FEAA
FACULTATEA DE ECONOMIE
ȘI ADMINISTRAREA AFACERILOR
UNIVERSITATEA "AL. I. CUZA" IAȘI

Proiect Introducere în R

Student: Dancă Alexandra-Simona
3104401050015M211018
DM11

IAȘI
2022

Cuprins

1.Introducere

Comportamentul criminal, în special comportamentul violent și antisocial, este considerat a fi o problemă socială majoră cu cauze complexe. Se știe că o multitudine de factori de mediu, sociali și psihologici sunt asociați cu risc crescut de condamnare pentru acest tip de criminalitate. Factorii interdependenți includ sărăcia, locuințe proaste, niveluri ridicate de inegalitate socială în societate, nivel scăzut de educație, alimentație proastă, stima de sine scăzută și impulsivitate.

Baza de date a fost descărcată de pe site-ul vincentarelbundock la https://vincentarelbundock.github.io/Rdatasets/articles/data.html. Această bază de date conține infracțiunile comise per persoană, anul în care s-au comis infracțiunile, probabilitatea de arest, probabilitatea de a fi condamnat la o anumită sentință, probabilitatea de a ajunge la închisoare, sentința medie, poliție per cap de locuitor, densitate, venituri fiscale pe cap de locuitor, regiunea de proveniență, zona statistică metropolitană standard, minorități procentuale din regiune, salariul saptamanal in constructii, salariu saptamanal in transporturi, utilitati, comunicatii, salariu săptămânal în comerțul cu ridicata și cu amănuntul, salariu săptămânal în finanțe, asigurări și imobiliare, salariu săptămânal în industria serviciilor, salariu săptămânal în producție, salariu săptămânal în guvernul federal, salariu săptămânal în guvernul de stat, salariul săptămânal în administrația locală, mix de infracțiuni: faţă în faţă/altul și procentul tinerilor de sex masculin din regiune.

Obiectivele acestui proiect sunt următoarele:

- Identificarea outlierilor și eliminarea acestora, dacă este cazul;
- Existența corelațiilor, asocierilor, concordantelor dintre variabile;
- Testarea mediilor populației;
- Compararea rezultatelor obținute ce indică existența unor legături între variabilele alese;
- Identificarea modului de influență a unei/ unor variabile asupra celorlalte.

Mijloacele prin intermediul cărora dorim să atingem obiectivele sunt:

- Utilizarea analizei grafice și numerică a variabilelor analizate
- Analiza statistică a variabilelor categoriale prin utilizarea analizei de asociere și cea de concordanță
- Estimarea și testarea mediilor
- Analiza de regresie și corelație.

2. Prezentarea bazei de date

Baza de date inițială are ca subiect crimele din regiunile din Carolina din Nord în perioada 1981 până în 1987.

Baza de date inițială este prezentată în Figura 1 și aceasta conține 23 de potențiali factori ce pot afecta rata criminalității pentru 90 de comitate.

	X	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	region	smsa	pctmin
1	1	1	81	0.0398849	0.289696	0.4020620	0.472222	5.61	0.0017868	2.3071590	25.69763	central	no	20.21870
2	2	1	82	0.0383449	0.338111	0.4330050	0.506993	5.59	0.0017666	2.3302540	24.87425	central	no	20.21870
3	3	1	83	0.0303048	0.330449	0.5257030	0.479705	5.80	0.0018358	2.3418010	26.45144	central	no	20.21870
4	4	1	84	0.0347259	0.362525	0.6047060	0.520104	6.89	0.0018859	2.3464200	26.84235	central	no	20.21870
5	5	1	85	0.0365730	0.325395	0.5787230	0.497059	6.55	0.0019244	2.3648960	28.14034	central	no	20.21870

În bază se afla variabile enumerate și în capitolul 1 precum densitatea, probabilitatea de a fi arestat, sentința medie, polițiști pe cap de locuitor, procentul de minorități, etc. După cum se poate observa in Figura 2, baza de date conține conține 25 de coloane.

```
criminalitate      630 obs. of 25 variables
```

Asupra acestei baze s-a efectuat o serie de transformări pentru a îndeplini obiectivele propuse. Mai întâi am verificat dacă în baza de date există valori lipsă (null).

```
> sum(is.na(criminalitate))
[1] 0
```

Rezultatul din consolă arată faptul că nu exista nici o astfel de valoare null.

În continuare, am realizat o selecție la nivelul bazei inițiale „crime”, impunând trei condiții:

- anul=1982 , în baza de date prezentat sub forma „82”;
- probabilitățile de a fi arestat ≤ 1;
- probabilitate de a fi condamnat ≤ 1.

Așadar, condițiile arată în felul următor în R:

```
criminalitate_v2    84 obs. of 25 variables
```

În urma acestei selectări în bază au mai rămas 84 de observații.

Variabilele au fost redenumite cu denumiri sugestive, s-au definit categoriile variabilelor nenumerice și s-au selectat din baza doar variabilele necesare analizei, astfel formând un nou dataframe sub numele de „criminalitate_v2”.

	X	county	year	criminalitate_loc	prob_arest	prbconv	prbpris	avgsen	polpc	densitate	taxpc	regiune	urban	minoritati	wcon	wtuc	wtrd
2	2	1	82	0.0383449	0.338111	0.433005	0.506993	5.59	0.0017666	2.3302540	24.87425	central	no	20.21870	212.7542	369.2964	189.5414
9	9	3	82	0.0190651	0.162218	0.772152	0.377049	5.71	0.0007047	0.9922780	35.64073	central	no	7.91632	186.9658	345.7217	156.8826
16	16	5	82	0.0123229	0.380000	0.736842	0.392857	9.70	0.0013555	0.4170213	19.52253	west	no	3.16053	147.9290	343.4066	160.7526
23	23	7	82	0.0173807	0.612335	0.582734	0.370370	6.11	0.0014165	0.4878049	55.98516	central	no	47.91610	308.4180	326.1756	181.7081
30	30	9	82	0.0087368	0.567164	0.640351	0.356164	9.95	0.0007389	0.5352113	17.20366	west	no	1.79619	395.9276	28.8577	149.3260
37	37	11	82	0.0198536	0.425856	0.196429	0.363636	4.32	0.0013588	0.5951417	22.88071	west	no	1.54070	215.9605	320.5128	158.7005
44	44	13	82	0.0363259	0.332885	0.430303	0.497653	5.32	0.0013436	0.5048426	22.52288	other	no	32.17940	205.6120	286.4157	172.4643
51	51	15	82	0.0200047	0.444706	0.370370	0.442857	9.48	0.0008473	0.3024251	23.05781	other	no	61.05400	188.5370	169.6833	179.1843
58	58	17	82	0.0211586	0.420155	0.476015	0.387597	5.50	0.0013450	0.3492605	24.06734	other	no	40.38900	166.1918	289.0900	173.3922
72	72	21	82	0.0367896	0.240573	0.297741	0.565517	7.01	0.0017689	2.4855840	24.50407	west	yes	9.62444	241.8530	369.2143	197.0567
79	79	23	82	0.0290211	0.286639	0.277419	0.470930	6.38	0.0013685	1.4702380	18.24509	west	no	7.93198	203.0730	336.8322	179.2207
86	86	25	82	0.0370917	0.256811	0.433850	0.626374	8.25	0.0016122	2.4505490	19.71914	central	no	15.09980	239.8677	319.1938	194.6381
93	93	27	82	0.0371544	0.303606	0.206897	0.364198	8.00	0.0013686	1.4479830	18.93122	west	no	6.45795	223.1021	350.7131	183.6174
100	100	33	82	0.0144477	0.485342	0.476510	0.478873	6.54	0.0006589	0.5046729	28.97830	central	no	43.91690	199.8002	205.1282	165.0402
107	107	35	82	0.0489555	0.239077	0.234968	0.457912	6.23	0.0017501	2.7247470	28.31856	central	no	10.08380	252.9197	358.0455	217.3804
114	114	37	82	0.0201309	0.325145	0.466667	0.390476	9.03	0.0011636	0.4844633	20.97966	central	no	27.80790	191.4752	316.7421	173.8058
121	121	39	82	0.0188128	0.258856	0.357895	0.441176	6.27	0.0011277	0.4314159	19.88560	west	no	3.94549	187.2659	342.4658	147.3551
128	128	41	82	0.0304537	0.423469	0.397590	0.560606	5.95	0.0017091	0.6923077	28.94018	other	no	42.64210	182.3607	299.1453	210.9959
135	135	45	82	0.0319727	0.219476	0.327731	0.533333	8.27	0.0011440	1.7927350	25.02926	central	no	21.74990	218.8868	346.4366	196.4433
142	142	47	82	0.0305205	0.336927	0.538000	0.371747	6.68	0.0018921	0.5479744	24.94390	other	no	33.40320	252.4503	260.3292	171.1498
149	149	49	82	0.0361096	0.310412	0.340426	0.408088	10.25	0.0013467	1.0485020	21.56757	other	no	29.90720	238.1145	291.8791	178.8462
156	156	51	82	0.0628671	0.223799	0.257295	0.417666	9.44	0.0014883	3.7793000	23.15848	other	yes	37.77920	221.8046	301.3916	177.5444
163	163	53	82	0.0199702	0.184211	0.690476	0.241379	7.43	0.0009635	0.4570313	24.72824	other	no	17.90960	203.9627	153.8462	174.8252

