Report on Timing Entries into Viable Investments

By Yin Kwong John Lee

1 Motivation

After identifying viable investment opportunities, the next challenge is to time our entries into these investments. However, one may conjecture that since the value of these investments always go up according to my selection scheme, wouldn't we profit regardless of when we strike? The answer is yes, but you probably wouldn't make as much as a well-timed strike. Let us use one of the viable stocks I've selected as an example, 3M, an industrial conglomerate with an intrinsic value of $200. Intrinsic value is the true value of a stock and is guaranteed to be attained overtime once the stock is selected using my method in project 1. Assume we close our eyes and buy this stock at $177 at some random time. There are two major outcomes. First, the stock can go straight up to say $200 overtime after it's bought, or second, the stock could go to $150 shortly after it is bought before bouncing up to reach $200. For the first outcome, you will make $200-$177=$23 right away, but for the second outcome, you will need to lose $177-150=$27 and wait for a while before making $23. You could have invested the $27 and time in some instruments with a more faithful return such as treasury bonds or spent it on a more thoughtful venue for a date night. Either way, you are doing something with the money rather than let it sit there to wait for a rebounce.

I will start by examining the severity of missing data and label imbalance issue, then move on to implement dimension reduction techniques, specifically Principal Component Analysis (PCA) to select the optimal number of components that captures the data variance and Exploratory Factor Analysis (EFA) to inspect possible latent variables that underlie the seeming relationships between any pair of variables. After that, I will use K-means Clustering mainly coupled with the Elbow and Silhouette methods to visualize the selected principal components. The results will be used to answer the following questions:

1) How do the indicators relate to each other under the real-world economic phenomena?
2) What are some important economic features that best represent the data?
3) How many economic groups are best represented by the data? (e.g. income class status, education level, age groups)

The exact timing of entries into the selected viable investments cannot be determined until models are fitted to predict the stock market's direction- uptrend or downtrend. The ideal entry point would be at the immediate transition from a downtrend to an uptrend. This project is primarily purposed to examine the importance of the economic indicators I've chosen and potential populations of concern, so that proper assumptions can be made when assessing statistical models.

2 Data Source

Realization for the following variables are collected via the Federal Reserve Economic Data website. All data is in csv format. The data of 13 economic indicators from 2000-09-01 to 2019-09-01 are gathered, each of which concerns the inevitable economic fate- inflation. In particular, according to a renowned financial author, Stephen Leeb, oil price is a significant indicator for the economic climate, in other words, how "hot" the economy is getting. The variables are concatenated by their index, date, using the pandas package in python. The final data set can be found under the file path "./data/fiscal_data.csv".

Variable	**# of Observations**	**Description**	**Source**
crude_oil_price (daily)	5218	Crude oil prices in the U.S.	https://fred.stlouisfed.org/series/DCOILWTICO
gasoline_price_index (monthly)	241	Consumer price index for urban gasoline consumption in the U.S.	https://fred.stlouisfed.org/series/CUUR0000SETB01
gold_fixing_price (daily)	5218	Gold price fixed in the London Bullion Market Association (LBMA)- in U.S. dollar	https://fred.stlouisfed.org/series/GOLDAMGBD228NLBM
usd_index (daily)	5218	A weighted average of the foreign exchange value of the U.S. dollar against the currencies of a broad group of major U.S. trading partners	https://fred.stlouisfed.org/series/TWEXB
energy_growth_rate (monthly)	241	Growth rate of the consumer price index for energy, electricity, and gasoline related goods from the previous year in the U.S.	https://fred.stlouisfed.org/series/CPGREN01USM657N
productivity_growth_rate (quarterly)	81	Growth rate of real output for the U.S. business sectors	https://fred.stlouisfed.org/series/PRS85006041
consumer_price_index (monthly)	241	Index that measures the average monthly change in the price for goods and services paid by urban consumers between any two time periods in the U.S.	https://fred.stlouisfed.org/series/CPIAUCSL
federal_funds_rate (daily)	7306	Interest rate set by the U.S. Federal Reserve to regulate the economy	https://fred.stlouisfed.org/series/EFFR
house_price_index (quarterly)	81	U.S. overall housing price index estimated using sales prices and appraisal data	https://fred.stlouisfed.org/series/USSTHPI

