A blue and white logo with a building in the background

Description automatically generated

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

[1. Увод 3](#_Toc181559862)

[2. Припрема података 4](#_Toc181559863)

[2.1 Уклањање неважних колона 30](#_Toc181559864)

[3. Анализа података 31](#_Toc181559865)

[4. Креирање модела 45](#_Toc181559866)

[4.1 Логистичка регресија 46](#_Toc181559867)

[4.1.1 Логистички модел и унакрсна валидација 47](#_Toc181559868)

[4.2 GLM модел 48](#_Toc181559869)

[4.2.2 GLM модел и унакрсна валидација 50](#_Toc181559870)

[4.3 Random Forest модел 51](#_Toc181559871)

[4.3.3 Random Forest модел и унакрсна валидација 52](#_Toc181559872)

[4.4 Резулати модела 53](#_Toc181559873)

[5. Закључак 54](#_Toc181559874)

[6. Литература 55](#_Toc181559875)

# 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.

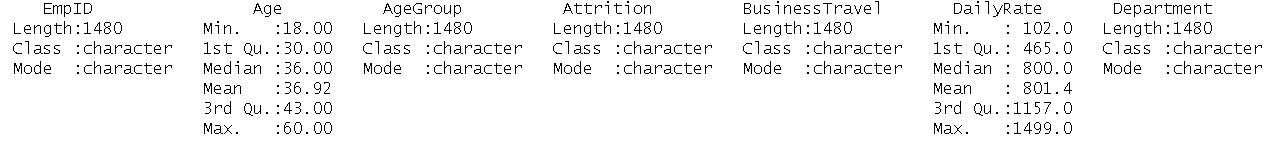
Основни циљ ове анализе је да се идентификују кључни фактори који утичу на одлазак запослених. Као што су старосна доб, удаљеност од посла, задовољство на послу, учесталост пословних путовања и сл.

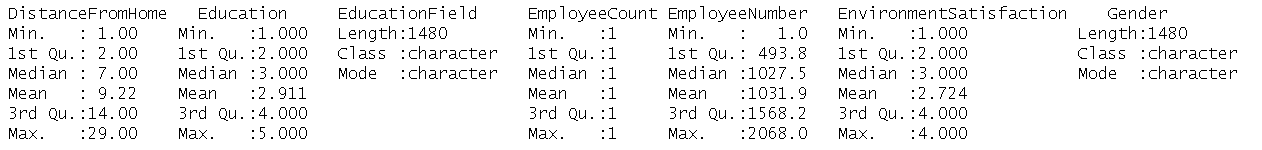
„**Attrition“** (у преводу осипање) или одлазак запослених подразумева када запослени својевољно или присилно напушта компанију. Висока стопа може довести до значајних трошкова за организацију, како новчаних тако и временских. Након одласка потребно је пронаћи, обучити и прилагодити нове запослене.

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

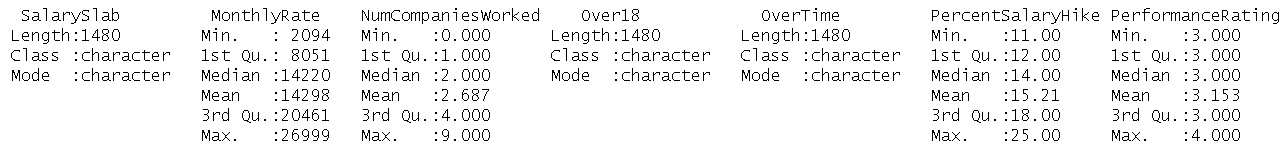
# Припрема података

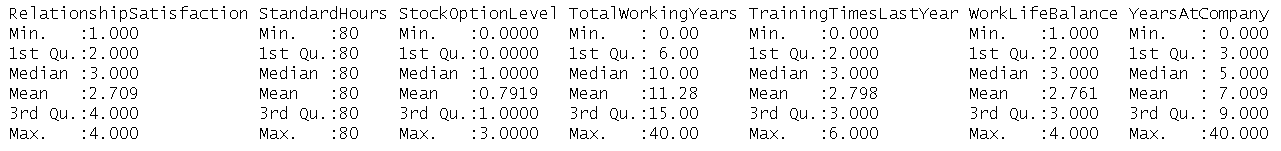
Скуп података ***HR\_Analytics*** обухвата 38 колона и садржи 1480 записа који пружају детаљан увид у различите аспекте живота запослених. Овај скуп података покрива како професионалне аспекте, попут задовољства послом, учешћа у обуци, и могућности за унапређење, тако и личне факторе као што су породични 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.



A close up of a white background

Description automatically generated





A black text on a white background

Description automatically generated

За анализу података креираћемо две функције („*plot\_histogram*“ и „*plot\_bar*“) помоћу којих ћемо добити графички приказ поделе података по колонама: A screenshot of a computer code

Description automatically generated

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. A close-up of a number

Description automatically generated



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.