Așadar, variabila „prob_arest”, adică probabilitatea de a fi condamnat va fi transformată din variabilă numerică în variabilă categorială . Astfel, probabilitatea de a fi condamnat a fost împărțită în 4 categorii:

- de la 0 până la 0,30 reprezintă probabilitate mica de a fi condamnat;
- de la 0,3 până la 0,5 reprezintă probabilitate medie;
- de la 0,5 până la 0,7 reprezintă probabilitate mare;
- de la 0,7 până la 1 reprezintă probabilitate foarte mare.

Prin aplicarea funcției class pentru toate variabilele prezente în baza de date criminalitate_v2 se observă că variabilele regiune și urban sunt definite ca date de tip character.

```
> sapply(criminalitate_v2,class)
criminalitate_loc       prob_arest        densitate          regiune            urban       minoritati
        "numeric"        "numeric"        "numeric"         "factor"         "factor"        "numeric"
   prob_arest.cat
         "factor"
```

Așadar este necesară convertirea acestora în variabile factor și definirea nivelelor.

Cu ajutorul codului *criminalitate_v2<-criminalitate_v2[-3]* am eliminat prob_arest, variabila care a fost transformată în variabilă categorială si salvată cu numele de prob_arest. De asemenea se vor elima si celelalte variabile care o sa ne trebuiasca.

Acum toate variabilele au categoriile definite și urmează exportarea bazei de date finală.

Descriere a bazei cu functii de descriere a bazei.

Pentru a face o descriere succintă a bazei de date se folosesc mai multe funcții din pachetul R.

```
> dim(criminalitate_v2)
[1] 84  6
```

Conform figurilor de mai sus se poate observa că baza de date, conține 6 variabile cu 84 de observații

```
> str(criminalitate_v2)
'data.frame':   84 obs. of  6 variables:
 $ criminalitate_loc: num  0.03834 0.01907 0.01232 0.01738 0.00874 ...
 $ densitate        : num  2.33 0.992 0.417 0.488 0.535 ...
 $ regiune          : Factor w/ 3 levels "central","other",..: 1 1 3 1 3 3 2 2 2 3 ...
 $ urban            : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
 $ minoritati       : num  20.22 7.92 3.16 47.92 1.8 ...
 $ prob_arest.cat   : Factor w/ 4 levels "mica","medie",..: 2 1 2 3 3 2 2 2 2 1 ...
```

În urma transformării variabilelor, se poate observa că baza de date are 3 variable numerice criminalitate_loc, densitate, minorități) și 3 categoriale (regiunea, urban și prob_arest.cat).

3. Analiza grafică și numerică a variabilelor analizate

În acest capitol analizeazăm o singură coloană de date numerice. Acestă analiză calculează statistici, efectuează teste de ipoteze și construiește diferite grafice.

Statisticile calculate se încadrează în mai multe categorii de bază:

- **măsuri ale tendinței centrale**- statistici care descriu centrul datelor, inclusiv media, mediana, modul și media geometrică.
- **măsuri de răspândire**- statistici care descriu dispersia datelor, inclusiv varianța, abaterea standard, intervalul și intervalul interquartil.
- **măsuri de formă**- statistici care compară forma datelor cu cea a unei distribuții normale, inclusiv asimetria și curtoza.

De asemnea se vor identifica valorile extreme cu ajutorul diagramei box plot.

3.1. Analiza descriptivă a variabilelor numerice și nenumerice

Statistica descriptivă implică expunerea cât mai sugestivă a datelor empirice cu ajutorul indicatorilor tendinței centrale, a indicatoilor variației și a valorilor minime și maxime. În această parte a capitolului se vor determina și se vor interpreta indicatorii statistici descriptivi.

3.1.1. Analiza descriptiva a variabilelor numerice

A. Rata de criminalitate

Pentru a realiza statisticile descriptive la nivelul bazei de date se va folosi functia summary().

```
> summary(rata_criminalitate)
 criminalitate_loc
 Min.   :0.008737
 1st Qu.:0.020254
 Median :0.031243
 Mean   :0.033909
 3rd Qu.:0.041065
 Max.   :0.089035
```

```
> describe(rata_criminalitate)
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 84 0.03 0.02   0.03    0.03 0.02 0.01 0.09  0.08    1      0.6  0
```

Rata de criminalitate în comitatele din Carolina de Nord, Statele Unite, este în **medie** egală cu 0,03% iar unitățile se **abat de la medie** cu 0,02%.

Mediana este egală cu 0,03 de unde rezultă că jumătate din comitatele din Carolina de Nord au o rata de criminalitate de până la 0,03% iar jumătate o rată de peste 0,03%.

Rata maximă este egală cu 0,09%.

Rata minima este egala cu 0,01.

Quartila 1 = 0,02 % de unde rezultă că 0,02% din comitate au valoarea ratei de criminalitate de până la 0,02 % și 75 % au mai mult de 0,02 % .

Quartila 3 = 0,04% așadar 75 % din comitate au valoarea ratei de crime până la 0,04 % și 25% au mai mult de 0,04 %.

Coeficientul de asimetrie (Skewness) = 1 arată că distribuția este asimetrică la dreapta. (>0).

Coeficientul de boltire (Kurtosis) = 0,6 arată că distribuția este leptocurtică. (>0).

B. Probabilitatea de arest

Chiar daca in noua baza de date „criminalitate_v2" probabilitatea de arest este de tip category, vom lua numerele din baza de date initiala „criminalitate"

```
> summary(prob_arest)
     prbarr
 Min.   :0.1371
 1st Qu.:0.2139
 Median :0.2769
 Mean   :0.3030
 3rd Qu.:0.3542
 Max.   :1.0000
```

```
> describe(prob_arest)
   vars  n mean   sd median trimmed mad  min max range skew kurtosis   se
X1    1 89  0.3 0.13   0.28    0.29 0.1 0.14   1  0.86 2.02     7.26 0.01
```

Probabilitatea de a fi arestat în comitatele din Carolina de Nord este în **medie** egală cu 0,3 iar unitățile se **abat de la medie** cu 0,11.

Mediana este egală cu 0,28 de unde rezultă că jumătate din comitatele din Carolina de Nord au o probabilitate de arest de până la 0,28 iar jumătate o probabilitate de peste 0,28.

Probabilitatea maximă este egală cu 0,61.

Probabilitatea minima este egală cu 0,14.

Quartila 1 = 0,21 așadar 25% din comitate au valoarea probabilității de arest de până la 0,21 și 75 % au mai mult de 0,21.

