A Mini Project Report on

# “Customer Segments Using PCA”

Submitted By

Shubham Kokane (BECOMPA-35)
Shreyas Soni (BECOMPA-68)
Piyush Wadi (BECOMPA-72)

(Final Year Computer Engineering)

Guided By

Dr. Archana Chaugule

PIMPRI CHINCHWAD EDUCATION TRUST
A Trusted Brand in Education Since 1990..

Department of Computer Engineering
Pimpri Chinchwad College of Engineering and
Research, Pune 412101 [2020-21]

PIMPRI CHINCHWAD EDUCATION TRUST'S

# Pimpri Chinchwad College of Engineering and Research, Pune 412101

# CERTIFICATE

This is to certify that *(Shubham Kokane, Shreyas Soni, Piyush Wadi) has* successfully completed the project entitled “Customer Segments Using PCA” in the fulfillment of B. E. (Computer Engineering) LP-III and this work has been carried out in my presence.

Date :

Place :

Dr. Archana Chaugule
Guide and HOD
(Dept. of Comp. Engg.)
Pimpri Chinchwad College of Engineering
and Research, Pune 412101

External

Prof. Dr. Tiwari H.U.
Principal,
Pimpri Chinchwad College of Engineering
and Research, Pune 412 101

# ACKNOWLEDGEMENT

This is a great pleasure & immense satisfaction to express my deepest sense of gratitude & thanks to everyone who has directly or indirectly helped me in completing my project work successfully.I express my gratitude towards Project guide (Dr. Archana Chaugule) , Head of Department of Computer Engineering, Pimpri Chinchwad college of Engineering and Research, Pune 412101 who guided & encouraged me in completing the project work in scheduled time. I would like to thank our Principal (Prof. Dr. Tiwari H.U), for allowing us to pursue my project in this institute.No words are sufficient to express my gratitude to our parents for their unwavering encouragement. We also thank all friends for being a constant source of my support.

Shubham Kokane
Shreyas Soni
Piyush Wadi

# ABSTRACT

Using PCA we can reduce the dimensionality of large datasets to reduce the variance of the data. We perform PCA on the products data, reducing the number of variables into principal components preserving as much data as possible. We reduce the dataset into 4 principal components to observe the variance of the dataset. We have calculated the variance for the same. Reducing the dimensionality of the dataset can help us reduce the variance in this dataset and then less computation time is required as well as less redundant features can be removed from the data. Thus, on the products dataset we try to remove redundant features and then divide the dataset into principal components so that we can take more useful data into consideration.

# INDEX

# List Of Figures

# 1. INTRODUCTION

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

# 2. MOTIVATION

Organizing information in the principal components way, will allow you to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables.

An important thing to realize here is that the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

Geometrically speaking, principal components represent the directions of the data that explain a **maximal amount of variance**, that is to say, the lines that capture most information of the data. The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has. To put all this simply, just think of principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.

# 3. PROBLEM STATEMENT

Apply the Principal Component Analysis for feature reduction on Customer segmentation data. Reduce the dimensionality of the data for reducing features and improving computational efficiency.

# 4. OBJECTIVES

Objectives of the project are:

- Reduce dimensionality of data using PCA in 4 principal components
- Observe the difference in variance as the change in principal components.
- Plots graphs of variance according to the change in variance.

# 5. SOFTWARE AND HARDWARE REQUIREMENTS

## 5.1 SOFTWARE REQUIREMENTS:

1. Python
2. Pandas
3. scikit-learn

## 5.2 HARDWARE REQUIREMENTS:

1. Hard Disk: minimum of 1GB of available hard disk space.
2. Memory: recommended 4GB
3. Operating System: *Windows / Ubuntu.*

# 4. Tool: Introduction

## 1. Python:

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

Python is a programming language that lets you work quickly and integrate systems more efficiently.

There are two major Python versions: **Python 2 and Python 3**.

## 2.Pandas:

### *What is Pandas?*

Pandas is a Python library used for working with data sets.

It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

### *Why Use Pandas?*

Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant.

Relevant data is very important in data science.

