

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

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

**COORDONATOR ȘTIINȚIFIC:  
PROF. DR. ASANDULUI MIRCEA**

**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 toleranță 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 prezenta 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 nenumerate din baza de date rezultată după operațiunile preliminare. Al patrulea capitol marchează începutul analizei inferențiale și se axează pe analiza variabilelor categoricale prin procedee de tabelare a datelor și realizarea analizelor de asociere și concordanță. Următorul capitol se va concentra pe analiza mediilor 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. Ultimul capitol va avea ca punct central crearea modelelor de regresie menite să identifice factorii ce influențează atitudinea față de imigranți.

## 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

```
rownames      int64
cntry          object
idno          int64
region        object
inwds         object
inwde         object
dweight       float64
pspwght       float64
pweight       float64
anweight      float64
prob          float64
stratum       int64
psu           int64
eu_vote       object
brnnorge      int64
agea          int64
imbgeco       float64
imueclt       float64
imwbcnt       float64
female        int64
eduyrs        float64
uempla        int64
polint        int64
hinctnta     float64
lrscale       float64
```

```
In [21]: len(bd)
Out[21]: 1411

In [22]: len(bd.columns)
Out[22]: 25
```

Cu ajutorul funcției „dtypes” pot ilustra clasele fiecărei variabile iar cu ajutorul funcției „len()” pot măsura numărul de observații și de coloane. Baza de date are 1411 observații și 25 de variabile dintre 5 de tip object iar 20 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 în care imigranții influențează societatea. Variabila *imbgeco* arată dacă respondentul consideră că imigranții influențează în bine sau în rău 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 attribute 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 în care a început interviul (*inwds*) și momentul în care s-a sfârșit interviul (*inwde*)

### # Redenumirea variabilelor

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.

```
Index(['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'],  
      dtype='object')
```

### # Crearea unei variabile categoriale "img\_bptl\_fact" bazate pe variabila numerică "img\_bptl"

Am creat o variabilă categorială, cu ajutorul funcției „cut()” conform cerințelor 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 să fie folosite în proiect având în vedere numărul de valori lipsa

rownames	0
country	0
id	0
region	0
int_start	0
int_end	0
dweight	0
pspwght	0
pweight	0
anweight	0
prob	0
stratum	0
psu	0
eu_vote	0
born_NO	0
age	0
img_econ	17
img_culture	9
img_bptl	11
female	0
educ_yrs	18
unemployed	0
pol_interest	0
household_income	64
lrscale	36
img_bptl_fact	18

În urma aplicării funcției pentru observarea numărului de valori lipsă am decis sa folosesc 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):** 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 nivele 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.

**# Din baza inițială se va face o selecție care să includă condiții pentru cel puțin două variabile. Cerințele proiectului vor fi executate pentru această selecție**

Am redefinit obiectul „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.

**# Definirea categoriilor variabilelor categoriale.**

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 „Categorical()” am transformat variabila în tipul categorial iar aceasta a fost împărțită automat în 7 categorii. De asemenea, variabila *img\_bptl\_fact* este deja categoriile definite.

**# Prezentarea bazei de date finale**

<i>img_bptl</i>	float64	In [47]: len(bd)
<i>age</i>	int64	Out[47]: 683
<i>educ_yrs</i>	float64	
<i>img_bptl_fact</i>	category	In [48]: len(bd.columns)
<i>eu_vote</i>	category	Out[48]: 5
dtype: object		

```
In [49]: bd['eu_vote'].cat.categories
Out[49]:
Index(['Blank Ballot', 'Don't Know', 'Join EU', 'Not Eligible',
      'Refuse to Answer', 'Remain Outside', 'Wouldn't Vote'],
      dtype='object')

In [50]: bd['img_bptl_fact'].cat.categories
Out[50]: Index(['Low Tolerance', 'Medium Tolerance', 'High Tolerance'], dtype='object')
```

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 efectuării 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

Cu ajutorul funcției „discribe()” putem analiza variabile de mai multe tipuri în același timp atât timp cât argumentul „include” este setat la valoarea „all”

#### 3.1. Analiza descriptiva numerica a variabilelor numerice si nenumarice

##### # Analiza indicatorilor descriptivi

	img_bptl	age	educ_yrs	img_bptl_fact	eu_vote
count	680.000000	683.000000	681.000000	680	683
unique	NaN	NaN	NaN	3	7
top	NaN	NaN	NaN	High Tolerance	Remain Outside
freq	NaN	NaN	NaN	345	444
mean	6.541176	52.784773	15.397944	NaN	NaN
std	2.006193	17.190129	3.776700	NaN	NaN
min	0.000000	16.000000	0.000000	NaN	NaN
25%	5.000000	40.000000	13.000000	NaN	NaN
50%	7.000000	55.000000	16.000000	NaN	NaN
75%	8.000000	67.000000	18.000000	NaN	NaN
max	10.000000	90.000000	28.000000	NaN	NaN

La nivelul celor 5 variabile se pot observa 8 valori lipsă și nici o eroare evidentă. Variabila categorială *img\_bptl\_fact* are 3 și 680 de observații iar variabila *eu\_vote* are 7 nivele și 683 observații. De asemenea indicatorii descriptivi ai variabilelor numerice 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.

