

Projekt Analizy Danych: Globalne Wzorce Szczęścia i Jakości Życia

Prezentacja wyników analizy danych

Łukasz Syguła

1. Cel Projektu

Celem projektu jest analiza danych z World Happiness Report 2023, aby zrozumieć, jakie czynniki wpływają na globalne różnice w poziomie szczęścia i jakości życia. Analiza ta pozwoli na identyfikację kluczowych wskaźników, takich jak PKB na mieszkańca, wsparcie społeczne i postrzeganie korupcji, oraz ich wpływu na poziom szczęścia w różnych krajach.



2. Cele Szczegółowe

Identyfikacja Wzorów Globalnych:

- Rozpoznanie wzorców i relacji w danych dotyczących szczęścia i jakości życia.

Segmentacja Krajów:

- Klasteryzacja krajów na podstawie cech takich jak PKB na mieszkańca, wsparcie społeczne, długość zdrowego życia, wolność wyboru, hojność i postrzeganie korupcji.

Redukcja Wymiarów:

- Użycie technik redukcji wymiarów (np. PCA) do uproszczenia danych i wizualizacji wyników klasteryzacji.

Analiza Porównawcza:

- Porównanie wyników klasteryzacji uzyskanych z różnych metod (np. KMeans, DBSCAN) i ocena ich spójności.

Wizualizacja i Interpretacja:

- Prezentacja wyników w formie wizualizacji, takich jak wykresy i mapy, w celu ułatwienia interpretacji i komunikacji wyników.

Rekomendacje:

- Formułowanie rekomendacji politycznych i ekonomicznych na podstawie analizy danych.

3. Oczekiwane Wyniki

- Grupowanie krajów o podobnych poziomach szczęścia.
- Jasne wizualizacje klastrów i wyników redukcji wymiarów.
- Wgląd w wpływ wskaźników na szczęście.
- Rekomendacje polityczne i strategiczne.



4. Proces Analizy

Wczytanie i Przygotowanie Danych

```
In [112]: 1 import numpy as np
          2 import pandas as pd
          3
          4 import matplotlib.pyplot as plt
          5 import seaborn as sns
          6 import plotly.express as px
          7 import plotly.graph_objects as go
          8
          9 from sklearn.preprocessing import StandardScaler
         10 from sklearn.decomposition import PCA
         11 from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN
         12 from scipy.cluster.hierarchy import dendrogram, linkage
         13 from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score
```

```
In [2]: 1 df = pd.read_csv('Desktop/WHR2023.csv')
```

```
In [3]: 1 df.head()
```

Out[3]:

	Country name	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perception of corruption
0	Finland	7.804	0.036	7.875	7.733	10.792	0.969	71.150	0.961	-0.019	0.1
1	Denmark	7.586	0.041	7.667	7.506	10.962	0.954	71.250	0.934	0.134	0.1
2	Iceland	7.530	0.049	7.625	7.434	10.896	0.983	72.050	0.936	0.211	0.6

Usuwanie Niepotrzebnych Kolumn

Removing unnecessary columns

```
In [4]: 1 columns_to_drop = ['upperwhisker', 'lowerwhisker', 'Ladder score in Dystopia',  
2                        'Explained by: Log GDP per capita', 'Explained by: Social support',  
3                        'Explained by: Healthy life expectancy', 'Explained by: Freedom to m  
4                        'Explained by: Generosity', 'Explained by: Perceptions of corruption  
5                        'Dystopia + residual']  
6  
7 df = df.drop(columns=columns_to_drop)
```

Uzupełnianie Brakujących Wartości

Fill missing values with the median

```
In [6]: 1 df['Healthy life expectancy'].fillna(df['Healthy life expectancy'].mean(), inplace = True)
```

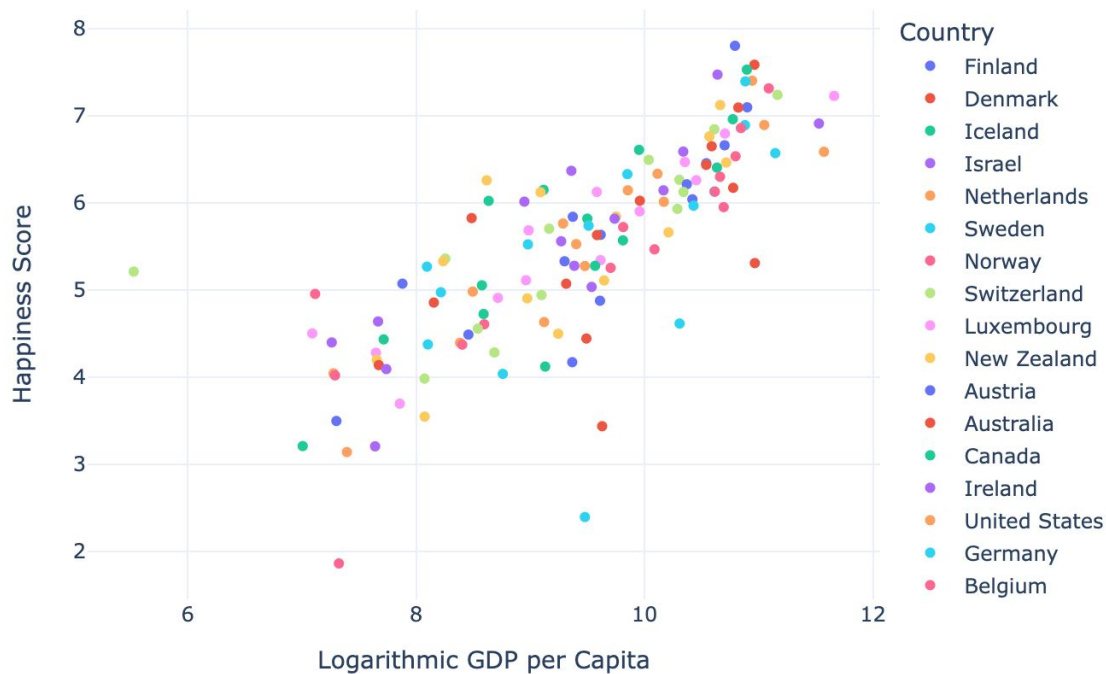
```
In [7]: 1 df.head()
```

Out[7]:

