Data mining project

Clustering and anomaly detection in wine Quality

Data Mining is the process of extracting knowledge from data which is done by combining various Machine learning

The wine quality dataset includes 2 datasets which are red and white wine datasets of the Portuguese "Vinho Verde" wine. For my project, I am only using white wine quality dataset, as white wine dataset has more instances and hence, it probably should help me makeup a better model.

The data features deals with describing the physiochemical properties of the wine while data also has an output label which describes the quality of wine on the scale of 1-10.

Chapter 1

1.1 Importing libraries

In this project, I have used various libraries like pandas, matplotlib, seaborn, scikit learn etc to perform various data analysis methods on white wine quality dataset hosted on UCI Machine Learning website. Libraries like matplotlib, seaborn are used for multidimensional visualisation of the data while libraries like scikit learn are used to apply various machine learning algorithms like PCA, K-means on the data.

1.2 Basics of the data

The data contains 4898 instances alongwith 11 features describing physiochemical properties of the wine while the last column deals with describes the Quality of the wine on the range of 1-10. After loading my dataset, the first thing I did was to remove the “Quality” label and also renaming all the column headers as column_0 to column_1 as it is meant to be a purely unsupervised task. But before removing the labels, in order to prevent repetition of identical values, I dropped off all the duplicate data.

“Duplicate data refers to the presence of identical or nearly identical records or observations within a dataset, indicating redundancy or repetition.”

Causes: Duplicate data may happen when we mistakenly enter the same information multiple times as a result of human error. This made me remove it.

1.3 Determine datatypes, using methods like head(), tail()

Next came, getting more general information about the data. All the features of my data are Quantitive continuous which can be seen using .head(), .tail(), .info() of pandas dataframe. Using methods like tail and head are always great way to start as they help us to get the very first glimpse of our data.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3961 entries, 0 to 4897
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   column_0   3961 non-null   float64
 1   column_1   3961 non-null   float64
 2   column_2   3961 non-null   float64
 3   column_3   3961 non-null   float64
 4   column_4   3961 non-null   float64
 5   column_5   3961 non-null   float64
 6   column_6   3961 non-null   float64
 7   column_7   3961 non-null   float64
 8   column_8   3961 non-null   float64
 9   column_9   3961 non-null   float64
 10  column_10  3961 non-null   float64
dtypes: float64(11)
memory usage: 371.3 KB
```

Figure 1info about data after duplicates were removed

1.4 Univariate Statistics

Univariate statistics is the most simple way to start analyzing data. It refers to the analysis and summary of data involving a single variable at a time. It focuses on examining the distribution and characteristics of individual variables independently, without considering the relationships or interactions with other variables. Common univariate statistical measures include measures of central tendency (such as mean, median, and mode), measures of dispersion (such as range, variance, and standard deviation), and frequency distributions. Univariate analysis provides insights into the characteristics and patterns of a single variable within a dataset.

For this, the describe methods puts everything all at once in front of you. The describe method in pandas provides a summary of descriptive statistics for a DataFrame, including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each numerical column.

	column_0	column_1	column_2	column_3	column_4	column_5	column_6	column_7	column_8	colu
count	3961.000000	3961.000000	3961.000000	3961.000000	3961.000000	3961.000000	3961.000000	3961.000000	3961.000000	3961.0
mean	6.839346	0.280538	0.334332	5.914819	0.045905	34.889169	137.193512	0.993790	3.195458	0.4
std	0.866860	0.103437	0.122446	4.861646	0.023103	17.210021	43.129065	0.002905	0.151546	0.1
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.2
25%	6.300000	0.210000	0.270000	1.600000	0.035000	23.000000	106.000000	0.991620	3.090000	0.4
50%	6.800000	0.260000	0.320000	4.700000	0.042000	33.000000	133.000000	0.993500	3.180000	0.4
75%	7.300000	0.330000	0.390000	8.900000	0.050000	45.000000	166.000000	0.995710	3.290000	0.5
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.0

Watching this table, we can interpret the data is already quite standardized and doesn't really requires any further standardization or normalization which will be proved further visually.

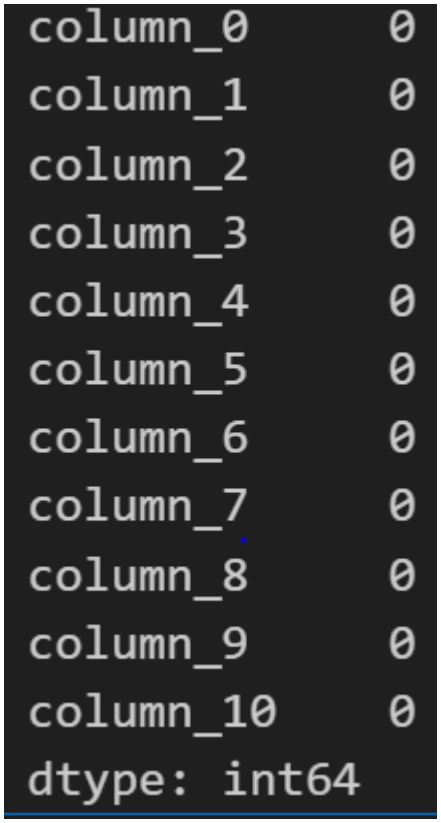


Figure 2 no missing values

isna.sum() of dataframe is like a king detector for any missing values. It simply checks all the columns for the presence of missing values and add them up and print them like a pandas series.

1.5 Distribution of data

For further analysis, Visual Univariate analysis is done by plotting the histograms of all the features of our data. Histogram is a great method to look at the distribution of data and not only data, it can also help in giving some hidden hints about the data. Like here, looking at the histogram, the empty spaces on the right side hints at the presence of some outliers in the data. Moreover, we can clearly observe many of our features have a rightly skewed distribution.

“Skewness measures the asymmetry of a distribution. A negative skewness indicates that the distribution is skewed to the left (tail on the left), while a positive skewness indicates a skew to the right (tail on the right).”

