Prediction of the number of ambulance calls for 2019 in the different neighbourhoods of The Hague based on a multiple regression model with socio-economic vulnerability indicators

Final Assignment - EPA 1316 Introduction to Urban Data Science

By:

Raquel Romão (5629608)

Sara Costa (5630096)

Francisco Cortez (5627427)

José Fonseca (5394406)

Professor: Trivik Verma

Abstract

The present work was motivated by a lack of explanatory capacity in literature of ambulance calls patterns in the region of The Hague, The Netherlands. Hence, thirteen of socio-economic indicators (see Appendix 1) were selected to measure the relationship between ambulance requests across the city neighbourhoods.

From Exploratory Data Analysis, it was concluded that four of these indicators have higher correlation (and so, more impact) with the number of ambulance calls in each neighbourhood: all offences, number of people that are 65 years or older, number of single-person households and people with disability benefit. Additionally, from the Exploratory Spatial Data Analysis, it was found that the neighbourhoods showed a relevant degree of clustering amongst each other that there was a big HH cluster in the centre of The Hague highlighting the importance of this particular area in the city when it comes to ambulance services management.

Further analysis by simple linear model for the predictor single person household indicator showed a R^2 equal to 0.542 which reveals a good explanatory capacity.

Finally, a multiple regression model was developed to predict the 2019 data for the number of ambulance calls based on four relevant indicators identified previously for which the metrics found were a value of R-squared = 0.7 and an RMSE = 153.99, suggesting that neighbourhoods with higher numbers of reported crimes, single person households and older than 65 years old or with some kind of disability will most likely be related with higher calling patterns.

Content

1. Introduction

The rise of Data Science, and related fields of Big Data, Machine Learning, and Deep Learning, have completely transformed the landscape of almost every society. With ever-increasing flow and recording of data, the panoply of potential applications is enormous ranging from the health care sector with Sanchez-Pinto et al., (2018) pointing to its relevance in the critical care setting and the sports sector where, for example, the usage of spatial data has been used to register the exact locations of every player and the ball at regularly-spaced time points to analyse dynamics in basketball and football aiming to improve both players' individual and teams' collective performances (Macdonald, 2020), just to name two. Opposite to Data Analytics where the objective is to get insights into what has already happened, Data Science is a way of understanding big data by creating models that can predict or analyse future outcomes. Of course, it is impossible to perfectly model all the dynamic complexities happening in cities nonetheless, as statistician George E.P. Box famously put it: "all models are wrong, but some are useful." Still, Data Science and the models developed can work as a bedrock to provide decision-makers with the support needed to make informed recommendations about key areas of uncertainty.

Naturally, considering its prominence, these tools have also been regarded as paramount in governance and planning of smart cities, given the ubiquitous digital technology embedded in its physical structure, with Kandt & Batty (2021), suggesting that big data analytics promises benefits in terms of real-time prediction, adaptation, higher energy efficiency, higher quality of life and greater ease of movement. Bibri (2021) goes even further arguing that "data-driven smart sustainable urbanism is shaped by socio-cultural and politico-institutional structures."

Following these considerations, it is clear that Data Science and Analysis can be used to predict the sizing and to allocate important services to the society, like emergency services. For instance, the relation between the number of emergency calls from a specific neighbourhood and the socio-economic factors of that neighbourhood can be a way to predict the number of resources (ambulances, police or fire stations) to allocate to that neighbourhood or region.

Although probably useful, this relation has not been used so far in The Netherlands to determine the number of ambulances and their right places. Noting the lack of studies about this matter, the following report aims to measure the relationship between ambulance calls across the neighbourhoods of The Hague, The Netherlands, with indicators of socio-economic vulnerability status of people and households in order to answer the following research question:

- Is it possible to use socio-economic vulnerability indicators to predict the number of ambulance calls in the region of The Hague?

This study contemplates a dataset based on three distinct data sources:

- A (not open-sourced) emergency calls dataset of The Netherlands collected from January 2017 to September 2020 provided by Dr Trivik Verma and obtained from 112-Nederland.
- Two online open data sources of social and economic indicators of The Netherlands - Den Haag Cijfers and CBS.

After an analytical examination of several socio-economic factors, the final model is based on a multi-regression of five indicators to predict the number of ambulance calls in the aforementioned regions. Actually, the R-squared of this model is approximately 0.7, which means that these indicators have a real influence on the number of ambulance calls in a neighbourhood. Also, the RMSE of the final model is 153.99, which, compared to the range of values for the ambulance calls (from 122 to 1260) is an acceptable value.

Consequently, it is possible to state that the prediction of the number of future ambulance calls in each neighbourhood based on socio-economic factors is possible, and with that is possible to predict the number of ambulances and the right places for their stations. The model used is quite reliable for the case, although the usage of more indicators or training the data over a longer period could help to increase the reliability of the model even more.

2. Related work

Socio-economic vulnerability is a factor that influences population's health worldwide and, thus, the way health services work and how they are organized. Since this report aims to study the influence of socio-economic vulnerability (through several factors described below) in the number of calls for ambulance services in the region of The Hague, this section of the report explores and explains the research conducted about this topic.

2.1. Ambulance calls in The Hague, The Netherlands

Generally, determining the correct location and number of resources and facilities of public services is always an important and sensitive challenge to any decision-maker due to the trade-off between limited resources and the necessity to properly serve the populations. Fundamentally, the process of locating and sizing the resources of emergency services, like ambulances or fire stations, must be treated uniquely once "emergency responders must reach urgent cases within mandatory timeframes" (Yu et al., 2020) but, at the same time, decision-makers must keep worrying about defining "the minimum-cost spatial arrangement of service facilities that adequately serves the entire user region". (Toregas et al., 1970)

Particularly in ambulance allocation, the historical record of emergency calls requesting for ambulance service could be helpful. According to the Dutch National Institute for Health and Environment (in Dutch, RIVM) - entity responsible for the allocation of ambulances in the Dutch territory -, the demand for ambulances in this region has been increasing in the past years. Despite the algorithm of RIVM being based on mathematical models which consider as the most important measure of the response time of a paramedic to reach every citizen after the call, as described by several authors like Swoveland et al. (1973) or Uyeno et al. (1984), and analysing the demand of ambulances for different times of the day, and different days in the week, as explained by Cantwell et al. (2012), it does not include socio-economic indicators which could assess the frequency of need of this type of service in each region or neighbourhood and, with this assessment, help to better allocate the ambulances.

2.2. Socio-economic indicators of vulnerability

Even though the relation of socio-economic indicators with ambulance calls or with the allocation of ambulance services in a region is not a widely researched topic, there are already some that analyse this topic in more depth. For instance, Aldrich et al. (1971) explored a similar question in the region of Los Angeles, USA. From that research, it was possible to identify many social and geographical patterns related to the volume of emergency calls. The authors concluded, for instance, that on average white people call the emergency services fewer times than non-white people; and married people or single women households generate more emergency calls than single men households. Further, they conclude that while regions with more housing density had more calls per capita than other regions, it was also notable that commercial areas had more emergency calls per capita than residential or industrial areas.

Moreover, research funded by the English National Institute for Health Research Health Protection Research Unit from Elliot et al. (2016) found that the quality and the composition of the air in different regions and different moments of the timeline have a direct impact on the number of emergency calls and the effect is even clearer when the same air conditions linger, once they are more evident when people stay in prolonged contact with determined conditions, for instance, the effects of pulmonary infections. Also, the impact of the air condition on emergency calls is different for different age groups.

Additionally, Bray et al. (2015) were capable of determining that the awareness of people for the possible effects of quick and effective intervention in the treatment of a stroke also influences the number of calls to ambulances. This research, done across several regions of Australia, concluded that

where the investment in the awareness campaigns was higher (which represents better-educated people), the number of emergency calls to ambulances had a higher increase.

2.3. Problem Statement

From the aforementioned examples, it is possible to understand that socio-economic indicators impact the number of calls to ambulances. However, that impact has not been proven yet in the region of The Hague, The Netherlands.

Indeed, this research focuses on socio-economic factors that express the vulnerability to disease or loneliness of populations of each neighbourhood of The Hague (description of each indicator in Appendix 1) to predict their influence on the volume of calls per neighbourhood. If the relation between the chosen socio-economic indicators and the number of calls to ambulances is considered relevant, and if the constructed model is reliable and has enough explanatory capacity, decision-makers could be capable of predicting the expected number of ambulance calls for the coming years using these indicators. Based on this, they can predict the best location of ambulance stations and the correct number of ambulances to allocate to each of these stations, improving the quality and equity of the distribution of this essential service to The Hague's people.

3. Exploratory Data Analysis

3.1. Data retrieval

The analysis conducted in the present report was made possible and driven by information (at the neighbourhood level) gathered from the following sources of data:

1. **Emergency calls data**: this source of data contains ambulance, firefighter, police, and coastguard emergency calls collected from January 2017 to September 2020 (calls made to these services during the period of lockdown and release cycles in the pandemic are included) in The Netherlands which we subsetted so it would only include ambulance calls in The Hague area for two years (2018 and 2019). The initial data only provided latitude and longitude indicators, which we used to derive the locations (points) of the neighbourhoods. A description of the indicators can be found in Appendix 1.

2. **The Hague in Cijfers data**: this accounts for one out of the two sources of data used to obtain the social and economic indicators, whose relationship will be measured with ambulance calls across the city. A description of the indicators can be found in Appendix 1.

3. **CBS data**: this accounts for one out of the two sources of data used to obtain the social and economic indicators, whose relationship will be measured with ambulance calls across the city. A description of the indicators can be found in Appendix 1.

4. **Shapefiles of The Hague**: the shapefile relative to The Hague neighbourhoods' locations (from class) was used to create geospatial representations of the data as part of exploratory data analysis.

3.2. Data scraping and cleaning

In order to obtain coherent dataframes from the previous data sources, the following modifications in the initial datasets were performed:

- There was a mismatch between the denomination of twelve (12) neighbourhoods in the sources of The Hague in Cijfers data and CBS data, demonstrated in Appendix 2. By analysing the table, it is visible that the names only differ in minor aspects but are, nonetheless, considered different in computational terms. A noteworthy case is the neighbourhood of 'Kraayenstein & Vroondaal' (as mentioned in The Hague in Cijfers data) because its denomination changed in the CBS data source from the year 2018 to 2019. Hence, in this data source, the same neighbourhood is referred to as 'Kraayenstein' (information relative to 2018) and as 'Kraayenstein en Vroondaal' (information relative to 2019). To homogenize the terms, we decided to alter the former - 'Kraayenstein' - to 'Kraayenstein en Vroondaal' and keep the CBS data denominations by changing them accordingly in The Hague in Cijfers data source.
- There was a mismatch between the denomination of one (1) neighbourhood in the Emergency calls data source and the others (already altered with the previous modification). The neighbourhood in question was referred to in the Emergency calls data source as 'Kraayenstein' which we changed to 'Kraayenstein en Vroondaal'.

Concerning the cleaning process, we came across with:

- Six (6) missing values in the indicator "Distance to GP", for three (3) neighbourhoods in both years (2018 and 2019), that were substituted by the average value of the distances of all the other (non-missing) neighbourhoods.

- Three (3) neighbourhoods for which the number of inhabitants was zero (0) - ‘Oostduinen', 'Tedingerbuurt' and 'Vliegeniersbuurt' - which we choose to not consider in the model by eliminating them from the dataframe.

The overall process resulted in three different dataframes which we used to carry on our analysis. The first included information for every socio-economic indicator, geographical information and number of ambulance calls concerning the area of The Hague for 2018 and 2019. The other two are merely variations of the former in which the information was divided so as to only include data from 2018 and 2019, exclusively.

3.3. Data Exploration and Scavenge for Relations

In order to find variables that would be good predictors for the number of ambulance calls, a regression plot was done between the number of calls per neighbourhood and each one of the variables thought to be eventual good predictors. The scores for each one of the cases were obtained, found in table 1.

Table 1: Comparison between the different correlations obtained between each variable and the number of calls per neighbourhood, after doing a linear regression.

	Private cars per address	**All offences**	**% Dutch Natives**	**Inhabitants 65 or older**	**Single person households**	**Households with children**
Score	0.009	0.430	0.083	0.430	0.542	0.360

	People with disability benefit	**Distance to GP**	**Low-income households**	**Wealth private households**	**Number of inhabitants**
Score	0.477	0.104	-0.009	0.015	0.509

Taking the obtained regression plots and score values into account, the variables were divided in two groups: good predictors and bad predictors. The variables with a bigger score were selected as eventual good predictors:

- All offences;
- Number of people that are 65 years or older;
- Single person households number;
- People with disability benefit;
- Number of inhabitants;
- Households with children.

The regression plots obtained for each one of these variables are shown in figure 1, evidencing the correlation between the 6 variables and the number of calls.

