UNIVERSITATEA „ALEXANDRU IOAN CUZA” DIN IAȘI FACULTATEA DE ECONOMIE ȘI ADMINISTRAREA AFACERILOR
SPECIALIZAREA: DATA MINING
DISCIPLINA: INTRODUCERE ÎN R

Analiza statistică a factorilor de influenta asupra atitudinii față de imigranți în Norvegia, anul 2022

COORDONATOR ȘTIINȚIFIC:
PROF. DR. VIORICĂ DANIELA

STUDENT:
COZMA ALEXANDRU-CRISTIAN

IAȘI
2024

1. Introducere

În contextul globalizării și al migrației internaționale, atitudinea populației față de imigranți reprezintă un subiect de mare actualitate, aceasta influențând pe de o parte, coeziunea socială, politicile publice și dinamica socio-economică a unei țări. Pe de altă parte, gradul de toleranța față de imigranți poate determina modul în care cetățenii unui stat se raportează la decizia de aderare la Uniunea Europeană, întrucât această afiliere va încuraja libera circulație a populației. Norvegia, cunoscută pentru politicile sale de incluziune socială și nivelul ridicat de trai, a devenit în ultimele decenii un punct de atracție pentru imigranți, fenomen ce a condus la diverse reacții în rândul populației locale, influențate de o varietate de factori, precum educația, veniturile, mediul de rezidență sau vârsta. Înțelegerea acestor factori constituie o necesitate pentru a fundamenta politici și intervenții eficiente.

Obiectivul principal al acestui proiect este de a analiza factorii care influențează atitudinea față de imigranți și testarea existenței unei asocieri între toleranța față de imigranți și atitudinea față de aderarea la UE, în Norvegia, în anul 2022, utilizând diverse instrumente statistice. Printr-o abordare descriptivă și inferențială, se urmărește identificarea relațiilor dintre variabilele socio-demografice și opiniile exprimate de populație. Studiul va oferi o perspectivă detaliată asupra dinamicii atitudinilor și va contribui la înțelegerea mai profundă a contextului social norvegian.

Studiul de față are o structură complexă, alcătuită din șapte capitole. După o scurtă introducere, voi prezentata baza de date care reprezintă suportul studiului și operațiunile preliminare și cele de transformare a variabilelor, mai apoi voi întocmi analiza descriptivă atât asupra variabilelor numerice cât și a celor nenumerice din baza de date rezultată după operațiunile preliminare. Al patrulea capitol marchează începutul analizei inferențiale și se axează pe analizarea variabilelor categoriale prin procedee de tabelare a datelor si realizarea analizelor de asociere și concordanță. Următorul capitol are ca obiectiv crearea modelelor de regresie menite să identifice factorii ce influențează atitudinea față de imigranți. În al șaselea capitol se vor analiza mediile anilor de educație a diferitelor categorii de persoane pentru a stabili importanța acestei variabile ca factor de influență atât pentru atitudinea față de imigranți cât și față de aderarea la UE.

2. Prezentarea bazei de date

Baza de date, denumită inițial, ”ESS10NO„ reprezintă un chestionar realizat de European Social Survey în anul 2022. Acest chestionar a vizat strict populația norvegiană și are ca scop construirea unei imagini de ansamblu asupra deschiderii cetățenilor norvegieni față de noțiunea de aderare la Uniunea Europeană.

 # Prezentarea bazei de date inițiale

>library(tidyverse)

>glimpse(ESS10NO)

```
Rows: 1,411
Columns: 25
$ rownames  <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 3…
$ cntry     <chr> "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", "NO", …
$ idno      <dbl> 50013, 50014, 50019, 50038, 50073, 50074, 50079, 50095, 50114, 50145, 50160, 50178, 50180, 50209, 50229, 502…
$ region    <chr> "Vestlandet", "Oslo og Viken", "Vestlandet", "Oslo og Viken", "Oslo og Viken", "Vestlandet", "Oslo og Viken"…
$ inwds     <dttm> 2021-09-20 17:47:07, 2021-09-10 10:04:36, 2021-10-11 10:04:38, 2021-11-09 10:14:59, 2021-07-08 15:11:23, 20…
$ inwde     <dttm> 2021-09-20 18:50:23, 2021-09-10 10:57:19, 2021-10-11 11:04:42, 2021-11-09 11:06:31, 2021-07-11 19:27:53, 20…
$ dweight   <dbl> 0.9965166, 1.0045637, 1.0006449, 0.9880579, 0.9880579, 0.9965166, 1.0065559, 0.9952779, 1.0065559, 0.9884048…
$ pspwght   <dbl> 0.3480659, 0.5267486, 0.3310604, 0.8391549, 0.3009387, 0.3480659, 0.3151087, 3.0190203, 0.4020469, 0.3725736…
$ pweight   <dbl> 0.3167809, 0.3167809, 0.3167809, 0.3167809, 0.3167809, 0.3167809, 0.3167809, 0.3167809, 0.3167809, 0.3167809…
$ anweight  <dbl> 0.11026063, 0.16686387, 0.10487361, 0.26582822, 0.09533162, 0.11026063, 0.09982039, 0.95636785, 0.12736078, …
$ prob      <dbl> 0.0008713102, 0.0008643306, 0.0008677155, 0.0008787695, 0.0008787695, 0.0008713102, 0.0008626199, 0.00087239…
$ stratum   <dbl> 1938, 1954, 1957, 1937, 1937, 1938, 1970, 1955, 1970, 1947, 1967, 1958, 1949, 1944, 1957, 1972, 1937, 1970, …
$ psu       <dbl> 16070, 15663, 16177, 16224, 15535, 15457, 16328, 15361, 16276, 16261, 16171, 15865, 15326, 16170, 16112, 152…
$ eu_vote   <chr> "Remain Outside", "Remain Outside", "Join EU", "Join EU", "Remain Outside", "Join EU", "Remain Outside", "Re…
$ brnnorge  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, …
$ agea      <dbl> 55, 58, 35, 18, 22, 57, 48, 46, 35, 30, 30, 32, 20, 74, 47, 51, 31, 41, 59, 17, 77, 19, 66, 38, 53, 74, 86, …
$ imbgeco   <dbl> 5, 7, 8, 5, 5, 9, 8, 7, 7, 8, 6, 6, 8, 9, 9, 10, 8, 3, 5, 7, 3, 7, 7, 4, 5, 5, 3, 8, 8, 5, 5, 6, 9, 8, 8, 10…
$ imueclt   <dbl> 7, 8, 8, 5, 5, 2, 10, 7, 7, 8, 6, 10, 10, 9, 8, 10, 10, 7, 5, 9, 6, 7, 6, 4, 5, 5, 6, 6, 8, 6, 5, 8, 9, 5, 1…
$ imwbcnt   <dbl> 5, 6, 8, 9, 6, 4, 7, 5, 5, 8, 5, 10, 9, 10, 9, 10, 9, 6, 5, 9, 5, 8, 8, 7, 5, 5, 7, 5, 8, 6, 6, 7, 7, 5, 10,…
$ female    <dbl> 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, …
$ eduyrs    <dbl> 10, 15, 19, 12, 13, 12, NA, 20, 15, 17, NA, 13, 14, 9, NA, 23, 17, 11, 12, 11, 14, 12, 16, 16, 12, 10, 16, 1…
$ uempla    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ polint    <dbl> 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, …
$ hinctnta  <dbl> 6, 8, 5, 9, 4, 4, 10, 10, 3, 8, 5, 5, 2, 2, 9, 3, 4, NA, 3, 6, 2, 8, 4, 3, 3, 1, 3, 5, 1, 3, 8, 8, 6, 9, 10,…
$ lrscale   <dbl> 6, 2, 4, 3, 8, 8, 7, 7, NA, 3, 6, 7, 2, 9, 4, 7, 2, 5, 7, 6, 8, 4, 9, 6, 4, 7, 5, 3, 3, 8, 5, 2, 7, 9, 5, 3,…
```

Cu ajutorul funcției „glimpse()” din pachetul „ dplyr” pot ilustra diferite caracteristici ale setului de date. Acesta are 1411 observații și 25 de variabile dintre care 2 sunt de tip șir de caractere, 2 de tip date-time iar 21 de tip numeric. Majoritatea acestor variabile definesc caracteristici socio-demografice ale indivizilor cum ar fi sexul (*female*), vârsta (*agea*), venitul gospodăriei (*hinctnta*), regiunea de proveniență (*region*), anii de educație (*educyrs*) ș.a.m.d. Alte variabile de interes sunt *imbgeco, imueclt și imwbcnt*. Acestea reprezintă gradul de toleranță a indivizilor, pe o scală de la 1 la 10, când vine vorba de diverse moduri in care imigranți impactează societate. Variabila *imbgeco* arată dacă respondentul consideră că imigranții sunt în general buni sau răi pentru economia Norvegiei, variabila *imueclt* arată dacă respondentul consideră că imigranții îmbogățesc sau subminează cultura Norvegiei și variabila *imwbcnt* arată dacă respondentul consideră că imigranții fac din Norvegia un loc mai bun pentru a trăi.

