Project
Mortality Rate Modeling

Thuy Ha Phuong Dang

DSCI 400. Machine Learning I
Professor Bevan Ferreira

SUMMARY

Nowadays, there are many ways to determine the overall condition of a country, one of those is to make a good observation of the vital statistics, such as birth and death rates, and those numbers may reveal a lot of underlying issues of a society, especially in the long run. In this project, we will apply machine learning methods in forecasting the mortality rate in Canada through demographic and socioeconomic characteristics.

INTRODUCTION

One of the most remarkable achievements of modern societies is the widespread reduction in mortality that has occurred over the past two centuries. For thousands of years before the nineteenth century, life expectancy at birth remained around 30 years. Today, life expectancy in the most advanced countries exceeds 80 years, and even in the countries with the worst average life expectancy has reached 53 years (World Population Ageing 2019, 2019,).

LITERATURE REVIEW

Governments can utilize mortality prediction to plan for the shifting needs of their societies in terms of health care and other services, as well as to estimate the future rate of population aging (European Commission, 2010).Several studies have been done on mortality, but one is the most well-known and has had the biggest impact on society thus far. In the early 1990s, the World Bank sponsored the 1990 Global Burden of Disease study implemented by researchers at Harvard University and the World Health Organization (WHO). The study provided the first comprehensive, global estimates of death and diseases by sex, and age, as well as projections of global mortality up to 2020 using models assumed that health trends are related to a set of predictors, such as income per person, tobacco use, etc (Murray, 1996). This project has been widely used by WHO and governments to plan ahead health policies although the statistics need updating over time with the current data as it is based on the 1990 estimates (Figure 1). As a consequence, the death rate plays an essential role in not only healthcare but also another perspective of society.

Over the last century, the annual number of deaths in Canada has gradually grown, from around 110,000 in 1926 to 240,000 in 2011 (Martel, 2013). Canada is anticipated to have the highest life expectancy at age 65 by 2030 for both genders, as compared with the US and UK. On June 16, 2023, the Canadian population celebrated by reaching 40 million, and on July 1, 2023, it had a growth of 1,158,705, approximately 2.9%, as compared to 2022. This marks the highest annual

population growth rate since 1957. Alberta experienced the greatest rise in demographic rate with 4.0% among all the provinces and territories, followed by Maritime provinces (Population estimates, quarterly, 2023). Despite experiencing such rapid population increase, the aging population issue in Canada continues happening as the population aged 65 or older is also expanding quickly. Facing the wave of an aging population, Canada must deal with many consequences that come with the issue, such as the increase in the need for healthcare facilities and costs, declining labor productivity, unsustainable pension commitments, etc. Furthermore, as the age structure of the population becomes older, a relatively greater proportion of the total population is found in the older age groups that experience higher rates of mortality. Kitty S. Chan et al suggested that demographic characteristics, especially the percentage older than 65 years, can be a great predictor of all-cause death.

Extensive literature has demonstrated and described the strong correlation between labor market disadvantage and health and mortality (Roelfs et al., 2011). However, determining wheter the link between unemployment and mortality is a causal relationship still remains difficult since there are numerous mediators along the pathway, such as poor health is a risk factor for both unemployment and mortality. Even so, it cannot be denied that the unemployment issue should always be assessed in terms of projecting mortality rates.

In North America, mortality rates are higher among individuals of lower socioeconomic status than among those of higher status. There are many reasons that can explain this phenomenon. For example, poverty and material hardship have been argued to be the main underlying cause of social inequalities in mortality and morbidity. Furthermore, income, more so than occupational class and particularly education, may be partly determined by pre-existing ill-health. Moreover, using data from Canada's census or national health surveys, Canadian researchers have observed that despite universal health insurance, mortality rates are greater among individuals of lower socioeconomic status compared with those of higher socioeconomic status (Finkelstein, 2018).

The social, economic, and environmental conditions that people experience throughout their lives are the most significant influences on their health. Education can impact health through various pathways. Educational attainment may be a sign of intra- and inter-personal skills that are required to produce and maintain good health (Lantz et al., 2010). A number of research have examined whether education has a causal impact on health, with some finding a positive effect and others no significant effect. In Western and eastern European countries, the association between education and mortality is well established (Mackenbach et al., 2008). However, there is little information available regarding the connection between health and education in Canada.

