University of Science and Technology of Hanoi

Principal component analysis report

Machine learning and Data mining II

Nguyễn Mạnh Hưng - 22BI13183

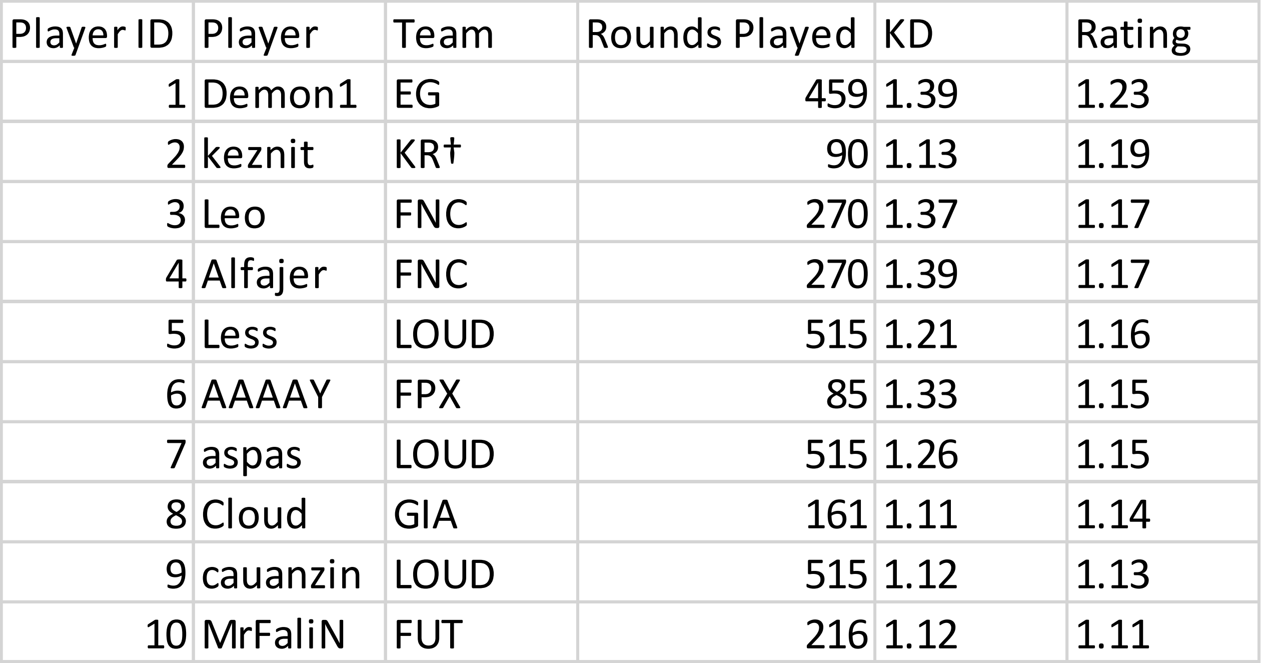
Nguyễn Trọng Minh - 22BI13304

TABLE OF CONTENT

I. Study the dataset

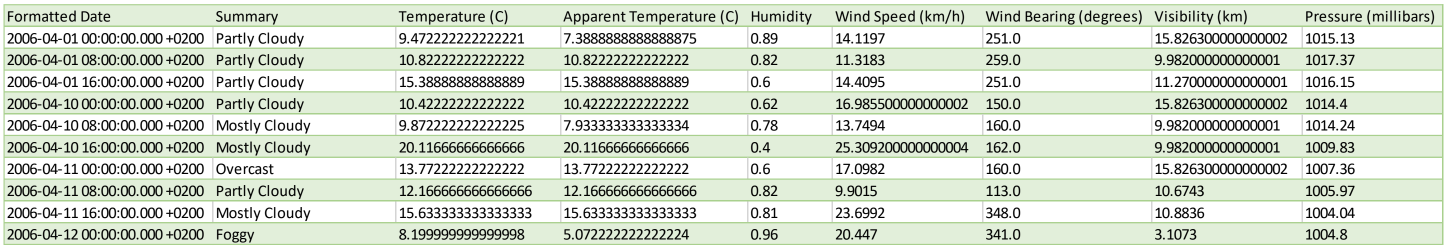
1. Choose the datasets

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.



Player ID	Player	Team	Rounds Played	KD	Rating
1	Demon1	EG	459	1.39	1.23
2	keznit	KR†	90	1.13	1.19
3	Leo	FNC	270	1.37	1.17
4	Alfajer	FNC	270	1.39	1.17
5	Less	LOUD	515	1.21	1.16
6	AAAAY	FPX	85	1.33	1.15
7	aspas	LOUD	515	1.26	1.15
8	Cloud	GIA	161	1.11	1.14
9	cauanzin	LOUD	515	1.12	1.13
10	MrFaliN	FUT	216	1.12	1.11

