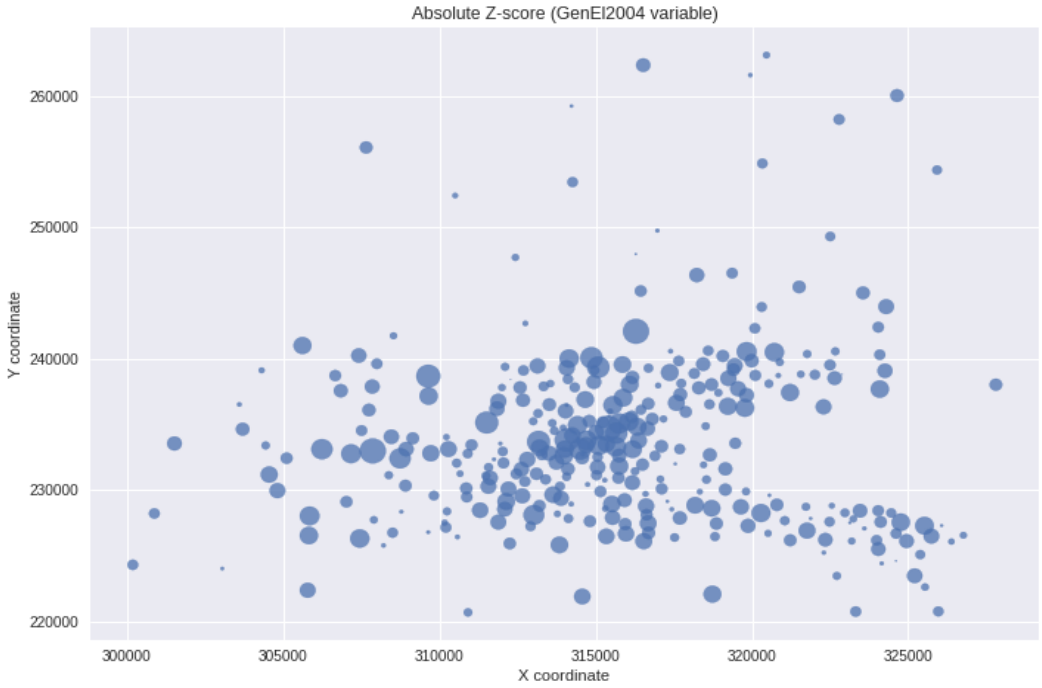
Dublin Voting Data

Data Set and Workflow

The dublin voting dataset contains the general election voter turnout rate in 322 division for the year 2002. Specific characteristic such as centroid of the division, percentage of people living in different address since 1 year ago, unemployment, housing and age demographsics were thought of having potential influence on the voting turnout rate. The task is to predict which of these variables actually influence the voter turnout. As with any dataset, initial screeing / visualization was performed to diagnose potential problems, outliers and issues. Then linear models with individual variable and multi variate model was fit. Later, geographically weighted regression was used to examine spatial variation. Cursory PCA was performed to examine collinearity.

