**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**🙡🕮🙣**

Logo

Description automatically generated

**Report Lab 03**

**Practice #3 - Scikit-learn, PCA**

**Môn: Phân tích thống kê dữ liệu nhiều biến**

**Giáo viên hướng dẫn:** Nguyễn Mạnh Hùng

Lý Quốc Ngọc

Phạm Thanh Tùng

**Lớp:** 20TGMT01

**Sinh viên thực hiện:** Giang Gia Bảo – 20127446

*Tp Hồ Chí Minh, 2 tháng 05 năm 2023*

**MỤC LỤC**

[I. Thông tin sinh viên 1](#_Toc133978928)

[II. Đánh giá mức độ hoàn thành 1](#_Toc133978929)

[III. Mô tả data CSV-file 1](#_Toc133978930)

[1. Giới thiệu về Dineout dataset 1](#_Toc133978931)

[2. Mô tả chi tiết 1](#_Toc133978932)

[3. Bảng tóm tắt các thông số của dataset 2](#_Toc133978933)

[IV. Data Visualization 2](#_Toc133978934)

[1. Biểu đồ dữ liệu da biến 2](#_Toc133978935)

[2. Biểu đồ thể hiện 3 thông số Rating, Votes và Cost 2](#_Toc133978936)

[3. Biểu đồ thể hiện chỉ số Rating và lượt votes tất cả nhà hàng của mỗi thành phố 3](#_Toc133978937)

[4. Biểu đồ thể hiện chỉ số Rating và Cost tất cả nhà hàng của mỗi thành phố 4](#_Toc133978938)

[5. Biểu đồ thể hiện số Votes và Cost tất cả nhà hàng của mỗi thành phố 4](#_Toc133978939)

[V. Tính toán các đại lượng thống kê cơ bản 5](#_Toc133978940)

[1. Mean, Standard deviation, Max, Min 5](#_Toc133978941)

[2. Tính các đại lượng thống kê cho một group cụ thể (City) 6](#_Toc133978942)

[3. Phương sai trong group 7](#_Toc133978943)

[4. Phương sai giữa các group 7](#_Toc133978944)

[5. Sự khả tách cho từng biến 7](#_Toc133978945)

[6. Hiệp phương sai trong group 7](#_Toc133978946)

[7. Hiệp phương sai giữa các group 7](#_Toc133978947)

[8. Biểu đồ heatmap tương quan cho dữ liệu đa biến 7](#_Toc133978948)

[9. Biểu đồ Hinton trực quan hóa ma trận trọng số 8](#_Toc133978949)

[10. Sự tương quan giữa các biến 9](#_Toc133978950)

[VI. PCA and visualization 9](#_Toc133978951)

[1. Tiêu chuẩn hóa các biến 9](#_Toc133978952)

[2. Phân tích thành phần chính 9](#_Toc133978953)

[3. Biểu đồ độ lệch chuẩn của các thành phần chính 10](#_Toc133978954)

[4. Biểu đồ scatter của 2 thành phần chính cho class “City” 10](#_Toc133978955)

[VII. LDA and visualization 11](#_Toc133978956)

[1. Tiêu chuẩn hóa các biến 11](#_Toc133978957)

[2. Phân tích thành phần chính 11](#_Toc133978958)

[3. Biểu đồ độ lệch chuẩn của các thành phần chính 12](#_Toc133978959)

[4. Biểu đồ scatter của 2 thành phần chính của class “City” 12](#_Toc133978960)

[VIII. Reference 13](#_Toc133978961)

1. Thông tin sinh viên

|  |  |  |  |
| --- | --- | --- | --- |
| **Họ và tên** | **MSSV** | **Lớp** | **Note** |
| Giang Gia Bảo | 20127446 | 20TGMT01 |  |

1. Đánh giá mức độ hoàn thành

|  |  |  |
| --- | --- | --- |
| **STT** | **Yêu cầu** | **Mức độ hoàn thành** |
| 1 | Find a data CSV file, describe the data info | 100% |
| 2 | Implement some basic multivariate analysis with visualization | 100% |
| 3 | Apply PCA & LDA by using Scikit-learn with the chose data | 100% |

1. Mô tả data CSV-file
   1. Giới thiệu về Dineout dataset

- Tên dataset: The Dineout dataset

- Link dataset: [The Dineout dataset](https://www.kaggle.com/datasets/arnabchaki/indian-restaurants-2023/download?datasetVersionNumber=1)

- Đây là tập dữ liệu nói về các đánh giá của khách hàng về các nhà hàng ở những thành phố chính của Ấn Độ.

