



Институт за математику и информатику  
Природно-математички факултет  
Универзитет у Крагујевцу

Семинарски рад

---

Представљање и тумачење скупа података  
**„HR Analytics“**

---

Ментор:  
др Бранко Арсић

Студенти:  
Јован Радовановић 85/2018  
Немања Тракић 130/2018

Новембар 2024

1.	Увод .....	4
2.	Припрема података .....	5
2.1	Преглед података.....	7
2.2	Уклањање дупликата и неважних колона.....	31
2.3	Исправљање структурних грешака .....	32
2.4	Уклањање екстремних вредности .....	32
2.5	Руковање недостајућим вредностима.....	33
2.6	Креирање нових променљивих .....	33
	DistanceFromHomeGroup.....	33
3.	Анализа података .....	34
4.	Креирање модела .....	49
4.1	Логистичка регресија.....	50
4.1.1	Логистички модел и унакрсна валидација .....	51
4.2	GLM модел.....	52
4.2.1	GLM модел и унакрсна валидација .....	54
4.3	Random Forest модел .....	55
4.3.1	Random Forest модел и унакрсна валидација .....	56
4.4	Неуронске мреже.....	57
4.5	Резултати модела .....	59
4.6	Resampling методе .....	61
4.6.1	Oversampling.....	61
4.6.2	Undersampling метода.....	62
4.6.3	Oversampling и Undersampling .....	62
4.6.4	Резултати .....	62
5.	Закључак.....	63
6.	Литература.....	64



# 1. Увод

Предвиђање напуштања запослених омогућава фирмама да на време препознају факторе који воде ка томе и омогућава им деловање у циљу спречавања одласка запослених. Ово може укључивати побољшање услова за рад, пружање додатне обуке, реорганизацију, проналазак улога који више одговарају свакоме појединачно, као и друге мере за повећање задовољства унутар организације.

Основни циљ ове анализе је да се идентификују кључни фактори који утичу на одлазак запослених. Као што су старосна доб, удаљеност од посла, задовољство на послу, учесталост пословних путовања и сл.

**„Attrition“** (у преводу осипање) или одлазак запослених подразумева када запослени својеволјно или присилно напушта компанију. Висока стопа може довести до значајних трошкова за организацију, како новчаних тако и временских. Након одласка потребно је пронаћи, обучити и прилагодити нове запослене.

На крају, прецизно предвиђање омогућава запосленима у Људским ресурсима да доносе информисане и стратешке одлуке које директно и позитивно утичу на задовољство запослених.

Скуп података можете погледати [овде \(HR Analytics Dataset \(kaggle.com\)\)](https://www.kaggle.com/datasets/hr-analytics/hr-analytics-dataset).

## 2. Припрема података

Скуп података **HR\_Analytics** обухвата 38 колона и садржи 1480 записа који пружају детаљан увид у различите аспекте живота запослених. Овај скуп података покрива како професионалне аспекте, попут задовољства послом, учешћа у обуци, и могућности за унапређење, тако и личне факторе као што су породични статус, баланс између пословног и приватног живота и задовољство радним окружењем. На тај начин, скуп података пружа комплетну слику о факторима који утичу на добробит и ангажованост запослених у компанији.

EmpID Length:1480 Class :character Mode :character	Age Min. :18.00 1st Qu.:30.00 Median :36.00 Mean :36.92 3rd Qu.:43.00 Max. :60.00	AgeGroup Length:1480 Class :character Mode :character	Attrition Length:1480 Class :character Mode :character	BusinessTravel Length:1480 Class :character Mode :character	DailyRate Min. : 102.0 1st Qu.: 465.0 Median : 800.0 Mean : 801.4 3rd Qu.:1157.0 Max. :1499.0	Department Length:1480 Class :character Mode :character
DistanceFromHome Min. : 1.00 1st Qu.: 2.00 Median : 7.00 Mean : 9.22 3rd Qu.:14.00 Max. :29.00	Education Min. :1.000 1st Qu.:2.000 Median :3.000 Mean :2.911 3rd Qu.:4.000 Max. :5.000	EducationField Length:1480 Class :character Mode :character	EmployeeCount Min. :1 1st Qu.:1 Median :1 Mean :1 3rd Qu.:1 Max. :1	EmployeeNumber Min. : 1.0 1st Qu.: 493.8 Median :1027.5 Mean :1031.9 3rd Qu.:1568.2 Max. :2068.0	EnvironmentSatisfaction Min. :1.000 1st Qu.:2.000 Median :3.000 Mean :2.724 3rd Qu.:4.000 Max. :4.000	Gender Length:1480 Class :character Mode :character
HourlyRate Min. : 30.00 1st Qu.: 48.00 Median : 66.00 Mean : 65.85 3rd Qu.: 83.00 Max. :100.00	JobInvolvement Min. :1.00 1st Qu.:2.00 Median :3.00 Mean :2.73 3rd Qu.:3.00 Max. :4.00	JobLevel Min. :1.000 1st Qu.:1.000 Median :2.000 Mean :2.065 3rd Qu.:3.000 Max. :5.000	JobRole Length:1480 Class :character Mode :character	JobSatisfaction Min. :1.000 1st Qu.:2.000 Median :3.000 Mean :2.725 3rd Qu.:4.000 Max. :4.000	MaritalStatus Length:1480 Class :character Mode :character	MonthlyIncome Min. : 1009 1st Qu.: 2922 Median : 4933 Mean : 6505 3rd Qu.: 8384 Max. :19999
SalarySlab Length:1480 Class :character Mode :character	MonthlyRate Min. : 2094 1st Qu.: 8051 Median :14220 Mean :14298 3rd Qu.:20461 Max. :26999	NumCompaniesWorked Min. :0.000 1st Qu.:1.000 Median :2.000 Mean :2.687 3rd Qu.:4.000 Max. :9.000	Over18 Length:1480 Class :character Mode :character	OverTime Length:1480 Class :character Mode :character	PercentSalaryHike Min. :11.00 1st Qu.:12.00 Median :14.00 Mean :15.21 3rd Qu.:18.00 Max. :25.00	PerformanceRating Min. :3.000 1st Qu.:3.000 Median :3.000 Mean :3.153 3rd Qu.:3.000 Max. :4.000
RelationshipSatisfaction Min. :1.000 1st Qu.:2.000 Median :3.000 Mean :2.709 3rd Qu.:4.000 Max. :4.000	StandardHours Min. :80 1st Qu.:80 Median :80 Mean :80 3rd Qu.:80 Max. :80	StockOptionLevel Min. :0.0000 1st Qu.:0.0000 Median :1.0000 Mean :0.7919 3rd Qu.:1.0000 Max. :3.0000	TotalWorkingYears Min. : 0.00 1st Qu.: 6.00 Median :10.00 Mean :11.28 3rd Qu.:15.00 Max. :40.00	TrainingTimesLastYear Min. :0.000 1st Qu.:2.000 Median :3.000 Mean :2.798 3rd Qu.:3.000 Max. :6.000	WorkLifeBalance Min. :1.000 1st Qu.:2.000 Median :3.000 Mean :2.761 3rd Qu.:3.000 Max. :4.000	YearsAtCompany Min. : 0.000 1st Qu.: 3.000 Median : 5.000 Mean : 7.009 3rd Qu.: 9.000 Max. :40.000
YearsInCurrentRole Min. : 0.000 1st Qu.: 2.000 Median : 3.000 Mean : 4.228 3rd Qu.: 7.000 Max. :18.000	YearsSinceLastPromotion Min. : 0.000 1st Qu.: 0.000 Median : 1.000 Mean : 2.182 3rd Qu.: 3.000 Max. :15.000	YearsWithCurrManager Min. : 0.000 1st Qu.: 2.000 Median : 3.000 Mean : 4.118 3rd Qu.: 7.000 Max. :17.000 NA's :57				

За анализу података креираћемо две функције („*plot\_histogram*“, „*plot\_bar*“ и „*plot\_boxplot*“) помоћу којих ћемо добити графички приказ поделе података по колонама:

```
plot_histogram <- function(data, column, binwidth=10) {  
  ggplot(data, aes_string(x = column)) +  
    geom_histogram(binwidth = binwidth, fill = "steelblue", color = "black", alpha = 0.7) +  
    labs(title = paste("Raspodela za", column), x = column, y = "Broj zaposlenih") +  
    theme_minimal()  
}  
  
plot_bar <- function(data, column) {  
  ggplot(data, aes_string(x = column)) +  
    geom_bar(fill = "steelblue", color = "black", alpha = 0.7) +  
    labs(title = paste("Raspodela za", column), x = column, y = "Broj zaposlenih") +  
    theme_minimal()  
}  
  
plot_boxplot <- function(data, column, name) {  
  data %>% ggplot() +  
    geom_boxplot(aes(y=column)) +  
    labs(y = name)  
}
```

За категоријских променљивих користићемо *plot\_bar*, док за приказ нумеричких променљивих користимо *plot\_histogram*. Уколико буде потребе, увешћемо додатне начине приказа расподеле како бисмо лакше увидели недостатке.

## 2.1 Преглед података

### Колона „EmpId“

Представља јединствени идентификатор запосленог. Уз помоћ следеће функције можемо проверити колико заиста јединствених вредности постоји унутар ове колоне.

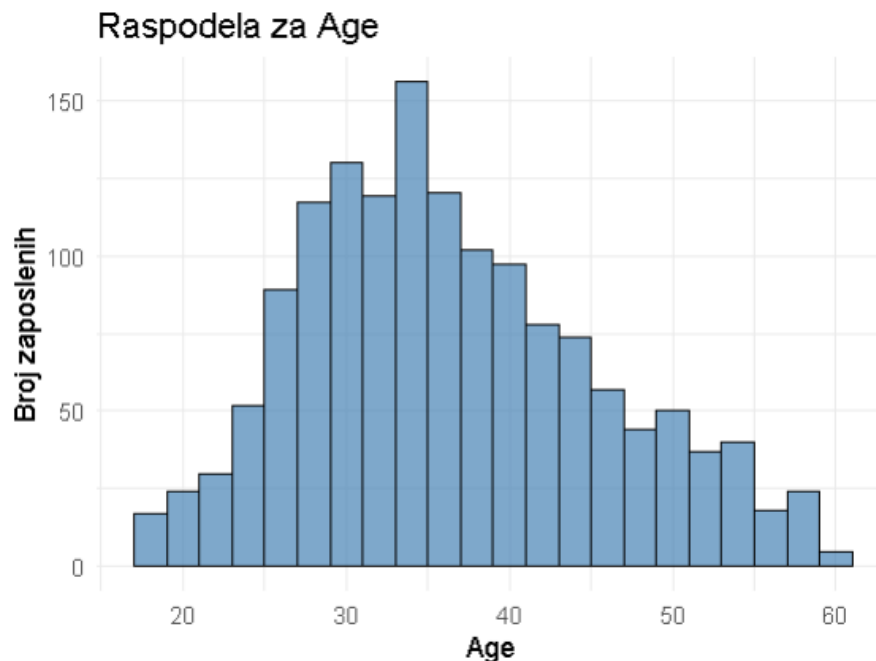
```
unique_values_summary <- hr_data %>%  
  summarise(across(everything(),  
    list(unique_values = ~ length(unique(.)),  
    sample_values = ~ paste(unique(.)[1:15], collapse = ", "))))  
  
print(unique_values_summary)  
  
> print(unique_values_summary$EmpID_unique_values)  
[1] 1470
```

Видимо да постоји 1470 јединствених вредности док у се скупу података налази 1480 података. Закључујемо да постоје дупликати. Дупликате и ирелевантне податке ћемо исправити у следећим корацима.

### Колона „Age“

Представља године запосленог. Најмлађи запослени има 18 година док најстарији 60 година. Средња вредност износи 36,92 године.

```
> summary(hr_data$Age)  
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   
  18.00   30.00   36.00   36.92   43.00   60.00
```



Расподела година запослених указује на релативно нормалну дистрибуцију, без грешака или екстремних вредности које би могле негативно утицати на наш модел.

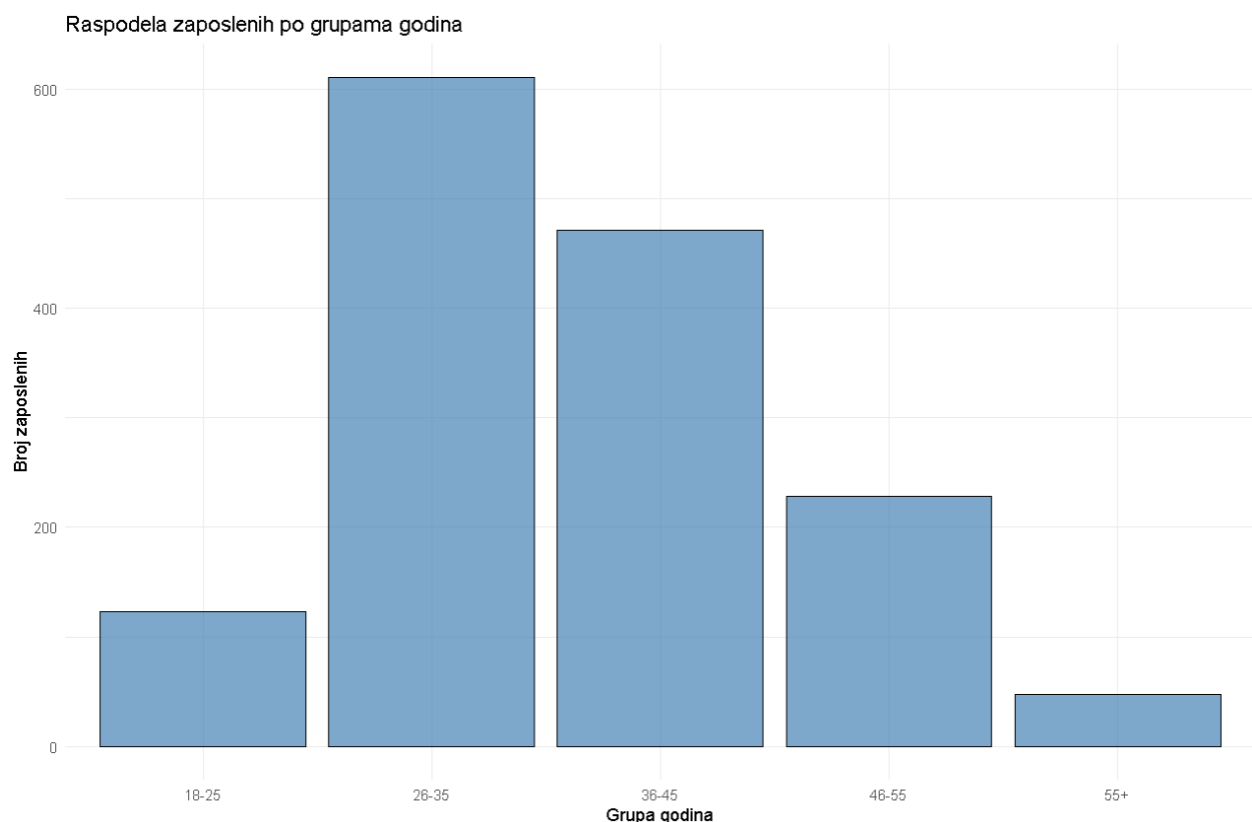
Видимо да су средња вредност као и медијана сличне што додатно потврђује да не постоје изузеци у подацима.

### Колона „AgeGroup“

Представља опсег година у којем се налази запослени.

```
> print(unique_values_summary$AgeGroup_unique_values)
[1] 5
> print(unique_values_summary$AgeGroup_sample_values)
[1] "18-25, 26-35, 36-45, 46-55, 55+"
```

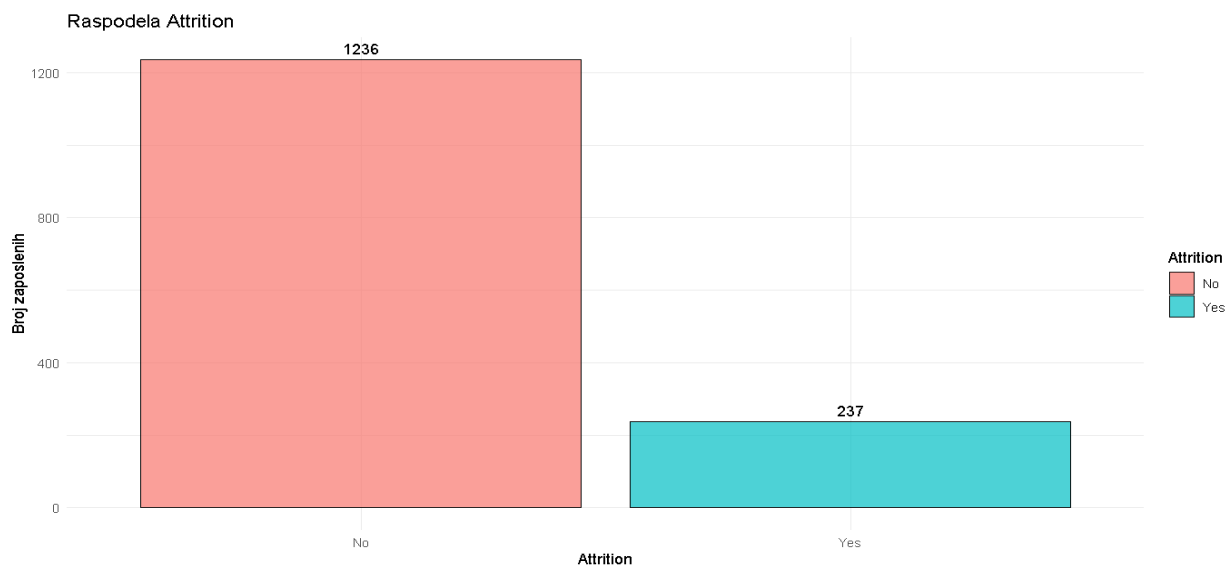
Садржи 5 различитих група: 18-25, 26-35, 36-45, 46-55, 55+. Расподелу по групама можемо приказати на следећи начин.





## Колона „Attrition“

Представља да ли је запослени напустио организацију, уједно и колона коју ћемо предвиђати у овом раду.



