TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐHQG - TPHCM

Khoa công nghệ thông tin
Môn: Phân tích thống kê dữ liệu nhiều biến

PRACTICE #03

Scikit-learn with PCA & LDA

Giáo viên hướng dẫn: Nguyễn Mạnh Hùng
Lý Quốc Ngọc
Phạm Thanh Tùng
Lớp: 20TGMT01
Sinh viên thực hiện: Phùng Nghĩa Phúc – 20127284

TP.Hồ Chí Minh, ngày 25 tháng 4 năm 2023

MỤC LỤC

MỤC LỤC 1

GIỚI THIỆU 3

BẢNG ĐÁNH GIÁ 4

CHƯƠNG 1: Dataset Travel times 5

1. Giới thiệu 5

2. Mô tả dữ liệu 5

3. Tóm tắt dữ liệu 5

4. Mô tả tương quan dữ liệu 6

CHƯƠNG 2: Trực quan hoá dữ liệu 7

1. Biểu đồ dữ liệu đa biến 7

2. Biểu đồ hiển thị quảng đường đi được và tốc độ tối đa của người lái xe trong tuần 7

3. Biểu đồ hiển thị tốc độ tối đa và trung bình tốc độ di chuyển của người lái xe trong tuần 8

4. Biểu đồ đường thể hiện thông số của tài xế 9

4.1 Biểu đồ hiển thị toàn bộ thống số của chuyến đi trong tuần 9

4.2 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed 9

4.3 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed 10

4.4 Biểu đồ hiển thị TotalTime, MovingTime 10

5. Thống kê các dữ liệu đa biến tính được trên tập data. 10

5.1 Với tính trung bình các mẫu trên data. 10

5.2 Với độ lệch chuẩn 11

5.3 Với số liệu lớn nhất 11

5.4 Với số liệu nhỏ nhất 11

6. Trung bình mẫu, độ lệch chuẩn và số lượng mẫu của mỗi nhóm data 11

6.1 Trung bình mẫu 11

6.2. Độ lệch chuẩn 12

6.3 Số lượng mẫu trong tuần 12

7. Within_groups và between_groups với một biến cụ thể: Biến Distance 12

8. Sự khả tách của từng biến 12

9. Within_groups và between_groups với 2 biến cụ thể: Biến AvgSpeed và AvgMovingSpeed 12

10. Sự tương quan cho dữ liệu đa biến 13

10.1 Tính sự tương quan 13

10.2 Ma trận tương quan 13

10.3 Headmap ma trận tương quan 13

11. Biểu đồ Hinton để trực quan hoá ma trận trọng số.....13
12. hệ số tương quan tuyến tính cho từng cặp biến theo thứ tự của hệ số tương quan.....14

CHƯƠNG 3: PCA (principal component analysis).....15
1. Tổng quan về PCA.....15
2. Các bước tính của PCA.....15
3. Thực thi.....15
3.1 Tiêu chuẩn hoá các biến..... 15
3.2 độ lệch chuẩn của từng thành phần chính 15
3.3 Tổng phương sai 16
4. Biểu đồ với các thành phần chính của data.....16
5. Biểu đồ phân tán của các thành phần chính 16

CHƯƠNG 4: LDA (Linear Discriminant Analysis) 18
1. Tổng quan về LDA 18
- Khác với PCA, LDA tìm phép chiếu sao cho tối đa hóa sự khác biệt giữa các lớp để có thể phân lớp hiệu quả. 2. Các bước tính LDA 18
3. Thực thi với code 18
3.1 Làm đẹp data bằng tính hệ số phân biệt tuyến tính..... 18
3.2 Chuẩn hoá dữ liệu theo nhóm 19
3.3 Tính sự khả tách đạt được bởi từng hàm phân biệt..... 19
3.4 Tính khoảng cách dưới dạng tỷ lệ của phương sai giữa các nhóm với phương sai bên trong các nhóm:..... 20
3.5 Tính tỷ lệ dấu vết cho từng phân biệt tuyến tính..... 20
4. Biểu đồ xếp chồng của các giá trị LDA với hàm phân biệt đầu tiên : LD1...20
4. Biểu đồ xếp chồng của các giá trị LDA với hàm phân biệt thứ 2 : LD2 21
5. Biểu đồ phân tán của LDA 22

CHƯƠNG 5: Bonus : Trực quan hoá dữ liệu với ICA và FA 24
1. ICA (Independent Component Analysis).....24
1.1 Tương quan về ICA..... 24
1.2 Biểu đồ..... 24
2. FA (Factor Analysis).....24
2.1 Tương quan về FA..... 24
2.2 Biểu đồ..... 25

CHƯƠNG 6: So sánh PCA và LDA.....26
1. Biểu đồ của PCA và LDA.....26
2. So sánh.....26

GIỚI THIỆU

1. **Sinh viên**
   - Họ và tên: Phùng Nghĩa Phúc
   - MSV: 20127284
   - Lớp: 20TGMT01
   - Môn học: Phân tích thống kê dữ liệu nhiều biến
2. **Chủ đề**
   - Sử dụng Scikit-learn để visualization và áp dụng PCA và PDA để đưa ra đánh giá

BẢNG ĐÁNH GIÁ

STT	Loại	Đánh giá
1	**Mô tả dữ liệu**	100%
2	**Trực quan hoá dữ liệu với các phép tính**	100%
3	**PCA**	100%
4	**LDA**	100%
5	**Mở rộng PCA**	100%
6	**Mở rộng LDA**	100%
7	**Bonus: ICA và FA**	100%

CHƯƠNG 1: Dataset Travel times

- Link dataset: https://openmv.net/info/travel-times

1. Giới thiệu

- Một người lái xe sử dụng một ứng dụng để theo dõi tọa độ GPS khi anh ta lái xe đi làm và quay về mỗi ngày. Ứng dụng thu thập dữ liệu vị trí và độ cao. Dữ liệu cho khoảng 200 chuyến đi được tóm tắt trong bộ dữ liệu này.

2. Mô tả dữ liệu

- Dữ liệu có 205 hàng và 13 cột

- Trong dữ liệu chứa các thuộc tính:

  - Date: ngày

  - StartTime: Thời gian bắt đầu khi lên xe

  - DayOfWeek: khi lên xe

  - GoingTo: hướng di chuyển

  - Distance: Quảng đường đi được

  - MaxSpeed: tốc độ nhanh nhất được ghi nhận (tất cả các chuyến đi đều trên đường cao tốc 407 trong một số đoạn)

  - AvgSpeed: tốc độ trung bình được ghi lại chỉ trong khi chiếc xe đang di chuyển

  - AvgMovingSpeed: the average speed recorded only while the car is moving

  - FuelEconomy: một ước tính sơ bộ về tiết kiệm nhiên liệu (nó không chính xác)

  - TotalTime: thời lượng của toàn bộ chuyến đi, tính bằng phút

  - MovingTime: khoảng thời gian khi ô tô được coi là đang di chuyển (nghĩa là không tính đến tắc nghẽn giao thông, tai nạn hoặc thời gian ô tô đứng yên)

  - Take407All: là Yes nếu đường cao tốc thu phí 407 đã được sử dụng cho toàn bộ chuyến đi. Tôi cố gắng tránh đi 407, chọn các tuyến đường quay lại chậm hơn để tiết kiệm chi phí. Nhưng một số ngày tôi đến muộn, hoặc chỉ lười biếng và làm mọi cách.

