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

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 својевољно 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.

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              Age          AgeGroup          Attrition         BusinessTravel       DailyRate       Department
 Length:1480        Min.   :18.00   Length:1480        Length:1480        Length:1480        Min.   : 102.0   Length:1480
 Class :character   1st Qu.:30.00   Class :character   Class :character   Class :character   1st Qu.: 465.0   Class :character
 Mode  :character   Median :36.00   Mode  :character   Mode  :character   Mode  :character   Median : 800.0   Mode  :character
                    Mean   :36.92                                                           Mean   : 801.4
                    3rd Qu.:43.00                                                           3rd Qu.:1157.0
                    Max.   :60.00                                                           Max.   :1499.0

 DistanceFromHome   Education     EducationField     EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender
 Min.   : 1.00    Min.   :1.000   Length:1480        Min.   :1     Min.   :   1.0   Min.   :1.000           Length:1480
 1st Qu.: 2.00    1st Qu.:2.000   Class :character   1st Qu.:1     1st Qu.: 493.8   1st Qu.:2.000           Class :character
 Median : 7.00    Median :3.000   Mode  :character   Median :1     Median :1027.5   Median :3.000           Mode  :character
 Mean   : 9.22    Mean   :2.911                      Mean   :1     Mean   :1031.9   Mean   :2.724
 3rd Qu.:14.00    3rd Qu.:4.000                      3rd Qu.:1     3rd Qu.:1568.2   3rd Qu.:4.000
 Max.   :29.00    Max.   :5.000                      Max.   :1     Max.   :2068.0   Max.   :4.000

   HourlyRate     JobInvolvement    JobLevel       JobRole          JobSatisfaction MaritalStatus      MonthlyIncome
 Min.   : 30.00   Min.   :1.00   Min.   :1.000   Length:1480        Min.   :1.000   Length:1480        Min.   : 1009
 1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Class :character   1st Qu.:2.000   Class :character   1st Qu.: 2922
 Median : 66.00   Median :3.00   Median :2.000   Mode  :character   Median :3.000   Mode  :character   Median : 4933
 Mean   : 65.85   Mean   :2.73   Mean   :2.065                      Mean   :2.725                      Mean   : 6505
 3rd Qu.: 83.00   3rd Qu.:3.00   3rd Qu.:3.000                      3rd Qu.:4.000                      3rd Qu.: 8384
 Max.   :100.00   Max.   :4.00   Max.   :5.000                      Max.   :4.000                      Max.   :19999

  SalarySlab         MonthlyRate    NumCompaniesWorked    Over18            OverTime         PercentSalaryHike PerformanceRating
 Length:1480        Min.   : 2094   Min.   :0.000      Length:1480        Length:1480        Min.   :11.00     Min.   :3.000
 Class :character   1st Qu.: 8051   1st Qu.:1.000      Class :character   Class :character   1st Qu.:12.00     1st Qu.:3.000
 Mode  :character   Median :14220   Median :2.000      Mode  :character   Mode  :character   Median :14.00     Median :3.000
                    Mean   :14298   Mean   :2.687                                           Mean   :15.21     Mean   :3.153
                    3rd Qu.:20461   3rd Qu.:4.000                                           3rd Qu.:18.00     3rd Qu.:3.000
                    Max.   :26999   Max.   :9.000                                           Max.   :25.00     Max.   :4.000

 RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
 Min.   :1.000            Min.   :80    Min.   :0.0000   Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
 1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
 Median :3.000            Median :80    Median :1.0000   Median :10.00     Median :3.000         Median :3.000   Median : 5.000
 Mean   :2.709            Mean   :80    Mean   :0.7919   Mean   :11.28     Mean   :2.798         Mean   :2.761   Mean   : 7.009
 3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
 Max.   :4.000            Max.   :80    Max.   :3.0000   Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000

 YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median : 3.000     Median : 1.000          Median : 3.000
 Mean   : 4.228     Mean   : 2.182          Mean   : 4.118
 3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :18.000     Max.   :15.000          Max.   :17.000
                                            NA's   :57
```

За анализу података креираћемо две функције („*plot_histogram*“ и „*plot_bar*“) 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()
}
```

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.

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. С обзиром да 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        TravelRarely           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
```

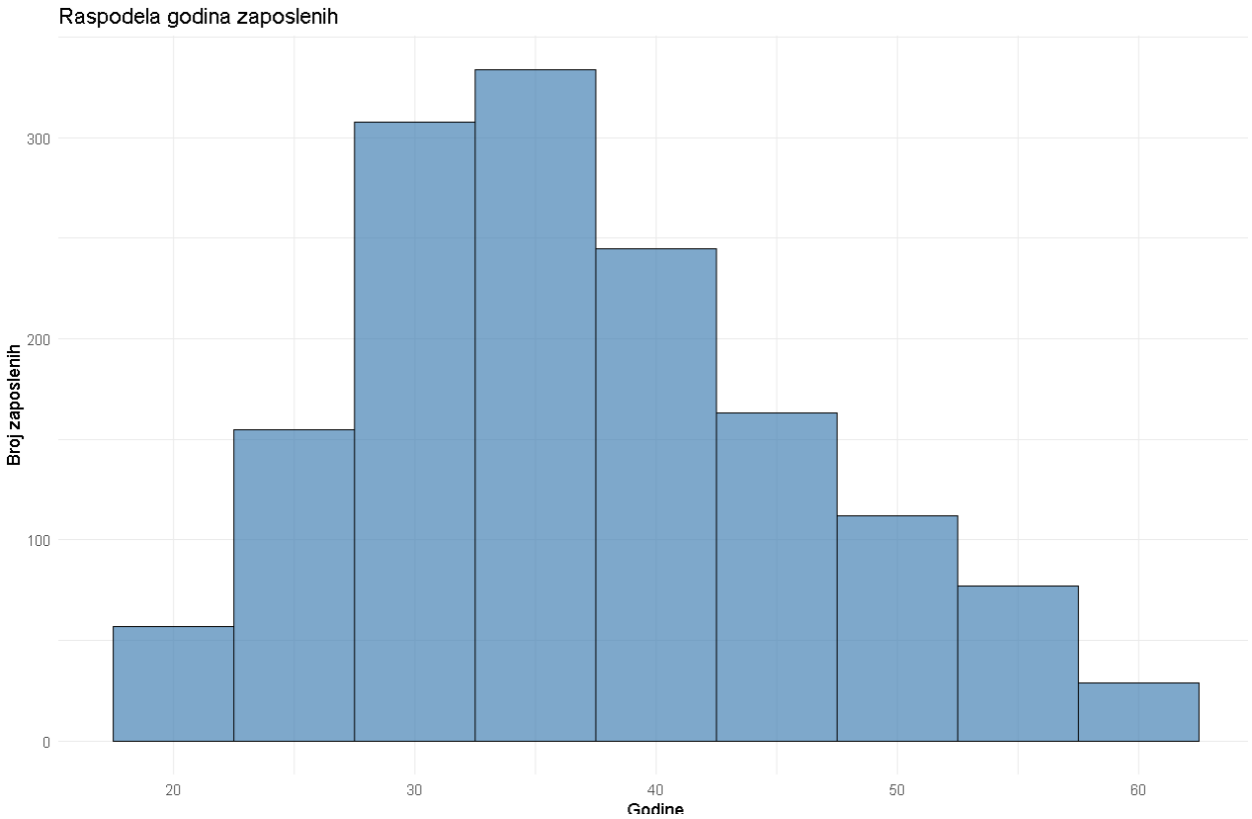
Успешно смо уклонили 7 редова, сада се у скупу налази 1473 јединствених редова података.

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.

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
```

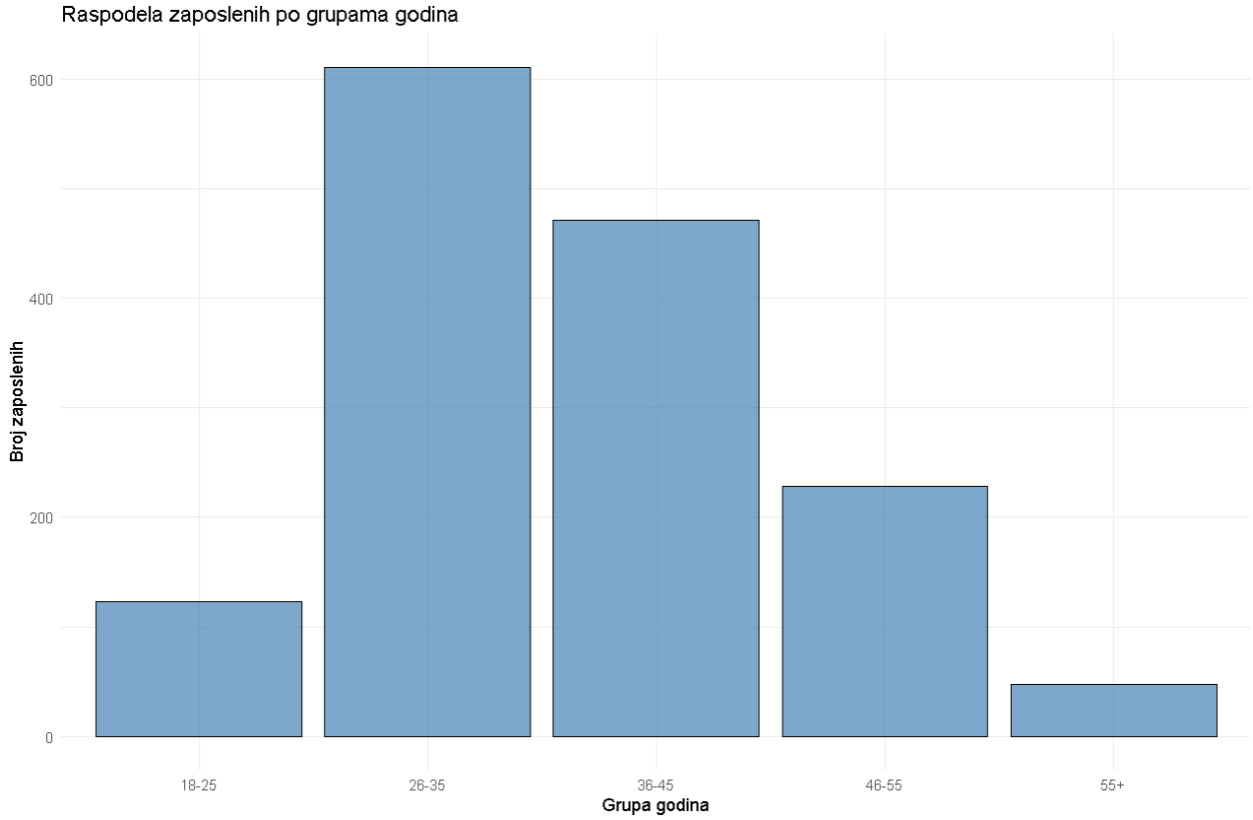


Колона „AgeGroup“

Представља опсег година у којем се налази запослени.

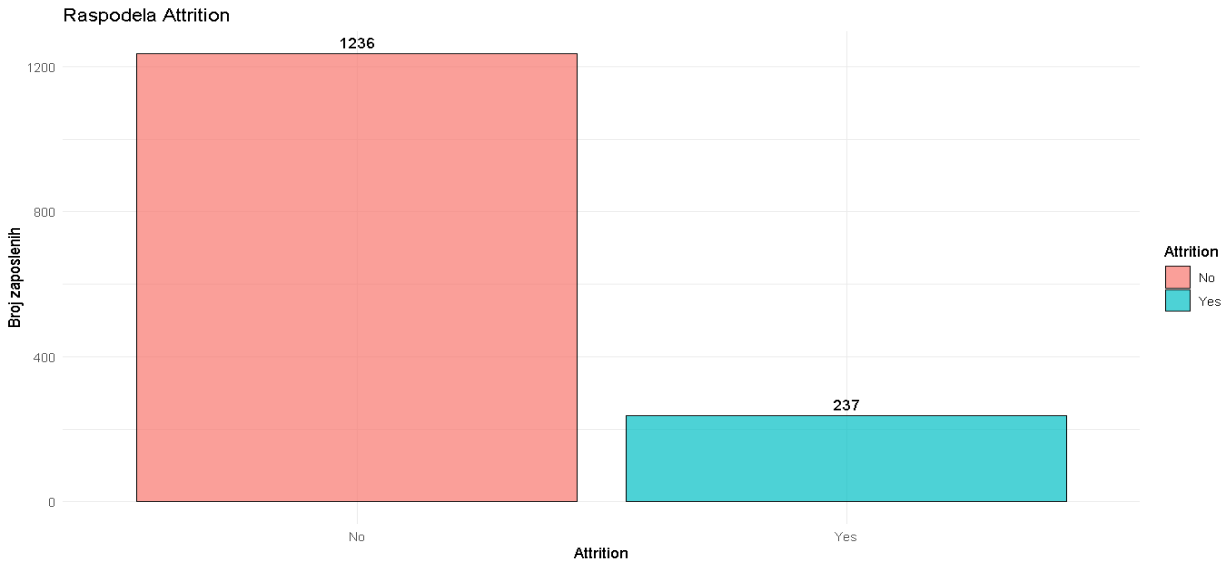
```
> print(unique_values_summary$AgeGroup_unique_values)
[1] 5
> print(unique_values_summary$AgeGroup_sample_values)
[1] "18-25, 26-35, 36-45, 46-55, 55+"
```

Садржи 5 различитих група: 18-25, 26-35, 36-45, 46-55, 55+. Расподелу по групама можемо прикази 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 флуктуације 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“

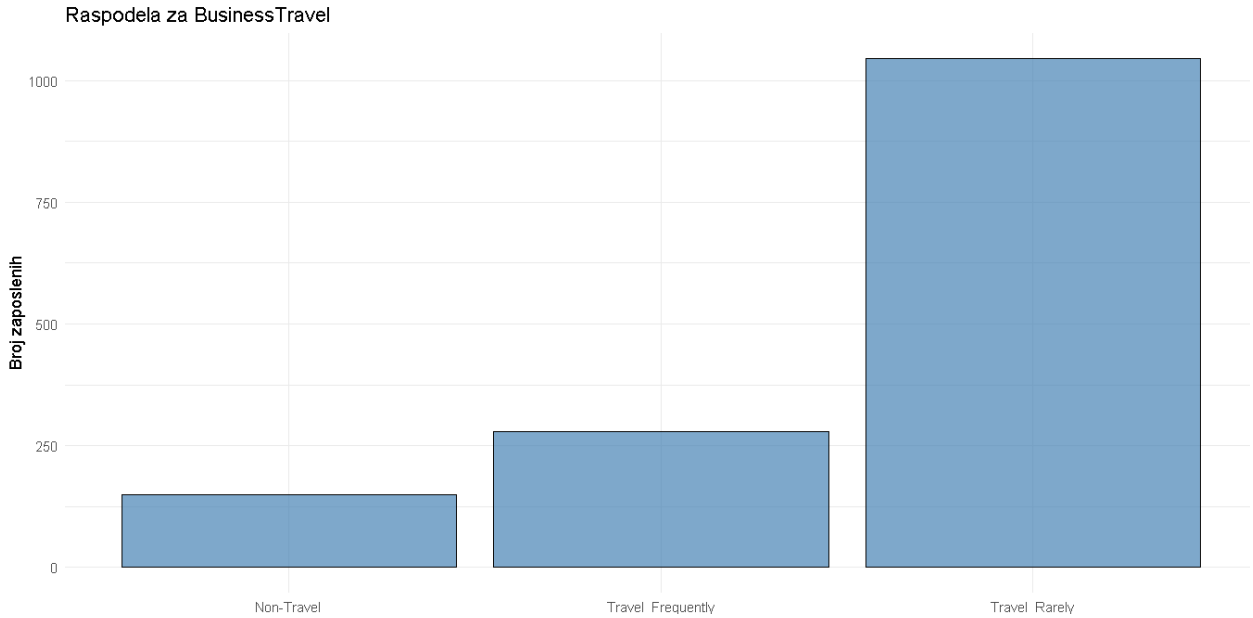
Представља информације о учесталости путовања запосленог. Садржи 4 могуће вредности: „Travel_Rarely“, „Travel_Frequently“, „Non-Travel“, „TravelRarely“.

