

Przetwarzanie, analiza i wizualizacja danych w systemie SAS

PAWEŁ LECHOWICZ

UNIWERSYTET EKONOMICZNY W KATOWICACH

WPROWADZENIE

Niniejsze opracowanie skupia się na analizie danych dotyczących zawałów serca oraz ich predykcji a także na chorobach związanych z sercem, wykorzystując zaawansowane technologie systemu SAS (Statistical Analysis System). Zdrowie serca stanowi fundamentalny obszar badań medycznych, a umiejętne przetwarzanie, analiza i wizualizacja danych mogą przyczynić się do identyfikacji istotnych wzorców oraz predykcji potencjalnych zagrożeń.

Zawały serca są jednym z najpoważniejszych problemów zdrowotnych na świecie, wpływając negatywnie na jakość życia oraz przynosząc znaczne koszty społeczeństwu. W moim projekcie skoncentruję się na analizie danych związanych z zawałami serca oraz chorobach związanymi z sercem korzystając z bazy danych zawierającej informacje.

Cele szczegółowe:

1. Wiek, a obecność ryzyka zawału serca
2. Płeć, a obecność ryzyka zawału serca
3. Śmiertelność chorób serca, a innymi chorobami
4. Palenie, a choroby serca
5. Położenie geograficzne, a ryzyko zawału serca:
 - a) Kraj
 - b) Kontynent

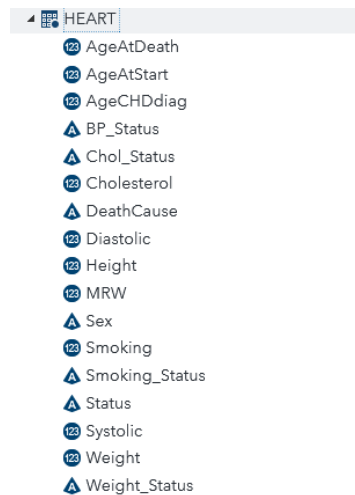
DANE ŹRÓDŁOWE

W mojej pracy korzystał będę z jednej dużej bazy danych, którą pozyskałem ze strony: [Kaggle.com](https://www.kaggle.com). Dane te pobrałem bezpłatnie i są one dostępne dla wszystkich. Same dane nie są prawdziwe a wygenerowane przez sztuczną inteligencję po to aby odzwierciedlać prawdziwe życie i przypadki w nim zachodzące. Na samym początku utworzyłem bibliotekę w programie.

```
1 libname PROJEKT '/home/u63619463/Student/Projekt SAS'
2 run;
```

Kod 1 - utworzenie nowej biblioteki

Jedną z tabeli jaką będą posługiwał się także podczas mojego projektu będzie już wbudowana tabela w SAS Studio- SASHELP.HEART której struktura wygląda następująco



AgeAtDeath	AgeAtStart	AgeCHDdiag	BP_Status	Chol_Status	Cholesterol	DeathCause	Diastolic	Height	MRW	Sex	Smoking	Smoking_Status	Status	Systolic	Weight	Weight_Status
------------	------------	------------	-----------	-------------	-------------	------------	-----------	--------	-----	-----	---------	----------------	--------	----------	--------	---------------

	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height
1	Dead	Other		. Female	29	62.5
2	Dead	Cancer		. Female	41	59.75
3	Alive			. Female	57	62.25
4	Alive			. Female	39	65.75
5	Alive			. Male	42	66
6	Alive			. Female	58	61.75
7	Alive			. Female	36	64.75
8	Dead	Other		. Male	53	65.5
9	Alive			. Male	35	71
10	Dead	Cerebral Vascular Disease		. Male	52	62.5
11	Alive			. Male	39	66.25
12	Alive			57 Male	33	64.25
13	Alive			55 Male	33	70
14	Alive			79 Male	57	67.25
15	Alive			66 Male	44	69
16	Alive			. Female	37	64.5
17	Alive			. Male	40	66.25

Następnie zaimportowałem dane z wcześniej pobranego pliku:

```

14
15 PROC IMPORT DATAFILE=REFFILE
16     DBMS=CSV
17     OUT=PROJEKT.'Dane_główne'.n;
18     GETNAMES=YES;
19 RUN;
20

```

Kod 2 - import danych z Kaggle

Z tak gotowymi danymi przystępujemy do poszczególnych analiz.

Kolumny Wierszy razem: 8763 Kolumn razem: 26 Wiersze 1-100

Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family
67	Male	208	158/88	72	0	
21	Male	389	165/93	98	1	
21	Female	324	174/99	72	1	
84	Male	383	163/100	73	1	
66	Male	318	91/88	93	1	
54	Female	297	172/86	48	1	
90	Male	358	102/73	84	0	
84	Male	220	131/68	107	0	
20	Male	145	144/105	68	1	
43	Female	248	160/70	55	0	
73	Female	373	107/69	97	1	
71	Male	374	158/71	70	1	
77	Male	228	101/72	68	1	
60	Male	259	169/72	85	1	
88	Male	297	112/81	102	1	
73	Male	122	114/88	97	1	
69	Male	379	173/75	40	1	
38	Male	166	120/74	56	1	
50	Female	303	120/100	104	1	
60	Male	145	160/98	71	1	
66	Male	340	180/101	69	1	

Kod 3 – prezentacja źródłowej tabeli

ANALIZA NUMER 1 - WIEK A OBECNOŚĆ RYZYKA ZAWAŁU SERCA

1. Wyjaśnienie celu szczegółowego:

Głównym celem analizy jest sprawdzenie czy występuje korelacja pomiędzy wiekiem a obecnością ryzyka zawału serca.

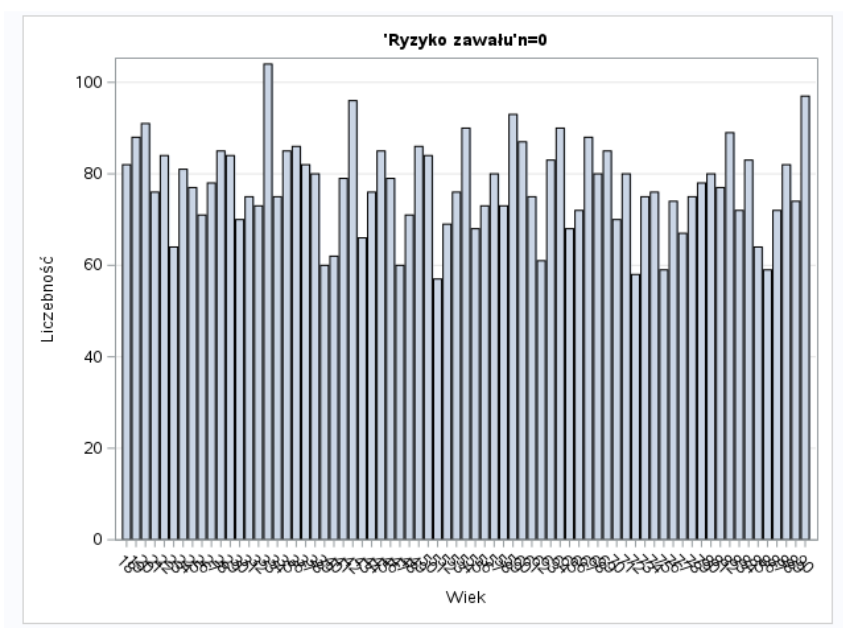
2. Opis procesu uzyskania danych wynikowych:

W celu łatwiejszej prezentacji danych stworzyłem tabele wyłącznie z wiekiem oraz ryzykiem zawału serca (gdzie 1 oznacza obecność ryzyka zawału serca, a 0 jest równoznaczne z jego brakiem)

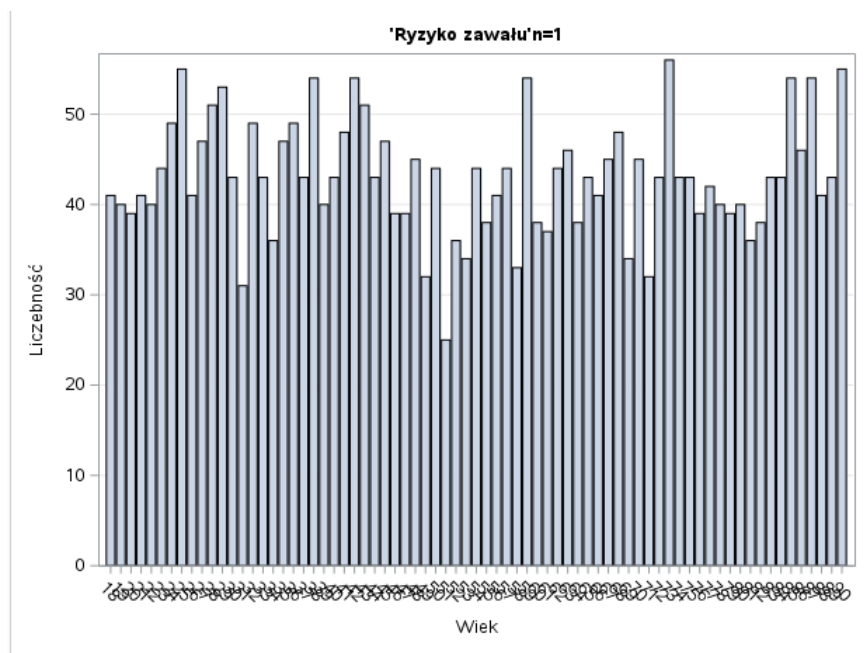
```
1 proc sql;  
2 create table Projekt.Wiek as  
3 select Age as Wiek, 'Heart Attack Risk'n as 'Ryzyko zawału'n  
4 from Projekt.dane_główne  
5 quit;
```

Kod 4 – tworzenie tabeli

Teraz chcielibyśmy przystąpić do badania zależności pomiędzy tymi dwoma cechami. Wydaje się, że wraz z wiekiem ryzyko zawału wzrasta i takiego wyniku też się spodziewamy po naszej analizie. Jednak, aby się upewnić utworzymy wykresy.



Wykres 1



Wykres 2

3. Wnioski uzyskane z przeprowadzonej analizy

Na wykresach została ukazana liczba osób na osi Y oraz wiek tych osób na osi X, jak widzimy ryzyko wystąpienia zawału serca nie jest zależne od wieku i wiele osób doświadcza go w młodym wieku. Aby upewnić się czy wnioski są prawidłowe sporządziłem jeszcze korelacje pomiędzy tymi danymi, czyli (Age – Wiek, Heart Attack Risk - obecnością ryzyka zawału serca)

Następującym kodem tworzymy tabele korelacyjną:

```
proc corr data=PROJEKT.WIEK pearson nosimple noprob plots=none;
var Wiek;
with 'Ryzyko zawału'n;
run;
```

1 Ze zmiennymi:	Heart Attack Risk
1 Zmienne:	Age

Współczynniki korelacji Pearsona, N = 8763	
	Age
Heart Attack Risk	0.00640

Tabela 1- Tabela korelacyjna

Współczynnik korelacji Pearsona służy do sprawdzenia czy dwie zmienne ilościowe są powiązane ze sobą związkiem liniowym. Podobnie jak inne współczynniki korelacji również wynik Pearsona może wahać się od -1 do 1. Wartości skrajne, czyli -1 i 1 oznaczają idealną, totalną korelację między zmienną A i zmienną B. Wynik równy “zero” oznacza brak współwystępowania wartości tych dwóch zmiennych w naturze (brak korelacji). Interpretując nasz wynik, czyli 0.00640 jestem w stanie stwierdzić, iż nie występuje korelacja pomiędzy tymi dwoma danymi bądź też jest bardzo słaba.

Podsumowując korelacja pomiędzy wiekiem a ryzykiem wystąpienia zawału serca nie istnieje bądź też jest bardzo mała, co wywnioskowałem na podstawie wykresu oraz analizy korelacji, w której skorzystałem z współczynnika korelacji Pearsona.

ANALIZA NUMER 2 - PŁEĆ A OBECNOŚĆ RYZYKA ZAWAŁU SERCA

1. Wyjaśnienie celu szczegółowego:

Głównym celem analizy jest sprawdzenie czy występuje korelacja pomiędzy płcią a obecnością ryzyka zawału serca.

2. Opis procesu uzyskania danych wynikowych:

Ponownie w celu ułatwienia pracy tworzymy nową tabelę, gdzie posiadamy tylko dwie kolumny: Płeć oraz Ryzyko zawału serca.

```
1 proc sql;  
2 create table Projekt.Płec as  
3 select Sex as Płeć, 'Heart Attack Risk'n as 'Ryzyko zawału'n  
4 from Projekt.dane_główne  
5 quit;
```

Kod 5- tworzenie tabeli

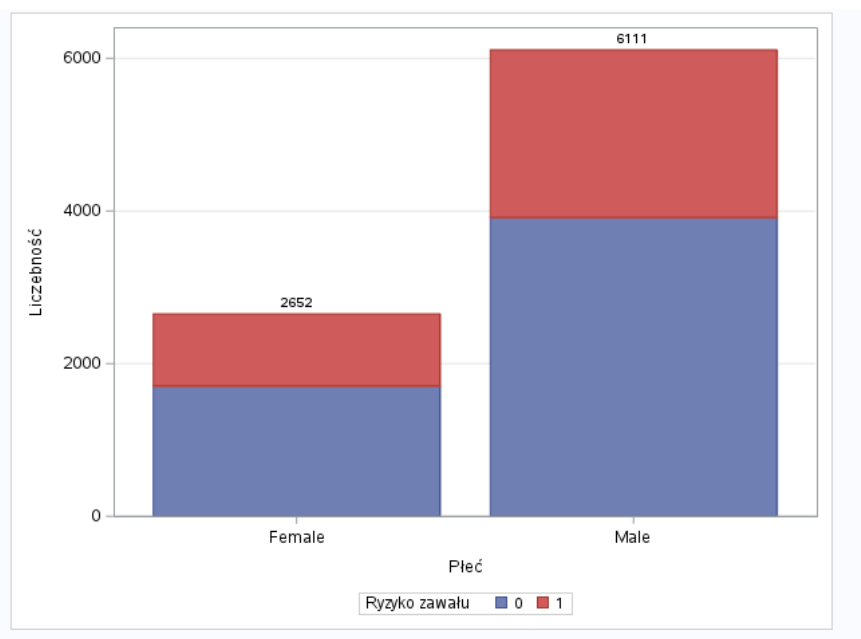
Po stworzeniu tabeli tworzymy wykres, który pomoże nam w ocenie danej korelacji.

```
ods graphics / reset width=6.4in height=4.8in imagemap;  
  
proc sgplot data=PROJEKT.'PłEC'n;  
  vbar 'Płeć'n / group='Ryzyko zawału'n groupdisplay=stack;  
  yaxis grid;  
run;  
  
ods graphics / reset;
```

Kod 6- tworzenie wykresu

Został zastosowany wykres słupkowy, aby każda z płci oddzielnie ukazywała dane.

