AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Ekonometria Przestrzenna

Projekt 3

Jakub Le Van, Mateusz Mulka

Informatyka i Ekonometria Rok V

Spis treści

1. Wstęp

Projekt obejmuje stworzenie modelu liniowego – w tym przypadku regresji liniowej, która pozwoli zbadać czynniki, które wpływają na wzrost lub zmniejszenie odpadów komunalnych w wybranych powiatach Polski, W analizie wybraliśmy sześć zmiennych objaśniających, które są następujące:

1) **Gęstość zaludnienia** (gestoscZaludnienia) – gęstość zaludnienia na kilometr kwadratowy,
2) **Liczba ludności** (ludnoscTys) – ludność wyrażona w tysiącach mieszkańców,
3) **Wskaźnik urbanizacji** (wskUrban) - procentowy udział obszarów miejskich w powierzchni powiatu,
4) **Turystyka** - liczba turystycznych obiektów noclegowych w regionie,
5) **Wynagrodzenie** - przeciętne miesięczne wynagrodzenie brutto w złotych,
6) **Bezrobocie** - stosunek liczby zarejestrowanych bezrobotnych do liczby cywilnej ludności aktywnej zawodowo,

Celem projektu jest określenie, które z wybranych czynników maja istotne statystycznie oddziaływanie na na wzrost/spadek odpadów komunalnych, zbadanie czy różnice przestrzenne istnieją w zależnościach wraz z interpretacja otrzymanych wyników. Każda z opisanych wyżej zmiennych objaśniających została włączona do modelu na podstawie jej potencjalnego wpływu, wedle naszej oceny, na ilość odpadów komunalnych w powiatach Polski. Gęstość zaludnienia odzwierciedla zagęszczenie populacji na danym obszarze, co może być kluczowe, ponieważ zagęszczone tereny często sprzyjają różnym wydarzeniom podczas którym produkowane są śmieci, dodatkowo wysoka gęstość zaludnienia często widziana jest w dużych miastach, gdzie generowana jest większa liczba śmieci. Wskaźnik urbanizacji ma na celu sprawdzenie tej zależności, podobnie jak liczba ludności w danym powiecie – co w naturalny sposób przyczynia się do liczby wyprodukowanych odpadów na tym obszarze. Turystyka również może być jednym z kluczowych czynników, turyści często generują znaczne ilości odpadów, dodatkowo nie zaliczają się oni do ludności danego powiatu, a są osobami przyjezdnymi. Turyści chętniej uczestniczą w różnych aktywnościach, a także jedzą na zewnątrz, co sprzyja produkcji odpadów. Kolejnym czynnikiem mogącym mieć istotny wpływ jest poziom wynagrodzenia – bogatsze osoby mogą przejawiać tendencję do produkcji większej ilości

odpadów z uwagi na większą konsumpcję. Bezrobocie natomiast może wskazywać pewne wartości odstające z wcześniejszej kategorii – osoby które nie zarabiają nic mogą przypuszczalnie produkować znacznie mniej śmieci, a z uwagi na fakt, że zmienna odpowiadająca za wynagrodzenie jest wartością uśrednioną to warto dodatkowo uwzględnić poziom bezrobocia.

2. Dane do projektu

Poniżej przedstawiliśmy dane dla pięciu pierwszych powiatów:

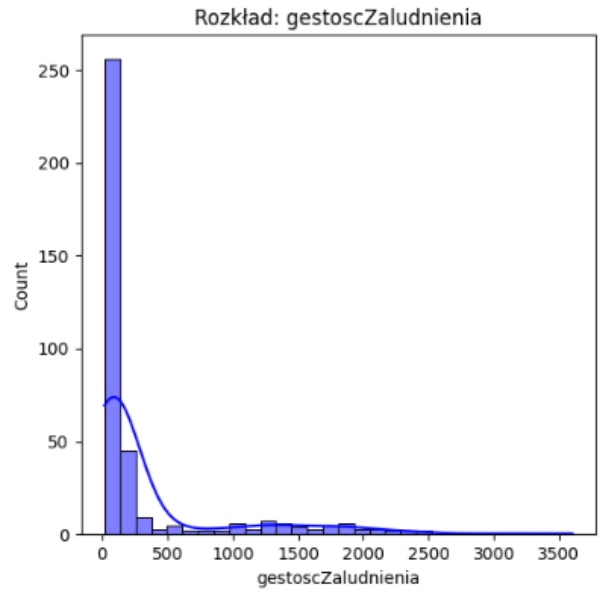
	gestoscZaludnienia	ludnoscTys	wskUrban	turystyka	wynagrodzenie	bezrobocie	odpadyTysTon
0	67.5	87.92	47.1	17	6181.60	3.1	10.56
1	200.3	95.86	78.2	19	5723.65	5.2	59.76
2	193.3	85.69	73.4	9	5856.76	6.4	17.89
3	44.3	32.74	41.1	1	5423.11	13.8	25.50
4	81.9	47.58	53.6	14	6142.84	10.2	31.36

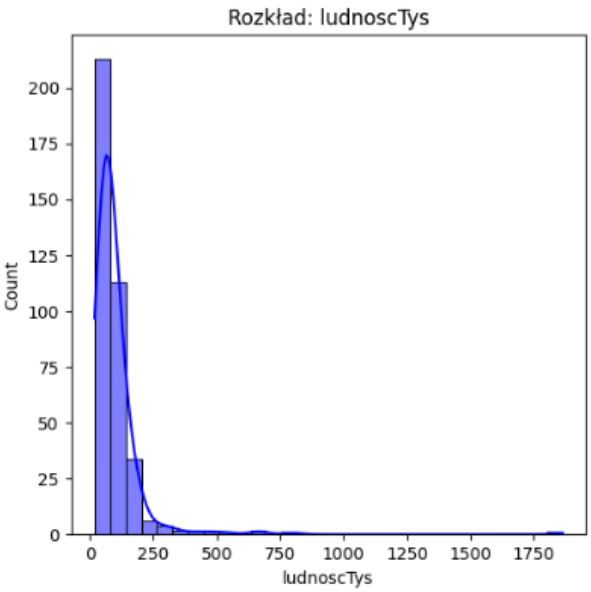
Sprawdziliśmy strukturę danych oraz obecność brakujących wartości. Okazało się, że dane były kompletne – wszystkie kolumny zawierały pełne zestawy informacji.

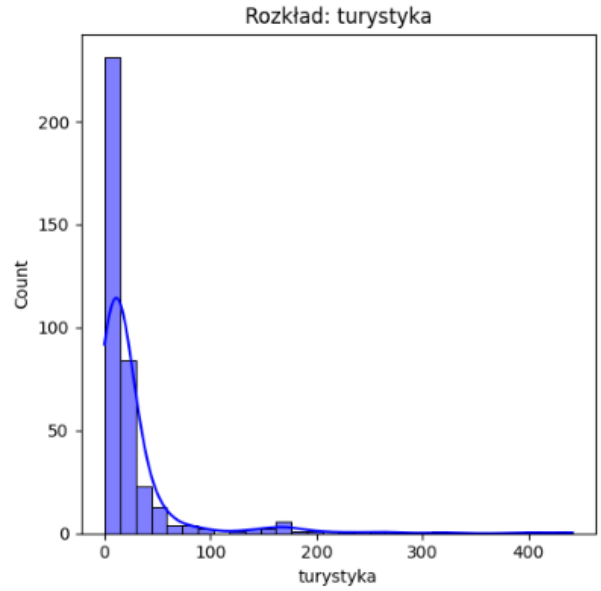
Obliczyliśmy podstawowe miary statystyczne dla każdej zmiennej.