Quartila 3 = 0,35 așadar 75 % din comitate au valoarea probabilității de arest până la 0,35 și 25% au mai mult de 0,35.

Coeficientul de asimetrie (Skewness) = 2,02 arată că distribuția este asimetrică la dreapta (>0).

Coeficientul de boltire (Kurtosis) = 7,26 arată că distribuția este leptocurtică (>0).

C. Procentul de minorități

```
> summary(minoritati)
   minoritati
 Min.   : 1.541
 1st Qu.:10.064
 Median :25.642
 Mean   :25.939
 3rd Qu.:38.326
 Max.   :61.942
> describe(minoritati)
   vars  n  mean    sd median trimmed   mad  min   max range skew kurtosis   se
X1    1 84 25.94 16.66  25.64   25.11 21.39 1.54 61.94  60.4 0.28    -0.94 1.82
```

Procentul de minorități în comitatele din Carolina de Nord este în **medie** egală cu 25,94% iar unitățile se **abat de la medie** cu 16,66%.

Mediana este egală cu 25,64% de unde rezultă că jumătate din comitatele din Carolina de Nord au un procent de minorități de până la 25,64% iar jumătate un procent de peste 25,64%.

Procentul maxim de minorități este egal cu 61,94%.

Procentul minim este egal cu 1,54%.

Quartila 1 = 10,06% așadar 25% din comitate au valoarea procentului de minorități de până la10,06 % și 75% au mai mult de 10,06%.

Quartila 3 = 38,33% așadar 75% din comitate au valoarea procentului de minorități până la38,33% și 25% au mai mult de 38,33%.

Coeficientul de asimetrie (Skewness)= 0,28 arată că distribuția este asimetrică la dreapta. (>0).

Coeficientul de boltire (Kurtosis)= -0,90 arată că distribuția este platicurtică. (<0).

D. Densitatea

```
> summary(densitate)
   densitate
 Min.   :0.2629
 1st Qu.:0.5494
 Median :0.9753
 Mean   :1.4125
 3rd Qu.:1.4958
 Max.   :7.9527
> describe(densitate)
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 84 1.41 1.43   0.98    1.08 0.71 0.26 7.95  7.69 2.45      6.2 0.16
```

Densitatea este în **medie** egală cu 1,41 pers/km^2 iar unitățile se **abat de la medie** cu 1,43.

Mediana este egală cu 0,98 pers/km^2de unde rezultă că jumătate din comitatele din Carolina de Nord au o densitate de până la 0,98 iar jumătate peste 0,98.

Densitatea maximă de persoane este egală cu 7,95.

Densitatea minima este egal cu 0,26.

Quartila 1 = 0,55. Așadar 25% din comitate au valoarea densității de până la 0,55 și 75 % au mai mult de 0,55.

Quartila 3 = 1,50 așadar 75 % din comitate au valoarea densității până la 1,50 și 25% au mai mult de 1,50.

Coeficientul de asimetrie (Skewness) = 2,45 arată că distribuția este asimetrică la dreapta. (>0).

Coeficientul de boltire (Kurtosis) = 6,2 arată că distribuția este leptocurtică. (>0).

E. Toate variabilele

```
> summary(criminalitate_v2)
 criminalitate_loc    densitate         regiune   urban      minoritati         prob_arest.cat
 Min.   :0.008737   Min.   :0.2629   central:33   no :76   Min.   : 1.541   mica       :49
 1st Qu.:0.020254   1st Qu.:0.5494   other  :34   yes: 8   1st Qu.:10.064   medie      :31
 Median :0.031243   Median :0.9753   west   :17            Median :25.642   mare       : 4
 Mean   :0.033909   Mean   :1.4125                        Mean   :25.939   foarte mare: 0
 3rd Qu.:0.041065   3rd Qu.:1.4958                        3rd Qu.:38.326
 Max.   :0.089035   Max.   :7.9527                        Max.   :61.942
> describe(criminalitate_v2)
                  vars  n  mean    sd median trimmed   mad  min   max range skew kurtosis   se
criminalitate_loc    1 84  0.03  0.02   0.03    0.03  0.02 0.01  0.09  0.08 1.00     0.60 0.00
densitate            2 84  1.41  1.43   0.98    1.08  0.71 0.26  7.95  7.69 2.45     6.20 0.16
regiune*             3 84  1.81  0.75   2.00    1.76  1.48 1.00  3.00  2.00 0.32    -1.20 0.08
urban*               4 84  1.10  0.30   1.00    1.00  0.00 1.00  2.00  1.00 2.71     5.40 0.03
minoritati           5 84 25.94 16.66  25.64   25.11 21.39 1.54 61.94 60.40 0.28    -0.94 1.82
prob_arest.cat*      6 84  1.46  0.59   1.00    1.40  0.00 1.00  3.00  2.00 0.83    -0.34 0.06
```

Din outputul functiei summary() se pot observa frecventele pt variabilele categoriale.

Variabila regiune prezintă faptul că cele mai multe comitate se afla în alta regiune față de cea centrala și cea de vest, 76 nu sunt din zona urbană(metropolitan) iar restul de 8 se află în zona urbană.

În ceea ce privește probabilitatea de condamnare, cele mai multe comitate au o probabilitate medie iar cele mai puține, și anume 10 au o probabilitate foarte mare.

3.1.2. Analiza descriptivă a variabilelor nenumerice

A. Analiza ratei de criminalitate în functie de zona(urbană/rurală)

```
 Descriptive statistics by group
group: no
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 76 0.03 0.01   0.03    0.03 0.01 0.01 0.07  0.06 0.72    -0.08  0
------------------------------------------------------------------------------
group: yes
   vars n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 8 0.07 0.02   0.06    0.07 0.01 0.04 0.09  0.05 -0.25    -1.02 0.01
```

Analiza descriptivă a variabilei categoriale pentru urban arată că 76 nu sunt din zona urbană(metropolitan) iar restul de 8 se află în zona urbană.

Din output se poate observa că rata criminalității este mai mare în zona urbană(0,07%) față de cea rurală, unde rata este egală cu 0,03%.

B. Analiza ratei de criminalitate în functie de regiune.

```
> describeBy(criminalitate_v2$criminalitate_loc,group=criminalitate_v2$regiune,digits= 4)

 Descriptive statistics by group
group: central
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 33 0.04 0.02   0.03    0.04 0.02 0.01 0.09  0.08 0.74    -0.34  0
------------------------------------------------------------------------------
group: other
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 34 0.04 0.01   0.03    0.03 0.02 0.02 0.08  0.06 0.82     0.10  0
------------------------------------------------------------------------------
group: west
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 17 0.02 0.01   0.02    0.02 0.01 0.01 0.04  0.03 0.15    -1.13  0
```

Analiza descriptivă a variabilei categoriale pentru regiune prezintă faptul că cele mai multe comitate se afla în alta regiune față de cea centrala și cea de vest.

Din output se observă că media ratei criminalității este egală cu 0,04 atât pentru regiunile din vest cât și pentru cele din centru, iar cea mai mica rata se găsește în alte regiuni cu un procent de 0,2. În schimb rata maximă se regăsește în zona centrală.