  - Comments

3. Tóm tắt dữ liệu

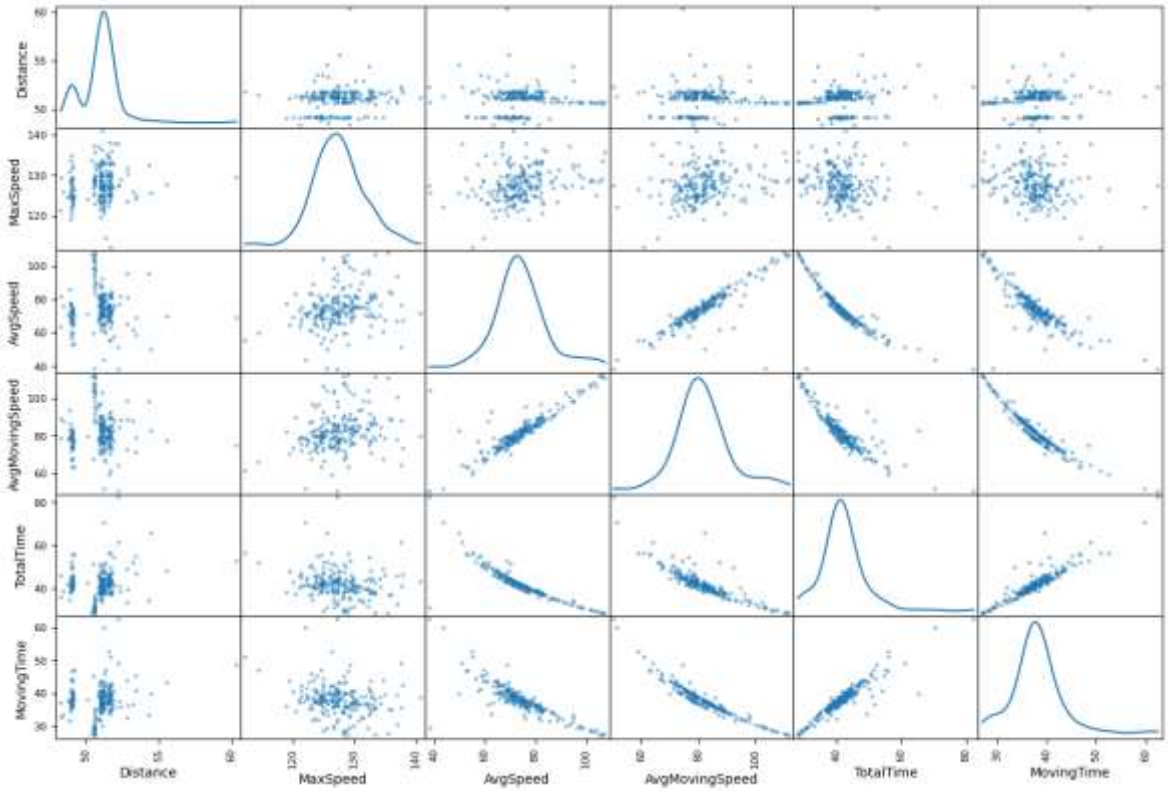
```
## Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            205 non-null    object
 1   StartTime       205 non-null    object
 2   DayOfWeek       205 non-null    object
 3   GoingTo         205 non-null    object
 4   Distance        205 non-null    float64
 5   MaxSpeed        205 non-null    float64
 6   AvgSpeed        205 non-null    float64
 7   AvgMovingSpeed  205 non-null    float64
 8   TotalTime       205 non-null    float64
 9   MovingTime      205 non-null    float64
 10  Take407All      205 non-null    object
 11  Comments        24 non-null     object
dtypes: float64(6), object(6)
memory usage: 19.3+ KB
```

4. Mô tả tương quan dữ liệu

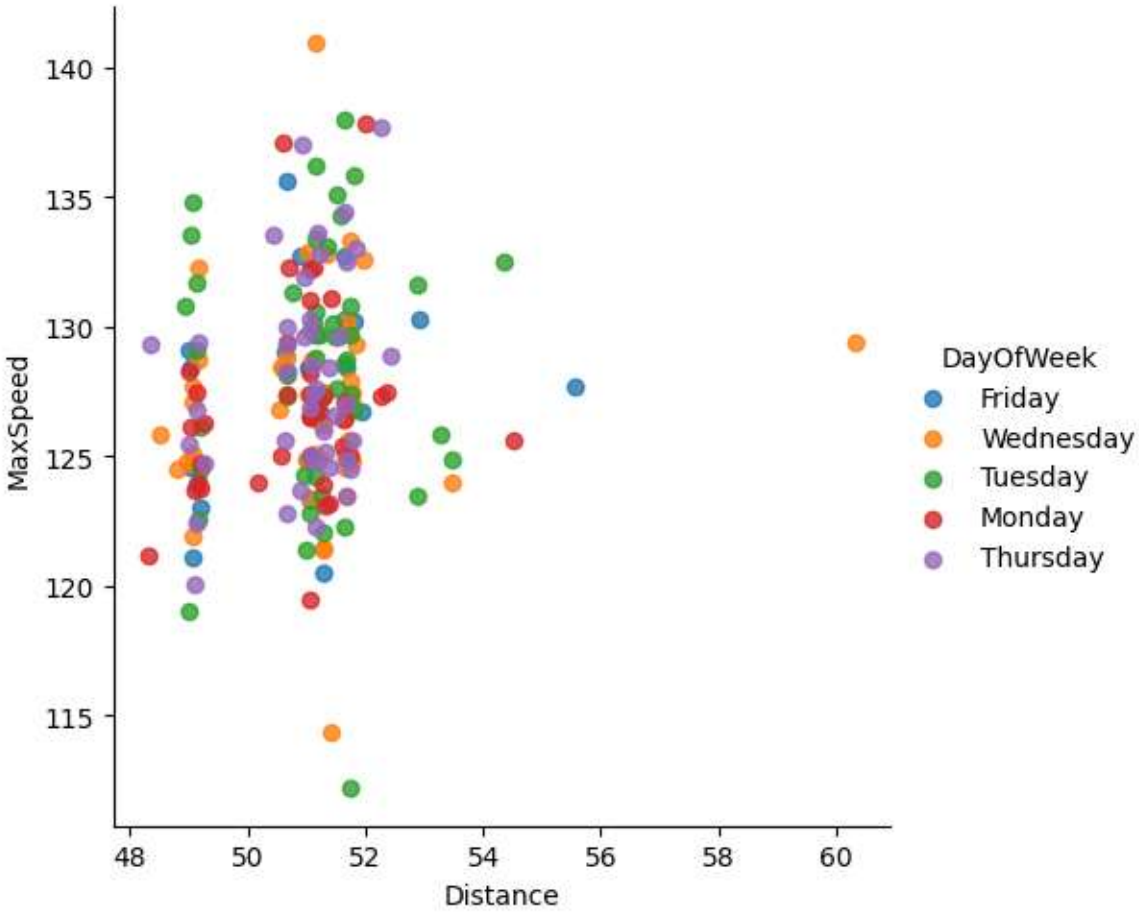
	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	TotalTime	MovingTime
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	50.981512	127.591707	74.477561	81.975610	41.904390	37.871707
std	1.321205	4.128450	11.409816	10.111544	6.849476	4.835072
min	48.320000	112.200000	38.100000	50.300000	28.200000	27.100000
25%	50.650000	124.900000	68.900000	76.600000	38.400000	35.700000
50%	51.140000	127.400000	73.600000	81.400000	41.300000	37.600000
75%	51.630000	129.800000	79.900000	86.000000	44.400000	39.900000
max	60.320000	140.900000	107.700000	112.100000	82.300000	62.400000