Wilkins et al. supported this point through their 16-year follow-up study in Canada, and they illustrated that the higher level of education group had a lower mortality rate for the majority of causes of death (Wilkins et al., 2009).

THE DATASET

The data set includes four predictors - that are, people with low-income percentage, a proportion of 65-year-olds and over population, and unemployment rate, one response variable - mortality rate, with 100 observations. Information was extracted from Statistics Canada along with their metadata. Data for all variables is in only available for 10 years, from 2010 to 2019, just before the onset of Covid-19.

- Region (named in the dataset) defined as official provinces of Canada (not including territories): Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan. Along with the province, the annual population estimates of each province from 2010 to 2019 were added as well. This data is from table 17-10-0005-01 (formerly CANSIM 051-0001) and released in 2022.
- Low-income percentage (Income) is a continuous variable, which is measured by taking estimates of the population with low income divided by the total population and multiplied by 100. Values were extracted from the percentage of Low-income cut-offs after tax (LICO-AT) 1992 base, in table: 11-10-0135-01 (formerly CANSIM 206-0041) of Centre for Income and Socioeconomic Well-being Statistics, Statistics Canada. LICO-AT 1992 base is defined as: “The low-income cut-offs after tax (LICO-AT) are income thresholds below which a family will likely devote a larger share of its after-tax income on the necessities of food, shelter, and clothing than the average family. The approach is essentially to estimate an income threshold at which families are expected to spend 20 percentage points more than the average family on food, shelter, and clothing, based on the 1992 Family Expenditures Survey. LICOs are calculated in this manner for seven family sizes and five community sizes.” (Low-income cut-offs (LICOs) before and after tax by community size and family size, in current dollars, 2023).
- Senior percentage (Age) is a continuous variable, measured by taking the estimates of people 65 years old and over devided by total population, and then multiplied by 100. Data were taken from table: 17-10-0005-01 (formerly CANSIM 051-0001).
- Unemployment percentage (Unemployment), which is a continuous variable, is the annual unemployed population divided by the total population, and multiplied by 100. Data is from the Labour Force Survey, Statistics Canada.

- Regarding the response variable, the annual mortality percentage is from Birth and Death Databases and the Centre for Demography, Statistics Canada.

DATA PREPARATION AND EXPLORATION

Before applying machine learning algorithms, we initially explore the quality and correlations in the dataset. Overall, there are no missing values in all of the database and the quality of data is ranked D, meaning “Acceptable”, and above. We then plot trends of variables over 10 years by provinces.

We can take a brief look at the relationship among variables. As seen in Figure 1 below, we found Death and Age have a moderately strong positive correlation (corr = 0.785), while Death is weakly correlated with Unemployment (corr = 0.493). Income is negatively correlated with Death (corr = -0.312), yet this plot shows a weaker correlation than might be expected. We can also notice there is a weak correlation between Age and Unemployment (corr = 0.432) which should be careful as collinearity can be an issue while building the model. Finally, as region is a categorical variable, we decide using dummy coding. It's a way to turn a categorical variable into a series of binary variables (variables that can have a value of zero or just one value).

