CA682 Data management and visualisation – Assignment 1

Name	Hugh Pearse
Student Number	18214570
Programme	GCAI1
Module Code	CA682
Assignment Title	Data Visualisation Assignment 1
Due Date	14th December 2018 before 23:59
Submission date	
Module coordinator	Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Signed:

_____

Name: Hugh Pearse
Date: 27nd November 2018

Visualising Summary Statistics for London Boroughs

Hugh Pearse
School of Computing
Dublin City University
Dublin, Ireland
hughpearse@gmx.co.uk
November 27th, 2018

Abstract The aim of this project is to take the results of a diverse collection of surveys relating to London and provide a method of visually investigating if the data contains any measure of linear correlation between multiple datasets. This will provide data researchers a method a simple way to visually find novel linear relationships. Finding linear relationships is a well-known strategy that has been revisited countless times. The focal point for this project is to trace back to the origins of these types of visualisations and introduce some small incremental improvements.

Keywords: Pearson Correlation, Linear Regression, Kolmogorov-Smirnov Test, Data Visualisation, Design Thinking

Introduction

Linear regression is a process that allows a data scientist to graph two numeric datasets and plot a line across the data. The slope of the line describes how much a change in one variable will result a change in the corresponding variable. Strong linear relationships can allow data scientists to create accurate predictive models. However having confidence in the accuracy of a predictive model is not always a straightforward task. By populating a cross-correlation matrix comparing every possible 2 pair dataset combination (${}_nC_2$) with a Pearson Correlation Coefficient (r) it is possible to colour code their linear relationship. This will give a data researcher a visual starting point for investigation.

Once a researcher has selected a data pair with a reasonable linear association they will need a method to see if the data is suitable to use with a linear model. Linear regression only works if the residuals of the data are normally distributed. To measure this a data researcher should have a mechanism to check for normal normality of the residuals. Outliers are basically abnormalities in your data. Outliers can have a strong influence on the normality of the residuals and a data researcher should have a mechanism to remove these to make more accurate linear models.

Once the normality of the residuals has been established a data researcher should have a method to create a linear model of the remaining data. This should clearly show the normal range for the data, the normal range for the linear model and the location of any outliers.

Literature Review

Use of colour coding for visualisation of a matrix can be traced back as far as 1873 when Loua summarised social statistics regarding the arrondissements of Paris using a shaded matrix display [1].

The development of the scatterplot for visualising linear regression can be traced back to the work of Francis Galton in 1886 [2] who graphed bivariate scatterplots, plotted regression lines, surrounded points using an ellipse and plotted vertical lines to show residual values. The contemporary principles of good graph design applied to this project are taken from the 6 principles outlined by Tufte [3].

Dataset

The London Datastore provides a one-stop shop for a diverse collection of datasets relating to London that have been collected from various government agencies around the UK. The approach for this project was use the London Datastore website [4] to source 13 interesting data sets [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] in various file formats and to analyse them without modification to the original files. The file formats included structured CSV files and semi-structured Microsoft Excel files. The semi-structured Excel files proved challenging as some of the cells had been merged, the data was spread across various sheets and did not have traditional column headers but instead had merged cells as headers for multiple columns. All of the sample data used to develop this project falls under the UK Open Government Licence (OGL v2).

Understanding the Data

From the collection of datasets each contained several columns. One column was always some type of reference key for the area similar to a postal code. The other columns were usually the statistics. Each row usually represented a statistic that was collected for a specific area. The areas were mostly at the level of granularity of the London boroughs but some datasets such as the life expectancy dataset [13] were at a very fine grained level by ward, and also included additional summary statistics at a national level for regions such as Wales or the North East.

