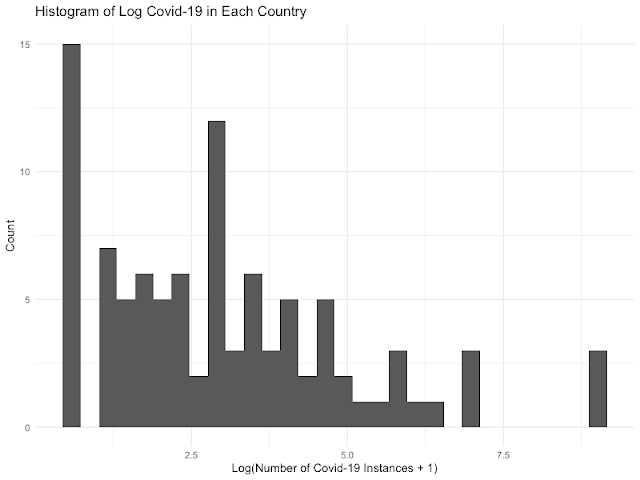
TLDR: I’ve found an association between number of people that tested positive for COVID-19 in a country and imports from China. In addition there are particular industries that are particularly correlated with COVID-19 rates. Iran is still an outlier taking this information into account.

*Intro:*

*=====*

Coronavirus (COVID-19) rates of infection outside of China have not applied to countries uniformly. As of 3/9/2020 Italy (9,172), South Korea (7,578), and Iran (7,161) have a disproportionate number of cases relative to other countries such as the US (605). Below is a histogram showing the distribution of cases in each country and it’s clear it is right skewed and non-uniform.



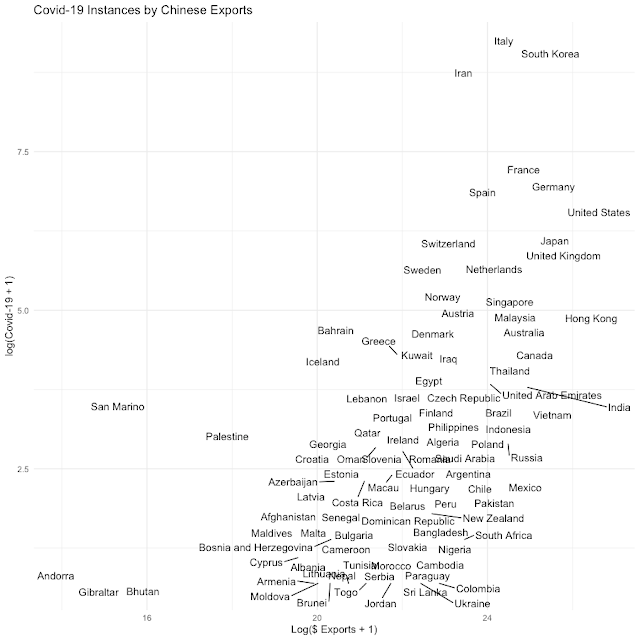
What causes such unequal distribution of COVID-19 cases among countries? I assume some degree of ‘connectives’ between countries would be a correlated with these rates. In particular I assume an increase in imports from China will lead to more cases in the importing country.

To investigate this I looked at data from <https://oec.world/en/resources/data/>. This includes bilateral trade data between countries and is broken down into hundreds of different categories of products. I’ll look at a few things 1) are the outliers of Iran, Italy, and South Korea explained by trade with china? and 2) are some categories of traded products more correlated with rates of infection and 3) if so which products are more correlated?

*1) Are Chinese Exports Correlated with COVID-19 Rates?*

Below is a plot of COVID-19 rates by total Chinese exports for each country in 2017 (the latest data available). One can see a fairly strong association between these two variables. However there are a few outliers. Notably South Korea, Italy, and Iran are still above expected along with smaller countries and territories like San Marino.

