# MTH 5401 Final Project

**David E. Nieves-Acaron ### MTH 5401 ### Dr. Nezamoddini-Kachouie Nezamoddin**

## Abstract

Learning to model data is a critical skill for many areas of research, particularly when said data presents an opportunity for real world intelligence-based decision as is seen in many top-performing companies and organizations. In this project, a rudimentary analysis of the `state.x77` database is performed by modelling two response variables, by grouping these responses, and by clustering the associated data points. In particular, income and life expectancy of all fifty states of the union is performed by using a variety of linear models. The grouping is used to then analyze the average life expectancy of three groups (low income, medium income, and high income). Finally, some additional groups are formed using K-Means and Fuzzy C-Means clustering. The results consist of the following: the `Income` and `Life Exp` data contains some important correlations which contributed to the development of the model as seen in the **Results** section; some models were developed using a combination of a brute force and a genetic algorithm for the interaction terms, with the Adjusted $R^2$ reaching values as high as 0.7876 and 0.7305 for `Income` and `Life Exp`, respectively; finally, the ANOVA results for the `HS Grad` and `Life Exp` values between different income groups showed that there is indeed a significant difference between the Medium Income and High Income groups and the Low Income groups in terms of Life Expectancy, as well as a significant difference in High School graduation rates; the results of the clustering show that the best number of clusters was around 5 or 6, with the best features involving the `Population`, `Area`, and `HS Grad`.

### Income Response modelling

The main model type used for modelling the income response consisted of linear regression models.

### Life Expectancy Response Modelling

### Grouping by Income

### K-Means and Fuzzy C-Means Clustering

## Introduction

1. Introduction (Arial 11 Bold) Introduce the problem, cite related works, explain the goal, and describe your approach.

The data for this project comes from a 1977 census by the United States Department of Commerce as seen in [1]. It has 8 columns consisting of the following: (state name); `Population Income`, the per capita income in 1974; `Illiteracy`, the illiteracy as a percent of the population in the year 1970; `Life`

`Exp`, the life expectancy in the years 1969 to 1971; `Area`, the land area in square miles; `Murder`, the murder and non-negligent manslaughter rate per 100,000 population in 1976; `HS Grad`, the High School graduation rate in 1970; finally, `Frost`, mean number of days with minimum temperature below freezing (1931-1960) in a capital or large city. The goal is to model two of the variables: `Life Exp` and `Income` with respect to the other predictors.

The importance in this type of exercise lies in using one's analytical skills to be able to make real-world intelligent decisions. Given the nature of the data being worked with, one could easily see this type of analysis being performed for policymakers who would like to perform more intelligence-based decision making.

## Methods

Up to two pages. (Arial 11 Double Space) 2. Methods (Arial 11 Bold) Explain the probability and stat theory that has been used in this work. Write relevant equations, probability distributions, and parameters.

Linear regression models

A simple linear regression model consists of a model of the form:

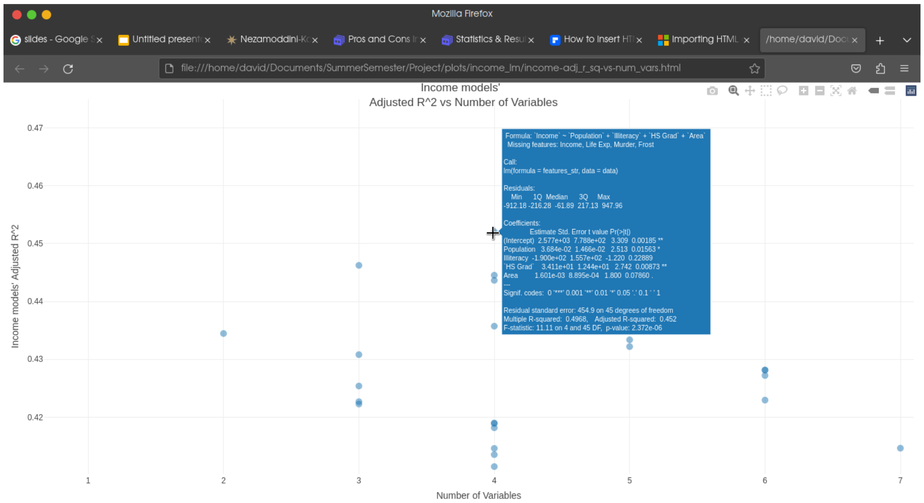
$$\hat{y_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

In general, the more predictors there are, the higher $R^2$, which can lead to misleading results. As a result, the Adjusted $R^2$ metric was developed since

$$\vec{y} = \vec{\beta} \cdot \vec{x}$$

$\ldots$ where $y_i$ is of size $n$, $\vec{\beta}$ is of size $n \times k$, and $x$ is of size $n \times 1$.

