

Семинарска работа по предметот Бизнис статистика

- Изборот на податочното множество е на следниот [линк](#)
- Овој сет на податоци содржи информации за 22700 вработени во различни Софтверски компании со различни карактеристики како :
 - **Плати (Salary) во евра (€)** – претставува сумата на пари што му се плаќа на работникот на редовна основа, обично во форма на плата или фиксен годишен приход.
Првично беа претставени во Индиска рупија (₹), но за полесна работа со податоците е претворена во евра.
 - **Име на компанија (Company Name)**
 - **Рејтинг на компанијата (Rating)** – со колкава оцена е оценета компанијата од интервалот [1, 5]
 - **Честота на пријавени плати** - Пријавени плати ги означува информациите или податоците поврзани со платите на вработените во една компанија или индустрија.
 - **Локација на компанијата**
 - **Статус на вработување (Employment Status)** – каков тип на договор има вработениот склучено со компанијата во која работи. Можни се: *Contractor, Full Time, Intern* и *Trainee*.
 - **Работни улоги (Job Title)** - официјалната ознака или позиција што ја има поединец во компанија или организација. Таа ја претставува специфичната улога или одговорностите што му се доделени на лицето во неговиот професионален капацитет.

Изработка на семинарската работа во програмскиот јазик R

Најпрво ја поврзувам дата базата во програмскиот јазик R. Тука се прикажани само првите 6 реда од базата.

```
> #gi pokazuva prвите 6 reda od dataset-ot
> head(rio_csv)
  Rating Company Name Job Title Salary Salaries Reported Location Employment Status Job Roles
1 3.9 KH TEC Test Engineer - Intern 72600 1 Hyderabad Intern Testing
2 3.6 Thapar University Software Development Engineer (SDE) 120000 1 New Delhi FullTime SDE
3 3.5 Koru UX Design Senior Front End Developer 110000 1 Pune FullTime Frontend
4 3.6 OASYS Cybernetics Senior Java Developer 110000 1 Chennai FullTime Java
5 3.8 Concentrix Oracle Database Administrator 110000 1 Bangalore FullTime Database
6 3.7 Nityo Infotech Lead UI Designer, Magento Front-end Developer 108900 1 Bangalore FullTime Frontend
```

За изработка на задачите ги користев податоците од обележјата **Rating, Salary**. Податоците од обележјето Salary се групирани според **Статус на вработување (Employment Status)** и добиени се резултати за секоја група на Employment Status вработени посебно.

А. Прв дел

1. Табела со распределба на честоти. Да се определат средните точки на интервалите, релативните фреквенции и кумулативните фреквенции.

За облевежјето Rating

Табела со распределба на честоти, релативните фреквенции и кумулативните фреквенции.

```
> #spojuvame vo tabelata TABELA ZA FREQ RELFREQ i CUMFREQ
> cbind(rating.freq, rating.relfreq, rating.cumfreq)
      rating.freq rating.relfreq rating.cumfreq
[1,1.5)         74          0.003          74
[1.5,2)          47          0.002         121
[2,2.5)         208          0.009         329
[2.5,3)          547          0.024         876
[3,3.5)        2279          0.100        3155
[3.5,4)        8244          0.362       11399
[4,4.5)       8533          0.375       19932
[4.5,5)       2166          0.095       22098
[5,5.5)         672          0.030       22770
```

Хистограм



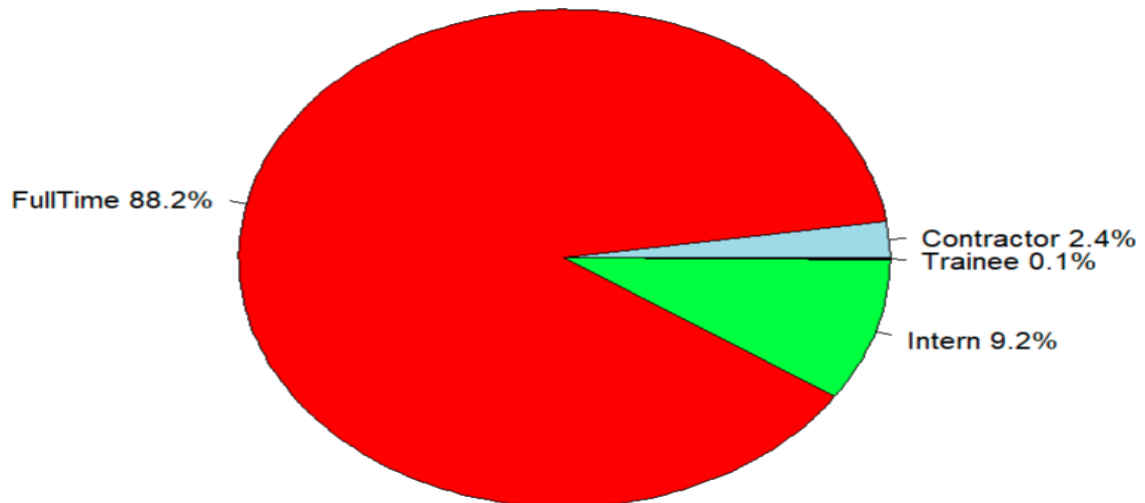
Полигон



За облевежјето *Salary* соодветно за *Employment Status*

Прегледност на застапеноста на различните статуси на вработени на кои ќе се врши анализа на платите.

Фреквенција на застапеност на видот на вработени



Согледуваме дека најзастапени се Fulltime работниците со 88.2%, потоа се Intern работниците или практикантите со 9.2% на застапеност, па работниците кои се Contractors со 2.4% на застапеност и најмалку се застапени работниците кои се Trainee со 0.1% на застапеност.

