COMP 381: Machine Learning

Term Project

Population and Household Projection

Name	*Student Id.*
Chaahna Chandiramani	300198011

I. Executive Summary

The proposed project aims to study the population demographics and household projections in different regional districts. We would use machine learning techniques to analyse the data and gain valuable insights into the characteristics and trends of these districts. By exploring questions related to age distribution, gender balance, and household patterns, we hope to understand how these factors influence regional populations.

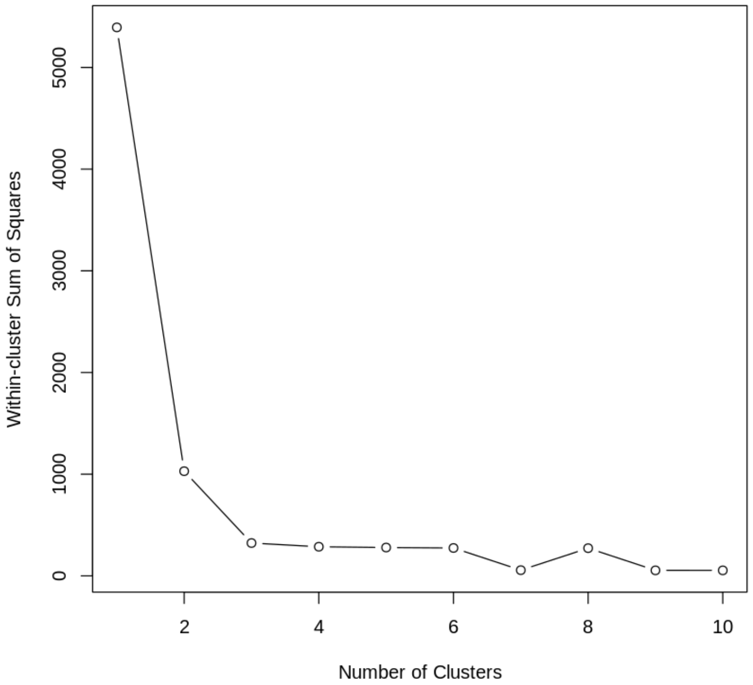
To achieve this, we will work with available datasets, clean the data, and apply machine learning models to make sense of the information. Our main goal is to identify groups of regional districts with similar age distributions and gender proportions. This analysis will help us understand demographic trends and predict future population changes. By using clustering, classification, and regression techniques, we will be able to make predictions and classify regional districts based on their population attributes.

The expected results of this project include valuable insights into changes in population composition, gender distribution, household sizes, and their impact on housing demands and infrastructure planning. We aim for a confidence level of at least 85% in our predictions. By discovering patterns and connections in the data, we hope to provide useful information for decision-making, policy development, and planning in regional districts.

II. Findings

Clustering:

K-means clustering was chosen as the clustering algorithm and the normalized gender distribution columns were selected as features for clustering. We calculated and plotted the within-cluster sum of squares (WCSS) for different numbers of clusters using the K-means clustering algorithm. The plot of the WCSS values allows for visual examination of the "elbow point" in the curve. This helps determine the optimal number of clusters to use in the K-means clustering analysis. The WCSS provides an indication of how well the data points within each cluster are grouped together. Lower WCSS values indicate tighter and more homogeneous clusters. By comparing the WCSS values for different numbers of clusters, it becomes possible to evaluate the trade-off between cluster quality and the practicality of having a higher or lower number of clusters.



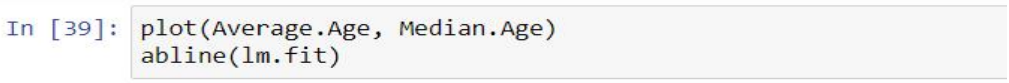
Classification:

The logistic regression model is used to predict the gender distribution of a regional district's population based on the age distribution. The model provides coefficient estimates for each age group variable (AGroup1 to AGroup6) and an intercept. These coefficients indicate the direction and significance of the association between each age group and the probability of gender distribution. The model's residual deviance and AIC are provided as measures of model fit.

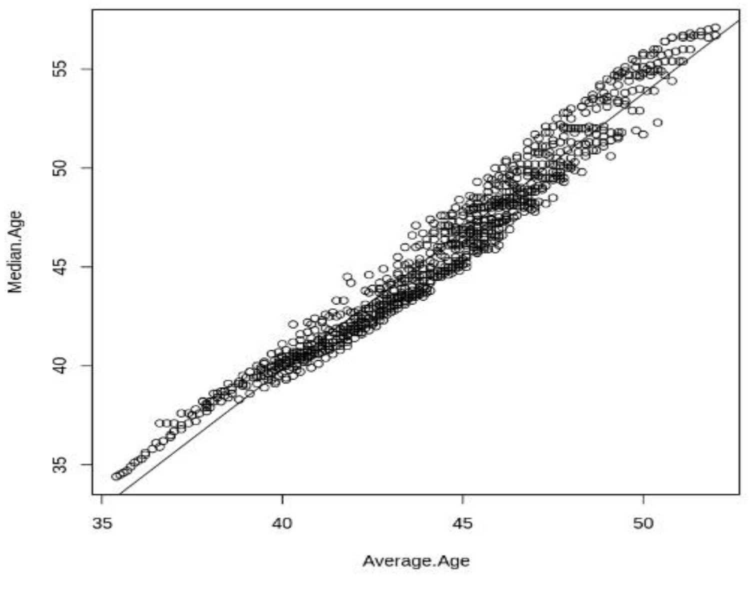
Logistic regression was also used to classify regional districts into different household types based on the average number of persons per household and the age distribution. Here, we merge 2 tables from our data and work specifically on Agroup4 and year 2022. The logistic regression model provides coefficients for district variables and the average number of persons per household (AGroup4). The coefficients indicate the influence of each predictor on the log-odds of belonging to a specific household type. The model's summary includes p-values, null deviance, residual deviance, and AIC as measures of model significance and fit.

