

**MAEER’S MAHARASHTRA INSTITUTE OF TECHNOLOGY, PUNE KOTHRUD, PUNE: 411021**

**DEPARTMENT OF COMPUTER ENGINEERING**

**A MINI PROJECT REPORT ON**

CLUSTERING AND DIMENTIONALITY REDUCTION USING PRINCIPAL COMPONENT ANALYSIS

SUBMITTED BY

**Chakor Pandkar – 71700425C**

**Rakshit Pensalwar – 71700489K**

**Shubham Patil – 71700447D**

**Machine Learning**

1. **TITLE:**

CLUSTERING AND DIMENSIONALITY REDUCTION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

1. **CONTRIBUTION:**

* Selection of topic:
  + Rakshit Pensalwar
* Dataset Analysis:
  + Chakor Pandkar
* Program Implementation:
  + Rakshit Pensalwar
  + Shubham Patil
* Presentation:
  + Chakor Pandkar
* Report:
  + Shubham Patil

1. **ABSTRACT:**
2. Objectives:

The objective of this report is to determine if a dimension reduction technique, namely principal component analysis, could be utilized in combination with a finite difference method for these problems. A number of numerical experiments were designed to examine the efficiency under different conditions.

1. Method:

An important area in wine factory concerns the amount of alcohol, alkalinity, ash, color intensity, hue, etc. When options depend on several underlying wine class, the complexity of the problem makes it difficult to solve using conventional finite difference methods. Instead, stochastic approaches are employed despite the extremely slow convergence of these methods. The results show that the proposed approach performs very well when the correlation between the underlying wine class is sufficiently high.

1. **MOTIVATION:**

Actual Design Process involves assignment of shapes and their dimensions which in turn serve as the guidelines for the manufactureres, developers, users, and the designers themselves. The various stages of development of a design, particularly the machinig fabrication and testing involve considering it as an approximation in the sense of defining tolerances for both shape and size parameters. These tolerances are inevitably encountrered due to the limitations of each of the stages of the design and development process. It is therefore only essential that the designer be equipped with tools which allow him to design and test with real-life considerations of tolerances.

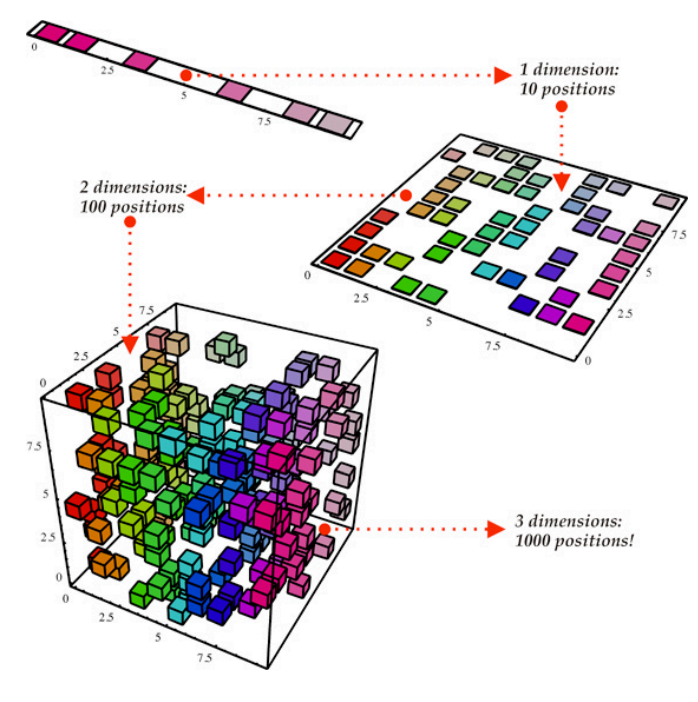
1. **INTRODUCTION:**

Principal Component Analysis (PCA) is the general name for a technique which uses sophisticated underlying mathematical principles to transforms a number of possibly correlated variables into a smaller number of variables called principal components. The origins of PCA lie in multivariate data analysis, however, it has a wide range of other applications, as we will show in due course. PCA has been called, ’one of the most important results from applied linear algebra and perhaps its most common use is as the first step in trying to analyze large data sets.

Some of the other common applications include; de-noising signals, blind source separation, and data compression. In general terms, PCA uses a vector space transform to reduce the dimensionality of large data sets. Using mathematical projection, the original data set, which may have involved many variables, can often be interpreted in just a few variables (the principal components). It is therefore often the case that an examination of the reduced dimension data set will allow the user to spot trends, patterns and outliers in the data, far more easily than would have been possible without performing the principal component analysis.

# **Dimensionality reduction:**

Reducing the dimensions of the feature space is called **dimensionality reduction**. Reduction of dimensions is needed when there are far too many features in a dataset, making it hard to distinguish between the important ones that are relevant to the output and the redundant or not-so important ones.



There are many ways to achieve dimensionality reduction, but most of these techniques fall into one of two classes:

* **Feature Elimination:** Eliminating features to reduce the feature space. As a disadvantage, information is lost due to dropped features.
* **Feature Extraction:**In feature extraction, we create **K** “new” independent variables, where each independent variable is a combination of each of the given old independent variables.

