STAT3503A

Exploratory analysis and preliminary results

November 27, 2020

School of Mathematics and Statistics

Carleton University

Table of Contents

List of Figures

Introduction

The goal of this project is to find the optimal model to estimate labour force from the engineered dataset from different verified sources. We will focus on labour force in Canada using the data from OpenGov Data Source. Upon researches, there are some variables that we believe to be good candidates for labour force estimator such as population distribution factors like by gender, province, educational level, marital status and change in population number from previous year; Moreover, we also consider integrating the economic factor which can be represented by Gross Domestic product (GDP).

Having a model with all of proposed variables can cause the degradation in output quality so having a model with combination of appropriated variables is more efficient approach. Hence, we aim to investigate the interactions between variables, eliminate auxiliary variable, and mutate new variable(s) if needed.

The sets of models I'm interested to test on this project is:

- The output model using stepwise with considerations of all permutation of variable's interaction.
- The output model using ANOVA result with considerations of all permutation of variable's interaction.
- The output model using simple multi-linear regression (lm) with consideration of only features seem to be most valuable to the output.

The model evaluation will be heavily relied on Akaike information criterion (AIC),Residual Sum of Square (SSE) and Adjusted R-square.

Methodology

Potential Variables of Labour Force estimator

From Population Dataset, which contains different aspect of Canadian population.

1. Year as ID for each sample row (1 variable)
2. Change in population each year (1 variable)
3. Population distribution of people per province in Canada (10 variables) : the number of people resided in each province.
4. Population distribution by gender (2 variables): the number of male or female in Canada.
5. Population distribution by educational degree holding (3 variables): even though the dataset provides more details into the different type of degree holders, but we should regroup to number of people who does not have any diploma, one with high school diploma, and one with post-education or higher degree.
6. Population distribution by age-group (3 variables): we also mutate the information to get concise data, so only consider young-adult(15-24 year old), mature-adult(25-55 year old) and senior (55 and up)

From Marital Dataset, we hypothesize that marriage can impact one's decision to be in or out labour force, so it is reasonable to include marital status to our final dataset.

1. Number of married people (1 variable)
2. Number of divorce people (1 variable)

We hypothesize the growth of economy can have some impacts to labor force, because well-growing economy will result in more job-opportunities which motivate people to get in the workforce.

From GDP dataset, to narrow down the research area, we select information of Gross Domestic Product from some essential industries using 2012 constant-price. The reason why we do not choose the census GDP value because we try dimmish the inflation impact but more try to represent the growth of economy accurately. The essential industry list is based on Statistics Canada's GDP Daily analysis. Those values are:

1. GDP value of Goods-producing industries [T002]
2. GDP value of Service-producing industries [T003]
3. GDP value of Industrial production [T010]
4. GDP value of Non-durable manufacturing industries [T011]
5. GDP value of Durable manufacturing industries [T012]
6. GDP value of Agriculture, forestry, fishing and hunting [11]
7. GDP value of Mining, quarrying, and oil and gas extraction [21]
8. GDP value of Utilities [22]
9. GDP value of Construction [23]
10. GDP value of Public administration [91]
11. GDP value of Federal government public administration [911]
12. GDP value of Educational services [61]

Response value: the number of labour force of Canada in each year

In general, our dataset contains 33 predictors for 1 response value.