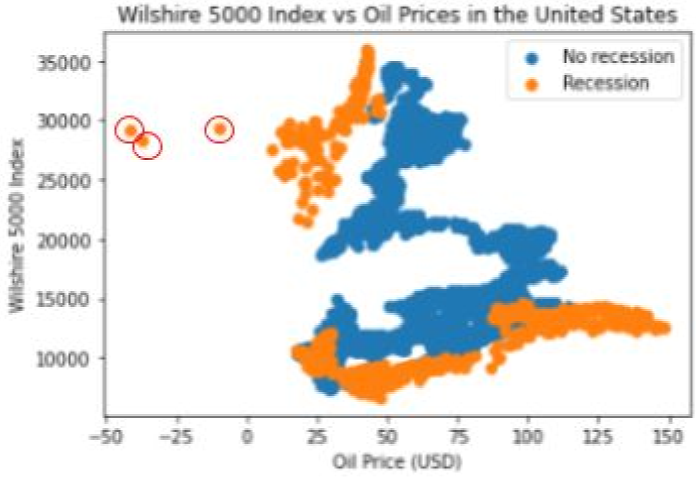
fixed_mortgage_rate (weekly)	1045	30-year fixed rate mortgage average in the U.S.	https://fred.stlouisfed.org/series/MORTGAGE30US
debt_service_percent (quarterly)	81	Household debt as a percentage of disposable personal income in the U.S.	https://fred.stlouisfed.org/series/TDSP
wilshire5000 (daily)	5218	Market capitalization-weighted index composed of 3,451 publicly traded companies in the U.S. This is a good reference for the overall stock market price	https://fred.stlouisfed.org/series/WILL5000INDFC
recession (daily)	7306	Binary symbol for recession status in the U.S.: 0=no recession, 1= recession	https://fred.stlouisfed.org/series/USRECD

3 Methods

3.1 Missing Data/ Frequency Adjustment

Since each variable is observed on its own frequency basis. For example, crude oil prices are measured daily as gasoline price index is measured monthly. These measures will be standardized to quote their daily performance using cubic spline interpolation, a third-order polynomial used to fill the missing gap between any interval. Meanwhile, the original data of each variable may consist of missing values; cubic spline interpolation will include this case and make sure every day from 2000-09-01 to 2019-09-01 is populated with an instance. Now that the data of each variable is converted to be presented on a daily basis, there are a total of 7306 observations .

