Linear Regression on Ambulance Calls in The Hague Neighbourhoods Based on Socio-Economic and Mental Health Indicators

EPA1316A Introduction to Urban Data Science

Final Project

Professor: Dr. Trivik Verma

Group 5

Lee Wen Qi - 5846463

Loh Ying Xin - 5850266

Titouan Fayet - 5845408

Varshana Kumanan - 5845882

Abstract

Globally, the demand for emergency medical services has been growing steadily over the years at a rate of 3-5% annually (Nehme et al., 2019). Research suggests that socio-economic factors are known to influence the demand for ambulances and can be used to predict it (Siler, 1975). We are also interested to know if the mental health of residents would affect this demand as mental health is a rising topic in recent times. The aim of this study is to understand whether such factors can be used to accurately predict the number of ambulance calls in neighbourhoods of The Hague, a city in The Netherlands. After looking at research work related to this topic, we obtained from CBS four indicators of socio-economic in the neighbourhoods of The Hague (open-source), and obtained from Professor Verma the private data on emergency calls made in The Netherlands. Exploratory data analysis and modelling was then done, with 2 predictors obtained in the end: percentage of single-person households and percentage of people under social benefits. The linear regression model was trained on data from 2017 to 2019, with a final R-squared value of 0.147 obtained. Despite this, the 2 predictors have statistically significant coefficients of 0.040 and 0.000 respectively (p-value < 0.05). This suggests that socio-economic factors could be used to improve the prediction of ambulance demand in The Hague, given that the relevant factors are chosen.

Contents

1. Introduction

In any city, different neighbourhoods may have varying living conditions. Examples of these living conditions are access to clean water, infrastructural connectedness and access to amenities. The socio-economic status (SES) of a neighbourhood is influenced by the above-mentioned living conditions, together with other factors, such as the average income and average education level attained of a neighbourhood. It plays a significant role in one's health (Braveman & Gottlieb, 2014).

With greater disparities in SES across neighbourhoods, inequities in health distribution are increasing as well (Wang & Geng, 2019). Neighbourhoods with a lower socio-economic status tend to have poorer physical infrastructure (CSDH, 2015) and fewer amenities. This indicates poorer access to healthcare services; it may take a longer time for an ambulance to reach a household with a health emergency, and there may be fewer clinics and hospitals in these neighbourhoods. Lower SES is also correlated with poorer health (American Psychological Association, 2014), suggesting a greater number of healthcare emergencies in the area.

Census from CBS has indicated that mental health has worsened in the Netherlands (CBS, 2022). Additionally, data from 1950 through 2021 has seen an increasingly high rise in suicide cases in the Netherlands as well (CBS, 2022). This translates to about 559 cases in 1950 to 1861 cases in 2021 itself. With rising mental health concerns, potentially life threatening incidents may occur which requires speedy assistance from emergency services. With the prediction of ambulance calls on this factor, some of these cases could be prevented.

Therefore, our research question is: Can socio-economic status and mental health be used to predict the number of ambulance calls in the neighbourhoods of The Hague?

To answer this question, these datasets are used:

1. Emergency calls dataset (ambulance, firefighter, police, and coastguards) of The Netherlands collected from January 2017 to September 2020, obtained from Professor Verma
2. Data from CBS on indicators of socio-economic in the neighbourhoods of The Hague

In this article, the following indicators are used to determine socio-economic status of a neighbourhood (Agarwal et al., 2019):

1. Extent of poverty: Percentage of low-income households

2. Access to healthcare services: Average distance to the nearest general practitioner (GP)
3. Extent of social isolation and loneliness: Percentage of single-person households
4. Extent of people requiring social assistance:
    a. Number of people receiving social assistance benefit / Number of inhabitants in a neighbourhood
    b. Number of people receiving disability benefit / Number of inhabitants in a neighbourhood
    c. Number of people receiving unemployment benefit / Number of inhabitants in a neighbourhood
    d. Number of people receiving pension / Number of inhabitants in a neighbourhood

If socio-economic status and mental health can be used to accurately predict the number of ambulance calls in a neighbourhood, the findings from this study will allow policymakers (of The Hague) to understand how healthcare resources, especially ambulance and other emergency healthcare services, can be better allocated to neighbourhoods based on socio-economic status.

The linear regression model had a final R-squared value of 0.147. Even though this value indicates that the model used is weak and unsuitable to determine ambulance demand in different neighbourhoods of The Hague, the removal of less relevant features lead to an increase in the R-squared value due to removal of noise from the data. Furthermore, the coefficients of the 2 final features in the model have statistically significant coefficients (p-value < 0.05), meaning they could be used as predictors for ambulance demand in The Hague. This suggests that socio-economic factors can improve the prediction of ambulance demand in The Hague, with further research conducted on such factors to obtain a more reliable and suitable model, together with the inclusion of other factors that are good predictors as well.

2. Related Work

It is important to understand how socio-economic status influences the frequency of ambulance calls in a neighbourhood, as it will allow for better resource allocation in a city, and for the needs of different neighbourhoods to be met appropriately.

In The Netherlands, the acute care network consists of various services, one of them being ambulance services. Over the years, especially due to the Covid-19 pandemic, the pressure on the network has increased. This indicates that ambulance services may not be distributed appropriately across neighbourhoods, where some may require more emergency medical services than others, depending on various factors. The Hague is a city with around 800,000 people, and a large number of healthcare providers. It is also a city with one of the largest proportions of low-income households in The Netherlands, meaning a greater proportion of neighbourhoods in The Hague have low socio-economic status (Minderhout et al., 2021). It is thus important for the city to plan and distribute ambulance services appropriately, so understanding whether socio-economic status can predict the number of ambulance calls in a neighbourhood will be beneficial to The Hague.

The Netherlands consists of 25 Regional Ambulance Service (RAV) regions, for which the Dutch Minister of Health, Welfare and Sport has designated a RAV in each region. Each RAV has a number of ambulance posts, determined by the National Reference Framework for Distribution & Availability (Ambulancezorg Nederland, 2019). It is an optimal theoretical distribution of posts calculated using different factors that do not include socio-economic factors (Kommer G.J. et al., 2020).

2.1 Social Factors

Plenty of research has been conducted on factors that influence frequency of health emergencies in a neighbourhood. One study on frequent callers of emergency medical services found that 67% of them were lonely, 43% had insufficient income to get through a month, and that frequent users of these services tended to be from vulnerable people with poorer health (Agarwal et al., 2019). Another study suggests that between 2 areas in the UK, the one with a lower socio-economic status had a greater number of ambulance calls for more serious health issues (Portz et al., 2013). These studies suggest that with lower socio-economic status and poorer mental health, people tend to rely more heavily on emergency medical services. However, what is yet to be found is whether these socio-economic and mental health factors can be used to accurately predict the frequency of ambulance calls, specifically, in The Hague.

2.2 Problem Statement

Thus, in this article, we investigate the following research question: Can socio-economic status and mental health be used to predict the number of ambulance calls in the neighbourhoods of The Hague? Answering this question allows policymakers of The Hague to consider altering the allocation of healthcare resources to neighbourhoods. The Dutch government will also be able to allocate ambulances more optimally across all neighbourhoods.

