Práctica 2 - PCA

Métodos Numéricos y Optimización

Mariano Villafuerte González - 156057

Objetivo

El objetivo de esta práctica es mostrar el uso del análisis de componentes principales (PCA por sus siglas en inglés). Para esto, se toma la base de datos de *Red Wine Quality* de Kaggle.

En esta práctica se mostrará todo el proceso de la aplicación de PCA, sin embargo, si se desea saber más sobre el tema se puede consultar el libro de Optimización de Erick Palacios.

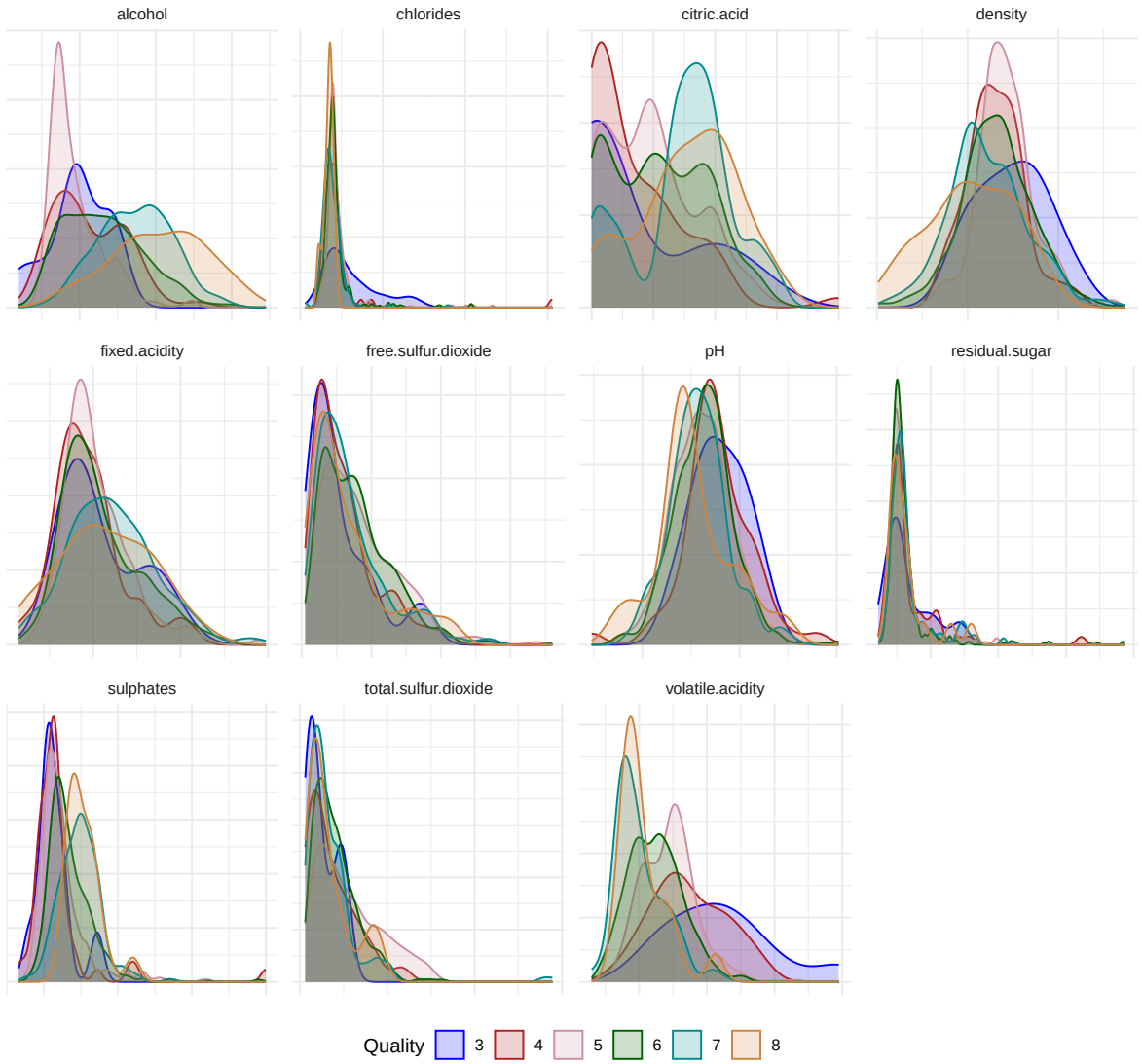
Análisis Exploratorio

La base de datos de *Red Wine Quality* cuenta con doce variables:

1. Fixed acidity
2. Volatile Acidity
3. Citric Acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol
12. Quality

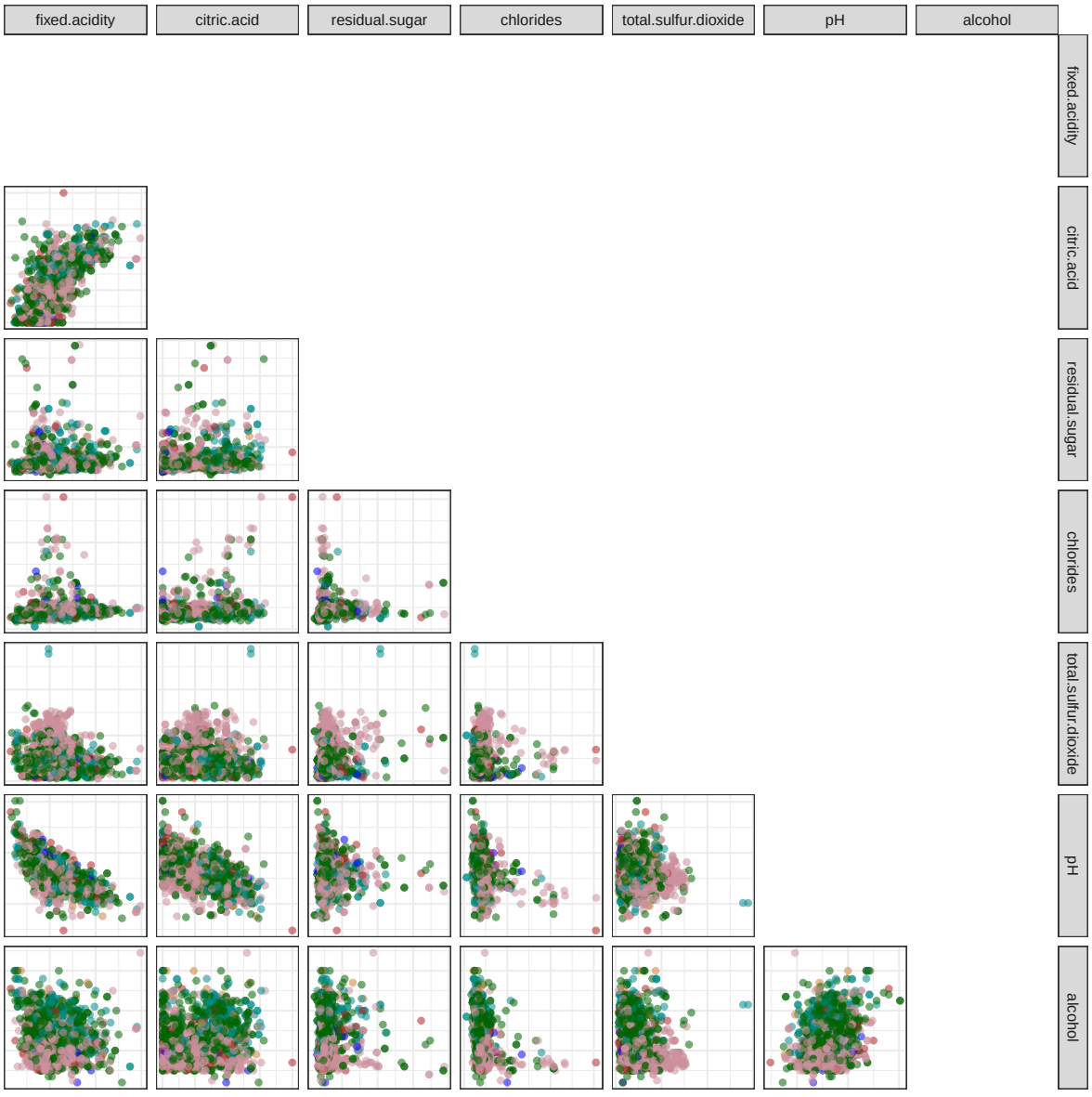
Ahora bien, para un experto en vinos, puede resultar evidente qué variables sirven para determinar la calidad de un vino rojo, sin embargo, intentaremos hacer uso de un análisis descriptivo para entender mejor qué relación existe entre las variables.

