Szymon Rewilak	Gr. 03	Data: 1.06.2021
Nr indeksu: 401145	Informatyka Techniczna WIMiIP IV semestr	Projekt zaliczeniowy

Temat projektu: Analiza statystyczna średniego wskaźnika poczucia szczęścia w krajach, w zależności od czynników demograficznych i geograficznych.

1. Opis badanego zbioru danych

Badanym zbiorem danych jest wynik badań przeprowadzonych przez *Gallup World Pull* w 2021 roku dotyczących poczucia szczęścia w 149 krajach podzielonych na 10 regionów geograficznych: Wschodnia Azja, Azja Południowa, Azja Południowo – Wschodnia, Europa Centralna i Wschodnia, Bliski Wschód, Europa Zachodnia, Ameryka Północna, Ameryka Łacińska i Karaiby, Ameryka Południowa, Afryka Subsaharyjska oraz Afryka Północna.

Zebrane dane dotyczą średniego krajowego poczucia szczęścia w skali *Ladder Scale* (skala liniowa o zakresie 1 – 10, gdzie 1 oznacza najgorsze wyobrażalne życie, a 10 oznacza najlepsze wyobrażalne życie). Reszta zebranych danych to czynniki demograficzne opisujące państwa: wskaźniki produkcji ekonomicznej, wsparcia socjalnego, przewidywanej długości życia, wolności, obecności korupcji oraz szczodrości obywateli.

2. Cel analizy statystycznej

Celem przeprowadzonej analizy statystycznej jest badanie zmienności średniego poczucia szczęścia w kraju w zależności od czynników demograficznych i geograficznych. Czynniki objaśniające poddane analizie to zmienne ilościowe: przewidywana długość życia, wskaźnik wsparcia socjalnego, wskaźnik PKB *per capita,* wskaźnik wolności wyborów życiowych a także jedna zmienna jakościowa: geograficzne położenie państwa.

3. Statystyczny opis struktury analizowanych cech, reprezentowanych przez dane liczbowe

3.1 Przewidywana wartość życia w latach

l.p.	**Minimalna wartość**	**Maksymalna wartość**	**Mediana**	**Średnia**
Przewidywana długość życia [LATA]	48,8 Czad	76,95 Singapur	66,6	64,99
	Kwantyl 1	**Kwantyl 2**	**Wariancja**	**Odchylenie standardowe**
Średni wskaźnik szczęścia [LADDER SCORE]	59,8	69,6	45,73	6,76

3.2 Wskaźnik wsparcia socjalnego, gdzie 1 oznacza państwo, gdzie mieszkańcy mają idealnie zapewnione wsparcie socjalne, a 0 – brak wsparcia socjalnego. Wsparcie socjalne to dostępność pomocy psychologicznej, dostępność służby zdrowia, wsparcie informacyjne (dostępność poradni) oraz wsparcie materialne (świadczenia socjalne i inne formy wsparcia materialnego).

l.p.	**Minimalna wartość**	**Maksymalna wartość**	**Mediana**	**Średnia**
Wskaźnik wsparcia socjalnego	0,46 Afganistan	0,98 Islandia oraz Turkmenistan	0,81	0,81
	Kwantyl 1	**Kwantyl 2**	**Wariancja**	**Odchylenie standardowe**
Wskaźnik wsparcia socjalnego	0,75	0,9	0,01	0,11

3.3 Logarytmiczna skala PKB *per capita* – wskaźnik pomiaru dobrobytu społeczeństwa, którego wartość jest obliczana jako stosunek Produktu Krajowego Brutto do ludności kraju. W analizie wykorzystaną daną jest skala logarytmiczna.

l.p.	**Minimalna wartość**	**Maksymalna wartość**	**Mediana**	**Średnia**
Logarytmiczy wskaźnik PKB *per capita* [USD]	6,64 Burundi	11,65 Luksemburg	9,57	9,43
	Kwantyl 1	**Kwantyl 2**	**Wariancja**	**Odchylenie standardowe**
Logarytmiczy wskaźnik PKB *per capita* [USD]	8,54	10,42	1,34	1,16