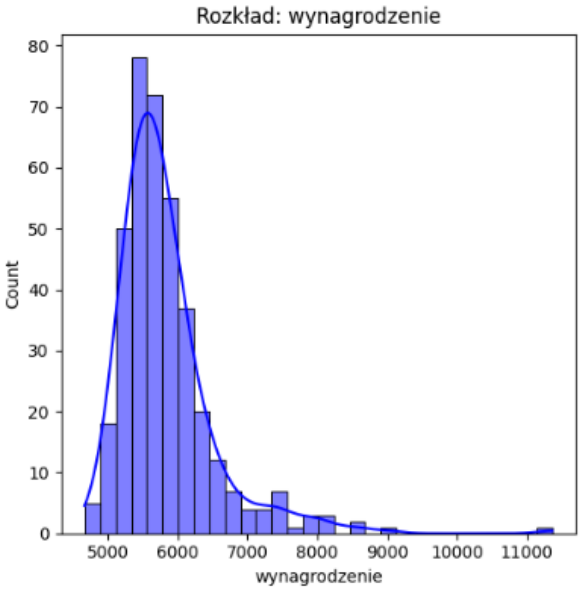
Analiza wykazała, że zmienne charakteryzowały się dużym zróżnicowaniem. Przykładowo, średnia gęstość zaludnienia wyniosła 355 osób/km², natomiast rozkład wartości zawierał zarówno bardzo niskie (18 osób/km²), jak i ekstremalnie wysokie wartości (3600 osób/km²)

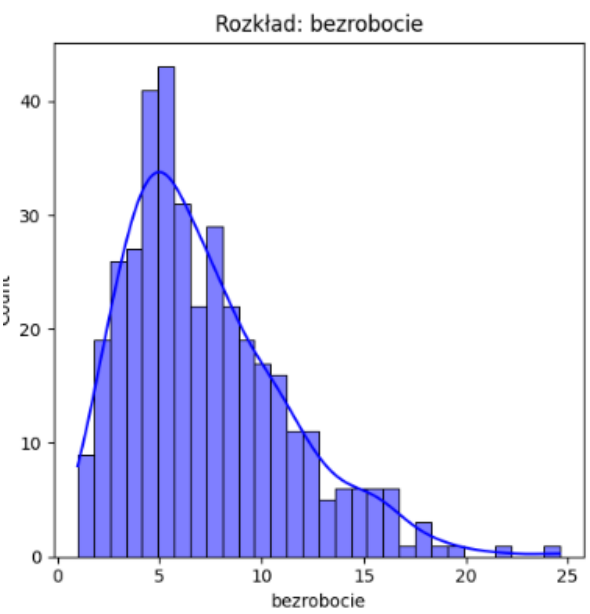
Przeprowadziliśmy wizualizację rozkładów zmiennych liczbowych za pomocą histogramów, co pozwoliło nam zidentyfikować ewentualne odstające wartości i lepiej zrozumieć charakterystykę danych.

