Efekt 3. Wstępna analiza danych

Damian Okoń, Mateusz Nowak, Robert Zamiar

# Opis zbioru danych, statystyki opisowe

W efekcie przekształceń danych otrzymano następujące zmienne:

1 Zmienne wykorzystane w projekcie

|  |  |  |
| --- | --- | --- |
| zmienna | wyjaśnienie | typ zmiennej |
| country | kraj | etykieta |
| GDP\_pc | Produkt krajowy brutto per capita w euro | zmienna numeryczna |
| high\_tech\_trade\_pc | Handel wysokimi technologiami per capita w euro | zmienna numeryczna |
| r&d\_gdp\_pct | Wydatki na badania i rozwój jako procent PKB | zmienna numeryczna |
| r&d\_bud\_pct | Wydatki na badania i rozwój jako procent budżetu | zmienna numeryczna |
| use\_cloud\_pct | Procent ludzi korzystających z chmury | zmienna numeryczna |
| weeknd\_work\_pct | Procent ludzi pracujących w weekendy (15-64) | zmienna numeryczna |
| emp\_deadline\_pct | Procent ludzi pracujących intensywnie i z deadlinem (15-64) | zmienna numeryczna |
| working\_pop\_pct | Procent populacji w wieku produkcyjnym (15-64) | zmienna numeryczna |
| phycisians\_per\_1000 | lekarze na 1000 mieszkańców | zmienna numeryczna |
| sea\_access | dostęp do morza | zmienna kategoryczna |
| joined\_EU | dekada dołączenia do Unii Europejskiej | zmienna kategoryczna |
| is\_euro\_currency | Czy walutą jest euro? | zmienna kategoryczna |
| nuclear\_electricity | Czy kraj posiada elektorwnie atomowe? | zmienna kategoryczna |

Dane numeryczne, etykietę z nazwami krajów (nazwy w języku angielskim), a także zmienne numeryczne zaczerpnięto z serwisu Eurostat, natomiast zmienne kategoryczne zostały utworzone przez zespół na podstawie źródeł z serwisów geopolitycznych. Zmienne te mogą być wykazane w przyszłych, bardziej wnikliwych analizach. Wszystkie zmienne numeryczne podane są w ujęciu procentowym na 1000 mieszkańców, bądź per capita. Zastosowane formuły przy łączeniu zbiorów można podejrzeć w arkuszu dane. Jedyną wartością brakującą w zbiorze była ilość lekarzy dla Czech. Brak ten w finalnym zbiorze uzupełniono obliczając średnią wszystkich wartości wskaźnika liczba lekarzy na 1000 mieszkańców

## Obliczanie zmienności

Przygotowane przez nas dane są przekrojowe – pochodzą z roku 2015, więc nie są szeregiem czasowym. Z powodu przyszłego wykorzystania danych do analizy regresji obliczono współczynnik zmienności będący ilorazem odchylenia standardowego i średniej danych zmiennych – wskaźnik ten obliczono zarówno dla zmiennych numerycznych, jak i binarnych. Tabela z wynikami znajduje się na następnej stronie:

2 Charakterystyka zmiennych

|  |  |  |  |
| --- | --- | --- | --- |
| zmienna | średnia | Odchylenie standardowe | Współczynnik zmienności |
| GDP\_pc | 23677 | 13148,34 | 55,53% |
| high\_tech\_trade\_pc | 863 | 567,75 | 65,81% |
| r&d\_gdp\_pct | 2 | 0,84 | 49,30% |
| r&d\_bud\_pct | 1 | 0,43 | 36,21% |
| use\_cloud\_pct | 25 | 9,00 | 36,09% |
| weeknd\_work\_pct | 24 | 7,98 | 32,68% |
| emp\_deadline\_pct | 43 | 8,68 | 20,07% |
| working\_pop\_pct | 65 | 1,76 | 2,70% |
| phycisians\_per\_1000 | 4 | 0,81 | 22,29% |
| sea\_access | 1 | 0,37 | 44,54% |
| is\_euro\_currency | 1 | 0,49 | 76,55% |
| nuclear\_electricity | 1 | 0,51 | 90,47% |

Możemy zauważyć, że największą zmiennością charakteryzują się zmienne związane z handlem wysokimi technologiami, a także PKB per capita. Zmienne te posiadają dosyć wysokie odchylenie standardowe z powodu zróżnicowania tych wartości w poszczególnych krajach Unii Europejskiej. Zmienne binarne, o najwyższych współczynnikach zmienności z kolei charakteryzują dwa popularne dzisiaj trendy: energię atomową oraz posiadanie przez kraj waluty Euro – czyli czynniki, które zależą od przyjętej przez państwo polityki nuklearnej oraz pieniężnej. Otrzymane przez nas współczynniki zmienności są w większości powyżej wartości granicznej(poza working\_pop\_pct), którą przy analizie regresji będziemy musieli odrzucić przy tworzeniu modelu.

Ponadto dla modelu w dodatku Rcmdr do programu Rstudio obliczono statystyki jak skośność, kurtoza oraz kwantyle zmiennych.