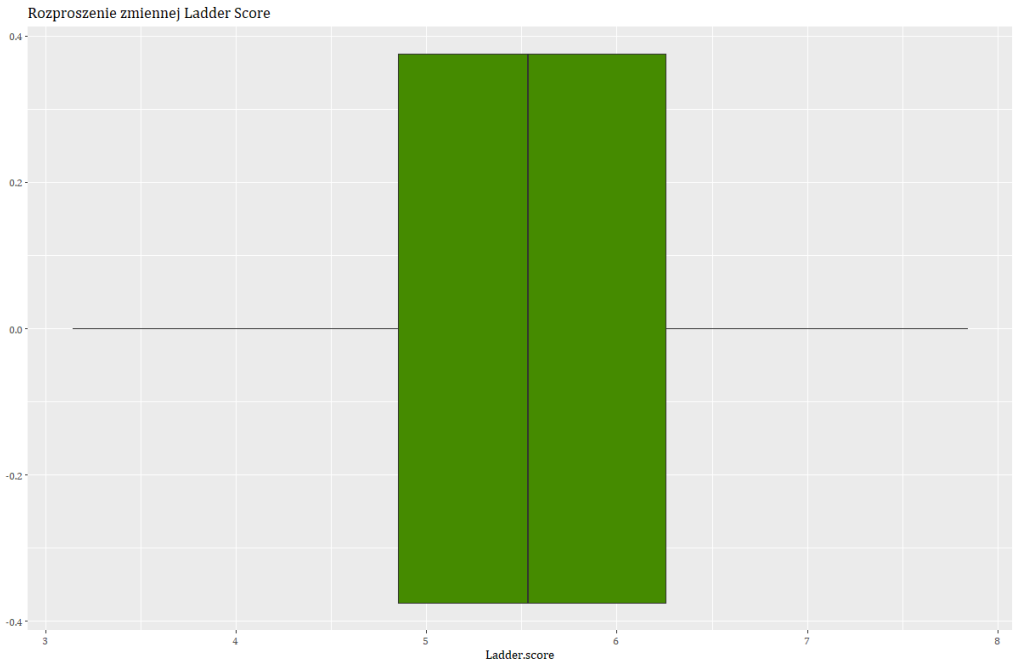
3.4 Wskaźnik wolności podejmowanych wyborów życiowych: wskaźnik obliczany na podstawie specjalistycznych testów. Wskaźnik przyjmuje wartości z zakresu <0 ; 1>, gdzie 0 oznacza brak wolności w podejmowaniu decyzji, a 1 oznacza pełną autonomię jednostki w podejmowaniu decyzji.

l.p.	**Minimalna wartość**	**Maksymalna wartość**	**Mediana**	**Średnia**
Wskaźnik wolności wyborów życiowych	0,32 Afganistan	0,97 Uzbekistan	0,72	0,79
	Kwantyl 1	**Kwantyl 2**	**Wariancja**	**Odchylenie standardowe**
Wskaźnik wolności wyborów życiowych	0,72	0,88	0,01	0,11

3.5 Wskaźnik szczęścia – w skali <1, 10>, gdzie 1 oznacza najgorsze wyobrażalne życie, a 10 najlepsze wyobrażalne życie

l.p.	**Minimalna wartość**	**Maksymalna wartość**	**Mediana**	**Średnia**
Średni wskaźnik szczęścia [LADDER SCORE]	2,52 Afganistan	7,84 Finlandia	5,53	5,53
	Kwantyl 1	**Kwantyl 2**	**Wariancja**	**Odchylenie standardowe**
Średni wskaźnik szczęścia [LADDER SCORE]	4,86	6,26	1,15	1,07

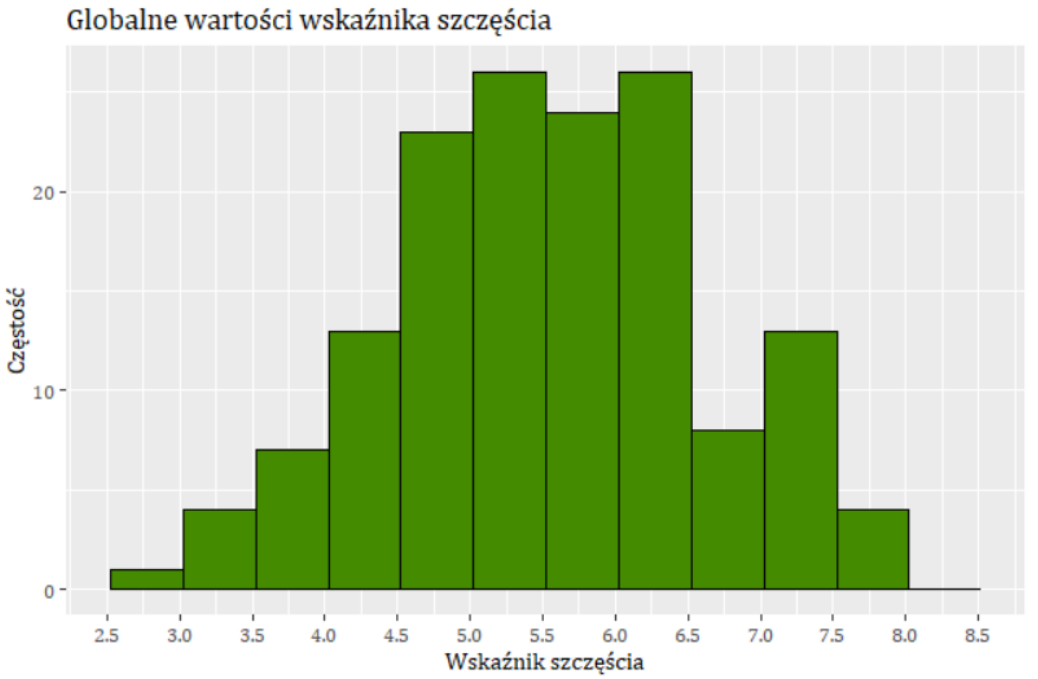
3.6 Wykres ramka – wąsy wskaźnika *Ladder Score*:



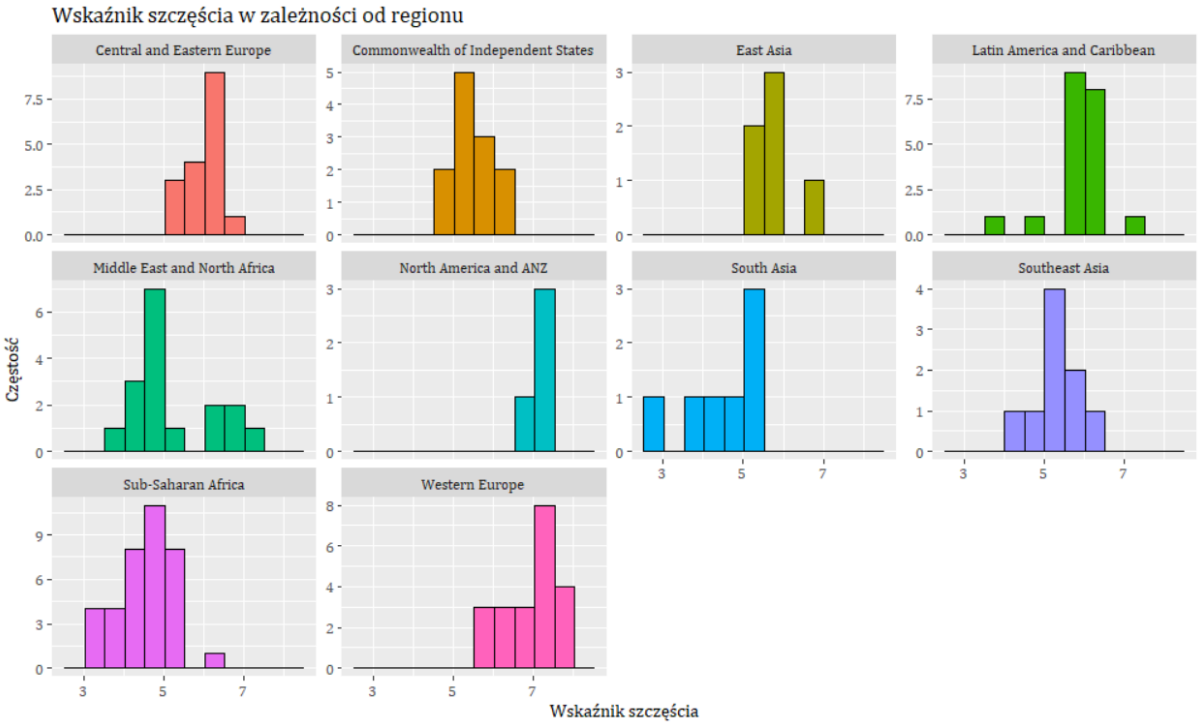
Wykres ramka wąsy dla wskaźnika Ladder Score.

Analizując wykres ramka wąsy można zauważyć, że rozkład zmiennej jest niemal symetryczny. Zmienna jest dość mocno rozproszona, co widać po długości wąsów z obyu stron pudełka wykresu. Można wnioskować, że jest bardzo wiele krajów, w których wskaźnik szczęścia jest bardzo niski – oraz wiele krajów, w których wskaźnik szczęścia jest wysoki. Zatem na świecie występuje zauważalne zróżnicowanie wskaźnika *Ladder Score.*

3.7 Globalny rozkład średniego krajowego wskaźnika *Ladder Score*:



Histogram (1). Rozkład globalnego średniego wskaźnika szczęścia w krajach.



