Proiect Analiză Statistică în R

Problema 1. **a)** Scrieți R Script-uri în care se utilizeaza structurile de control *if, for, while*, câte două pentru fiecare structură, şi explicați care este scopul fiecăreia.

- ```
  j<-0
  while(j<=20)
  {
  print(j)
  j<-j+3
  } #afiseaza, din 3 in 3, numerele de la 0 la 20
  ```

- ```
  n<-5
  factorial<-1
  i<-1
  while (i<=n)
  {
  factorial=factorial*i
  i=i+1
  }
  print(factorial)
  #calculeaza 5!
  ```

- ```
  x<-sample(30:70,20,replace=TRUE)
  x
  a<-array(x,dim=c(4,5))
  a
  vanzari<-data.frame(luni=a[,1], marti=a[,2], miercuri=a[,3], joi=a[,4], vineri=a[,5])
  vanzari
  medie_vanzari<-
  data.frame(market1=mean(a[1,]),market2=mean(a[2,]),market3=mean(a[3,]),market4=
  mean(a[4,]))
  medie_vanzari
  s<-0
  for(i in seq_len(nrow(a)))
  for(j in seq_len(ncol(a))) s<-s+a[i,j]
  s #s returneaza suma tuturor vanzarilor a celor 4 magazine in cele 5 zile lucratoare
  ```

- ```
  a<-0
  for(i in seq_len(ncol(medie_vanzari))) if(medie_vanzari[i] > 50) a<-a+1
  ```

a #afiseaza numarul magazinelor cu vanzari de peste 50

- medie_saptamanala<-s/20
  medie_saptamanala
  if(medie_saptamanala>30 & medie_saptamanala<60) print("media saptamanala a vanzarilor este in intervalul (30;60)")

b) Creaţi câte o funcţie pentru a calcula şi afişa:
1) media şi abaterea standard,

```
media<-function(x){M=sum(x)/length(x); M}
abaterest<-function(x){A=sqrt(sum((x-media(x))^2)/(length(x)-1)); A}
```

2) valoarea testului Student pentru ipoteza privind valoarea medie H0: $\overline{X}$ =c, unde c este o constantă,

```
tcalc<-function(x,c){T=(media(x)-c)/abaterest(x);T}
```

3) coeficientul de corelaţie liniară între 2 variabile; nu se va utiliza funcţia *cov*.

```
mediaxy<-function(x,y){N=sum(x*y)/length(x);N}
mediaxpatrat<-function(x){P=sum(x^2)/length(x);P}
beta2<-function(x,y){B=(mediaxy(x,y)-media(x)*media(y))/(mediaxpatrat(x)-
(media(x))^2); B}
corelatielin<-function(x,y){C=beta2(x,y)*abaterest(x)/abaterest(y); C}
```

Apelaţi cele trei funcţii pentru un set de date, şi comparaţi rezultatele cu cele din funcţiile *mean, sd, t.test,cor.*

c) Se dă un vector x care conţine n valori; n este multiplu de 7 şi n>14. Aceste valori redau vânzările dintr-un produs în cele nr_days=7 zile ale săptămânii la nr_markets=k magazine; primele 7 valori sunt vânzările zilnice aferente primului magazin, urmatoarele 7 sunt vânzările zilnice la al 2-lea magazin…. Se cere: scrieţi un R Script care afişează valoarea medie a vânzărilor în fiecare zi, într-un tabel cu două linii (ziua pe linia 1, media pe linia 2). Scrieţi codul pentru un vector x oarecare; atribuiţi apoi valori vectorului x şi rulaţi codul pentru acest vector. Prezentaţi două soluţii alternative; una dintre rezolvări va utiliza structuri de control iar cealaltă o funcţie din familia apply.

