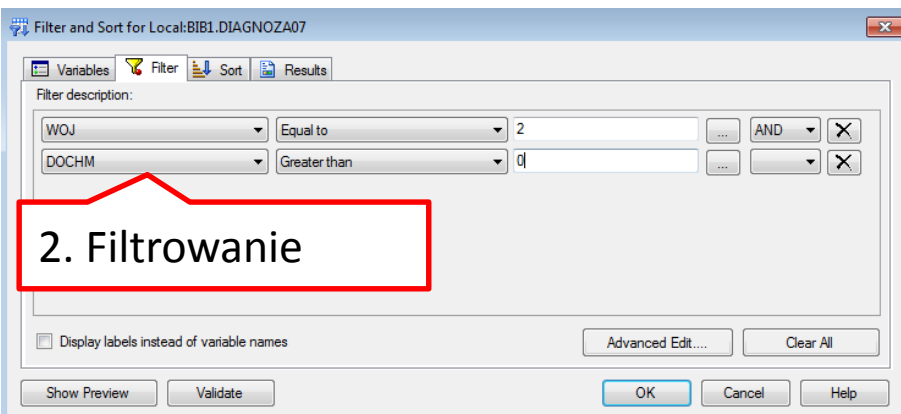
DIAGNOZA07

	ID_GOSP	KMZ	WOJ	POW
1	1	5	2	-8
2	14	6	2	-8
3	15	1	2	-8
4	15	1	2	-8
5	24	5	2	-8
6	26	6	2	-8



Wybierz gospodarstwa z województwa dolnośląskiego osiągające dochód większy od zera

Przesortuj zbiór wynikowy po dochodach Gospodarstw domowych

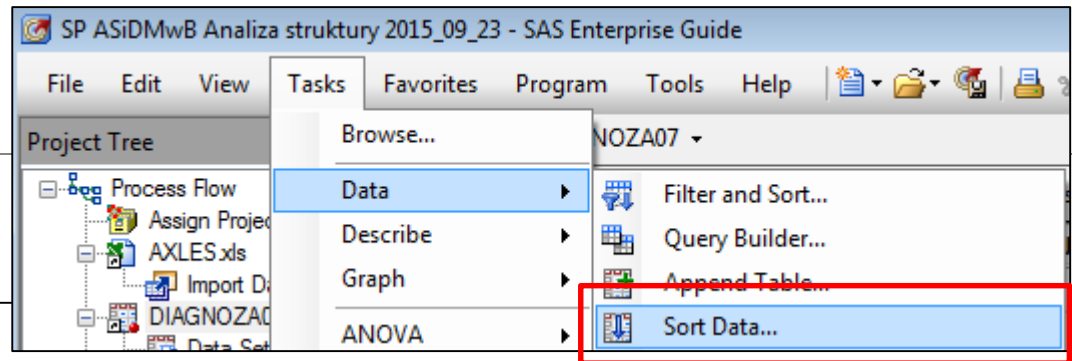


Filtrowanie i sortowanie danych (2)

Proc Step

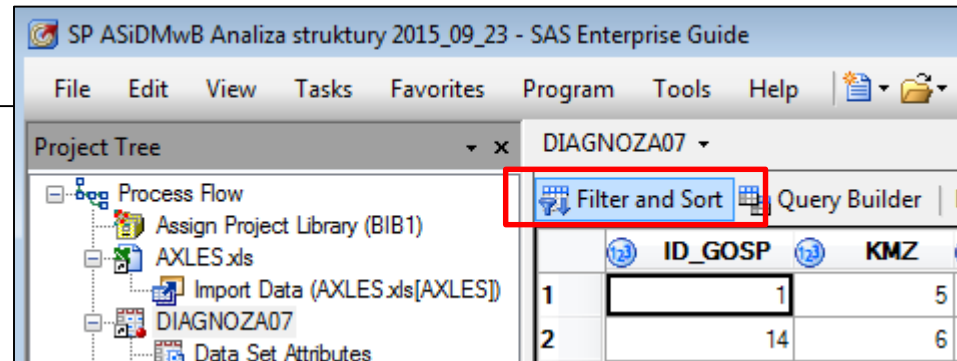
options LOCALE=Polish sortseq=Polish;

```
proc sort data=bib1.diagnoza07 (where = (woj = 2 and dochm > 0))
    out=_diagnoza07;
    by dochm;
run;
```



SQL

```
proc sql;
    create table _diagnoza07 as
    select * from bib1.diagnoza07 t1
    where t1.woj = 2 and t1.dochm > 0
    order by t1.dochm;
quit;
```



Ćwiczenie

1. Utwórz zbiór danych obejmujący gospodarstwa domowe osiągające dochód większy od 1000 zł i nie większy 3000 zł. Zbiór wynikowy uporządkuj malejąco zgodnie z wiekiem głowy gospodarstwa domowego.

5. Budowa zapytań SQL (Query Builder)

The screenshot displays the SAS Enterprise Guide interface. The 'Project Tree' on the left shows a project named 'DIAGNOZA07' with several steps, including 'Import Data (AXLES.xls[AXLES])' and 'DIAGNOZA07'. The 'Query Builder' window is open, showing a query named 'Query Builder' with an output name 'WORK.QUERY_FOR_DIAGNOZA07'. The 'Computed Columns' list on the left includes 'ID_GOSP', 'KMZ', 'WOJ', 'POW', 'ROK', 'PLEC', 'SC', 'EDU', 'LEDU', 'PJ', 'TEL', 'HTYG', 'WYMP', 'BEZR', 'STAZP', and 'DOCHM'. The 'Select Data' tab is active, showing a table with columns 'ID_GOSP', 'KMZ', and 'WOJ'. The 'Filter Data' and 'Sort Data' tabs are also visible. The 'Column Name', 'Source C...', 'S...', and 'Format' columns are present in the table. The 'Drop a column here to add it to the query.' message is displayed in the center. The 'Run', 'Save and Close', 'Cancel', and 'Help' buttons are at the bottom.

Biblioteka i nazwa zbioru wynikowego

Łączenie zbiorów danych

Filtrowanie i sortowanie danych

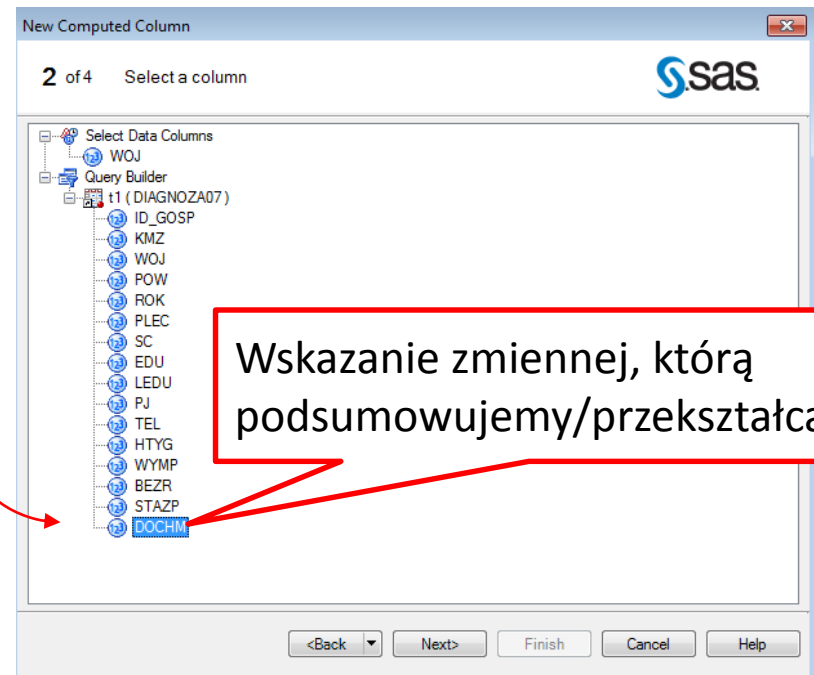
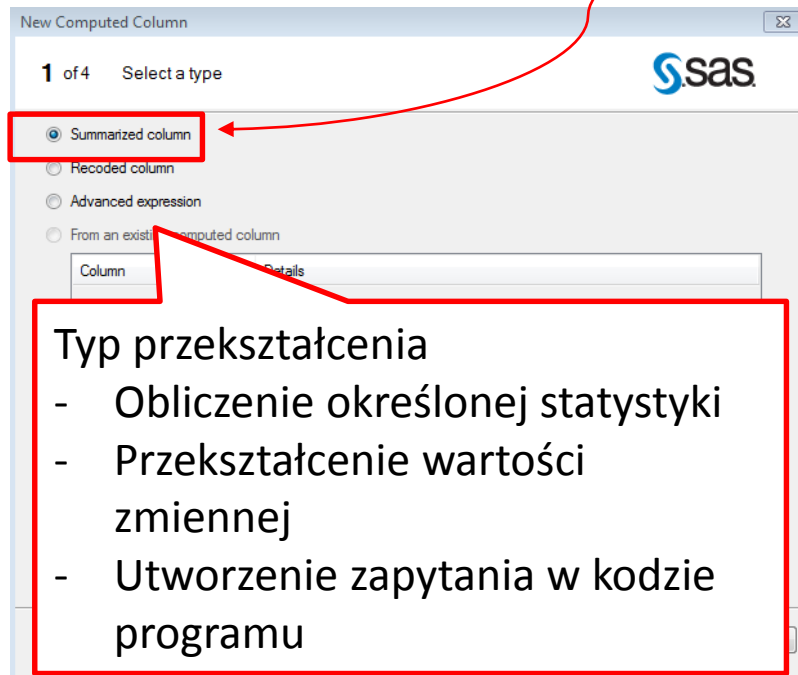
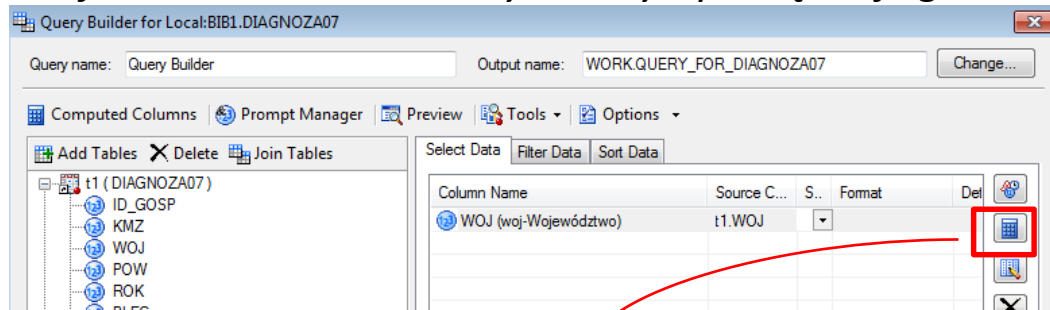
Zmienne w zbiorze wejściowym

Zmienne w zbiorze wynikowym

Tworzenie nowych zmiennych

Zapytanie SQL [1]

Utwórz zbiór danych zawierających sumę dochodów oraz średni dochód w poszczególnych województwach. Zbiór wynikowy uporządkuj zgodnie z numerem województwa.



Zapytanie SQL [2]

New Computed Column

3 of 4 Modify additional options

Column Name: Suma

Label: Suma dochodów

Summary: SUM

Expression: t1.DOCHM

Format: Change...

<Back Next> Finish Cancel Help

Nazwa nowej zmiennej oraz
wskazanie statystyki, która ma
zostać obliczona

New Computed Column

4 of 4 Summary of properties

Column Name: Suma
Label: Suma dochodów
Format: Default
Length: Default
Summary: SUM
Expression: None

<Back Next> Finish Cancel

Zbiór wynikowy Z1A

	WOJ	Sum	Mean
1	2	124147	1149.5092593
2	4	133319	1666.4875
3	6	123812	1629.1052632
4	8	66102	1120.3728814
5	10	137828	1498.1304348
6	12	207184	1801.6
7	14	353451	2155.1890244
8	16	88135	1335.3787879
9	18	113500	1233.6956522
10	20	90786	1592.7368421
11	22	177789	2020.3295455
12	24	267864	1750.745098
13	26	96409	1161.5542169
14	28	90130	1201.7333333
15	30	205716	1700.1322314
16	32	102283	1440.6056338

Query Builder for Local:BIB1.DIAGNOZA07

Query name: Query Builder Output name: WORK.QUERY_FOR_DIAGNOZA07

Computed Columns Prompt Manager Preview Tools Options

Add Tables X Delete

t1 (DIAGNOZA07)

- ID_GOSP
- KMZ
- WOJ
- POW
- ROK
- PLEC
- SC
- EDU
- LEDU
- PJ
- TEL
- HTYG
- WYMP
- BEZR
- STAZP
- DOCHM

Computed Columns

- Sum
- Mean

Select Data Filter Data Sort Data

Column Name	Source C...	S...	Format
WOJ (woj-Województwo)	t1.WOJ		
Sum	Computed	S...	
Mean	Computed	M..	

Summary groups

☒ Automatically select groups

t1.WOJ

☐ Select distinct rows only

Run Save and Close Cancel Help

Zapytanie SQL *Advanced expression*

Utwórz zmienną tekstową zawierającą wartość „Unknown” dla nieznanej liczby lat edukacji głowy gospodarstwa domowego oraz wartość „x Years of Education”, gdzie x jest liczbą lat edukacji, dla znanej liczby lat edukacji.

