

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH



## Report Lab 03

### Practice #3 - Scikit-learn with PCA & LDA

**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, 02 tháng 05 năm 2023*

# MỤC LỤC

<b>I. Thông tin sinh viên .....</b>	<b>1</b>
<b>II. Đánh giá mức độ hoàn thành .....</b>	<b>1</b>
<b>III. Mô tả data CSV-file.....</b>	<b>1</b>
1. Giới thiệu về Dineout dataset.....	1
2. Mô tả chi tiết.....	1
3. Bảng tóm tắt các thông số của dataset .....	2
<b>IV. Data Visualization.....</b>	<b>2</b>
1. Biểu đồ thể hiện 3 thông số Rating, Votes và Cost.....	2
2. Biểu đồ thể hiện dữ liệu đa biến.....	2
3. Biểu đồ thể hiện chỉ số Rating và lượt votes tất cả nhà hàng của mỗi thành phố.....	3
4. Biểu đồ thể hiện chỉ số Rating và Cost tất cả nhà hàng của mỗi thành phố.....	4
5. Biểu đồ thể hiện số Votes và Cost tất cả nhà hàng của mỗi thành phố ....	4
<b>V. Tính toán các đại lượng thống kê cơ bản.....</b>	<b>5</b>
1. Mean, Standard deviation, Max, Min .....	5
2. Tính các đại lượng thống kê cho một group cụ thể (City) .....	6
3. Phương sai trong group .....	7
4. Phương sai giữa các group .....	7
5. Sự khả tách cho từng biến.....	7
6. Hiệp phương sai trong group .....	7
7. Hiệp phương sai giữa các group .....	7
8. Biểu đồ heatmap tương quan cho dữ liệu đa biến .....	7
9. Biểu đồ Hinton trực quan hóa ma trận trọng số.....	8
10. Sự tương quan giữa các biến.....	9

<b>VI. PCA and visualization .....</b>	<b>9</b>
1. Tiêu chuẩn hóa các biến .....	9
2. Phân tích thành phần chính .....	9
3. Biểu đồ độ lệch chuẩn của các thành phần chính .....	10
4. Biểu đồ scatter của 2 thành phần chính cho class “City” .....	10
<b>VII. LDA and visualization .....</b>	<b>11</b>
1. Tiêu chuẩn hóa các biến .....	11
2. Phân tích thành phần chính .....	11
3. Biểu đồ độ lệch chuẩn của các thành phần chính .....	12
4. Biểu đồ scatter của 2 thành phần chính của class “City” .....	12
<b>VIII. Reference .....</b>	<b>13</b>

## I. Thông tin sinh viên

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

## II. 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%

## III. Mô tả data CSV-file

### 1. Giới thiệu về Dineout dataset

- Tên dataset: The Dineout dataset
- Link dataset: [The Dineout dataset](https://www.dineout.co.in/)
- 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. Trong đó bao gồm Rating, Votes,
- Tập dữ liệu được thu thập từ trang web công khai <https://www.dineout.co.in/>

### 2. 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) (theo đánh giá của khách)
    - Votes: số lượng lượt vote của nhà hàng
    - Cost: giá trung bình của nhà hàng (theo đánh giá của khách)

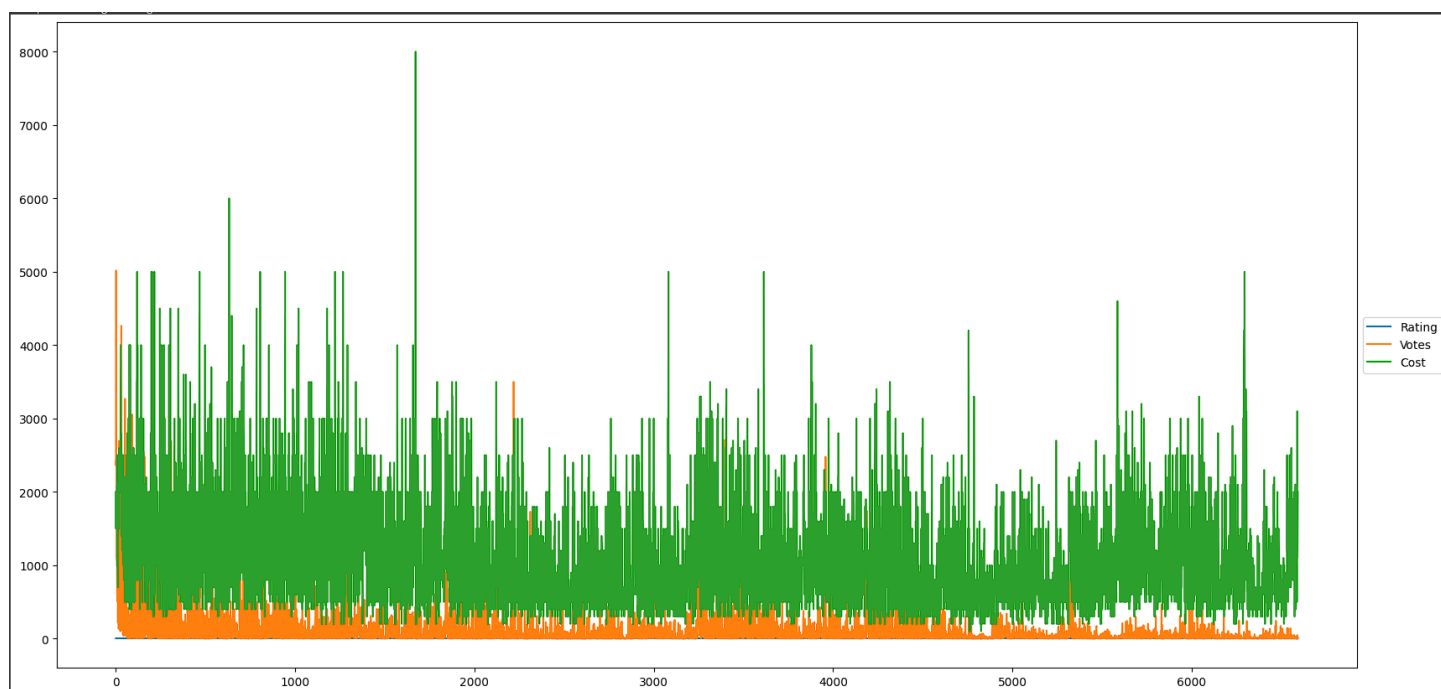
### 3. Bảng tóm tắt các thông số của dataset

