Abstract

To answer a curiosity that we have whenever we see an economics analysis on the news on which factors contribute to the fluctuations of Vietnamese economics, we decided on this project of finding the correlations between the Vietnam's GDP from 1991 to 2021 and factors such as Import, Export, Seafood, Poultry, Cattle, World Oil Price, Unemployment rate, and Gold Price from the same period. It is frequently stated that the data from the annual GDP can be used as a representation of the economic growth of a nation. However, our society has numerous aspects that can be argued to potentially influence this data. Therefore, this project aims to test the relations between some variables to see if claims and analysis on the news or in articles are correct. In order to do this, we plan to conduct this by putting our collected data into two different data analysis techniques: Principal Component Analysis (PCA) and Linear Regression, which are both very common tools in this field. From the results of these tests, we can observe the differences of the correlations between different variables and compare them to answer the question: what the GDP per capita is constituted by in the previous years for Vietnam. Moreover, we plan to compare the two techniques of analysis to see which can give us a better result on this topic. In the end, there is not enough evidence to help us reach any conclusion about the correlation between GDP and the factors. This can be attributed to the limitations of the techniques we applied in the project. From this project, the awareness on how the data of GDP relating to other factors is calculated and analyzed will hopefully increase.

1. Introduction

1.1 Inspiration

When reading newspapers or watching the news in Vietnam, there are numerous articles reporting on the current trend of the nation's GDP, whether increasing or decreasing. Each source will have a different analysis and reasonings behind it as the fluctuation of GDP can be depending on numerous factors from many aspects of society. This phenomenon prompted us to do this project.

1.2 Main objectives

The main goal for this project is to assess whether certain factors have any correlation to GDP and might be considered a cause for the growth in Vietnamese economics. While doing so, we want to compare the two tools that we are going to be using to test the correlation, which are Principal Component Analysis (PCA) and Linear Regression. With the given results from the two techniques, we can compare and discuss the relations between the factors, as well as finding out what the growth of Vietnam's economics is compiled of. In addition to that, we are able to apply two important data analysis's techniques and see which will give us a more precise answer.

1.3 Approach

As stated above, we will be testing the correlation between some factors and the GDP of Vietnam. Specifically, we are doing our project on the relations between GDP and 8 other factors (Import, Export, Seafood, Poultry, Cattle, World Oil Price, Unemployment rate, and Gold Price). And all of the data we gather are from the database of World Bank and range from 1991 to 2021, which gives us a population of 30 years.

We will be conducting our analysis on the calculation of these two techniques: Principal Component Analysis (PCA) and Linear Regression. From the analysis of these two tools, we can observe the correlations among the variables and understand more about the components that make up the growth of a nation's economics. Additionally, the comparison between the two results of the tools can tell us the precision of these techniques in data analyzing.

2. Principal Component Analysis

As stated in the introduction section, in this project, we investigate the factors that contribute the most to GDP in Vietnam, hence to its economic growth in the last 30 years. The first tool we used in this project is called Principal Component Analysis (PCA) which is also known as dimensionality reduction technique in statistics and data analysis. By using a covariance matrix, PCA measures how each variable in the dataset is correlated with one another. The most important parameters used in PCA are the eigenvalues and eigenvectors. The eigenvectors indicate the directions of new feature space whereas the eigenvalues indicate their magnitude or length. We can determine which direction captures the most variability of our dataset by determining the eigenvector that has the largest eigenvalue. This eigenvector is called the principal component and the principal component that has the eigenvalue smaller than 1 will be deducted in PCA to simplify complex and high-dimensional dataset to a lower-dimensional dataset but still retains the general trends and patterns. Regarding our dataset, eigenvalue for each principal component has been computed using Python and shown in the table 1 below.

Principal component	1	2	3	4	5	6	7	8	9
Eigenvalue	7.620	1.030	0.340	0.218	0.056	0.019	0.010	0.006	0.0006

Table 1: Eigenvalue of each principal component

Since only principal components 1 and 2 have the eigenvalues greater than 1, we deduct from 9 principal components to 2 principal components. These two principal components capture cumulatively 93% of variance of our dataset as shown in figure 1 below. This means 93% of the variance in our dataset can be represented in a 2 dimensional-space.