**Std:** Valorile se abat, de la valoarea medie cu 2.006 puncte.

##### ➤ **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.

**Std:** Vârsta se abate de la valoarea medie cu 17.19 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.

**Std:** Valorile anilor de studiu se abat de la medie cu 3.77 ani.

Suplimentar, pot fi analizați indicatorii de asimetrie și boltire pentru a crea o imagine de ansamblu asupra distribuțiilor celor 3 variabile.

```
In [60]: skew(bd['age'])
Out[60]: -0.24198339164335583

In [61]: skew(bd['img_bptl'], nan_policy='omit')
Out[61]: -0.26514766608203616

In [62]: skew(bd['educ_yrs'], nan_policy='omit')
Out[62]: -0.3846880507451694

In [63]: kurtosis(bd['age'])
Out[63]: -0.7964322817051905

In [64]: kurtosis(bd['img_bptl'], nan_policy='omit')
Out[64]: 0.0716466603078052

In [65]: kurtosis(bd['educ_yrs'], nan_policy='omit')
Out[65]: 0.7609268440914954
```

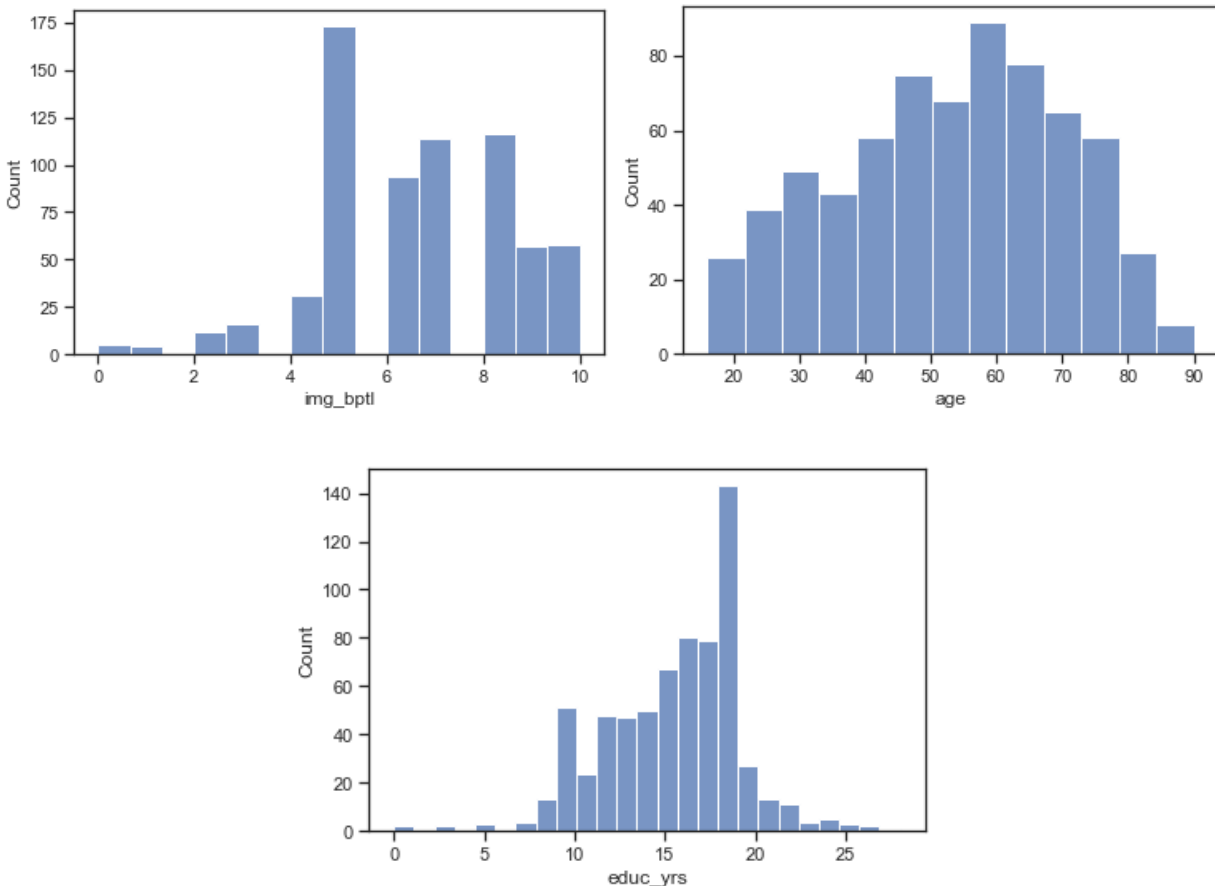
- **skew:** Toate cele 3 distribuții ale variabilelor sunt asimetrice la stânga întrucât indicatorul de asimetrie este mai mic decât 0.



- **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.

### 3.2. Analiza grafica a variabilelor numerice si nenumerice

Pentru a putea analiza grafic variabilele numerice mă voi folosi de histogramme, 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.

