מעבדה בניתוח נתונים עם R

עבודת סיום קורס

HR Analytics: Job Change of Data Scientists

Predict who will move to a new job

מגישים:

רותם לוי 311577126

אלון קרסניצקי 204498091

מרצה הקורס:

אבי זכאי

תוכן עניינים

רשימת תרשימים וטבלאות

חלק ב'

1. תיאור הבעיה

חברה הפעילה התחום ה Big Data רוצה להעסיק מדעני נתונים.
כדי להתקבל לחברה על המועמדים לעבור הכשרות וקורסים.
החברה רוצה לדעת מי מבין המועמדים האלה באמת רוצה לעבוד תקופה ממושכת בחברה לאחר ההכשרה – במטרה להפחית בעלויות גיוס העובדים וההכשרה שלהם. הדבר משפיע על איכות ההכשרה, תכנון הקורסים וסיווג המועמדים.

<u>תיאור משתנה המטרה:</u>

- target: 0 - Not looking for job change, 1 - Looking for a job change ***(Categorical)***

<u>תיאור הפיצ'רים:</u>

- Enrollee id : Unique ID for candidate ***(Categorical)***
- city: City code ***(Categorical)***
- city_ development _index : Development index of the city ***(Numeric)***
- gender: Gender of candidate ***(Categorical)***
- Relevant experience: Relevant experience of candidate ***(Categorical)***
- Enrolled university: Type of University course enrolled if any ***(Categorical)***
- Education level: Education level of candidate ***(Categorical)***
- Major discipline: Education major discipline of candidate ***(Categorical)***
- experience: Candidate total experience in years ***(Numeric)***
- Company size: No of employees in current employer's company ***(Categorical)***
- Company type : Type of current employer ***(Categorical)***
- Last new job: Difference in years between previous current job ***(Numeric)***
- Training hours: training hours completed ***(Numeric)***

```
Rows: 19,158
Columns: 14
$ ı..enrollee_id         <int> 8949, 29725, 11561, 33241, 666, 21651, 28806, 402...
$ city                   <fct> city_103, city_40, city_21, city_115, city_162, c...
$ city_development_index <dbl> 0.920, 0.776, 0.624, 0.789, 0.767, 0.764, 0.920, ...
$ gender                 <fct> Male, Male, NA, NA, Male, NA, Male, Male, Male, N...
$ relevent_experience    <fct> Has relevent experience, No relevent experience, ...
$ enrolled_university    <fct> No Enrollment, No Enrollment, Full time, NA, No E...
$ education_level        <fct> Graduate, Graduate, Graduate, Graduate, Masters, ...
$ major_discipline       <fct> STEM, STEM, STEM, Business Degree, STEM, STEM, NA...
$ experience             <dbl> 21, 15, 5, 0, 21, 11, 5, 13, 7, 17, 2, 5, 21, 2, ...
$ company_size           <fct> NA, 50-99, NA, NA, 50-99, NA, 50-99, <10, 50-99, ...
$ company_type           <fct> NA, Pvt Ltd, NA, Pvt Ltd, Funded Startup, NA, Fun...
$ last_new_job           <dbl> 1, 5, 0, 0, 4, 1, 1, 5, 1, 5, 0, 1, 3, 0, 0, 5, 0...
$ training_hours         <int> 36, 47, 83, 52, 8, 24, 24, 18, 46, 123, 32, 108, ...
$ target                 <fct> 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0...
```

מטרת המחקר:

מטרת המחקר היא לחזות אילו מבין המועמדים לעבודה עתידים להישאר בה לאורך זמן.
במהלך המחקר אנו נבנה מודלים שונים אשר יחזו האם העובד לא מחפש שינוי עבודה (0) או שהעובד כן מחפש שינוי עבודה (1).

במהלך המחקר נבחן מספר מודלים שונים ומניפולציות שונות על הנתונים וננסה לראות אילו תכונות הן המשפיעות ביותר על האם העובד מחפש / לא מחפש שינוי עבודה.

מטרת המחקר היא (חיזוי), לבנות מודל שיחזה האם מועמד מסוים מחדש שינוי עבודה.
המטרה הנוספת במחקר היא לדעת (הסקה סטטיסטית) אילו תכונות / פיצ'רים קשורים למשתנה המטרה – לכך שהאם העובד יחפש עבודה אחרת.

במהלך המחקר אנו נבדוק מספר מודלים שונים כמו Logistic Regression, KNN, Decision Tree & Random Forest

ונבדוק איזה מודל הוא הטוב ביותר לחיזוי הבעיה ואילו תכונות משפיעות הכי הרבה על ההחלטה.

General :

summery:

```
> summary(df)
  enrollee_id          city       city_development_index    gender
 Min.   :    1   city_103:4355   Min.   :0.4480         Female: 1238
 1st Qu.: 8554   city_21 :2702   1st Qu.:0.7400         Male  :13221
 Median :16983   city_16 :1533   Median :0.9030         other :  191
 Mean   :16875   city_114:1336   Mean   :0.8288         NA's  : 4508
 3rd Qu.:25170   city_160: 845   3rd Qu.:0.9200
 Max.   :33380   city_136: 586   Max.   :0.9490
                 (Other) :7801
              relevent_experience    enrolled_university       education_level
 Has relevant experience:13792    Full time    : 3757    Graduate      :11598
 No relevant experience : 5366    No Enrollment:13817    High School   : 2017
                                  Part time    : 1198    Masters       : 4361
                                  NA's         :  386    Phd           :  414
                                                         Primary School:  308
                                                         NA's          :  460

       major_discipline   experience      company_size                company_type
 Arts           :  253   Min.   : 0.0   50-99    :3083   Early Stage Startup: 603
 Business Degree:  327   1st Qu.: 4.0   100-500  :2571   Funded Startup     :1001
 Humanities     :  669   Median : 9.0   10000+   :2019   NGO                : 521
 No Major       :  223   Mean   :10.1   10-49    :1471   Other              : 121
 Other          :  381   3rd Qu.:16.0   1000-4999:1328   Public Sector      : 955
 STEM           :14492   Max.   :21.0   (Other)  :2748   Pvt Ltd            :9817
 NA's           : 2813   NA's   :65     NA's     :5938   NA's               :6140
  last_new_job training_hours   target
 Min.   :0     Min.   :  1.00   0:14381
 1st Qu.:1     1st Qu.: 23.00   1: 4777
 Median :1     Median : 47.00
 Mean   :2     Mean   : 65.37
 3rd Qu.:3     3rd Qu.: 88.00
 Max.   :5     Max.   :336.00
 NA's   :423
```

Table 1 - summery data

Data Frame:

```
> str(df)
'data.frame':   19158 obs. of  14 variables:
 $ enrollee_id           : int  8949 29725 11561 33241 666 21651 28806 402 27107 699 ...
 $ city                  : Factor w/ 123 levels "city_1","city_10",..: 6 78 65 15 51 58 50 84
 6 6 ...
 $ city_development_index: num  0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...
 $ gender                : Factor w/ 3 levels "Female","Male",..: 2 2 NA NA 2 NA 2 2 2 NA ...
 $ relevent_experience   : Factor w/ 2 levels "Has relevant experience",..: 1 2 2 2 1 1 1 1 1
 1 ...
 $ enrolled_university   : Factor w/ 3 levels "Full time","No Enrollment",..: 2 2 1 NA 2 3 2
 2 2 2 ...
 $ education_level       : Factor w/ 5 levels "Graduate","High School",..: 1 1 1 1 3 1 2 1 1
 1 ...
 $ major_discipline      : Factor w/ 6 levels "Arts","Business Degree",..: 6 6 6 2 6 6 NA 6 6
 6 ...
 $ experience            : num  21 15 5 0 21 11 5 13 7 17 ...
 $ company_size          : Factor w/ 8 levels "<10","10-49",..: NA 6 NA NA 6 NA 6 1 6 5 ...
 $ company_type          : Factor w/ 6 levels "Early Stage Startup",..: NA 6 NA 6 2 NA 2 6 6
 6 ...
 $ last_new_job          : num  1 5 0 0 4 1 1 5 1 5 ...
 $ training_hours        : int  36 47 83 52 8 24 24 18 46 123 ...
 $ target                : Factor w/ 2 levels "0","1": 2 1 1 2 1 2 1 2 2 1 ...
```

Table 2 - variables

number of NA in each Colum:

```
sapply(df, function(x) sum(is.na(x)))
          enrollee_id                  city city_development_index
                    0                     0                       0
               gender    relevent_experience    enrolled_university
                 4508                     0                     386
     education_level       major_discipline             experience
                  460                  2813                     65
        company_size           company_type           last_new_job
                 5938                  6140                    423
       training_hours                target
                    0                     0
```

Table 3 - NA`s

סטטיסטיקה תיאורית של הפיצ'רים:

city:

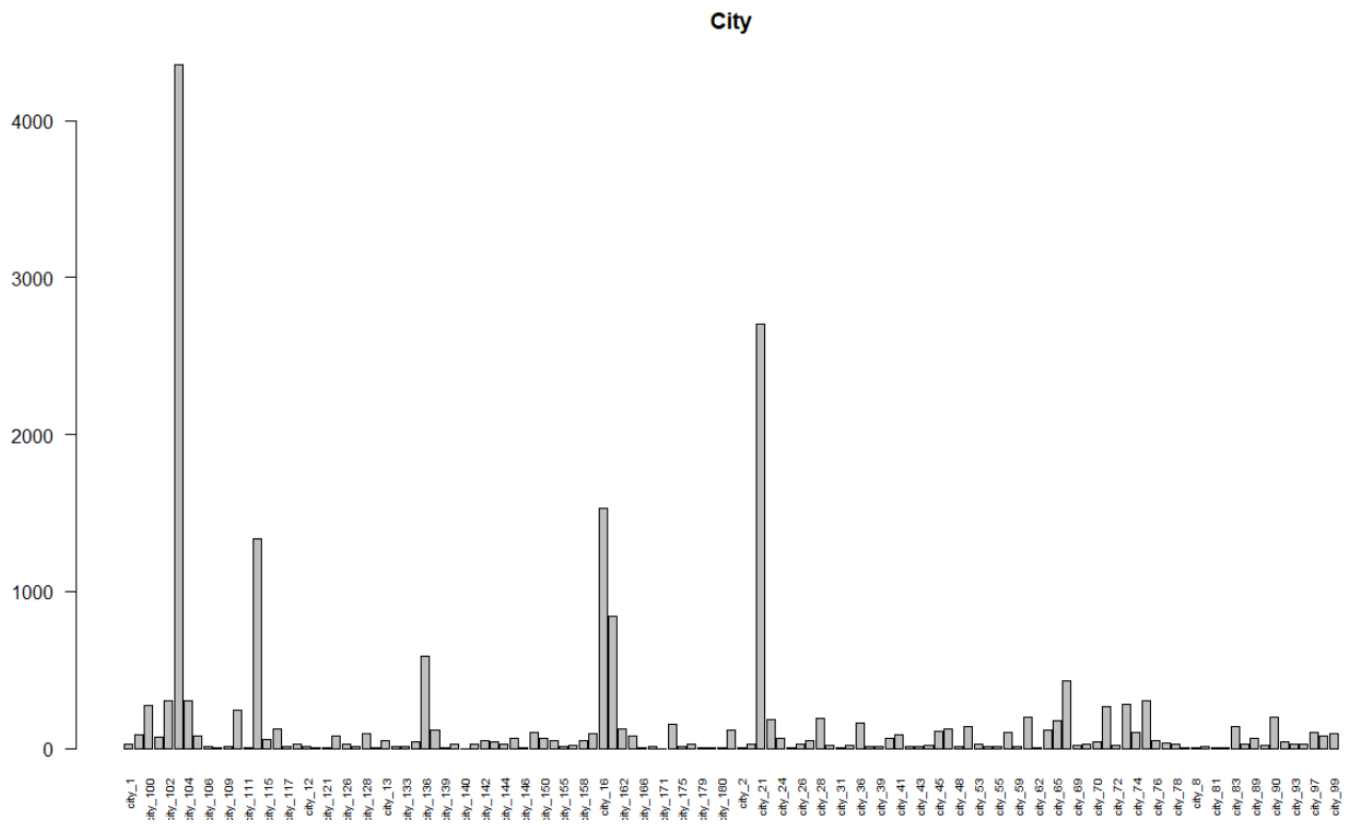


Figure 1 – city

city development index:

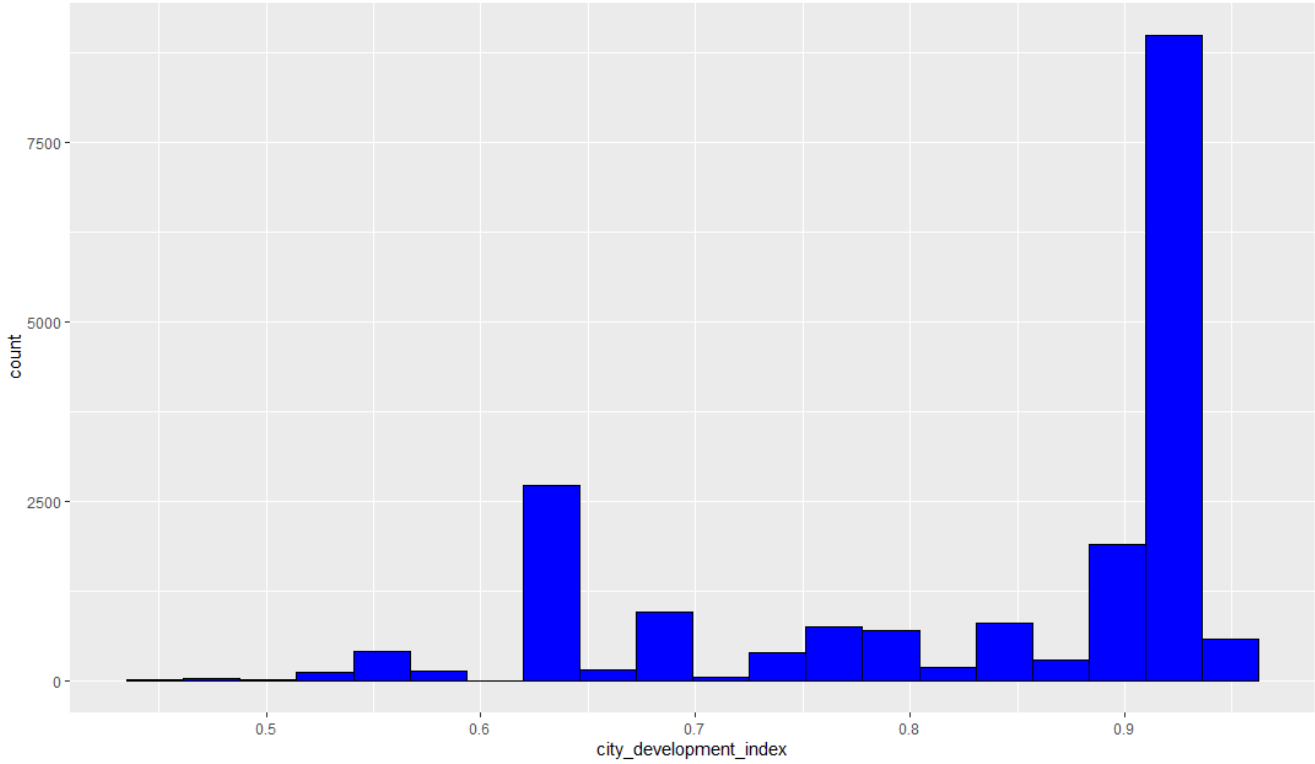


Figure 2- city development index

```
summary(df$city_development_index)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4480  0.7400  0.9030  0.8288  0.9200  0.9490 
```

