AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ ZARZĄDZANIA

Sprawozdanie z Projektu

Praktyczne wykorzystanie testów statystycznych

Autorzy: *Izabela Guła*

Kierunek: *Informatyka i ekonometria (stacjonarne)*

Przedmiot: *Rachunek prawdopodobieństwa i statystyka matematyczna*

Kraków, 2024

Spis treści

Wstęp

Celem projektu było wykonanie statystycznej analizy na temat wybranego zagadnienia. W pierwszej części należało samodzielnie wybrać dane z Internetu. W tym celu posłużyłam się stroną https://www.gapminder.org/, wybierając dane samodzielnie i tworząc z nich zbiór danych. Kolejnym krokiem było scharakteryzowanie i opisanie wybranych zmiennych za pomocą statystyk opisowych lub wizualizacji. Do badania należało użyć testów statystycznych, a ewentualnie także modelowania regresji liniowej lub metod analizy wizualnej. Analizę należało przeprowadzić w dowolnym oprogramowaniu statystycznym, ja posłużyłam się językiem R oraz programem JASP.

Wybranie i przetworzenie danych

Przy użyciu Gapmindera postanowiłam wybrać zmienne, które miały sprostać pytaniom badawczym. Moim założeniem było zbadanie aspektów związanych z rozwojem społeczno-gospodarczym krajów i czynników, które mogą wpływać na pewne zmienne. Wybrane zmienne to: populacja, PKB per capita, procentowy udział rolnictwa w PKB, procentowy udział przemysłu w PKB, poziom urbanizacji, emisja dwutlenku węgla na osobę, produkcja energii odnawialnej. Proces wczytania oraz oczyszczenia danych zaczyna się od wczytania każdego pliku CSV. Ze względu na brak danych w niektórych zmiennych ustawiam badany zakres na lata 1990 – 2015. W zbiorze danych postanowiłam połączyć kontynent Ameryki Północnej oraz Ameryki Południowej. Kolejnym aspektem, na który się zdecydowałam było usunięcie wszystkich państw z Australii oraz Oceanii, zdecydowałam się na taki krok przez duże braki danych na tych kontynentach oraz wartości odstające, które mogłyby zaburzyć analizę i interpretację danych. Ostatnim krokiem było usunięcie brakujących danych.

Opisanie i scharakteryzowanie zmiennych

W tej części zajęłam się przedstawieniem i przeanalizowaniem każdej zmiennej oraz jej rozkładu przy pomocy statystyk opisowych oraz wizualizacji. Przedstawione nazwy w nawiasach symbolizują nazwy, które będą używane podczas analizy danych w RStudio.

- Zmienna populacja (pop)

Zmienna ta oznacza całkowitą populacje oznaczająca liczbę mieszkańców na terytorium danego państwa.

Statystyki populacji

Descriptive Statistics

	pop_milion												
	1990	1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014
Mean	47.556	47.573	48.043	42.521	45.438	45.783	46.670	47.955	48.287	49.189	50.360	50.887	51.779
Std. Deviation	152.451	152.579	152.962	143.481	147.332	148.718	151.281	154.970	156.390	159.526	162.662	164.682	167.786
Minimum	0.262	0.277	0.292	0.269	0.274	0.281	0.288	0.292	0.304	0.318	0.318	0.321	0.328
Maximum	1150.000	1180.000	1210.000	1230.000	1250.000	1260.000	1280.000	1300.000	1310.000	1330.000	1350.000	1370.000	1390.000

Przy użyciu oprogramowania JASP stworzyłam tabele obliczającą podstawowe statystyki opisowe dotyczące populacji w wybranych latach. Dla ułatwienia przedstawiłam populacje w milionach. W populacji można zauważyć ciągły wzrost. W moim zbiorze danych posiadam państwa o bardzo dużej populacji, jak i takie o bardzo małej. Odchylenie standardowe w 2014 roku wynosiło 167, 79 co świadczy o dużej różnorodności.

- <u>Zmienna przedstawiająca PKB per capita (PKB)</u>

Produkt krajowy brutto na osobę. Dane skorygowane pod względem różnic w sile nabywczej. PKB wyrażone jest w międzynarodowych dolarach, skorygowane pod względem inflacji bazując na roku 2017 - parytet siły_nabywczej bazowany jest na ICP z roku 2017.

Statystyki PKB

Descriptive Statistics

	PKB												
	1990	1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014
Mean	13654.191	13557.968	13755.505	15304.500	16532.806	17558.535	17927.512	18924.899	20356.160	21101.508	21133.348	22162.597	22607.726
Std. Deviation	15331.936	15601.802	15889.505	16729.529	17588.175	18779.197	18882.915	19783.416	20412.772	20065.421	19454.256	21073.890	21026.936
Minimum	893.000	482.000	507.000	532.000	614.000	653.000	678.000	727.000	768.000	815.000	842.000	507.000	614.000
Maximum	82100.000	82500.000	79800.000	86300.000	87100.000	94700.000	99400.000	104000.000	111000.000	117000.000	114000.000	112000.000	113000.000

	kontynent	mean	sd
1	Africa	5663.456	5326.472
2	Americas	13817.262	10865.608
3	Asia	18607.446	21576.128
4	Europe	31794.200	19190.954

Statystyki dotyczące populacji wskazują na ciągły wzrost gospodarczy państw w naszym zbiorze danych, wykazuje to przede wszystkim średnia, która z biegiem lat jest coraz większa. Dane charakteryzują się wysokim odchyleniem standardowym co świadczy o tym, że w badanym zbiorze danych są państwa o bardzo wysokim PKB oraz takie o bardzo niskim. Najwyższym średnim PKB mogą pochwalić się państwa w Europie, natomiast największe zróżnicowanie występuje na kontynencie azjatyckim.