Pe lângă aceste variabile mai sunt prezente și atribute care țin strict de modul în care a fost ales eșantionul pentru a se asigura ca acesta este reprezentativ, cum ar fi probabilitatea ca

un individ să fie inclus în eșantion (prob). De asemenea este specificat momentul in care a început interviul (*inwds*) și momentul în care s-a sfârșit interviul (*inwde*).

#Redenumirea variabilelor

```
colnames(ESS10NO)
nume_col<- c("rownames", "country", "id", "region", "int_start", "int_end", "dweight",
             "pspwght", "pweight", "anweight", "prob", "stratum", "psu", "eu_vote",
             "born_NO", "age", "img_econ", "img_culture", "img_bptl", "female",
             "educ_yrs", "unemployed", "pol_interest", "household_income", "lrscale")
names(ESS10NO) <- nume_col
```

Prima etapă a procesului de preprocesare este redenumirea coloanelor pentru ca acestea să fie mai ușor de înțeles. Acest proces este necesar întrucât denumirile inițiale mi s-au părut neprietenoase și neintuitive.

 # Crearea unei variabile categoriale "img_bptl_fact" bazate pe variabila numerica "img_bptl"

```
ESS10NO$img_bptl_fact <- cut(ESS10NO$img_bptl, c(0, 3, 6, 10), c('Low Tolerance',
                                'Medium Tolerance', 'High Tolerance'))
```

Am creat o variabilă categorială conform cerențeilor obligatorii pentru dezvoltarea proiectului. Variabila nou-creată este de tip factor și repartizează respondenții, în funcție de valoarea variabilei *img_bptl*, pe 3 nivele de toleranță a imigranților:

- **Low Tolerance:** în această categorie se încadrează oameni care au valori de la 0 la 3 pentru variabila *img_bptl.*
- **Medium Tolerance:** categoria aceasta înglobează respondenți care au valori de la 4 la 6 pentru variabila *img_bptl.*
- **High Tolerance:** această ultimă categorie cuprinde persoanele care au valori de la 7 la 10 pentru variabila *img_bptl*.

#Alegerea variabilelor ce urmează sa fie folosite in proiect având in vedere numărul de valori lipsa si valorile eronate

```
summary(ESS10NO)
which(is.na(ESS10NO$region))
which(is.na(ESS10NO$eu_vote))
```

În urma aplicării funcțiilor pentru observarea numărului de valori lipsă și a valorilor eronate am decis sa aleg următoarele variabile:

- **Img_bptl (numeric)**: această variabilă arată dacă respondentul consideră că imigranții fac din Norvegia un loc mai bun pentru a trăi. Aceasta a fost aleasă în detrimentul celorlalte două variabile legate de imigranți întrucât reflectă mai bine atitudinea față de problematica imigrării și prezintă mai puține valori lipsă
- **age (numeric):** această variabilă ilustrează vârsta respondentului și a fost aleasă deoarece nu conține valori lipsă
- **educ_yrs (numeric):** arată indică numărul de ani de studiu și a fost aleasă în detrimentul variabilei *household_income* pentru ca are mult mai puține valori lipsă
- **img_bptl_fact (nenumeric):** creată pe baza variabilei *img_bptl*. Este o variabilă categorială cu 3 care împarte respondenții în categorii în funcție de toleranța lor față de imigranți
- **eu_vote (nenumeric):** ilustrează modul în care ar vota respondentul dacă ar avea opțiunea de a intra în UE. Variabila a fost aleasă pentru a ilustra legătura dintre percepția față de imigranți și percepția față de UE

#Operatii preliminare

#Din baza inițiala se va face o selecție care sa includă condiții pentru cel puțin doua variabile. Cerințele proiectului vor fi executate pentru aceasta selecție.

```
bd <- ESS10NO |>
  filter(born_NO == 1 & pol_interest == 1)|>
  select(img_bptl, age, educ_yrs, img_bptl_fact, eu_vote)
```

Am construit un nou obiect denumit „bd” în care am inclus doar variabilele ce vor face parte din analiza viitoare. De asemenea am făcut selecția obligatorie ce limitează eșantionul la persoane care au un interes în politică și care totodată au fost născute în Norvegia.

#Noua baza va fi exportata

```
library(writexl)
write_xlsx(bd, 'C:/Users/User/Desktop/Master/R/bd.xlsx')
```

#Definirea categoriilor variabilelor categoriale.

```
bd$eu_vote <- as.factor(bd$eu_vote)
levels(bd$eu_vote)
```

```
> levels(bd$eu_vote)
[1] "Blank Ballot"     "Don't Know"       "Join EU"
"Not Eligible"     "Refuse to Answer"
[6] "Remain Outside"   "Wouldn't Vote"
```

Variabila *eu_vote* nu este percepută ca variabilă categorială, ci este de tip șir de caractere deși valorile acesteia pot fi împărțite în categorii. După aplicarea funcției ”as.factor()„ am transformat variabila în tipul categorial iar aceasta a fost împărțită automat în 7 categorii. De asemenea, variabila *img_bptl_fact* este deja definită ca fiind categorială.

#Prezentarea bazei de date finale

glimpse(bd)

levels(bd$eu_vote)

levels(bd$img_bptl_fact)

```
> glimpse(bd)
Rows: 683
Columns: 5
$ img_bptl      <dbl> 5, 6, 6, 4, 8, 10, 10, 9, 5, 8, 5, 5, 7, 8, 6, 7, 7, 5, 10, 3, 4, 3, 5, 6, 5, 6, 5, 7, 2, 8, 5, 5, 6, 9…
$ age           <dbl> 55, 58, 22, 57, 30, 32, 74, 31, 77, 66, 53, 74, 86, 24, 61, 64, 58, 57, 69, 71, 32, 55, 57, 57, 32, 77,…
$ educ_yrs      <dbl> 10, 15, 13, 12, 17, 13, 9, 17, 14, 16, 12, 10, 16, 16, 14, 17, 13, 14, 14, 16, 15, 14, 15, 15, 13, 12, …
$ img_bptl_fact <fct> Medium Tolerance, Medium Tolerance, Medium Tolerance, Medium Tolerance, High Tolerance, High Tolerance,…
$ eu_vote       <fct> Remain Outside, Remain Outside, Remain Outside, Join EU, Remain Outside, Remain Outside, Remain Outside…
> levels(bd$eu_vote)
[1] "Blank Ballot"     "Don't Know"       "Join EU"          "Not Eligible"     "Refuse to Answer" "Remain Outside"
[7] "Wouldn't Vote"
> levels(bd$img_bptl_fact)
[1] "Low Tolerance"    "Medium Tolerance" "High Tolerance"
```

În urma operațiunilor preliminare setul de date a ajuns de la 25 de variabile la 5, iar numărul de observații a ajuns de la 1411 la 683. În componența setului de date sunt incluse 3 variabile numerice care indică vârsta respondentului (*age*), ani de educație (*educ_yrs*) și părerea respondentului cu privire la ideea că imigranți fac din Norvegia un loc mai bun pentru a trăi, pe o scala de la 0 la 10 (*img_bptl*). De asemenea, în setul de date sunt prezente 2 variabile categoriale, una care indică modul în care ar vota respondentul în cazul efectuări unui referendum cu privire la aderarea la UE (*eu_vote*) și nivelul de toleranță față de imigranți (*img_bptl_fact*).

3. Analiza grafica si numerica a variabilelor analizate

#Crearea unui nou set de date ce conține doar variabilele numerice

bd_num <- bd|>

select(img_bptl, age, educ_yrs)

Baza de date asupra căreia va fi executată analiza descriptivă a variabilelor numerice se numește ”bd_num„ și conține variabilele *img_bptl, age* și *educ_yrs.*

3.1. Analiza descriptiva numerica a variabilelor numerice si nenumerice

#Analiza indicatorilor descriptivi

summary(bd_num)

```
> summary(bd_num)
    img_bptl          age            educ_yrs
 Min.   : 0.000   Min.   :16.00   Min.   : 0.0
 1st Qu.: 5.000   1st Qu.:40.00   1st Qu.:13.0
 Median : 7.000   Median :55.00   Median :16.0
 Mean   : 6.541   Mean   :52.78   Mean   :15.4
 3rd Qu.: 8.000   3rd Qu.:67.00   3rd Qu.:18.0
 Max.   :10.000   Max.   :90.00   Max.   :28.0
 NA's   :3                        NA's   :2
```

La nivelul celor 3 variabile se pot observa 5 valori lipsă și nici o eroare evidentă, de asemenea indicatorii descriptivi ai variabilelor pot fi interpretați în felul următor:

- **Img_bptl:**

Min: Valoarea cea mai mică a variabilei este 0

1st Qu: Primi 25% din oamenii din eșantion au scoruri de sub 5 iar 75% au scoruri de peste 5 în cazul indicelui de toleranță față de imigranți.