CHƯƠNG 2: Trực quan hoá dữ liệu

1. Biểu đồ dữ liệu đa biến

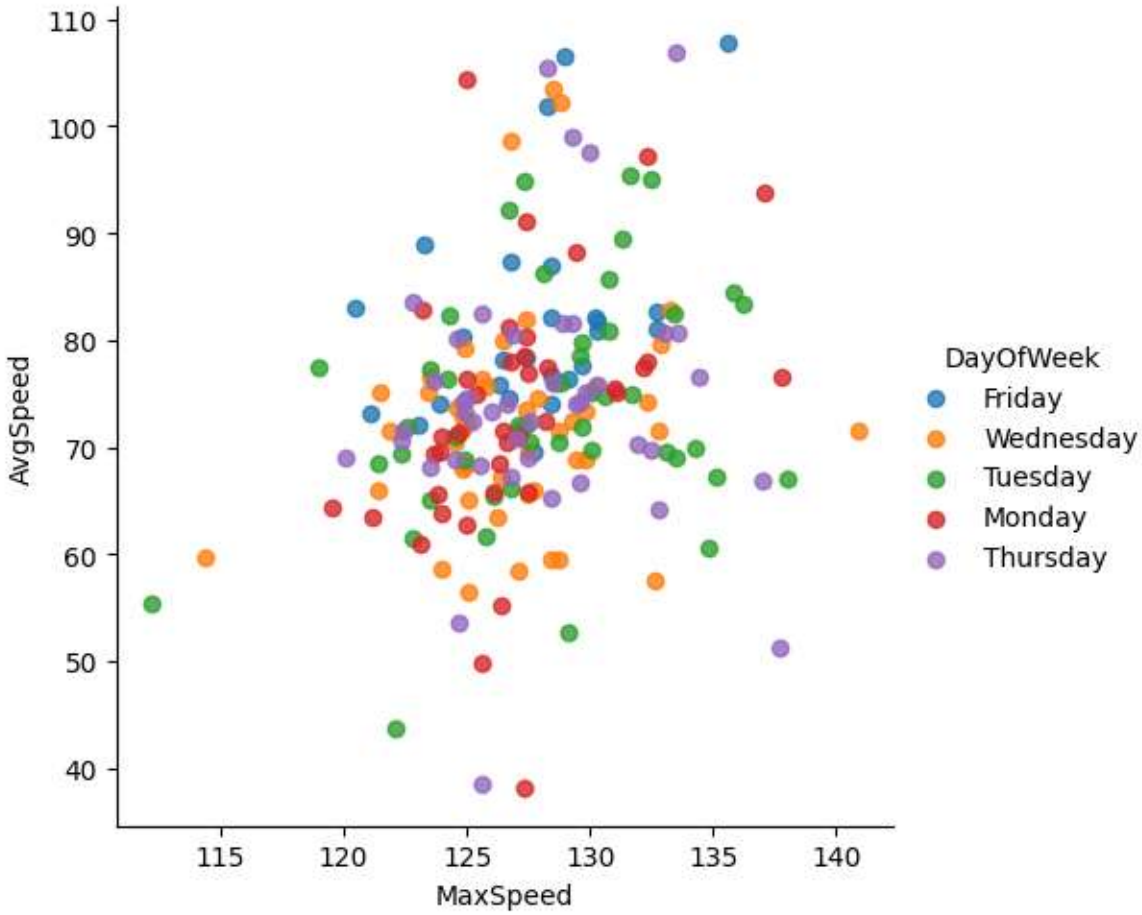


2. Biểu đồ hiển thị quảng đường đi được và tốc độ tối đa của người lái xe trong tuần



- Sơ lược vào biểu đồ có thể thấy được tốc độ lớn nhất mà tái xe đi được nằm trong khoảng 125 -> 130. Tốc độ này được duy trì trên tổng quảng đường di chuyển được là 50 đến 52

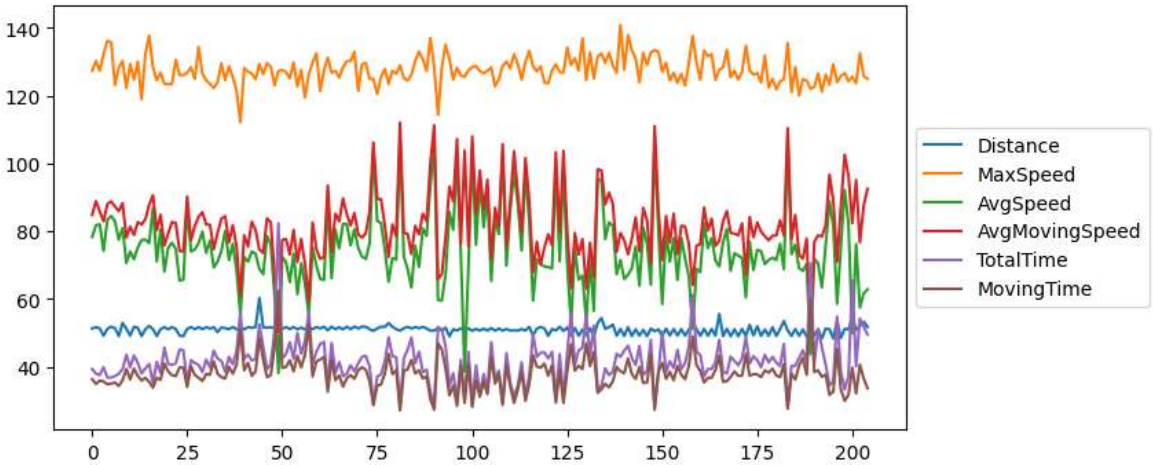
3. Biểu đồ hiển thị tốc độ tối đa và trung bình tốc độ di chuyển của người lái xe trong tuần



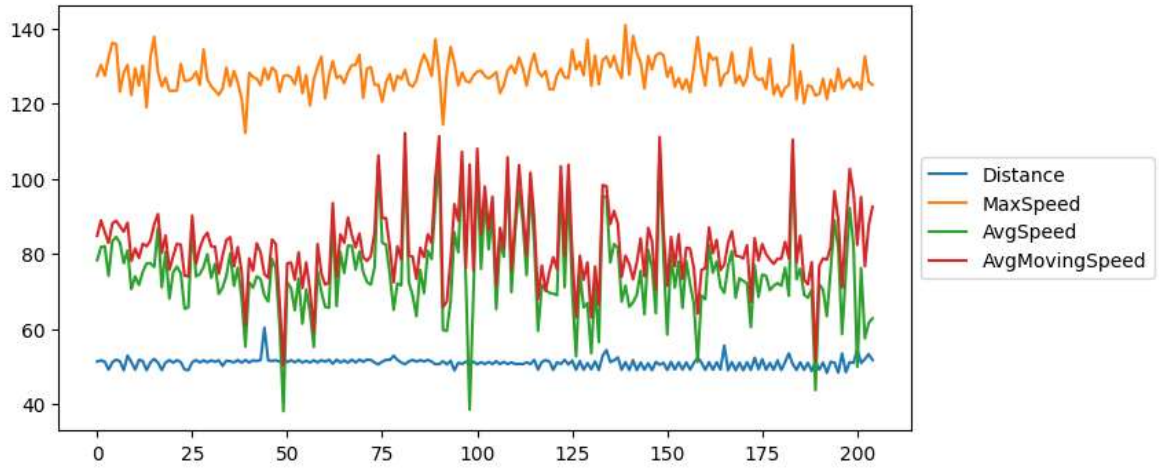
- Dựa vào biểu đồ, thấy được tốc độ trung bình và tốc độ lớn nhất mà tài xế này di chuyển khá tương đồng nhau

4. Biểu đồ đường thể hiện thông số của tài xế

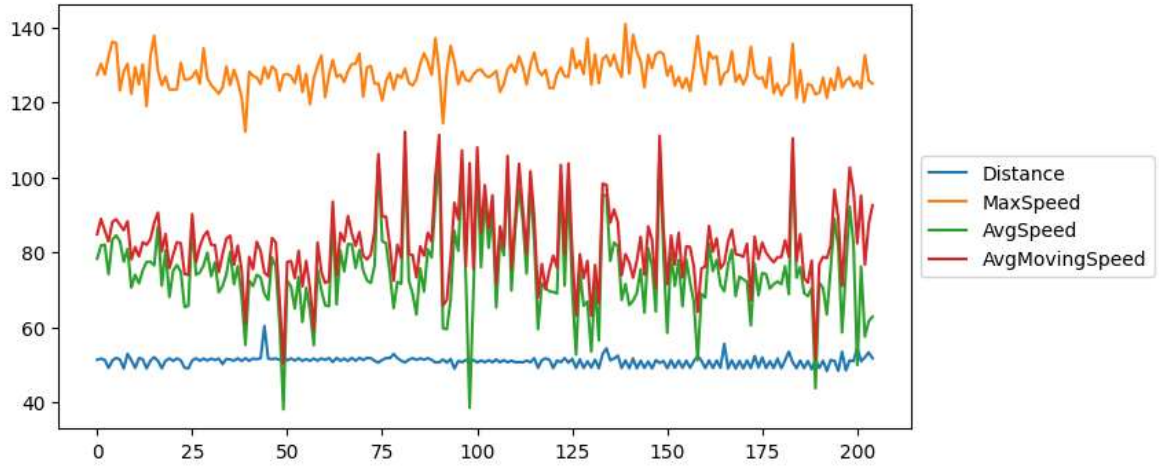
4.1 Biểu đồ hiển thị toàn bộ thống số của chuyển đi trong tuần



