**IS-733 Data Mining**

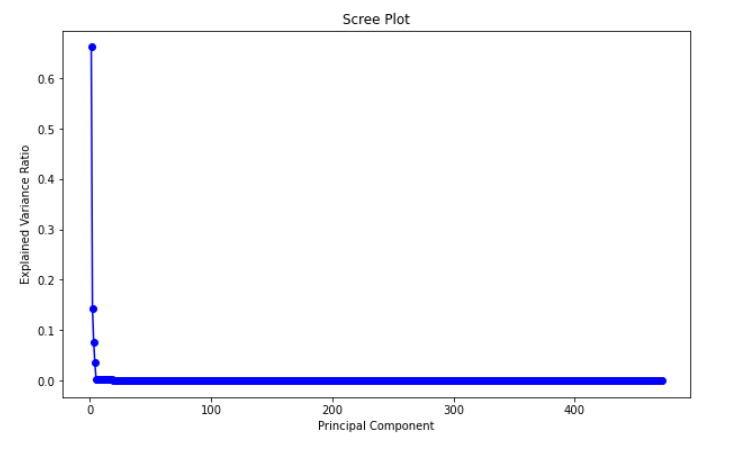
**Homework-3**

**Problem-1**

**A). PCA model to log returns**

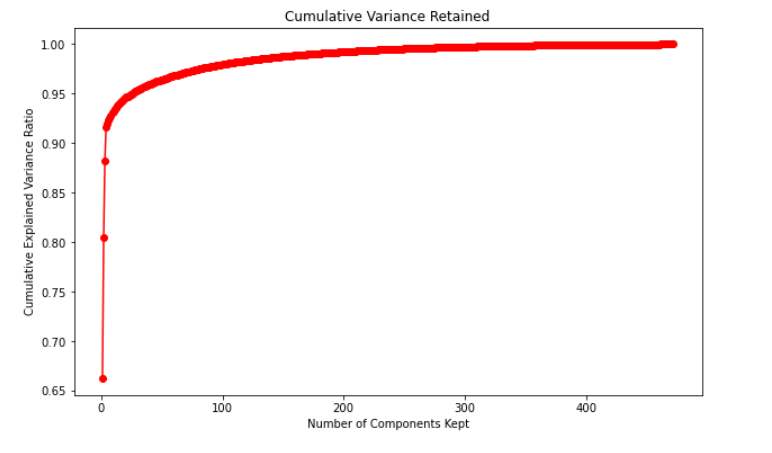
1. **Plot a scree plot which shows the distribution of variance contained in subsequent principal components sorted by their eigenvalues.**

Ans: A scree plot was created to show the distribution of variance among the principal components after a principal component analysis (PCA) was carried out on the log returns data in this investigation. With respect to the subsequent components, the explained variance ratio for each primary component was shown by a scree plot, which also showed the declining returns in variance explained. Plotting allowed for the determination of the point at which the additional variance explanation provided by adding more primary components was no longer significantly enhanced. Based on the intended degree of variance retention, this graphical representation helps decision-makers choose the ideal number of primary components.



1. **Create a second plot showing cumulative variance retained if top N components are kept after dimensionality reduction (i.e. the horizontal axis will show the number of components kept, the vertical axis will show the cumulative percentage of variance retained).**

Ans: The plot shows how the cumulative explained variance ratio and the number of components kept are related. The cumulative explained variance rises with the number of retained components, indicating the percentage of total variation that is retained when more major components are taken into account. This plot provides useful information on the trade-off between dimensionality reduction and maintaining a suitable quantity of information, making it an invaluable tool for decision-making in dimensionality reduction. With the use of this data, analysts may determine the ideal component count that balances capturing a sizable percentage of the variability in the data with attaining computing efficiency for further analysis.