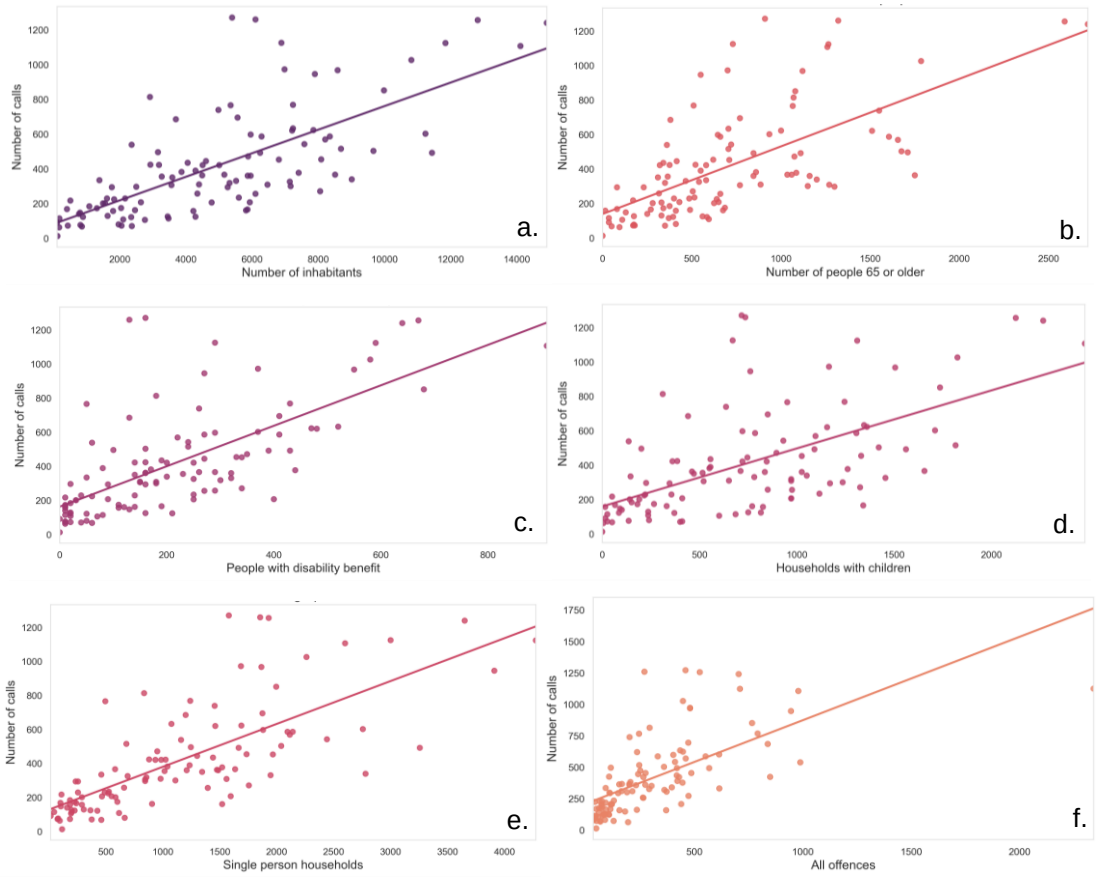


Figure 1: Regression plots for Number of calls vs. each one of the chosen variables as good predictors: a) Number of inhabitants , b) Number of people 65 or older, c) People with disability benefit, d) Households with children, e) Single person households and f) All offences(excluding outliers, see section 3.3.3).

As it would be to expect, it is shown that a bigger number of inhabitants per neighbourhood translates into a bigger number of emergency calls (figure 1, a).

Looking at variables that are directly related to a health vulnerability group and therefore with a higher probability to contribute to the number of calls of a neighbourhood, like the number of people that are 65 years or older, the number of people with disability benefits and the number of households with children, the results were also satisfactory, since the number of calls increases with a higher number for these variables (figure 1, b-d).

A higher number of single-person households also translated into a superior number of calls (figure 1, e). When in need, a person that lives alone was assumed to have a higher tendency of calling 112 instead of being helped by a familiar or cohabitant.

It was also verified that a neighbourhood with a higher number of offences, has a higher probability to have a bigger number of ambulance calls (figure 1, f).

Taking these variables into account, a further prediction was done in section 4.

The plots for the remaining variables, classified as bad predictors, can be found in figure 2.

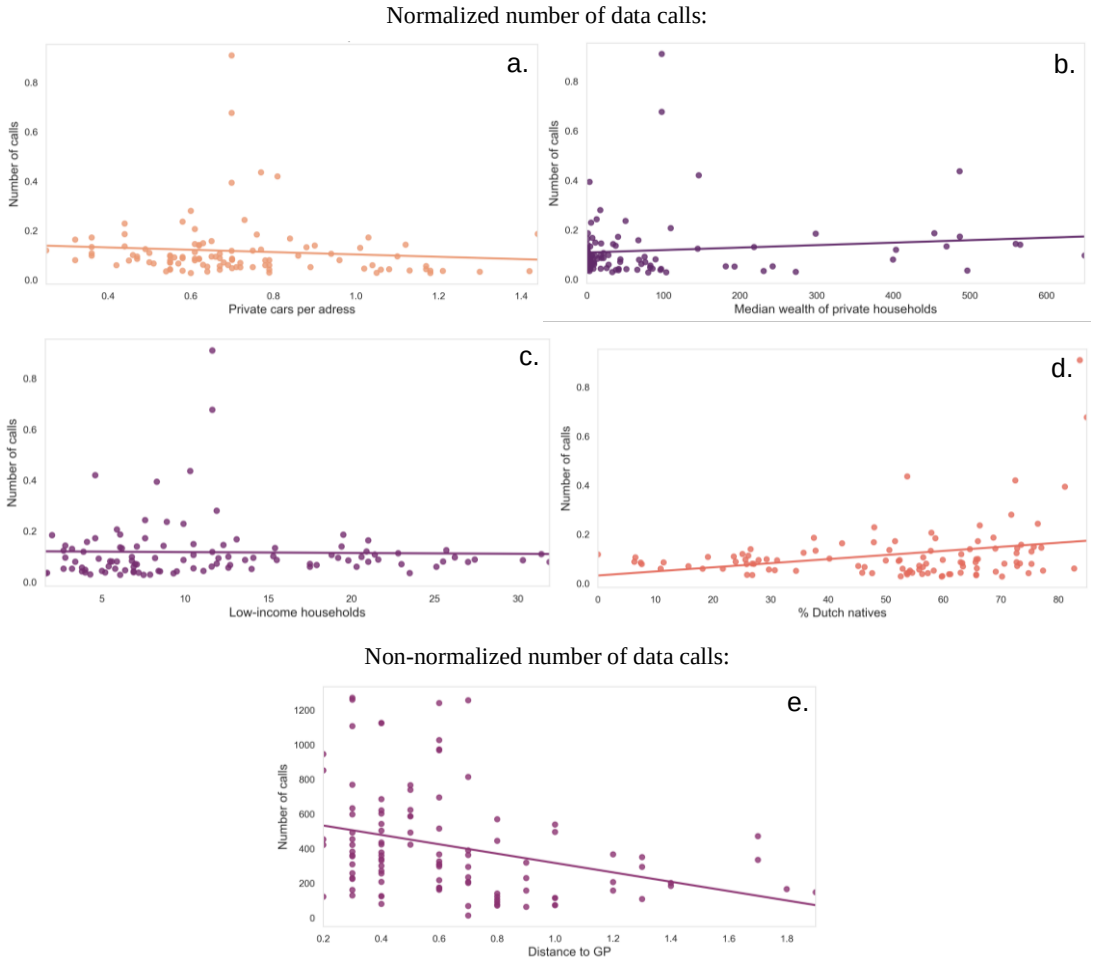


Figure 2: Regression plots for Number of calls vs. each one of the bad predictors: a) Private cars per address, b) Median wealth of private households, c) Low-income households, d) Percentage of Dutch natives and e) Distance do general practitioner (excluding outliers, see section 3.3.3).

For the variables shown above it was expected better results. For instance, it was expected that someone with a car would be more likely to head for the hospital instead of calling for an ambulance and therefore result in an indirect correlation between the number of calls and the number of private

cars per address. Indeed, the results in figure 2-a show a slight decrease in the number of calls with the increase of possession of cars, however not with the desired correlation.

Moreover, it was hypothesized that a smaller median wealth of private households and a higher percentage of low-income households would lead to a higher number of emergency calls, yet those correlations were impossible to unravel (figure 2, b-c).

The number of calls was also compared with the percentage of Dutch natives per neighbourhood. It was expected that foreigners would call less, taking into account one possible existent language barrier or an absence of the needed information to know who to reach for care. It was in fact what it was possible to observe (figure 2, d), revealing a positive correlation, although a small one.

Finally, it was conjectured that a higher distance to a general practitioner (GP) would reveal a higher number of ambulance calls since, with a GP nearby, one can more likely seek care there rather than calling an ambulance. Nonetheless, was obtained the opposite (figure 2, e).

It was decided to show the results for these variables to notice that further analysis is needed to understand why these relations were obtained and what other factors could be contributing to it.

3.3.1. Outliers

After doing the first exploration of the data some outliers were detected, which were excluded and not used for further analysis. In the regression plot obtained for the number of calls and the number of inhabitants (figure 3), for example, it can be clearly detected three outliers - with a high number of calls.

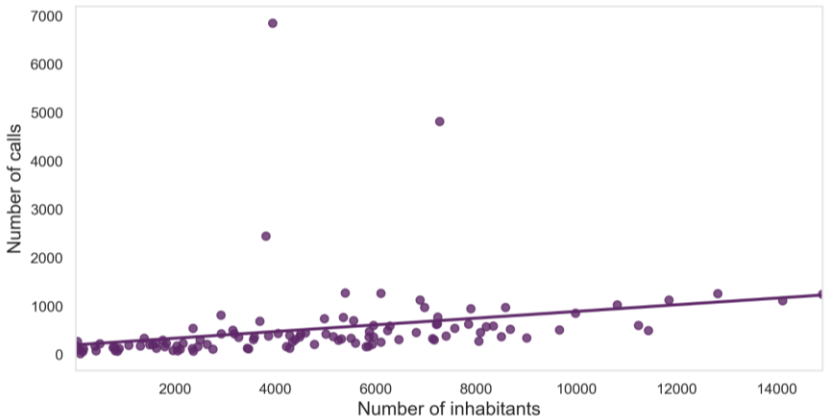


Figure 3: Regression plots for Number of calls vs. Number of inhabitants, with outliers.

In order to further analyse this occurrence, an analysis of the number of calls variable was done recurring to an histogram and a boxplot, shown in figure 4.

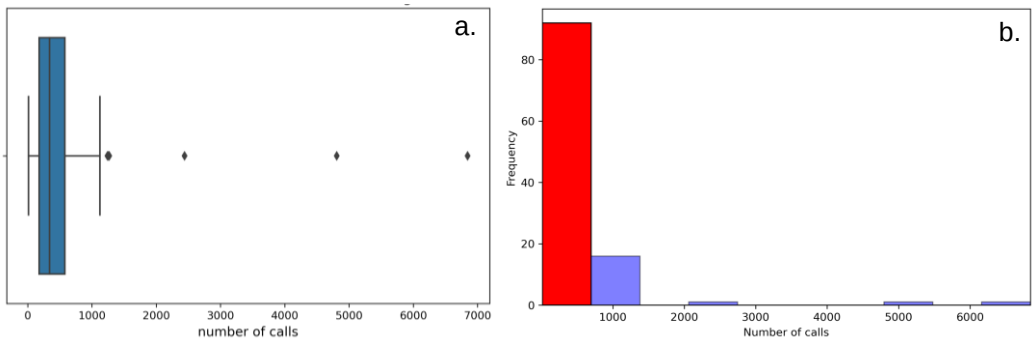


Figure 4: Boxplot (a) and histogram (b) for the number of calls.

Evident outliers can be identified for a number of calls higher than 2000. Therefore, it was chosen to exclude neighbourhoods with a number of calls higher than this value. Those neighbourhoods were:

- 'Waalsdorp';
- 'Kortenbos';
- 'Transvaalkwartier-Noord'.

The *regplots* for the remaining variables can be found in appendix 3, for which it can be noticed that once again these neighbourhoods correspond to outliers.

Moreover, the analysis done for the number of calls was equally performed for the other variables (in appendix 4). However, no important outliers were found and, other than the already excluded neighbourhoods, it was decided to keep the remaining ones.

It is to mention that further analysis should be performed in order to understand why there was obtained such a difference in the number of calls for these three neighbourhoods.

3.4. Exploratory Spatial Data Analysis

As part of the exploratory data analysis (EDA), we conducted exploratory spatial data analysis (ESDA). From the literature, it is known that the range of ESDA methods is very wide and spans from less sophisticated approaches like choropleths and general table querying to more advanced and robust

methodologies that include statistical inference and explicit recognition of the geographical dimension of the data. Regarding the latter method, we know that ESDA techniques are usually divided into two main groups: tools to analyze global, and local spatial autocorrelation. Hence, we will start by the former and then perform a more detailed global and local analysis with the latter method.

3.4.1. Choropleth Analysis

In this part of the analysis, choropleths were plotted for each indicator to provide a geographical representation and/or visual perception of how the variables behave in the different neighbourhoods. To construct them, the *"quantiles"* approach was followed, to avoid the potential problem of sparse classes, which means that the numbers of values in each class are roughly equal (as opposed to the "*equal_interval"* approach, for example, where the range of attribute values is divided into equal-sized subranges which can potentially result in one or more classes being sparse - low representativity - despite the simplicity and ease of interpretation). We also took into consideration the colour palette chosen so as to be inclusive of colourblind people, in which yellow represents the highest values and dark purple represents the lowest. Finally, the neighbourhoods whose number of inhabitants was zero (0) - 'Oostduinen', 'Tedingerbuurt' and 'Vliegeniersbuurt' -, were coloured grey.