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

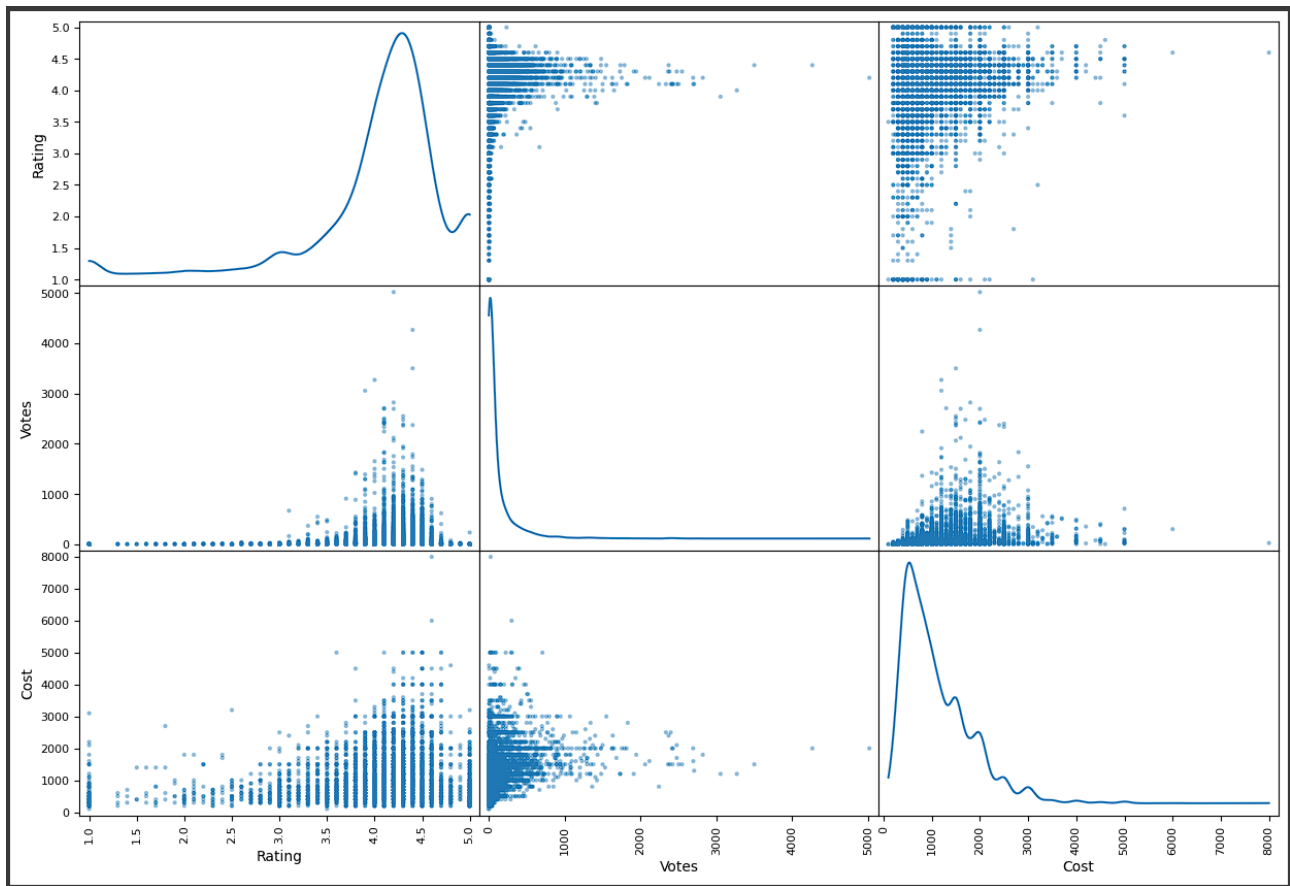
	Rating	Votes	Cost
count	6593.000000	6593.000000	6593.000000
mean	4.088200	119.420143	1102.798271
std	0.670031	261.849704	716.935212
min	1.000000	1.000000	100.000000
25%	3.900000	6.000000	500.000000
50%	4.200000	31.000000	900.000000
75%	4.400000	115.000000	1500.000000
max	5.000000	5016.000000	8000.000000

