Understanding Relationships between Economic Indicators and Financial Markets

by Adri KATYAYAN

Submission date: 16-Nov-2023 07:23PM (UTC+0800)
Submission ID: 2218973820
File name: report_1.docx (3.69M)
Word count: 5893
Character count: 33780

Understanding Relationships between Economic Indicators and Financial Markets

Project Report

MANIPAL ACADEMY OF HIGHER EDUCATION

For Partial Fulfilment of the Requirement for the

Award of the Degree

Of

Bachelor of Technology

In

Computer and Communication Engineering

By

Soumili Acharya

Reg. No.: 210953210

Adri Katyayan

Reg. No.: 210953218

Priyanshi Agarwal

Reg. No.: 210953228

Under the guidance of

Mr. Chetan Sharma
Assistant Professor-Senior Scale
Department of I&CT
Manipal Institute of Technology
Manipal, Karnataka, India

Dr. Kaliraj S.
Assistant Professor-Senior Scale
Department of I&CT
Manipal Institute of Technology
Manipal, Karnataka, India

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Unit of MAHE, Manipal

INDEX

INTRODUCTION:

In today's world where a country's economy is used as a way of comparison with the other countries, understanding the complicated and intricate relationship between various economic factors that shape a country's economy is pivotal for making informed decisions and predictions.

This demands advanced analytical approaches that can be used to unravel these patterns and dependencies. In this era of data-driven decision-making, the of understanding relationships between economic indicators and financial markets has never been more critical. Our project delves into this study and understanding, focusing on a comprehensive set of economic factors from various countries using clustering and regression analysis methods.

LITERATURE SURVEY:

PAPER 1

"This paper is a review paper, talking about various clustering techniques and developments".

PROBLEM STATEMENT:

- The paper talks about how regression analysis is an important tool when it comes to explanation and prediction models. However, regression analysis consists of least square variables that may perform poorly in presence of outliers and biases. Therefore, many studies have talked about various ways to deal with this problem.
- But since these studies have taken place in disconnected regions and the dataset has never been ample, these efforts have majorly failed to accomplish this aim.
- Therefore, this paper has studied all the studies made on M-estimators in various disconnected areas and presented it as a review paper.

METHODS USED:

1. **L-estimators-** The Linear estimators are statistical estimators which are usually used to depict linear parameters such as population, location or scales.
2. **R-estimators-** Robust estimators are those designed to be resistant towards the outliers present in data.
3. **M-estimators-** The Maximum-Likelihood type estimators are typically used to maximise likelihood functions.

ADVANTAGES AND DISADVANTAGES:

L-estimators:

- Advantages:
  - They are effective against outliers and are therefore useful when data has extreme and unpredictable values.
  - They are simple and straightforward as compared to the R-estimators and M-estimators.
  - Solutions (closed form) are possible for various datasets and problems.
- Disadvantages:

- The performance of these estimators is relatively sensitive towards assumptions regarding distribution as compared to R-estimators and M-estimators.
- The efficiency of L-estimators are not at par with the other two in the study.

R-estimators:

- Advantages:
  - As the name suggests, these parameters are specifically made to deal with outliers and extreme data values, and thus can easily detect potential outliers in the data provided.
- Disadvantages:
  - These were found to be less effective than M-estimators in the study, where data was close to the assumed model.
  - The complexity of R-estimators is high since it includes several iterative algorithms.

M-estimators:

- Advantages:
  - Highly efficient as compared to R and L estimators.
  - Has lower complexity as compared to R-estimators.
- Disadvantages:
  - Unlike R-estimators, M-estimators are sensitive to outliers.
  - Requires intensive data preprocessing to produce accurate results.

RESULTS AND CONCLUSION:

- Comparing the 3 estimators, the study found that L-estimators are simple and are computationally less complex than the other two but lack robustness due to shear simplicity.
- R-estimators help address l-estimator's issue with robustness and is immune to outliers but has high complexity and therefore must be used only when dealing with outliers is the utmost priority.
- M-estimators, has lower complexity than r-estimators but is sensitive to outliers and extensively rely on assumptions related to distribution. However, being a wider category of estimators than the other two, if efficiency is in question and distribution assumptions are well-understood, M-estimators are the most suitable.

The paper acknowledges that although these robust regression techniques and estimators are not as widely used today, but the demand for M-estimators may increase in future when R-estimators becomes more complex than what computation capacity of a system may permit.

PAPER 2

“This paper is about portfolio formation as well as optimization using continuous realignment.”

PROBLEM STATEMENT:

Four facets of the issue have been explored in the literature on portfolio optimization, namely:

i. The stock selection for the portfolio
ii. Defining the constraints and the objective function(s)
iii. Formula is used to calculate how much each stock in the portfolio is worth.
iv. Evaluating the portfolio's performance.

METHODS USED:

By using both vertical and horizontal clustering algorithms, the authors of this research provide a novel method for building investment portfolios. They impose exposure limitations on each stock while using a clustering algorithm to ensure a diversified selection of equities in the portfolio. Utilizing a variable-length Non-dominated Sorting based Genetic Algorithm (NSGA-II), the near Pareto optimum portfolios are produced. A single goal Genetic Algorithm (GA) based Markowitz model is then used to calculate quarterly weights for the stocks in each portfolio, allowing for dynamic adjustment based on the macroeconomic environment.

When these dynamic portfolios' performance is compared to a benchmark portfolio, the findings show that the suggested approach continuously beats the benchmark index return over the course of the study.

ADVANTAGES AND DISADVANTAGES:

Advantages:

- Vertical and horizontal clustering are both used in this method to ensure that the portfolio contains a varied range of stocks. Diversification can aid in risk distribution and lessen the negative effects of underperforming assets on the portfolio as a whole.
- Exposure Limits: The methodology reduces risk by preventing over-concentration in a single stock or industry by putting exposure limits on each stock. For investors looking to build a well-balanced portfolio, this risk management function is essential.
- Dynamic Realignment: The portfolio can adjust to shifting macroeconomic conditions by using quarter-wise weights and dynamic realignment. Insecure financial markets can benefit from this flexibility.
- Outperformance: The methodology allegedly consistently outperforms a benchmark portfolio during the course of the study. This indicates that it may have the ability to produce better.
- Comparison and Validation: The effectiveness of the methodology is thoroughly assessed and compared to four well-known clustering algorithms, giving it more support.

Disadvantages:

- The process seems to involve a lot of steps, including clustering, genetic algorithms, and dynamic realignment. Investors who are unfamiliar with these tactics may find it more difficult to access due to its complexity.
- Availability and quality of data, particularly historical stock prices, macroeconomic indicators, and other pertinent financial data, are essential to the success of this methodology. The results could be impacted by data gaps or inaccuracies.
- Genetic algorithms frequently require parameter tuning, and the choice of parameters may have an impact on how well the methodology works. It can be difficult to determine the best parameter settings.
- Transaction expenses: It appears that the technique does not take into account the transaction expenses involved in realigning a portfolio. Transaction costs may increase as a result of frequent realignment, which could reduce overall returns.

RESULTS AND CONCLUSION:

The complexity of having many objectives and various constraints has received most of the attention in the literature on multi-objective portfolio optimization. They have offered solutions using innovative algorithms that, when applied to various criteria, have produced effective solutions. Some have placed an emphasis on shorter computation times, while others have compared their methods to others of a similar nature and calculated GD and IGD as well as Hyper volume, and epsilon metrics.

PAPER 3

"This paper is a survey about node-attributed social networks using Community detection."

EMPACT STATEMENT:

The primary goal of the paper is to describe and clarify the current state of affairs in the field of community detection within node-attributed social networks. This suggests that the authors want to provide an overview of the existing approaches and contribute to the understanding of this specific area of research.

METHODS USED:

The authors plan to achieve their objective by conducting an exhaustive search of known methods in the field. This means they are going to review and analyze existing techniques used for community detection in such networks.

Early Fusion Methods: These methods for community detection incorporate node properties and network structure before the community detection process really starts. The data will be preprocessed with the intention of making it compatible with traditional community detection algorithms so that researchers can use existing software implementations for their investigation.

Simultaneous Fusion Methods: This class of community identification techniques involves the simultaneous fusion or integration of network structure and node attribute information during the community detection procedure itself. Contrary to early fusion and late fusion techniques, which mix these two forms of data with the community detection algorithm as a separate preprocessing phase, simultaneous fusion techniques incorporate these two types of data simultaneously.

Late Fusion Methods are a class of community detection techniques in which the network structure and node attributes are fused or integrated after the community detection procedure itself. In other words, node attributes (i.e., extra data connected with nodes) and the network topology (i.e., connections between nodes) are first treated separately while doing community discovery. Late fusion approaches combine the individual partitions that were obtained based on these two factors to produce the final partition.

ADVANTAGES AND DISADVANTAGES:

Advantages:

- They are simple to develop since they are compatible with the current classical community detection algorithms.
- Integrative: Offer a comprehensive perspective by taking into account the network topology and node properties right away.
- Simplicity: Compared to simultaneous or late fusion approaches, these are frequently easier to implement.
- Comprehensive: Take into account qualities and structure more thoroughly, potentially capturing subtle community tendencies.
- Allow for customized fusion approaches based on the unique specifications of the issue.
- Better Accuracy: When interactions between structure and attributes are complex, this may result in more accurate community detection.