1. **PRINCIPAL COMPONENT ANALYSIS**
2. Brief Primer:

Principal component analysis (PCA) is a statistical procedure that uses an [orthogonal transformation](https://en.wikipedia.org/wiki/Orthogonal_transformation) to convert a set of observations of possibly correlated variables into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables called principal components. The number of distinct principal components is equal to the smaller of the number of original variables or the number of observations minus one. This transformation is defined in such a way that the first principal component has the largest possible [variance](https://en.wikipedia.org/wiki/Variance) (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is [orthogonal](https://en.wikipedia.org/wiki/Orthogonal) the preceding components. The resulting vectors are an uncorrelated [orthogonal basis set](https://en.wikipedia.org/wiki/Orthogonal_basis_set). PCA is sensitive to the relative scaling of the original variables.

1. History

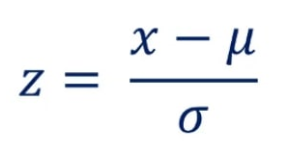
PCA was invented in 1901 by [Karl Pearson](https://en.wikipedia.org/wiki/Karl_Pearson) as an analogue of the principal axis theorem in mechanics; it was later independently developed and named by [Harold Hotelling](https://en.wikipedia.org/wiki/Harold_Hotelling) in the 1930s.

1. Mathematics Behind PCA:

PCA can be thought of as an unsupervised learning problem. The whole process of obtaining principle components from a raw dataset can be simplified in six parts:

* Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.
* Compute the mean for every dimension of the whole dataset.
* Compute the covariance matrix of the whole dataset.
* Compute eigenvectors and the corresponding eigenvalues.
* Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix **W.**
* Use this d × k eigenvector matrix to transform the samples onto the new subspace.

1. **Steps involved in PCA**:
   1. **Standardization:**Calculate the mean of all the dimensions of the dataset, except the labels. Scale the data so that each variable contributes equally to analysis.



* 1. **Covariance Matrix Computation:**We can compute the covariance of two variables **X** and **Y** using the following formula:



* 1. **Compute Eigenvectors and corresponding Eigenvalues:**  In [linear algebra](https://en.wikipedia.org/wiki/Linear_algebra), an eigenvector, or characteristic vector of a [linear transformation](https://en.wikipedia.org/wiki/Linear_map), is a nonzero [vector](https://en.wikipedia.org/wiki/Vector_space) that changes at most by a [scalar](https://en.wikipedia.org/wiki/Scalar_(mathematics)) factor when that linear transformation is applied to it. The corresponding eigenvalue is the factor by which the eigenvector is scaled.

In general, the eigenvector of a matrix ***A***is the vector for which the following holds:



where ***lambda***is a scalar value called the **‘eigenvalue’.** This means that the linear transformation is defined by *lambda*and the equation can be re-written as:



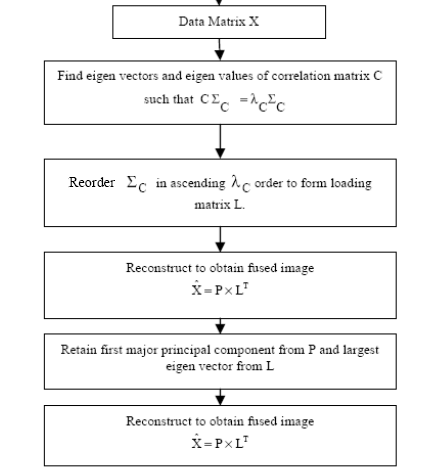
where ***I***is the identity matrix.

It’s important to notice that these eigenvectors are both unit eigenvectors, i.e. their lengths are both 1. These eigenvectors give us the patterns in the data, in order for us to extract the most useful ones.

* 1. **Choose k eigenvectors with the largest eigenvalues:**Sort the eigenvectors with respect to their decreasing order of eigenvalues, choosing *k*out of them, where *k* is the number of dimensions you wish to have in the new dataset.
  2. **Recasting data along Principal Components’ axes:**