A close-up of a document

Description automatically generated

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

A close-up of a computer screen

Description automatically generated

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.

A close up of a number

Description automatically generated A graph with a number of blue bars

Description automatically generated with medium confidence

**Колона „AgeGroup“**

Представља опсег година у којем се налази запослени.

A close-up of a text

Description automatically generated

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

A graph of blue rectangular bars

Description automatically generated with medium confidence

**Колона „Attrition“**

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

A graph with a red rectangle and blue rectangle

Description automatically generated

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“.

A close-up of blue text

Description automatically generated

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

A computer code on a white background

Description automatically generated

Сада можемо погледати расподелу ове колоне. A graph with blue rectangles

Description automatically generated

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.

A screenshot of a graph

Description automatically generated

Пошто колона не садржи аутлејере можемо погледати и њену расподелу.

**Колона „Department“**

Представља ком одељењу припада запослени.

A graph of a bar

Description automatically generated with medium confidence

**Колона „DistanceFromHome“**

Представља удаљеност места пребивалишта од фирме. Можемо проверити да ли ова колона садржи екстремне вредности који би негативно утицали на наш модел.

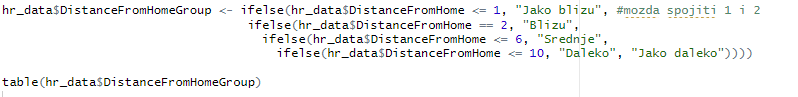
A graph with a line

Description automatically generatedС 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 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)



**Колона „Education“**

Представља ниво образовања запосленог.

A graph of a bar graph

Description automatically generated

**Колона „EducationField“**

Представља поље образовања запосленог.

A graph of a bar

Description automatically generated with medium confidence

**Колона „EnvironmentSatisfaction“**

Представља задовољство пословним окружењем.

A graph showing a bar graph

Description automatically generated with medium confidence

**Колона „Gender“**

Представља пол запосленог.

A graph of a bar graph

Description automatically generated with medium confidence

**Колона „HourlyRate“**

Представља уговорени износ запосленог по радном сату. Као и за DailyRate колону, можемо проверити екстремне вредности. A graph with a line

Description automatically generated

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. A graph of a bar graph

Description automatically generated with medium confidence

**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:

A graph of a bar

Description automatically generated with medium confidence

**Колона „JobRole“**

Представља звање запосленог унутар компаније.

A graph of blue rectangular bars

Description automatically generated with medium confidence

**Колона „JobSatisfaction“**

Представља задовољство запосленог на пословном плану. A graph of blue rectangular bars

Description automatically generated with medium confidence

**Колона „MartialStatus“**

Представља брачно стање запосленог.

A graph of a bar chart

Description automatically generated with medium confidence

**Колона „MonthlyIncome“**

Представља месечне приходе запосленог. Можемо проверити расподелу ове колоне на боксплоту и проверити екстремне вредностиA graph with a white line

Description automatically generated.

**Колона „SalarySlab“**

Представља категорију у којој се налази запослени на основу месечног прихода.A graph with blue squares

Description automatically generated

**Колона „MonthlyRate“**

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

Description automatically generated with medium confidence

**Колона „NumCompaniesWorked“**

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

A graph with a line

Description automatically generated

**Колона „OverTime“**

Представља индикатор за прековремени рад.

A graph showing a number of blue squares

Description automatically generated

**Колона „PercentSalaryHike“**

Представља проценат повишице.

A graph of a graph

Description automatically generated

**Колона „PerformanceRating“**

Представља оцену перформансе запосленог.

A diagram of a pie chart

Description automatically generated

**Колона „RelationshipStatus“**

Представља тренутни статус везе запослених. Ова информација може бити значајна за анализу различитих аспеката радне динамике и задовољства запослених. A graph showing a bar graph

Description automatically generated with medium confidence

**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. A graph of a bar graph

Description automatically generated with medium confidence

**Колона „TotalWorkingYears“**

Представља укупан радни стаж запосленог. A graph of a graph

Description automatically generated with medium confidence

**Колона „TrainingTimesLastYear“**

Представља колико је пута радник био на професионалном обучавању у претходној години.A graph with a bar graph

Description automatically generated with medium confidence

**Колона „WorkLifeBalance“**

Представља ниво којим запослени успева да се посвети како пословном тако и приватном животу. A graph of a bar graph

Description automatically generated with medium confidence

**Колона „YearsAtCompany“**

Представља колико година је запослени провео у фирми. A graph with a blue line

Description automatically generated with medium confidence

**Колона „YearsInCurrentRole“**

Представља колико година је запослени провео на истој позицији.A graph with blue bars

Description automatically generated with medium confidence

**Колона „YearsSinceLastPromotion“**

Представља колико година је прошло од последње промоције.A graph with a blue line

Description automatically generated

**Колона „YearsWithCurrentManager“**

Ова колона показује број година које запослени провео под тренутним менаџером. A close-up of a number

Description automatically generated

Видимо да ова колона садржи 57 НА 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. A screenshot of a computer

Description automatically generated

Сада можемо погледати расподелу за ову колону.