### *What Can Pandas Do?*

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is the average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

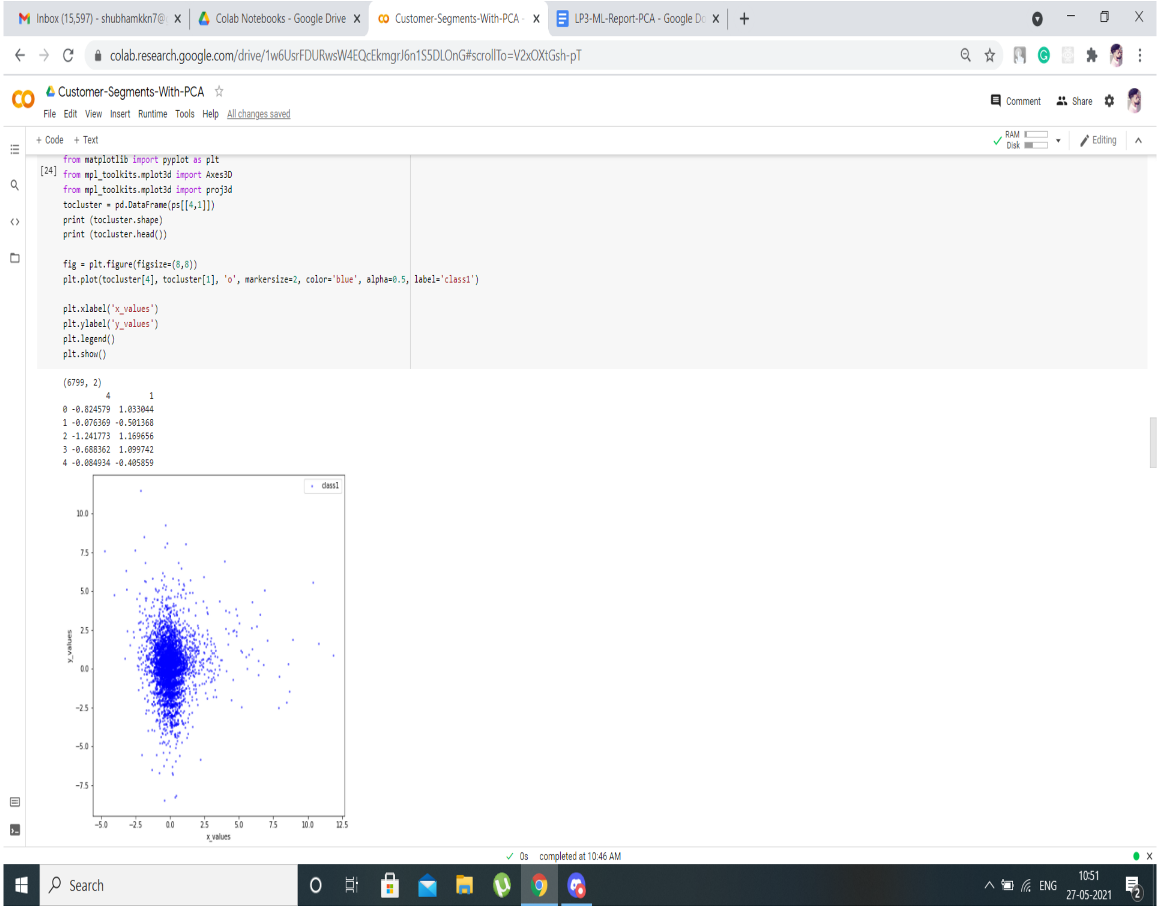
## 3. Scikit-learn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

# 7. STEPS OF IMPLEMENTATION WITH OUTPUT

- **7.1 Steps of implementation :**
    1. Load the customer segmentation dataset..
    2. Apply PCA with specified no. of principal components.
    3. Visualize and get total extended variance.

