What factors affect life expectancy in London and how can it be improved?

1 Introduction

Life expectancy (LE) is a commonly used metric to assess the mortality rate in a population. The phrase "life expectancy" denotes the average number of years a person can expect to live (Roser, Ortiz-Ospina and Ritchie, 2013). LE is often used as a summary measure of a population's overall health and well-being, as it provides a clear and easy-to-understand indicator of mortality, which allows for meaningful comparisons of mortality rates across populations, as it adjusts for variations in the proportion of older or younger individuals in each population (Modig, Rau and Ahlbom, 2020). At present time, health and well-being is a major concern of the government and the general public, and improving well-being is a shared goal of all people. Therefore, in order to identify the elements that can impact a population's health and well-being, life expectancy can serve as a quantifiable indicator. This research aims to determine the possible relationship between social factors and life expectancy by implementing linear regression model in Python.

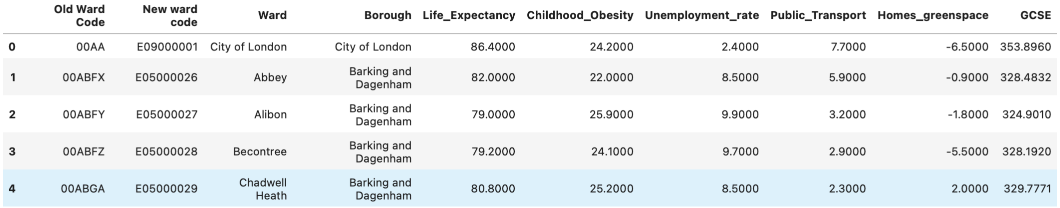
2 Literature Review

There have been a number of studies conducted on identifying factors that affect life expectancy and potential ways to improve it. Some of the most commonly studied factors include genetics, lifestyle choices, access to healthcare, and socio-economic status. For example, as Roser, Ortiz-Ospina and Ritchie (2013) found out in their research, countries that spend more money per person on healthcare tend to have a higher average life expectancy, and the biggest improvements in life expectancy are seen in countries with lower spending and a lower standard of living. Additionally, they also examined the correlation between life expectancy and GDP, and found that countries with higher GDP display higher LE as well. The relationship between GDP and life expectancy is logarithmic, meaning that as GDP increases, the correlation between GDP and life expectancy decreases, which may be due to the fact that, in general, poorer countries tend to have less access to healthcare, education, and other resources that can impact life expectancy. As a country becomes wealthier, the impact of GDP on life expectancy tends to decrease. Their research has also shown that better science and technology can improve life expectancy, because it means more interventions such as vaccination, disease management, and preventative care. Advances in medical technology and treatments have indeed played a role in improving life expectancy. In Mathers *et al.* (2015)'s research, they examined how life expectancy has changed at older age, which is defined as aged 60 years or older.

The data analyzed in their study showed that over the past three decades, there has been a significant change in the progress of population health status. Specifically, mortality rates have been decreasing in older age groups in many countries. This change can be attributed to a number of factors such as advances in medical technology and treatments, declined exposure to risk factors such as blood pressure and tobacco, as well as improved access to healthcare and better living conditions.

3 Data

Apparently, many research focus on personal health factors, while in this project, health factors such as childhood obesity, social factors such as unemployment rate, access to public transport, access to green open spaces and nature, and educational factors such as GCSE score are examined. The data used in this research is from LondonDatastore (2013). The following is the first several rows of the data:



	Old Ward Code	New ward code	Ward	Borough	Life_Expectancy	Childhood_Obesity	Unemployment_rate	Public_Transport	Homes_greenspace	GCSE
0	00AA	E09000001	City of London	City of London	86.4000	24.2000	2.4000	7.7000	-6.5000	353.8960
1	00ABFX	E05000026	Abbey	Barking and Dagenham	82.0000	22.0000	8.5000	5.9000	-0.9000	328.4832
2	00ABFY	E05000027	Alibon	Barking and Dagenham	79.0000	25.9000	9.9000	3.2000	-1.8000	324.9010
3	00ABFZ	E05000028	Becontree	Barking and Dagenham	79.2000	24.1000	9.7000	2.9000	-5.5000	328.1920
4	00ABGA	E05000029	Chadwell Heath	Barking and Dagenham	80.8000	25.2000	8.5000	2.3000	2.0000	329.7771

