University of Exeter coursework header sheet

Extension
211419

BEMM460J Statistics and Mathematics for Business Analytics (A, TERM2 2021/2)

1072885

720010221

Coursework: Individual report

Submission Deadline: Fri 13th May 2022 12:00

Personal tutor: Justin Tumlinson

Marker name: N/A

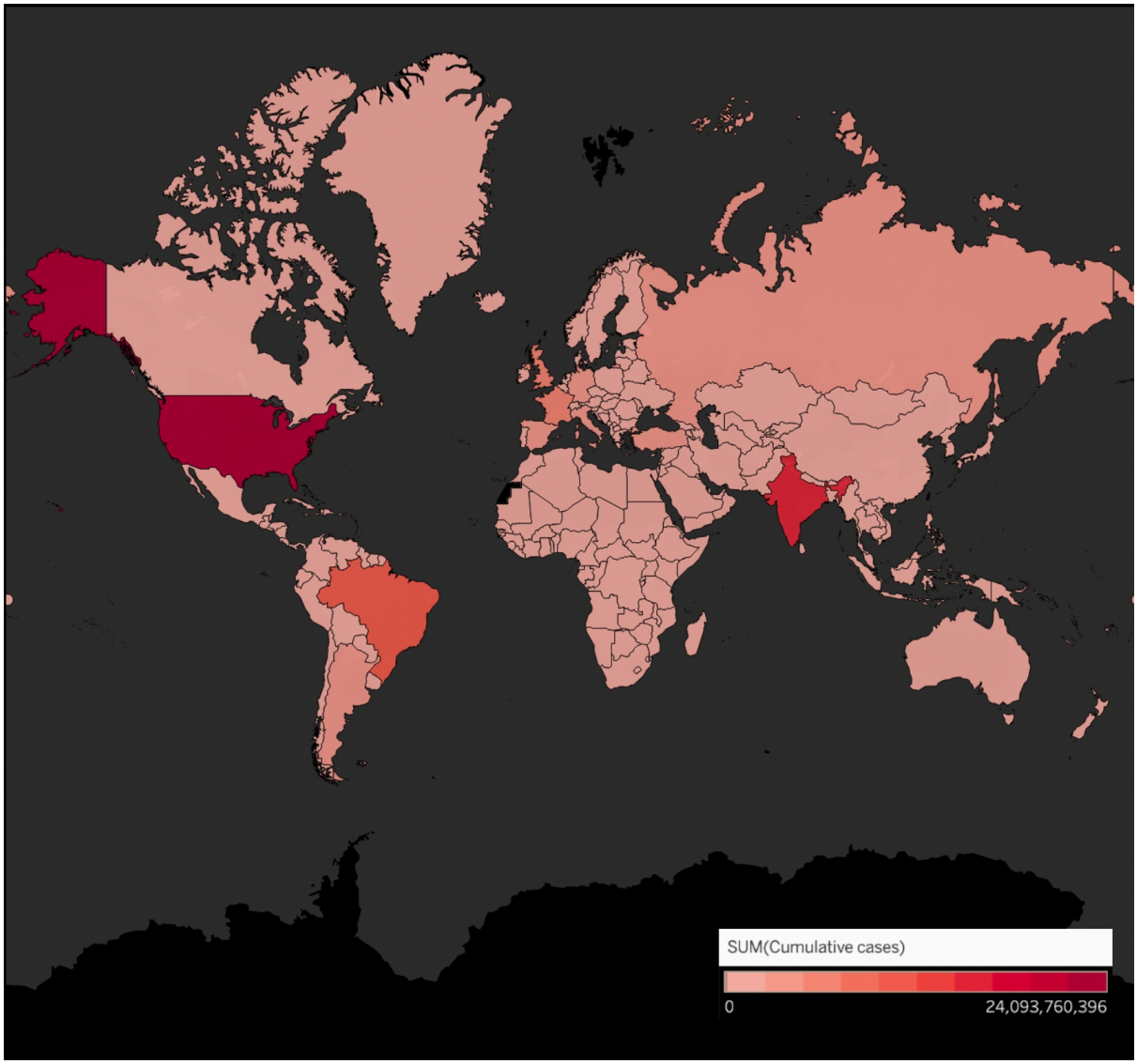
Word count: 3000

By submitting coursework you declare that you understand and consent to the University policies regarding plagiarism and mitigation (these can be seen online at www.exeter.ac.uk/plagiarism, and www.exeter.ac.uk/mitigation respectively), and that you have read your school's rules for submission of written coursework, for example rules on maximum and minimum number of words. Indicative/first marks are provisional only.

Statistical Analysis of Economic Indicators Pre and Post COVID-19

Statistical Analysis of Economic Indicators Pre and Post COVID-19

Sachin Sharma



Abstract

In this report I analyse data from WHO's COVID-19 data bank The Organisation for Economic Co-operation and Development (OECD) about Economic Indicators and do analysis pre and post COVID-19 pandemic announcement. It was necessary that the data be in time series format (monthly) because data about COVID-19 cases is new and only available from Jan 2020 to now. But

due to the large magnitude of effects that were caused by this pandemic to the world economies, its effects were very apparent. In report I first correlated the Indicators to find the most proportionate attributes with the help of different Statistical tools followed by a comparison between COVID-19 cases and most significant indicators.

Keywords: COVID-19, Inflation, Business Confidence Index, Consumer Confidence Index, Composite Leading Values, Long Term Interest Rates, Total Manufacturing Inflation, Domestic Manufacturing Inflation, Share Prices, Energy Inflation, Short Term Interest Rates, Unemployment Rate.

---

1 Introduction

On December 31, 2019, the World Health Organization (WHO) was alerted about a cluster of cases of pneumonia of unknown aetiology identified in Wuhan City, Hubei Province, China. On January 12, 2020, it was stated that a new coronavirus had been discovered in case samples and that early examination of virus genetic sequences showed that this was the source of the epidemic. The virus was named SARS-CoV-2 and the disease COVID-19. Ever since that atrocious day the world has never been the same. Nearly every field of work has been affected due to the lockdown protocols implemented by the world's governments. No one had any idea about the scale on which this disease would spread. To date, COVID-19 has been infected and reported in 519 million people and taken the lives of 6.26 million. Many cases are believed to go unreported or undiscovered because people are fearful of prejudice or testing themselves so that they do not have to disclose it. This fear has made the spread even more aggressive. But perhaps now, in a hopeful sigh of relief, we can see that marking the 2 year 4 months anniversary has brought some elective comfort. There is a less urgent need to wear masks in public places. Vaccinations are readily accessible in most developed countries, and some degree of herd immunity is at work, which has begun to phase out lockdown procedures. To celebrate this jubilee, in this research, I will analyze the countries that "Carpe" ed their "Diem"s by studying the economic indicators The Organisation for Economic Co-operation and Development (OECD)'s data bank provides in a monthly format find significant statistical details.

