**UNIVERSITY OF INFORMATION TECHNOLOGY**

**INFORMATION SYSTEM FACULTY**

*IS403.O11.TMCL*

**FINAL PROJECT**

**BUSINESS ANALYTICS**

**Analyze macroeconomic factors affecting the labor market**

**Lecturer: Tran Van Hai Trieu**

**Students:**

| Phan Công Thành | 20521923 |
| --- | --- |
| Võ Kiều My | 20521624 |
| Trần Trọng Nghĩa  Hà Minh Thông  Phan Phi Thuyền  Nguyễn Xuân Yến  Lê Anh Tuấn  Hồ Chí Minh, 11/2023 | 20521287  20521979  20522005  20520357  20520948 |

Analysis and prediction of Vietnam's labor market

**Final Report**

***Team 3 - IS403.O11.TMCL***

# **Feedback**

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

# **Table of contents**

[**Feedback 3**](#_qzv2w1viubej)

[**Table of contents 4**](#_oxd8xjbhxu9j)

[**ACKNOWLEDGEMENT 6**](#_yhnuatj3041j)

[**Chapter 1: Introduction 7**](#_w9s0tsdkdoho)

[I. Rationale of the study 7](#_pkdj5t1ru883)

[II. Aims of the study 7](#_zcj0hmtce054)

[**Chapter 2: Research overview 10**](#_74g27r2d12va)

[**Chapter 3: Research methodology 12**](#_yna2jifmmb18)

[I. Data collection 12](#_yayl2rh0h6np)

[1. Data sources 12](#_pfnya2rxiwux)

[2. Data cleaning and preprocessing procedures 12](#_j4uwljcq80cu)

[2.1. Data collection and preparation procedures: 12](#_ye0c52d80jwf)

[II. Statistical and analytical methods 13](#_qxp0f42x6764)

[1. Explore data 13](#_yejln4mpocf7)

[1.1. Central indicators 13](#_racx8rivitfl)

[1.2. Dispersion (Variation) indices 13](#_drgcngynqk40)

[1.3. Shape indicators 13](#_eyg3qyxtys2e)

[1.4. Graph 14](#_6qel8dxtmo5h)

[2. Hypothesis testing methods 14](#_sh90ixq7bkaa)

[2.1. Testing equality of two population means 14](#_lzdhmoos3tne)

[2.2. Comparing two population variances using F-test 14](#_40pkskqhi1q8)

[2.3. Chi-square Test 15](#_gxu7b3es9v2l)

[2.4. One-way ANOVA 15](#_b4kqfun5nso1)

[3. Analysis model 18](#_nay8bqhzogu2)

[3.1. Linear Regression 18](#_vec5nmoxyfq5)

[3.2. Autoregressive Integrated Moving Average (ARIMA) 19](#_pl6nsv4uc4vo)

[3.3. XGBoost 20](#_3pxotldhcsfv)

[**Chapter 4: Implementation and research results 22**](#_r0g4ih51ixrb)

[I. Descriptive statistics 22](#_jvze6oscu01)

[1. Result descriptive statistics 22](#_q8143w207nu0)

[2. Graph 23](#_u0o0xsjmzwqq)

[2.1. Histogram 23](#_n0306yab2n5v)

[2.1. Line 27](#_knn4qqly74wn)

[2.1.1. CPI 27](#_1nh6zx4ypvks)

[2.1.2. FDI In & FDI Out 30](#_jma4qvmi0ksn)

[2.1.3. Workers & Unemployed rate 31](#_okhi9pd48pos)

[II. Hypothesis testing 32](#_2o7h71dihjdq)

[2.1 Testing equality of two population means 32](#_q1j187rjpqvc)

[2.2 Comparing two population variances using F-test 33](#_7u1cslprgsjf)

[2.3 Chi-square Test 36](#_7u1cslprgsjf)

[2.4 One-way ANOVA 38](#_5dnc1stcwhpo)

[III. Execution Analysis 41](#_1athqsklamoy)

[3.1. Linear regression 41](#_gn8l8jg6lwmu)

[3.1.1. Implementation steps 41](#_qdg9z0hlnf79)

[3.1.2. Result 48](#_yiysif9gqy3h)

[3.2. Multiple Linear Regression 51](#_uxlqj8bbeym2)

[3.2.1. Implementation steps 51](#_xj9noz2pgo2w)

[3.2.2. Result 56](#_ypzn1ni6qnf)

[3.3. ARIMA 58](#_dqay78pqz9hl)

[3.3.1. Implementation steps 58](#_yjylsq7d4lbp)

[3.3.2. Result 62](#_j4hvkc977xyt)

[3.4. XGBoost 66](#_kghsmocfurru)

[3.4.1. Implementation steps 66](#_i6ipigsrp1a1)

[3.4.2. Result 71](#_eg3k31q1qdm5)

[**Chapter 5: Conclusion 72**](#_wk0s985qwy7d)

[I. Strengths and Limitations 72](#_rnpx7vr4txf)

[II. Future Development Directions 72](#_3943waeq0lxu)

[**References 73**](#_vl5cs1mq3t50)

# 

# **ACKNOWLEDGEMENT**

Dear lecturer,

We would like to express our deep gratitude to the teachers who supported us during the project implementation process. With a spirit of dedication and high responsibility, the teachers helped us overcome difficulties and challenges and successfully complete our project research work.

First, we would like to thank the teachers for conveying their professional knowledge clearly, thoroughly and enthusiastically. The answers to questions, useful advice and sincere feedback from teachers have helped us grasp knowledge more quickly and effectively.

In addition, we also want to express our gratitude to our teachers for their important support in editing, evaluating and approving our project. The teachers' comments, corrections and suggestions for improvement helped us understand our mistakes and carry out our projects better.

We believe that the knowledge and experience that teachers have imparted will help us develop our careers and become useful people for society. Once again, we would like to sincerely thank our teachers for all the good things they have brought to us during the past time.

Wishing all teachers good health, success and happiness!

Best regards,

***-Team 3-***

# **Chapter 1: Introduction**

## **I. Rationale of the study**

The contemporary labor market in Vietnam is subject to multifaceted influences stemming from technological advancements, economic trends, and socio-cultural dynamics. Recognizing the pivotal role of employment in individual livelihoods and overall societal development, this study endeavors to analyze and forecast the unemployment rate in Vietnam over the next 5 to 10 years.

This research is prompted by the critical need for accurate and forward-looking insights into the labor market, particularly with regard to unemployment. As a key indicator of economic health, unemployment rates not only reflect the challenges faced by job seekers but also hold implications for social well-being and national economic stability. By examining historical patterns, identifying influencing variables, and employing predictive modeling techniques, this study aims to provide stakeholders, including policymakers, businesses, and educational institutions, with a nuanced understanding of the potential shifts in Vietnam's labor market.

In addition to historical factors, the study recognizes the importance of considering emerging trends and potential disruptors. The increasing integration of Vietnam into the global economy, technological advancements, and evolving industry landscapes necessitate a comprehensive approach to predicting unemployment. By investigating the impact of these factors, the research aspires to assist in proactive workforce planning, skill development initiatives, and the formulation of policies that can mitigate the challenges associated with unemployment.

In summary, the rationale for this study is deeply rooted in the urgency of comprehending and predicting the unemployment rate in Vietnam. Through a meticulous analysis and forecasting of unemployment over the next 5 to 10 years, this research seeks to equip stakeholders with valuable insights, fostering informed decision-making and facilitating the development of strategies to address the evolving dynamics of the Vietnamese labor market.

## **II. Aims of the study**

This study aims to contribute to discussions on the current labor market issues in Vietnam by providing a reliable macroeconomic study. It includes predictions about the labor market (unemployment rate and average income) in Vietnam for the next 5 to 10 years, employing scientific methods such as linear regression, multiple linear regression, and ARIMA models, and relying on reputable data. The study seeks to clarify issues related to unemployment rates stemming from economic fluctuations and contribute to the development of macroeconomic policies regarding unemployment in Vietnam.

* **Methodological Integration:** The primary objective of this study is to integrate multiple quantitative methodologies, including linear regression, multiple linear regression, and ARIMA (AutoRegressive Integrated Moving Average), to predict the unemployment rate in Vietnam. By employing a combination of these models, the research aims to leverage their respective strengths in capturing linear relationships, multivariate dependencies, and time-series dynamics for a more robust and accurate forecasting framework.
* **Enhanced Predictive Modeling:** Building upon established statistical techniques, the study seeks to enhance predictive modeling precision. The utilization of linear regression and multiple linear regression will enable the exploration of complex relationships between multiple variables, while ARIMA, with its time-series analysis capabilities, will account for temporal dependencies, ensuring a more nuanced and dynamic prediction of unemployment trends in the Vietnamese labor market.
* **Variable Significance Assessment:** The research aims to extend beyond model application by conducting a rigorous assessment of variable significance within the context of unemployment prediction. Through the regression models, the study seeks to identify which factors contribute most significantly to variations in the unemployment rate, offering insights into the key drivers that warrant focused attention from policymakers and stakeholders.
* **Optimized Forecasting:** With a focus on practical applications, the study aims to optimize the forecasting accuracy of unemployment rates using the selected models. This involves refining model parameters, validating against historical data, and iteratively improving the models to ensure their reliability for future predictions, thereby supporting evidence-based decision-making.
* **Policy Implications from Model Insights**: The integration of linear regression, multiple linear regression, and ARIMA is not only a methodological choice but also a strategic one to derive actionable policy insights. By extrapolating from model outputs, the study aims to generate policy recommendations tailored to the specific influences identified by each model, providing a more nuanced and targeted approach to addressing unemployment challenges in Vietnam.
* **Practical Application of Academic Insights:** Beyond academic contribution, the study aims to demonstrate the practical application of advanced statistical models in addressing real-world challenges. By showcasing the utility of linear regression, multiple linear regression, and ARIMA in forecasting unemployment, the research aims to bridge the gap between theoretical advancements and actionable insights for policymakers and practitioners in Vietnam.

In summary, this study aspires to methodologically innovate by integrating linear regression, multiple linear regression, and ARIMA to refine unemployment predictions in Vietnam. Through this integration, the research aims to provide a sophisticated and practical framework for forecasting, variable significance assessment, and policy recommendation development, thereby contributing to both academic discourse and the effective management of labor market dynamics.

## 

## 

# 

# **Chapter 2: Research overview**

Some relevant studies:

* **Vietnam's labor market in the fourth quarter of 2022 by the Department of Statistics.** It states that the labor market is slowly recovering, with an increase in the number of employed individuals and average monthly income compared to the previous quarter and the same period last year. However, the document also mentions an increase in the unemployment rate and job shortage among the working-age population compared to the previous quarter, although it has decreased compared to the same period last year. Additionally, the labor force participation rate in the fourth quarter of 2022 was 2.32%.
* **Analysis of GDP using Linear Regression** **by Ryan Mayolo in 2018**. The article demonstrates the use of linear regression to analyze the relationship between GDP and the stock S&P 500. The results show a moderate positive correlation between GDP and the stock S&P 500, with some industries having stronger relationships. However, the stock S&P 500 is not the best index to predict GDP, as industries such as finance and industry have better relationships. This study provides a basis for policymakers, investors, and business leaders to use this information to make informed business decisions.

[*Analysis of GDP using Linear Regression (lynchburg.edu)*](https://digitalshowcase.lynchburg.edu/cgi/viewcontent.cgi?article=1062&context=utcp)

* **Application of ARDL model for examining the relationship between**

**unemployment and economic growth in Vietnam by Nguyen Thi Thu Ha.** The document is a study examining the relationship between unemployment and economic growth in Vietnam from 1986 to 2019. The study uses the ARDL model to analyze data. The results indicate that unemployment has a negative impact on economic growth in both the short and long term. However, the study also concluded that there is no cause-and-effect relationship between the two variables. The study provides recommendations to balance this relationship.

[*Applying the ARDL model to evaluate the relationship between unemployment and economic growth in Vietnam (ou.edu.vn)*](https://journalofscience.ou.edu.vn/index.php/econ-vi/article/view/959/1498?fbclid=IwAR13VcB6oBWJz1T8Xa8kMGbxw3_sKws73bvZdEFtW3d17D-O6xCUqD23rdQ)

* **Multiple linear regression used to analyze the correlation between GDP and some variables** **by Prof. Univ. Dr. Constantin Anghelache in 2016**. The research article focuses on the application of multiple linear regression analysis to examine the relationship between GDP and various other variables. The author explains the simple linear regression method and the multiple linear regression method, providing an example of a multiple linear regression model to predict GDP based on independent variables such as inflation rate, average annual income, and average annual net income. The author also presents the methodology and data used in the study and concludes with the findings of the identified multiple linear regression model.
* **The Effects of Unemployment and Insecure Jobs by Jonas Voßemer and Michael Gebel.** The article discusses the impact of unemployment and insecure jobs on human development and health, with a focus on the role of labor market policies. The authors used linear random-intercept models and pooled linear regression models with country and round fixed effects to test their hypotheses. The study found that labor market policies play a crucial role in shaping the experience of unemployment, but are less important for workers in insecure jobs. The article also suggests different explanations for these finding
* **Using linear regression to establish empirical relationships by Marno Verbeek in 2017**. The research article introduces the use of linear regression to determine the empirical relationship between one variable and a set of other variables. The linear regression is estimated using the ordinary least squares method, which is the best tool for linear prediction within a given sample. The article also emphasizes the need for caution in understanding and interpreting the linear regression model, especially in determining causal relationships.
* **Economic Growth Analysis and Forecasting towards unemployment rate using Arima mode** **by Nuwairah Syazwani Mohamed Ally Jinnah Sahib** focuses on the economic growth analysis and forecasting of the unemployment rate. The study employs the ARIMA (AutoRegressive Integrated Moving Average) model to analyze and predict trends in unemployment.
* **Forecasting Unemployment in Russia Using Machine Learning Methods by Urmat Dzhunkeev**. Research is centered around forecasting unemployment in Russia through the application of machine learning methods. The study likely explores the utilization of advanced computational techniques to predict and analyze trends in unemployment rates within the Russian context.

## 

# 

# **Chapter 3: Research methodology**

## **I. Data collection**

### **1. Data sources**

* Macroeconomic data sets are taken from many sources and articles reporting economic data, but the majority of data is taken from [finance.vietstock.vn/du-lieu-vi-mo](http://finance.vietstock.vn/du-lieu-vi-mo) for the period from January 2000 to September 2023.

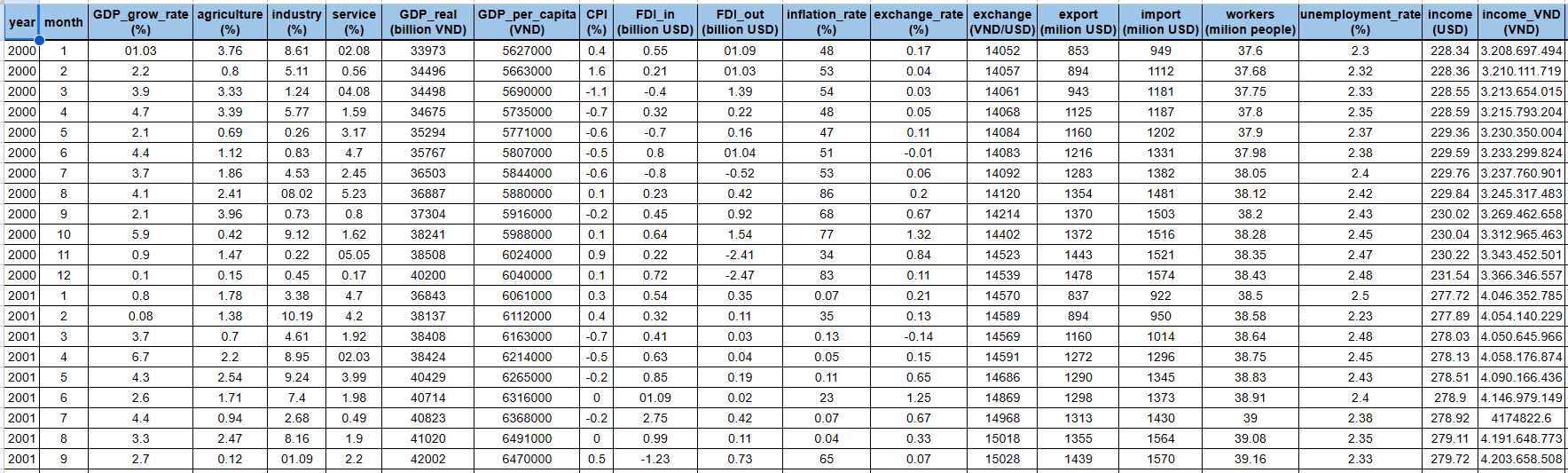
### **2. Data cleaning and preprocessing procedures**

#### ***2.1. Data collection and preparation procedures:***

* Setting up the data structure:

| **Name** | **Type** | **Label** | **Measures** |
| --- | --- | --- | --- |
| GDP\_grow\_rate | Numeric | GDP grow rate | % |
| industry | Numeric | Industry | % |
| agriculture | Numeric | Agriculture | % |
| service | Numeric | Service | % |
| GDP\_real | Numeric | GDP real | billion VND |
| GDP\_per\_capita | Numeric | GDP per capita | VND |
| CPI | Numeric | Consumer Price Index | % |
| FDI\_in | Numeric | Direct Foreign Investment In | billion VND |
| FDI\_out | Numeric | Direct Foreign Investment Out | billion VND |
| inflation\_rate | Numeric | Inflation rate | % |
| exchange\_rate | Numeric | Exchange rate | % |
| exchange | Numeric | Exchange | VND/USD |
| export | Numeric | Export | million VND |
| import | Numeric | Import | million VND |
| workers | Numeric | Workers | million people |
| unemployment\_rate | Numeric | Unemployment rate | % |
| income | Numeric | Income | USD |
| income\_VND | Numeric | Income VND | VND |

***2.2. The resulting data were cleaned:*** [team3\_data\_cleaned](https://docs.google.com/spreadsheets/d/1vEOoAVwO2Vn2XBTMRDXFqMM8AfKLVzBd4Xw05IWa_XU/edit?usp=sharing)



## **II. Statistical and analytical methods**

### **1. Explore data**

#### ***1.1. Central indicators***

* Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.
* Mode is defined as the value that has a higher frequency in a given set of values. It is the value that appears the greatest number of times.
* Median is the middle value of the given list of data when arranged in an order.

#### ***1.2. Dispersion (Variation) indices***

* Variance is a measure of how data points differ from the mean. In other words, a variance measures how far a set of data (numbers) is spread out from their mean (average) value.
* Standard deviation is the positive square root of the variance and a measure of how spread out the data is.
* Range is the difference between the lowest and highest values.

#### ***1.3. Shape indicators***

* Skewness is a measure of the asymmetry of a distribution.
* Kurtosis is a measure of the tailedness of a distribution. Excess kurtosis is the tailedness of a distribution relative to a normal distribution.

#### ***1.4.*** *Graph*

* Histograms: Histograms that represent the distribution of a set of data. It represents the frequency of values within defined intervals.
* Line: chart form to show the development process and development dynamics of an object or a group of objects over time.

### **2. Hypothesis testing methods**

#### ***2.1. Testing equality of two population means***

Hypothesis testing is to evaluate a hypothesis about the population to accept or reject based on the sample obtained.

When testing the equality of two population means, you typically use a t-test. The two-sample t-test is a statistical method used to determine if there is a significant difference between the means of two independent groups.

* Some notation for hypothesis testing.
  + The null hypothesis is denoted by H0 and the alternative hypothesis is denoted by Ha .
  + The researcher wishes to support Ha and doubts H0 .
  + Assume H0 is true, find a standard for the testing of T (which is a known random quantity to be distributed).
* Issue:

H0 : m1 = m2

Ha : m1 > m2

* + The domain accepts H0 : T <= za
  + The domain rejects H0 : T > za

#### *2.2. Comparing two population variances using F-test*

When comparing two population variances, you typically use the F-test for equality of variances. The F-test is a statistical test that assesses whether the variances of two populations are significantly different.

