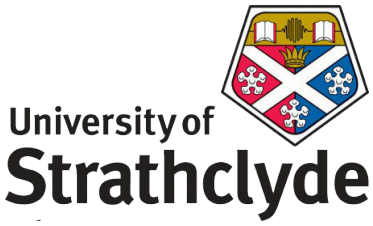


CS982: Big Data Technologies

"Stay at home". Domestic abuse and COVID-19: the Italian case

Contents

List of Figures

1 Introduction

The outbreak of COVID-19 has brought an unprecedented global public health crisis. Severely affected countries have adopted preventive measures to combat the spread of the disease: lock-downs, travel restrictions, mandatory face masks, social distancing. In the background, another threat to global public health has unfortunately silently intensified. "Stay home!" began to sound like a prayer that citizens made to protect, be protected, and save lives. However, home can be a dangerous to be forced in. Domestic violence, in particular against women and children, is on a rising trajectory. The World Health Organisation (2021) reports that 1 in 3 women experience violence globally. Most of the victims of feminicide, the intentional killing of women because they are women, had been sexually, physically, financially or psychologically abused by friends, partners, or family members (Wehnam et al., 2020).

As a survey conducted by the non-profit Women's Aid (2020) shows, 67% of survivors who were still experiencing abuse and violence at the time of the lockdown reported that it got significantly worse, and 72% had seen their abusers getting more control over their life.

Bertolucci (2021) outlines that in Italy almost 7 million women (or one in three) have experienced violence in their lifetime. According to the Italian National Istitute of Statistics (ISTAT) (2021), only in March 2019 there was, on average, one victim of gender-based violence against women (GBVAW) every fifteen minutes. An investigation conducted by the Italian Ministry of Internal Affairs (2021), "Violated Lives", reports that 77 women were killed by hands of their partners in the first semester of 2021, against the 55 victims of the first semester of 2020. The social and human cost of GBVAW is incommensurate and incommensurable, and it therefore fundamental to conduct a research in order to recommend policy implementations that counteract GBVAW, offer women safe ways, methods, and means of communication, to escape from their abusers, and to inform about the dynamics that are behind this form of often unnoticed and unreported violence.

Italy was the first country in Europe to face the COVID-19 emergency. The Italian government led by PM Giuseppe Conte, from 9 March 2020 up to the 18 May 2020, implemented a series of restrictive lockdown measures to contain the spread of COVID-19. The paper will use publicly available data offered by the ISTAT and proceed in its estimations by looking at the numbers of calls received by the anti-violence helpline 1522. A number of studies have been conducted in a similar manner. For example, McCrary and Sanga (2021) use 911 call records and phone location to find that domestic violence in the US icreased by 12% and that the rate of first-off abuse increased by 16%. investigates With an emphasis of intimate partner violence (IPV), this paper will investigate the change in reported case of domestic violence in Italy during the lockdown. The aim is twofold. First, it aims to gain a deeper understanding of how gender-based violence against women presented itself on a territoriality, economic, and social scale, and, second, to investigate the possible factors that affect the rates of domestic abuse and IPV.

2 Dataset - The helpline 1522 during the pandemic (Quarterly data at Q2 2021)

2.1 Source

The data used in this report is freely available and accessible on the Italian National Institute of Statistics' analytics website. The data is provided directly from the call-centers to the Italian Ministry of Equal Opportunities and National Institute of Statistics. The dataset consists of 21 tables in Excel format. These have been cleaned and then converted for the sake of a more convenient analysis. Table 1 offers an overview of the tables used in the analysis. All numerical time series data are in absolute values. The dataset spans data from Q1 2018 to Q2 2021. It offers information about the number of calls received by the 1522 helpline across time for different reasons and regions, the number of victims, a categorical aggregation of the

emergence of violence for the first time by region and quarter, socio-economic data in regards of both the victims and the perpetrators, and a categorical aggregation of the number of the victims by the places were the abuse happened.

Table	Description
Tab1	Calls to 1522 for reasons of the calls and quarter
Tab16	Place of the violence
Tab8	Calls from victims by region of origin
Tab21	Social Variables of the abuser
Tab12	Social Variables of the victim

2.2 Analysis

This subsection presents an overview and a preliminary analysis of each of the tables taken into consideration.

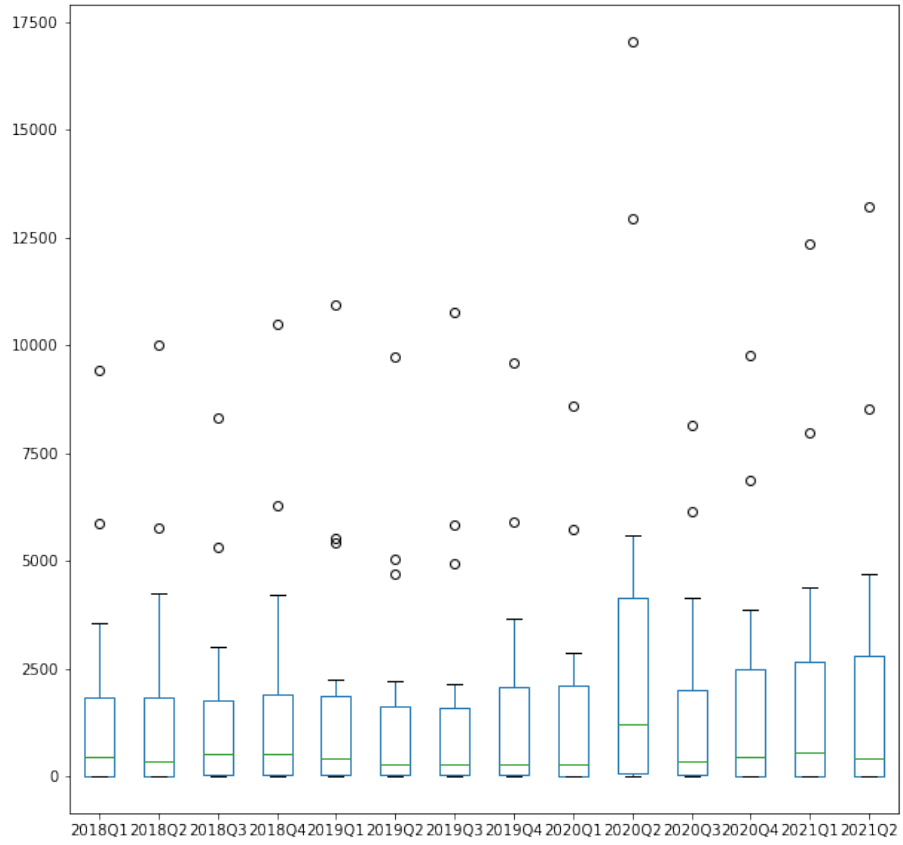
2.2.1 Number of calls received by the 1522 helpline with motivation related to GBVAW: 2018Q1-2021Q2

Table 1 reports the numbers, in absolute values, of valid and not valid calls (joke calls, interrupted calls) to the 1522 helpline by the reason of the call and time period, which spans from Q1 2018 to Q2 2021. The reasons for the call are the following fourteen:

- victim of violence seeking for help
- asking information about the helpline 1522
- asking information about national shelters for victims of violence
- reporting violence or abuse
- asking for useful phone numbers for out of target calls
- victim of stalking seeking for help
- asking legal information
- emergency
- information for professionals on the procedures to be followed in the event of violence
- reporting public services malfunctions
- reporting of media misinformation
- information on legal responsibility of the public services workers
- international after hours calls
- victim of discrimination seeking for help

Figure 1 and Figure 2 show a peak of total calls in Q2 2021, as per hypothesis. Figure 2 shows a comparison between total calls received and total calls by confirmed victims. Figure 3 plots the number of calls by reasons (some are left out as not of interest for the current analysis). Total calls in Q2 2020 increased by 75% with respect to the same period the previous year. The number of calls by victims of violence asking for help sees a sharp increase beginning in Q1 2020, and, after reaching a peak in Q2 2020 (a 159% increase against Q2 2019), it falls, slightly, to a permanent higher level. The number of emergency calls follow a similar path until Q3 2020, after which it rises steeply. The most frequent reason for calling in Q2 2021 is the reporting of violence: 25% of the total (Figure 3, Figure 4). Abstaining from causal inferences, it is reasonable to assume that the national lockdown of March-May 2020 has acted as the driver of this structural break in the data.

Figure 1: Calls received by the 1522 helpline with motivation related to GBVAW: 2018Q1-2021Q2



2.2.2 Place of the violent act reported by victims and quarter: 2018Q1-2021Q2

Table 16 presents the number, in absolute values, of the places where the violent act reported happened. The table has 9 rows, the places, and 14 columns, the yearly quarters. The places are the victim's home, the road, the workplace, other people's home, a public place, and unspecified others. As Figure 5 shows, after being constant

