HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN 1

BÁO CÁO BÀI TẬP
MÔN HỌC: LẬP TRÌNH PYTHON

Giảng viên hướng dẫn : KIM NGỌC BÁCH
Họ và tên sinh viên : CHU MINH QUANG
Mã sinh viên : B23DCVT354
Lớp : D23CQCE04-B
Nhóm : 01

Hà Nội – 2025

REPORT

Statistical Analysis of Premier League Players 2024–2025 using Python

TABLE OF CONTENTS

1. Introduction

1.1 Rationale

As a student pursuing data science and analytics, I was drawn to this topic because football is both a personal interest and a compelling domain to apply real-world data skills. Football statistics offer a rich, high-dimensional dataset where insights can have actual implications in sports strategy and economics. Through this project, I aim to simulate a mini data analytics pipeline involving data collection, cleaning, analysis, clustering, and value estimation.

1.2 Objectives and Requirements

The main goals of this project are:

- To build a complete pipeline for web-scraping and integrating multiple statistical tables from fbref.com.
- To clean and normalize the resulting dataset for consistency and usability.
- To explore descriptive statistics across players and teams.
- To identify standout players via top/bottom comparisons.
- To cluster players based on performance data and interpret the characteristics of each group.
- To collect external market value data and propose a framework for estimation.

1.3 Tools and Libraries

- Python 3.10 as the programming language
- Selenium for automated browsing and scraping dynamic content
- BeautifulSoup for parsing HTML and handling nested table comments
- Pandas and NumPy for data manipulation
- Matplotlib for plotting histograms and scatter plots
- Scikit-learn for normalization, clustering (KMeans), and dimensionality reduction (PCA)

2. Data Collection (BTL1.py)

The first step was to retrieve player statistics from fbref.com. This site organizes data by category: standard, shooting, passing, possession, defense, and more. These tables are often hidden inside HTML comments, so I built a function to parse both visible and commented-out tables.

After cleaning duplicate entries and ensuring that each row corresponds to an individual player, I merged the data using 'Player' as the key. I then filtered out players with fewer than 90 minutes of playtime to maintain statistical relevance. The final dataset contained 78 standardized metrics for 397 players, which was saved as results.csv.

3. Descriptive Statistical Analysis (BTL2.py)

To begin analyzing, I first converted all string-based statistics to numeric types where applicable. This included removing percent signs, commas, and dealing with 'N/a' values.

3.1 Top/Bottom Performer Analysis

Using Pandas, I extracted the top 3 and bottom 3 players for each metric. For instance:

Top Goalscorer: Mohamed Salah (28 goals)

Top Pass Accuracy: William Saliba (94.2%)

Least Minutes Played (above 90 threshold): Billy Gilmour (98 mins)

These rankings were saved to a file called top_3.txt.

3.2 Team and League Summary Statistics

I calculated mean, median, and standard deviation both across the league and for each team. These insights were exported to results2.csv, which is useful for team comparisons.

3.3 Visualization

For each metric, I plotted histograms showing the distribution of player performance—both for the entire league and split by team. These were saved to the histograms/ folder.

4. Clustering with KMeans and PCA (BTL3.py)

After identifying the numeric columns, I applied StandardScaler to normalize the data. Then I tested k-values from 2 to 10 to evaluate clustering performance.

4.1 Choosing Optimal k

- Elbow Method: The inertia graph showed an inflection at k=3.
- Silhouette Score: Peaked at k=3, indicating strong cohesion and separation.

I finalized k=3 and fit the KMeans model accordingly.

4.2 Dimensionality Reduction with PCA

Using PCA, I reduced the dataset to two principal components. This allowed me to visualize the clusters on a scatter plot (pca_clusters.png) with color-coded labels.

4.3 Cluster Interpretation

Based on cluster centroids (see cluster_comments.txt):

- Cluster 0: Midfielders with balanced contributions across metrics.
- Cluster 1: Attackers with high xG, shots, and touches in the final third.
- Cluster 2: Goalkeepers and defenders with high save percentages and clearances.

5. Transfer Value and Estimation (BTL4.py)

To evaluate economic value, I scraped market values from footballtransfers.com, filtering for the Premier League.

5.1 Value Parsing and Merging

Values were stored as text (e.g., €12.5m), so I converted them to numeric values in euros. I then merged these values with the results.csv dataset, restricting to players with >900 minutes. The merged file is saved as transfer_values.csv.

