UCL DEPARTMENT OF GEOGRAPHY

YEAR 2016-17

EXAM CANDIDATE ID:	**QRSH1**
MODULE CODE:	**GEOGG125**
MODULE NAME:	**Principals of Spatial Analysis**
COURSE PAPER TITLE:	**Understanding London using Open Data**
WORD COUNT:	**1963**

Are you registered as dyslexic with UCL Student Disability Services (SDS) and been given labels to 'flag' your written work

~~YES~~ / NO *(please delete as applicable)*

Understanding London using Open Data

Introduction

This report covers the multivariate analysis of demographic, economic, and social indictors within Greater London, focusing on exploring the factors influencing the crime rate per thousand population at both borough and ward scales.

The aim of this report is to analyse crime rates in London, determine with which other variables a statistically significant relationship exists, and create an estimation model based on the conclusions. Initially, correlation analysis will be conducted between all available variables at a borough scale, and indicators showing strong positive or negative correlation with crime rate will be selected for further analysis. The selected indicators must also be recorded on a ward scale to allow direct comparison between the two datasets.

Regression analysis will be conducted using the selected variables to determine which factors best explain the variation in crime rate across Greater London boroughs, with the aim of combining these statistics to create an estimation model for crime rate across London. The process will then be repeated at a ward scale to determine whether the same model can be applied to crime rate measured on a smaller scale. The statistical analysis detailed in this report is carried out using R.

This data used for this analysis has been obtained from the London Datastore (London Datastore, 2015). The borough profile dataset comprises 74 numeric demographic, economic, social, and environmental indictors within Greater London over 33 boroughs, whilst the ward profile dataset comprises 64 variables covering mainly demographic indictors over 658 wards within Greater London.

Analysis

Variable Selection

Initial correlation analysis between crime rate and the remaining indicators on a borough scale shows Pearson correlation coefficients ranging from 0.86 to -0.80 (Figure 2), with the greatest positive correlation occurring between crime rate, and the average Public Transport Accessibility (PTA) score (2014). This indicates that the higher the PTA score of the borough, the higher the crime rate. The greatest negative correlation exists between crime rate, and the percentage of homes being bought by mortgage or loan, although this information is not available on a ward scale. Other notable correlations exist between crime rate and the number of jobs per area (2013), and the average number of cars per household (2011), with correlation coefficients of 0.85 and -0.75 respectively. These indicators have equivalent measurements on a ward scale and so have been selected for further analysis. The relationships between these variables are further explored in the correlation analysis section of this report.

Univariate Analysis

The four variables selected for analysis in conjunction with crime rate appear normally distributed on a borough scale with varying degrees of weak positive skew (Figure 1), with the exception of total number of cars per household which shows a uniform distribution. On a ward scale, PTA score and the total number of cars per household show a normal distribution with weak positive skew. Both crime rate and the number of jobs per area show significant outliers falling beyond the upper end of

the interquartile range. To normalise the dataset they have been plotted as the logarithm of the recorded values. This clearly shows a log-normal distribution, also with weak positive skew.

The maximum crime rate recorded on a borough scale is 212.4, compared to the 1212.0 measured for the ward of West End in Westminster. Values of greater than 1000 indicate that more than one crime is recorded per resident. Similarly high values of 990 and 656.4 are recorded for the wards of St. James's (also in Westminster), and the City of London respectively. The next highest recorded crime rate is 314.0, which has been recorded for Bloomsbury in Camden. The PTA score for these wards is also high and ranges from 7.6 to 8.0, with 8.0 being the highest recorded for any ward.