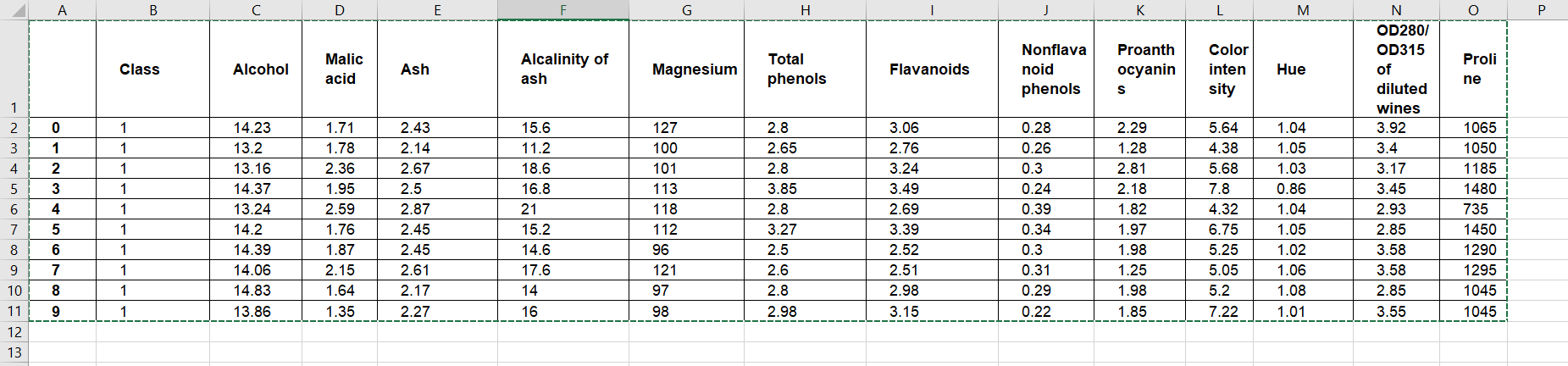
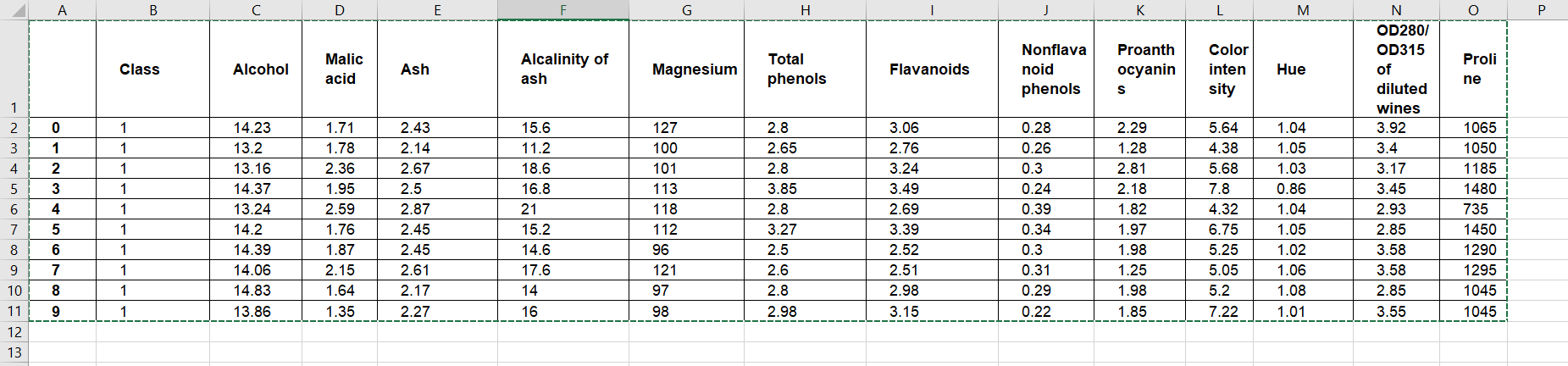
In the last step, we need to transform our samples onto the new subspace by re-orienting data from the original axes to the ones that are now represented by the principal components.

*Final Data= Feature-Vector\*Transpose(Scaled(Data))*

****

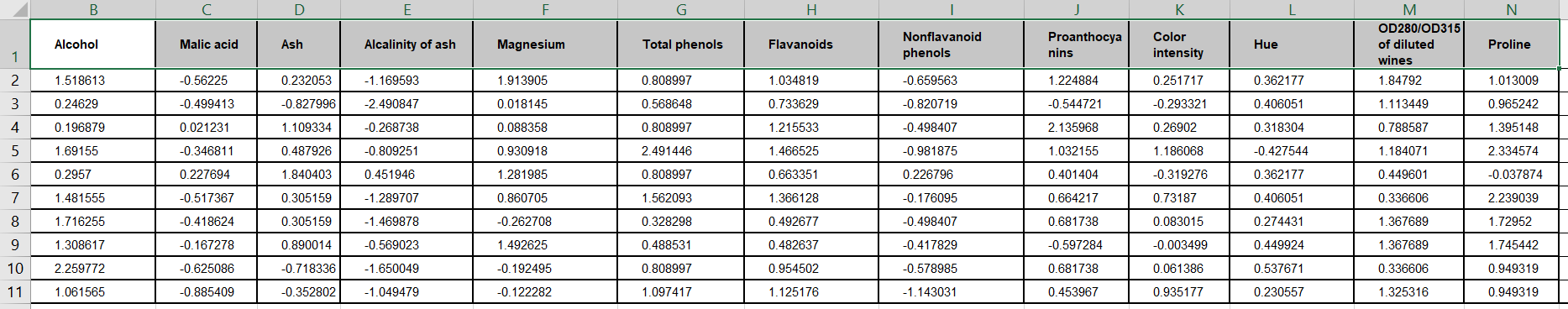
1. **SOFTWARE/HARDWARE REQUIR.:**
   1. *Software*:
      1. Python – 3.7.0
      2. Jupyter Notebook (IDE)
   2. *Hardware*:
      1. 64-Bit O.S. Platform (Windows)
2. **INPUT:**

