“OPTIMIZING STOCK TRADING STRATEGY WITH K-MEANS CLUSTERING”

Mini Project

Big Data Analytics [BDA]

(Fourth Year/ Sem VII)

Submitted in fulfilment of the requirement of
University of Mumbai For the Degree of

Bachelor Of Engineering
(Computer Engineering)

By

1.	**AMEY THAKUR**	**BE-COMPS B-50**	**TU3F1819127**
2.	**HASAN RIZVI**	**BE-COMPS B-51**	**TU3F1819130**
3.	**MEGA SATISH**	**BE-COMPS B-58**	**TU3F1819139**

Under the Guidance of
Prof. D.M. Bavkar

Department of Computer Engineering
TERNA ENGINEERING COLLEGE
Plot no.12, Sector-22, Opp. Nerul Railway station,
Phase-11, Nerul (w), Navi Mumbai 400706
UNIVERSITY OF MUMBAI
2020-2021

Internal Approval Sheet

Terna Engineering College

NERUL, NAVI MUMBAI

CERTIFICATE

This is to certify that

1. **AMEY MAHENDRA THAKUR**
2. **HASAN MEHDI RIZVI**
3. **MEGA SATISH**

Has satisfactorily completed the requirements of the

Mini Project (Fourth Year/ Sem VII)

entitled

“OPTIMIZING STOCK TRADING STRATEGY WITH K-MEANS CLUSTERING”

As prescribed by the University of Mumbai

Under the guidance of **Prof. D. M. Bavkar**

GUIDE	**APC**	**HOD**

Approval Sheet

Project Report Approval

This Mini Project Report - I entitled

“OPTIMIZING STOCK TRADING STRATEGY WITH K-MEANS CLUSTERING”

by the following students are approved for the degree of Bachelor in

"Computer Engineering (Sem-VII)".

Submitted by:

AMEY THAKUR	**TU3F1819127**
HASAN RIZVI	**TU3F1819130**
MEGA SATISH	**TU3F1819139**

Examiners Name & Signature:

1. --

2. --

Date: 24-10-2021
Place: MUMBAI

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced The cartoon sources task-specific. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

AMEY THAKUR	**TU3F1819127**	Amey
HASAN RIZVI	**TU3F1819130**	
MEGA SATISH	**TU3F1819139**	

Date: 24-10-2021

Place: MUMBAI

ACKNOWLEDGEMENT

We would like to express our sincere gratitude towards our guide **Prof. D.M. Bavkar** for the help, guidance and encouragement she provided during the project development. This work would have not been possible without her valuable time, patience and motivation. We thank her for making our stint thoroughly pleasant and enriching. It was great learning and an honour being her student.

We are deeply thankful to **Dr Archana Mire** (H.O.D Computer Department) and the entire team in the Computer Department. They supported us with scientific guidance, advice and encouragement, they were always helpful and enthusiastic and this inspired us in our work.

We take the privilege to express our sincere thanks to **Dr L. K. Ragha** our Principal for providing encouragement and much support throughout our work.

ABSTRACT

We propose to determine a single exchange-traded fund (SPY) investing strategy that will maximise our total wealth. We compare our findings to the tried-and-true “buy-and-hold” and Moving average convergence divergence (MACD) strategies. It’s a Machine Learning model which integrates Data Science and Web Development. We have deployed the app on the Heroku Cloud Application Platform. Here, we intend to base an evaluation on every basic criterion that is taken into account when establishing the pricing. As mentioned, we intend to forecast future price fluctuations for a specific stock. Prices from previous days, as well as financial media stories connected to the firm of interest, are used to create these forecasts. Reinforcement learning is all about taking the right steps to maximise your reward in a given situation. It is used by a variety of software and computers to determine the best feasible action or path in a given situation. The Reinforcement Machine Learning model is employed in this work to forecast the closure price using past data. We develop a predictor for multiple firms using these trained models that forecasts the every day close stock prices. By providing inputs such as open, high, and low prices, stock volume, and the latest events about each firm, this predictor may be used to determine the price at which the stock value will close for a certain day. The goal of this project is to learn and get hands-on experience in Data Analytics and Machine Learning.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

We've always been captivated by the stock market's seeming unpredictability. There are hundreds of stocks to select from, and day traders can trade almost any of them. So, for a day trader, deciding what to trade is the first and most important step. The following stage is to come up with some ways to benefit from the trading opportunity (one stock, several stocks, exchange-traded funds, ETFs, etc.). Intraday traders use a number of ways to profit from price movements in a particular asset. Day traders should look for equities with plenty of liquidity, moderate to high volatility, and a large number of followers. Isolating the present market trend from any surrounding noise and then capitalising on that trend is the key to finding the appropriate stocks for intraday trading.

This has traditionally been done in conjunction with the trade plan and current events. Various research techniques have been explored to automate this laborious procedure since the emergence of Data Science and Machine Learning. This automated trading method will assist in providing recommendations at the appropriate moment and with more accurate estimates. Mutual funds and hedge funds would benefit greatly from an automated trading approach that maximises profits. The type of profitable returns that may be expected will be accompanied by some risk. It's difficult to come up with a lucrative automated trading technique.

Every human being aspires to make as much money as possible in the stock market. It's critical to devise a well-balanced, low-risk plan that will benefit the majority of individuals. One such technique proposes the use of K means clustering to generate automated trading strategies based on previous data. This project was made because we were intrigued and we wanted to gain hands-on experience with the Machine Learning Project.

CHAPTER 2

LITERATURE SURVEY

Clustering very large datasets is a challenging problem for data mining and processing. MapReduce is considered a powerful programming framework that significantly reduces executing time by dividing a job into several tasks and executing them in a distributed environment.

K-Means which is one of the most used clustering methods and K-Means based on MapReduce is considered as an advanced solution for very large dataset clustering. However, the execution time is still an obstacle due to the increasing number of iterations when there is an increase in dataset size and the number of clusters.