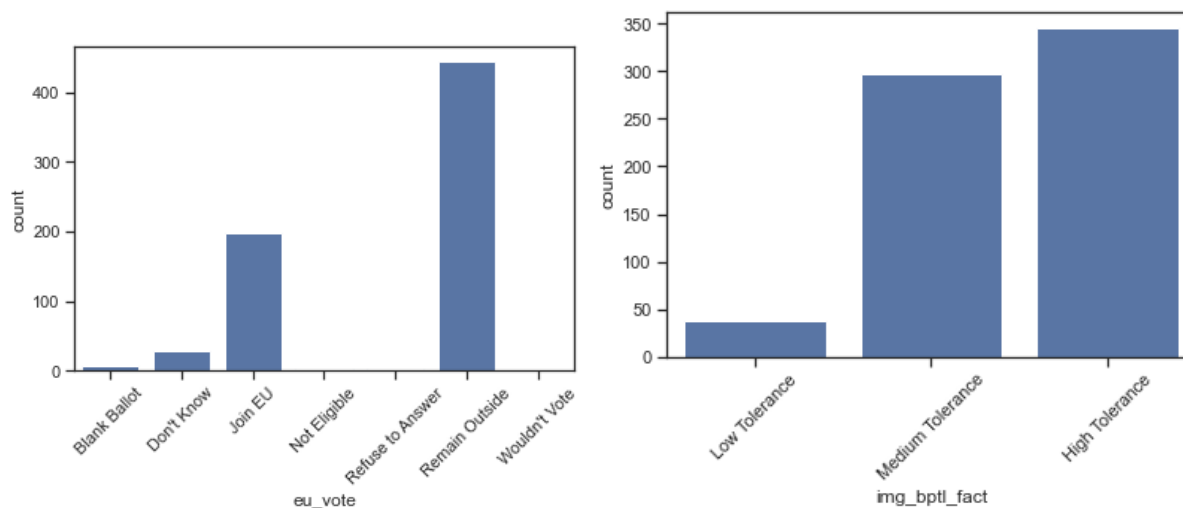
Histograma variabilei *educ\_yrs*, care măsoară anii de educație, arată o distribuție aproximativ simetrică, cu un vârf în jurul valorii de 15-17 ani de studiu. Acest model este coerent cu un nivel de educație ridicat al populației norvegiene, reflectând atât accesul larg la educație,

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

Pentru a putea construi barplot-urile variabilelor categoriale am apelat la pachetele „seaborn” și „matplotlib.pyplot”.

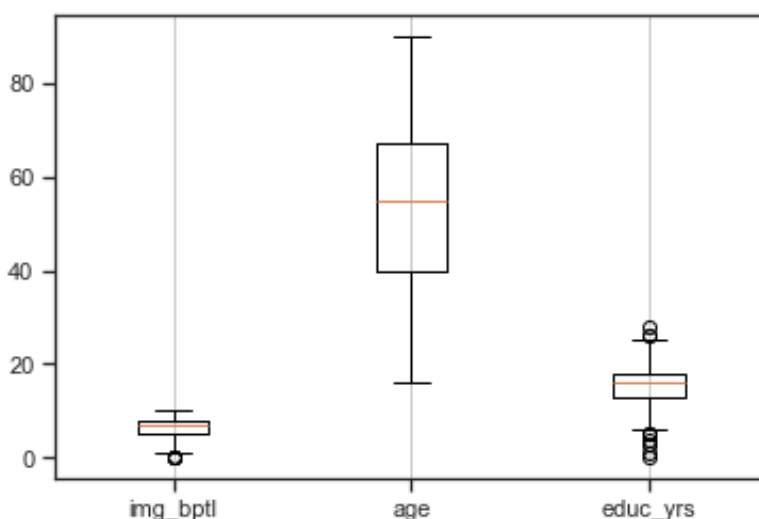


În primul barplot, care analizează variabila *eu\_vote*, observăm o distribuție clară a preferințelor respondenților. Cea mai mare proporție dintre aceștia susține menținerea Norvegiei în afara Uniunii Europene, această opțiune având un număr semnificativ mai mare de respondenți comparativ cu celelalte 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)



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 intercuartil 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 procedurii 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 frecvențe marginale, condiționate, parțiale)

#### # Frecvența marginală

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*.

```
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
dtype: int64
```

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 frecvențele variabilei *img\_bptl*.

```
img_bptl_fact
Low Tolerance      37
Medium Tolerance   298
High Tolerance     345
dtype: int64
```

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

img_bptl_fact	Low Tolerance	Medium Tolerance	High Tolerance
eu_vote			
Blank Ballot	0.000000	1.006711	1.159420
Don't Know	2.702703	2.684564	5.507246
Join EU	8.108108	26.510067	33.043478
Not Eligible	0.000000	1.006711	0.000000
Refuse to Answer	0.000000	0.000000	0.579710
Remain Outside	89.189189	68.791946	59.420290
Wouldn't Vote	0.000000	0.000000	0.289855

O primă observație se referă la susținerea pentru opțiunea de „Rămânere în afara UE” („Remain Outside”), care prezintă frecvențe extrem de ridicate, în special în rândul celor cu toleranță scăzută față de imigranți (89.19%). Această proporție scade semnificativ în cazul celor cu toleranță medie (68.79%) și ridicată (59.42%), sugerând o corelație negativă între toleranța față de imigranți și opoziția față de aderarea la UE. Pe de altă parte, opțiunea „Aderare la UE” („Join EU”) câștigă mai mult sprijin pe măsură ce nivelul de toleranță față de imigranți crește. Frecvența acestui răspuns este de doar 8.11% în grupul cu toleranță scăzută, dar crește semnificativ la 26.51% pentru cei cu toleranță medie și ajunge la 33.04% în cazul celor cu toleranță ridicată. Acest tipar sugerează o asociere pozitivă între deschiderea față de imigrație și susținerea pentru integrarea europeană.