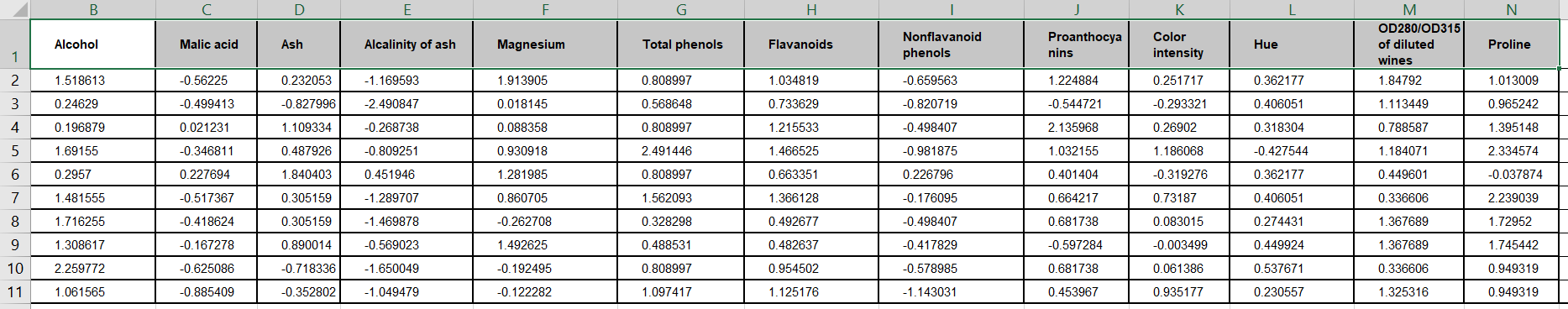
Wine dataset

****

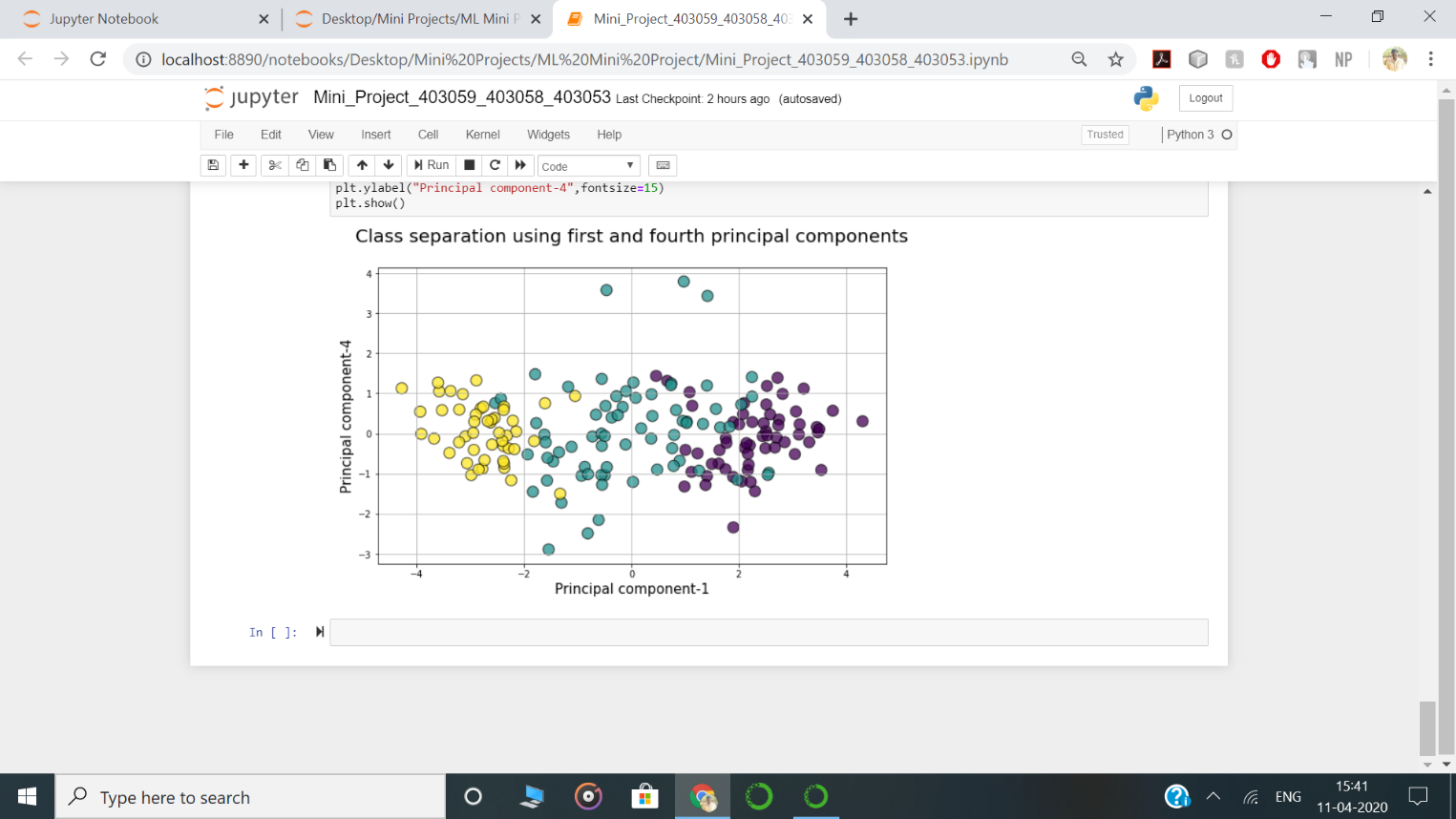