To get a better understanding of K-Means Clustering and its uses for the stock trading strategy, we referred to the following papers:

"Fast K-Means Clustering for Very Large Datasets Based on MapReduce Combined with a New Cutting Method", by Duong Van Hieu and Phayung Meesad. This paper gives an overview of the importance of the K-Means Clustering Algorithm for Big Data. This paper presents a new approach for reducing the number of iterations of the K-Means algorithm which can be applied to very large dataset clustering.

"Enhancing stock prediction clustering using K-means with genetic algorithm", by E. N. Desokey, A. Badr and A. F. Hegazy. The main objectives of this paper are to optimize the clustering of stock market prediction and to examine the impact of applying genetic algorithm optimization with the k-means clustering algorithm. The evaluation shows that using genetic algorithm and k-means clustering algorithm with Chi-square similarity measure achieved the highest accuracy with the least sum of square distances.

CHAPTER 3

PROBLEM STATEMENT

The main emphasis and objective of our project is to analyse given raw data and do exploratory data analysis in order to fully comprehend and identify patterns. Then, using a Neural Network approach, construct a model and train it to get the desired outcomes. Finally, it will be deployed as a web application.

CHAPTER 4

METHODOLOGY

4.1 K-MEANS CLUSTERING

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known or labelled outcomes.
A cluster refers to a collection of data points aggregated together because of certain similarities.

In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

To process the learning data, the K-means algorithm in data mining starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

- ➔ The centroids have stabilized — there is no change in their values because the clustering has been successful.
- ➔ The defined number of iterations has been achieved.

4.2 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process. Another use of PCA is to compress the data and hence save computational time.

PCA in conjunction with k-means is a powerful method for visualizing high dimensional data.

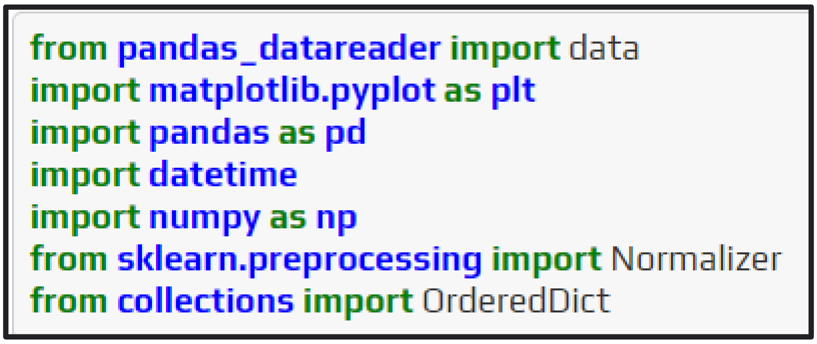
CHAPTER 5

IMPLEMENTATION

5.1 STEPS TO CREATE AND DEPLOY MODEL

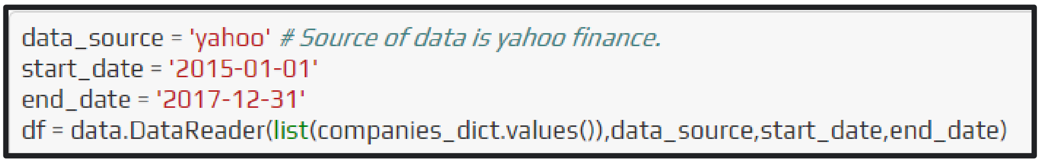
1. Import necessary Libraries.
2. Load Dataset.
3. Perform Exploratory Data Analysis.
4. Create An Environment.
5. Prepare Data.
6. Train Data.
7. Test Data.
8. Evaluate Model.
9. Deploy Model.

5.2 IMPORT NECESSARY LIBRARIES



```
from pandas_datareader import data
import matplotlib.pyplot as plt
import pandas as pd
import datetime
import numpy as np
from sklearn.preprocessing import Normalizer
from collections import OrderedDict
```

5.3 LOAD DATASET

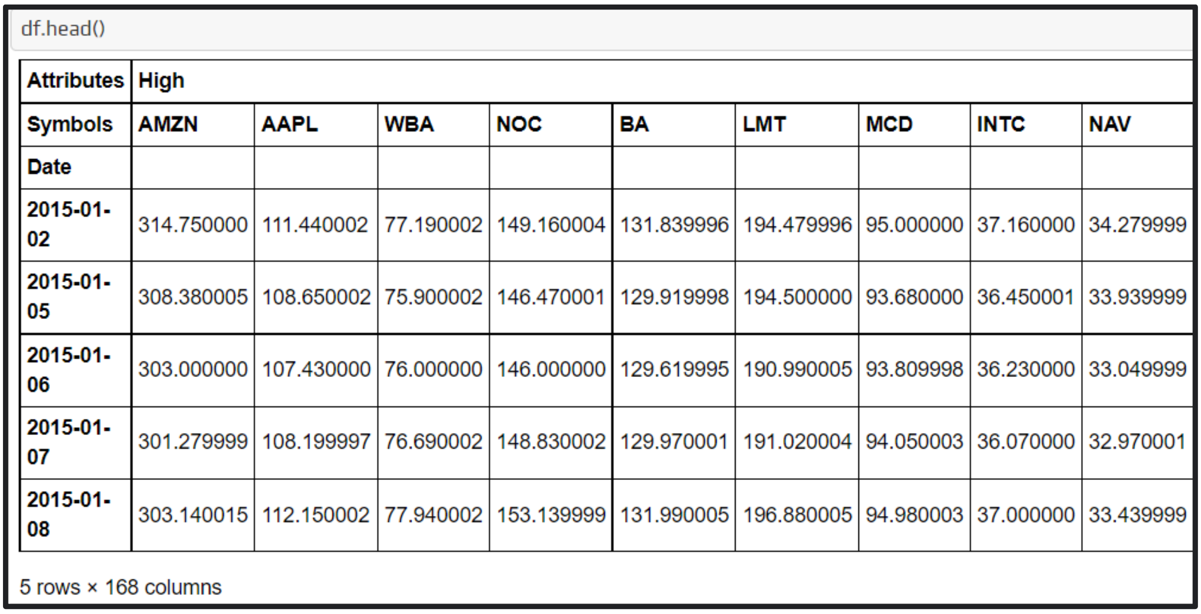


```
data_source = 'yahoo' # Source of data is yahoo finance.
start_date = '2015-01-01'
end_date = '2017-12-31'
df = data.DataReader(list(companies_dict.values()),data_source,start_date,end_date)
```

The following is a quick description of each component:

1. DATE - The day on which stock is exchanged.
2. OPEN - Any listed stock's daily opening price is the price at which it is initially traded.