3. Exploratory Data Analysis

This chapter will discuss how the data was cleaned and how the different variables we will be working with were explored. The list of socio-economic and mental health indicators were mentioned in section 1 and this data was retrieved from open data sources which was then cleaned to retrieve the time period and factors needed.

After the initial analysis, a geospatial analysis was done to analyse the spatial distribution of the socio-economic, mental health factors and ambulance calls. This analysis was performed by plotting the various choropleth distribution for The Hague based on each variable.

This segment will conclude with the discussion on the limitations of the data and the study.

3.1 Data Retrieval

Three sources of data were used for the years **2017**, **2018** and **2019** for the study.

1. **CBS data:** the indicators mentioned in section 1 were all retrieved from the public database of CBS. This data is open source and provided by the Dutch government, hence proves its reliability and robustness. The dataset was filtered to only include entries from The Hague.
2. **Emergency Calls data:** the dataset included all emergency calls of The Netherlands from January 2017 to September 2020. This dataset included the date, time, location of the calls and which emergency service was called. The dataset was filtered to all ambulance calls in The Hague only. This dataset is not open source and was retrieved through permission from the course coordinator, Trivik Verma.
3. **The Hague Neighbourhood ShapeFiles**: the shapefiles were provided by the course coordinator as well for course assignments. These shape files enabled us to create spatial visualisations of the data and indicators that could aid in our analysis and modelling.

3.2 Data Cleaning

Data cleaning was done primarily using the pandas package in Python on the CBS data and Ambulance Calls data. Some basic analyses were done on the datasets to look for outliers or missing data before deciding on what preparations were needed.

CBS Data

All numerical data was typecast into a float to prepare the data for modelling. The CBS dataset was cleaned to only include data from The Hague. The dataset contains missing data from several neighbourhoods that were consistent throughout the years. It is interesting to note that these neighbourhoods have little to no inhabitants and only account for 7% of the dataset therefore these neighbourhoods will be dropped for the analysis. The neighbourhoods are: Oostduinen, Vliegeniersbuurt, Tedingerbuurt and De Rivieren.

The columns representing the indicators are:

1. 'low_income_perc': Percentage of low-income households
2. 'dist_gp' : Average distance to the nearest general practitioner (GP)
3. 'single_hh_perc': Percentage of single-person households
4. 'social_security_perc': Number of people requiring social assistance

The social_security_perc indicator consist of the aggregation of the following mentioned indicators from the CBS dataset:

- Number of people receiving social assistance benefit / Number of inhabitants in a neighbourhood
- Number of people receiving disability benefit / Number of inhabitants in a neighbourhood
- Number of people receiving unemployment benefit / Number of inhabitants in a neighbourhood
- Number of people receiving pension / Number of inhabitants in a neighbourhood

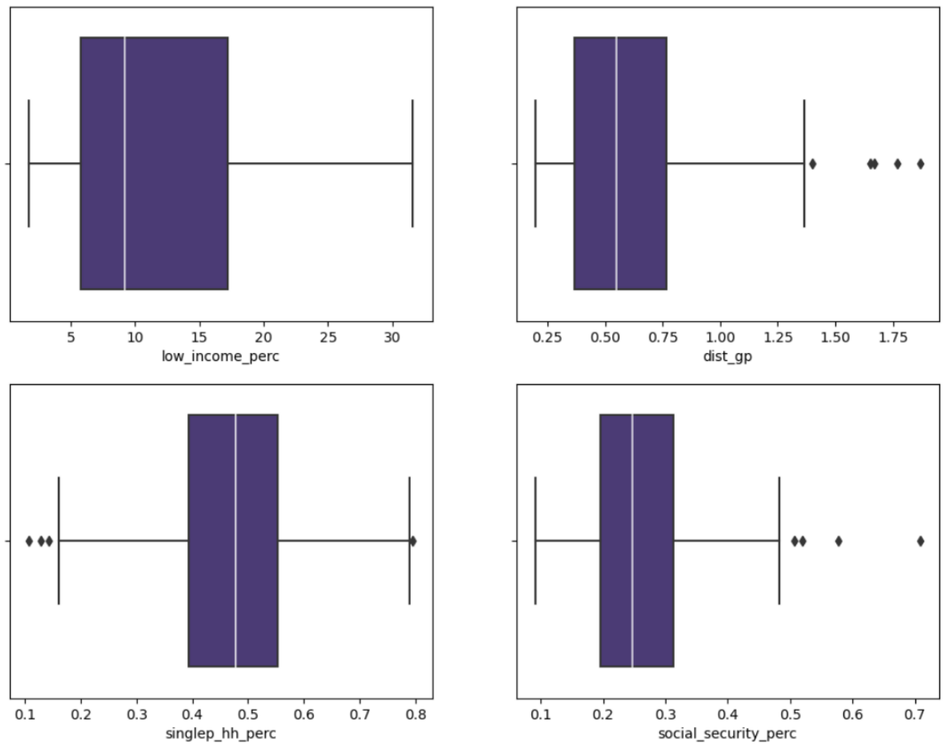


Figure 1: Boxplot of indicators

We then take a look at the breakdown of indicators in the dataset. It can be observed that the indicators do not have many outliers (**Fig 1**). The low_income_perc indicator has a lower quartile of around 5 while an upper quartile of around 17. The dist_gp indicator has a lower quartile of around 0.30 while an upper quartile of around 0.75, with several outliers. This data indicates that most households have a short distance to their GPs. The singlep_hh_perc indicator has a lower quartile of around 0.4 while an upper quartile of around 0.6. The social_security_perc indicator has a lower quartile of around 0.2 while an upper quartile of around 0.3, indicating that the population of people under social benefit is not high.

Emergency Calls Data

As the study is done on neighbourhoods in The Hague, each ambulance call data needed to be mapped to the neighbourhood the call was made in. Only the exact longitude and latitude was included in the dataset, thus a geometry Point object was made of these two points and joined with The Hague Neighbourhood ShapeFiles dataset to retrieve the neighbourhood of each call.

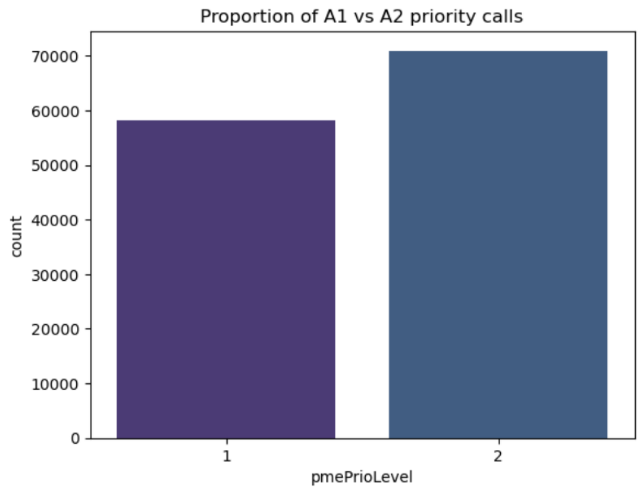


Figure 2: Barplot of priority calls

The pmePrioLevel variable in this dataset measures the conversion of priority to a normalised level. We observe that class 2 has approximately 10,000 more calls than class 1 (**Fig 2**).