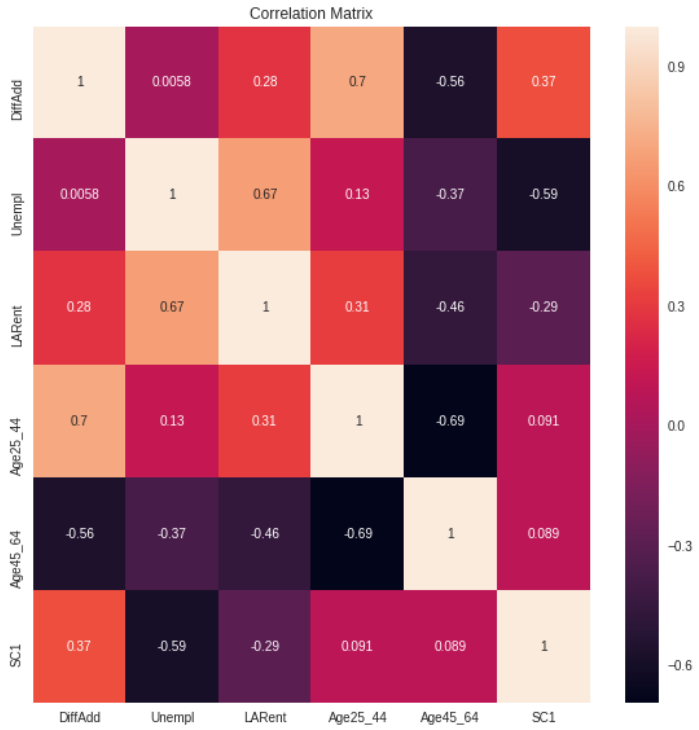
Visualizing Location

Initial data screening of individual variables showed no problematic values. The bubble chart below shows the absolute value of the Z-score (of GenEl variable), we can see the higher turnout rate is clustered mainly at the heart of Dublin City, indicating there is a gradual spatial gradient. But at the same time there are these divisions scattered all over the place with very small/large Z-score.



Visualing Correlation

On checking for collinearity, we found strong correlation between Unemployment rate and percentage of people renting from local authorities. Moreover, people between age 25 and 44 are highliy mobile due to work / family, which is also visible here. Strangely there seems to be additional negative correlation between age group 25-44 and 45-64 which suggest multicollineariy issues.



The condition number of all the above conflicting variable is 239 which is greater than 30 confirming our suspicion of strong multicollinearity. But the unemployment and social class 1 are not so strongly related (thus their collinralty should be ignored).

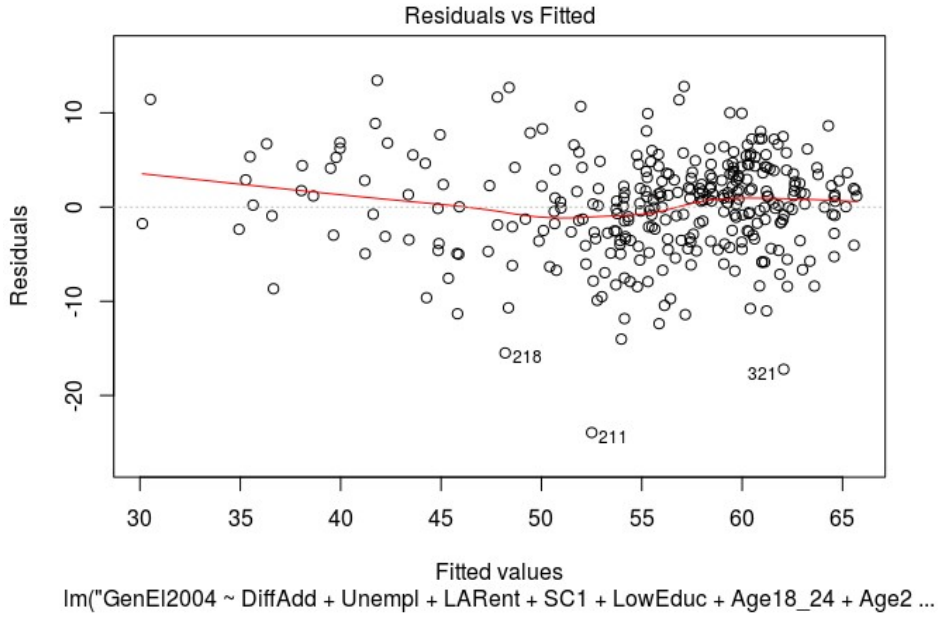
Variables	Condition Number
Age 25 -44, Age 45-64, DiffAdd	114.91 > 30
LARent, Unempl	25.83
Unempl, SC1	2.67
All above variables togather	239 > 30