C. Analiza ratei de criminalitate în functie de probabilitatea de condamnare.

```
> describeBy(criminalitate_v2$criminalitate_loc,group=criminalitate_v2$prob_arest.cat,digits= 4)

 Descriptive statistics by group
group: mica
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 49 0.04 0.02   0.03    0.04 0.02 0.02 0.09  0.07 0.66    -0.52  0
------------------------------------------------------------------------------
group: medie
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 31 0.03 0.01   0.03    0.03 0.01 0.01 0.05  0.04 0.13    -0.72  0
------------------------------------------------------------------------------
group: mare
   vars n mean   sd median trimmed  mad  min  max range skew kurtosis se
X1    1 4 0.02 0.01   0.01    0.02 0.01 0.01 0.03  0.02 0.37    -2.03  0
------------------------------------------------------------------------------
group: foarte mare
NULL
```

Cum era de așteptat, cea mai mare valoare a ratei de criminalitate se regăsește în comitatele unde probabilitatea de a fi condamnat este cea mai mica iar cea mai scăzuta rată este regăsită în comitatele unde probabilitatea de a fi condamnat este foarte mare, adică între.

```
> tapply(criminalitate_v2$criminalitate_loc, list(criminalitate_v2$prob_arest.cat, criminalitate_v2$regiune), me
an)
             central      other       west
mica      0.04490150 0.03896904 0.02579282
medie     0.02749623 0.03122597 0.02175888
mare      0.01738070 0.02724600 0.00904490
foarte mare        NA         NA         NA
```

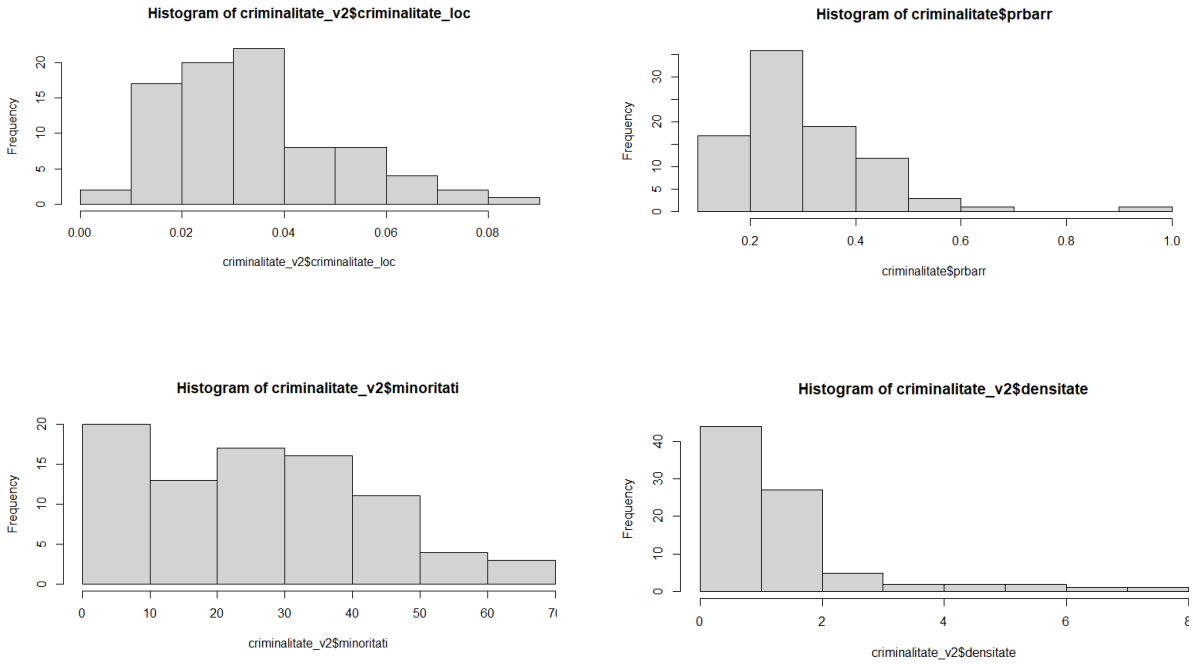
Din analiza mediei variabilei rata de criminalitate având în vedere grupurile probabilitatea de condamnare și tipul regiunii, am constatat că cele mai mari rate ale criminalității se regăsesc unde probabilitatea de condamnare este mica, pentru toate regiunile.

Cele mai mici rate se gasesc acolo unde probabilitatea de a fi condamnat este foarte mare.

3.2. Analiza grafica a variabilelor numerice si nenumerice

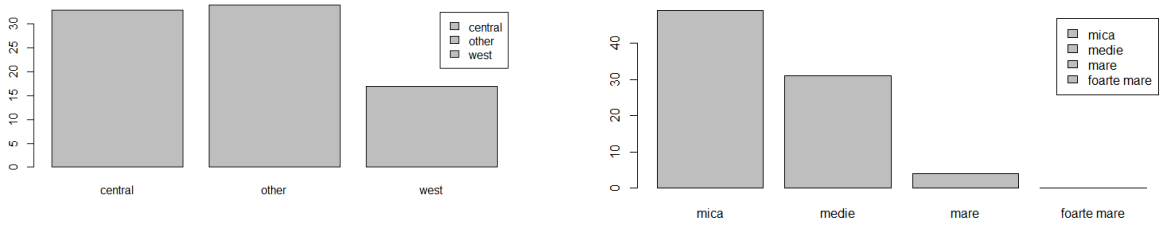
În acest subcapitol se vor analiza atât variabilele numerice cât și cele categoriale pe cale grafică, cu ajutorul funcțiilor boxplot și plot.

3.2.1. Analiza grafică a variabilelor numerice



Din histogramele de mai sus, se poate observa că distribuția celor 4 variabile numerice, rata criminalității, procentul de minorități, probabilitatea de arest și densitatea persoanelor/km², sunt asimetrice la dreapta.

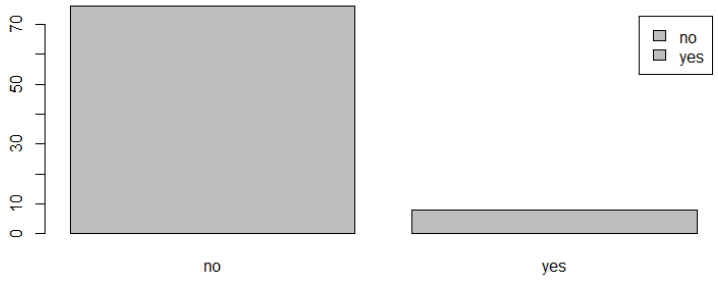
3.2.2. Analiză grafică variabile nenumerice



Se observă că cele mai multe comitate au o probabilitate de condamnare medie, fiind urmate

de cele cu probabilitate mare iar în cele din urma cu o probabilitate mică și foarte mare.

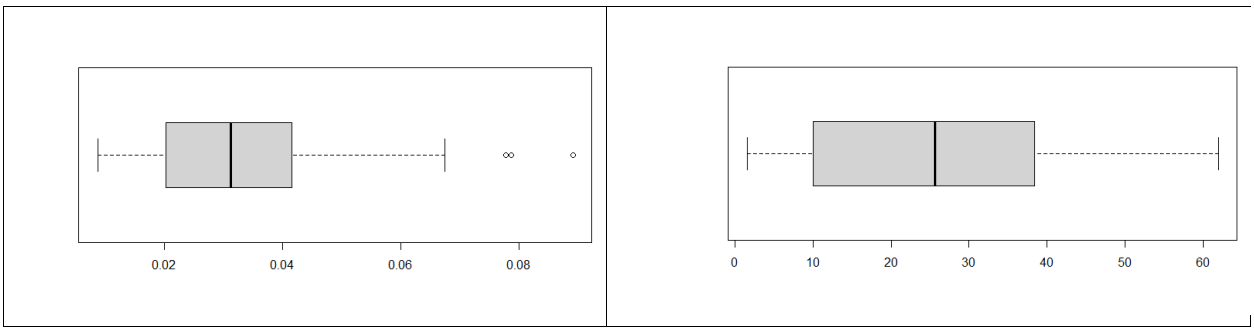
În ceea ce privește regiunea se remarcă faptul că cele mai multe unități din bază se regăsesc în altă regiune față de vest sau regiunea centrală. Cele mai puține unități din bază sunt regăsite în regiunea de vest.



După cum se poate observa și din grafic majoritatea comitatelor prezente în bază se află în zona rurală și foarte puține se regăsesc în zona urbană.