Figure 2: Comparison Total calls and Total victims

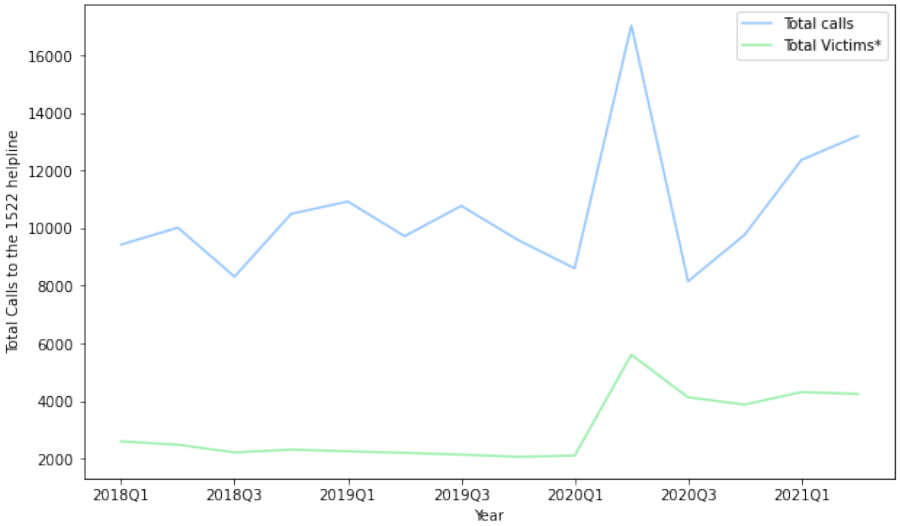


Figure 3: Calls by reasons

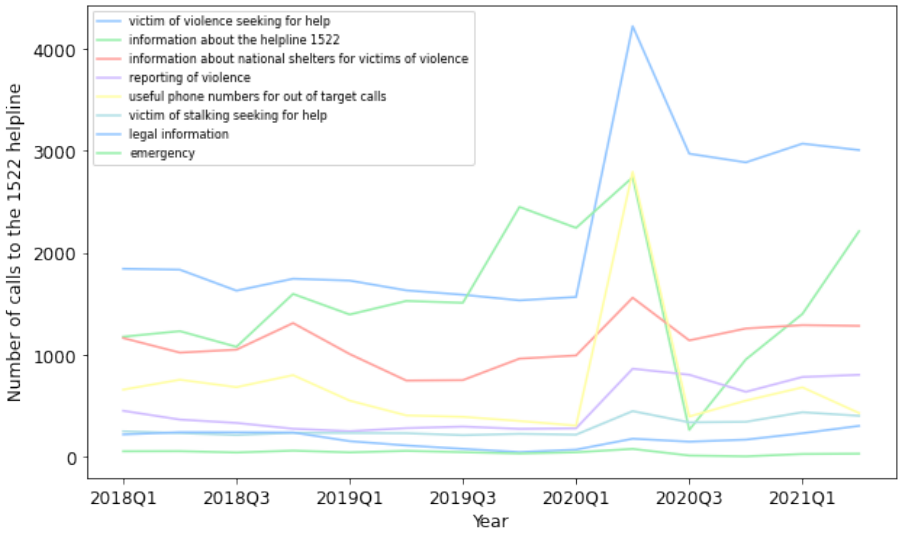
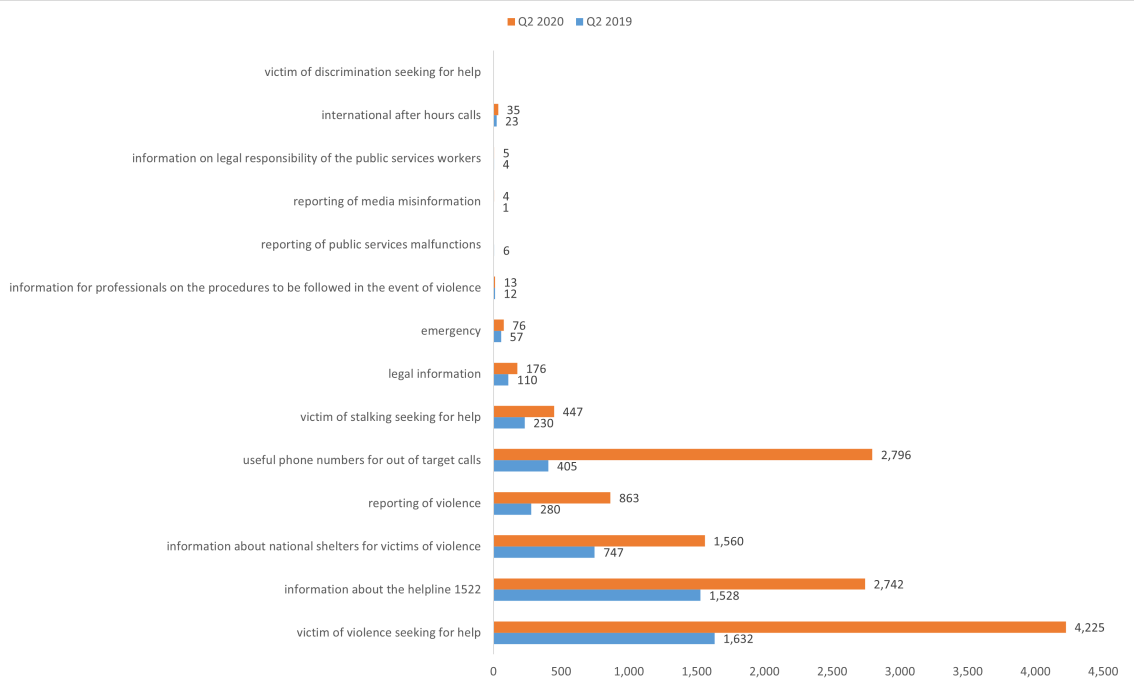
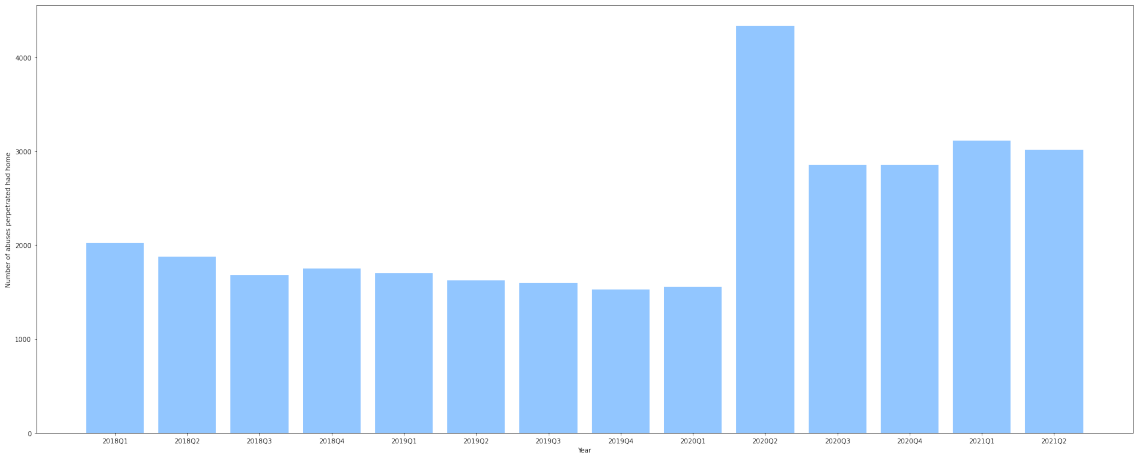


Figure 4: Calls by reasons - Comparison Q2 2019 - Q2 2020



if not slightly declining in the previous quarters, the number of abuses perpetrated at home drastically rises in Q2 2020, and as we also observed for the number of calls by victim asking for help, it has not yet returned to the pre-shock levels (see also Figure 6 and Figure 7). In 2020, 93,4% of the victims was abused at home.

Figure 5: Number of abuses perpretated at home: 2018 - Q2 2021



2.2.3 Calls from victims by region of origin, year and quarter: 2018Q1-2021Q2

Table 8 (22 rows, 15 columns) reports quarterly data, in absolute values, from Q1 2018 to Q2 2021, of the calls from victims by region of origin. The regions are 22 in total. The table also reports the cases where the region was unknown in order to