3.2 Label Imbalance Issue



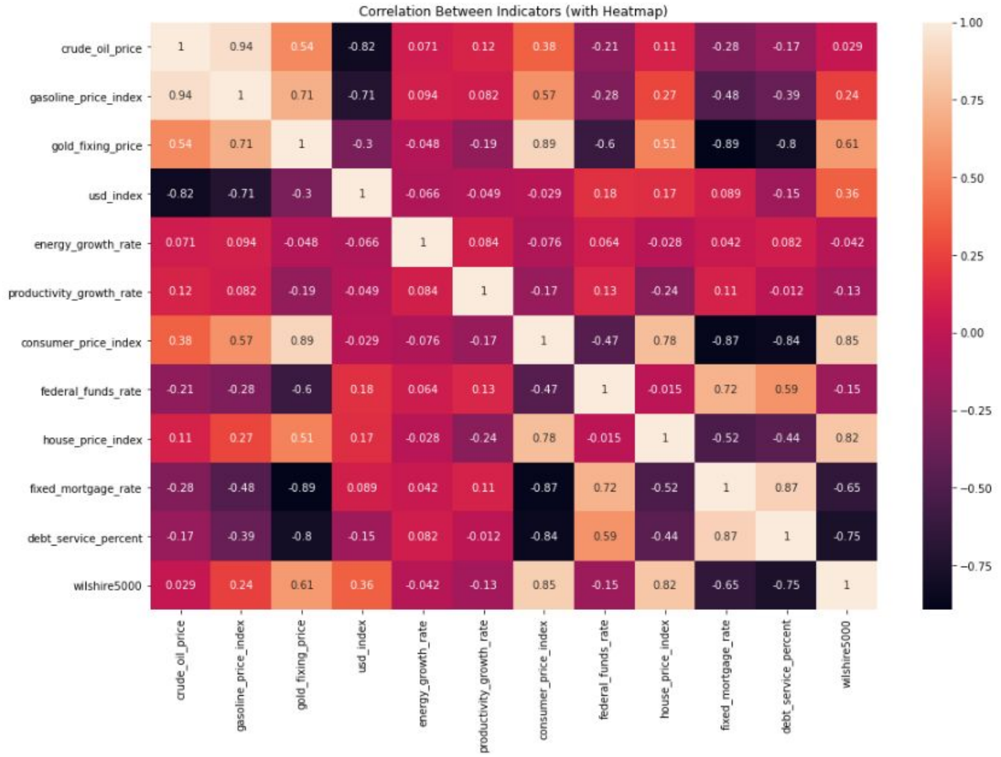
A scatter plot is used to determine the relationship between oil prices and the Wilshire 5000 Index. There seems to be near to no direct correlation between the two variables. In addition, there may be 3 potential outliers (circled in red) as the oil price turned negative in 2019. The outliers can be truncated by replacing them with the closest reasonable value, $0 in this case for the purpose of model fitting. What caught my eye is the imbalanced distribution of recession and non-recession labels in the data. According to the data, the economy is classified to be in recession only 13.359% of the time throughout the 20 year period; the rest is classified as a healthy economy with no recession. This would result in a biased model because the training dataset will only have a 13.359% chance of accounting for recession if random sampling is assumed. A leeway around this is to upsample the minority (recession) data or down sample the majority (non-recession) data. This way the model can fairly choose between the recession class and the non-recession class during the training stage. Since I will emphasize more on exploratory data analysis for this project (and not model fitting), I will avoid the recession variable when performing dimension reduction and clustering. This section presumes my concern when entering the model fitting stage in the future.

3.3 Standardization

Since the variables are measured in different units, for instance, oil price is measured in a USD per barrel basis and federal funds rate is measured in percentage basis, they may note contribute to our analysis equally, hence creating a possible bias. To cope with this, we scale the values of each variable so that they are centered (mean) at 0 with spreads (standard deviations) of 1. *sklearn.preprocessing.scale()* from the sklearn package in python is used to transform these values.

4 Analysis and Results

4.1 Question 1: How do the indicators relate to each other under the real-world economic phenomena?

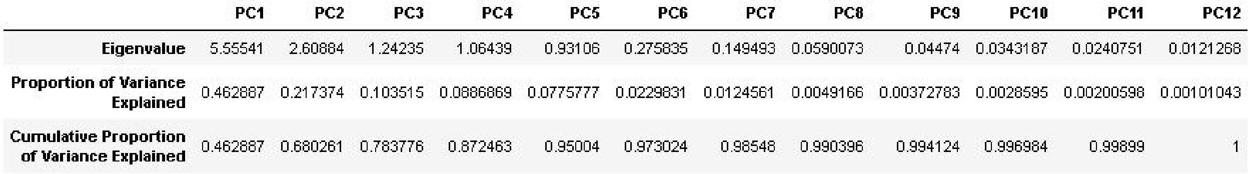


One clarification for the above economic variables is that they are merely indicators used to support relationships underlying the modern economic theory. By no means am I suggesting that they are fine predictors for economic climates. According to Stephen Leeb, a renewable energy and climate change theorist, the decrease in oil prices will speed up energy consumption majorly because a myriad of industrial and technology companies depend on it; this will in turn heat up the economy, thus applying inflationary pressure and bumping up consumer prices. Let us try to trace such a relationship in Leeb's proposition. According to the correlation heatmap, from a consumption point of view, the price of gasoline has a near perfect positive correlation with the oil price at 0.94. This is not surprising since gasoline is a by-product of crude oil. At the same time, as gasoline price increases, gold price rises the same degree as the U.S. dollar drops (0.71 and -0.71 in correlation, respectively). This makes sense because people will be holding more gold as a hedge when the purchasing power of the U.S. dollar weakens, which drives up the gold price. In addition, the slight inflationary pressure will drive prices of goods and services, hence the Consumer Price Index (CPI), up. This can be seen through the gold price's 0.89 correlation with CPI. From a real estate point of view, the fixed mortgage rate has relatively strong negative correlation with gold price and CPI, -089 and -0.87 , respectively; this is because low mortgage rates will stimulate the housing market, thus encouraging people to own homes as hard assets. Stronger demand in houses will further home prices, thus adding to the current inflationary pressure. Another evidence for this is the debt service payment's strong negative correlation with gold price and CPI, -0.8 and -0.84, respectively; lower mortgage rate allows tenants to have opportunities to spend more on commodities, leading the price hike for goods and services in consumer-based markets. Following this, it is not surprising that there is a fairly strong positive correlation of 0.85 between CPI and the Wilshire 5000 Index. Therefore, oil price and fixed mortgage rate are two major roots for explaining the economic climates.