Given the relatively cheap computational performance of $\approx O(2^n)$ combinations of linear regression models on a modern day computer, all the combinations were attempted for the *base* models, with these referring to ones which contain no interaction terms. From there, a the use of the Plotly library ensued to display not only the selected axes corresponding to the models' metrics (with these being Adjusted $R^2$ vs number of variables), but also the printout of each model's summary using the hovertext feaure of the library. An example of this can be seen below:



```
Formula: `Income` ~ `Population` + `Illiteracy` + `HS Grad` + `Area`
 Missing features: Income, Life Exp, Murder, Frost

Call:
lm(formula = features_str, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-912.18 -216.28  -61.89  217.13  947.96

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.577e+03  7.788e+02   3.309  0.00185 **
Population   3.684e-02  1.466e-02   2.513  0.01563 *
Illiteracy  -1.900e+02  1.557e+02  -1.220  0.22889
`HS Grad`    3.411e+01  1.244e+01   2.742  0.00873 **
Area         1.601e-03  8.895e-04   1.800  0.07860 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 454.9 on 45 degrees of freedom
Multiple R-squared:  0.4968,    Adjusted R-squared:  0.452
F-statistic: 11.11 on 4 and 45 DF,  p-value: 2.372e-06
```

A list of all combinations of base variables was made using the `combn` function from R. This allowed for the creation of all possible combination of predictors in a single formula as a string object. Once that collection of string objects was created, the members of the collection were used to analyze the performance of the linear regression model. An interactive representation of the models' performances was created using the Plotly library in R. The benefit of Plotly mainly lies in the customizability and the feature of allowing for hover labels which showcase a printout of each model's summary when hovering over a specific model's dot (see picture below).

The reader is *highly* encouraged to inspect these plots using a web browser.

The Adjusted $R^2$ metric was used due to it being able to account for model sizes better.

After this was performed, the best model was chosen from the "base model", i.e. the one which had no interaction terms. From this base model, some interaction terms were added to test and see how they would affect model performance. Given the large amount of combinations present in the interaction terms, a simple algorithm was used for this, whereby out of all the possible set of interactions, the interaction which best affected the model was chosen to further enhance it and hopefully improve. Specifically, the model with the interaction term which most increased the adjusted $R^2$ was chosen. From this, the model with the maximum adjusted $R^2$ was chosen.

For clustering, a similar approach was taken by seeking to minimize the within cluster sum of squares while maximizing the between cluster sum of squares over the total sum of squares.

An ANOVA test attempts to determine if there is a significant difference between the averages of two groups. It does so by calculating what is known as an $F$ statistic and testing it. In its essence, it is a ratio of two variances. The motivation behind it lies in the fact that in a set of data points organized into different groups, the variation within the data can be broken into variation between the groups and variation between the groups. As a result, a data distribution that exhibits a relatively small within-group variation and a relatively large between-group variation should indicate the significance of the groups, because if the groups were not significant, the within-group variation would not differ much from the between-group variation.

For that reason, one tests the ratio of the within-group variation and between-group variation to see if it is significant enough using the $F$ distribution.

Regarding the F-distribution, for any ratio of two chi-squared random variables $\chi_1^2$ and $\chi_2^2$, with $n_1$ and $n_2$ degrees of freedom respectively, the ratio of the two like so:

$$F = \frac{\frac{\chi_1^2}{n_1}}{\frac{\chi_2^2}{n_2}}$$

has an $F$ distribution with $n_1$ and $n_2$ degrees of freedom, respectively.

