APRENDIZAJE NO SUPERVISADO

Análisis de Componentes Principales

Sistemas de Inteligencia Artificial - 2024

TABLA DE CONTENIDOS

05

ANÁLISIS DE COMPONENTES PRINCIPALES

05.1

INTRODUCCIÓN

EJEMPLO

Se cuenta con un conjunto de datos tomado de una competencia de natación. Los nadadores debieron recorrer 4 tramos y se tomaron los siguientes tiempos (min):

NADADOR	TR1	TR2	TR3	TR4
1	10	10	13	12
2	12	12	14	15
3	11	10	14	13
4	9	9	11	11
5	8	8	9	8
6	8	9	10	9
7	10	10	8	9

EJEMPLO

Para una investigación se seleccionan al azar 3500 pacientes hipertensos sobre los que se midieron las siguiente variables:

- Edad
- Duración de los síntomas (cantidad de días)
- Colesterol (mg/dl)
- Sexo
- Peso (kg)
- Presión arterial media (mm Hg)
- Medida del estrés

Se desea saber si el paciente presenta una enfermedad arterial o no

MEDIDAS DESCRIPTIVAS

Dado un conjunto de datos de una variable $X = \{x_1, x_2, \dots, x_n\}$

MEDIA MUESTRAL

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

VARIANZA MUESTRAL

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

NOTA: Se toma una parte de la población

MEDIDAS DESCRIPTIVAS

Dado un conjunto de datos con n variables y m observaciones.
La covarianza muestral mide la asociación lineal entre las variables X_i y X_k

$$\begin{bmatrix} X_1 & X_2 & \dots & X_n \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

COVARIANZA MUESTRAL

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{m} ((x_{ji} - \overline{x_i})(x_{jk} - \overline{x_k}))$$

MATRIZ DE COVARIANZAS

Las covarianzas forman una matriz simétrica definida positiva:

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{12} & \dots & s_{1n} \\ \dots & & & \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{bmatrix}$$

INTERPRETACIÓN COVARIANZA MUESTRAL

Sea s_{ik} la covarianza muestral entre las variables X_i y X_k

- $\mathbf{s_{ik} > 0}$ indica asociación lineal **positiva** entre los datos de las variables
- $\mathbf{s_{ik} < 0}$ indica asociación lineal **negativa** entre los datos de las variables
- $\mathbf{s_{ik} = 0}$ indica que las variables son **independientes**

Nota:

La varianza muestral es la covarianza muestral entre los datos de la variable X_i con ella misma (se puede denotar como $\mathbf{s_{ii}}$)

CORRELACIÓN MUESTRAL

La correlación muestral es otra medida de asociación lineal.
Es la covarianza muestral con las variables estandarizadas.

VARIABLE ESTANDARIZADA

$$\tilde{X}_i = \frac{X_i - \bar{X}_i}{s_i}$$

La **correlación muestral** para las variables X_i y X_k se define como:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

05.2

PCA

ANÁLISIS DE COMPONENTES PRINCIPALES

Si las variables de un conjunto de datos están *muy correlacionadas* entonces posee **información redundante**

El objetivo del método de *Principal Component Analysis* **(PCA)**
es **eliminar la redundancia**

¿CÓMO?

ANÁLISIS DE COMPONENTES PRINCIPALES

Transformar el conjunto de variables original en otro conjunto

Tendrá variables que son combinaciones lineales de las anteriores
pero no están correlacionadas entre sí

CONJUNTO DE COMPONENTES PRINCIPALES

PCA - HISTORIA

Técnica publicada por H. Hotelling en 1933.

Las primeras versiones se encuentran en los ajustes ortogonales por cuadrados mínimos introducidos por K. Pearson en 1901.