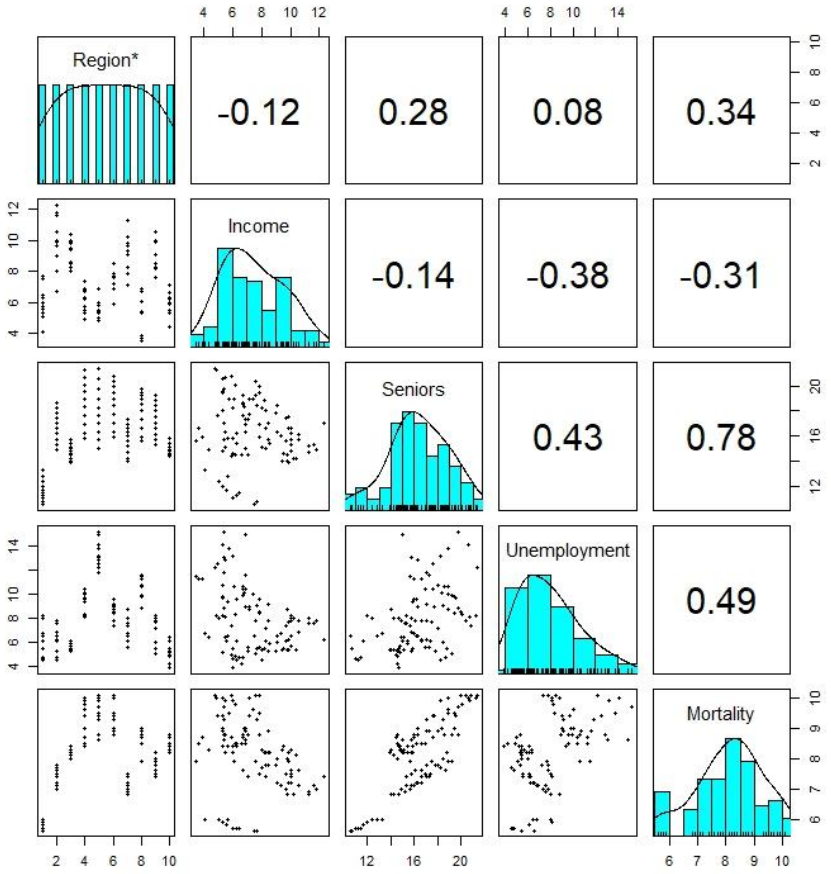

Figure 1. Plotting the Mortality dataset with ggpairs

RESULT

Since the response variable is continuous and most of the predictors are continuous variables, we use linear regression to build the model. We initially applied linear regression in our original dataset. This model gave an R-Squared (R^2) score of 0.9647, RSE of 0.2339, and coefficients as follows:

```
Call:
lm(formula = Mortality ~ ., data = project_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97490 -0.10470  0.00758  0.10616  0.53417

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.25616    0.56688   9.272 1.25e-13 ***
Income        0.01536    0.02970   0.517   0.6068
Seniors       0.21491    0.02597   8.275 7.69e-12 ***
Unemployment -0.02751    0.03664  -0.751   0.4553
AB           -1.89865    0.15204 -12.488  < 2e-16 ***
BC           -1.48516    0.18902  -7.857 4.35e-11 ***
MB           -0.18778    0.15977  -1.175   0.2440
NB            0.10338    0.24102   0.429   0.6693
NL            0.54133    0.34084   1.588   0.1169
NS            0.45673    0.23548   1.940   0.0566 .
ON           -1.46914    0.17582  -8.356 5.49e-12 ***
PE           -0.25494    0.27603  -0.924   0.3590
QC           -1.15023    0.19324  -5.952 1.07e-07 ***
SK                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2339 on 67 degrees of freedom
Multiple R-squared:  0.9647,    Adjusted R-squared:  0.9583
F-statistic: 152.5 on 12 and 67 DF,  p-value: < 2.2e-16
```

Figure 2. Performance of linear model from the original dataset

The summary of the model demonstrates comprehensive information about each regressor and how they contribute to the response variable. We can notice that the algorithm could not offer us the coefficient for the last predictor, and the contributions of half of the predictors are not significant. We can presume that there exists collinearity among the independent variables in all of the data. This leads us to the next step, which is to exam the use of PCA .

After scaling and centering the dataset, we performed PCA to reduce dimensions and mitigate the impact of multicollinearity. The Scree Plot in Figure 3 illustrates how much each PC accounted for percentage variance. As can be seen, there are huge gaps between PC1 - PC2, PC9 - PC10. Thus, we should consider at the table of PCA summary below for the Cumulative Proportion, which helps us determine that we should retain the top 9 principal components. In the end, we decided to develop a new model based on new variables created by PCA.