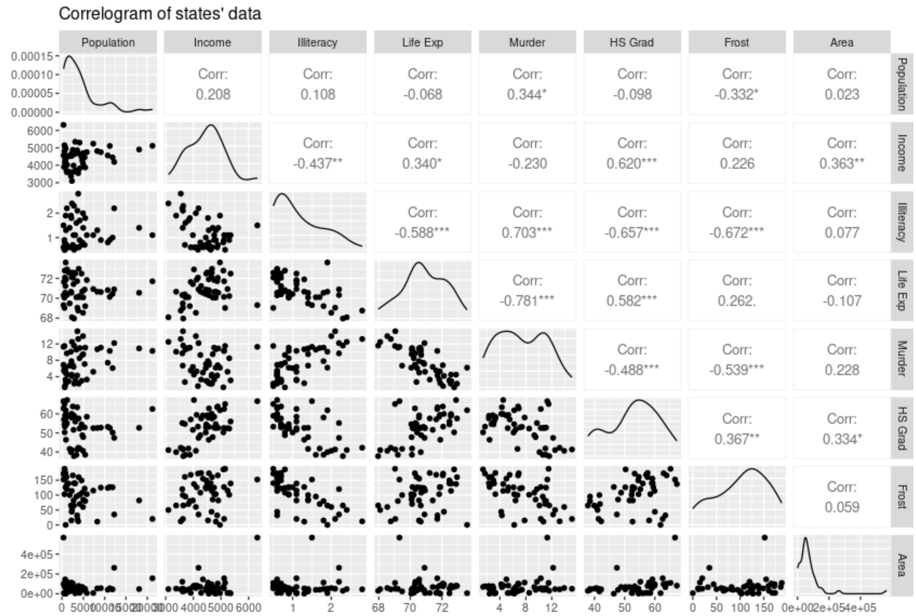
The $F$-statistic consists of

$$F = \frac{\text{between-group variation}}{\text{within-group variation}}$$

## Results

Up to two pages. (Arial 11 Double Space) 3. Results (Arial 11 Bold) Include relevant figures/graphs/plots/tables (ordered numbers) and explain each of them separately in the text.

One of the first tasks undertaken in this project was analyzing the data by visualizing a correlation matrix using the `ggpairs` function. This resulted in the following plot seen below.



As can be seen from this correlogram, there are some notable correlations for the two response variables of interest, `Income`, and `Life Exp`. For `Life Exp`, the strongest correlations correspond to `Illiteracy` (-0.588), `Murder` (-0.781), `HS Grad` (0.582), and `Income` (0.340). For the `Income` variable, the strongest correlated variable is `Illiteracy` (0.620), but there are other terms such as `Life Exp` (0.340), `Illiteracy` (-0.437), and `Area` (0.363) which still have a medium to strong correlation. For this reason, it is no surprise that later on, the models with some of the strongest results made use of these predictors. In addition to these correlations, it can be seen from this plot that the distributions of each of the predictors and responses are far from normal, which undermines the assumptions of linear regression.

The best model for the `Life Exp` response was the following:

$$\begin{aligned}
\text{Life Exp } & \text{Population} + \text{Murder} + \text{HS Grad} + \text{Frost}+ \\
& \text{Population} : \text{Murder} + \text{Population} : \text{Frost}+ \\
& \text{Population} : \text{Area}+ \\
& \text{Murder} : \text{Frost}+ \\
& \text{Illiteracy} : \text{Murder}+ \\
& \text{Illiteracy} : \text{HS Grad}+ \\
& \text{Murder} : \text{HS Grad}+ \\
& \text{Income} : \text{Frost}+ \\
& \text{Income} : \text{Murder}+ \\
& \text{Frost} : \text{Area}+ \\
& \text{HS Grad} : \text{Area}
\end{aligned}$$

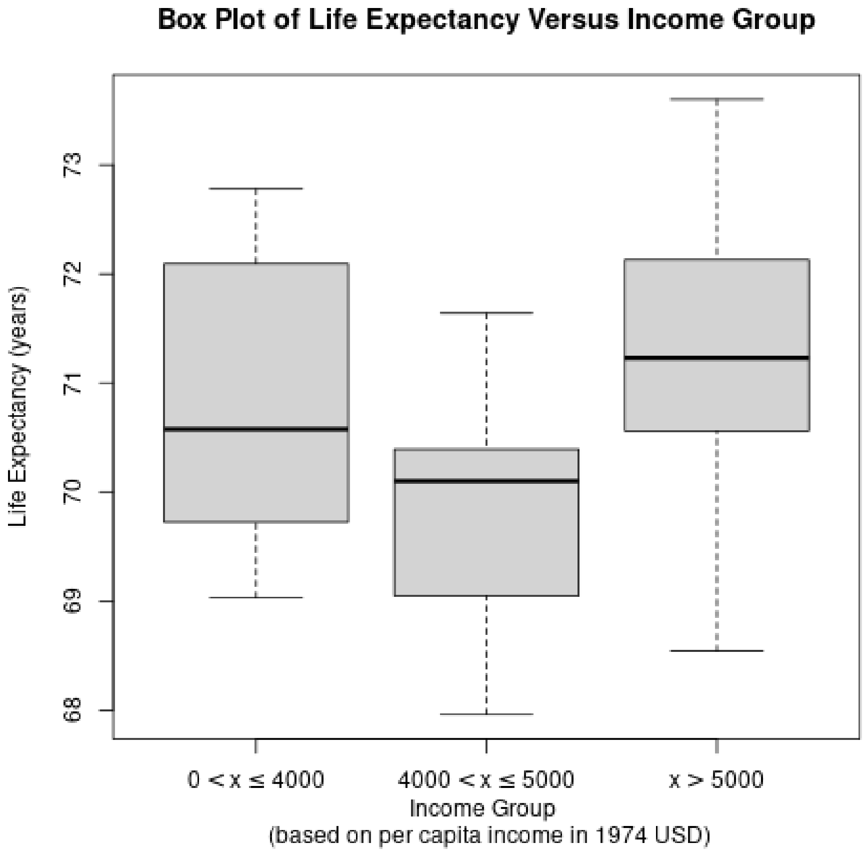
It obtained an $R^2$ score of 0.7876 and an Adjusted $R^2$ score of 0.8526.