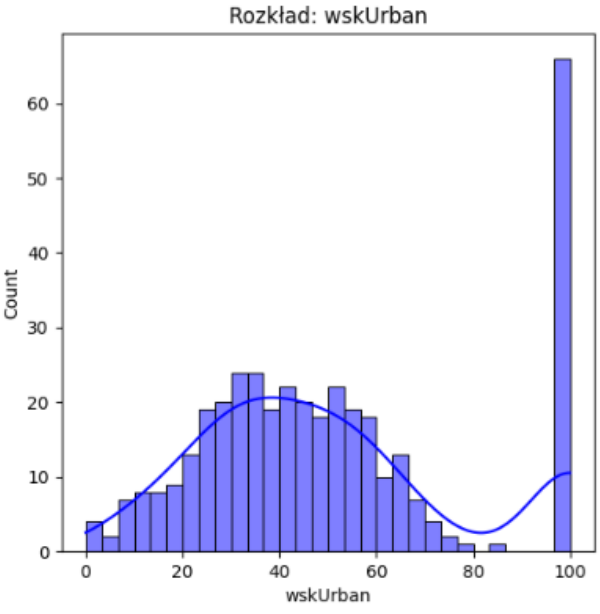

Rozkład: gestoscZaludnienia
Count
gestoscZaludnienia


Rozkład: ludnoscTys
Count
ludnoscTys

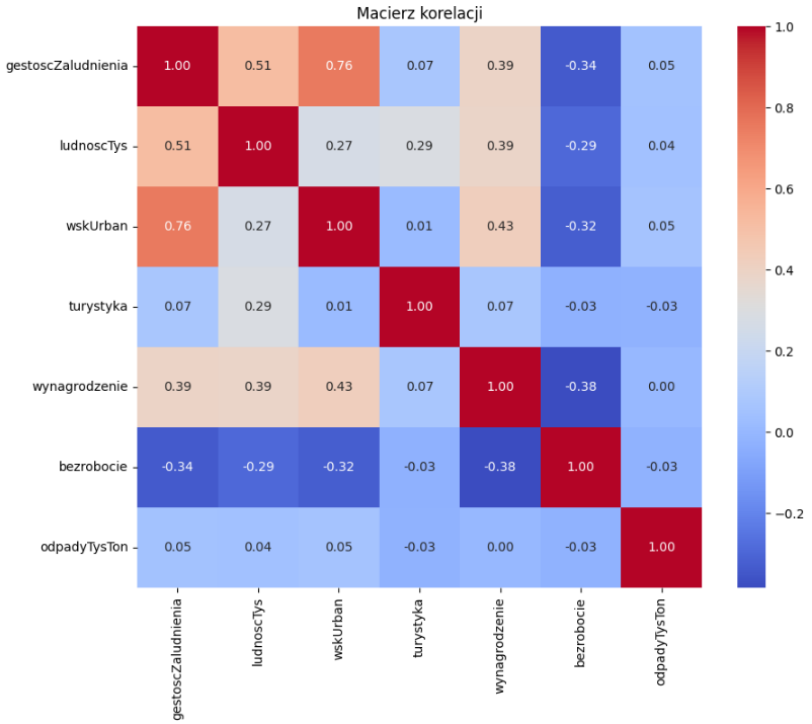

Rozkład: turystyka
Count
turystyka


Rozkład: wynagrodzenie
Count
wynagrodzenie


Rozkład: bezrobocie
bezrobocie


Rozkład: wskUrban
Count
wskUrban

Po zwizualizowaniu rozkładów zmiennych przeanalizowaliśmy ich wzajemne powiązania, generując macierz korelacji. Wyniki wykazały, że niektóre zmienne były silnie skorelowane między sobą, co mogłoby negatywnie wpłynąć na wyniki regresji liniowej. Szczególną uwagę zwróciła wysoka korelacja między wskaźnikiem urbanizacji a gęstością zaludnienia. Aby uniknąć problemów współliniowości, podjęliśmy decyzję o usunięciu kolumny „wskUrban" z zestawu zmiennych objaśniających.



Regresja liniowa jest wrażliwa na różne skale danych wejściowych, dlatego wszystkie zmienne liczbowe poddaliśmy standaryzacji przy użyciu funkcji StandardScaler z biblioteki scikit-learn. Proces ten przekształcił dane tak, aby każda zmienna miała średnią 0 i odchylenie standardowe równe 1. Dzięki temu zredukowaliśmy wpływ różnic w skalach poszczególnych zmiennych na wyniki modelu.

Po standaryzacji usunęliśmy wspomnianą kolumnę i zweryfikowaliśmy wyniki przekształceń.

	gestoscZaludnienia	ludnoscTys	turystyka	wynagrodzenie	bezrobocie	odpadyTysTon
0	-0.459219	-0.092757	-0.176942	0.503620	-1.043607	-0.463797
1	-0.247407	-0.028522	-0.136266	-0.145999	-0.517253	0.457919
2	-0.258572	-0.110797	-0.339647	0.042823	-0.216478	-0.326477
3	-0.496222	-0.539163	-0.502352	-0.572326	1.638296	-0.183910
4	-0.436251	-0.419108	-0.237956	0.448638	0.735973	-0.074129

3. Budowanie modelu

Zbudowaliśmy model regresji liniowej, którego celem było określenie wpływu zmiennych objaśniających na ilość wytwarzanych odpadów komunalnych (odpadyTysTon). W analizie wykorzystaliśmy dane standaryzowane, a do zmiennych objaśniających należały: gęstość zaludnienia, liczba ludności, turystyka, wynagrodzenie i bezrobocie. Model został zbudowany przy użyciu metody najmniejszych kwadratów (OLS).