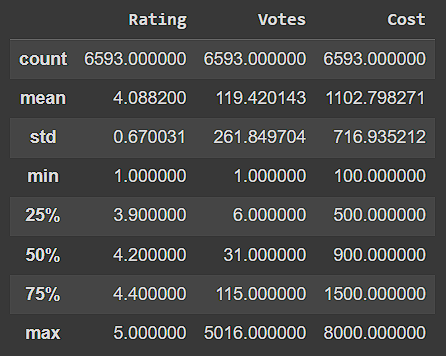
- Tập dữ liệu được thu thập từ trang web công khai https://www.dineout.co.in/

* 1. Mô tả chi tiết

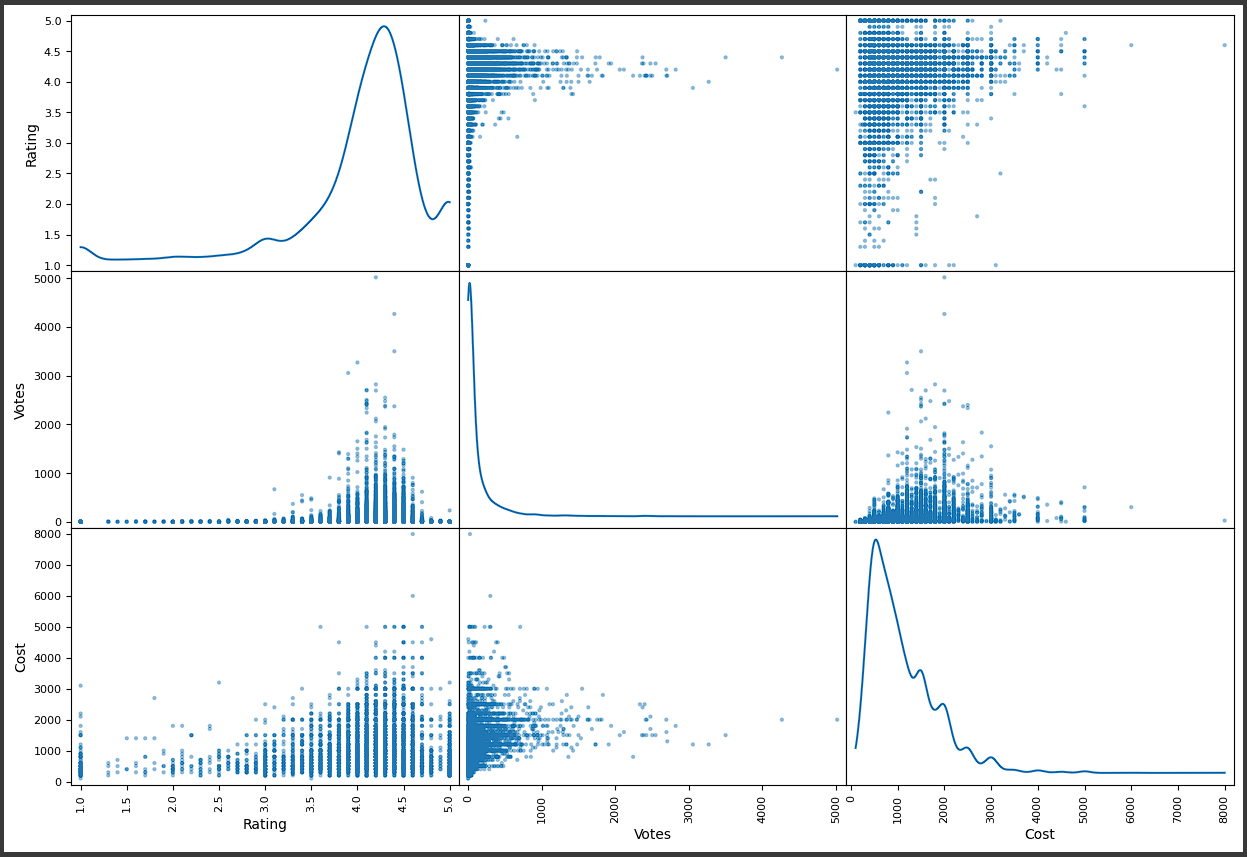
- Trong tập dữ liệu có: 8 cột và 6594 dòng