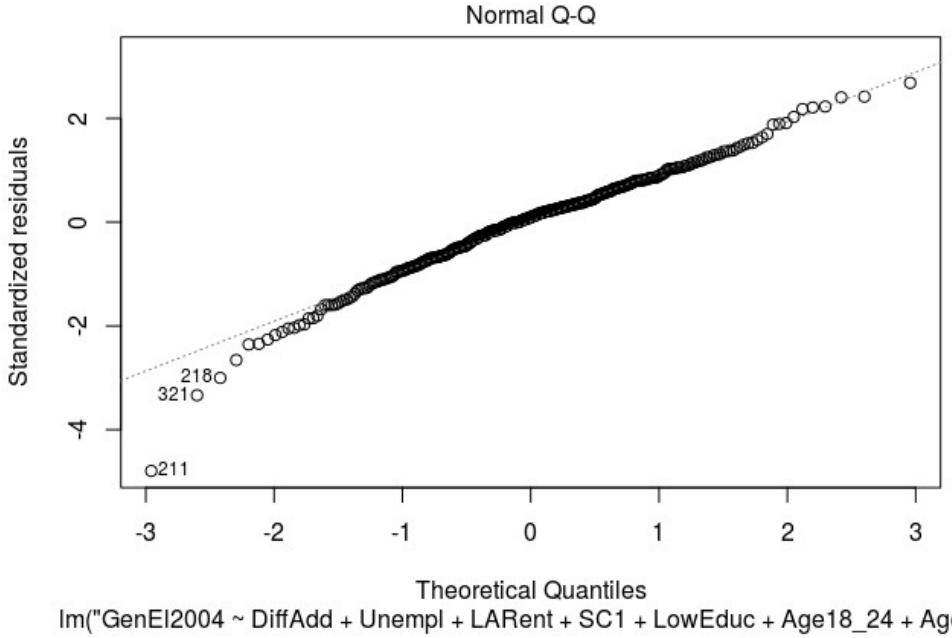
Linear Regression Models

```
# model with 1 variable
m1 = lm('GenEl2004 ~ DiffAdd', data=Dub.voter)
m2 = lm('GenEl2004 ~ Unempl', data=Dub.voter) # lowest AIC of 2110
m3 = lm('GenEl2004 ~ LARent', data=Dub.voter) # AIC of 2112
m4 = lm('GenEl2004 ~ SC1', data=Dub.voter)
m5 = lm('GenEl2004 ~ LowEduc', data=Dub.voter)
m6 = lm('GenEl2004 ~ Age18_24', data=Dub.voter)
m7 = lm('GenEl2004 ~ Age25_44', data=Dub.voter)
m8 = lm('GenEl2004 ~ Age45_64', data=Dub.voter)

# model with 2 variables
m10 = lm('GenEl2004 ~ Unempl + LARent', data=Dub.voter) # lowest AIC of 2052

# model with all variables

dubvoter_ols = lm('GenEl2004 ~ DiffAdd + Unempl + LARent + SC1 + LowEduc + Age18_24 +
Age25_44 + Age45_64', data=Dub.voter) # AIC of 2000, Adjusted R^2 of 0.62
```

From these models, we found that Unempl and LARent had lowest AIC. So the next step was to combine the Unempl and LARent togather and we got a lower AIC score. But the model with all variables is will lowest AIC score of 2000, and it is a better model than model with just 2 variabes, which is confirmed by ANOVA.

The model with all variables indicated that Unemployment, LARent, Age18_24, Age25_44 were significant variables, but ANOVA says we need the complex model with all variales. Follwing diagnostic plots indicate a few outliers in residual v/s fit plot and almost correct QQ plot. These outliers and slight curvature might be due to the spatial aspect of the data, which is examined in the next section.

