Data Visualization Using Bokeh

Project submitted to the

SRM University – AP, Andhra Pradesh

for the course project of

CSE338 L Applied Data Science Lab

Submitted by

Ganga Srinivas Gollapalli (AP21110010262)

Madhukar Sai Babu Gadde (AP21110010277)

Maruti Mahesh Chowdary Gadde (AP21110010289)

Sai Ram Mallipeddi(AP21110010296)

Kowshik Veldhi (AP21110010301)

Rohith Kamal Kumar Yenduri (AP21110010303)

Under the Guidance of

Sabyasachi Dutta

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

May,2024

Acknowledgements

We would like to acknowledge all those without whom this project would not have been successful. Firstly, we would like to thank our Professor Dr Sabyasachi Dutta who guided us throughout the project and gave his immense support. He made us understand how to successfully complete this project and without him, the project would not have been complete.

This project has been a source to learn and bring our theoretical knowledge to the real-life world. So, we would really acknowledge his help and guidance for this project.

Table of Contents

Abstract

This project utilizes Bokeh, a renowned Python library for interactive data visualization, to analyze the Wine dataset, which comprises data from three cultivars of wine grown in the same region in Italy, featuring measurements of various chemical constituents. Through a series of visualizations including scatter plots, pairwise feature comparisons, and Principal Component Analysis (PCA) projections, intricate patterns within the dataset are unveiled. Bokeh's capabilities are showcased via interactive plots, facilitating detailed exploration of feature relationships and cultivar clustering. By leveraging Bokeh's tools, the project offers a nuanced understanding of the Wine dataset's characteristics, highlighting the effectiveness of Bokeh in visualizing complex datasets.

Introduction

In today's era of data-driven decision-making, understanding and interpreting data is paramount for optimizing strategies and achieving success. This project focuses on harnessing the power of data visualization techniques to gain insights into the Wine dataset.

Using Bokeh, a versatile Python library renowned for its ability to create interactive visualizations, our goal is to delve into various facets of the Wine dataset and uncover hidden patterns and trends within the data. Bokeh offers an intuitive platform for generating dynamic visualizations that can be explored and manipulated directly in a web browser. By presenting output graphs in a web page format, we ensure accessibility and ease of use for stakeholders and analysts alike.

A key technique employed in this project is Principal Component Analysis (PCA), a powerful dimensionality reduction method widely used in data analysis. PCA enables us to identify underlying patterns and relationships within high-dimensional datasets by transforming the original variables into a new set of uncorrelated variables known as principal components. By visualizing sales data based on segment and profit data based on category using PCA, we aim to gain deeper insights into the underlying structure of the Wine dataset.

Through this project, we aim to showcase the efficacy of Bokeh and data visualization techniques in extracting valuable insights from complex datasets. By empowering businesses with actionable insights, we endeavor to facilitate data-driven decision-making and enhance competitiveness in today's competitive landscape.

System Requirements

2.1. Hardware Requirements

- Windows 10 or Greater Version
- Stable Internet Connection

2.2. Software Requirements

- Python 3.0
- Python IDE (Anaconda Jupyter / Google Colab)
- Libraries (NumPy, Pandas, Seaborn, Matplotlib,Sklearn,Warnings)

Methodology

3.1. Description about the dataset:

The Wine dataset, a well-known dataset in the realm of machine learning and data science, is directly imported from the scikit-learn library. It consists of measurements from 178 samples of wines produced in the same region in Italy, each characterized by 13 features: Alcohol content, Malic Acid, Ash, Alcalinity of Ash, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280/OD315 of diluted wines, and Proline. Additionally, each sample is labeled with one of three cultivars: 'class_0', 'class_1', or 'class_2'. With its comprehensive feature set and labeled data, the Wine dataset is widely utilized for classification and regression tasks, making it a valuable resource for exploratory analysis and predictive modeling.

3.2. Programming Language

The entire project was implemented using the Python programming language. Python's simplicity, versatility, and extensive library ecosystem make it an ideal choice for data analysis and visualization tasks. Python's intuitive syntax facilitated rapid development and easy debugging, enabling seamless execution of the project.

