Analiza zależności pomiędzy otyłością a czynnikami społeczno-ekonomicznymi w różnych krajach

Mroczko Tomasz, 266604

I Wstęp

Wprowadzenie do problemu

Problemem wybranym do badań jest zależność otyłości w różnych krajach od innych czynników. Inspiracją do tego rodzaju rozważań były niepokojące trendy dotyczące rosnącego problemu otyłości na całym świecie. Obserwuje się coraz większą liczbę osób dotkniętych otyłością, zarówno wśród dzieci, młodzieży, jak i dorosłych. Ten dynamiczny wzrost otyłości jest poważnym wyzwaniem dla zdrowia publicznego i wpływa negatywnie na naszą jakość życia.

Cel projektu

Celem tego projektu jest przeprowadzenie analizy zależności między otyłością a różnymi czynnikami w różnych krajach. Skupimy się na zrozumieniu wpływu trzech czynników:

- Podaż kaloryczna na osobę
- PKB per capita
- Wskaźnik urbanizacji kraju

Metodyka badania

Badanie rozpoczęto się od wstępnej analizy danych, w której została przeanalizowana dynamika otyłości świata na przestrzeni lat. Przeanalizowano trend tego zjawiska w czasie. Następnie skupiono się na analizie otyłości w roku 2016 oraz zmiennych niezależnych, takich jak podaż kaloryczna na osobę, PKB per capita oraz wskaźnik urbanizacji. Przedstawiono rozkłady poszczególnych danych, wartości średnie, minimalne oraz maksymalne.

W kolejnym kroku przystąpiono do modelowania, wykorzystując dostępne dane z roku 2016. Stworzono modele regresyjne oraz inne techniki modelowania w celu predykcji poziomu otyłości na podstawie analizowanych zmiennych niezależnych.

II Dane

Kod źródłowy zawierający pozyskanie, oczyszczenie i ujednolicanie danych zawarty jest w pliku *gather_and_clean_data.ipynb*.

Pozyskanie danych

- **Wskaźnik otyłości:** Dane dotyczące wskaźników otyłości w danych krajach na przestrzeni lat pobrane zostały z World Health Organization (WHO). Dane dotyczą wskaźnika otyłości z podziałem na rok, kraj oraz płeć.
- **Podaż kaloryczna na osobę:** Dane dotyczące średniej podaży kalorycznej na osobę w danych krajach na przestrzeni lat zostały zdobyte poprzez scrapowanie strony internetowej Wikipedia za pomocą biblioteki Beautiful Soup. Dane zawierają informacje o średniej podaży kalorii z podziałem na kraje. Po przetworzeniu zostały dane o średniej dziennej podaży kalorycznej 171 krajów.

- **PKB per capita:** Dane dotyczące wskaźnika PKB per capita w danych krajach na przestrzeni lat zostały pobrane ze źródła danych World Bank. Zawierają informacje o PKB per capita różnych krajów na przestrzeni lat.
- **Urbanizacja:** Dane dotyczące wskaźnika urbanizacji w danych krajach również pobrane zostały z World Bank.

Przetwarzanie wstępne

- **Wskaźnik otyłości:** Ze zbioru usunięto kolumny dotyczące wskaźników z podziałem na płeć. Usunięto także zakresy nadwagi (interesuje nas tylko średnia wartość). Tabela została transponowana, a następnie oraz pozbyto się krajów, dla których nie było wartości (były 4 takie kraje). Po przetworzeniu pozostały dane o wskaźnikach otyłości dla 191 krajów, w latach 1975 - 2016.
- **Podaż kaloryczna na osobę:** Ze zbioru usunięto kolumny dotyczące pozycji w rankingu oraz roku pozyskania danych (dane pochodzą z okolic 2016 roku, dla którego będzie analizowana otyłość). Przygotowany zbiór zawiera średnią podaż 171 krajów.
- **PKB per capita:** Tabela została transponowana, a następnie usunięto informacje o roku pozyskania danych. W celu analizy skupiono się na danych z roku 2016 lub wcześniejszego (dla 98% wpisów udało się pozyskać dane właśnie z 2016 roku). Następnie pozbyto się wartości, które nie dotyczyły krajów (kontynenty, obszary). Po przetworzeniu pozostały dane o PKB per capita 211 krajów.
- **Urbanizacja:** Dane dotyczące urbanizacji pozsykane zostały z tego samego źródła co dane o PKB per capita. W związku z tym proces przygotowania informacji był bardzo podobny. Po przetworzeniu pozostały wskaźniki urbanizacji 212 krajów.

