Семинарска работа по предметот Бизнис статистика

Тема: Статистичка обработка на податоците од Светскиот извештај на среќа во државите од 2019 година изработена во програмскиот јазик R

Изработил
Николов Мартин, индекс 193113

Ментор
проф. д-р Наташа Илиевска

Вовед

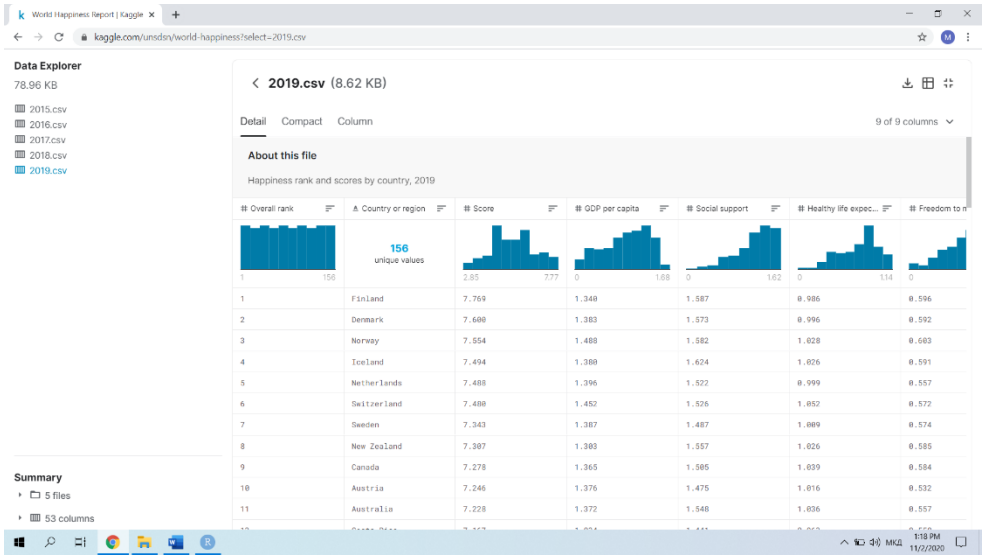
Семинарската работа е изработена за Светскиот извештај за среќа од 2019 година со користење на податочното множество од следниот линк: https://www.kaggle.com/unsdsn/world-happiness?select=2019.csv.

Податочното множество е составено од 156 единки (во случајот, секоја единка претставува соодветна држава) и истото е анализирано според 6 квантитативни обележја:

1. Бруто домашен производ
2. социјална поддршка,
3. очекувања за здрав живот,
4. слобода на избор
5. великодушност
6. перцепции за корупција.

Според вредностите на горенаведените обележја, се рангираат државите по соодветно ниво и се рангираат по факторот среќа.

Од интерес на оваа анализа се следните две обележја: GDP и социјална поддршка. Сите пресметки ќе бидат извршени на овие две обележја.