3. HIGH - The maximum price at which a stock can be bought or sold during a trading day is known as the high.
4. LOW - The minimum price at which a stock can be bought or sold during a trading day is known as the low.
5. CLOSE - During a standard trading session, the last price at which a stock trades is referred to as the closing price.
6. VOLUME - The total number of shares or contracts exchanged in a securities or market within a certain time period.
7. NAME - The name of the stock.



```
df.head()
```

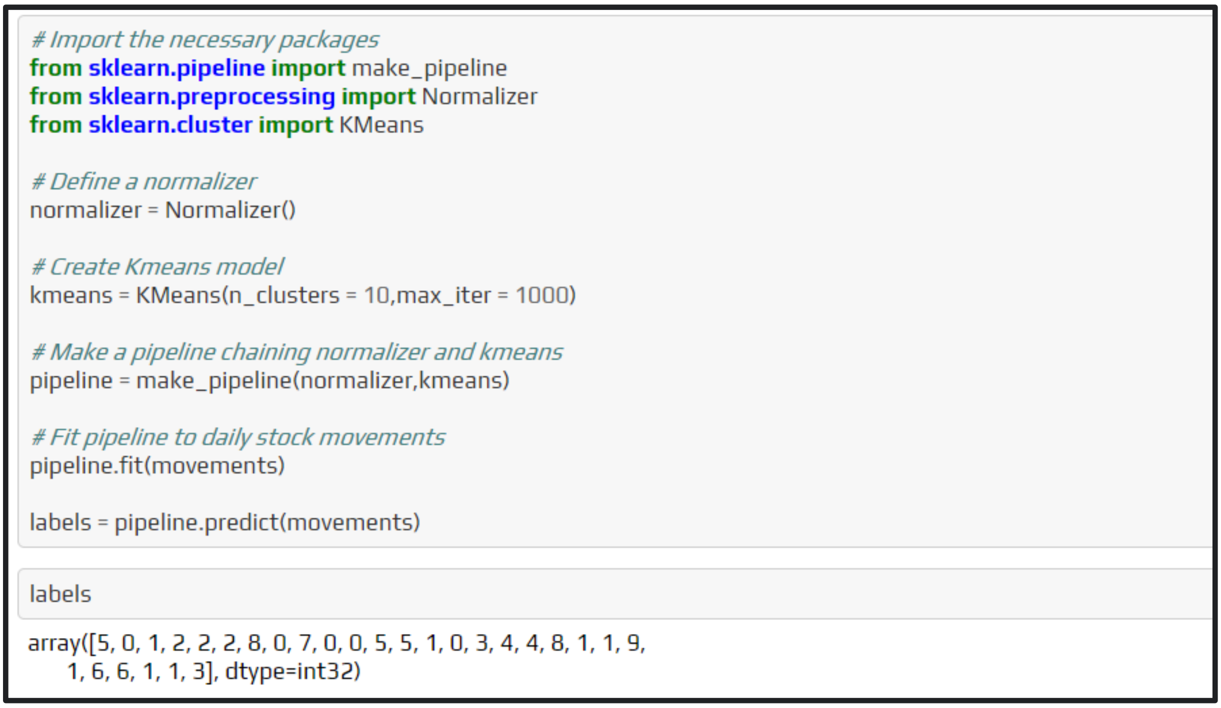
Attributes	High								
Symbols	AMZN	AAPL	WBA	NOC	BA	LMT	MCD	INTC	NAV
Date									
2015-01-02	314.750000	111.440002	77.190002	149.160004	131.839996	194.479996	95.000000	37.160000	34.279999
2015-01-05	308.380005	108.650002	75.900002	146.470001	129.919998	194.500000	93.680000	36.450001	33.939999
2015-01-06	303.000000	107.430000	76.000000	146.000000	129.619995	190.990005	93.809998	36.230000	33.049999
2015-01-07	301.279999	108.199997	76.690002	148.830002	129.970001	191.020004	94.050003	36.070000	32.970001
2015-01-08	303.140015	112.150002	77.940002	153.139999	131.990005	196.880005	94.980003	37.000000	33.439999

5 rows × 168 columns

5.4 EXPLORATORY DATA ANALYSIS

In every Data Analysis or Data Science project, exploratory data analysis, or EDA, is a critical stage. EDA is the investigation of a dataset to find patterns and anomalies as well as to generate hypotheses based on our knowledge of the information.