* Trong đó 8 cột đại diện cho các thuộc tính như:
* Name: tên của nhà hàng
* Location: vị trí của nhà hàng
* Locality: vị trí tương đối ở thành phố mà nhà hàng trực thuộc
* City: tên thành phố nhà hàng trực thuộc
* Cuisine: xu hướng ẩm thực của nhà hàng
* Rating: điểm vote của nhà hàng (0 – 5)
* Votes: số lượt vote của nhà hàng
* Cost: giá trung bình của nhà hàng
  1. Bảng tóm tắt các thông số của dataset

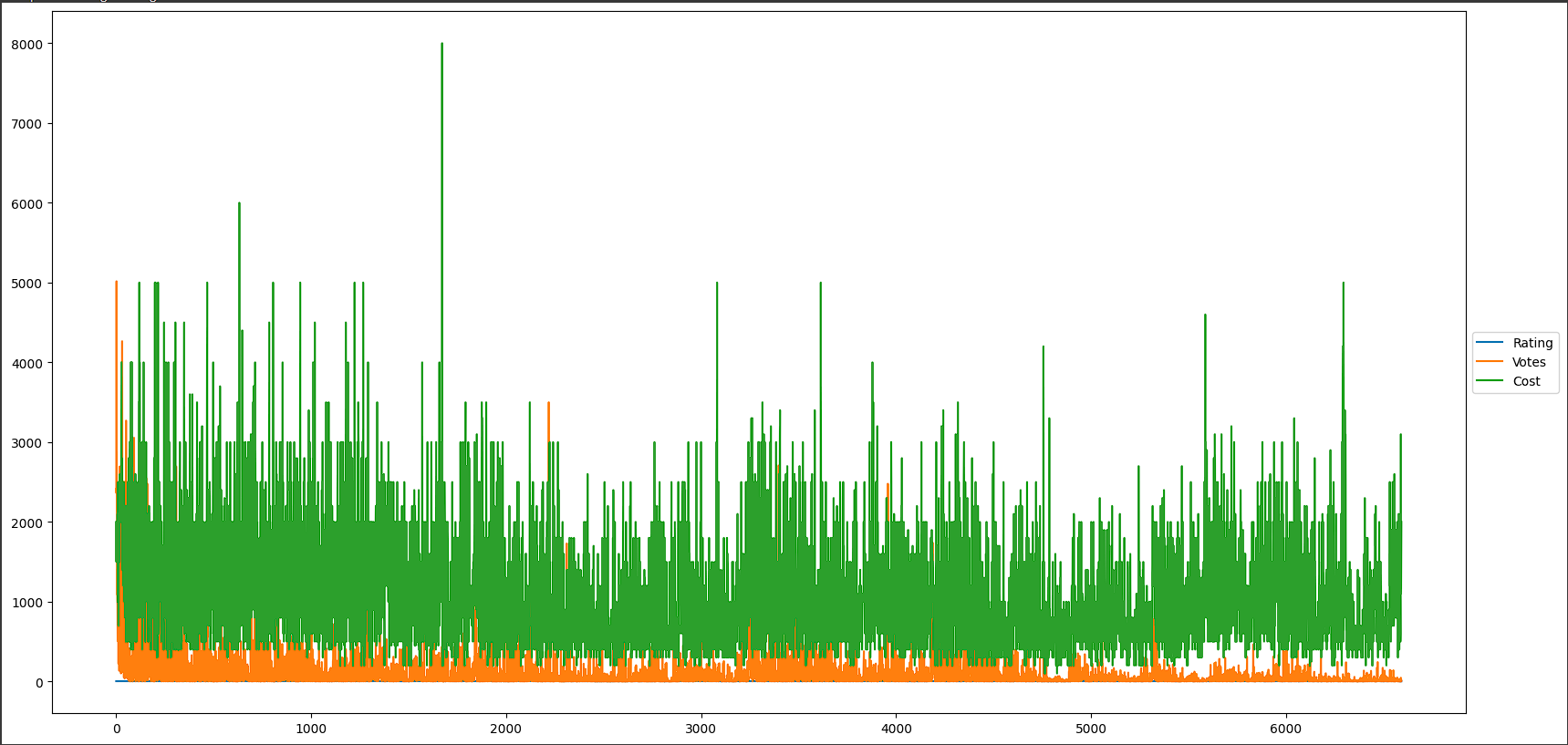
Bảng tóm tắt của dataset về các thông số như: số lượng, trung bình, độ lệch chuẩn, min/max



