Семинарска работа по предметот Бизнис статистика

Извор: [Find Open Datasets and Machine Learning Projects | Kaggle]

Избрано е податочното множество [Petrol/Gas Prices Worldwide | Kaggle], кое ни дава информации за најновите цени на горивото и потрошувачката на гориво на светско ниво. Се состои од категориски податоци кои ги претставуваат државите во светот односно во овој случај тоа се 181 држава и нумерички податоци со кои се претставени обележјата:

- Дневна потрошувачка на гориво во барели
- Годишна потрошувачка по жител во галони
- Цена по галон во долари
- Цена по литар во долари

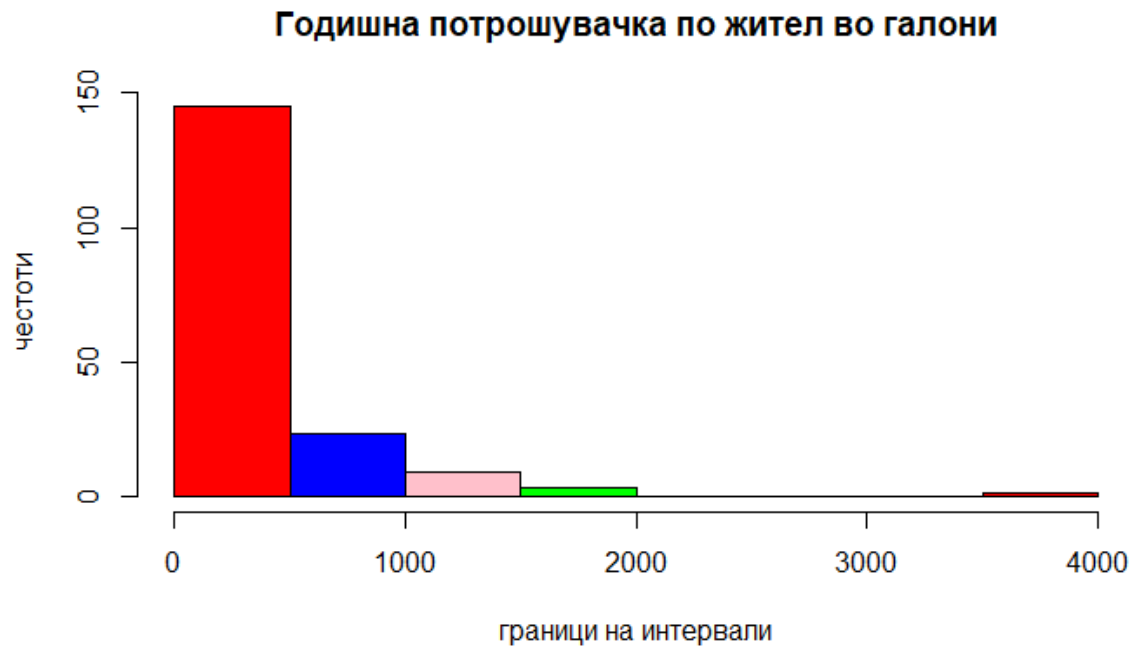
А. Прв дел

1. Табела со распределба на честоти

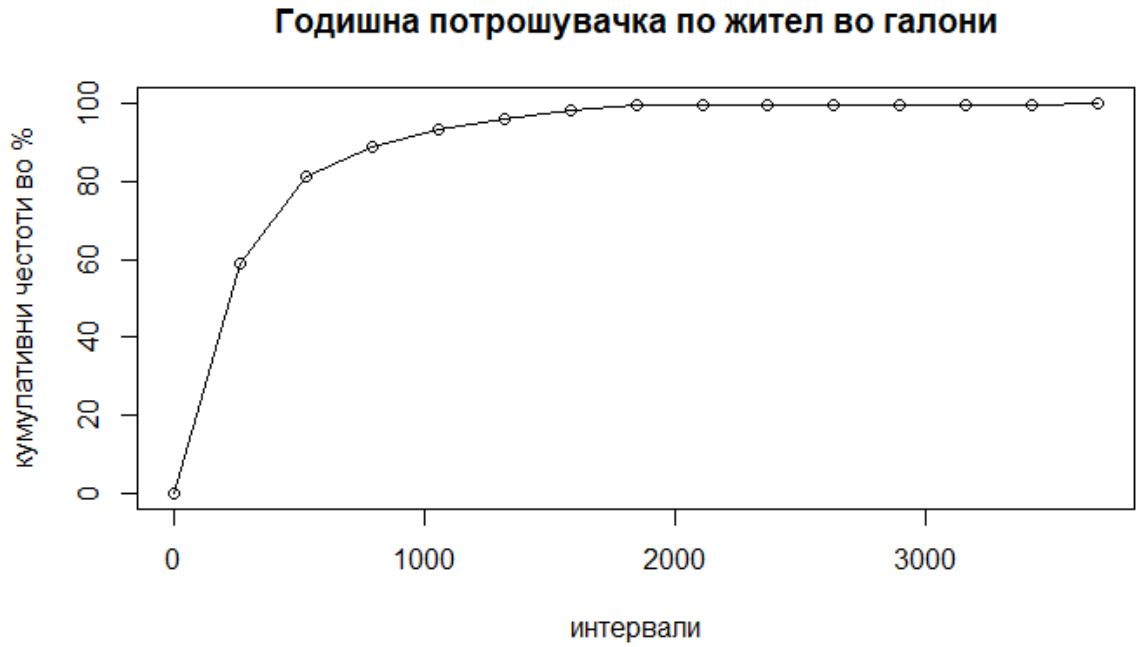
Избрано е обележјето “годишна потрошувачка по жител во галони”. Обемот на примерокот е 181 (n=181). Најголемиот податок е 3679.5 , а најмалиот е 2.2. Рангот е разлика меѓу најголемиот и најмалиот податок, во овој случај ќе изнесува 3677.3. Бројот на интервали (k) е квадратен корен од обемот на примерокот $\sqrt{181}$ што е приближно 14. Ширината на интервалите се пресметува со формулата w≥R/k, во овој случај 3677.3 / 14 = 262.66 или приближно 263. Табелата со распределба на честоти добиена во R изгледа вака:

Интервали	Средни точки	Честоти	Релативни честоти	Релативни честоти во %	Кумулативни честоти	Релативни кумулативни честоти	Релативни кумулативни честоти во %
[2,265)	133,7	107	0,59	59,12	107	0,59	59,12
[265,528)	396,7	40	0,22	22,10	147	0,81	81,22
[528,791)	659,7	14	0,08	7,73	161	0,89	88,95
[791,1054)	922,7	8	0,04	4,42	169	0,93	93,37
[1054,1317)	1185,7	5	0,03	2,76	174	0,96	96,13
[1317,1580)	1448,7	4	0,02	2,21	178	0,98	98,34
[1580,1843)	1711,7	2	0,01	1,10	180	0,99	99,45
[1843,2106)	1974,7	0	0,00	0,00	180	0,99	99,45
[2106,2369)	2237,7	0	0,00	0,00	180	0,99	99,45
[2369,2632)	2500,7	0	0,00	0,00	180	0,99	99,45
[2632,2895)	2763,7	0	0,00	0,00	180	0,99	99,45
[2895,3158)	3026,7	0	0,00	0,00	180	0,99	99,45
[3158,3421)	3289,7	0	0,00	0,00	180	0,99	99,45
[3421,3684)	3552,7	1	0,01	0,55	181	1,00	100,00
Вкупно		181	1,00	100,00			

Хистограм за годишната потрошувачка по жител во галони:



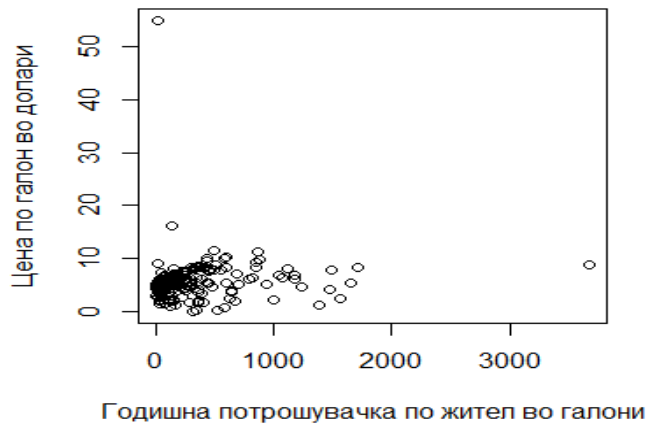
Полигон на кумулативни честоти во проценти:



2. Стебло-лист дијаграм за обележјето “Цена по галон во долари”, каде што лево од “|” е цел број во долари, а десно е децимала, односно во овој случај сите цени се заокружени на една децимала.

```
 0 | 1128
 1 | 12346688
 2 | 00001222344779
 3 | 01244666667899999
 4 | 012333444555666677888888899999
 5 | 00111111222223333344444444455566777899999
 6 | 11112333444455566678889 99
 7 | 0122566678899
 8 | 00112333334444677
 9 | 1358
10 | 002
11 | 44
12 |
15 |
16 | 2
17 |
18 |
51 |
52 |
53 |
54 | 9
```

3. График на расејување на обележјата годишна потрошувачка по жител во галони и цена по галон во долари. Може да забележиме r е приближно 0, па затоа може да се заклучи дека нема поврзаност меѓу овие обележја.



4. Се разгледува обележјето цена по галон во долари.

Просек: 5.7 долари по галон

Медијана: 5.28

Мода: За да се одреди модата ја искористив следната функција:

```
mode = function(){
  return(sort(-table(pricePerGallon))[1])
}
```

Како резултат е добиен 8.27 кое се појавува четири пати.