Слика 1. Приказ на соодветното податочно множество

Прв дел

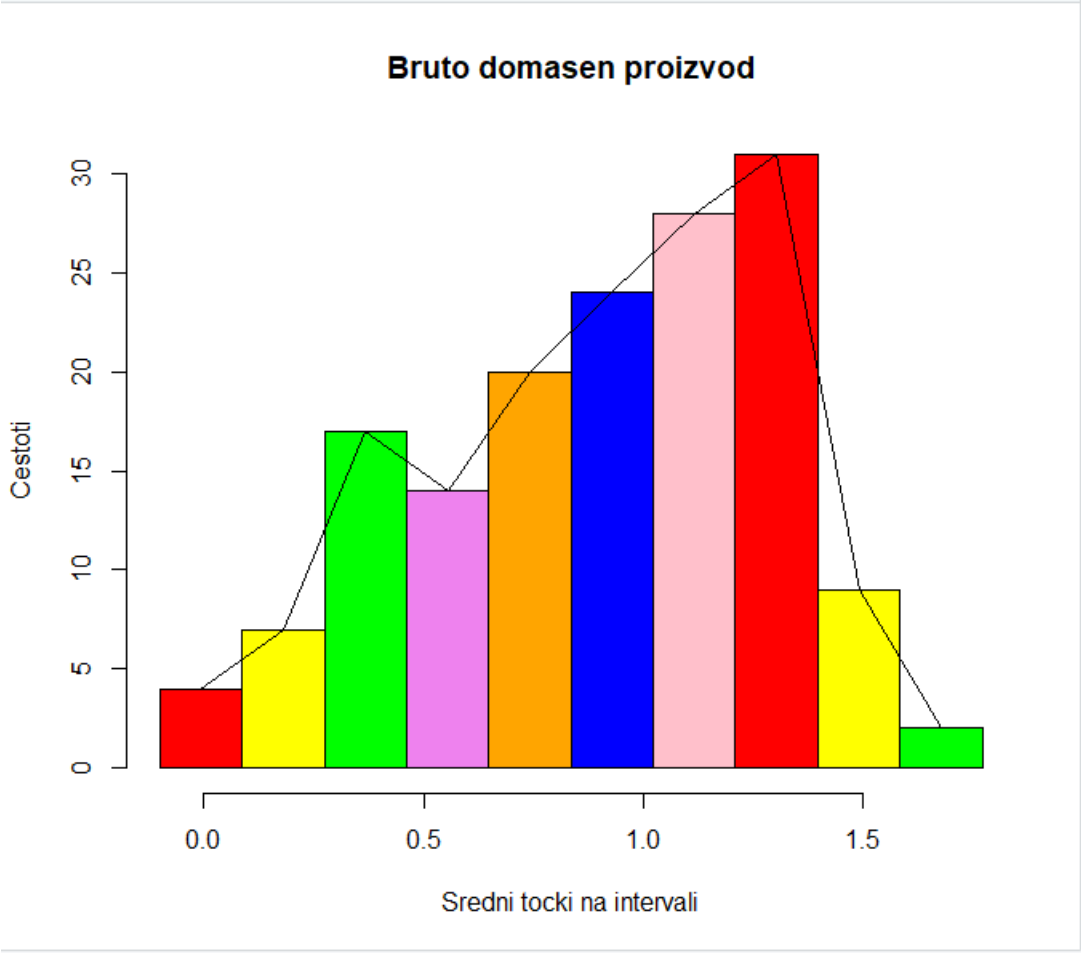
2.1 Табела за честоти, хистограми и полигони

Како прво обележје од интерес е обележјето GDP и истото е сместено во векторот GDP, додека обележјето Social Support е сместено во векторот Social. Дветте обележја имаат премногу мали вредности на податоци во интервал од [0.0, 1.68] за првото обележје и [0.0, 1.62] за второто обележје. Исто така, вредностите на податоците и за дветте обележја се распоредени во 10 интервали бидејќи обемот на примерокот изнесува 156 единки. Малиот опсег на податоци повлекува и ширината на интервалите да биде мала односно не поголема од 0.2 за дветте обележја. Интервалите на дветте обележја се поместени за 0.1 во лево од минимумот и 0.1 во десно од максимумот со цел во истите да бидат застапени соодветниот минимум и максимум.

