**MACHINE LEARNING AND DATA MINING II**

**LABWORK 1 REPORT**

BI12-130 Võ Trường Giang

BI12-144 Bùi Ngọc Hải The first plot (using the highest variance principal components) shows a well-separated structure, capturing most of the dataset's variability.

The second plot (using the lowest variance principal components) appears more compressed and noisy, as these components contain the least amount of meaningful variation. This confirms that the highest variance principal components are the most informative for analysis and visualization.

* In this labwork, I use the datasets [Best Ever Basketball Players Stats](https://www.kaggle.com/datasets/akulvaishnavi/best-ever-basketball-players-stats) to study and analyze.

1. **Study the dataset**

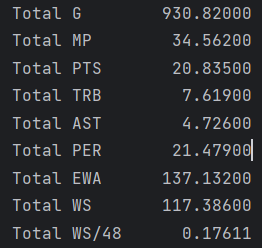
* [Best Ever Basketball Players Stats](https://www.kaggle.com/datasets/akulvaishnavi/best-ever-basketball-players-stats)
* For each dataset, determine which feature is discrete or continuous? quantitative or qualitative? numerical or categorical? Explain.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Discrete/Continuous** | **Quantitative/Qualitative** | **Numerical/Categorical** | **Explanation** |
| **Name** | N/A | Qualitative | Categorical | Names are labels, not numerical values. |
| **Pos** | N/A | Qualitative | Categorical | Positions (e.g., F, C, PG) are categories, not numerical values. |
| **Drafted** | Discrete | Quantitative | Numerical | The year drafted is a discrete numerical value (whole numbers). |
| **Retired** | N/A | Qualitative | Categorical | Binary variable (Y/N), not numerical. |
| **Pick** | Discrete | Quantitative | Numerical | Draft pick number is a discrete numerical value (whole numbers). |
| **Team** | N/A | Qualitative | Categorical | Team names are categories, not numerical values. |
| **Total G** | Discrete | Quantitative | Numerical | Total games played is a discrete numerical value (whole numbers). |
| **Total MP** | Continuous | Quantitative | Numerical | Total minutes played is a continuous numerical value (can be fractional). |
| **Total PTS** | Continuous | Quantitative | Numerical | Total points scored is a continuous numerical value (can be fractional). |
| **Total TRB** | Continuous | Quantitative | Numerical | Total rebounds is a continuous numerical value (can be fractional). |
| **Total AST** | Continuous | Quantitative | Numerical | Total assists is a continuous numerical value (can be fractional). |
| **Total PER** | Continuous | Quantitative | Numerical | Player Efficiency Rating is a continuous numerical value (can be fractional). |
| **Total EWA** | Continuous | Quantitative | Numerical | Estimated Wins Added is a continuous numerical value (can be fractional). |
| **Total WS** | Continuous | Quantitative | Numerical | Win Shares is a continuous numerical value (can be fractional). |
| **Total WS/48** | Continuous | Quantitative | Numerical | Win Shares per 48 minutes is a continuous numerical value (can be fractional). |

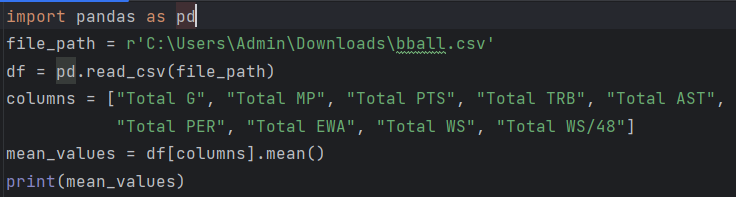
* **What are the issues lying within data? for example: missing data, incorrect data? Do you need to prepare data?**

The dataset appears complete with no missing or incorrect values. However, verifying accuracy against a reliable source is recommended. If missing data were present, we could either remove affected rows or impute values using statistical methods for numerical features and mode or a placeholder for categorical ones.

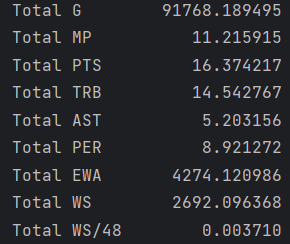
* **Is there any label in the datasets? Explain.**
* **Labels**: The dataset does not have a specific label column that indicates a target variable for prediction. However, if this dataset were used for a predictive task, a label could be created based on a specific feature.
* **Calculate mean, variance, covariance, correlation of the selected datasets. How do you calculate these measures for categorical features (if any)?**
* The mean value of the datasets is:



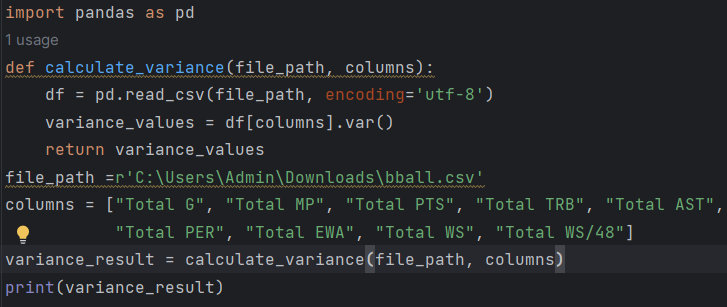
* Code to calculate the mean value:



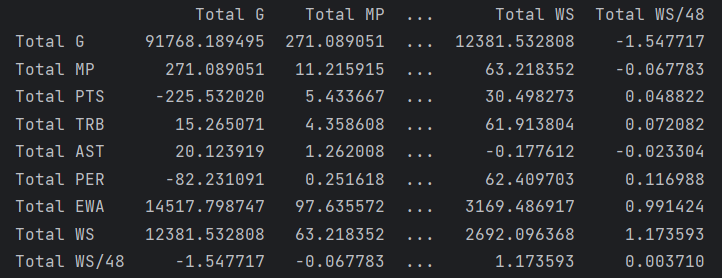
* The variance of the datasets is:



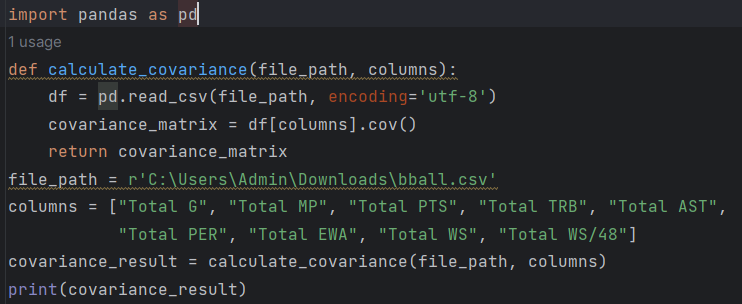
* Code to calculate the variance:



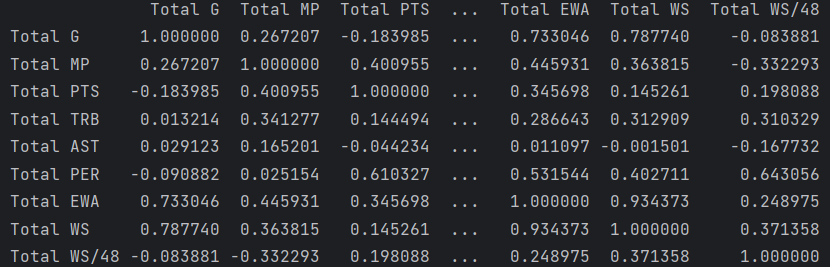
* The covariance of the datasets is:



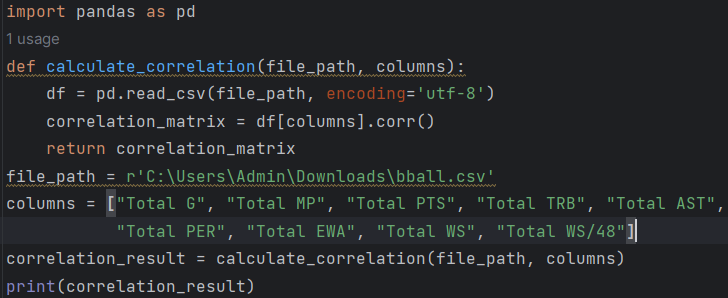
* Code to calculate the covariance:



* The of the correlation datasets is:



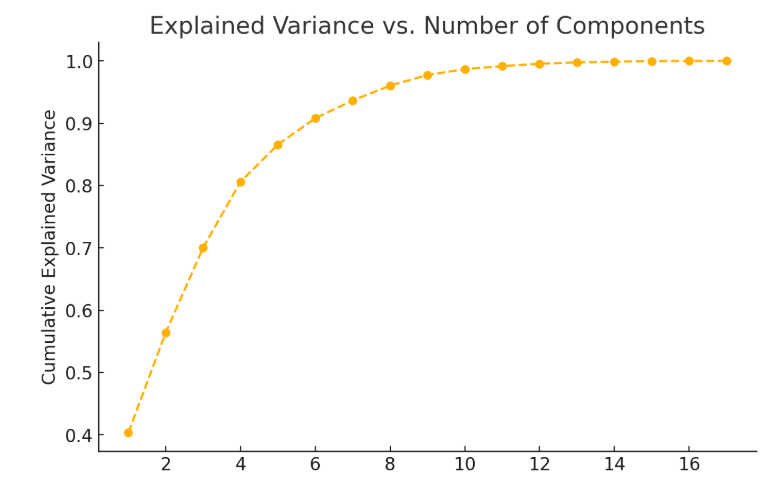
* Code to calculate the correlation:



* **Find the most corelated couple of features of each dataset. Comment on the results**
* Total PTS and Total MP are correlated because players who play more minutes tend to score more points.
* Total TRB and Total AST are also correlated for players who are involved in both rebounding and assisting.
* **Is there any missing data? How do you handle this issue?**
* As mentioned earlier, there doesn't appear to be any missing data in this dataset.