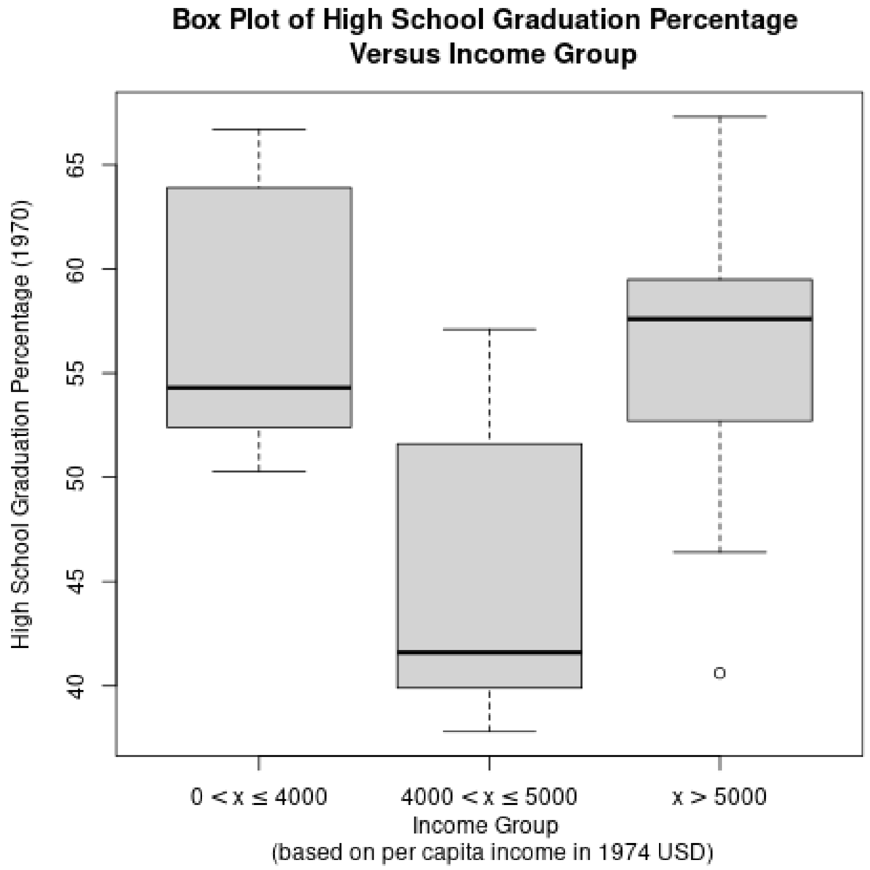
The best model for modelling the `\text{Income}` response was the following:

$$\begin{aligned}
\text{Income} \sim {} & \text{Population} + \text{Murder} + \text{HS Grad} + \text{Frost} \\
& \text{Murder} : \text{HS Grad} + \text{Frost} : \text{Area}+ \\
& \text{Life Exp} : \text{Area} + \text{Murder} : \text{Area}+ \\
& \text{Illiteracy} : \text{Murder} + \text{Life Exp} : \text{Frost}+ \\
& \text{Life Exp} : \text{Murder} + \text{Illiteracy} : \text{HS Grad}+ \\
& \text{Population} : \text{Illiteracy} + \text{HS Grad} : \text{Frost}+ \\
& \text{Population} : \text{Frost} + \text{Population} : \text{Area}+ \\
& \text{Life Exp} : \text{HS Grad}
\end{aligned}$$

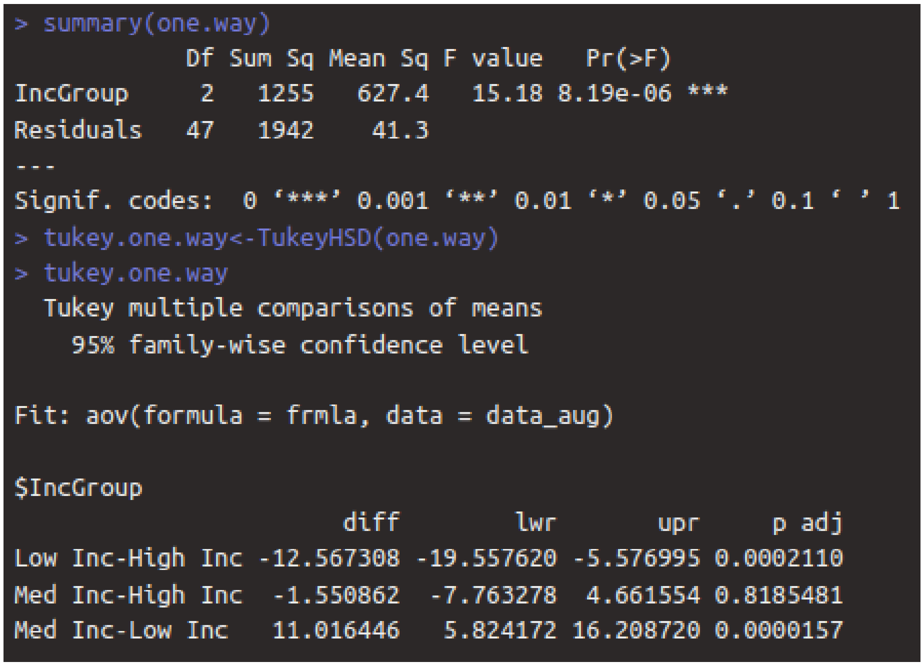
It obtained an $R^2$ score of 0.824 and an Adjusted $R^2$ score of 0.7305.