- Rezolvare folosind functii
  a<-function(n){A=array(c(1:(7*(n+1)), 7*(1+n), replace=FALSE), dim=c(n,7));A}
  a(4)

```
vanzari<-function(n){D=data.frame(luni=a(n)[,1], marti=a(n)[,2], miercuri=a(n)[,3],
joi=a(n)[,4], vineri=a(n)[,5], sambata=a(n)[,6], duminica=a(n)[,7]); D}
vanzari(4)

media_vanzarilor<-function(n){E=data.frame(luni=mean(a(n)[,1]),
marti=mean(a(n)[,2]), miercuri=mean(a(n)[,3]), joi=mean(a(n)[,4]),
vineri=mean(a(n)[,5]), sambata=mean(a(n)[,6]), duminica=mean(a(n)[,7]));E}
media_vanzarilor(4)
```

- Rezolvare folosind apply

```
valori<-1:28
valori
v<-matrix(valori, nrow=4, ncol=7)
v
vanzari<-data.frame(luni=v[,1], marti=v[,2], miercuri=v[,3], joi=v[,4], vineri=v[,5],
sambata=v[,6], duminica=v[,7])
vanzari
apply(vanzari, MARGIN=2, FUN=mean)
```

d) Realizaţi o analiză comparativă, într-un tabel, a funcţiilor din familia apply privind obiectele acceptate ca input (arguments) respectiv tipul obiectului output (value). Indicaţi sursele bibliografice parcurse.

Functie	**Input**	**output**
Apply	List, array, data frame, matrix	Vector, list, array, matrix
Lapply	List, vector, data frame	List
Sapply	List, vector, data frame	Vector, matrix, array, list

Sursa: https://towardsdatascience.com/dealing-with-apply-functions-in-r-ea99d3f49a71
https://www.guru99.com/r-apply-sapply-tapply.html

e) Descarcaţi de pe Eurostat evoluţia anuală a unei variabile, pentru mai multe tari, in Excel. Transformati setul de date în format tidy data.

```
install.packages("tidyverse")
library(tidyverse)

format_wide<-migr_imm8
format_wide
```

```
format_long<-format_wide %>%
  gather(key = an,
         value = valori,
         starts_with('migr'),
         convert=TRUE)
view(format_long)

format_long<-format_long %>%
  separate(an,
           into=c('variabila', 'an'),
           sep="_",
           convert=TRUE)
view(format_long)
```

Calculaţi valoarea medie a variabilei într-un an; se pot folosi na.omit=TRUE şi na.rm=TRUE, dacă este cazul.

```
medie_migr<-format_long %>%
  filter(an==2017) %>%
  summarize(medie=mean(valori))
medie_migr
```

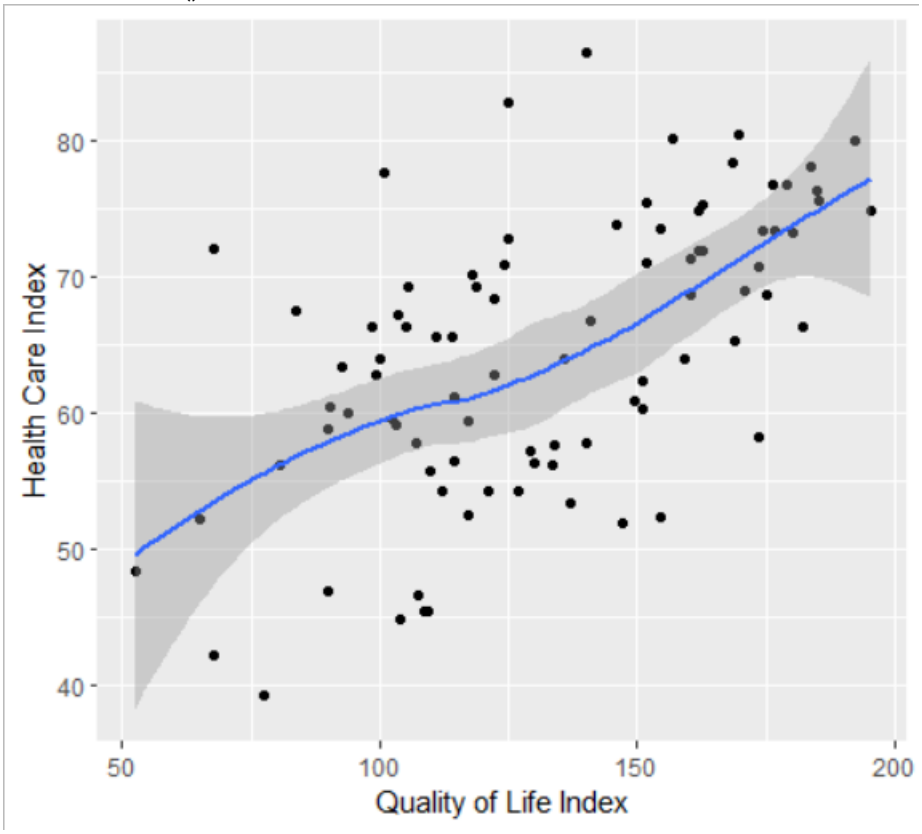
Calculati valoarea medie a variabilei în fiecare an; rezultatele se afişează într-un tabel.

```
medie_an<-function(n){medie_an=format_long %>%
    filter(an==n) %>%
    summarize(medie=mean(valori));
  medie_an}

medii<-
array(data=c(medie_an(2015),medie_an(2016),medie_an(2017),medie_an(2018),medie_an(20
19),medie_an(2020) ))
medii

media_migratiei_pe_ani<-data.frame(an2015=medii[[1]],an2016=medii[[2]],
an2017=medii[[3]], an2018=medii[[4]], an2019=medii[[5]], an2020=medii[[6]])
media_migratiei_pe_ani
```

