Introduction

Intuitively, tobacco use increases the risk of lung cancer and of all adults in the UK. This project uses the linear model and Poisson regression model to analyse the relationship of the number of lung cancer deaths per year in relation to age and smoking status. Poisson regression model may be the better choice for this project, and if someone who smoke cigarettes and cigars or pipes and if his age is more than 80, he will have the highest probability to get the lung cancer.

Aim

The aim of this project is to investigate whether the age and smoking status influence the risk of lung cancer, and this project also want to construct a predictive model to answer a question: how many people will die because of lung cancer for a group of people when given the age and smoke status for each person.

Data

Four variables including age, smoking status, population and deaths are recorded in the data set with 36 observations. Age is a categorical variable and is placed into groups of 5 year increments from the age 45 to the age 80. Individuals above the age of 80 are placed into a group. Therefore, there are a total 9 age groups. Smoking status is a categorical variable with four levels, "no" is for people who do not smoke, "cigarPipeOnly" is the people who smoke cigar or pipes only, "cigarretteOnly" is the people who smoke cigarettes only and 'cigarrettePlus' is the people who smoke cigarettes and cigars or pipes. Population is a numerical variable which means number of individuals in hundreds of thousands. Deaths is a count data; it indicates the number of lung cancer deaths per year.

Method

Liner regression model

Linear regression is the simplest model which always be considered at first, but the variable death is a count data. Since one of the main assumptions of linear models and analysis of variance is that the residual errors follow a normal distribution, a transformation of the variable death is required. Based on the population and death, the death rate is being calculated as:

$$Deathrate_i = \frac{Death_i}{Population_i}$$

where i means the i^{th} group in the dataset

The death rate means the percentage of deaths in each group, so fit a linear regression model of the form:

$$Deathrate_i = \beta_0 + \beta_1 age_i + \beta_2 smoke_i + \epsilon_i$$

where $\beta_0, \beta_1, \beta_2$ are coefficients for variables, ϵ is the error and i means the i^{th} group in the dataset.

Comparing to all other models, the linear regression model is easy to interpret, and many complex models are expanding from the linear regression model.

Poisson regression model

Alternatively, the Poisson regression model is appropriate when the response is count data, so Deaths are treated as the response and fitted a Poisson regression model of the form:

$$\log(deaths_i) = \log(population_i) + \beta_1 + \sum_{k=2}^{9}(\beta_k Age_{ki}) + \sum_{j=10}^{12}(\beta_j smoking\ status_{ij})$$

Where Age_{ki} including eight age groups from 45 to 80+, and $smoking\ status_{ij}$ including cigarPipeOnly, cigarretteOnly and cigarrettePlus. β_n (n=1 ,2 …12) is the coefficient of the n^{th}covariates.

The log population is offset here, since the size of the different groups are different.

The goodness of the model could be reflected by the Pearson chi-square statistics, which could be calculated as

$$X^2 = \sum {r_i}^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

where o_i is the i^{th} observation value and e_i is the i^{th} fitted value

if the X^2 is smaller than chi-squared distribution with residual degree of freedom, the model fits the data well.

Result

The report began with the bean plot, which can show the actual density of the data. The bean plot of death rate versus smoking status showed that the pattern of four status seemed to be similar, and the mean of each groups were also approximately the same. The group of “no” was unimodal and symmetric for their density curve, and the density of the group “cigarretteOnly” and “cigarrettePlus” displayed slightly skewed but both were also unimodal. The group “cigarPipeOnly” revealed potential multimodal in the density but it is not very clear. The bean plot of death rate versus age showed that the death rate increased with age, and this growing trend accelerated for the elderly group. The group “40-44” and “45-49” displayed symmetric and unimodal distributions, while the group “50-54”, “55-59” “60-64” and “65-69” showed one obvious peak value and another potential peak value, these models might be the unimodal distribution. For the group “70-74”, “75-70” and “80+”, these groups displayed clearly unimodal distributions.