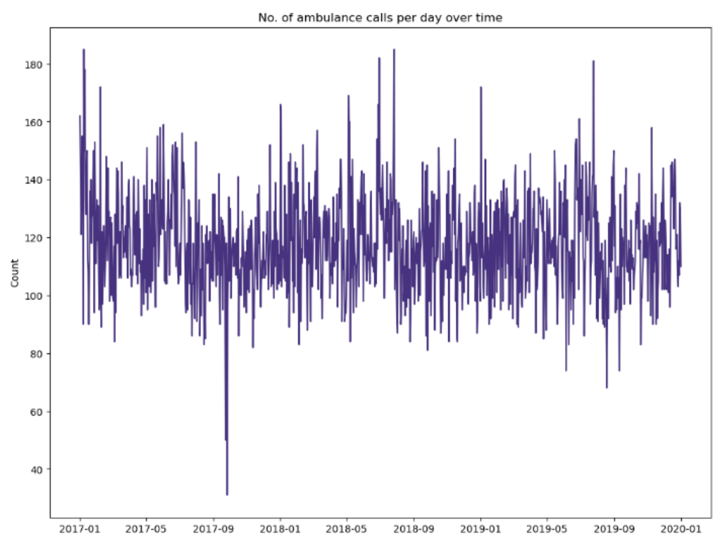


Figure 3: Plot of ambulance calls over time

When studying the frequency of ambulance calls across a time period, a time trend analysis is vital. It can be observed that throughout the years the majority of calls fall between 80 to 160 per day (**Fig 3**).

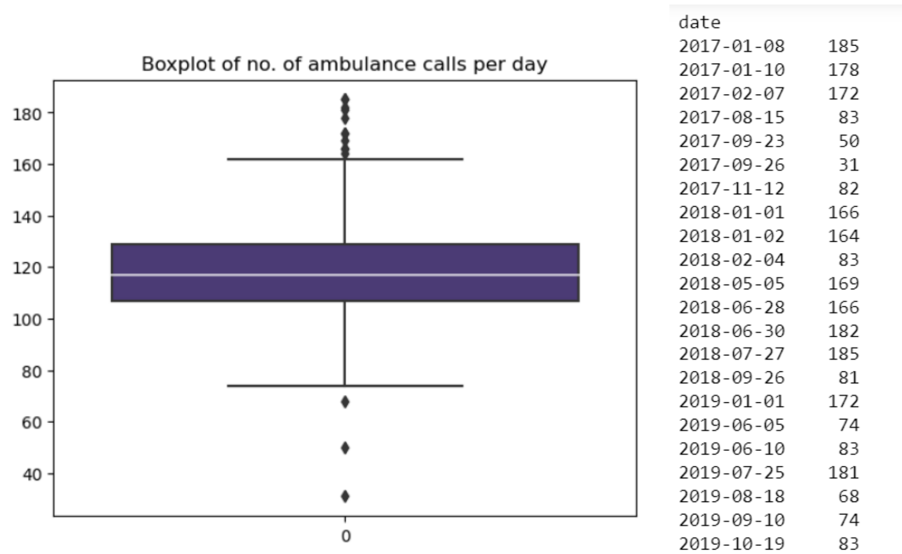


Figure 4: Boxplot of number of ambulance calls per day and list of outliers

A boxplot of the number of calls per day was generated and many outliers were observed (**Fig 4**). These outliers were printed out to observe any trends and to decide how to handle the outliers for our analyses.

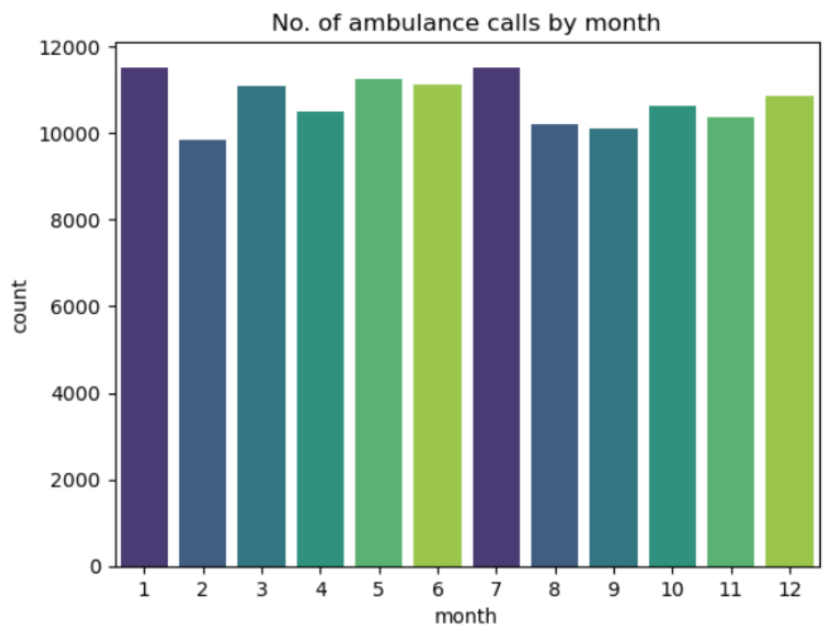


Figure 5: Bar plot of number of ambulance calls by month

Finally, we want to observe if there are any trends across the different months or seasons. We see that the number of calls per month do not have a significant difference, which leads us to conclude that the time of the year is a weak factor with regards to the number of ambulance calls being made (**Fig 5)**.