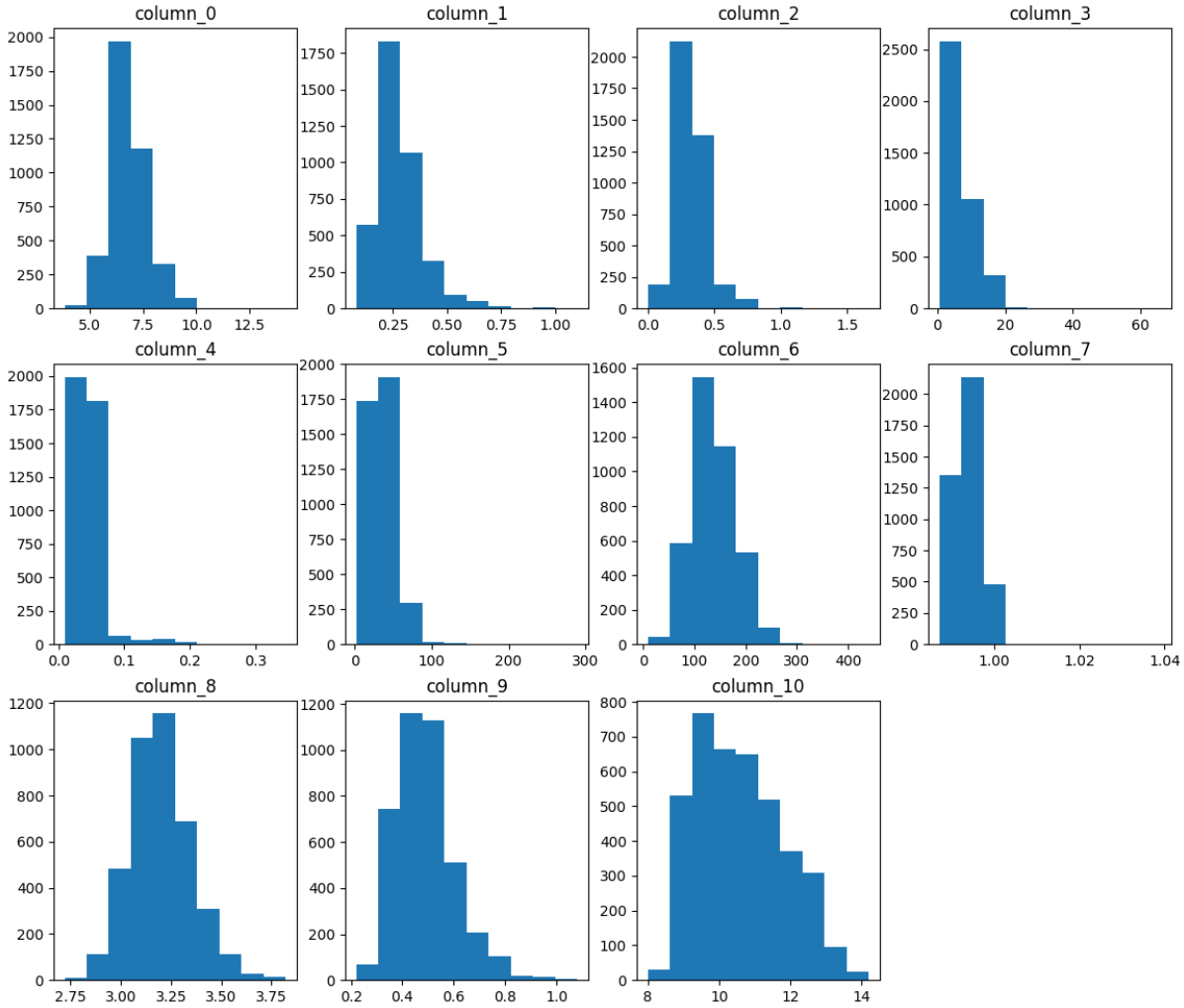


Figure 3 Histograms of all attributes

Skeweness is one of the statistical measures of the distribution which tells us about the shape of the distribution. It determines the symmetry of dataset. Symmetricity of Distribution: A distribution is called symmetric if Mean=Median=Mode otherwise non-symmetric. Another one is called kurtosis,

"Kurtosis measures the "tailedness" of a distribution. It indicates whether the tails of a distribution are heavier or lighter than a normal distribution."

skewness and kutosis of the data can be find out using the stats module of sklearn and furthermore, the distribution can be viewed even better with the help of kde plots from the seaborn library.

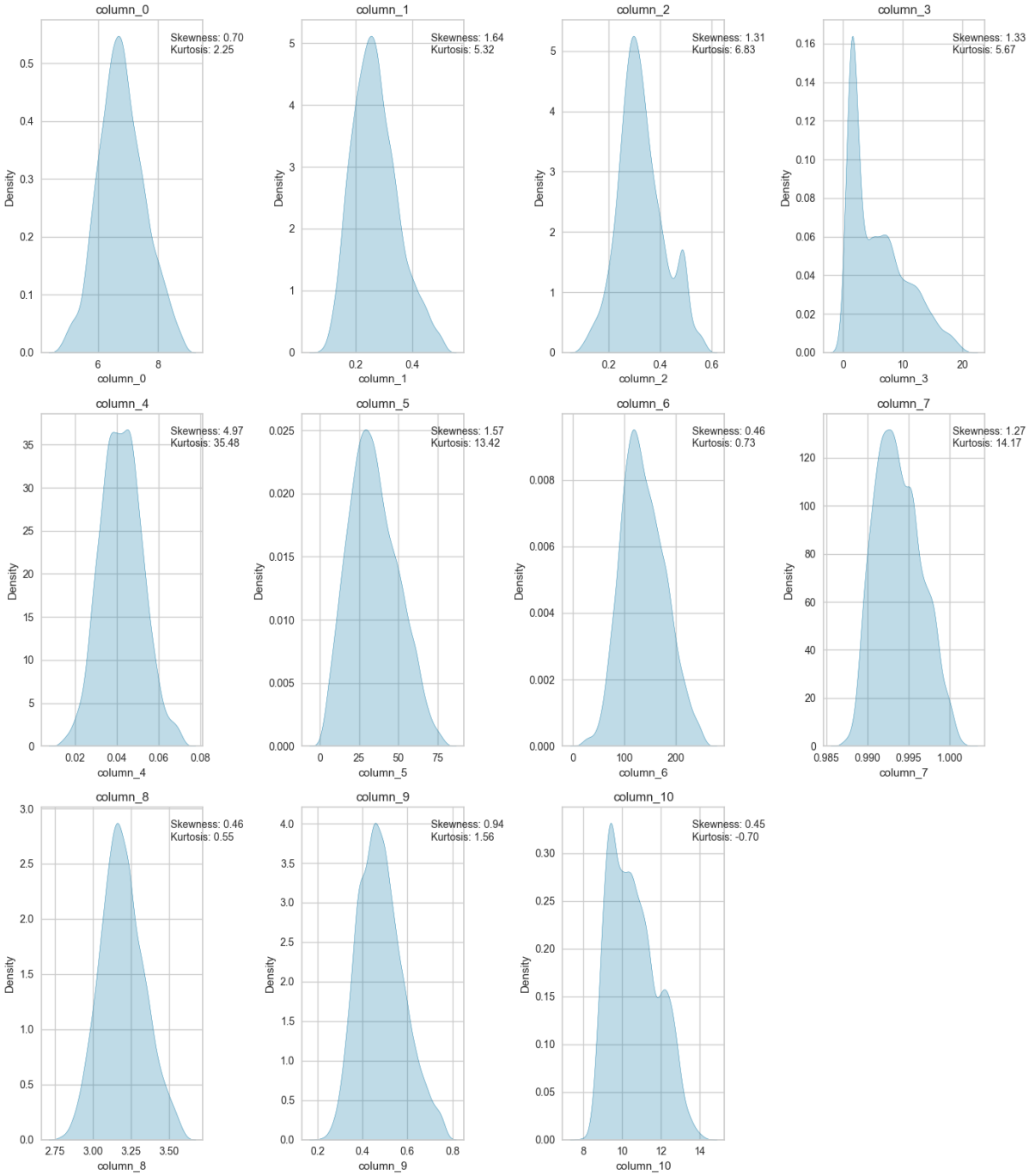


Figure 4 kernel density plots for determining distribution of data

Kernel density plots of our features using seaborn with their skewness and kurtosis written over the plot. Some columns like column_3 appears to have a multimodal distribution as well.