Wszystkie nazwy krajów zostały ujednolicone używając dataprep.clean.clean_country

Połączenie danych

Po odpowiednim przygotowaniu danych, połączenie ich polegało na scaleniu wszystkich danych na podstawie kolumny zawierającej nazwę kraju, wykorzystując funkcję pandas.merge. Po przeprowadzeniu scalenia, uzyskaliśmy przygotowaną do analizy ramkę zawierającą następujące dane:

- Nazwę kraju
- Wskaźnik otyłości w %
- Średnią dzieną podaż kaloryczną na osobę w kilokaloriach
- Wskaźnik PKB per capita w $ US
- Wskaźnik urbanizacji w %

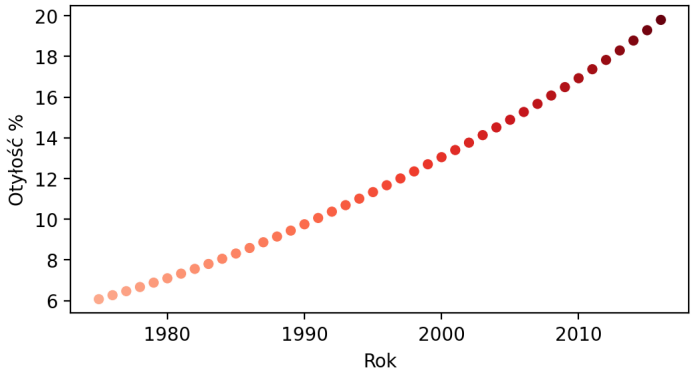
W finalnej ramce danych znajdują się wpisy o 166 krajach. Po zebraniu i przygotowaniu danych, zostały one zserializowane do pliku *.pkl* w celu dalszej analizy. Warto zaznaczyć, że analizowana będzie zależność otyłości od różnych czynników w roku 2016 i tylko te wpisy pozostały w scalonej ramce.

III Wstępna analiza danych

Kod źródłowy w którym przeprowadzona jest wstępna analiza danych zawarty jest w pliku *exploratory_data_analysis.ipynb*. Zaczęto od przeanalizowania otyłości na przestrzeni lat. Następnie skupiono się na najnowszych danych, czyli na roku 2016.

Otyłość na przestrzeni lat

Analizując trend otyłości na przestrzeni lat, można zauważyć, że przez cały badany okres, od 1975 do 2016 roku, średni poziom otyłości systematycznie wzrastał. Wykres ukazuje roczne zmiany w średnim poziomie otyłości w badanych krajach, a widoczne dane jednoznacznie wskazują na kontynuację trendu wzrostowego przez wszystkie lata analizy.



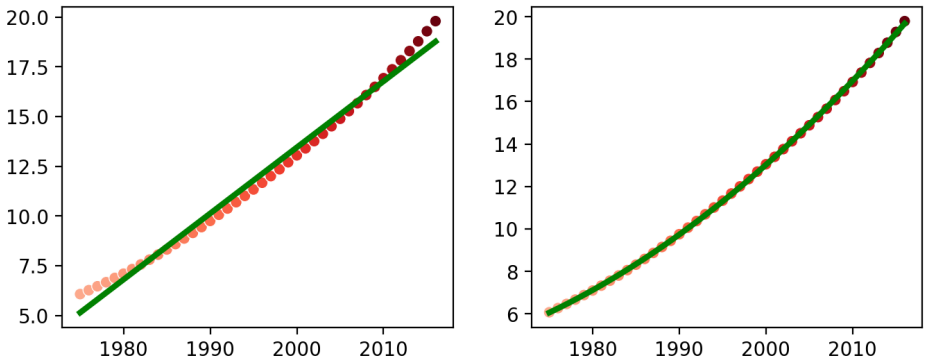
Rysunek 1: Średni % otyłości krajów na przestrzeni lat

Wartość	Rok	Otyłość %
Min	1975	6.074 %
Rok	2016	19.79 %

Tabela 1: Skrajne wartości poziomu otyłości

Przedstawione dane odnoszą się do średniej otyłości na poziomie krajowym, co nie jest równoznaczne z procentowym udziałem otyłej populacji na całym świecie.