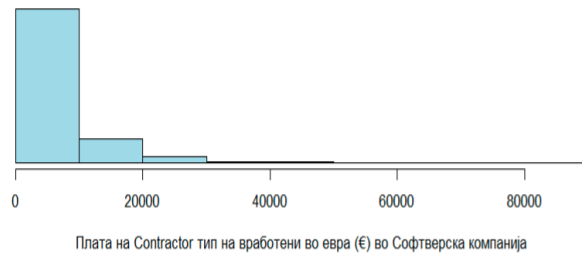
Табела со распределба на честоти за Contractors.

Првата колона ги покажува под интервалите во која е поделена платата на работниците кои се со статус Contractors, а во последната колона фреквенциите на платите во соодветниот под интервал

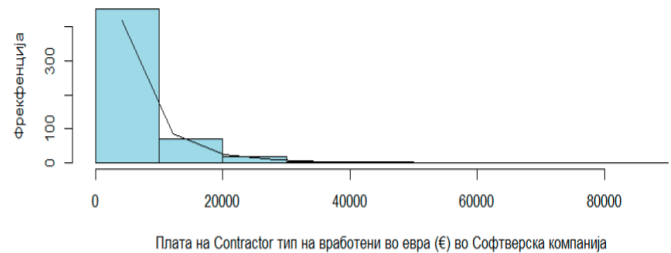
```
> print(table_data)
      Interval      Frequency.Var1 Frequency.Freq
1  [132, 8152)      [132,8.15e+03)           421
2  [8152, 16172) [8.15e+03,1.62e+04)           87
3  [16172, 24192) [1.62e+04,2.42e+04)           25
4  [24192, 32212) [2.42e+04,3.22e+04)            9
5  [32212, 40232) [3.22e+04,4.02e+04)            2
6  [40232, 48252) [4.02e+04,4.83e+04)            2
7  [48252, 56272) [4.83e+04,5.63e+04)            1
8  [56272, 64292) [5.63e+04,6.43e+04)            0
9  [64292, 72312) [6.43e+04,7.23e+04)            0
10 [72312, 80332) [7.23e+04,8.03e+04)            1
> |
```

Хистограм и Полигон на честота за Contractor employees

Хистограм на платите на Contract тип на вработени во Софтверска компанија



Полигон на честоти на платите на Contract тип на вработени во Софтверска компанија

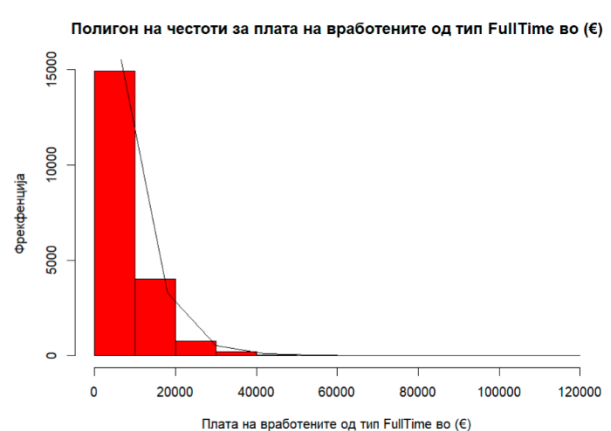
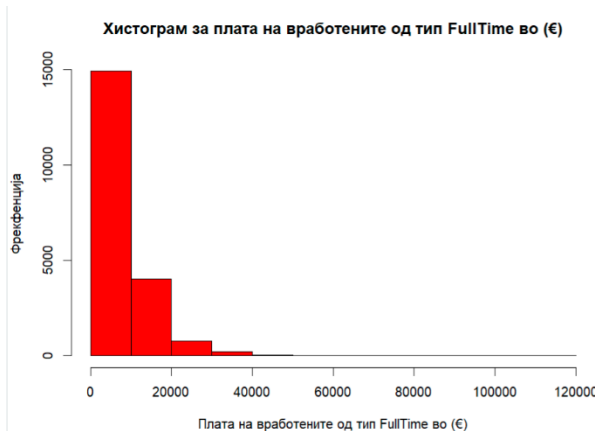


Табела со распределба на честоти за FullTime.

Првата колона ги покажува под интервалите во која е поделена платата на работниците кои се со статус Fulltime, а во последната колона фреквенциите на платите во соодветниот под интервал.

```
> print(table_data)
      Interval      Frequency.Var1 Frequency.Freq
1      [90, 12081)      [90,1.21e+04)      16074
2     [12081, 24072) [1.21e+04,2.41e+04)      3300
3     [24072, 36063) [2.41e+04,3.61e+04)       524
4     [36063, 48054) [3.61e+04,4.81e+04)       116
5     [48054, 60045) [4.81e+04,6e+04)         35
6     [60045, 72036) [6e+04,7.2e+04)          9
7     [72036, 84027) [7.2e+04,8.4e+04)          5
8     [84027, 96018) [8.4e+04,9.6e+04)          9
9     [96018, 108009) [9.6e+04,1.08e+05)          5
10    [108009, 120000) [1.08e+05,1.2e+05)          5
> |
```

Хистограм и Полигон на честота за FullTime employees



Табела со распределба на честоти за Intern.