Un alt aspect interesant este proporția celor care „Nu știu” („Don't Know”), care rămâne relativ constantă în rândul celor cu toleranță scăzută și medie (aproximativ 2.7%), dar aproape se dublează pentru respondenții cu toleranță ridicată (5.51%). Aceasta poate indica o incertitudine mai mare în rândul celor cu o atitudine deschisă față de imigrație, posibil din cauza percepțiilor

complexe legate de beneficiile și riscurile aderării. Opțiunea „Blank Ballot” și răspunsurile precum „Not Eligible” sau „Refuse to Answer” au frecvențe scăzute, dar ele par să fie ușor mai frecvente în rândul celor cu toleranță medie sau ridicată.

### # Frecvențe parțiale

În cadrul tabelului anterior se puteau observa anumite celule cu valori doar pe o anumită coloană, celelalte coloane având valoarea 0. 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.

```
In [115]: tabel.loc["Wouldn't Vote", 'High Tolerance']
Out[115]: 1

In [116]: tabel.loc['Refuse to Answer', 'High Tolerance']
Out[116]: 2

In [117]: tabel.loc['Not Eligible', 'Medium Tolerance']
Out[117]: 3
```

În urma selectării celulelor care indică un procentaj atipic 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 „chi2\_contingency()” care efectuează un test Chi-square.

### 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 nu sunt independente).

```
Chi2ContingencyResult(statistic=24.30260193615078, pvalue=0.018496214651125675, dof=12,
expected_freq=array([[3.80882353e-01, 3.06764706e+00, 3.55147059e+00],
[1.52352941e+00, 1.22705882e+01, 1.42058824e+01],
[1.06647059e+01, 8.58941176e+01, 9.94411765e+01],
[1.63235294e-01, 1.31470588e+00, 1.52205882e+00],
[1.08823529e-01, 8.76470588e-01, 1.01470588e+00],
[2.41044118e+01, 1.94138235e+02, 2.24757353e+02],
[5.44117647e-02, 4.38235294e-01, 5.07352941e-01]]))
```

**P-value = 0.0185** < 0.05 se respinge H0

**Interpretare:** Cu un risc asumat de 5%, putem afirma că între poziționarea norvegienilor față de imigrare și modul în 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 egale.

#### Formularea ipotezelor:

**H0:** distribuția variabilei *img\_bptl\_fact* este egalitară.

**H1:** distribuția variabilei *img\_bptl\_fact* nu este egalitară.

```
chisquare(observe, np.sum(observe)/np.sum(asteptate) * asteptate)  
Power_divergenceResult(statistic=242.93235294117648, pvalue=1.7697414422601153e-53)
```

**P-value = 1.76e-53** < 0.01, se respinge H0.

**Interpretare:** Cu un risc asumat de 1% putem afirma că distribuția variabilei *img\_bptl\_fact* nu este egalitară.

#### Formularea ipotezelor:

**H0:** distribuția variabilei *eu\_vote* este egalitară.

**H1:** distribuția variabilei *eu\_vote* nu este egalitară.

```
chisquare(observe2, np.sum(observe2)/np.sum(asteptate2) * asteptate2)  
Power_divergenceResult(statistic=1747.9062957540264, pvalue=0.0)
```

**P-value = 0**, se respinge H0

**Interpretare:** Cu un risc asumat de 1% putem afirma că distribuția variabilei *eu\_vote* nu este egalitară.

## 5. Estimarea și testarea mediilor

Acest capitol se concentrează atât pe relevanța anilor de studiu din punct de vedere social, cât și pe diferențele între diversele grupuri de respondenți printr-o prismă a anilor medii de studiu.

### 5.1. Estimarea mediei prin interval de încredere

```
In [123]: IC(educ_yrs, 0.95)  
Out[123]: (15.113785627119892, 15.682102772292737)
```

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

## 5.2. Testarea mediilor populației

### 5.2.1. 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.

```
...: stats.ttest_1samp(educ_yrs, 12, alternative= 'greater')
Out[199]: TtestResult(statistic=23.478873592143305,
pvalue=4.927408677743371e-90, df=680)
```

**P-value = 4.92e-90**, se respinge ipoteza nula.

**Interpretare:** Putem afirma, cu un risc asumat de 1%, că numărul mediu a anilor de studiu este semnificativ mai mare de 12, ceea ce înseamnă că în medie, respondenții au studii superioare.

### 5.2.2. Testarea diferenței dintre doua medii (cu eșantioane independente)

#### # Testul de omogenitate a varianțelor

##### Formularea ipotezelor:

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

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

```
In [127]: levene(Join_EU, Remain_Outside)
Out[127]: LeveneResult(statistic=3.123707817830356, pvalue=0.07763825379918621)
```

**p-value = 0.0776** > 0.05 se refuza H0.

**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 nu 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.

```
In [128]: stats.ttest_ind(Join_EU, Remain_Outside, equal_var= True)
Out[128]: TtestResult(statistic=2.9975445559431226, pvalue=0.0028272702756926165, df=639.0)
```

**P-value = 0.0028** < 0.05, se respinge ipoteza H0.



**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.

### 5.2.3. Testarea diferenței dintre trei sau mai multe medii

#### Formularea ipotezelor:

**H0:** Nu exista nici o diferență între 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.

	df	sum_sq	mean_sq	F	PR(>F)
img_bptl_fact	2.0	259.975089	129.987545	9.339674	0.0001
Residual	675.0	9394.502787	13.917782	NaN	NaN

**p-value = 0.0001**, se respinge ipoteza H0.

**Interpretare:** Putem garanta, cu un risc de 1%, că exista cel puțin 2 nivele de toleranță a căror ani medii de studiu diferă semnificativ.