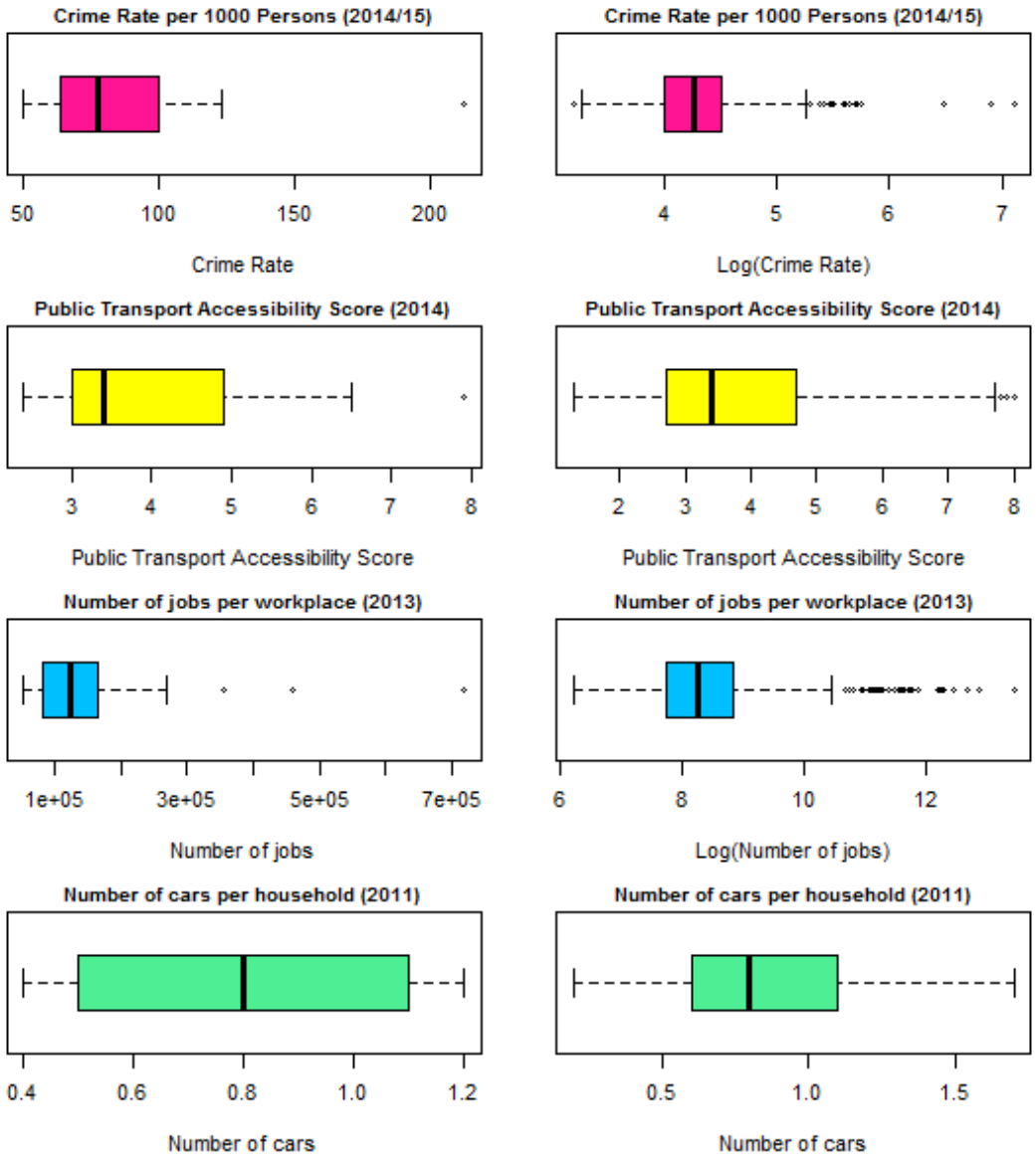

Figure 1: Box and whisker plots showing statistical characteristics of the demographic variables analysed in this study at both borough (left) and ward (right) scales. All variables, except number of cars per household, per borough, show normal or log-normal distributions and natural to weak positive skew. Normal distributions are required for successful modelling through least-squares analysis. Most outliers, if any, occur beyond the 3rd quartile.

Correlation Analysis

Initial correlation analysis, as detailed previously, was carried out using the Pearson correlation test. Each relationship is statistically significant, with P-values of well below 0.001.

As mentioned previously, crime rate shows high positive correlation with the other variables when analysing the relationships using a Pearson's correlation test at borough scale. On a ward scale these factors also correlate well with crime rate, however to a lesser extent, with PTA score, total jobs per area, and average cars per household showing Pearson's correlation coefficient values of 0.52, 0.50, and -0.38 respectively (Figure 2). Although not directly applicable to the objective of this study, with the exception of total job numbers in relation to mean car number, all other variables show strong positive or negative intercorrelation.