A multinomial logistic regression model is used to predict the regional district of a household based on its demographic characteristics, such as age distribution and average number of persons per household. It includes coefficients and standard errors for each predictor variable (AGroup1 to AGroup6) in relation to the districts. The residual deviance being high shows that this model may not be the best fit for this prediction.

Regressions:
To answer the question if we can predict the future population growth rate of a regional district on its current median age and average age, we plotted a graph taking y-axis as median age and x-axis as average age based on the regional district. Upon studying the outcome in the graph, we can see that there is significant growth in the median age when the average age is increasing. So, this answers the above answers as yes, we can predict the future population based on the regional district. Using the same graph below



```
In [39]: plot(Average.Age, Median.Age)
         abline(lm.fit)
```



we can estimate more further future population growth of a regional district.
To find if there is a correlation between the median age and the average household size, we calculated the correlation coefficient to quantify the strength and direction of the relationship between the median age and average household size. We found that the coefficient was close to 0 which determines that there is no correlation between median age and average household size.
To estimate the future population size of a regional district based on its current household projections we used the same method as above to create a graph between the household projections and regional district, but the predictions were not as accurate and it seemed difficult to predict if we can compare these two in order to figure out a different outcome. So, the answer to this general question would be no, we cannot estimate the future population size of the regional district based on its current household projections.

III. Data sets used in this project:

Population and household projection by Regional District for 2019-2028; from BC Data Catalogue

Access to the data set on the BC Data Catalogue:
https://catalogue.data.gov.bc.ca/dataset/population-and-household-projections-2019-2028-/resource/d39bda1b-b9a9-4d2c-8c83-2eb1d8b6dbcc

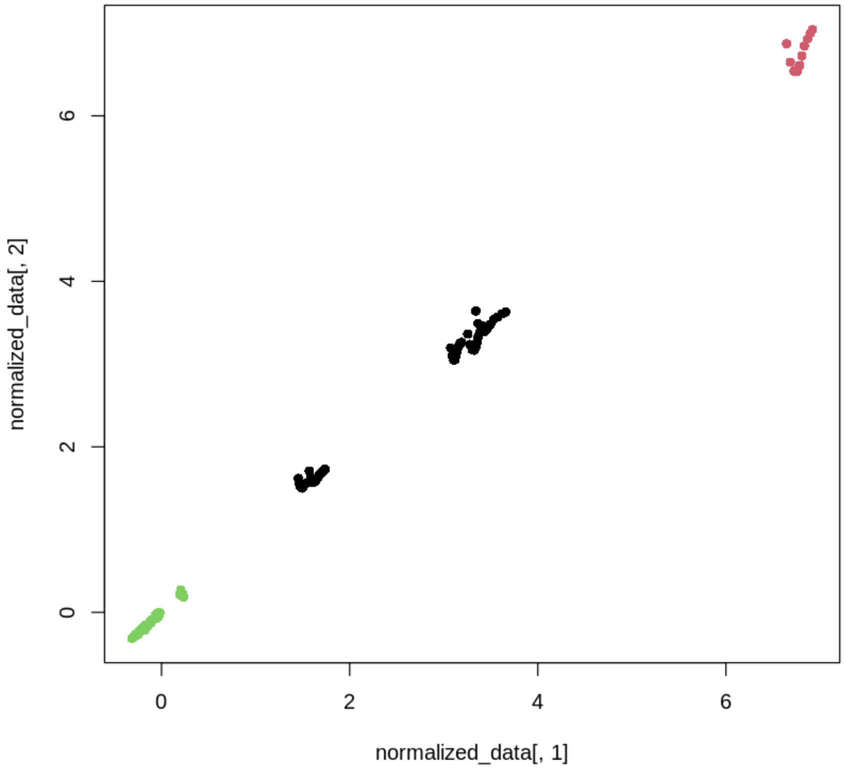
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page_Download-Telecharger.cfm?Lang=E&Tab=1&Geo1=PR&Code1=59&Geo2=PR&Code2=01&SearchText=&SearchType=Begins&SearchPR=01&B1=Housing&TABID=1&type=1

Access to the data set on google sheet:
https://docs.google.com/spreadsheets/d/1fS4KwSlAIrqx1ir9hlW-OGjwFpAbuyGQ/edit?usp=sharing&ouid=110023239984027275402&rtpof=true&sd=true