Several tests like Shapiro wilk test and p-test are also used to check if our data is normally distributed but since our data is quite big, and these tests are quite sensitive to big datasets. Hence visual univariate analysis is the main component for determining distribution of our different features.

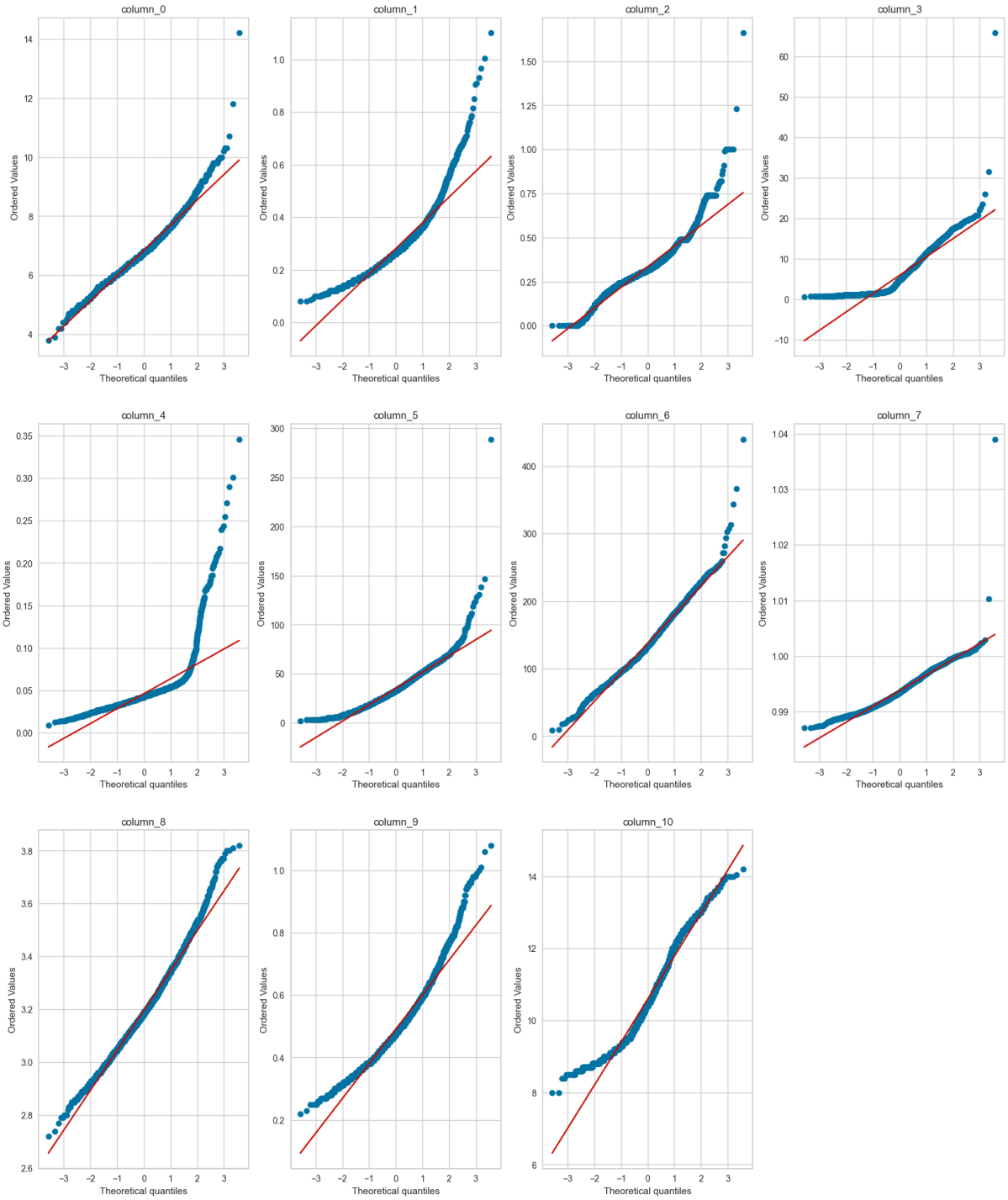


Figure 5 Probplot test

This is probplot and it is also one of the methods to test if our data is normally distributed. Some of them like the column_0, Column_8, column_6 are more likely to be normally distributed than columns like column_1, column_4.

1.6 Outlier detection

Previously, the histograms hinted towards the presence of outliers, so now we are going to visualize the outliers with the help of boxplot.

"A boxplot is a graphical representation that displays the distribution, central tendency, and variability of a dataset, highlighting potential outliers."