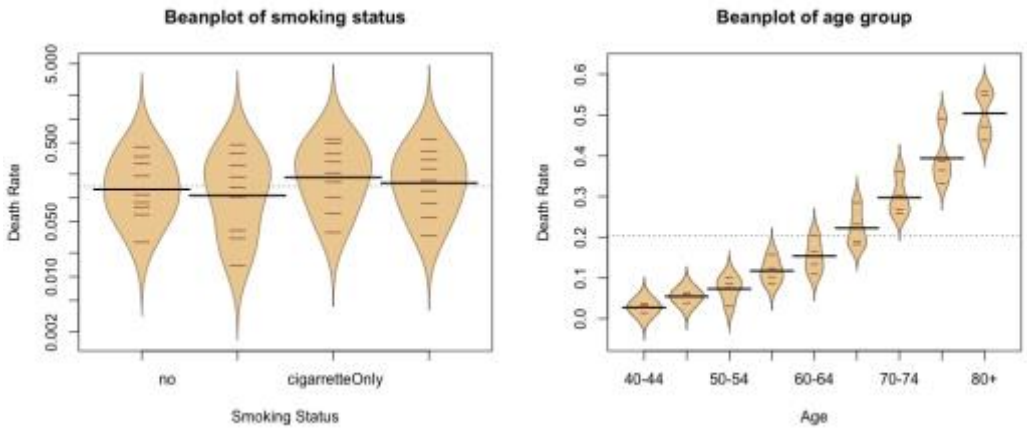

Figure 1: the bean plot the smoking status and age

Then, assessing the linear regression model and Poisson regression model by comparing to the observed value and the fitted value which indicated the goodness of fit for the models. According to the plots of observed values versus the fitted values, the Poisson regression model fitted better than the linear regression model, because most fitted value points and observed value points coincided in the plot of Poisson regression model, while only a few points coincided in the plot of the linear regression model.

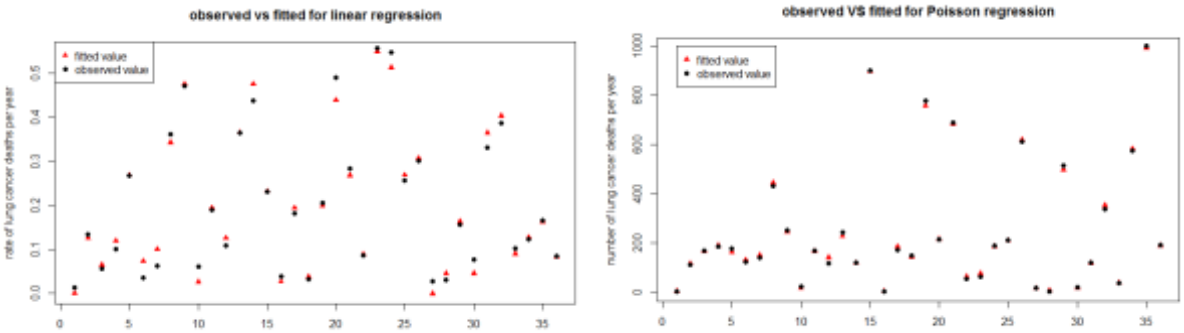

Figure 2: the plots of observed versus fitted for linear regression model and Poisson regression model

One of the basic assumptions for both model is normality which could be accessed by residual plot and Normal quantile-quantile(QQ) plot. According to the figures for the linear regression model, all points are located randomly in the residual plot and in QQ plot all point are located around the straight line, both indicating well fit. As for the figures for the Poisson regression model, almost all points are located around zero in the residual plot, and all points are located around the straight line in the QQ plot, both indicating well fit. Therefore, both models are valid.