IV. Implementing Machine Learning concepts

A. Clustering

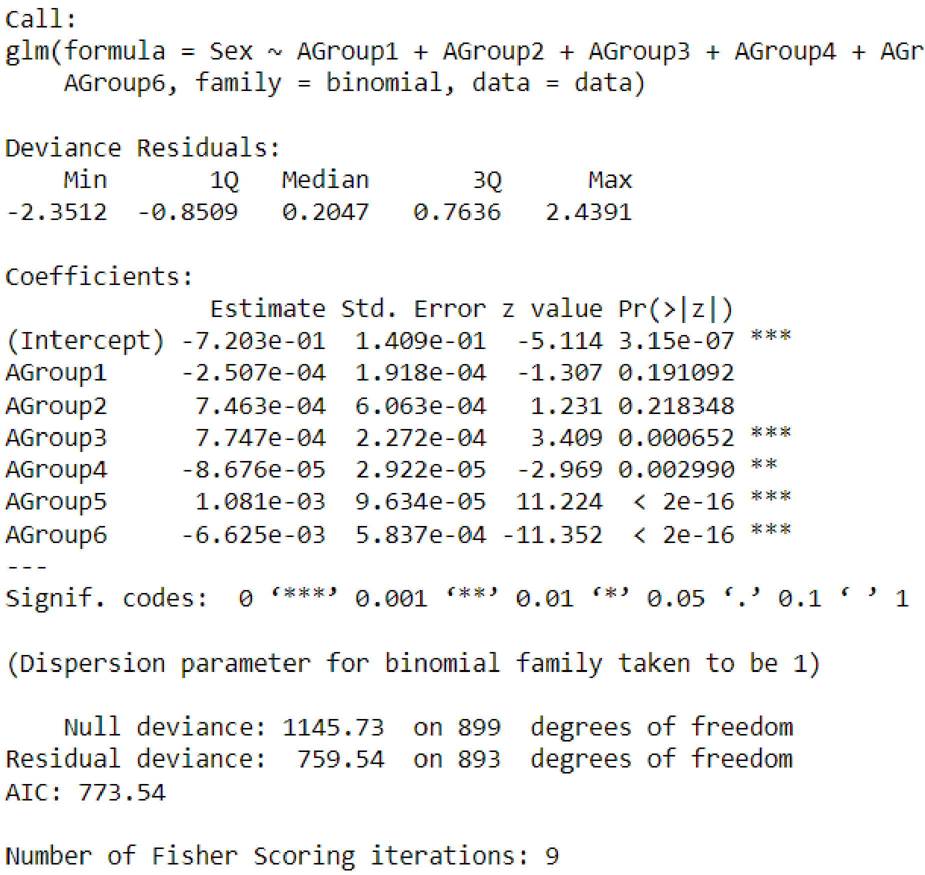
As per our proposal, we used K-means clustering. The clustering models analysed patterns and similarities in the population and household data to create clusters. We performed clustering analysis to explore patterns and identify clusters of regional districts based on their household projections. We loaded the "Population.csv" file and selected relevant columns, including regional district information and attributes related to age groups or household projections. To ensure meaningful comparisons, we normalized the selected data using the scale() function. This step involved scaling the data to have zero mean and unit variance. within-cluster sum of squares (WCSS) was then calculated for different numbers of clusters using the K-means algorithm. This allowed us to evaluate the clustering quality and determine the optimal number of clusters. We then aggregated and summarized the normalized data attributes within each cluster. Additionally, the clusters are visualized using scatter plots, where the color of the points represented the assigned cluster labels.



B. Classification

Testing for possibilities to predict the gender distribution of a regional district's population based on the age distribution? Results interpreted based on the model summary are as follows:

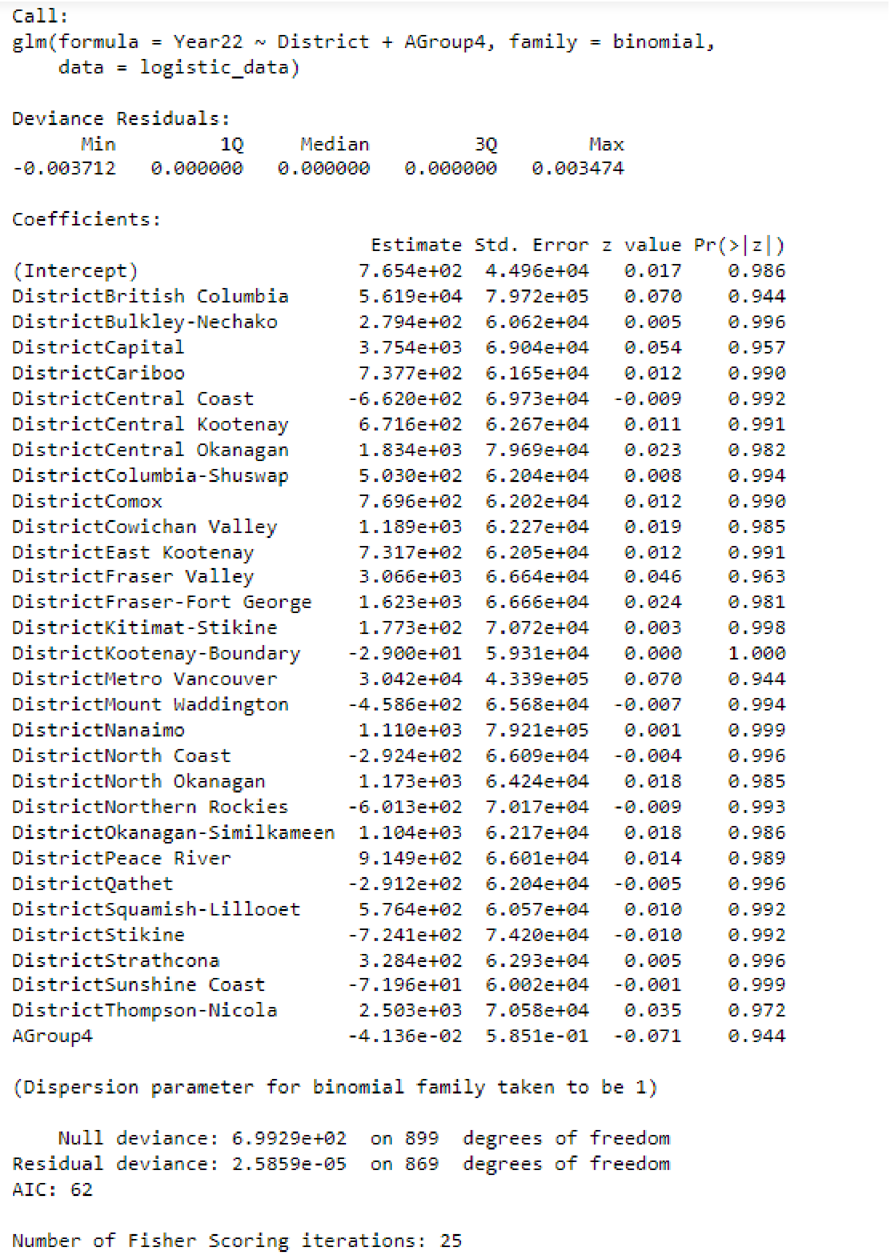
- The estimated intercept coefficient is -0.7203, indicating the log-odds of the baseline category of the dependent variable (Sex) when all age groups are zero.
- The coefficient estimate for AGroup1 and AGroup2 suggests that there is no significant association between AGroup1 and the probability of the dependent variable (Sex).
- The coefficient estimate for AGroup3 indicates that AGroup3 has a positive and statistically significant association with the probability of the dependent variable (Sex).
- The coefficient estimate for AGroup4 and AGroup6 suggests that they have a negative and statistically significant association with the probability of the dependent variable (Sex).
- The coefficient estimate for AGroup5 indicates that AGroup5 has a positive and highly statistically significant association with the probability of the dependent variable (Sex).
- Overall, the coefficient estimates indicate the direction and significance of the associations between each age group and the probability of the dependent variable.



```
Call:
glm(formula = Sex ~ AGroup1 + AGroup2 + AGroup3 + AGroup4 + AGroup5 + 
    AGroup6, family = binomial, data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.3512  -0.8509   0.2047   0.7636   2.4391  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -7.203e-01  1.409e-01  -5.114 3.15e-07 ***
AGroup1     -2.507e-04  1.918e-04  -1.307 0.191092    
AGroup2      7.463e-04  6.063e-04   1.231 0.218348    
AGroup3      7.747e-04  2.272e-04   3.409 0.000652 ***
AGroup4     -8.676e-05  2.922e-05  -2.969 0.002990 ** 
AGroup5      1.081e-03  9.634e-05  11.224  < 2e-16 ***
AGroup6     -6.625e-03  5.837e-04 -11.352  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1145.73  on 899  degrees of freedom
Residual deviance:  759.54  on 893  degrees of freedom
AIC: 773.54

Number of Fisher Scoring iterations: 9
```