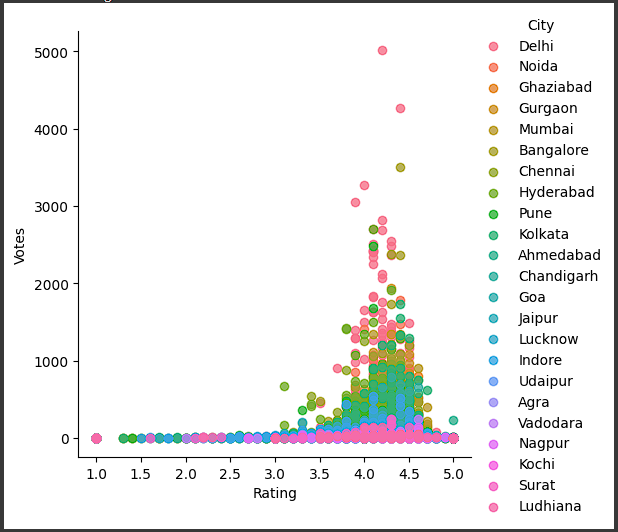
1. Data Visualization
   1. Biểu đồ dữ liệu da biến



* 1. Biểu đồ thể hiện 3 thông số Rating, Votes và Cost



* 1. Biểu đồ thể hiện chỉ số Rating và lượt votes tất cả nhà hàng của mỗi thành phố



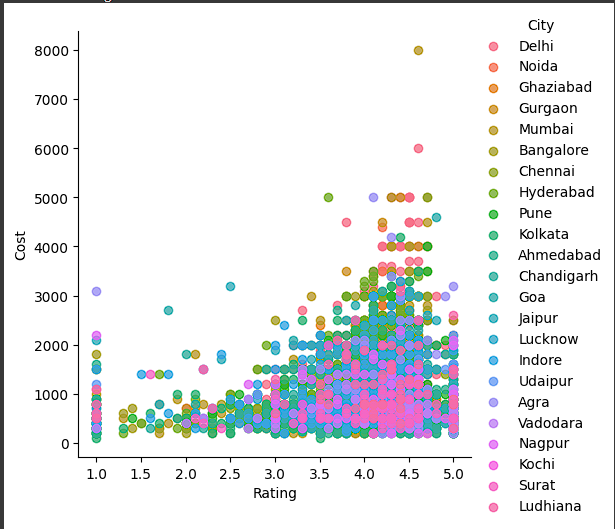
Nhận xét:

- Biên độ dao động của Rating ở khoảng [3.7, 4.6]