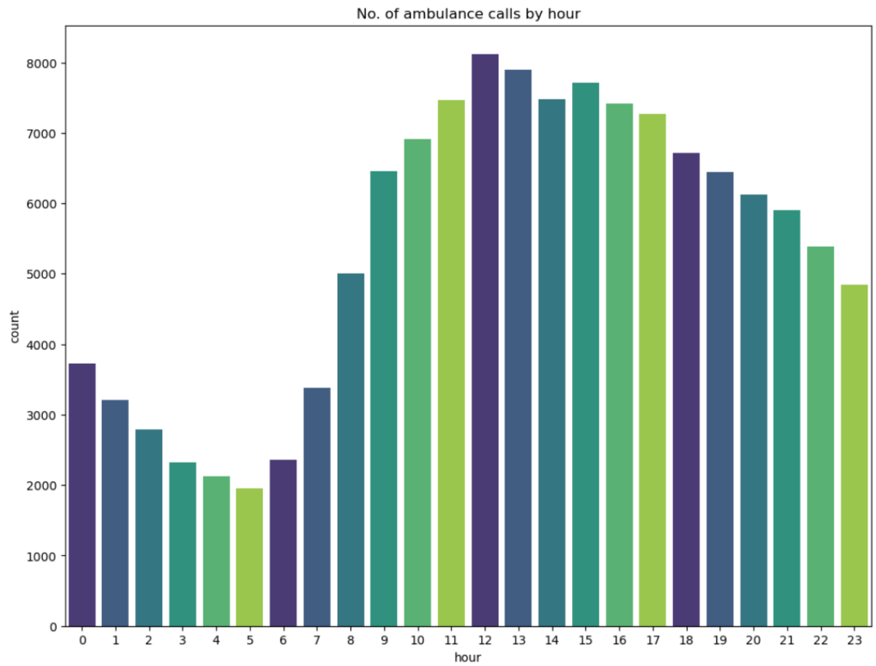


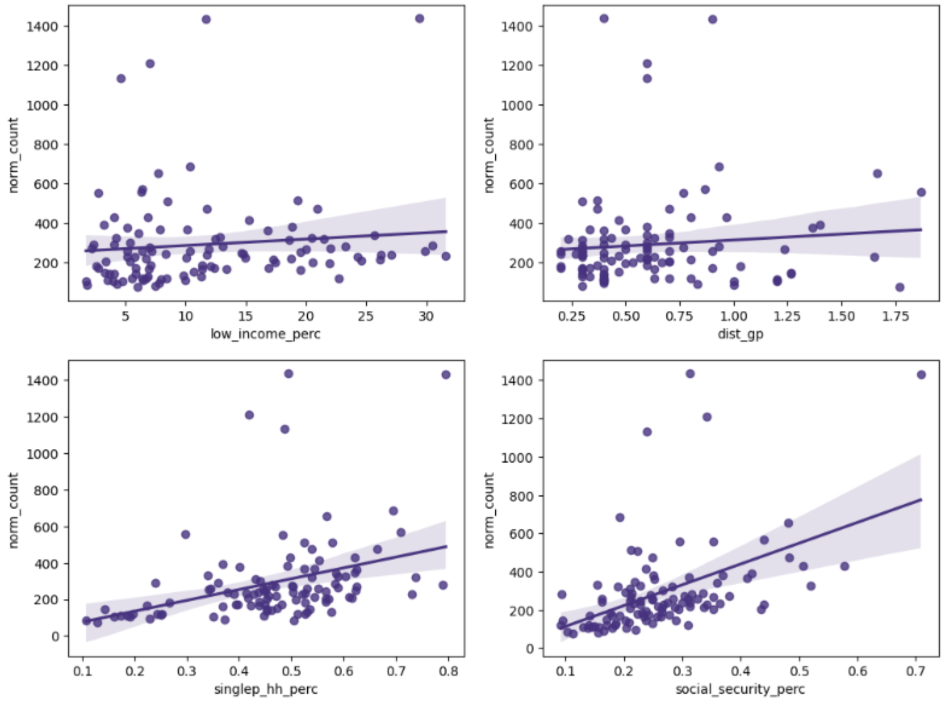
Figure 6: Bar plot of number of ambulance calls by hour

It can be observed that ambulance calls across the time period studied appear to happen more often in the afternoon to evening, with significantly fewer calls in the early morning hours (**Fig 6**). This is interesting to note as it would be expected that over time there would not be significant differences in when emergency services are needed as health ailments do happen randomly depending on person and situation.

3.3 Investigating Relationships

After cleaning the datasets and merging the CBS set with the Ambulance Calls data, exploratory data analysis can be performed on the indicators. For each of the indicators, the correlation between the indicator and the number of ambulance calls was analysed and recorded as scatterplots and from the scale of -1 to 1. A normalised count of the ambulance calls per neighbourhood was created for the purpose of this scatter plot analysis. This was done as ambulance call count depends on the population, as some neighbourhoods may have fewer residents and some may have a significantly larger population.

It can be observed that the percentage of low income households and the average distance to a GP both have the lowest positive and negative correlation with the number of ambulance calls respectively (**Fig 7, 8**). The percentage of the population receiving social benefits has a clear linear correlation with the number of ambulance calls while the percentage of single persons households have a weak linear correlation with the number of ambulance calls. As such, we will focus on the percentage of the population receiving social benefits and the percentage of single persons households in our analysis.




```
Correlation coefficient between percentage of low income households and no. of ambulance calls: 0.10411727276757513
Correlation coefficient between average distance to GP and no. of ambulance calls: 0.0908795076602941
Correlation coefficient between percentage of single person household and no. of ambulance calls: 0.3570575006827822
Correlation coefficient between percentage of population receiving social benefits and no. of ambulance calls: 0.48205894523340
365
```

Figure 7: Scatterplots of all indicators and correlation coefficient of all indicators

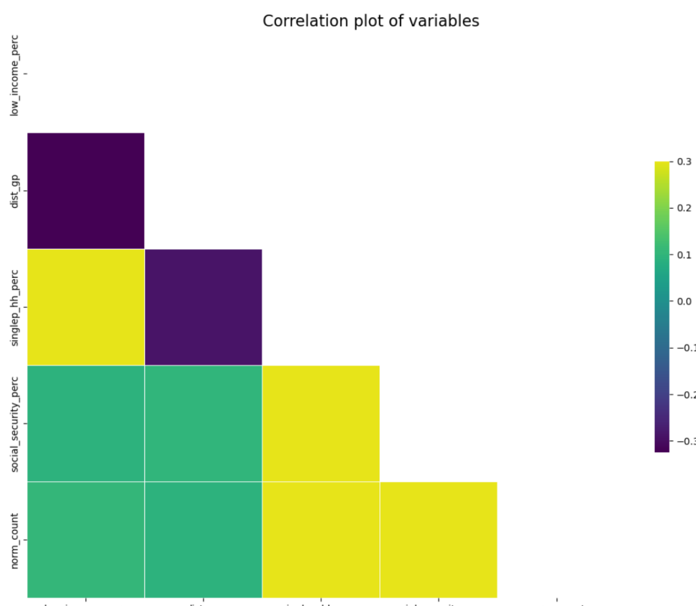


Figure 8: Correlation matrix of all indicators

It is interesting to observe that with a larger population under social security, the number of ambulance calls increases. It would be expected that this population may be on the poorer end in terms of income thus requiring social security, thus being less able to afford healthcare. It is also interesting that neighbourhoods with higher percentages of single person households experience more ambulance calls. This could be related to mental health issues that can cause emergency situations.

The positively correlated indicators used will be the percentage of low income households and single person households.

3.4 Spatial Analysis

In this section the geo-spatial distribution of the indicators and number of ambulance calls will be analysed. This is done through choropleths created by the shapefiles of The Hague neighbourhoods. The purple colour indicates the neighbourhoods with the highest values while the yellow colour indicates neighbourhoods with the lowest values. This colour scheme was used as it is colour-blind friendly. For clearer images of choropleths, please refer to Appendices. To identify neighbourhoods in our analysis, we also referred to a visual map of The Hague's neighbourhoods which can be found under Appendices as well.

The choropleth for the percentage of low income households show that neighbourhoods around South of The Hague experience a larger proportion of low income households (**Fig 9**). When we compare this to the choropleth of the high number of ambulance calls, it cannot be observed that this area of neighbourhoods experience higher or lower number of ambulance calls. The two choropleths do not have many similarities, which could suggest that the percentage of low income households is not very useful as an indicator to the number of ambulance calls the neighbourhood might have.