Figure 1: Data Overview

These data are gathered by the scale of London wards, with six indicators in total. The first one is life expectancy, which is the response variable in this project. Childhood obesity is the prevalence of obesity by area of child residence, in which case children whose BMI is greater than or equal to 95th percentile of the British 1990 growth reference BMI distribution have been classified as obese. Home_greenspace denotes homes with access to public open space and the proportion of that is greenspace. Together with unemployment rate, accessibility to public transport and GCSE points score, each figure is the average number of that indicator in the specified London ward over the period of 2009-2013. The summary statistics are as follows:

Table 1: Summary Statistics

	LE	Child_Obesity	Unemployment	Transport	Homes_greenspace	GCSE
count	625	625	625	625	625	625
mean	82.01	21.27	5.98	3.69	-0.22	342.49
std	2.26	4.79	3.15	1.36	6.80	17.12
min	76.00	6.10	0.50	1.30	-22.30	306.20
25%	80.40	18.20	3.40	2.60	-5.00	329.31
50%	81.90	21.90	5.50	3.30	-0.10	340.21
75%	83.40	24.90	8.10	4.50	4.50	354.10
max	90.30	35.80	19.70	8.00	18.30	389.38

To determine which quantitative method should be used, life expectancy data is plotted against every indicator. The figures presented below suggest that there is a correlation between various indicators and life expectancy. However, it appears that accessibility to open greenspaces is not related to life expectancy in the same way as the other indicators. It is important to note that this does not necessarily mean that access to open greenspaces has no impact on overall health and well-being, but rather that it may not have a direct correlation with life expectancy in this specific study or data set.