Analysis of a complex of statistical variables into principal components.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441.
https://doi.org/10.1037/h0071325

The problem is stated in detail, a method of analysis is derived and its geometrical meaning shown, methods of solution are illustrated and certain derivative problems are discussed. (To be concluded in October issue.) (APA PsycInfo Database Record (c) 2016 APA, all rights reserved)

Paper

PCA

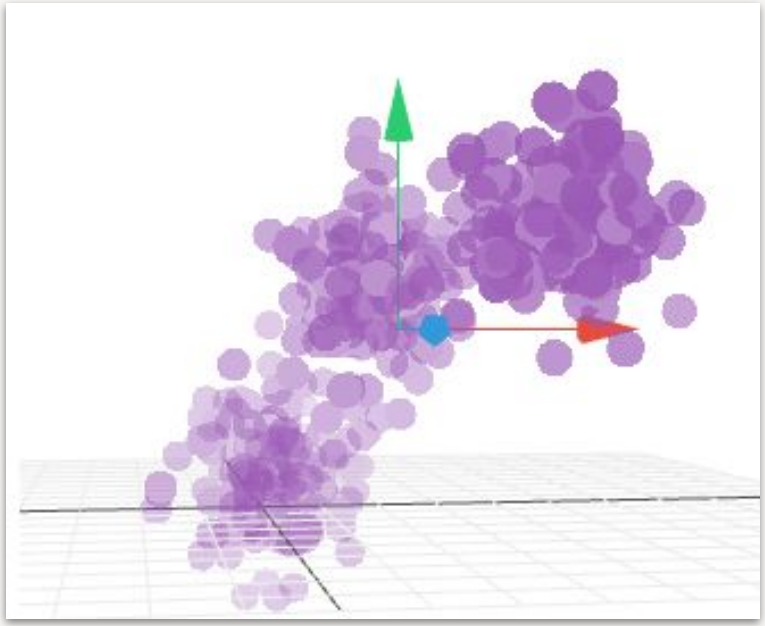
Dadas *p* variables originales:

Se desean encontrar ***q < p*** variables que sean combinaciones lineales de las *p* originales, recogiendo la **mayor** parte de la información o **variabilidad** de los datos.

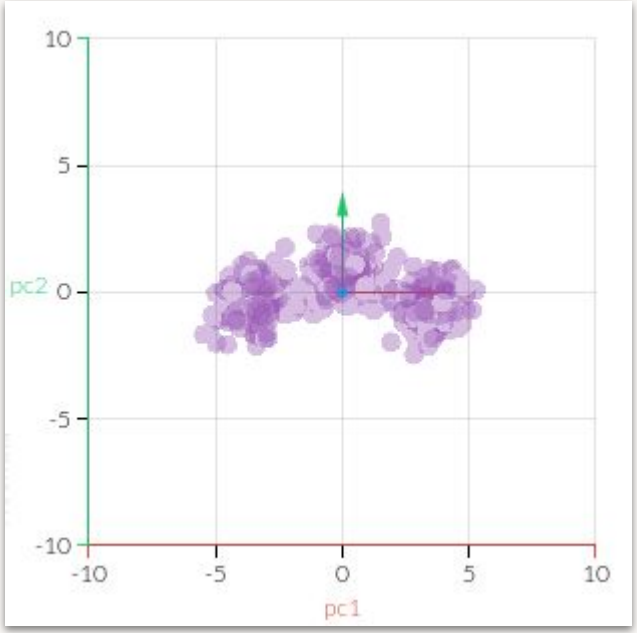
Si las variables originales no están correlacionadas, entonces no tiene sentido realizar un Análisis de Componentes Principales.

Se hallan *p* componentes principales

PCA



PCA



Interactive PCA Example

PCA - VARIABILIDAD

¿Cómo sabemos que la variable “es principal”?

Se toma la característica que **maximiza la variabilidad,** es un buen factor para diferenciar objetos en un conjunto de datos.

Ejemplo:

Se tiene un conjunto de datos de vehículos.