1.1 Valorant players dataset



Formatted Date	Summary	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Pressure (millibars)
2006-04-01 00:00:00.000 +0200	Partly Cloudy	9.472222222222221	7.3888888888888875	0.89	14.1197	251.0	15.826300000000002	1015.13
2006-04-01 08:00:00.000 +0200	Partly Cloudy	10.822222222222222	10.822222222222222	0.82	11.3183	259.0	9.982000000000001	1017.37
2006-04-01 16:00:00.000 +0200	Partly Cloudy	15.38888888888889	15.38888888888889	0.6	14.4095	251.0	11.270000000000001	1016.15
2006-04-10 00:00:00.000 +0200	Partly Cloudy	10.422222222222222	10.422222222222222	0.62	16.985500000000002	150.0	15.826300000000002	1014.4
2006-04-10 08:00:00.000 +0200	Mostly Cloudy	9.872222222222225	7.933333333333334	0.78	13.7494	160.0	9.982000000000001	1014.24
2006-04-10 16:00:00.000 +0200	Mostly Cloudy	20.11666666666666	20.11666666666666	0.4	25.309200000000004	162.0	9.982000000000001	1009.83
2006-04-11 00:00:00.000 +0200	Overcast	13.77222222222222	13.77222222222222	0.6	17.0982	160.0	15.826300000000002	1007.36
2006-04-11 08:00:00.000 +0200	Partly Cloudy	12.166666666666666	12.166666666666666	0.82	9.9015	113.0	10.6743	1005.97
2006-04-11 16:00:00.000 +0200	Mostly Cloudy	15.633333333333333	15.633333333333333	0.81	23.6992	348.0	10.8836	1004.04
2006-04-12 00:00:00.000 +0200	Foggy	8.199999999999998	5.072222222222224	0.96	20.447	341.0	3.1073	1004.8

1.2 Weather dataset

2. 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.

	Discrete	Continuous	Quantitative	Qualitative	Numerical	Categorical
Player	x			x		x
Team	x			x		x
Rounds Played	x		x		x	
KD		x	x		x	
Rating		x	x		x	

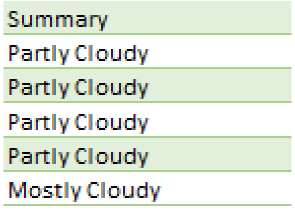
2.1 Features classification for player dataset

- 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.

	Discrete	Continuous	Quantitative	Qualitative	Numerical	Categorical
Date	x			x	x	
Summary	x			x		x
Temperature		x	x		x	
Apparent Temperature		x	x		x	
Humidity		x	x		x	
Wind Speed		x	x		x	
Wind Bearing		x	x		x	
Visibility		x	x		x	
Pressure		x	x		x	

2.2 Features classification for weather data

3. 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.



Summary
Partly Cloudy
Partly Cloudy
Partly Cloudy
Partly Cloudy
Mostly Cloudy

3.1 The “Summary” column of weather dataset

4. Mathematical properties.

Weather dataset:

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

	Mean	Var
Temperature	12.07	91.73176185
Apparent Temperature	10.97	114.7639079
Humidity	0.73	0.038266621
Wind Speed	10.98	44.44204563
Wind Bearing	190.42	10936.21513
Visibility	10.44	18.05168174
Pressure	1016.70	60.07562734

- Covariance:

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

	Temperature	Apparent Temperature	Humidity	Wind Speed	Wind Bearing	Visibility	Pressure
Temperature	91.73176185	101.8682828	-1.202437303	1.21531412	26.25978635	16.088	-23.19171987
Apparent Temperature	101.8682828	114.7639079	-1.283593887	-3.176705674	30.79778766	17.56712	-24.16582573
Humidity	-1.202437303	-1.283593887	0.038266621	-0.285053587	0.157402448	-0.3138	0.082599377
Wind Speed	1.21531412	-3.176705674	-0.285053587	44.44204563	59.35524969	3.172369	-14.19607629
Wind Bearing	26.25978635	30.79778766	0.157402448	59.35524969	10936.21513	24.16369	-62.0106811
Visibility	16.08800329	17.56711851	-0.313801672	3.172368748	24.16369168	18.05168	-5.316683493
Pressure	-23.19171987	-24.16582573	0.082599377	-14.19607629	-62.0106811	-5.31668	60.07562734

- Correlation:

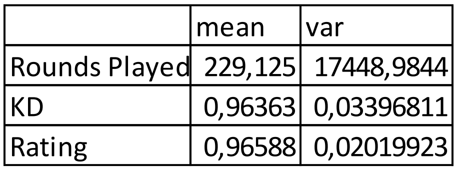
The correlation of the weather dataset is performed by the matrix below:

	Temperature	Apparent Temperature	Humidity	Wind Speed	Wind Bearing	Visibility	Pressure
Temperature	1	0.992833571	-0.641789364	0.01903407	0.026217894	0.395352	-0.312409284
Apparent Temperature	0.992833571	1	-0.612512446	-0.044481282	0.027490544	0.385957	-0.291038081
Humidity	-0.641789364	-0.612512446	1	-0.218584535	0.007694278	-0.37756	0.054477548
Wind Speed	0.01903407	-0.044481282	-0.218584535	1	0.08513901	0.112003	-0.27474016
Wind Bearing	0.026217894	0.027490544	0.007694278	0.08513901	1	0.054384	-0.076503943
Visibility	0.395351549	0.385957098	-0.377560334	0.112002524	0.054383973	1	-0.161448054
Pressure	-0.312409284	-0.291038081	0.054477548	-0.27474016	-0.076503943	-0.16145	1

