Prediction of Indiginous and Non-Indiginous Status of Communities in Canada from Socio-Economic Well-Being Scores

Jagraj Singh Gill

Abstract

Recent studies have indicated strong correlations between socioeconomic status of Canadian communities and the health of the individuals that comprise them. The objective of this study is to explore the classification between the Indigenous and Non-Indigenous Status of Communities in Canada from the Community Well-Being (CWB) scores, to highlight any similarities and contrasts in the data between communities, and to suggest if there is enough predictive power in the models that may - in conjunction with health statistics - assist in labeling potential at-risk communities as it pertains to mortality and morbidity rates. Logistic regression, PCA, and KNN algorithms will be employed to explore classification capabilities, efficacy of dimensionality reduction in modeling, and for remedying class imbalances. Performance summaries of the logistic models will then be used to scrutinize and justify viability of use. The non-PCA logit model was determined to be viable candidate for making the aforementioned predictions due to its high accuracy and kappa values of .94 and .87 respectively. Exploratory analysis provided a concrete and vivid understanding of Canadian Indigenous and Non-Indigenous communities in terms of their socioecomic backbones, and validated the current obstacles that First Nations communities are facing. This predictor may be employed with health studies to look at longitudinal trends that occur in First Nations communities.

Introduction

The correlation between socioeconomic status and mortality rates among Aboriginal groups has been highlighted as a concern by the Canadian Journal of Public Health and the Canadian Medical Association Journal. It's proposed that in studies comparing the health status of Aboriginal and Non-Aboriginal populations, there is a higher rate of mortality and morbidity in Aboriginal populations (Oliver et al., 2017). The high mortality rates may be partially explained from individual socioeconomic characteristics in First Nations' populations; however, there are stark differences that arise when examining the communities' impact on health.

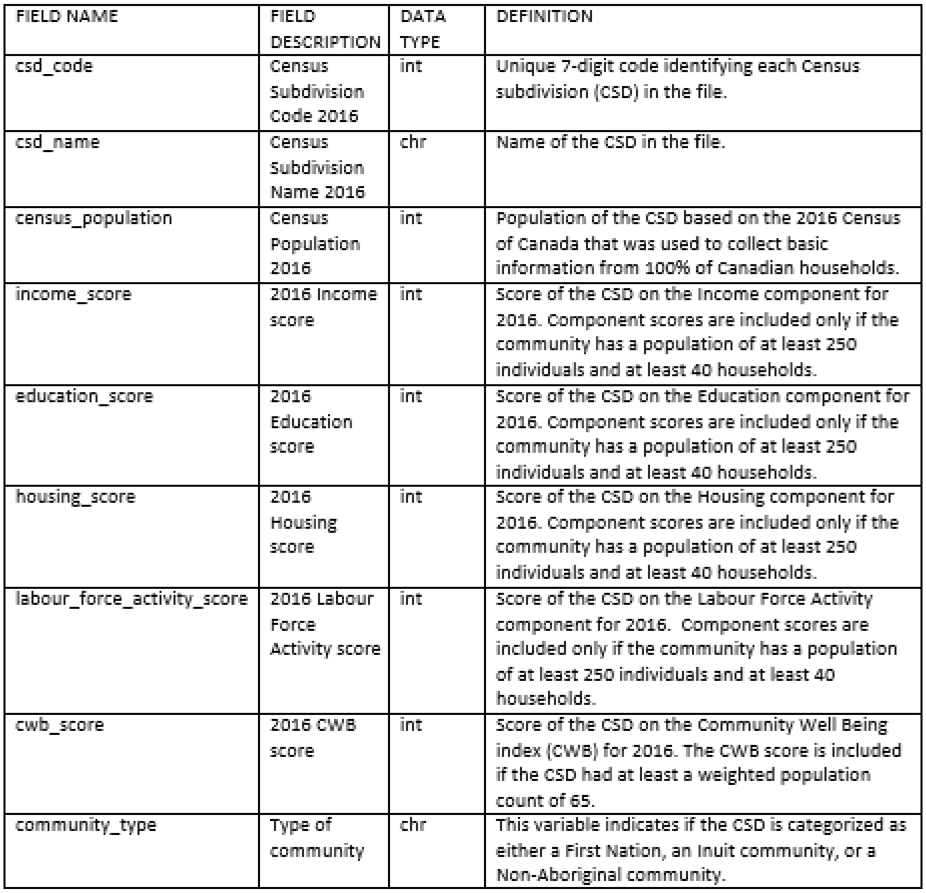
The CWB index is a measure of socio-economic well-being for communities across Canada. There are four underlying factors that contribute to the index: education, per capita community income, housing, and labour force participation; scores that range from 0 (lowest) to 100 (highest). It was developed complimentary to The United Nations' modeled Registered Indian Human Development Index (HDI), which was an index derived from life expectancy, education and wealth of approximately 170 countries, which compared Registered First Nations at the national level. The goal of the CWB was to establish an index to track the well-being of First Nations and Inuit against non-Native counterparts at the community level over time.

It also attempts to capture the non-monetary aspects of well-being. This poses a challenge as equally important factors such as mental and physical fortitude are omitted from reporting. Another challenge for this model was having similar measures for varying community values, as it's suggested that Aboriginal cultures put less emphasis on the collection of material wealth, and to gauge across communities, hence cultures, with modern economic indicators could imply that the standard of well-being lies with communities who have assimilated or who put more importance to these measures (K. et al., 2007). This does not discount the use of these community factors in the ability to identify the threshold to those who may face health-related hardships; this is what this study aims to explore.

Data

The CWB 2016 dataset was sourced from the Open Data portal provided by The Government of Canada's Open Government initiative, which seeks to provide enhanced transparency, accountability, and motivation in citizens' participation in policymaking through open information and dialogue. The publisher of the dataset is the Crown-Indigenous Relations and Northern Affairs Canada; Information Management Branch, Business Decision Support, Geomatics Services which authorizes use under the Open Government Licence - Canada (Community Well-Being Index 2015).