A graph of a graph

Description automatically generated

## 2.1 Уклањање неважних колона

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



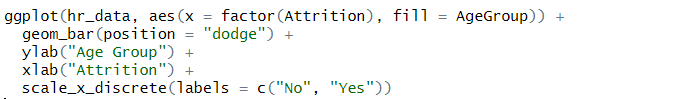
* Даљом анализом приметили смо да су вредности колона **Over18** и **StandardHours** исте за све запослене, па ћемо их такође избацити:

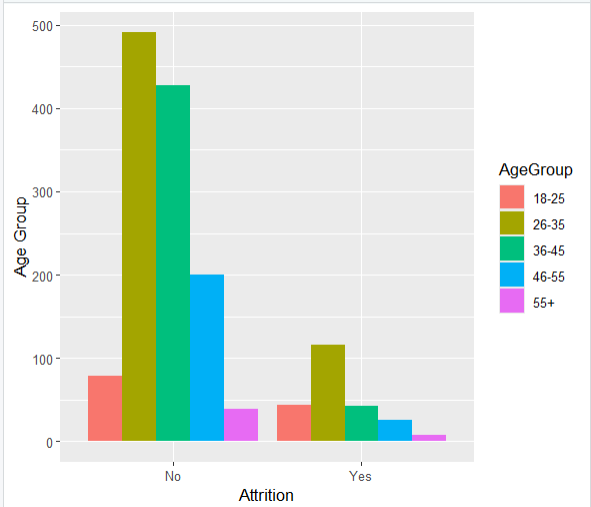


# Анализа података

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

* **Age group и Attrition**



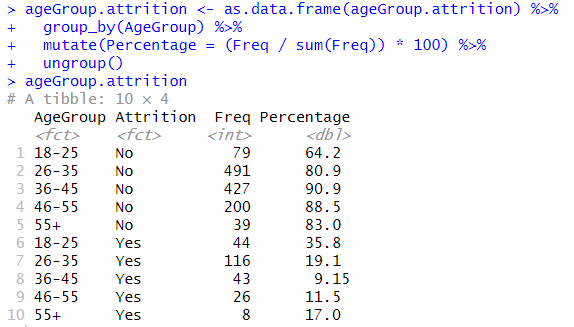


На основу графика можемо да закључимо да запослени који процентуално више напуштају компанију припадају млађим старосним групама. Тај тренд је најизраженији у размаку од 18 до 35 година, након тога је процентуално мања одлазност из компаније.

* **Age и Monthly Income**

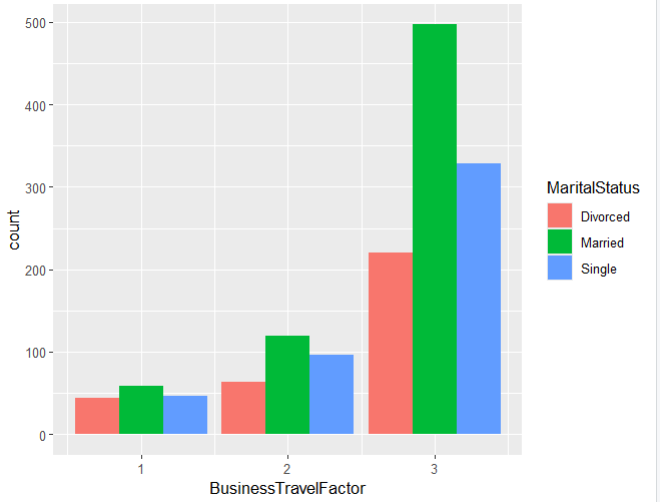






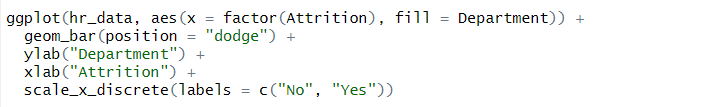
У овом примеру смо упоредили године запослених и месечна примања и примећујемо да повезаност постоји. Најизраженија је за млађе запослене и можемо приметити како са порастом година расту и приманја уз доста изузезака након 30-те године.