f) Importaţi/citiţi în R un HTML table care conţine un set de date, de pe Wikipedia sau alta pagina Web. Transformaţi setul de date în forma tidy data, dacă este necesar, şi efectuaţi două prelucrări statistice (medie..., grafice, tabele de frecvență, regresie, previziuni, ...).

```
library(rvest)

pagina<-read_html("https://www.numbeo.com/quality-of-life/rankings_by_country.jsp")
class(pagina)
view(pagina)
tabele<-pagina %>% html_nodes("table")
length(tabele)
qol<-html_table(tabele[[2]])
view(qol)
```

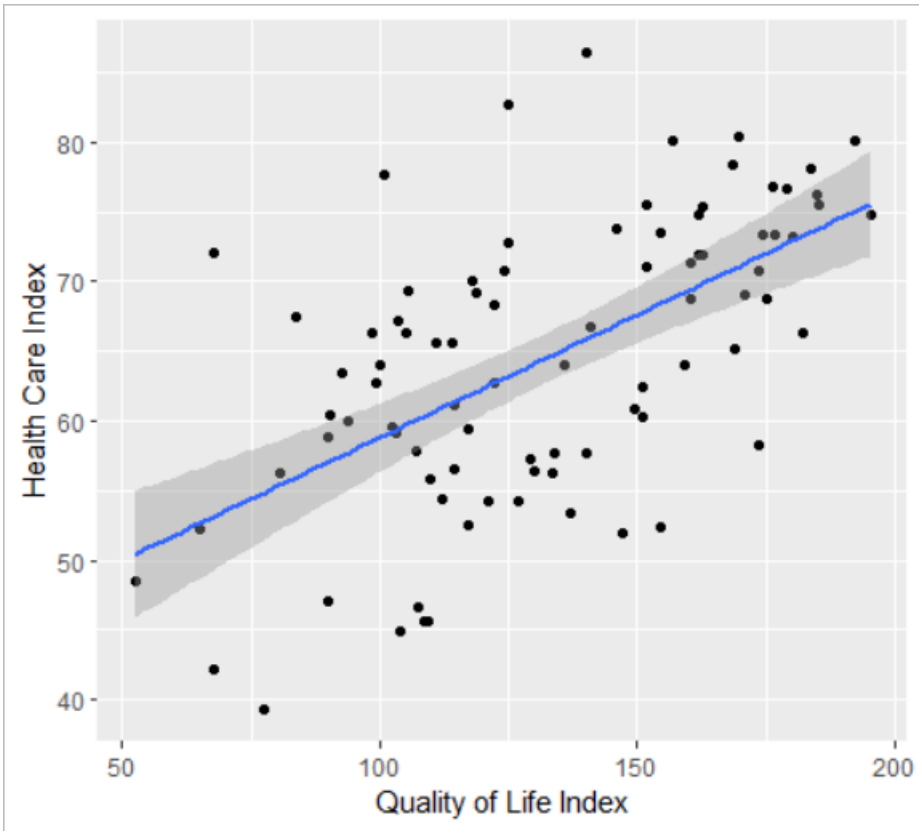
```
ggplot(qol, aes(x = `Quality of Life Index`, y =`Health Care Index`)) +
  geom_point() +
  stat_smooth()
```



cor(qol$`Quality of Life Index`,qol$`Health Care Index`) => ***0.5863177*** (corelatie directa si de intensitate medie intre calitatea vietii si indicele sanatatii.

