**Hypothesis Testing:**

**Visualising the Correlations and Simple Linear Regressions between Environmental Variables and Health:**

*Note: A detailed step-by-step iPython notebook for these visualisations is available in our GitHub Repository:*

[*https://github.com/Ftracy/QMGroup22/blob/c038376844095d25d3450ad5665a179da7ee3c77/QMM2%20Environmental%20Data%20and%20Health.ipynb*](https://github.com/Ftracy/QMGroup22/blob/c038376844095d25d3450ad5665a179da7ee3c77/QMM2%20Environmental%20Data%20and%20Health.ipynb)

*Sources: Waskom, a, 2022; Waskom, b, 2022*

**Correlation between Environmental Variables and Health**

[**Figure 1: A Heatmap to Show the Correlation Coefficients between Environmental Variables and Health]**

Background pattern

Description automatically generated

[**Figure 2: A Pairplot to Visualise the Correlations between Environmental Variables and Health]**

A picture containing text, shoji, crossword puzzle, building

Description automatically generated

**Explanation for both figures:**

The heat map shows the Pearson Correlation Coefficient between the 3 environmental variables we have chosen - green space, air pollution, and noise pollution - and health status as a whole. How health status is calculated has been outlined in our methodology section. The further away a value is from 0, the stronger the relationship between two variables. The pairplot visualises the correlations between the aforementioned environmental variables and health. The diagonals are histograms to show the distribution of data.

As we can observe from these two visualisations, none of the individual variables related to environmental factors have a great impact on the health score in the London boroughs. The one with the highest correlation is the air pollution score (-0.4). This correlation is negative, suggesting that the less air pollution in a borough, the poorer its health. This can be seen in boroughs such as City of London, which is the one with both the highest health score and the highest air pollution score. The figures further indicate that there is a positive correlation between green space score and health, and noise pollution score and health. However, these correlations are very weak, since their correlation coefficient values, 0.037 and 0.095 respectively, are both close to 0.

**Correlation between BSe and BSh**

*Note: A detailed step-by-step iPython notebook for this visualisation is available in our GitHub Repository:*

[*https://github.com/Ftracy/QMGroup22/blob/8f815d891bb9f3a6c730655aceb4086e96ea71f7/BSh%20and%20Bse%20Data%20Regression.ipynb*](https://github.com/Ftracy/QMGroup22/blob/8f815d891bb9f3a6c730655aceb4086e96ea71f7/BSh%20and%20Bse%20Data%20Regression.ipynb)

*Sources: Waskom,c, 2022*

**Figure 3: A Linear regression with Marginal Distributions to Visualise the Correlation between BSe and BSh**

**Chart, scatter chart

Description automatically generated**

R Squared value (calculation made in Microsoft Excel): 0.0351

[**Explanation:** This visualisation consists of two elements: A linear regression between the variables BSe and BSh, and the marginal distribution of the variables contained in each subset. The regression shows that there is a negative correlation between BSh and Bse - as Bse increases, BSh decreases. Therefore, according to this visualisation and the factors and boroughs studied, the less environmentally friendly a borough is, the better its overall health outcome is. This can be seen by using City of London as an example. Indeed, City of London has the lowest borough score for all health factors (BSh), indicating poorest health, and the highest borough score for environmental factors (BSe), indicating it is the worst borough environmentally, using our considered variables. However, considering the low R Squared value means that less than 5% of the data points is represented by the model, therefore the relationship between both variables is not explained by the linear regression. ]

**Multi-linear Regression on Environmental Factors and Health**

**Figure 4: A Multi-linear Regression on Environmental Factors and BSh**

Table

Description automatically generated

The data for hypothesis testing comes from the raw data collected on the 33 boroughs after the normalisation process outlined in the methodology section.

A multi-linear regression (MLR) was conducted to statistically analyse the accumulative effect of the three environmental variables outlined on health (BSh). This was done using Google Collab, which has a built-in MLR model.

The R squared value is 0.162, meaning that approximately 16% of the data points fit in the regression model, which indicates the noise of data and suggests that linear models might not be the best model to simulate the correlation between overall environmental factors and health.

If we set the significance value at 0.05, then the p-value is higher than this value, indicating the acceptance of Hypothesis 1 (and subsequent rejection of Null Hypothesis 1) that there is no correlation between overall environmental factors and health.

**3 independent variables on health:** Whilst the p-values for  green space score (0.775) and noise pollution score (0.805) are above the significance level (0.05), the p-value for air pollution score (0.029) is below the significance level, indicating it is the only factor that has a correlation with overall health status.

*Note that the interpretation of these coefficients is not significant in this research as we calculated scores through the normalisation of our datasets.*

**In conclusion,** there is virtually no correlation between BSe and BSh. Therefore, when regarding the variables chosen for this study, environmental factors do not have a great direct (causal) impact on the health of the residents in the 33 London boroughs studied.