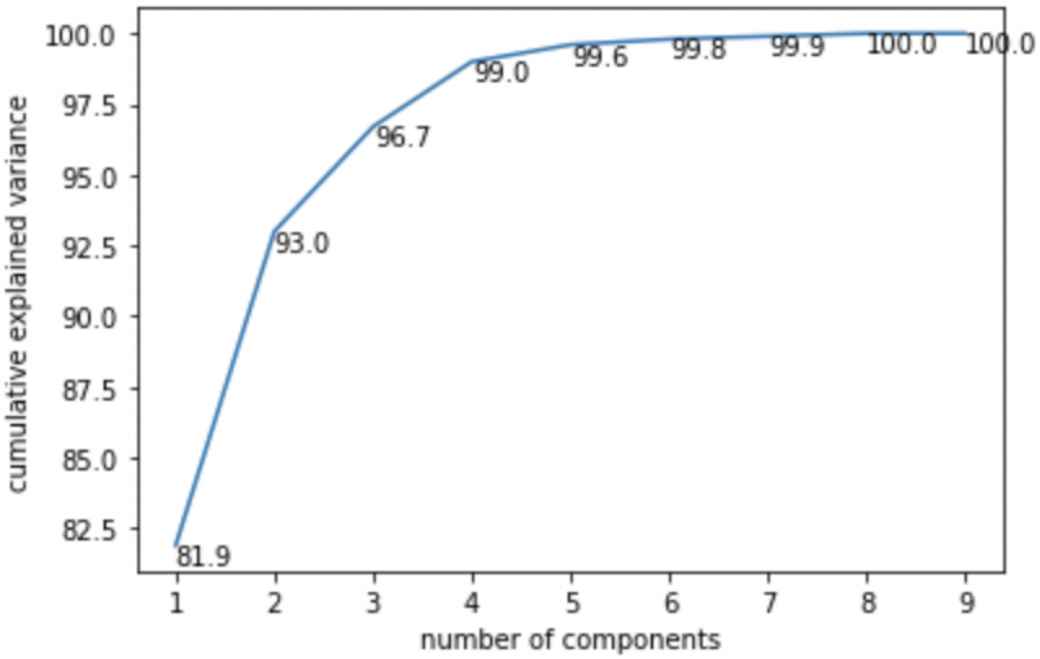


Figure 1: Cumulative explained variance

The correlation between variables can be demonstrated in the correlation circle diagram shown below.

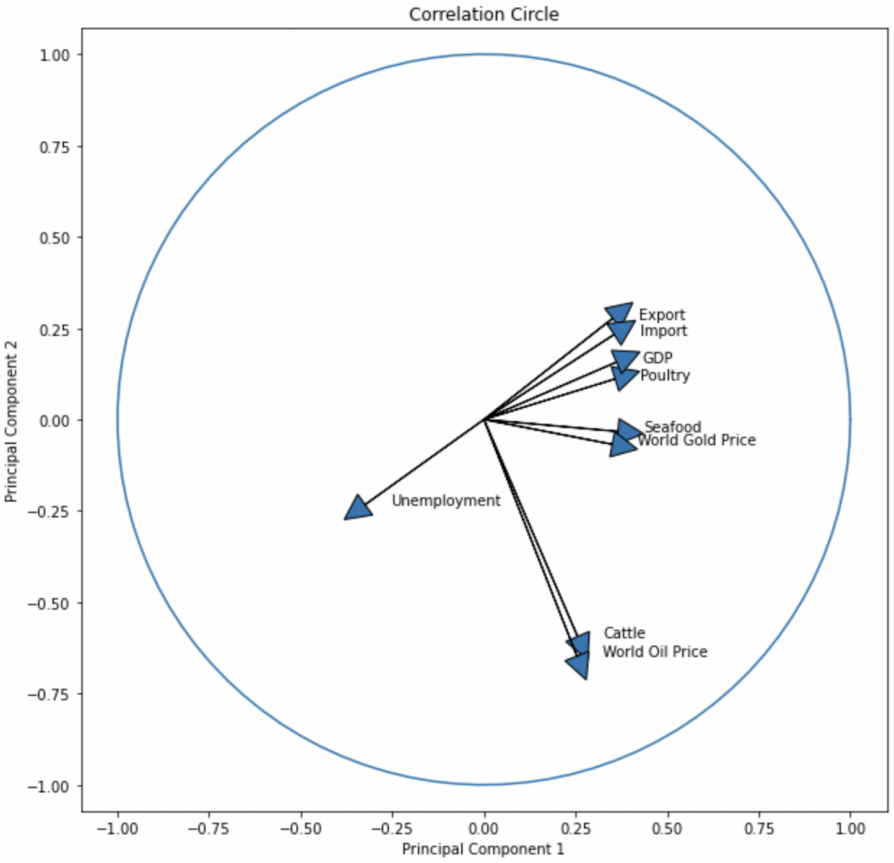


Figure 2: Correlation circle

In the correlation circle, two variables are positively correlated if their vectors are pointing in the same direction. If their vectors are pointing in opposite directions, they are negatively correlated. If their vectors are perpendicular, they are not correlated. The figure 2 shows that GDP is negatively correlated with the unemployment rate since they have opposite directions. However, it is positively correlated with export, import, poultry, seafood, and world gold price as they are all pointing to the right. Poultry has the strongest correlation among the four variables as it is closest to GDP and its magnitude is more fairly equal to GDP. Nonetheless, GDP has no correlation with cattle and world oil prices as they are at a 90 degrees angle.