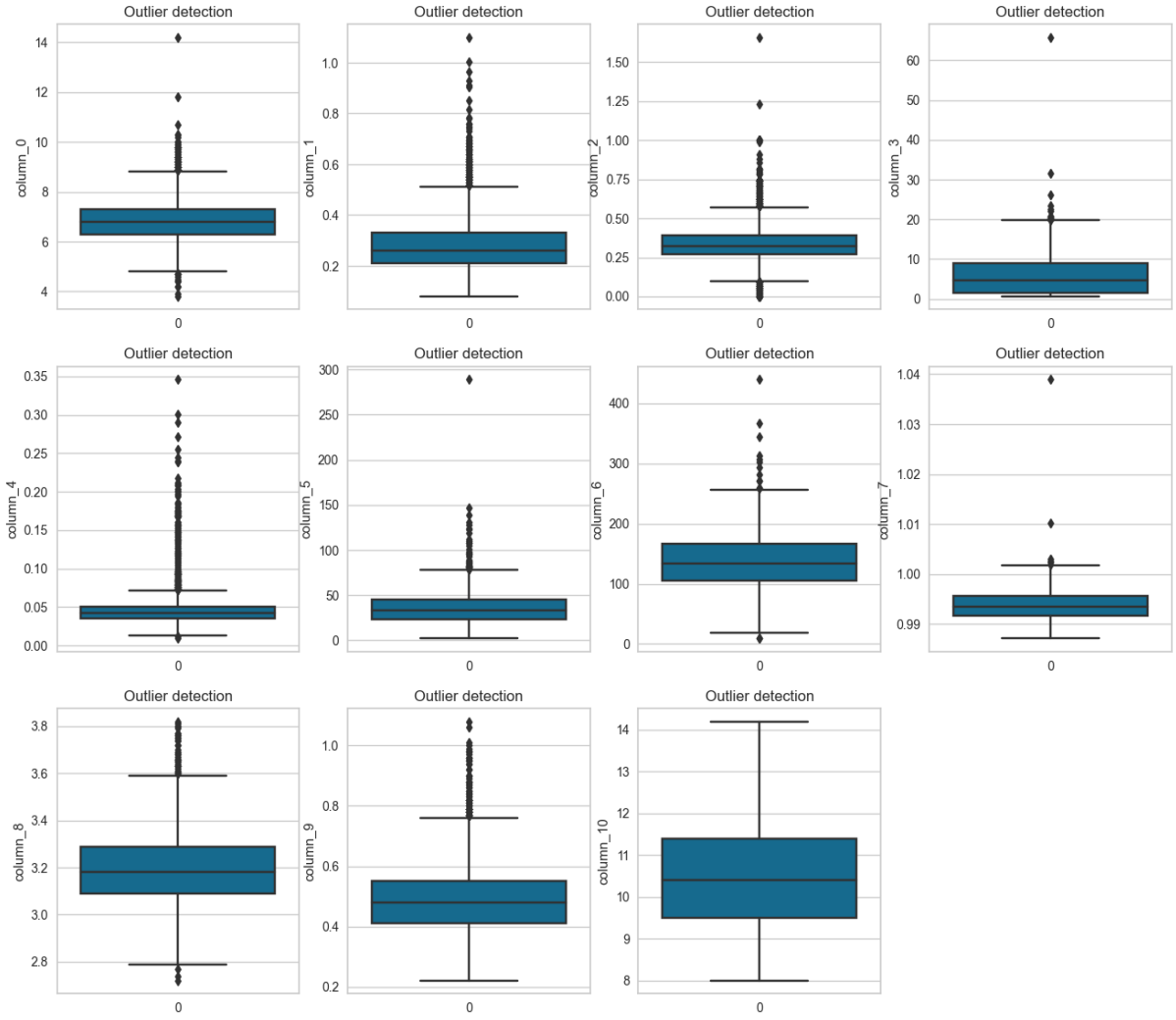


Figure 6 Boxplots visualising outliers

As suspected, there are definitely some outliers in the data. The no of outliers can be identified using various methods such as z score methods, IQR range etc. for my analysis, I used both od them but the outliers weren't too much and removing outliers and filling them or all together removing outliers, both lead to creation of sparse scatter matrix. So, I decided to carry out my analysis without outliers removal. In the last chapter of my analysis, I will further use DSBAN to detection anomalies in my data.

column_0	33
column_1	69
column_2	70
column_3	9
column_4	89
column_5	24
column_6	10
column_7	3
column_8	27
column_9	38
column_10	0
dtype: int64	

column_0	106
column_1	133
column_2	223
column_3	16
column_4	178
column_5	44
column_6	14
column_7	6
column_8	46
column_9	96
column_10	0
dtype: int64	

Figure 7 Outlier detection using zscore normalization and on the right, outliers detected using IQR technique

Since our data has around 4000 instances, these outliers remain insignificant and hence, we can simply tolerate some amount of error by keeping them.

Moreover, since our data is already standardized, further removal of outliers lead to unwanted results, one example of this can be seen during visual bivariate analysis.

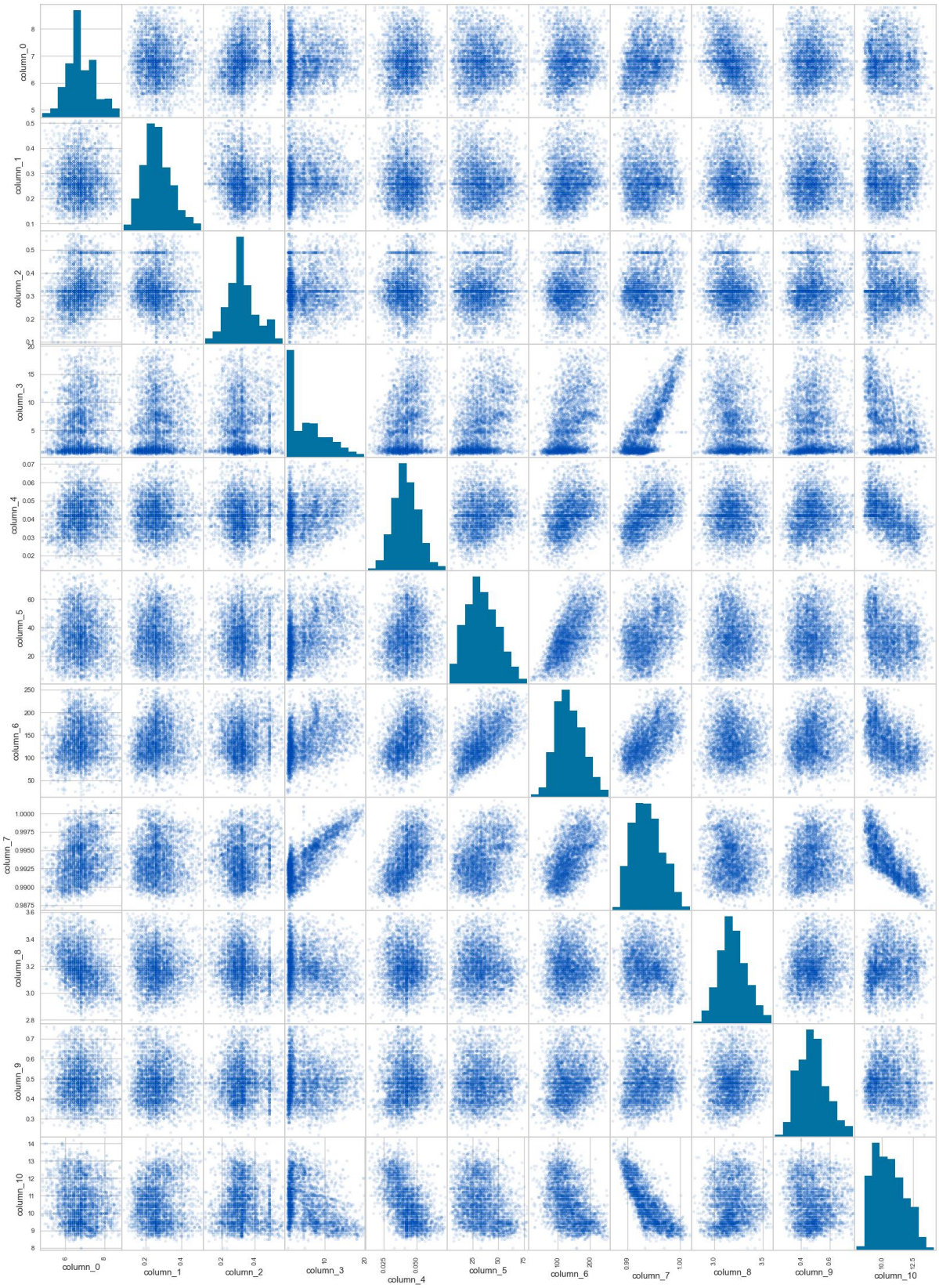


Figure 8 after removing outliers, we get obscure relations between features

This is what happen (the relationship between features look so obscure~ unclear) when we try filling our data with either best measure of central tendancy or even if we try removing the outliers.

