Projekt 4

Kacper Prorok, Popkiewicz Szymon

2024-04-25

Opis problemu

Badanym przez nas problemem jest ilość samochodów osobowych o napędzie „zielonym" w powiatach. Dane te w GUSie mają kategorię : Pojazdy według rodzajów stosowanego paliwa – pozostałe. Na początku do przewidywania ilości bierzemy pod uwagę takie zmienne objaśniające jak: ludność na 1km2, ludność powiatu, przystanki autobusowe, odsetek osób w wieku powyżej 65 lat, ceny mieszkań, liczba parków spacerowo-wypoczynkowych oraz dochody na 1 mieszkańca. Należy również wziąć pod uwagę, że liczba samochodów elektrycznych będzie zależeć od aktualnego trendu globalnego oraz od wielu innych czynników takich jak liczba ładowarek (brak danych w GUS), bliskość dużych miast oraz polityki powiatu, co może dosyć utrudniać działanie naszego modelu.

Zmienna samochody zielone

Naszą zmienną **objaśnianą** jest liczba samochodów o napędzie „elektrycznym" (mogą tam również wchodzić inne kategorie jak np.: napęd wodorowy). Zmienna jest typu numerycznego i oznacza ile takich samochodów (w sztukach) zostało zarejestrowanych w powiecie w 2022 roku.

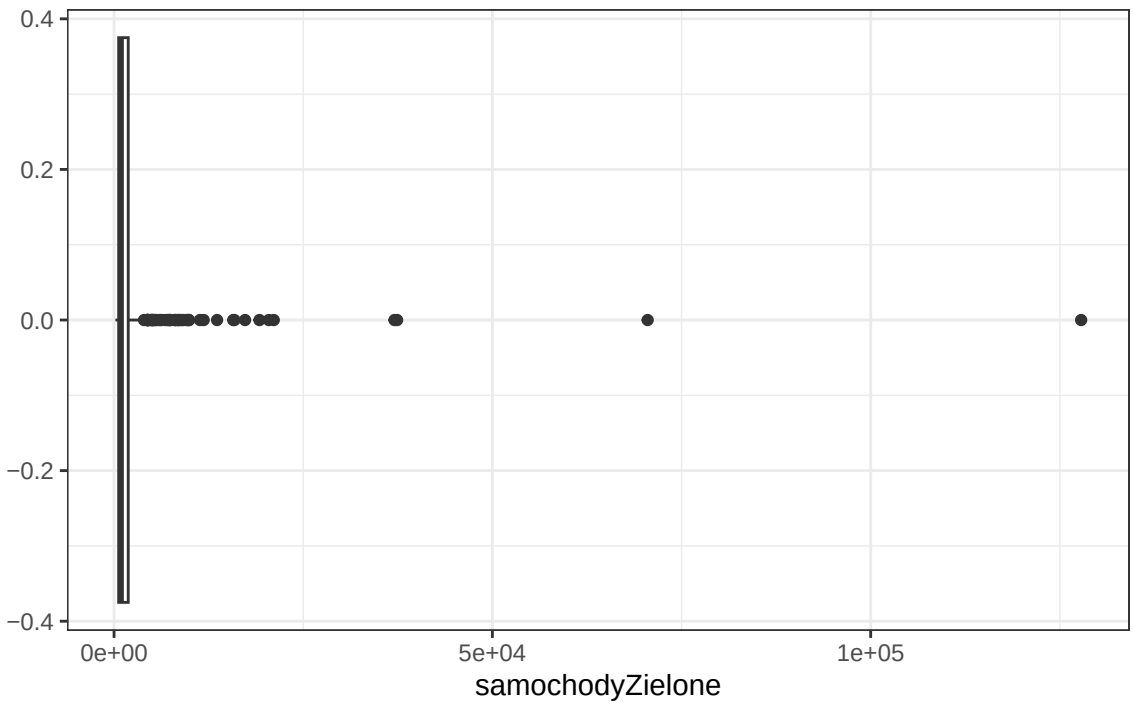


Table 1: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	7601.606	1668.675
odchylenie	18552.860	2469.456

skośność	4.934550	3.919893
kurtoza	29.82515	22.00635

Table 2: Razem

średnia	2699.13158
odchylenie	8314.71672
skośność	10.94684
kurtoza	149.18203

Table 3: Wsp.zmienności

razem	grodzkie	ziemskie
3.080516	2.44065	1.47989

W tym momencie do danych okrojonych o outlier'y, zapisujemy rekordy o ilości samochodów mniejszej niż 50 000.

Zmienna ludność na 1 km2

Pierwszą zmienną objaśniającą jest ludność na 1km2. Mówi nam ona ile osób przypada na 1km2 w powiecie w 2022 roku. Jest to zmienna typu numerycznego.

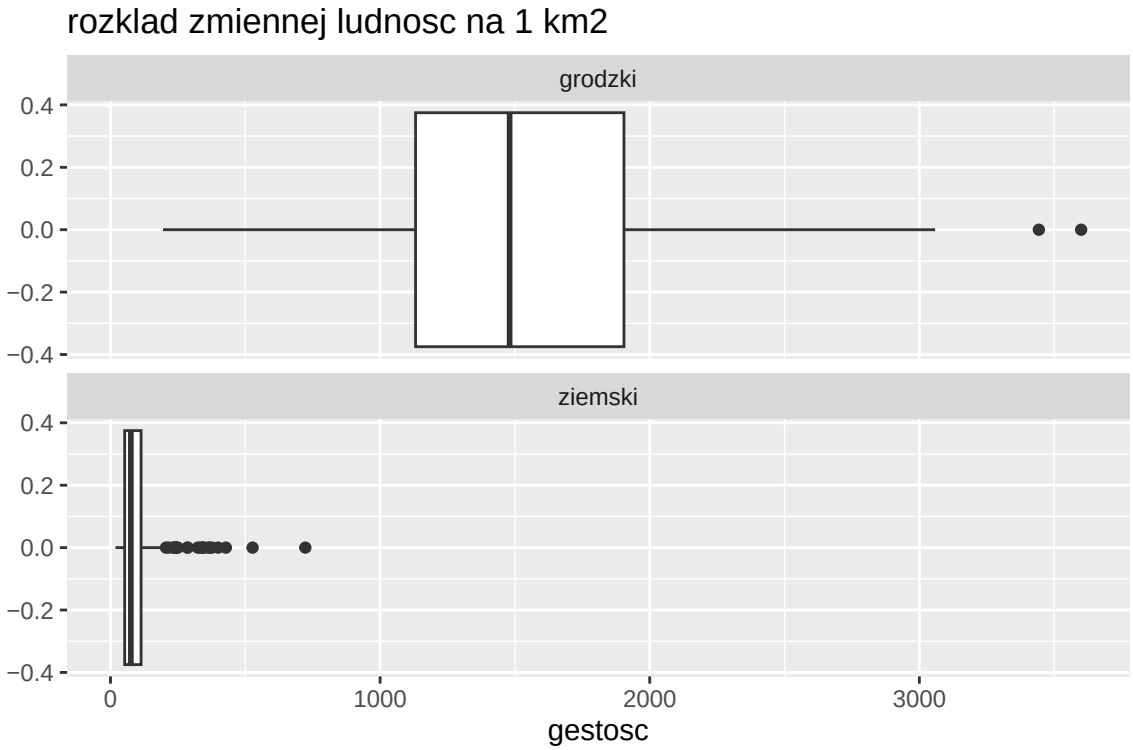


Table 4: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	1572.03939	99.69363
odchylenie	668.85858	81.33817
skośność	0.7172131	3.1728089
kurtoza	3.897312	17.805402

