Изработено од Никита Вуевски индекс: 196141

Семинарска работа по предметот Бизнис статистика

Податочното множество кое е користено во овој труд се наоѓа на следниот линк: NBA Players | Kaggle. Сетот на податоци содржи повеќе од две децении податоци за секој играч кој бил дел од списокот на НБА тимови. Ги опфаќа демографските променливи како што се возраста, висината, тежината и местото на раѓање, биографските детали како тимот за кој играл, годината во која бил одбран да игра во НБА и на кое место. Покрај тоа, има основна статистика за бодови, како што се одиграни натпревари, просечен број на поени, скокови, асистенции итн.

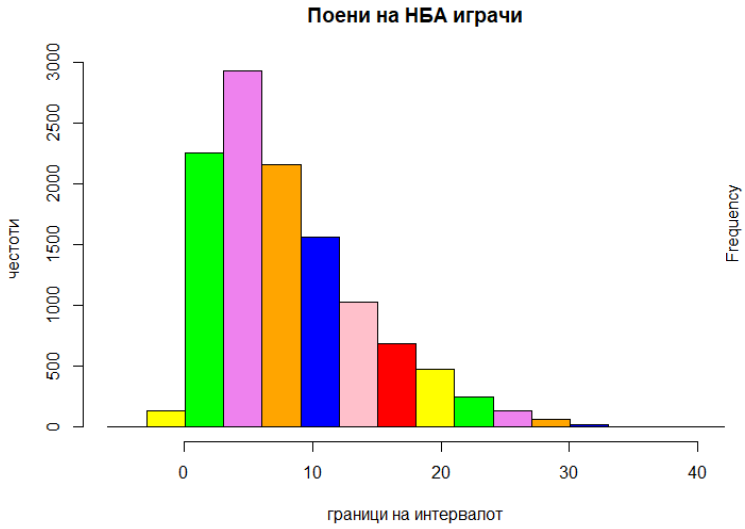
Дел А

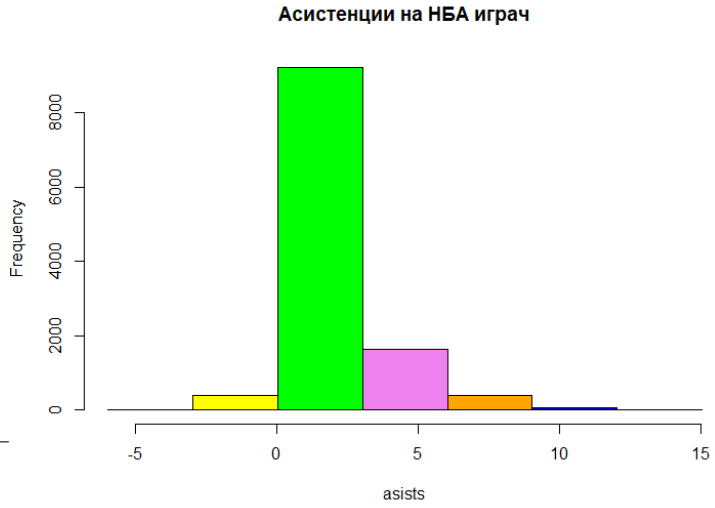
Како обележје за истражување ќе ги одберам колоните “pts” и “ast”. Бидејќи податочното множество содржи 11700 индекси, тоа значи дека обемот на двата примероци е n = 11700. Најголемиот податок е 36.1 а најмалиот податок е 0 во колоната pts, а во колоната ast најголем податок е 11.7 а најмал е 0. Според формулата дека рангот на податоците е разлика помеѓу најголемиот и најмалиот податок, рангот на pts е 36.1 – 0 = 36.1 и на ast е 11.7 – 0 = 11.7. Бројот на интервали (k) во кој се групираат податоците е 1 + 3.322 (log_10n). После пресметка оваа пресметка се добива к = 14.09956. А кое ќе може да се заокружи на 14. Најчесто бројот на интервали се зема да е помеѓу 5 и 15 а нашиот број припаѓа во тој интервал. Ширината на интервалот се пресметува преку формулата w ≥ R / k. Во нашиот примерок тоа е 36.1 / 14 = 2.578 но која ќе ја земеме како 3. Бидејќи вкупната ширина на целиот интервал е (w * K = 14 * 3 = 42) (42 - интервалот се проширува за 2.95 од двете страни.