Histogram (2). Rozkład średniego wskaźnika szczęścia w krajach w zależności od regionu.

Obserwując histogramy *(1)* zauważalna jest delikatna lewostronna skośność. Najwięcej zaobserwowanych średnich wskaźników szczęścia to wskaźniki z przedziału <5 ; 6>. Można wnioskować, że średni wskaźnik dla znacznej większości krajów jest w zakresie przeciętnego poczucia szczęścia.

Na podstawie histogramu *(2)* można dostrzec silną zależność pomiędzy rejonem geograficznym kraju a jego średnim wskaźniku *Ladder Score*. Zdecydowanie najwyższe wskaźniki są obserwowane w krajach w rejonie Europy Zachodniej oraz Ameryki Północnej. Najniższe wskaźniki zostały zaobserwowane w rejonach Bliskiego Wschodu, Azji Południowej, Afryki Północnej oraz Afryki Subsaharyjskiej. Zauważalna tendencja prowadzi do wniosku, że ludzie w krajach bardziej rozwiniętych gospodarczo wykazują średnio większe zadowolenie z życia niż ludzie zamieszkujący kraje uboższe oraz posiadające mniej sprzyjające warunki klimatyczne (kraje Afryki).

4. Wnioskowanie statystyczne

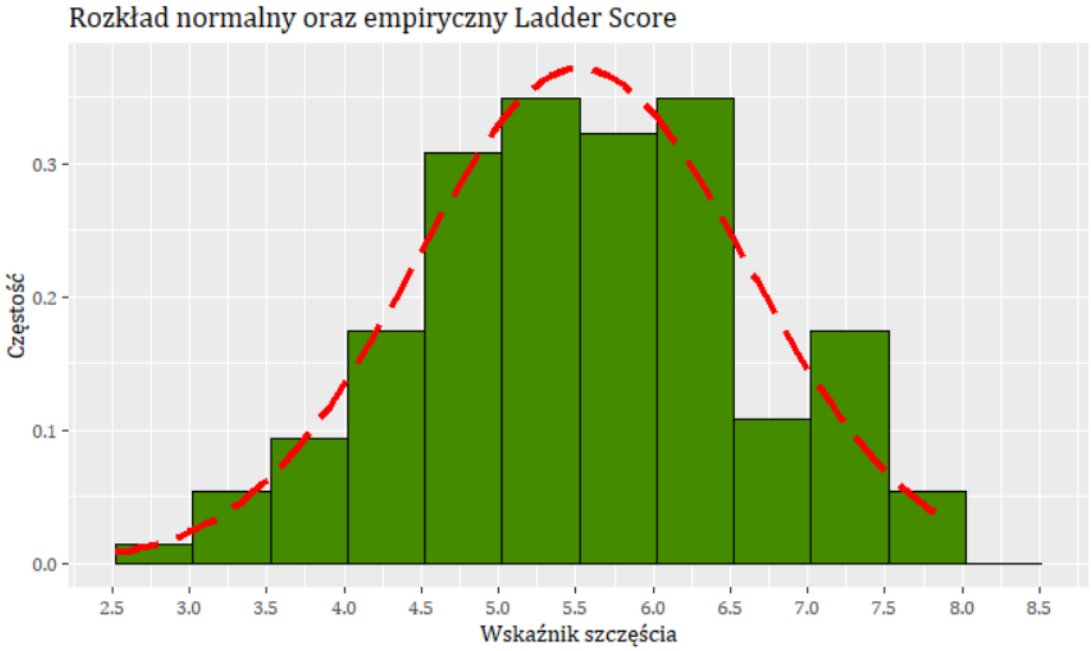
4.1 Przedział ufności dla wartości przewidywanej wskaźnika szczęścia w skali *Ladder Scale.*

Przyjęty współczynnik ufności $\alpha = 0{,}05$.

```
> round(5.533+c(-1,1)*1.073*pnorm(0.975)/sqrt(nrow(data)),3)
[1] 5.460 5.606
```