- Característica 1: Cantidad de ruedas
- Característica 2: Longitud del vehículo (Mayor variabilidad)

PCA - TRANSFORMACIÓN

Sea X^{nxp} una matriz con:

- n elementos de la población
- p variables $\{x_1, x_2, ..., x_p\}$

Sea realiza una **transformación** de X de forma tal que la varianza del nuevo conjunto de variables sea máxima

LA PRIMERA COMPONENTE

Sea y_1 la primera componente principal, y un vector de cargas $\overline{a} = (a_{11}, a_{12}, ..., a_{1p}) \epsilon \mathbb{R}^p$

$$y_1 = \sum_{j=1}^{p} a_{1j}(x_j - \overline{x_j})$$

$$y_1 = a_{11}(x_1 - \overline{x_1}) + ... + a_{1p}(x_p - \overline{x_p})$$

Nota: si las variables están estandarizadas no es necesario restar la media

CARGAS

El conjunto de componentes principales es una combinación lineal de las variables originales.

Se desea encontrar la carga $\overline{a_1} \ / ||\overline{a_1}|| = 1$ y la $Var(y_1)$ resulte máxima

Los coeficientes $a_{ji}, i = 1...p, j = 1...p$ se denominan **cargas (loadings)**

¿Cómo hallar las cargas?

Son los autovectores de la matriz de covarianzas (o correlaciones)

BUSCAMOS AyA

Para hallar las componentes principales se deben calcular los autovectores v_i correspondientes a las cargas

$$det(S_x - \lambda_i \mathbb{I}) = 0 \qquad S_x \overline{v_i} = \lambda_i \overline{v_i}$$

Entonces las componentes se calculan como:

$$y_1 = v_{11} x_1 + ... + v_{n1} x_n$$

$$y_i = v_{1i} x_1 + ... + v_{ni} x_n$$

BUSCAMOS AyA

El autovalor λ_i se corresponde con la **varianza** de la componente i

Ordenando los autovalores de mayor a menor se logra **reducir la dimensionalidad** tomando los autovectores correspondientes a los primeros *q* autovalores, que son los que proveen mayor información (en términos de variabilidad).

05.3

COVARIANZA vs CORRELACIÓN

COVARIANZA vs CORRELACIÓN

Si alguna de las variables toma valores mayores a las demás entonces tendrá mayor varianza, pero no quiere decir que tenga mayor variabilidad.

Cuando las escalas de medida de las variables son muy distintas, la maximización de la **varianza depende decisivamente de estas escalas de medida** y las variables con valores más grandes tendrán más peso en el análisis.

COVARIANZA vs CORRELACIÓN

Para evitar el problema se deben **estandarizar** las variables cuando calculamos las componentes principales

De esta manera las magnitudes de los valores numéricos de las variables originales serán comparables

COVARIANZA vs CORRELACIÓN

Esto equivale a aplicar el análisis de componentes principales utilizando la **matriz de correlaciones** en lugar de la matriz de covarianzas.

Cuando las variables tienen las mismas unidades, ambas alternativas son posibles.

EJERCICIO

Sea la matriz de correlaciones:

$$S = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

correspondiente a las (X_1, X_2, X_3) variables aleatorias

Ejercicio:
Escribir las variables (Y_1, Y_2, Y_3) de componentes principales y calcular qué proporción de la varianza total explica cada componente.

SOLUCIÓN

Sabemos que $Y_1 = a_1X_1 + a_2X_2 + a_3X_3$

El primer autovector es $v_1 = (a_1, a_2, a_3)$ que se corresponde con el mayor autovalor λ_1

El mayor autovalor es la varianza de Y_1

SOLUCIÓN

AUTOVALORES $\lambda_1 = 6, \lambda_2 = 3, \lambda_3 = 2$

AUTOVECTORES

$$V = \begin{bmatrix} -0.40 & -0.57 & 0.70 \\ -0.40 & -0.57 & -0.70 \\ -0.81 & 0.57 & 0 \end{bmatrix}$$

$$Y_1 = -0.40X_1 - 0.40X_2 - 0.81X_3$$

$$Y_2 = -0.57X_1 - 0.57X_2 + 0.57X_3$$

$$Y_3 = 0.70X_1 - 0.70X_2$$

SOLUCIÓN

AUTOVALORES $\lambda_1 = 6, \lambda_2 = 3, \lambda_3 = 2$

La proporción de la varianza de cada componente es:

De Y1 = 6/11
De Y2 = 3/11
De Y3 = 2/11

PROCEDIMIENTO

1. Construir la matriz **X** a partir del dataset, poniendo las variables en columnas
2. **Estandarizar** las variables X
3. Calcular la matriz de correlaciones **Sx**
4. Calcular **AyA**
5. Ordenar los autovalores de mayor a menor
6. Construir la matriz **V** con los autovectores con mayor autovalor
7. Calcular las nuevas variables **Y** como **combinación lineal** de las originales

05.4

INTERPRETACIÓN

INTERPRETACIÓN DE PC1

EJEMPLO:

Los datos de una encuesta de presupuestos familiares en distintas provincias, presentan los gastos medios de las familias utilizando seis variables:

- X1: alimentación
- X2: vestido y calzado
- X3: vivienda
- X4: mobiliario doméstico
- X5: salud
- X6: educación y cultura

INTERPRETACIÓN DE PC1

Calculamos el primer autovector y obtenemos PC1:

$$Y_1 = 0.50X_1 + 0.22X_2 + 0.35X_3 + 0.33X_4 + 0.48X_5 + 0.49X_6$$

¿Qué información nos brinda?

INTERPRETACIÓN DE PC1

Calculamos el primer autovector y obtenemos PC1:

$$Y_1 = 0.50X_1 + 0.22X_2 + 0.35X_3 + 0.33X_4 + 0.48X_5 + 0.49X_6$$

- Y1 es una suma ponderada de todos los gastos, con mayor carga, de las variables: X1 alimentación, X5 salud y X6 educación.
- El menor peso lo tiene el gasto en X2 vestido y calzado.

INTERPRETACIÓN DE PC1

Si calculamos los valores de Y1 para cada provincia y las ordenamos por esta nueva variable quedan ordenadas por sus gastos.

Explicación inmediata: muestra la **capacidad de gasto**

INTERPRETACIÓN DE PC1

- Si la carga (coeficiente o loading) de una variable en la componente principal es positiva, significa que la variable y la componente tienen una correlación positiva.

- Si la carga es negativa, la variable se correlaciona en forma negativa con la primera componente.

- La primera componente representa un **índice** (o una característica) por el cual se pueden ordenar los registros.

05.5

EJERCICIO CHATBOT

EJEMPLO

Una compañía B2B brinda soluciones de servicio al cliente a través del desarrollo de chatbots.

Se desea optimizar el flujo del chatbot con el objetivo de:

- automatizar el proceso de ventas de un cliente
- recurrir a asistencia humana lo menos posible
- mejorar la experiencia de usuario

EJEMPLO

Para ello se recopilan datos sobre las interacciones de los usuarios con el chatbot:

1. Longitud del input del usuario (cantidad de palabras/tokens)
2. Longitud del mensaje enviado por el chatbot (cantidad de palabras/tokens)
3. Índice de complejidad de la respuesta del chatbot
4. Tiempo de respuesta
5. Frecuencia de solicitudes de input hechas por el bot al usuario
6. Satisfacción del usuario (1 al 10)
7. Frecuencia de input types (text, images, voice, emojis)
8. Cantidad de interacciones del usuario por sesión
9. Número de excepciones
10. Solicitudes de asistencia humana