By using PCA, we have dimensionally reduced our dataset and are able to find the correlation between factors and GDP through a correlation circle. However, the result from PCA is also very subjective and has some drawbacks. For instance, at principal components 1

and 2, GDP is not weighting the most. It weighs the most at principal component 7 with a loading score of 0.729. This indicates that by just using eigenvalues, PCA might deduct our desired principal component. Thus, might lead to a subjective correlation result. In addition, PCA can only reveal which factor is associated with GDP but can not reveal accurately how strong the correlation is. Thus, to assess the result accuracy as well as to determine the correlation scale of each factor to GDP, we have used an additional tool called linear regression to compare and evaluate their correlations.

3. Linear Regression

3.1 Methodology

Continuing working on the same source of data, after selecting variables associated with GDP, we use another tool called ordinary least squares (OLS) to come up with a regression formula for GDP with all of the variables gathered and double-check the results of variables affecting GDP from the PCA method mentioned above.

OLS is one type of method for choosing the unknown parameters in a linear regression model. The logic behind OLS is the usage of rotations of the lines, then finding the rotation that has the least squared difference and putting that as the best fit model. To achieve this, an imaginary straight line with slope 0 is drawn at first. Then, each data point is evaluated in terms of its residual, meaning distance from the point to the line. After that, the sum of the squared residuals are taken for all the data points. With this calculation done, the line is rotated to a certain degree, and the procedure is repeated until the line with the least sum of the squared residuals is found. This is then called the linear regression line.

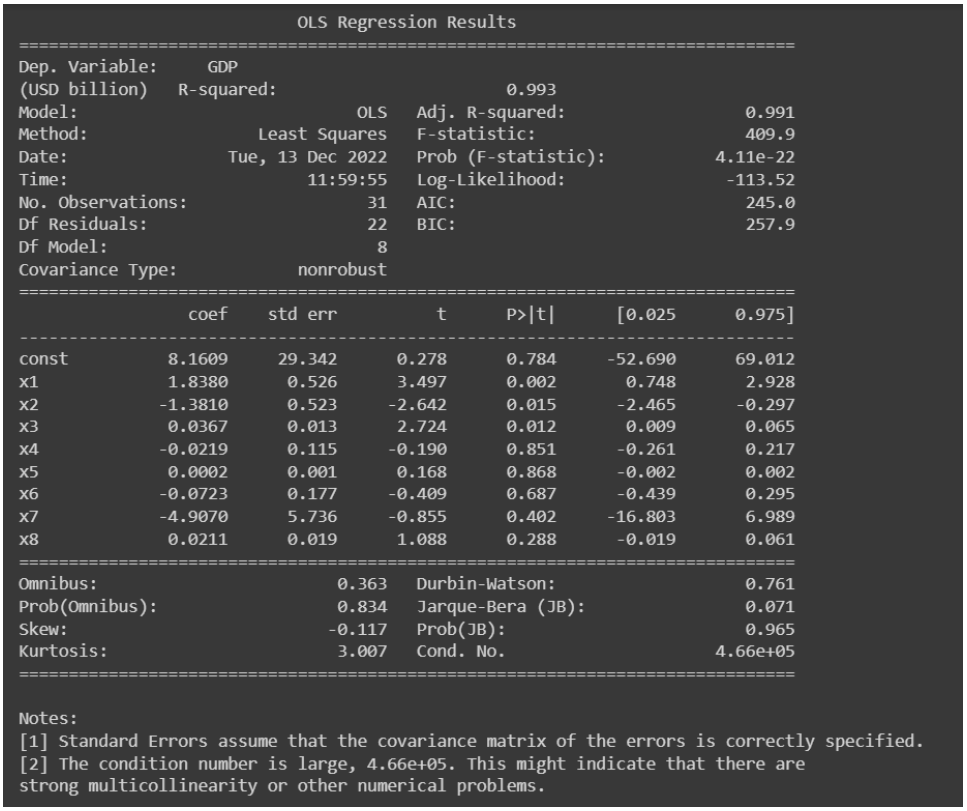
For the specific method of how the OLS Regression Line is produced, we first uploaded our data file, then defined the variables. Afterwards, we stack the data points year

by year, so that there are 31 arrays corresponding to our 31 years of data. Finally, we define the model, then perform the OLS fit to reap the results.