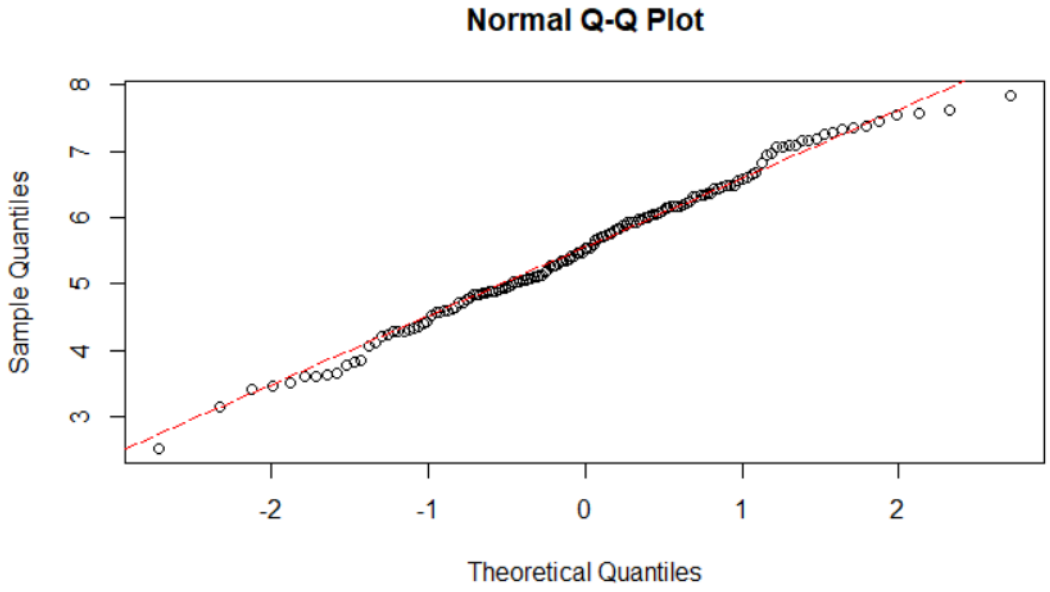
Na poziomie ufności 0.95 można stwierdzić, że wartość oczekiwana średniego wskaźnika poczucia szczęścia w kraju znajduje się w zakresie (5,46 ; 5.606). Oznacza to, że oczekiwaną wartością średniego wskaźnika szczęścia jest nieco powyżej wartości 5 – czyli wartości życia określanego jako przeciętne.

4.2 Zgodność rozkładu zmiennej *Ladder score* z rozkładem normalnym



Histogram (3). Rozkład zmiennej Ladder Score z nałożoną linią rozkładu normalnego.

Zbadano zbieżność rozkładu empirycznego zmiennej objaśnianej z rozkładem normalnym:



Wykres rozkładu normalnego zmiennej Ladder Score z nałożoną linią rozkładu normalnego.

Przyjęto hipotezy: *H0:* wskaźnik *Ladder Score* ma rozkład normalny, *H1:* wskaźnik *Ladder Score* nie ma rozkładu normalnego.

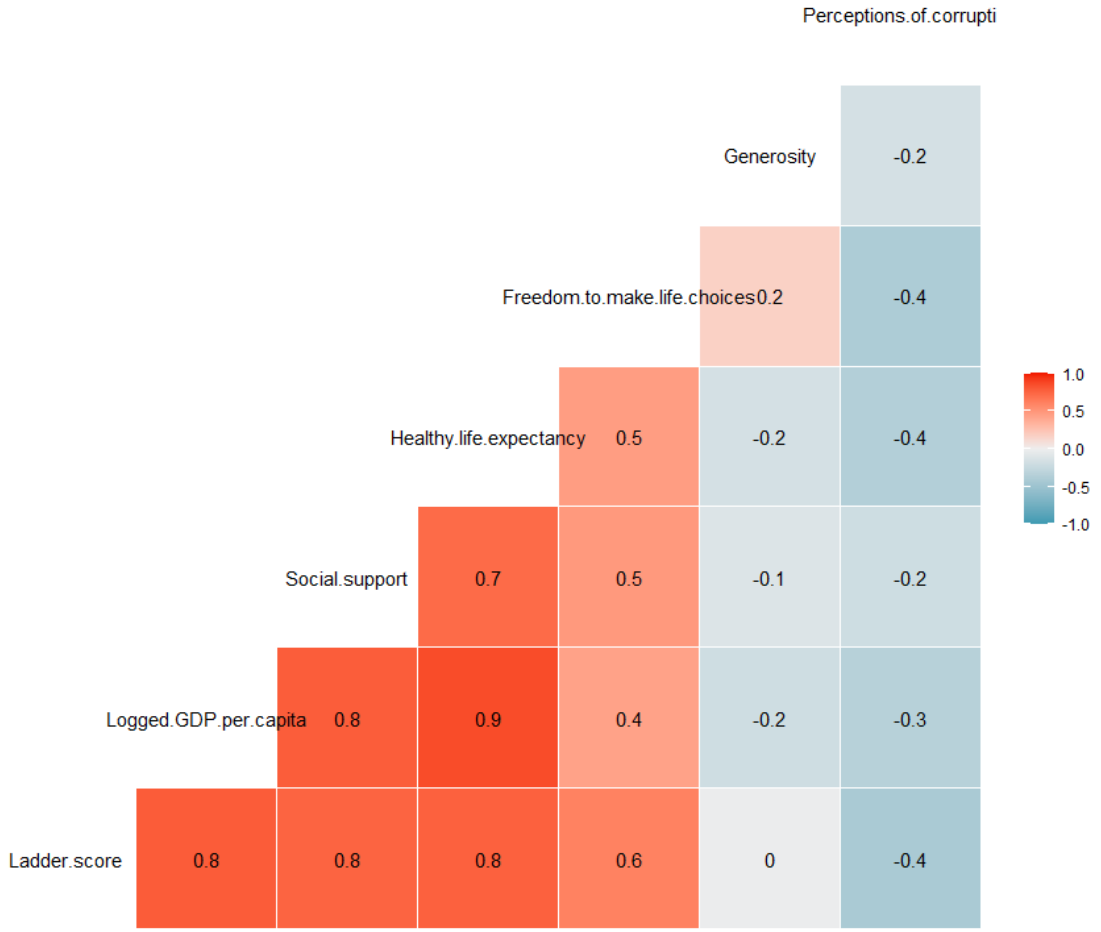
```
        Shapiro-Wilk normality test

data:  data$Ladder.score
W = 0.99125, p-value = 0.4893
```

