Komputerowa analiza szeregów czasowych raport 1

Natalia Klepacka, Joanna Kołaczek

21.12.2022

Spis treści

1 Wstęp

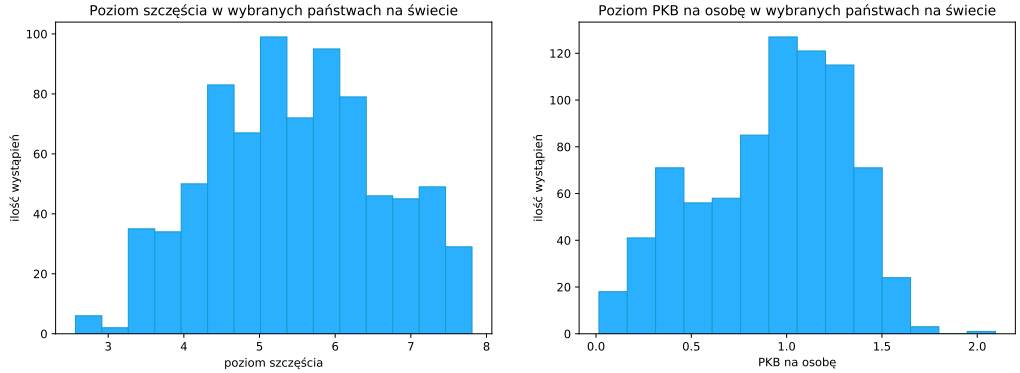
Niniejszy raport powstał na potrzeby realizacji laboratorium z Komputerowej Analizy Szeregów Czasowych, prowadzonych przez mgr Justynę Witulską, do wykładu prof. Agnieszki Wyłomańskiej. Będziemy analizować dane dotyczące poziomu szczęścia w wybranych krajach na świecie oraz jego związku z wartością PKB na osobę (w dalszej części raportu, będziemy je określać skrótowo jako szczęście i PKB). Po usunięciu wartości brakujących dysponujemy próbami o wielkości 791. Dane pochodzą *z tej strony*. Są to wyniki uzyskane przez Instytut Gallupa, w ankietach badających poziom szczęścia oraz jego możliwe indykatory, zebrane w latach 2015-2020. W raporcie przeprowadzimy analizę jednowymiarową dla dwóch zmiennych oraz zwizualizujemy je przy pomocy histogramu, dystrybuanty empirycznej oraz boxplotu. Następnie wyestymujemy współczynniki w klasycznym modelu regresji, aby ostatecznie sprawdzić, czy uzyskane residua spełniają oczekiwane założenia.

Życzymy Czytelnikowi miłej lektury.

2 Analiza jednowymiarowa zmiennej zależnej oraz zmiennej niezależnej

2.1 Wizualizacja danych

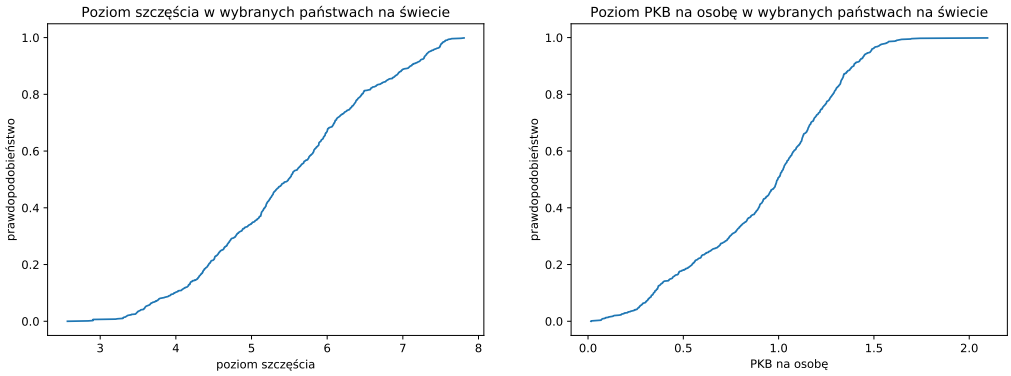
W przeprowadzanej przez nas analizie zmienną zależną jest **szczęście**, natomiast zmienną niezależną jest **PKB**. Rozkład danych możemy zobaczyć na histogramach [1]. Zauważmy, że rozkład szczęścia wydaje się bardziej symetryczny niż rozkład PKB, który sprawia wrażenie lewoskośnego. Jednakże, niestety na pierwszy rzut oka, nie przypominają nam one żadnego ze znanych rozkładów.