5. Квартили за обележјето цена по галон во долари.

```
  25%   50%   75%
4.15  5.28  6.76
```

Ранг: Најголемиот елемент е 54.89, а најмалиот 0.08 па рангот ќе биде 54.89 – 0.08 = 54.81.

Интерквартални распон е разликата меѓу третиот и првиот квартил, односно 6.76 – 4.15 = 2.61

6. Дисперзијата на обележјето цена по галон во долари е 19, а стандардната девијација 4.4

7. Во оваа точка е пресметан коефициентот на корелација за обележјата разгледувани во точка 3. Се потврдува тоа што го заклучивме од графикот на расејување, дека скоро и да нема поврзаност меѓу обележјата. Коефициентот на корелација изнесува 0.057.

Б. Втор дел

1. Ќе се разгледува обележјето “цена по литар во долари” и ќе се разгледува параметарот математичко очекување. Целата популација се состои од 193 држави кои се признаени од Обединетите нации. Ние располагаме со информации за 181 држава. Од овие 181 држава со помош на функцијата sample() во R ќе избереме случајно 30 држави што значи дека обемот на нашиот примерок ќе биде 30. За да одредиме интервал на доверба прво треба да земеме во предвид дека имаме голем примерок и

непозната дисперзија. Ова ни дава до знаење дека може да ја искористиме формулата $\overline{X} \pm z_{a/2} \frac{S}{\sqrt{n}}$. Следно треба да ги пресметаме просекот на примерокот и стандардната девијација на примерокот. Во R е добиено дека просекот е 1.541667, а стандардната девијација е 0.960065. Стандардната грешка $\frac{S}{\sqrt{n}}$ изнесува 0.2814685. За 95% интервал на доверба маргиналната грешка во R ја добиваме на следниот начин: MarginalnaGreska = qnorm(.975)*standardnaGreska = 0.5516682.

Според пресметките погоре 95% интервал на доверба за математичкото очекување на обележјето "цена по литар во долари" е интервалот (0.9899985, 2.0933349)

2. Заради поедноставување да претпоставиме дека очекуваната цена по литар е просекот од 181 држава што е многу блиску до очекувањето за целата популација. Тоа би значело дека очекуваната цена по литар за целата популација е 1.505138, а дисперзијата не е позната. Ќе го искористиме истиот примерок од претходната точка за со ниво на значајност 0.05 ги тестираме хипотезите:

H_0: μ_0 = 1.505138

H_a: $\mu_0 \neq$ 1.505138

Бидејќи имаме голем примерок и непозната дисперзија ја користиме формулата:

Z = $\frac{\overline{X} - \mu_0}{S} \sqrt{n}$, каде што $\overline{X}$ (просек на примерок) = 1.541667,

μ_0 (математичко очекување на популацијата) = 1.505138,

S (стандардна девијација на примерокот) = 0.960065

n (обем на примерокот) = 30

По пресметката добиваме дека вредноста за Z тест статистиката е 0.2083974

Бидејќи алтернативната хипотеза е сложена, критичниот домен ќе биде двостран и за ниво на значајност 0.05 ќе биде (-1.96, 1.96).

0.2083974 $\in$ (-1.96, 1.96) -> ја отфрламе нултата хипотеза што значи дека очекуваната цена по литар во долари е различна од 1.505138

3. Ќе тестираме дали обележјето X-цена по литар во долари има нормална распределба.

H_0: X има нормална распределба

H_a: X нема нормална распределба

Бидејќи за секој примерок што е поголем од 30 тежи кон тоа да има нормална распределба во овој случај од ќе избереме случаен примерок од 15 држави.

Користиме Shapiro-Wilk тест за нормална распределба. Со ниво на значајност 0.05 се добива дека p има вредност што е помала од 0.05 ја орфрламе нултата хипотеза и можеме да кажеме дека обележјето X-цена по литар во долари нема нормална распределба.

4. Не е можно да се направи тест за независност бидејќи немаме две категориски обележја.

5. Правиме регресиона анализа на обележјата:

-цена по литар во долари (зависно обележје)

-цена по галон во долари (независно обележје)

Добиената права на регресија е следна: y = 0.0004812 + 0.2641746x.

Коефициентот на детерминираност е 0.9999935 што значи дека има силна линеарна поврзаност. Истото може да се забележи и од графикот на расејување. Оваа вредност беше очекувана со оглед на тоа ако подобро ги анализираме обележјата ќе заклучиме дека цената на горивото за секоја поединечна единка од примерокот е иста и кај двете обележја, но разликата е во мерните единици, кај едното обележје во литри, а кај другото во галони. Така со оваа регресиона анализа и добиената права на регресија добивме еден олеснувачки начин за добивање на цената по литар што е мерна единица која се користи на нашите простори.