Exploratory data analysis

Variable's Visualization from the primitive datasets

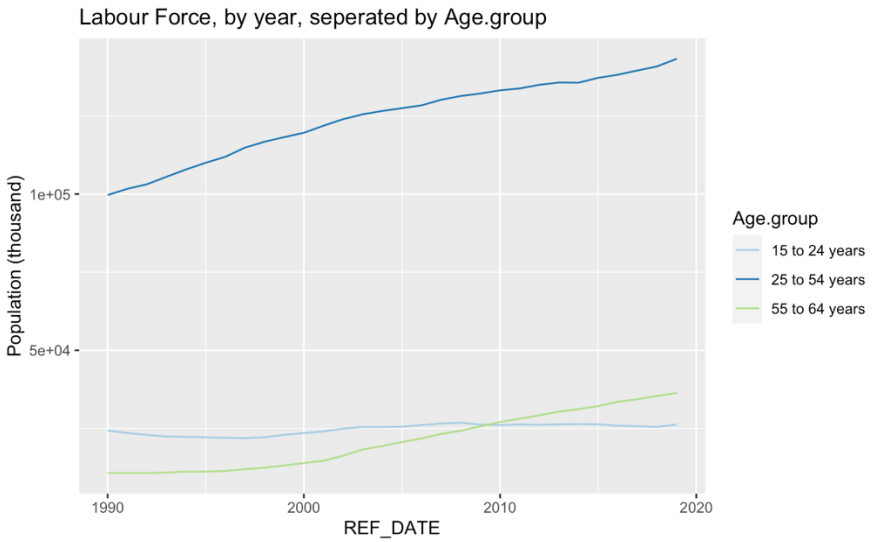


Figure 1. *Age distribution in Labour Force*

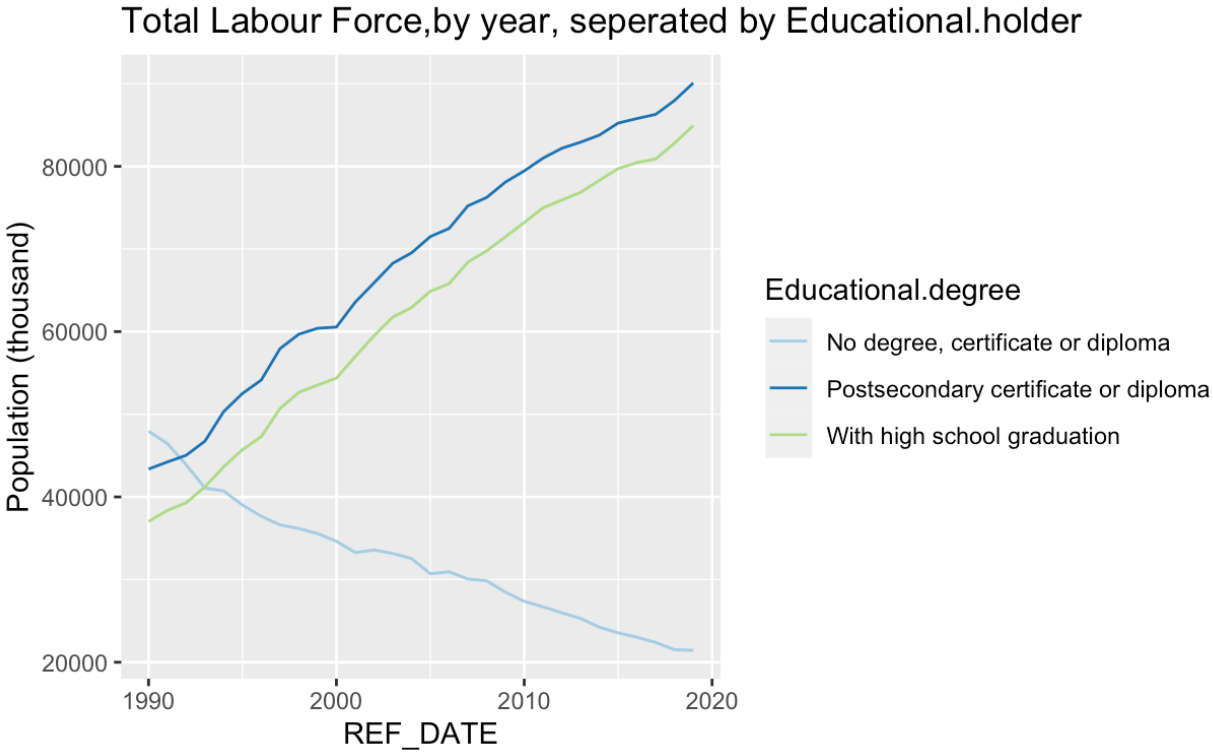


Figure 2*. Education Level in Labour Force*

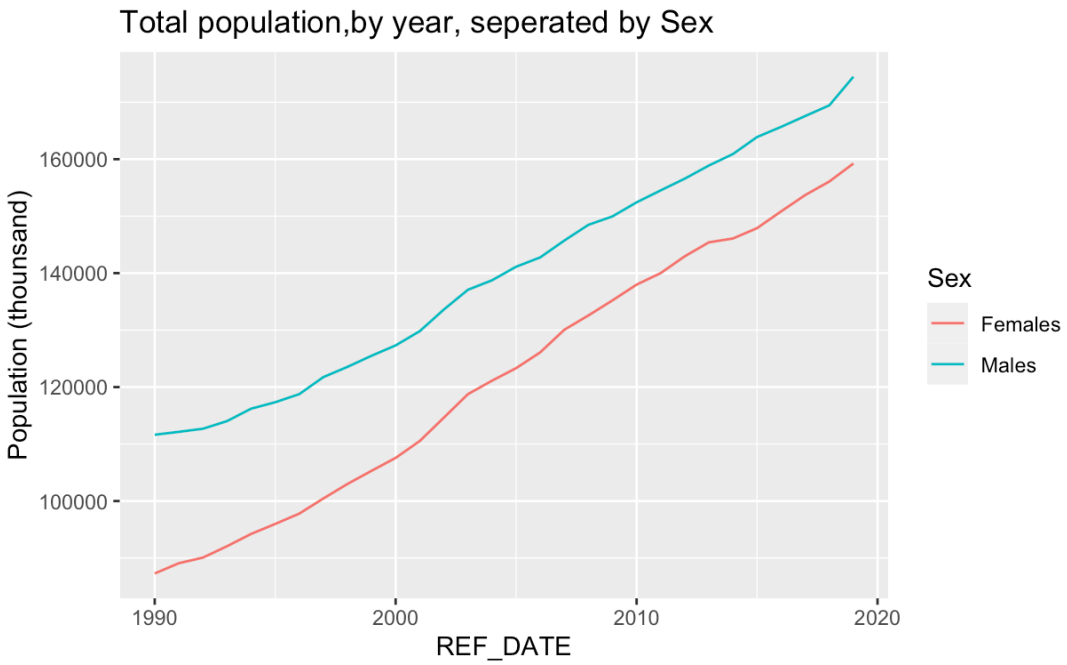


Figure 3. *Gender Participation in Labour Force*

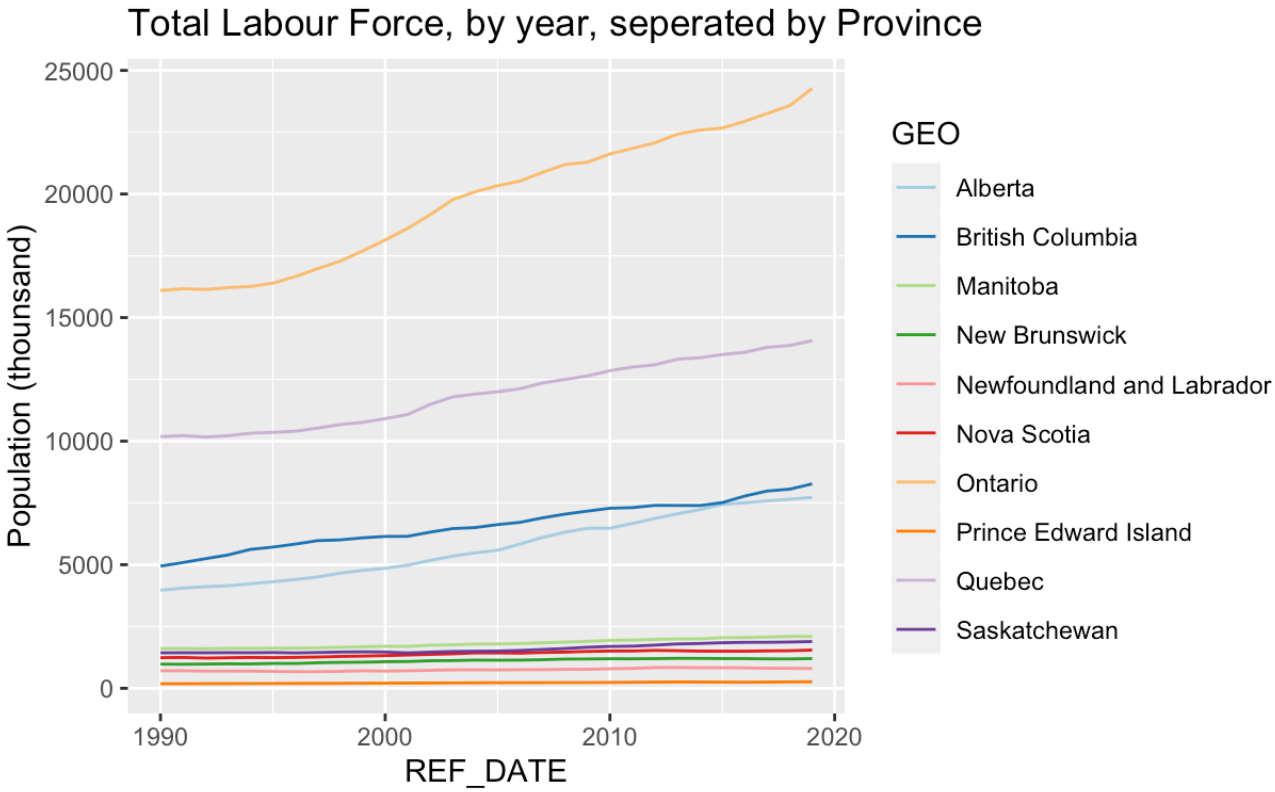


Figure 4. *Labour Force distribution by Province*

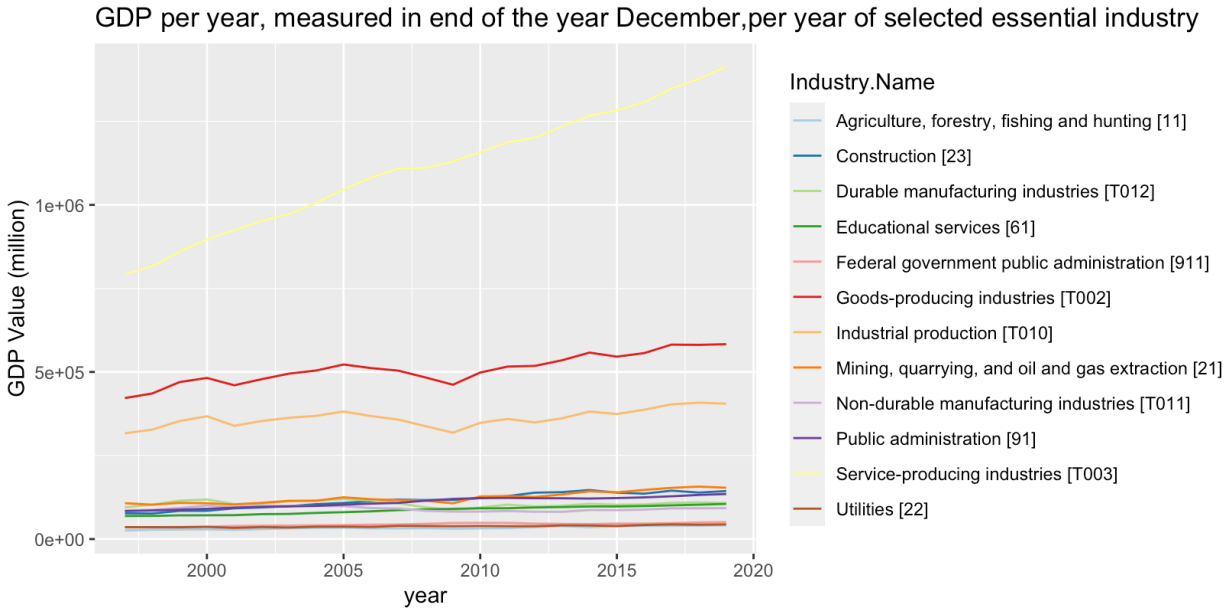


Figure 5. *GDP value per year*

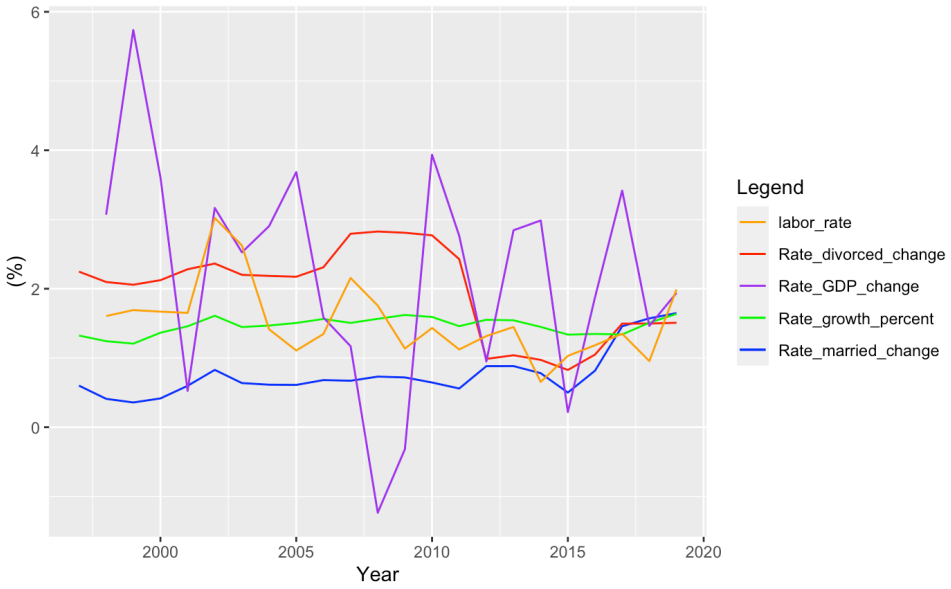


Figure 6. *Percentage of change in Labour Force Rate, Married Change, Divorced Change, GDP change, and Population Change.*

Build the Model

Limitation

I engineered my dataset from three datasets from OpenGov, where each column is uniquified by year. Unfortunately, the three dataset's timelines do not algin, so my final dataset ends up with up to 55 variables but only from year 1998 to 2020 (23 samples).

```
all_model = lm(LabourForce~.,model_data)
summary(all_model)
```

```
Call:
lm(formula = LabourForce ~ ., data = model_data)

Residuals:
ALL 22 residuals are 0: no residual degrees of freedom!

Coefficients: (33 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.799e+06         NA      NA       NA
X15.years.and.over        -1.491e+04         NA      NA       NA
X15.to.24.years            1.691e+04         NA      NA       NA
X25.years.and.over         3.292e+04         NA      NA       NA
X25.to.54.years           -1.541e+04         NA      NA       NA
X55.years.and.over        -7.913e+03         NA      NA       NA
X55.to.64.years           -6.362e+03         NA      NA       NA
X65.years.and.over        -6.448e+03         NA      NA       NA
total_no_degree           -2.352e+00         NA      NA       NA
total_highschool_grad     -4.486e+00         NA      NA       NA
total_postsecondary        3.206e+00         NA      NA       NA
Alberta                   -1.111e+03         NA      NA       NA