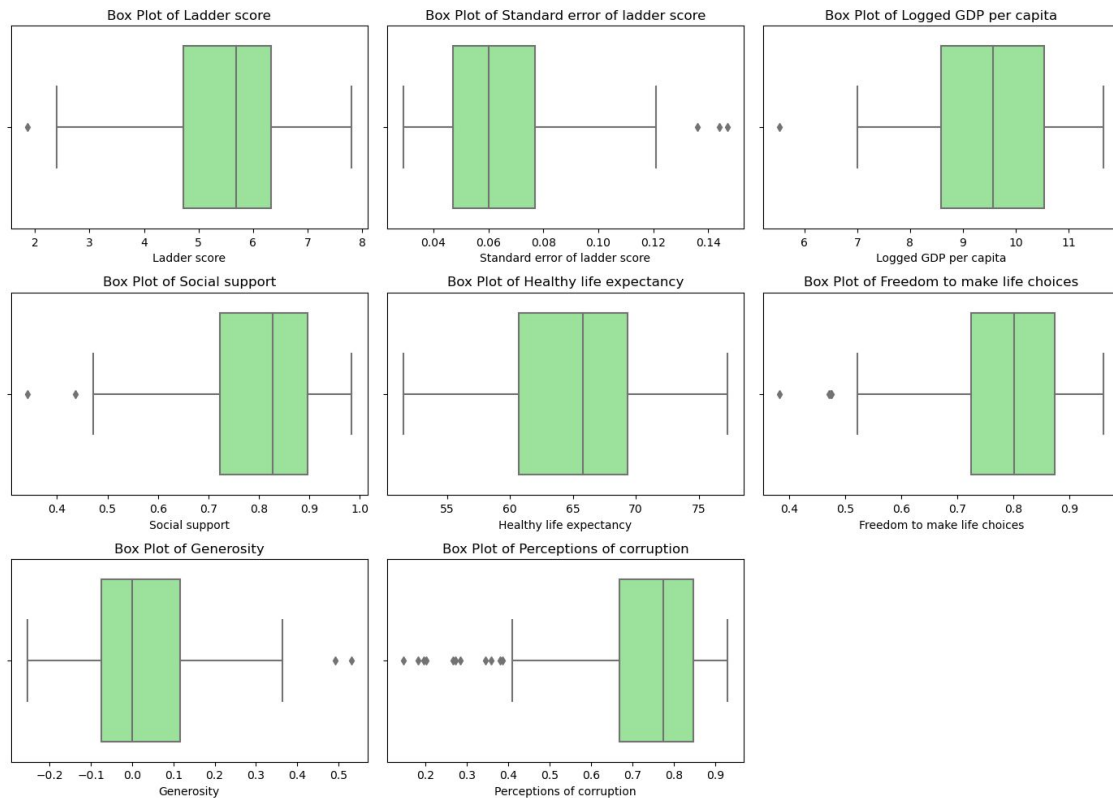
	Country name	Ladder score	Standard error of ladder score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	Finland	7.804	0.036	10.792	0.969	71.150	0.961	-0.019	0.182
1	Denmark	7.586	0.041	10.962	0.954	71.250	0.934	0.134	0.196
2	Iceland	7.530	0.049	10.896	0.983	72.050	0.936	0.211	0.668
3	Israel	7.473	0.032	10.639	0.943	72.697	0.809	-0.023	0.708
4	Netherlands	7.403	0.029	10.942	0.930	71.550	0.887	0.213	0.379

Wykres ilustruje zależność między szczęściem a PKB na mieszkańca w różnych krajach.

Happiness Score vs. Logarithmic GDP per Capita



Wykresy pudełkowe przedstawiają rozkład i wartości odstające dla różnych wskaźników szczęścia w krajach.



Adnotacje odnośnie wartości odstających:

Wynik w Rankingu Szczęścia (Ladder Score): Afganistan ma bardzo niski wynik szczęścia. Potwierdziłem, że ten niski poziom jest zgodny z trudnymi warunkami życia w tym kraju.

Błąd Standardowy Wyniku Szczęścia: Mauretania, Liberia i Botswana mają wysokie błędy standardowe. Analiza wykazała, że są one wynikiem małej liczby próbek lub dużej zmienności, co może wpływać na dokładność wyników.

Zalogowane PKB na Osobę (Logged GDP per Capita): Wenezuela ma wysoką wartość PKB. Potwierdziłem, że ta wartość jest dokładna i zgodna z rzeczywistymi danymi ekonomicznymi.