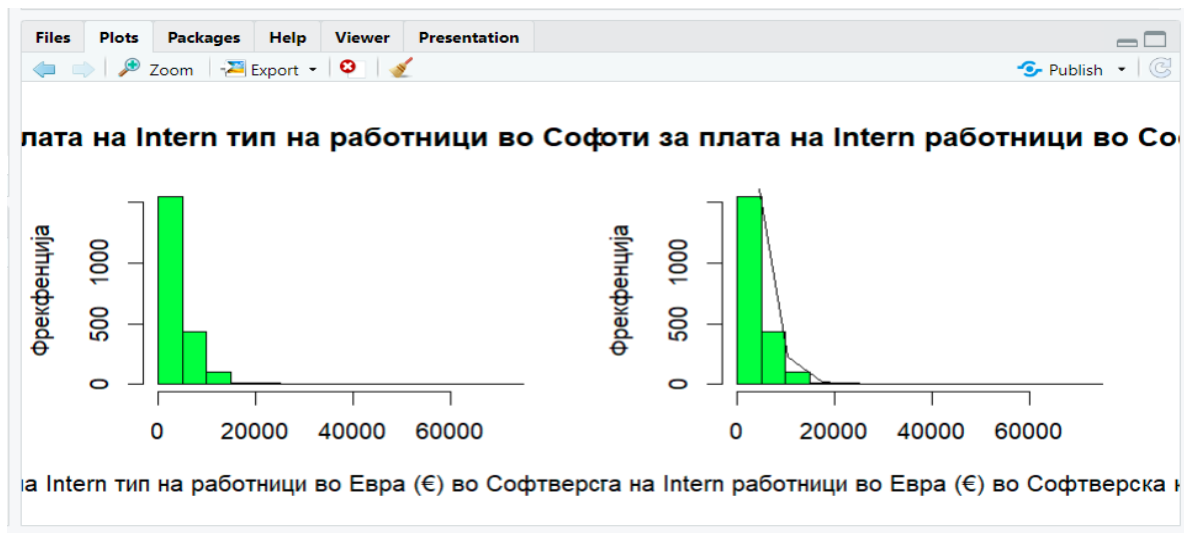
Првата колона ги покажува под интервалите во која е поделена платата на работниците кои се со статус Intern, а во последната колона фреквенциите на платите во соодветниот под интервал.

```
> #tabela na cestoti za intern salary
> print(table_data)
```

	Interval	Frequency.Var1	Frequency.Freq
1	[20, 7020)	[20,7.02e+03)	1854
2	[7020, 14020)	[7.02e+03,1.4e+04)	222
3	[14020, 21020)	[1.4e+04,2.1e+04)	19
4	[21020, 28020)	[2.1e+04,2.8e+04)	6
5	[28020, 35020)	[2.8e+04,3.5e+04)	1
6	[35020, 42020)	[3.5e+04,4.2e+04)	1
7	[42020, 49020)	[4.2e+04,4.9e+04)	1
8	[49020, 56020)	[4.9e+04,5.6e+04)	1
9	[56020, 63020)	[5.6e+04,6.3e+04)	0
10	[63020, 70020)	[6.3e+04,7e+04)	0
11	[70020, 77020)	[7e+04,7.7e+04)	1

```
> hist(ris.csv$Salary, ris.csv$`Employment Status`)
```

Хистограм и Полигон на честота за Intern employees



Табела со распределба на честоти за Trainee.

Првата колона ги покажува под интервалите во која е поделена платата на работниците кои се со статус Trainee, а во последната колона фреквенциите на платите во соодветниот под интервал.

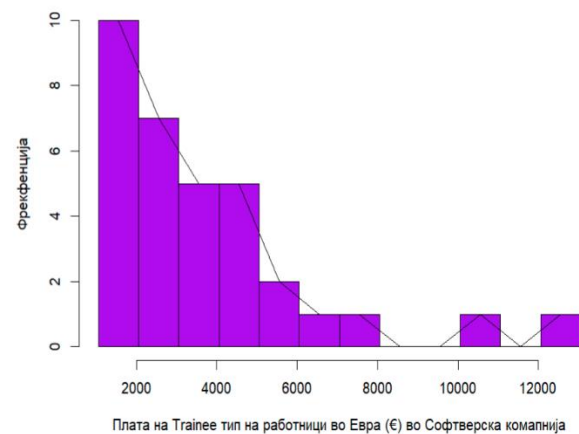
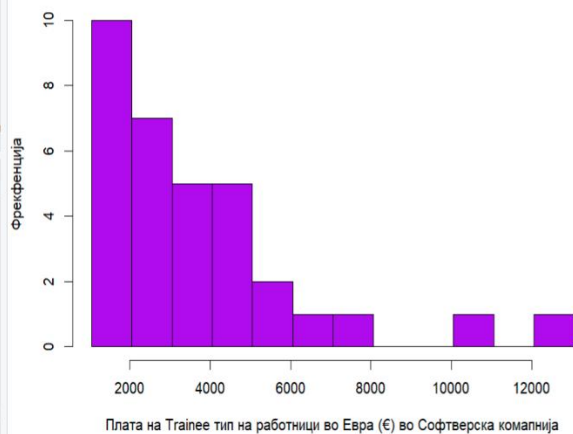
```

+ )
> print(table_data)
      Interval      Frequency.Var1 Frequency.Freq
1    [1056, 2056) [1.06e+03,2.06e+03)          10
2    [2056, 3056) [2.06e+03,3.06e+03)           7
3    [3056, 4056) [3.06e+03,4.06e+03)           5
4    [4056, 5056) [4.06e+03,5.06e+03)           5
5    [5056, 6056) [5.06e+03,6.06e+03)           2
6    [6056, 7056) [6.06e+03,7.06e+03)           1
7    [7056, 8056) [7.06e+03,8.06e+03)           1
8    [8056, 9056) [8.06e+03,9.06e+03)           0
9    [9056, 10056) [9.06e+03,1.01e+04)           0
10   [10056, 11056) [1.01e+04,1.11e+04)           1
11   [11056, 12056) [1.11e+04,1.21e+04)           0
12   [12056, 13056) [1.21e+04,1.31e+04)           1
>