British.Columbia          -8.705e+02         NA      NA       NA
Manitoba                  -1.281e+03         NA      NA       NA
New.Brunswick             -1.352e+03         NA      NA       NA
Newfoundland.and.Labrador -1.320e+03         NA      NA       NA
Nova.Scotia               -7.795e+02         NA      NA       NA
Ontario                   -1.066e+03         NA      NA       NA
Prince.Edward.Island      -1.781e+03         NA      NA       NA
Quebec                    -1.147e+03         NA      NA       NA
Saskatchewan              -7.559e+02         NA      NA       NA
total_male                 3.017e+02         NA      NA       NA
total_female                      NA         NA      NA       NA
Diff_pop_growth                   NA         NA      NA       NA
Rate_growth_percent               NA         NA      NA       NA
Diff_married                      NA         NA      NA       NA
Rate_married_change               NA         NA      NA       NA
```

From Stack Overflow, I realize that my sample size can give me a lot of limitations in this project. The possible explanation is from Green (1991) makes two rules of thumb for the minimum acceptable sample size.

1. First based on whether you want to test the overall fit of your regression model (i.e. test the R2). If you want to test the model overall, then he recommends a minimum sample

size of 50 + 8k, where k is the number of predictors. So, with five predictors, you'd need a sample size of 50 + 40 = 90.

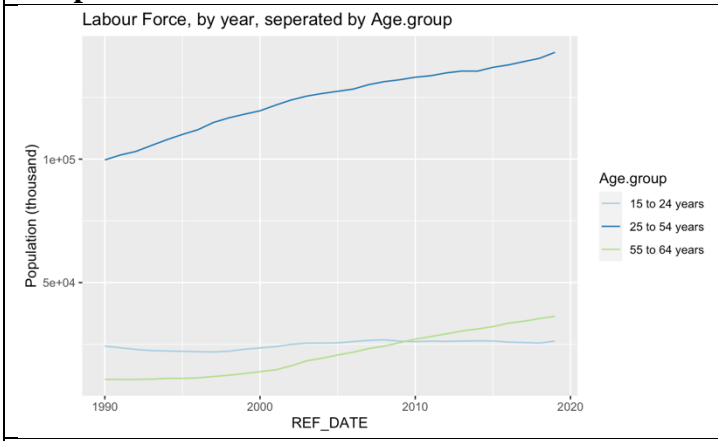
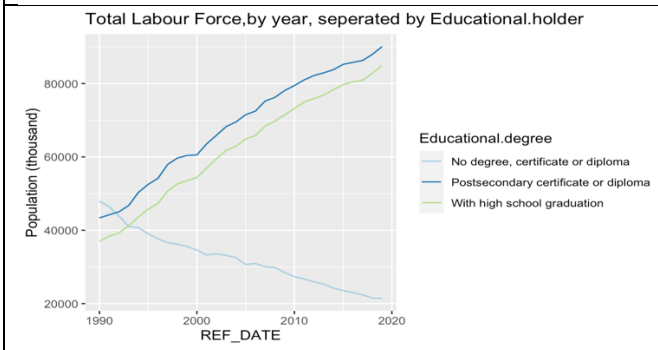
2. Depends on whether you want to test the individual predictors within the model If you want to test the model overall, then he recommends a minimum sample size of 50 + 8k, where k is the number of predictors. So, with five predictors, you'd need a sample size of 50 + 40 = 90. If you want to test the individual predictors then he suggests a minimum sample size of 104 + k, so again taking the example of 5 predictors you'd need a sample size of 104 + 5 = 109.