Table 5: Razem

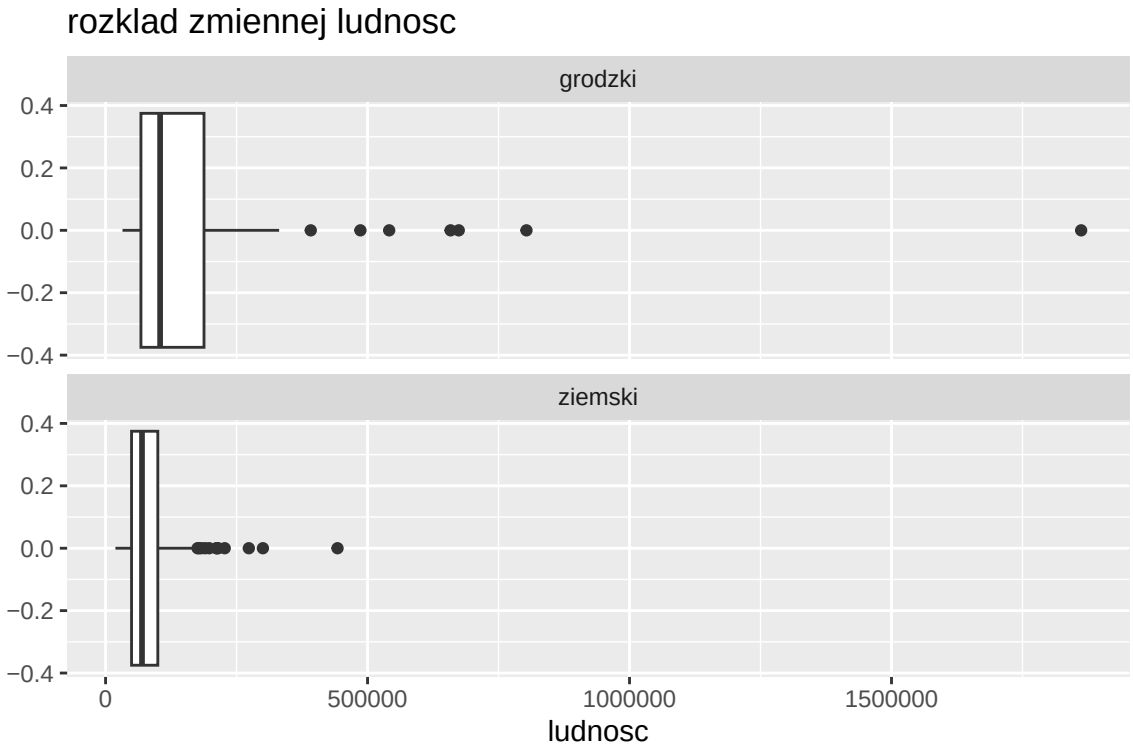
średnia	355.416842
odchylenie	627.797215
skośność	2.482827
kurtoza	8.890055

Table 6: Wsp.zmienności

razem	grodzkie	ziemskie
1.766369	0.4254719	0.8158813

Zmienna ludność

Zmienna ludność mówi nam ile ludzi mieszkało w powiecie w 2022 roku. Zmienna typu numerycznego.



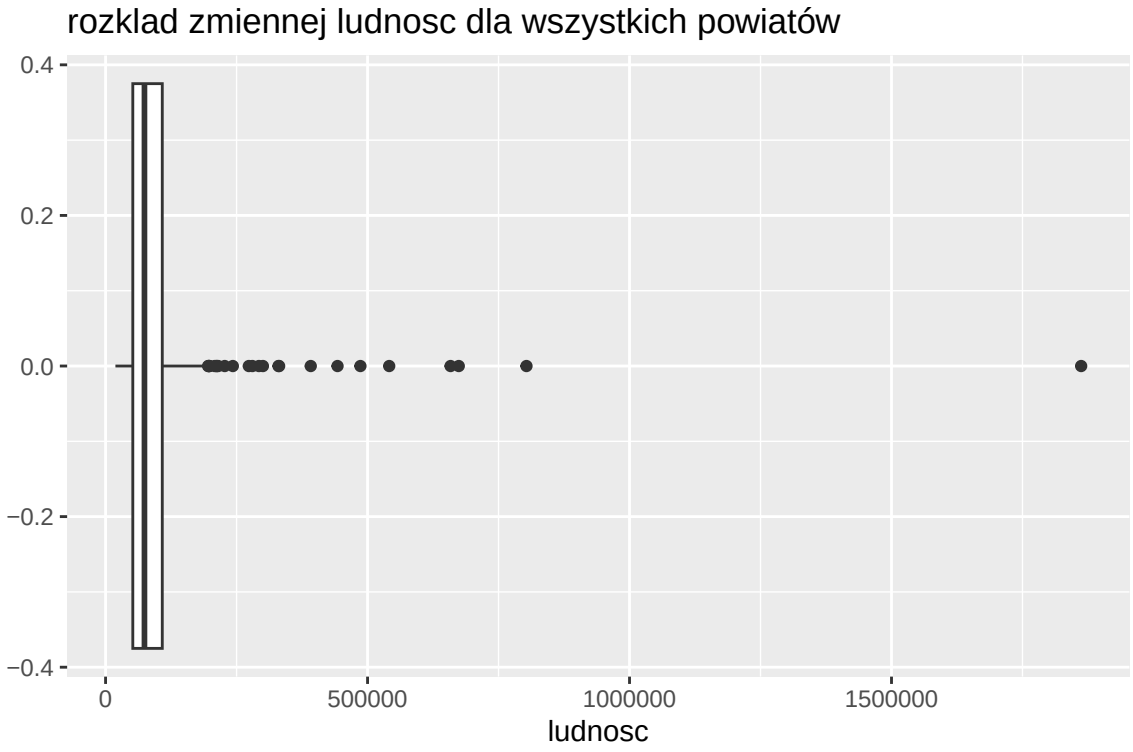


Table 7: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	187405.74	80883.91
odchylenie	262731.9	47306.2

skośność	4.422882	2.549822
kurtoza	26.70903	15.25961

Table 8: Razem

średnia	9.938507e+04
odchylenie	1.237721e+05
skośność	8.983691e+00
kurtoza	1.149807e+02

Table 9: Wsp.zmienności

razem	grodzkie	ziemskie
1.245379	1.401942	0.5848654

Dane okrojone filtrujemy dla rekordów, z ludnością mniejszą niz 1 000 000.

Zmienna przystanki autobusowe i tramwajowe

Mówi nam ile przystanków komunikacji miejskiej występuje w powiecie w 2022 roku. Zmienna typu numerycznego.