```

Хистограм и Полигон на честота за Intern employees

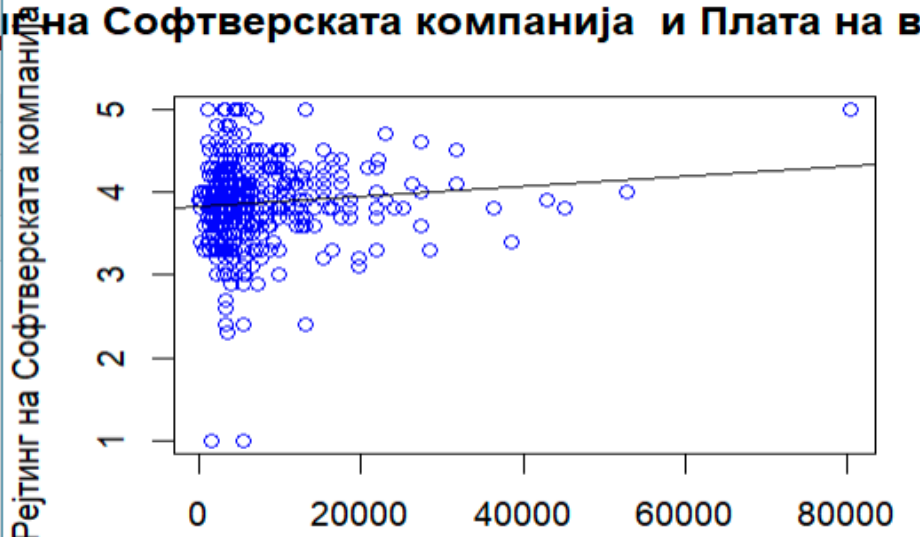
Хистограм на платите на вработените кои се Trainee во Софтверска компанија Полигон на платите на вработените кои се Trainee во Софтверска компанија



3a Trainee

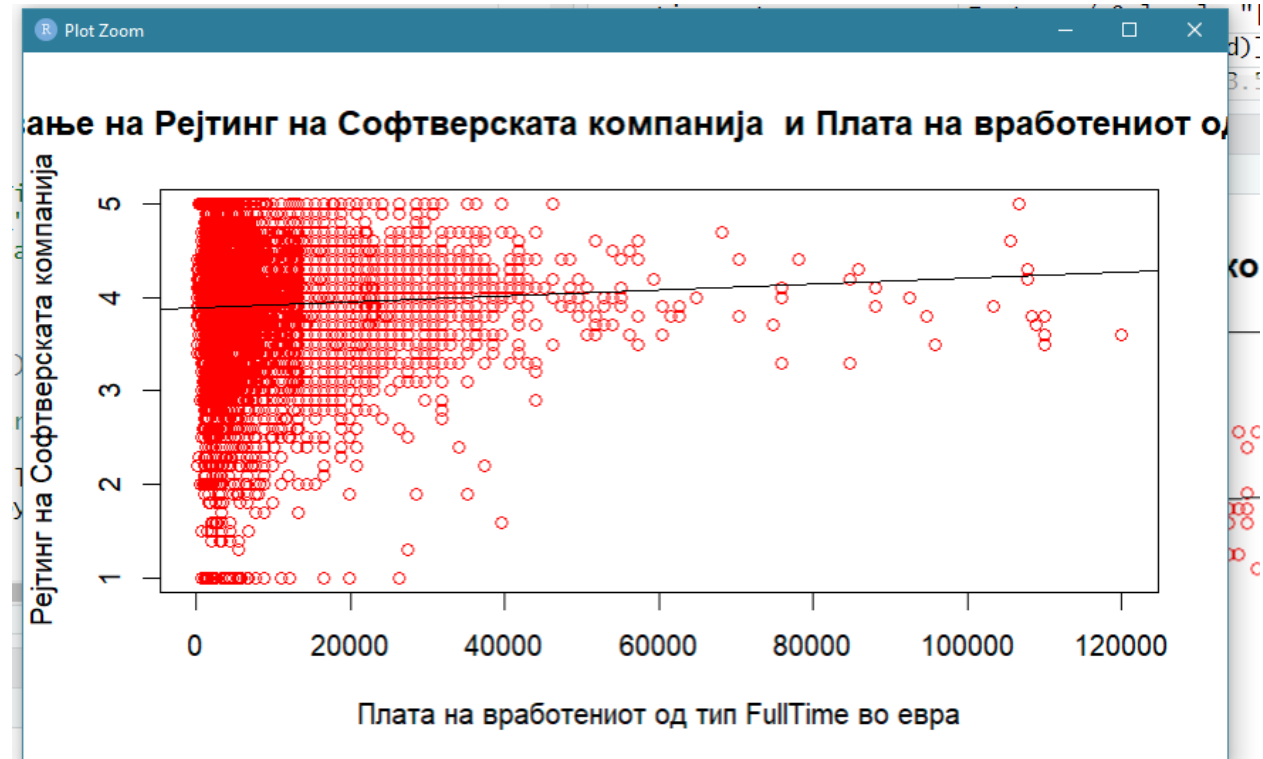
3. Графика на расејување

Регресија на Софтверската компанија и Плата на вработениот од тип Contractor во евра