Median: Prima jumătate a respondenților din eșantion au scoruri sub 7 iar cealaltă jumătate au scoruri peste 7 în cazul indicelui de toleranță față de imigranți.

Mean: Valoarea medie pentru eșantion a toleranței față de imigranți este de 6.541

3rd Qu: Primi 75% din respondenți au scoruri sub 8 în timp ce restul de 25% au scoruri peste 8 în cazul indicelui de toleranță față de imigranți.

Max: Valoarea maximă înregistrată pentru variabilă este 10.

- **age:**

Min: Cea mai mică vârstă a unui respondent este de 16 ani.

1st Qu: Primi 25% din respondenți au sub 40 de ani, iar restul de 75% au peste 40 de ani.

Median: Prima jumătate a eșantionului este compusă din respondenți cu sub 55 de ani, iar cealaltă jumătate au peste 55 de ani.

Mean: Vârsta medie a eșantionului este de 52.7 ani.

3rd Qu: : Primi 75% din respondenți au sub 67 de ani, în timp ce restul de 75% au peste 67 de ani.

Max: Cea mai mare vârstă a unui respondent este de 90 de ani.

- **educ_yrs:**

Min: Ani de studiu a unui respondent încep de la valoarea 0.

1st Qu: Primi 25% din respondenți au sub 13 ani de studiu, iar restul de 75% au peste 13 ani de studiu.

Median: Prima jumătate din respondenții din eșantion au sub 16 ani de studiu, iar cealaltă jumătate au peste 16 ani de studiu.

Mean: În medie, un respondent din eșantion are 15.4 ani de studiu.

3rd Qu: Primi 75% din respondenți sub 18 ani de studiu, iar restul de 25% au peste 18 ani de studiu.

Max: Valoarea maximă a anilor de studiu pentru respondenții din eșantion este de 28 de ani.

Pentru a analiza și alți indicatori descriptivi relevanți voi folosi funcția ”describe()„.

library(psych)

describe(bd_num)

```
> bd_num <- bd|>
+   select(img_bptl, age, educ_yrs)
> library(psych)
> describe(bd_num)
         vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 680  6.54  2.01      7    6.58  2.97   0  10    10 -0.26     0.06 0.08
age         2 683 52.78 17.19     55   53.32 19.27  16  90    74 -0.24    -0.80 0.66
educ_yrs    3 681 15.40  3.78     16   15.53  2.97   0  28    28 -0.38     0.75 0.14
```

- **mean & trimmed:** Comparând cei doi indicatori putem observa că nu este o diferență foarte mare între medie și media ajustată ceea ce indică o lipsă de outlieri.
- **skew:** Toate cele 3 distribuții ale variabilelor sunt asimetrice la stânga întrucât indicatorul de asimetrie este mai mic decât 0.
- **kurtosis:** Distribuțiile variabilelor *img_bptl* și *educ_yrs* sunt leptocurtice întrucât indicatorul de boltire este mai mare decât 0, iar în cazul distribuției variabilei *age* distribuția este platicurtică pentru că indicatorul este mai mic decât 0.

#Analiza indicatorilor descriptive pe grupe (în funcție de variabila eu_vote)

describeBy(bd_num, group = bd$eu_vote, digits = 4)

În urma analizei pe grupe se poate observa discrepanțe destul de mari între numărul de observații per categorie. În timp ce în categoria ”Wouldn't Vote„ nu există decât o singură observație, în categoria ”Remain Outside„ există 444.

```
 Descriptive statistics by group
group: Blank Ballot
         vars n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 7  7.57  2.51      8    7.57  2.97   5  10     5 -0.07    -2.15 0.95
age         2 7 49.29 23.61     48   49.29 32.62  24  79    55  0.14    -1.97 8.92
educ_yrs    3 7 15.14  4.56     18   15.14  0.00   8  18    10 -0.75    -1.58 1.72
------------------------------------------------------------
group: Don't Know
         vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 28  7.39  1.87    8.0    7.50  1.48   2  10     8 -0.72     0.32 0.35
age         2 28 49.89 16.26   52.5   49.71 16.31  23  86    63  0.04    -0.91 3.07
educ_yrs    3 28 16.75  3.48   17.0   16.88  2.97   9  24    15 -0.45    -0.12 0.66
------------------------------------------------------------
group: Join EU
         vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 196  6.96  1.86    7.0    6.93  2.97   1  10     9 -0.08    -0.58 0.13
age         2 198 54.28 16.53   56.5   54.84 17.05  17  90    73 -0.26    -0.65 1.17
educ_yrs    3 198 16.03  3.36   17.0   16.08  2.97   5  25    20 -0.18     0.33 0.24
------------------------------------------------------------
group: Not Eligible
         vars n  mean   sd median trimmed mad min max range  skew kurtosis   se
img_bptl    1 3  5.67 0.58      6    5.67   0   5   6     1 -0.38    -2.33 0.33
age         2 3 17.00 0.00     17   17.00   0  17  17     0   NaN      NaN 0.00
educ_yrs    3 3 11.00 0.00     11   11.00   0  11  11     0   NaN      NaN 0.00
------------------------------------------------------------
group: Refuse to Answer
         vars n mean   sd median trimmed  mad min max range skew kurtosis  se
img_bptl    1 2  8.0 1.41    8.0     8.0 1.48   7   9     2    0    -2.75 1.0
age         2 2 55.5 6.36   55.5    55.5 6.67  51  60     9    0    -2.75 4.5
educ_yrs    3 1 18.0   NA   18.0    18.0 0.00  18  18     0   NA       NA  NA
------------------------------------------------------------
group: Remain Outside
         vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 443  6.28  2.03    6.0    6.33  1.48   0  10    10 -0.28     0.20 0.10
age         2 444 52.66 17.21   53.5   53.20 20.02  16  89    73 -0.24    -0.86 0.82
educ_yrs    3 443 15.07  3.91   16.0   15.21  2.97   0  28    28 -0.38     0.83 0.19
------------------------------------------------------------
group: Wouldn't Vote
         vars n mean sd median trimmed mad min max range skew kurtosis se
img_bptl    1 1    8 NA      8       8   0   8   8     0   NA       NA NA
age         2 1   17 NA     17      17   0  17  17     0   NA       NA NA
educ_yrs    3 1   10 NA     10      10   0  10  10     0   NA       NA NA
```

Pentru prima grupă și anume aceea a oamenilor care și-ar anula votul în cazul unui referendum ce vizează aderarea la UE, putem observa că acest grup este relativ mai tolerant la ideea de Imigrare comparativ cu media eșantionului, având un scor de 7.39 comparativ cu 6.54 (trebuie menționat că deviația standard este mai mare decât media eșantionului deci valorile se pot abate mai mult de la medie). În același timp, media de vârstă este mai mică cu 3 ani comparativ cu media eșantionului, iar anii medii de educație este de asemenea mai mică. Diferența dintre medie și media ajustată nu este semnificativă ceea ce subliniază lipsa outlierilor pe acest segment al eșantionului.

În cazul celui de-al 2-lea grup format din 28 de respondenți care sunt indecișI în legătură cu modul în care ar vota, situația este similară, media toleranței față de imigranți este mai mare (și abaterea standard este mai mică deci pentru această categorie valorile se abat mai puțin de

la medie decât în cazul eșantionului), vârsta medie este mai mică dar anii medii de educației sunt mai mari decât cei din eșantion.

Al treilea grup, și unul dintre cele mai consistente ca număr de observații, valorile medii a tuturor variabilelor sunt peste valorile medii ale eșantionului și de asemenea, abaterile standard sunt mai reduse ceea ce înseamnă ca dispersia de la nivelul grupului este mai mare decât cea de la nivelul acestui grup pentru toate variabilele. Un alt aspect ce merită precizat este diferența mică dintre medie și media ajustată, fapt care reflectă faptul că outlieri nu sunt o problemă pentru nici una dintre variabile pentru această categorie.

Grupurile ”Not Eligible„, ”Refuse to Answer„ și ”Wouldn't Vote„ au 3, 2 și respectiv 1 observație, acestea reprezentând o parte nesemnificativă a eșantionului. Ultima și cea mai mare categorie ca număr de observații este grupul ”Remain Outside„ constituită din persoane care ar vota ca Norvegia să nu adere la UE în cazul unui referendum. Acest grup este unul din două grupuri care au media toleranței față de imigranți mai mică decât media eșantionului. Numărul mare de observații duce la cel mai mare range, la nivelul celor 3 variabile, comparativ cu celelalte grupuri.

