Statistical Analysis of GCSE results in London, 2009 - 2013

Introduction

Many factors influence the students' GCSE performance. For example, household income, quality of infrastructure, even geographic location may play a significant role in predicting how well a student will pass their exam. However, because it is impossible to account for all factors, we shall focus on two covariates - the percentage of unauthorised absence in all schools and reported public transport accessibility measure on a scale of 1 to 10.

The purpose of this paper is to analyse the GCSE results in 6 London boroughs – Camden, Hackney, Haringey, Islington, Tower Hamlets and Westminster – and discover which factors have the most significant influence on the students' exam results. Additionally, we shall build a model that would predict an average GCSE based on the two independent variables mentioned earlier, as well as two additional non-numeric independent variables, which are borough to which award belongs, and the year (we treat years as a categorical variable rather than a numeric one because the data is only available for five years).

Data description and transformation

We obtained the data from the official website of the Government of London. The data set is in XLS format, and it contains 664 observations, each corresponding to a specific London ward, and 64 variables.

To proceed with statistical analysis, we first need to turn the Excel file into a "tidy" data frame. We have done this using R packages "readxl", "dplyr" and "reshape2". Here is an excerpt from the prepared data frame.

```
##          Ward     Borough Year     GCSE   Absence  Abscence Transport
## 1  Abbey Road Westminster 2009 307.4700 0.9300000 0.9300000  5.302046
## 2  Abbey Road Westminster 2010 305.9800 1.0800000 1.0800000  5.302046
## 3  Abbey Road Westminster 2011 361.2800 1.2700000 1.2700000  5.358394
## 4  Abbey Road Westminster 2012 354.5000 0.8881690 0.8881690  5.414741
## 5  Abbey Road Westminster 2013 337.4031 0.8276471 0.8276471  5.356010
## 6   Alexandra    Haringey 2009 376.3900 1.2300000 1.2300000  2.978702
## 7   Alexandra    Haringey 2010 360.5200 0.8800000 0.8800000  2.978702
## 8   Alexandra    Haringey 2011 375.6800 1.0700000 1.0700000  3.006827
## 9   Alexandra    Haringey 2012 377.7028 0.5745884 0.5745884  3.034952
## 10  Alexandra    Haringey 2013 385.2711 0.5610289 0.5610289  3.260597
```

We shall summarise the data from all numeric variables in our dataset - GCSE, Absence and Transport. Since sample means and sample median are close to each other in all three variables, one may assume that they are typically distributed. However, the Shapiro-Wilk test of normality (see Shapiro and Wilk (1965)) shows that such an assumption is wrong, as the null hypothesis is rejected in all three cases ($p = 0.02$ for GCSE and Absence, $p < 0.01$ for Transport).

```
##                 Min.    1st Qu.    Median       Mean    3rd Qu.       Max.
## GCSE      270.760000 321.300000 334.51403 333.489069 344.274934 401.033333
## Absence     0.260000   1.009108   1.18000   1.204669   1.392473   2.115000
## Transport   2.853333   4.232832   5.21875   5.205537   6.015322   7.996289
```

Analysis

As it was said in the introduction, we treat GCSE results as a dependent variable. Below are three boxplots demonstrating the distribution of GCSE by borough, by year, and by the combination of borough and year. Note that the last graph is done in black and white because, with many boxes, colours will be uninterpretable.