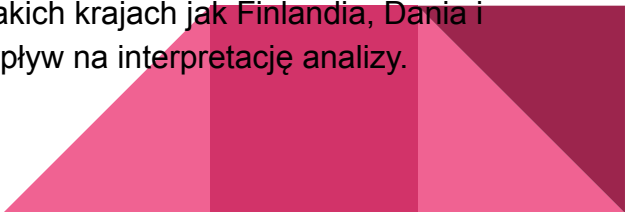
Wsparcie Społeczne (Social Support): Benin i Afganistan mają niskie wartości wsparcia społecznego. Potwierdziłem, że te wartości odzwierciedlają rzeczywistość i mogą wpływać na inne zmienne w analizie.

Oczekiwana Długość Życia w Zdrowiu (Healthy Life Expectancy): Nie wykryto wartości odstających. Ta kolumna nie wymaga dalszej analizy.

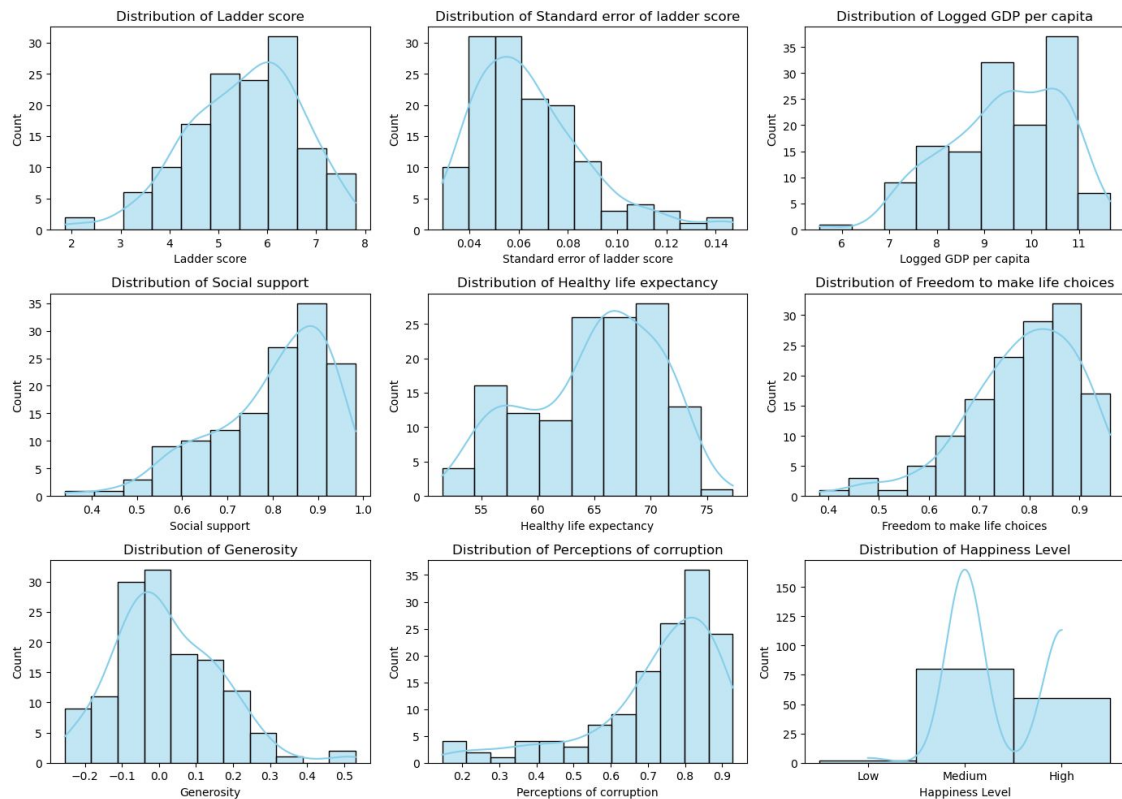
Wolność w Podejmowaniu Decyzji Życiowych (Freedom to Make Life Choices): Turcja, Komory, Liban i Afganistan mają niskie wartości. Potwierdzenie danych wskazuje na niską wolność wyboru w tych krajach, co może wpływać na wyniki.

Hojność (Generosity): Indonezja i Myanmar mają wysokie wartości hojności. Potwierdziłem, że te wartości są zgodne z danymi i powinny być uwzględnione w analizie.