Rysunek 1: Histogramy

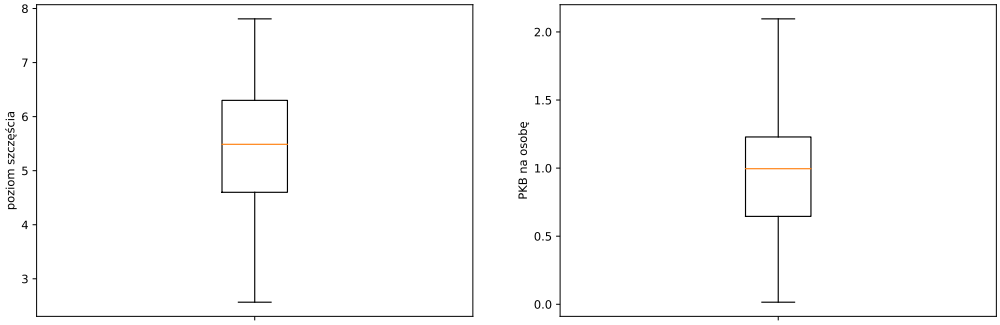
Dystrybuanty empiryczne [2] pokazują, z jakim prawdopodobieństwem natrafimy na kolejno szczęście i PKB mniejsze bądź równe od danej wartości. Gdy spojrzymy na dystrybuantę szczę-

ścia, wydaje się, że ma ona ogony lżejsze niż rozkład normalny. Dystrybuanta PKB, tak jak jego histogram, nie jest symetryczna.



Rysunek 2: Dystrybuanta empiryczna

Wykres pudełkowy (ang. *boxplot*) [3] jest to graficzna reprezentacja mediany oraz kwartyli. Końce wąsów wskazują ostatnią wartość odległą od końca "pudełka" o co najwyżej półtorej wartości rozstępu międzykwartylowego. Co ciekawe, w naszych danych nie pojawiły się wartości odstające, zatem możemy sądzić, że dane zostały rzetelnie zebrane, a ryzyko przeprowadzenia błędnej analizy (nieodzwierciedlającej rzeczywistych trendów) jest mniejsze.



Rysunek 3: Boxploty

2.2 Podstawowe miary

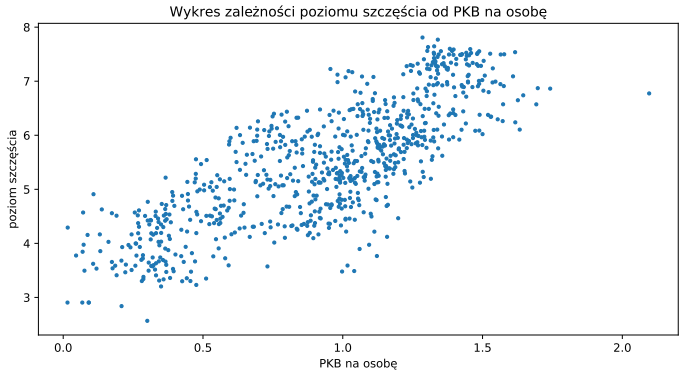
W tabeli [1] podsumowałyśmy najistotniejsze miary położenia, rozproszenia, spłaszczenia i skośności. Rzeczywiście skośność, którą odczytałyśmy z histogramu PKB, jest ujemna (zatem rozkład jest lewoskośny), natomiast skośność szczęścia jest niewielka. Przypuszczenia z wykresu dystrybuanty empirycznej również się sprawdziły — kurtoza szczęścia jest mniejsza od 3 więc rozkład ten jest platokurtyczny.