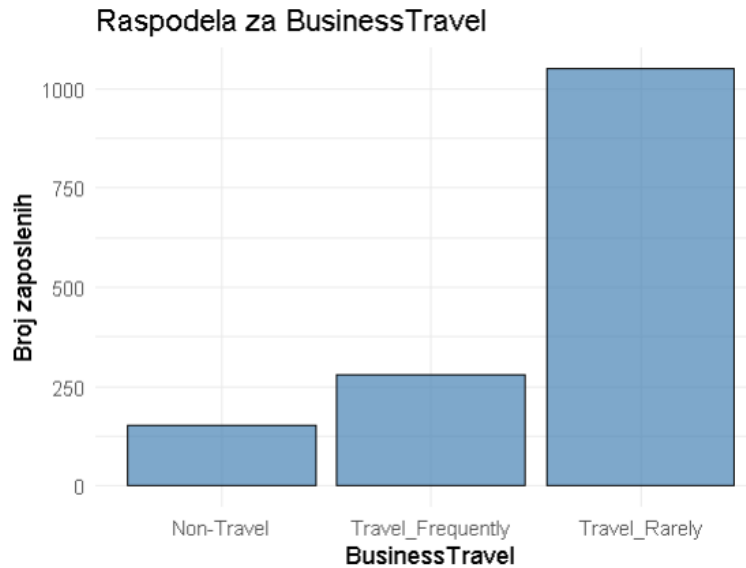
Из доступних података можемо закључити да је укупно 237 запослених напустило компанију. Овај податак указује на одређени ниво флукуације радне снаге, што може бити резултат различитих фактора, као што су незадовољство послом, боље прилике у другим компанијама, или неусклађеност између личних и професионалних потреба запослених. Анализа разлога за одлазак ових запослених може пружити драгоцене увиде у области које треба побољшати у оквиру компаније, као и у стратегије задржавања радне снаге.

## Колона „BusinessTravel“

Представља информације о учесталости путовања запосленог. Садржи 4 могуће вредности: „Travel\_Rarely“, „Travel\_Frequently“, „Non-Travel“, „TravelRarely“.

```
> print(unique_values_summary$BusinessTravel_unique_values)
[1] 4
> unique_values_summary$BusinessTravel_sample_values <- strsplit(unique_values_summary$BusinessTravel_sample_values[1], ", ")[1]
> print(unique_values_summary$BusinessTravel_sample_values[[1]][1:unique_values_summary$BusinessTravel_unique_values])
[1] "Travel_Rarely"      "Travel_Frequently" "Non-Travel"       "TravelRarely"
```

Након исправљања структурних грешака, које ћемо објаснити у даљем току рада, расподела за ову колону је:



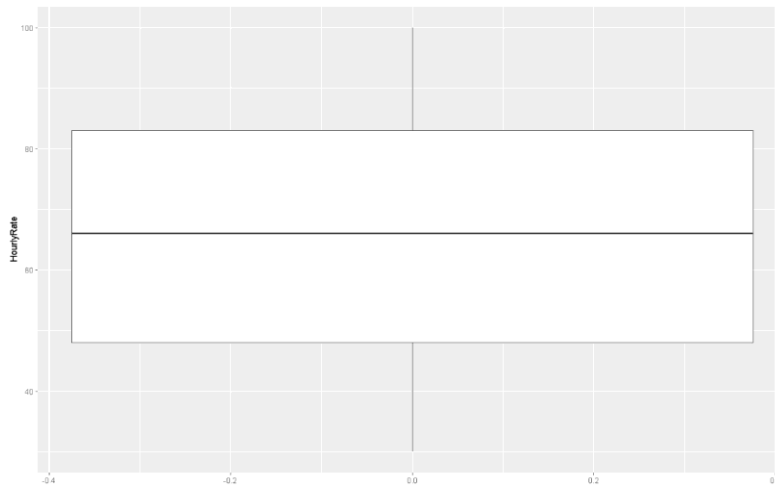
Већина запослених ретко има обавезу да иде на пословни пут, док нешто више од 250 запослених често путује због пословних обавеза. Најмањи број чине они који никада не иду на службена путовања, око 150 запослених.

### Колона „StandardHours“

Ова колона указује на уговорени број радних сати на месечном нивоу. Сви запослени имају уговорено 80 сати, што значи да ова колона неће имати значаја у нашем даљем истраживању. Поступак уклањања објаснићемо у даљем тексту.

## Колона „HourlyRate“

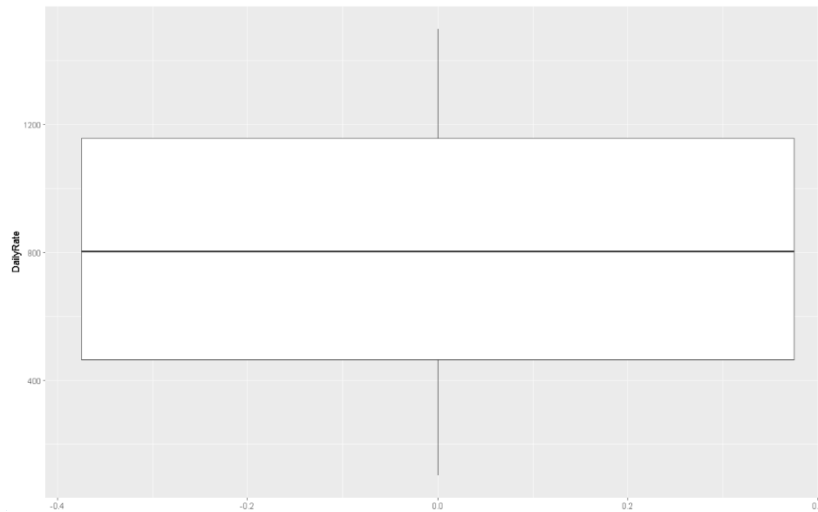
Представља уговорени износ запосленог по радном сату. Можемо проверити расподелу ове колоне као и да ли ова колона садржи екстремне вредности који би негативно утицали на наш модел.



Видимо да ова колона не садржи изузетке и да задржава нормалну дистрибуцију.

## Колона „DailyRate“

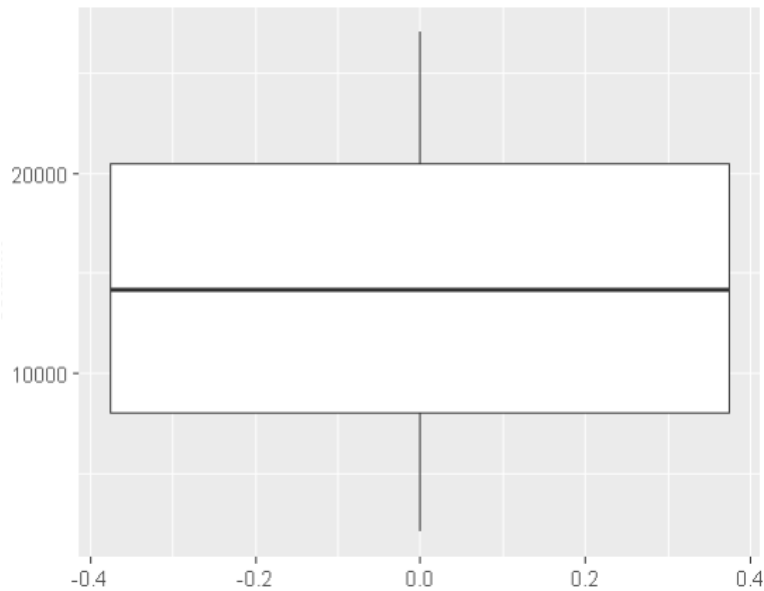
Представља уговорени износ дневне накнаде запосленог. Као и за „HourlyRate“ колону, можемо проверити екстремне вредности и њену расподелу.



Добијамо сличну расподелу без изузетака као код „HourlyRate“, што је и логично, с обзиром на то да сви запослени имају исти уговорени број сати.

### Колона „MonthlyRate“

Представља уговорену месечну зараду запосленог.

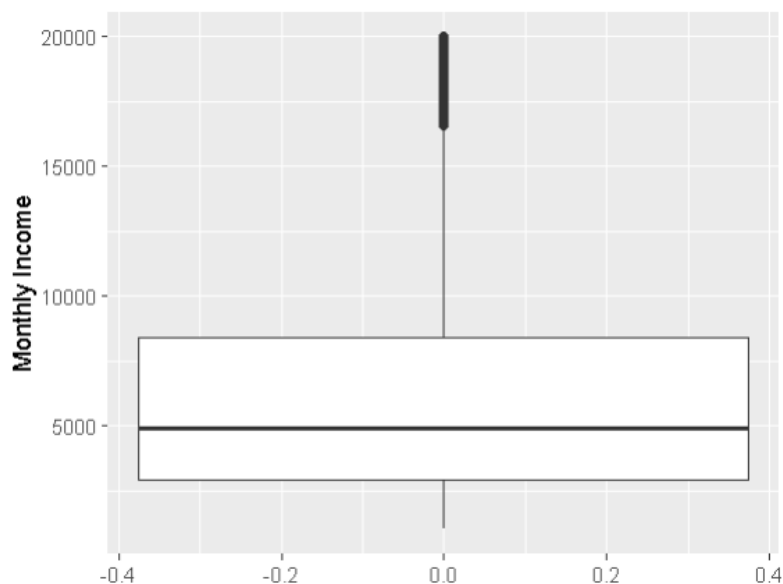


Као и код осталих типова ове колоне тако ни ова колона не садржи екстремне вредности.

### Колона „MonthlyIncome“

Представља укупне месечне приходе запосленог.

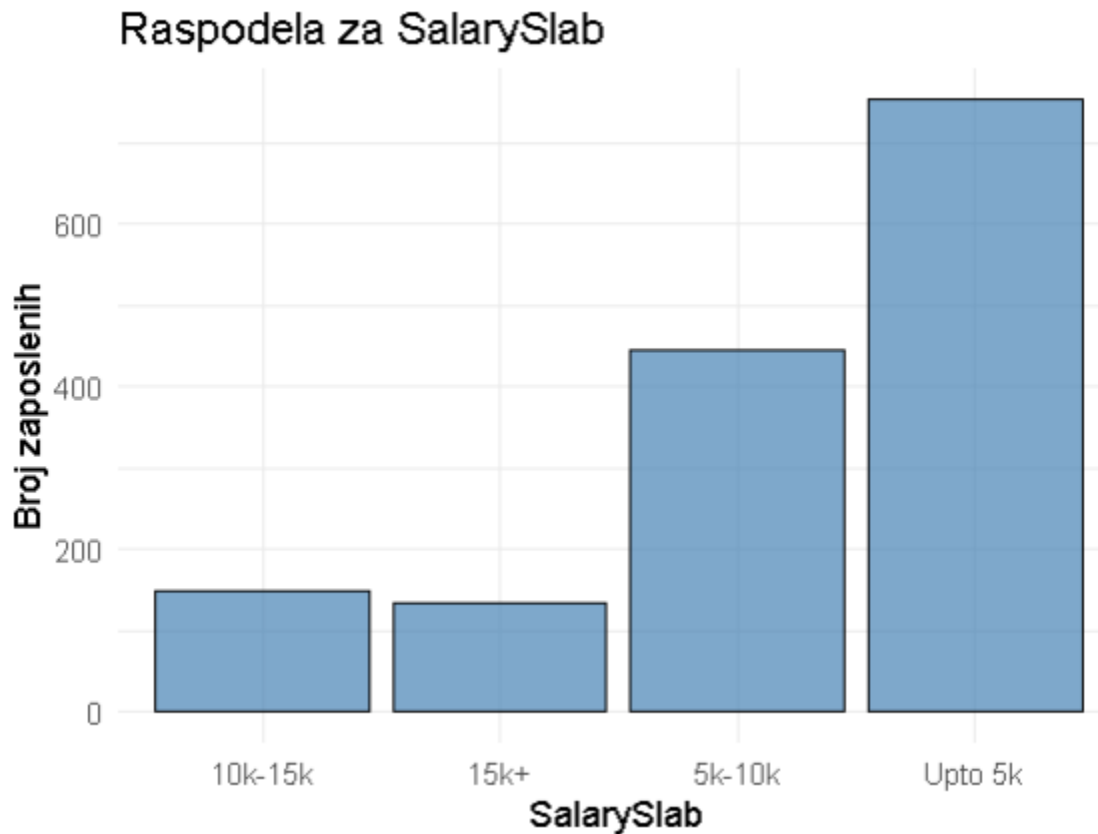
Можемо проверити расподелу ове колоне на боксплоту и проверити екстремне вредности.



Видимо да постоје вредности које се налазе изван граница ( $Q3+1.5 \times IQR$ ) проверићемо у даљем раду да ли је потребно да избацимо ове вредности.

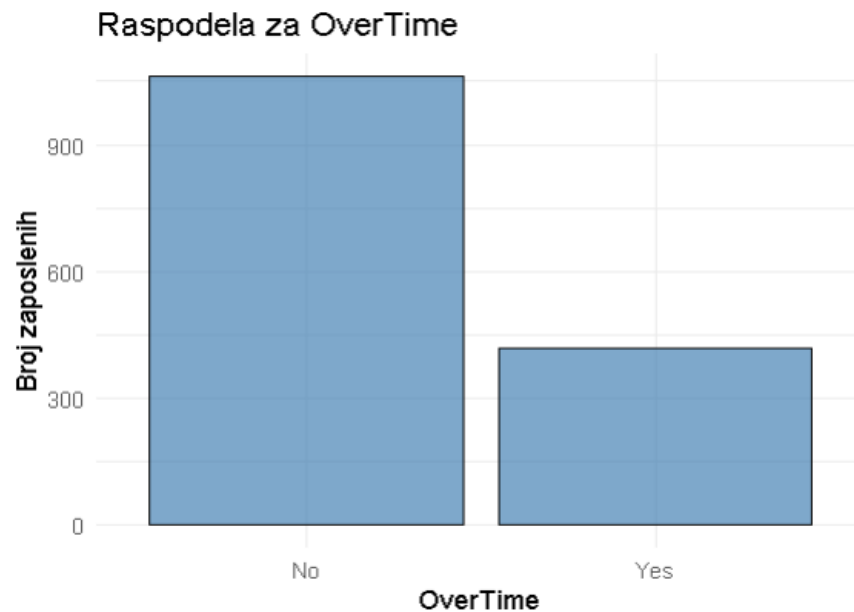
### Колона „SalarySlab“

Представља категорију у којој се налази запослени на основу укупног месечног прихода.



## Колона „OverTime“

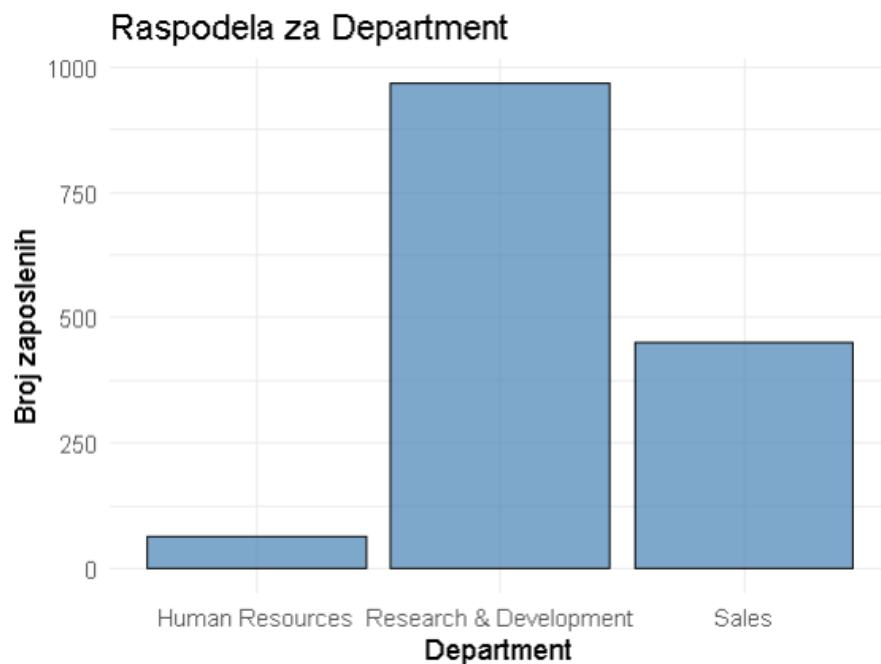
Представља увид да ли запослени ради прековремено или не.



Можемо закључити да већина запослених не ради прековремено али такође да нам ово можда може помоћи при креирању модела.

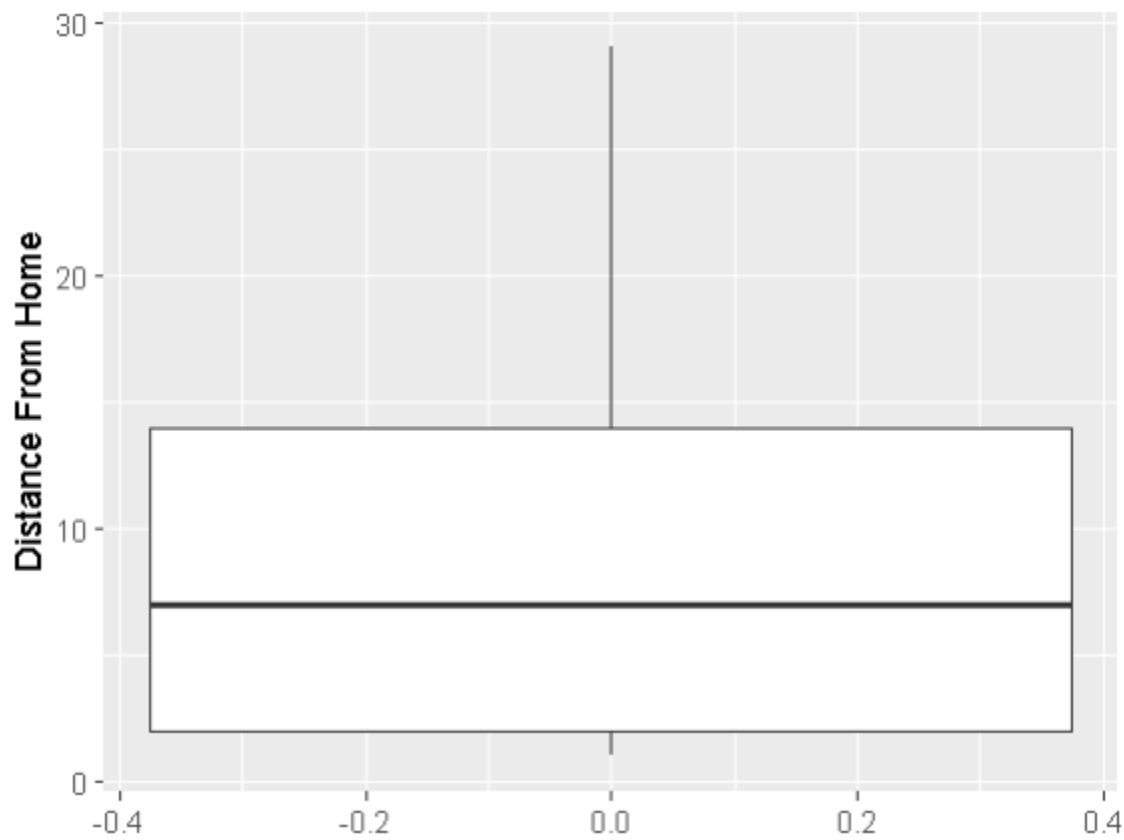
## Колона „Department“

Представља ком одељењу припада запослени.