After that, some beginning modelling was performed using the `Life Exp` predictor. This allowed for the easy inspection of models to be able to

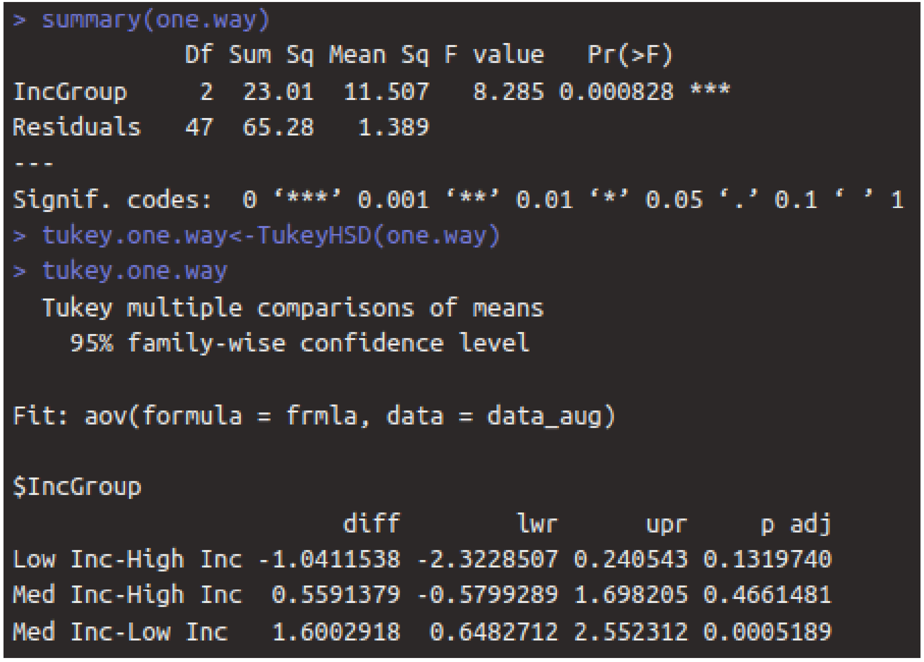

Box Plot of Life Expectancy Versus Income Group
Life Expectancy (years)
68
69
70
71
72
73
0 < x ≤ 4000
4000 < x ≤ 5000
x > 5000
Income Group
(based on per capita income in 1974 USD)



From the results of the above plot, there is clearly a difference between the averages of the middle income group and the other two groups. This is further validated by the results of the ANOVA test seen below.