1. **OUTPUT:**

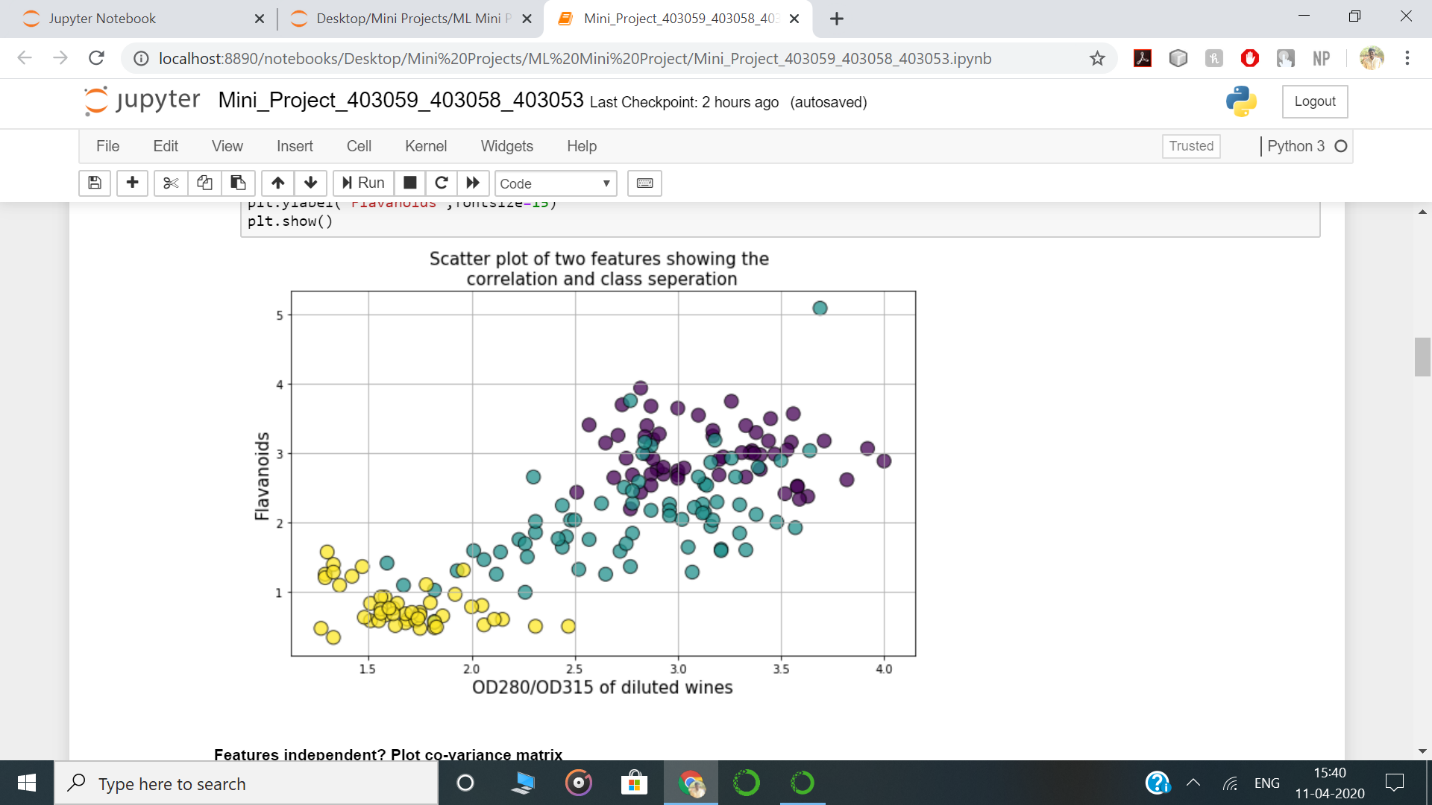
