K-means Clustering of S&P 500 Stock Data

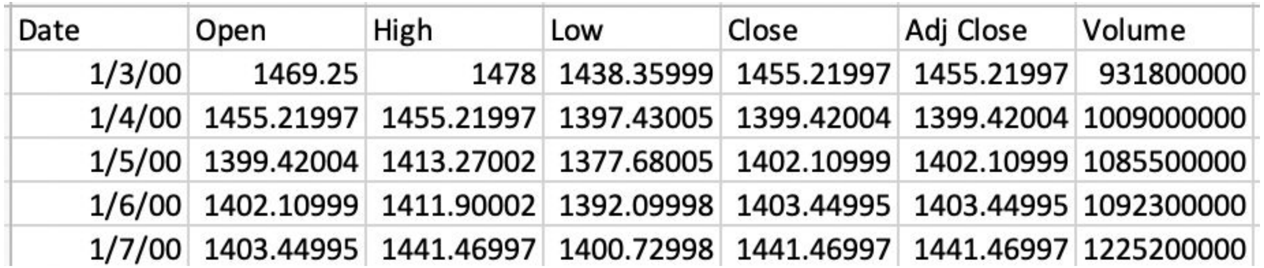
Jessica Kwok, Melody Zhao, Kewei Zhou

Goal/Motivation

- Connect significant events to stock performance trend during that period of time
- Which events resulted in similar reactions? Different reactions?
- How did countries respond differently?
  - E.g. COVID-19 US vs China reaction
- Predict future trend based on current trends
  - Pattern finding

Data Collection

- Used Yahoo Finance and Barchart's services on historical data
- Collected data for S&P 500
- Found daily historical data from 2000 to current with the following information:
    - Date, Open price, high, low, closing price, volume



Date	Open	High	Low	Close	Adj Close	Volume
1/3/00	1469.25	1478	1438.35999	1455.21997	1455.21997	931800000
1/4/00	1455.21997	1455.21997	1397.43005	1399.42004	1399.42004	1009000000
1/5/00	1399.42004	1413.27002	1377.68005	1402.10999	1402.10999	1085500000
1/6/00	1402.10999	1411.90002	1392.09998	1403.44995	1403.44995	1092300000
1/7/00	1403.44995	1441.46997	1400.72998	1441.46997	1441.46997	1225200000

Methodology

- Data Pre-processing
- Principal Component Analysis (PCA)
- K-means Clustering
- Markov Model

Data pre-processing

Historical Data

Calculate Percent Change

Matching Standard Date

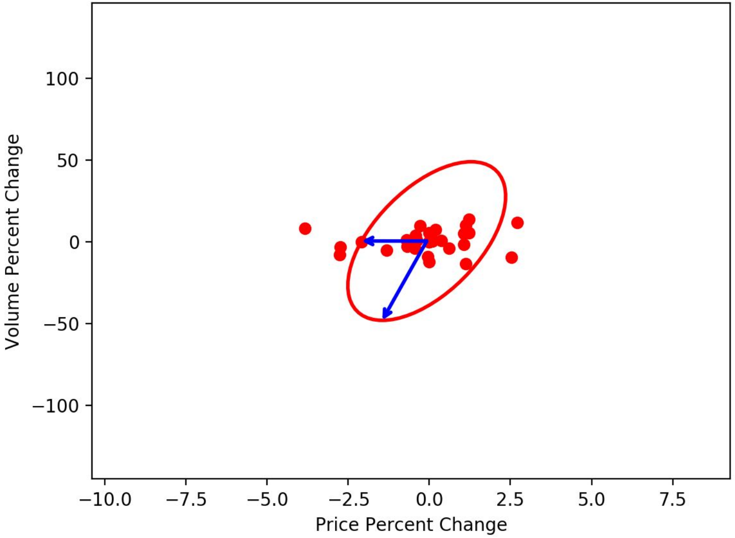
Get historical data from Yahoo Finance or Barchart.