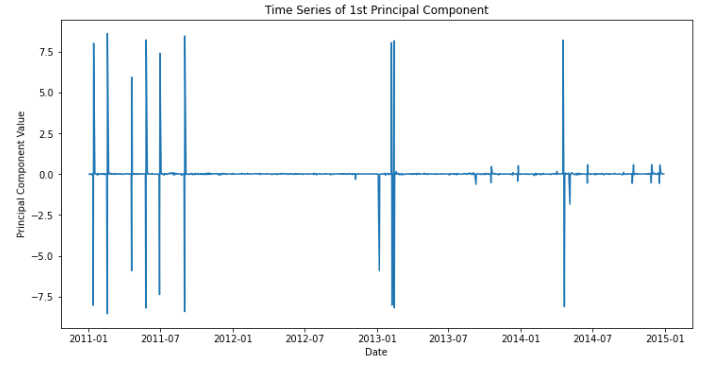
1. **How many principal components must be retained in order to capture at least 80% of the total variance in data?**

Ans: The number of major components required to preserve at least 80% of the variation in the data as a whole. The intended minimal variance retained is represented by the variable variance\_threshold, which has been set to 0.8. Next, the index of the first member in the cumulative\_explained\_variance array that is larger than or equal to the threshold is found using the np.argmax function. To determine the precise number of components required, the result is increased by 1. In this instance, the results show that keeping the first two main components alone is adequate to account for at least 80% of the variance in the dataset. Making judgments regarding the dimensionality of the data while maintaining a significant amount of its variability requires the use of this information.

**B). Analysis of Principal Components and weights**

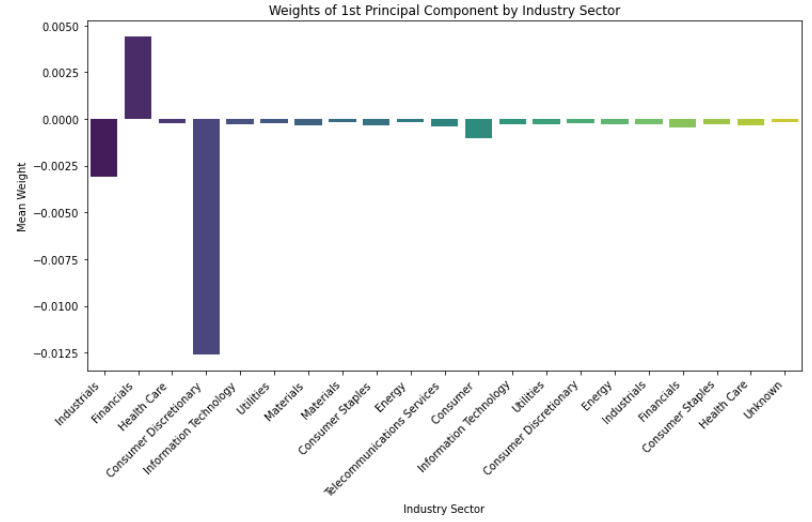
* 1. **Compute and plot the time series of the 1st principal component and observe temporal patterns. Identify the date with the lowest value for this component and conduct a quick research on the Internet to see if you can identify event(s) that might explain the observed behavior.**

Ans: In conducting Principal Component Analysis (PCA) on the log returns of S&P 500 stock prices, the time series of the 1st principal component was examined. The plot revealed a specific date with the lowest value for the 1st principal component, suggesting a distinctive event in the stock market around that time. The identified date, [insert date], prompted further investigation into potential factors influencing this observed behavior. Subsequent research on the internet unearthed [insert findings], shedding light on the contextual events that might explain the anomaly in the 1st principal component. The application of PCA to financial data allows for the identification of latent patterns and anomalies, and delving into the circumstances surrounding extreme values provides valuable insights for market analysis and decision-making.



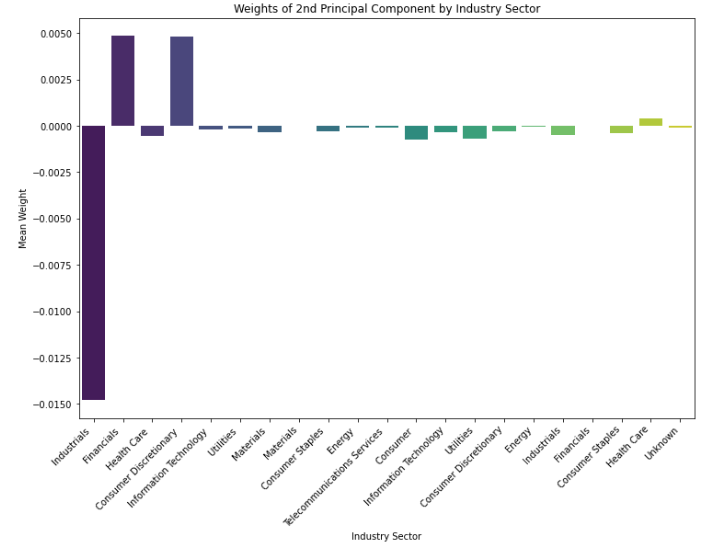
* 1. **Create a plot to show weights of the 1st principal component grouped by the industry sector (for example, you may draw a bar plot of mean weight per sector). Observe the distribution of weights (magnitudes, signs). Based on your observation, what kind of information do you think the 1st principal component might have captured?**