Pierwsze wyniki wskazywały na bardzo słabe dopasowanie modelu do danych:

- R-kwadrat wyniósł 0.005, a skorygowane R-kwadrat -0.008, co oznaczało, że model praktycznie nie wyjaśniał zmienności zmiennej zależnej.
- Statystyka F (0.396) oraz jej wartość p (0.852) sugerowały brak istotności statystycznej modelu.
- Współczynniki dla większości zmiennych objaśniających miały bardzo wysokie wartości p (>0.05), co oznaczało brak ich istotności statystycznej.

Słabe wyniki modelu najpewniej wynikają z faktu, że mamy doczynienia z danymi panelowymi, jednak tak prosty model jak regresja liniowa nie uwzględnia tej zależności, niemniej jednak wyniki te mogły też wynikać z obecności wartości odstających w danych. W związku z tym zdecydowaliśmy się na ich usunięcie.

Wartości odstające zidentyfikowaliśmy za pomocą metody IQR (interquartile range). Dla każdej zmiennej obliczyliśmy wartości Q1 i Q3 oraz rozstęp międzykwartylowy (IQR = Q3 - Q1). Usunęliśmy obserwacje spoza zakresu [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR].

Po usunięciu wartości odstających liczba obserwacji zmniejszyła się z 380 do 248. Następnie ponownie wystandaryzowaliśmy dane i zbudowaliśmy nowy model regresji.

Nowy model regresji

Po oczyszczeniu danych i ponownej budowie modelu wyniki nieznacznie się poprawiły:

- R-kwadrat wzrosło do 0.036 co oznacza, że model wyjaśnia jedynie 3.6% zmienności zmiennej zależnej. To niski poziom dopasowania, który sugeruje, że istnieją inne czynniki wpływające na ilość odpadów, niewykorzystane w modelu, a skorygowane R-kwadrat do 0.016, co oznaczało minimalne zwiększenie wyjaśnianej zmienności zmiennej zależnej.
- Statystyka F wyniosła 1.818, ale jej wartość p (0.110) nadal wskazywała na brak istotności modelu jako całości.
- Spośród wszystkich zmiennych objaśniających tylko wynagrodzenie miało istotny statystycznie wpływ (p = 0.008). Wskazuje to, że wzrost wynagrodzenia może wiązać się ze wzrostem ilości odpadów komunalnych.

Interpretacja współczynników w modelu regresji

1) **Wyraz wolny** - Współczynnik wynosi -0.0216. Oznacza to przewidywaną wartość zmiennej zależnej (ilości odpadów komunalnych w tysiącach ton) w sytuacji, gdy wszystkie zmienne objaśniające (gęstość zaludnienia, liczba ludności, turystyka, wynagrodzenie, bezrobocie) mają wartość 0. Jego wartość nie ma dużego znaczenia interpretacyjnego w praktyce, ponieważ takie warunki są nierealistyczne.
2) **Gęstość zaludnienia** - Współczynnik wynosi -0,0415, co oznacza, że wzrost gęstości zaludnienia o 1 jednostkę na km² wiąże się z przeciętnym spadkiem ilości odpadów komunalnych o 0,0415 tys. ton, przy założeniu, że pozostałe zmienne pozostają stałe. Wartość ta nie jest istotna statystycznie (p = 0,636), więc jej wpływ może być przypadkowy.
3) **Liczba ludności** - Współczynnik wynosi 0,0550, co sugeruje, że wzrost liczby ludności o 1 tysiąc osób wiąże się z przeciętnym wzrostem ilości odpadów

komunalnych o 0.055 tys. ton, przy założeniu stałości innych zmiennych. Wartość p = 0.560 wskazuje, że zmienna ta nie jest istotna statystycznie.