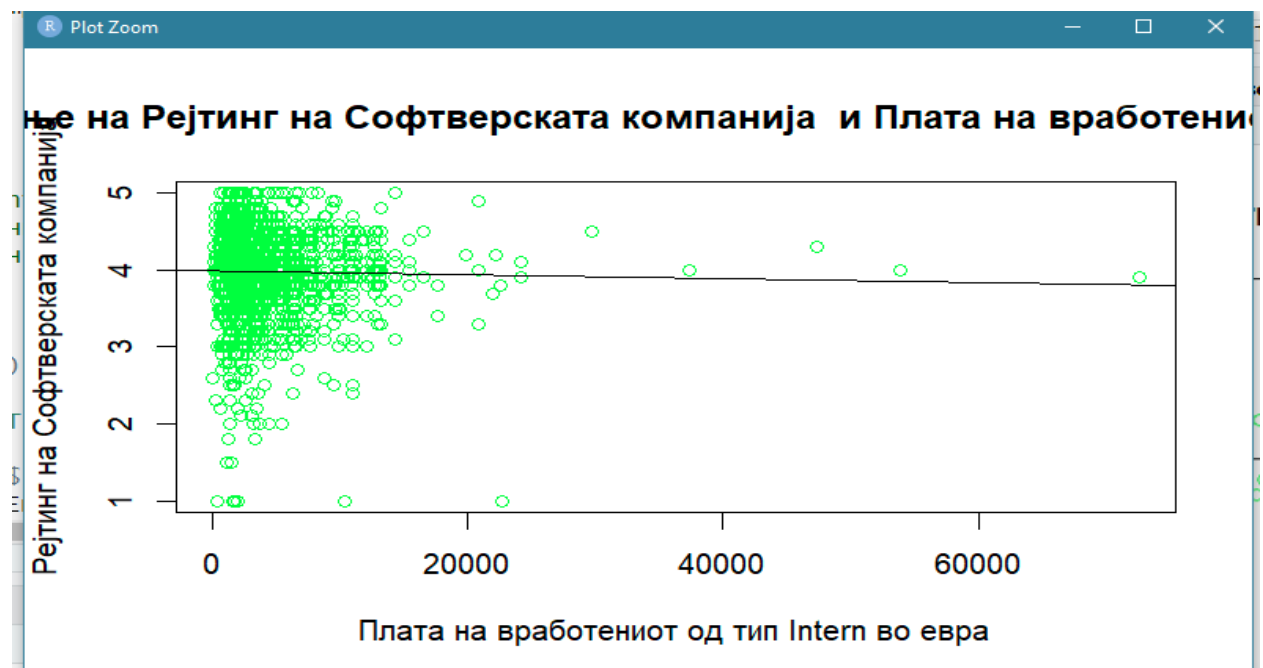


Регресија на Софтверската компанија и Плата на вработениот од тип Contractor во евра

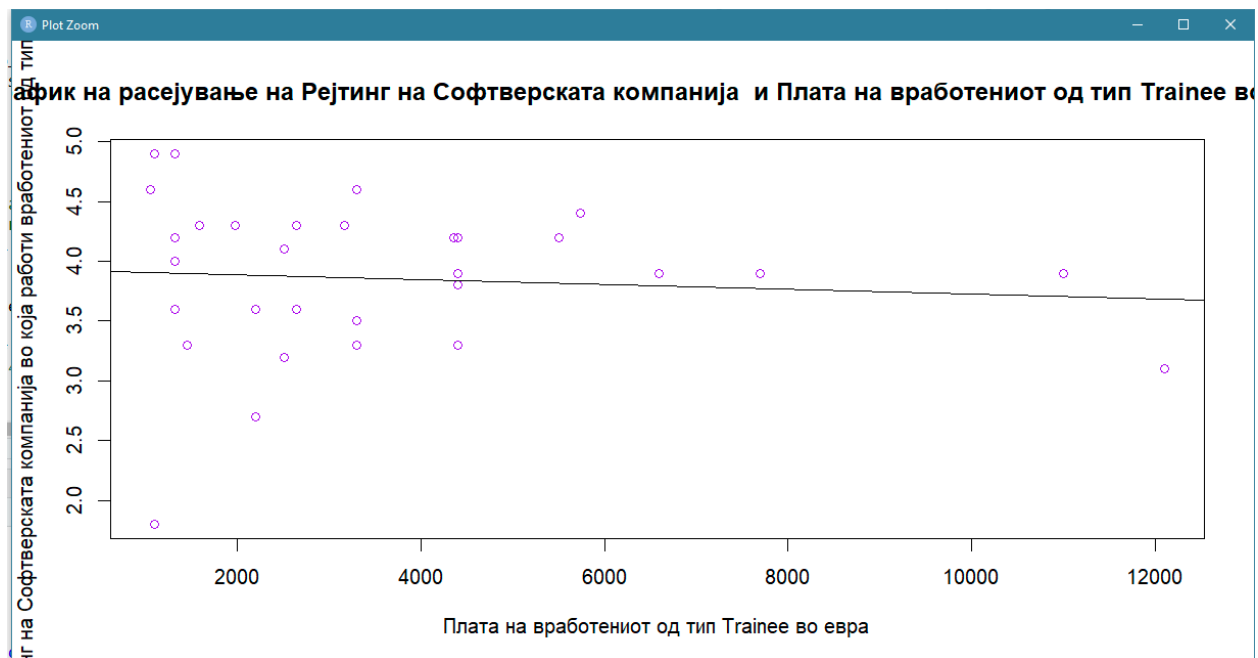
За FullTime



За Intern



Za Trainee



4. Определување на мода, медијана и просек на податоците

ПРОСЕК

```
> ## * PROSEK *
> #PROSEK ZA RATING
> mean(rating)
[1] 4
> #PROSEK za plata na COUNTRACTOR employees
> mean(salary.contractor)
[1] 6901
> #PROSEK za plata na FullTime employees
> mean(full_time_salaries)
[1] 8023
> #PROSEK za plata na Intern employees
> mean(intern_salaries)
[1] 3928
> #PROSEK za plata na Trainee employees
> mean(trainee_salaries)
[1] 3567
> |
```