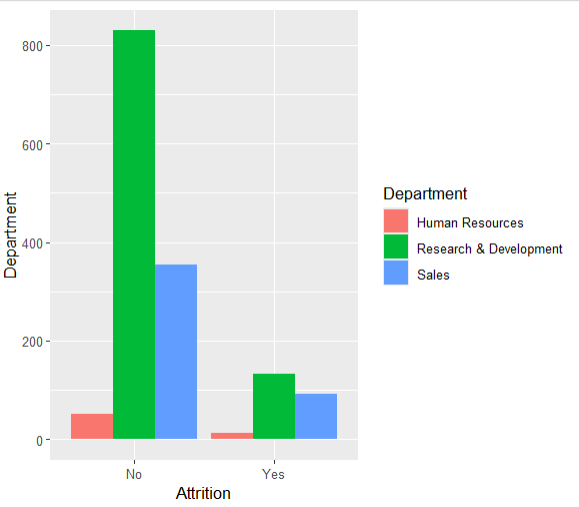
* **Martial status и Bussines travel factor**

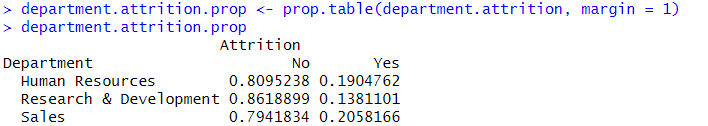


На овом графику можемо видети однос брачног статуса и учесталости пословних путовања запослених у компанији. Генерално примећујемо пре свега да највећи број запослених често путује. Затим можемо приметити да у свим категоријама има највише ожењених, али да су сразмерно броју по катаегорији сви уједначени,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

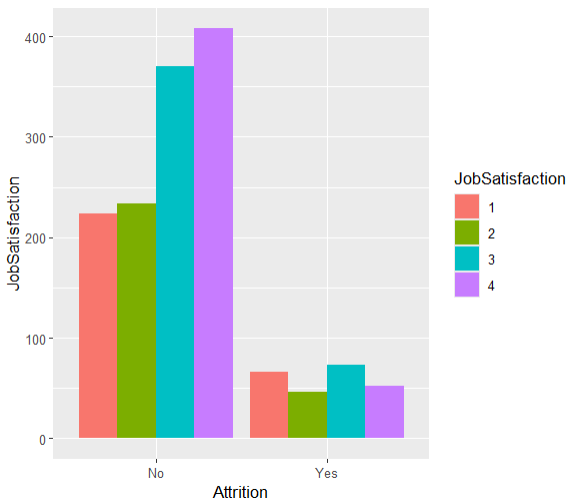
* **Attrition и Department**

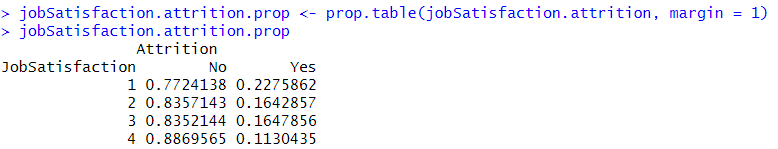




Примећујемо да је мала разлика у процентима међу одељењима, али да ”Sales” има највећи проценат људи који напуштају фирму.

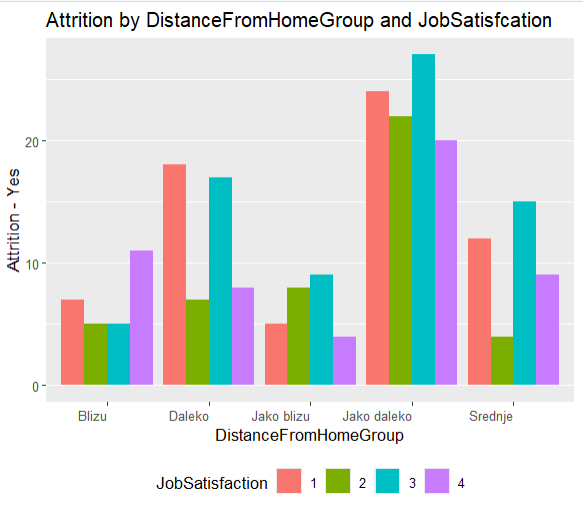
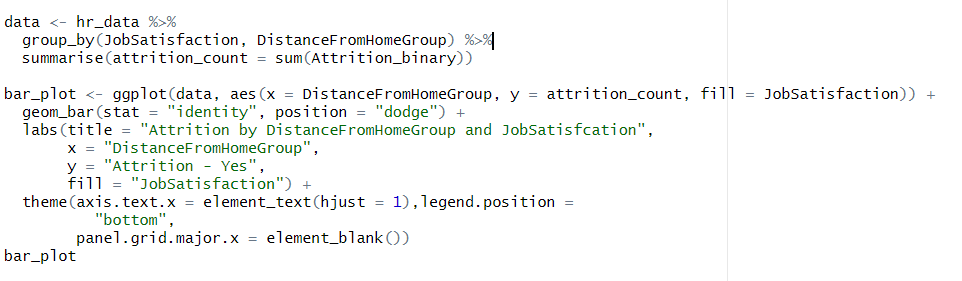
* **Job Satisfaction и Attrition**