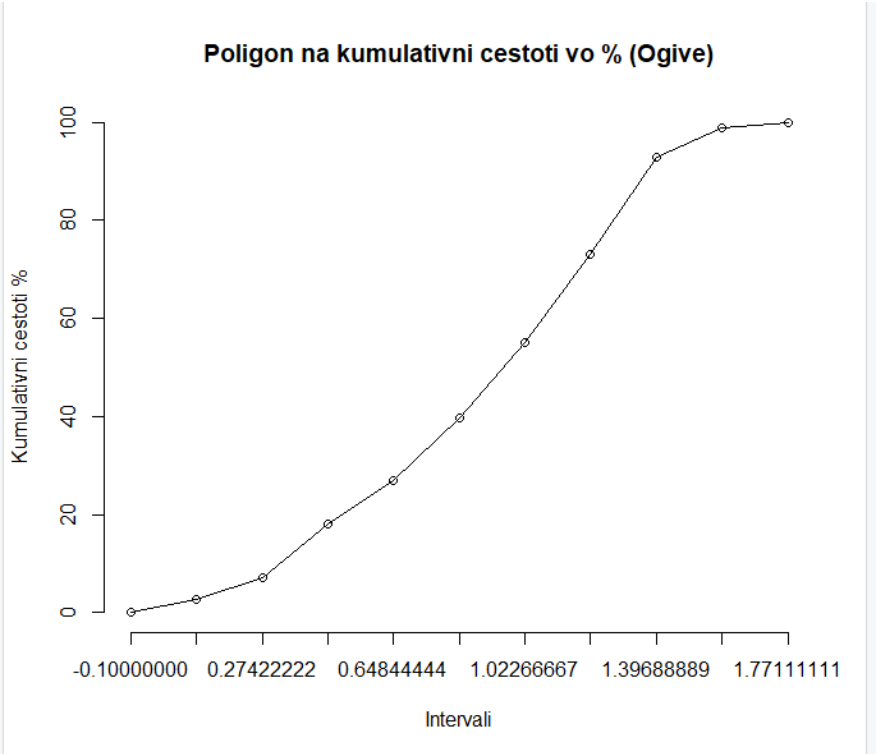
За првото обележје, интервалот (1.21, 1.4] содржи најголем дел од вредностите односно приближно 20% (31 вредности) од податоците, а првиот интервал најмалку, само 4 вредности (0.03%). Со оглед на кумулативната честота во проценти , може да забележиме дека вредностите на податоците приближно 50% се поголеми од половината (0.84) на најголемата вредност (1.68).

```
> GDP.table
                 sredniGDP          freq         Rfreq       Cumfreq         Pfreq     P_Cumfreq
(-0.1,0.0871]  -0.006444444   4.000000000  0.025641026   4.000000000   2.564102564   2.564102564
(0.0871,0.274]  0.180666667   7.000000000  0.044871795  11.000000000   4.487179487   7.051282051
(0.274,0.461]   0.367777778  17.000000000  0.108974359  28.000000000  10.897435897  17.948717949
(0.461,0.648]   0.554888889  14.000000000  0.089743590  42.000000000   8.974358974  26.923076923
(0.648,0.836]   0.742000000  20.000000000  0.128205128  62.000000000  12.820512821  39.743589744
(0.836,1.02]    0.929111111  24.000000000  0.153846154  86.000000000  15.384615385  55.128205128
(1.02,1.21]     1.116222222  28.000000000  0.179487179 114.000000000  17.948717949  73.076923077
(1.21,1.4]      1.303333333  31.000000000  0.198717949 145.000000000  19.871794872  92.948717949
(1.4,1.58]      1.490444444   9.000000000  0.057692308 154.000000000   5.769230769  98.717948718
(1.58,1.77]     1.677555556   2.000000000  0.012820513 156.000000000   1.282051282 100.000000000
>
```