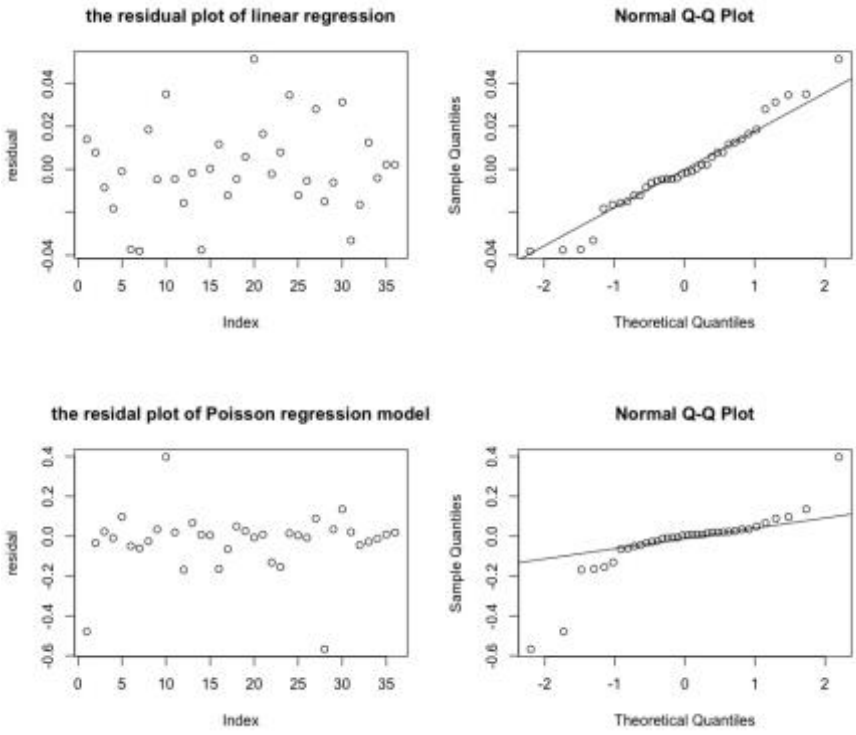

Figure 3: the residual plot (top left) and the QQ plot (top right) for the linear regression model and the residual plot (bottom left) and QQ plot (bottom right) for the Poisson regression model

The most important assumption for the Poisson regression model is the equality of mean and variance. From the plot of the variance versus mean, most points were located around the straight line but there were a small number of points below the line. A slight underdispersion appeared in the Poisson regression model.

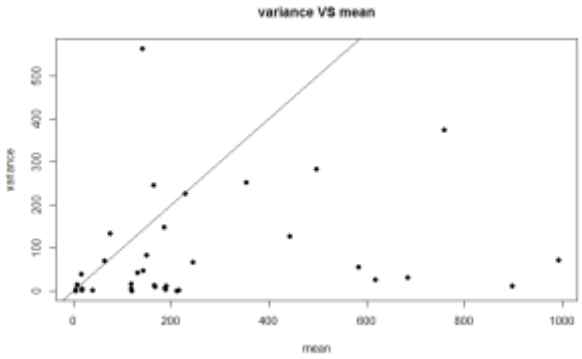

Figure 4: the plot of variance versus mean for the passion regression model

Additional, the Pearson chi-squared statistic was calculated to assess the goodness of the Poisson model. The result was 14.88, since the residual degree of freedom is 23 which meant the $\chi^2(23)$ is 35.17. The Pearson chi-squared statistic was smaller than $\chi^2(23)$, so it indicated a good fit for the Poisson regression model.

From the **Table 1**, the intercept is age group 40-44 and smoking status "no", it was clear that all terms except "cigarPipeOnly" had a very small P-value (i.e. <0.01) which indicated a significant effect. The 95% confidence interval of rate ratio included 1 for the "cigarPipeOnly", so there was no significant difference in number of death between the group of people who do not smoke and the group of people who only smoke cigars or pipes. For the group of people who smoke cigarettes only, the number of death was 1.52 times higher than the group of people who do not smoke, and for the group of people who smoke cigarettes and cigars or pipes, the number of death was 1.24 times higher than the group of people who do not smoke. As for the age groups, the number of death for age group from 45 to 49 was 1.74 times the age group from 40 to44 while the number of death for age group over 80 was 17.23 times than the age group from 40 to 44. The rate ratio for age groups increased gradually with age.

Table 1: the table of coefficients in the Poisson regression model

Term	Estimated value	Standard error	P-value	Rate Ratio	95% CI
Age 45-49	0.554	0.080	<0.01	1.74	(1.49, 2.04)
Age 50-54	0.980	0.077	<0.01	2.67	(2.29, 3.10)
Age 55-59	1.380	0.065	<0.01	3.97	(3.50, 4.51)
Age 60-64	1.654	0.063	<0.01	5.23	(4.63, 5.91)
Age 65-69	1.998	0.063	<0.01	7.38	(6.52, 8.34)
Age 70-74	2.271	0.064	<0.01	9.69	(8.54, 11.00)
Age 75-79	2.559	0.068	<0.01	12.92	(11.31, 14.75)
Age 80+	2.847	0.072	<0.01	17.23	(14.95, 19.86)
cigarPipeOnly	0.048	0.047	0.31	1.05	(0.96, 1.15)
cigarretteOnly	0.417	0.040	<0.01	1.52	(1.40, 1.64)
cigarrettePlus	0.218	0.039	<0.01	1.24	(1.15, 1.34)

