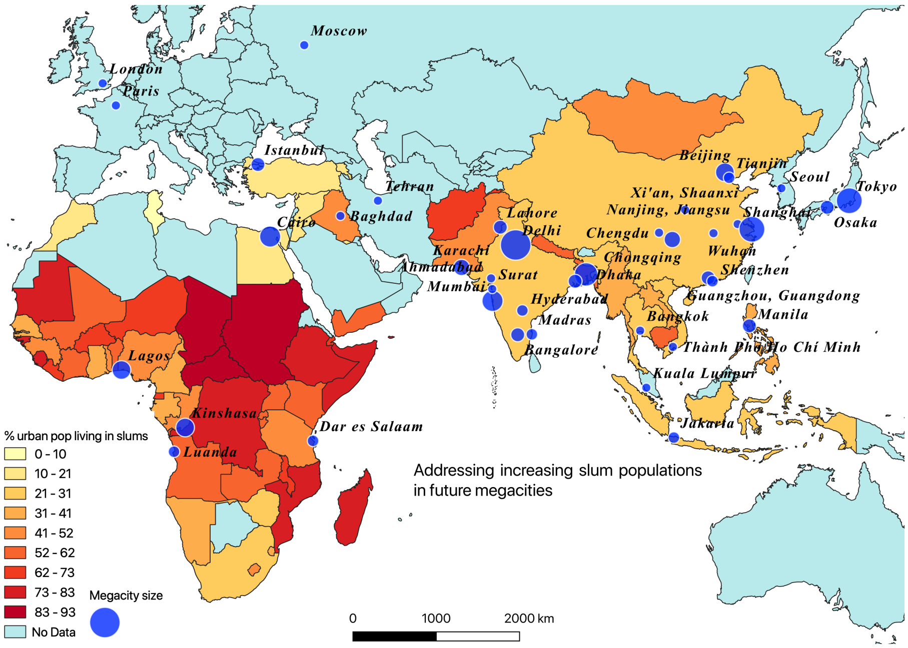
GIS Final Assessment: Parts 1, 2, & 3

January 4, 2019

Part 1 - Presentation and Critical Evaluation of Maps Produced Using Different Software Packages

Qgis Map:



R Map:

Benefits and drawbacks of GUI vs Command line generated maps

The process involved in building the same map using QGIS, a GUI-based software platform, and in R, a code-based statistics software, had its advantages and drawbacks when comparing the two softwares. Building a map in a GUI-based software was easier to practice than using code-based software. Editing the data attributes visually makes breaking down the workflow easier than using command line generated maps. The first map was built in Qgis 3.4.0, as the GUI-based software. The advantages I saw using the Layers tab in ArcMap was directly dragging the data, as its imported in from ArcCatalog side-tab, in addition to the map visually displaying immediately, without having to convert it frequently in command line. Another advantage about the Layers are they can manually be turned off and on in ArcMap. In R, calculating statistics has a fluid flow and can be stored as set variables for later use. Editing a small element to change the final output in a code script is easier than having to repeat a whole process in GUI-based software.

Data sources

The goal was to build a map on a global scale. The world countries shapefile polygon data was downloaded from a reliable websource, The Database of Global Administrative Areas, a firm that "provides maps and spatial data for all countries and their sub-divisions". Megacities data was downloaded from the United Nations population surveys database as an .xls file on total populations and projections of cities over 300k around the world. The percent of each countries' urban population is living in slums was downloaded at ourworldindata.org, as a .csv file, from an article, "Urbanization", by Hannah Ritchie and Roser. Most current survey data is from 2014. It is worthy to note the data on the share of urban population living in slums is missing data for good chunk of our world.

Workflow

The process involved in building a final map visualization was broken up into three main themes: finding and importing data, cleaning the data, building the map, and finally polishing up the map visualization i.e. legends, scale bar, etc… and exporting it. The maps built, visualizes a choropleth density maps of the percent of each countries' urban population living in slums with added layers of 2020 population spread raster layer estimates and 2035 megacities points estimates for Africa and Asia. The first step was importing the world countries vector layer shapefile to set up as base polygons. To add on, importing in the cities as a .csv file and set geometry as latitude and longitude inputs, and importing in the share of urban population living in slums as a .csv file attribute table with no geometry. Join world countries and share of urban pop living in slums. Filter out all cities > 10 million for 2035 to set as megacities estimates. Visualize world countries map with % of urban pop living in slums as the polygon fill, and project megacities population estimates of 2035 as points.

Assessment of Maps

The Qgis map displays the megacities as a scale to their size, rather in R, the points are manually scaled to one size. The R generated map is interactive and the user is able to click on the polygons and points to display a pop-up information tab.

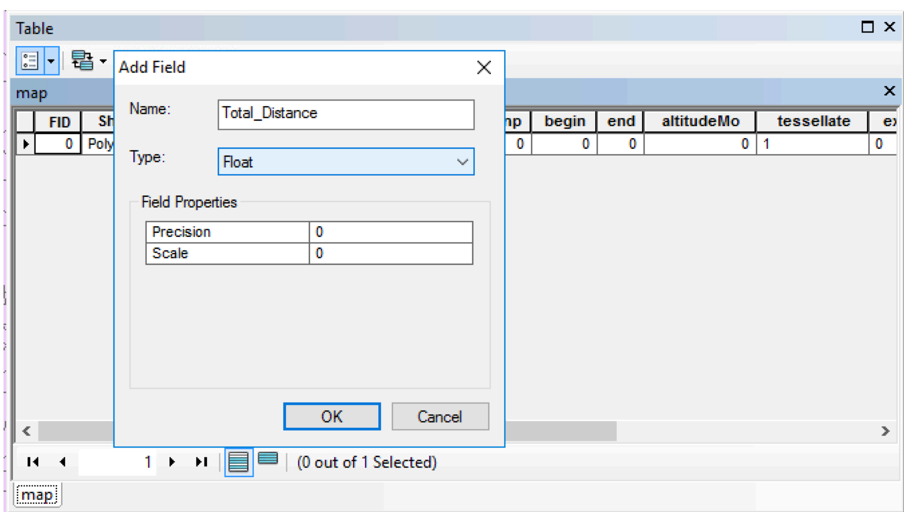
Conclusion

In conclusion, GUI-based software is visually easier to work with than command line, yet GUI-based still needs development. For example, to build a 3D map in Qgis, the 3D view plug-in cannot view the map in layout view i.e. add legend and scale-bar. Using command line, its advantage is building interactive maps, in addition to cleaning and converting large datasets.

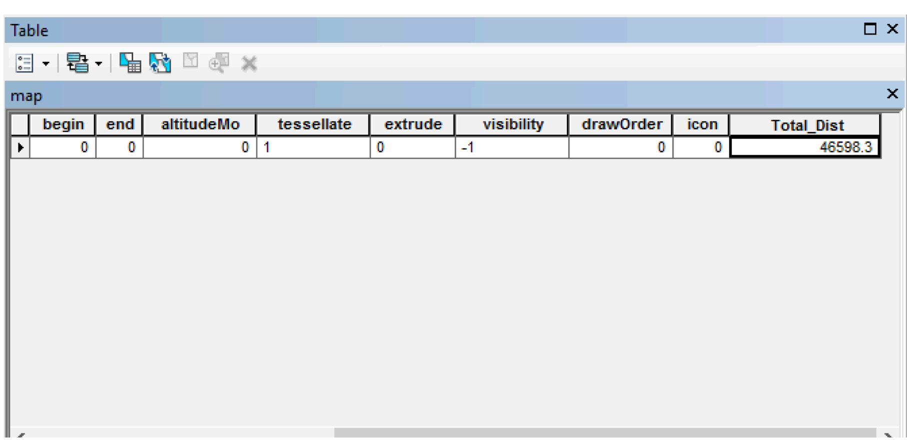
Part 2 - Scavanger Hunt Analysis

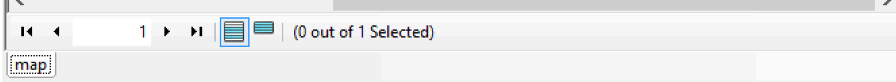
How far did you travel?

Opened up the attribute table for the route and created new field as "Total_Distance", and set it as a float.



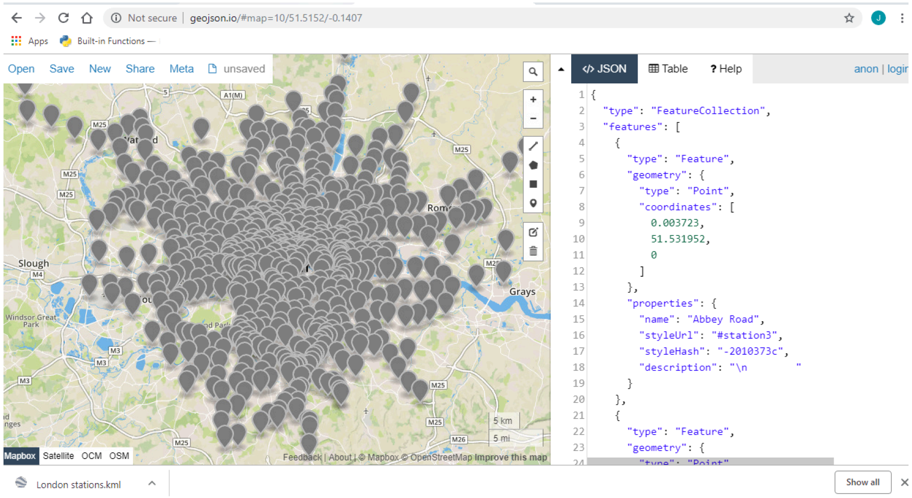
After right clicking on the new field, we can calculate the geometry for Total_Distance. The distance covered was 46598.3 Meters.