Слика 2.1.1 Приказ на табелата за распределба на честоти за GDP



Слика 2.1.2 Приказ на хистограм на честота на обележје GDP

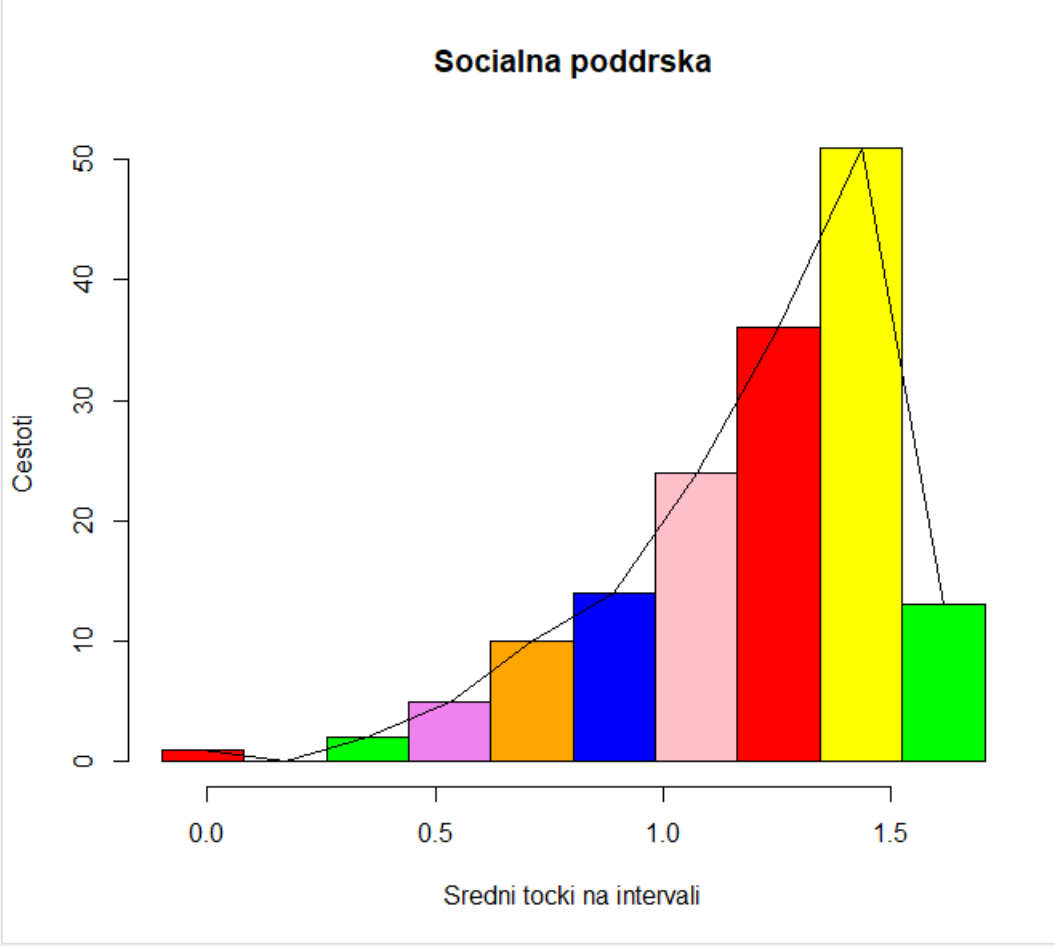


Слика 2.1.3 Приказ полигон на кумулативни честоти на обележје GDP

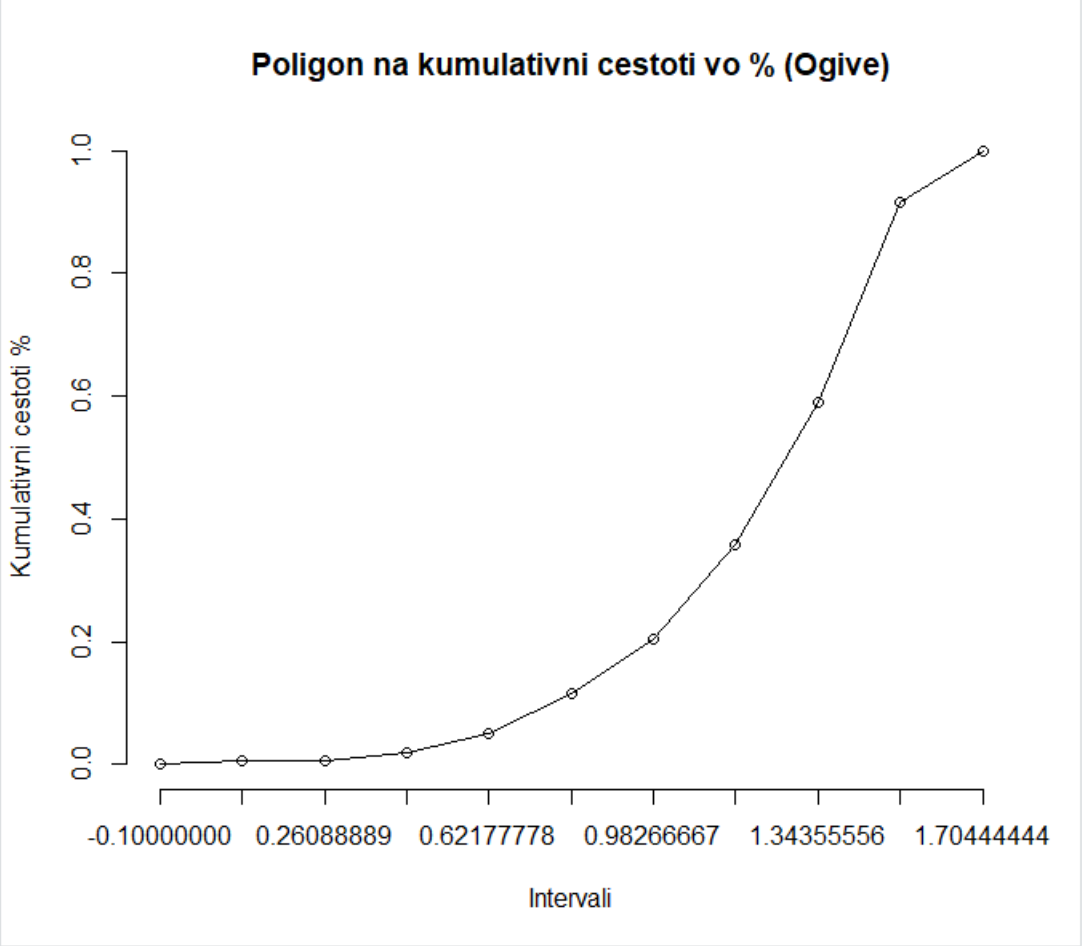
За второто обележје, можеме да заклучиме дека најголема честота(51) имаме во интервалот (1.34, 1.52] односно се содржат приближно 33% од вредностите на податоците, додека во интервалот помеѓу (0.0804, 0.261] не се содржи ниту една вредност. Со оглед на процентот на кумулативна честота, податоците се поголеми од средната вредност(0.812) на максимумот(1.624) и податоците се распределени поблиску до максимумот.

```
> Social.table
             sredniSocial       cestota      RCestota    kumCestota      pCestota          pKum
(-0.1,0.0804]   -0.009777778   1.000000000   0.006410256   1.000000000   0.641025641   0.641025641
(0.0804,0.261]   0.170666667   0.000000000   0.000000000   1.000000000   0.000000000   0.641025641
(0.261,0.441]    0.351111111   2.000000000   0.012820513   3.000000000   1.282051282   1.923076923
(0.441,0.622]    0.531555556   5.000000000   0.032051282   8.000000000   3.205128205   5.128205128
(0.622,0.802]    0.712000000  10.000000000   0.064102564  18.000000000   6.410256410  11.538461538
(0.802,0.983]    0.892444444  14.000000000   0.089743590  32.000000000   8.974358974  20.512820513
(0.983,1.16]     1.072888889  24.000000000   0.153846154  56.000000000  15.384615385  35.897435897
(1.16,1.34]      1.253333333  36.000000000   0.230769231  92.000000000  23.076923077  58.974358974
(1.34,1.52]      1.433777778  51.000000000   0.326923077 143.000000000  32.692307692  91.666666667
(1.52,1.7]       1.614222222  13.000000000   0.083333333 156.000000000   8.333333333 100.000000000
>
```