Disadvantages:

- Complexity: Because of their interconnected nature, they frequently call for specialist software implementations.
- Cost of Computing: This method could be computationally more expensive than early or late fusion techniques.
- Flexibility Issues: May not be as effective at capturing intricate links between structure and attributes as simultaneous or late fusion methods.
- Algorithm Selection: There is no one-size-fits-all fusion algorithm, therefore selecting the right one can be crucial.

RESULTS AND CONCLUSION:

- The study demonstrates the diversity of methodologies for community recognition in node-attribute social networks, describing 75 major strategies and citing related ones. This illustrates the diversity of the field and its ongoing investigation of various tactics, from early to late fusion. Insights regarding technique selections based on study objectives and dataset features are provided by the survey, which is a useful tool for academics and industry professionals. It emphasizes how dynamic this discipline is, constantly changing as new theories and techniques are developed.

PAPER 4

"This paper is a Survey on Multiview Clustering."

PROBLEM STATEMENT:

- To provide new categorization of existing model-view-control methods and introduce representative algorithms in each category.
- Also, pointing out open problems to investigate advance the MVC study.

METHODS USED:

1. Generative Modelling: Generative modelling is a type of machine learning approach that involves modelling the distribution of collected dataset to generate new samples that resembles original data.

   Under generative approach we have mixture models and CMM (categorical mixture model).

2. Discriminative Modelling: Instead of explicitly modelling the probability distribution of the data, discriminative techniques, on the other hand, concentrate on directly optimizing a measure of cluster quality or separation.

   Based on the available data, they seek to identify a decision boundary or distinction between clusters.

ADVANTAGES AND DISADVANTAGES:

1. First, since generative approaches rely on data distribution, they should work well if data follows the distribution assumed by the study previously.

2. Second, the number of clusters does not need to be predetermined.

3. Multiview generative clustering is a direction that we think is underutilized; more work can be done in this area in the future.

4. Discriminative strategies are typically easier to apply and conceptually simpler. They concentrate on achieving well-separated clusters by maximizing a distinct objective function.

5. Efficiency: Since these methods avoid simulating the complex data distribution, they frequently use less computer power than generative approaches.

6. Direct Cluster Separation: When the main objective is to identify unique and well-separated clusters, discriminative approaches, which are designed to directly optimize the separation between clusters, can be useful.

RESULTS AND CONCLUSION:

- Generative approaches have made far less progress than discriminative algorithms. Despite its intrinsic limitations, it satisfactorily handles with missing data and, therefore it merits additional consideration.

PAPER 5

"This paper uses meta regression analysis to understand the effects of urbanization on the wage gap in China."

PROBLEM STATEMENT:

- The wage gap in urban and rural population in China has been widening since 2000s.
- This study uses meta regression analysis to understand this problem by means of the URIG indicators.

METHODS USED:

1. Meta-Regression Analysis (MRA) Method-
   - **a.** It is a statistical regression technique that includes quantitative synthesis from multiple studies on a specific topic.
   - **b.** According to the study, this is one of the meta-analysis techniques which are generally used to integrate various economic estimates to provide a reasonable conclusion and summary to quantitative research.

ADVANTAGES AND DISADVANTAGES:

1. The advantage of Meta-regression analysis is that it helps define heterogenous sources among results in a study. This helps a person understand why a single study may produce different effects and conclusions and may help researchers identify the main moderation factors.
2. Larger the dataset, more accurate is the conclusion given by the MRA.
3. However, it depends on availability and quantity of good data since noisy and inconsistent data can limit analysis and accuracy.
4. This may also lead to overfitting of data.

RESULTS AND CONCLUSION:

Study found that more than half of the studies indicated a negative association between urbanization and URIG. It was found that a higher level of urbanization resulted in the reduction of URIG in China.

Researchers acknowledge the fact that although the MRA has valuable results and insights, one must carefully consider its limitations such as data quality, overfitting, public bias involved with it.

PAPER 6

"The selection method for K-value of the K-means Clustering algorithm."

METHODS USED:

1. An Elbow Method Algorithm: It uses sum of squared errors as performance indicator. Degree of convergence of each cluster is indicated by smaller values.
2. The Gap Statistic Algorithm: It uses the Monte Carlo sampling method. It uses the output of the algorithm, comparing the total intra-cluster variation to determine optimal number of clusters.
3. The Silhouette Coefficient Algorithm: It comprises two factors: cohesion and separation. Silhouette values range between -1 and 1. It indicates how the relationship is there between the object and the cluster.

The Canopy Algorithm: It forms a Canopy which is created by dividing the data into overlapping subsets. Here, each subset is a cluster.

ADVANTAGES AND DISADVANTAGES:

ADVANTAGES:

1. The elbow method: It has a simple complexity.
2. Canopy algorithm: The fault tolerance and the noise immunity are increased by adding overlapping datasets.

DISADVANTAGES:

1. Gap Statistic Algorithm: Not desirable for large scale datasets.

The Silhouette Coefficient Method: The computational complexity us O(n2).

PERFORMANCE:

NO.	NAME	K VALUE	EXECUTION TIME
1	ELBOW METHOD	2	1.830 s
2	GAP STATISTICS	2	9.763 s
3	SILHOUETTE COEFFICIENT	2	8.648 s
4	CANOPY	2	2.120 s

RESULTS AND CONCLUSION:

Here, four methods are used and the best one is the Canopy Algorithm because problems caused by larger clusters and computations are avoided by using this algorithm.

PAPER 7

"The paper emphasizes Performance evaluation using Silhouette Analysis in Machine Learning. Furthermore, it focuses on the K-means clustering methods."

METHODS USED:

The minimizing cluster distance or the maximizing of distance between cluster systems helps to optimize the objective function. K-means identifies clusters that are linearly clustered and kernel k-means identifies the ones which are non-linearly separable.

1. K-means:

   Euclidean distance is calculated between each element and all clusters separately. Process is repeated till Euclidean values are not constant

2. Kernel K-means:

   The objects that are not linearly separable are grouped using kernel k-means which is an extension of k-means.

3. Silhouette Index
4. Weighted Clustering using Silhouette Method: The results of the performances of different clustering methods are combined to give a single outcome.
5. Simulation

ADVANTAGES AND DISADVANTAGES:

Advantages:

1. Kernel K-means: nonlinearly separable clusters can be grouped using this method.
2. Silhouette Index: Doesn't require a training set.
3. The Gaussian kernel performs better than the other kernels.

No disadvantages were mentioned explicitly.

RESULTS AND CONCLUSION:

Five methods are used to find the result. The results are combined by three different kernels.

PAPER 8

"This paper is about classification of Antineutrophil Cytoplasmic Antibody-Associated Vasculitis Using Cluster Analysis of Clinical Phenotypes: A Retrospective Cohort Study from a Single Centre"

METHODS USED:

1. Agglomerative hierarchical clustering method is used here
2. Several Korean patients diagnosed with AAV classified into mutually exclusive clusters. The outcomes of the resulting clusters are analysed in order to investigate classification's clinical significance.

ADVANTAGES AND DISADVANTAGES:

Advantages:

The agglomerative clustering method is one of the topmost methods to classify data based on the similarity of the inputs.

RESULTS AND CONCLUSION:

In terms of clinical relevance, one method, here, is appropriate for AAV revealing the phenotypic diversity. A simple distance-based algorithm is used because easy modification can be done for specific clinical needs.

PAPER 9

“Unsupervised K-Means Clustering Algorithm”

PROBLEM STATEMENT:

To develop unsupervised algorithm for k-means algorithm, called u-k means algorithm, that is free of initializations of K.

METHODS USED:

1. expectation and maximization (EM) algorithm
2. U-k-means clustering algorithm
3. X-means Algorithm

ADVANTAGES AND DISADVANTAGES:

1. A clear disadvantage of the EM algorithm is that it sensitive to the choice of initial parameter values, which, in turn, impacts the final estimates and accuracy.
2. The disadvantage of k-means algorithm is that it cannot be implemented without initializations or parameter selection.
3. X-means can handle datasets where the underlying number of clusters might vary across different regions. This flexibility has been beneficial when dealing with complex data distributions. However, it was evident that this algorithm is prone to overfitting, especially when true underlying cluster structure is not well-defined.
4. The advantage of U-k-means is that free of initializations and parameters. Additionally, it’s applicable to different cluster volumes and automatically finds the number of clusters.

RESULTS AND CONCLUSION:

1. K-means is common for implementing clustering analysis by portraying similarities within same clusters and dissimilarities between different clusters.
2. However, k-means requires initialization and parameter selection. Also, it cannot simultaneously find frequency of clusters in a given dataset. The so proposed algorithm is developed to overcome this hurdle and is applied on various datasets, yielding high accuracy rates for all of them.

PAPER 10

"The paper talks about uncertain M- estimation and it's application in field of uncertain regression models.

PROBLEM STATEMENT:

There exists a problem of estimating unknown parameters in uncertain distributions of population based on the likelihood measured by uncertainty theory.

This study uses M- estimation in uncertain regression analysis to find correlation between explanatory and response variables where imprecise observations are considered as uncertain variables.

While studies have previously explored methods for improvement, this paper contributes a new approach, leveraging the principles of uncertainty theory and maximum likelihood estimation.

METHODS USED:

1. Likelihood function- It is the basis in statistical modelling that measures likelihood of different attributes given in a dataset.
2. Maximum Likelihood estimator- The Maximum-Likelihood type estimators are typically used to maximise likelihood functions.
3. Uncertain Regression Analysis- models the relationship between variables in case of inaccurate observations and conclusions.