Nuestro primer acercamiento es ver la distribución de cada variable para las calificaciones de calidad que tenemos disponibles:



De este primer vistazo, podemos identificar algunos comportamientos interesantes, por ejemplo, el nivel de alcohol parece ser más alto en los vinos de mayor puntuación de calidad. La acidez volátil es mayor en los vinos con peor calificación. El ácido cítrico parece ser más elevado en vinos con mejor puntuación.

Encontrar todas las relaciones entre las variables que pudieran explicar la calidad del Vino Rojo requeriría de un análisis exploratorio más profundo y de otro tipo de pruebas que salen del enfoque de este trabajo, sin embargo, podemos estudiar cómo se ve la nube de datos tras aplicar el análisis de componentes principales. Para esto, primero vemos cómo se ven las nubes de datos sin ningún tipo de transformación:



Preparación de los Datos

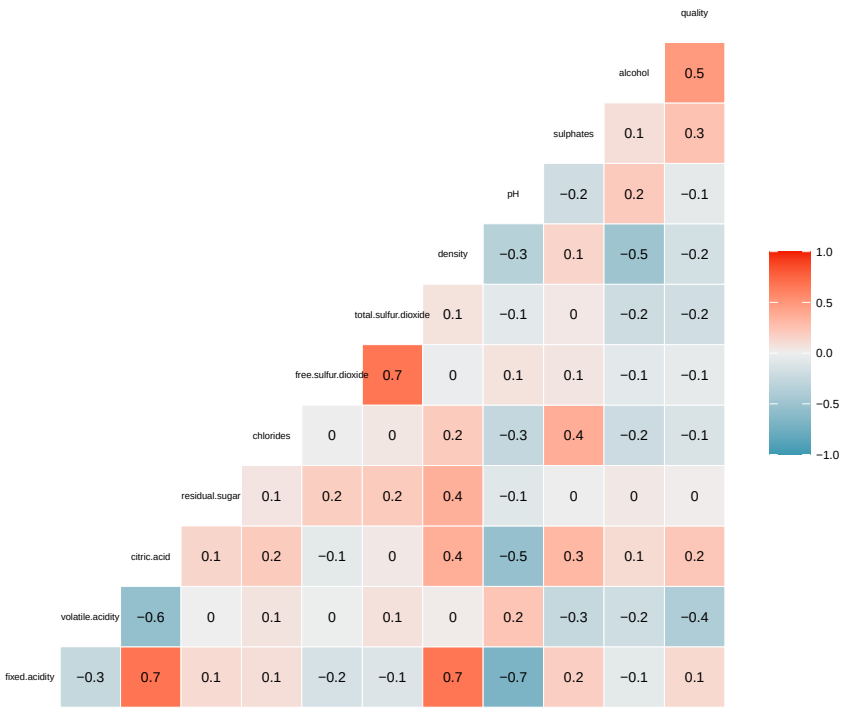
El primer paso para este ejercicio consiste en centrar nuestros datos alrededor del cero y controlar sus escalas. Para esto, a cada variable se le resta su media y se divide entre la varianza. Esto se realiza con el objetivo de que las variables estén en la misma escala y al utilizar métodos que representen distancias, éstas no se vean afectadas meramente por la naturaleza de las variables si no por su comportamiento.

Para hacer el centrado de los datos y el escalamiento, para cada columna (variable) se calcula para cada observación (x_i):

$$z_i = \frac{x_i - \mu}{\sigma}$$

Con μ =promedio de la variable, y σ =desviación estándar de la variable.

Y podemos visualizar cómo se comporta la matriz de correlaciones de las variables centradas y escaladas.



Con este mapa de calor, podemos observar que existe una relación lineal fuerte entre algunas variables, por ejemplo: entre *fixed acidity* y *citric acid.* Para los siguientes pasos, dejaremos fuera la variable de *quality* y veremos, al final de la aplicación de componentes principales, si estos logran captar la estructura de la calidad.