EJEMPLO

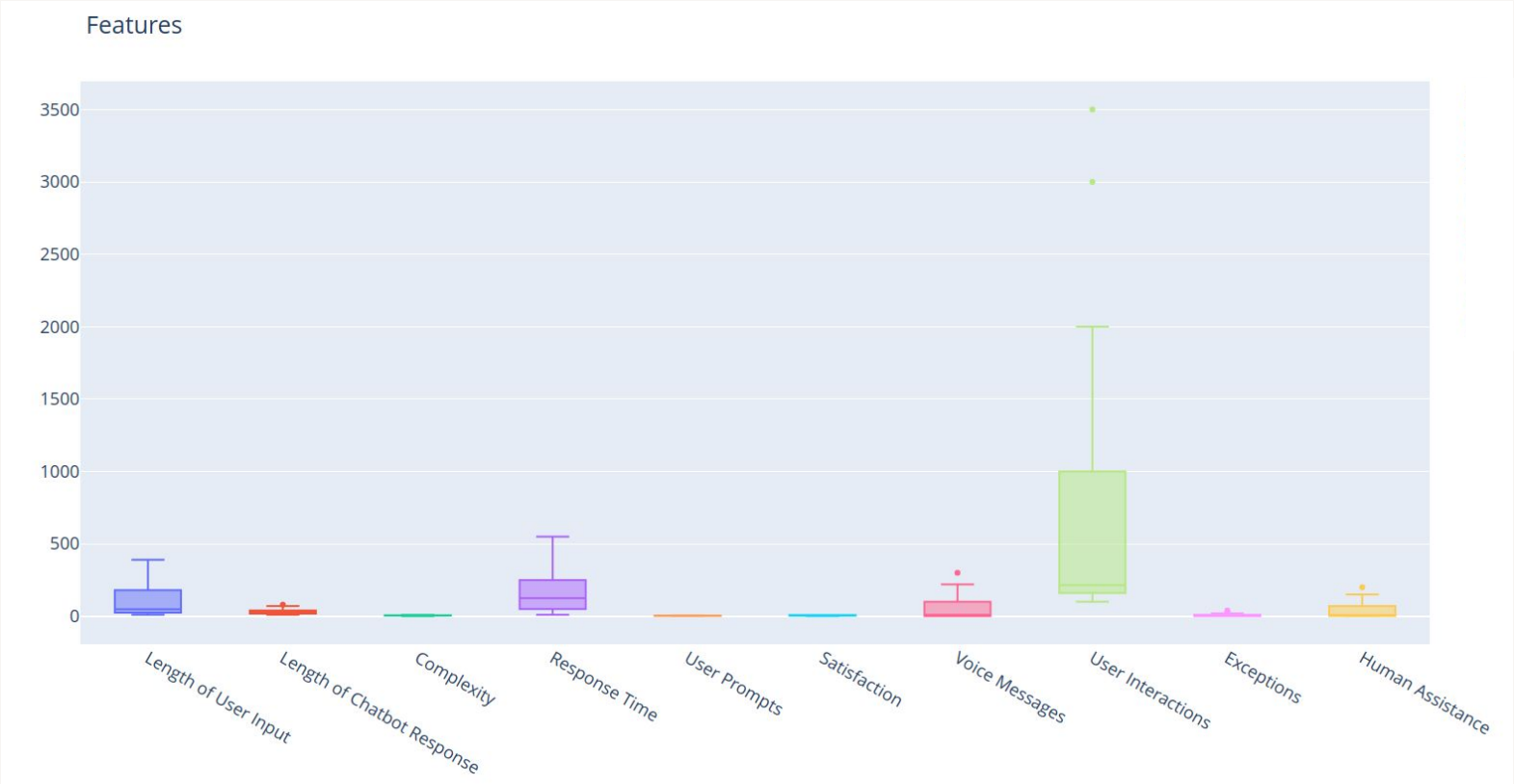
Se cuenta con un conjunto de datos con:

- 10 variables o atributos (columnas)
- 26 observaciones (filas)