ADVANTAGES AND DISADVANTAGES:

1. The advantage of Likelihood function is that it provides a basic framework required for Maximum likelihood estimator or the MLE method used in this study. However, its disadvantage is that it is sensitive to data with extreme values and distributional assumptions. This makes it incapable to be interpreted in complex models.
2. The advantage of MLE method is that it is highly efficient than other estimators generally used such as L-estimator or R-estimator and has relatively lower complexities. It is theoretically well understood today but is still sensitive to outliers and is not advisable to be used when data is noisy or has extreme values.

3. Advantages of Uncertain Regression Analysis include accurate observations when given good data and accommodation of situations where precise measurements are impossible to make. However, it requires good data and analysis methods in order to do this, which in turn increases the complexity of this algorithm.

RESULTS AND CONCLUSION:

- The study resulted in a new method that proposed the maximum likelihood that determines unknown parameters in unpredictable regression models.

METHODOLOGY:

Part1- Clustering Analysis

Data collection and data cleaning are the very first steps of the Project. This is done using three authentic sources-Quandl API, World Bank Data and Kaggle datasets. The two key features of the tables are: “countryname” and “year”. Target value for the project is- “GDP”.

We start by examining correlations among the chosen economic indicators. These include data for Consumer Price Index (CPI), Gross Domestic Product (GDP), 10-year treasury yield, unemployment rates, housing statistics, and corporate profits from many countries acquired using the Quandl API. Through correlation matrix we discern patterns of dependencies, associativity, and correlation between these factors. A heatmap is generated to visualize the correlation matrix, providing insights into the strength and direction of relationships.

As we progress, our attention shifts to clustering methods such as agglomerative hierarchal clustering revealing clusters of influence within the dataset. The resulting dendrograms serve

as visual representation of the relationships among economic indicators, offering insights into the structural organization of the global economic system.

Further, Cross-correlation functions (CCFs) are computed to examine the time-lagged relationships between GDP and other economic indicators.

Since the data we are using is Linear data, we go beyond correlation and employ the Granger causality tests, aiming to discern the causal relationships between GDP and other critical economic variables. Granger causality tests are performed to assess if past values of one economic indicator provide information about future values of another. These tests are conducted for GDP with a few other indicators selected after finding out the correlations.

After the causality tests, Spectral clustering is employed to group countries based on their economic indicators. The optimal clusters for the clustering analysis are chosen based on the dendrogram and domain knowledge. The time series data for each cluster chosen is then visualized to observe trends and patterns. Rolling averages (1-year and 2-year) are computed for each cluster to identify long-term trends. Qualitative implications are provided based on the cluster characteristics.

Economic indicators showing visible results are then used for part 2 of the project.

Part 2- Using the data to predict GDP using Regression analysis.

AIM: to predict the GDP for various countries based on GDPs of various countries. The GDP of countries is impacted by various social, economic, cultural parameters. We are analysing those parameters that are selected by us in part 1 of the study.

Since most of these models require clean data, we apply the following data cleaning methods on the dataset:

- DATA MERGING: We load the chosen datasets to be merged using innerjoin using columns countrycode and countryname.
- HANDLING MISSING DATA: Since we have numeric data, we replace null values with mean of corresponding columns as follows:

```
print(GDP_Combine['Women_Informed_Choices'].isnull().sum())
GDP_Combine['Women_Informed_Choices'].fillna(value=GDP_Combine['Women_Informed_Choices'].mean(),inplace=True)
print(GDP_Combine['Women_Informed_Choices'].notnull().sum())
```

```
print(GDP_Combine['RuralPopulation_PerCent'].isnull().sum())
GDP_Combine['RuralPopulation_PerCent'].fillna(value=GDP_Combine['RuralPopulation_PerCent'].mean(),inplace=True)
print(GDP_Combine['RuralPopulation_PerCent'].notnull().sum())
```

```
print(GDP_Combine['LegalRights_Strength'].isnull().sum())
GDP_Combine['LegalRights_Strength'].fillna(value=GDP_Combine['LegalRights_Strength'].mean(),inplace=True)
print(GDP_Combine['LegalRights_Strength'].notnull().sum())
```

```
print(GDP_Combine['CreditTo_PrivateSector'].isnull().sum())
GDP_Combine['CreditTo_PrivateSector'].fillna(value=GDP_Combine['CreditTo_PrivateSector'].mean(),inplace=True)
print(GDP_Combine['CreditTo_PrivateSector'].notnull().sum())
```

```
print(GDP_Combine['BirthsAttendedby_SkilledStaff'].isnull().sum())
GDP_Combine['BirthsAttendedby_SkilledStaff'].fillna(value=GDP_Combine['BirthsAttendedby_SkilledStaff'].mean(),inplace=Tr
print(GDP_Combine['BirthsAttendedby_SkilledStaff'].notnull().sum())
```

```
print(GDP_Combine['ATMMachines_Ratio'].isnull().sum())
GDP_Combine['ATMMachines_Ratio'].fillna(value=GDP_Combine['ATMMachines_Ratio'].mean(),inplace=True)
print(GDP_Combine['ATMMachines_Ratio'].notnull().sum())
```

```
print(GDP_Combine['Agricultural_Machines'].isnull().sum())
GDP_Combine['Agricultural_Machines'].fillna(value=GDP_Combine['Agricultural_Machines'].mean(),inplace=True)
print(GDP_Combine['Agricultural_Machines'].notnull().sum())
```

```
print(GDP_Combine['LiteracyRate_Adult'].isnull().sum())
GDP_Combine['LiteracyRate_Adult'].fillna(value=GDP_Combine['Agricultural_Machines'].mean(),inplace=True)
print(GDP_Combine['LiteracyRate_Adult'].notnull().sum())
```

```
print(GDP_Combine['AccountsRatio_FinancialInst'].isnull().sum())
GDP_Combine['AccountsRatio_FinancialInst'].fillna(value=GDP_Combine['AccountsRatio_FinancialInst'].mean(),inplace=True)
print(GDP_Combine['AccountsRatio_FinancialInst'].notnull().sum())
```

```
GDP_Data = pd.merge(GDP_Country, population_per_country, on= ['Country Code','Country Name'], how='inner')
GDP_Data.head()
```

After completing the given steps, Exploratory Data analysis (EDA) is done for better understanding of variable correlation. Descriptive statistics are generated, and correlation matrix is made with the help of heatmap. Pair Pilots are also created in this step categorizing by strength of legal rights.

In the next step, we perform the actual Regression modelling. Data is split into testing and training sets. We are using the following regression algorithms:

- Multiple Linear Regression
  - An extension of Linear Regression.
  - $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + €$
  - Purpose is finding interdependence of the variables.
- Polynomial Regression
  - An extension of Linear Regression where X and Y are modelled as an n-degree polynomial.
  - $Y = b_0 + b_1X + b_2X^2 + \ldots + €$
  - You don't get a straight line since variables are represented by polynomials.
- Decision Tree Regression
  - Uses decision trees as predictive model to map features to a target variable by splitting the original dataset into subsets based on significant attributes.
  - Tree is constructed by this splitting of dataset at each node of the tree.
- Random Forest Regression

        - The method constructs multiple decision trees during training and gives the mean of the predictions prediction for all trees formed as output.
        - effective against overfitting and improving accuracy.
    - Ridge Regression
        - also known as L2 regularization, is linear regression but with a regularization term to prevent overfitting in the model.
        - It is used when there is multicollinearity or high correlation in the data.
    - Lasso Regression
        - Also known as L1 regularization.
        - Same as ridge regression, but has additional coefficients.
        - Used when feature selection is important.
    - Elastic Net Regression
        - Combination of Ridge and Lasso Regressions.

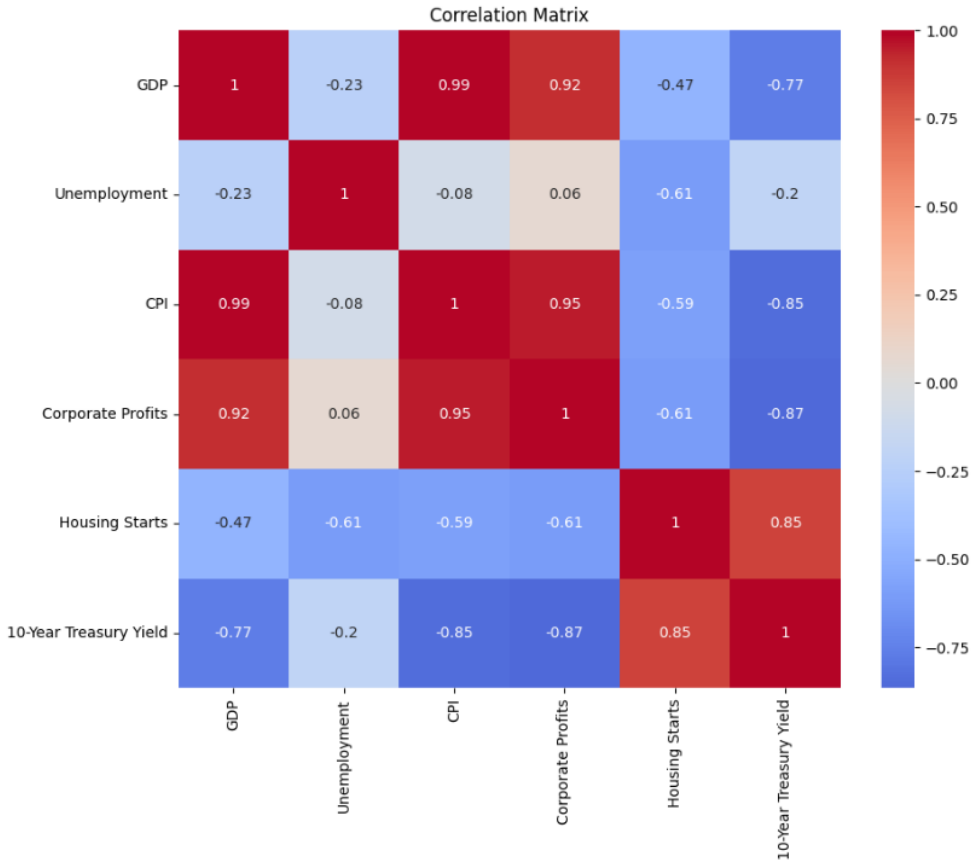
For both the training and testing sets, we employ the RMSE, R squared values for both training and testing sets. K-fold validation scores are also used in each algorithm. These two form the evaluation metrics for the models. The results are later stored and compared for further analysis.