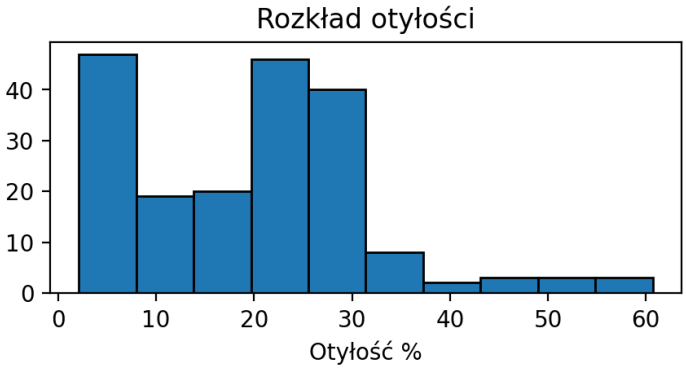
Postanowiono dopasować wielomian do średnich poziomów otyłości w celu zbadania trendu i wzrostu w danych. Wykorzystanie wielomianu pozwala jeszcze bardziej wyszczególnić rosnącą skalę tego zjawiska na przestrzenia lat.



Rysunek 2: Średni % otyłości krajów na przestrzeni lat, po lewej wielomian stopnia pierwszego, po prawej stopnia drugiego

W przypadku modelu liniowego, który zakłada liniową zależność między latami a poziomem otyłości, średni błąd kwadratowy (MSE) wynosi 0.1237. W przypadku modelu kwadratowego, różnica między wartościami rzeczywistymi a predykcjami jest znacznie mniejsza, ze średnim błędem kwadratowym (MSE) wynoszącym 0.0008.