МЕДИЈАНА

```
R 4.3.1 · D:/Seminarska_BS/ ↗
> ## * MEDIJANA *
> #MEDIJANA ZA RATING
> median(rating)
[1] 4
> #MEDIJANA za plata na COUNTRACTOR employees
> median(salary.contractor)
[1] 5082
> #MEDIJANA za plata na FullTime employees
> median(full_time_salaries)
[1] 5500
> #MEDIJANA za plata na Intern employees
> median(intern_salaries)
[1] 3036
> #MEDIJANA za plata na Trainee employees
> median(trainee_salaries)
[1] 2640
> |
```

МОДА

Мода за ратинг

```
> print(most_frequent_rating)
[1] 4
```

Мода за плата на Контрактори

```
> print(most_frequent_contractor_salary)
[1] 3300
```

Мода за плата на FullTime вработени

```
> print(most_frequent_fulltime_salary)
[1] 3300
```

Мода за платата на Intern вработени

```
> print(most_frequent_intern_salary)
[1] 1320
```

Мода за плата на Trainee вработени

```
> print(most_frequent_trainee_salary)
```

```
[1] 1320
```

5. Кварталите, опсегот и интеркварталниот распон на податоците

КВАРТИЛИ

Rating

```
quantile(rating)
```

0%	25%	50%	75%	100%
1	4	4	4	5

Contractor employee salary

```
quantile(salary.contractor)
```

0%	25%	50%	75%	100%
132	3300	5082	7700	80300

FullTime employee salary

```
quantile(full_time_salaries)
```

0%	25%	50%	75%	100%
93	3300	5500	10956	120000

Intern employee salary

```
quantile(intern_salaries)
```

0%	25%	50%	75%	100%
23	1584	3036	5016	72600

Trainee employee salary

```
quantile(trainee_salaries)
```

0%	25%	50%	75%	100%
1056	1584	2640	4400	12100

ОПСЕГ

Rating

```
> max(rating) - min(rating)
```

```
[1] 4
```

Contractor employee salary

```
> max(salary.contractor) - min(salary.contractor)
```

```
[1] 80168
```

FullTime employee salary

```
> max(full_time_salaries) - min(full_time_salaries)
```

```
[1] 119907
```

Intern employee salary

```
> max(intern_salaries) - min(intern_salaries)
[1] 72577
```

Trainee employee salary

```
> max(trainee_salaries) - min(trainee_salaries)
[1] 11044
```

ИНТЕРКВАРТИЛЕН РАСПОН

Rating

```
IQR(rating)
[1] 0.5
```

Contractor employee salary

```
> IQR(salary.contractor)
[1] 4400
```

FullTime employee salary

```
> IQR(full_time_salaries)
[1] 7656
```

Intern employee salary

```
> IQR(intern_salaries)
[1] 3432
```

Trainee employee salary

```
> IQR(trainee_salaries)
[1] 2816
```

6. Дисперзија и стандардна девијација

Rating: Дисперзија= 0.3 ; Стандардна девијација = 0.5

Contractor employee salary: Дисперзија= 49225887; Стандардна девијација = 7016

FullTime employee salary: Дисперзија= 55285443; Стандардна девијација = 7435

Intern employee salary: Дисперзија= 15331493; Стандардна девијација = 3916

Trainee employee salary: Дисперзија= 7033431; Стандардна девијација = 2652

7. Коефициент на корелација

Contractor employee salary: Коефициент на корелациј = 0.1

FullTime employee salary: Коефициент на корелација = 0.05

Intern employee salary: Коефициент на корелациј = -0.02

Trainee employee salary: Коефициент на корелациј = -0.08

Б. Втор дел

1. Интервал на доверба за просечната платата на Contractor вид на вработен за 99% ниво на доверба. Ја содржи точната но непозната вредност за просечната плата.

```
> #golemina na primerokot n>=30
> length(salary.contractor)
[1] 548
> # 1) za CONTRACTOR employee salary prosek
> mean(salary.contractor)
[1] 6901
> # 2.1 marginata na greshka i procenka na intervalot so nivo na doverba o
d 99%
> n = length(salary.contractor)
> # presmetka na S (standardna devijacija na primerokot)
> S = sd(salary.contractor)
> print(S)
[1] 7016
> #standardna procenka na greshka
> SE = S/sqrt(n); SE
[1] 300
> #margina na greshka (greshka na primerokot)
> alpha <- 0.01 # 1 -  $\alpha$  = 0.99
> alpha_2 = alpha / 2 #0.005
> #so 3 decimiali
> print(alpha_2) # 0.005

[1] 0.02
> z_alpha_2 <- qnorm(1 - alpha_2) #qnorm(0.995) = 2.58
> print(z_alpha_2) # dobivam 2.58
[1] 2
> E = qnorm(1 - alpha_2) * SE; E #dobivam 772
[1] 772
> # prosek na primerokot
> xbar = mean(salary.contractor)
> #intervalot na doverba
> xbar + c(-E, E)

[1] 6129 7673
```

2. Тестирање на хипотези

CONTRACTOR employee salary

Во примерок од 548 случајно избрани вработени со статус на вработување Contractor во Софтверска компанија покажува дека просечната плата е $\bar{x} = 6901$ и стандардната девијација е $S = 7016$. Сакаме со ниво на значајност од 0.01 да откриеме дали просечната годишна плата би можела да бие 12 000€. Се тестира хипотезата дека просечната плата на вработени со статус на

вработување Contractor во Софтверска компанија е 12 000€, а спротивната е дека таа не би можела да биде толку.