Heatmaps, pair plots, and graphs are used for results visualization. We now move on to model training and visualization. Principle Component analysis or PCA is also used for dimensionality reduction.

After all this prediction of new data can take place by entering the name of a specific country (India, United States, Japan, etc) to predict the GDP of that country. The value so received is compared with historical values to test the reliability of models based on the evaluation metrics spoken about previously. This way the aim of the study shall be achieved.

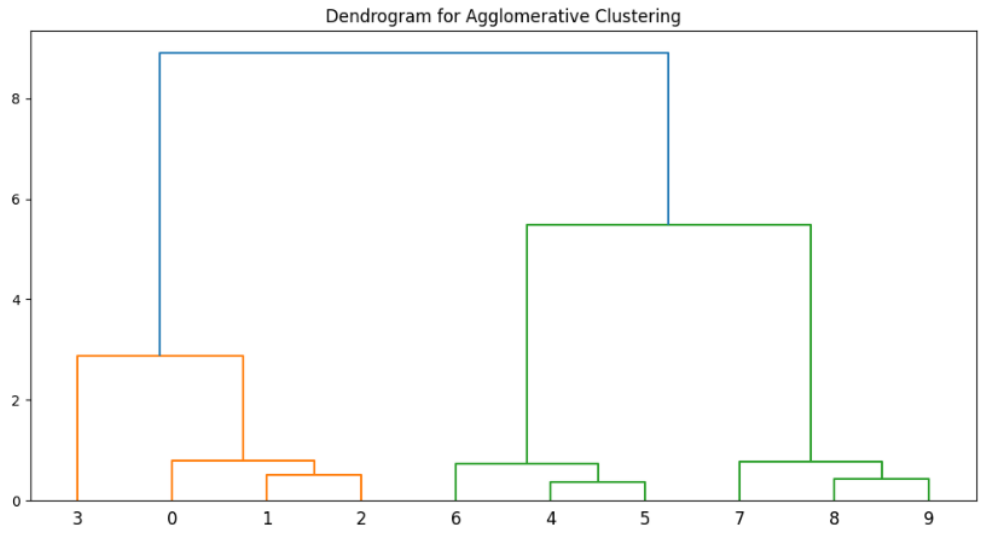
DISCUSSION AND RESULTS:

On examining correlations among the chosen economic indicators that include Consumer Price Index [CPI], Gross Domestic Product [GDP], 10-year treasury yield, unemployment rates, housing statistics, and corporate profits we get the following correlation matrix:



This heatmap shows high correlation GDP and CPI and GDP and Corporate Profits. These two indicators are later used in regression models along with percentage working population, population per country, literacy rates and domestic credit to private sector. Other indicators that is unemployment, housing starts, and treasury yield are tested further with granger causality before being eliminated from the study and not being used further.

Here is a dendrogram for better visualization.



On employing Granger causality tests- which are basically used to test hypothesis. These tests are conducted for GDP and unemployment, GDP and CPI, and GDP and the 10yr treasury yield.

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.2174  , p=0.6458  , df_denom=21, df_num=1
ssr based chi2 test:   chi2=0.2484  , p=0.6182  , df=1
likelihood ratio test: chi2=0.2472  , p=0.6191  , df=1
parameter F test:         F=0.2174  , p=0.6458  , df_denom=21, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.7088  , p=0.5055  , df_denom=18, df_num=2
ssr based chi2 test:   chi2=1.8114  , p=0.4042  , df=2
likelihood ratio test: chi2=1.7437  , p=0.4182  , df=2
parameter F test:         F=0.7088  , p=0.5055  , df_denom=18, df_num=2