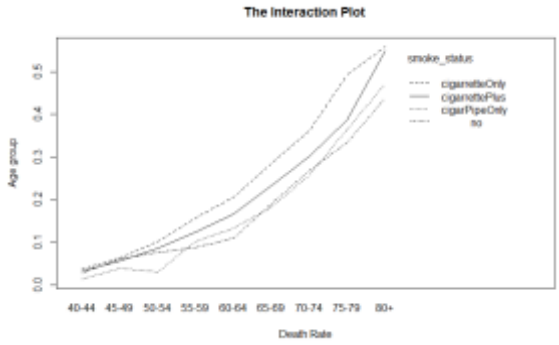

Figure 5: the interaction plot of age group versus death rate

To see whether there was any interaction among different smoking status group, the report used the interaction plot. Based on the **Figure 5**, only two lines (i.e. cigarretteOnly and no) intercepted at a few points and the other two lines were approximately parallel.

Conclusion and Discussion

According to the result, both models fit the data well, but comparing to the linear regression model, the Poisson regression model performs better according to the plots of observed versus fitted value. Besides, the project transforms the response in the linear regression which lead to difficulties of interpreting the model. In conclusion, the Poisson model is selected as the final model.

The result shows that cigarettes are more harmful than the cigars or pipes, and smoking is more harmful to older people. This conclusion suggests that smoking less cigarette, or smoking cigars or popes instead of cigarettes is a better option. It also suggests the elderly population should smoke less, especially for those who are older than 60, as their health suffer five times more risk than people ages from 40 to 45.

Though the Poisson shows great property here, but there is a slightly underdispersion, which means the mean is higher than the variance. The regression parameter estimates are consistent in the presence of underdispersion, but their standard errors will be wrong. One approach is use the generalized Poisson regression model to fit the data, but this may exceed the range of this course and still need more exploration.

Additionally, the model shows that the number of deaths increase approximately exponentially with the increase of age. But considering the natural mortality rate, it also grows exponentially with the increase of age therefore, this result might be biased. Using the direct standardization or indirect standardization might be helpful.

Another variable that could be added in is the total length of years that they have smoked. Because the health effect of smoking has long term consequences and are compounded.

In this dataset, the age group starts from 40, but ignore the age group under 40. When discussing the relationship between the age and the number of lung cancer death per year, those age groups are indispensable. Besides, electronic cigarettes are increasingly popular, it could be a new group for the smoking status.

Introduction

The Chicago Police Department (CPD) opened the data set which includes all the reported incidents of crime except of murdering that occurred in the City of Chicago from 2002 to 2015. The aim of this project is to use the clustering method to divide the communities into appropriate number of clusters. This could be interesting to find the distribution of crime types. The project selected three common crimes including narcotics, arson, theft and battery which occurred in 2011. Comparing the hierarchical clustering and K-means, the K-means is chosen for this project with k equalling to 2.

Data

The resource of the data is from the CPD and all crime record in this report occurred in 2011. The project includes four types of crime in the data set as subset of the whole data set provided by CPD, which are narcotics, arson, theft and battery. The reason for choosing these four crimes is because these crimes have many records in 2011 and all of them are representative. There are 77 communities, and the data set record the number of crimes occurred in each community and population as well. Considering the population among different communities, the crime rate may be more meaningful to consider, thus it is defined as:

$$crime\ rate = \frac{the\ number\ of\ crime\ \ record\ in\ each\ community}{population\ in\ each\ community}$$

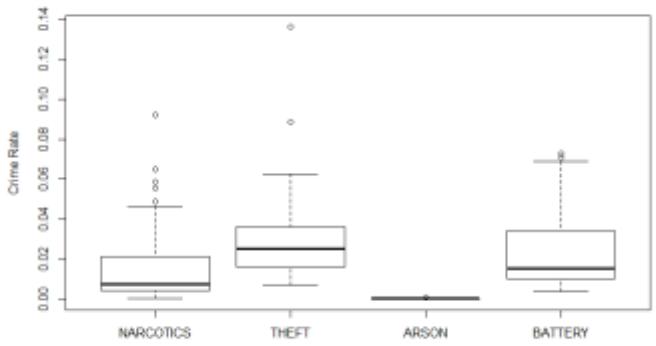

Figure 1: The boxplot of the crime rate

In term of the Figure 1, the mean among crimes varies widely, so the original data need to standardization.