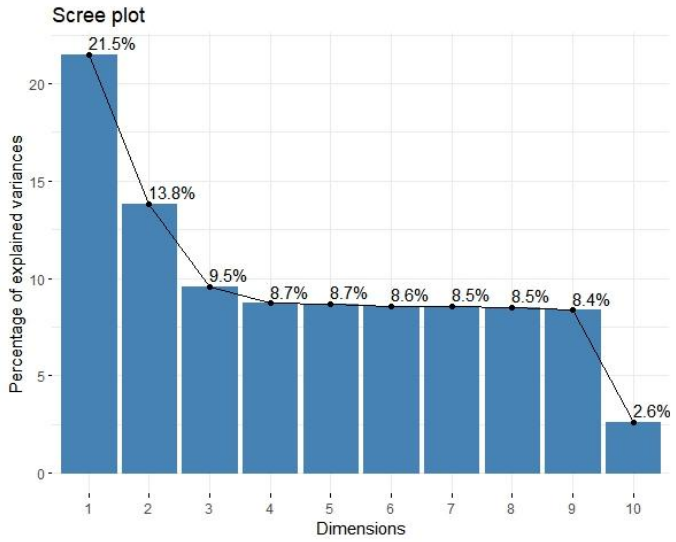

Figure 3. Scree Plot for PCA

The linear model has shown a noticeable improvement post PCA as compared to the former one Firstly, the model is now capable of measuring sufficient coefficients for the predictors, which is a good sign of successful solution to the problem of collinearity. Furthermore, all the predictors have made significant contributions to the response variable. Although Residual Standard Error (RSE) has increased and R^2 has decreased, these changes are all explicable as we have only selected the important principal components.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.82918 -0.13950  0.00671  0.15953  0.49546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.19125    0.03061 267.570  < 2e-16 ***
PC1         -0.45112    0.01843 -24.479  < 2e-16 ***
PC2         -0.19661    0.02299  -8.552 1.76e-12 ***
PC3         -0.28846    0.02765 -10.433 6.67e-16 ***
PC4          0.16625    0.02895   5.742 2.22e-07 ***
PC5          0.34961    0.02903  12.044  < 2e-16 ***
PC6         -0.21506    0.02919  -7.366 2.67e-10 ***
PC7         -0.20833    0.02923  -7.128 7.28e-10 ***
PC8          0.26699    0.02930   9.111 1.65e-13 ***
PC9          0.15997    0.02955   5.413 8.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2738 on 70 degrees of freedom
Multiple R-squared:  0.9438,    Adjusted R-squared:  0.9366
F-statistic: 130.7 on 9 and 70 DF,  p-value: < 2.2e-16
```

Figure 4. Performance of linear model from the dataset after perorming PCA

To ensure the absence of redundant regressor, we utilize an additional method with a less subjective approach to feature selection, which allow another algorithm to evaluate relevant features. We use filter method to compare each predictor to the response variable. As per Figure 5w, which ranks the degree of information that each independent variable can contribute to the model. As a result, we decide to retain all nine predictors.

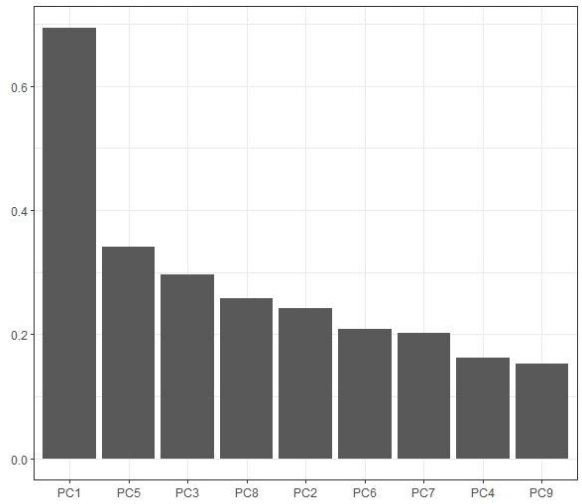

Figure 5. Feature selection for the new linear model

Following feature selection, it is necessary to verify 4 assumptions of linear regression: Linearity, Equal Variance, Independence, and Normality. Given that PCA was used to eliminate collinearity, it is acceptable to skip the Independence test.

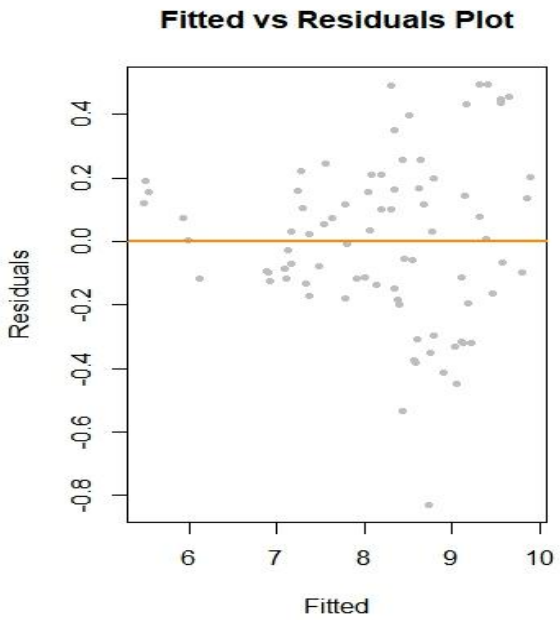

Figure 6. Fitted vs Residuals Plot

Fitted versus Residuals Plot in Figure 6 is used for checking both the Linearity and Equal Variance assumptions. The Linearity test reveals whether the relationship between dependent and independent variables is linear. We observe that at any fitted value, the mean of the residuals is roughly 0. For this reason, we can conclude that the linearity assumption is not violated, and it indicates we have a linear relationship between predictors and response variables. On the other hand, there is heteroscedasticity in this model since the spread of the residuals is not consistent across all variables. Thus this violation of the Equal Variance assumption indicates the presence of heteroscedasticity.