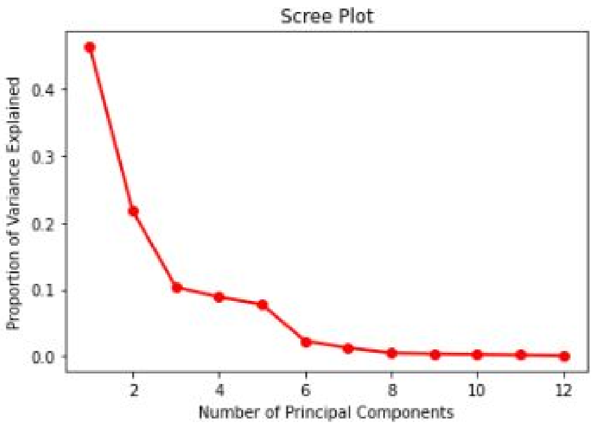
4.2 Question 2: What are some important economic features that best represent the data?

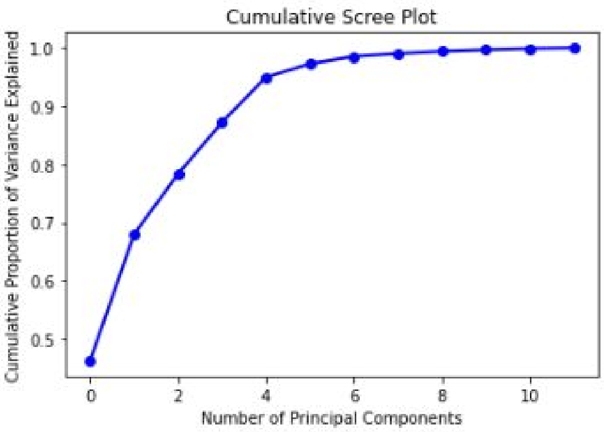
4.2.1 Eigenvalues, Scree Plot, and Cumulative Scree Plot

In this section, I seek a Principal Component Analysis (PCA) approach to choose the appropriate number of principal components to reduce the dimension of the raw data set, as well as to identify some of the most important features in each of these principal components.



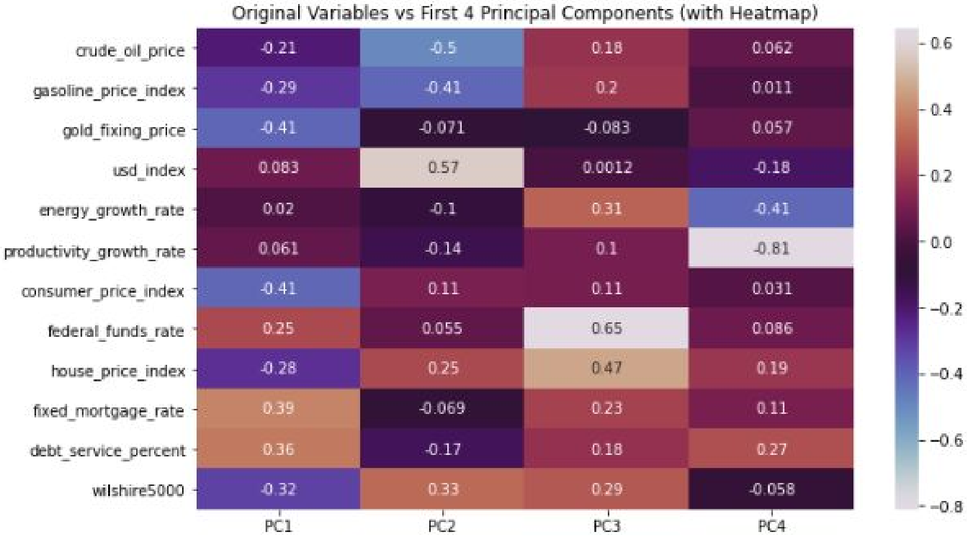
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Eigenvalue	5.55541	2.60884	1.24235	1.06439	0.93106	0.275835	0.149493	0.0590073	0.04474	0.0343187	0.0240751	0.0121268
Proportion of Variance Explained	0.462887	0.217374	0.103515	0.0886869	0.0775777	0.0229831	0.0124561	0.0049166	0.00372783	0.0028595	0.00200598	0.00101043
Cumulative Proportion of Variance Explained	0.462887	0.680261	0.783776	0.872463	0.95004	0.973024	0.98548	0.990396	0.994124	0.996984	0.99899	1



