UNIVERSIDAD NACIONAL DE COLOMBIA

PREGRADO EN ESTADISTICA

Departamento de Estadística
Facultad de Ciencias

— Introducción al Análisis Multivariado —

Analisis de componentes principales para medir la calidad del vino

Multivariados

Datos: wine quality
tarea número 5

Integrantes:

Ricardo William Salazar Espinal C.C. 1017219472

Medellín, Colombia

Medellin, agosto 23 de 2024

Índice

Índice de Figuras **2**

Índice de Tablas **2**

1 Introducción **2**

2 Descripcion de la base de datos **2**

3 Objetivo del Análisis **3**

4 Metodología **4**

5 Analisis descriptivo del Data-set **4**

6 Matriz de correlaciones **7**

7 Valores propios y su varianza **7**

8 vectores propios **10**

8.1 Componente Principal 1 (PC1) 10

8.2 Componente Principal 2 (PC2) 10

8.3 Componente Principal 3 (PC3) 10

8.4 Componente Principal 4 (PC4) 11

9 Circulos de correlaciones **11**

10 Comunalidades **13**

11 Graficos Biplot **15**

12 Conclusiones **17**

Referencias **17**

Índice de figuras

1 box-plot datos sin normalizar 6
2 box-plot datos normalizados 6
3 grafico de codos con valores propios 9
4 grafico de codos varianza explicada 9
5 Contribuciones de Variables por Componente Principal 11
6 circulo de correlacion CP1 VS CP2 12
7 circulo de correlacion CP1 VS CP3 13
8 Biplot CP1 VS CP2 16
9 Biplot CP1 VS CP3 16

Índice de cuadros

1 Encabezado de tipo de vino rojo 4
2 Estadísticas Descriptivas de las Variables del Dataset 5
3 Matriz de Correlación de Variables Normalizadas 7
4 Resumen del Análisis de Componentes Principales (PCA) 8
5 Vectores Propios de las Componentes Principales 10
6 Cargas de las Primeras Cuatro Componentes Principales 14
7 Comunalidad Total Capturada por las Primeras Cuatro Componentes Principales 15

1 Introducción

El análisis de componentes principales (PCA) es una técnica estadística que se utiliza para reducir la dimensionalidad de un conjunto de datos mientras se conserva la mayor cantidad posible de variabilidad presente en los datos originales. En el contexto del análisis de vinos, el PCA puede ser una herramienta valiosa para identificar patrones y relaciones subyacentes entre diversas características químicas y sensoriales de los vinos.

2 Descripcion de la base de datos

Para el desarrollo del estudio consideraremos una base de datos con las siguientes variables:

acidezF (Acidez Fija): Mide la acidez total del vino, excluyendo el ácido volátil. Es una característica importante que influye en el sabor y la estabilidad del vino.

acidezV (Acidez Volátil): Representa la cantidad de ácido acético y otros ácidos volátiles presentes en el vino, los cuales pueden afectar negativamente el aroma y el sabor.

acidezC (Ácido Cítrico): Un ácido presente en el vino que contribuye al sabor y puede ayudar a estabilizar el vino.

azucar (Azúcar Residual): La cantidad de azúcar que queda en el vino después del proceso de fermentación, influyendo en la dulzura del vino.

Cl (Cloruros): La concentración de sales de cloro en el vino, que puede influir en el sabor y en la percepción de salinidad.

dioxidoAL (Dióxido de Azufre Libre): Un conservante utilizado para prevenir la oxidación y el crecimiento microbiano en el vino.

totaldioxidos (Dióxido de Azufre Total): La cantidad total de dióxido de azufre, que incluye tanto el libre como el combinado.

densidad (Densidad): La densidad del vino, que puede proporcionar información sobre el contenido de alcohol y azúcar.

PH: El pH del vino, que afecta su acidez y estabilidad.

Sulfatos: Compuestos que pueden influir en el sabor del vino y en su preservación.

Alcohol: El contenido de alcohol en el vino, un factor clave en la percepción del sabor y cuerpo del vino.

Calidad: Una calificación subjetiva que evalúa la calidad global del vino.

3 Objetivo del Análisis

El objetivo del análisis de componentes principales (PCA) en esta base de datos es identificar las dimensiones subyacentes que capturan la mayor parte de la variabilidad en las características químicas y sensoriales de los vinos. Mediante la reducción de dimensionalidad, el PCA ayudará a:

Identificar Patrones: Revelar patrones y relaciones entre las variables que no son evidentes a simple vista.

Reducir la Complejidad: Simplificar el conjunto de datos al reducir el número de variables necesarias para describir las características del vino, lo que facilita el análisis posterior y la visualización de los datos.

Clasificar y Agrupar: Ayudar en la identificación de grupos o clusters de vinos similares basados en sus características químicas y sensoriales.

Visualizar: Ofrecer una representación visual de la estructura de los datos en un espacio de menor dimensión, lo que facilita la interpretación y el análisis.

4 Metodología

El PCA se llevará a cabo en los siguientes pasos:

Normalización de Datos: Es esencial normalizar los datos para que todas las variables tengan una escala comparable.