### Колона „DistanceFromHome“

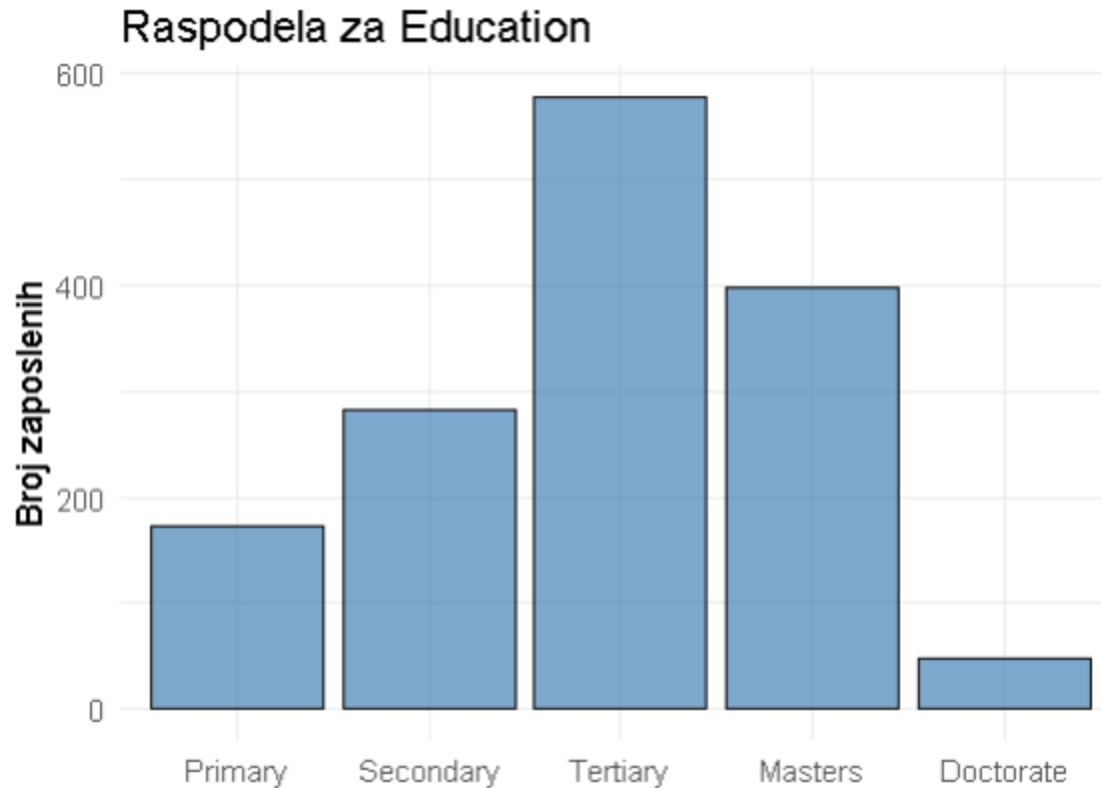
Представља удаљеност места пребивалишта од фирме. Можемо проверити да ли ова колона садржи екстремне вредности који би негативно утицали на наш модел. У наставку рада ћемо додатно обрадити ову колону.



### Колона „Education“

Представља ниво образовања запосленог.

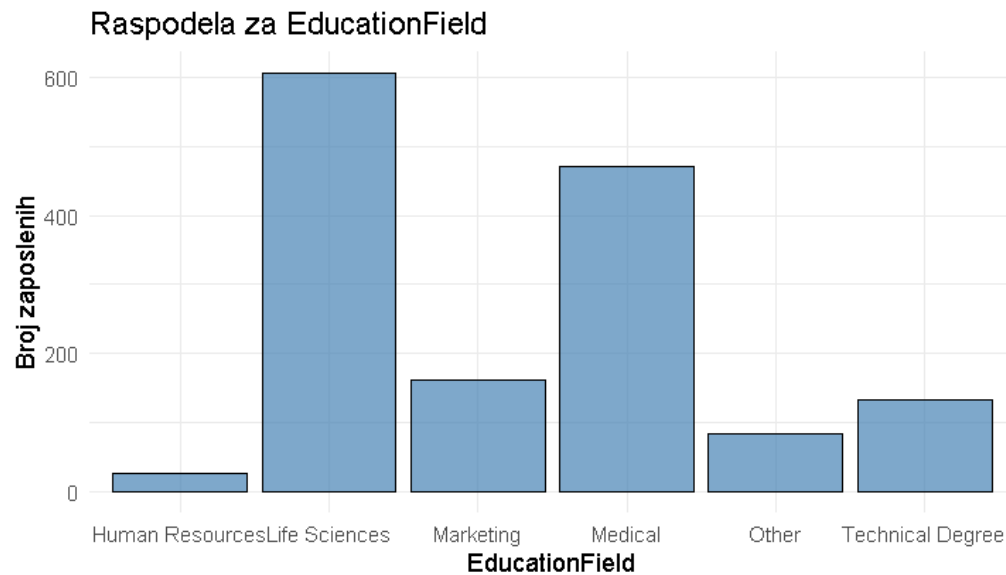




Видимо да и ова колона има униформну расподелу. Највише запослених има завршене факултете, потом мастер студије и средњу школу. Такође видимо да 15% запослених има завршену само основну школу, а веома мали број запослених има докторску дисертацију.

## Колона „EducationField“

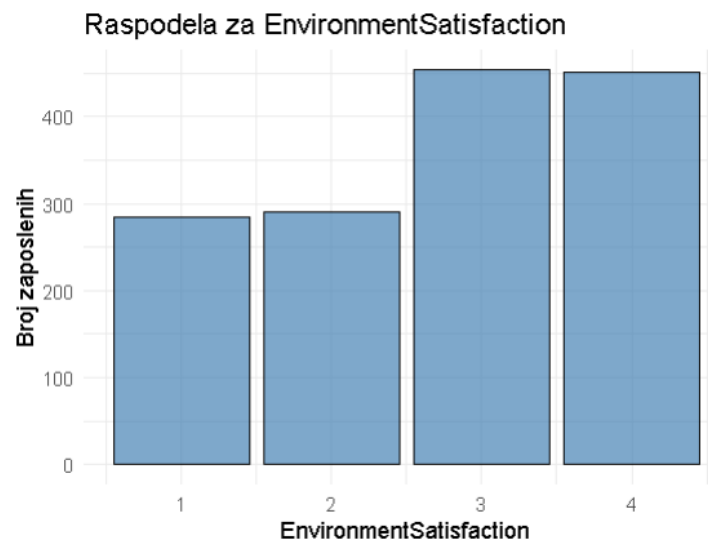
Представља поље образовања запосленог.



Највише запослених долази из сфера природних наука и медицине.

## Колона „EnvironmentSatisfaction“

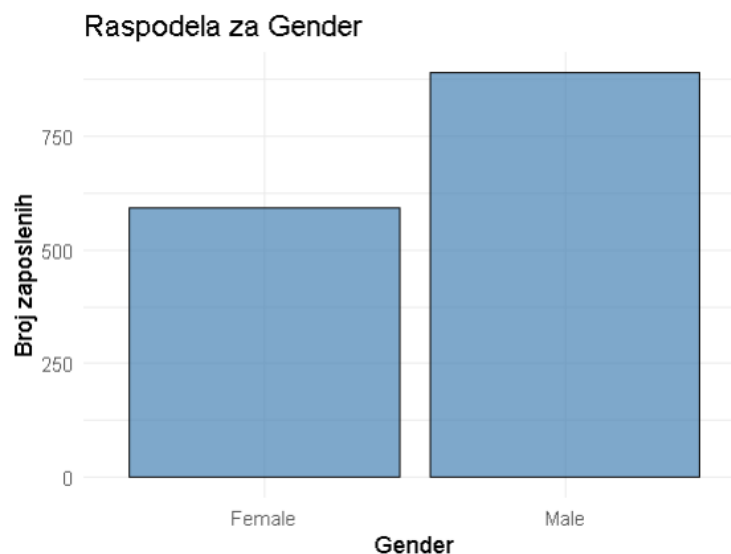
Представља задовољство пословним окружењем. Оцене су приказане на скали од 1 до 4, где оцена 1 означава да запослени уопште није задовољан док оцена 4 означава да је веома задовољан.



Запослени су углавном задовољни својим пословним окружењем. Може бити добар предиктор.

## Колона „Gender“

Представља пол запосленог.



Видимо да трећину запослених чине особе женског пола.

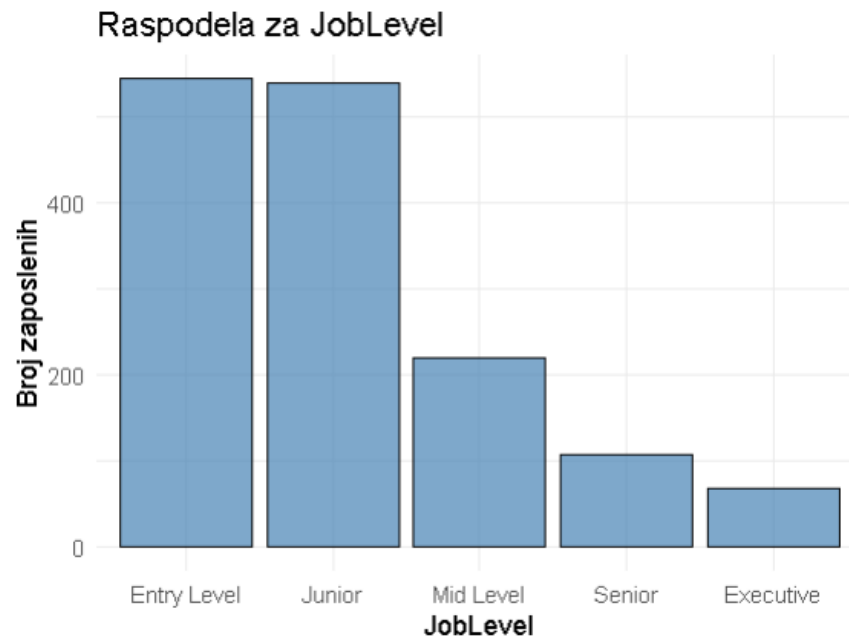
## Колона „JobInvolvement“

Ова колона одражава степен укључености запосленог у активности и одлуке у компанији.



## Колона „JobLevel“

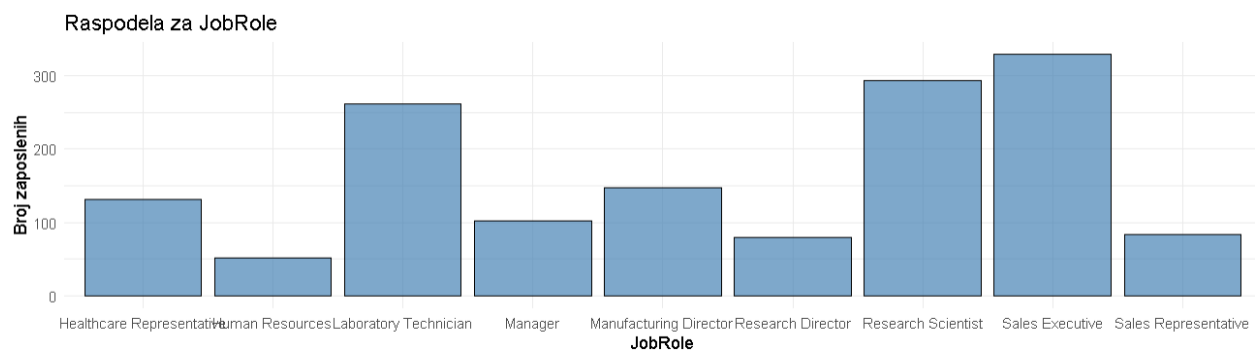
Представља ниво позиције запослених у компанији. Ова колона обично категоризује запослене у различите нивое, који могу указивати на степен одговорности, сложеност посла и положај у организацији. Подела скупа података на основу ове колоне изгледа:



Добијамо очекивану расподелу, где се највише људи налази на нижим позицијама, а како позиције расту тако и број запослених опада.

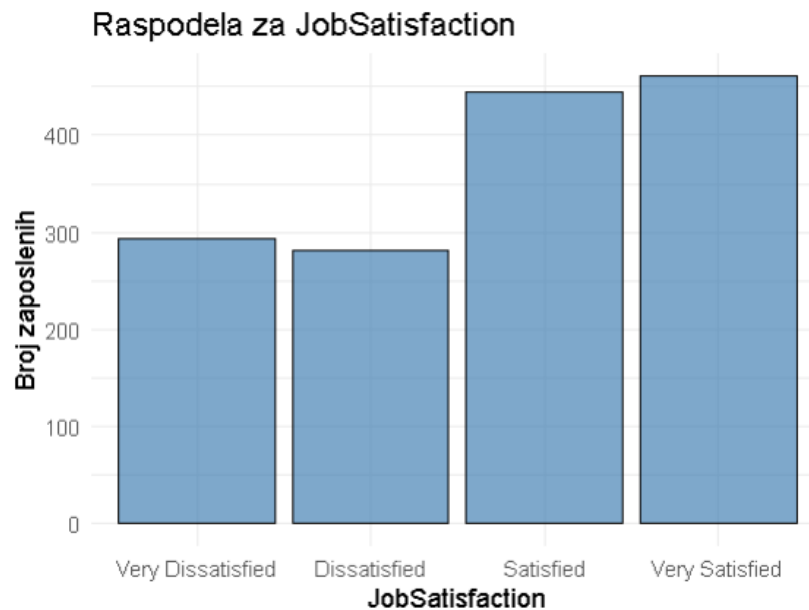
## Колона „JobRole“

Представља звање запосленог унутар компаније.



## Колона „JobSatisfaction“

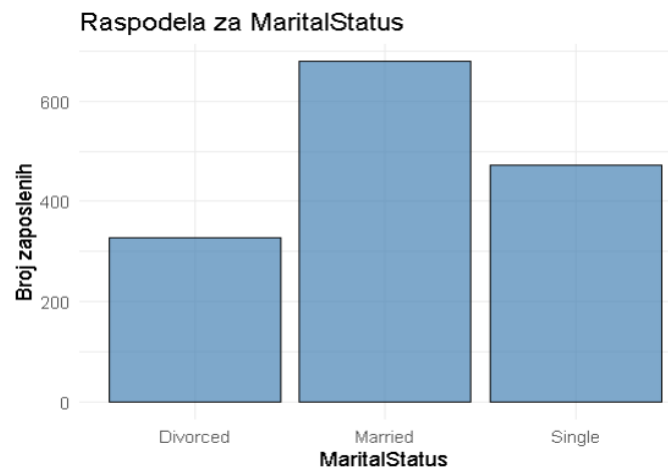
Представља задовољство запосленог на пословном плану.



Више запослених је задовољно самим послом, може бити један од важнијих предиктора.

## Колона „MaritalStatus“

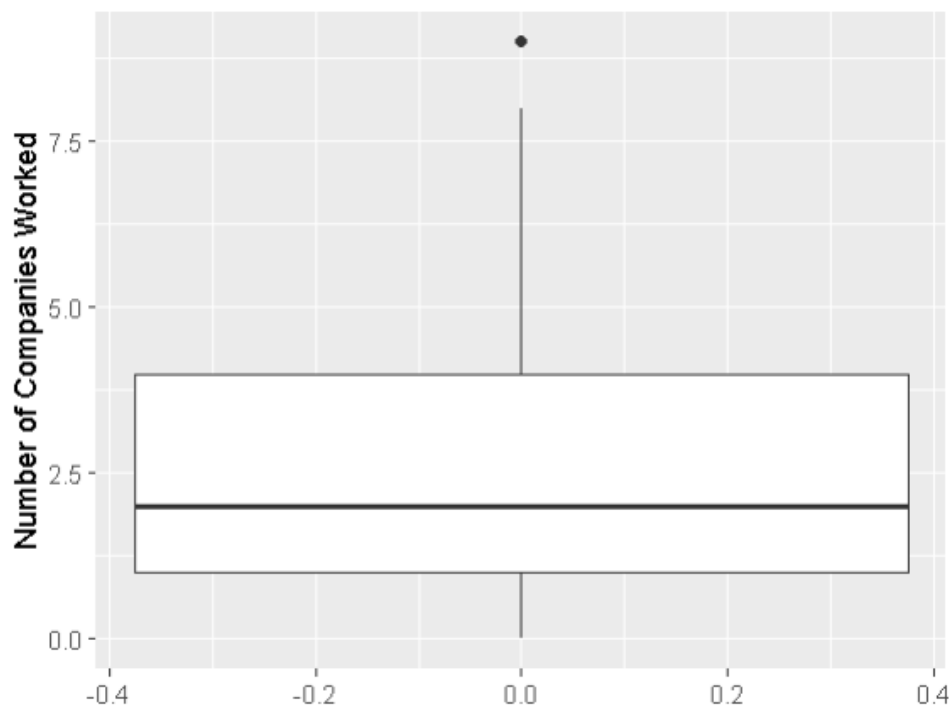
Представља брачно стање запосленог.



Највише има запослених који су у браку. Можда у комбинацији са „WorkLifeBalance“ подацима можемо добити корисне информације за наш модел.

### Колона „NumCompaniesWorked“

Ова колона указује на број компанија у којима је запослени раније радио. Не укључујући тренутну компанију.



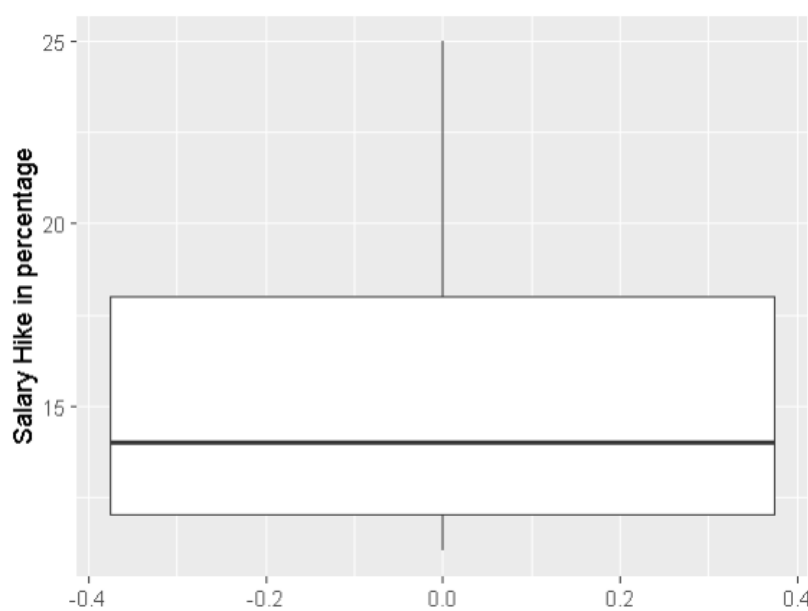
Видимо да се у скупу података налази запослени који је чак 8 пута променио организацију. Можемо га избацити из овог скупа.