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

Видимо да је потребно извршити трансформацију података с обзиром да постоје уноси за TravelRarely и Travel_Rarely који представљају исту учесталост.

```
> hr_data <- hr_data %>%
+   mutate(BusinessTravel = ifelse(BusinessTravel == "TravelRarely", "Travel_Rarely", BusinessTravel))
> unique_values_summary <- hr_data %>%
+   summarise(across(everything(),
+                    list(unique_values = ~ length(unique(.)),
+                         sample_values = ~ paste(unique(.)[1:5], collapse = ", "))))
> print(unique_values_summary$BusinessTravel_unique_values)
[1] 3
> print(unique_values_summary$BusinessTravel_sample_values)
[1] "Travel_Rarely, Travel_Frequently, Non-Travel, NA, NA"
```

Сада можемо погледати расподелу ове колоне.



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.

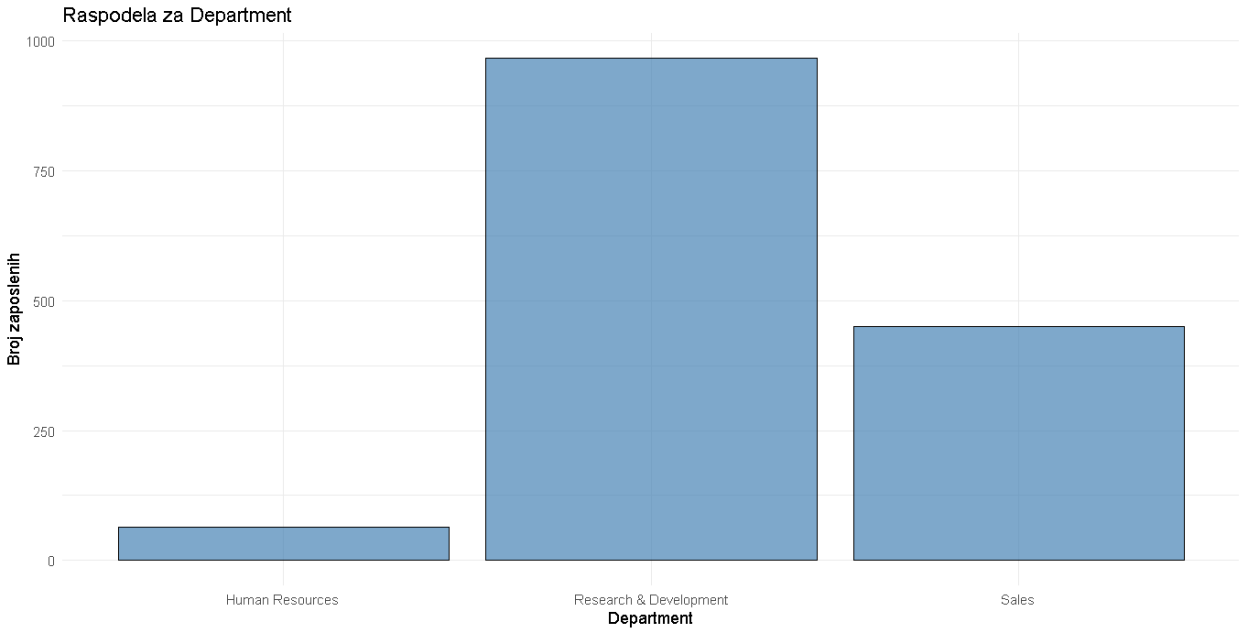
Колона „DailyRate“

Представља уговорени износ дневне накнаде запосленог. Можемо проверити да ли ова колона садржи екстремне вредности који би негативно утицали на наш модел.

Пошто колона не садржи аутлејере можемо погледати и њену расподелу.

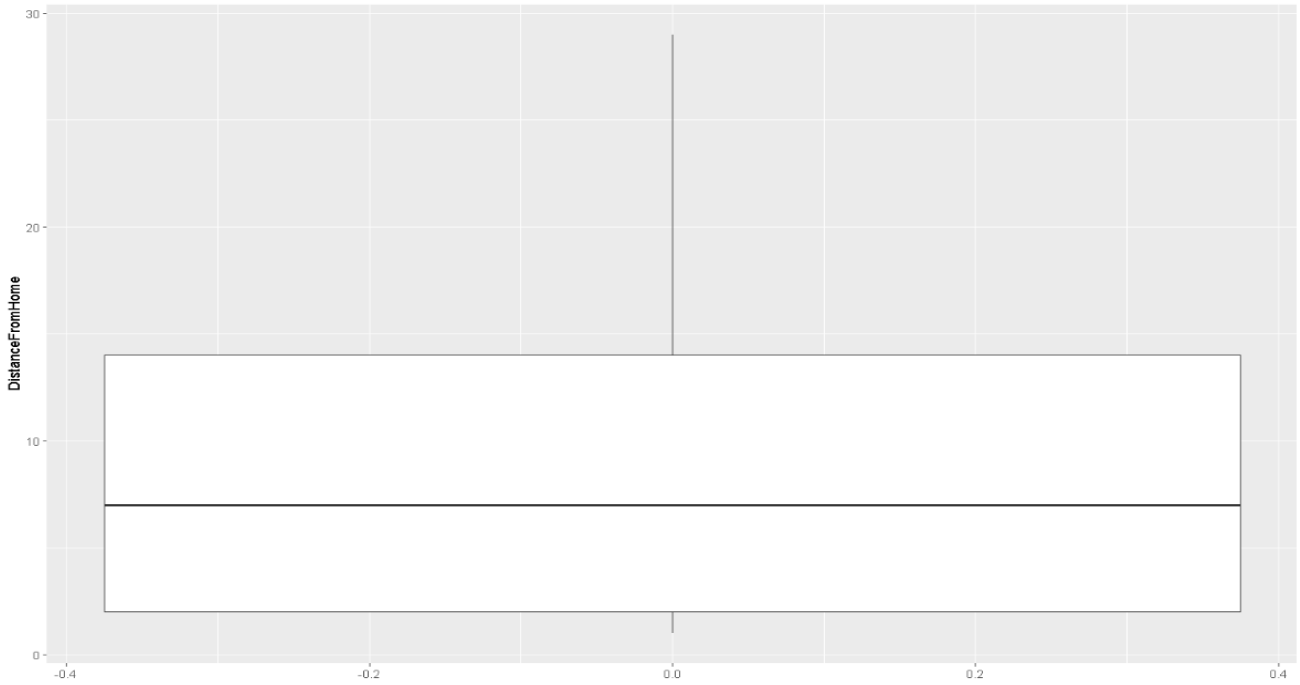
Колона „Department“

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.



С обзиром да ни ова колона не садржи можемо погледати њену расподелу.

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 података удаљености запосленог од куће.

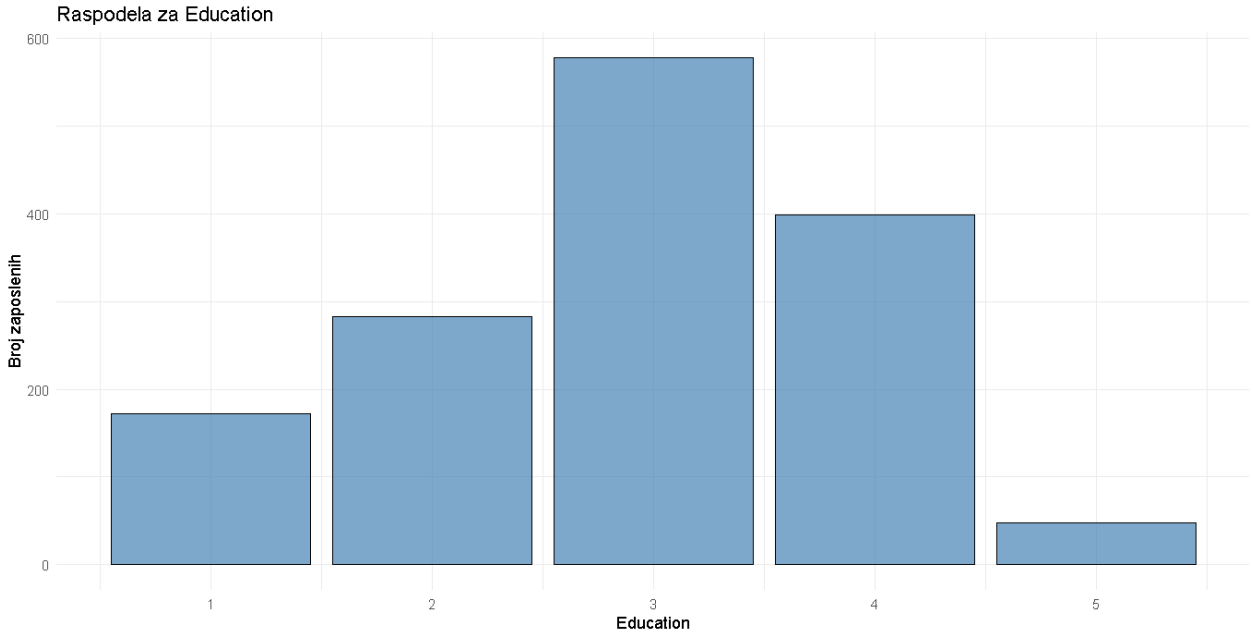
Подела на групе:

- 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
> 
```

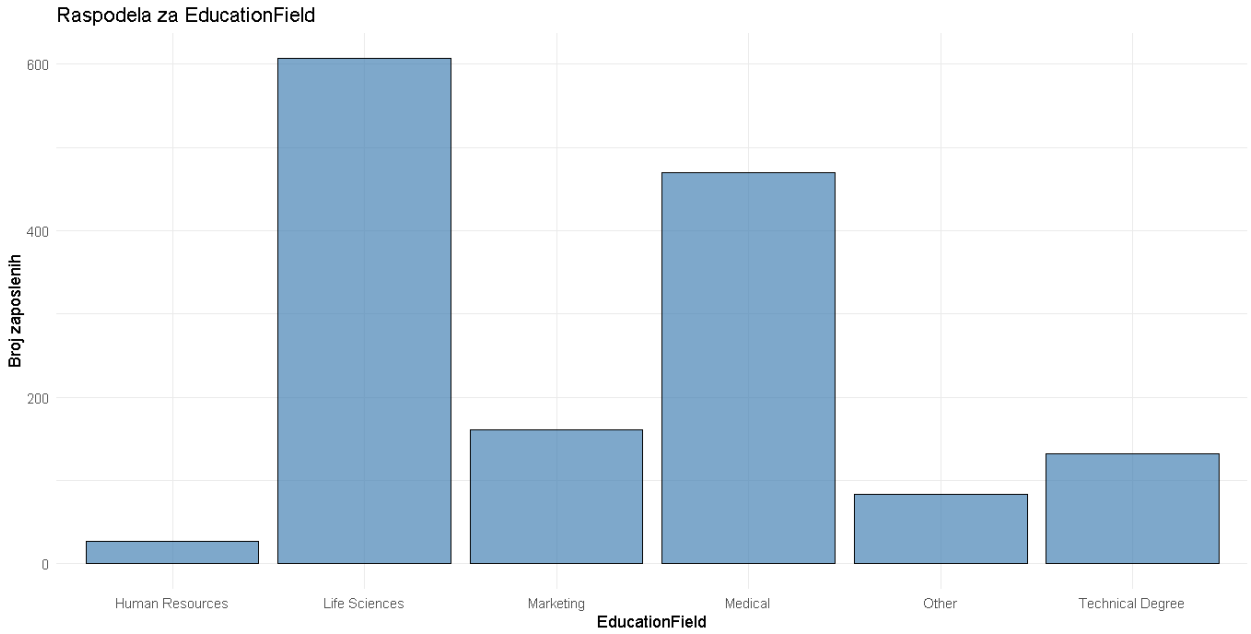
Колона „Education“

Представља ниво образовања запосленог.



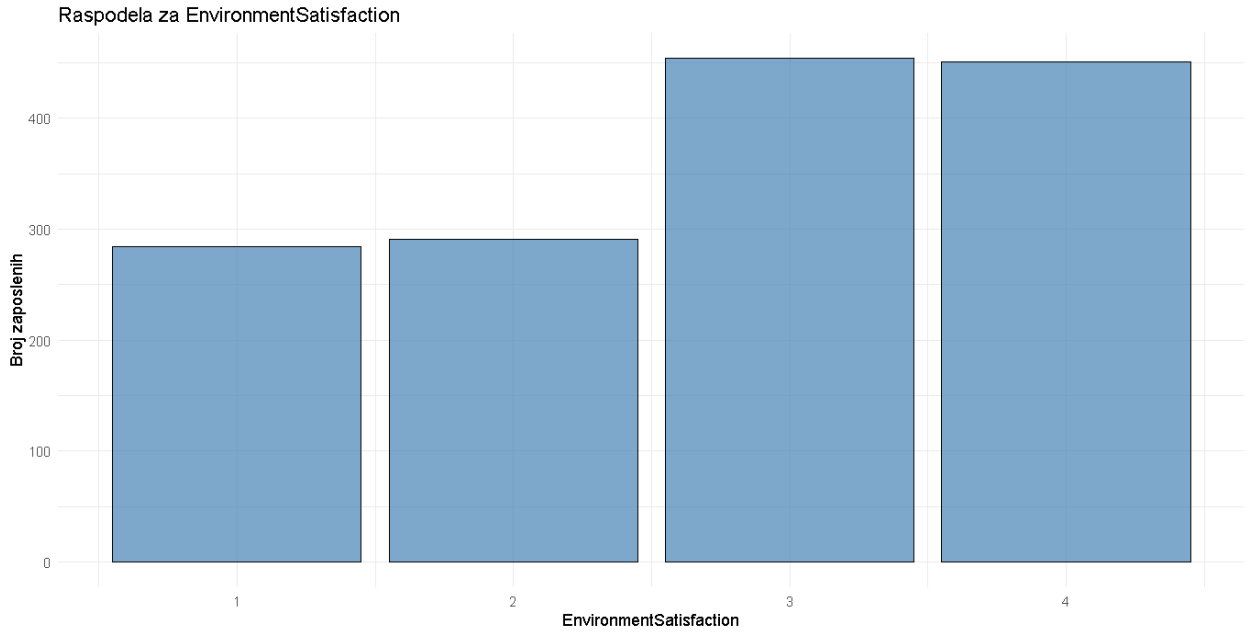
Колона „EducationField“

Представља поље образовања запосленог.



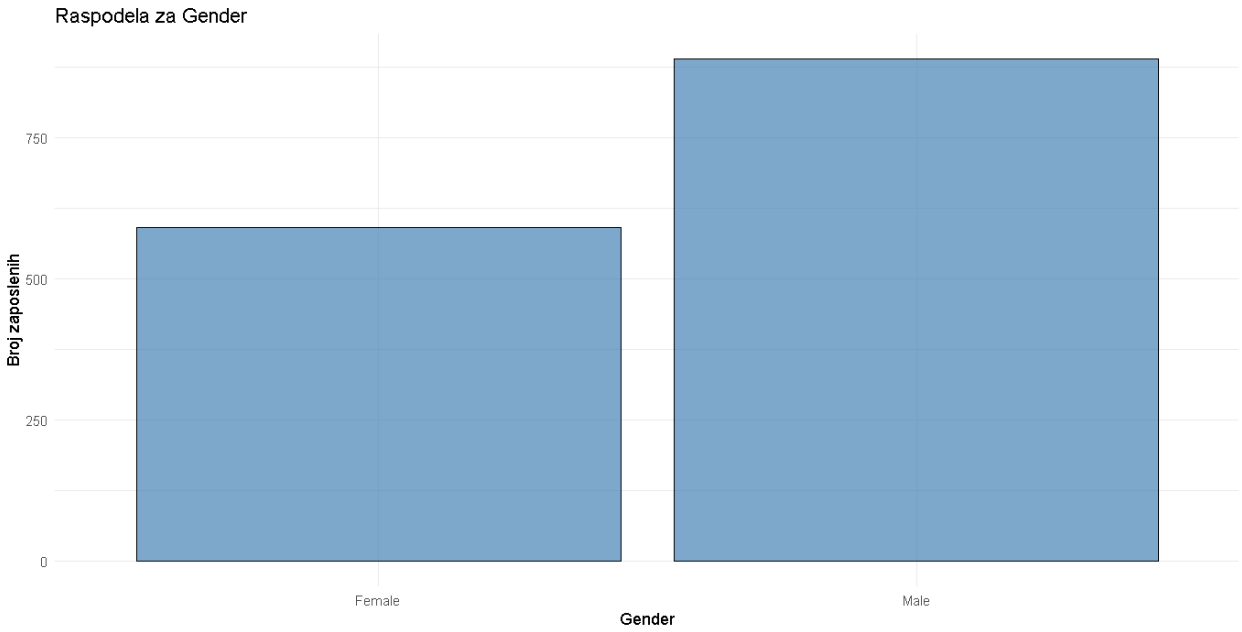
Колона „EnvironmentSatisfaction“

Представља задовољство пословним окружењем.



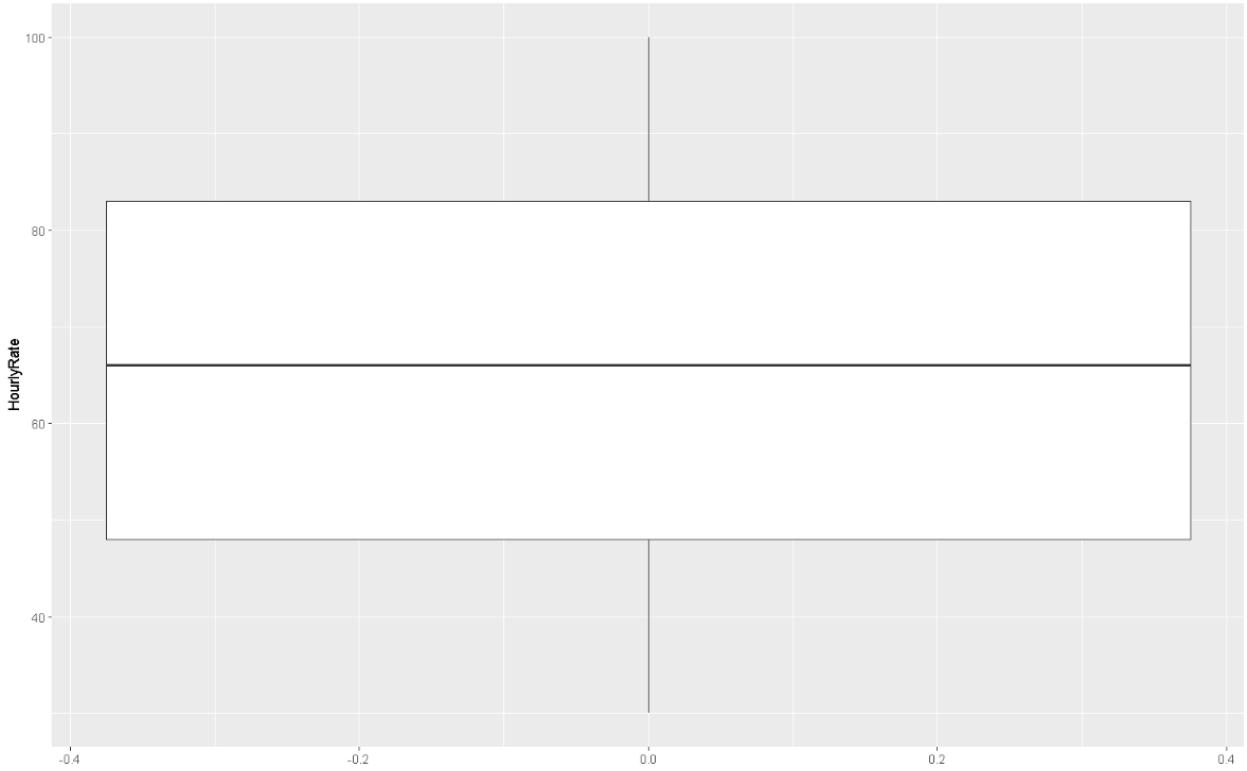
Колона „Gender“

Представља пол запосленог.



Колона „HourlyRate“

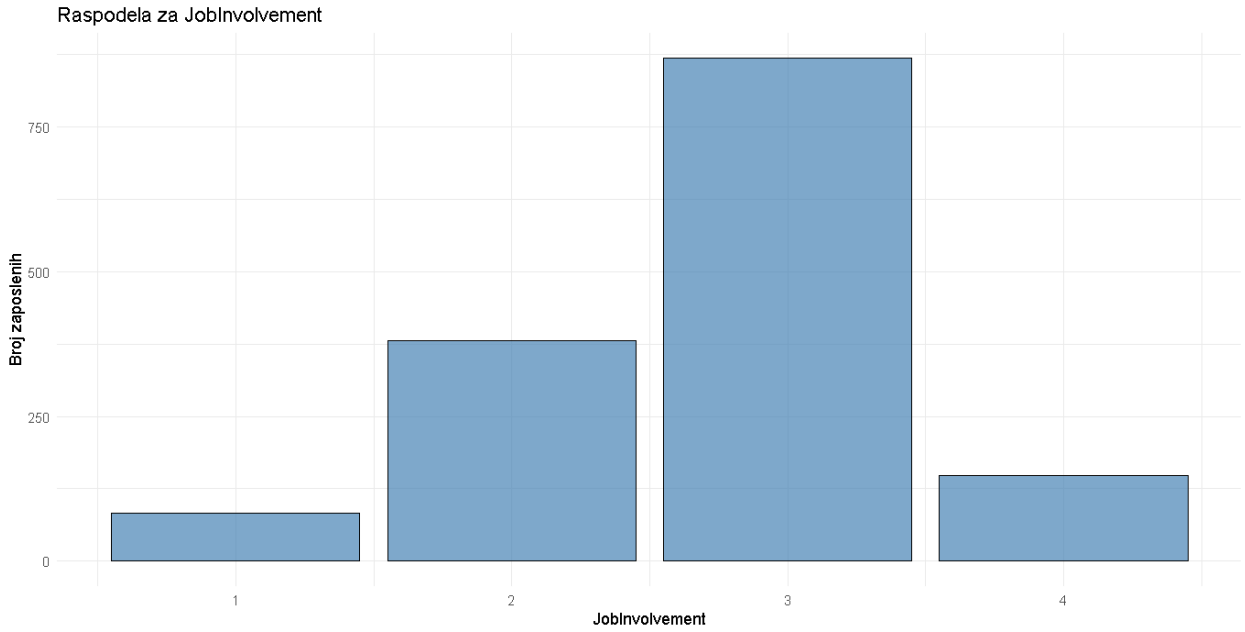
Представља уговорени износ запосленог по радном сату. Као и за DailyRate колону, можемо проверити екстремне вредности.



Добијамо сличан боксплот као код DailyRate, што је и логично, с обзиром на то да сви запослени имају исти уговорени број сати.

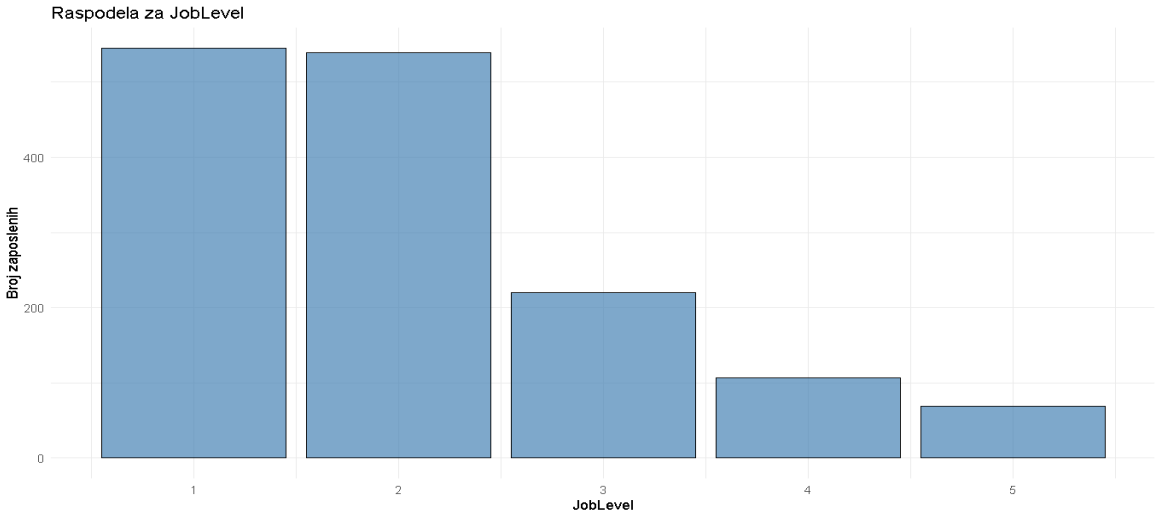
Колона „JobInvolment“

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



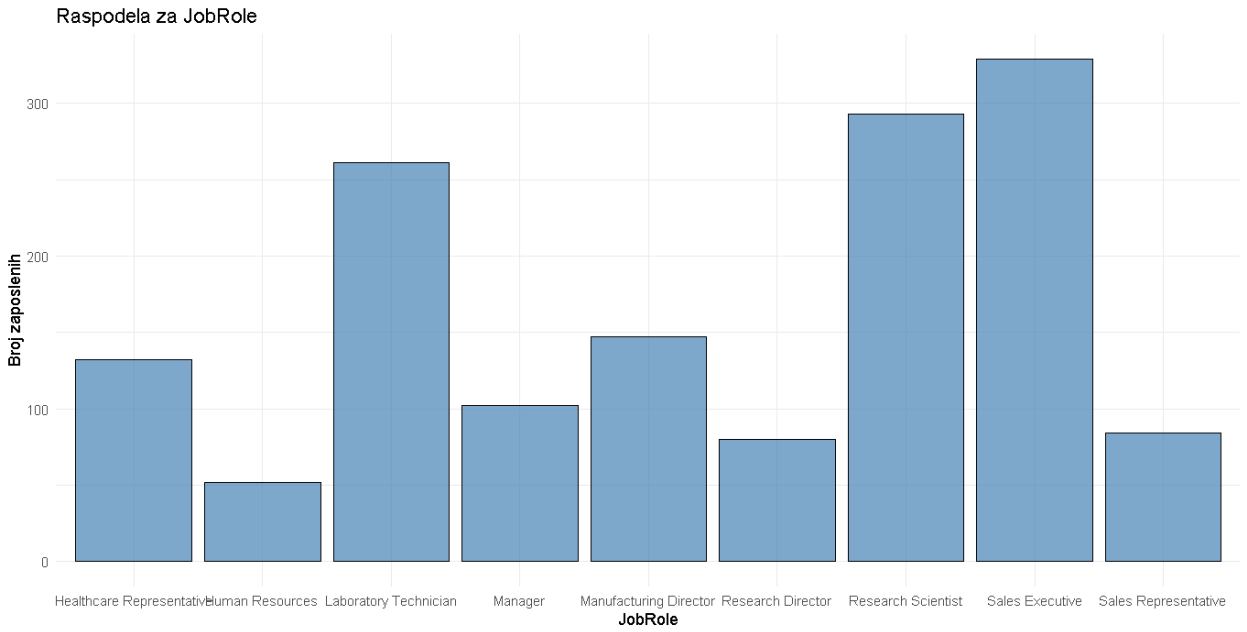
Колона „JobLevel“

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. Приказано на скали од 1 до 5, подела скупа података на основу ове колоне изгледа:



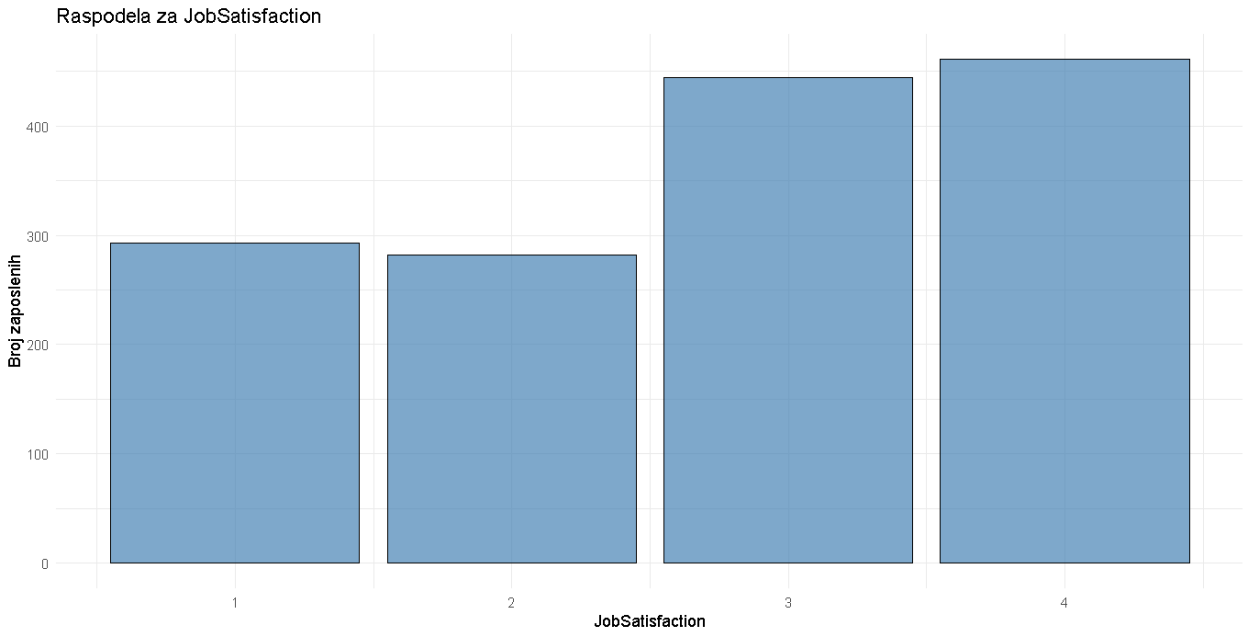
Колона „JobRole“

Представља звање запосленог унутар компаније.