2 Background Information

2.1 Pearsons Correlation Test

Pearson's correlation coefficient (r) is a measure of two variables' linear relationship. A scatter diagram is used to graphically show the relationship between data pairs in correlation analysis. Correlation coefficient values range from –1 to +1. Positive correlation coefficient values suggest a propensity for one variable to rise or fall in tandem with another. Negative correlation coefficients imply that a rise in one variable's value is connected with a reduction in the other variable 's value, and vice versa. Correlation coefficients around zero suggest a weak linear relationship between two variables, whereas those near –1 or +1 indicate a strong linear relationship between two variables. Formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

2.2 Chi-Square Test

A chi-square statistic is a test that evaluates how a model compares to actual observed data. A chi-square requires data that is: random, raw, mutually exclusive, drawn from independent variables, and drawn from a big enough sample. Tossing a fair coin, for example, meets these conditions. It allows you to assess how well a data sample fits the (known or assumed) features of the total population that the sample is meant to represent. This is also called quality of fit. If the sample data do not match the predicted attributes of the population in question, we should not utilise this sample to draw inferences about the wider population. Formula :

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

2.3 Linear Regression

Linear regression is a method of modelling a relationship between two variables by fitting a linear equation to observed data, with one variable serving as an explanatory variable and another as a dependent variable.

When two variables have a linear connection, linear regression analysis is used to explain or predict one from the other. A straight line may be used to represent a linear connection between two variables in the same way that the average can be used to summarise a single variable. There is variability in the straight line, just as there is variability in the mean (for univariate data) (for bivariate data). Because of this uncertainty, the straight line, like the average, is a helpful but incomplete description. Formula: Sample regression line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \tag{2}$$

Population regression line:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{3}$$

3 Data Formation

I looked for a data set that satisfied specific criteria when seeking data for this project. The first criteria I set was for it to come from a reliable source. Hence for COVID-19, I took the data set from the World Health Organization and, for the economic indicators, I took the data from Organisation for Economic Co-operation and Development (OECD).

The next feature I looked for was the presence of a time series. Since we are aware that COVID-19 has only been around for past two years now, I determined that this time series should be monthly at the very least. Eventually, this resulted in multiple separate data files, which I had to link by period and nation (economic indicator set) as well as COVID-19 cases and deaths.

All the data pre-processing and analysis was done in python on pandas. A ipynb notebook is attached here for further refferences.

4 Exploratory Data Analysis (EDA)

To begin my EDA, I first looked at the distribution of values amongst columns (shown in the table "Descriptive Statistics List"). As you can see there is quite a bit of deviation amongst certain values. As I'll explain later, I will try to draw conclusions based on the relationships between these columns. Next *Normalization* was done to not bias a particular column, these values must fall within an acceptable range.The act of structuring data into tables in a way that the outcomes of the

database are always clear and as expected is called Database Normalisation. Relational database theory requires such standardisation.

Descriptive Statistics List				
	Mean	Min	Median	Max
Business Confidence	99.84	83.28	100.14	106.55
Consumer Confidence	99.85	100.05	100.058	105.47
Composite Leading	100.00	91.63	100.21	109.17
Inflation	2.78	-6.54	1.99	69.97
Long Term Interest	3.48	-0.97	2.97	29.24
Total Manufacturing Inflation	103.46	79.90	102.10	154.42
Domestic Manufacturing Inflation	105.40	71.81	102.60	569.69
Share Prices	103.07	26.86	99.25	373.21
Energy Inflation	3.73	-26.36	3.30	120.88
Short Term Interest	1.56	-0.85	0.32	21.91
Unemployment - Age 25-74	6.62	1.50	5.90	25.90
Unemployment - Age 15-24	17.64	3.10	1 6.20	62.90
Unemployment Rate	7.80	1.70	7.00	28.00
Unemployment in Youth Men	17.80	3.10	16.15	57.90
Unemployment in Youth Women	17.86	2.70	15.41	72.70
New cases	2.53e+03	0.00	2.40e+01	1.25e+06
Cumulative cases	6.31e+05	0.00	1.17e+04	8.06e+07
New deaths	30.88	0.00	0.00	11447
Cumulative deaths	11830.82	0.00	151.00	986698

Next, I wanted to get a general understanding of the flows of each data set. That is why, I plotted the average per month indifferent to country (normalized) (Fig. 1 and Fig. 2).

In Fig. 1 we can observe notable dips in our chosen economic indicator values around the time that COVID-19 first occurred(Dec19 -Jan20). This leads me to assume that these attributes were most likely affected by COVID-19, therefore any analysis, of these said indicator values, before and after COVID-19 could be interesting. In Fig. 2 information is as expected. In recent times we observe an increase in COVID-19 cases, with a significantly lower fatality rate. Out of these attributes, I chose cumulative COVID-19 cases to observe. The values seem less sporadic, but also capture the effect of a global event (the spike in values around January 2022).

Furthermore, I wanted to see in general how the economic values of the top ten GDP ranking countries compared to that of the grouping. As you can see in Fig. 3 all seemingly have a dip due to COVID-19, followed by a sharp increase in values. However, for some countries, the magnitude of the dip drastically differs (ex. between India and Canada). Additionally, we also observe that a country like Japan has a greater spread of attribute values compared to the mean attributes over time that we previously saw (Fig. 2). Therefore, any meaningful analysis should be done per country.

Anticipating correlation analytics that I would have to perform, I thought it would be useful to generate a quick insight into the whole grouping of data. A heatmap with Pearson correlation values can be seen in Fig. 2. In this, we expectedly view a dark blue region, signifying high correlation, in the bottom right corner. All these values relate to unemployment and are logically expected to correlate. It doesn't bring much meaningful information to do analysis between these attributes, and therefore in model generation, they should not be correlated with one another.