### Колона „PercentSalaryHike“

Представља проценат последње повишице. Проценат је израчунат на основу уговорене месечне зараде.

```
> summary(hr_data$PercentSalaryHike)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.00  12.00   14.00   15.21  18.00   25.00
```

Видимо да је у последњој години најмања повишица износила 11% док је максимална износила 25% уговорене месечне зараде. Изгледа да немамо екстремне вредности али вредело би погледати и расподелу ове колоне.

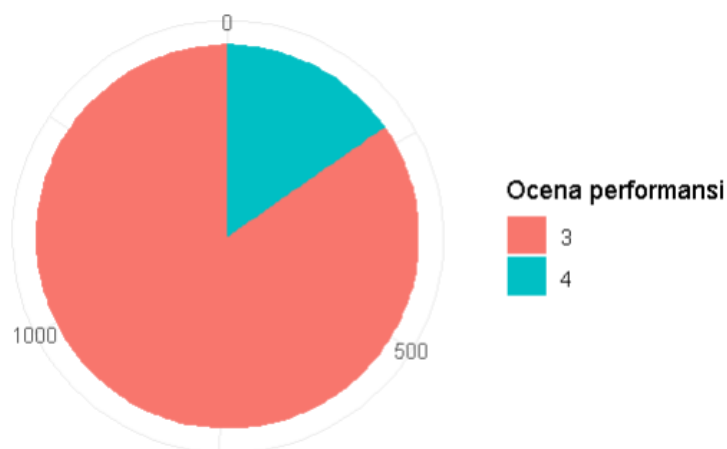


Нема екстремних вредности. С обзиром да ова колона утиче позитивно на запослене може бити добар предиктор.

### Колона „PerformanceRating“

Представља оцену перформансе запосленог. Оцене су приказане на скали 1 до 4.

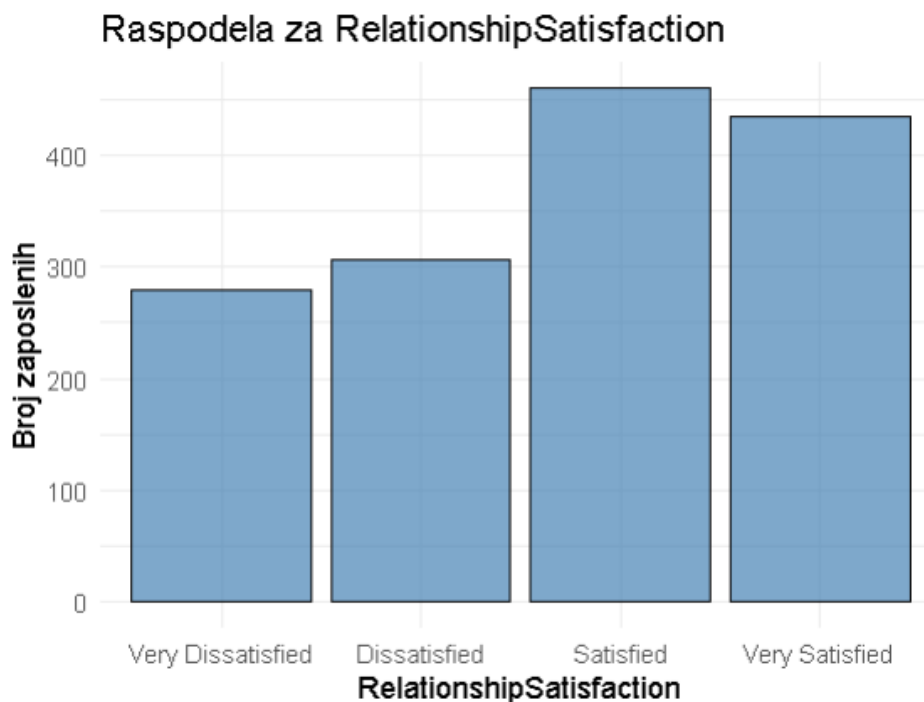
Raspodela ocena performansi zaposlenih



Видимо да су дате оцене само 3 и 4. Као и да осмину скупа података чине запослени који имају изузетне перформансе. Можемо уклонити ову колону из даљих опсервација.

### Колона „RelationshipSatisfaction“

Представља тренутни статус везе запослених на приватном плану. Ова информација може бити значајна за анализу различитих аспеката радне динамике и задовољства запослених.

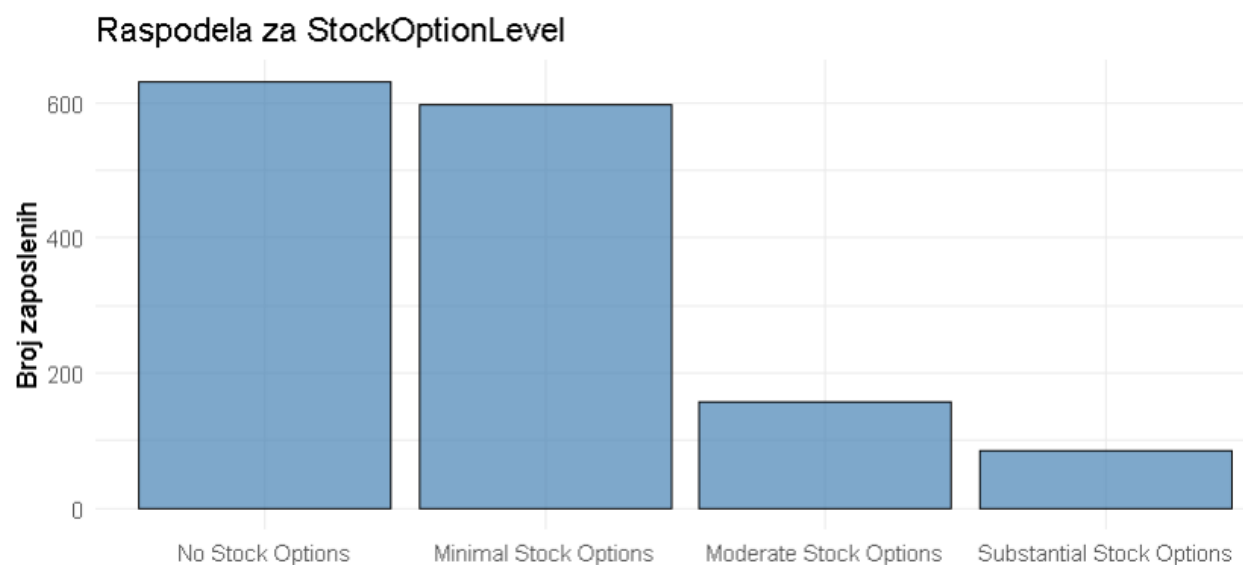


Чини се да је расподела приближна и да ће нам можда користити у комбинацији са још неким предиктором.

### Колона „StockOptionLevel“

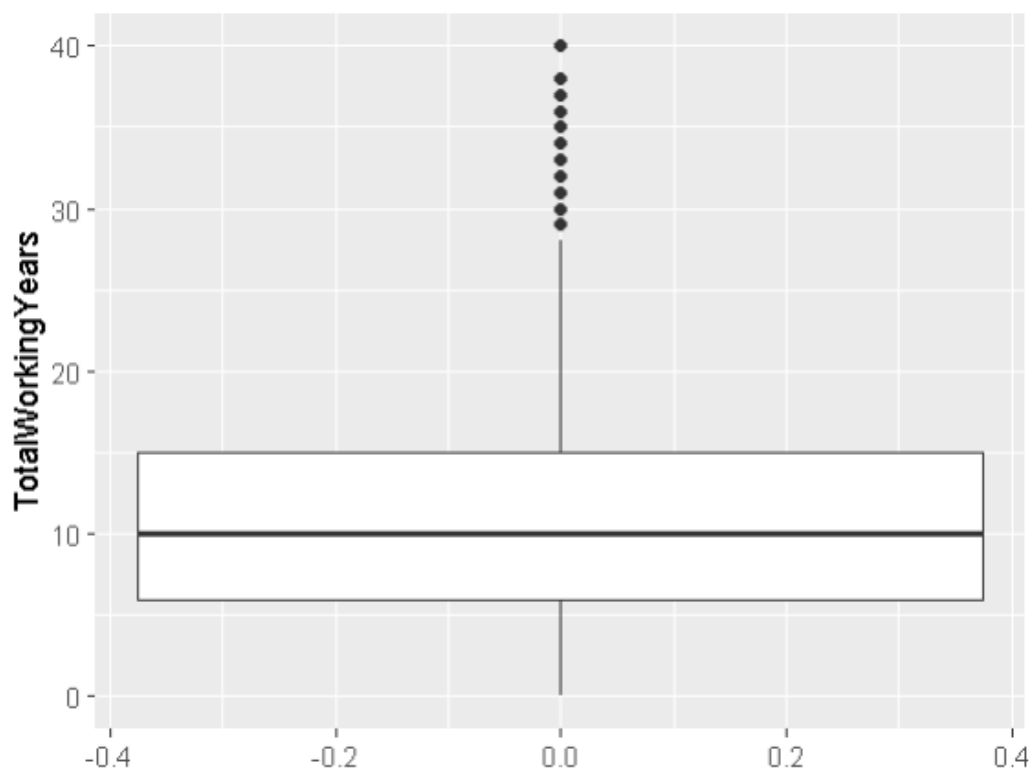
Показује да ли запослени има могућност да купује акције компаније. Највећи број запослених нема могућност поседовања акција компаније (0), опције 1, 2, 3 представљају ниво могућности поседовања акција компаније. Виши нивои представљају услове за привилеговане запослене на вишим позицијама и дужим радним стажом, док нижи нивои представљају ограничену могућност куповине акција и мање повољне услове. Како бисмо лепше приказали односе, уместо нумеричких ознака користићемо текстуалне ознаке. Оцену 0 ћемо описати „Без опција“, 1 „Минимално опција“, 2 „Умерено“ и 3 „Значајно“.





### Колона „TotalWorkingYears“

Представља укупан радни стаж запосленог.



Видимо да постоје екстремне вредности, њих ћемо обрадити у следећем делу рада.

### Колона „TrainingTimesLastYear“

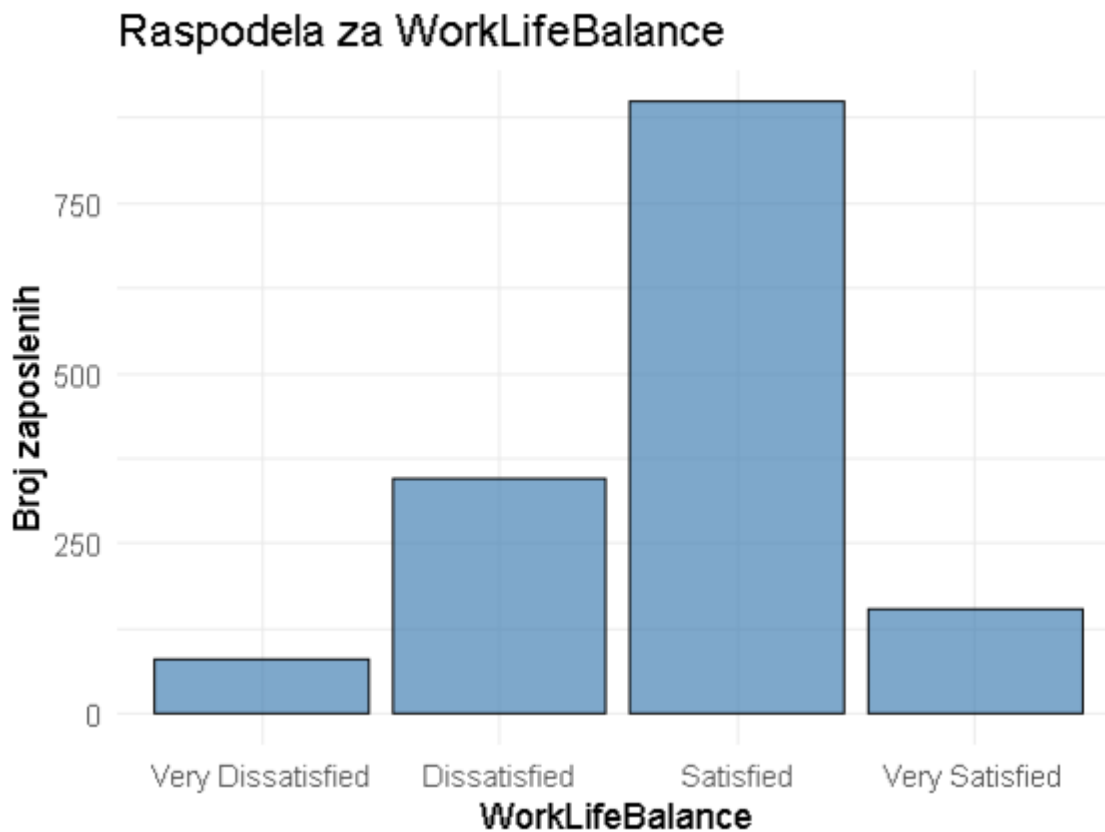
Представља колико је пута радник био на професионалном обучавању у претходној години.



Видимо да је расподела униформна и да не постоје екстремне вредности.

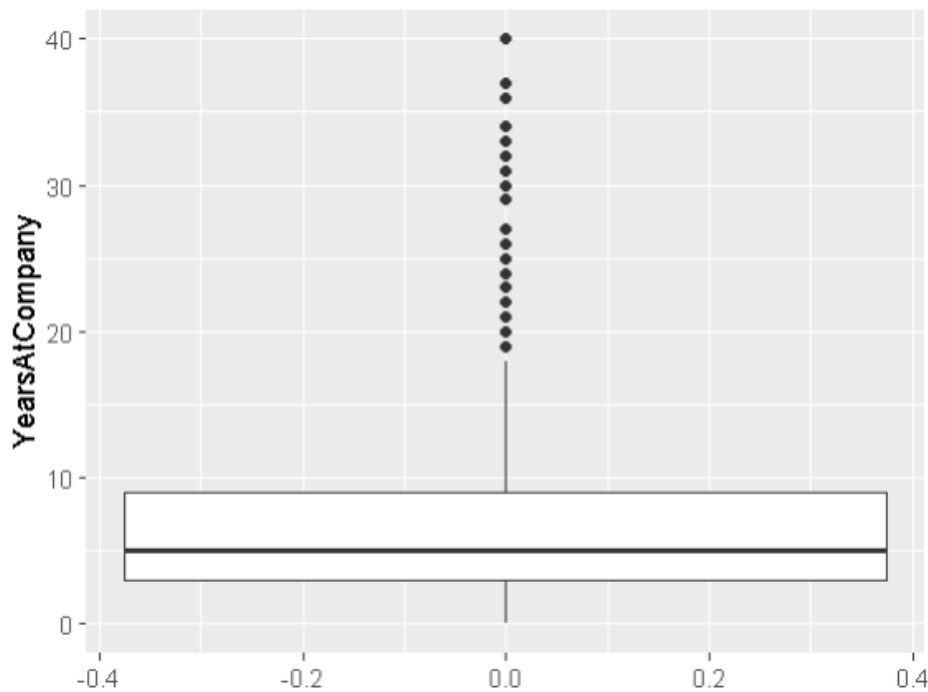
### Колона „WorkLifeBalance“

Представља ниво којим запослени успева да се посвети како пословном тако и приватном животу.



### Колона „YearsAtCompany“

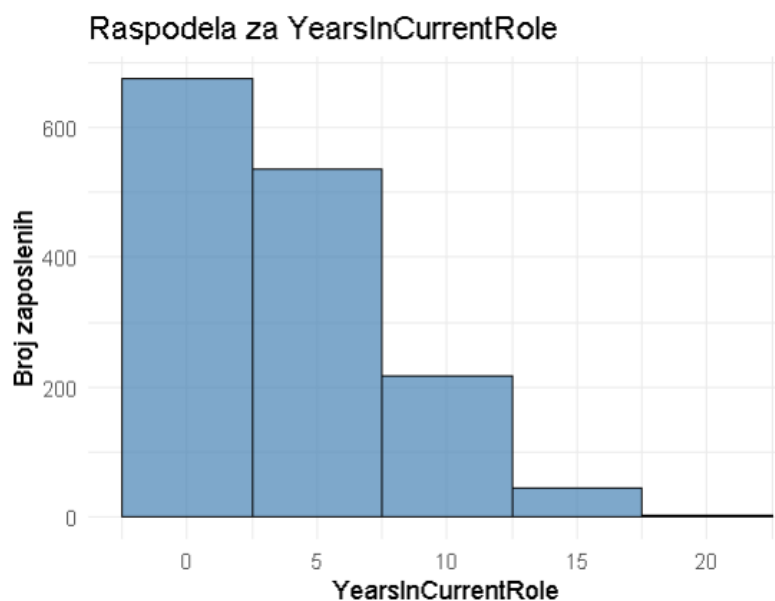
Представља колико година је запослени провео у фирми.



Видимо да и овде постоје екстремне вредности. То су запослени који су цео радни век провели у једној или две компаније.

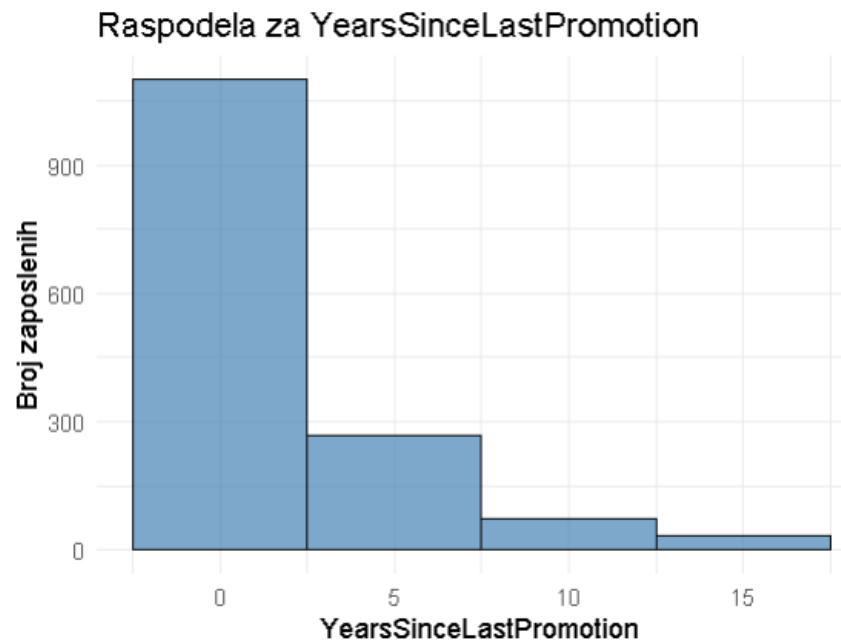
### Колона „YearsInCurrentRole“

Представља колико година је запослени провео на истој позицији.