A limitation of the Pearson correlation test is that it assumes a linear relationship between the variables, and so does not account for non-linearity in the trends. On a borough scale, each independent variable showed either strong positive or negative correlation between itself and the crime rate yet, on a ward scale, the correlation between these factors was found to be reduced. A Spearman's rank test was implemented to measure the strength of the relationships, without the assumption that the relationships are linear. The Spearman's rank test is also less susceptible to extreme outliers and can reveal trends masked by their presence. The resulting coefficients saw a drastic increase in the correlation between crime rate, PTA score, and the average number of cars on both borough and ward scales compared to the Pearson coefficient. A reduction in the correlation coefficient was seen between crime rate and total number of jobs per area on a borough scale, indicating that the previously assumed relationship between the variables was likely overestimated.

Regression Analysis

Linear regression was performed to model the relationship between crime rate and the other predictor variables, and determine the equations that best represents their relationships via minimisation of the residuals. Linear regression assumes that the relationship between the variables is both linear, and normally distributed. The resulting R^2 value obtained through linear regression is an indicator of how well the model fits the data; a value of around ± 1.0 indicated excellent fit, whilst a value of around 0 indicates very poor fit. The following described linear models are shown in Figure 3. The three prediction variables are statistically significant for each model, with all calculated coefficients falling within 99.9% confidence levels.

On a borough scale, the PTA score explains around 75% of the variation in crime rate, as represented by the R^2 value of 0.75 for the model. On a ward scale, however, PTA score only accounts for 27% of the difference. This is likely influenced by the presence of three significant outliers showing very high crime rate compared to the resident population.

The number of jobs also appears to provide a good explanation of the variance in crime rate across boroughs with an R^2 value of 0.73. This, however, appears to be based purely on one borough with a recorded crime rate of over 200, and a high number of available jobs. This figure is also nearly double that of the next highest crime rate, and so is potentially anomalous. On a ward scale, the number of jobs is a relatively poor explanation for crime rate variance, with an R^2 value of just 0.25.

Of the three independent variables analysed, the average number of cars per household gives the least best estimation of crime rate for both boroughs and wards, with R^2 values of 0.53 and 0.15 respectively.