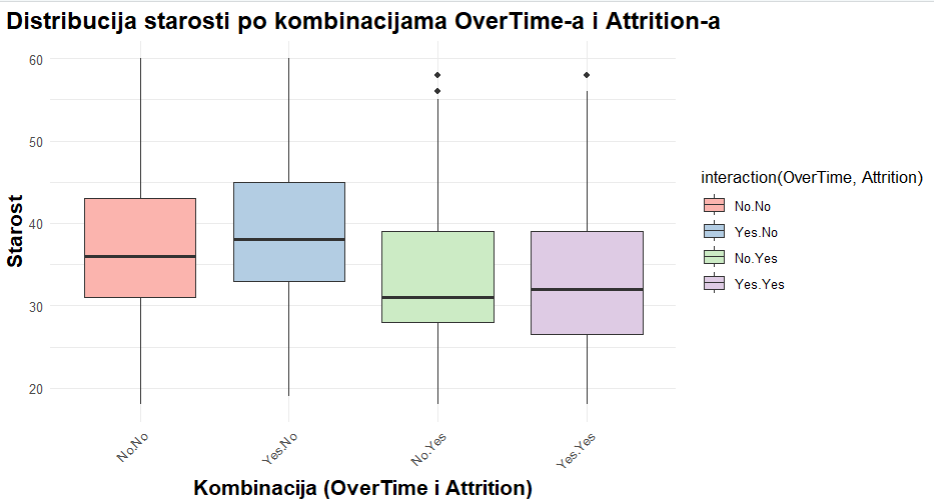
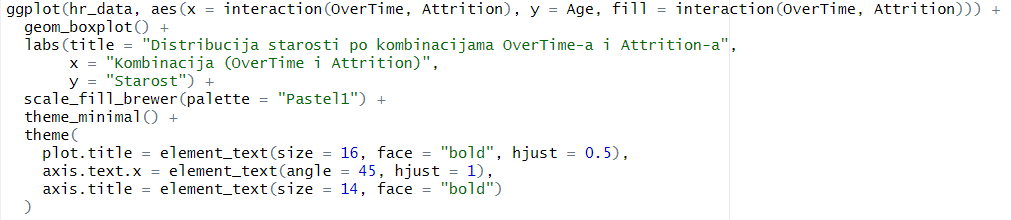
Запослени који су незадовољни послом чешће напуштају фирму, али то није превише изражено.

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



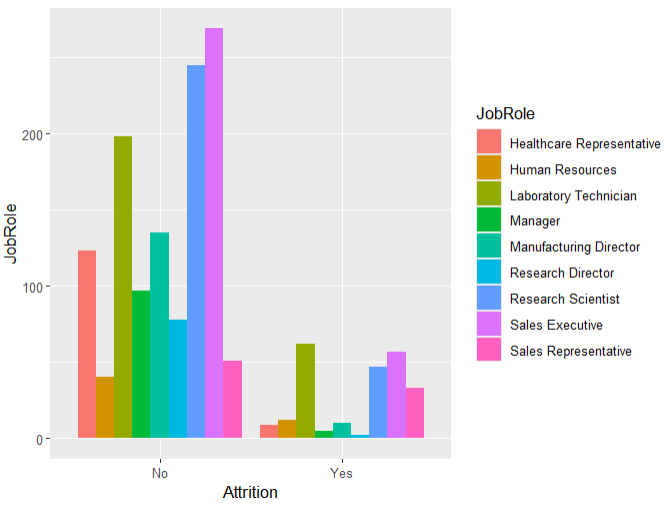
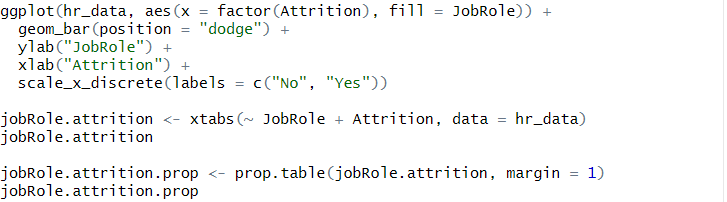
Из овог графика можемо да закључимо да запослени који живе далеко од фирме чешће напуштају исту. Удаљеност од посла од куће може бити јако добар предиктор.

* **Age, OverTime и Attrition**



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.