- <u>Zmienna przedstawiająca procentowy udział rolnictwa w PKB (agriculture)</u>

Jest to procentowy udział rolnictwa w wartości PKB. Obejmuje on leśnictwo, łowiectwo i rybołówstwo, a także uprawę roślin i produkcję zwierzęcą.

Statystyki rolnictwo

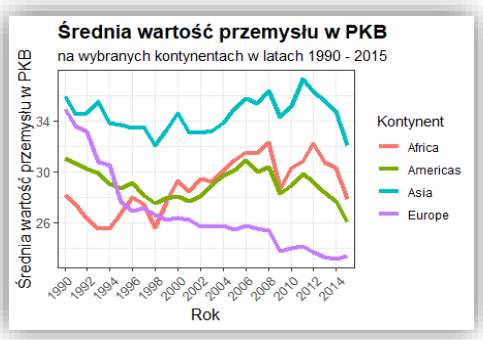
Descriptive Statistics

	agriculture												
	1990	1992	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014
Mean	18.653	18.187	17.530	15.981	15.090	13.613	13.386	12.657	11.681	11.439	11.158	10.965	10.513
Std. Deviation	13.591	13.938	13.667	13.502	13.266	12.766	12.908	11.974	11.716	11.845	11.294	11.161	10.282
Minimum	0.323	0.197	0.173	0.150	0.111	0.087	0.071	0.056	0.049	0.041	0.036	0.033	0.035
Maximum	56.900	63.800	56.500	54.200	61.400	76.100	79.000	65.000	63.800	65.200	52.900	50.800	51.800

Statystyki dotyczące rolnictwa wskazują na spadek średniego udziału rolnictwa w PKB. W zbiorze danych można zauważyć, że istnieją państwa, w których rolnictwo posiada ponad połowę udziału w PKB.

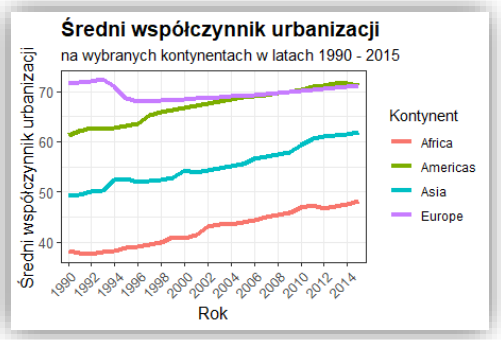
- Zmienna przedstawiająca procentowy udział przemysłu w PKB (industry)

Jest to procentowy udział przemysłu w PKB. Obejmuje wartość dodaną w górnictwie, przemyśle przetwórczym (również wykazywanym jako odrębna podgrupa), budownictwie, energii elektrycznej, wodzie i gazie. Największe udziały przemysłu w PKB można obserwować w Azji. Znaczny spadek po roku 1990 zarejestrowała Europa, jednak patrząc na najwcześniejsze dane, można zaobserwować, że na każdym kontynencie utrzymuje się trend spadkowy.



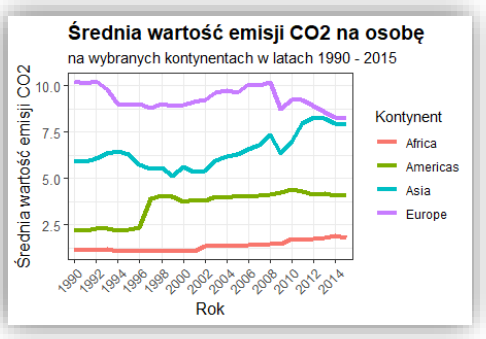
- Zmienna przedstawiająca poziom urbanizacji (urban)

Jest to procentowy udział mieszkańców miast w ogólnej liczbie ludności. Na każdym z badanych kontynentów można zaobserwować wzrost współczynnika urbanizacji, oznacza to, że coraz więcej osób mieszka w miastach, a co za tym idzie badane państwa rozwijają się. Najniższy współczynnik urbanizacji posiada Afryka, co było do przewidzenia. Natomiast najwyższe wartości tego współczynnika zarejestrowały państwa europejskie.



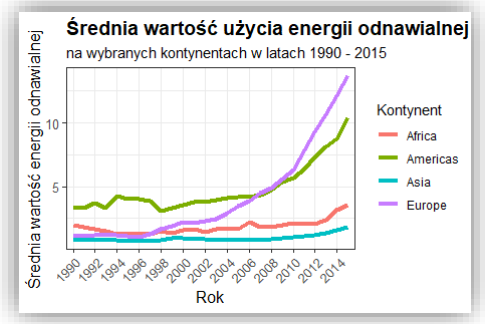
- Zmienna przedstawiająca emisję dwutlenku węgla na osobę (CO2)

Emisja dwutlenku węgla pochodzącego ze spalania paliw kopalnianych w tonach na osobę (metryczne tony dwutlenku węgla na osobę). Niestety najwyższe wartości emisji dwutlenku węgla można zaobserwować w Europie, jednak po roku 2010 można zobaczyć tendencje spadkową, co może być już optymistyczną wiadomością. W Azji średnia emisja dwutlenku węgla ciągle rośnie i prawdopodobnie teraz prześcignęła już Europę. Najniższe wartości posiada Afryka.