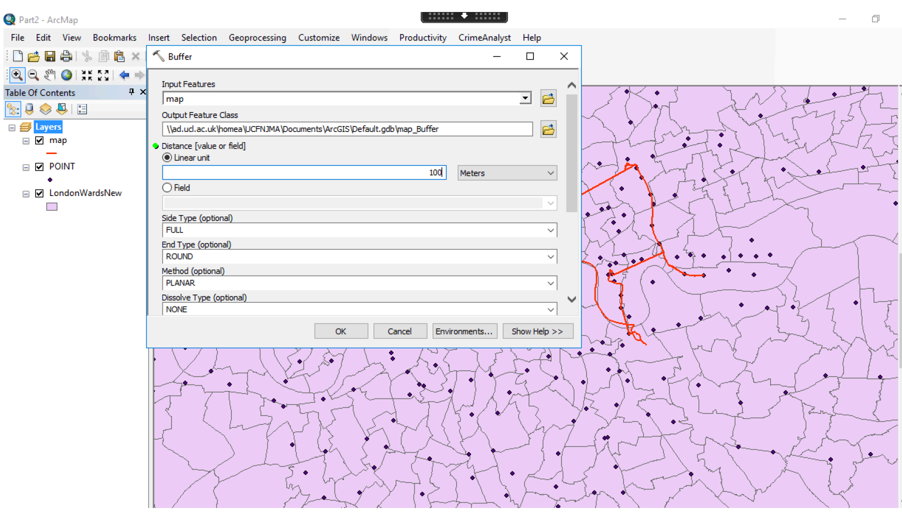


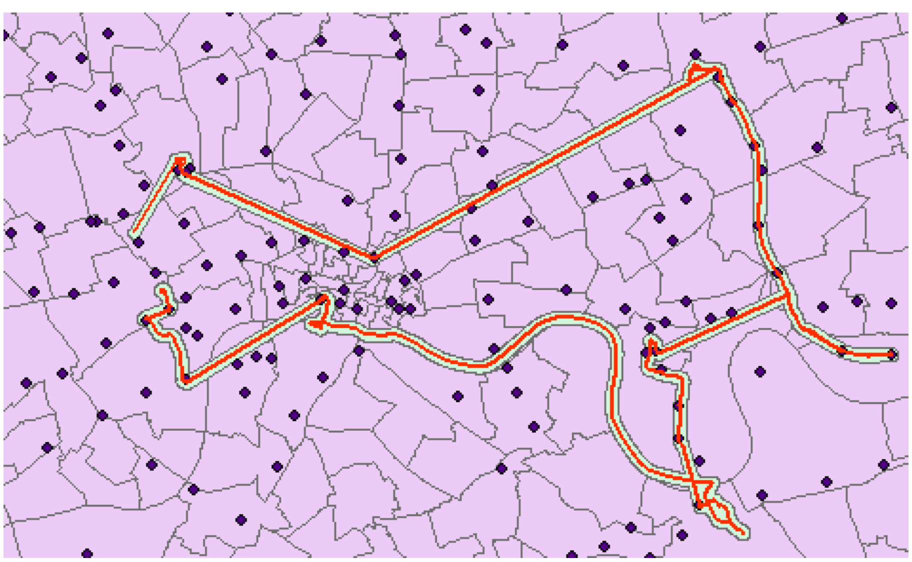
How many TfL station did your route pass within 100 metres distance?

Download the London stations .xml file as a .kml file to the desktop. Open up geojson, upload the downloaded London stations .kml file, and save it as a shapefile. Add the newly downloaded shapefile to the layers tab in ArcMap.

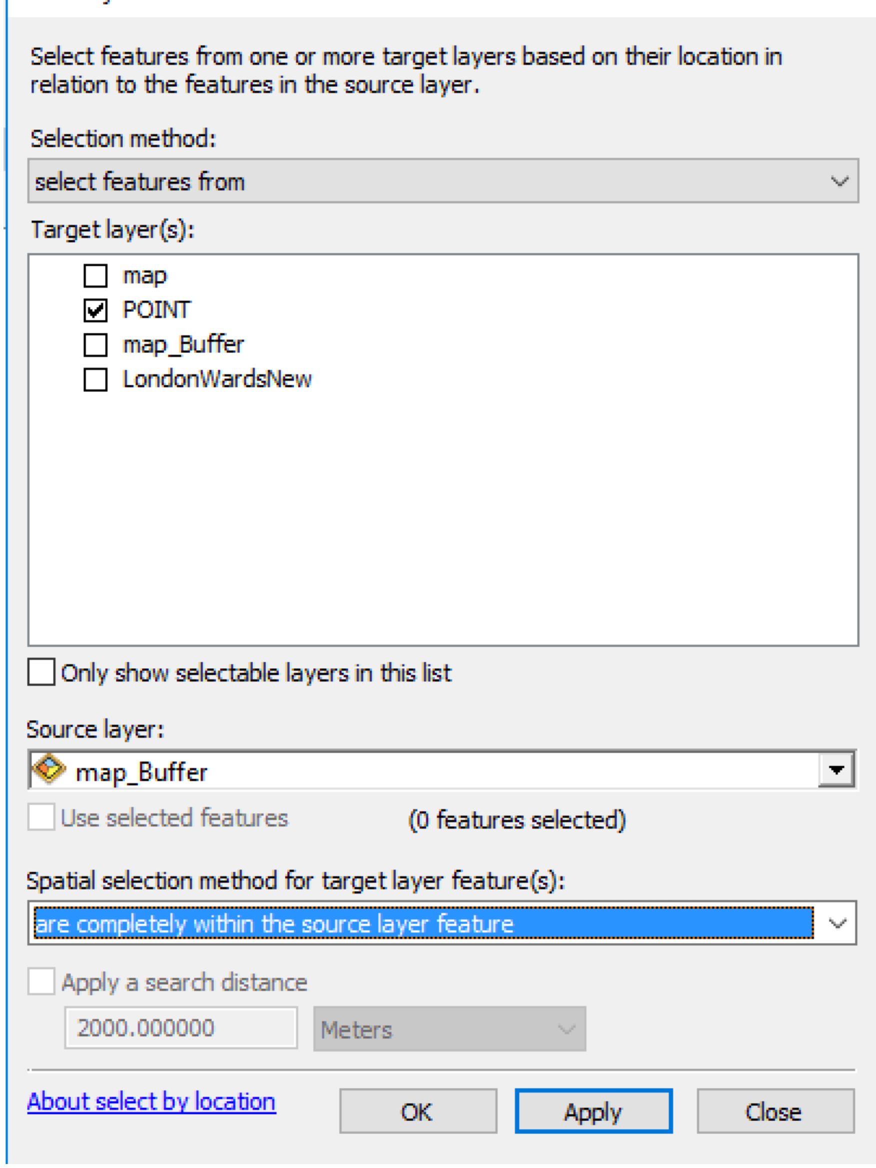


Under geoprocessing, create a buffer around the route layer and set it at 100 meters.

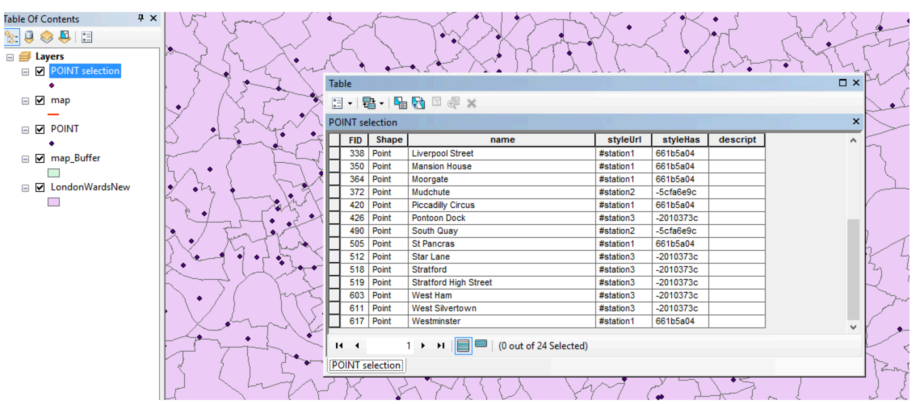




After selecting by location, we can select features from the London stations points. Choose the source layer as London stations and set spatial selection are completely within the source layer feature.



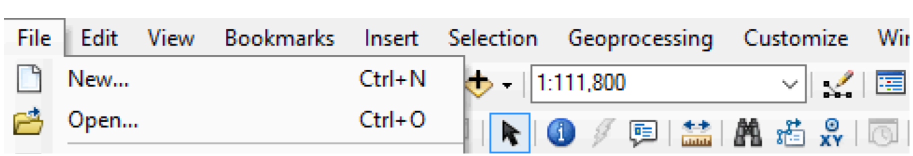
Once the London station points are highlight that represent within 100 meters of the route, right-click the points layer and create a new layer from the selection. A new layer is added to the Layers tab. The attribute table displays the number of points as 24 stations.