3.3. Libraries Used

The project leveraged several key Python libraries to perform data visualization using Bokeh 2.1.1 for the Wine dataset:

- Pandas: Pandas played a crucial role in data manipulation and transformation. It allowed for efficient structuring of the dataset, making it easy to work with its features and labels.
- NumPy: NumPy provided essential support for numerical computations. It enabled the project to perform various mathematical operations required for data analysis and visualization.
- Bokeh 2.1.1: Bokeh is a powerful Python library for creating interactive visualizations. With its extensive range of chart types and interactive features, Bokeh enabled the creation of dynamic and visually appealing plots directly in Python code.

3.4. Dataset Source - scikit-learn:

The Wine dataset used in this project was imported directly from the scikit-learn library. Scikit-learn provides a wide range of datasets for educational and testing purposes, including the Wine dataset. Importing the dataset from scikit-learn ensured its reliability and consistency, making it a suitable choice for data visualization and analysis using Bokeh 2.1.1.

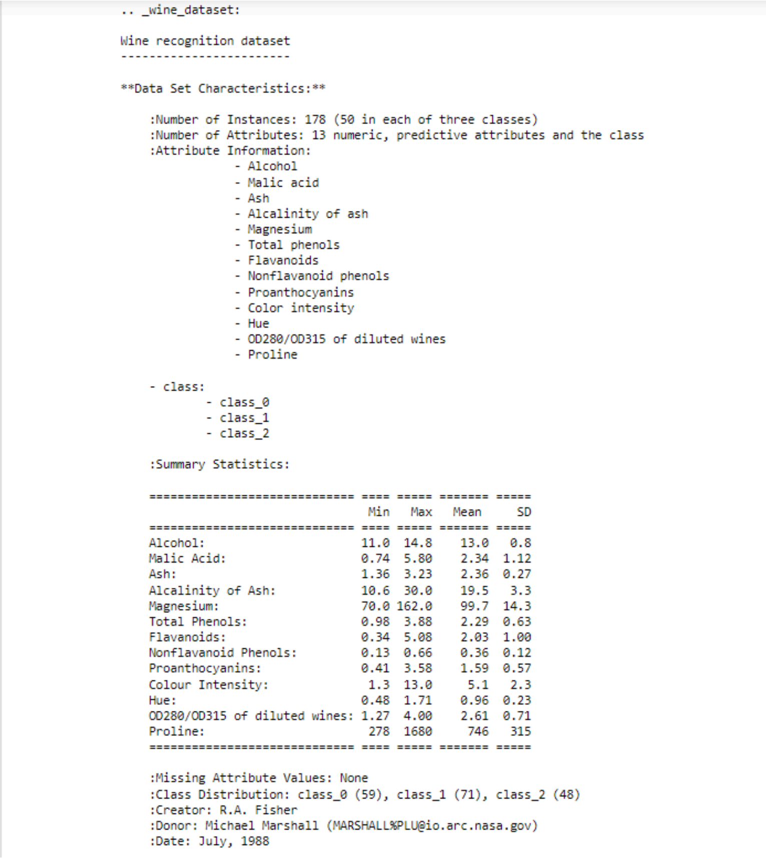
Implementation & Results

Import Data Set

```
from sklearn.datasets import load_wine

wine = load_wine()

print(wine.DESCR)

X = wine.data

y = wine.target
```

Output:

The code Imports the data set from scikit-learn library and gives the data set characteristics