model<-lm(`Quality of Life Index` ~`Health Care Index`, data=qol)
model => **quality of life index = *1.951* * health care index + *6.322***

```
ggplot(qol, aes(x = `Quality of Life Index`, y =`Health Care Index`)) +
  geom_point() +
  stat_smooth(method=lm)
```



summary(model)

```
> summary(model)

Call:
lm(formula = `Quality of Life Index` ~ `Health Care Index`, data = qol)

Residuals:
    Min      1Q  Median      3Q     Max
-79.118 -21.892   6.112  21.840  53.744

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.3216    19.0856   0.331    0.741
`Health Care Index`   1.9511     0.2924   6.673 2.42e-09

(Intercept)
`Health Care Index` ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.79 on 85 degrees of freedom
Multiple R-squared:  0.3438,    Adjusted R-squared:  0.336
F-statistic: 44.53 on 1 and 85 DF,  p-value: 2.419e-09
```

Problema 2. Selectați un set de date CSV din https://www.kaggle.com/datasets care conține cel puțin șase variabile, dintre care cel puțin două sunt cantitative și două calitative (categoriale sau text); selectați variabilele. Se cere:

https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data

a) Descrieți succint datele și variabilele selectate; indicați tipul variabilelor. Importați datele în R, cu o funcție din readr. Realizați cleaning data: verificați tipul/valorile variabilelor și transformați-le adecvat pentru a putea fi prelucrate (parse, mutate, as. ...); transformați variabilele categoriale în factor; recodificați daca este cazul.

Setul de date ales contine informatii cu privire la cererile pentru a beneficia de un credit. Baza de date este alcatuita din 690 de observatii si 16 variabile. Avem 5 variabile **cantitative**: Age (varsta), Debt (datoria acumulata), YearsEmployed (vechimea la locul de munca masurata in ani), CreditScore si Income (venit). Gender (0-F; 1-M), Married (0-singur/divortat/etc; 1-casatorit), BankCustomer (0-nu este client al bancii; 1-este client), PriorDefault (0-fara antecedente in neplata ratelor; 1-antecedente de neplata), Employed (0-somer; 1-angajat), DriversLicense (0-fara permis de conducere; 1-cu permis de conducere) si Approved (0-aplicatie respinsa; 1-credit acordat) sunt variabile **ordinale**. Iar Industry (industria din care face parte job-ul), Ethnicity (etnia), Citizen (motivul pentru care detine cetatenie), ZipCode (cod postal) sunt variabilele **nominale (categoriale)** ale bazei de date.

```
date<-read.csv("C:/Users/suciu/Desktop/semestrul 6/Analiza in R/proiect/clean_dataset.csv")
view(date)
```

Am transformat variabilele categoriale (exceptand variabila ZipCode=cod postal) in factor:

```
etnie<-c("White", "Black", "Latino", "Other", "Asian")
date$Ethnicity<-parse_factor(date$Ethnicity, levels=etnie)
str(date$Ethnicity)

cetatenie<-c("ByBirth", "ByOtherMeans", "Temporary")
date$Citizen<-parse_factor(date$Citizen, levels=cetatenie)
str(date$Citizen)

industrie<-c("Industrials", "Materials", "CommunicationServices", "ConsumerDiscretionary", "ConsumerStaples", "Education", "Energy", "Financials", "Healthcare", "InformationTechnology", "Real Estate", "Research", "Transport", "Utilities" )
```

```
date$Industry<-parse_factor(date$Industry, levels=industrie)
str(date$Industry)
```

