Amazon Web Services

Munagala Jayavardhan Reddy
Department of AI
Amrita Vishwa Vidyapeetham
India, Bengaluru
bl.en.u4aid23033@bl.students.amrita.edu

Maligi Thanuj Kumar
Department of AI
Amrita Vishwa Vidyapeetham
India, Bengaluru
bl.en.u4aid23030@bl.students.amrita.edu

***Abstract*—This paper outlines a comprehensive procedure for importing data from an Excel spreadsheet, normalizing numerical data, and performing Principal Component Analysis (PCA). The analysis is demonstrated using sales data, providing detailed steps and results for each phase of the process.**

***Index Terms*—amazon web services, matlab, principal component analysis, predictive modelling**

I. Introduction

Principal Component Analysis (PCA) is a powerful statistical technique used to reduce the dimensionality of data while retaining most of its variability. This paper presents a MATLAB script designed to import, normalize, and analyze sales data from an Excel spreadsheet using PCA. The objectives are to demonstrate the step-by-step process of PCA and to provide insights from the resulting principal components.The results demonstrate the effectiveness of PCA in maintaining the essential characteristics of the dataset, making it suitable for further predictive modeling and classification tasks. This work provides a foundational understanding of data prepro- cessing and analysis techniques, with implications for broader applications in machine learning and data science.

AWS (Amazon Web Services) is a leading cloud computing platform that offers a broad range of scalable services including computing power, storage, and databases.It is a widely adopted cloud platform used in computer services, storage services, databases, networking, analytics, machine learning, developer tools, security and identity, managment and monitoring,etc. It provides flexible and cost-effective solutions for building, deploying, and managing applications and services. Key features include virtual servers through EC2, managed databases via RDS, and serverless computing with Lambda. AWS supports global scalability, high availability, and robust security, making it suitable for businesses of all sizes. It also offers tools for analytics, machine learning, and application development.

The integration of PCA with AWS services provides a robust solution for managing and analyzing complex datasets. By leveraging AWS's scalable infrastructure, organizations can efficiently perform dimensionality reduction, enhance data insights, and drive informed decision-making. The combination of PCA and AWS tools offers a powerful approach to tackling the challenges of big data, ultimately advancing research and applications across diverse fields.

Identify applicable funding agency here. If none, delete this.

II. Related Works

AWS (Amazon Web Services) is a comprehensive cloud computing platform provided by Amazon that offers a wide range of services to support various IT needs. It enables businesses to deploy and manage applications and infrastructure in the cloud, providing scalability, flexibility, and cost efficiency.It also provides a flexible and scalable cloud environment suitable for a wide range of use cases, from small startups to large enterprises.Principal Component Analysis (PCA) is widely used in sales data analysis to reduce dimensionality, uncover patterns, and enhance decision-making processes.

A. Data Preprocessing

PCA is frequently employed to reduce the number of features in sales datasets, which often contain numerous variables such as sales volume, product categories, and customer demographics. For instance, one study demonstrated that PCA could simplify complex sales data by extracting the most influential components, making it easier to interpret and visualize trends (Jolliffe, 2002).

B. Machine Learning and Model Training

PCA is often used as a preprocessing step to transform features into a lower-dimensional space, which can help in training more efficient and effective machine learning models. By reducing dimensionality, PCA can help to alleviate the curse of dimensionality and potentially improve model perfor- mance and training times.

C. Sales Forecasting

Research has shown that PCA can improve sales forecasting models by removing redundant or less informative features. By focusing on principal components, models such as ARIMA or regression algorithms can become more accurate and less prone to overfitting (Wold et al., 1987).

D. Customer Segmentation

PCA has been used to identify distinct customer segments by reducing dimensionality in customer behavior data. This approach allows businesses to group customers based on their purchasing patterns and preferences, leading to more targeted marketing strategies (Klein et al., 2012).

