Final Report for NeverGiveUp Team

Jinhan Cheng, Ming Li

12/15/2017

Contents

1 Introduction: Trade Around the World

Trade is an whole-world concept, it could be explained by macroeconomics and other economics theory. In our project, we want to investigate and find out the relationship between goods, whether they are complementary or substitute. For example, tea and coffee are substitute goods, customers often buy one kind of them. As for the relationship between milk and tea, or milk and coffee, they could be complementary goods, for those who what to have a bottle of bubble milk tea or like to have a cup of coffee with milk. We can track into the data to find these similar relationship between other goods. After doing this, we could offer some commercial suggestions to companies who sell those products, and to strike a balance between different target customers.

We also want to find out the tendency of trading, for example we want to know which product are becoming more popular among international trade, apart from what is universally known, personal digital products for instance. It is important to find some potential commercial opportunities that many companies may ignore.

And for a specific country, we are trying to compute its trading structure and to find some potential relationship between other indicators. If the result of our research is strongly concerned with one or more indicators of this country, we could offer suggestions for them to help build a better and healthier society.

2 Sources of Data

We obtain the data from these websites:

2.1 United States Census Bureau

Firstly, for the main data, we obtain the information about trade volumn between the United States and other countries since year 1996 till now from *United States Census Bureau*.

2.2 World Trade Organization

Secondly, for the target goods, we've done some text mining on the reports of *World Trade Organization*. We download reports since year 2007. Each report includes the trading situation of that year and the trade in this globalizing, including the causes, distributional consequences of trade and policy implications of global integration and WTO etc.

2.3 The World Bank Organization

Finally, we use data from *The World Bank Organization*. We uses indicators such as the ratio of death caused by communicable diseases and nutrition conditions, GDP per person employed, prevalence of overweight(% of children under 5), prevalence of undernourishment(% of population), wage and salaried female workers(% of female employment), wage and salaried male workers(% of male employment) and the population growth rate.

The resourses of data are reliable because they come from reliable data bank. As for accuracy, they mostly meet our standard, besides some of the data has missing value, and the country names for the same country code has different formats. Different data has different responsibility to serve. All of the data offer information about year, the first and the last one provide us with further information about countries. And for the first one, it gives us more about the trading data including. As for the last resourse, it gives us country code, region and income level of a country, and a country's life index. In this way, all sources of data are representative of the population we are studying.

3 Examination of the Data

3.1 Data obtained from *United States Census Bureau*

We check the data and it is of good quality. The data structure is as below:

```
colnames(dat3)
```

```
 [1] "Year"                          "SITC"
 [3] "sitc_sdesc"                    "Country"
 [5] "ExportsFASValueBasisJan"       "GenImportsCustomsValBasisJan"
 [7] "GenImportsCIFValBasisJan"      "ExportsFASValueBasisFeb"
 [9] "GenImportsCustomsValBasisFeb"  "GenImportsCIFValBasisFeb"
[11] "ExportsFASValueBasisMar"       "GenImportsCustomsValBasisMar"
[13] "GenImportsCIFValBasisMar"      "ExportsFASValueBasisApr"
[15] "GenImportsCustomsValBasisApr"  "GenImportsCIFValBasisApr"
[17] "ExportsFASValueBasisMay"       "GenImportsCustomsValBasisMay"
[19] "GenImportsCIFValBasisMay"      "ExportsFASValueBasisJun"
[21] "GenImportsCustomsValBasisJun"  "GenImportsCIFValBasisJun"
[23] "ExportsFASValueBasisJul"       "GenImportsCustomsValBasisJul"
[25] "GenImportsCIFValBasisJul"      "ExportsFASValueBasisAug"
[27] "GenImportsCustomsValBasisAug"  "GenImportsCIFValBasisAug"
[29] "ExportsFASValueBasisSep"       "GenImportsCustomsValBasisSep"
[31] "GenImportsCIFValBasisSep"      "ExportsFASValueBasisOct"
[33] "GenImportsCustomsValBasisOct"  "GenImportsCIFValBasisOct"
[35] "ExportsFASValueBasisNov"       "GenImportsCustomsValBasisNov"
[37] "GenImportsCIFValBasisNov"      "ExportsFASValueBasisDec"
[39] "GenImportsCustomsValBasisDec"  "GenImportsCIFValBasisDec"
[41] "ExportsFASValueBasisYtdDec"    "GenImportsCustomsValBasisYtdDec"
[43] "GenImportsCIFValBasisYtdDec"   "CTY_CODE"
```

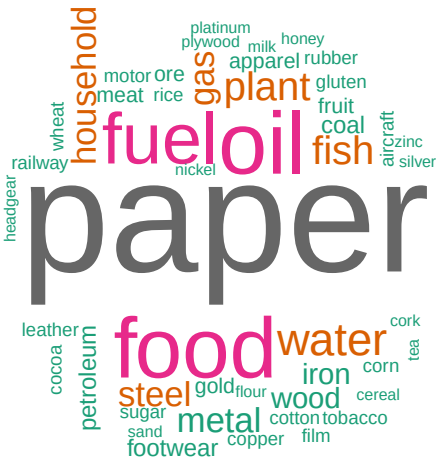
SITC includes the goods, and *sitc_sdesc* has the codes for each goods. We use data of the total trade of a year, so we use the *41th* and *42th* columns.