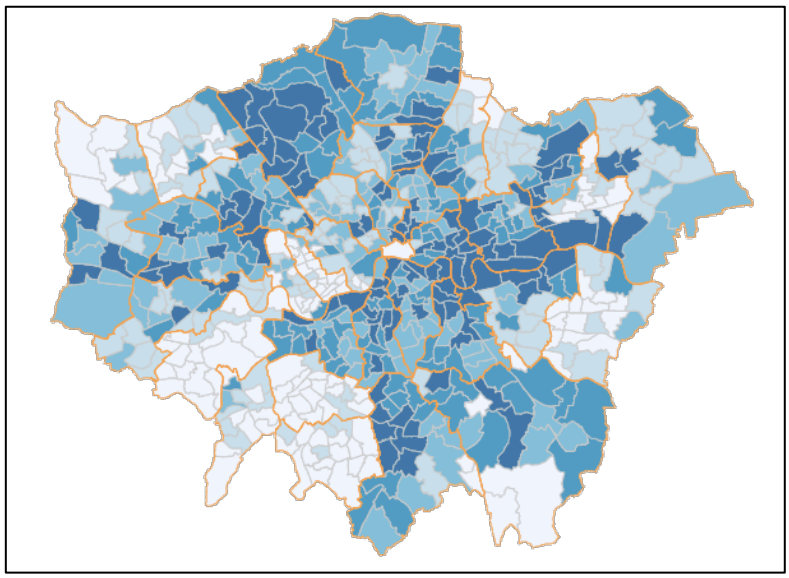

Figure 1. Map showing Boroughs (major) and Wards (minor) in London

Regarding the time period when the data was gathered, not all of the datasets were collected at the same time. Some were collected several times throughout the year [6] and others may just have been collected once at the beginning or end of a year. Additionally not all datasets covered the same years. The business survival rate dataset [10] only had full coverage between 2002-2011, and the GCSE

results dataset only covered between 2015-2016. For this reason it was determined that 2011 had the best coverage and all data would be collected as close as possible to this period to reduce any error.

Data Cleansing Challenges

The statutory homelessness dataset [15] was the worst offender in terms of structure. The file type was a Microsoft Excel file format that contained multiple sheets. The columns headers were with a main title, subtitle, super-headers and sub-headers. The relevant column sub-headers that were of interest were located in separate columns from the data itself. It also contained sub-sub-headers located far from the top and mixed within the data. The worst offence was that the main column headers that described the data were split across several rows.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1							Decisions made during the year April 2011 - March 2012									
2																
3																
4							Numbers accepted as being homeless and in priority need									
5						Number of households (2008 mid-year estimate) thousands										Eligible homeless and in priority need but intentionally	
6	Former ONS Code	Current ONS Code	County and Local Authority area		Number of quarters covered		White	Black or Black British	Asian or Asian British	Mixed	Other Ethnic Origin	Ethnic Group Not Stated	Total		No. per 1,000 h'holds	
7																
8																
9																
10																
11																
12			**England**			21,731	**33,410**	**7,130**	**3,410**	**1,640**	**2,270**	**2,430**	**50,290**		**2.31**	**7,920**
13																
14	H	E12000007	**London**			3,244	4,260	4,420	1,360	520	1,260	910	12,720		3.92	1,890
15																
16			**Inner London**													
17	00AG	E09000007		Camden	4	103	53	46	23	3	9	2	136		1.32	25
18	00AA	E09000001		City of London	4	7	9	4	1	1	2	0	17		2.43	0
19	00AM	E09000012		Hackney	4	90	140	297	36	25	146	42	686		7.62	94
20	00AN	E09000013		Hammersmith and Fulham	4	76	70	94	15	12	12	0	203		2.67	10
21	00AP	E09000014		Haringey	4	98	180	280	43	25	33	12	573		5.85	70
22	00AU	E09000019		Islington	4	87	172	137	31	26	29	18	413		4.75	37
23	00AW	E09000020		Kensington and Chelsea	4	85	197	105	58	38	116	20	534		6.28	61
24	00AY	E09000022		Lambeth	4	126	202	605	37	42	49	24	959		7.61	73
25	00AZ	E09000023		Lewisham	4	115	128	314	19	38	64	4	567		4.93	24
26	00BB	E09000025		Newham	4	92	67	93	60	8	5	15	248		2.70	46
27	00BE	E09000028		Southwark	4	124	53	151	1	0	291	22	518		4.18	154
28	00BG	E09000030		Tower Hamlets	4	93	105	53	219	12	15	0	404		4.34	91
29	00BJ	E09000032		Wandsworth	4	**126**	201	218	105	22	25	20	591		4.69	45
30	00BK	E09000033		Westminster	4	**120**	--	--	--	--	--	--	561		4.67	108
31			**Outer London**													
32	00AB	E09000002		Barking and Dagenham	4	68	88	86	15	3	3	4	199		2.93	12
33	00AC	E09000003		Barnet	4	137	134	98	46	17	26	18	339		2.47	35
34	00AD	E09000004		Bexley	4	93	254	71	8	7	5	1	346		3.72	43
35	00AE	E09000005		Brent	4	98	99	228	74	17	30	35	483		4.93	51
36	00AF	E09000006		Bromley	4	133	429	112	13	19	25	36	634		4.77	55