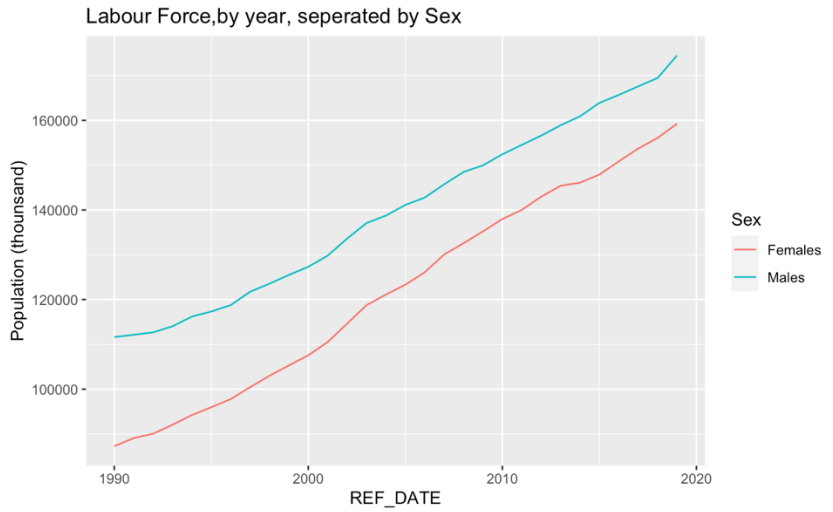
Source VanVoorhis, C. R. W., & Morgan, B. L. (2007).

This has steered my research direction. *My most important step is to select a group of up to 20 variables out of 55 variables available in my dataset*. I can't do stepwise selection at this point since the all-in model AIC is infinity right now.

Selecting Variables

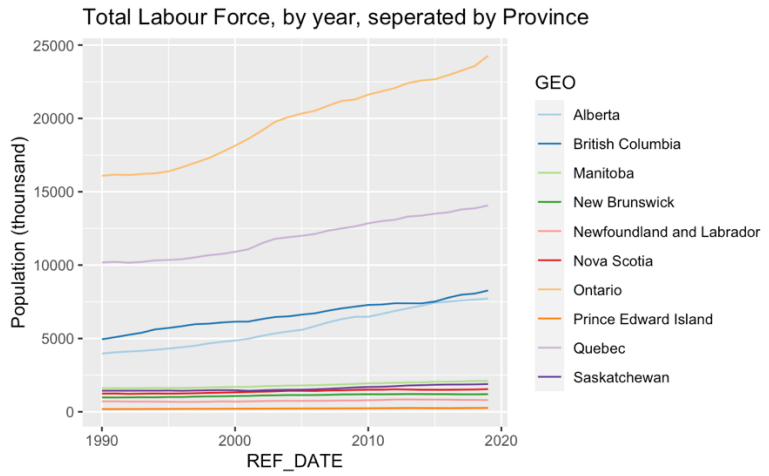
I will pick the variables of interested based on the graphs I presented last section. The naïve solution is to pick the dominant variables out of the group of variables I mentioned in variable description.

Graph	Dominant Variable
	Select the number of people belong to age group of 25-54.
	Select the number of people with postsecondary certificate or diploma and the one with high school graduation.



I'm not too convinced that the number of male and female within the population can have any impact to the labour force.

The trend is very steady.

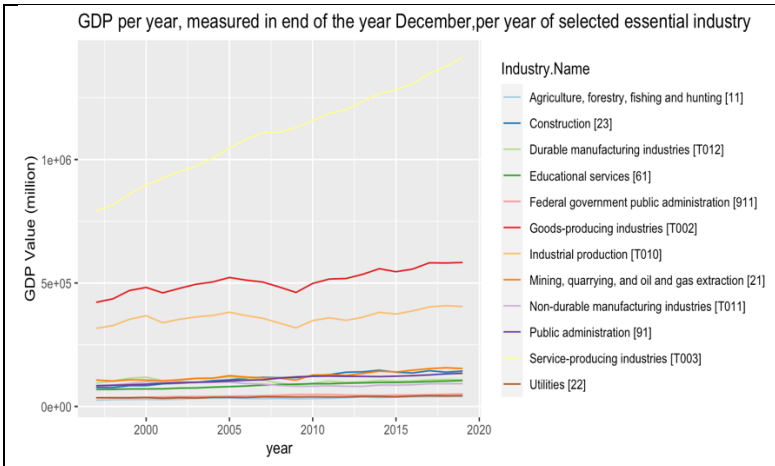


Number of people reside in Ontario seems to have a noticeable impact to our model.

GDP rate change seems to have a big impact to rate of labour force change every year.

The difference in population doesn't seem align to any difference of labour force every year.

It is worth to add divorced rate to our model.



As we saw earlier, GDP rate of change can impact the number of labour force.

It is a good hint to consider the change of GDP of the industry that has a great contribution to the country's GDP.

We can try to incorporate the rate of GDP change of Service Producing Industry, Good-Producing Industry, Industrial Production and Construction.

In conclusion, there are 13 variables I'm testing out for this project is:

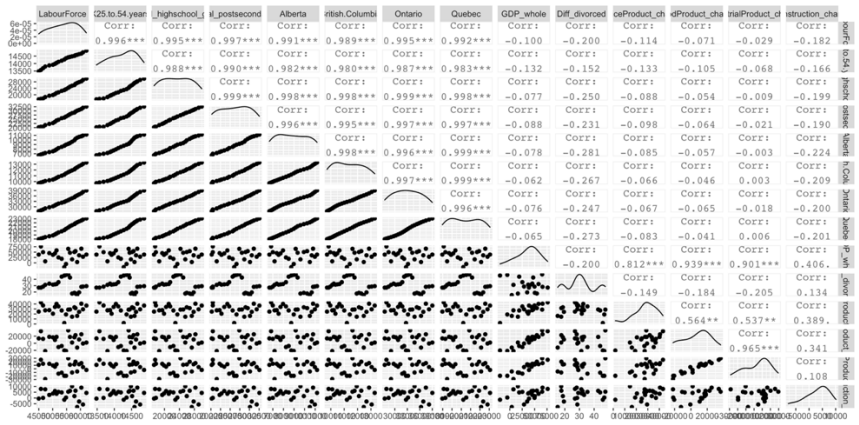
- The number of people belong to age group of 25-54 in Canada
- The number of people with postsecondary certificate or diploma
- The number of people with high school graduation
- The number of people resides in Ontario
- The number of people resides in Quebec
- The number of people resides in Alberta
- The number of people resides in BC
- The change of country's GDP annually
- The change in number of divorced people annually
- The change of Service Producing Industry's GDP annually
- The change of Good-Producing Industry's GDP annually
- The change of Industrial Production's GDP annually
- The change of Construction's GDP annually