Why would we want to compare two population variances? There are many situations, such as in quality control problems, where you may want to choose the process with smaller variability for a variable of interest

H0 : σ1 = σ2

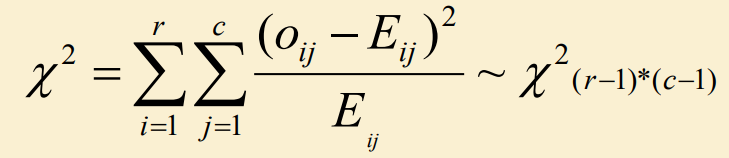
Ha : σ1 σ2

* The domain accepts H0 : F < fα (fα : F Critical)
* The domain rejects H0 : F >= fα

It's important to note that the F-test assumes normality of the populations being compared. If the populations are not normally distributed, alternative tests or transformations of the data may be considered.

#### *2.3. Chi-square Test*

To test the independence between two variables, we use the Chi-square random distribution based on the Pearson chi-square statistical parameter to compare number of observed results with the expected results by using the formula as follows:



Testing for independence between two variables

* H0 : Two variables are independent (if the condition of formula above happens).
* Ha : Two variables are related to each other.

The chi-square test is widely used in fields such as biology, social sciences, and business to analyze the relationship between categorical variables and to assess goodness of fit. It is important to ensure that the assumptions of the test are met, such as the categories being mutually exclusive and the observations being independent.

#### *2.4. One-way ANOVA*

An ANOVA test is a way to detect if survey or experiment results are significant. In other words, it helps us to figure out if we need to reject the null hypothesis or accept the alternate hypothesis.

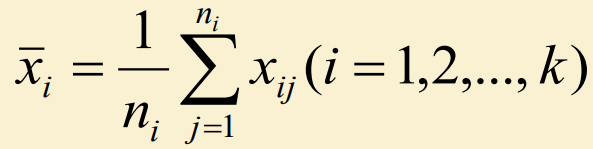
One-way ANOVA analysis of variance is commonly called a one-factor analysis of variance.

* One-factor analysis of variance: only one factor (control variable) is considered to determine its influence on another factor. The considered factor is used to classify observations into different subgroups.

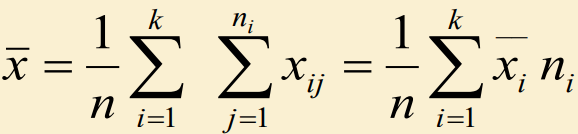
One-factor analysis of variance is used to test the hypothesis of homogeneity of mean values. We have the following steps.

**Step 1: Compute the average value (xi ) for each group and the average value (x) for all.**

The average value (xi ) for each group

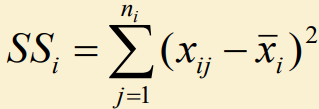


The average value (x) for all.

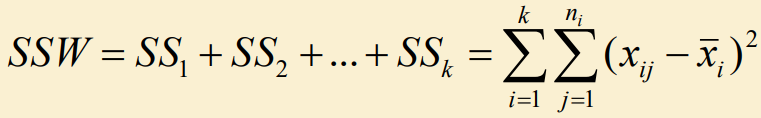


**Step 2: Compute the quantities representing variation of SSW and SSG.**

SSi (Sum of Square) is a quantity representing the internal variation of i group which is computed by the formula as follows:

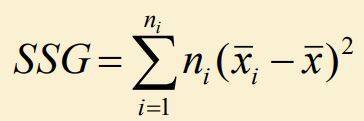


Sum of Squares Within Group (SSW)



SSW is the sum of squares of the differences between each observation and the mean of the group to which that observation belongs. SSW are variations that are not caused by a control factor used to divide the groups.

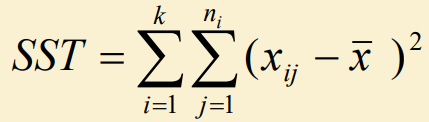
Sum of Squares between Groups (SSG)



SSG indicates the variation due to the differences between groups, i.e., the variation caused by the factor under study.

Sum of Square Total (SST)

* SST is the sum of squared deviations between each observation and the average of all observations.



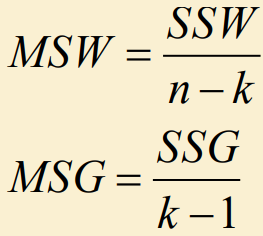
We have, **SST = SSG + SSW**

Formula’s meaning: Variation of values compared to the mean value (SST) = variation due to research factors (SSG) + variation due to other factors (SSW).

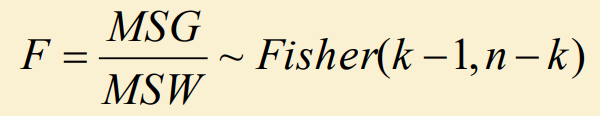
**Step 3: Compute estimates for a variance of population (k), MSW, and MSG.**

It needs to divide SSW and SSG by the corresponding number of degrees of freedom, so we have:

* Mean Sum of Squares Within (MSW)
* Mean Square Between Groups (MSG)



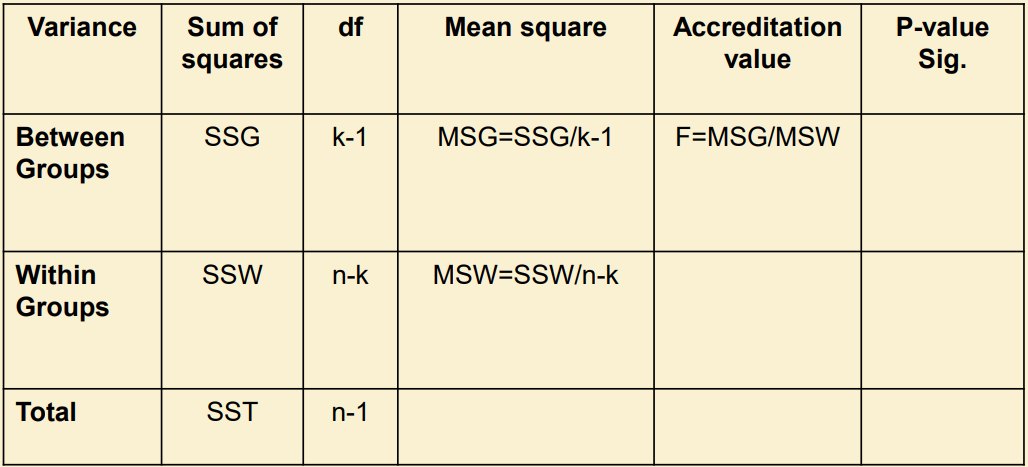
**Step 4: Compute the F-test, so we have the following formula as follows:**



With df = (k-1, n-k), significance level ( = 5%)

* Use the F-table to look up f (k-1,n-k), that denotes fcritical
  + If Fstat > fcrit : reject H0
  + Vice versa, if Fstat < fcrit : accept H0

The results of one-factor variance analysis are shown in the following table:



ANOVA allows for the comparison of means across multiple groups while controlling the overall Type I error rate. It is commonly used in experimental designs with more than two conditions or levels of a factor. The choice between one-way ANOVA and other methods depends on the study design and the nature of the data.

### **3. Analysis model**

#### ***3.1. Linear Regression***

* Linear Regression is an algorithm of machine learning, based on supervised learning. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable used to predict the other variable's value is called the independent variable.
  + Independent variables are the variables used to explain or predict the changes in the dependent variable. They are factors, characteristics, or conditions that are believed to have a relationship with the dependent variable.
  + The dependent variable is the variable that depends on one or more other variables and is used to measure the changes in the independent variables. It is the outcome or phenomenon that we want to explain using the independent variables.

The multivariate linear regression formula:

y = β0 + β1X1 + β2X2 + β3X3 + … + βkXk + ε

where:

● **Y** is the dependent variable

● **X1, X2, X3**, … are the independent variables

● **β0**is the intercept term

● **β1,β2, β3**, … is the coefficient of the independent variables

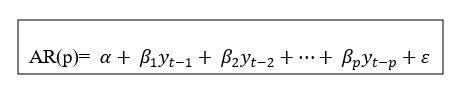
● **ε** is the error term

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

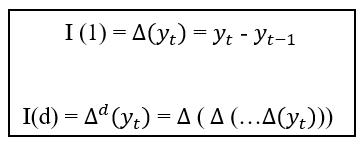
* The regression coefficient that leads to the smallest overall model error. The "Least Squares" method is used to determine the error that (residuals) to be the minimal, i.e., the standard deviation of the residuals to be minimized.
* The t statistic of the model.
* The associated p-value, which indicates how likely it is that the t statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables were true.

#### ***3.2. Autoregressive Integrated Moving Average (ARIMA)***

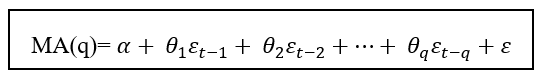
* ARIMA (Autoregressive Integrated Moving Average) model is one of the most popular time series analysis methods, used to predict future values based on past values of time series. ARIMA is used to model and predict non-stationary time series, i.e. trends, divergences, or trends in variance.
* ARIMA consists of three main components:
  + AR (Autoregression): The autoregressive part, which models the relationship between present and past values of a time series. The p-order delay is the backward value p time step of the sequence. Long or short delay in AR process depends on delay parameter p.



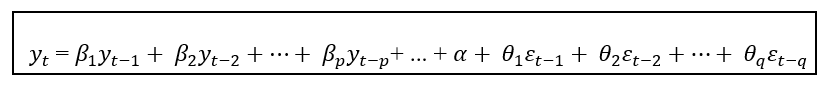
* + I (Integrated): Is the process of co-integration or taking the difference. The general requirement of the algorithms in the time series is that the sequence must be stationary. To form a stationary series, one of the simplest methods is to take the difference.



* + MA (Moving Average): Moving average, which is the process of shifting or changing the mean of a series over time.



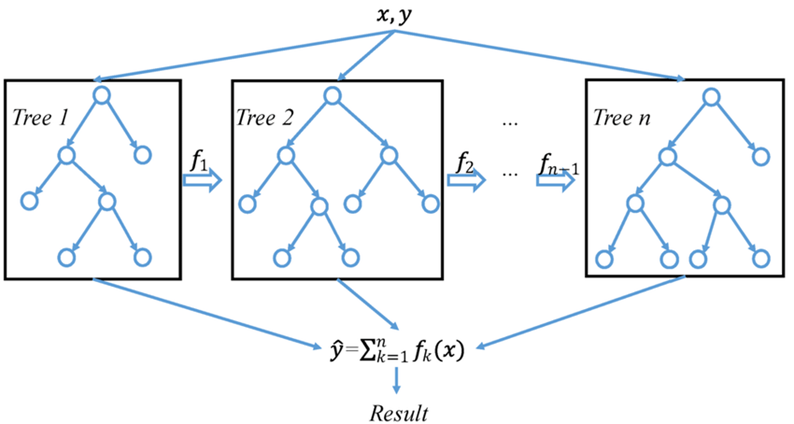
Regression equations for ARIMA (p, d, q):



**Conclude**: ARIMA is a combined model of two autoregressive processes and moving average. Past data will be used to forecast future data. Before training the model, it is necessary to convert the series to a stationary series by taking the first difference or logarithm.

#### ***3.3. XGBoost***

* XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. The same code runs on major distributed environments (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.
* Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset’s features.



* XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed. With XGBoost, trees are built in parallel, instead of sequentially like GBDT. It follows a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set.

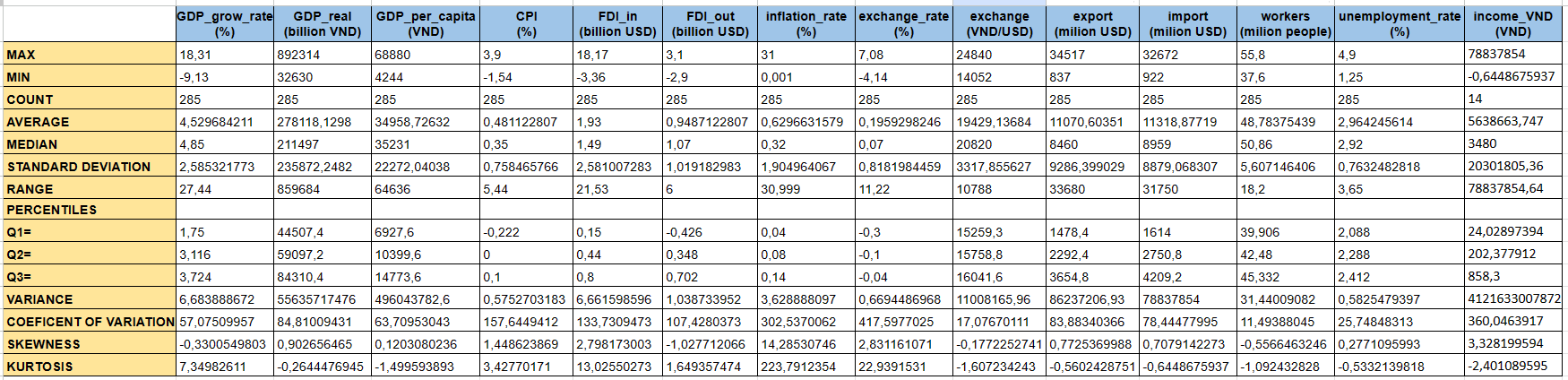
# 

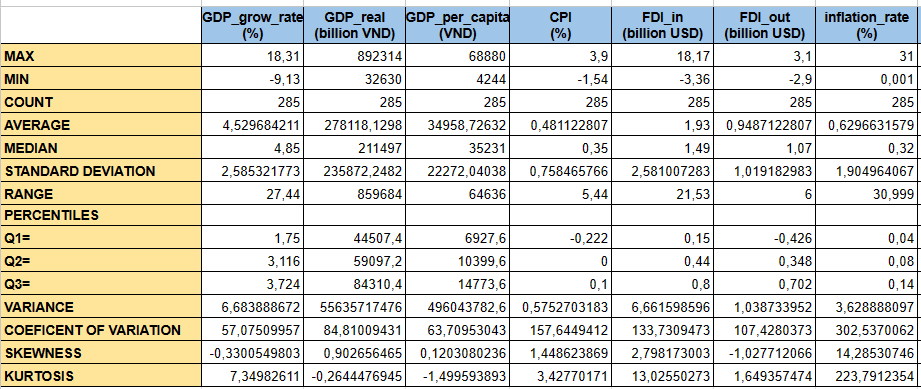
# **Chapter 4: Implementation and research results**

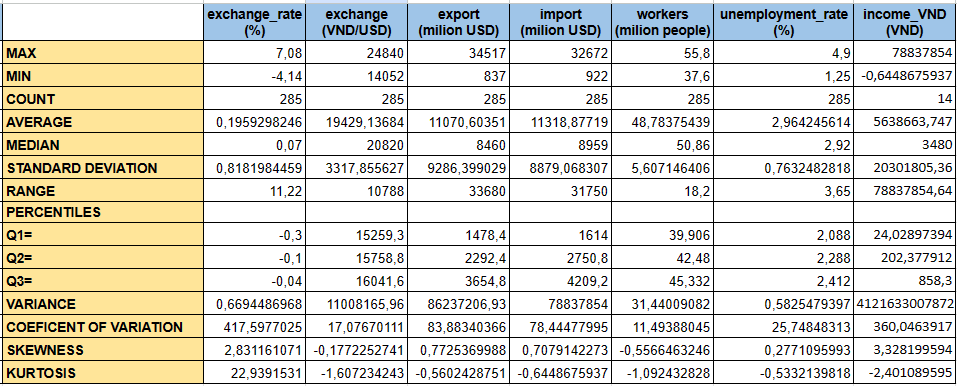
## **I. Descriptive statistics**

### **1. Result descriptive statistics**

* Detailed information visit the link: [team3\_result\_descriptive statistical](https://docs.google.com/spreadsheets/d/1hTgtVUOdGmxpCsISj3CdO3-Ofvg_WwwZAHRB9dkWoow/edit?usp=sharing)