Valorant Player dataset:

- Mean and variance:

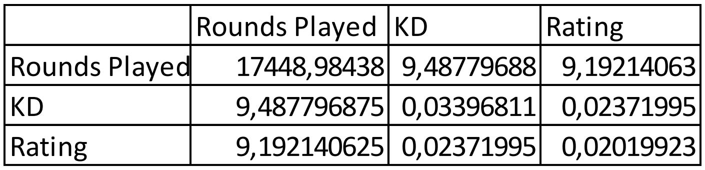
We can calculate the mean and variance of numerical variables of Valorant Player dataset as follow:



	mean	var
Rounds Played	229,125	17448,9844
KD	0,96363	0,03396811
Rating	0,96588	0,02019923

- Covariance:

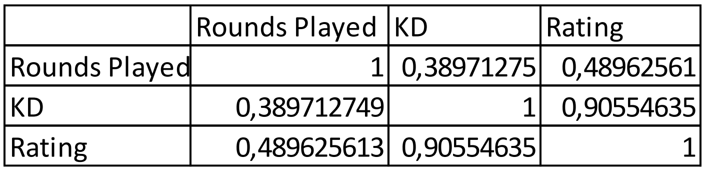
The covariance matrix of the dataset:



	Rounds Played	KD	Rating
Rounds Played	17448,98438	9,48779688	9,19214063
KD	9,487796875	0,03396811	0,02371995
Rating	9,192140625	0,02371995	0,02019923

- Correlation:

We calculated the correlation matrix of the dataset:



	Rounds Played	KD	Rating
Rounds Played	1	0,38971275	0,48962561
KD	0,389712749	1	0,90554635
Rating	0,489625613	0,90554635	1

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.