El conjunto de datos es multivariado

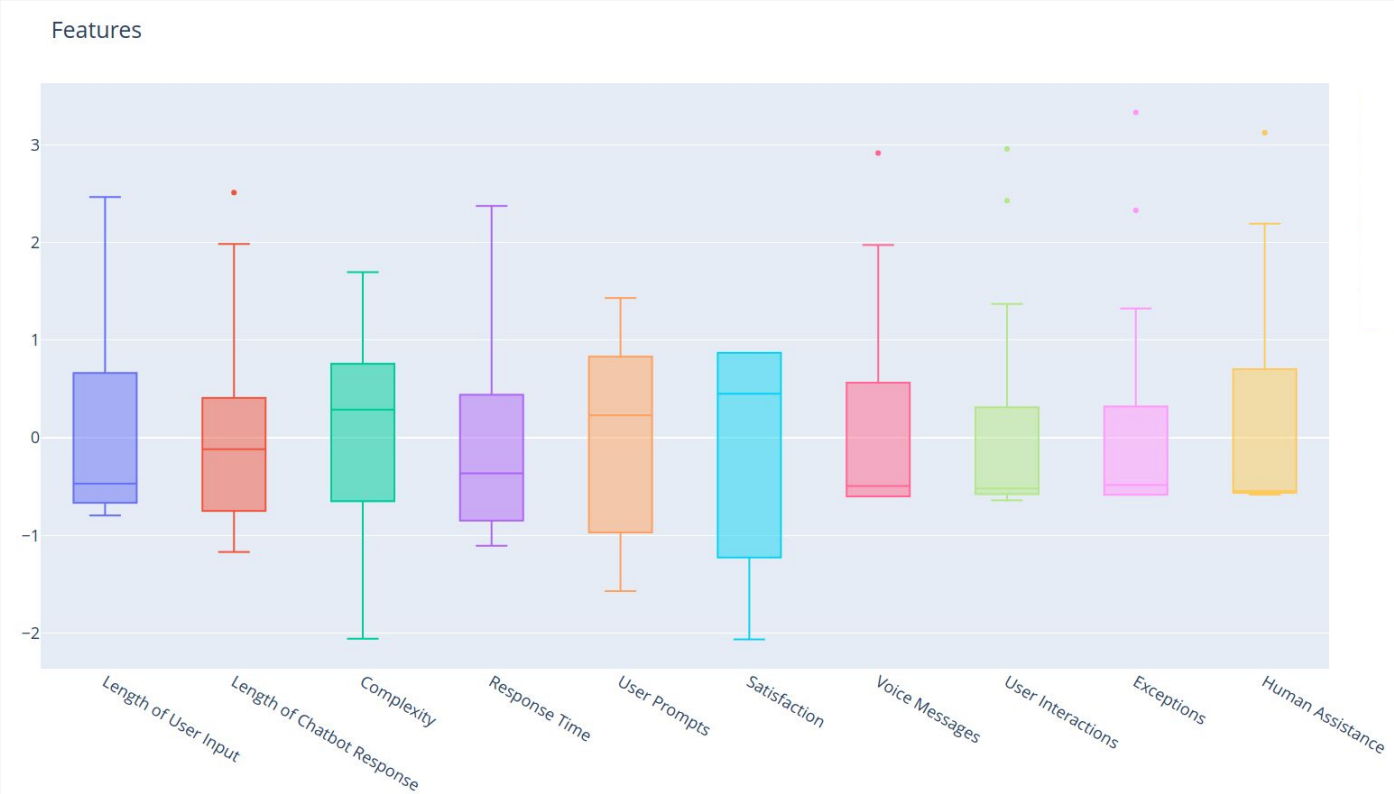
¿Cuáles son las variables que más influyen en la satisfacción del usuario?