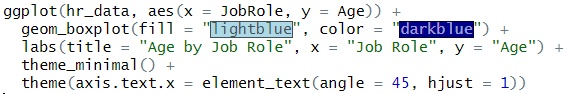
* **Job role и Attrition**

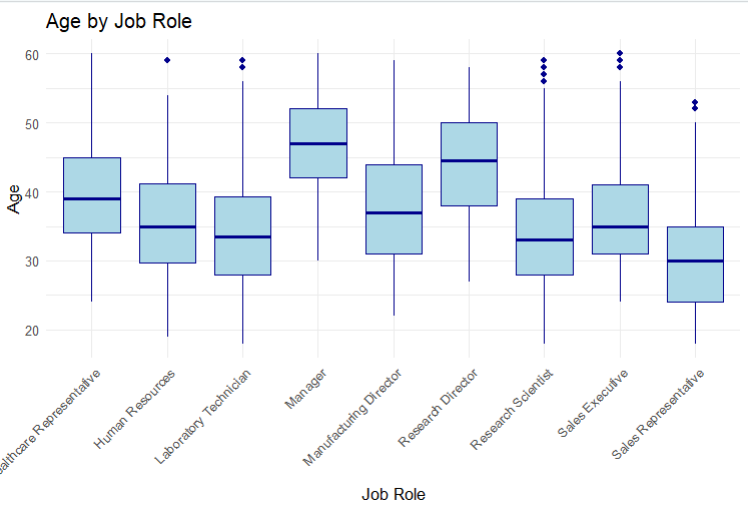


Позиције које најчешће напуштају фирму јесу Sales Representative затим Laboratory Technician и Human Resources, односно најниже позиције у фирми.

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

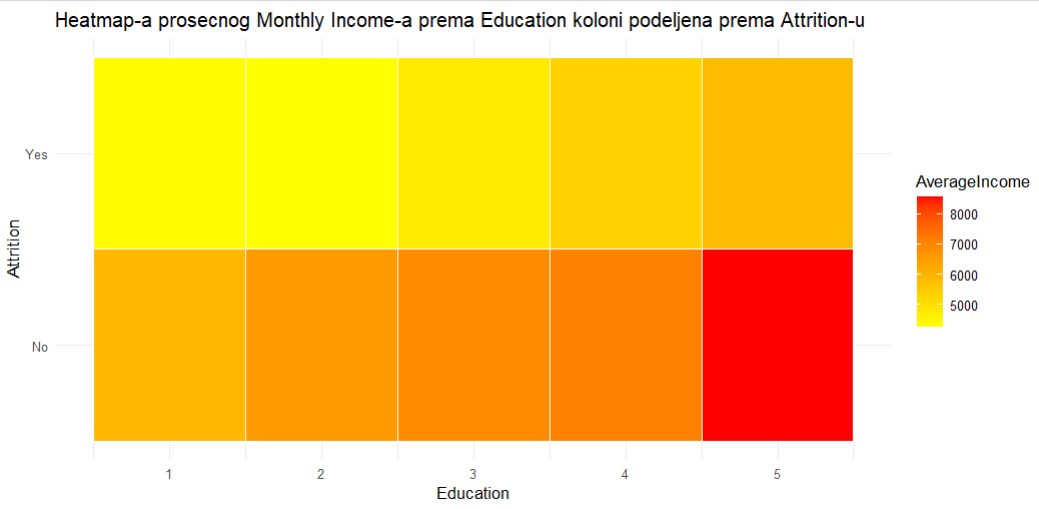
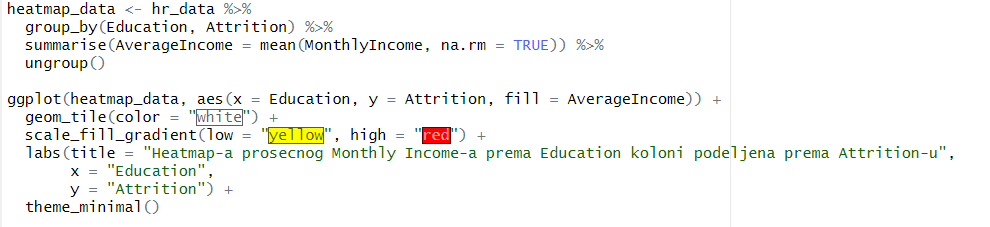
* **Job role и Age**