## IV. Data Visualization

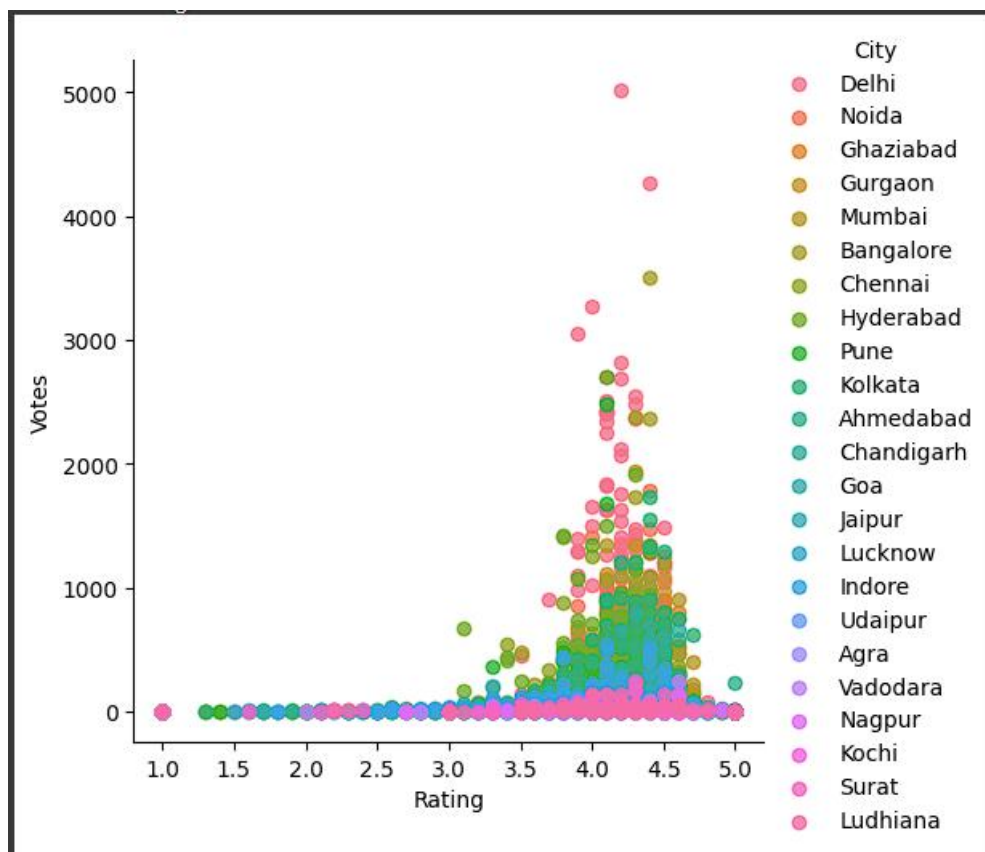
### 1. Biểu đồ thể hiện 3 thông số Rating, Votes và Cost



### 2. Biểu đồ thể hiện dữ liệu đa biến



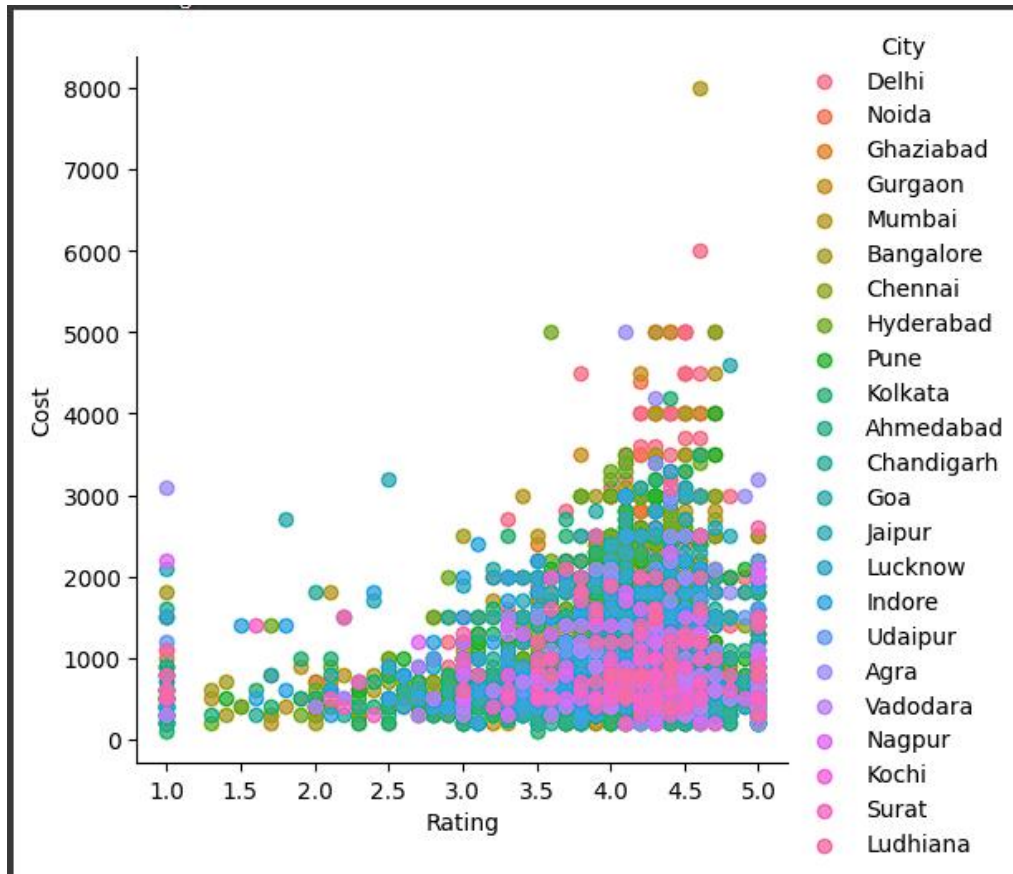
### 3. Biểu đồ thể hiện chỉ số Rating và lượt votes tất cả nhà hàng của mỗi thành phố