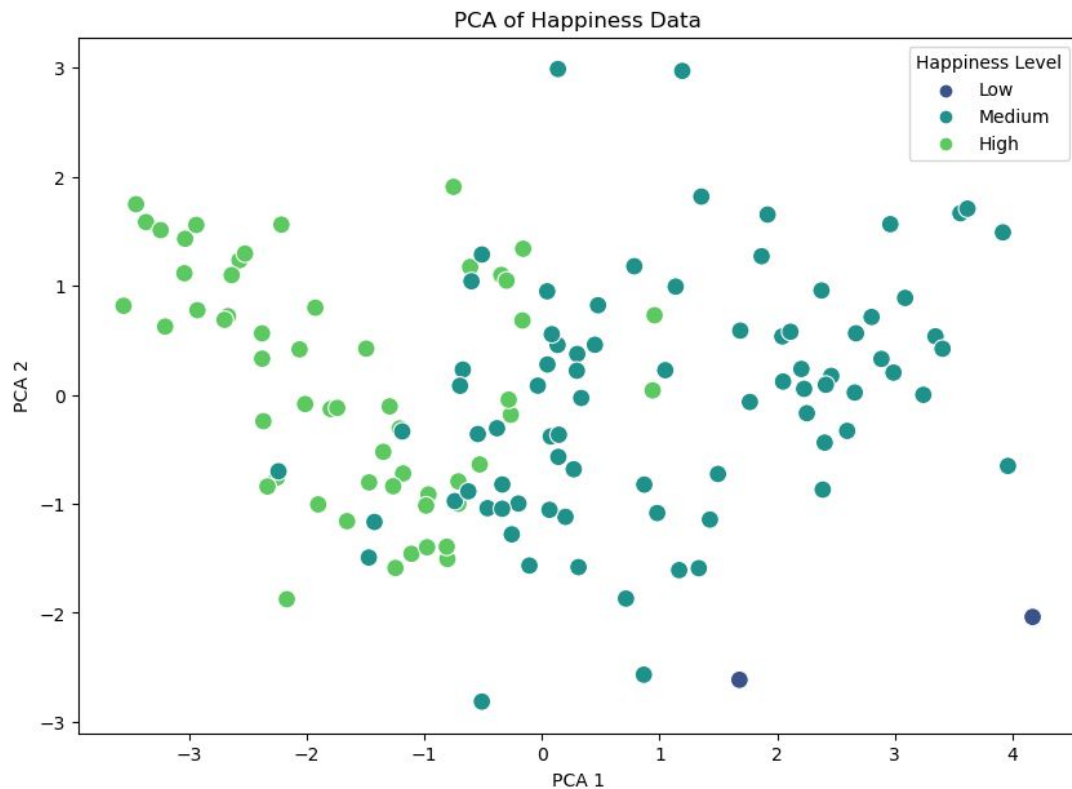
Postrzeganie Korupcji (Perceptions of Corruption): Wysokie wartości występują w takich krajach jak Finlandia, Dania i Holandia. Potwierdzenie danych wskazuje, że te wysokie wartości są dokładne i mają wpływ na interpretację analizy.



Histogramy pokazują rozkład wartości dla różnych wskaźników szczęścia, z dodatkową krzywą KDE ilustrującą gęstość rozkładu.