Figure 2. Overview of homelessness dataset

Aggregating Using Primary Keys

Joining tables in SQL is as simple as specifying the primary and foreign key to join the tables with. However with semi-structured data it does not always have a common value to link together with. To solve this problem a bespoke dataset was required to join all of the other data together on a common factor. This bespoke dataset needed to be a CSV file with 3 columns: the modern borough code, the old borough code and the primary care trust code to accommodate for the immunisation dataset [16]. The primary care trust (PCT) code was linked to its corresponding borough code by searching the National Statistics Postcode Lookup (NSPL) dataset [18]. The NSPL dataset is a >700MB CSV file so a program with a low memory footprint is required to search for the PCT code and its corresponding borough code.

Visualising the Data

The process of visualising the data follows a four step process: (1.) visualising the correlations to show the strength of the relationships between all of the datasets, (2.) testing if the data follows a normal distribution and identify any outliers, (3.) removing outliers and (4.) visualising the linear model. A careful design process for these visualisations is required to convey the information in a simple and effective manner.

Cross-Correlation Matrix

The cross-correlation matrix is a set of tabular data where each column represents a variable and each row also represents a variable. The cells are a representation of f(X,Y) for every ${}_nC_2$ combination. In this case the function f(X,Y) is the Pearson correlation coefficient r. The Pearson correlation coefficient is a symmetric function meaning that r= f(X,Y) and is the same as r= f(Y,X). This results in the upper triangular portion of the matrix being a reflection of the lower triangular portion of the matrix and the diagonal across the middle being a self-comparison.