Perhaps had we studied more narrow health variables related to these environmental factors such as lung cancer (for air pollution) or hearing loss (for noise pollution), the correlation between the BSe and the BSh would be bigger, but our study focused on general health indicators because its purpose is to compare the impact of environmental and socioeconomic factors in overall general health.

**Visualising the Correlations and Simple Linear Regressions between Socioeconomic Variables and Health:**

*Note: A detailed step-by-step iPython notebook for these visualisations is available in our GitHub Repository:*

[*https://github.com/Ftracy/QMGroup22/blob/a7ccb36fb9a3a8d9fdaae39962539c87830fdab0/Socioeconomic%20Variables%20and%20Health.ipynb*](https://github.com/Ftracy/QMGroup22/blob/a7ccb36fb9a3a8d9fdaae39962539c87830fdab0/Socioeconomic%20Variables%20and%20Health.ipynb)

*Sources: Waskom, a, 2022; Waskom, b, 2022*

**Correlation between Socioeconomic Variables and Health**

[**Figure 5: A Heatmap to Show the Correlation Coefficients between Socioeconomic Variables and Health]**

**Background pattern

Description automatically generated**

[**Figure 6: A Pairplot to Visualise the Correlations between Socioeconomic Variables and Health]**

**A picture containing text, shoji, building, crossword puzzle

Description automatically generated**

**Explanation for both figures:**

The heat map shows the Pearson Correlation Coefficient between the 3 environmental variables we have chosen - income, education, and race and migration - and health status as a whole. How these have been calculated is outlined in our methodology section. The further away a value is from 0, the stronger the relationship between two variables. The pairplot visualises the correlations between the aforementioned socioeconomic variables and health. The diagonals are histograms to show the distribution of data.

As we can observe from these two visualisations, race and migration has a very weak, positive correlation with health status (0.092), whilst income has a stronger positive correlation with health (0.31), suggesting that the wealthier one is, the better their health. The socioeconomic variable with the strongest correlation with health is education (0.52). This correlation is positive, indicating that the boroughs with the greatest educational attainment, taking into consideration the indicators we have outlined in our methodology, have better health outcomes.

**Correlation between BSse and BSh**

*Note: A detailed step-by-step iPython notebook for this visualisation is available in our GitHub Repository:*

[*https://github.com/Ftracy/QMGroup22/blob/7186ea806ae78e718bc2e0c71b708ecd2fa349a5/BSh%20and%20Bsse%20Data%20Regression.ipynb*](https://github.com/Ftracy/QMGroup22/blob/7186ea806ae78e718bc2e0c71b708ecd2fa349a5/BSh%20and%20Bsse%20Data%20Regression.ipynb)

*Sources: Waskom,c, 2022*

**Figure 7: A Linear Regression with Marginal Distributions to Visualise the Correlation between BSse and BSh**

**Chart

Description automatically generated**

R Squared value (calculation made in Microsoft Excel): 0.205

[**Explanation:** This visualisation consists of two elements: A linear regression between the variables BSse and BSh, and the marginal distribution of the variables contained in each subset. The regression shows that there is a positive correlation between BSh and Bse - as Bse increases, so does BSh. Therefore, the worse a borough is socioeconomically, the poorer its overall health outcomes. The Rsquared value of 0.205 means that 20% of the data of one variable is explained by the other, which while not being very high, can be considered significant when working with real world data]

**Multi-linear Regression on Socioeconomic Factors and Health**

**Figure 8: A Multi-linear Regression on Socioeconomic Factors and BSh**

**Table

Description automatically generated**

The data for hypothesis testing comes from raw data collected on 33 boroughs after normalisation in the data section.

A multi-linear regression (MLR) was conducted to statistically analyse the accumulative effect of the three socioeconomic variables outlined on health (BSh). This was done using Google Collab, which has a built-in MLR model.

The R squared value is 0.350, meaning that approximately 35% of the data points fit in the regression model, which indicates the noise of data and suggests that multi-linear models might not be the best model to simulate the correlation between overall socioeconomic factors and health. If we set the significance value at 0.05, then the p-value is significantly lower than this value, rejecting Hypothesis 2 that states there is no correlation between overall socioeconomic factors and health. Hence, the overall socio-economic factors have a correlation with health, and we accept Null Hypothesis 2.

**3 independent variables on health:** Whilst the p-values for income (0.112) and race and migration (0.133) are above the significance level (0.05), the p-value for education score (0.002)  is below the significance level, indicating it is the only individual factor that has a strong correlation with health independently.

*Note that the interpretation of coefficients is not significant in this research as we calculated scores through the normalisation of our datasets.*

In conclusion, there is a correlation between the BS**SE** and the BSh, therefore, considering the variables chosen for this study, the socio-economic variables of income, education and race and migration have a great direct impact on the health of the residents of the 33 London boroughs studied. Amongst them, education seems to exert the most direct (causative) influence.