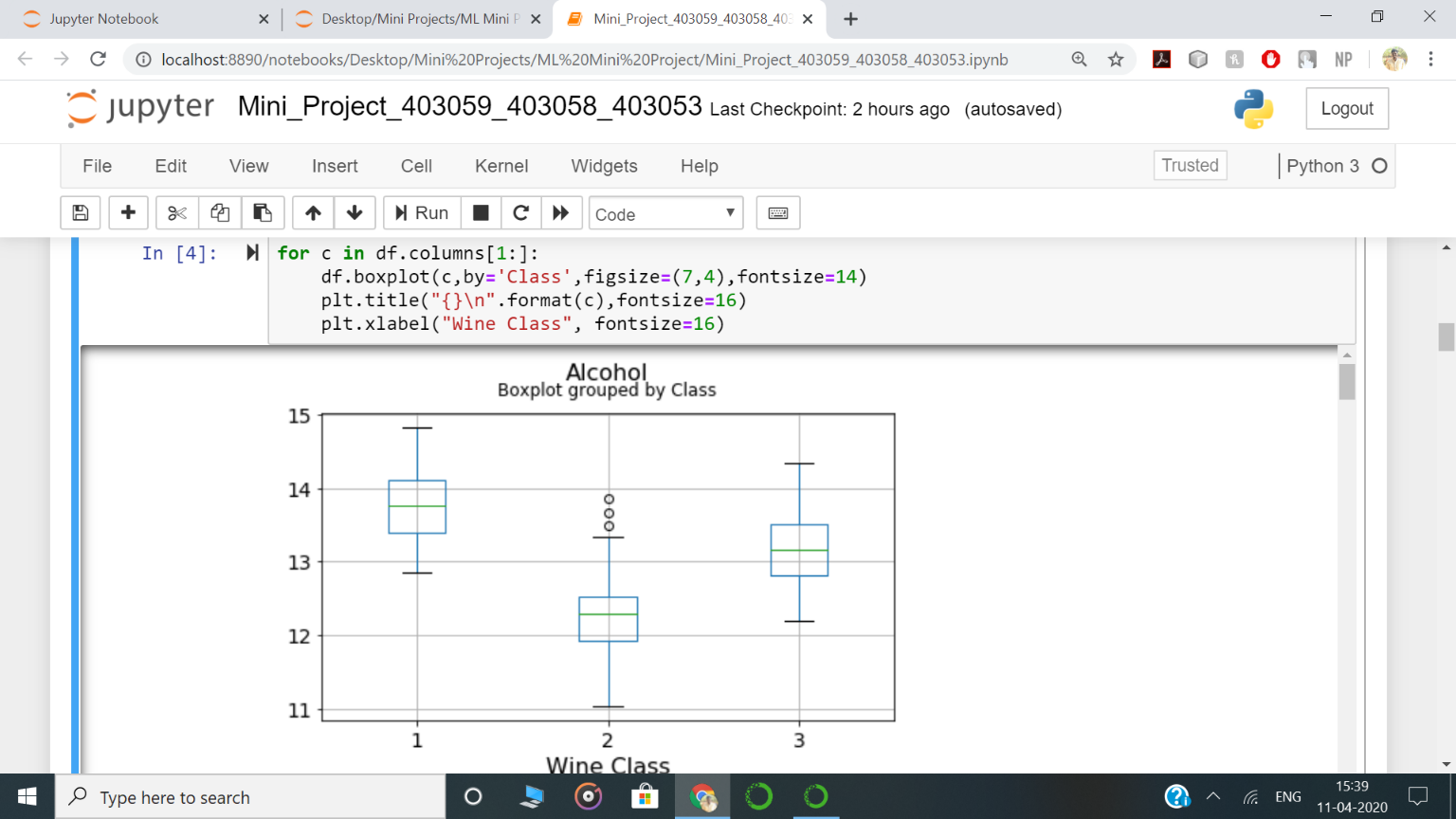
After PCA Clustering & Dimensionality reduction