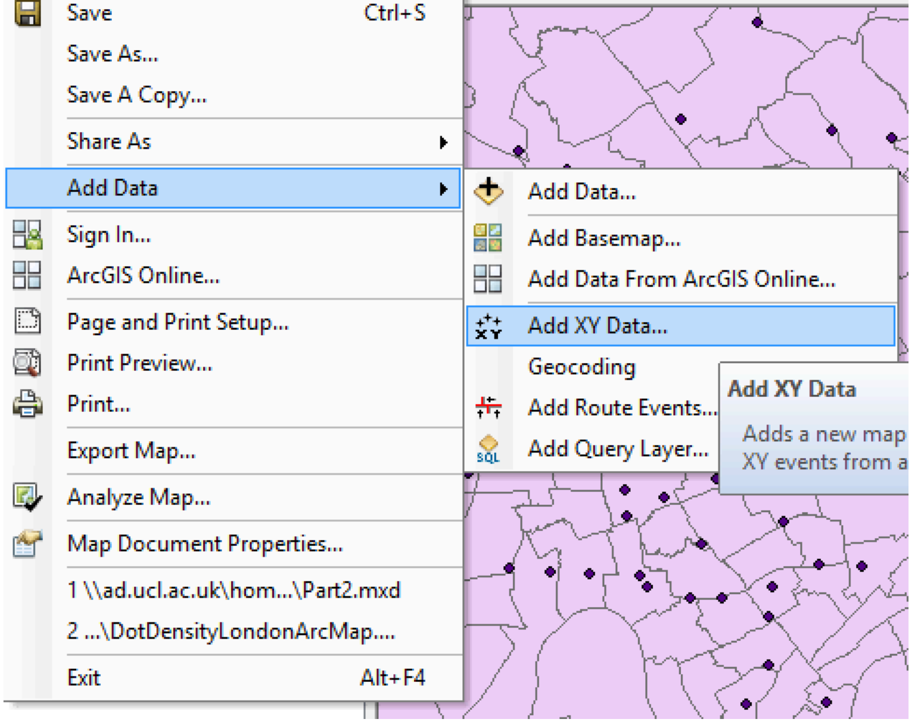


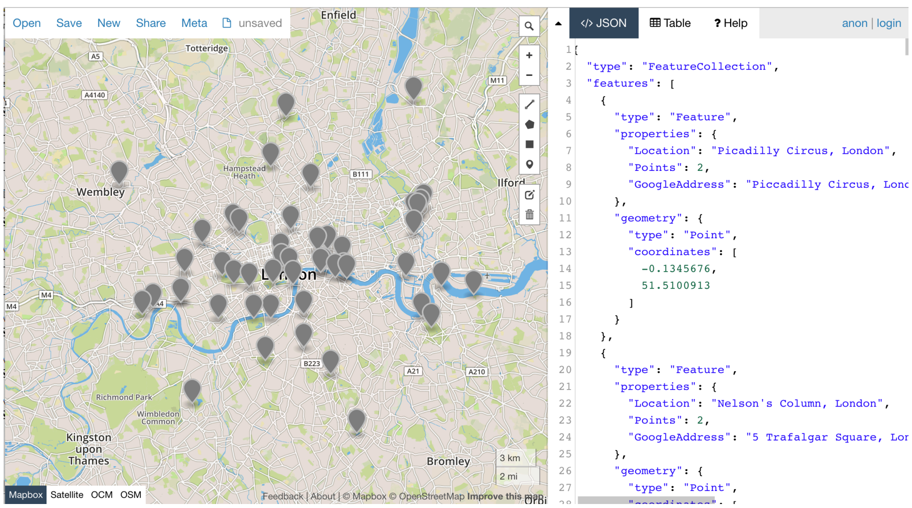
FID	Shape	name	styleUrl	styleHas	descript
338	Point	Liverpool Street	#station1	661b5a04	
350	Point	Mansion House	#station1	661b5a04	
364	Point	Moorgate	#station1	661b5a04	
372	Point	Mudchute	#station2	-5cfa6e9c	
420	Point	Piccadilly Circus	#station1	661b5a04	
426	Point	Pontoon Dock	#station3	-2010373c	
490	Point	South Quay	#station2	-5cfa6e9c	
505	Point	St Pancras	#station1	661b5a04	
512	Point	Star Lane	#station3	-2010373c	
518	Point	Stratford	#station3	-2010373c	
519	Point	Stratford High Street	#station3	-2010373c	
603	Point	West Ham	#station3	-2010373c	
611	Point	West Silvertown	#station3	-2010373c	
617	Point	Westminster	#station1	661b5a04	

How many points did you score based on treasure hunt locations they managed to get within 300 metres of?

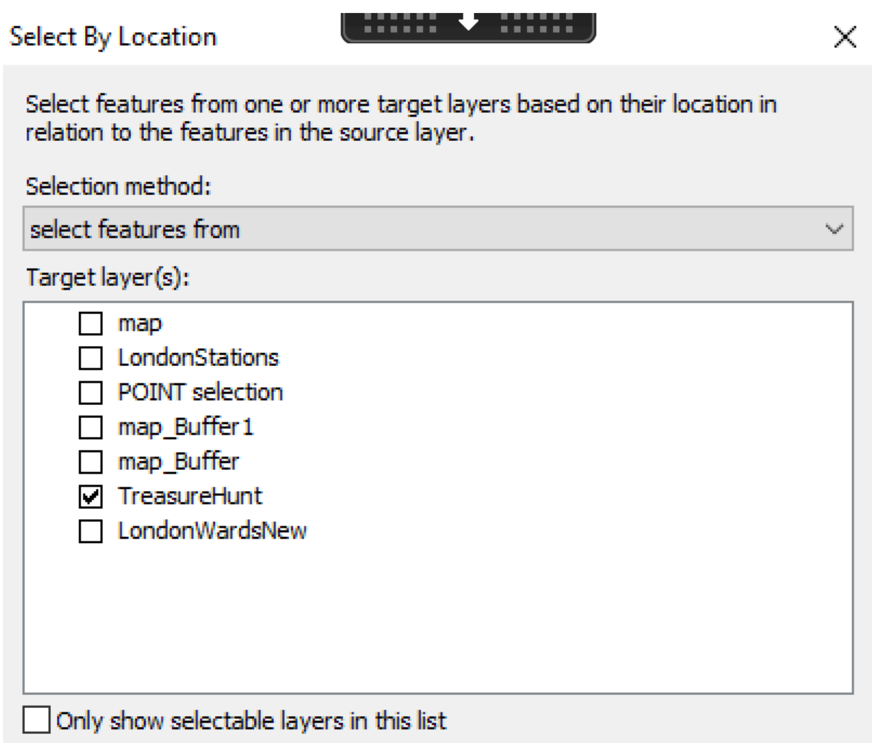
Treasure hunt locations were downloaded in R as a dataframe from the source and written as a .csv to the desktop. Add data into ArcMap as XY data and set longitude and latitude accordingly was projecting the data incorrectly onto the map, even after fixing the projection. Therefore, uploading the .csv file to geojson was necessary to export as a shapefile. The treasure hunt shapefile was added to the Layers tab and displayed as points.

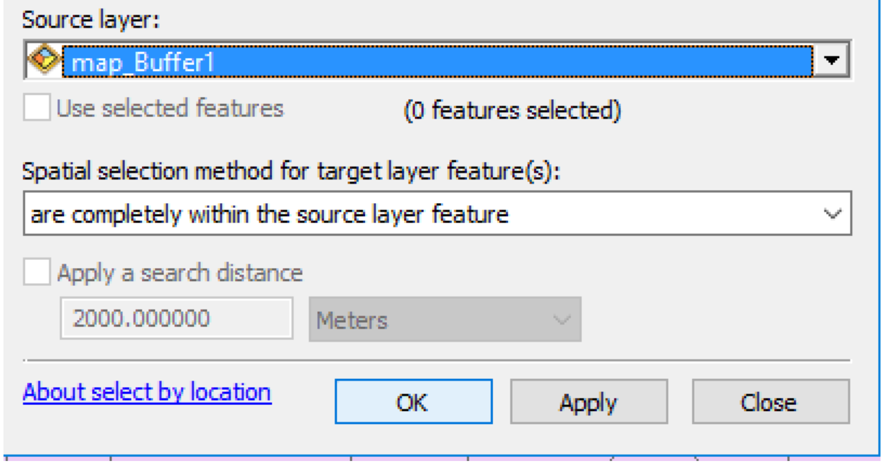




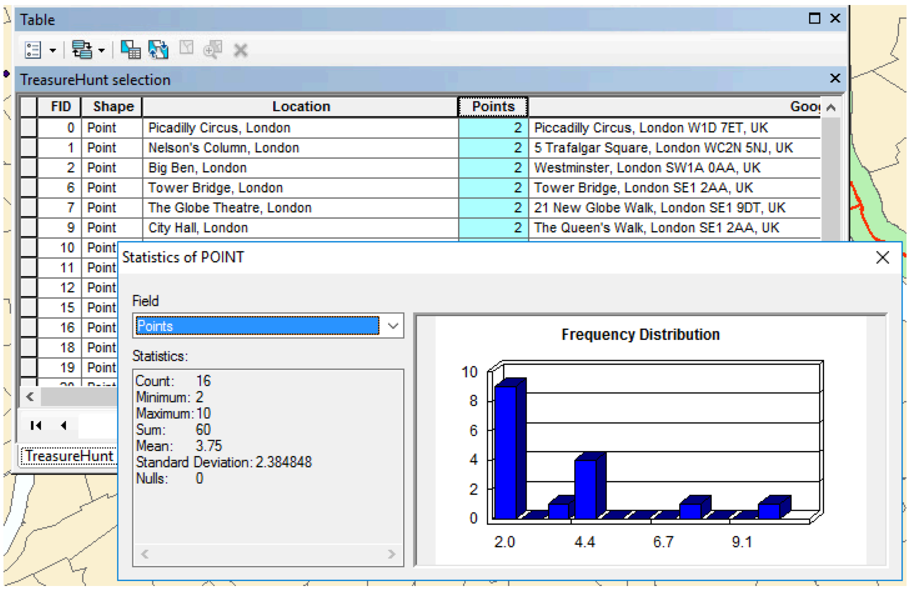


Once again, repeat similar steps above to create a buffer, but with 300 meters instead of 100 meters. From selection by location, select features from TreasureHunt points, set source layer as the newly created buffer, and are completely within the buffer.



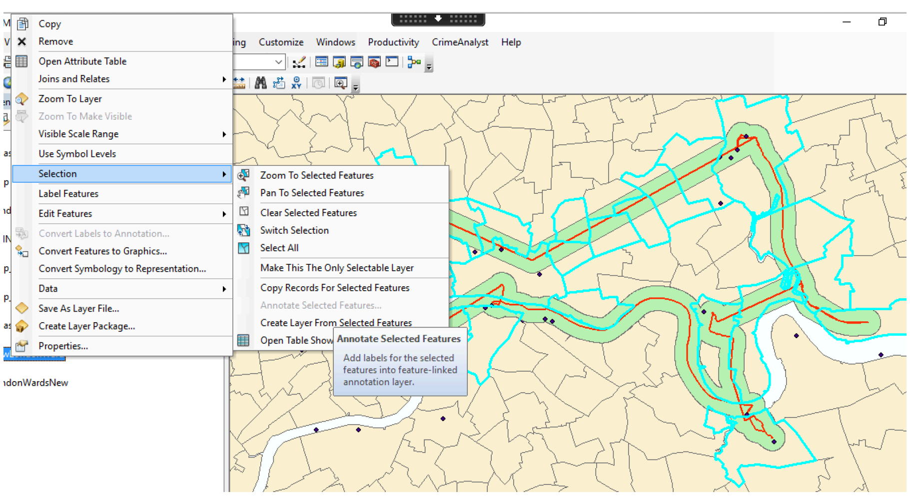


Set the selected treasure hunt locations within the buffer as a new layer. After opening up the attribute table, and selecting the points (representing scores for locations) field, the Statistics tab displays a sum of 60 points scored within 300 meters of the route.