#Analiza indicatorilor descriptive pe grupe (în funcție de variabila img_bptl_fact)

describeBy(bd_num, group = bd$img_bptl_fact, digits = 4)

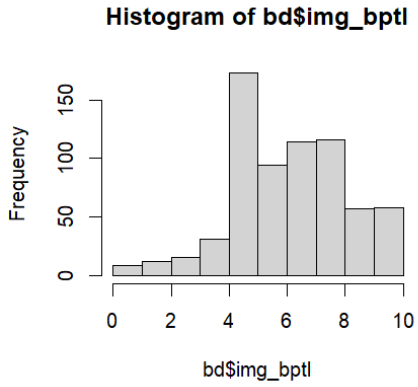
Diferitele niveluri de toleranță au fost create luându-se în considerare valorile variabilei *img_bptl* lucru ce va face va face ca mediile acestei variabile să crească odată cu nivelul de toleranță. Ca în cazul variabilei *eu_vote,* variabila *img_bptl_fact* prezintă discrepanțe între numărul de observații per categorie.

```
 Descriptive statistics by group
group: Low Tolerance
         vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 37  2.05  1.05      2    2.16  1.48   0   3     3 -0.80    -0.65 0.17
age         2 37 57.89 13.49     58   58.35 11.86  23  84    61 -0.34     0.00 2.22
educ_yrs    3 37 13.49  3.29     13   13.45  4.45   7  20    13  0.00    -0.98 0.54
------------------------------------------------------------------------------------
group: Medium Tolerance
         vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 298  5.21  0.61      5    5.26  0.00   4   6     2 -0.15    -0.54 0.04
age         2 298 54.16 17.47     56   55.06 19.27  16  87    71 -0.39    -0.68 1.01
educ_yrs    3 298 15.05  3.76     15   15.13  4.45   3  28    25 -0.12     0.65 0.22
------------------------------------------------------------------------------------
group: High Tolerance
         vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
img_bptl    1 345  8.17  1.07      8    8.09  1.48   7  10     3  0.48    -1.02 0.06
age         2 345 51.11 17.09     52   51.26 20.76  16  90    74 -0.08    -0.92 0.92
educ_yrs    3 343 15.91  3.75     16   16.12  2.97   0  25    25 -0.69     1.37 0.20
```

Primul lucru de remarcat referitor la analiza indicatorilor descriptivi este modul în care evoluează mediile variabilelor în funcție de grup. Media vârstei urmează un trend descendent odată cu creșterea nivelului de toleranță în timp ce media anilor de studii urmează un trend ascendent. La nivelul tuturor grupurilor se poate observa cum distribuțiile variabilelor sunt leptocurtice întrucât indicatorul de boltire este mai mare decât 0.

3.2. Analiza grafica a variabilelor numerice si nenumerice

Pentru a putea analiza grafic variabilele numerice mă voi folosi de histograme, întrucât acestea permit o vizualizare precisă a datelor.



În cazul variabilei *age*, histograma indică o distribuție asimetrică spre stânga, cu o concentrație mai mare de respondenți în intervalul de vârstă 60–80 de ani. De asemenea, se observă o scădere treptată a frecvenței în rândul vârstelor mai tinere, ceea ce sugerează o participare mai redusă a respondenților sub 40 de ani. Acest lucru ar putea reflecta structura demografică a populației care a răspuns la chestionar.

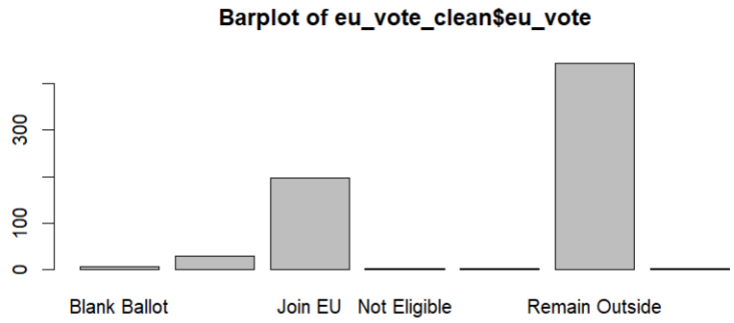
Histograma variabilei *edu_yrs*, care măsoară anii de educație, arată o distribuție aproximativ simetrică, cu un vârf în jurul valorii de 15-17 ani de studiu. Acest model este coerent cu un nivel de educație ridicat al populației norvegiene, reflectând atât accesul larg la educație, cât și politica țării în această direcție. Extremele acestei variabile sunt mai puțin frecvente, ceea ce sugerează că un număr mic de respondenți fie au o educație redusă (sub 10 ani), fie au studii superioare îndelungate (peste 20 de ani).

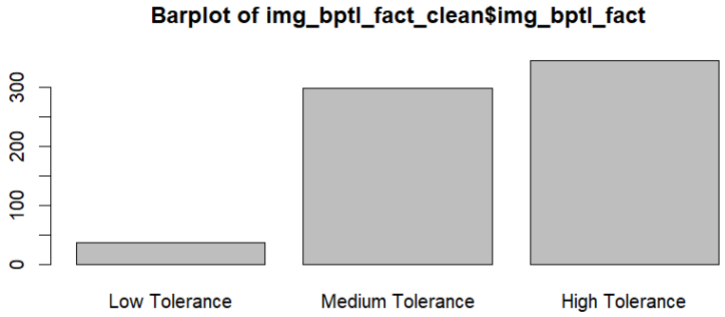
Pentru variabila *img_bptl*, care evaluează percepția impactului imigrației asupra condițiilor de trai, histograma arată o distribuție asimetrică spre stânga. Majoritatea respondenților evaluează acest impact în intervalul 5–8, ceea ce indică o percepție progresistă față de imigranți. Valori scăzute (sub 5) și ridicate (peste 8) sunt mai puțin frecvente, ceea ce sugerează că opiniile extreme sunt relativ rare.

#Analiza grafică a variabilelor categoriale

```
eu_vote_clean = bd|>
        filter(!is.na(bd$eu_vote))
        plot(eu_vote_clean$eu_vote)
img_bptl_fact_clean <- bd|>
        filter(!is.na(bd$eu_vote))
        plot(img_bptl_fact_clean$img_bptl_fact)
```

Pentru a putea construi barplot-urile variabilelor categoriale trebuie mai întâi să fac două selecție din data-frame-ul principal care să conțină pe de o parte, observațiile fără valori lipsă pentru variabila *eu_vote*, iar pe de altă parte, observațiile fără valori lipsă pentru variabila *img_bptl. În acest scop, am creat două obiecte noi eu_vote_clean* și *img_bptl_fact_clean* pe baza cărora vor fi construite graficele.



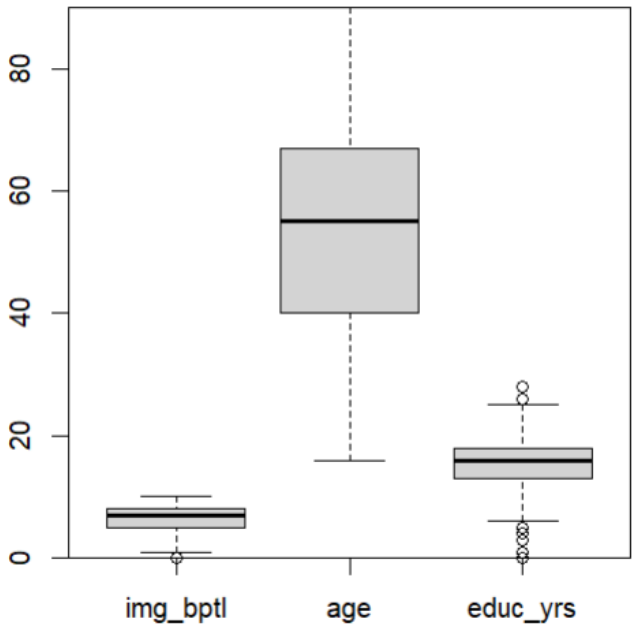


În primul barplot, care analizează variabila *eu_vote*, observăm o distribuție clară a preferințelor respondenților. Cea mai mare proporție dintre aceștia susține menținerea Norvegiei în afara Uniunii Europene, această opțiune având un număr semnificativ mai mare de respondenți comparativ cu celelalte categorii. Următoarea opțiune preferată, cu o frecvență semnificativ mai mică, este aderarea la Uniunea Europeană. Categoriile „Blank Ballot” și „Not Eligible” au o prezență marginală, indicând că doar o mică parte dintre respondenți fie nu și-ar exprima opinia, fie nu ar fi eligibili pentru vot. Această distribuție sugerează o opoziție majoritară a cetățenilor norvegieni față de integrarea în Uniunea Europeană, o atitudine ce poate fi asociată cu factori culturali, economici sau politici specifici contextului norvegian.