The choropleth of the percentage of single person households in the neighbourhood do show the majority of The Hague in a purple zone (**Fig 10**). However, Zuiderpark and Binckhorst have the highest percentages of single persons households. When we compare this with the choropleth of ambulance calls, Zuiderpark has the most ambulance calls as well, though Binckhorst does not have an exceptionally high number of ambulance calls. This seems to suggest that single person households could be experiencing more instances of emergency health issues. This is interesting as we hypothesised that mental health issues like loneliness could be more prevalent in single person households which could potentially lead to life threatening situations that can perhaps explain the high frequency of ambulance calls.

Figure 9: Choropleth of low income households

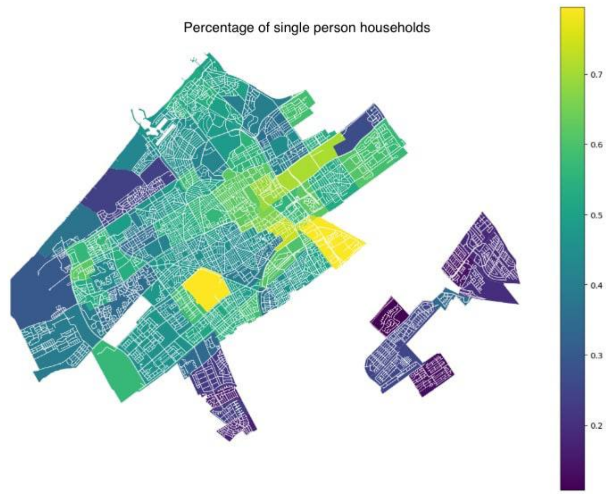
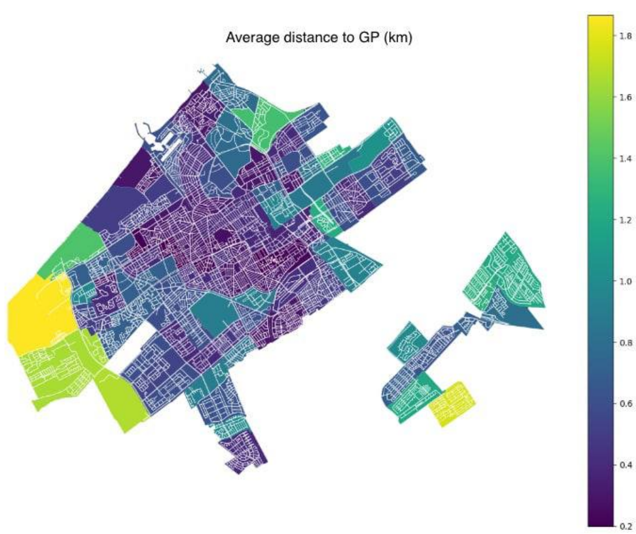


Figure 10: Choropleth of single households

The choropleth for average distance to GP in km is shown below. As we have mentioned in the previous section 3.3, there is almost no linear correlation between average distance to GP and the number of ambulance calls being made. On the right is the plot of all healthcare facilities within The Hague. It can be observed that there are more healthcare facilities clustered around the centre of the Hague and these facilities tend to appear more spread out around the periphery of The Hague. This corresponds to the information on the choropleth. However, when we compare this information to the choropleth of the ambulance calls, we do not see any similarities as the number of ambulance calls do not follow the trend of significant difference between the centre and the periphery of the Hague. Therefore, this seems to suggest to us that this indicator may not be as useful in helping us predict the number of ambulance calls in a neighbourhood.



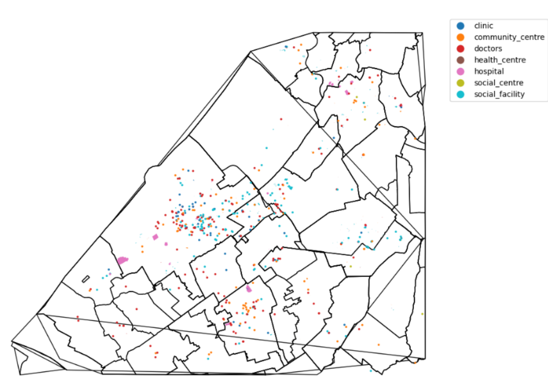


Figure 11: Choropleth of average distance to GP *Figure 12: Visualisation of healthcare facilities*

Finally, from the choropleth of the percentage of households with social benefits, it can be observed that the choropleth has the most similar pattern with that of the number of ambulance calls (**Fig 13**). It can be observed that Zuiderpark has the highest percentage of households under social benefits while it also experiences the highest number of ambulance calls. The centre of the Hague appears to have more households under social benefits compared to the neighbourhoods in the periphery of the Hague. This is consistent with the number of ambulance calls being higher in the centre and fewer the further away from the centre (**Fig 14**). This could suggest that people under social benefits could have more health issues which could prevent them from working and thus lowering their ability to spend and also requiring medical assistance more often.

Figure 13: Choropleth of households with social benefits

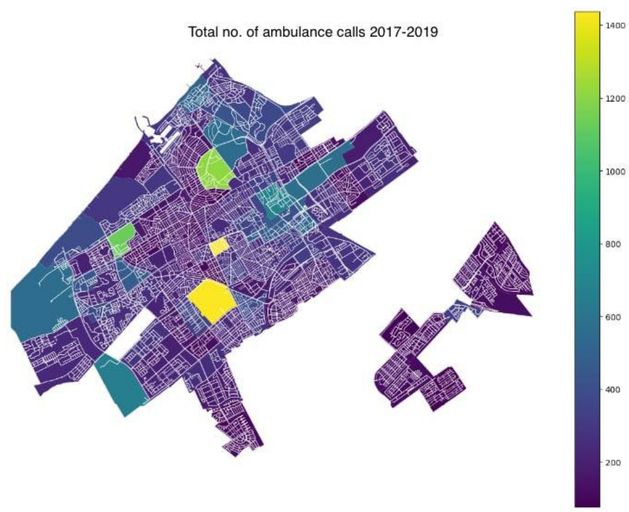


Figure 14: Choropleth of ambulance calls

3.5 Spatial Autocorrelation Analysis

In this section we will perform Local Spatial Autocorrelation to observe the overall geographical pattern present in the data. The moran plot assists with the visualisation of spatial data to explore the nature and strength of spatial autocorrelation. The variables of interest, in this case % single person households and % households under social benefit, are standardised and a moran plot is created.