Otyłość w 2016

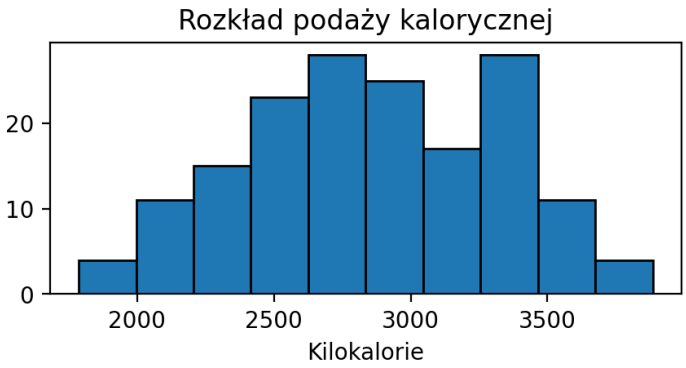


Rysunek 3: Histogram otyłości

Min	Max	Średnia	Unikalne wartości
2.1 %	60.7 %	19.79 %	137

Tabela 2: Statystyki poziomu otyłości

Podaż kaloryczna na osobę

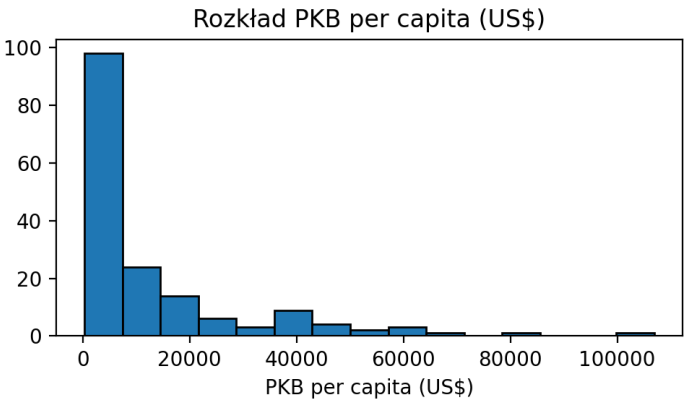


Rysunek 4: Histogram otyłości

Min	Max	Średnia	Unikalne wartości
1786 kcal	3885 kcal	2868.03 kcal	154

Tabela 3: Statystyki podaży kalorycznej

PKB per capita

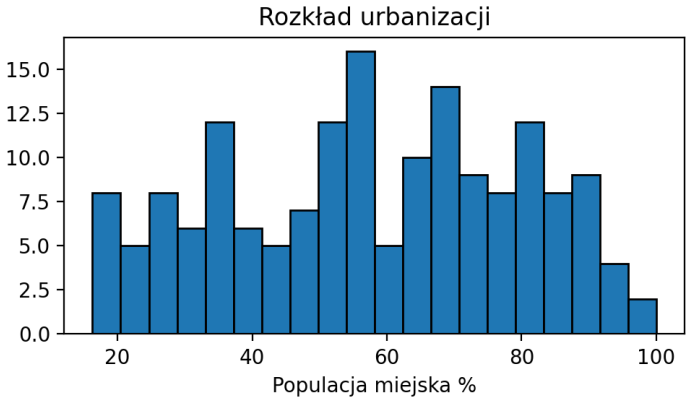


Rysunek 5: Histogram PKB per capita

Min	Max	Średnia	Unikalne wartości
312.143 $	106899.294 $	12703.687 $	166

Tabela 4: Statystyki PKB per capita

Urbanizacja



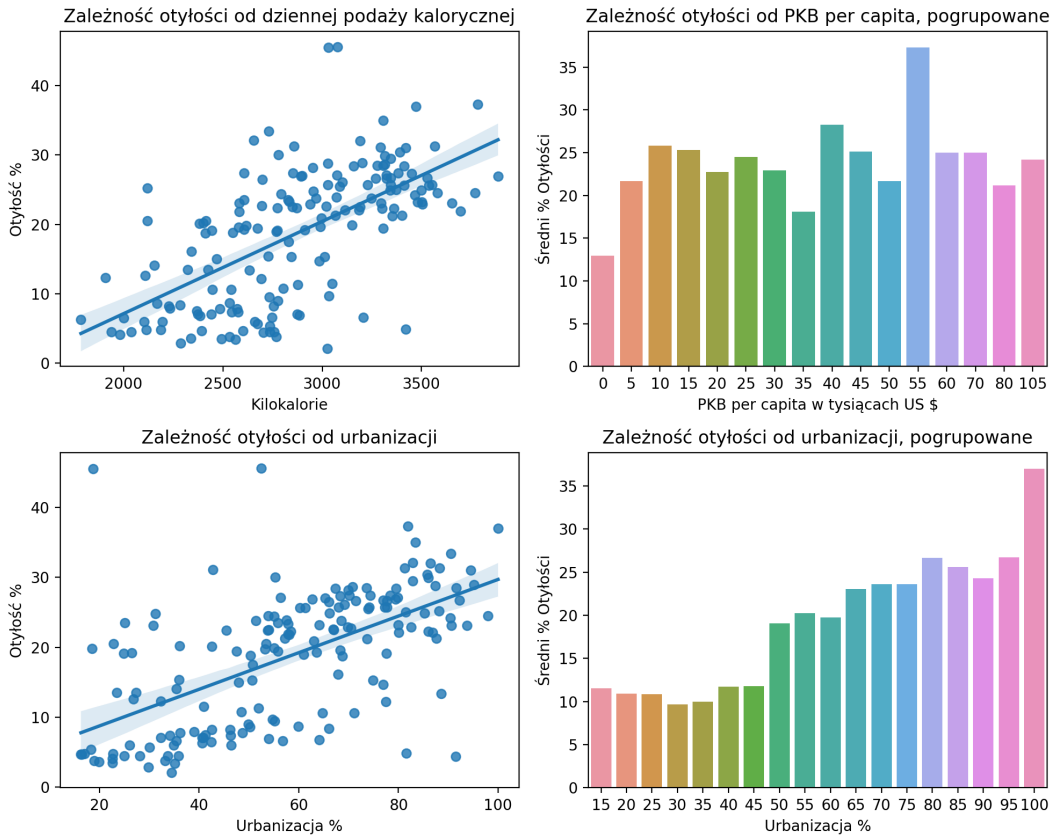
Rysunek 6: Histogram urbanizacji

Min	Max	Srednia	Unikalne wartości
16.29 %	100.0 %	57.898 %	166

Tabela 5: Statystyki urbanizacji

Podstawowe zależności pomiędzy danymi

W ramach analizy w tej sekcji raportu przeprowadzono wizualizację danych za pomocą bibliotek seaborn i matplotlib. Wykorzystano różne typy wykresów, takie jak wykresy słupkowe i wykresy punktowe z linią regresji, aby przedstawić zależności między zmiennymi. Dodatkowo, dla niektorych danych, aby ułatwić interpretację wyników, wykorzystano pogrupowanie danych.