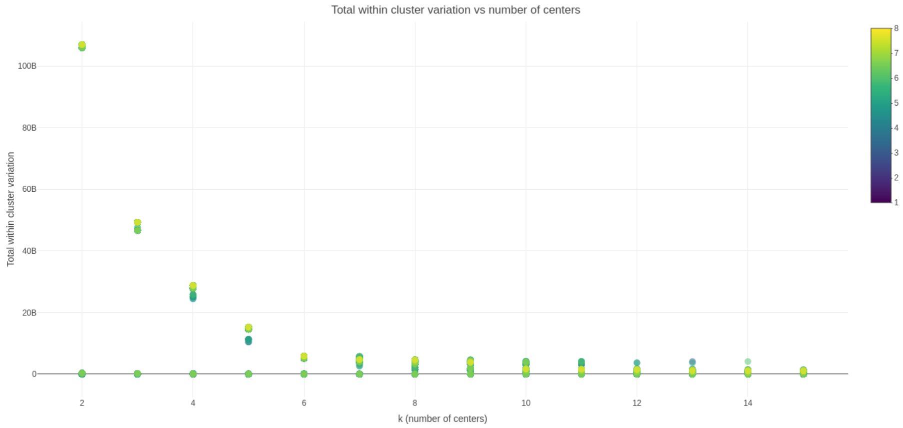


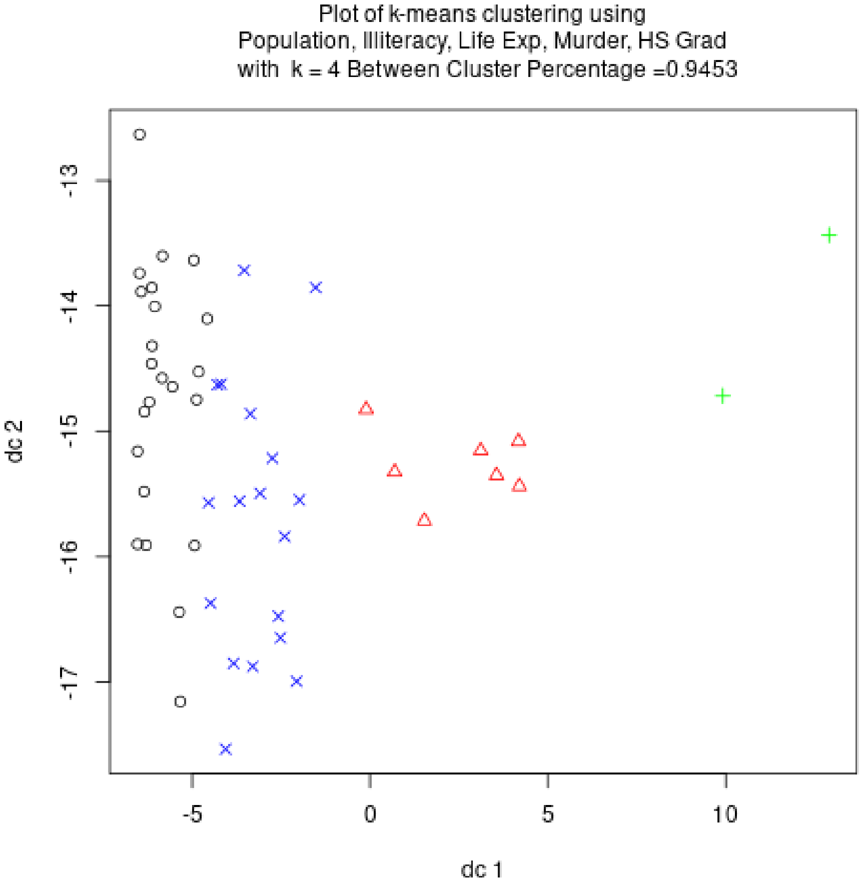
```
> summary(one.way)
            Df Sum Sq Mean Sq F value   Pr(>F)    
IncGroup     2   1255   627.4   15.18 8.19e-06 ***
Residuals   47   1942    41.3                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> tukey.one.way<-TukeyHSD(one.way)
> tukey.one.way
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = frmla, data = data_aug)

$IncGroup
                          diff        lwr       upr     p adj
Low Inc-High Inc -12.567308 -19.557620 -5.576995 0.0002110
Med Inc-High Inc  -1.550862  -7.763278  4.661554 0.8185481
Med Inc-Low Inc   11.016446   5.824172 16.208720 0.0000157
```