The data was originally sourced from the 2016 Census of Population (Canada). It comprises of a sample size of 5,162 observations with 9 variables. The data dictionary is referenced as follows, with the only pre-processing being renaming the columns to exclude French, standardizing the database object naming convention, and adding in a data type column:



FIELD NAME	FIELD DESCRIPTION	DATA TYPE	DEFINITION
csd_code	Census Subdivision Code 2016	int	Unique 7-digit code identifying each Census subdivision (CSD) in the file.
csd_name	Census Subdivision Name 2016	chr	Name of the CSD in the file.
census_population	Census Population 2016	int	Population of the CSD based on the 2016 Census of Canada that was used to collect basic information from 100% of Canadian households.
income_score	2016 Income score	int	Score of the CSD on the Income component for 2016. Component scores are included only if the community has a population of at least 250 individuals and at least 40 households.
education_score	2016 Education score	int	Score of the CSD on the Education component for 2016. Component scores are included only if the community has a population of at least 250 individuals and at least 40 households.
housing_score	2016 Housing score	int	Score of the CSD on the Housing component for 2016. Component scores are included only if the community has a population of at least 250 individuals and at least 40 households.
labour_force_activity_score	2016 Labour Force Activity score	int	Score of the CSD on the Labour Force Activity component for 2016. Component scores are included only if the community has a population of at least 250 individuals and at least 40 households.
cwb_score	2016 CWB score	int	Score of the CSD on the Community Well Being index (CWB) for 2016. The CWB score is included if the CSD had at least a weighted population count of 65.
community_type	Type of community	chr	This variable indicates if the CSD is categorized as either a First Nation, an Inuit community, or a Non-Aboriginal community.

Table 1: Data dictionary of the CWB 2016 dataset (Community Well-Being Index 2015). The education score comprises of functional literacy and a "high school plus" score weighted at 2/3, and 1/3 respectively. Functional literacy is the percentage of individuals 15 and over who have completed a grade 9 education, and the high school plus score accounts for individuals 20 and over who have completed secondary school

studies. The labour force participation score comprises of the labour force participation (i.e. the percentage of population over 20 who are active in the workforce) and the employment rate which is the percent ratio of the employed labour force and the total labour force for individuals over 15. The housing score takes into consideration both the quantity (percentage of population with dwellings with occupancies no greater than one occupant per room) and quality (percentage of population living in dwellings that don't require major repairs) of living. Finally, the income score is derived from the income per capita, and by taking the logarithmic, the diminishing marginal utility of income is taken into account. (K. et al., 2007)

According to Indigenous Services Canada, communities will not appear in the 2016 CWB index if there was a population of less than 65, if there were issues of 'data quality', or they were not fully enumerated in the census. Communities are defined as municipalities, or census subdivisions (CSD), as defined by provincial or territorial legislation, or areas that are treated as municipal equivalents, where the two primary CSD types affiliated with First Nations or Indigenous bands are Indian reserves and settlements. 96% of the aforementioned primary CSDs are represented by Indian reserves. Inuit communities do not have the same legal status as First Nations groups hence they are categorized as a separate entity in the study; all other CSDs are listed as non-Indigenous (Government of Canada; Indigenous Services Canada, 2019).

Logistic regression, PCA, and KNN algorithms will be employed on the dataset to explore classification capabilities, efficacy of dimensionality reduction, and to remedy any class imbalances.

Methods

Data Ingestion

The CWB_2016_Data.csv dataset was imported via the read() function. The dimensions, classes, and first few rows were examined using the dim() and str() functions respectively.

EDA and Data Pre-Processing

The class occurrences of the community_type column were determined to be 3249 and 386 for the Non-Indigenous and Indigenous communities respectively, hence this is considered an imbalanced dataset.There were no outliers present as all scoring columns fell within the range of 0 to 100, and the only other numerical feature values used for analysis - the population - was unremarkable. NA-values were present; any rows containing an NA-value were dropped with the na.omit() function, and as a result, no imputation was required. The Inuit Community class was removed from the dataset as they were not a community for consideration in this study. The CSD unique identifier and name were also removed as they are neither a feature or target variable.

Taking a look at the density distributions, Pearson correlations, and box-whisker plots of the feature variables using the GGally library, we have:

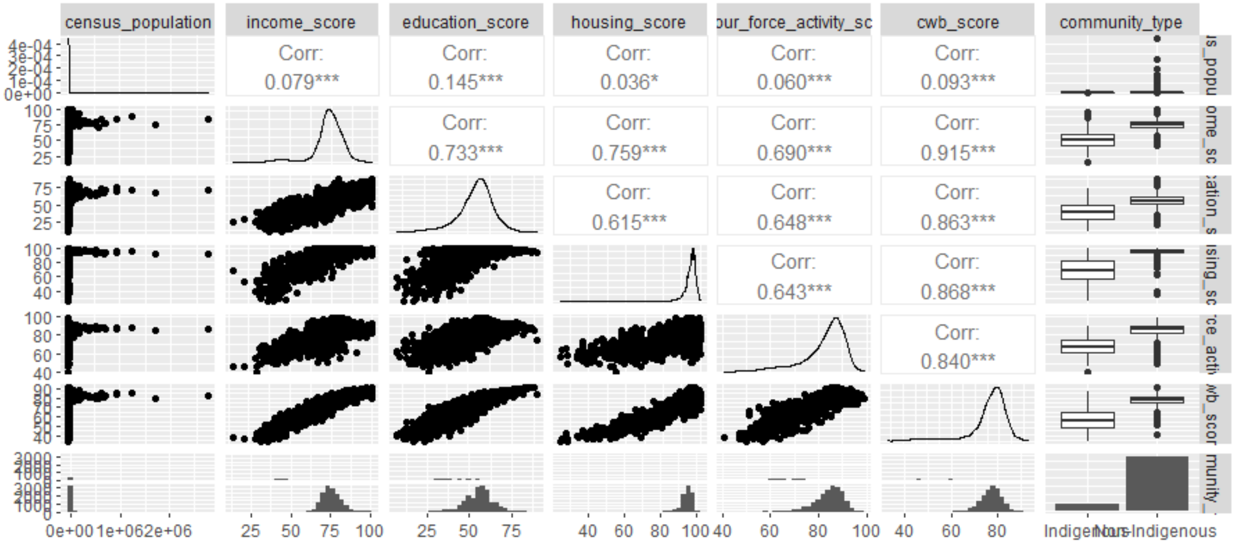


Figure 1: The upper triangle correlation matrix highlights a few variables that experience a moderate degree of positive correlation, namely the education and income, and housing and income scores. The CWB score has a high positive correlation with the other feature variables ; this is logical as the well-being index is comprised (i.e. is a function of) the aforementioned scores. This also suggests that one should address the issue of multicollinearity before model application. There are no negative correlations present in among the feature variables. The density distributions for the labour force activity and cwb score appear left-skewed, while the education, income and housing appear to be Gaussian . This should not be an issue for implementing logistic regression as the model makes no assumption of multivariate normality. Visual study of the histograms breaking down the variable distributions across the classes suggests that the housing, labour force, education and income scores may all be important features to distinguish the Indigenous and Non-Indigenous communities. The box-whisker plot shows that each variable contains outliers occurring beyond the minimum and maximum ranges; however none warrant any attention as they fall within the range of acceptable scores and population.