- Zmienna przedstawiającą produkcję energii odnawialnej (renewable)

Jest to procent użycia energii odnawialnej w stosunku do całkowitego użycia energii. Na podstawie wykresu można stwierdzić, że najmniejszy średnią produkcję energii odnawialnej posiada Azja. Optymistyczną wiadomość, którą można wywnioskować z tego wykresu jest fakt, że na każdym kontynencie obserwuje się wzrost użycia odnawialnych źródeł energii. Najwyższe wartości odnotowała Europa.



Problem badawczy I

Zależność między udziałem przemysłu i rolnictwa w PKB a wykorzystaniem odnawialnej energii. Który z tych czynników przyczynia się do większego wykorzystywania źródeł odnawialnej energii przez wybrane państwa. Badanie ma na celu sprawdzenie, czy stopień sektorów gospodarki może wpływać na wykorzystywanie źródeł energii przez państwa.

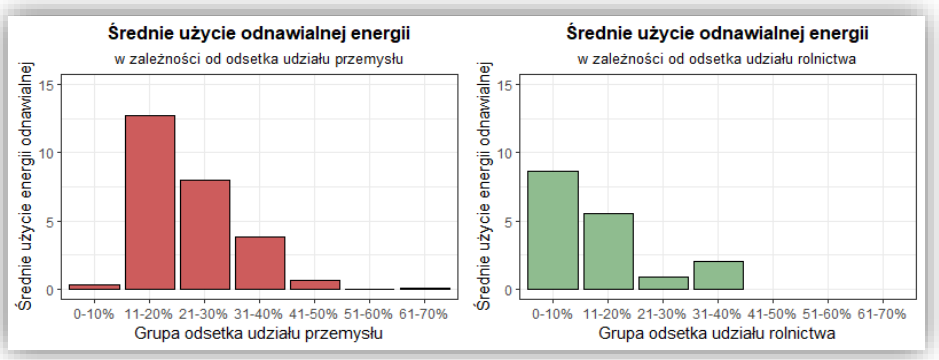
- Statystyki opisowe

W tej części zdecydowałam się użyć statystyk opisowych do zbadania przedstawionego problemu. Zdecydowałam się wybrać po 30 państw z 2015 roku z najwyższym i najniższym udziałem przemysłu w PKB oraz z najwyższym i najniższym udziałem rolnictwa w PKB. Następnie dla każdej grupy obliczyłam średnia arytmetyczną, wartość minimalną, wartość maksymalną, medianę oraz odchylenie standardowe dla wartości użycia energii odnawialnej. Państwa z najwyższą wartością przemysłu w gospodarce charakteryzują się niską średnią użycia energii odnawialnej, gdyż jest to zaledwie 2%. Istnieje jednak państwo w tym przedziale osiągające aż 25% użycia energii odnawialnej. Państwa z najniższą wartością przemysłu gospodarce posiadają już wyższą średnią wartość dotyczącą użycia energii odnawialnej. Zbiór ten charakteryzuje się wyższym odchyleniem standardowym, co świadczy o różnych wartościach użycia energii odnawialnej. Najwyższa wartość rolnictwa w PKB charakteryzuje się średnią wynoszącą prawie 4% oraz bardzo niską mediana, są to natomiast wartości wyższe niż w kategorii z najwyższym przemysłem. Jednak wartość maksymalna jest już wyższa, tak samo jak odchylenie standardowe, nie jest to zadziwiający aspekt, gdyż przy takiej niskiej średniej kategoria ta posiada państwo osiągające nawet 48% użycia energii odnawialnej. Najniższa wartość rolnictwa posiada drugą najwyższą średnią użycia energii odnawialnej. Państwa te posiadają taką sama wartość maksymalna jak w przypadku państw z najniższym udziałem przemysłu, odchylenie standardowe jest również podobne w tej kategorii. Wspólnym punktem każdej z tych kategorii jest to, że znalazło się w nich przynajmniej jedno państwo posiadające zerowe użycie źródeł odnawialnej energii. Podsumowując badanie przy użyciu statystyk opisowych uważam, że największy potencjał do wysokiego użycia źródeł energii odnawialnej mogą posiadać państwa z najniższym udziałem przemysłu w PKB oraz z najniższym udziałem rolnictwa w PKB.

	Kategoria	srednia	mediana	min	max	odchylenie_standardowe
1	Najwyższy przemysł	1.817620	0.04065	0	25.1	4.916836
2	Najniższy przemysł	12.291900	7.33000	0	65.4	15.265637
3	Najwyższe rolnictwo	3.975013	0.20000	0	48.3	11.377968
4	Najniższe rolnictwo	10.307743	6.23500	0	65.4	13.829989

- Wizualizacje

W drugiej części badania problemu zdecydowałam się za pomocą wizualizacji zbadać, w których grupach odsetka dany sektor gospodarki posiada najwyższe użycie źródeł odnawialnej energii. Do wykonania tej wizualizacji najpierw stworzyłam grupy odsetków, które pokażą mi w jakim przedziale procentowym danego sektora gospodarki występują dane wartości dotyczące odnawialnej energii. Następnie utworzyłam wykresy, w których dla danego odsetka obliczana jest średnia z użycia źródeł odnawialnej energii.