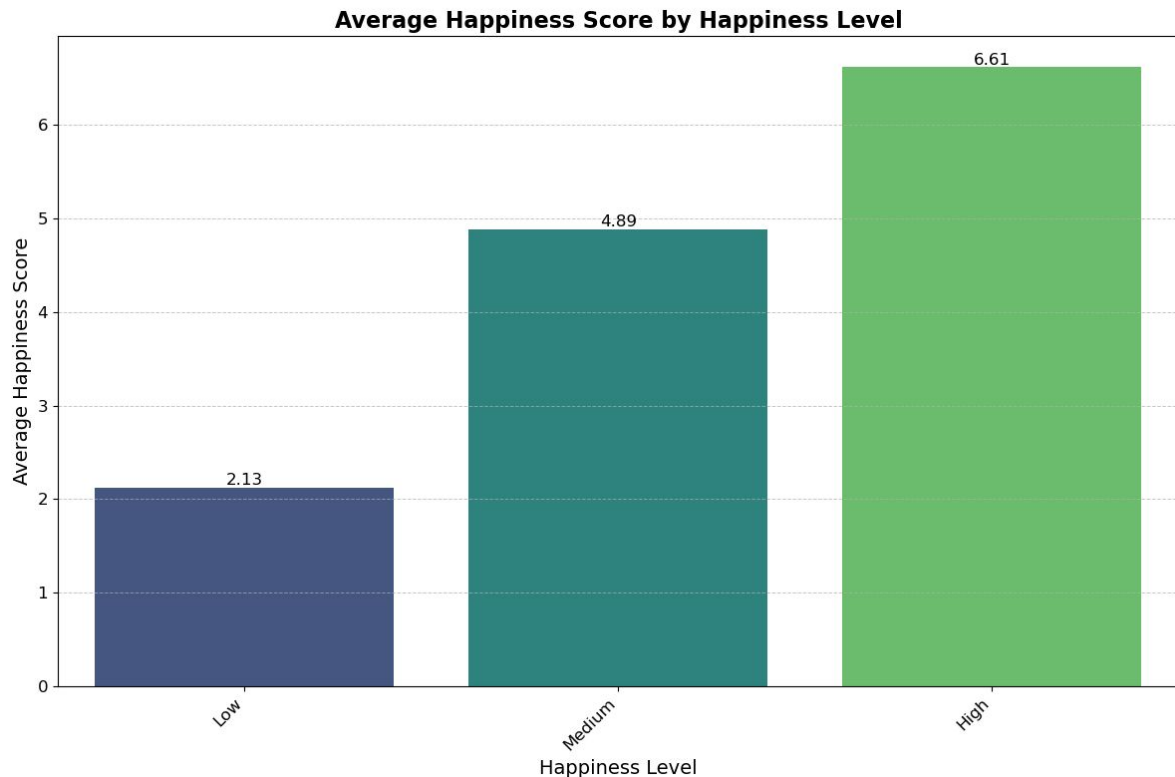


Redukcja Wymiarów za pomocą PCA



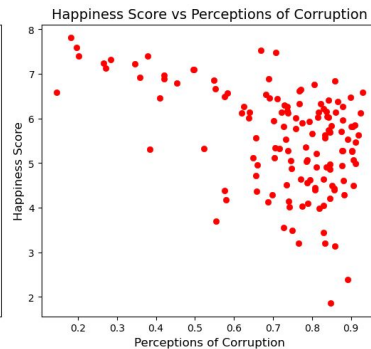
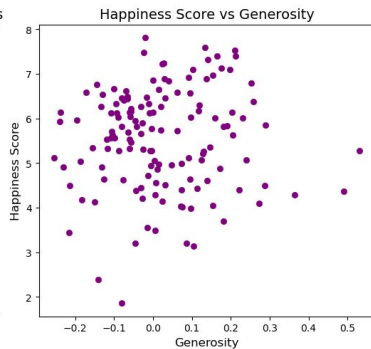
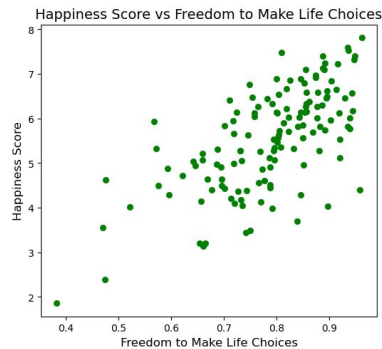
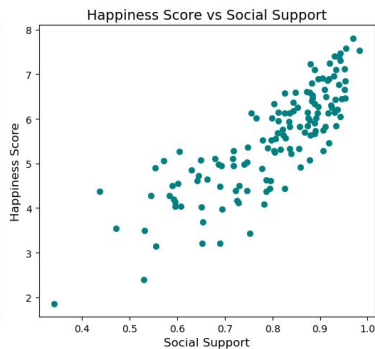
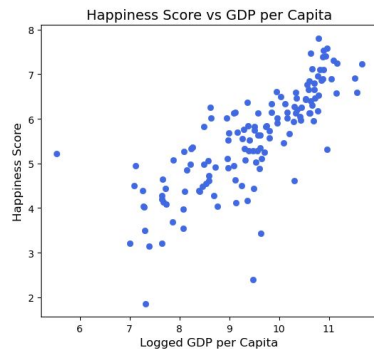
Wykres przedstawia dwuwymiarową reprezentację danych o szczęściu, uzyskaną za pomocą analizy składowych głównych (PCA). Oś X (PCA 1) i oś Y (PCA 2) ilustrują główne kierunki zmienności w danych. Kolorowanie punktów według poziomu szczęścia (Happiness Level) pozwala na łatwe rozróżnienie krajów o różnych poziomach szczęścia.

Wykres słupkowy ilustruje średni poziom szczęścia w zależności od poziomu szczęścia.



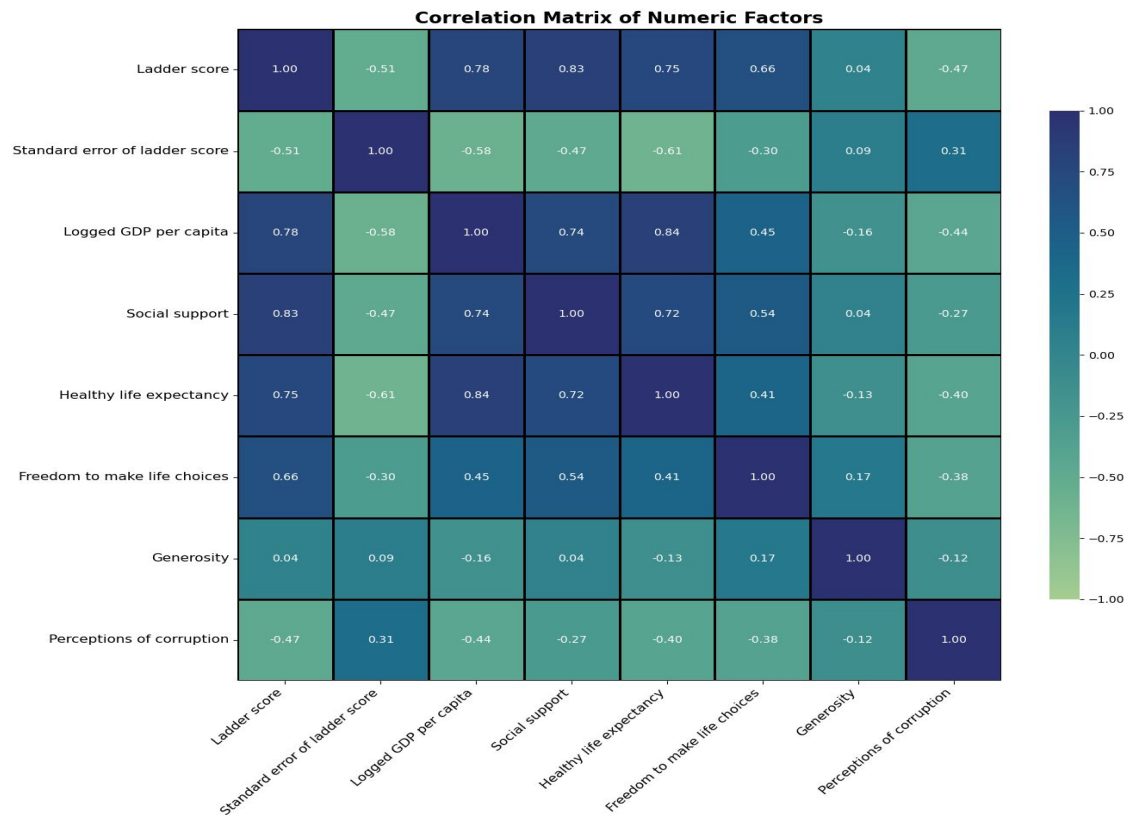
Każdy słupek reprezentuje średnią ocenę szczęścia dla danego poziomu, z wartościami wyraźnie oznaczonymi nad słupkami.