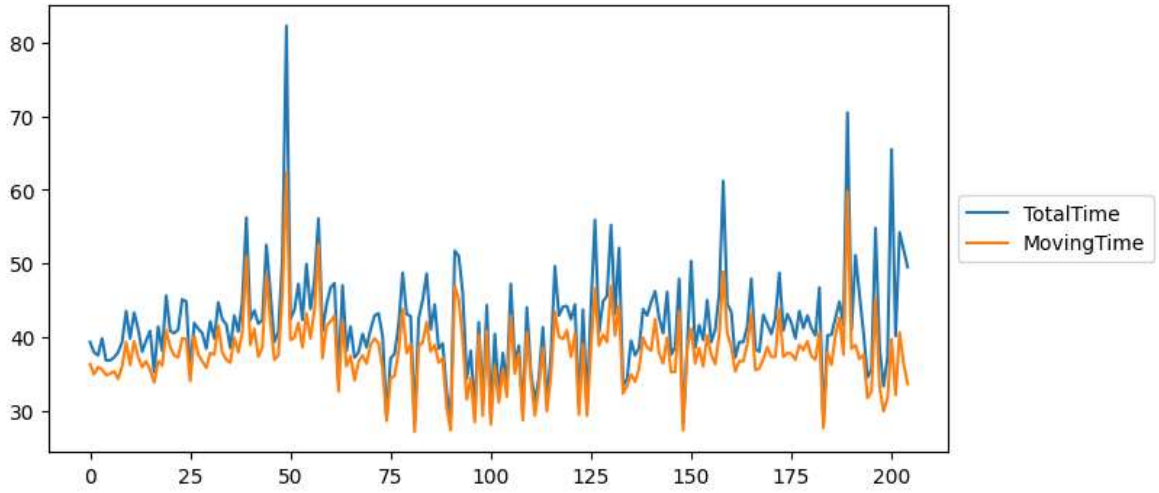
4.2 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed



4.3 Biểu đồ hiển thị Distance, MaxSpeed, AvgSpeed, AvgMovingSpeed



4.4 Biểu đồ hiển thị TotalTime, MovingTime



5. Thống kê các dữ liệu đa biến tính được trên tập data.

5.1 Với tính trung bình các mẫu trên data.

Distance 50.981512

MaxSpeed 127.591707

AvgSpeed 74.477561

AvgMovingSpeed 81.975610

TotalTime 41.904390

MovingTime 37.871707

5.2 Với độ lệch chuẩn

Distance 1.317979

MaxSpeed 4.118368

AvgSpeed 11.381953

AvgMovingSpeed 10.086852

TotalTime 6.832750

MovingTime 4.823265

5.3 Với số liệu lớn nhất

Distance 60.32

MaxSpeed 140.90

AvgSpeed 107.70

AvgMovingSpeed 112.10

TotalTime 82.30

MovingTime 62.40

5.4 Với số liệu nhỏ nhất

Distance 48.32

MaxSpeed 112.20

AvgSpeed 38.10

AvgMovingSpeed 50.30

TotalTime 28.20

6. Trung bình mẫu, độ lệch chuẩn và số lượng mẫu của mỗi nhóm data

6.1 Trung bình mẫu

```
## Means:
           Distance    MaxSpeed   AvgSpeed  AvgMovingSpeed  TotalTime
DayOfWeek
Friday     50.958889  127.559259  81.659259       87.937037  37.922222
Monday     50.795897  127.017949  73.197436       81.405128  43.197436
Thursday   50.902727  127.986364  74.365909       82.809091  41.177273
Tuesday    51.127500  128.235417  73.781250       80.893750  42.520833
Wednesday  51.073191  127.059574  72.229787       79.348936  43.170213

                MovingTime
DayOfWeek
Friday       35.114815
Monday       38.146154
Thursday     37.418182
Tuesday      38.427083
Wednesday    39.085106
```

6.2. Độ lệch chuẩn

```
## Standard deviations:
           Distance  MaxSpeed   AvgSpeed  AvgMovingSpeed  TotalTime
DayOfWeek
Friday     1.363691  3.422286   9.705564        9.115154   4.097093
Monday     1.179099  3.710199  12.137365       10.463513   9.121445
Thursday   0.923879  3.921595  12.094457       10.413689   5.955703
Tuesday    1.198276  4.893682  10.619570        9.817446   6.804195
Wednesday  1.739046  3.980840  10.001849        8.707074   5.610663

                MovingTime
DayOfWeek
Friday        3.403452
Monday        5.765051
Thursday      4.408065
Tuesday       5.061610
Wednesday     4.058664
```

6.3 Số lượng mẫu trong tuần

```
## Sample sizes:
              0
DayOfWeek
Friday       27
Monday       39
Thursday     44
Tuesday      48
Wednesday    47
```

7. Within_groups và between_groups với một biến cụ thể: Biến Distance

- Within_groups:

```
## v_w:
1.7652520213013938
```

- between_groups:

```
## v_b:
0.7621567398083797
```

8. Sự khả tách của từng biến

```
variable Distance Vw= 1.7652520213013938 Vb= 0.7621567398083797 separation= 0.43175520017050945
variable MaxSpeed Vw= 17.120387112102343 Vb= 13.22962000463854 separation= 0.7727407048691421
variable AvgSpeed Vw= 124.19872607641817 Vb= 429.4428913010443 separation= 3.4577076985219044
variable AvgMovingSpeed Vw= 97.37191936383405 Vb= 345.81354400341866 separation= 3.5514709606500885
variable TotalTime Vw= 44.80275573761345 Vb= 152.54372531444946 separation= 3.404784433525007
variable MovingTime Vw= 22.33945408915602 Vb= 75.3012711519549 separation= 3.3707749012769073
```

9. Within_groups và between_groups với 2 biến cụ thể: Biến AvgSpeed và AvgMovingSpeed

- Within_groups:

```
## cov_w:
95.16225377143297
```

- between_groups:
```
## cov_b:
373.4992626478646
```

10. Sự tương quan cho dữ liệu đa biến

10.1 Tính sự tương quan

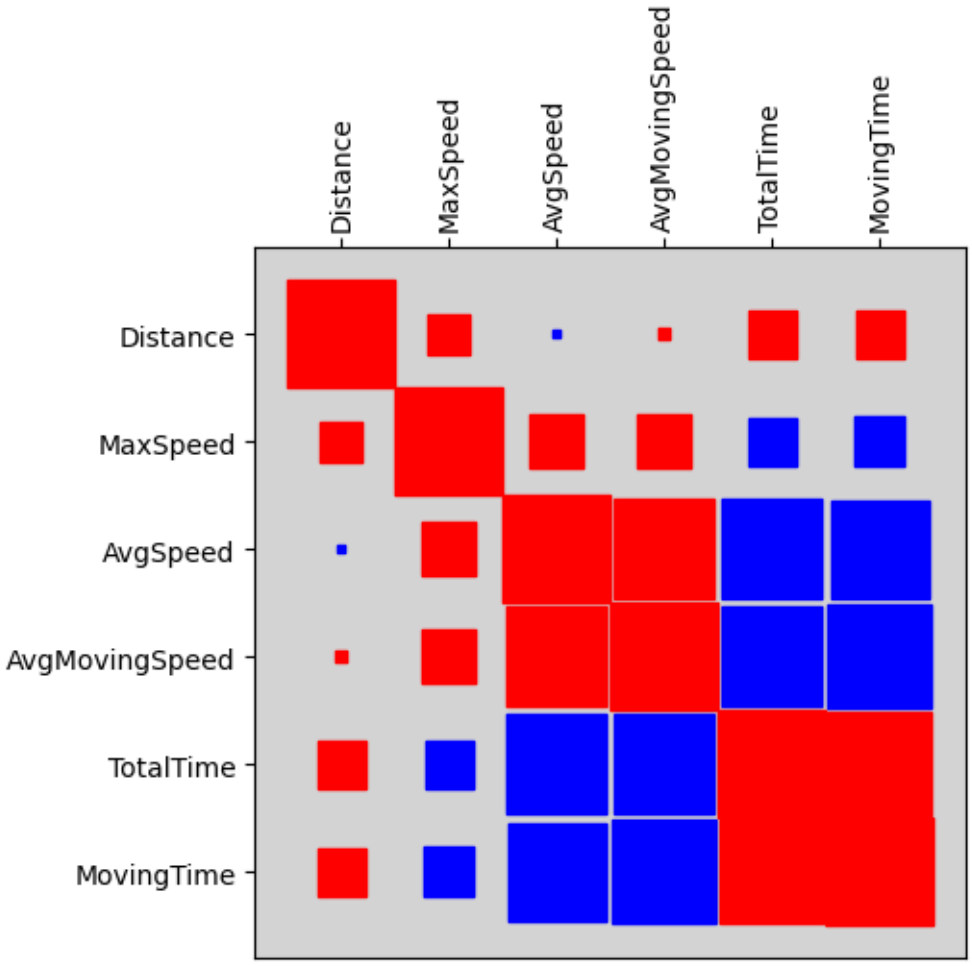
- Giá trị p-value: 0.00023974233393727256
- Giá trị cor: 0.2538685434138951

10.2 Ma trận tương quan

	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	TotalTime
Distance	1.000000	0.145091	-0.006445	0.011874	0.197207
MaxSpeed	0.145091	1.000000	0.253869	0.257823	-0.198775
AvgSpeed	-0.006445	0.253869	1.000000	0.872143	-0.877806
AvgMovingSpeed	0.011874	0.257823	0.872143	1.000000	-0.856986
TotalTime	0.197207	-0.198775	-0.877806	-0.856986	1.000000
MovingTime	0.197044	-0.222574	-0.835814	-0.944433	0.920935

10.3 Headmap ma trận tương quan

11. Biểu đồ Hinton để trực quan hoá ma trận trọng số.



12. hệ số tương quan tuyến tính cho từng cặp biến theo thứ tự của hệ số tương quan.