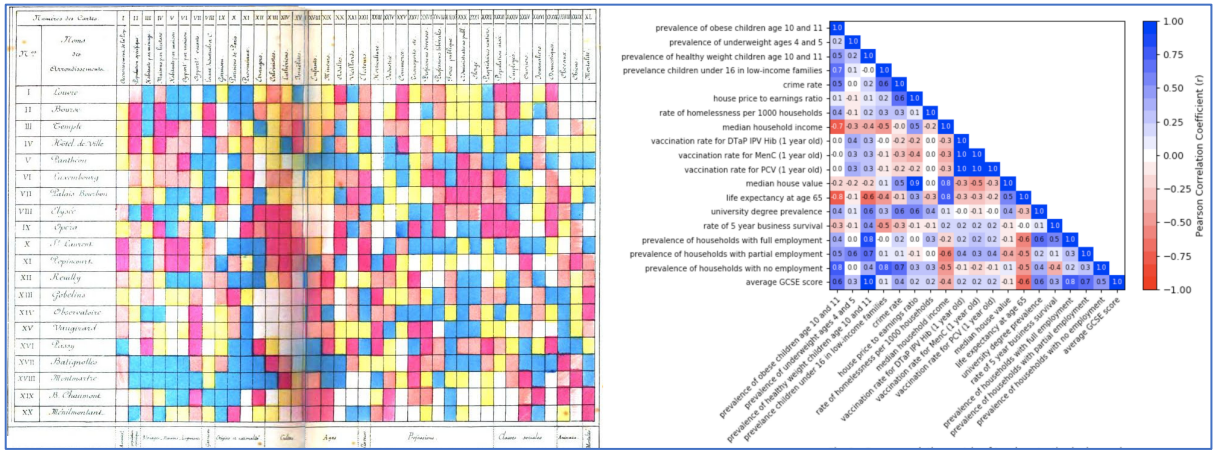

Figure 3. Visualisation of cross correlation matrix based on original work by Toussaint Loua

By applying Tufte's rule for maximising the data-to-ink ratio, the upper or lower triangular portion of the correlation matrix can be removed. As the cell colour represents the strength of the Pearson correlation coefficient in the range of -1.0 to +1.0 a diverging colour palette from blue to red was selected. White was chosen to represent the absence of correlation. This palette was divided in to 21 unique colours. The text placed on top of the colour was black. As colour saturation increased for the stronger correlation values the black text became harder to read. For the extreme values the text colour was automatically changed from black to white to make it more legible. The correlation matrices were tested using a colour blindness simulator. The original matrix by Loua resulted in a problem differentiating between red and blue cells with the Protanopia (Red-Blind) simulation and with the Achromatopsia (Monochromacy) simulation. However, the new matrix with the blue-red diverging colour palette worked well with most types of colour blindness and only encountered minor problems with Achromatopsia. The new matrix had an additional feature of including text superimposed on top of the cell which mitigated problems related to Achromatopsia.

Normality of Residuals

In order for linear regression to produce accurate and reliable results, the distribution of the residuals from the mean must be normally distributed. There are a number of tests which allow a researcher to verify if their dataset has normally distributed residuals. Some of these tests include the Kolmogorov-Smirnov test [19], the Anderson-Darling test [20] and the Shapiro-Wilk test [21]. For the purposes of visualising the proximity of the fit of the data to a normal distribution the Kolmogorov-Smirnov test is most suitable. The data is first of all normalised to have a mean of zero. The residuals are calculated as the distance from the mean. This can be easily seen in a stem plot. The distribution of the z-score of the residuals can then be visualized in a histogram aggregated in "bins" by proximity in to their combined probability densities. The normal curve of best fit is then superimposed on top of the

histogram. Some summary statistics of the data distribution can be displayed as a horizontal box plot underneath the fitted standard normal curve.