Cálculo de eigenvalores y eigenvectores

Para realizar la reducción de dimensionalidad mediante PCA, lo primero que se hace es calcular los eigenvalores y eigenvectores de la matriz de varianza-covarianza. Obtenemos:

eigen valores
3.10
1.93
1.55
1.21
0.96
0.66
0.58
0.42
0.34
0.18
0.06

Varianza Acumulada

Y dados estos eigenvalores, podemos calcular la varianza explicada de la siguiente manera:

$$V_E = \frac{\lambda_i}{\sum \lambda_i}$$

De modo que se obtienen los siguientes resultados:

componente	eigen valores	var explicada	var acum
1	3.10	28.17%	28.17%
2	1.93	17.51%	45.68%
3	1.55	14.10%	59.78%
4	1.21	11.03%	70.81%
5	0.96	8.72%	79.53%
6	0.66	6.00%	85.52%
7	0.58	5.31%	90.83%
8	0.42	3.85%	94.68%
9	0.34	3.13%	97.81%
10	0.18	1.65%	99.46%
11	0.06	0.54%	100.00%

Esto querría decir que la variabilidad total de los datos originales son explicados en un 28.17% por el primer componente. Es decir, el primer componente basta para explicar la mitad de la variabilidad total de los datos.

Ahora bien, lo que se busca es reducir la dimensionalidad del problema, no obstante, cuántos componentes considerar no es una pregunta con respuesta trivial y dependerá de las necesidades de cada problema. Lo más común es determinar un umbral, por ejemplo el 95% y considerar los componentes necesarios para alcanzar mínimo este umbral establecido. De haber usado esa regla, en este caso nos quedaríamos con 8 componentes en total; de modo que se estaría perdiendo aproximadamente el 5% de la varianza de los datos.

Contribuciones por componente

Ahora bien, los nuevos componentes son una combinación lineal de las variables originales, por lo que puede resultar de interés analizar las variables que predominan en los primeros componentes (aquellos que explican la mayor variabilidad de los datos).

Para determinar esto, basta con analizar los eigenvectores que ya se calcularon anteriormente:

	X1	X2	X3	X4	X5	X6	X7	X8
fixed.acidity	0.49	−0.11	−0.12	0.23	0.08	−0.10	0.35	0.18
citric.acid	0.46	−0.15	0.24	0.08	0.06	−0.07	−0.11	0.38
pH	−0.44	0.01	0.06	0.00	−0.27	0.52	0.03	0.56
density	0.40	0.23	−0.34	0.17	−0.16	0.39	0.17	0.24
sulphates	0.24	−0.04	0.28	−0.55	−0.23	0.38	0.45	−0.37
volatile.acidity	−0.24	0.27	−0.45	−0.08	−0.22	−0.41	0.53	0.08
chlorides	0.21	0.15	−0.09	−0.67	−0.25	−0.30	−0.37	0.36
residual.sugar	0.15	0.27	0.10	0.37	−0.73	−0.05	−0.29	−0.30
alcohol	−0.11	−0.39	0.47	0.12	−0.35	−0.36	0.33	0.22
free.sulfur.dioxide	−0.04	0.51	0.43	0.04	0.16	0.01	0.12	0.20
total.sulfur.dioxide	0.02	0.57	0.32	0.03	0.22	−0.14	0.09	−0.02

Otro método es si usamos descomposición de valores singulares (SVD) con nuestra base de datos centrada y escalada. Recordamos que SVD regresa tres matrices:

$$SVD \longrightarrow X = u\Sigma v$$

Dada esta descomposición, podemos apreciar que la matriz v contiene los pesos de las combinaciones lineales para los componentes principales (se muestran los primeros 5 componentes):