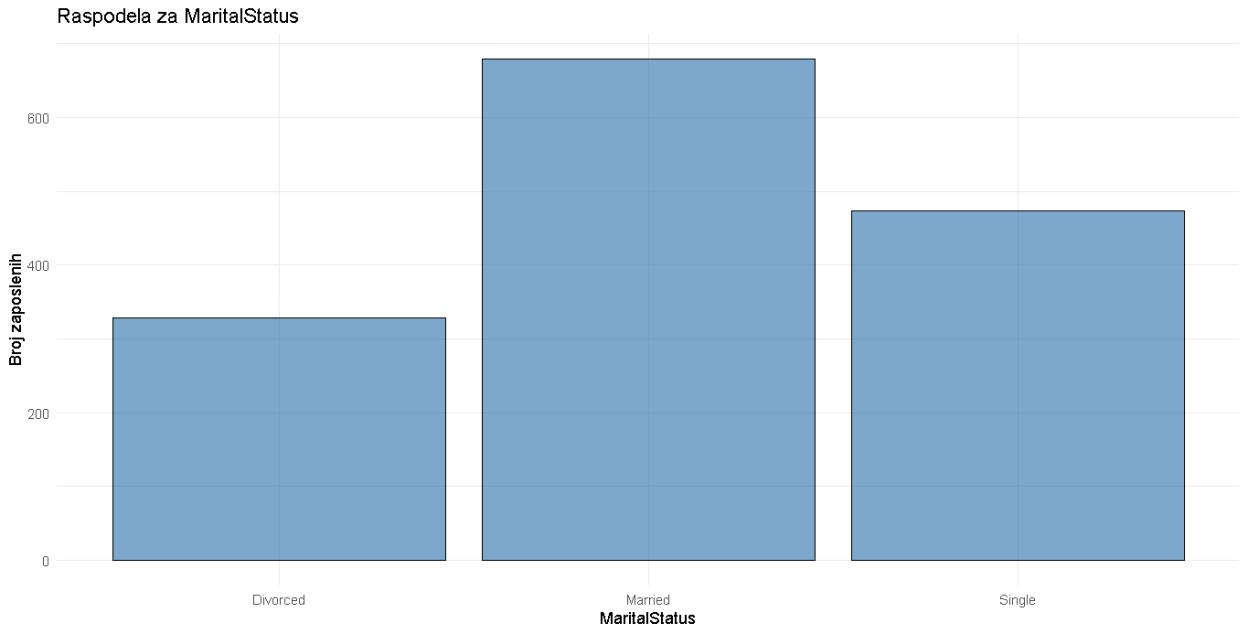
Колона „JobSatisfaction“

Представља задовољство запосленог на пословном плану.



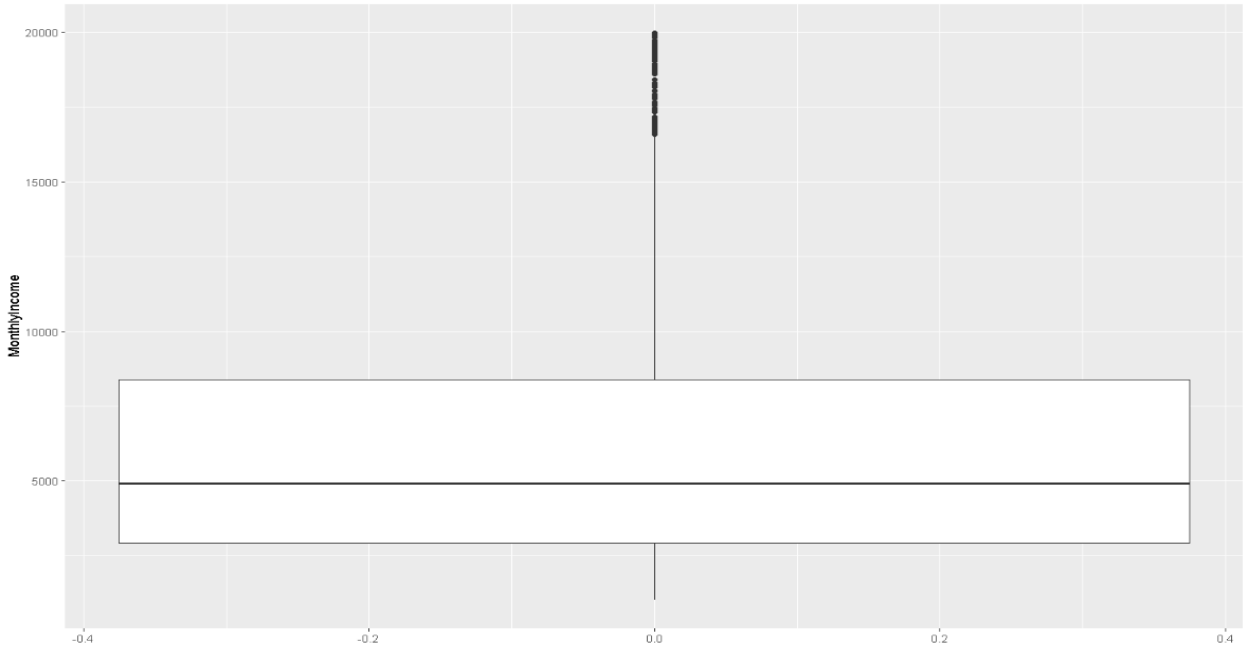
Колона „MartialStatus“

Представља брачно стање запосленог.



Колона „MonthlyIncome“

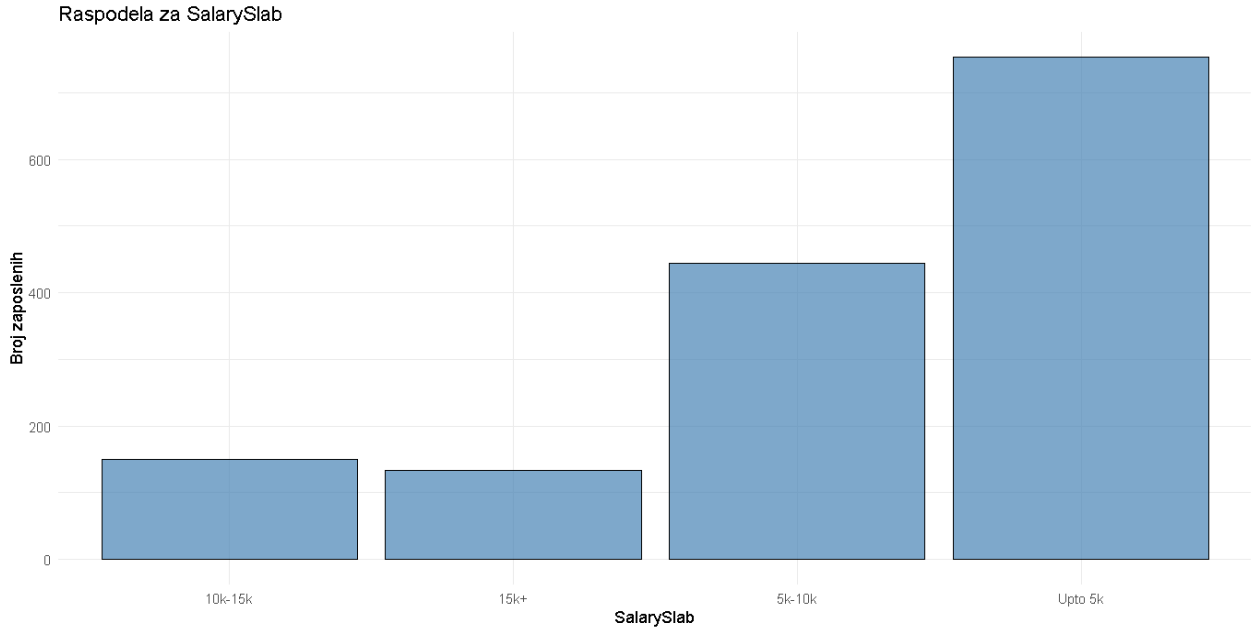
Представља месечне приходе запосленог. Можемо проверити расподелу ове колоне на боксплоту и проверити екстремне вредности



.

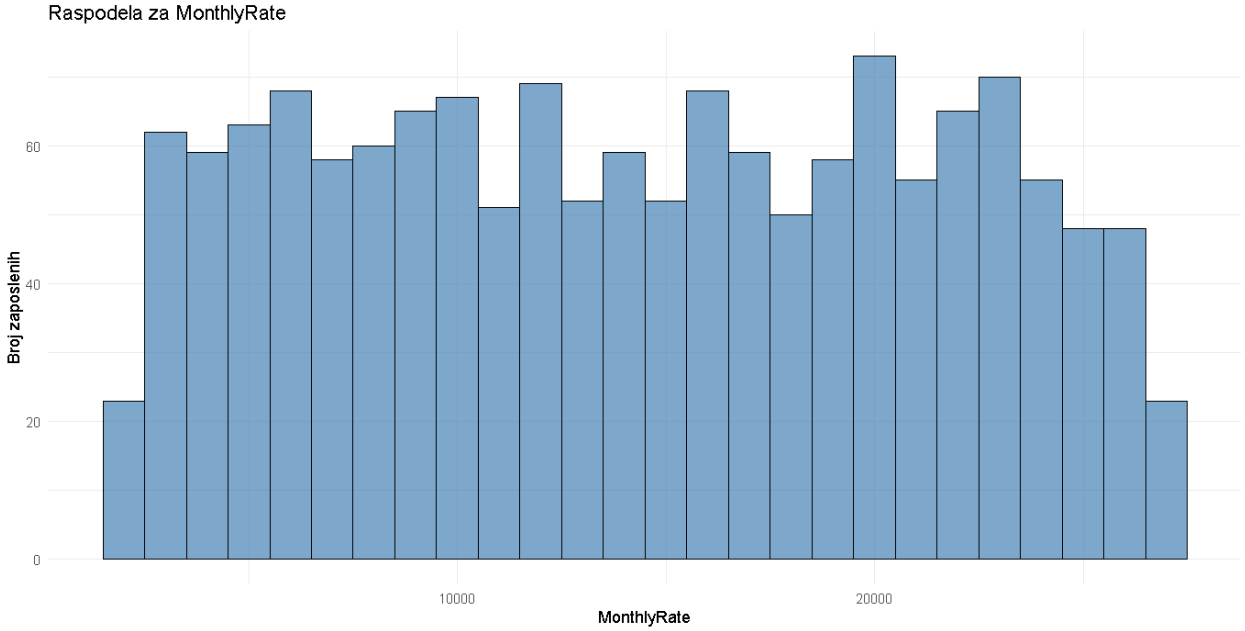
Колона „SalarySlab“

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



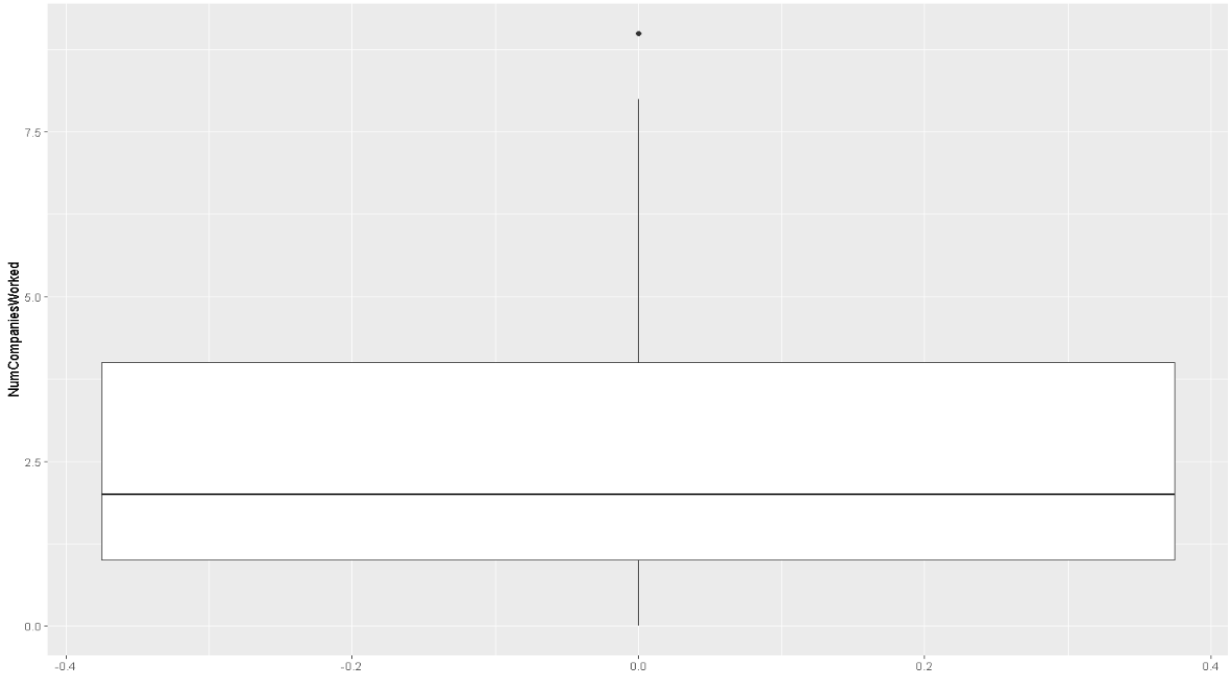
Колона „MonthlyRate“

Представља уговорену месечну зараду запосленог. Као и код осталих типова ове колоне тако ни ова колона не садржи екстремне вредности.



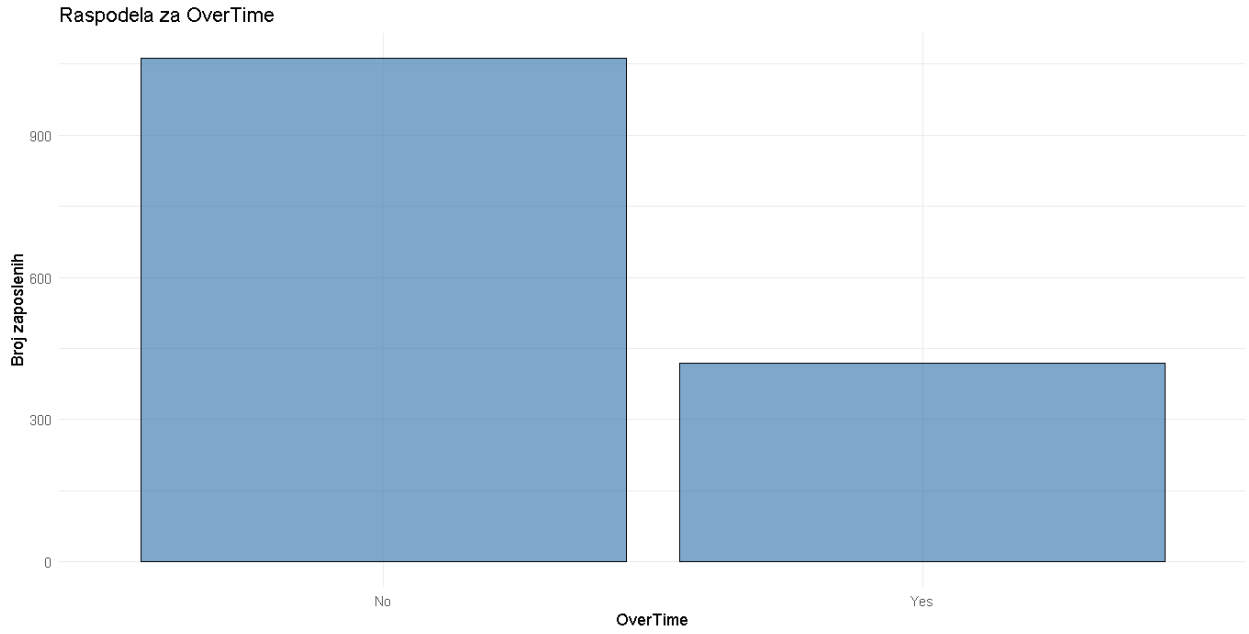
Колона „NumCompaniesWorked“

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



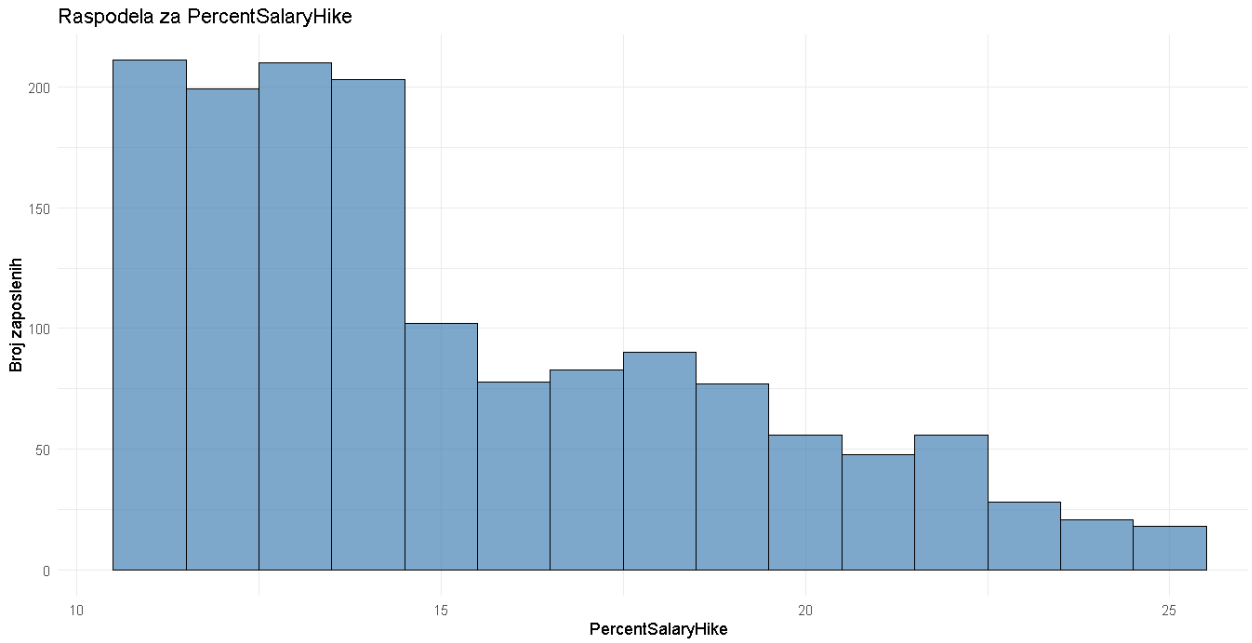
Колона „OverTime“

Представља индикатор за прековремени рад.



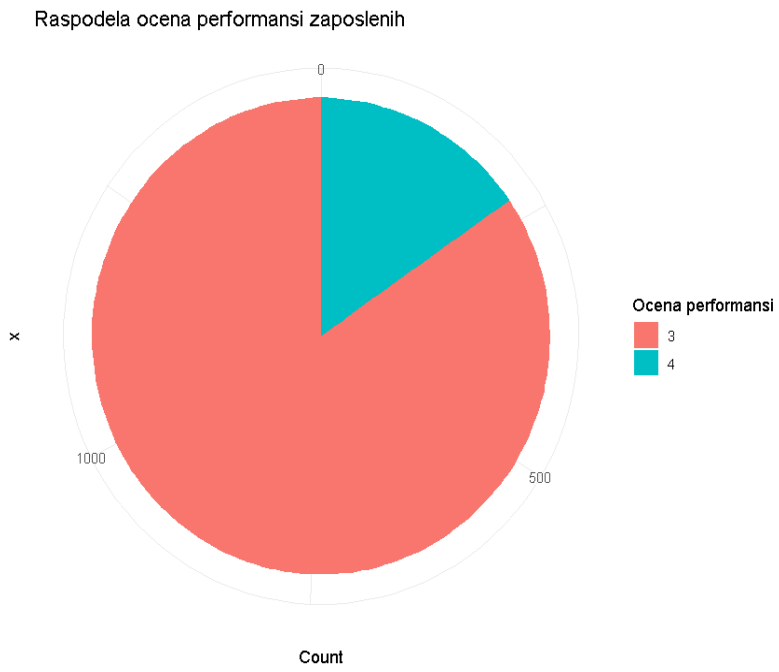
Колона „PercentSalaryHike“

Представља проценат повишице.



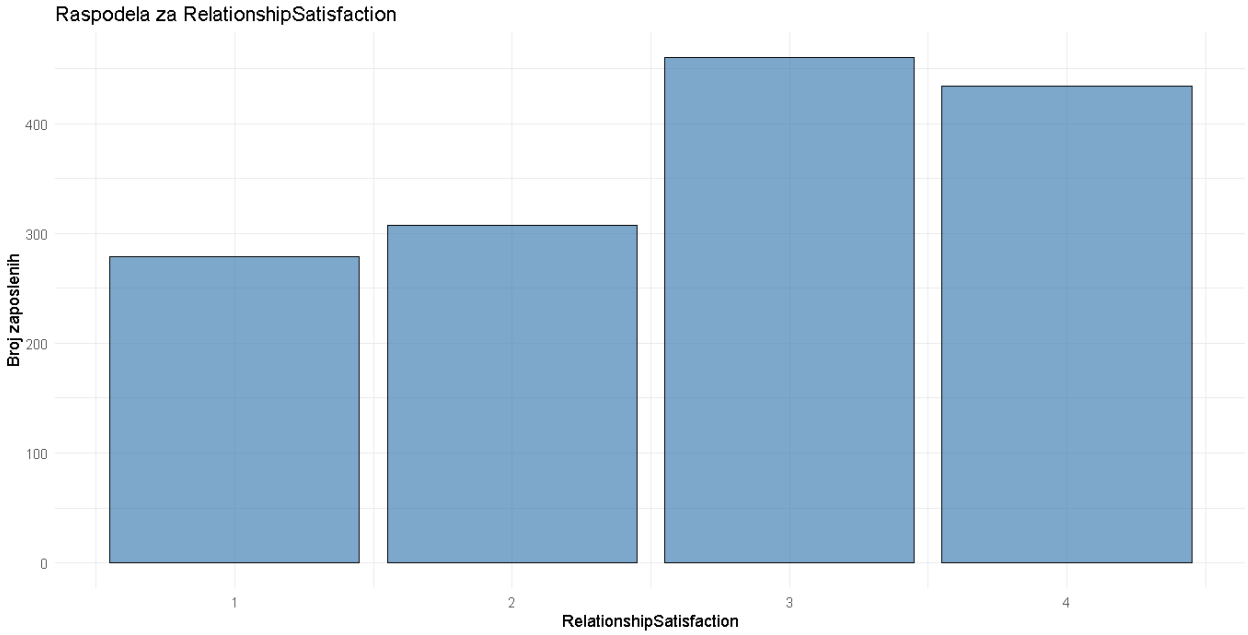
Колона „PerformanceRating“

Представља оцену перформансе запосленог.