	FirstVariable	SecondVariable	Correlation
0	AvgMovingSpeed	MovingTime	-0.944433
1	TotalTime	MovingTime	0.920935
2	AvgSpeed	TotalTime	-0.877806
3	AvgSpeed	AvgMovingSpeed	0.872143
4	AvgMovingSpeed	TotalTime	-0.856986
5	AvgSpeed	MovingTime	-0.835814
6	MaxSpeed	AvgMovingSpeed	0.257823
7	MaxSpeed	AvgSpeed	0.253869
8	MaxSpeed	MovingTime	-0.222574
9	MaxSpeed	TotalTime	-0.198775

CHƯƠNG 3: PCA (principal component analysis)

1. Tổng quan về PCA

- PCA là kĩ thuật giảm chiều dữ liệu từ n chiều sang dữ liệu m chiều (m<n) mà vẫn giữ được nhiều thông tin nhất có thể.

2. Các bước tính của PCA

Bước 1: Tính giá trị trung bình: $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n.$

Bước 2: Chuẩn hóa: $\hat{x} = x_n - \bar{x}.$

Bước 3: Tính ma trận hiệp phương sai: $S = \frac{1}{N}\hat{X}\hat{X}^T.$

Bước 4: Tính các trị riêng λ_i và vector riêng v_i: $Sv_i = \lambda_i v_i.$

Bước 5: Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trận U_K có các cột tạo thành một hệ trực giao. K vector này, còn được gọi là các thành phần chính, tạo thành một không gian con gần với phân bố của dữ liệu ban đầu đã chuẩn hoá.

Bước 6: Chiếu dữ liệu ban đầu đã chuẩn hoá X xuống không gian con tìm được. Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới: $Z = U_K^T\hat{X}$

3. Thực thi

3.1 Tiêu chuẩn hoá các biến

```
Distance          5.545699e-15
MaxSpeed         -3.795338e-15
AvgSpeed         -5.502374e-16
AvgMovingSpeed    6.498866e-16
TotalTime        -2.664535e-16
MovingTime        8.231898e-16
dtype: float64
Distance          1.0
MaxSpeed          1.0
AvgSpeed          1.0
AvgMovingSpeed    1.0
TotalTime         1.0
MovingTime        1.0
dtype: float64
```

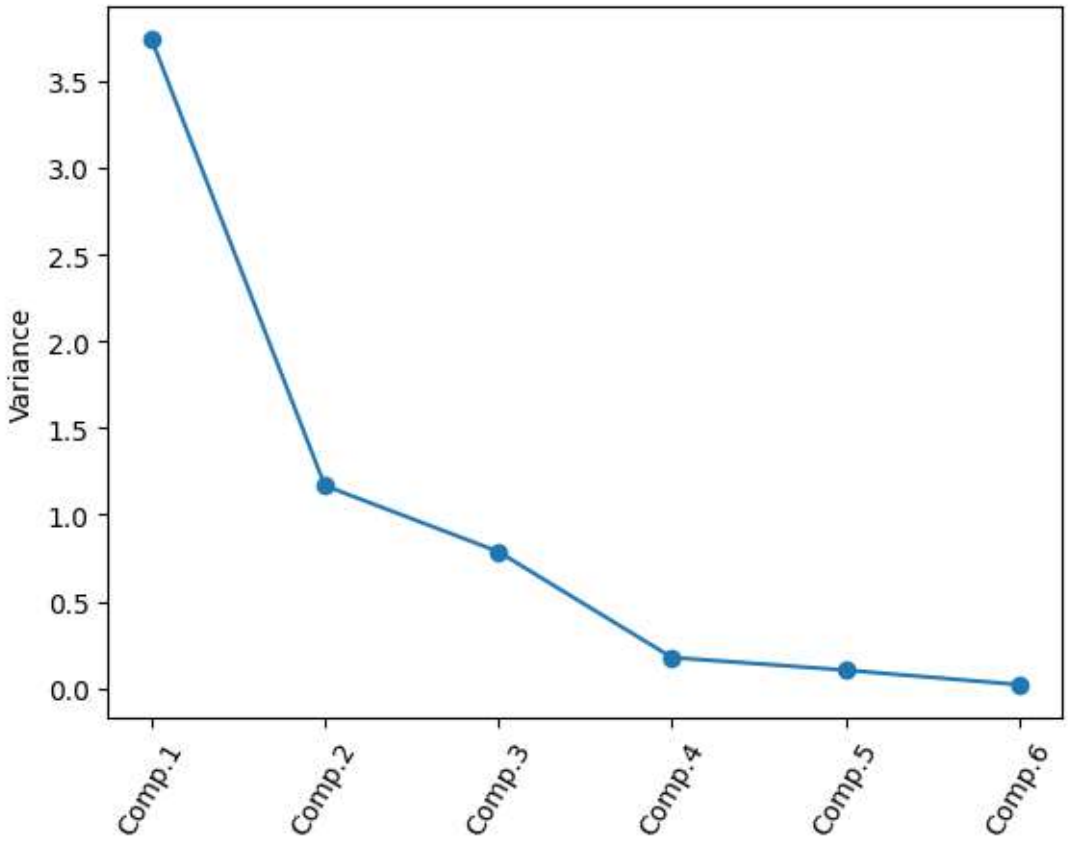
3.2 độ lệch chuẩn của từng thành phần chính

```
Importance of components:
                    sdev                 varprop                 cumprop
    Standard deviation Proportion of Variance Cumulative Proportion
PC1           1.829638               0.866207               0.866207
PC2           0.930571               0.050363               0.916571
     Standard deviation
PC1            1.829638
PC2            0.930571
```

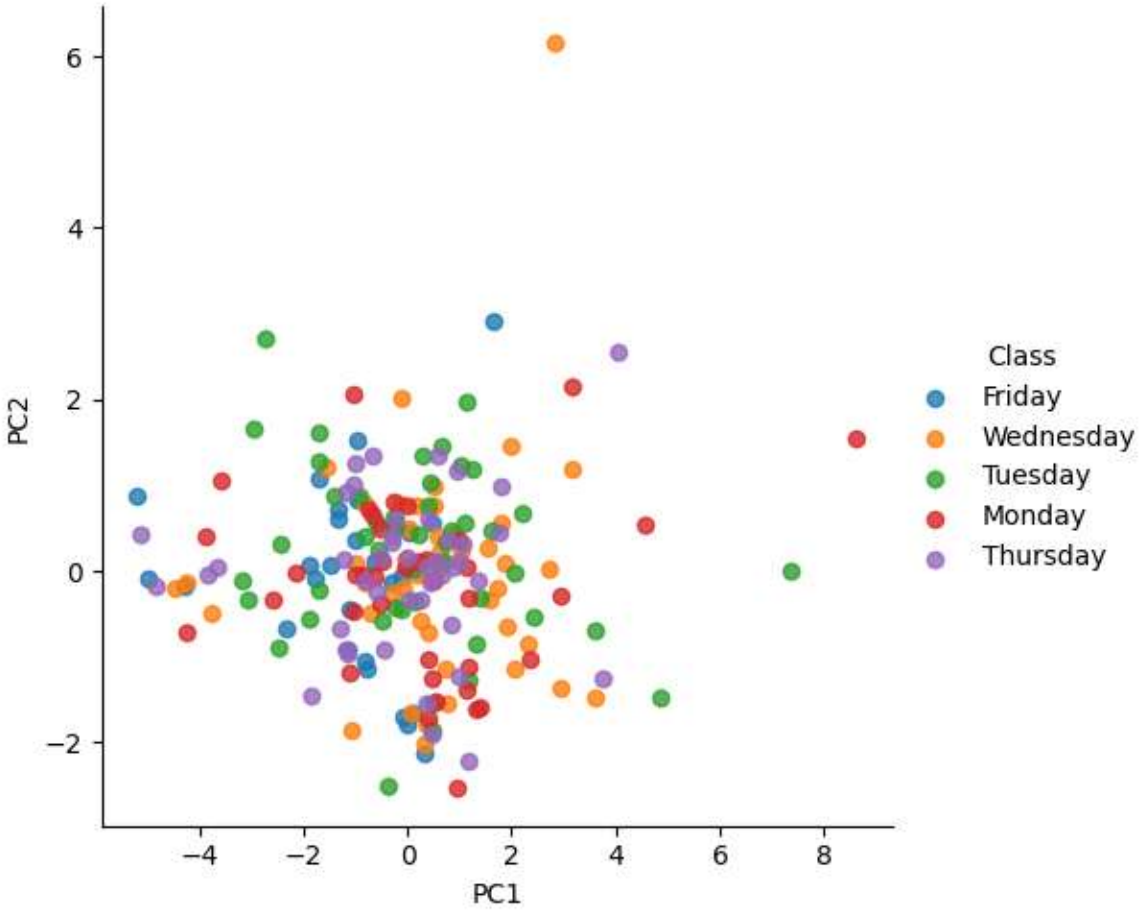
3.3 Tổng phương sai

```
Standard deviation    4.213537
dtype: float64
```