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

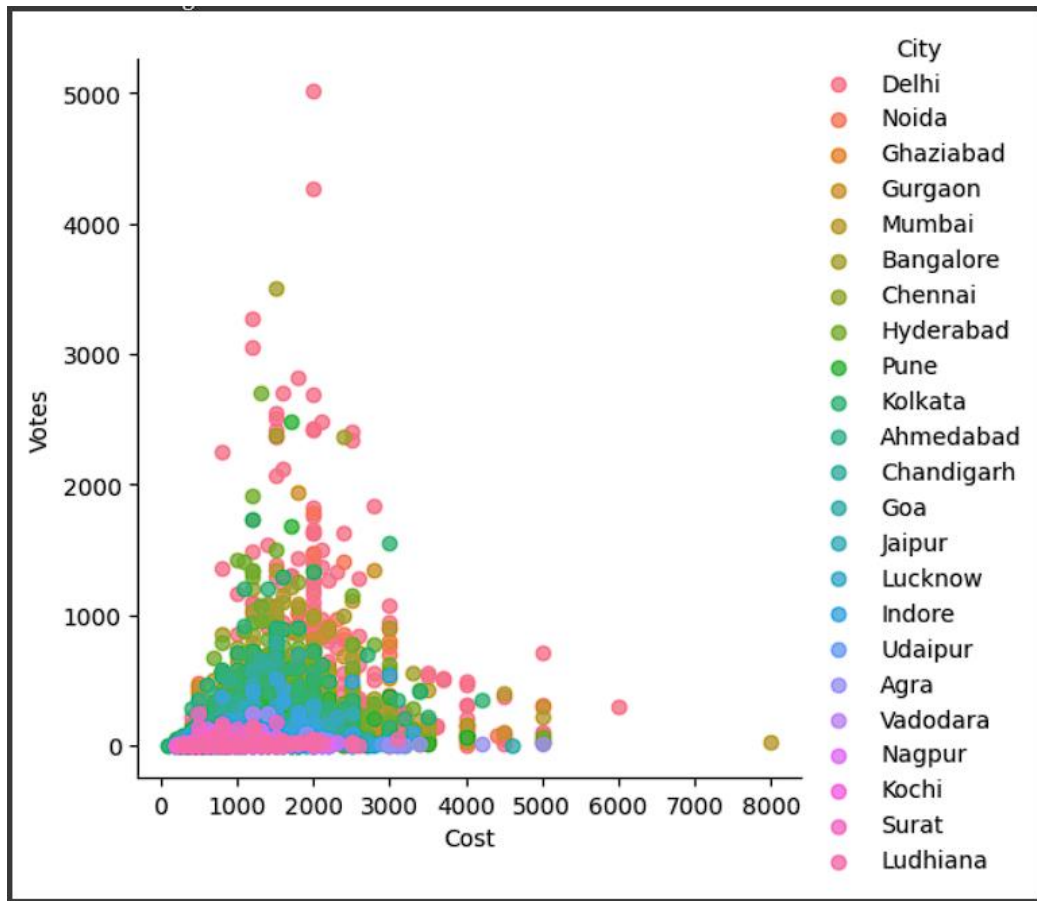
#### 4. Biểu đồ thể hiện chỉ số Rating và Cost tất cả nhà hàng của mỗi thành phố



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

#### 5. 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 nhiều nhất có giá nằm ở mức 1000 - 2000

## V. Tính toán các đại lượng thống kê cơ bản

### 1. Mean, Standard deviation, Max, Min



```

Mean:
Rating      4.088200
Votes       119.420143
Cost        1102.798271
dtype: float64
Standard deviation:
Rating      0.669980
Votes       261.829845
Cost        716.880839
dtype: float64
Max:
Rating      5.0
Votes       5016.0
Cost        8000.0
dtype: float64
Min:
Rating      1.0
Votes       1.0
Cost        100.0
dtype: float64

```

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

```

## Means:

```

	Rating	Votes	Cost
City			
Agra	4.238667	19.613333	1320.000000
Ahmedabad	4.202899	50.422705	777.294686
Bangalore	4.029931	100.210010	924.288518
Chandigarh	4.137500	53.083333	984.848485
Chennai	4.025258	80.074742	937.113402
...	...	...	...
Noida	4.193836	200.753425	1399.315068
Pune	4.078632	95.957265	1096.723647
Surat	3.964062	33.500000	709.375000
Udaipur	4.072093	19.953488	1120.930233
Vadodara	4.131868	16.000000	612.087912

```

## Standard deviations:

```

	Rating	Votes	Cost
City			
Agra	0.620300	32.618560	959.027285
Ahmedabad	0.581968	71.383245	396.593496
Bangalore	0.809752	237.602081	607.696272
Chandigarh	0.743912	109.561941	539.410431
Chennai	0.688852	132.274578	610.656160
...	...	...	...
Noida	0.445867	295.512547	728.245835
Pune	0.630854	191.129126	590.423896
Surat	0.745060	46.291535	334.345793
Udaipur	0.745631	24.966912	511.067346
Vadodara	0.833750	30.408934	278.076266

```
## Sample sizes:
      0
City
Agra      75
Ahmedabad 414
Bangalore 1019
Chandigarh 264
Chennai   388
...      ...
Noida     146
Pune      351
Surat     64
Udaipur   43
Vadodara  91
```