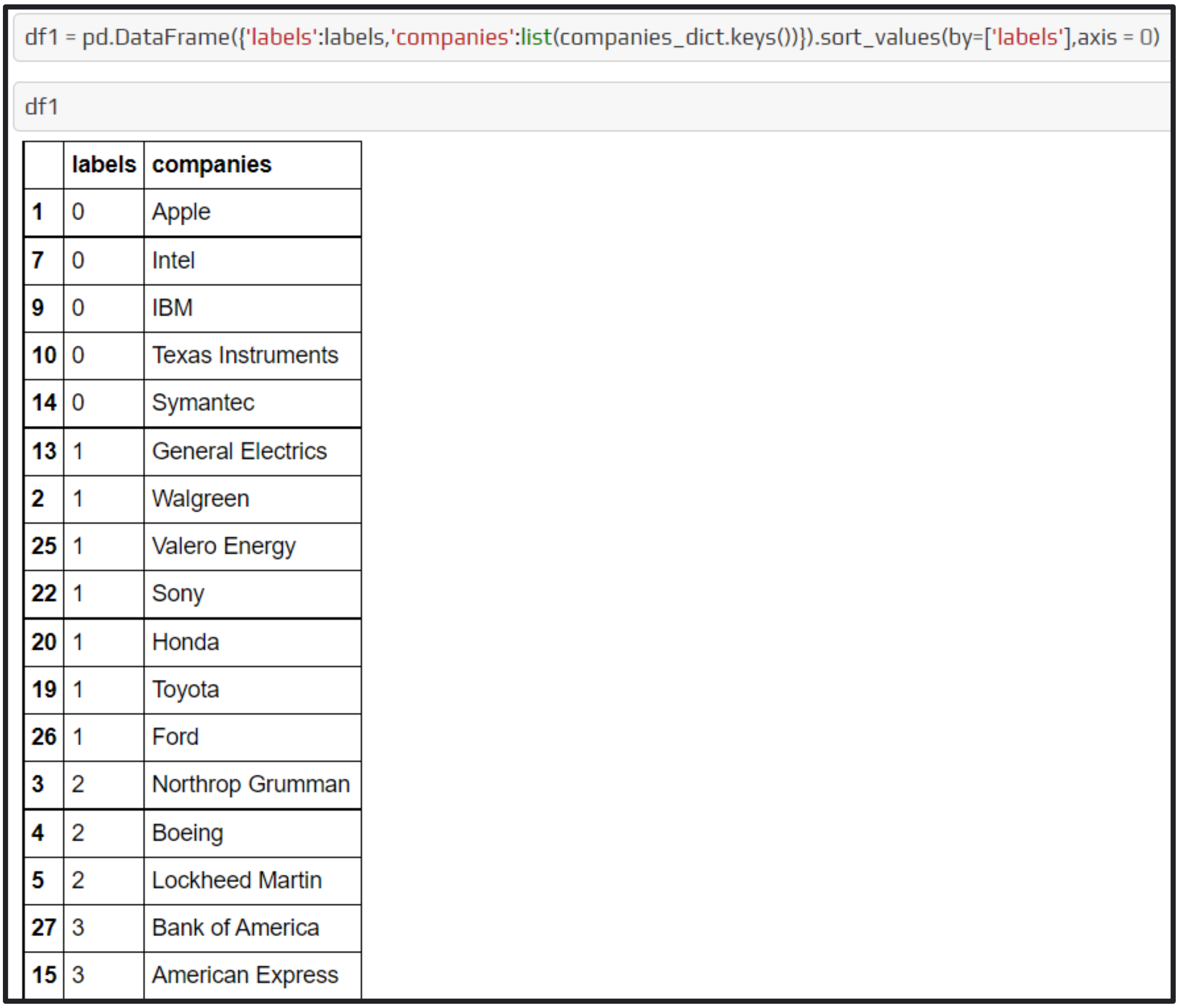
During the course of a data science or machine learning project, EDA consumes around half of the time spent on data analysis, feature selection, feature engineering, and other processes. Because it is the most essential element or backbone of a data science project, where one has to perform several tasks like data cleaning, dealing with missing values, handling outliers, treating unbalanced datasets, handling categorical features, and many others. To automate our tasks in exploratory data analysis, we may utilise python packages like dtale, pandas profiling, sweetviz, and autoviz.



```
# Import the necessary packages
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import Normalizer
from sklearn.cluster import KMeans

# Define a normalizer
normalizer = Normalizer()

# Create Kmeans model
kmeans = KMeans(n_clusters = 10,max_iter = 1000)

# Make a pipeline chaining normalizer and kmeans
pipeline = make_pipeline(normalizer,kmeans)

# Fit pipeline to daily stock movements
pipeline.fit(movements)

labels = pipeline.predict(movements)
```

```
labels
```

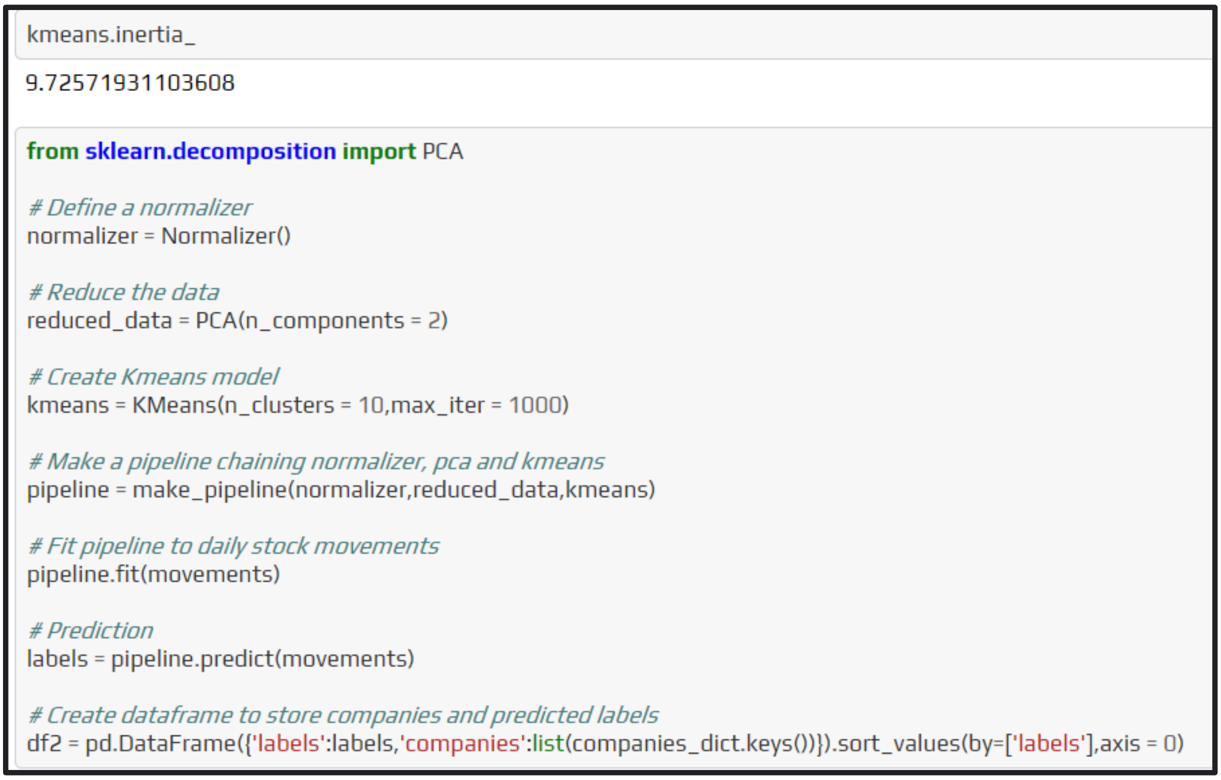
```
array([5, 0, 1, 2, 2, 2, 8, 0, 7, 0, 0, 5, 5, 1, 0, 3, 4, 4, 8, 1, 1, 9,
      1, 6, 6, 1, 1, 3], dtype=int32)
```



```
df1 = pd.DataFrame({'labels':labels,'companies':list(companies_dict.keys())}).sort_values(by=['labels'],axis = 0)
```

```
df1
```