In order to check for Normality assumption, we determine whether residuals are normally distributed by performing a Q-Q Plot and a Histogram Plot. In Figure 7, the points of Q-Q plot do not closely follow the line, suggesting that the errors may not follow a normal distribution. The histogram appears not to display a bell-shaped curve, with its skewness and kurtosis is -0.2440806 and 3.33, respectively, inferring a slightly left-skewed and fat-tailed graph. These observations lead us to suspect that a violation of normality. Since graph visualization is insufficient to draw a conclusion,the Shapiro-Wilk test should be used with hypothesis H0: The distribution follows a normal distribution. This normality test yielded a p-value of 0.2592, which is greater than alpha of 0.05, we cannot reject the hypothesis that the distribution follows a normal distribution.

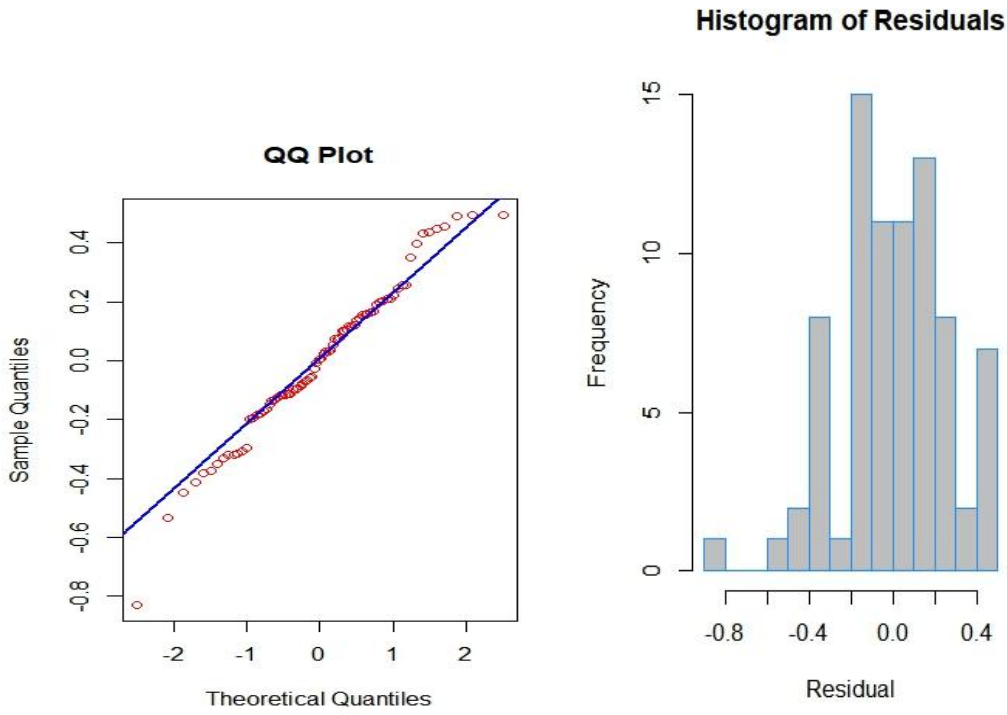

Figure 7. Q-Q Plot and Histogram of Residuals

We now predict mortality rate in testing set using the newly constructed model and then compare results to the observed rates. The association between the predicted and observed values is subsequently validated by implementing a linear regression. Both Root Mean Squared Error (RMSE) and R^2 are used to examine the model. R^2 offers information about the goodness of fit of a model. It is the rate of the sum of explained variation to the total variation:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

As a proportion, the range of R^2 is from 0 to 1. In the case where R^2 = 1, every data point lies perfectly on the regression line, meaning the predicted values the observations are identical. On the other hand, an R^2 value of 0 implies that the our model to predict mortality rate fails. In this case, the R^2 is 0.9796, which is pretty close to 1, indicating that 97.96% of the observed values can be accounted for by the predicted values.

In contrast, RMSE evaluates the standard deviation of residuals. RMSE is a good measure of how accurately the model predicts the response, and is the most important criterion for fit if the main purpose of the model is prediction. Lower value of RMSE denotes a better fit. The formula of RMSE is as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_i - \hat{y})^2}$$

As can be seen, there is no thumb of rule regarding good or bad RMSE because it depends on the absolute value of dependent variable, which is the observed mortality rate. Our mean, min and max of mortality rate are approximately 8, 5 and 10, respectively, while RMSE is 0.2, roughly equal to 2.5% of the response value. It means that there is only 2.5% of discrepancy between the predicted and observed values. Figure 8 depicts the data points are distributed closely around the regression line, indicating a degree of variance in observed values in charge in predicted ones. In general, the graph suggest a robust positive linear regression between those two variables.

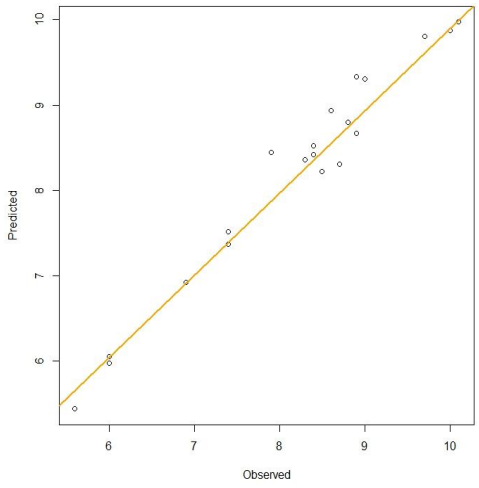

Figure 8. Linear model between predicted values and observed values

Outcomes and Limitations

Some limitations remain with regard to this model. To begin with, it has low interpretability as despite Principal components are linear combinations of the features from the original data, they are not as easy to interpret. While the model effectively predicts the mortality rate using the provided dataset, it is not easily determine the individual contribution of each regressor to the mortality. In lieu of PCA, we can employ backward or stepwise selection using Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Secondly, there is still an issue with the condition of heteroscedasticity, which requires a transformation of response variables using Box-Cox method, which aims to find a more exact form for the relation. Finally, the potential concern about overfitting may happen due to the small size of the dataset. In spite of all these constraints, we have constructed a linear model to predict the mortality rate of Canada with give predictors.