- **7.2 Output Screenshots**



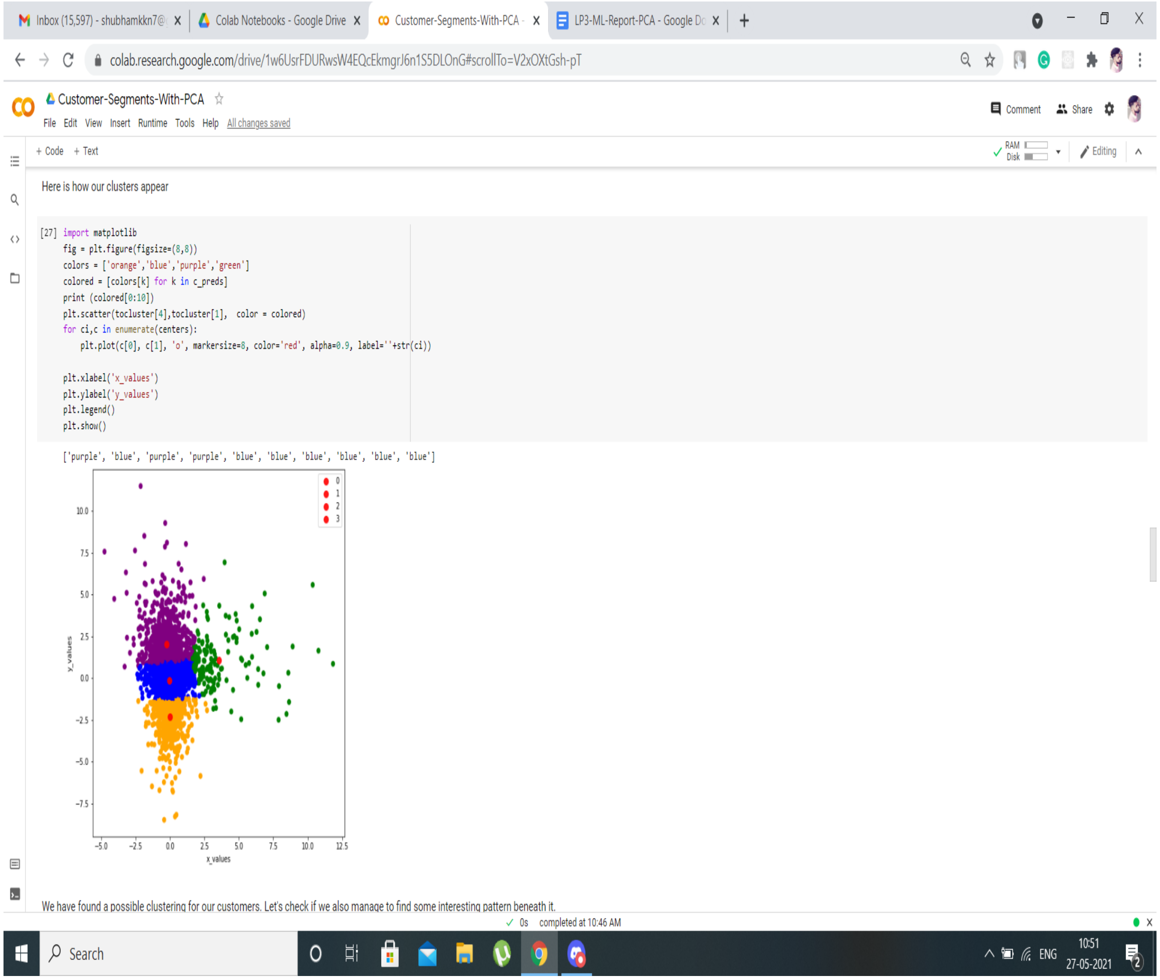
Inbox (15,597) - shubhamkkn7@
Colab Notebooks - Google Drive
Customer-Segments-With-PCA
LP3-ML-Report-PCA - Google Dc
colab.research.google.com/drive/1w6UsrFDURwsW4EQcEkmgrJ6n1S5DLOnG#scrollTo=V2xOXtGsh-pT
Customer-Segments-With-PCA
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
[24]

```
from matplotlib import pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from mpl_toolkits.mplot3d import proj3d
tocluster = pd.DataFrame(ps[[4,1]])
print (tocluster.shape)
print (tocluster.head())

fig = plt.figure(figsize=(8,8))
plt.plot(tocluster[4], tocluster[1], 'o', markersize=2, color='blue', alpha=0.5, label='class1')

plt.xlabel('x_values')
plt.ylabel('y_values')
plt.legend()
plt.show()
```

```
(6799, 2)
          4         1
0 -0.824579  1.033044
1 -0.076369 -0.501368
2 -1.241773  1.169656
3 -0.688362  1.099742
4 -0.084934 -0.405859
```


Inbox (15,597) - shubhamkkn7@
Colab Notebooks - Google Drive
Customer-Segments-With-PCA
LP3-ML-Report-PCA - Google Do
colab.research.google.com/drive/1w6UsrFDURwsW4EQcEkmgrJ6n1S5DLOnG#scrollTo=V2xOXtGsh-pT
Customer-Segments-With-PCA
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
0
1
2
3
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
y_values
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
x_values
✓ 0s completed at 10:46 AM
Search
ENG
10:51
27-05-2021