The logistic regression model fitted is aimed at classifying regional districts into different household types based on the average number of persons per household and the age distribution. Here, only year 2022 and age group 4 is taken into account. The model summary provides information about the coefficients and their statistical significance and is interpreted below:

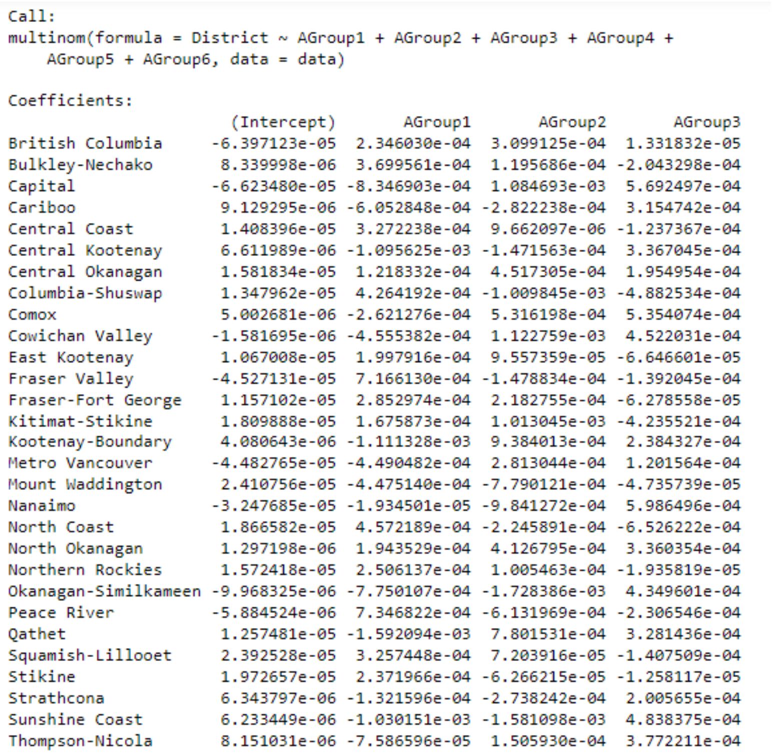
- The intercept term represents the baseline log-odds of belonging to a particular household type when all other predictors are zero.
- District coefficients: The coefficients associated with each district indicate how they influence the log-odds of belonging to a specific household type compared to the reference district
- For example, the coefficient for "District - British Columbia" indicates that this district has a higher log-odds of belonging to a particular household type compared to the reference district.
- AGroup4 coefficient: A negative coefficient suggests that an increase in AGroup4 tends to decrease the log-odds of belonging to the household type, although it is not statistically significant in this case.
- Pr(>|z|): A lower p-value indicates stronger evidence against the null hypothesis. Here, it has no effect.
- A lower residual deviance suggests a better fit to the data.