Al doilea barplot, care ilustrează distribuția variabilei *img_bptl_fact*, evidențiază nivelurile de toleranță față de imigranți. Se remarcă o distribuție echilibrată între nivelurile „Medium Tolerance” și „High Tolerance”, care au frecvențe similare și dominante în cadrul eșantionului. Pe de altă parte, categoria „Low Tolerance” este mult mai puțin frecventă, indicând o atitudine general pozitivă sau moderată față de imigranți. Această tendință poate fi interpretată ca un indicator al unui grad ridicat de deschidere culturală și acceptare socială în rândul populației norvegiene. Totuși, diferențierea dintre nivelurile de toleranță medie și ridicată ar putea reflecta diverși factori personali sau socio-economici care influențează percepția asupra imigrației.

3.3. Identificarea outlierilor si eliminarea acestora din baza (sau înlocuirea lor cu valori lipsa)

boxplot(bd[-c(4, 5)])



Pentru variabila *age*, distribuția arată o gamă largă de valori, reflectată printr-un interval intercuartilic considerabil și o mediană plasată aproximativ la mijlocul acestui interval. Valoarea medianei indică faptul că vârsta tipică a respondenților se situează în jurul unei valori centrale (aproximativ 50-60 de ani). Extremitățile sugerează prezența unei varietăți semnificative în vârste, deși nu se observă outlieri notabili, ceea ce denotă o distribuție relativ omogenă fără valori extreme în afara limitelor așteptate.

Pentru variabila *edu_yrs*, boxplotul arată o mediană bine definită, indicând că majoritatea respondenților au un număr mediu de ani de educație situat în jurul valorii de 12-15 ani. Intervalul intercuartilic este mai restrâns decât la variabila *age*, sugerând o variabilitate mai mică în anii de studii ai respondenților. Totuși, prezența unor puncte individuale în afara intervalului principal indică existența unor outlieri. Deși există outlieri, aceștia nu influențează negativ analiza datelor, fiind valori reale și relevante, astfel încât nu vor fi eliminate din analiză.

În ceea ce privește variabila *img_bptl*, boxplotul arată că majoritatea respondenților au acordat un scor ridicat impactului imigrației asupra condițiilor de trai, mediană situându-se în intervalul superior al scalei de 0-10. Variabilitatea în acest caz este mult mai redusă, așa cum indică lățimea foarte mică a intervalului intercuartilic. Totodată, există câteva puncte izolate,

corespunzând unor respondenți care au acordat scoruri mult mai mici, ceea ce sugerează percepții divergente, deși acestea sunt rare și se află la marginea distribuției.

4. Analiza statistica a variabilelor categoriale

Acest capitol are ca scop analizarea variabilelor categoriale *img_bptl_fact* și *eu_vote*. Obiectivul va fi atins cu ajutorul procedeului de tabelare a datelor, mai precis obținerea și interpretarea frecvențelor marginale, condiționate și parțiale. Ulterior vor fi elaborate analiza de asociere și cea de concordanță pentru a se studia existența unei relații între cele două variabile.

4.1. Tabelarea datelor (obținere frecvente marginale, condiționate, parțiale)

#frecvețe marginale

Pentru a putea analiza frecvențele marginale a fiecărei variabile categoriale în parte am creat două obiecte noi pe baza tabelului de frecvență a celor două variabile. Primul obiect, denumit *frecv_marg_eu_vote* însumează valorile tabelului de frecvență, pe linie, subliniind astfel frecvențele variabilei *eu_vote*.

```
tabel <- table(bd$eu_vote, bd$img_bptl_fact)
frecv_marg_eu_vote <- rowSums(tabel)
as.data.frame(frecv_marg_eu_vote)
```

```
> tabel <- table(bd$eu_vote, bd$img_bptl_fact)
> frecv_marg_eu_vote <- rowSums(tabel)
> as.data.frame(frecv_marg_eu_vote)
                frecv_marg_eu_vote
Blank Ballot                     7
Don't Know                      28
Join EU                        196
Not Eligible                     3
Refuse to Answer                 2
Remain Outside                 443
Wouldn't Vote                    1
```

Tabelul de frecvență prezentat ilustrează preferințele de vot ale respondenților norvegieni în cazul unui referendum ipotetic privind aderarea la Uniunea Europeană. Se observă o predominanță a respondenților care ar vota împotriva aderării (443), reprezentând o majoritate clară. Un număr semnificativ de respondenți (196) s-au declarat în favoarea aderării la UE, în timp ce un procent mic (7) au indicat că și-ar anula votul. De asemenea, un număr relativ mic de respondenți nu au o opinie formată (28) sau au refuzat să răspundă (2).

Al doilea obiect, denumit *frecv_marg_img_bptl* a fost creat pe premise asemănătoare. Singura diferență este că, în loc să însumez liniile tabelului de frecvență, am însumat coloanele, subliniind astfel frecventele variabilei *img_bptl*.

```
frecv_marg_img_bptl <- colSums(tabel)
as.data.frame(frecv_marg_img_bptl)
```

```
> frecv_marg_img_bptl <- colSums(tabel)
> as.data.frame(frecv_marg_img_bptl)
                  frecv_marg_img_bptl
Low Tolerance                      37
Medium Tolerance                  298
High Tolerance                    345
```

Tabelul rezultat oferă o imagine asupra distribuției respondenților norvegieni în funcție de nivelul de toleranță. Se observă o predominanță a persoanelor cu toleranță ridicată (345), urmate de cei cu toleranță medie (298). Persoanele cu toleranță scăzută reprezintă un segment redus din eșantion (37). Această distribuție sugerează o tendință generală către un nivel de toleranță ridicat în rândul respondenților norvegieni.

#frecvențe condiționate

```
round(prop.table(tabel, margin = 1), 4) * 100
```

```
                Low Tolerance Medium Tolerance High Tolerance
Blank Ballot             0.00             42.86          57.14
Don't Know               3.57             28.57          67.86
Join EU                  1.53             40.31          58.16
Not Eligible             0.00            100.00           0.00
Refuse to Answer         0.00              0.00         100.00
Remain Outside           7.45             46.28          46.28
Wouldn't Vote            0.00              0.00         100.00
```

Analizând datele, se observă o polarizare a opțiunilor de vot pe măsură ce nivelul de toleranță față de imigranți crește. În rândul respondenților cu toleranță scăzută, un procent foarte mic (1,53%) ar vota pentru aderarea la UE, iar o pondere ceva mai mare (7,45%) ar prefera ca Norvegia să rămână în afara Uniunii Europene. Legat de atitudinea față de aderarea la UE, se poate observa că la nivelul opțiunilor pe care foarte puțini respondenți le-ar alege (cum ar fi ”Refuse to Answer„, ”Wouldn't Vote„ sau ”Don't Know„) 0% din respondenți o un nivel scăzut de toleranță față de imigranți.

În cazul respondenților cu un nivel mediu de toleranță, se remarcă o creștere semnificativă a procentului celor care ar sprijini aderarea la UE (40,31%). În același timp, ponderea celor care ar prefera ca Norvegia să rămână în afara Uniunii Europene crește la 46,28%. Respondenții cu un nivel ridicat de toleranță manifestă o tendință clară de sprijin pentru aderarea la UE, cu un procent majoritar de 58,16%. În mod interesant, procentul celor care ar opta pentru rămânerea în afara Uniunii Europene este de doar 46,28%.

 # frecvențe parțiale

tabel["Wouldn't Vote", 'High Tolerance']

tabel['Refuse to Answer', 'High Tolerance']

tabel['Not Eligible', 'Medium Tolerance']

În cadrul tabelului anterior se puteau observa anumite celule cu valoarea de 100%. Acest lucru poate fi explicat de numărul redus de respondenți care ar alege anumite opțiuni. Cu ajutorul frecvențelor parțiale putem identifica numărul de persoane aflate în aceste celule cu procentaje atipice.

```
> # frecv partiale
> tabel["Wouldn't Vote", 'High Tolerance']
[1] 1
> tabel['Refuse to Answer', 'High Tolerance']
[1] 2
> tabel['Not Eligible', 'Medium Tolerance']
[1] 3
```

În urma selectării celulelor care indică un procentaj de 100% în tabelul de frecvențe procentuale, din tabelul cu frecvențe absolute se poate observa un număr insignifiant pentru respondenți care nu ar vota în cazul unui referendum ipotetic și de asemenea au un grad crescut de toleranță, pentru persoanele care au refuzat să răspundă și care au un grad ridicat de toleranță față de imigranți și pentru cei care nu sunt eligibili și au un grad mediu de toleranță.