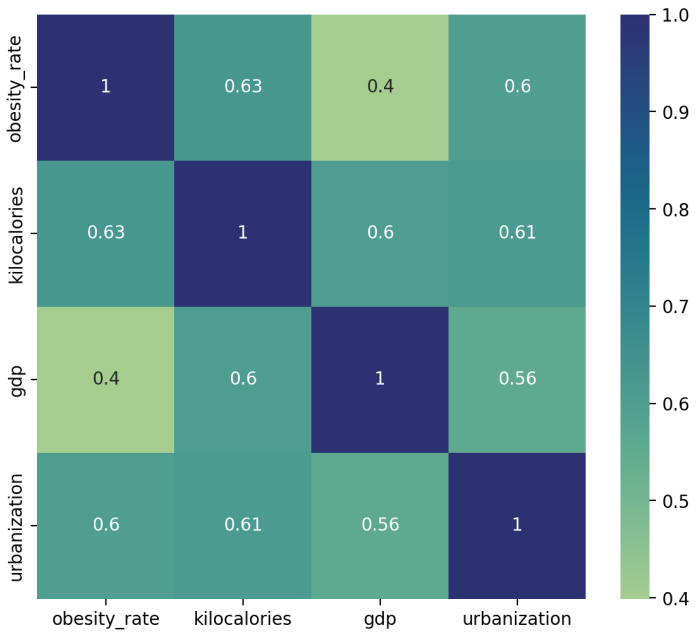


Rysunek 7: Histogram urbanizacji

Pierwszy rzut oka na powyższe wykresy, sugeruje, że zarówno dzienna podaż kalorii, jak i stopień urbanizacji kraju mają wpływ na poziom otyłości w badanych krajach. Na wykresach słupkowych oraz punktowych z linią regresji można zauważyć tendencję wzrostową, co sugeruje, że im większa podaż kaloryczna oraz wyższy poziom urbanizacji, tym wyższy jest odsetek osób otyłych w danym kraju. Natomiast, jeśli chodzi o zależność między PKB per capita a poziomem otyłości, na pierwszy rzut oka nie widać wyraźnej zależności. Wykres słupkowy pogrupowanych danych ze względu na PKB per capita nie sugeruje wyraźnego wzorca, jednak to tylko ogólne wrażenie.

Korelacja danych

Wstępną analizę danych zakończono utworzeniem mapy cieplnej (heatmap) korelacji, aby zbadać wzajemne związki między zmiennymi. Mapa cieplna korelacji umożliwiła nam wizualną identyfikację silnych i słabych zależności między poszczególnymi zmiennymi. Na podstawie mapy cieplnej można było wstępnie ocenić, które zmienne mają największy wpływ na otyłość.



Rysunek 8: Histogram urbanizacji

Analiza mapy cieplnej korelacji wykazała następujące wnioski: Istnieje dość wysoka korelacja między otyłością a dostępnością kalorii (0.63). Kraje o wyższej podaży kalorii mają tendencję do wykazywania wyższych wskaźników otyłości. Zauważalna jest również stosunkowo niska dodatnia korelacja między otyłością a PKB per capita (0.4). Mapa sugeruje, że znaczący wpływ na wskaźnik otyłości ma również poziom urbanizacji kraju (0.6).

IV Modelowanie

Użyte modele

- **Linear Regression** - prosty model, który zakłada, że istnieje liniowa zależność między zmiennymi niezależnymi a zmienną zależną. Model ten stara się znaleźć prostą, która najlepiej pasuje do danych.
- **Generalized Linear Model** - model podobny do regresji liniowej ale używa wielomianu danego stopnia, zamiast prostej. To oznacza, że GLM jest w stanie modelować nieliniowe zależności w danych, które nie są idealnie dopasowane do linii prostej.
- **Support Vector Regression** - model regresji, który wykorzystuje rzutowanie danych na przestrzeń o wyższej wymiarowości w celu dopasowania nieliniowych zależności. SVR znajduje optymalną krzywą lub powierzchnię dopasowania, koncentrując się na tzw. wektorach nośnych.

Metody ewaluacji modeli

Aby ocenić jakość otrzymanych modeli regresyjnych, zastosowano różne metryki. Oto trzy wykorzystane miary:

- **Mean Absolute Error (MAE)** - średni błąd bezwzględny między przewidywanymi wartościami a rzeczywistymi danymi. MAE mierzy średnią różnicę, wyrażoną w jednostkach war-

tości przewidywanej, dla każdej próbki. Oznacza to, że MAE ocenia średnią wielkość błędu przewidywanej wartości dla każdego punktu danych.

- **Mean Squared Error (MSE)** - średni kwadrat błędu między prognozowanymi wartościami a rzeczywistymi danymi. Mierzy średnią kwadratową różnicę między wartościami przewidywanymi przez model a wartościami rzeczywistymi.
- **R-squared (R^2)** - wskaźnik określający, jak dobrze model regresyjny dopasowuje się do danych w porównaniu do prostego modelu, który zawsze przewiduje średnią wartość. Im bliżej wartości 1, tym lepsze dopasowanie modelu.