III. System Implementation

A. Data Import

The first step involves setting up import options and reading data from the Excel file. The following MATLAB code configures the import options, specifies the sheet and data range, and assigns appropriate variable names and types.Standardization is crucial for PCA as it scales the data so that each feature contributes equally to the analysis.

B. Data Normalization

Next, the numerical variables Sales, Quantity, Discount, and Profit are normalized to the range [0, 1]. Normalization is crucial to ensure that all variables contribute equally to the PCA, preventing variables with larger scales from dominating the analysis.

C. Covariance Matrix Computation

The covariance matrix is computed to understand the relationships between the normalized numerical variables. The covariance matrix provides insights into how the variables vary together.

D. Eigenvalue and Eigenvector Computation

Compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvalues indicate the amount of variance captured by each principal component, while the eigenvectors determine the direction of these components.

E. Sorting and Selecting Principal Components

Sort the eigenvalues and their corresponding eigenvectors in descending order. The eigenvectors corresponding to the largest eigenvalues are the principal components.

F. Projection of Data onto Principal Components

Project the original data onto the selected principal components to obtain a reduced-dimensionality representation.

G. Evaluation of PCA Results

Evaluate the effectiveness of PCA by visualizing the transformed data, analyzing the explained variance ratio, and applying downstream machine learning algorithms or statistical analysis on the reduced dataset.

IV. Results and Analysis

The application of normalization and PCA to the sales dataset has provided a clearer view of the data's structure and relationships. The covariance analysis and principal components offer valuable insights into how sales metrics interact and vary across different dimensions. Future work could focus on further analysis of these components to uncover deeper insights or apply clustering techniques to identify distinct patterns within the data. The dataset, derived from the Excel workbook Book2.xlsx, consists of 19 variables capturing sales transaction details across different geographical regions and time periods. The dataset includes 63 records spanning various countries and regions, with key attributes such as Order ID, Order Date, Customer Information, and sales metrics (Sales, Quantity, Discount, and Profit).

The numerical variables—Sales, Quantity, Discount, and Profit—were normalized to a range of [0, 1]. This normalization was performed to ensure comparability and to prepare the data for further analysis, such as Principal Component Analysis (PCA).These projections reveal how the original data can be represented in a reduced-dimensional space, highlighting the primary patterns and structures within the data.