REFERENCES

Bhatia, R. (2020). Predictions of COVID-19 related unemployment on suicide and all-cause mortality. https://doi.org/10.1101/2020.05.02.20089086

Canada at a glance. (2022, November 23). Statistics Canada: Canada's national statistical agency / Statistique Canada : Organisme statistique national du Canada. https://www150.statcan.gc.ca/n1/pub/12-581-x/12-581-x2022001-eng.htm?utm_source=rddt&utm_medium=smo&utm_campaign=statcan-all-content-22-23

Chan, K. S., Roberts, E., McCleary, R., Buttorff, C., & Gaskin, D. J. (2014). Community characteristics and mortality: The relative strength of association of different community characteristics. *American Journal of Public Health*, *104*(9), 1751-1758. https://doi.org/10.2105/ajph.2014.301944

Closing the gap in a generation: Health equity through action on the social determinants of health : Commission on social determinants of health final report. (2008). World Health Organization. https://www.who.int/publications/i/item/WHO-IER-CSDH-08.1

Dealing with the impact of an ageing population in the EU. (2009 Ageing Report). (2009). COMMISSION OF THE EUROPEAN COMMUNITIES. https://ec.europa.eu/economy_finance/publications/pages/publication14992_en.pdf

Finkelstein, M. (2018, April). *Relationship between income and mortality in a Canadian family practice cohort*. Can Fam Physician. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5897086/

Lantz, P. M., Golberstein, E., House, J. S., & Morenoff, J. (2010). Socioeconomic and behavioral risk factors for mortality in a national 19-year prospective study of U.S. adults. *Social Science & Medicine*, *70*(10), 1558-1566. https://doi.org/10.1016/j.socscimed.2010.02.003

Mackenbach, J. P., Stirbu, I., & Roskam, A. I. (2008). Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine*, *359*(12), e14. https://doi.org/10.1056/nejmx080032

Martel, L. (2013, July 9). *Mortality: Overview, 2010 and 2011*. Statistics Canada: Canada's national statistical agency / Statistique Canada : Organisme statistique national du Canada. https://www150.statcan.gc.ca/n1/pub/91-209-x/2013001/article/11867-eng.htm

Murray, C. J., & Lopez, A. D. (1996). *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020 ; Summary*. THE GLOBAL BURDEN OF DISEASE.

Population estimates, quarterly. (2023, September 27). Statistics Canada. https://doi.org/10.25318/1710000901-eng

Roelfs, D. J., Shor, E., Davidson, K. W., & Schwartz, J. E. (2011). Losing life and livelihood: A systematic review and meta-analysis of unemployment and all-cause mortality. *Social Science & Medicine*, *72*(6), 840-854. https://doi.org/10.1016/j.socscimed.2011.01.005

Wilkins, R., & Tjepkema, M. (2009). *The Canadian census mortality follow-up study, 1991 through 2001, summary: Summary of results*. Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/82-003-x/2008003/article/10681-eng.pdf?st=1LLEG8bZ

World Population Ageing 2019. (2019). United Nations. https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Report.pdf

APPENDIX