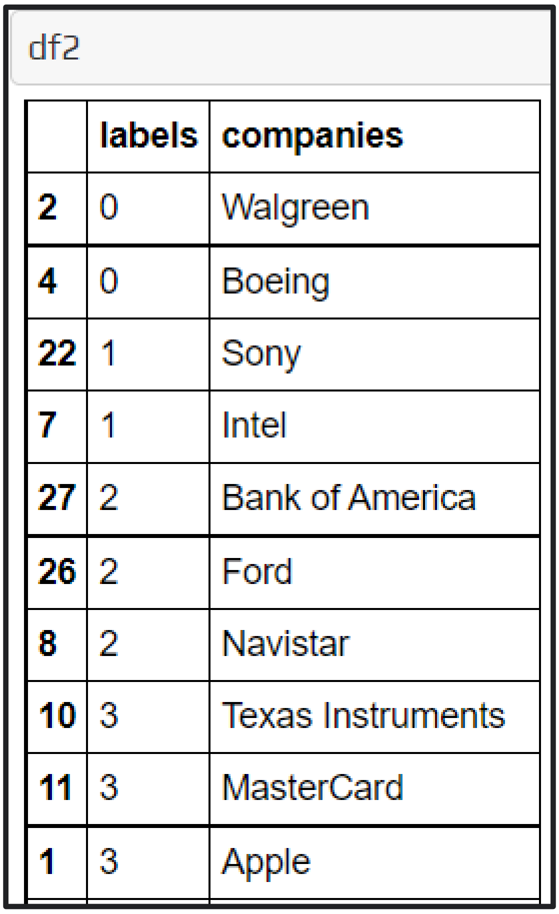
	labels	companies
1	0	Apple
7	0	Intel
9	0	IBM
10	0	Texas Instruments
14	0	Symantec
13	1	General Electrics
2	1	Walgreen
25	1	Valero Energy
22	1	Sony
20	1	Honda
19	1	Toyota
26	1	Ford
3	2	Northrop Grumman
4	2	Boeing
5	2	Lockheed Martin
27	3	Bank of America
15	3	American Express



```
kmeans.inertia_
```

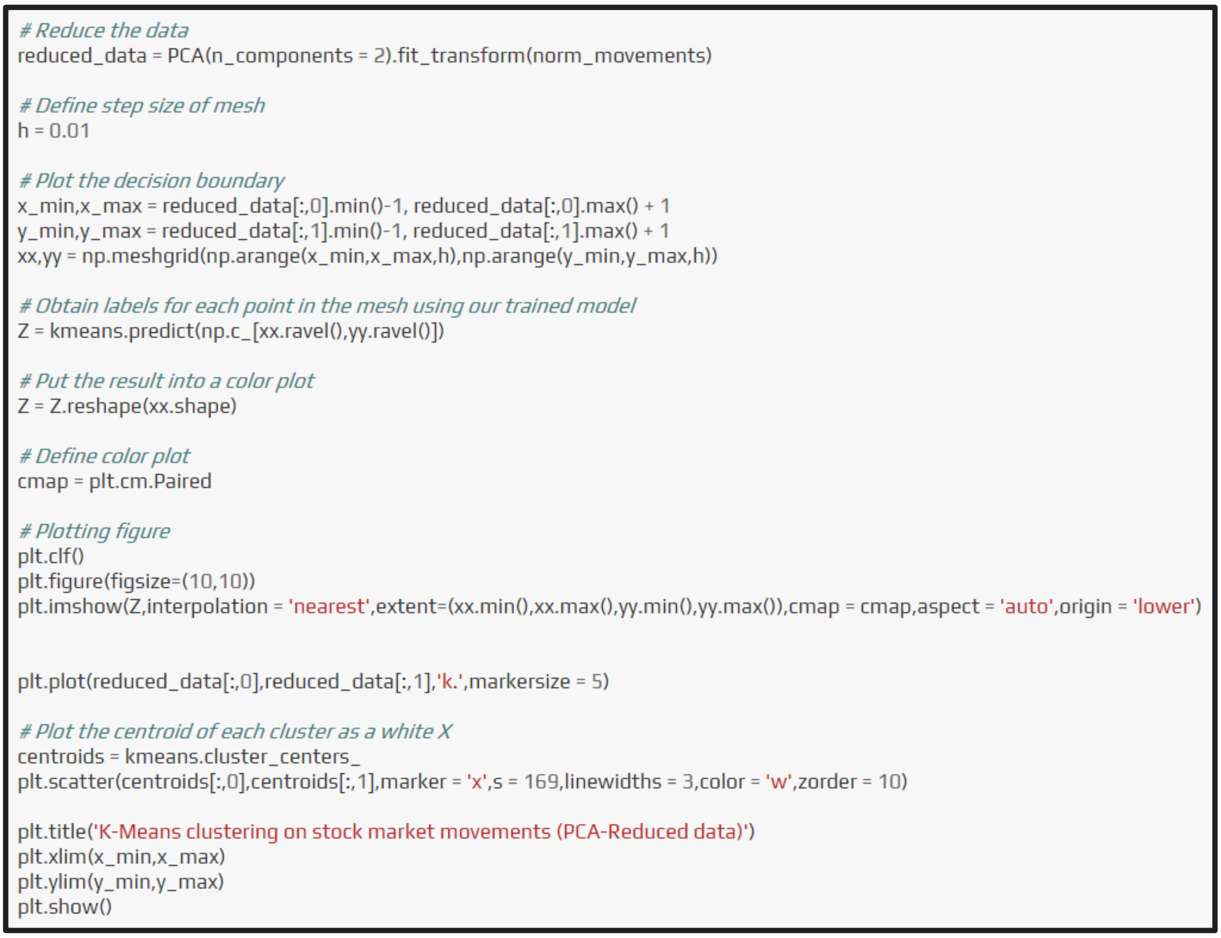
```
9.72571931103608
```

```
from sklearn.decomposition import PCA

# Define a normalizer
normalizer = Normalizer()

# Reduce the data
reduced_data = PCA(n_components = 2)

# Create Kmeans model
kmeans = KMeans(n_clusters = 10,max_iter = 1000)

# Make a pipeline chaining normalizer, pca and kmeans
pipeline = make_pipeline(normalizer,reduced_data,kmeans)

# Fit pipeline to daily stock movements
pipeline.fit(movements)

# Prediction
labels = pipeline.predict(movements)

# Create dataframe to store companies and predicted labels
df2 = pd.DataFrame({'labels':labels,'companies':list(companies_dict.keys())}).sort_values(by=['labels'],axis = 0)
```