1.7 Bivariate analysis and correlation

“Bivariate analysis is the examination of two variables to understand the relationship between them. It helps explore how changes in one variable might be associated with changes in another, providing insights into potential connections or patterns.”

It includes things plotting the correlation matrix between different attributes, visualizing the relationships between different features by methods like scatterplot as scatterplot is the best visual analysis to determine any linear or non-linear relationship between any 2 features.

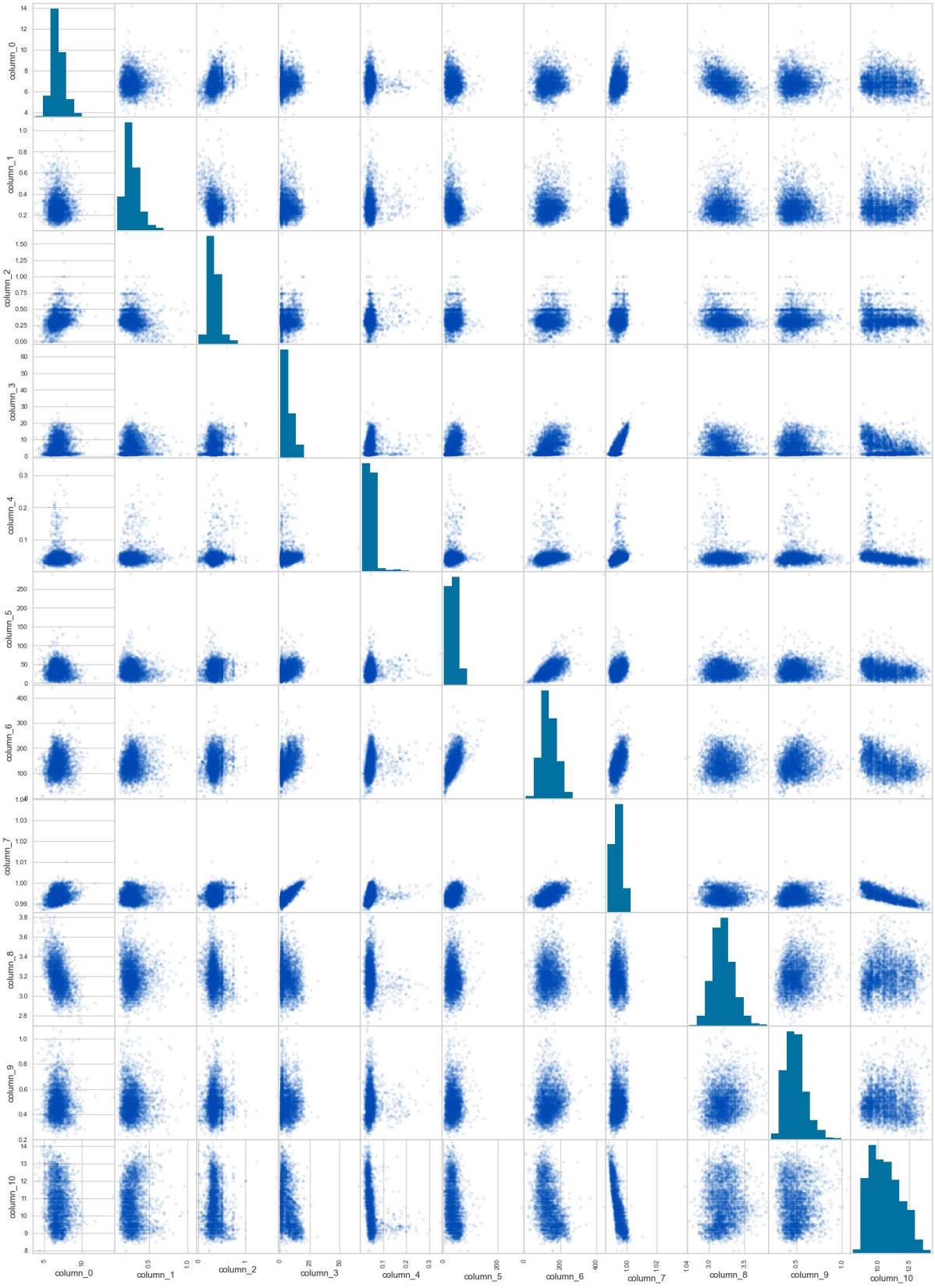


Figure 9 A scatterplot showing relationship between every 2 features of our data

This scatterplot which was plotted without removing outliers looks much better than when we plotted a scatter matrix after removing the outliers

Besides, all this the main part is now we can clearly observe the relationships between different features. On the basis of our observation, column_3 and column_7 seems to be linearly dependent on each other. Other then that, most features shows kind of an independent relationship from the other ones.