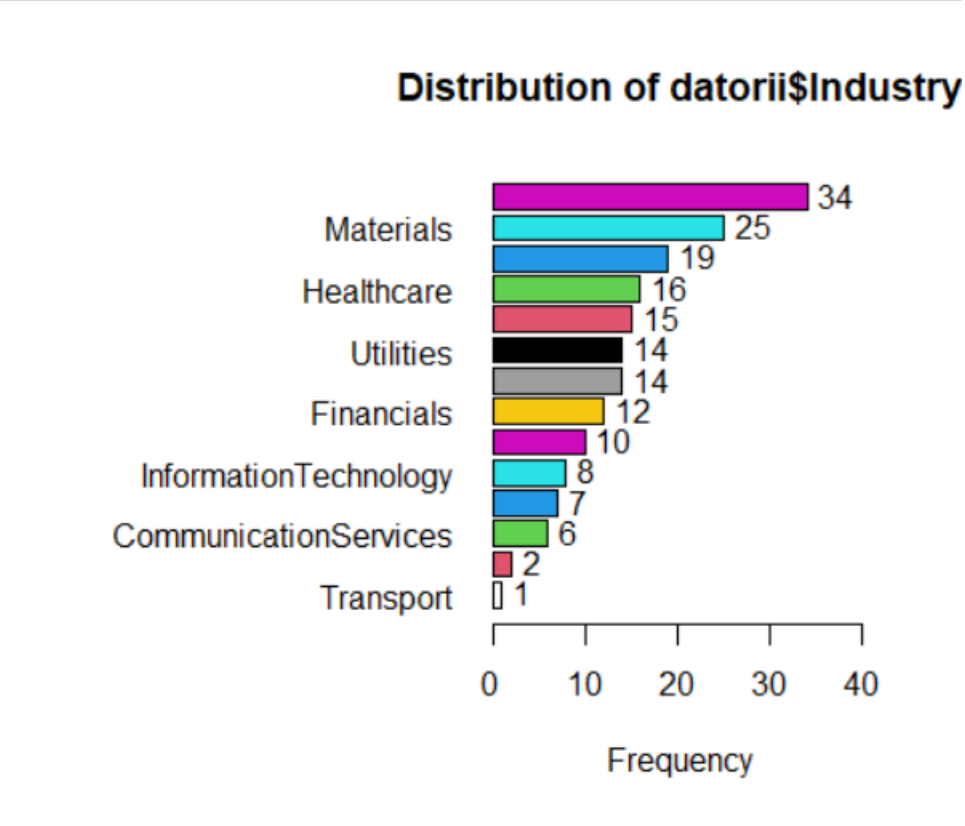
b) Utilizați funcția filter şi o altă funcție la alegere (dintre cele discutate) din pachetul dplyr; explicați obiectivele şi rezultatele obținute.

Filter

Am folosit functia Filter pentru a vedea ce varsta au si in ce industrie profeseaza persoanele cu datorii substantiale la banca. Avand in vedere tabelul de frecventa creat mai jos, putem observa faptul ca, cei mai multi angajati care nu si-au platit ratele creditelor lucreaza in domeniul energetic (34 din 183), urmat de industria materialelor samd.

```
datorii<- date %>%
  select(Debt,Age, Industry) %>%
  filter(Debt>=7)

datorii

tab1(datorii$Industry, sort.group = "decreasing", cum.percent = FALSE )
```

datorii	***Frequency***	***Percent***
Energy	34	18.6
Materials	25	13.7
ConsumerStaples	19	10.4
Healthcare	16	8.7
Education	15	8.2
Utilities	14	7.7
Industrials	14	7.7
Financials	12	6.6
ConsumerDiscretionary	10	5.5
InformationTechnology	8	4.4
Real Estate	7	3.8
CommunicationServices	6	3.3
Research	2	1.1
Transport	1	0.5
Total	183	100



In continuare am modificat tabelul ‘datorii’ astfel incat sa afiseze descrescator observatiile in functie de datoria acumulata la banca. Functia ***arrange*** a fost folosita.
Head(arrange(datorii, desc(Debt)))

– faptul ca am inclus functia ‘head’ imi va permite sa afisez primele 6 observatii din tabelul ‘datorii’ care are un total de 183 de observatii.

Debt	***Age***	***Industry***
28	56.42	Energy
26.335	48.75	Healthcare
25.210	43.25	Materials
25.125	35.17	Utilities
25.085	48.25	Industrials
22.290	76.75	Education

c) Aplicaţi 6 funcţii din familia **str**_... din pachetul stringr, în care utilizaţi următoarele meta-caractere: . \ | { } [ ] ^ $ - * + ? pentru a construi regular espressions.