New Computed Column

1 of 4 Select a type

☐ Summarized column
☐ Recoded column
☒ Advanced expression
☐ From an existing computed column

Column Details

☐ Convert to an advanced expression

<Back Next> Finish Cancel Help

Edit Computed Column

1 of 2 Build an advanced expression

Enter an expression:

```
case
when LEDU < 0 then "Unknown"
when LEDU > 0 then trim(left(put(LEDU.best.))) || " Years of Education"
end
```

Home Next Back End Undo Redo Edit Favorites Validate

+ - * / ** || (x) 'x' "x" , 'abc'n

Functions
 Tables
 t1 (DIAGNOZA07)
 Selected Columns
 Computed Columns

<Back Next> Finish Cancel Help

Ćwiczenie

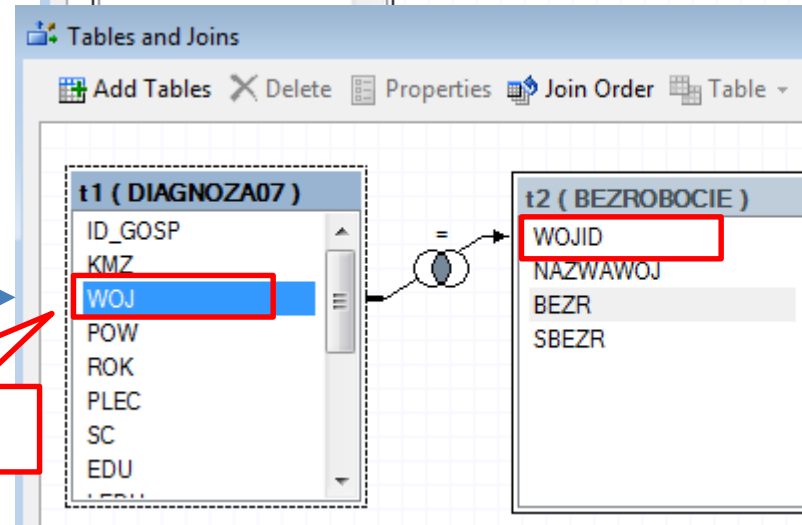
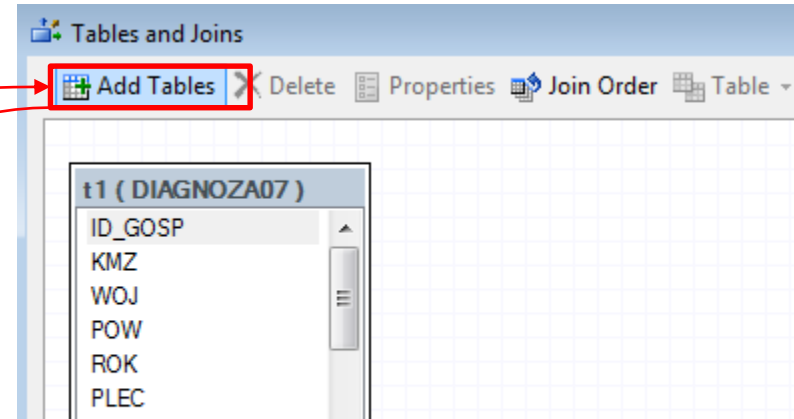
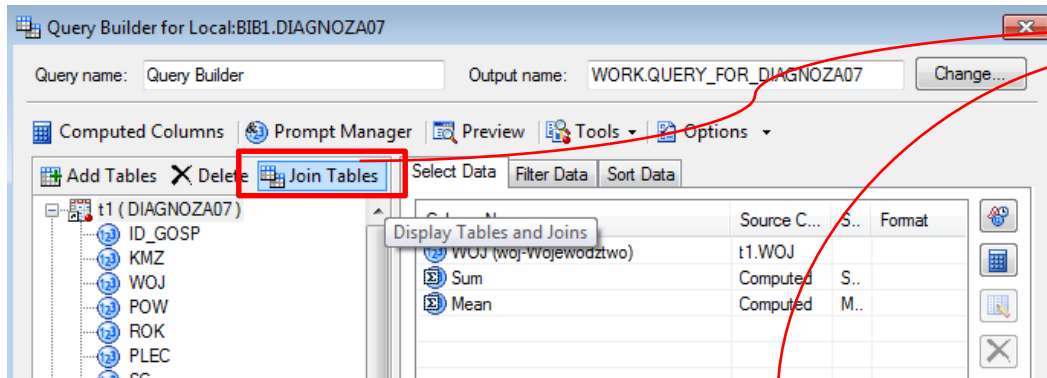
1. Utwórz zbiór danych zawierający maksymalny dochód w poszczególnych województwach.

6. Łączenie zbiorów danych [1]

Połącz zbiory danych „Z1A” oraz „bezrobocie”, tak by w zbiór wynikowy zawierał Pełną nazwę województwa, przeciętny dochód w każdym z województw oraz stopę bezrobocia.

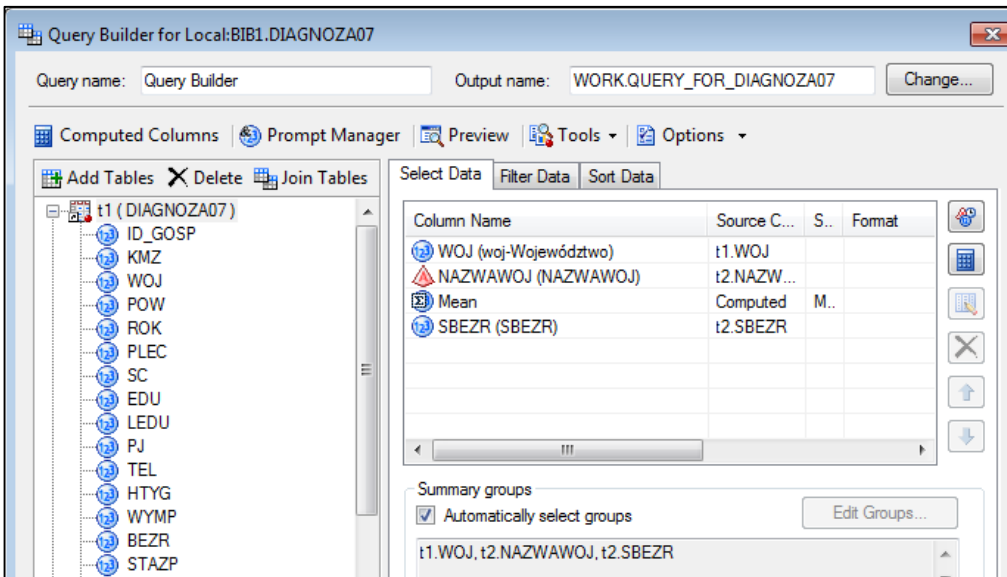


bezrobocie



Zmienna klucz

Łączenie zbiorów danych [2]



Zbiór wynikowy Z1B

	WOJ	NAZWAWOJ	Mean	SBEZR
1	2	DOLNOŚLĄSKIE	1149.5092593	11.4
2	4	KUJAWSKO-POMORSKIE	1666.4875	14.9
3	6	LUBELSKIE	1629.1052632	13
4	8	LUBUSKIE	1120.3728814	14
5	10	ŁÓDZKIE	1498.1304348	11.2
6	12	MAŁOPOLSKIE	1801.6	8.7
	14	MAZOWIECKIE	2155.1890244	9
	16	OPOLSKIE	1335.3787879	11.9
	18	PODKARPACKIE	1233.6956522	14.2
	20	PODLASKIE	1592.7368421	10.4
	22	POMORSKIE	2020.3295455	10.7
	24	ŚLĄSKIE	1750.745098	9.2
	26	ŚWIĘTOKRZYSKIE	1161.5542169	14.9
	28	WARMIŃSKO-MAZURSKIE	1201.7333333	18.7
	30	WIELKOPOLSKIE	1700.1322314	7.8
	32	ZACHODNIOPOMORSKIE	1440.6056338	16.4

```
proc sql;
  create table z1b as
  select t1.woj
    ,t2.nazwawoj
    ,(mean(t1.dochm)) label="mean" as mean
    ,t2.sbezr
  from bib1.diagnoza07 t1
  inner join zbiory.bezrobocie t2 on (t1.woj =
t2.wojid)
  group by t1.woj,t2.nazwawoj,t2.sbezr;
quit;
```

Warunkowe łączenie tabel [1]

Zbiory „P1” i „P2” zawierają dane o wyniku pomiaru pewnych parametrów dokonane w kolejnych wizytach. Zadanie polega na złączeniu tabel w taki sposób, aby każdy wiersz zawierał informację o wielkości parametru p1 i parametru p2 otrzymanej w danym dniu dla danej jednostki. W wyniku zachowaj wszystkie wiersze ze zbioru „P1”.

Zbiór P1

	id	wizyta	data	wynik
1	1	1	11OCT2015	5
2	1	2	12OCT2015	6
3	1	3	13OCT2015	4
4	1	4	14OCT2015	3
5	1	5	15OCT2015	1
6	1	99	17OCT2015	1

Zbiór P2

	id	wizyta	data	wynik
1	1	1	11OCT2015	4
2	1	2	12OCT2015	5
3	1	3	13OCT2015	3
4	1	.	14OCT2015	2
5	1	5	15OCT2015	3

Łączymy zbiory po numerze jednostki, numerze wizyty i dacie. Jeżeli numer wizyty jest brakujący łączymy zbiory z pominięciem numeru wizyty.

```
proc sql;
  create table param as
  select
    a.*,b.wynik as wynik_b,b.wizyta as wizyta_b
  from p1 a left join p2 b
  on (a.id = b.id and
    ((nmiss(a.wizyta,b.wizyta)=0 and a.wizyta=b.wizyta and a.data=b.data) or
    (nmiss(a.wizyta,b.wizyta)>0 and a.data=b.data)))
  order by a.wizyta;
quit;
```

Warunkowe łączenie tabel [2]

a) `from p1 a left join p2 b on (a.id = b.id and a.wizyta=b.wizyta)`

	id	wizyta	data	wynik	wynik_b	wizyta_b
1	1	1	11OCT2015	5	4	1
2	1	2	12OCT2015	6	5	2
3	1	3	13OCT2015	4	3	3
4	1	5	15OCT2015	1	3	5

Brak wyniku dla 14 października 2015

b) `from p1 a left join p2 b
on (a.id = b.id and ((nmiss(a.wizyta,b.wizyta)=0 and a.wizyta=b.wizyta and
a.data=b.data) or (nmiss(a.wizyta,b.wizyta)>0 and a.data=b.data)))`

	id	wizyta	data	wynik	wynik_b	wizyta_b
1	1	1	11OCT2015	5	4	1
2	1	2	12OCT2015	6	5	2
3	1	3	13OCT2015	4	3	3
4	1	4	14OCT2015	3	2	.
5	1	5	15OCT2015	1	3	5
6	1	99	17OCT2015	1	.	.



c) `from p1 a inner join p2 b`

	id	wizyta	data	wynik	wynik_b	wizyta_b
1	1	1	11OCT2015	5	4	1
2	1	2	12OCT2015	6	5	2
3	1	3	13OCT2015	4	3	3
4	1	4	14OCT2015	3	2	.
5	1	5	15OCT2015	1	3	5

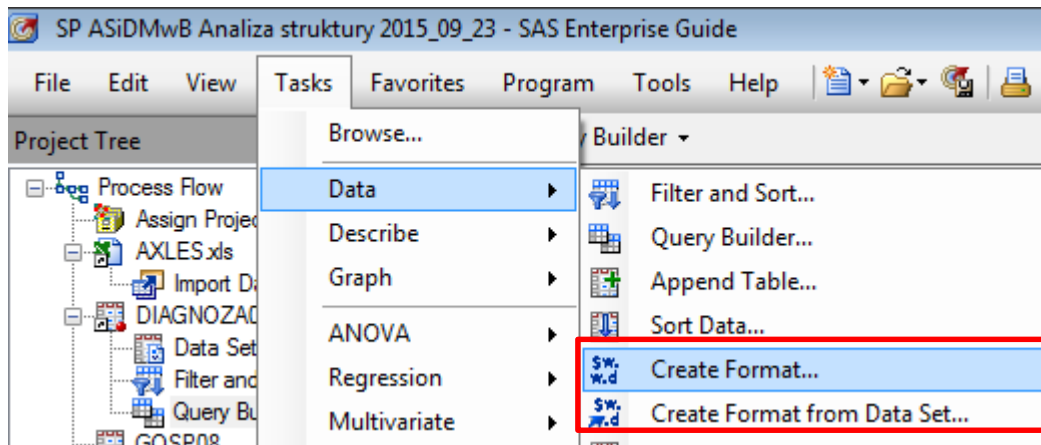
Brak wizyty 99

Ćwiczenie

1. Połącz zbiór danych „*Diagnoza07*” ze zbiorem „*Z1B*”. W zbiorze wynikowym zachowaj wszystkie wiersze ze zbioru „*Diagnoza07*”.

7. Formaty

W zestawieniu zawierającym średni miesięczny dochód gospodarstw domowych (zbiór „Z1B”) uwzględnij klasę miejscowości zamieszkania (KMZ). Sformatuj otrzymany zbiór danych w taki sposób, aby województwo oraz klasa miejscowości zamieszkania były wyświetlane w postaci pełnej nazw, natomiast średni dochód był podany w złotych.



Zbiór Z1B po uzupełnieniu
 o zmienną KMZ

	WOJ	KMZ	Mean
1	2	1	863.9
2	2	3	1156.1
3	2	4	1081.2727273
4	2	5	1227
5	2	6	1300.8947368
6	4	2	1542.3333333
7	4	3	456.5
8	4	4	1400.8333333
9	4	5	1847.8333333
10	4	6	1904.3235294
11	6	2	2470.375
12	6	4	1058
13	6	5	1212.25
14	6	6	1601.6666667

Tworzenie formatu *Create format*

Typ formatu

Nazwa formatu

Format name: klmzfmt

Format type: ☒ Numeric

Specify format width: ☐

Specify fuzz factor: ☐

Fuzz factor: 1E-12

Server: Local

Format catalog locations: EG TASK

Currently assigned libraries: WORK (Temporary)

Type a name for the format that you are creating. This name will be used to refer to the format each time you apply it. For more information about specifying a format name, see the online help.

Preview code Run Save Cancel Help

Please enter a valid format name. Information on format names can be found in the help.