BOXPLOT

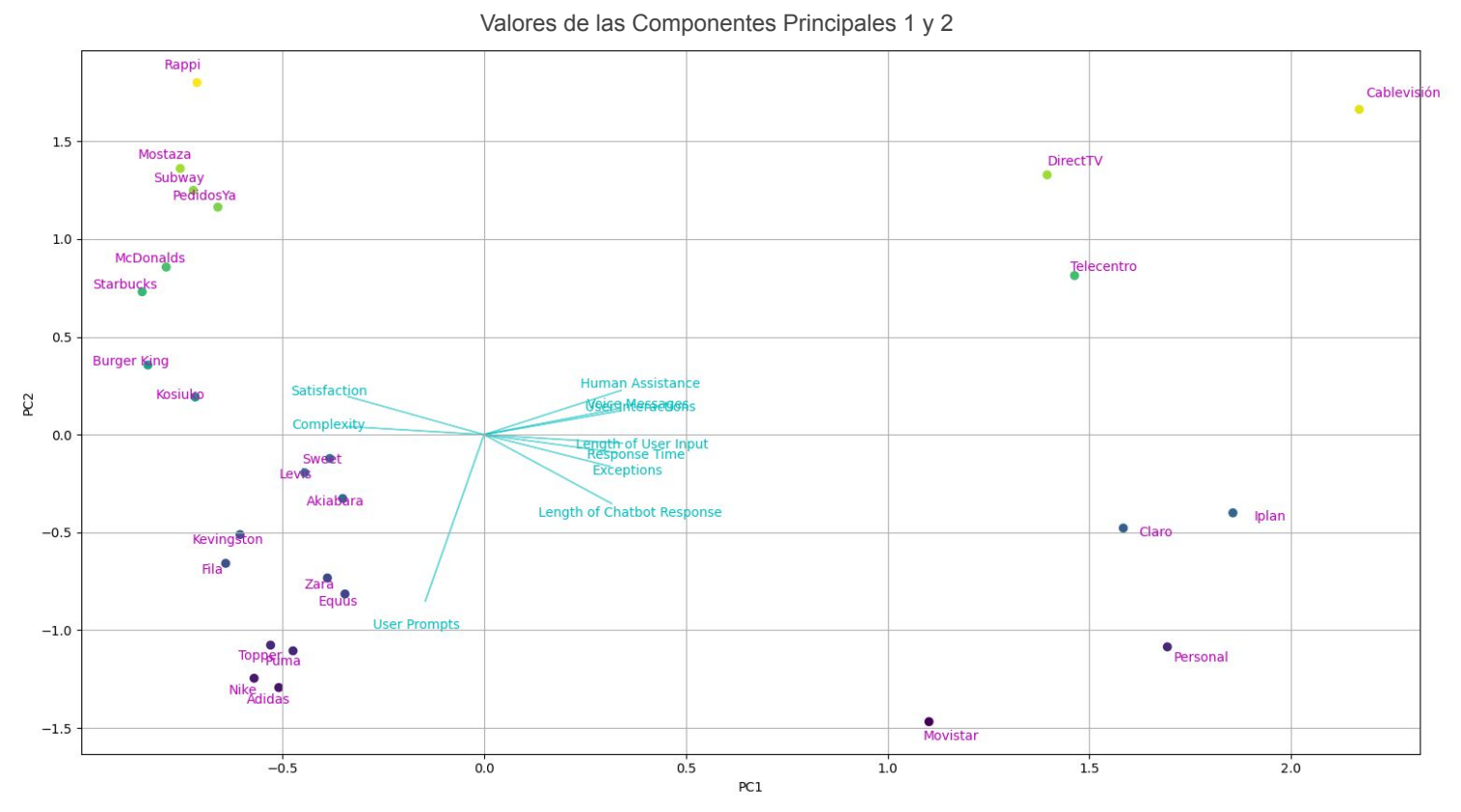

Features
3500
3000
2500
2000
1500
1000
500
0
Length of User Input
Length of Chatbot Response
Complexity
Response Time
User Prompts
Satisfaction
Voice Messages
User Interactions
Exceptions
Human Assistance