Cálculo de Componentes Principales: Se calcularán los componentes principales y se evaluará la varianza explicada por cada componente.

Interpretación de Componentes: Se interpretarán los componentes principales para entender qué variables contribuyen más a cada componente.

Visualización: Se utilizarán gráficos de biplots o gráficos de scores para visualizar la distribución de los vinos en el espacio de componentes principales.

El PCA proporcionará una visión comprensiva de las características del vino y ayudará a mejorar la comprensión de cómo las distintas propiedades se relacionan con la calidad del vino.

5 Analisis descriptivo del Data-set

primero realizamos una visualizacion de algunos de los datos que conforman nuestro conjunto de datos, asi como algunos estadisticos utiles que permitan comprender el comportamiento de los datos.

Tabla 1: Encabezado de tipo de vino rojo

acidezF	acidezV	acidezC	azucar	Cl	dioxidoAL	totaldioxidos	densidad	ph	sulfatos	alcohol
7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4
7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8
7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8
11.2	0.28	0.56	1.9	0.075	17	60	0.9980	3.16	0.58	9.8
7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4
7.4	0.66	0.00	1.8	0.075	13	40	0.9978	3.51	0.56	9.4

Tabla 2: Estadísticas Descriptivas de las Variables del Dataset

	Media	Varianza	Maximo	Minimo
acidezF	8.3196373	3.0314160	15.90000	4.60000
acidezV	0.5278205	0.0320624	1.58000	0.12000
acidezC	0.2709756	0.0379475	1.00000	0.00000
azucar	2.5388055	1.9878970	15.50000	0.90000
Cl	0.0874665	0.0022151	0.61100	0.01200
dioxidoAL	15.8749218	109.4149000	72.00000	1.00000
totaldioxidos	46.4677924	1082.1020000	289.00000	6.00000
densidad	0.9967467	0.0000036	1.00369	0.99007
pH	3.3111132	0.0238352	4.01000	2.74000
sulfatos	0.6581488	0.0287326	2.00000	0.33000
alcohol	10.4229831	1.1356470	14.90000	8.40000

ahora obtenemos algunos estadisticos de interes para nuestro conjunto de datos como son las medias, las varianzas los valores maximos y minimos de cada una de nuestras variables, donde podemos indicar lo siguiente:

El análisis estadístico de las variables del conjunto de datos de vinos rojos revela varias características importantes. La acidez fija presenta una variabilidad notable, con valores que oscilan entre 4.60 y 15.90, y una media de 8.32, indicando una diversidad en la acidez entre los vinos. La acidez volátil y el ácido cítrico muestran una menor variabilidad, con medias bajas y varianzas pequeñas, sugiriendo que estas características son relativamente estables entre los vinos. Por otro lado, el azúcar residual y el dióxido de azufre libre muestran una amplia gama de valores y alta variabilidad, lo que indica una gran diversidad en el contenido de azúcar y dióxido de azufre entre los vinos. La concentración de cloruros es baja y muestra poca variabilidad, mientras que el dióxido de azufre total también presenta una alta variabilidad. La densidad y el pH tienen varianzas bajas, sugiriendo poca variabilidad en estas características. Los sulfatos tienen una media moderada con una gama amplia, mientras que el contenido de alcohol presenta una variabilidad moderada y un rango relativamente amplio. Finalmente, la calidad del vino varía entre 3 y 8, con una media de 5.64, lo que indica una distribución bastante uniforme de las calificaciones.

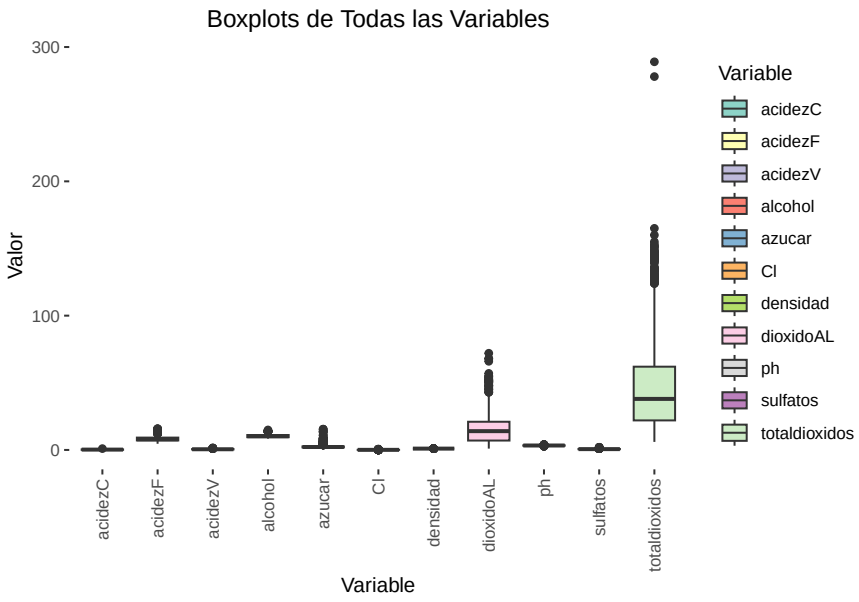

Figura 1: box-plot datos sin normalizar

observando el grafico de box-plot podemos notar que algunas medidas tienen diferentes medidas, por lo tanto tendremos que normalizar las variables para poder aplicar de manera adecuada el metodo de componentes principales.