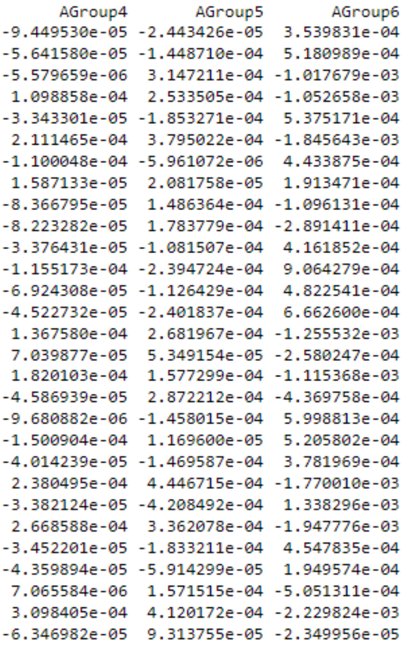
```
Call:
glm(formula = Year22 ~ District + AGroup4, family = binomial,
    data = logistic_data)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.003712   0.000000   0.000000   0.000000   0.003474

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   7.654e+02  4.496e+04   0.017    0.986
DistrictBritish Columbia      5.619e+04  7.972e+05   0.070    0.944
DistrictBulkley-Nechako       2.794e+02  6.062e+04   0.005    0.996
DistrictCapital               3.754e+03  6.904e+04   0.054    0.957
DistrictCariboo               7.377e+02  6.165e+04   0.012    0.990
DistrictCentral Coast        -6.620e+02  6.973e+04  -0.009    0.992
DistrictCentral Kootenay      6.716e+02  6.267e+04   0.011    0.991
DistrictCentral Okanagan      1.834e+03  7.969e+04   0.023    0.982
DistrictColumbia-Shuswap      5.030e+02  6.204e+04   0.008    0.994
DistrictComox                 7.696e+02  6.202e+04   0.012    0.990
DistrictCowichan Valley       1.189e+03  6.227e+04   0.019    0.985
DistrictEast Kootenay         7.317e+02  6.205e+04   0.012    0.991
DistrictFraser Valley         3.066e+03  6.664e+04   0.046    0.963
DistrictFraser-Fort George    1.623e+03  6.666e+04   0.024    0.981
DistrictKitimat-Stikine       1.773e+02  7.072e+04   0.003    0.998
DistrictKootenay-Boundary    -2.900e+01  5.931e+04   0.000    1.000
DistrictMetro Vancouver       3.042e+04  4.339e+05   0.070    0.944
DistrictMount Waddington     -4.586e+02  6.568e+04  -0.007    0.994
DistrictNanaimo               1.110e+03  7.921e+05   0.001    0.999
DistrictNorth Coast          -2.924e+02  6.609e+04  -0.004    0.996
DistrictNorth Okanagan        1.173e+03  6.424e+04   0.018    0.985
DistrictNorthern Rockies     -6.013e+02  7.017e+04  -0.009    0.993
DistrictOkanagan-Similkameen  1.104e+03  6.217e+04   0.018    0.986
DistrictPeace River           9.149e+02  6.601e+04   0.014    0.989
DistrictQathet               -2.912e+02  6.204e+04  -0.005    0.996
DistrictSquamish-Lillooet     5.764e+02  6.057e+04   0.010    0.992
DistrictStikine              -7.241e+02  7.420e+04  -0.010    0.992
DistrictStrathcona            3.284e+02  6.293e+04   0.005    0.996
DistrictSunshine Coast       -7.196e+01  6.002e+04  -0.001    0.999
DistrictThompson-Nicola       2.503e+03  7.058e+04   0.035    0.972
AGroup4                      -4.136e-02  5.851e-01  -0.071    0.944

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.9929e+02  on 899  degrees of freedom
Residual deviance: 2.5859e-05  on 869  degrees of freedom
AIC: 62

Number of Fisher Scoring iterations: 25
```

Prediction of the regional district of a household based on its demographic characteristics, such as age distribution and average number of persons per household? Here, a multinomial logistic regression model has been used to predict the "District" variable based on the predictors "AGroup1" to "AGroup6."