Podział danych

Do podziału danych na dane treningowe i testowe użyto *train_test_split*. Dane treningowe stanowią 80% zbioru. Dla powtarzalności eksperymentów użyto stałego ziarna losowości o wartości 0.

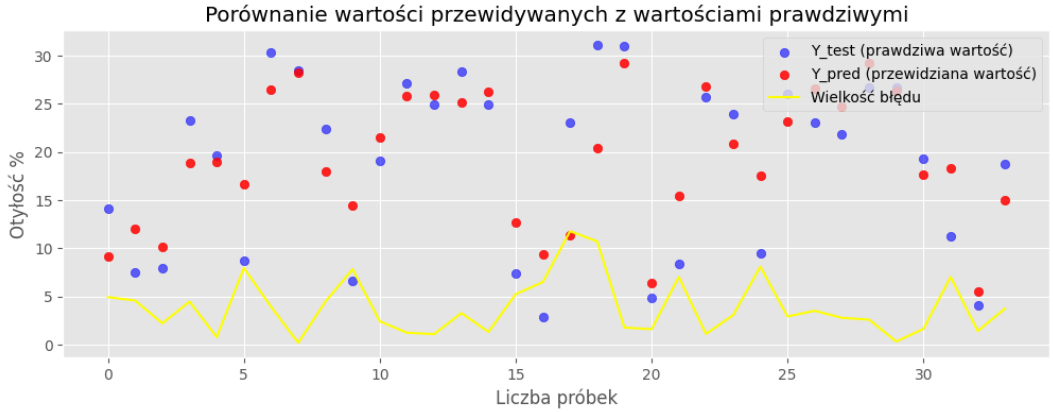
Linear Regression

W pierwszym etapie modelowania przetestowano regresję liniową, która po przeprowadzeniu testów na dostępnych danych osiągnęła następujące wyniki.

MAE	MSE	R2
3.927	23.822	0.696

Tabela 6: Wyniki regresji liniowej

Poniżej przedstawiono wykres porównujący przewidziane wartości przez model z rzeczywistymi danymi, co pozwala na wizualną ocenę jakości dopasowania modelu do danych.



Rysunek 9: Wykres przewidywań regresji liniowej

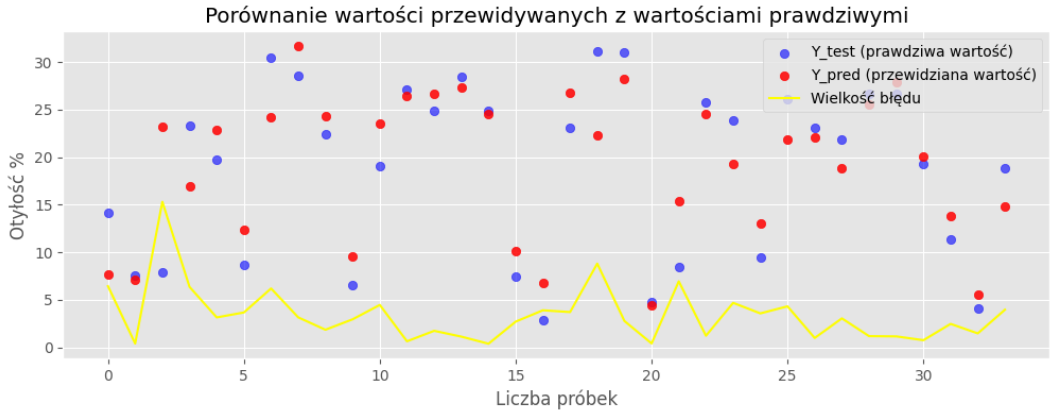
Generalized Linear Model

Następnie zastosowano Generalized Linear Model (GLM). Po przetestowaniu różnych stopni wielomianu w modelu, zaobserwowano, że najlepsze przybliżenie danych uzyskano przy użyciu wielomianu stopnia trzeciego.

MAE	MSE	R2
3.405	20.176	0.742

Tabela 7: Wyniki GLM stopnia 3

Jak można zauważyć, model o stopniu 3 osiągnął lepsze przybliżenie danych niż model liniowy. Dla danych testowych, był w stanie oszacować procent osób otyłych z danego kraju myląc się średnio o 3.4 punkty procentowe.