## Колона „YearsSinceLastPromotion“

Представља колико година је прошло од последње промоције.

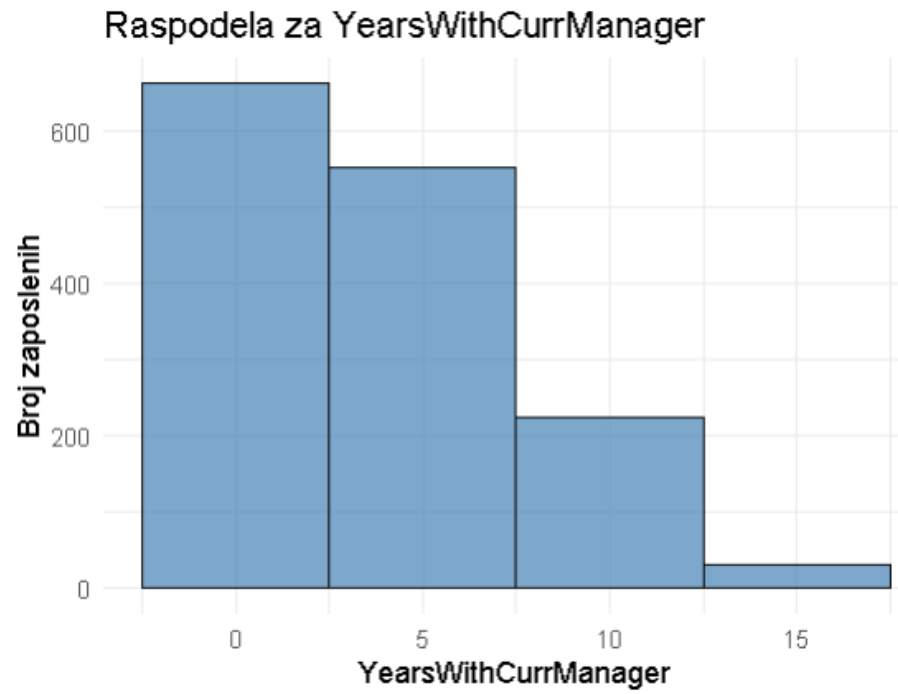


## Колона „YearsWithCurrentManager“

Ова колона показује број година које запослени провео под тренутним менаџером.

```
> summary(hr_data$YearsWithCurrManager)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's 
 0.000   2.000   3.000   4.118   7.000  17.000    57
```

Након попуњавања недостајућих вредности, где ћемо детаљне кораке представити у даљем раду, можемо погледати коначну расподелу.



Видимо да број запослених опада, што смо и очекивали.

## 2.2 Уклањање дупликата и неважних колона

С обзиром да смо раније уочили дуплиране уносе података колоне „EmpID“ и како нам ова колона није од превелике користи за наш модел, можемо проверити да ли су цели редови дупликати.

```
> print(hr_data[duplicated(hr_data), ])  
# A tibble: 7 × 38  
  EmpID    Age AgeGroup Attrition BusinessTravel    DailyRate Department DistanceFromHome Education EducationField  
  <chr>    <dbl> <chr>    <chr>    <chr>    <dbl> <chr>    <dbl>    <dbl> <chr>  
1 RM1468    27 26-35    No      Travel_Rarely    155 Research ...      4      3 Life Sciences  
2 RM1461    29 26-35    No      Travel_Rarely    468 Research ...     28      4 Medical  
3 RM1464    31 26-35    No      Non-Travel      325 Research ...      5      3 Medical  
4 RM1470    34 26-35    No      Travel_Rarely    628 Research ...      8      3 Medical  
5 RM1463    39 36-45    No      Travel_Rarely    722 Sales      24      1 Marketing  
6 RM1469    49 46-55    No      Travel_Frequently 1023 Sales      2      3 Medical  
7 RM1462    50 46-55    Yes     Travel_Rarely    410 Sales      28      3 Marketing  
# i 28 more variables: EmployeeCount <dbl>, EmployeeNumber <dbl>, EnvironmentSatisfaction <dbl>, Gender <chr>,  
#   HourlyRate <dbl>, JobInvolvement <dbl>, JobLevel <dbl>, JobRole <chr>, JobSatisfaction <dbl>,  
#   MaritalStatus <chr>, MonthlyIncome <dbl>, SalarySlab <chr>, MonthlyRate <dbl>, NumCompaniesWorked <dbl>,  
#   Over18 <chr>, OverTime <chr>, PercentSalaryHike <dbl>, PerformanceRating <dbl>,  
#   RelationshipSatisfaction <dbl>, StandardHours <dbl>, StockOptionLevel <dbl>, TotalWorkingYears <dbl>,  
#   TrainingTimesLastYear <dbl>, WorkLifeBalance <dbl>, YearsAtCompany <dbl>, YearsInCurrentRole <dbl>,  
#   YearsSinceLastPromotion <dbl>, YearsWithCurrManager <dbl>
```

Пронашли смо укупно 7 дуплираних редова које ћемо уклонити из скупа података уз помоћ:

```
> hr_data <- hr_data[!duplicated(hr_data), ]  
> nrow(hr_data)  
[1] 1473
```

Успешно смо уклонили 7 редова, сада се у скупу налази 1473 јединствених редова података.

Три преостала реда можемо сврстати под грешку при уносу колоне EmpID. С обзиром на то да ова колона не игра улогу у нашем моделу, можемо их оставити у скупу података без потребе за уклањањем.

За почетак уклонимо податке везане за идентификацију запослених попут **EmployeeCount**, **EmpID**, **EmployeeNumber**:

```
hr_data$EmployeeCount <- NULL  
hr_data$EmpID <- NULL  
hr_data$EmployeeNumber <- NULL
```

Даљом анализом приметили смо да су вредности колоне **Over18** и **StandardHours** исте за све запослене, па ћемо их такође избацити:

```
hr_data$StandardHours <- NULL  
hr_data$Over18 <- NULL
```

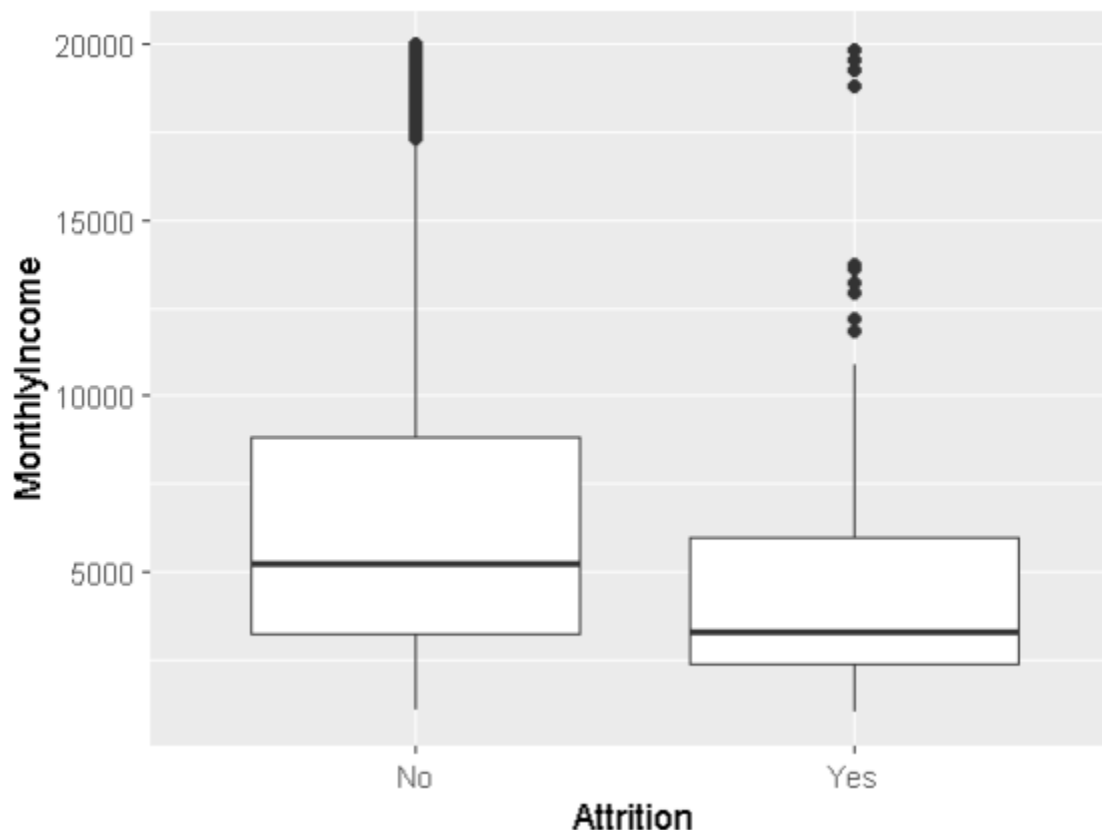
## 2.3 Исправљање структурних грешака

Видели смо да је потребно извршити трансформацију података с обзиром да постоје уноси за TravelRarely и Travel\_Rarely који представљају исту учесталост. Уз помоћ функције „mutate“ претворићемо „TravelRarely“ у „Travel\_Rarely“ како би унос био конзистентан са осталим вредностима.

```
> hr_data <- hr_data %>%  
+   mutate(BusinessTravel = ifelse(BusinessTravel == "TravelRarely", "Travel_Rarely", BusinessTravel))  
  
> print(unique_values_summary$BusinessTravel_unique_values)  
[1] 3  
> unique_values_summary$BusinessTravel_sample_values <- strsplit(unique_values_summary$BusinessTravel_sample_values[1], ", ")[1]  
> print(unique_values_summary$BusinessTravel_sample_values[[1]][1:unique_values_summary$BusinessTravel_unique_values])  
[1] "Travel_Rarely" "Travel_Frequently" "Non-Travel"
```

## 2.4 Уклањање екстремних вредности

С обзиром да се на графику налазе екстремне вредности можемо проверити да ли би нам ове вредности негативно утицале на модел.



Колико видимо са графика, постоји веза између прихода и колоне која описује напуштање запослених где је просечни месечни приход мањи код људи који



су напустили организацију. Пошто се екстремне вредности налазе и код људи који су напустили фирму и оних који нису, нема потребе за избацивањем изузетака.

## 2.5 Руковање недостајућим вредностима

Ова колона садржи 57 Н/А вредности. Како бисмо што прецизније одредили средњу вредност, податке ћемо груписати на основу „AgeGroup“ колоне а затим за сваку групу извући њену просечну вредност којом ћемо заменити НА вредности.

```
> hr_data <- hr_data %>%
+   group_by(AgeGroup) %>%
+   mutate(YearsWithCurrManager = if_else(is.na(YearsWithCurrManager),
+                                         mean(YearsWithCurrManager, na.rm = TRUE),
+                                         YearsWithCurrManager)) %>%
+   ungroup()
> summary(hr_data$YearsWithCurrManager)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   2.000   3.000   4.116   7.000  17.000
```

## 2.6 Креирање нових променљивих

### DistanceFromHomeGroup

Примећујемо да највећи број људи живи на удаљености мањој од 1, док је удаљеност 10+ јако ретка. Због тога ћемо направити нову променљиву DistanceFromHomeGroup која ће садржати 5 категорија удаљености ради лакше анализе податка удаљености запосленог од куће.

Подела на групе:

- Јако близу (0 до 1)
- Близу (1 до 2)
- Средње (2 до 6)
- Далеко (6 до 10)
- Јако далеко (10 и више)

```
hr_data$DistanceFromHomeGroup <- ifelse(hr_data$DistanceFromHome <= 1, "Jako blizu", #mozda spojiti 1 i 2
                                         ifelse(hr_data$DistanceFromHome == 2, "Blizu",
                                         ifelse(hr_data$DistanceFromHome <= 6, "Srednje",
                                         ifelse(hr_data$DistanceFromHome <= 10, "Daleko", "Jako daleko"))))

table(hr_data$DistanceFromHomeGroup)
```

Blizu	Daleko	Jako blizu	Jako daleko	Srednje
211	335	208	445	274

```
> |
```

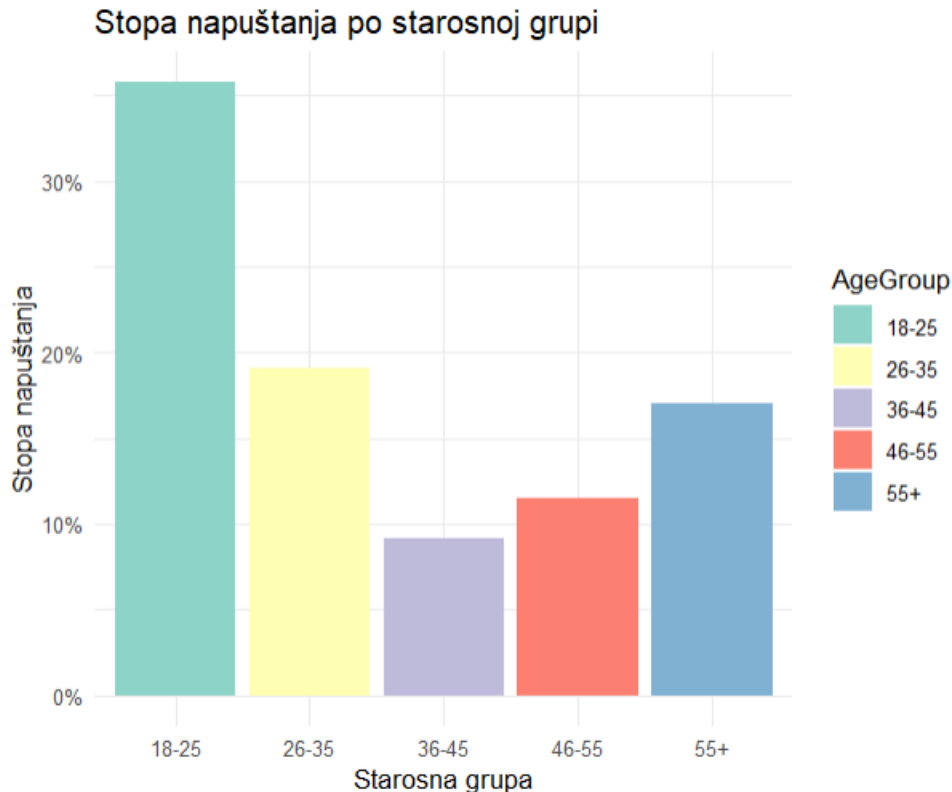
### 3.      **Анализа података**

У овом поглављу ћемо покшати да уочимо повезаност између колона на основу досадашње анализе.

- **Age group и Attrition**

```
attrition_rate_by_age <- hr_data %>%
  group_by(AgeGroup) %>%
  summarise(attrition_rate = mean(Attrition == "Yes"))

ggplot(attrition_rate_by_age, aes(x = AgeGroup, y = attrition_rate, fill = AgeGroup)) +
  geom_bar(stat = "identity") +
  ylab("Stopa napuštanja") +
  xlab("Starosna grupa") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Stopa napuštanja po starosnoj grupi") +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal()
```



```
ggplot(hr_data, aes(x = factor(Attrition), fill = AgeGroup)) +
  geom_bar(position = "dodge") +
  ylab("Age Group") +
  xlab("Attrition") +
  scale_x_discrete(labels = c("No", "Yes"))
```

```
> ageGroup.attrition <- as.data.frame(ageGroup.attrition) %>%
+   group_by(AgeGroup) %>%
+   mutate(Percentage = (Freq / sum(Freq)) * 100) %>%
+   ungroup()
```

```
> ageGroup.attrition
# A tibble: 10 × 4
  AgeGroup Attrition Freq Percentage
  <fct>    <fct>    <int>    <dbl>
1 18-25    No         79      64.2
2 26-35    No        491      80.9
3 36-45    No        427      90.9
4 46-55    No        200      88.5
5 55+      No         39      83.0
6 18-25    Yes         44      35.8
7 26-35    Yes        116      19.1
8 36-45    Yes         43       9.15
9 46-55    Yes         26      11.5
10 55+     Yes          8      17.0
```

На основу графика можемо да закључимо да запослени који процентуално више напуштају компанију припадају млађим старосним групама. Тај тренд је најизраженији (35.8%) у размаку од 18 до 25 година

и од 26 до 35 у мањој мери (19.1%), након тога је процентуално мања одлазност из компаније. Истиче се и група најстаријих од 55+ са степеном одлазности од 17% али њих има знатно мање, па нам овај податак није превише важан јер се вероватно ради о особама које одлазе у пензију или напуштају посао због година.

55+ 39 8

- **Age и Monthly Income**

```
ggplot(data = hr_data) +  
  geom_point(aes(x = MonthlyIncome, y = Age, color = Attrition))
```



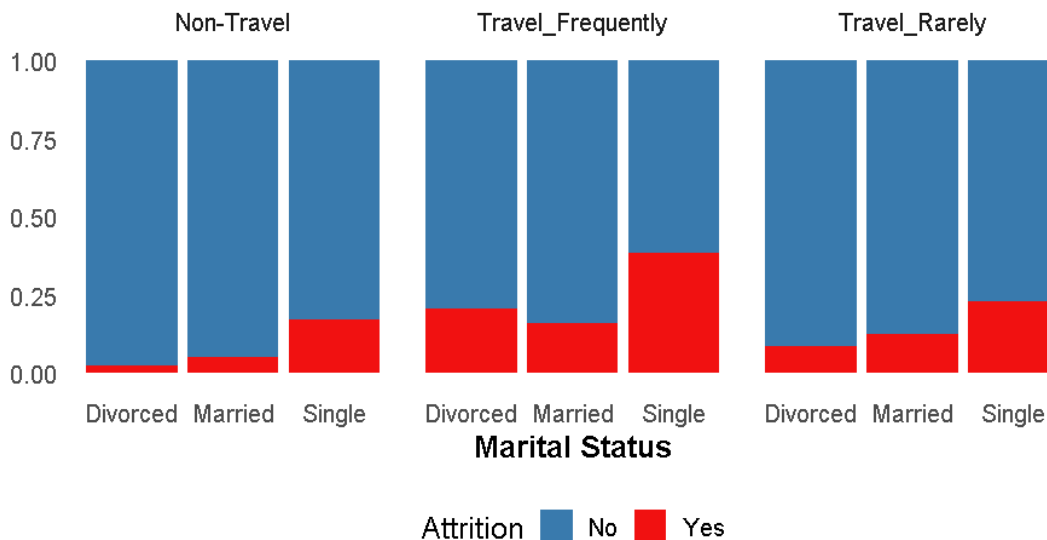
```

> ageGroup.attrition <- as.data.frame(ageGroup.attrition) %>%
+   group_by(AgeGroup) %>%
+   mutate(Percentage = (Freq / sum(Freq)) * 100) %>%
+   ungroup()
> ageGroup.attrition
# A tibble: 10 x 4
  AgeGroup Attrition   Freq Percentage
  <fct>     <fct>     <int>     <dbl>
1 18-25     No         79      64.2
2 26-35     No        491      80.9
3 36-45     No        427      90.9
4 46-55     No        200      88.5
5 55+       No         39      83.0
6 18-25     Yes         44      35.8
7 26-35     Yes        116      19.1
8 36-45     Yes         43       9.15
9 46-55     Yes         26      11.5
10 55+      Yes          8      17.0

```

У овом примеру смо упоредили године запослених и месечна примања и примећујемо да повезаност постоји. Најизраженија је за млађе запослене и можемо приметити како са порастом година расту и приманја уз доста изузезака након 30-те године.