3.2 Data obtained from *World Trade Organization*

As for information obtained from WTO reports, we do some text mining of them. We keep the names of *goods* in our data table and build a simple world cloud of them to serve as a resource for us to target the objectives we want to do research on.

To achieve this, we want to do some text processing of the *goods* we have from *USCB*. Fisrtly we combine each product's name and remove all the punctuation and numbers, at the same time we make them lowercase. Secondly we split the long sentence we've created into short words, then remove those words whose length are less then 3. Thirdly we match the texts we've chosen with the texts from the *WTO* reports and get the frequence table. The result of the whole 11 years since year 2007 is as below:

```
wordcloud(text.cleaned$Term, text.cleaned$Frequence, random.order = FALSE,
    scale = c(5, 0.5), colors = brewer.pal(8, "Dark2"))
```



Finally, we choose goods from three categories to inverstigate: cloth, cereal and drinks. The specific goods we want to deal with are as followed:

```
s.keywords
```

```
 [1] "Silk"          "Cotton"        "Wheat"         "Rice"
 [5] "Barley"        "Maize"         "Other.Cereals" "Coffee"
 [9] "Cocoa"         "Tea"           "Soft.Drinks"   "Alcohol"
[13] "Milk"          "Juices"
```

3.3 Data obtained from *The World Bank Organization*

Each of the data of the indicators includes 3 data files, one is the main data, one is a description of the country including its region and income level, the last one is just an introduction. We use only the first two data files. When examining these data, our main concern is that there exist some missing values when we decide to directly merge the first with the second data(later we will talk about merge it with data from *USCB*). Most country names in these two data files share the same formats, some of them are written in different ways however. So we merge them using country code instead because code are more universally used and can shared by different formats of names. After this, we've gained data structure like this:

```
colnames(total.keydat)
```

```
 [1] "Country"       "Year"          "Type"          "Silk"
 [5] "Cotton"        "Wheat"         "Rice"          "Barley"
 [9] "Maize"         "Other.Cereals" "Coffee"        "Cocoa"
[13] "Tea"           "Soft.Drinks"   "Alcohol"       "Milk"
[17] "Juices"
```

When merging with the data we gain from *USCB*, another problem is that there are no country codes in it, and also some country names are written in different ways from *WBO*. To deal with this problem we decide to do text processing using grep function. For example, there are *WorldTotal* in *USCB* data and *World* in *WBO* data, we use text processing to match them and add the related country code to the former data. After this, the data structure has become like this:

```
colnames(total.keydat)
```

```
 [1] "Country Code"  "Country"       "Year"          "Type"
 [5] "Silk"          "Cotton"        "Wheat"         "Rice"
 [9] "Barley"        "Maize"         "Other.Cereals" "Coffee"
[13] "Cocoa"         "Tea"           "Soft.Drinks"   "Alcohol"
[17] "Milk"          "Juices"        "Region"        "IncomeGroup"
```