Zaczynając od przemysłu, największe użycie energii odnawialnej posiadają państwa, które posiadają udział przemysłu w PKB między 11 – 20%. Wartości zaczynają się stopniowo zmniejszać z zwiększeniem się udziału tego sektora. Może to potwierdzać, że państwa z większym odsetkiem przemysłu w PKB wykorzystują mniej odnawialnej energii. Jeśli chodzi o rolnictwo, to największe zauważalne użycie odnawialnych źródeł energii występuje w najniższym odsetku wynoszącym od 0 do 10%. W tym przypadku jest podobnie, że im odsetek się zmniejsza to maleje wykorzystywanie czystej energii, poza odsetkiem 31 – 40%, w którym jak się domyślam mogą występować wartości odstające. Podsumowując badanie przy użyciu wizualizacji może wykazywać, że wraz z wzrostem udziału danego sektora gospodarki, średnie użycie źródeł odnawialnej energii zmniejsza się . Najwyższe wartości można zaobserwować w grupie z 11 – 20% udziale przemysłu w gospodarce oraz w niskim udziale rolnictwa w gospodarce.

- Testy statystyczne

W trzeciej części badania mojego pytania badawczego posłużę się testem statystycznym. Do swojego badania postanowiłam użyć testu t – Studenta dla prób niezależnych. Test t Studenta dla prób niezależnych to parametryczny test służący do porównania średnich między dwiema niezależnymi od siebie grupami oraz, aby stwierdzić, czy wyniki w jednej grupie są większe

bądź mniejsze niż w drugiej grupie i czy te różnice są istotne statycznie. Postanowiłam przeprowadzić dwa takie testy, jeden dla przemysłu, a drugi dla rolnictwa.

```
        Welch Two Sample t-test

data:  combined_selected_values_industry$industry_low and combined_selected_values_industry$industry_high
t = -0.16308, df = 188.76, p-value = 0.8706
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.305258  1.953196
sample estimates:
mean of x mean of y
 3.867227  4.043258
```

```
        Welch Two Sample t-test

data:  combined_selected_values_agriculture$agriculture_low and combined_selected_values_agriculture$agriculture
_high
t = 4.6484, df = 129.82, p-value = 8.113e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.949311 4.838120
sample estimates:
mean of x mean of y
 4.283246  0.889531
```

Badanie przeprowadziłam wybierając wartości od roku 2000, gdyż wtedy użycie odnawialnych źródeł energii zaczynało rosnąć. W każdym dla grupy z niskim udziałem oraz wysokim danego sektora, założyłam, że niski udział to taki poniżej 20%, a średni / wysoki to tauodnawialnej energii. Dane z przemysłu oraz rolnictwa będą analizowane oddzielnie w celu sprawdzenia, czy różnice są istotne statycznie. Wynik t – statystyki dla przemysłu informuje o różnicy między grupami w jednostkach standardowych. Wartość p- value wynosi 0,87 i jest większą niż założony poziom istotności 0.05, może to sugerować, że nie mamy podstaw do odrzucenia hipotezy zerowej o braku różnicy między średnimi. Na podstawie wyników testu, nie ma statystycznie istotnej różnicy między średnimi wartościami w grupie. Wyniki testu t – studenta wskazują, że poziom udziału przemysłu w gospodarce nie musi mieć wpływu na użycie odnawialnej energii. Przy danych dotyczących rolnictwa wartosc statystyki t jest już większa, oznacza to większą różnicę między średnimi w badanuch grupach. Wartość p – value jest bardzo mała i bliska zeru, a co najistotniejsze jest mniejsza niż założony poziom istotności. Może to sugerować, że mamy podstawy do odrzucenia hipotezy o braku różnicy między średnimi. Test wykazuje, że poziom udziału rolnictwa w gospodarce może mieć wpływ na użycie odnawialnych źródeł energii.

- Podsumowanie

Podsumowując wszystkie wykorzystane metody uważam, że mogę stwierdzić, że poziom danego sektora gospodarki wpływa na wykorzystywanie przez państwa źródeł odnawialnej energii. Potwierdzają to badania przeprowadzone na statystykach opisowych oraz wizualizacje, które pokazały że największy potencjał do wysokiego użycia źródeł energii odnawialnej mogą posiadać państwa z najniższym udziałem przemysłu oraz z najniższym udziałem rolnictwa w

PKB. Na koniec test t – Studenta wykazuje, że w przypadku rolnictwa poziom udziałów w PKB może mieć wpływ na użycie źródeł odnawialnej energii, jednak przy przemyśle już nie.