	X1	X2	X3	X4	X5	X6	X7	X8
fixed.acidity	0.49	0.11	−0.12	0.23	−0.08	−0.10	0.35	−0.18
citric.acid	0.46	0.15	0.24	0.08	−0.06	−0.07	−0.11	−0.38
pH	−0.44	−0.01	0.06	0.00	0.27	0.52	0.03	−0.56
density	0.40	−0.23	−0.34	0.17	0.16	0.39	0.17	−0.24
sulphates	0.24	0.04	0.28	−0.55	0.23	0.38	0.45	0.37
volatile.acidity	−0.24	−0.27	−0.45	−0.08	0.22	−0.41	0.53	−0.08
chlorides	0.21	−0.15	−0.09	−0.67	0.25	−0.30	−0.37	−0.36
residual.sugar	0.15	−0.27	0.10	0.37	0.73	−0.05	−0.29	0.30
alcohol	−0.11	0.39	0.47	0.12	0.35	−0.36	0.33	−0.22
free.sulfur.dioxide	−0.04	−0.51	0.43	0.04	−0.16	0.01	0.12	−0.20
total.sulfur.dioxide	0.02	−0.57	0.32	0.03	−0.22	−0.14	0.09	0.02

Ahora nos enfocamos en los primeros dos componentes principales. Se puede apreciar que las variables con mayor peso para el primer componente son:

- *Fixed acidity*
- *Citric Acid*
- *pH*
- *Density*

Y para el segundo componente son:

- *Total sulfur dioxide*
- *Free sulfur dioxide*
- *Alcohol*
- *Volatile acidity*
- *Residual sugar*

Así pues, observamos que las variables que más aportan al primer componente (si observamos la matriz de correlaciones) tienen correlaciones altas entre ellas. Esto tiene sentido, el nuevo eje de mayor varianza podríamos considera la dispersión de estas variables que parecen ir en un mismo sentido.
Ahora bien, en el segundo componente las correlaciones entre las variables dominantes son mucho menores, sin embargo, se incluyen variables que durante el análisis exploratorio se identificaron como relevantes, tal como el nivel de alcohol.