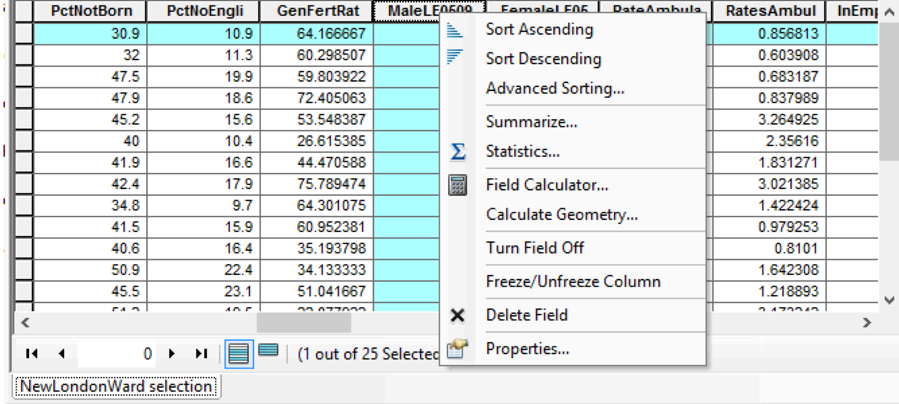


Which Wards did you pass through that had the (a) lowest and (b) the highest rates of Male Life Expectancy?

Select by location tab and setting the source layer as LondonWards intersecting through the route. A new layer was created from the selected wards.

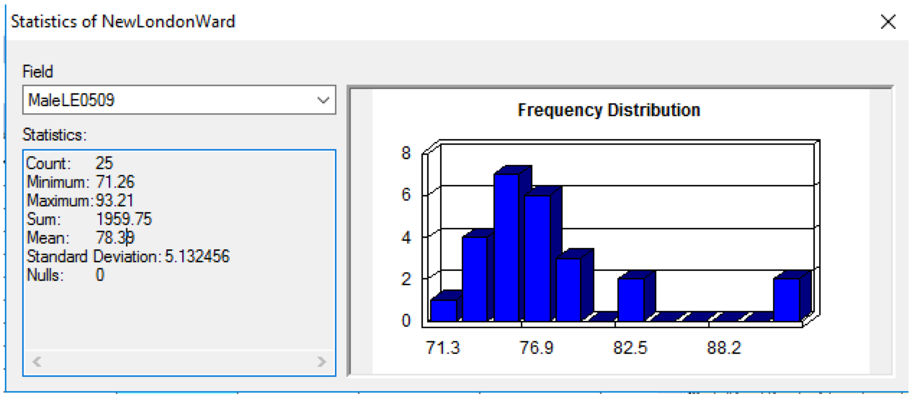


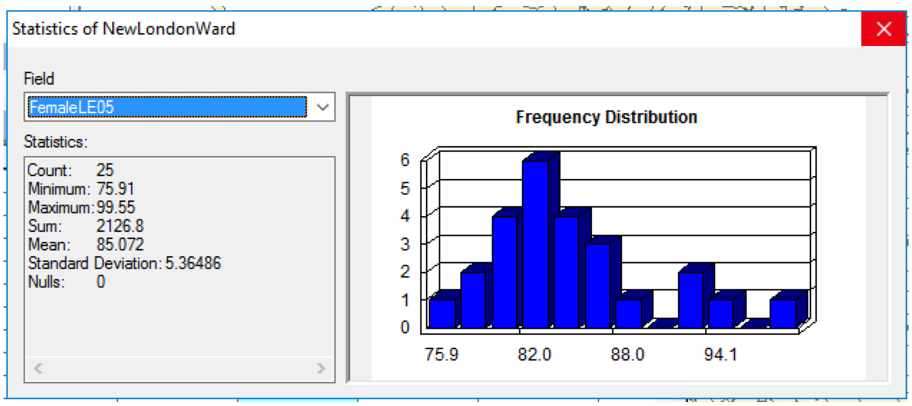
The attribute table can sort fields by ascending or descending giving the ability to see the highest and lowest rates of Male Life Expectancy that are intersecting through the route. The lowest ward, Bow East, has a male life expectancy rate of 71.26 years, while West End has the highest rate of 93.21 years.



Taking the average of all Wards that you passed through, what was the average life expectancy at birth for babies born in those wards along the whole route?

In the attribute table for the intersecting London Wards, we can highlight the Male life expectancy and Female life expectancy individually and clicking the Statistics tab to reveal the averages.

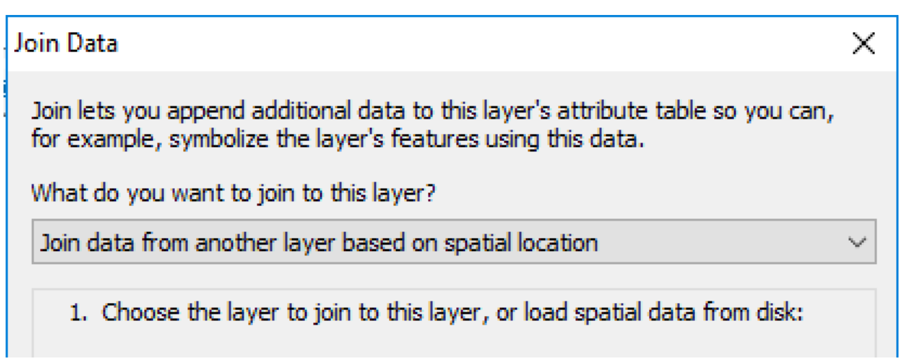


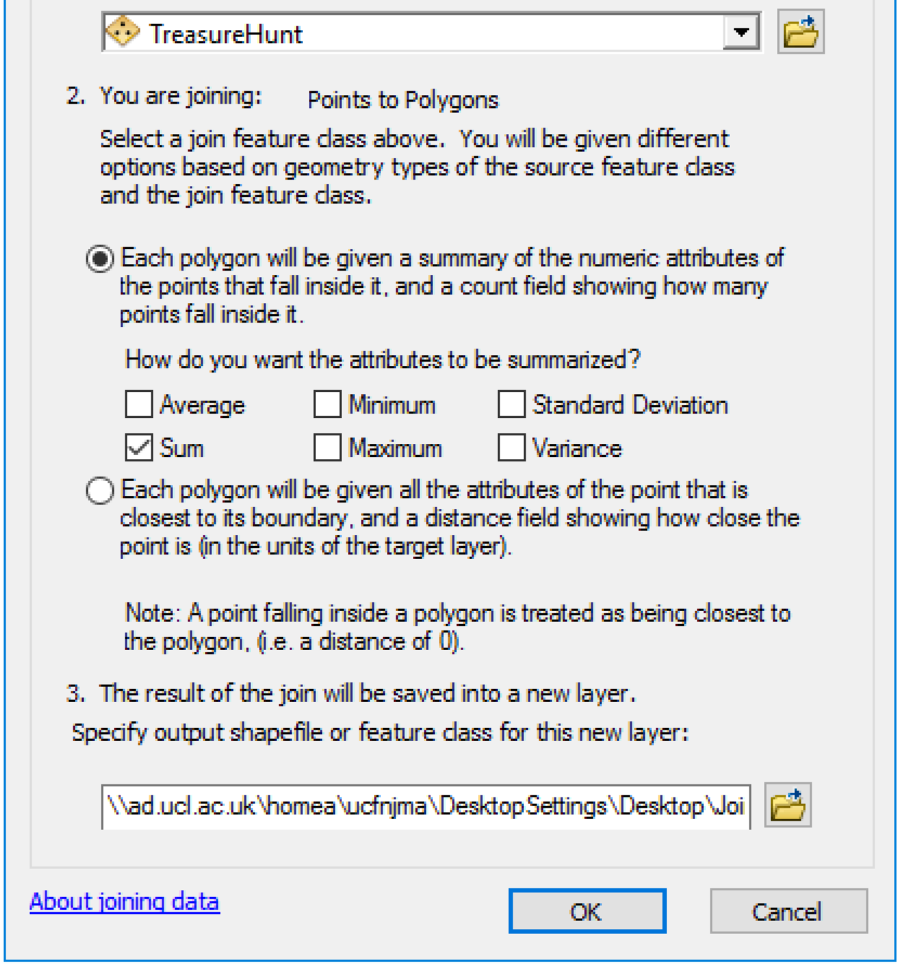


Male life expectancy has a lower average of 78.39 years compared to the average Female life expectancy rate of 85.072 years.

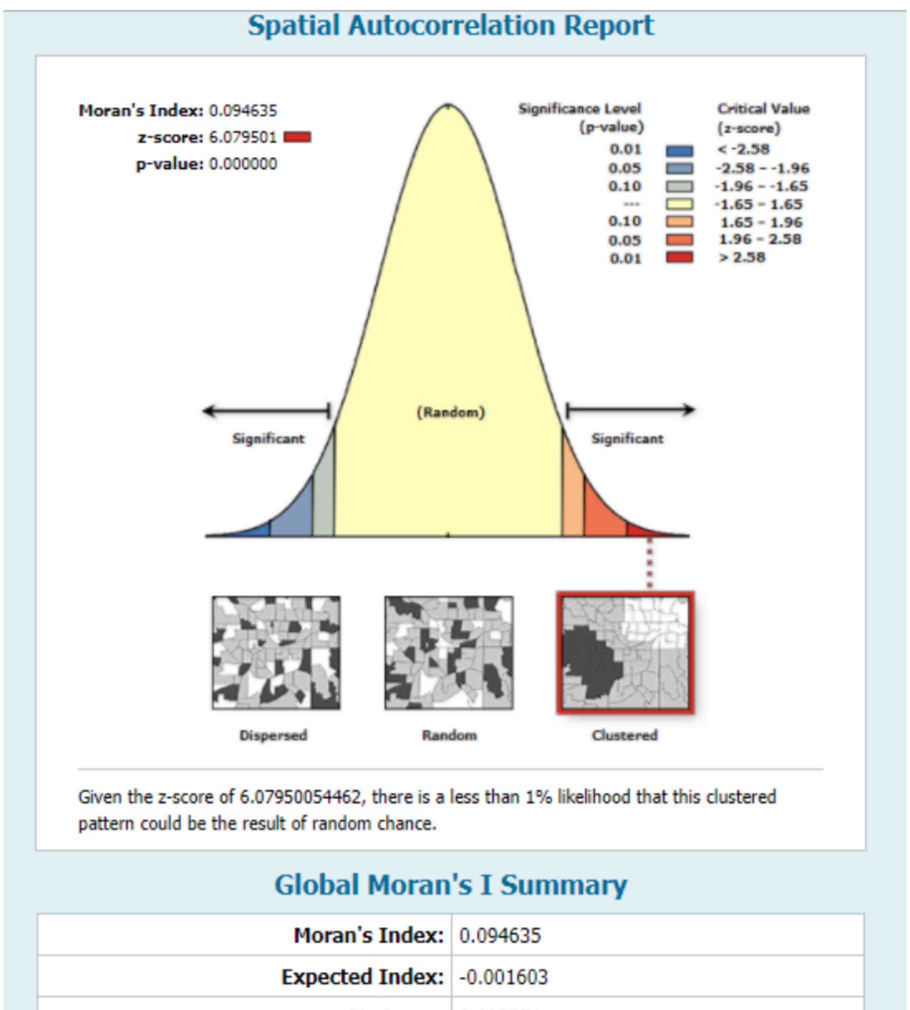
Is there any spatial patterns for CASA Treasure Hunt locations or are they randomly distributed?