## 6. Analiza de regresie si corelație

### 6.1. Analiza de corelație

#### # Matricea de corelație

Pentru a analiza corelația dintre variabilele numerice se va folosi funcția „corr()” cu ajutorul căreia poate fi construită matricea de corelație. Având în vedere faptul că relația de corelație se poate stabili doar între variabile numerice, am creat un obiect nou denumit „bd\_num” care conține doar variabilele numerice și pe baza căruia se va construi matricea.

	img_bptl	age	educ_yrs
img_bptl	1.000000	-0.124276	0.184038
age	-0.124276	1.000000	-0.116874
educ_yrs	0.184038	-0.116874	1.000000

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 variabile, *age* și *educ\_yrs*.

Având în vedere imposibilitatea folosirii variabilelor ce prezintă valori NaN, voi crea o bază de date care nu conține observațiile cu NaN-uri.

#### Formularea ipotezelor:

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

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

```
In [136]: bd_clean = bd.dropna()
...: stats.pearsonr(bd_clean['img_bptl'], bd_clean['age'])
Out[136]: PearsonRResult(statistic=-0.12440878189114188, pvalue=0.0011701776347102157)
```

**p-value = 0.00117** < 0.05, se respinge H0.

**Interpretare:** Putem afirma cu un risc de 5% că între toleranța față de imigranți și vârstă există corelație.

#### 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.

```
In [137]: stats.pearsonr(bd_clean['img_bptl'], bd_clean['educ_yrs'])
Out[137]: PearsonRResult(statistic=0.1840378564276139, pvalue=1.4036734366748627e-06)
```

**P-value = 1.404e-06** < 0.01, se respinge H0.

**Interpretare:** Putem afirma, cu un risc asumat de 1%, că între toleranța față de imigranți și anii de studiu există corelație.

## 6.2. 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.

### 6.2.1. Regresie liniară simplă și multiplă

#### # Regresia liniară simplă

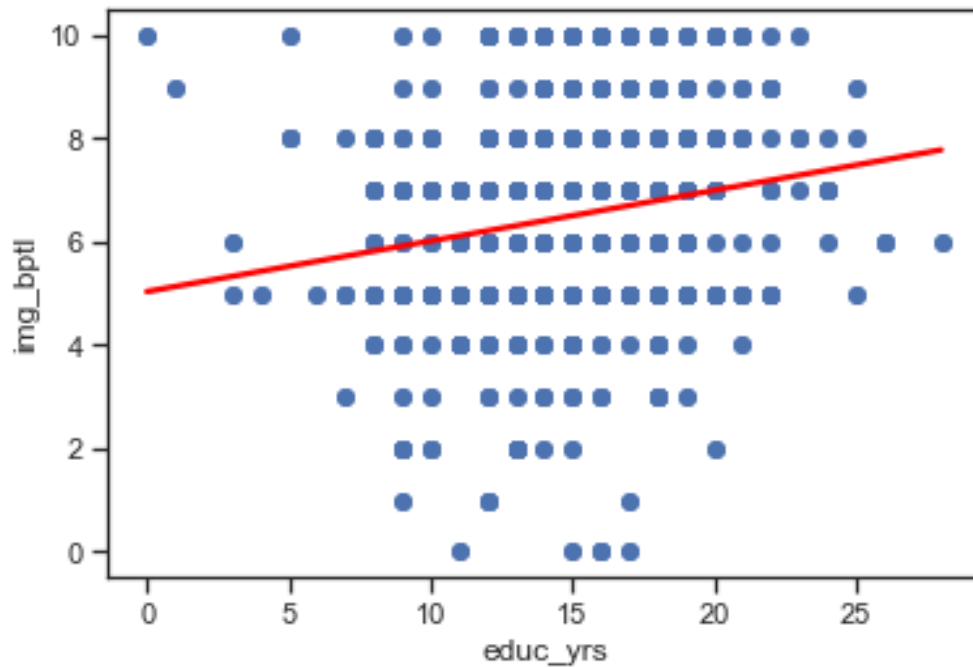
Deoarece putem garanta cu o probabilitate mai mare existența unei relații de corelație între toleranța față de imigranți și anii de studiu, comparativ cu relația dintre toleranță și vârstă, am ales să folosesc prima combinație de variabile în acest model.

Ecuția modelului la nivelul eșantionului este  $Y = b_0 + b_1 \cdot x$ , unde:

- **Y:** variabila dependentă (*img\_bptl*)
- **b<sub>0</sub>:** constanta modelului, unul dintre coeficienții de regresie
- **b<sub>1</sub>:** al doilea parametru al modelului (aferent variabile *educ\_yrs*)

- **x:** variabila independentă (*educ\_yrs*)

În urma creării modelului, am construit scatterplotul celor două variabile peste care am generat dreapta de regresia, pe care am colorat-o în roșu 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

Ecuatia estimată a modelului: **Toleranta = 5.032 + 0.979 \* Ani de studiu**

OLS Regression Results						
=====						
Dep. Variable:	img_bptl		R-squared:	0.034		
Model:	OLS		Adj. R-squared:	0.032		
Method:	Least Squares		F-statistic:	23.70		
Date:	Sat, 04 Jan 2025		Prob (F-statistic):	1.40e-06		
Time:	17:30:13		Log-Likelihood:	-1422.9		
No. Observations:	678		AIC:	2850.		
Df Residuals:	676		BIC:	2859.		
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	5.0322	0.319	15.782	0.000	4.406	5.658
educ_yrs	0.0979	0.020	4.868	0.000	0.058	0.137
=====						
Omnibus:		5.714	Durbin-Watson:		2.025	
Prob(Omnibus):		0.057	Jarque-Bera (JB):		5.585	
Skew:		-0.218	Prob(JB):		0.0613	
Kurtosis:		3.089	Cond. No.		66.9	
=====						

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 0.0979, 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 (0,034) și R-squared ajustat (0,032) 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.