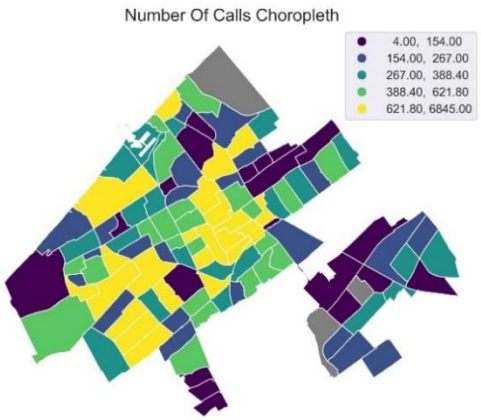


Figure 5: Number of Calls Choropleth

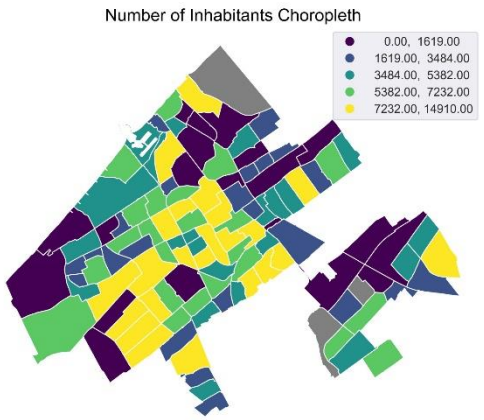


Figure 6: Number of inhabitants Choropleth

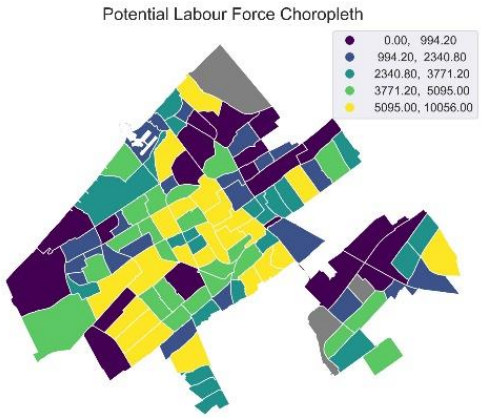


Figure 7: Potential Labour Force Choropleth

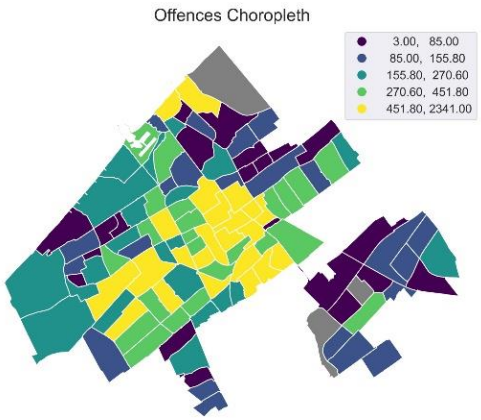


Figure 8: Offences Choropleth

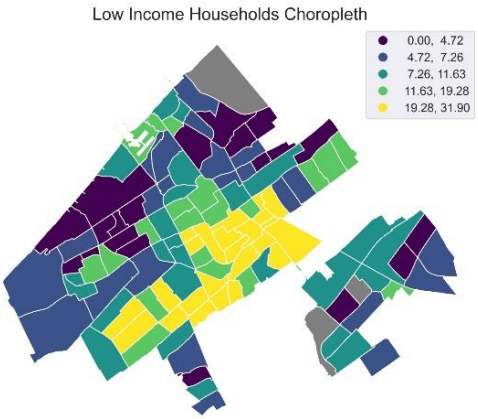


Figure 9: Low Income Households Choropleth

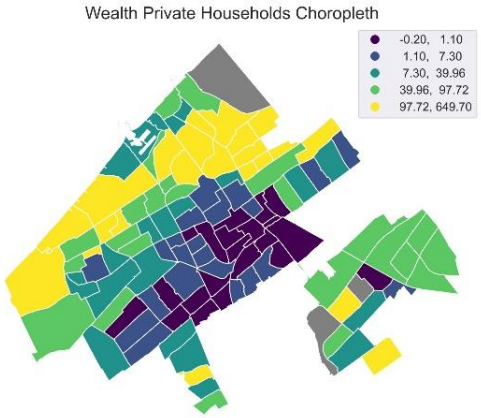


Figure 10: Wealth Private households Choropleth

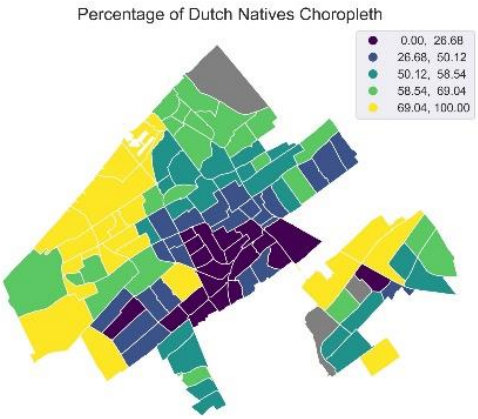


Figure 11: Percentage of Dutch Natives Choropleth

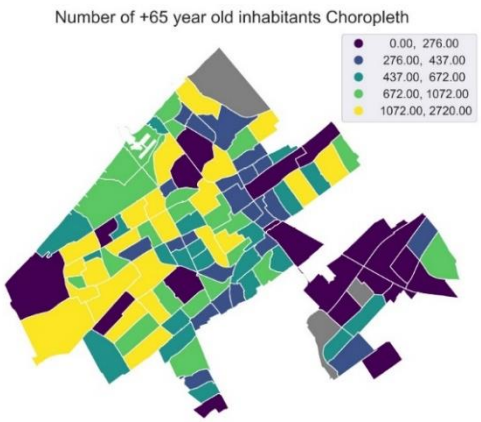


Figure 12: Number of 65+ year old inhabitants Choropleth

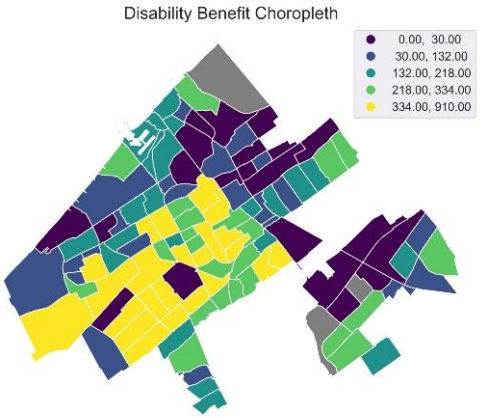


Figure 13: Disability Benefit Choropleth

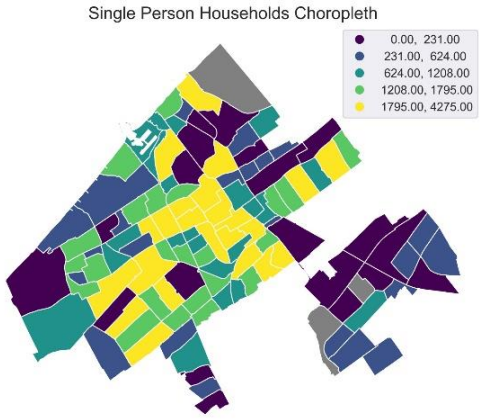


Figure 14: Single Person Households Choropleth

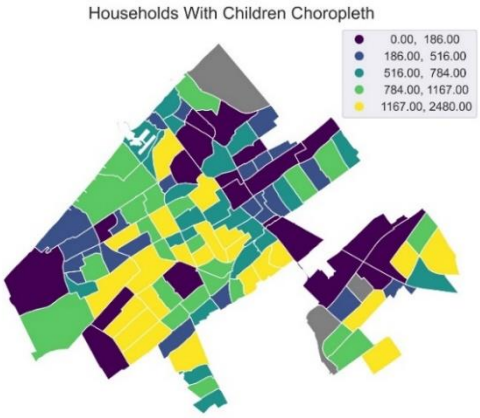


Figure 15: Households With Children Choropleth

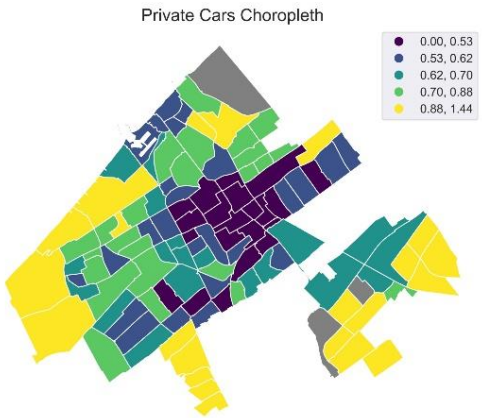


Figure 16: Private Cars Choropleth

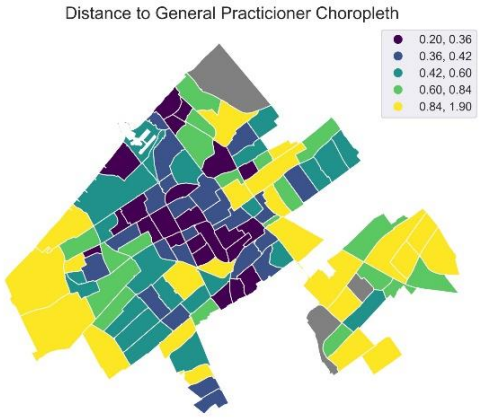


Figure 17: Distance to General Practitioner Choropleth

The Number of Calls Choropleth depicted in Figure 5, shows that the measurement does not follow any trend in terms of neighbourhood locations. In fact, and even though the central area of The Hague does show a concentration of neighbourhoods that are regarded as the top 40% neighbourhoods for ambulance calls, there are some coastal areas in the north-western and others in the southern part of the city that also indicate high ambulance requesting patterns. Still noteworthy is the eastern part of the city (island-shape) which is majorly composed of neighbourhoods that made 267 calls or less during 2018.

Figures 6 and 7 show the Number of Inhabitants and Potential Labour Force Choropleths, respectively. When compared, these choropleths differ on the classification of an extremely low number of neighbourhoods (an expected relation between the indicators) and are significantly like the previous spatial distribution - Number of Calls Choropleth. As in the latter, the measurements do not follow any substantial geospatial tendency other than the broader concentration of the indicator around central areas of The Hague. This is in line with the expected because, naturally, the number of calls requesting ambulance services is foreseen to increase with the population and of the active population. Also with a considerably similar geospatial distribution, there is the Offences Choropleth (Figure 8). The figure shows a high concentration of neighbourhoods with most offences in the centre and southern part of The Hague, thus identifying northern and eastern areas as "safer". Hence, based on these studies, it is reasonable to conclude that the amount of emergency calls is highly dependent on the population of each neighbourhood. The Offences Choropleth also shows a high predictability capacity when estimating the number of calls which can be justified by the intrinsic relation amongst the two: offences can result in situations where medical assistance is required.

Exhibiting a more distinctive trend in values, there are the Low-Income Households and Wealth Private Households Choropleths in Figures 9 and 10, respectively. The spatial distribution for these measurements goes hand in hand with the first, showing a higher concentration of low-income household neighbourhoods in the central-eastern and central-southern part of the city while the second coherently identifies the northern and north-western areas with higher private household wealth. Interestingly, these latter areas are the same areas identified, in the Percentage of Dutch Natives Choropleth (Figure 11), as having higher percentages of Dutch natives and roughly the same "safer" areas mentioned previously.

Concerning the Number of +65 year old inhabitants Choropleth (Figure 12), it is clear that the eastern part of the city (island-shape) is majorly composed of neighbourhoods with a low number of people belonging to this age group (only excluding three neighbourhoods where the number is 437 or higher). It is more difficult to conclude anything concrete regarding the geospatial arrangement of the other areas other than there is a concentration of older than sixty-five (65) inhabitants in the southern part of the city. Also centring around this area (suggesting some overlapping amongst the choropleths

and thus some inexplicit relation), but in a more consolidated manner, there is the number of people that have access to disability benefits. Contrarily, the northern area (also including some costal sections), shows a tendency for neighbourhoods with a higher number of people with low access to disability benefits, as it's visible in the Disability Benefit Choropleth (Figure 13).

Figures 14 show the Single Person Households Choropleth which suggests a considerable concentration of this measurement around the city centre with a tendency to diminish the further we move from this point. This fact can partially be explained by the higher availability of jobs and other professional reasons that usually lead people to move closer to these areas. Roughly sustaining and corroborating this view is the geographical arrangement seen in the Households With Children Choropleth (Figure 15) depicting a higher number of households with children neighbourhoods in the periphery of the city.

Finally, we took into consideration the Private Cars and Distance to General Practitioner Choropleths in Figures 16 and 17, respectively, which show high correspondence amongst each other. Expectedly, both representations show that the number of cars and the distance increase with the distance to the city centre. Our initial assumption when considering the former measurement was that neighbourhoods with a higher value in the private cars measurement would, in turn, have fewer reported ambulance calls. Nonetheless, the lack of trends in the Number of Calls Choropleth makes it hard to have any concrete conclusion. Concerning the latter, the values obtained are logical since amenities are usually concentrated in highly populated areas – such as the city centre.

3.4.2. Moran Plot and Moran's I statistic

In order to perform a more detailed global analysis, we studied the degree of clustering in the dataset to check if the values follow a particular pattern in their geographical distribution - understand if similar values are closer to other similar values or not. To do it, we used two global spatial autocorrelation tools: the Moran Plot (visual tool) and the Moran's I statistic (numeric tool).

For the analysis, we decided to compute the weight matrix based on the distance criterion instead of the contiguity one. This allowed us to have more control over the number of neighbourhoods taken into account due to the fact that in the contiguity approach, the criterion is the existence of common boundaries (in the queen approach the areas only need to share a vertex wherein the rook approach the must share boundary lines). After row-standardizing the matrix we were able to compute the spatial lag for the ambulance emergency calls (variable of interest) . From here we also standardized this variable in order to be able to interpret values as above or below the mean, and their quantities in terms of standard deviations. For five (5) neighbourhoods, the results obtained are represented in Figure 18.

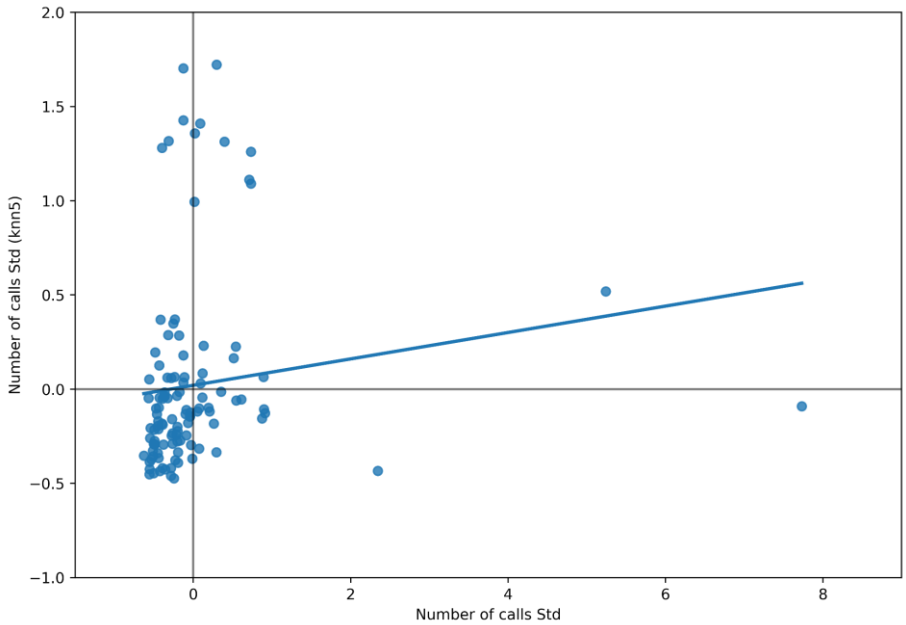


Figure 18 - Moran Plot for Ambulance Emergency Calls Indicator (with outliers)

Overall, it is visible that the majority of the points are concentrated around the origin despite some distinctive more distant points in the right part of the graph. The line, which represents the best linear fit to the scatter plot or, in other words, what is the best way to represent the relationship between the two variables, displays a positive relationship between the number of ambulance emergency calls and its spatial lag. Fundamentally, this is associated with a positive spatial autocorrelation, suggesting that similar values tend to be located close to each other. The slope of the line in the Moran Plot corresponds to the value of Moran's I, which is represented in Table 2, along with the p-value statistic.