To remedy the class imbalance, the SMOTE() (Synthetic Minority Oversampling Technique) function was applied from the bimba library to the train set. As described by He and Ma in Imbalanced learning, SMOTE initially selects a minority class instance X at random and finds its K nearest minority class neighbors (KNN). The instance is then created by choosing one of the K nearest neighbors Y at random and connecting a line segment from X and Y in the feature space. The convex combination of both instances X and Y create the synthetic observations (He & Ma, 2013). The K-value was set at 8.

An approach to deal with the correlation of predictor variables would be to omit them in the pre-processing stage and then run the logit model with the reduced number of features, or the utilization of Principle Component Analysis (PCA) using the prcomp() function. The latter method was employed and the logit performance will be compared to the original list of features. The cumulative proportion of explained variance plot is generated to find the optimal number of principle components. The principle components here are the eigenvectors of the covariance matrix of the cwb predictor matrix, that contain the largest corresponding eigenvalues. PCA will also allow for visual inspection of the loading vectors and their contribution to the primary principle components. Although multicollinearity may not necessarily introduce bias in the model, there may be problems in model convergence which PCA may remedy.

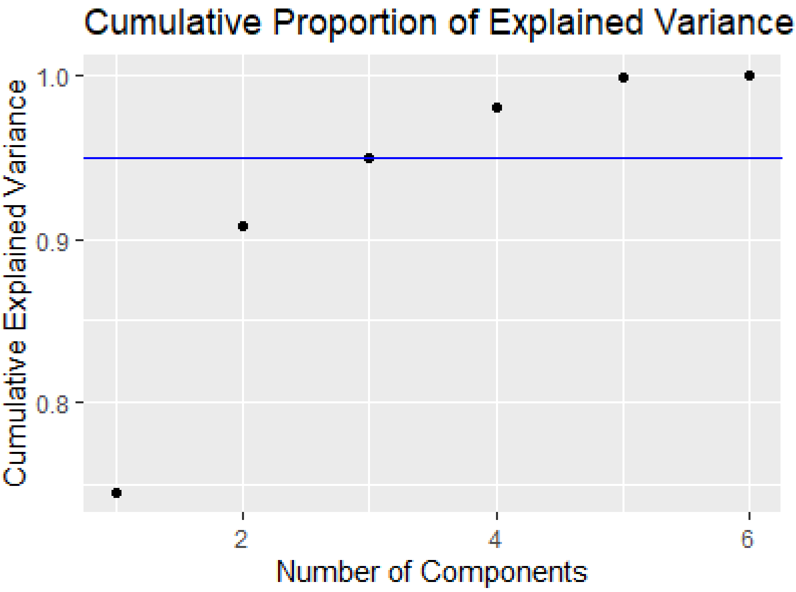


Figure 2: It's observed above that the first three principle component account for approximately 95% of the explained variance and the first five principle components account for nearly 100%. All the principle components will be used in the pca-transformed logit model.

Density Distributions of the first Five Principle Components and Loadings Plot

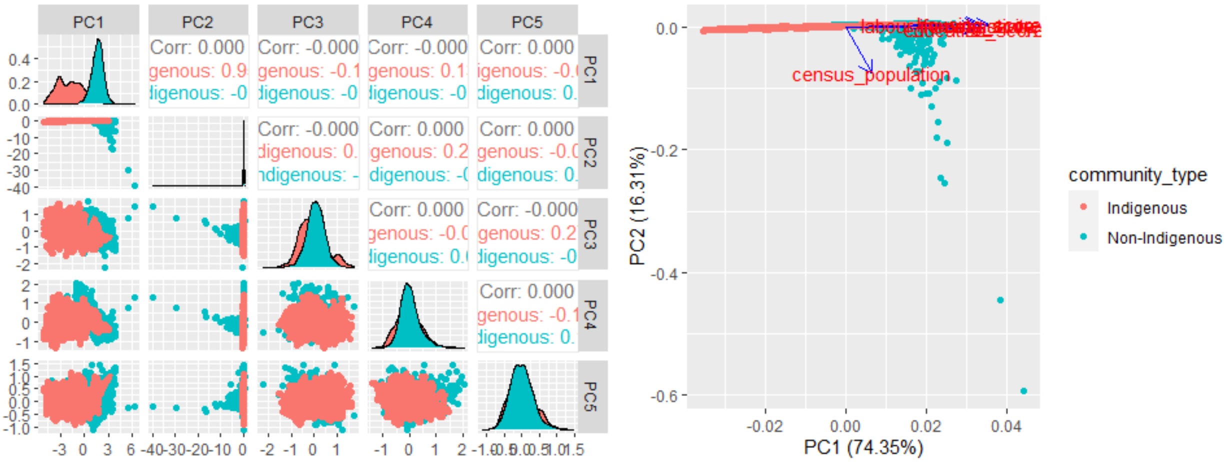


Figure 3: Showcasing the ability of the first five principle components to separate the Indigenous and Non-Indigenous classes from their explained variance (left), and observing the feature importance that arise from the first two principle component vectors (right). The first principle component accounts for 74% of the explained variance and the second principle component accounts for 16%. The succeeding three principle components do not showcase much capability in class separation. Income, housing, education, labour force, and overall cwb scores all have positive loadings and similar influence on PC1, where the census population has a negative loading on PC2.