- Các thành phố có lượt vote cao nhất thường có điểm rating ở mức 3.8 – 4.6

- Delhi là thành phố có tỉ lệ giữa Rating và Votes cao nhất

* 1. Biểu đồ thể hiện chỉ số Rating và Cost tất cả nhà hàng của mỗi thành phố

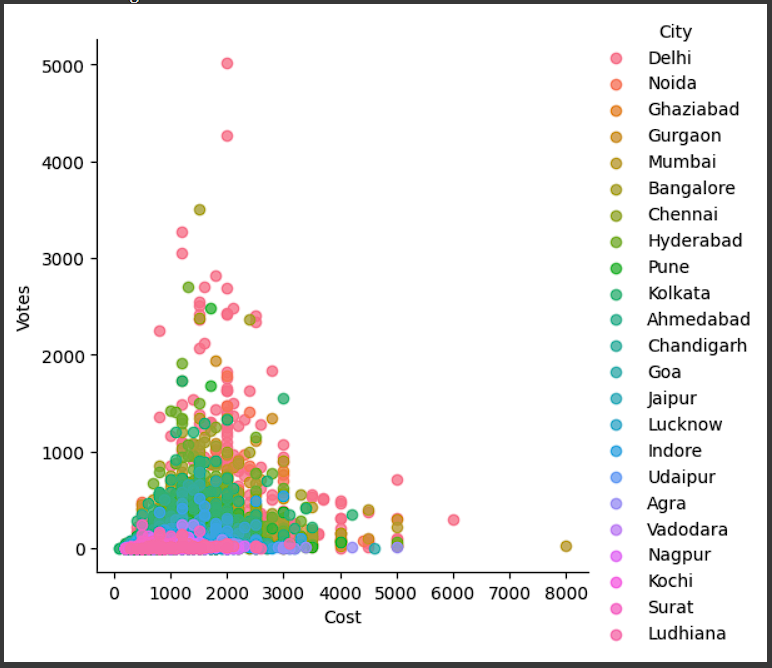


Nhận xét:

- Biên độ dao động của Rating đa số ở khoảng [3.0, 4.8]

- Giá của hầu hết nhà hàng trong mỗi thành phố rơi vào mức 200 - 3000

* 1. Biểu đồ thể hiện số Votes và Cost tất cả nhà hàng của mỗi thành phố

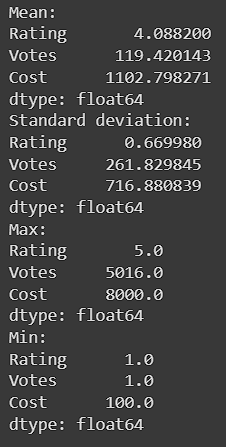


Nhận xét:

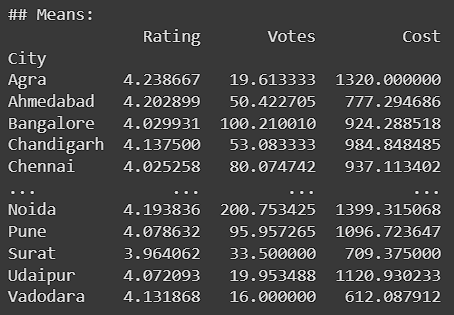
- Giá của các nhà hàng trong thành phố dao động từ 200 - 3000

- Các thành phố có lượt vote nhiểu nhất có giá nằm ở mức 1000 - 2000

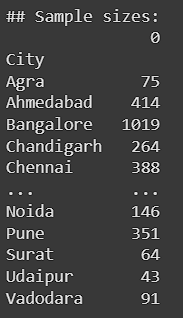
1. Tính toán các đại lượng thống kê cơ bản
   1. Mean, Standard deviation, Max, Min



* 1. Tính các đại lượng thống kê cho một group cụ thể (City)