- **Martial status и Bussines travel factor**



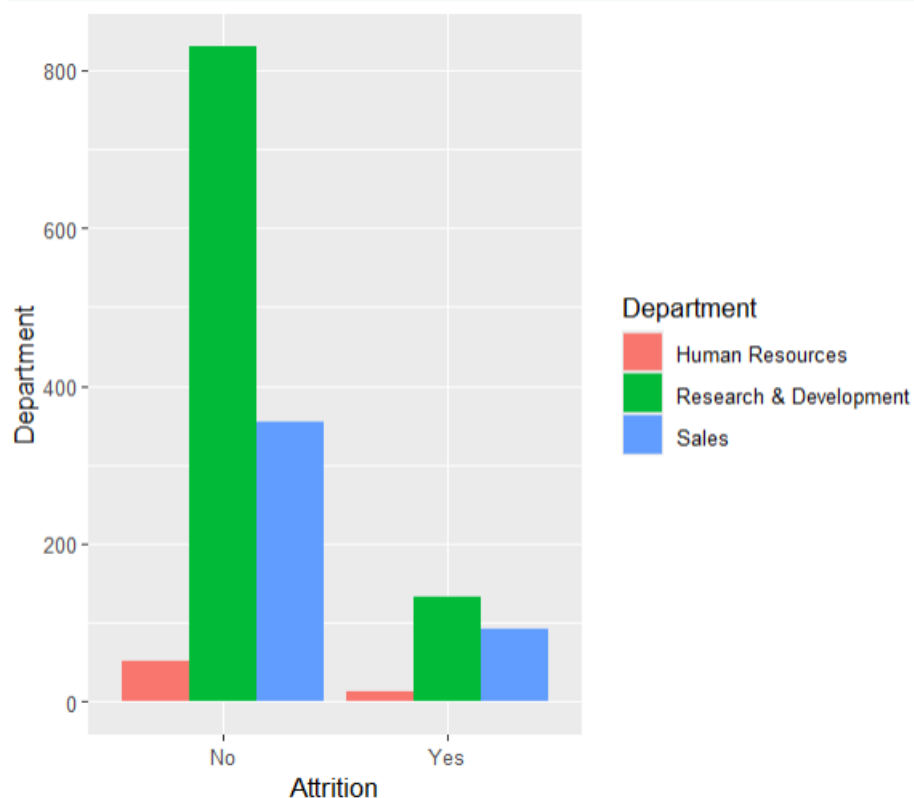
На овом графику можемо видети скалиран однос брачног статуса и учесталости пословних путовања запослених као и удео запослених који су напустили компанију.

Видимо да су највећа удела оних који су отишли управо они који најчешће путују.

Процентуално најчешће компанију напуштају људи који (често) путују а нису у браку.

- **Attrition и Department**

```
ggplot(hr_data, aes(x = factor(Attrition), fill = Department)) +  
  geom_bar(position = "dodge") +  
  ylab("Department") +  
  xlab("Attrition") +  
  scale_x_discrete(labels = c("No", "Yes"))
```

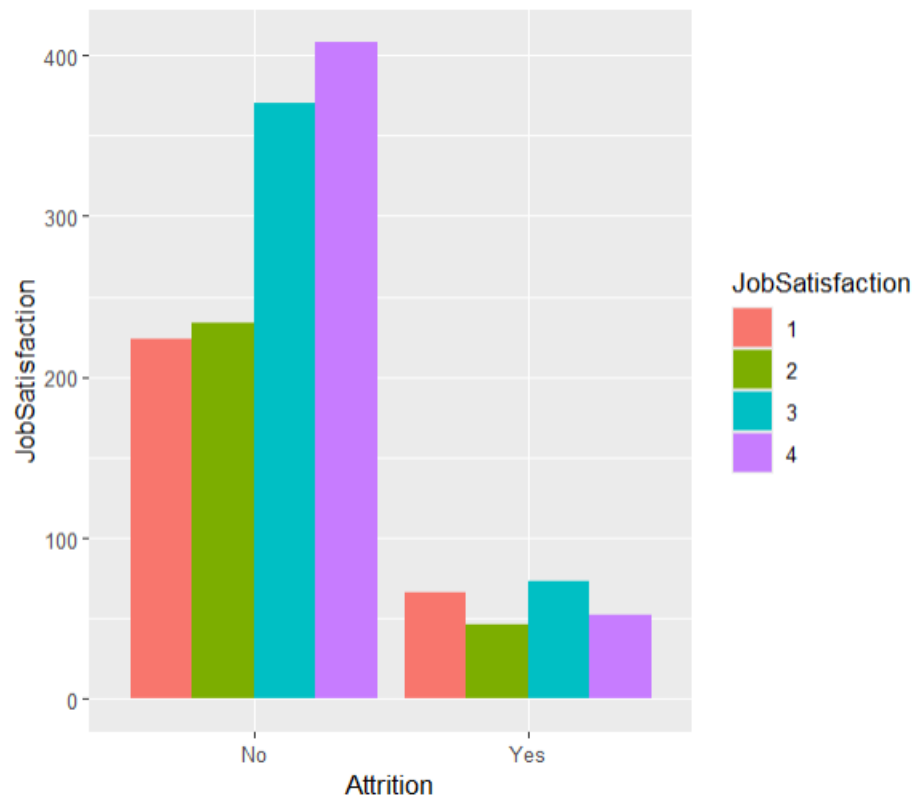


```
> department.attrition.prop <- prop.table(department.attrition, margin = 1)  
> department.attrition.prop
```

	Attrition	
Department	No	Yes
Human Resources	0.8095238	0.1904762
Research & Development	0.8618899	0.1381101
Sales	0.7941834	0.2058166

Примећујемо да је мала разлика у процентима међу одељењима, али да "Sales" има највећи проценат људи који напуштају фирму.

- **Job Satisfaction и Attrition**



```
> jobSatisfaction.attrition.prop <- prop.table(jobSatisfaction.attrition, margin = 1)
> jobSatisfaction.attrition.prop
```

	Attrition	
JobSatisfaction	No	Yes
1	0.7724138	0.2275862
2	0.8357143	0.1642857
3	0.8352144	0.1647856
4	0.8869565	0.1130435

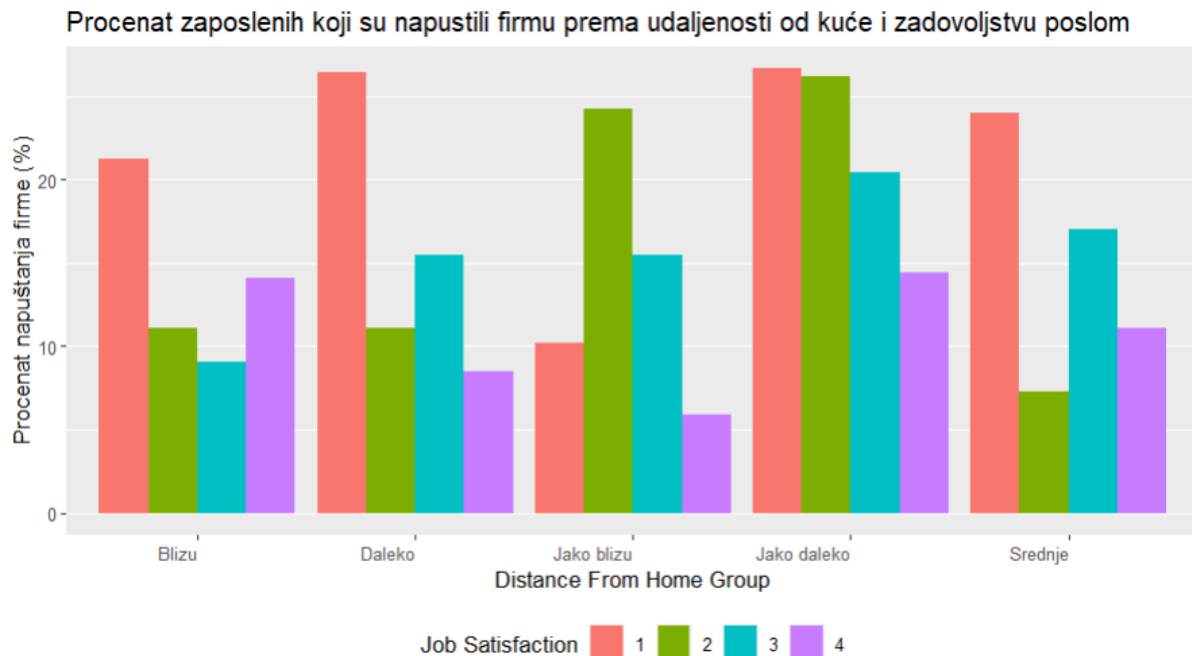
Запослени који су незадовољни послом чешће напуштају фирму, али то није превише изражено.

- Distance from home и Job satisfaction за запослене који напуштају фирму (Attrition-Yes)

```
data <- hr_data %>%
  group_by(DistanceFromHomeGroup, JobSatisfaction, Attrition) %>%
  summarise(count = n()) %>%
  mutate(percent = count / sum(count) * 100) %>%
  filter(Attrition == "Yes") %>%
  arrange(DistanceFromHomeGroup, JobSatisfaction)

bar_plot <- ggplot(data, aes(x = DistanceFromHomeGroup, y = percent, fill = JobSatisfaction)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Procenat zaposlenih koji su napustili firmu prema udaljenosti od kuće i zadovoljstvu poslom",
       x = "Distance From Home Group",
       y = "Procenat napuštanja firme (%)",
       fill = "Job Satisfaction") +
  theme(axis.text.x = element_text(hjust = 1),
        legend.position = "bottom",
        panel.grid.major.x = element_blank())

bar_plot
```

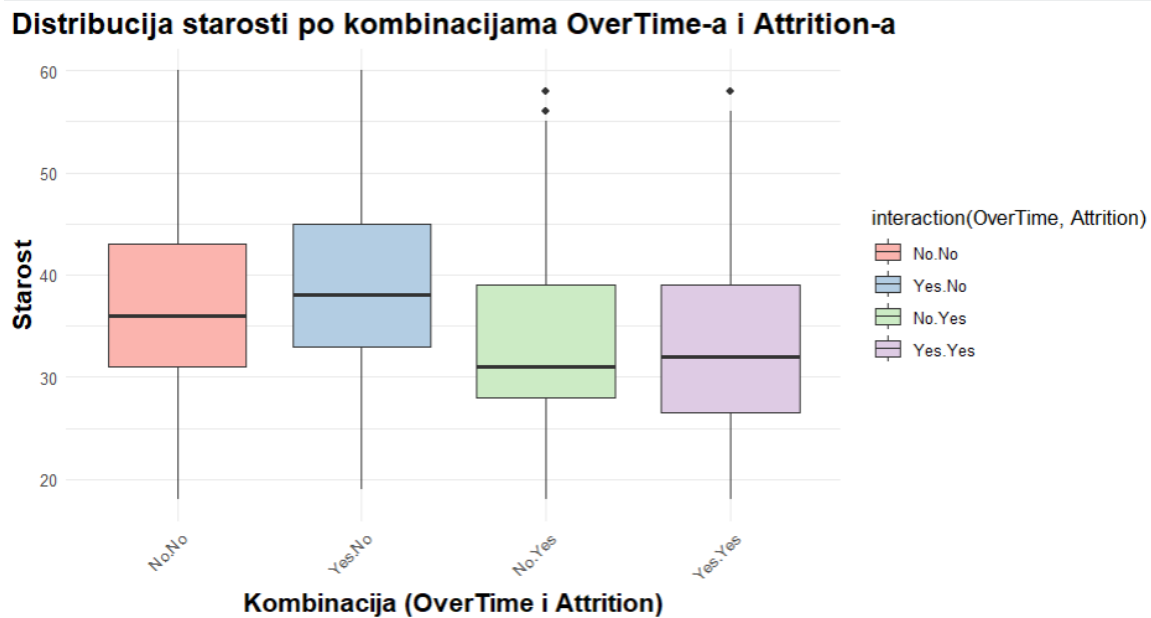


Из овог графика можемо да закључимо да запослени који живе јако далеко од фирме чешће напуштају исту. Такође су то у највећем проценту запослени који нису задовољни послом (ниво 1 и 2). Удаљеност од посла од куће може бити јако добар предиктор.



- **Age, OverTime и Attrition**

```
ggplot(hr_data, aes(x = interaction(OverTime, Attrition), y = Age, fill = interaction(OverTime, Attrition))) +
  geom_boxplot() +
  labs(title = "Distribucija starosti po kombinacijama OverTime-a i Attrition-a",
       x = "Kombinacija (OverTime i Attrition)",
       y = "Starost") +
  scale_fill_brewer(palette = "Pastel1") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title = element_text(size = 14, face = "bold")
  )
)
```



Примећујемо да комбинација предиктора прековремени рад и напуштање посла у односу на старост запосленог може бити корисна за креирање модела. Старије особе које раде прековремено не напуштају фирму, док млађе особе напуштају фирму без обзира да ли раде прековремено или не.

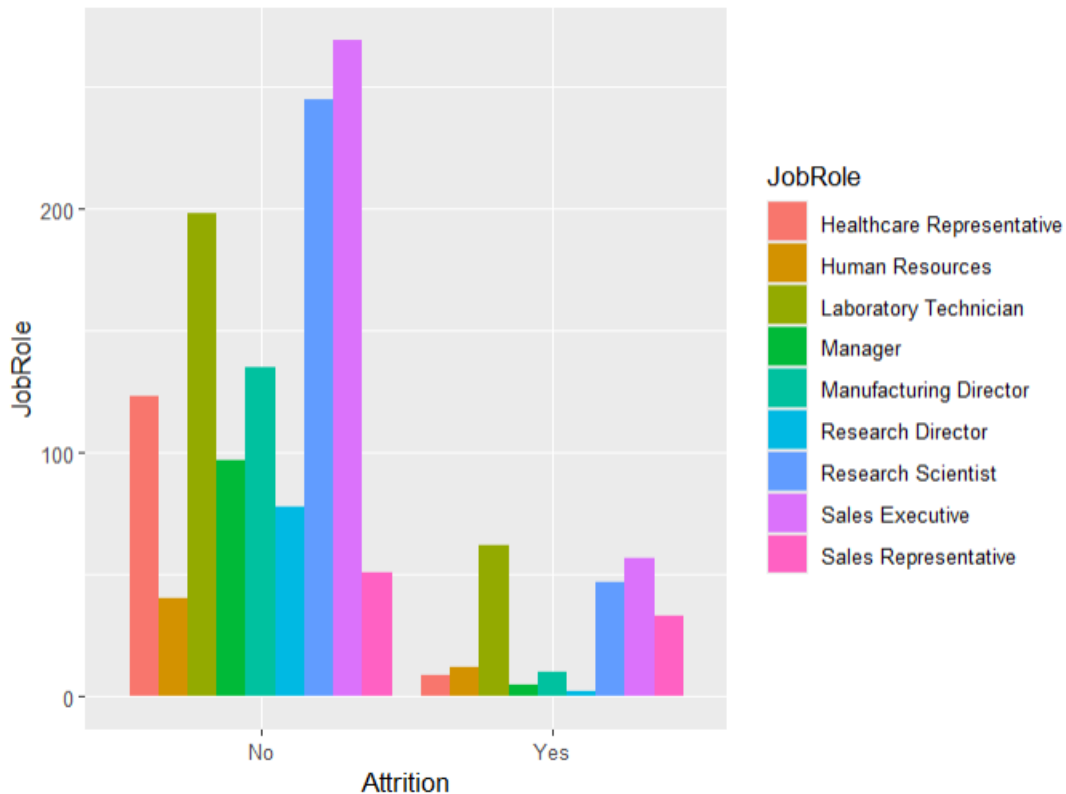
Постоји неколико аутлајер-а у категоријама запослених који напуштају фирму, што указује да постоје и случајеви старијих у тим категоријама.

- Job role vs Attrition

```
ggplot(hr_data, aes(x = factor(Attrition), fill = JobRole)) +
  geom_bar(position = "dodge") +
  ylab("JobRole") +
  xlab("Attrition") +
  scale_x_discrete(labels = c("No", "Yes"))

jobRole.attrition <- xtabs(~ JobRole + Attrition, data = hr_data)
jobRole.attrition

jobRole.attrition.prop <- prop.table(jobRole.attrition, margin = 1)
jobRole.attrition.prop
```



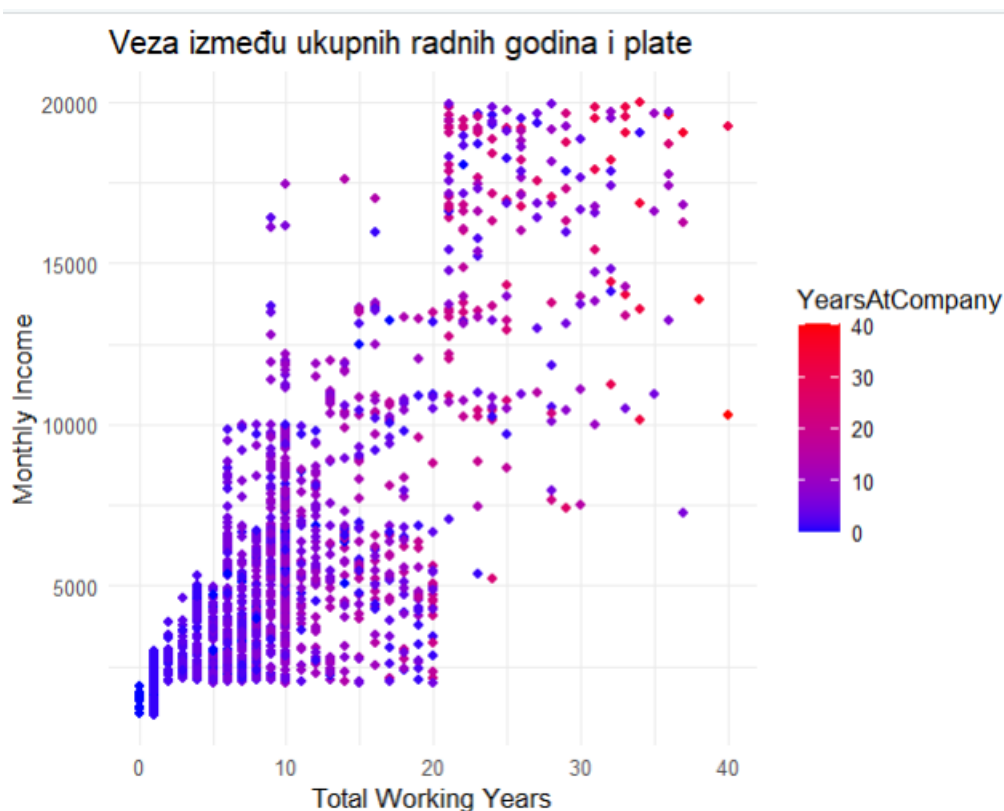
JobRole	Attrition	
	No	Yes
Healthcare Representative	0.93181818	0.06818182
Human Resources	0.76923077	0.23076923
Laboratory Technician	0.76153846	0.23846154
Manager	0.95098039	0.04901961
Manufacturing Director	0.93103448	0.06896552
Research Director	0.97500000	0.02500000
Research Scientist	0.83904110	0.16095890
Sales Executive	0.82515337	0.17484663
Sales Representative	0.60714286	0.39285714

Позиције које у највећем проценту напуштају фирму јесу Sales Representative (39.28%), Laboratory Technician (23.84%) и Human Resources (23.07%).