Define Formats

Format definition:

Label	Ranges
miasto >= 500 tys.	1
miasto 200-500 tys.)	2
miasto (100-200 tys.)	3
miasto [20-100 tys.)	4
miasto > 20 tys.	5
miasto wieś	6

Range definitions for "miasto >= 500 tys.":

Type	Values
Discrete	1

New Delete

Run Save Cancel Help

Przypisanie wartościom zmiennej KZM etykiet (formatów)

Query Builder for Local:BIB1.DIAGNOZA07

Query name: Query Builder Output name: WORKQUERY_FOR_DIAGNOZA07

Computed Columns: ID_GOSP, KMZ, WOJ, POW, ROK, PLEC, SC, EDU, LEDU, PJ, TEL, HTYG, WYMP, BEZR, STAZP, DOCHM

Sum Mean

Properties for KMZ (klmz-klasa miejscowości zamieszkania we wszystkich próbach)

Column Name: klmz-klasa miejscowości zamieszkania we wszystkich próbach

Label: klmz-klasa miejscowości zamieszkania we wszystkich próbach

Format: **Change...**

Summary: **Formats**

Categories	Formats
None	DOCHFMT.
Numeric	KMZFMF.
Date	PLECFMT.
Time	WOJFMT.
Date/Time	

Przypisanie formatu zmiennej KMZ za pomocą narzędzia Query Builder

Wynik formatowania

KMZ
miasto >= 500 tys.
miasto (100-200 tys.)
miasto [20-100 tys.)
miasto > 20 tys.
miasto wieś

Tworzenie formatu na podstawie zbioru danych

Create format from Data Set

	WOJID	NAZWAWOJ	BEZR	SBEZR
1	2	DOLNOŚLĄSKIE	127.5	11.4
2	4	KUJAWSKO-POMORSKIE		
3	6	LUBELSKIE		
4	8	LUBUSKIE		
5	10	ŁÓDZKIE		
6	12	MAŁOPOLSKIE		
7	14	MAZOWIECKIE		
8	16	OPOLSKIE		
9	18	PODKARPACKIE		
10	20	PODLASKIE		
11	22	POMORSKIE		
12	24	ŚLĄSKIE		
13	26	ŚWIĘTOKRZYSKIE		
14	28	WARMIŃSKO-MAZURSKIE		
15	30	WIELKOPOLSKIE		
16	32	ZACHODNIOPOMORSKIE		

Create Format from Data Set

Source data set: Local: ZBIORY.BEZROBOCIE

Format name: wojfmt

Format type: Numeric

Output Libraries

☐ Format catalog locations WORK

☒ Currently assigned libraries WORK (temporary)

Value types

☒ Discrete/Look up

☐ Ranges

Variables

Discrete values: WOJID

Labels: NAZWAWOJ

Maximum label length: 30

☐ Specify label for other values

Label: Other

☐ Generate a summary report

Run Save Cancel Help

Look Up Value	Label
2	DOLNOŚLĄSKIE
4	KUJAWSKO-POMORSKIE
6	LUBELSKIE
8	LUBUSKIE
10	ŁÓDZKIE
12	MAŁOPOLSKIE
14	MAZOWIECKIE
16	OPOLSKIE
18	PODKARPACKIE
20	PODLASKIE
22	POMORSKIE
24	ŚLĄSKIE
26	ŚWIĘTOKRZYSKIE
28	WARMIŃSKO-MAZURSKIE
30	WIELKOPOLSKIE
32	ZACHODNIOPOMORSKIE

Formaty walutowe

Formats

Categories: None, Numeric, Date, Time, Date/Time, **Currency**, User Defined, All

Formats: NLMNLMXNw.d, NLMNLMYRw.d, NLMNLNOKw.d, NLMNLNZDw.d, **NLMNLRURw.d**, NLMNLRPLNw.d, NLMNLRRLw.d, NLMNLRUBw.d, NLMNLRURw.d

Attributes
Overall width: 12 Min: 8 Max: 32
Decimal places: 2 Min: 0 Max: 11

Description
format waluty Polski

Example
Value: 123.1
Output: 123.10 zł

OK Cancel

Zbiór Z1C

	WOJ	KMZ	Mean
1	DOLNOŚLĄSKIE	miasto >= 500 tys.	863,90 zł
2	DOLNOŚLĄSKIE	miasto (100-200 tys)	1 156,10 zł
3	DOLNOŚLĄSKIE	miasto (20-100 tys.)	1 081,27 zł
4	DOLNOŚLĄSKIE	miasto > 20 tys.	1 227,00 zł
5	DOLNOŚLĄSKIE	miasto wieś	1 300,89 zł
6	KUJAWSKO-POMORSKIE	miasto 200-500 tys.)	1 542,33 zł
7	KUJAWSKO-POMORSKIE	miasto (100-200 tys)	456,50 zł
8	KUJAWSKO-POMORSKIE	miasto (20-100 tys.)	1 400,83 zł
9	KUJAWSKO-POMORSKIE	miasto > 20 tys.	1 847,83 zł
10	KUJAWSKO-POMORSKIE	miasto wieś	1 904,32 zł

Formaty warunkowe

```
proc format lib=work;
    value dochfmt
        /*LOW - 2000 = orange*/
        2000 <- HIGH = green;
;
run;
```

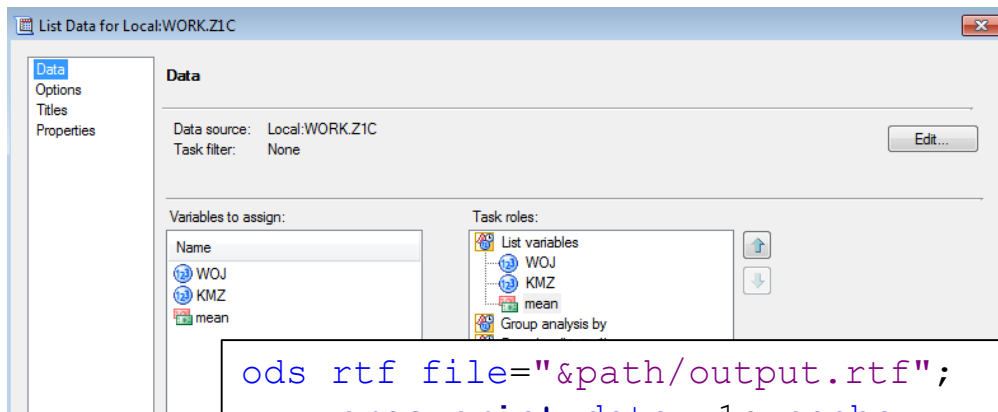
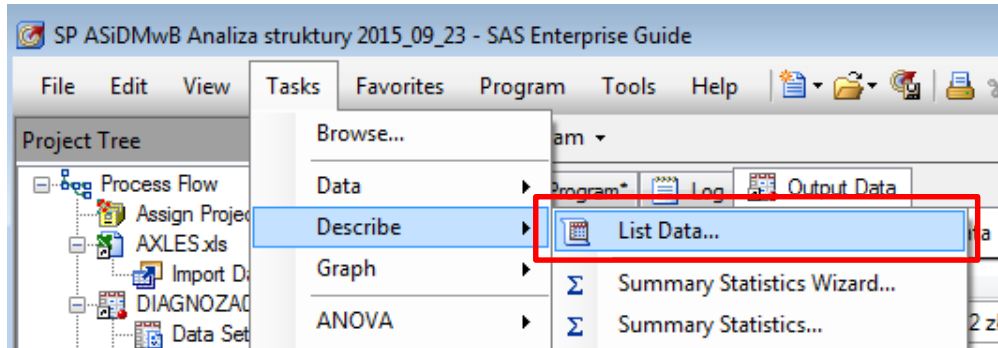
Wartości wyższe od
2000 zaznacz
kolorem zielonym

Ćwiczenie

1. Utwórz odpowiedni format a następnie sformatuj zmienną PLEC ze zbioru „*diagonza07*” w taki sposób, aby w zbiorze wyświetlane były pełne nazwy płci.

8. Raportowanie

Utwórz raport HTML i RTF na podstawie zbioru „Z1C”.



WOJ	KMZ	mean
DOLNOŚLĄSKIE	miasto >= 500 tys.	863,90 zł
DOLNOŚLĄSKIE	miasto [100-200 tys.)	1 156,10 zł
DOLNOŚLĄSKIE	miasto [20-100 tys.)	1 081,27 zł
DOLNOŚLĄSKIE	miasto > 20 tys.	1 227,00 zł
DOLNOŚLĄSKIE	miasto wieś	1 300,89 zł
KUJAWSKO-POMORSKIE	miasto 200-500 tys.)	1 542,33 zł
KUJAWSKO-POMORSKIE	miasto [100-200 tys.)	456,50 zł
KUJAWSKO-POMORSKIE	miasto [20-100 tys.)	1 400,83 zł
KUJAWSKO-POMORSKIE	miasto > 20 tys.	1 847,83 zł
KUJAWSKO-POMORSKIE	miasto wieś	1 904,32 zł
LUBELSKIE	miasto 200-500 tys.)	2 470,38 zł
LUBELSKIE	miasto [20-100 tys.)	1 058,00 zł
LUBELSKIE	miasto > 20 tys.	1 212,25 zł
LUBELSKIE	miasto wieś	1 601,67 zł

```
ods rtf file="&path/output.rtf";
proc print data=z1c noobs;
    var woj kmz;
    var mean / style (data) = [background=dochfmt.];
run;
ods rtf close;
```

Wybrane opcje definiujące wygląd raportu

```
options nonumber nodate;
ods rtf file="&path\out.rtf" ;
title justify=center 'Dochody gospodarstw domowych';
footnote justify=left "Zestawienie z dnia %SYSFUNC(today()), yymmddp10.";

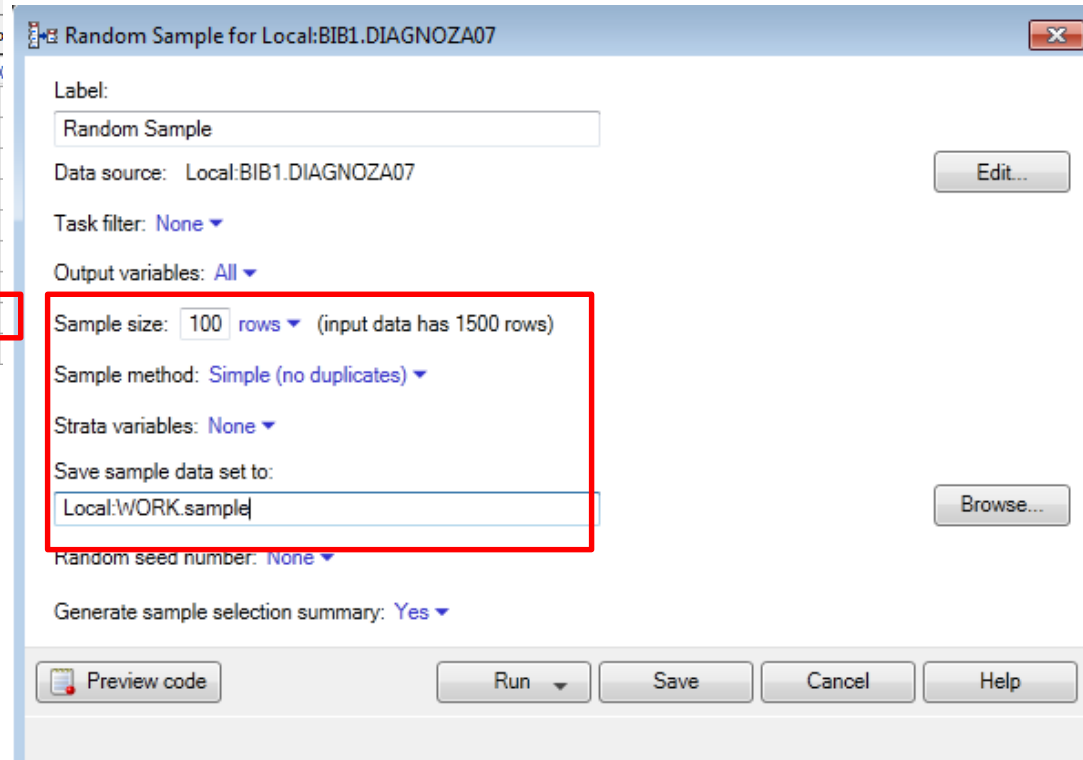
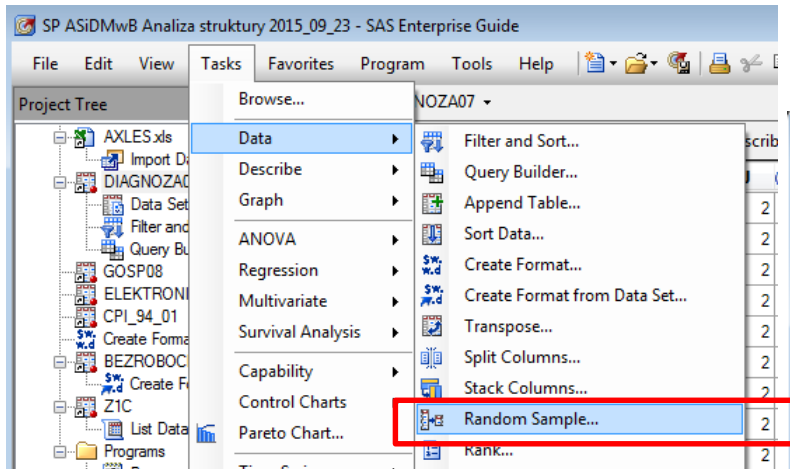
proc print data=z1c style(header)=[background=yellow] label noobs;
  var WOJ / style (data) = [just=left];
  var MEAN / style (data) = [background=dochfmt.];
  var KMZ / style (header) = [font_style=italic foreground=red]
  style(data) = [font_style=italic foreground=red background=none];
run;
```

Dochody gospodarstw domowych

woj-Województwo	mean	klmz-klasa miejscowości zamieszkania we wszystkich próbach
DOLNOŚLĄSKIE	863,90 zł	miasto >= 500 tys.
DOLNOŚLĄSKIE	1 156,10 zł	miasto [100-200 tys)
DOLNOŚLĄSKIE	1 081,27 zł	miasto [20-100 tys.)
DOLNOŚLĄSKIE	1 227,00 zł	miasto > 20 tys.
DOLNOŚLĄSKIE	1 300,89 zł	miasto wieś
KUJAWSKO-POMORSKIE	1 542,33 zł	miasto 200-500 tys.)
KUJAWSKO-POMORSKIE	456,50 zł	miasto [100-200 tys)
KUJAWSKO-POMORSKIE	1 400,83 zł	miasto [20-100 tys.)
KUJAWSKO-POMORSKIE	1 847,83 zł	miasto > 20 tys.
KUJAWSKO-POMORSKIE	1 904,32 zł	miasto wieś
LUBELSKIE	2 470,38 zł	miasto 200-500 tys.)
LUBELSKIE	1 058,00 zł	miasto [20-100 tys.)

Próbkowanie

Pobierz próbę losową prostą liczącą $n=100$ obserwacji ze zbioru „Diagnoza07”.