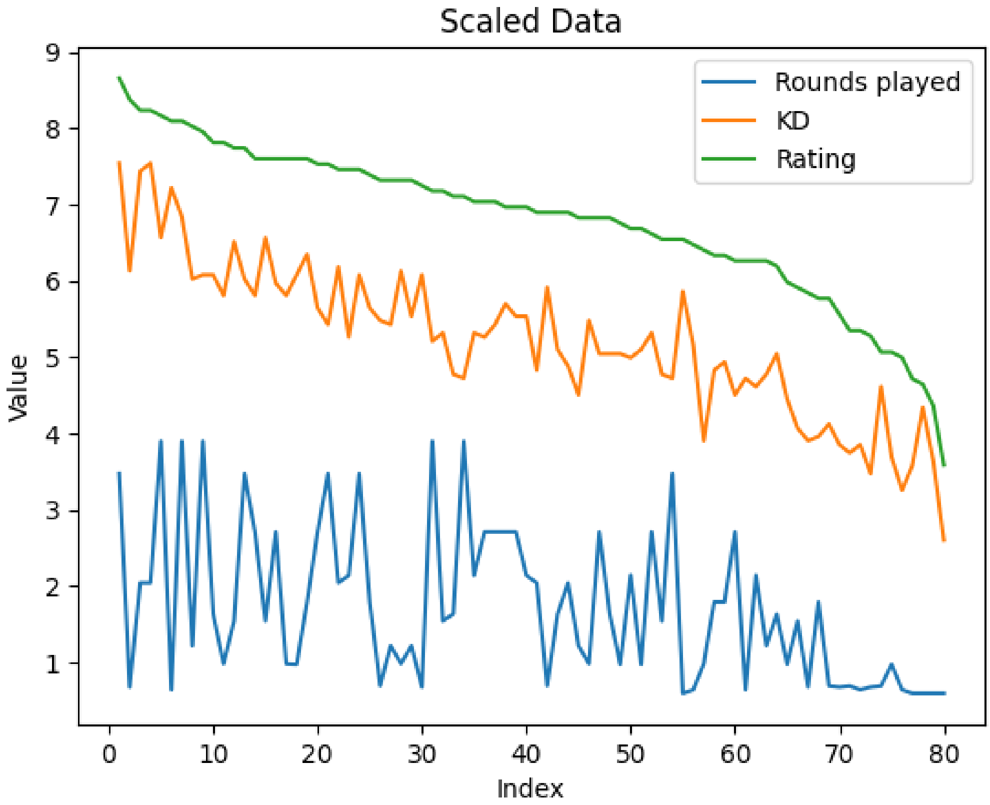
5.Missing Data

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

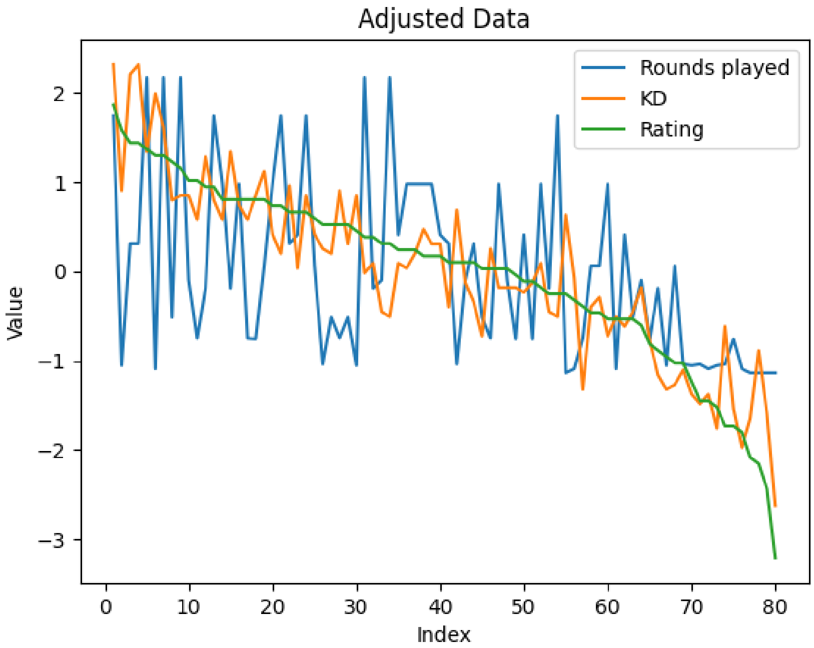
II. PCA – Principal Component Analysis

1. Apply PCA into the Valorant players datasets

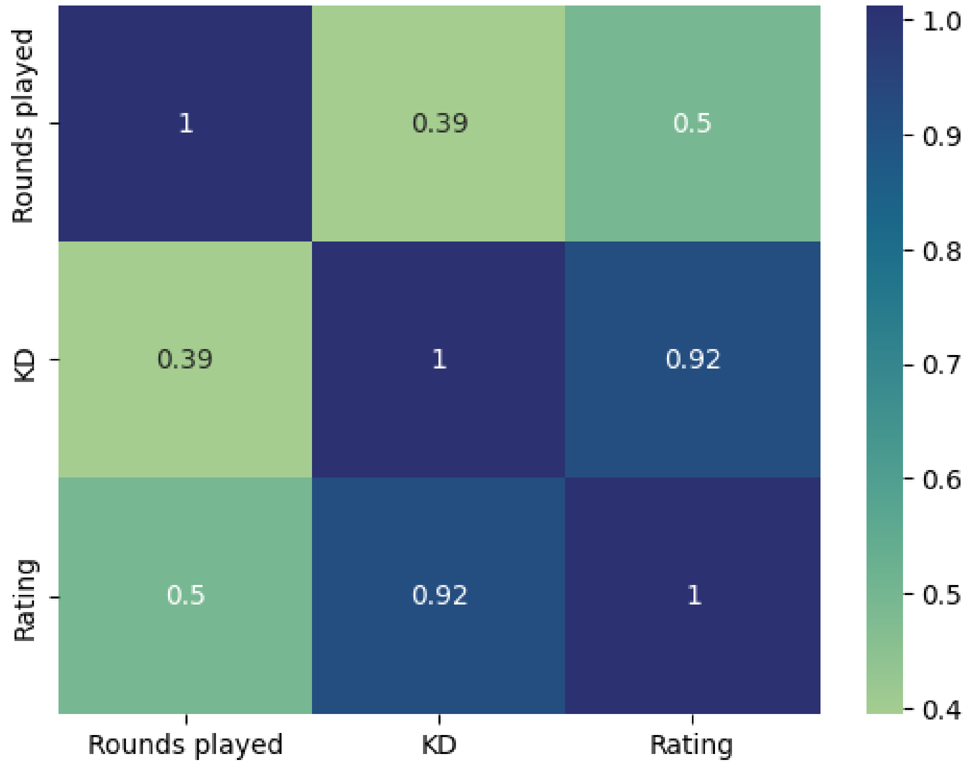
- For the Valorant players dataset, we load in the csv file as pandas Data Frame.
- Then, we drop 3 first columns since they are categorical variables.
- Because the data in 3 numerical categories are not in the same range, we have to scale them so that one with large number doesn’t account for majority 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.