Table 2 - Moran's I and p-value statistics for Ambulance Emergency Calls Indicator Moran Plot (with outliers)

Measurement	**Value**
Moran's I statistic	0.07452
p-value	0.062

These statistics corroborate our previous findings. The slope of the line in Figure 18 has a value of 0.07452 which suggests a positive but nonetheless weak relationship between the variables. Accordingly, the p-value statistic shows that the graph still displays more spatial pattern than we would expect if the values had been randomly allocated to a particular location despite not being an excellent value for the metric. In the end, it seems that the outliers previously identified are hampering these statistics and drive down the possibly higher positive spatial autocorrelation amongst other

neighbourhoods. Hence, to test this idea, we conducted a global analysis without them. The results are represented in Figure 19 and Table 3.

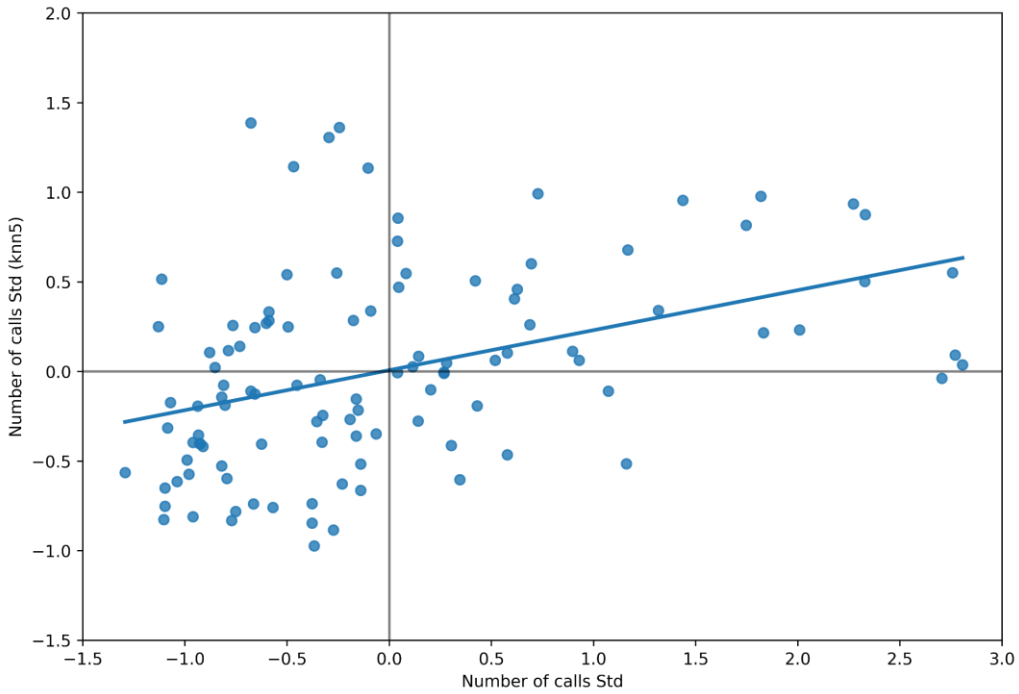


Figure 19 - Moran Plot for Ambulance Emergency Calls Indicator (without outliers)

The adjusted graph is clearly associated with the presence of *positive* spatial autocorrelation, as the former was but, in this case, it is also visible that the best linear fit to the scatter plot - straight line - is now more closely related with the overall tendency of the points thus suggesting a higher degree of clustering amongst the neighbourhoods over space. This finding is confirmed by the metrics in Table X for which we obtained a value of 0.22321 for the Moran's I statistic (whereas previously it was 0.07452) and the value of 0.001 for the p-value statistic (whereas previously it was 0.062).

Table 3 - Moran's I and p-value statistics for Ambulance Emergency Calls Indicator Moran Plot (without outliers)

Measurement	Value
Moran's I statistic	0.22321
p-value	0.001

3.4.3. Local Indicators of Spatial Association (LISAs)

As an enhancement and building upon the global analysis performed with the Moran Plot and Moran's I statistic - good tools to inform us about its degree of clustering -, we can perform local analysis. By doing so, we can identify areas within the map where specific values are located - inform

us about where the clusters are - and not only if values are clustered overall. To do that, we need to use a local/more granular measure of spatial autocorrelation. Thus, at the core of the Local Indicators of Spatial Association (LISAs) method is a classification of the observations in a dataset into four groups derived from the Moran Plot: high values surrounded by high values (HH), low values nearby other low values (LL), high values among low values (HL), and vice versa (LH), commonly known as "quadrants". Despite the effect the outliers identified previously shown to have on the results, we decided to conduct this analysis with them. The results are represented in Figure 20.

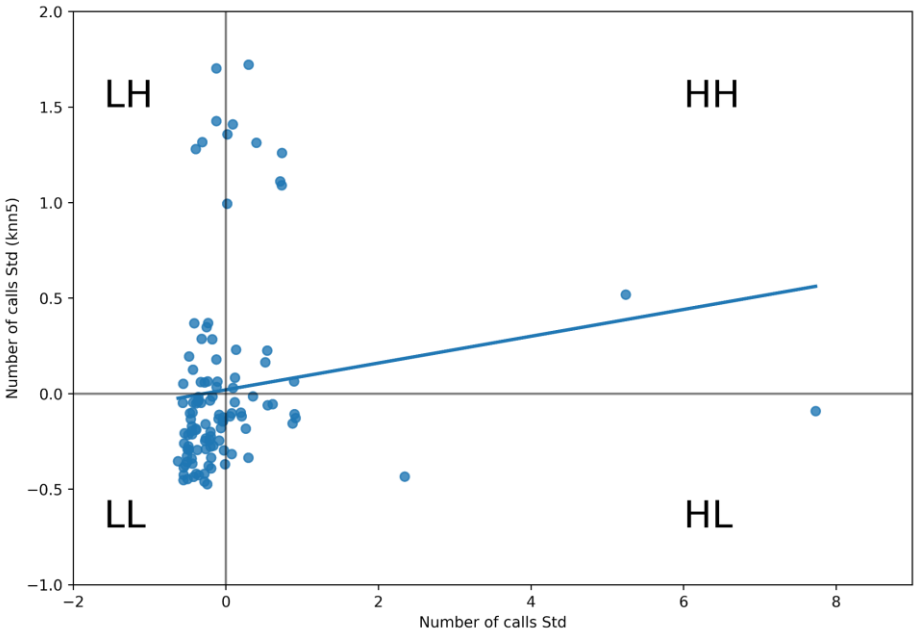


Figure 20 - LISAs for Ambulance Emergency Calls Indicator (with outliers)

The previous graph shows a rough classification of the neighbourhoods in terms of the clusters they belong to since it only differentiated the quadrants. Hence, and since we are only interested in neighbourhoods where the strength with which the values are concentrated is unusually high and find whether each of the locations is a *statistically significant* cluster of a given kind, we LISAs. Figure 21 shows the relevant observations for a significance level of 0.05.

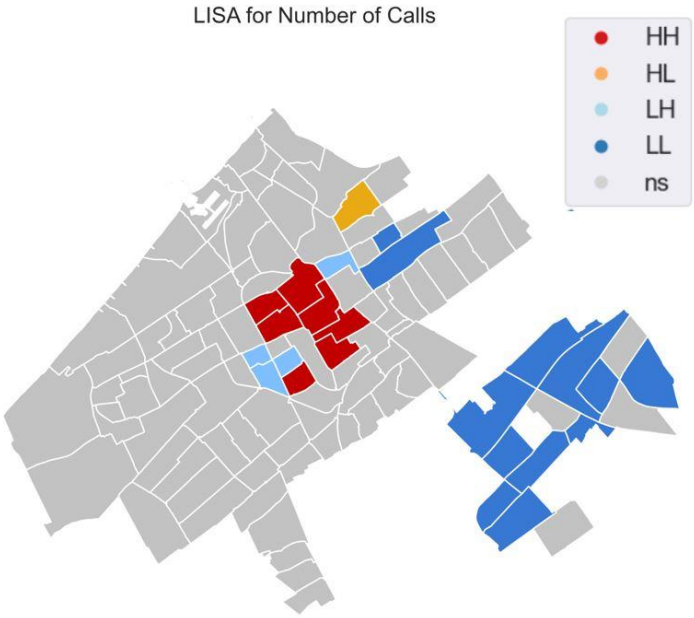


Figure 21 - LISA cluster Map for Number of Calls indicator

Analysing the results more closely, in red, around the city centre we find an unusual concentration of neighbourhoods that show high ambulance services requesting patterns. Closely located in its surroundings, in light blue, we find four LH spatial outliers suggesting that these neighbourhoods do not follow the ambulance calls trend (less calls were made) of their closer neighbourhoods (high ambulance requesters). In dark yellow, dissociated more from the city centre, we find the only HL outlier - calling patterns in this area were high when compared to calling patterns of its neighbourhoods. Finally, in dark blue, located in the eastern part of the city (island-shape), we find those neighbourhoods with an unusual concentration of neighbourhoods with a low number of ambulance calls requests - LL cluster.

These findings are in line with the results obtained in the choropleths analysis (big HH cluster in the centre of The Hague), suggesting the neighbourhoods located in the centre as a particular area of interest, emphasizing the relevance of this particular area in the city when it comes to ambulance services management.

3.5. EDA Limitations

The Exploratory Data Analysis performed in this research incurs a few assumptions required for its development, but that possibly will limit or change the final results of the analysis:

- At first, the indicator “Ambulance emergency calls” was normalized by the “Number of inhabitants” in order to apply to all the socio-economic indicators that are percentages of this number. This methodology may have two main implications in the study:
    - It may cause misinterpretations of the final results and conclusions once the research uses the indicator “Ambulance emergency calls” normalized to study the impact of some indicators and the same indicator not-normalized when studying the impact of other indicators. Thus, the explanation of the final results and conclusions requires special care.
    - By normalizing the ambulance emergency calls only by the number of inhabitants, we are not considering all other people that move through a neighbourhood during the day - like employed persons in that neighbourhood, tourists or drivers - and that may also need to call an ambulance.

- Since all the research values are based on the number of inhabitants of each neighbourhood, the neighbourhoods without inhabitants were considered irrelevant for the study and removed from the dataframe used. In that sense, this relates to the previous limitation pointed out because we are not considering all the people that may call an ambulance. If we were, there was no reason to exclude these neighbourhoods from the analysis.

- Further, the six missing values of the indicator “Distance to GP” - corresponding to distances between the same three neighbourhoods and the nearest GP in 2017 and 2018 - were filled in by the mean value of the indicator. A better approach could be, for instance, to use the mean value of this indicator considering only the nearest neighbours of each neighbourhood; in this case, the results would probably be more reliable and close to the real values.

4. Analysis

In this chapter, the indicators selected after the Exploratory Data Analysis were used to create models of prediction of the ambulance calls for the year of 2019 based on the data from 2018. Two supervised learning algorithms were implemented and will be briefly described.

4.1. Simple linear regression model

Firstly, a simple linear regression model was created for the predictor with the higher R^2 (see table 1) which was the single person household indicator, with an R^2 of 0.542. This is one of the socio-economic indicators of vulnerability. The direct correlation between this indicator and the number of

calls support our initial argument that the ambulance calls are higher where vulnerability indicators are also higher and with this, it is possible to predict the number of ambulance calls for other data frames.

This model can be written as a simple equation of the type:

$$y = \beta_0 + \beta_1 \times x,$$

in which the y is the variable to predict, the number of ambulance calls, and the x is the independent variable, the predictor, in this case, single person households.

For that, the dataframe from 2018 was split into two dataframes, a test one (with 20% of the data) and a train one (with the remaining 80%). This was done to verify if the model was good to predict the data from the same year, because if so, it would more likely be a good one to predict the data from 2019. After doing the regression with the training group, the values of R^2 for both the training and the testing groups were obtained and are shown in table 4. The values for the coefficients β_0 and β_1 were also obtained and can be seen in table 5.

After that, the prediction for the number of calls of the test dataframe was made and the results were compared with the actual values of this variable (see appendix 5). The Root Mean Squared Error (RMSE) is shown in table 5.

After this, the prediction for the year of 2019 took place, as the model was verified as a good one based on the R^2 (both above 0.4) and the RMSE (low when compared to the values of the number of calls that vary between 122 and 1260). For the prediction for the year of 2019 (the new testing data frame), the values obtained for the R^2 are in table 4 and the values for the coefficients and for the RMSE are in table 5. Figure 23 shows the values predicted and the real values of the number of calls. The regression model and its comparison with the actual data is shown in figure 24.

Table 4 - Squared R of simple linear regression model for several data tested.

Data frame	**R^2**
Train (80% of 2018 data)	0.5603
Test (20% of 2018 data)	0.4628
Train (2018 data)	0.5415
Test (2019 data)	0.5488

Table 5 - Coefficients of simple linear regression model for prediction of the number of calls in 2018 and 2019 and RMSE obtained.