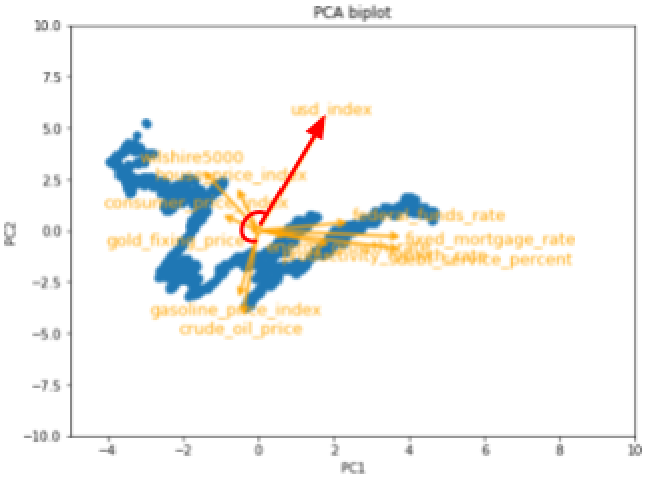


I start by finding the first 12 principal components using the *sklearn.decomposition* module in python. Eigenvalues for each principal component is obtained. Eigenvalues are coefficients that give the amount of variance carried in each principal component. I would select the first 4 principal components, PC1, PC2, PC3, and PC4 based on the Kaiser criterion, which states that principal components with eigenvalues larger than 1 carry relatively more information. The proportions (including cumulative proportions) of variance explained by each principal component is also listed in the table, and are displayed in the fashion of scree plot and cumulative scree plot. From the scree plot, the first, second, third and fourth principal component explains 46.29%, 21.74%, 10.35%, and 8.87% variance of the data, respectively. An "elbow" shape is starting to form after the 4th principal component, this is because the later principal components start adding less significant information. From the cumulative scree plot, the first four principal components capture a total of 87.25% variance in the data. The remaining principal components do not add much information to the data because there seems to be a slight plateau in the plot after accounting for the first four principal components, meaning that information accumulated from the fifth component on are not as important. At this point, it is safe to keep the first four principal components and discard the rest for our analysis.