Слика 2.1.4 Приказ на табелата за распределба на честоти за Social Support



Слика 2.1.5 Приказ на хистограм на честота на обележје Social Support



Слика 2.1.6 Приказ на полигонот за кумулативни честоти на обележје Social Support

2.2 Стебло - лист дијаграм

Стебло – лист дијаграмот за првото обележје покажува дека вредностите на податоците се меѓу 0.0 и 1.68, додека за второто обележје се од 0.0 до 1.62 .

```
>     Obelezje1 <-Funkcija(GDP,0,2)
0   |   0 3 5 7 9 14 19 20 27 27 27 28 29 31 31 32 33 33 34 35 35 36 37 38 38 39 45 45 48 49 49 51 55 55 56 57 57 57 58 61 62 64 66 67 68 69 69 70 71
  74 74 76 76 78 79 80 80 81 81 81 82 83 84 85 86 88 88 89 91 91 92 93 94 95 95 95 96 96 96 98 98 99
1   |   0 0 0 1 3 3 4 4 4 4 5 5 5 6 7 7 9 9 10 12 12 15 16 16 16 17 18 18 18 19 20 21 22 22 23 24 24 25 26 26 26 27 28 29 29 30 30 30 32 33 33 34 36 3
6 36 37 37 37 38 38 38 39 40 40 43 44 45 49 50 50 50 57 61 68
>
```