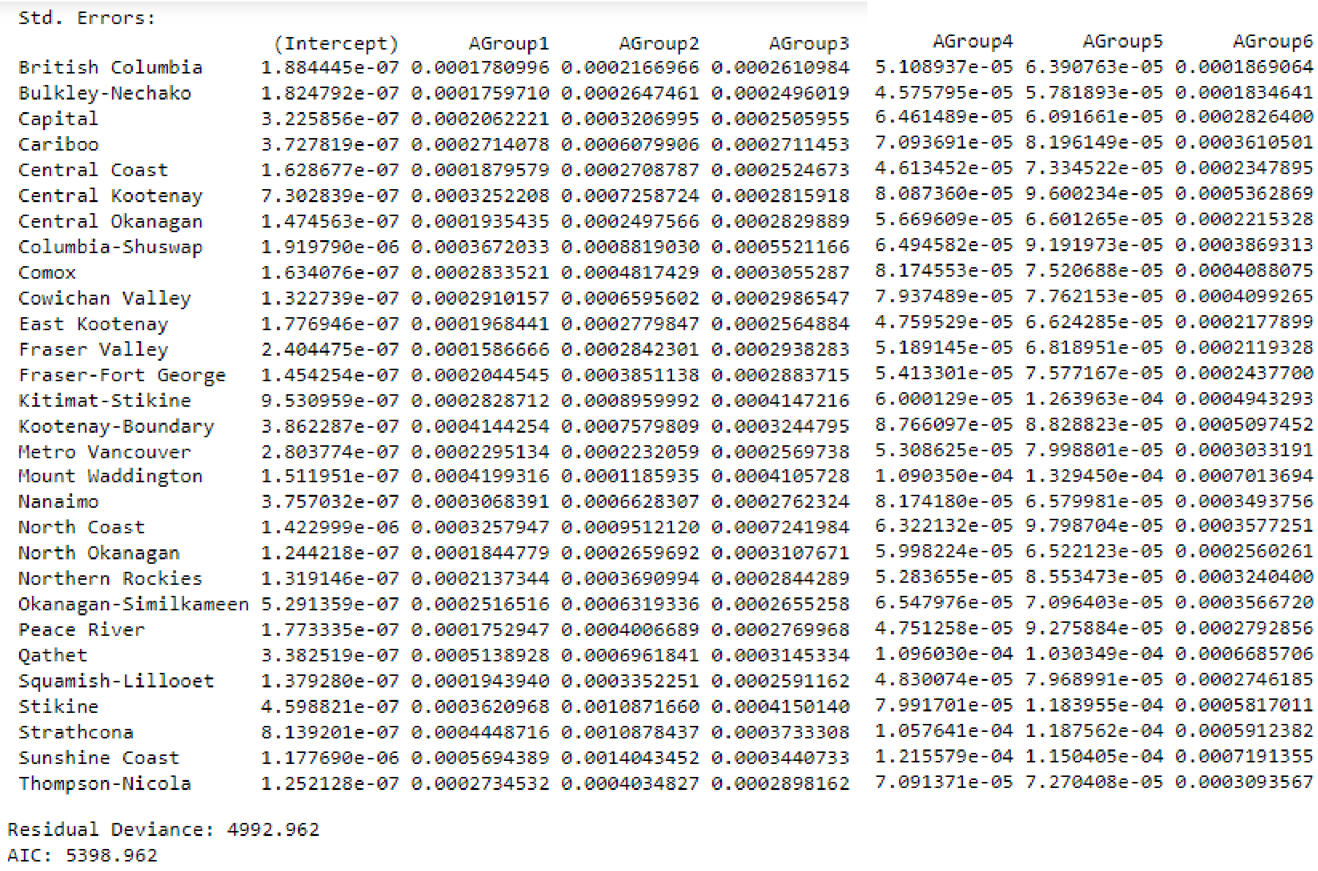
- The coefficients indicate the relationship between each predictor and the log-odds of belonging to a specific district.
- For example, the coefficient for "AGroup1" in the "British Columbia" district is 0.0002346, suggesting that a one-unit increase in "AGroup1" is associated with an increase in the log-odds of belonging to the "British Columbia" district.
- Std. Errors: These are the standard errors associated with each coefficient estimate. They provide a measure of uncertainty or variability in the coefficient estimates.
- A lower residual deviance and AIC indicated better fit of the model.



```
Call:
multinom(formula = District ~ AGroup1 + AGroup2 + AGroup3 + AGroup4 + 
    AGroup5 + AGroup6, data = data)

Coefficients:
                       (Intercept)       AGroup1       AGroup2       AGroup3
British Columbia     -6.397123e-05  2.346030e-04  3.099125e-04  1.331832e-05
Bulkley-Nechako       8.339998e-06  3.699561e-04  1.195686e-04 -2.043298e-04
Capital              -6.623480e-05 -8.346903e-04  1.084693e-03  5.692497e-04
Cariboo               9.129295e-06 -6.052848e-04 -2.822238e-04  3.154742e-04
Central Coast         1.408396e-05  3.272238e-04  9.662097e-06 -1.237367e-04
Central Kootenay      6.611989e-06 -1.095625e-03 -1.471563e-04  3.367045e-04
Central Okanagan      1.581834e-05  1.218332e-04  4.517305e-04  1.954954e-04
Columbia-Shuswap      1.347962e-05  4.264192e-04 -1.009845e-03 -4.882534e-04
Comox                 5.002681e-06 -2.621276e-04  5.316198e-04  5.354074e-04
Cowichan Valley      -1.581695e-06 -4.555382e-04  1.122759e-03  4.522031e-04
East Kootenay         1.067008e-05  1.997916e-04  9.557359e-05 -6.646601e-05
Fraser Valley        -4.527131e-05  7.166130e-04 -1.478834e-04 -1.392045e-04
Fraser-Fort George    1.157102e-05  2.852974e-04  2.182755e-04 -6.278558e-05
Kitimat-Stikine       1.809888e-05  1.675873e-04  1.013045e-03 -4.235521e-04
Kootenay-Boundary     4.080643e-06 -1.111328e-03  9.384013e-04  2.384327e-04
Metro Vancouver      -4.482765e-05 -4.490482e-04  2.813044e-04  1.201564e-04
Mount Waddington      2.410756e-05 -4.475140e-04 -7.790121e-04 -4.735739e-05
Nanaimo              -3.247685e-05 -1.934501e-05 -9.841272e-04  5.986496e-04
North Coast           1.866582e-05  4.572189e-04 -2.245891e-04 -6.526222e-04
North Okanagan        1.297198e-06  1.943529e-04  4.126795e-04  3.360354e-04
Northern Rockies      1.572418e-05  2.506137e-04  1.005463e-04 -1.935819e-05
Okanagan-Similkameen -9.968325e-06 -7.750107e-04 -1.728386e-03  4.349601e-04
Peace River          -5.884524e-06  7.346822e-04 -6.131969e-04 -2.306546e-04
Qathet                1.257481e-05 -1.592094e-03  7.801531e-04  3.281436e-04
Squamish-Lillooet     2.392528e-05  3.257448e-04  7.203916e-05 -1.407509e-04
Stikine               1.972657e-05  2.371966e-04 -6.266215e-05 -1.258117e-05
Strathcona            6.343797e-06 -1.321596e-04 -2.738242e-04  2.005655e-04
Sunshine Coast        6.233449e-06 -1.030151e-03 -1.581098e-03  4.838375e-04
Thompson-Nicola       8.151031e-06 -7.586596e-05  1.505930e-04  3.772211e-04
```



```
      AGroup4       AGroup5       AGroup6
-9.449530e-05 -2.443426e-05  3.539831e-04
-5.641580e-05 -1.448710e-04  5.180989e-04
-5.579659e-06  3.147211e-04 -1.017679e-03
 1.098858e-04  2.533505e-04 -1.052658e-03
-3.343301e-05 -1.853271e-04  5.375171e-04
 2.111465e-04  3.795022e-04 -1.845643e-03
-1.100048e-04 -5.961072e-06  4.433875e-04
 1.587133e-05  2.081758e-05  1.913471e-04
-8.366795e-05  1.486364e-04 -1.096131e-04
-8.223282e-05  1.783779e-04 -2.891411e-04
-3.376431e-05 -1.081507e-04  4.161852e-04
-1.155173e-04 -2.394724e-04  9.064279e-04
-6.924308e-05 -1.126429e-04  4.822541e-04
-4.522732e-05 -2.401837e-04  6.662600e-04
 1.367580e-04  2.681967e-04 -1.255532e-03
 7.039877e-05  5.349154e-05 -2.580247e-04
 1.820103e-04  1.577299e-04 -1.115368e-03
-4.586939e-05  2.872212e-04 -4.369758e-04
-9.680882e-06 -1.458015e-04  5.998813e-04
-1.500904e-04  1.169600e-05  5.205802e-04
-4.014239e-05 -1.469587e-04  3.781969e-04
 2.380495e-04  4.446715e-04 -1.770010e-03
-3.382124e-05 -4.208492e-04  1.338296e-03
 2.668588e-04  3.362078e-04 -1.947776e-03
-3.452201e-05 -1.833211e-04  4.547835e-04
-4.359894e-05 -5.914299e-05  1.949574e-04
 7.065584e-06  1.571515e-04 -5.051311e-04
 3.098405e-04  4.120172e-04 -2.229824e-03
-6.346982e-05  9.313755e-05 -2.349956e-05
```