4.2.2 Loadings

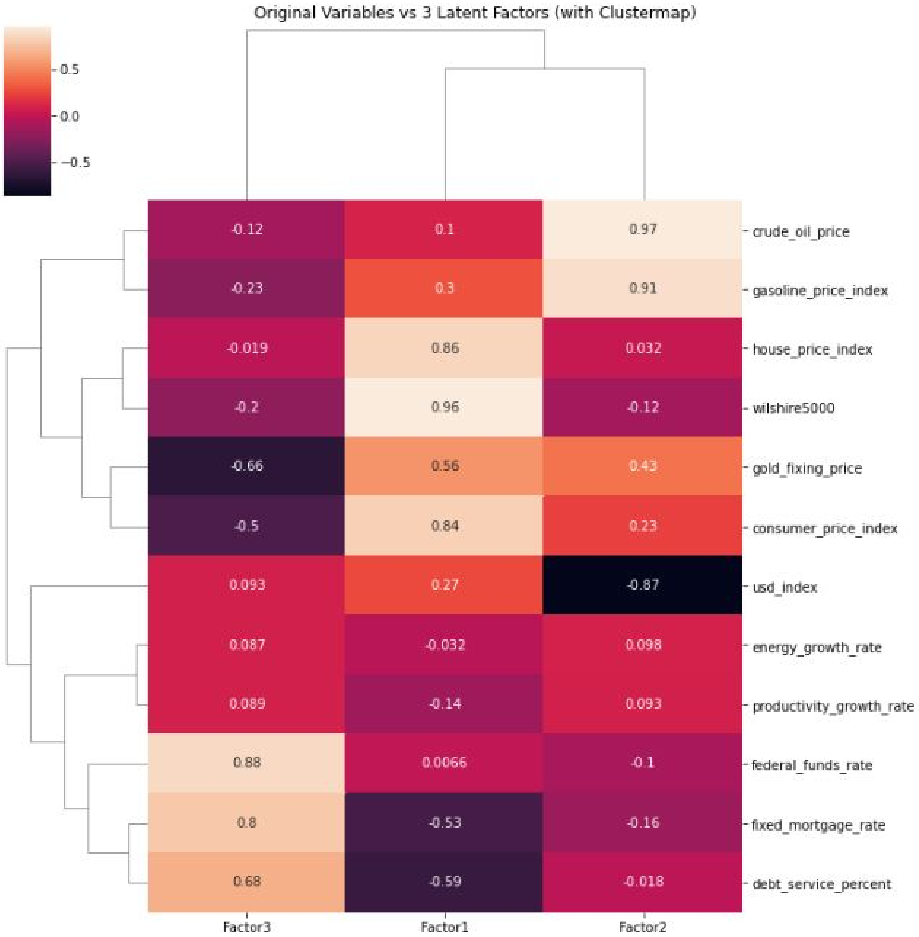


Now that we know the appropriate number of principal components to keep, it is of interest to study the loadings, in other words, importance of each economic feature in the selected principal components. I plot a correlation heatmap here to exhibit the direction and magnitude of coefficient to each economic feature. Direction is determined by the sign of a coefficient, whereas magnitude is determined by the absolute value of a coefficient- the larger the more important its corresponding feature. For instance, *gold_fixing_price* and *consumer_price_index* are two most important measures in PC1, with correlations of -0.41; *usd_index* is the most important measure in PC2, with the correlation of 0.57; *federal_funds_rate* is the most important measure in PC3, with the correlation of 0.65; *productivity_gowth_rate* is the most important measure in PC4, with the correlation of -0.81. These relationships may also be represented in a two-dimensional biplot below.



Take the first two principal components, PC1 and PC2, as an example. On top of a normal scatter plot, loadings between these principle components are plotted. This shows how strongly each economic characteristic influences a principal component. For example, since the vector (the red arrow) for *usd_index* stretches the furthest from the origin on the PC2 axis, this confirms that *usd_index* in fact influences the second principal component the most. Also, the angle formed by the vectors for any two variables reflects their actual pairwise correlation. For example, the large angle (the red angel) between the vectors for *crude_oil_price* and *usd_index* resembles the large -0.81 correlation between these variables. I shall note that vectors that diverge to form an angle more than 180° are negatively correlated.

4.2.3 Exploratory Factor Analysis (EFA)



A correlation matrix only infers the direct relationships between any two variables. It may be easy for a well-trained economist to explain such relationships, but difficult for non-professionals to interpret. For example, it may be challenging for one to understand the direct positive relationship between interest rates and treasury bond yields, when underlying factors are unknown. In disclosure to macro-economic theory, as interest rate goes up, people tend to shift their assets from treasury bonds to banks to seek higher interest returns; this pushes the demand, hence the prices of treasury bonds downwards, which to account for this, bond yield has to be lowered to align with the discounted bond prices. In this case, price and demand for treasury bonds may be two underlying factors in explaining the interest rate-bond yield relationship. By no means am I inferring that this relationship is true in every case, especially under variant economic climates. Yet again, I shall hereby issue the disclaimer that the stated underlying factors, price and demand for treasury bonds, are only two of the many factors in the whole universe.