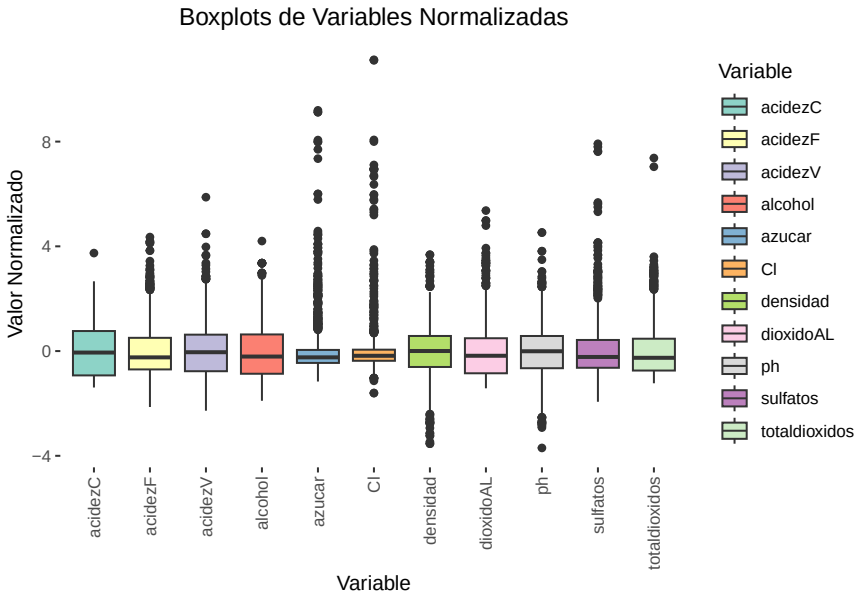

Figura 2: box-plot datos normalizados

Al momento de centralizar los datos observamos que algunas de ellas presentan gran catidad de datos atipicos especialemente en la variable **CL** y **Azucar** donde la gran cantidad de datos por atipicos hace que el tamaño de su caja se reduzca respecto a las otras.

6 Matriz de correlaciones

Como ya dijimos anteriormente algunas de las variables presentes tenian una alta variabilidad mientras que otras tenian una mas baja y uniforme, por lo tanto nuestro interes ahora es ver que tan relacionadas estan las variables entre si para ellos construiremos una matriz de correlaciones entre las variables presentes

Tabla 3: Matriz de Correlación de Variables Normalizadas

	acidezF	acidezV	acidezC	azucar	Cl	dioxidoAL	totaldioxidos	densidad	pH	sulfatos	alcohol
acidezF	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06
acidezV	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20
acidezC	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11
azucar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04
Cl	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22
dioxidoAL	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07
totaldioxidos	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21
densidad	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21
sulfatos	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00

La matriz de correlación muestra las relaciones entre diferentes características de los vinos. Entre las correlaciones más destacadas, se observa una relación positiva fuerte entre el dióxido de azufre libre y el dióxido de azufre total (0.67), lo que indica que a medida que aumenta el dióxido de azufre libre, también lo hace el nivel total de dióxido de azufre. También hay una correlación positiva notable entre la densidad y la acidez fija (0.67), sugiriendo que un aumento en la acidez fija está asociado con un incremento en la densidad del vino. Por otro lado, la matriz revela una relación negativa fuerte entre la densidad y el pH (-0.68), lo que implica que a medida que la densidad aumenta, el pH tiende a disminuir, un patrón que podría reflejar cambios en la acidez del vino. Otras correlaciones moderadas incluyen la relación positiva entre el azúcar residual y los sulfatos (0.37), y la relación negativa entre el alcohol y la acidez volátil (-0.26). En contraste, algunas relaciones son muy débiles, como la correlación entre los cloruros y el azúcar residual (0.09), y entre el dióxido de azufre libre y los cloruros (0.05). Este análisis proporciona una visión integral de cómo las variables relacionadas con el vino están interconectadas, lo cual puede ser útil para construir modelos predictivos y entender mejor la estructura de los datos.

7 Valores propios y su varianza

Ahora calcularemos los valores propios y realizaremos algunos analisis que nos llevaran a concluir cuantas componentes principales seran las que estudiaremos.

Las primeras 4 componentes juntas explican el 70.81% de la varianza total en el conjunto de datos. Esto indica que al considerar solo estas 4 componentes,se esta capturando una parte significativa de la variabilidad de los datos.

Componente 1 tiene un valor propio de 3.10 y explica 28.17% de la varianza total. Esto indica que esta componente captura una parte significativa de la variabilidad en los datos.

Tabla 4: Resumen del Análisis de Componentes Principales (PCA)

Component	Eigenvalue	Variance_Explained	Cumulative_Variance_Explained
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

Componente 2 explica 17.51% de la varianza, lo que hace que esta componente también sea importante, pero con menor impacto que la primera. **Componente 3** y **Componente 4** explican 14.10% y 11.03% de la varianza, respectivamente. Aunque estos valores son menores, todavía contribuyen de manera significativa a la descripción de la variabilidad en el conjunto de datos.