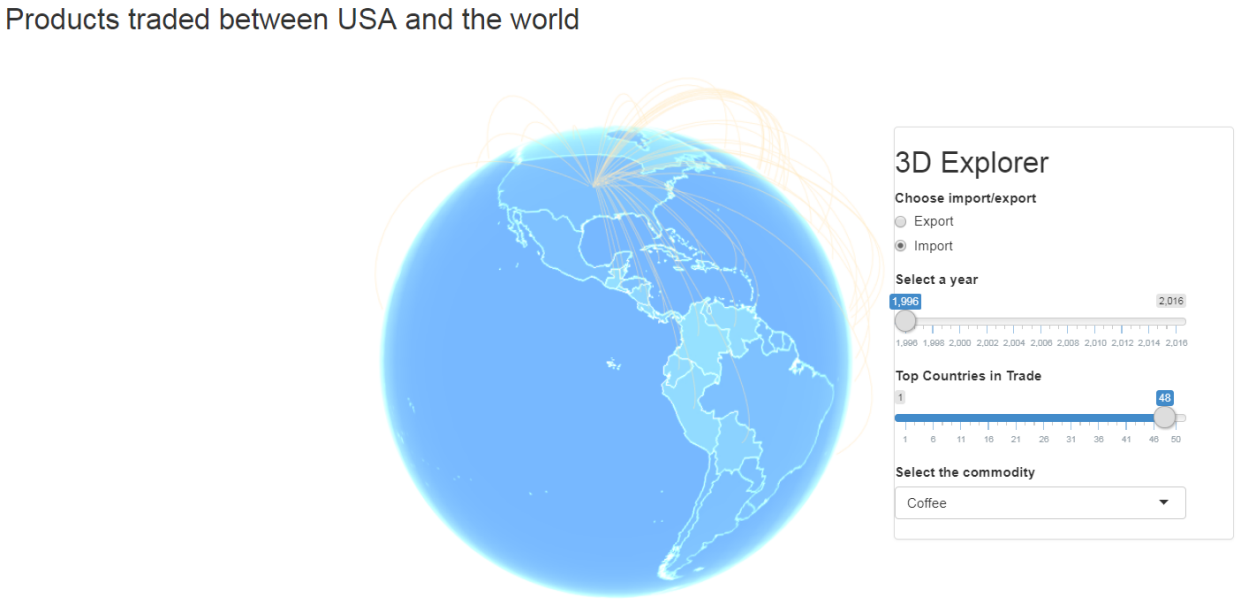


Figure 1: 3D Globe Panel, powered by Google Map API

4 Investigation and Interpretation

4.1 Data Visualization

4.1.1 Globe Panels

As the data has many dimensions in countries and the products, so the first thing we need to do is to get a more explicit visualization of the data. Inspired by the data visualization example in *BeautifulData* book, we decided to combine the spacial data with the trade data, then using an embedded JavaScript script to power the interactive part.

```
kable(head(input_data[1:7], 3))
```

	Year	Country	longitude	latitude	type	value	To
1	1996	WorldTotal	77.69170	12.97985	Export	342906629	US
2	1996	Canada	-106.34677	56.13037	Export	35594313	US
4	1996	Guatemala	-90.23076	15.78347	Export	464471	US

From the chart above, we can see that we combine the basic spacial data, i.e. the longitute and latitude value to each country and then use the points for plot of lines connects between countries.

Thanks to Javascript supported shiny app, we are able to see when it comes to the specific trade type(Expoer or Import) for some specific product, the top largest trade partner for one specific year. From the rotatable 3D Globe in the Panel, we can clearly find which countries are the largest trade partners with USA and the spacial distributions when comes to different products in different year, which also lays the foundation for the exploratory modelling of the data.

As that API can be knitr out, so I just put a screenshot here.

Besides the data visualization of 3D globe, to have a more convenient and more perspicuous review of the

Figure 2: 2D Globe Panel, powered by Leaflet Map Library

trade record, we decide to split the data for different trade levels, then show them on plain map of the world.

Thanks to leaflet library is support by Shiny App's JavaScript version, we are able to easily deploy this Panel just by downloading the R version of leaflet library and deploy it into our app. As the above chart shows, we have the spacial data assigned for each country, so we make use of the spacial data again and mark the different countries with different Dollar signs.

We now have a general sense of the trade, and how the proportions of different products distributes around the world. So we will introduce the more numeric way of data analysis: Market Share analysis and the Clustering Analysis.

4.1.2 Market Share

From the globe panels, we can get a general insights for the trade numbers, but to be more precisely, here comes to the summarize numbers. One of the most used statistics in trade report is the market share, namely, how much proportion for each country's product has in the total trade sum worldwide.

Thanks to the powerful R packages again, here we use the treemap package to plot the market share plot. From the areas of the squares, we can easily judge which country plays a important role in the trade.

```
input$year_tree = 2017
input$number_countries_tree = 15
input$com_tree = "Rice"
input$type = "Import"
```

From these we obtain some interesting insights. In the other hand, some well-known conclusion seems to have changed as time passed. In our childhood, we all know the rice export is one of the most trade income source for China, but from the data it seems the time has changed. When we verify the conclusion online for more detailed data, it appears that Thailand and India has much more market share as for Rice when it comes to trade with USA.

So we want to have a more closer look at the trade numbers, how much gap it has for different countries' market share, so we also conclude a ggplot into the Market share Panel.

4.1.3 Clustering Analysis

```
input$number_clusters = 5
input$year_cluster = 2017
```

Here, we use the simplest clustering method, Kmeans Clustering, which is also deployed by the R function $kmeans(data, centers)$.

K Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In k means clustering, we have the specify the number of clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps:

1. Reassign data points to the cluster whose centroid is closest.
2. Calculate new centroid of each cluster.

These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the euclidean distance between the data points and their respective cluster centroids.

By Clustering, we can see which countries plays the most important role, and which plays the less important role in the trade game. From this part, we can decrease the perdictor variables of the linear model in the next part. Also, we can reclean the data by deleting the smartest proportion countries and the model more clear.