Problem badawczy II

Zależność pomiędzy współczynnikiem urbanizacji a emisją gazów cieplarnianych. Analiza ma na celu zbadanie, czy państwa, które posiadają wyższy współczynnik urbanizacji, emitują więcej gazów cieplarnianych. Współczynnik urbanizacji stanowi kluczowy wskaźnik stopnia urbanizacji danego obszaru, będący proporcją ludności miejskiej do całkowitej populacji. Emisja gazów cieplarnianych, z kolei, jest jednym z głównych czynników wpływających na zmiany klimatyczne. W ramach analizy, uwzględniono okres czasowy obejmujący lata 1990–2015.

- <u>Najwyższe i najmniejsze wartości CO2</u>

CO2	urban	kraj	rok
0.016	35.60	Congo, Dem. Rep.	2001-01-19
0.017	36.10	Congo, Dem. Rep.	2002-01-19
0.019	35.10	Congo, Dem. Rep.	2000-01-19
0.020	36.50	Congo, Dem. Rep.	2003-01-19
0.022	37.00	Congo, Dem. Rep.	2004-01-19
0.024	32.00	Haiti	1994-01-19

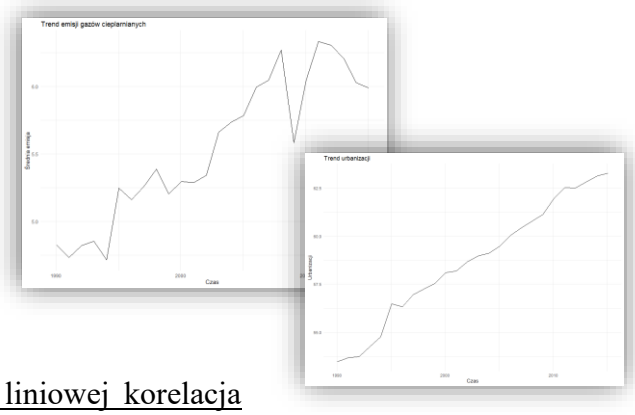
CO2	urban	kraj	rok
47.5	100.0	Singapore	2013-01-19
45.5	100.0	Singapore	2012-01-19
41.6	100.0	Singapore	2011-01-19
40.9	100.0	Singapore	2008-01-19
34.7	82.6	UAE	2006-01-19
33.9	100.0	Singapore	2014-01-19

Podjęłam decyzję o wstępnej analizie danych w tabeli.

Warto zauważyć, że najwyższe poziomy emisji gazów cieplarnianych występują w krajach, gdzie wskaźnik urbanizacji jest szczególnie wysoki. Szczególnie dominujący jest Singapur, gdzie wskaźnik ten osiąga 100. Obserwujemy, że najniższe poziomy emisji dwutlenku węgla (CO2) występują w krajach, takich analizowanych obszarów o największych emisjach, gdzie ten wskaźnik sięgał nawet 100. To spostrzeżenie stanowi istotny aspekt analizy, sugerując, że obszary o mniejszej emisji CO2 charakteryzują się niższym stopniem urbanizacji. Takie kontrastowe zestawienie wskazuje na potencjalną odwrotną zależność między poziomem emisji a intensywnością urbanizacji, co może być kluczowym elementem dalszej analizy

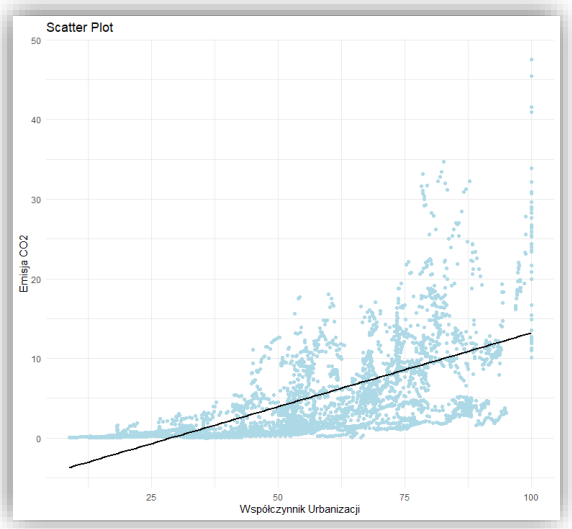
- Trendy czasowe

Obserwowane dane wskazują na istnienie dodatnich trendów czasowych zarówno w emisji gazów cieplarnianych, jak i procesie urbanizacji. W miarę upływu czasu obie zmienne wykazują tendencję wzrostową, co stanowi interesujący punkt wyjścia do dalszej analizy.



- Wykres punktowy i model regresji liniowej korelacja Pearsona

Na etapie wstępnego badania, przeprowadzę analizę za pomocą wykresu Scatter Plot, który uwzględni zarówno dane punktowe, jak i linię regresji liniowej. Obserwując prezentowany wykres, można zauważyć, że w miarę wzrostu współczynnika urbanizacji, następuje równoczesny wzrost emisji dwutlenku węgla (CO2). Warto zaznaczyć, że to spostrzeżenie znajduje potwierdzenie w formie linii regresji, która wyraźnie manifestuje rosnący trend wraz ze wzrostem wartości współczynnika urbanizacji na naszym wykresie. To odkrycie sugeruje, że istnieje pozytywna korelacja między stopniem urbanizacji a emisją CO2, co może stanowić punkt wyjścia do głębszej analizy tej zależności.