Model Building Workflow

Build the Basic Model

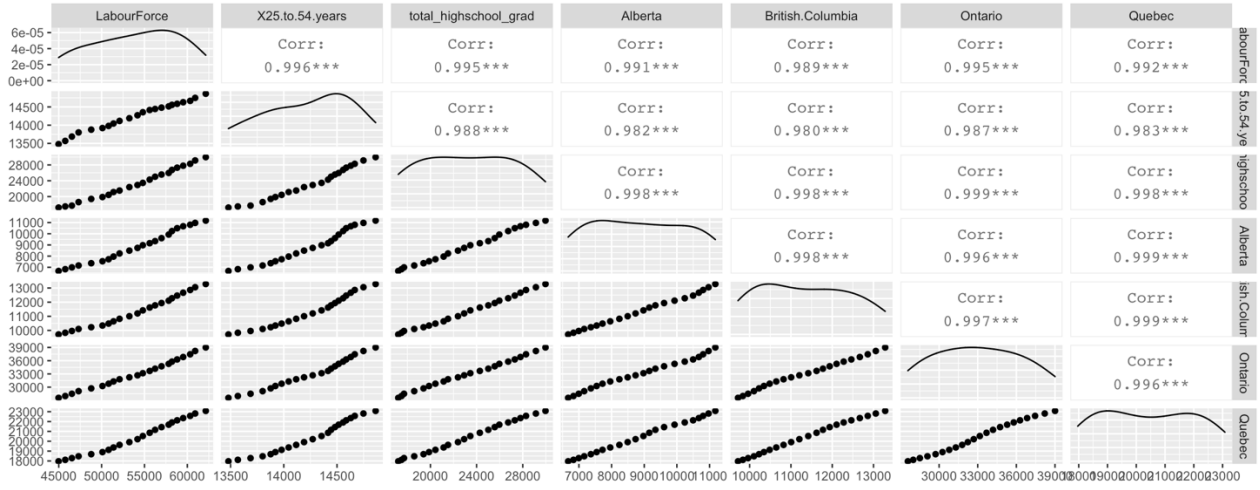
1. Variable correlation and possible interactions

```
library(GGally)
ggpairs(data_keep)
```



Comment: So far our variables have decent distributions, each variable has some relationships (either positive and negative) with others. It is worth to notice that the population-related variables have a very strong linear relationship within each other. I suspect it can be due to the fact that they all come from one dataset.

The interactions that I need to work on further, they might be/be not abundant when placing together. Further result will be discussed in the final paper. I might start using Type 2 Sum of Square Approaches to investigate this matter.



2. Basic Model Information

```
basic_model = lm(formula = LabourForce ~ .,data_keep)
summary(basic_model)
AIC(basic_model)
```

Call:
lm(formula = LabourForce ~ ., data = data_keep)

Residuals:

Min	1Q	Median	3Q	Max
-123.007	-56.508	4.015	33.888	220.960

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-2.997e+04	1.181e+04	-2.537	0.034869 *
X25.to.54.years	1.230e+00	1.225e+00	1.005	0.344511
total_highschool_grad	-7.971e-01	5.341e-01	-1.493	0.173910
total_postsecondary	-2.023e-02	4.784e-01	-0.042	0.967307
Alberta	1.095e+00	5.727e-01	1.912	0.092254 .
British.Columbia	-7.649e+00	1.861e+00	-4.110	0.003391 **
Ontario	2.443e+00	3.668e-01	6.659	0.000159 ***
Quebec	3.993e+00	1.202e+00	3.323	0.010489 *
GDP_whole	-8.920e-02	4.396e-02	-2.029	0.076969 .
Diff_divorced	2.384e+01	9.111e+00	2.616	0.030833 *
ServiceProduct_change	7.464e-02	4.631e-02	1.612	0.145711
GoodProduct_change	1.299e-01	5.127e-02	2.533	0.035092 *
IndustrialProduct_change	-3.542e-02	2.323e-02	-1.524	0.165913
Construction_change	-1.810e-02	2.045e-02	-0.885	0.401969

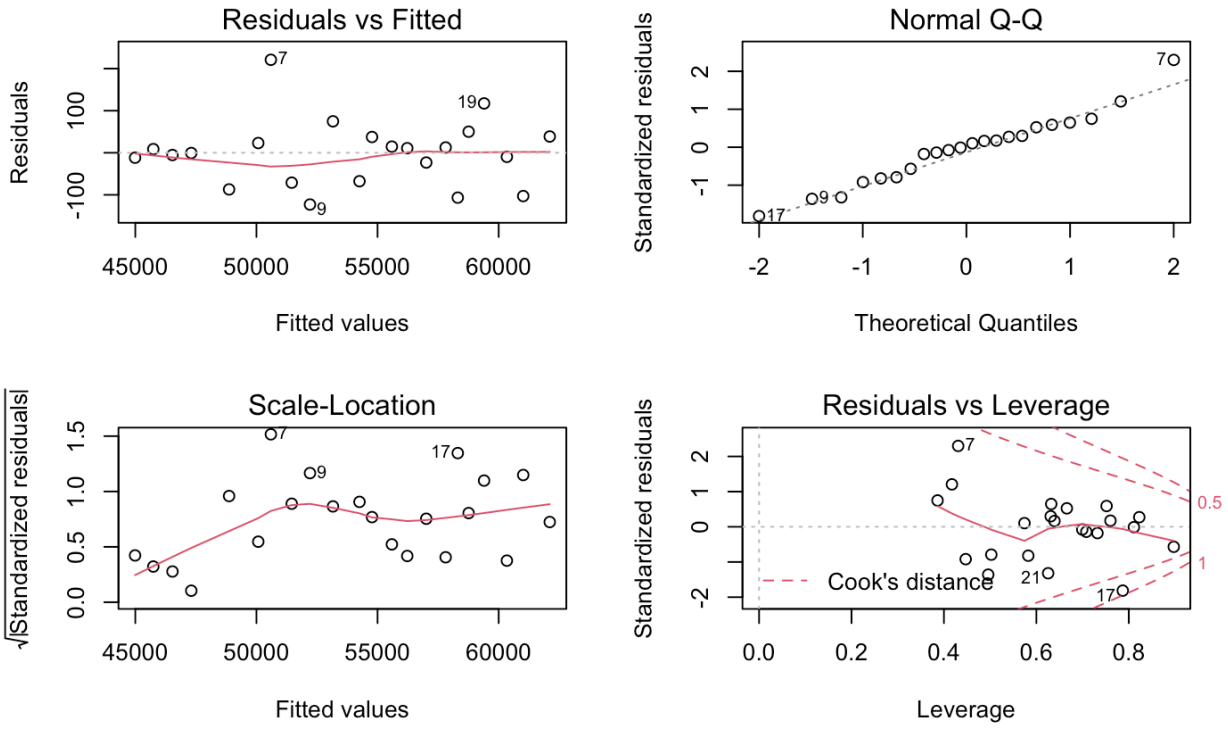
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127.4 on 8 degrees of freedom
Multiple R-squared: 0.9998, Adjusted R-squared: 0.9994
F-statistic: 2706 on 13 and 8 DF, p-value: 4.38e-13

The model is not too bad in estimating the labour force, we have R-Square is 99.98%, decent standard error in perspective to the values we have in this dataset. AIC for this current model is 283.4509. It is not too bad.