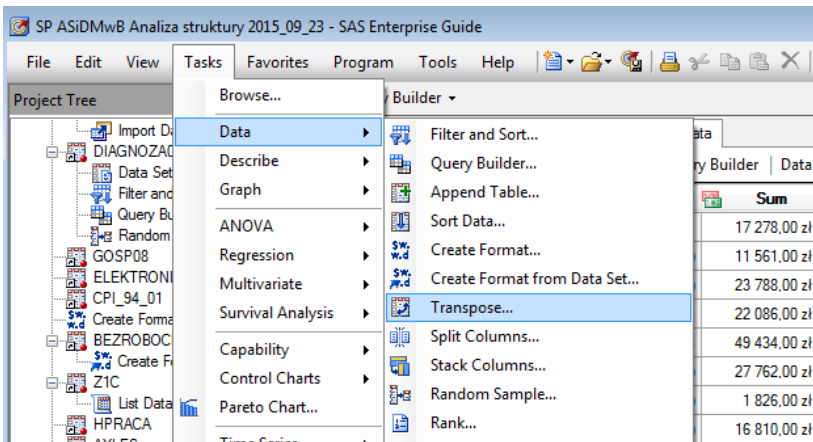


```
proc surveyselect
data=bib1.diagnoza07
out=sample
method=srs
n=100;
run;
```

Ćwiczenie

1. Wylosuj 1500 obserwacji ze zbioru „*gosp_2007*” zgodnie ze schematem losowania prostego.

9. Transpozycja zbioru danych



```
proc sql;
  create table zld as
  select t1.woj format=wojfmt.,
  t1.kmz format=kmzfmt.,
  (sum(t1.dochm)) format=nlmnlpn12.2 as sum
  from bib1.diagnoza07 t1
  group by t1.woj,
  t1.kmz;
quit;
```

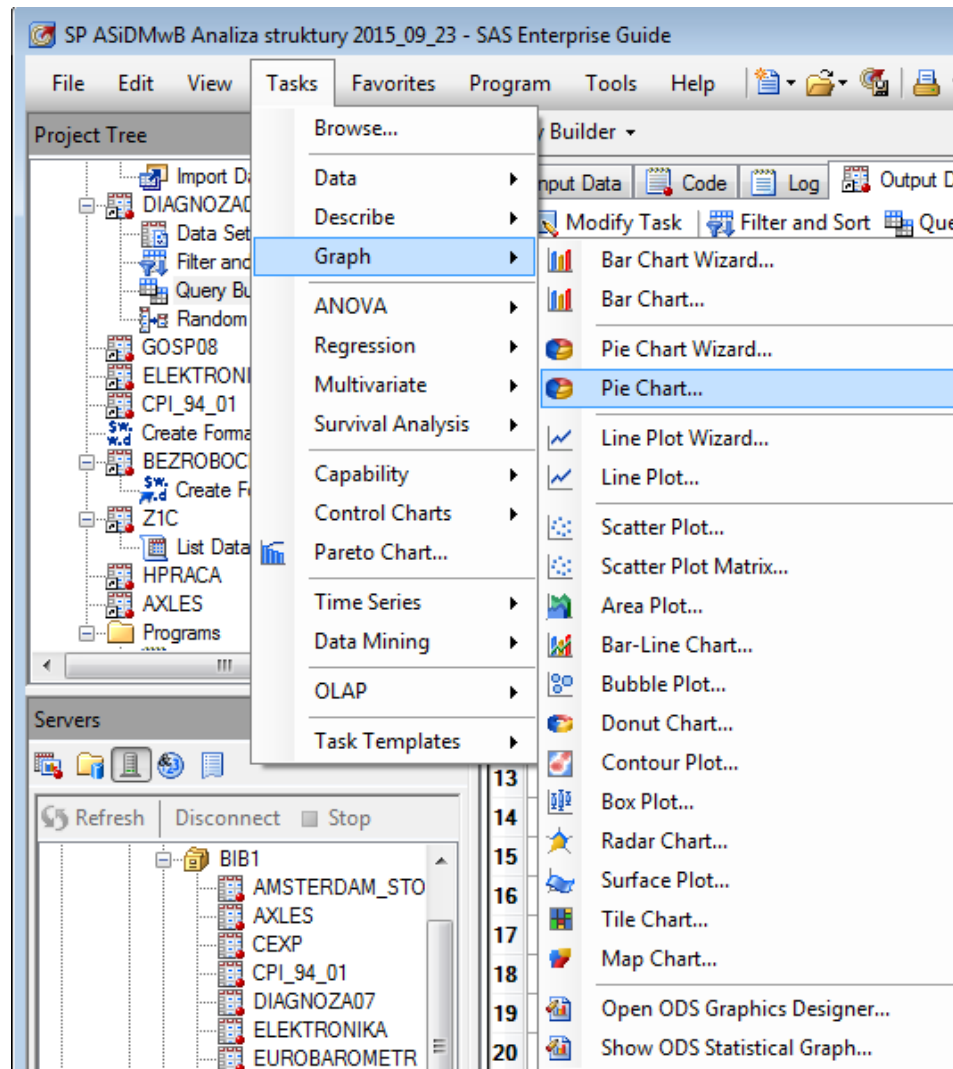
Suma dochodów w poszczególnych województwach i klasach miejscowości zamieszkania

```
proc transpose data=zld out=zld_t prefix=column;
  by woj;
  var sum;
run;
```

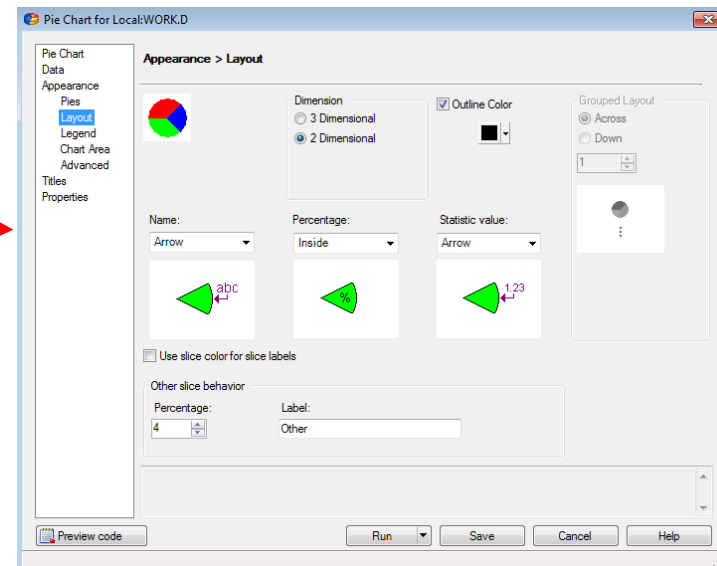
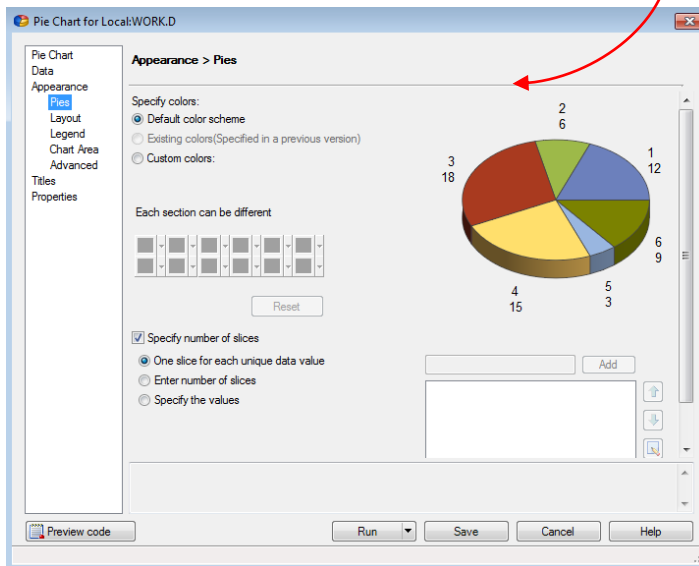
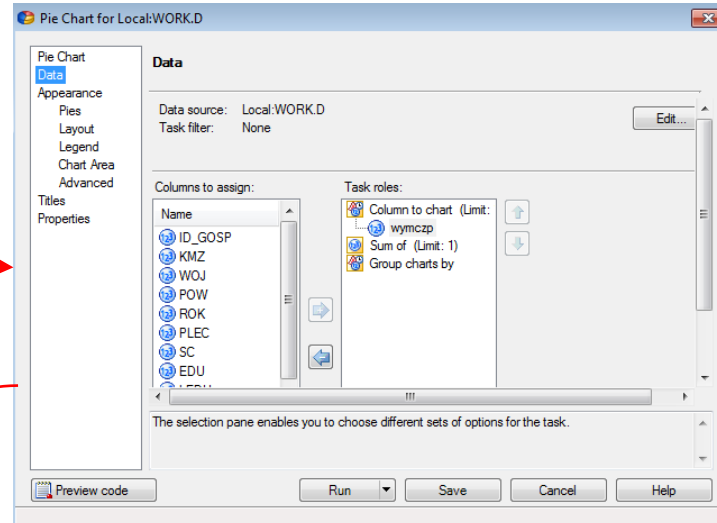
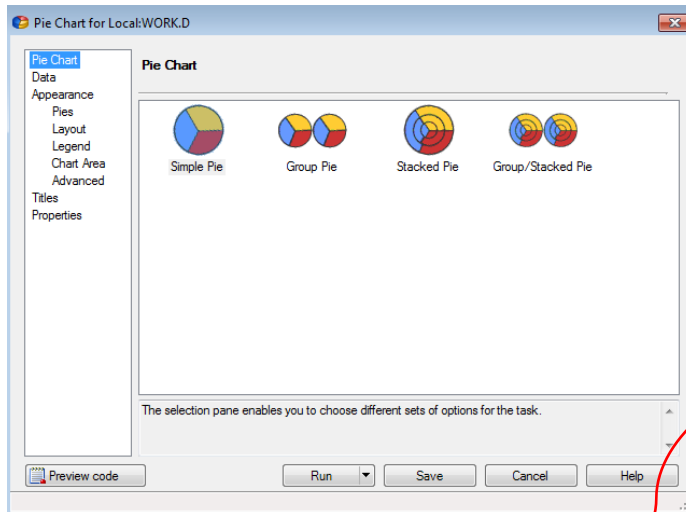
	WOJ	KMZ	Sum
1	DOLNOŚLĄSKIE	miasto >= 500 tys	17 278,00 zł
2	DOLNOŚLĄSKIE	miasto [100-200 tys	11 561,00 zł
3	DOLNOŚLĄSKIE	miasto [20-100 tys.	23 788,00 zł
4	DOLNOŚLĄSKIE	miasto > 20 tys	22 086,00 zł
5	DOLNOŚLĄSKIE	miasto wieś	49 434,00 zł
6	KUJAWSKO-POMORSKIE	miasto 200-500 tys.)	27 762,00 zł
7	KUJAWSKO-POMORSKIE	miasto [100-200 tys)	1 826,00 zł
8	KUJAWSKO-POMORSKIE	miasto [20-100 tys.)	16 810,00 zł
9	KUJAWSKO-POMORSKIE	miasto > 20 tys..	22 174,00 zł
10	KUJAWSKO-POMORSKIE	miasto wieś	64 747,00 zł
11	LUBELSKIE	miasto 200-500 tys.)	39 526,00 zł
12	LUBELSKIE	miasto [20-100 tys.)	16 928,00 zł
13	LUBELSKIE	miasto > 20 tys..	9 698,00 zł
14	LUBELSKIE	miasto wieś	57 660,00 zł

	WOJ	Source	Label	Column1	Column2	Column3	Column4	Column5	Column6
1	DOLNOŚLĄSKIE	Sum	Suma dochod...	17 278,00 zł	11 561,00 zł	23 788,00 zł	22 086,00 zł	49 434,00 zł	
2	KUJAWSKO-POMORSKIE	Sum	Suma dochod...	27 762,00 zł	1 826,00 zł	16 810,00 zł	22 174,00 zł	64 747,00 zł	
3	LUBELSKIE	Sum	Suma dochod...	39 526,00 zł	16 928,00 zł	9 698,00 zł	57 660,00 zł		
4	LUBUSKIE	Sum	Suma dochod...	4 904,00 zł	3 168,00 zł	22 364,00 zł	35 666,00 zł		
5	ŁÓDZKIE	Sum	Suma dochod...	27 450,00 zł	50 162,00 zł	12 418,00 zł	47 798,00 zł		
6	MAŁOPOLSKIE	Sum	Suma dochod...	68 886,00 zł	1 492,00 zł	14 542,00 zł	32 976,00 zł	89 288,00 zł	
7	MAZOWIECKIE	Sum	Suma dochod...	87 606,00 zł	6 932,00 zł	7 984,00 zł	27 083,00 zł	63 968,00 zł	159878,00 zł
8	OPOLSKIE	Sum	Suma dochod...	32 172,00 zł	11 926,00 zł	4 822,00 zł	39 215,00 zł		
9	PODKARPACKIE	Sum	Suma dochod...	11 350,00 zł	31 296,00 zł	15 642,00 zł	55 212,00 zł		
10	PODLASKIE	Sum	Suma dochod...	10 064,00 zł	18 342,00 zł	16 552,00 zł	45 828,00 zł		
11	POMORSKIE	Sum	Suma dochod...	54 147,00 zł	52 670,00 zł	25 490,00 zł	45 482,00 zł		
12	ŚLĄSKIE	Sum	Suma dochod...	7 192,00 zł	54 158,00 zł	54 874,00 zł	47 550,00 zł	19 880,00 zł	84 210,00 zł
13	ŚWIĘTOKRZYSKIE	Sum	Suma dochod...	22 836,00 zł	5 820,00 zł	8 565,00 zł	59 188,00 zł		
14	WARMIŃSKO-MAZURSKIE	Sum	Suma dochod...	1 400,00 zł	13 596,00 zł	8 434,00 zł	14 688,00 zł	52 012,00 zł	
15	WIELKOPOLSKIE	Sum	Suma dochod...	22 346,00 zł	7 484,00 zł	51 236,00 zł	21 096,00 zł	103554,00 zł	
16	ZACHODNIOPOMORSKIE	Sum	Suma dochod...	15 860,00 zł	3 476,00 zł	10 694,00 zł	22 522,00 zł	49 731,00 zł	

10. Wykresy



Wykres kołowy *Graph* → *Pie chart* [1]



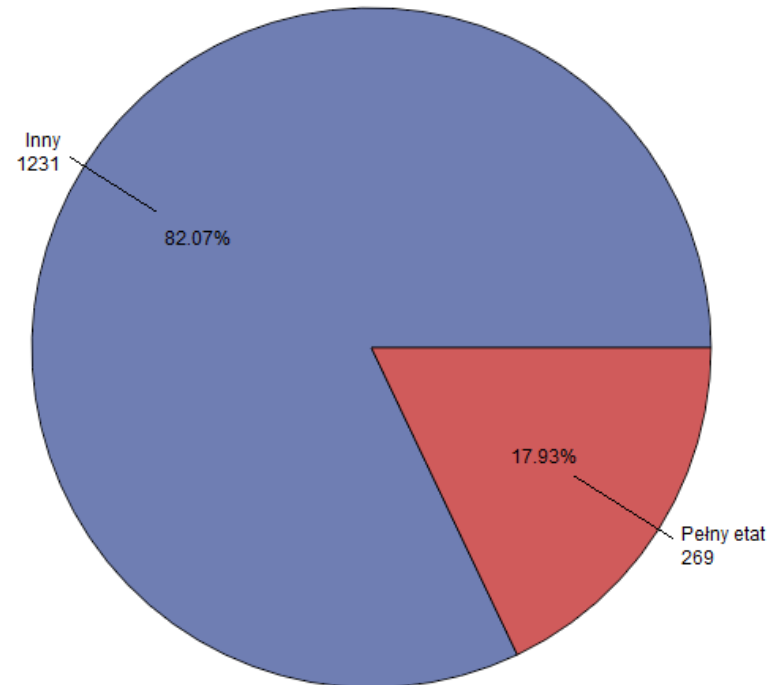
Wykres kołowy [2]

Wykres kołowy prezentujący udział pracujących na pełny i niepełny etat w próbie „Diagnoza07”.