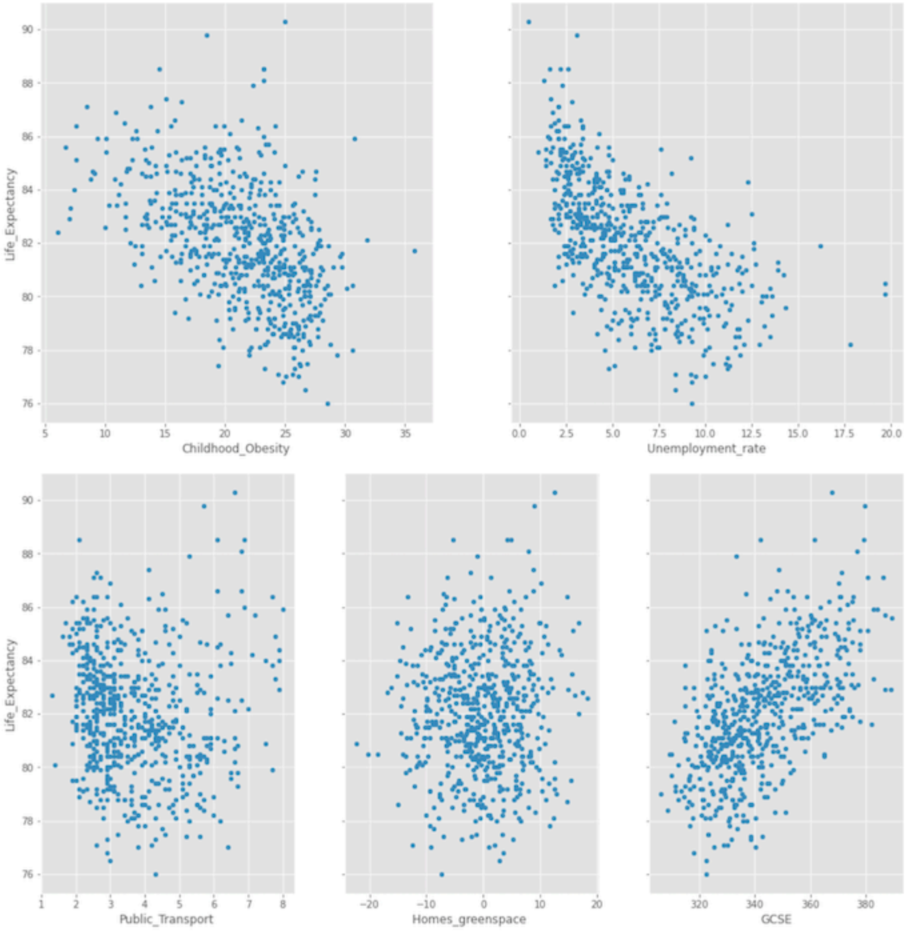


Figure 2: Plots of life expectancy and each variable

A box plot is also produced to check for outliers. Before plotting, the data set is normalized to produce a more comparable and understandable plot, and also to enable comparison between regression coefficients to conclude which factors affects life expectancy the most. As seen below, there are some outliers within every group of data except GCSE scores, but after observing the original data, these outliers do not appear to be mistakenly recorded data. Therefore, these points were not removed.

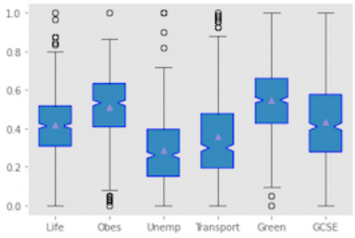


Figure 3: Box Plot

4 Methodology

After plotting the indicators and life expectancy, it can now be sure to use multiple linear regression model to further investigate into their relationship. Firstly, a correlation matrix is produced between each of the indicators.

Table 2: Correlation Matrix

	LE	Child_Obesity	Unemployment	Transport	Homes_greenspace	GCSE
LE	1.00	-0.49	-0.62	-0.08	0.05	0.58
C_Obesity	-0.49	1.00	0.66	0.32	-0.06	-0.66
Unemploy	-0.62	0.66	1.00	0.21	-0.06	-0.70
Transport	-0.08	0.32	0.21	1.00	-0.15	-0.19
H_green	0.05	-0.06	-0.06	-0.15	1.00	0.01
GCSE	0.58	-0.66	-0.70	-0.19	0.01	1.00

It can be seen above that each variable has some level of correlation, with accessibility to public open greenspaces and public transport to be the least correlated ones with life expectancy. While childhood obesity seem to have relatively strong correlation with unemployment rate and GCSE points score, and correlation can also be observed between these two indicators. Therefore, it is necessary to check for multicollinearity before regression to maintain accuracy of the model.

The second step is to check for multicollinearity and drop the variable that displays strong correlation with other variables. The approach to detect multicollearity used in this research is variance inflation factors, with the threshold to be 5. It will produce predictors with mitigated multicollearity. After handling the predictors, those who are left will be used as an input into the linear regression model. Assume the model to be :

$$y = a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_3 + a_4 * x_4 + a_5 * x_5,$$

where x_1 is Childhood_obesity, x_2 is Unemployment_rate, x_3 is Public_Transport, x_4 is Home_greenspace, x_5 is GCSE score, and y is life expectancy.

The hypothesis for this model is as follows:

$$H_0 : a_1 = a_2 = a_3 = a_4 = a_5 = 0;$$
$$H_1 : At\ least\ one\ of\ a_i(i = 1, 2, 3, 4, 5)\ is\ not\ 0.$$

With 95% confidence level, predictors whose p-values of their coefficients are greater than 0.05 will be removed, because it means they are not significant in the model. By repeating the above steps of model regression, a significant model may be produced. Next, residuals need to be checked by plotting residuals vs. fitted values to make sure they meet the assumptions of multiple linear regression:

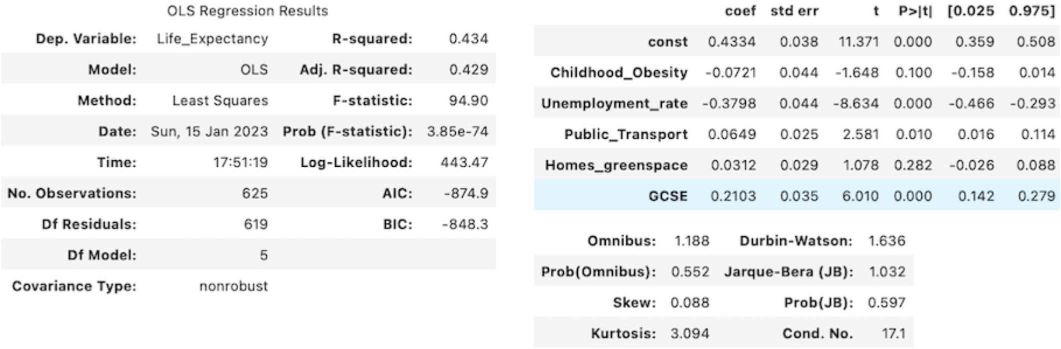
1. Linear relationship exists, which can be identified by checking if residuals scatter around residual=0 randomly;
2. Errors are independent with no systematic patterns;
3. Errors are normally distributed;
4. Equal variances.

Q-Q plot for residuals may be plotted as well to check for normality.

In the last step, logistic regression is used to predict when life expectancy exceeds 85. This threshold is chosen here because in the original data, 85 can be counted as a rather high life expectancy, which means this number can act like a goal when improving life expectancy.

5 Empirical Results

Since correlation displays within the predictors, VIF is used to mitigate it. After running the process, no predictors were dropped. Therefore, all of them can be used to fit the regression model. The regression result is below. The p-values of Childhood_Obesity and Home_greenspace are greater than 0.5, so they should be dropped.



OLS Regression Results

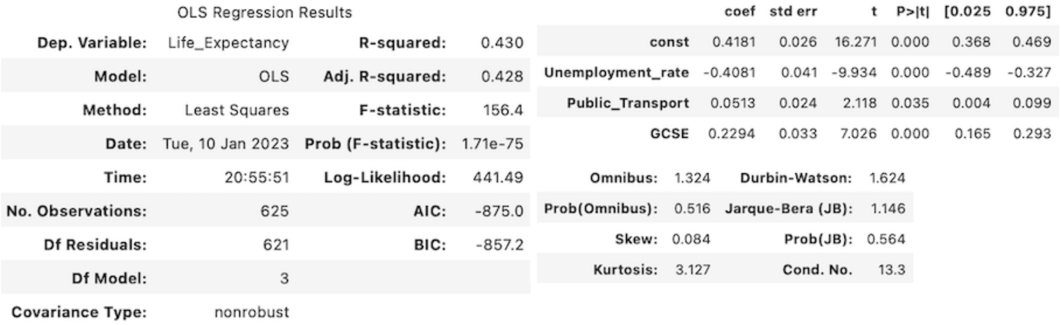
Dep. Variable:	Life_Expectancy	R-squared:	0.434
Model:	OLS	Adj. R-squared:	0.429
Method:	Least Squares	F-statistic:	94.90
Date:	Sun, 15 Jan 2023	Prob (F-statistic):	3.85e-74
Time:	17:51:19	Log-Likelihood:	443.47
No. Observations:	625	AIC:	-874.9
Df Residuals:	619	BIC:	-848.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4334	0.038	11.371	0.000	0.359	0.508
Childhood_Obesity	-0.0721	0.044	-1.648	0.100	-0.158	0.014
Unemployment_rate	-0.3798	0.044	-8.634	0.000	-0.466	-0.293
Public_Transport	0.0649	0.025	2.581	0.010	0.016	0.114
Homes_greenspace	0.0312	0.029	1.078	0.282	-0.026	0.088
GCSE	0.2103	0.035	6.010	0.000	0.142	0.279

Omnibus:	1.188	Durbin-Watson:	1.636
Prob(Omnibus):	0.552	Jarque-Bera (JB):	1.032
Skew:	0.088	Prob(JB):	0.597
Kurtosis:	3.094	Cond. No.	17.1

Figure 4: First Regression

Then the regression model is run again with the remaining predictors:



OLS Regression Results

Dep. Variable:	Life_Expectancy	R-squared:	0.430
Model:	OLS	Adj. R-squared:	0.428
Method:	Least Squares	F-statistic:	156.4
Date:	Tue, 10 Jan 2023	Prob (F-statistic):	1.71e-75
Time:	20:55:51	Log-Likelihood:	441.49
No. Observations:	625	AIC:	-875.0
Df Residuals:	621	BIC:	-857.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4181	0.026	16.271	0.000	0.368	0.469
Unemployment_rate	-0.4081	0.041	-9.934	0.000	-0.489	-0.327
Public_Transport	0.0513	0.024	2.118	0.035	0.004	0.099
GCSE	0.2294	0.033	7.026	0.000	0.165	0.293

Omnibus:	1.324	Durbin-Watson:	1.624
Prob(Omnibus):	0.516	Jarque-Bera (JB):	1.146
Skew:	0.084	Prob(JB):	0.564
Kurtosis:	3.127	Cond. No.	13.3

Figure 5: Second Regression

The p-values of the model and each predictor are smaller than 0.05, suggesting the model is significant. By comparing the coefficients, unemployment rate seems to be the most influential predictor and it affects life expectancy negatively. R^2 equals 0.43, meaning that 43% of the total variance of life expectancy can be explained by the variables used in the model. The conditional number is small, suggesting multicollinearity has been successfully mitigated.

Next, residual analysis is conducted to test if the model satisfies the conditions. The result below shows the residuals satisfy all conditions, as the plot against fitted values shows randomly distributed residuals around reidual=0 with no significant patterns and equal variances for all x, and the Q-Q plot shows they follow the normal distribution.

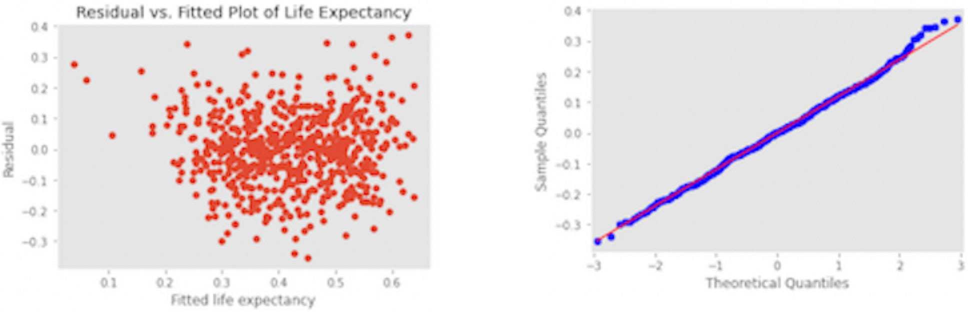


Figure 6: Residual vs. Fitted

Figure 7: Q-Q Plot

Figure 8: Residual Analysis

Therefore,the model can be expressed as:

$$y = 0.4181 - 0.4081 * x_1 + 0.0513 * x_2 + 0.2294 * x_3 + \epsilon, \epsilon \sim N(0, \sigma^2),$$

where σ is a constant, x_1 is Unemployment_rate, x_2 is Public_Transport, x_3 is GCSE score, and y is life expectancy.

Lastly, the logistic regression result shows 91.52% accuracy, and the confusion matrix is produced as well. It shows the number of times each combination of actual and predicted categories occurred. For example, there are 11 samples that are correctly predicted as Type 1 (i.e. over 85).