4.2 Modelling

To inverstigate our questions, firstly we add several indicators into our main data, as we have talked about in the introduction part. The data structure then becomes like this:

```
colnames(total.keydat)
```

```
 [1] "Country Code"      "Year"              "Country"
 [4] "Region"            "IncomeGroup"       "Type"
 [7] "Silk"              "Cotton"            "Wheat"
[10] "Rice"              "Barley"            "Maize"
[13] "Other.Cereals"     "Coffee"            "Cocoa"
[16] "Tea"               "Soft.Drinks"       "Alcohol"
[19] "Milk"              "Juices"            "Population.Growth"
[22] "Undernourishment"  "Overweight"        "Death"
[25] "Income_Male"       "Income_Female"     "GDP.per.person"
[28] "P.G.U.S"           "U.N.U.S"           "O.W.U.S"
[31] "D.U.S"             "I.M.U.S"           "I.F.U.S"
[34] "Gdp.P.U.S"
```

The explanation of the columns we've added is as below: *Death* stands for death rates under communicable diseases and nutrition conditions, *P.G.U.S* stands for population growth rate in the United States, *U.N.U.S* stands for prevalence of undernourishment(% of population) in the United States, *O.W.U.S* stands for prevalence of overweight(% of children under 5) in the united states, *D.U.S* stands for death rates under communicable diseases and nutrition conditions in the United States, *I.M.U.S* stands for wage and salaried male workers(% of male employment) in the United States, *I.F.U.S* stands for wage and salaried female workers(% of female employment) in the United States, *Gdp.P.U.S* stands for GDP per person employed in the United States.

Secondly, we build some models and decide to do linear regression of them. When building our model, we use ratios rather than real sizes of trade. We originally used the real sizes of trading in our model, and the results were not good, after consideration we turned to look at ratios because ratios are more operable while the real numbers are enormous and hard to handle. Suppose that the volumn of import and export equals to what has been consumed in reality, these ratios could serve as the structures of what percentage a country would consume certain goods in certain category. Specifically for a certain category, the ratio is the volumn of trade of one goods out of the total volumn of trade of all the goods in that category. For example, the ratio of *cotton* equals to

$$\frac{cotton}{cotton + silk} \tag{1}$$

4.2.1 Model Overweight (Export)

We've built a very interesting model for the prevalence of *overweight* of children under 5. Firstly we introduce *region*, *income group*, *population growth*, *year*, *cereal*, *drinks* into our variables. We've found only *region*, *population growth* and *tea* to be statistically significant. Then we keep those variables in our model, the result is as followed:

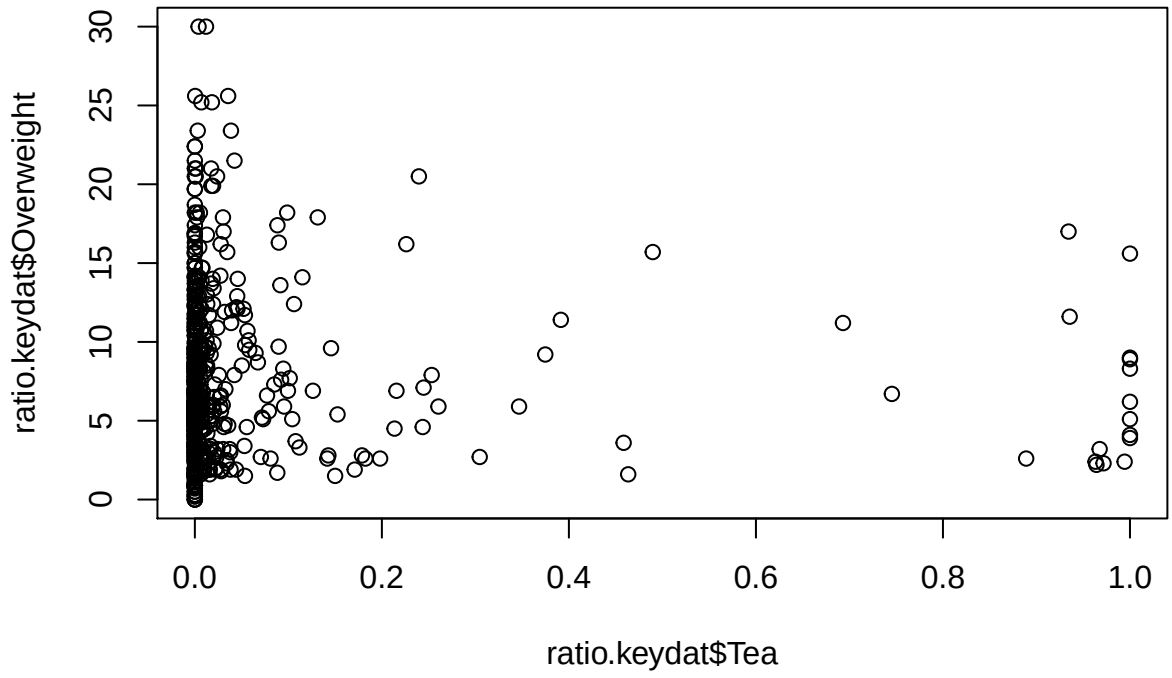
```
overweight.formula <- " Overweight ~ Region + Population.Growth + Tea"
overweight.model <- fit.model(dat = ratio.keydat[Type ==
    "Export", ], the.initial.formula = overweight.formula,
    model.type = "linear")
kable(overweight.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	6.828	0.760	8.980	0.000	5.338	8.318
RegionEurope & Central Asia	6.414	1.036	6.192	0.000	4.384	8.445
RegionLatin America & Caribbean	1.320	0.847	1.558	0.120	-0.340	2.980
RegionMiddle East & North Africa	6.228	1.024	6.084	0.000	4.221	8.234
RegionSouth Asia	-3.058	1.128	-2.711	0.007	-5.268	-0.847
RegionSub-Saharan Africa	0.402	0.912	0.441	0.659	-1.385	2.189
Population.Growth	-0.763	0.254	-2.998	0.003	-1.261	-0.264
Tea	10.416	3.580	2.909	0.004	3.399	17.433

Looking at the table, we could see that *region Europe & Central Asia*, *region Middle East & North Africa*, *region South Asia* are statistically significant, so we assume that if a baby is born in *Europe*, *Central Asia*, *Middle East* and *North Africa* region, he or she has higher probability to be overweight when he or she grows older. If a baby is born in *South Asia* region, he or she has lower probability to suffer from overweight as he or she turns older. And *population growth* is also statistically significant, it has negative correlationship with the prevalence of *overweight*. As for foods and drinks, we first believe that people's main *cereal* may affect the prevalence of *overweight*, however the former result shows that they are not statiscally significant, and there might be some other reasons. As for *drinks*, only the ratio of *tea* is statistically significant, which is opposite to our intuition. Scientists often think *tea* to be a kind of drink that is good for human's health, however in our result, the more ratio you consume *tea*, prevalence of *overweight* in children under 5 is more likely to happen than any other factors in our model.

To interpret this unusual result, we have two ways. For the first one, we think that it is not appropriate to put *tea* in our model if we don't know the consuming structure of *tea* for children under 5, and a child is not like to be allowed to drink tea because he or she is too young. Also, the stand error of the coefficient is very large, it's not proper to put *tea* in our model. For the second one, we suppose that the prevalence of *overweight* is approximately the same within each age level, and we assume that most people drink *tea* together with high heat foods such as sugar and cream. For example,*bubble milk tea* is very popular among teenagers and it turns out to be a kind of unhealthy drink.