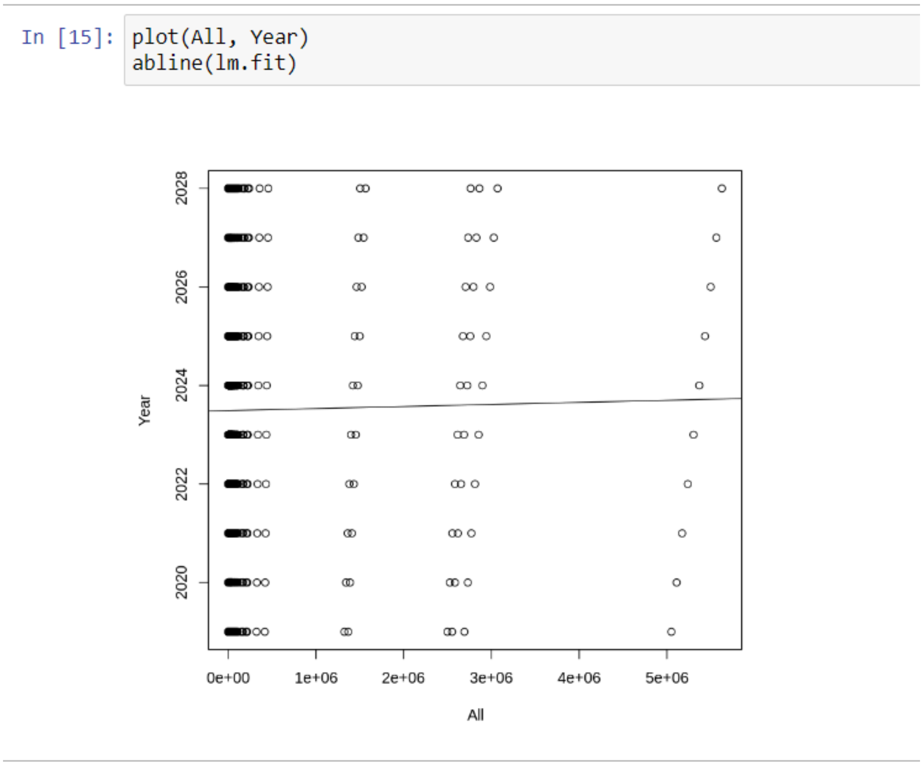
```
Std. Errors:
                      (Intercept)     AGroup1      AGroup2      AGroup3      AGroup4      AGroup5      AGroup6
British Columbia     1.884445e-07 0.0001780996 0.0002166966 0.0002610984 5.108937e-05 6.390763e-05 0.0001869064
Bulkley-Nechako      1.824792e-07 0.0001759710 0.0002647461 0.0002496019 4.575795e-05 5.781893e-05 0.0001834641
Capital              3.225856e-07 0.0002062221 0.0003206995 0.0002505955 6.461489e-05 6.091661e-05 0.0002826400
Cariboo              3.727819e-07 0.0002714078 0.0006079906 0.0002711453 7.093691e-05 8.196149e-05 0.0003610501
Central Coast        1.628677e-07 0.0001879579 0.0002708787 0.0002524673 4.613452e-05 7.334522e-05 0.0002347895
Central Kootenay     7.302839e-07 0.0003252208 0.0007258724 0.0002815918 8.087360e-05 9.600234e-05 0.0005362869
Central Okanagan     1.474563e-07 0.0001935435 0.0002497566 0.0002829889 5.669609e-05 6.601265e-05 0.0002215328
Columbia-Shuswap     1.919790e-06 0.0003672033 0.0008819030 0.0005521166 6.494582e-05 9.191973e-05 0.0003869313
Comox                1.634076e-07 0.0002833521 0.0004817429 0.0003055287 8.174553e-05 7.520688e-05 0.0004088075
Cowichan Valley      1.322739e-07 0.0002910157 0.0006595602 0.0002986547 7.937489e-05 7.762153e-05 0.0004099265
East Kootenay        1.776946e-07 0.0001968441 0.0002779847 0.0002564884 4.759529e-05 6.624285e-05 0.0002177899
Fraser Valley        2.404475e-07 0.0001586666 0.0002842301 0.0002938283 5.189145e-05 6.818951e-05 0.0002119328
Fraser-Fort George   1.454254e-07 0.0002044545 0.0003851138 0.0002883715 5.413301e-05 7.577167e-05 0.0002437700
Kitimat-Stikine      9.530959e-07 0.0002828712 0.0008959992 0.0004147216 6.000129e-05 1.263963e-04 0.0004943293
Kootenay-Boundary    3.862287e-07 0.0004144254 0.0007579809 0.0003244795 8.766097e-05 8.828823e-05 0.0005097452
Metro Vancouver      2.803774e-07 0.0002295134 0.0002232059 0.0002569738 5.308625e-05 7.998801e-05 0.0003033191
Mount Waddington     1.511951e-07 0.0004199316 0.0001185935 0.0004105728 1.090350e-04 1.329450e-04 0.0007013694
Nanaimo              3.757032e-07 0.0003068391 0.0006628307 0.0002762324 8.174180e-05 6.579981e-05 0.0003493756
North Coast          1.422999e-06 0.0003257947 0.0009512120 0.0007241984 6.322132e-05 9.798704e-05 0.0003577251
North Okanagan       1.244218e-07 0.0001844779 0.0002659692 0.0003107671 5.998224e-05 6.522123e-05 0.0002560261
Northern Rockies     1.319146e-07 0.0002137344 0.0003690994 0.0002844289 5.283655e-05 8.553473e-05 0.0003240400
Okanagan-Similkameen 5.291359e-07 0.0002516516 0.0006319336 0.0002655258 6.547976e-05 7.096403e-05 0.0003566720
Peace River          1.773335e-07 0.0001752947 0.0004006689 0.0002769968 4.751258e-05 9.275884e-05 0.0002792856
Qathet               3.382519e-07 0.0005138928 0.0006961841 0.0003145334 1.096030e-04 1.030349e-04 0.0006685706
Squamish-Lillooet    1.379280e-07 0.0001943940 0.0003352251 0.0002591162 4.830074e-05 7.968991e-05 0.0002746185
Stikine              4.598821e-07 0.0003620968 0.0010871660 0.0004150140 7.991701e-05 1.183955e-04 0.0005817011
Strathcona           8.139201e-07 0.0004448716 0.0010878437 0.0003733308 1.057641e-04 1.187562e-04 0.0005912382
Sunshine Coast       1.177690e-06 0.0005694389 0.0014043452 0.0003440733 1.215579e-04 1.150405e-04 0.0007191355
Thompson-Nicola      1.252128e-07 0.0002734532 0.0004034827 0.0002898162 7.091371e-05 7.270408e-05 0.0003093567

Residual Deviance: 4992.962
AIC: 5398.962
```