Wyniki regresji liniowej można przedstawić w formie tabeli zawierającej konkretne wartości liczbowe dla współczynników, błędów standardowych, statystyk t i p-value. Jak można zauważyć, wartość estimate wynosi około *0,1854*, co oznacza, że wraz ze wzrostem współczynnika urbanizacji o jednostkę, oczekujemy wzrostu emisji dwutlenku węgla o 0.1854 jednostki. Wartość p-value równa 0 wskazuje na silne dowody na to, że współczynnik regresji dla urban jest statystycznie istotnie różny od zera, co wskazuje na zależność między współczynnikiem urbanizacji a emisją dwutlenku węgla.

	term	estimate	std.error	statistic	p.value
1	urban	0.1854229	0.003947845	46.96812	0

- Test korelacji Pearsona

```
        Pearson's product-moment correlation

data:  dane$urban and dane$CO2
t = 46.968, df = 3187, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6185922 0.6596358
sample estimates:
      cor
0.6395696
```

W analizie zależności między stopniem urbanizacji a emisją gazów cieplarnianych zastosowałam test korelacji Pearsona. Otrzymane rezultaty ukazują silną, dodatnią korelację między wspomnianymi zmiennymi, gdzie oszacowany współczynnik korelacji wyniósł 0,64. Statystyka *t = 46,968* dla *df = 3917* (w odniesieniu do liczby obserwacji) jest znacząco wysoka, co sugeruje, że obserwowana korelacja między urbanizacją a emisją CO2 jest niewykluczona.. Statystyczna istotność tego związku została potwierdzona bardzo niską wartością *p-value < 2,2e-16* (mniejszą od 0,05), co daje nam podstawy do odrzucenia hipotezy o braku związku między stopniem urbanizacji a emisją gazów cieplarnianych. Wnioskiem z analizy jest stwierdzenie, że zwiększający się stopień urbanizacji jest skorelowany z wyższą emisją gazów cieplarnianych.

- Test korelacji Tau-Kendalla

```
        Kendall's rank correlation tau

data:  dane$CO2 and dane$urban
z = 47.018, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5560915
```

Kierując się hipotezą zerową, zakładającą brak zależności (Tau-Kendall równa zero) między emisją gazów cieplarnianych a współczynnikiem urbanizacji, wyniki naszego testu są znaczące. Wartość p-value, która nie przekracza poziomu istotności 0,05 (a nawet jest znacznie mniejsza), dostarcza mocnych dowodów na odrzucenie hipotezy zerowej. Otrzymane wyniki, w tym wartość współczynnika Tau (0.5560915), potwierdzają istnienie statystycznie istotnej zależności między badanymi zmiennymi. Wzrost jednej zmiennej jest skorelowany ze wzrostem drugiej, co jest dodatkowo potwierdzone statystyką testu z = 47.018. Wartość ta informuje nas o odległości estymowanej wartości Tau od zera, wyrażonej w jednostkach odchylenia standardowego. Te rezultaty umożliwiają nam założenie, że istnieje pozytywna, statystycznie istotna zależność pomiędzy emisją gazów cieplarnianych a współczynnikiem urbanizacji. Wzrost urbanizacji jest skorelowany ze wzrostem emisji gazów cieplarnianych, co może być istotnym aspektem w kontekście.

- Podsumowanie

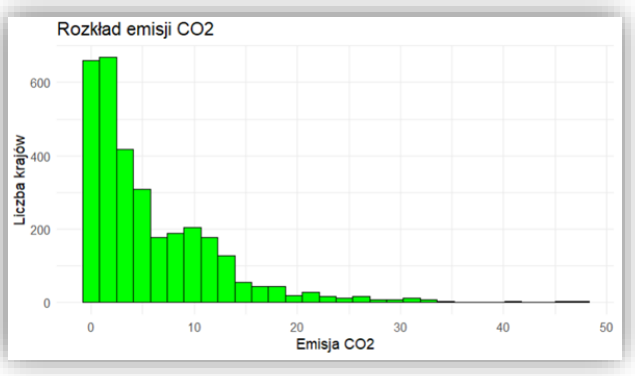
Analiza wskazuje, że istnieje zauważalna zależność pomiędzy stopniem urbanizacji a emisją gazów cieplarnianych. Najwyższe poziomy emisji występują w krajach o wysokim wskaźniku urbanizacji, co sugeruje, że rozwój miejski może wpływać na zwiększenie emisji gazów cieplarnianych. Wartości współczynnika korelacji oraz wyniki testów statystycznych potwierdzają istnienie tej zależności. Jednakże, choć obserwowana korelacja jest istotna, należy podkreślić, że urbanizacja to tylko jeden z wielu czynników wpływających na emisję gazów cieplarnianych. Kompleksowość tej relacji wymaga uwzględnienia innych czynników, takich jak technologie, polityki energetyczne, czy świadomość ekologiczna społeczeństwa.