To take this a step further, one of the analytics I plan to present is a based on correlation pre

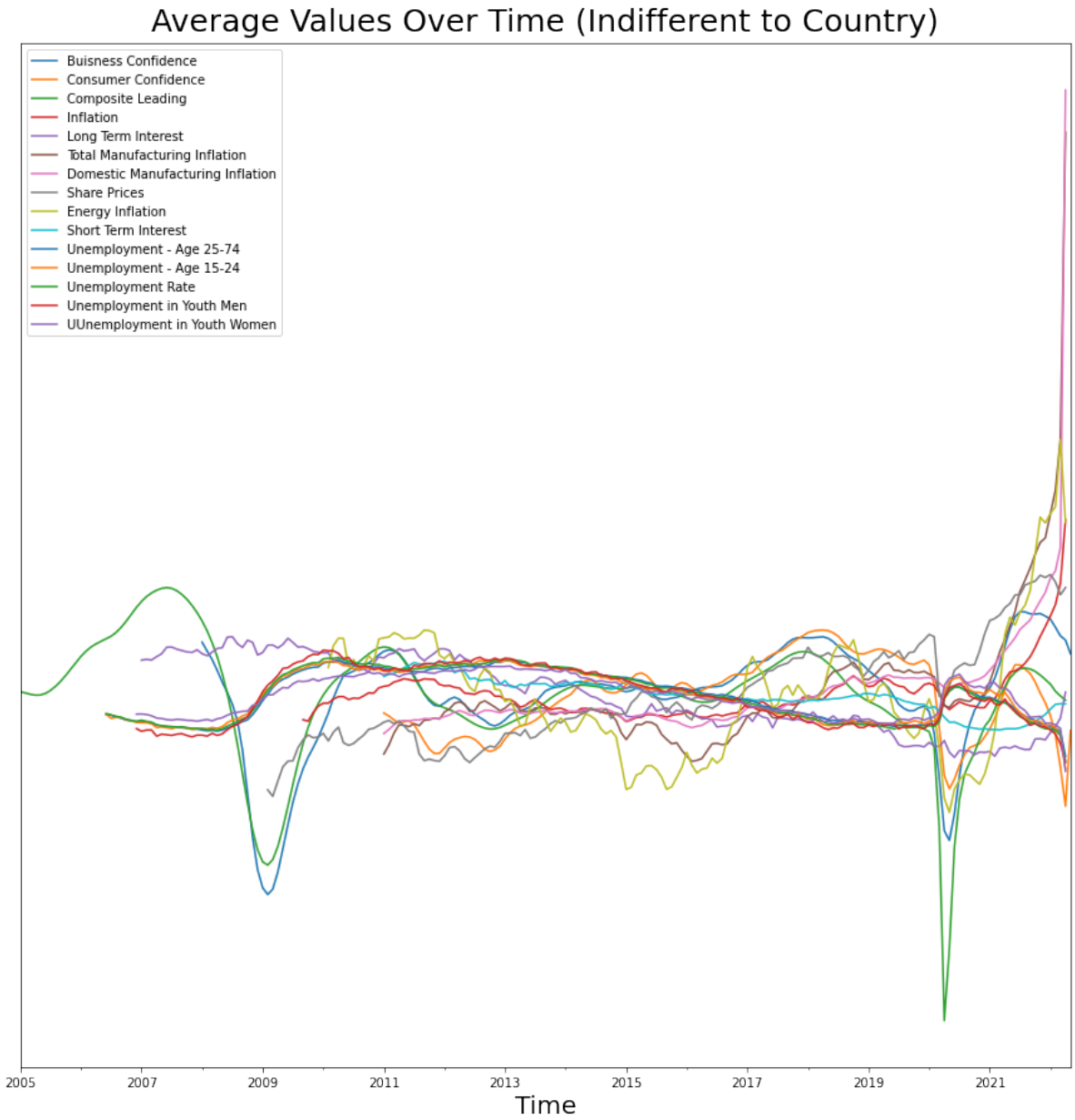


Figure 1: Economic Attributes Over Time

and post COVID-19. Therefore, as an exploration, I decided to look on an aggregated level how the correlation values for these two groupings compare (Fig. 5 and Fig. 6). Although its not overtly apparent, certain attributes inverted colors (dark tiles became light and light tiles became dark). Observing this allows me to believe that plotting a linear regression model pre and post COVID-19 will not have post COVID-19 values exactly fit. But, the information obtained will be nonetheless valuable.

5 Analysis

There are two essential ways in which I analyzed the data. The first was economic versus economic attribute linear regression. For this, the goal was to find the most correlated economic attributes pre-COVID-19, fit the data with a linear regression model, and see how the regression line performs for post-COVID-19 data. The second model was a COVID-19 attribute versus economic attribute