Ecuția modelului la nivelul eșantionului este  $Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$ , 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 *age*)
- **x1**: variabila independentă (*educ\_yrs*)
- **x2**: variabila independentă (*age*)

### # Estimarea modelului de regresie liniară multiplă

Ecuția estimată a modelului: **Toleranța = 5.774 + 0.911 \* Ani de studiu -0.012 \* Vârsta**

OLS Regression Results						
=====						
Dep. Variable:	img_bpt1		R-squared:	0.044		
Model:	OLS		Adj. R-squared:	0.042		
Method:	Least Squares		F-statistic:	15.68		
Date:	Sat, 04 Jan 2025		Prob (F-statistic):	2.21e-07		
Time:	17:33:27		Log-Likelihood:	-1419.1		
No. Observations:	678		AIC:	2844.		
Df Residuals:	675		BIC:	2858.		
Df Model:	2					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	5.7741	0.418	13.813	0.000	4.953	6.595
educ_yrs	0.0912	0.020	4.519	0.000	0.052	0.131
age	-0.0121	0.004	-2.727	0.007	-0.021	-0.003
=====						
Omnibus:	6.923		Durbin-Watson:	2.039		
Prob(Omnibus):	0.031		Jarque-Bera (JB):	6.817		
Skew:	-0.240		Prob(JB):	0.0331		
Kurtosis:	3.101		Cond. No.	318.		
=====						

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 -0.0121, 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 (0.044) și R-pătratul ajustat (0.042) 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.

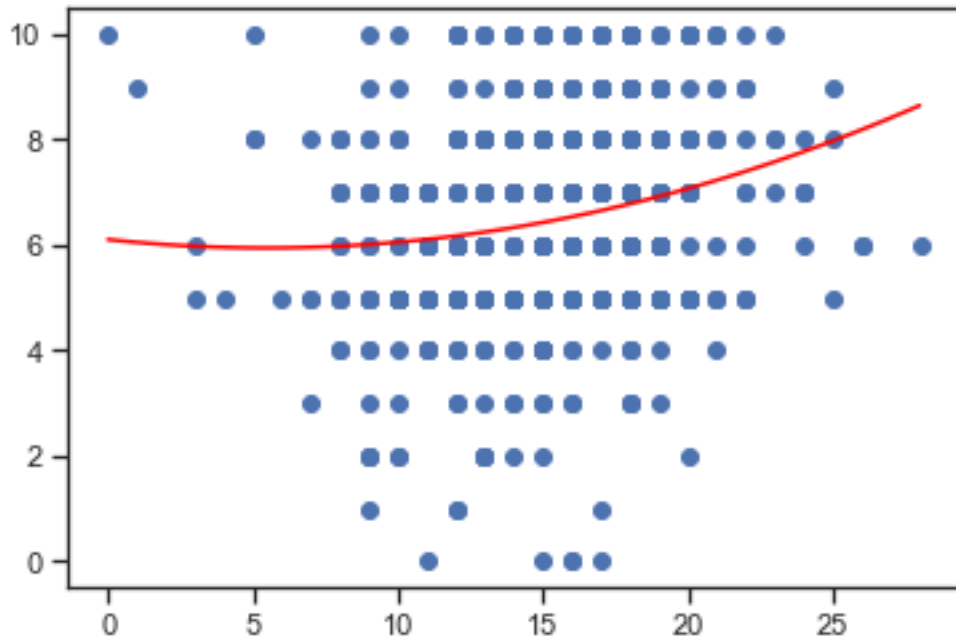
### 6.2.2. Regresia neliniara

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

Ecuția modelului la nivelul eșantionului este  $Y = b_0 + b_1 \cdot x + b_2 \cdot x^2$ , 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*)

În urma creării modelului, am construit scatterplotul al celor două variabile peste care am generat curba de regresie, pe care am colorat-o în roșu 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

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

OLS Regression Results						
Dep. Variable:	img_bptl	R-squared:	0.038			
Model:	OLS	Adj. R-squared:	0.035			
Method:	Least Squares	F-statistic:	13.19			
Date:	Sat, 04 Jan 2025	Prob (F-statistic):	2.41e-06			
Time:	17:42:18	Log-Likelihood:	-1421.5			
No. Observations:	678	AIC:	2849.			
Df Residuals:	675	BIC:	2863.			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	6.0992	0.732	8.333	0.000	4.662	7.536
educ_yrs	-0.0582	0.099	-0.591	0.555	-0.252	0.135
educ_yrs_sq	0.0053	0.003	1.619	0.106	-0.001	0.012
Omnibus:	5.807	Durbin-Watson:	2.029			
Prob(Omnibus):	0.055	Jarque-Bera (JB):	5.741			
Skew:	-0.224	Prob(JB):	0.0567			
Kurtosis:	3.041	Cond. No.	2.69e+03			

Din tabelul coeficienților, constanta (6.0992) 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 „educ\_yrs” (-0.0582) și pentru „educ\_yrs^2” (0.0053) 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 (0.038) și R-pătrat ajustat (0.035) arată că doar aproximativ 3.5-3.8% din variația toleranței față de imigranți poate fi explicată de acest model.