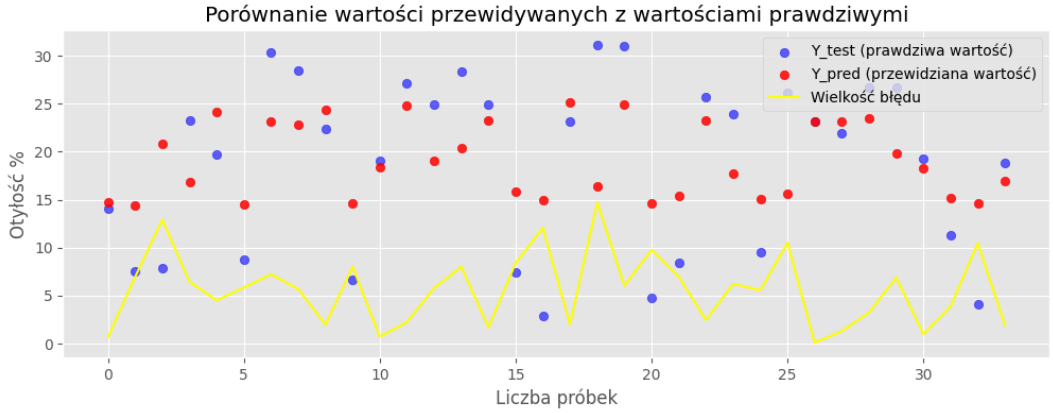
Rysunek 10: Wykres przewidywań GLM stopnia 3

Support Vector Regression

W kolejnym etapie modelowania przetestowano model Support Vector Regression (SVR). Model korzystał z jądra *rbf*. Osiągnął on gorsze wyniki niż poprzednie modele.

MAE	MSE	R2
5.648	45.935	0.413

Tabela 8: Wyniki SVR

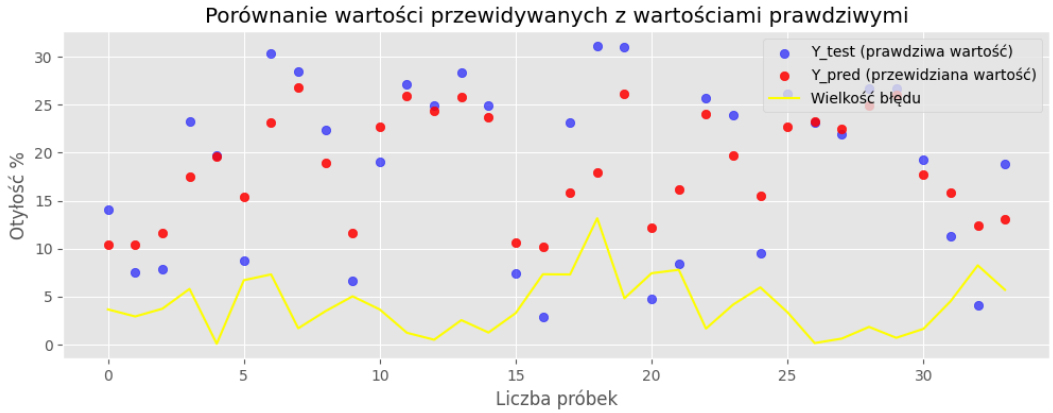


Rysunek 11: Wykres przewidywań SVR

Po uzyskaniu niezadowalających wyników za pomocą modelu Support Vector Regression (SVR), podjęto decyzję o przeskalowaniu danych i ponownym przetestowaniu modelu. Dane wejściowe przeskalowane zostały za pomocą *StandardScaler*, a eksperyment powtórzono. Po przeskalowaniu model znacząco zyskał na jakości.