1 Skośność, kurtoza, kwantyle modelu

Obraz zawierający stół

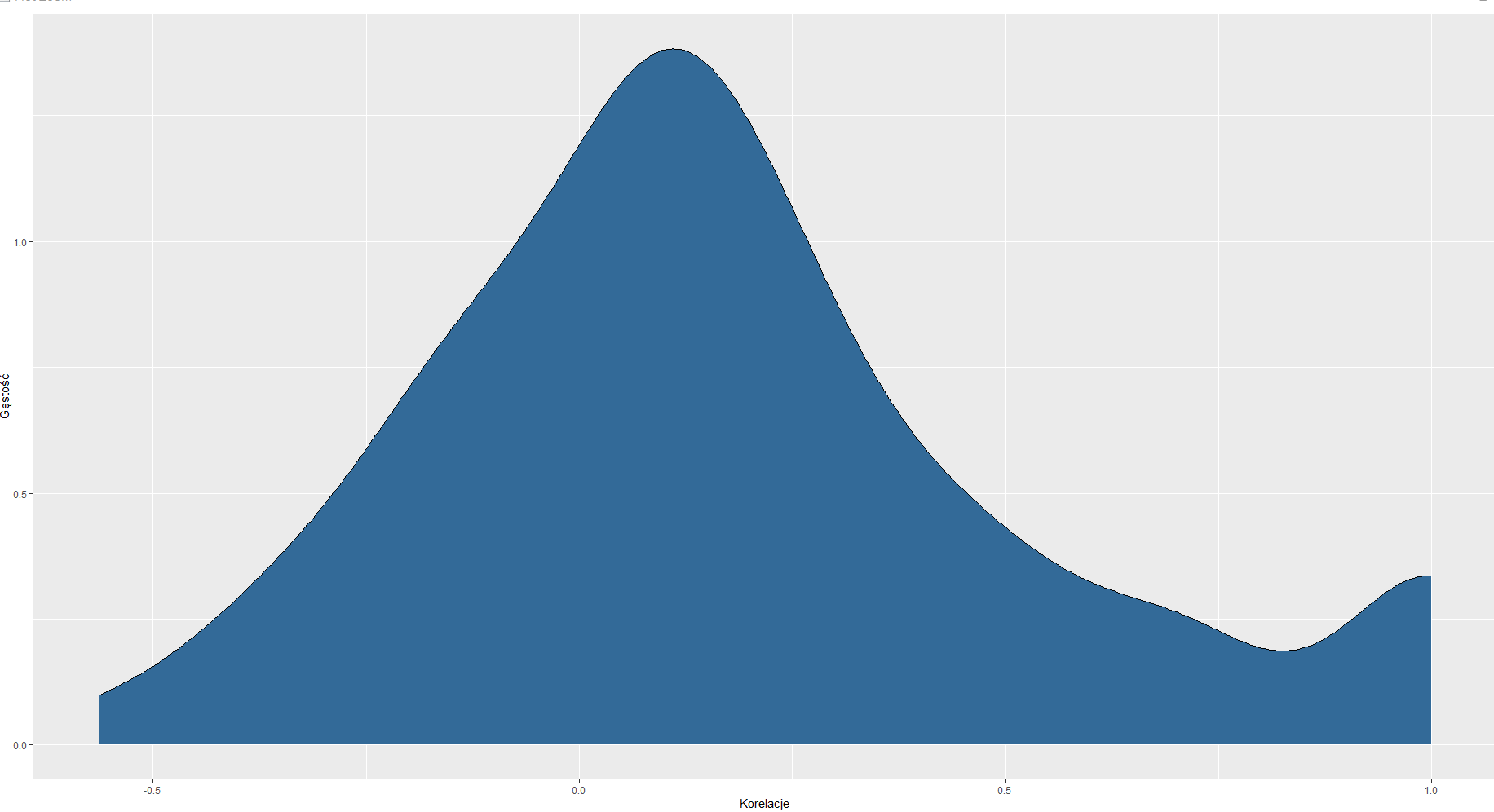
Opis wygenerowany automatycznie

Z powyższych statystyk opisowych możemy dowiedzieć się, że zmienne posiadają asymetryczny rozkład – z czego większość posiada prawostronny. Z kolei wyniki kurtozy informują nas, że w większości badanych przez nas zmiennych rozkład jest mniej wysmukły niż rozkład normalny(k<0) – istnieje mniej dodatnich wartości niż w przypadku rozkładu normalnego.

## Korelacje i ich istotność

Na początku – w celu lepszej wizualizacji macierzy korelacyjnej, na jej podstawie wykonano wykres gęstości korelacji między zmiennymi.

2 Wykres gęstości korelacji



Dzięki wykonanej wizualizacji możemy dowiedzieć się, że wartości korelacji posiadają rozkład raczej symetryczny, a ich najwięcej wartości statystyki zawiera się w przedziale [0;0.3]. Następnie wyznaczono macierz koleracyjną.