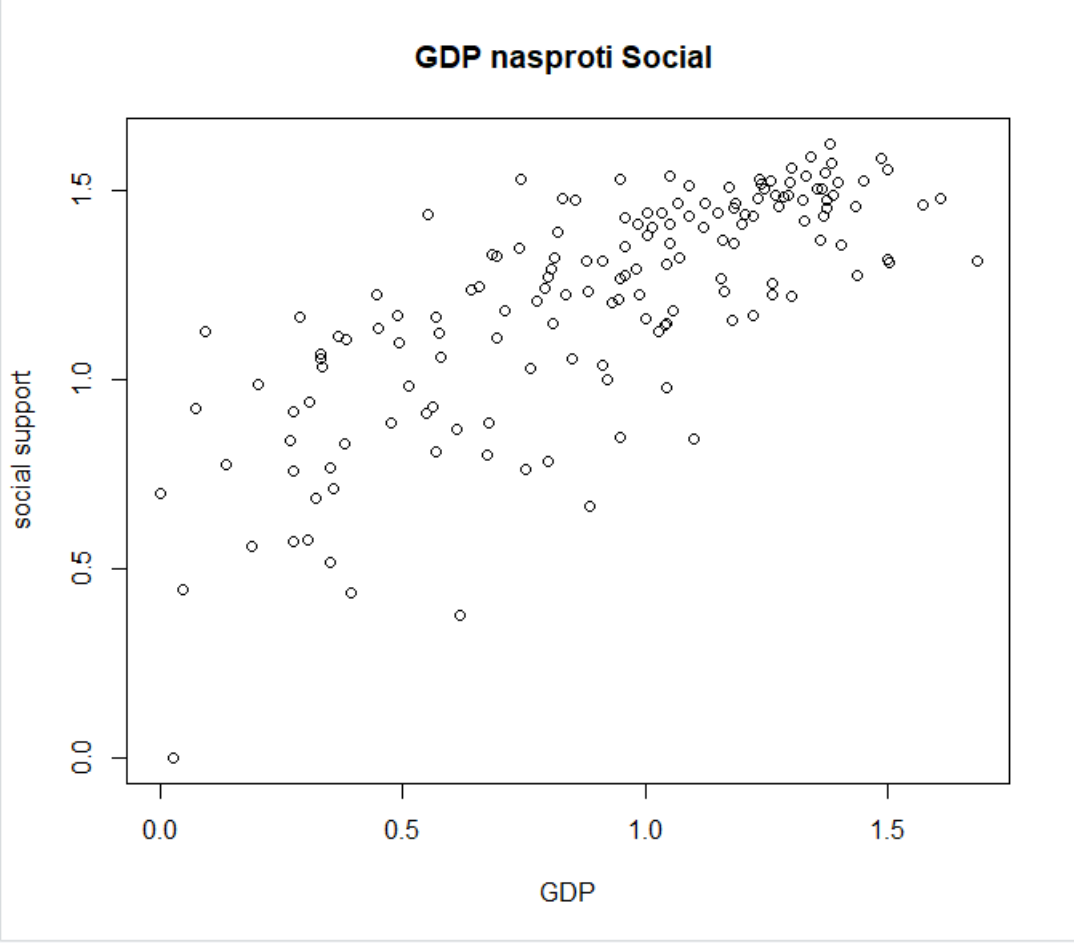
Слика 2.2.1 Приказ на стебло-лист дијаграмот на обележје GDP

```
>   Obelezje2 <- Funkcija(Social,0,2)
0   |   0 38 44 45 52 56 57 58 67 69 70 71 76 76 77 77 78 80 81 83 84 84 85 87 88 89 91 92 92 93 94 98 98 99
1   |   0 3 3 4 5 6 6 7 10 10 11 11 12 12 12 13 15 15 15 16 16 16 17 17 17 18 18 20 21 21 22 22 22 23 23 23 23 24 24 25 25 26 27 27 27 28 29 29 30 31
  31 31 31 32 32 32 32 33 35 35 36 36 36 37 37 38 39 40 40 41 41 41 42 43 43 43 43 44 44 44 44 44 45 45 46 46 46 47 47 47 47 48 48 48 48 48 48 49 49
  49 50 50 50 51 51 51 52 52 52 53 53 53 53 54 54 55 55 56 57 58 59 62
>
```

Слика 2.2.2 Приказ на стебло-лист дијаграмот за обележје SocialSupport

2.3 График на расејување

Според графикот на расејување даден подолу и според премногу малиот интервал на вредности за двете обележја (0,2), со мало децимално покачување на вредноста на бруто домашен производ , се зголемува и децимално вредноста на обележјето социјална поддршка.



Слика 2.3 Приказ на дијаграм на расејување на GDP и Social Support

2.4 Мода, медијана и просек

За првото обележје најзастапен вредност е 0.96 што воедно е и медијана, односно 50% од вредностите се пред 0.96, а останатите 50% од вредностите се после 0.96 . Аритметичката средина за првото обележје изнесува 0.905 .

```
>     medijana
[1] 0.96
>     moda
[1] 0.96
>     prosek
[1] 0.9051474
>
```

Слика 2.4.1 Приказ на медијана, мода и просек на обележје GDP

За второто обележје најзастапен вредност е 1.47, а медијаната има вредност 1.2716, односно 50% од вредностите се пред 0.96, а останатите 50% од вредностите се после 0.96 . Аритметичката средина за првото обележје изнесува 1.209 .

```
>     medijana2
[1] 1.2715
>     moda2
[1] 1.465
>     prosek2
[1] 1.208814
>
```