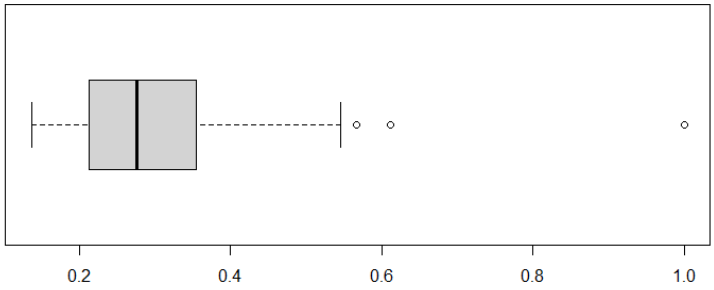
3.3. Identificarea valorilor extreme

Pentru a identifica valorile extreme pentru variabilele studiate se va folosi graficul Box plot.

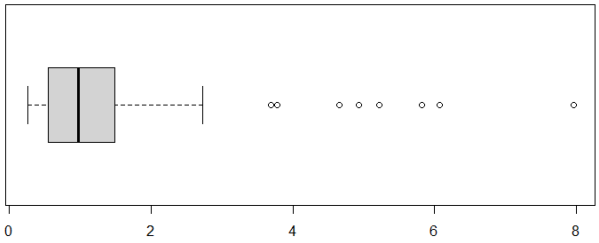


Din Box plot-ul ratei criminalității (prima figura) se pot observa 3 valori extreme apropiate care nunecesită eliminare.

Pentru procentul de minorități nu există nicio valoare extremă.



În ceea ce privește probabilitatea de arest se poate observa din box plot că există doar trei valori extreme însă acestea sunt apropiate și nu trebuie eliminate din bază.



Pentru densitate se observă mai multe valori extreme insa în urma analizei corelației s-a luat decizia de a nu se elimina outlierii întrucât există corelații mai puternice atunci când valorile extreme sunt prezente.

4. Analiza statistica a variabilelor categoriale

În acest capitol se vor tabela datele, se va realiza analiza de asociere și în cele din urmă analiza de concordanță.

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

```
> table(criminalitate_v2$prob_arest.cat,criminalitate_v2$urban)

             no yes
  mica       41   8
  medie      31   0
  mare        4   0
  foarte mare 0   0
```

Din tabel se poate observa că nu există comitate din zona urbană care sa aibă probabilitate de condamnare mare sau foarte mare, 8 comitate din zona metropolitană au o probabilitate mică.

În ceea ce privește zona rurală, 31 de observații se regăsesc la probabilitate medie, 4 la mare, 41 la mică și deloc la foarte mare.

```
> prop.table(contin)#tabel de frecvente

                    no       yes
  mica        0.48809524 0.09523810
  medie       0.36904762 0.00000000
  mare        0.04761905 0.00000000
  foarte mare 0.00000000 0.00000000
```

Cele mai multe unități din baza de date sunt din zona rurală și au o probabilitate de condamnare mica(48,8%), această categorie este urmată de cele din zona rurală cu probabilitate medie(36,9%), iar în cele din urmă sunt cele cu probabilitate mare(4,8%). Cele mai puține unități din baza se afla în zona urbană, unde nu există unități cu probabilitate mare, medie sau foarte mare, ci doar mic. Cea mica este regăsită în proporție de 9,5%

```
> margin.table(contin, 1) # Frecvente marginale pentru prob_arest.cat

      mica       medie        mare foarte mare
        49          31           4           0
> margin.table(contin, 2) # Frecvente marginale pentru urban

 no yes
 76   8
```

Din tabelul cu frecvențe marginale se poate observa că cele mai multe unități din bază au o probabilitate mica de arest fiind urmate de cele cu probabilitate medie și în cele din urmă cele cu probabilitate mare.

```
> prop.table(contin, 1) # frecvente conditionate dupa prob_arest.cat(pe linie)

                    no       yes
  mica         0.8367347 0.1632653
  medie        1.0000000 0.0000000
  mare         1.0000000 0.0000000
  foarte mare
> prop.table(contin, 2)

                     no        yes
  mica         0.53947368 1.00000000
  medie        0.40789474 0.00000000
  mare         0.05263158 0.00000000
  foarte mare  0.00000000 0.00000000
```

Frecvențe condiționate pe rând(probabilitate condamnare)

Din totalul comitatelor ce fac parte din categoria probabilitate mică de arest, 16,32% sunt din zona urbană și 83,67% din zona rurală.

Frecvențe condiționate pe coloană

Din totalul comitatelor ce fac parte din zona urbană, 100% au o probabilitate mică de condamnare. În ceea ce privește procentul din totalul zonelor rurale, 54% dintre comitate au o probabilitate mică, 40,80% au o probabilitate medie si 5,26% au o probabilitate mare. Deci din totalul comitatelor din zona urbană cele mai multe au o probabilitate mică si de asemenea si cele din zona rurală.

```
> addmargins(contin) # frecvente absolute marginale

             no yes Sum
  mica       41   8  49
  medie      31   0  31
  mare        4   0   4
  foarte mare 0   0   0
  Sum        76   8  84
> addmargins(prop.table(contin)) # frecvente relative partiale si marginale

                    no        yes        Sum
  mica         0.48809524 0.09523810 0.58333333
  medie        0.36904762 0.00000000 0.36904762
  mare         0.04761905 0.00000000 0.04761905
  foarte mare  0.00000000 0.00000000 0.00000000
  Sum          0.90476190 0.09523810 1.00000000
```

Din primul output(tabel frecvente absolute marginale) se observă că 49 comitate au o probabilitate mică de arest, 31 o probabilitate medie, 4 o probabilitate mare si niciuna o probabilitate foarte mare. În ceea ce privește zona, 8 unități se află în zona urbană și 76 în zona rurală.

Cele mai multe unități din bază au o probabilitate de condamnare mica (58,33%), fiind urmate de cele cu o probabilitate medie (40%) iar în cele din urmă cele cu probabilitate mare(5%). 90,5% din comitate se află în zona rurală și 9,5 in zona meptropolitană.

4.2. Analiza de asociere

A. Analiza de asociere între regiune și probabilitatea de condamnare

```
Number of cases in table: 84
Number of factors: 2
Test for independence of all factors:
        Chisq = 4.82578237, df = 6, p-value = 0.566344344
```

- **Formularea ipotezelor**

  H0: între regiune și probabilitatea de condamnare nu există o asociere semnificativă

  H1: între regiune și probabilitatea de condamnare există o asociere semnificativă
- **Regula de decizie**

  Sig < 0.05, se respinge H0, cu un risc asumat de 5%

  Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%
- **P-value** = 0.56 > 0.05, nu se respinge H0
- **Interpretare**: pentru o probabilitate de 95%, nu există o asociere semnificativă între regiune și probabilitatea de condamnare.

B. Analiza de asociere între trei variabile

```
Number of cases in table: 84
Number of factors: 3
Test for independence of all factors:
        Chisq = 29.55693547, df = 17, p-value = 0.0297230972
        Chi-squared approximation may be incorrect
```

- **Formularea ipotezelor**

H0: între regiune, zonă și probabilitatea de condamnare nu există o asociere semnificativă

H1: între cele trei variabile există o asociere semnificativă
- **Regula de decizie**

  Sig < 0.05, se respinge H0, cu un risc asumat de 5%

  Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%
- **P-value** = 0.029 < 0.05, se respinge H0
- **Interpretare:** Cu un risc asumat de 5%, există o asociere semnificativă între cele trei variabile

C. Analiza de asociere între zona urbană și probabilitatea de condamnare

```
Number of cases in table: 84
Number of factors: 2
Test for independence of all factors:
        Chisq = 16.98836917, df = 3, p-value = 0.000710645684
```

- **Formularea ipotezelor**

  H0: între zonă și probabilitatea de condamnare nu există o asociere semnificativă

  H1: între zonă și probabilitatea de condamnare există o asociere semnificativă

- **Regula de decizie**

  Sig < 0.05, se respinge H0, cu un risc asumat de 5%

  Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

- **P-value** = 0.0007 < 0.05, se respinge H0
- **Interpretare**: Cu un risc asumat de 5%, există o asociere semnificativă între zonă și probabilitatea de condamnare.