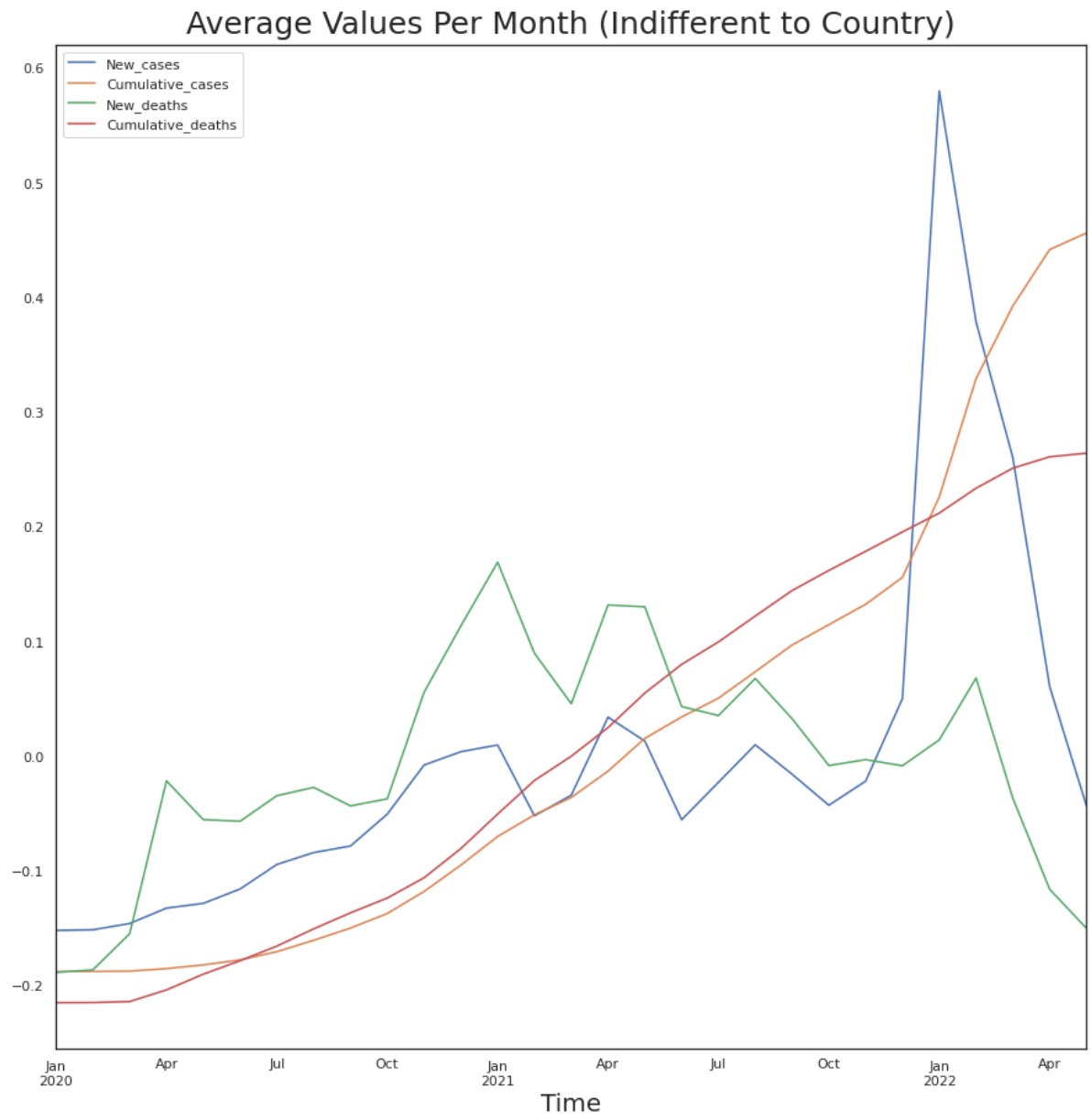


Figure 2: COVID-19 Attributes Over Time

linear regression. For this model, all data was from the start of COVID-19 on. The goal of this model was to take out a test set of data, fit COVID-19 Cumulative cases versus the most correlated economic attribute, and once again see how this test set performs. Both analyses were done on a per-country basis, as to ensure that the actions of one country would not disturb the actions and effects of another.

5.1 Economic Indicator versus Economic Indicator Linear Regression

5.1.1 Correlation

Since analysis is done on a per-country basis, I wanted to see which economic attributes correlated most with one another. Once again, this was done by applying Pearson correlation to corresponding columns. This was also filtered to exclude matching of columns containing "Manufacturing", "Unemployment", "Inflation", and "Interest", with themselves. Using this information I generated a histogram (Fig. 7), counting the number of occurrences for each combination of most correlated

attributes, per country.

From Fig. 7 you can see that there are two highest correlated attribute combinations: (Business Confidence, Consumer Confidence) and (Business Confidence, Composite Leading). Since there are two highest attributes, the following linear regression analysis was performed on both of these combinations. It should also be noted that you can also see that the count for these is not significantly higher than the other correlation combinations. This leads me to believe that the correlation values I see for the max combinations do not deviate much from the lower-ranked combinations.

5.1.2 Regression Model

Following the previous correlation analysis, I decided to plot the top ten GDP ranking countries for the regression model analysis. In Fig. 8, you can observe the plotted Business Confidence versus Consumer Confidence linear regression, and in Fig. 9 you can observe the plotted Business Confidence versus Composite Leading linear regression. While I attempted to plot the top ten GDP ranking countries, some countries did not have data for a particular attribute or did not have data relating to the particular attribute in the time frame that was necessary. Due to such, we observe eight countries, rather than ten, in the Business Confidence versus Consumer Confidence Linear Regression (Per Country) figures (Fig. 8).

As expected (due to quick glances at correlation values from Fig. 4 excluding overlap "Manufacturing", "Unemployment", "Inflation", and "Interest" columns), in Fig.8 and Fig. 9, you can see that the values between attributes are sporadic and the linear regression models don't even seem to fit the training set well, nonetheless the post COVID-19 data points.

5.1.3 Results

To start, it is important to discuss the variation of data points in the training set. As mentioned previously, visually, they seem more sporadic, almost circular or elliptical at times. However, as you will see in the Table: Business Confidence versus Consumer Confidence Variation and Chi-Square, variance in x and y are independently relatively small. With .22, .78, .61, and .54 being variance in x, y, for the Business Confidence versus Consumer Confidence model, and x, y for the Business Confidence versus Composite Leading model. While variance does represent the spread of a data set, it might have difficulty capturing the circular shapes that we are seeing.

As another method of analysis of results, to quantify how the difference between what was expected (our linear regression model) and what was observed (post COVID-19 data), chi-square tests were performed. As you can see in the tables below, most countries have a large chi-square value, indicating that the values we are observing are far away from what we would have expected. At first sight, these values are seemingly small, having a value of five is typically not considered a large value. However, everything is based on how you scale it - keep in mind that we performed standardization for each column previously, leaving the mean at zero, and std at one. Upon further inspection of individual values, we see that Italy has the lowest absolute chi value for the Business Confidence versus Consumer Confidence model, and Germany has the lowest absolute chi value for the Business Confidence versus Composite Leading model. If you reference back to Fig. 8, you can visually see that. You can also see the countries that have the largest chi values: France, and India, respective to the models mentioned previously.

Overall, the mean Chi value for the Business Confidence versus Consumer Confidence model was -5.36, and the mean Chi value for the Business Confidence versus Composite was 5.06. This allows us to quantize that the models built for the Business Confidence versus Composite were marginally better at predicting the data points we see due to COVID-19.

Business Confidence versus Consumer Confidence Variation and Chi Square			
Country	Variance X	Variance Y	Chi Square Value
Italy	0.158	0.718	-.3
China	0.112	2.359	-11.88
Japan	0.074	0.112	-5.49
France	0.225	0.419	-18.87
United States	0.119	0.444	2.23
United Kingdom	0.285	1.002	0.62
Brazil	0.618	1.089	2.3
Germany	0.175	0.062	-11.5

Business Confidence versus Composite Leading and Chi Square			
Country	Variance X	Variance Y	Chi Square Value
Italy	0.288	0.455	.31
China	0.733	0.573	-5.87
Japan	0.247	0.196	-.15
France	0.548	0.446	-.12
United States	0.299	0.383	.38
United Kingdom	0.781	0.763	-.36
India	1.152	0.554	57.34
Brazil	0.870	1.051	-.30
Germany	0.463	0.614	-.08
Canada	0.670	0.349	-.62

5.2 COVID-19 Cases versus Economic Indicator Linear Regression

5.2.1 Correlation

Similar to the correlation metrics performed in Section 5.1.1, I started by obtaining correlation values between economic attributes and cumulative COVID-19 cases. A histogram was then generated to identify the most correlated attribute amongst the countries (Fig. 10).

Ultimately, the most correlated attribute was determined to be: Domestic Manufacturing Inflation.

5.2.2 Regression Model

The regression model of Cumulative COVID-19 Cases versus Domestic Manufacturing Inflation can be seen in Fig. 11).

Also similar to the previous analysis, while there is a clear "winner" for a correlated attribute, it does not win by much. The following correlation bucket has simply one less country in it. This means it is also possible that we would see valuable information if we were to correlate with that attribute.

5.2.3 Results

While there are fewer countries (out of the top ten GDP countries) that have this corresponding economic attribute, the training set values seemingly fit the linear regression model better. Comparably, to the previous linear regression models, the linear regression fit for these models is a good decision. Likewise, the test set has a strikingly lower deviation from the linear regression model.

On analysis of variation we, however, can see that the variance in x values are extremely higher than that of the variance in y values, and likewise the variance we observed in the other linear regression models. This can seemingly be due to the natural variation of the cumulative cases attribute. Even though the column was standardized, the standardization had to capture values that steadily ranged from 0 to 8.06e+07.