From Table 6 it's observed that the score rotation values for PC1 are between .43 to .47 hence a moderately strong correlation to the first principle component and for the census population we have .08, not a significant correlation. PC2 shows rotations for the scores ranging from -.02 to .07, an insignificant correlation as compared to PC1; however the census population, with a value of -.99, shows a strong correlation.

Scaling the data was not required as all score occurrences occur from a range of 0 to 100; however scaling and centering were performed in conjunction with the prcomp() transformation function.

Model Implementation

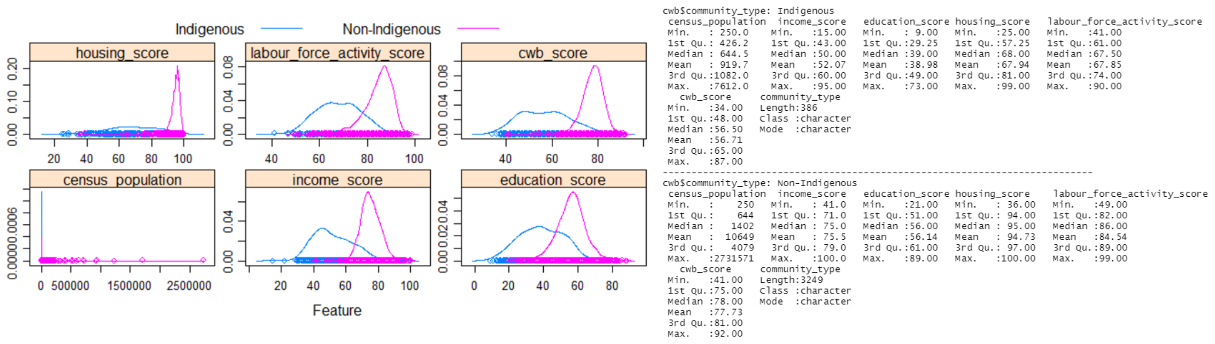
After setting the seed, the CWB data was split into the train-test set (70/30) using the createDataPartition() function, which is defaulted to perform a stratified random split. As previously noted, SMOTE was applied to the train set. The PCA-transformed and CWB data model was created using the train() function with a k-fold cross validation of 5 to reduce the likelihood of overfitting.

RStudio desktop version 1.4.1717 was used for all analysis and modeling work (RStudio Team 2021) in conjunction with R version 4.1.0.

Results and Discuission

A notable observation is that the top three populated metropolises, Toronto (2,731,571), Montreal (1,704,694), and Calgary (1,239,220), which are considered to be major economic cores in Canada, are indexed in the study. Pelican Narrows and Wood Buffalo regions are the only Non-Indigenous communities to score in the bottom 10 of all scoring categories, namely in labour force and housing respectively. Non-indigenous communities have better scores in all scoring metrics overall, with Indigenous communities having wider density distributions.

Density Distribution and Summary Statistics of Predictor Variables by Class



```
cwb$community_type: Indigenous
 census_population  income_score   education_score housing_score   labour_force_activity_score
 Min.   : 250.0    Min.   :15.00   Min.   : 9.00   Min.   :25.00   Min.   :41.00
 1st Qu.: 426.2    1st Qu.:43.00   1st Qu.:29.25   1st Qu.:57.25   1st Qu.:61.00
 Median : 644.5    Median :50.00   Median :39.00   Median :68.00   Median :67.50
 Mean   : 919.7    Mean   :52.07   Mean   :38.98   Mean   :67.94   Mean   :67.85
 3rd Qu.:1082.0    3rd Qu.:60.00   3rd Qu.:49.00   3rd Qu.:81.00   3rd Qu.:74.00
 Max.   :7612.0    Max.   :95.00   Max.   :73.00   Max.   :99.00   Max.   :90.00
   cwb_score      community_type
 Min.   :34.00   Length:386
 1st Qu.:48.00   Class :character
 Median :56.50   Mode  :character
 Mean   :56.71
 3rd Qu.:65.00
 Max.   :87.00
------------------------------------------------------------------------------
cwb$community_type: Non-Indigenous
 census_population  income_score   education_score housing_score    labour_force_activity_score
 Min.   :     250   Min.   : 41.0   Min.   :21.00   Min.   : 36.00   Min.   :49.00
 1st Qu.:     644   1st Qu.: 71.0   1st Qu.:51.00   1st Qu.: 94.00   1st Qu.:82.00
 Median :    1402   Median : 75.0   Median :56.00   Median : 95.00   Median :86.00
 Mean   :   10649   Mean   : 75.5   Mean   :56.14   Mean   : 94.73   Mean   :84.54
 3rd Qu.:    4079   3rd Qu.: 79.0   3rd Qu.:61.00   3rd Qu.: 97.00   3rd Qu.:89.00
 Max.   :2731571   Max.   :100.0   Max.   :89.00   Max.   :100.00   Max.   :99.00
   cwb_score      community_type
 Min.   :41.00   Length:3249
 1st Qu.:75.00   Class :character
 Median :78.00   Mode  :character
 Mean   :77.73
 3rd Qu.:81.00
 Max.   :92.00
```

Figure 4: It's noted that these are not heavily skewed distributions so it is appropriate to look at the predictor means over the median. There is approximately a tenfold difference in mean population between Indigenous and Non-Indigenous communities in the study. The mean income scores vary by 23 points, mean education score by 17 points, mean housing score by 27 points, mean labour force activity by 17 points, and the overall mean CWB score by 21 points. Non-indigenous communities are unequivocally far better off socioeconomically than Indigenous communities.