Model - Linear regression	β_0	β_1	**RMSE**
Prediction of 2018 test data	128.872	0.247	219.00
Prediction of 2019 data	127.791	0.252	196.81

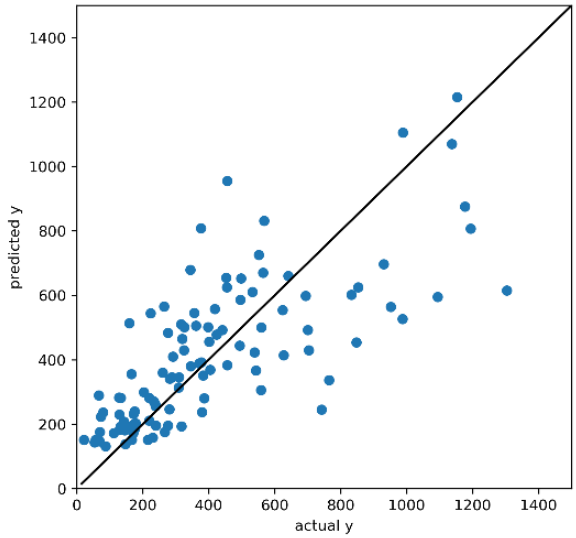


Figure 22 - Predicted y compared with actual values, y being the number of ambulance calls related to the data frame from 2019.

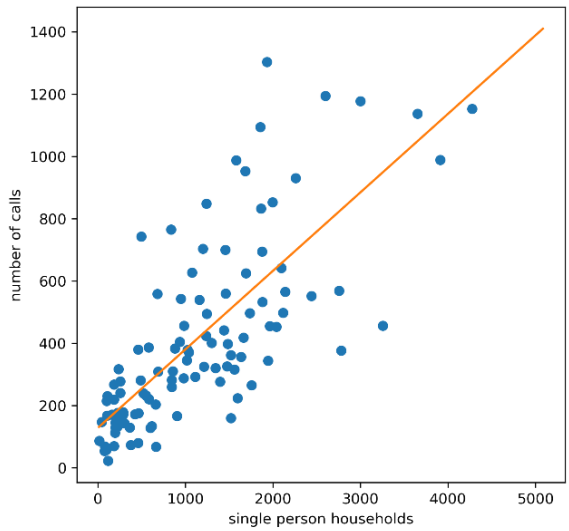


Figure 23 - Predicted data compared with actual values. The orange line represents the predicted values and the blue dots the actual data from the data frame of 2019.

As it can be seen, the coefficients are very similar for both models which is valid because the same indicators were used and the only difference relies on the fact that for the prediction of 2019 data, the training data was the 100% data frame from 2018. It can also be acknowledged that the squared R is better for the test data of 2019, being 0.55 instead of 0.46. This is because there is a lot more data in this model and so small deviations from what is expected have less impact.

As for the absolute error of this model, Figure 25 shows all the residuals for the prediction of 2019 (difference from predicted and actual data), and the maximum of the absolute residuals is 688 in neighbourhood Houtwijk.

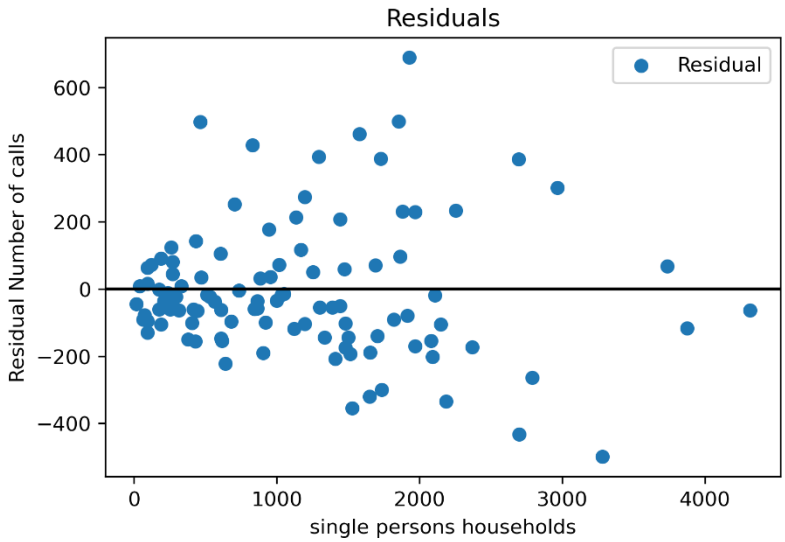


Figure 24 - Residuals from the simple linear regression model for the prediction of the 2019 ambulance calls data.

This analysis was also done for the training/testing initial model and it can be seen in appendix 5, figure 5.3.

A polynomial fit of the data of degree 2 was also achieved, having led to very similar results (see appendix 6).

4.2. Final model

In order to have a better model for the prediction of the number of ambulance calls of 2019 than the simple linear regression, a multiple regression model was created based on the predictors: all offences, people with 65 years or older, single persons households, and people with a disability benefit. These were the variables that lead to a multiple regression model with a higher R^2. Firstly, the variable number of inhabitants was chosen to be a part of the model, but then it was left out for two reasons. The overall model led to the worst results (worst RMSE and R^2) and this variable does not give insight about the socio-economic aspects of each neighbourhood.

The equation that describes the final model obtained is

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + \beta_4 \times x_4,$$

where y is the number of ambulance calls, x_1 is the number of offences, x_2 is the number of people with 65 years or older, x_3 is the number of single persons households, and x_4 is the number of people with a disability benefit.

These four indicators were considered very relevant because they are indicators of social vulnerability which is the main focus of this study. The relation between social vulnerability and the need to call the ambulance is well established by this model. The economic factors were not considered in this model, but their relationship with the ambulance calls was also explored previously and other models could be done taking into account other economic vulnerability indicators (as the low income one).

This model was first tested for 2018 (see appendix 8) and then used to predict the data from 2019 (figure 26).

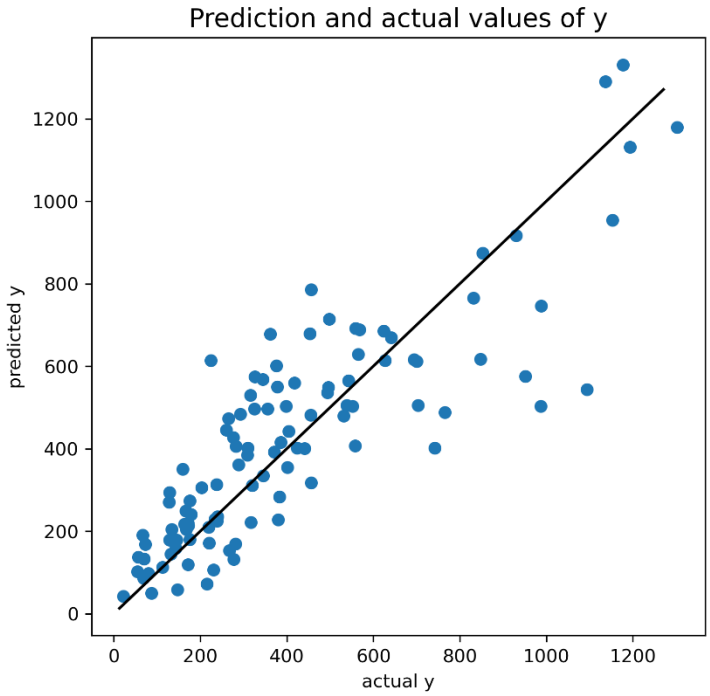


Figure 25 - Predicted y compared with actual values based on multiple regression model, y being the number of ambulance calls related to the test data frame from 2019.

Table 5 shows the obtained coefficients of this model for the prediction of the data of 2019 and the RMSE. The values of the R^2 are shown in table 6 and, as we can see, they are both above 0.7 which means the multiple regression model is a very good prediction one and could even be used to predict for other years whose data are not available yet.

Table 6 - Coefficients of multiple regression model for prediction of the number of calls in 2019 and RMSE obtained.

Model - Multiple regression	β_0	β_1	β_2	β_3	β_4	**RMSE**
Prediction of 2018 test data	23.657	0.397	0.230	0.029	0.348	153.99

Table 7 - Squared R of multiple linear regression model for the train (2018) and test (2019) data.

Data frame	**R^2**
Train (2018 data)	0.7126
Test (2019 data)	0.7238

By analysing the coefficients of this model, what it reveals is that if the coefficient is higher than 0 it means that the number of calls will increase if the predictor is also higher. In this case, all 4 coefficients are higher than 0, which was what was already expected. These are supporting points of our initial argument.

This means that the neighbourhoods with higher number of offences, older people, more people living alone and more disabled people, will have more ambulance requests, which makes these neighbourhoods more vulnerable and more in need of, for instance, of investment in health services. With this said, policies can be made based on these analyses.

For the next step, the accuracy of the multiple regression model was studied. This model has a high R2 and a low RMSE (given the interval of number of calls). This leads to the conclusion that its accuracy is high. This accuracy can also be quantified spatially, and it can be seen for which neighbourhoods the model developed predicts better and for which it predicts worst. Figure 27 shows the distribution per neighbourhood of three variables: the predicted values of the number of ambulance calls for 2019; the actual values for these calls and the absolute error in the number of calls of our prediction.

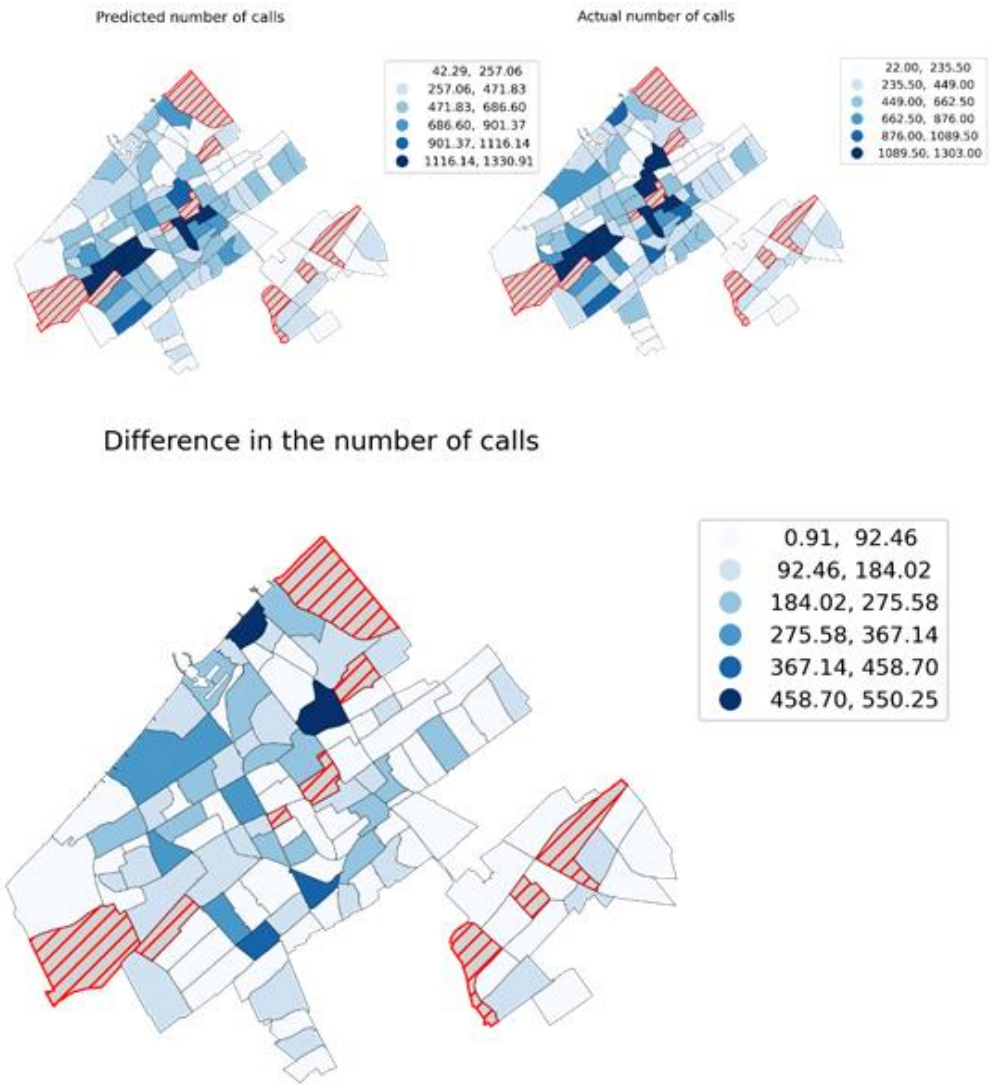


Figure 26 - Spatial distribution of ambulance calls in The Hague in 2019. Upper left are the predicted values based on the multiple regression model. Upper right are the actual values from the data frame of 2019. Lower is the distribution of the absolute error (residuals) between the predicted and the actual values. The hashed grey and red neighbourhoods are the ones that were removed from the data frame (as discussed previously).

In figure 26 it is possible to observe that some neighbourhoods were taken out from our experiment. These neighbourhoods were evident outliers that were removed in the chapter of the Exploratory Data Analysis. No more neighbourhoods were removed from the dataframe after the learning algorithms were developed.

The neighbourhood with the worst prediction is Archipelbuurt, with 550.25 calls of absolute error, which is equivalent to 1.5 calls per day. On the other hand, the neighbourhood with the best prediction is *Rietbuurt* with an absolute error of only 0.9 calls in the whole year. This last neighbourhood is a good neighbourhood that typifies our pattern because our model predicts perfectly

well for its number of calls. This means that the vulnerability indicators chosen were very good predictors for the number of ambulance calls. The small absolute error (the pattern) provides evidence for the supporting points that each of the four socio-economic indicators are good predictors.