Adicionalemte los componentes con valores propios menores a 1 (componentes 5 a 11) explican una proporción muy pequeña de la varianza. Estos componentes generalmente se consideran menos significativos para la reducción de dimensionalidad, y se optara por no incluirlos en el análisis ya que buscamos una simplificación que conserve la mayor cantidad posible de información.

Para estar mas seguros en cuantas componentes principales usaremos realizaremos la regla del codo (scree plot) que es el punto donde el decrecimiento de la varianza explicada se vuelve menos pronunciado.

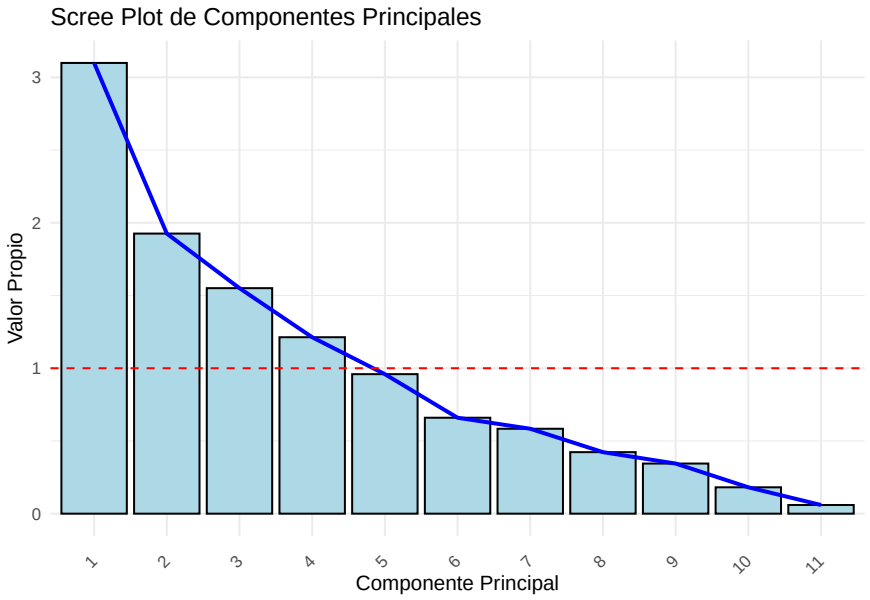

Figura 3: grafico de codos con valores propios

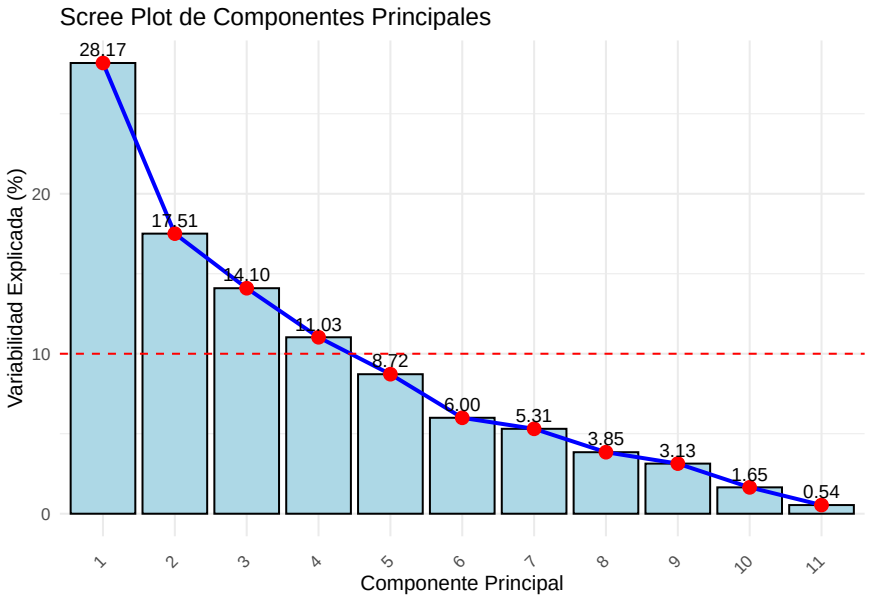

Figura 4: grafico de codos varianza explicada

Luego de observar el grafico podemos concluir que las primeras 4 componentes principales explican un aproximado del 70, en este lugar el codo del grafico indica que el numero adecuado puede ser de 4 a 5 componentes principales, teniendo encuenta la componente principal numero 5 no cumple con el hecho de que $\lambda_5 > \frac{1}{11}$ no la consideraremos como una de las componentes principales a estudiar, por lo tanto el analisis se realizara solo en base a las primeras 4 componentes principales como se habia indicado anteriormente.