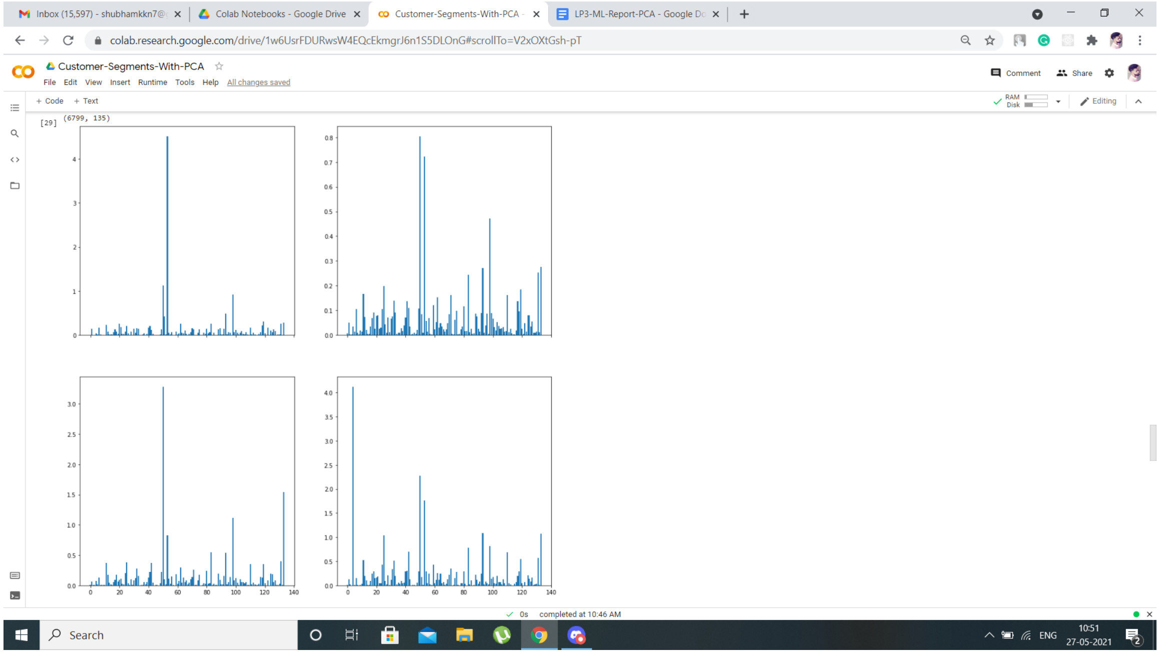
Here is how our clusters appear

```
[27] import matplotlib
     fig = plt.figure(figsize=(8,8))
     colors = ['orange','blue','purple','green']
     colored = [colors[k] for k in c_preds]
     print (colored[0:10])
     plt.scatter(tocluster[4],tocluster[1],  color = colored)
     for ci,c in enumerate(centers):
         plt.plot(c[0], c[1], 'o', markersize=8, color='red', alpha=0.9, label=''+str(ci))

     plt.xlabel('x_values')
     plt.ylabel('y_values')
     plt.legend()
     plt.show()
```

```
['purple', 'blue', 'purple', 'purple', 'blue', 'blue', 'blue', 'blue', 'blue', 'blue']
```

We have found a possible clustering for our customers. Let's check if we also manage to find some interesting pattern beneath it.


Inbox (15,597) - shubhamkkn7@
Colab Notebooks - Google Drive
Customer-Segments-With-PCA -
LP3-ML-Report-PCA - Google Do
colab.research.google.com/drive/1w6UsrFDURwsW4EQcEkmgrJ6n1S5DLOnG#scrollTo=V2xOXtGsh-pT
Customer-Segments-With-PCA
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
0 1 2 3 4
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 20 40 60 80 100 120 140
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0 20 40 60 80 100 120 140
0s completed at 10:46 AM
Search
ENG 10:51 27-05-2021

```
[29] (6799, 135)
```

## 8. CONCLUSION

Hence, we can see that as the number of principal components in the data increases the variance in the data increases and hence we can see that it becomes difficult to analyze the data if principal components increase. Thus, it is important to reduce dimensionality of the dataset to reduce computational power and remove redundant features.