5. Conclusions & Discussion

As it happens around the world, decision-makers in the city of The Hague have the difficult task of distributing correctly and efficiently the resources of public and emergency services, such as ambulance services. As in the whole Netherlands, The Hague's allocation of ambulances is done by mathematical models that measure the time of response of paramedics to reach every citizen after the call (RIVM, 2017&2018). Looking for a more complete analysis, this research aimed to study if the impact or influence of socio-economic vulnerability of each neighbourhood of The Hague in the number of calls for ambulance services of that region is significant and if with that information it is possible to allocate The Hague's ambulance resources efficiently.

Hence, to study this possible correlation between ambulance calls and socio-economic characteristics of The Hague's society, thirteen socio-economic indicators were analysed (see Appendix 1) From the Exploratory Data Analysis performed to which indicator, it was clear that four of these indicators have higher correlation (and so, more impact) with the number of ambulance calls in each neighbourhood:

1. All offences
2. Number of people that are 65 years or older
3. Number of single-person households
4. People with disability benefit

Further, from the appliance of the LISAs method was possible to conclude that the neighbourhood located in the centre of The Hague were a particular area of interest for this research, since there is a particularly high concentration of neighbourhoods with a high volume of ambulance calls, and so could be a relevant area in terms of ambulance service management.

Consequently, the multi-regression used as the final research model is based on these four indicators. Training this multi-regression with data from 2018 and then testing it with data from 2019, the model scored an $R^2 = 0.7$ and an RMSE = 153.99. Evidently, with an R^2 near 1, the model shows that these socio-economic indicators impact the number of ambulances. Indeed, the relationship between these indicators and the number of emergency ambulance calls is not difficult to understand. In neighbourhoods with higher numbers of reported crimes, there is a higher probability of people getting hurt and, thus, of needing an ambulance. In the same way, people older than 65 years old or with

some kind of disability are also highly susceptible to getting suddenly sick or severely hurt (compared to the rest of the population) and that is why the number of ambulance calls is higher from neighbourhoods with a higher number of people older than 65 years or a higher number of people with a disability benefit. Finally, the tendency for neighbourhoods with more single-person households to have a higher number of calls can be explained by the lack of some familiar or cohabitant to help them and, thus, a higher dependency on calling the 112 emergency number.

As so, we may conclude that our model shows promising results, and that we can make predictions for the number of ambulance calls based on the referred indicators.

However, in order to introduce possible improvements in future research about topics in the same or similar scope, the limitations of this research and some future research topics will be enumerated below.

5.1. Limitations & Future Research

- At first, this research model was trained based on information from 2018 and tested for the year 2019. Likely, this model does not fit the data of the pandemic years (2020 and 2021), when suddenly the habits and routines of the population suffered a complete change. Furthermore, any other year with a sudden catastrophe will not be possible to predict with this model.

- The thirteen socio-economic indicators were a limitation once they were predefined before the analysis. The study of the correlation between the number of ambulance calls and some more socio-economic indicators could be useful and reveal more influential indicators worthy of being used in the final model multi-regression.

- The outliers of the dataset were excluded from the final multi-regression once they completely distorted its main conclusions. These three neighbourhoods (*Waalsdorp*, *Kortenbos* and *Transvaalkwartier-Noord*) have more than two thousand calls registered per year, which are divergent numbers compared with the rest of the dataset. Even so, these neighbourhoods are part of The Hague region. So, when making a decision, it is important to take these neighbourhoods and the ones without inhabitants in consideration. Should be performed, in future researches, some analysis to understand the reason for such number of calls in these neighbourhoods.

- The system does not consider some factors like the number of calls under different weather conditions, daytime versus nighttime or on different days of the week/year (weekends versus weekdays; general holidays versus workdays; Summer versus Winter). This type of information would increase the quality of the model and then increase the conclusions to take from it.

- In this research, only three models were tested: simple linear regression, polynomial regression and multi-linear regression. Therefore, it is doubtful whether other nonlinear model would be better suitable for the actual dataset.

- The model would be improved if tested for more regions of The Netherlands and/or with data from more years.

Overall, from the research, it is possible to conclude that socio-economic factors impact the number of ambulance calls in the region of The Hague and that our model is reliable to predict the number of ambulance calls and can be used to help decision-makers to allocate the right number of ambulances to each station of the region and the right investment in health services nearby each neighbourhood.

References

[1] Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big Data and Data Science in Critical Care. Chest, 154(5), 1239–1248. https://pubmed.ncbi.nlm.nih.gov/29752973/

[2] Macdonald, B. (2020). Recreating the Game: Using Player Tracking Data to Analyze Dynamics in Basketball and Football. Harvard Data Science Review, 2(4). https://hdsr.mitpress.mit.edu/pub/kxks56er/release/4

[3] Kandt, J., & Batty, M. (2019). Smart cities, big data and urban policy: Towards urban analytics for the long run. Cities, 109, 102992. https://www.sciencedirect.com/science/article/pii/S0264275120313408

[4] Bibri, S. E. (2021). Data-driven smart sustainable urbanism: the intertwined societal factors underlying its materialization, success, expansion, and evolution. GeoJournal, 86(1), 43–68. https://link.springer.com/article/10.1007/s10708-019-10061-x

[5] Yu, D., Yin, J., Wilby, R., Lane, S., Aerts, J., Lin, N., Liu, M., Yuan, H., Chen, J., Prudhomme, C., Guan, M., Baruch, A., Johnson, C., Tang, X., Yu, L. & Xu, S. (2020). Disruption of emergency response to vulnerable populations during floods. Nature Sustainability, 3, 728-736 https://www.nature.com/articles/s41893-020-0516-7

[6] Toregas, C., Swain, R., ReVelle, C. & Bergman, L. (1971). The Location of Emergency Service Facilities. Operations Research, 19(6),1363-1373. https://pubsonline.informs.org/doi/abs/10.1287/opre.19.6.1363

[7] RIVM. Referentiekader spreiding en beschikbaarheid ambulancezorg 2018 (2018-0128). https://www.rivm.nl/bibliotheek/rapporten/2018-0128.pdf

[8] C. Swoveland, C., Uyeno, D., Vertinsky, I., Vickson, R. (1973). Ambulance Location: A Probabilistic Enumeration Approach. Management Science, 20(4.2), 686-698. https://pubsonline.informs.org/doi/abs/10.1287/mnsc.20.4.686

[9] Uyeno, D. & Seeberg, C. (1984). A practical methodology for ambulance location. Simulation 79, 43(2), 79-87 https://journals.sagepub.com/doi/10.1177/003754978404300202

[10] Cantwell K., Dietze, P., Morgans, A. & Smith, K. (2012). Ambulance demand: random events or predicable patterns? Emergency Medicine Journal, 30(11), 883-887 https://pubmed.ncbi.nlm.nih.gov/23184922/

[11] Aldrich, C., Hisserich, J., Lave, L. (1971). An analysis of the demand for emergency ambulance service in an urban area. American Journal of Public Health, 61(6), 1156-1169 https://pubmed.ncbi.nlm.nih.gov/5112963/

[12] Elliot, A., Smith, S., Dobney, A., Thornes, J., Smith, G., Vardoulakis, S. (2016) Monitoring the effect of air pollution episodes on health care consultations and ambulance call-outs in England during March/April 2014: A retrospective observational analysis. Environmental Pollution, 214, 903-911 https://pubmed.ncbi.nlm.nih.gov/27179935/

[13] Bray, J., Straney, L., Barger, B. & Finn, J. (2015) Effect of Public Awareness Campaigns on Calls to Ambulance Across Australia. American Heart Association, 46(5), 1377-1380 https://www.ahajournals.org/doi/full/10.1161/strokeaha.114.008515

Appendices

Appendix 1 - Indicators used and respective source plus description

Source	Indicator	Description
Course material	Ambulance emergency calls	Number of ambulance calls in each neighbourhood.
Den Haag Cijfers	Potential Labour force	Number of 15 to 64-year-olds.
Den Haag Cijfers	All offences	All reports of crimes and of nuisance.
Den Haag Cijfers	Average number of private cars per address (excluding lease cars):	Average number of registered private cars per address in The Hague neighbourhoods (excluding leased cars).
Den Haag Cijfers	% Dutch natives:	Percentage of Dutch natives in the population.
CBS	Number of inhabitants	Number of inhabitants
CBS	Number of people that are 65 years or older	Number of residents who are 65 years of age or older on 1 January.
CBS	Single person households number	Number of private households consisting of one person.
CBS	Households without children	Number of multi-person households without children, which consist of unmarried couples without children, couples without children and other households.
CBS	Households with children	Number of multi-person households with children, which consist of unmarried couples with children, couples with children and single-parent households.
CBS	People with disability benefit	Number of people who receive a disability benefit under the Disability Insurance Act (WAO), the Self-employed Disability Insurance Act (WAZ), the Work and

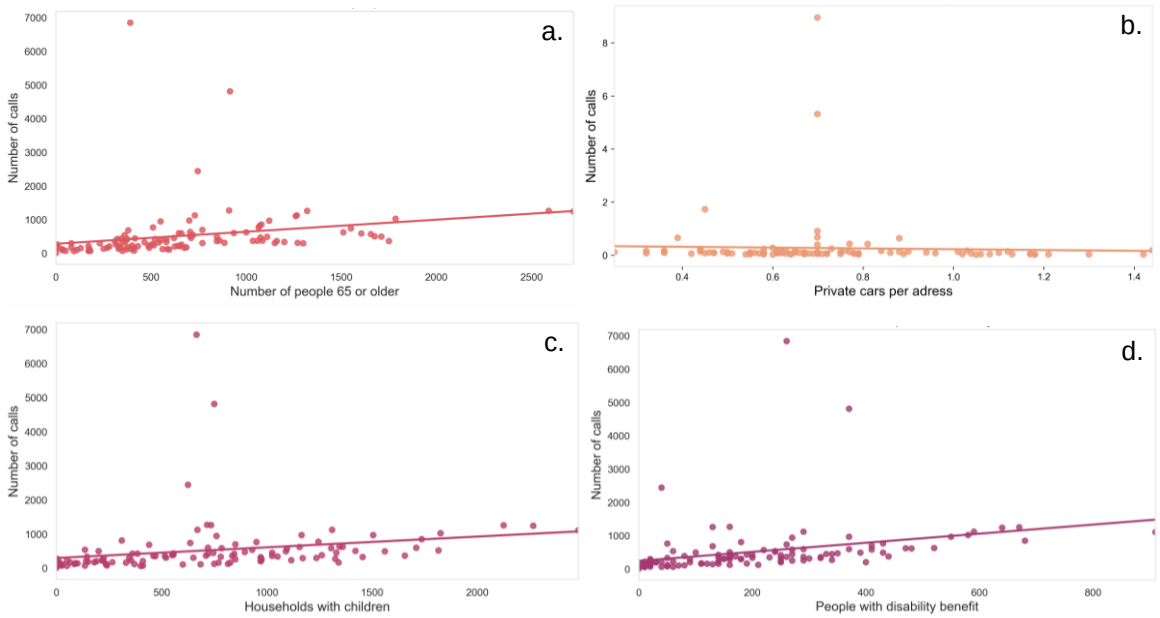
		Employment Support for Young Persons with Disabilities Act (Wajong Act) and the Work and Income According to Work Capacity Act (WIA).
CBS	Distance to GP	The average distance of all residents in an area to the nearest general practitioner in kilometers.
CBS	Low-income households	Percentage of low-income households which concerns student households on the one hand and households with an incomplete annual income on the other.
CBS	Median wealth of private households	Median wealth of neighbours (x 10^3). Wealth is the balance of assets and liabilities. Assets are formed by bank and savings balances, securities, the owner-occupied home, other real estate, business capital, substantial interest and other assets while debts include debts for an owner-occupied home and consumer credit.

Appendix 2 - Neighbourhood denominations in the The Hague in Cijfers and CBS data sources

Denomination in The Hague in Cijfers Data	**Denomination in CBS Data**
'Kerketuinen/Zichtenburg'	'Kerketuinen en Zichtenburg'
‘Tedingerbroek'	'Tedingerbuurt'
‘Koningsplein eo'	'Koningsplein en omgeving'
'Venen/Oorden/Raden'	‘Venen, Oorden en Raden'
'Parkbuurt Oosteinde'	'Parkbuurt oosteinde'
'v Hoytemastraat eo'	'Van Hoytemastraat en omgeving'
'Zijden/Steden/Zichten'	'Zijden, Steden en Zichten'

'Bohemen/Meer en Bos'	'Bohemen en Meer en Bos'
'Sweelinckplein eo'	'Sweelinckplein en omgeving'
'Eykenduynen'	'Eykenduinen'
'v Stolkpark/Schev Bosjes'	‘Van Stolkpark en Scheveningse Bosjes'
'Kraayenstein & Vroondaal'	'Kraayenstein en Vroondaal' (information relative to 2019)
	'Kraayenstein' (information relative to 2018)

Appendix 3 - Regression plots obtained for all the studied variables as eventual predictors (with outliers)