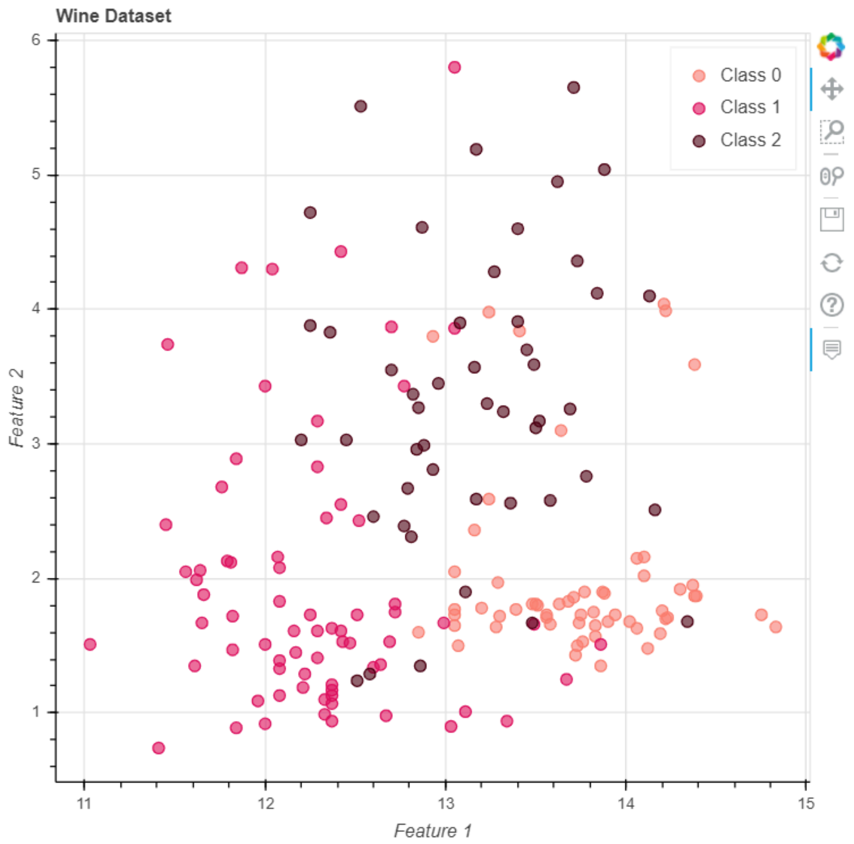


```
.. _wine_dataset:

Wine recognition dataset
------------------------

**Data Set Characteristics:**

    :Number of Instances: 178 (50 in each of three classes)
    :Number of Attributes: 13 numeric, predictive attributes and the class
    :Attribute Information:
                - Alcohol
                - Malic acid
                - Ash
                - Alcalinity of ash
                - Magnesium
                - Total phenols
                - Flavanoids
                - Nonflavanoid phenols
                - Proanthocyanins
                - Color intensity
                - Hue
                - OD280/OD315 of diluted wines
                - Proline

    - class:
            - class_0
            - class_1
            - class_2

    :Summary Statistics:

    ============================= ==== ===== ======= =====
                                   Min   Max   Mean     SD
    ============================= ==== ===== ======= =====
    Alcohol:                      11.0  14.8    13.0   0.8
    Malic Acid:                   0.74  5.80    2.34  1.12
    Ash:                          1.36  3.23    2.36  0.27
    Alcalinity of Ash:            10.6  30.0    19.5   3.3
    Magnesium:                    70.0 162.0    99.7  14.3
    Total Phenols:                0.98  3.88    2.29  0.63
    Flavanoids:                   0.34  5.08    2.03  1.00
    Nonflavanoid Phenols:         0.13  0.66    0.36  0.12
    Proanthocyanins:              0.41  3.58    1.59  0.57
    Colour Intensity:              1.3  13.0     5.1   2.3
    Hue:                          0.48  1.71    0.96  0.23
    OD280/OD315 of diluted wines: 1.27  4.00    2.61  0.71
    Proline:                       278  1680     746   315
    ============================= ==== ===== ======= =====

    :Missing Attribute Values: None
    :Class Distribution: class_0 (59), class_1 (71), class_2 (48)
    :Creator: R.A. Fisher
    :Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
    :Date: July, 1988
```