Са друге стране позиције попут Research Director (2.5%), Manager (4.9%) и Manufacturing Director (6.89%) запослени скоро па не напуштају, што је и логично јер су то највише позиције.

- **Total working years и Years at company према Monthly Income**

```
ggplot(hr_data, aes(x = TotalWorkingYears, y = MonthlyIncome, color = YearsAtCompany)) +  
  geom_point() +  
  scale_color_gradient(low = "blue", high = "red") +  
  labs(title = "Veza između ukupnih radnih godina i plate",  
        x = "Total Working Years",  
        y = "Monthly Income") +  
  theme_minimal()
```



Са следећег графика закључујемо да посотји веза између радног стажа и плате. Пораст година радног стажа прати и пораст плате. Закључујемо

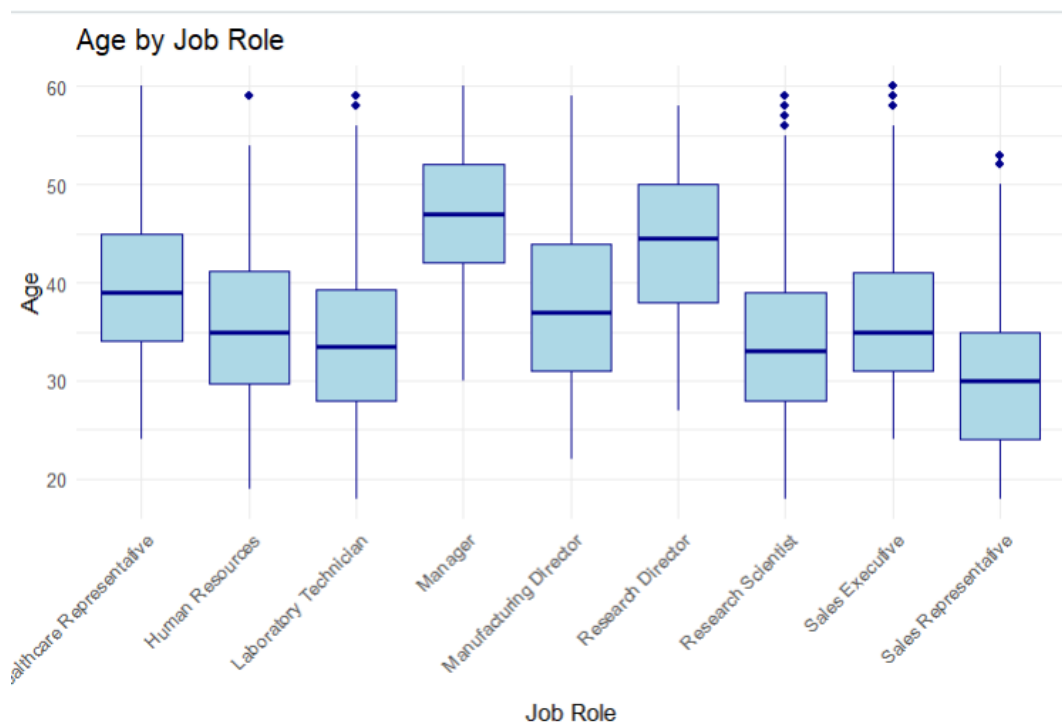
и да број година у тренутној фирми не гарантује већу плату, већ је битније укупно радно искуство.

Након 20-те године радног стажа плате углавном достижу горњу границу (15000-20000).

Још један закључак са графика јесте да запослених са мањим радним искуством (0-10) има више и да имају ниже плате.

- **Job role и Age**

```
ggplot(hr_data, aes(x = JobRole, y = Age)) +  
  geom_boxplot(fill = "lightblue", color = "darkblue") +  
  labs(title = "Age by Job Role", x = "Job Role", y = "Age") +  
  theme_minimal() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

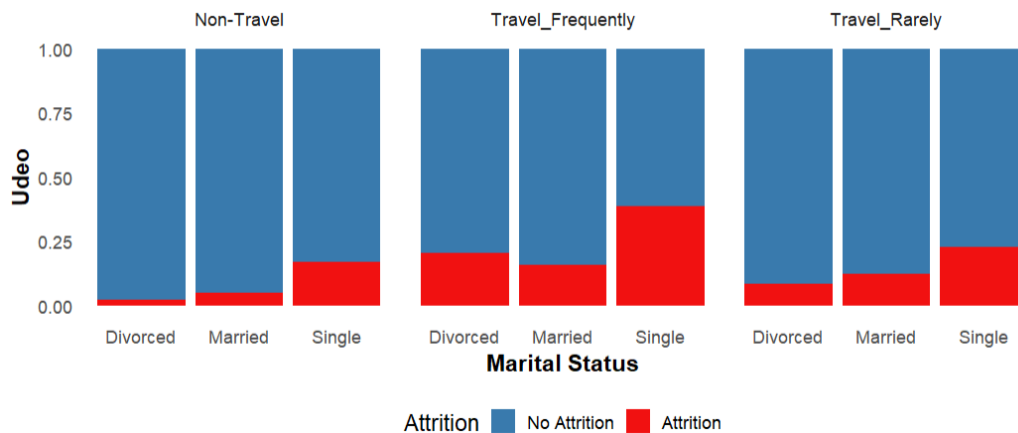


За више позиције у оквиру фирме, потребне су старије особе, млађе особе најчешће раде на нижим позивијама које се чешће напуштају.

- **Marital status, Business Travel и Attrition**

```
summary_data <- hr_data %>%
  group_by(MaritalStatus, BusinessTravel, Attrition) %>%
  summarise(Count = n(), .groups = 'drop')

summary_data %>% ggplot(aes(x = MaritalStatus, y = Count, fill = factor(Attrition))) +
  geom_bar(stat = "identity", position = "fill") + # Stacked position
  labs(title = "Udeo Attrition prema Marital Status i Business Travel kolona",
       x = "Marital Status",
       y = "Udeo",
       fill = "Attrition") +
  scale_fill_manual(values = c("Yes" = "#F11111", "No" = "#397AAD"),
                    labels = c("Yes" = "Attrition", "No" = "No Attrition")) +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    axis.title = element_text(face = "bold"),
    legend.position = "bottom",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  ) +
  facet_wrap(~ BusinessTravel)
```



Са графика можемо закључити да посао најчешће напуштају запослени који често путују, а најчешће они који су слободни. Можемо и закључити и да неvezано за учесталост путовања, слободни запослени најчешће напуштају, што је и логично.

У наставку се налази још један график који нам је помогао да донесемо ове закључке.

```
ggplot(hr_data, aes(x = BusinessTravelFactor, y = MaritalStatus, color = Attrition)) +
  geom_jitter() + theme_minimal() +
  labs(title = "Одлазак запослених у односу на брачни статус и учесталост путовања") +
  theme(plot.title = element_text(hjust = 0.5))
```

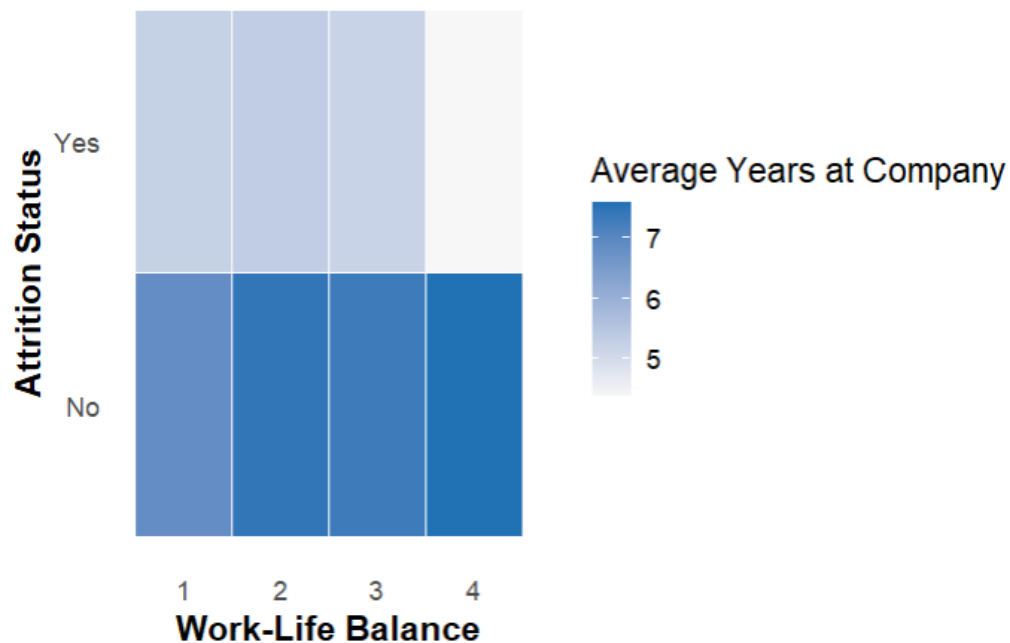
### Одлазак запослених у односу на брачни статус и учесталост путовања



- **Work life balance, Average years и Attrition**

```
summary_data <- hr_data %>%
  group_by(WorkLifeBalance, Attrition) %>%
  summarise(AverageYears = mean(YearsAtCompany, na.rm = TRUE), .groups = 'drop')

ggplot(summary_data, aes(x = WorkLifeBalance, y = factor(Attrition), fill = AverageYears)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "#f7f7f7", high = "#2171b5") +
  labs(title = "Heatmap of Average Years at Company by Work-Life Balance and Attrition",
       x = "Work-Life Balance",
       y = "Attrition Status",
       fill = "Average Years at Company") +
  theme_minimal(base_size = 15) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()
  )
```



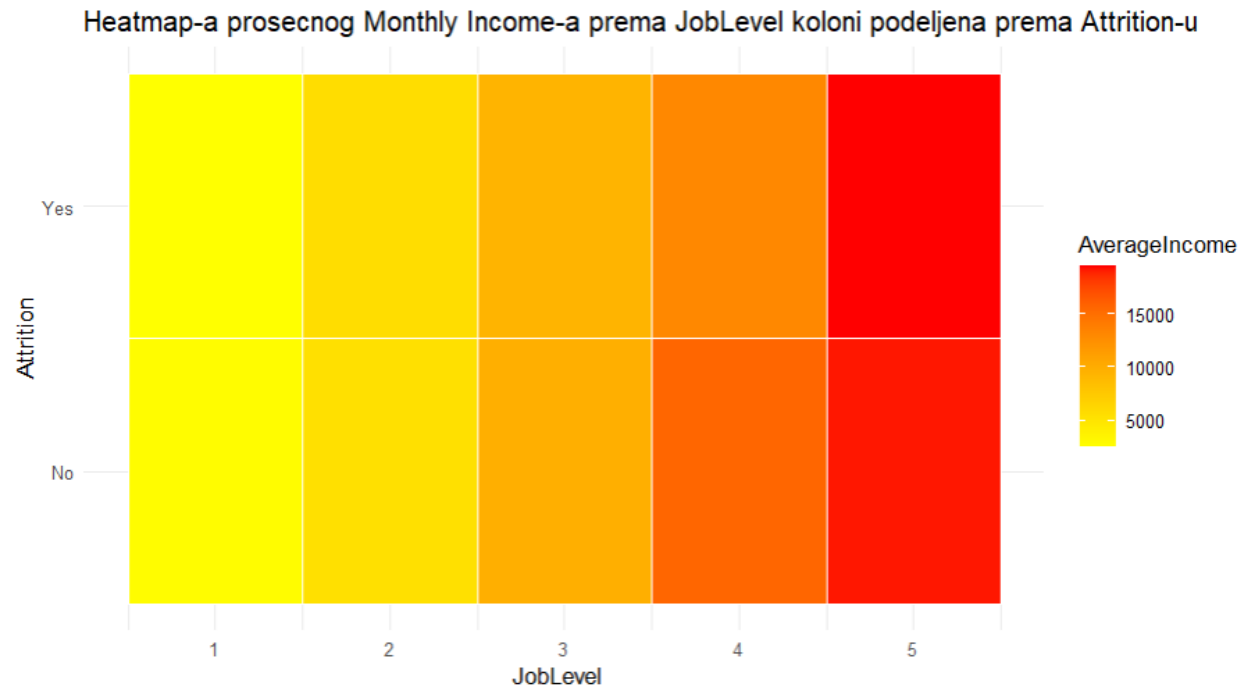
Закључујемо да запослени који су пронашли добар баланс дуже остају у фирми, што значи и да је мања вероватноћа да ће је напустити.

- Job level и Monthly income

```
heatmap_data <- hr_data %>%
  group_by(JobLevel , Attrition) %>%
  summarise(AverageIncome = mean(MonthlyIncome, na.rm = TRUE)) %>%
  ungroup()

ggplot(heatmap_data, aes(x = JobLevel , y = Attrition, fill = AverageIncome)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "yellow", high = "red") +
  labs(title = "Heatmap-a prosecnog Monthly Income-a prema JobLevel koloni podeljena prema Attrition-u",
        x = "JobLevel ",
        y = "Attrition") +
  theme_minimal()

cor(hr_data$JobLevel, hr_data$MonthlyIncome, method = "spearman")
```



На основу графика закључујемо да више позиције имају већа примања, али и такође да пораст месечних примања јесте пропорцијалан расту висини позиције и да ове две колоне имају велику корелацију која износи 0.919878, због чега колону JobLevel нећемо узимати у обзир у даљој анализи.



## 4. Креирање модела

Пре креирања модела, потребно је извршити поделу скупа података на податке за тренинг као и податке за валидацију модела.

```
train_index <- createDataPartition(hr_data.final$Attrition_binary,  
                                   p = 0.8,  
                                   list = FALSE)  
train_data <- hr_data.final[train_index, ]  
test_data <- hr_data.final[-train_index, ]
```

Као први корак користимо команду „set.seed(int)“ која нам помаже да реплицирамо случајни избор тестних и тренинг података.

Податке делимо уз помоћ функције „*createDataPartition*“ из библиотеке „*caret*“. Уз помоћ ове функције поделили смо, случајним одабиром, овај скуп на два дела у односу 80:20. Где 80 процената скупа припада валидационом скупу док 20 тестном. Овај проценат смо одабрали због ограниченог обима података у нашем скупу. Потребно је обезбедити више података како би модел постао функционалан и пружио прецизније резултате.

Развијаћемо 4 модела: модел логистичке регресије, стабло одлучивања, „*Random Forest*“ модел и модел Неуронских мрежа.

Користићемо различите предикторе за сва три модела како бисмо постигли оптималне резултате. Као праг сигурности модела поставићемо вредност од 0,5, а за приказ резултата модела користићемо матрицу конфузије.

## 4.1 Логистичка регресија

За предикторе овог модела изабрали смо колоне „Age“, „OverTime“ и „YearsInCurrentRole“.

```
model1_formula <- Attrition_binary ~ OverTime + Age + YearsInCurrentRole
log_reg_model <- glm(model1_formula, data = train_data, family = binomial)

y_pred_model1 <- predict(log_reg_model, test_data, type = "response")
y_pred_model1_class <- ifelse(y_pred_model1 > 0.5, 1, 0)

confusionMatrix(as.factor(y_pred_model1_class), as.factor(test_data$Attrition_binary))
```

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	246	43
1	1	4

Accuracy : 0.8503  
95% CI : (0.8043, 0.8891)  
No Information Rate : 0.8401  
P-Value [Acc > NIR] : 0.3511

Kappa : 0.127

Mcnemar's Test P-Value : 6.37e-10

Sensitivity : 0.99595  
Specificity : 0.08511  
Pos Pred Value : 0.85121  
Neg Pred Value : 0.80000  
Prevalence : 0.84014  
Detection Rate : 0.83673  
Detection Prevalence : 0.98299  
Balanced Accuracy : 0.54053

'Positive' Class : 0

*1 Резултати логистичког модела*

Према резултатима модела видимо да овај модел има високу прецизност као и сензитивност док је специфичност веома ниска. Балансирана прецизност модела износи 54% што је поприлично ниско у односу на пријављену прецизност.

Један од разлога за ове резултате представља то да је у тренинг скуп ушло само 5 особа које су напустиле компанију. Ово је веома мали узорак података на коме модел није успео да уочи довољно карактеристика.

#### 4.1.1 Логистички модел и унакрсна валидација

Како бисмо побољшали резултате претходног модела можемо искористи десетоструку унакрсну валидацију. Као још један параметар овој функцији можемо проследи и параметар „*sampling*“ с вредношћу „*up*“ која нам помаже да избалансирамо податке.

```
> control <- trainControl(method = "cv", number = 10, sampling = "up")
> model1_formula <- Attrition_binary ~ OverTime + Age + YearsInCurrentRole
> logistic_model <- train(model1_formula, data = train_data, method = "glm", family = "binomial", trControl = control)
> y_pred_model1 <- predict(logistic_model, test_data)
> confusionMatrix(y_pred_model1, as.factor(test_data$Attrition_binary))
```

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	185	17
1	62	30

Accuracy : 0.7313

95% CI : (0.6768, 0.7811)

No Information Rate : 0.8401

P-Value [Acc > NIR] : 1

Kappa : 0.2791

McNemar's Test P-Value : 7.407e-07

Sensitivity : 0.7490

Specificity : 0.6383

Pos Pred Value : 0.9158

Neg Pred Value : 0.3261

Prevalence : 0.8401

Detection Rate : 0.6293

Detection Prevalence : 0.6871

Balanced Accuracy : 0.6936

'Positive' Class : 0

У поређењу са првим моделом, модел са унакрсном валидацијом пружа бољу осетљивост и специфичност. Специфичност је знатно побољшана самим тим модел пружа прецизније резултате.

Иако је укупна тачност опала, *Карра* вредност је значајно боља што се осликава у томе да модел боље рефлектује стварне податке.

Следећи корак може бити смањење прага (енг. „*threshold*“) како бисмо повећали осетљивост модела.

## 4.2 GLM модел

Како бисмо постигли боље резултате, можемо размотрити коришћење GLM модела који нуди већу флексибилност у односу на логистички модел.

Такође, пробаћемо са другачијим предикторима, у овом случају „*OverTime*“, „*MonthlyIncome*“, „*JobInvolvement*“.