Method

Dendrograms

A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. It shows the closeness of either individual data points or clusters in a visual way as the y-axes is a measure of distance. Basically, there are three cluster distance methods including single linkage, complete linkage and average linkage. These three linkages are commonly used in the dendrogram, and the one with the clearest cluster in the dendrogram and a distance value that will yield an appropriate number of clusters will be chosen. The goodness of the fit for dendrogram could be assessed by cophenetic correlation coefficient. This is the correlation between the original distances and those that result from the cluster configuration. Values above 0.75 are thought to be good. The cophenetic correlation coefficient and clustering situation is used to find the

best linkage. Comparing to the K-means, dendrogram is easier to decide on the number of clusters.

K-means

K-means clustering aims to partition all observations into k clusters in which each observation belongs to the cluster with the nearest mean. The problem for the K-means clustering is choosing the number of k. The projects define k from 2 to 3. In order to find the most appropriate k, the silhouette plot and the silhouette score are used. The silhouette score closes to 1 indicates that the data instance is close to the centre of the cluster and instances possessing the silhouette scores close to 0 are on the border between two clusters. Then the project uses the principle component analysis to interpret the clustering. K-means may be more appropriate to deal with large number of variables.

Result

According to **Table 1**, all three dendrograms are fitted well as the cophenetic correlation coefficients are above 0.75, and the dendrogram with average linkage might be the best dendrogram. Based on the figure below, four clusters might be appropriate when choosing the height around 15. There was a large cluster which accounted for around two thirds of communities, one cluster which accounted for around one fourth of communities, one cluster which only had one community named Loop and one cluster which includes small number of communities. The dendrogram with complete linkage might have three clusters which merge the Loop with a few communities though it still has large distances with other communities in the cluster. The dendrogram with single linkage showed poor clustering, as there was no clear number of clusters in the dendrogram.

Table 1 cophenetic correlation coefficient for three linkages

	Complete	Average	Single
cophenetic correlation coefficient	0.854	0.883	0.809

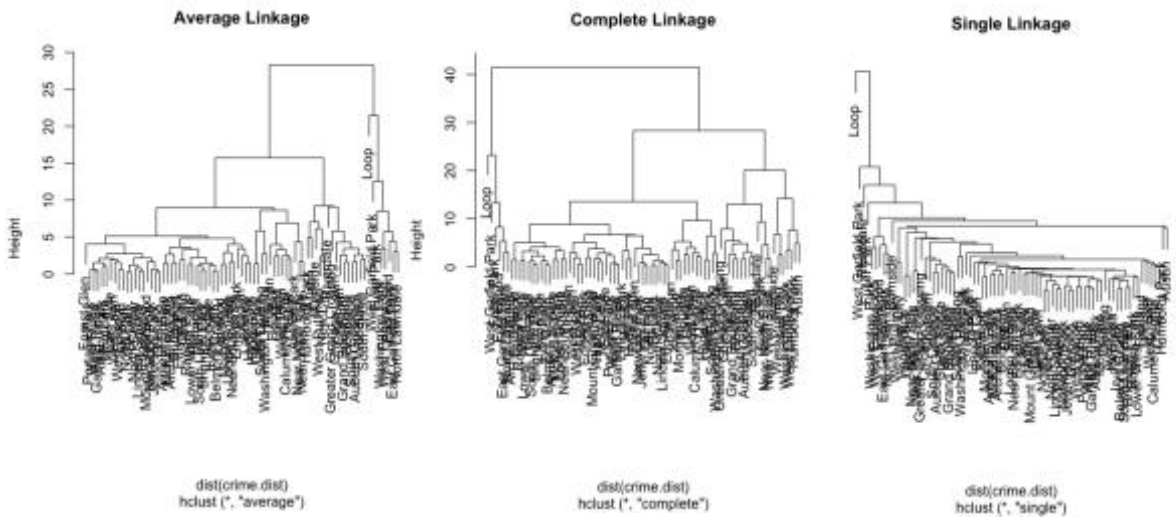

Figure 2: the dendrogram for the average linkage, complete linkage and single linkage.