Miary		Poziom szczęścia	PKB na osobe
położenia	średnia arytmetyczna	5.47	0.93
	średnia geometryczna	5.23	0.58
	średnia harmoniczna	5.35	0.81
	średnia ucinana 10%	5.47	0.94
	mediana Q2	5.48	0.99
	Q1	4.59	0.64
	Q3	6.30	1.22
rozproszenia	rozstęp	5.24	2.08
	rozstęp międzykwartylowy	1.70	0.58
	wariancja nieobciążona	1.26	0.14
	odchylenie standardowe	1.12	0.38
	współczynnik zmienności	20.52	41.33
spłaszczenia	kurtoza	2.25	2.31
skośności	skośność	-0.01	-0.36

Tabela 1: Zestawienie statystyk.

3 Analiza zależności liniowej pomiędzy zmienną zależną a zmienną niezależną

3.1 Wykres rozproszenia i określenie zależności



Rysunek 4: Wykres rozproszenia

Na podstawie wykresu rozproszenia [4] możemy stwierdzić, że zależność pomiędzy zmiennymi prawdopodobnie jest liniowa.

3.2 Punktowa estymacja współczynników

Współczynniki regresji liniowej wyliczamy ze wzorów

$$\begin{cases} \hat{\beta}_1 = r\frac{S_y}{S_x} \\ \hat{\beta}_0 = \bar{y} - r\beta_1\bar{x} \end{cases},$$

gdzie

$$r = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_xS_y},$$

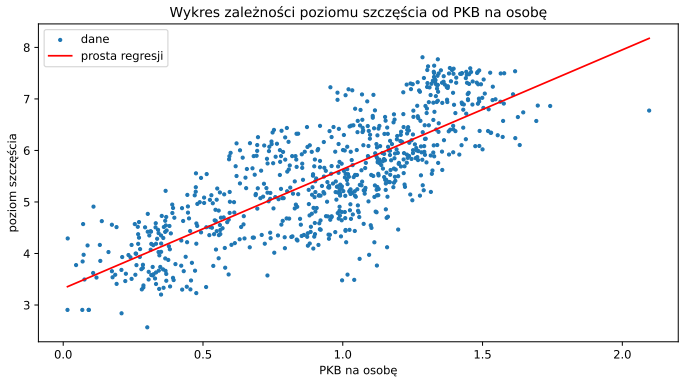
$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Wzory te zostały wyznaczone przy pomocy metody najmniejszych kwadratów.



Rysunek 5: Wykres rozproszenia z zaznaczoną prostą regresji

Z powyższych wzorów otrzymaliśmy współczynniki o wartościach

$$\begin{cases} \hat{\beta}_1 \approx 2.32 \\ \hat{\beta}_0 \approx 3.32 \end{cases}.$$

Współczynniki te opisują prostą widoczną na wykresie [5].

3.3 Przedziałowa estymacja współczynników

Z założenia o normalności rozkładu $\{\varepsilon\}_{i=1}^{n}$, przy braku znajomości jego wariancji możemy stwierdzić, że unormowane parametry β_0, β_1 mają rozkład t-studenta z n-2 stopniami swobody. Przedziały ufności przyjmują wtedy postać

$$\begin{cases} P\left(\hat{\beta}_0 - t_{n-2}(1-\frac{\alpha}{2})S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} < \beta_0 < \hat{\beta}_0 + t_{n-2}(1-\frac{\alpha}{2})S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha \\ P\left(\hat{\beta}_1 - t_{n-2}(1-\frac{\alpha}{2})\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} < \beta_1 < \hat{\beta}_1 + t_{n-2}(1-\frac{\alpha}{2})\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha \end{cases},$$

gdzie

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}.$$

Przyjmując $\alpha = 0.05$ otrzymujemy zatem, że z 95% prawdopodobieństwem

$$\begin{cases} \beta_1 \in (2.19, 2.44) \\ \beta_0 \in (3.20, 3.44) \end{cases}.$$

3.4 Ocena poziomu zależności

Poziom zależności liniowej pomiędzy zmiennymi można ocenić przy pomocy wielu różnych współczynników. Jednym z nich jest współczynnik korelacji Pearsona

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y},$$

gdzie

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

W opisywanym przypadku współczynnik ten przyjmuje wartość $r \approx 0.79$, co oznacza, że w danych występuje dosyć silna pozytywna zależność liniowa.