4.2. Analiza de asociere

Analiza de asociere a fost realizată pe baza tabelului de frecvențe absolute, cu ajutorul funcției summary() care, în acest context, efectuează un test Chi-square.

summary(tabel)

Formularea ipotezelor:

H0: Între atitudinea față de aderarea la UE și toleranța față de imigranți nu există o asociere semnificativă (variabilele sunt independente).

H1: Între atitudinea față de aderarea la UE și toleranța față de imigranți există o asociere semnificativă (variabilele sunt independente).

```
> summary(tabel)
Number of cases in table: 680
Number of factors: 2
Test for independence of all factors:
        Chisq = 24.303, df = 12, p-value = 0.0185
        Chi-squared approximation may be incorrect
```

P-value = 0.0185 < 0.05 se respinge H0, cu un risc de 5%.

Interpretare: Cu o probabilitate de 95% putem afirma că între poziționarea norvegienilor față de imigrare și modul in care votează pentru aderarea la UE există o asociere semnificativă.

4.3. Analiza de concordanță

În acest subcapitol sunt efectuate teste de concordanță atât pentru variabila *img_bptl_fact*, cât și pentru variabila *eu_vote* pentru a se verifica dacă distribuțiile acestora sunt sau nu egalitare.

chisq.test(table(bd$img_bptl_fact))

Formularea ipotezelor:

H0: distribuția variabilei *img_bptl_fact* este egalitară.

H1: distribuția variabilei *img_bptl_fact* nu este egalitară.

```
> chisq.test(table(bd$img_bptl_fact))

        Chi-squared test for given probabilities

data:  table(bd$img_bptl_fact)
X-squared = 242.93, df = 2, p-value < 2.2e-16
```

P-value < 2.2e-16, se respinge cu un risc de 1% H0.

Interpretare: Interpretare: Cu o probabilitate de 99% putem afirma că distribuția variabilei *img_bptl_fact* nu este egalitară.

chisq.test(table(bd$eu_vote))

Formularea ipotezelor:

H0: distribuția variabilei *eu_vote* este egalitară.

H1: distribuția variabilei *eu_vote* nu este egalitară.

```
> chisq.test(table(bd$eu_vote))

        Chi-squared test for given probabilities

data:  table(bd$eu_vote)
X-squared = 1747.9, df = 6, p-value < 2.2e-16
```

P-value < 2.2e-16, se respinge cu un risc de 1% H0.

Interpretare: Cu o probabilitate de 99% putem afirma că distribuția variabilei *eu_vote* nu este egalitară.

5. Analiza de regresie si corelație

5.1. Analiza de corelație

 # Matricea de corelație

Pentru a analiza corelația dintre variabilele numerice se va folosi funcții cor() cu ajutorul căreia poate fi construită matricea de corelație. Întrucât există valori lipsă pentru două din cele trei variabile voi dai valoarea ”complete” argumentului use, iar metoda de calcul este aceia a coeficientului de corelație Pearson.

cor(bd[-c(4, 5)], use = 'complete')

```
> cor(bd[-c(4, 5)], use = 'complete')
            img_bptl        age   educ_yrs
img_bptl   1.0000000 -0.1244088  0.1840379
age       -0.1244088  1.0000000 -0.1227818
educ_yrs   0.1840379 -0.1227818  1.0000000
```

Din matricea corelației pot fi trase următoarele concluzii:

- Între toleranța față de imigranți și vârstă există o legătură indirectă de intensitate mică
- Între toleranța față de imigranți și ani de educația există o legătură directă de intensitate mică

 # Test de corelație

Pentru a înțelege mai profund intensitatea legăturilor dintre variabilele numerice voi efectua două teste de corelație între variabila *img_bptl* și celelalte două bariabile *age* și *educ_yrs*.

cor.test(bdimg_bptl, bdage)

Formularea ipotezelor:

H0: Între toleranța față de imigranți și vârstă nu există corelație.

H1: Între toleranța față de imigranți și vârstă există corelație.

```
> cor.test(bd$img_bptl, bd$age)

        Pearson's product-moment correlation

data:  bd$img_bptl and bd$age
t = -3.2612, df = 678, p-value = 0.001165
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.19761538 -0.04955411
sample estimates:
       cor
-0.1242765
```

P-value = 0.001165 < 0.05, se respinge H0 cu un risc de 5%.

Interpretare: Putem afirma cu o probabilitate de 95% că între toleranța față de imigranți și vârstă există corelație.

cor.test(bdimg_bptl, bdeduc_yrs)

Formularea ipotezelor:

H0: Între toleranța față de imigranți și anii de studiu nu există corelație.

H1: Între toleranța față de imigranți și anii de studiu există corelație.

```
> cor.test(bd$img_bptl, bd$educ_yrs)

        Pearson's product-moment correlation

data:  bd$img_bptl and bd$educ_yrs
t = 4.8681, df = 676, p-value = 1.404e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1102696 0.2557896
sample estimates:
      cor
0.1840379
```

P-value = 1.404e-06< 0.01, se respinge H0 cu un risc de 1%.

Interpretare: Putem afirma cu o probabilitate de 99% că între toleranța față de imigranți și anii de studiu există corelație.

5.2. Analiza de regresie

Modelul de regresie este cel mai sofisticat instrument statistic existent. Analiza de regresie presupune construirea unuia sau a mai multor modele de regresie care să aibă capacitatea de a determina factorii de influență și intensitatea pe care o au aceștia asupra unui anumit fenomen.

5.2.1. Regresie liniara simpla si multipla

 # Regresia liniară simplă

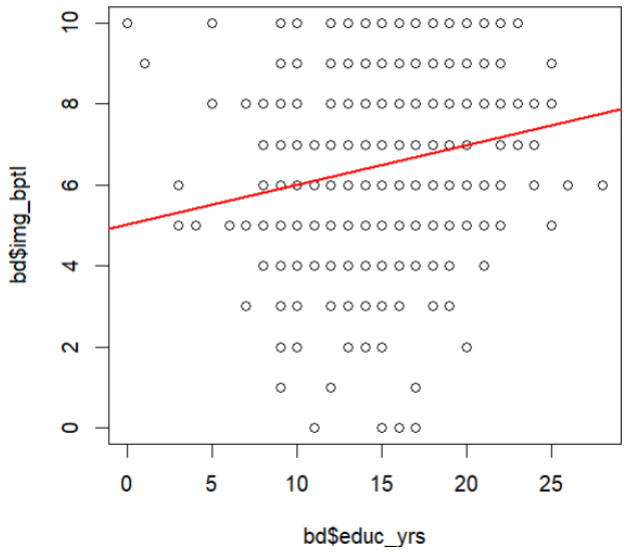
Deoarece putem garanta cu o probabilitate mai mare existența unei relații de corelație între toleranța față de imigranți și anii de studiu, comparativ cu relația dintre toleranță și vârstă, am ales să folosesc prima combinație de variabile în acest model.

Ecuația modelului la nivelul eșantionului este **Y = b0 + b1*x**, unde:

- **Y:** variabila dependentă (img_bptl).
- **b0:** constanta modelului, unul dintre coeficienții de regresie.
- **b1:** al doilea parametru al modelului (aferent variabile educ_yrs).
- **x:** variabila independentă (educ_yrs) .

```
simp_reg <- lm(bd$img_bptl ~ bd$educ_yrs)
plot(bd$img_bptl ~ bd$educ_yrs)
abline(simp_reg, col = 'red', lwd = 2)
```

În urma creări modelului, am construit scatterplotul al celor două variabile peste care am generat dreapta de regresia, pe care am colorat-o în roșu și i-am mărit grosimea pentru a o evidenția. Pe graficul de mai jos se poate observa că dreapta de regresie urmează un trend ascendent ceea ce sugerează că creșterea anilor de studiu duc la creșterea toleranței față de imigranți.



 # Estimarea modelului de regresie liniară simplă

summary(simp_reg)

Ecuația estimată a modelului: **Toleranța = 5.032 + 0.979 * Ani de studiu**

```
Call:
lm(formula = bd$img_bptl ~ bd$educ_yrs)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6966 -1.5008  0.0096  1.4013  4.9678

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.03222    0.31885  15.782  < 2e-16 ***
bd$educ_yrs  0.09791    0.02011   4.868  1.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.976 on 676 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.03387,   Adjusted R-squared:  0.03244
F-statistic:  23.7 on 1 and 676 DF,  p-value: 1.404e-06
```

Constanta modelului are o valoare estimată de 5.0322, ceea ce indică faptul că, în absența anilor de studiu, nivelul mediu al toleranței față de imigranți este de aproximativ 5,03 unități. Coeficientul asociat variabilei independente, anii de studiu, este estimat la 0.09791, ceea ce semnifică o creștere medie de 0.097 unități în nivelul toleranței față de imigranți pentru fiecare an suplimentar de studiu. Acest coeficient este pozitiv și semnificativ din punct de vedere statistic, având o valoare $p < 0.001$, ceea ce susține existența unei relații pozitive între educație și toleranța față de imigranți.