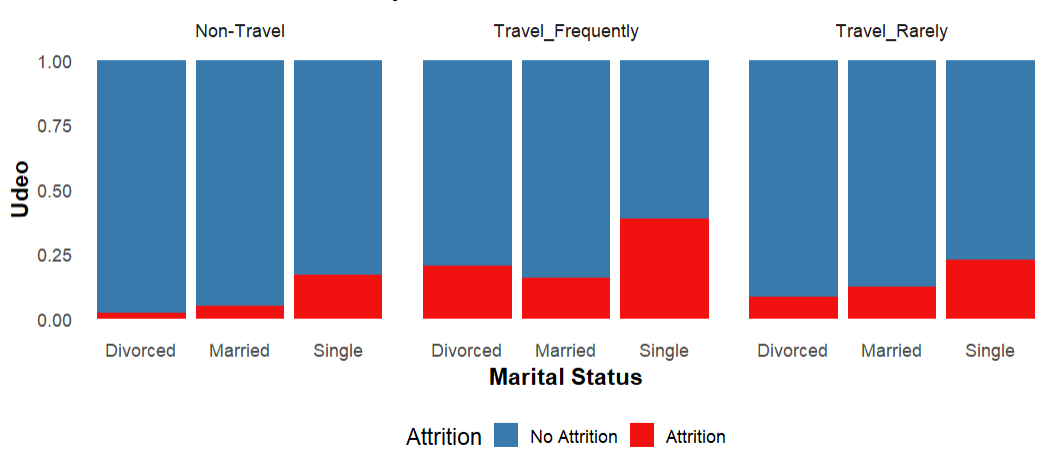
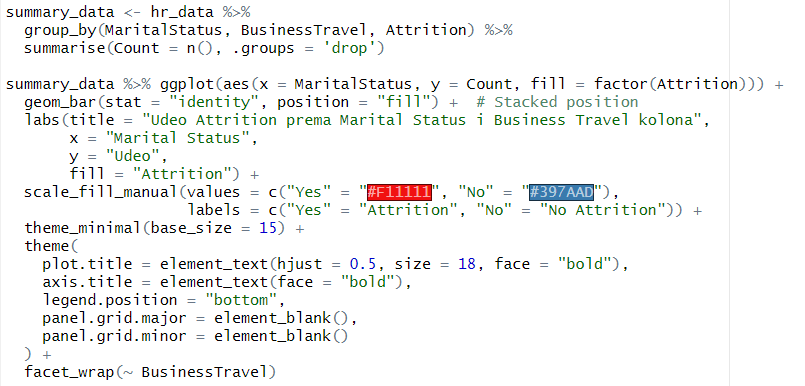
За више позиције у оквиру фирме, потребне су старије особе, млађе особе најчешће раде на нижим позивијама које се чешће напуштају.

* Monthly income, Education и Attrition



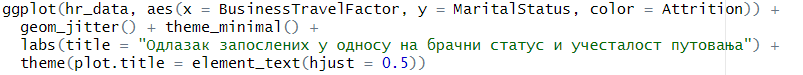
На осову графика можемо закључити да фирму напуштају запослени са ниским месечним приходима и високим степеном едукације.

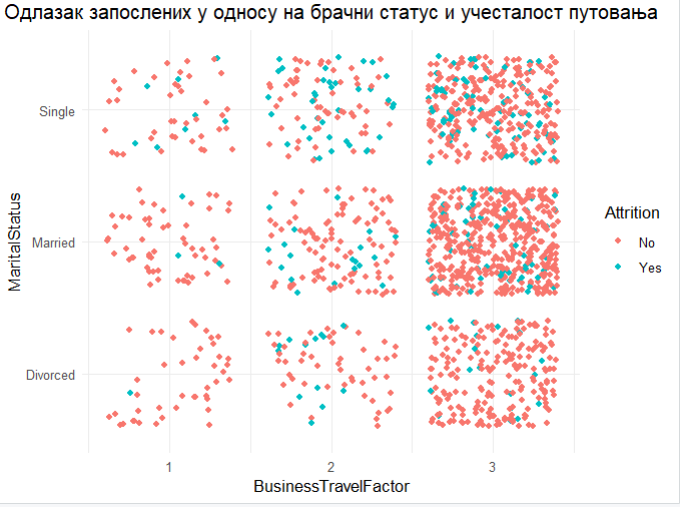
* **Martial status, Business Travel и Attrition**



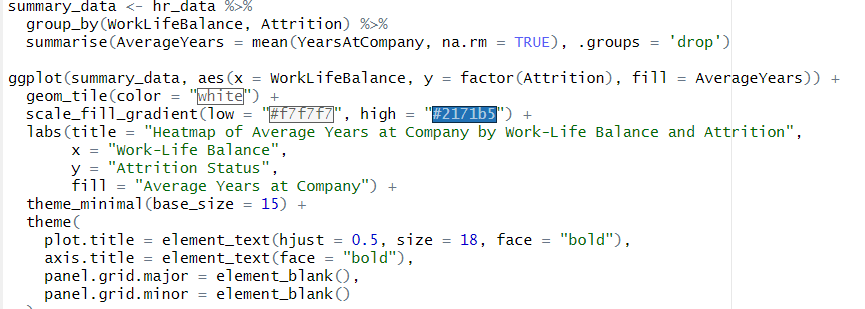
Са графика можемо закључити да посао најчешће напуштају запослени који често путују, а најчешће они који су слободни. Можемо и закључити и да невезано за учесталост путовања, слободни запослени најчешће напуштају, што је и логично.