Interactive Scatter Plot for Wine data set

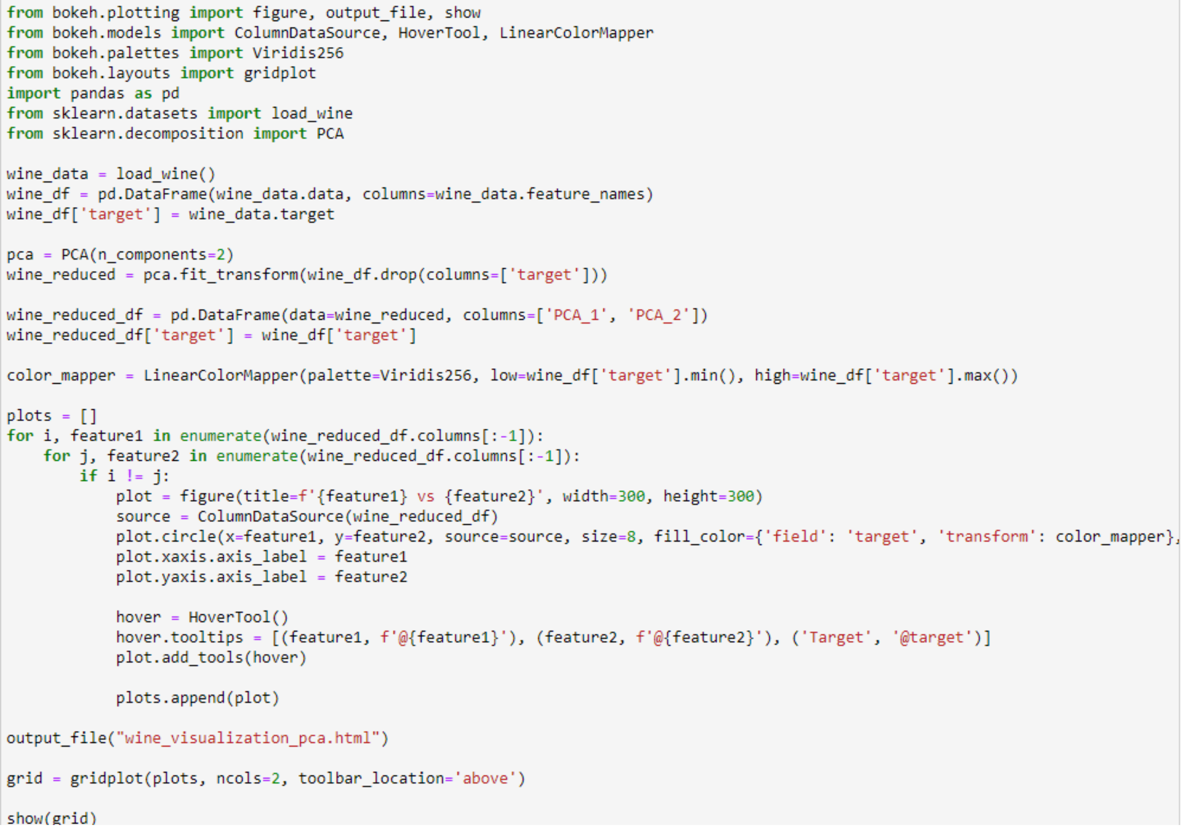
```
from bokeh.plotting import figure, show
from bokeh.models import ColumnDataSource, HoverTool
from sklearn.datasets import load_wine

wine = load_wine()

class_colors = ['#FA8072', '#E0115F', '#4C0013']
colors = [class_colors[target] for target in wine.target]

source = ColumnDataSource(data=dict(
    x=wine.data[:, 0],
    y=wine.data[:, 1],
    cultivar=[f'Class {i}' for i in wine.target],
    color=colors
))

plot = figure(title="Wine Dataset", x_axis_label='Feature 1', y_axis_label='Feature 2')

plot.circle('x', 'y', size=8, source=source, legend_field='cultivar',
            color='color', fill_alpha=0.6)

hover = HoverTool()
hover.tooltips = [("Cultivar", "@cultivar"), ("Feature 1", "@x"), ("Feature 2", "@y")]
plot.add_tools(hover)

show(plot)
```

Output:



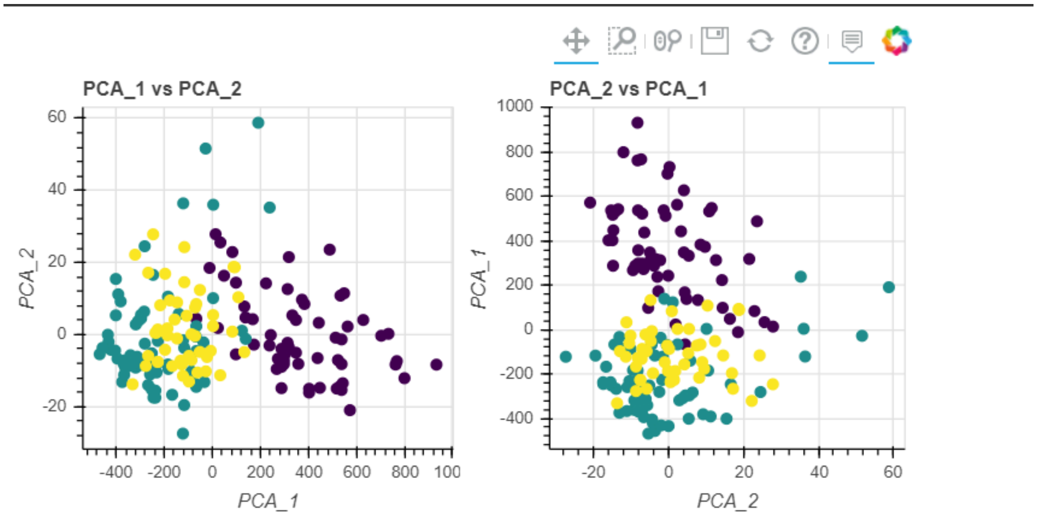
PCA Visualization of Wine Data Set Features

- No:of Principal Components : 2

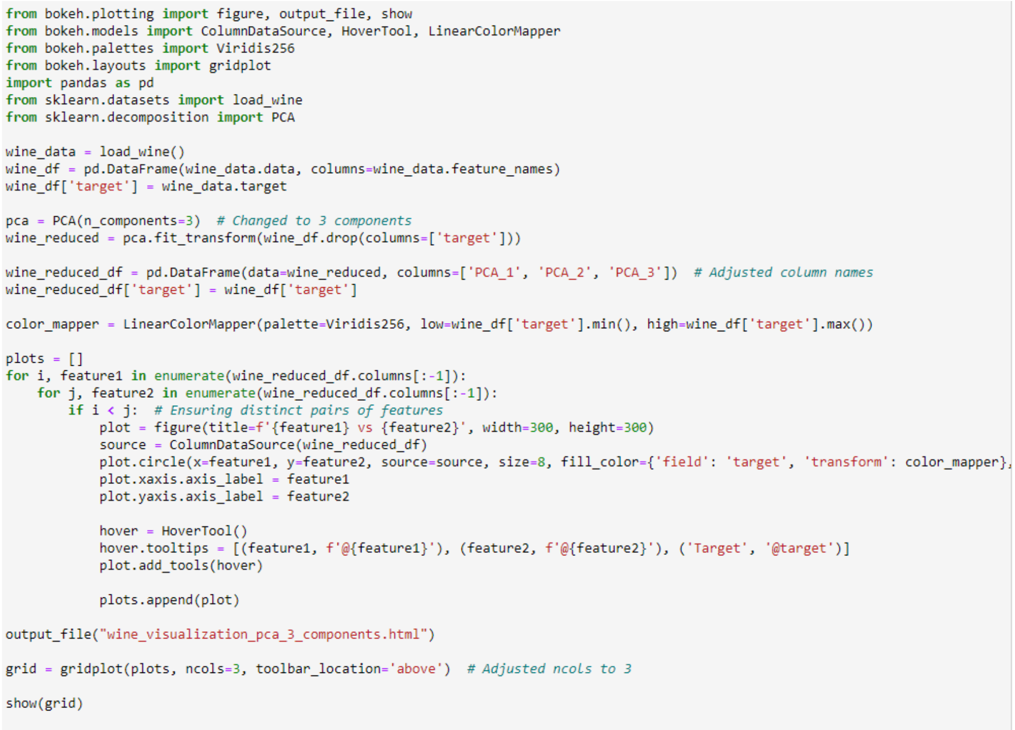


```
from bokeh.plotting import figure, output_file, show
from bokeh.models import ColumnDataSource, HoverTool, LinearColorMapper
from bokeh.palettes import Viridis256
from bokeh.layouts import gridplot
import pandas as pd
from sklearn.datasets import load_wine
from sklearn.decomposition import PCA

wine_data = load_wine()
wine_df = pd.DataFrame(wine_data.data, columns=wine_data.feature_names)
wine_df['target'] = wine_data.target

pca = PCA(n_components=2)
wine_reduced = pca.fit_transform(wine_df.drop(columns=['target']))

wine_reduced_df = pd.DataFrame(data=wine_reduced, columns=['PCA_1', 'PCA_2'])
wine_reduced_df['target'] = wine_df['target']

color_mapper = LinearColorMapper(palette=Viridis256, low=wine_df['target'].min(), high=wine_df['target'].max())

plots = []
for i, feature1 in enumerate(wine_reduced_df.columns[:-1]):
    for j, feature2 in enumerate(wine_reduced_df.columns[:-1]):
        if i != j:
            plot = figure(title=f'{feature1} vs {feature2}', width=300, height=300)
            source = ColumnDataSource(wine_reduced_df)
            plot.circle(x=feature1, y=feature2, source=source, size=8, fill_color={'field': 'target', 'transform': color_mapper},
            plot.xaxis.axis_label = feature1
            plot.yaxis.axis_label = feature2

            hover = HoverTool()
            hover.tooltips = [(feature1, f'@{feature1}'), (feature2, f'@{feature2}'), ('Target', '@target')]
            plot.add_tools(hover)

            plots.append(plot)

output_file("wine_visualization_pca.html")

grid = gridplot(plots, ncols=2, toolbar_location='above')

show(grid)
```