### 3. Phương sai trong group

```
## v_w:
0.4403532117485856
```

### 4. Phương sai giữa các group

```
## v_b:
1830116.0236454236
```

### 5. Sự khả tách cho từng biến

```
variable Rating Vw= 0.4403532117485856 Vb= 3.0136967332167104 separation= 6.84381685613172
variable Votes Vw= 62666.61973894469 Vb= 1830116.0236454236 separation= 29.204000970042475
variable Cost Vw= 456108.3642440714 Vb= 17801378.300760847 separation= 39.02883546164267
```

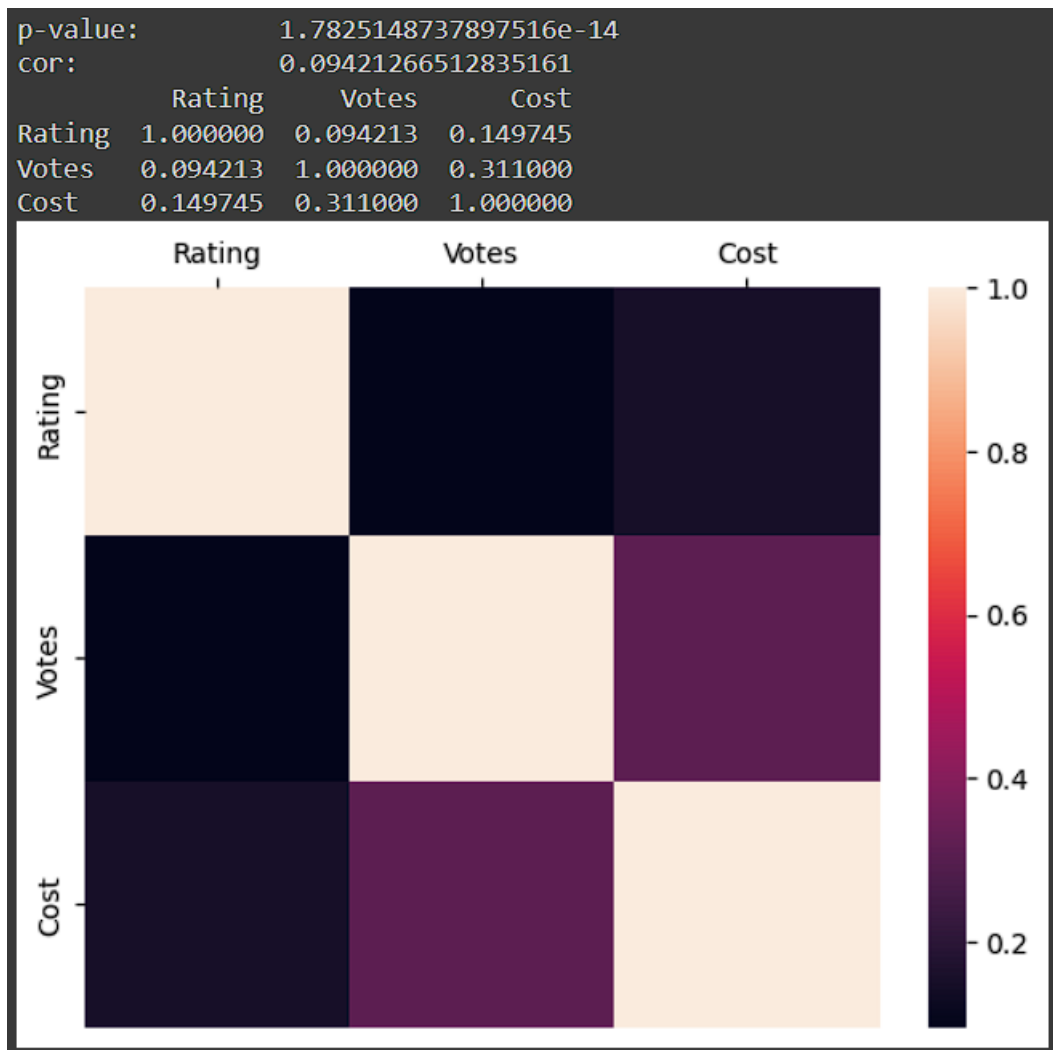
### 6. Hiệp phương sai trong group

```
## cov_w:
13.778933911232212
```

### 7. Hiệp phương sai giữa các group

```
## cov_b:
837.90869525341
```

### 8. Biểu đồ heatmap tương quan cho dữ liệu đa biến



## 9. Biểu đồ Hinton trực quan hóa ma trận trọng số



## 10. Sự tương quan giữa các biến

	FirstVariable	SecondVariable	Correlation
0	Votes	Cost	0.311000
1	Rating	Cost	0.149745
2	Rating	Votes	0.094213
3	Rating	Rating	0.000000
4	Votes	Rating	0.000000
5	Votes	Votes	0.000000
6	Cost	Rating	0.000000
7	Cost	Votes	0.000000
8	Cost	Cost	0.000000

## VI. PCA and visualization

### 1. Tiêu chuẩn hóa các biến

```

Rating    -3.793585e-16
Votes     -1.724357e-17
Cost      -1.207050e-16
dtype: float64
Rating     1.0
Votes      1.0
Cost       1.0
dtype: float64

```

### 2. Phân tích thành phần chính

Importance of components:

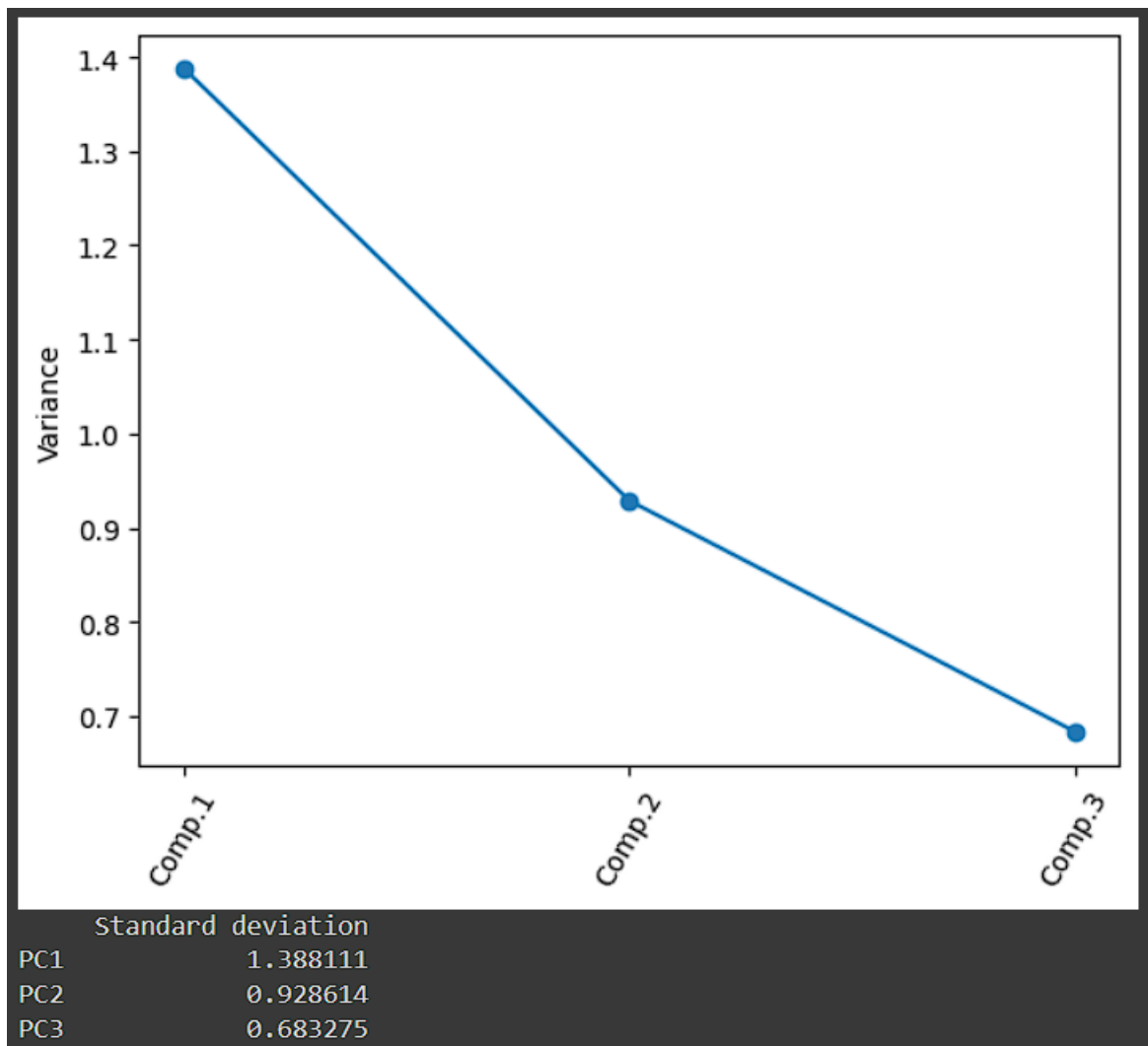
	sdev	varprop	cumprop
	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	1.178181	0.462704	0.462704
PC2	0.963646	0.309538	0.772242
PC3	0.826604	0.227758	1.000000