### 6.2.3 Testarea ipotezelor ce vizează erorile de regresie

#### # Media erorilor trebuie să fie egală cu 0

##### Formularea ipotezelor:

**H0:** Media erorilor modelului de regresie este egală cu 0.

**H1:** Media erorilor modelului de regresie nu este egală cu 0.

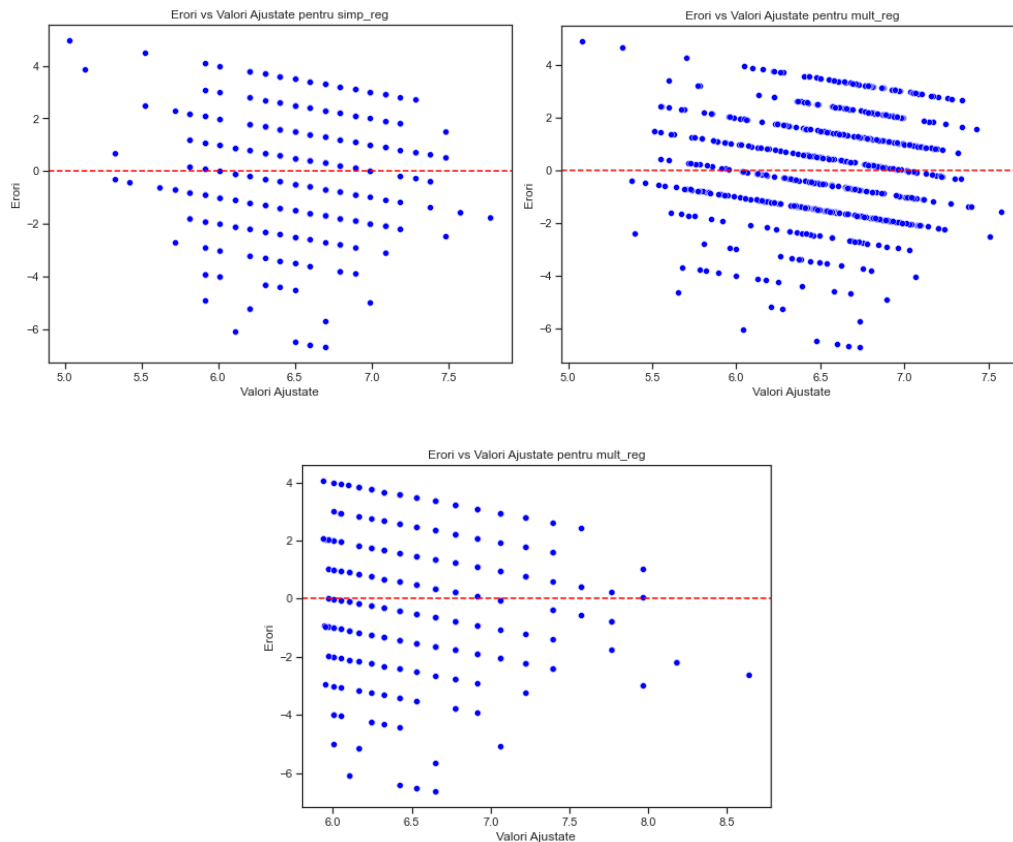
```

In [148]: stats.ttest_1samp(err_simp_reg, 0) #pvalue=0.999
Out[148]: TtestResult(statistic=1.2782609370272606e-14, pvalue=0.9999999999999898, df=677)

In [149]: stats.ttest_1samp(err_mult_reg, 0) #pvalue=0.999
Out[149]: TtestResult(statistic=9.518024032089017e-15, pvalue=0.9999999999999925, df=677)

In [150]: stats.ttest_1samp(err_nonl_reg, 0) #pvalue=0.999
Out[150]: TtestResult(statistic=2.218920713794624e-11, pvalue=0.9999999999823022, df=677)

```



**p-value (simp\_reg) = 0.99**  $> 0.05$ , nu se respinge ipoteza  $H_0$ .

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului de regresie liniară simplă nu diferă semnificativ de valoarea 0.

**p-value (mult\_reg) = 0.99**  $> 0.05$ , nu se respinge ipoteza  $H_0$ .

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului de regresie liniară multiplă nu diferă semnificativ de valoarea 0.

**p-value (nonl\_reg) = 0.99**  $> 0.05$ , nu se respinge ipoteza  $H_0$ .



**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului de regresie neliniară nu diferă semnificativ de valoarea 0.

Toate cele 3 modele de regresie create anterior validează ipoteza de medie a erorilor. Acest lucru ne arată că în medie, nici un model nu a subestimat nici nu a supraestimat valorile variabilei dependente.