Through the result of the OLS, we can first determine the linear regression line through which the independent variables are used to show the dependent variables. Next, through the coefficients of each variable, we would be able to see the direction as well as the proportion to which the independent variables affect the dependent variable. The strength of each correlation can be shown through the results of individual two-variable regression lines, which is an indicator of whether there is an effect of the independent variable to the dependent variable. Lastly, the p-value of each variable with regards to their effect on the independent variable determines whether or not it is conclusive to say that that specific variable has an effect on the dependent variable. Linking it to our investigation, we would calculate and evaluate the values of the variables' coefficients, their r^2 values and their p-value in order to answer the questions of how these variables affect GDP and whether or not they do affect GDP.

3.2 Results

By conducting the linear regression of data across our 8 independent variables, here are the OLS Regression Results.



OLS Regression Results

Dep. Variable: (USD billion)	GDP	R-squared:	0.993
Model:	OLS	Adj. R-squared:	0.991
Method:	Least Squares	F-statistic:	409.9
Date:	Tue, 13 Dec 2022	Prob (F-statistic):	4.11e-22
Time:	11:59:55	Log-Likelihood:	-113.52
No. Observations:	31	AIC:	245.0
Df Residuals:	22	BIC:	257.9
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.1609	29.342	0.278	0.784	-52.690	69.012
x1	1.8380	0.526	3.497	0.002	0.748	2.928
x2	-1.3810	0.523	-2.642	0.015	-2.465	-0.297
x3	0.0367	0.013	2.724	0.012	0.009	0.065
x4	-0.0219	0.115	-0.190	0.851	-0.261	0.217
x5	0.0002	0.001	0.168	0.868	-0.002	0.002
x6	-0.0723	0.177	-0.409	0.687	-0.439	0.295
x7	-4.9070	5.736	-0.855	0.402	-16.803	6.989
x8	0.0211	0.019	1.088	0.288	-0.019	0.061

Omnibus:	0.363	Durbin-Watson:	0.761
Prob(Omnibus):	0.834	Jarque-Bera (JB):	0.071
Skew:	-0.117	Prob(JB):	0.965
Kurtosis:	3.007	Cond. No.	4.66e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.66e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 3: OLS Regression Results

For the convenience of data interpretation, here is a clarification of the variables represented in the OLS Regressions Results

Variable	Description
x1	Export quantity (USD billion)
x2	Import quantity (USD billion)
x3	Seafood (thousand tonnes)
x4	Poultry (thousand animals)
x5	Cattle (thousand animals)
x6	Global crude oil price (USD per barrel)
x7	Unemployment (%)
x8	Global gold price (USD per tael)

Table 2: Representations on the OLS Regression Results table and their corresponding variables

Given the results from the table, the regression line that we have deduced is:

$$y = 1.838x1 - 1.381x2 + 0.0367x3 - 0.0219x4 + 0.0002x5 - 0.0722x6 - 4.907x7 + 0.0211x8$$

3.3 Interpretation

i. Coefficients analysis

From the regression results, here are the interpretations of the 8 independent variables

Variable	Interpretation
Export quantity (USD billion)	A rise of $1 billion in Vietnamese exports lead to an increase of $1.838 billion in GDP
Import quantity (USD billion)	A rise of $1 billion in Vietnamese imports lead to a decrease of $1.381 billion in GDP
Seafood (thousand tonnes)	A rise of 1000 tonnes of seafood lead to an increase of $36.7 million in GDP
Poultry (thousand animals)	A rise of 1000 poultry animals lead to a decrease of $21.9 million in GDP
Cattle (thousand animals)	A rise of 1000 cattles leads to an increase of $200,000 in GDP
Global crude oil price (USD per barrel)	A rise of 1 USD per barrel of oil leads to a decrease of $72.3 million in GDP
Unemployment (%)	A rise of 1% in unemployment leads to a decrease of $4.907 billion in GDP
Global gold price (USD per tael)	A rise of 1 USD per tael of gold leads to an increase of $21.1 million in GDP

Table 3: Interpretation the OLS Regression Results of corresponding variables

ii. Standard error evaluation

From this table, it can be seen that the standard error value of most variables is not significant, except the variable x7 which represents the value of unemployment rate. Having a high value of standard error like that makes the data of unemployment not reliable. Based on the result illustrated on the table, x1 representing the data of export, x3 representing the data of seafood and x8 representing the data of gold price are the three variables that correlate to GDP the most.

iii. r^2 evaluation

Following the analysis of the standard error, we would have another indicator to show which shows the strength of the correlation between each individual independent variable and the dependent variable to demonstrate the reliability of the impact of change. To reap the coefficient of determination for each variable, we perform separate regressions with each variable, then reap the pearson coefficient r to which we could further evaluate in terms of r^2

The table below shows the coefficient of determination for each of the variables. Their individual regression graphs are included in the appendix part of this paper.