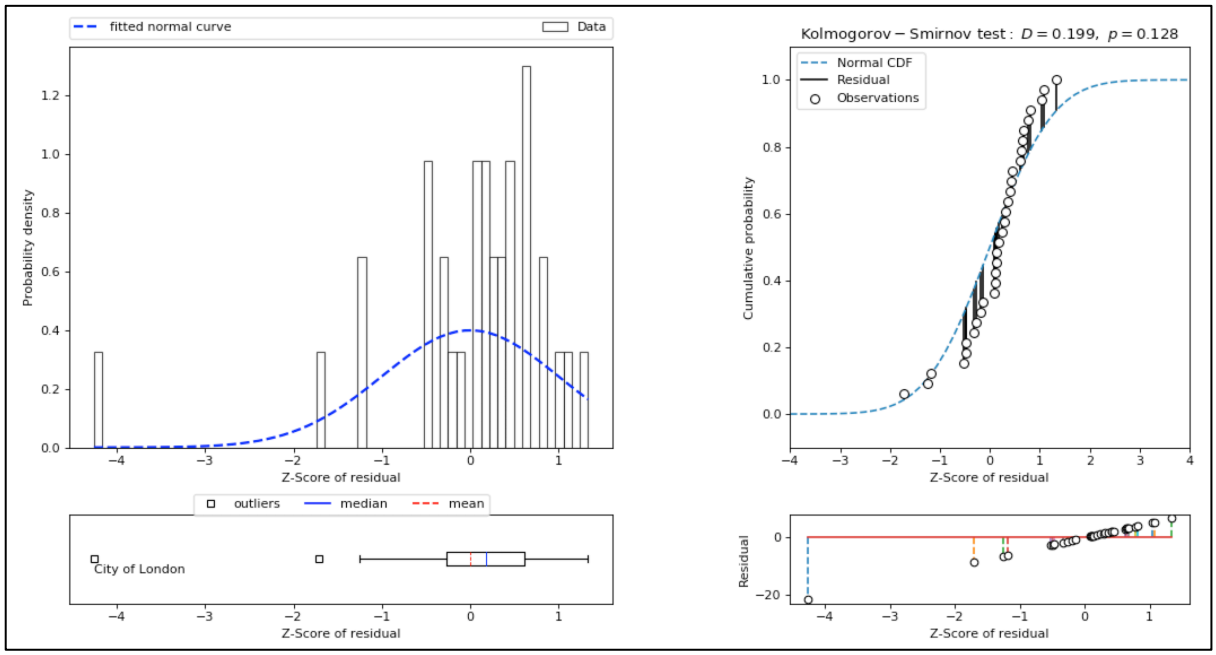

Figure 4. Visualising the distribution of the residuals of a dataset

When a box plot is only used to display a single variable it lends itself well to annotations as the annotations do not risk of covering other variables. This means that annotations can be added to label outliers. One of the drawbacks of the box plot is that it requires some experience to read an effort is made to introduce a legend explaining some of the symbols. Lastly the Kolmogorov-Smirnov graph is displayed showing the ideal cumulative probability curve and the actual residual values superimposed on top and around the curve. Finally by applying Tufte's rule for maximising the data-to-ink ratio, the fill area inside the histogram and points can be removed.

Outliers

As outliers can influence the accuracy of a predictive model, there should always be a mechanism to identify and remove them. To identify them is a two-step process. Firstly they must be identified mathematically. The rule of thumb is that is the data falls below -2.698 standard deviations or above +2.698 standard deviations from the mean then there is a strong chance that it is an outlier. Once the data items have been identified mathematically it must then be communicated visually. The approach for this has been to label outliers by superimposing annotations beside areas of interest on the graphs. This requires careful calculation as multiple outliers may require multiple annotations and the text may end up overlapping.

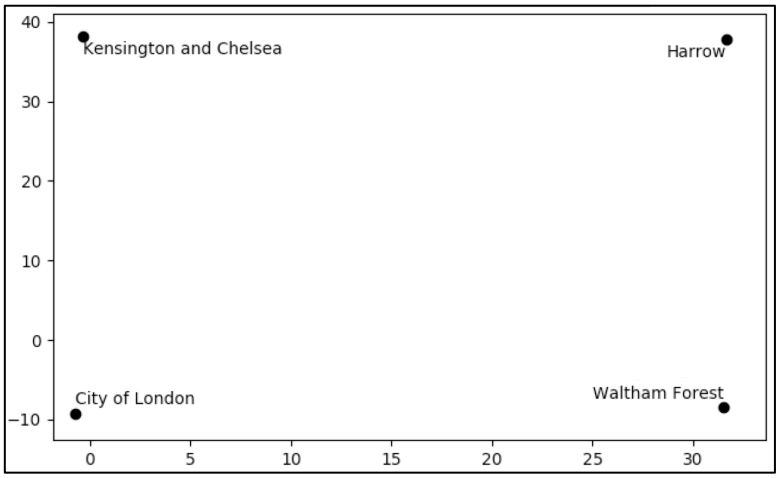

Figure 5. Shifting annotations to avoid escaping the graph

If a point falls on the left of a graph and the annotation is centred around the point, then annotation may escape the left side of the graph. To remedy this the annotation should be displayed to the right of the point, and vice versa for the right. Also when a point is near the top of a graph, the annotation should be displayed underneath the point to prevent it escaping from the top, and vice versa for the bottom.