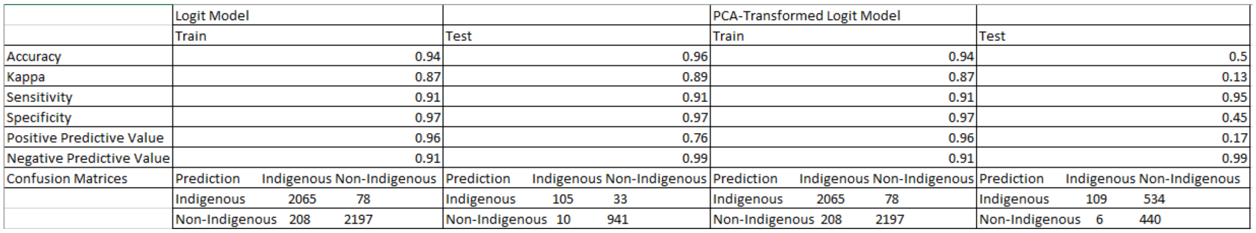
Interpreting the Logit and PCA-Logit Model Summary Reports

Coefficients: The logit model report (please refer to Table 3) suggests that the housing score, education score and the census population are statistically significant to the community type class (alpha=0.05 and statistically significant is considered p<0.05). The coefficient estimate of housing score is .23 which means an increase in the housing score is associated with an increased probability of the community class being Non-Indigenous. The coefficient estimate of the education score of -.13 suggests that an increase in education score is associated with a decreased probability of the community class being Non-Indigenous.The coefficient estimate of the census population is .0004 which means an increase in the population is associated with an increased probability of the community class being Non-Indigenous. Looking at the odds ratio, a unit increase in the housing, or census population values will increase the likelihood of being a Non-Indigenous community by 1.23 and 1.004 times respectively (Euler's number raised to the respective variable coefficients). The odds ratio for the education score is .88. These individual predictions are based on holding other factors constant. The summary report for the PCA model suggests the first five principle components are statistically significant.

Deviance: The null deviance (6305) as compared to the residual deviance (1565) suggests that the response is better predicted with the feature variables included as compared to the intercept alone.

AIC (Akaike Information Criterion): The same AIC scores of 1579 for both the logit, and PCA-transformed logit model suggest that both may be excellent or equally poor candidate model choices for the dataset, as there are no more comparable models to base their performance.

Interpretation of the Logit and PCA-Logit Model Performance and Confusion Matrices



	Logit Model		PCA-Transformed Logit Model	
	Train	Test	Train	Test
Accuracy	0.94	0.96	0.94	0.5
Kappa	0.87	0.89	0.87	0.13
Sensitivity	0.91	0.91	0.91	0.95
Specificity	0.97	0.97	0.97	0.45
Positive Predictive Value	0.96	0.76	0.96	0.17
Negative Predictive Value	0.91	0.99	0.91	0.99
Confusion Matrices	Prediction Indigenous Non-Indigenous	Prediction Indigenous Non-Indigenous	Prediction Indigenous Non-Indigenous	Prediction Indigenous Non-Indigenous
	Indigenous 2065 78	Indigenous 105 33	Indigenous 2065 78	Indigenous 109 534
	Non-Indigenous 208 2197	Non-Indigenous 10 941	Non-Indigenous 208 2197	Non-Indigenous 6 440

Table 2: The accuracy and kappa values suggest that the preferred model is the non-PCA transformed logit model. The test set performance was equal if not better than the train set in all areas except the positive predictive value, which are instances where predictions of Indigenous communities are actually Indigenous. The performance metrics do not suggest that this model was overfitted. One possible solution to improving the performance would be to undersample the majority class instead of creating synthetic observations from SMOTE, or running an iterative model with multiple KNN k-values and observing the changes not only in model performance, but the cluster patterns. The PCA-transformed logit model performed identically for the train set. This was expected as all of the principle components were used encompassing ~100% of the variance from the original dataset plus the synthetic observations. This model outperformed the standard logit model in sensitivity; however the kappa, specificity, and positive predictive values were surprisingly disappointing, at .13, .45, and .17 respectively. This occurred with a predicted probability cutoff of 0.5. This means that the PCA model is poor at predicting a Non-Indigenous outcome for an observation that is Non-Indigenous, and also instances where predicting an Indigenous community is actually Indigenous. The low kappa score is indicative of a model that has poor observed to expected accuracy. These measures suggest that the model is overfitting despite k-fold cross validation.

Refer to Table 4 and Table 5 for the full confusion matrix reports.

Significance of Results and Key Takeaways

Utilizing PCA to eliminate multicollinearity proved effective as nearly 100% of the explained variance from the original dataset was able to be showcased with the first five principle components. Using KNN SMOTE to oversample the minority class proved useful to the traditional logit model; however the PCA-transformed model test confusion matrix was indicative of poor performance with this technique, and demonstrates potential overfitting.

The ability to distinguish Non-Indigenous and Indigenous communities opens the avenue to compare the health challenges posed from their socioeconomic background. This may be accomplished by a longitudinal classification study, and comparing the class labels as analogous to communities who are at-risk and not at risk for higher morbidity and mortality rates when cross referenced with historical health data.

Conclusion

The objective of this study was to explore the classification between the Indigenous and Non-Indigenous Status of Communities in Canada from the Community Well-Being (CWB) scores, to highlight any similarities and contrasts in the data between communities, and to suggest if there is enough predictive power in the models that may assist in labeling potential at-risk communities as it pertains to mortality and morbidity rates. The non-PCA logit model is a viable candidate for making the aforementioned predictions due to its high accuracy and kappa values of .94 and .87 respectively. The dataset provided a rich understanding of Canadian Indigenous and Non-Indigenous communities in terms of their socioecomic backbones, and validated the current obstacles for First Nations as described in Canada's Continuing Challenge.

One of the data limitations of this study was the data source publishers, the Crown-Indigenous Relations of Canada, who relied solely on the census aggregates to update their scoring metrics and population figures. Although census completion is a mandatory task in Canada, approximately 25% of households receive a long-form, or fully detailed version to complete (Government of Canada, 2020). The sampling, along with no governing body to audit the responses may have led to skewed scores. Another major limitation to the study was that communities with low populations (<65) were not accounted for on the census, and as the analysis highlights, it's the smaller communities that are scoring lower, and hence may be at greater risk. One recommendation to improve the study for PCA model would be to change the predicted probability cutoff to observe sensitivity and specificity trade offs. Another recommendation would be to code an iterative model using different k-values for the KNN algorithm to compare the various synthetic observations' performance, or undersample the majority class instead.

References

Crown-Indigenous Relations and Northern Affairs Canada. (2015, April 2). Community Well-Being Index. Open Government Portal. https://open.canada.ca/data/en/dataset/56578f58-a775-44ea-9cc5-9bf7c78410e6.