Sam wykres prezentuje się następująco.



Wykres 3

Na osi Y wykres znajduje się liczba ludzi, natomiast na osi X występują dwie płci: mężczyźni oraz kobiety. Legenda informuje nas o ryzyku zawału, gdzie wartości są to 0 oraz 1: w naszym przypadku 0 oznacza brak ryzyka a 1 oznacza ryzyko zawału serca.

Aby liczbowo przedstawić te dane tworzę nową tabelę z poszczególnymi liczebnościami:

```

1 proc sql;
2 create table Projekt.LiczbaOsob as
3 select Płeć,
4         sum('Ryzyko Zawału'n=0) as BrakRyzyka,
5         sum('Ryzyko Zawału'n=1) as ObecnośćRyzyka
6 from projekt.plec
7 group by Płeć;
8 quit;
```

Kod 6- tworzenie tabeli

Wierszy razem: 2 Kolumn razem: 3

Wiersze 1-2

	Płeć	BrakRyzyka	ObecnoscRyzyka
1	Female	1708	944
2	Male	3916	2195

Tabela 2 – Liczebność osób zależnie od płci

Obliczając procentowo obie płcie wyniki są następujące w bazie danych 56% mężczyzn ma ryzyko zawału serca natomiast 55% posiada ryzyko zawału serca.

3. Wnioski uzyskane z przeprowadzonej analizy:

Brak związku statystycznie istotnego: Na podstawie przeprowadzonej analizy danych stwierdzono, że nie istnieje statystycznie istotny związek między płcią a ryzykiem wystąpienia zawału serca. Procentowy rozkład ryzyka zawału serca wśród osób różnych płci jest zbliżony, co sugeruje, że płci nie można uznawać za kluczowy czynnik predykcyjny ryzyka zawału serca w badanej populacji.

Różnice indywidualne: Pomimo braku związku ogólnego, warto zauważyć, że ryzyko zawału serca to złożony problem, a wpływ różnych czynników może być zróżnicowany dla każdej osoby. Wyniki analizy wskazują, że inne czynniki, takie jak styl życia, genetyka czy obecność innych chorób, mogą mieć większy wpływ na ryzyko zawału serca niż sama płeć.

ANALIZA NUMER 3 - ŚMIERTELNOŚĆ ZWIĄZANA Z CHOROBA SERCA A INNymi CHOROBAМИ

1. Wyjaśnienie celu szczegółowego:

Głównym celem analizy jest zbadanie, w jaki sposób choroby serca porównują się do innych chorób pod względem ryzyka zgonu oraz ustalenie, jaki jest przeciętny wiek przeżycia dla osób dotkniętych chorobami serca w porównaniu z innymi schorzeniami.

2. Opis procesu uzyskania danych wynikowych:

Najpierw, aby zobaczyć ogólne dane naszych chorób stworzyłem tabelę chi2 za pomocą kodu:

```
1 proc freq data=sashelp.heart;  
2     tables DeathCause*Status / chisq;  
3 run;  
4  
5
```

Kod 7- tworzenie tabeli chi-2

Dane jej prezentują się następująco

Procedura FREQ

Liczebność Procent Proc. wier. Proc. kol.	Tabela DeathCause od Status			
	DeathCause(Cause of Death)	Status		
		Alive	Dead	Suma
	Cancer	0 0.00 0.00 .	539 27.07 100.00 27.07	539 27.07
	Cerebral Vascular Disease	0 0.00 0.00 .	378 18.99 100.00 18.99	378 18.99
	Coronary Heart Disease	0 0.00 0.00 .	605 30.39 100.00 30.39	605 30.39
	Other	0 0.00 0.00 .	357 17.93 100.00 17.93	357 17.93
	Unknown	0 0.00 0.00 .	112 5.63 100.00 5.63	112 5.63
	Suma	0 0.00	1991 100.00	1991 100.00
Liczebność braków danych = 3218				

Statystyki dla tabeli przedstawiającej DeathCause od Status
Zerowa suma wiersza lub kolumny. Nie wyliczono statystyk dla tej tabeli.
Liczebność próby = 1991
Liczebność braków danych = 3218

Kod 8 – Tabela przyczyna zgonu od Statusu przeżycia

Tabela przedstawia 5 kategorii i są to: *Rak*, *Choroba naczyniowa mózgu*, *Choroba wieńcowa serca*, *Inne* oraz *Nieznane*. Sama liczebność próby wynosiła 1991. Największa liczba zgonów czyli aż 605 przypada właśnie, chorobie wieńcowej serca co tym samym stanowi 30,39% naszej próby. Na drugim miejscu jest Rak z liczbą 539 zgonów co procentowo można przedstawić jako 27,07% liczebności próby. Natomiast choroba naczyniowa mózgu była na miejscu trzecim pod względem śmiertelności, ponieważ liczb zgonów wyniosła 378 co była równoznaczne z wartością procentową 18,99%.

Po przeanalizowaniu danych ogólnych przeszedłem do analizy testu przeżycia, gdzie skorzystałem z procedury „lifetest”. Ogólnie rzecz biorąc, procedura lifetest jest używana w analizie przeżycia do oceny funkcji przeżycia w zależności od różnych zmiennych, w tym czasu, zdarzeń i stratyfikacji. Analiza przeżycia jest często stosowana w badaniach medycznych do oceny czasu do wystąpienia pewnych zdarzeń, takich jak śmierć czy nawrót choroby. Stratyfikacja pozwala na analizę różnic w przeżyciu między różnymi podgrupami. Jednak, żeby uzyskać wynik potrzebowałem zamienić kolumnę *Status* na kolumnę binarną, gdzie 0 będzie oznaczało martwy, a 1 będzie oznaczało osobę żywą.

```
1 data binarna_heart;  
2   set sashelp.heart;  
3   if Status = "Dead" then Status_Binary = 1;  
4   else if Status = "Alive" then Status_Binary = 0;  
5 run;
```

Kod 9- tworzenie tabeli z kolumna binarną

Gdy już miałem tabelę oraz wiedziałem z jakiej procedury chciałem skorzystać przeszedłem do skonstruowania testu czasowego.

```
1 proc lifetest data=projekt.binarna_heart;  
2   time AgeAtDeath*Status_Binary(0);  
3   strata DeathCause;  
4 run;  
5
```

Kod 10- procedura lifetest

Sama analiza testu przeżycia prezentuje się następująco:

Procedura LIFETEST					
Warstwa 3: Cause of Death = Coronary Heart Disease					
Oceny przeżycia Kaplana-Meiera					
AgeAtDeath	Przeżycie	Niepowodzenie	Błąd standardowy przeżycia	Liczba nieudanych	Liczba pozostałych
0.0000	1.0000	0	0	0	605
38.0000	0.9983	0.00165	0.00165	1	604
43.0000	0.9967	0.00331	0.00233	2	603
45.0000	.	.	.	3	602
45.0000	0.9934	0.00661	0.00329	4	601
46.0000	0.9917	0.00826	0.00368	5	600
47.0000	.	.	.	6	599
47.0000	0.9884	0.0116	0.00435	7	598
48.0000	.	.	.	8	597
48.0000	0.9851	0.0149	0.00492	9	596
49.0000	.	.	.	10	595
49.0000	.	.	.	11	594
49.0000	0.9802	0.0198	0.00567	12	593
50.0000	.	.	.	13	592
50.0000	.	.	.	14	591
50.0000	0.9752	0.0248	0.00632	15	590
51.0000	.	.	.	16	589
51.0000	.	.	.	17	588
51.0000	.	.	.	18	587
51.0000	.	.	.	19	586
51.0000	.	.	.	20	585
51.0000	0.9653	0.0347	0.00744	21	584
52.0000	.	.	.	22	583
52.0000	.	.	.	23	582
52.0000	0.9603	0.0397	0.00794	24	581
53.0000	.	.	.	25	580
53.0000	0.9570	0.0430	0.00825	26	579
54.0000	.	.	.	27	578
54.0000	.	.	.	28	577
54.0000	.	.	.	29	576
54.0000	.	.	.	30	575
54.0000	.	.	.	31	574
54.0000	.	.	.	32	573
54.0000	.	.	.	33	572
54.0000	.	.	.	34	571
54.0000	.	.	.	35	570
54.0000	.	.	.	36	569
54.0000	.	.	.	37	568
54.0000	.	.	.	38	567
54.0000	.	.	.	39	566
54.0000	.	.	.	40	565
54.0000	.	.	.	41	564
54.0000	.	.	.	42	563
54.0000	.	.	.	43	562
54.0000	.	.	.	44	561
54.0000	.	.	.	45	560
54.0000	.	.	.	46	559
54.0000	.	.	.	47	558
54.0000	.	.	.	48	557
54.0000	.	.	.	49	556
54.0000	.	.	.	50	555
54.0000	.	.	.	51	554
54.0000	.	.	.	52	553
54.0000	.	.	.	53	552
54.0000	.	.	.	54	551
54.0000	.	.	.	55	550
54.0000	.	.	.	56	549
54.0000	.	.	.	57	548
54.0000	.	.	.	58	547
54.0000	.	.	.	59	546
54.0000	.	.	.	60	545
54.0000	.	.	.	61	544
54.0000	.	.	.	62	543
54.0000	.	.	.	63	542
54.0000	.	.	.	64	541
54.0000	.	.	.	65	540
54.0000	.	.	.	66	539
54.0000	.	.	.	67	538
54.0000	.	.	.	68	537
54.0000	.	.	.	69	536
54.0000	.	.	.	70	535
54.0000	.	.	.	71	534
54.0000	.	.	.	72	533
54.0000	.	.	.	73	532
54.0000	.	.	.	74	531
54.0000	.	.	.	75	530
54.0000	.	.	.	76	529
54.0000	.	.	.	77	528
54.0000	.	.	.	78	527
54.0000	.	.	.	79	526
54.0000	.	.	.	80	525
54.0000	.	.	.	81	524
54.0000	.	.	.	82	523
54.0000	.	.	.	83	522
54.0000	.	.	.	84	521
54.0000	.	.	.	85	520
54.0000	.	.	.	86	519
54.0000	.	.	.	87	518
54.0000	.	.	.	88	517
54.0000	.	.	.	89	516
54.0000	.	.	.	90	515
54.0000	.	.	.	91	514
54.0000	.	.	.	92	513
54.0000	.	.	.	93	512
54.0000	.	.	.	94	511
54.0000	.	.	.	95	510
54.0000	.	.	.	96	509
54.0000	.	.	.	97	508
54.0000	.	.	.	98	507
54.0000	.	.	.	99	506
54.0000	.	.	.	100	505
54.0000	.	.	.	101	504
54.0000	.	.	.	102	503
54.0000	.	.	.	103	502
54.0000	.	.	.	104	501
54.0000	.	.	.	105	500
54.0000	.	.	.	106	499
54.0000	.	.	.	107	498
54.0000	.	.	.	108	497
54.0000	.	.	.	109	496
54.0000	.	.	.	110	495
54.0000	.	.	.	111	494
54.0000	.	.	.	112	493
54.0000	.	.	.	113	492
54.0000	.	.	.	114	491
54.0000	.	.	.	115	490
54.0000	.	.	.	116	489
54.0000	.	.	.	117	488
54.0000	.	.	.	118	487
54.0000	.	.	.	119	486
54.0000	.	.	.	120	485
54.0000	.	.	.	121	484
54.0000	.	.	.	122	483
54.0000	.	.	.	123	482
54.0000	.	.	.	124	481
54.0000	.	.	.	125	480
54.0000	.	.	.	126	479
54.0000	.	.	.	127	478
54.0000	.	.	.	128	477
54.0000	.	.	.	129	476
54.0000	.	.	.	130	475
54.0000	.	.	.	131	474
54.0000	.	.	.	132	473
54.0000	.	.	.	133	472
54.0000	.	.	.	134	471
54.0000	.	.	.	135	470
54.0000	.	.	.	136	469
54.0000	.	.	.	137	468
54.0000	.	.	.	138	467
54.0000	.	.	.	139	466
54.0000	.	.	.	140	465
54.0000	.	.	.	141	464
54.0000	.	.	.	142	463
54.0000	.	.	.	143	462
54.0000	.	.	.	144	461
54.0000	.	.	.	145	460
54.0000	.	.	.	146	459
54.0000	.	.	.	147	458
54.0000	.	.	.	148	457
54.0000	.	.	.	149	456
54.0000	.	.	.	150	455
54.0000	.	.	.	151	454
54.0000	.	.	.	152	453
54.0000	.	.	.	153	452
54.0000	.	.	.	154	451
54.0000	.	.	.	155	450
54.0000	.	.	.	156	449
54.0000	.	.	.	157	448
54.0000	.	.	.	158	447
54.0000	.	.	.	159	446
54.0000	.	.	.	160	445
54.0000	.	.	.	161	444
54.0000	.	.	.	162	443
54.0000	.	.	.	163	442
54.0000	.	.	.	164	441
54.0000	.	.	.	165	440
54.0000	.	.	.	166	439
54.0000	.	.	.	167	438
54.0000	.	.	.	168	437
54.0000	.	.	.	169	436
54.0000	.	.	.	170	435
54.0000	.	.	.	171	434
54.0000	.	.	.	172	433
54.0000	.	.	.	173	432
54.0000	.	.	.	174	431
54.0000	.	.	.	175	430
54.0000	.	.	.	176	429
54.0000	.	.	.	177	428
54.0000	.	.	.	178	427
54.0000	.	.	.	179	426
54.0000	.	.	.	180	425
54.0000	.	.	.	181	424
54.0000	.	.	.	182	423
54.0000	.	.	.	183	422
54.0000	.	.	.	184	421
54.0000	.	.	.	185	420
54.0000	.	.	.	186	419
54.0000	.	.	.	187	418
54.0000	.	.	.	188	417
54.0000	.	.	.	189	416
54.0000	.	.	.	190	415
54.0000	.	.	.	191	414
54.0000	.	.	.	192	413
54.0000	.	.	.	193	412
54.0000	.	.	.	194	411
54.0000	.	.	.	195	410
54.0000	.	.	.	196	409
54.0000	.	.	.	197	408
54.0000	.	.	.	198	407
54.0000	.	.	.	199	406
54.0000	.	.	.	200	405
54.0000	.	.	.	201	404
54.0000	.	.	.	202	403
54.0000	.	.	.	203	402
54.0000	.	.	.	204	401
54.0000	.	.	.	205	400
54.0000	.	.	.	206	399
54.0000	.	.	.	207	398
54.0000	.	.	.	208	397
54.0000	.	.	.	209	396
54.0000	.	.	.	210	395
54.0000	.	.	.	211	394
54.0000	.	.	.	212	393
54.0000	.	.	.	213	392
54.0000	.	.	.	214	391
54.0000	.	.	.	215	390
54.0000	.	.	.	216	389
54.0000	.	.	.	217	388
54.0000	.	.	.	218	387
54.0000	.	.	.	219	386
54.0000	.	.	.	220	385
54.0000	.	.	.	221	384
54.0000	.	.	.	222	383
54.0000	.	.	.	223	382
54.0000	.	.	.	224	381
54.0000	.	.	.	225	380
54.0000	.	.	.	226	379
54.0000	.	.	.	227	378
54.0000	.	.	.	228	377
54.0000	.	.	.	229	376
54.0000	.	.	.	230	375
54.0000	.	.	.	231	374
54.0000	.	.	.	232	373
54.0000	.	.	.	233	372
54.0000	.	.	.	234	371
54.0000	.	.	.	235	370
54.0000	.	.	.	236	369
54.0000	.	.	.	237	368
54.0000	.	.	.	238	367
54.0000	.	.	.	239	366
54.0000	.	.	.	240	365
54.0000	.	.	.	241	364
54.0000	.	.	.	242	363
54.0000	.	.	.	243	362
54.0000	.	.	.	244	361
54.0000	.	.	.	245	360
54.0000	.	.	.	246	359
54.0000	.	.	.	247	358
54.0000	.	.	.	248	357
54.0000	.	.	.	249	356
54.0000	.	.	.	250	355
54.0000	.	.	.	251	354
54.0000	.	.	.	252	353
54.0000	.	.	.	253	352
54.0000	.	.	.	254	351
54.0000	.	.	.	255	350
54.0000	.	.	.	256	349
54.0000	.	.	.	257	348
54.0000	.	.	.	258	347
54.0000	.	.	.	259	346
54.0000	.	.	.	260	345
54.0000	.	.	.	261	344
54.0000	.	.	.	262	343
54.0000	.	.	.	263	342