8 vectores propios

Tabla 5: Vectores Propios de las Componentes Principales

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
acidezF	0.489	0.111	-0.123	0.230	-0.083	-0.101	0.350	-0.178	-0.194	0.250	-0.640
acidezV	-0.239	-0.275	-0.450	-0.079	0.219	-0.411	0.534	-0.079	0.129	-0.366	-0.002
acidezC	0.464	0.152	0.238	0.079	-0.059	-0.070	-0.105	-0.378	0.381	-0.622	0.071
azucar	0.146	-0.272	0.101	0.373	0.732	-0.049	-0.291	0.300	-0.008	-0.093	-0.184
Cl	0.212	-0.148	-0.093	-0.666	0.247	-0.304	-0.370	-0.357	-0.111	0.218	-0.053
dioxidoAL	-0.036	-0.514	0.429	0.044	-0.159	0.014	0.117	-0.205	-0.635	-0.248	0.051
totaldioxidos	0.024	-0.569	0.322	0.035	-0.222	-0.136	0.094	0.019	0.592	0.371	-0.069
densidad	0.395	-0.234	-0.339	0.174	0.157	0.391	0.170	-0.239	-0.021	0.240	0.567
ph	-0.439	-0.007	0.058	0.004	0.268	0.522	0.025	-0.561	0.168	0.011	-0.341
sulfatos	0.243	0.038	0.280	-0.551	0.226	0.381	0.447	0.375	0.058	-0.112	-0.070
alcohol	-0.113	0.386	0.472	0.122	0.351	-0.362	0.328	-0.218	-0.038	0.303	0.315

8.1 Componente Principal 1 (PC1)

La primera componente principal (PC1) captura la mayor parte de la variabilidad en los datos, con una carga significativa de variables como **acidezF** (acidez fija) y **acidezC** (acidez cítrica). Estas variables muestran altas cargas positivas, indicando que PC1 está fuertemente influenciada por la acidez en el vino. Por otro lado, **ph** tiene una carga negativa considerable, lo que sugiere una relación inversa con la acidez. Densidad también tiene una carga positiva relevante, sugiriendo que PC1 refleja características generales relacionadas con la acidez y la densidad del vino.

8.2 Componente Principal 2 (PC2)

La segunda componente principal (PC2) está dominada por cargas negativas en **dioxidoAL** (dióxido de azufre libre) y **totaldioxidos** (dióxido de azufre total). Esto sugiere que PC2 está asociada principalmente con la concentración de dióxidos de azufre en el vino, donde niveles más altos de estos compuestos tienden a reducir la puntuación en esta componente. Las cargas en **azucar** y **Cl** (cloruros) también son negativas, aunque en menor medida, lo que indica que estas variables están moderadamente relacionadas con PC2.

8.3 Componente Principal 3 (PC3)

La tercera componente principal (PC3) se caracteriza por una alta carga positiva en alcohol y dioxidoAL, junto con una carga negativa en **acidezV** (acidez volátil). Esto sugiere que PC3 podría estar relacionada con la presencia de alcohol y dióxidos de azufre libre en contraste con la acidez volátil. Las variables con cargas positivas elevadas en PC3 indican una fuerte asociación con características como el contenido de alcohol, mientras que las cargas negativas indican una relación inversa con la acidez volátil.

8.4 Componente Principal 4 (PC4)

La cuarta componente principal (PC4) tiene una carga negativa importante en **Cl** (cloruros) y **sulfatos**, junto con una carga positiva en azucar. Esto sugiere que PC4 está relacionada con una combinación de características químicas del vino, como los niveles de cloruros y sulfatos, y su contenido de azúcar. Las cargas negativas indican que mayores concentraciones de cloruros y sulfatos están asociadas con una menor puntuación en PC4, mientras que un mayor contenido de azúcar se relaciona positivamente con esta componente.

Acontinuacion mostramos un grafico donde se puede observar la contribucion de cada variables en cada uno de las 4 primeras componentes principales que estamos estudiando.

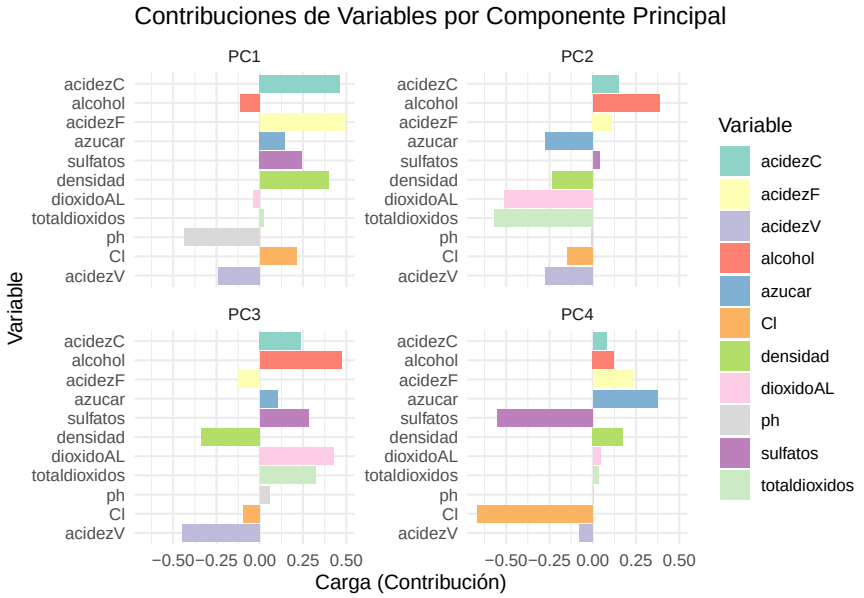

Figura 5: Contribuciones de Variables por Componente Principal

9 Circulos de correlaciones

A continuacion realizamos algunos circulos de correlacion donde podemos observar como cada variable se proyecta sobre los ejes de las componentes principales estudiadas.