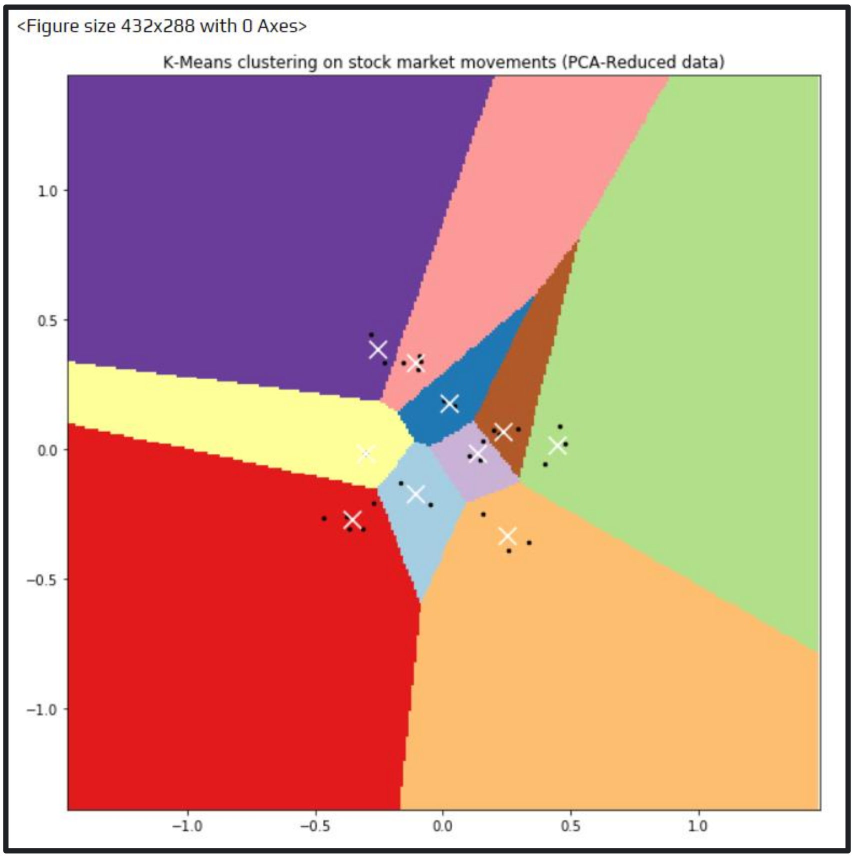


```
df2
```

	labels	companies
2	0	Walgreen
4	0	Boeing
22	1	Sony
7	1	Intel
27	2	Bank of America
26	2	Ford
8	2	Navistar
10	3	Texas Instruments
11	3	MasterCard
1	3	Apple



```
# Reduce the data
reduced_data = PCA(n_components = 2).fit_transform(norm_movements)

# Define step size of mesh
h = 0.01

# Plot the decision boundary
x_min,x_max = reduced_data[:,0].min()-1, reduced_data[:,0].max() + 1
y_min,y_max = reduced_data[:,1].min()-1, reduced_data[:,1].max() + 1
xx,yy = np.meshgrid(np.arange(x_min,x_max,h),np.arange(y_min,y_max,h))

# Obtain labels for each point in the mesh using our trained model
Z = kmeans.predict(np.c_[xx.ravel(),yy.ravel()])

# Put the result into a color plot
Z = Z.reshape(xx.shape)

# Define color plot
cmap = plt.cm.Paired

# Plotting figure
plt.clf()
plt.figure(figsize=(10,10))
plt.imshow(Z,interpolation = 'nearest',extent=(xx.min(),xx.max(),yy.min(),yy.max()),cmap = cmap,aspect = 'auto',origin = 'lower')


plt.plot(reduced_data[:,0],reduced_data[:,1],'k.',markersize = 5)

# Plot the centroid of each cluster as a white X
centroids = kmeans.cluster_centers_
plt.scatter(centroids[:,0],centroids[:,1],marker = 'x',s = 169,linewidths = 3,color = 'w',zorder = 10)

plt.title('K-Means clustering on stock market movements (PCA-Reduced data)')
plt.xlim(x_min,x_max)
plt.ylim(y_min,y_max)
plt.show()
```


<Figure size 432x288 with 0 Axes>

CHAPTER 6

MODEL EVALUATION

Model evaluation is a step in the model creation process that is often overlooked. This is the stage where the model's performance is determined. As a result, it's important to think about the model's results in terms of every conceivable assessment technique. Using various approaches can result in a variety of viewpoints.