Gender:

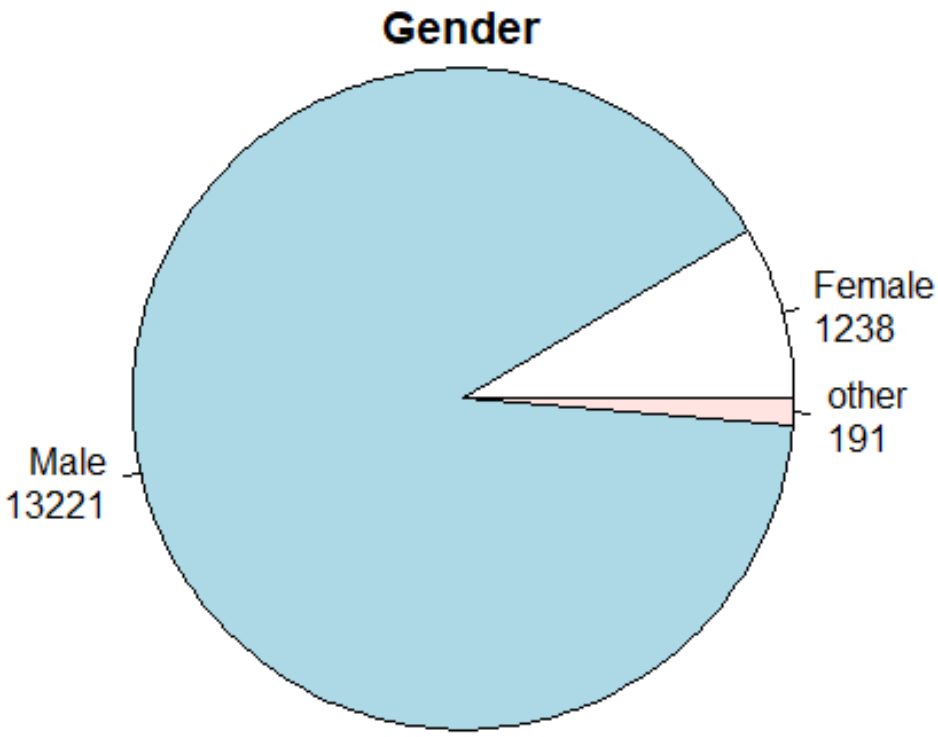


Figure 3- Gender

```
> ft
  Gender  Freq Rel.Freq
1   Male 13221    90.25
2 Female  1238     8.45
3  other   191     1.30
```

```
> summary(df$gender)
Female   Male  other   NA's
  1238  13221    191   4508
```

relevent experience:

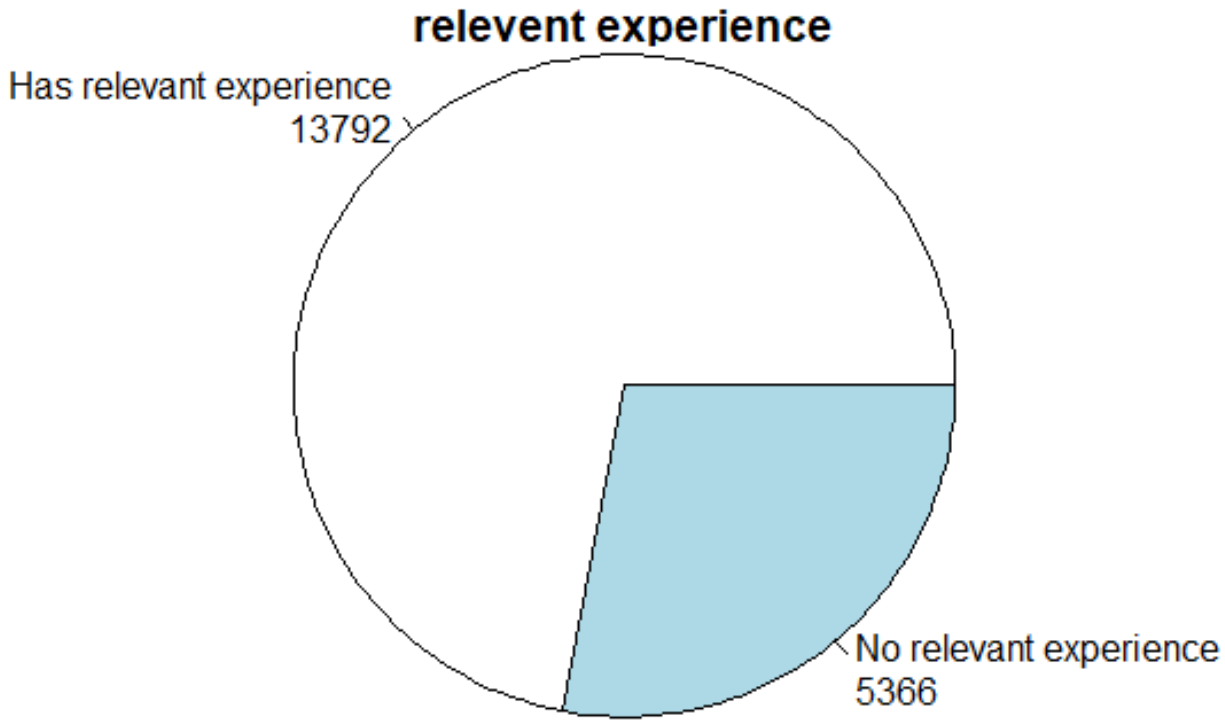


Figure 4 - relevent experience

```
> ft
      relevent experience  Freq Rel.Freq
1 Has relevant experience 13792    71.99
2  No relevant experience  5366    28.01
> summary(df$relevent_experience)
Has relevant experience  No relevant experience
                  13792                    5366
```

enrolled university:

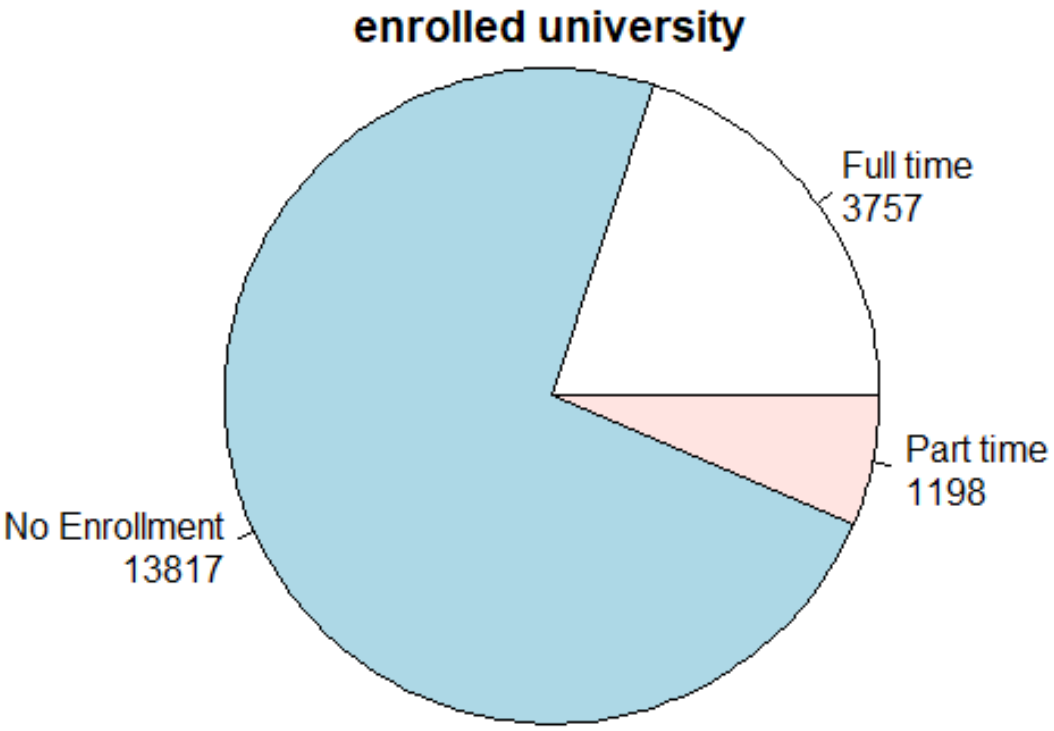


Figure 5 - enrolled university

```
> ft
  enrolled university  Freq Rel.Freq
1       No Enrollment 13817    73.60
2           Full time  3757    20.01
3           Part time  1198     6.38

> summary(df$enrolled_university)
    Full time No Enrollment     Part time          NA's
         3757         13817          1198           386
```

education level:

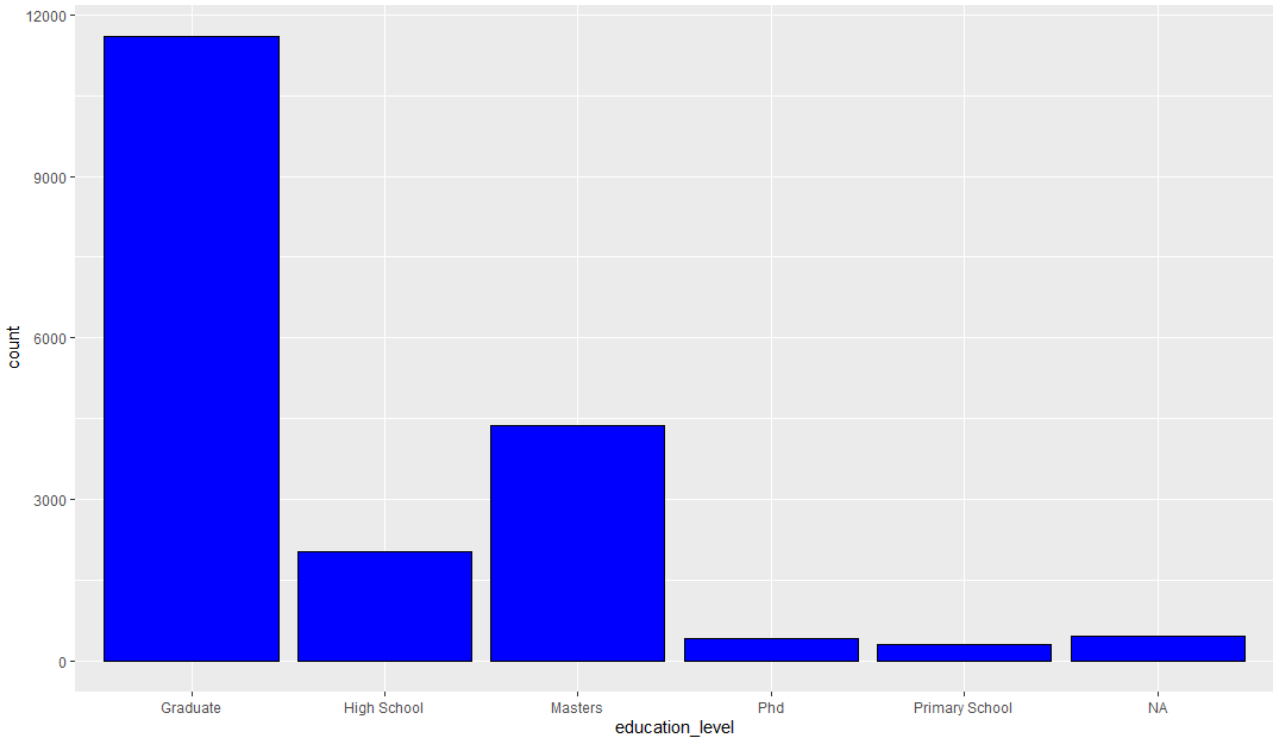


Figure 6 - education level

```
> ft
  education_level  Freq Rel.Freq
1        Graduate 11598    62.03
2         Masters  4361    23.32
3     High School  2017    10.79
4             Phd   414     2.21
5  Primary School   308     1.65
```

```
summary(df$education_level)
    Graduate     High School         Masters             Phd Primary School            NA's
       11598            2017            4361             414            308             460
```

major discipline:

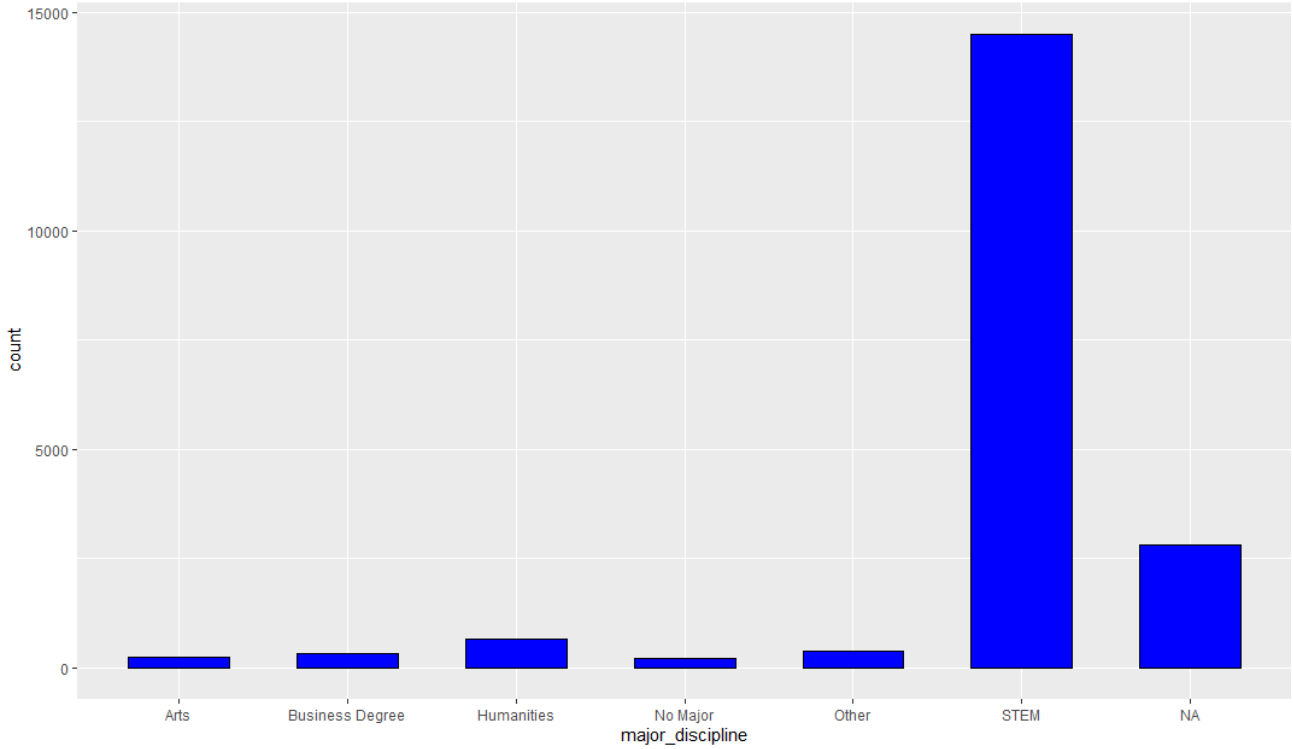


Figure 7 - major discipline

```
> ft
  major discipline  Freq Rel.Freq
1             STEM 14492    88.66
2       Humanities   669     4.09
3            Other   381     2.33
4  Business Degree   327     2.00
5             Arts   253     1.55
6         No Major   223     1.36
```

```
summary(df$major_discipline)
         Arts Business Degree      Humanities        No Major
          253            327             669             223
        Other           STEM            NA's
          381          14492            2813
```

Experience:

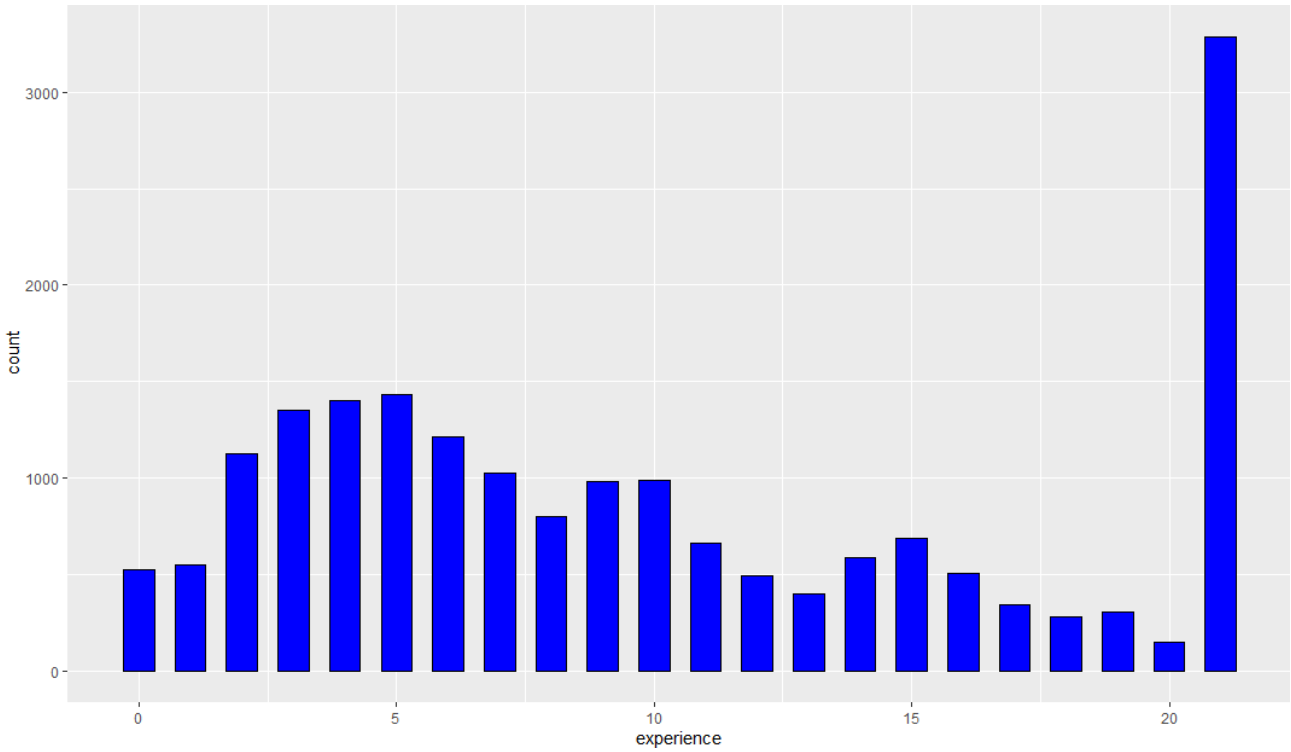


Figure 8 - Experience

```
summary(df$experience)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.0     4.0     9.0    10.1    16.0    21.0      65 
```

company size:

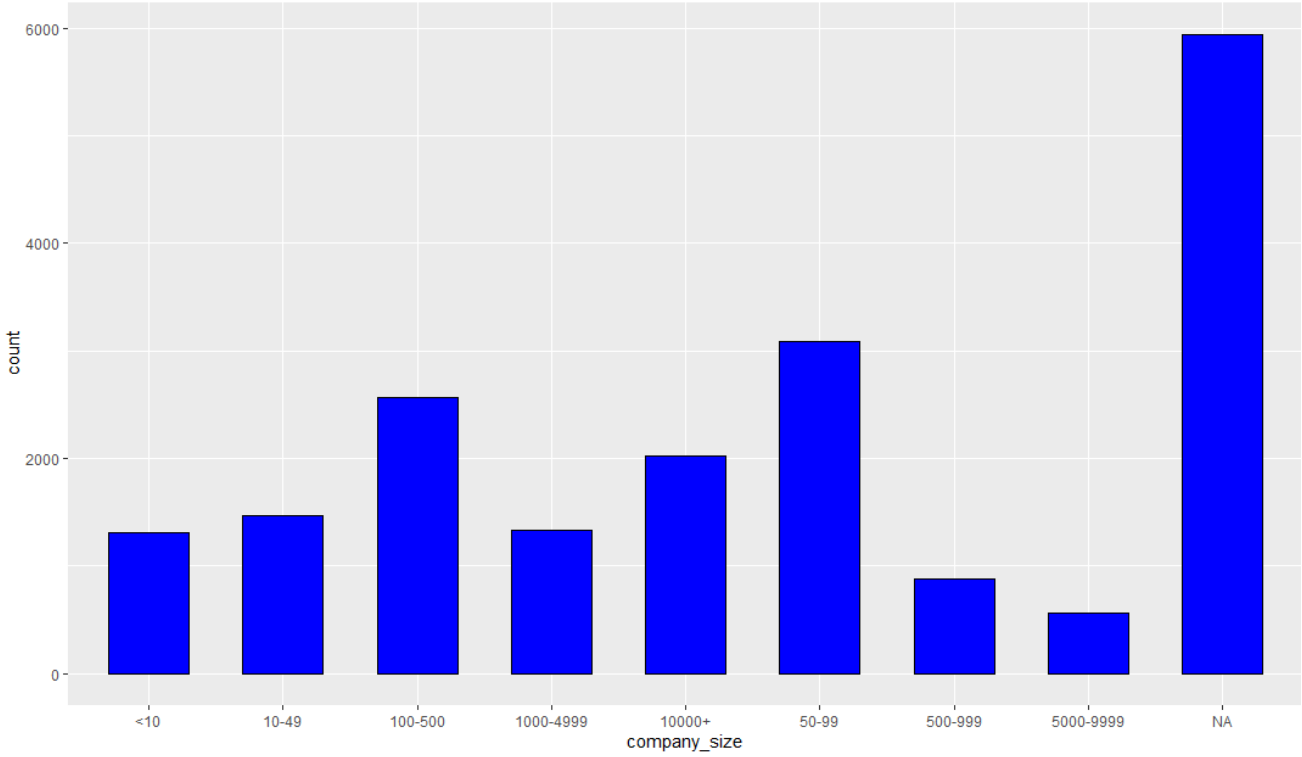


Figure 9 - company size

```
> ft
  company size Freq Rel.Freq
1        50-99 3083    23.32
2      100-500 2571    19.45
3       10000+ 2019    15.27
4        10-49 1471    11.13
5    1000-4999 1328    10.05
6          <10 1308     9.89
7      500-999  877     6.63
8    5000-9999  563     4.26
```

```
summary(df$company_size)
    <10     10-49   100-500 1000-4999    10000+
   1308      1471      2571      1328      2019
  50-99   500-999 5000-9999      NA's
   3083       877       563      5938
```

company type:

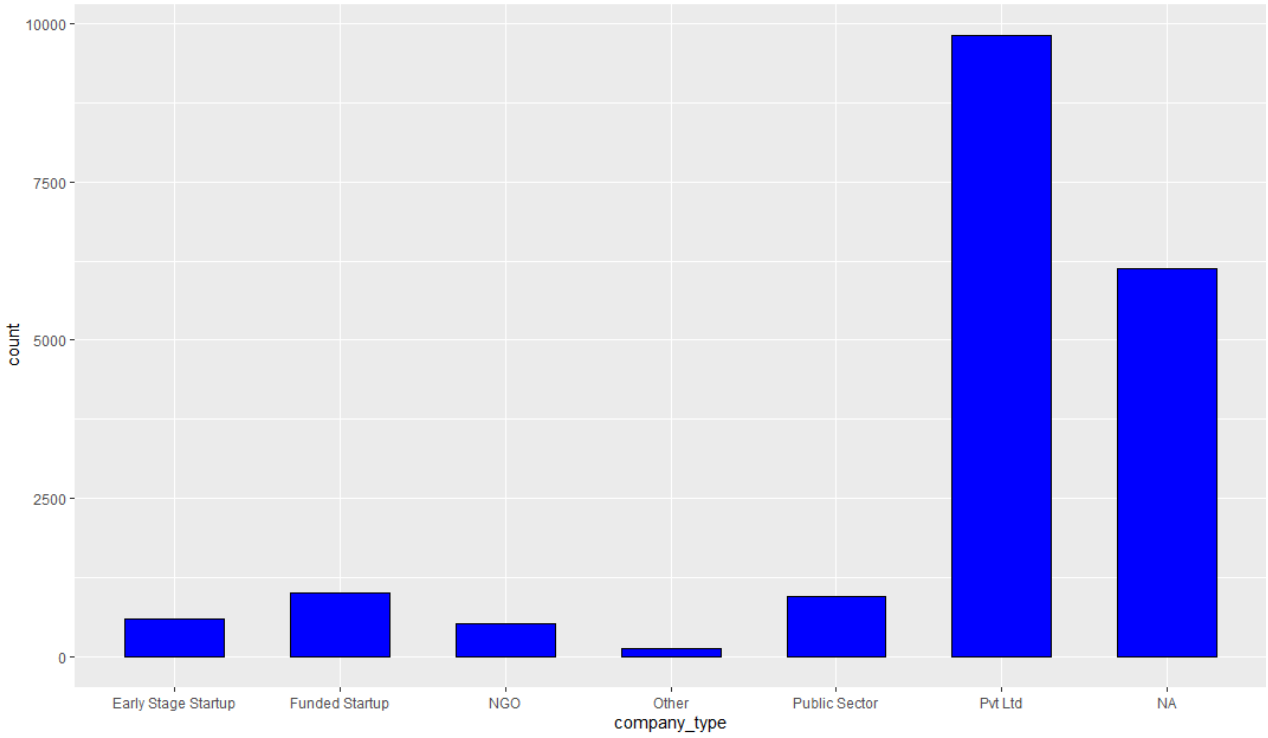


Figure 10 - company type

```
> ft
         company type Freq Rel.Freq
1             Pvt Ltd 9817    75.41
2      Funded Startup 1001     7.69
3       Public Sector  955     7.34
4 Early Stage Startup  603     4.63
5                 NGO  521     4.00
6               Other  121     0.93
```

```
> summary(df$company_type)
Early Stage Startup      Funded Startup                 NGO
                603                1001                 521
              Other       Public Sector             Pvt Ltd
                121                 955                9817
               NA's
               6140
```

last new job:

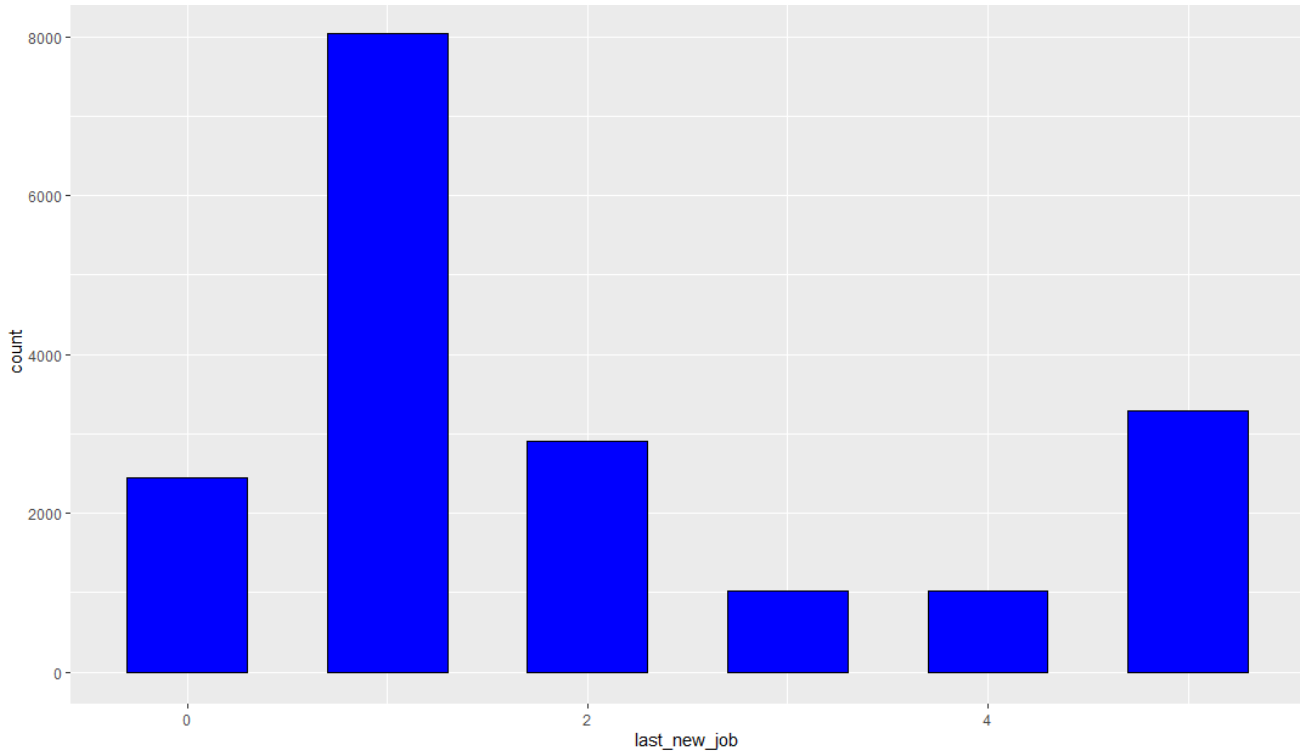


Figure 11 - last new job

```
> ft
  last new job Freq Rel.Freq
1            1 8040    42.91
2            5 3290    17.56
3            2 2900    15.48
4            0 2452    13.09
5            4 1029     5.49
6            3 1024     5.47
```

```
summary(df$last_new_job)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0       1       1       2       3       5     423
```

training hours:

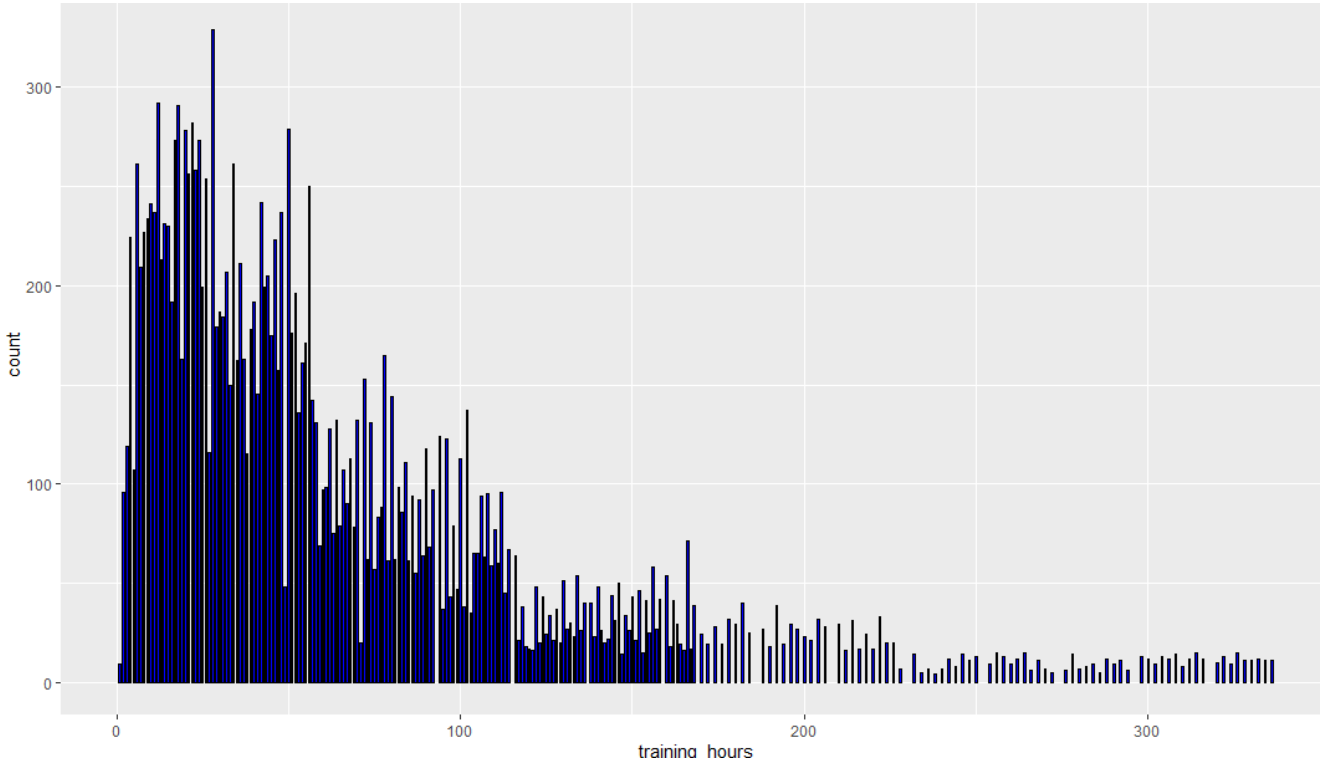
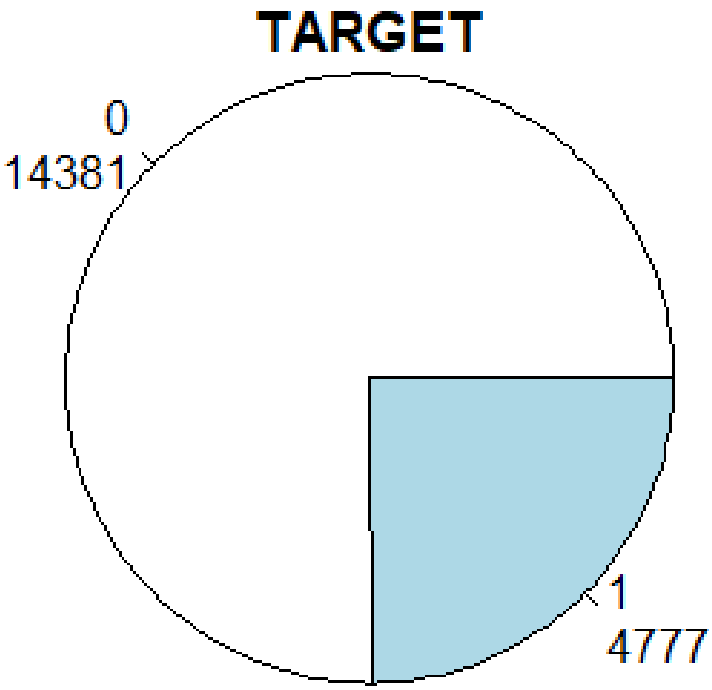


Figure 12 - training hours

```
summary(df$training_hours)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   23.00   47.00   65.37   88.00  336.00
```

Target:



Target13 Figure

```
> ft
  target  Freq Rel.Freq
1      0 14381    75.07
2      1  4777    24.93
```

Relationships between the features and target:

City & target - correlation:

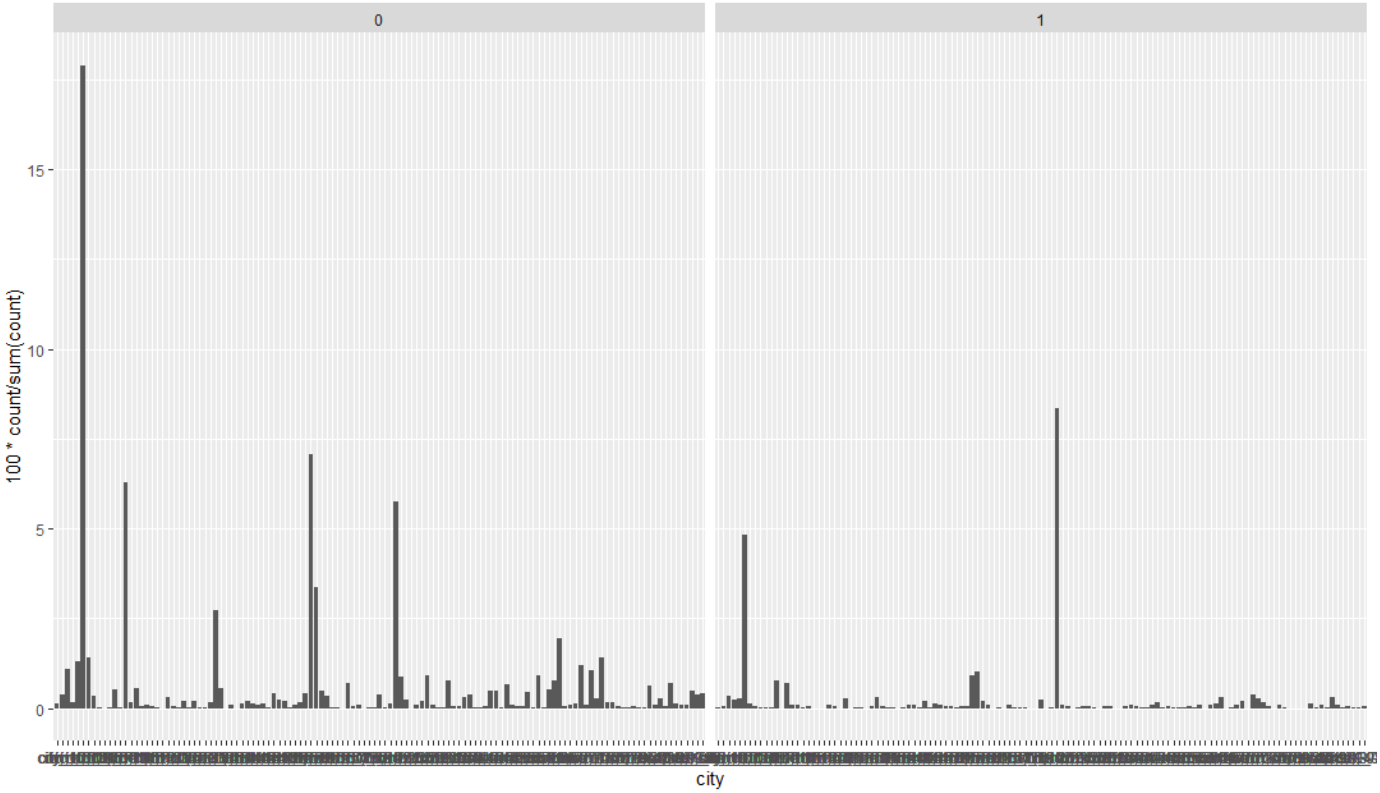


Figure 14 - City & target - correlation

```
> chisq.test(df$city, df$target)

        Pearson's Chi-squared test

data:  df$city and df$target
X-squared = 2998.8, df = 122, p-value < 2.2e-16
```

city development index & target - correlation:

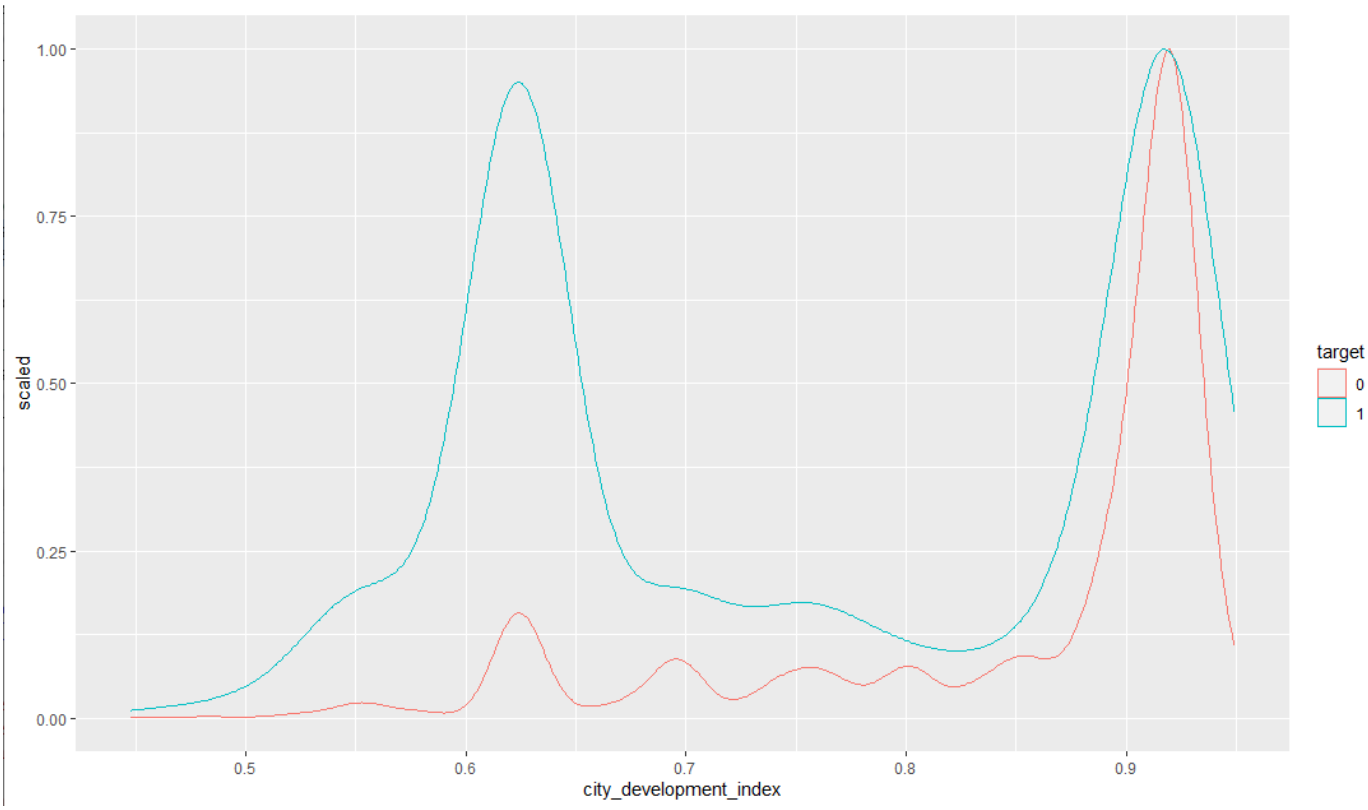


Figure 15 - city development index & target - correlation

```
> t.test(df$city_development_index ~ df$target, mu = 0, alternative = "t
wo.sided", var.equal = T)

        Two Sample t-test

data:  df$city_development_index by df$target
t = 50.316, df = 19156, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0936249 0.1012150
sample estimates:
mean in group 0 mean in group 1
      0.8531394       0.7557195
```