Obraz zawierający stół

Opis wygenerowany automatycznie3 Macierz korelacyjna

Najbardziej skorelowane były: zmienna GDP\_pc oraz zmienne odpowiadające min. Handlami wysokimi technologiami oraz użyciem technologii „cloud” czy wydatkami na badania rozwojowe. Przyjęte przez nas założenia zgadzały się z wizualizacją na wykresie gęstości – korelacje w większości przypadków są słabe. Założenia potwierdza macierz wartości p-value wykonana w programie Rcmdr. Można z niej odczytać, że najwięcej istotnych korelacji zachodzi między PKB per capita, a ww. zmiennymi.

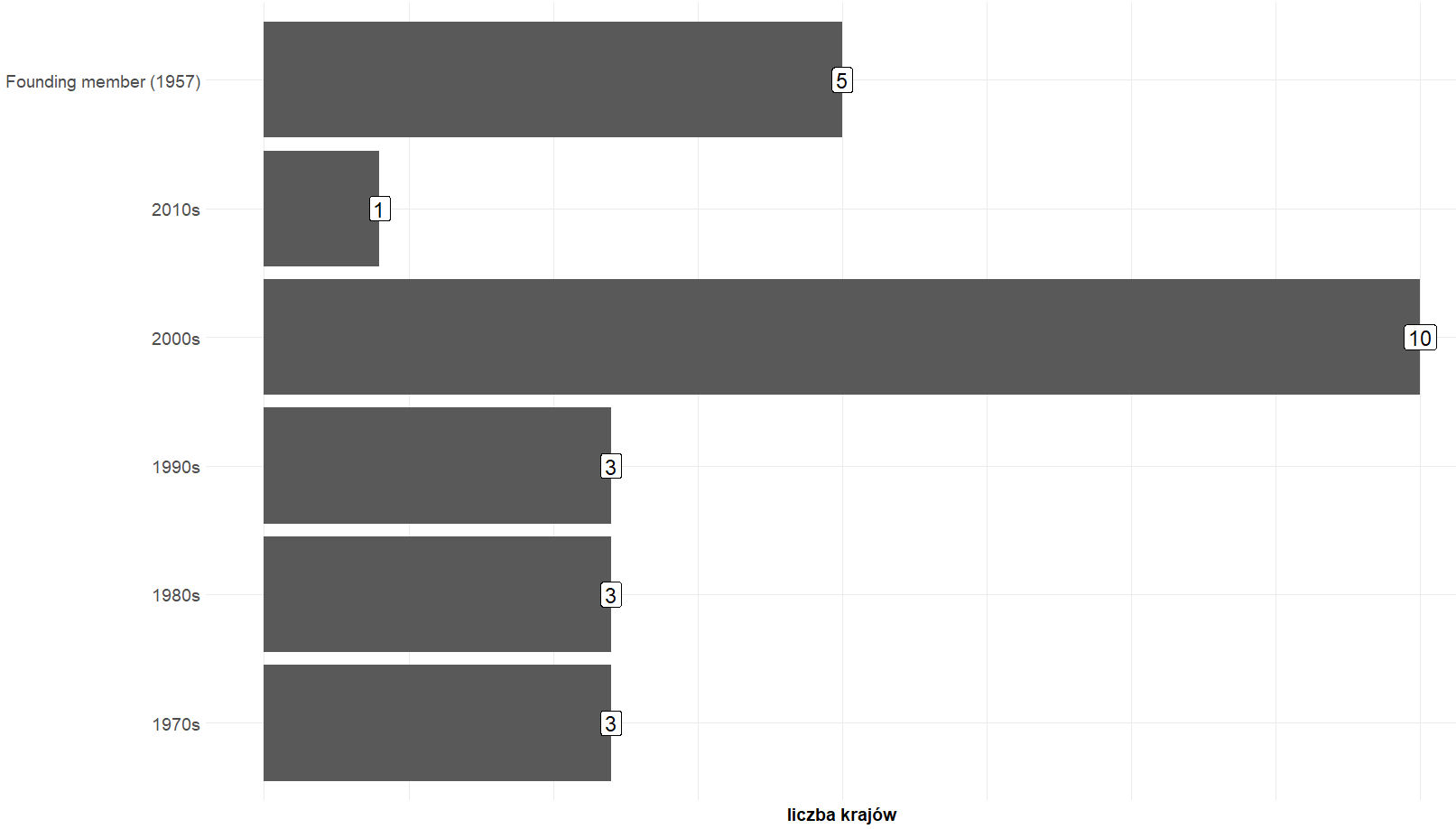
Obraz zawierający stół

Opis wygenerowany automatycznie4 Macierz p-value

## Wstępny pomiar zmiennej kategorycznej

Zmienna „joined\_EU” jest zmienną utworzoną w celu grupowania krajów, zmienna ta została utworzona w celu przyszłej analizy skupień lub klasyfikacji. Dla zmiennej tej niemożliwe jest wykonanie ww. analiz statystycznych, lecz można utworzyć histogram pokazujący rozkład zmiennej – zliczanie rekordów.

3 Liczba krajów w zależności od czasu dołączenia do Unii Europejskiej



Jak możemy zauważyć – większość krajów obecnej Unii Europejskiej, to kraje, które dołączyły do Wspólnoty w 1. Dekadzie lat 2000.