Indicatorii de ajustare ai modelului arată că puterea explicativă a acestuia este relativ redusă. Valoarea R-squared (0,03387) și R-squared ajustat (0,03244) sugerează că doar aproximativ 3,2% din variația toleranței față de imigranți poate fi explicată prin variația anilor de studiu. Deși această proporție este mică, semnificația statistică a coeficientului arată că relația este robustă la nivel de eșantion.

 # Regresia liniară multiplă

Pentru a construi un model de regresie multiplu am adăugat o nouă variabilă independentă modelului precedent și anume vârsta (*age*). Este de menționat că modelul multiplu cu trei parametri este imposibil de transpus într-un grafic bidimensional.

Ecuația modelului la nivelul eșantionului este **Y = b0 + b1*x1 + b2*x2**, unde:

- **Y:** variabila dependentă (img_bptl).
- **b0:** constanta modelului, unul dintre coeficienții de regresie.
- **b1:** al doilea parametru al modelului (aferent variabilei educ_yrs).
- **b2:** al treilea parametru al modelului (aferent variabilei age).
- **x1:** variabila independentă (educ_yrs).
- **x2:** variabila independentă (age).

```
mult_reg <- lm(bd$img_bptl ~ bd$educ_yrs + bd$age)
summary(mult_reg)
```

Ecuația estimată a modelului: **Toleranța = 5.774 + 0.911 * Ani de studiu -0.012 * Vârstă**

```
Call:
lm(formula = bd$img_bptl ~ bd$educ_yrs + bd$age)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7373 -1.4589  0.0046  1.3945  4.9144

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.774130   0.418032  13.813  < 2e-16 ***
bd$educ_yrs  0.091156   0.020170   4.519 7.32e-06 ***
bd$age      -0.012080   0.004431  -2.727  0.00657 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.967 on 675 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.04439,   Adjusted R-squared:  0.04156
F-statistic: 15.68 on 2 and 675 DF,  p-value: 2.208e-07
```

Constanta modelului are o valoare estimată de 5.7741, indicând nivelul mediu al toleranței față de imigranți pentru un respondent cu 0 ani de studiu și 0 ani. Această valoare, deși lipsită de

interpretare practică deoarece o persoană nu poate avea vârsta 0, oferă punctul de referință pentru interpretarea coeficienților. Coeficientul asociat anilor de studiu este estimat la 0.0912, ceea ce înseamnă că fiecare an suplimentar de studiu este asociat cu o creștere de 0,091 unități în nivelul toleranței față de imigranți, atunci când vârsta este 0, această interpretare este de asemenea lipsită de aplicabilitate practică. Acest coeficient a rămas pozitiv și semnificativ statistic ($p < 0.001$), confirmând influența pozitivă a educației asupra toleranței. În schimb, coeficientul asociat vârstei este de -0.0211, indicând o relație negativă între vârstă și toleranță. Mai precis, fiecare an în plus de vârstă este asociat cu o scădere de 0,021 unități în nivelul toleranței când persoana are 0 ani de studiu, iar această relație este semnificativă statistic ($p < 0,01$). R-pătratul modelului (0,0444) și R-pătratul ajustat (0,0416) indică faptul că aproximativ 4,2% din variația toleranței față de imigranți poate fi explicată prin variația anilor de studiu și a vârstei, ceea ce nu reprezintă o creștere mare față de modelul de regresie simplu.

5.2.2. Regresia neliniara

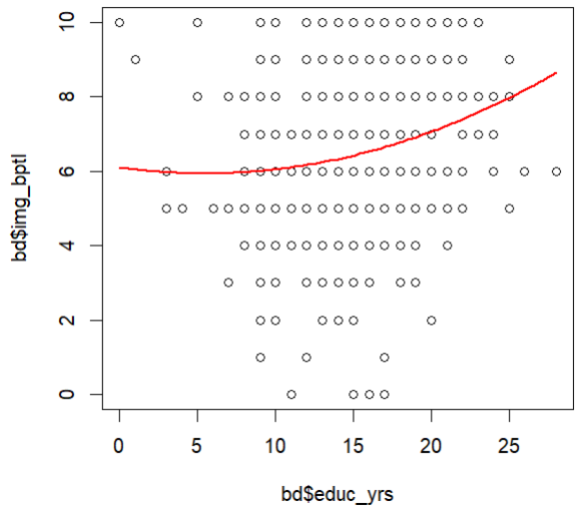
Modelul neliniar ales este unul parabolic iar variabila independentă este reprezentată de anii de educație din același motiv ca în cazul modelului liniar simplu.

Ecuația modelului la nivelul eșantionului este **Y = b0 + b1*x + b2*x^2**, unde:

- **Y:** variabila dependentă (img_bptl).
- **b0:** constanta modelului, unul dintre coeficienții de regresie.
- **b1:** al doilea parametru al modelului (aferent variabilei educ_yrs).
- **b2:** al treilea parametru al modelului (aferent variabilei educ_yrs^2).
- **x:** variabila independentă (educ_yrs).

```
nonl_reg <- lm(bd$img_bptl ~ bd$educ_yrs + I(bd$educ_yrs^2))
plot(bd$educ_yrs, bd$img_bptl)
curve(6.099205340 - 0.058222421 * x +  0.005319936 *x^2, add = T, col = 'red', lwd = 2)
```

În urma creări modelului, am construit scatterplotul al celor două variabile peste care am generat curba de regresie, pe care am colorat-o în roșu și i-am mărit grosimea pentru a o evidenția. Pe graficul de mai jos se poate observa că dreapta de regresie urmează un trend ascendent ceea ce sugerează că creșterea anilor de studiu duc la creșterea toleranței față de imigranți. Se poate observa de asemenea un punct de minim ceea ce subliniază existența unui interval a anilor de studiu în care toleranța scade, nu crește.



#Estimarea modelului de regresie neliniar

summary(nonl_reg)

Ecuația estimată a modelului: **Toleranța = 6.099 -0.058 * Ani de studiu +0.005 * Ani de studiu^2**

```
Call:
lm(formula = bd$img_bptl ~ bd$educ_yrs + I(bd$educ_yrs^2))

Residuals:
    Min      1Q  Median      3Q     Max
-6.6469 -1.4229 -0.0559  1.4704  4.0589

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.099205   0.731963   8.333 4.42e-16 ***
bd$educ_yrs    -0.058222   0.098507  -0.591    0.555
I(bd$educ_yrs^2)  0.005320   0.003286   1.619    0.106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.974 on 675 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.03761,   Adjusted R-squared:  0.03476
F-statistic: 13.19 on 2 and 675 DF,  p-value: 2.407e-06
```

Din tabelul coeficienților, constanta(6.099205) este semnificativ statistic ($p < 0.001$), ceea ce indică faptul că, atunci când valorile variabilelor independente sunt egale cu zero, toleranța față de imigranți are o valoare medie de aproximativ 6.09. Totuși, coeficienții pentru variabila „bd$educ_yrs” (-0.05822) și pentru „bd$educ_yrs^2” (0.00532) nu sunt semnificativi din punct de vedere statistic ($p > 0.05$). Acest lucru sugerează că nici nivelul educațional, măsurat prin numărul de ani de studiu, și nici componenta sa non-lineară , nu au un efect semnificativ asupra toleranței față de imigranți, în cadrul acestui model. În plus, valoarea R-pătrat (0.03761) și R-pătrat ajustat

(0.03476) arată că doar aproximativ 3.4-3.8% din variația toleranței față de imigranți poate fi explicată de acest model.

5.2.3. Compararea a doua modele de regresie si alegerea celui mai bun model

Formularea ipotezelor

H0: modelul *nonl_reg* (cu mai multi parametri) nu este semnificativ mai bun decât modelul *simp_reg.*
H1: modelul *nonl_reg* (cu mai multi parametri) este semnificativ mai bun decât modelul *simp_reg.*

anova(simp_reg, nonl_reg)

```
Analysis of Variance Table

Model 1: bd$img_bptl ~ bd$educ_yrs
Model 2: bd$img_bptl ~ bd$educ_yrs + I(bd$educ_yrs^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    676 2639.9
2    675 2629.7  1    10.211 2.6211 0.1059
```

P-value = 0.1059 > 0.05, cu o probabilitate de 95% nu se refuză ipoteza H0.