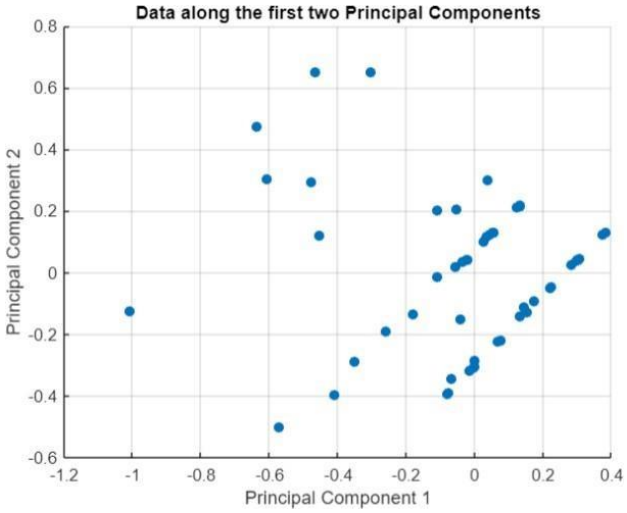

Fig. 1. Data along the first two Principal Components

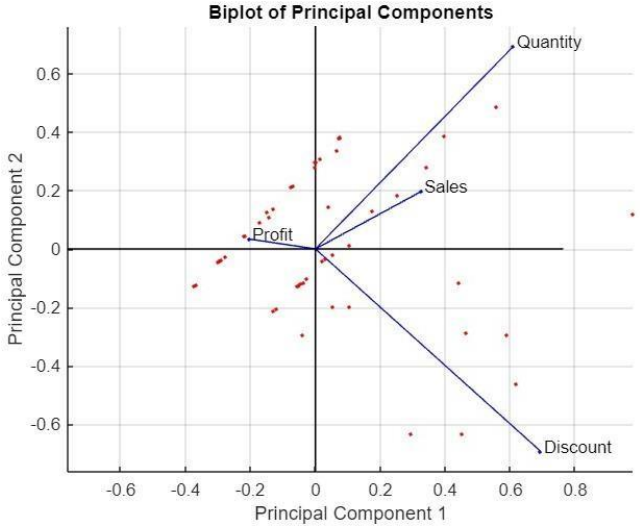

Fig. 2. Biplot of Principal Components

V. ACKNOWLEDGEMENT

Mentor: Dr. Sarada jayan Associate professor and chair Department of mathematics, Amrita School Engineering Amrita Vishwa Vidyapeetham, Bengaluru, Karnataka.

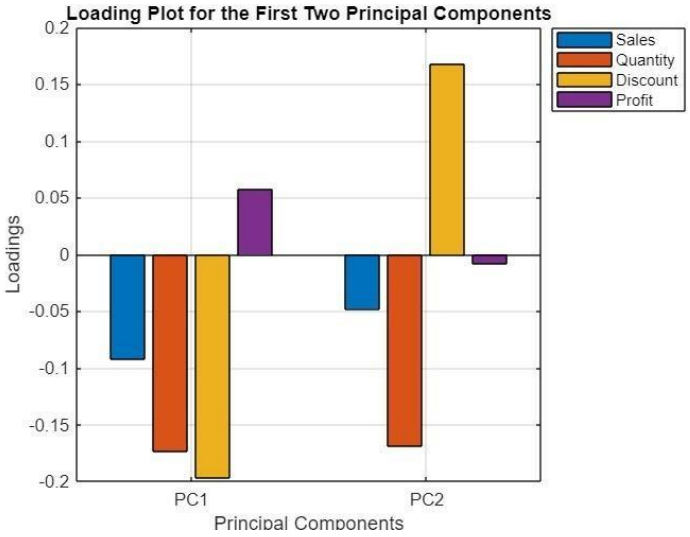

Fig. 3. Loading Plot for the first two Principal Components

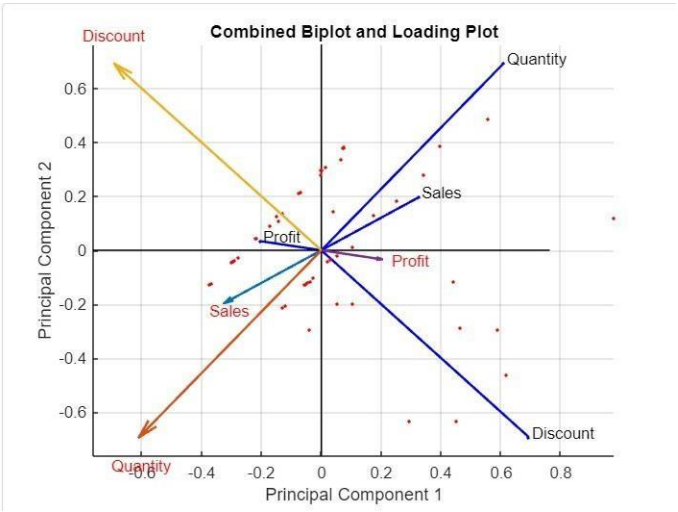

Fig. 4. Combined Biplot and Loading Plot

REFERENCES

[1] K. Pearson, ”On lines and planes of closest fit to systems of points in space,” Philosophical Magazine, Series 6, vol. 2, no. 11, pp. 559-572, 1901.

[2] H. Hotelling, ”Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, vol. 24, no. 6, pp. 417-441, 1933.

[3] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.

[4] J.E. Jackson, ”A User’s Guide to Principal Components”. New York: John Wiley Sons, 1991.

[5] S. Wold, K. Esbensen, and P. Geladi, ”Principal Component Analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37-52, 1987.

[6] H. Abdi and L.J. Williams, ”Principal Component Analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 2010.

[7] C.M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006, pp. 574-596.