Ги иницијализирам вредностите за просек и дисперзија на примерок, за обемот на примерокот, за нивото на значајност и за претпоставената вредност од хипотезите која ја тестираме.

```
x_crt Contractor_salary = mean(salary.contractor)
> x_crt Contractor_salary
[1] 6901
> s_contractor_salary = sd(salary.contractor)
> s_contractor_salary
[1] 7016
> n_salary_contractor = length(salary.contractor)
> n_salary_contractor
[1] 548
> alpha_contractor_salary <- 0.01
> alpha_contractor_salary
[1] 0.01
> mi0 <- 12000
```

Следуваат чекорите за тестирање на хипотези:

1) Нека обележјето X – просечна плата на вработени во Софтверски компании со статус на вработување Contractor и нека μ е математичкото очекување на оваа случајна променлива.

2) Дефинирање на Хипотезите:

Но: $\mu = \mu_0$

На: $\mu \neq \mu_0$

Но: $\mu = 12\,000$

На: $\mu \neq 12\,000$

3) Тест статистиката

$n \geq 30$ -> обемот на примерокот е поголем од 30 и изнесува 548

$S = 7016$ -> позната ни е само стандардната девијација на примерокот

$\bar{x} = 6901$

$$Z_0 = \frac{\bar{x} - \mu_0}{S} \sqrt{n}$$

```
> # Z0 = ((x_crt - mi0) / s) * sqrt(548) = -0.727 * 23.409 = -17.018
> options(digits = 3) # zaokružuva na 3 decimali
> dropka = ((x_crt Contractor_salary - mi0) / s_contractor_salary) ; dropka
[1] -0.727
> koren = sqrt(n_salary_contractor); koren
[1] 23.4
> Z0 = dropka * koren; Z0 # Z0 = -17.018
[1] -17
```

$Z_0 = -17.018$

4) Од таблица за $N(0,1)$ распределба (нормална нормирана) , се чита $z_{\alpha/2} = z_{0.005}$.

$\Phi(z_{0.005}) = 1 - \alpha/2 = 0.995$.

Наоѓаме $z_{0.005} = 2.58$

```

> # 4) Od tablicata gledame
> # z_alfa_polovina = z_0.01/2 = z_0.005
> z_alfa_polovina = alpha_contractor_salary / 2; z_alfa_polovina #0.005
[1] 0.005
> #  $\Phi(z_{0.005}) = 1 - 0.005 = 0.995$ 
> fi_z_alfa_polovina = 1 - 0.005; fi_z_alfa_polovina
[1] 0.995
> #vrednost na z_alfa_polovina ja baramе
> vrednost_z_alfa_polovina = qnorm(fi_z_alfa_polovina); vrednost_z_alfa_polovina
[1] 2.58

```

5) Критичен домен – област на отфрлање на нултата хипотеза

$$C = (-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, +\infty)$$

$$C = (-\infty, -2.58) \cup (2.58, +\infty)$$

6) Заклучок:

- Вредноста на тест статистиката $Z_0 = -17.018 \in C$ (критичниот домен).
- Занчи дека H_0 (нултата хипотеза) се **ОТФРЛА**, таа е не точна.
- На (алтернативната хипотеза) е точна.

Претпоставката дека просечната плата на вработени во Софтверски компании со статус на вработување Contractor не би можела да биде 12 000€ годишно.

4. Тестираат хипотези за независност. Во случај тоа да не е можно, да се образложи зошто не е можно.

Во примерок од 548 случајно избрани вработени со статус на вработување Contractor покажува дека просечната плата е $\bar{y} = 6901$ и стандардната девијација е $S_y = 7016$. Примерок од 20083 случајно избрани вработени во Софтверска компанија со статус на вработување FullTime покажува дека просечната плата на овие вработени е $\bar{x} = 8023$ и стандардната девијација е $S_x = 7435$.

Да докажеме дека FullTime вработените имаат поголема просечна плата од Contractor вработените во Софтверските компании, наспроти тврдењето дека земаат помалку.

Со ниво на значајност $\alpha = 0.1$ да се тестира хипотезата дека Fulltime вработените имаат поголема просечна плата од Contractor вработените во Софтверските компании, наспроти хипотезата дека имаат помалку.

1) Да ги поставиме обележјата кои ги разгледуваме:

X – просечна плата на Fulltime employees во Софтверски компании.

Y – просечна плата на Contractor вработените во Софтверските компании.

2) Поставување на хипотези:

Но: просечната плата на FullTime employees е **поголема** од просечна плата од Contractor вработените во Софтверските компании.

На: просечната плата на FullTime employees е **помала** од просечна плата од Contractor вработените во Софтверските компании.

$$H_0: \mu_x > \mu_y$$

Ha: $\mu_x < \mu_y$

- 3) Бидејќи дисперзиите на обележјата не се познати, но имаме големи примероци се користи следната тест статистика:

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{S_x^2}{n}\right) + \left(\frac{S_y^2}{m}\right)}}$$

- 4) Да ги поставиме вредностите:

$$\bar{X} = 8023$$

$$\bar{Y} = 6901$$

$$S_x = 7435$$

$$S_x^2 = 55279225$$

$$S_y = 7016$$

$$S_y^2 = 49224256$$

$$n=20083$$

$$m=548$$

```
> razlika_proseci = x_cрта_salary_fulltime - y_cрта_salary_contractor; razli
ka_proseci
[1] 1122
> disperzija_fulltime_salaries = var(full_time_salaries); disperzija_fulltime
_salaries
[1] 55285443
> disperzija_contractor_salary = var(salary.contractor); disperzija_contracto
r_salary
[1] 49225887
> pod_koren_sobirok1 = disperzija_fulltime_salaries / n_fulltime; pod_koren_s
obirok1
[1] 2753
> pod_koren_sobirok2 = disperzija_contractor_salary / m_contractors; pod_kore
n_sobirok2
[1] 89828
> koren = sqrt(pod_koren_sobirok1 + pod_koren_sobirok2); koren
[1] 304
> z0 = razlika_proseci / koren; z0
[1] 3.69
```