Correlated variable	Coefficient of determination	Strength of correlation
Export quantity (USD billion)	0.96870	Strong positive
Import quantity (USD billion)	0.97210	Strong positive
Seafood (thousand tonnes)	0.95185	Strong positive
Poultry (thousand animals)	0.95368	Strong positive
Cattle (thousand animals)	0.34372	Weak positive
Global crude oil price (USD per barrel)	0.33025	Weak positive
Unemployment (%)	-0.76861	Strong negative
Global gold price (USD per tael)	0.85069	Strong positive

Table 4: Coefficient of determination for each respective variables

Since the coefficient of determination denotes how much of the dependent variable data can be explained using the independent variable, it shows the extent to which one independent variable correlates with the dependent variable, hence giving us a hint on

whether they do affect the dependent variable. Looking at the coefficient of determination for the 8 independent variables, we can say that for most of the variables, the strength of the correlations are high. Therefore, it is inconclusive to say that they do not have any effects on GDP. For the two variables “cattle” and “global crude oil price”, their coefficient of determination with GDP are relatively low. Therefore, it suggests that the impacts that these variables have on GDP are not consistent, and are closer towards having no impact.

iv. p-value evaluation

From the “P > absolute value t” column of the OLS Regression Results, the variables that have an effect on GDP can be decided.

The theory of the p-value evaluation says that if the p-value of the respective variable distribution is smaller than our coefficient of significance, 0.05 in our case, then there is enough evidence to reject the null hypothesis that the variable does not have an effect on the dependent variable, which then answers our question of which variables do have an effect on the dependent variable.

From the results table, it is said that import, export, and seafood are the three variables that have the p-value smaller than our coefficient of significance, which means that there is enough evidence to reject the null hypothesis that these variables don’t have effect on GDP. Based solely on the regression formulas, it can be concluded that import, export and seafood harvest per year respectively have effects on GDP. For the other variables, it is inconclusive to say that they do have effects on GDP.

4. Discussion

To sum up, here are some highlighted points in terms of methodology that you should take away from our project. Firstly, the three methods that we use are Principal Component Analysis (PCA), simple linear regression, and ordinary least squares (OLS), which are used

to measure the correlation of each economic factor to Vietnam's GDP. In the correlation circle of PCA, two variables are positively correlated if their vectors are pointing in the same direction; negatively correlated if their vectors are pointing in opposite directions, and not correlated if their vectors are perpendicular. Regarding the next method, OLS uses rotations of lines to identify which rotation produces the smallest squared difference; having the result, we use it as the best fit model. Finally, the results we get from these methods are not necessarily consistent with each other as each of them, more or less, has its limitations.

4.1 Compare and contrast

To make it easier for conducting any comparisons, here are the brief results of the PCA, simple linear regression and OLS method:

v. PCA: GDP is negatively correlated with the unemployment rate; positively correlated with export, import, poultry, seafood, and world gold price, among which poultry has the strongest correlation; and has no correlation with cattle and world oil prices.

vi. Simple linear regression: Except for unemployment rate with a strong negative correlation to GDP, the remaining variables are positively correlated to GDP with all correlation coefficients above 0.5. The strongest positive correlation to GDP is the cumulative imports, with its correlation coefficient equaling 0.98595.

vii. OLS: Exports, seafood, cattle, and world gold price positively correlate to GDP. Imports have a negative correlation to GDP. Poultry, world oil price, and unemployment rate negatively correlate to GDP. Still, since their standard errors are higher than their coefficients (with that of the unemployment rate being the highest), we are yet to deliver any significant conclusion.

In conclusion, there are mostly contradictions within the two linearity methods as well as between the PCA test and linear regression methods. Those conflicts make it almost impossible to explain or conclude the relationships between each variable to Vietnam's GDP. Nevertheless, the results share their similarities: the positive correlations of export, seafood, and world gold price to GDP.

4.2 Methodological concerns and limitations

In this project, we have only observed how each factor correlates to Vietnam's GDP and how strong these correlations are by applying the learned tools which are PCA, simple linear regression, and OLS. Using simple linear regression means making many assumptions and eliminating the effects of potential confounders; therefore, we have faced many contradictions when interpreting the results between the methods. The OLS method may help us capture more variables that contribute to GDP, but it fails to explain whether the correlations are causational or not. Moreover, some variables in the OLS regression result have relatively large standard errors - even higher than the coefficients, which leads to no conclusion being made. We also find that PCA is not an effective tool since it has some drawbacks: By just using eigenvalues, PCA might deduct our desired principal component; hence, it might lead to a subjective correlation result. In addition, PCA can only reveal which factor is associated with GDP but can not show accurately how strong the correlation is. Lastly, our lack of interpreting skills and applied tools lead this project not to explain anything but rather have a general observation of some factors of Vietnam's economy, as well as understand how to apply the tools we have learned to analyze real-world situations.