Gender & target – correlation:

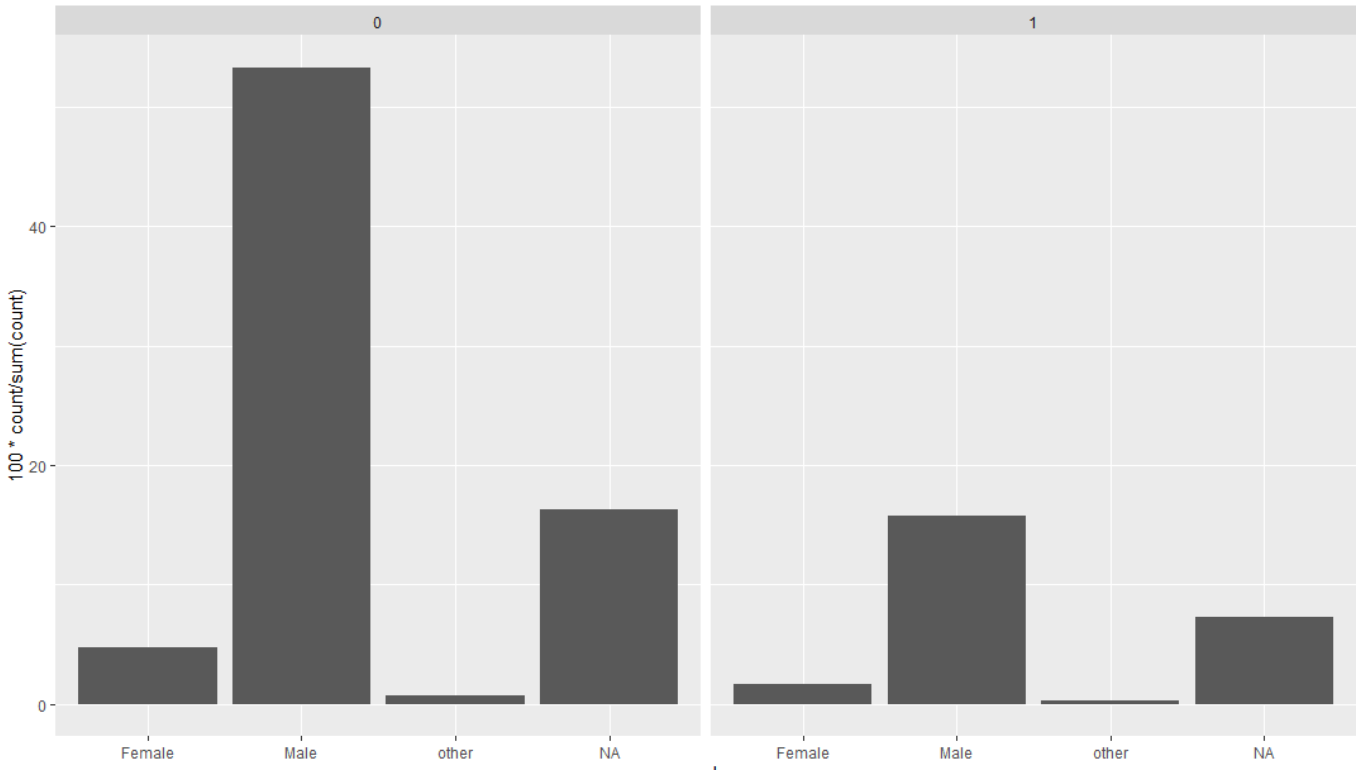


Figure 16 - Gender & target – correlation

```
             | df$target
   df$gender |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
      Female |       912 |       326 |      1238 |
             |     1.656 |     5.504 |           |
             |     0.737 |     0.263 |     0.085 |
             |     0.081 |     0.096 |           |
             |     0.062 |     0.022 |           |
-------------|-----------|-----------|-----------|
        Male |     10209 |      3012 |     13221 |
             |     0.204 |     0.678 |           |
             |     0.772 |     0.228 |     0.902 |
             |     0.906 |     0.889 |           |
             |     0.697 |     0.206 |           |
-------------|-----------|-----------|-----------|
       other |       141 |        50 |       191 |
             |     0.231 |     0.769 |           |
             |     0.738 |     0.262 |     0.013 |
             |     0.013 |     0.015 |           |
             |     0.010 |     0.003 |           |
-------------|-----------|-----------|-----------|
Column Total |     11262 |      3388 |     14650 |
             |     0.769 |     0.231 |           |
-------------|-----------|-----------|-----------|
```

```
> chisq.test(df$gender, df$target)

        Pearson's Chi-squared test

data:  df$gender and df$target
X-squared = 9.0422, df = 2, p-value = 0.01088
```

relevent experience & target – correlation:

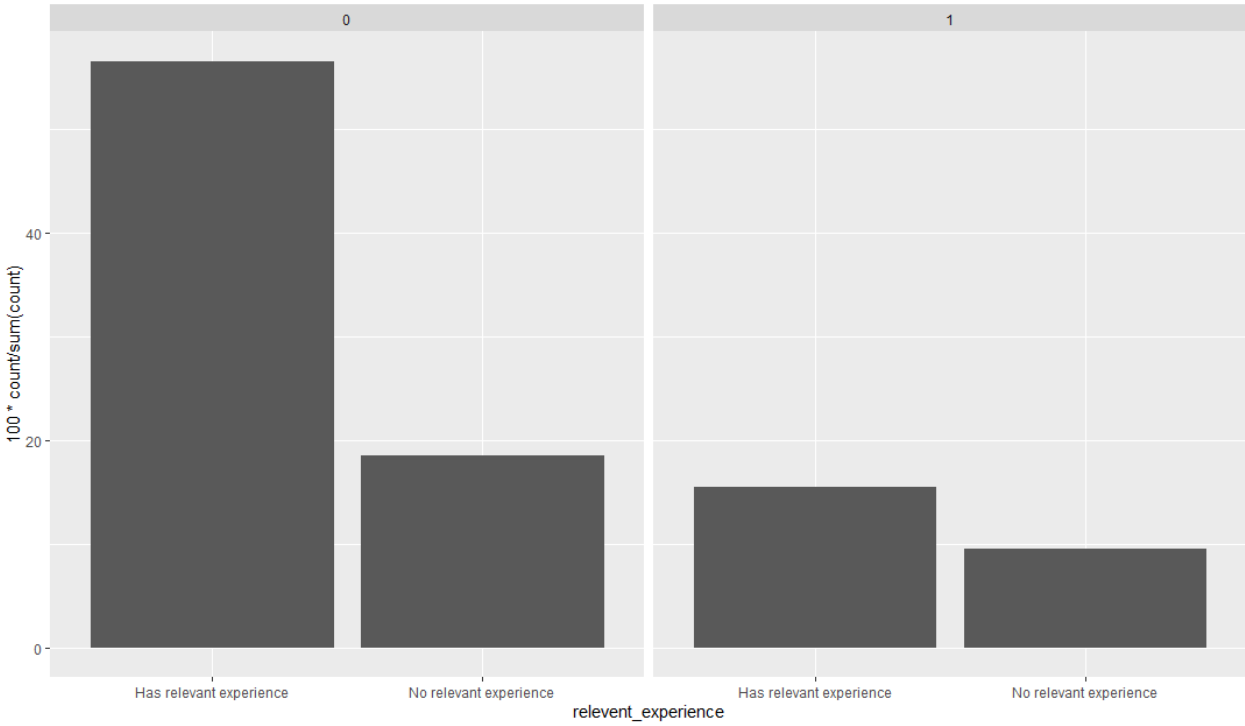


Figure 17 - relevant experience & target – correlation

```
                        | df$target
 df$relevent_experience |         0 |         1 | Row Total |
------------------------|-----------|-----------|-----------|
Has relevant experience |     10831 |      2961 |     13792 |
                        |    22.069 |    66.439 |           |
                        |     0.785 |     0.215 |     0.720 |
                        |     0.753 |     0.620 |           |
                        |     0.565 |     0.155 |           |
------------------------|-----------|-----------|-----------|
 No relevant experience |      3550 |      1816 |      5366 |
                        |    56.724 |   170.766 |           |
                        |     0.662 |     0.338 |     0.280 |
                        |     0.247 |     0.380 |           |
                        |     0.185 |     0.095 |           |
------------------------|-----------|-----------|-----------|
           Column Total |     14381 |      4777 |     19158 |
                        |     0.751 |     0.249 |           |
------------------------|-----------|-----------|-----------|
```

```
> chisq.test(df$relevent_experience, df$target)

        Pearson's Chi-squared test with Yates'
        continuity correction

data:  df$relevent_experience and df$target
X-squared = 315.34, df = 1, p-value < 2.2e-16
```

enrolled university & target – correlation:

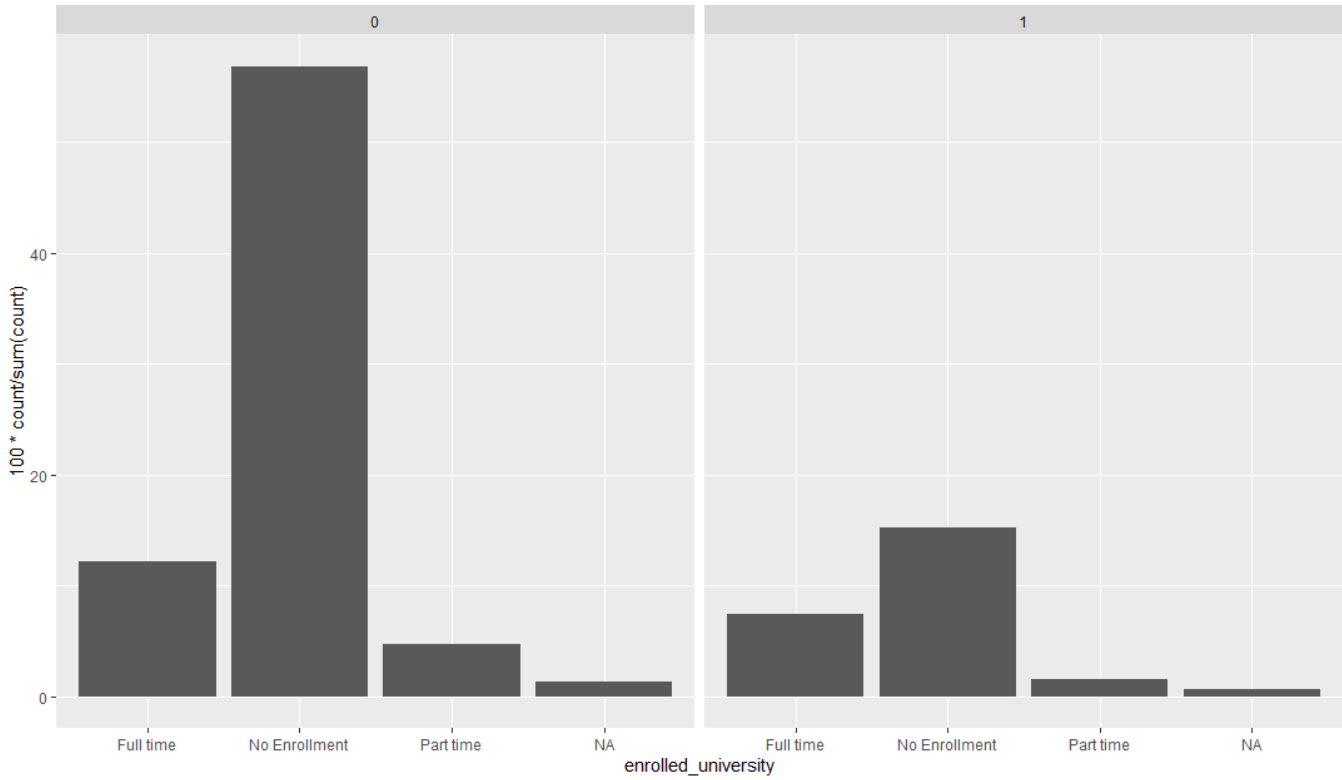


Figure 18 - enrolled university & target – correlation

```
                        | df$target
df$enrolled_university |         0 |         1 | Row Total |
-----------------------|-----------|-----------|-----------|
             Full time |      2326 |      1431 |      3757 |
                       |    88.321 |   267.923 |           |
                       |     0.619 |     0.381 |     0.200 |
                       |     0.165 |     0.307 |           |
                       |     0.124 |     0.076 |           |
-----------------------|-----------|-----------|-----------|
         No Enrollment |     10896 |      2921 |     13817 |
                       |    24.498 |    74.314 |           |
                       |     0.789 |     0.211 |     0.736 |
                       |     0.772 |     0.628 |           |
                       |     0.580 |     0.156 |           |
-----------------------|-----------|-----------|-----------|
             Part time |       896 |       302 |      1198 |
                       |     0.028 |     0.084 |           |
                       |     0.748 |     0.252 |     0.064 |
                       |     0.063 |     0.065 |           |
                       |     0.048 |     0.016 |           |
-----------------------|-----------|-----------|-----------|
          Column Total |     14118 |      4654 |     18772 |
                       |     0.752 |     0.248 |           |
-----------------------|-----------|-----------|-----------|
```

```
> chisq.test(df$enrolled_university, df$target)

        Pearson's Chi-squared test

data:  df$enrolled_university and df$target
X-squared = 455.17, df = 2, p-value < 2.2e-16
```

education level & target – correlation:

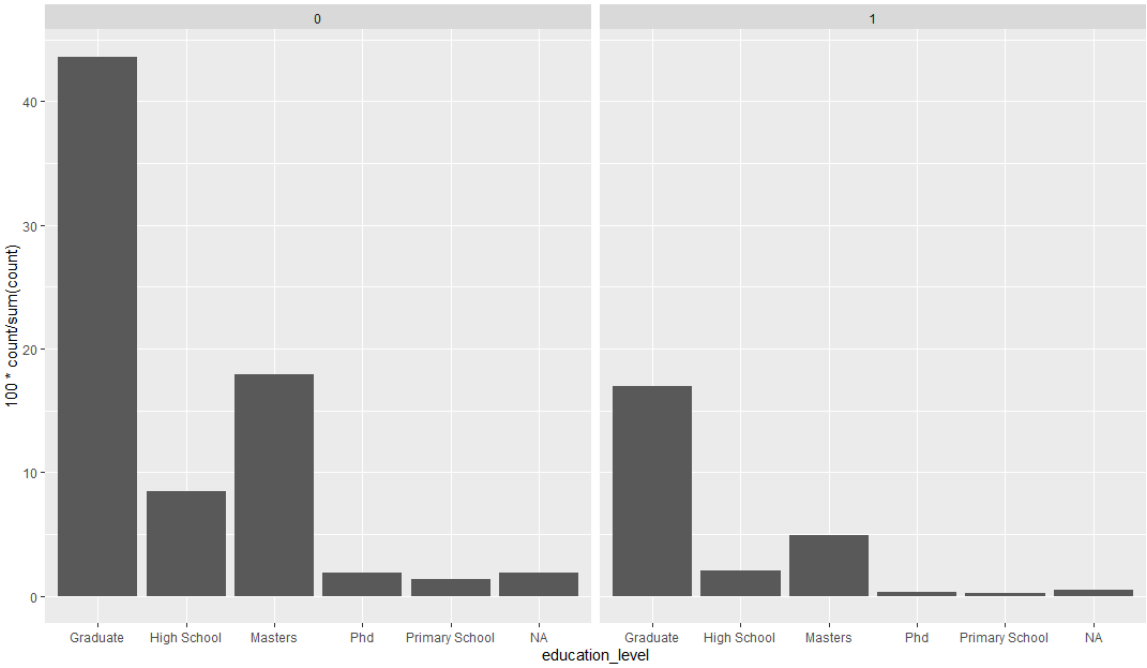


Figure 19 - education level & target – correlation

df$education_level	df$target 0	1	Row Total
Graduate	8353	3245	11598
	13.796	41.405	
	0.720	0.280	0.620
	0.596	0.694	
	0.447	0.174	
High School	1623	394	2017
	8.011	24.042	
	0.805	0.195	0.108
	0.116	0.084	
	0.087	0.021	
Masters	3426	935	4361
	7.335	22.015	
	0.786	0.214	0.233
	0.244	0.200	
	0.183	0.050	
Phd	356	58	414
	6.657	19.980	
	0.860	0.140	0.022
	0.025	0.012	
	0.019	0.003	
Primary School	267	41	308
	5.602	16.813	
	0.867	0.133	0.016
	0.019	0.009	
	0.014	0.002	
Column Total	14025	4673	18698
	0.750	0.250	

```
> chisq.test(df$education_level, df$target)

        Pearson's Chi-squared test

data:  df$education_level and df$target
X-squared = 165.66, df = 4, p-value < 2.2e-16
```

major discipline & target – correlation:

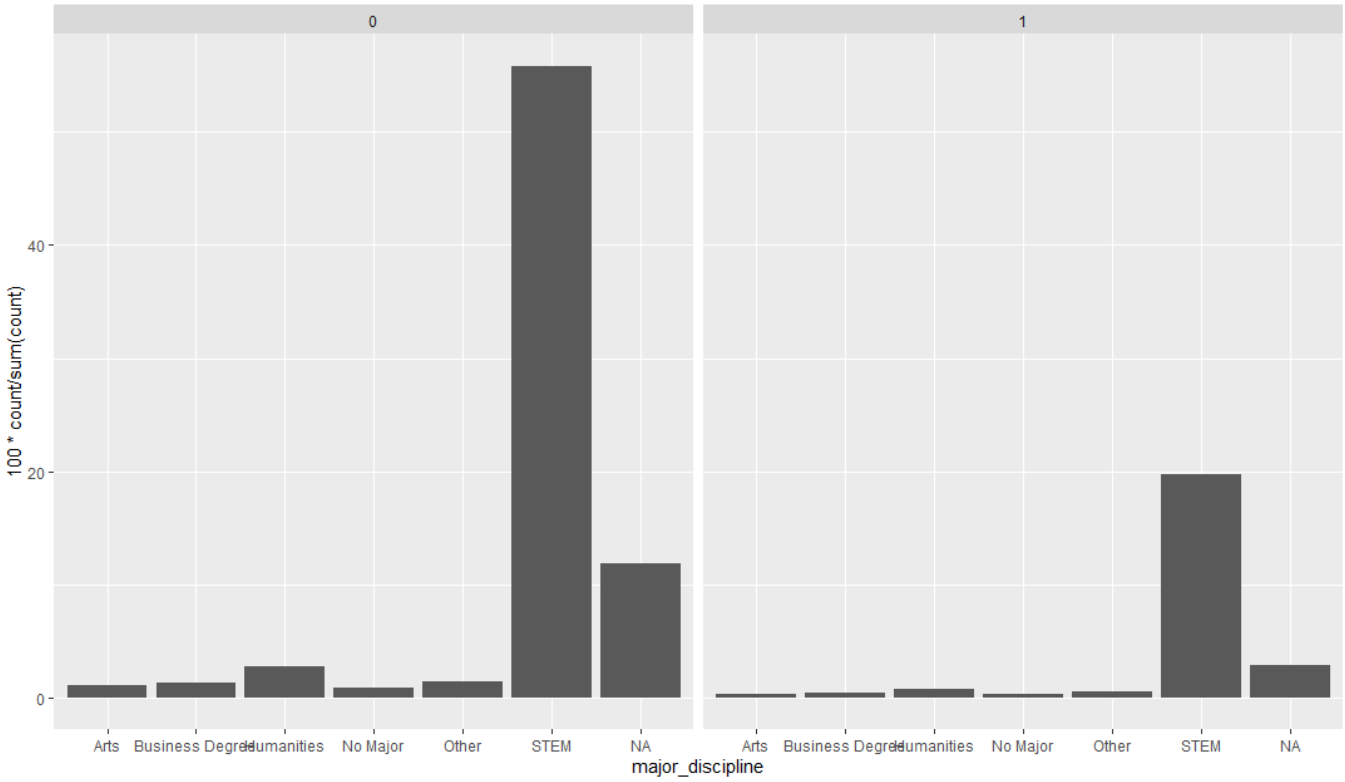


Figure 20 - major discipline & target – correlation

```
> chisq.test(df$major_discipline, df$target)

        Pearson's Chi-squared test

data:  df$major_discipline and df$target
X-squared = 12.207, df = 5, p-value = 0.03206
```

experience & target – correlation:

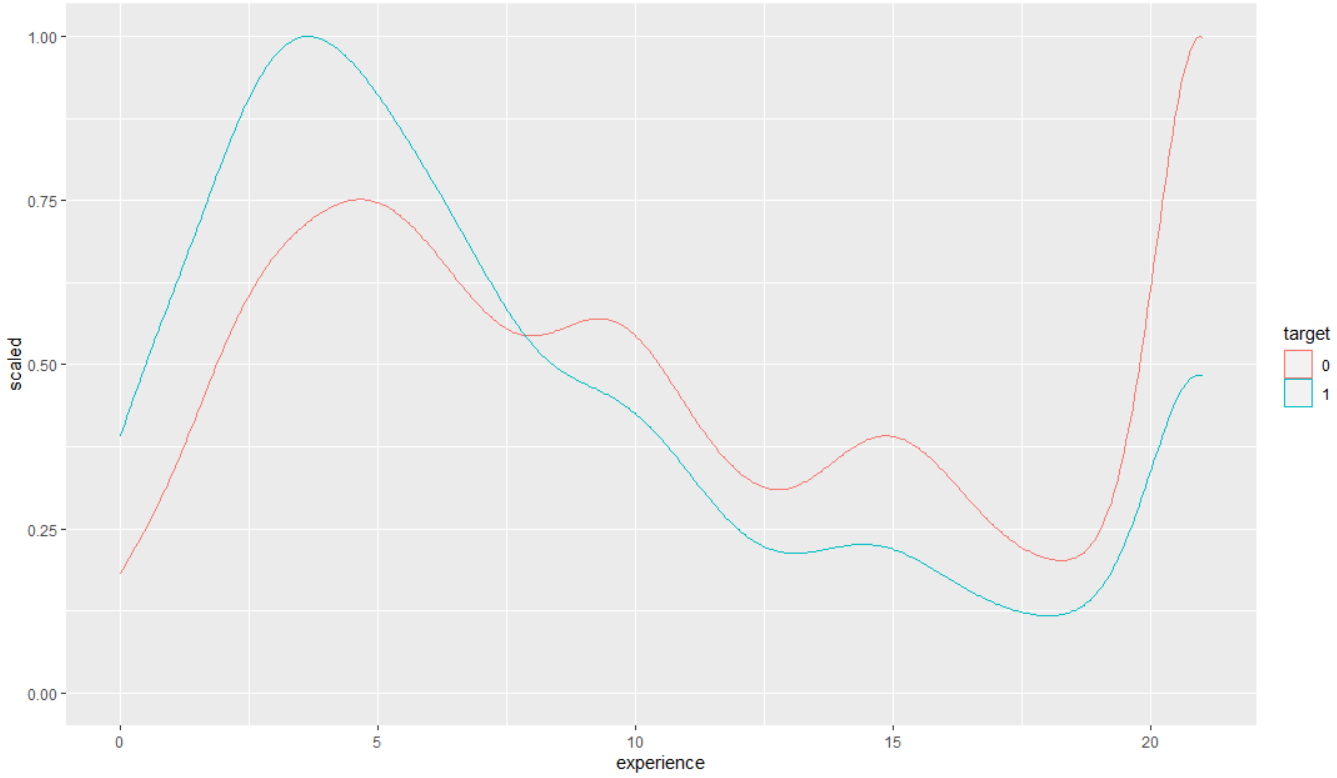


Figure 21 - experience & target – correlation

```
> t.test(df$experience ~ df$target, mu = 0, alternative
 = "two.sided", var.equal = T)

        Two Sample t-test

data:  df$experience by df$target
t = 24.808, df = 19091, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
 equal to 0
95 percent confidence interval:
 2.550722 2.988359
sample estimates:
mean in group 0 mean in group 1
      10.789734        8.020194
```

company size & target – correlation:

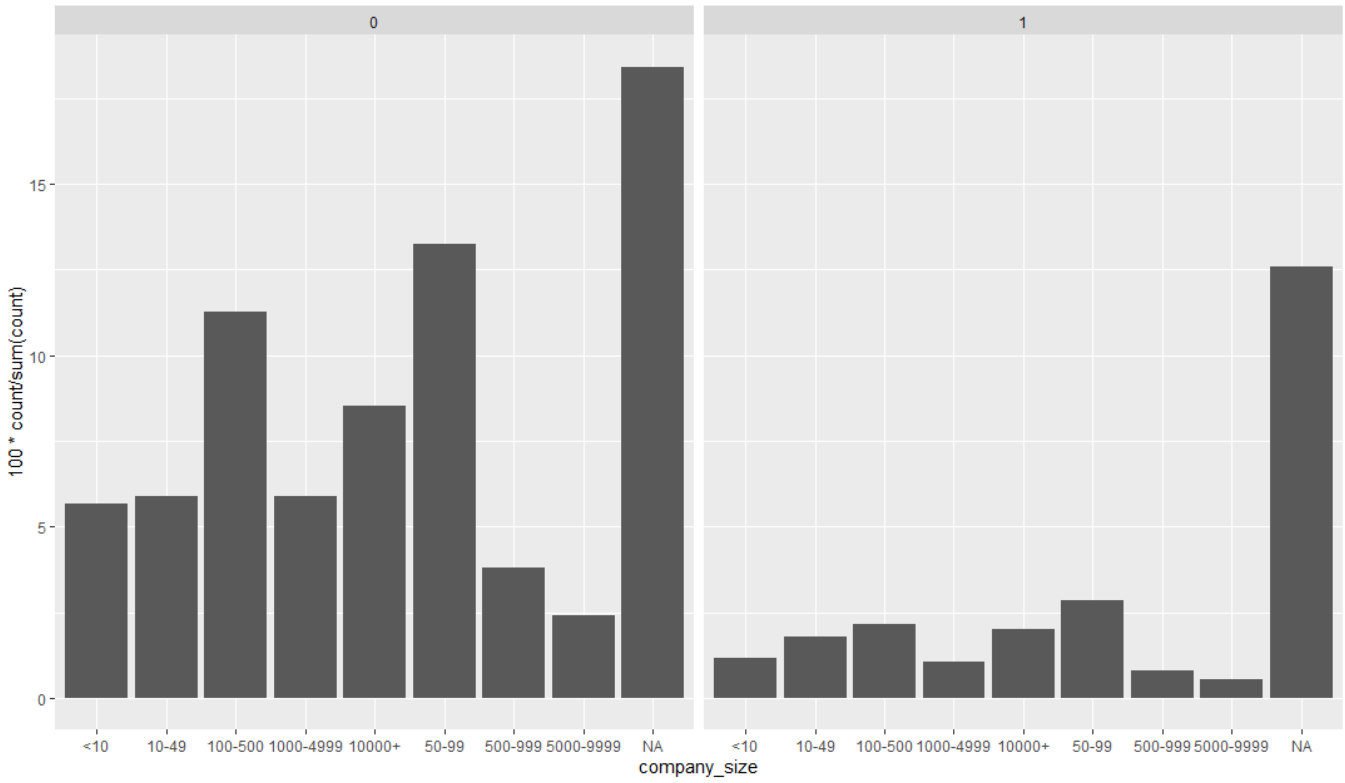


Figure 22 - company size & target – correlation

```
> chisq.test(df$company_size, df$target)

        Pearson's Chi-squared test

data:  df$company_size and df$target
X-squared = 45.532, df = 7, p-value = 1.078e-07
```

company type & target – correlation:

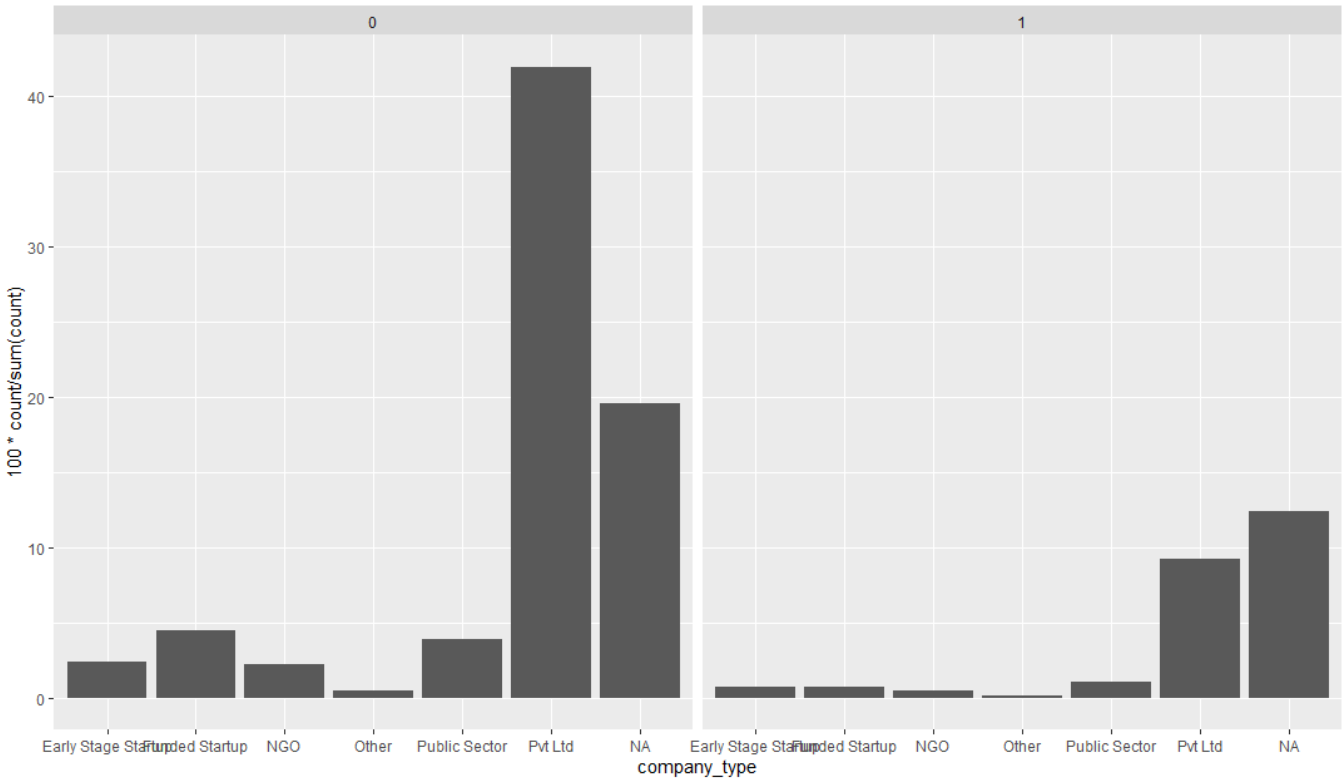


Figure 23 - company type & target – correlation

```
> chisq.test(df$company_type, df$target)

        Pearson's Chi-squared test

data:  df$company_type and df$target
X-squared = 35.035, df = 5, p-value = 1.48e-06
```

last new job & target – correlation:

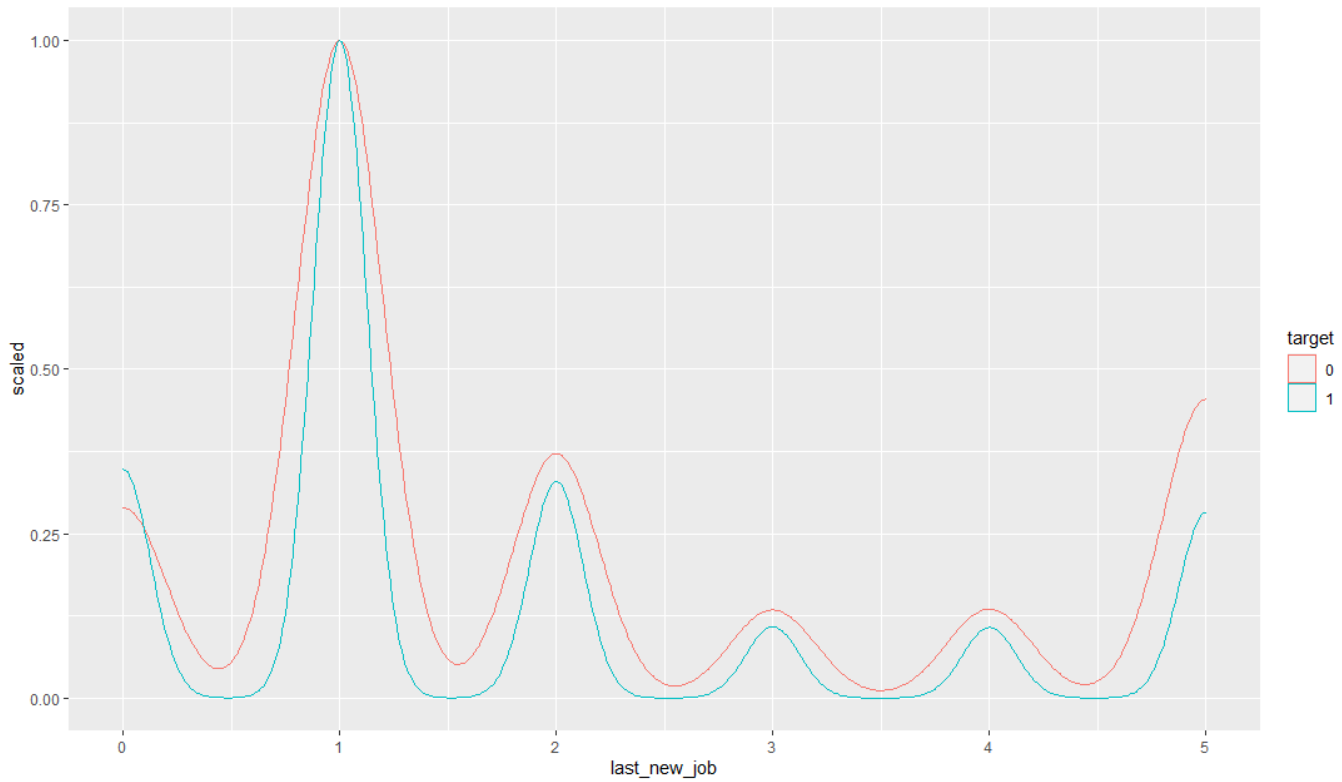


Figure 24 - last new job & target – correlation

```
> t.test(df$last_new_job ~ df$target, mu = 0, altern
ative = "two.sided", var.equal = T)

        Two Sample t-test

data:  df$last_new_job by df$target
t = 11.345, df = 18733, p-value < 2.2e-16
alternative hypothesis: true difference in means is
 not equal to 0
95 percent confidence interval:
 0.2655824 0.3765180
sample estimates:
mean in group 0 mean in group 1
       2.079649        1.758598
```

training hours & target – correlation:

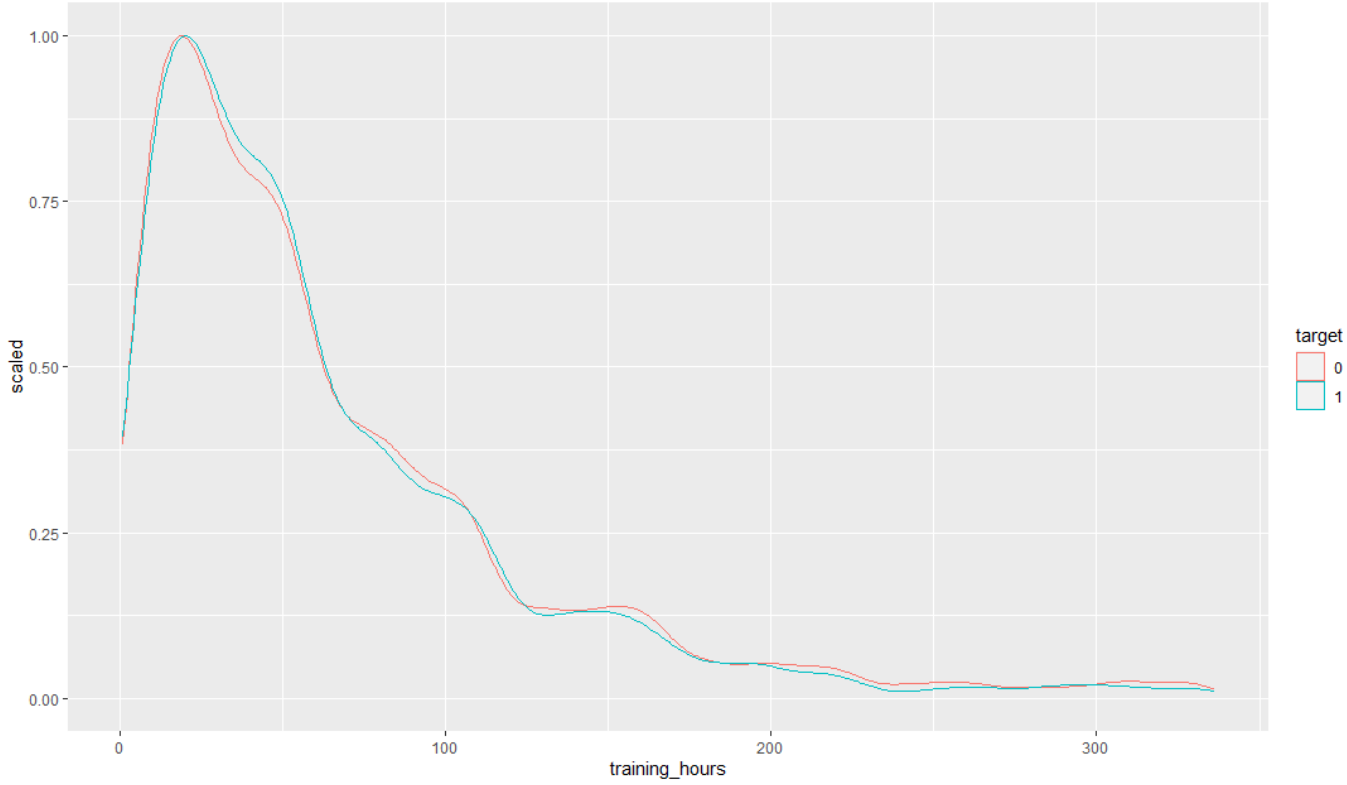


Figure 25 - training hours & target – correlation

```
> t.test(df$training_hours ~ df$target, mu = 0, alte
rnative = "two.sided", var.equal = T)

        Two Sample t-test

data:  df$training_hours by df$target
t = 2.9871, df = 19156, p-value = 0.00282
alternative hypothesis: true difference in means is
 not equal to 0
95 percent confidence interval:
 1.029823 4.960731
sample estimates:
mean in group 0 mean in group 1
       66.11376        63.11848
```

3. שלב המידול

הכנה לפני תחילת המידול:

1. כתיבת פונקציית חיזוי.
2. הורדת עמודות מרובות בערכים חסרים (מגדר, גודל חברה, סוג חברה, תחום עיקרי).
3. הורדת שורות עם ערכים חסרים (NA.Omit).
4. חלוקה לקבוצת אימון וקבוצת ביקורת.

המודלים:

Logistic Regression Model:

```
> summary(logistic)

Call:
glm(formula = target ~ . - city, family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8376  -0.6823  -0.5335  -0.3303   2.5823

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              3.6724114  0.1359105  27.021  < 2e-16 ***
city_development_index                  -5.3345141  0.1684496 -31.668  < 2e-16 ***
relevent_experienceNo relevant experience 0.6210007  0.0517203  12.007  < 2e-16 ***
enrolled_universityNo Enrollment        -0.3619029  0.0554612  -6.525 6.78e-11 ***
enrolled_universityPart time            -0.3875530  0.0921555  -4.205 2.61e-05 ***
education_levelHigh School              -0.9106392  0.0792596 -11.489  < 2e-16 ***
education_levelMasters                  -0.2010428  0.0517175  -3.887 0.000101 ***
education_levelPhd                      -0.5505973  0.1760720  -3.127 0.001765 **
education_levelPrimary School           -1.4125714  0.2319653  -6.090 1.13e-09 ***
experience                              -0.0179727  0.0040527  -4.435 9.22e-06 ***
last_new_job                             0.0221661  0.0148533   1.492 0.135610
training_hours                          -0.0011257  0.0003576  -3.148 0.001643 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16062  on 14410  degrees of freedom
Residual deviance: 14135  on 14399  degrees of freedom
AIC: 14159

Number of Fisher Scoring iterations: 4
```

Table 4 - Logistic Regression

Decision Tree Model:

```
> summary(tree)

Classification tree:
tree(formula = target ~ . - city, data = training_set)
Variables actually used in tree construction:
[1] "city_development_index" "relevent_experience"
Number of terminal nodes:  3
Residual mean deviance:  0.9838 = 14170 / 14410
Misclassification error rate: 0.2166 = 3121 / 14411
```