For a clustering analysis in ArcMap, first, a spatial join was performed between the treasure hunt locations and London wards.

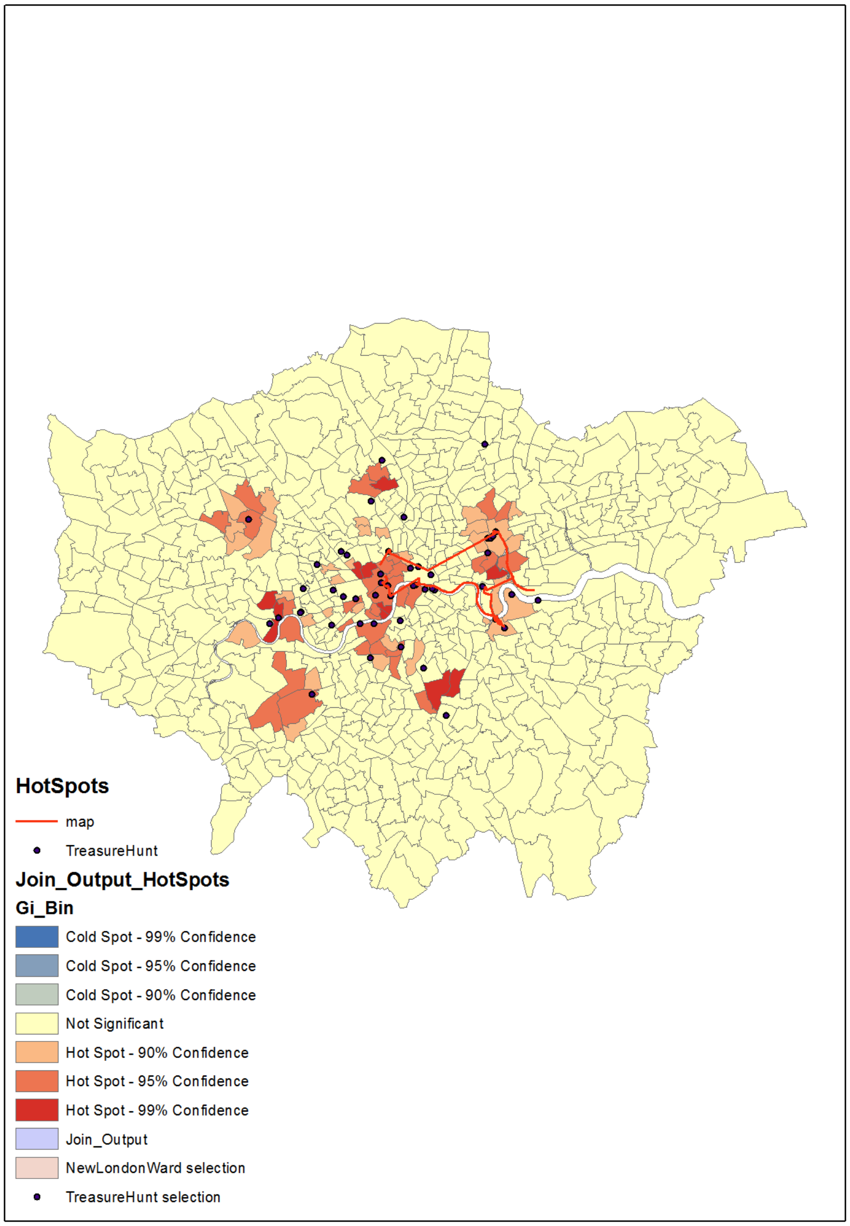




For this analysis, a Morans I and a Hot spot (Getis-ord Gi*) analyses were performed with results displayed below. Moran I test shows our values are clustered and a hot-spot test maps the potential clusters.



z-score:	6.079501
p-value:	0.000000



Part 3 - London, UK Opportunity Index

Introduction

In 2009, The Kirwan Institute conducted an in-depth analysis on neighborhood opportunity for multiple cities and states in the USA. This was the first introduction into “Opportunity mapping”; the use of quantitative methods as a way to visualize and map opportunities that are experiences throughout cities and neighborhoods. Opportunity in this case can be defined as environmental (neighborhoods) conditions that determine access to resources and infrastructure that provide stability and ability to excel in society. Fair access to sustainable employment, quality education, health care, and safe and affordable housing are examples of opportunity. This report explores opportunity mapping for London using the Kirwan Institute’s study as a foundational model. The report will pull data representing indicators of opportunity such as:

Educational variables:
Economic variables:
Neighborhood and Housing quality varibales:

The data will help explore and address the following research topics: What does opportunity look like in London?

Are there disparities and/or inequalities in opportunity based upon geography?

Benefits of research

Analyzing census themed data can provide information to municipalities or institutions that are focused on alleviating the opportunity gap. Making relationships with categories that go hand-in-hand have the benefit to raise questions about neighborhood quality of life—creating these linkages is what alleviates this gap. Equity mapping provides data shown in visual terms which helps municipalities, foundations, housing initiatives, etc… strengthen their organizational missions that target investments that improve neighborhood conditions. Equity mapping can guide policymaking decisions, such as fixing public transit or develop affordable housing in close proximity to employment districts (HUD User, 2015).

Literature Review

Theoretical Overview

Geography of opportunity as presented by Galster and Killen argue spatial differentiation happens in the social systems, markets, and institutions that make up the urban opportunity system of upward (and downward) mobility. Spatial differentiation in cities are arguably growing in areas such as one's socioeconomic status, therefore a larger spatial socioeconomic gap means greater urban inequality. Even though there may be a wide socioeconomic demographic in a metropolitan area, but if it is not creatively spatially diversified and instead is homogenized, the consequences are greater inequality.

Previous Work in the UK

The government of the UK has conducted a similar research study named, "The English Indeces of Deprivation", for the last 3 to 5 years since the year 2000, with 2015 being the latest update. Their study uses similar indicators for their index and mainly focused on the whole country of England rather than just London projected on super-lower output areas.

Methodologies

Data Sources

Data was derived from multiple reliable sources. Boundary data for census wards was downloaded from the https://data.london.gov.uk/, an open data source website developed by the Greater London Authority (GLA). The variables used in the index uses UK census data derived from https://www.nomisweb.co.uk/census/2011 and https://data.london.gov.uk/:

Educational variables:
Average GCSE scores
Percent Level 4 qualified
Unauthorized absence from school
Economic variables:
Employment rate
Job Seekers Allowance claim
Number of jobs (normalized)
Neighborhood and Housing quality variables:
Percent housing owned
Crime rate (normalized)

Z-scores

Calculations and visualizations produced in this project was performed in the statistical coding software: R. The method to create the opportunity index used z-scores as a scale for the variables, and took the average of the z-scores. Z-scores is a method of standardizing data to create a common scale between the observations and reports how many standard deviations away an observation is from the mean value. To ensure the validity of the results of the z-scores, it is important to have the variables used in the calculation to be normally distributed. Skewed data can have an effect on the overall composite index; the more extreme the value, the greater the effect. Although, if the distributions for the data have similar bell-shaped curve, then the z-scores are comparable. Therefore, this report visualizes a set of descriptive statistics such as distributions and density maps for all variables in the index and their summary statistics. Before calculating the z-scores, some variables had to be normalized by population to be used as a comparable variable. The formula and code lines function to calculate the z-score are displayed below:

$$I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t}}{\sigma_{qc=\bar{c}}^{t}}$$

```
z_score <- function(input, colID){

 attach(input)

 data <- colID

 mean <- mean(data)
 sd <- sd(data)

 for(i in 1:length(data)){

   data[i] <-(data[i] - mean)/sd

 }

 return(input <- cbind(data))
}
```

Principal Components Analysis, Correlation Matrix, and Weights

A correlation matrix was calculated to visualize the relationships between the variables used in the opportunity index. For this report, understanding the weights given by each variable in the index was calculated using a Principal Components Analysis (PCA). The PCA analysis helps explain how much of the overall variance is being influenced by.

Results

Below are summary stats for the variables used in the index:

Table 1 Summary statistics for opportunity map variables

```
   Population        AvgGCSE         PctLvl4Qual      UnauthAbsc        Employment
 Min.   : 5110    Min.   :245.0    Min.   :12.50    Min.   :0.2463    Min.   :45.02
 1st Qu.:11197    1st Qu.:332.3    1st Qu.:27.30    1st Qu.:0.8215    1st Qu.:61.85
 Median :12979    Median :343.7    Median :35.50    Median :1.1364    Median :65.95
 Mean   :13078    Mean   :345.8    Mean   :37.66    Mean   :1.1286    Mean   :65.46
 3rd Qu.:14862    3rd Qu.:358.3    3rd Qu.:47.00    3rd Qu.:1.4105    3rd Qu.:69.24
 Max.   :23084    Max.   :409.1    Max.   :68.70    Max.   :2.4675    Max.   :81.48
    No_Jobs          JSA_Claim       PctHousingOwned    CrimeRate          LogNoJobs
 Min.   :   600    Min.   : 0.00    Min.   :11.90    Min.   :  25.75    Min.   :-1.2767
 1st Qu.:  2100    1st Qu.: 6.45    1st Qu.:34.30    1st Qu.:  64.09    1st Qu.:-0.7777
 Median :  3500    Median :10.43    Median :48.30    Median :  84.83    Median :-0.5795
 Mean   :  7091    Mean   :11.73    Mean   :49.67    Mean   : 101.05    Mean   :-0.5274
 3rd Qu.:  6000    3rd Qu.:15.72    3rd Qu.:64.40    3rd Qu.: 107.57    3rd Qu.:-0.3463
 Max.   :382700    Max.   :48.67    Max.   :90.80    Max.   :2100.20    Max.   : 1.7151
```