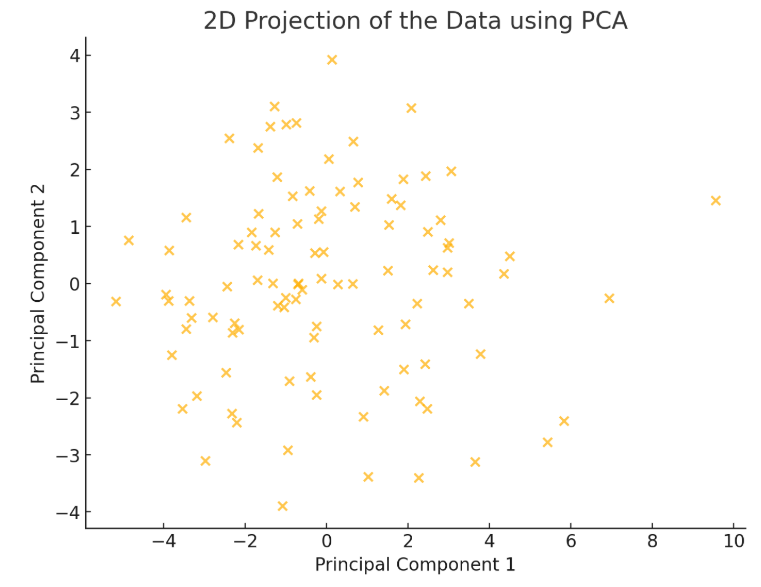
1. **PCA**

* Apply PCA on the two selected datasets. Describe how you select the principal components. Is there any difference while using principal components with highest and lowest values?
* Vary the number of used principal components, analyze and comment on the obtained results.
* Visualize data distribution in 2D.

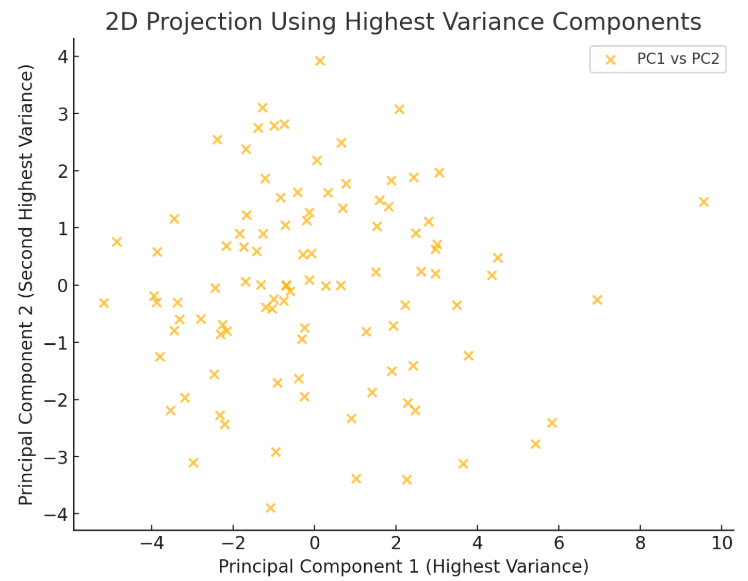
****

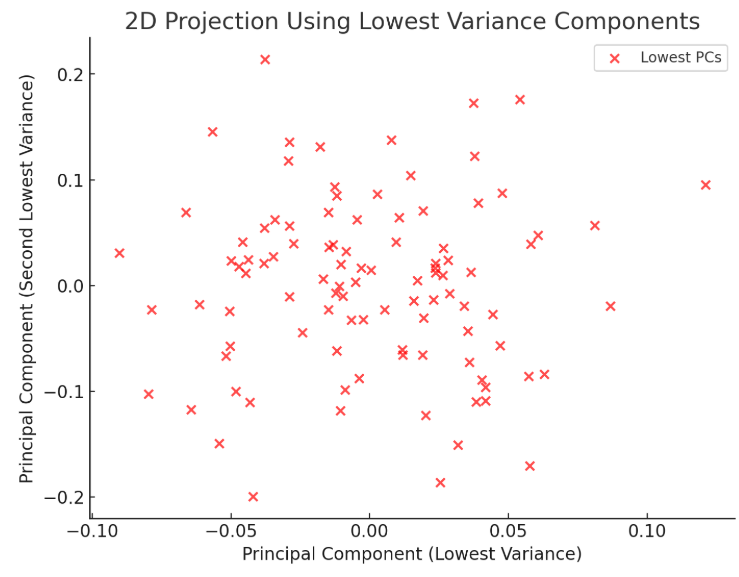
The plot shows the cumulative explained variance as the number of principal components increases.

This helps determine how many components are needed to capture most of the variance in the data.

****

The scatter plot represents the data distribution in 2D using the first two principal components. This projection captures the most variance in the dataset.





The first plot (using the highest variance principal components) shows a well-separated structure, capturing most of the dataset's variability.

The second plot (using the lowest variance principal components) appears more compressed and noisy, as these components contain the least amount of meaningful variation. This confirms that the highest variance principal components are the most informative for analysis and visualization.