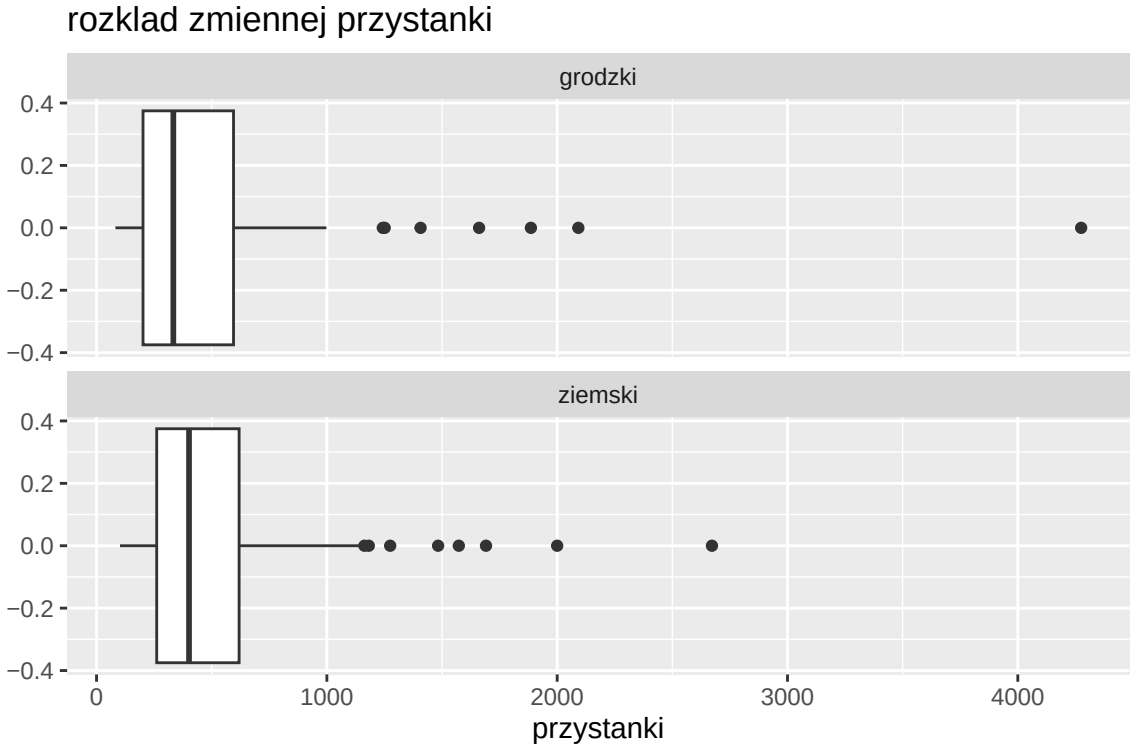


Table 10: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	524.697	472.328
odchylenie	628.3422	305.3518
skośność	3.813892	2.419769
kurtoza	21.10187	13.63831

Table 11: Razem

średnia	481.423684
odchylenie	380.932465
skośność	4.067471
kurtoza	32.369216

Table 12: Wsp.zmienności

razem	grodzkie	ziemskie
0.7912624	1.197534	0.6464826

Filtrujemy o liczę przystanków mniejszą niż 2500.

Zmienna odsetek osób w wieku >=65 lat

Przedstawia odsetek osób w wieku równym lub powyżej 65 lat w ogólnej populacji powiatu w 2022 roku. Zmienna ta będzie w przedziale od 0 do 1 i jest typu numerycznego

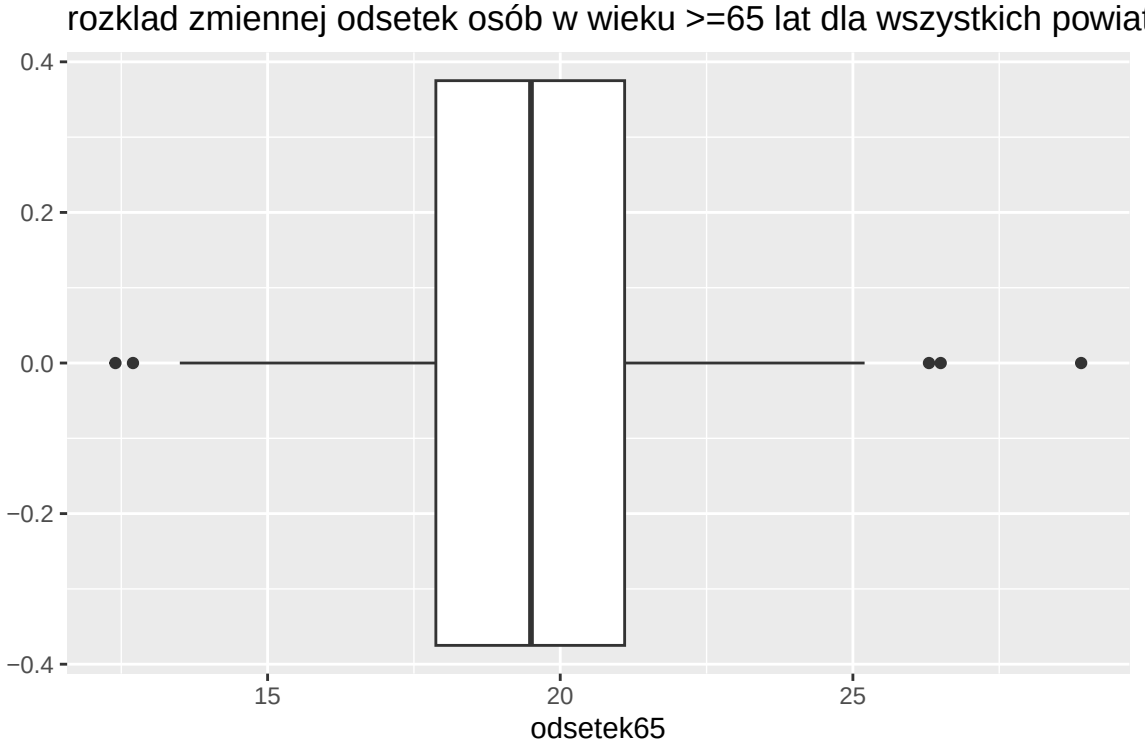


Table 13: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	22.04697	18.92516
odchylenie	2.068981	2.174431