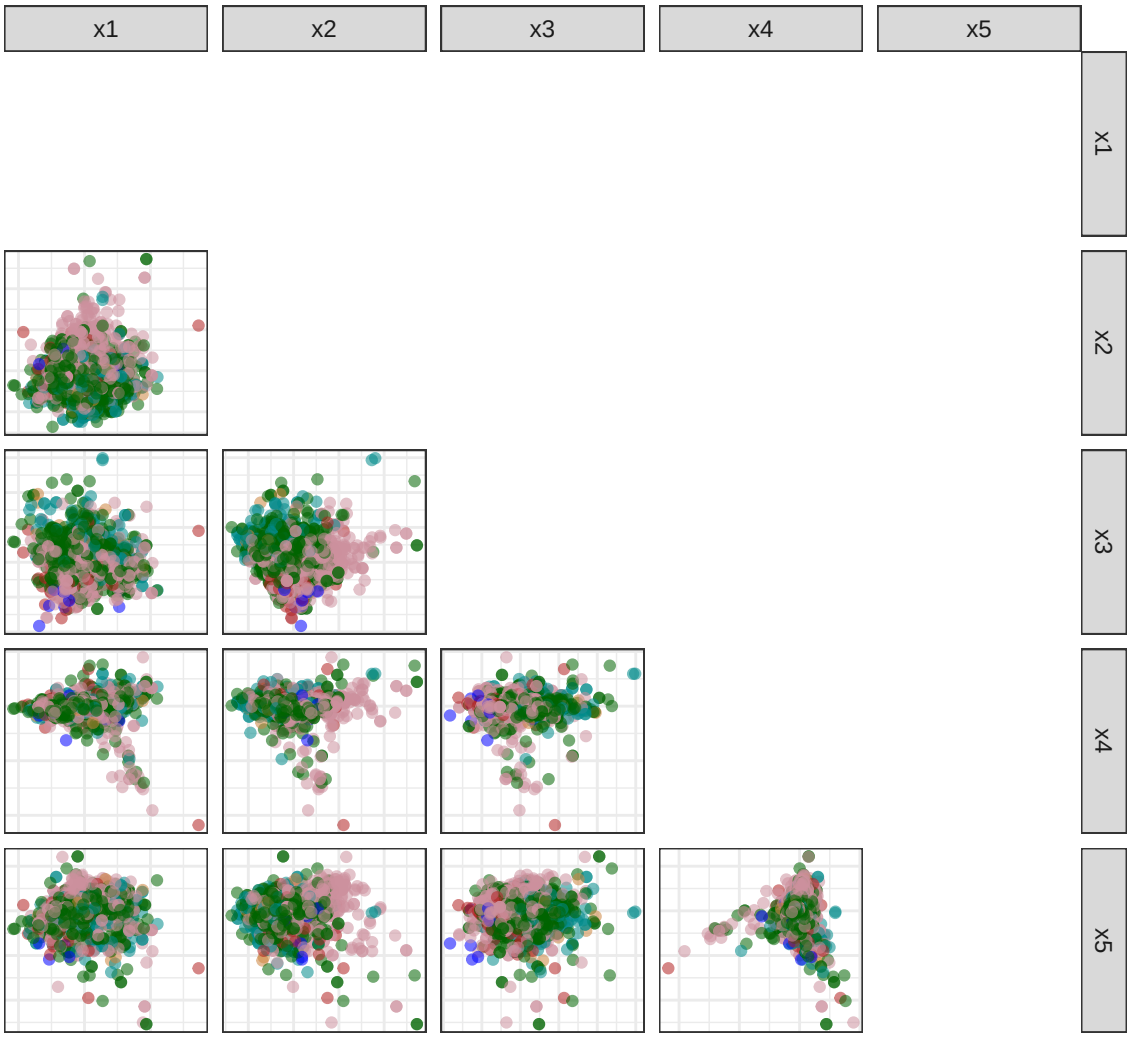
Visualización en los Nuevos Componentes

Lo que resta es ver el comportamiento de las variables en los nuevos componentes y estudiar si los nuevos componentes ayudan a segmentar de alguna manera los niveles de calidad (la variable que se excluyó para el análisis de PCA). Para hacer la proyección de nuestra base escalada y centrada al nuevo espacio de Componentes Principales, basta calcular:

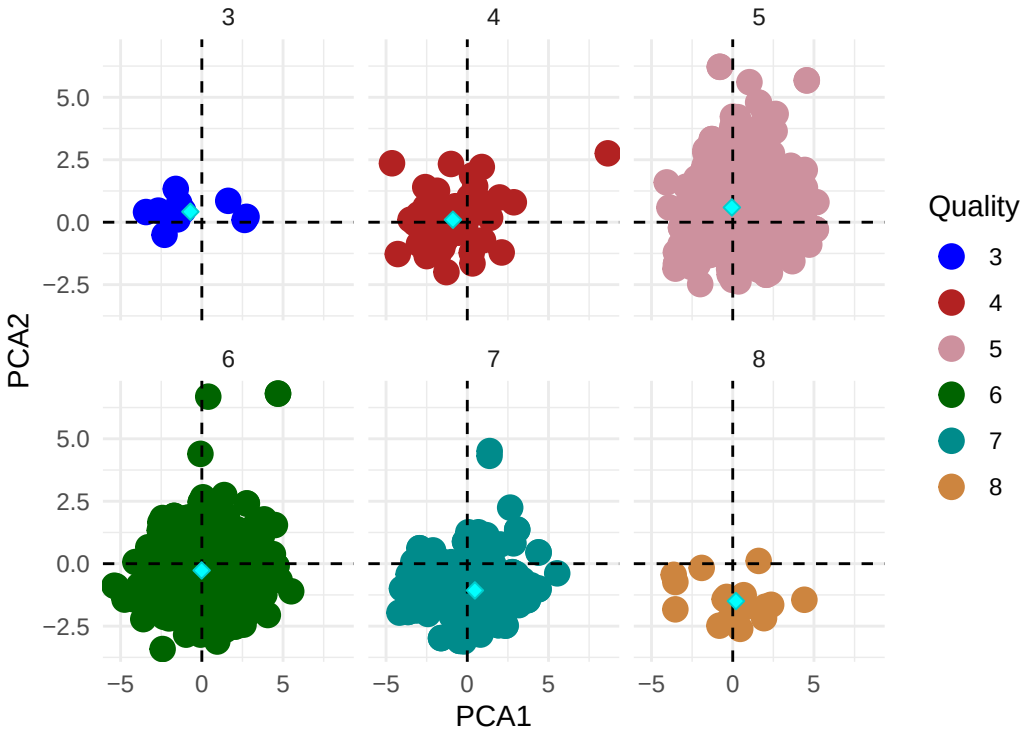
$$Y = XV$$

Con $V =$ los eigenvectores calculados.