Слика 2.4.2 Приказ на медијана, мода и просек на обележје Social Support

2.5 Квартали, опсег и интерквартален опсег

Првиот квартал на обележјето GDP покажува дека 25% од вредностите на податоците се помали од 0.5945, а 75% се поголеми од таа вредност. Вториот квартал на оваа обележје е ист со вредноста на медијаната односно 0.96 . Опсегот изнесува 1.684 што во ова обележје е еднаков на максимумот бидејќи минимумот е еднаков на 0 и податоците се во ранг од [0,1.684] . Интерквартален распон има вредност 0.6395 .

```
>     Kvartal1.GDP
[1] 0.5945
>     Kvartal2.GDP
[1] 0.96
>     Opseg.GDP
[1] 1.684
>    Inter.Raspon.GDP
[1] 0.6395
>
```

Слика 2.5.1 Приказ на вредностите на кварталите, опсег и интерквартален опсег на обележје GDP

За обележјето Social support, вредноста на првиот квартал е поголема и изнесува 1.055 односно 25% се помали од оваа вредност, а 75% од вредностите на податоците се поголеми од истата. И овде медијаната е еднаква на вредноста на вториот квартал односно 1.2715 . Опсегот

повторно има вредност на максимумот од вредностите на податоците бидејќи и во овој случај минимумот е 0.0 и вредностите на податоците се во опсег [0, 1.624]

Интерквартални от опсег изнесува 0.3975 што е помал во споредба на претходното обележје.

```
>     Kvartal1.Social
[1] 1.0555
>     Kvartal2.Social
[1] 1.2715
>     Inter.Raspon.Social
[1] 0.3975
>     Opseg.Social
[1] 1.624
>
```

Слика 2.5.2 Приказ на вредностите на кварталите, опсег и интерквартален опсег на обележје Social Support

2.6 Дисперзија и стандардна девијација

Првото обележје има поголемо варирање на вредностите на податоците околу просекот односно 0.399 додека второто обележје има стандардна девијација од 0.299 . Вредностите на податоците на второто обележје се поблиску до просекот отколку вредностите на податоците на првото обележје.

```
>     St.Devijacija.GDP
[1] 0.3983895
>     Disperzija.GDP
[1] 0.1587142
>     St.Devijacija.Social
[1] 0.2991914
>     Disperzija.Social
[1] 0.08951549
>
```

Слика 2.6 Приказ на дисперзија и стандардна девијација на GDP и Social Support обележјата

2.7 Коефициент на корелација

Според вредноста на коефициентот на корелација, 0.755 , што е поблиску до 1, има не толку силна линеарна зависност меѓу овие две обележја

```
>     r<-cor(GDP,Social)
>     r
[1] 0.7549057
>
```

Слика 2.7 Приказ на коефициент на корелација на GDP и Social Support

Втор дел

3.1 Интервал на доверба

Интервал на доверба се тестира во примерот на обележјето GDP и параметар е очекување на истото обележје. Со 95% интервал на доверба очекуваме дека просекот на популацијата ќе има вредност меѓу [0.88024 , 0.93005]. Поради големината на примерокот (156) и централната гранична теорема , овде е искористен z-test со дисперзијата на примерокот место стандардната девијација на обележје.

```
>   test1 <- simple.z.test(GDP,Disperzija.GDP)
>   test1
[1] 0.8802416 0.9300533
>
```

Слика 3.1 Приказ на интервал на доверба за очекување на обележје GDP

3.2 Хипотеза за тестиран параметар

Ја тестираме нултата хипотеза дека очувањето на обележјето ќе е 1. Согласно направениот t-test добиваме дека p-вредноста за овој настан е 0.003413 и интервал на доверба 0.95, се отфрла нултата хипотеза бидејќи 0.05 > 0.0034. Доколку истиот тест го направиме со очекување од 0.9, добиваме p-вредност од 0.872 која е поголема од 0.05 и не ја отфрламе нултата хипотеза. Ниво на значајност на тест ни е 0.05 .

```
>   t.test(GDP,alternative = "two.sided",mu=1,conf.level = 0.95)

        One Sample t-test

data:  GDP
t = -2.9737, df = 155, p-value = 0.003413
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 0.8421391 0.9681557
sample estimates:
mean of x
0.9051474

>
```

Слика 3.2.1 Приказ на хипотезата за тестирање на просек=1 на обележје GDP

```
>   t.test(GDP,alternative = "two.sided",mu=0.9,conf.level = 0.95)

        One Sample t-test

data:  GDP
t = 0.16138, df = 155, p-value = 0.872
alternative hypothesis: true mean is not equal to 0.9
95 percent confidence interval:
 0.8421391 0.9681557
sample estimates:
mean of x
0.9051474
```