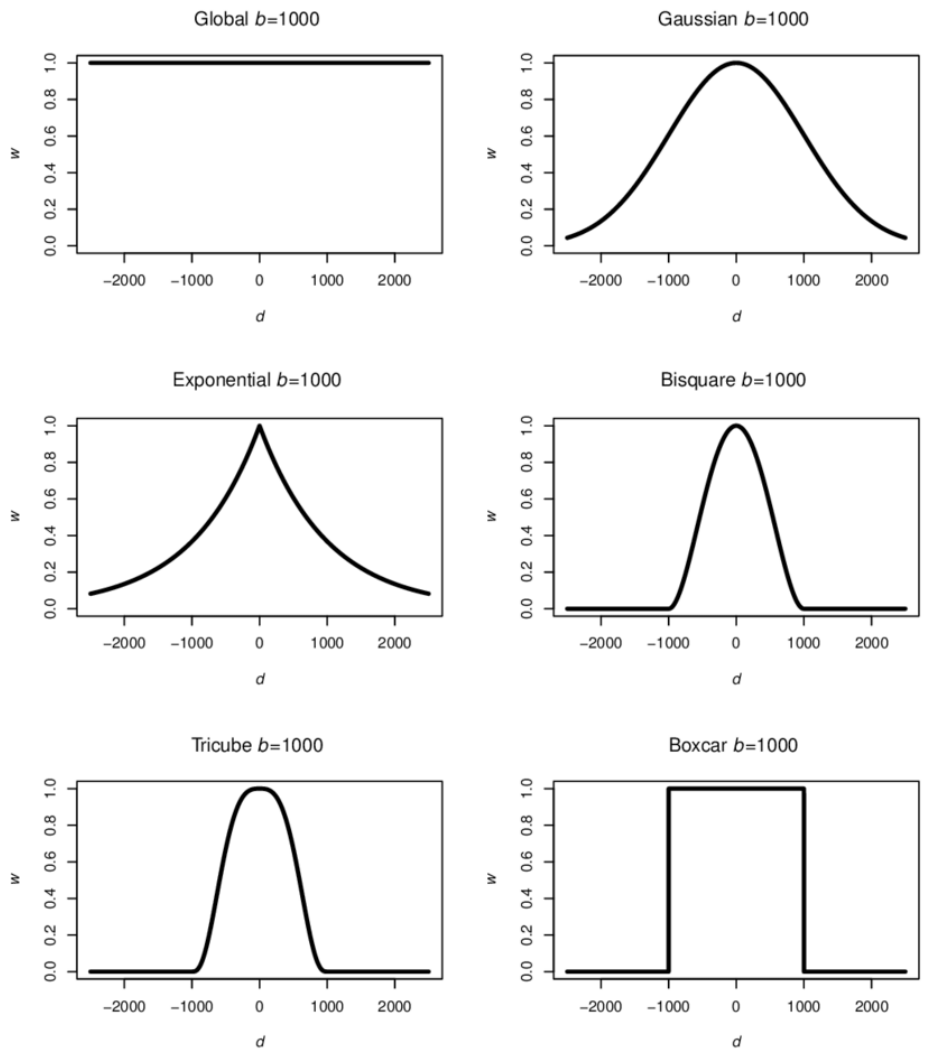



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm("GenEl2004 ~ DiffAdd + Unempl + LARent + SC1 + LowEduc + Age18_24 + Age2 ...
218
321
211
-3
-2
-1
0
1
2
3
-4
-2
0
2

Geographically weighted Regression

In this section we will explore the effect of different kernels when using geographically weighted regression. Due to irregualr distribution of data point, adapative kernel is enables so we can account for same number of observation in each location. Thus in the centre of the data cluster, where the points are located vary close to one another, the kernel will shrink and the opposite occours when points are vary far from one another. This method incorportes the spatial influence of locations into our model. The euclidean distance is a metric to find distance between points.

Kernel	AICc	Number of Nearest Neghbours
Bisquare	1921	109
Gaussian	1938	25
Tricube	1920	109
BoxCar	1920	56 *
Exponential	1943	24



As we can see boxcar and tricube give the same AICc but very different number of neighbours. Since our dataset is small, 109 seems like a vary large number of observations. Let's see the effect of regression using these 2 kernels.

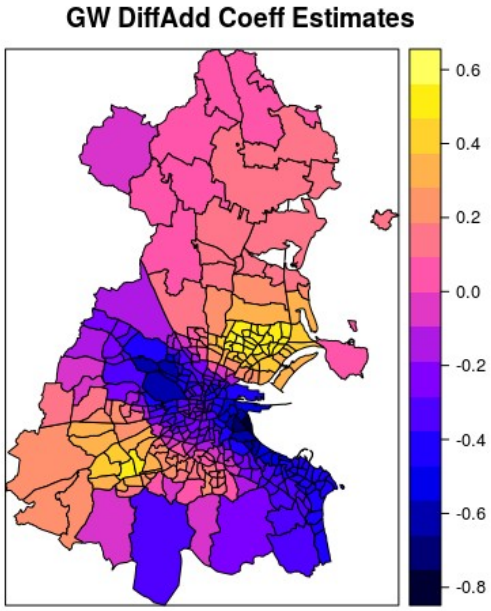
Kernal	Effective number of parameters	Adjusted $R^{\wedge 2}$	AICc	AIC
Tricube	73	0.75	1920	1831
BoxCar	50	0.75	1921	1848