4.3 Suggested further research

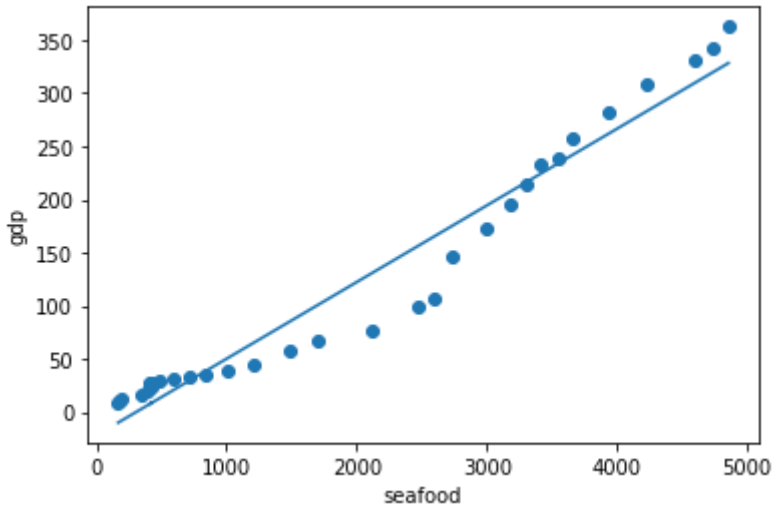
In further studies, we would want to try better approximation models that are capable of explaining the contradictions between our used methods. Moreover, we will learn to use tools

and obtain interpretation skills to analyze the data; therefore, will be able to determine whether the correlation is causation or not. Lastly, it is important to search for the reasons, which we are all curious about: Why do we have such results? How did Vietnam's economy change from 1991 to 2021 regarding the effect of each variable?

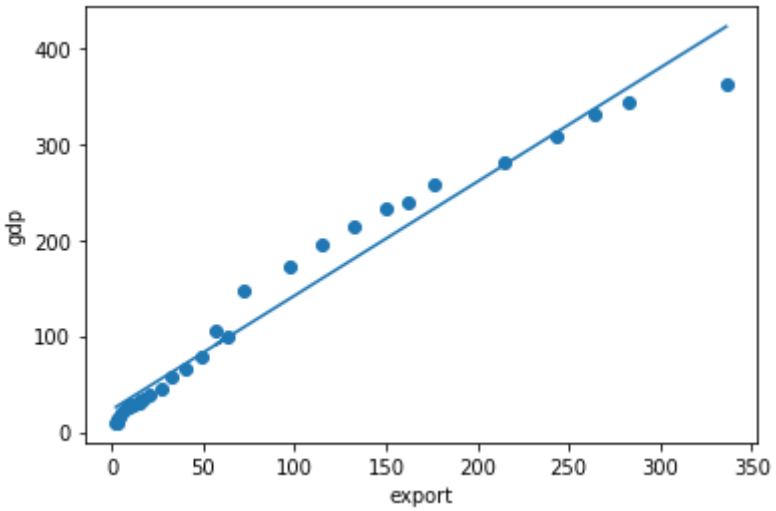
5. Conclusion

Based on the basic statistics concepts and practical coding skill in Python and the encourage to crave for stepping out of the comfort zone, we applied Principal Component Analysis (PCA) and Linear Regression tool; however, there is insufficient evidence from our data analysis to assess which factors most affect the growth of GDP in Vietnam over the last 30 years. It could be due to the crudeness and inadequacies of the techniques we employed in this research, and we need more sophisticated statistical approaches as well as much professional knowledge to answer this question.