4.3. Analiza de concordanță

```
> chisq.test(table(criminalitate_v2$prob_arest.cat))

        Chi-squared test for given probabilities

data:  table(criminalitate_v2$prob_arest.cat)
X-squared = 76.857, df = 3, p-value < 2.2e-16
```

- **Formularea ipotezelor**

H0: între cele două distribuții nu există diferențe semnificative (există concordanță de structură)

H1: cele două distribuții diferă semnificativ (nu există concordanță de structură)

- **Testul** folosit este Chi-square (Chi patrat)
- **Regula de decizie**

Sig < 0.05, se respinge H0, cu un risc asumat de 5%

Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

- **Interpretare**: Sig= 2.2e-16 <0,05 deci pentru un risc asumat de 5% se va respinge ipoteza nulă. Așadar nu există concordanță de structură.

5. Analiza de regresie și corelație

5.1. Analiza de corelație

Matricea corelațiilor prezintă valorile coeficienților de corelație dintre rata criminalității și celelalte 3 variabile numerice și aceasta arată astfel:

```
> cor(crime)
                rata_crime      prob_arest        densitate       minoritati
rata_crime  1.000000000000 -0.467204800243  0.7926618322397  0.1612101414192
prob_arest -0.467204800243  1.000000000000 -0.3767862763074  0.2247378328083
densitate   0.792661832240 -0.376786276307  1.0000000000000 -0.0814816180208
minoritati  0.161210141419  0.224737832808 -0.0814816180208  1.0000000000000
```

După cum se poate remarca și în matrice, coeficientul de corelație (0,792) arată existența unei legături puternice și directe între *rata criminalității* și *densitatea populației*, astfel cu cât aceasta crește cu atât crește și rata criminalității. În ceea ce privește relația dintre rata de crime și *probabilitatea de arest* se observă o legătură medie și inversă (-0,467) iar între rata de crime și *procentul de minorități* există o legătură slabă și directă. Între variabilele independente se poate remarca o corelație slabă și inversă între densitate și probabilitate de arest,

Testarea coeficientului de corelație

```
> cor.test(crime_v2$rata_crime, crime_v2$prob_arest)

        Pearson's product-moment correlation

data:  crime_v2$rata_crime and crime_v2$prob_arest
t = -4.7851, df = 82, p-value = 7.454e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6195426 -0.2809515
sample estimates:
       cor
-0.4672048
```

- **Formularea ipotezelor**

H0: ρ=0 (coeficientul de corelație nu are o valoare semnificativă la nivelul populației)

H1: ρ≠0 (coeficientul de corelatie are o valoare semnificativă la nivelul populației)

- **Testul folosit:** testul t
- **Regula de decizie**

Sig (p-value) < 0.05, se respinge H0, cu un risc asumat de 5%

Sig (p-value) > 0.05, nu se respinge H0, cu o probabilitate de 95%

- Decizie statistică: Având în vedere că p-value este 7.454e-06, valoare mai mică decât α asumat de 0,05, se va respinge ipoteza nulă cu un risc asumat de 5%. Coeficientul de corelație dintre rata de criminalitate și probabilitatea de arest are o valoare semnificativă la nivelul populației.

5.2. Analiza de regresie

În acest subcapitol vor fi realizate mai multe tipuri de modele de regresie, se vor analiza și se vor compara între ele.

5.2.1. Regresie liniară simplă și mutiplă

Variabila dependentă (Y) este reprezentată de rata de crime iar cea independentă (X) de densitate.

```
> lm(rata_crime~densitate,data=crime_v2)

Call:
lm(formula = rata_crime ~ densitate, data = crime_v2)

Coefficients:
  (Intercept)      densitate
0.02062539209  0.00940434542
```

Ecuația modelului de regresie: Yi= β0+ βi*xi+ ε

Rata crime = 0,020+0,009*densitatea

Interpretare: β0 = 0,020 de unde rexultă că atunci când valoarea densității ia valoarea 0, variabila dependentă, rata de crime, are o valoare medie de 0,02%.

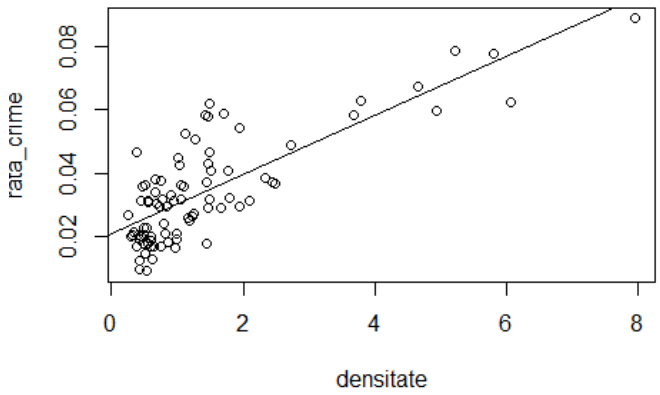
β1= 0,0094 arată că la o creștere a densității cu o pers/km^2, rata de criminalitate ,crește în medie, cu 0,0094 procente.

```
> summary(ls)

Call:
lm(formula = rata_crime ~ densitate, data = crime_v2)

Residuals:
          Min            1Q        Median            3Q           Max
-0.01692190403 -0.00734589362 -0.00303143237  0.00566734263  0.02705557370

Coefficients:
                Estimate     Std. Error  t value   Pr(>|t|)
(Intercept) 0.020625392094 0.001601076957 12.88220 < 2.22e-16 ***
densitate   0.009404345419 0.000798771413 11.77351 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0104113822 on 82 degrees of freedom
Multiple R-squared:  0.62831278,        Adjusted R-squared:  0.623780009
F-statistic: 138.615603 on 1 and 82 DF,  p-value: < 2.220446e-16
```

Din output se poate observa că atât constanta modelului cât și coeficientul variabilei independente sunt semnificativi statistic, ambii având probabilități mai mici decât 0,05.

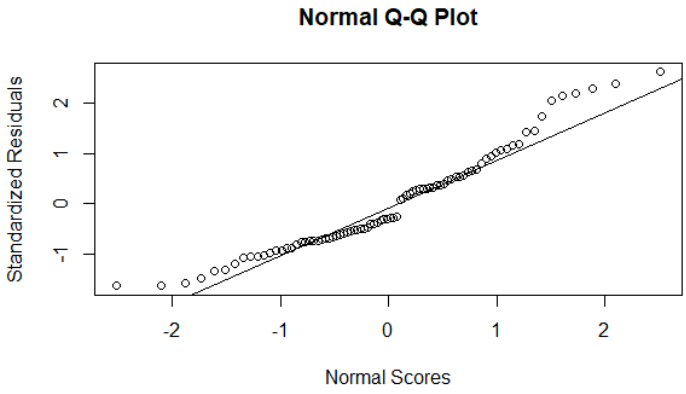
De asemenea, valoarea testului F.este mai mică decât 0,05 de unde rezultă că modelul este corect specificat statistic.

R^2 = 0.628=0.628%



Așadar 62,8% din variația totală a ratei criminalității este explicată de densitatea populației.

Punctele sunt adunate în jurul dreptei de regresie de unde rezulta o putere explivstivă mare a modelului.




```
> summary(ls.resid)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.016922 -0.007346 -0.003031  0.000000  0.005667  0.027056
```

Erorile modelului liniar simplu sunt normal distribuite iar media acestora este egală cu 0.

5.2.1. Regresie liniara multiplă

Variabila dependentă (Y) este reprezentată de rata de crime iar cele independente de densitate,probabilitatea de arest și procentul de minorități.

```
Call:
lm(formula = rata_crime ~ densitate + prob_arest + minoritati,
    data = crime_v2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.016048 -0.005287 -0.000081  0.004506  0.023354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.70e-02   3.75e-03    7.18  3.1e-10 ***
densitate    8.51e-03   7.37e-04   11.54  < 2e-16 ***
prob_arest  -4.12e-02   1.01e-02   -4.08  0.00011 ***
minoritati   2.83e-04   6.02e-05    4.70  1.1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0089 on 80 degrees of freedom
Multiple R-squared:  0.735,     Adjusted R-squared:  0.725
F-statistic: 73.9 on 3 and 80 DF,  p-value: <2e-16
```

Rata crime = 0.027+ 0.00851*densitate -0.0412*probabilitate arest+ 0.000283* minorități

Atunci când toate variabilele independente sunt egale cu 0, variabila rata de criminalitate este, în medie, de 0.027 procente.

La o creștere cu o unitate a densității, rata criminalității crește cu 0.00851% atunci când probabilitatea de arest și minoritățile rămân constante.

La o creștere cu un procent a probabilității de arest, rata criminalității scade cu 0.0412% atunci când densitatea și minoritățile rămân constante.

La o creștere cu un procent a minorității, rata criminalității crește cu 0.000283% atunci când densitatea și probabilitatea de arest rămân constante.

De asemenea se poate observa că toți coeficienții modelului de regresie sunt semnificativi statistici, toți având probabilități mai mici decât 0,05.

De asemenea, valoarea testului F pentru model este mai mică decât 0,05 de unde rezultă că modelul este corect specificat statistic.

R^2 = 0.735=73,5%

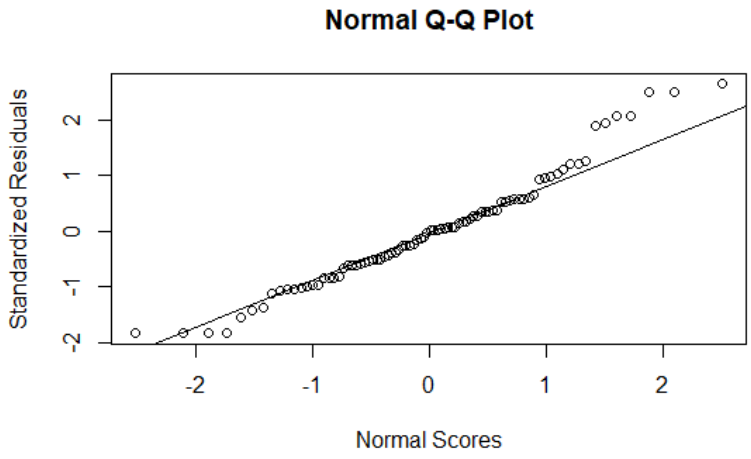
Așadar 73,5% din variația totală a ratei criminalității este explicată de densitatea populației, procentul minorităților și probabilitatea de arest.

```
> lm.beta(lm)

Call:
lm(formula = rata_crime ~ densitate + prob_arest + minoritati,
    data = crime_v2)

Standardized Coefficients::
(Intercept)   densitate  prob_arest  minoritati
    0.00000     0.71762    -0.25928     0.27795
```

Factorul cu cel mai mare impact asupra ratei de criminalitate este densitatea care la o creștere cu o abatere standard duce la o creștere a ratei de criminalitate cu 0,77 abateri standard, în condițiile în care probabilitatea de arest și minoritățile rămân constante. Acest factor este urmat de minorități iar cel din urmă este probabilitatea de arest.

```
> summary(lm.resid)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.60e-02 -5.29e-03 -8.07e-05  0.00e+00  4.51e-03  2.34e-02
```

Din acest output se poate observa că media erorilor este egală cu 0.



Întrucât punctele se află în apropierea liniei există mari șanse ca erorile să fie normal distribuite.

5.2.2. Regresia neliniară

A fost creat un model de regresie polinomial de ordinul 2.(Quadratic).

```
Call:
lm(formula = rata_crime ~ densitate + I(densitate^2), data = crime_v2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.018276 -0.007717 -0.001153  0.005941  0.025484

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0172491  0.0024353   7.083 4.66e-10 ***
densitate       0.0137232  0.0024991   5.491 4.47e-07 ***
I(densitate^2) -0.0006780  0.0003723  -1.821   0.0723 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01027 on 81 degrees of freedom
Multiple R-squared:  0.6429,    Adjusted R-squared:  0.6341
F-statistic: 72.92 on 2 and 81 DF,  p-value: < 2.2e-16
```

Rata crime = 0,0172+0,0137*densitatea-0,0007*densitatea^2

Întrucat coeficientul beta1 este pozitiv iar beta2 este negativ, curba prezinta punct de maxim.

De asemenea, valoarea testului F.este mai mică decât 0,05 de unde rezultă că modelul este corect specificat statistic.

R^2 = 0.643=64,3%

Acest model are un coeficient R^2 mai mare față de cel linear simplu însă densitatea^2 nu este semnificativa din punct de vedere statistic întrucât sig=0,072, valoare mai mare decât 0,05.

```
Call:
lm(formula = rata_crime ~ log(densitate), data = crime_v2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0225475 -0.0095361 -0.0006851  0.0074301  0.0298906

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.033623   0.001187   28.32   <2e-16 ***
log(densitate) 0.017147   0.001565   10.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01088 on 82 degrees of freedom
Multiple R-squared:  0.5942,    Adjusted R-squared:  0.5892
F-statistic: 120.1 on 1 and 82 DF,  p-value: < 2.2e-16
```

Model logaritmic(cu variabila independentă logaritmată)

Rata criminalitate=0,0336+0,0171*ln(densitate)

Interpretare: β0 = 0,0336 . Pentru o valoare densității de o pers/km^2, estimăm o valoare medie a ratei de criminalitete de 0,0336%.

β1= 0,0171

Atunci când densitatea crește cu 1%, rata criminalității crește în medie cu 0,0171/100=0.000171%.

B1 arată variația medie absolută a variabilei dependente la o variație relativa a lui X cu o unitate.

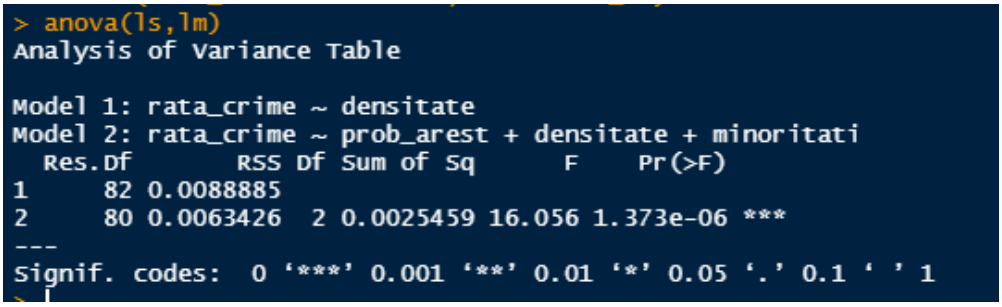
Din output se poate observa că atât constanta modelului cât și coeficientul variabilei independente sunt semnificativi statistic, ambii având probabilități mai mici decât 0,05.

De asemenea, valoarea testului F.este mai mică decât 0,05 de unde rezultă că modelul este corect specificat statistic.

R^2 = 0.59=59% Așadar 59% din variația totală a ratei criminalității este explicată de densitatea populației logaritmata. Deci coeficientul R^2 a scăzut față de modelul liniar si cel parabolic, acesta având o valoare mai scăzută decât cele două modele.

5.2.3. Comparare a 2 modele de regresie și alegerea celui mai bun model

În acest subcapitol se vor compara mai multe modele de regresie și se va alege cel mai bun model. Am început prin a compara modelul liniar simplu cu cel multiplu cu ajutorul funcției ANOVA, de unde a rezultat următorul output:



H0: modelul mai complex (cu mai multi parametri-modelul multiplu) nu este semnificativ mai bun decât modelul mai simplu(modelul liniar simplu)

H1: modelul mai complex (cu mai multi parametri) este semnificativ mai bun decat modelul mai simplu

Interpretare: p-value=0.000001<0,05 ceea ce înseamnă că vom respinge ipoteza nulă, un un risc asumat de 5%.