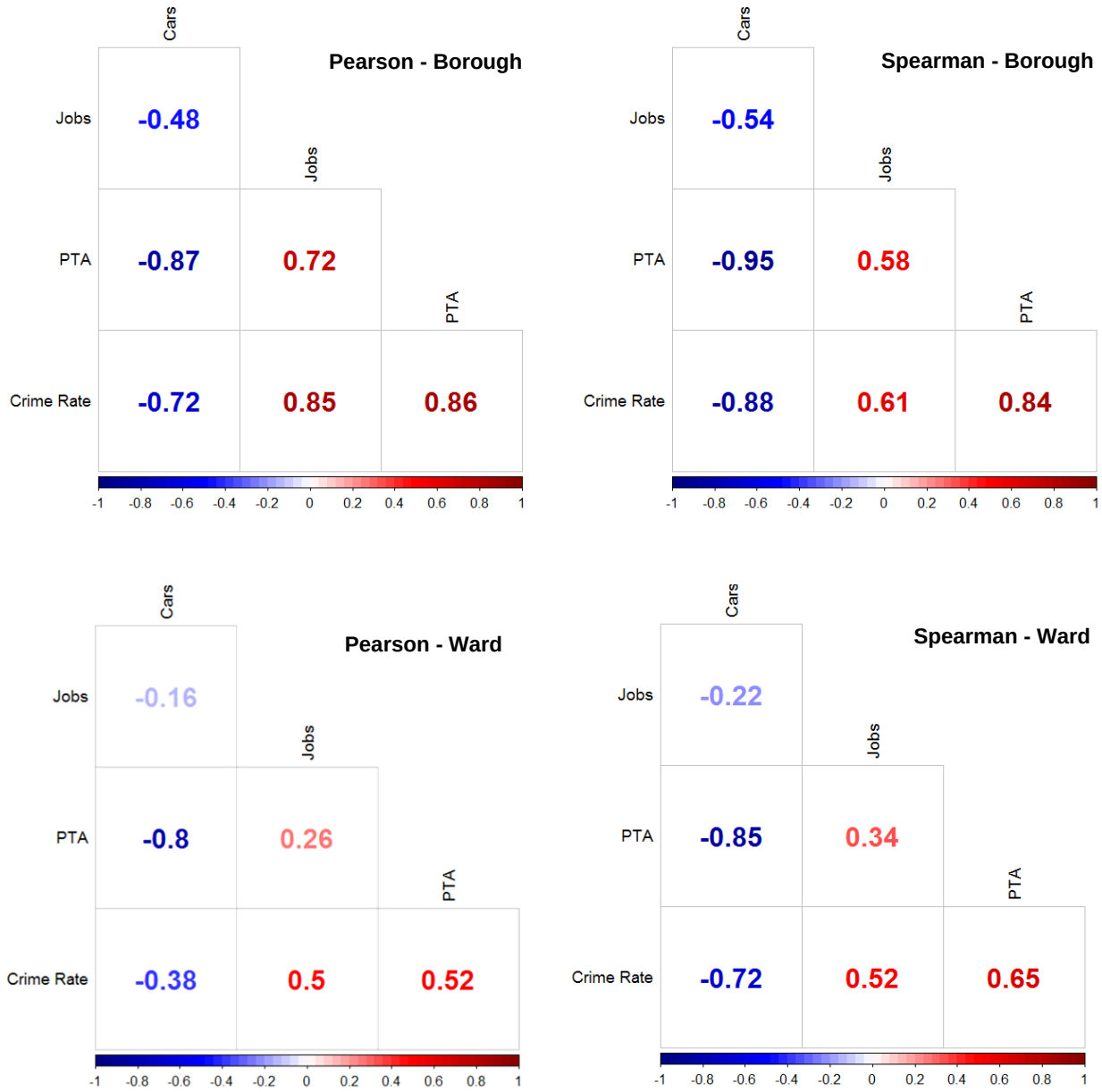

Figure 2: Correlation matrices showing both Pearson (left) and Spearman's Rank (right) correlation coefficients for the relationships between crime rate, average PTA score, total number of jobs per area, and average number of cars per household, on both borough (top) and ward (bottom) scales across London. The highest correlation is found between crime rate and PTA score, with the least significant relationship seen between the number of jobs per area and the average number of cars per household. Spearman's Rank coefficients that are greater than the Pearson values indicate a potentially non-linear relationship between the variables, or that potential outliers exist in the dataset, masking the true relationship.

From the scatterplots it is clear that both PTA score and the number of available jobs are good predictors of crime rate on a borough scale. Multiple regression was carried out to determine if combining the variables into one model can improve the least squares estimate. The resulting models account for 89% of the variance in crime rate across boroughs, but only 41% of the variance across wards. Both models are a significant improvement on the previous models and the resulting coefficients fall within the 99.9% confidence limits, therefore indicating that both variables remain statistically significant to the prediction of crime rate.