```
model2_formula <- Attrition_binary ~ OverTime + MonthlyIncome + JobInvolvement  
decision_tree_model <- train(model2_formula, data = train_data, method = "rpart")  
y_pred_model2 <- predict(decision_tree_model, test_data)  
confusionMatrix(y_pred_model2, as.factor(test_data$Attrition_binary))
```

### Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	200	27
1	47	20

Accuracy : 0.7483  
95% CI : (0.6946, 0.7969)  
No Information Rate : 0.8401  
P-Value [Acc > NIR] : 1.0000

Kappa : 0.2007

Mcnemar's Test P-Value : 0.0272

Sensitivity : 0.8097  
Specificity : 0.4255  
Pos Pred Value : 0.8811  
Neg Pred Value : 0.2985  
Prevalence : 0.8401  
Detection Rate : 0.6803  
Detection Prevalence : 0.7721  
Balanced Accuracy : 0.6176

'Positive' Class : 0

Модел показује прецизност од 74% али Карра вредност је ниска док је специфичност боља. Матрица конфузије потврђује да модел није успео да препозна ни један случај напуштања запослених, што указује на проблем у препознавању ове класе.

С обзиром да овај модел пружа знатно лошије резултате него претходни, можемо проверити да ли је проблем у моделу или је у предикторима. Покушаћемо да уз помоћ истих предиктора као у логистичком модела дођемо до бољих резултата.

```
model2_formula <- Attrition_binary ~ OverTime + Age + YearsInCurrentRole
decision_tree_model <- train(model2_formula, data = train_data, method = "rpart")
y_pred_model2 <- predict(decision_tree_model, test_data)
confusionMatrix(y_pred_model2, as.factor(test_data$Attrition_binary))
```

#### Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	234	33
1	13	14

Accuracy : 0.8435  
95% CI : (0.7969, 0.8831)  
No Information Rate : 0.8401  
P-Value [Acc > NIR] : 0.475465

Kappa : 0.2963

Mcnemar's Test P-Value : 0.005088

Sensitivity : 0.9474  
Specificity : 0.2979  
Pos Pred Value : 0.8764  
Neg Pred Value : 0.5185  
Prevalence : 0.8401  
Detection Rate : 0.7959  
Detection Prevalence : 0.9082  
Balanced Accuracy : 0.6226

'Positive' Class : 0

У овом случају добијамо мало боље резултате. Покушаћемо, исто као и са првим моделом, да уз помоћ десетоструке унакрсне валидације и *up-sampling-a* дођемо до бољих резултата.

### 4.2.1 GLM модел и унакрсна валидација

```
control <- trainControl(method = "cv", number = 10, sampling = "up")
model2_formula <- Attrition_binary ~ OverTime + Age + YearsInCurrentRole
decision_tree_model <- train(model2_formula, data = train_data, method = "rpart", trControl = control)
y_pred_model2 <- predict(decision_tree_model, test_data)
confusionMatrix(y_pred_model2, as.factor(test_data$Attrition_binary))
```

#### Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	175	20
1	72	27

Accuracy : 0.6871  
95% CI : (0.6307, 0.7397)  
No Information Rate : 0.8401  
P-Value [Acc > NIR] : 1

Kappa : 0.1954

Mcnemar's Test P-Value : 1.054e-07

Sensitivity : 0.7085  
Specificity : 0.5745  
Pos Pred Value : 0.8974  
Neg Pred Value : 0.2727  
Prevalence : 0.8401  
Detection Rate : 0.5952  
Detection Prevalence : 0.6633  
Balanced Accuracy : 0.6415

'Positive' Class : 0

Као и код логистичко модела овде видимо знатно боље резултате у погледу специфичност и могућности модела да детектује запослене који су напустили компанију.

Прецизност овог модела износи 68.71%, док *Sensitivity* износи 70.85% а „*Specifity*“ је 57.45%. У овом случају *Kappa* вредност износ 0.195 што је ниже него побољшани први модел.

## 4.3 Random Forest модел

Знајући да *Random Forest* може боље обрадити сложеније податке и да већ користи унакрсну валидацију у сваком стаблу, што нам може помоћи да дођемо до боље генерализације, следећи корак ће нам бити управо креирање оваквог модела.

```
model3_formula <- Attrition_binary ~ OverTime + Age + YearsInCurrentRole
random_forest_model <- randomForest(model3_formula, data = train_data)
y_pred_model3 <- predict(random_forest_model, test_data)
confusionMatrix(y_pred_model3, as.factor(test_data$Attrition_binary))
```

### Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	246	44
1	1	3

Accuracy : 0.8469  
95% CI : (0.8006, 0.8861)  
No Information Rate : 0.8401  
P-Value [Acc > NIR] : 0.4123

Kappa : 0.095

Mcnemar's Test P-Value : 3.825e-10

Sensitivity : 0.99595  
Specificity : 0.06383  
Pos Pred Value : 0.84828  
Neg Pred Value : 0.75000  
Prevalence : 0.84014  
Detection Rate : 0.83673  
Detection Prevalence : 0.98639  
Balanced Accuracy : 0.52989

'Positive' Class : 0

Видимо да поново долазимо до истих мана модела па ћемо онда одмах прећи на унакрсну валидацију уз примену *sampling-up* технике. Иако овај модел већ

користи крос валидацију, она нам може помоћи у смањењу варијансе и пружити нам флексибилност у подацима, што је у нашем случају веома битно с обзиром на неуравнотеженост података.

#### 4.3.1 Random Forest модел и унакрсна валидација

```
control <- trainControl(method = "cv", number = 10, sampling = "up")
model3_formula <- Attrition_binary ~ OverTime + Age + YearsInCurrentRole
random_forest_model <- train(model3_formula, data = train_data, method = "rf", trControl = control)
y_pred_model3 <- predict(random_forest_model, test_data)
confusionMatrix(y_pred_model3, as.factor(test_data$Attrition_binary))
```

##### Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	190	24
1	57	23

Accuracy : 0.7245  
95% CI : (0.6696, 0.7748)  
No Information Rate : 0.8401  
P-Value [Acc > NIR] : 0.9999998

Kappa : 0.2014

Mcnemar's Test P-Value : 0.0003772

Sensitivity : 0.7692  
Specificity : 0.4894  
Pos Pred Value : 0.8879  
Neg Pred Value : 0.2875  
Prevalence : 0.8401  
Detection Rate : 0.6463  
Detection Prevalence : 0.7279  
Balanced Accuracy : 0.6293

'Positive' Class : 0

Као и у свим претходним случајевима, видимо опадање прецизности и осетљивости модела али пораст свих других вредности.



## 4.4 Неуронске мреже

Неуронске мреже су показале највећу прецизност у поређењу са осталим моделима. Резултати показују да мрежа најбоље препознаје запослене који напуштају компанију, са сензитивношћу од 81,78% и прецизношћу од 79,59%. Упркос dobrim резултатима, специфичност је и даље нешто нижа (68,09%).

```
library("nnet")

control <- trainControl(method = "cv", number = 6, sampling = "up") ## 10, 8

grid <- expand.grid(
  size = c(2, 4, 6),
  decay = c(0.1, 0.01, 0.001)
)

set.seed(123)
nn_model <- train(
  Attrition_binary ~ .,
  data = train_data,
  method = "nnet",
  trControl = control,
  tuneGrid = grid,
  trace = FALSE
)

print(nn_model)
```

Resampling results across tuning parameters:

size	decay	Accuracy	Kappa
2	0.001	0.4731857	0.09800358
2	0.010	0.5175612	0.13108191
2	0.100	0.6893127	0.27883134
4	0.001	0.6363379	0.15744148
4	0.010	0.7591529	0.30845080
4	0.100	0.7641426	0.33466252
6	0.001	0.7405796	0.25438569
6	0.010	0.6506070	0.21252627
6	0.100	0.7683902	0.32903025

Accuracy was used to select the optimal model using the largest value.  
The final values used for the model were size = 6 and decay = 0.1.

```

predictions <- predict(nn_model, newdata = test_data)

confusionMatrix(as.factor(predictions), as.factor(test_data$Attrition_binary))

```

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	202	15
1	45	32

```

          Accuracy : 0.7959
          95% CI   : (0.7453, 0.8405)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.9817550

          Kappa : 0.3963

McNemar's Test P-Value : 0.0001812

```

```

          Sensitivity : 0.8178
          Specificity : 0.6809
          Pos Pred Value : 0.9309
          Neg Pred Value : 0.4156
          Prevalence : 0.8401
          Detection Rate : 0.6871
    Detection Prevalence : 0.7381
    Balanced Accuracy : 0.7493

```

```

'Positive' Class : 0

```

```

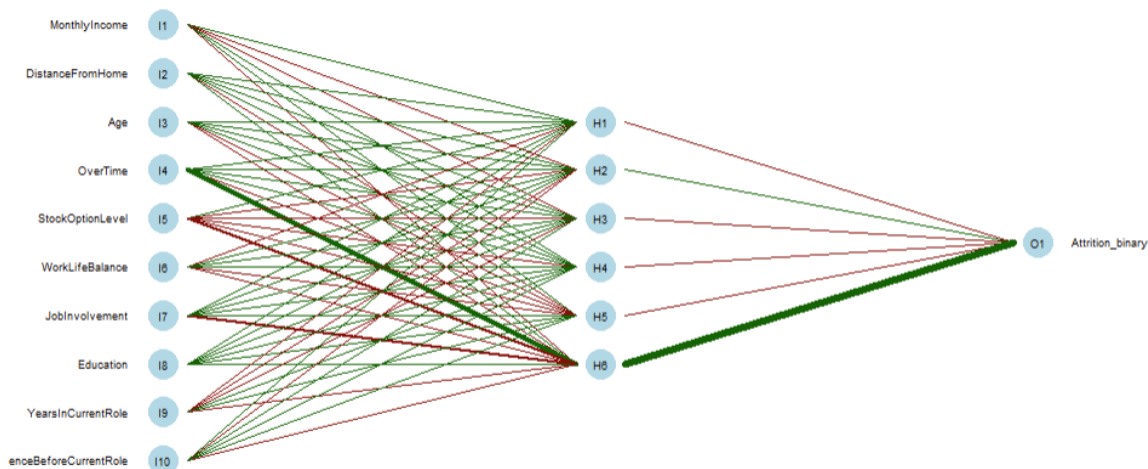
library(DescTools)
library(NeuralNetTools)

```

```

plotnet(mod_in = nn_model, # nnet object
  pos_col = "darkgreen", # positive weights are shown in green
  neg_col = "darkred", # negative weights are shown in red
  bias = FALSE, # do not plot bias
  circle_cex = 4, # reduce circle size (default is 5)
  cex_val = 0.6)

```



На графику је приказана структура неуронске мреже која се користила у моделу. Мрежа се састоји од улазног слоја, једног скривеног слоја са 10 неурона и излазног слоја. Сваки чвор (неурон) у скривеном слоју повезан је са свим чворовима у претходном и следећем слоју.

Као најважнији предиктори показали су се „OverTime“ и „MonthlyIncome“, док су остали предиктори допринели мањем степену варијације у резултатима.

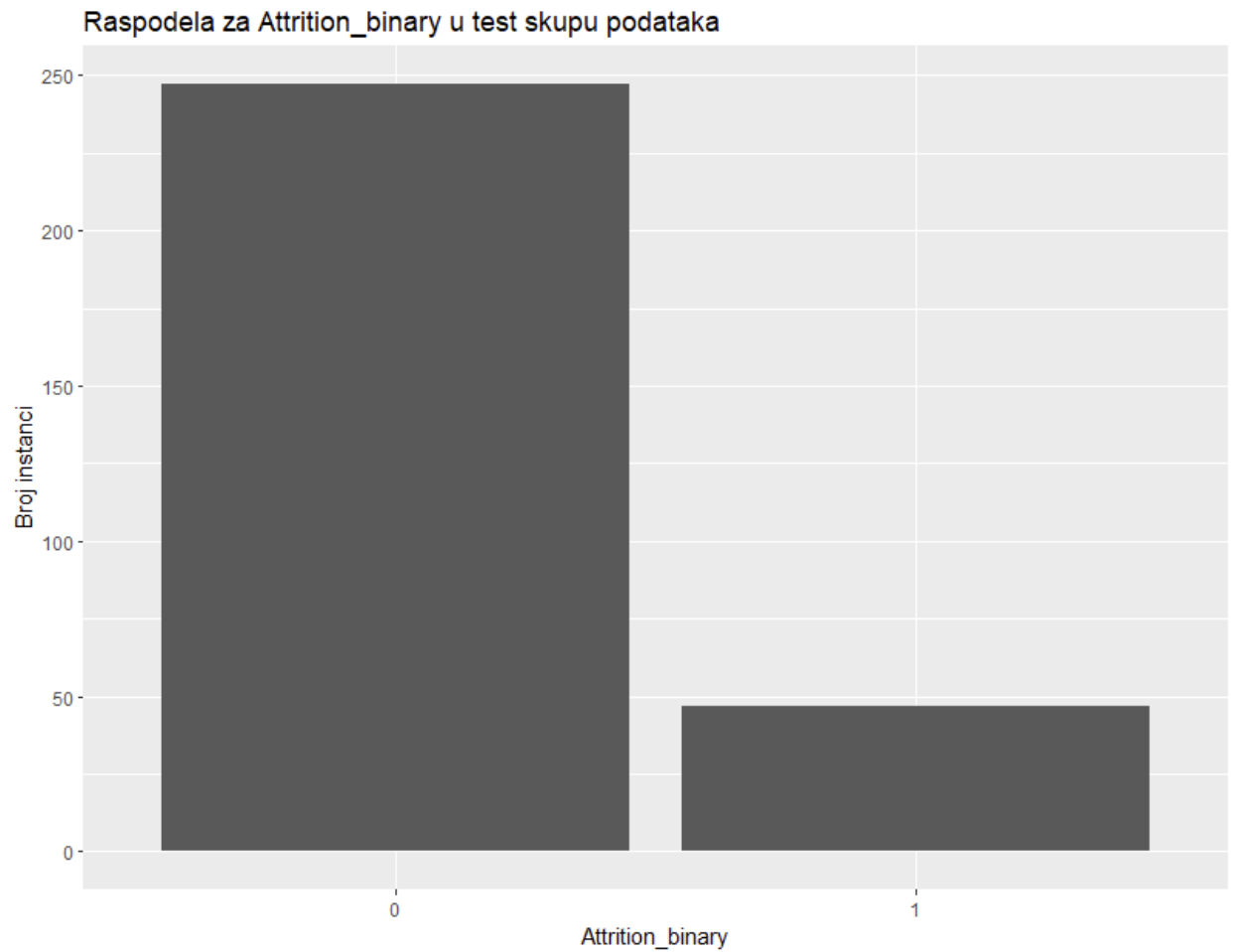
## 4.5 Резултати модела

Да бисмо упоредили ова три модела на једноставан начине, направићемо табелу у којој ће колоне садржати резултате појединачних модела а редови одговарајуће метрике.

	Логистичка регресија	ГЛМ модел	Random forest	Неуронске мреже
<b>Прецизност</b>	73,13%	68,71%	72,45%	79,59%
<b>Балансирана прецизност</b>	69,36%	64,15%	62,93%	74,93%
<b>Сензитивност</b>	74,9%	70,85%	76,92%	81,78%
<b>Специфичност</b>	63,83%	57,45%	48,94%	68,09%
<b>Карра вредност</b>	0,2791	0,1954	0,2014	0,3963

Неуронске мреже се показују као најбољи модел према свим кључним метрикама.

## 4.6 Resampling методе



Имамо приближно 84% негативних случајева и 16% позитивних што указује на небалансиране податке колоне Attrition. Кључни проблем небалансираних података може бити пристрасност модела према већинској класи и имати потешкоћа у идентификацији мањинске.

Покушаћемо да добијемо прецизније моделе коришћењем Resampling метода из ROSE библиотеке.

с

### 4.6.1 Oversampling

Коришћењем методе oversampling-a ћемо уједначити број негативних узорака са бројем позитивних.

```
> train_data <- ROSE(Attrition_binary ~ ., data = train_data, seed = 43)$data
> #provera odnosa nakon oversampling-a
> table(train_data$Attrition_binary)
```

```
  0    1
596 583
```

#### 4.6.2 Undersampling метода

Коришћењем ове методе прилагодили смо број узорака класе са позитивним вредностима броју негативних.

```
> train_data <- ovun.sample(Attrition_binary ~ ., data = train_data, method = "under", N = 380)$data
> #provera odnosa nakon undersampling-a
> table(train_data$Attrition_binary)
```

```
  0    1
190 190
```

#### 4.6.3 Oversampling и Undersampling

Овом методом смо избалансирали вредности класе и сада имамо скоро идентичне бројеве негативних и позитивних вредности.

```
> train_data <- ovun.sample(Attrition_binary ~ ., data = train_data, method = "both", p=0.5, N=1289, seed = 1)$data
> table(train_data$Attrition_binary)
```

```
  0    1
673 616
```

#### 4.6.4 Резултати

Resampling методе коришћене су за решавање проблема небалансираности података. Најпре је примењен oversampling, где је број узорака мањинске класе повећан како би се изједначио са већинском. Затим је примењен undersampling, где је смањен број већинске класе. Комбиновањем обе методе, постигнута је равнотежа класа.

Иако је баланс постигнут, примена resampling метода није значајно побољшала параметре модела. Наши модели показали су сличне резултате пре и након примене ових метода. Ово указује да небалансираност није била доминантан фактор у прецизности модела, већ да је квалитет података и избор предиктора имао значајнији утицај.

## 5. Закључак

Анализа је усмерена на предвиђање запослених код којих постоји највећа могућност да напусте компанију користећи податке Људских ресурса из претходно урађених анкета.

Израђени модели укључују логистичку регресију, *GLM* модел, модел *Random Forest* и модел неуронских мрежа. Ови модели пружили су нам увид у вероватноћу задржавања запослених са различитим процентима тачности.

Након упоређивања перформанси модела на основу кључних метрика, закључили смо да модел логистичке регресије даје најпрецизније резултате. Овај резултат сугерише да је однос између варијабли унутар скупа података линеаран.

## 6. Литература

- *Увод у науку о подацима* - вежбе, предавања и материјали предмета
- *GGPlot2* - [документација](#)
- *Caret* - [документација](#)
- *GLM* - [документација](#)
- *Neural Networks* - [документација](#)