Oliver, L. N., Peters, P. A., & Penney, C. (2017, July 12). Health reports the influence of Community well-being on mortality among registered First Nations People; Statistics Canada. https://www150.statcan.gc.ca/n1/pub/82-003-x/2016007/article/14646-eng.htm.

Government of Canada; Indigenous Services Canada. (2019, May 24). About the Community Well-Being Index. Government of Canada; Indigenous Services Canada. https://www.sac-isc.gc.ca/eng/1421245446858/1557321415997.

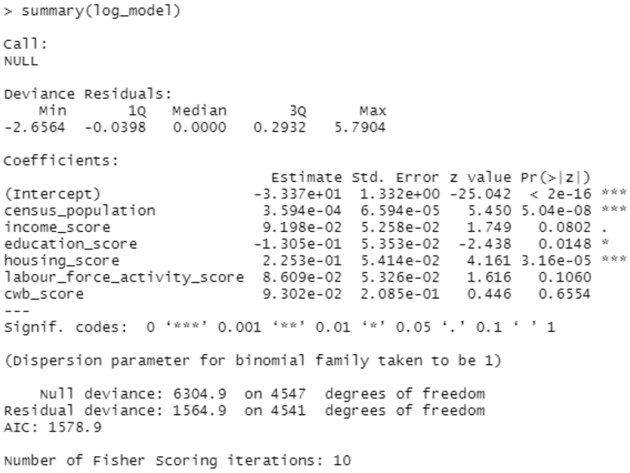
Government of Canada, S. C. (2020, July 17). Census of population. Surveys and statistical programs. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3901.

He, H., & Ma, Y. (2013). Imbalanced learning: Foundations, algorithms, and applications. Wiley-IEEE Press.

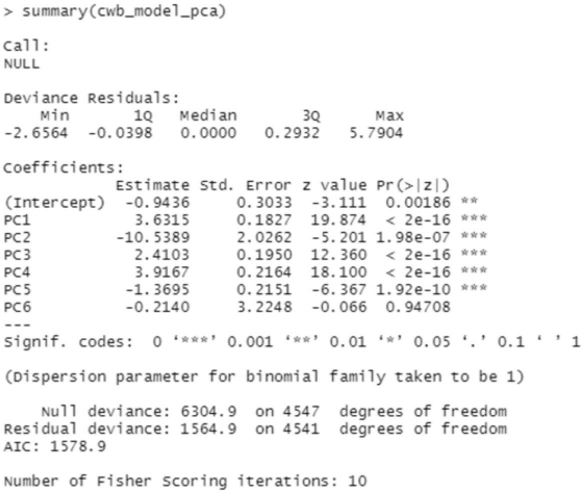
K., B. D. J., Spence, N., & White, J. (2007). Aboriginal well-being: Canada's continuing challenge. Thompson Educational Pub.

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Appendix

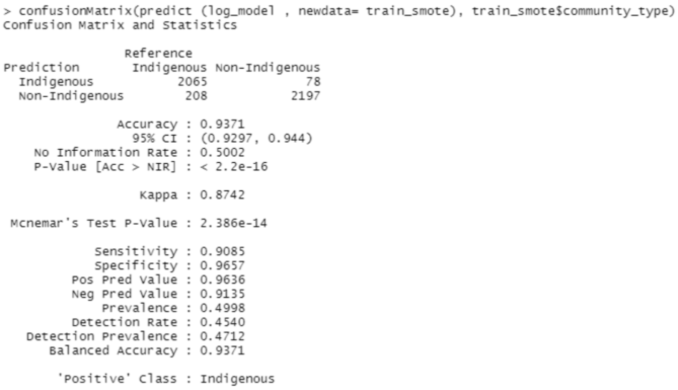


```
> summary(log_model)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6564  -0.0398   0.0000   0.2932   5.7904

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.337e+01  1.332e+00 -25.042  < 2e-16 ***
census_population           3.594e-04  6.594e-05   5.450 5.04e-08 ***
income_score                9.198e-02  5.258e-02   1.749   0.0802 .
education_score            -1.305e-01  5.353e-02  -2.438   0.0148 *
housing_score               2.253e-01  5.414e-02   4.161 3.16e-05 ***
labour_force_activity_score 8.609e-02  5.326e-02   1.616   0.1060
cwb_score                   9.302e-02  2.085e-01   0.446   0.6554
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6304.9  on 4547  degrees of freedom
Residual deviance: 1564.9  on 4541  degrees of freedom
AIC: 1578.9

Number of Fisher Scoring iterations: 10
```

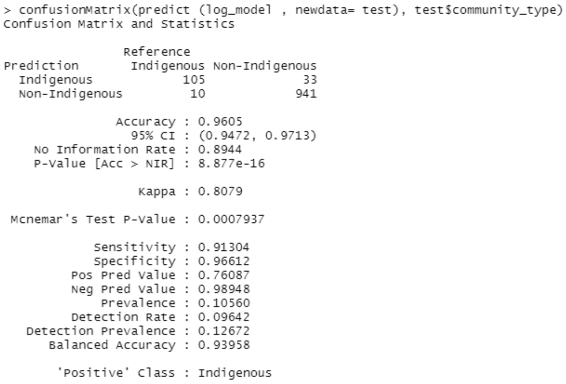


```
> summary(cwb_model_pca)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6564  -0.0398   0.0000   0.2932   5.7904

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9436     0.3033  -3.111  0.00186 **
PC1           3.6315     0.1827  19.874  < 2e-16 ***
PC2         -10.5389     2.0262  -5.201 1.98e-07 ***
PC3           2.4103     0.1950  12.360  < 2e-16 ***
PC4           3.9167     0.2164  18.100  < 2e-16 ***
PC5          -1.3695     0.2151  -6.367 1.92e-10 ***
PC6          -0.2140     3.2248  -0.066  0.94708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6304.9  on 4547  degrees of freedom
Residual deviance: 1564.9  on 4541  degrees of freedom
AIC: 1578.9

Number of Fisher Scoring iterations: 10
```

Table 3: Summary reports for the cwb logit (left), and PCA-logit models (right)