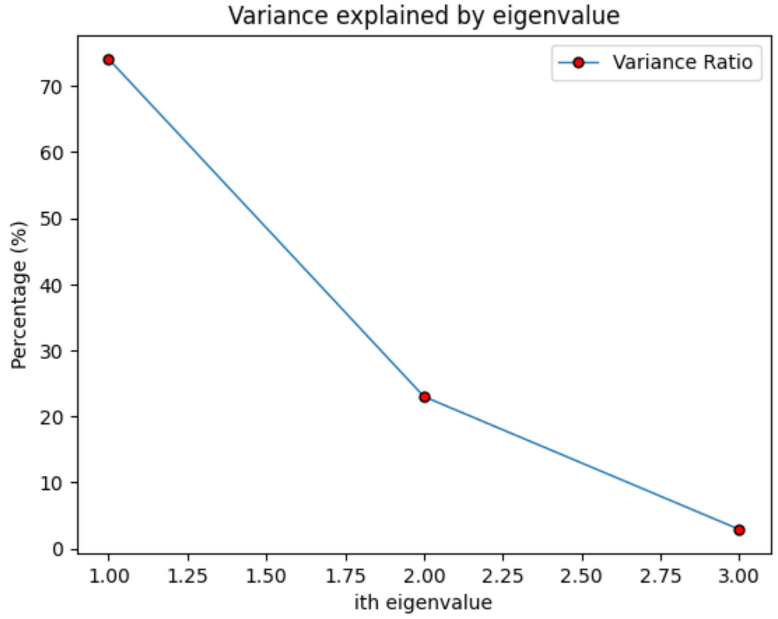


- ○ We continue by calculate covariance matrix from the centered data.

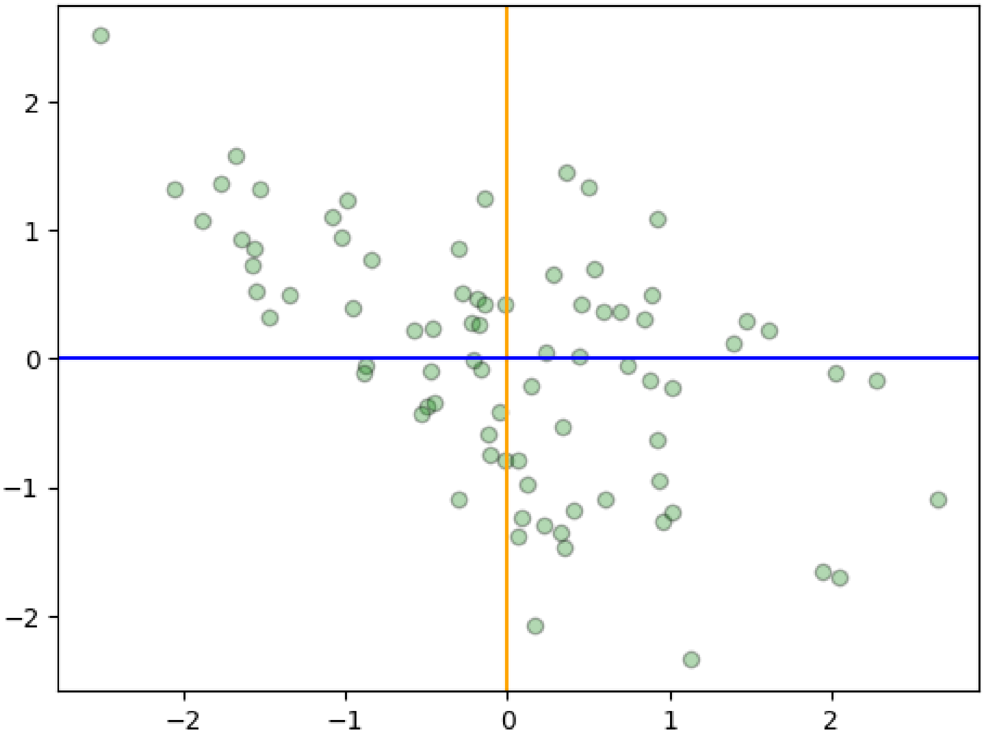


Covariance matrix of the dataset

- ○ And follow up by compute eigenvalues and eigenvectors of that covariance matrix above.
- ○ From that, we can plot a scree graph that presents the variance explained by eigenvalue.