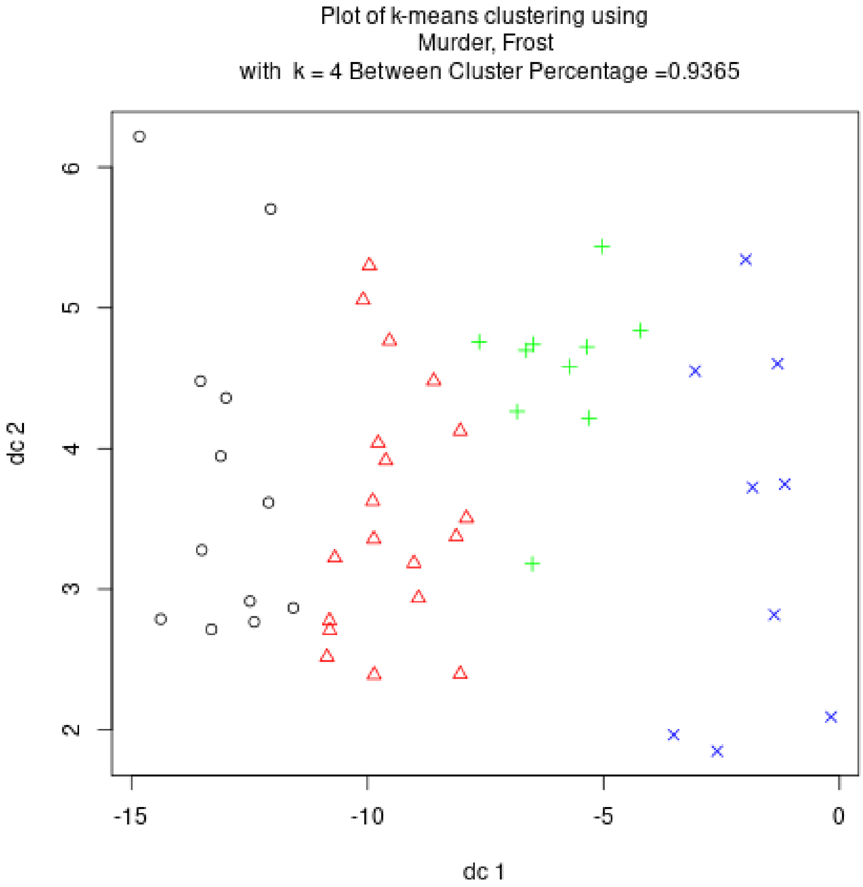


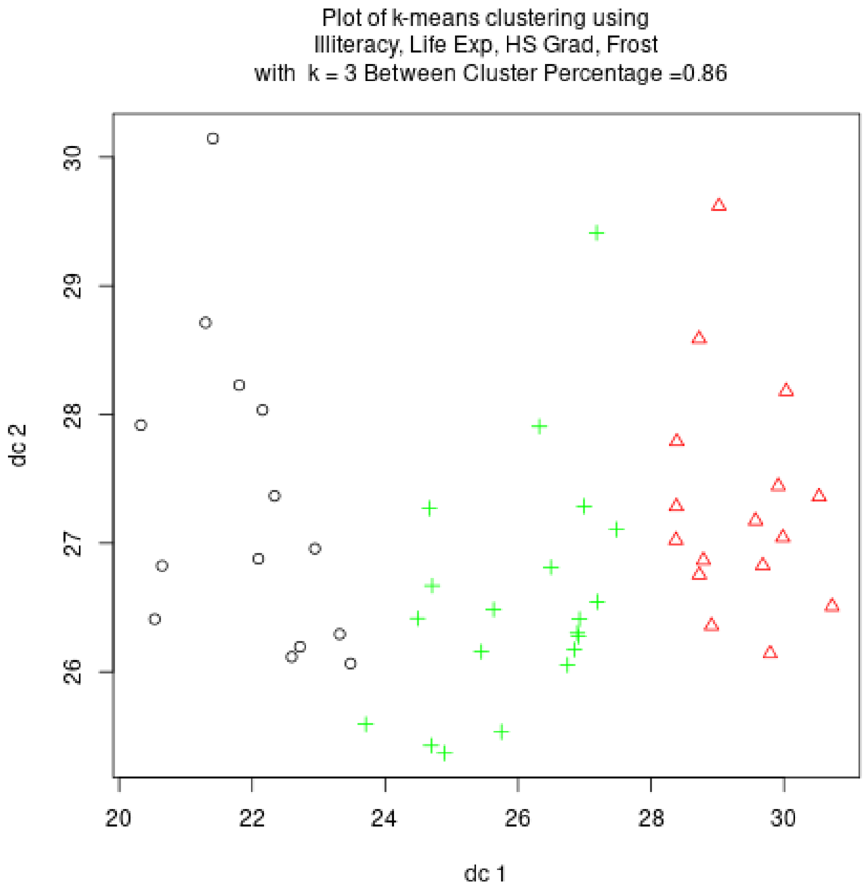
```
> summary(one.way)
            Df Sum Sq Mean Sq F value   Pr(>F)    
IncGroup     2  23.01  11.507   8.285 0.000828 ***
Residuals   47  65.28   1.389                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> tukey.one.way<-TukeyHSD(one.way)
> tukey.one.way
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = frmla, data = data_aug)

$IncGroup
                        diff        lwr      upr     p adj
Low Inc-High Inc -1.0411538 -2.3228507 0.240543 0.1319740
Med Inc-High Inc  0.5591379 -0.5799289 1.698205 0.4661481
Med Inc-Low Inc   1.6002918  0.6482712 2.552312 0.0005189
```