- ***str_detect(cetatenie, "By.")*** #arata daca sirurile de caractere din vector incep cu secventa "By"
- ***str_view(etnie, "^Latino$")*** #afiseaza pe ecran toate sirurile vectorului 'etnie' si le evidentiaza pe cele care contin doar cuvantul 'Latino'
- ***str_subset(industrie, ".on")*** #returneaza sirurile de caractere din vectorul 'industrie' care contin secventa 'on'

- ***str_replace(industrie, "Education", "educ")*** #cauta in vector cuvantul "Education" si il inlocuieste cu "educ", iar apoi afiseaza noul sir cu cuvantul inlocuit
- ***str_locate(etnie, "ite")*** #analizeaza fiecare sir de caractere din vectorul dat si evidentiaza al catelea cuvant contine secventa 'ite' si pe ce pozitie incepe si pe ce pozitie se incheie
- ***str_split(nume_coloane, "\\b")***

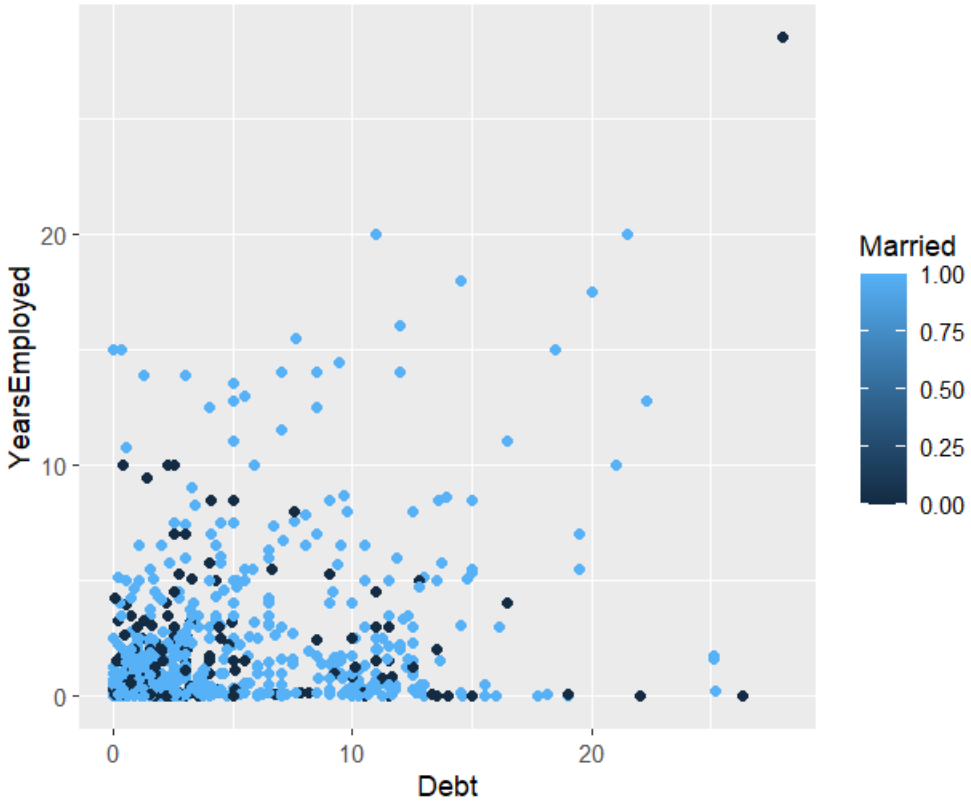
d) Formulaţi obiective (întrebări) asupra datelor şi găsiţi răspunsul, utilizând următoarele analize preliminare (de tipul analiza exploratorie a datelor): grafice, descriptive, corelaţii, medii condiţionate şi anova, testul chi-square, regresie.

→ ***Care sunt statisticile descriptive ale varstei clientilor bancii respective?***

Min. 1st Qu. Median Mean 3rd Qu. Max.
13.75 22.67 28.46 31.51 37.71 80.25

De unde reiese faptul că, în medie, clienții băncii care depun cerere pentru un credit au vârsta de 31 de ani. Deci, am putea spune că aceștia deja sunt pe piața muncii cu o anumită experiență și, cel mai probabil, doresc acest credit pentru a investi în familie sau locuință. Mediana este egală cu aproximativ 28, astfel concluzionăm că 50% din persoane au până în 28 de ani, iar cealaltă jumătate a acestora au peste 28 de ani.