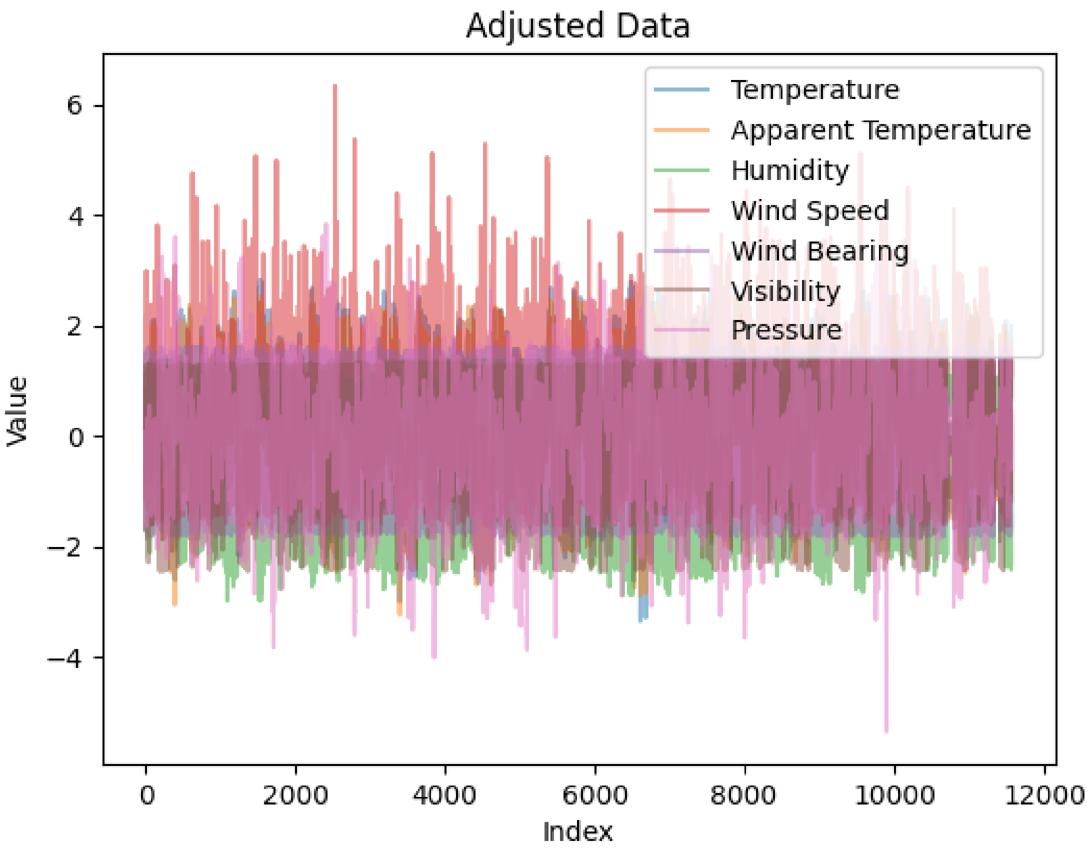
- We can see that the first and second components are the most important. They can present 97% of the data. These 2 will be used to form a new space.
- Because eigenvector present a proportion of original data, we must use the highest value from this graph to choose principal components, or else we will lose most of the data.
- Last step is to project the adjusted data onto the new space, and we got this nicely centered graph.



Since we used 2 principal components that accounted for 97% of the data, the result is accurate if we compare to the original data.

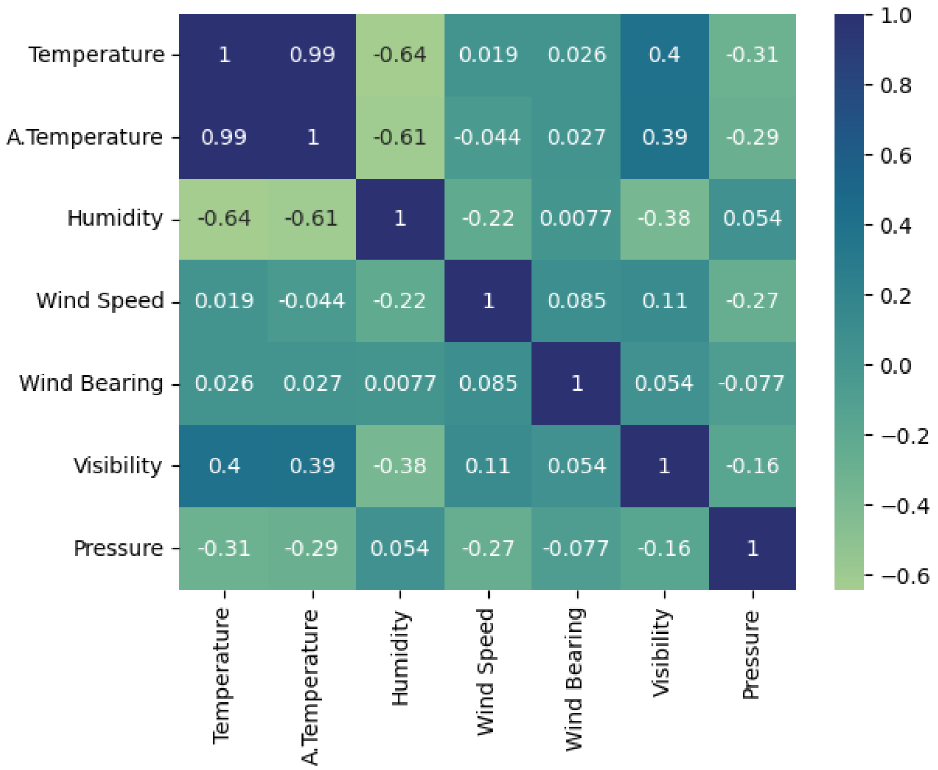
2. Apply PCA into weather dataset

- For the weather dataset, the process is almost the same, just with different number of dimensions.
- We loaded the csv file into data frame and then dropped off all non-numerical variables.
- Then we performed data scaling and adjust them to the center of gravity.