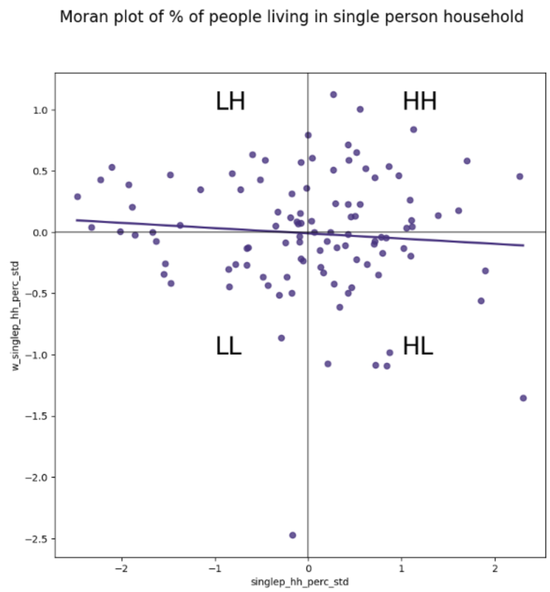


Figure 15: Moran plot of single person household

Figure 16: LISA map of single person households

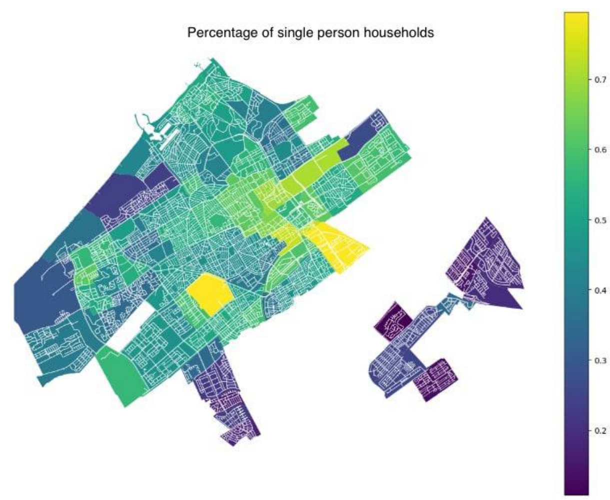


Figure 17: Choropleth of single person households

Through local spatial autocorrelation, a LISA cluster map is built. The purple part of the map shows us the neighbourhood with a high percentage of single person households surrounded by other neighbourhoods with high percentages of single person households (**Fig 16**). Comparing this LISA map to the choropleth previously discussed, the LISA map helps to identify areas that might have interesting observations that are not significant on the choropleth (**Fig 17**). We can also zoom into Zuiderpark, which is highlighted in green on the LISA map. This tells us what we already know, that Zuiderpark has an exceptionally high percentage of single person households but is surrounded by neighbourhoods that do not share the same characteristics.

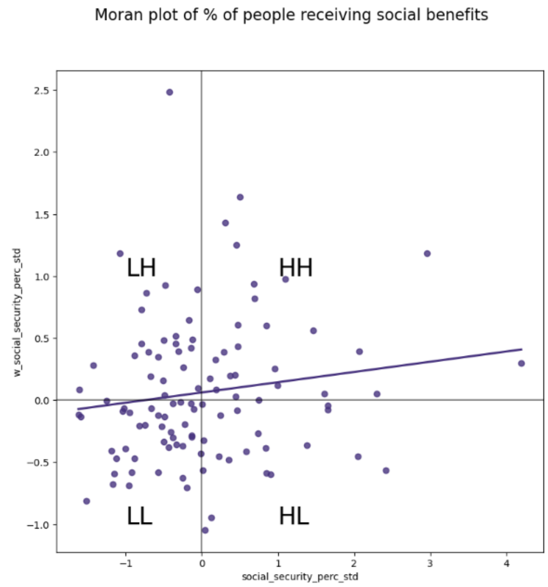


Figure 18: Moran plot of households with social benefits

Figure 19: LISA map of households with social benefits *Figure 20: Choropleth of households with social benefits*

From the above LISA map plot, the purple parts of the map shows us the neighbourhood with a high percentage of households under social benefit surrounded by other neighbourhoods with high percentages of households under social benefit (**Fig 19**). The yellow part shows us those with low percentages and surrounded by other low percentages. It is interesting to note that these neighbourhoods highlighted in purple or yellow do not have extreme values in the choropleths (**Fig 20**). This can be explained by the correlation values. The correlation values for this indicator is generally low, around 0.4, thus when LISA takes into account standardising correlations, these neighbourhoods were highlighted despite not having the highest or lowest percentages. It is still useful in our analysis to identify neighbourhoods of interest.

3.6 Limitations

- Missing values: For all 3 years, there was missing data for all the indicators for 4 neighbourhoods. These neighbourhoods were removed from the dataset when it was found that there was no data for other years as well, thus we could perform any model to estimate a value for the required years.
- Empty neighbourhoods: Some neighbourhoods do not have residents but there are still ambulance calls present in the dataset. Therefore, the indicators we chose would not be able to predict the amount of ambulance calls in these neighbourhoods.
- The indicators chosen were measured from the residents who live in the neighbourhoods. However, ambulance calls could be made by visitors or non-residents that we do not take account of. In some neighbourhoods, emergencies may be prevalent due to events or celebrations with large crowds that may cause the number of ambulance calls over the years to be high. These factors are not related to the socio-economic status of the residents living there, thus might affect the interpretation of the results.
- We only performed regression but other models could be more appropriate to predict ambulance calls but we could not attempt this with the time frame of the project.

4. Analysis

The preliminary study on the existing relationships in the dataset, including correlations between each of the factors and the number of ambulance calls, and the spatial analysis led us to infer that some variables had a significant influence on the number of ambulance calls. In the regressions we performed, we included:

- The percentage of low-income households
- The average distance of each household to their nearest GP
- The percentage of single-person households
- The percentage of people receiving social benefits

We first split the data set into two subsets : train_data and test_data using the train_test_split function of sklearn and perform an OLS regression on the train_data dataset. Moreover, we normal-standardised the indicators so that we can compare the coefficients across indicators to get their relative importance in predicting ambulance calls.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             norm_count   R-squared:                       0.364
Model:                            OLS   Adj. R-squared:                  0.331
Method:                 Least Squares   F-statistic:                     11.28
Date:                Wed, 09 Nov 2022   Prob (F-statistic):           2.72e-07
Time:                        18:50:18   Log-Likelihood:                -547.78
No. Observations:                  84   AIC:                             1106.
Df Residuals:                      79   BIC:                             1118.
Df Model:                           4
Covariance Type:            nonrobust
========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                 279.2736     18.758     14.888      0.000     241.936     316.611
low_income_perc       -21.7988     20.648     -1.056      0.294     -62.898      19.301
dist_gp                30.7769     19.655      1.566      0.121      -8.346      69.899
singlep_hh_perc        68.1497     24.356      2.798      0.006      19.671     116.628
social_security_perc   76.2897     20.613      3.701      0.000      35.260     117.320
==============================================================================
Omnibus:                       70.593   Durbin-Watson:                   2.355
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              489.064
Skew:                           2.585   Prob(JB):                    6.33e-107
Kurtosis:                      13.631   Cond. No.                         2.02
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 21: Multilinear regression with all hypothesised variables

The coefficient analysis **(Fig 21)** clearly emphasises that the percentage of low-income households and the GP distance have a non-significant impact compared to the two others features, considering the difference of order of magnitude. The low impact of GP distance is evident in the spatial analysis section- the concentration of medical facilities in the centre compared to other districts had no visible correlation with the number of ambulance calls.

The percentage of people benefiting from social benefit seems to have the greatest impact on the number of ambulance calls. We can link this observation with our previous spatial analysis on Zuiderpark, as the district has the characteristics to host a high percentage of people under social benefit and to experience an exceptionally high number of ambulance calls each year. However, if this relation works well on Zuiderpark, it does not work anymore for the second district with the highest number of calls considering the low number of people under social benefits. Before further analysis on the R-squared and fit test on test_data, we get a hint that our model will fail to explain the number of ambulance calls as a linear function of the selected features.

```
lr_model = LinearRegression().fit(X_train, y_train)
lr_model.score(X_test, y_test)
```

```
0.09680937349160978
```

Figure 22: R-squared value on test data

We notice a low R-squared value meaning that our features fail to fully explain the variation of the number of ambulance calls in The Hague **(Fig 22)**. The weakness of our model on the train_data can be seen in the correlation test performed on X_test and Y_test, as we obtain a score of 0.097.