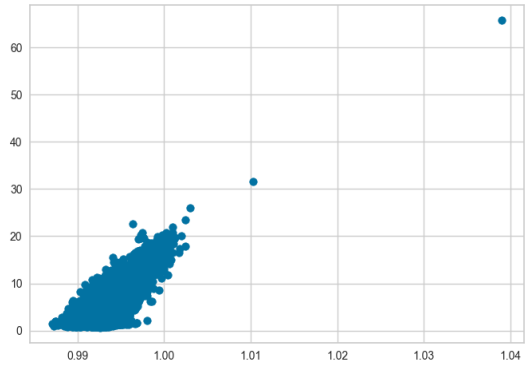
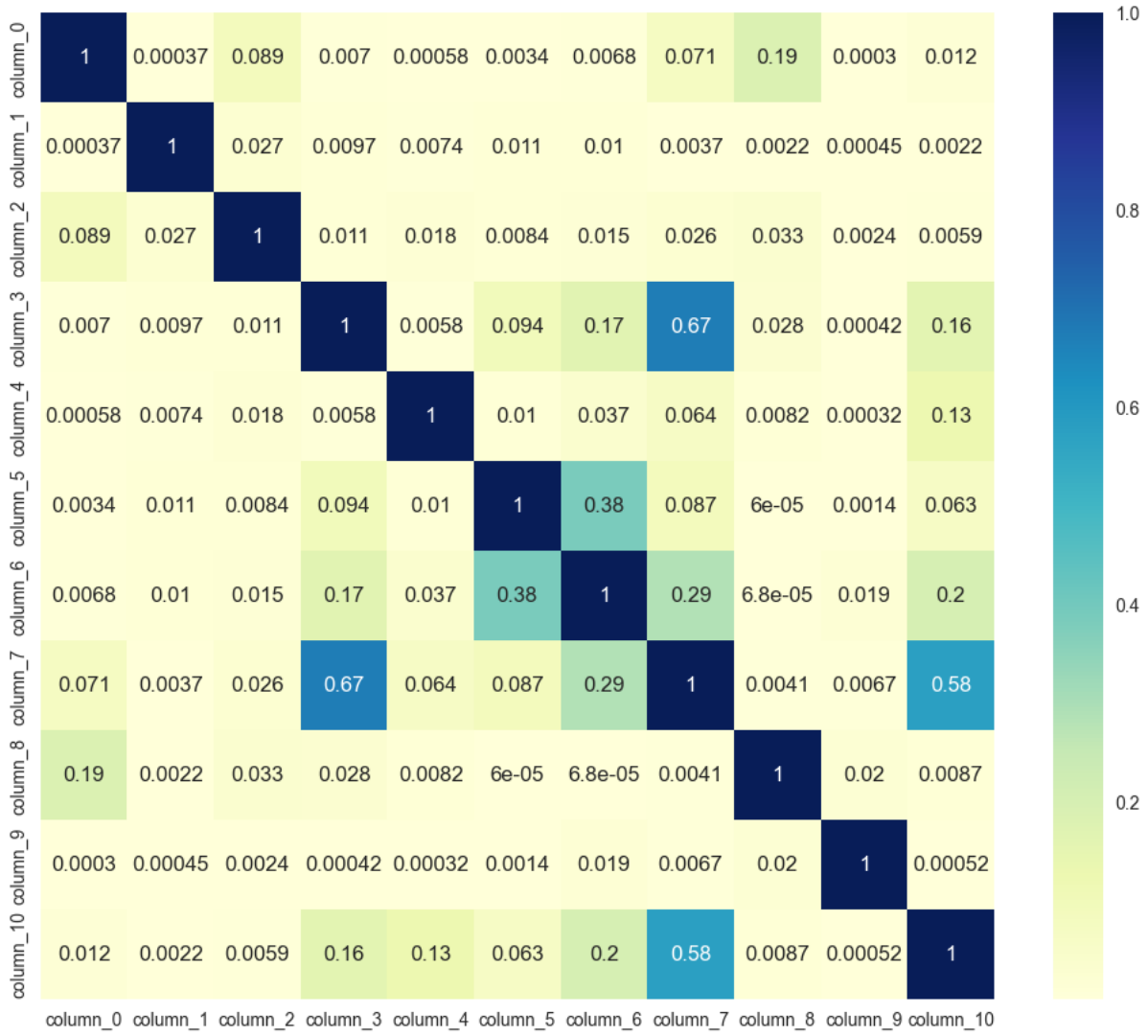


Figure 10 A better view of column_3 vs column_7

Furthermore, the numerical correlation between correlation between 2 variables can be shown by the use of heatmap. I have used the method of pearson correlation to determine the linear relationships between all attributes.

"Pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data."

I have shown squared peareson correlation coefficient in my data by the use of heatmap and The heatmap shown on next page uses colors to represent the magnitude of correlations, with brighter colors indicating stronger correlations, and annotations provide the exact correlation values for each pair of variables, facilitating the identification of patterns and relationships in the data. And the squared pearson correlation refers to coefficient of determination obtained by the method of squaring the pearson correlation such that 0 signifies no correlation and 1 signifies maximum correlation. However, not being non linearly correlated never means 2 features are not related at all, they can be non-linearly related.



As suspected earlier, based on the heatmap, we can say that column_3 and column_7 are somewhat correlated with value 0.67. However, most features here doesn't look like any linearly correlated at all.

Chapter 2: Dimensionality reduction

"It refers to decrease the number of dimensions of our data to overcome certain problems."

why is it important?

There are various reasons regarding why dimensionality reduction is important such as:

- To get rid of the well-known curse of dimensionality
- To be able to visualize data in 2 or 3 dimensions

"Curse of dimensionality refers to the challenges and issues that arise as the number of features or dimensions in a dataset increases. In high-dimensional spaces, data points become sparse, and the volume of the space grows exponentially, leading to increased computational complexity, a need for more data to maintain statistical significance."

In other words, exponential growth in number of examples are required to maintain the sample density as the number of dimensions increases.

Secondly, if the data will have only fewer dimensions, it can be easily visualized.