```
plot(ratio.keydat$Tea, ratio.keydat$Overweight)
```



We then plot the *tea* ratio and prevalence of *overweight*, from the result we could see that it is not suitable to put *tea* in our model.

4.2.2 Model Death (Export)

When building our model for *death* rate under communicable diseases and nutrition conditions, we first introduce variables including *region*, *income group*, *cereals* and *drinks*. The result is not good because both items in *cereals* and *drinks* are not statistically sigificant though we thought they would have some significant effect on our model. Then we remove those facotrs from our model and the result is as below:

```
death.formula <- "Death ~ Region + IncomeGroup"
death.model <- fit.model(dat = ratio.keydat[Type == "Export",
    ], the.initial.formula = death.formula, model.type = "linear")
kable(death.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	12.484	1.465	8.524	0.000	9.613	15.354
RegionEurope & Central Asia	-9.465	1.529	-6.191	0.000	-12.462	-6.469
RegionLatin America & Caribbean	1.951	1.554	1.255	0.210	-1.095	4.997
RegionMiddle East & North Africa	-1.409	1.669	-0.845	0.399	-4.680	1.861
RegionNorth America	-7.434	4.945	-1.503	0.133	-17.125	2.258
RegionSouth Asia	10.202	2.431	4.197	0.000	5.438	14.966
RegionSub-Saharan Africa	34.394	1.569	21.918	0.000	31.318	37.469
IncomeGroupLow income	14.679	1.790	8.202	0.000	11.171	18.186
IncomeGroupLower middle income	11.665	1.356	8.601	0.000	9.007	14.323
IncomeGroupUpper middle income	0.906	1.242	0.729	0.466	-1.528	3.339

From the table above we could see that: *region Europe & Central Asia*, *region South Asia*, *region Sub-Saharan Africa*, *low income group*,*lower middle income group* are statistically significant. For *region*, if you live in *Europe* and *Central Asia* area, you are less exposed to death under communicable diseases and nutrition conditions than *South Asia* and *Sub-Saharan Africa* area. We assume that *World Health Organization* should take action to do more research on people's health of these areas. As for income group, *lower income* group should be more taken care of because they have high potential to suffer from communicable diseases and bad nutrition conditions.