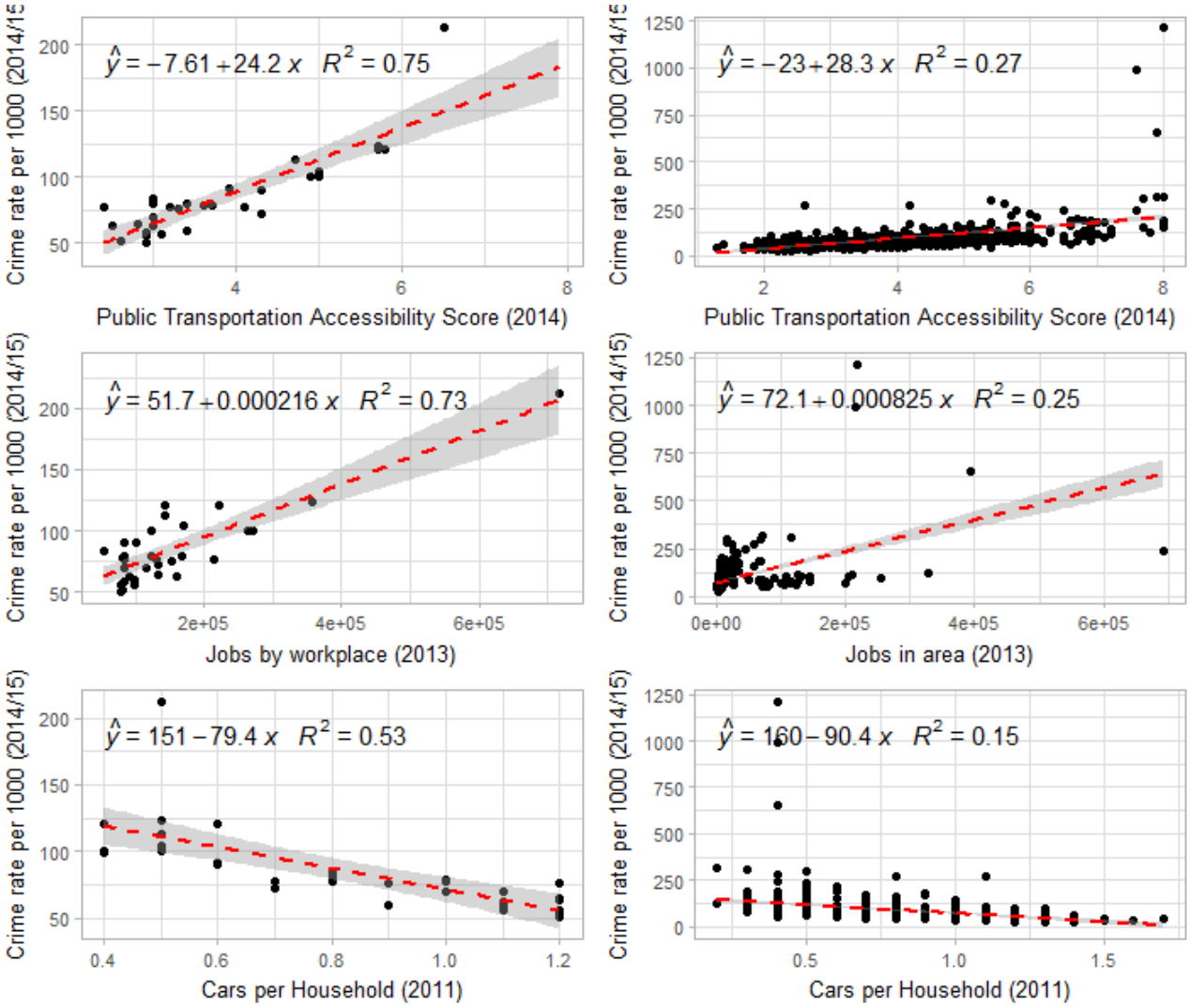

Figure 3: Scatterplots showing the relationship between crime rate and the listed independent variables on both a borough (left) and ward (right) scale. The least square estimate model and R^2 values are also shown, along with the 95% confidence intervals for each model. PTA score and job by workplace explain the greatest variance in crime rate, as shown by their large R^2 values.

Discussion and Conclusion

The aim of this study was to explore the relationship between crime rate and other demographic factors across London on both a borough and ward scale. Analysis of the dataset has concluded that the most significant estimators of crime rate are PTA score, and the number of jobs per area, with the linear regression model combining these variables accounting for 89% of the variance in crime rate across boroughs. These variables are plausible influencing factors of crime rate since a greater number of working individuals in a region, coupled with easy access via public transport, could potentially be seen as an attractive target by criminals, particularly for petty theft.

A limitation in the modelling of crime rate are the three wards for which anomalously high crime rates have been recorded. A potential explanation for the existence of these anomalously high crime rates is that although crime rate is measured per thousand of the population, larger crime rates are not necessarily indicative of the criminal population. These figures can therefore become significantly inflated for areas where there are frequently disproportionately large numbers of visitors compared to residents. Given that these three wards have high PTA scores, and are also known tourist hot-

spots, this high crime rate is likely a result of there being more people present in these wards amongst which crime can occur, as opposed to the criminal population of these wards being significantly larger than average. As Spearman's rank analysis is less sensitive to anomalous data points than Pearson correlation, and since the Spearman's rank values for these relationships are significantly greater than the corresponding Pearson correlation values, it is likely that the fit of the model is affected by these anomalies. A better model could be obtained through exclusion of outliers, or reweighting of these ward records to minimise their influence on the model.

An additional consideration is that, although PTA score and number of jobs per area on a ward scale do not account for the same percentage of variance in crime rate on a borough scale, this may not necessarily indicate that these variables are any less related to crime rate on a ward scale. Instead, it is possible that crime rate on a smaller scale is more greatly influenced by other demographic characteristics or, most likely, the characteristics of neighbouring wards. It is probable, therefore, that crime rate cannot be fully explained without accounting for spatial variations in the dataset.

Bibliography

London Datastore, 2015. *London Datastore.* [Online]
Available at: https://data.london.gov.uk/dataset/
[Accessed 03 01 2017].