BOXPLOT

Si estandarizamos las variables

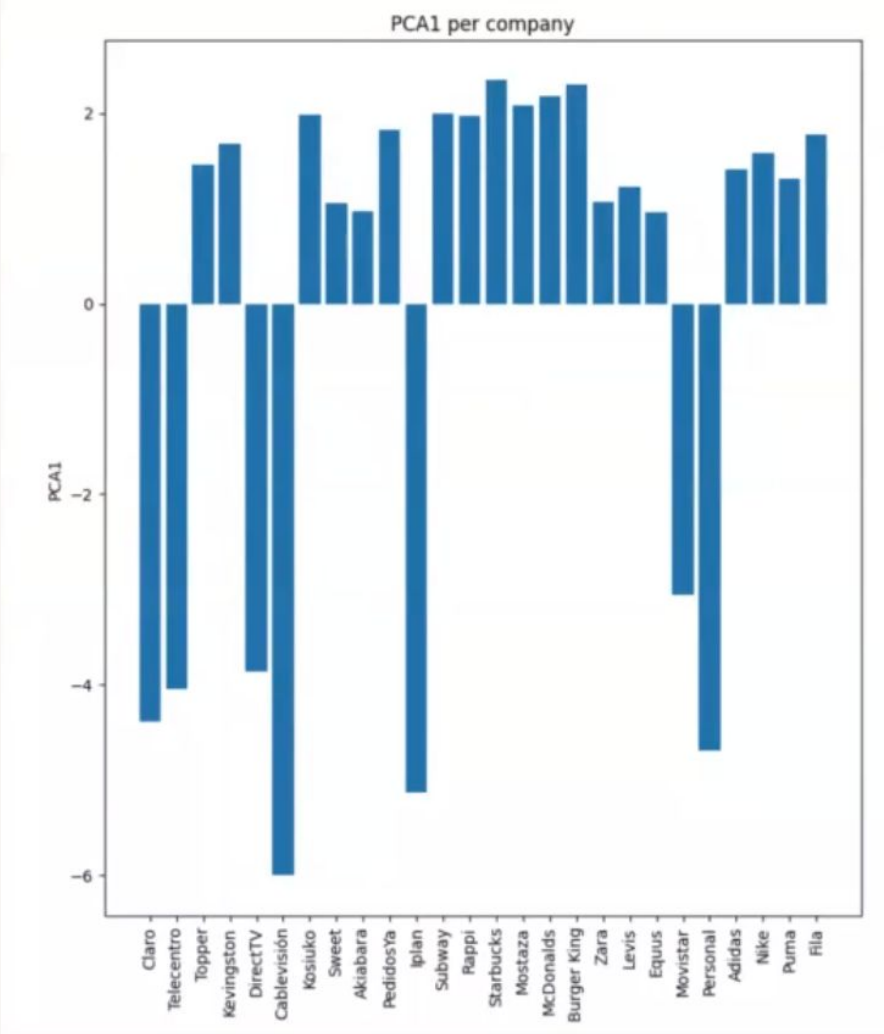


BIPLOT


Valores de las Componentes Principales 1 y 2
Rappi
Mostaza
Subway
PedidosYa
McDonalds
Starbucks
Burger King
Kosiuko
Kevingston
Fila
Topper
Puma
Nike
Adidas
Zara
Equus
Akiabara
Sweet
Levis
Cablevisión
DirectTV
Telecentro
Iplan
Claro
Personal
Movistar
Satisfaction
Complexity
Human Assistance
Voice Messages
User Interactions
Length of User Input
Response Time
Exceptions
Length of Chatbot Response
User Prompts
PC1
PC2

BIPLOT

Índice a partir del cálculo de la PC1



EJERCICIO OBLIGATORIO

Utilizar una librería para calcular las componentes principales e interpretar la PC1 (gráfica y teóricamente)

El conjunto de datos europe.csv corresponde a características económicas, sociales y geográficas de 28 países de Europa. Las variables son:

- **Country** : Nombre del país.
- **Area:** área.
- **GDP:** producto bruto interno.
- **Inflation:** inflación anual.
- **Life.expect:** expectativa de vida media en años.
- **Military:** presupuesto militar.
- **Pop.growth:** tasa de crecimiento poblacional.
- **Unemployment:** tasa de desempleo.