4.2.3 Model Undernourishment (Export)

We've built a model for the prevalence of *undernourishment*. We introduce variables including *region*, *income group*, *cereals*. We remove *income group* in our model because it raises some perplexing result. The result of linear regression is as followed:

```
undernourishment.formula <- "Undernourishment ~ Region + Wheat + Rice +Barley + Maize + Other.Cereals"
undernourishment.model <- fit.model(dat = ratio.keydat[Type ==
    "Export", ], the.initial.formula = undernourishment.formula,
    model.type = "linear")
kable(undernourishment.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	19.200	1.655	11.600	0.000	15.956	22.445
RegionEurope & Central Asia	-6.157	1.641	-3.752	0.000	-9.373	-2.941
RegionLatin America & Caribbean	-2.206	1.156	-1.908	0.057	-4.472	0.060
RegionMiddle East & North Africa	-3.354	1.284	-2.612	0.009	-5.870	-0.838

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
RegionSouth Asia	4.785	1.719	2.784	0.005	1.416	8.154
RegionSub-Saharan Africa	7.167	1.103	6.498	0.000	5.005	9.329
Wheat	-4.580	1.578	-2.901	0.004	-7.674	-1.486
Rice	-3.976	1.684	-2.362	0.018	-7.276	-0.677
Barley	22.660	9.441	2.400	0.017	4.155	41.165
Maize	-2.259	1.611	-1.402	0.161	-5.417	0.899

As we can see in the table, *region Europe & Central Asia*, *region Middle East & North Africa*, *region South Asia*, *region Sub-Saharan Africa*, *wheat*, *rice*, *barley* are all statistically significant. As for region, the prevalence of *undernourishment* is severe in *South Asia* and *Sub-Saharan Africa* region. And as for *cereals*, consuming more cereal in *wheat* and *rice* will reduce the prevalence of *undernourishment*, we assume that eating them would provide much more energy than *barley* for human's body, and we think that it is a very good topic for biologists to do research on. In the meantime, we also suggest organizations suchs as *World Health Organization* and *Red Cross* to take action on helping reduce the prevalence of *undernourishment* in these areas, maybe introducing cereals like *wheat* and *rice* in their consuming structure would work.

4.2.4 Model goods

4.2.4.1 Model Coffee (Export)

We've built a model for the ratio of *coffee*, adding variables such as *region*, *income groups*, *salaried ratio* by different *genders*, and *population growth*. The result of the linear regression is interesting: the effect of *region* and *population growth* are not statistically significant, which means these two are not likely to affect the ratio of *coffee* in a consuming structure. Also the effect of *income* growth of *male* is not statistically significant. We remove the factors of *region* and *population growth* out of our model and the result is as followed.

```
Coffee.formula <- " Coffee ~ IncomeGroup + Income_Female + Income_Male"
Coffee.model <- fit.model(dat = ratio.keydat[Type == "Export",
    ], the.initial.formula = Coffee.formula, model.type = "linear")
kable(Coffee.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	-0.062	0.034	-1.794	0.073	-0.129	0.006
IncomeGroupLow income	0.164	0.041	4.025	0.000	0.084	0.243
IncomeGroupLower middle income	0.084	0.021	4.103	0.000	0.044	0.125
IncomeGroupUpper middle income	0.040	0.012	3.217	0.001	0.016	0.064
Income_Female	0.002	0.001	3.609	0.000	0.001	0.004
Income_Male	-0.001	0.001	-1.517	0.130	-0.003	0.000

All of the *income group* have positive correlationships with the ratio of *coffee*. The *salaried ratio* growth of *female* would have positive affect on the ratio of coffee from our investigation. Now we could suppose that the consuming of coffee is very populatio among any level of income groups. And female is a very stable resouce of the consuming of coffee.

Our assumption is that, if a coffee company invest more into making advertisement to attract the attention of female consumers, it will gain more money. For example, Starbucks put on new and decorative wrap of their coffee cups when there come Christmas and other holidays, many female consumers buy coffee and take selfies with the lovely cups, that must increase their turnover. However, we do suggest that those coffee company invest a little on exploring male customers, because male are also a large potential cash pool.

