目录

[Part 1 Situation Understanding 2](#_Toc527389343)

[1. An understanding of situations for adequate employment and economic growth 2](#_Toc527389344)

[*2.* *Situation Assessment* 2](#_Toc527389345)

[*3.* *Business Objectives* 4](#_Toc527389346)

[*4.* *Project plan* 5](#_Toc527389347)

[*5.* *Data mining objectives* 5](#_Toc527389348)

[*6.* *Project plan table* 6](#_Toc527389349)

[Part 2 Data understanding 6](#_Toc527389350)

[*1.* *Collect and describe data* 8](#_Toc527389351)

[*2.* *Explore the data* 13](#_Toc527389352)

[*3.* *Verify the data quality* 16](#_Toc527389353)

[Part 3 Data preparation 18](#_Toc527389354)

[*1.* *Select the Data* 18](#_Toc527389355)

[*2.* *Clean the Data* 21](#_Toc527389356)

[*3.* *Construct and Format the Data* 25](#_Toc527389357)

[*4.* *Integrate Various Data Sets* 28](#_Toc527389358)

[Part 4 Data transformation 32](#_Toc527389359)

[*1.* *Reduce the data* 33](#_Toc527389360)

[*2.* *Data Project* 34](#_Toc527389361)

[Part 5 Data-Mining method selection 38](#_Toc527389362)

[Part 6 Data-Mining algorithm (s) selection 39](#_Toc527389363)

[Part 7 Data Mining 47](#_Toc527389364)

[1. Logical test and design 47](#_Toc527389365)

[2. Data Mining Result Interpretation 48](#_Toc527389366)

[3. Pattern searching 52](#_Toc527389367)

[Part 8 Interpretation 55](#_Toc527389368)

[Iteration Improvement 57](#_Toc527389369)

**Can adequate employment promote economic growth for the left behinds?**

*The software applied in iteration 4 is Pyspark.*

# Part 1 Situation Understanding

## An understanding of situations for adequate employment and economic growth

As the authoritative branch in the United Nations, the World Bank has been long working for well-being across nations and races, promoting the development of the left-behinds and enhance the sustainable development for the advanced economy.

Among 17th sustainable goals published by the World Bank, a decent work and economic growth could be the foundation of a sustainable growth in economy. Our intuition also suggests that there might be somehow a relationship between employment and the economic growth.

Nevertheless, the journey to the developed world may not be plain sailing. The world bank claim that although many least developed economics have developed in last decade, but few of them have achieved the 7% annual growth rate for sustainable development[[1]](#footnote-1). (Note here, the growth rate may be impressive compare to the low digits increase for the developed world. But meanwhile, the developing world may suffer from a high interest of their borrowed money together with a high inflation rate in currency. If the left behind grows in a mild rate way below 7%, they could sacrifice their future.)

## Situation Assessment

Just as the world bank suggests, the economic growth of those left behinds may impede by less productive employmees and high unemployment rate[[2]](#footnote-2). The remedy could be improving the knowledge and skill of labour force, making schools and college much cheaper and enjoyable for the young people and providing more jobs in the those least developed countries. This approach could, in the long run, upgrade the industry of the left behind and providing decent works as well as boosting the efficiency and the competitiveness of those economies.

**Available resources and requirement indexes:**

Our first impression of lower than expected issues could help us to narrow down our research questions and indexes.

Literatures have long claims that a sustainable economic growth could be vital for the left-behinds (developing countries) to catch up. In 1956, Robert Solow has discovered two external variables which may result in different equilibrium in economics. These two variables are savings and the population. He found a convincing relation between the population growth and the economic growth. The higher the population growth, the poor the country would be. In other words, the long run equilibrium economic-growth rate would be lower if the population growth-rate is high.

However, the saving index may be impractical in reality because only part of the money in the bank are used for reproduction in the economy. To make things more practical, we will use FDI, net financial account and gross capital formation indexes to replace the saving ratio in the economy. These three indexes are directly link to the reproduction in economy.

Based on the early 1950s Solow model, Mankiw, Romer, and Weil[[3]](#footnote-3) [1992] has adopted a new external variable: human factor into the model. And they prove the relation that the higher the human resource and capital accumulated, the faster the economy is expected to grow with good data support. Also, they give prediction for the convergence of the economic growth rate when growth rate and capital growth is constant. Many papers have focused on these issues even though the original theory is quite old but still work well with the prediction matches. And many new indexes has added into the model, which include R&D by Nunneman and Vanhout [1996][[4]](#footnote-4) and health and longevity introduced by Knowles and Owen [1995][[5]](#footnote-5). Case study in relevant to our report is Theodore R’s work [2004][[6]](#footnote-6) on institution and education to overcome poverty, etc.

Nowadays, the Solow model has become a common believes among economists that a high employment rate is the crucial requirement to obtain a rapid growth in economy.

Our research range would contain independent variables such as labour force, unemployment rate, the distribution of age & education level within the group and the technology involved in production, etc, and the response variable is the growth of economy.

These indexes carry out in large-scale and require continuous recording. Also, across countries data over decades deal with joining database with different standard. All these complexity demands our data should come from NGOs which are a highly authoritative and non-biased and non-Political driven organizations. And World Bank Database would be the best source to carry out this research.

**Assumption:**

To be specific, the most famous rules to reveal the relation between unemployment and economic growth is the Solow model. Nevertheless, this model has only a short validation, which means a weak and short term of relation could be explained by traditional Solow Model. Also, the traditional model has limited and restricted variables, risking at underestimate the real complex underlying relations.

Mankiw, Romer, and Weil [1992] have adopted an improved model called augmented Solow model (denote as ASM), encompassing both accumulation of human and physical capital. The ASM provides better explanation of cross-country data over OECD countries. In this project, we are interested at revealing whether this model could be applied to the least developed counties.

**Constrains and contingency:**

The overall story could be sounding and encouraging, all of these models is studied in “developed world”, such as the US and Europe. It is arguable whether this could also work for the developing world.

The real situations could be quite controversy. Many of these developing countries might have a much higher unemployment rate with remarkable growth rate. The only problem is the number of growths is less than 7%. One without economic background could argue this growth rate is higher than those developed country and this could be a propaganda for the world bank’s interest. But in fact, all those developed countries had experienced a vast economic growth in history and gradually converged approximate to 2%. This is predicted by endogenous model because the revenue-return decreases as per capital increase in production.

If a country sits long below 7%, the potential ceiling for them could be quite low. And eventually they will never be in the club of developed world and the poor growth would be furtherly eaten up by a rapid population growth. In this sense, individuals in these countries may not benefit from the economic growth.

Our guess is that such tragedy could be simply caused by the greedy money borrower setting a too high interests, which is over the range of self-generate of most business. Or alternatively, these countries have lost their advantage of cheap labour and young aged workers in the global competitions because their workers are low skilled or there are many surplus in labour markets among poor countries.

One fact we know is the low skill and bad facilities in least developed countries made young labours (aged 15 to 24[[7]](#footnote-7)) struggle for works. Globalization is supposed to bring opportunities and investment to them, but in reality, they often ended in low-productivity and low-quality job[[8]](#footnote-8).

Despite the cruel reality, the refined model requires available data across countries. But many of those poor countries could have little or no access for the human capital development and bad record for those data as well. This could bring a failure to apply a model requiring more variables. Although the result might be implicit, we could make a good guess of what is available and adjust our indexes we apply in our model for a better and reasonable fit.

## Business Objectives

*Does low employment in the least developed countries impede a sustainable growth rate above 7%? If this is true, in what extend it impede the economic growth? In other word, we need to find out the marginal impact per variables on the economic growths.*

*If things have improved, how many percent of economic growth we could expected with one percent increase of those explanatory variables? And how much improve we need to achieve a growth rate above 7%?*

*Nonetheless, these digits just serve for the development goals. In real-life scenario, we need to focus on the most important variables, and start to fix it in order to solve the problem. For each targeted variable, it is also vital to know what may cause such impede for each variable. And this conclusion could be a guidance government to set up policies to tackle with the problems.*

## Project plan

In this research paper, we suspect the employment rate strongly related to the economic growth. But as many macroeconomic cases, there might be numbers underlying of variables which may not have direct relation but could be revealed through analysis. To determining every related problem could be quite difficult before running through the data, we instead adapt a guess and test methodology by applying the simplest model and refine it step by step, until bringing up an accurate model to explain what really impede the economic growth in the least developed countries.

My research plan is to brought up simple linear regress first, where the employment rate is the only explanatory variables and the economic growth is the response variable. Not surprisingly, this simple model might be poorly predicted the overall variation of the gross growth. Then, we adapt different economical growth model to add new variables to increase the explanatory power of out model. Eventually, our final model may also not be linear form. One example could be the classic Solow growth model which follows a power law relation (, where K represents for the capital in production, A stands for the knowledge (or technology) in production and L is the labour force measured by employment rate here). However, as the complexity of real cases, the gross labour number might change due to the change of the population. So, the L (labour number available) need to be adjusted with the changes of population and the employment.

Moreover, the Solow model is the product of production function model which imply the Inada condition meaning the elasticity (1st degree of derivatives) of substitution close to infinity at the beginning and convergence to 0 when x is large. But this also could be unlikely in real case because of the business cycle or external factors for the least developed countries. Details of model modification could be discussed in detail later in this report.

## Data mining objectives

Data mining is the method serving for the business objectives. The data mining objective is to find the simplest accurate model (Occam’s razor) to explain the reasons which leads to the changes of economic growth. By extracting valuable data, the most appropriate model could be built through the try and refine iterations. Eventually, we could expect the final model to well explain the business objectives, which is the reason why the least developed countries face lower economic growth rates. Later on, these knowledges could conduct the action to improve the situations in the least developed countries.

## Project plan table

Figure 1-1 Project Plan

# Part 2 Data understanding

Before start Part2 I have something to complain for the inconvenience in Pyspark.

*There are three issues with my AWS Juypter notebook.*

*First is the upload button not work. I try many times with my console running at different place, but the button simply not working. I suppose the bottom coding is not redirect to the correct direction. Thanks to Mr. Diego, and I uploaded the all my data sets at GitHub and pull all of my data sets with git pull command.*

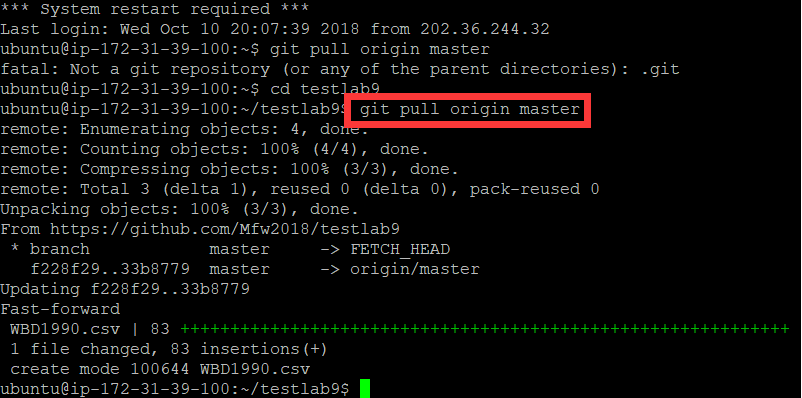
**

Figure 2-1 Git Pull Command Show

*Second is the input function is are not auto-adjusted. When I firstly input data into Pyspark, the auto detection is not functioning since every variable is converted to String which is actually numeric except the first 4 columns. I have checked the external material. There is a solution to assign a special type to each of the columns. But it still not working because some of the content contains some regular expression code. The World bank data are construct in R programming. And R supports object-oriented programming S4, which means the object name are taken as vector automatically, allowing something like “$, -, (), |, etc…”. Unfortunately, the pyspark just can’t recognized it even with inferSchema. Even though I use R to process the data and change the format, I still need to put a great effort to resubstitute and make every data sets coherent in Series Name column over 5000 rows because my data are hierarchy structure or aka. Long type. This procedure consumes most of my time in writing and that’s why I ask for this extension mainly. After all the non-word name converted to word characters, I can to transfer to wide form. I tried to input pandas’ packages but the same code for pandas report errors and the error report is not helpful because the exceptional throws do not tell me where is the error locating in my data sets.*

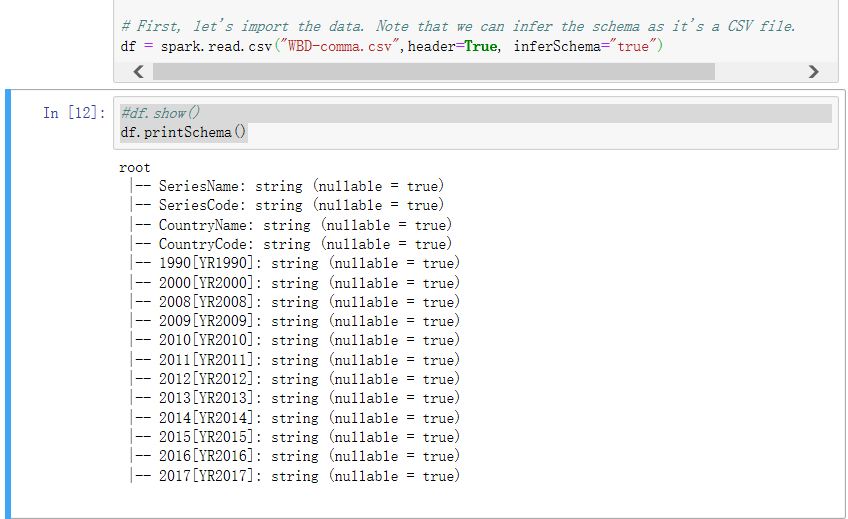


Figure 2-2 Wrong Auto-classification of Variables

Originally, all the variables start with year should be float or double and the four first columns are the strings. But this is not meaningful because this is a long form. The detailed format classification is related to the wide form later in section 2.

*Alternatively, I used R in my iteration 3 and I apply the same software in iteration 4 for data reshape because the R just process the datasets much faster with all the datasets listed in the studio with high toleration which is easy to change and observe. Also, the error report and instruction books are more helpful and in details.*

*The third issue is that we only have python 3 available in AWS. Since I found the API for R on* [*http://spark.apache.org/*](http://spark.apache.org/)*, but the pre-setting of environment is quite complex and at a risk not running on markers AWS console, I have to run the data processing part in R studio on my laptop and after I transfer the long type table to wide table, I uploaded to GitHub and pull from GitHub into AWS Jupiter.*

*The following data processing part is carrying out in R code but the entire R file is uploaded to Jupiter. Anyone who mark this file please run the R script in Rstudio with correct working directory setting (containing my original data sets WBD.csv), the entire code should work without any error reported (I have been tested without a bug before turning in).*

## Collect and describe data

In this report, all of the initial data sets are derived from the country grouped streams in World Bank Data Base[[9]](#footnote-9). The data includes the 81 countries with low income and lower middle income[[10]](#footnote-10) and 68 detailed variables mainly focus in the following fields from 2008 to 2017 (plus 1990 and 2000 as two baselines):

**Response Variables**

* **Economic growth rate annually**, numeric type

(correspond to **double type** in pyspark, shown in red square)

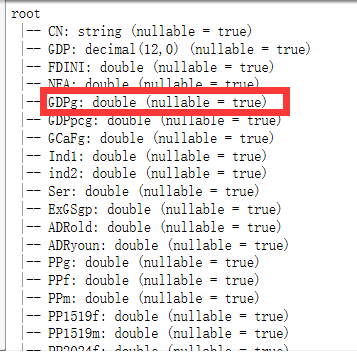


Figure 2-3a List of variables

Note here:

1) The variable name in the programming only takes for abbreviations because pyspark cannot auto-adjust the width of the print. If the name is long, everything would be a mass.

2) Numeric type is the type in R. Originally, the world bank data base process data using R. That is the reason why I suggested R because the process is much faster with great tolerance. Pyspark has lots of problems because it is built on Java API. The intermediate java compiler would make things much more complicated than R originated from s-plus language.

**Explanatory Variables** (possible modified in later iteration, but the general explanatory power of variables would only be stronger not weaker through iterations)

* **Employment rate** (adjusted by population) double in pyspark, named Labtot[[11]](#footnote-11)
* **Population** (or labours in the market) double in pyspark, named PPg
* **Age11** (used for identifying the change of structure in labour forces) double in pyspark
* **Investment** (Capital) double in pyspark, named GCaFg
* **Technology** (R&D) double in pyspark, dropped in figure 2-3, too many missing values

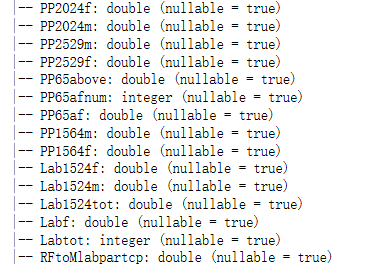


Figure 2-3b List of variables(continues)

The variable is originally constructed in partial hierarchical structure (aka. Long-format):

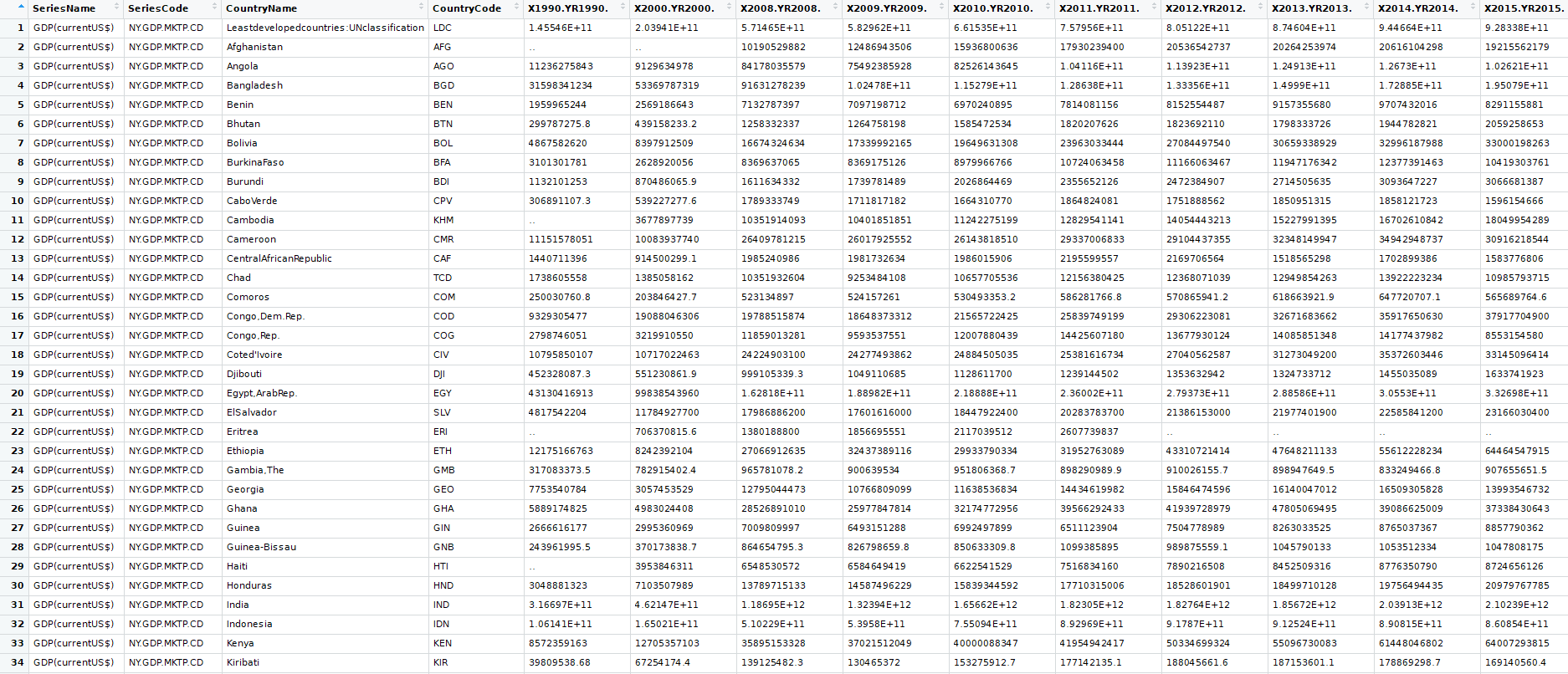


Figure 2-4 Long format table heads

This a long table shown above has 14 columns and 5581 rows. The problem arises when we only have years as columns and all the variables merged into one column called “Series Name” show in the left most column. Although it is possible to analysis row-wisely, but every function has to redefined as row-wisely and this could be low efficient in Jupiter notebook. As a remedy, we transfer to this long-format into wide-format data. This can be done by treated the content in Series Names as factor in data frame (automatically transfer to factor by smart recognition function as default) and then we reshape this data frame by factors in series name.