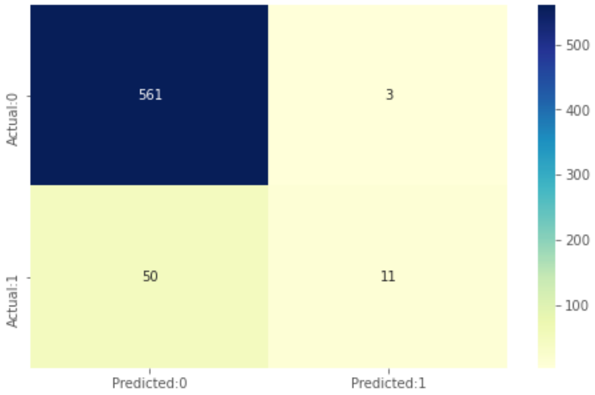
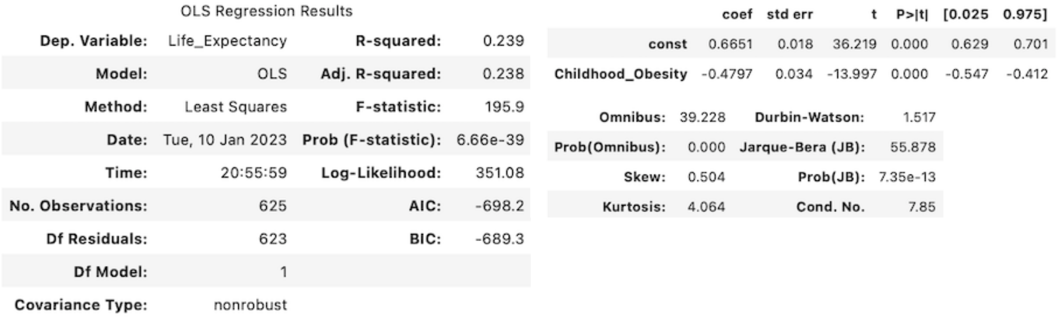


Figure 9: Confusion Matrix

6 Discussion

In this project, I found that low correlation does not necessarily mean zero relationship. For example, the correlation coefficient between life expectancy and accessibility to public transport is only -0.08, but its coefficient shows significance in the regression result. One limitation of my research is that in the plot of childhood obesity against life expectancy, there displays a relationship between them. However, the regression result shows otherwise. When I regress only childhood obesity on life expectancy, its coefficient is significant, but R^2 is relatively low. I presume it is because its correlation with unemployment rate and GCSE score. Though their correlation is not strong enough for VIF to drop them, it affects the model to some extent after all. Secondly, dropping the insignificant predictors did not improve the adjusted R^2 much. More future work may be done to deal with this issue.



OLS Regression Results

Dep. Variable:	Life_Expectancy	**R-squared:**	0.239
Model:	OLS	**Adj. R-squared:**	0.238
Method:	Least Squares	**F-statistic:**	195.9
Date:	Tue, 10 Jan 2023	**Prob (F-statistic):**	6.66e-39
Time:	20:55:59	**Log-Likelihood:**	351.08
No. Observations:	625	**AIC:**	-698.2
Df Residuals:	623	**BIC:**	-689.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.6651	0.018	36.219	0.000	0.629	0.701
Childhood_Obesity	-0.4797	0.034	-13.997	0.000	-0.547	-0.412

Omnibus:	39.228	**Durbin-Watson:**	1.517
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	55.878
Skew:	0.504	**Prob(JB):**	7.35e-13
Kurtosis:	4.064	**Cond. No.**	7.85

Figure 10: Regression of Childhood Obesity Solely

7 Conclusion

The model result shows that out of five predictors, three turned out to be influential on life expectancy, including unemployment rate, accessibility to public transport and GCSE score. Among the three predictors, unemployment rate is the most influential one. Therefore, to improve life expectancy, the government can start by reducing unemployment rate, building more public transport facilities in a more efficient way, and raising the overall educational level of the public.

Word Count: 1667

References

LondonDatastore (2013) ‘London ward well-being scores’. Available at: https://data.london.gov.uk/dataset/london-ward-well-being-scores.

Mathers, C. D. *et al.* (2015) ‘Causes of international increases in older age life expectancy’, *The Lancet*, 385(9967), pp. 540–548. doi: https://doi.org/10.1016/S0140-6736(14)60569-9.

Modig, K., Rau, R. and Ahlbom, A. (2020) ‘Life expectancy: What does it measure?’, *BMJ Open*, 10(7). doi: 10.1136/bmjopen-2019-035932.

Roser, M., Ortiz-Ospina, E. and Ritchie, H. (2013) ‘Life expectancy’, *Our World in Data*.