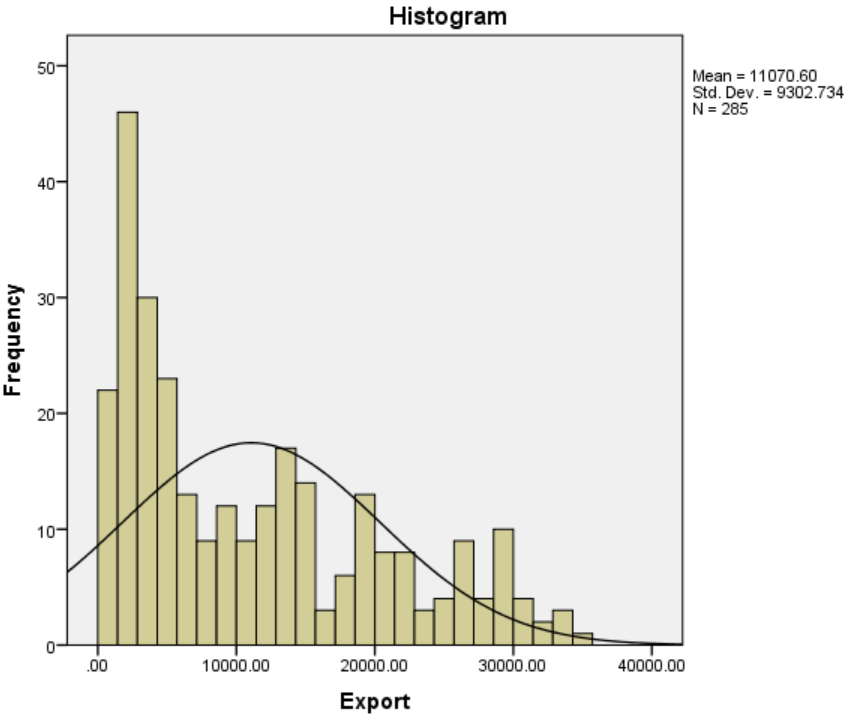




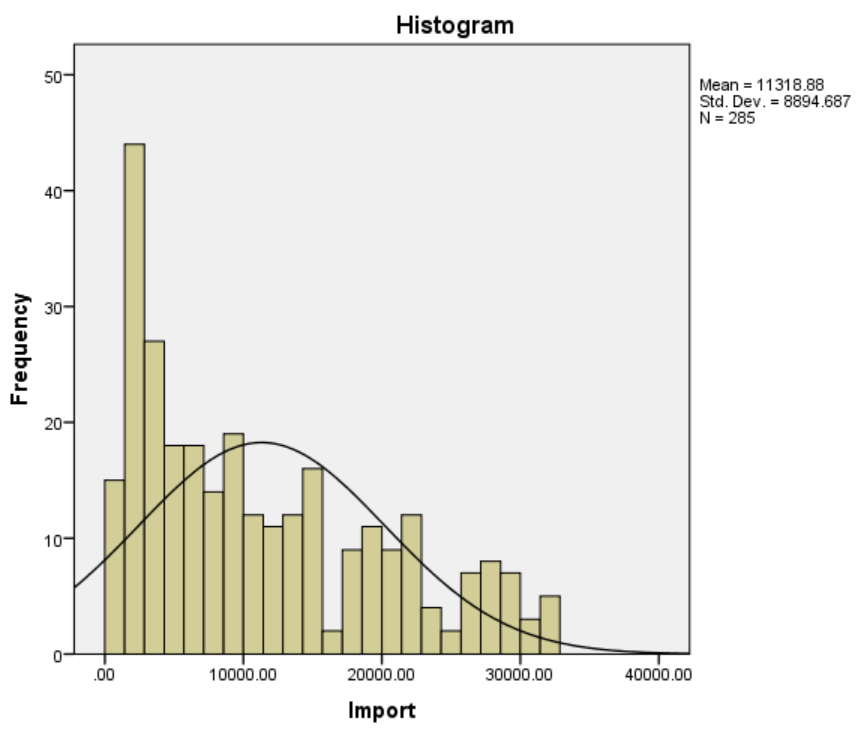


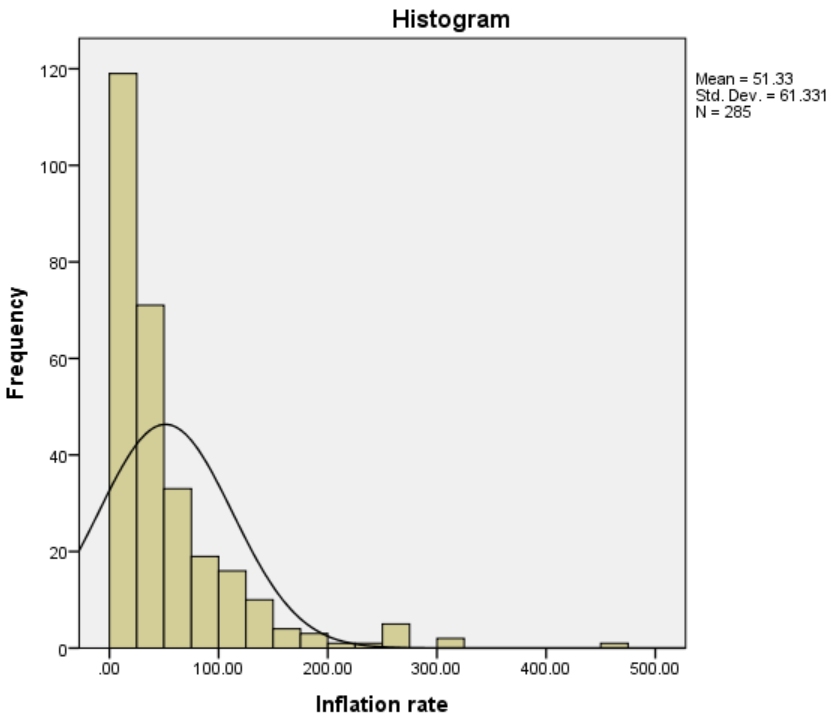
### **2. Graph**

### **2.1. Histogram**

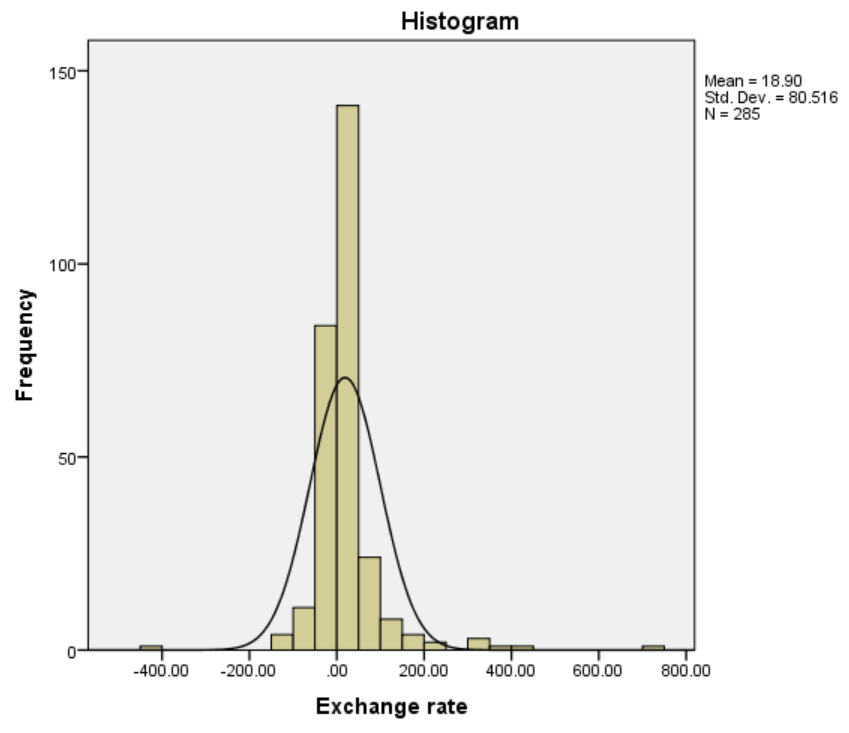


* The histogram has a mean of 11070.60, a standard deviation of 9302.734, and a sample size of 285.
* The chart shows that the majority of the data falls between 0 and 20000, with a smaller amount of data falling between 20000 and 40000. This may indicate that the majority of months of the year have export levels in the range between 0 and 20000, while there are only a few months with higher export levels, ranging from 20000 to 40000.
* The black curve on the graph is the normal curve, a type of continuous probability distribution that has a bell shape. This curve shows that the data tends to be concentrated around the mean and dispersed by the standard deviation.





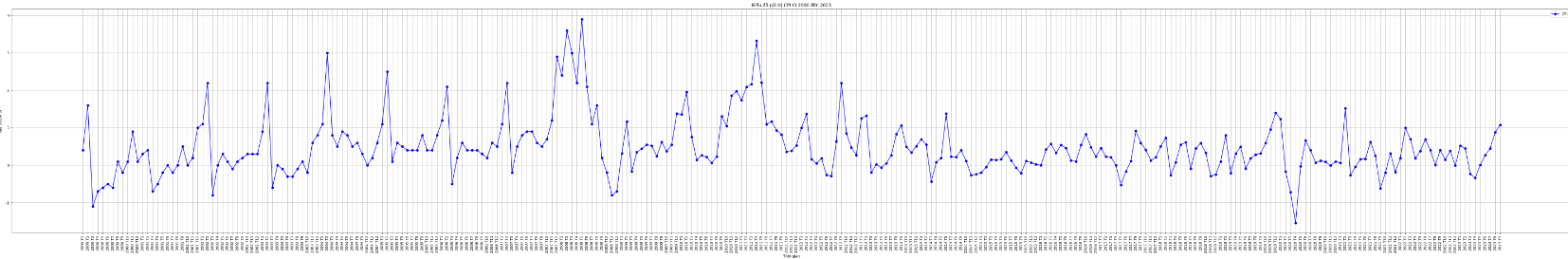
* The histogram has a mean of 51.33, a standard deviation of 63.31, and a sample size of 285.
* The chart shows that the highest frequency of the inflation rate is around 100 and decreases as the inflation rate increases.



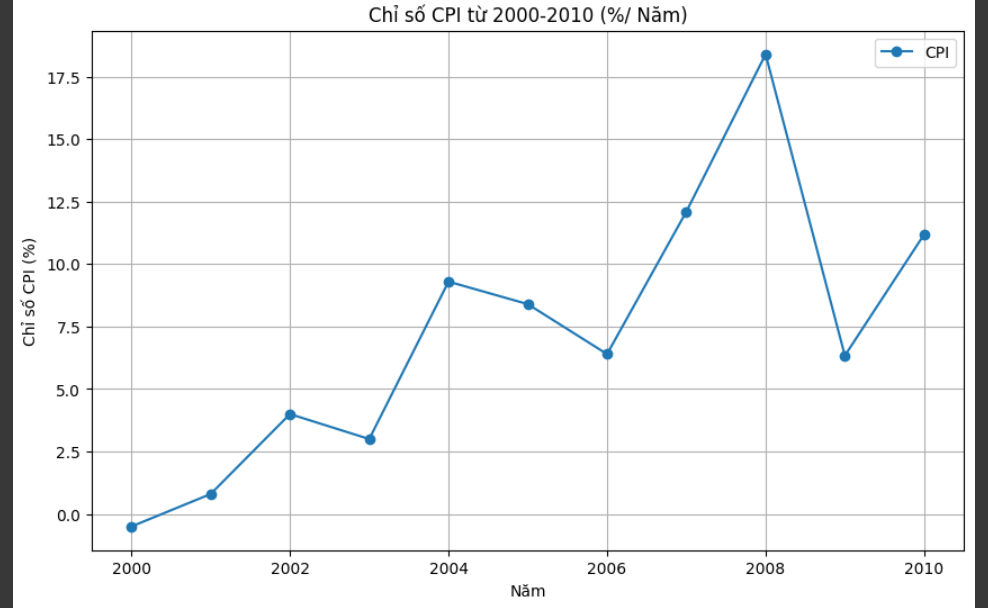
### **2.1. Line**

#### *2.1.1. CPI*

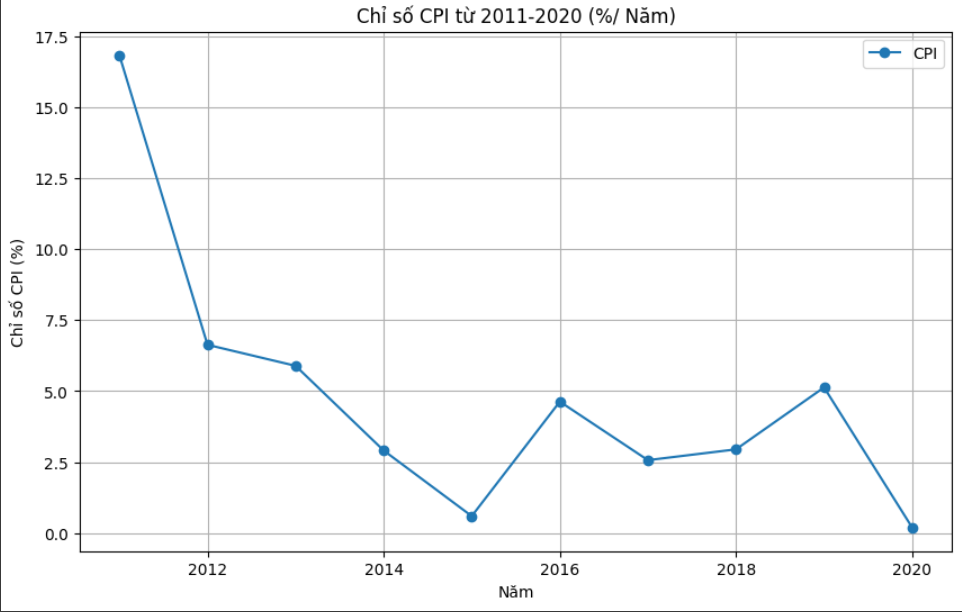
* General trend: Vietnam's CPI tends to increase in the period 2000-2023, with an average increase of 6.2%/year. However, the CPI growth rate has tended to decrease in recent years, from 16.8% in 2008 to 3.1% in 2023.
* Impacting factors: Vietnam's CPI is influenced by many factors, including:
* Internal factors: Prices of domestic goods and services, exchange rates, monetary and fiscal policies.
* External factors: Prices of imported goods and services, prices of goods and services on the world market.



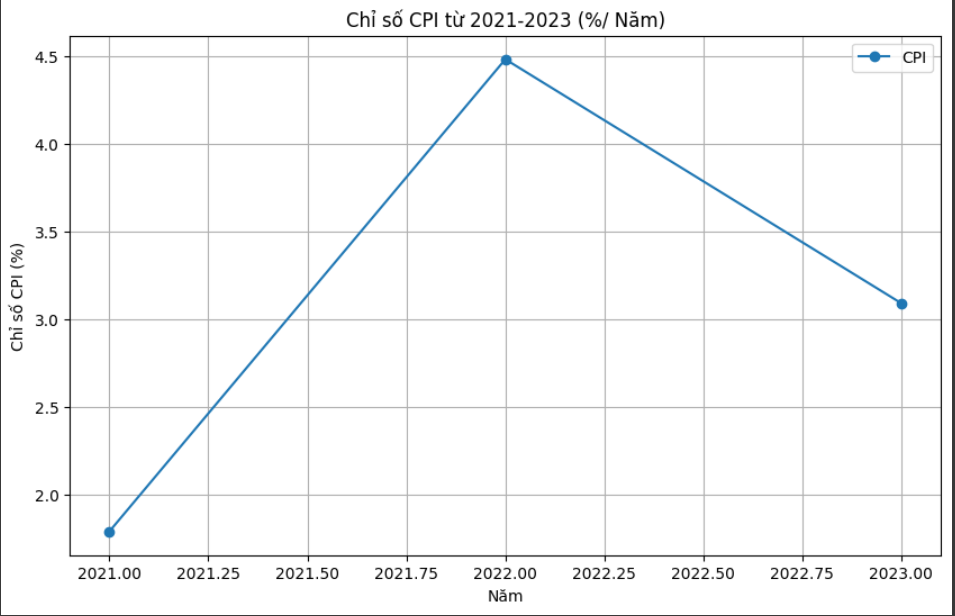
* Specifically divided into 3 stages:
  + **Phase 1 (2000-2010):** During this period, Vietnam experienced several major economic upheavals, including the 2008 global financial crisis³. This has led to fluctuations in CPI. Vietnam's CPI index during this period increased sharply, with the inflation rate in 2010 being 9.19%².



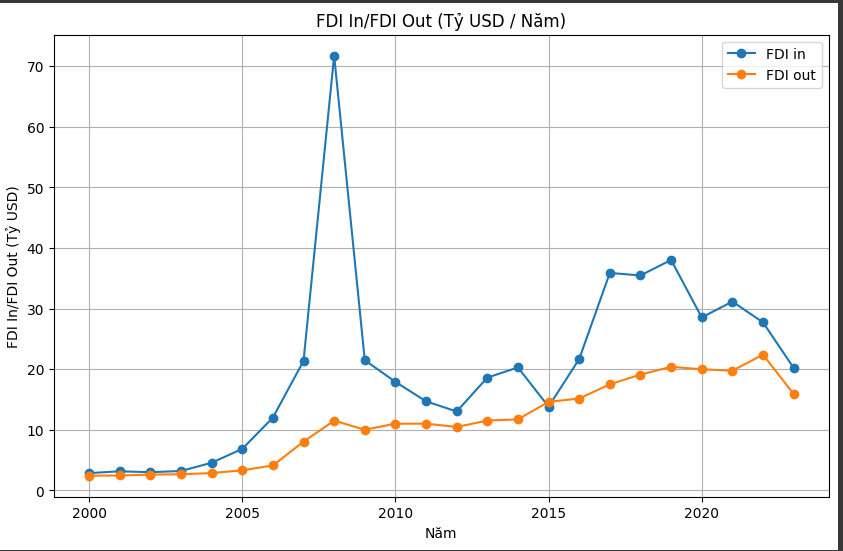
* + **Phase 2 (2011-2020):** This period saw more economic stability, with CPI rising steadily². However, 2011 had an inflation rate of 18.58%, the highest in this period. After that, the inflation rate gradually decreased and remained stable at about 4% in the period from 2016 to 2020.



* + **Phase 3 (2021-2023):** During this period, CPI continued to increase. According to data from the General Statistics Office of Vietnam, CPI in August 2023 increased by 2.02% compared to December 2022 and increased by 2.96% over the same period last year. On average in the first 8 months of 2023, CPI increased by 3.1% over the same period last year. Forecasts suggest that the average CPI in 2023 will increase between 3.2% and 3.6%.



#### *2.1.2. FDI In & FDI Out*



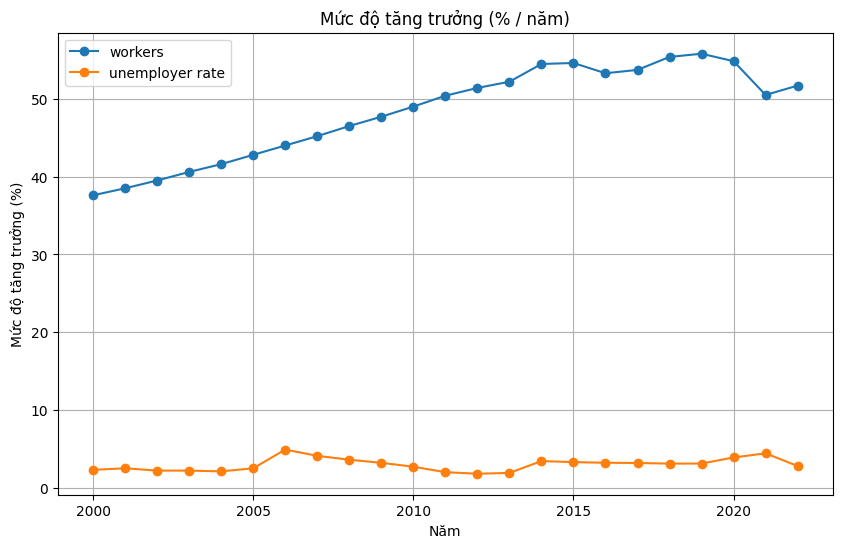
* Based on the data in the figure, it can be seen that Vietnam's FDI In has grown steadily over the period 2000-2023, with an average annual growth rate of 22.2%. Vietnam's FDI Out also grew during this period, but at a slower rate, with an average annual growth rate of 11.3%.

**THE REASONS IN 2008 FDI IN VIETNAM WAS A RECORD HIGH?**

* In 2008, registered FDI capital in Vietnam reached 71.7 billion USD, a record high in the history of attracting foreign direct investment. There are several key reasons for this growth, including:
  + Joining the WTO: Vietnam officially joined the WTO in November 2006, opening up a large market for foreign investors.
  + Stable economic growth: Vietnam's economy grew steadily during this period, with an average annual GDP growth rate of 8%.
  + Preferential policies: The Vietnamese government has issued many preferential policies to attract foreign investment, including tax, land, and labor incentives.
* In addition, a number of specific events also contributed to promoting FDI capital flows into Vietnam in 2008, including:
  + Samsung Group opens a mobile phone factory in Bac Ninh: Samsung's factory is the largest FDI project ever implemented in Vietnam, with a total investment of 5.8 billion USD.
  + Vingroup Corporation received an investment of 3 billion USD from TPG Group: This investment has helped Vingroup become one of the largest private corporations in Vietnam.
  + Global financial crisis: The global financial crisis has caused foreign investors to look for new markets to invest in, and Vietnam is one of the markets benefiting from this trend.

=> Overall, 2008 was a successful year for Vietnam in attracting foreign investment. Record FDI capital inflows have contributed to promoting Vietnam's economic growth and development.

#### *2.1.3. Workers & Unemployed rate*



* General comment:
  + The number of workers in Vietnam is increasing, from 36.5 million people in 2000 to 61.2 million people in 2020. The average annual growth rate is 2.5%.
  + The unemployment rate in Vietnam is also increasing, from 2.6% in 2000 to 3.5% in 2020. The average annual growth rate is 0.2%.
* Details:
  + The number of workers has increased steadily over the past 20 years, with the highest growth rate being 3.6% in 2001 and the lowest growth rate being 1.8% in 2014.
  + The unemployment rate increased in the period 2000-2010, then decreased in the period 2010-2015 and increased again in the period 2015-2020. The highest growth rate was 0.5% in 2000 and the lowest growth rate was 0% in 2015.