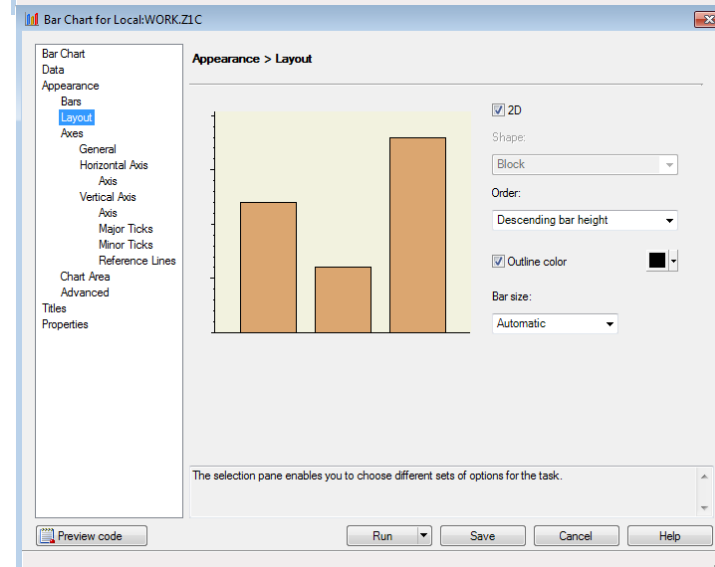
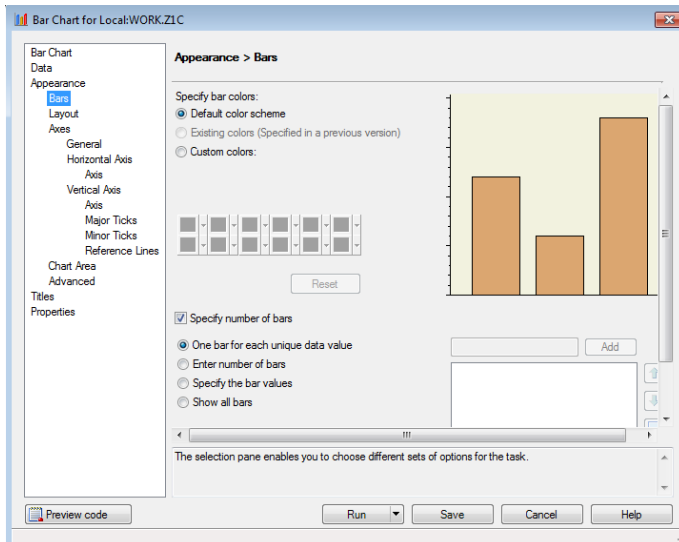
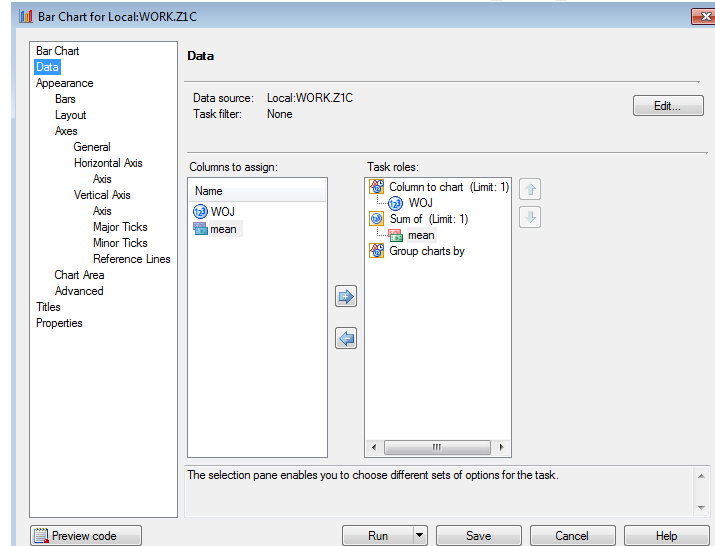
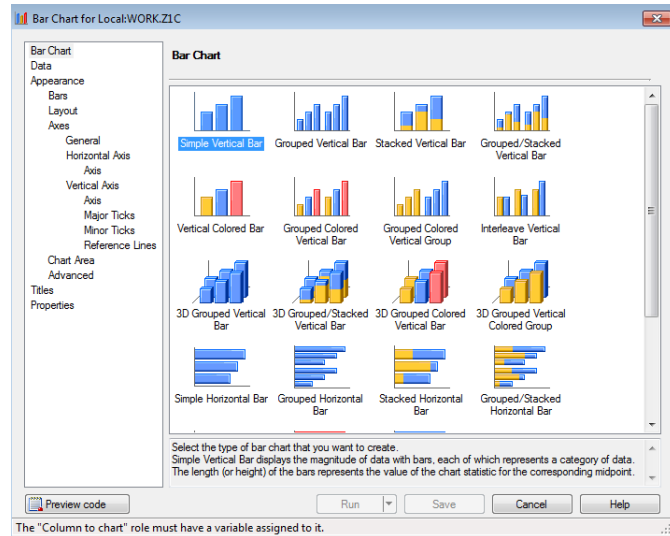
```

proc format lib=work;
    value wymczpfmt
    1="Pełny etat"
    0="Inny"
;
run;
data d;
    set bib1.diagnoza07;
    wymczp=(htyg>=40);
    format wymczp wymczpfmt.;
run;

proc gchart data=d;
    pie wymczp /
    slice=arrow /*linia do
statystyki*/
    percent=inside
    discrete;
run;
    
```



Wykres słupkowy *Graph* → *Bar Chart* [1]

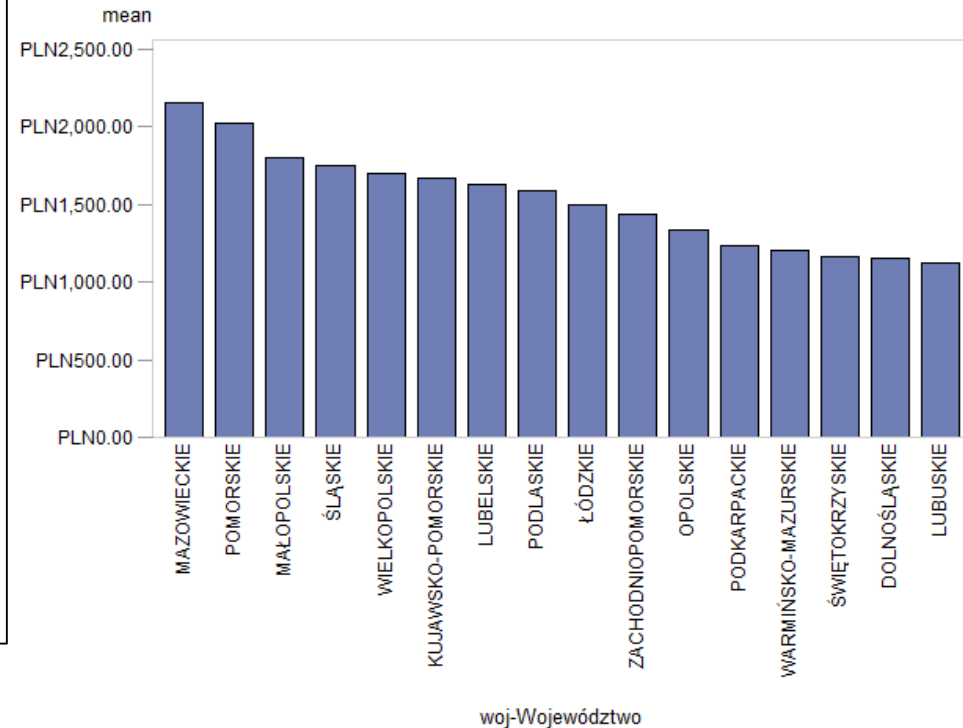


Wykres słupkowy [2]

Wykres słupkowy prezentujący średni dochód gospodarstw domowych w poszczególnych województwach otrzymany na podstawie próby „Diagnoza07”.

```

proc sql;
  create table z1c as
  select woj format=wojfmt.
  ,mean(dochm) label="mean,"
  format=NLMLPLN12.2 as mean
  from bib1.diagnoza07
  group by woj      ;
quit;
proc gchart data=z1c;
  vbar woj / sumvar=mean frame
  discrete descending
  coutline=black;
  format mean nlmnipln12.2;
run;
    
```



Wykres mapowy *Graph* → *Map chart* [1]

Łączenie ze zbiorem *MAPS.POLAND2*

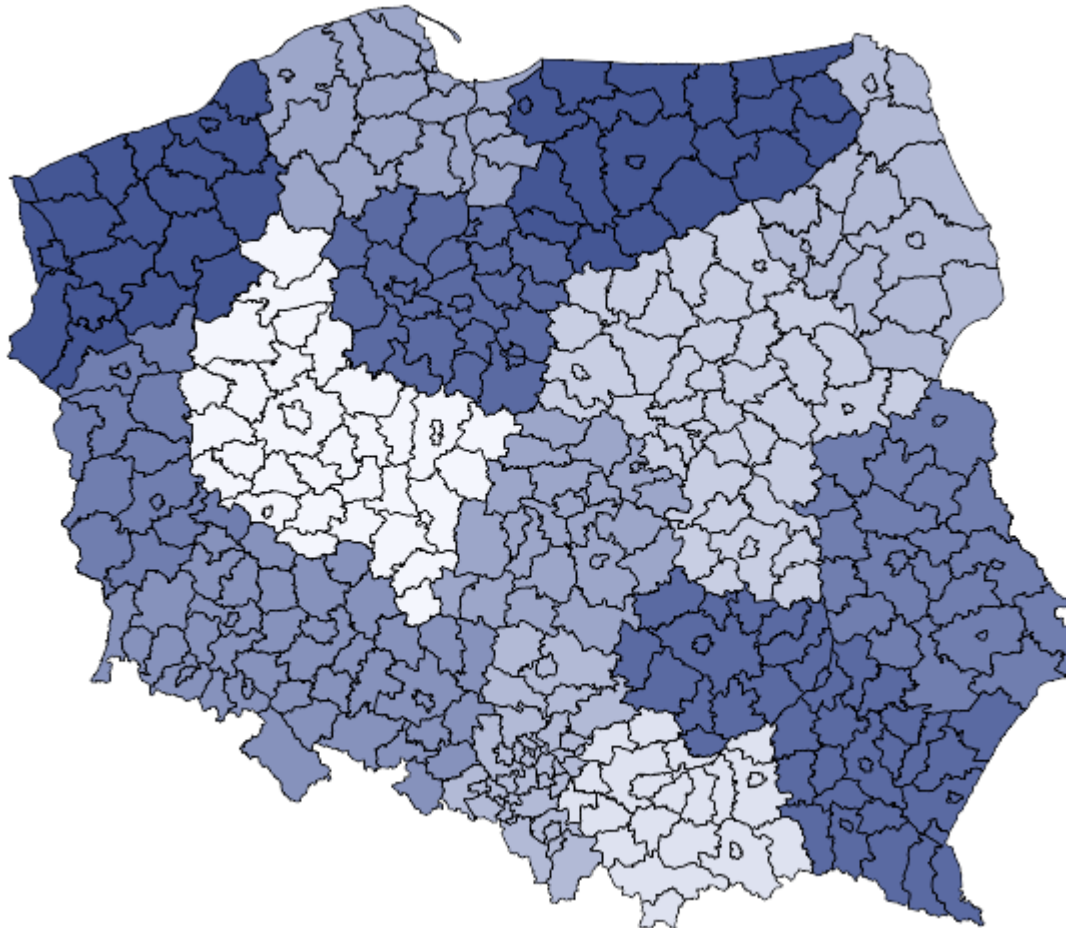
The screenshot shows the SAS Query Builder interface. At the top, a data grid displays columns: woj_id, WOJID, NAZWAWOJ, BEZR, and SBEZR. Below the grid, the 'Query Builder (2) for Local:WORK.B' window is open. It shows a tree view of tables on the left, including 't1 (B)' and 't2 (POLAND2)'. The 'Tables and Joins' window is also open, showing the join between 't1 (B)' and 't2 (POLAND2)' with a join condition. The 't1 (B)' table has columns: woj_id, WOJID, NAZWAWOJ, BEZR, SBEZR. The 't2 (POLAND2)' table has columns: _MAP_GEOMETRY_, COUNTRY, ID, IDNAME, WOJID, PROVNAME, PROVNAME2.

Łączenie ze zbiorem *MAPS.POLAND*

The screenshot shows the 'Map Chart for: WORK:QUERY_FOR_B' configuration window. The 'Data' tab is selected. The 'Map data source' is set to 'MAPS.POLAND'. The 'Response data source' is set to 'WORK.QUERYFOR_B'. The 'ID variable(s) role' section shows 'ID' assigned to the 'ID' task role. The 'Other roles' section shows 'BEZR', '_MAP_GEOMETRY_', and 'COUNTRY' assigned to the 'Response (Limit: 1)' task role, and 'SBEZR' assigned to the 'Group charts by' task role. The 'Preview code' button is at the bottom left, and 'Run', 'Save', 'Cancel', and 'Help' buttons are at the bottom right.