Methods:

Dimensionality reduction can be done in mainly 2 different ways:

Feature selection:

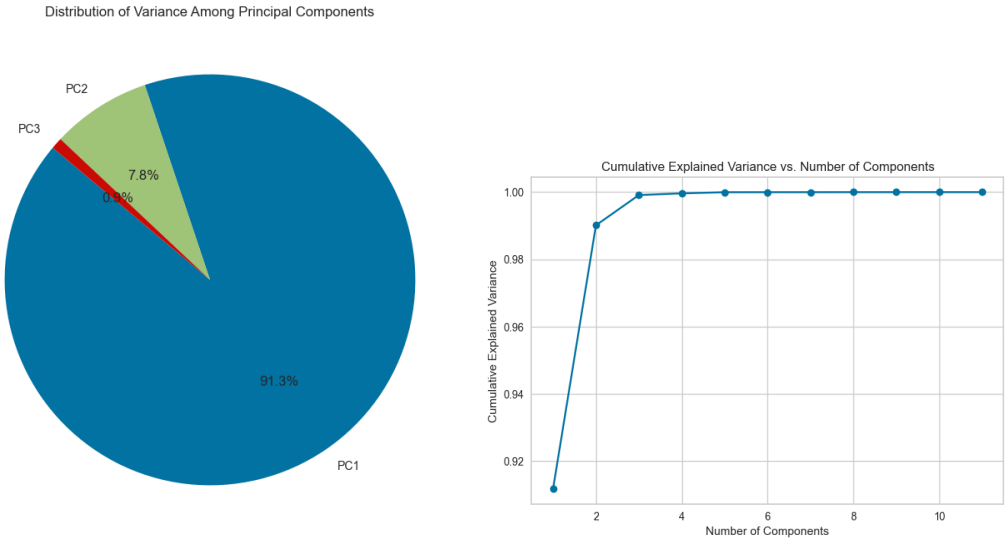
Under this, by looking at our previously scatterplot and correlation matrix, we can determine the highly correlated features and out of which one feature can be discarded as one features simply represents the other one well enough. Therefore, we end up selecting just one out of 2 highly closed related features and simple remove the other feature. However, the only closely correlated feature in my data was found between column_3 and column_7 and they are that too much correlated to be just kept one out of them.

Feature extraction:

This is the another way of performing dimensionality reduction. After performing feature selection, feature extraction can be a great way to perform further dimensionality reduction.

Feature extraction methods aim at selecting and combining features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set, without losing important or relevant information. This way it reduces the dimensions and also helps in effective visualization. The most famous algorithm for this purpose is called PCA. It identifies the principal components (linear combinations of the original features) that capture the maximum variance in the data. By transforming the data

into a new coordinate system defined by these principal components, PCA allows for a reduction in dimensionality while retaining the essential information.



While employing PCA, we need to selection to how many dimensions, we can reduce our features to. For selecting the number of principal components, the amount of variation we wish to be preserve can be taken into account. The above pie chart shows how much variation we will be preserving to the ratio of number of principal components selected. The plots above show that by clustering our data to either 2 or 3 will preserve us good amount of variance.

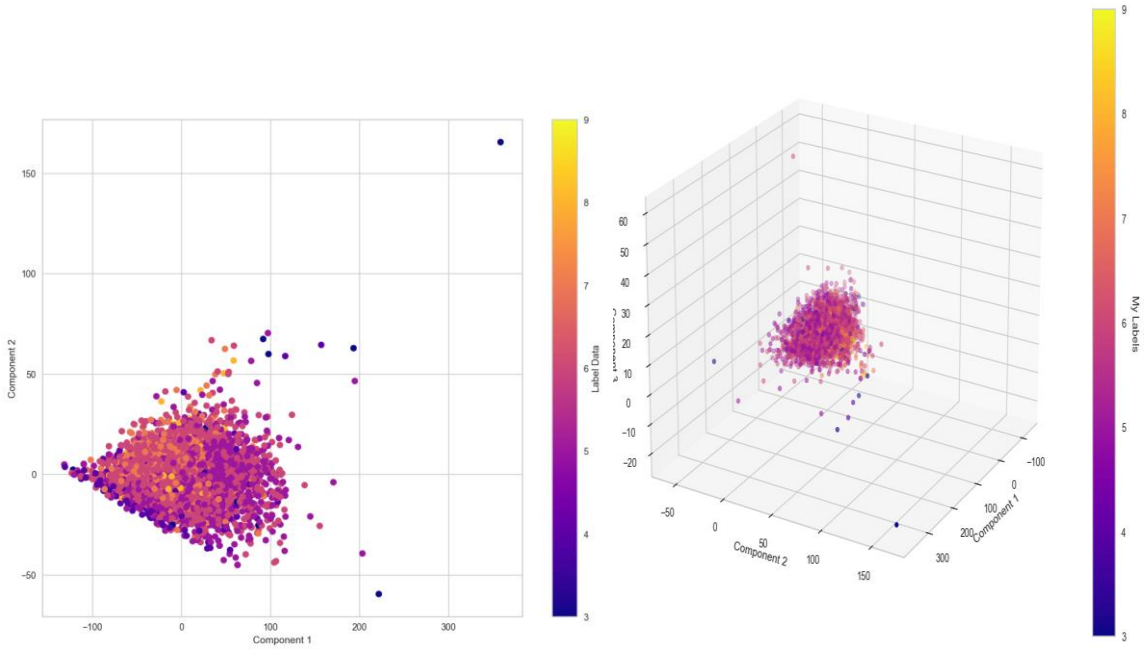


Figure 11Observation of data points using PCA in 2D and 3D

Hence, PCA helped us visualizing the data in 2D as well as 3D. Moreover, the color bars represent the wine Quality value. Since, we have been successfully able to reduce our dimensions to 2D as well as 3D, we can move on to the next part of performing clustering.

Chapter 3

“Clustering is a machine learning technique that involves grouping similar data points together based on certain characteristics or features.” The primary goal of clustering is to identify inherent patterns or structures within the data, allowing for the categorization of data points into distinct clusters or groups. We will use 2 of them here, one is kmeans and other one is DSBCAN.

Kmeans clustering

k-means clustering measures similarity using ordinary straight-line distance (Euclidean distance, in other words). It creates clusters by placing a number of points, called **centroids**, inside the feature-space. Each point in the dataset is assigned to the cluster of whichever centroid it's closest to. The "k" in "k-means" is how many centroids (that is, clusters) it creates. You define the k yourself.

The optimal value of k can be find out by using various methods,

1.) **Inertia: The elbow curve method**

Description: Inertia measures the sum of squared distances between data points and their assigned cluster center. It reflects how compact the clusters are.

Interpretation: Lower inertia values indicate tighter, more compact clusters. It is often used to identify the optimal number of clusters using the "elbow method."

2.) **Davies-Bouldin Score:**

Description: The Davies-Bouldin score quantifies the ratio of the average distance between clusters to the average cluster size. Lower values indicate better-defined clusters.

Interpretation: A lower Davies-Bouldin score suggests more distinct and well-separated clusters.