4) **Turystyka** - Współczynnik wynosi -0.0636, co oznacza, że wzrost liczby obiektów turystycznych o 1 jedna jednostkę wiąże się z przeciętnym spadkiem ilości odpadów komunalnych o 0.0636 tys. ton, przy założeniu stałości pozostałych zmiennych. Wartość p = 0.408 wskazuje, że zmienna ta również nie jest istotna statystycznie.
5) **Wynagrodzenie** - Współczynnik wynosi -0.1942, co oznacza, że wzrost wynagrodzenia o 1 jednostkę wiąże się z przeciętnym spadkiem ilości odpadów komunalnych o 0.1942 tys. ton, przy założeniu stałości pozostałych zmiennych. Ta zmienna jest istotna statystycznie (p = 0.008), co może sugerować, że w regionach o wyższych wynagrodzeniach odpady są efektywniej zarządzane.
6) **Bezrobocie** - Współczynnik wynosi -0.0129, co oznacza, że wzrost bezrobocia o 1% wiąże się z przeciętnym spadkiem ilości odpadów komunalnych o 0.0129 tys. ton. Jednak wartość p = 0.864 wskazuje, że zmienna ta nie jest istotna statystycznie i jej wpływ może być losowy.

Stabilność modelu

Aby ocenić stabilność parametrów modelu, przeprowadziliśmy test Chow. Dane zostały podzielone na dwie równe części, a następnie oszacowaliśmy modele dla obu podzbiorów oraz całego zbioru danych. Wyniki testu:

- Statystyka F wyniosła 0.489, a wartość p = 0.8163.
- Wartość p większa niż 0.05 wskazuje, że nie ma podstaw do odrzucenia hipotezy zerowej o stabilności modelu. Parametry modelu są więc takie same w obu podzbiorach.

H_o (hipoteza zerowa): Parametry modelu są takie same w obu podzbiorach (Model regresji jest stabilny, a dane mogą być analizowane za pomocą jednego zestawu parametrów.).

H_1 (hipoteza alternatywna): Parametry modelu różnią się między podzbiorami (model nie jest stabilny).

Statystyka testu:

Statystyka F jest obliczana jako:

$$F = \frac{(SSE_{\text{całość}} - (SSE_1 + SSE_2))/k}{(SSE_1 + SSE_2)/(n_1 + n_2 - 2k)}$$

Gdzie:

- $SSE_{\text{całość}}$: Suma kwadratów reszt dla całego modelu.
- SSE_1, SSE_2: Sumy kwadratów reszt dla podzbiorów danych.
- k: Liczba parametrów w modelu (w tym stała).
- n_1, n_2: Liczba obserwacji w podzbiorach danych.

P-value wynosi 0.68 zatem jest brak podstaw do odrzucenia Hipotezy zerowej, przyjmujemy więc że model **jest stabilny**.

4. Podsumowanie

Ostateczny model regresji liniowej opisuje zależność ilości odpadów komunalnych w tysiącach ton od pięciu zmiennych objaśniających: gęstości zaludnienia, liczby ludności, liczby obiektów turystów, wynagrodzenia oraz bezrobocia. W trakcie analizy przeprowadzono oczyszczenie danych z wartości odstających oraz testy stabilności parametrów modelu.

Wyniki wskazują, że jedyną zmienną istotną statystycznie jest wynagrodzenie – im wyższe średnie wynagrodzenie brutto, tym mniejsza ilość odpadów komunalnych, co może sugerować, że w regionach o wyższych dochodach odpady są efektywniej zarządzane. Pozostałe zmienne, takie jak gęstość zaludnienia, liczba ludności, liczba turystów oraz bezrobocie, nie wykazały istotnego statystycznie wpływu na ilość wytwarzanych odpadów, co może wynikać z ich niewielkiego wpływu na zmienną zależną lub potrzeby uwzględnienia dodatkowych czynników w analizie.

Model regresji wykazuje bardzo niski poziom dopasowania (R-kwadrat wynosi 0.036), co wskazuje, że większość zmienności ilości odpadów komunalnych nie została wyjaśniona przez analizowane zmienne.