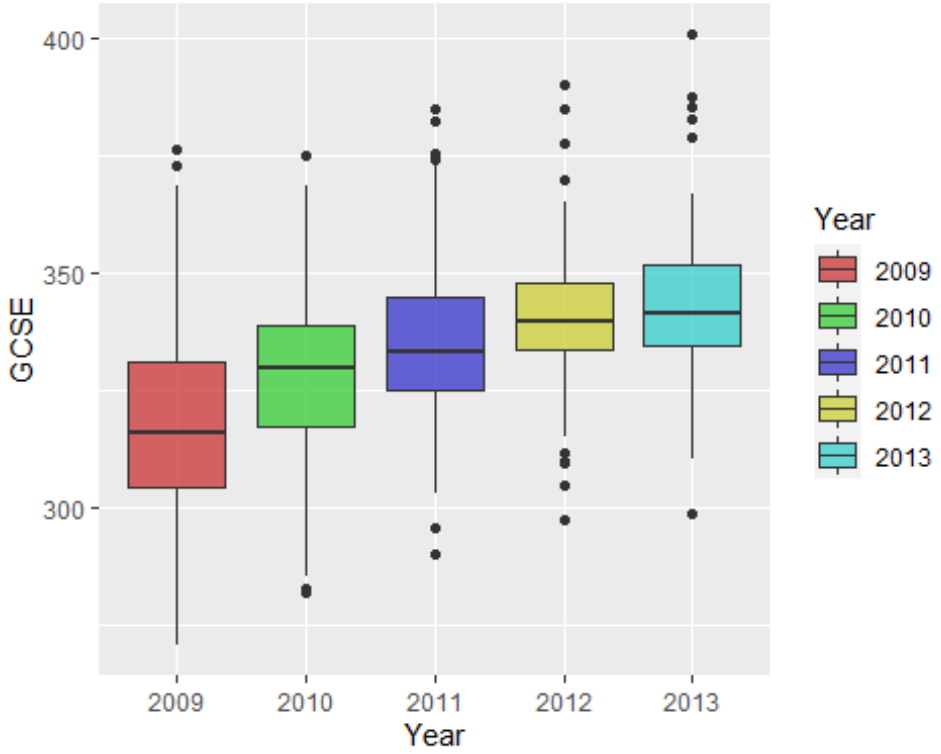

Figure 1. Box plot with average GCSE result by year

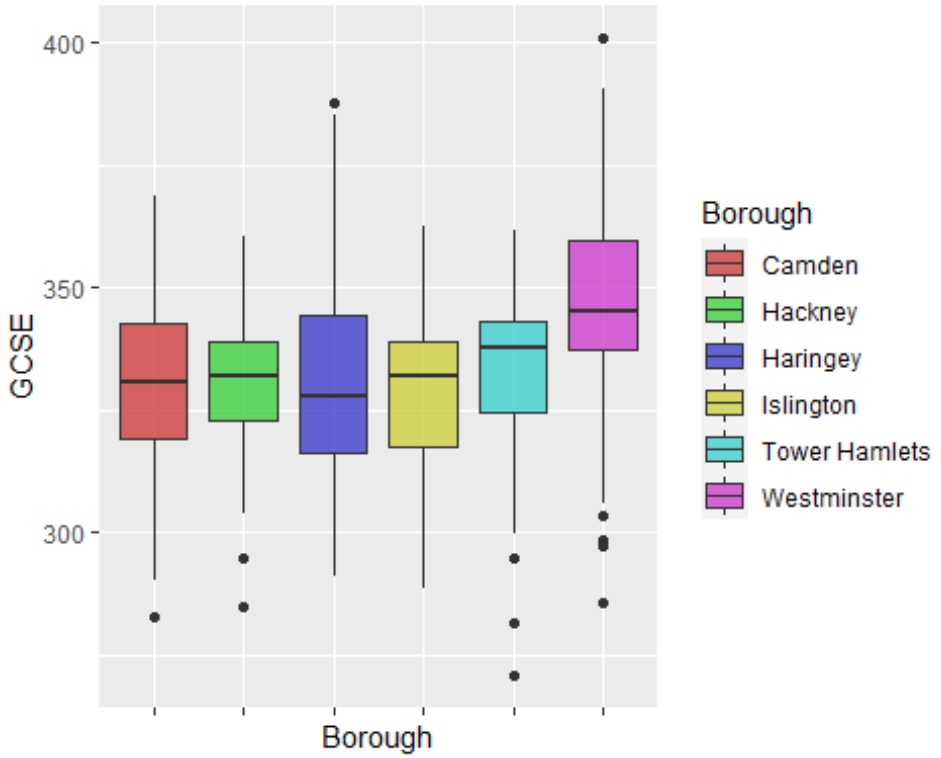

Figure 2. Box plot with average GCSE result by borough, 2009 - 2013

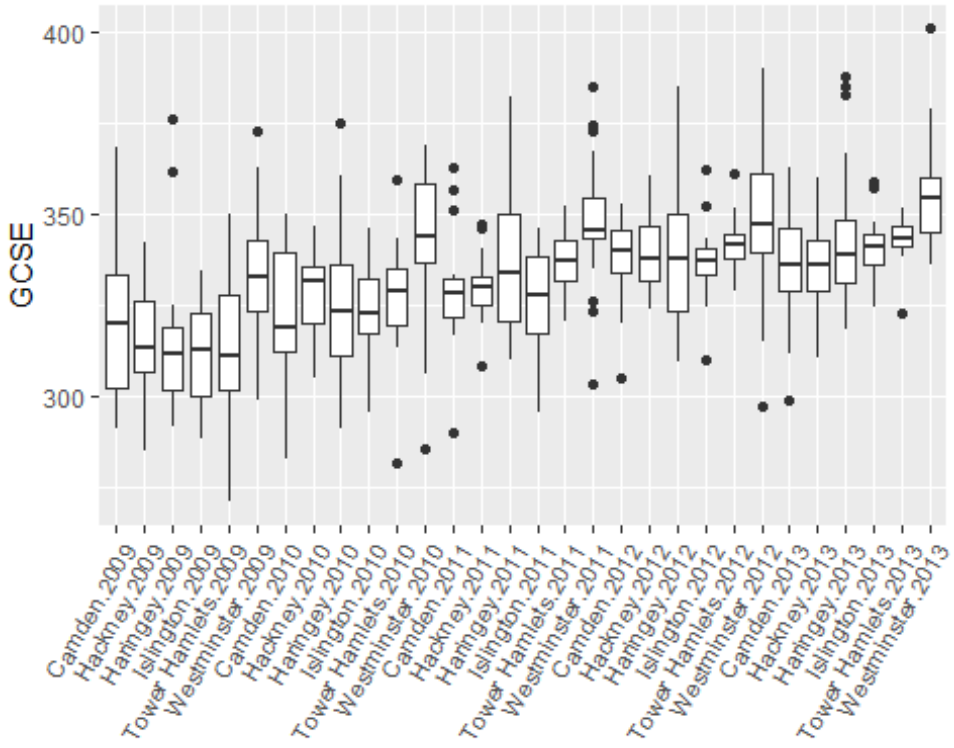

Figure 3. Box plot with borough-year interaction

Notice that, in the first graph, the line inside the box, which represents the median value, increases every year. As a result, the averages grew, although the growth slowed down in 2012-2013.

```
##   Year Mean GCSE
## 1 2009  319.1650
## 2 2010  328.6812
## 3 2011  335.3045
## 4 2012  341.1190
## 5 2013  343.1757
```

Graphical comparison of GCSE results between boroughs is rather complicated. For example, although the average GCSE score in Westminster is higher than in all other boroughs of London, there is no way to conclude whether this difference is statistically significant just by looking at the graph. So instead, we shall use the Kruskal-Wallis test to check if the means are significantly different across groups. Kruskal-Wallis test is essentially a nonparametric version of ANOVA, and we use it because ANOVA requires that the data follow a normal distribution, an assumption our data violates.

The summary of this statistical procedure is printed out below.

```
##
##  Kruskal-Wallis rank-sum test
##
## data:  GCSE by Borough
## Kruskal-Wallis chi-squared = 62.694, df = 5, p-value =
## 0.000000000003368
```

Since the p-value is very close to zero, we know that we can reject the null hypothesis of the equality of averages. However, the Kruskal-Wallis test does not indicate which sample pairs differ from each other, and to tackle this problem, we shall run *post-hoc Mann-Whitney tests* on all possible pairs. Furthermore, to adjust for multiple comparisons, Holm-Bonferroni correction is implemented (see Holm (1979) for more details). Below are the results of the post-hoc tests.

```
##                       Pair p-value
## 1   Westminster - Haringey  0.0000
## 2  Westminster - Islington  0.0000
## 3     Westminster - Camden  0.0000
## 4    Westminster - Hackney  0.0000
## 5  Westminster - T.Hamlets  0.0000
## 6     Haringey - Islington  1.0000
## 7        Haringey - Camden  1.0000
## 8       Haringey - Hackney  1.0000
## 9     Haringey - T.Hamlets  1.0000
## 10      Islington - Camden  1.0000
## 11     Islington - Hackney  1.0000
## 12   Islington - T.Hamlets  0.1305
## 13        Camden - Hackney  1.0000
## 14      Camden - T.Hamlets  1.0000
## 15     Hackney - T.Hamlets  0.2816
```