3.)Calinski-Harabasz Score:

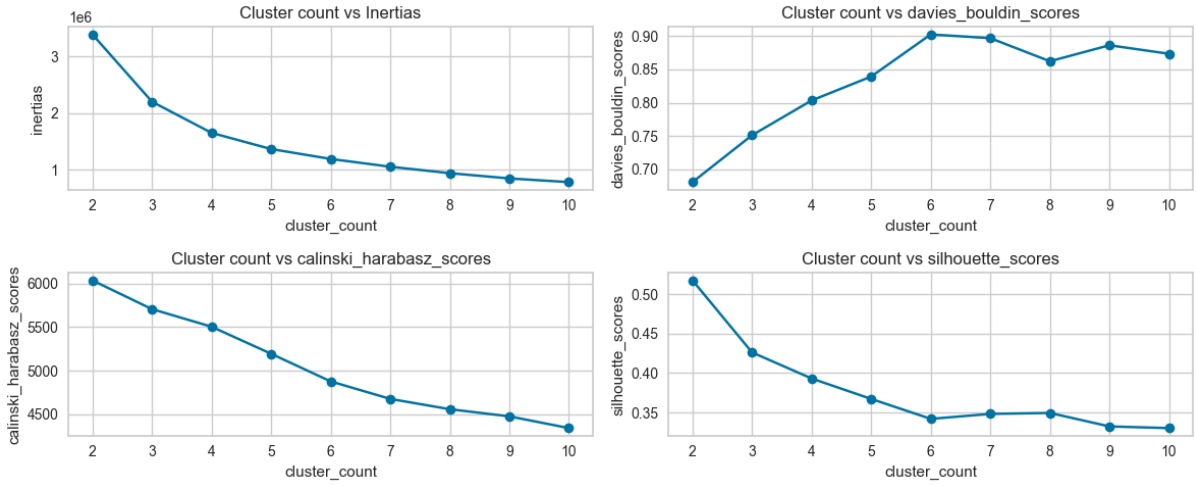
The Calinski-Harabasz score, also known as the Variance Ratio Criterion, measures the ratio of between-cluster variance to within-cluster variance. Higher scores indicate better-defined clusters.

Interpretation: Higher Calinski-Harabasz scores correspond to more distinct clusters. It is used to evaluate the clustering performance.

4.)Silhouette Score:

Description: The Silhouette Score quantifies how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

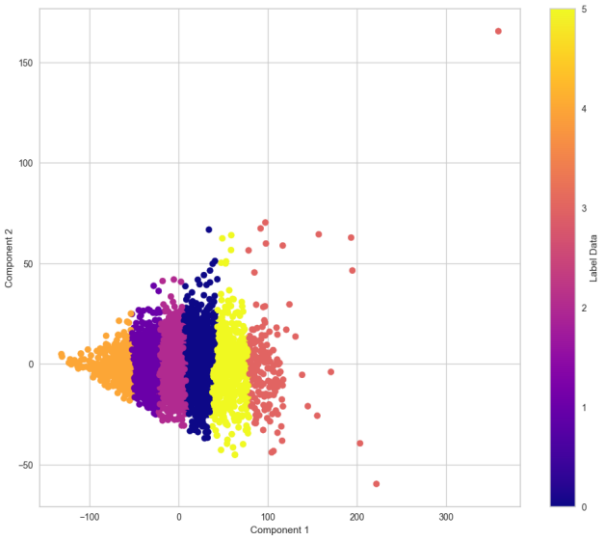
Interpretation: A Silhouette Score near 1 indicates that the object is appropriately placed within its cluster, and the clusters are well-separated. A score around 0 means overlapping clusters, and negative values suggest that the object may be assigned to the wrong cluster.



Different metrices described above are shown for my kmean clustering

Interpreting the results can be tricky as different metrices may show different results. Like here, CH index determines the best possible clustering to be of 2 clusters while the elbow point in inertia and David bouldin index shows different. However, for k means clustering, David bouldin index is the preferred one in general. Lastly, the sihoulette score is nearest to one for the clsuster count of 2 and hence showing good clustering for n=2.

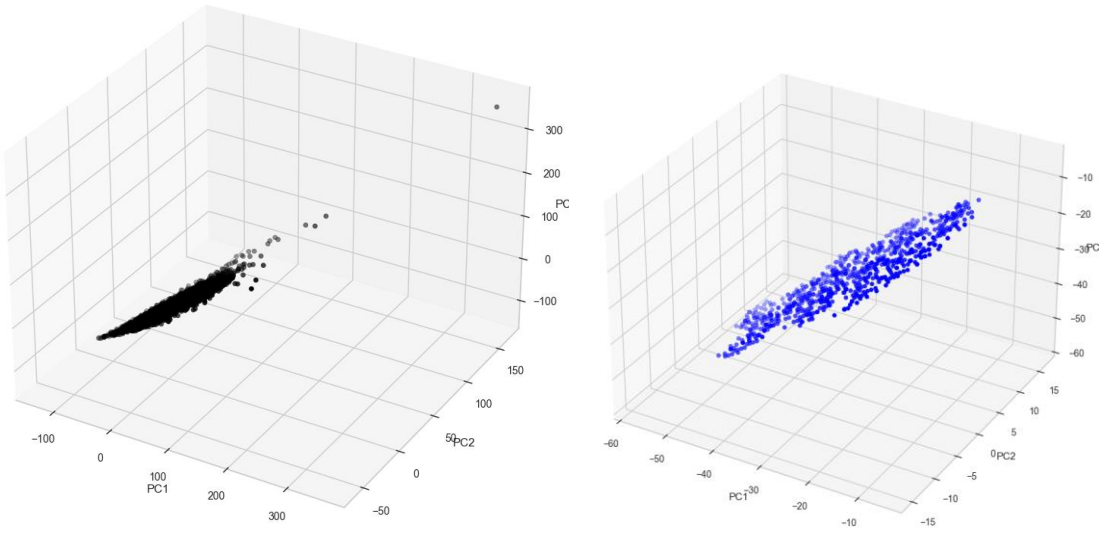
Another perimeter is the number of unique wine qualities in our dataset is 6. So, applying kmeans with 6 clusters is shown on the left side here.



Evaluating the model we got no accuracy at all. The main reason could be the true label itself. The reason behind this is wine quality only have 6 values out of a score of 1-10. And while the predicted values by the clustering have a totally different scale of 0—5.

```
True labels are [6 5 7 8 4 3 9]
Predicted labels are [0 1 2 3 4 5]
```

Density-Based Clustering refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms. The data points in the region separated by two clusters of low point density are considered as noise. The surroundings with a radius ε of a given object are known as the ε neighborhood of the object. If the epsilon neighborhood of the object comprises at least a minimum number, MinPts of objects, then it is called a core object.



12data points detected

13outliers detected

When I applied the algorithm on the data, I these black points were detected as the outliers by the algorithm while the figure on the right represents the good data points in blue.

Name – Ajeesh Garg

Course - IMAPP