Колона „RelationshipStatus“

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

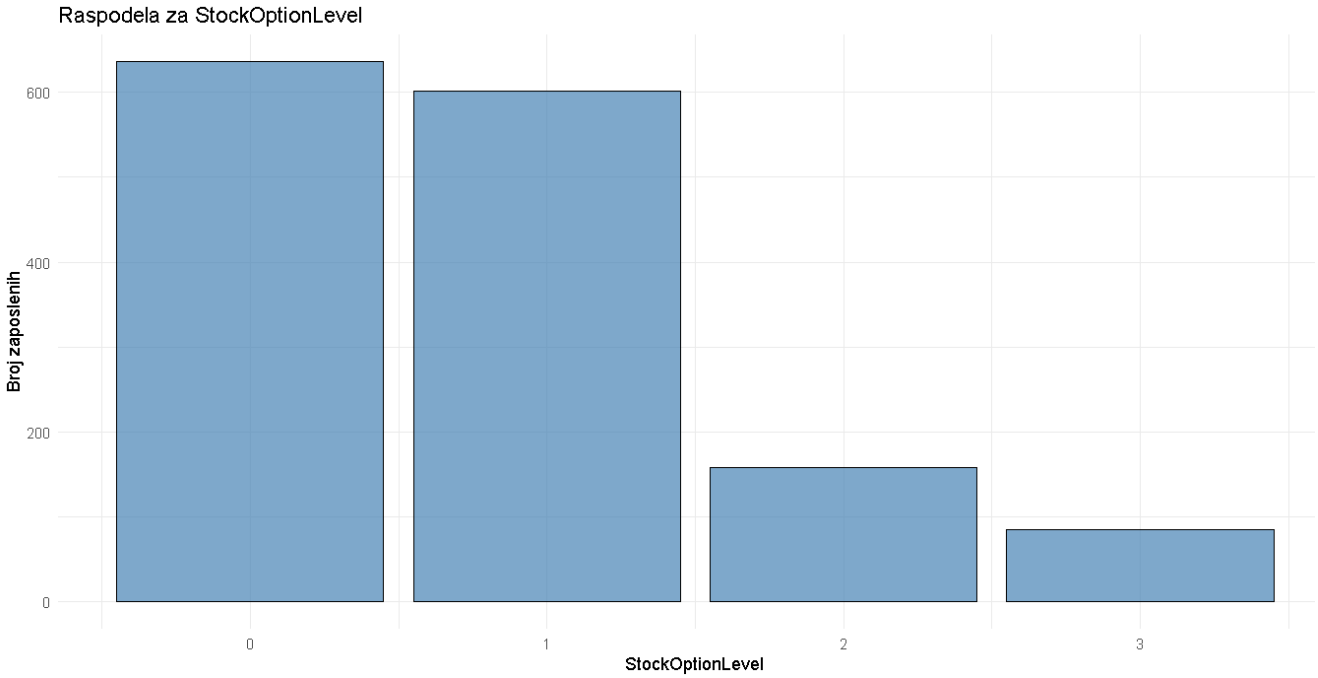


Колона „StandardHours“

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.

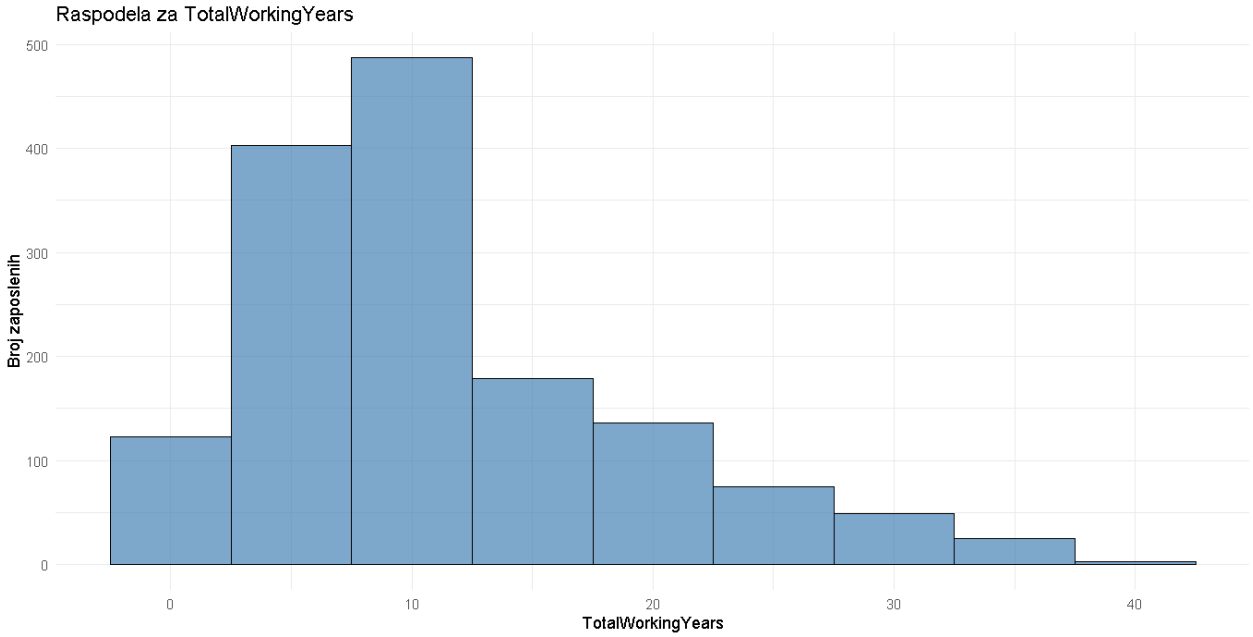
Колона „StockOptionLevel“

Показује да ли запослени има могућност да купује акције компаније.



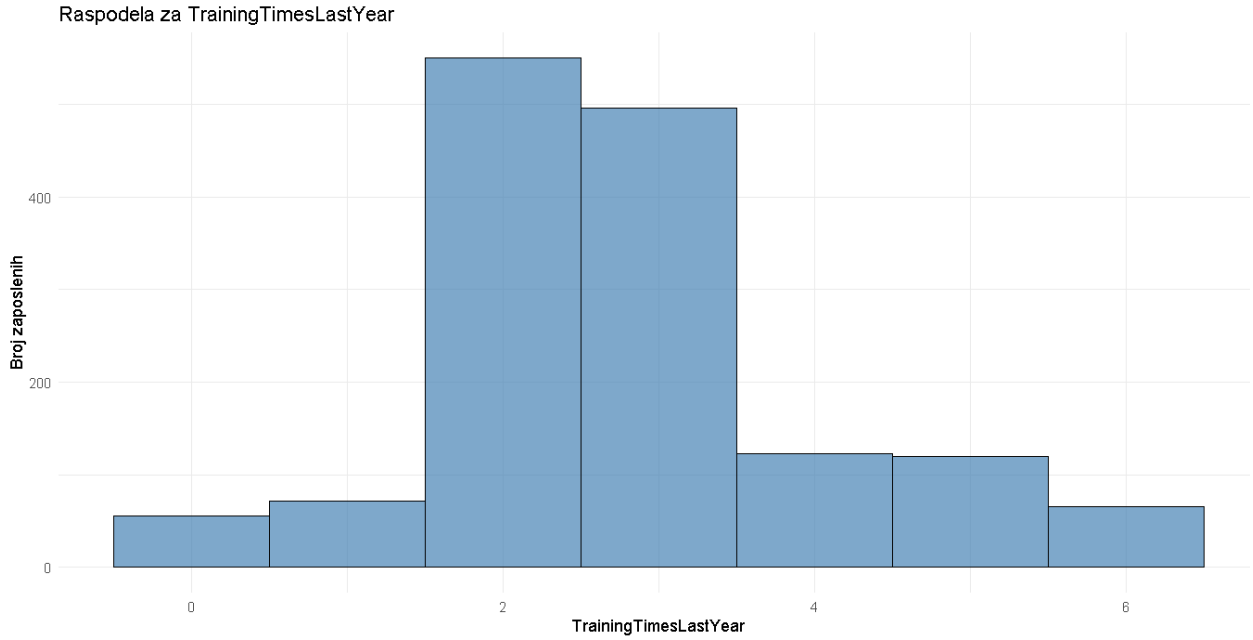
Колона „TotalWorkingYears“

Представља укупан радни стаж запосленог.



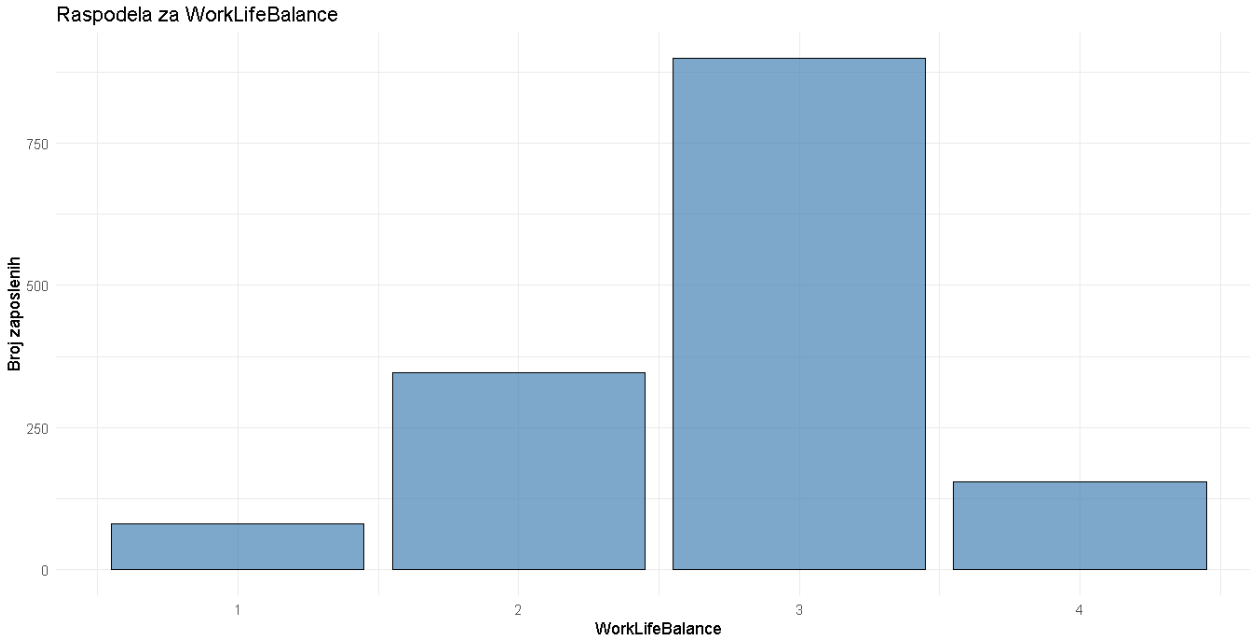
Колона „TrainingTimesLastYear“

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



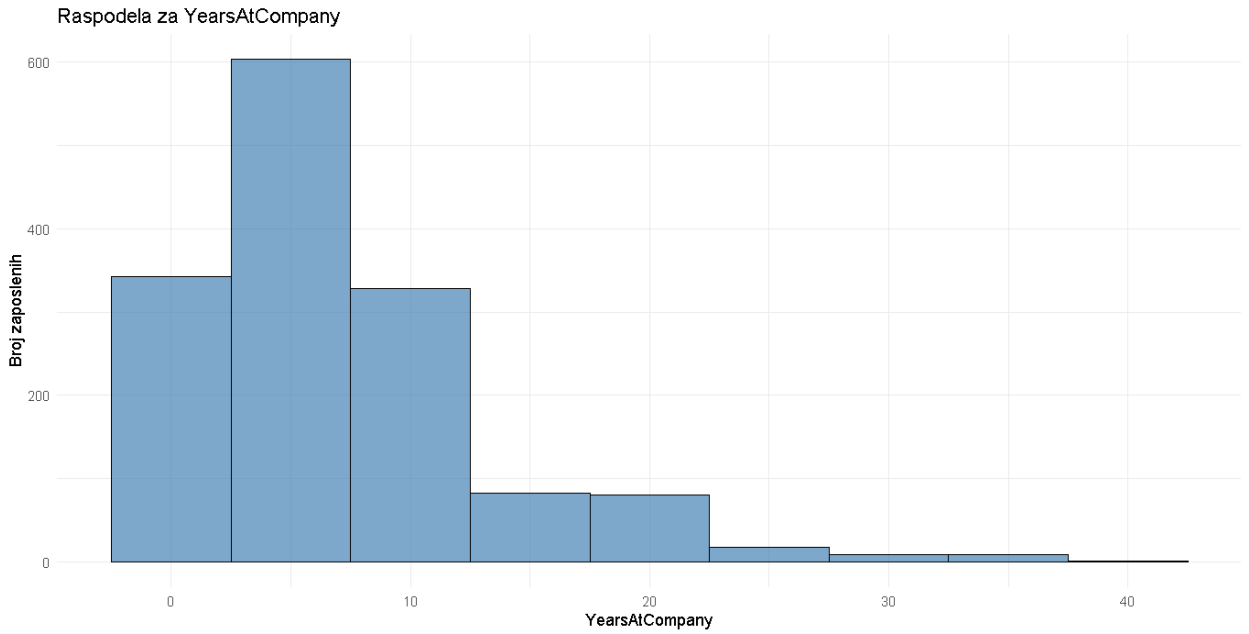
Колона „WorkLifeBalance“

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



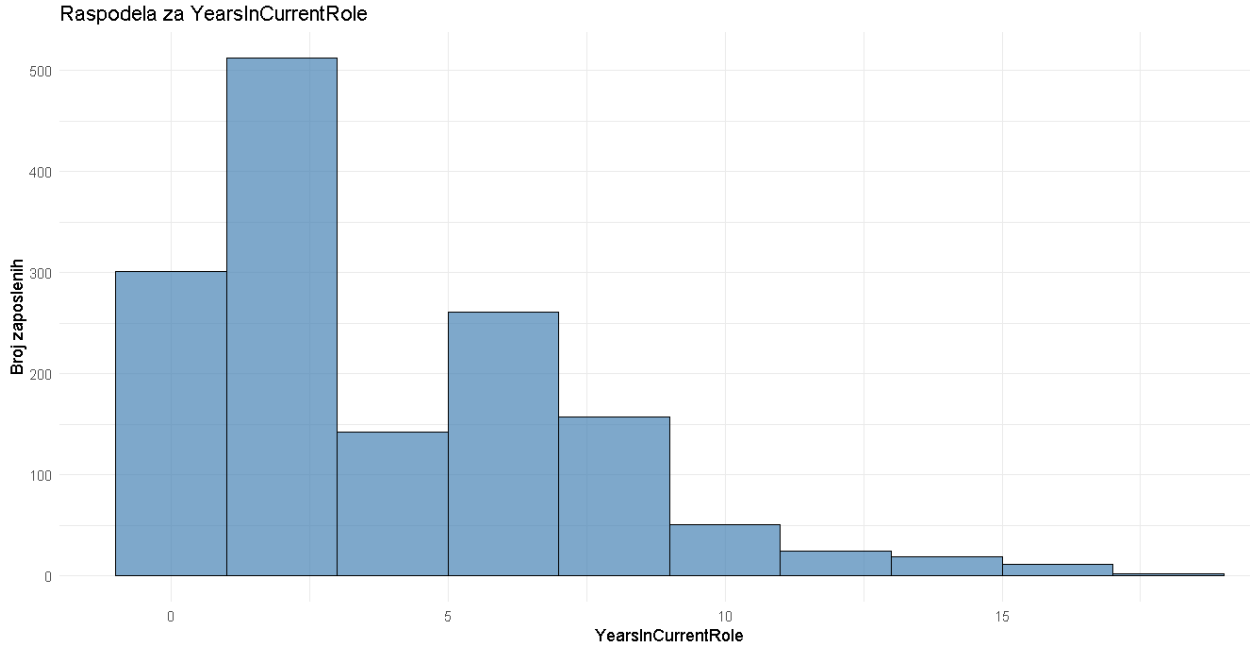
Колона „YearsAtCompany“

Представља колико година је запослени провео у фирми.



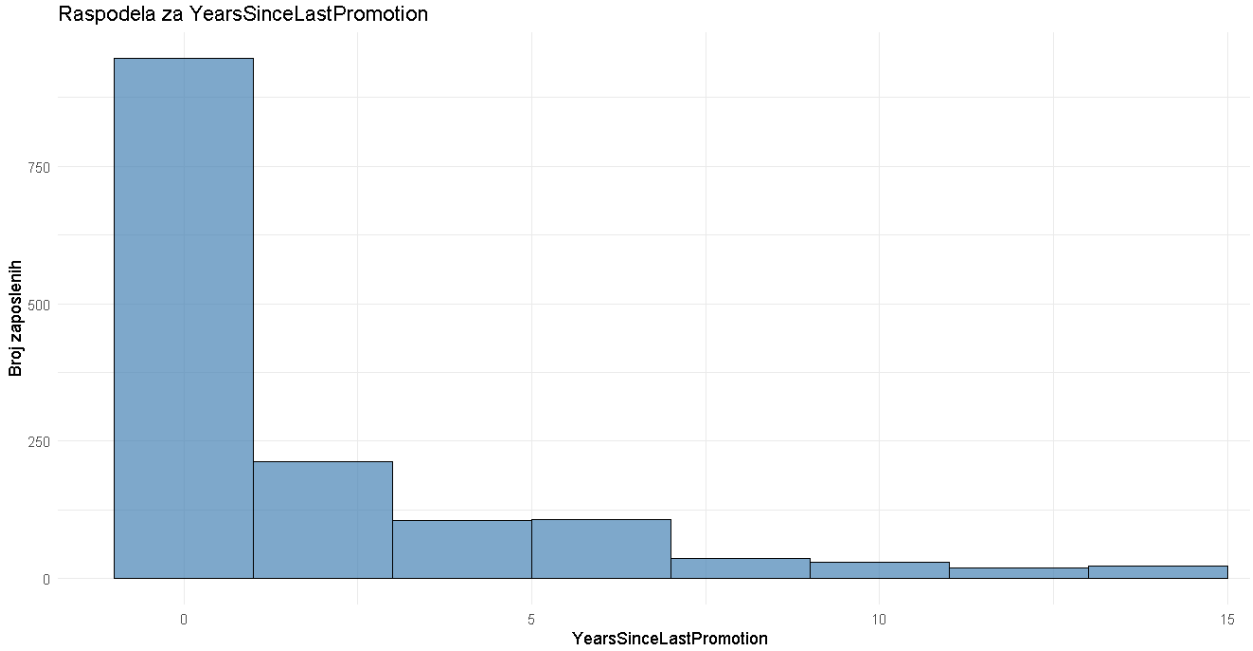
Колона „YearsInCurrentRole“

Представља колико година је запослени провео на истој позицији.



Колона „YearsSinceLastPromotion“

Представља колико година је прошло од последње промоције.



Колона „YearsWithCurrentManager“

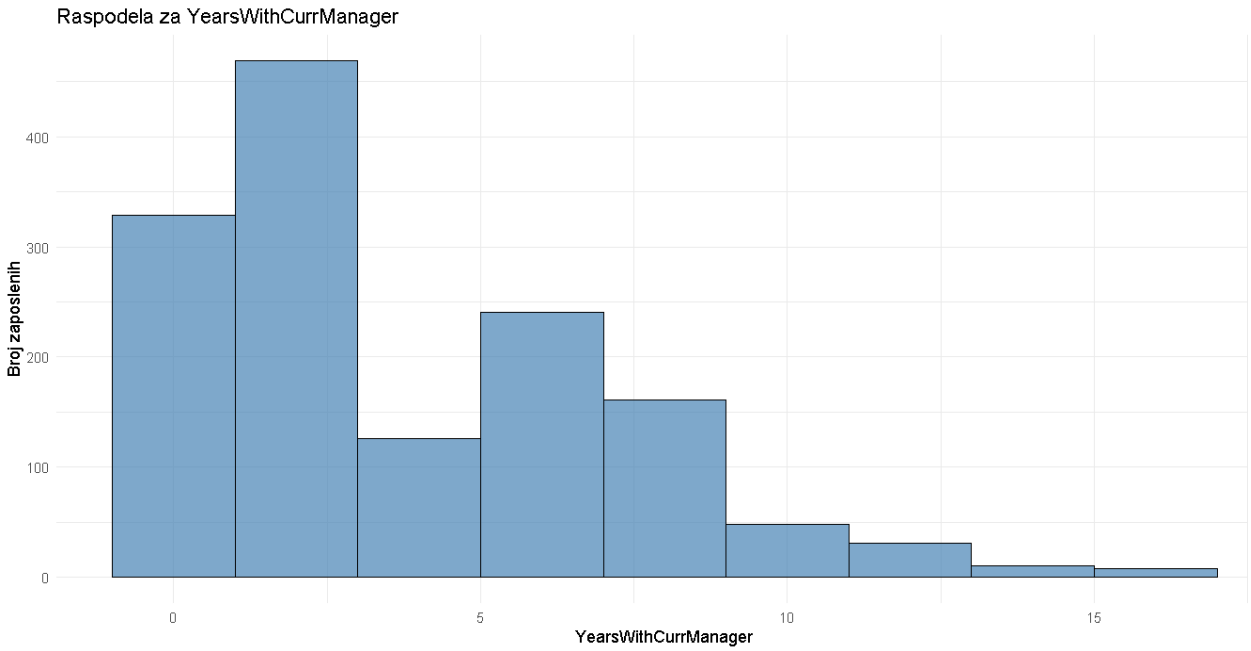
Ова колона показује број година које запослени провео под тренутним менаџером.

