Principal component analysis report

Machine learning and Data mining II

Nguyễn Mạnh Hưng - 22BI13183

Nguyễn Trọng Minh - 22BI13304

**TABLE OF CONTENT**

1. **Study the dataset**
2. Choose the dataset
3. Features of the dataset
4. Labels
5. Mathematical properties
6. Missing data
7. **PCA**
8. Apply PCA
9. Principle components
10. Visualize
11. **Study the dataset**
12. Choose the datasets

For this lab work, we chose 2 datasets from Kaggle. One is the dataset of 80 Valorant players in pro league, contains data present their performance on the game. And the second dataset is data about weather (mostly rainy and snowy) recorded in some days from 2006 to 2016, there are total of 12056 rows.



* 1. Valorant players dataset



* 1. Weather dataset

1. Features of the dataset

* For the Valorant players dataset, the “Player” and “Team” columns indicate names and a team, these features are qualitative and discrete because they present properties of the data and can only be some fixed value. The next column – “Rounds played” indicates the number of rounds that player has played in 2023. This feature is discrete and quantitative because it can only be integer. The last 2 columns are “KD” and “Rating”, these stats can be any number in their range, so they are quantitative and continuous.

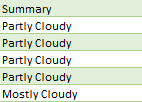


2.1 Features classification for player dataset

* For weather dataset, only the first 2 columns, “Formatted Date” and “Summary” are qualitative and discrete data because they present the time and weather state at that time. The rest of the data set are all quantitative and continuous since they present a value measured at a time.  
  
  1. Features classification for weather data

1. Labels

* For the Valorant Player dataset, there is no label since all the columns are separated and do not need to use any of the other columns to predict or get in conclusion about the data of 1 column.
* For the weather dataset, we can see that there is one column is the label which is “Summary” that uses the data of other columns to predict what the weather will be like in that day.



* 1. The “Summary” column of weather dataset

1. Mathematical properties.

**Weather dataset:**

* Mean and variance: we can calculate the mean and variance of each column as the following table:



* Covariance:

For the weather dataset, we can calculate the covariance of the variables using variance matrix:



* Correlation:

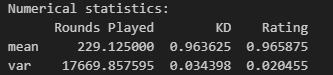
The correlation of the weather dataset is performed by the matrix below:



**Valorant Player dataset**:

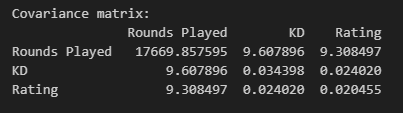
* Mean and variance:

We can calculate the mean and variance of numerical variables of Valorant Player dataset as follow:



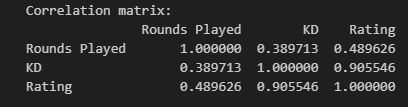
* Covariance:

The covariance matrix of the dataset:



* Correlation:

We calculated the correlation matrix of the dataset:



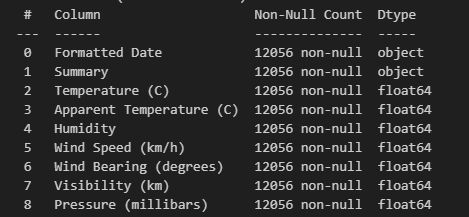
For categorical features we can still calculate the mean by using mode function, but we cannot calculate the variance, covariance, and correlation between these features because they don’t have natural ordering or numerical values for the calculation.

As we can see in the correlation matrix of the two datasets. We can simply see that in the weather dataset, the most corelated couple of features is “apparent Temperature” and “Temperature” with the correlation of 0.99. This is normal phenomenal since they are almost the same in term of temperature.

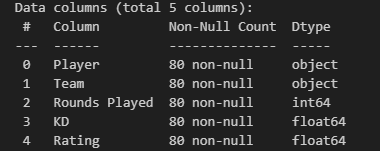
- Regarding the Valorant Player dataset, the most corelated couple of features is “KD” with “Rating” with the correlation of 0.9. This result is also common sense since the better K/D ratio, the better rating that player get.

5.Missing Data

- Using the **info()** function in python, we can see that there are no null data in both datasets.



5.1 Weather dataset

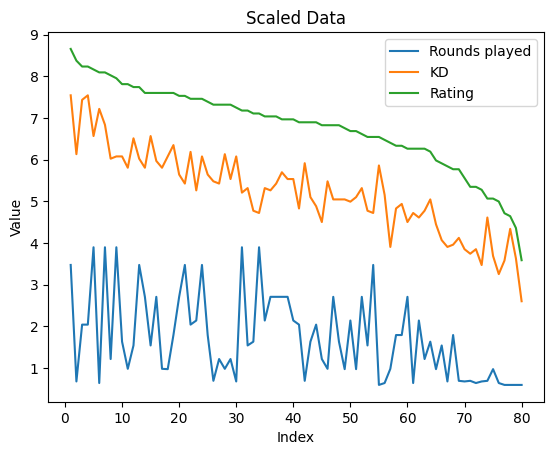


5.2. Valorant Player dataset

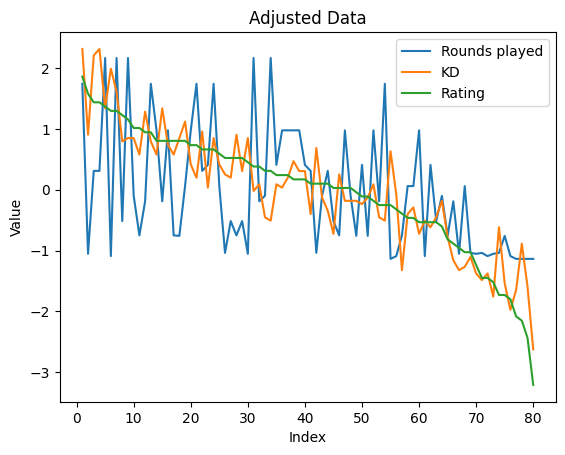
But in case that there are some missing data in the dataset we can handle this by some of the following solution:

* Check the data in the collection source: we can go back to the source where we collected our data and find the one that we need.
* Drop the missing value: we can drop rows of the data that contain the missing value.
* Replace missing value: besides dropping the data, we can also replace the missing one with the value that we choose, normally with numerical data, we can replace the missing data with the average value of the column, or with the categorical data, we can replace it with the data that is mode in the column.
* Leave it empty: we can also let the data empty.

1. **PCA – Principal Component Analysis**
2. Apply PCA into the datasets
   * For the Valorant player dataset, we load in the csv file as pandas Data Frame.
   * Then, we drop 3 first columns since they are categorical variables.
   * Because the data in 3 numerical categories are not in the same range, we have to scale them so that one with large number doesn’t account for majority of the ratio between each category. For example, “KD” and “Rating” can only be around 1 to 2 while “Round played” gets up to 400.
   * We can simply do this by divide each value for their standard deviation, so that large value reduces more than small value.



* The next step is to center the data by subtract them with their mean value.



* We continue by calculate covariance matrix from the centered data.