### # Ipoteza de normalitate a erorilor

#### Formularea ipotezelor:

**H0:** Erorile modelului urmează o distribuție normală.

**H1:** Erorile modelului nu urmează o distribuție normală.

```
In [167]: stats.jarque_bera(err_simp_reg)
Out[167]: SignificanceResult(statistic=5.585485332554159, pvalue=0.06125298682373662)

In [168]: stats.jarque_bera(err_mult_reg)
Out[168]: SignificanceResult(statistic=6.816796134899452, pvalue=0.03309417256788046)

In [169]: stats.jarque_bera(err_nonl_reg)
Out[169]: SignificanceResult(statistic=5.741218179938732, pvalue=0.05666440234764242)
```

**p-value (simp\_reg) = 0.061** >0.05, nu se respinge ipoteza H0.

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului de regresie liniară simplă sunt normal distribuite.

**p-value (mult\_reg) = 0.033** <0.05, se respinge ipoteza H0.

**Interpretare:** Putem garanta cu un risc de 5%, că erorile modelului de regresie liniară multiplă nu sunt normal distribuite.

**p-value (nonl\_reg) = 0.056** >0.05, nu se respinge ipoteza H0.

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului de regresie neliniară sunt normal distribuite.

Dintre cele trei modele testate, doar două îndeplinesc condiția de normalitate a erorilor, modelul liniar simplu și modelul neliniar. Nerespectarea acestei ipoteze semnifică o putere de predicție mai scăzută a modelului, modelul liniar multiplu având astfel o precizie mai mică decât celelalte modele.

### # Ipoteza de heteroscedasticitate a erorilor

#### Formularea ipotezelor:

**H0:** Variația erorilor modelului este constantă (homoscedasticitate).

**H1:** Variația erorilor modelului nu este constantă (heteroscedasticitate).

```

In [170]: het_breuschpagan(err_simp_reg, simp_reg.model.exog)
Out[170]:
(4.617741448571836,
 0.031642876880971775,
 4.635692698452859,
 0.03166659305073058)

In [171]: het_breuschpagan(err_nonl_reg, nonl_reg.model.exog)
Out[171]:
(2.2202614656212862,
 0.3295158797220695,
 1.1088498247555312,
 0.3305384616802989)

In [172]: het_breuschpagan(err_mult_reg, mult_reg.model.exog, robust=False)
Out[172]:
(5.7306158659256425,
 0.05696558684900387,
 2.7371789941354656,
 0.06547157735328028)

```

**p-value (simp\_reg) = 0.0316** < 0.05, se respinge ipoteza H0.

**Interpretare:** Putem garanta cu un risc de 5% că varianța erorilor modelului liniar simplu nu este constantă.

**p-value (mult\_reg) = 0.065** > 0.05, nu se refuză ipoteza H0.

**Interpretare:** Putem garanta cu o probabilitate de 95%, că varianța erorilor modelului liniar multiplu este constantă.

**p-value (nonl\_reg) = 0.330** > 0.05, nu se respinge ipoteza H0.

**Interpretare:** Putem garanta cu o probabilitate de 95%, că varianța erorilor modelului neliniar este constantă.

În urma testării homoscedasticității erorilor se poate observa faptul că două modele îndeplinesc această ipoteză. Modelul liniar simplu este mai puțin eficient la estimare după criteriul homoscedasticității decât celelalte două modele.

### # Ipoteza de independență a erorilor

#### Formularea ipotezelor:

**H0:** Nu există autocorelare a erorilor la nivelul modelului.

**H1:** Există autocorelare a erorilor la nivelul modelului.

```

In [175]: acorr_breusch_godfrey(simp_reg)
Out[175]:
(4.7992592443523865,
 0.9041778524556424,
 0.4747925044097407,
 0.9066576622748669)

In [176]: acorr_breusch_godfrey(mult_reg)
Out[176]: (5.180416393726668, 0.878804606459353, 0.512020902150811,
0.8822421720269527)

In [177]: acorr_breusch_godfrey(nonl_reg)
Out[177]:
(4.797583815899046,
 0.9042828536539809,
 0.47391292141475994,
 0.9071992269033569)

```

**p-value (simp\_reg) = 0.906** > 0.05, nu se respinge ipoteza  $H_0$ .

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului liniar simplu nu sunt autocorelate.

**p-value (mult\_reg) = 0.882** > 0.05, nu se respinge ipoteza  $H_0$ .

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului liniar multiplu nu sunt autocorelate .

**p-value (nonl\_reg) = 0.907** > 0.05, nu se respinge ipoteza  $H_0$ .

**Interpretare:** Putem garanta cu o probabilitate de 95%, că erorile modelului neliniar nu sunt autocorelate.

Toate modele validează această ipoteză, așa că erorile sunt independente unele de altele, neinfluențându-se.

## 6.2.4 Compararea a doua modele de regresie si alegerea celui mai bun model

### Formularea ipotezelor

**H0:** modelul *mult\_reg* (cu mai multi parametri) nu este semnificativ mai bun decât modelul *simp\_reg*

**H1:** modelul *mult\_reg* (cu mai multi parametri) este semnificativ mai bun decât modelul *simp\_reg*

```

In [179]: sms.stats.anova_lm(simp_reg, mult_reg)
Out[179]:

```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	676.0	2639.877736	0.0	NaN	NaN	NaN
1	675.0	2611.120602	1.0	28.757134	7.433998	0.006566

**P-value = 0.0065** < 0.01, se refuză ipoteza  $H_0$

**Interpretare:** Putem garanta cu un risc asumat de 1% că modelul mai complex (*nonl-reg*) este semnificativ mai bun decât modelul mai simplu (*sim\_reg*).

## 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 nenumерice 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 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.

Capitolul 6 se concentrează 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.