Note that, if p = 0, the adjusted p-value is less than 10^{-4}*.*

As we can see from the table, Westminster contributes to the statistical significance, as only the pairs with Westminster have p-values less than 0.05.

Now we shall proceed to an essential part of the research - building a predictive model. We shall build three models; the first will include the Absence variable, the second will consist of the Transport variable, and the final model will contain both independent variables. Thus, models 1 and 2 are simple linear regressions, whereas model 3 is multiple linear regression.

```
## lm(formula = GCSE ~ Absence, data = final)
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  368.387      3.002  122.70   <2e-16 ***
## Absence      -28.969      2.419  -11.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.29 on 573 degrees of freedom
## Multiple R-squared:  0.2002, Adjusted R-squared:  0.1988
## F-statistic: 143.4 on 1 and 573 DF,  p-value: < 2.2e-16

##
## lm(formula = GCSE ~ Transport, data = final)
##


## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 329.0821     3.4517  95.339   <2e-16 ***
## Transport     0.8466     0.6448   1.313     0.19
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.31 on 573 degrees of freedom
## Multiple R-squared:  0.003,  Adjusted R-squared:  0.00126
## F-statistic: 1.724 on 1 and 573 DF,  p-value: 0.1897


## lm(formula = GCSE ~ Absence + Transport, data = final)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 365.5656     4.3515  84.010   <2e-16 ***
## Absence     -28.8652     2.4224 -11.916   <2e-16 ***
## Transport     0.5180     0.5783   0.896    0.371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.3 on 572 degrees of freedom
## Multiple R-squared:  0.2013, Adjusted R-squared:  0.1985
## F-statistic: 72.07 on 2 and 572 DF,  p-value: < 2.2e-16
```

In all the models, the Absence coefficient is insignificant, but the Transport coefficient is not. GCSE ~ Absence should be considered the best of the three models we constructed because of the highest adjusted R-squared. An increase of absence from a school by 1% leads to a decrease of an average GCSE score by almost 29 points. In an ideal situation, when there is no unauthorised absence whatsoever, the expected GCSE score is 368.4.