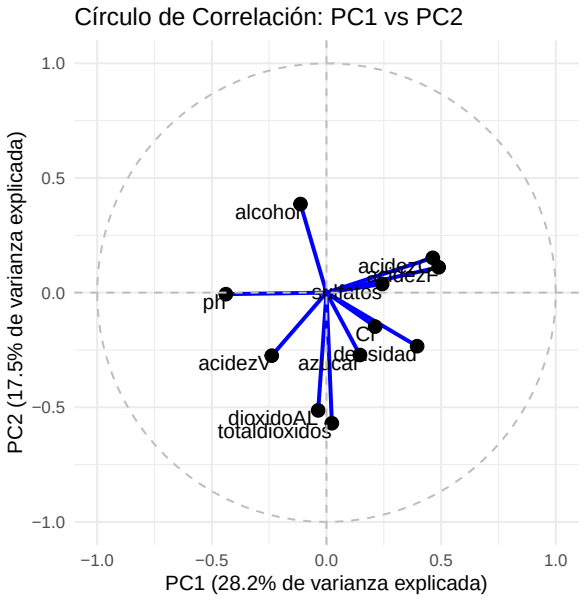

Figura 6: circulo de correlacion CP1 VS CP2

En el gráfico de círculos de correlación para las primeras dos componentes principales (PC1 y PC2), se observan patrones claros en las proyecciones de las variables:

Componente PC2:

Variables con alta proyección positiva: La variable **alcohol** muestra una fuerte proyección positiva en la PC2, indicando que está positivamente relacionada con esta componente. Esto sugiere que un aumento en el nivel de alcohol tiende a estar asociado con un incremento en la PC2. Variables con alta proyección negativa: Por otro lado, las variables **dioxidosAL**, **totaldioxidos** y **azucar** tienen una proyección negativa significativa en la PC2. Esto implica que estas variables están inversamente relacionadas con esta componente; es decir, un aumento en estas variables está asociado con una disminución en el valor de la PC2.

Componente PC1:

Variables con alta proyección: En cuanto a la PC1, se destaca la influencia de variables como **pH** y las tres variables de acidez (**acidezF**, **acidezV** y **acidezC**), que tienen una proyección destacada. Esto indica que estas variables contribuyen significativamente a la variabilidad explicada por la PC1.

Variables con baja contribución: En contraste, la variable **sulfatos** tiene una proyección muy baja en ambas componentes, lo que sugiere que no aporta de manera significativa a la variabilidad explicada por las primeras dos componentes principales.

En resumen, el círculo de correlación muestra que la variabilidad explicada por las primeras dos componentes principales está principalmente influenciada por el alcohol y las medidas de acidez y pH, mientras que los sulfatos tienen una contribución menor en este contexto.

Ahora analizaremos el circulo de correlacion para PC1 Y PC3

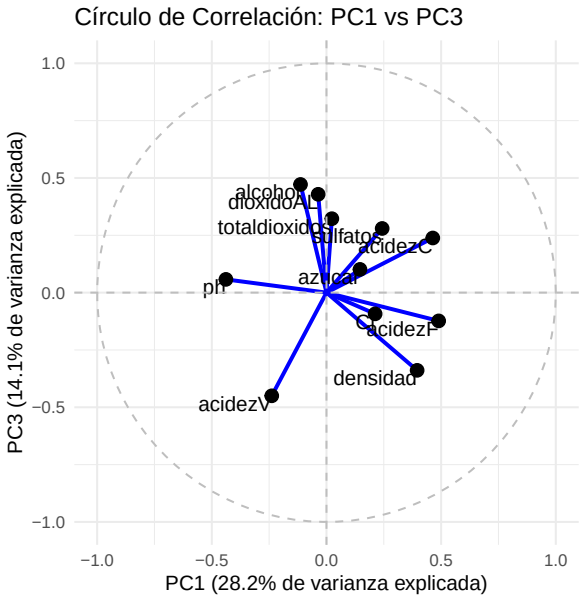

Figura 7: circulo de correlacion CP1 VS CP3

Como se mencionó anteriormente, los ácidos muestran una alta proyección en la primera componente principal (PC1). Sin embargo, al examinar la proyección en la tercera componente principal (PC3), notamos que la **acidez volátil** (acidezV) también tiene una contribución significativa. Además, las variables **alcohol**, **dióxidos de azufre libre** (dioxidosAL) y **total de dióxidos de azufre** (totaldioxidos) tienen una alta contribución en PC3.

La **densidad** muestra una contribución positiva en PC1 y una contribución negativa en PC3. En contraste, la **azúcar** es la variable con menor contribución en ambas componentes principales, tanto en PC1 como en PC3.