Zależność poziomu szczęścia od kluczowych wskaźników



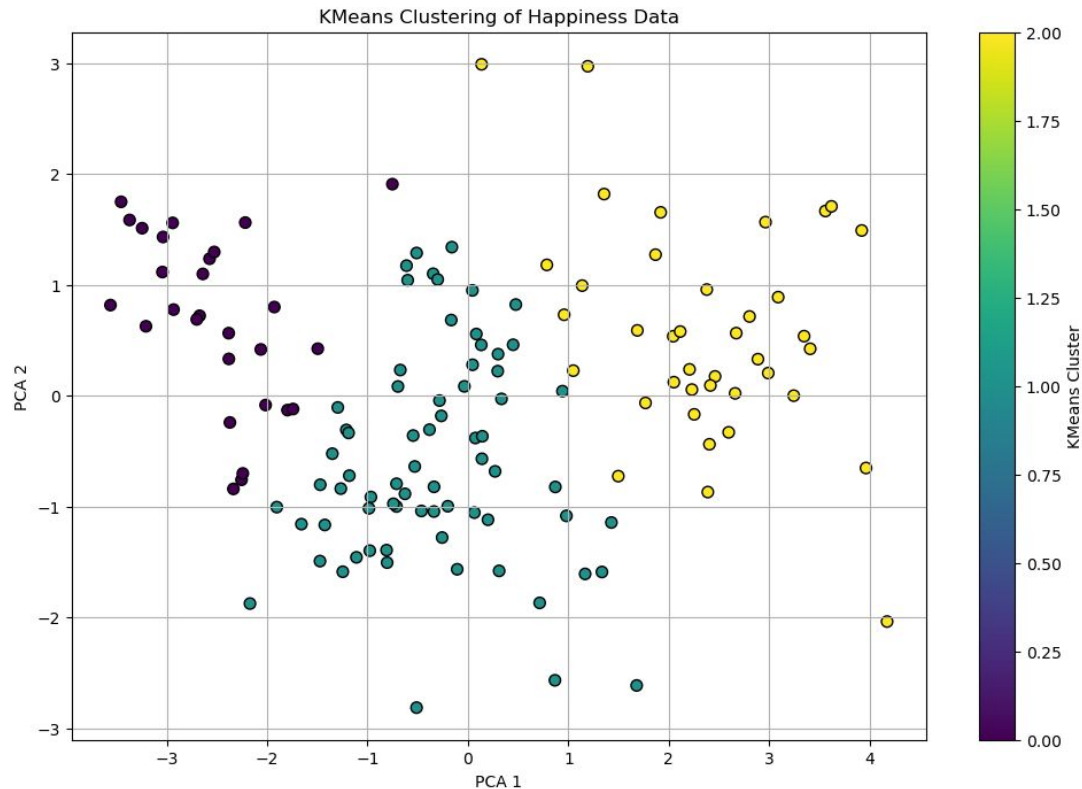
Wykresy ilustrują, jak różne czynniki, takie jak PKB, wsparcie społeczne i długość życia, wpływają na poziom szczęścia. Każdy wykres pokazuje zależność między poziomem szczęścia a jednym z wybranych wskaźników.

Macierz Korelacji Czynników Numerycznych



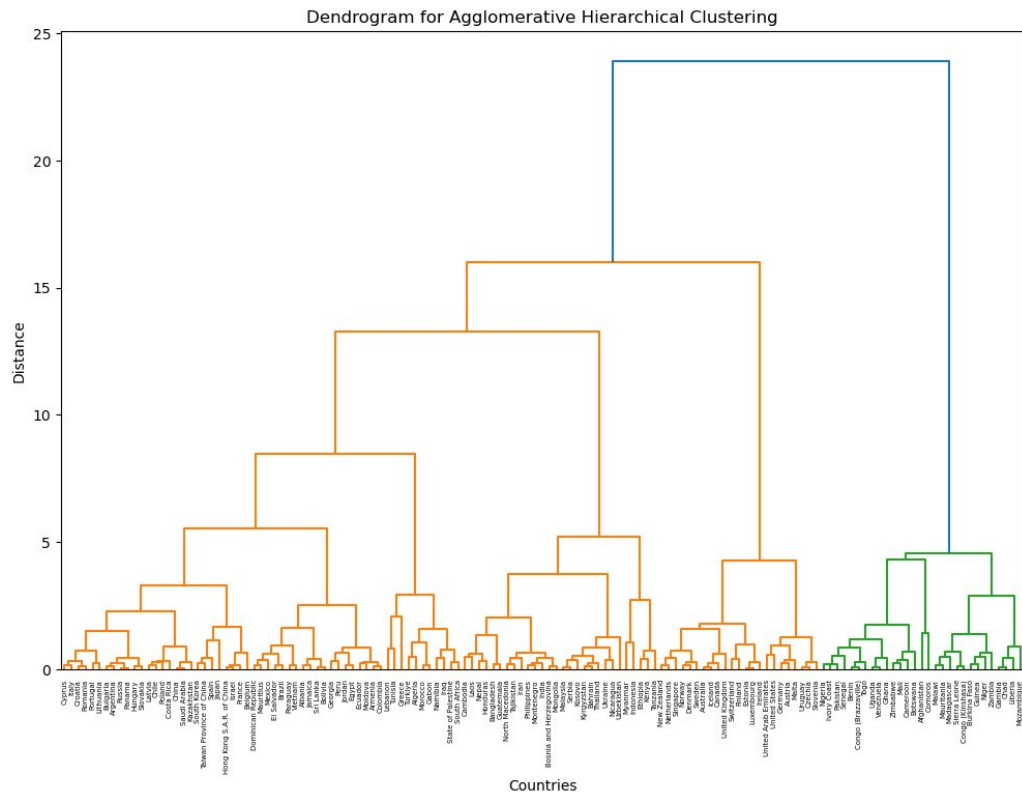
Wykres ilustruje korelacje między zmiennymi numerycznymi, pokazując zależności w zakresie od -1 do 1.