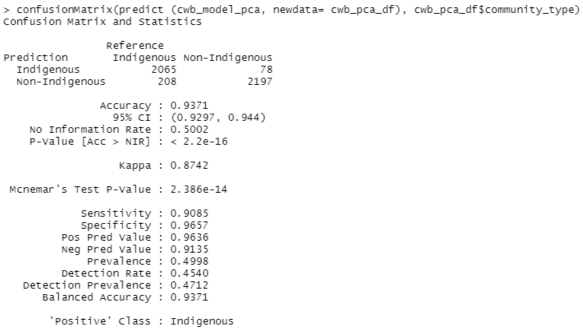


```
> confusionMatrix(predict (log_model , newdata= train_smote), train_smote$community_type)
Confusion Matrix and Statistics

                Reference
Prediction       Indigenous Non-Indigenous
  Indigenous           2065             78
  Non-Indigenous        208           2197

               Accuracy : 0.9371
                 95% CI : (0.9297, 0.944)
    No Information Rate : 0.5002
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8742

 Mcnemar's Test P-Value : 2.386e-14

            Sensitivity : 0.9085
            Specificity : 0.9657
         Pos Pred Value : 0.9636
         Neg Pred Value : 0.9135
             Prevalence : 0.4998
         Detection Rate : 0.4540
   Detection Prevalence : 0.4712
      Balanced Accuracy : 0.9371

       'Positive' Class : Indigenous
```

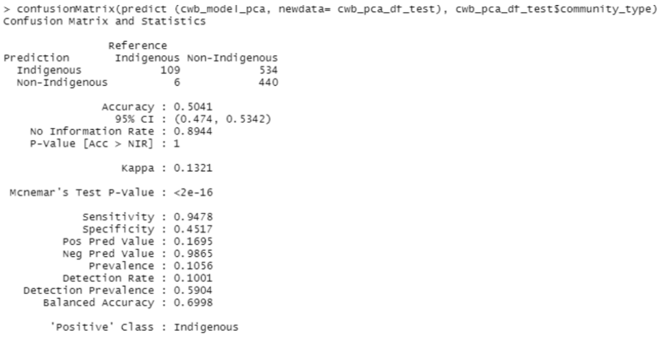


```
> confusionMatrix(predict (log_model , newdata= test), test$community_type)
Confusion Matrix and Statistics

                Reference
Prediction       Indigenous Non-Indigenous
  Indigenous            105             33
  Non-Indigenous         10            941

               Accuracy : 0.9605
                 95% CI : (0.9472, 0.9713)
    No Information Rate : 0.8944
    P-Value [Acc > NIR] : 8.877e-16

                  Kappa : 0.8079

 Mcnemar's Test P-Value : 0.0007937

            Sensitivity : 0.91304
            Specificity : 0.96612
         Pos Pred Value : 0.76087
         Neg Pred Value : 0.98948
             Prevalence : 0.10560
         Detection Rate : 0.09642
   Detection Prevalence : 0.12672
      Balanced Accuracy : 0.93958

       'Positive' Class : Indigenous
```

Table 4: Confusion matrices for the train (left) and test (right) sets for the cwb logit model.

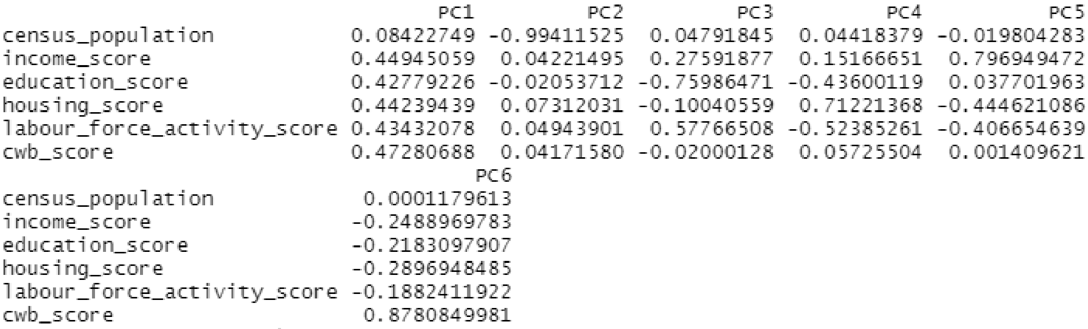


```
> confusionMatrix(predict (cwb_model_pca, newdata= cwb_pca_df), cwb_pca_df$community_type)
Confusion Matrix and Statistics

                Reference
Prediction       Indigenous Non-Indigenous
  Indigenous           2065             78
  Non-Indigenous        208           2197

               Accuracy : 0.9371
                 95% CI : (0.9297, 0.944)
    No Information Rate : 0.5002
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8742

 Mcnemar's Test P-Value : 2.386e-14

            Sensitivity : 0.9085
            Specificity : 0.9657
         Pos Pred Value : 0.9636
         Neg Pred Value : 0.9135
             Prevalence : 0.4998
         Detection Rate : 0.4540
   Detection Prevalence : 0.4712
      Balanced Accuracy : 0.9371

       'Positive' Class : Indigenous
```



```
> confusionMatrix(predict (cwb_model_pca, newdata= cwb_pca_df_test), cwb_pca_df_test$community_type)
Confusion Matrix and Statistics

                Reference
Prediction       Indigenous Non-Indigenous
  Indigenous            109            534
  Non-Indigenous          6            440

               Accuracy : 0.5041
                 95% CI : (0.474, 0.5342)
    No Information Rate : 0.8944
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1321

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9478
            Specificity : 0.4517
         Pos Pred Value : 0.1695
         Neg Pred Value : 0.9865
             Prevalence : 0.1056
         Detection Rate : 0.1001
   Detection Prevalence : 0.5904
      Balanced Accuracy : 0.6998

       'Positive' Class : Indigenous
```

Table 5: Confusion matrices for the train (left) and test (right) sets for the PCA transformed logit model.



```
                                   PC1         PC2         PC3         PC4          PC5
census_population           0.08422749 -0.99411525  0.04791845  0.04418379 -0.019804283
income_score                0.44945059  0.04221495  0.27591877  0.15166651  0.796949472
education_score             0.42779226 -0.02053712 -0.75986471 -0.43600119  0.037701963
housing_score               0.44239439  0.07312031 -0.10040559  0.71221368 -0.444621086
labour_force_activity_score 0.43432078  0.04943901  0.57766508 -0.52385261 -0.406654639
cwb_score                   0.47280688  0.04171580 -0.02000128  0.05725504  0.001409621
                                      PC6
census_population            0.0001179613
income_score                -0.2488969783
education_score             -0.2183097907
housing_score               -0.2896948485
labour_force_activity_score -0.1882411922
cwb_score                    0.8780849981
```

Table 6: Rotation values for all six predictors for PC1-PC6.