→ ***Cum sunt distribuite persoanele în funcție de scorul datoriei, experienței la muncă și statutul marital?***



Din grafic putem observa că până la pragul de 10 ani de experiență la locul de muncă și scorul datoriei de 15 avem atât persoane căsătorite, cât și singure. Dar perosanele cu peste 10 ani de experiență pe piața muncii apar a fi căsătorite. Avem de a face cu un fenomen interesant în care unul dintre clienții băncii este necăsătorit, are aproape 30 de ani de vechime la locul de muncă și o datorie foarte mare. Având în vedere maximul vârstei aflat la subpunctul anterior, putem spune că această persoană are 80 de ani.

→ ***Se diferențiază venitul persoanelor în funcție de etnie?***

Ethnicity	*mean*	*sd*
1 White	*776.*	*3231.*
2 Black	*968.*	*3309.*
3 Latino	*435.*	*1448.*
4 Other	*4389.*	*18908.*
5 Asian	*1762.*	*6936.*

În acest tabel s-a calculat media și abaterea standard pentru grupele de persoane în funcție de etnie.
Vom rula testul ANOVA pentru a determina dacă venitul persoanelor în funcție de etnie.

	Df	***Sum Sq***	***Mean Sq***	***F value***	***Pr(>F)***
Ethnicity	*4*	*3.944e+08*	*98609555*	*3.689*	*0.00555*
Residuals	*685*	*1.831e+10*	*26727859*		

Formulăm ipoteza nulă: Venitul nu diferă în cele 5 grupe de persoane.
Probabilitatea rezultată în urma testului este mai mică decât pragul de 1%, deci ipoteza alternativă se acceptă, adică venitul persoanelor se diferențiază în cele 5 grupe.

e) Pornind de la rezultatele anticipate la d), propuneţi şi estimaţi două modele de regresie pentru aceeasi variabilă dependentă, ce includ cel puţin o variabilă categorială. Alegeţi-l pe cel mai potrivit după criteriile mape si rmse. Testaţi semnificativitatea, interpretaţi coeficienţii, salvaţi predicţiile în setul de date şi analizaţi comportamentul erorilor de predicţie (reziduurilor) din perspectiva ipotezelor econometrice.

Primul model creat are ca variabilă dependentă venitul, iar ca variabile explicative: varsta și etnia.
Coeficienții regresiei:

(Intercept)	*Age*
714.516	*2.099*
factor(Ethnicity)Black	*factor(Ethnicity)Latino*
186.550	*-359.673*
factor(Ethnicity)Other	*factor(Ethnicity)Asian*
3599.211	*971.545*

REGRESIA: venit = 714.52 + 2.1*varsta + 186.6*negru(etnie) +3599.2*altele(etnie) -359.67*latino(etnie) +971.54*asian(etnie)

Al doilea model are aceeași variabilă explicată, iar la variabilele independente am mai adăugat vechimea la locul de muncă și statutul marital.

Coeficienții regresiei:

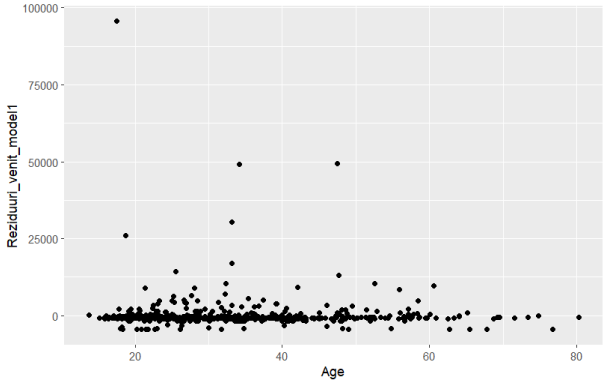
(Intercept)	*Age*
895.062	*-5.436*
YearsEmployed	*factor(Ethnicity)Black*
71.310	*87.027*
factor(Ethnicity)Latino	*factor(Ethnicity)Other*
-284.500	*3566.311*
factor(Ethnicity)Asian	*factor(Married)1*
937.706	*-110.024*

REGRESIA: venit = 895.06 + 71.31*ani_vechime -284.5*latino +937.71*asian +3566.31*altele -5.436*varsta -110.02*casatorit