Problem badawczy III

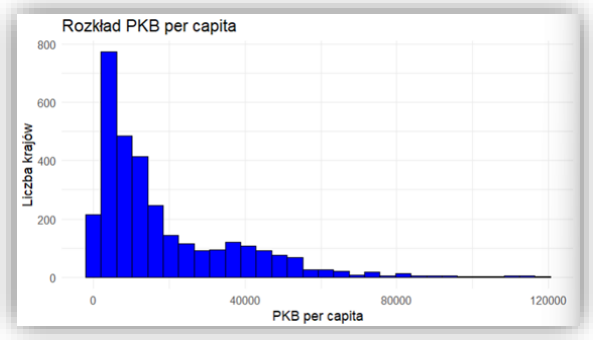
Badanie ma na celu **zbadanie zależności między wielkością PKB per capita a emisją dwutlenku węgla (CO2) w krajach.** Rozważę to, czy kraje o wyższym PKB per capita emitują więcej gazów cieplarnianych. Przeprowadzę testy statystyczne, takie jak Test Wilcoxona oraz analizy korelacji Pearsona i rangowego Spearmana, aby lepiej zrozumieć związki między tymi dwiema zmiennymi.

- Analiza Rozkładu Emisji CO2

W badaniu wykorzystam dane obejmujące informacje o emisji CO2 dla różnych krajów. Rozkład emisji CO2 zaprezentowałam na wykresie. Histogram pokazuje, jak liczba krajów rozkłada się w zależności od poziomu emisji dwutlenku węgla. Zauważamy, że większość krajów ma niższą emisję, ale istnieje również pewna liczba krajów o wyższych poziomach emisji.



- Analiza Rozkładu PKB per Capita



W badaniu wykorzystam dane obejmujące informacje o PKB per capita dla różnych krajów. Wykres prezentuje rozkład PKB per capita wśród krajów. Podobnie jak w przypadku emisji CO2, histogram ten ukazuje, jak rozmieszczone są kraje ze względu na swoje PKB per capita. Widzimy, że większość krajów ma niższe PKB per capita, ale istnieje pewna liczba krajów o wyższych wartościach.

- Test Wilcoxona

Następnym krokiem w projekcie było wykonanie Testu Wilcoxona do porównywania PKB per capita i emisji dwutlenku węgla dla różnych krajów. Celem testu było sprawdzenie, czy istnieje istotna statystycznie różnica między medianami rozkładów.

```
        Wilcoxon rank sum test with continuity correction

data:  dane$PKB and dane$CO2
W = 10169721, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Wyniki testu Wilcoxona są prezentowane w formie statystyki testowej (W), gdzie wartość W wynosi 10,169,721 (Mierzy siłę efektu i kierunek różnicy między medianami. Im wyższa wartość W, tym większa jest różnica między grupami, sugeruje większą różnicę w rangach między zmiennymi) oraz wartości p-wartości (p-value), gdzie wartość p jest bardzo mała, poniżej poziomu istotności 0,05, a dokładnie wynosi < 2.2e-16 (znacznie mniejsze niż 0,05). (Na tej podstawie mogę stwierdzić, że istnieją istotne statystycznie różnice między rozkładami PKB a emisją CO2) P-value jest na tyle niskie, że odrzucamy hipotezę zerową, sugeruje to, że istnieje pewne rzeczywiste statystyczne przesunięcie (różnica w medianach) między grupami, a nie jest to przypadek.

- Analizy Korelacji Pearsona

W przypadku badania związku między PKB a emisją CO2, chciałam ocenić, czy istnieje statystycznie istotna liniowa korelacja między tymi dwiema zmiennymi.

```
        Pearson's product-moment correlation

data:  dane$PKB and dane$CO2
t = 102.04, df = 3187, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8666256 0.8829090
sample estimates:
      cor
0.8750146
```

Statystyka t (t-value): wynosi 102.04. To miara tego, jak bardzo różnice między obserwowanymi danymi a wartościami oczekiwanymi są duże w porównaniu z błędami losowymi, co oznacza, że dane różnią się od wartości oczekiwanych, jednak nie są to ogromne różnice. Stopnie swobody (df): wynoszą 3187, co odnosi się do liczby obserwacji (po odjęciu dwóch). Wartość p (p-value) została zinterpretowana wyżej. Jej wartość jest taka sama. Hipoteza alternatywna: stwierdza, że rzeczywista korelacja nie jest równa zeru.
Odrzucenie hipotezy zerowej wskazuje, że istnieje statystycznie istotna liniowa korelacja między PKB a emisją CO2. Przedział ufności (confidence interval): wynosi od 0.8666 do 0.8829. Oznacza to, że z 95% pewnością możemy stwierdzić, że rzeczywista korelacja między PKB a emisją CO2 mieści się w tym przedziale. Korelacja (sample estimate): wynosi 0.8750. Oznacza to, że obserwowane dane sugerują silną dodatnią korelację między PKB a emisją CO2. Na tej podstawie mogę wywnioskować, że **kraje o wyższym PKB per capita mają tendencję do wykazywania wyższych poziomów emisji dwutlenku węgla**, istnieje statystycznie istotna zależność liniowa między tymi dwiema zmiennymi.

- Analiza korelacji rangowego Spearmana

Korelacja rangowa jest używana, gdy zależność między zmiennymi może być nieliniowa, a analiza oparta na rangach danych jest bardziej odporna na wpływ odstających obserwacji. Oceniając korelację rangową, możemy uzyskać pełniejszy obraz związku między PKB a emisją CO2, niezależnie od ewentualnych nieliniowych trendów lub skoków w danych. W tym przypadku statystyka S (Spearman's rank correlation): wynosi 483,799,232, wartość p (p-value) oraz alternatywna hipoteza są takie same jak w poprzednich korelacjach. Korelacja rangowa