Klasteryzacja KMeans



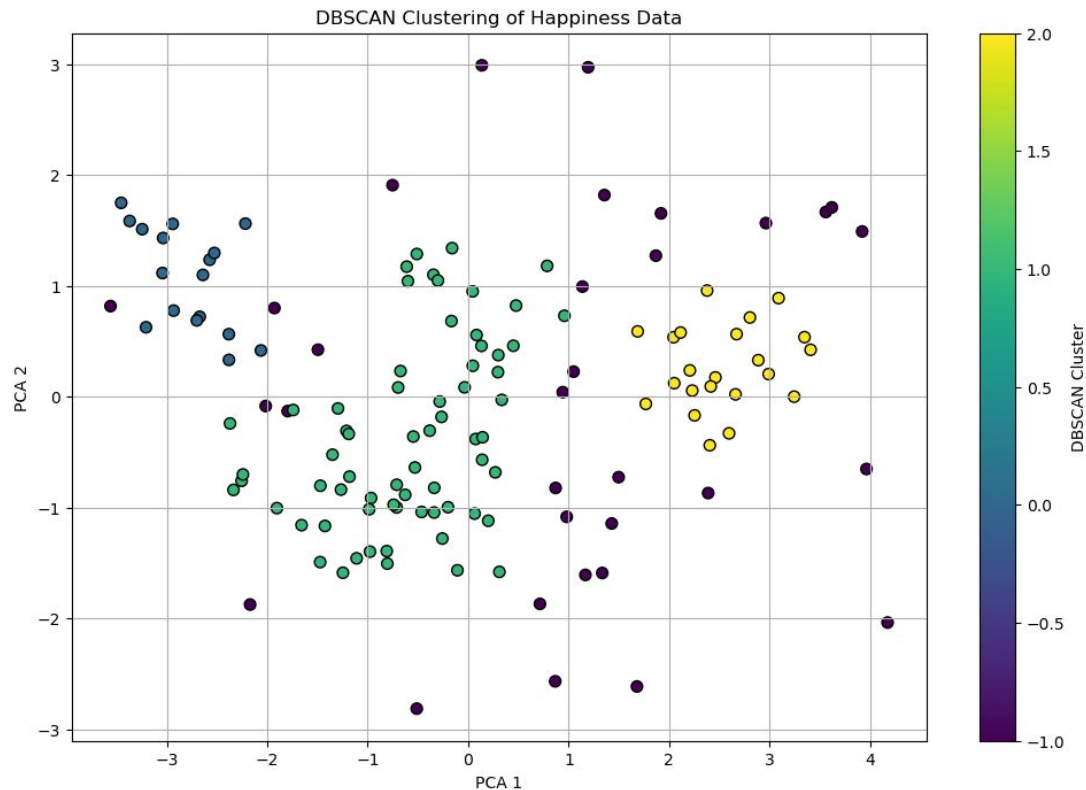
Wykres ilustruje wyniki klasteryzacji KMeans na podstawie analizy składowych głównych (PCA) danych o szczęściu. Punkty są kolorowane zgodnie z przypisanymi klastrami, co pozwala zobaczyć, jak dane są grupowane w przestrzeni dwóch głównych komponentów. Kolorowanie według klastrów (KMeans Cluster) ułatwia identyfikację struktur i wzorców w danych.

Klasteryzacja AHC



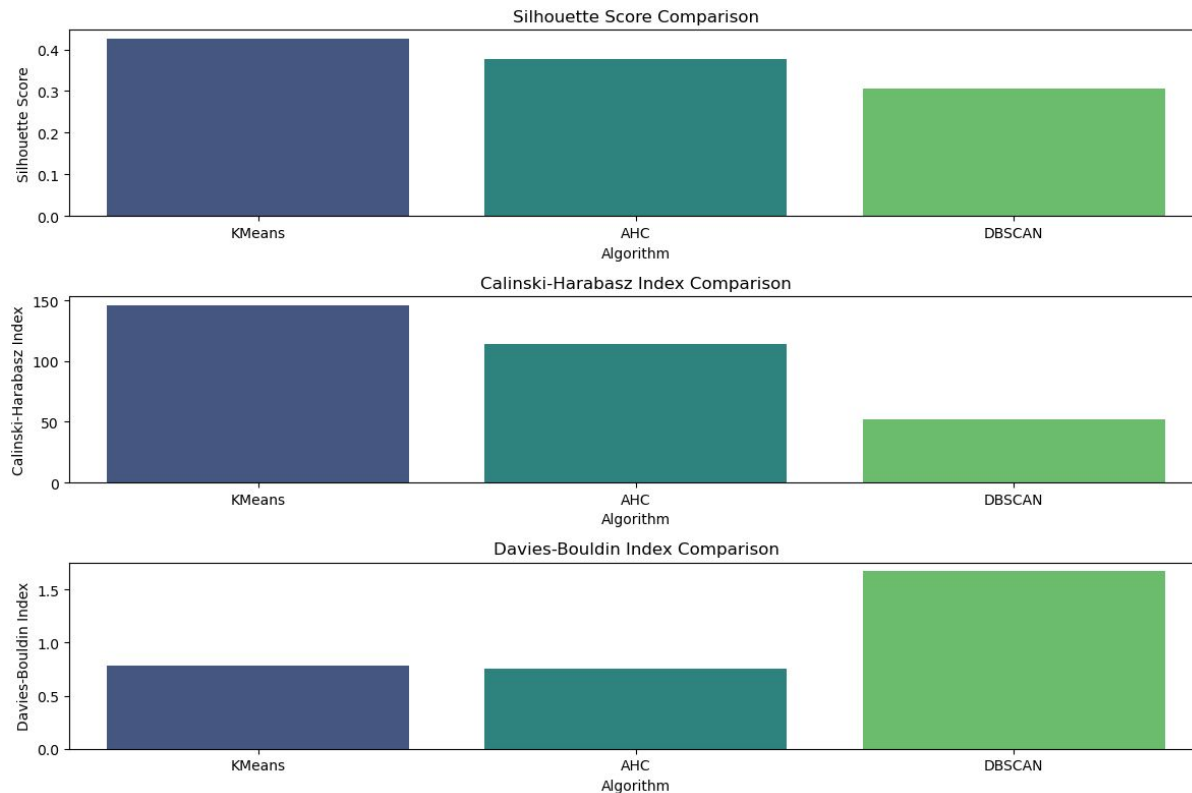
Dendrogram przedstawia hierarchiczną strukturę klasteryzacji aglomeracyjnej na podstawie odległości między krajami. Wykres ilustruje, jak poszczególne kraje łączą się w grupy w miarę wzrostu odległości, co pozwala na ocenę liczby i struktury klastrów w analizowanych danych. Wysokość poziomych linii wskazuje odległość między połączonymi klastrami.