skośność	0.38533389	0.05042761
kurtoza	3.649474	3.294303

Table 14: Razem

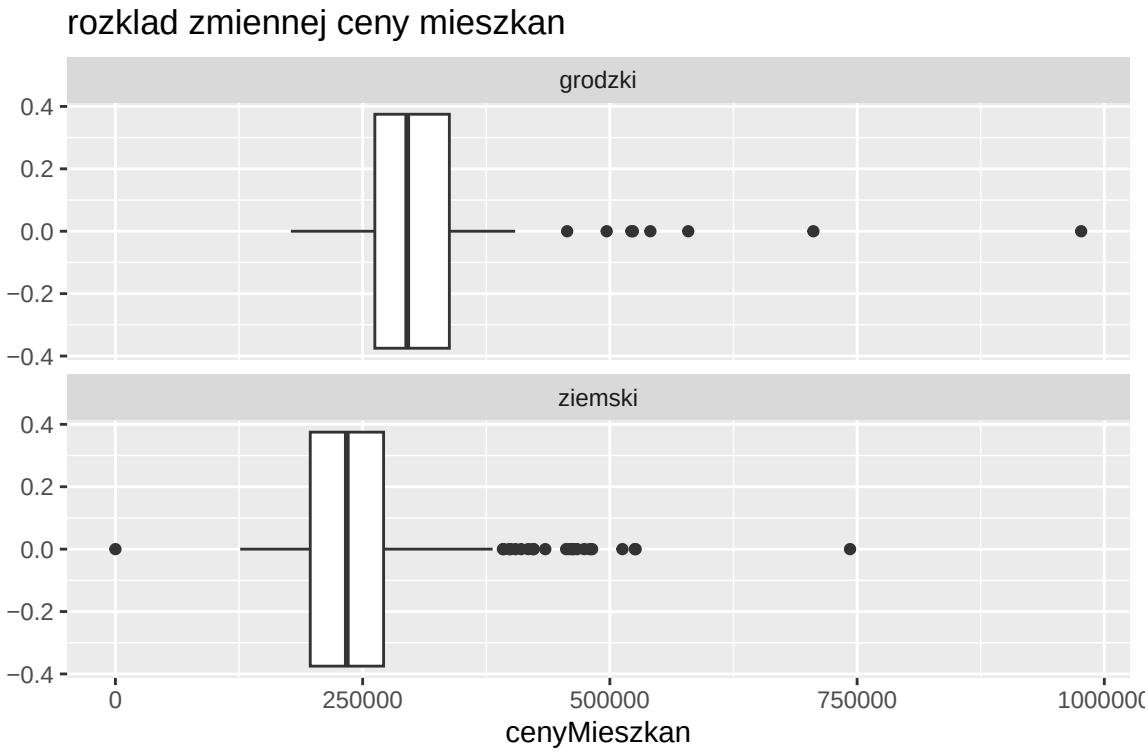
średnia	19.4673684
odchylenie	2.4579070
skośność	0.2154753
kurtoza	3.3232580

Table 15: Wsp.zmienności

razem	grodzkie	ziemskie
0.1262578	0.0938442	0.1148963

Bez większych anomalii (na pewno w porównaniu do poprzednich zmiennych).

Zmienna ceny mieszkań



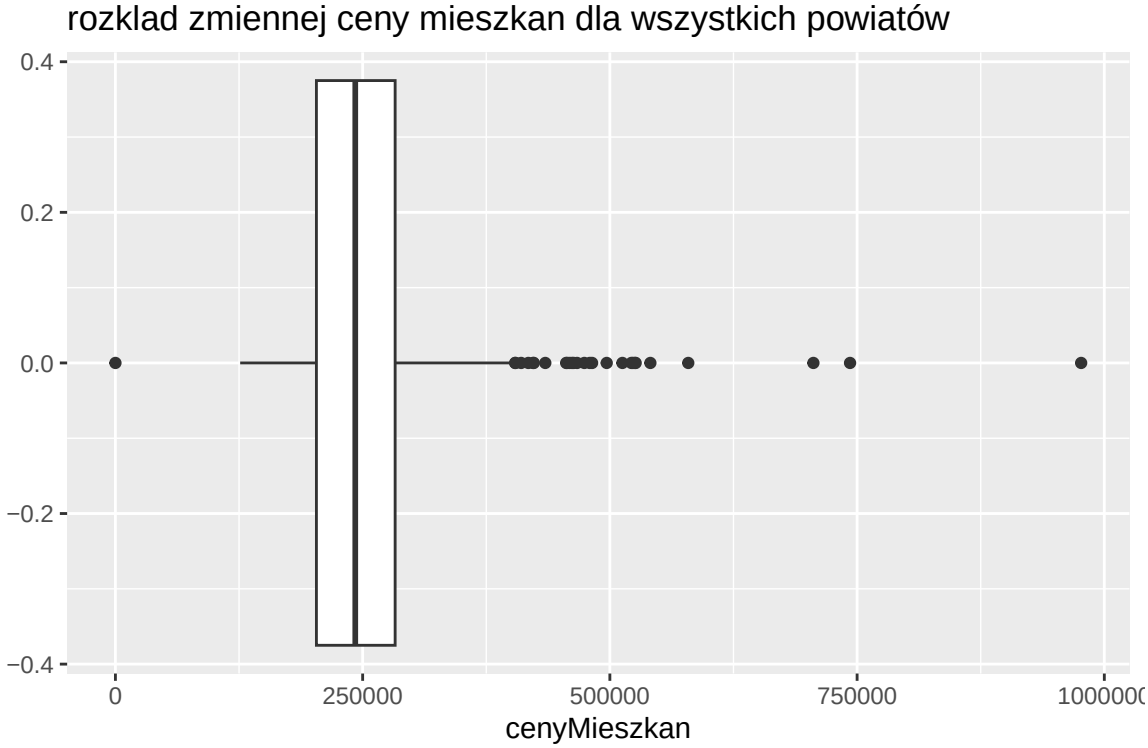


Table 16: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	323787.5	248017.8
odchylenie	128457.88	78469.35
skośność	2.693348	1.917105
kurtoza	12.612599	9.220497

Table 17: Razem

średnia	2.610470e+05
odchylenie	9.336348e+04
skośność	2.578783e+00
kurtoza	1.501647e+01