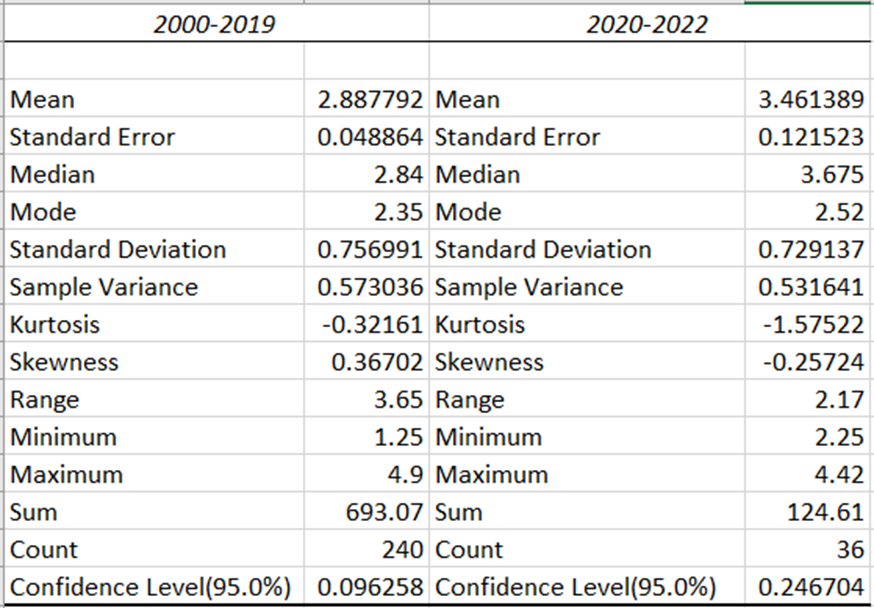
**Conclusion:** It can be seen that the growth of the labor force in Vietnam is occurring faster than the growth of employment, leading to an increase in the unemployment rate. This shows that Vietnam needs solutions to solve the unemployment problem, such as economic development, creating more jobs, improving the quality of human resources,...

## **II. Hypothesis testing**

### **2.1 Testing equality of two population means**

A newspaper said that Vietnam's unemployment rate increased sharply due to the impact of the Covid 19 pandemic. To test this hypothesis, I will compare the unemployment rate around 2000-2019 (a time when the Covid epidemic had not yet occurred). ) and from 2020 - 2022 (around the time the covid 19 epidemic is raging).

We use Excel to compute the Mean, Standard Deviation and Count of each interval. The results are: *(Hypothesis\_Result.xlsx [Sheet 1]).*



* 2020 – 2022

n1 = 36

1 = 3.46

S1 = 0.73

* 2000 - 2019

n2 = 240

2 = 2.89

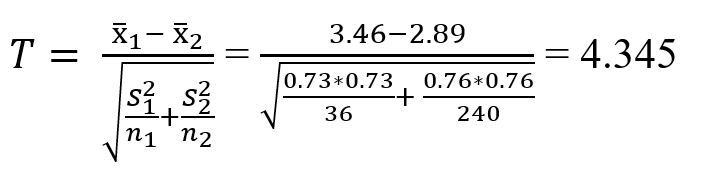
S2 = 0.76

Test at the 5% level of significance *(= 0.05).*

H0 : μ1 = μ2 :unemployment rate not increased during the Covid 19 Pandemic.

H: μ1 > μ2 : unemployment rate increased during the Covid 19 Pandemic.

Two samples are independent and large. The test statistics is followed by this formula:



α =  *0.05 → Z*α *= Z0.05 = 1.64*

H is “>”, so this is a right-tailed test. We have T > Zα (4.345 > 1.64), so the rejection region is [1.64,∞)

The decision is to reject H0 . Evidence indicates the unemployment rate increased during the Covid 19 Pandemic.

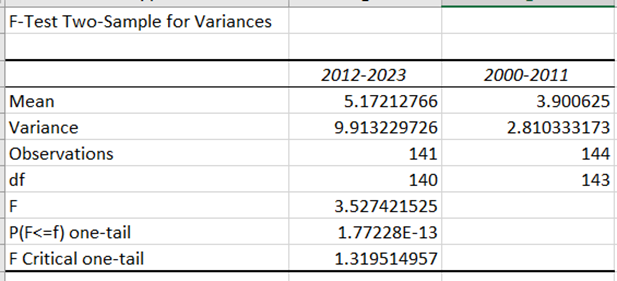
### **2.2 Comparing two population variances using F-test**

I want to compare the variance of the GDP growth rate in the two time periods 2000-2011 and 2012-2023 to compare the stability between the two periods and evaluate the impact of economic stability on the labor market.

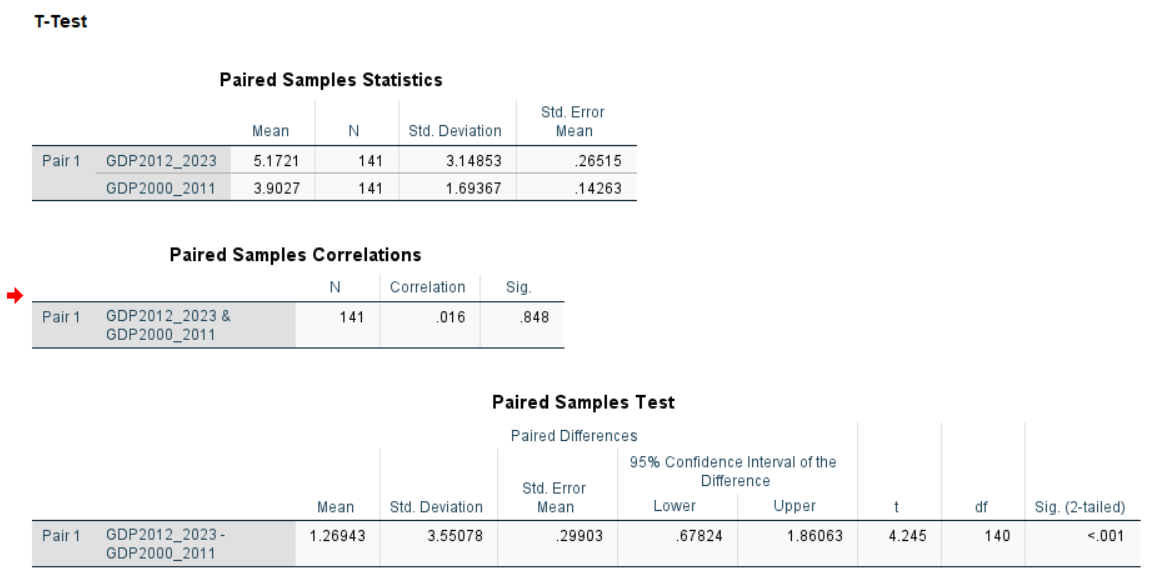
H0 : σ1 = σ2 : the stability in economics between the two periods is the same.

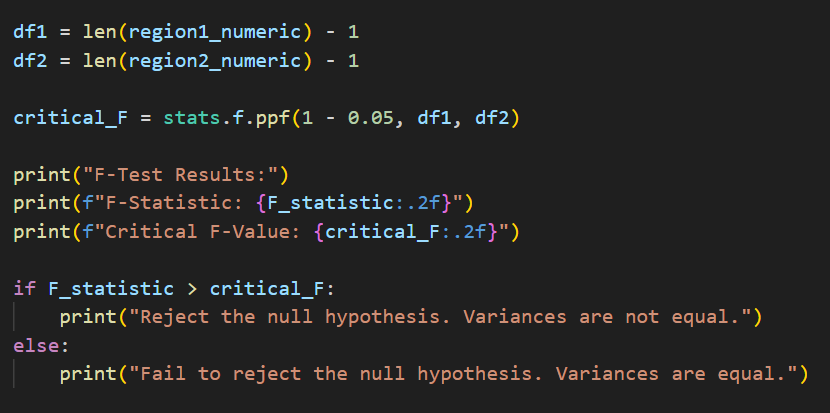
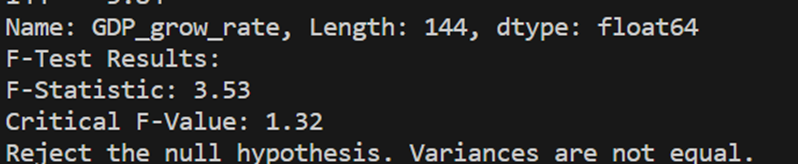
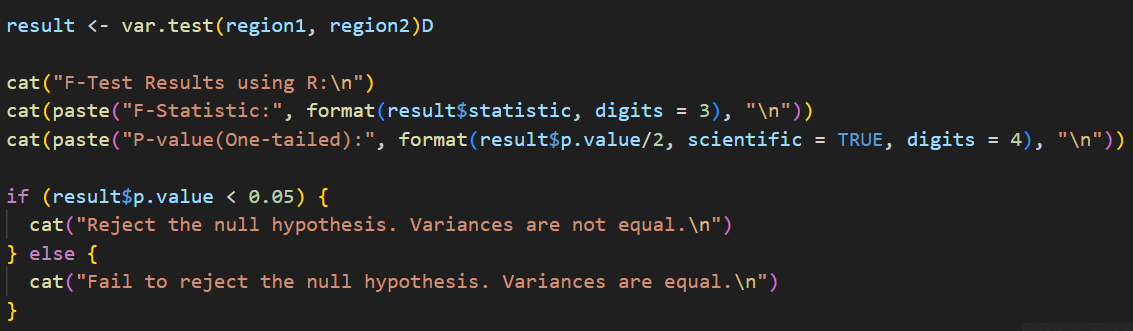
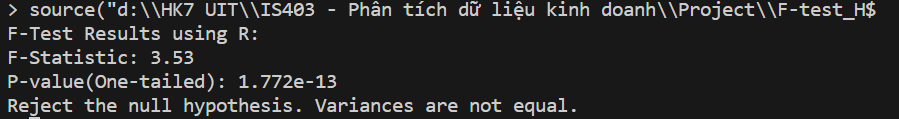
H : σ2 > σ1:  the economic at 2000-2011 is more stable than the 2012-2023.

* Using Excel: *(Hypothesis\_Result.xlsx [Sheet 2])*



* Using SPSS: *(F-test-spss-output.spss)*



* Using Python:
  + Code: *(F-test\_Hypothesis.py)*
  + Result:
* Using R:
  + Code: (*F-test\_Hypothesis.r)*
  + Result:

According to the results, F is equal to 3.527 and the (F critical) = 1.319.

So that F >Fcrit  (3.527 > 1.319) , we reject H0 and accept .

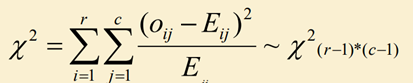
The two standard deviations are not equal, and we are 95% confident that the first period (2000-2011) is more stable than the second period (2012-2023). Based on the results, it is evident that during the period of 2012-2023, when the Covid-19 pandemic occurred, the GDP growth rate was not stable so that the economic is not stable, significantly impacting the labor market.

### **2.3 Chi-square Test**

Testing the independence between two variables, such as GDP growth rate and Consumer Price Index (CPI), in the context of the labor market can provide valuable insights into the economic dynamics and relationships between these factors.

GDP growth is often associated with job creation. However, the relationship between GDP growth and employment can be complex, and inflation can play a role in shaping this relationship. Analyzing the independence between GDP growth and CPI can help identify whether changes in employment levels are primarily driven by economic expansion or are influenced by inflationary pressures.

To test the independence between two variables: GDP Growth Rate (%) and CPI (%). We use the Chi-square random distribution based on the Pearson chi-square statistical parameter to compare number of observed results with the expected results by using the formula as follows:

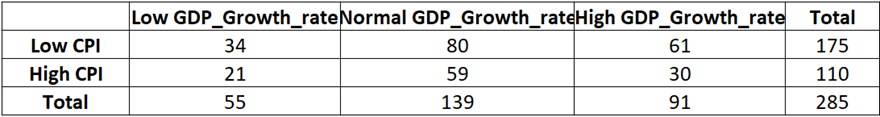


Testing for independence between two variables: GDP Growth Rate (%) and CPI (%).

H0 : Two variables are independent (if the condition of formula (1) happens).

Ha : Two variables are related to each other.

After synthesizing data from the two columns GDP\_growth\_rate and CPI in the data\_month.csv file, we obtain the following data table: *(Hypothesis\_Result.xlsx [Sheet 3]).*



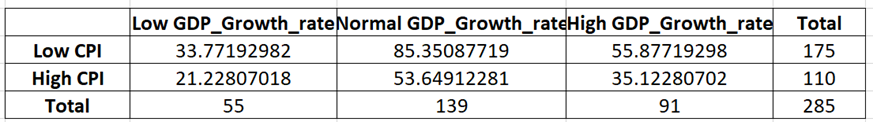
In particular, Low CPI value is determined by values ​​less than 0.05 and High CPI value is determined by values greater than or equal to 0.05. In addition, Low GDP\_growth\_rate values are values less than 3, Medium GDP\_growth\_rate are values in the range [3;5.5), and High GDP\_growth\_rate are values greater than or equal to 5.5.

After reviewing Table of the survey’s results, we have some key information as follows:

Table 1 has two dimensions that include r middle columns and c middle rows. The term Oij of this table indicates the number of GDP Growth rate levels i in a CPI level j.

* 285 (data): 175 data (61.4% has low CPI %) and 110 data (38.6% has High CPI %).
* 55 data (19.3% has low GDP Growth Rate), 139 data (48.8% has Medium GDP Growth Rate) and 91 data (31.9% has High GDP Growth Rate).

State the H0: All different levels of GDP Growth Rate have the same low and high levels of CPI. That means 61.4% of data has low CPI and 37.6% of data has CPI. If this hypothesis is correct then we will have this Table. *(Hypothesis\_Result.xlsx [Sheet 3]).*

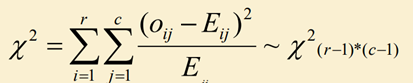


The low GPD\_growth\_rate group has 55 data. To be consistent with H0 hypothesis, 61.4% must has low CPI, and 37.6% has high CPI, so we have the below result.

* 0.614 \* 55 = 33.77 data has low CPI.
* 0.376 \* 55 = 21.23 data has high CPI.

We argue the same of two other groups and get the results in Table above.

To test this hypothesis, we use the testing criteria as follows:



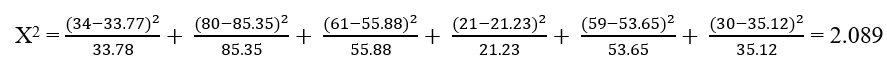
This standard follows a chi-square distribution based on degrees of freedom calculated as follows:

df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2.

With confidence level (95%), we apply the chi-squared percentile table to compute

X2\* = X2 (α, df) = X2 (0.05, 2) = 5.99.

From the data of the above example, we have the below result.



⇒ X2 <X2\* , we accept H0.

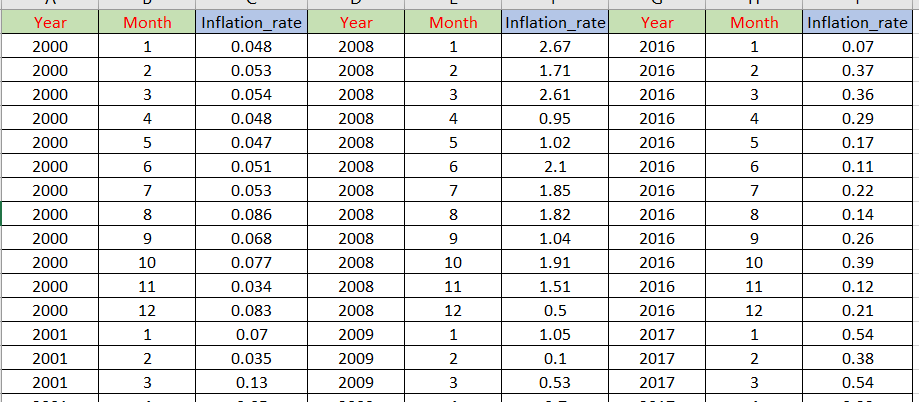
Conclusion:

The two characteristics "GDP\_Growth\_Rate" and "CPI" are independent. This means that there is no statistically significant relationship between the growth rate of GDP and the Consumer Price Index. In other words, the changes in GDP growth rate and CPI values do not appear to be associated with each other based on the analyzed data.

### **2.4 One-way ANOVA**

Using One-way ANOVA to analyze the inflation rate across different periods may provide insights into potential effects on the labor market.

First, we split the "Inflation Rate" column into three periods (2000-2007, 2008-2015, 2016-2023) and use One-way ANOVA to test the differences in inflation levels between these time periods.