We can visualize each variable as choropleth density maps:

Fig. 1 Average GCSE scores for London wards

Fig. 2 % Level 4 Qualification

Fig. 3 Unauthorized Abscences

Fig. 4 Employment rate

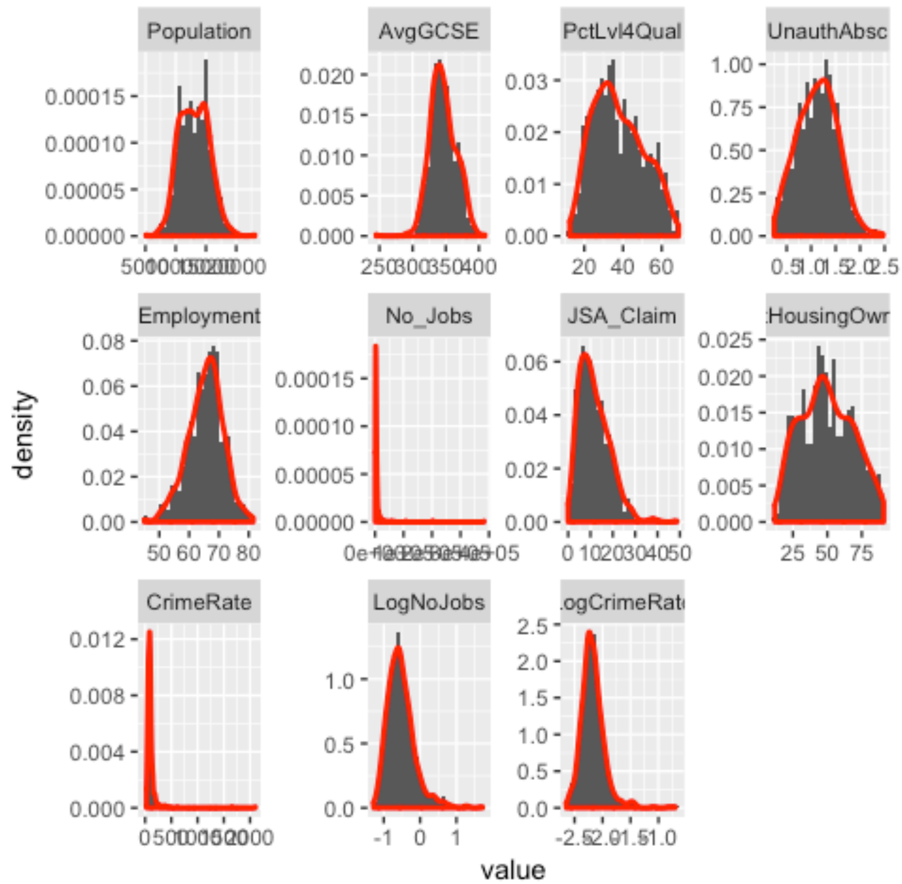
Fig. 5 Number of jobs per London ward

Fig. 6 Crime rate

Fig. 7 Job seekers allowance claims

Fig. 8 % of houwing owned

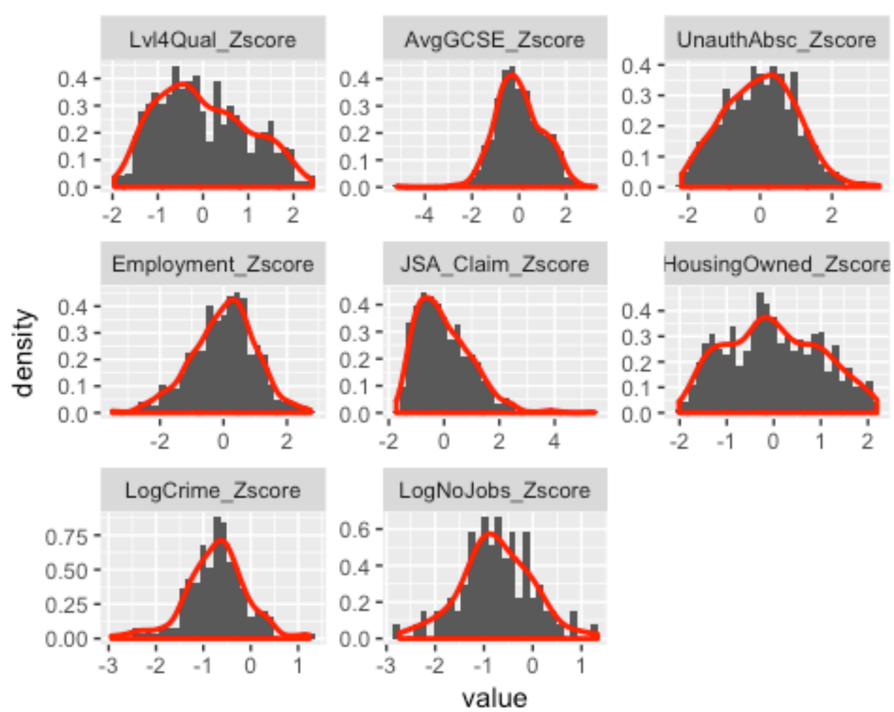
Plotting distributions is another way to visualize the patterns we see in the values for the variables in the maps above. Distributions below are represented across 625 census tracts in London. Many variables are normally distributed following the bell-curve shape, yet some variables remain skewed. Normalized variables were logged, using a log10() transformation, to visualize the distributions clearer.



Plot 1 Distributions of variables in index

Opportunity Map

The visualization below shows distribution charts for the z-scores of the variables used in the index. Many of the variables in the distributions fall below the mean of the scores.

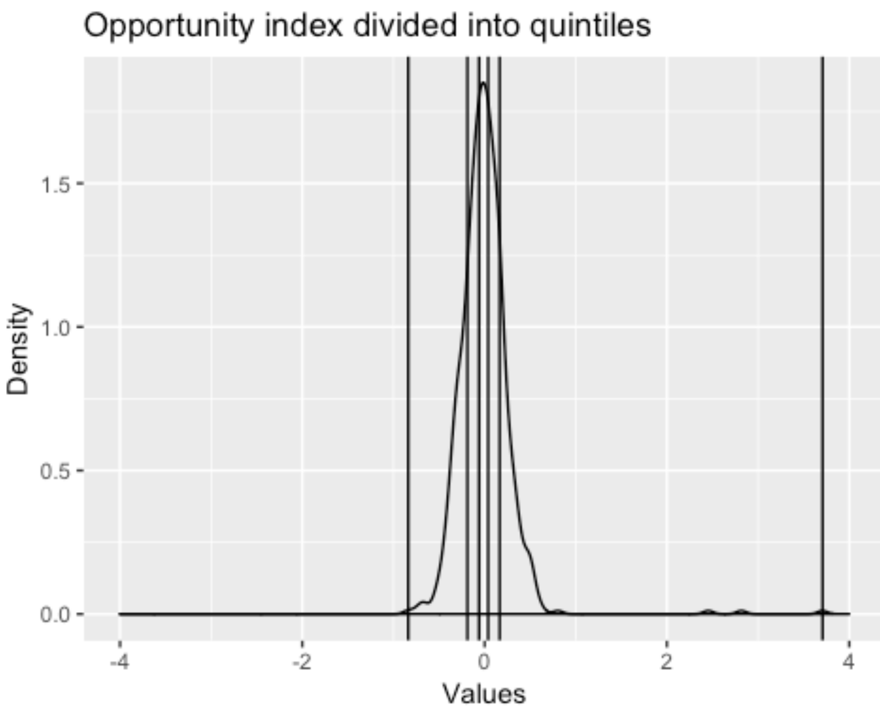


Plot 2 Distributions of variables after z-score calculations for the index

Once the z-score were calculated for all the variables, and the average of them all to represent the index, we can visualize the index as an interactive opportunity map. The user is able to click on each spatial census tract to identify the correlating ward name. Five categories were created, in conjunction to the Kirwan Institute study, to represent the level of opportunity - the darker-shaded regions represent higher opportunity, while lower-shaded regions represent lower opportunity. Directly below the map we can visualize the distribution of the index.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.83770	-0.15637	-0.01406	0.00000	0.13292	3.70684

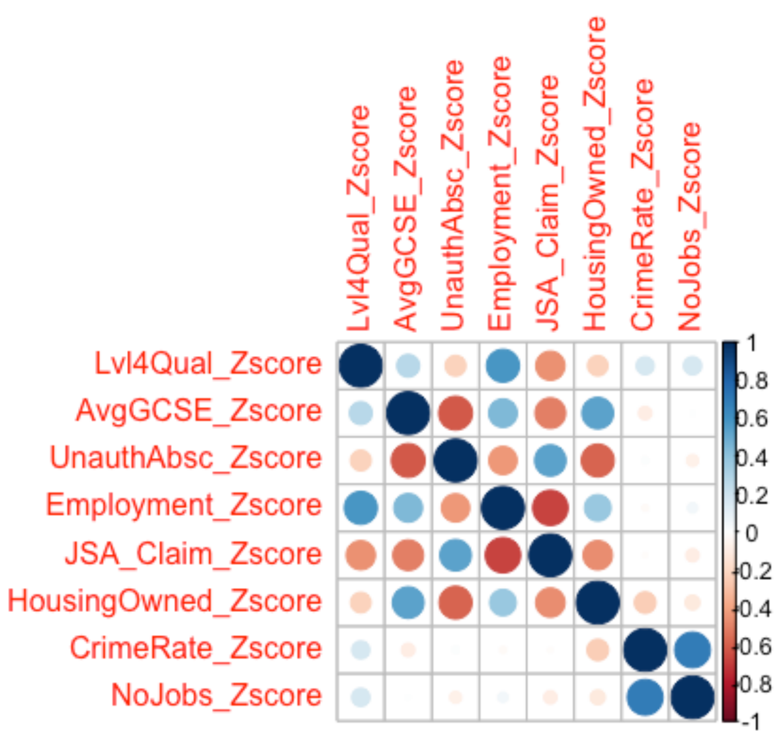
Fig. 9 Opportunity map for London wards



Plot 3 Index distribution

Correlation Matrix

The correlation matrix below shows the relationships between the variables used in the index.