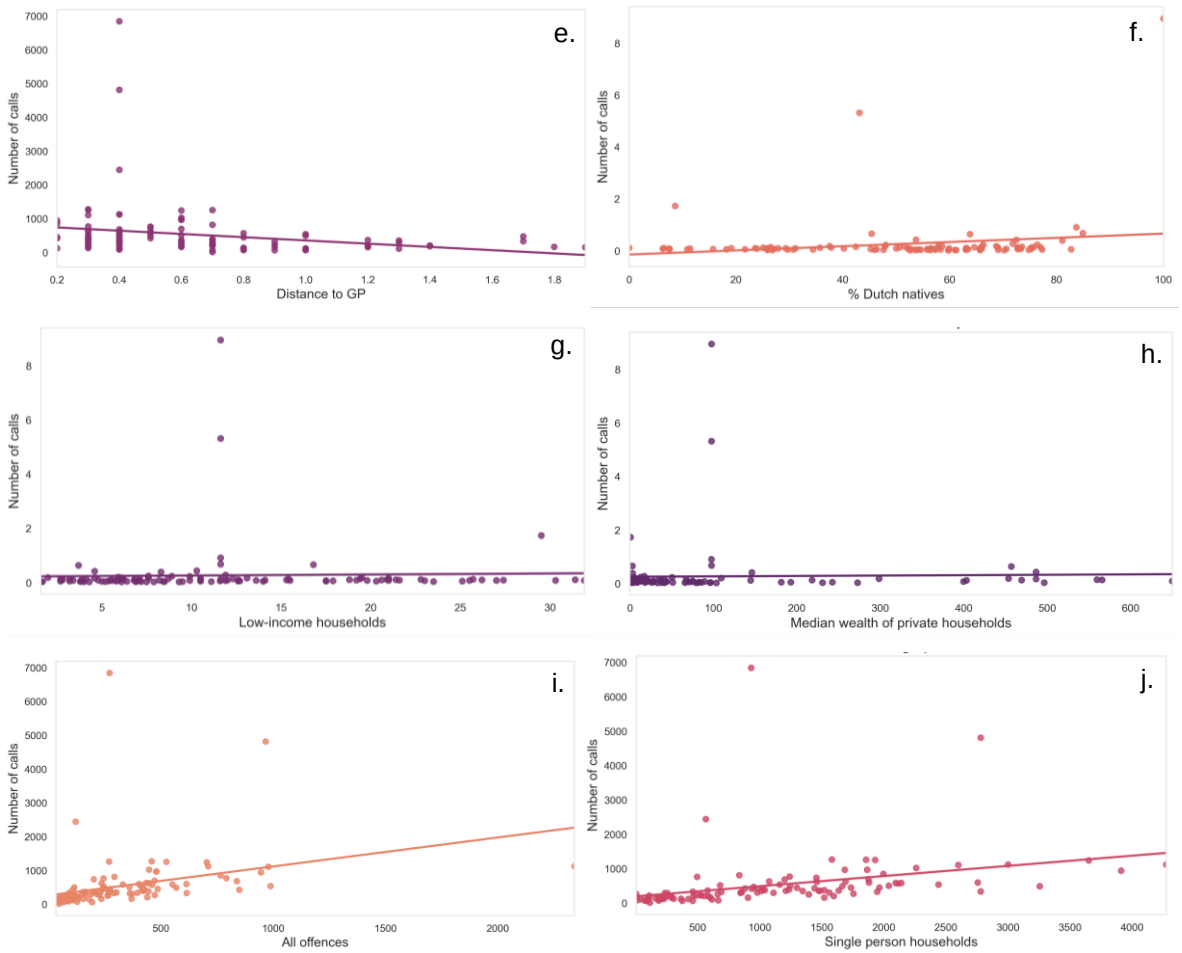


Figure 3.1 - Regression plots for Number of calls vs. each one of the studied variables, with outliers: a) Number of people 65 or older, b) Private cars per address, c) Households with children, d) People with disability benefit, e) Distance to GP, f) Percentage of dutch natives, g) Low-income households, h) Median-wealth of private households, i) All offences, j) Single person households.

Appendix 4 - Boxplots and histograms for each studied variable

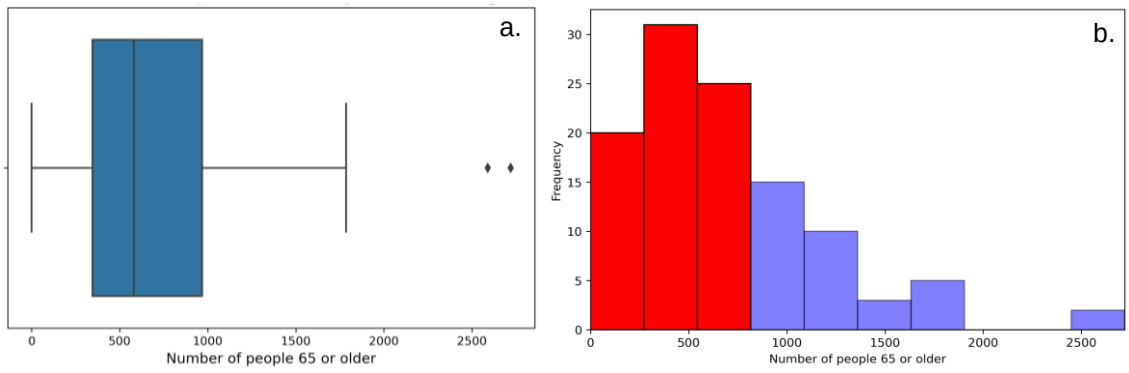


Figure 4.1 - Boxplot (a) and histogram (b) for the number of people 65 or older.

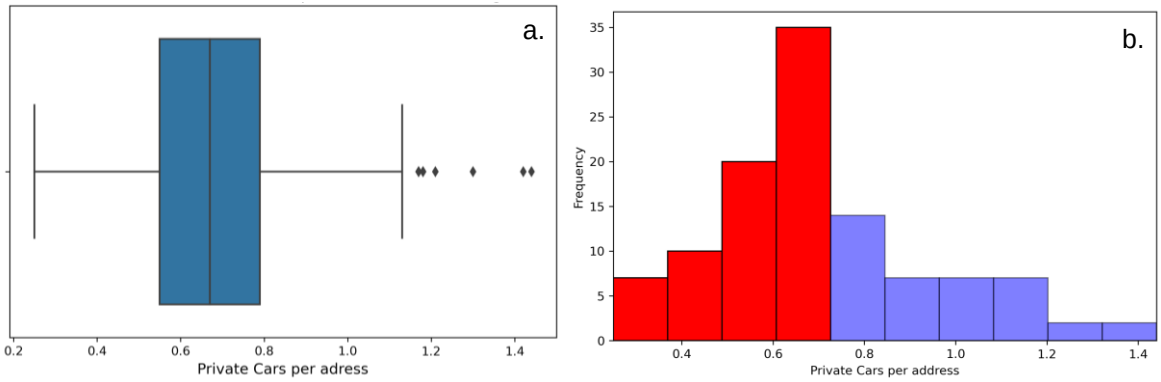


Figure 4.2 - Boxplot (a) and histogram (b) for the private cars per address.

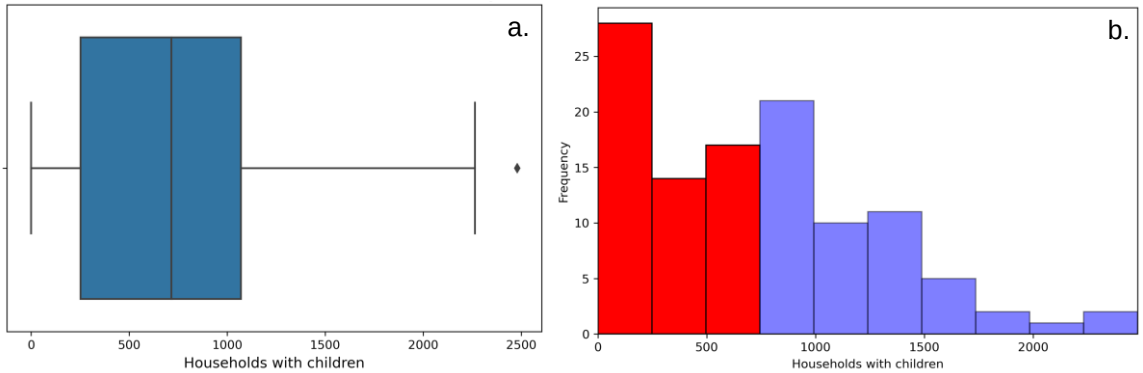


Figure 4.3 - Boxplot (a) and histogram (b) for the number of households with children.

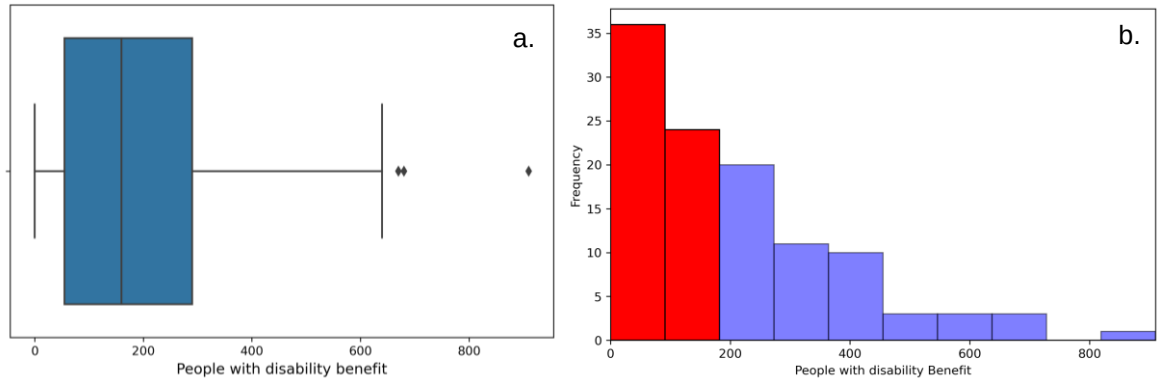


Figure 4.4 - Boxplot (a) and histogram (b) for the number of people with disability benefit.

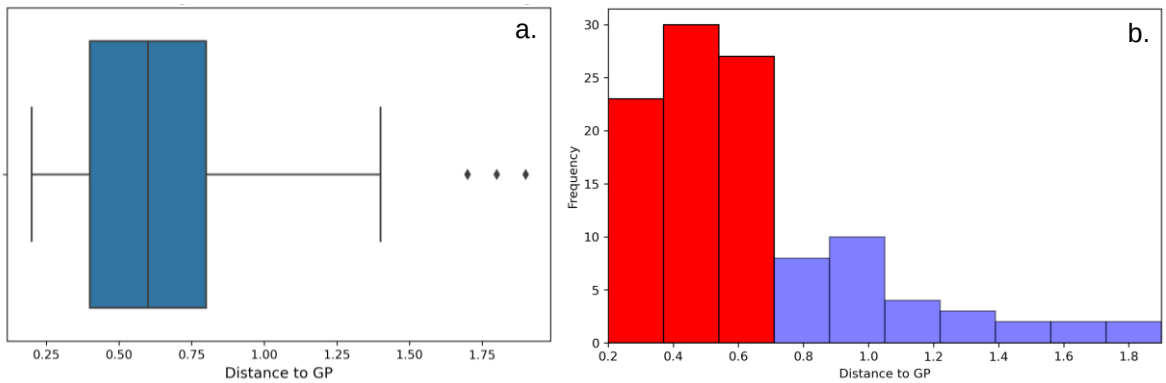


Figure 4.5 - Boxplot (a) and histogram (b) for the distance to GP.

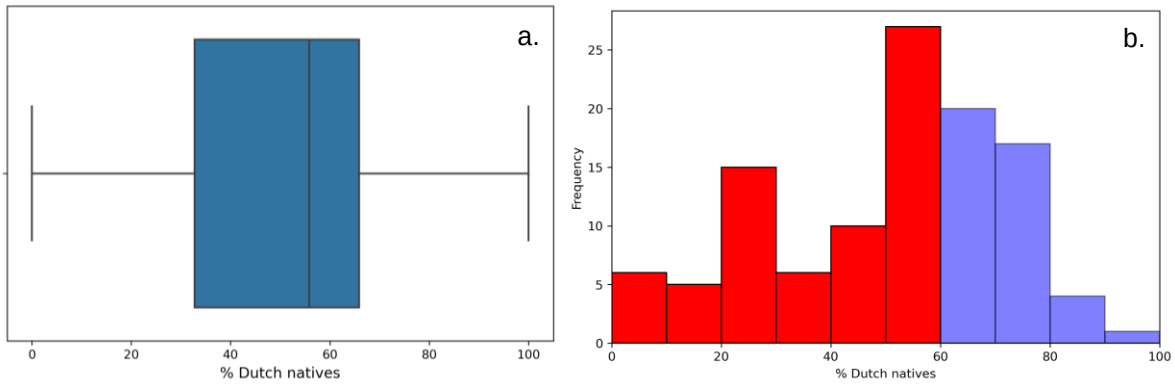


Figure 4.6 - Boxplot (a) and histogram (b) for the percentage of dutch natives.

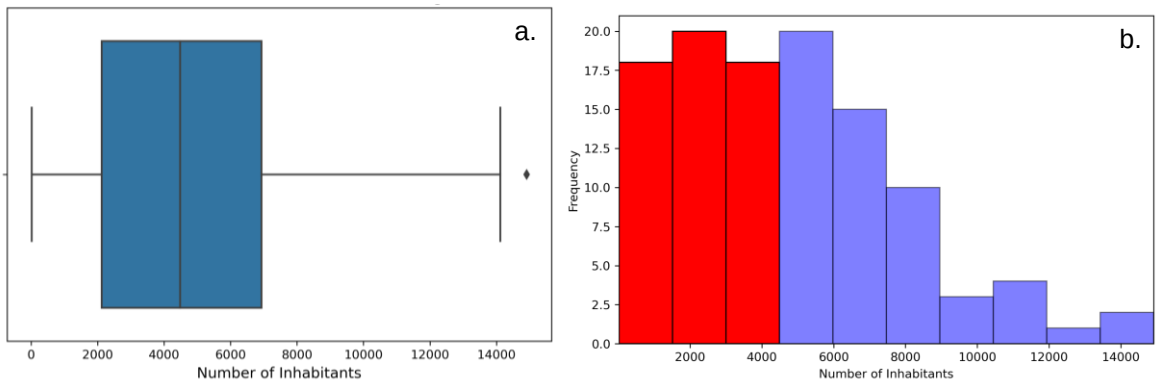


Figure 4.7 - Boxplot (a) and histogram (b) for the number of inhabitants.