Wykres mapowy [2]

Wykres mapowy prezentujący stopę bezrobocia w poszczególnych województwach otrzymany na podstawie zbioru „Bezrobocie”.





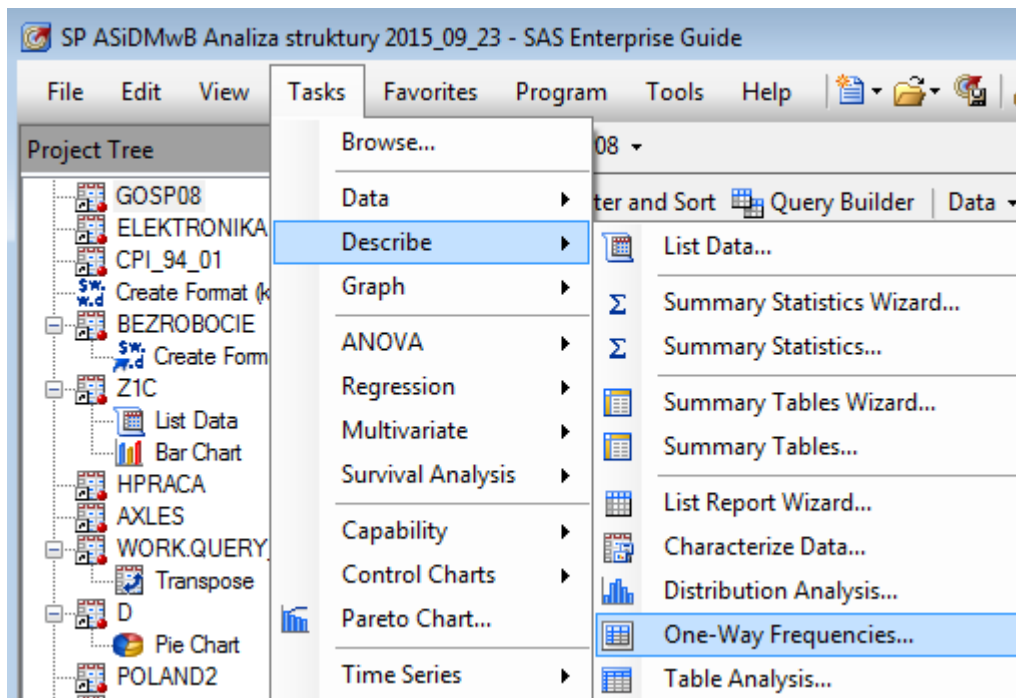
Analiza struktury

1. Analiza częstości i liczebności [1]

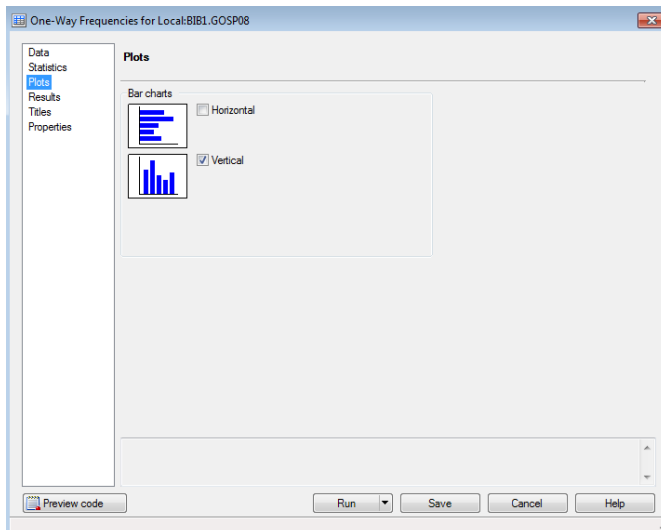
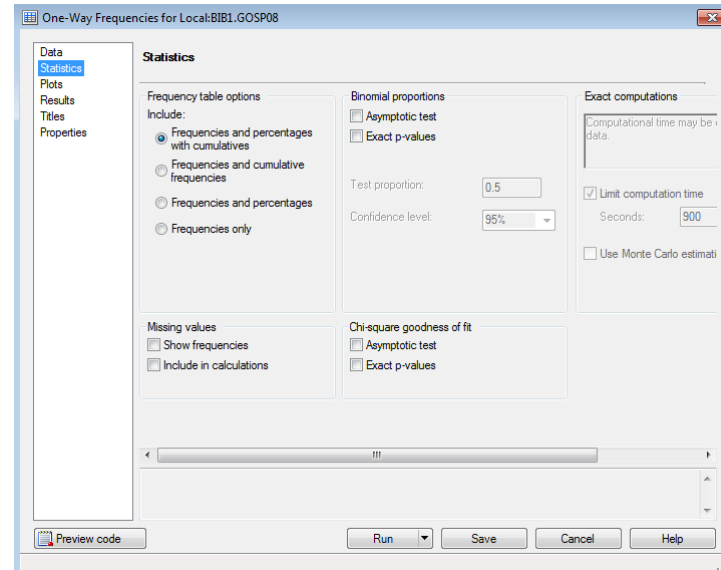
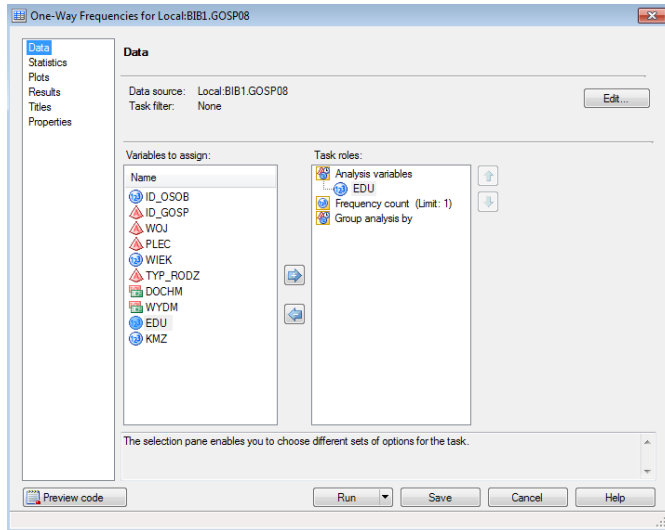
Analiza częstości i liczebności – poziom wykształcenia (EDU)



gosp08



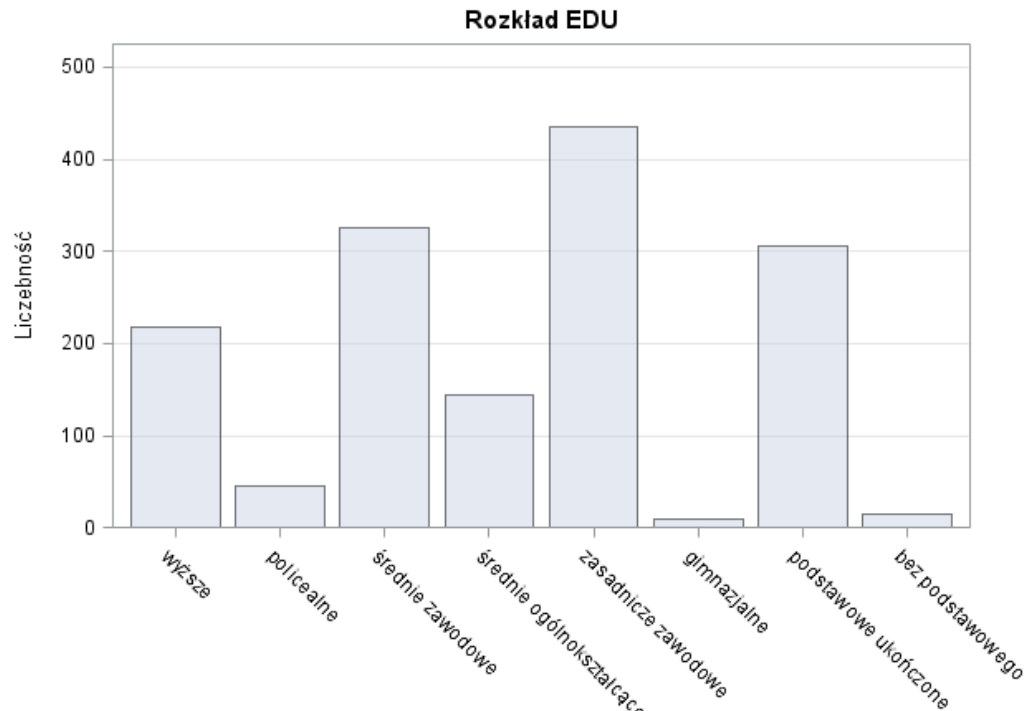
Analiza częstości i liczebności [2]



```
proc freq data=bib1.gosp08;
    tables edu / plots=freqplot;
    format edu edufmt.;
run;
```

Analiza częstości i liczebności [3]

wykształcenie				
EDU	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
wyższe	217	14.47	217	14.47
policealne	46	3.07	263	17.53
średnie zawodowe	325	21.67	588	39.20
średnie ogólnokształcące	145	9.67	733	48.87
zasadnicze zawodowe	436	29.07	1169	77.93
gimnazjalne	10	0.67	1179	78.60
podstawowe ukończone	306	20.40	1485	99.00
bez podstawowego	15	1.00	1500	100.00

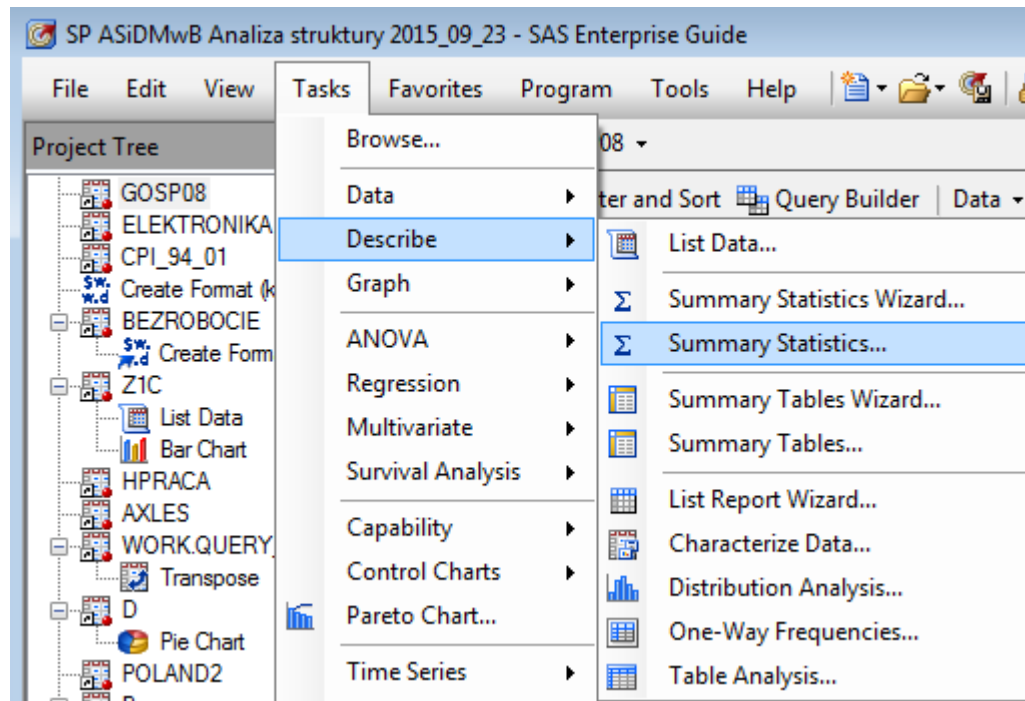
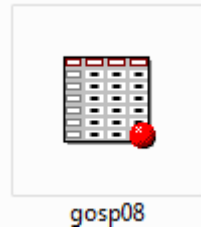