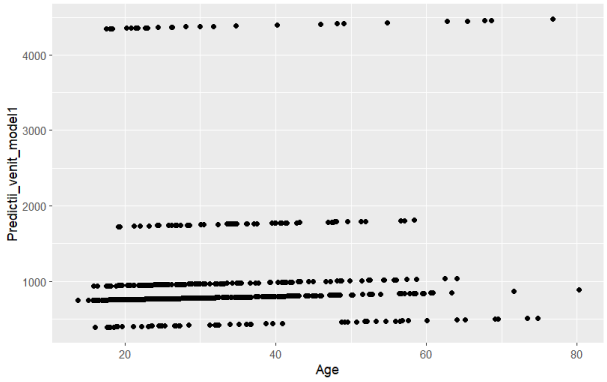
	Model 1	*Model 2*
R^2	0.02111	0.02286
$\overline{R^2}$	0.01396	0.01283
MSE	26533599	26486166
RMSE	5151.077	5146.471

Conform RMSE-ului, acesta fiind un criteriu de minim, vom alege cel de-al doilea model.

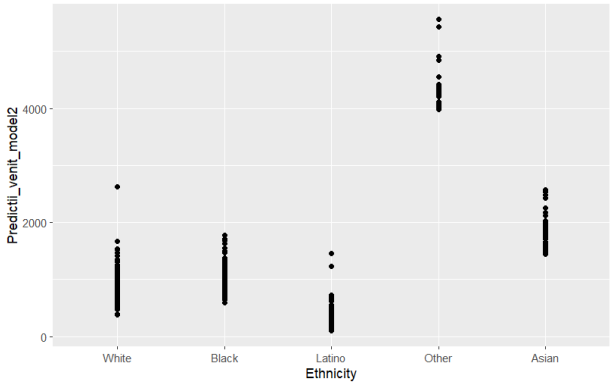
- ggplot(date, aes(Age, Reziduuri_venit_model1)) + geom_point()



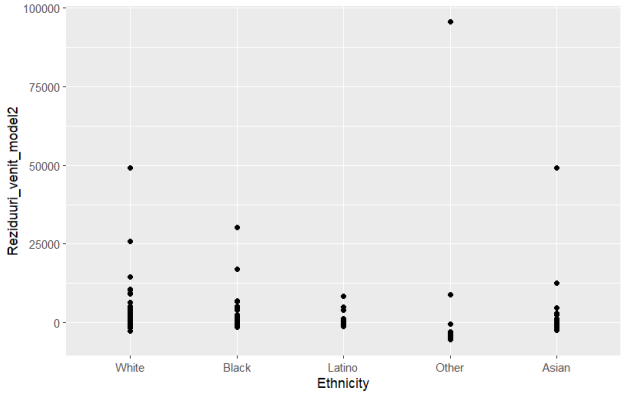
- ggplot(date, aes(Age, Predictii_venit_model1)) + geom_point()



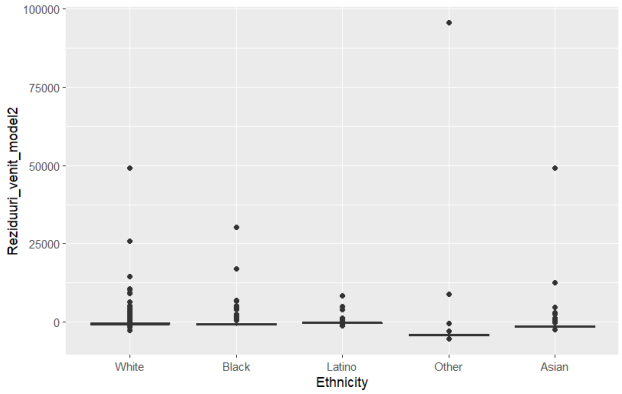
- ggplot(date, aes(Ethnicity, Predictii_venit_model2)) + geom_point()



- ggplot(date, aes(Ethnicity, Reziduuri_venit_model2)) + geom_point()



- ggplot(date, aes(Ethnicity, Reziduuri_venit_model2)) + geom_boxplot()


100000
75000
50000
25000
0
Reziduuri_venit_model2
White
Black
Latino
Other
Asian
Ethnicity