CHAPTER 7

MODEL DEPLOYMENT

The project is hosted on Heroku which is a cloud Platform as a container-based Service (PaaS). Heroku is used by developers to launch, manage, and grow contemporary programs. Heroku is an open-source software platform for machine learning and data science that makes it simple to develop and publish attractive, bespoke web apps. The benefit of web apps is that they are platform agnostic and may be operated by anybody with an Internet connection. Their code is run on a back-end server, which processes incoming requests and answers using a common protocol that all browsers can understand.

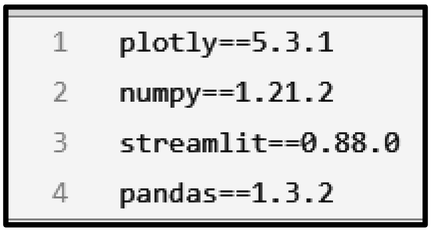
Necessary Files:

1. app.py
2. requirement.txt - Contains a list of all the dependencies that your code requires in order to function properly.

3. Procfile - In an app, a Procfile is a list of process types.
4. setup.sh

requirement.txt

This file contains all of the libraries that must be installed in order for the project to run. This file may be manually produced by walking through the project and identifying all of the libraries used. However, we used the pipreqs module instead, which generated a requirement.txt file for us.

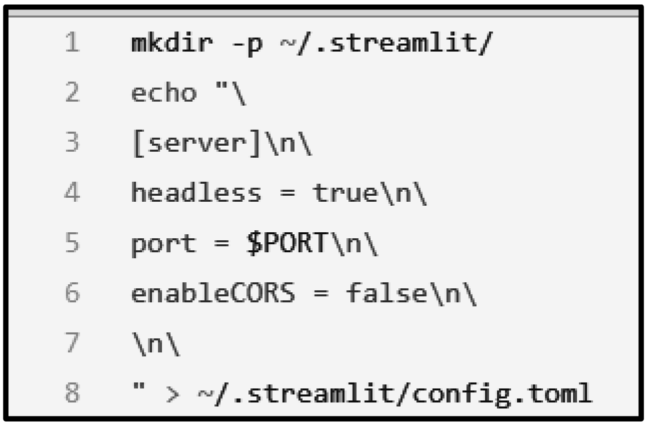


```
plotly==5.3.1
numpy==1.21.2
streamlit==0.88.0
pandas==1.3.2
```

Procfile and setup.sh

Using these two files, we tell Heroku the needed commands to start our application. In the setup.sh file we created a streamlit folder with credentials.toml and a config.toml file.

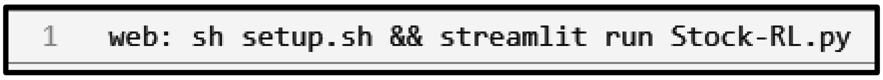
Following is the command to be pasted in the setup.sh file.



```
mkdir -p ~/.streamlit/
echo "\
[server]\n\
headless = true\n\
port = $PORT\n\
enableCORS = false\n\
\n\
" > ~/.streamlit/config.toml
```

The setup.sh script is first performed using the Procfile, and then the application is executed using streamlit run.

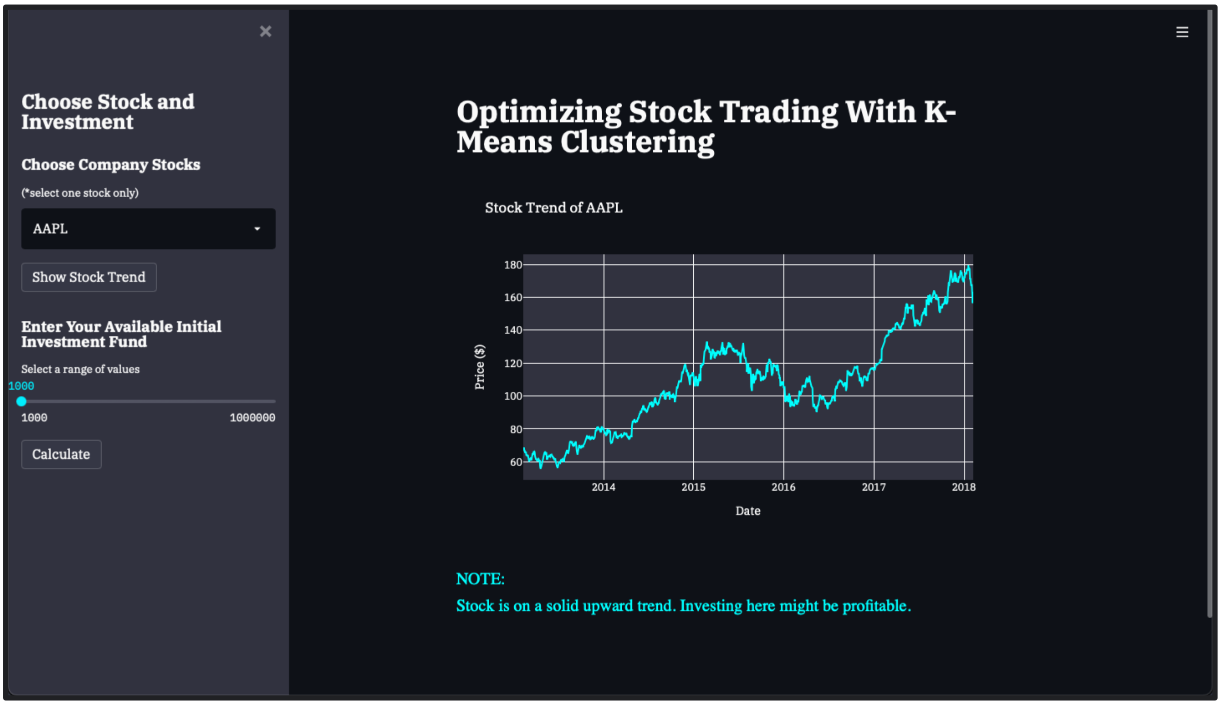
Procfile looks like this.



```
web: sh setup.sh && streamlit run Stock-RL.py
```