#read in data

[1] WBD=read.csv("WBD-comma.csv", header = TRUE)

#remove the replicated or useless columns

[2] WBD=WBD[, c(-2, -4)]

#rename the variable name

[3] colnames(WBD)= c("Variable", "countryName","1990","2000","2008", "2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2017")

#explore the datasets

[4] summary(WBD)

#select the new df separated by each year column

[5] WBD1990 = WBD[,1:3]

#reshape the data frame into wide format

[6] WBD1990w = reshape(WBD1990, idvar = "countryName", timevar = "Variable",direction = "wide")

#remove the quote and blanks from the data region (quote generated automatically in output)

[7] WBD1990w = WBD1990w[1:82,1:69]

#change the weird dot by the null content

[8] WBD1990w[WBD1990w==".."] = ""

#remove some weird 0 generated by World Bank database system

[9] index = which(WBD1990w[,11]==""|WBD1990w[,12]=="")

[10] WBD1990w[index, 10] = ""

#export the csv file

[11] write.table(WBD1990w, file = "WBD1990.csv", row.names = FALSE, col.names = TRUE, na = "", sep=",")

Highlight here, this code is not only just transforming the shape, but also did a little more thing for matching the stiff pyspark format:

Line 2: extra variables as series code and country code is removed due to replication.

Line 3: rename the colnames in WBD because the original names shown in the screen shot could be quite misleading and the year is repeated serval times in the source table.

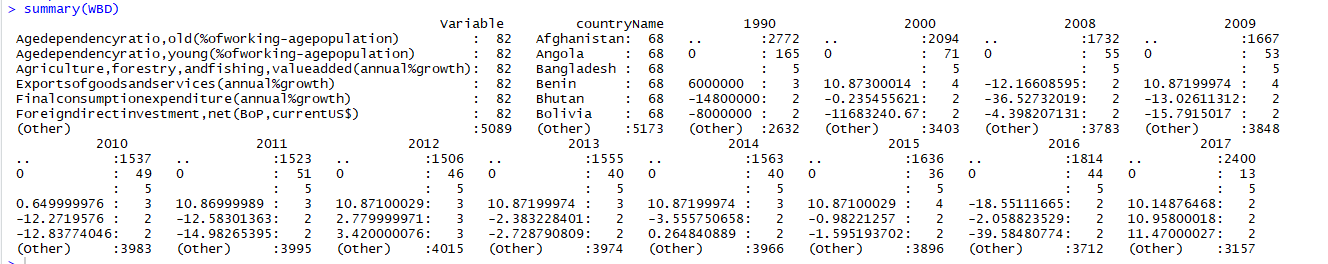
Line 4: inspect the original long-format table. The same procedure will carry out later, here is just to see how many weird values in each column. The weird content here is “..”, 0 (may be 

Figure 2-5 Summary by Columns

generated by data base automatically), and “” (sign for missing values, will be handled in the data cleaning section).

Line 5: since we have hierarchy data structure and our data frame is normally two dimensional, we have to divide the original data into different table each within a certain year. Since the third column is the year 1990’s data and the first two columns are variables and countries names, we only extract 1st to 3rd columns from the original table and reshape to country vs variables in 1990.

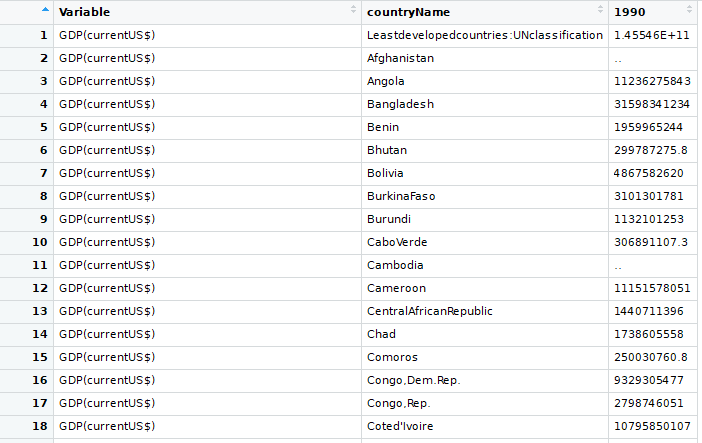


Figure 2-6 Long Format of WBD (name of my data sets) Datasets

The long form of the data before reshape.

Line 6: the following figure shows the wide-form after transformation.



Figure 2-7 Wide Format of WBD Datasets (separated by year)

Line 8: although we will apply clean process in section 3, the weird “..” values would actually stop the columns be recognized as numeric in pyspark. So, we substitute “..” with ”” a true null values.

Line 9 and 10: while exploring data (distribution of data will be shown in graph in next section), I found GDP growth rate is 0 while the annual GDP data is actually missing for some countries. This 0 might be generated automatically by the world bank database. But we could not remove this directly from the data sets because 0 could be meaningful if the GDP is actually the same if two GDP data present or as an initial year. And other columns as net investment could also be 0 showing no investment present in certain years. So, we apply the logical statement that if the following columns is null (which is GDP value), then replace the growth rate with “” because an GDP growth rate cannot be present when there is no data of GDP. After all the process, the final trimmed-nice-wide-form is showed in figure below:

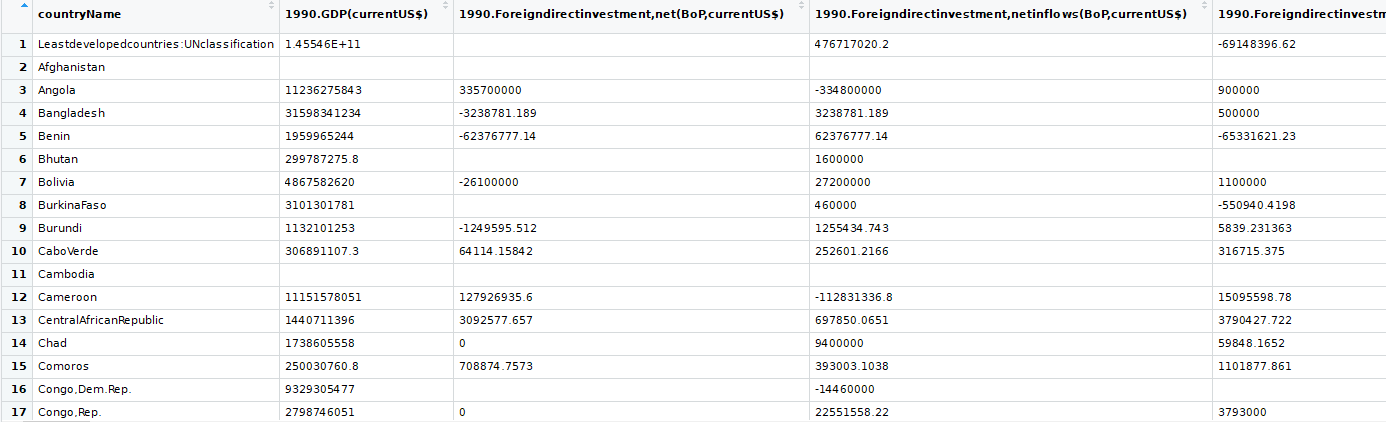


Figure 2-7 Wide Format of WBD Datasets (after removing weird values)

The dimension of this new table has:

82 rows(countries) and 68 columns (variables).

The structure of the data frame is shown as below:

Index (Variables)

GDP

Capital

Unemployment

Country Name

Values separated by years

Long format

Wide format

Country Name

GDP

Capital

Capital

Years: 1990~2017

Figure 2-8 Relational View of Data Structure

By far, we get the wide form of data sets in year 1990. For simplicity reason, we are not showing the rest years because they are all similar.

After transform to the trimmed data into pyspark, the type matching seems much better (only the head of variables is shown).

Notice there is still some type coerce to string type (we only accept the Country Name as string). By inspection, this is because all the content in the column is empty. If we coerce to double here, 0 problem may be generated. We will handle this in data cleaning process.

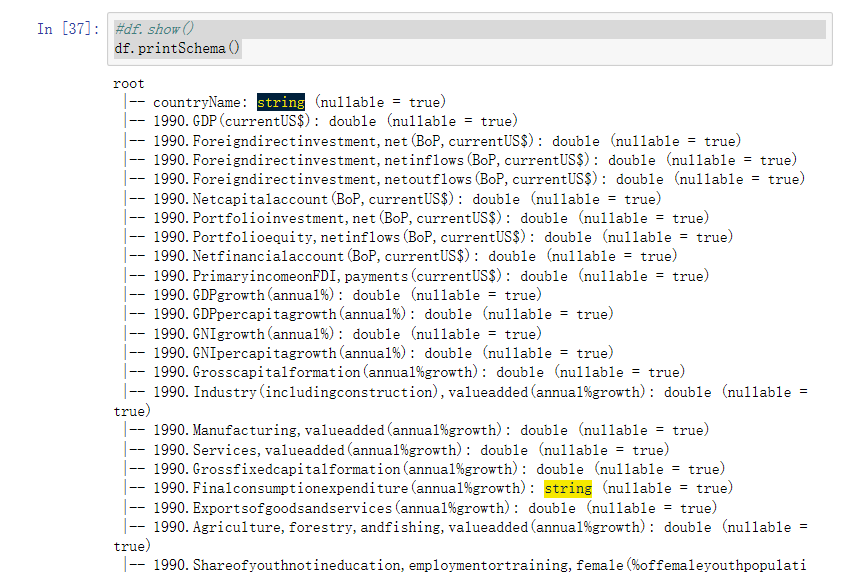


Figure 2-9 Schema View of Original Datasets

## Explore the data

In individual year’s sheet, the heads of the table seem okay but there are still some missing values. We will handle this in data cleaning section. There are more than 30 variables in the table and there is only part of them may relevant to our research questions. We select 8 variables just for exploration (the regression part will include more variables through iterations. The second table in 2-10 is the basic statistics for those 8 variables.

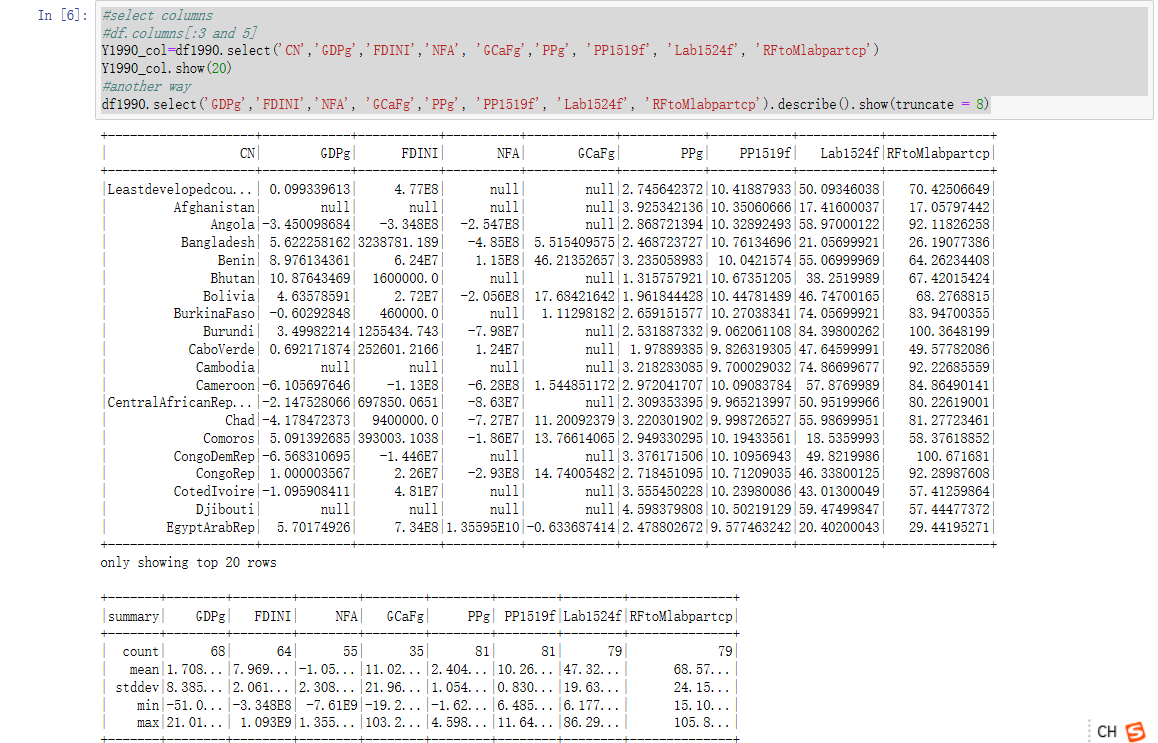
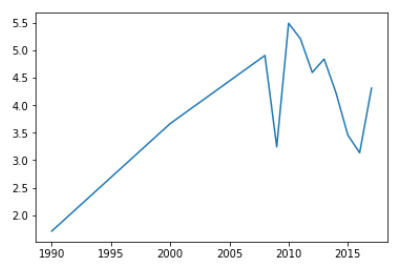


Figure 2-10 the heads of the data sets and basic statistics

By taking the mean of each variables in the table, we can plot the growth rate of the economy over the years



2008

GDP Growth Rate

Figure 2-11 the average GDP growth rate for least developed countries

According the world bank’s classification, the least development countries has experienced a significant growth in average from 1990. The GDP growth rate peaks in 2010, at the rate of 5.49%. These countries had not reached the 7% target in average. This result indicating there must be something impede the economic development or simply the goal sets too high.

But things getting worse if we measure the GDP per capita:



Figure 2-12 the average GDP growth rate per capita for least developed countries

This showing the population growth in these countries actually eaten up the gross increments. the average increase in GDP per capita fall down below 3 percent and further down less than 2% in the last two years. This data is not adjusted by the increment of CPI, overall, there might be no actual increase for the purchase power per person in these countries, just as if in the Malthus trap.

Surprisingly, the overall unemployment rate in these countries is not high as we expected, but the young people seem to suffer more in the labour market:

Figure 2-13 The Employment Rate in Least Developed Countries

And there is no obvious pattern between the employment rate and the economic growth. The simple model seems not work on the macro level. Maybe there is a huge difference between counties.

Figure 2-14 Unemployment (Youth total) VS GDP Per Capita Growth in Least Developed Countries

We expected there might be some age group people in the labor force in relate to the years which jitters most that affect the overall economic growth. Through data inspection, we find a common feature that the young people dominate the labor market. So, there might be a high unemployment rate for these people because they are not only the majority, but also low skills due to missing adequate education at early age. Therefore, these young labors are vulnerable in the global competitions and results in a high unemployment rate just all above 10%. This unemployment level can be considered as high. Therefore, we suspect the reason might be early working age sacrifices the human capital in the long run for these countries. Although all the years have lower than expected economic growth rate and high unemployment rate, there seems to be no significant relations between the GDP per capita and the unemployment rate. A further analysis is certainly needed.

## Verify the data quality

In general, as an authoritative international financial institution, the World Bank data could be a reliable source to investigate the left-behinds. But this is not always the case. Although the procedure of a rigid data collection has been secured, the table overall still might be biased due to the missing values present by various of limitation in reality. This missing data presents in all of these 68 variables. In the transposed long table, among all the 66972 lines of the table, there are more than 20000 missing values, especially for the R&D investment and technicians’ number, etc, which are not available in most of the low-income nations. This can be caused by a laggard statistical system, resulting missing record in the early year data across developing countries. If we count the missing values by cell, the odds of missing are one thirds.

The errors problem is also present as some weird value present in the data sets. Some of them may mean no values but printed as non-standard character not recognized by the pyspark, or some converting problems for non-UTF-8 code. I all substitute the errors by correct value and ensure everything works with pyspark. But in general, the data has been well depicted and well maintained in world bank databases. My work is more to make the required data in a consistent format.

In this case, supposing we apply a complete case analysis (which is removing all the observation of missing values), the result could be with only few observations remains. So, this missing value method is not acceptable.

To fix this problem, we could have two approaches:

The first is use linear approximation method. This indicates filling missing values with the linear function or the moving averages based on the value of non-missing cells next to it. Also, this method has limitations. If we apply this linear function to an all missing rows, there would be still NAs after applying the method.

Then, we can apply the second method by available-case analysis. This method is imbedded within R as default when calling functions (quite convenient).

Number of missing values before process: (False: value is not missing, true: value is missing, sorted by years)

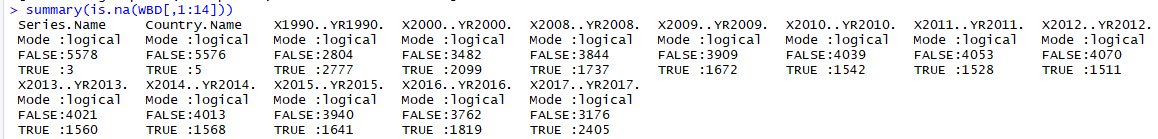


Figure 2-15 the Summary for Missing Values before Processing

The missing of country name is actually some blank rows in the data sets, will be removed during the process.

Number of missing values after process:

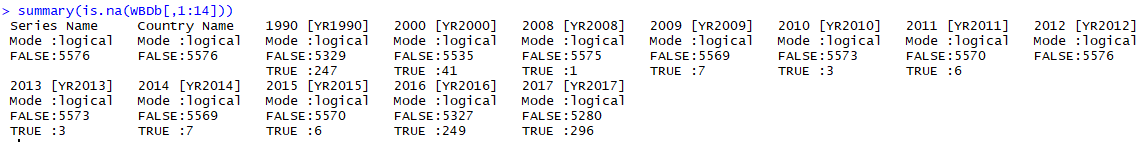


Figure 2-16 the Summary for Missing Values after Processing

All the missing values in the series names (names of variables, such as “GDP”, “unemployment rate”, etc) and country names has been removed. The year sorted groups also have significant reduction from the missing values (1990: the number of missing values is deduced from 2777 to 247, …).

Also inspecting each missing value, the individual substitution seems okay because each filling consider column wise anuual effect together with row wise moving average.

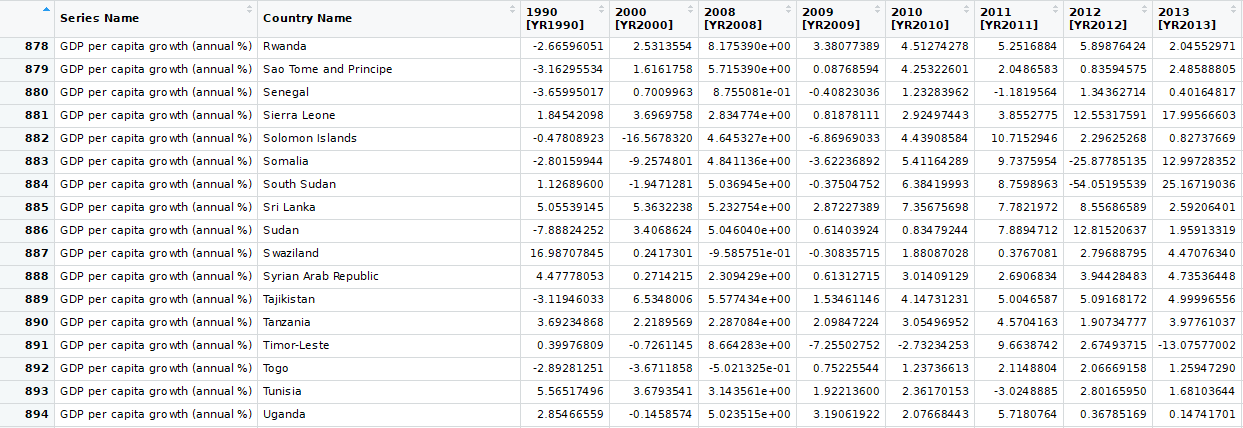


Figure 2-17 the Table View after Filling

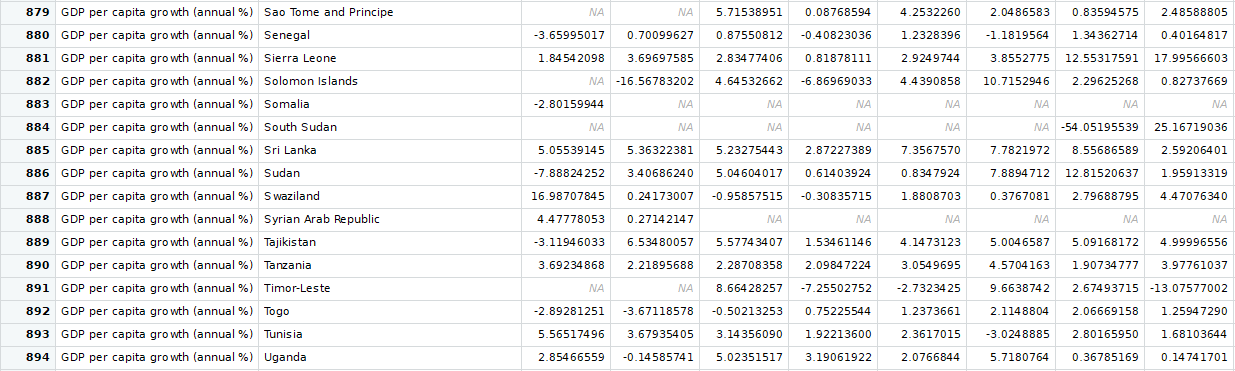


Figure 2-18 the Table View before Filling

The reason why we stated bit of data cleaning issues is this quality of my datasets is poor due to too many missing data. But as the most common thing we apply to fill the data is the variable wise filling. But this is not work for our complex data sets. If we see the value after filling, it is good to match the missing values by other countries, but the situation between countries are quite different. So, the previous filling method need to change a little bit to make it works. The detail will be explained in the data cleaning part.

Part 3 Data preparation

## Select the Data

As mentioned in part 2, we went through different combinations variables specified for our research goals. The observation we took is the least developed countries below the average. In the data sets, we have 81 these samples to investigate. Just a call back, our goal is to reveal the underlying reasons for slower than expected growth rate in these countries. To achieve that, the related variables could be distributed in fields including the indexes of employment rate, population, age, investment and technology.

Unfortunately, many of the variables are not direct depicted in world bank data bases. In order to save the time and avoid download the datasets again, my initial datasets have included as many relevant variables as possible to those categories and then deleting the unnecessary variables.

Initially, there are 68 categories (variables) in the data frame but as the analysis carrying on, some categories could be removed by two conditions: explained by other existing variables or having too many missing values. For instance, the replicated variable Country Code can be removed since we got the country name (the highlight variables are the remaining variables within each duplicated groups):

> levels(WBD1990$Series.Name)

[1] "countryName"

[2] "1990.GDP(currentUS$)"

[3] "1990.Foreigndirectinvestmentnet(BoPcurrentUS$)"

[4] "1990.Foreigndirectinvestmentnetinflows(BoPcurrentUS$)"

[5] "1990.Foreigndirectinvestmentnetoutflows(BoPcurrentUS$)"

[6] "1990.Netcapitalaccount(BoPcurrentUS$)"

[7] "1990.Portfolioinvestmentnet(BoPcurrentUS$)"

[8] "1990.Portfolioequitynetinflows(BoPcurrentUS$)"

[9] "1990.Netfinancialaccount(BoPcurrentUS$)"

[10] "1990.PrimaryincomeonFDIpayments(currentUS$)"

[11] "1990.GDPgrowth(annual%)"

[12] "1990.GDPpercapitagrowth(annual%)"

[13] "1990.GNIgrowth(annual%)"

[14] "1990.GNIpercapitagrowth(annual%)"

[15] "1990.Grosscapitalformation(annual%growth)"

[16] "1990.Industry(includingconstruction)valueadded(annual%growth)"

[17] "1990.Manufacturingvalueadded(annual%growth)"

[18] "1990.Servicesvalueadded(annual%growth)"

[19] "1990.Grossfixedcapitalformation(annual%growth)"

[20] "1990.Finalconsumptionexpenditure(annual%growth)"

[21] "1990.Exportsofgoodsandservices(annual%growth)"

[22] "1990.Agricultureforestryandfishingvalueadded(annual%growth)"

[23] "1990.Shareofyouthnotineducationemploymentortrainingfemale(%offemaleyouthpopulation)"

[24] "1990.Shareofyouthnotineducationemploymentortrainingmale(%ofmaleyouthpopulation)"

[25] "1990.Shareofyouthnotineducationemploymentortrainingtotal(%ofyouthpopulation)"

[26] "1990.Unemploymenttotal(%oftotallaborforce)(modeledILOestimate)"

[27] "1990.Unemploymentmale(%ofmalelaborforce)(modeledILOestimate)"

[28] "1990.Unemploymentfemale(%offemalelaborforce)(modeledILOestimate)"

[29] "1990.Unemploymentwithadvancededucation(%oftotallaborforcewithadvancededucation)"

[30] "1990.Unemploymentwithbasiceducation(%oftotallaborforcewithbasiceducation)"

[31] "1990.Unemploymentwithintermediateeducation(%oftotallaborforcewithintermediateeducation)"

[32] "1990.Unemploymentyouthmale(%ofmalelaborforceages15-24)(modeledILOestimate)"

[33] "1990.Unemploymentyouthfemale(%offemalelaborforceages15-24)(modeledILOestimate)"

[34] "1990.Unemploymentyouthtotal(%oftotallaborforceages15-24)(modeledILOestimate)"

[35] "1990.Agedependencyratioold(%ofworking-agepopulation)"

[36] "1990.Agedependencyratioyoung(%ofworking-agepopulation)"

[37] "1990.Populationgrowth(annual%)"

[38] "1990.Populationfemale(%oftotal)"

[39] "1990.Populationmale(%oftotal)"

[40] "1990.Populationages15-19female(%offemalepopulation)"

[41] "1990.Populationages15-19male(%ofmalepopulation)"

[42] "1990.Populationages20-24female(%offemalepopulation)"

[43] "1990.Populationages20-24male(%ofmalepopulation)"

[44] "1990.Populationages25-29male(%ofmalepopulation)"

[45] "1990.Populationages25-29female(%offemalepopulation)"

[46] "1990.Populationages65andabove(%oftotal)"

[47] "1990.Populationages65andabovefemale"

[48] "1990.Populationages65andabovefemale(%oftotal)"

[49] "1990.Populationages15-64male(%oftotal)"

[50] "1990.Populationages15-64female(%oftotal)"

[51] "1990.Laborforceparticipationrateforages15-24female(%)(modeledILOestimate)"

[52] "1990.Laborforceparticipationrateforages15-24male(%)(modeledILOestimate)"

[53] "1990.Laborforceparticipationrateforages15-24total(%)(modeledILOestimate)"

[54] "1990.Laborforcewithadvancededucationfemale(%offemaleworking-agepopulationwithadvancededucation)"

[55] "1990.Laborforcewithbasiceducation(%oftotalworking-agepopulationwithbasiceducation)"

[56] "1990.Laborforcewithadvancededucationmale(%ofmaleworking-agepopulationwithadvancededucation)"

[57] "1990.Laborforcewithadvancededucation(%oftotalworking-agepopulationwithadvancededucation)"

[58] "1990.Laborforcewithbasiceducationmale(%ofmaleworking-agepopulationwithbasiceducation)"

[59] "1990.Laborforcewithbasiceducationfemale(%offemaleworking-agepopulationwithbasiceducation)"

[60] "1990.Laborforcewithintermediateeducation(%oftotalworking-agepopulationwithintermediateeducation)"

[61] "1990.Laborforcewithintermediateeducationfemale(%offemaleworking-agepopulationwithintermediateeducation)"

[62] "1990.Laborforcewithintermediateeducationmale(%ofmaleworking-agepopulationwithintermediateeducation)"

[63] "1990.Laborforcefemale(%oftotallaborforce)"

[64] "1990.Laborforcetotal"

[65] "1990.Ratiooffemaletomalelaborforceparticipationrate(%)(modeledILOestimate)"

[66] "1990.Ratiooffemaletomalelaborforceparticipationrate(%)(nationalestimate)"

[67] "1990.Researchanddevelopmentexpenditure(%ofGDP)"

[68] "1990.ResearchersinR&D(permillionpeople)"

[69] "1990.TechniciansinR&D(permillionpeople)"

Also, to identify each individual case, we need a key term to distinguish our countries in certain years. Therefore, we pick the country name as our record ID (shown in the next snap shot).

Also, since this is a large data sets, the source may vary as well. Within the same field of indexes, we tend to pick only one from any similar source (the preference here is the data from ILO, which might eliminate the policy make-up in countries own report.).

In general, we take capital formation and net financial account as the status of capital and savings in reproduction. The direct age measurement in the labour force is missing in the data bases. Instead, the percent of different age group (15-19, 20-24, 25-29,…) could help us have a good representation of the age distribution. Although the industry data is not directly related to the unemployment rate, a well distribution between 1st and 2nd and 3rd sector could help us better understand the driven force in economic growth. Furthermore, our data has selected a great proportion of indexes indicating the variables by gender. For instance, we expected the female young labour could have a vital contribution for improving employments because they might be more discriminated in some developing world by cultural or other reasons. And even though the social status has improved a lot, they are still more vulnerable in employment. These indexes could help us work on the most effective marginal to improve the employment rate. One drawback for our datasets is the R&D data are completely missing. This could be the result of poor statistics system in the least developed countries or simply they do not have sufficient money or adequate technology or human capital to support research and development. Any of these reasons could stop the region to report the R&D index. Nonetheless, the structure of the industry can somehow, providing us other perspectives to reveal the research level as the higher the 3rd Industry (which is the service here), the higher the investment in research is expected.

Although there might be missing of desired variables, there is always a way to combine alternative indexes in order to achieve a similar research goal. Simple and effective is the best.

## Clean the Data

Firstly, I apply the cleaning process in Spark. I carefully follow the instructions in tutorial.

For an Na.drop() command, it will remove any rows with missing values inside. Not surprisingly, the remaining of observation is poor. The result shows we have deleted too many observations and nothing left. This because we have 68 variables and only 81 observations. We simply have a high chance to miss it the values.



Figure 3-1 The Result after Complete Removing NAs

As a remedy, we can substitute the explanatory variables only when we need to input into the model, this method of removing the missing values is called “available-case analysis”. In this case, the only columns need to be removed is the GDP percentage because any model will have GDP as the only response variable, and it make no sense to keep samples with no response variables.

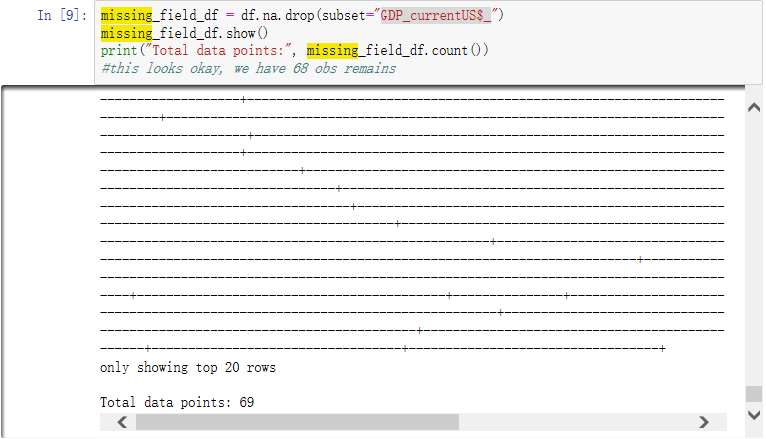


Figure 3-2 The Result of Available Case NA Removed

Eventually, this seems okay since 68 observations out of 81 is remained in year 1990.

Alternatively, if deleting NAs would generate information loss, why not filling missing values with mean values? The answer for our case is simply no because this method also has limitations. Stated before, the quality of our data is poor with 1/3 values missing. In some random line, there might be a complete miss or one present the other miss. This random missing entire row could generate difficulties for interpolation. And any replacement by mean would omit the fact that the economy and its related indexes are in growth over a period. Furthermore, each observation is much unique from others because each represent distinct countries with different population, geographic locations and industry structure, etc.

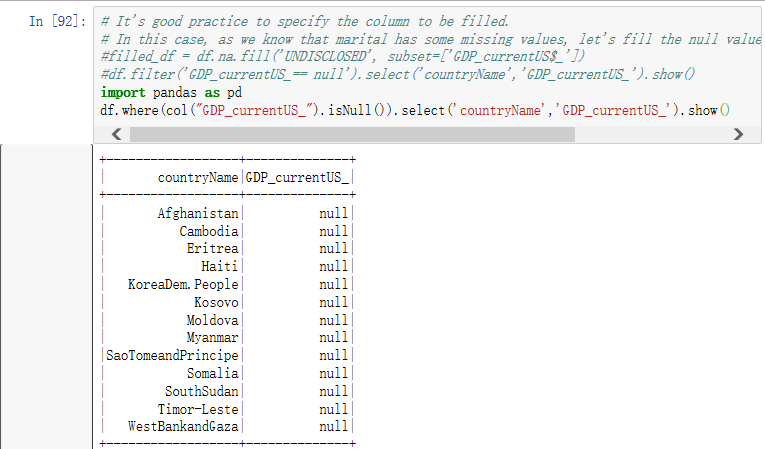


Figure 3-3 The Observations Missing GDP

The country missing GDP is shown above. The mean of the GDP through all the country is show below. We notice that the mean number is quite large, which is more than 16 billion. The missing countries are more likely to be small or in turmoil. Simply substitution by mean may not valid in our report.

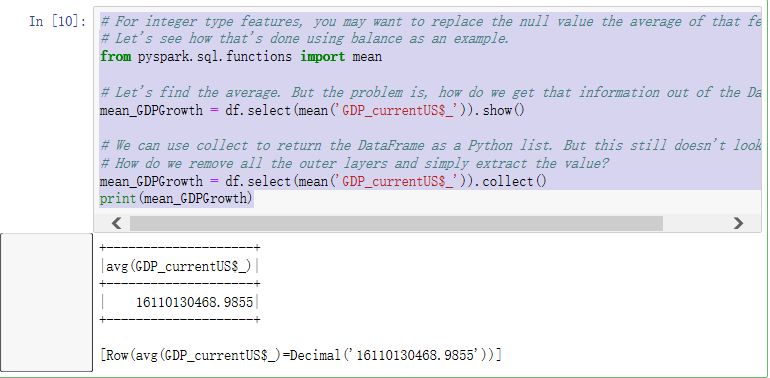


Figure 3-4 The Average GDP for the Least Developed Countries

This part could be similar to the 2-3. The original data could be quite ununiformed due to the unmatured and unsophisticated statistical system in the developing countries. Although our data have lots of elements (82\*68\*12=66912), many of the variables for individual countries or year is missing for various reasons (21867 elements missing). In this case, we cannot apply the deleting method cause one third of the cases might gone for this process. Instead, we could to use the imputation method later, like spline quadratic fit or Hot-deck imputations. But before our final data frame has been constructed, we may not mark the correct closest data and many of the missing values could be replicated and removed eventually.

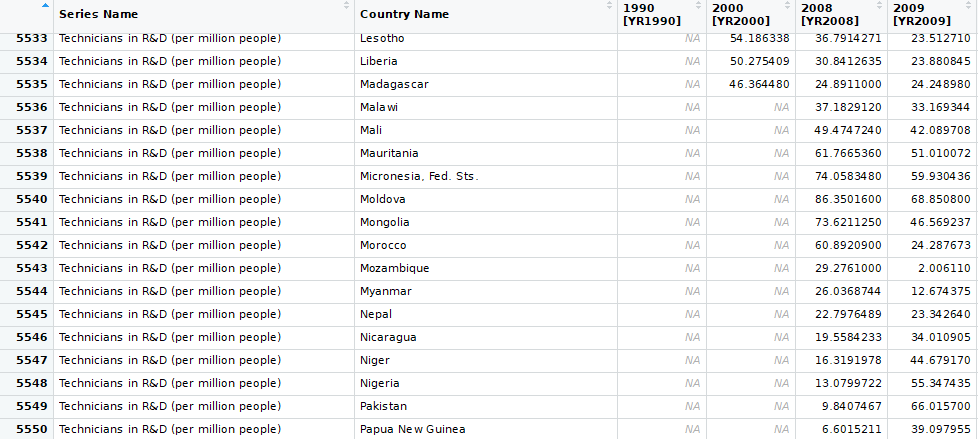


Figure 3-5 Cluster of Missing Values at the End of the Table

Also notice that for the linear function we apply in 2-3, there is still 866 missing values left. Many of these missing values occurred at the end of the table and in the year of 1990, 2000 (transpose back to the wide form to make it better for inspection). Although we fixed 99% of the missing values, mending all the missing values accurately could be costly and inefficient. At this stage, we could mark the missing values as N/A and consider to analysis year 1990 and 2000 separately or drop the R&D series due to a high missing rate.



But everything would be okay with last iterations because R have a tolerate method for regression. The pyspark got a low tolerance for the missing values. So, we apply the same method in 2-3 which is quick and dirty to fill with every NAs in the table.

Nevertheless, we have made a little improvement for the two-stage filling process.

At stage 1, we won’t rely on the in-column correction thoroughly. Instead, we apply a time series smooth which estimates the changes from little to maximum in linear smoothing. The result of the smoothing is much better than the column wised in the aspect of consistency of indexes. (The year-column wise in 2-3 could randomly got quite a different values because the adjacent cells just have quite a different values. So, the column wise interpolation is not robust).

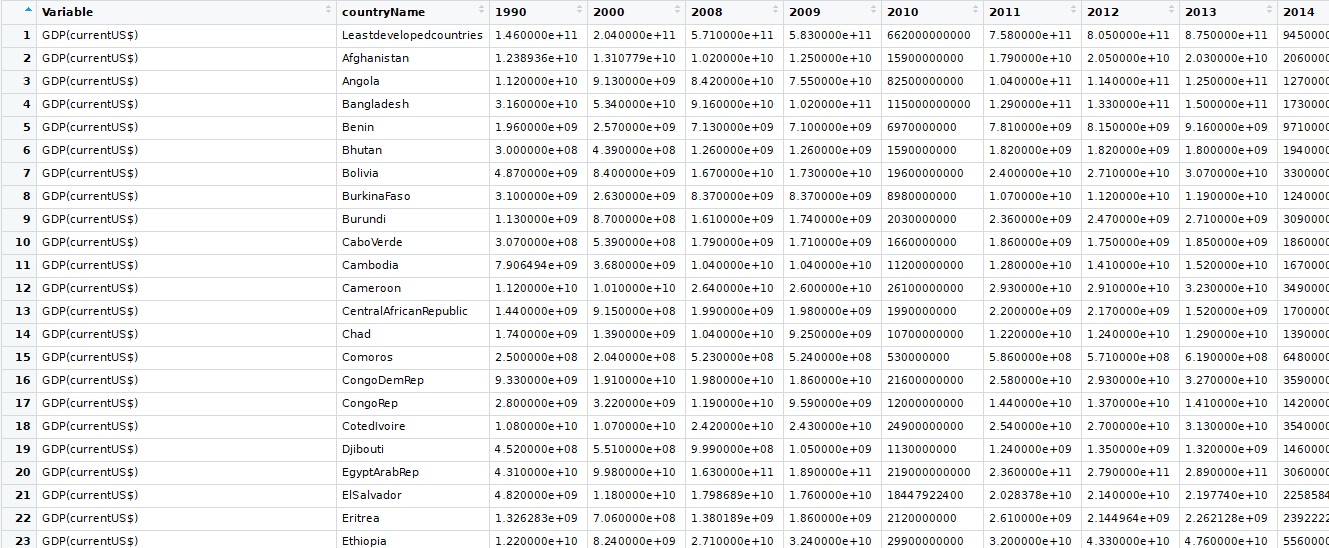


Figure 3-6 Missing Values Filled after Linear Time Smoothing + Mean filled

At step 2, we fill all the still NAs cell with row mean values. This is because the time smoothing only works for row at least 2 values presents. For many rows only one cell has data, the mean fill method is applied due to the pyspark require all filled values to carry out analysis.

However, the real case is more complex than we expected. In certain years, there could be some yearly effect like the financial crises, the political turmoil and other seasonal or yearly event which could greatly change the data by factor column wisely. The time smoothing method, therefore, will not always be the better method because the missing yearly effect will be missing using the smooth function. Therefore, the column wise interpolation could be better. But we cannot tell ahead of time which method is better. So, for the sake of better model for prediction, we apply different missing values filling method. And this is the reason why I upload many data sets on GitHub, just serving for better modelling.

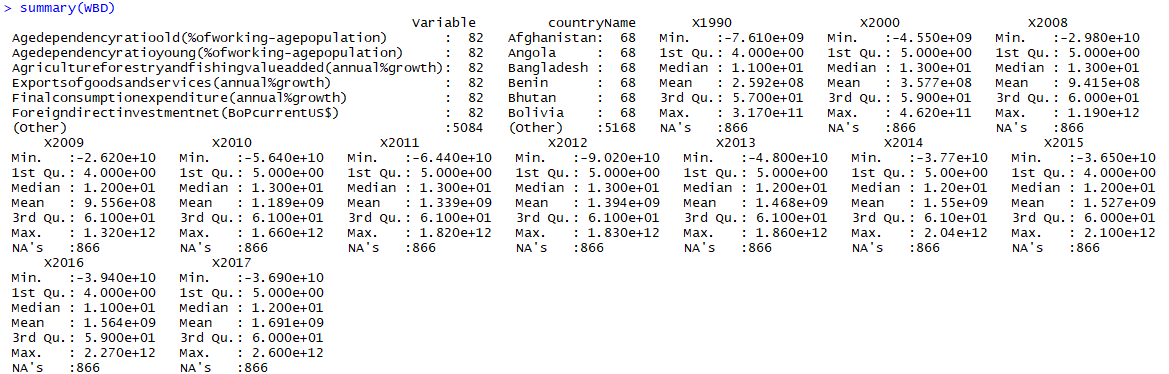


Figure 3-6b The Missing Values after Processing

After two steps of filling, the remain missing values decrease from 61/81=75% to 15%. This result can be furtherly removed in pyspark and the percent losing in na.drop() is acceptable since the missing rate is decrease dramatically. The remaining missing values could be the missing values from typical countries caused by no statistic system or other reasons.

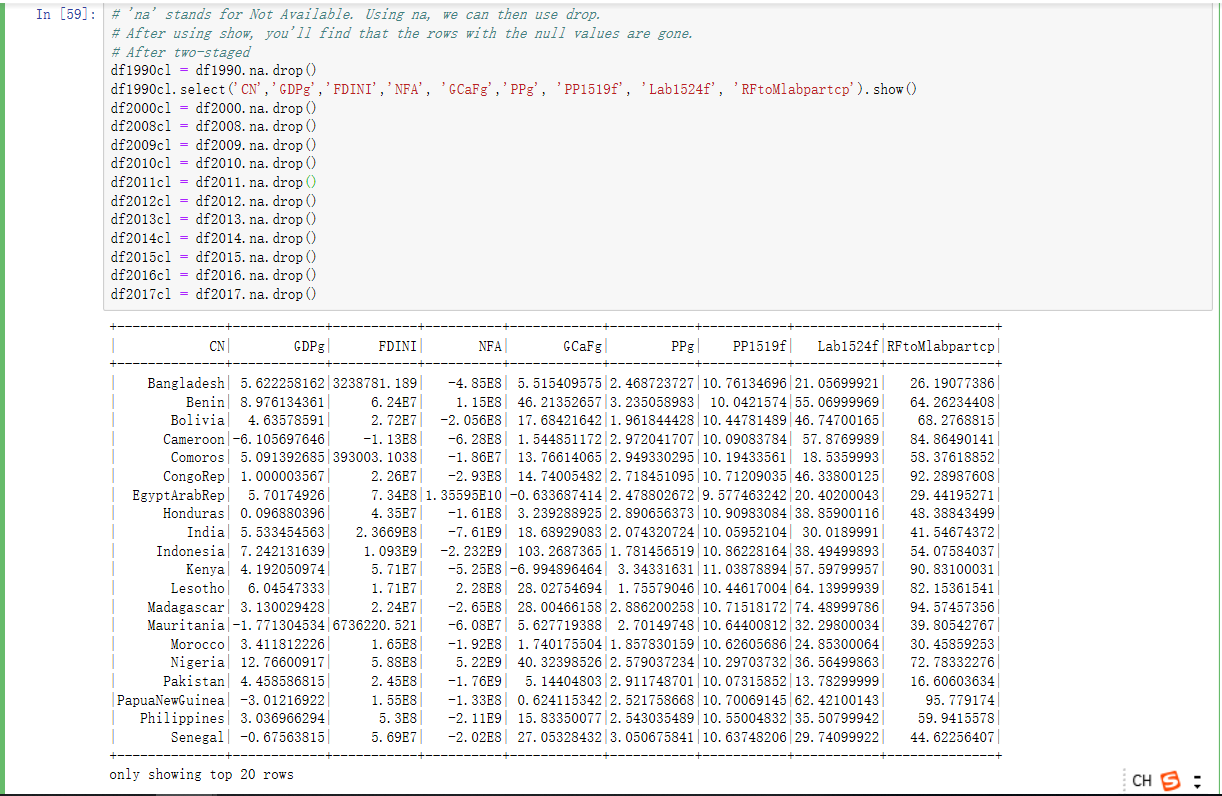


Figure 3-6c the last Deletion of NAs in Pyspark

After this removal, we have no NAs in our discrete datasets. No NAs is the prerequisite for analysis later on. But this procedure an inconsistency of the size of data frame. This could lead to a difficulty we join data frame mentioned later.

## Construct and Format the Data

The database of world bank is quite sophisticated but the output selection is only 2 dimensional. This could produce an issue with the hierarchy data like what we have in this project. To find out whether employment rate have effect with the economic growth, we need panel data from 81 countries and 68 variables across year 1990, 2000, 2008-2017. The scale of our dataset is at least 3 dimensional. When we input this data, we can either do a hierarchy data or depict our data by year and then merge it. For the former ways of input, I tried in R but the error message pops up not every time but quite regularly, saying the content in the first or second column (variable name, to be specific) is replicated and cannot be transpose to the variable name. The hierarchy for our source data is like this:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| County name | Series Name | Values | Year 1990 | Year 2000 | … | Year 2017 |
| Country 1 | GDP | A% | 2 | 3 | 1.9 |
| Country 2 | GDP | B% | 3 | 2.5 | 2.3 |
| … | | | | | | |
| Country 1 | Unemployment | C% | 4.5 |  |  |  |
| Country 2 | Unemployment | D% | 5.0 |  |  |  |

The data hierarchy we would like to construct:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| County name | Year | Values | GDP | Investment | … | Unemployment  Rate |
| Country 1 | 1990 | A% | 2 | $9999 | 4.5 |
| Country 2 | B% | 3 | $8888 | 5.0 |

However, the second row has the same content for different countries and replication. This can generate problems applying normal active analysis. If we want to extract wanted variable, we have to search and locate the variables in certain lines and since it is just in the middle of the table, it is bit hard to locate them in a single column and all the data of the wanted variables may spread across different column and rows. Thus, we may not be surprised to find out pyspark actually struggling to define the replicated variables in one row. Alternatively, we then switch our procedure by inputting separate year data. In this way, we only have two-way dimensional data and quite easy to transpose. What we get is shown below:

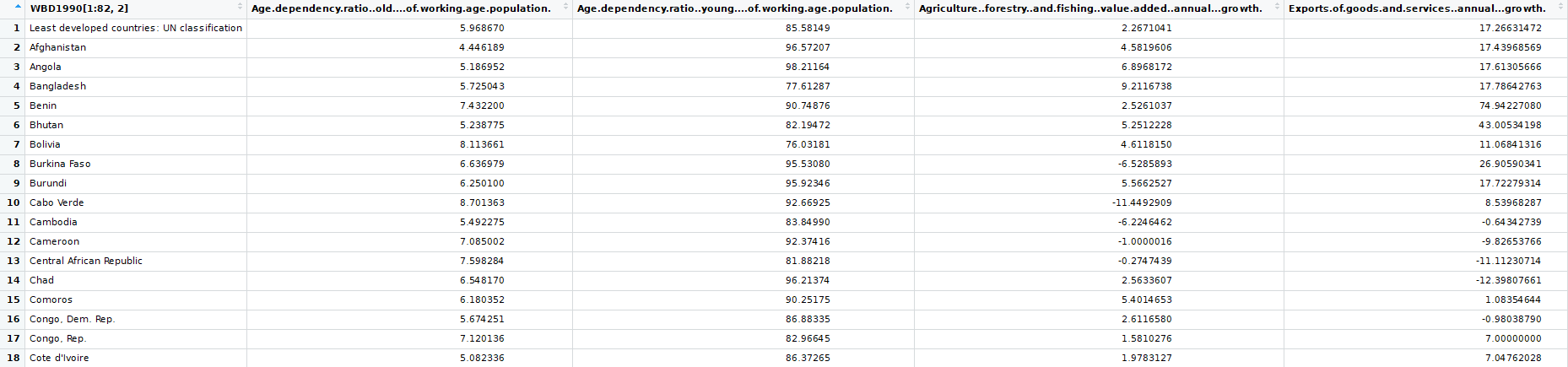


Figure 3-7 Long Format with long name after filling NAs

The architecture is:

**Year: 1990, 2000, 2008, 2009, … ,2017**

GDP independent Variable 1 Variable 2 Variable 3 Variable 4

Country Name

**Values**

**Values**

**Values**

**Values**

**Values**

**…**

**…**

**…**

**…**

**…**

Figure 3-8 Long Format in conceptual view

This structure might seem no simpler than the previous. But as we remove all the overlap rows for countries, each country could only run one line. This is prompt for explicit analysis and efficient to process the independent and dependent variables over a certain time of period. Our project goal is to find the relation between the economic growth and employment rate. This latter table could present the relations quite obvious because the changes of response variable is easy to catch since it all move to the column and all the replicated row of countries has been rearranged in different columns. Also, the association improves in the sense that he independent variables are just next to our dependent variables, which would convenient our analysis and exploration.

Highlighted here, the pyspark has a rigid rule of available character used in the name of variables. Characters as “-”, “(”, “)”, “$”, “’”, would cause a misclassification problem. So, we rename all the variable after reshape the variables to long form. Mention above, all the variables in the yellow highlighters are originated from variable Series Name in the long format. After transformation, the new name is constructed in the same sequence shown in the table from left to the right but with a shorter and abbreviate title. Shown below:

|  |  |
| --- | --- |
| Abbreviation | Long Format |
| CN | countryName |
| GDP | 2014.GDP(currentUS$) |
| FDINI | 2014.Foreigndirectinvestmentnetinflows(BoPcurrentUS$) |
| NFA | 2014.Netfinancialaccount(BoPcurrentUS$) |
| GDPg | 2014.GDPgrowth(annual%) |
| GDPpcg | 2014.GDPpercapitagrowth(annual%) |
| GCaFg | 2014.Grosscapitalformation(annual%growth) |
| Ind1 | 2014.Industry(includingconstruction)valueadded(annual%growth) |
| ind2 | 2014.Manufacturingvalueadded(annual%growth) |
| Ser | 2014.Servicesvalueadded(annual%growth) |
| ExGSgp | 2014.Exportsofgoodsandservices(annual%growth) |
| ADRold | 2014.Agedependencyratioold(%ofworking-agepopulation) |
| ADRyoun | 2014.Agedependencyratioyoung(%ofworking-agepopulation) |
| PPg | 2014.Populationgrowth(annual%) |
| PPf | 2014.Populationfemale(%oftotal) |
| PPm | 2014.Populationmale(%oftotal) |
| PP1519f | 2014.Populationages15-19female(%offemalepopulation) |
| PP1519m | 2014.Populationages15-19male(%ofmalepopulation) |
| PP2024f | 2014.Populationages20-24female(%offemalepopulation) |
| PP2024m | 2014.Populationages20-24male(%ofmalepopulation) |
| PP2529m | 2014.Populationages25-29male(%ofmalepopulation) |
| PP2529f | 2014.Populationages25-29female(%offemalepopulation) |
| PP65above | 2014.Populationages65andabove(%oftotal) |
| PP65afnum | 2014.Populationages65andabovefemale |
| PP65af | 2014.Populationages65andabovefemale(%oftotal) |
| PP1564m | 2014.Populationages15-64male(%oftotal) |
| PP1564f | 2014.Populationages15-64female(%oftotal) |
| Lab1524f | 2014.Laborforceparticipationrateforages15-24female(%)(modeledILOestimate) |
| Lab1524m | 2014.Laborforceparticipationrateforages15-24male(%)(modeledILOestimate) |
| Lab1524tot | 2014.Laborforceparticipationrateforages15-24total(%)(modeledILOestimate) |
| Labf | 2014.Laborforcefemale(%oftotallaborforce) |
| Labtot | 2014.Laborforcetotal |
| RFtoMlabpartcp | 2014.Ratiooffemaletomalelaborforceparticipationrate(%)(modeledILOestimate) |

Table 3-9 The Short and Long Name Corresponding Table

The year in the long format is just distinguishes the year. Each year the start of the long format is different but the rest are the same. Also, we cannot start variable names with numbers in pyspark. This is another reason why we change the name of variables.

## Integrate Various Data Sets

In economic fields, there are always the case when more than 3 dimensions of datasets across time, objectives and other cluster dimension measurement serving the goals.

As we dump many of these complexity into a 2-dimensional data sets, we need to merge our multi-dimensional datasets into a common key (country name) and join all the rest variables with their values. By doing these, the three or higher One signle “Series Name” representing all the variables, as mentioned above, could be difficult while carrying out variable analysis.

This combined name might be neat to observe but hard to analysis. For each level or factor in series names indicating a variable. But these variables are sorted by the country name. If we want to find the across region average, the structure works by adding all the same index nearby, but if we search for the indexes within a typical country, the close featural row are separated by other 80 countries through rows, which are completely a mess if we compare between indexes. So there is a need for splitting the data once it merges.

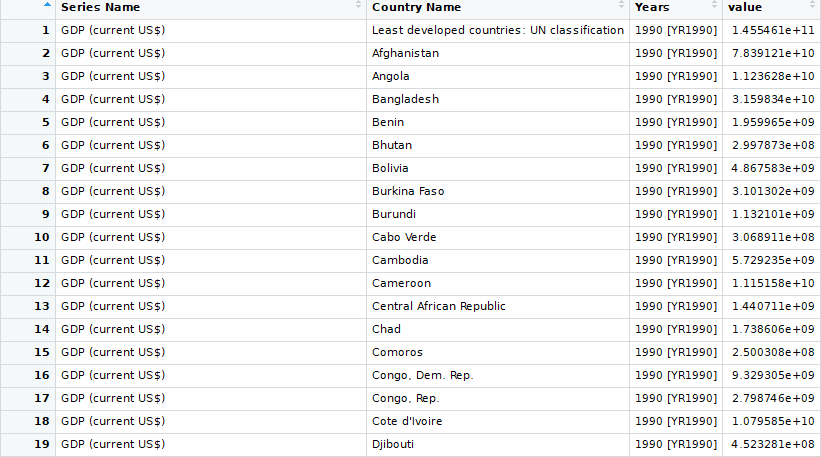


Figure 3-10 Long Format with Single Series Name

Accordingly, we need to transpose our single column source data at every specific year to one country name on rows versus all the variables in columns under a certain year marked at the column. We can set all the requirement in the transpose dialog. Shown in the picture below, the key variable (country name) is treated as index, the source column to derive 68 different variable names is added to the field category, and finally, all the values named after certain year (here 2011) fits in value category. After transpose, each data set remains the sequence from row wise to column wise.

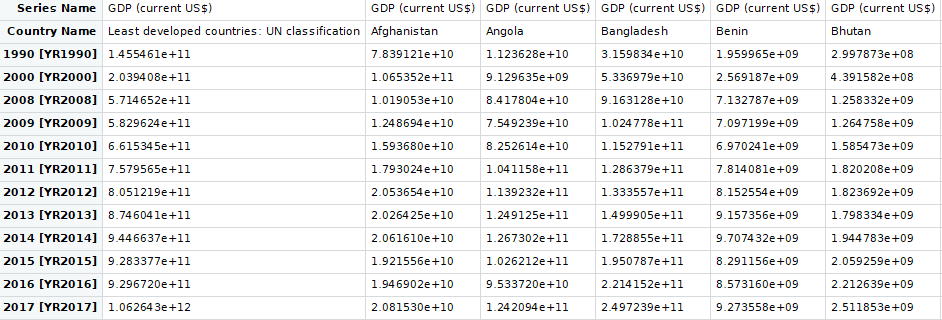
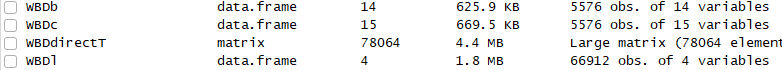


Figure 3-11 Wide Format with each factor of Single Series as independent variables

Also, by transpose or reshape, we have a risk of enlarge the datasets. This is because the long format requires an external long storage for the names and indexes. And each name repeated many unnecessary times which could consume much more space. It would be much efficient if we expand the number of columns to reduce the unnecessary times of repetition.

We get 7 times larger data sets with no gaining information.



And reading and basic operation becomes really slow in the matrix because all of the columns are stored with one long sequence in memory and we have to search a long distance to find the desired value. This is really low efficient. Instead, we can boost our efficiency by splitting the long form of hierarchy data by series name. This is like multitasking. And in order to split by variables in series name, we need to factor the series name and then split by the factor. The process time is 80 times quicker than the traditional.

#year 1990

WBD1990=data.frame(WBDb[,1:3])

WBD1990r=split(WBD1990[,3], as.factor(WBD1990$Series.Name))

WBD1990r=data.frame(WBD1990r)

WBD1990j=cbind(WBD1990[1:82,2],WBD1990r)

colnames(WBD1990j)[1]="CountryName"

#year 2000

WBD2000=data.frame(WBDb[,c(1:2,4)])

WBD2000r=split(WBD2000[,3], as.factor(WBD2000$Series.Name))

WBD2000r=data.frame(WBD2000r)

WBD2000j=cbind(WBD2000[1:82,2],WBD2000r)

colnames(WBD2000j)[1]="CountryName"

….

Merging the data could be also quite obvious and efficient. Unlike the SPSS softwires, dataframe operates in a single line code, just as shown in the above. The result screen shot shows the country name matches with the individual values of variable by joining the same sorted length of matrix column wise.



Figure 3-12 the result for Combining Columns in Data Frame

After merging, we obtain 68 rows plus one country name data frame across 81 countries. And we reproduce the same table for each individual year (68\*12 = 816 fields). If we are interested in certain countries cases, we can just check the specific rows and obtain all the information we needed. Also, this format is convenient for comparison between countries.

After exporting the reshaped wide format datasets to the pyspark, the filled data for each individual table have the same rows each represents a country. The country name acting as the key identity columns between tables. This key has the same length and it is ready for joining. This is the reason why our input just has replacement with NAs, not replacing it. If we remove the arbitrary data, the result would be different dimension of subtitles which cannot be easily re-join.

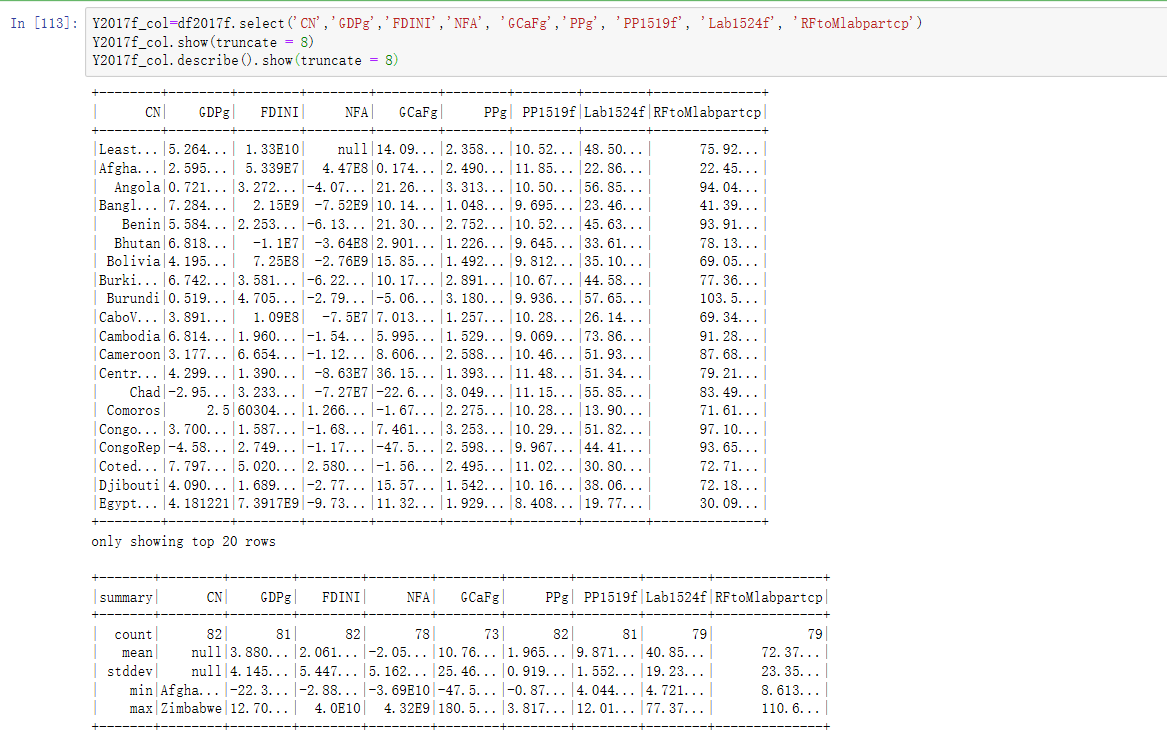


Figure 3-13 Input the Reformatting Data Frame with trimmed content

Most of the NAs are removed as stated before by two step filling. But in the first row, there is still a missing value in the gross (the headings are the aggregates number, will be NAs if there are missing values for each countries). This could be handled by na.rm when ever the column is specified into a model. But for the matter joining and splitting process, we just keep the very few uncovered NAs.

Also, for the round issues, we actually trim the data once we put in. But since the we have to show the process, here we just use a tricky show truncate command to show it as if it is not trimmed.

Here we apply a round to keep everything readable by 2 decimals:

import pyspark.sql.functions as func

SCol=['GDPg','FDINI','NFA', 'GCaFg','PPg', 'PP1519f', 'Lab1524f', 'RFtoMlabpartcp']

df1990na = df1990.na.drop()

Y1990\_col=df1990na.select('CN','GDPg','FDINI','NFA', 'GCaFg','PPg', 'PP1519f', 'Lab1524f', 'RFtoMlabpartcp')

Y1990\_col=Y1990\_col.withColumn('GDPg',func.round(Y1990\_col['GDPg'], 2))

Y1990\_col=Y1990\_col.withColumn('FDINI',func.round(Y1990\_col['FDINI'], 2))

Y1990\_col=Y1990\_col.withColumn('NFA',func.round(Y1990\_col['NFA'], 2))

Y1990\_col=Y1990\_col.withColumn('GCaFg',func.round(Y1990\_col['GCaFg'], 2))

Y1990\_col=Y1990\_col.withColumn('PPg',func.round(Y1990\_col['PPg'], 2))

Y1990\_col=Y1990\_col.withColumn('PP1519f',func.round(Y1990\_col['PP1519f'], 2))

Y1990\_col=Y1990\_col.withColumn('Lab1524f',func.round(Y1990\_col['Lab1524f'], 2))

Y1990\_col=Y1990\_col.withColumn('RFtoMlabpartcp',func.round(Y1990\_col['RFtoMlabpartcp'], 2))

Y1990\_col.show(20)

#another way

df1990.select('GDPg','FDINI','NFA', 'GCaFg','PPg', 'PP1519f', 'Lab1524f', 'RFtoMlabpartcp').describe().show(truncate = 8)

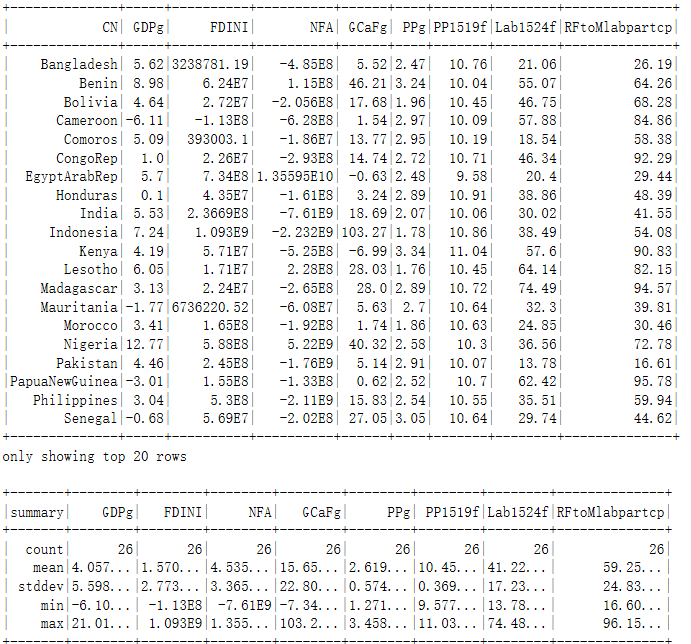


Figure 3-14 The Result Keeping 2 Decimals and with some statistics shown

As mentioned above, each table is one-year data. This separated sheet can be treated as data from different sources because the missing values are randomly generated and interpolated based on the random missing sequence. As the split is more efficient in analysis, we will keep the data separation during the modelling process. But the sperate of sheet has advantage that we naturally split the data into the test and training data by fixed year gap. And the result in each year eventually sum up through iterations and produce a less biased model and improve the accuracy in the model.

Part 4 Data transformation

## Reduce the data

The merging data sets has 816 fields. There are still many of combination left. We reduce the data by horizontal first (by variables).

In linear model, usually we keep several variables as the simplest and most powerful model. Hence, there is a need to reduce our variables. But just as stated before, the variables belong to different fields. Some of the variables are complementary to each other, then a removal is possible if you know the gross. A full set of variables contains all the necessary fields with at least one variable cover each side. Therefore, the combination of variables might be more than one sets. Here I just illustrate one of many possible combinations.

Also, many of other variables are removed due to duplication or some rare lines filling with missing values after two round of NA fills. Stated before, we discard the R&D variables since there is to many missing values and some year the available data is only several percent. Other section like gender divided education and employment rate are not available in most of the developing countries. Indexes like GDP net values is removed since we considered the growth percentage already. Similarly, we remove the GNP measurement because GDP and GNP are quite similar as a dependent measurement. All of the removal operation is carried within the filter node. We can call back and change our filter if the result of our first set of variables does not work well.

countryName

GDP(currentUS$)

Foreigndirectinvestmentnetinflows(BoPcurrentUS$)

Netfinancialaccount(BoPcurrentUS$)

GDPgrowth(annual%)

GDPpercapitagrowth(annual%)

Grosscapitalformation(annual%growth)

Industry(includingconstruction)valueadded(annual%growth)

Manufacturingvalueadded(annual%growth)

Servicesvalueadded(annual%growth)

Exportsofgoodsandservices(annual%growth)

Agedependencyratioold(%ofworking-agepopulation)

Agedependencyratioyoung(%ofworking-agepopulation)

Populationgrowth(annual%)

Populationfemale(%oftotal)

Populationmale(%oftotal)

Populationages15-19female(%offemalepopulation)

Populationages15-19male(%ofmalepopulation)

Populationages20-24female(%offemalepopulation)

Populationages20-24male(%ofmalepopulation)

Populationages25-29male(%ofmalepopulation)

Populationages25-29female(%offemalepopulation)

Populationages65andabove(%oftotal)

Populationages65andabovefemale

Populationages65andabovefemale(%oftotal)

Populationages15-64male(%oftotal)

Populationages15-64female(%oftotal)

Laborforceparticipationrateforages15-24female(%)(modeledILOestimate)

Laborforceparticipationrateforages15-24male(%)(modeledILOestimate)

Laborforceparticipationrateforages15-24total(%)(modeledILOestimate)

Laborforcefemale(%oftotallaborforce)

Laborforcetotal

Ratiooffemaletomalelaborforceparticipationrate(%)(modeledILOestimate)

Even though the data now looks much smaller and compact, it may be redundant to answer our model efficiently. So, it seems not enough just reduce the data by the quality of each variables but also by the relevance of our data against the response variable. The further deduction is after the projection part.

Call back to the figure 3-14, the part we shown is also a quick and dirty ways to explore the feature of the explanatory variables. These variables include only the highlight one shown above.

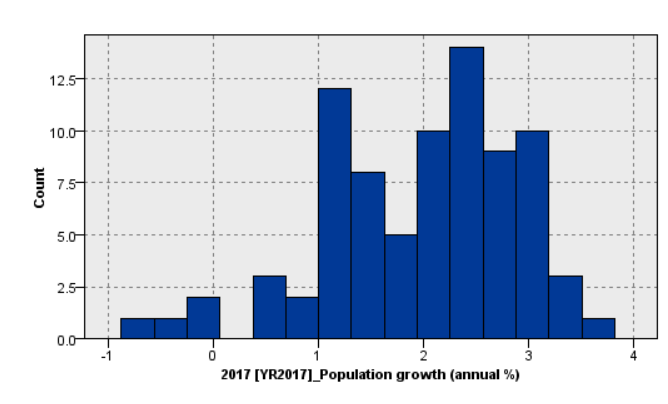
## Data Project

The quality after processing is quite good compare to the original raw data since many of the missing values have been removed (The rate of missing values decreases from 75% to 15%).

The majority of our data has been standardized. This means we scale the GDP, number of populations, the gross output, the number of labours all to variation of percentage changes over the period. The major merit from this mathematical process is that we do not need to worry about the scale or significance of our variables. And the marginal proportion could be more accurately measured for multilinear relations. And for some data original in scale of 0 to 1 and 1 to 10, transfer to the log scale could be misleading and negate the associations which could be misleading. In additive models, the log for dividing multi additive component could also be complex. This is the reason why we choose a standardized simple additive model than a logscale multiplication or exponent model. And simple is always the best.

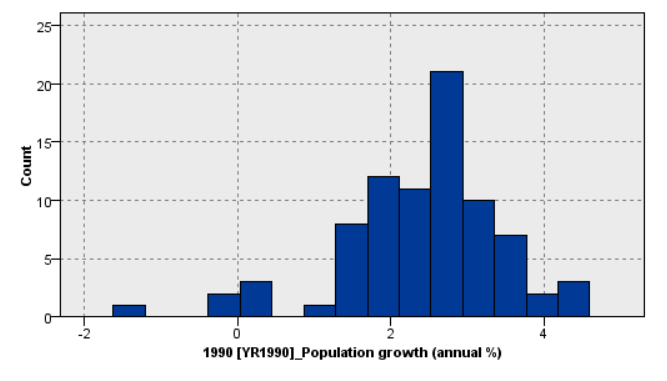
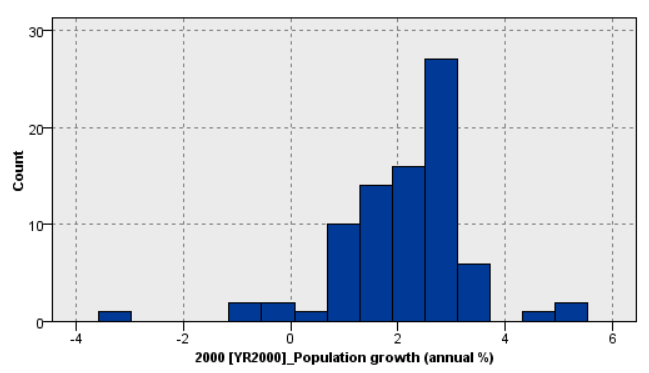
Also, the data published by world bank are likely to be checked for any invalid values. So, I didn’t find a major issue with the odd or invalid data. But this doesn’t mean I did check for these values. For instance, one of the relatively good quality variables, say, the population, marked yellow in the snap shots, have one outlier shown by the system. But if we check the min and max, the maximum number is 3.817% might be quite low for developed countries if you consider the high investment, emerging market and high inflation in the economy.

But for those regions with war and conflict, refugees could generate a vast and instant increase or decrease of the nations. Just maybe the odds even out the reginal growth. It is reasonable to argue the odds might be rare and thus fluctuations could would have little impact on average or the long term of the growth. Should we remove the so-called outliers immediately? Possibility not. If we check the overall distribution, it seems a right skewed distribution with central sunken for some reasons unknown. Neither the two end of the distribution looks unnatural. Instead, they follow the distribution overall. So, I keep the outliers in this situation.



For each variable category, there is some trend could be interesting:

1. Population growth in this least or lower developing countries cluster somewhere between 1 and 3 in 2017. And similar overall distribution found in the year of 1990 and 2000.



This may impose the endogenous factor of their economy has not changed much over nearly 30 years. This could be very suggestive for the unsuccessful in the development due to a vast population growth eaten up the return of the economic growth.

1. The employment rate in these countries are mainly right skewed.

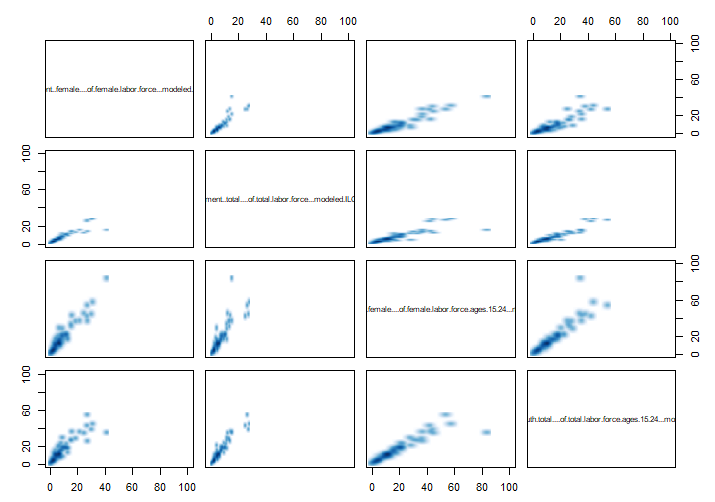


Figure 4-1 The Distribution among labour force

The average of the total unemployment rate for less developed countries are 6.97% in 2017. The male tends to have a lower unemployment rate than female. Unfortunately, there is an excessively high unemployment rate for young labours (aged 15~24, not in school, seeking for full- time job). The female young labours, especially have a hard time with unemployment rate rank 16.45%. Just for comparison, the north American seem to have a lower unemployment rate. (the table is taken from ilo website[[12]](#footnote-12)) But we also notice like developed countries Canada, enjoying 3.0% economic growth (which is high in developed world) with relatively high unemployment rate (7.1%, above the less developed countries’ average).

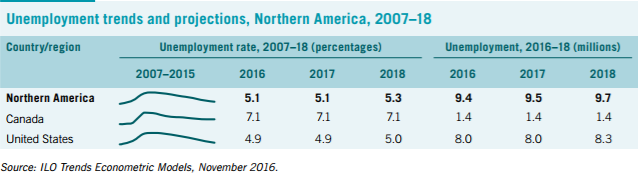
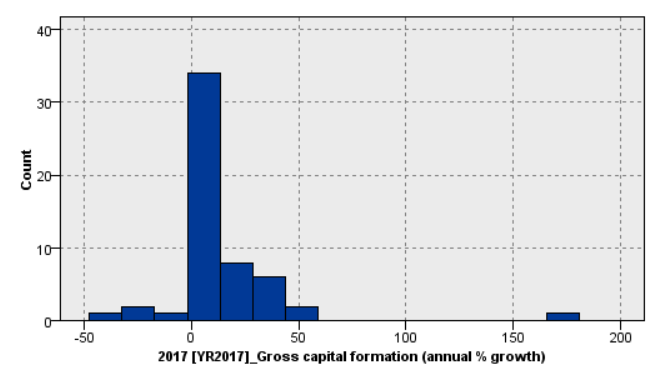


Figure 4-2 the Unemployment rate for some Developed Economies

Therefore, we might suggest the unemployment rate overall in lest and less developed countries may not be unacceptably high. But what matters is the unemployment rate for young labours are too high. And these issues could be relating to many aspects which would discussed in the following sections.

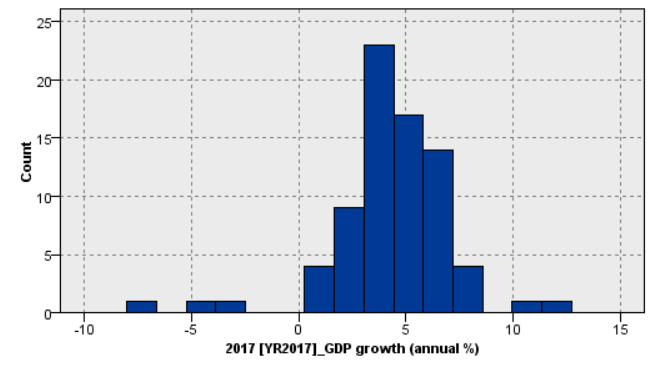
1. Capital formation



Many theory and economics affirm the fundamental role for the capital devotion to high economical growth. This could be true on certain extent. But the overall distributions shown many countries actually earns little on capital level but the overall economic growth is still significant. This could be an interesting and further association analysis could be carried on.

1. Economic growth

We also need to project our dependent variable, the growth rate.



As our target, these left behinds need to achieve 7 percent annual and only 5 nations in our list achieve the goal. The median and mean growth rate, among these countries, lies between 3 and 5%. Here we confirm again that the less developed countries could improve their economic growth by working on the most related fields.

Part 5 Data-Mining method selection

Based on the tip of sup in 4.2 section, the histography just provide a general view of the shape of distribution of our variables. Our data mining goal is to find the convincing relations between the economic growth and other variables, especially for the population, unemployment rate and the capital available in the market. To achieve this, we most likely need to apply regression to find out the correlation and explanatory power of our independent variables. This process is called the active analysis, in which we could easily tune and predict if one variable or sets of variable changes, how the result would affect respectively. The advantage of this method is the understanding of its mechanism thoroughly. But this method can also be weak. If we construct a hierarchy data structure or 3D dimensional data sets, the expression of regression might be slow and complex. Suppose we cannot determine what is the major effect, then we can firstly put related values or variables into a cluster or classification.

In order to do the classification, in the part 2 and 3, we have made some guesses with the possible combination of variables. The variables we selected are the goods and the most possible variables which have the major effect by our mind or called it empirical intuitions. But statistics or academia could trust such as our conclusion. Instead, we need to apply the proper classification method. And the best classification model for our datasets is the tree partition or call it decision tree model. The reason we apply tree partition is not only it provide us a good solution of cluster by a greedy approach, but also trees enable us to deal much with missing values much better, possible better than another existing models.

However, since the method we apply is greedy (locally optimal), the story may look the same to the previous linear regression model because the process is still arbitrary by the threshold we choose to start the partition. To tackle this, we introduce a random forest model. In a random forest, we could traverse all the possible starting points and find the most effective splits and result in the most repetitive combinations of variables serving our goal, which are the major effect on the economic growth rate.

These are directed learnings.

But there might be more to be found if we apply undirected path by NN.

In the Neuro network model, we can apply the passive and undirected machine learning processes to identify patterns in this fields. The method classification might help because the vast data may vary from time to time. Under certain year, there might be seasonal or sudden event which may reduce the generality of our gross model. Considering this factor, the classification by tree and by mind may not come out with a best solution simply there is too complicated and beyond our perspective of human mind to understand. And this is particular true for complicated fields like the overall performance of the nations’ economy. Nero network working in a black box, within which multilayers of nodes has been plotted and each representing some variables derived from the already know variables. clustering might be important to identify the unknown patterns in our data sets. And by setting connection between unknown nodes, the connection between already seen nodes could be build and a better and more precise model could be built to giving better explanation of what happens in the past.

Random Forest

Cluster

D

**Supervised learning**

General Linear Regression

**Unsupervised learning**

Input

Output

Neuro Network

Figure 5 the Conceptual of the Model Methodology in this paper

Based on our data sets, bunch of different categories into similar small groups, and possibly we can conclude the something rational and convincing hopefully in the end.

Part 6 Data-Mining algorithm (s) selection

Some of the exploratory process has been mentioned in section 4.

The first step we apply here empirical based selection of the variables stated before.

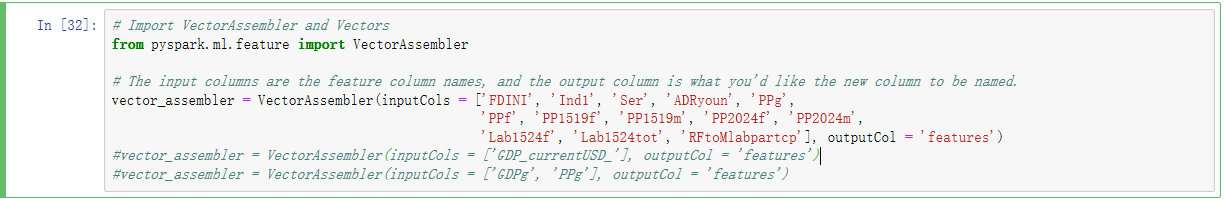


Figure 6-1 Vectorize the input variables

The combinations of variables have been all been chosen are: 'FDINI', 'Ind1', 'Ser', ADRyoun', 'PPg', 'PPf', 'PP1519f', 'PP1519m', 'PP2024f', 'PP2024m', 'Lab1524f', 'Lab1524tot', 'RFtoMlabpartcp'.

In order to carry out model analysis, we need to transfer our dater into a vector and assign an extra attribute to the array of features in pyspark. Comparied with R, the R support vectorized computation inherently without assigning and creating a special vector format. So it is much more efficient to carry out in R. But since we are doing models in pyspark, we have to change to vector and then assemble it.

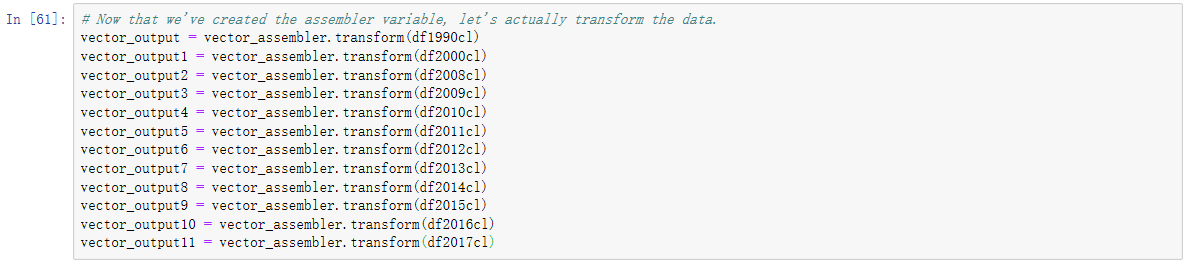


Figure 6-2 Vector Assembler

Notice here we sperate the vector output by years because each year have rather random NAs. And after the remove of NAs, the length of columns and rows are different. So, it is quite complex to merge here and we keep the separation till the train and sets merging.



Figure 6-3 Print Schema for original datasets

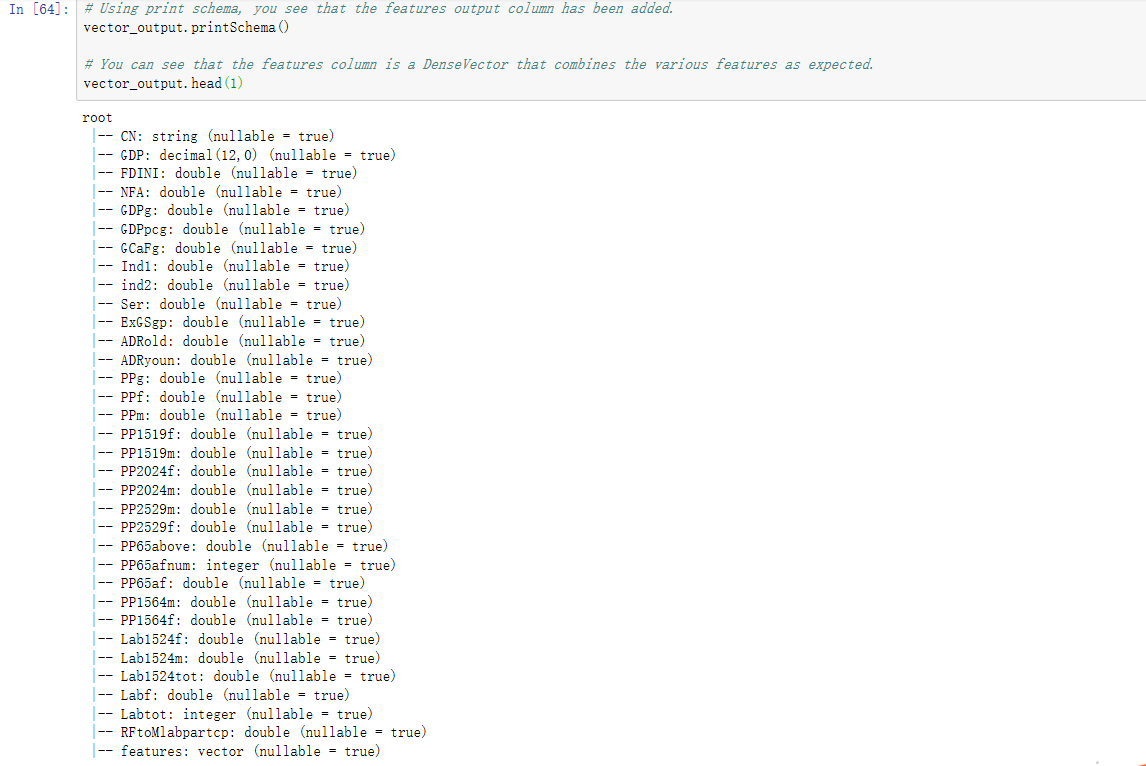


Figure 6-3b Print Schema for the vector output

The features match so we could have a merge after transforming to a dense vector.



Figure 6-4 The Dense Vector Format

Each of the feature cells hold an array. And the columns are reformatted and we can apply the merge of vector row wise now.

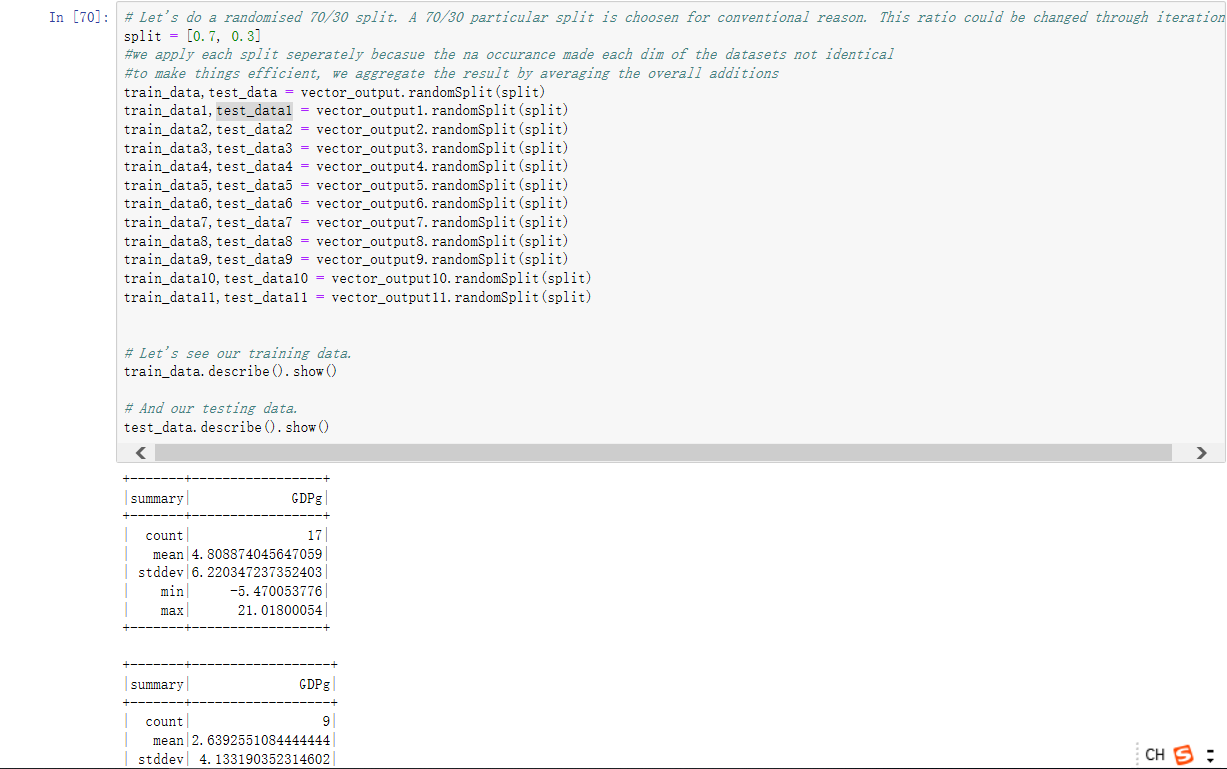


Figure 6-5 The Splits

The split will be mentioned in section 7. Showing here is just for showing an consecutive procedure.

Then we merge the training data and test data to one.

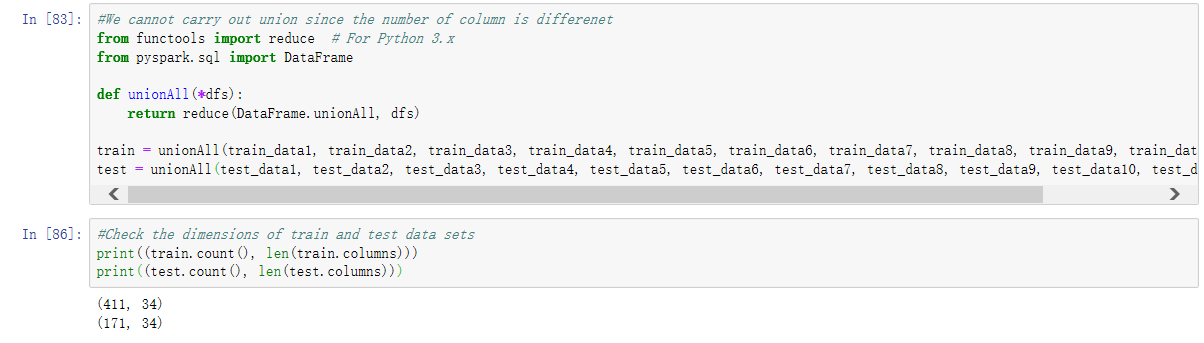


Figure 6-6 Merge train and test data

The dimension of them is roughly 411/(411+171)=0.706, which is the same as 70/30. This proves the division of the overall output could be roughly similar if we carry out by years. This could generate some split pivot problem which means the miss classification problem near the split. But since we have lots of data, this problem is not significant.

We start with a quite complex sets of variables avoiding missing any good things in the sections. But the many variables we pick may not be good because the overuse of the variables may have multi correlations. This means variables could explain each other. They are not independent. If we include such, the overall explanatory power would decrease. Therefore, selecting the minimum variable sets but nearly have the same overall explanatory power is vital. But since this has been a good guess for it, we may have little improvement or no improvement since it is my best guess. So the second step is vital to improve our model’s performance.

Secondly, we apply the undirected or unsupervised model since we may not know the patterns. Then we apply the directed and supervised because we can test our hypotheses (test our empirical picked variable sets) and to answer our research goal “whether unemployment rate has devoted to the economic growth”. At the same time, we can inspect all the variables through our panel data to find out anything significant towards GDP growth.

Inspired by precious SPSS result, we know that the random forest could find us the right combinations.

Ranked by correlations, the most significant variables are not the labour in our exploration. Take the most recent year, the significant variables from the most to the least (positive effect) are: Industry value added; gross capital formation; manufacturing value added (third industry); and the forth place comes the labour force participation rate for youngers (15-24); follows the same participation rate but for females.

But the connections might be weak since the correlation number is not significant. We exclude the agriculture because its major distribution is nearly half below 0 and correlation is smaller than 0.2 (for social science, 0.2 may still have meanings so the keep variable with correlation> 0.2).

For detail and prediction, we then apply the neural network model. The overall accuracy is 62.5%, it is acceptable.

The overall prediction for GDP using the given combination are okays (both have same scale).

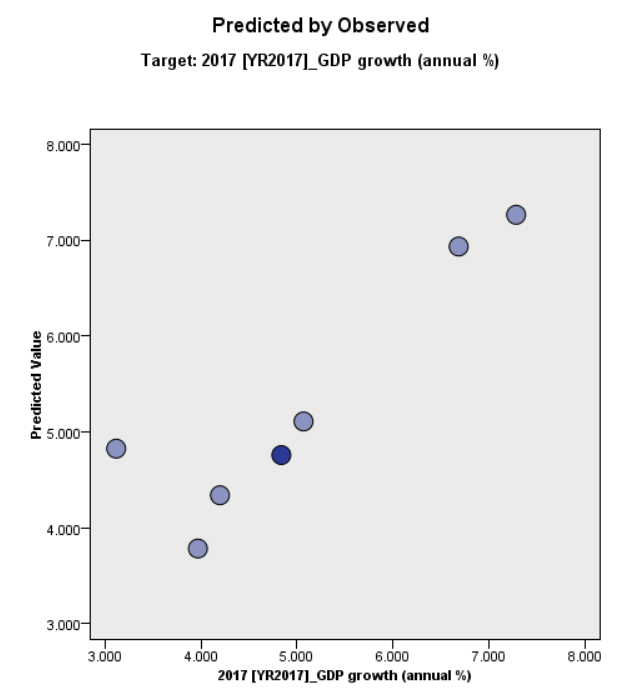
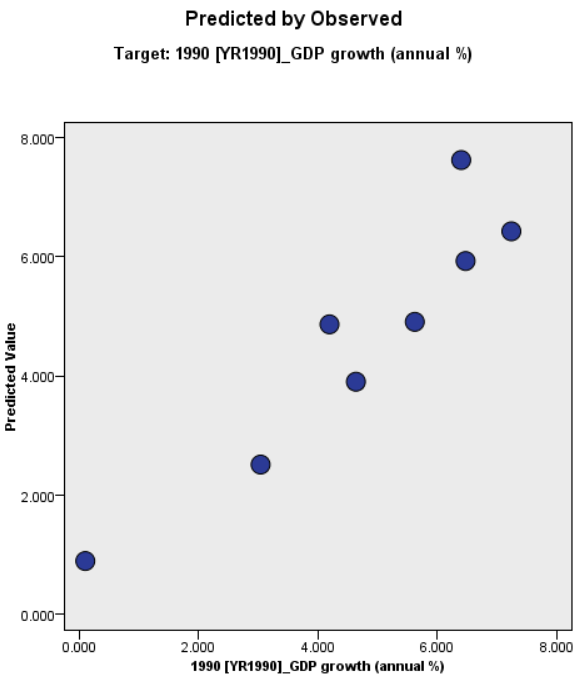
 

Figure 6-7 The predicated result by grid search method

Call back to the SPSS, they apply a grid search for the best solution among all of the variables we inputted. Although it is subjected to the first round of our variables selection, the prediction is acceptable because it predicted matches what we observed in data sets.

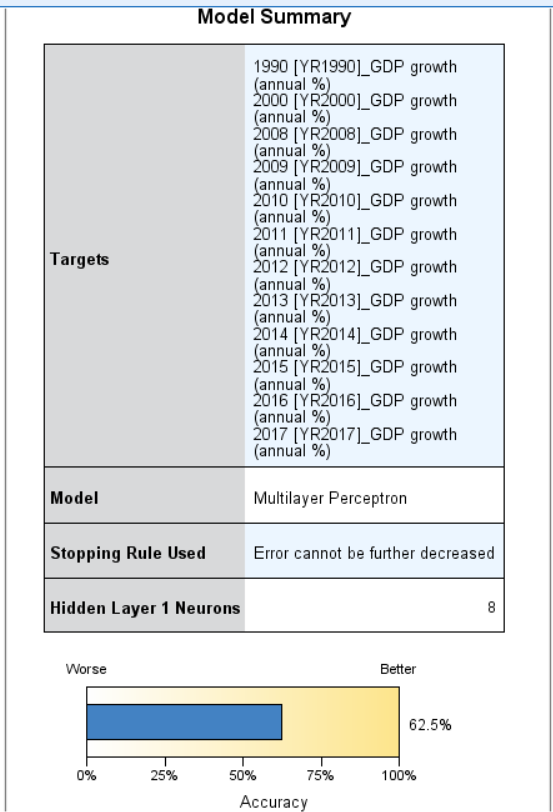


Figure 6-8 The details for regression

The prediction power is acceptable for economic fields. Above 60% variation of GDP growth has been explained by our sets of variables.

Although the overall prediction is good, the prediction importance shown in the following seems to fail in NN model. Maybe we need to remove more related variables or analysis each year individually or change the predicted models.

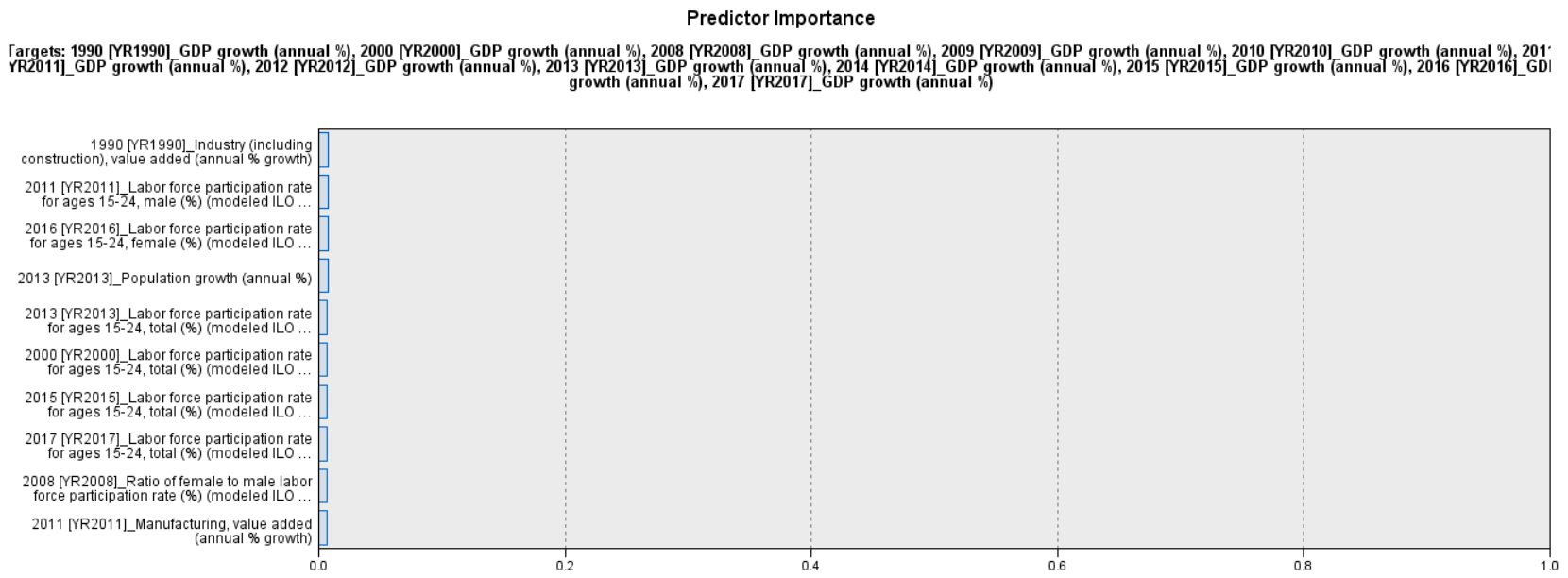


Figure 6-8 the importance of sets of variables in Nero Network Models

Suppose we choose the year 2017 and run a supervised model: random tree for prediction importance, we get some correlations:

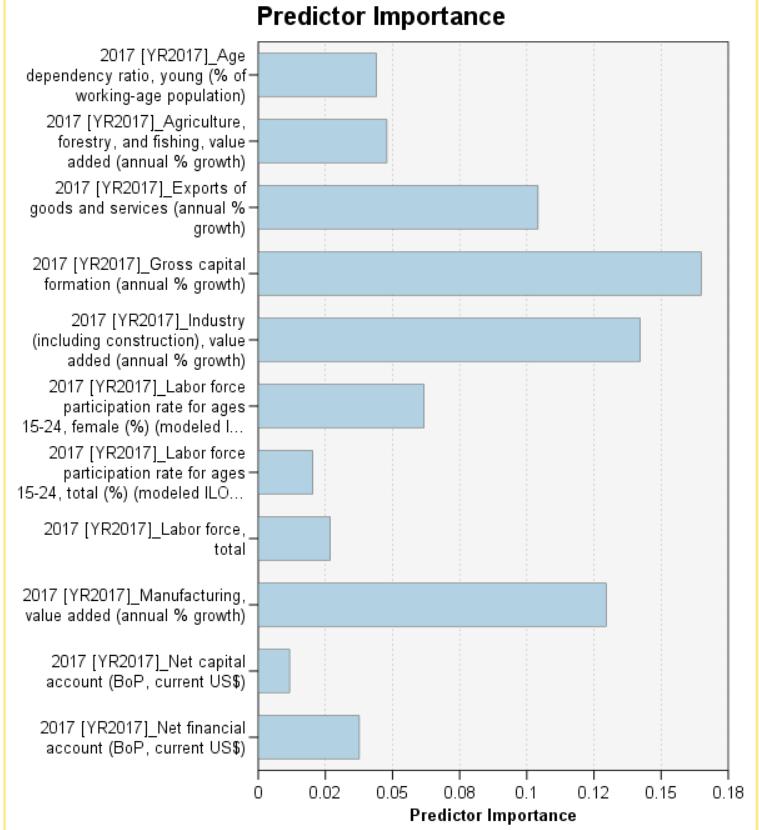


Figure 6-9 Predictor Importance

Different from what we get from the exploration, random forest gives us the strongest factor is actually the Gross capital formation, following by industry value added, manufactural and exports good and services. All these important variables are kinds of capital data rather than employment data.

Knowing that capital variables have proportion, we want to analyse the individual affect on employment. So, we kick off all the most relevant variables

Then we apply the regression model.

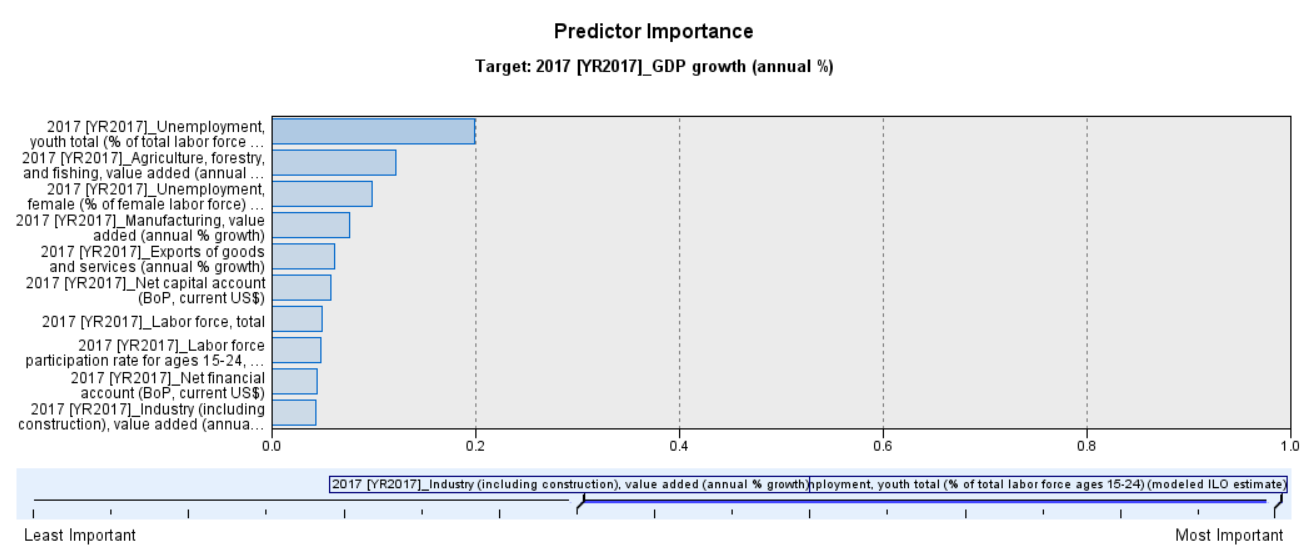


Figure 6-10 2nd Iteration for linear regression

In this iteration, we remove the industry data and capital indexes, which are the most significant in 1st iteration. The result could be interesting because the unemployment rate pops up. The most importance predictor is the youth unemployment rate followed by the agriculture output (first industry). The next rank is the female unemployment rate and then the second industry output and export and capital.

The best algorithm might be hard to identify by human mind, because each model may return different importance of variable sets.

But if we just apply the process we apply in section 6, a better and accurate sets of variables here could be produced. And with better sets, we can testify the hypothesis in the supervised model, which is more helpful for us to answer question we interested, such as whether unemployment matters.

All of the procedure above has proven that a multi combination of model would improve our model accuracy and the best is to learn from potential better combinations and developed a better linear model.

Our model selection is based on the undirected and directed sequence, targeting on getting the best combination of models. Our baseline for the research is just simply the general regression model. All the random tree or the NN is just serves like a grid searching purpose targeting for the best combination.

Random Forest

Cluster

D

**Supervised learning**

General Linear Regression

**Unsupervised learning**

Guess sets1

Best Sets

Neuro Network

Figure 6-11 Model Selection Logic Graph

If we have to pick one best model, we take the **general linear regression** as ours model. This format is the most matches the format of Solow Model which explains the reasons for economic growth between capital, population and the classic economic endogenous growth model. And the supervised model has the advantage to manipulated the variables individually and measuring the elasticity by changing the marginal values.

Part 7 Data Mining

## 1. Logical test and design

Splitting the data could be useful when building a passive and undirected model. This method could be informative to split the data to the training data sets and predict data set by 70:30.

This initial ratio is just pick by convention. But the conventional believe could also have theoretical basis. In real time, the misclassification rate is around 69% of the data in datasets. So, if we apply 70% for train data and 30% for the test data could be optimum.

The other ratios may be better or worse. We don’t know the best ratio in each case before applying such ratio to the model. What we can expect is the more we train, the higher the explanatory power for the test data. On the other hand, the size of test data left might be small, the chance of getting odds observations could also increase. Since our goal is to reveal the relations, we need to eliminate the split bias as much as we can.

However, this ratio of split may generate leave out because the situation between countries could be different. If we divide some odd patterns in the test data, these types of features would not be learned by the model constructed by running data. Since we only care about the relation between the growth rate and the other explanatory variables, if we divide manually by variables, this would produce some what a bias. We should be aware of selecting 30% of the counties may generate a risk of insufficient study with the observations. Between the countries, many variables could be different, not only the location but also the proportion of the industry and the labour force and education. Therefore, the classification of country group and find the representative to train and predict the performance of the economy may be the small dishes to use split. Instead, we could divide by year.

Nonetheless, if we divide by year, we also get a chance to miss the yearly or cycled features since we take years as training by gap. Hence, there might be no best solution for the partition. The method we apply here is not scratches on earth to find the best split. Alternatively, if our model is good enough, we will keep apply the 70/30 conventional splits.

## 2. Data Mining Result Interpretation

There are more than 1 iteration run in our programme. But for simplicity reason, we only show first few in section 7. Iteration result would be presented in section 8.

Call back to figure 6-1 of this report, based on the set of selection of explanatory variables and the splits 70:30

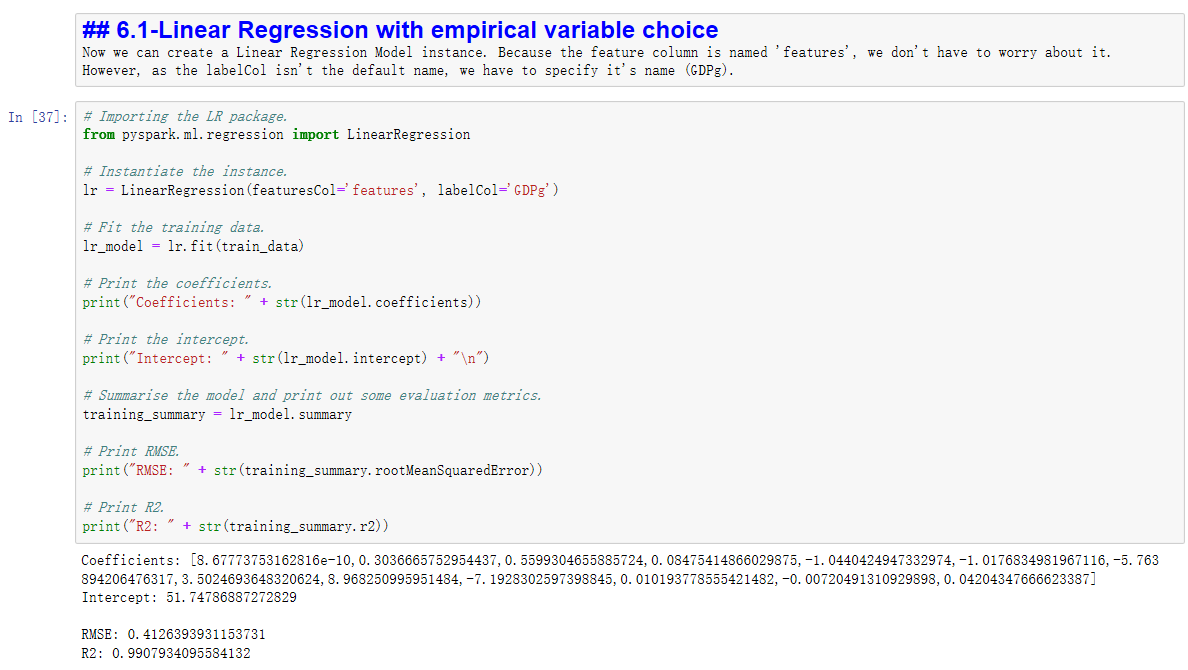


Figure 7-1 the result of Linear Regression by single year

The overall explanatory power for the training datasets is really good, with little RMSE is acceptable. The overall result looks good. The expected mean of the GDPs is some where close to 4%, which is the same as we expected in the mean of the values from the observed data points.

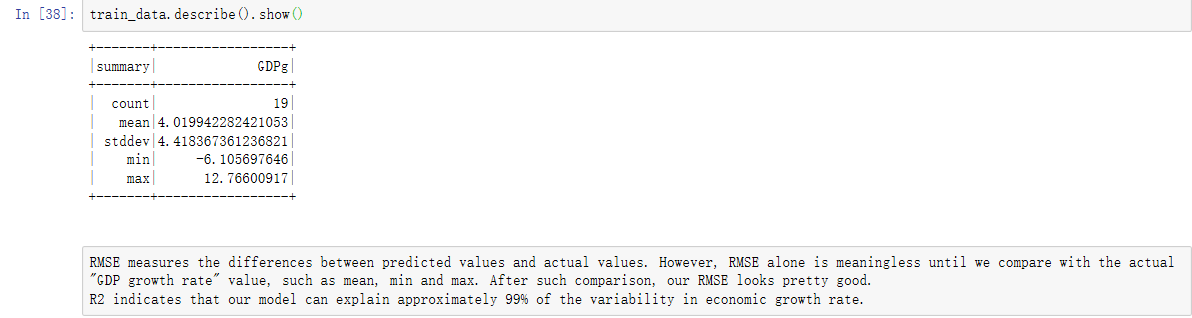


Figure 7-2 the result of Linear regression with 1990 data

Notice here carry out the explanatory variable analysis just in the year 1990, which may not be anything really help us answer our research question. The reason we apply this here is it provides a comparison for the cross-year analysis.

Then, we apply the linear model with data from all years. Call back to the combined data sets in 6-6, we apply these merged sets into our linear model:

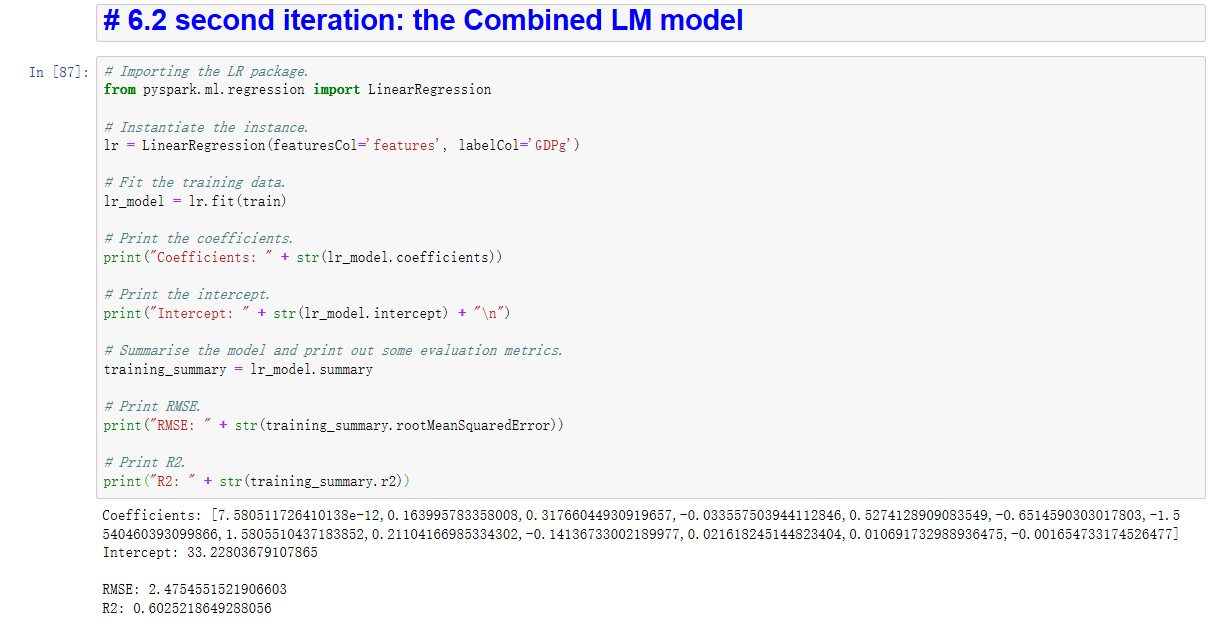


Figure 7-3 Linear Regressions based on the data from 1990 to 2017

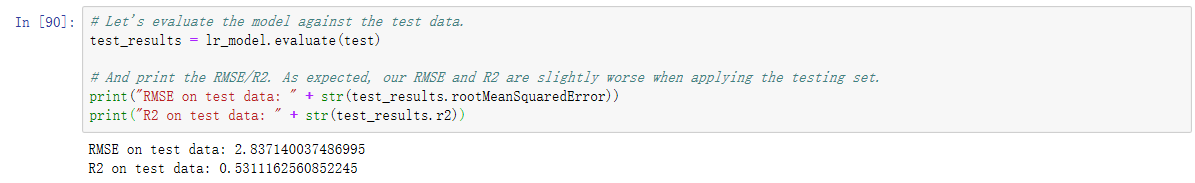


Figure 7-4 The result of Prediction by Linear Regressions Model (1990-2017)

The model two has a larger 2.48 RMSE and less R2. The R2 here represents the ratio of variation explained by the sets of our selected variables. The R2 decreases from nearly perfect 99.0% down to 60.3%. And the RMSE also increases from 2.48 to 2.84. This is a dramatic drop together with much more errors left behind. There is most likely to have some unknown yearly affect on the issues. The split and types of model seems have no problem since we got quite high prediction R2 in the first round. The suddenly decrease in the predicted result could be most likely caused by yearly affect. Although the decrease is huge, but 50% of the variation explained by our variables doesn’t mean we cannot accept it. But the result could be improved if we change the sets of variables which we hope could flatten out the cycle effect.

Then we apply a decision tree and random forest to check if there are relations between the countries and the explanatory variables. This is to check for the omitted correlations in the residuals.

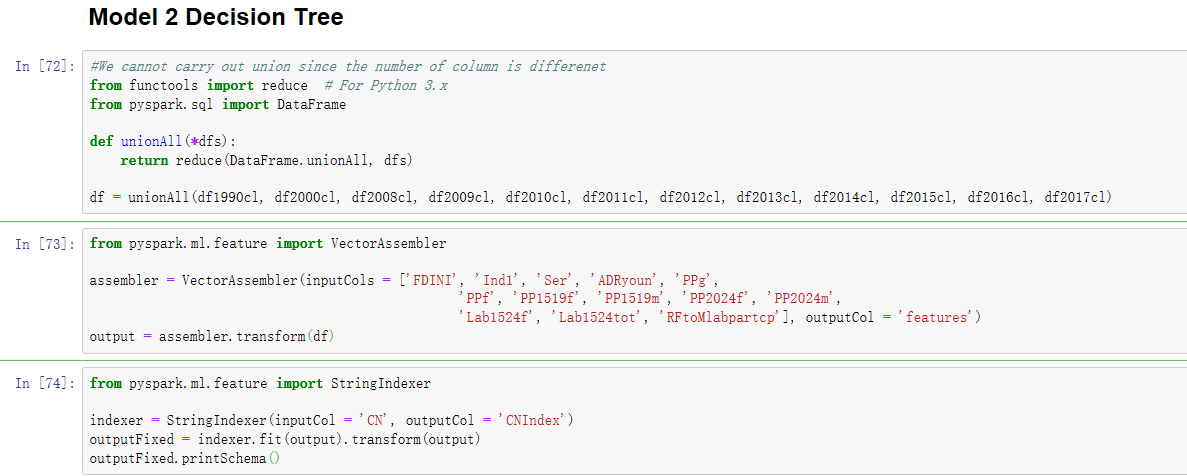
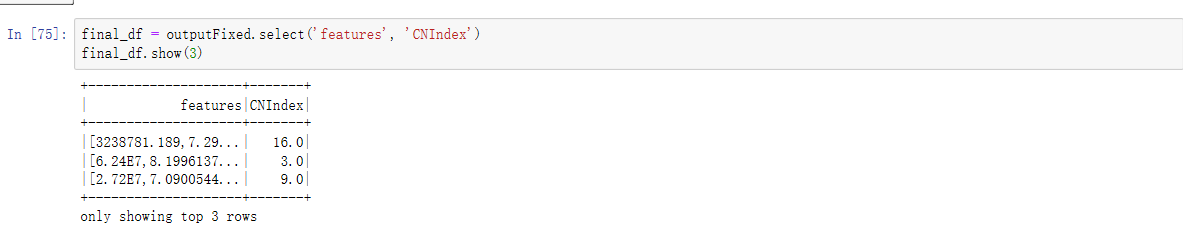


Figure 7-5 same classification for random tree construction

The variable selection here is the same as the previous selection in linear regressions. But we remove the GDPp and use the country name as the new response variable in order to see if there is other effect by countries.



We factorized the countries into three levels.



Figure 7-6 The Result of Classification Tree and the Random Forest

The result of decision tree is quite bad. This indicates the cross-country effect is not the major issues. The omitted factor is not a major issue. However, if we add the result of the random tree and the result of the R2 in linear regression prediction. The overall sum is much closer to 1 indicating the failure of explaining the variation of GDPs may indicating the changes to other factors through the identity of rows.

And if we change the response variables back to GDP growth rate. Similarly, we can construct trees in a similar process.





Figure 7-7 The Result of GDPg Tree partition

The Random Tree here has also weak predictor. This can be deduced by the overfitting problem. We will fix this through iterations.

## 3. Pattern searching

As mentioned in last section, we choose the regression model to reveal and testify our hypothesis. By applying the Occam’s Razor, we can deduce the unimportant variables.



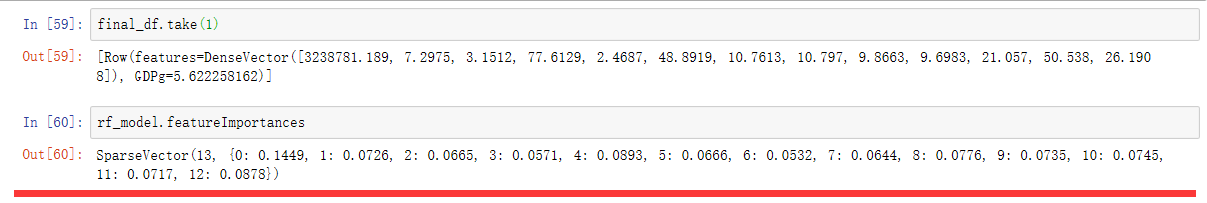


Figure 7-8 The Feature Selection Importance

The pyspark gives the importance of each variables.

|  |  |
| --- | --- |
| Variables | Importance |
| FDINI | 0.1449 |
| PPg | 0.0893 |
| RFtoMlabpartcp | 0.0878 |
| PP2024f | 0.0776 |
| Lab1524f | 0.0745 |
| PP2024m | 0.0735 |
| Ind1 | 0.0726 |
| Lab1524tot | 0.0717 |
| PPf | 0.0666 |
| Ser | 0.0665 |
| PP1519m | 0.0644 |
| ADRyoun | 0.0571 |
| PP1519f | 0.0532 |

Table 7-9 The importance of the feature sorted by importance

So, the most important three factors are foreign directed investment, followed by the population growth and the gender proportion in labour forces.

If we compared to the result through first and second iteration, we find that different software has different toleration of missing values could produce different combination of variables. For instance, in SPSS, we keep the youth unemployment rate because the high missing rate is acceptable by SPSS analysis. But in pyspark, the missing values will report error and stops us from finding its significance.

Another interesting thing to highlight is the result in SPSS is quite different. This is because we picked different factors in SPSS and the sets really affect the result of analysing.



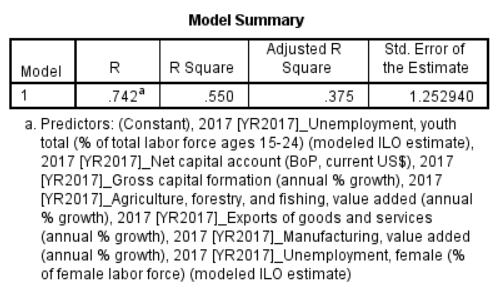


Figure 7-10 the end of year feature importance

In SPSS, the overall prediction power for the regression model is acceptable with R square 0.55, which is bit higher than what we get in pyspark. The reason we check this because although the gross R2 is low, the year start in 1990 has perfect result in R2. But the overall result deteriorates due to the bad estimation in 2017.

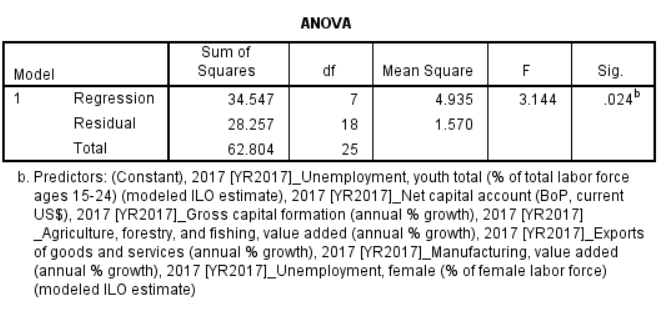


Figure 7-11 Error Check

The multi-correlation between variables seems okay, since the p value is small and the overall explainable variable is statistically significant. And ideally, the result seems to be robust because no matter on which platform, the overall result should be similar.

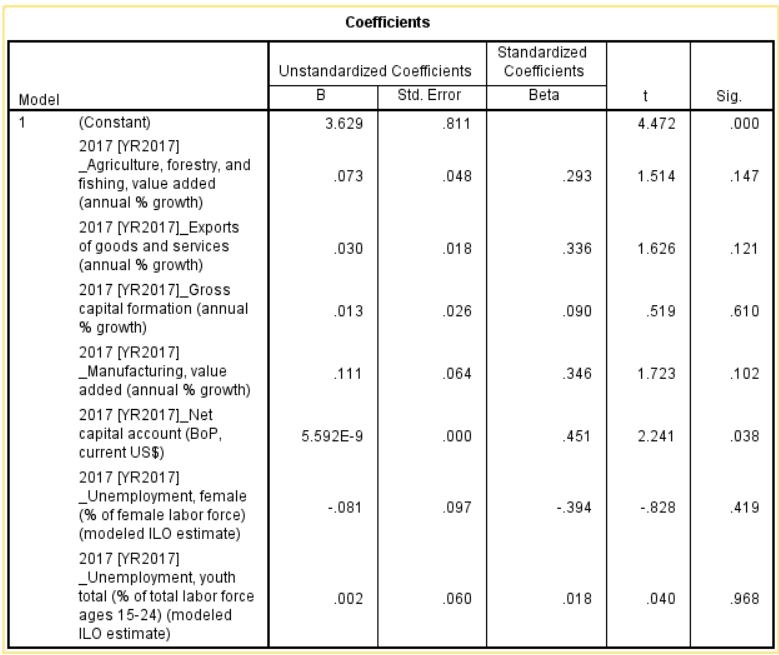


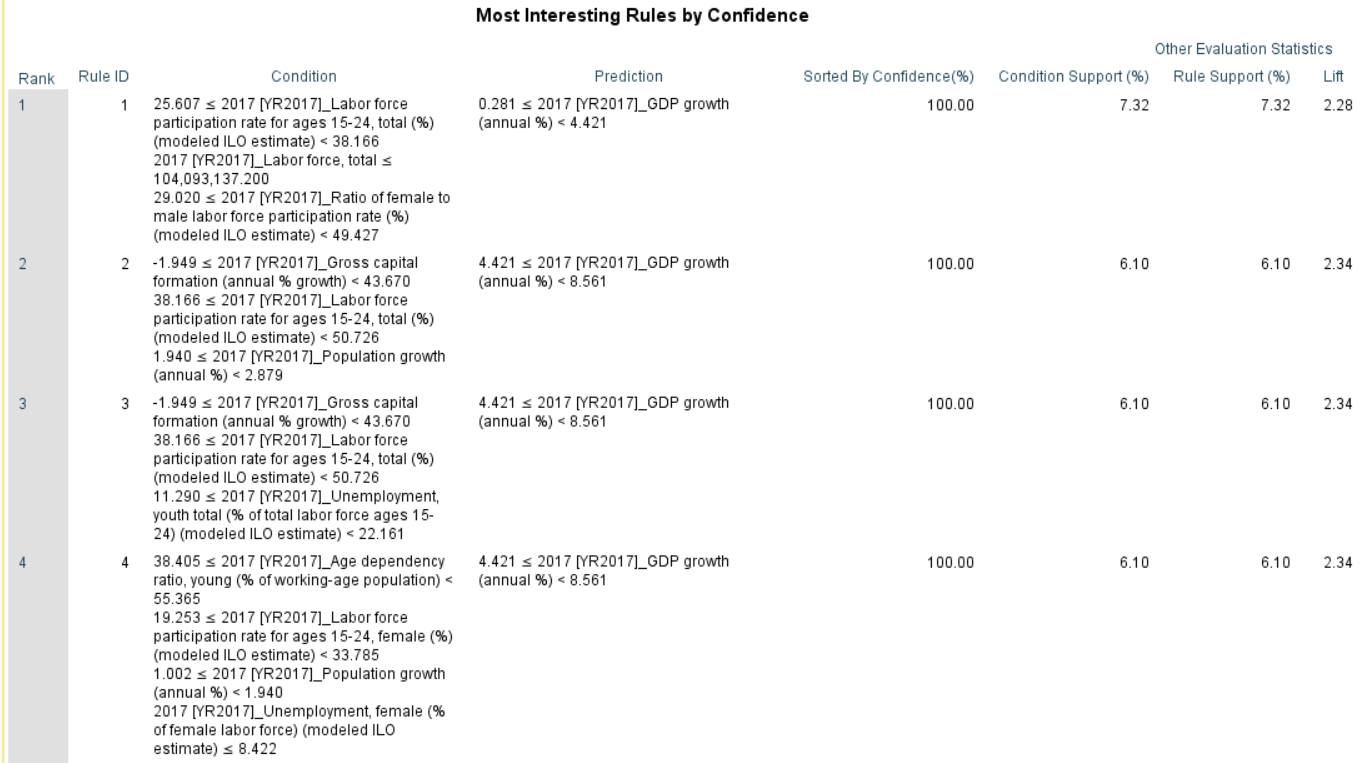
Figure 7-12 the last year coefficient

However, the result from last iteration is not good and have much bias left out in the constant. This time we are more experienced and the result will not easily miss out the major factors.

The improvement through iteration is that we are no longer to blame the software or the model predicted poorly. Instead, we know the partition, the choice of variables, could greatly affects the result and such act as a vital role in data analysis especially for the supervised study. What is important in one set may not have significant power in explaining the sets of variables.

Part 8 Interpretation

Our pattern is discovering association between our input variables and the target GDP growth rate (response variables). And by empirical knowledge we could classify our variables into certain group of different fields like industry fields and labour sector variable group, etc. But to evaluate this is not enough. Here we adapt association rules to discover any difference between the ability of prediction among these variables.

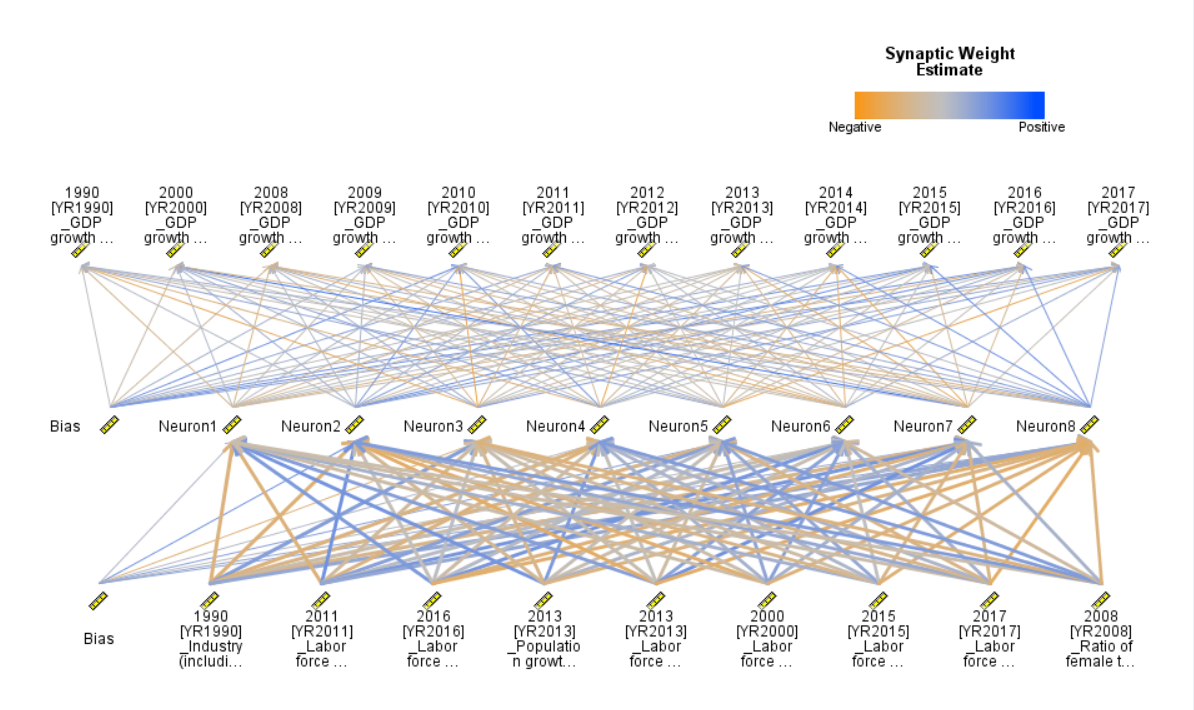


The labour force participation variable seems to be distinct from others in the sense it predicts much lower GDP growth rate while others remains the same confidence interval for prediction. This could be meaningful because the negative relation between the uneducated labour may actually decrease the efficiency in economy if more corporations hired more youth works.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rule Statisticsa,b** | | | | |
| Measurements | Minimum | Maximum | Mean | Standard Deviation |
|
| Condition Support (%) | 6.10 | 17.07 | 9.53 | 2.20 |
| Confidence (%) | 85.71 | 100.00 | 89.28 | 4.71 |
| Rule Support (%) | 6.10 | 14.63 | 8.47 | 1.91 |
| Lift | 2.01 | 2.34 | 2.09 | 0.11 |
| Deployability (%) | 0.00 | 2.44 | 1.06 | 0.46 |
| a. Number of Rules is 153 | | | | |
| b. Number of Valid Events Data Source Records is 82 | | | | |

The evaluation table shows the overall distribution patterns for our data ranged from 6.1% to 14.63 % increase for GDP, and the mean is bit over 7. But we know from the reality that is unlikely. So, the association rule maybe good to identify different groups of variables and discovery rules. But each rule may not predict accurate relationships.

There is a tradeoff between the models with more generality and accuracy. The neuro network model produces good result for prediction across the different years, but the neuro structure has interlayers with nodes we do not know its actual meanings. With these intermediate layers, all the significant variables could be insignificant since they only relate to intermediate level. Notice here the code in pyspark for Neuro Network is not working. Since the system keeping throwing the bugs, we take the similar result in SPSS procedure to help us improve through iterations.

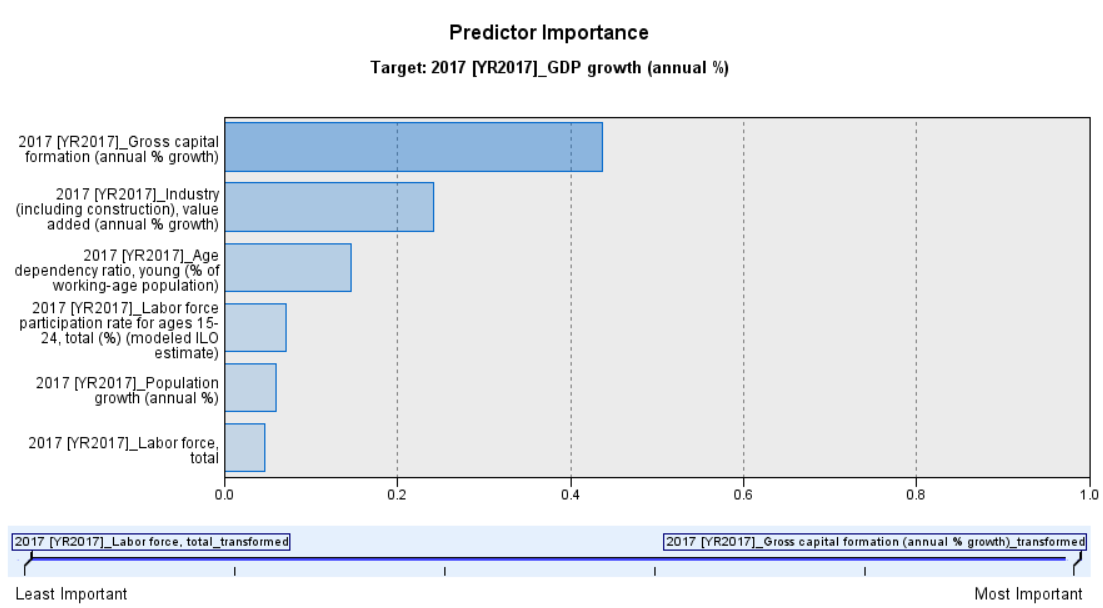


Some years the labor factor is significant and other year the population or industry affects more. But across year, we find the labor variables have more nodes than the rest.

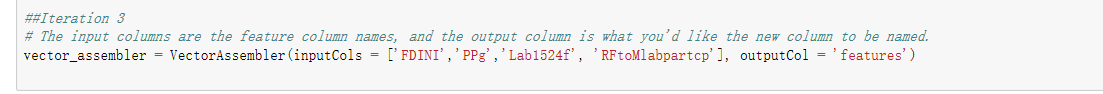
## Iteration Improvement

To fix this the drawback of the choice of combination of variables, we have to take the previous iterations and re-adapt the model again and again. Our previous iteration just did it and resulting a better importance through iterations.

Predicted by neuro network, we also went back to the massive variables’ iteration. In the linear regression model, we apply the new variables and achieve the overall 50% of explanatory power of the variation of GDP growth rate changes explained by the linear model. Note here all the extremes have been trimmed in this model.



The predictor importance of the variables improves to over 0.4 and the most important variables are mainly the capital industry and the labour force and population. But here we do not have the interest for capital. We are interested in our research question: whether the unemployment matters? So, we drop the replicated variables which belongs to the gender group and produce a very simple formula only with Capital, population and the distribution of labels and the gender ratio within labour force participation rate.



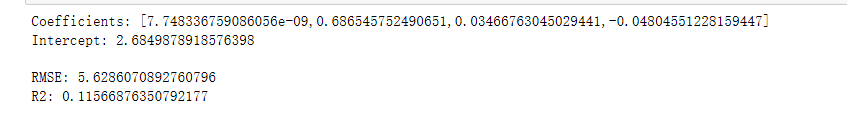


Figure 9-1 The Iteration 3 result (Not applicable)

This simplified set, however, result RMSE which is higher than the previous result and R2 only 0.12 which is not acceptable in real life application. After simplify my variables’ combination, there seem to be little explanation power left which is no good at all. This is because the importance table have very close importance of the variables. Although we already choose the different variables belong to different category, the result is still not presenting the overall variance of GDP numbers. The removing for the relatively important or insignificant may not be valid for multi variants models like this. So, we have to go back again to gird searching for other combinations.

But limited by the time and complexity of recombining the variable and times of trial. We back to the first combination given by the grid searching method which is more reliable by our human interpretation.

Other potential improvements could be trying other partition. But also, we have to sacrifices the time for something or even no gain eventually. The unsupervised model is then much more powerful to serve the best grid searching problem. Therefore, we take the NN models for granted.

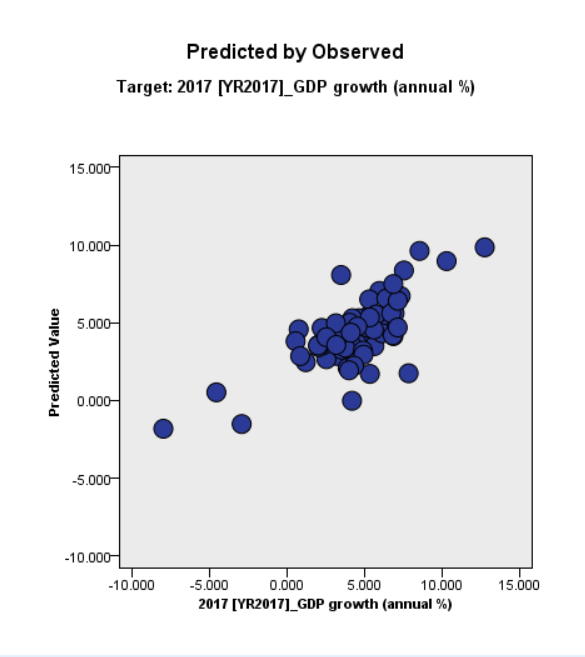
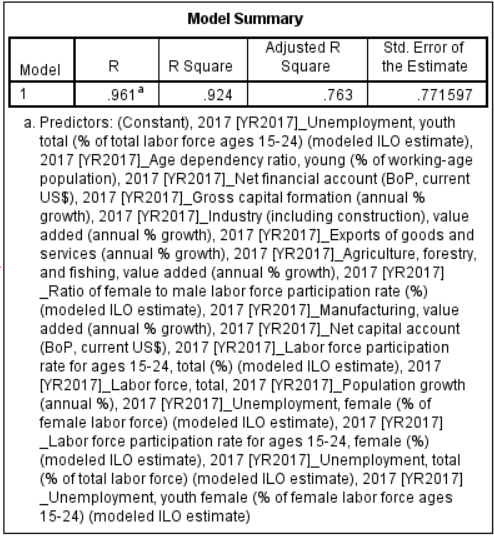


Figure 9-2 the prediction of GDP growth rate by linear regression model

The prediction by linear regression seems linear and cluster somewhere close to 5% for most cases. This coincide with the reality.

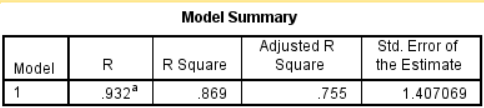
Deducing again the regression model, we can focus on the combination of the original massive variables’ set.



The total explaining power after refilter the data increases to 92.4% the variance of GDP growth rate. The coefficient shows the significance. The significant term is the agriculture, industries, and manufactural output (at ).

However, economists could argue these significant may contained by the GDP calculation itself. This may wash out the other effect. As a response, these industry sector and export data is removed to see the labour effects.

This result after removing the industry seems good enough for making our conclusion.



Our explained power decreases a bit to 86.9% after removing the industry output variables. But this is still better than expected and enough for carrying on our conclusion part.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficients** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | 2.435 | 3.302 |  | .738 | .472 |
| Age dependency ratio, young (% of working-age population) | -.236 | .050 | -1.503 | -4.697 | .000 |
| Gross capital formation (annual % growth) | .042 | .027 | .180 | 1.533 | .146 |
| Labor force participation rate for ages 15-24, female (%) (modeled ILO estimate) | -.178 | .097 | -1.175 | -1.837 | .086 |
| Labor force participation rate for ages 15-24, total (%) (modeled ILO estimate) | .197 | .092 | 1.038 | 2.141 | .049 |
| Labor force, total | 2.045E-8 | .000 | .700 | 2.007 | .063 |
| Net capital account (BoP, current US$) | 1.039E-8 | .000 | .459 | 3.507 | .003 |
| Net financial account (BoP, current US$) | 2.197E-10 | .000 | .597 | 1.542 | .144 |
| Population growth (annual %) | 3.985 | 1.066 | 1.234 | 3.739 | .002 |
| Ratio of female to male labor force participation rate (%) (modeled ILO estimate) | .073 | .035 | .610 | 2.056 | .058 |
| Unemployment, female (% of female labor force) (modeled ILO estimate) | .717 | .217 | 1.918 | 3.297 | .005 |
| Unemployment, total (% of total labor force) (modeled ILO estimate) | -.955 | .294 | -2.122 | -3.252 | .005 |
| Unemployment, youth female (% of female labor force ages 15-24) (modeled ILO estimate) | -.482 | .165 | -2.267 | -2.925 | .010 |
| Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate) | .596 | .245 | 2.366 | 2.432 | .028 |

At level, labour force participation rate for 15-24, net capital account, population growth, unemployment rate for female, youth female, youth total and all total devote to affect the overall performance on GDP growth rate.

Based on the grid search, the regression model reveals the significant variables. Among them, we can see the trend:

The employment rate for youth overall and females, population growth, participation rate for young worker, and net capital account have positive devotion to the country’s GDP growth rate, while age dependency ratio, unemployment rate have a negative impact on the GDP growth rate.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name of significant variables** | Unstandardized Coefficients |  | Standardized Coefficients |  |  |
| **B** | **Std. Error** | **Beta** | **t** | **Sig.** |
| % of total labor force ages 15-24) (modeled ILO estimate) | .596 | .245 | 2.366 | 2.432 | .028 |
| Unemployment, female (% of female labor force) (modeled ILO estimate) | .717 | .217 | 1.918 | 3.297 | .005 |
| Population growth (annual %) | 3.985 | 1.066 | 1.234 | 3.739 | .002 |
| Labor force participation rate for ages 15-24, total (%) (modeled ILO estimate) | .197 | .092 | 1.038 | 2.141 | .049 |
| \_Net capital account (BoP, current US$) | 1.039E-8 | .000 | .459 | 3.507 | .003 |
| \_Age dependency ratio, young (% of working-age population) | -.236 | .050 | -1.503 | -4.697 | .000 |
| \_Unemployment, total (% of total labor force) (modeled ILO estimate) | -.955 | .294 | -2.122 | -3.252 | .005 |
| \_Unemployment, youth female (% of female labor force ages 15-24) (modeled ILO estimate) | -.482 | .165 | -2.267 | -2.925 | .010 |
| (Constant) | 2.435 | 3.302 |  | .738 | .472 |

To conclude, the relation we found in this assignment has verify our hypothesis that the higher employment rate would promote the economic growth and unemployment would impede the growth. And we also highlighted that the higher proportion of young people in the job market, the lower expected GDP growth it would be. This can be interpreted as low efficiency for these kinds of workers due to the lack of skills which should be obtained in school. And early working age may set a lower ceiling of the productivity for individual could achieve in the job market.

Recall the familiar Solow Model for economic growth, the classic term involved with capital, knowledge and labor force. Here we can depict the roughly model by the result of regression:

This result shows that the employment of youth and female indeed has a positive impact on economic growth, and the number of population and labor force of youth and capital also devote to economic growth. While the age dependent ratio of young people (ADR) and the unemployment of total and youth female have a negative relation with economy growth. Therefore, these results indicate a higher ratio of youth out of work (back to school to improve the quality to knowledge) and larger available ratio of labor force could stimulate the growth of the economy. Less unemployment for the entire labor force could also promote to a higher economic growth rate. The government should subside the basic education facility and guide the young people to gain knowledge rather than entering the labor market too early. But the unemployment of youth female seems to have positive relation may indicates female may not gain much from education. This is contradicted to our common sense and need further check if it is a bias or something else.

And the sophisticated and detailed analysis through all the iterations will be present in the final report.

1. SDG8.1, <http://datatopics.worldbank.org/sdgatlas/SDG-08-decent-work-and-economic-growth.html> [↑](#footnote-ref-1)
2. Atlas of Sustainable Development Goals 2017, P45, 8b [↑](#footnote-ref-2)
3. Mankiw, N.G., Romer, D., Weil, D.N., 1992. A contribution to the empirics of economic growth. Quarterly Journal of Economics, Vol. 107 (2), pp. 407–437. [↑](#footnote-ref-3)
4. Nunneman, W., Vanhout, P., 1996. A further augmentation of the Solow model and the empirics of economic growth for the OECD. Quarterly Journal of Economics, Vol.111, pp. 943–953 [↑](#footnote-ref-4)
5. Knowles, S., Owen, P.D., 1995. Health capital and cross-country variation in income per capita in the Mankiw–Romer–Weil model. Economics Letters, Vol. 48, pp. 99–106 [↑](#footnote-ref-5)
6. Breton, Theodore R, 2004. Can institutions or education explain world poverty? An augmented Solow model provides some insights. Journal of Socio-Economics, 2004, Vol.33(1), pp.45-69. [↑](#footnote-ref-6)
7. This young age standard is given by the world bank. [↑](#footnote-ref-7)
8. Atlas of Sustainable Development Goals 2017, P44 Goal 8 [↑](#footnote-ref-8)
9. https://data.worldbank.org/ [↑](#footnote-ref-9)
10. By the World Bank Atlas method, 2017, where the low-income represents GNI per capita less than 995$; the lower middle income counties between 996$ and 3895$. [↑](#footnote-ref-10)
11. No direct employment and age number available in developing countries, instead, we take a series of Labour force/population ratio, participation ratio indexes for employment, age distribution for age as a replacement. [↑](#footnote-ref-11)
12. <https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_541211.pdf>, page 23, table 4. [↑](#footnote-ref-12)