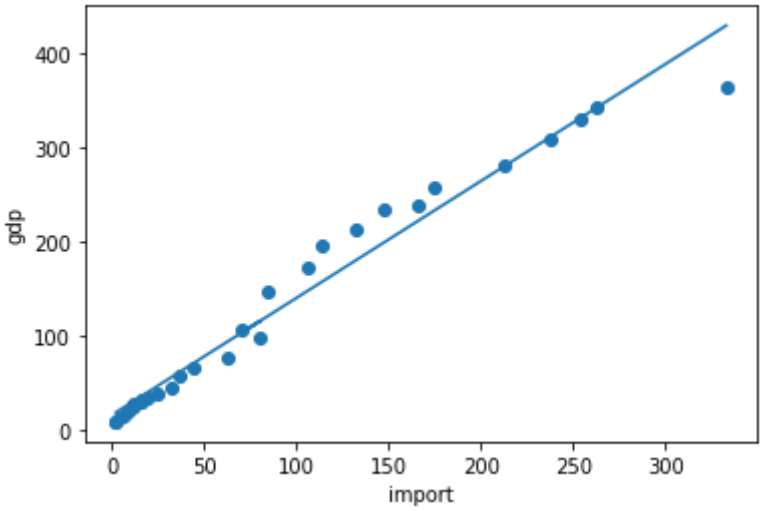
6. Appendix



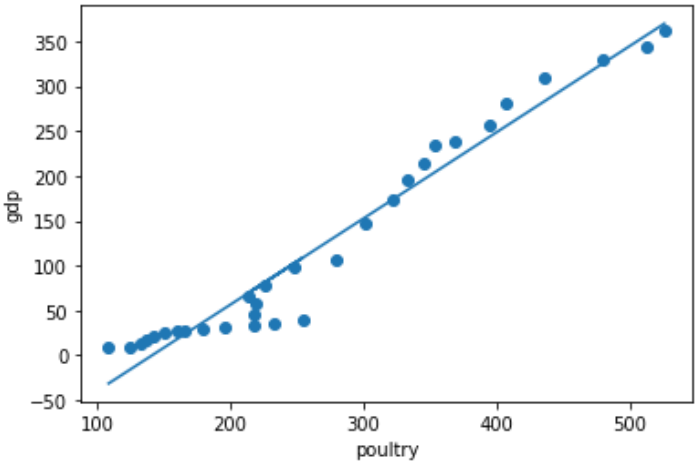
Appendix 1: Seafood by GDP linear regression



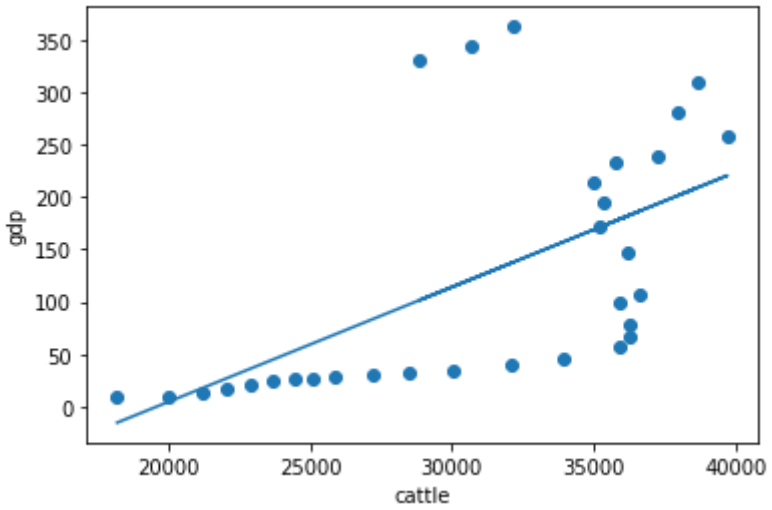
Appendix 2: Export by GDP linear regression



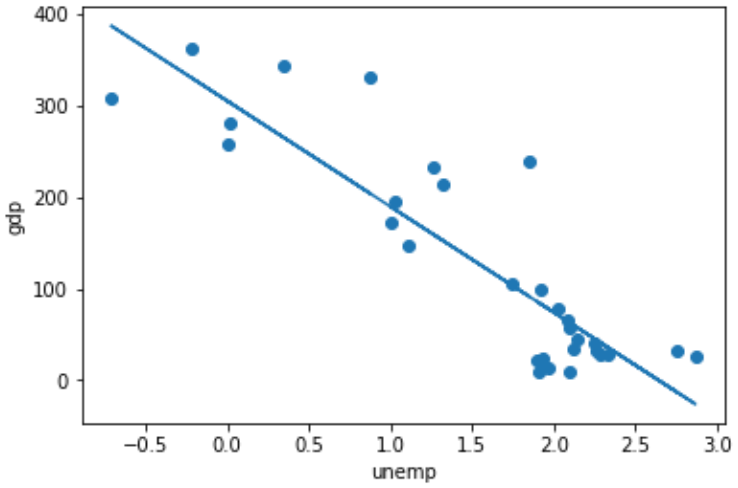
Appendix 3: Import by GDP linear regression