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.1118  , p=0.7415  , df_denom=21, df_num=1
ssr based chi2 test:   chi2=0.1277  , p=0.7208  , df=1
likelihood ratio test: chi2=0.1274  , p=0.7212  , df=1
parameter F test:         F=0.1118  , p=0.7415  , df_denom=21, df_num=1
```

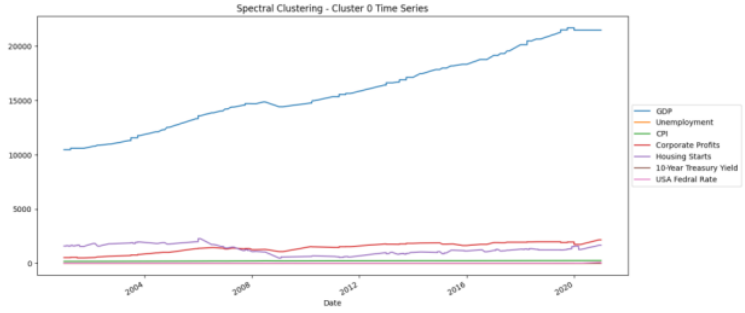
The above figure shows the results when granger causality tests are applied on GDP and 10-year treasury yield, GDP and unemployment, and GDP and CPI respectively.

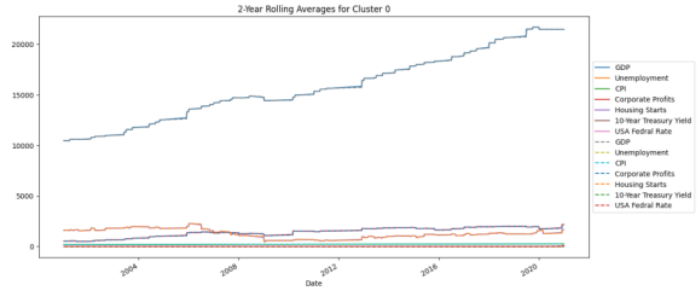
1. granger causality tests on GDP and 10-year treasury yield:
    - SSR based F test and chi2 test along with likelihood ratio test and parameter F test has been employed here.

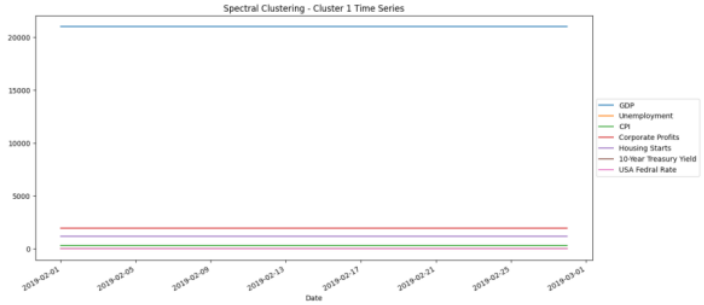
- P is used as an indicator to whether to accept or reject the hypothesis. If p>=0.5, we reject the hypothesis else, we accept it.
- Since the value of p is greater than 0.6 through all 4 tests, we reject the hypothesis that “10-year treasury yield has no effect on the GDP of a country”.

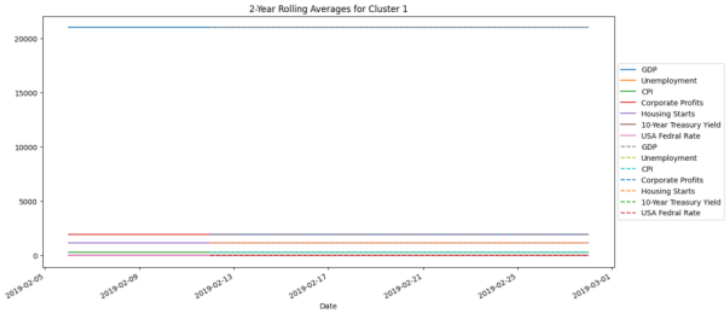
2. granger causality tests on GDP and unemployment:
    - SSR based F test and chi2 test along with likelihood ratio test and parameter F test has been employed here.
    - P is used as an indicator to whether to accept or reject the hypothesis. If p>0.5, we reject the hypothesis else, we accept it.
    - Since the value of p is less than or equal to 0.5 through all 4 tests, we accept the presumed hypothesis that “Unemployment has no effect on the GDP of a country”.
3. granger causality tests on GDP and CPI:
    - SSR based F test and chi2 test along with likelihood ratio test and parameter F test has been employed here.
    - P is used as an indicator to whether to accept or reject the hypothesis. If p>0.5, we reject the hypothesis else, we accept it.
    - Since the value of p is greater than 0.7 through all 4 tests, we reject the hypothesis that “CPI has no effect on the GDP of a country”.

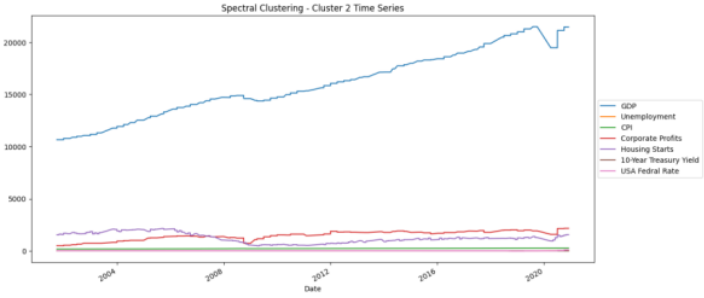
We now perform Spectral clustering on the time-series data and each cluster formed in agglomerative hierarchal clustering along with their 2-year rolling averages.

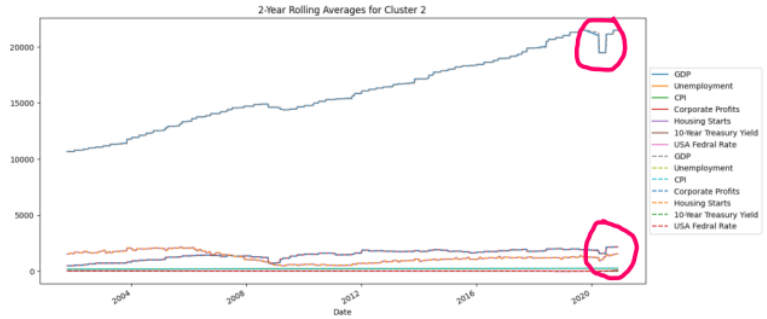


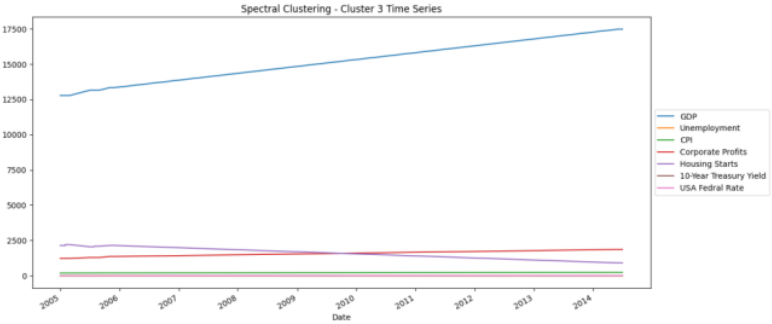

2-Year Rolling Averages for Cluster 0
20000
15000
10000
5000
0
2004
2008
2012
2016
2020
Date
GDP
Unemployment
CPI
Corporate Profits
Housing Starts
10-Year Treasury Yield
USA Fedral Rate
GDP
Unemployment
CPI
Corporate Profits
Housing Starts
10-Year Treasury Yield
USA Fedral Rate

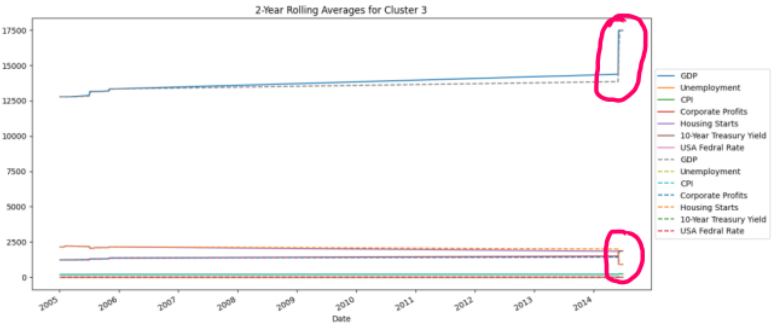

Spectral Clustering - Cluster 1 Time Series
20000
15000
10000
5000
0
2019-02-01
2019-02-05
2019-02-09
2019-02-13
2019-02-17
2019-02-21
2019-02-25
2019-03-01
Date
GDP
Unemployment
CPI
Corporate Profits
Housing Starts
10-Year Treasury Yield
USA Fedral Rate


2-Year Rolling Averages for Cluster 1
20000
15000
10000
5000
0
2019-02-05
2019-02-09
2019-02-13
2019-02-17
2019-02-21
2019-02-25
2019-03-01
Date
GDP
Unemployment
CPI
Corporate Profits
Housing Starts
10-Year Treasury Yield
USA Fedral Rate
GDP
Unemployment
CPI
Corporate Profits
Housing Starts
10-Year Treasury Yield
USA Fedral Rate


Spectral Clustering - Cluster 2 Time Series
20000
15000
10000
5000
0
2004
2008
2012
2016
2020
Date
GDP
Unemployment
CPI
Corporate Profits
Housing Starts
10-Year Treasury Yield
USA Fedral Rate







For cluster 0, the graph is fairly linear, and no significant dip has been found for the 2-year rolling average. We do not consider this cluster and move on to the second cluster. For cluster 1, the graph is still linear, and no significant dip has been found for the 2-year rolling average. We do not consider this cluster and move on to the next cluster. For cluster 2, we finally get to see a significant dip towards the end in the graph for rolling average. For cluster 3, we get to see another significant change in the graph for rolling average. From this we may conclude that the data we use is significant starting from cluster 2.

We now move on to Regression Analysis based on the findings of clustering analysis. Here, we add and combine a few features to the merged dataset to check correlation now.

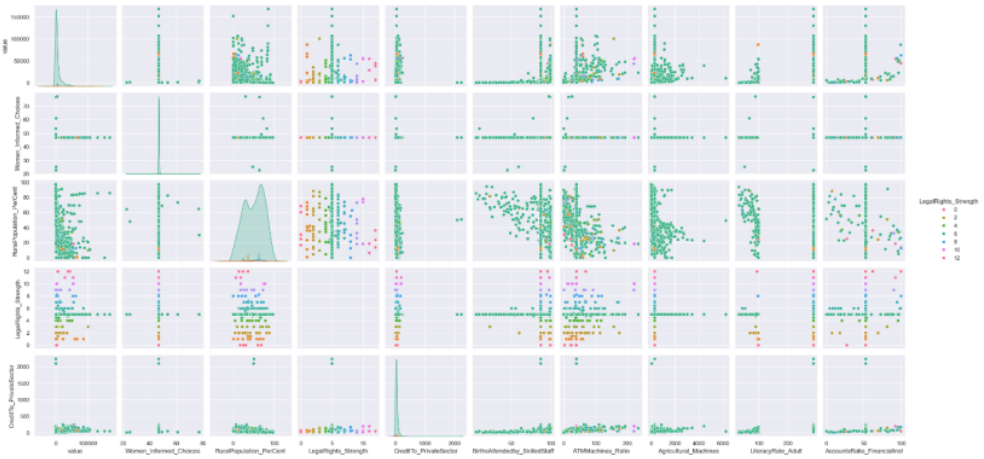
```
GDP_Combine['Women_Informed_Choices'] = Women_Informed_Choices.value
GDP_Combine['RuralPopulation_PerCent'] = RuralPopulation_PerCent.value
GDP_Combine['LegalRights_Strength'] = LegalRights_Strength.value
GDP_Combine['CreditTo_PrivateSector'] = CreditTo_PrivateSector.value
GDP_Combine['BirthsAttendedby_SkilledStaff'] = BirthsAttendedby_SkilledStaff.value
GDP_Combine['ATMMachines_Ratio'] = ATMMachines_Ratio.value
GDP_Combine['Agricultural_Machines'] = Agricultural_Machines.value
GDP_Combine['LiteracyRate_Adult'] = LiteracyRate_Adult.value
GDP_Combine['AccountsRatio_FinancialInst'] = AccountsRatio_FinancialInst.value