4.2.4.2 Model Tea (Export)

Now let's take a look at one of the substitute goods of *coffee*, *tea*. There is a model for the ratio of *tea*, we add variables including *region*, *income groups*, *salaried ratio* by different *genders*, and *population growth*. Like what we've seen in *coffee* model, *population growth* is not statistically significant. However, *income groups* is not statistically significant in this model, and some *region* is significant. Removing *income groups* and *population growth* out of our model, the result is as below:

```
Tea.formula <- " Tea ~ Region + Income_Female + Income_Male"
Tea.model <- fit.model(dat = ratio.keydat[Type == "Export",
    ], the.initial.formula = Tea.formula, model.type = "linear")
kable(Tea.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	-0.024	0.019	-1.239	0.215	-0.062	0.014
RegionEurope & Central Asia	0.019	0.011	1.758	0.079	-0.002	0.040
RegionLatin America & Caribbean	0.016	0.012	1.355	0.176	-0.007	0.039
RegionMiddle East & North Africa	0.095	0.016	5.915	0.000	0.063	0.126
RegionNorth America	0.022	0.024	0.926	0.355	-0.025	0.069
RegionSouth Asia	0.012	0.031	0.397	0.691	-0.048	0.073
RegionSub-Saharan Africa	0.051	0.019	2.756	0.006	0.015	0.088
Income_Female	-0.001	0.000	-3.273	0.001	-0.002	-0.001
Income_Male	0.002	0.001	3.586	0.000	0.001	0.003

As we can see in the table, *region Middle East & North Africa* and *region Sub_Saharan Africa* are statistically significant and so are *salaried ratio* growth with both *female* and *male*. Both these two regions have positive correlationships with the ratio of *tea*, we suggest that tea companies make more trade with these two regions. As for *salaried ratio* growth with different genders, when the salary of a *female* increases, she has the tendency of consuming less tea in her drinking structure, maybe she will buy more coffee as we have assumpted before. When the salary of a *male* increases, he is more likely to buy more *tea* drinks.

Like what we have talked before, we think *tea* and *coffee* companies should focus more on gender problems and make different target consumers. At the same time we encourage them to invest a little on exploring new customers.

4.2.4.3 Model Milk (Export)

We've built another model for trade volumn of *Milk*, we consider it to be one of the supplementary goods. So in this model we use the real volumn of trade in our model. To simplify the formula we introduce only *coffee* and *tea* as the variables. And the result is as below:

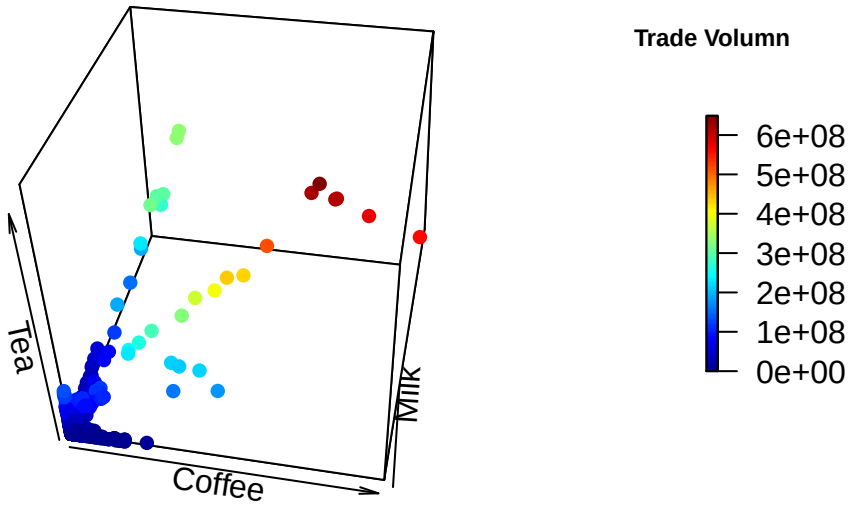
```
Milk.formula <- " Milk ~ Coffee + Tea"
Milk.model <- fit.model(dat = total.keydat[Type == "Export",
    ], the.initial.formula = Milk.formula, model.type = "linear")
kable(Milk.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	1937807.607	720709.950	2.689	0.007	525242.061	3350373.153
Coffee	-1.391	0.029	-47.732	0.000	-1.448	-1.334
Tea	14.345	0.132	108.530	0.000	14.086	14.604

As we can see in the table, both *coffee* and *tea* are statistically significant. As for correlationships, *coffee* is negative and *tea* is positive. We assume that the increaing trade volumn of *tea* would accompany the

increasing trade volumn of *milk*, for example, *bubble milk tea* is quite popular among teenagers. However when we increase the export of *coffee*, we may export less *milk*, though we previously thought *coffee* and *milk* are good mate. And we've make a 3D plot for the trading volumn of these three goods.

```
scatter3D(x = total.keydat$Coffee, y = total.keydat$Milk,
    z = total.keydat$Tea, xlab = "Coffee", ylab = "Milk",
    zlab = "Tea", pch = 16, cex = 1, clab = "Trade Volumn",
    theta = 10, d = 2, colkey = list(length = 0.5, width = 0.3,
        cex.clab = 0.75))
```



4.2.4.4 Model Silk (Export)

We build the model of the ratio of *silk*. We don't build a model for cotton because in the consuming structure of cloth these are only two kinds of goods: *cotton* and *silk*. When we gain a conclusion of one goods, it's not hard to gain the other. We originally introduce variables including *region*, *income groups*, *population growth*, *GDP per person*, and *salaried ratio* by different genders. However the first 3 factors are not statistically significant in the result, then we remove them from our model. The result of linear regression is as below:

```
Silk.formula <- "Silk ~ GDP.per.person + Income_Female + Income_Male"
Silk.model <- fit.model(dat = ratio.keydat[Type == "Export",
    ], the.initial.formula = Silk.formula, model.type = "linear")
kable(Silk.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	-0.036	0.038	-0.959	0.338	-0.109	0.038
GDP.per.person	0.000	0.000	-2.128	0.034	0.000	0.000
Income_Female	-0.001	0.001	-1.046	0.296	-0.004	0.001
Income_Male	0.003	0.001	2.043	0.041	0.000	0.006

We could see from the table that: *GDP per person* and *salaried ratio* growth with *male* are statistically significant, however the coefficient of the former one is 0. And for *male*, he is likely to consume more silk product when his salary increases. Though we don't know whether a man buys silk product for himself or not, we do make the assumption that a company that sells *silk* products should focus mainly on appealing *male* customers.

4.2.4.5 Model Silk (Import)

We build a similar model for *silk* ratio, and it is for goods import. However there is no variable that is statistically significant in the result.