Figure 6: Abuses by place over time 1a: 2018 - Q2 2021

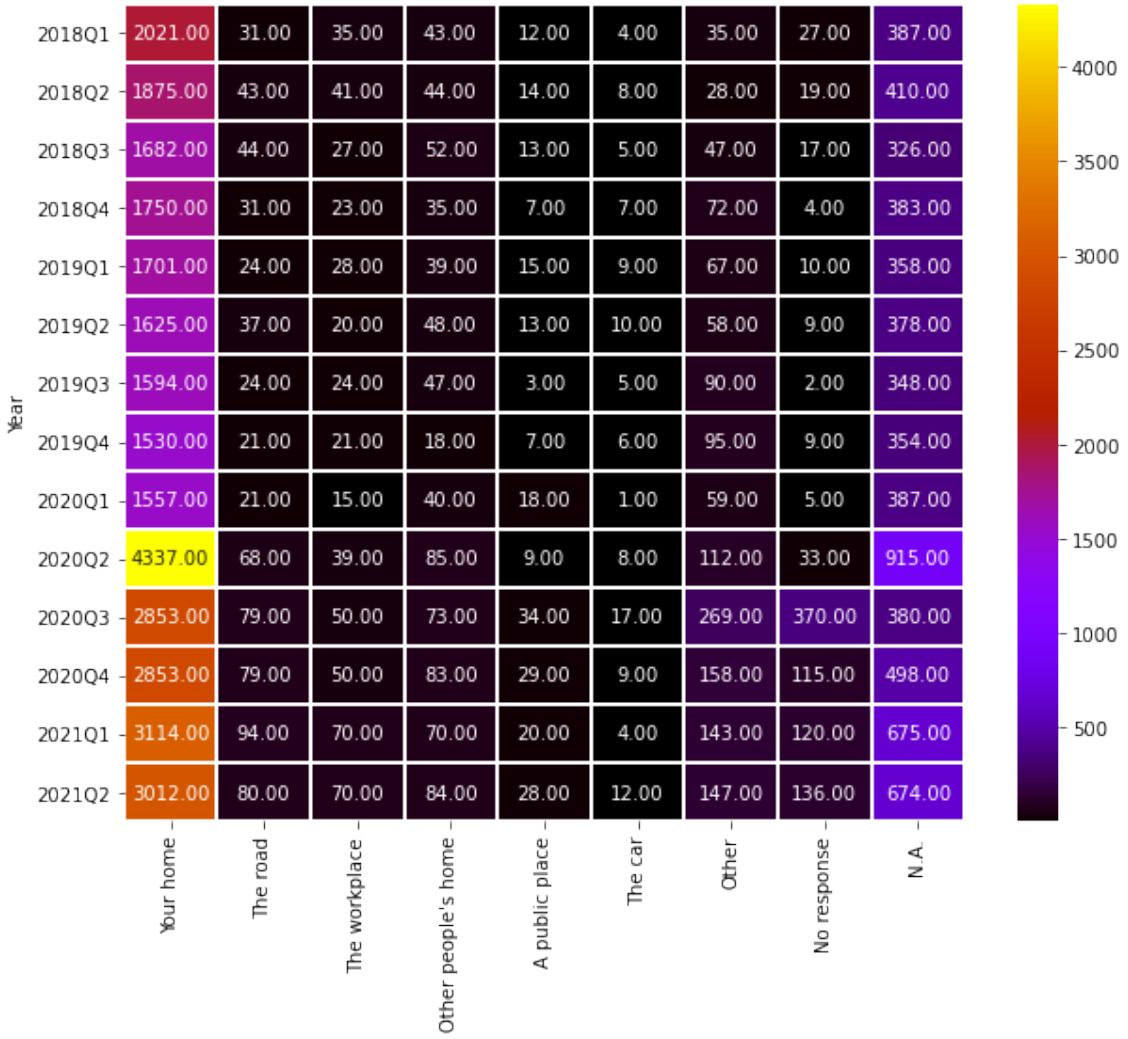
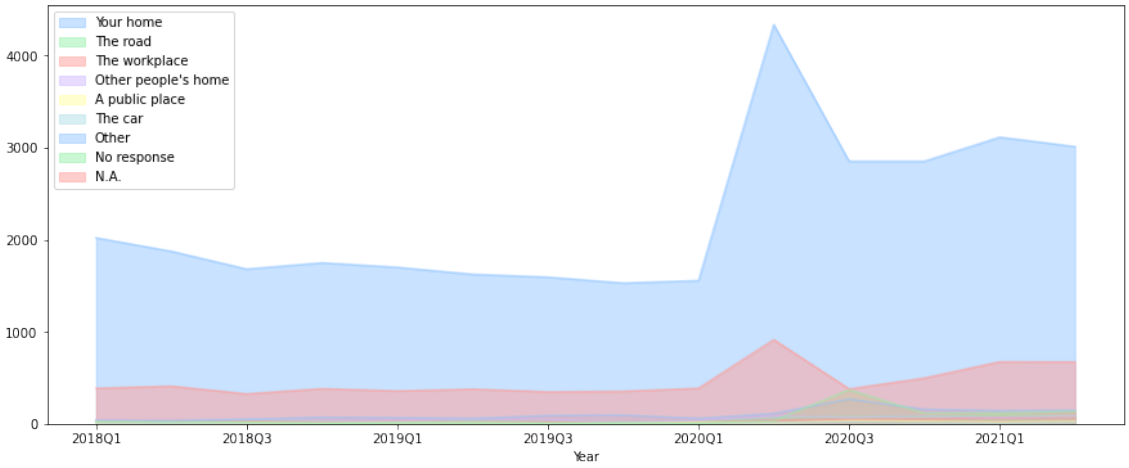
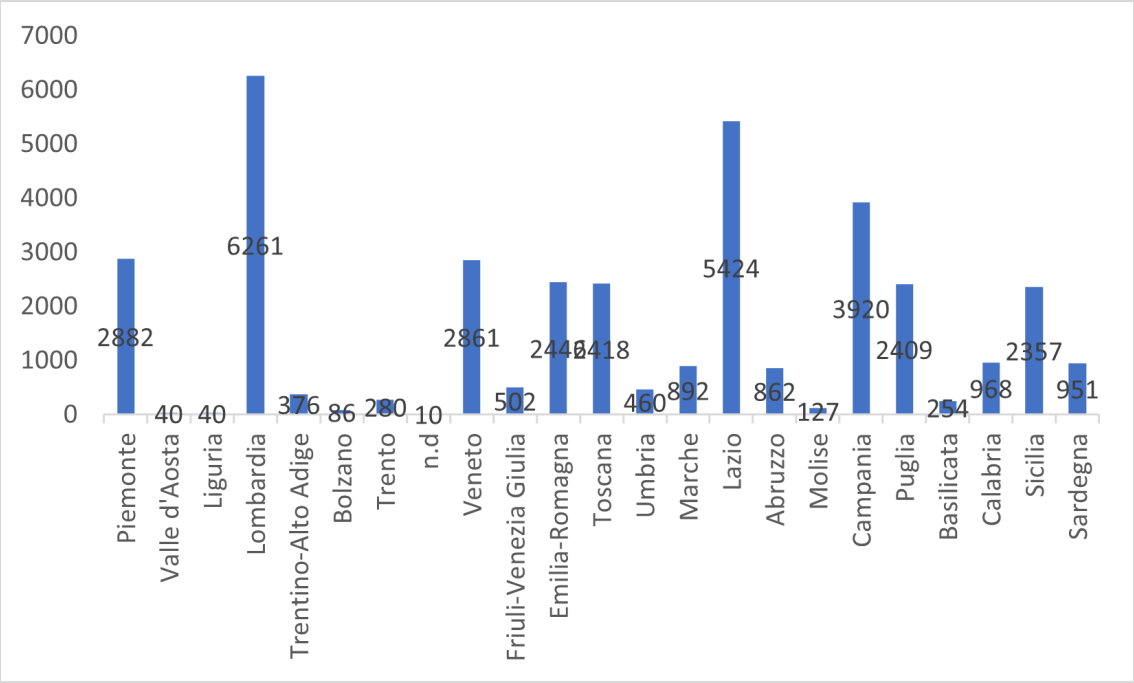


Figure 7: Abuses by place over time 1b: 2018 - Q2 2021



derive nation-wide data. As shown by Figure 8, the regions with the highest number of calls are Lombardia (6,261) and Lazio (5,424). In Q2 2020, in Lombardia, the region that has been hit the hardest by COVID-19 and the first to go in a lockdown, there have been 872 calls (329 in Q2 2019). All the regions have experienced an increase in the number of calls in Q2 2020 versus Q2 2019 (Figure 9) The region with the lowest number of calls is Valle D'Aosta: 40 in total between Q1 2018 and Q2 2021. The dataset does not include data on the population size, and it is therefore not possible to scale the data to draw more accurate comparisons.

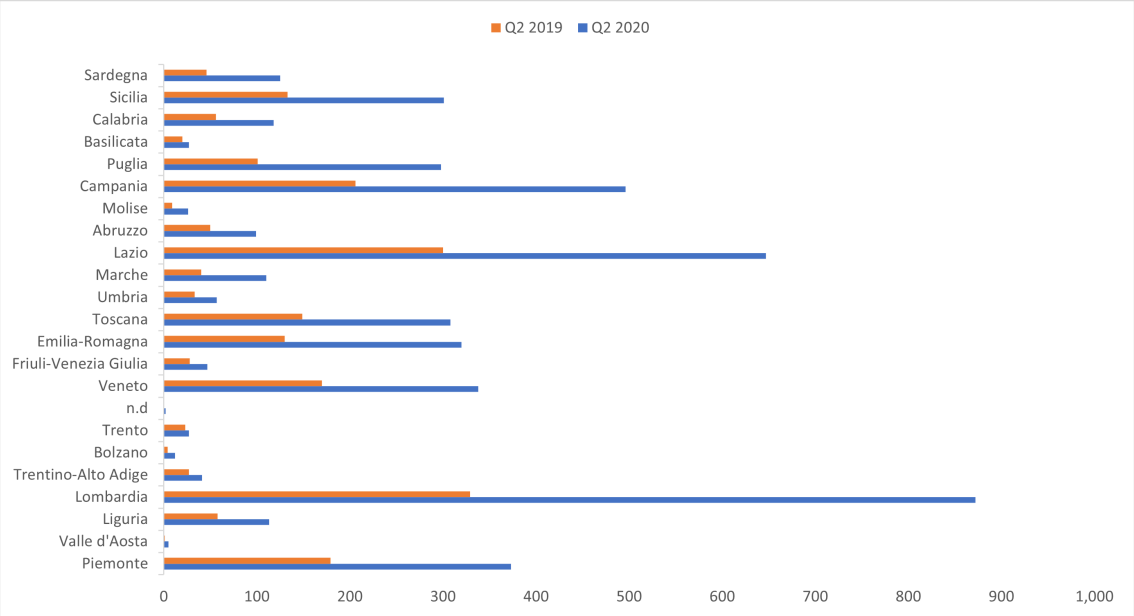
Figure 8: Calls by region of origin - Total



2.2.4 Social variables of the perpetrator and the victim. Gender, age group, marital status, employment status, educational qualifications, citizenship

Table 21 and Table 12 (48 rows, 15 columns) present, as above, a time series, in absolute values, of social variables- gender, age group, marital status, employment status, educational qualifications, citizenship- associated with the the perpetrators and the victims. The striking majority of the victims (41081) are women (men are 1411) (Figure 13). The most affected women belong to the 35-44 age group (Figure 10) and have a high school diploma (Figure 12). As of the marital status, the most affected are married women. As Figure 11 shows, the has been a sharp increase in victims for all the age groups. Figure 14 reports, as expected (trivially), an increase in the number of abusers for each of the social variables across time, with a peak at Q2 2020 for most. The majority of the abusers belong to the 35-54 age group, are employed and married. The matrix in Figure 15 shows a matrix of the correlation between the social variables and supports these findings. Looking at data from a supporting table (25 rows, 10 columns) present in the dataset, we also observe that

Figure 9: Calls by region of origin - Comparison Q2 2019 - Q2 2020



the number of abuses by relationship type is higher for people in a relationship, and that the number of abuses perpetrated by the victim's husband (or wife) has risen by 158% in Q2 2020 in comparison to Q2 2019.