CHAPTER 8

RESULTS



Choose Stock and Investment

Choose Company Stocks

(*select one stock only)

AAPL

Enter Your Available Initial Investment Fund

Select a range of values

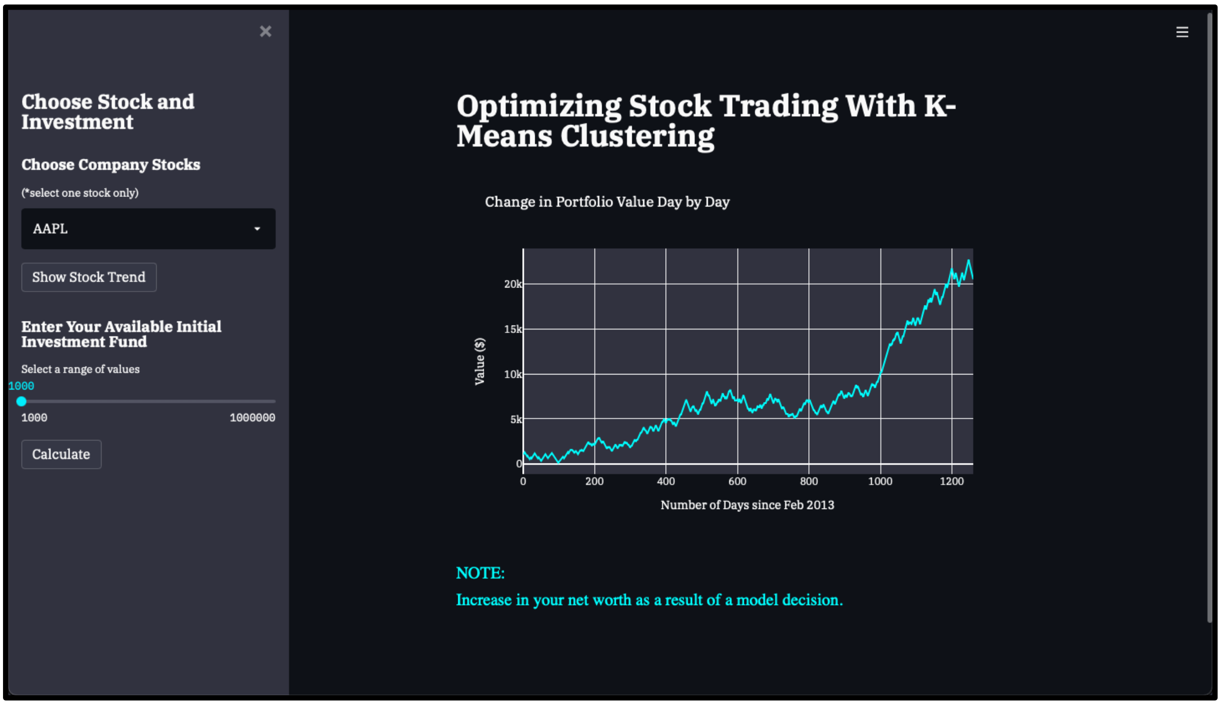
1000

Optimizing Stock Trading With K-Means Clustering

Stock Trend of AAPL

NOTE:

Stock is on a solid upward trend. Investing here might be profitable.



Choose Stock and Investment

Choose Company Stocks

(*select one stock only)

AAPL

Enter Your Available Initial Investment Fund

Select a range of values

1000

Optimizing Stock Trading With K-Means Clustering

Change in Portfolio Value Day by Day

NOTE:

Increase in your net worth as a result of a model decision.

CHAPTER 9

CONCLUSION

In this case study, we found a fuzzy K-means clustering being the most stable to group stock trading customers and used it to classify three tiers of customers (Normal, Best, and VIP level) based on the total trade amount over a 3-month period. For each group, a different brokerage commission rate is assigned. This approach is different from the existing graded commission policy in that the proposed policy adopts the idea of the graded commission based on the historically accumulated transaction amount made by the customer. This new approach is expected to bring more profit by treating loyal customers in a better manner and subsequently retaining them in the longer term.

REFERENCES

1. Segmentation of stock trading customers according to potential value (2004) https://doi.org/10.1016/j.eswa.2003.12.002

2. Customer Profitability Analysis, Cost System Purposes and Decision Making Process: A Research Framework

3. https://www.researchgate.net/publication/325092899_Customer_Profitability_Analysis_Cost_System_Purposes_and_Decision_Making_Process_A_Research_Framework

4. Van Hieu D., Meesad P. (2015) Fast K-Means Clustering for Very Large Datasets Based on MapReduce Combined with a New Cutting Method. In: Nguyen VH., Le AC., Huynh VN. (eds) Knowledge and Systems Engineering.

Advances in Intelligent Systems and Computing, vol 326. Springer, Cham. https://doi.org/10.1007/978-3-319-11680-8_23

5. E. N. Desokey, A. Badr and A. F. Hegazy, "Enhancing stock prediction clustering using K-means with a genetic algorithm," *2017 13th International Computer Engineering Conference (ICENCO)*, 2017, pp. 256-261, DOI: 10.1109/ICENCO.2017.8289797.