As we can see there is a difference of greater than 2 between the AIC of these models. So we pick the model with lowest AIC which is Tricube and make further comments.

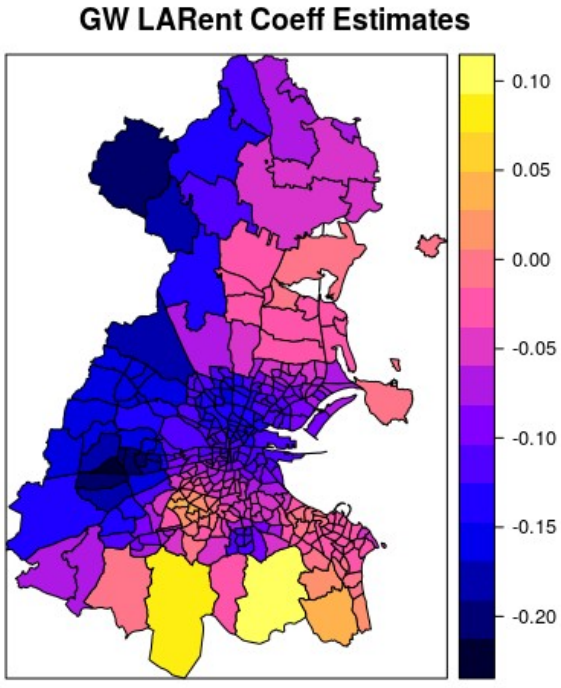
```
***********************************************************************
*          Results of Geographically Weighted Regression              *
***********************************************************************

*********************Model calibration information*********************
Kernel function: tricube
Adaptive bandwidth: 109 (number of nearest neighbours)
Regression points: the same locations as observations are used.
Distance metric: Euclidean distance metric is used.

****************Summary of GWR coefficient estimates:******************
                  Min.     1st Qu.      Median     3rd Qu.      Max.
Intercept  53.7808262 72.8157781 82.1737362 95.8088772 117.3724
DiffAdd    -0.7562393 -0.3582695 -0.1697161  0.1711809   0.5633
Unempl     -2.3781468 -1.1714262 -0.7709499 -0.4743775  -0.0875
LARent     -0.2132590 -0.1214840 -0.0814984 -0.0371046   0.0934
SC1        -0.1693264  0.0262118  0.3107902  0.4342629   0.9572
LowEduc    -7.9146467 -0.6785253  0.5897242  1.9167590   3.5055
Age18_24   -0.4089751 -0.2628863 -0.1449846 -0.0017141   0.3830
Age25_44   -1.1042407 -0.7184230 -0.4658252 -0.3070263   0.2493
Age45_64   -0.9517535 -0.4085001 -0.1036690  0.0447150   0.5213
************************Diagnostic information*************************
Number of data points: 322
Effective number of parameters (2trace(S) - trace(S'S)): 73.25304
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 248.747
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 1920.715
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 1831.654
Residual sum of squares: 4636.832
R-square value:  0.8095094
Adjusted R-square value:  0.7531857
```