```
Silk.us.formula <- "Silk ~ Gdp.P.U.S + I.F.U.S + I.M.U.S"
Silk.us.model <- fit.model(dat = ratio.keydat[Type == "Import",
    ], the.initial.formula = Silk.us.formula, model.type = "linear")
kable(Silk.us.model)
```

Variable	Estimate	Std. Error	t value	p.value	Coef.Lower.95	Coef.Upper.95
(Intercept)	1.184	9.315	0.127	0.899	-17.073	19.440
Gdp.P.U.S	0.000	0.000	0.232	0.816	0.000	0.000
I.F.U.S	-0.035	0.121	-0.287	0.774	-0.272	0.203
I.M.U.S	0.023	0.055	0.427	0.669	-0.084	0.131

5 Limitations and Uncertainties

1. We apply general linear regression models to all countries. However different countries have their own economics structures, a general function is not valid to all of them. At the same time, we only have limited data of years, we can hardly build a suitable specialized model for a certain country using the available data.
2. Secondly, there are lots of missing values in the dataset. For missing values in data obtained from *WBO*, we assign *NA* for them because there is no way to fill in the blank of a missing *income group* and *region*. As for missing value in data obtained from *USCB*, we assign *0* value for them. We only have no more than 200 effective countries in the former data after we cleaned them. As for the latter one, we find some points gathering together around the value of *0* when we plot the data. We don't know whether it's the best way to deal with those missing values. And also, lack of data is a very serious limitation in our project.
3. Thirdly, we focused mainly on exports, our model mainly present the effect of consuming goods from the United States. However we don't know how much United States' goods contribute to their consuming structure and the results are not quite convincing. At the meantime, we have only used data of sixteen years, which is not enough to model the consuming structure and other indicators for a single country, and the results of linear regressions are not good for the United States as a consequence, so we haven't talked much about model for *Import* in the report.
4. Finally, we assume that the consumption of a certain goods equals to the trade volumn of it. We don't know whether they truly consume them in one year or some goods may be stored for years.

6 Areas of Future Inverstigation

6.1 Variable Selection for Linear Models

We have introduced some variables in this project, including 14 kinds of goods traded between 247 countries, as well as 6 data dimensions including indicators obtained from *WBO*. For each model, we have options for several predictor variables, and we can further use Lasso method to decrease the number of predictor variables and make our models much more clear and statistically significant.

6.2 Economical Models Validation

As we become more interested in digging out our own insights from the data, we haven't pay much attention to the wellknown macroeconomic models on textbooks. With these multi-dimensional and time-series data on trade, we can easily apply the models to data and valid the models. As the outcome is quite significant, we haven't spend much time on that in the process.

6.3 More Profound Text Mining

When it comes to trade, there are much information presented by news, trade reports and policy files, which requires the text mining techs to play with them. Even though in this project we did the text mining part and draw some conclusion including what's the mostly mentioned products in the annual trade report, it still has long way to go for the text processing part. Following are some of Ming's ambitions and plans for text part but not completed due to the limited time.

(This ideas came from a personal talk with a data scientist in Mckinsey, who informed Ming how the text data process and machine learning algorithms helps the traders in GS)

1. Web Scrape the news and reports related to some products, let's say, Iron, from the mainstream media like the Wall Street Journal and completed the sentimental analysis for the text data.
2. Load the data traded from countries to countries of those above products, as well as the stock price of some main companies in that industry and perform the time-series analysis based on a much shorter time interval, let's say, 2 weeks, than years.
3. combine the more detailed analysis part with the yearly-analysis part, and let's see how the news, reports and policy files will influence the daily trade and the annual trade.

This is really a large-workload task so we didn't finish it due to the limited time and limited data skills. But it sounds really interesting and the whole trade numbers can be cut into more and more smaller time interval, from years to quarters, to weeks, to days, to hours and to minutes. That sounds really nice but really needs time and skills to accomplish.

6.4 Macro-Economics is really COMPLICATED

Both two contributors are not so familiar with the macro-enonomics field, so we just select the variables by some short essays, blogs or even online news. The data source still hides many secrets to be digged.

7 References

1. Beautiful Data: The Stories Behind Elegant Data Solutions by Toby Segaran
2. How trade has influence the world, by Yale University
3. Storytelling with Data: A Data Visualization Guide for Business Professionals
4. Spark R Documentation
5. tm Documentation
6. Shiny App Development Tutorial
7. HTML Tutorial by Udemy
8. Google Maps API for R
9. Leaflet Documentation
10. NLTK Documentation
11. Long-only Trading Strategy with NLP derived Social Media Sentiment
12. Joseph Wang's Quora Page
13. Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100–108.
14. Applied Linear Regression Models, by Nelson Li, 2004
15. Rmarkdown Documentation
16. Handling and Processing Strings in R, by Gaston Sanchez
17. Pattern Recognition and Machine Learning
18. Revealjs Package Documentation
19. Macroeconomics (9th Edition 2016) by N. Gregory Mankiw
20. International Economics 10th edition by Krugman