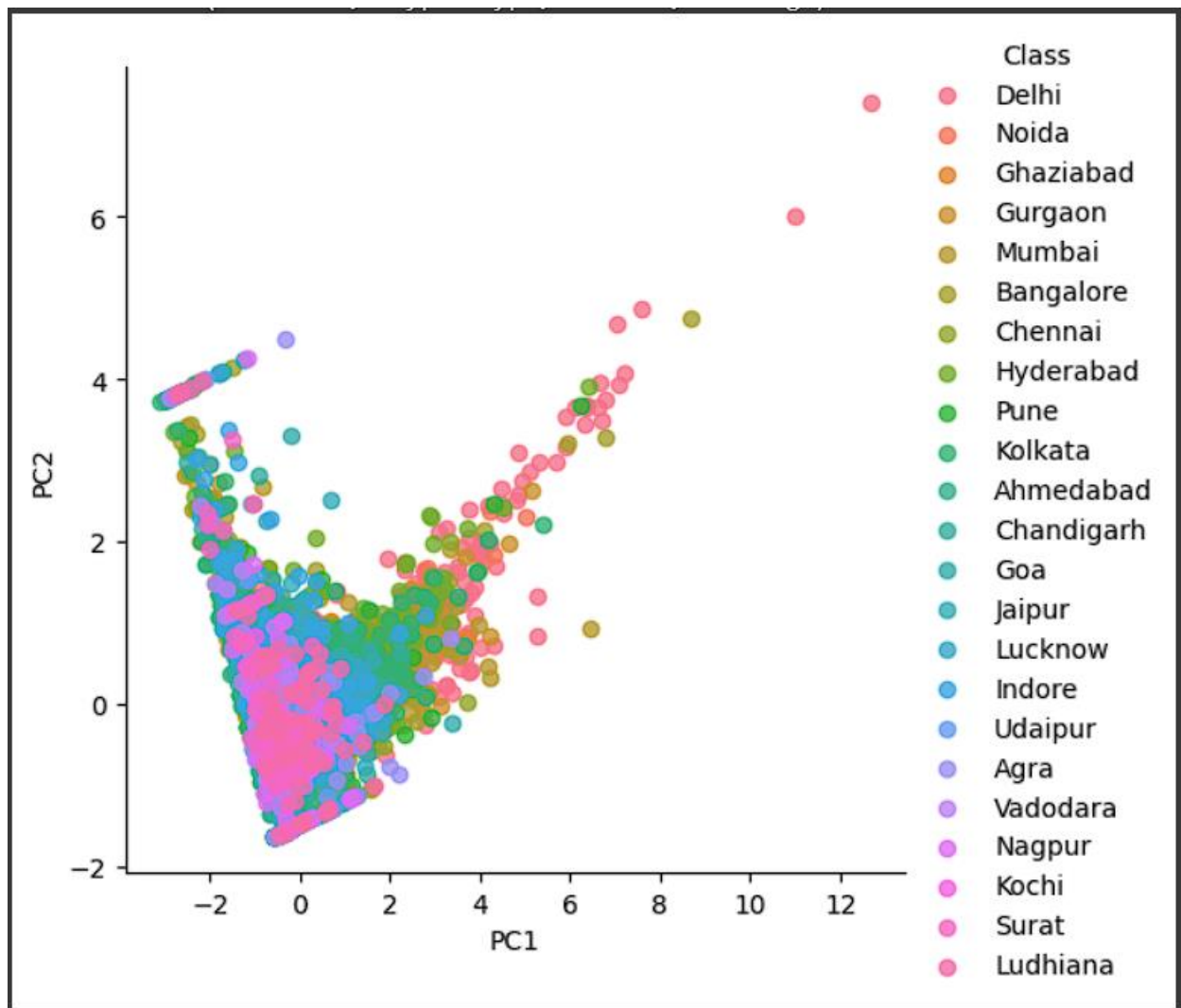
	Standard deviation
PC1	1.178181
PC2	0.963646
PC3	0.826604

Standard deviation 3.0  
dtype: float64

### 3. Biểu đồ độ lệch chuẩn của các thành phần chính



### 4. Biểu đồ scatter của 2 thành phần chính cho class “City”



## VII. LDA and visualization

### 1. Tiêu chuẩn hóa các biến

```
Rating    -3.793585e-16
Votes     -1.724357e-17
Cost      -1.207050e-16
dtype: float64
Rating     1.0
Votes      1.0
Cost       1.0
dtype: float64
```