Figure 10: Victims by age - Total

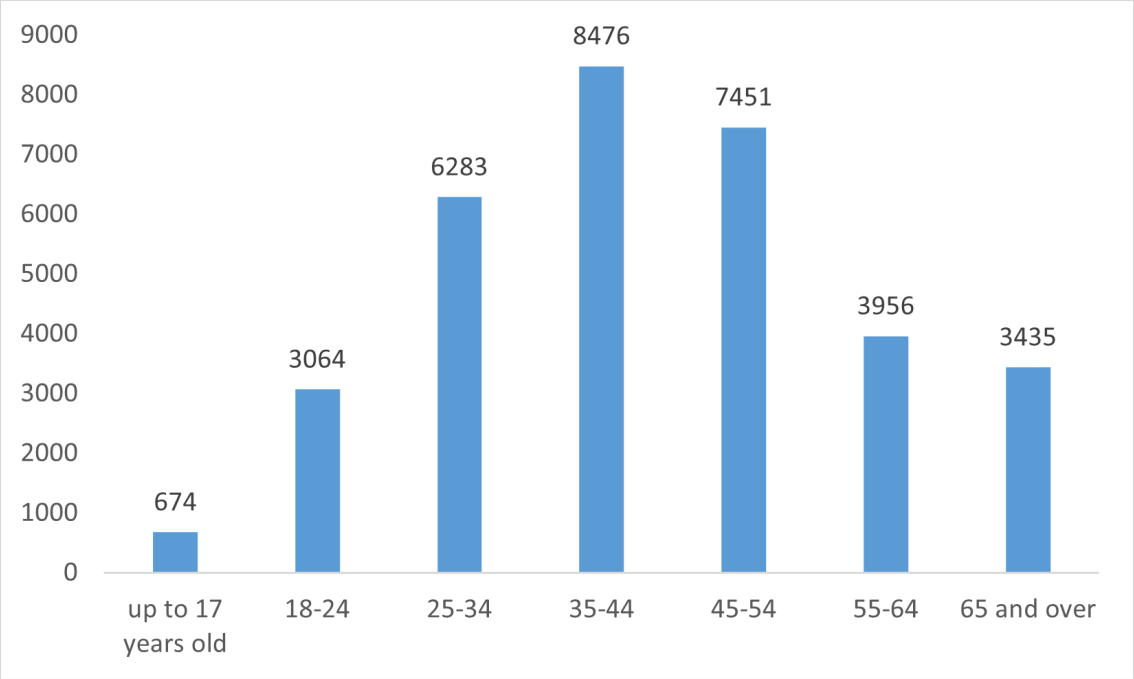


Figure 11: Victims by age - Comparison Q2 2019 - Q2 2020

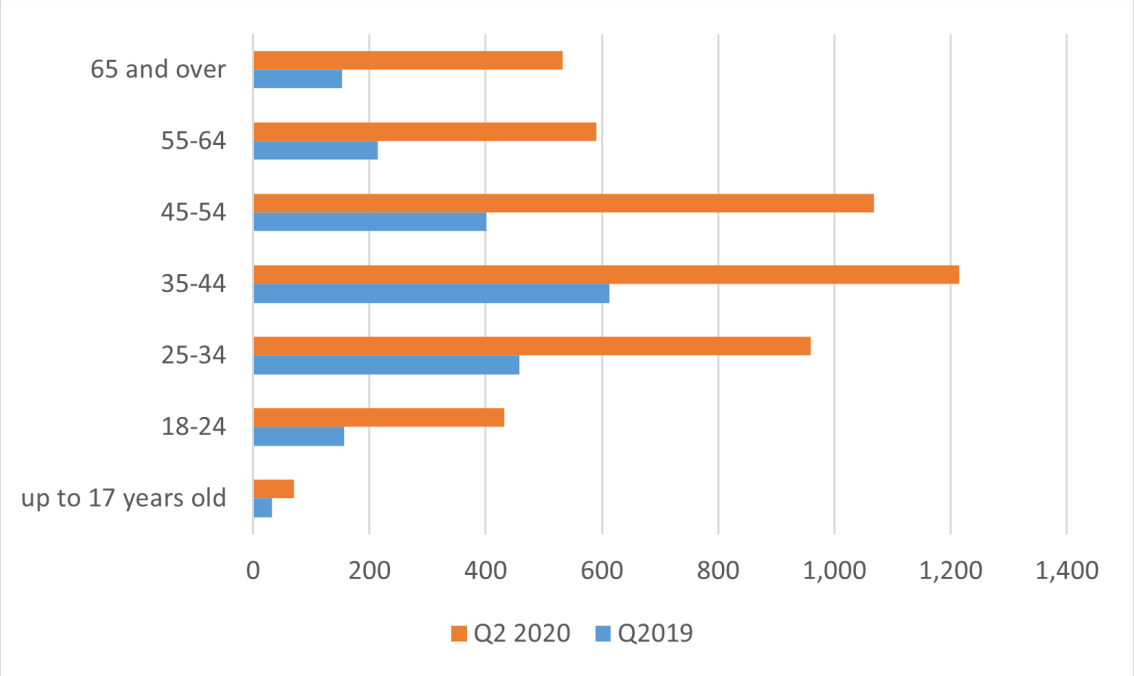


Figure 12: Victims by eduction - Comparison Q2 2019 - Q2 2020

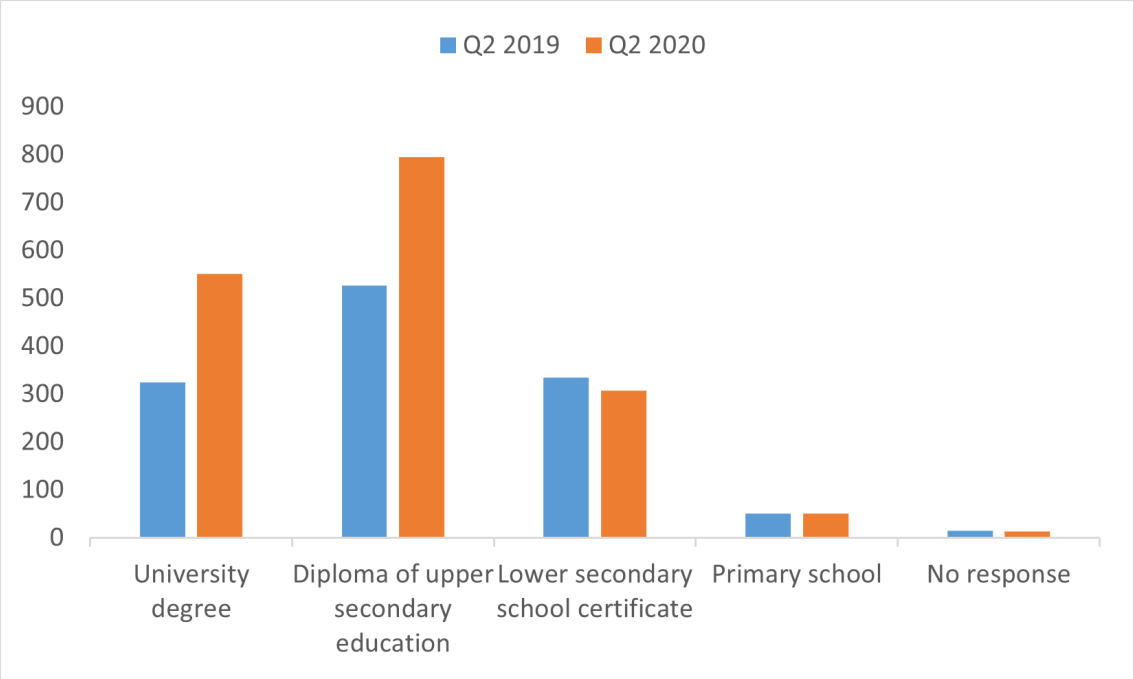


Figure 13: Victims by gender

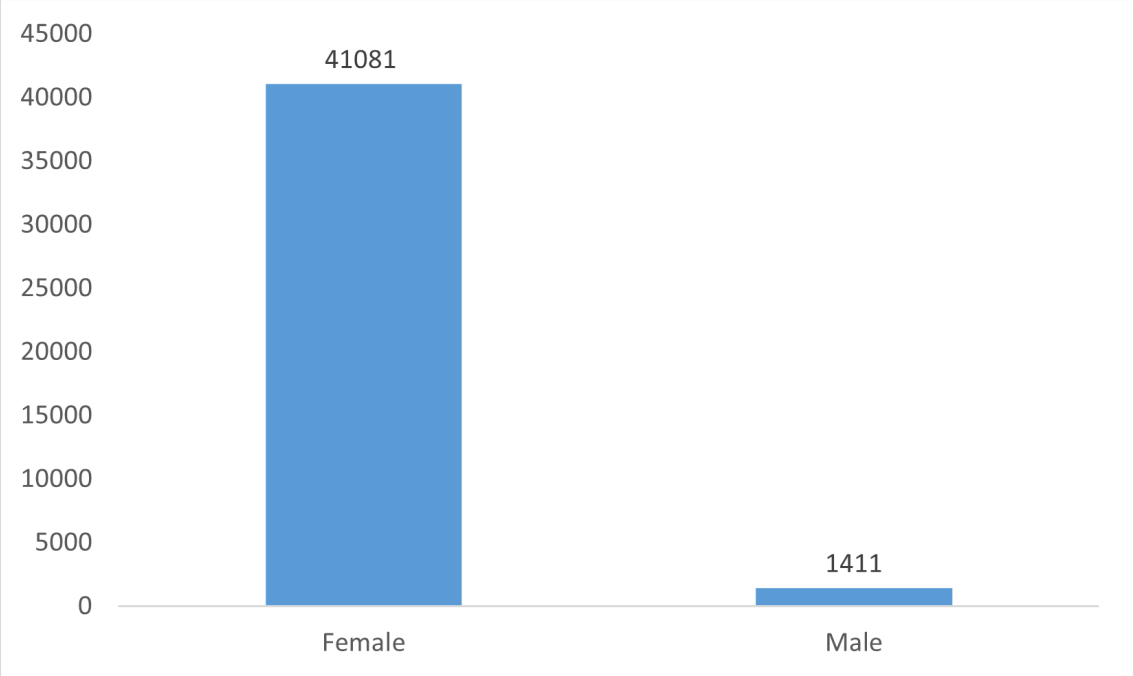


Figure 14: Abusers by social variables over time: 2018 - Q2 2021

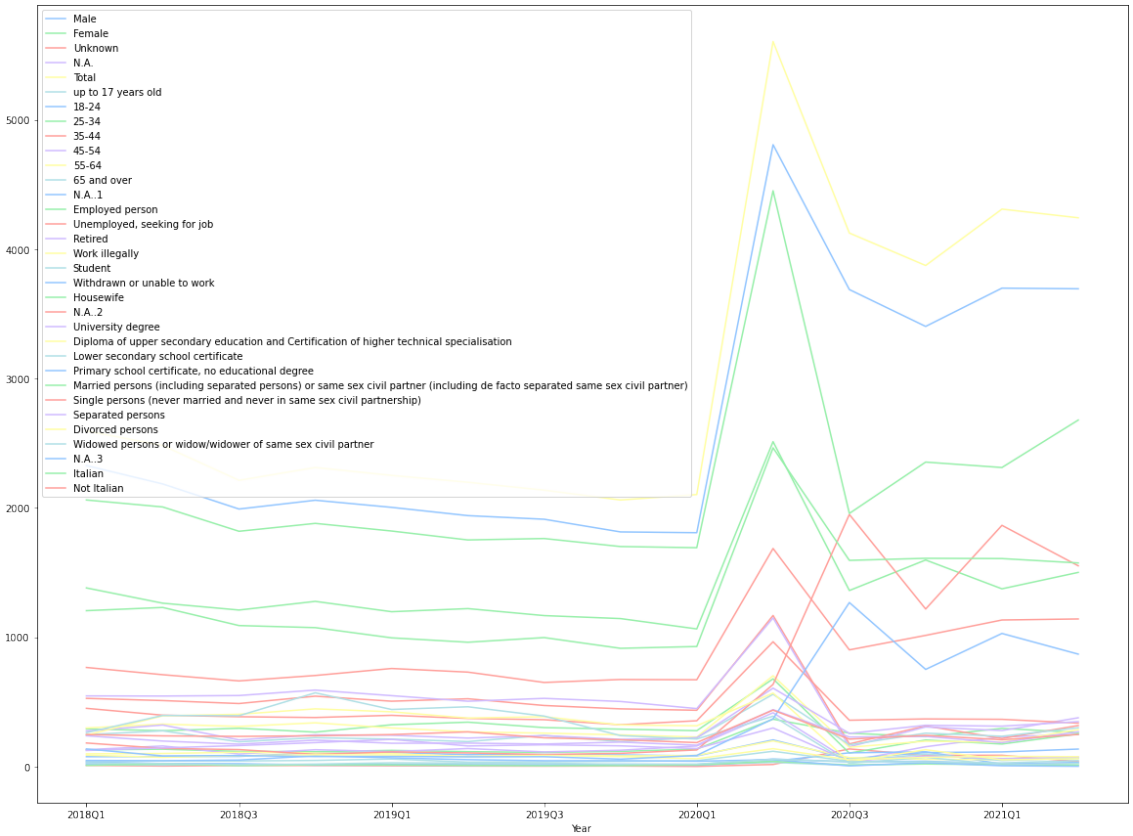


Figure 15: Correlation matrix of the social characteristics of the abusers

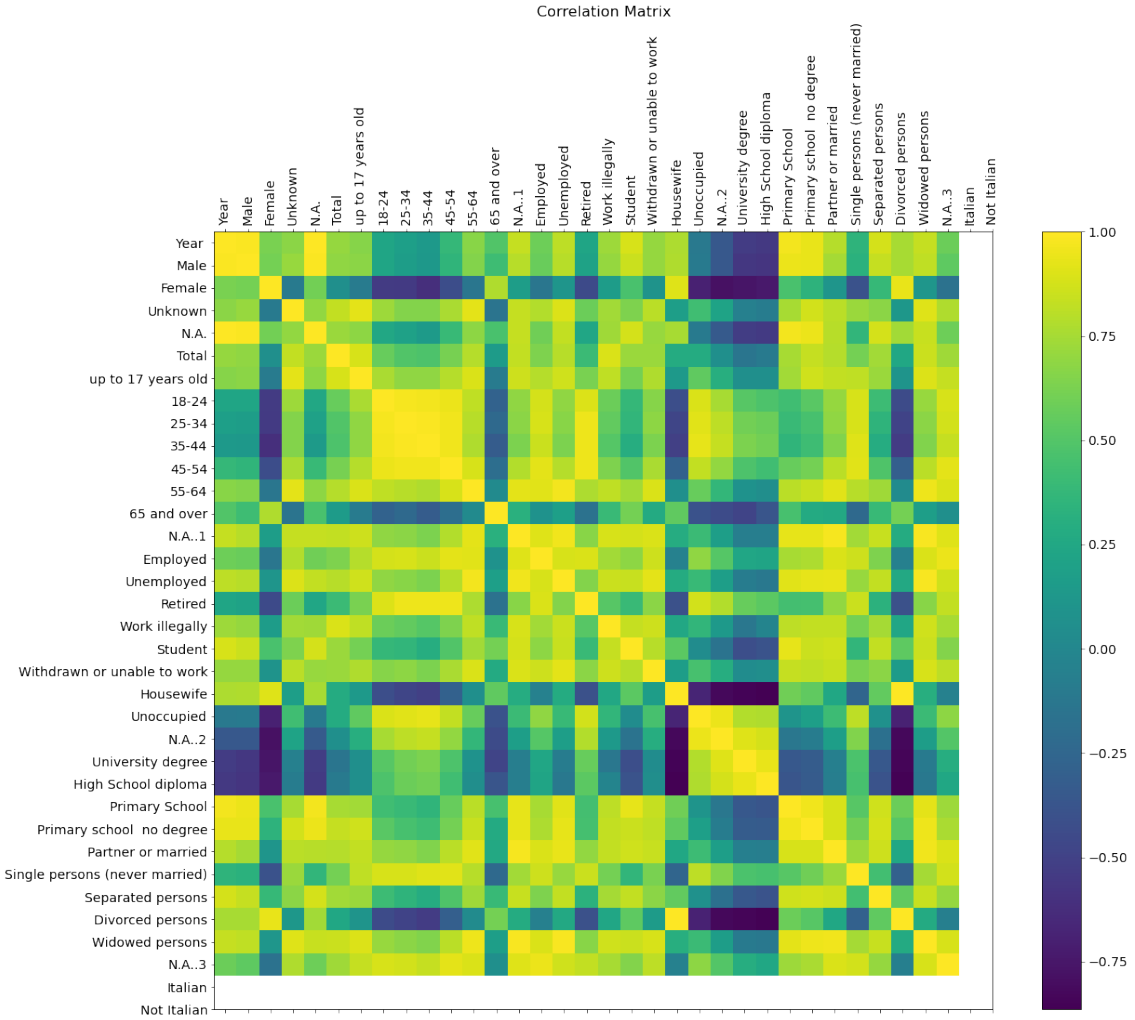


Figure 16: Abusers by employment status

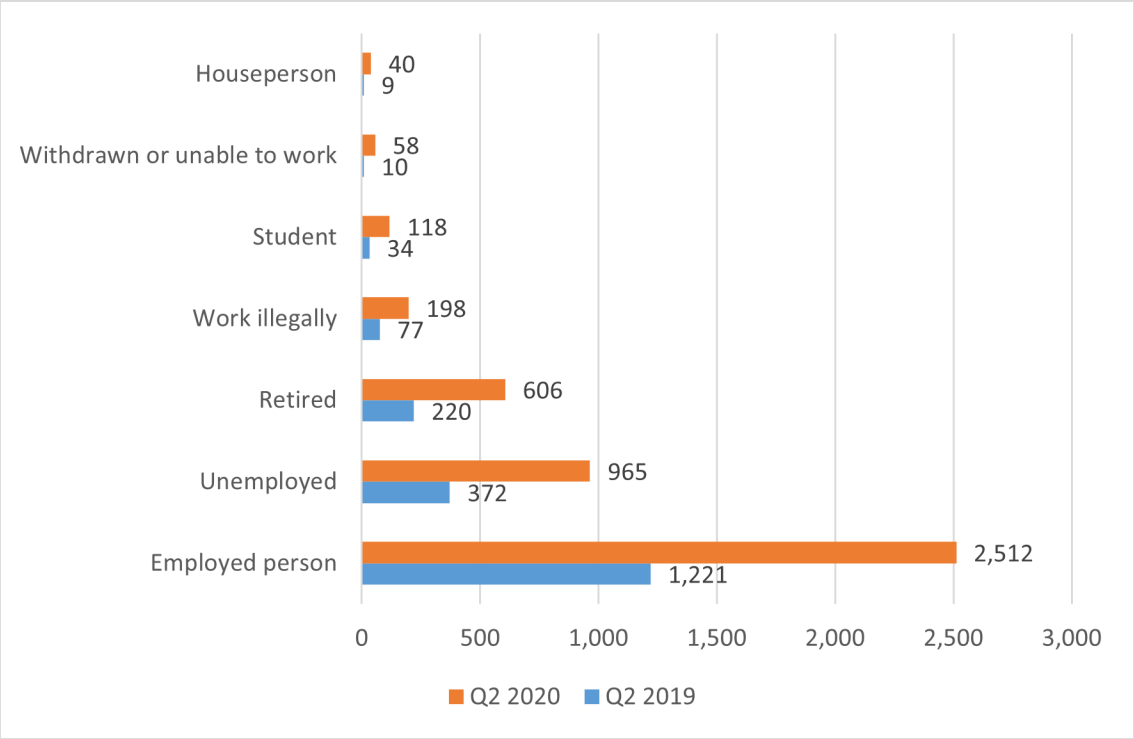


Figure 17: Abusers by age - Comparison Q2 2019 - Q2 2020

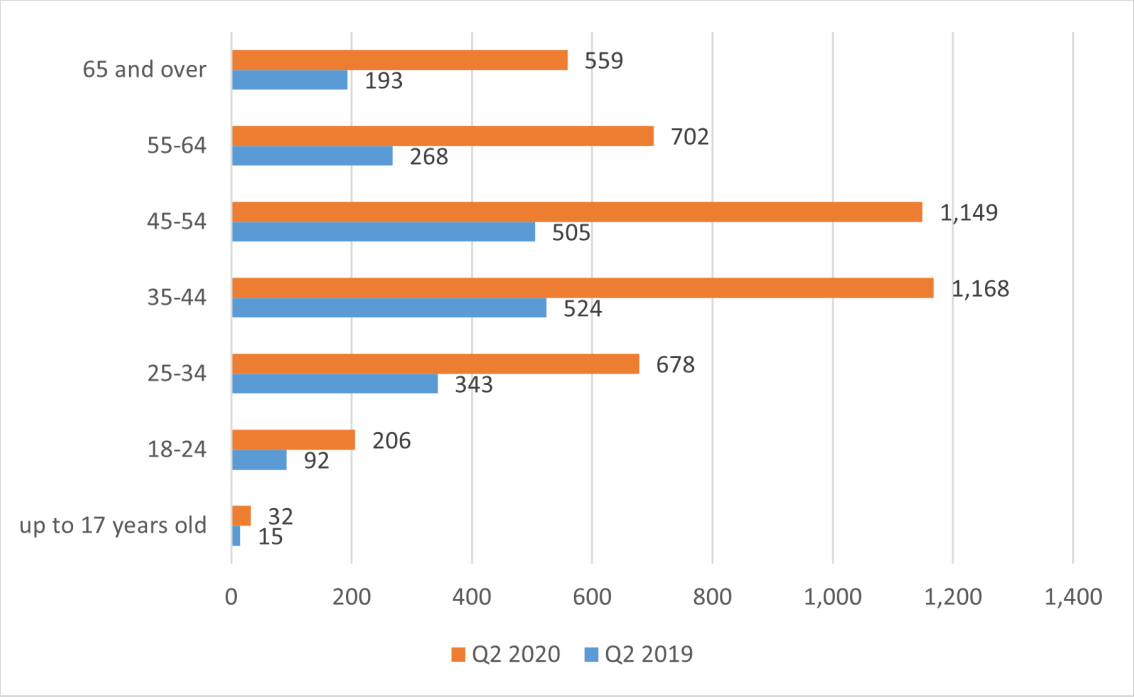


Figure 18: Abuses by relationship type between the abuser and the victim

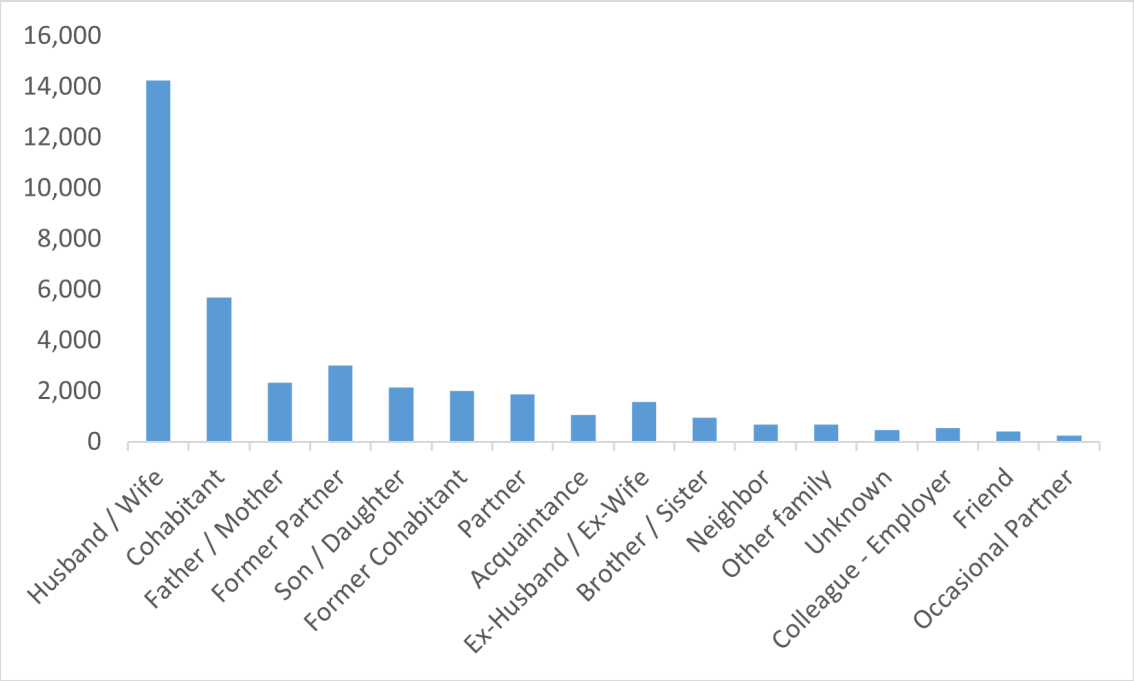
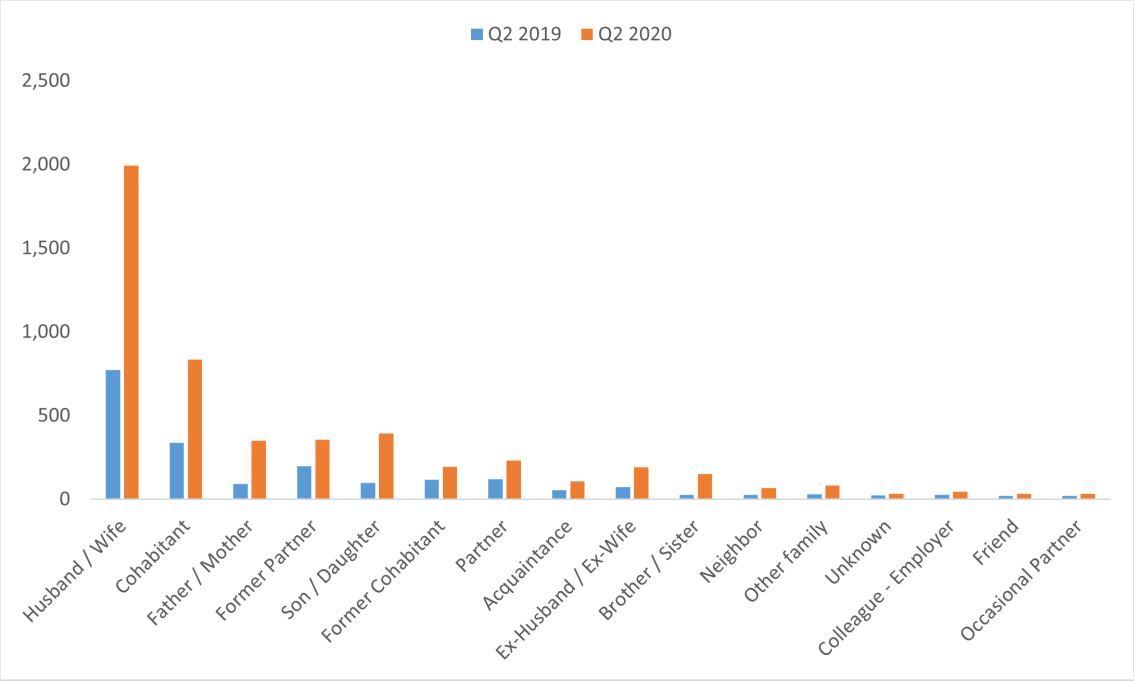


Figure 19: Abuses by relationship type between the abuser and the victim - Comparison Q2 2019 - Q2 2020



3 Unsupervised Learning: learning to make mistakes or making mistakes to learn

3.1 Clustering

The aim of the part of the analysis is to find groupings of data-points, in order, for example, to find which regions have a similar incidence of Gender Based Violence Against Women and which are the social characteristics that the victims have in common and the