If we look at the Chi-square values for these models, it corresponds to the visual analysis of the regression lines fitting better. The max Chi-square value is hardly over a value of one, for the country Japan, with the min Chi-square value at -.02 for Spain. The average Chi value for this grouping is .25, compared to the -5.36 and 5.06 values that we saw for the previous model.

Cumulative COVID-19 Cases versus Domestic Manufacturing Inflation Chi Square Values			
Country	Variance X	Variance Y	Chi Square Value
Italy	896.69	0.103	0.048
Japan	195.90	0.041	1.074
France	1154.14	0.068	0.030
Spain	918.36	0.245	-0.024
United Kingdom	3043.14	0.074	0.014
Germany	1611.91	0.096	0.356

6 Conclusion

There are several statistical methods of analysis available. They're most typically employed in real-world data exploration to figure out why something happened or why something is true. That was ultimately how I handled this issue set's investigation. I was curious to see how different properties were changed before and after COVID-19, as well as whether COVID-19 may be related to another property. Wouldn't it be interesting to know how many individuals would be affected if some other attribute changed?

Finally, because the linear regression models did not fit appropriately, the examination of economic features looks to be fruitless. Individual data points have too much volatility to justify fitting with a linear regression model (even though it was given a valiant shot). As a result, the average Chi-Square values we saw were significantly higher than the standard deviation of each parameter. In contrast to the economic attribute analysis, the cumulative COVID-19 case analysis appears to have worked well. The data points were low enough invariance and had a high enough correlation value to fit a linear regression model satisfactorily.

7 Appendix

All the data preprocessing and analysis was done on python. here is a link to the jupyter notebook for further refferences. ://drive.google.com/file/d/1GGqktQeUpL4JwBzJJl0uBnVqZUWT8TPK/view?usp=sharing

References

.