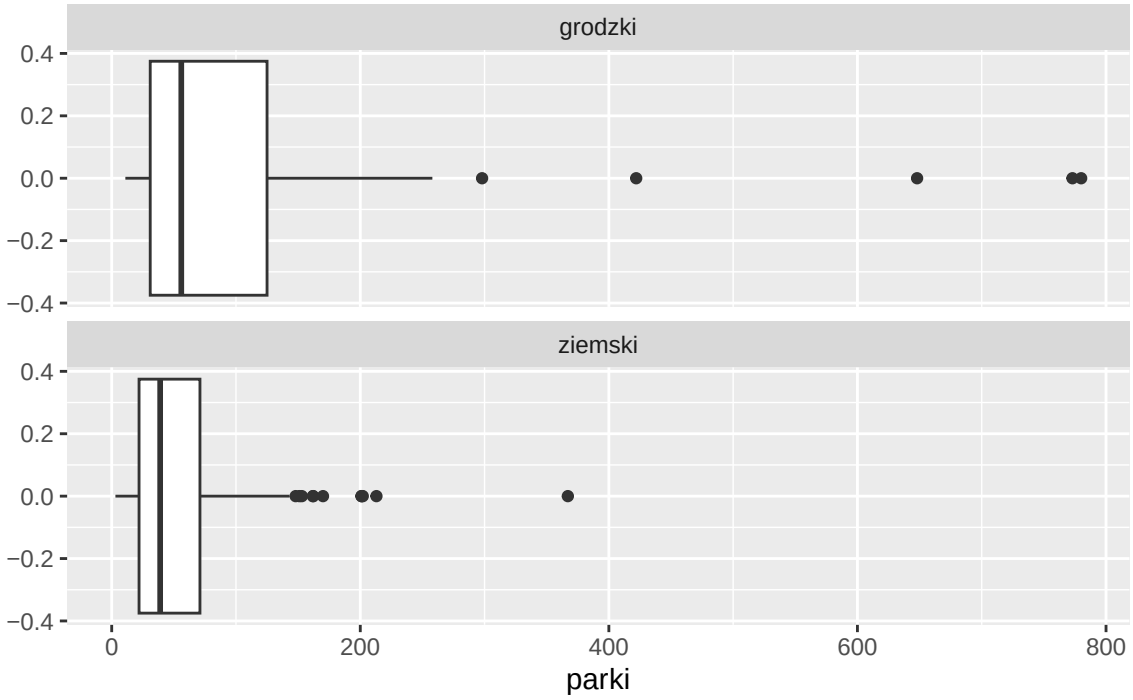
Table 18: Wsp.zmienności

razem	grodzkie	ziemskie
0.3576501	0.3967352	0.3163859

Z wszystkich danych pozbywamy się rekordu o cenie = 0.

Zmienna parki

Rozklad zmiennej parki



Rozklad zmiennej parki dla wszystkich powiatów

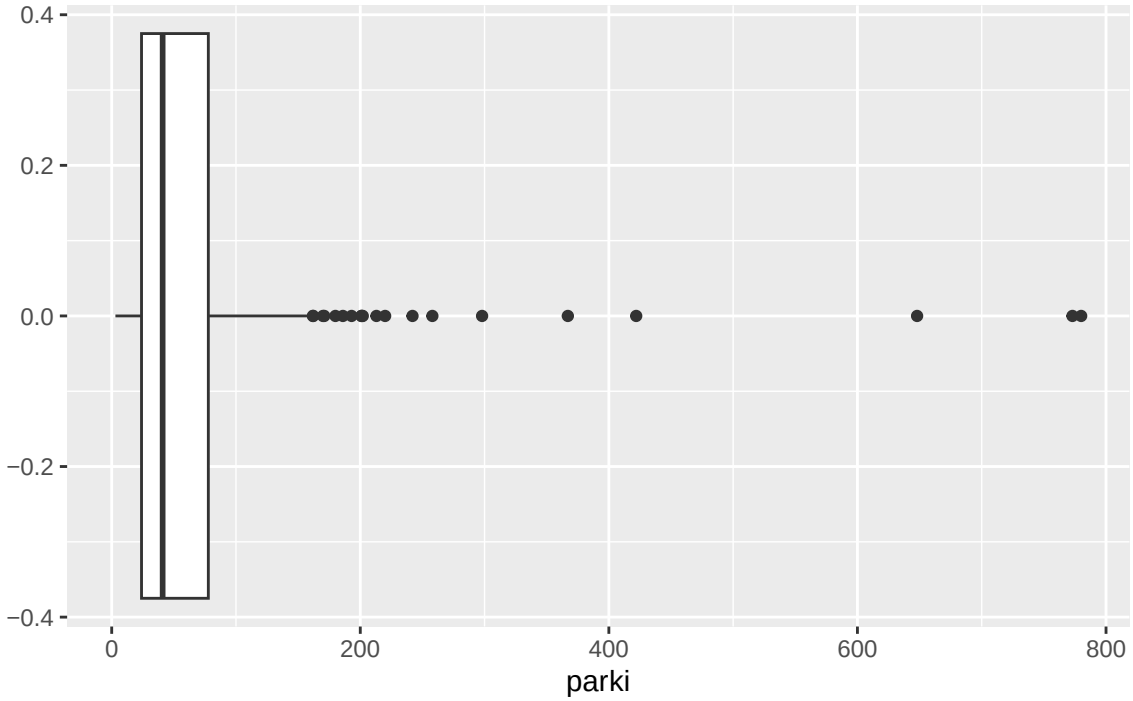


Table 19: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	115.21538	51.57827
odchylenie	158.54066	43.83566
skośność	2.966332	2.192902
kurtoza	11.95290	12.02056

Table 20: Razem

średnia	62.521164
odchylenie	80.220888
skośność	5.412044
kurtoza	43.226127

Table 21: Wsp.zmienności

razem	grodzkie	ziemskie
1.2831	1.376037	0.8498861

Filtrujemy o parki w ilości mniejszej niż 500.

Zmienna dochody brutto na mieszkańca

Table 22: Oddzielnie

rodzaj_powiat	grodzki	ziemski
średnia	6373.743	5705.178
odchylenie	752.7233	617.9805
skośność	1.010243	3.509884
kurtoza	3.354334	26.929773

Table 23: Razem

średnia	5818.674350
odchylenie	689.199519
skośność	2.556991
kurtoza	15.509817

Table 24: Wsp.zmienności

razem	grodzkie	ziemskie
0.1184461	0.1180975	0.1083192

Dla wszystkich danych usuwamy rekord z zerową wartością. Dla danych okrojonych filtrujemy dla rekordów o dochodach poniżej 9000.

Zebrane statystyki opisowe przed oraz po okrojeniu danych

Przed:

Table 25: Wszystkie dane

	samochodyZielone	gęstość	ludnosc	przystanki
średnia	2714.95756	355.011141	9.975072e+04	483.899204
odchylenie	8345.84329	629.236939	1.241824e+05	381.425260
skośność	10.90572	2.485276	8.957797e+00	4.068997
kurtoza	148.06175	8.883204	1.142704e+02	32.350203

Table 26: Wszystkie dane

	odsetek65	cenyMieszkan	parki	ogółem
średnia	19.4522546	2.612562e+05	62.610080	5818.674350
odchylenie	2.4527018	9.339878e+04	80.308840	689.199519
skośność	0.2157754	2.577595e+00	5.405515	2.556991
kurtoza	3.3423821	1.500925e+01	43.127781	15.509817

Table 27: Grodzkie

	samochodyZielone	gęstość	ludnosc	przystanki
średnia	7811.531250	1602.9000000	1.913480e+05	535.48438
odchylenie	18805.695758	651.7864910	2.658631e+05	635.16295
skośność	4.857495	0.8198633	4.363895e+00	3.76431
kurtoza	28.944302	3.9582763	2.602996e+01	20.59918

Table 28: Grodzkie

	odsetek65	cenyMieszkan	parki	ogółem
średnia	22.0234375	3.260003e+05	116.562500	6373.743281
odchylenie	2.0800444	1.282187e+05	159.418574	752.723275
skośność	0.4059703	2.721143e+00	2.941246	1.010243
kurtoza	3.6829048	1.269668e+01	11.778238	3.354334

Table 29: Ziemskie

	samochodyZielone	gęstość	ludnosc	przystanki
średnia	1672.846645	99.851757	81021.562300	473.351438
odchylenie	2472.302332	81.420061	47318.926362	305.300941
skośność	3.914315	3.168285	2.549627	2.422387
kurtoza	21.948904	17.765784	15.260240	13.654796