La primera visualización es similar al análisis exploratorio realizado en un inicio. Vemos la dispersión dentro de los primeros 5 componentes:

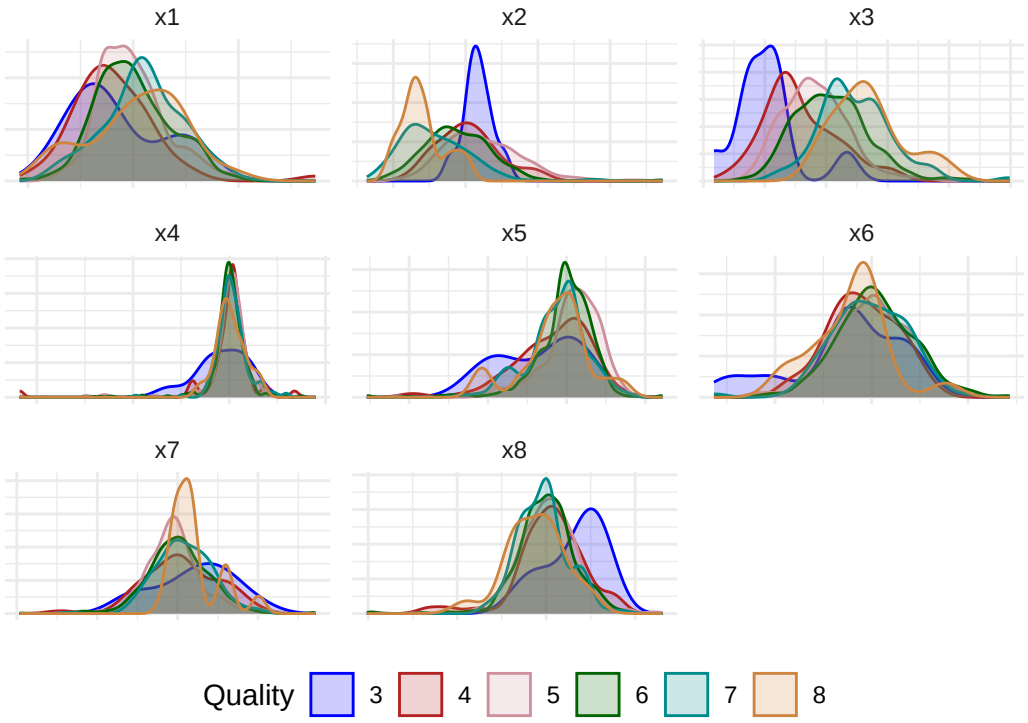


Podemos apreciar una mayor dispersión dentro de los primeros componentes. Podemos hacer otro análisis de dispersión:



Pareciera que, conforme aumenta la calidad la nube de datos tiende a moverse a la derecha y hacia abajo. Se incluye el promedio de cada nube, representado por el diamante color cian, y se puede apreciar el movimiento de este estadístico con la dinámica ya descrita.

Podemos hacer un análisis de dispersión para los ocho componentes principales:



Y podemos apreciar que, en efecto, parece ser que las calificaciones más altas están asociados a un PCA_1 más alto, a un PCA_2 más bajo, y a un PCA_3 más alto. En los otros componentes el comportamiento parece no ser tan distintivo entre niveles de calidad.

Conclusión

Este análisis de componentes principales podría ayudar con la segmentación y calificación de vinos. Es decir, estamos abstrayendo información valiosa. Considero que con 3 componentes principales se podría desarrollar un modelo de clasficicación o de conglomerados que discrimine de mejor manera los vinos de alta y menor calidad. Lo podemos apreciar por las nubes de puntos que parecieran ya estar más separados. Esto ayudaría a los modelos predictivos a hacer predicciones más "correctas" es decir, más apegadas a los valores que podría asignar, por ejemplo, un experto de vinos.

Datos y código

Para este ejercicio, se usa la base de datos de *Red Wine Quality* de Kaggle. Este set de datos está disponible aquí.

Este reporte se hace utilizando R, no obstante, se hizo igualmente una versión en Python para los lectores que lo prefieran. Ambas versiones se pueden consultar aquí.

Referencias

Palacios, Erick. (2022). *Optimización*. México: Jupyter Notebook. Disponible en: https://itam-ds.github.io/analisis-numerico-computo-cientifico/README.html