3. Diagnostic plot for Basic Model

```
par(mfrow = c(2, 2))
plot(basic_model)
#Cook Distance
plot(basic_model,4)
```



Comment:

- **Scale-Location:** This plot shows how residual variance varies with the predicted values. The red line shows a slight upward curve, but nothing particularly concerning. The variance appears to be approximately constant with respect to the fitted values.
- **Residual vs Fitted:** There does not appear to be any relationship between the residuals and the predicted values. Anything other than a flat line here would suggest a missing nonlinear relationship. Here we don't see that
- **Normal QQ**: This compares the ordered residuals with what they would be if they came from a normal distribution. Ideally this is a linear relationship with intercept 0 and slope 1. There is one point that is lower than a normal distribution would suggest, but the problem is not particularly concerning.
- **Residuals vs Leverage:** This plot shows observations which could have a large impact on the regression slopes. Highly influential outliers that impact slopes are considered high leverage. Those high leverage points would be outside of the dashed lines. There are no points outside of those points in this plot.

Stepwise selection from variables presented in all-in model

1. Defining the better model

```
stepmodel = step(lm(LabourForce ~ .,data_keep))
summary(stepmodel)
```

```
Step:  AIC=217.02
LabourForce ~ X25.to.54.years + total_highschool_grad + Alberta +
    British.Columbia + Ontario + Quebec + GDP_whole + Diff_divorced +
    ServiceProduct_change + GoodProduct_change + IndustrialProduct_change +
    Construction_change

Residuals:
     Min       1Q   Median       3Q      Max
-124.254  -56.812    3.888   34.530  220.467

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -3.019e+04  1.002e+04  -3.012 0.014665 *
X25.to.54.years           1.256e+00  1.004e+00   1.251 0.242349
total_highschool_grad    -8.155e-01  2.898e-01  -2.814 0.020253 *
Alberta                   1.087e+00  5.093e-01   2.134 0.061588 .
British.Columbia         -7.585e+00  1.014e+00  -7.483 3.76e-05 ***
Ontario                   2.437e+00  3.249e-01   7.502 3.69e-05 ***
Quebec                    3.958e+00  8.184e-01   4.837 0.000925 ***
GDP_whole                -8.851e-02  3.855e-02  -2.296 0.047327 *
Diff_divorced             2.354e+01  5.495e+00   4.284 0.002037 **
ServiceProduct_change     7.384e-02  3.984e-02   1.853 0.096816 .
GoodProduct_change        1.288e-01  4.233e-02   3.043 0.013949 *
IndustrialProduct_change -3.494e-02  1.914e-02  -1.826 0.101146
Construction_change      -1.784e-02  1.844e-02  -0.968 0.358411
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120.1 on 9 degrees of freedom
Multiple R-squared:  0.9998,      Adjusted R-squared:  0.9995
F-statistic:  3297 on 12 and 9 DF,  p-value: 7.273e-15
```

Comment: the only different between this model and the basic one is this model eliminate the “total_postsecondary” variable. This model has lower AIC than the basic model (basic model is 283.4509 and step model is 217.02), the standard error here is lower too (120.1 compared to 127.4 from the full model). Hence, there are enough evidences to claim our step model has better estimation than the basic one.

2. Interaction

We can test for interaction using:

```
all_interaction = lm(formula = LabourForce ~(X25.to.54.years +
total_highschool_grad + Ontario + GDP_whole + Diff_divorced)^2,data=
data_keep)
summary(all_interaction)
```

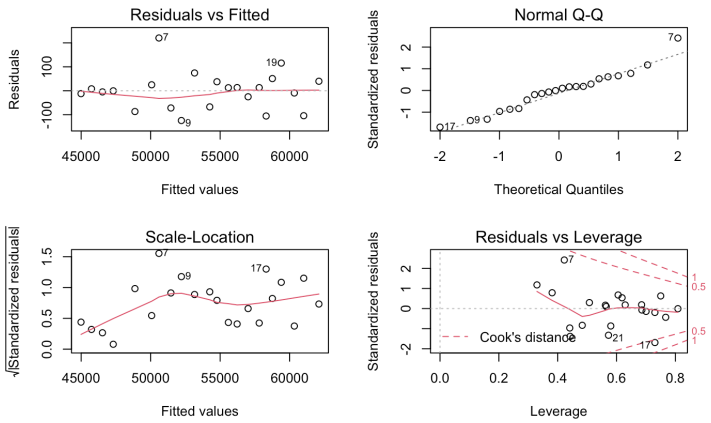
```
Call:
lm(formula = LabourForce ~ (X25.to.54.years + total_highschool_grad +
    Ontario + GDP_whole + Diff_divorced)^2, data = data_keep)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-244.96  -52.14  -10.67   24.17  386.01

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              2.929e+05  2.512e+05   1.166    0.288
X25.to.54.years                         -2.866e+01  2.059e+01  -1.392    0.213
total_highschool_grad                    1.042e+00  2.199e+01   0.047    0.964
Ontario                                 -6.482e+00  2.089e+01  -0.310    0.767
GDP_whole                               -2.425e-01  9.719e-01  -0.249    0.811
Diff_divorced                            1.554e+03  3.293e+03   0.472    0.654
X25.to.54.years:total_highschool_grad    1.478e-04  1.677e-03   0.088    0.933
X25.to.54.years:Ontario                  8.831e-04  1.486e-03   0.594    0.574
X25.to.54.years:GDP_whole               -1.353e-05  7.589e-05  -0.178    0.864
X25.to.54.years:Diff_divorced           -3.095e-02  2.349e-01  -0.132    0.899
total_highschool_grad:Ontario           -1.369e-04  8.701e-05  -1.573    0.167
total_highschool_grad:GDP_whole         -2.618e-05  2.074e-05  -1.262    0.254
total_highschool_grad:Diff_divorced      6.767e-02  7.846e-02   0.863    0.422
Ontario:GDP_whole                        3.180e-05  2.131e-05   1.492    0.186
Ontario:Diff_divorced                   -8.125e-02  9.242e-02  -0.879    0.413
GDP_whole:Diff_divorced                 -1.274e-04  4.286e-04  -0.297    0.776

Residual standard error: 219 on 6 degrees of freedom
Multiple R-squared:  0.9995,      Adjusted R-squared:  0.9982
F-statistic: 792.9 on 15 and 6 DF,  p-value: 1.291e-08
```

Comment: So far I haven't spotted any good result yet. As I said ealier, I would do type 2 error on the variables to yield the result.

3. Diagnostic plot

```
par(mfrow = c(2, 2))
plot(stepmodel)
#Cook Distance
plot(stepmodel,4)
```



Comment: not much different from the basic model.