Plot 4 Correlation matrix of variables in index

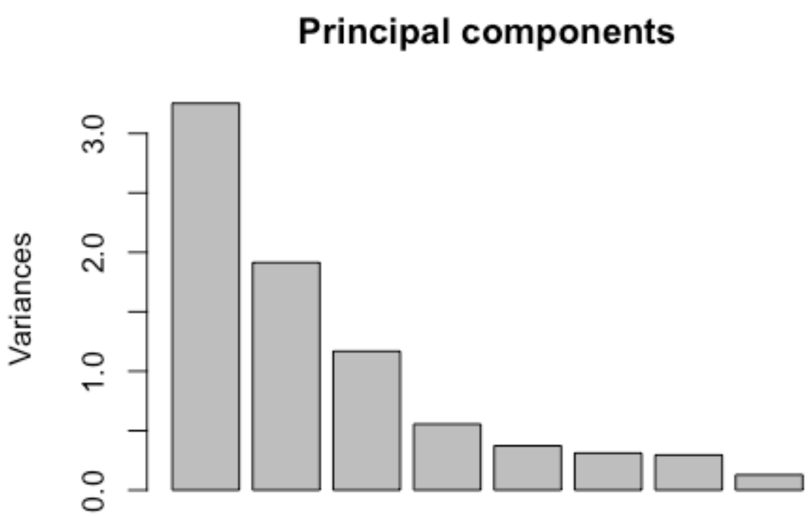
Principal Components Analysis (PCA)

Another way to see the variation in the index is conducted in the PCA results below. The first two principal components weight heavier in effecting the variance of the final index.

Table 2 Principal Compenent Analsysis

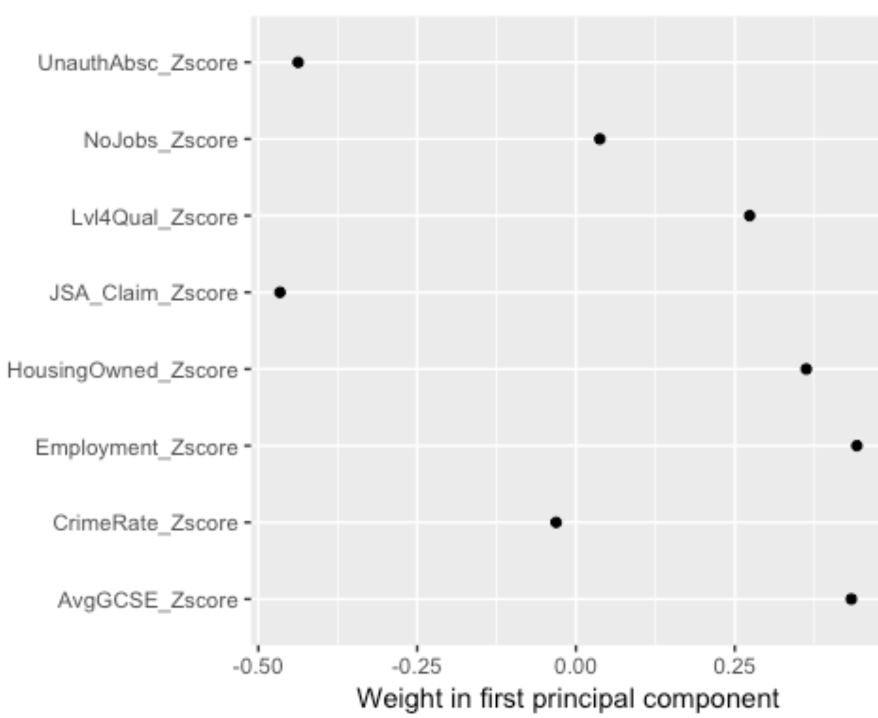
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.8038	1.3835	1.081	0.74533	0.61051	0.55803	0.54348	0.35906
Proportion of Variance	0.4067	0.2393	0.146	0.06944	0.04659	0.03893	0.03692	0.01612
Cumulative Proportion	0.4067	0.6460	0.792	0.86145	0.90804	0.94696	0.98388	1.00000

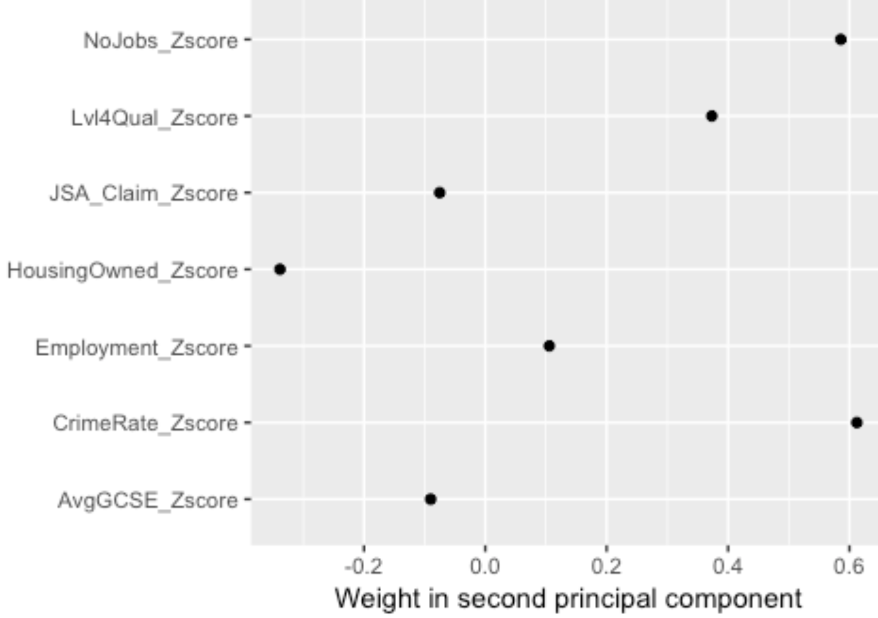


Plot 5 Principal Compenent Analsysis results as a bar graph

Weights



Plot 6 Weights of variables in PCA1



Plot 7 Weights of variables in PCA2

K-means Clustering Analysis Plot and Map of clusters

A k-means test was ran twice, so that the three central london wards could be removed as outliers.

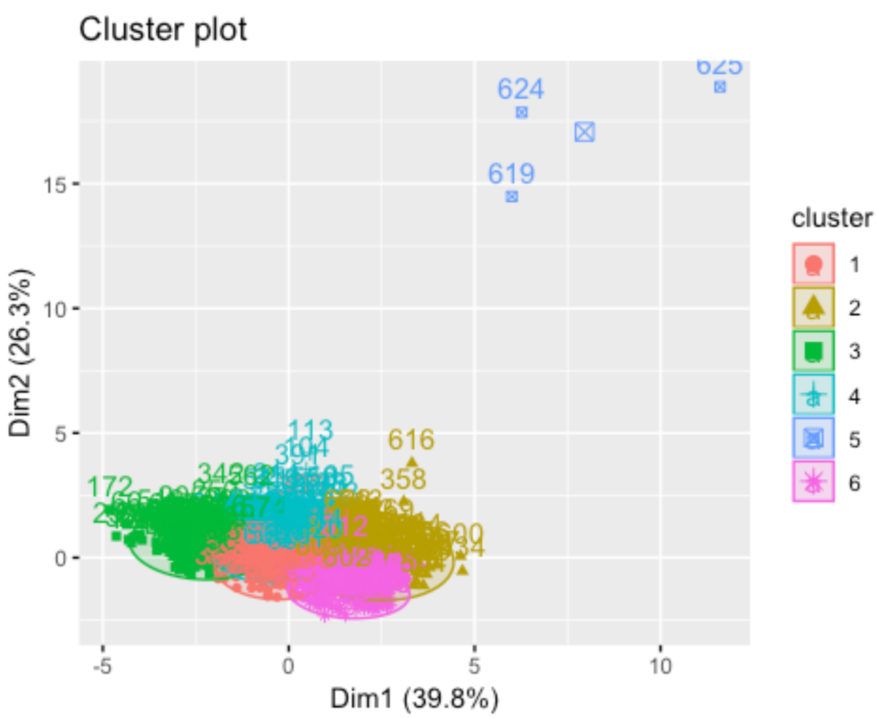
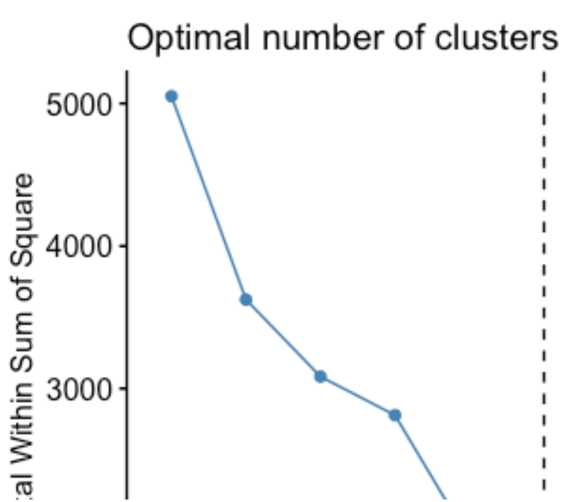
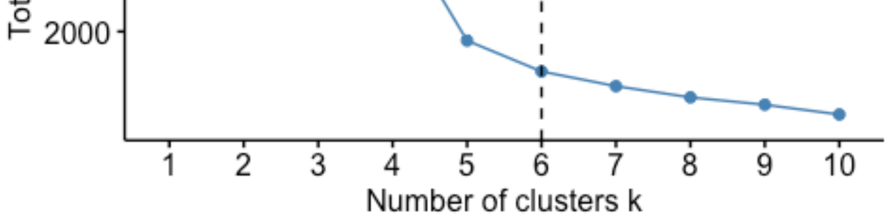


Fig. 10 K-means clustering (6 clusters)





Plot 8 K-means optimal clusters

	cluster	Lvl4Qual_Zscore	AvgGCSE_Zscore	UnauthAbsc_Zscore	Employment_Zscore	JSA_Claim_Zscore
1	1	-0.7455199	-0.1611365	0.1102871	-0.2039717	0.048610499
2	2	1.4253157	0.7999539	-0.4901852	1.1910952	-0.971637177
3	3	-0.7389497	-0.8345700	0.9191364	-1.2369778	1.250694762
4	4	0.7579927	-0.5218331	0.4015457	0.2244286	-0.008883548
5	5	1.5534107	-0.3576896	-0.9903582	0.7554306	-0.566263559
6	6	-0.2913317	1.0401431	-1.1897567	0.4702610	-0.721733379