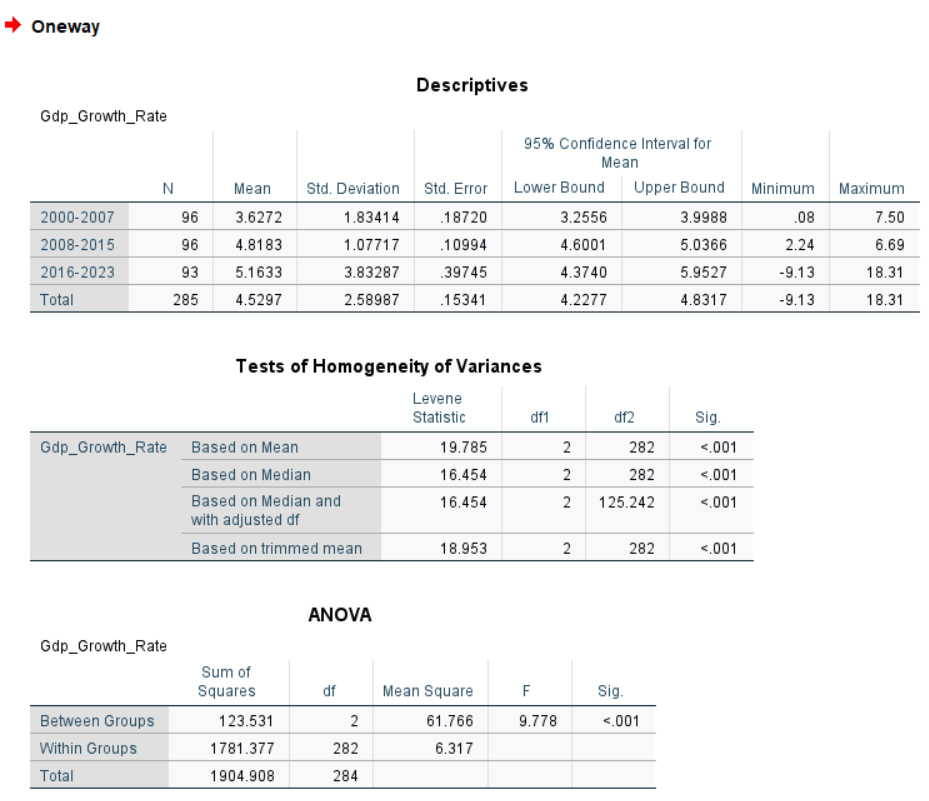


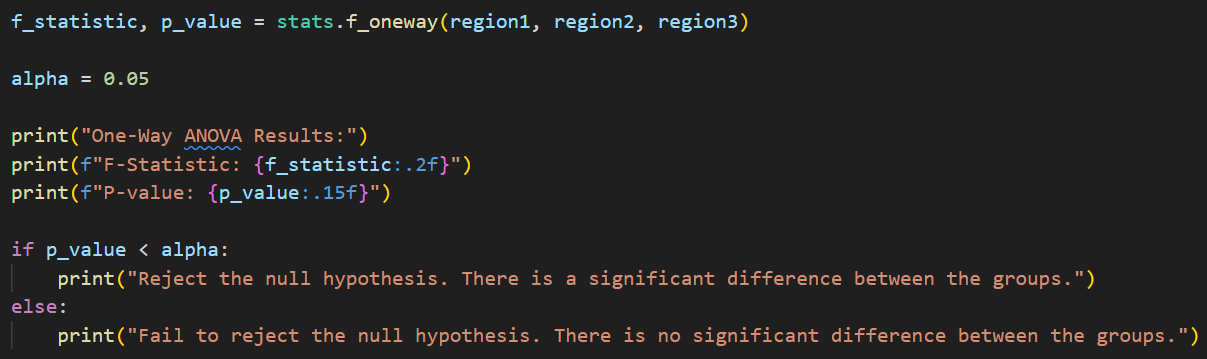
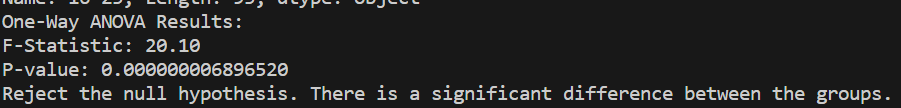
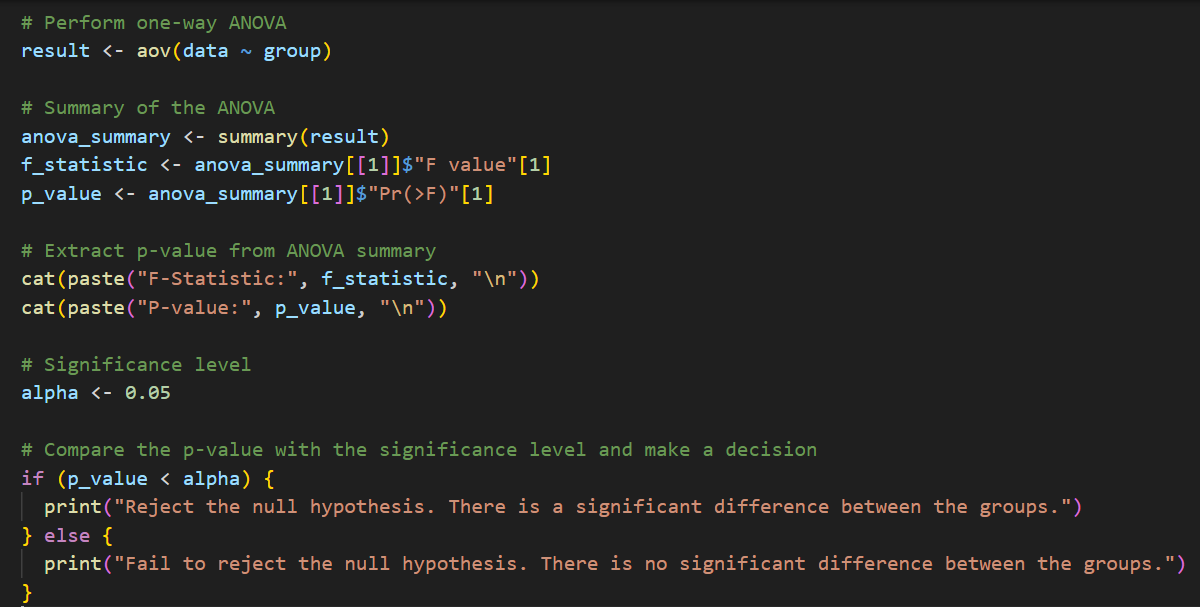
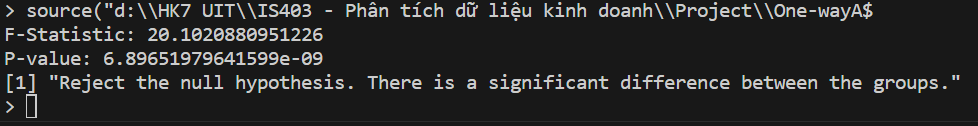
Here is the result after we use ANOVA: One-way factor analysis:

Using Excel: *(Hypothesis\_Result.xlsx [Sheet 4])*



* Using SPSS: (ANOVA\_Output.spss)



* Using Python: (One-wayANOVA.py)
  + Code:
  + Result:
* Using R: (One-wayANOVA.r)
  + Code:
  + Result:

Fstat = 20.102

fcrit  = 3.028

→Fstat > fcrit : reject H0 and accept Ha . We can conclude that there is a difference in the sample mean among three periods of inflation rate.

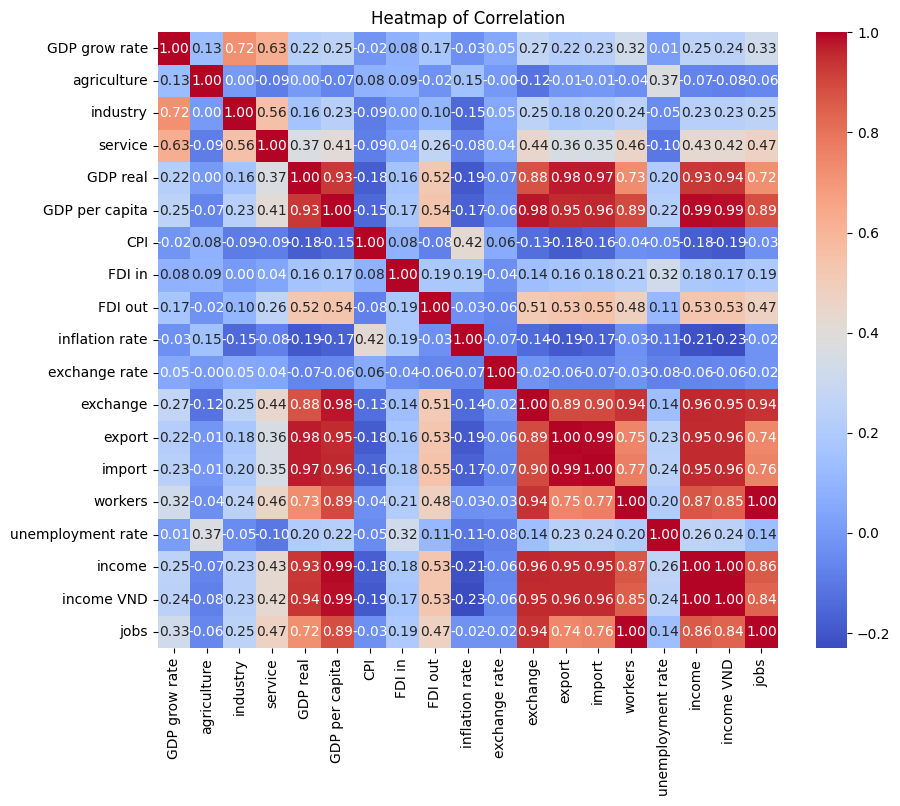
In summary, while One-way ANOVA can identify differences in inflation rates across periods, understanding the implications for the labor market requires a more comprehensive analysis that considers correlation, causation, time-series dynamics, and potential external factors.

## **III. Execution Analysis**

### **3.1. Linear regression**

#### ***3.1.1. Implementation steps***

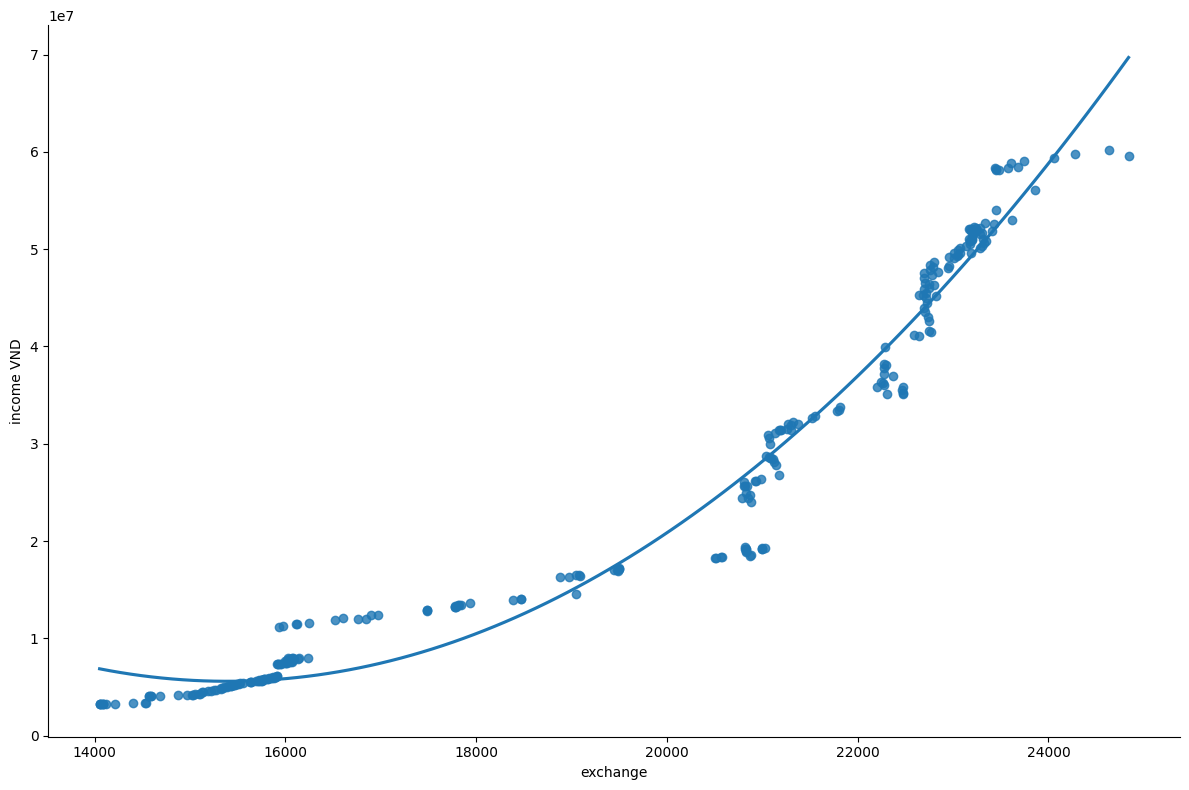
* Heatmap:



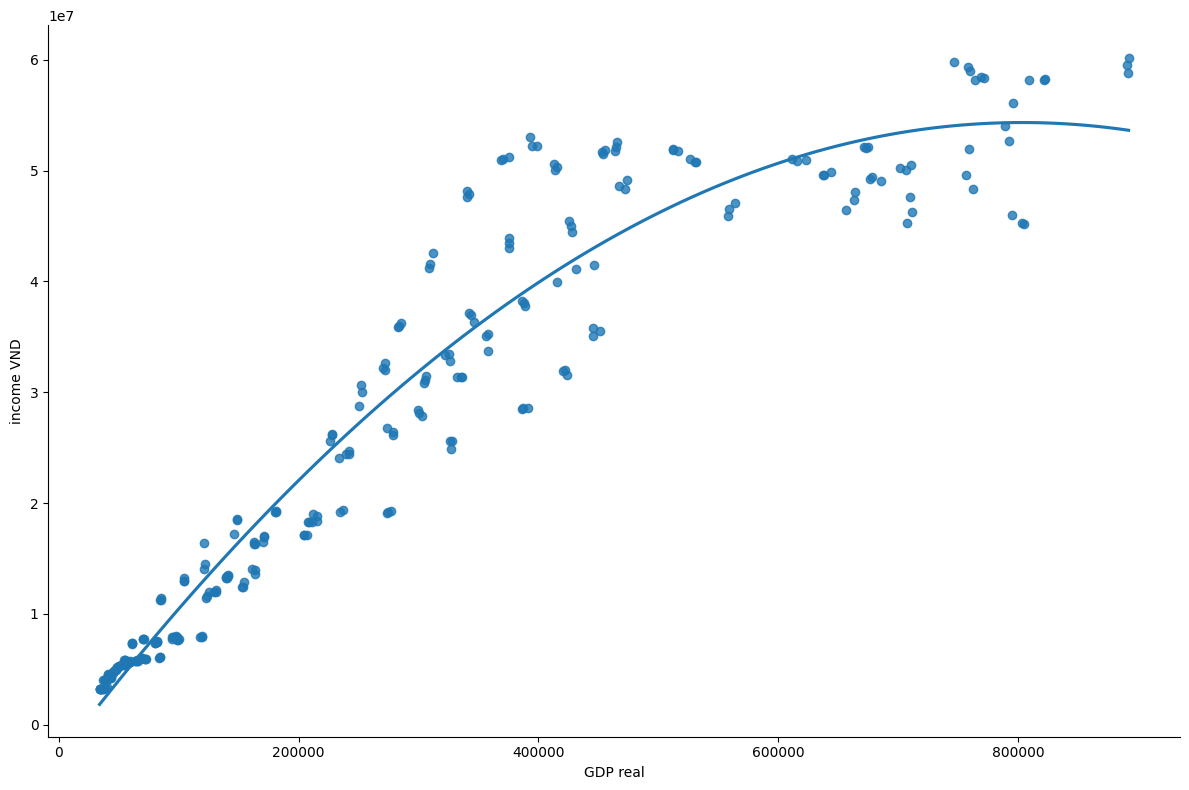
Based on the heatmap of the correlation chart, we can see that the correlation between per capita income and value columns such as: real value of GDP, GDP per capita, exchange, total exports and total imports is quite high.

Therefore, we will draw a scatter plot to see how correlated it is

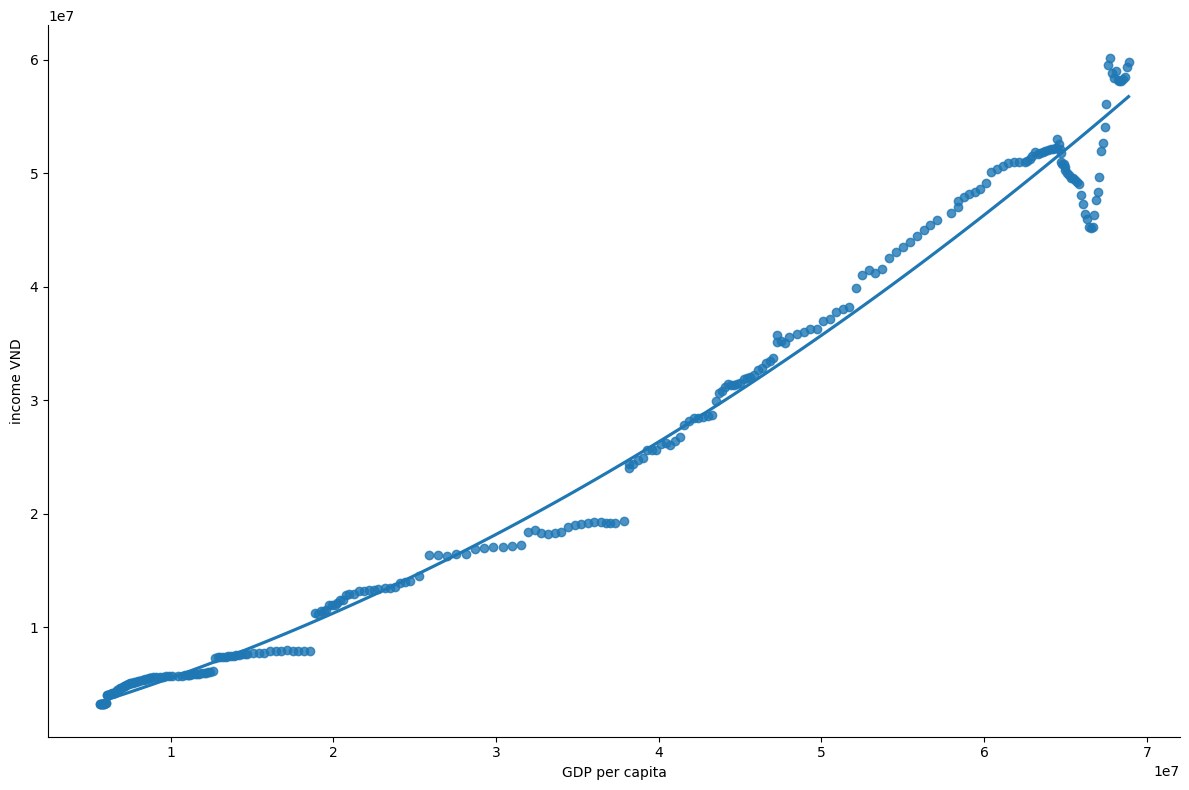
* Annual household Income vs Exchange:



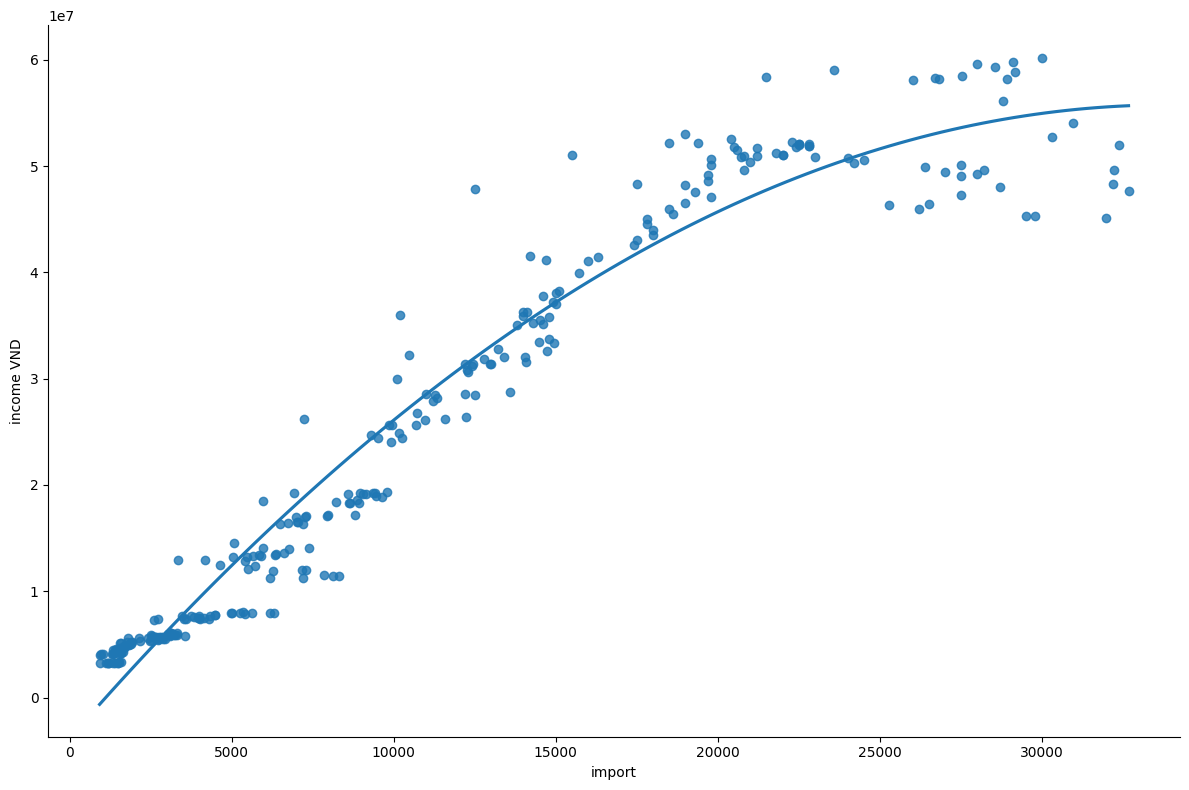
* Annual household income and total GDP value:



* Annual household income and GDP per capita:

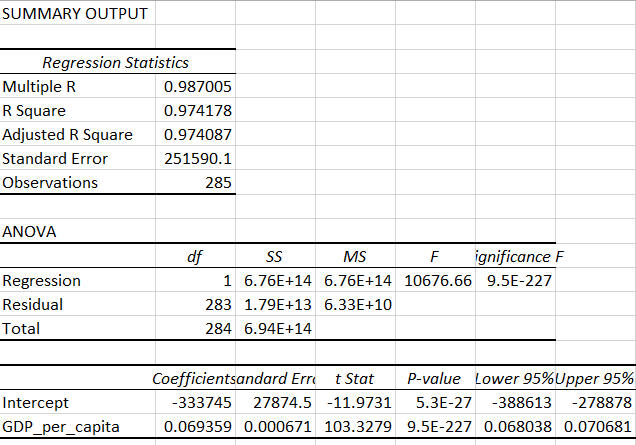


* Annual household income and total import/total export:



We can see that the scatterplots tend to be linear and most of them are nonlinear.