```
library(mlr)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(fastDummies)
library(psych)

project_ln_full = read.csv("D:\\OC\\Semester 2\\MLI 400\\Project\\Data\\Project_lr.csv")
project_ln_full_Tib = as.tibble(project_ln_full[-1])
summary(project_ln_full_Tib)
library(GGally)
project_plot=read.csv("D:\\OC\\Semester 2\\MLI 400\\Project\\Data\\Project_lr _ Acronym province.csv")
ggpairs(project_plot,upper = list(continuous="cor",combo = "facetdensity"),lower = list(continuous=wrap("smooth", alpha = 0.3, size=0.1),combo = "box_no_facet"))
 +theme(axis.text.x = element_text(angle = 90, hjust = 1))
pairs.panels(project_plot,smooth=FALSE,lm=TRUE)

project_ln_full_Tib_dummy = dummy_cols(project_ln_full_Tib,select_columns="Region")
project_ln_full_Tib_dummy
project_ln_full_Tib_dummy_noregion=project_ln_full_Tib_dummy[,-1]
project_full= project_ln_full_Tib_dummy_noregion[sample(1:nrow(project_ln_full_Tib_dummy)),]
project_train=project_full[1:80,]
project_test = project_full[81:100,]
colnames(project_train)=c("Income","Seniors","Unemployment","Mortality","AB","BC","MB","NB","NL","NS","ON","PE","QC","SK")
colnames(project_test)=c("Income","Seniors","Unemployment","Mortality","AB","BC","MB","NB","NL","NS","ON","PE","QC","SK")

project_train.lm= lm(Mortality~.,data=project_train)
summary(project_train.lm)
```

#PCA