Table 30: Ziemskie

	odsetek65	cenyMieszkan	parki	ogółem
średnia	18.9265176	2.480178e+05	51.578275	5705.177827
odchylenie	2.1777796	7.846935e+04	43.835661	617.980521
skośność	0.0485062	1.917105e+00	2.192902	3.509884
kurtoza	3.2844884	9.220497e+00	12.020558	26.929773

Po:

Table 31: Wszystkie dane

	samochodyZielone	gęstość	ludnosc	przystanki
średnia	2093.622642	333.453908	90981.951482	460.735849
odchylenie	3530.082176	589.248893	70088.135560	293.858185
skośność	4.856498	2.491810	3.750846	1.883308
kurtoza	36.312768	8.798274	23.464059	8.369964

Table 32: Wszystkie dane

	odsetek65	cenyMieszkan	parki	ogółem
średnia	19.4544474	2.583551e+05	56.159030	5785.419272
odchylenie	2.4635177	8.908663e+04	49.807131	595.836224
skośność	0.2166935	2.643222e+00	2.068828	1.536870
kurtoza	3.3279664	1.668726e+01	9.376056	6.284012

Table 33: Grodzkie

	samochodyZielone	gęstość	ludnosc	przystanki
średnia	4325.633333	1544.9566667	1.451626e+05	430.083333
odchylenie	6315.184869	601.5350148	1.253156e+05	360.395747
skośność	3.157574	0.7051955	2.221973e+00	2.410691
kurtoza	14.819662	3.8532347	8.107905e+00	9.910628

Table 34: Grodzkie

	odsetek65	cenyMieszkan	parki	ogółem
średnia	22.196667	3.127593e+05	80.616667	6288.685333
odchylenie	2.026014	1.167901e+05	68.494239	668.054523
skośność	0.381249	3.443302e+00	1.277761	0.962355
kurtoza	3.936009	1.900796e+01	3.995957	3.299008

Table 35: Ziemskie

	samochodyZielone	gęstość	ludnosc	przystanki
średnia	1663.009646	99.723473	80529.096463	466.649518
odchylenie	2476.440835	81.640777	46866.991747	279.515726
skośność	3.929166	3.167131	2.602118	1.676719
kurtoza	22.003693	17.707802	15.872718	7.328807

Table 36: Ziemskie

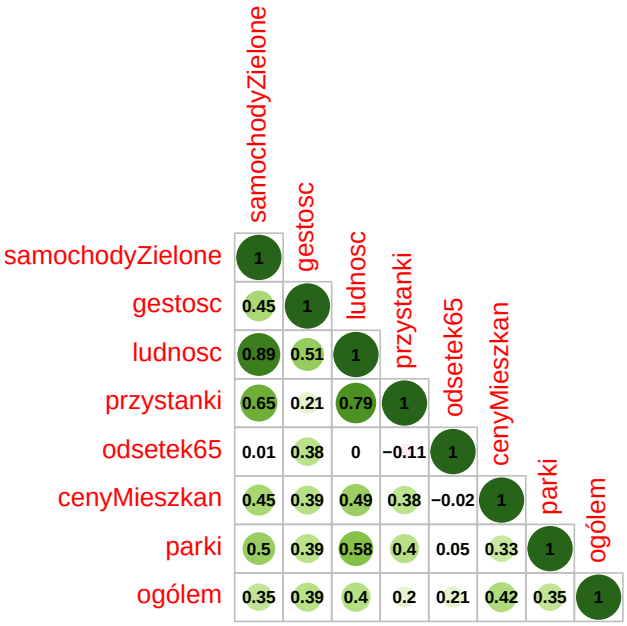
	odsetek65	cenyMieszkan	parki	ogółem
średnia	18.925402	2.478591e+05	51.440514	5688.326141
odchylenie	2.173958	7.869646e+04	43.930541	529.448160
skośność	0.048230	1.918686e+00	2.200649	1.774967
kurtoza	3.313367	9.188917e+00	12.020544	8.369073

N.T współczynników zmienności

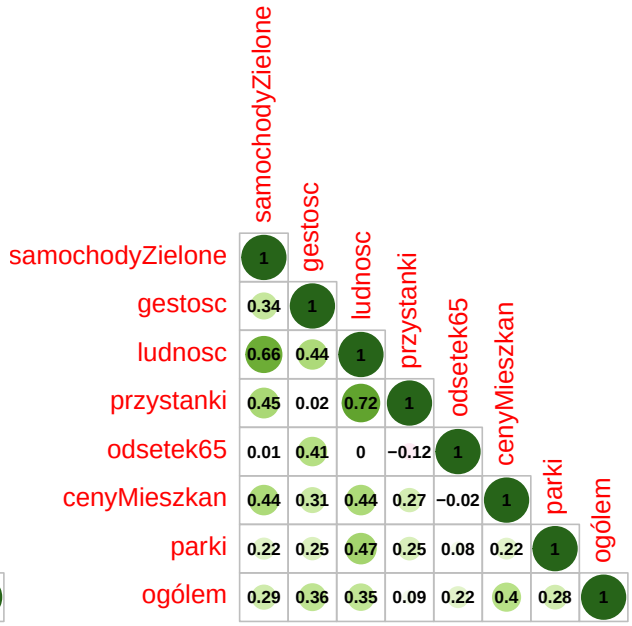
Nie da się ukryć, że dane mają współczynniki zmienności con. duże. Są również bardzo 'skośne', tak więc, wpływ na regresję liniową mogą mieć znaczący. W przypadku regresji liniowej zakłada się, że dane mają rozkład normalny, co oznacza, że są symetryczne i równomiernie rozłożone wokół średniej. Jednak, gdy dane są bardzo skośne, istnieje kilka konsekwencji jak chociażby obciążone estymatory współczynników.

Korelacje (wszystkie dane)