Sama procedura podzielona jest na 5 kategorii natomiast my zajmiemy się tylko tą dotyczącą *Choroby wieńcowej serca*. W naszej analizie głównie będą interesowały nas 3 kolumny: z prawdopodobieństwem przeżycia, prawdopodobieństwem do niej przeciwnym czyli niepowodzenie oraz AgeAtDeath czyli wiekiem śmierci. Pierwszą wartością jest 0 gdzie prawdopodobieństwem przeżycia jest równe 100%. Jak widzimy prawdopodobieństwa te maleją wraz z wiekiem aż do 92 lat, gdzie prawdopodobieństwo przeżycia wynosi 0%. Oczywiście jest, że w prawdziwym życiu zdarzają się przypadki gdzie osoby z tą chorobą dożywają takiego wieku i nie można brać takiej danej jako pewnik. Na koniec każdej tabeli umieszczona jest również inna tabela:

Statystyki agregujące zmiennej czasowej AgeAtDeath				
Oceny kwartylowe				
Procent	Ocena punktowa	Przedział ufności 95%		
		Przekształcenie	[Dolna	Górna]
75	77.0000	LOGLOG	76.0000	79.0000
50	70.0000	LOGLOG	69.0000	72.0000
25	64.0000	LOGLOG	63.0000	65.0000

Średnia	Błąd standardowy
70.3289	0.3949

3. Wnioski uzyskane z przeprowadzonej analizy:

Dominująca rola choroby wieńcowej serca: Choroba wieńcowa serca wydaje się być głównym czynnikiem wpływającym na śmiertelność w analizowanej grupie. Wysoka śmiertelność wskazuje na znaczący wpływ tej choroby na zdolność przeżycia pacjentów.

Średni wiek dożycia: Średni wiek 70 lat dla osób z chorobą wieńcową serca jest istotnym wskaźnikiem, sugerującym, że pomimo wysokiej śmiertelności, część pacjentów jest w stanie przeżyć do zaawansowanego wieku. To może być punktem wyjścia do dalszych badań nad czynnikami wpływającymi na długość życia tych pacjentów.

ANALIZA NUMER 4 – PALENIE A WYSTĘPOWANIE CHOROÓB SERCA

1. Wyjaśnienie celu szczegółowego:

Celem tej analizy jest zgłębienie wpływu palenia tytoniu na ryzyko występowania chorób serca. Palenie jest powszechnie uznawane za istotny czynnik ryzyka w rozwoju chorób serca, takich jak choroba wieńcowa serca czy choroba naczyniowa mózgu.

2. Opis procesu uzyskania danych wynikowych:

W ramach analizy skoncentrujemy się na zmiennych *Smoking_Status* (status palenia), oraz *Status* (status przeżycia) a także *CauseOfDeath*. Analiza odsetka oraz test chi-kwadrat zostaną zastosowane, aby ocenić związki między tymi zmiennymi. Najpierw przeprowadzimy test chi-kwadrat dla statusu palenia a statusu przeżycia.

```
1 proc freq data=sashelp.heart;  
2   tables Status * Smoking_Status / chisq;  
3 run;  
4
```

Kod 7- tworzenie testu chi-kwadrat

Taki kod pozwoli nam utworzyć odpowiednią tabelę i wyciągnąć już pierwsze wnioski.