1. National Institute of Standards and Technology | NIST. NIST. (2022). Retrieved 13 May 2022, from **??**

2. OECD data. theOECD. (2022). Retrieved 13 May 2022, from ://data.oecd.org/.

3. WHO Coronavirus (COVID-19) Dashboard. Covid19.who.int. (2022). Retrieved 13 May 2022, from ://covid19.who.int/.

Figure 3: Economic Distribution Per Country

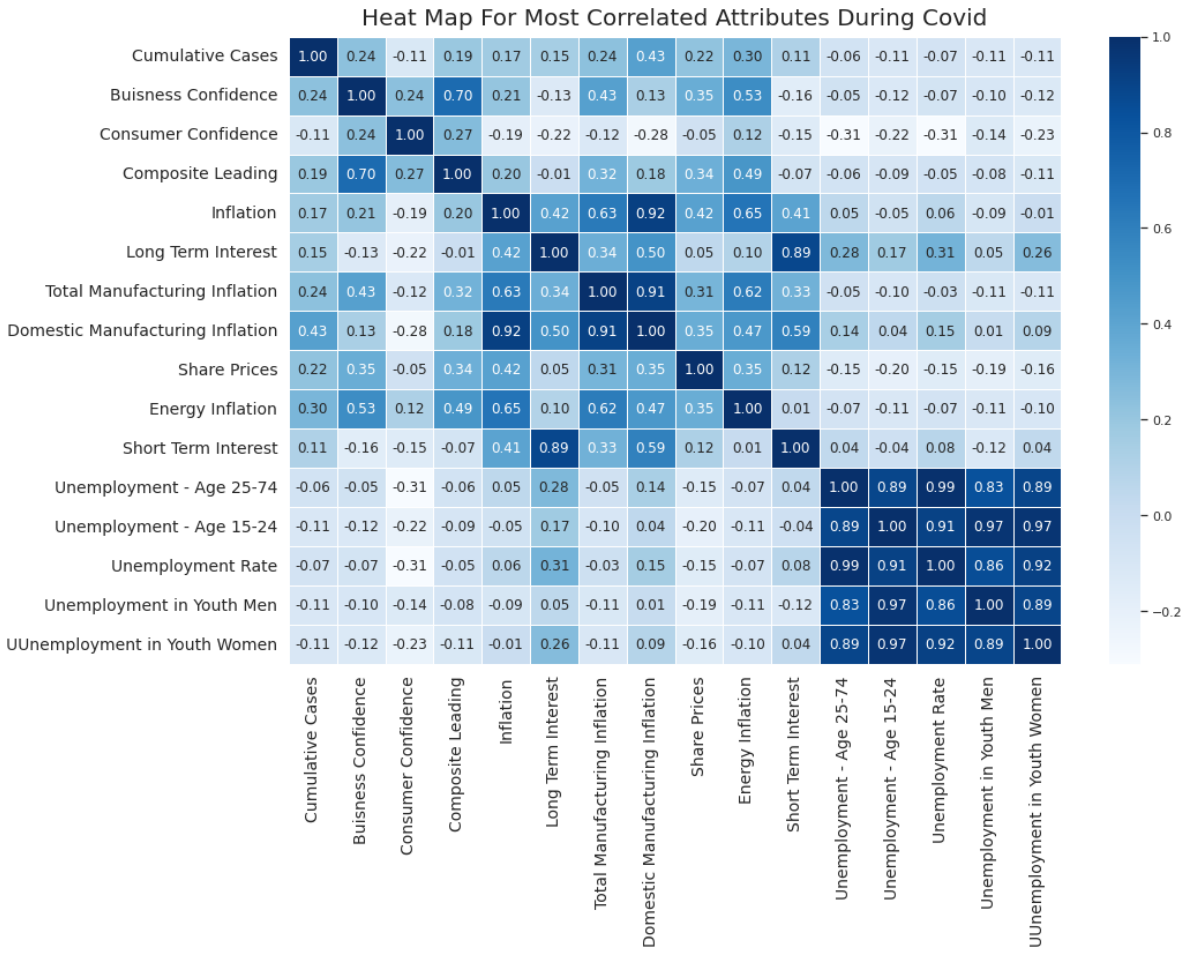


Figure 4: Correlation Heat Map

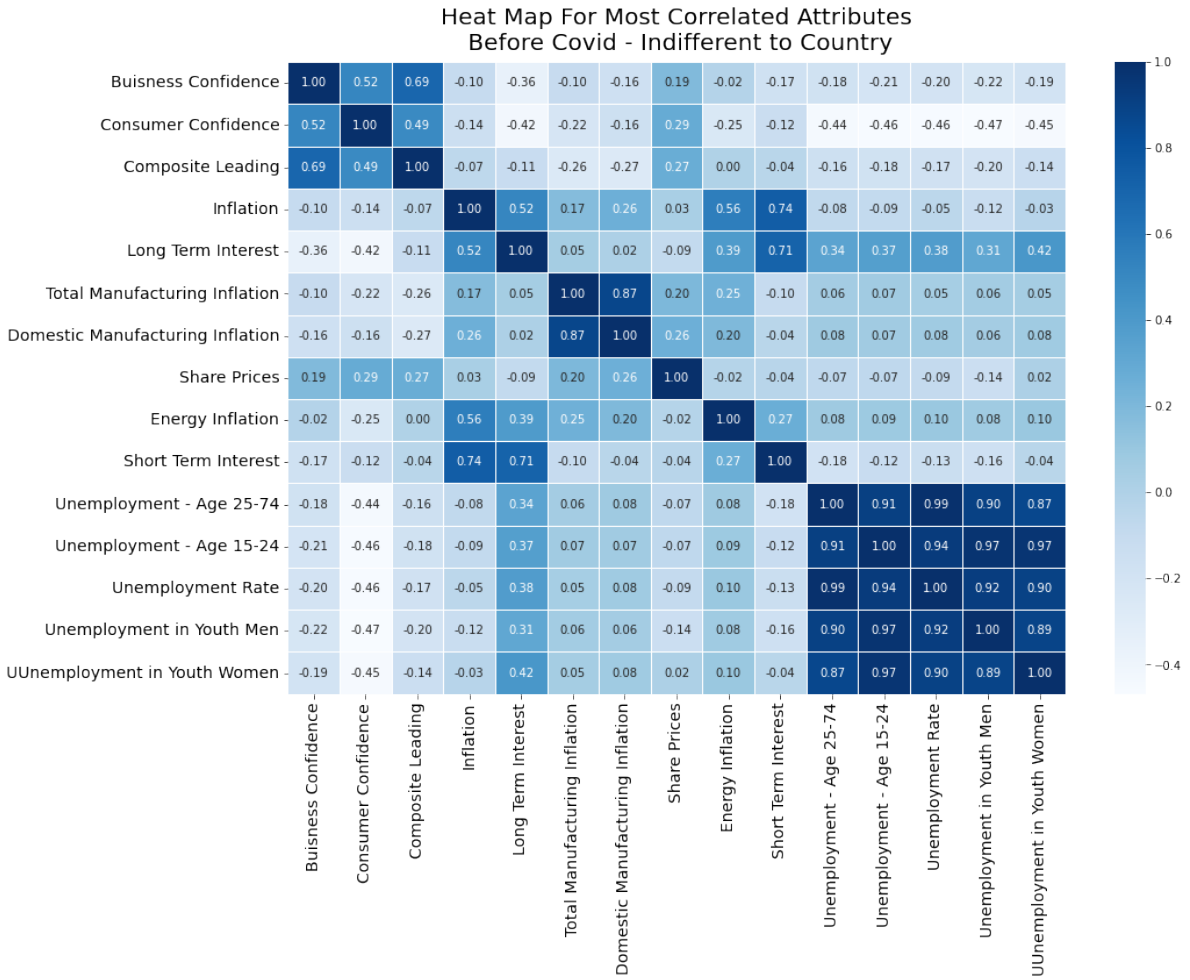


Figure 5: Correlation Heat Map - Before COVID-19

Figure 6: Correlation Heat Map - After COVID-19

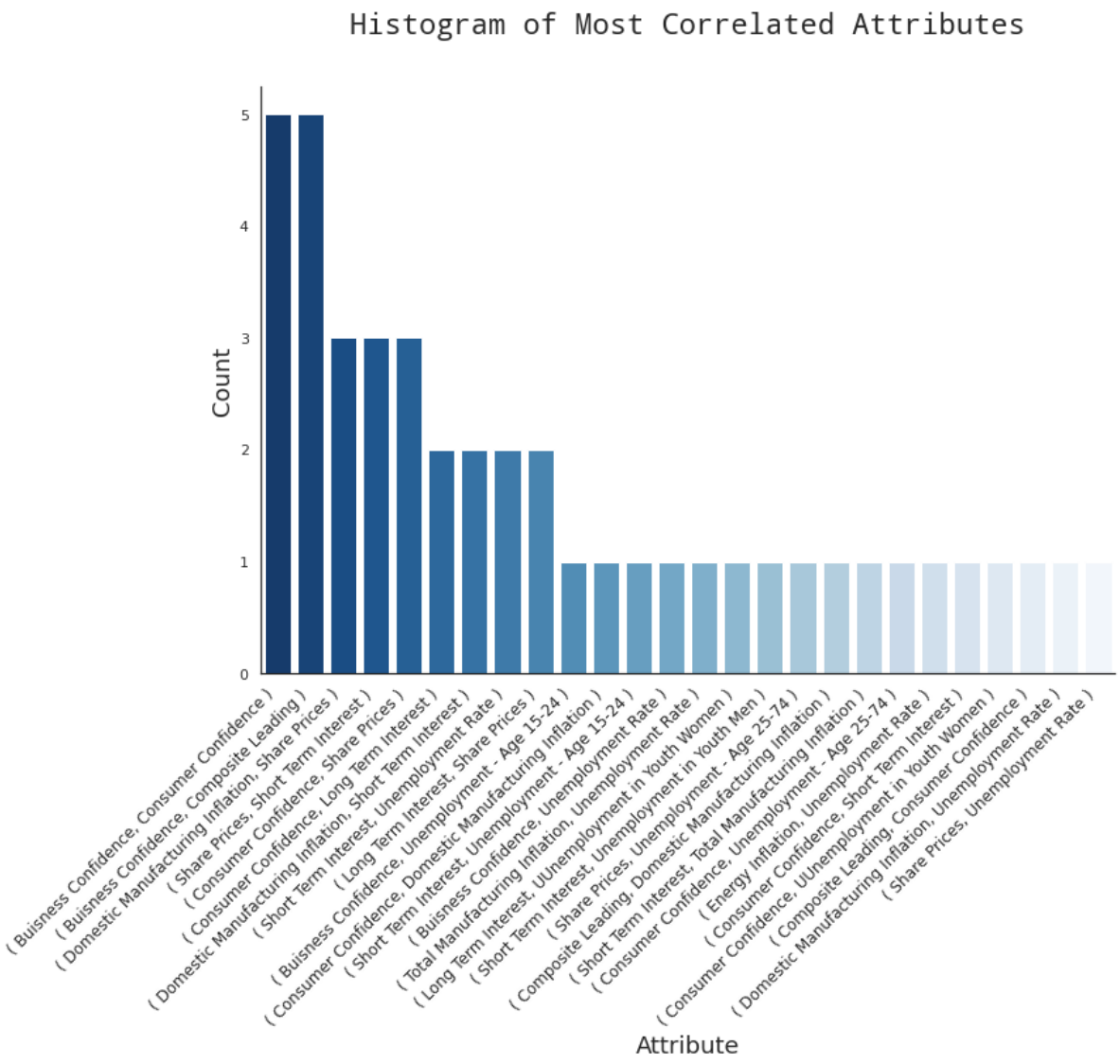


Figure 7: Economic Attribute Correlation Histogram (Per Country)