- Percent change (T) = (Price(T) - Price(T-1)/ Price (T-1)
- Done directly on Excel

- Used standard dates to simplify comparison in the future
- Default to 0 if any standard date’s information is not found

Principal Components Analysis (PCA)

- Center
- Largest variance
- Second largest variance (orthogonal to the first)
- 1st principal component
- 2nd principal component



K-Means Clustering

- Unsupervised learning
- Label data points into k clusters
- Steps:
    - Normalize Data
    - Distance Function
    - Determine optimal K

K-Means Clustering - Normalize Data

- Input:

$$\begin{bmatrix} p_1 & v_1 & \lambda_{prin,1} & \lambda_{minor,1} & \theta_1 \\ p_2 & v_2 & \lambda_{prin,2} & \lambda_{minor,2} & \theta_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_n & v_n & \lambda_{prin,n} & \lambda_{minor,n} & \theta_n \end{bmatrix}$$

- Used Min-Max Normalization Method

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

K-Means Clustering - Distance Function

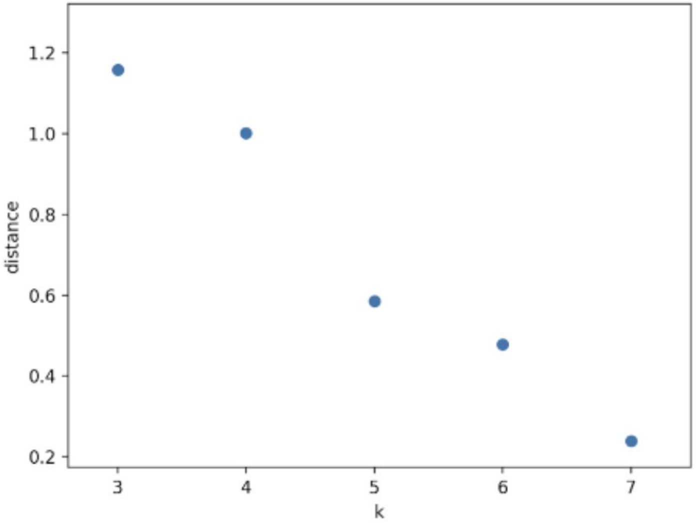
$$D = W_{center} \cdot \Delta center + W_{\lambda_1} \cdot \Delta\lambda_1 + W_{\lambda_2} \cdot \Delta\lambda_2 + W_\theta \cdot \Delta\theta$$

$$W_{center} + W_{\lambda_1} + W_{\lambda_2} + W_\theta = 1$$

- Determined that:
    - $(W_{center}, W_{\lambda_1}, W_{\lambda_2}, W_\theta) = (0.2, 0.78, 0.015, 0.005)$

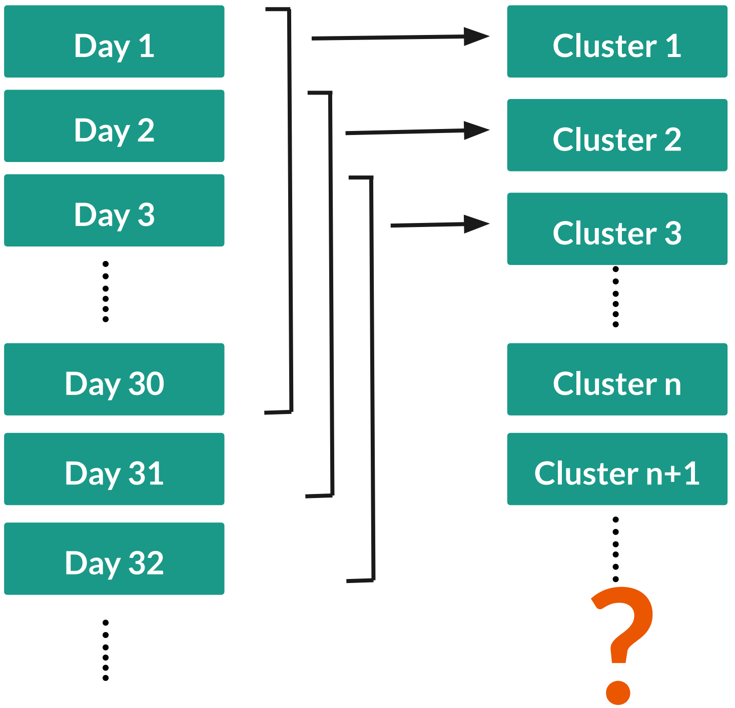
K-Means Clustering - Optimal K

- Ran K-Means Clustering on different K values
- Plotted K vs average distance within the clusters
- 7 has the lowest distance → optimal K