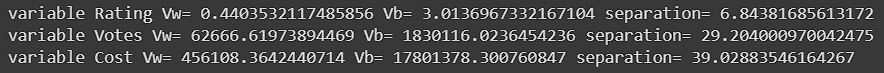
* 1. Phương sai trong group



* 1. Phương sai giữa các group



* 1. Sự khả tách cho từng biến



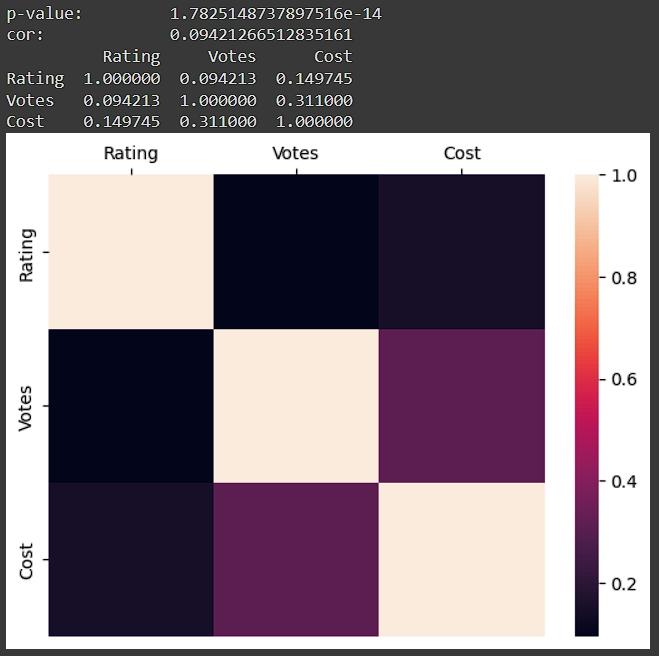
* 1. Hiệp phương sai trong group



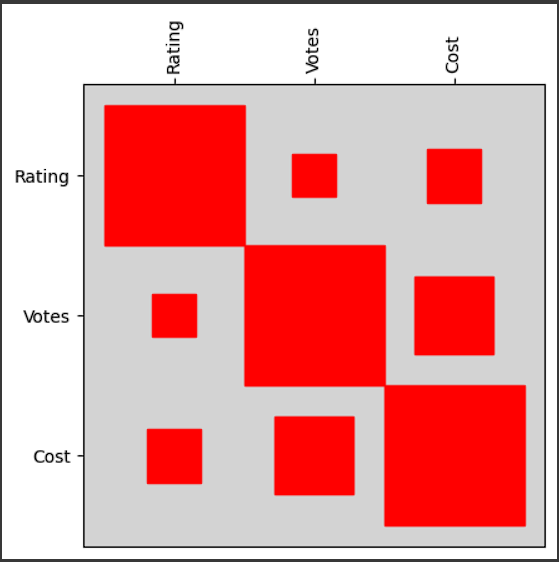
* 1. Hiệp phương sai giữa các group



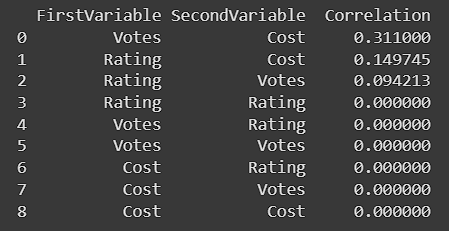
* 1. Biểu đồ heatmap tương quan cho dữ liệu đa biến



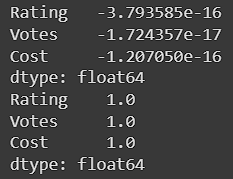
* 1. Biểu đồ Hinton trực quan hóa ma trận trọng số