	HousingOwned_Zscore	CrimeRate_Zscore	NoJobs_Zscore	Index
1	0.46303209	-0.14484665	-0.10518069	-0.09234071
2	0.06784558	-0.01063881	-0.01439115	0.24966976
3	-0.79802904	-0.01941697	-0.10199541	-0.19501346
4	-0.91139996	0.17649715	0.06589508	0.02303033
5	-1.01432896	12.41662428	12.14228593	2.99238890
6	1.33800339	-0.29424768	-0.11756189	0.02922202

Table 3 Mean of clusters

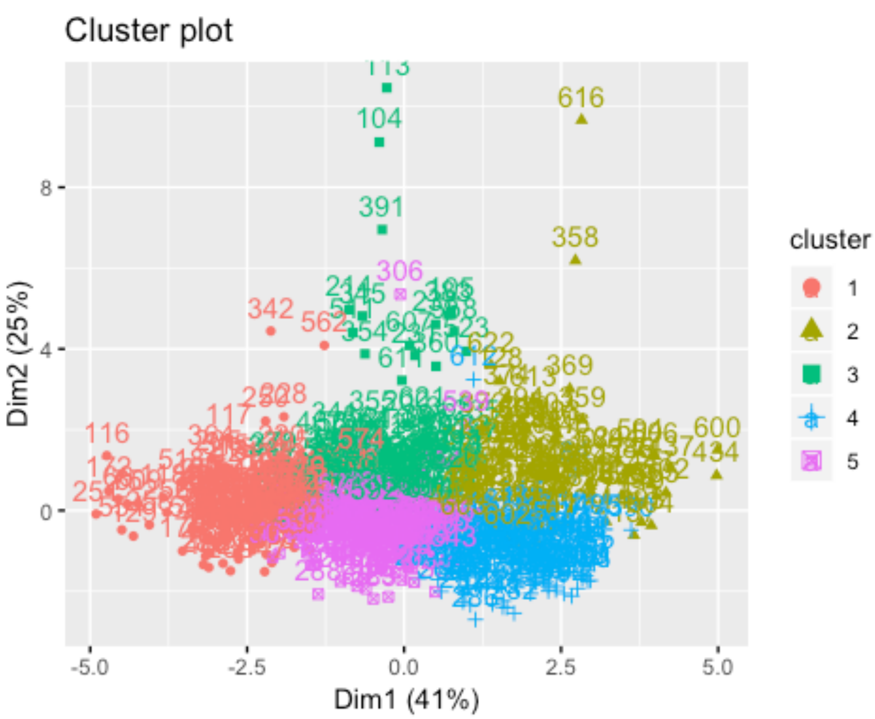
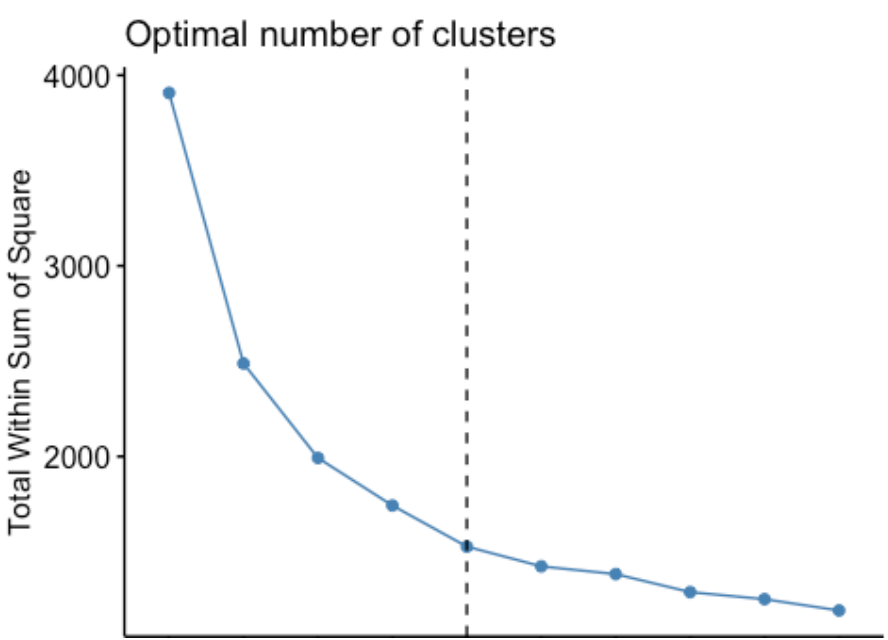
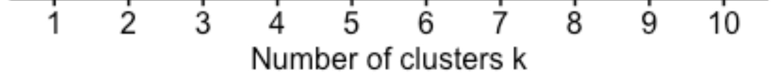


Fig. 11 K-means clustering (5 clusters)





Plot 9 K-means optimal clusters

	cluster	Lvl4Qual_Zscore	AvgGCSE_Zscore	UnauthAbsc_Zscore	Employment_Zscore	JSA_Claim_Zscore
1	1	-0.7389497	-0.8345700	0.9191364	-1.2369778	1.250694762
2	2	1.4253157	0.7999539	-0.4901852	1.1910952	-0.971637177
3	3	0.7579927	-0.5218331	0.4015457	0.2244286	-0.008883548
4	4	-0.2913317	1.0401431	-1.1897567	0.4702610	-0.721733379
5	5	-0.7455199	-0.1611365	0.1102871	-0.2039717	0.048610499

	HousingOwned_Zscore	CrimeRate_Zscore	NoJobs_Zscore	Index
1	-0.79802904	-0.01941697	-0.10199541	-0.19501346
2	0.06784558	-0.01063881	-0.01439115	0.24966976
3	-0.91139996	0.17649715	0.06589508	0.02303033
4	1.33800339	-0.29424768	-0.11756189	0.02922202
5	0.46303209	-0.14484665	-0.10518069	-0.09234071

Table 4 Mean of clusters

Fig. 12 Map of k-means clusters

Discussion

The opportunity map for London (Figure 9) is separated into five categories ranging from very low, regions shaded in light-yellow, to very high, regions shaded dark-red, opportunity. The index has a minimum of -0.837 (lowest opportunity) for King's Cross ward, and a maximum of 3.70 (highest opportunity) which is the City of London ward. The index has a mean of 0 and about 53% of London wards fall below that. Inequalities exist throughout London, but are not visually concentrated or clustered in a pocket, which is positive to avoid concentrating poverty, with the exception of the three highest opportunity wards in central london. Opportunity tends to decrease in the direction away from the city center. Central London can be seen as experiencing high inequalities. Within a small geographic location (minutes away driving), one can go from the highest opportunities in London (West End, etc..) to the lowest opportunity in London (King's Cross). We can also see there are "islands" of low opportunity neighborhoods surrounded by higher opportunity neighborhoods, such as Roehampton and Putney Heath, Hoxton, Evelyn, and a few others. The distribution for the index came out normally distributed with central london neighborhoods exhibiting three extreme cases that typically would be labeled as outliers and deleted, but are necessary to keep in the final index calculation, as they are part of the city.

Re-examining the distributions of the variables that go in the final index, we can see some of the variables fall below the mean of their scores such as JSA claims and Number of jobs. Mapping the variables visualizes the distribution of variables across the London wards and their patterns can be reflected in the correlation matrix. Average GCSE scores and % of housing owned are highly correlated with each other and their spatial patterns across London. Crime rate and Number of jobs are extremely correlated, yet this case is a bit extreme considering the spatial patterns visualize most of the data occurs in central london neighborhoods i.e. City of London has normalized number of jobs of 51 while the second most, in West End, is 19. The results of the PCA show that first two principal components weight heavier in impacting the variance of the final index. The first principal component (PC1) explains about 41% of the overall variance in the data, while PC2 explains about 24% - the first two bars in Plot 5. Plots 6 and 7 visualize the weights of each variable in PC1 and PC2. In PC1, crime rate and number of jobs have little influence (weight almost at 0), while average GCSE scores and employment are strong positive weights. JSA claims and unauthorized absences are strong negative weights and counterbalance the positive ones, which parallels roughly the patterns in the maps above.

The K-means clustering results of the composite index calculated an optimal number of 6 clusters of London Wards as seen in Plot 8. It is obvious to see that the three wards with the highest z-scores fall into a cluster of their own by far. After removing those data points and running another k-means test with 5 clusters, we have a closer look. In cluster 3 in table 3, it has the lowest mean of -0.2 for its index. Level 4 qualifications, average GCSE scores, and employment are all very low averages. For example, Bunhill falls in cluster 3 (figure 12) and has a very low average gcse score of 317 and a low employment rate of 57%. In contrast, cluster 2 has the highest cluster mean for average gcse score of 0.8 and employment rate mean of 1.19. Marylebone high street has an average gcse score of 372 and an employment rate at 69%. In the same cluster, but in contrast, Hampstead Town has a low gcse score of 322, but a high employment rate of 73%. This variance can be explained by the high volume number of jobs in Hampstead Town of 10,800.

Conclusion

The limitations in the data are not robust, but do affect the composite index. If the variables in the index data are not normally distributed then the z-scores can be harder to interpret across the board. Incorporating qualitative data such as images of neighborhoods or interviews with local residents will help explain *why* a neighborhood is experience low or high levels of opportunity. Further research on the topic would include an in-depth analysis of the k-means results. Running regression models and a t-test on the index would help explain the influence of the relationships between the variables to the index. The opportunity map can be used as a tool for local municipalities in addressing local urban issues that require attention.

References

Galster, George C., and Sean P. Killen. “Housing Policy Debate.” The Geography of Metropolitan Opportunity: A Reconnaissance and Conceptual Framework. Knowledge Plex. The Urban Institute, 1995. Web. May 12, 2016. http://content.knowledgeplex.org/kp2/cache/kp/788.pdf

Hannah Ritchie and Max Roser. “Urbanization”. Published online at OurWorldInData.org. September 2018. Retrieved from: ‘https://ourworldindata.org/urbanization’

HUD User. ”Equity Mapping and the Geography of Opportunity | HUD USER." PD&R Edge Home Page. HUD User, 2015. Web. May 12, 2016. https://www.huduser.gov/portal/pdredge/pdr_edge_featd_article_042114.html

Reese, Jason, and David Norris. “The Geography of Opportunity: A Framework to Boost Collaborative Efforts and Expand Opportunity across the Greater New Orleans Region.” Kirwan Institute. 2012. Web. May 12, 2016. http://kirwaninstitute.osu.edu/ki-gnof-opportunity.pdf.