Visualizing the GW Coefficients of Variables

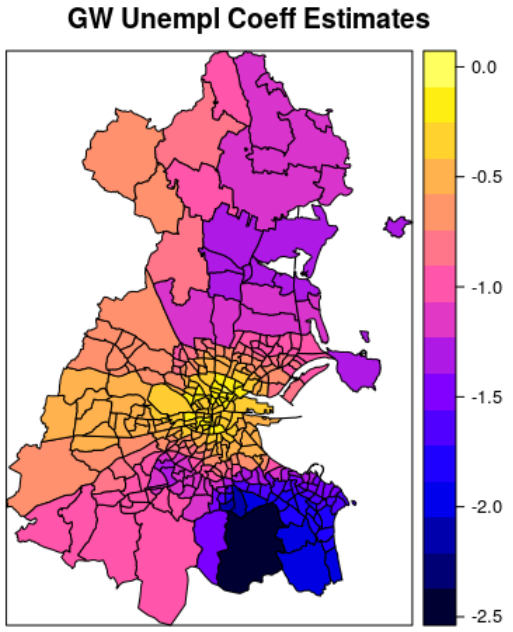
People in the center of Dublin have negative relationship with GenEl, while people who live on the outskirts of dublin have positive relationship with GenEL. Maybe it's because centre of city is expensive and people cannot afford the houses, so they move outside, and this increases mobility in the center of dublin.



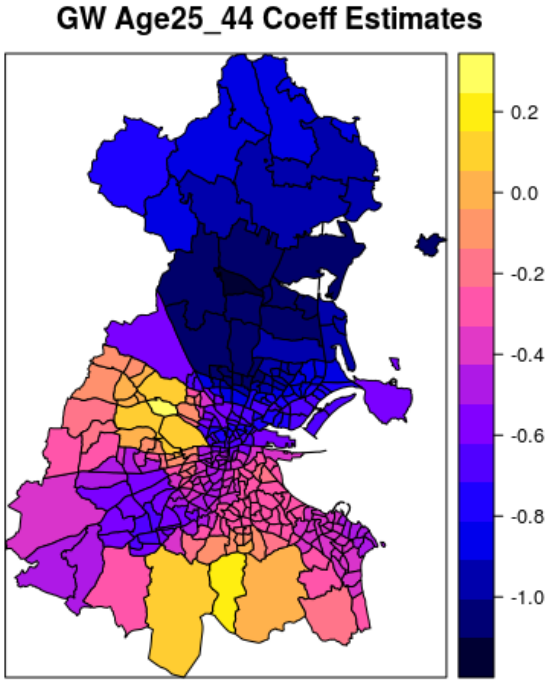
Dublin for most part has negative relationship with renting, so as the renting percentage in each ED increases, there is reduced voter turnout rate except the 2 southern ED, which are outliers.



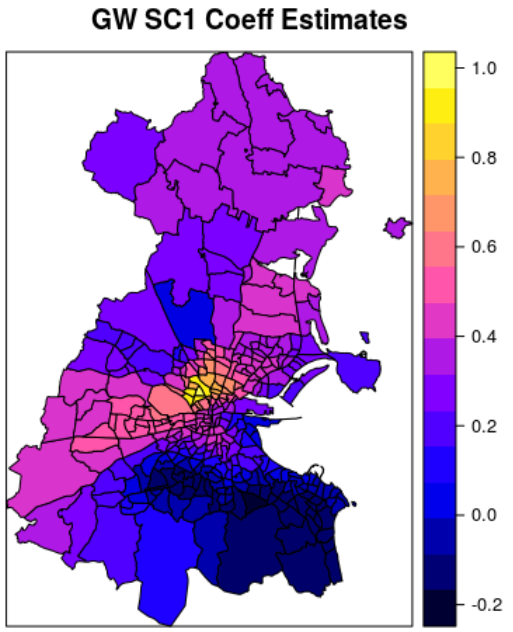
For the most part, Dublin is an affluent place, as unemployment rate are almost near zero. But if the unemplyment increases by 1 unit, then voter turnout rate decreases or stays the same.