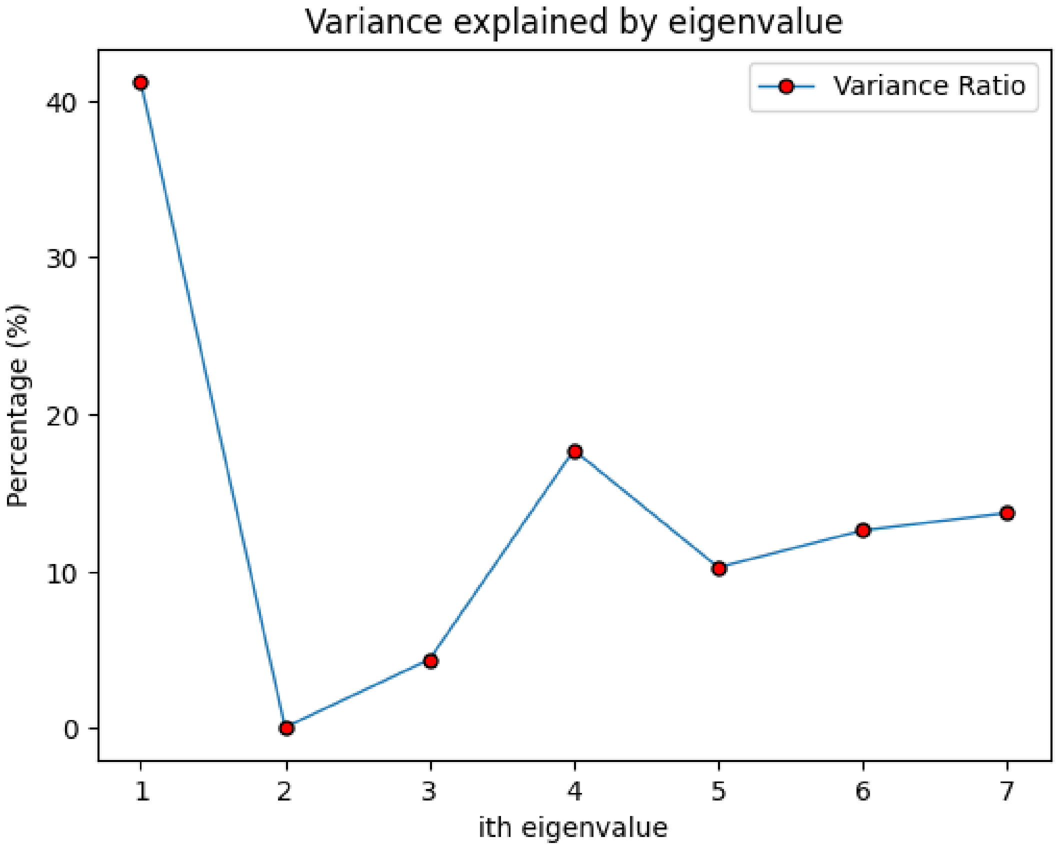


There is a huge amount of data in this dataset

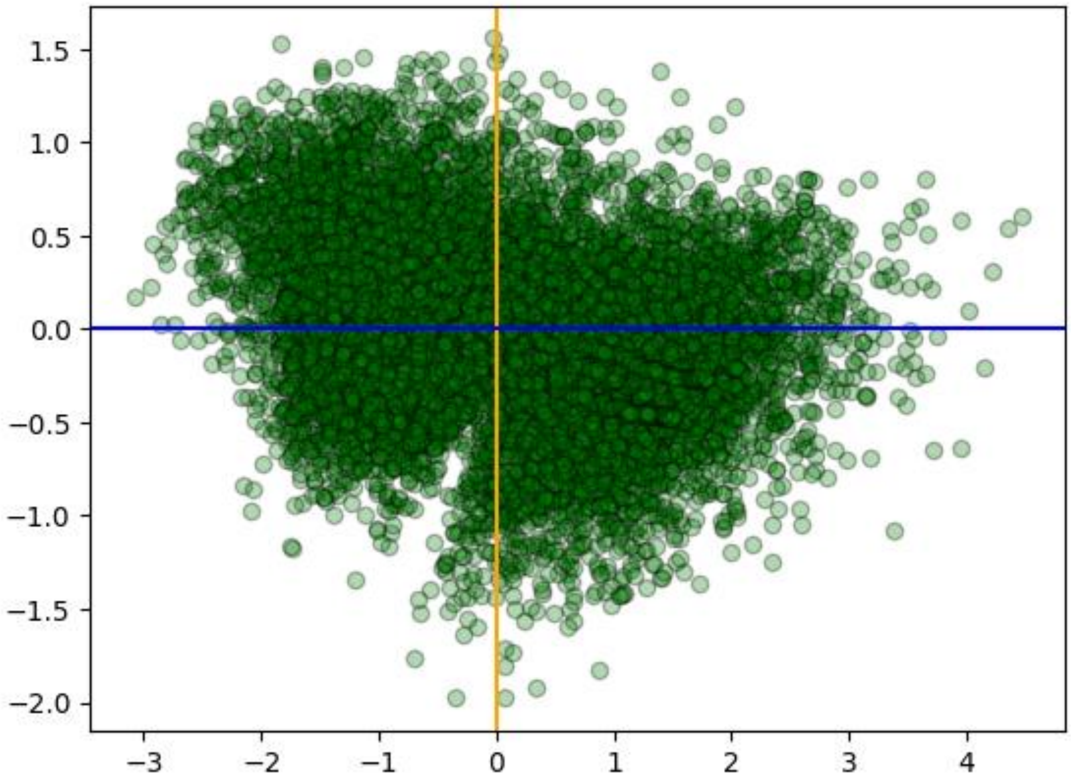
- Next, we computed the covariance matrix for this centered data.