* In the case of annual household income and GDP per capita:



ANOVA conducts an F-test to determine whether variation in Y is due to varying levels of X.

ANOVA is used to test for significance of regression:

* H0 : GDP per capita is not a significant variable
* H1 : GDP per capita is a significant variable

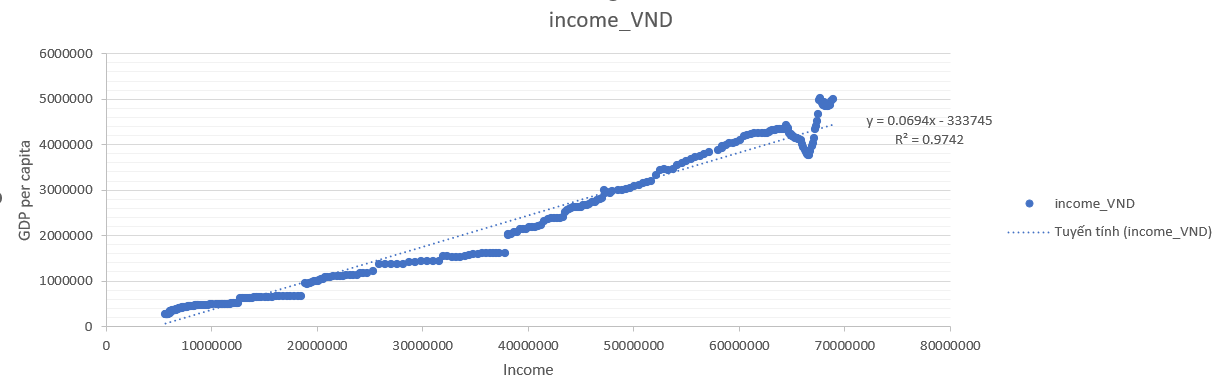
p-value = 9.5 x 10-227 < 0.05

=> Reject H0 .   
- p-value for test on the intercept = 5.3 x 10-27

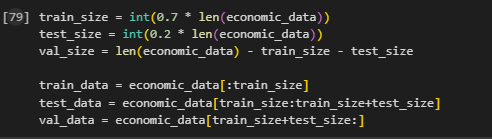
- p-value for test on the slope = 9.5 x 10-227

=> Both tests reject their null hypotheses and both the intercept and slope coefficients are significantly different from zero.

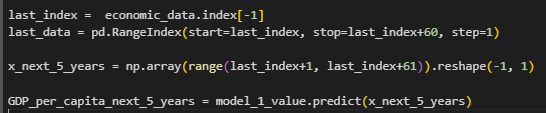
The linear regression is: Y = 0.06935X - 333745

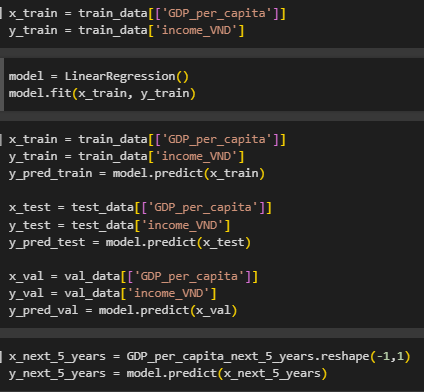


* Step 1: Split our dataset into three subsets: training, validation, and test datasets. You can allocate 70% of your data for training, 20% for validation, and 10% for testing. This is commonly known as a 7-2-1 split.



* Step 2: Choose 'income' as a dependent variable and 'GDP per capita' as your independent variable. Use a linear regression model to predict the future values of 'GDP per capita' for the next 5 years based on historical data. Train the regression model on this historical data to learn the relationship between GDP per capita and time. Once the model is trained. We have GDP per capita for each of the next 5 years.

And now we have GDP per capita for the next 5 years. Use the predicted GDP per capita values for 5 years into the future to estimate the 'income' after 5 years.



#### ***3.1.2. Result***

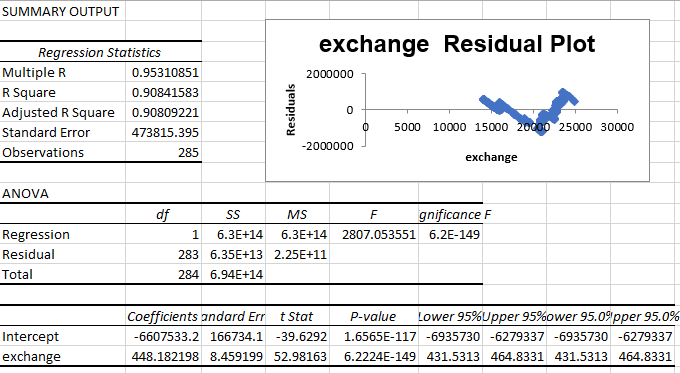
### 

### 

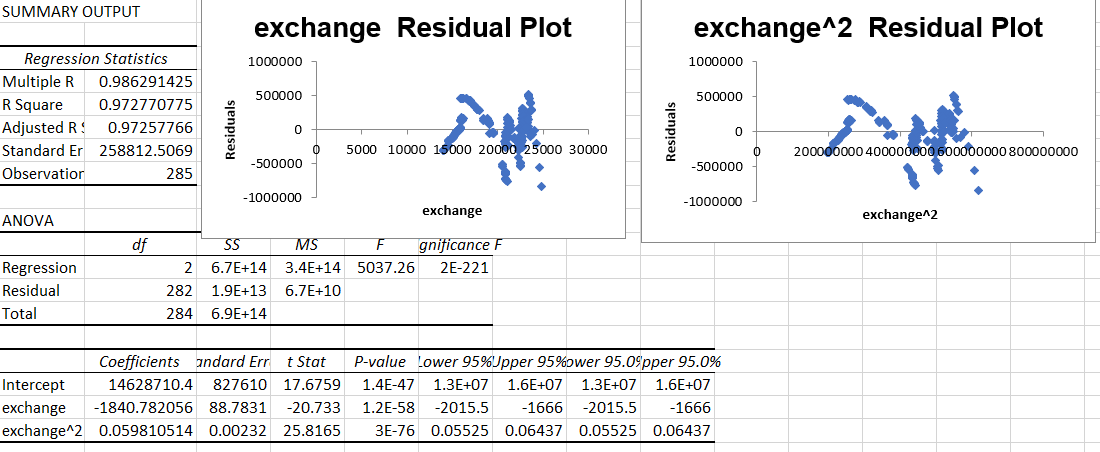
### 

* Based on forecasts using the linear regression model, we find that GDP per capita has a significant impact on the average monthly income of a household. Prediction results show that in the first year, average income will decrease. However, following the trend, in the next 4 years, the average household income is expected to increase, which corresponds to the growth of GDP per capita.
* This linear regression model not only shows the link between GDP per capita and household income but also shows the stability and recovery of the economy in the period after experiencing an initial decline in value. This may be a result of economic stimulus policies and various factors that positively affect household income.
* In summary, based on the linear regression model, we can expect that after the initial period of depreciation, average monthly household income will recover and grow, which reflects the trend Corresponding to the development of GDP per capita in the next period.

In the case income vs exchange:

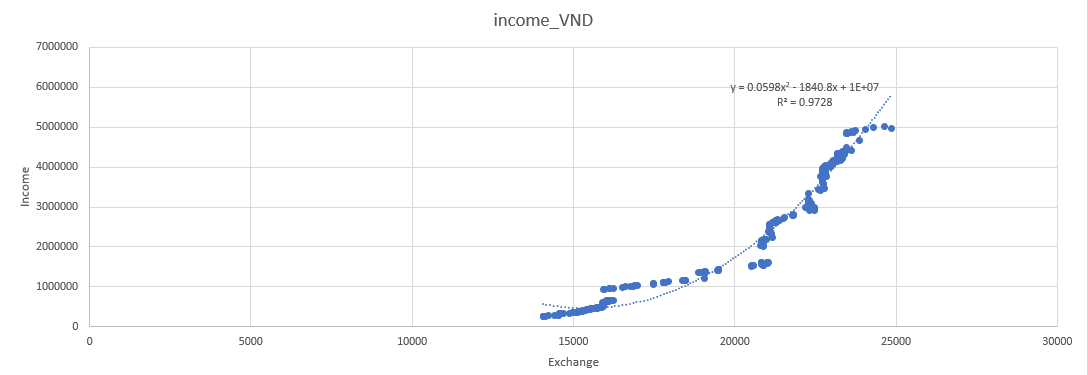


You can see the U-shape residual plot. And now we square the exchange and summary.



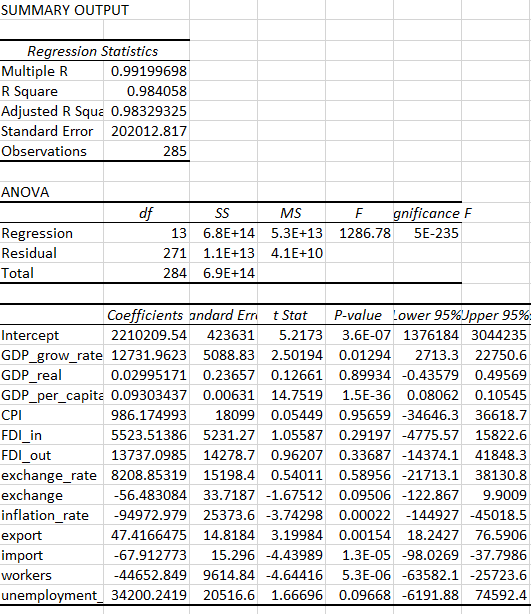
The result show us the formula of nonlinear regression:

y = 0.0598x2 - 1840.8x + 1E+07

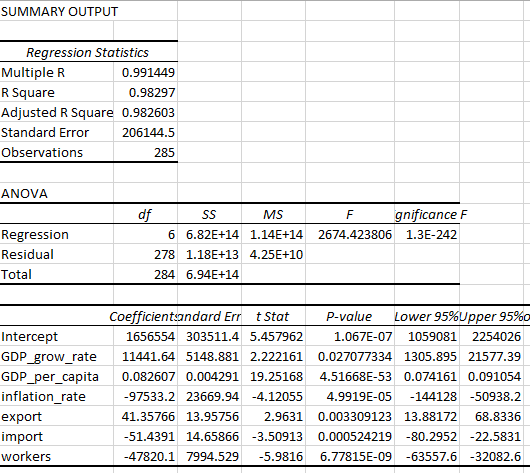


### **3.2. Multiple Linear Regression**

#### ***3.2.1. Implementation steps***

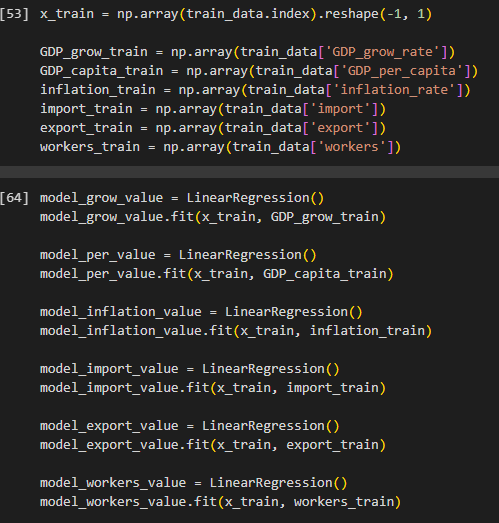


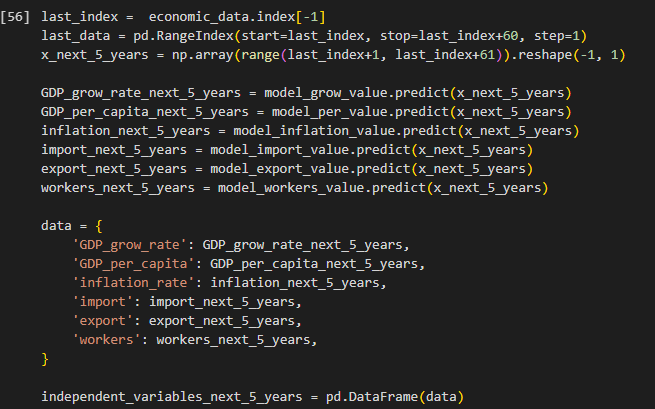
There are p values of slope coefficient p < 0.05 such as the data columns GDP per capita, GDP growth rate, inflation rate, total export, total import and workers. Therefore we see that these are statistically significant values in the model.



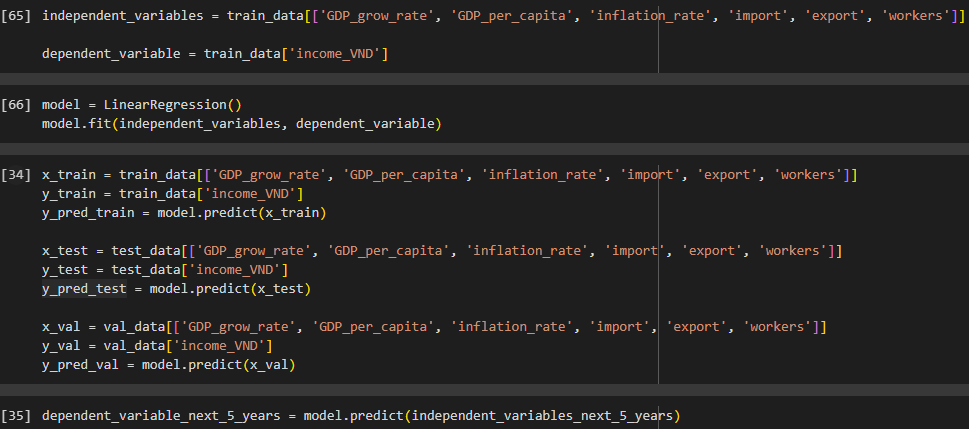
And now all the variables are significant.

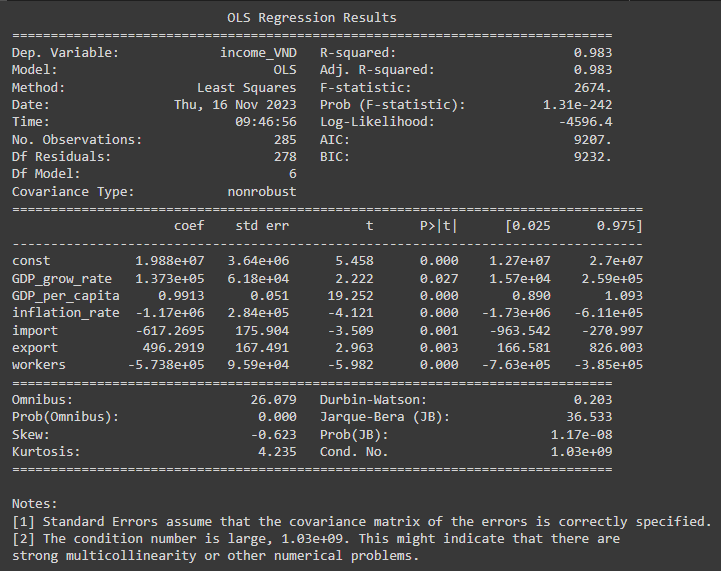
Similar to single linear regression. Then we will predict the independent variables in 5 years by training the regression model with old data.





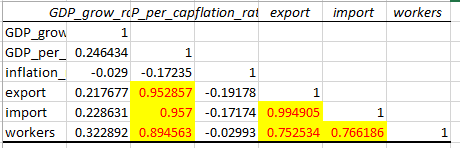
After we have the data of the independent variables for the next 5 years. We will start training the multiple regression prediction model with the corresponding independent variables.



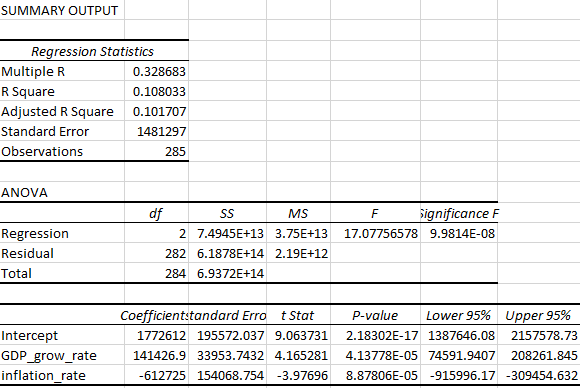


To use multivariable linear regression with the dependent variable 'income', we first select independent variables with a correlation level ranging from -0.7 to +0.7 to avoid multicollinearity.

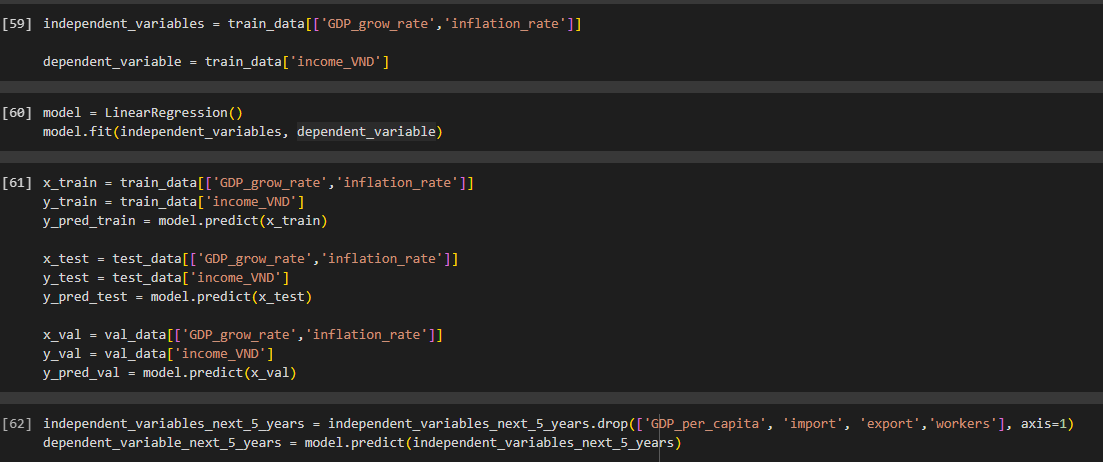
* Multicollinearity:
  + It occurs when there are strong correlations among the independent variables.
  + Makes it difficult to isolate the effects of independent variables.
  + Signs of slope coefficients may be opposite of the actual value and p-values can be inflated.



The results show that only two variables, 'inflation rate' and GDP growth rate', have a correlation level within the allowable range so that multicollinearity does not occur.



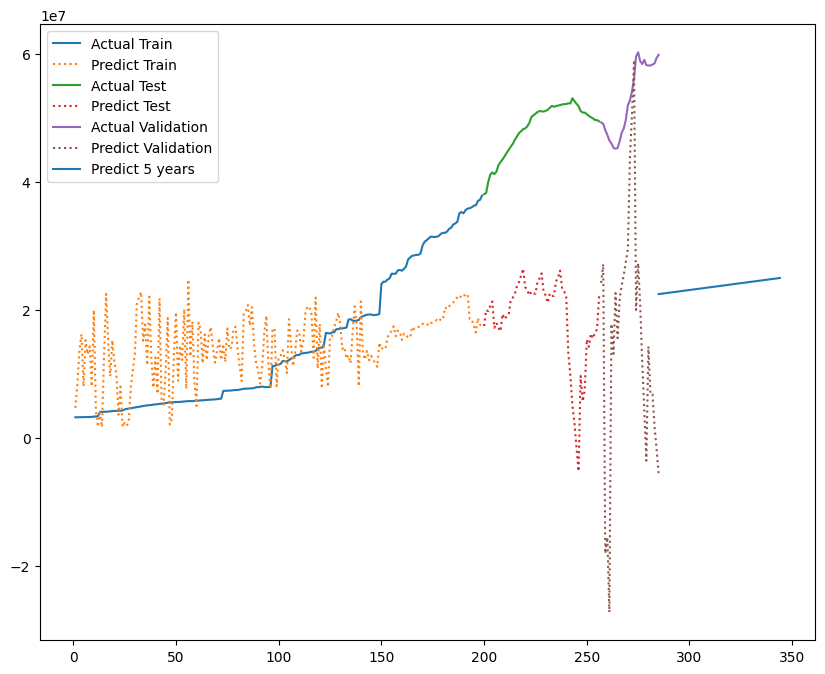
So now we will do the same as above but we will delete the highly correlated columns in the upcoming 5-year independent variables data and apply it to the multiple regression model to predict.