R-Code

```r
#import required libraries
library(dplyr)
library(tidyr)
library(caret)
library(bimba)
library(GGally)
library(tidyverse)
library(devtools)
library(ggfortify)
library(ggplot2)
library(combinat)
library(maptree)
library(factoextra)
library(rsample)

#import the Community Well-Being (CWB) Index 2011 dataset
cwb <- read.csv('CWB_2016_Data.csv')

#viewing the dimension of the dataframe
dim(cwb)

#viewing the first few rows and checking the class of each column
str(cwb)

#renaming the columns to exclude French, standardize the database objects, and
# provide more context of the feature
names(cwb) = c(
  "csd_code",
  "csd_name",
  "census_population",
  "income_score",
  "education_score",
  "housing_score",
  "labour_force_activity_score",
  "cwb_score",
  "community_type"
)

#checking for NA-values
sapply(cwb, function(x) sum(is.na(x)))

#drop rows with NA-values
cwb <- na.omit(cwb)

#checking again for NA-values
sapply(cwb, function(x) sum(is.na(x)))

#removing the Inuit Community class from the study
cwb<- cwb[ grep("Inuit Community", cwb$community_type, invert = TRUE) , ]

#rename the community_type values to Indigenous and Non-Indigenous
cwb$community_type[cwb$community_type == 'Non-Indigenous Community
/ Communauté non-Autochtone'] <- 'Non-Indigenous'
cwb$community_type[cwb$community_type =='First Nations Community
/ Communauté des Premières Nations'] <- 'Indigenous'

#counting class occurrences of the community_type column
dplyr::count(cwb, community_type, sort = TRUE)

#summary of the dataset
summary(cwb)

#exploratory analysis using ggpairs
GGally::ggpairs(cwb[c(-1,-2)], progress=F)

#looking closer at the density distributions for each diagnosis with featurePlot
featurePlot(x=cwb[,3:8], y=as.factor(cwb$community_type),
plot="density", scales=list(x=list(relation="free"), y=list(relation="free")),
auto.key=list(columns=2))

#set seed and split into train-test set (70/30)
set.seed(42)
cwb_numeric <- cwb[c(-1,-2)]
cwb_train <- createDataPartition(cwb_numeric$community_type, p = 0.7,
list = FALSE, times=1)
train <- cwb_numeric[cwb_train,]
test <- cwb_numeric[-cwb_train,]

#summary statistics grouped by community_type
by(cwb_numeric, cwb_numeric$community_type, summary)

#create balanced dataset with SMOTE
train$community_type = as.factor(train$community_type )
train_smote <- SMOTE(train, perc_min = 50, k=5)

#view if the response variable is balanced
round(prop.table(table(train_smote$community_type)), 2)
```

```r
#implement logistic model and view the summary
log_model = train(
  form = community_type~ .,
  data = train_smote,
  trControl = trainControl(method = "cv", number = 5),
  method = "glm",
  family = "binomial")
summary(log_model)

#application to test set and viewing the resultant table
log_model_test <- predict(log_model, test)
log_model_table<- table(log_model_test, test$community_type)
log_model_table

#confusion matrix of train and test set
confusionMatrix(predict (log_model , newdata= train_smote),
train_smote$community_type)
test$community_type = as.factor(test$community_type )
confusionMatrix(predict (log_model , newdata= test), test$community_type)

#Using PCA to address multicollinearity
cwb_pca <- prcomp(train_smote[,1:6], retx=TRUE, center=TRUE, scale=TRUE)

#plotting the cumulative proportion of variance explained
cwb_pca_var <- cwb_pca$sdev^2
cwb_pca_var_perc <- cwb_pca_var / sum(cwb_pca_var)
cum_cwb_pca_var_perc <- cumsum(cwb_pca_var_perc)
cwb_pca_var_perc_table <- tibble(comp = seq(1:ncol(cwb_numeric[,1:6])),
cwb_pca_var_perc, cum_cwb_pca_var_perc)

ggplot(cwb_pca_var_perc_table, aes(x = comp, y = cum_cwb_pca_var_perc)) +
  geom_point() +
  geom_abline(intercept = 0.95, color = "blue", slope = 0) +
  labs(x = "Number of Components", y = "Cumulative Explained Variance",
  title = "Cumulative Proportion of Explained Variance")+
  scale_x_continuous(breaks=c(0,2,4,6,8,10))

#visualizing the first five principle components
pca_vis <- cbind(as_tibble(train_smote$community_type), as_tibble(cwb_pca$x))
GGally::ggpairs(pca_vis, columns = 2:6, ggplot2::aes(color = value))

#feature importance in the first two PCA vectors
autoplot(cwb_pca, data = train_smote,  colour = 'community_type',
loadings = FALSE, loadings.label = TRUE, loadings.colour = "blue")

#repeating logit model for PCA transformed dataset
#build PCA dataframe
cwb_pca_df <- data.frame(cwb_pca$x, train_smote)

#Using the first 5 principle components as the predictors , create the logit
#model from glm function
cwb_model_pca = train(
  form = community_type~PC1 + PC2 + PC3 + PC4 + PC5 + PC6,
  data = cwb_pca_df,
  trControl = trainControl(method = "cv", number = 5),
  method = "glm",
  family = "binomial")

#viewing the summary report
summary(cwb_model_pca)

#application to test set
cwb_pca_test <- prcomp(test[,1:6], retx=TRUE, center=TRUE, scale=TRUE)
cwb_pca_df_test <- data.frame(cwb_pca_test$x, test)
cwb_model_test <- predict(cwb_model_pca, cwb_pca_df_test)

#confusion matrix of train and test set
confusionMatrix(predict (cwb_model_pca, newdata= cwb_pca_df),
cwb_pca_df$community_type)
confusionMatrix(predict (cwb_model_pca, newdata=cwb_pca_df_test),
cwb_pca_df_test$community_type)
```