10 Comunalidades

Componente Principal 1 (PC1): La primera componente principal (PC1) captura la mayor parte de la varianza en el conjunto de datos y es la dirección en la cual los datos varían más. En la matriz de cargas, **acidezF** (0.489) y **acidezC** (0.464) tienen las cargas más altas en esta componente, indicando que estas variables contribuyen significativamente a la varianza explicada por PC1. En contraste, **ph** (-0.439) tiene una carga negativa considerable, lo que sugiere que esta variable está inversamente relacionada con PC1. La "comunalidad total" para cada variable en PC1 muestra cuánto de la varianza de la variable es explicada por esta componente. Variables como **acidezF** y **acidezC** con valores altos en PC1 indican que estas variables son bien representadas por esta componente, lo que es útil para interpretar la principal dirección de variabilidad en el conjunto de datos.

Componente Principal 2 (PC2): La segunda componente principal (PC2) explica la varianza adicional en los datos que no se captura en PC1. Aquí, **dioxidoAL** (-0.514) y **totaldioxidos**

Tabla 6: Cargas de las Primeras Cuatro Componentes Principales

	PC1	PC2	PC3	PC4
acidezF	0.4893142	0.1105027	-0.1233016	0.2296174
acidezV	-0.2385844	-0.2749305	-0.4499625	-0.0789598
acidezC	0.4636317	0.1517914	0.2382471	0.0794183
azucar	0.1461072	-0.2720802	0.1012834	0.3727926
Cl	0.2122466	-0.1480516	-0.0926138	-0.6661948
dioxidoAL	-0.0361575	-0.5135668	0.4287929	0.0435378
totaldioxidos	0.0235749	-0.5694870	0.3224145	0.0345771
densidad	0.3953530	-0.2335755	-0.3388714	0.1744998
ph	-0.4385196	-0.0067108	0.0576974	0.0037877
sulfatos	0.2429213	0.0375539	0.2797862	-0.5508724
alcohol	-0.1132321	0.3861810	0.4716732	0.1221811

(-0.569) tienen las cargas más altas en valor absoluto, lo que indica que estas variables tienen una fuerte relación con PC2. En cambio, **acidezF** (0.111) y **Cl** (-0.148) tienen cargas relativamente bajas en PC2, sugiriendo que su relación con esta componente es menos pronunciada. Las comunalidades altas en PC2 para algunas variables indican que PC2 captura una porción significativa de la varianza de esas variables, proporcionando una segunda dimensión importante para interpretar los datos.

Componente Principal 3 (PC3): La tercera componente principal (PC3) captura la varianza adicional en los datos que no está explicada por PC1 y PC2. En esta componente, **alcohol** (0.472) y **acidezV** (-0.450) tienen las cargas más altas, lo que indica que estas variables tienen una fuerte influencia en PC3. **ph** (0.058) y **acidezC** (0.238) tienen cargas menores, sugiriendo que su relación con esta componente es menos significativa. Las comunalidades altas en PC3 para ciertas variables muestran que PC3 añade valor a la interpretación al capturar una dimensión adicional de la varianza en los datos.

Componente Principal 4 (PC4): La cuarta componente principal (PC4) explica la varianza residual en el conjunto de datos que no se captura en las tres primeras componentes. En esta componente, **Cl** (-0.666) y **sulfatos** (-0.551) tienen las cargas más altas, indicando una fuerte influencia de estas variables en PC4. **acidezV** (-0.079) y **acidezC** (0.079) tienen cargas relativamente bajas, lo que sugiere una menor contribución de estas variables en PC4. Las comunalidades altas en PC4 para algunas variables indican que esta componente es importante para entender la varianza residual y proporciona una visión más completa de la estructura de los datos.

En nuestro análisis de componentes principales, hemos calculado la comunalidad total de las variables, que representa la proporción de la varianza de cada variable explicada por las primeras cuatro componentes principales. Los resultados muestran que la variable **Cl** tiene la mayor comunalidad total, con un valor de 0.519, indicando que una gran parte de su varianza es explicada por las cuatro primeras componentes. Por el contrario, la variable **ph** tiene la comunalidad total más baja, 0.196, lo que sugiere que las primeras cuatro componentes explican solo una fracción de su varianza total. Las variables como **acidezF** y **acidezV** tienen comunalidades totales de 0.320 y 0.341, respectivamente, indicando que estas variables están moderadamente bien representadas por las componentes principales

Tabla 7: Comunalidad Total Capturada por las Primeras Cuatro Componentes Principales

	Variable	ComunalidadTotal
acidezF	acidezF	0.320
acidezV	acidezV	0.341
acidezC	acidezC	0.301
azucar	azucar	0.245
Cl	Cl	0.519
dioxidoAL	dioxidoAL	0.451
totaldioxidos	totaldioxidos	0.430
densidad	densidad	0.356
ph	ph	0.196
sulfatos	sulfatos	0.442
alcohol	alcohol	0.399

seleccionadas. Variables como **dioxidoAL** y **sulfatos**, con valores de 0.451 y 0.442, tienen una buena representación, pero no tan alta como **Cl**. En general, los resultados sugieren que las primeras cuatro componentes principales capturan de manera efectiva una parte significativa de la varianza para la mayoría de las variables, aunque para algunas, como **ph**, se podría considerar la inclusión de más componentes para una representación más completa esto se podria desarrollar mas extensamente en un analisis mas profundo donde se consideren mas componentes de las estudiadas.