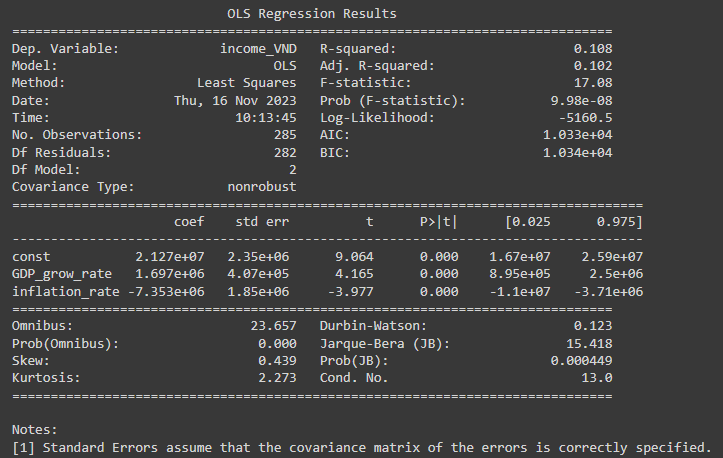
#### ***3.2.2. Result***

### 

* The prediction results from the multivariate linear regression model show a pretty good match between the training, testing, and validation data sets with real data. However, when we turn to predictions for the next five years, it appears that there is a significant error in predicting future values of the independent variable. This results in the dependent variable, i.e. the average predicted income according to the data set, not being estimated accurately.
* Forecast data for the next five years shows a notable trend. During the initial period of 1-2 years, there are clear signs of a sharp decline in average income, however, after that, the trend turns to gradual growth over the next 4 years. Large deviations in predictions can be explained by considering the unknown and unpredictable factors underlying the independent variable.
* To improve prediction quality, it may be necessary to reconsider the model or add additional independent variables to more accurately capture the factors that influence average earnings in the future.

****

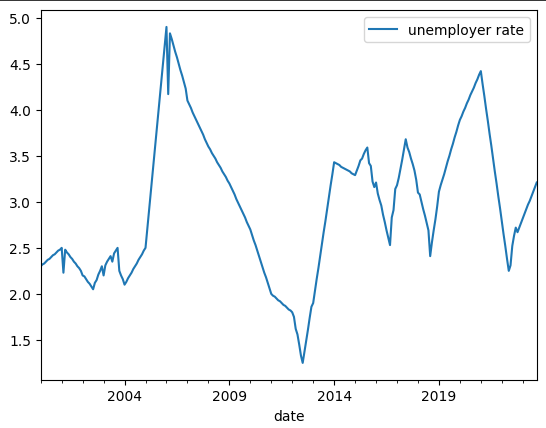
* The prediction results from the multivariate linear regression model after eliminating some variables, retaining only 2 independent variables: GDP growth rate and inflation rate, show that the assessment seems to be inaccurate for training, testing, and validation datasets. It can be assumed that the years 2019-2020, when the GDP growth rate decreased suddenly due to the negative impact of the COVID-19 pandemic, and then recovered strongly, created a large margin in the data , causing significant errors in prediction.
* Projections suggest that average household income will experience a significant decline followed by a gradual increase over time, but this increase appears to be slower than it actually is. This may be because the model is not flexible enough to be stable for independent variables with few and large amplitudes.
* To improve model accuracy, it may be necessary to reconsider adding additional variables or adjust the model to handle the unique characteristics of the 2019-2020 data. At the same time, the use of different modeling approaches can be considered to determine which model best fits this particular data.

****

### **3.3. ARIMA**

#### ***3.3.1. Implementation steps***

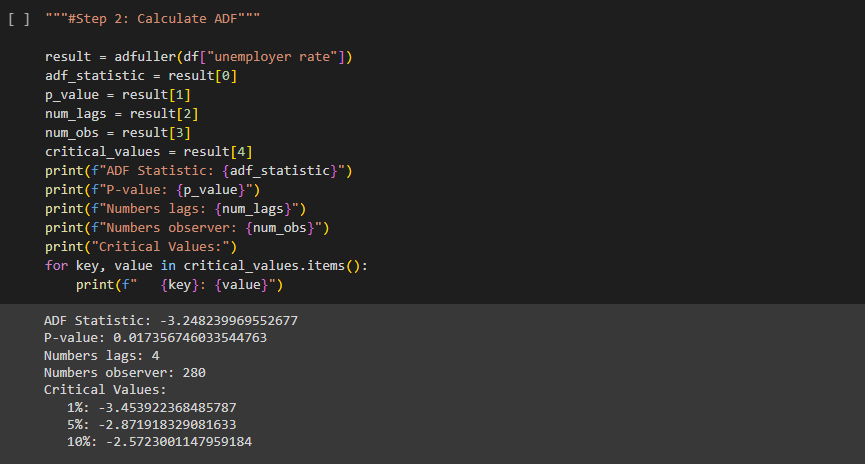
* **The reason why we choose the ARIMA model: Current status of the unemployment rate in Vietnam from 2000-2023:** First, we must analyze the unemployment rate index through the unemployment rate chart over time from 2000-2023.



* Looking at the chart, we see that the unemployment rate chart has no increasing or decreasing trend, in other words, with this chart we can see that this is a stationary series. So this is a reasonable condition to use the ARIMA model to predict the unemployment rate next year.

In particular, we can see years with a sudden decrease or increase such as 2008 and 2013. These years were strongly influenced by the world economic situation while the remaining periods were evenly distributed. and **we can see that time is related to the unemployment rate.**

* **Steps:**



The Augmented Dickey-Fuller (ADF) test is applied to the "unemployment rate" time series data to check for stationarity.

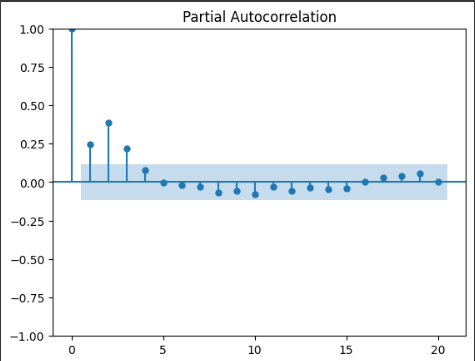
The result of the ADF test is stored in the result variable.

Relevant information such as the ADF statistic, p-value, number of lags, number of observations, and critical values is extracted from the test result.

The extracted information is then displayed using print statements.

The ADF statistic is a key metric in determining stationarity. A smaller p-value suggests stronger evidence against the null hypothesis of non-stationarity.

This code snippet is a crucial step in deciding whether the time series data (in this case, the unemployment rate) is stationary or not. The obtained p-value will guide you in determining the appropriate differencing order d for your ARIMA model. If the p-value is below a significance level (commonly 0.05), you may conclude that the time series is stationary, and you can **choose d = 0 for your ARIMA model.**



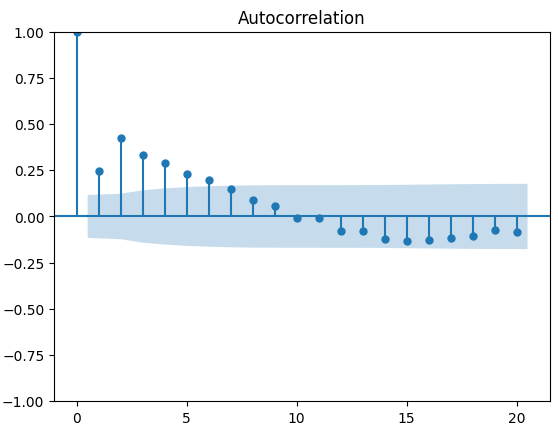
The provided code segment generates a Partial Autocorrelation Function (PACF) plot for the differenced "unemployment rate" time series. The plot\_pacf function from the statsmodels library is utilized, with np.diff applied to obtain the first-order difference of the "unemployment rate" series. Differencing helps eliminate trends and better focuses on the oscillatory components of the time series.

The lags=15 parameter specifies the number of lags (time delays) displayed on the PACF plot. In this case, 15 lags are visualized to observe the relative correlations between the current value and preceding values.

* **Selecting p:**

Interpreting the PACF plot is essential for determining the appropriate value of 'p' in the ARIMA model. Peaks in the PACF plot indicate significant autocorrelations. If you observe a single prominent peak followed by a rapid decrease, selecting p = 1 or p = 2 might be appropriate. If there are multiple abrupt peaks, you may consider experimenting with p = 3.

The choice of 'p' reflects the number of lags in the PACF that significantly influence the current value of the time series. This analysis aids in selecting the optimal autoregressive order for the ARIMA model.

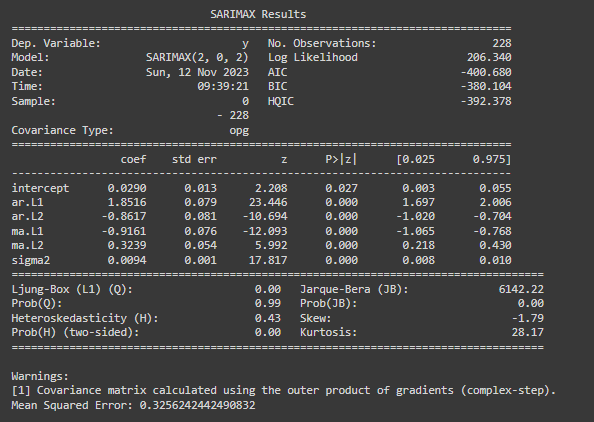


**We will choose q can be 1, 2, 3, 4, 5, 6, 7.**

And now we will choose ARIMA model from

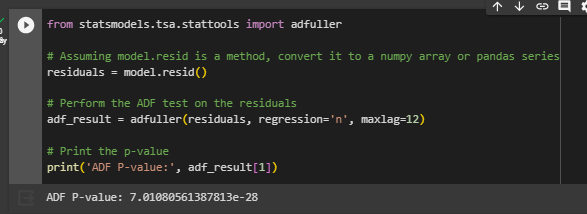
(1, 0, 1),(1, 0, 2),(1, 0, 3),(1, 0, 4),(1, 0, 5),(1, 0, 6),(1, 0, 7),(2, 0, 1),(2, 0, 2),(2, 0, 3),(2, 0, 4),(2, 0, 5),(2, 0, 6),(2, 0, 7),(3, 0, 1),(3, 0, 2),(3, 0, 3),(3, 0, 4),(3, 0, 5),(3, 0, 6),(3, 0, 7)

After evaluating the models, we choose **the model (2,0,2)** with the best index.

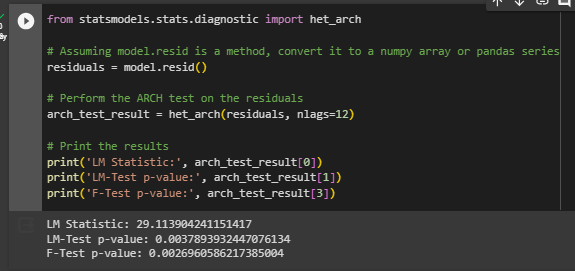


#### ***3.3.2. Result***

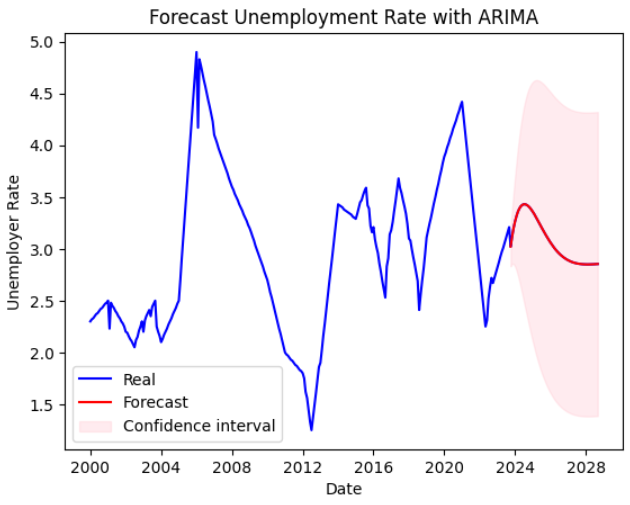
* Check stationarity for residuals:

****

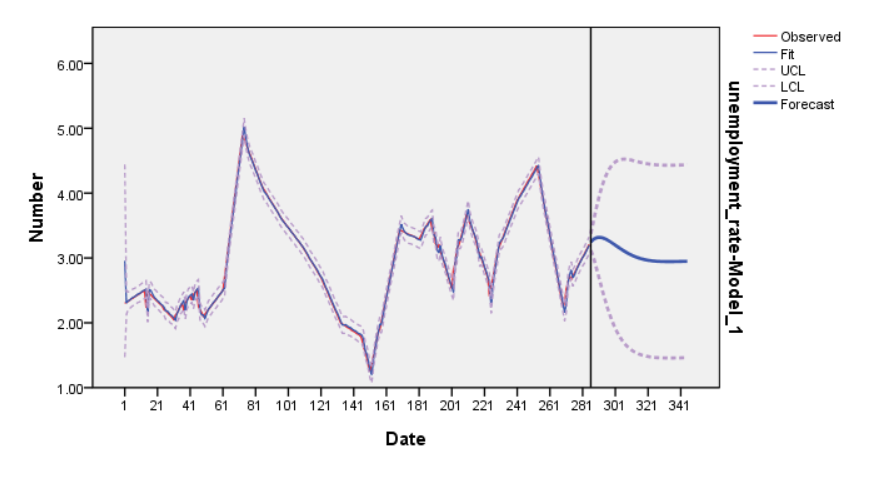
* The ARCH test is employed to assess whether there is evidence of conditional heteroskedasticity, which implies that the variance of the residuals is not constant across observations.
* The results of the ARCH test are printed to the console, including the LM (Lagrange Multiplier) Statistic, the p-value associated with the LM-Test, and the p-value associated with the F-Test.

****

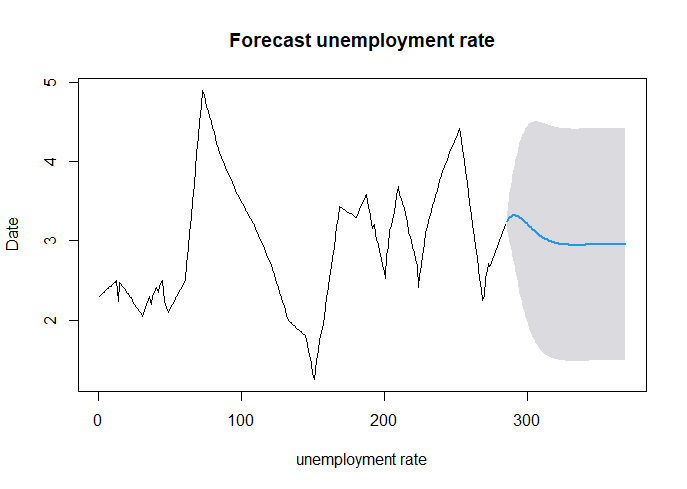
* Using Python:

****

* Using SPSS:

****

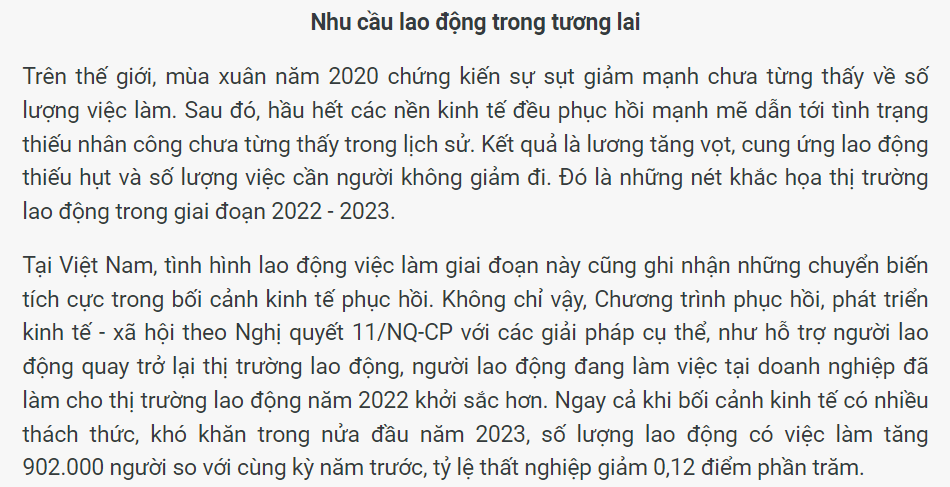
* Using R:

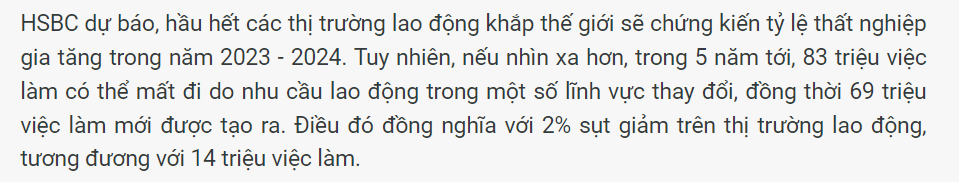
******

* With this result, we can predict that in 2025 and 2026, the unemployment rate will continue to increase. Starting from 2027 onwards, there will be signs of gradual decline, showing economic recovery after the world economic crisis in general and Vietnam's economy in particular in 2023.
* According to 61% of chief economists surveyed in the report, the global economy will weaken in 2024, with uncertainty being the main cause of this situation as published in May 2023. each topic comes. Fears of a global economic slowdown are fading, but concerns are growing as China's economy slips into deflation.
* We can see that the economic crisis of 2022 lasts until 2023 and continues to greatly affect the unemployment rate. From there we can see that economic crises seem to be cyclical and are currently entering a decline cycle.
* In the electronic newspaper THE PEOPLE'S REPRESENTATIVE, we can see that the labor market by the end of 2023 will face many difficulties.

******

* In the BIDDING newspaper, we can see that the forecast from ARIMA seems to be in line with the future labor market trend. In 2020, there will be a sharp increase in the unemployment rate and will gradually recover in 2022-2023. But according to experts' predictions, the unemployment rate will increase sharply before stably recovering. We can conclude that the ARIMA model produces quite reliable results.