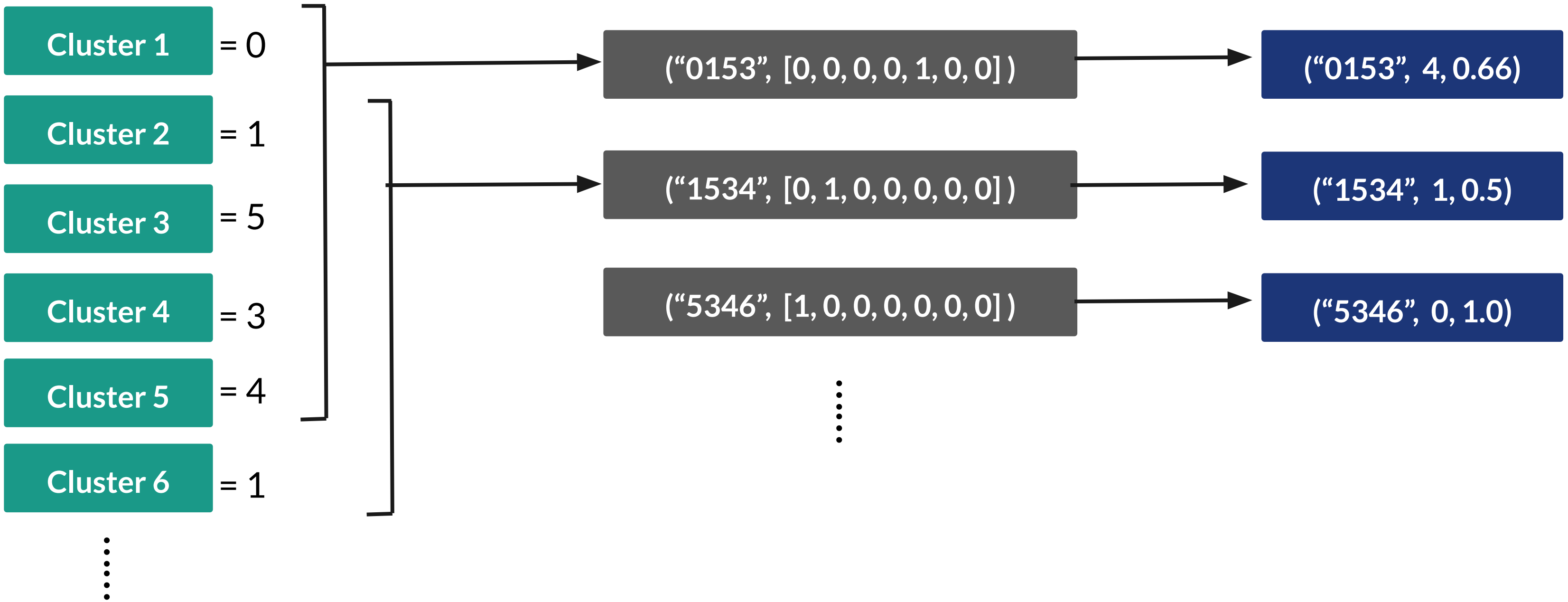


Markov Model

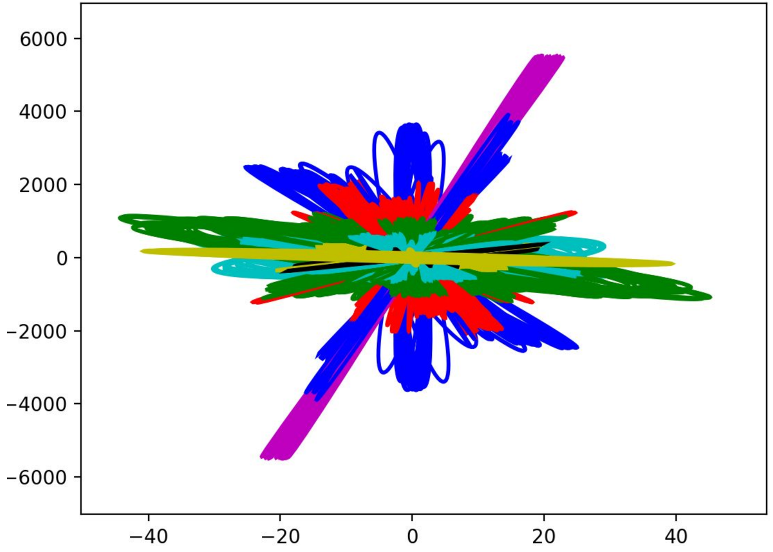
- Predict upcoming clusters

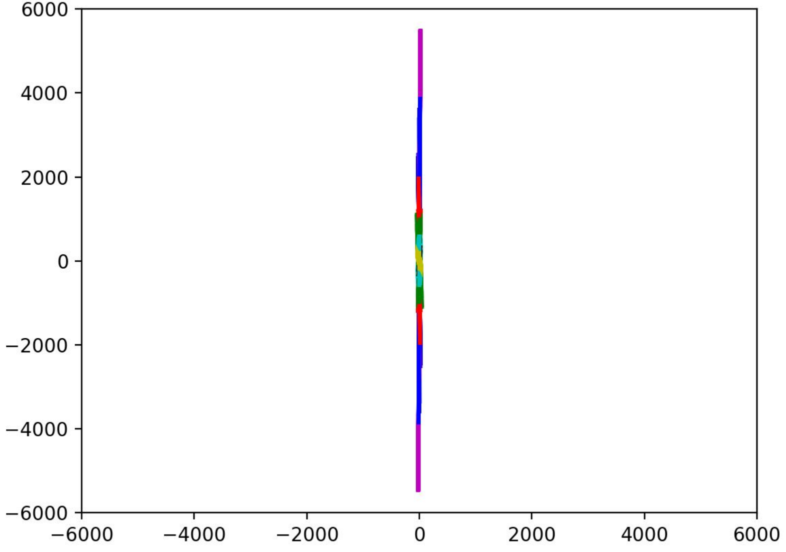


Markov Model



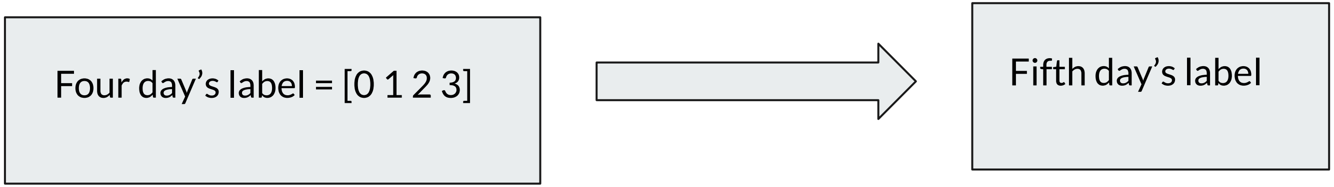
Results of K-Means Clustering





X-axis rescaled

Results of Markov Model



Four day’s label = [0 1 2 3]

Fifth day’s label

Accuracy: **92%**

- May suggest k-means is working correctly, can be applied to find a longer sequence of trend
- Possible to happen by luck, need more extensive testing

Conclusion

- K-means clustering algorithm seems to work, but must find alternative normalization method to make variance of each variable similar
    - Parameters still need to be optimized
- K-means clustering results can be applied in Markov model with high accuracy
- More extensive testing must be done before conclusion

Future Work

- Revising Markov Model and clustering method
- Other time series prediction models
- Finding correlation between the trends in different countries in the same period
- Identifying similarities between different countries or different companies across different periods
- Better method for pre-processing
- More data

References

- Garbade, Michael J. “Understanding K-Means Clustering in Machine Learning.” Medium, Towards Data Science, 12 Sept. 2018, towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. \newline
- Irshad, Hira. “Relationship Among Political Instability, Stock Market Returns and Stock Market Volatility.” Studies in Business and Economics, vol. 12, no. 2, 2017, pp. 70–99., doi:10.1515/sbe-2017-0023.\newline
- “Sklearn.decomposition.PCA.” Scikit, scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.\newline
- McCaffrey03/27/2018, James. “Data Clustering with K-Means Using Python.” Visual Studio Magazine, visualstudiomagazine.com/articles/2018/03/27/clustering-with-k-means-using-python.aspx.