C. Regression:

As mentioned previously in the project proposal, we used linear regression as our main approach in predicting data for our chosen dataset. The basic idea of linear regression is that the relationships between the independent and dependent variables are linear. This implies that the value of the dependent variable varies consistently and proportionally as the values of the independent variables do.

In the screenshot below of the graph we plotted, we have used Year on the y-axis and All ages (all population number) on the x -axis. The tilted line running along the graph represents the relation between the variables Year and All population and predict how the population will be growing over the coming years.

To make the prediction even more accurate we can modify the data by minimizing the differences between the predicted values and the actual values of the data. After modifying the data we can use that to predict the total population values that are not present in the actual dataset. This allows us to make predictions based on the linear relationship observed in the data.



```
In [15]: plot(All, Year)
         abline(lm.fit)
```

V. Ethical Concerns and Mitigations in the Project

a. Respect for Human Dignity:
The data was sourced from government websites and therefore respect for human dignity was upheld in the project. We followed the ethical guidelines and regulations to protect the privacy and confidentiality of individuals represented in the dataset.

b. Well-being of all Parties in question:
To prioritize the well-being of all residents of BC, the project focused on utilizing publicly available data for analysis and modelling purposes only. The research aimed to be used solely for college project purposes, while minimizing any potential risks or harm to individuals.

c. Minimum biases
The project ensured fairness by treating all regional districts equally in the analysis and modelling process. No district was favoured or disadvantaged, and efforts were made to maintain impartiality in the selection and treatment of districts involved in the study.

d. Integrity in Research:
We put efforts to maintain research integrity by following established scientific practices, using appropriate data analysis methods. Transparency was maintained in reporting the results, providing an accurate representation of the insights derived from the data.

e. Respect for Indigenous Knowledge and Culture:
As the project solely utilized data from government websites, there was no direct involvement of Indigenous knowledge or culture. Therefore, specific considerations related to Indigenous communities were maintained as out of scope of this project.

In the nutshell, the project maintained ethical standards by respecting privacy and confidentiality, prioritizing the well-being of participants, ensuring fairness, upholding research integrity, and operating within the bounds of publicly accessible data from government websites.

VI. Conclusions

Clustering:

While we attempted to determine the optimal number of clusters by evaluating the within-cluster sum of squares (WCSS) and inspecting the plot, this approach is subjective and may not always provide a clear indication of the ideal number of clusters. The quality of clustering results heavily depends on the choice of features used for clustering. It's important to choose meaningful features to capture the true patterns and characteristics of the data. We chose the columns and features based on our dataset, which did not present the absolute result we hoped for.

Classification:

The findings suggest that certain age groups (AGroup3 and AGroup5) have a significant influence on the gender distribution, while other age groups (AGroup1, AGroup2, AGroup4, and AGroup6) do not.
The associations between age groups and gender distribution can provide insights into the demographic characteristics of regional districts.
The coefficients for age groups and districts can be used to predict the probability of gender distribution and household types in different regions.

However, the limitations include:

- The absence of statistical significance for some coefficients suggests caution in drawing strong conclusions about their associations.
- The model's performance and accuracy in predicting gender distribution and household types should be further evaluated using appropriate validation techniques and additional data if available.

Regression:

After all the results we got while estimating the population using regression, we say that linear regression shows us how a dependent variable can help us in prediction of various things using the data provided. For example, we calculated the correlation coefficient to express the strength and direction of the association between the median age and average household size in order to determine whether there is a correlation between the two variables. Using the above examples, we can predict many different things using the data present in the dataset.

There are some limitations to making predictions such as when we used the same technique as above to make a graph between the household projections and regional district in order to estimate the future population size of a regional district based on its current household projections, but the predictions were not as accurate and it seemed difficult to predict if we could compare these two in order to figure out a different outcome. So, this may be the case with different variables and their predictions.

References

A. *Interpret all statistics and graphs for Cluster K-means*. Minitab. (n.d.). https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/#:~:text=the%20cluster%20centroid.-,Interpretation,a%20large%20sum%20of%20squares

B. *Multinomial and Ordinal Logistic Regression In R*. (2016, February 1). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/

C. *K-means cluster analysis*. K-means Cluster Analysis UC Business Analytics R Programming Guide. (n.d.). https://uc-r.github.io/kmeans_clustering

D. *K-means clustering: Explain it to me like I'm 10*. Medium. Rao, S. (2023, February 27). https://towardsdatascience.com/k-means-clustering-explain-it-to-me-like-im-10- e0badf10734a

E. *Scale: Scaling and centering of matrix-like objects*. RDocumentation. (n.d.). https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale

F. Singh, D. (2019, November 6). *Deepika Singh*. Pluralsight. https://www.pluralsight.com/guides/normalizing-data-r