4.1 Removal of Features

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             norm_count   R-squared:                       0.326
Model:                            OLS   Adj. R-squared:                  0.309
Method:                 Least Squares   F-statistic:                     19.55
Date:                Wed, 09 Nov 2022   Prob (F-statistic):           1.18e-07
Time:                        18:50:18   Log-Likelihood:                -550.22
No. Observations:                  84   AIC:                             1106.
Df Residuals:                      81   BIC:                             1114.
Df Model:                           2
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  279.3194     19.006     14.696      0.000     241.503     317.136
singlep_hh_perc         46.8132     22.406      2.089      0.040       2.232      91.394
social_security_perc    88.4696     20.154      4.390      0.000      48.369     128.571
==============================================================================
Omnibus:                       67.633   Durbin-Watson:                   2.230
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              403.485
Skew:                           2.520   Prob(JB):                     2.42e-88
Kurtosis:                      12.480   Cond. No.                         1.67
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 23: Multilinear regression with limited hypothesised variables

Following this analysis, we tried to refine the model by dropping the features with low correlation, namely: percentage of households with low income and average GP distance. We notice in this new model that dropping these two variables slightly reduces the R-squared on the training data from 0.364 to 0.326, which confirms the low importance of these variables **(Fig 23)**.

```
lr_model = LinearRegression().fit(X_train_sub, y_train)
lr_model.score(X_test_sub, y_test)
```

```
0.1471310087793326
```

Figure 24: R-squared value on test data

When we try to fit the test_data using this new model, we get a slightly better score of 0.147 > 0.097 **(Figure 24)**. Even if the model has been improved, it is still insufficient to use these features alone to predict reliably the number of ambulance calls in a specific area.

As already stated before, some important explanatory features could have been missing in the dataset leading to the weakness of the model. Furthermore, as our dataset is rather small, it is possible that the non-homogeneity of the train and test dataset has led to a large difference in model performance (R-squared of 0.326 vs 0.147).

5. Discussion & Conclusion

5.1 Limitations

Before concluding on our findings, we can highlight the limitations of our model.

First, on the quality of the data itself, we did only drop 7% of the dataset in neighbourhoods with a low number of inhabitants which is reasonable. The dataset is nevertheless rather small, which could lead to difficulties in finding statistical relations. We highlighted for instance the non-homogeneity between the trained data and the test data that could explain the significant difference between the R-squared for the same model.

The dataset may also omit some important features having a high impact on R-squared in our linear regression model. For instance, it would have been reasonable to think that the number of car accidents in a neighbourhood could have an impact on the number of ambulance calls, but as stated before, the hypothesis focuses exclusively on socio-economic factors.

We only tested a linear regression model to perform our analysis but other models could be considered to fit better the number of calls as a function of our features.

Thus, the following explanation for findings have to be interpreted while keeping in mind these limitations and the non-accurate results we obtained.

5.2 Possible Explanation for Findings

The percentage of low-income households and the average distance between a household and the nearest GP should not be used to predict the number of ambulance calls in a neighbourhood. However, the percentage of the population under social benefit and the percentage of single-person households seem to be relatively better predictors.

The finding that having a larger population under social benefit correlates with a larger number of ambulance calls is likely due to these individuals being less able to take good care of their health in general. Research suggests that firstly, individuals under social benefit are likely to already have had pre-existing health conditions, leading to job loss and hence allowing them to qualify for social benefit programmes. Secondly, individuals with psychological issues tend to experience greater socioeconomic disadvantage, leading to them being under social benefit (Shahidi et al., 2019). With worse overall health, they are more likely to have emergency health situations, contributing to a larger number of ambulance calls in their neighbourhood.

Next, neighbourhoods with higher percentages of single person households experiencing more ambulance calls could be due to mental health issues caused by loneliness, such as depression due to living alone (Jacob et al., 2019), causing poorer mental and physical health as well. This may lead to a greater number of life-threatening situations. As we have discovered before, the number of suicide cases rise year-on-year in the Netherlands. This is strongly linked to mental health and should be something to be taken more seriously. Not only are these people likely to require immediate medical attention that may strain ambulance resources, these people will always be at the risk of harming themselves.

5.3 Implications of Findings

Our findings suggest that the percentage of population under social benefit and the percentage of single person households could be included in the prediction of the number of ambulance calls – while these factors alone are not sufficient, it is possible that they increase the prediction accuracy. This further indicates that socio-economic factors could be taken into account when predicting ambulance call patterns across neighbourhoods. However, the factors need to be chosen carefully since not all socio-economic factors are good predictors of the number of ambulance calls, as per our findings.

It was expected that the average GP distance for a household is not a good indicator. As highlighted in **Figure 1** in section 3.2, most households have a short distance to the nearest GP in the dataset used, hence not much variation can be observed for this feature. This is perhaps why it is not a good indicator for the number of ambulance calls.

However, it is interesting to note that the percentage of low-income households is not a good indicator. Typically, individuals with low income are less able to take care of their well-being, due to poor living conditions (Cunningham, 2018) which might lead to emergency health situations. In The Netherlands, Dutch citizens are also required by the government to have health insurance. However, the cost to patients has been increasing faster than income over the years and this has led to the decreasing affordability of healthcare services (Scott, 2020). This might result in increased hesitance for the low income to use healthcare services, even during emergencies. This demonstrates why considering the percentage of low-income households may not be suitable, since other factors affect the likelihood of these households using ambulance services.

5.4 Conclusion

In conclusion, some socio-economic and mental health factors can be used to improve the prediction of the number of ambulance calls in neighbourhoods of The Hague. The RAV of The Hague will be able to better understand how ambulances should be distributed across the city, allowing for more efficient escorting of patients and to better deal with emergency medical situations. To take it a step further, they may also be able to provide insight into issues that need to be tackled by policymakers or the government body in charge of the city, such as providing more support for those living alone through neighbourhood-wide social activities to prevent loneliness, which may allow them to have better physical and mental health in the long run and thus reduce the number of emergency health situations in an area.

Some of the above-mentioned features, while increasing the predictive capability, are not strongly indicative of ambulance demand. However, this could be due to the limitations discussed, especially the need to include other factors when investigating ambulance demand. In order to attain a good predictive model that allows ambulances to be distributed more efficiently in The Hague, more research has to be done on this topic with larger amounts of data that can be studied.

