Final Prject

Xun Zhang

12/2/2019

Abstract

The world has witnessed a rapid growth in service industry. The focus of this project is to use the rate of male employment in service as an indicator of the degree of prosperity in service industry. We make use of R with the tool of data wrangling, data visualization and machine learning to study the relationships between variables and employ model linear regression to make prediction.

Key Words: Service industry, World Bank, employment, economy, development

Introduction

My goal is to find proper predictor for the response variable male employment rate in service. Is the model useful and helpful? That is, is the independent variables I picked that provides information in other areas better than the simple predictor, the mean of datas over last several years? How strong the relationship is to the target? In this report I will go through the background, data, analysis and results to present an overview of my project. In each part I would like to recruit the actual content like the data and graph. And finally I evaluate my work in all and point out what can be improved next time.

Problem Statement and Background

Now I am interested in service industry. I decide to take the rate of male employment in service as an indicator of the degree of prosperity of service industry, meaning the higher the rate of male engaging in service is, the more developed the service industry is in this country. I isolate my project from gender discussion(like the comparison of the rate between males and females). The target - the rate of male employment I have chosen is just one of the interpretations for the state of service industry in every country. Also, I believe other things like the volume of production yielded in service industry will work in the same way. Other categories of data are selected intuitively, but not randomly. They are rural population, GDP, population density, international tourism and secure internet server. First, I thought the development of service is closely related urbanisation. When rural population drops down, that means more and more people are freed from farming (might be the result of mechanized harvesting) and turn to work in cities and urban-rural fringe area. Much open labor force for the service will be provided due to the spatial distribution of population. Thus, at first glance rural population is a meaningful variable to be taken into consideration. Second, population

density. The assumption I made is that the more population is concentrated, the clearer the division of labor is and the more people are needed to serve this huge population. Thus, the industry is promoted. Third, GDP is a straightforward indicator of the development state of one country. Experience has taught us that the when a country's economy achieves a lot, its service industry will reach a higher level, too. Forth, International tourism. The industry of tourism has flourished these years, especially when it comes to traveling abroad. From my perspective, international tourism will boost the service industry and consequently result in higher rate of male employment in service.

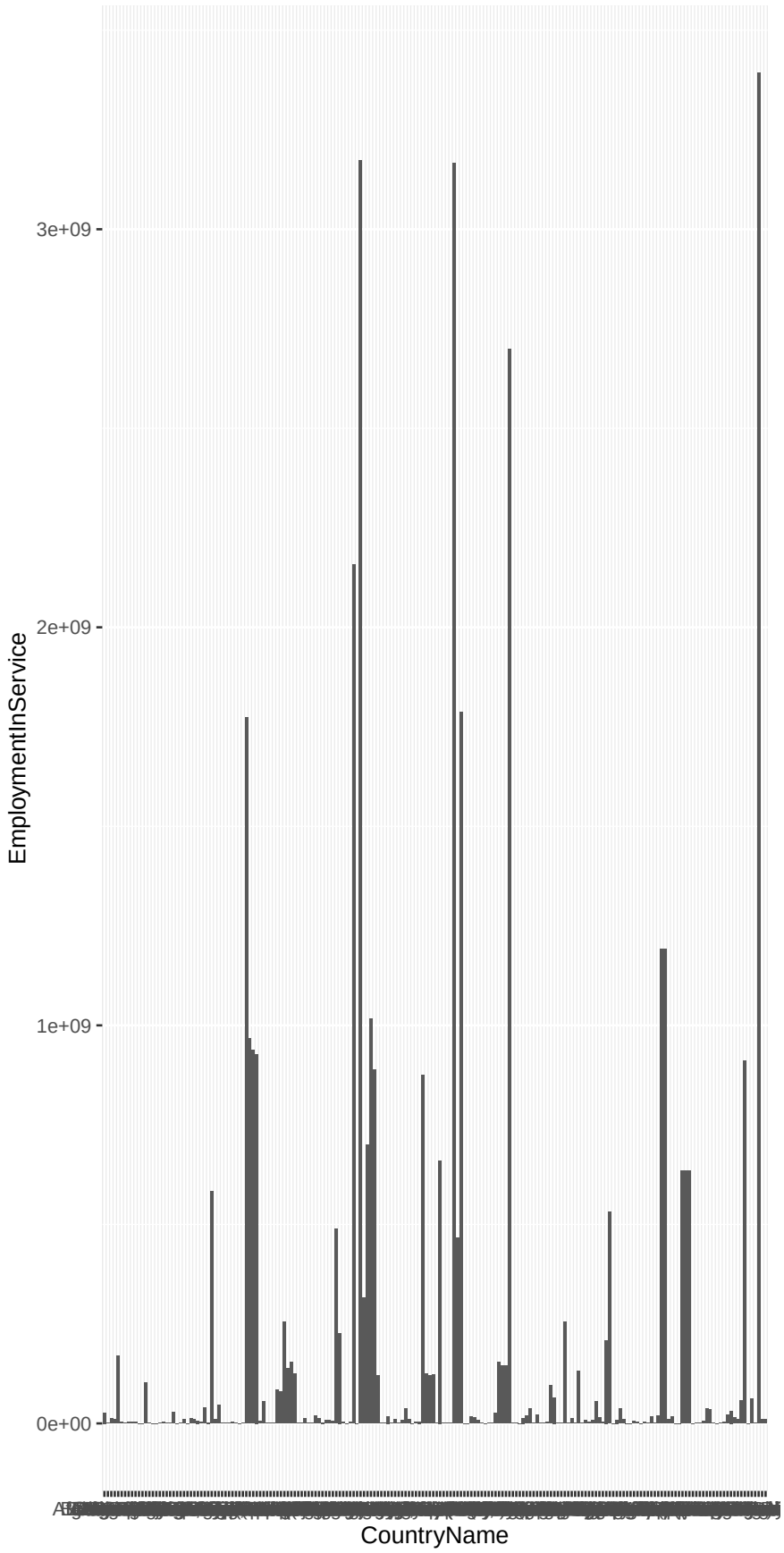
Data

I downloaded the data I need from the database of Wold Bank website, which has huge and reliable database.

All charts are saved in CSV format, which walk me through the input work with much ease.

Data Wrangling The first thing I need to do is to install and require the package tidyverse. I use function Select() to pick up data from 2013-2017, because I need unified time frame for all categories of data I have. Select() helps me to keep only the variables I mention. Then I reshaped my chart by using function Gather(), which takes multiple columns and collapses into key-value pairs. Next, I employ function Full_Join() to line up my variables. It helps me to merge my frames. Finally, I drop NA by applying function Omit_NA to make it look cleaner and pave the way for next steps. ## Analysis There three main aspects in my work: First, the correlation study of the relationship between variables. Second, cross validation when training model, the reason why I did this is to reduce the bias. Third, model linear regression to predict the rate of male employment in service. The reason why I want to create models is to analysis the relationship of variables by calculating the correlation

```
##   CountryName Year InternationalTourism EmploymentInService
## 1 Afghanistan 2013            1.3487539             24404522
## 2      Angola 2013            0.6397658              9885476
## 3     Albania 2013           26.0776739              1291587
## 4  Arab World 2013            7.7883120            161751073
## 5   Argentina 2013            9.3253954              3693179
## 6     Armenia 2013           18.3880040              1067354
##   PopulationDensity SecureInternetServer RuralPopulation      GDP
## 1          49.42804             1.425491    2.056105e+10 27.61100
## 2          20.86771             4.574147    1.367100e+11 41.83300
## 3         105.66029            35.577453    1.277628e+10 36.61800
## 4          33.80379            20.498471    2.867270e+12 58.44397
## 5          15.42116            73.857422    5.520250e+11 90.85700
## 6         101.77675            54.528186    1.112147e+10 48.78600
```


3e+09
2e+09
1e+09
0e+00
EmploymentInService
CountryName

Results

As it shows in the graph, male employment in service spreads very unevenly across countries all over the world. There exists huge difference between countries. Some are at high level, but most are at low level. The result is the same as my expectation. The problem here is that I have too many countries at hand, even when I chose to only select the year of 2017, their names still crowd together under the horizontal axis. This may negatively affect the effectiveness of the information communication with message recipients, which goes against the intent of why we decide to make use of data visualization.

Table 1: Data summary

Name	train
Number of rows	716
Number of columns	8
Column type frequency:	
character	2
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CountryName	0	1	4	52	0	208	0
Year	0	1	4	4	0	5	0

Variable type: numeric

Table 3: Data summary

skim_variable	n_missing	complete_rate	mean	sd	p0	
InternationalTourism	0	1	6.150000e+00	3.75000e+00	0.32	3.690
EmploymentInService	0	1	1.833887e+08	5.27788e+08	0.00	1.063
PopulationDensity	0	1	3.569400e+02	1.81714e+03	1.85	3.533
SecureInternetServer	0	1	1.772470e+03	5.32727e+03	0.17	1.408
RuralPopulation	0	1	3.360290e+12	1.01577e+13	302925488.10	1.479
GDP	0	1	6.207000e+01	2.41300e+01	2.97	4.204

Name	test
Number of rows	304

Number of columns	8
Column type frequency:	
character	2
numeric	6
Group variables	None

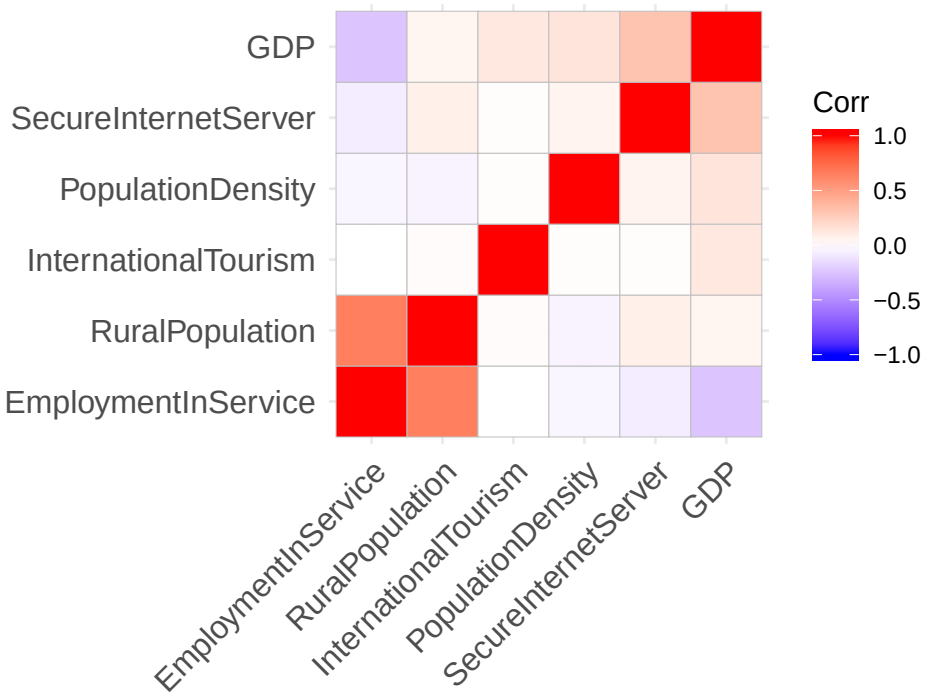
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CountryName	0	1	4	52	0	172	0
Year	0	1	4	4	0	5	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	
InternationalTourism	0	1	6.690000e+00	4.470000e+00	0.58	3.75
EmploymentInService	0	1	1.779583e+08	5.104278e+08	0.00	1.06
PopulationDensity	0	1	1.819200e+02	6.170000e+02	1.89	3.22
SecureInternetServer	0	1	2.059930e+03	6.802280e+03	0.15	1.37
RuralPopulation	0	1	2.634841e+12	7.254206e+12	318265835.30	1.92
GDP	0	1	6.249000e+01	2.435000e+01	12.75	4.10

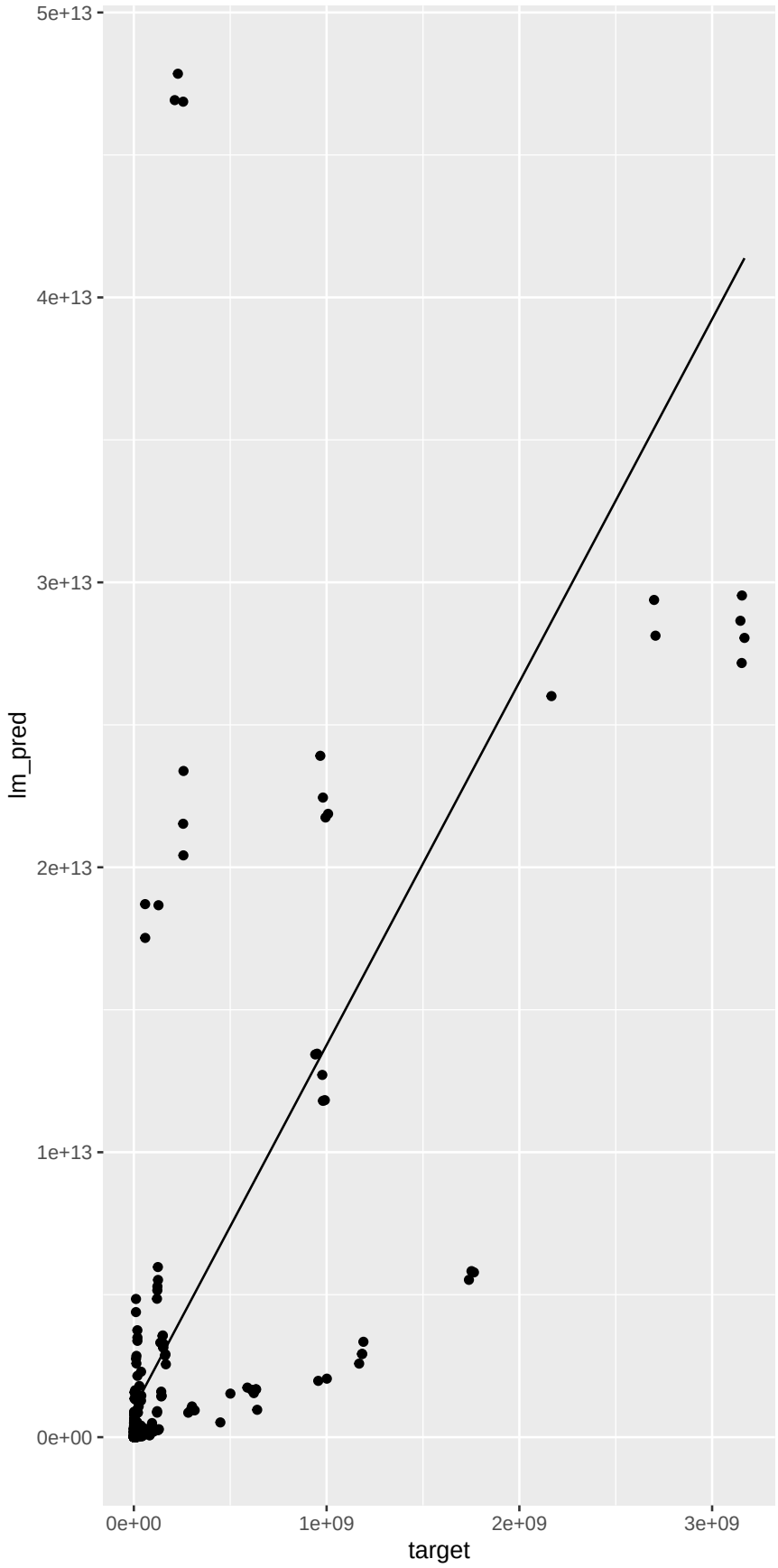
Because I am going to see how my data looks and want to explore it, I try to split them randomly, assuming 70% as train data, and the rest 30% as test data. Here comes the result. The left is predicted value, the right is true value. The correlation is one of our aims for the project.



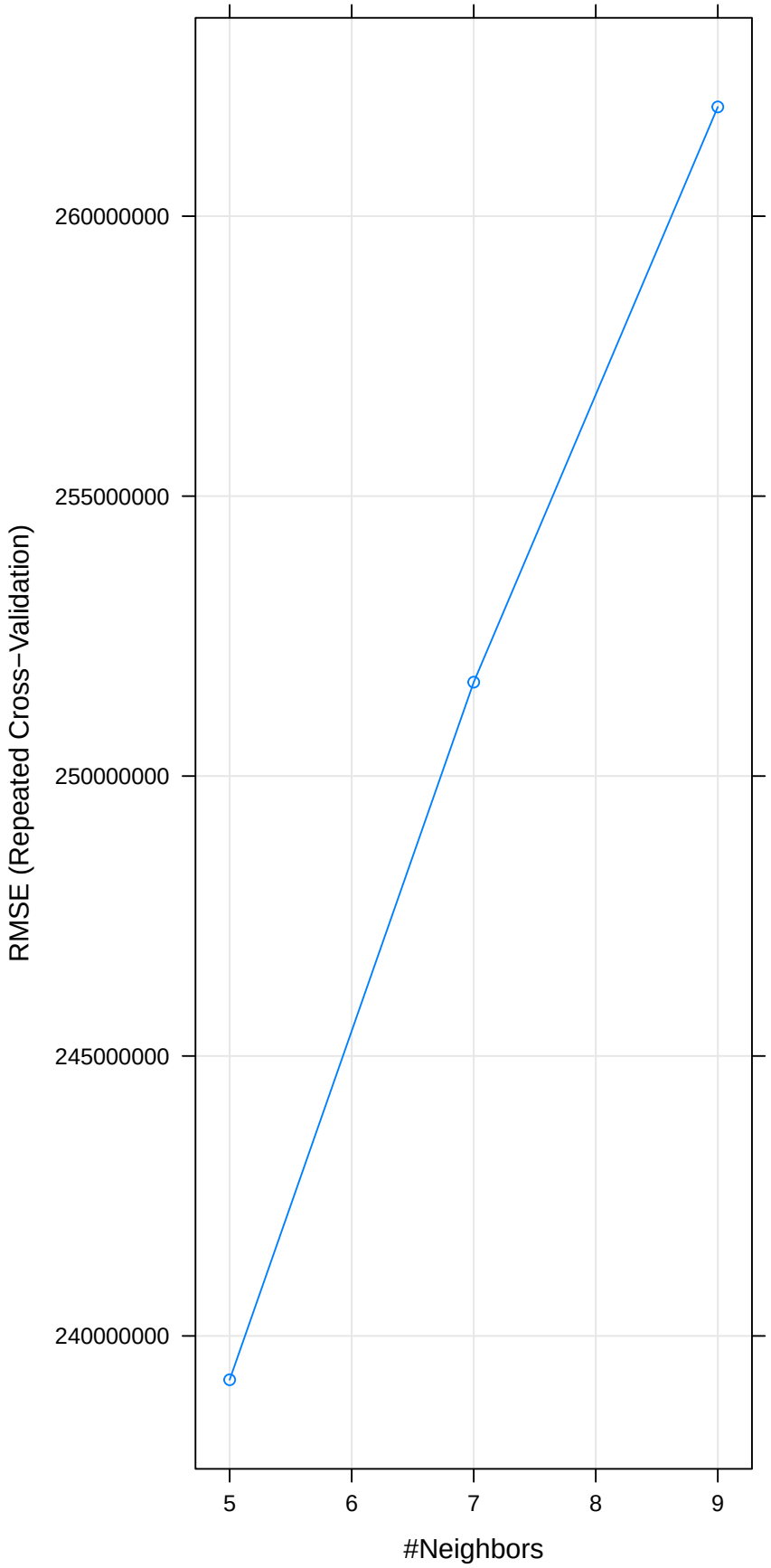
At the very first, we need to omit country name and years, because they belong to class - character rather than numeric. In the plot, the darker the color, the stronger the correlation. In the scale from -1 to 1, corresponding negative relationship and positive relationship. As we could tell from the results show on the plot above, rural population is in obvious correlation with the rate of male employment in service. The possible reason behind this has been stated in previous section: people who are attracted from rural area are likely to take on work in service in cities (this remains to be examined in the future work). As we assumed before, GDP also in a relatively strong relationship with the target and the connection between the male employment in service and secure Internet server is poor. What surprise me most is that international tourism and the target we set not really correlate with each other (if we are still interested in the topic of tourism, it might be helpful to have a look at the domestic). Because the correlation is strongest between the rate of male employment in service and rural population, in the next we will focus on rural population.

At the start, I use the ten-fold cross validation on the training set. The process will be repeated five times to complete. The reason why to do this is to reduce the bias of dataset and make it better.

```
##         lm_pred    target
## 4  3.084668e+12 161751073
## 6  1.037859e+12   1067354
## 7  1.066975e+12   3353116
## 18 1.048397e+12   1894621
## 25 1.029949e+12    446392
## 26 1.033308e+12    710090
```


5e+13
4e+13
3e+13
2e+13
1e+13
0e+00
lm_pred
0e+00
1e+09
2e+09
3e+09
target

For most columns of my data are numeric, and target column is numeric, too, I decided to create regression model. Each time we use 644 or 645 observations to train the model. The line is Linear regression. We can tell that train data does not have a good performance here. Rural population and the rate of male employment in service are not quite linearly related to each other. As shown in the graph, there are some point locating very far away from the line we drew.


RMSE (Repeated Cross−Validation)
260000000
255000000
250000000
245000000
240000000
5
6
7
8
9
#Neighbors

Discussion

In this part I am going to talk part how successful is my project. With the help of models and plots in R, we have figured out how strong the relationship is between the variables I picked. Let's look at the disadvantages first. At beginning, the year of data I use is from 2013 to 2017. Thus, the timeliness is not as good as I expected before I stated this project. Second, the tool I wanted to use but failed exists, that is random forest. Until the end I have no idea why the model can not run and export the result. At first, I thought it was because of NA(not applicable values) that I didn't clean up. Then I wondered if the data type is the reason, like I by mistake did not transform all character into numeric. I gave it a try and wanted to fix the problem, but it did not work out. This hurts the efficiency of my project. All in all I have achieved what I set out to do before the project. The correlation between variables are examined and the model works well. Where to go next: In the future work, I would like to add more variables into my work, put them into the model.