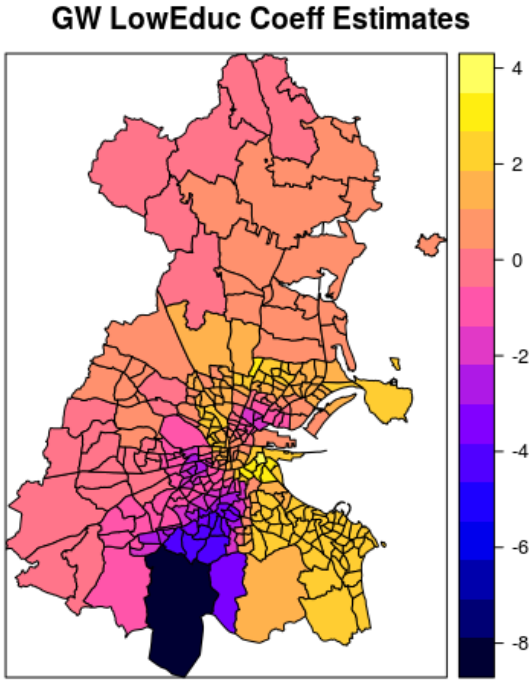
Here there is a stark contrast, between north and south dublin. In the northern part, if the population in Age 25 – 44 has increasing value, then the voter turn out decreases and vice varsa for southern part.



For the most part, Dublin is having negative relationship except very specific places in the heart. But the effect of this variable is quite uniform in each location.



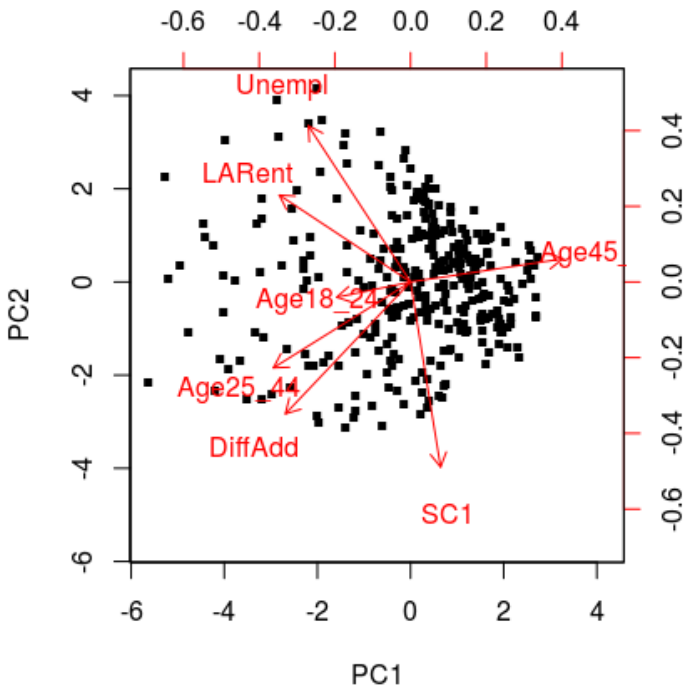
As the percentage of lower education increases, the voter turnout increases, which is quite strange. Although there is an outlier division.



PCA for Collinearity

PCA was performed to examine if we can collapse the data set to smaller dimension. After looking at the correlation heatmap shown in the Visualizing Correlation section, we selected specific 7 variables and performed a non GW-PCA. Following loadings were obtained, their inter[retation is given below. A biplot is provided for additional visual explanation.

```
Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
Age18_24 -0.244         0.917 -0.231                 0.199
Age25_44 -0.453 -0.283 -0.314 -0.294 -0.434 -0.207  0.547
LARent   -0.433  0.287         0.627 -0.116 -0.549 -0.134
Age45_64  0.503         0.220  0.275 -0.761  0.105  0.167
DiffAdd  -0.413 -0.435         0.107 -0.327  0.437 -0.572
Unempl   -0.338  0.516         0.206         0.664  0.360
SC1              -0.611         0.581  0.327  0.104  0.399

                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SS loadings      1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var   0.143  0.143  0.143  0.143  0.143  0.143  0.143
Cumulative Var   0.143  0.286  0.429  0.571  0.714  0.857  1.000
```



The first four principal components account for 90% of the variance.