Jakość dopasowania modelu możemy sprawdzić też przy pomocy współczynników SST, SSE i SSR.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})$$

Wiemy, że $SST = SSR + SSE$. Ponadto im mniejsze SSE, a zarazem SSR bliższe SST, tym lepiej dopasowany jest model. W opisywanym przypadku otrzymaliśmy $SST \approx 997.74$, $SSE \approx 370.75$, oraz $SSR \approx 626.98$. Jak widać, SSE jest znacząco mniejsze od SSR, zatem nasz model jest dosyć dobrze dopasowany do danych.

3.5 Predykcja oraz jej przedziały ufności

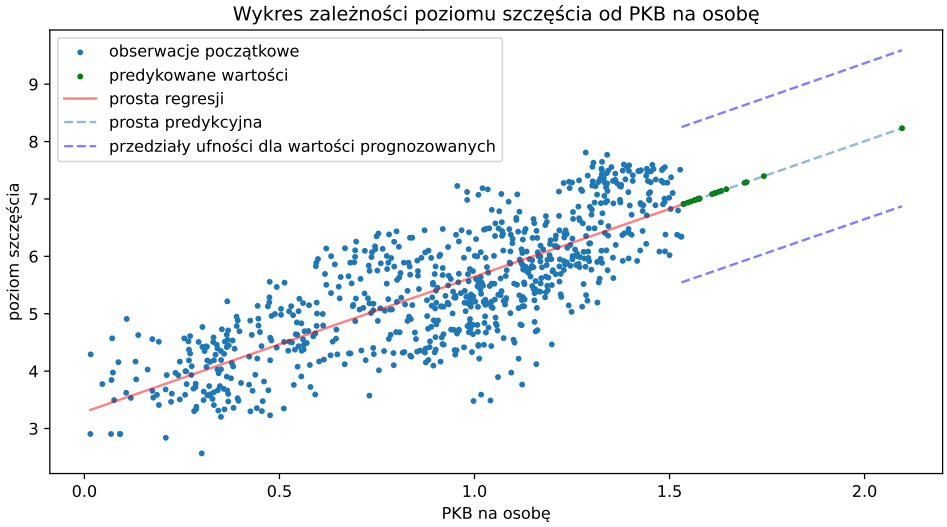
W tej części wykonamy predykcję wartości zmiennej zależnej dla 20 największych obserwacji zmiennej niezależnej. W tym celu estymujemy współczynniki ze wzorów [3.2], używając jednak tylko 771 obserwacji. Otrzymujemy w ten sposób

$$\begin{cases} \beta_1 \approx 2.36 \\ \beta_0 \approx 3.29 \end{cases}.$$

Następnie wartości dla ostatnich 20 obserwacji wyliczamy ze wzoru $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Przy założeniu, że residua modelu mają rozkład normalny o nieznanej wariancji, możemy stwierdzić, że przedziały ufności dla Y_0 przyjmują postać

$$P\left(\hat{Y}_0 - t_{n-2}\left(1-\frac{\alpha}{2}\right) S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x}_n)^2}{\sum_{i=1}^n (x_i-\bar{x}_n)^2}} \leqslant Y_0 \leqslant \hat{Y}_0 + t_{n-2}\left(1-\frac{\alpha}{2}\right) S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x}_n)^2}{\sum_{i=1}^n (x_i-\bar{x}_n)^2}}\right) = 1-\alpha.$$

Rezulaty powyższej procedury można zobaczyć na wykresie [6].