Interpretare: Putem garanta cu un risc de 5% că modelul mai complex (*nonl-reg*) nu este semnificativ mai bun decât modelul mai simplu (*sim_reg*).

6. Estimarea si testarea mediilor

6.1. Estimarea mediei prin interval de încredere

t.test(bd$educ_yrs)

```
> t.test(bd$educ_yrs)

        One Sample t-test

data:  bd$educ_yrs
t = 106.4, df = 680, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 15.11379 15.68210
sample estimates:
mean of x
 15.39794
```

Interpretare: Cu o încredere de 95%, putem afirma că anii medii de studii sunt acoperiți de intervalul [15.113, 15.682].

6.2. Testarea mediilor populației

6.2.1. Testarea unei medii cu o valoare fixa.

Formularea ipotezelor:

H0: numărul mediu de ani de educație nu este semnificativ mai mare de 12.

H1: numărul mediu de ani de educație este semnificativ mai mare de 12.

t.test(bd$educ_yrs, mu = 10, alternative = 'greater')

```
> t.test(bd$educ_yrs, mu = 12, alternative = 'greater')

        One Sample t-test

data:  bd$educ_yrs
t = 23.479, df = 680, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 12
95 percent confidence interval:
 15.15957      Inf
sample estimates:
mean of x
 15.39794
```

P-value < 2.2e-16, se respinge ipoteza nula, cu un risc de 5%.

Interpretare: Putem afirma, cu un risc asumat de 5%, că numărul mediu a anilor de studiu este semnificativ mai mare de 12, ceea ce înseamnă că în medie, respondenții au făcut studii universitare.

6.2.2. Testarea diferenței dintre doua medii (cu eșantioane independente)

#Testul de omogenitate a variațelor

Formularea ipotezelor:

H0: Variațele grupurilor celor pro-aderare și a celor anti-aderare nu diferă semnificativ.

H1: Variațele grupurilor celor pro-aderare și a celor anti-aderare diferă semnificativ.

bartlett.test(educ_yrs ~ eu_vote, data = bd, eu_vote %in% c('Join EU', 'Remain Outside'))

```
        Bartlett test of homogeneity of variances

data:  educ_yrs by eu_vote
Bartlett's K-squared = 6.1034, df = 1, p-value = 0.01349
```

p-value = 0.01349 < 0.05 se refuza H0 cu un risc de 5%.

Interpretare: Putem garanta cu un risc de 5% că varianța grupului de persoane care au votat că ar vrea să intre în UE și varianța grupului de persoane care au votat că ar dori să rămână în afara UE diferă semnificativ.

 # Testarea diferențelor dintre medii

Formularea ipotezelor:

H0: mediile grupurilor celor pro-aderare și a celor anti-aderare nu diferă semnificativ.

H1: mediile grupurilor celor pro-aderare și a celor anti-aderare diferă semnificativ.

t.test(educ_yrs ~ eu_vote, bd, eu_vote %in% c('Join EU', 'Remain Outside'), var.equal = F)

```
        Welch Two Sample t-test

data:  educ_yrs by eu_vote
t = 3.1771, df = 436.95, p-value = 0.001593
alternative hypothesis: true difference in means between grou
p Join EU and group Remain Outside is not equal to 0
95 percent confidence interval:
 0.3662449 1.5544063
sample estimates:
       mean in group Join EU mean in group Remain Outside
                    16.03030                      15.06998
```

P-value = 0.001593 < 0.05, se respinge ipoteza H0 cu un risc asumat de 5%.

Interpretare: Putem garanta, cu un risc asumat de 5%, că anii medii de studiu a grupului celor pro-aderare variază semnificativ de anii medii de studiu a celor anti-aderare.

6.2.3. Testarea diferenței dintre trei sau mai multe medii

Formularea ipotezelor:

H0: nu exista nici o diferența intre anii medii de studiu în funcție de nivelul de toleranță.
H1: exista cel puțin 2 nivele de toleranță a căror ani medii de studiu diferă semnificativ.

aov <- aov(educ_yrs ~ img_bptl_fact, data = bd)

anova(aov)

```
Analysis of Variance Table

Response: educ_yrs
              Df Sum Sq Mean Sq F value    Pr(>F)
img_bptl_fact   2  260.0 129.988  9.3397 9.976e-05 ***
Residuals     675 9394.5  13.918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value = 9.976e-05, se respinge cu un risc de 5% ipoteza H0.

Interpretare: Putem garanta, cu un risc de 5%, că exista cel puțin 2 nivele de toleranță a căror ani medii de studiu diferă semnificativ.

7. Concluzii

Acest studiu are la bază un set de date ce contorizează răspunsurile unui eșantion de 1411 cetățeni norvegieni, la un chestionar ce vizează deschiderea acestora față de aderarea Norvegiei la Uniunea Europeană. Acest chestionar a fost realizat de European Social Survey în anul 2022 și vizează atât problematica de aderare a Norvegiei, cât și anumite problematici adiacente cum ar fi toleranța față de imigrare și nivelul de educație. Baza de date, cu 1411 observații inițiale și 25 de variabile, a fost supusă unui proces riguros de preprocesare. Acest proces a inclus selecția și transformarea variabilelor, reducând setul la cinci variabile relevante pentru studiu. Variabilele numerice alese (vârsta, anii de educație și scorul toleranței față de imigranți) reflectă caracteristicile socio-demografice, iar cele categoriale (opțiunea de vot privind UE și nivelul de toleranță față de imigranți) au fost selectate pentru analiza relațiilor interdependente.

Ulterior am efectuat analiza descriptivă a variabilelor numerice și nenumerice selectate, utilizând indicatori statistici și reprezentări grafice. Rezultatele arată că distribuțiile variabilelor numerice sunt caracterizate de diverse grade de asimetrie și dispersie. Variabila „img_bptl” (percepția impactului imigrației) prezintă o distribuție ușor asimetrică, indicând o percepție predominant favorabilă, dar cu variații semnificative în opinii. Vârsta respondenților are o distribuție concentrată în jurul intervalului 50–60 de ani, reflectând participarea predominantă a adulților maturi, iar anii de educație demonstrează un nivel ridicat de acces la educație în rândul populației norvegiene. Analiza grafică a variabilelor categoriale evidențiază opoziția majoritară față de aderarea la UE și niveluri semnificative de toleranță față de imigranți, cu o predominanță a categoriilor „High Tolerance” și „Medium Tolerance.” Aceste observații oferă o bază solidă pentru explorările inferențiale ulterioare.

În continuare, studiul s-a axat pe explorarea relațiilor dintre variabilele categoriale utilizând metode statistice precum tabelarea frecvențelor și analiza de asociere. Rezultatele indică o asociere semnificativă între atitudinea față de imigranți și opțiunea privind aderarea Norvegiei la Uniunea Europeană. Persoanele cu un nivel mai ridicat de toleranță față de imigranți prezintă o probabilitate mai mare de a susține aderarea la UE, sugerând existența unor legături între deschiderea culturală și percepțiile politice. Totodată, analiza de concordanță confirmă că distribuțiile variabilelor img_bptl_fact și eu_vote nu sunt egalitare, reflectând opinii predominant distincte în rândul respondenților.

Capitolul 5 reprezintă punctul de început al analizei inferențiale, mai specific pe identificarea factorilor determinanți ai toleranței față de imigranți, utilizând modele de regresie

liniară și neliniară. Rezultatele arată că anii de studiu influențează pozitiv nivelul toleranței, în timp ce vârsta exercită un efect negativ moderat. Deși semnificația statistică a coeficienților confirmă validitatea relațiilor, puterea explicativă a modelelor este limitată, sugerând că toleranța față de imigranți este influențată și de alți factori neincluși în analiză. Modelul neliniar indică o posibilă relație complexă între educație și toleranță, dar coeficienții săi nu sunt semnificativi. Testele ipotezelor privind erorile modelelor arată că modelul liniar simplu și cel neliniar îndeplinesc majoritatea criteriilor de validitate, deși modelul multiplu oferă o ajustare ușor mai bună.

Capitolul 6 se axează pe mediile eșantionului. Rezultatele arată că media anilor de educație în rândul respondenților este semnificativ mai mare decât pragul de 12 ani, indicând un nivel educațional ridicat, specific contextului norvegian. Diferențele semnificative între mediile grupurilor pro-aderare și anti-aderare la UE, precum și între nivelurile de toleranță față de imigranți, confirmă influența variabilelor socio-demografice asupra poziționărilor politice și culturale. Aceste descoperiri subliniază importanța educației în formarea opiniilor și evidențiază variațiile atitudinilor în funcție de nivelurile de toleranță.