Table 5 - Decision Tree Model

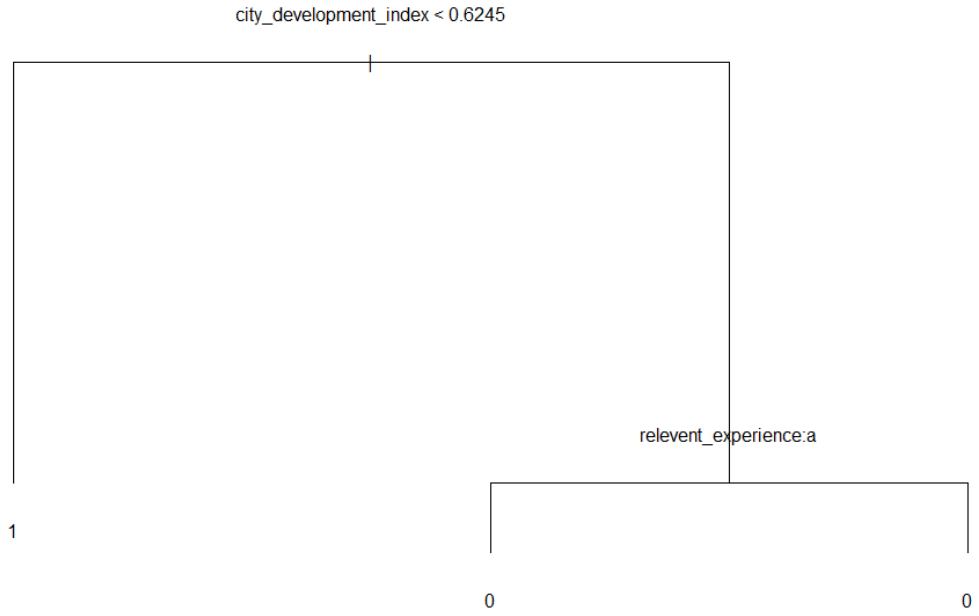


Figure26 - Decision Tree Model

Random Forest Model:

```
> summary(Random_forest)
                Length Class  Mode
call                 6 -none- call
type                 1 -none- character
predicted        14411 factor numeric
err.rate           900 -none- numeric
confusion            6 -none- numeric
votes            28822 matrix numeric
oob.times        14411 -none- numeric
classes              2 -none- character
importance           7 -none- numeric
importanceSD         0 -none- NULL
localImportance      0 -none- NULL
proximity            0 -none- NULL
ntree                1 -none- numeric
mtry                 1 -none- numeric
forest              14 -none- list
y                14411 factor numeric
test                 0 -none- NULL
inbag                0 -none- NULL
terms                3 terms  call
```

```
> importance(Random_forest)
                      MeanDecreaseGini
city_development_index        437.50883
relevent_experience            57.77663
enrolled_university            66.15813
education_level                59.18746
experience                    119.47218
last_new_job                   44.13822
training_hours                 99.40828
```

Table 6 - Random Forest Model

Table 7 - Random Forest Model – importance

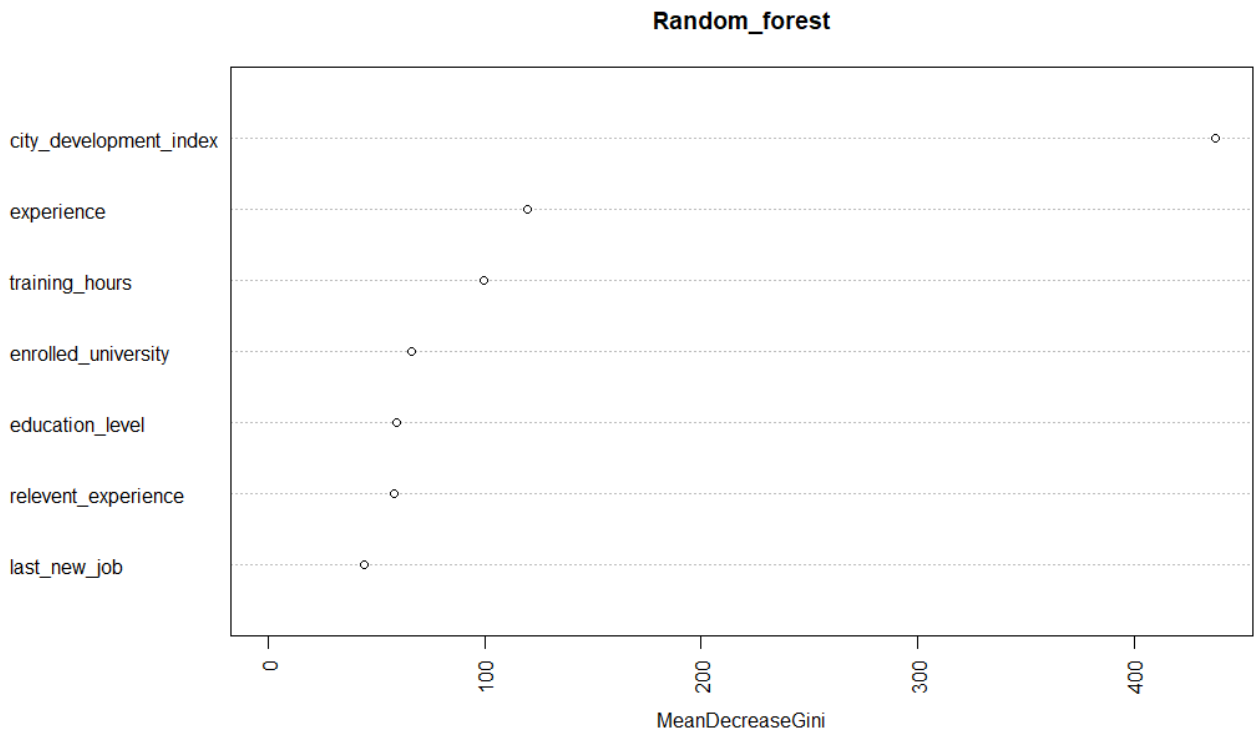


Figure 27 - Random Forest Model

KNN Model:

1. הכנת קובץ נתונים ללא עמודות קטגוריאליות.
2. ביצוע נרמול על כל העמודות (המספריות).
3. ביצוע Cross Validation – לבחירת ערך ה K המתאים.
4. ביצוע המודל עם ערך ה K הרצוי

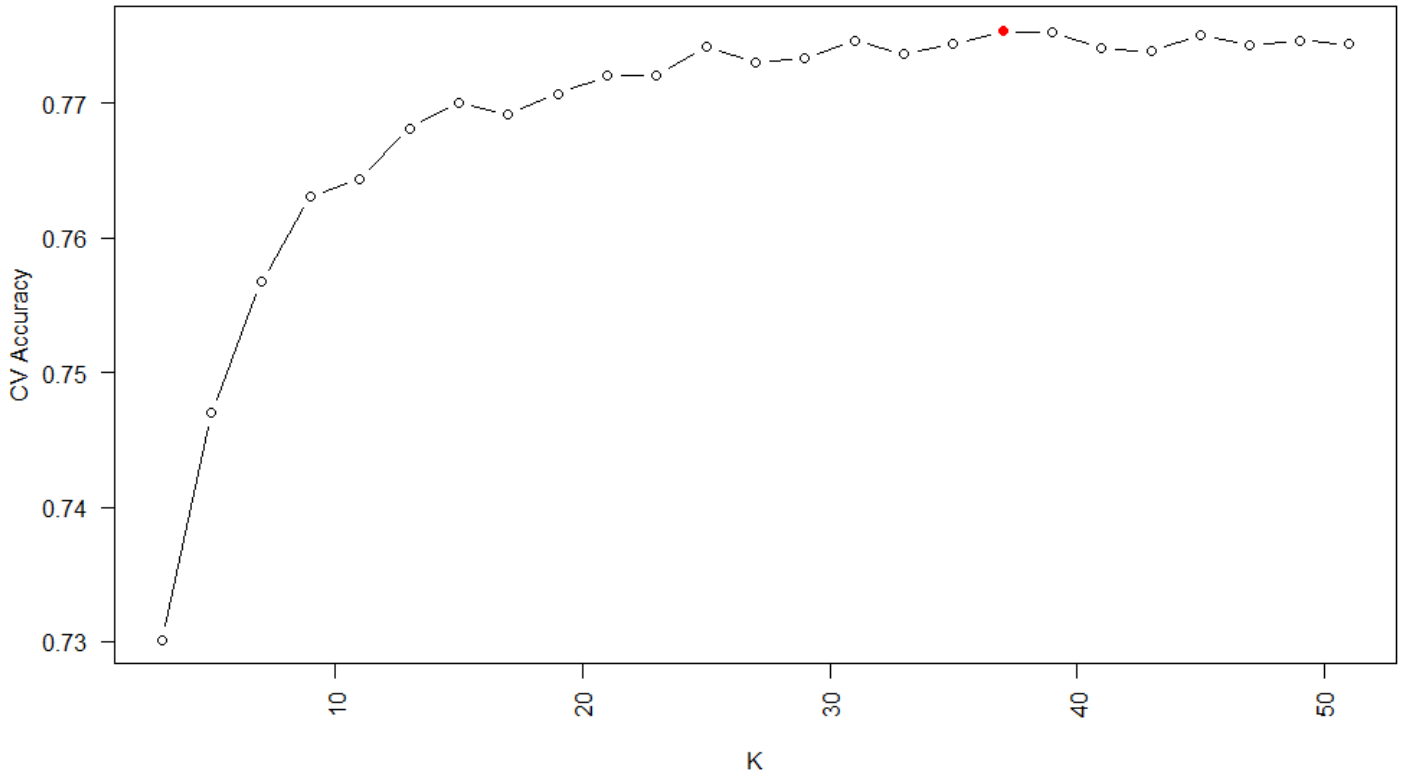


Figure 28 - Cross Validation KNN

ה K הטוב ביותר = K=37

4. טיב המידול:

Logistic Regression Model:

```
> accuracy_LR = confusion_Matrixs(confusion_logistic)

predictions    0    1
          0 2573  690
          1  146  194
[1] "accuracy:" "0.768"
[1] "sensitivity:" "0.2195"
[1] "specificity:" "0.9463"
[1] "F1Score:" "0.317"
[1] "PPV:"    "0.5706"
[1] "NPV:"    "0.7885"
```

Figure 29 - Logistic Regression - Prediction

Decision Tree Model:

```
> accuracy_Tree = confusion_Matrixs(confusion_tree)

test.predictions    0    1
               0 2459  498
               1  260  386
[1] "accuracy:" "0.7896"
[1] "sensitivity:" "0.4367"
[1] "specificity:" "0.9044"
[1] "F1Score:" "0.5046"
[1] "PPV:"    "0.5975"
[1] "NPV:"    "0.8316"
```

Figure 30 - Decision Tree - Prediction

Random Forest Model:

```
> accuracy_RF = confusion_Matrixs(confusion_RF)
   test.predictions
       0    1
  0 2604  115
  1  713  171
[1] "accuracy:" "0.7702"
[1] "sensitivity:" "0.5979"
[1] "specificity:" "0.785"
[1] "F1Score:" "0.2923"
[1] "PPV:"    "0.1934"
[1] "NPV:"    "0.9577"
```

Figure 31 - Random Forest - Prediction

KNN Model:

```
> accuracy_KNN = confusion_Matrixs(confusion_KNN)
                 test.y
test.predictions    0    1
               0 2485  572
               1  234  312
[1] "accuracy:" "0.7763"
[1] "sensitivity:" "0.3529"
[1] "specificity:" "0.9139"
[1] "F1Score:" "0.4364"
[1] "PPV:"    "0.5714"
[1] "NPV:"    "0.8129"
```

Figure 32 - KNN - Prediction

5. מסקנות

- **שלב הכנת הנתונים:**
  - סוגי הפיצ'רים: צורת ההתייחסות לפיצ'רים פעם כמספר ופעם כקטגוריה משפיעה על החיזוי.
  - ערכים חסרים: ריבוי של ערכים חסרים משפיע מאוד על החיזוי. טיפול נכון בערכים אלה יכול לשפר מאוד את טיב החיזוי.
  - עמודות בעיתיות: חשוב לבחור בקפידה איזה עמודות חשובות, בין איזה עמודות יש קורלציה גבוהה ולטפל בהן, ולהתאים את העמודות למודל הנבחר.
  - קיים קושי בבדיקת הקורלציה בין משתנים שאינם מאותו סוג (כמותי מול קטגוריאלי).

- **שלב המידול:**
  - רגרסיה לוגיסטית
    - הפיצ'ר city development index מאוד משפיע על משתנה המטרה.
  - עץ החלטה
    - מודל זה מתעלם מרוב הפיצ'רים. הוא בוחר רק שני פיצ'רים.
    - הפיצ'ר העיקרי שמכריע הוא city development index.
  - יער אקראי
    - לוקח את כל הפיצ'רים אך שוב ניתן לראות לפי מדד GINI שהפיצ'ר city development index הוא החזק ביותר.
    - חשוב לשים לב למספר העצים שממנו מורכב המודל.
  - KNN
    - צריך השתמש רק בפיצ'רים הכמותיים.
    - חשוב לא לשכוח לנרמל את הנתונים.
    - חשוב להריץ מספר מודלים על מנת לבחור את ה-K המתאים ביותר.

- **כלליות:**
  - **משתנים חשובים:**
    - הפיצ'ר city development index מאוד משפיע על משתנה המטרה בכל המודלים.
    - הקורלציה בין city development index ו- city גבוהה ולכן החלטנו לוותר על city.

  - **המודל הטוב ביותר:**
    - למרות שה accuracy של העץ הוא הטוב ביותר, זהו המודל הגרוע ביותר לחיזוי. מודל זה מתבסס למעשה על פיצ'ר אחד בלבד.
    - יער אקראי הוא מודל לא טוב, מכיוון שניתן לראות שהוא לא חוזה בצורה טובה את ערך הניבוי חיובי. (PPV)
    - הרגרסיה הלוגיסטית חוזה בצורה טובה יחסית את שני ערכי הניבוי, אך עם זאת יכולת החיזוי הכללית שלה לא טובה.
    - KNN הוא המודל היציב ביותר לדעתנו, אך יש לציין כי במודל זה אנו חוזים על פי הפיצ'רים הכמותיים בלבד, ומתעלמים מהאחרים.

  - **שיפור תוצאות החיזוי**:
    - טיפול בערכים החסרים בצורה טובה יותר, השלמת הערכים החסרים בצורה חכמה בעזרת Imputation.
    - החלפת כל הפיצ'רים הקטגוריאליים לכמותיים או להיפך (Encode / Scale).
    - שינוי העמודות (הוספה / הורדת מימד).
    - בחירת העמודות לחיזוי בצורה טובה יותר.
    -

- **סיכום:**
  - לפי התוצאות שקיבלנו נראה כי קשה מאוד לקבוע איזה מודל טוב יותר.
  - כחלק מתהליך העבודה עשינו הרבה 'ניסוי ותהייה' בנושא התאמת הפיצ'רים, ניקוי הנתונים וניסיונות בבחירת המודלים שלדעתנו יתאימו.
  - טיפול בערכים החסרים היה שלב הכרחי שהשפיע המון על המודלים.
  - לבסוף קיבלנו מודלים עם מעל 0.7 אחוז חיזוי שזה נחשב טוב.