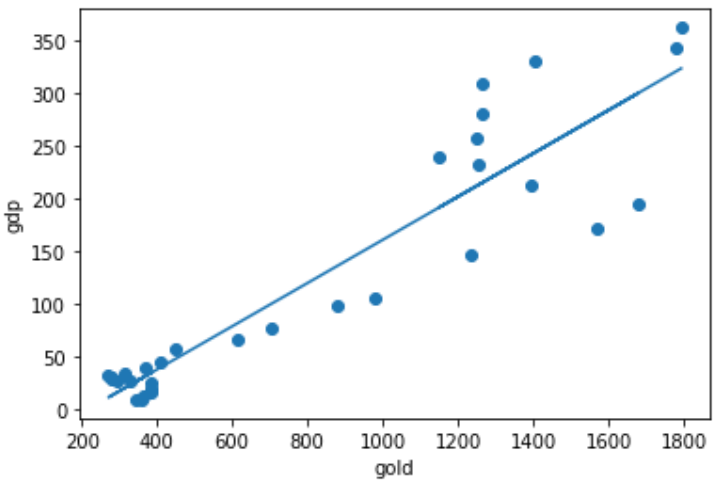
Appendix 4: Poultry by GDP linear regression



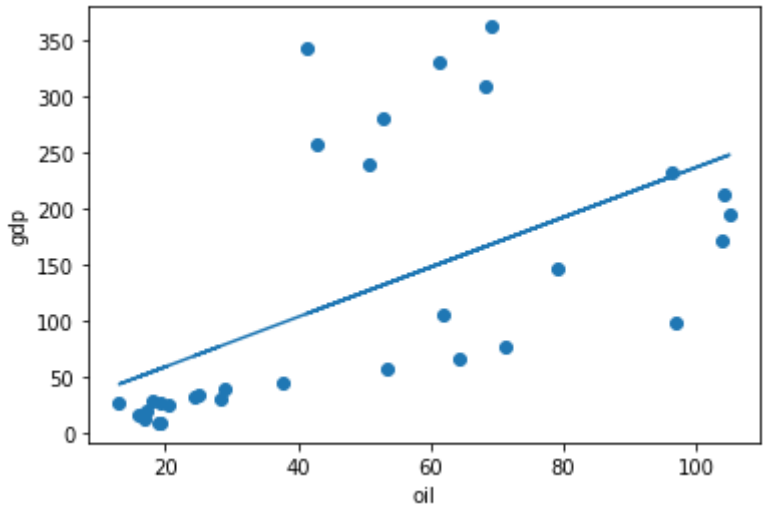
Appendix 5: Cattles by GDP linear regression



Appendix 6: Unemployment by GDP linear regression



Appendix 7: Gold price by GDP linear regression



Appendix 8: Oil price by GDP linear regression

Collab link of PCA section:

https://colab.research.google.com/drive/15aXKmunj9OqcO8A4dxzToGbo2igqNVs7

Collab link of Linear Regression section:

https://colab.research.google.com/drive/1_FJc6ncyqQXaYgd_WJxMwb_B1MEJX4WI

7. References

Commodity Markets. (n.d.). World Bank. https://www.worldbank.org/en/research/commodity-markets

Dịch vụ. General Statistics Office of Vietnam. (2022, February 22). Retrieved December 2022, from https://www.gso.gov.vn/thuong-mai-dich-vu/

GDP (current US$) - Vietnam | Data. (n.d.). https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=VN

joshstarmer. (2017, July 24). Linear regression, clearly explained!!! YouTube. Retrieved December 16, 2022, from https://www.youtube.com/watch?v=nk2CQITm_eo&t=483s

Lao Động. General Statistics Office of Vietnam. (2022, February 22). Retrieved December 2022, from https://www.gso.gov.vn/lao-dong/

Monthly Gold Prices (1979-2021). (2021, August 22). Kaggle. https://www.kaggle.com/datasets/odins0n/monthly-gold-prices

Nông, Lâm Nghiệp và Thủy Sản. General Statistics Office of Vietnam. (2022, February 22). Retrieved December 2022, from https://www.gso.gov.vn/nong-lam-nghiep-va-thuy-san/

PCA problem / How to compute principal components / KTU Machine learning. (2020, October 10). YouTube. https://www.youtube.com/watch?v=MLaJbA82nzk

Principal Component Analysis (PCA): Illustration with practical example in Minitab. (2020, February 29). YouTube. https://www.youtube.com/watch?v=f0_UWY3R8CY

Principal Components Analysis (PCA) & Interest Rate Modeling. (2012, May 19). YouTube. https://www.youtube.com/watch?v=OV9zVbytujs

StatQuest: Principal Component Analysis (PCA), Step-by-Step. (2018, April 2). YouTube. https://www.youtube.com/watch?v=FgakZw6K1QQ

Unemployment, total (% of total labor force) (modeled ILO estimate) - Vietnam | Data. (n.d.). https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?locations=VN

Vu, T. (2017, June 15). Bài 27: Principal Component Analysis (phần 1/2). Tiep Vu's blog. https://machinelearningcoban.com/2017/06/15/pca/