```
> 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
```

Видимо да ова колона садржи 57 NA 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 NA 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
```

Сада можемо погледати расподелу за ову колону.


Raspodela za YearsWithCurrManager
Broj zaposlenih
400
300
200
100
0
0
5
10
15
YearsWithCurrManager

2.1 Уклањање неважних колона

- За почетак уклонићемо 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 колона **Over18** и **StandardHours** исте за све запослене, па ћемо их такође избацити:

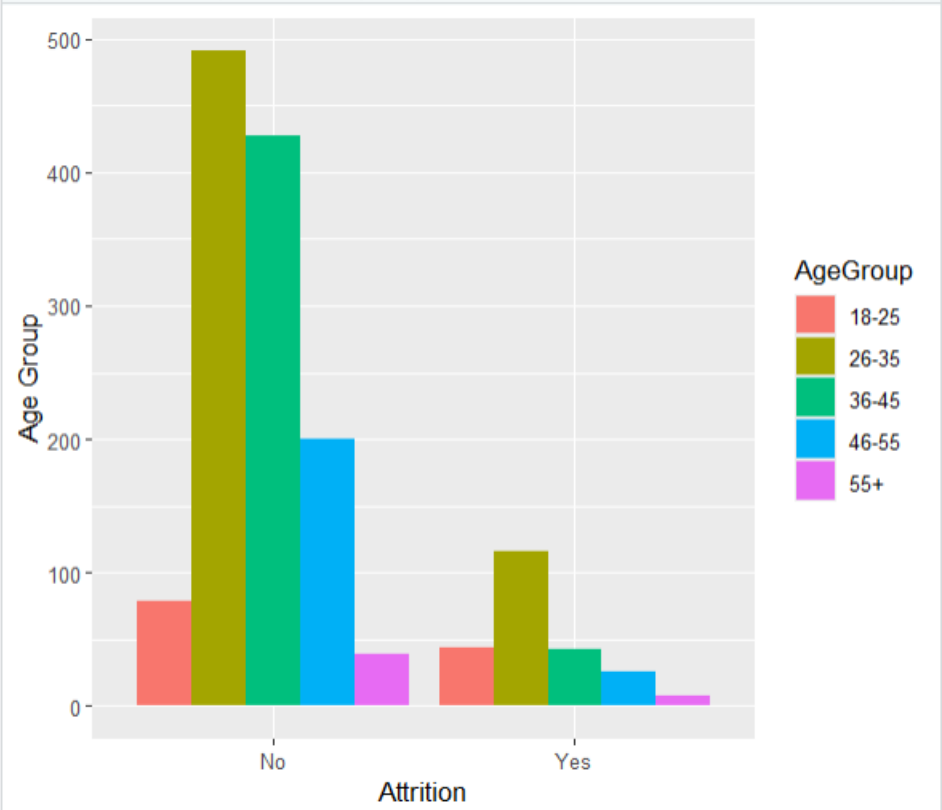
```
hr_data$StandardHours <- NULL
hr_data$Over18 <- NULL
```

3. Анализа података

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

- **Age group и Attrition**

```
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"))
```



На основу графика можемо да закључимо да запослени који процентуално више напуштају компанију припадају млађим старосним групама. Тај тренд је најизраженији у размаку од 18 до 35 година, након тога је процентуално мања одлазност из компаније.