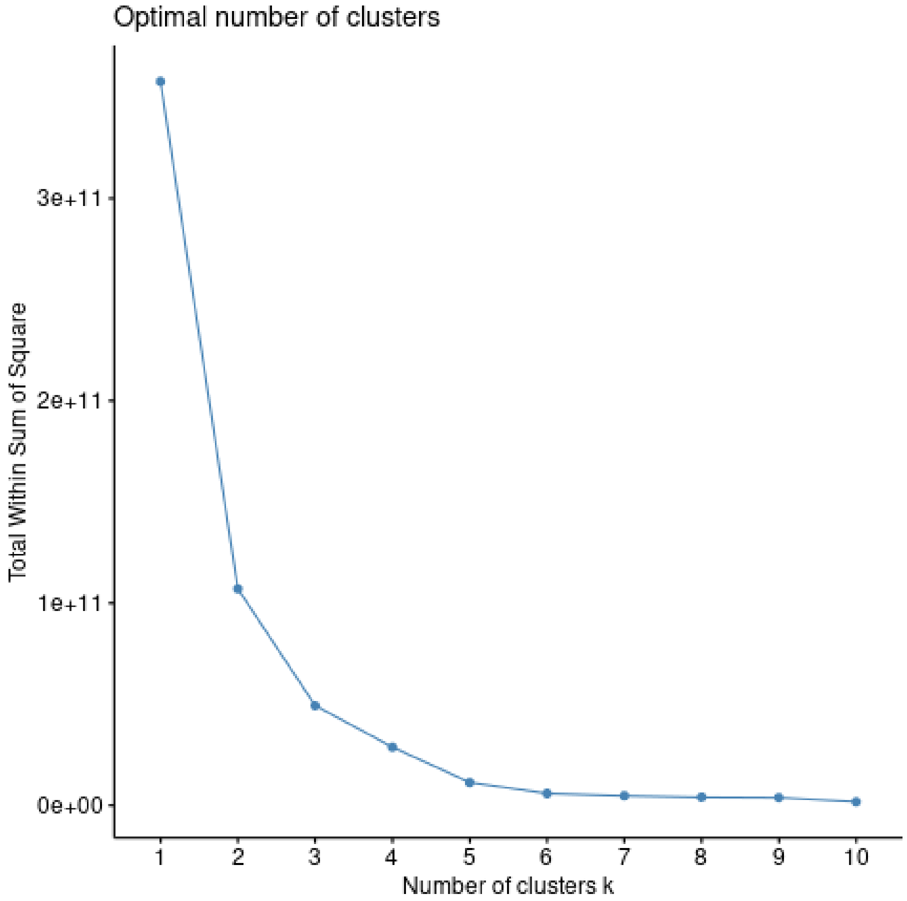
Moreover, the results of the Tukey Honest Significant Differences test from using the `TukeyHSD` function reveals that the main difference in the `HS Grad` between income groups lies between the Low Income and High Income as well as the Medium Income and Low Income.


Total within cluster variation vs number of centers
Total within cluster variation
100B
80B
60B
40B
20B
0
2
4
6
8
10
12
14
k (number of centers)
8
7
6
5
4
3
2
1


Plot of k-means clustering using
Population, Illiteracy, Life Exp, Murder, HS Grad
with k = 4 Between Cluster Percentage =0.9453
dc 2
-13
-14
-15
-16
-17
-5
0
5
10
dc 1


Plot of k-means clustering using
Murder, Frost
with k = 4 Between Cluster Percentage =0.9365
dc 2
dc 1
-15
-10
-5
0
2
3
4
5
6


Plot of k-means clustering using
Illiteracy, Life Exp, HS Grad, Frost
with  k = 3 Between Cluster Percentage =0.86
dc 2
30
29
28
27
26
20
22
24
26
28
30
dc 1



Up to 10 pages. (Arial 11 Double Space) 4. Discussion and Conclusions (Arial 11 Bold) Discuss your observations, explain the outcomes, and propose the future work to further

In conclusion, the results of the `Income` and `Life Exp` response modelling show that the best predictors for both were `Population`, `Murder`, `HS Grad`, and `Frost`. Adding interaction terms improved the performance of the models significantly. As can be seen in the modelling of the `Income` response variable, it caused an increase in the Adjusted $R^2$ score from 0.452 to 0.7305. For the ANOVA results, it was seen that there exists a significant difference between the High School Graduation Percentage, and the Life Expectancy of different income groups. With regards to the clustering results, it was seen that the optimum number of clusters was around 5 or 6, and the `Population`, `Area`, and `HS Grad` predictors were some of the most important ones. For future efforts, a deep dive could be performed so as to analyze the normality of the residuals of each model to take it into account with the results of the other models and thus

In addition, ANOVA tests for all the models as well as a more comprehensive investigation of the interaction terms could be undertaken to take into account other relationships that could exist between the different predictors such as perhaps nonlinear relationships.

Finally, the following will attempt to assign some meaning to the results of this project. Following the bygone yet still present political movement known as rationalism, one's actions and decisions should be based on reason. While this can hold some meaning in one's personal life, this is mainly with respect to the actions undertaken by governments around the world, as their actions can have consequences on many people, even outside of their own jurisdiction. For that reason, gathering intelligence and analyzing data is ever so crucial, which is why this project is relevant. One conclusion that could be obtained from this project is that it seems that if one were to wonder where to focus on raising High School graduation rates with respect to income groups, the low income groups would be the natural choice as they have been shown to have a different mean High School graduation rate with respect to the others.

address the problem. Up to four pages. (Arial 11 Double Space) References (Arial 11 Bold) One page. (Arial 11 Double Space)

## References

U.S. Department of Commerce, Bureau of the Census (1977) Statistical Abstract of the United States, and U.S. Department of Commerce, Bureau of the Census (1977) County and City Data Book