### 2. Phân tích thành phần chính

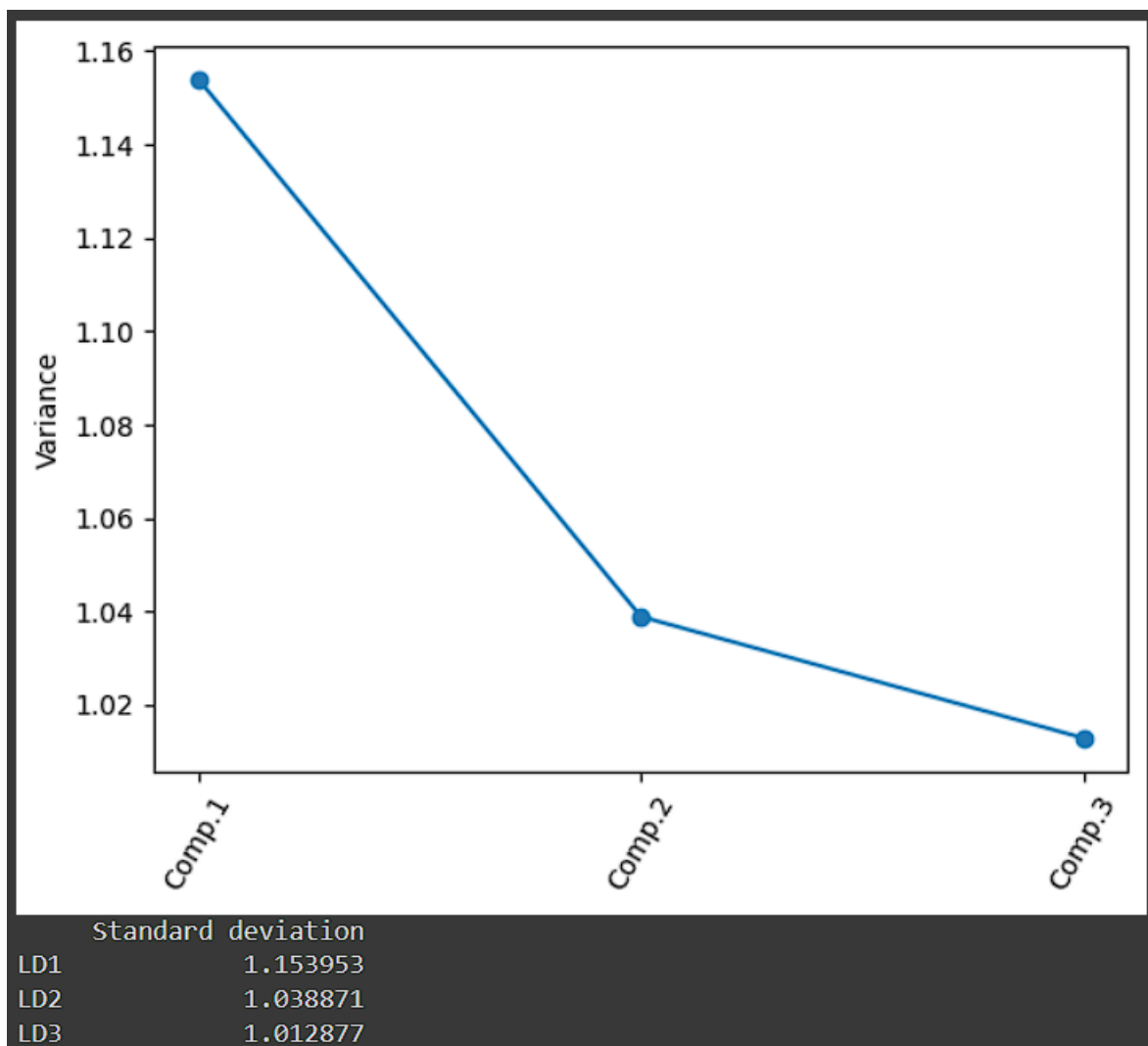
```

Importance of components:
              sdev              proportion
Standard deviation Proportion of Explained Variance
LD1              1.074222              0.728332
LD2              1.019250              0.195958
LD3              1.006418              0.075709

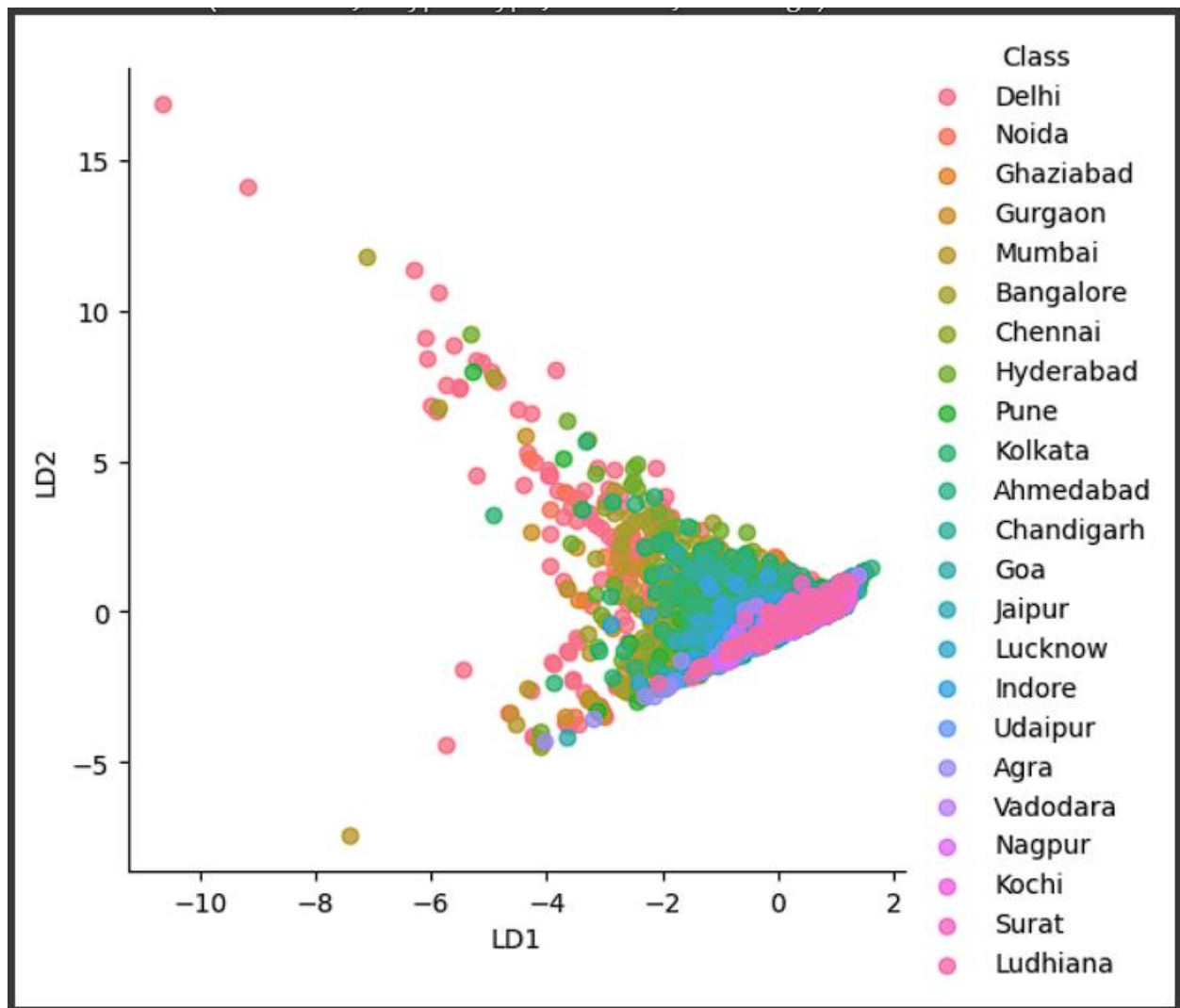
              cumprop
Cumulative Proportion of Explained Variance
LD1              0.728332
LD2              0.924291
LD3              1.000000
Standard deviation
LD1              1.074222
LD2              1.019250
LD3              1.006418
Standard deviation    3.205702
dtype: float64

```

### 3. Biểu đồ độ lệch chuẩn của các thành phần chính



### 4. Biểu đồ scatter của 2 thành phần chính của class “City”



## VIII. Reference

1. [Sample source](#)
2. [Dataset](#)