Interactive Cleansing

The Greater London area is not considered the same thing as the City of London. The Greater London area is managed by the Greater London Authority which is headed by the mayor of London (currently Sadiq Khan). The Greater London area includes the ancient roman City of London, however the City of London has a special status, with its own government, its own police force, its own mayor and most importantly is managed by The City of London Corporation. This means that The City of London is not always involved in all surveys of the Greater London area and often statistics are missing data for the borough of the City of London. In order to facilitate for this issue a mechanism has been introduced to manually remove outliers when appropriate.

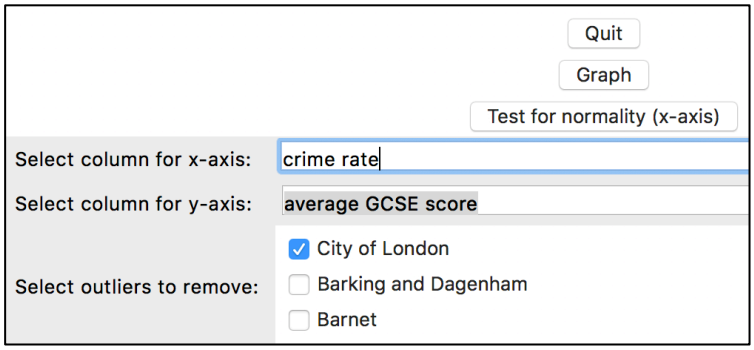

Figure 6. Interactive data cleansing mechanism to remove outliers

Linear Model

Once the data is confirmed to have residuals that are normally distributed, then a line can be plotted across a scatterplot of two variables. The regression line is plotted in such a way as to minimise the sum of the residuals as much as possible. An ellipse is used to display a quantile of interest such as two standard deviations capturing 95% of the data. This can allow a researcher to easily see if there

are outliers in the data. As well as the regression line it is also possible to add an enhancement to this to fit lines at various quantiles from the regression line. This provides additional information about points that are outliers from the rest of the data but are accounted for by the model. Outliers that fall outside a specific quantile of both the data and the model should be annotated as they may be negatively influencing the accuracy of the model and may need to be removed from the analysis.

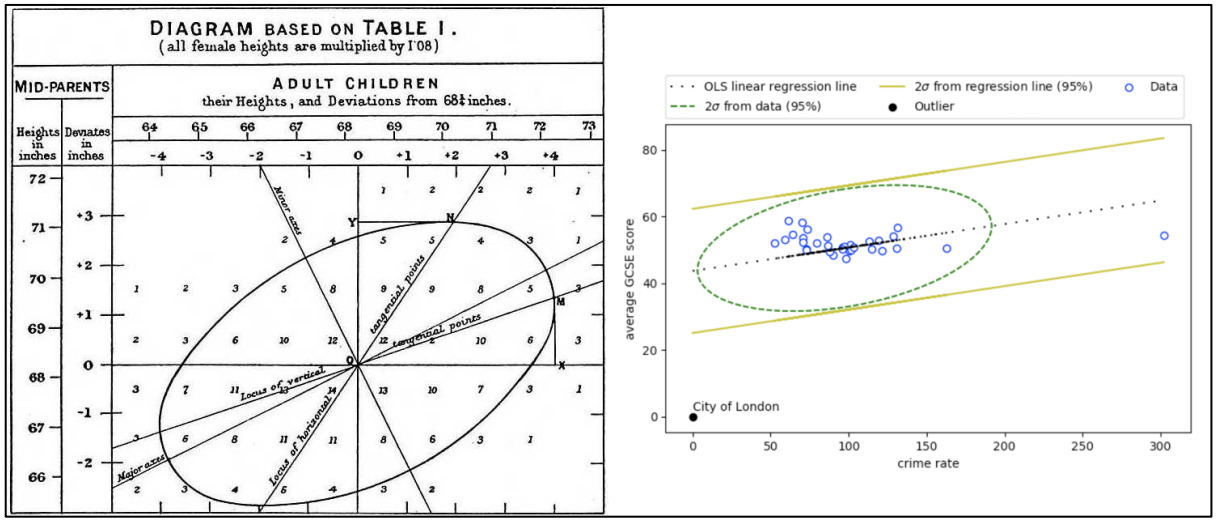

Figure 7. Visualisation of linear model based on work by Francis Galton

The goal for formatting the points was to minimise the use of ink and accommodate printing in black and white. Only points of interest are filled, while the majority are hollow.