The clustermap designates 3 underlying factors and their corresponding weights to each of the economic variables. One noticeable relationship is Factor 1's strong positive correlation with *wilshire5000*. There is an extremely broad range of factors that may assort to such a relationship because almost anything in the economy can be positively associated with the stock market- emotions, company's earnings, and litigation issues to name a few. I would speculate a bit and assume that this factor is the employment rate during a fiscal period because as more people are put to work, the market would be more optimistic in general. Another noticeable relationship is Factor 2's strong positive correlation with *crude_oil_price*. This factor may be the economic tension between the U.S. and the Middle East countries, as friction between the two may limit oil export from the rich grounds in the Middle East. It is also interesting to know that the form of economies ruled by monarchies in regions like Saudi Arabia, due their reserved ways of living, are less likely to offer good deals in oil trades, thus making oil prices hard and inflexible . One more noticeable relationship is Factor 3's strong positive correlation with *federal_funds_rate*. This factor may be the year-round inflation rate, as the The Federal Reserve always tries to curb inflation by increasing the cost of borrowing. While every factor helps explain a certain variance in observer variables, not all of them are easily measurable. Here are a few examples: economic tension, cultural values, social responsibility, etc.

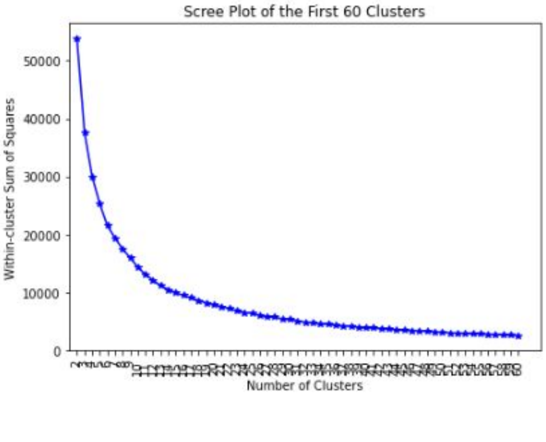
4.3 Question 3: How many economic groups are best represented by the data? (e.g. income class status, education level, age groups)

4.3.1 Choose Number of Clusters using the Rule of Thumb

Using the rule of thumb, the number of clusters can be approximated by $\sqrt{n/2} = \sqrt{7306/2} = 60.44$, where n is the total number of observations. We will round down to the nearest whole number and estimate that the observations can be grouped into 60 clusters.

4.3.2 Choose Number of Clusters using Scree Plot and Elbow Method

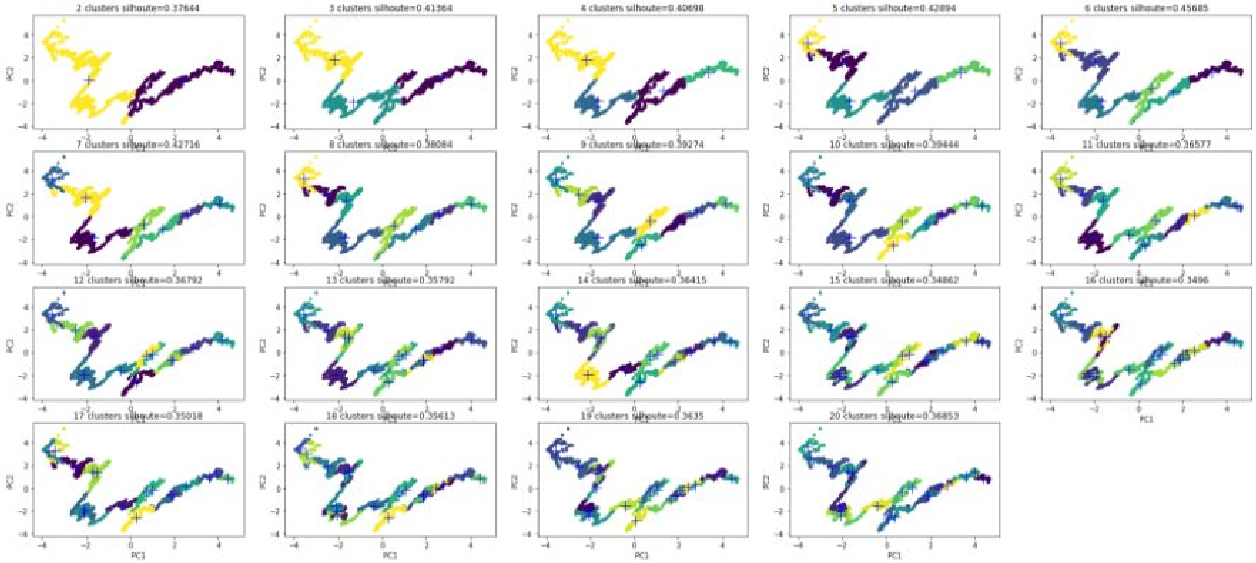
A scree plot is used to track the within-cluster sum of squares from 2 to 60 clusters (displayed in the frequency of 10 clusters). Within-cluster sum of squares indicates the variability of observations within each cluster. My goal here is to find the number of clusters, after which the data is grouped, that minimizes the dispersion in each group.