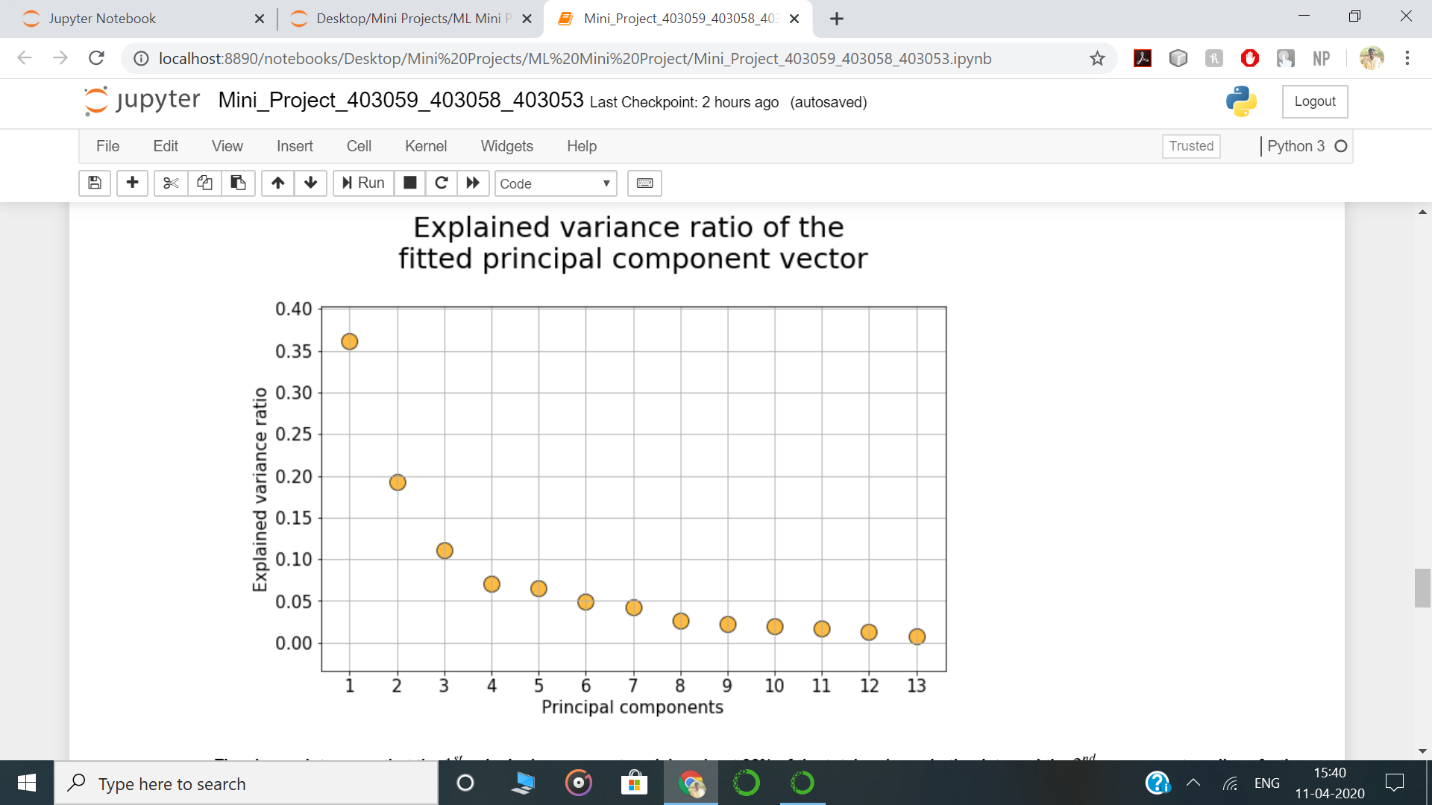
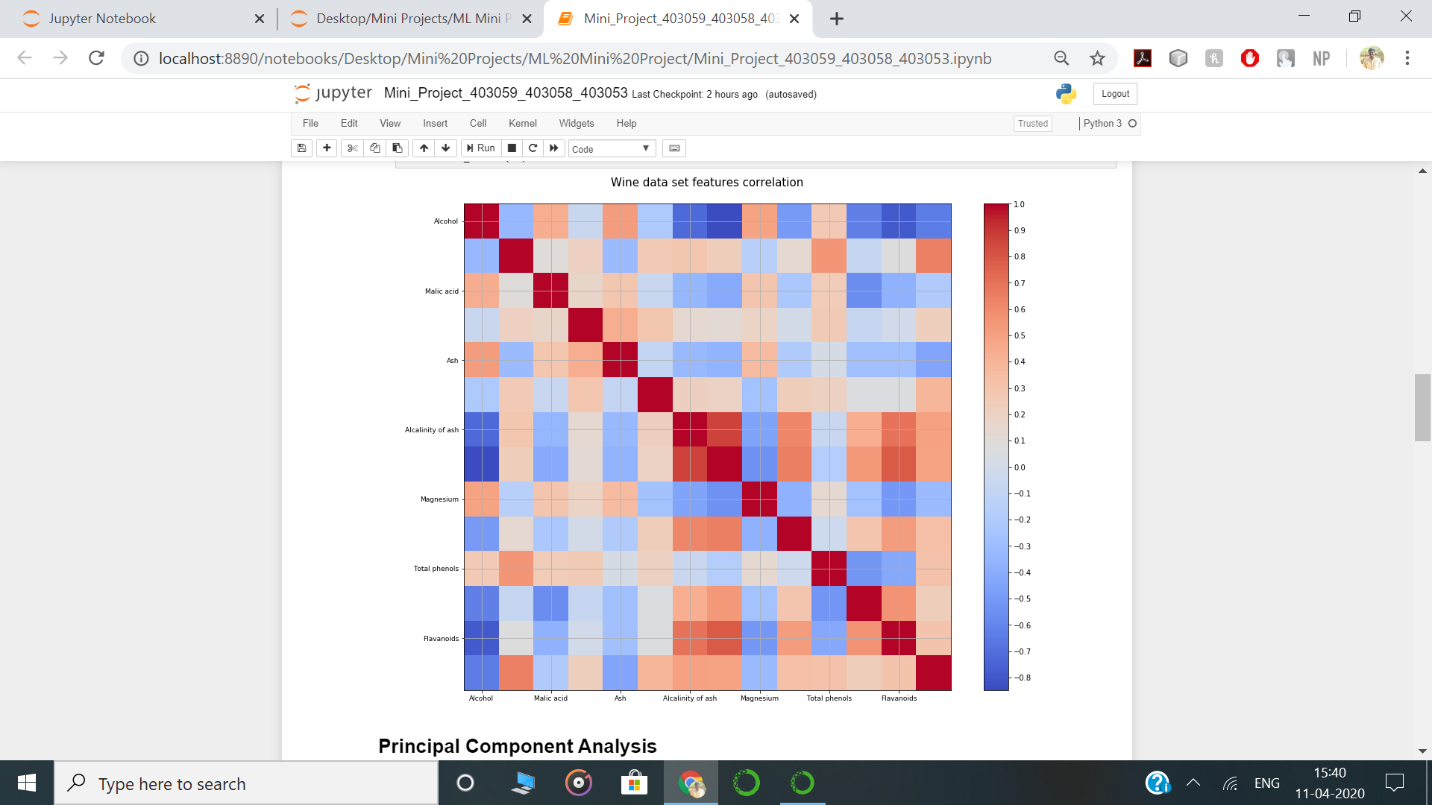
****