W celu zweryfikowania hipotez wykorzystano test Shapiro – Wilka. Ponieważ *pvalue* > α nie ma podstaw do odrzucenia hipotezy *H0.* Można przyjąć, że zmienna przyjmuje rozkład normalny. Dalej, można wyciągnąć wniosek, że globalny wskaźnik szczęścia jest zróżnicowany w sposób zbliżony do losowego rozkładu.

4.3 Model regresji

Utworzono macierz korelacji między zebranymi danymi. Zmienną objaśnianą w modelu jest wskaźnik szczęścia *Ladder Score*.

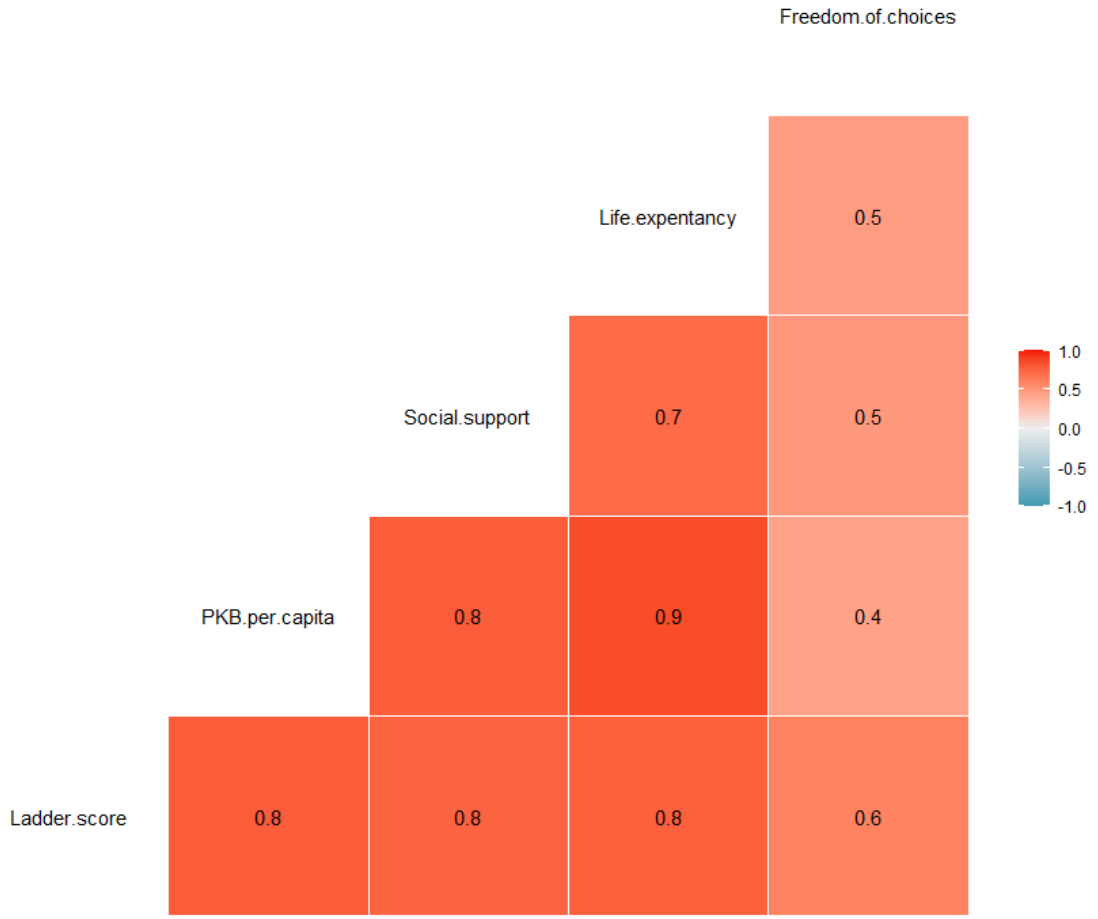


Macierz korelacji między badanymi zmiennymi.

Dla otrzymanych współczynników korelacji między zmiennymi objaśniającymi i zmienną objaśnianą zbadano ich istotność statystyczną:

```
> cbind(cor.test(data$Ladder.score, data$Logged.GDP.per.capita)$p.value,
+       cor.test(data$Ladder.score, data$Social.support)[[3]],
+       cor.test(data$Ladder.score, data$Healthy.life.expectancy)[[3]],
+       cor.test(data$Ladder.score, data$Freedom.to.make.life.choices)[[3]],
+       cor.test(data$Ladder.score, data$Generosity)[[3]],
+       cor.test(data$Ladder.score, data$Perceptions.of.corruption)[[3]])
             [,1]         [,2]         [,3]         [,4]      [,5]         [,6]
[1,] 5.228089e-33 6.051517e-29 2.961721e-30 2.054574e-16 0.8294125 8.881143e-08