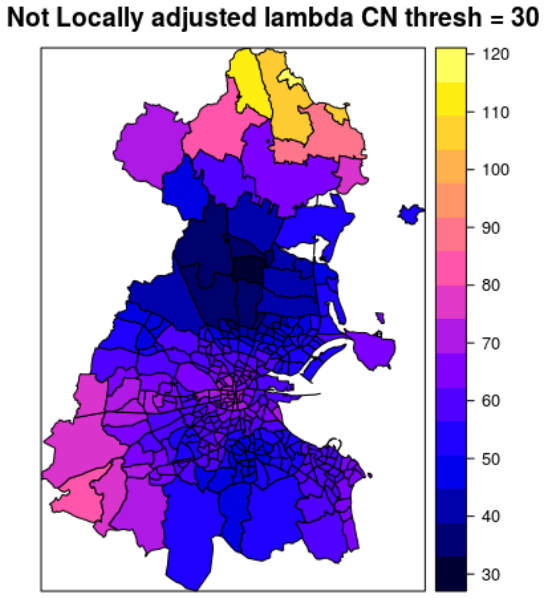
The first princical component have large positve magnitute for Age 45_64 predictir and large -ve magnitude for other preditctors, so it compares Age45_64 with other variables.

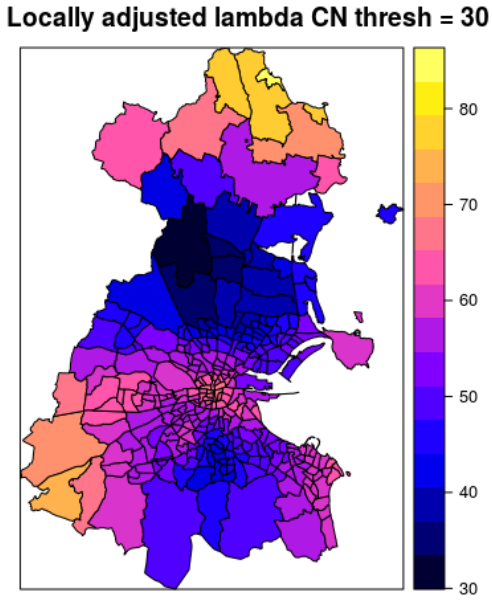
The second principal component has high negative values for Age18_24, SC1, DiffAdd and Age25_44 but +ve values for LARent and Unempl. It informs us about the age and mobility.

GW Ridge Regression for Collinearity

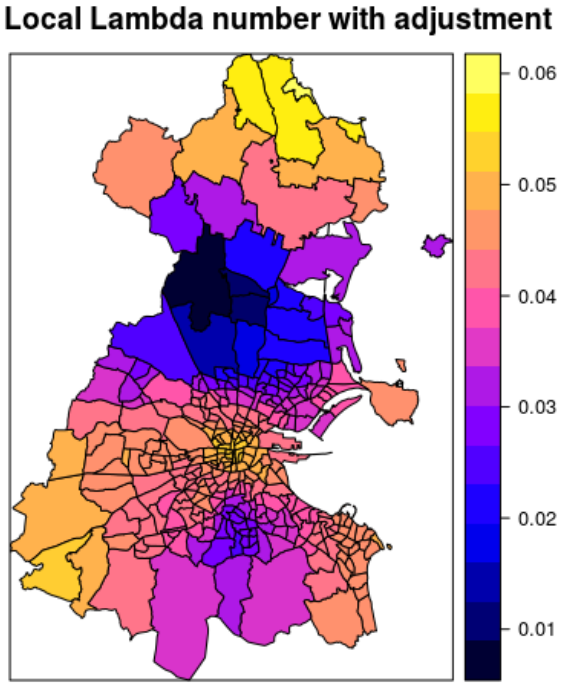
Local multi collinearity is a major problem, I used GW ridge regression to deal with it.

Model	AIC	AICC
Non adjusted Lambda	2002	2044
Adjusted Lambda	2005	2029





The above 2 graphs show the condition number before and after adjusting the lambda parameter for the regression. There is very large range of the CN before adjusting which is reduced to 30 to 80 after adjusting the lambda. Also the local lmabda is shown below for each disivison.



Conclusion

In conclusion, specifiing significant variables depends upon the specific division in consideration. The geographically component makes the interpretation complex and we cannot say which variable is important in general sense as it depends on the division. Although visualizing coefficients from the GW model gives some sense of overall stark contrats between the various divisions. Further work includes applying goegraphically weighted PCA and geographcailly weighed correlation for each division to make a more complex model.