```
        Spearman's rank correlation rho

data:  dane$PKB and dane$CO2
S = 483799232, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9104938
```

(sample estimate): (rho) wynosi 0,9104938. Oznacza to, że dane sugerują bardzo silną pozytywną korelację rangową między PKB a emisją CO2.

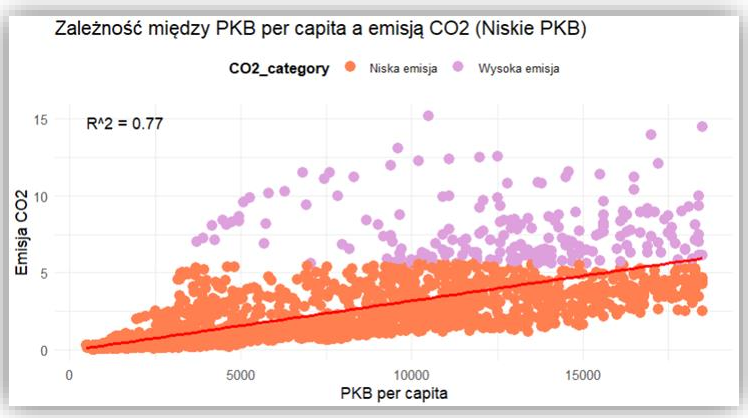
Wartość rho wynosząca 0.91 wskazuje, że **kraje o wyższym PKB per capita mają tendencję do zajmowania wyższych miejsc w rankingu emisji dwutlenku węgla**.

- Podział ze względu na wielkość PKB per capita

W trakcie analizy dodałam również kategorie do danych w celu stworzenia dodatkowego wymiaru analizy. Kategorie PKB są określone na podstawie wartości średniej PKB, dzieląc kraje na "Wysokie PKB" i "Niskie PKB". Podobnie, kategorie CO2 są określone na podstawie średniej emisji CO2, dzieląc kraje na "Wysoką emisję" i "Niską emisję". Kategorie stworzyłam, aby pomogły w porównywaniu związku między PKB a emisją CO2 w różnych kontekstach gospodarczych.

- Wysokie PKB per capita VS emisja CO2

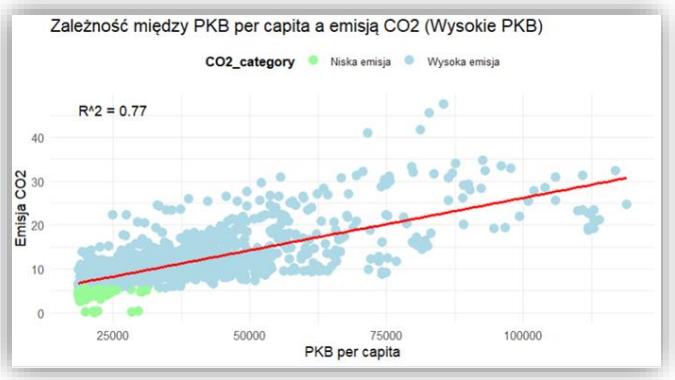
Na wykresie dla kategorii "Wysokie PKB" obserwujemy punkty danych przedstawiające kraje o wysokim PKB per capita. Kolor punktów wskazuje, czy dany kraj ma wysoką czy niską emisję CO2.



Zauważamy ogólny trend wzrostu emisji CO2 wraz ze wzrostem PKB per capita, co potwierdza pozytywną zależność między tymi zmiennymi. Czerwona linia reprezentuje dopasowaną prostą za pomocą modelu regresji liniowej. Tekst z wartością R2 umieszczony jest w górnym lewym rogu. R2 to wartość współczynnika determinacji (R2=0.76) wskazuje, że około 76% zmienności emisji CO2 można wyjaśnić za pomocą zmienności PKB per capita. Wysoki współczynnik sugeruje, że model regresji liniowej dobrze dopasowuje się do danych, wyjaśniając znaczną część zmienności emisji CO2 na podstawie PKB per capita.

- Niskie PKB per capita VS emisja CO2

Na wykresie dla kategorii "Niskie PKB" obserwujemy punkty danych dla krajów o niższym PKB per capita. Ponownie, kolor punktów wskazuje, czy dany kraj ma wysoką czy niską emisję CO2. Mimo niższego PKB per capita, obserwujemy podobny trend wzrostu emisji CO2 wraz ze wzrostem PKB. Czerwona linia to dopasowana prosta z modelu regresji liniowej. Tekst z wartością R2 znajduje się w górnym lewym rogu. Współczynnik R2 również utrzymuje się na stosunkowo wysokim poziomie, co może wskazywać na istnienie związku między tymi zmiennymi, nawet dla krajów o niższych dochodach. Wartość współczynnika determinacji (R2=0.76) wskazuje, że około 76% zmienności emisji CO2 można wyjaśnić za pomocą zmienności PKB per capita.



- Podsumowanie

Przeprowadzone badanie potwierdziło istotną statystycznie zależność między PKB per capita a emisją dwutlenku węgla (CO2) w badanych krajach. Analiza uwzględniła różnorodne testy statystyczne, korelacje oraz model regresji, wskazując na silną dodatnią relację między poziomem rozwoju gospodarczego a emisją CO2, co zostało dodatkowo potwierdzone wykresami dla różnych kategorii ekonomicznych.