Rysunek 6: Predykcja dla ostatnich 20 obserwacji zmiennej niezależnej

4 Analiza residuów

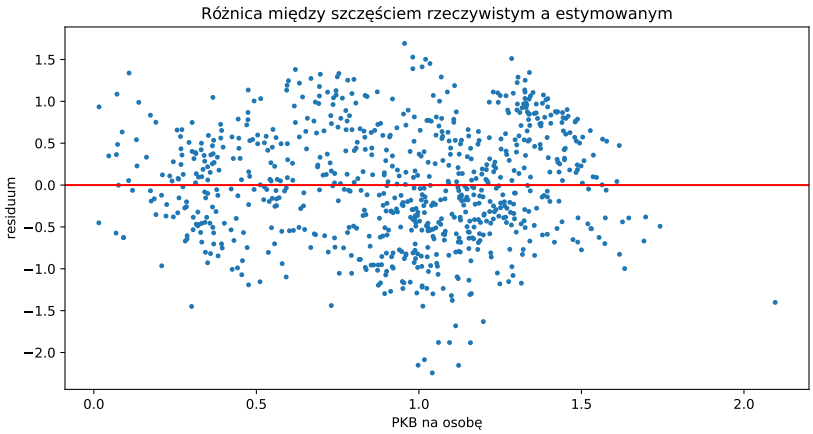
4.1 Wstęp

W tej części przeprowadzimy analizę residuów. Dzięki temu możemy ocenić czy model regresji liniowej jest dobrym wyborem dla naszego zbioru danych. Pod nazwą residuum mamy na myśli różnicę między wartością wyestymowaną a rzeczywistą.

$$e_i = y_i - \hat{y}_i$$

W klasycznym modelu regresji zakładamy, że residua mają średnią równą 0, stałą wariancję, są niezależne od siebie nawzajem i mają rozkład normalny.

4.2 Średnia i wariancja



Rysunek 7: Residua

Na wykresie [7] przedstawione zostały residua z regresji liniowej szczęścia w zależności od PKB. Czerwona linia oznacza średnią, która jest bliska zeru. Na pierwszy rzut oka widzimy, że wariancja jest raczej równa na całej długości osi zmiennej niezależnej. Aby upewnić się, czy nasze przewidywanie jest słuszne, wykonamy test Levene'a jednorodności wariancji [1].

Test ten sprawdza hipotezę zerową "wszystkie próbki pochodzą z populacji o tej samej wariancji." przeciwko hipotezie alternatywnej "Przynajmniej dwie próbki pochodzą z populacji o różnych wariancjach.". W tym celu dzieli próbkę na mniejsze podgrupy i porównuje ze sobą ich mediany. Test ten, w przeciwieństwie do testu Bartletta można stosować dla danych niepochodzących z rozkładu normalnego.

```
stat, p_value = scipy.stats.levene(random.sample(residuals,350),random.sample(
  residuals,350))
if p_value > 0.05:
  print("Wariancja jest raczej stala.")
else:
  print("Wariancja raczej nie jest stala.")

#output
Wariancja jest raczej stala.
```

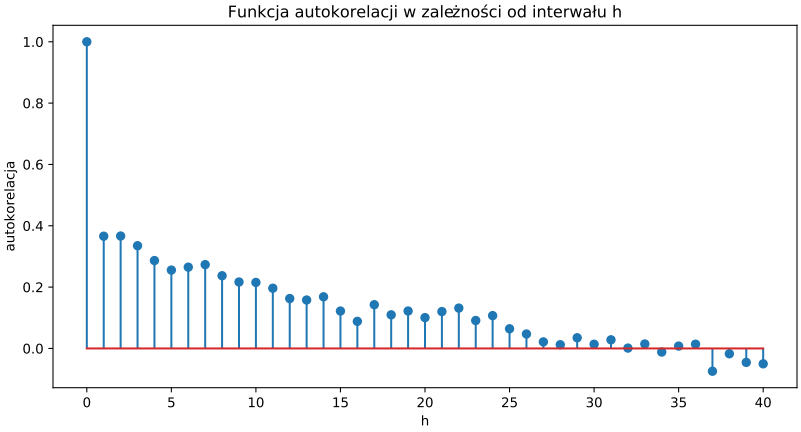
Kod 1: Test Levene'a

Po wykonaniu testu nie mamy podstaw, aby odrzucić hipotezę zerową.