- **Age и Monthly Income**

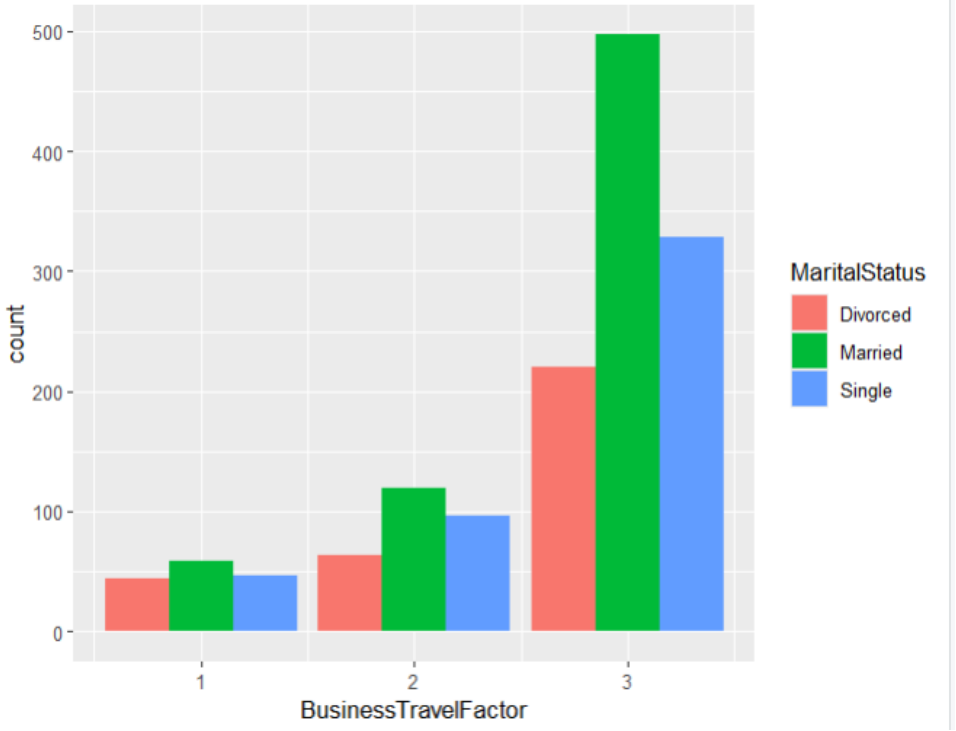
```
ggplot(data = hr_data) +
  geom_point(aes(x = MonthlyIncome, y = Age, color = Attrition))
```


```
> 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 примања уз доста изузезака након 30-те године.

- **Martial status и Bussines travel factor**

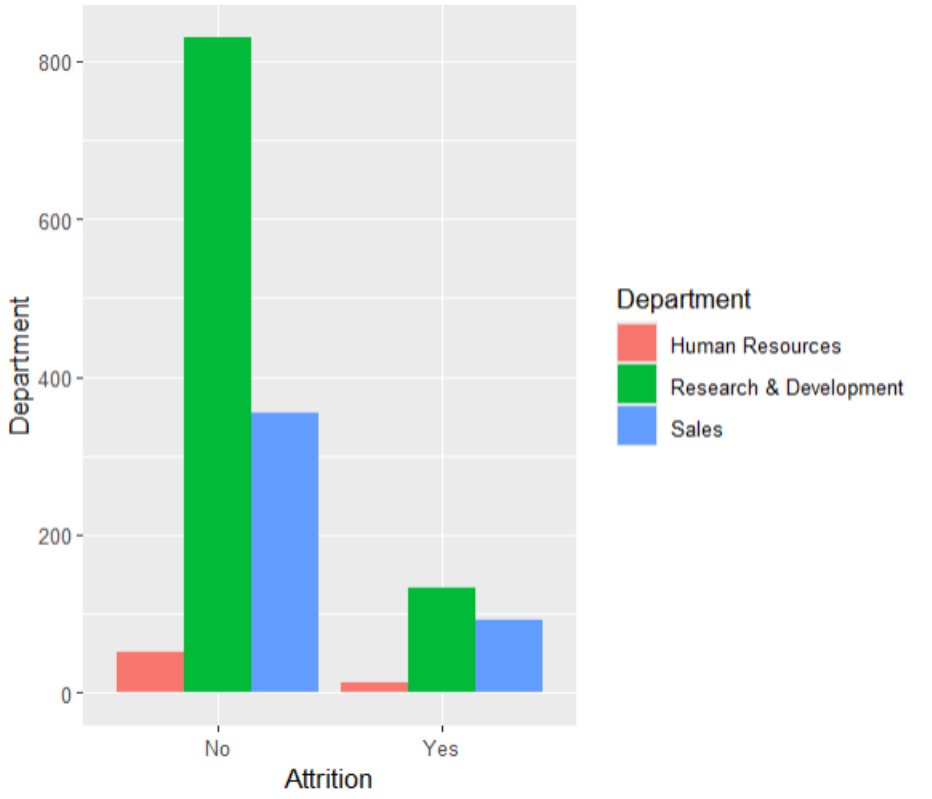
```
ggplot(data = hr_data) +
  geom_bar(aes(x = BusinessTravelFactor, fill = MaritalStatus), position = "dodge")
```



На овом графику можемо видети однос брачног статуса и учесталости пословних путовања запослених у компанији. Генерално примећујемо пре свега да највећи број запослених често путује. Затим можемо приметити да у свим категоријама има највише ожењених, али да су сразмерно броју по катаегорији сви уједначени, па из ове две променљиве не можемо уочити зависност.

- **Attrition и Department**

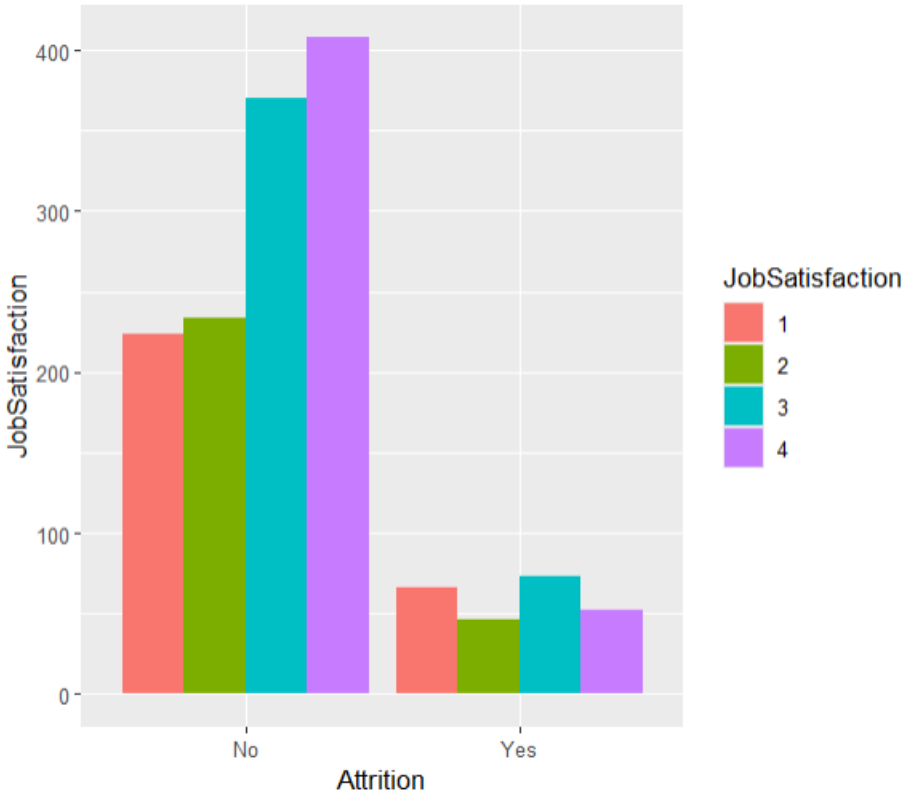
```
ggplot(hr_data, aes(x = factor(Attrition), fill = Department)) +
  geom_bar(position = "dodge") +
  ylab("Department") +
  xlab("Attrition") +
  scale_x_discrete(labels = c("No", "Yes"))
```




```
> 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**

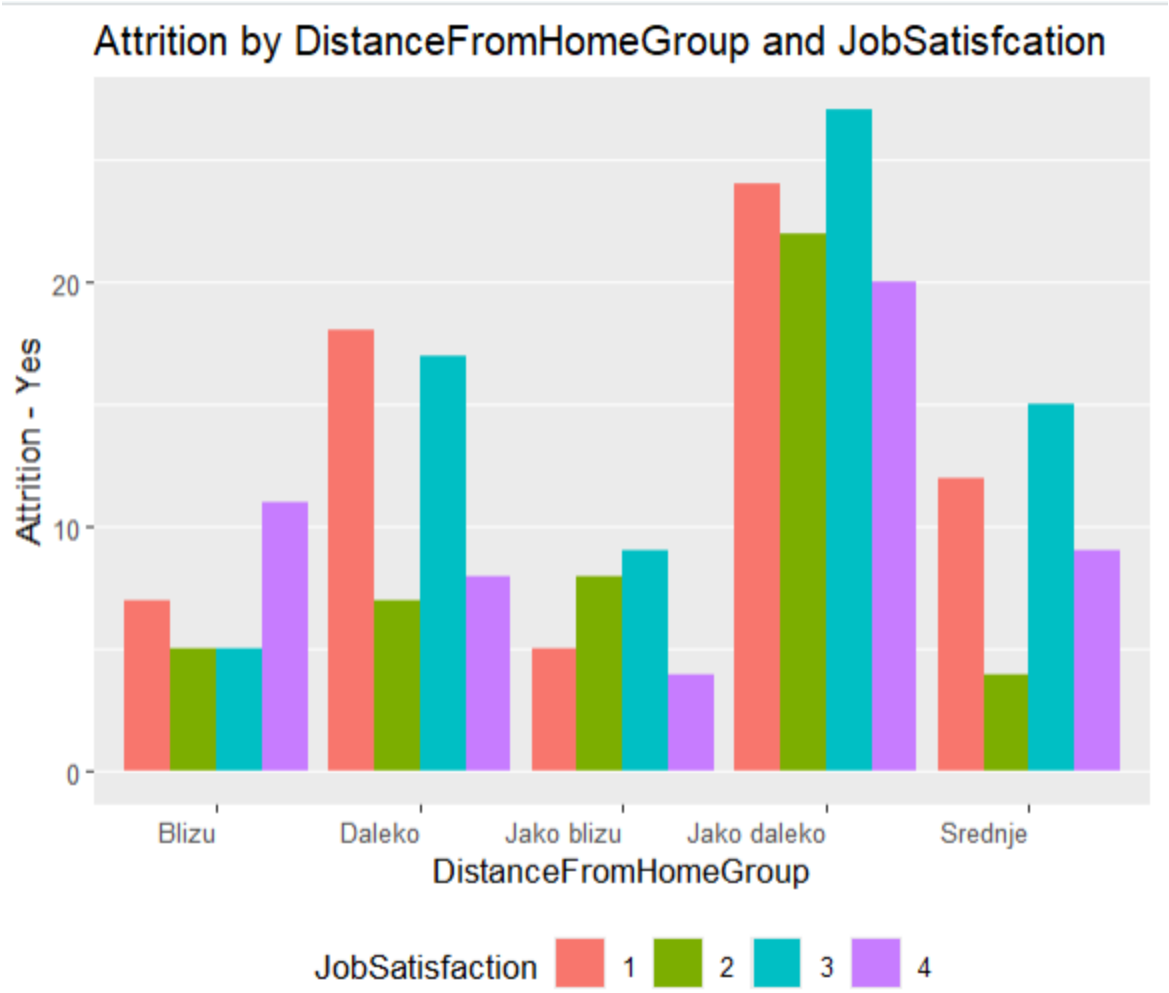



```
> jobSatisfaction.attrition.prop <- prop.table(jobSatisfaction.attrition, margin = 1)
> jobSatisfaction.attrition.prop
               Attrition
JobSatisfaction        No       Yes
              1 0.7724138 0.2275862
              2 0.8357143 0.1642857
              3 0.8352144 0.1647856
              4 0.8869565 0.1130435
```

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

- **Distance from home, Attrition и Job satisfaction**

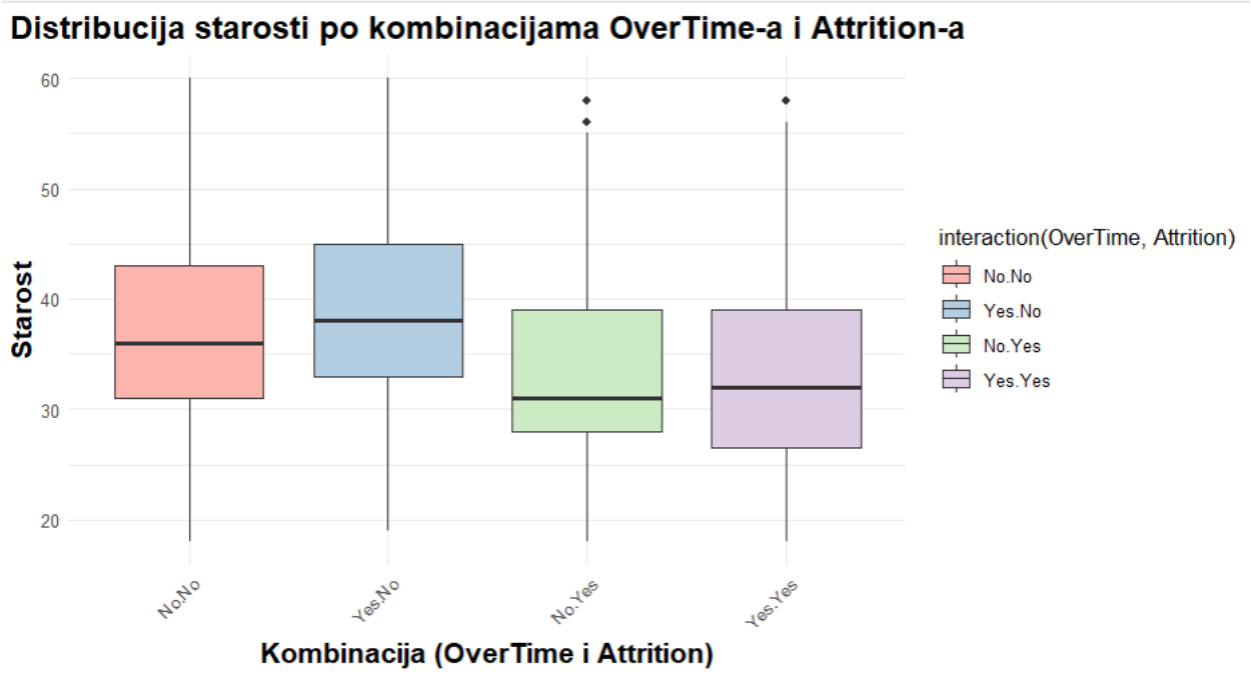
```
data <- hr_data %>%
  group_by(JobSatisfaction, DistanceFromHomeGroup) %>%
  summarise(attrition_count = sum(Attrition_binary))

bar_plot <- ggplot(data, aes(x = DistanceFromHomeGroup, y = attrition_count, fill = JobSatisfaction)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Attrition by DistanceFromHomeGroup and JobSatisfcation",
       x = "DistanceFromHomeGroup",
       y = "Attrition - Yes",
       fill = "JobSatisfaction") +
  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 далеко од фирме чешће напуштају исту. Удаљеност од посла од куће може бити јако добар предиктор.

- **Age, OverTime и Attrition**

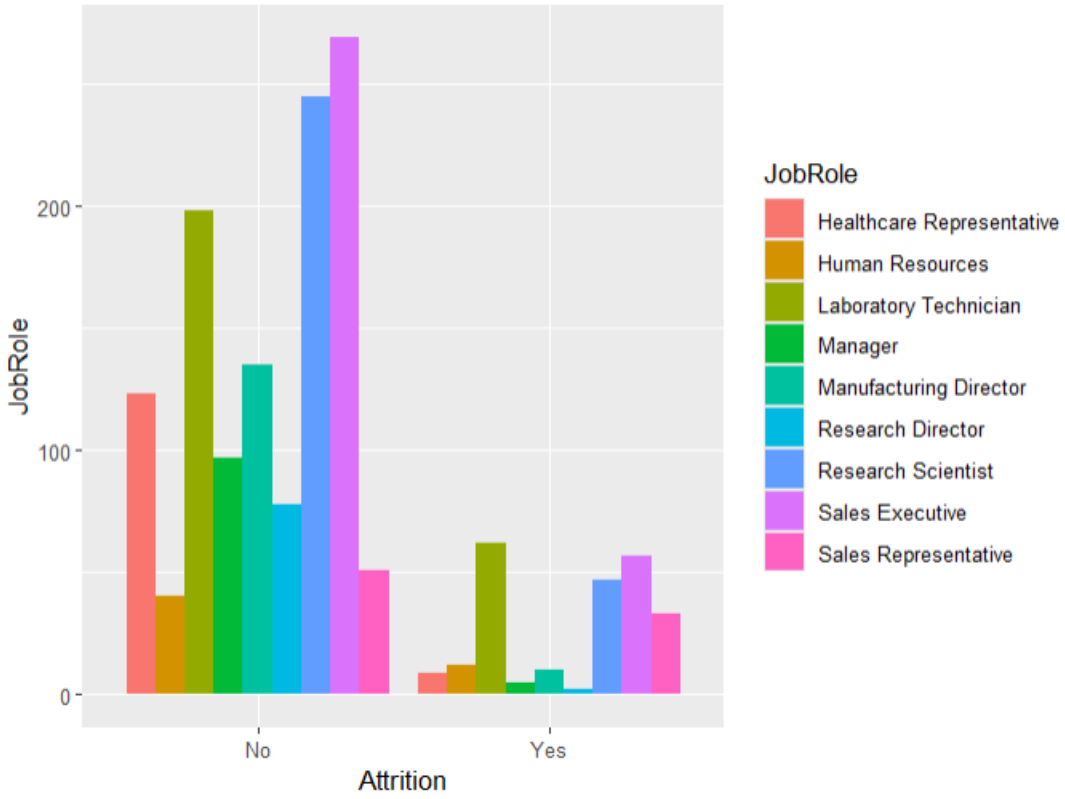
```
ggplot(hr_data, aes(x = interaction(OverTime, Attrition), y = Age, fill = interaction(OverTime, Attrition))) +
  geom_boxplot() +
  labs(title = "Distribucija starosti po kombinacijama OverTime-a i Attrition-a",
       x = "Kombinacija (OverTime i Attrition)",
       y = "Starost") +
  scale_fill_brewer(palette = "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. Млађи запослени који раде прековремено углавном најчешће напуштају фирму, док старије особе које раде прековремено не напуштају фирму.