GDP_Combine.describe()
```

✓ 0.0s

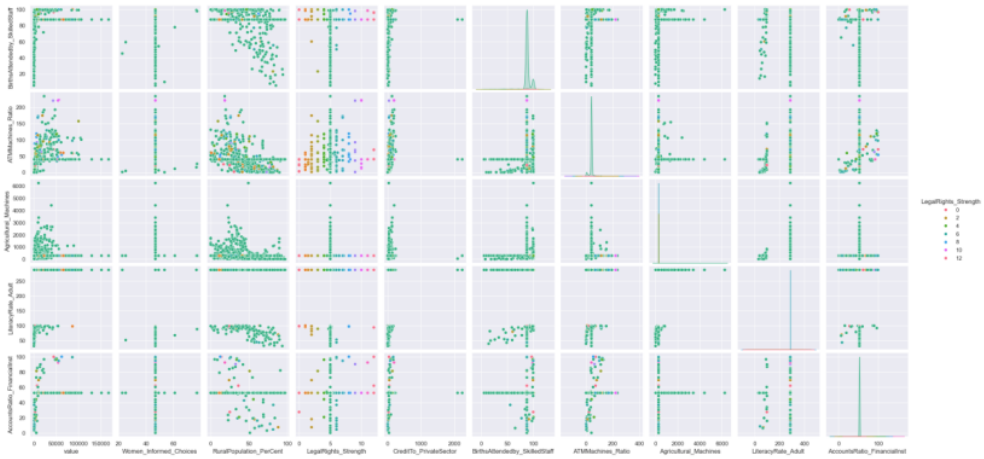
```
GDP_Combine_new=GDP_Combine
GDP_Combine_new['Literacy_creditToPriva']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['CreditTo_PrivateSector'];
GDP_Combine_new['Literacy_RuralPop']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['RuralPopulation_PerCent'];
GDP_Combine_new['Literacy_AgriMach']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['Agricultural_Machines'];
GDP_Combine_new['Literacy_AccountRa']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['AccountsRatio_FinancialInst'];
GDP_Combine_new['Literacy_ATM']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['ATMMachines_Ratio'];
GDP_Combine_new['Literacy_BirthAT']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['BirthsAttendedby_SkilledStaff'];
GDP_Combine_new['Literacy_Legal']=GDP_Combine['LiteracyRate_Adult']*GDP_Combine['LegalRights_Strength'];
GDP_Combine_new['Literacy_Woman']=GDP_Combine['LiteracyRate_Adult']*GDP_Comb
GDP_Combine_new['Woman_Rural']=GDP_Combine['Women_Informed_Choices']*GDP_Com
GDP_Combine_new['Woman_CreditToPriv']=GDP_Combine['Women_Informed_Choices']*GDP_Combine['CreditTo_PrivateSector'];
GDP_Combine_new['Woman_AgriM']=GDP_Combine['Women_Informed_Choices']*GDP_Combine['Agricultural_Machines'];
GDP_Combine_new['Woman_ATM']=GDP_Combine['Women_Informed_Choices']*GDP_Combine['ATMMachines_Ratio'];
GDP_Combine_new['Woman_BirthAT']=GDP_Combine['Women_Informed_Choices']*GDP_Combine['BirthsAttendedby_SkilledStaff'];
```

(variable) GDP_Combine: DataFrame

Now we check correlation between the attributes. Pair plot Correlation of the Attributes added are as follows:

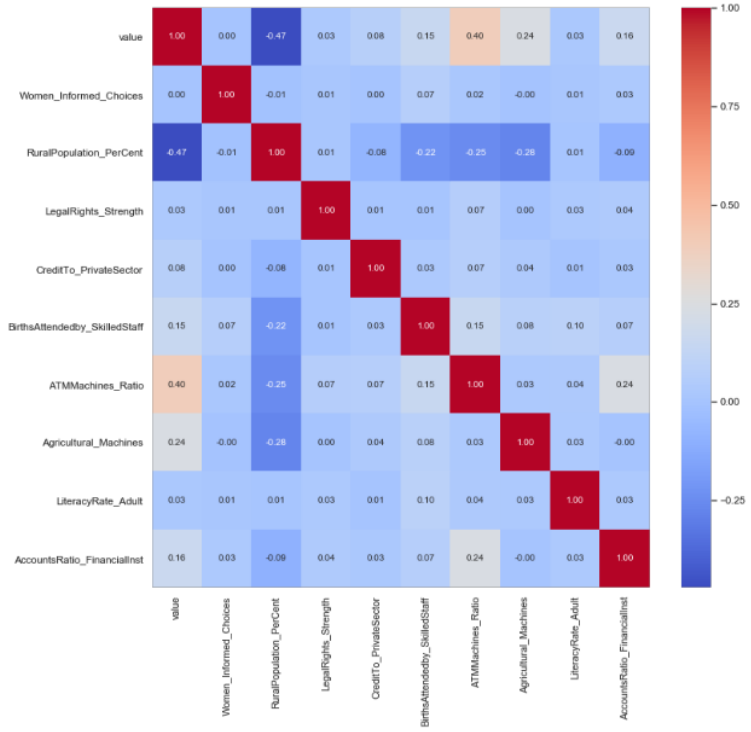


Pair plot 1

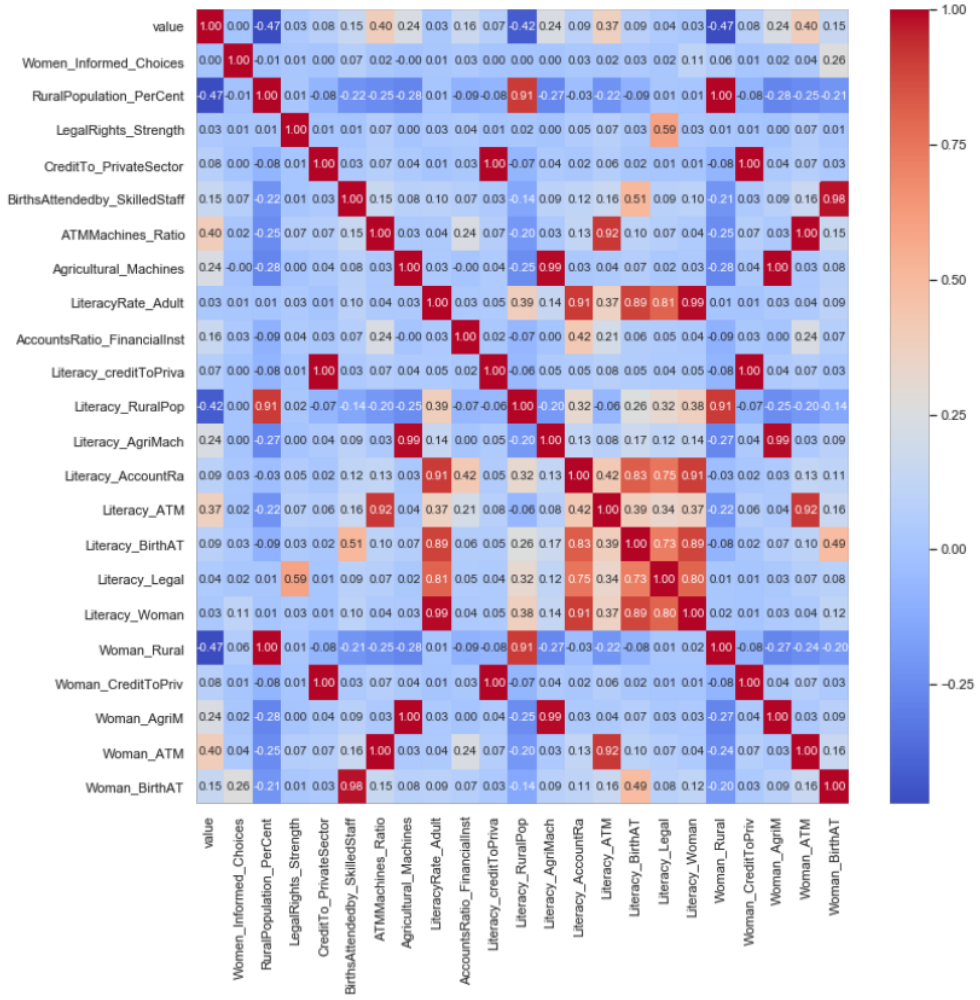


pair plot 2

Features Correlation Matrix & Combining Features are as follows:



Matrix 1



Matrix 2

The first matrix represents the correlation matrix with some new features added. The second matrix represents the actual correlation of the features. So far, there is no significant change between correlation of features with GDP value.

The data collected and processed is now split into training set as well as testing set. The seven Regression models are trained now. The seven are used to evaluate the performances:

- Multiple Linear Regression
- Polynomial Regression
    - deg=4
- Decision Tree Regression
    - max_depth = 30
- Random Forest Regression
    - max_features used = 10

    - n_estimators used = 30
- Ridge Regression
    - [alpha] = 0.0000001
    - normalize = True
- Lasso Regression
    - [alpha] = 0.0000001
    - normalize = True
- Elastic Net Regression
    - [alpha] = 0.001
- PCA
    - Explained_variances = 95%
- Ordinary Least Square Method (OLS)

Model validation methods for the models are as follows:

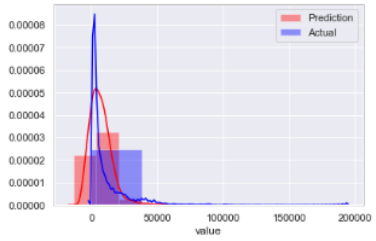
- Root Mean Squared Error: Measures the difference between values predicted by the model being tested and the values statistically observed.
- Test & Train Split: The data is divided into test & train with parameters:
    - Test_size = 0.20
    - Random_state = 40
- Cross validation: k-Fold cross validation method:
    - Number of splits of data = 10
    - Random state = None
    - Shuffle = True
- R - Squared: This represents the proportion of variance for dependent variable that's explained by an independent variable or variables in a regression model, also known as coefficient of determination.
-

MODEL TRAINING:

1. Multiple linear Regression:

Result of 10k-fold cross-validation for a Multiple Linear Regression model, evaluating its performance using negative mean squared error, and average RMSE over the 10 folds is as follows:

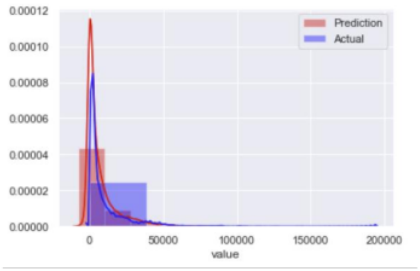
```
Linear Regression-1 MSE:189063907.76161993, STD: 29634843.82342346, RMSE: 13709.367622306865
Linear Regression-2 MSE:189100921.88687503, STD: 23066385.47197286, RMSE: 13725.078677108697
Linear Regression-3 MSE:189137128.78931373, STD: 26769490.405529916, RMSE: 13717.763289131277
Linear Regression-4 MSE:189032804.0006031, STD: 18573745.25761638, RMSE: 13732.549268254868
Linear Regression-5 MSE:188881998.55894637, STD: 23617118.188600965, RMSE: 13716.794362465163
Linear Regression-6 MSE:188957070.93601662, STD: 35930647.89835548, RMSE: 13683.422748382383
Linear Regression-7 MSE:189059032.25060222, STD: 32742095.53794235, RMSE: 13693.54605074524
Linear Regression-8 MSE:189156086.00132883, STD: 27993200.861664057, RMSE: 13716.43963925588
Linear Regression-9 MSE:189044024.4141727, STD: 23291834.81317933, RMSE: 13722.761644192822
Linear Regression-10 MSE:188910514.20184627, STD: 17906101.239903003, RMSE: 13729.967444968328
13714.769074681153
```



2. Polynomial Regression

The results of average RMSE value and negative mean-squared error after performing 10k-fold cross-validation for a Polynomial Regression model of degree 4 are as follows:

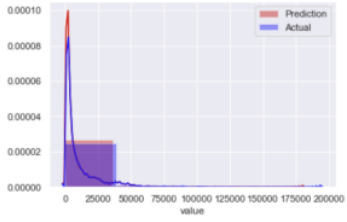
```
Degree 4-1polynomial MSE: 178829082.21622905, STD: 21446808.98547562, RMSE: 13348.045851742692
Degree 4-2polynomial MSE: 178508301.64848226, STD: 27338037.19128364, RMSE: 13319.512736816341
Degree 4-3polynomial MSE: 178453173.0153547, STD: 39910751.47157032, RMSE: 13279.232991009476
Degree 4-4polynomial MSE: 178926677.4312752, STD: 27725317.309963275, RMSE: 13335.41367682993