```

Dla każdego ze współczynników przyjęto hipotezy: *H0:* współczynnik korelacji = 0, *H1:* współczynnik korelacji nie jest równy zero i jest istotny statystycznie. Ponieważ *pvalue* są bardzo małe można odrzucić hipotezy zerowe. Jedynym współczynnikiem korelacji, który nie jest istotny statystycznie jest współczynnik korelacji między zmienną objaśnianą a zmienną objaśniającą opisującą hojność obywateli. Aby zwiększyć istotność statystyczną modelu usunięto tę zmienną z modelu. Usunięto również zmienną wskaźnika korupcji, którego współczynnik korelacji był niski.



Macierz korelacji po wytypowaniu zmiennych do modelu regresji.

Utworzony model globalny:

Ladder.score=0,29*(PKB.per.capita)+2,17(Wsparcie.socjalne)+0.03(Przew.długość.życia)+2,5(wolność.decyzji)-3,11

Weryfikacja statystyczna utworzonego modelu

5. Weryfikacja statystyczna przyjętego modelu

5.1 Istotność statystyczna modelu

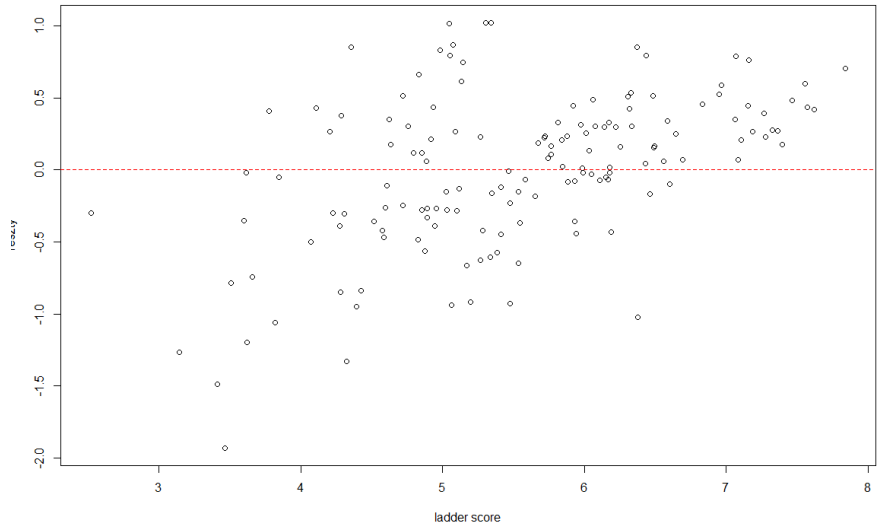
Przyjęto hipotezy: *H0:* model nie jest istotny statystycznie, *H1:* model jest istotny statystycznie

pvalue < *α,* zatem istnieją podstawy do odrzucenia *H0.* Model można uznać za istotny statystycznie.

```
p-value: < 2.2e-16
```