Output

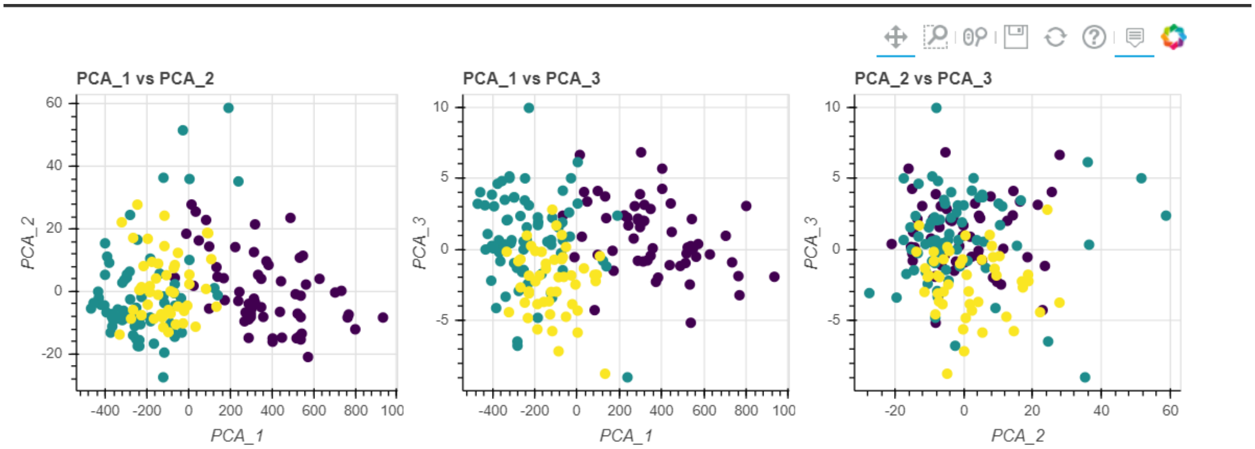


- No:of Principal Components : 3

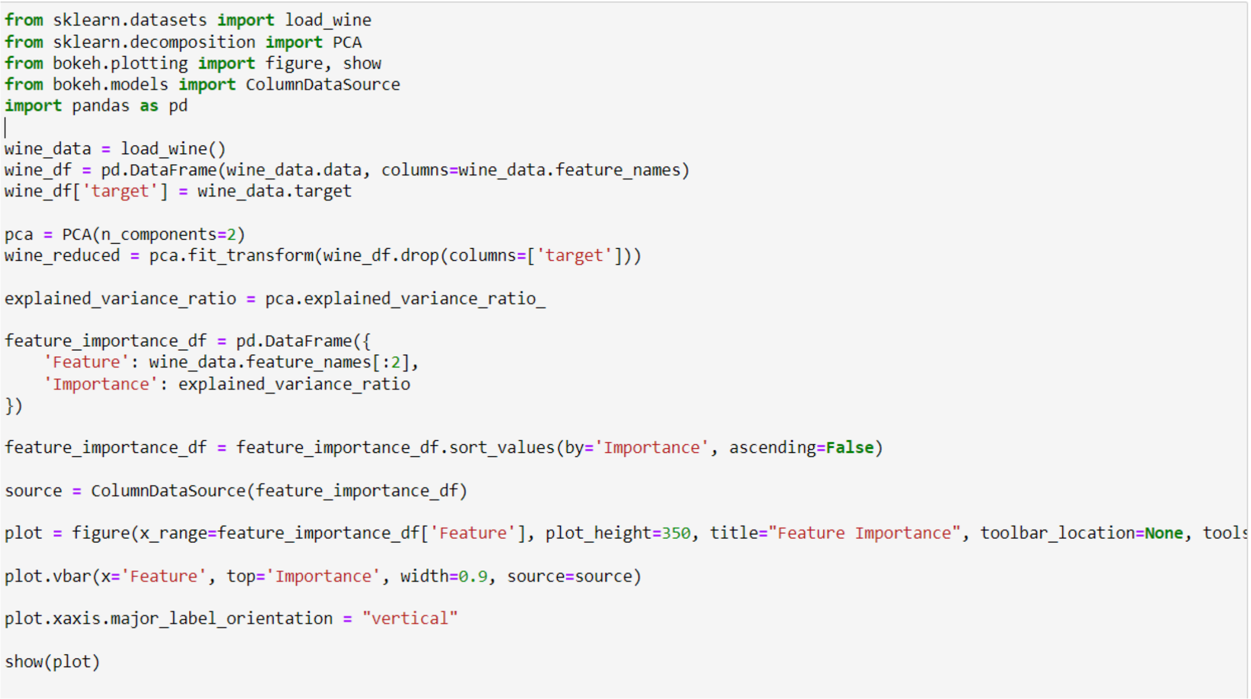


```
from bokeh.plotting import figure, output_file, show
from bokeh.models import ColumnDataSource, HoverTool, LinearColorMapper
from bokeh.palettes import Viridis256
from bokeh.layouts import gridplot
import pandas as pd
from sklearn.datasets import load_wine
from sklearn.decomposition import PCA

wine_data = load_wine()
wine_df = pd.DataFrame(wine_data.data, columns=wine_data.feature_names)
wine_df['target'] = wine_data.target

pca = PCA(n_components=3)  # Changed to 3 components
wine_reduced = pca.fit_transform(wine_df.drop(columns=['target']))

wine_reduced_df = pd.DataFrame(data=wine_reduced, columns=['PCA_1', 'PCA_2', 'PCA_3'])  # Adjusted column names
wine_reduced_df['target'] = wine_df['target']

color_mapper = LinearColorMapper(palette=Viridis256, low=wine_df['target'].min(), high=wine_df['target'].max())

plots = []
for i, feature1 in enumerate(wine_reduced_df.columns[:-1]):
    for j, feature2 in enumerate(wine_reduced_df.columns[:-1]):
        if i < j:  # Ensuring distinct pairs of features
            plot = figure(title=f'{feature1} vs {feature2}', width=300, height=300)
            source = ColumnDataSource(wine_reduced_df)
            plot.circle(x=feature1, y=feature2, source=source, size=8, fill_color={'field': 'target', 'transform': color_mapper},
            plot.xaxis.axis_label = feature1
            plot.yaxis.axis_label = feature2

            hover = HoverTool()
            hover.tooltips = [(feature1, f'@{feature1}'), (feature2, f'@{feature2}'), ('Target', '@target')]
            plot.add_tools(hover)

            plots.append(plot)

output_file("wine_visualization_pca_3_components.html")

grid = gridplot(plots, ncols=3, toolbar_location='above')  # Adjusted ncols to 3

show(grid)
```