Се добива $Z_0 = 3.69$

- 5) Од таблица за нормална нормирана распределба гледаме:

$$z_\alpha = z_{0.1} . \text{ Така што } \Phi(z_{0.1}) = 1 - 0.1 = 0.9$$

$$\text{Наоѓаме } z_{0.1} = 1.28$$

- 6) Критичен домен

Критичниот домен зависи од алтернативната хипотеза и се чита од таблица на нормална нормирана распределба во случајов ова е обликот на критичниот домен $C = (-\infty, -z_\alpha)$.

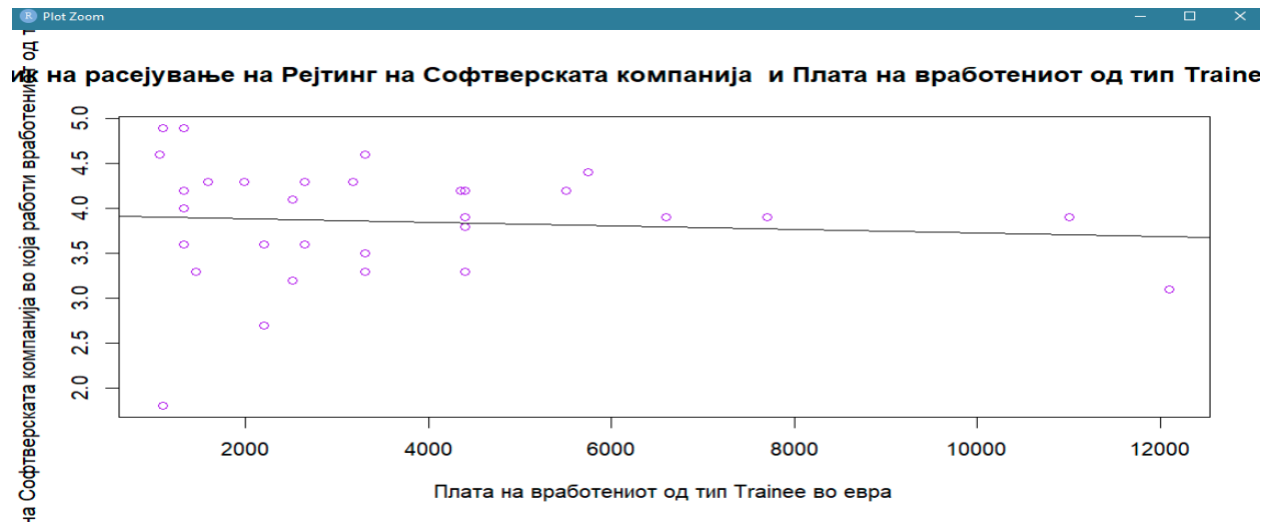
$$Z_0 = 3.69$$

$$C = (-\infty, -Z_\alpha) = (-\infty, 3.69)$$

- 7) Заклучок: $Z_0 = 3.69 \notin (-\infty, -Z_\alpha) = (-\infty, 1.28)$, па оттука заклучуваме дека нултата хипотеза се прифаќа.

Со ниво на значајност $\alpha = 0.1$ заклучуваме дека просечната плата на Fulltime е **поголема** од просечна плата од Contractor вработените во Софтверските компании.

5. Регресиона анализа



Определување на правата на регресија која ја изразува зависноста на платата на вработените од тип Trainee во Софтверска компанија (x-оска) и оценката која ја има компанијата во која е вработен (y-оска).

Обликот на *regresionata prava* е: $y = \beta_0 + \beta_1 x$

Коефициент на корелација (r) = -0.0842 , **покажува кон негативна асоцијација помеѓу променливите. Вредноста на r е блиску до 0 зборуваат за слаба линеарна поврзаност.**

- Ги определуваме потребните вредности за определување на правата на регресија:
 - $SSx = 225069797$
 - $SSy = 13.1$
 - $SSxy = -4574$
 - $\beta_1 = \frac{SSxy}{SSx} = -0.0000203$
 - $\beta_0 = \bar{y} - \frac{SSxy}{SSx} \bar{x} = 3.93$

Облик на правата на регресија: $y = 3.93 - 0.0000203x$

Интерпретација на резултатите:

$y_i - \bar{y}$ = вкупно отстапување на податокот од примерокот

$\hat{y}_i - \bar{y}$ = објаснето отстапување (или отстапување што се должи на моделот) и покажува за колку се намалува вкупното отстапување кога ќе се постави регресионата права на податоците.

$y_i - \hat{y}$ = необјаснето отстапување, односно дел од вкупното варирање кој не е објаснет со воведувањето на регресионата права.

SST = вкупна сума на квадрати = 13.1. Го мери варирањето на y_i околу нивниот просек.

SSE = сума на квадрати на грешки = 13. Варирање што се должи на други причини надвор од релацијата меѓу x и y .

SSR = сума на квадрати на регресија = 0.0929. Објаснето варирање што се должи на линеарната врска на x и y .

Коефициент на детерминираност (R^2) = ја мери јачината на совпаѓањето на правата на регресија со податоците.

$$R^2 = 0.00708$$

Слаба е јачината на совпаѓање на правата на регресија со податоците.