5.2 Analiza reszt

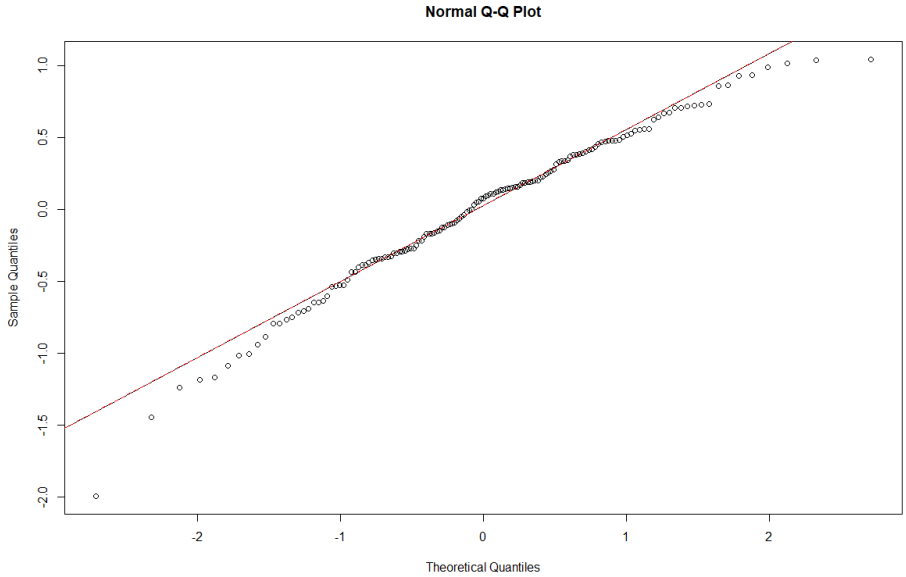
- Zbadano losowość odchyleń reszt modelu



Wykres odchylenia reszt.

Analizując wykres odchylenia reszt można przyjąć, że model spełnia założenie o losowym odchyleniu reszt.

- Normalność rozkładu reszt.



Wykres normalności rozkładu reszt.

Przyjęto hipotezy: *H0:* reszty mają rozkład normalny, *H1:* reszty nie mają rozkładu normalnego.
Hipotezy zweryfikowano testem Shapiro – Wilka:

```
        Shapiro-Wilk normality test

data:  model_all$residuals
W = 0.97967, p-value = 0.02629
```

pvalue < *α*, zatem istnieją podstawy do odrzucenia hipotezy *H0.* Można przyjąć, że reszty nie spełniają założenia o rozkładzie normalnym.

- Nieobciążoność reszt

```
> mean(model_all$residuals)
[1] 1.910524e-17
```

Wartość oczekiwana reszt jest bliska zeru, zatem można przyjąć, że założenie o nieobciążoności reszt jest spełnione.

- Założenie o występowaniu homoscedastyczności

```
        studentized Breusch-Pagan test

data:  model
BP = 15.079, df = 4, p-value = 0.00454
```

Przyjęto hipotezy: *H0:* występuje homoscedastyczność, *H1:* nie występuje homoscedastyczność. Ponieważ *pvalue* < *α* istnieją podstawy do odrzucenia hipotezy *H0.* Można przyjąć, że występuje heteroscedastyczność rozkładu reszt. Może to oznaczać, że istnieje jakiś nieuwzględniony czynnik, który wpływa na rozkład reszt.

Zakłócenia współliniowości zmiennych

```
> vif(model_all)
      Logged.GDP.per.capita                Social.support      Healthy.life.expectancy Freedom.to.make.life.choices
                   4.882355                      2.826338                      4.025146                      1.352807
```

Wartości współczynników współliniowości zmiennych są niskie, zatem można przyjąć, że współliniowość zmiennych nie wpływa na istotność statystyczną modelu.

Interpretacja i analiza modelu

```
Multiple R-squared:  0.7442,
```

Współczynnik determinacji wynosi 0,7442. To oznacza, że model regresji w około 74% wyjaśnia zmienność zmiennej objaśnianej, którą w modelu uzależniono od pięciu zmiennych objaśniających. Można przewidywać, że wartość współczynnika szczęścia *Ladder Score* zależy od zmiennych objaśniających w następującej relacji:

Ladder.score=0,29*(PKB.per.capita)+2,17(Wsparcie.socjalne)+0.03(Przew.długość.życia)+2,5(wolność.decyzji)-3,11

Należy jednak pamiętać, że model regresji nie spełnił założeń o normalności rozkładu reszt oraz o homoscedastyczności. Należy założyć, że pomimo wysokiego współczynnika determinacji istnieją nieuwzględnione czynniki, które wpływają na zmienność współczynnika szczęścia.