Appendix

```
#import and minor format dataset

library(RColorBrewer)

library(tidyverse)
temp <- tempfile()
download.file("https://www150.statcan.gc.ca/n1/tbl/csv/14100118-
eng.zip",temp)
# download as a temporary file
(file_list <- as.character(unzip (temp, list = TRUE) $Name)) # unzip
the file
population <- read_csv(unz(temp, "14100118.csv"))
unlink(temp) # Delete temporary file

martial <-
read.csv("https://www.dropbox.com/s/6i09vk9t0r70msd/marriage_sheet_mod
ified.csv?dl=1")

temp <- tempfile()
download.file("https://www150.statcan.gc.ca/n1/tbl/csv/36100434-
eng.zip",temp)
# download as a temporary file
(file_list <- as.character(unzip(temp, list = TRUE)$Name)) # unzip the
file
GDP <- read_csv(unz(temp, "36100434.csv"))
unlink(temp) # Delete temporary file

#Basic process of each dataset

population <- population %>%
  drop_na(VALUE) %>%
  rename_all(make.names)

martial <- martial %>%
  rename_all(make.names)

martial <- martial %>% select(Year,Divorced,Married)
martial[1:3] <- lapply(martial[1:3], as.numeric)
martial$Divorced <- martial$Divorced/1000
martial$Married <- martial$Married/1000

GDP <- GDP %>%
  separate(REF_DATE, c("year", "month"), "-") %>%
```

```
  mutate(year = as.numeric(year))%>%
  drop_na(VALUE) %>%
  rename_all(make.names)

names(GDP)[names(GDP) ==
"North.American.Industry.Classification.System..NAICS."] <-
'Industry.Name'

# List unique value of every column for each dataset.

list_unique_value <- function(arr) {
  for (val in names(arr))
    {
    print(val)
    message(unique(arr[val]))
    #print(typeof(a))
    }
}

#list unique value of three datasets
list_unique_value(population%>%select(REF_DATE,GEO,Labour.force.charac
teristics,Educational.degree,Sex,Age.group))

list_unique_value(martial%>%select(REF_DATE,GEO,Type.of.marital.status
,Marital.status,Sex,Age.group))

list_unique_value(GDP%>%select(year,month,GEO,Prices,Industry.Name))



#provincial distribution of Canadian population
population%>% subset(GEO != "Canada" & Labour.force.characteristics ==
"Labour force" & Educational.degree == "Total, all education levels" &
Sex == "Both sexes") %>%
           group_by(REF_DATE, GEO) %>%
           summarize(total = sum(VALUE)) %>%
           ungroup() %>%
ggplot(aes(x = REF_DATE, y = total, group = GEO)) +
  labs(y = "Population (thounsand)") +
  geom_line(aes(colour = GEO)) +
  scale_color_brewer(palette = "Paired")+
  ggtitle("Total Labour Force, by year, seperated by Province")

selected_group <- c("15 to 24 years","25 to 54 years","55 to 64
years")
```

```
population%>% subset(Labour.force.characteristics == "Labour force" &
Age.group %in% selected_group ) %>%
           group_by(REF_DATE, Age.group) %>%
           summarize(total = sum(VALUE)) %>%
           ungroup() %>%
ggplot(aes(x = REF_DATE, y = total, group = Age.group)) +
  labs(y = "Population (thousand)") +
  geom_line(aes(colour = Age.group)) +
  scale_color_brewer(palette = "Paired")+
  ggtitle("Labour Force, by year, seperated by Age.group")

degree_group <- c("No degree, certificate or diploma","Postsecondary
certificate or diploma","With high school graduation")
population%>% subset(Labour.force.characteristics == "Labour force" &
Educational.degree %in% degree_group) %>%
           group_by(REF_DATE, Educational.degree) %>%
           summarize(total = sum(VALUE)) %>%
           ungroup() %>%
ggplot(aes(x = REF_DATE, y = total, group = Educational.degree)) +
  labs(y = "Population (thousand)") +
  geom_line(aes(colour = Educational.degree)) +
  scale_color_brewer(palette = "Paired")+
  ggtitle("Total Labour Force,by year, seperated by
Educational.holder")


population%>% subset(Labour.force.characteristics == "Labour force" &
Sex != "Both sexes" ) %>%
           group_by(REF_DATE, Sex) %>%
           summarize(total = sum(VALUE)) %>%
           ungroup() %>%
ggplot(aes(x = REF_DATE, y = total, group = Sex)) +
  labs(y = "Population (thounsand)") +
  geom_line(aes(colour = Sex)) +
  ggtitle("Total population,by year, seperated by Sex")


df <- martial %>%
  select(Year, Married_thousand,Divorce_thousand ) %>%
  gather(key = "variable", value = "Population_In_Thousand", -Year)
head(df)

ggplot(df, aes(x = Year, y = Population_In_Thousand)) +
  geom_line(aes(color = variable, linetype = variable)) +
  scale_color_manual(values = c("darkred", "steelblue"))
Essential_names <- c(
```

```
"Goods-producing industries [T002]",
"Service-producing industries [T003]",
"Industrial production [T010]",
"Non-durable manufacturing industries [T011]",
"Durable manufacturing industries [T012]",
"Agriculture, forestry, fishing and hunting [11]",
"Mining, quarrying, and oil and gas extraction [21]",
"Utilities [22]",
"Construction [23]",
"Public administration [91]",
"Federal government public administration [911]",
"Educational services [61]")

GDP %>% subset(GEO == "Canada" & month == "12" & Industry.Name %in%
Essential_names & Prices == "2012 constant prices" &
Seasonal.adjustment == "Seasonally adjusted at annual rates") %>%
  ggplot(aes(x = year, y = VALUE, group = Industry.Name)) +
  labs(y = "GDP Value (million)") +
  geom_line(aes(colour = Industry.Name)) +
  scale_color_brewer(palette = "Paired")+

  ggtitle("GDP per year, measured in end of the year December,per year
of selected essential industry")


population_total = population%>% subset(GEO == "Canada" &
Labour.force.characteristics == "Population" & Educational.degree ==
"Total, all education levels" & Sex == "Both sexes")

labour_force_total = population%>% subset(GEO == "Canada" &
Labour.force.characteristics == "Labour force" & Educational.degree ==
"Total, all education levels" & Sex == "Both sexes")

y_sample = labour_force_total %>% group_by(REF_DATE) %>%
  summarise(total=sum(VALUE))

# select the columns of interest
age_population_wide = population_total %>%
      select(REF_DATE, GEO, Age.group, VALUE) %>%
      pivot_wider(names_from = Age.group, values_from = VALUE) %>%
  # rebuild the variable names so that they do not have spaces
      rename_all(make.names)

sex_population_wide = population %>% subset(GEO == "Canada" &
Labour.force.characteristics == "Population" & Educational.degree ==
"Total, all education levels") %>%
```

```
      select(REF_DATE,GEO, Age.group,Sex, VALUE) %>%
      pivot_wider(names_from = Sex, values_from = VALUE) %>%
  # rebuild the variable names so that they do not have spaces
      rename_all(make.names)

sex_population_wide = sex_population_wide  %>% group_by(REF_DATE) %>%
  summarise(total_male=sum(Males),total_female=sum(Females))