Critical Analysis

The weakest factor of this work was the colour of the cross correlation matrix. If a researcher decides to print the matrix out on paper using black and white, the matrix uses a lot of ink and the user loses the ease of finding positive and negative correlations based on colour.

The second weakest factor was the design of the legend in the linear model. The central regression line provides density information. If the density is more dispersed it can become difficult to tell the difference between the regression line and the ellipse. This means that to differentiate between the ellipse and the and the regression line a marker would have to be superimposed on top such as a triangle or square. Adding a circular marker would cause confusion with the point data. However adding any markers at all may result in cluttering the image and take away from its ease of use.

Conclusion

The process of researching, designing and creating simple graphs that clearly convey complex ideas requires a considerable amount of effort. A lot of the problems have been solved by previous researchers and since their inception the same visualisations have gradually improved incrementally. The tool that was created as a result of this research follows sound design principles and can be used in the future to find new relationships in interesting data sets with little modification.

References

[1] T. Loua, *Atlas statistique de la population de Paris*. J. Dejey & cie, 1873.
[2] F. Galton, 'Regression towards mediocrity in hereditary stature.', *J. Anthropol. Inst. G. B. Irel.*, vol. 15, pp. 246–263, 1886.
[3] E. Tufte and P. Graves-Morris, *The visual display of quantitative information.* 1983.
[4] The Mayor of London, 'London Datastore'. [Online]. Available: https://data.london.gov.uk/. [Accessed: 22-Nov-2018].
[5] Metropolitan Police Service, 'Recorded Crime: Borough Rates'. Jan-2017.
[6] Land Registry, 'Average House Prices by Borough, Ward, MSOA & LSOA'. Dec-2017.
[7] Department for Education, 'GCSE Results by Borough'. Jul-2017.
[8] Office for National Statistics (ONS), 'Qualifications of Working Age Population (NVQ), Borough'. Dec-2016.
[9] Greater London Authority (GLA), 'Household Income Estimates for Small Areas'. Mar-2013.
[10] Office for National Statistics (ONS), 'Business Demographics and Survival Rates, Borough'. Jan-2015.
[11] Office for National Statistics (ONS), 'Workless Households, Borough'. Dec-2017.
[12] Department of Health, 'Prevalence of Childhood Obesity, Borough, Ward and MSOA'. Aug-2017.
[13] Office for National Statistics (ONS), 'Life Expectancy at Birth and Age 65 by Ward'. Dec-2014.
[14] HM Revenue & Customs, 'Children in Poverty, Borough and Ward'. Aug-2014.
[15] Ministry of Housing, Communities & Local Government (MHCLG), 'Homelessness Provision, Borough'. Mar-2017.
[16] Department of Health, 'Immunisation Rates for Children at 1st, 2nd and 5th Birthdays'. Dec-2017.
[17] Office for National Statistics (ONS), 'Ratio of House Prices to Earnings, Borough'. Dec-2017.
[18] Office Of National Statistics, 'National Statistics Postcode Lookup UK'. .
[19] N. V. Smirnov, 'Approximate laws of distribution of random variables from empirical data', *Uspekhi Mat. Nauk*, no. 10, pp. 179–206, 1944.
[20] T. W. Anderson and D. A. Darling, 'Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes', *Ann. Math. Stat.*, vol. 23, no. 2, pp. 193–212, Jun. 1952.
[21] S. S. Shapiro and M. B. Wilk, 'An analysis of variance test for normality (complete samples)', *Biometrika*, vol. 52, no. 3–4, pp. 591–611, Dec. 1965.