	freq	Rfreq	Cumfreq	Pfreq	P_Cumfreq
[-2.95,0.05)	132	1.128205e-02	132	1.128205128	1.128205
[0.05,3.05)	2260	1.931624e-01	2392	19.316239316	20.444444
[3.05,6.05)	2935	2.508547e-01	5327	25.085470085	45.529915
[6.05,9.05)	2163	1.848718e-01	7490	18.487179487	64.017094
[9.05,12.1)	1560	1.333333e-01	9050	13.333333333	77.350427
[12.1,15.1)	1031	8.811966e-02	10081	8.811965812	86.162393
[15.1,18.1)	682	5.829060e-02	10763	5.829059829	91.991453
[18.1,21.1)	476	4.068376e-02	11239	4.068376068	96.059829
[21.1,24.1)	244	2.085470e-02	11483	2.085470085	98.145299
[24.1,27.1)	133	1.136752e-02	11616	1.136752137	99.282051
[27.1,30.1)	65	5.555556e-03	11681	0.555555556	99.837607
[30.1,33)	16	1.367521e-03	11697	0.136752137	99.974359
[33,36)	2	1.709402e-04	11699	0.017094017	99.991453
[36,39)	1	8.547009e-05	11700	0.008547009	100.000000

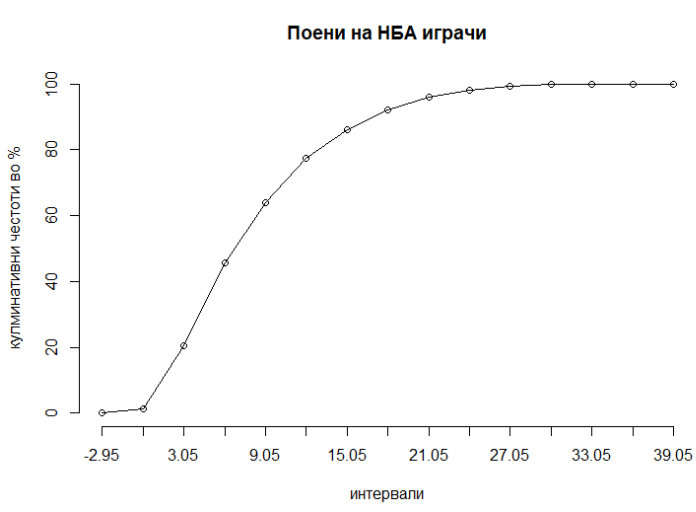
Изработено од Никита Вуевски индекс: 196141

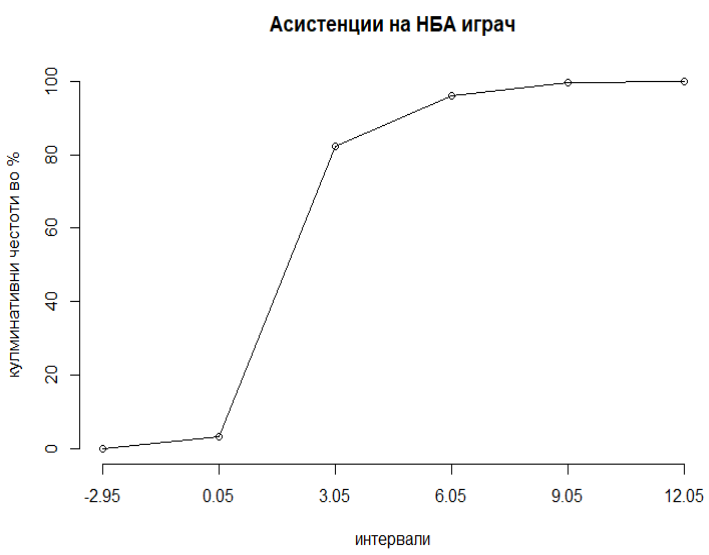
Хистограм за поените на НБА играчите





Полигон на Кулминативни честоти во проценти





2. Стебло лист дијаграмо. Стеблото ги претставува цифрите пред децималната запирка, а листот цифрите после децималната запирка кој се наоѓа после “l”. Некои од податоците во стебло лист дијаграмот од двата примероци.

```
 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7
 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8
 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9
 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
8  |  0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5 6 6 6 6 6 6 6 6 6 7 7 7 7 7
 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9
7  |  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 9 9 9 9 9 9 9 9 9
5  |  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8
 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
10  |  0 0 0 1 1 2 2 2 3 3 4 4 5 5 5 5 5 7 7 7 7 7 7 8 8 9
6  |  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9
 9 9 9 9 9 9 9
9  |  0 0 0 0 0 0 1 1 1 1 1 2 2 2 3 3 3 4 6 6 7 7 8 8 8 8 9 9
>
```

```
24  |  0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 6 6 7 7 7 8 8 9 9
16  |  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3
  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
  5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
  8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
10  |  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
  3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5
  5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
  6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
  8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
26  |  0 0 0 1 1 1 1 2 3 3 4 4 4 4 4 4 4 4 4 5 5 6 6 6 6 6 7 7 8 8 8 8 8 8 9 9 9 9 9
23  |  0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 4 4 4 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 7 8 8 8 8 8 8 9 9 9
28  |  0 0 0 0 1 1 1 1 2 2 3 3 4 4 4 5 5 7 7 7 8 8 9 9
31  |  1 3 4 4 6 6
30  |  0 0 0 1 1 2 4 5 7
32  |  0 0 1
33  |  0
35  |  4
36  |  1
34  |  3
```

3. График на Расејување

Точките во ова претставување на Графикот на расејување, се чинат дека генерално паѓаат по линија. Во нашиот случај имаме позитивна врска, ова го гледаме во графикот како линерана шема која се крева од лево кон десно. Коефицентот на корелација е помеѓу 0 и 1, што значи дека има силна линерана поврзаност, односно обележјата се силно поврзани.