- **Job role и 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
```

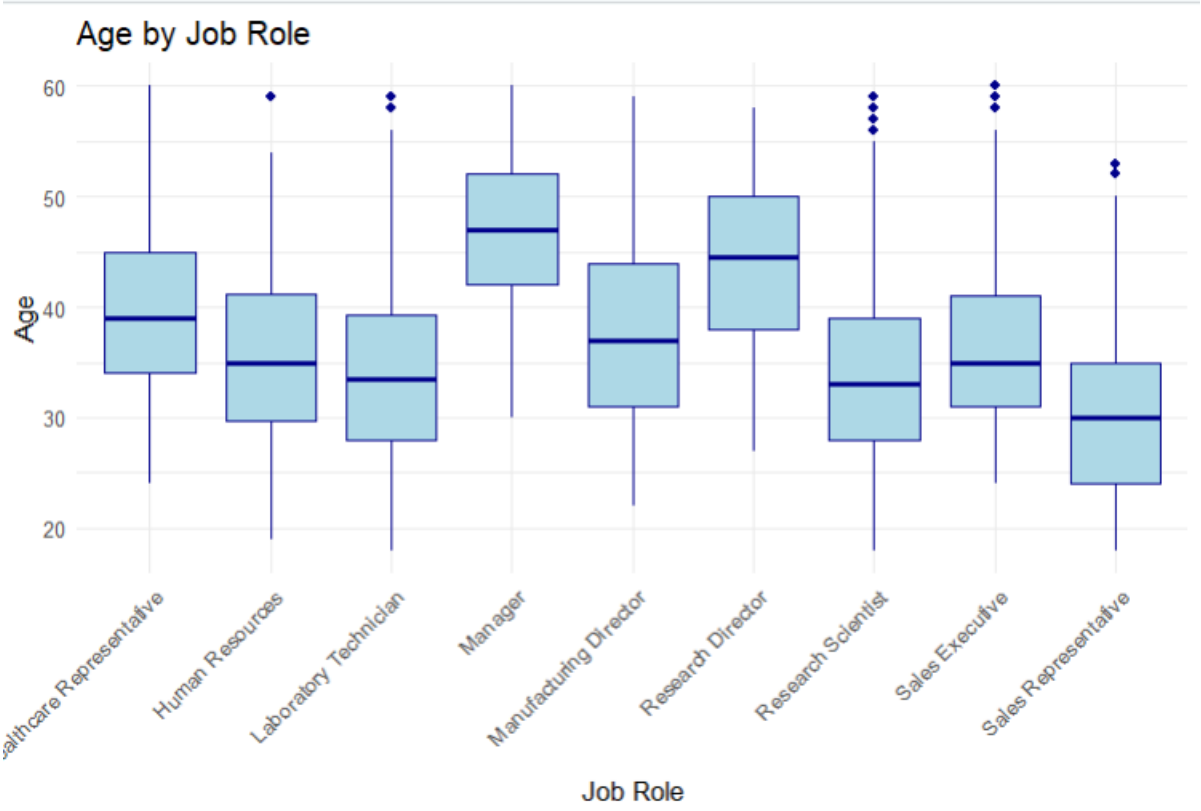


Позиције које најчешће напуштају фирму јесу Sales Representative затим Laboratory Technician и Human Resources, односно најниже позиције у фирми.

Запослени на позицијама Manager, Research Director i Manufacturing Director најмање напуштају фирму, што је и логично.

- **Job role и Age**

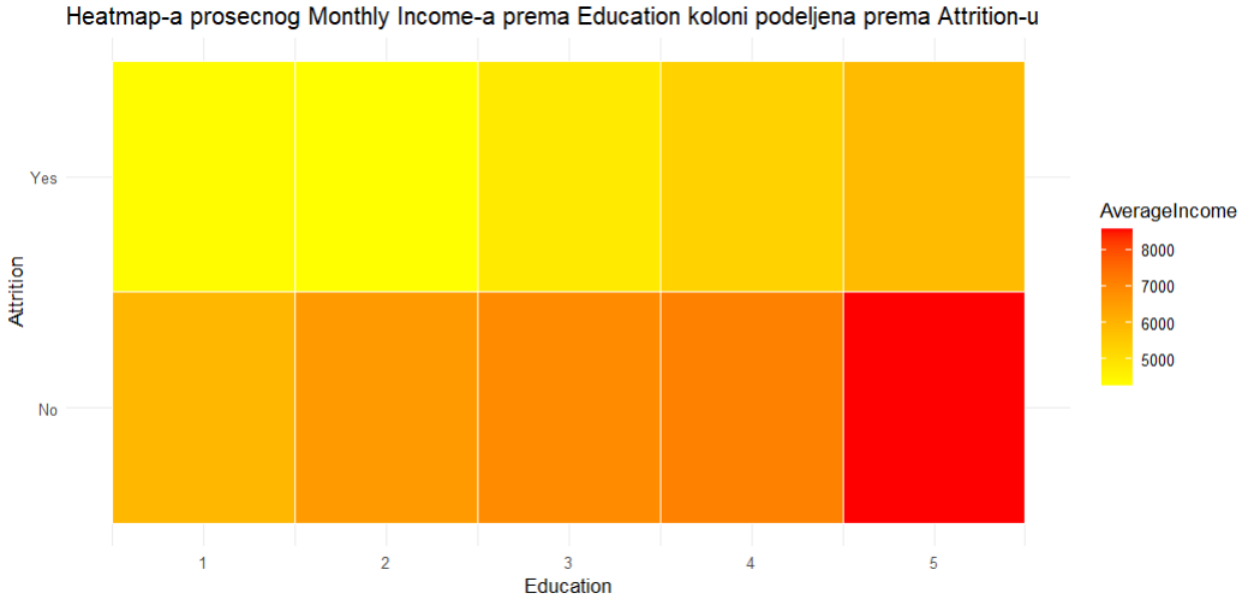
```
ggplot(hr_data, aes(x = JobRole, y = Age)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Age by Job Role", x = "Job Role", y = "Age") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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

- Monthly income, Education и Attrition

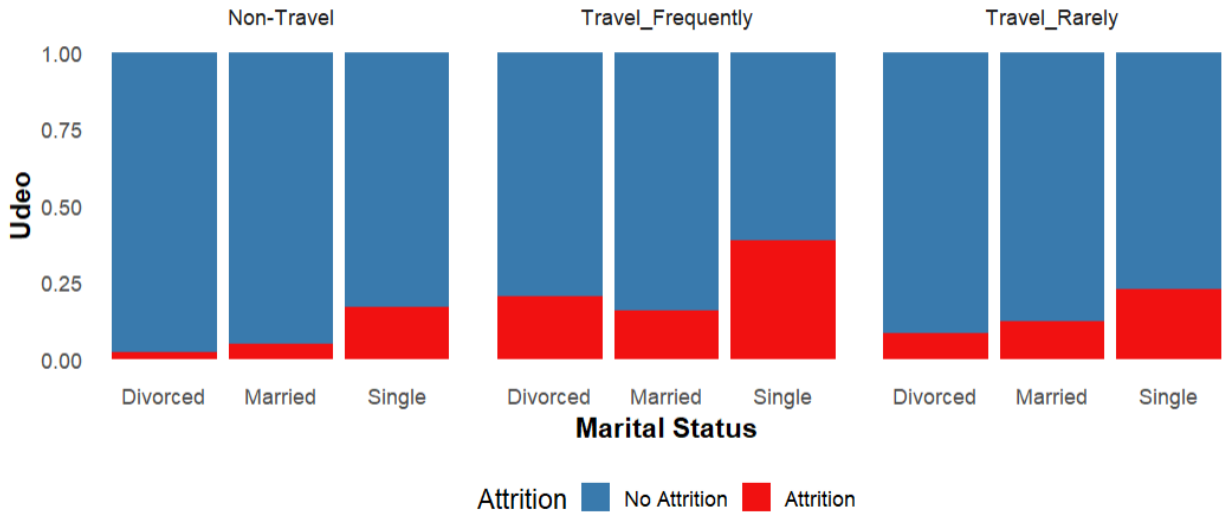
```
heatmap_data <- hr_data %>%
  group_by(Education, Attrition) %>%
  summarise(AverageIncome = mean(MonthlyIncome, na.rm = TRUE)) %>%
  ungroup()

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



На осову графика можемо закључити да фирму напуштају запослени са ниским месечним приходима и високим степеном едукације.

- **Martial 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)
```



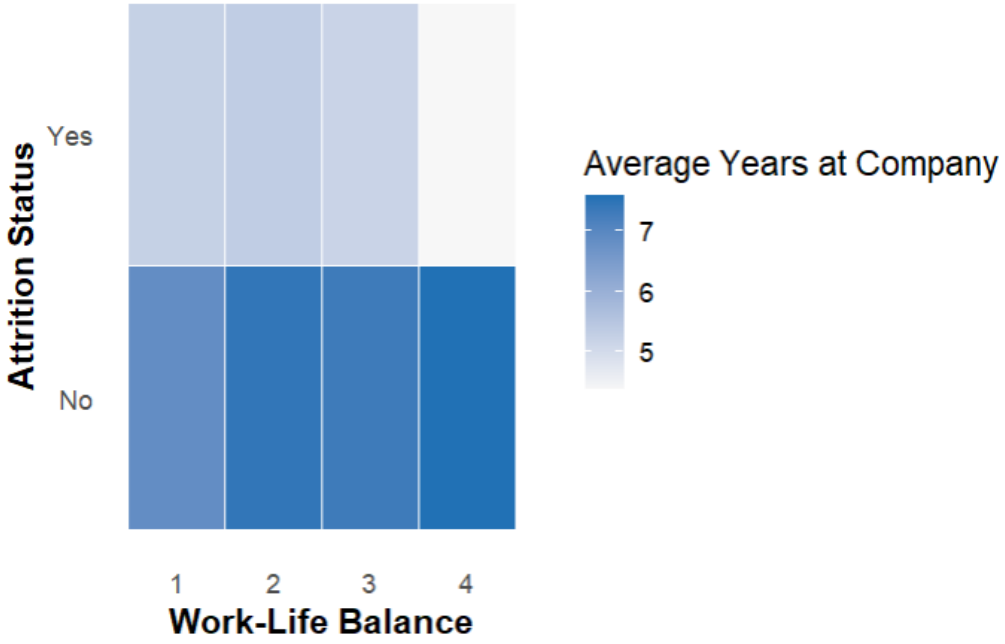
Са графика можемо закључити да посао најчешће напуштају запослени који често путују, а најчешће они који су слободни. Можемо и закључити и да невезано за учесталост путовања, слободни запослени најчешће напуштају, што је и логично.

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))
```

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

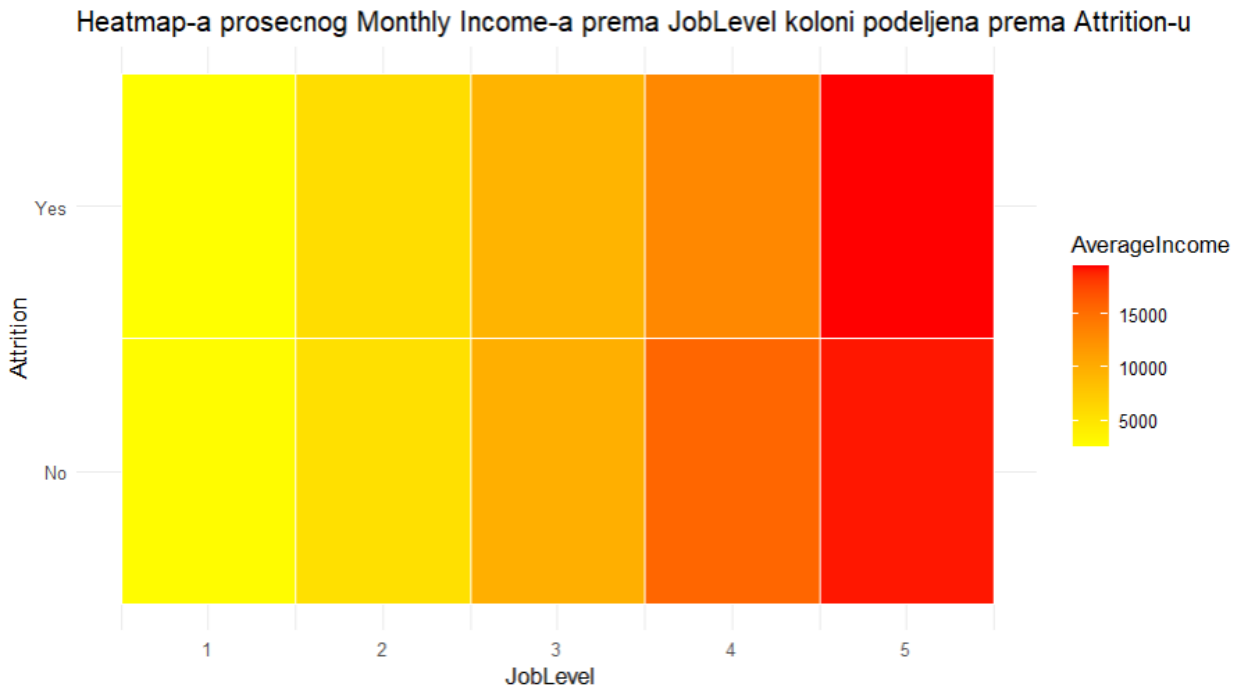
```
summary_data <- hr_data %>%
  group_by(WorkLifeBalance, Attrition) %>%
  summarise(AverageYears = mean(YearsAtCompany, na.rm = TRUE), .groups = 'drop')

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.

```
set.seed(21)

train_index <- createDataPartition(hr_data.final$Attrition_binary,
                                   p = 0.6,
                                   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 60:40. Где 60 процената скупа припада валидационом скупу док 40 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 3 модела: модел логистичке регресије, стабло одлучивања и „*Random Forest*“ модел.

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
         0 491  86
         1   3   8

               Accuracy : 0.8486
                 95% CI : (0.8171, 0.8766)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.3097

                  Kappa : 0.123

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99393
            Specificity : 0.08511
         Pos Pred Value : 0.85095
         Neg Pred Value : 0.72727
             Prevalence : 0.84014
         Detection Rate : 0.83503
   Detection Prevalence : 0.98129
      Balanced Accuracy : 0.53952

       'Positive' Class : 0
```

1Резултати логистичког модела

Према резултатима модела видимо да овај модел има високу прецизност као и сензитивност док је специфичност веома ниска. Балансирана прецизност

модела износи 53% што је поприлично ниско у односу на пријављену прецизност.

Један од разлога за ове резултате представља то да је у тренинг скуп ушло само 11 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 + MonthlyIncome
> 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
         0 349  30
         1 145  64

               Accuracy : 0.7024
                 95% CI : (0.6636, 0.7391)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.259

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7065
            Specificity : 0.6809
         Pos Pred Value : 0.9208
         Neg Pred Value : 0.3062
             Prevalence : 0.8401
         Detection Rate : 0.5935
   Detection Prevalence : 0.6446
      Balanced Accuracy : 0.6937

       'Positive' Class : 0
```

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

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

Следећи корак може бити смањење прага (енг. „treshold“) 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
         0 494  94
         1   0   0

               Accuracy : 0.8401
                 95% CI : (0.808, 0.8688)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.5275

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8401
         Neg Pred Value :    NaN
             Prevalence : 0.8401
         Detection Rate : 0.8401
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

Модел показује прецизност од 84% али Kappa вредност је 0 као и специфичност која износи 0%. 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
         0 494  94
         1   0   0

               Accuracy : 0.8401
                 95% CI : (0.808, 0.8688)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.5275

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8401
         Neg Pred Value :    NaN
             Prevalence : 0.8401
         Detection Rate : 0.8401
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

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

4.2.2 GLM модел и унакрсна валидација

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 374  46
         1 120  48

               Accuracy : 0.7177
                 95% CI : (0.6794, 0.7537)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.203

 Mcnemar's Test P-Value : 1.463e-08

            Sensitivity : 0.7571
            Specificity : 0.5106
         Pos Pred Value : 0.8905
         Neg Pred Value : 0.2857
             Prevalence : 0.8401
         Detection Rate : 0.6361
   Detection Prevalence : 0.7143
      Balanced Accuracy : 0.6339

       '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 71.77%, док *Sensitivity* износи 75.71% а „*Specifity*“ је 51.06%. У овом случају *Kappa* вредност износ 0.203 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
         0 492  88
         1   2   6

               Accuracy : 0.8469
                 95% CI : (0.8153, 0.8751)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 0.3509

                  Kappa : 0.095

 Mcnemar's Test P-Value : <2e-16

            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 бито с обзиром на неуравнотеженост података.

4.3.3 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
         0 357  46
         1 137  48

               Accuracy : 0.6888
                 95% CI : (0.6496, 0.726)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1676

 Mcnemar's Test P-Value : 2.872e-11

            Sensitivity : 0.7227
            Specificity : 0.5106
         Pos Pred Value : 0.8859
         Neg Pred Value : 0.2595
             Prevalence : 0.8401
         Detection Rate : 0.6071
   Detection Prevalence : 0.6854
      Balanced Accuracy : 0.6167

       'Positive' Class : 0
```

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

4.4 Резулати 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
Прецизност	70,24%	71,77%	68,88%
Балансирана прецизност	69,37%	63,39%	61,67%
Сензитивност	70,65%	70,65%	72,27
Специфичност	68.09%	51,06%	51,06
***Kappa* вредност**	0,259	0,203	0,1676

Логичка регресија се показује као најбољи модел према већини кључних метрика.

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 и модел *Random Forest.* 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