- In the following step, we calculated eigenvalue and eigenvectors of the covariance matrix, and got this scree plot of variance explained by eigenvalue.



- As you can see, we can see that eigenvalue number 1 and 4 accounted for the highest amount of proportion. We simply pick these 2 for dimensional deduction and plot a nicely centered data from the original in 2D.



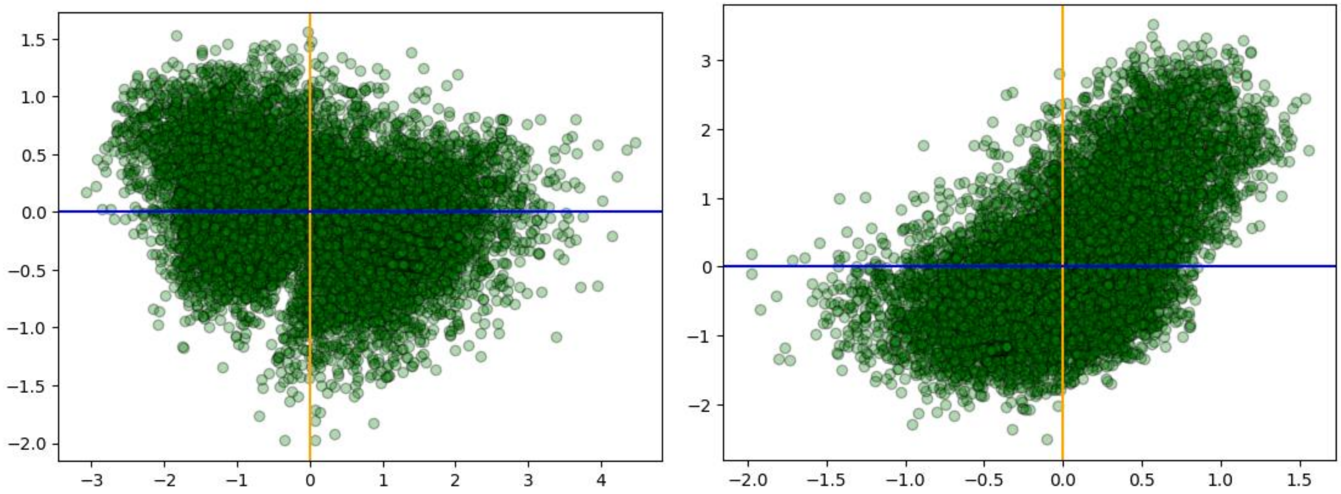
2D representation of weather dataset

Since we used 2 principal components that only present 58.9% of the dataset, this graph is not very accurate, but this is the best we can perform on this dataset with PCA.

3. Vary the components

Now we can try vary the principal components on both datasets to see what will happen.

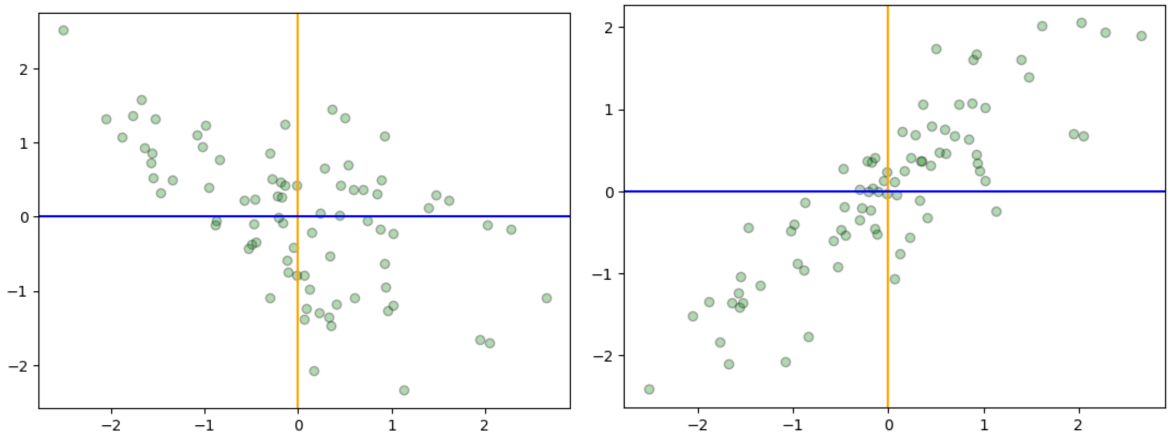
For the first attempt, I chose the 2 least proportion eigenvectors in the weather dataset. And plotted this 2D graph of the original dataset:



Before and after vary the principal components

As we can see, the data from 2 pictures is far different from each other. But for the left figure, it only presents 4.4% of the dataset, this is considered to be wrong.

For the second attempt, I chose 2 principal components with highest and lowest value to represent the data.



Before and after principal components adjustment

We can see that by changing 1 principal component, the plotted dataset rotated 90 degrees since these eigenvectors are orthogonal. However, we’ve just only reduced accuracy from 97% to 76%, the result is still remained some of the clustered areas.