4. Biểu đồ với các thành phần chính của data



5. Biểu đồ phân tán của các thành phần chính


6
4
2
0
−2
PC2
−4
−2
0
2
4
6
8
PC1
Class
Friday
Wednesday
Tuesday
Monday
Thursday

CHƯƠNG 4: LDA (Linear Discriminant Analysis)

1. Tổng quan về LDA

- LDA là một thuật toán học có giám sát, giảm chiều dữ liệu.

- Mục đích LDA là tìm sự khác nhau giữa các thành phần trong 1 class (within-class) là nhỏ và sự khác nhau giữa các classes là lớn.

- Khác với PCA, LDA tìm phép chiếu sao cho tối đa hóa sự khác biệt giữa các lớp để có thể phân lớp hiệu quả.

2. Các bước tính LDA

Bước 1: Tính ma trận phân tán giữa các nhóm :

$$S_B = \sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T$$

μ_i là giá trị trung bình của từng lớp.

μ là giá trị trung bình của tất cả dữ liệu.

Bước 2: Tính ma trận phân tán tích lũy ứng với từng nhóm

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$$

Bước 3 : Xây dựng hàm tiêu chí tách lớp

$$W = S_W^{-1} S_B$$

Bước 4 : Dự đoán nhãn của mẫu dữ liệu nhập (so sánh vector trung bình của từng nhóm – gần vector trung bình nhất).

3. Thực thi với code

3.1 Làm đẹp data bằng tính hệ số phân biệt tuyến tính

	LD1	LD2	LD3	LD4
Distance	-0.059257	0.338918	-0.409285	0.092612
MaxSpeed	-0.054925	0.089484	0.090202	0.203417
AvgSpeed	0.038729	-0.094171	-0.177913	0.081632
AvgMovingSpeed	0.062466	0.013750	0.223432	-0.083727
TotalTime	-0.004036	-0.411381	-0.102382	0.207191
MovingTime	0.001315	0.363934	0.214856	-0.266318

3.2 Chuẩn hoá dữ liệu theo nhóm

	LD1	LD2	LD3	LD4
Distance	-0.078730	0.450296	-0.543787	0.123047
MaxSpeed	-0.227261	0.370255	0.373227	0.841676
AvgSpeed	0.431616	-1.049478	-1.982740	0.909744
AvgMovingSpeed	0.616395	0.135678	2.204766	-0.826191
TotalTime	-0.027015	-2.753574	-0.685292	1.386831
MovingTime	0.006214	1.720119	1.015511	-1.258740

3.3 Tính sự khả tách đạt được bởi từng hàm phân biệt

	LD1	LD2
0	0.325162	0.265671
1	0.541205	0.433609
2	0.538850	0.519124
3	-0.098758	-0.114835
4	0.261894	1.053335
...	...	...
200	-1.118476	-5.700063
201	1.104809	-1.715843
202	-1.366173	-1.759229
203	-0.227345	-2.702526
204	0.266874	-3.413159

205 rows × 2 columns

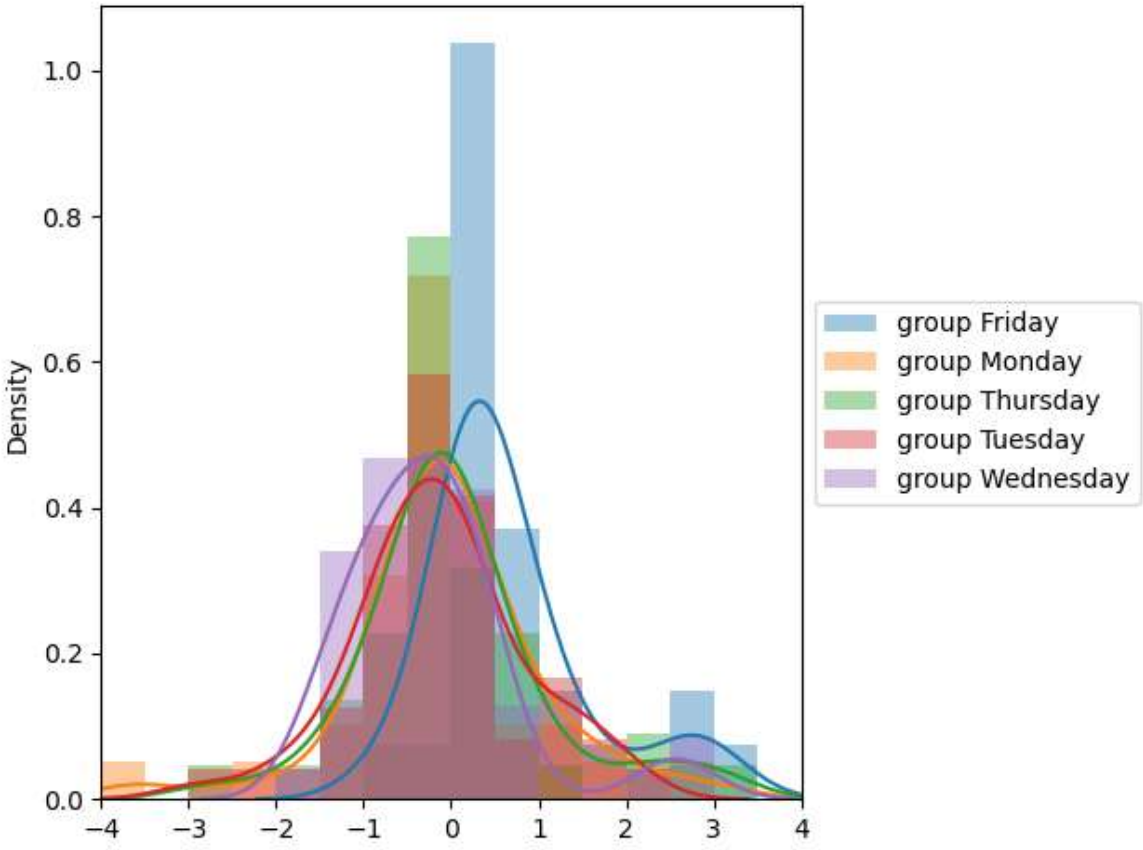
3.4 Tính khoảng cách dưới dạng tỷ lệ của phương sai giữa các nhóm với phương sai bên trong các nhóm:

```
variable LD1 Vw= 1.000000000000001 Vb= 3.8913881314556273 separation= 3.891388131455623
variable LD2 Vw= 1.0000000000000002 Vb= 2.347929717180682 separation= 2.3479297171806817
```