Output

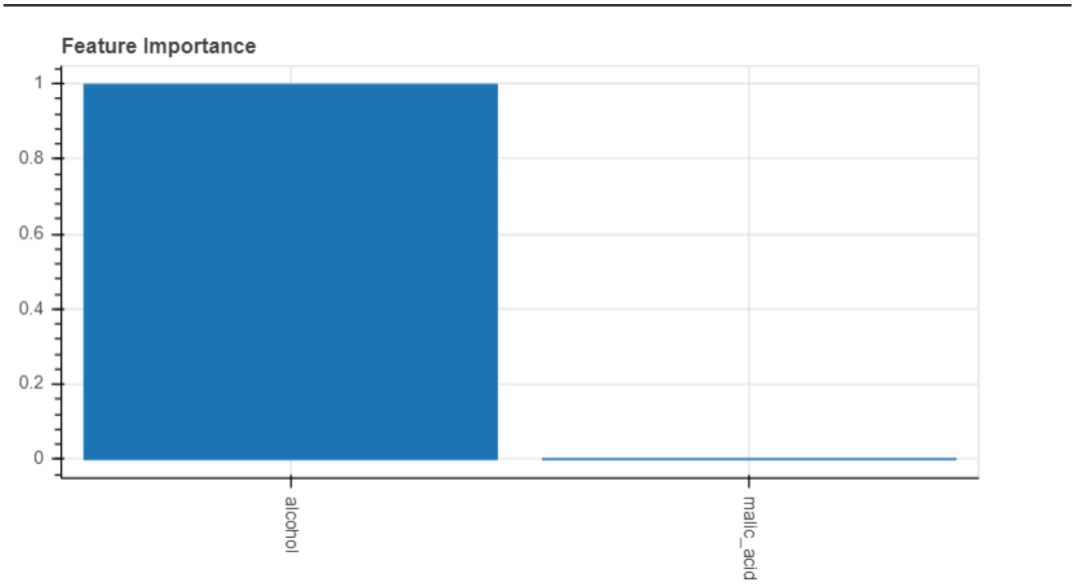


Feature Importance for PCA with N=2



```
from sklearn.datasets import load_wine
from sklearn.decomposition import PCA
from bokeh.plotting import figure, show
from bokeh.models import ColumnDataSource
import pandas as pd

wine_data = load_wine()
wine_df = pd.DataFrame(wine_data.data, columns=wine_data.feature_names)
wine_df['target'] = wine_data.target

pca = PCA(n_components=2)
wine_reduced = pca.fit_transform(wine_df.drop(columns=['target']))

explained_variance_ratio = pca.explained_variance_ratio_

feature_importance_df = pd.DataFrame({
    'Feature': wine_data.feature_names[:2],
    'Importance': explained_variance_ratio
})

feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=False)

source = ColumnDataSource(feature_importance_df)

plot = figure(x_range=feature_importance_df['Feature'], plot_height=350, title="Feature Importance", toolbar_location=None, tools

plot.vbar(x='Feature', top='Importance', width=0.9, source=source)

plot.xaxis.major_label_orientation = "vertical"

show(plot)
```

Output



Conclusion

Through the utilization of Bokeh for data visualization on the Wine dataset, this project has successfully demonstrated the power of interactive visualizations in gaining insights from complex datasets. By leveraging Bokeh's capabilities, we have been able to explore and analyze various features of the Wine dataset, including the relationships between different attributes and the distribution of data points.

The visualizations created using Bokeh have provided a clear and intuitive representation of the Wine dataset, allowing us to identify patterns, trends, and anomalies in the data. From scatter plots showcasing the relationships between sepal length and width to PCA visualizations highlighting the principal components contributing to the variance in the dataset, Bokeh has enabled us to uncover valuable insights that may not have been apparent from the raw data alone.

Furthermore, Bokeh's interactive features, such as hover tooltips and zooming capabilities, have enhanced the exploratory data analysis process, allowing users to interactively explore the data and gain a deeper understanding of its underlying structure.

Overall, the use of Bokeh for data visualization on the Wine dataset has been instrumental in facilitating data-driven decision-making and enhancing our understanding of the dataset. By visualizing complex datasets in an intuitive and interactive manner, Bokeh empowers users to extract meaningful insights and drive actionable outcomes.

References

Bokeh Documentation : https://docs.bokeh.org/en/latest/index.html

Scikit-learn Documentation : https://scikit-learn.org/stable/documentation.html