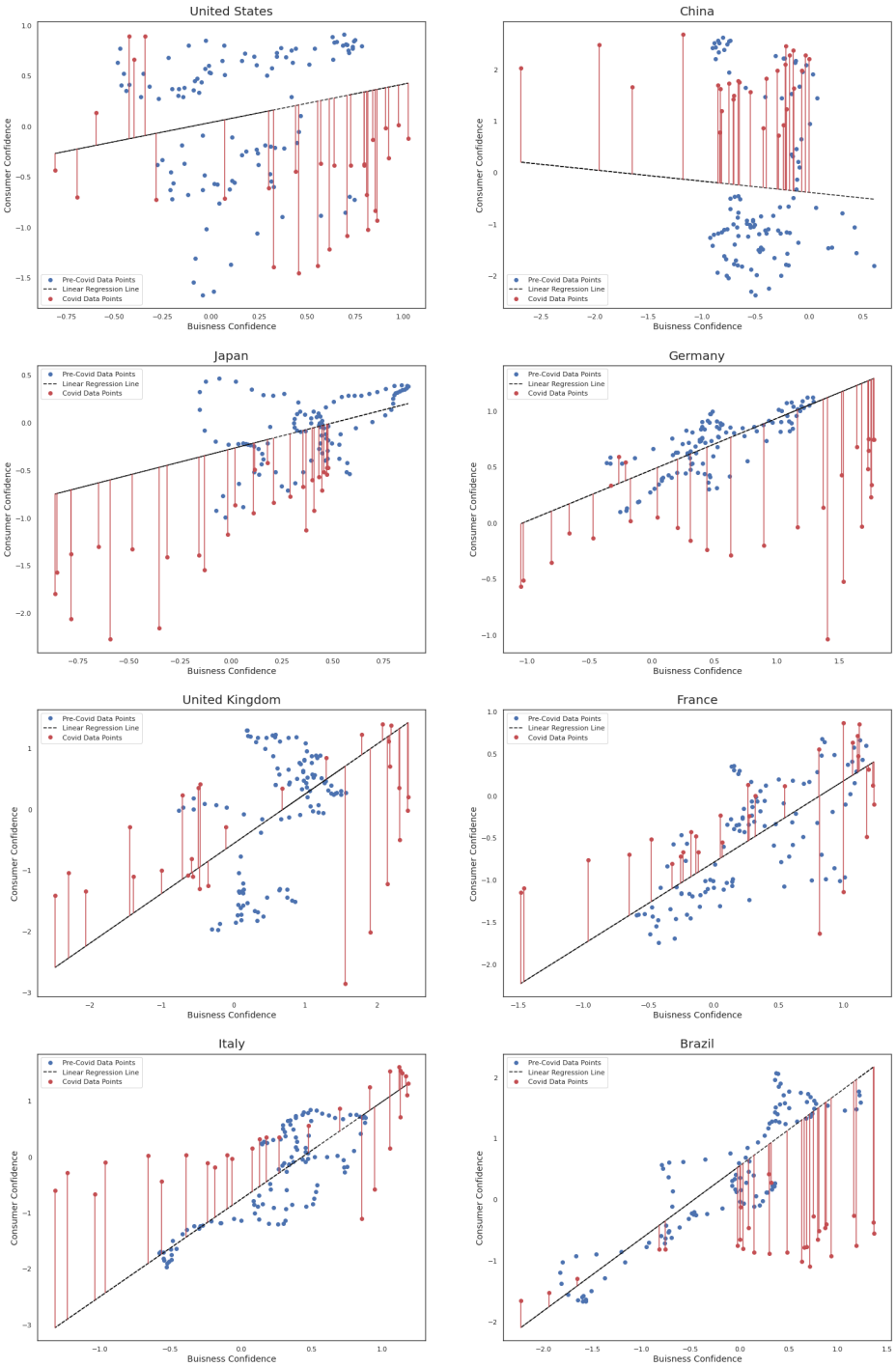


Figure 8: Business Confidence versus Consumer Confidence Linear Regression (Per Country)