```
pca = select (project_train, -Mortality,) %>%
  prcomp(center=TRUE, scale=TRUE)
pca
summary(pca)
library(factoextra)
pcaDat <- get_pca(pca)
fviz_pca_biplot(pca, label = "var")+ theme_bw()
fviz_pca_var(pca)
fviz_screeplot(pca, addlabels = TRUE, choice = "eigenvalue")
fviz_screeplot(pca, addlabels = TRUE, choice = "variance")

projectPca = project_train%>%
  mutate(PC1 = pca$x[, 1], PC2 = pca$x[, 2],PC3 = pca$x[, 3],PC4 = pca$x[, 4],
         PC5 = pca$x[, 5],PC6 = pca$x[, 6],PC7 = pca$x[, 7],PC8 = pca$x[, 8],
         PC9 = pca$x[, 9])
project_train_Pca = as.tibble(projectPca[,c("Mortality","PC1","PC2","PC3","PC4"
                                  ,"PC5","PC6","PC7","PC8","PC9")])
ggplot(project_train_Pca, aes(PC1, PC2,  col = Mortality)) +
  geom_point() +
  theme_bw()

project_train_Pca
project_train_Pca.lm=lm(Mortality~PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9,
                        data=project_train_Pca)
summary(project_train_Pca.lm)
```

#Feature Selection

```
projectTask = makeRegrTask(data=project_train_Pca,target = "Mortality")
filterVals <- generateFilterValuesData(projectTask,
                                method = "linear.correlation")
filterVals$data
plotFilterValues(filterVals) + theme_bw()
```

#Check for normality

```
library(moments)
hist(resid(project_train_Pca.lm),
    xlab = "Residual",
    main = "Histogram of Residuals",
    col = "grey",
    border = "dodgerblue",
    breaks = 15)
skewness(resid(project_train_Pca.lm))
kurtosis(resid(project_train_Pca.lm))

par(mfrow = c(1, 2))
project_train_Pca.lm$residuals
project_train_Pca.lm$fitted.values

plot(fitted(project_train_Pca.lm), resid(project_train_Pca.lm), col = "grey", pch = 20, xlab = "Fitted", ylab = "Residuals", main = "Fitted vs Residuals Plot")
abline(h = 0, col = "darkorange", lwd = 2)
qqnorm(resid(project_train_Pca.lm), main="QQ Plot", col="red")
qqline(resid(project_train_Pca.lm),col="blue",lwd=2)

shapiro.test(resid(project_train_Pca.lm))
```

#Cross Validation

```
library(caret)
CV= trainControl(method="repeatedcv",number=8,repeats = 10)
CV_train=train(Mortality~PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9,data=project_train_Pca,method="lm",trControl=CV)
CV_train
```

#Predict

```
project_test_pca = predict(pca,project_test)
project_test_pca_df=as.data.frame(project_test_pca)
```

```
predict_trial=predict(project_train_Pca.lm,newdata =project_test_pca_df)
predict_trial
library(Metrics)
data.frame( R2 = R2(predict_trial, project_test$Mortality),
        RMSE = RMSE(predict_trial, project_test$Mortality),
        MAE = MAE(predict_trial, project_test$Mortality))
project_test_predict_full=mutate(project_test,predict_trial)
colnames(project_test_predict_full)[15]="Mortality_predicted"

observed_predict.lm = lm(Mortality~Mortality_predicted,project_test_predict_full)
summary(observed_predict.lm)
plot(project_test_predict_full$Mortality,project_test_predict_full$Mortality_predicted,  ylab  =
"Observed",xlab = "Predicted")
+abline(observed_predict.lm,col="orange", lwd=2)
```