ones that, instead, the abusers have in common. However, as we can observe from the figures above, the data-set itself is not well suited for this particular kind of analysis and appears already well grouped. Moreover, applying clustering to time series data will be more appropriate to detect anomalies -as in the case of stock price variation over time, for example (Rebbrapragada et al, 2009). Given the limited scope of the dataset's dimensions and data points, as seen above, it was not Nonetheless, this section will proceed in illustrating why this is the case by reporting the results of a disastrous attempt at clustering.

3.2 K-Means

The K-Means clustering algorithm is an unsupervised learning method that groups the dataset into a certain, specified, number of clusters, making sure, by selecting random points, called centroids, and making them the "centre of gravity", that all the most inner points are similar to each other as much as possible. It does so by minimising the Euclidean distance, the sum of the squared distance between the centroid and the data (Na et al., 2018). This is also defined the Ward method to calculate distance.

Given the limited usability of the dataset, stemming from the formatting and general organisation of the data, which prevented a correct and functional concatenation of the tables and a more general investigation, the clustering was performed only on Table12 (Section 2.2.4) for 2 to 12 clusters. Figure 20 shows the data clustered in 4 groups in a pairwise representation across time. The Silhouette Coefficient, a measure of intra-cluster similarity and inter-cluster distance ranging from -1 (wrong cluster) to 1 (best assignment), with values near 0 indicating overlapping clusters, was computed for each sample. Figure 21 shows the Silhouette score for different numbers of clusters: as this increases, the score decreases. The Silhouette Score Visualiser (Figure 22), shows that for 3, 4, and 5 clusters the results are relatively suboptimal. Clustering the data in 2 groups would be, therefore, the best option, as its score is closer to one. The Elbow Method, which computes the distortion score, defined as the Euclidean distance of the data point to its centroid, for all clusters, was also computed. As a rule of thumb, once the scores are plotted, is it possible to visually determine the best value of k as the point at which the line flexes (Figure 23). The Elbow Methods supports the results obtained above by suggesting two as the best number. Compared to the Elbow Methods, the Silhouette score appears to be more useful, as it can be applied to dimensions higher than three which cannot be visualised.

The results of this analysis are counter intuitive with respect to the preliminary

analysis performed above.