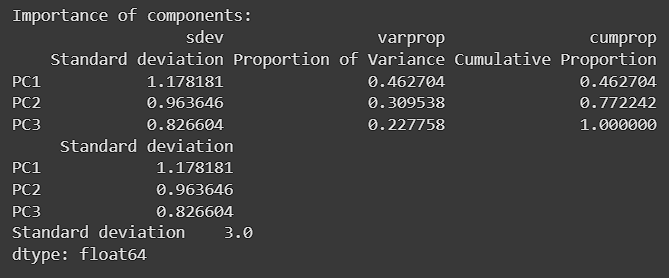
* 1. Sự tương quan giữa các biến



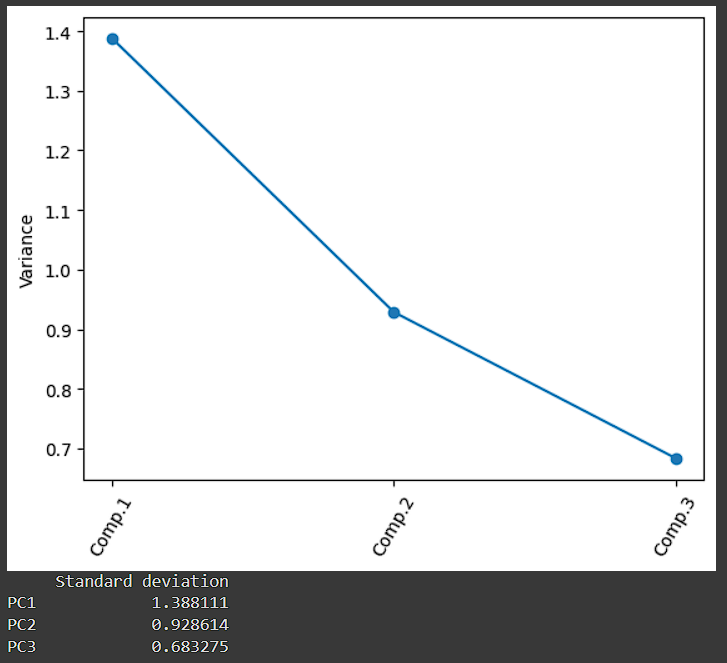
1. PCA and visualization
   1. Tiêu chuẩn hóa các biến



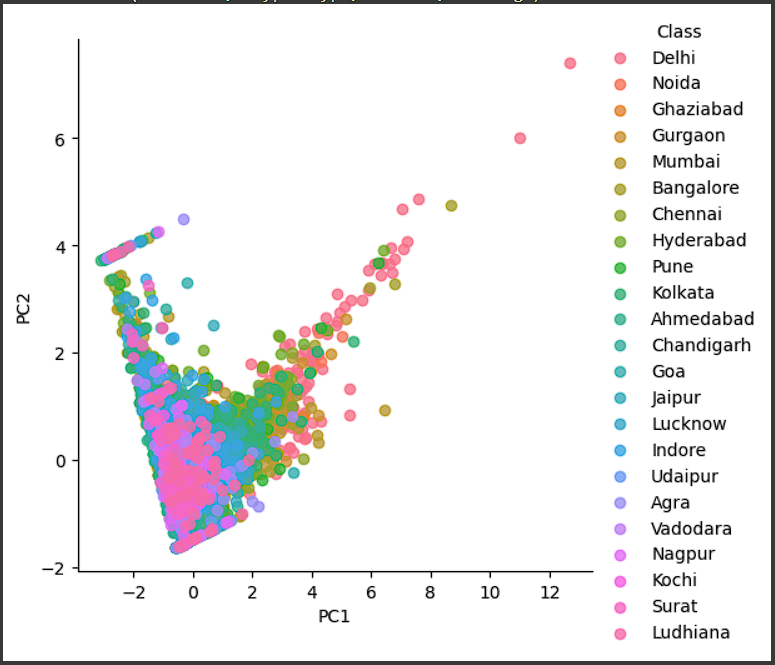
* 1. Phân tích thành phần chính



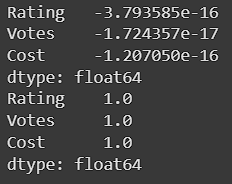
* 1. Biểu đồ độ lệch chuẩn của các thành phần chính



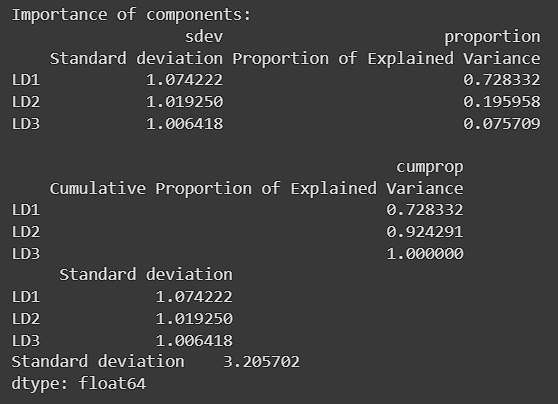
* 1. Biểu đồ scatter của 2 thành phần chính cho class “City”