4. Мода, Медијана и Средна Вредност

Средната вредност на поени обележјето изнесува 8.169299. Медијаната на ова обележје изнесува 6.7.

Бидејќи R нема функција за Мода, направив своја и таа изнесува 2.

```
  moda = function(x) {
                        unique_x <- unique(x)
                        tabulate_x <- tabulate(match(x, unique_x))
                        unique_x[tabulate_x == max(tabulate_x)]
                        }
moda(points)
```

Средна вредност на Асистенции обележјето изнесува 1.811179, модата е 0.3 и медијаната 1.2.

5.Квартили за истото обележје

Поени	Асистенции
Q1 = P25 = 3.6	Q1 = P25 = 0.6
Q2 = P50 = 6.7	Q1 = P25 = 1.2
Q3 = P75 = 11.5	Q1 = P25 = 2.4

Рангот е разликата помеѓу најголемата со најмалата вредност. А тие ги наоѓаме преку функцијата range(points) = 36.1 и range(assists) = 11.7

Интерквартални распон е раслика меѓу третиот и првиот квартил.

IQR(points) = 7.9 IQR(assists) = 1.8

6. Дисперзија и Стандардна Девијација

Дисперзијата на истите обележја се: 35.47531 и 3.211683

Стандардните Девијации се: 5.956115 и 1.792117

7. Коефицент на Корелација

Коефицентот на корелација помеѓу поени и асистенции на НБА играч е 0.6565643

Дел Б

1. Обележјето кое ќе се разгледува во овој дел е “ast” колоната, односно бројот на асистенции на НБА играчот таа сезона. Големината на ова обележје изнесува 11700, стандардната девијација 1.792117 и средната вредност 1.811179. Поради фактот што обемот на примерокот е поголем од 30, тој е голем, позната е дисперзијата односно и стандардната девијација. Се користи Z статистика. Маргината на грешка претставува производ од Z нормирано за 97.5 % ниво на доверба со стандардната грешка. Горната/долната граница на интервалот на доверба се средната вредност +/- маргина на грешка. Интервалот е (1.778707, 1.843652)

2. Според предходното барање познато е дека, големината на обележјето е 11700, стандардната девијација е 1.792117, дисперзијата 3.211683 и очекувањето 1.811179 (која ја добив како збир од интервалите на доверба, поделен со 2, која е). На случаен начин се одбираат 887 единки од истото обележје и се пресметува средната вредност на истото обележје за тој примерок, таа е 1.858039. Со ниво на значајност 0.05 треба да се провери дали бројот на асистенции отстапува од стандардите. Се поставува нулта хипотеза H0 : μ = 1.811179и се тестира дали е точна. Се проверува двостран тест, односно алтернативната хипотеза е Ha : $\mu \neq$ 1.811179. Со помош на централната гранична теорема, за Z се добива дека е 0.7743354. Критичниот домен е (-бескрај, -1.96) унија (+1.96, +бескрај). Па бидејќи z тестот не припаѓа во критичниот домен, нултата хипотеза не се отфрла. Односно бројот на асистенциите оваа сезона е според стандардите.

3. Тест на распределба. Ќе тестираме дали обележјето X – број на асистенции на НБА играч во таа сезона, има нормална распределба. Се поставуваат следните хипотези: H0 : X има нормална распределба. Ha : X нема нормална распределба. Може да се користи Shapiro-Wilk тест, и да се селектираат првите 5000 елементи од примерокот или со Anderson-Darling тест. И со двата теста, p-вредноста се добива дека е помала од алфа, што значи нултата хипотеза се отфрла. Односно обележјето X нема нормална распределба.

4. Проверка за зависност на две обележја. Како обележја за истражување ќе ги земам. X: поени на играч. Y : асистенции на играч. Со ниво на значајност алфа =

0.05, п-вредноста се добива 0. Бидејќи п вредноста е помала од алфа, нултата хипотеза се отфрла односно двете обележја се зависни.

5. Регресиона анализа на обележјата процент на офанзивни скокови наспроти процент на дефанзивни скокови. Гледаме дали има линеарна поврзаност помеѓу нив. Коефицентот на детерминираност е 0.4310767. Тој број е помеѓу 0 и 1, што значи дека има послаба линеарна поврзаност меѓу x и y. Ова е прикажано и визуелно.