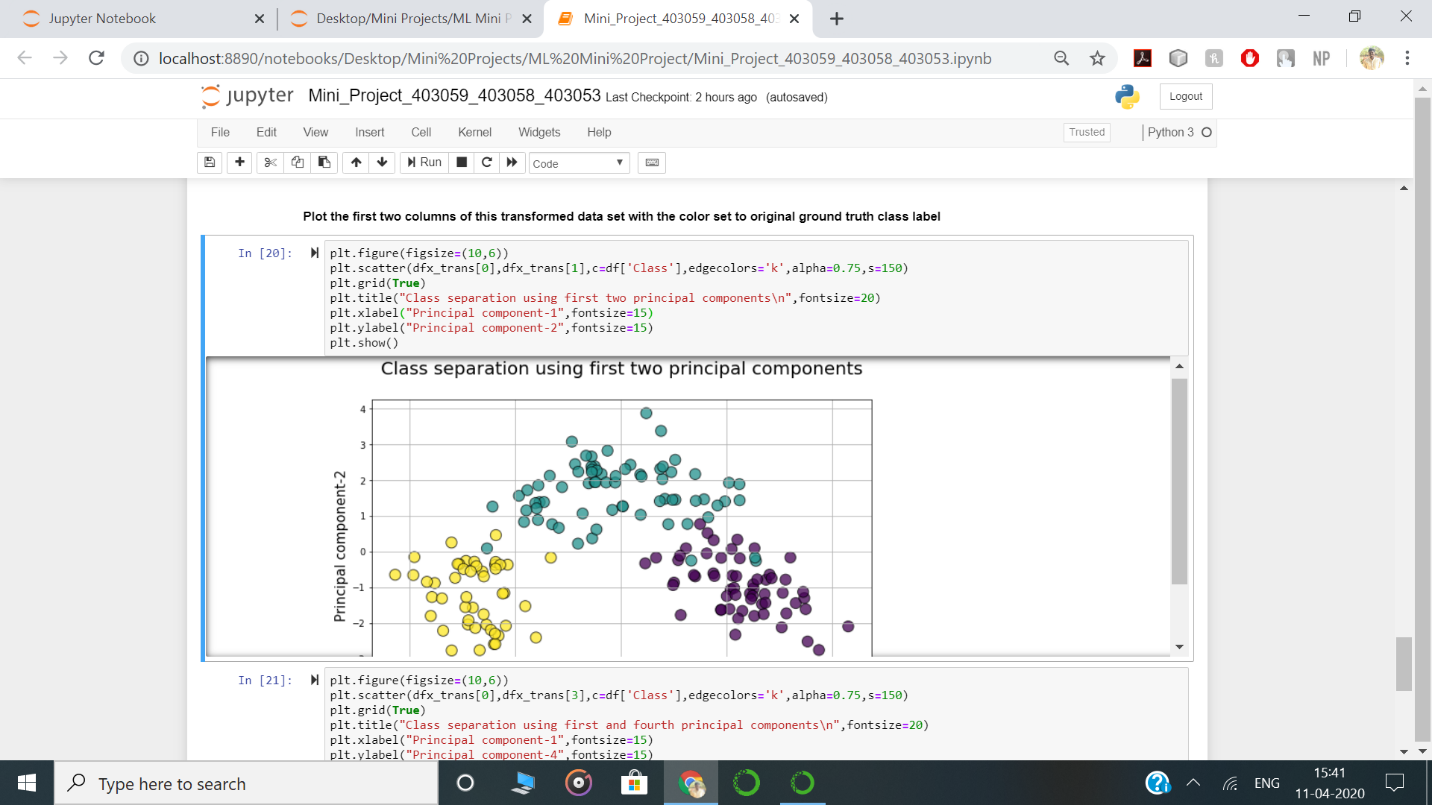
****

1. **SNAPSHOTS:**









1. **CONCLUSION:**

In this Project, we learned the fundamentals of working with principal component analysis (PCA), including the mathematics behind it. Despite being widely used and strongly supported, it has its share of advantages and disadvantages.

One benefit of PCA is that we can examine the variances associated with the principle components. Often one finds that large variances associated with the first principal components, and then a precipitous drop-off. One can conclude that most interesting dynamics occur only in the first k dimensions. In the example of the alcohol, k = 1. This process of throwing out the less important axes can help reveal hidden, simplified dynamics in high dimensional data.

1. **REFERENCES:**
2. Ian T. Jolliffe. Principal Component Analysis. Springer-Verlag New York Inc., 1986.
3. Lay, David. (2000). Linear Algebra and It’s Applications. Addison-Wesley, New York.
4. Abdi H. and Williams L.J. (2010) Principal component analysis. Statistics & Data Mining Series, Vol. 2, John Wiley & Sons, p.433-459.
5. Everitt BS, Dunn G (2002) Applied Multivariate Data Analysis. 2nd Edition, Arnold Publisher, London.
6. Ma, Y. Z., 2011, Lithofacies clustering using principal component analysis and neural network: applications to wireline logs, Math. Geosciences, 43(4):401-419.