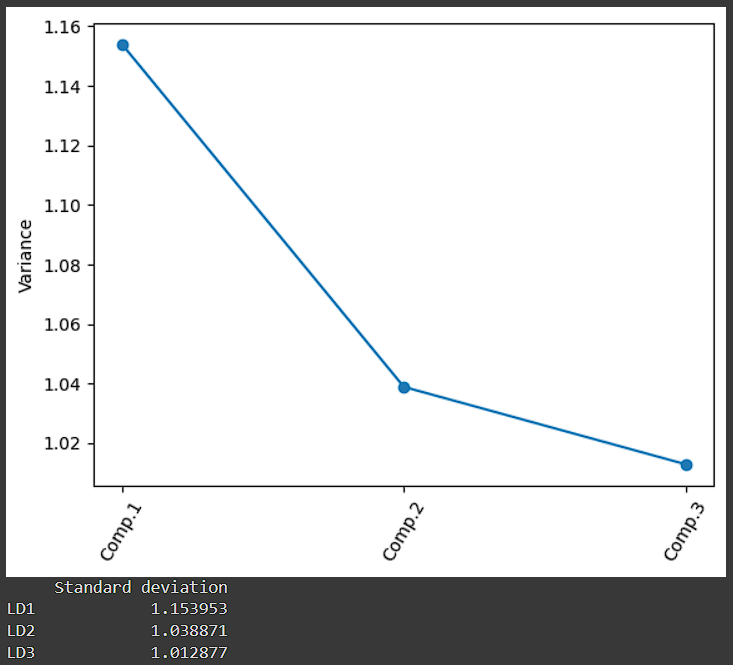
1. LDA and visualization
   1. Tiêu chuẩn hóa các biến



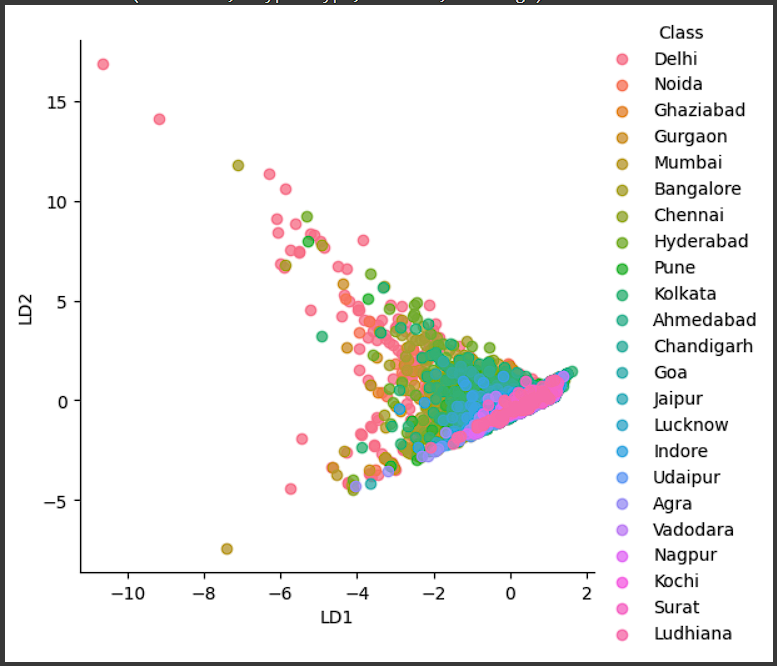
* 1. Phân tích thành phần chính



* 1. Biểu đồ độ lệch chuẩn của các thành phần chính



* 1. Biểu đồ scatter của 2 thành phần chính của class “City”



1. Reference

1. [Sample source](https://courses.ctda.hcmus.edu.vn/mod/resource/view.php?id=66863)

2. [Dataset](https://www.kaggle.com/datasets/arnabchaki/indian-restaurants-2023/download?datasetVersionNumber=1)