4.3 Niezależność

Następnie będziemy chciały sprawdzić, czy residua są niezależne. W tym celu wykonamy wykres [8], który wykorzystuje funkcję `acf` z pakietu `statsmodels.tsa.stattools`, aby policzyć auto-

korelację.

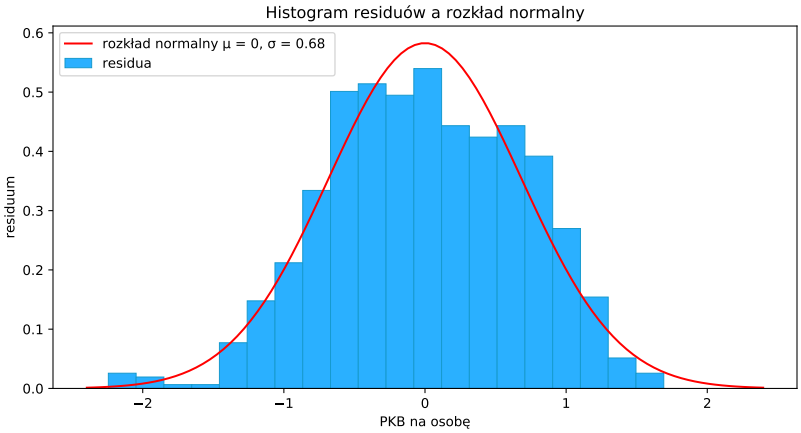


Rysunek 8: Autokorelacja residuów

Jak widać na wykresie [8], wartości funkcji autokorelacji dla $h \neq 0$ są małe, możemy zatem uznać, że residua są niezależne.

4.4 Normalność rozkładu

Na koniec, sprawdzimy jeszcze, czy rozkład residuów jest normalny. W tym celu porównamy ich histogram z gęstością rozkładu normalnego [9].



Rysunek 9: Rozkład residuów

Ponieważ możemy mieć wątpliwości wynikające z różnic na wykresie, wykonamy test D'Agostino-Pearsona sprawdzający normalność rozkładu. Polega on na porównaniu skośności i kurtozy danych z tymi samymi statystykami dla rozkładu normalnego. Wynik wykonanego przez nas testu [2] wskazuje, że residua prawdopodobnie nie mają oczekiwanego rozkładu.

```
stat, p = scipy.stats.normaltest(residuals)
if p > 0.05:
  print('Dane raczej pochodza z rozkladu normalnego')
else:
  print('Dane raczej nie pochodza z rozkladu normalnego')

#output
Dane raczej nie pochodza z rozkladu normalnego
```

Kod 2: Test D'Agostino-Pearsona

5 Podsumowanie

Na podstawie przeprowadzonej analizy możemy stwierdzić, że między poziomem szczęścia a PKB per capita dla danego kraju występuje dosyć silna zależność liniowa. Niestety badanie residuów wykazało, że nie mają one rozkładu normalnego. Oznacza to, że co prawda punktowe estymacje zostały przez nas wykonane poprawnie, jednak wszystkie estymacje przedziałowe oparte były o założenie o normalności rozkładu residuów, zatem nie mają zastosowania w tym modelu. Niestety nie znając rozkładu ϵ, nie możemy wyznaczyć faktycznych przedziałów ufności.

6 Źródła

- Wykłady

- https://www.kaggle.com/datasets/eliasturk/world-happiness-based-on-cpi-20152020
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.html
- http://tofesi.mimuw.edu.pl/~cogito/smarterpoland/samouczki/testyNormalnosci/testyNormalnosci.pdf str.11