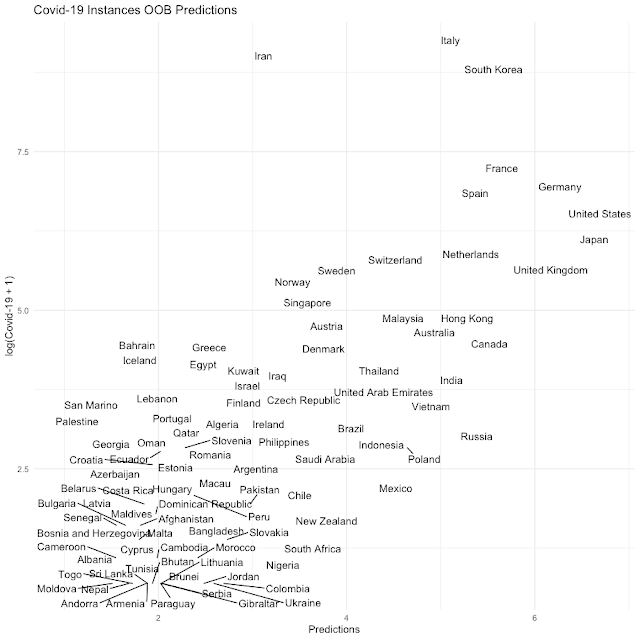
Running a random forest of log(COVID-19-rates)~log(total\_chinese\_exports) yields an out-of-bag (OOB) r^2 of .12. So while relationship appears to be highly correlated there are enough outliers to decrease the ‘explained’ variance in infection rates.



*2) Are there product categories among Chinese Exports that are more correlated with COVID-19?*

To test this I ran a random forest of log(COVID-19-rates) on all export amounts broken out by categories. This is a very wide data set with 97 countries (observations) and over 1200 categories (explanatory variables). Even this wide dataset produced an OOB r^2 of .49 meaning that it captured much of variance not captured by previous univariate model.

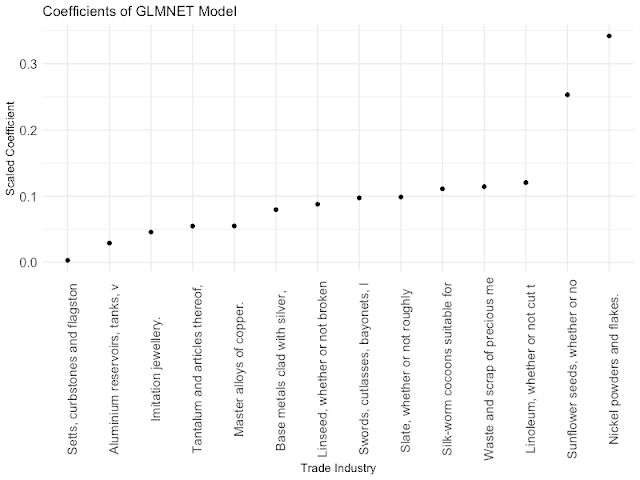
Below a scatterplot of COVID-19 rates by these OOB random forest predictions. San Marino does not look like a large outlier anymore but Italy and South Korea do. Iran looks like a much larger outlier.



*3) Which product categories are more correlated with COVID-19?*

To answer this question I setup a model of same form as the random forest in 2) but used a lasso regression to apply variable selection. Below are the scaled coefficients kept by lasso regression.

One can see Chinese nickel powder and sunflower exports are more correlated with COVID-19 rates in importing country.



I looked at these and couldn’t really make a story behind it. For instance the 7th most important feature had to do with swords and bayonets. I have no idea what this means :).

*Conclusion:*

It appears trade is highly correlated with COVID-19 rates. In addition the evidence suggests that some industries are more correlated with COVID-19 rates than others (do to increase in r^2 between 1) and 2) ).

However, it’s still not clear why countries like Iran have such high rates. Perhaps the data is stale since it was from 2017 or maybe Iran-Chinese trade is under reported in this dataset due to US sanctions.