Software development for data analysis project

Project theme - criminality

Chivu Monica Gabriela

Catalog

Software development for data analysis project 1
Project theme - criminality 1
Description of the database 3
Intentional Homicide (country and largest city)database 3
Sexual Violence (Rape and Sexual Assault) database 3
Sexual Assault database 3
Theft database 3
Crime index and safety index database 3
Romania Crime Rate 1990-2022 database 3
Introduction 4
Processing performed on data 4
Reading from excel and csv files 4
Writing in excel files 4
Merging two files for creating one database 4
Calculating the sum by country for all years and creating a new column for it 5
Use the numpy library 6
Use the pandas library 6
Graphs 7
-Pie chart Belgium types of crimes 7
- Plotting the criminality index and safety index 7
- Sexual violence and rape correlation in year 2020 8
- Fifth graph : Romania Crime Rate & Statistics 1990-2022 8
Exploratory Factor Analysis 9
- Correlogram of Kaiser-Meyer-Olkin indices 9
-Correlogram of correlation factors 10
Table of figures 12
Bibliography 13

Description of the database

Intentional Homicide (country and largest city)database

Definition: Unlawful death inflicted upon a person with the intent to cause death or serious injury. Data on intentional homicide should also include murder, honour killing, serious assault leading to death, death as a result of terrorist activities, dowry-related killings, femicide, infanticide, voluntary manslaughter, extrajudicial killings, killings caused by excessive use of force by law enforcement/state officials. It should exclude death due to legal interventions, justifiable homicide in self-defence, attempted intentional homicide, homicide without the element of intent is non-intentional homicide, non-negligent or involuntary manslaughter, assisting suicide or instigating suicide, illegal feticide, euthanasia.

Sexual Violence (Rape and Sexual Assault) database

Definition: Unwanted sexual act, attempt to obtain a sexual act, or contact or communication with unwanted sexual attention without valid consent or with consent as a result of intimidation, force, fraud, coercion, threat, deception, use of drugs or alcohol, or abuse of power or of a position of vulnerability. It should exclude acts of abuse of a position of vulnerability, power or trust, or use of force or threat of force, for profiting monetarily, socially or politically from the prostitution or sexual acts of a person, coercion, prostitution offences, pornography offences and other acts against public order sexual standards such as incest not amounting to rape and exhibitionism, assaults and threats, slavery and exploitation not amounting to injurious acts of a sexual nature, TiP(trafficking in human beings) for sexual exploitation, harassment and stalking.

In practice sexual violence figures are the sum of rape and sexual assault.

Sexual Assault database

Definition: Sexual violence not amounting to rape. It includes an unwanted sexual act, an attempt to obtain a sexual act, or contact or communication with unwanted sexual attention not amounting to rape. It also includes sexual assault with or without physical contact including drug-facilitated sexual assault, sexual assault committed against a marital partner against her/his will, sexual assault against a helpless person, unwanted groping or fondling, harassment and threat of a sexual nature

Theft database

Definition: Unlawfully taking or obtaining of property with the intent to permanently withhold it from a person or organization without consent and without the use of force, threat of force or violence, coercion or deception.“Theft” excludes burglary, housebreaking and robbery, which are recorded separately. Excludes possession of stolen goods or money; receiving, handling, disposing, selling or trafficking stolen goods; using stolen parts for producing other goods; concealment of stolen goods, obtaining money or other benefit or evading a liability through deceit or dishonest conduct,robbery, property damage , theft after unauthorized access to premises, theft of intellectual property, identity theft.“Theft” excludes burglary, housebreaking and robbery, which are recorded separately.

Crime index and safety index database

Romania Crime Rate 1990-2022 database

Introduction

Crime data recorded from various sources are widely used for a range of purposes. Police forces use data about calls for police services and police-recorded offenses to identify geographic areas with high crime rates and design targeted policing strategies. Policymakers use estimates of crime obtained from victim surveys and police records to evaluate the impact of crime prevention policies and other legal and socio-economic changes.

Detailed information that pinpoints the types and locations of the crime problems of a jurisdiction or organization is the key to the design of a prevention program.

Crime statistics must be gathered as a first step in the design of a prevention program.

I chose this topic for my project because I am interested in the topic of crime prevention and understanding the complexity of this.

Processing performed on data

Reading from excel and csv files

```
# read the Project_python_crime.xlsx file
theftData = pd.read_excel('Project_python_crime.xlsx', sheet_name='Theft',na_values=':')
print(theftData)

rapeData = pd.read_excel('Project_python_crime.xlsx', sheet_name='Rape',na_values=':')
# print(rapeData)

sexualViolenceData = pd.read_excel('Project_python_crime.xlsx', sheet_name='Sexual violence',na_values=':')
# print(sexualViolenceData)
```

Figure 1 - reading from .xlsx and .csv files

Writing in excel files

```
homicideData.to_excel("HomData.xlsx")
```

Figure 2 - Writing in .xlsx files

Merging two files for creating one database

```
homicideData1=pd.read_excel("2011-2016.xlsx",na_values=':')
homicideData2=pd.read_excel("2016-2020.xlsx",na_values=':')

homicideData=pd.merge(homicideData1,homicideData2,on=["TIME"])
print(homicideData)
```

Figure 3- Merging two files for creating one database

Calculating the sum by country for all years and creating a new column for it

In the HomData.xlsx

```
# taking the values from my table
numericalValuesHomicide=homicideData.values
# deleting first column to only have the numerical values
numericalValuesHomicide=np.delete(arr=numericalValuesHomicic
print(numericalValuesHomicide)
dfHomicide = pd.DataFrame(numericalValuesHomicide)
# dfHomicide.fillna(0)
dfHomicide = dfHomicide.replace(np.nan, 0)
homicideData['sum']=dfHomicide.sum(axis=1)
```

Figure 4 - Calculating the sum by country for all years and creating a new column for it

	A	B	C	D	E	F	G	H	I	J	K	L	M
		TIME	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	sum
0	0	Belgium	214	206	204	210	231	175	198	190	146	142	1916
1	1	Bulgaria	128	141	109	112	126	79	95	92	80	66	1028
2	2	Czechia	83	95	90	81	88	65	40	55	81	57	735
3	3	Denmark	49	43	41	62	52	53	61	54	53	51	519
4	4	Germany (	689	619	623	645	655	747	738	632	586	719	6653
5	5	Estonia	65	63	52	41	50	33	29	25	23	37	418
6	6	Ireland	42	52	51	53	30	34	39	39	35	34	409
7	7	Greece	184	165	141	105	86	81	78	94	76	73	1083
8	8	Spain	385	364	302	323	302	294	307	289	333	298	3197
9	9	France	874	814	802	765	880	884	813	779	861	879	8351
10	10	Croatia	49	51	46	36	37	44	46	24	33	40	406
11	11	Italy	552	530	506	487	471	404	376	359	317	285	4287
12	12	Cyprus	8	19	11	10	12	11	7	14	13	15	120
13	13	Latvia	69	97	69	77	81	111	109	101	91	93	898
14	14	Lithuania	189	181	172	155	168	142	113	97	84	99	1400
15	15	Luxembou	4	3	2	4	5	5	2	3	4	2	34
16	16	Hungary	142	113	138	129	202	91	85	83	58	77	1118
17	17	Malta	3	10	6	6	4	5	9	6	4	6	59
18	18	Netherlanc			147	144	120	108	158	119	125	121	1042
19	19	Austria	80	88	63	43	42	49	61	73	74	54	627
20	20	Poland	449	377	296	282	287	256	278	265	205	261	2956
21	21	Portugal	114	122	144	92	100	66	76	81	73	81	949
22	22	Romania	335	378	336	298	291	247	256	267	255	256	2919
23	23	Slovenia	17	14	12	17	20	10	19	10	10	11	140
24	24	Slovakia	96	75	78	72	48	60	80	67	76	63	715
25	25	Finland	110	88	89	88	82	74	67	86	88		772
26	26	Sweden	81	68	87	87	112	106	113	108	111	124	997
27	27	Iceland	3	1	1	2	3	1	3	2	1	3	20
28	28	Liechtenst	0	0	0	1	0	0	1	1	0	1	4

Sheet1

Use the numpy library

```
#4th graph - PIE belgium theft years 2011-2020
belgium=df[df['Country']=='Belgium']
print(belgium)

y = np.asarray(belgium.values)
ynorm=y[1:].flatten()
# print(type(y))
print(ynorm)
print(ynorm.ndim)
valuesY=np.array([242399,237921,224109, 208443, 246853, 232504, 223563, 221955, 223946, 155760])
mylabels = np.array([2011,2012,2013,2014,2015,2016,2017,2018,2019,2020])
# print(belgium.columns)
# mylabels=belgium.columns[1:]
myl=list(mylabels)
print(mylabels)
# mylabels = np.asarray(belgium.columns)
figureObject, axesObject = plotter.subplots()

# Draw the pie chart

axesObject.pie(valuesY,

               labels=mylabels,

               autopct='%1.2f',
```

Figure 5 - Use the numpy library

Use the pandas library

```
# First graph

df=pd.read_excel("TheftData2.xlsx")
plt.figure(figsize=(8,5))

#PLOT the number of thefts by country in 2020 and 2016
plt.scatter(df['Country'],df['2020'])
plt.scatter(df['Country'],df['2016'])
# plt.show()
```

Figure 6 - Use the pandas library

Graphs

-Pie chart Belgium types of crimes

- to compare the sizes of categories to the entire dataset
- the most common type of crime in Belgium are the robberies
- the most uncommon type of crime in Belgium are the vehicle thefts

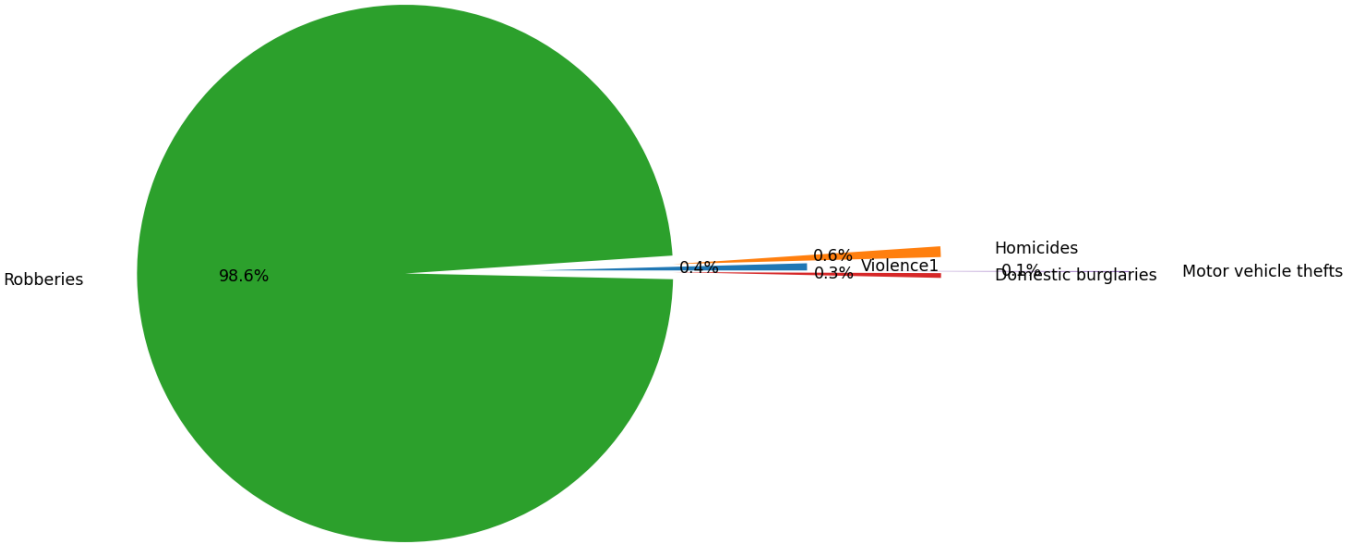

Figure 7 - Pie chart Belgium types of crimes

- Plotting the criminality index and safety index

The Crime Index is an annual index of crimes based on the occurrences of murder and non-negligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny (theft), and motor vehicle theft. This means that when the crime index is high, the country is not safe.
This graph shows that the safest countries have a very low crime rate and vice versa.

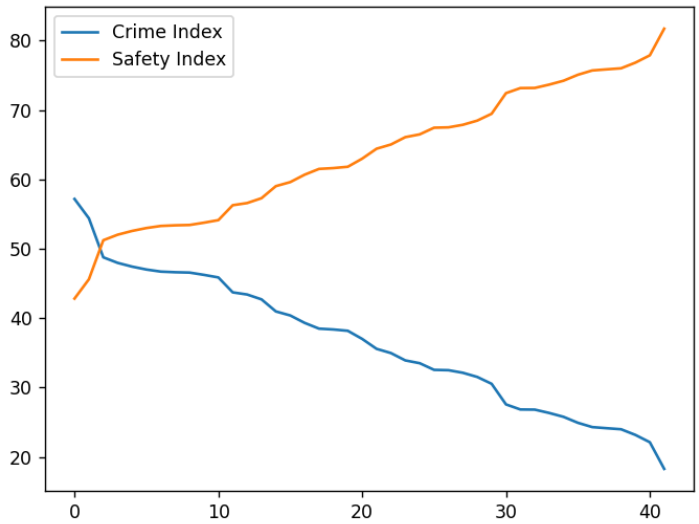

Figure 8 - Plotting the criminality index and safety index

- Sexual violence and rape correlation in year 2020

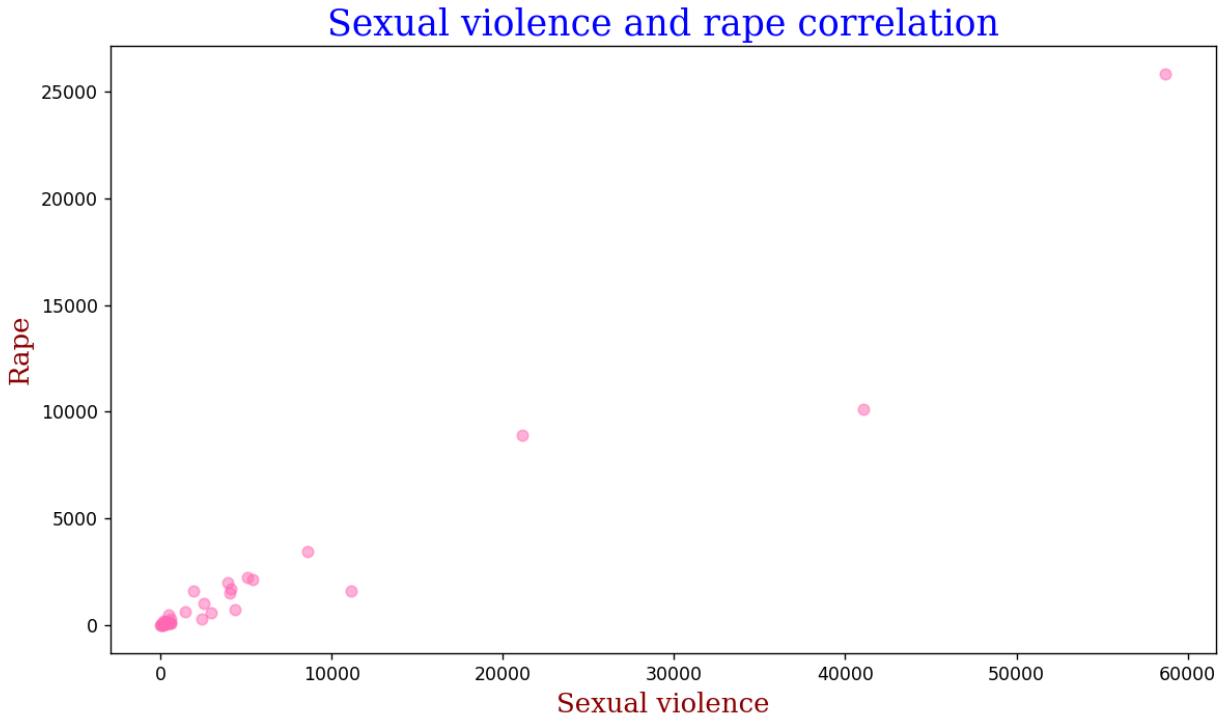

Figure 9 - Sexual violence and rape correlation in year 2020

- Fifth graph : Romania Crime Rate & Statistics 1990-2022

Romania crime rate & statistics for 2020 was 1.46, a 15.93% increase from 2019.

Romania crime rate & statistics for 2019 was 1.26, a 1.69% decline from 2018.

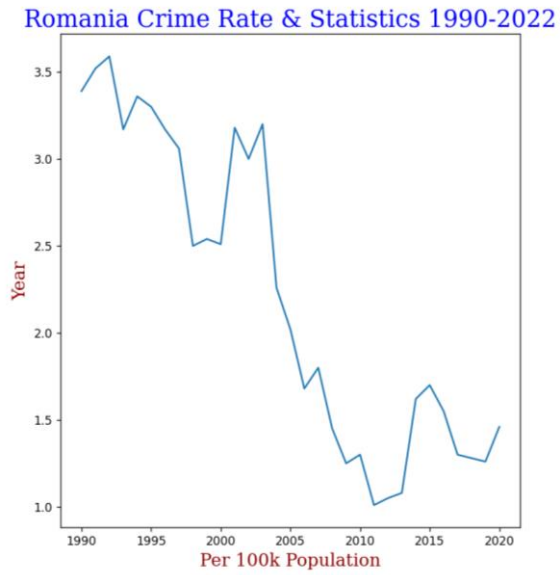

Figure 10 - Romania Crime Rate & Statistics 1990-2022

Exploratory Factor Analysis

Psychometric test development by psychologists involves the statistical technique known as exploratory factor analysis. It can help researchers formulate questions about their research topics, comprehend the correlations between variables, and find latent variables. This approach places a strong emphasis on the interrelationships between common factors and manifest variables. Before continuing with the next phase of their research, it aids researchers in establishing links between indicators and variables.

- Correlogram of Kaiser-Meyer-Olkin indices

I followed the seminar and I did the Correlogram of Kaiser-Meyer-Olkin indices with data from the database tipuri_crime_romania.

KMO values between 0.8 and 1 indicate the sampling is adequate. KMO values less than 0.6 indicate the sampling is not adequate and that remedial action should be taken. In contrast, others set this cutoff value at 0.5.

A KMO value close to zero means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which would be a large problem for factor analysis.

Our indices are all above 0.8, which means we have good and very factorability.

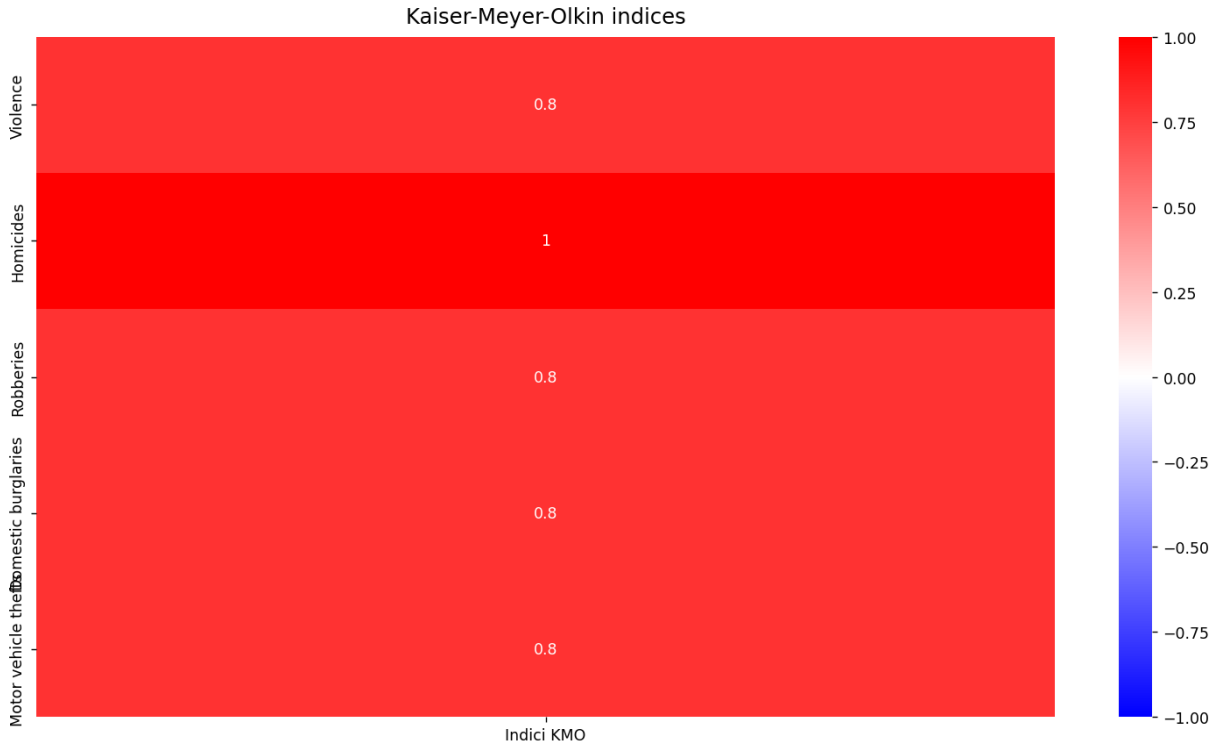

Figure 11 - Correlogram of Kaiser-Meyer-Olkin indices

A good model is one in which the estimated covariance matrix, derived from the model, nears as much as possible the covariance matrix of the observed, causal variables.
In order to test the model, statistical tests are employed, such goodness-of fit test is the one proposed by Bartlett, a chi-square test.

-Correlogram of correlation factors

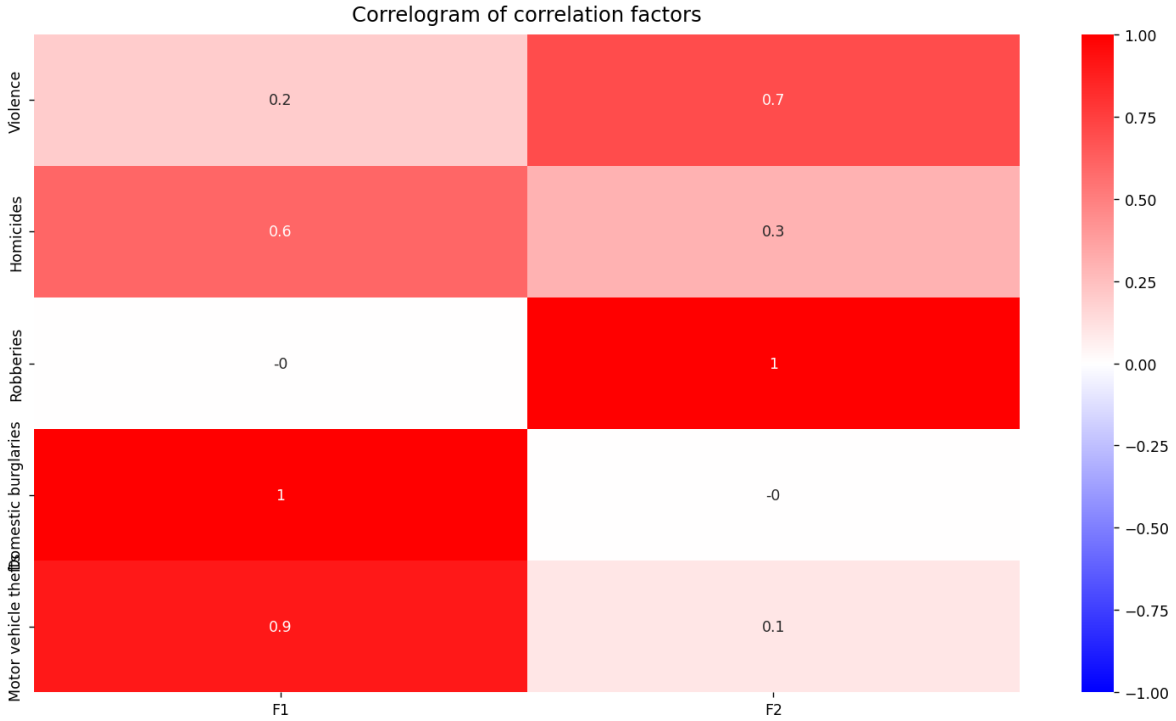

Figure 12 - Correlogram of correlation factors

Positive correlations are displayed in red and negative correlations in blue. The intensity of the color is proportional to the correlation coefficient so the stronger the correlation (the closer to -1 or 1), the darker the boxes.
Consider that a negative correlation indicates that the two variables are changing in the opposite directions of one another, i.e., if one variable rises, the other falls, and vice versa. A positive correlation suggests that the two variables are varying in the same direction; for example, if one variable grows, the other increases as well, and vice versa if one variable drops. Additionally, the relationship between the two variables is stronger the higher the correlation.

	F1	F2
Violence	0.239722	0.682160
Homicides	0.594224	0.257473
Robberies	-0.019977	0.999137
Domestic burglaries	0.980204	-0.003002
Motor vehicle thefts	0.852470	0.090240

Figure 13 - Correlogram of correlation factors highlighted

We highlight all the loading factors whose absolute value is greater than 0.4.
From glancing at the solution, we see that Domestic burglaries has the highest correlation with Factor 1 and Robberies the lowest. Vehicles theft and homicide rate have large positive loadings on factor 1. Also, Domestic burglaries is highly positively correlated with Factor 1 while Robberies is negatively correlated with Factor 1.

Similarly, we see that Violence and Robberies have the highest correlation with Component 2 and Domestic burglaries the lowest. Robberies is highly positively correlated with Factor 2 while Domestic burglaries is negatively correlated with Factor 1

Ideally, we would like to see that each variable is highly correlated with only one factor. As we can see, this is the case in our example, except that Motor vehicle thefts is correlated with both Factor 1 and 2.

Table of figures

Figure 1 - reading from .xlsx and .csv files 4
Figure 2 - Writing in .xlsx files 4
Figure 3- Merging two files for creating one database 4
Figure 4 - Calculating the sum by country for all years and creating a new column for it 5
Figure 5 - Use the numpy library 6
Figure 6 - Use the pandas library 6
Figure 7 - Pie chart Belgium types of crimes 7
Figure 8 - Plotting the criminality index and safety index 7
Figure 9 - Sexual violence and rape correlation in year 2020 8
Figure 10 - Romania Crime Rate & Statistics 1990-2022 8
Figure 11 - Correlogram of Kaiser-Meyer-Olkin indices 9
Figure 12 - Correlogram of correlation factors 10
Figure 13 - Correlogram of correlation factors highlighted 11

Bibliography

https://www.macrotrends.net/countries/ROU/romania/crime-rate-statistics

https://journals.sagepub.com/doi/10.1350/pojo.2013.86.3.625

https://www.numbeo.com/crime/rankings_by_country.jsp?title=2022-mid&region=150

http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

https://ec.europa.eu/eurostat/databrowser/view/ilc_mddw06/default/table?lang=en

https://stats.oarc.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/