Klasteryzacja DBSCAN



Wykres rozrzutu przedstawia wyniki klasteryzacji DBSCAN w przestrzeni wymiarów PCA. Punkty są kolorowane według przypisanych klastrow DBSCAN, co ukazuje grupy o różnej gęstości i ich rozmieszczenie w analizowanych danych. Kolor na pasku obok wykresu wskazuje przynależność do poszczególnych klastrow.

Porównanie Algorytmów Klasteryzacji



Wykresy porównują trzy algorytmy klasteryzacji: **KMeans**, **AHC** i **DBSCAN**, używając trzech metryk:

- **Silhouette Score:** Wyższe wartości oznaczają lepszą separację klastrów.
- **Calinski-Harabasz Index:** Wyższe wartości wskazują lepszą jakość klastrów.
- **Davies-Bouldin Index:** Niższe wartości oznaczają lepszą jakość klasteryzacji.

Wyniki przedstawione są w postaci wykresów słupkowych dla każdej z metryk.

Oto podsumowanie wniosków dla każdego z algorytmów klasteryzacji:

- **KMeans** okazał się najbardziej efektywnym algorytmem dla tego zbioru danych, zapewniając wyraźne rozdzielenie i dobrze zdefiniowane klastry. Sprawdza się doskonale w sytuacjach, gdy klastry mają być zwarte i wyraźnie oddzielone.
- **AHC** dostarcza satysfakcjonujących wyników, ale nie przewyższa KMeans w żadnym z ocenianych wskaźników. Może być preferowany, gdy potrzebna jest hierarchiczna struktura lub elastyczny pomiar odległości w klasteryzacji.
- **DBSCAN** jest skuteczny w przypadku zbiorów danych z szumem i nieregularnymi kształtami klastrów. W sytuacjach, gdy klastry są dobrze oddzielone i zwarte, DBSCAN może nie wypaść tak dobrze jak KMeans lub AHC.



Wnioski

- **Klaster KMeans:** Analiza wykazała trzy wyraźne klastry. Klaster 1 składa się z krajów o wysokim PKB na mieszkańca i wysokim wsparciu społecznym, Klaster 2 z krajów o średnim poziomie tych wskaźników, a Klaster 3 z krajów o niskich wskaźnikach.
- **Klaster AHC:** Klasteryzacja aglomeracyjna dostarczyła podobnych wyników, ale z bardziej płynnie definiowanymi granicami między klastrami.
- **Klaster DBSCAN:** Wykrył klastry o różnej gęstości, co może wskazywać na bardziej zróżnicowane grupy krajów.



Rekomendacje

- **Polityka Ekonomiczna:** Krajom o niższym PKB na mieszkańca i wsparciu społecznym zaleciłbym inwestycje w te obszary w celu poprawy poziomu szczęścia.
- **Wsparcie Społeczne:** Wzmocnienie programów wsparcia społecznego może przynieść korzyści w krajach o niskim poziomie wsparcia.



Problemy Napotkane w Projekcie

Redukcja Wymiarów:

- Proces redukcji wymiarów za pomocą PCA wymagał ode mnie szczególnej staranności, aby upewnić się, że istotne cechy danych są zachowane i dobrze reprezentowane w mniejszej liczbie wymiarów.



Podsumowanie

Kluczowe Wnioski:

- **Związek PKB i Wsparcia Społecznego ze Szczególnym Poziomem Szczęścia:** Analiza wykazała silną korelację między wysokim poziomem PKB i wsparcia społecznego a wyższym poziomem szczęścia.
- **Wyniki Klasteryzacji:** Klasteryzacja ujawniła różne grupy krajów, co sugeruje, że każda z nich może korzystać z odmiennych strategii politycznych, aby poprawić poziom szczęścia obywateli.

Dalsze Badania:

- **Rozszerzenie Analizy:** Rekomenduje się dalsze badania, które uwzględnią wpływ innych czynników, takich jak polityka zdrowotna i edukacyjna, na poziom szczęścia, aby uzyskać pełniejszy obraz determinantów szczęścia w różnych krajach.