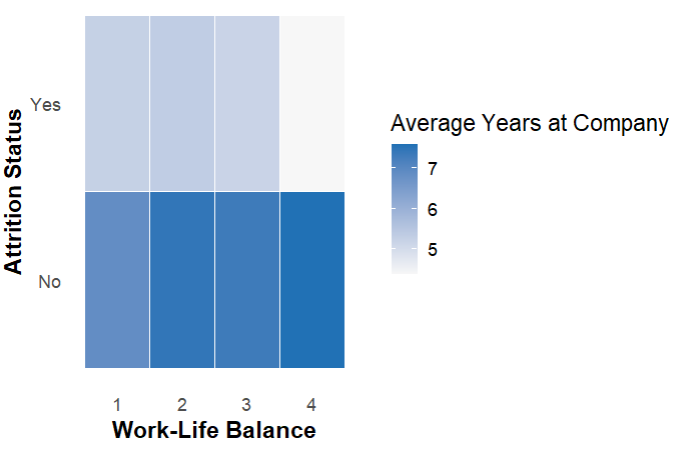
У наставку се налази још један график који нам је помогао да донесемо ове закључке.





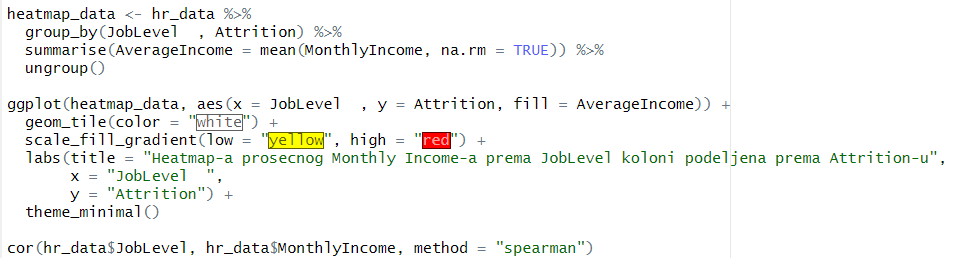
* **Work life balance, Average years и Attrition**

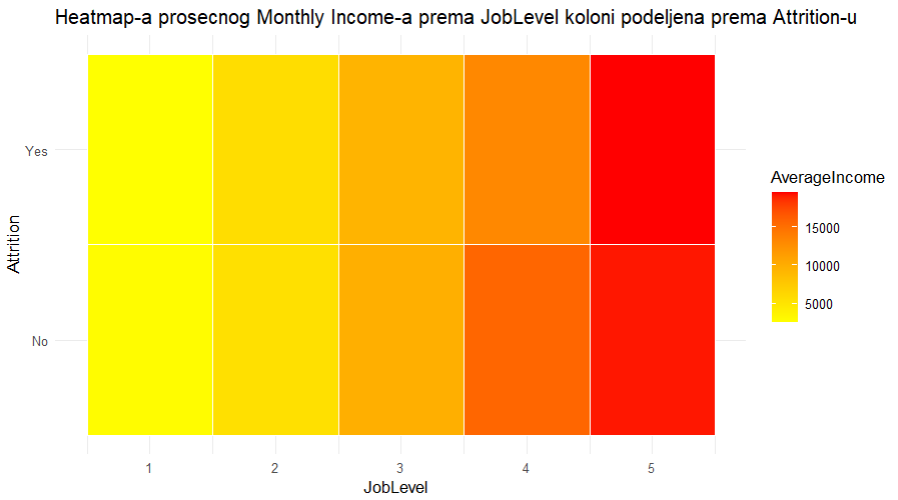




Закључујемо да запослени који су пронашли добар баланс дуже остају у фирми, што значи и да је мања вероватноћа да ће је напустити.

* Job level и Monthly income





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.

# Креирање модела

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

A screenshot of a computer code

Description automatically generated

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.

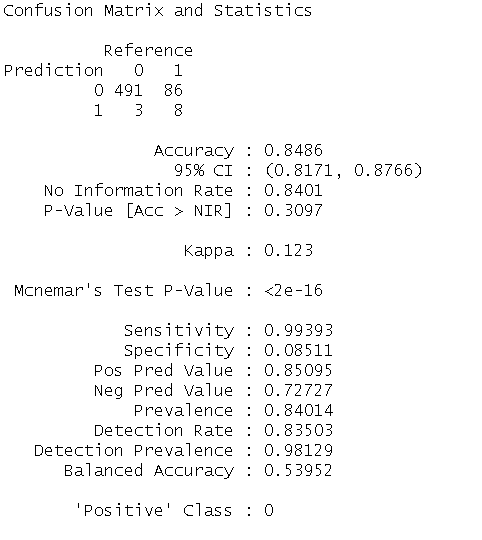
Развијаћемо 3 модела: модел логистичке регресије, стабло одлучивања и „*Random Forest*“ модел.

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“.

A screenshot of a computer code

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

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*“.

A black text on a white background

Description automatically generated A screenshot of a computer

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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.

A black text on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

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

A close up of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

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.

# 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.

# Литература

* *Увод у науку о подацима* - вежбе, предавања и материјали предмета
* *GGPlot2* - [документација](https://ggplot2.tidyverse.org/reference)
* *Caret* - [документација](https://topepo.github.io/caret/)
* *GLM* - [документација](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm)