Multivariate Statistical Analysis - MS-E2112.

Multiple Correspondence Analysis of domestic violence in Finland, 2019.

Anton Saukkonen

April 11, 2021
Aalto University, Espoo

Contents

1 **Introduction** **2**
1.1 Description of the data set and data collection 2
1.1.1 Data collection and refining. 2
1.2 Research aim and questions 2

2 **Univariate analysis** **3**
2.1 Victim's sex 3
2.2 Mode of housing 3
2.3 Relation between victim and the suspect 4
2.4 Victim's age 5
2.5 Suspect's sex 6
2.6 Offence 6

3 **Bivariate analysis** **7**
3.1 Victim's sex vs. Suspect's sex 7
3.2 Victim's sex vs. Victim's age 8
3.3 Victim's age vs. Victim-Suspect relationships 9
3.4 Victim's age vs. Offence 10
3.5 Victim's sex vs. Relationships 11
3.6 Victim's sex vs. Offence 12
3.7 Victim's sex vs. Housing Mode 13
3.8 Housing Mode vs. Relationships 13
3.9 Housing Mode vs. Offence 14
3.10 Victim's age vs. Housing Mode 15
3.11 Suspect's sex vs. Housing Mode 15
3.12 Suspect's sex vs. Relationships 16
3.13 Relationships vs. Offence 17
3.14 Victim's age vs. Suspect's sex 18
3.15 Suspect's sex vs Offence 18

4 **Multivariate analysis** **19**
4.1 Component analysis of explained variance 19
4.2 Analysis of column profiles 21
4.3 Qualitative discussion of results 22
4.3.1 MCA density plot (**) 23
4.4 Answering research questions 24
4.4.1 General trends 24
4.4.2 Outliers 24
4.4.3 Open questions 24
4.4.4 Words of warning 24

5 **Suggestion for further improvements** **25**

6 **Sources of bias** **25**
6.1 Data selection bias 25
6.2 Detection bias 25
6.3 Confirmation bias and belief bias 26

To the evaluator: *Dear evaluator, unfortunately, at the first glance this report may seem to be excessively large and scare you. This is very friendly note to warn you that it is actually within the limit and the majority of space is taken by uncompressed ggplots, which I simply did not want to compress, since they will loose their quality and thus meaning. Essential part of the analysis is only chapters 4, 5 and 6. Univariate and bivariate parts could be skimmed through very quickly.*

1 Introduction

The aim of this project is to perform multiple correspondence analysis (MCA) of the data about domestic violence (DV) and intimate partner violence that has happened in Finland during the year 2019 and was reported as an offence. The purpose of this report is to provide information on the data set and variable selection process, define the research questions and describe the employed statistical methods. Furthermore, the report provides qualitative analysis of the results and discussion of the research questions, identifies sources of biases and suggests potential ways to achieve stronger results.

1.1 Description of the data set and data collection

The data set "Domestic violence and intimate partner violence reported as an offence" was retrieved from open archive of "Statistics Finland" agency (Tilastokeskus). In particular, this data set contains categorical data about characterization of domestic violence and intimate partner violence offences, that happened in Finland during the period of 2009-2019 according to the following criteria (factors): Information on number of victims, Year, Victim's sex, Mode of Housing, Relation between victim and the suspect, Victim's age, Suspect's sex and Offence. Each of the aforementioned categorical variables is further divided into several categories (levels). For the sake of convenience, all of the categories are not listed in the report. Instead, they could be found in the reference [1].

1.1.1 Data collection and refining.

The original data set [1] contains 810810 cells, where each cell represents the sum of the frequencies of observations belonging to a certain category. Due to extensively high number of levels and observations, it was necessary to construct a sample. During the construction, only 2019 year data was used and only categories allowing to make the most generalized and informative inferences were selected. Namely, too specific or rare categories were excluded. Moreover, categories, which do not carry particular qualitative information such as "other" were excluded from the sample. In addition, it was necessary to refine data into factor-level format and cumulative frequency table to enable an efficient use of computational tools. As a result, the constructed sample contained 6 factors further divided into 28 levels: 1) **Victim's sex**: Male Victim's, Female Victim's, 2) **Type of household**: Same household unit, Different household unit, 3) **Type of relations**: Suspect is the parent of the victim, Siblings, Spouses and cohabiting partners, Former spouses, Cousin, 4) **Victim's age**: 0-4, 5-9, 10-14, 15-17, 18-20, 25-34, 35-44, 45-54, 55-64, 65-, 5) **Suspect's sex**: Suspect males, Suspect Females, 6) **Offence**: Rape, Attempted manslaughter, Assault, Robbery, Extortion, Deprivation of personal liberty, Menace.

1.2 Research aim and questions

The aim of the research is to explore and characterize association between selected categories and as a result, reveal the underlying structure of the domestic violence in Finland. In other words, we wish to know how various characteristics such as type of an offence, victim's age, victim's sex, suspect's sex, type of relationships and even housing mode are related to each other and if such relation exists, what could we infer about it's structure. The research questions are:

1) Is there a structural association between the Victim's sex, Type of household, Victim-Suspect relationships, Victim's age, Suspect's sex and Offence?
2) What is the structure of this association?

To answer the research questions, multiple correspondence analysis was applied to the data set.

2 Univariate analysis

2.1 Victim's sex

The variable *Victim's sex* describes the mass distribution of victims according to their gender. In particular, we observe that roughly 67% of victims are of female gender, while victims of male gender constitute only 33% of the individuals.

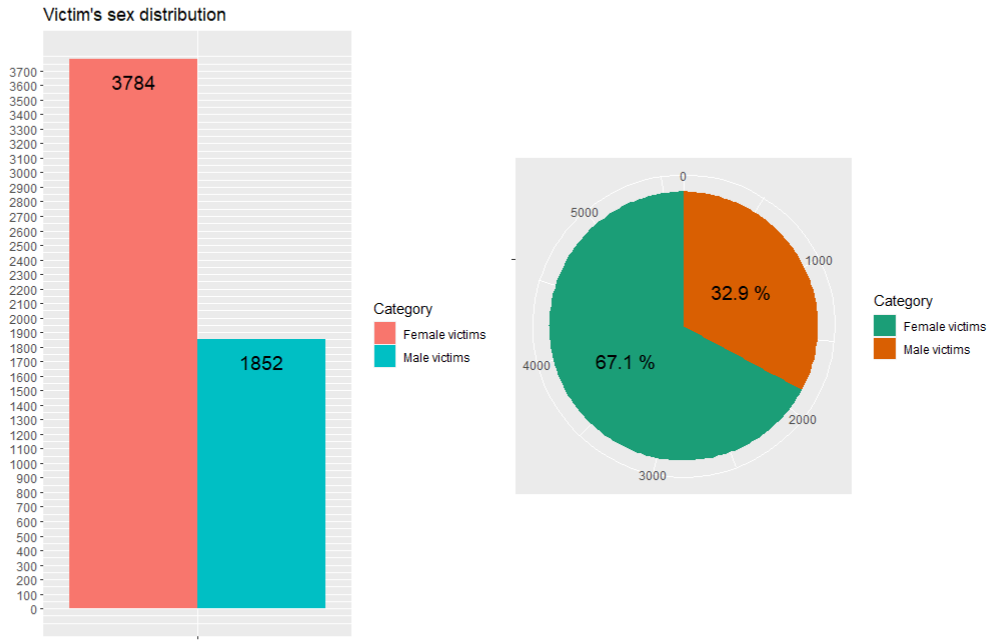


Figure 1: Victim's sex distribution

2.2 Mode of housing

The variable *Mode of housing* describes the mass distribution of the household-dwelling units, where the reported offence has been committed. Accordingly, roughly 60% of DV offences are committed by people living in the same household-dwelling units, while 40% are related to offences from non-shared units.

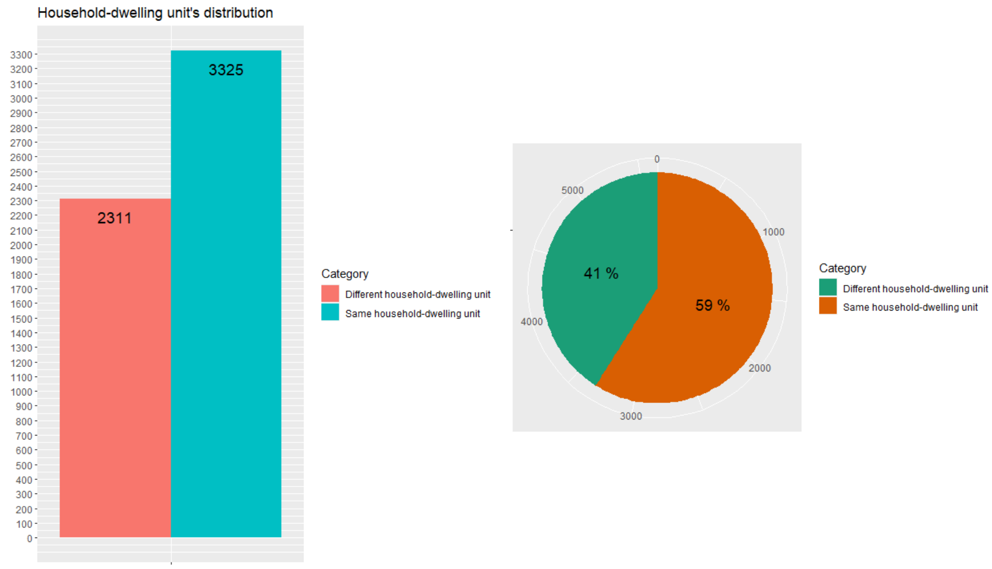


Figure 2: Household-dwelling unit's distribution

2.3 Relation between victim and the suspect

The variable *Relations between victim and suspect* contains five modalities. Namely: *Spouses and cohabiting partners*, *Suspect of is the parent of the victim*, *Former spouse*, *Siblings* and *Cousin*. Accordingly, almost 47% and 35% of relationships between suspect and victim are in the form of partnership or parent-child respectively. Consequently, observation suggests that partners or spouses are the most DV inclined relationship category, while the second most frequent category is relationship between child and parent.

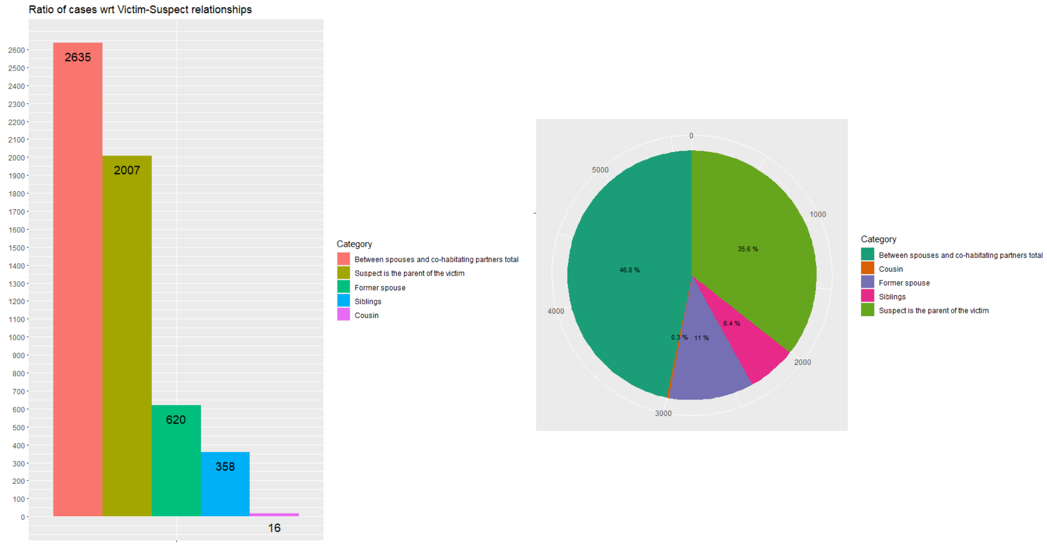


Figure 3: Distribution of cases with respect to relationship between suspect and the victim

2.4 Victim's age

Variable *Victim's age* consists of 10 modalities, which describe the different age groups of the victim. The modalities are: *0-4*, *5-9*, *10-14*, *15-17*, *18-20*, *25-34*, *35-44*, *45-54*, *55-64*, *65-*. It is observable that the most frequent categories are *25-34* and *35-44*, which constitute roughly to the 21% and 19% of the relative frequency. Other relatively frequent categories are *5-9*, *45-54*, 10-14, *0-4*.

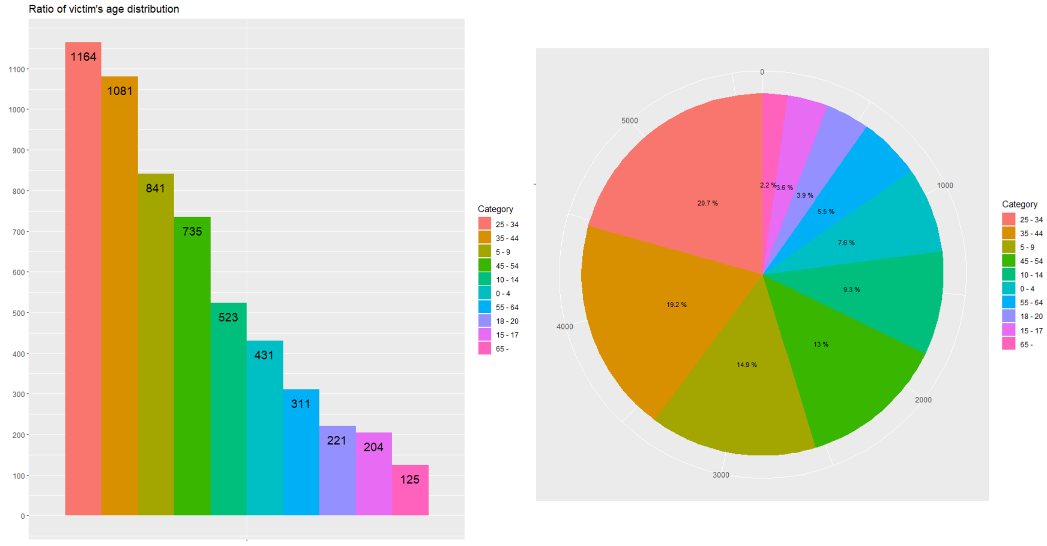


Figure 4: Distribution of cases with respect to the age of the victim

2.5 Suspect's sex

Another characteristic of DV offences is classification of suspects with respect to the gender. According to the mass distribution of observations, roughly 77% of all offences are committed by male suspects, while the 23% are committed by female suspects.

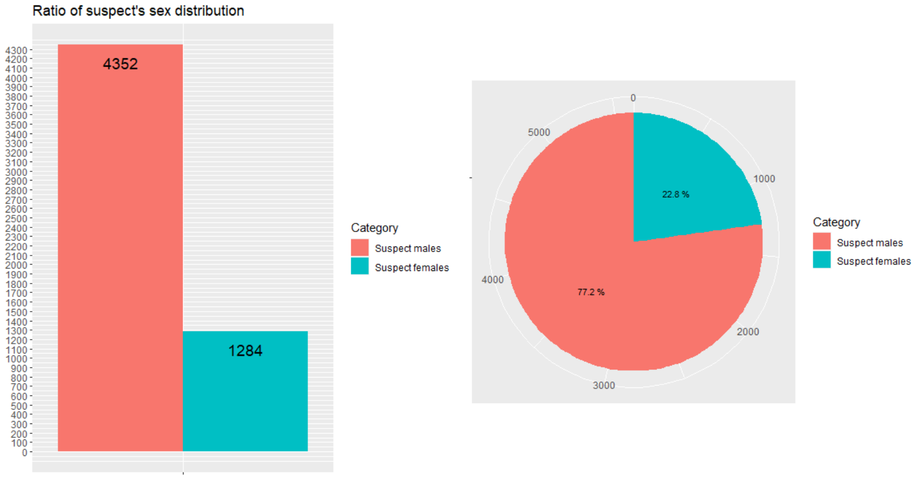


Figure 5: Distribution of cases with respect to gender of the suspect.

2.6 Offence

Variable *Offence* contains the following modalities: *Assault*, *Menace*, *Rape*, *Deprivation of personal liberty*, *Attempted manslaughter*, *Extortion and Robbery*. Accordingly, it is observable that almost 80% of the cases are related to the category *Assault*. The next most frequent category is *Menace*, which constitute roughly 18% of total relative frequency. In the order of decreasing frequency, the next categories are: *Rape*, *Deprivation of personal liberty*, *Attempted manslaughter*, *Extortion and Robbery*.

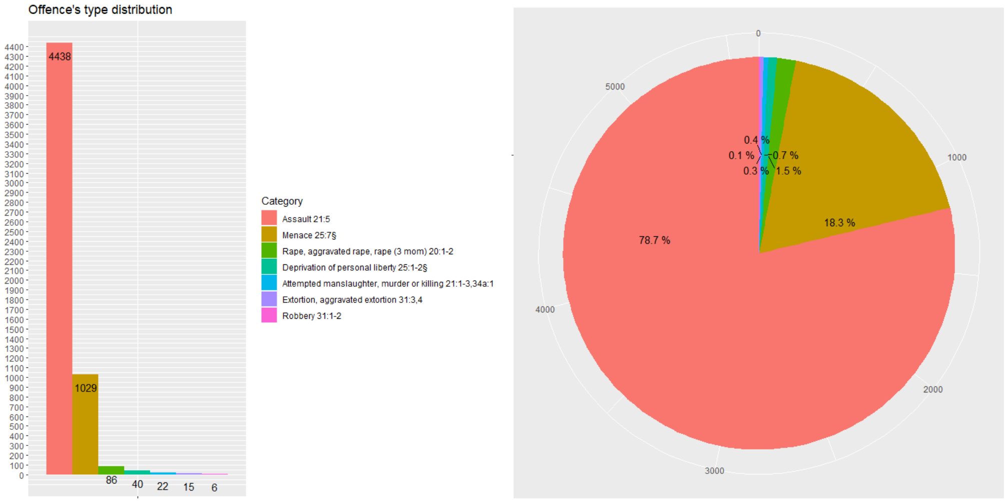


Figure 6: Distribution of cases with respect to offence type.

3 Bivariate analysis

3.1 Victim's sex vs. Suspect's sex

According to bivariate statistics, category *Suspect males* is significantly more frequent for the category *Female victims*, constituting 60% of total relative frequency, while for category *Male victims* proportion of categories *Male suspects* and *Female suspects* is roughly the same.

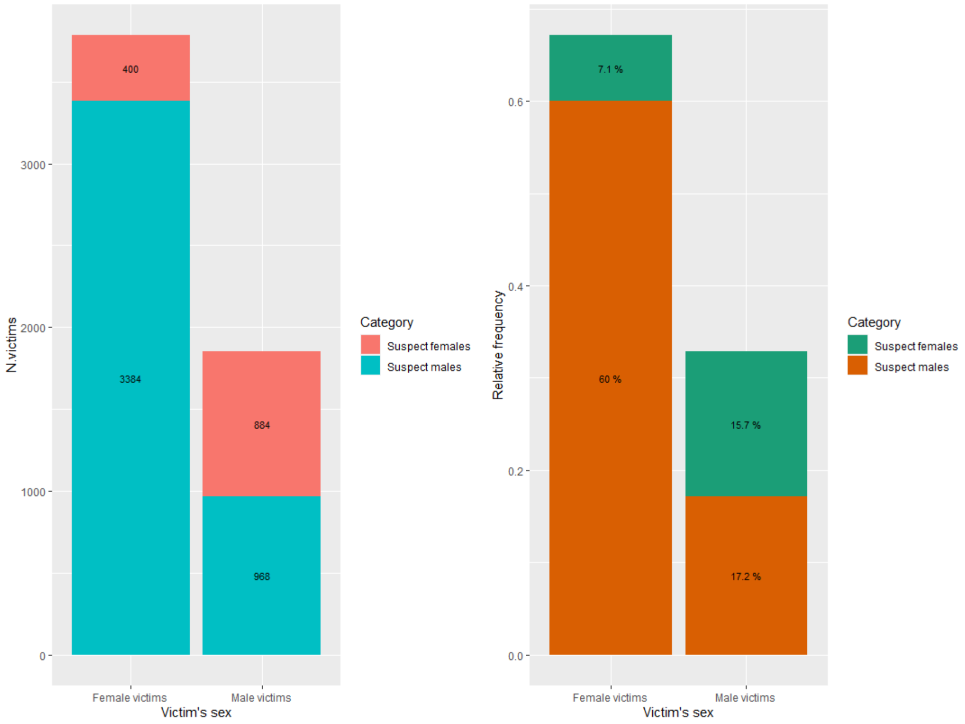


Figure 7: Frequency of cases with respect to victim's sex and suspect's sex.

3.2 Victim's sex vs. Victim's age

According to the bivariate statistics of the variables *Victim's sex* and *Victim's age*, it is observable that in the age groups *25-34*, *35-44* and *45-54*, victims of female gender are highly dominant. Moreover, together they constitute roughly 43% of the relative frequency. On the other hand, it is notable that for the age groups 0-4, 5-9 and 10-14, the proportion of males and females is either roughly the same or males are more inclined to become victims of DV (note the category *5-9*).

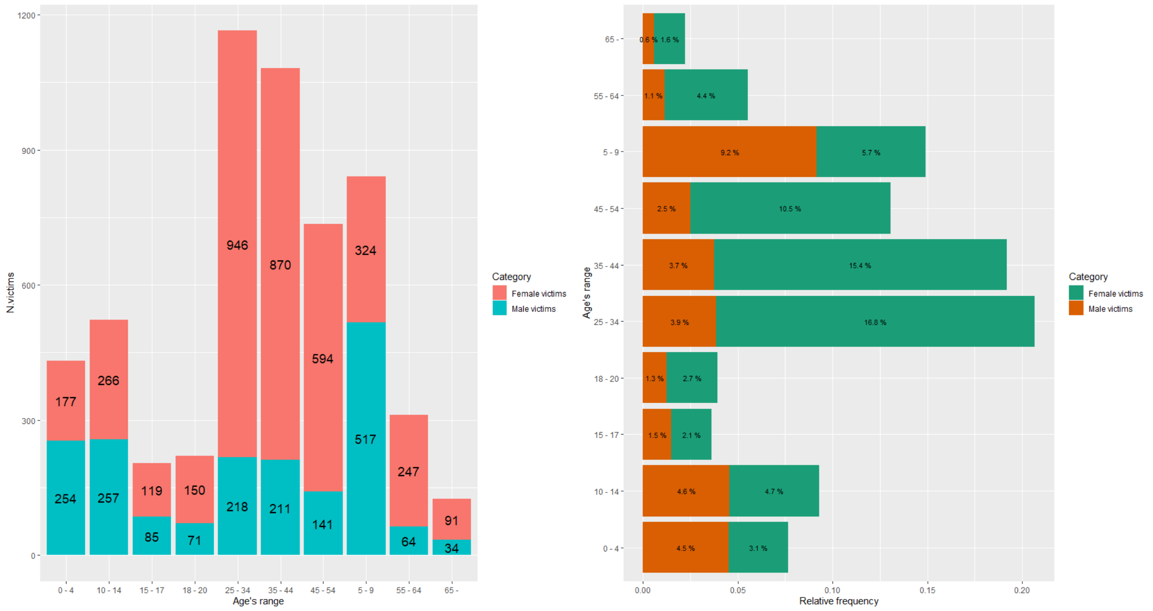


Figure 8: Classification of cases with respect to Victim's sex and Victim's age

3.3 Victim's age vs. Victim-Suspect relationships

The bivariate statistics of the variables reveals high mutual frequency in the middle age groups *18-20*, *25-34*, *35-44*, *45-54*, *55-64*, *65-* and young age groups *0-4*, *5-9*, *10-14*, *15-17*. For instance, category *Between spouses and cohabiting partners* is the most frequent for groups ranging from 18 to 65- years, which account for almost 47% of the relative frequency. On the other hand, young age groups ranging from 0 to 17 have higher mutual frequency with the category *Suspect is the parent of the victim* and constitute almost 34% of relative frequency.

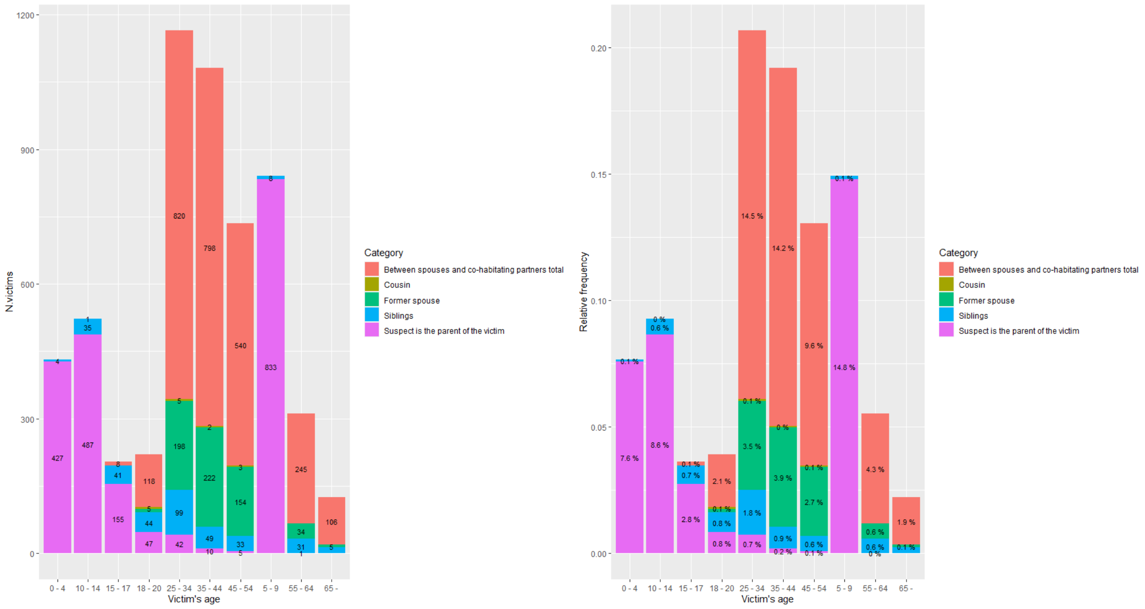


Figure 9: Distribution of cases with respect to Victim's age and Victim-Suspect relationships.

3.4 Victim's age vs. Offence

According to the univariate analysis of variable *Offence*, it was noted that almost 80% of committed DV crimes are classified as an *Assault* and around 18% as a *Menace*. It is observable that this trend propagates to the bivariate case too. Category Menace is frequent for the categories *25-34*, *35-44*, *45-54*, while other infrequent types of offences are approximately equally distributed, excluding possibly *Rape*, which is more typical for the category *25-34*.

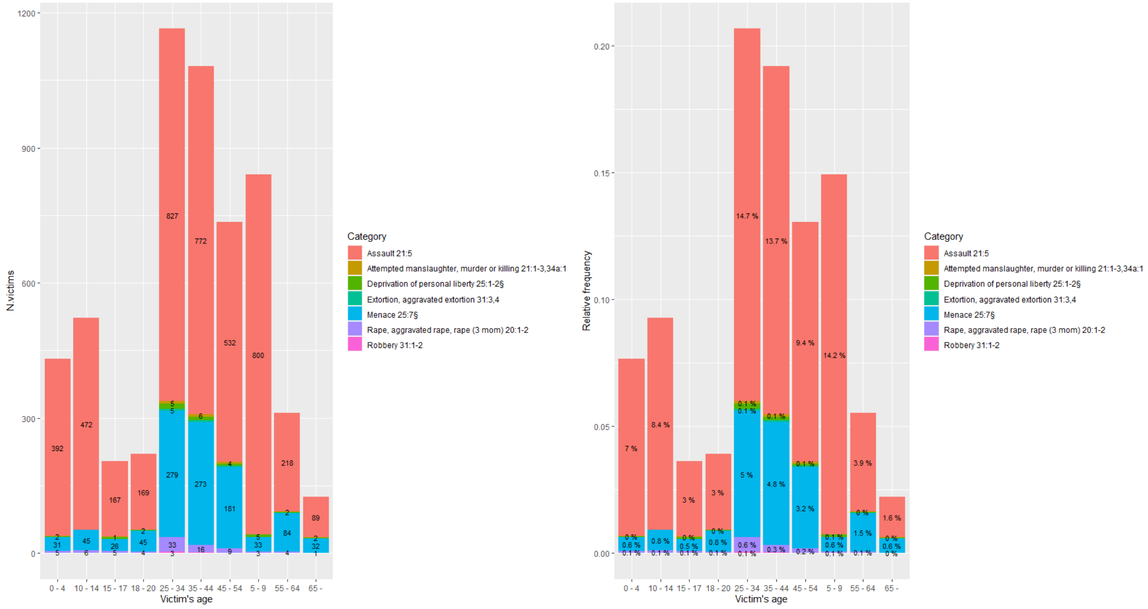


Figure 10: Distribution of cases with respect to Victim's age and offence type.

3.5 Victim's sex vs. Relationships

According to the mutual distribution of the variables *Victim's sex* and *Victim-suspect relationships*, it is observable that category *Female victims* is tend towards the category *Between spouses and co-habiting partners* which constitute to almost 40% of relative frequency. On the other hand, category *Male victims"* is lean towards category *Suspect is the parent of the victim*, constituting 20% of the relative frequency.

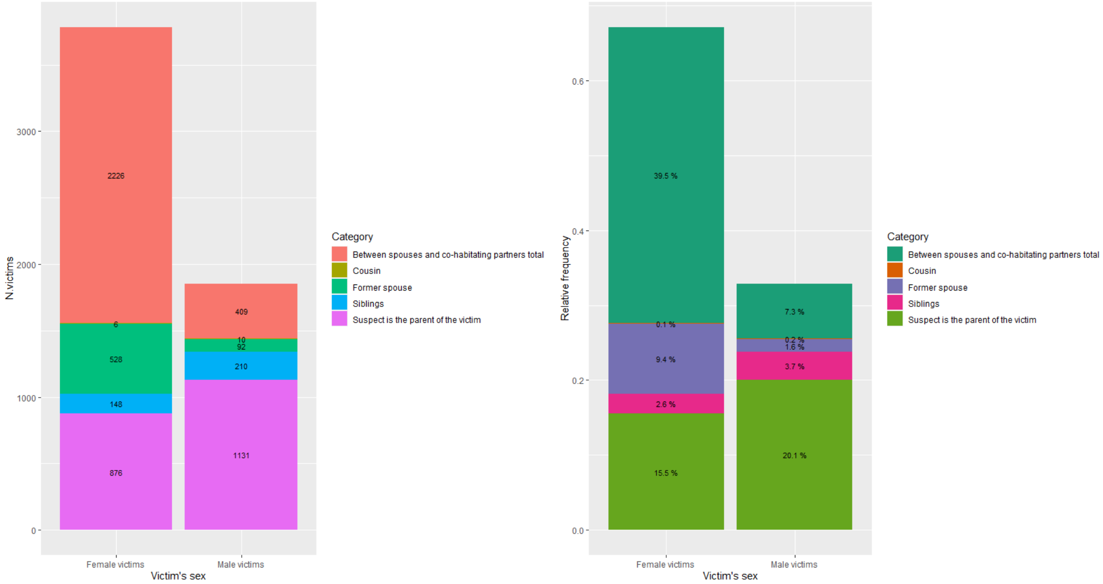


Figure 11: Classification of cases with respect to Victim's sex and Victim-suspect relationships.

3.6 Victim's sex vs. Offence

The variable *Offence* is prevailed by the categories *Assault* and *Menace*. 50% and 28% of all assaults are committed against female and male individuals respectively. Relative frequency of the modality *Menace* is 14.3% and 4% for the females and males respectively.

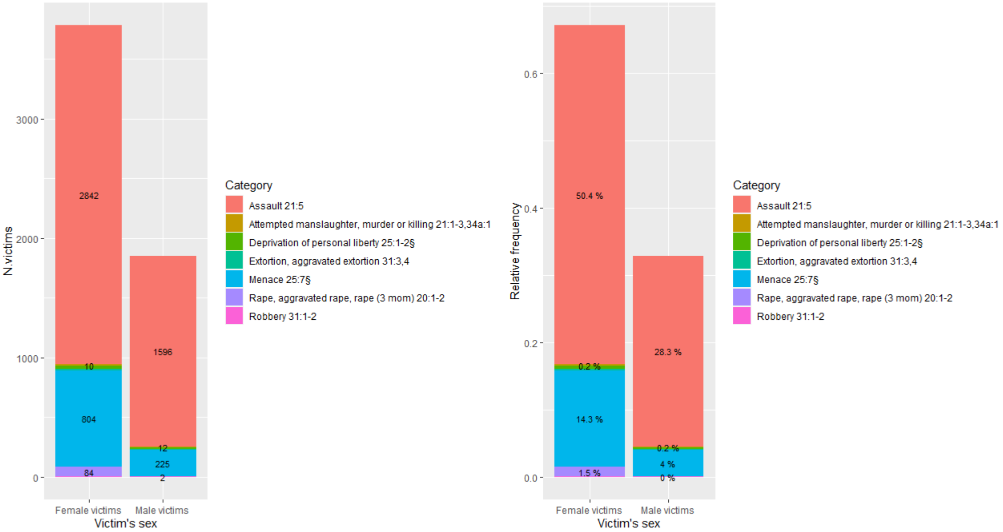


Figure 12: Classification of cases with respect to Victim's sex and offence type.

3.7 Victim's sex vs. Housing Mode

Pairwise dependency between victim's sex and housing is approximately equal for female and male individuals. It is observable that frequency is higher in the same household unit for both genders.

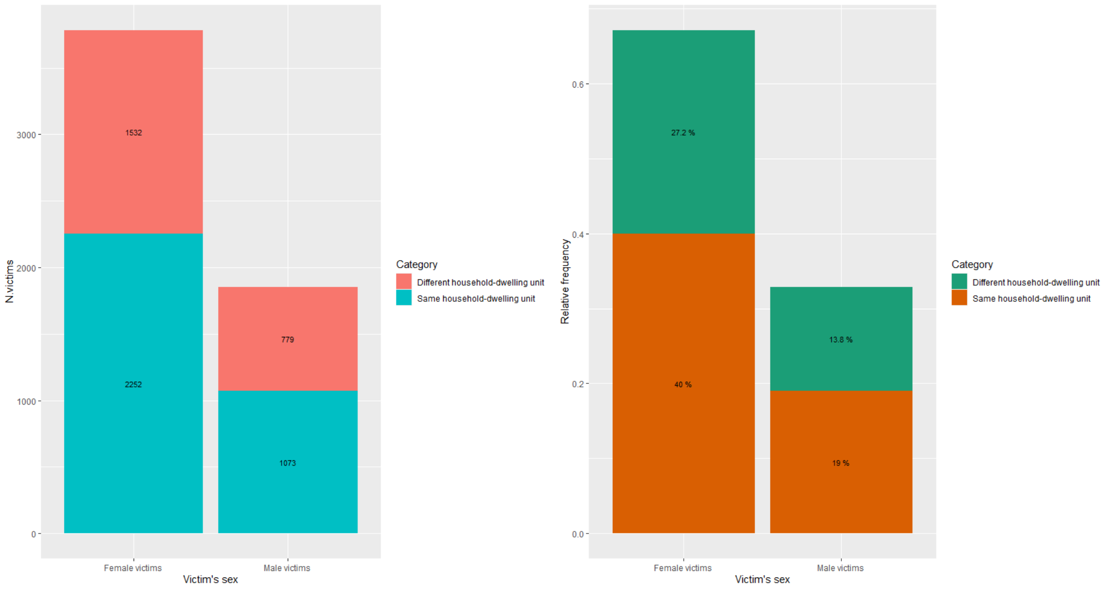


Figure 13: Classification of cases with respect to Victim's sex and Housing mode.

3.8 Housing Mode vs. Relationships

According to the classification of housing mode vs. victim-suspect relationships, major proportion of DV for category *Between spouses and co-habiting partners* as well as for the category *Suspect is the parent of the victim* is frequent for the same household units. On the other hand categories *Siblings*, *Cousin* and *Former spouses* occur more often in non-shared household dwelling units.

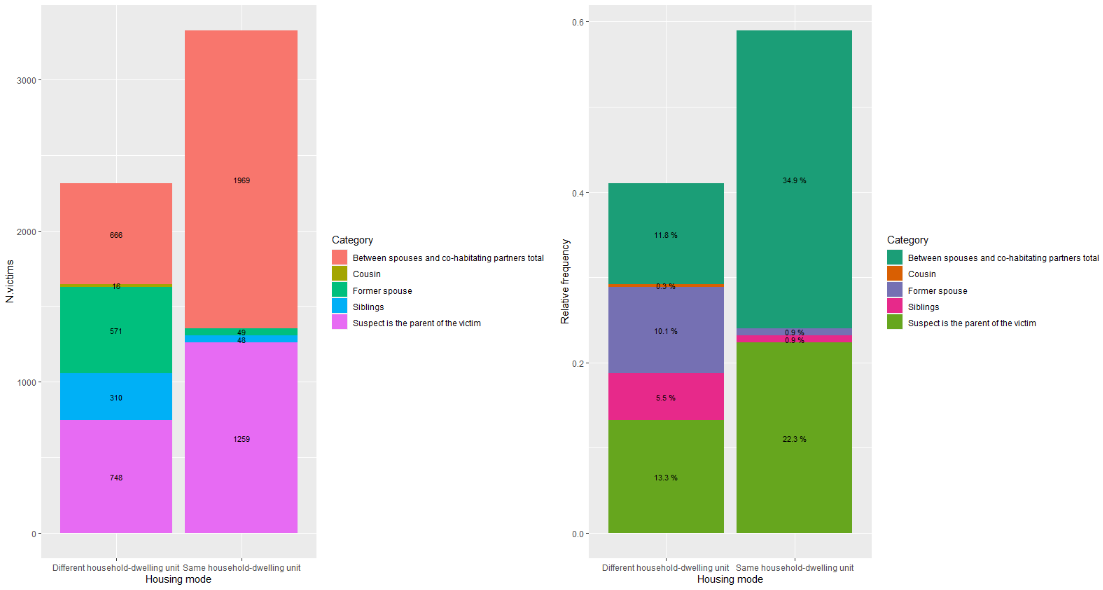


Figure 14: Classification of cases with respect to Housing mode and Victim-suspect relationships.

3.9 Housing Mode vs. Offence

Bivariate statistics of the *Housing mode* and *Offence* is dominated by the categories *Assault* and *Menace.*Accordingly, assaults in same household units constitute 51% of the relative total frequency, while assaults in different household units make up 27%.

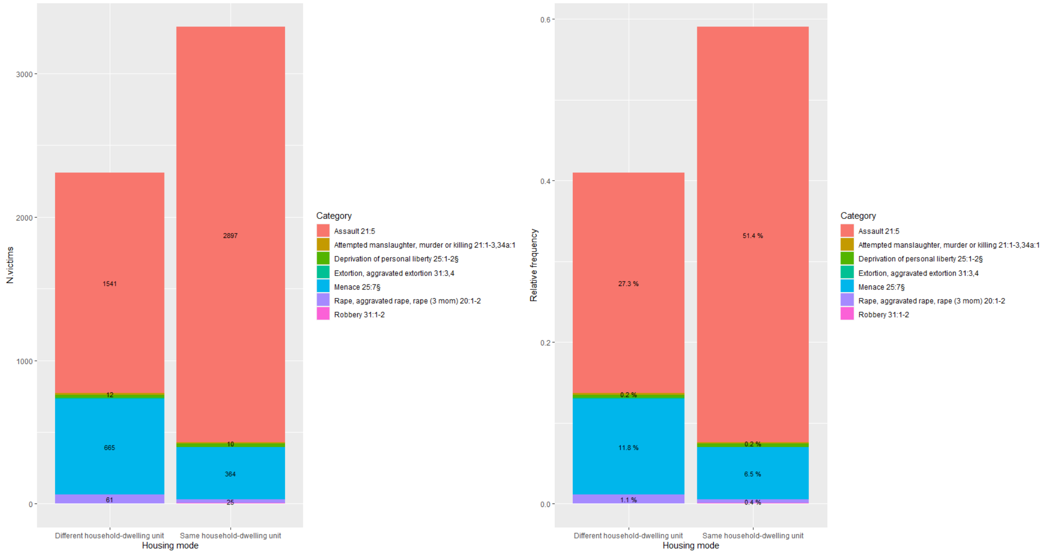


Figure 15: Classification of cases with respect to Housing mode and Offence type.

3.10 Victim's age vs. Housing Mode

According to pairwise dependency between variables *Victim's age* and *Housing mode*, it is observable that almost all age groups are more lean towards modality *Same household-dwelling unit*.

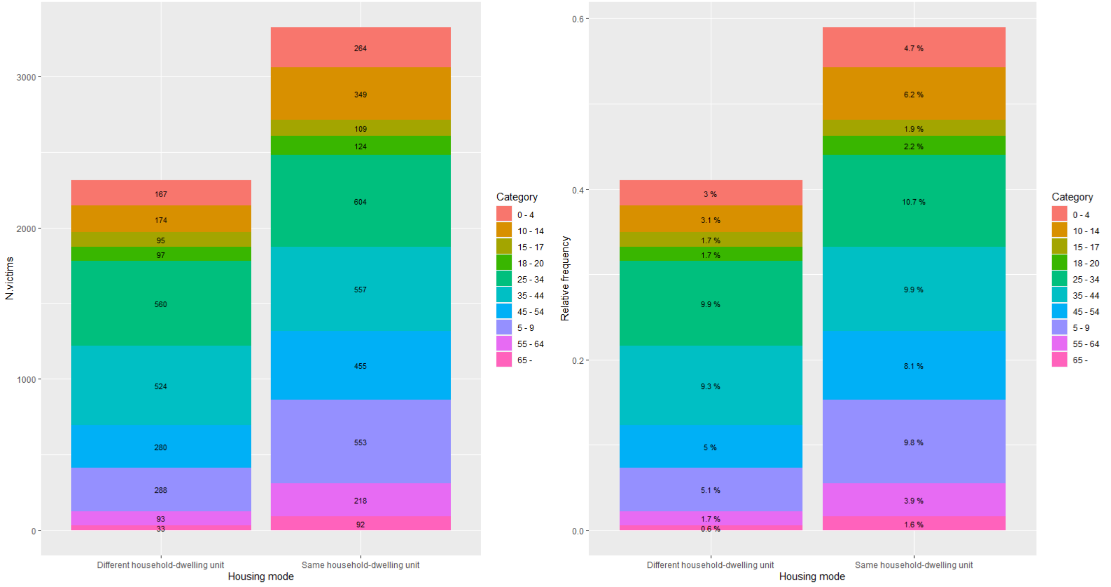


Figure 16: Classification of cases with respect to the victim's age and housing mode.

3.11 Suspect's sex vs. Housing Mode

For the variables *Suspect's sex* and *Housing Mode*, it is observable that category *Suspect males* is the most frequent for the both categories of variable *Housing mode*.

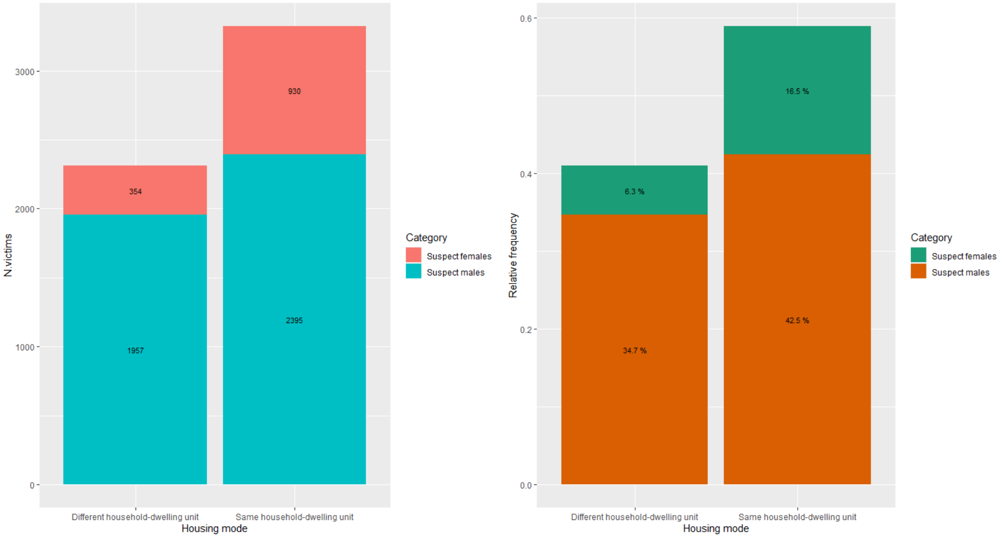


Figure 17: Classification of cases with respect to the suspect's sex and housing mode.

3.12 Suspect's sex vs. Relationships

According to the mutual distribution of variables *Suspect's sex* and *Victim-suspect relationships*, modalities *Suspect males* and *Between spouses and cohabiting partners* are attracted to each other. It is observable that in roughly 40% of cases, suspect individuals are of male gender and have partnership-type relationship with the victim. The category *Between spouses and cohabiting partners* also prevails, when consideration is restricted only to the male suspects. On the other hand, it is notable that category *Suspect is the parent of the victim* is more frequent among the category *Suspect females*.

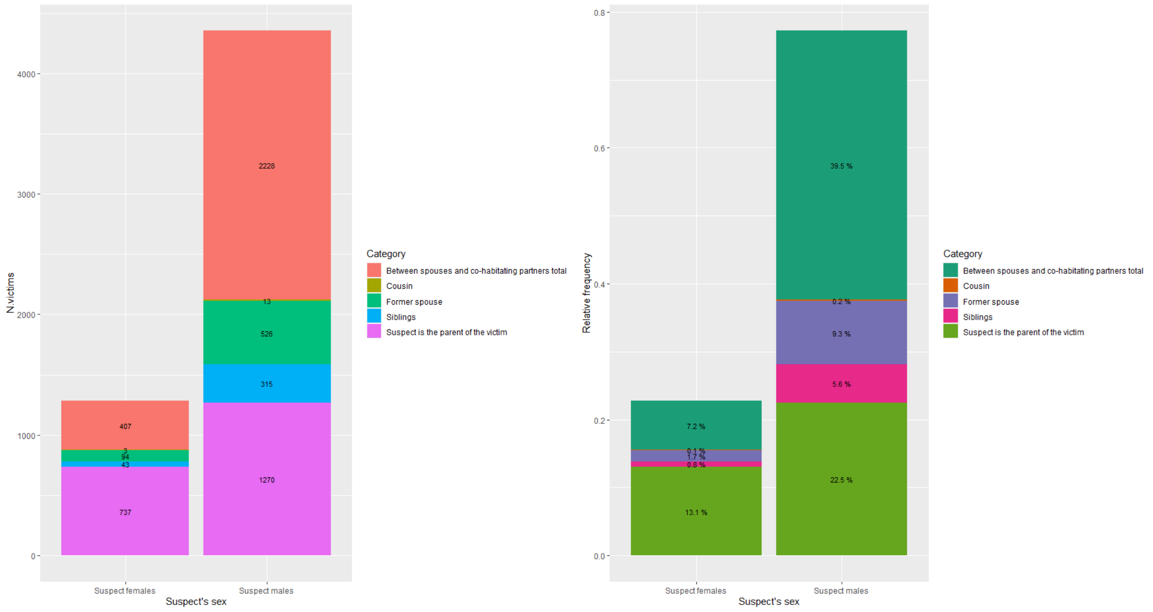


Figure 18: Classification of cases with respect to the suspect's sex and victim-suspect relationships.

3.13 Relationships vs. Offence

According to the bivariate statistics, category *Assault* dominates among all categories of the variable *Relationship between suspect and the victim.*

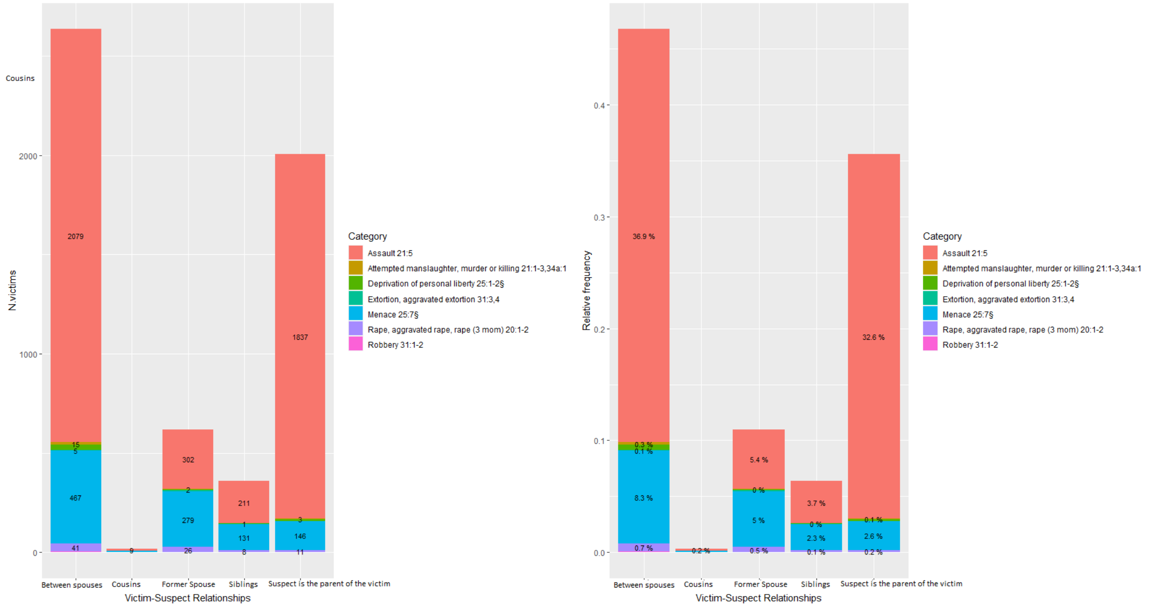


Figure 19: Classification of cases with respect to the offence type and victim-suspect relationships.

3.14 Victim's age vs. Suspect's sex

According to the mutual distribution of the variables *Suspect's sex* and *Victim's age*, it is observable that category *Suspect males* prevails in all age groups.

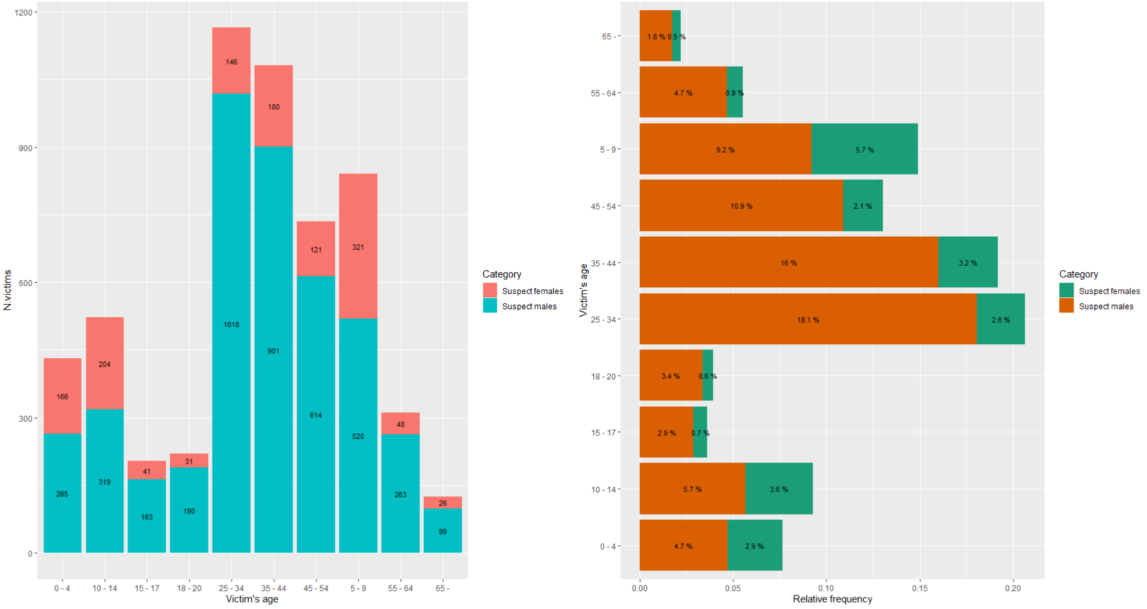


Figure 20: Classification of cases with respect to the suspect's sex and victim's age.

3.15 Suspect's sex vs Offence

Bivariate dependency between variables *Suspect's sex* and *Offence* reveals that modality *Assault* is the most frequent for both modalities *Suspect males* and *Suspect females*. Second most frequent modality for both categories is *Menace*.

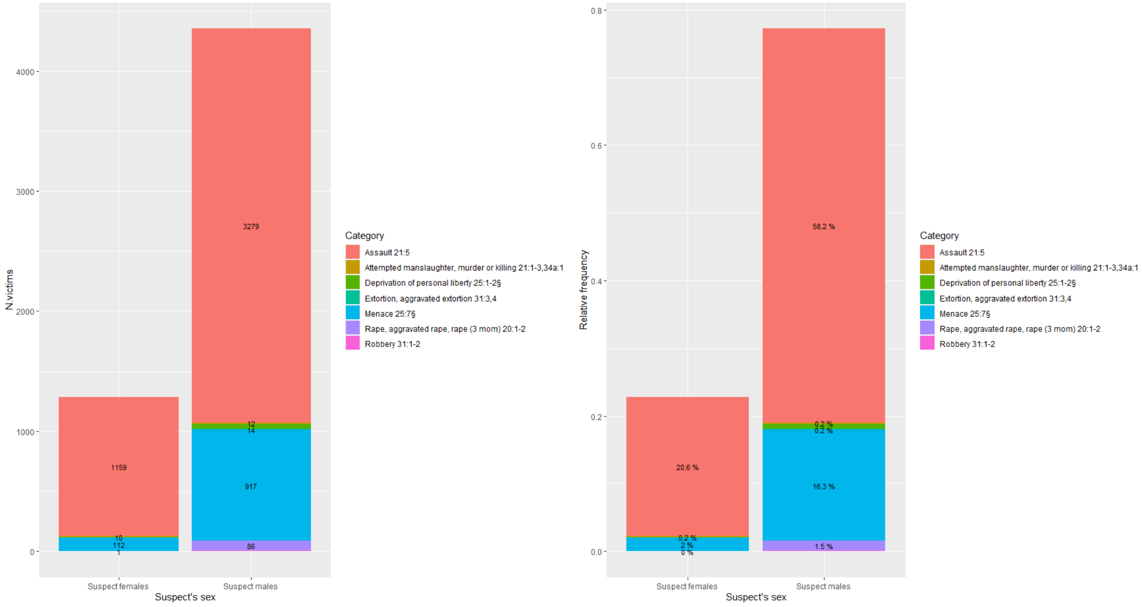


Figure 21: Classification of cases with respect to the suspect's sex and offence type.

4 Multivariate analysis

4.1 Component analysis of explained variance

The sample, composed of the data retrieved from [1] contained 28 dimensions. During the construction of principal components, the dimensionality was reduced to 22. It is of interest to analyze the proportion of variance explained by each newly obtained 22 principle components and especially, by the first 2 principle components. In order to gain these insights, we shall look into the following plots:

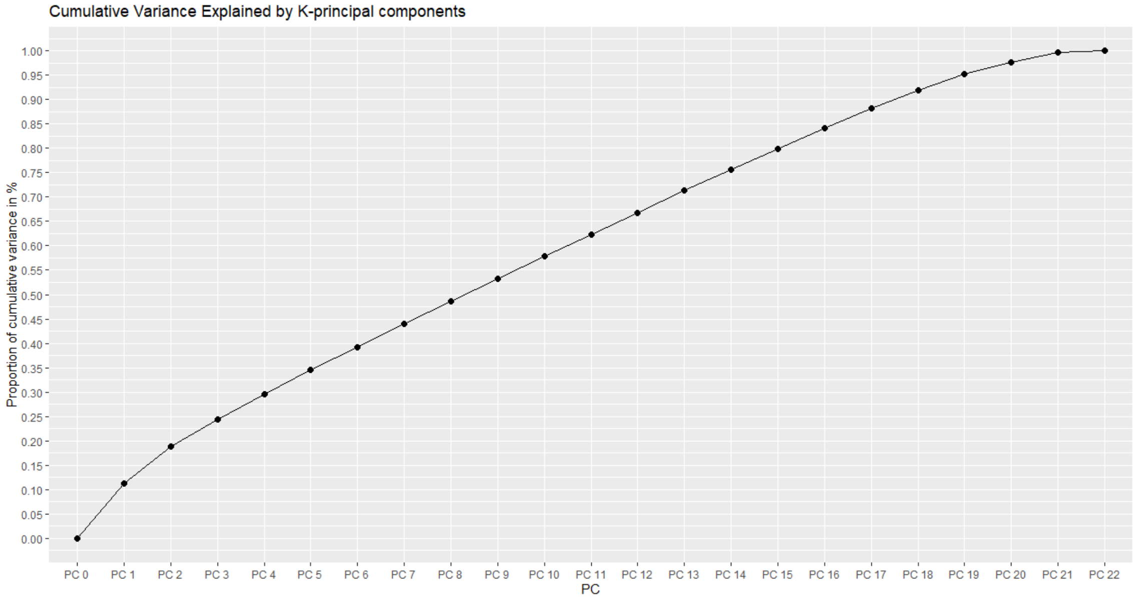


Figure 22: Cumulative variance plot.

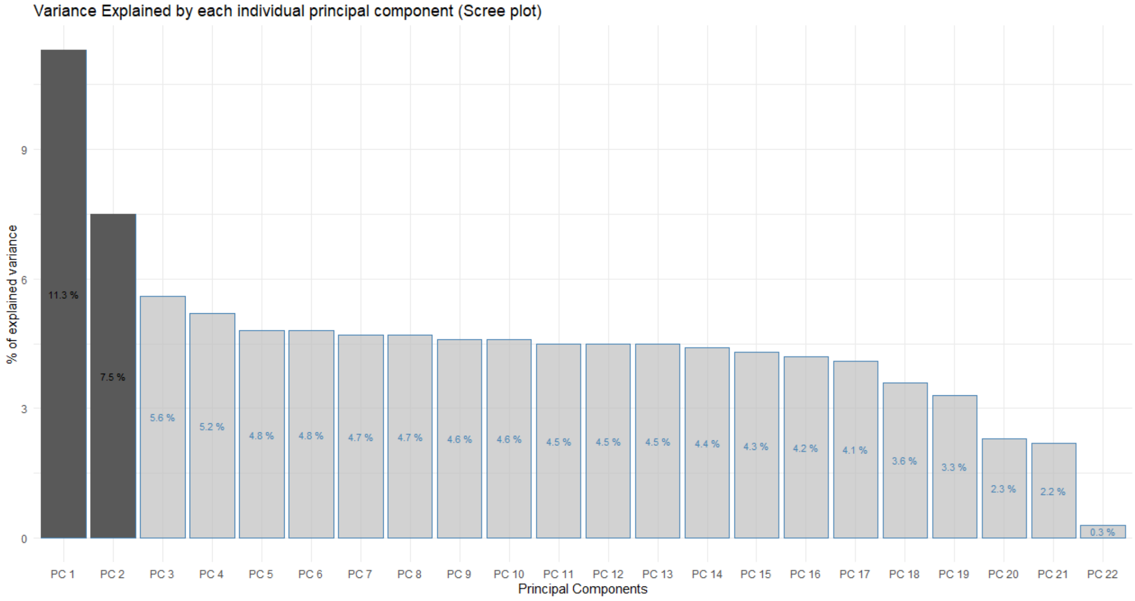
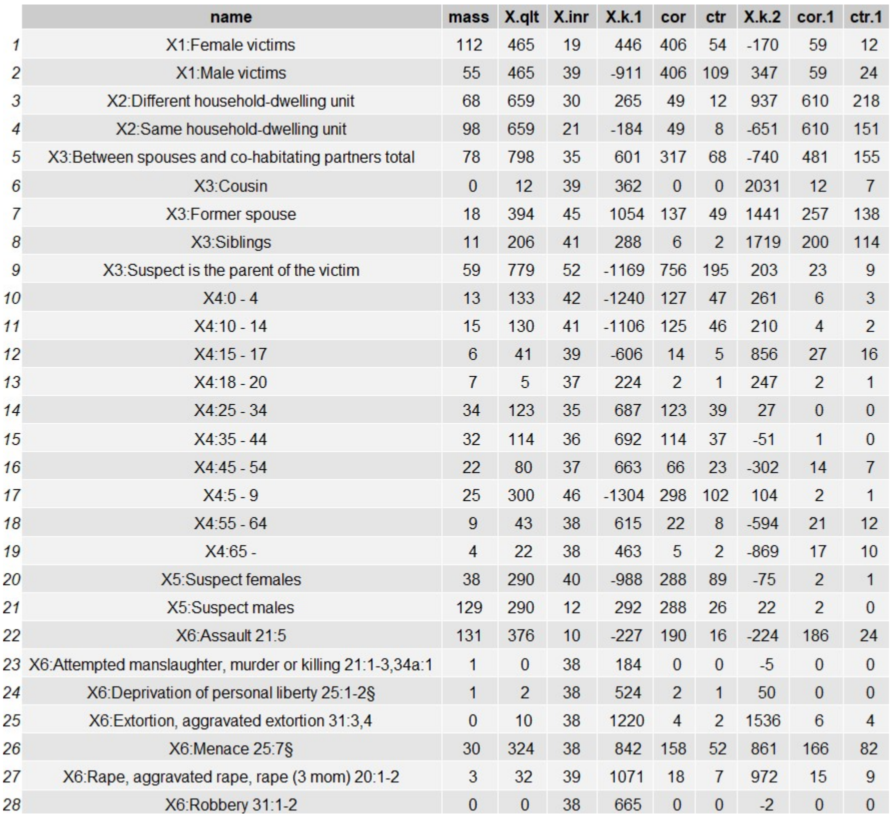


Figure 23: Scree plot.

According to the Scree and Cumulative variance plot, the first and second principal components explain 11.8% and 7.5% of the total variance respectively, which implies that their combination accounts for roughly 19% of total variance. In other words, after dimensionality reduction, we were able to get only $\frac{1}{5}$ of the information from data. In addition, it is observable that components 5 to 17 have difference in the explainable variance only up to 0.6%, which implies that in order to interpret at least 90% of the variance, we would need roughly 17 components. This potentially suggest the idea that either there is a caveat in the choice of the variables/modalities, or even the nature of original data is difficult itself. Nevertheless, J.Hair [2] states that information type in social sciences often has low degree of precision, which implies lower satisfactory levels. Therefore, it is a must continue analysis with this consideration in mind.

4.2 Analysis of column profiles



	name	mass	X.qlt	X.inr	X.k.1	cor	ctr	X.k.2	cor.1	ctr.1
1	X1:Female victims	112	465	19	446	406	54	-170	59	12
2	X1:Male victims	55	465	39	-911	406	109	347	59	24
3	X2:Different household-dwelling unit	68	659	30	265	49	12	937	610	218
4	X2:Same household-dwelling unit	98	659	21	-184	49	8	-651	610	151
5	X3:Between spouses and co-habitating partners total	78	798	35	601	317	68	-740	481	155
6	X3:Cousin	0	12	39	362	0	0	2031	12	7
7	X3:Former spouse	18	394	45	1054	137	49	1441	257	138
8	X3:Siblings	11	206	41	288	6	2	1719	200	114
9	X3:Suspect is the parent of the victim	59	779	52	-1169	756	195	203	23	9
10	X4:0 - 4	13	133	42	-1240	127	47	261	6	3
11	X4:10 - 14	15	130	41	-1106	125	46	210	4	2
12	X4:15 - 17	6	41	39	-606	14	5	856	27	16
13	X4:18 - 20	7	5	37	224	2	1	247	2	1
14	X4:25 - 34	34	123	35	687	123	39	27	0	0
15	X4:35 - 44	32	114	36	692	114	37	-51	1	0
16	X4:45 - 54	22	80	37	663	66	23	-302	14	7
17	X4:5 - 9	25	300	46	-1304	298	102	104	2	1
18	X4:55 - 64	9	43	38	615	22	8	-594	21	12
19	X4:65 -	4	22	38	463	5	2	-869	17	10
20	X5:Suspect females	38	290	40	-988	288	89	-75	2	1
21	X5:Suspect males	129	290	12	292	288	26	22	2	0
22	X6:Assault 21:5	131	376	10	-227	190	16	-224	186	24
23	X6:Attempted manslaughter, murder or killing 21:1-3,34a:1	1	0	38	184	0	0	-5	0	0
24	X6:Deprivation of personal liberty 25:1-2§	1	2	38	524	2	1	50	0	0
25	X6:Extortion, aggravated extortion 31:3,4	0	10	38	1220	4	2	1536	6	4
26	X6:Menace 25:7§	30	324	38	842	158	52	861	166	82
27	X6:Rape, aggravated rape, rape (3 mom) 20:1-2	3	32	39	1071	18	7	972	15	9
28	X6:Robbery 31:1-2	0	0	38	665	0	0	-2	0	0

Figure 24: Column profiles information from the summary of mjca.

Analysis of the column profiles performed by reading the summary on columns obtained from mjca. Firstly, it is observable that highest row mass belongs to the modalities *Female victim's*, *Same household-dwelling units*, *Between spouses and cohabiting partners*, *Suspect males* and *Assault*. It implies that a lot of individuals are falling into these categories, which coincides with observations from uni- and bivariate cases. Parameter X.qlt is the sum of the parameters cor and cor1 and roughly describes how well the particular modality is represented by combination of chosen PC. For instance, it is notable that the first five modalities and modality *Suspect is the parent of the victim* have relatively satisfactory level of representation, while modalities with X.qlt $<$ 250 are poorly represented. Finally, parameters ctr and ctr.1 represent contributions to the construction of the particular PC. For example, it is observable that modalities *Male victims*, *Suspect is the parent of the victim* and *5-9* have greatest contribution towards construction of the component 1, while modalities numbered 3, 4, 5, 7 and 8 are the most important in the construction of component 2.

4.3 Qualitative discussion of results

Qualitative interpretation of results is based on the careful analysis of modalities' projections plot onto the plane spanned by first two principle components. Moreover, qualitative analysis is supported by technical considerations from 4.1, 4.2 and general background information from uni- and bivariate analyzes.

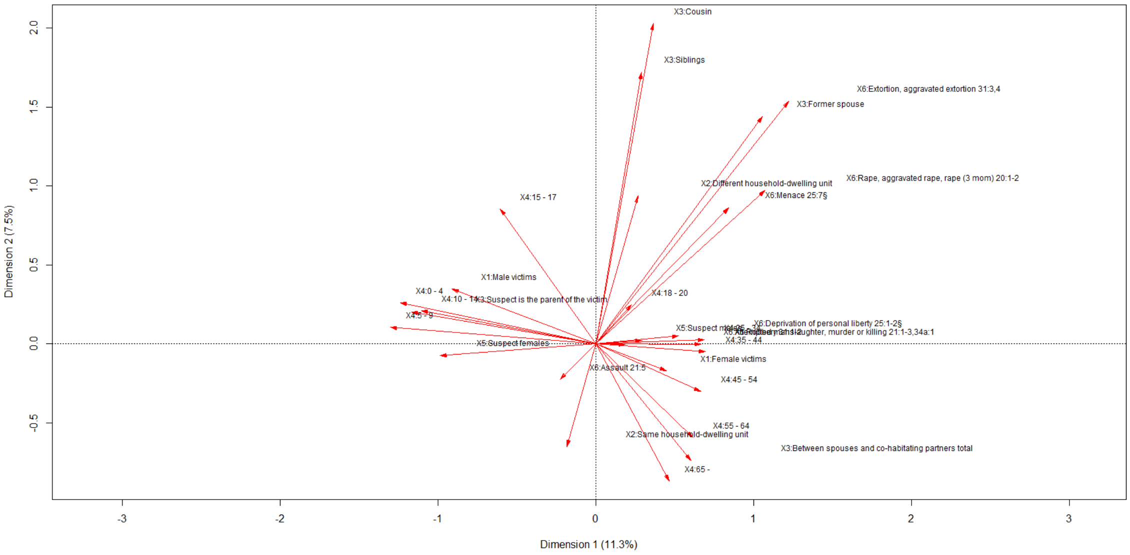


Figure 25: Projection onto plane spanned by first two principle components.

To start with, we note that a single vector represents the profile of particular modality. Modalities possess similar profile, if their representative vectors are close to each other (have sharp angle in between), while the opposite holds for modalities with different profiles. In addition, rare modalities are represented by vectors with large norms (far away from the origo), while more average-inclined modalities are represented by vectors with smaller norm (closer to the origo). Analysis is guided by this rule.

Firstly, observe that the modalities *5-9*, *0-4*, *10-14* and *Suspect is the parent of the victim* possess similar profiles. Moreover, modalities *Suspect females* and *Male victims* are also lean towards these categories, even though they seem to have slightly higher deviation comparing to the cluster formed by aforementioned groups. In addition, based on the length of representative vectors, we observe that these modalities are not far away from average profile, but also are not identical to it.

Secondly, it is observable that modalities *Suspect males*, *Deprivation of personal liberty*, *25-34*, *35-44*, *Robbery* and *Attempted manslaughter* possess similar profiles. In addition, modality *Female* *victims* and *45-54* also exhibit similarity with these categories, even though they have higher deviation from the profile formed by the cluster of aforementioned groups. Moreover, it is observable that modality *Suspect males* seems to be very close to the average profile, as well as modality *Attempted manslaughter*.

Furthermore, we observe that there exist modalities with similar, but very rare profiles. For instance, consider modalities *Cousin* and *Siblings*, which seems to be far away from origo, but closer to each other. In addition, consider modalities *Extortion* and *Former spouses*, which exhibit similar behavior, but also seem to be slightly attracted to the modalities *Rape* and *Menace*.

Finally, it is notable that some of the modalities are harder to interpret in terms of associations: they do not form any groups. Consider, for instance *Assault* or *Same household-dwelling units*, which seem to be relatively close to each other and the average profile, but far away from any other groups. The same holds for modalities *Between spouses and cohabiting partners*, *55-64*, *65-*.

To conclude, it is also worth an attempt to interpret the dimensions. For instance, it seems to be that the first dimension splits the offences according to victims age, victims sex and suspect's sex: victims from 0-14, male victims, suspect parents and suspect females on the left while victims from 25-44, female victims and suspect males are on the right.
Second dimension is harder to interpret, but we might note that it splits the offences with respect to the frequent characteristics such as *Assault*, *Same household unit*, *Between spouses and co-habiting partners* (on the bottom) and rare characteristics such as *Extortion*, *Rape* and *Different household unit*.

4.3.1 MCA density plot (**)

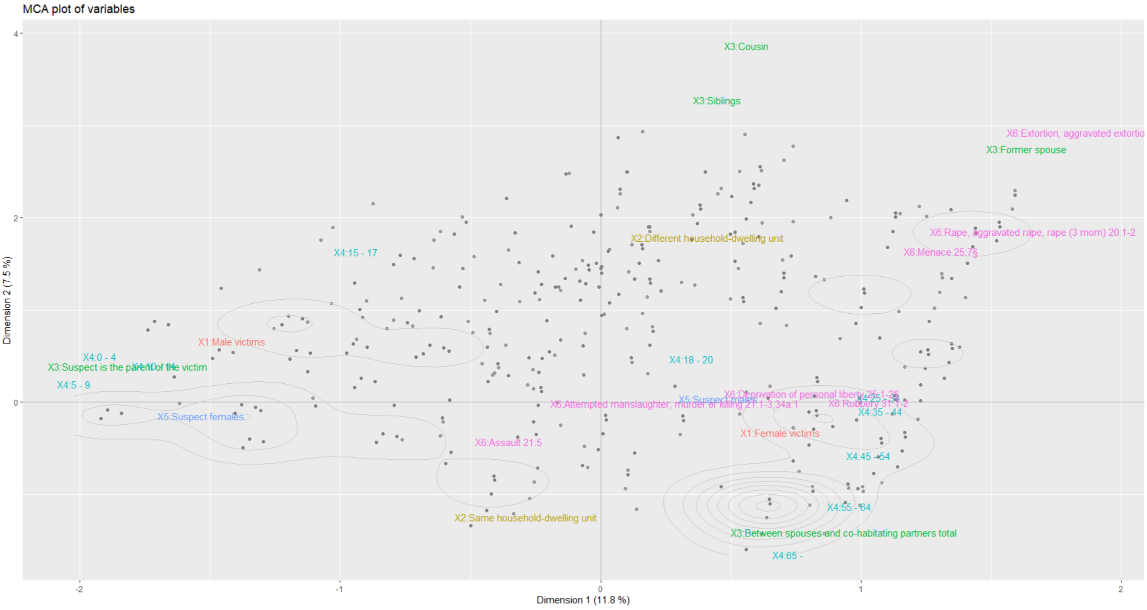


Figure 26: MCA density plot.

To complement the understanding about density of distributions, it is useful to consider the MCA density plot. The density curves indicate the regions of high concentration of individuals. For instance, it is observable that a lot of observations are located near category *Between spouses and co-habiting partners* although on the arrow plot it is challenging to associate this category to some group. In addition to that, it is observable that main cluster formed to the right and left side of the origo also correspond to the high density areas, which was also seen in the arrow plot.

4.4 Answering research questions

Multiple correspondence analysis revealed several underlying properties, that seem to emerge in the nature of domestic violence in Finland. We discuss observed general trends, outliers and questions that are still left open.

4.4.1 General trends

Firstly, it seems to be, that there is correspondence between age of the victim, sex of the victim and sex of the suspect. Namely, we have observed that children from 0-14 are tend to be victims of DV caused by parents and male victims are more frequent in this age then when individual grows older. On the other hand, according to the analysis, female individuals are more frequent victims of DV in the age 25-65- and in this case suspects are mainly male individuals. Even more general observation is that people of opposite genders tend to cause more DV to each other than people of the same gender. However, in general, it seems that male individuals are substantially more prone to DV.

4.4.2 Outliers

In addition to the general structure, MCA revealed rare characteristics of DV offences. For instance, it seems to be that offences committed by siblings and cousins have a lot of similar properties, but on the other hand represent rare characteristics. In addition, we observed that former spouses are prone extortion crimes, however also seem to occur relatively rare. Finally, another rarely meet crime is Rape.

4.4.3 Open questions

Despite of the revealed structural correspondence, there are many open question and properties that are left uninterpreted. For instance, the MCA have managed to characterize the age groups 0-4,5-9 10-14 and 25-34, 35-44, 45-54, but there is a challenge to understand structural correspondence of the group 65- and especially groups 15-17, 18-20. In addition, univariate analysis has revealed high frequencies of the offences *Assault* and *Between spouses and co-habiting partners*, but it is hard to derive strong structural association to other crimes. Taking into account the fact that modality *Between spouses and co-habiting partners* has very high quality of representation according to MCA summary statistics, this is especially interesting observation. Finally, it is also difficult to understand the role of the household unit type. Household types are seemed to be well separated by the second dimension, however, it is a question whether they associate significantly with any other characteristic of DV crimes.

4.4.4 Words of warning

Firstly, it is good to remember the results obtained from analysis of PC and variance. Current interpretation is able to explain only $\frac{1}{5}$ of the information in the data, while the heavy major part

slips away due to possibly technical implementation and/or data related reasons. This possibly suggests that very strong and general conclusions should be considered with proper care.
Secondly, the statistics covers only reported cases. There exist a vast of other research findings, which suggests that at the best case, only 40% of DV crimes are reported, while in some cases 20% or even less. Therefore, there is a reason to believe that largest proportion of information were not even present in the data. Accordingly, there is high probability that true picture of the phenomena is far away from the conclusion derived in the current research.
Thirdly, one should be very careful with interpretation of crimes such as robbery, deprivation of personal liberty and attempted manslaughter. It is observable that these crimes seem to have a lot of similarities with average profile and group formed by *Suspect males* and other categories. However, their frequency of occurrence is less than 1%. It is possible that presence of such variable has affected MCA [3] and therefore, it is better do not rush to associate them as similar with the average profile.

5 Suggestion for further improvements

One caveat of this analysis is variable-modality ratio. For instance, many of the variables contain only 2 modalities, such as victim's sex or suspect's sex, while some other modalities such as victim's age contain 10 modalities. Such non-uniform distribution of modalities across categorical variables has effect on analysis and interpretation of results. Potential improvement is to cluster age group together such that as a result one would obtain at least twice less groups that current data set has. However, this approach may lead to loss of more precise information about age categories and their correspondence. This is especially important to consider when we have observed that this correspondence exists.
Another improvement point is to eliminate rare modalities. For example, some types of offences such as robbery, extortion, deprivation of personal liberty and attempted manslaughter have very low frequency. For instance, frequency of modality *Robbery* constitute only 0.1% of relative frequency. Presence of such modalities makes MCA quite a non-robust method and has significant impact on the analysis.[3] Moreover, elimination of rare modalities will also improve variable-modality ratio.

6 Sources of bias

6.1 Data selection bias

One of the most common statistical biases is data selection bias. Apparently, this is very relevant for the research in question. During the variable selection and sample construction process, the emphasis was made on so-called common sense characteristics. However, there are no means to guarantee that the selected subset of variables and categories more or less represents the true population and is actually the most valid for generalized inference. In addition, no expert knowledge to support the selection process was involved. Therefore, there is a strong reason to consider the variables were chosen in a biased way.

6.2 Detection bias

Detection bias occurs, when the researcher is aware that certain subset of study subjects will most likely exhibit the phenomena. Therefore, the researcher may emphasise this specific subset of study subjects, disregarding the other. It is necessary to mention that for the research in question, researcher was aware that it is more likely to observe phenomena from certain study group. Therefore, to increase objectivity, it is good idea to reconsider analysis and interpretation of results by evolving expertise of other individuals.

6.3 Confirmation bias and belief bias

These are types of cognitive biases which occur when researcher has preconceptions and emphasises the only findings confirming these preconceptions. For the research in question, this type of bias is also relevant due to researcher prior beliefs. One is encouraged to complement and re-evaluate findings as well as to emphasize potential sources of confirmatory preconceptions.

References

[1] Domestic violence and intimate partner violence reported as an offence, 2009-2019.

[2] J. Hair. *Multivariate Data Analysis(p. 114).* 2014.

[3] Paulina Ilmonen. Lectures on multiple correspondence analysis, March 2021.