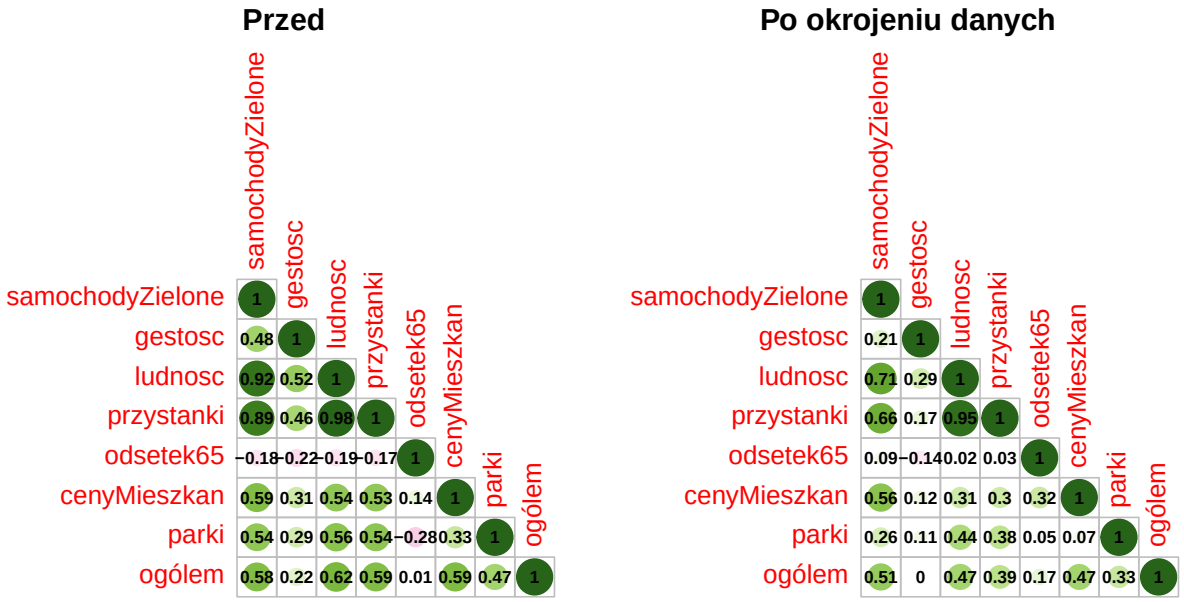
Przed



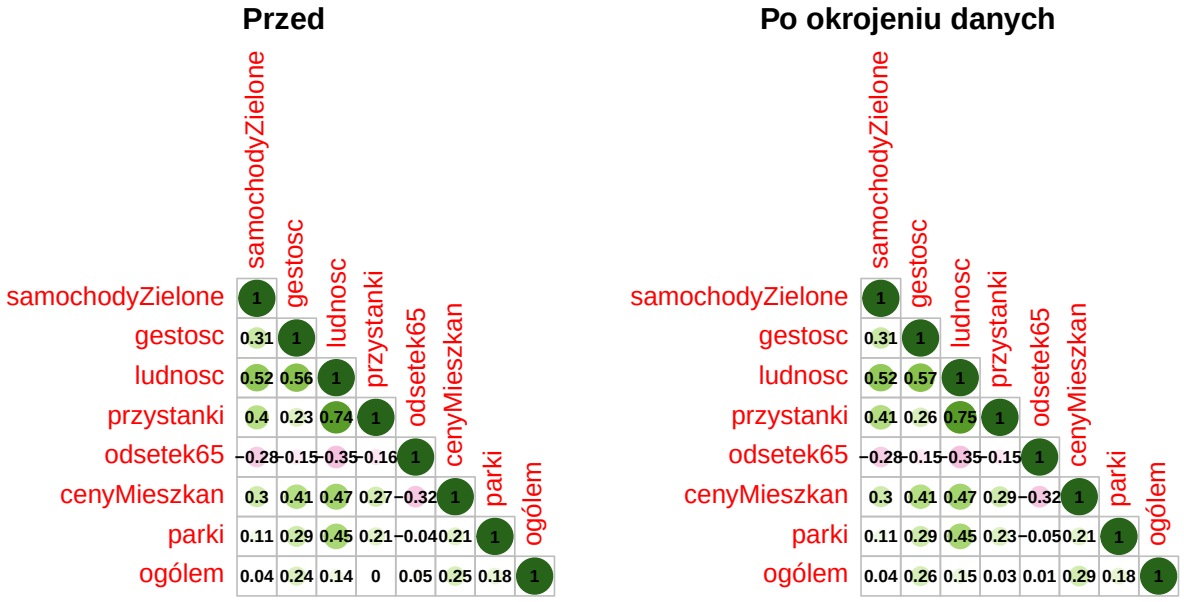
Po okrojeniu danych



Korelacje dla regionów grodzkich



Dla rejonów ziemskich



Jak jeszcze w przypadku obszarów grodzkich można dopatrzeć się korelacji (niestety są również duże korelacje miedzy samymi zmiennymi objaśniającymi, ale o tym potem), tak dla obszarów ziemskich wygląda to nie za pocieszająco.

Metoda Hellwiga

Wszystkie dane

Do analizy zmiennych metodą Hellwiga dla wszystkich danych użyjemy tylko danych okrojonych.

Najwyższe otrzymane wartości:

```
## $gęstość_ludnosc_przystanki_cenyMieszkan
## [1] 0.4288966
##
## $ludnosc_odsetek65_cenyMieszkan
## [1] 0.4351178
##
## $ludnosc
## [1] 0.436369
##
## $ludnosc_odsetek65
## [1] 0.4364408
##
## $ludnosc_cenyMieszkan
## [1] 0.4370459
```

Ciężko zaprzeczyć, że liczba ludności absolutnie dominuje pozostałe zmienne (oczywiście, mogliśmy się tego spodziewać). Zakładając, że chcielibyśmy wziąć kombinację conajmniej 3-cechową, wg. metody Hellwiga wzielibyśmy np. kombinację ludność-odsetek65-cenyMieszkan.

Grodzkie

```
## $ludnosc_cenyMieszkan_ogółem
## [1] 0.592523
##
## $gęstość_ludnosc_przystanki_cenyMieszkan_ogółem
## [1] 0.6040844
##
## $ludnosc_przystanki_cenyMieszkan
## [1] 0.6074616
##
## $ludnosc_przystanki_cenyMieszkan_ogółem
## [1] 0.6109227
##
## $ludnosc_cenyMieszkan
## [1] 0.6200802
```

Sytuacja bardzo podobna, a współczynniki nawet wyższe. Ludność dalej dominuję, lecz wydaje się, że nieco mniej. Zakładając, że chcielibyśmy wziąć kombinację conajmniej 3-cechową, wg. metody Hellwiga wzielibyśmy np. kombinację ludność-przystanki-cenyMieszkan-ogółem(wynagrodzenie).

Ziemskie

```
## $ludnosc_przystanki_odsetek65_cenyMieszkan
## [1] 0.2718565
```

```
##
## $ludnosc
## [1] 0.2738081
##
## $ludnosc_przystanki_odsetek65
## [1] 0.2742738
##
## $gęstość_ludnosc_przystanki_odsetek65_cenyMieszkan
## [1] 0.2746154
##
## $gęstość_ludnosc_przystanki_odsetek65
## [1] 0.2799816
```

Metoda Hellwiga bez ludności

Wszystkie dane

Do analizy zmiennych metodą Hellwiga dla wszystkich danych użyjemy tylko danych okrojonych.