5.2 Toward a Predictive Model

Although a regression model was not implemented in this iteration, future steps could include using metrics like xG, assists, age, and minutes played as features to train a model that predicts market value. Linear Regression or Random Forests would be suitable starting points.

6. Evaluation and Conclusion

6.1 Key Findings

- Data from multiple fbref tables was successfully merged and filtered into a usable dataset.
- KMeans clustering revealed three meaningful player archetypes based on roles.
- PCA visualization helped confirm cluster separability.
- Top players like Mohamed Salah and Joško Gvardiol appeared in several metrics’ top 3.

6.2 Limitations

- Several players lacked available transfer value data.
- No predictive model was trained to estimate market value.
- Positional encoding, injury history, and salary were not considered due to data availability.

7. Appendix

Below are key visualizations included in the analysis:

- PCA Clusters
- Elbow Method
- Silhouette Score
- results.csv: Consolidated player statistics
- top_3.txt: Metric-wise performance rankings
- results2.csv: Team-based summary stats
- pca_clusters.png, elbow.png, silhouette.png: Diagnostic visualizations
- transfer_values.csv: Combined performance and market data

8. References

https://fbref.com

https://www.footballtransfers.com

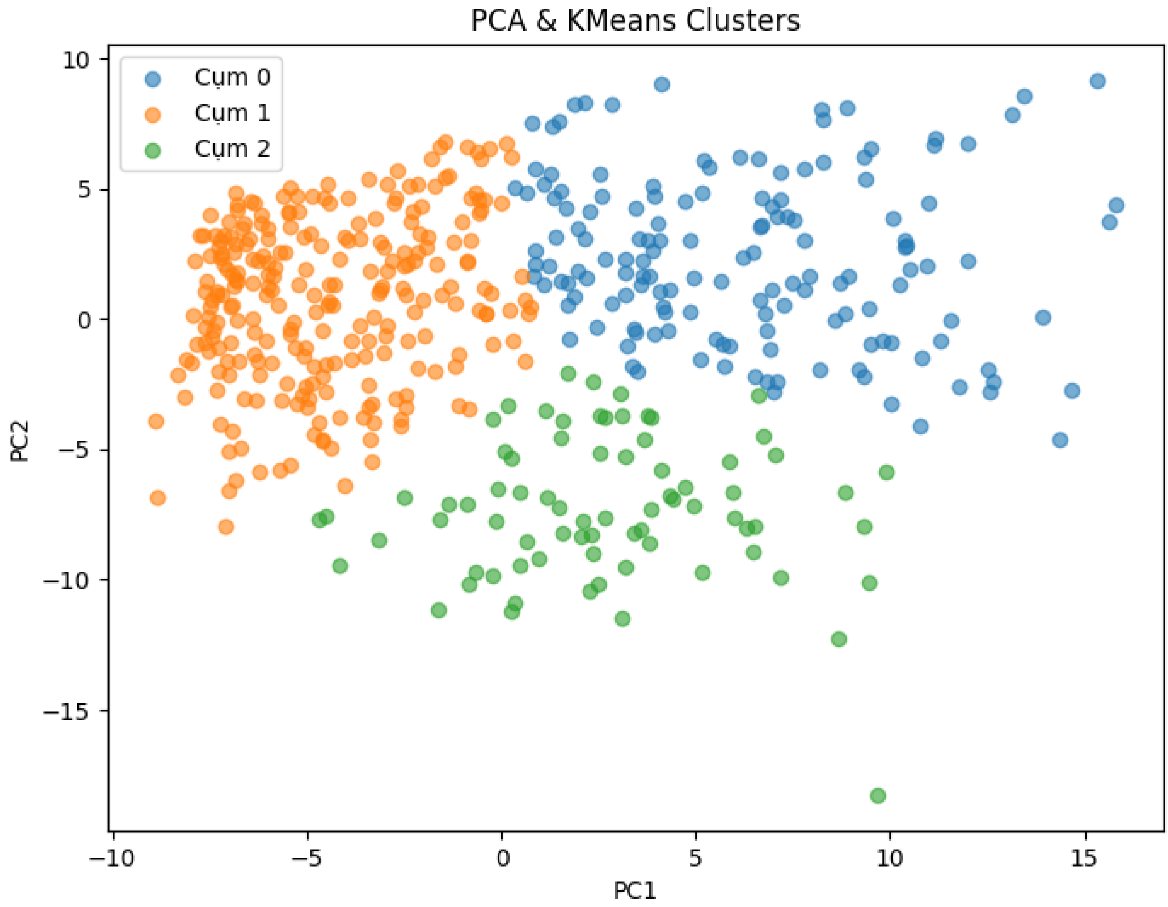
https://scikit-learn.org

https://pandas.pydata.org

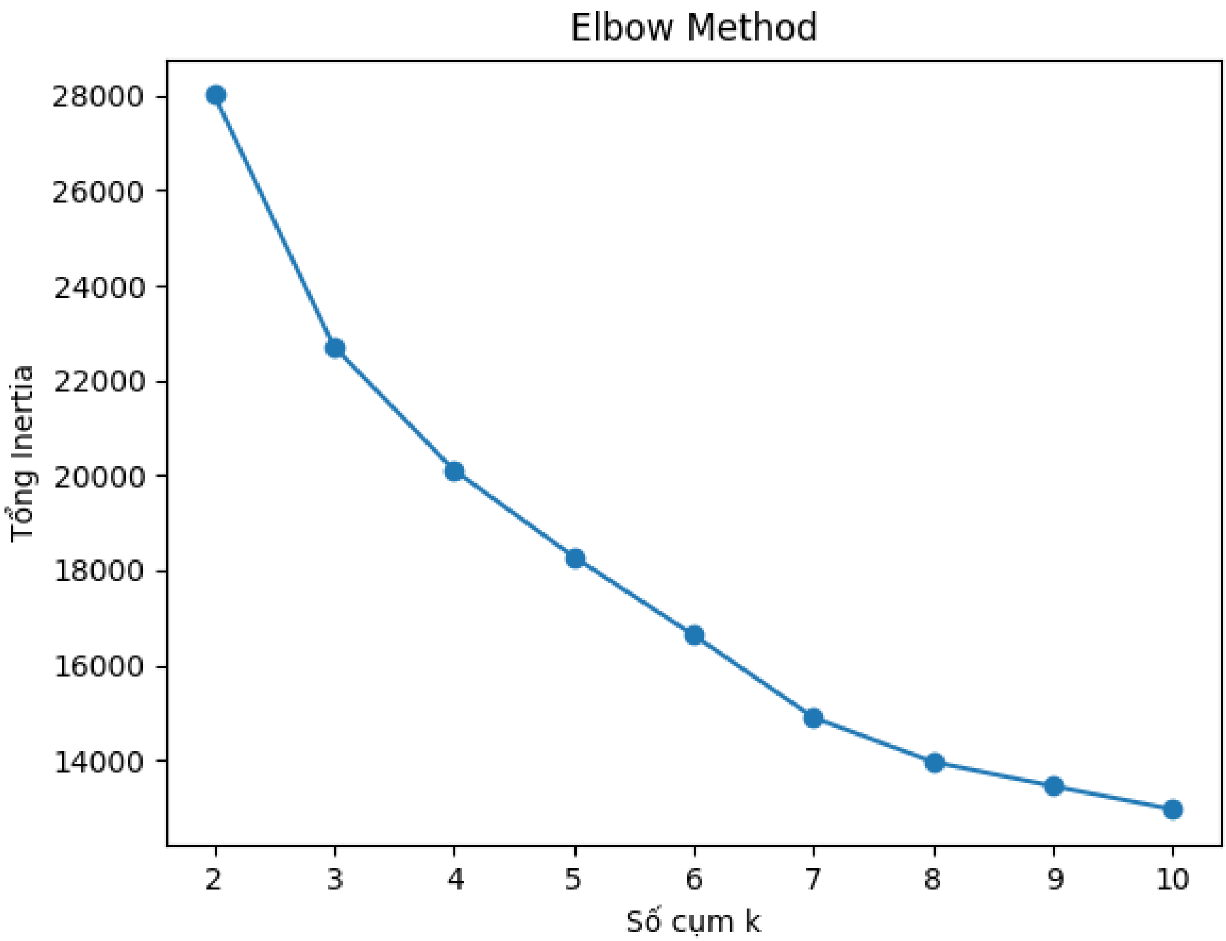
https://matplotlib.org

https://selenium.dev

PCA Clusters



Elbow Method



Silhouette Score