Degree 4-5polynomial MSE: 178531180.84071812, STD: 36230751.27753511, RMSE: 13287.356779059945
Degree 4-6polynomial MSE: 178818110.22104782, STD: 28890212.57778097, RMSE: 13325.527141428049
Degree 4-7polynomial MSE: 178403929.0341698, STD: 26402124.238841336, RMSE: 13319.324375944536
Degree 4-8polynomial MSE: 178614235.8548877, STD: 23721328.04828006, RMSE: 13335.628862474443
Degree 4-9polynomial MSE: 6534460259.760158, STD: 19067569092.605515, RMSE: 37251.02829530883
Degree 4-10polynomial MSE: 178722163.47905368, STD: 33160572.36152907, RMSE: 13311.820700328715
13714.769074681153
```



3. Decision Tree Regression

We perform 10-fold cross-validation for a Decision Tree Regressor. Within 10 iterations, the Decision Tree Regressor (dreg) is trained on the training data The average RMSE is computed for each iteration. Result of the code is as follows:

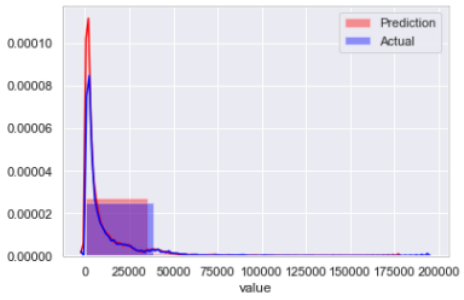
```
Decision Tree Regressor-1 MSE:236067439.8593858, STD: 26797114.37914973, RMSE: 15339.60252555681
Decision Tree Regressor-2 MSE:238392157.78007156, STD: 26855259.846822403, RMSE: 15415.733946987055
Decision Tree Regressor-3 MSE:241051044.96210617, STD: 16113932.89527418, RMSE: 15516.648865120282
Decision Tree Regressor-4 MSE:237729249.18356386, STD: 46597264.12339979, RMSE: 15345.572246145812
Decision Tree Regressor-5 MSE:241071891.83701095, STD: 30606340.562440865, RMSE: 15493.029819056212
Decision Tree Regressor-6 MSE:234208405.57693157, STD: 19947221.64251077, RMSE: 15290.333413952816
Decision Tree Regressor-7 MSE:237948022.71400633, STD: 19110002.685967308, RMSE: 15412.94166733202
Decision Tree Regressor-8 MSE:235410855.35561663, STD: 31843904.90312563, RMSE: 15307.47867419284
Decision Tree Regressor-9 MSE:238619516.75665307, STD: 13339506.27012279, RMSE: 15441.241785626975
Decision Tree Regressor-10 MSE:230672032.54431328, STD: 42160100.80829336, RMSE: 15122.282738674263
15368.486568264509
```



4. Random Forest Regression

We now perform 10-fold cross-validation for Random Forest Regression. The model's negative mean-squared error and average RMSE is computed for each iteration. Results are as follows:

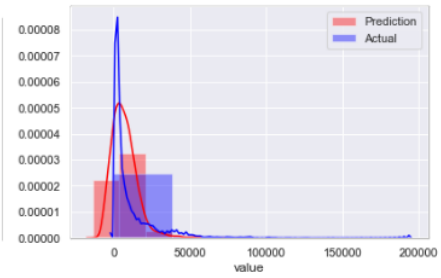
```
Random Forest Regressor-1 MSE:136817988.10562104, STD: 11465590.04122934, RMSE: 11686.58256580321
Random Forest Regressor-2 MSE:136402206.55276176, STD: 14707591.132086398, RMSE: 11662.082899250649
Random Forest Regressor-3 MSE:138107778.19044104, STD: 21525206.150575977, RMSE: 11719.415525609744
Random Forest Regressor-4 MSE:136208364.58803338, STD: 15751765.650027253, RMSE: 11651.24726654372
Random Forest Regressor-5 MSE:136217245.09501383, STD: 19866407.7849291, RMSE: 11640.36431584106
Random Forest Regressor-6 MSE:134640960.87257013, STD: 14575044.247819642, RMSE: 11585.789663335643
Random Forest Regressor-7 MSE:136386657.8623649, STD: 21511447.705672175, RMSE: 11642.133049684893
Random Forest Regressor-8 MSE:136005538.5234038, STD: 16882204.939195435, RMSE: 11638.64332367115
Random Forest Regressor-9 MSE:133682978.68564165, STD: 13940798.641548533, RMSE: 11546.318777069837
Random Forest Regressor-10 MSE:135593384.35511583, STD: 12325516.04283535, RMSE: 11632.478625070817
11640.505601188072
```



5. Ridge Regression

We conduct 10-fold cross-validation for Ridge Regression. Within 10 iterations, the Decision Tree Regressor (dreg) is trained on the training data The average RMSE value is computed for all iterations . Result of the code is as follows:

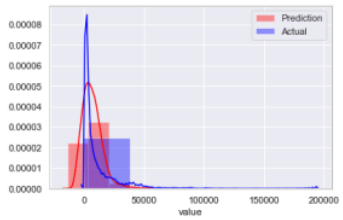
```
ElasticNet Regression-1 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-2 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-3 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-4 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-5 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-6 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-7 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-8 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-9 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
ElasticNet Regression-10 MSE:194933968.40582803, STD: 189308260.99618274, RMSE: 12424.70157640104
12424.701576401041
```



6. Lasso Regression

We conduct 10-fold cross-validation for Lasso Regression. Within 10 iterations, the Decision Tree Regressor (dreg) is trained on the training data The average RMSE value is computed for each iteration. Result of the code is as follows:

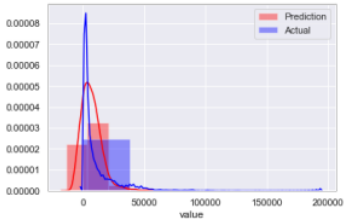
```
Lasso Regression-1 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-2 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-3 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-4 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-5 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-6 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-7 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-8 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-9 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
Lasso Regression-10 MSE:194933968.41059196, STD: 189308260.9970274, RMSE: 12424.701576607877
12424.701576607878
```



7. Elastic Net Regression

We conduct 10-fold cross-validation for Elastic Net Regression. Within 10 iterations, the Decision Tree Regressor (dreg) is trained on the training data The average RMSE value is computed for each iteration. Result of the code is as follows:

```
Ridge Regression-1 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-2 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-3 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-4 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-5 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-6 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-7 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-8 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-9 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
Ridge Regression-10 MSE:194933968.22434145, STD: 189308261.9489679, RMSE: 12424.701550124008
12424.701550124008
```



MODEL EVALUATION:

Listing down all the parameters and evaluation metric for the seven models, we get the result as shown below:

```
perf = pd.DataFrame({"Model": Models})
perf["Train.RMSE"] = train_rmses.values()
perf["Test.RMSE"] = test_rmses.values()
perf['Kfold'] = Kfold.values()
perf['RSquare_train'] = RSquare_train.values()
perf['RSquare_test'] = RSquare_test.values()
perf
```

	Model	Train.RMSE	Test.RMSE	Kfold	RSquare_train	RSquare_test
0	Multiple Linear Regression	11098.558146	11357.502081	13714.769075	0.377350	0.391068
1	Polynomial Regression	9165.460690	161698.415494	15711.289141	0.575362	0.550000
2	Decision Tree Regression	2255.186797	6300.789568	15368.486568	0.974292	0.812590
3	Random Forest Regression	2902.347557	5441.296505	11610.635767	0.957420	0.860232
4	Ridge Regression	11098.558146	11357.502088	12424.701550	0.377350	0.391068
5	Lasso Regression	11098.558146	11357.502081	12424.701577	0.377350	0.391068
6	Elastic Net Regression	11098.558146	11357.503028	12424.701576	0.377350	0.391068

From the results, it is clear that RMSE is the least for [Decision Tree Regression] and [Random Forest Regression] model. Also, R-squared test value is also the highest for these two models.

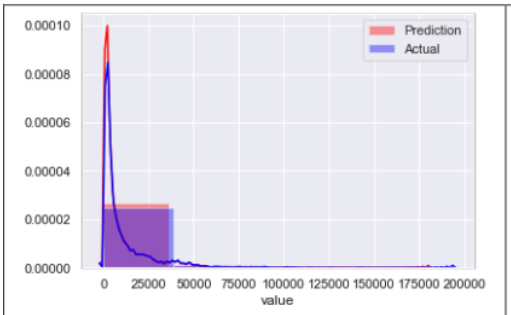


Fig. 1-decision tree training set graph

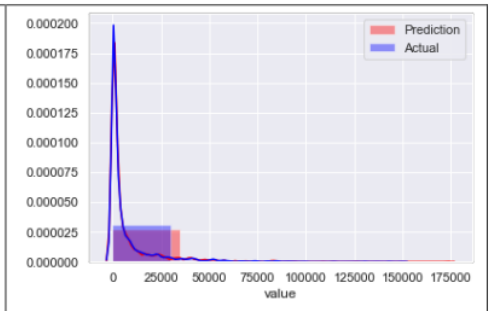
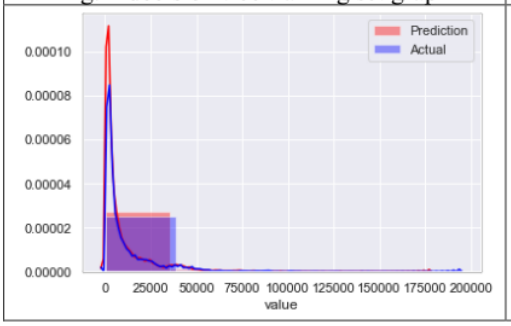


Fig. 2- decision tree testing set graph



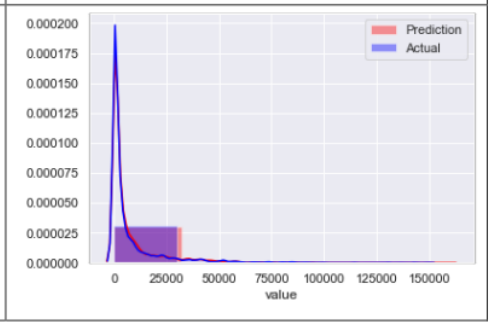


Fig. 3- Random forest training set graph	Fig. 4- Random forest testing set graph

The model is now fully trained and can now be used for predicting GDP of any country.

Sample 1- Predicting GDP for India:

```
#Country = India
X_sample = GDP_Combine_X
X_sample['Year'] = 1960
X_sample['Women_Informed_Choices'] = 46.753333
X_sample['RuralPopulation_PerCent'] = 82.076000
X_sample['LegalRights_Strength'] = 5.01029
X_sample['CreditTo_PrivateSector'] = 7.949170
X_sample['BirthsAttendedby_SkilledStaff'] = 87.418299
X_sample['ATMMachines_Ratio'] = 40.080656
X_sample['Agricultural_Machines'] = 2868.92997
X_sample['LiteracyRate_Adult'] = 28699.92997
X_sample['AccountsRatio_FinancialInst'] = 521.493345
```

```
Rfreg_y_pred_sample = Rfreg.predict(X_sample)
print('The predicted GDP valus is:',Rfreg_y_pred_sample.max())
```

```
The predicted GDP valus is: 9784.033532572013
```

Sample 2- Predicting GDP for USA:

```
#Country = United States
X_sample2 = GDP_Combine_X
X_sample2['Year'] = 2016
X_sample2['Women_Informed_Choices'] = 46.753333
X_sample2['RuralPopulation_PerCent'] = 18.212000
X_sample2['LegalRights_Strength'] = 11.000000
X_sample2['CreditTo_PrivateSector'] = 192.165500
X_sample2['BirthsAttendedby_SkilledStaff'] = 87.418299
X_sample2['ATMMachines_Ratio'] = 40.080656
X_sample2['Agricultural_Machines'] = 286.92997
X_sample2['LiteracyRate_Adult'] = 286.92997
X_sample2['AccountsRatio_FinancialInst'] = 52.493345
```

```
Rfreg_y_pred_sample2 = Rfreg.predict(X_sample2)
print('The predcited GDP value is :',Rfreg_y_pred_sample2.max())
```

```
The predcited GDP value is : 57414.72022395733
```

CONCLUSION:

- We performed Data gathering and Data preprocessing for consistent and accurate data.
- We explored and performed 3 different types of clustering analysis for selecting with tables and features to use in order to provide results as accurate as possible.
- We performed Granger causality tests to test our hypothesis and bias against the target value (GDP).
- We explored and performed 7 different Regression and compared their performances with each other to get the 2 best fitted models (that is, Random Forest model and decision tree model).
- We evaluated, trained, and tested our model using various validation techniques and tested the accuracy of our model by picking random samples from the data set (India and USA) and predicted their respective GDP values.

REFERENCES

- Yuan Yuan et al., "Urbanization's effects on the urban-rural income gap in China: A meta-regression analysis", Land Use Policy Volume 99, ScienceDirect, 2020
- Ramen Pal et al. "Portfolio formation and optimization with continuous realignment: A suggested method for choosing the best portfolio of stocks using variable length NSGA-II", Expert Systems with Applications
- Volume 186, Elsevier, 2021
- Petr Chunaev et al., "Community detection in node-attributed social networks: A survey", Computer Science Review Volume 37, Elsevier, 2020
- Guoqing Chao et al., "A Survey on Multiview Clustering", IEEE Transactions on Artificial Intelligence Volume 2, IEEE,2021
- Waichon Lio et al., "Uncertain maximum likelihood estimation with application to uncertain regression analysis", Soft Computing, Springer Link, 2020
- Kristina P. Sinaga et al., "Unsupervised K-Means Clustering Algorithm", IEEE Access, 2020
- Lucy Eunju Lee et al., "Antineutrophil cytoplasmic antibody-associated vasculitis classification by cluster analysis based on clinical phenotypes: a single-center retrospective cohort study", Clinical Rheumatology, Springer Link, 2023
- Meshal Shutaywi et al., "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering",Entropy, MDPI, 2021
- Chunhui Yuan et al., "K-Value Selection Method of K-Means Clustering Algorithm", J, MDPI, 2019.
- D.Q.F. de Menezes et al., "A review on robust M-estimators for regression analysis", Computers & Chemical Engineering Volume 147, Elsevier, 2021.
- https://www.kaggle.com/
- https://data.worldbank.org/

Understanding Relationships between Economic Indicators and Financial Markets

ORIGINALITY REPORT

6%	4%	4%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	www.coursehero.com Internet Source	1%
3	Bikash Sadhukhan, Somenath Mukherjee, Shounak Banerjee, Raj Kumar Samanta. "Multifractal, nonlinear, and chaotic nature of earthquake and global temperature", Arabian Journal of Geosciences, 2021 Publication	1%
4	Meshal Shutaywi, Nezamoddin N. Kachouie. "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering", Entropy, 2021 Publication	<1%
5	link.springer.com Internet Source	<1%
6	dokumen.pub Internet Source	<1%

7	easy-forex-review.com Internet Source	<1 %
8	Guoqing Chao, Shiliang Sun, Jinbo Bi. "A Survey on Multiview Clustering", IEEE Transactions on Artificial Intelligence, 2021 Publication	<1 %
9	Zhen Zhou, Changbin Zhao, Bolin Cai, Manting Ma, Shaofen Kong, Jing Zhang, Xiquan Zhang, Qinghua Nie. "Myogenic differentiation potential of chicken mesenchymal stem cells from bone marrow", Research Square Platform LLC, 2021 Publication	<1 %
10	ci.desoto.tx.us Internet Source	<1 %
11	medworm.com Internet Source	<1 %
12	"On the Move to Meaningful Internet Systems: OTM 2019 Conferences", Springer Science and Business Media LLC, 2019 Publication	<1 %
13	Lucy Eunju Lee, Jung Yoon Pyo, Sung Soo Ahn, Jason Jungsik Song, Yong-Beom Park, Sang-Won Lee. "Antineutrophil cytoplasmic antibody-associated vasculitis classification by cluster analysis based on clinical phenotypes:	<1 %

a single-center retrospective cohort study", Clinical Rheumatology, 2023
Publication

14	ebin.pub Internet Source	<1%
15	Biao Sun, Chuanglin Fang, Xia Liao, Menghang Liu, Zhitao Liu, Xiaomin Guo. "Revealing the heterogeneous effects of new urbanization on urban-rural inequality using geographically weighted quantile regression", Applied Geography, 2023 Publication	<1%
16	youngpioneers.manipal.edu Internet Source	<1%
17	Kittisak Kerdprasop, Nittaya Kerdprasop, Pairote Sattayatham. "Chapter 48 Weighted K-Means for Density-Biased Clustering", Springer Science and Business Media LLC, 2005 Publication	<1%
18	Yuan Yuan, Mingshu Wang, Yi Zhu, Xianjin Huang, Xuefeng Xiong. "Urbanization's effects on the urban-rural income gap in China: A meta-regression analysis", Land Use Policy, 2020 Publication	<1%
19	dspace.plymouth.ac.uk Internet Source	

<1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		