Najwyższe otrzymane wartości:

```
## $gęstość_przystanki_odsetek65_cenyMieszkan
## [1] 0.3282138
##
## $gęstość_przystanki_cenyMieszkan_parki_ogółem
## [1] 0.334001
##
## $gęstość_przystanki_cenyMieszkan_parki
## [1] 0.3381717
##
## $gęstość_przystanki_cenyMieszkan_ogółem
## [1] 0.3566373
##
## $gęstość_przystanki_cenyMieszkan
## [1] 0.3636173
```

W przypadku metody Hellwiga gdy nie bierzemy pod uwagę ludności powiatu widzimy wyraźne pogorszenie informacji przenoszonej przez zmienne. Najlepszą kombinacją zmiennych objaśniających jest: gęstość-przystanki-ceny mieszkań

Grodzkie

```
## $gęstość_przystanki_cenyMieszkan_parki_ogółem
## [1] 0.5422372
##
## $gęstość_przystanki_cenyMieszkan
## [1] 0.5482351
##
## $przystanki_cenyMieszkan
## [1] 0.5716165
##
```

```
## $przystanki_cenyMieszkan_ogółem
## [1] 0.5744857
##
## $gęstość_przystanki_cenyMieszkan_ogółem
## [1] 0.5749817
```

Podobnie jak w wersji z ludnością widzimy znaczny wzrost współczynników. W tym przypadku wybralibyśmy model z następującymi zmiennymi: przystanki-cenyMieszkań-ogółem

Ziemskie

```
## $gęstość_przystanki_odsetek65_cenyMieszkan_parki
## [1] 0.2298593
##
## $gęstość_przystanki_odsetek65_cenyMieszkan_ogółem
## [1] 0.2350637
##
## $gęstość_przystanki_odsetek65_ogółem
## [1] 0.2403353
##
## $gęstość_przystanki_odsetek65_cenyMieszkan
## [1] 0.2480628
##
## $gęstość_przystanki_odsetek65
## [1] 0.2526456
```

Powiaty ziemskie ponownie wykazują najgorsze wskaźniki pojemności informacji.

Wnioski

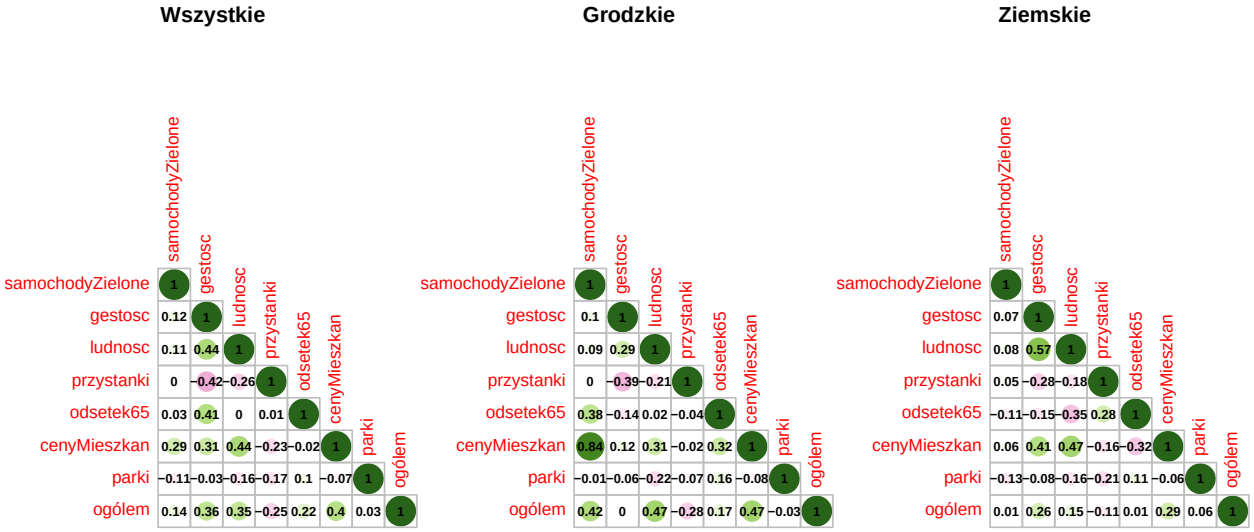
Dane które posiadamy okazują się nie być najlepszymi do tworzenia modelu przewidującego liczbę samochodów zielonych. Jest tak z kilku powodów:

- niskie korelacje między potencjalnymi zmiennymi objaśniającymi a zmienną objaśnianą
- średnie korelacje między zmiennymi objaśniającymi, co może prowadzić do efektu katalizy
- nieistotność zmiennych - dane które podejrzewaliśmy, że mogą mieć wpływ na liczbę aut zielonych tak naprawdę nie mają dużego wpływu.
- brak dostępności zmiennych, które mogłyby mieć większy wpływ - brak informacji na GUSie o liczbie ładowarek elektrycznych w powiatach oraz brak danych o np.: liczbie paneli słonecznych w gospodarstwach domowych. Dane te mogłyby poprawić wyniki naszego modelu
- duża zmienność liczby samochodów elektrycznych w powiatach - np.: w mieście Kraków według danych w 2022 roku było około 30 tysięcy aut elektrycznych, a w mieście Wrocław, który ma mniejszą ludność niż Kraków, było ich około 70 tysięcy. Jest to tylko jeden przykład z wielu i ciężko jest nam wskazać na naszych danych przyczyny tego zjawiska.

Najlepszym doborem zmiennych objaśniających jaki możemy uzyskać na naszych danych była by model : auta zielone~ludność, ceny mieszkań.

Model przeskalowany o ludność

W tym rozdziale trochę przebudujemy nasz model i zbadamy ilość samochodów na 10 000 mieszkańców. Zeskalujemy zmienną samochodów, parków oraz przystanków. (wykorzystamy dane okrojone)



Hellwig

Wszystkie dane

Najwyższe otrzymane wartości:

```
## $gęstość_cenyMieszkan_parki
## [1] 0.08189204
##
## $odsetek65_cenyMieszkan
## [1] 0.08307936
##
## $cenyMieszkan
## [1] 0.08376722
##
## $odsetek65_cenyMieszkan_parki
## [1] 0.08772752
##
## $cenyMieszkan_parki
## [1] 0.08919514
```

Po przeskalowaniu danych przez ludność widzimy bardzo duży spadek pojemności informacyjnej naszych danych. Wszystkie wskaźniki są poniżej 0.1.

Grodzkie

```
## $przystanki_cenyMieszkan_parki
```

```
## [1] 0.6463961
##
## $odsetek65_cenyMieszkan
## [1] 0.6473657
##
## $cenyMieszkan_parki
## [1] 0.6565217
##
## $przystanki_cenyMieszkan
## [1] 0.6954489
##
## $cenyMieszkan
## [1] 0.7071838
```

Co ciekawe widzimy wzrost pojemności informacyjnej w przypadku powiatów grodzkich. Najwyższy wskaźnik miałby model z samą ceną mieszkań (ponieważ powstała bardzo znaczna korelacja) i wynosiłby 0.69, podczas gdy bez skalowania najwyższy wskaźnik był równy 0.61.

Ziemskie

```
## $odsetek65_cenyMieszkan_parki
## [1] 0.02675427
##
## $gęstość_odsetek65_parki_ogółem
## [1] 0.02687461
##
## $gęstość_odsetek65_cenyMieszkan_parki
## [1] 0.02699794
##
## $odsetek65_parki
## [1] 0.02712078
##
## $gęstość_odsetek65_parki
## [1] 0.02832488
```

Zupełnie na odwrót ma się sytuacja dla powiatów ziemskich - tutaj wyszły wskaźniki równe prawie 0. Oznacza to, że skalowanie zmiennych przez ludność działa dobrze w przypadku powiatów grodzkich, a bardzo pogarsza wskaźniki dla powiatów ziemskich