Procedura FREQ

Liczebność Procent Proc. wier. Proc. kol.	Tabela Status od Smoking_Status						
	Status	Smoking_Status(Smoking Status)					Suma
		Heavy (16-25)	Light (1-5)	Moderate (6-15)	Non-smoker	Very Heavy (> 25)	
Alive		603	392	363	1610	234	3202
		11.66	7.58	7.02	31.12	4.52	61.90
		18.83	12.24	11.34	50.28	7.31	
		57.65	67.70	63.02	64.37	49.68	
Dead		443	187	213	891	237	1971
		8.56	3.61	4.12	17.22	4.58	38.10
		22.48	9.49	10.81	45.21	12.02	
		42.35	32.30	36.98	35.63	50.32	
Suma		1046	579	576	2501	471	5173
		20.22	11.19	11.13	48.35	9.10	100.00
Liczebność braków danych = 36							

Statystyki dla tabeli przedstawiającej Status od Smoking_Status

Statystyka	DF	Wartość	Prawd.
Chi-kwadrat	4	52.8985	<.0001
Chi-kw. ilorazu wiarygodn.	4	52.1677	<.0001
Chi-kwadrat Mantela-Haenszela	1	0.0022	0.9629
Współczynnik FI		0.1011	
Współczynnik kontyngencji		0.1006	
V Cramera		0.1011	

Liczebność próby = 5173
Liczebność braków danych = 36

Tabela 3- Tabela chi-2

Informacje jaką możemy wywnioskować jest większa liczba osób martwych niż żywych(237 do 234) w kategorii osób palących ponad 25 papierosów.

```
proc sql;
  create table HeartSummary as
  select Smoking_Status, DeathCause, count(*) as Count
  from sashelp.heart
  group by Smoking_Status, DeathCause;
quit;
```

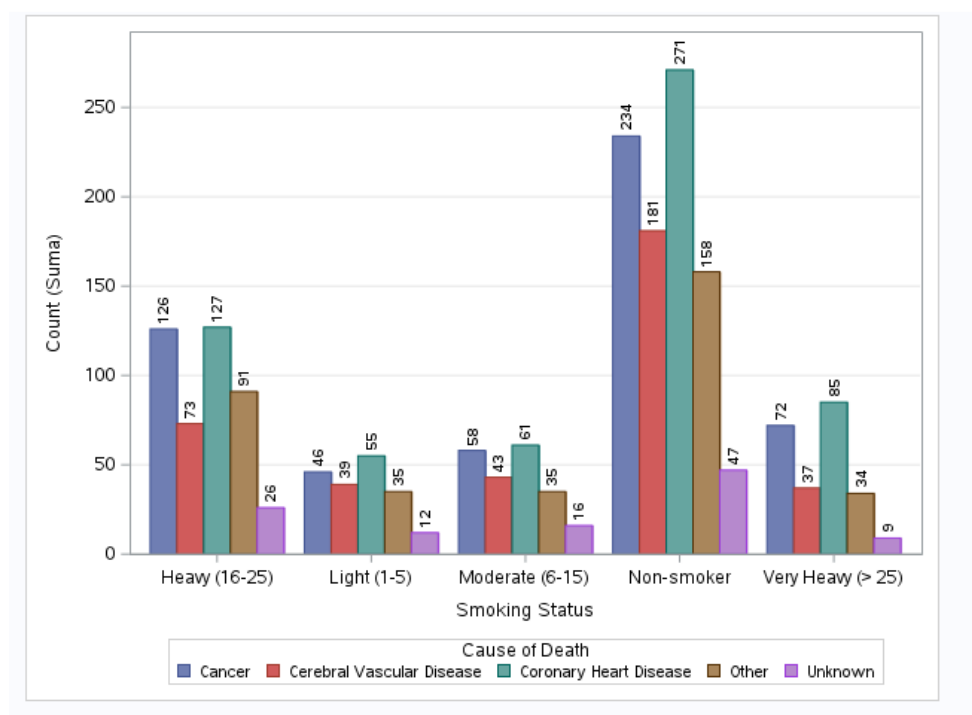
W tym kodzie najpierw używamy proc sql do stworzenia nowej tabeli HeartSummary, w której używamy funkcji count(*) do policzenia liczby przypadków dla każdej kombinacji Smoking_Status i DeathCause.. Następnie używamy procedury sgplot z nowo utworzoną tabelą HeartSummary do stworzenia wykresu słupkowego.

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=PROJEKT.HEARTSUMMARY;
  vbar Smoking_Status / response=Count group=DeathCause groupdisplay=cluster
  datalabel;
  yaxis grid;
run;

ods graphics / reset;
```


Dzięki temu uzyskujemy taki wykres słupkowy:



Wykres 4- Wykres słupkowy wpływu palenia na poszczególne choroby

Wykres ten nie odpowiada nam jednak na nasze pytanie, czyli jak palenie wpływa na rozwój chorób, ponieważ niepalących osób jest znacznie więcej niż palących, stąd też sam wynik nie ma sensu gdyż największą śmiertelność uzyskały osoby niepalące. W takim przypadku musimy zrobić odsetek osób zmarłych na chorobę wieńcową serca dla każdej z 5 kategorii a następnie je porównać. Skorzystamy, więc z tego wykresu oraz wcześniej utworzonej tabeli korelacyjnej aby poobliczać odsetki dla każdej z grup:

Heavy: $127/1046 = 12\%$

Light: $55/579 = 9\%$

Moderate: $61/576 = 11\%$

Non-Smoker: $271/2501 = 11\%$

Very Heavy: $85/471 = 18\%$

3. Wnioski uzyskane z przeprowadzonej analizy:

Analizując uzyskane odsetki możemy wyciągnąć parę zaskakujących informacji. Pierwszą z nich jest fakt, że osoby palące małą liczbę papierosów dziennie miały mniejszą liczbę zgonów na chorobę wieńcową serca od wszystkich innych grup. Następnie wszystkie trzy grupy, czyli Moderate Heavy oraz Non-smoker zbliżyły się do siebie wynikiem co również może być informacją dosyć kontrowersyjną. Ostatnim wysnutym wnioskiem jest fakt, że kategoria Very Heavy uzyskała odsetek zgonów na chorobę serca równy 18% tym samym czyniąc go największym. Odpowiadając na tezę postawioną w celu szczegółowym tej analizy: Tak papierosy wpływają na występowanie chorób serca. W naszym przypadku tylko grupa Very Heavy odbiegała znacznie od reszty i widać to zarówno w przypadku tabeli korelacyjnej jak i w przypadku wyliczonych odsetków zmarłych na chorobę wieńcową.

ANALIZA 5 - POŁOŻENIE GEOGRAFICZNE A RYZYKO ZAWAŁU SERCA

1. Wyjaśnienie celu szczegółowego:

Celem tego szczegółowego badania jest zrozumienie, czy istnieją różnice w średnim ryzyku zawału serca w poszczególnych regionach geograficznych, zarówno według państw, jak i kontynentów. W tym kontekście analiza zostanie przeprowadzona, aby zbadać, czy lokalizacja geograficzna może być potencjalnym czynnikiem wpływającym na ryzyko zawału serca.

Rozważając ryzyko zawału serca na poziomie państw, analiza ta umożliwi identyfikację ewentualnych obszarów, gdzie poziom ryzyka może być wyższy lub niższy niż średnia. Natomiast analiza na poziomie kontynentów pozwoli na szersze spojrzenie na zróżnicowanie między obszarami geograficznymi o różnych charakterystykach kulturowych, społeczno-ekonomicznych i środowiskowych.

2. Opis procesu uzyskania danych wynikowych:

Najpierw zajmiemy się krajami i w tym celu utworzymy procedurę *MEANS*:

Procedura MEANS

Zmienna analizowana: Heart Attack Risk		
Country	N obs.	Średnia
Argentina	471	0.3694268
Australia	449	0.3741648
Brazil	462	0.3528139
Canada	440	0.3590909
China	436	0.3555046
Colombia	429	0.3776224
France	448	0.3520179
Germany	477	0.3605870
India	412	0.3131068
Italy	431	0.3155452
Japan	433	0.3325635
New Zealand	435	0.3471264
Nigeria	448	0.3973214
South Africa	425	0.3388235
South Korea	409	0.3985330
Spain	430	0.3488372
Thailand	428	0.3761682
United Kingdom	457	0.3501094
United States	420	0.3952381
Vietnam	425	0.3482353

Ukazuje ona liczbę obserwacji dla danego kraju oraz średnie ryzyko zawału serca dla każdego z nich. Sam kod do procedury prezentuje się następująco:

```

2 proc means data=Projekt.Dane_Główne mean;
3   class Country;
4   var "Heart Attack Risk";
5   output out=AvgRiskByCountry mean=średnie_ryzyko_zawału_serca;
6 run;

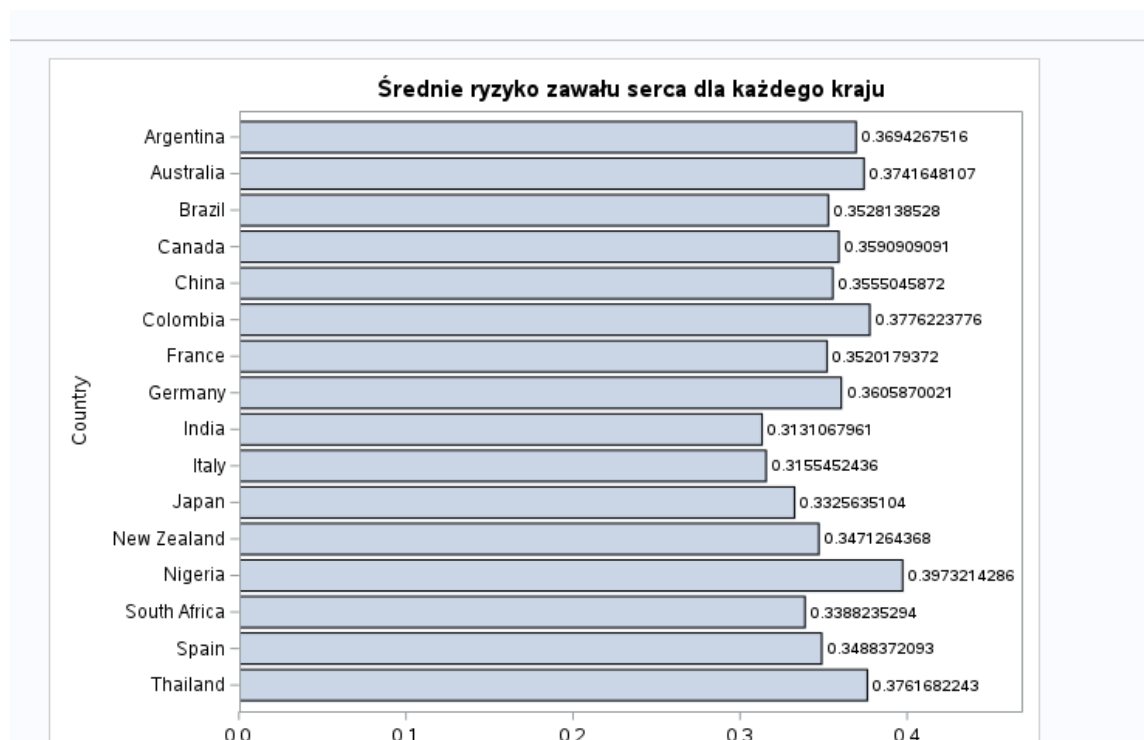
```

Aby dane były bardziej przejrzyste utworzymy jeszcze wykres do tej tabeli.

```

/* Wykres dla każdego kraju */
proc sgplot data=AvgRiskByCountry;
  hbar Country / response=średnie_ryzyko_zawału_serca datalabel;
  title 'Średnie ryzyko zawału serca dla każdego kraju';
run;

```



Z wykresu tego odczytujemy że średnie ryzyko zawału serca jest najmniejsze dla Indii a największe dla Nigerii. Można także stwierdzić że średnie nie odchylają się bardzo od siebie. Teraz ponowimy czynności dla kontynentu i najpierw stworzymy tabelę.

```

proc means data=Projekt.prawidlowe_dane mean;
  class Continent;
  var "Heart Attack Risk";
  output out=AvgRiskByContinent mean=średnie_ryzyko_zawału_serca;
run;

```

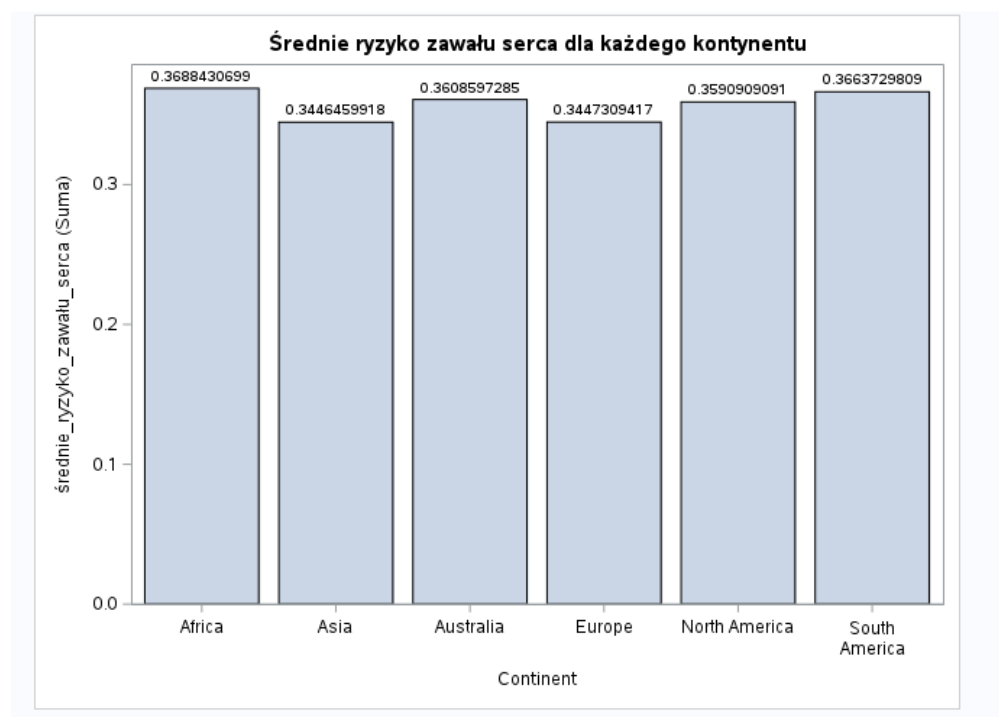
Kod – tworzenie tabeli ze średnimi dla kontynentu

Procedura MEANS

Zmienna analizowana: Heart Attack Risk		
Continent	N obs.	Średnia
Africa	873	0.3688431
Asia	1709	0.3446460
Australia	884	0.3608597
Europe	1784	0.3447309
North America	440	0.3590909
South America	1362	0.3663730

Od razu dla tabeli tworzymy również wykres bardzo podobnym kodem jak tym wcześniejszym.