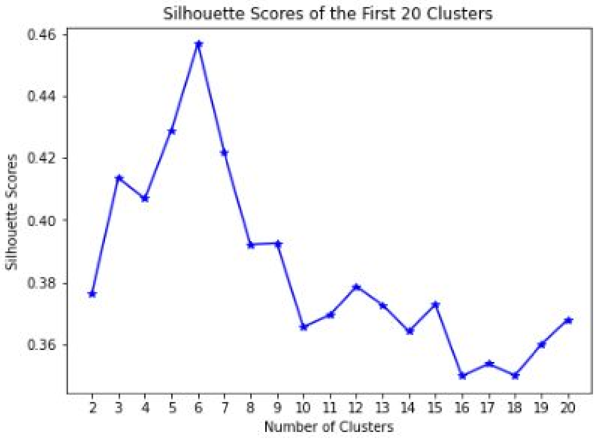
The within-cluster variability peaks when there are only two clusters and starts to flatten at around 20 clusters- again, seen by the "elbow" shape in the scree plot. This means that assigning more than 20 groups to the observations would have little to no significant improvement on the clustering quality. Moreover, parsimony is achieved with the minimum number of clusters and within-cluster variability. It is thus safe to assume 20 groupings of the observations to be optimal.

4.3.3 Choose Number of Clusters using the Silhouette Method

I choose the final optimal number of clusters using the silhouette score (silhouette coefficient) metric. Silhouette score is used to compare the goodness of the K-means clustering technique, in particular, by with the mean distance to the other instances in the same cluster and that to the nearest instances of the other cluster.



Visualizations of the clustered groups, from 2 groups to 20 groups are displayed above, with their associate silhouette scores computed in the title bars. The cross, "+", in the plots represents the centroids, the center point of each cluster. It seems that 6 is the optimal number of clusters to use because it yields the highest silhouette score of 0.45685, which means that it is the likeliest way to separate datums into their correct clusters. This can also be visualized by plotting the silhouette scores by the number of clusters in the trend plot below.



Silhouette scores seem to have an overall uptrend from 0.37644 in a 2-cluster separation to the peak of 0.45685 in a 6-cluster separation, then a sharp drop to 0.39444 in a 10-cluster separation, followed by some directionless fluctuation through the 20-cluster separation. There is strong evidence that 6 clusters is an optimal choice to the economic data. Therefore, the whole data set best represents 6 population groups in the economy. This result can be extended to the inspecting homogeneity of the data. For example, if data for the 6 economic phases- boom, peak, recession, trough, depression, and recovery, were collected, in the case that all observations with the same economic phase label are in the same cluster, homogeneity is present, and assumptions for the models shall be changed.

5 Conclusion

Based on the correlation heatmap, we obtained a preliminary understanding as to how the economic indicators relate to each other under the real-world economic phenomena. For example oil price is strongly, negatively correlated with the value of the U.S. dollar because people rather hang on to alternatives such as gold rather than dollars when they experience inflationary pressure. From PCA, the only the first four principal components can best represent the entire data, specifically, gold price and consumer price index are the two most important features from the first principal component. As far as population groups are concerned, optimally 6 clusters can be formed- each including different social-economic groups, to capture most part of the economic data. There are challenges in dealing with imbalanced labels, in particular recession status of the economy, where the majority of the data is labelled as non-recession. I have proposed to upsample or downsample the corresponding labels to resolve this issue, but haven't gotten to do so for the convenience of this study.