When k equalled to 2 in the K-mean clustering, the average silhouette score is 0.52 which indicated a reasonable clustering. The first cluster had 23 communities and the second cluster had 54 communities. The silhouette score for the first cluster was only 0.23 and there were some bars with negative value which indicated some communities might be more similar with the second cluster. The silhouette score for the second cluster was 0.64, and it showed a good clustering.

The silhouette score for the K-means when k equal to 3 and 4 are 0.43 and 0.27 respectively, which were low, and many negative bars appeared in the plots, so the K-means clustering for the k equal to 3 or 4 were not very good.

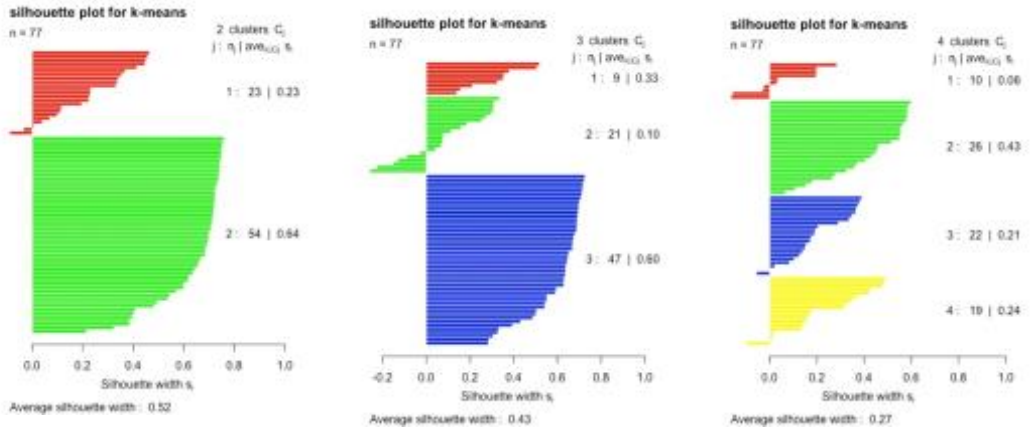

Figure 3: the silhouette plots for k=2,3,4 in the K-means clustering

The silhouette plot suggested that k equal to 2 might be the best choice, and the principal components analysis (PCA) is used to interpret the cluster.

According to the **Figure 4**, the two components explained 86.08 percent of the point variability, and there was no overlap between clusters. In the first component, two clusters were separate clearly while in the second component, the range of the first cluster contained the second cluster, so the second component was not informative.

In term of the **Table 2**, the first component could be written as

PC1=-0.578*Narcotics-0.3*Theft-0.458*Arson-0.583*Battery

The second component could be written as

PC2=0.826*Theft-0.558*Arson

Looking at the x axes of **Figure 4**, the four crime rates in the first cluster were higher than the second cluster.

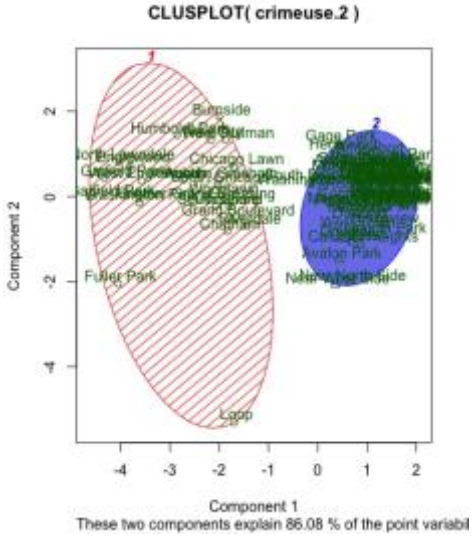

Figure 4: the clusplot of two component analysis

Table 2: the loading of PCA

Component 1	-0.578	-0.341	-0.458	-0.583
Component 2	0	0.826	-0.556	0

Then using the scatter plot to compare dendrogram and K-means clustering, the K-means showed a clearer separation than the dendrogram.