```
proc sgplot data=AvgRiskByContinent;
  vbar Continent / response=średnie_ryzyko_zawału_serca datalabel;
  title 'Średnie ryzyko zawału serca dla każdego kontynentu';
run;
```



Ponownie średnie są zbliżone do siebie. Największą średnią uzyskała Afryka, która wyniosła w zaokrągleniu 37%. Natomiast wynik najmniejszy uzyskał region Azji który miał średnią od Europy mniejsza o tylko 0,001. Na koniec, aby przedstawić zarówno kraje jak i kontynenty

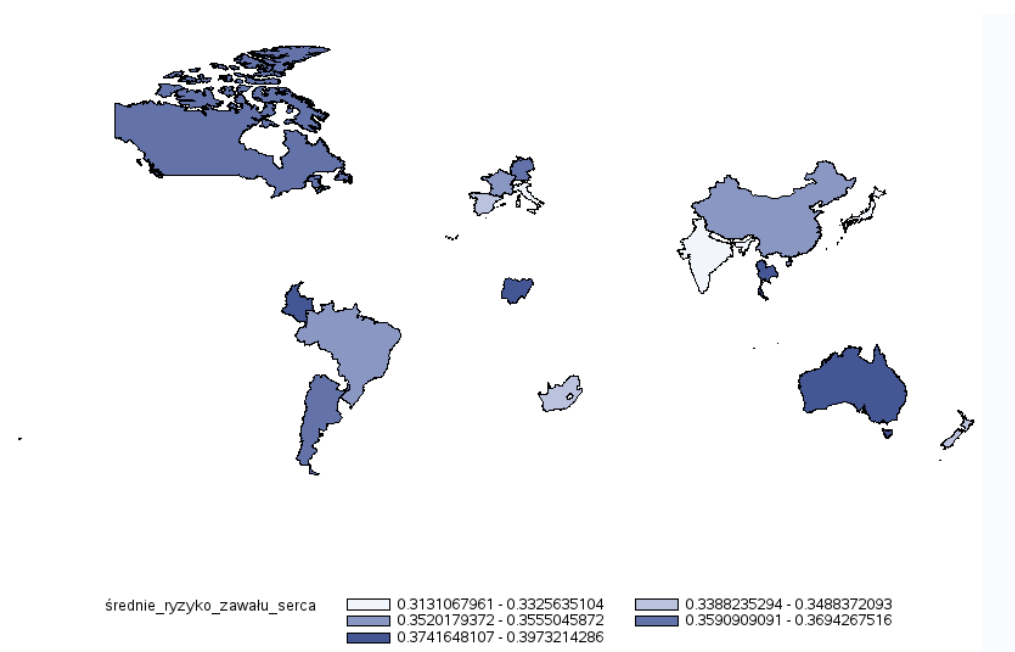
utworzymy mapę choropleth. Zaczniemy od przygotowania nowej tabeli gdzie przygotujemy zmienną pod ID mapy.

```
1 data MapDataNew;  
2   set AvgRiskByCountry(rename=(Country=IDNAME));  
3   format IDNAME $55.;  
4 run;
```

Następnie tworzymy samą mapę następującym kodem:

```
proc gmap data=MapDataNew map=mapsgfk.world;  
  id IDNAME;  
  choro średnie_ryzyko_zawału_serca / levels=5; / Definiuje poziomy kolorów */  
  title 'Mapa choropleth dla średniego ryzyka zawału serca w krajach';  
run;
```

Dzięki temu uzyskujemy mapę:



Sama mapa nie jest uzupełniona o wszystkie państwa ponieważ tak wygląda nasza baza danych.

3. Wnioski uzyskane z przeprowadzonej analizy:

W zakończeniu tego szczegółowego badania stwierdzono, że średnie ryzyko zawału serca w poszczególnych regionach geograficznych wykazuje podobne wartości, różniące się między sobą średnio o około 3 punkty procentowe w przypadku kontynentów oraz 8 punktów procentowych w przypadku państw. Pomimo że różnice te nie są znaczące statystycznie, warto zwrócić uwagę na istniejące subtelności, które mogą mieć wpływ na zdrowie publiczne.

Analiza na poziomie państw wskazała, że Nigeria wykazuje najwyższe średnie ryzyko zawału serca, podczas gdy Indie prezentują najniższe wartości. Różnice te, choć niewielkie, mogą wynikać z lokalnych czynników, takich jak zwyczaje żywieniowe, poziom aktywności fizycznej czy dostępność opieki zdrowotnej.

Na poziomie kontynentów, Afryka charakteryzuje się największym średnim ryzykiem zawału serca, natomiast Azja prezentuje najniższe wartości. Różnice te, choć subtelne, mogą wynikać z różnorodności czynników kulturowych, społeczno-ekonomicznych i środowiskowych.

Mimo niewielkich rozbieżności między regionami, warto kontynuować badania w celu głębszego zrozumienia specyfiki czynników wpływających na ryzyko zawału serca w poszczególnych lokalizacjach.

PODSUMOWANIE

Podsumowując wszystkie pięć analiz. Nie jesteśmy w stanie wskazać jednego konkretnego czynnika który by miał znaczący czy też dominujący wpływ na ryzyko zawału serca. Sam człowiek to bardzo skomplikowany organizm i wiele czynników składa się na występowanie czy to choroby czy też właśnie zawału serca. Pomimo to udało się wyznaczyć pewne zależności czy też ich braki w mojej pracy. Pierwszą cechą badaną był wiek co do ryzyka zawału serca gdzie stwierdziłem, że nie odgrywa on jakiejś istotnej funkcji w tym przypadku. Następnie badałem czy płeć ma wpływ na to czy występuje ryzyko zawału serca i ponownie zarówno dla kobiet jak i mężczyzn dane prezentowały się podobnie do siebie, więc eliminujemy zależność pomiędzy tymi dwoma cechami. Trzecią przeprowadzoną analizą była śmiertelność osób na chorobę wieńcową serca a innymi chorobami. Wysnutym wnioskiem z tej analizy był fakt że najwięcej osób w naszej grupie badanych zmarło właśnie na chorobę serca oraz średni wiek zgonu wyniósł 70 lat. Czwartą analizą była analiza wpływu palenia na choroby serca. Jednym z wniosków było to że ponad połowa osób palących od 19-25 papierosów już nie żyje. Natomiast drugi wniosek był dosyć zaskakujący ponieważ reszta grup palaczy którzy zmarli na chorobę wieńcową nie odbiegała od grupy osób niepalących. Jedną z grup palaczy: osoby palące od 1-5 papierosów posiadała nawet mniejszy odsetek liczby zgonów na chorobę serca niż grupa osób niepalących. Ostatnią przeprowadzoną analizą była analiza ryzyka zawału serca w zależności od położenia geograficznego. Stwierdziłem w niej, że położenie geograficzne może mieć lekki wpływ na zawał serca ale mówimy tutaj bardziej w przypadku zależnym od kraju niż samego kontynentu.