Figure 20: Clusters of victims by social characteristics: 2018Q1-2021Q2

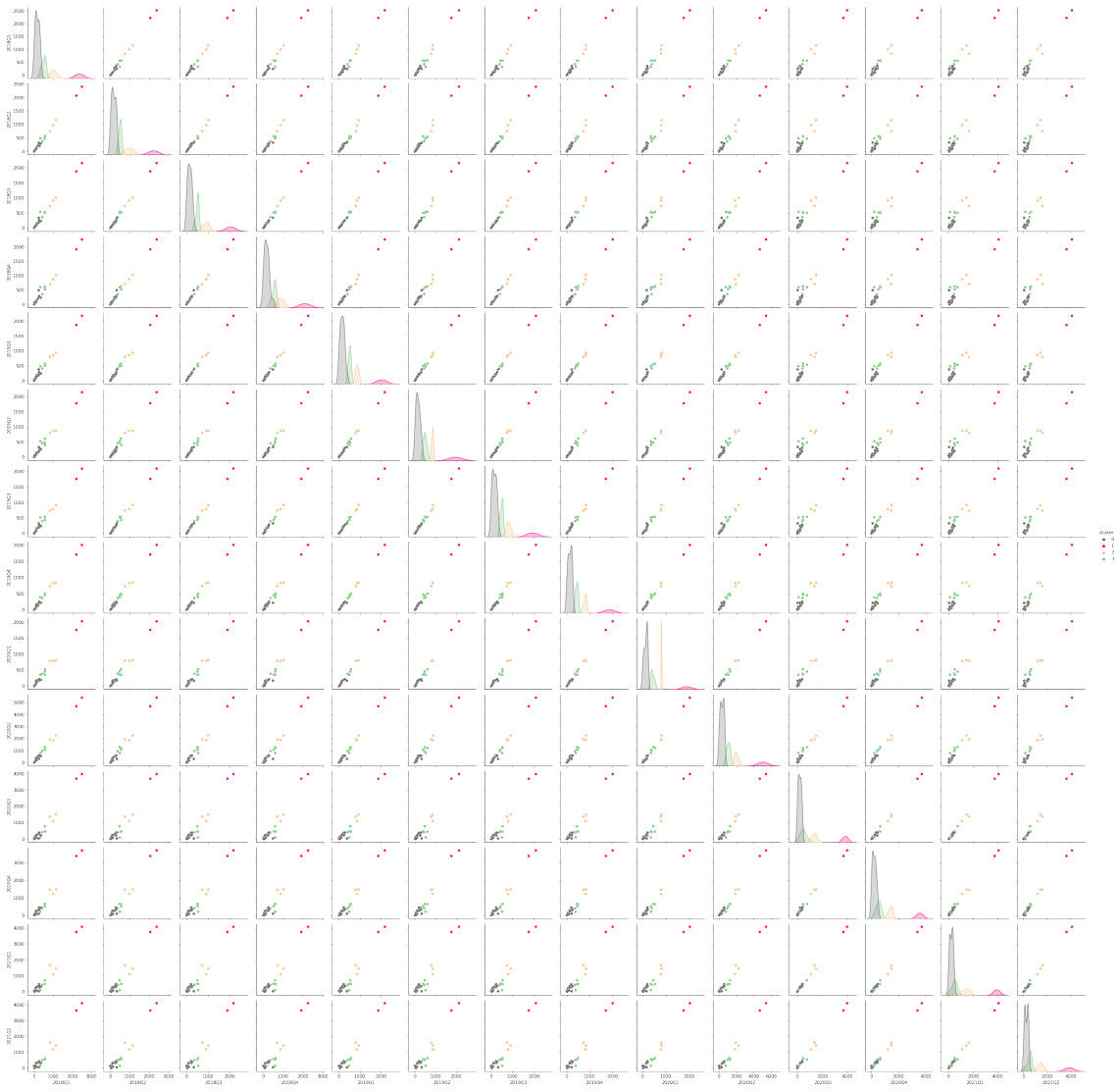


Figure 21: Silhouette Score 1

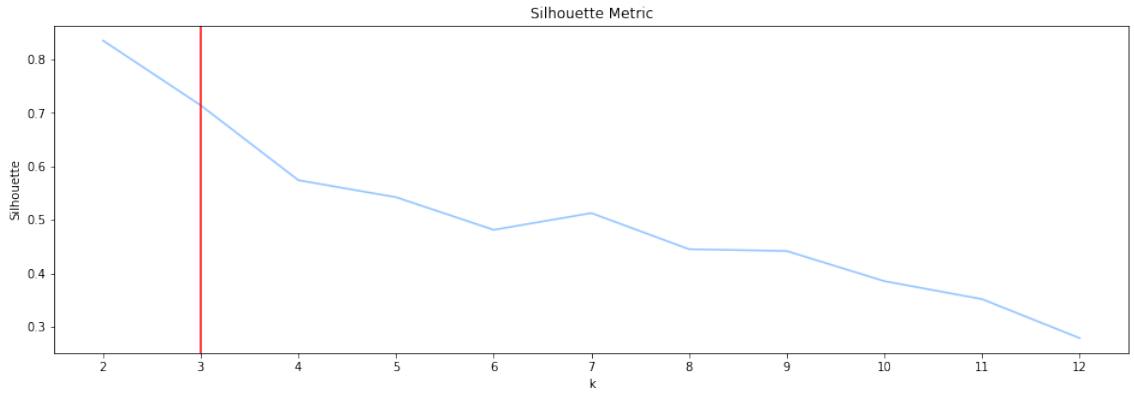


Figure 22: Silhouette Score 2

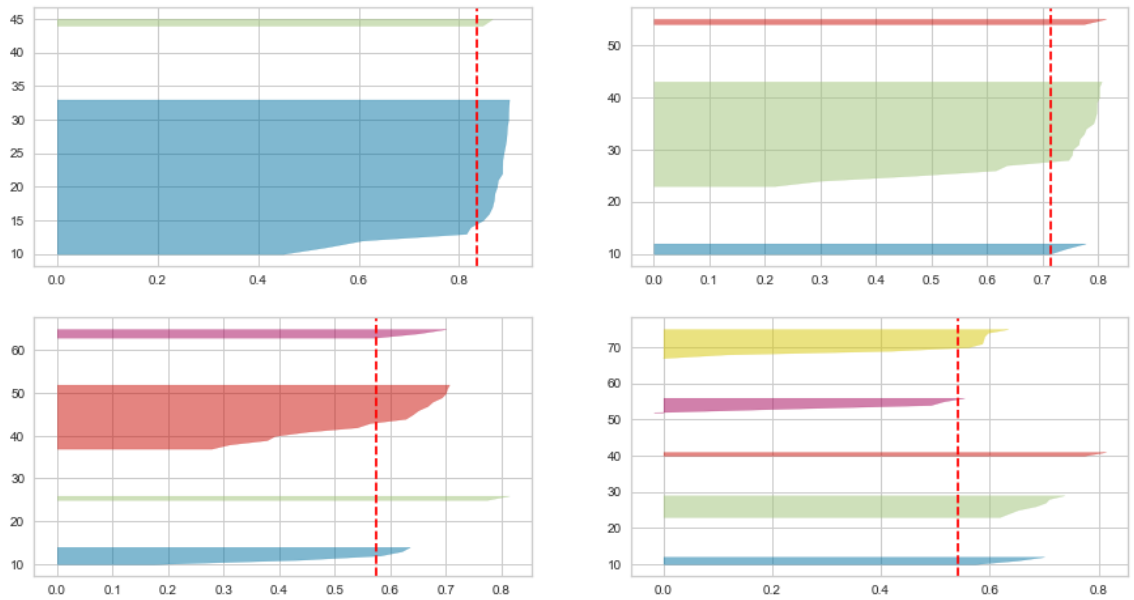
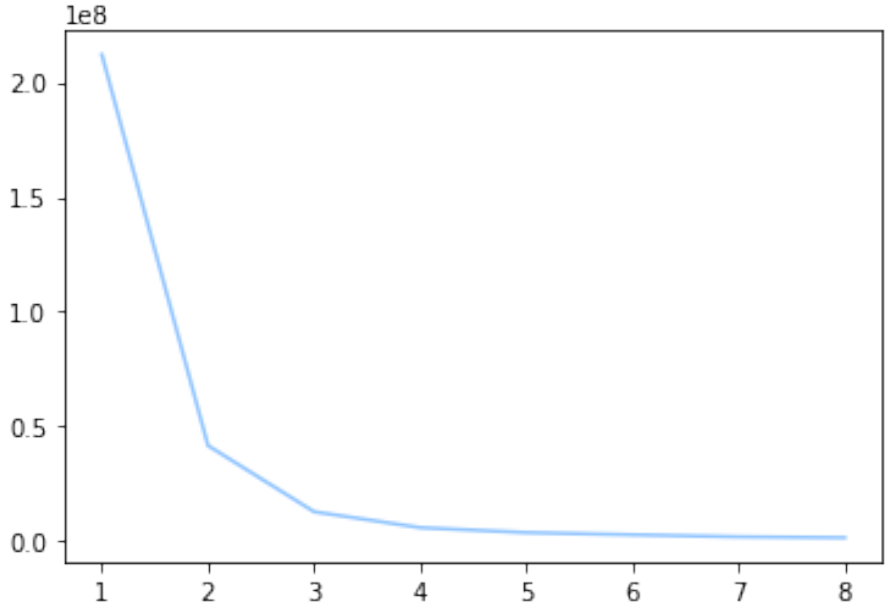


Figure 23: Elbow Method



4 Supervised Learning

4.1 Times Series Analysis and Forecasting

This section will approach time series as a supervised learning problem. The aim is to analyse Table1 in order to understand if it possible to make predictions about the number of calls received by the 1522 day by day in 2020 (Figure 24). The main issue when dealing with time series analysis is the serial autocorrelation of the regression residuals (Verbeek, 2017). This means that the value of the calls today might depend on the number of the calls in the past. A regression or analysis that does not solve for autocorrelation will inevitably produce biased estimates. In contrast to standard linear regression, time series analysis allows to give data an ordering and in that ordering itself relies information. Tab1 provides mono-dimensional time series data. In order to make the data applicable to a supervised learning methods, twelve lagged values of the time variant data points were generated, following the sliding window methods. Then windows statistics have be produced by calculating the mean of the lagged values. Two stationarity tests have been conducted to check whether the mean and the variance, among other statistics, of the data change over time. The stationary tests conducted both on the raw and the 12-lagged data show that the data is not stationary(Figure 25). After de-trending the data, the test statistics rejects the null hypothesis that the data is stationary at the 99% confidence interval (Figure 26).

The Simple Exponential Smoothing method, which computes the predicted values of the data using weighted averages, has been implemented (Figure 26). The training has conducted with a 80-20 split between test and train data. The Root Mean Squared Error of our forecasts with smoothing level of 0.8 (Closer to observed data) is 52.88, while the Root Mean Squared Error of our forecasts with auto

optimization is 53.41. The SES does not predict data optimally.

Then, the Holtz Linear Trend method, which adopts the same functionalities as the SES but using two smoothing parameters, which allows to capture trends, has been applied to the data. This method is optimal for data series that presents a trend, and therefore might help to better handle the increasing number of calls that have been observed from Q2 2020. The training has conducted with a 80-20 split between test and train data. The Root Mean Squared Error of the Holts Linear trend is 16.45, while the Root Mean Squared Error of the Exponential trend 19.28. As Figure 28 also shows, the HLT performs better than the SES in fitting the data. However, the trends it captures are taken forward in time too dramatically, in a way that most likely will not reflect reality.

Finally, the Holt-Winters' Seasonal Method has been implemented. This allows to capture seasonal variations as well as trends, and includes a parameter for seasonal smoothing. The type of seasonality chosen is of the multiplicative type: this allows to have seasonal variations that depends directly on the general changes in the data. The training has conducted with a 80-20 split between test and train data. The Root Mean Squared Error of the additive trend and multiplicative seasonality is 46.4. As Figure 29 shows, the HTWS is the only method able to forecast an increase in daily calls at Q3 2020, and it is the one the best fits the observed data, and predict time variation.

Being able to predict the number of daily calls to the 1522 is fundamental in order to make better administrative decisions internal to the call centers that manage the calls: it could serve as a way to be prepared to what appears an increasing number of victims asking for help and thus to be better equipped to serve the users of the service.