glimpse(sex_population_wide)

degree_population_wide = population %>% subset(GEO == "Canada" &
Labour.force.characteristics == "Population" & Sex == "Both sexes")
%>%
      select(REF_DATE,GEO, Age.group,Educational.degree, VALUE) %>%
      pivot_wider(names_from = Educational.degree, values_from =
VALUE) %>%
  # rebuild the variable names so that they do not have spaces
      rename_all(make.names)

degree_population_wide = degree_population_wide  %>%
group_by(REF_DATE) %>%
  summarise(
    total_no_degree=sum(No.degree..certificate.or.diploma),
    total_highschool_grad=sum(With.high.school.graduation),
    total_postsecondary = sum(Postsecondary.certificate.or.diploma))
glimpse(degree_population_wide)



provincal_population_wide = population %>% subset(GEO != "Canada" &
Labour.force.characteristics == "Population" & Educational.degree ==
"Total, all education levels" & Sex == "Both sexes")   %>%
  group_by(REF_DATE, GEO) %>%
  summarize(total = sum(VALUE))


provincal_population_wide = provincal_population_wide %>%
select(REF_DATE,GEO,total) %>%
      pivot_wider(names_from = GEO, values_from = total) %>%
  # rebuild the variable names so that they do not have spaces
      rename_all(make.names)



library(dplyr)
pop_growth = population_total %>% group_by(REF_DATE) %>%
summarise(total=sum(VALUE))
```

```
pop_growth = pop_growth %>%
  mutate(Diff_pop_growth = total - lag(total),
         Rate_growth_percent = (Diff_pop_growth /total) * 100) #
growth rate in percent

martial_growth = martial %>%
  mutate(Diff_married = Married - lag(Married),
         Rate_married_change = (Diff_married /Married)*100,
         Diff_divorced = Divorced - lag(Divorced),
         Rate_divorced_change = (Diff_divorced /Divorced) * 100)


GDP_essential = GDP %>% subset(GEO == "Canada" & month == "12" &
Industry.Name %in% Essential_names & Prices == "2012 constant prices"
& Seasonal.adjustment == "Seasonally adjusted at annual rates")

GDP_total = GDP %>% subset(GEO == "Canada" & month == "12" &
Industry.Name == "All industries [T001]" & Prices == "2012 constant
prices" & Seasonal.adjustment == "Seasonally adjusted at annual
rates")

GDP_whole_growth = GDP_total %>% select(year,Industry.Name,VALUE) %>%
  mutate(GDP_whole = VALUE - lag(VALUE),
         Rate_GDP_change = (GDP_whole /VALUE)*100)

GDP_industry_rate = GDP_essential %>% select(year,Industry.Name,VALUE)
%>%
  pivot_wider(names_from = Industry.Name, values_from = VALUE)

colname <-
c("REF_DATE","GoodProduct","ServiceProduct","IndustrialProduct","NonDu
rableManu","DurableManu","Arigiculutre","Mining","Utilities","Construc
tion","Educational","PublicAdmin","Federal")

GDP_industry_rate <- GDP_industry_rate %>%
  mutate(GoodProduct_change = GoodProduct - lag(GoodProduct),
        GoodProduct_rate = ((GoodProduct_change / GoodProduct)*100),
         ServiceProduct_change = ServiceProduct - lag(ServiceProduct),
         ServiceProduct_rate = ((ServiceProduct_change
/ServiceProduct)*100),
         IndustrialProduct_change = IndustrialProduct -
lag(IndustrialProduct),
         IndustrialProduct_rate = ((IndustrialProduct_change
/IndustrialProduct)*100),
         NonDurableManu_change = NonDurableManu - lag(NonDurableManu),
```

```
        NonDurableManu_rate = ((NonDurableManu_change
/NonDurableManu)*100),
        DurableManu_change = DurableManu - lag(DurableManu),
        DurableManu_rate = ((DurableManu_change /DurableManu)*100),
        Arigiculutre_change = Arigiculutre - lag(Arigiculutre),
        Arigiculutre_rate = ((Arigiculutre_change /Arigiculutre)*100),
        Mining_change = Mining - lag(Arigiculutre),
        Mining_rate = ((Mining_change /Mining)*100),
        Utilities_change = Utilities - lag(Utilities),
        Utilities_rate = ((Utilities_change /Utilities)*100),
        Construction_change = Construction - lag(Construction),
        Construction_rate = ((Construction_change /Construction)*100),
        Educational_change = Educational - lag(Educational),
        Educational_rate = ((Educational_change /Educational)*100),
        PublicAdmin_change = PublicAdmin - lag(PublicAdmin),
        PublicAdmin_rate = ((PublicAdmin_change /PublicAdmin)*100),
        Federal_change = Federal - lag(Federal),
        Federal_rate = ((Federal_change /Federal)*100))


names(GDP_industry_rate) <- colname

names(martial_growth)[names(martial_growth) == "Year"] <- "REF_DATE"
names(GDP_whole_growth)[names(GDP_whole_growth) == "year"] <-
"REF_DATE"
names(y_sample)[names(y_sample) == "total"] <- "LabourForce"

mergeCols <- c("REF_DATE")
test <- merge(age_population_wide, degree_population_wide,by =
mergeCols)
test <- merge(test, provincal_population_wide,by = mergeCols)
test <- merge(test, sex_population_wide,by = mergeCols)
test <- merge(test, pop_growth,by = mergeCols)
test <- merge(test, martial_growth,by = mergeCols)
test <- merge(test, GDP_whole_growth,by = mergeCols)
test <- merge(test, GDP_industry_rate,by = mergeCols)
test <- merge(test, y_sample,by = mergeCols)

uselessCol =
c("GEO","total","Divorced","Married","Industry.Name","VALUE","GoodProd
uct","ServiceProduct","IndustrialProduct","NonDurableManu","DurableMan
u","Arigiculutre","Mining","Utilities","Construction","Educational","P
ublicAdmin","Federal")
names.use <- names(test)[!(names(test) %in% uselessCol)]
final_dataset <- test[, names.use]
```

```
martial_g = merge(martial_growth,pop_growth,by = mergeCols)
martial_g = merge(martial_g,GDP_whole_growth,by = mergeCols)
martial_g = merge(martial_g,y_sample,by = mergeCols)

martial_g <- martial_g %>% mutate(labor_change = LabourForce -
lag(LabourForce),
                                  labor_rate =
(labor_change/LabourForce)*100)


glimpse(martial_g)

library(ggplot2)
colors <- c("Rate_married_change" = "blue", "Rate_divorced_change" =
"red", "labor_rate" =
"orange","Rate_growth_percent"="green","Rate_GDP_change" = "purple")

ggplot(martial_g, aes(x = REF_DATE)) +
    geom_line(aes(y = Rate_married_change, color =
"Rate_married_change"), size = 1.0) +
    geom_line(aes(y = Rate_divorced_change, color =
"Rate_divorced_change"), size = 1.0) +
  geom_line(aes(y = Rate_growth_percent, color =
"Rate_growth_percent"), size = 1.0) +
   geom_line(aes(y = Rate_GDP_change, color = "Rate_GDP_change"), size
= 1.0) +
    geom_line(aes(y = labor_rate, color = "labor_rate"), size = 1.0) +
    labs(x = "Year",
         y = "(%)",
         color = "Legend") +
    scale_color_manual(values = colors)

write.csv(final_dataset,'final_dataset.csv')
```