Ćwiczenie

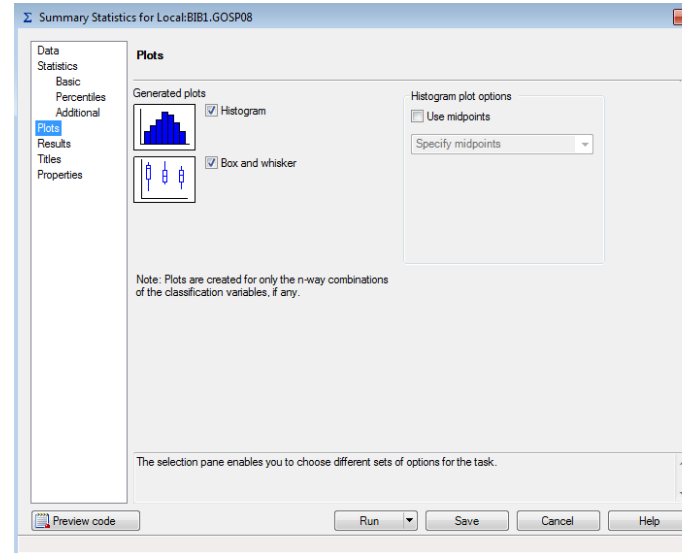
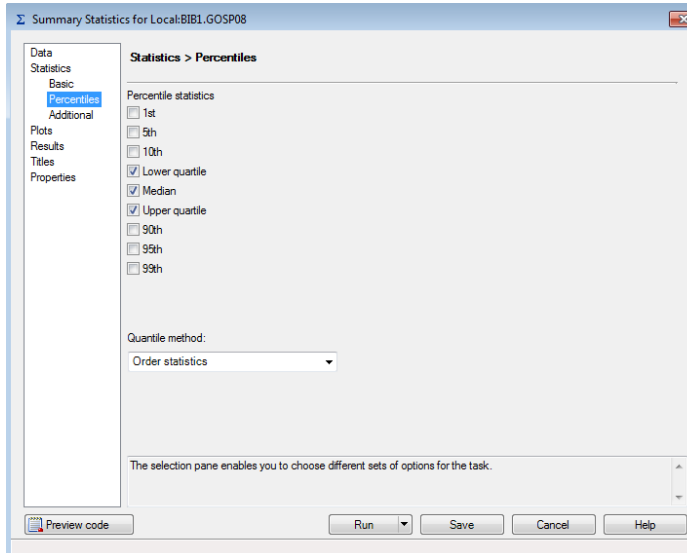
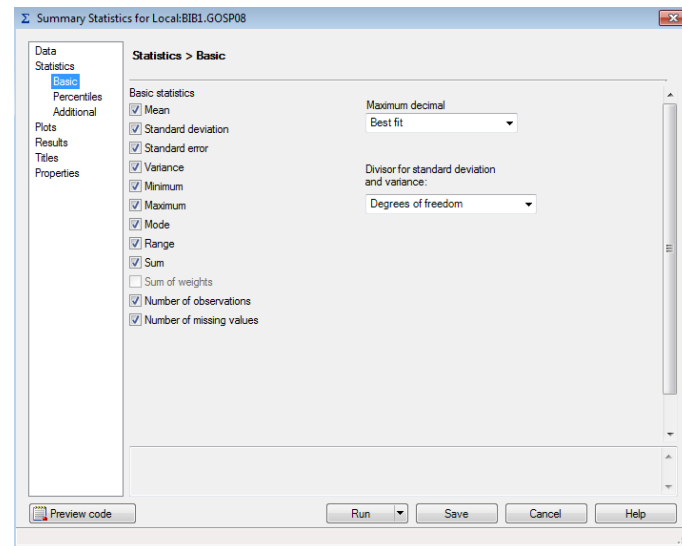
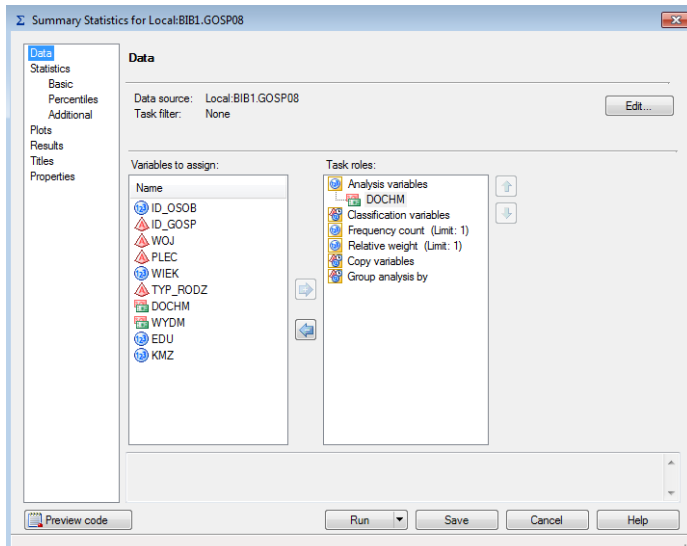
1. Sporządź raport zawierający liczebności i częstości klas miejscowości zamieszkania w próbie „*Gosp08*”. Wyniki przedstaw w tabeli oraz na wykresie słupkowym.

2. Miary położenia rozkładu, zróżnicowania asymetrii oraz koncentracji

Analiza rozkładu dochodów gospodarstw domowych (DOCHM).



Analiza rozkładu dochodów gospodarstw domowych [1]



Analiza rozkładu dochodów gospodarstw domowych [2]

Zmienna analizowana: DOCHM DOCHM										
Średnia	Odch. std.	Błąd std.	Wariancja	Minimum	Maksimum	Moda	Rozstęp	Suma	N	N braków
3438.83	2702.95	69.7898492	7305934.57	94.2000000	34900.40	3000.00	34806.20	5158246.22	1500	0

Dolny kwartył	Mediana	Górny kwartył	Kurtoza	Skośność
1910.00	2837.50	4200.04	36.5373720	4.5242410

```
proc means data=bib1.gosp08 vardef=df
    mean std stderr var min max
    mode range sum n nmiss
    q1 median q3;
    var dochm;

run;
proc univariate data=bib1.gosp08 noprint;
    var dochm;
    histogram;

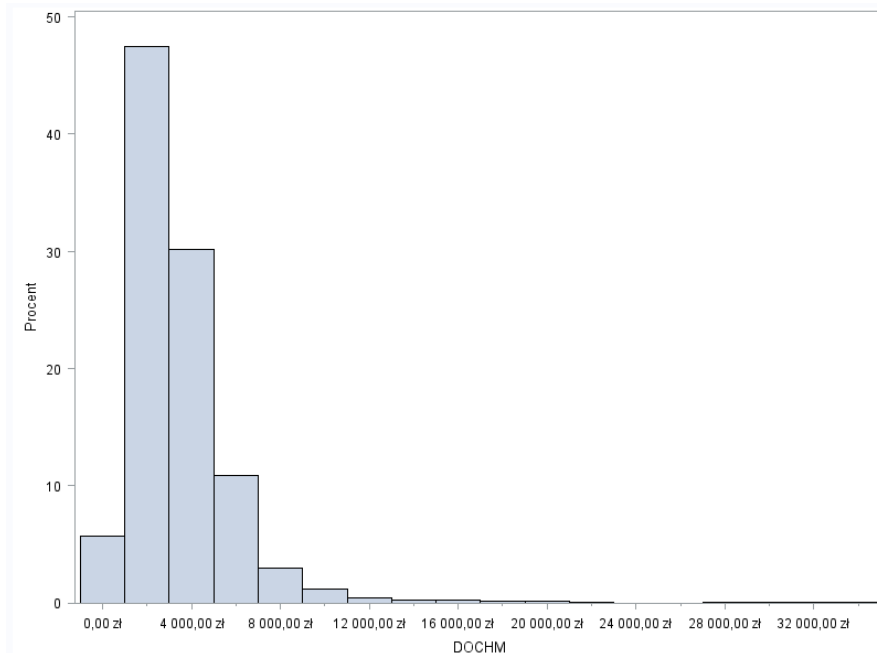
run;

proc sgplot data=bib1.gosp08;
    vbox dochm;

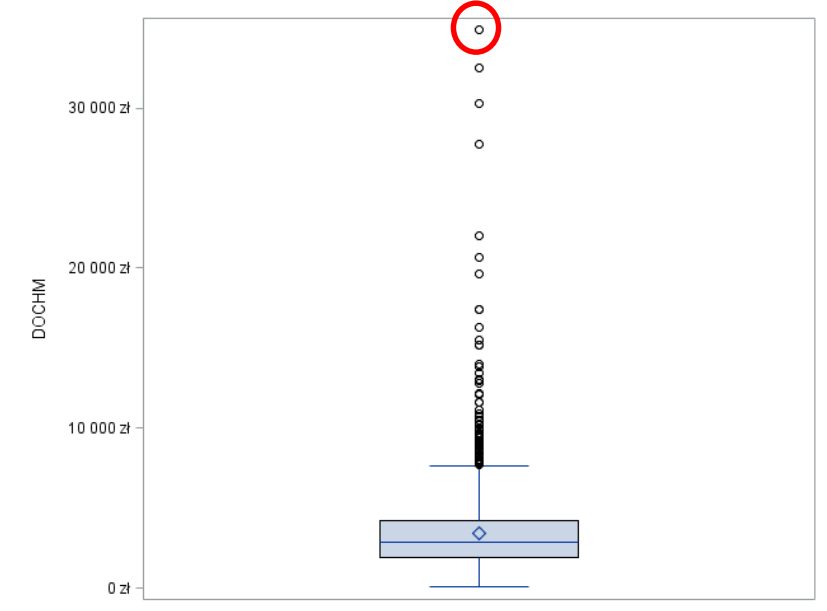
run;
```

Analiza rozkładu dochodów gospodarstw domowych [3]

Histogram



Wykres pudełkowy

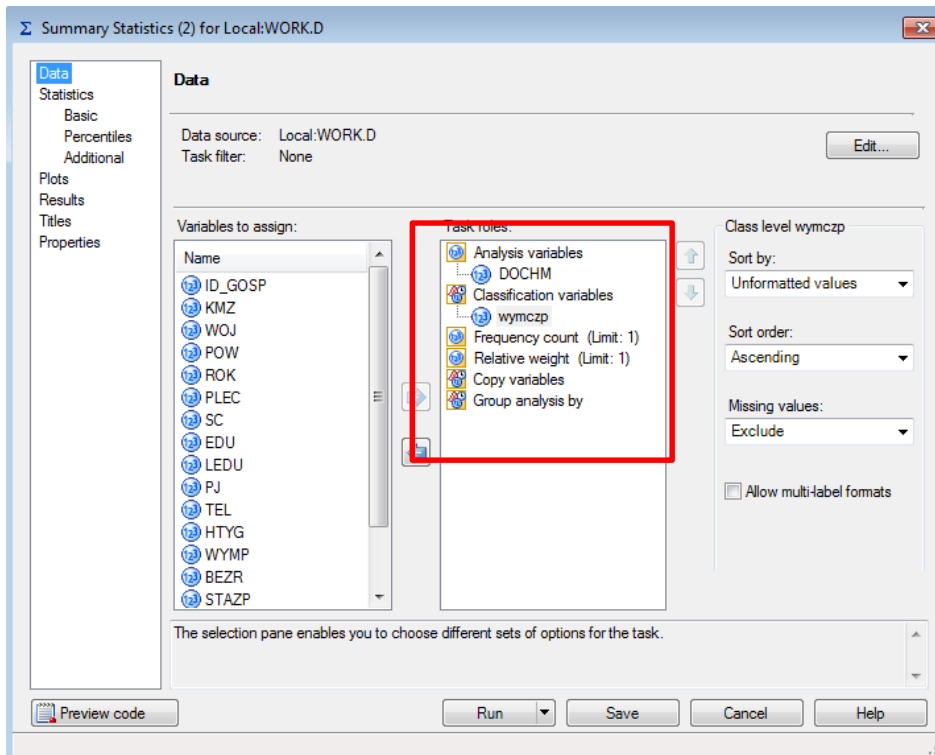


Ćwiczenie

1. Przeanalizuj rozkład wydatków gospodarstw domowych w próbie „*Gosp08*”. Oblicz poszczególne miary opisujące rozkład wydatków oraz podaj ich interpretacje. Rozkład zilustruj za pomocą histogramu oraz wykresu pudełkowego.

Analiza porównawcza *Describe* → *Summary Statistics* [1]

Przeanalizuj rozkład rozkładu dochodów gospodarstw domowych w próbie „Diagnoza07” z podziałem na osoby pracujące na pełny etat oraz pozostałe osoby.



```
data d;

    set bib1.diagnoza07;
    wymczp=(htyg>=40);
    format wymczp wymczpfmt.;

run;

proc means =d vardef=df
    mean std stderr var min max
    mode range sum n nmiss
    q1 median q3;
    var dochm;
    class wymczp;
    where dochm>0;

run;

proc sgplot =d ;
    vbox dochm / group=wymczp;

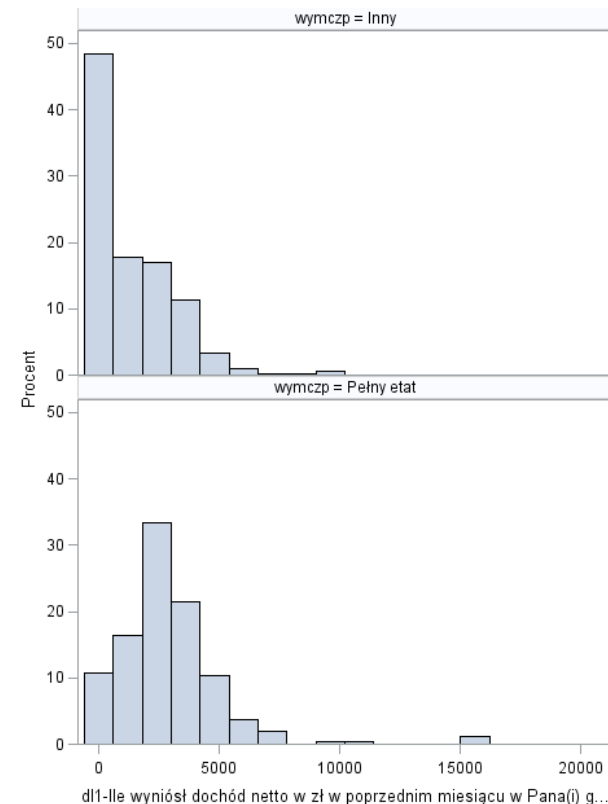
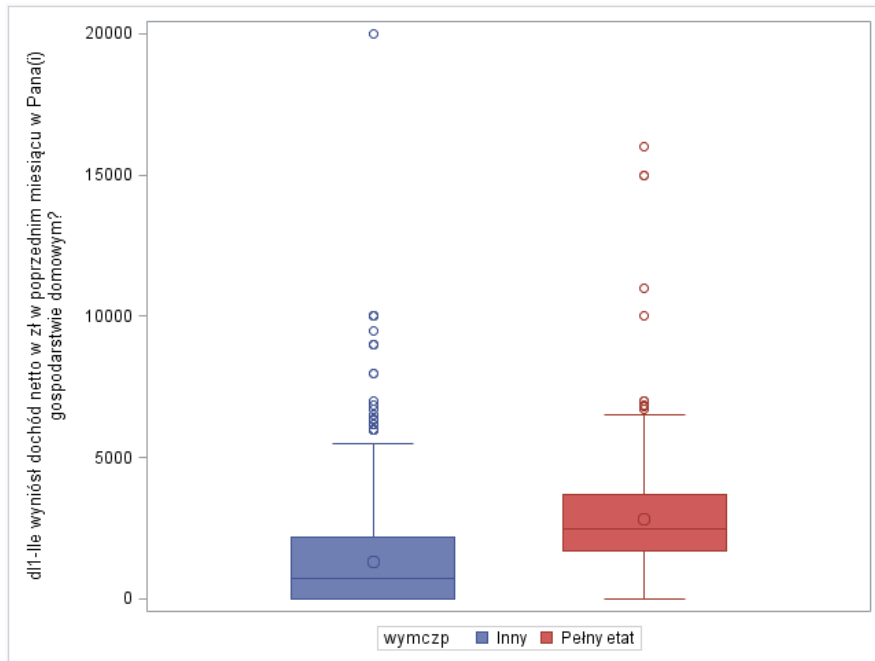
run;

proc sgpanel =d;
    panelby wymczp / columns=1;
    histogram dochm ;

run;
```

Analiza porównawcza [2]