Figure 24: Daily calls to the 1522 helpline

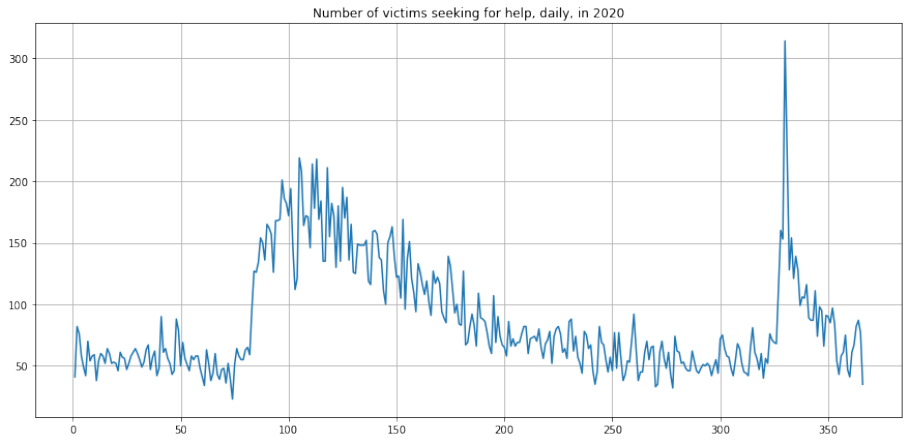


Figure 25: Stationarity test on raw data and 12-lagged data

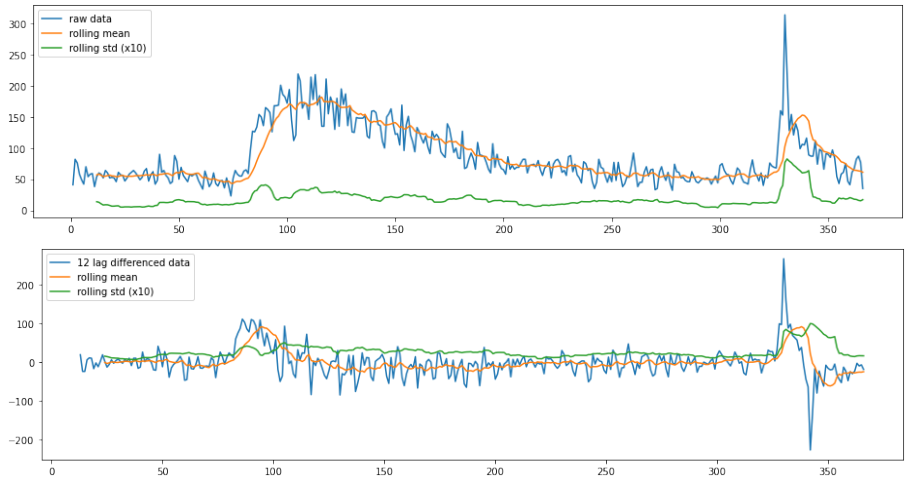


Figure 26: Stationarity test on de-trended raw data and 12-lagged data

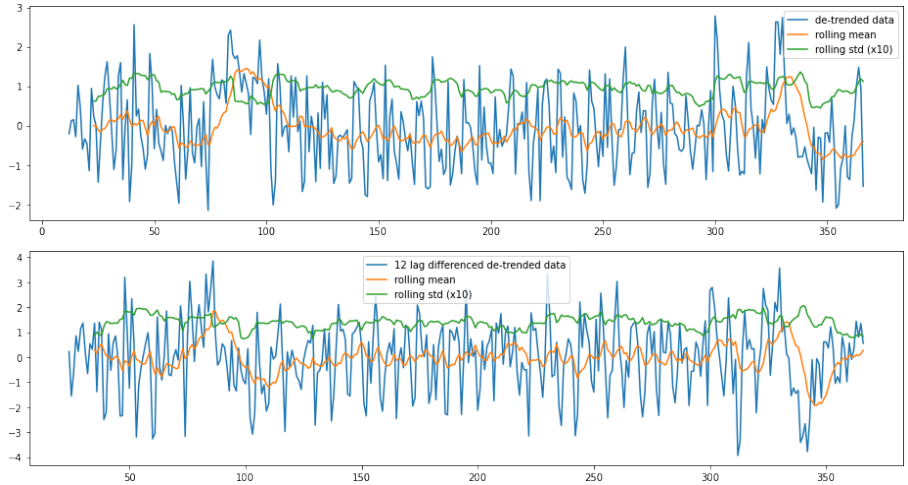


Figure 27: Simple Esponential Smoothing

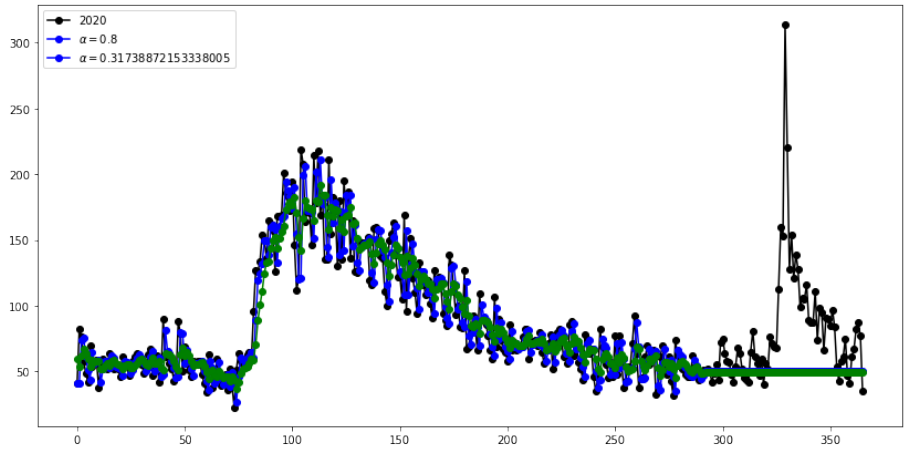


Figure 28: Holtz's Linear Trend

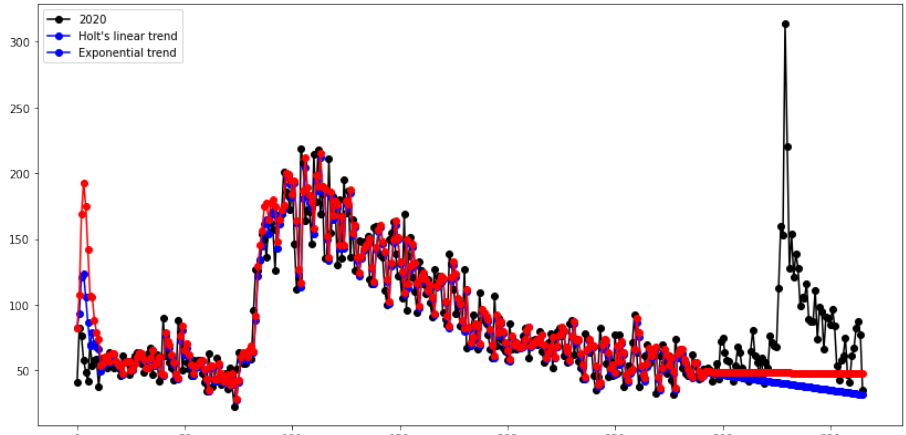
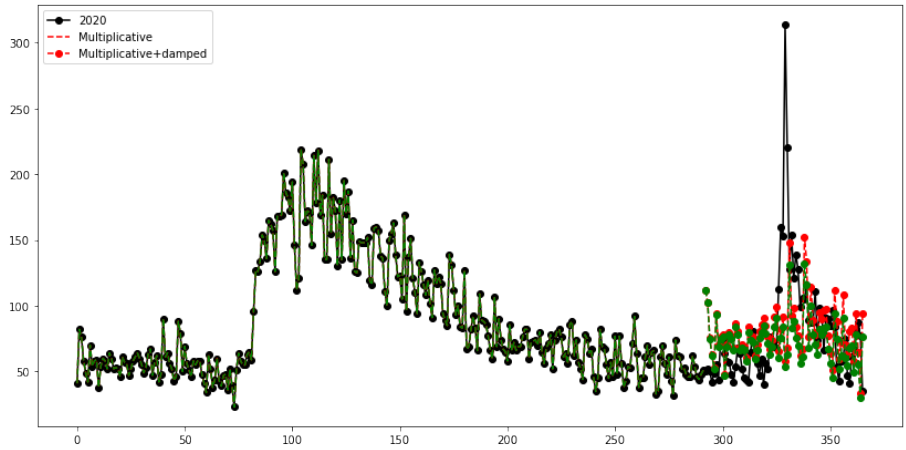


Figure 29: Holtz-Winter's Seasonal Method



5 Discussion

The results of this analysis, albeit its limitations, show that during the lockdown imposed in Italy in Q2 2020 there has been a significant increase in the number of calls to the anti-violence helpine 1522 for all the reasons of call with respect to Q2 2019 (an increase by 73% only from March to April 2020). Furthermore, it is has been observed a dramatic surge in intimate partner violence, as hypothesised. The fact that the majority of the violent acts has been perpetrated in a domestic environment is accompanied by the evidence that the cases where the abuser was a spouse increased by 158% in Q2 2020 in comparison to Q2 2019. 64% of the abuses were witnessed by children. The increase in number of calls during and after Q2 2020 might be the positive sign of the spreading of information related to the service, which, in turn, with time, could help to counteract the instances of under-reporting of domestic violence, or the negative sign of the emergence of new victims during the time period of interest. The increase in domestic abuse might be due to the increase in stress levels, economic and health uncertainty experienced during the period of interest. However, there are structural social motives that go beyond the scope of this paper and of data analysis that put women in the unfortunate position of victims. Notwithstanding the impossibility of discerning with certainty the reasons for the surge in calls from the reasons to the surge in violence, this report offers hints that the higher number of reported cases of gender based violence against women experience in Q2 2020 has been an unfortunate side effect of and can be seen as highly correlated with the restrictive policies on movement implemented to contrast the spreading of COVID-19. Given the results from the Supervised Analysis in Section 3, it is likely that the number of calls are predictable. An effort to implement more sophisticated forecasting methodologies might be useful to inform the decision and policy-making of the institutional entities that are in charge of the 1522.

5.1 Reflection

If I could turn back time I would not choos this dataset. The Excel file that was provided by the Italian National Institute of Statistics has been complex to work with, and I found myself having to manually clear data and concatenate tables to performs the task the assignment asked. The data points were collected in random formats and in a way that would have been more suited for a presentation than it was for data analysis. Time series have been interesting to investigate. I am relived I managed to implement, even with limited results, more sophisticated analysis methods using such a dataset. Next time I will prioritise usability in my choice of the dataset, other than intellectual interest in the subject. My initial aim - when looking for a dataset - was to draw a profile of the abuser and the victims, and understand what are the variables that influence the most the likelihood of being an abuser or victim, and to investigate the effects of the lookdown on the number of abuses reported. Manually creating tables from the tables in the daset and performing more specific analysis helped navigating the assignment. I feel I learnt more about data handling and about the importance of having a well structure data source. I

understood my limitations in coding and learnt a lot by mistake.

References

Bertolucci, F (2021). Domestic violence against women escalating in Italy. 3 Jun 2021. Available at: https://independentaustralia.net/politics/politics-display/domestic-violence-against-women-escalating-in-italy,15150

ISTAT. IL NUMERO VERDE 1522 DURANTE LA PANDEMIA (DATI TRIMESTRALI AL II TRIMESTRE 2021) https://www.istat.it/it/archivio/262039

McCrary, Justin and Sanga, Sarath, The Impact of the Coronavirus Lockdown on Domestic Violence (June 2, 2020). Available at SSRN: https://ssrn.com/abstract=3612491 or http://dx.doi.org/10.2139/ssrn.3612491

WHO. Devastatingly pervasive: 1 in 3 women globally experience violence. Press release, 9 Mar 2021. https://www.who.int/news/item/09-03-2021-devastatingly-pervasive-1-in-3-women-globally-experience-violence

WOMENSAID. Survivors say domestic abuse is escalating under lockdown. 28 Apr 2020. Available at: https://www.womensaid.org.uk/survivors-say-domestic-abuse-is-escalating-under-lockdown/