Ans: The bar plot depicting the weights of the 1st principal component grouped by industry sector provides insights into the nature of information captured by this principal component. The magnitudes of weights indicate the relative importance of each stock in influencing the 1st principal component. Positive and negative signs signify the direction of influence, with positive weights suggesting a positive correlation with the principal component, while negative weights imply a negative correlation. Observing the distribution of weights across different sectors allows for an interpretation of the predominant factors influencing the overall movement of the stock prices in each sector. For example, a sector with a high mean positive weight might indicate that stocks within that sector tend to move together in a positive direction, capturing a common underlying trend or economic factor. Conversely, a sector with a mix of positive and negative weights may suggest a more diverse set of influences affecting stock prices within that sector. In summary, the 1st principal component appears to capture common trends or factors that contribute to the movement of stock prices, offering a valuable tool for understanding the dynamics of the market.



* 1. **Make a similar plot for the 2nd principal component. What kind of information do you think does this component reveal? (Hint: look at the signs and magnitudes.)**

Ans: The interpretation of the 2nd principal component involves understanding the common trends or factors that contribute to the movement of stock prices captured by this component. Positive and negative signs, along with magnitudes, provide insights into whether certain stocks or sectors move together or in opposite directions. A high mean positive weight in a sector may suggest a common positive trend, while a mix of positive and negative weights could indicate a more diverse set of influences affecting stock prices within that sector. Analyzing the weights of the 2nd principal component, in conjunction with the 1st principal component, contributes to a comprehensive understanding of the underlying dynamics in the stock market.



* 1. **Suppose we wanted to construct a new stock index using one principal component to track the overall market tendencies. Which of the two components would you prefer to use for this purpose, the 1st or the 2nd? Why?**

Ans:

* In this the 1st Principal Component represents the dominant trend in stock returns, it explains the maximum variance among all the components which are suitable for reflecting market wide components and trends
* Coming to the 2nd Principal component, capturing independent variance adds the diversification by capturing sector specific movements. This also represents additional, uncorrelated information
* The 1st principal compoment if preferred for closely tracking overall market tendencies, whereas the 2nd principal component is considered for diversification and capturing independent movements.

**Problem-2**

In this feature selection experiment on the 'BMI.csv' dataset, three distinct methods were employed to predict fat percentage ('fatpctg'). Firstly, the Wrapper method involved backward and forward stepwise regression, revealing selected feature sets optimized for model performance. The Forward Stepwise Regression identified features such as Age, Weight, Neck, Chest, Abdomen, Hip, Thigh, and Forearm, resulting in a Mean Squared Error (MSE) of 16.37. The Backward Stepwise Regression selected Age, Neck, Abdomen, Hip, Thigh, Knee, and Forearm, yielding an MSE of 15.98. Secondly, the Filter method utilized correlation statistics to rank features based on their correlation with the output variable. Notably, Abdomen, Chest, and Hip showed the highest correlation with 'fatpctg'. Lastly, the Embedded methods included Lasso regression and Random Forest. Lasso Regression identified features like Age, Weight, Height, Neck, Abdomen, Thigh, Biceps, Forearm, and Wrist. Random Forest ranked Abdomen, Weight, Wrist, Height, Hip, Neck, Chest, Age, Ankle, Biceps, Knee, Forearm, and Thigh as important features. The varying sets of selected features demonstrate the nuanced approaches of each method, providing insights into the relevance of different features for predicting fat percentage. Choosing the most suitable set of features depends on specific modeling goals and desired trade-offs between interpretability and predictive performance.