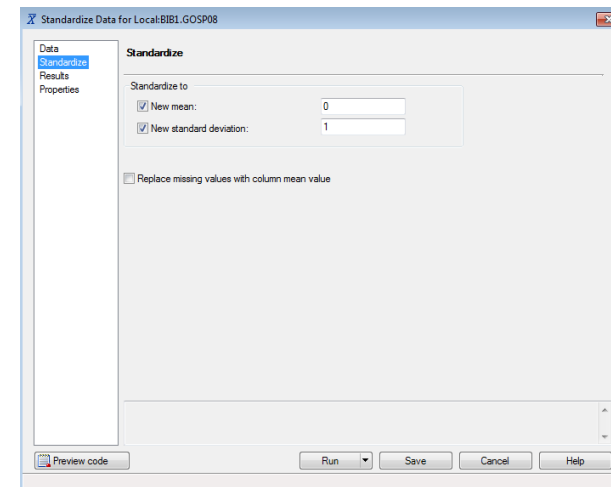
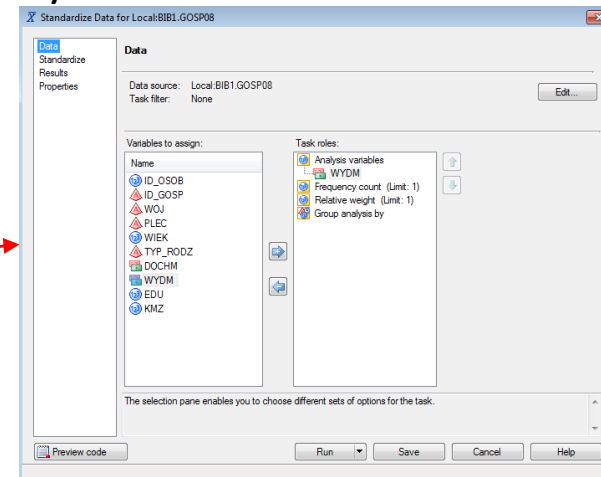
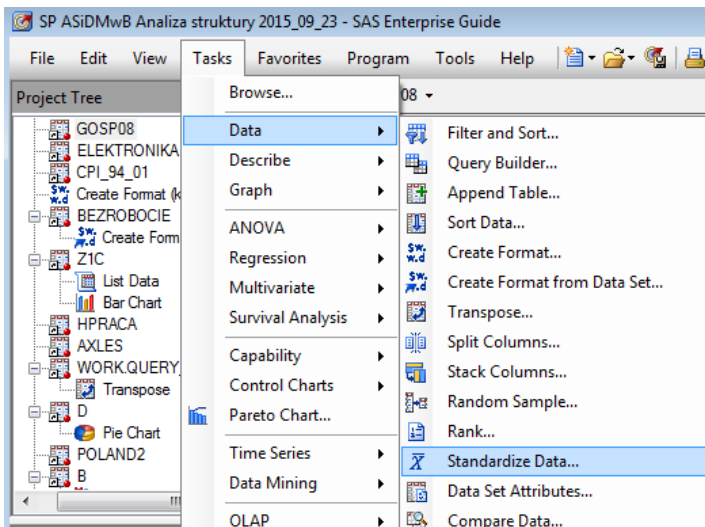
Zmienna analizowana: DOCHM dl1-Ile wyniósł dochód netto w zł w poprzednim miesiącu w Pana(i) gospodarstwie domowym?															
wymczp	N obs.	Średnia	Odch. std.	Błąd std.	Wariancja	Minimum	Maksimum	Moda	Rozstęp	Suma	N	N braków	Dołny kwartył	Mediana	Górny kwartył
Inny	655	2485.55	1675.07	65.4503164	2805852.27	160.0000000	20000.00	3000.00	19840.00	1628032.00	655	0	1400.00	2100.00	3100.00
Pełny etat	241	3131.82	2064.30	132.9730919	4261324.21	230.0000000	16001.00	2000.00	15771.00	754769.00	241	0	2000.00	2600.00	3800.00



Obserwacje odstające

Korzystając z reguły **trzech sigm** sprawdź czy w zbiorze danych „Gosp08” znajdują obserwacje odstające pod względem poziomu wydatków.

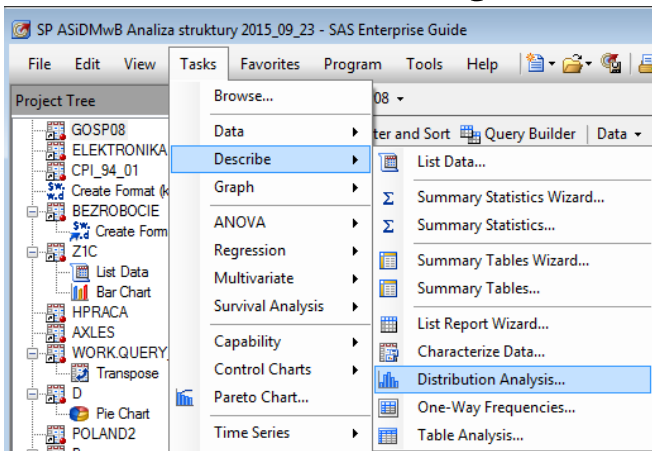
Standaryzacja
$$u = \frac{x - \bar{x}}{s}$$



Lista obserwacji, dla których $u > 3$

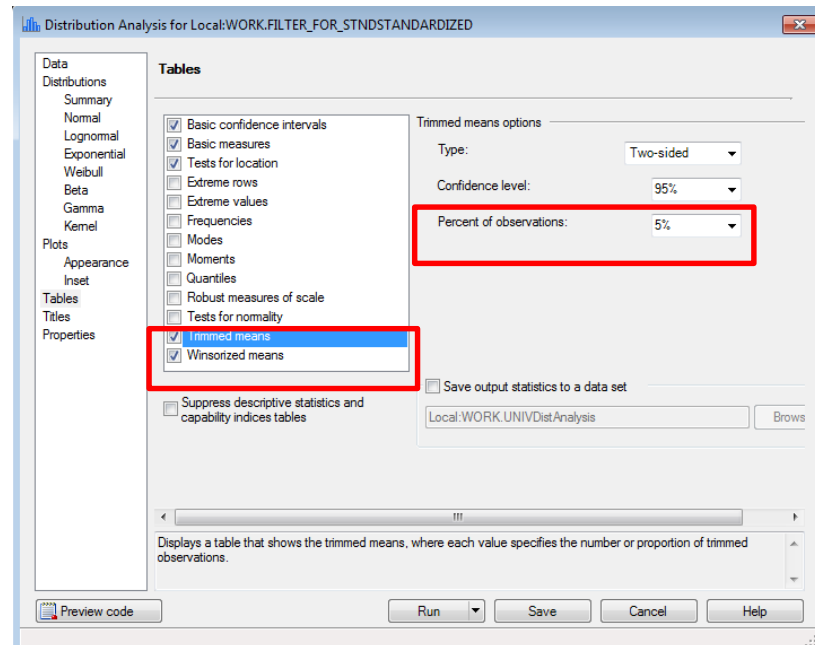
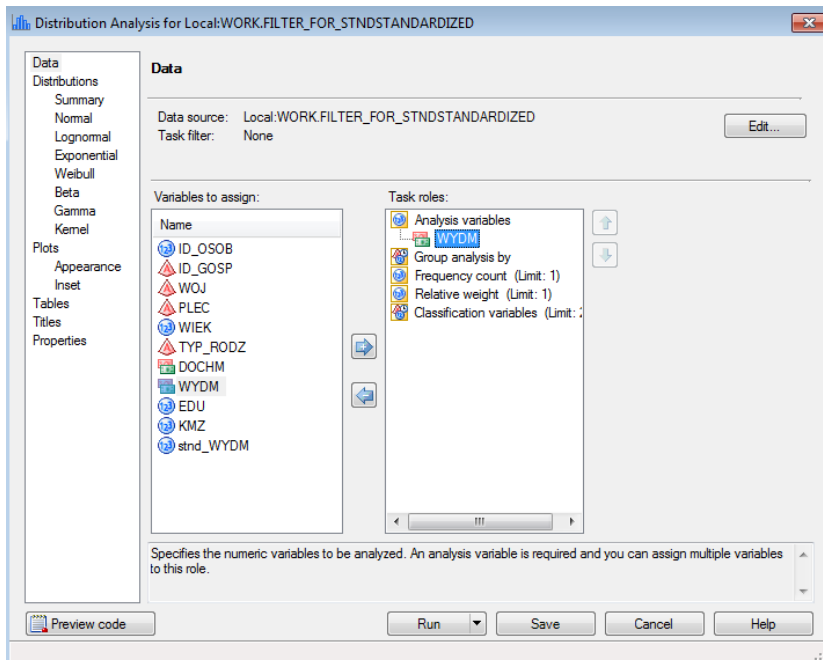
1478	1492	206650521	Pomorskie	Meczczyzna	53	9:Pozostale	9 076,68 zł	8 359,77 zł	1	2	3.0160365956
1479	29332	222610611	Lubuskie	Kobieta	81	9:Pozostale	2 830,99 zł	8 384,77 zł	5	6	3.0294430585
1480	28545	222450811	Lodzkie	Meczczyzna	49	6:Malzenstwo +4+	7 156,50 zł	8 424,72 zł	5	6	3.0508665861
1481	24667	221670221	Mazowieckie	Meczczyzna	58	9:Pozostale	10 870,63 zł	8 522,11 zł	5	6	3.1030928029
1482	24326	221580621	Mazowieckie	Kobieta	63	9:Pozostale	9 853,20 zł	8 969,13 zł	7	6	3.3428110844
1483	24293	221571211	Opolskie	Kobieta	29	9:Pozostale	7 986,45 zł	9 020,80 zł	1	6	3.3705195619
1484	3806	207340311	Malopolskie	Kobieta	24	3:Malzenstwo +1	22 032,30 zł	9 631,72 zł	1	1	3.698130614
1485	2977	207090921	Mazowieckie	Kobieta	30	3:Malzenstwo +1	10 045,76 zł	9 652,39 zł	1	1	3.7092150775
1486	4767	207621121	Lubelskie	Kobieta	24	2:Malzenstwo bez...	4 529,00 zł	10 549,89 zł	1	2	4.1905070949
1487	8405	214020311	Slaskie	Kobieta	49	9:Pozostale	5 351,91 zł	10 789,15 zł	7	3	4.3188123072
1488	2077	206820111	Mazowieckie	Meczczyzna	31	2:Malzenstwo bez...	7 040,00 zł	10 815,24 zł	3	1	4.3328032919
1489	2145	206831222	Mazowieckie	Kobieta	41	4:Malzenstwo +2	12 100,45 zł	11 009,95 zł	1	1	4.4372181873
1490	28074	222370721	Lodzkie	Kobieta	46	2:Malzenstwo bez...	15 180,00 zł	11 474,83 zł	1	6	4.6865140459
1491	10211	214540111	Pomorskie	Meczczyzna	48	3:Malzenstwo +1	14 000,00 zł	11 998,62 zł	1	2	4.9674008936
1492	30774	222920921	Kujawsko-po...	Kobieta	60	9:Pozostale	32 523,40 zł	12 251,50 zł	1	6	5.1030099469
1493	25234	221790411	Mazowieckie	Meczczyzna	31	4:Malzenstwo +2	6 679,65 zł	12 807,14 zł	3	6	5.4009766282
1494	22983	221310921	Podkarpackie	Kobieta	20	9:Pozostale	351,00 zł	12 873,93 zł	6	6	5.4367933345
1495	22978	221310921	Podkarpackie	Meczczyzna	52	9:Pozostale	351,00 zł	12 873,93 zł	3	6	5.4367933345
1496	11892	214961011	Mazowieckie	Kobieta	22	4:Malzenstwo +2	3 769,78 zł	13 268,59 zł	4	3	5.6484331201
1497	23864	221471111	Opolskie	Meczczyzna	37	3:Malzenstwo +1	27 712,80 zł	15 993,39 zł	2	6	7.1096303223
1498	8580	214090811	Slaskie	Meczczyzna	56	9:Pozostale	7 618,49 zł	16 384,25 zł	1	3	7.3192323255
1499	3767	207330221	Malopolskie	Meczczyzna	60	9:Pozostale	11 117,30 zł	18 022,76 zł	1	1	8.1978972652
1500	2657	206990621	Mazowieckie	Kobieta	24	9:Pozostale	17 431,34 zł	24 080,69 zł	4	1	11.446513812

Średnia obcięta i średnia winsorowska



```
proc univariate data = bib1.gosp08
    trimmed=      0.05
    winsorized=   0.05;
var wydm;

run;
```



Średnia obcięta i średnia winsorowska wydatków gospodarstw domowych dla $p=0,05$ (zbiór danych „Gosp08”, zmienna WYDM).

Średnie obcięte								
Procent obcięty w ogonie	Liczba obcięta w ogonie	Średnia obcięta	Średnia bł. std.	Przedział ufności 95%		DF	t dla H0: $\mu_0=0.00$	Pr. > t
5.00	75	2534.952	38.68359	2459.065	2610.838	1349	65.53042	<.0001

Średnie w oknie								
Procent winsoryzowany w ogonie	Liczba winsoryzowana w ogonie	Średnia winsoryzowana	Średni bł. std.	Przedział ufności 95%		DF	t dla H0: $\mu_0=0.00$	Pr. > t
5.00	75	2617.919	38.68502	2542.029	2693.808	1349	67.67267	<.0001

Momenty			
N	1500	Suma wag	1500
Średnia	2735.54864	Suma obserwacji	4103322.96
Odchylenie std.	1864.77225	Wariancja	3477375.56
Skośność	3.31377742	Kurtoza	21.5543457
Niesk. suma kw.	1.64374E10	Skoryg. suma kw.	5212585962
Wsp. zmienności	68.1681264	Błąd std. śr.	48.1482126

Kwantyle (definicja 5)	
Poziom	Kwantyl
100% Maks.	24080.69
99%	10101.14
95%	5828.77
90%	4757.93
75% Q3	3370.11
50% Mediana	2279.82
25% Q1	1586.51
10%	1111.17
5%	918.08
1%	648.77
0% Min.	272.52

Bazowe miary statystyczne			
Położenie		Zmienność	
Średnia	2735.549	Odchylenie std.	1865
Mediana	2279.820	Wariancja	3477376
Moda	741.050	Rozstęp	23808
		Rozstęp międzykwartkowy	1784

Obserwacje ekstremalne			
Najniższe		Najwyższe	
Wartość	Obs.	Wartość	Obs.
272.52	411	13268.6	571
354.45	917	15993.4	1150
369.04	981	16384.3	413
386.34	925	18022.8	200
424.04	1066	24080.7	135



DZIĘKUJĘ ZA UWAGĘ!