However, it is possible to go even further and include boroughs as factors in our linear regression model. In such a case, these new coefficients are interpreted as the conditional difference between Camden - the base factor, and one of the other five boroughs. Since the adjusted R-square has increased to approximately 0.295, this model is even more accurate.

```
## lm(formula = GCSE ~ Absence + factor(Borough), data = final)

## Coefficients:
##                              Estimate Std. Error t value          Pr(>|t|)
## (Intercept)                   364.418      3.135 116.252           < 2e-16 ***
## Absence                       -32.864      2.531 -12.983           < 2e-16 ***
## factor(Borough)Hackney          2.042      2.329   0.877          0.380976
## factor(Borough)Haringey        14.423      2.481   5.814 0.000000010195472 ***
## factor(Borough)Islington        8.436      2.546   3.313          0.000981 ***
## factor(Borough)Tower Hamlets    8.832      2.432   3.631          0.000308 ***
## factor(Borough)Westminster     17.350      2.300   7.545 0.000000000000181 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.22 on 568 degrees of freedom
## Multiple R-squared:  0.3028, Adjusted R-squared:  0.2954
## F-statistic: 41.11 on 6 and 568 DF,  p-value: < 2.2e-16
```

Finally, we present a scatter plot that demonstrates clearly a negative dependence between GSCE results and absence from school.

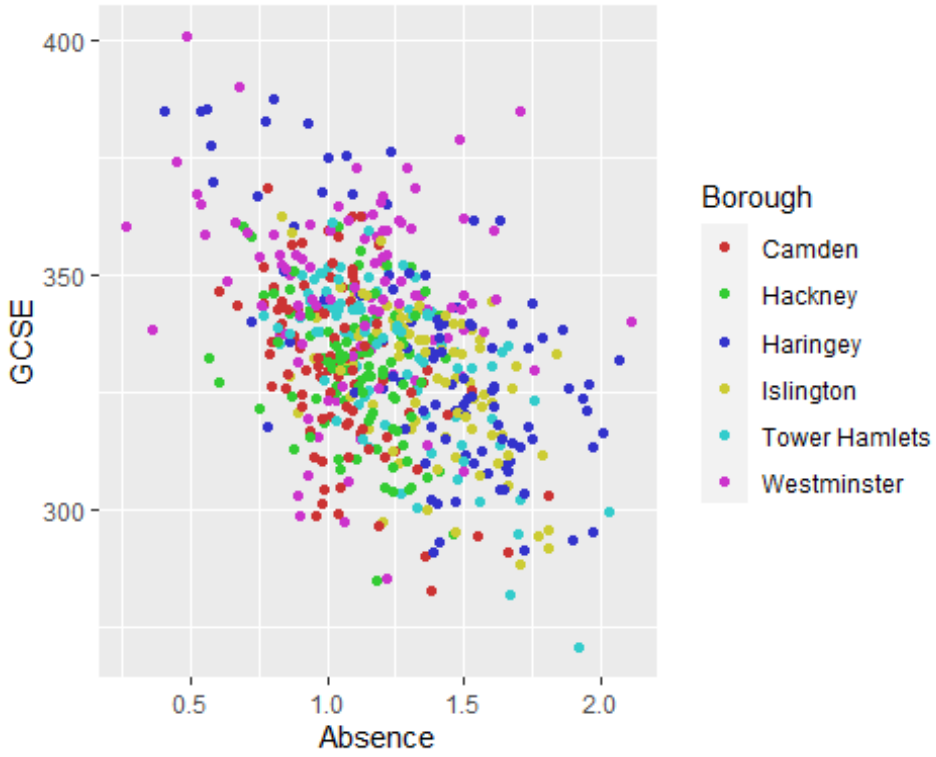

Figure 4. Absence – GCSE scatter plot

Conclusion

We have used Camden, Hackney, Haringey, Islington, Tower Hamlets and Westminster as a case study to demonstrate that GCSE results differ across various London boroughs, at least those we have investigated. The results in Westminster are the highest, and the mean GCSE result in Westminster is significantly different from that in all other boroughs. On the other hand, additional pairs of boroughs did not differ considerably in terms of average GCSE results.

Unauthorised absence in schools turned out to be an essential factor that negatively correlates with average GCSE scores.

Finally, we have built a predictive model that predicts GCSE scores based on the absence from schools and adjusts the prediction by borough.