6. References

Agarwal, G., Lee, J., McLeod, B., Mahmuda, S., Howard, M., Cockrell, K., & Angeles, R. (2019). Social factors in frequent callers: A description of isolation, poverty and quality of life in those calling emergency medical services frequently. *BMC Public Health*, *19*(1), 1–8. https://doi.org/10.1186/s12889-019-6964-1

Ambulancezorg Nederland. (2019). *Ambulance Care Sector Overview*. https://www.ambulancezorg.nl/en/ambulance-care-sector-overview/facts-figures-2019

American Psychological Association. (2014). Children, youth, families & socioeconomic status. *American Psychological Association*. https://www.apa.org/pi/ses/resources/publications/children-families

CBS. (2022). deaths; suicide (residents), various themes. https://opendata.cbs.nl/#/CBS/en/dataset/7022eng/table

CBS. (2022). Mental health has worsened among young people. CBS. https://www.cbs.nl/en-gb/news/2022/22/mental-health-has-worsened-among-young-people

CSDH. (2015). *How Do Neighborhood Conditions Shape Health?* 1–16. http://www.buildhealthyplaces.org/content/uploads/2015/09/How-Do-Neighborhood-Conditions-Shape-Health.pdf

Cunningham, P. J. (2018). Why Even Healthy Low-Income People Have Greater Health Risks Than Higher-Income People. *The Commonwealth Fund*. https://www.commonwealthfund.org/blog/2018/healthy-low-income-people-greater-health-risks

Gottliebformat, Laura Braveman, P. (2014). The Social Determinants of Health: It's Time to Consider the Causes of the Causes. *Public Health Reports*, *129*, 19–31. https://doi.org/10.1177/00333549141291S206

Jacob, L., Haro, J. M., & Koyanagi, A. (2019). Relationship between living alone and common mental disorders in the 1993, 2000 and 2007 National Psychiatric Morbidity Surveys. *PLoS ONE*, *14*(5), 1–12. https://doi.org/10.1371/journal.pone.0215182

Kommer G.J. et al. (2020). *Doorontwikkeling referentiekader ambulancezorg 2020*. www.rivm.nl

Minderhout, R. N., Baksteen, M. C., Numans, M. E., Bruijnzeels, M. A., & Vos, H. M. M. (2021). Effect of COVID-19 on health system integration in the Netherlands: a mixed-methods study. *Journal of the American College of Emergency Physicians Open*, *2*(3). https://doi.org/10.1002/emp2.12433

Nehme, E., Nehme, Z., Cameron, P. A., & Smith, K. (n.d.). Drivers of Increasing Emergency Ambulance Demand. *Prehospital Emergency Care*, *24*(3), 1–11. https://doi.org/http://dx.doi.org/10.1080/10903127.2019.1635670

Portz, K., Newell, R., & Archibong, U. (2013). Rising Ambulance Life-Threatening Call Demand in High and Low Socioeconomic Areas. *Journal of Psychological Issues in Organizational Culture*, *3*(S1), 306–320. https://doi.org/10.1002/jpoc.21095

Scott, D. (2020). The Netherlands has universal health insurance — and it's all private. *Vox*. https://www.vox.com/platform/amp/policy-and-politics/2020/1/17/21046874/netherlands-universal-health-insurance-private

Shahidi, F. V, Ramraj, C., Sod-erdene, O., Hildebrand, V., & Siddiqi, A. (2019). *The impact of social assistance programs on population health : a systematic review of research in high-income countries*. 1–11.

Siler, K. F. (1975). Predicting demand for publicly dispatched ambulances in a metropolitan area. *Health Services Research*, *10*(3), 254–263. http://www.ncbi.nlm.nih.gov/pubmed/1225867%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1071860

Wang, J., & Geng, L. (2019). Effects of socioeconomic status on physical and psychological health: Lifestyle as a mediator. *International Journal of Environmental Research and Public Health*, *16*(2). https://doi.org/10.3390/ijerph16020281

Appendices

[Link to Notebook]
[Link to HTML]
Ambulance call dataset fields

pmeId - Related to the id on the website

pmeTimeStamp - Time of call registered

pmeProtocol1 - P2000 protocol related info

pmeProtocol2 - P2000 protocol related info

pmeTarget - P2000 protocol related info

pmeMessage - Original message

pmePrio - Priority of the message

pmePrioLevel - Conversion of priority to a normalized level

pmeDienst - A=Ambulance, B=Brandweer, P=Politie

pmeStrippedMessage - Cleaned Message of the call

pmeStraat - Street

pmeHouseNumber - Mid-point of the street closest to location of call (for data privacy)

pmeHouseNumberAdd - Addition to house number

pmeANRoad = A=highway, N=Provincial Way

pmeHectometerpaal = Location on A or N way

pmeZip - Zip code of the area

pmeRegionName - Name of region

pmeLatitude = wgs84

pmeLongitude = wgs84

pmeHash - Ignore internal indexing (for collection purposes only)

pmeGeoAccuracy = H=Housenumber, S=Street

pme_strId = Street Id

pme_wplId = Woonplaats Id

pme_gemId = Gemeente Id

pme_proId = Province Id

pme_vrgId = Veiligheidsregio Id

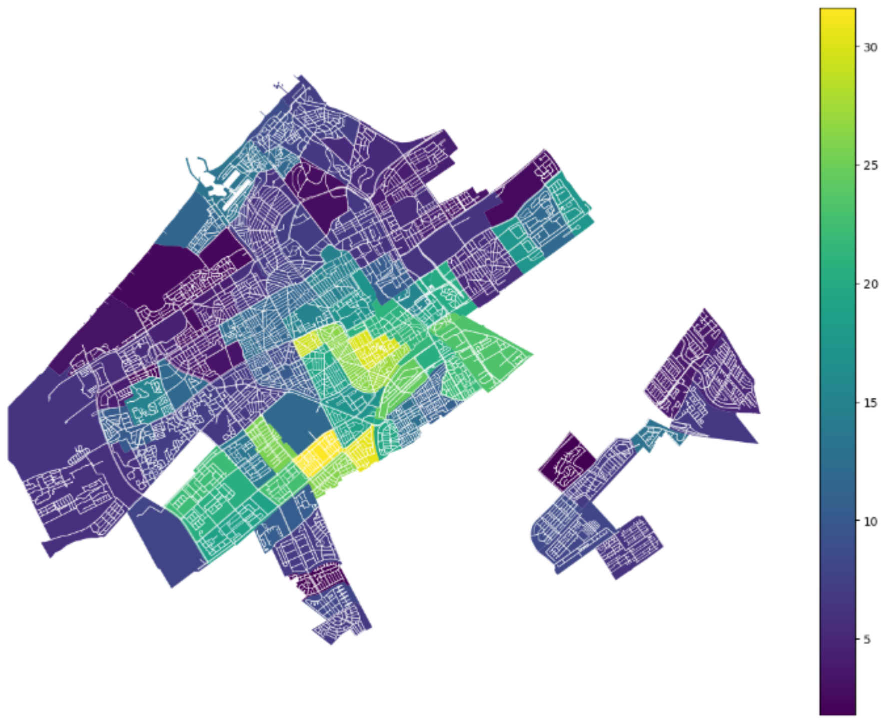
pmeCapCodes = Capcodes

pmeLifeLiner - True/False (Extraction by air) wplName = Woonplaats

gemName = Gemeente / Municipality

Choropleths

Percentage of low income households



Average distance to GP (km)

Percentage of single person households

Percentage of households with social benefits

Total no. of ambulance calls 2017-2019

Map reference for neighbourhood names