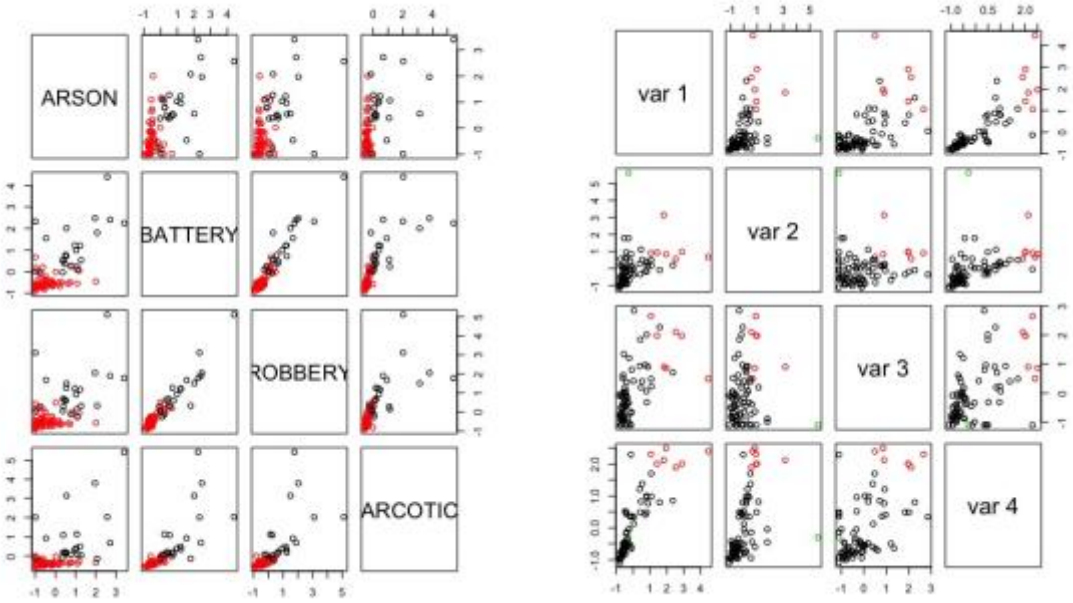

Figure 5: the scatter plots of the K-means (left) and the dendrogram (right)

Conclusion and Discussion

According to the result, K-means clustering is better than dendrograms, as the clusters in the K-means shows better separation than the dendrogram. As for the dendrogram, there is a strange community Loop which has a large distance with all other communities. The existence of Loop causes additional cluster. Besides, 77 communities are quite large, so the dendrogram looks disorganized, and K-Means produces tighter clusters than dendrogram.

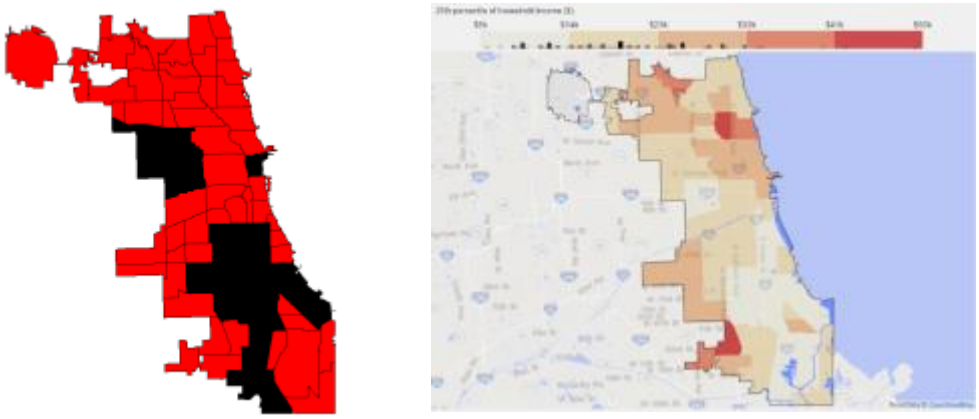

Figure 6: the map of Chicago which divide by the K-means and Map of Household Income by Neighbourhood in Chicago

In the first figure of the **Figure 6** shows the two clusters by K-means for the 77 communities in Chicago, the black is the first cluster and the red is second cluster. So those communities in black have higher crime rate for narcotics, arson, theft and battery. In term of the map of household Income by Neighbourhood in Chicago which published by the website Statistics Atlas, the most communities with high crime rate have low household income, it suggests that the low income is the one reason of high crime rate.

Additionally, if the project can sub-categorized these variable, for example, the number of arson which occurred during the working time for police and the number of arson which occurred during the off time for police. Besides, the variable the distance between and CPD and community could be included in the data set. These variables might reflect the effective of the CPD, so can be considered in the data set.