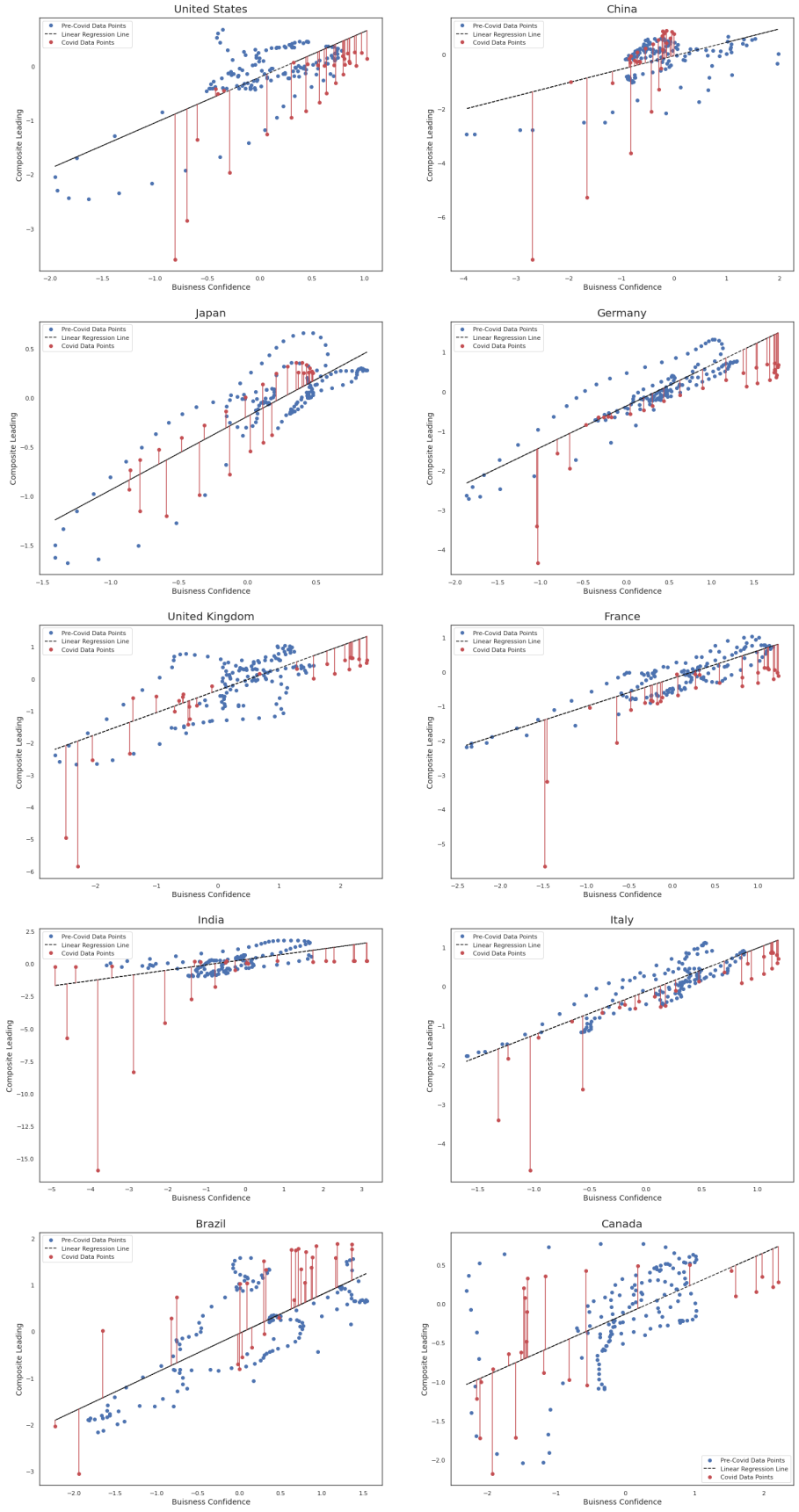


Figure 9: Business Confidence versus Composite Leading Linear Regression (Per Country)

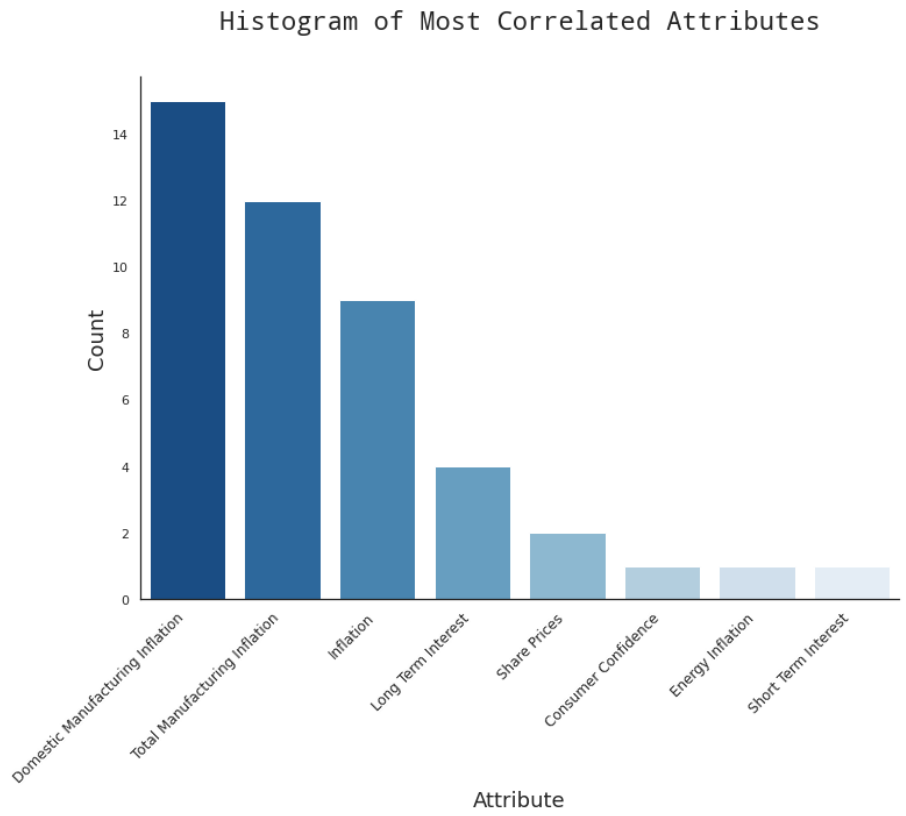


Figure 10: Cumulative COVID-19 Cases and Economic Attribute Correlation Histogram (Per Country)

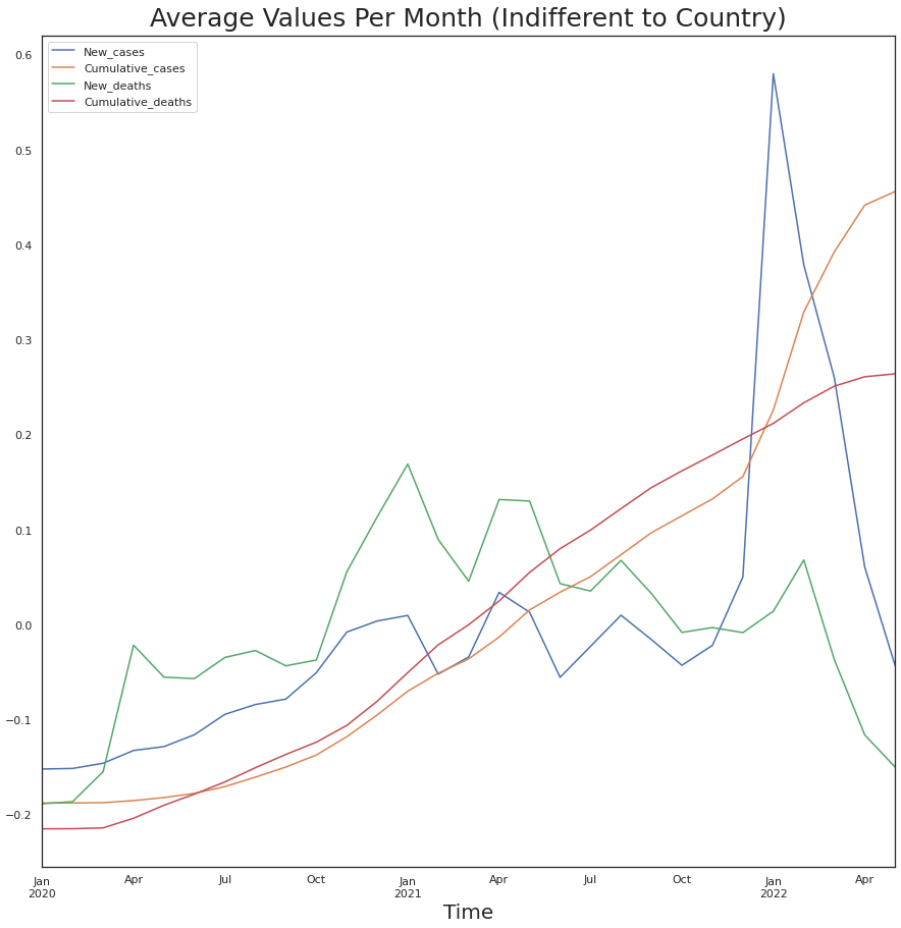


Figure 11: Cumulative COVID-19 Cases versus Domestic Manufacturing Inflation (Per Country)