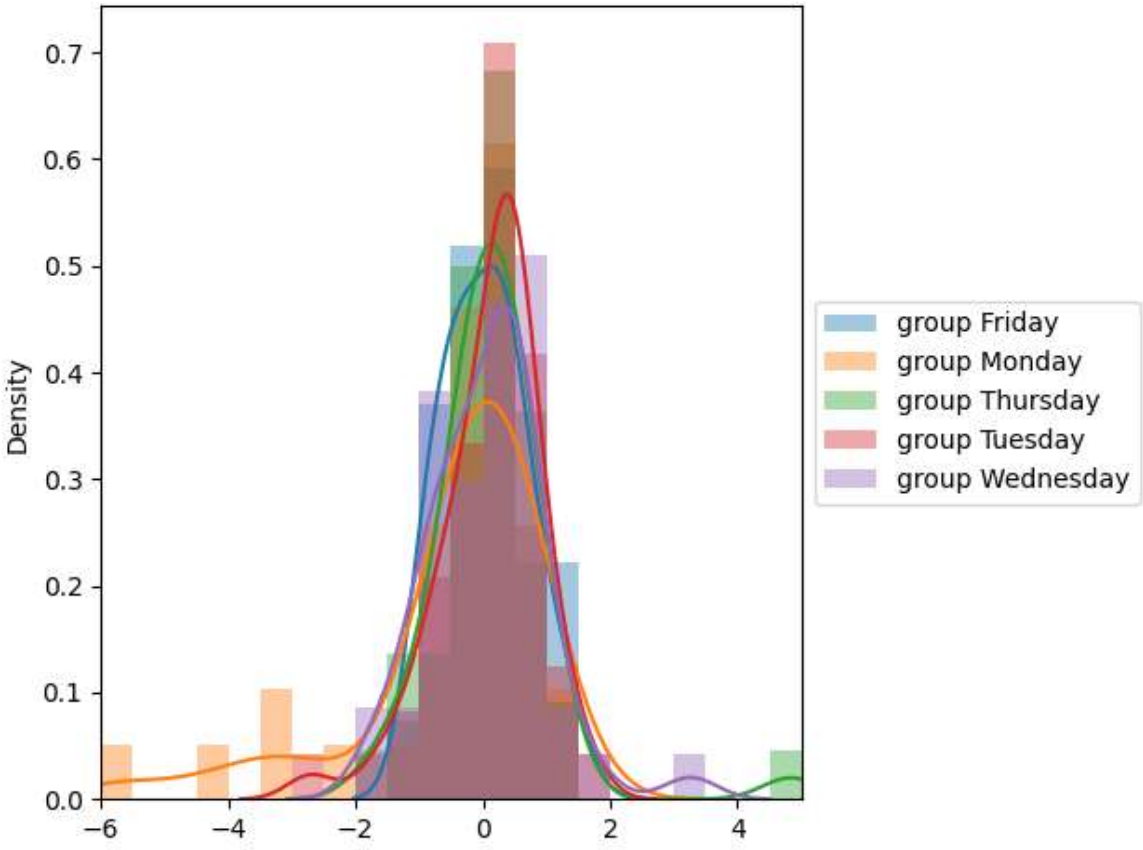
3.5 Tính tỷ lệ dấu vết cho từng phân biệt tuyến tính

```
Proportion of trace:
   LD1    LD2    LD3    LD4
0.4963 0.2995 0.1409 0.0633
```

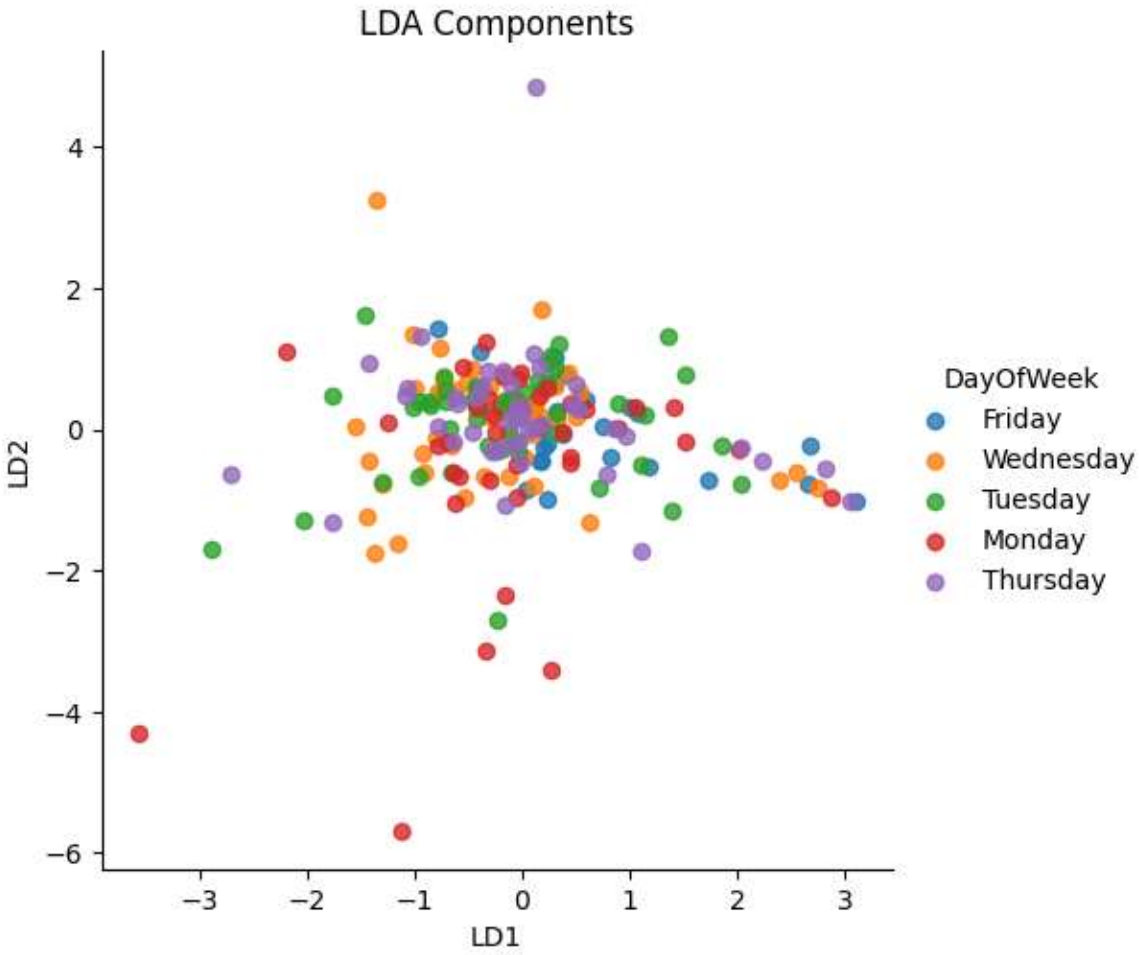
4. Biểu đồ xếp chồng của các giá trị LDA với hàm phân biệt đầu tiên : LD1



4. Biểu đồ xếp chồng của các giá trị LDA với hàm phân biệt thứ 2 : LD2



5. Biểu đồ phân tán của LDA


LDA Components
LD2
LD1
4
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3
DayOfWeek
Friday
Wednesday
Tuesday
Monday
Thursday