MAE	MSE	R2
4.09	25.144	0.679

Tabela 9: Wyniki SVR po przeskalowaniu

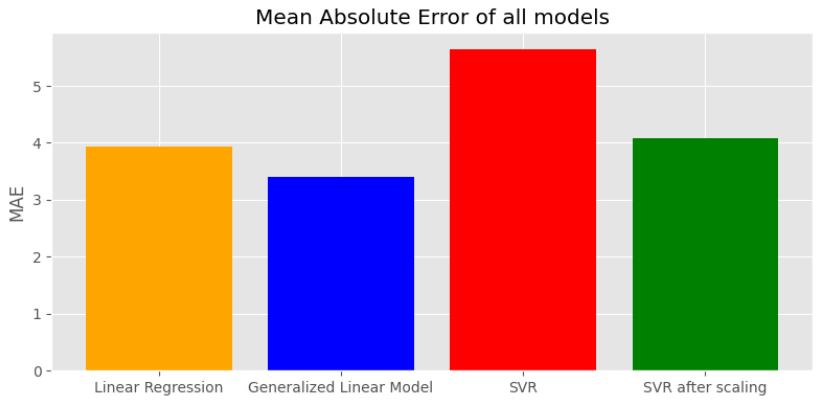


Rysunek 12: Wykres przewidywań SVR po przeskalowaniu

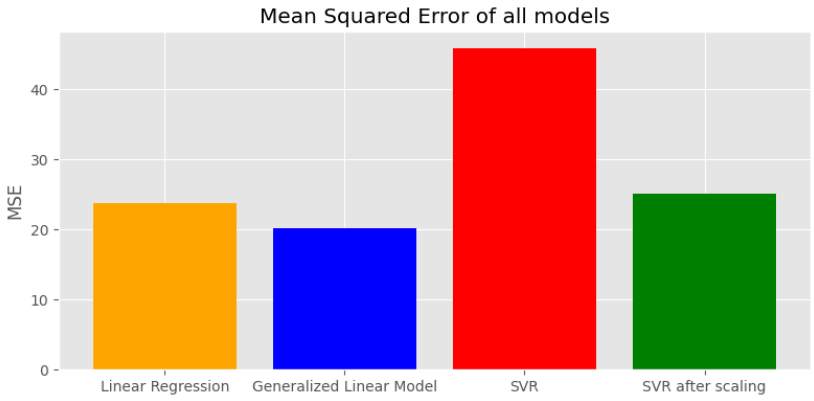
Jak widać, po przeskalowaniu danych wejściowych model lepiej przewidywał zmienną zależną.

V Wyniki

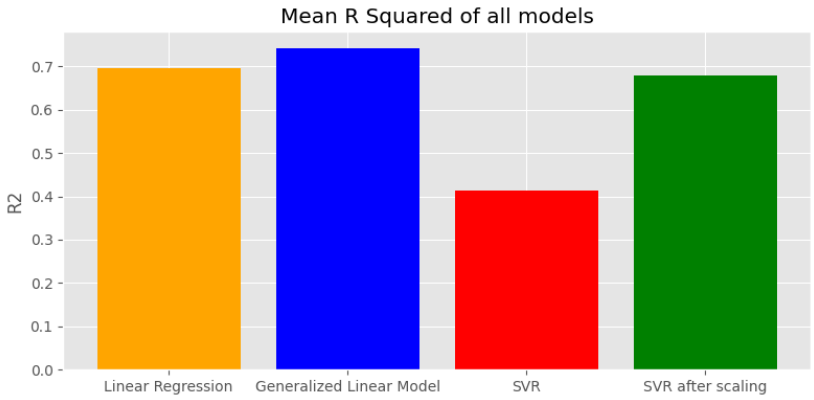
Po przetestowaniu poszczególnych modeli wyniki zostały zestawione i zwizualizowane.



Rysunek 13: Wykres porównawczy MAE



Rysunek 14: Wykres porównawczy MSE



Rysunek 15: Wykres porównawczy R2

VI Wnioski

W wyniku przeprowadzonej analizy oraz modelowania danych dotyczących zależności między otyłością a trzema czynnikami - podażą kaloryczną na osobę, PKB per capita i wskaźnikiem urbanizacji - udało się wyciągnąć kilka istotnych wniosków.

Po pierwsze, zauważono znaczący wpływ podaży kalorycznej na osobę oraz wskaźnika urbanizacji na poziom otyłości. Analiza danych wykazała, że kraje o wyższej podaży kalorycznej oraz większym stopniu urbanizacji częściej wykazywały wyższy poziom otyłości. Jest to zgodne z przypuszczeniami i potwierdza rolę diety i stylu życia związanych z urbanizacją w kształtowaniu epidemii otyłości.

Wpływ PKB per capita na poziom otyłości okazał się niewielki. Oznacza to, że czynnik ten nie jest głównym determinantem otyłości w badanych krajach. Choć wyższe PKB per capita może wpływać na dostęp do żywności i styl życia, inne czynniki wydają się mieć większe znaczenie dla występowania otyłości.

W kontekście modelowania danych, najlepszym modelem predykcyjnym okazał się uogólniony model liniowy stopnia 3. Natomiast model Support Vector Regression (SVR) nie poradził sobie dobrze z danymi na początku, ale po przeskalowaniu danych osiągnął znacznie lepsze wyniki. Wskazuje to na znaczenie skalowania danych w przypadku tego modelu.

W ramach analizy udało się znaleźć pewną zależność otyłości od badanych czynników. Jednak jakość modeli różniła się znacząco w zależności od podziału zbioru. Z powodu małej ilości danych, wybór

punktów miał istotne znaczenie, i mógł zaburzyć dokładność modelu.