******

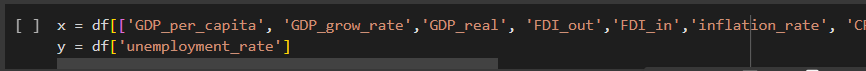
******

### **3.4. XGBoost**

#### ***3.4.1. Implementation steps***

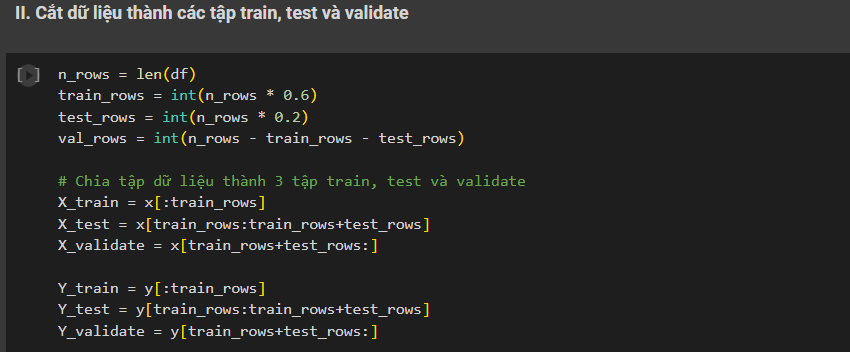
Prior to deploying the XGBoost algorithm, data preparation plays a crucial role in ensuring the effectiveness and accuracy of the model. The following is a detailed description of this process:

* **Feature Selection:** this process begins with selecting important features from the dataset. These features are chosen based on assumptions regarding their influence on the target feature - the unemployment rate. The selected features include 'GDP\_per\_capita', 'GDP\_grow\_rate', 'GDP\_real', 'FDI\_out', 'FDI\_in', 'inflation\_rate', 'CPI', 'import', 'export', 'exchange\_rate', 'exchange', 'workers', and 'income\_VND'.
* **Creating the Feature Matrix:** with the chosen features, a new feature matrix (denoted as X) is formed. This matrix represents all observations in the dataset and includes values of the selected features.
* **Selecting the Target Feature:** the target feature chosen is 'unemployment\_rate,' representing the variable to be predicted. This vector (denoted as y) contains the values of 'unemployment\_rate' corresponding to each observation.
* **Integrating Features and Target Feature:**
  + The feature matrix X and the target vector y are then integrated to create a complete dataset. This serves as the database for the XGBoost model training process.
  + Through this step, we have successfully processed the data, selected important features, and prepared the dataset for applying the XGBoost algorithm, with the hope that the model will exhibit good predictive capabilities on the unemployment rate based on important economic and societal factors.

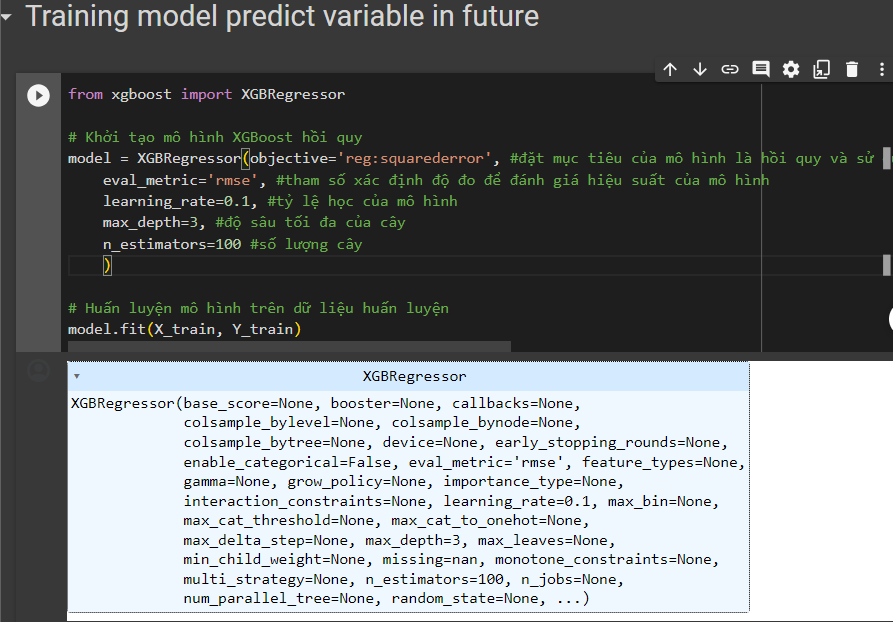


Next, dividing the Dataset into Training, Testing, and Validation Sets:

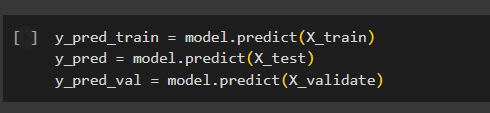
* **Determine Dataset Sizes:**
  + Calculate the total number of rows in the dataset (n\_rows).
  + Allocate proportions for the training, testing, and validation sets: 60% for training (train\_rows), 20% for testing (test\_rows), and the remaining 20% for validation (val\_rows).
* **Split the Dataset:** Utilize the calculated row indices to divide the feature matrix X and target vector y into corresponding sets:
  + X\_train and Y\_train represent the training set.
  + X\_test and Y\_test represent the testing set.
  + X\_validate and Y\_validate represent the validation set.
* **Intuitive Data Splitting:**
  + The dataset is split based on the calculated proportions, ensuring a representative distribution of data across the training, testing, and validation sets.
  + This step establishes a systematic division of the data, enabling the subsequent training, testing, and fine-tuning of the XGBoost model on distinct subsets for robust performance evaluation.



Next step,

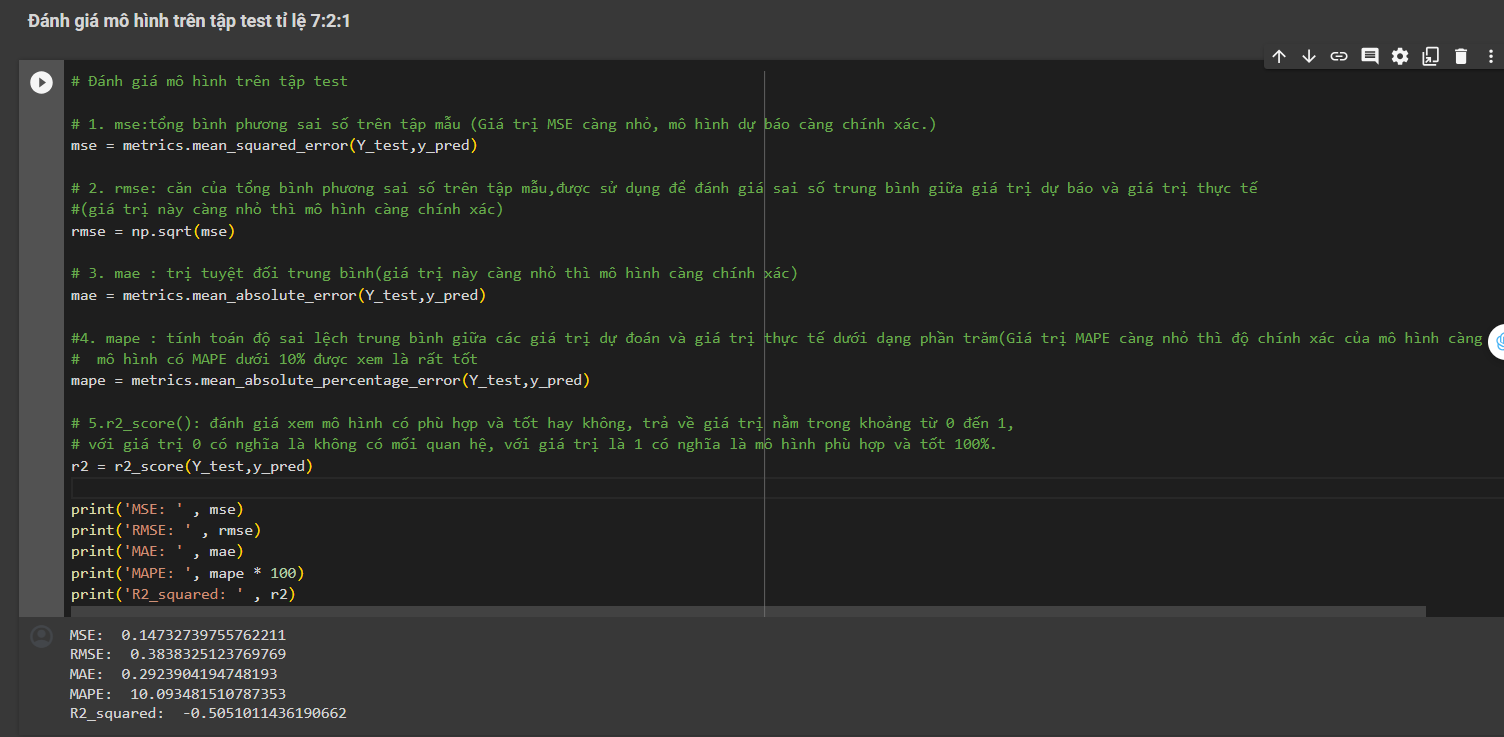
* **Import XGBoost and Initialize the Model:**
  + Import the XGBoost regressor from the XGBoost library.
  + Initialize the XGBoost regression model (XGBRegressor) with specific parameters:
    - objective='reg:squarederror': Specifies the regression objective and utilizes squared error loss to assess the difference between predicted and actual values.
    - eval\_metric='rmse': Determines the metric used to evaluate the model's performance, with 'rmse' (Root Mean Squared Error) measuring the root of the average squared differences.
    - learning\_rate=0.1: Sets the learning rate of the model.
    - max\_depth=3: Defines the maximum depth of the trees in the ensemble.
    - n\_estimators=100: Specifies the number of trees in the ensemble.
* **Train the Model on the Training Data:**
  + Utilize the fit method to train the XGBoost regression model on the training data.
  + The training data consists of the feature matrix (X\_train) and the corresponding target vector (Y\_train).
  + This step establishes the XGBoost regression model with specified parameters and initiates its training on the provided training dataset. The resulting trained model can subsequently be used for predictions and performance evaluation.

Following the training of the XGBoost regression model, predictions are generated for the training, testing, and validation sets. For the training set, the model predicts the target variable (unemployment\_rate) using the feature matrix (X\_train), and the results are stored in y\_pred\_train. Similarly, predictions for the testing set are obtained and stored in y\_pred. The validation set predictions are stored in y\_pred\_val. These predicted values provide insights into the model's performance on both the training data and previously unseen datasets, facilitating the evaluation of its predictive capabilities.

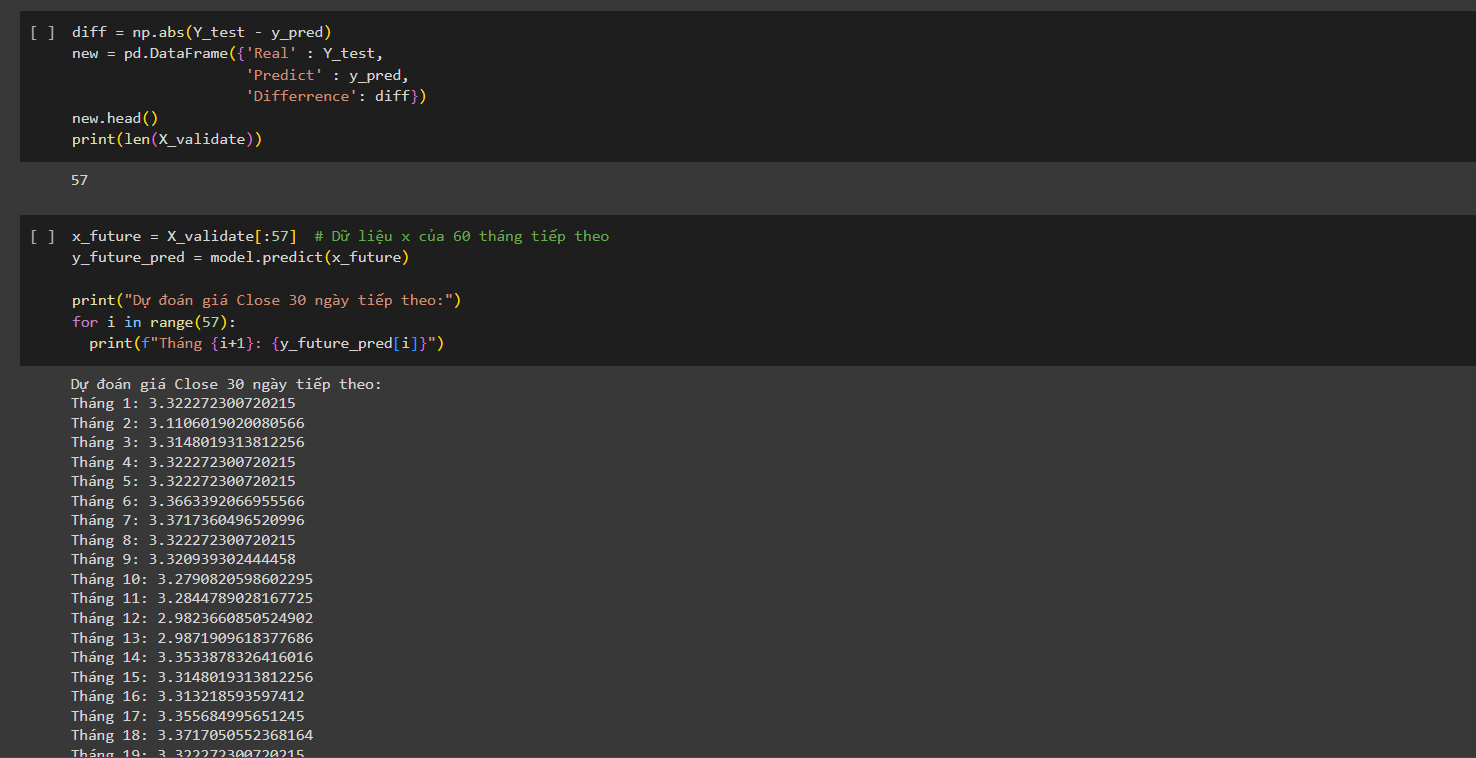


**And now, Model Evaluation on the Test Set (Ratio 7:2:1):** The model is rigorously assessed on the test set using various evaluation metrics to gauge its predictive performance. The evaluation metrics and their interpretations are succinctly described as follows:

* **Mean Squared Error (MSE):**
  + MSE is calculated as the sum of squared errors on the test dataset.
  + A smaller MSE indicates a more accurate model prediction.
* **Root Mean Squared Error (RMSE):**
  + RMSE is the square root of the MSE and is employed to assess the average error between predicted and actual values.
  + A smaller RMSE implies higher model accuracy.
* **Mean Absolute Error (MAE):**
  + MAE is the mean absolute difference between predicted and actual values.
  + A smaller MAE signifies greater model accuracy.
* **Mean Absolute Percentage Error (MAPE):**
  + MAPE computes the average percentage difference between predicted and actual values.
  + Lower MAPE values indicate higher model accuracy, with values below 10% considered very good.
* **R-squared (R2) Score:**
  + R2 evaluates how well the model fits the data, ranging from 0 to 1.
  + A value of 0 suggests no relationship, while 1 indicates a perfect fit.
  + These metrics collectively provide a comprehensive overview of the model's performance on the test set, offering insights into its accuracy, precision, and overall suitability for predicting the unemployment rate based on the selected features. The printed results of MSE, RMSE, MAE, MAPE, and R2 further quantify the model's effectiveness.



And the last step, we will forecast in the next 60 months:



#### ***3.4.2. Result***

### 

* The prediction results of the unemployment rate using the xgboost model show that the unemployment rate will increase again in recent years. Although in the next 1-2 years the unemployment rate will decrease to 3% and then increase and stabilize at 3.5%.
* This forecast aligns with the broader economic analysis indicating a prolonged impact of the 2022 economic crisis until 2023, with signs of gradual recovery thereafter. The cyclicality of economic crises becomes apparent as we observe the current trajectory entering a phase of decline.
* The insights gathered from the xgboost model reinforce the notion that economic uncertainties, both globally and within Vietnam, play a pivotal role in shaping future employment trends. It is crucial for policymakers and businesses to remain vigilant and adaptive, considering the cyclical nature of economic challenges.
* In conclusion, while the short-term outlook may offer a reprieve with a temporary reduction in the unemployment rate, the overarching trend points towards a subsequent increase, underscoring the need for strategic measures to navigate the evolving economic landscape. This period of economic flux necessitates proactive and flexible approaches to mitigate the impact on employment and foster sustainable recovery.

### 

# **Chapter 5: Conclusion**

**Future development directions:**

* Economic development: Economic growth is an important driving force to create more jobs. The government needs to have policies to promote economic development, especially industries and services that can create many jobs.
* Create more jobs: The government needs to have policies to support business development and create more jobs for workers. These policies may include capital support, vocational training,...
* Improving the quality of human resources: To improve the competitiveness of workers in the labor market, it is necessary to improve the quality of human resources through vocational training, training to improve educational qualifications, etc. .
* Strengthen social policies: The government needs to have policies to support unemployed workers, such as unemployment insurance, social benefits, etc. These policies help workers have time to search. new jobs and reduce difficulties when unemployed.

# **References**

[1] Ryan Mayolo , "Analysis of GDP using Linear Regression" (2018). Undergraduate Theses and Capstone Projects. 56.

[2] Ghosal, V., & Ye, Y. (2015). Uncertainty and the Employment Dynamics of Small and Large Businesses. Small Business Economics, 44(3), 529-558. doi:<http://dx.doi.org/10.1007/s11187-014-9614-0>

[3] Marno Verbeek “Using linear regression to establish empirical relationships”. IZA World of Labor 2017: 336 doi: 10.15185/izawol.336

[4] Angrist, J. D., and A. B. Krueger. “Empirical strategies in labor economics.” In: Ashenfelter, O., and D. Card (eds). Handbook of Labor Economics, Volume 3. Amsterdam: Elsevier, 1999; pp. 1277–136

[5] Prof. Univ. Dr. Constantin Anghelache, Drd. Cristina Sacala “Multiple linear regression used to analyse the correlation between GDP and some variables”, Romanian Statistical Review - Supplement nr. 9 / 2016

[6] Anghelache, C., Manole, A., Anghel, M.G. (2015). *Analysis of final consumption and gross investment influence on GDP – multiple linear regression model*, Theoretical and Applied Economics, No. 3/2015 (604), Autumn, pg 137-142, ISSN 1841-8678

[7] Urmat Dzhunkeev, Lomonosov Moscow State University; RANEPA “Forecasting Unemployment in Russia Using Machine Learning Methods”: https://www.researchgate.net/profile/Urmat-Dzhunkeev-2/publication/359169888\_Forecasting\_Unemployment\_in\_Russia\_Using\_Machine\_Learning\_Methods/links/62cd8bfcc276426014a96342/Forecasting-Unemployment-in-Russia-Using-Machine-Learning-Methods.pdf

[8] Yasin KÜTÜK - Prediction of Transition Probabilities from Unemployment to Employment for Turkey Via Machine Learning and Econometrics: A Comparative Study: <https://dergipark.org.tr/en/download/article-file/704638>

[9] Íñigo Martínez de Rituerto de Troya, Ruqian Chen - Predicting, explaining, and understanding risk of long-term unemployment: https://www.sciencedirect.com/science/article/pii/S1877050920311467

[10] [D Didiharyono](https://scholar.google.com/citations?user=G4ujGyMAAAAJ&hl=vi&oi=sra), [M Syukri](https://scholar.google.com/citations?user=PjBaNZ8AAAAJ&hl=vi&oi=sra) - Int. J. Sci. Technol. Res, 2020 - osf.io, [Forecasting with arima model in anticipating open unemployment rates in south sulawesi](https://osf.io/f5zh6/download).

[11] [MO Adenomon](https://scholar.google.com/citations?user=fqNq8EcAAAAJ&hl=vi&oi=sra) - FUW Trends in Science and Technology Journal, 2017, [Modelling and forecasting unemployment rates in Nigeria using ARIMA model](http://www.ftstjournal.com/uploads/docs/21B%20Article%2035.pdf).

[12] E Rublikova, M Lubyova - Forecast. Pap. Progn. Pr, 2013, [Estimating ARIMA-ARCH model rate of unemployment in Slovakia](https://www.prog.sav.sk/wp-content/uploads/PP_5_2013_3_clanok3_Rublikova_Lubyova.pdf).

[13] Rami Alamoudi, Mohammed Balubaid, Amir Siddiquir -Use of Multiple Regression Analysis to Identify Factors that Affect the Unemployment Rate in the Kingdom of Saudi Arabia: <https://www.lifesciencesite.com/lsj/life120715/003_28932life120715_15_22.pdf>