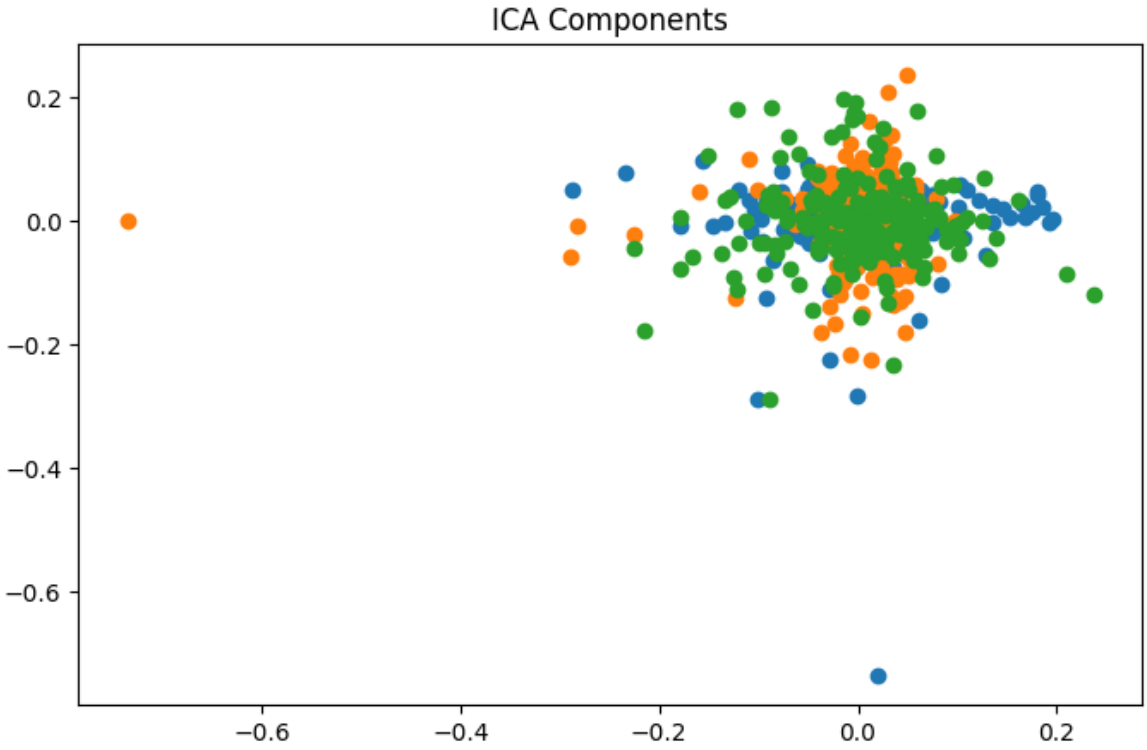
CHƯƠNG 5: Bonus : Trực quan hoá dữ liệu với ICA và FA

1. ICA (Independent Component Analysis)

1.1 Tương quan về ICA

- Independent Component Analysis (phân tích thành phần độc lập) là một phương pháp thống kê được xây dựng để tách rời tín hiệu nhiều chiều thành các thành phần tín hiệu độc lập ẩn sâu bên dưới dữ liệu. Kỹ thuật này đòi hỏi phải đặt ra giả thuyết tồn tại các nguồn tín hiệu bên dưới nongaussianity và độc lập thống kê từng đôi một. Thuật toán ICA có nhiều ứng dụng rộng rãi trong nhiều bài toán khác nhau như xử lý tín hiệu, kinh tế học, sinh tin học,…

1.2 Biểu đồ



2. FA (Factor Analysis)

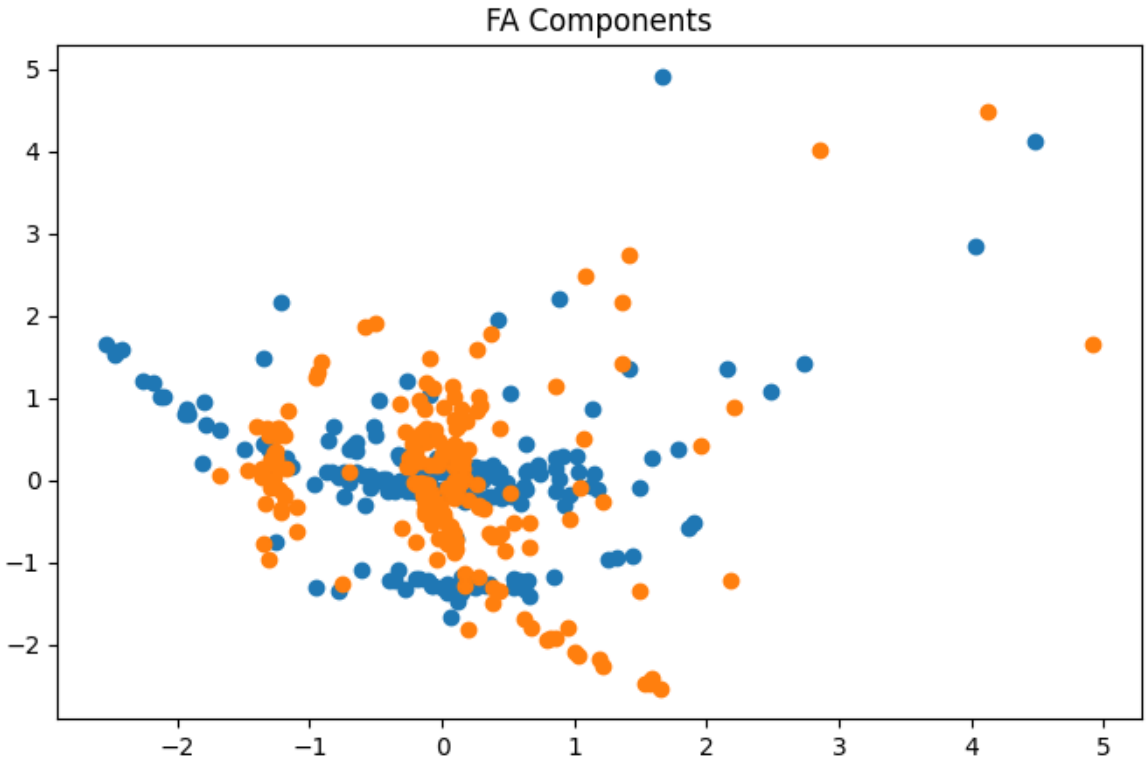
2.1 Tương quan về FA

- Phân tích nhân tố (FA) là một phương pháp thống kê được sử dụng để xác định các mẫu giữa các biến quan sát và giải thích cấu trúc cơ bản của các biến đó. Nó nhằm mục đích giảm số lượng các biến bằng cách nhóm chúng thành các yếu tố cơ bản.

- FA giả định rằng các biến quan sát bị ảnh hưởng bởi một hoặc nhiều yếu tố cơ bản và các yếu tố này độc lập với nhau.

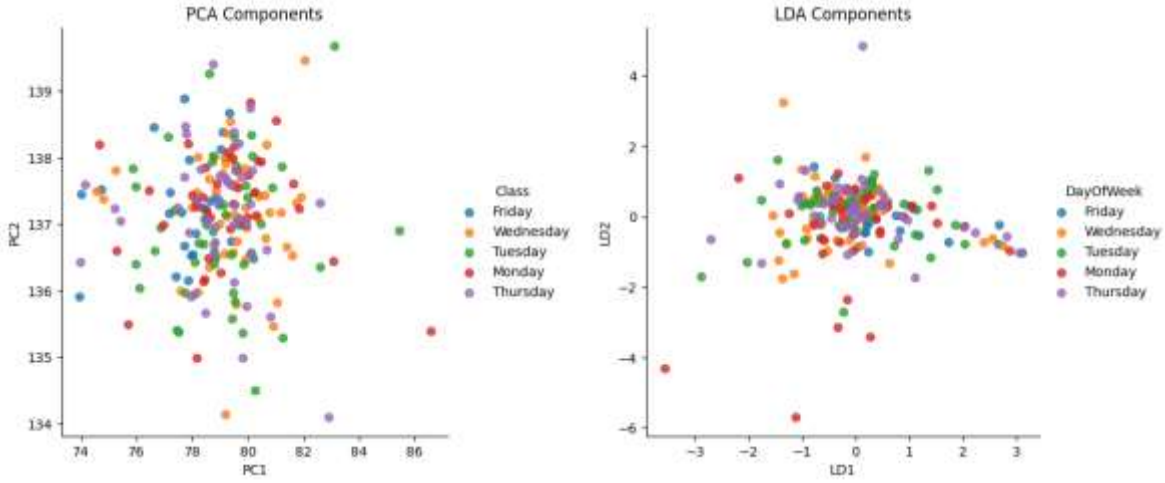
- FA được sử dụng cho mục đích khám phá hoặc xác nhận và thường được sử dụng trong các lĩnh vực như tâm lý học, xã hội học và nghiên cứu tiếp thị để xác định các cấu trúc tiềm ẩn giải thích dữ liệu quan sát được

2.2 Biểu đồ



CHƯƠNG 6: So sánh PCA và LDA

1. Biểu đồ của PCA và LDA



2. So sánh

- Dựa vào biểu đồ có thể thấy được, dữ liệu ở PCA cho ra nhiều hơn so với LDA nhưng ở PCA rải rác khắp nơi khó nhận định được đâu là thông tin cần thiết để triển khai so với LDA. LDA thì hiển thị những thông tin cần thiết nhất để có thể trực quan được về dữ liệu hơn.

- Lí do vì sao LDA làm được điều đó là vì LDA còn xét trên phương sai còn PCA thì không. LDA được giám sát trong khi PCA không được giám sát và PCA bỏ qua các nhãn lớp.