Слика 3.2.2 Приказ на хипотезата за тестирање на очекување=0.9 на обележје GDP

3.3 Тест на распределба

Бидејќи p-вредноста е еднаква на 1 и истата е поголема од 0.05 бидејќи се користи интервал на доверба 95%, следува дека дветте обележја имаат иста распределба. Исто така, бидејќи станува збор за премногу мали опсези на податоци, но и на премногу мали очекувани вредности кои може да не бидат вклучени во пресметката на хи – квадратниот тест.

```
>   chisq.test(rbind(GDP,Social))

        Pearson's Chi-squared test

data:  rbind(GDP, Social)
X-squared = 9.2526, df = 155, p-value = 1

Warning message:
In chisq.test(rbind(GDP, Social)) :
  Chi-squared approximation may be incorrect
```

Слика 3.3 Приказ на тест на распределба на обележја GDP и SocialSupport

3.4 Хипотези за независност

Според p-вредноста која изнесува 0.452 и истата е поголема од 0.05, со што следува дека нултата хипотеза не се отфрла односно дветте обележја се независни.

```
>   chisq.test(tabelaZaTestiranje)

        Pearson's Chi-squared test

data:  tabelaZaTestiranje
X-squared = 20904, df = 20880, p-value = 0.452

Warning message:
In chisq.test(tabelaZaTestiranje) :
  Chi-squared approximation may be incorrect
```

Слика 3.4 Приказ на тест за независност на обележја GDP и Social support

3.5 Регресиона анализа

Правата на регресија на обележјето GDP по обележјето Social е:

GDP = 1.0052 * Social - 0.3099

Наклонот на правата е 1.0052 и тоа значи за секое зголемување на вредноста на обележјето Social за една единица, вредноста на GDP се очекува да се зголеми за 1.0052 .

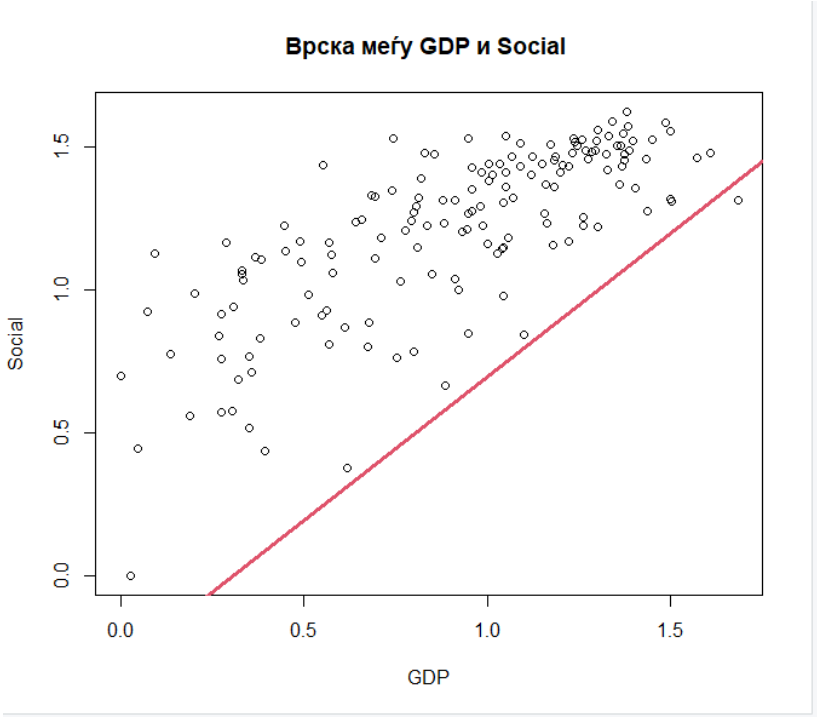
```
Call:
lm(formula = GDP ~ Social)

Coefficients:
(Intercept)       Social
    -0.3099       1.0052

> |
```

Слика 3.5.1 Приказ на проценетата права

Согласно претставениот график на расејување и вредноста на R која изнесува 0.5699, но и малиот опсег на вредностите на податоците, имаме послаба линеарна поврзаност на обележјата GDP и Social.



Слика 3.5.2 Приказ на дијаграм на расејување