Modelul complex, ce include si variabile probabilitatea de arest si minoritățile, este semnificativ mai bun decât modelul simplu.

Comparare model liniar simplu cu modelul cu variabila independentă logaritmată

Dacă e să comparăm cele două modele putem observa că ambele au valoarea testului F mai mică decât 0,05 de unde rezultă că modelele sunt corect specificate statistic. De asemenea, amebele au coeficienții variabilelor semnificativ din punct de vedere statistic. Însă, atunci cân comparăm R^2 observăm că modelul liniar simplu (0.628) are o valoare mai mare față de cel logaritmic(0.59), de unde rezultă că modelul liniar simplu este mai bun decât modelul logarithmic.

6. Estimarea si testarea mediilor

6.1. Estimarea mediei prin interval de incredere

```
> t.test(crime_v2$rata_crime,conf.level = 0.95)

        One Sample t-test

data:  crime_v2$rata_crime
t = 18.309, df = 83, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.03022556 0.03759278
sample estimates:
 mean of x
0.03390917
```

Cu o probabilitate de 95% se poate afirma că intervalul de încredere al variabilei rata de criminalitate este acoperit de valorile [0,03022;0,03759].

6.2. Testarea mediilor populației

6.2.1. Testarea unei medii cu o valoare fixă

```
> t.test(crime_v2$rata_crime,mu=0.04)

        One Sample t-test

data:  crime_v2$rata_crime
t = -3.2887, df = 83, p-value = 0.001478
alternative hypothesis: true mean is not equal to 0.04
95 percent confidence interval:
 0.03022556 0.03759278
sample estimates:
 mean of x
0.03390917
```

6

- **Ipoteze statistice**

H0=media este egală cu 0.04
H1=media este diferită de 0.04

- **Regula de decizie**

Sig < 0.05, se respinge H0, cu un risc asumat de 5%
Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

- **Decizie statistica:** P-value = 0,0014 < 0.05, se respinge H0

- **Interpretare**: Cu o probabilitate de 95%, se va respinge ipoteza nulă , astfel, media variabilei ratei de criminalitate, este diferită de valoarea 0,04%.

6.2.2. Testarea diferentei dintre doua medii

```
> bartlett.test(rata_crime~urban,crime_v2)

        Bartlett test of homogeneity of variances

data:  rata_crime by urban
Bartlett's K-squared = 0.49044, df = 1, p-value = 0.4837
```

- **Ipoteze statistice**

H0: variațele sunt egale

H1: variațele nu sunt egale

- **Regula de decizie**

Sig < 0.05, se respinge H0, cu un risc asumat de 5%

Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

- **Decizie statistica:** P-value = < 0.05, se respinge H0
- **Interpretare:** p-value= 0.48>0.05, nu se respinge H0, deci ipoteza de omogenitate a variațelor este verificată (variațele celor două grupuri sunt egale).

```
> t.test(crime_v2$rata_crime~crime_v2$urban,var.equal=TRUE)

        Two Sample t-test

data:  crime_v2$rata_crime by crime_v2$urban
t = -6.9872, df = 82, p-value = 6.792e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04511115 -0.02511669
sample estimates:
 mean in group no mean in group yes
      0.03056499       0.06567891
```

- **Ipoteze statistice**

H0:$\mu_1-\mu_2=0$

H1: $\mu_1-\mu_2\neq0$

- **Regula de decizie**

Sig < 0.05, se respinge H0, cu un risc asumat de 5%

Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

- **Decizie statistica:** P-value = 6.792e-10< 0.05, se respinge H0

- **Interpretare:** p-value= 6.792e-10<0.05 deci se respinge H0 pentru un risc asumat de 5%, rata criminalității în zona urbană diferă semnificativ de rata criminalității al comitatelor din zona rurală.

6.2.3. Testarea diferenței dintre trei sau mai multe medii

Pentru a testa diferența dintre trei sau mai multe medii va fi folosită ANOVA. Mai întâi va fi testată diferența ratei criminalității în funcție de regiune(centrala, vest, alta) iar apoi în funcție de probabilitatea de condamnare(mica, medie, mare, foarte mare).

Intai de toate trebuie sa testam omogenitatea celor trei grupuri.

```
> bartlett.test(rata_crime~prob_cond.Cat,crime_v2)

        Bartlett test of homogeneity of variances

data:  rata_crime by prob_cond.Cat
Bartlett's K-squared = 8.3349, df = 3, p-value = 0.03957
```

- **Ipoteze statistice**

H0: variațele sunt egale

H1: variațele nu sunt egale

- **Regula de decizie**

Sig < 0.05, se respinge H0, cu un risc asumat de 5%

Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

Decizie statistica: P-value = < 0.05, se respinge H0

Interpretare: p-value= 0.03<0.05, se respinge H0, deci ipoteza de omogenitate a variațelor nu este verificată (variațele celor două grupuri nu sunt egale).

```
Analysis of Variance Table

Response: rata_crime
              Df   Sum Sq    Mean Sq F value    Pr(>F)
prob_cond.Cat  3 0.004777 0.00159233  6.6565 0.0004534 ***
Residuals     80 0.019137 0.00023921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Ipoteze statistice**

H0: între rata de criminalitate ale celor patru grupuri nu există diferențe semnificative

H1: între rata de criminalitate ale celor patru grupuri există diferențe semnificative

- **Testul folosit:** F
- **Valoarea calculată a testului:** Fcalc=6.6565

- **Regula de decizie**

Sig < 0.05, se respinge H0, cu un risc asumat de 5%

Sig > 0.05, nu se respinge H0, cu o probabilitate de 95%

- **Decizie statistica:** P-value = 0.00476< 0.05, se respinge H0
- **Interpretare:** p-value=0.0004534<0.05, se respinge H0, deci pentru un risc asumat de 5%, între rata de criminalitate ale celor patru grupuri există diferențe semnificative. Rata de criminalitate este semnificativ influențată de probabilitatea de condamnare.

7. Concluzii

În această lucrare am analizat fenomenul de criminalitate al comitatelor din North Carolina, Statele Unite. Am ales să analizez datele pentru anul 1982. Prin noţiunea de „criminalitate” se înţelege săvârşirea de crime; totalitatea infracţiunilor săvârşite pe un teritoriu, într-o anumită perioadă.

Cu ajutorul estimării modelului de regresie liniară multiplă am constatat că criminalității este explicată de densitatea umană, proporția minorităților și probabilitatea de arest.

Cele mai multe comitate au o probabilitate de condamnare medie, 41,6%, regiunile fiind altele față de vest sau regiunea centrală din în zonele rurale.

Rata de criminalitate este semnificativ influențată de regiune și de probabilitatea de condamnare.

De asemenea, în urma testării ipotezelor privind erorile s-a observat că toate ipotezele sunt respectate, mai ales pentru procentul de minorități la care nu există nicio valoare extremă.

Din matrices corelațiilor s-a observat existența unei legături puternice și directe între rata criminalității și densitatea populației (0,79), astfel că cu cât aceasta crește cu atât crește și rata criminalității. În ceea ce privește relația dintre rata de crime și probabilitatea de arest se observă o legătură medie și inversă (-0,47) iar între rata de crime și procentul de minorități există o legătură slabă și directă. Între variabilele independente se poate remarca o corelație slabă și inversă între densitate și probabilitate de arest.

Am testat intervalul de încredere al variabilei rata de criminalitate și astfel am aflat că cu o probabilitate de 95% se poate afirma că intervalul de încredere al variabilei rata de criminalitate este acoperit de valorile [0,03022; 0,03759].

Cu ajutorul testului ANOVA am putut observa dacă sunt diferențe semnificative între mediile variabilelor studiate, adică zona urbana și probabiltatea de arest.

În concluzie, se poate afirma că criminalitatea este un fenomen complex, ușor de influențat de o multitudine de factori externi.