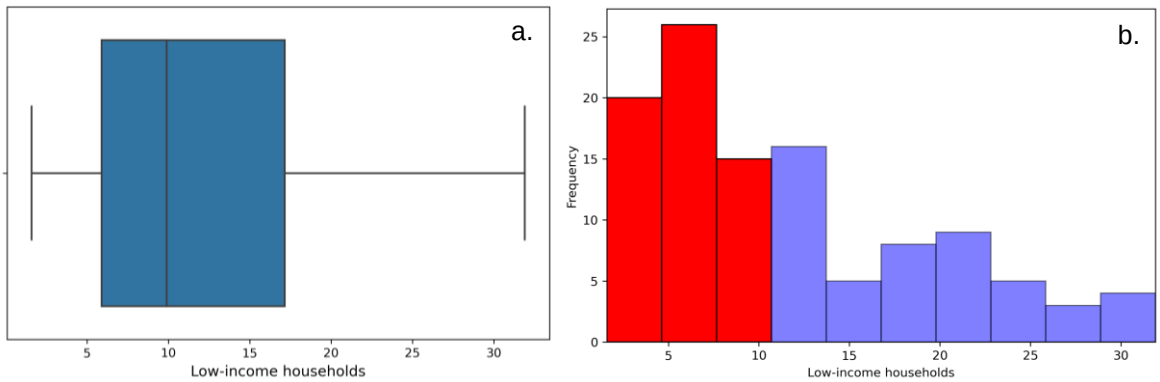


Figure 4.8 - Boxplot (a) and histogram (b) for the percentage of low-income households.

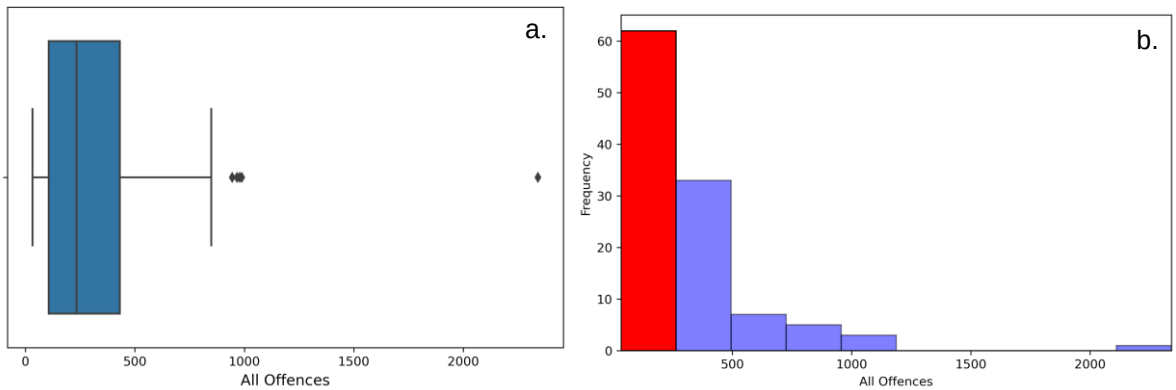
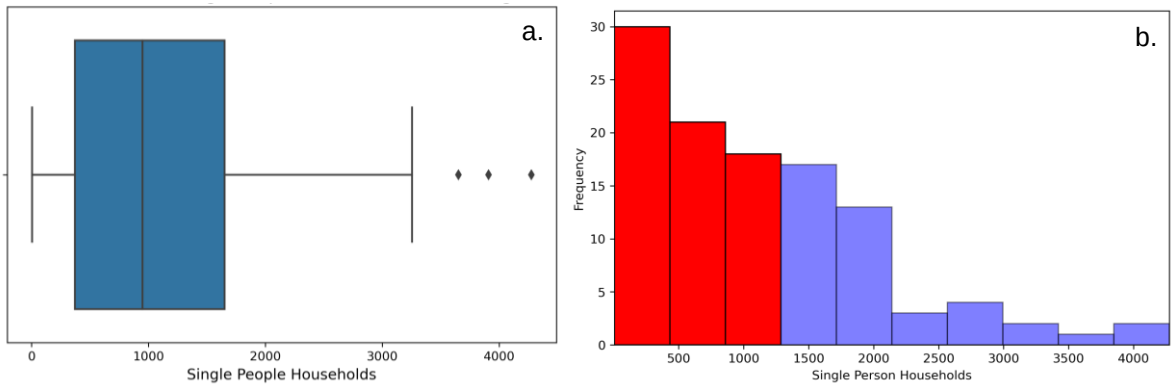


Figure 4.9 - Boxplot (a) and histogram (b) for the number of all offences.



igure 4.10 - Boxplot (a) and histogram (b) for the number of single person households.

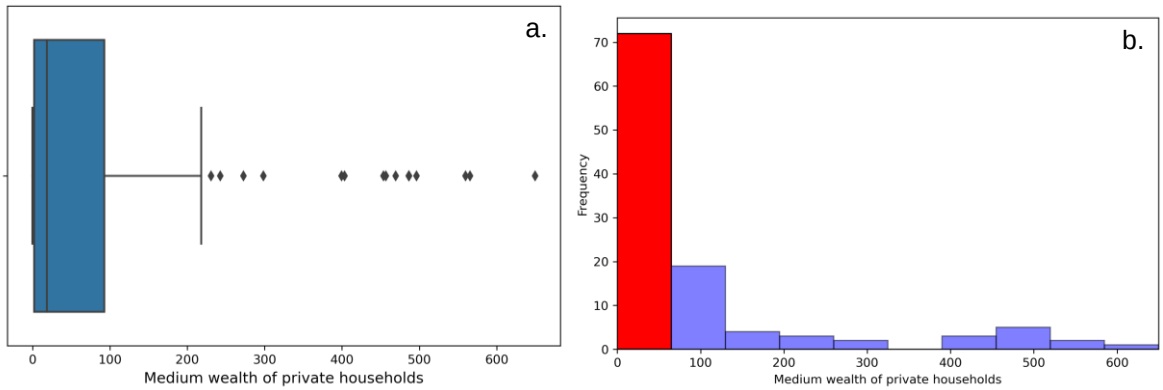


Figure 4.11 - Boxplot (a) and histogram (b) for the medium wealth of private households.

Appendix 5 - Results of linear regression prediction model for 2018 test data, based on 2018 train data.

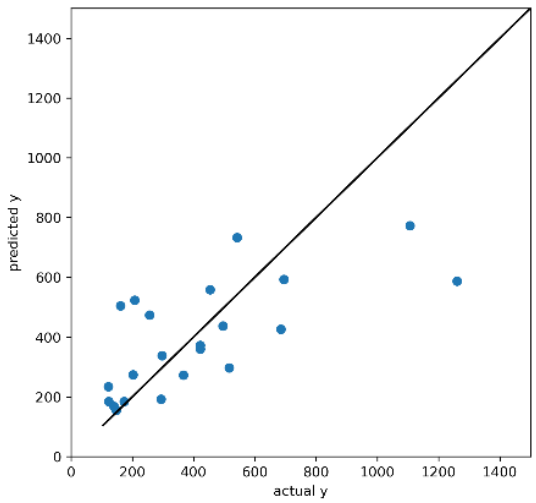


Figure 5.1 - Predicted y compared with actual values, y being the number of ambulance calls related to the test data frame from 2018.

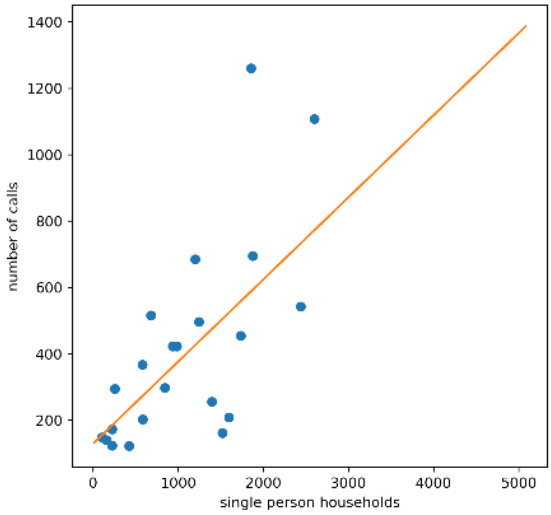


Figure 5.2 - Predicted data compared with actual values. The orange line represents the predicted values and the blue dots the actual data from the test data frame of 2018.

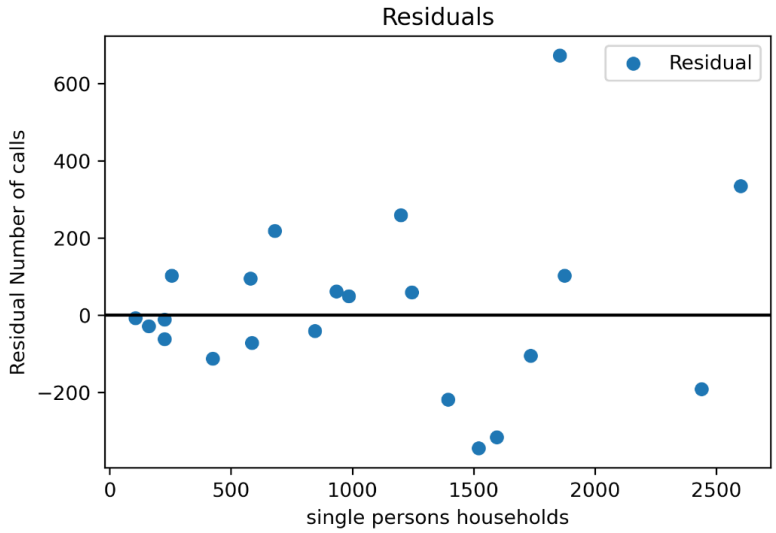


Figure 5.3 - Residuals from the simple linear regression model for the prediction of the 2018 test data frame.

Appendix 6 - Results of the polynomial regression model

Table 6.1 - Squared R of polynomial regression model of degree 2 for the prediction of 2019 ambulance calls data.

Data frame	**R^2**
Train (2018 data)	0.5512
Test (2019 data)	0.5511

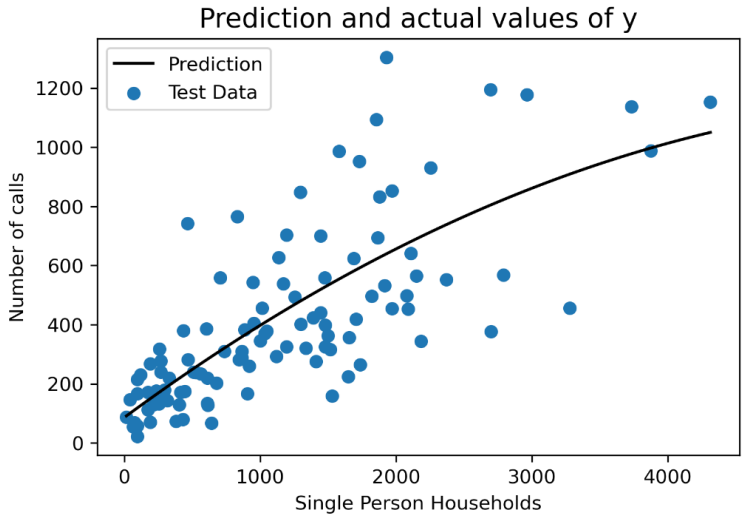


Figure 6.1 - Predicted data compared with actual values for polynomial model of degree 2. The black curve represents the predicted values and the blue dots the actual data from the data frame of 2019.

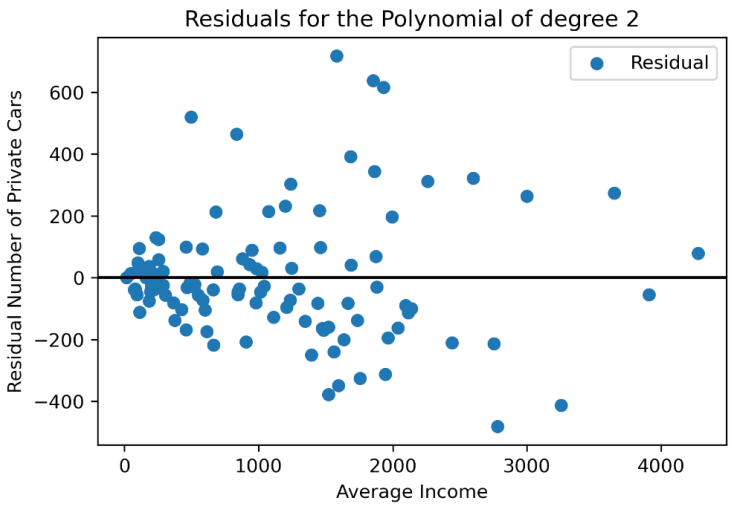


Figure 6.2 - Residuals from the polynomial regression model of degree 2 for the prediction of the 2019 number of ambulance calls.

Appendix 8 - Results for multiple regression model: testing/training learning algorithm

Table 8.1 - Squared R of a multiple linear regression model for several data tested.

Data frame	**R^2**
Train (80% of 2018 data)	0.7491
Test (20% of 2018 data)	0.5479

Table 8.2 - Coefficients of multiple regression model for prediction of the number of calls in 2018 and RMSE obtained.

Model - Multiple regression	β_0	β_1	β_2	β_3	β_4	**RMSE**
Prediction of 2018 test data	20.738	0.389	0.211	0.017	0.475	200.90

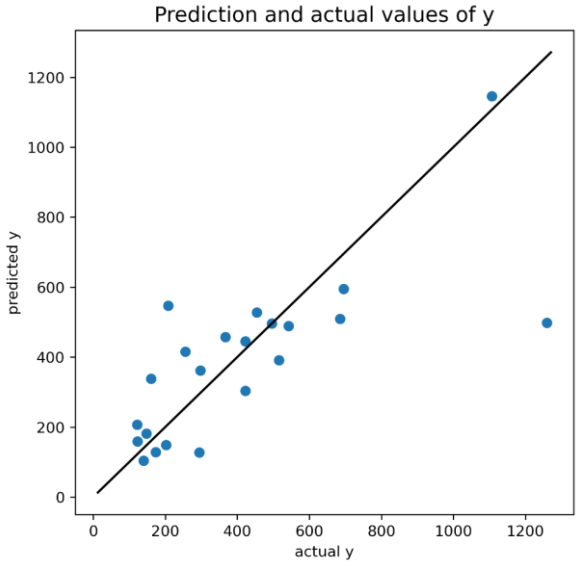


Figure 8.1 - Predicted y compared with actual values based on multiple regression model, y being the number of ambulance calls related to the test data frame from 2018.