11 Graficos Biplot

para el analisis de estos graficos tendremos encuenta que los puntos grises representan las observaciones individuales (datos originales proyectados en las nuevas dimensiones), Los puntos rojos, numerados, representaN observaciones atípicas o datos que se destacan mucho, realizaremos algunos graficos donde podemos observar el peso que tiene cada una de las variables en cada uno de los componentes principales.

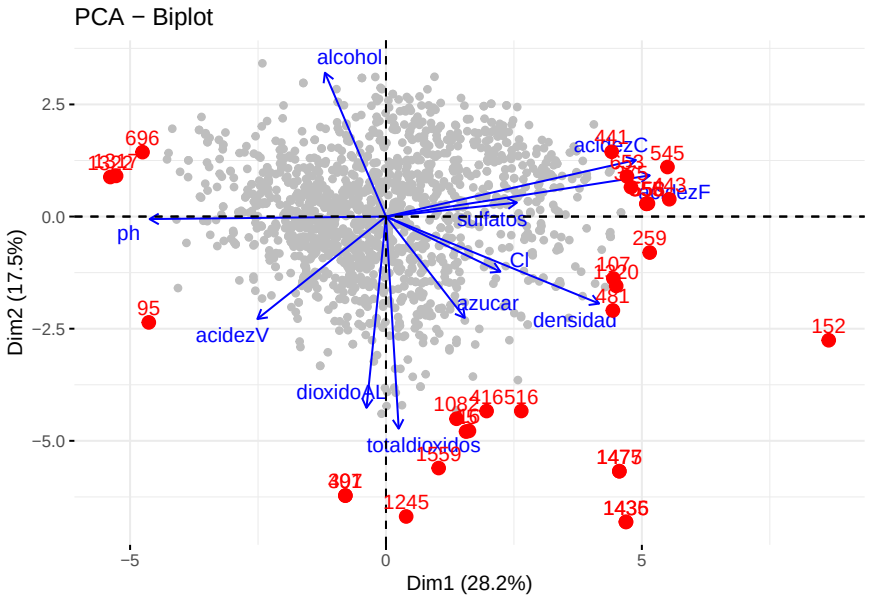

Figura 8: Biplot CP1 VS CP2

Podemos observar como **pH** esta fuertemente correlacionada con dim1, observamos que tanto *acidezF* como **acidezC** tambien estan con gran correlacion con esta componente, mientras que *alcohol*, *dioxidoAL* y *totaldioxidos* tienen una alta correlacion con la segunda componente, adicionalmente podemos observar que los datos 152, 1475 y 1436 son los datos con mayor dispercion respecto a los otros.

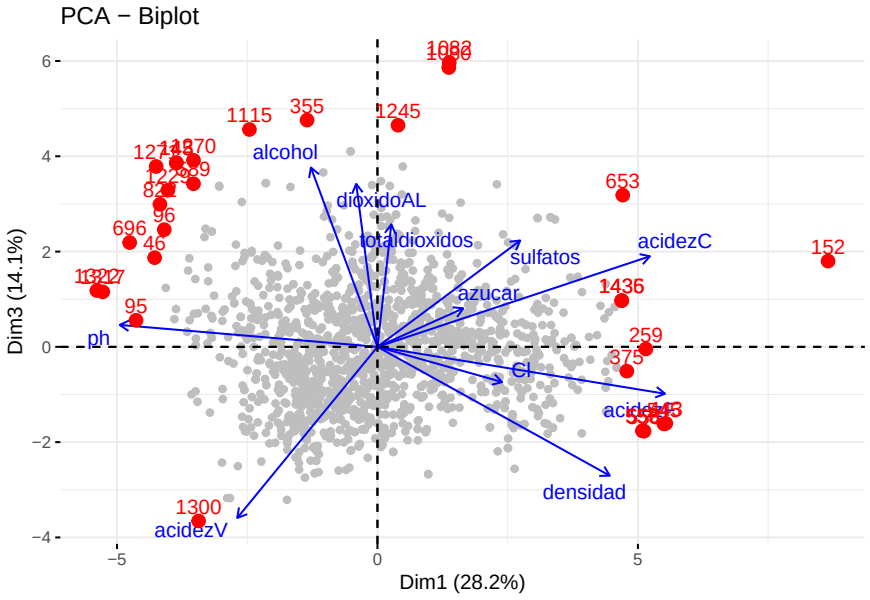

Figura 9: Biplot CP1 VS CP3

en este segundo grafico podemos ver nuevamente como el dato 152 nuevamente es un dato

con gran variabilidad, podemos notar nuevamente que *pH* esta fuertemente correlacionada con la primera componente principal, nuevamente notamos que los diferentes tipos de acidez tambien tiene una fuerte correlacion con la primera componente principal, nuevamente notamos que la componente numero 1 no captura de manera adecuada la variabilidad de variables como *alcohol*, *dioxidoAL* y *totaldioxidos* siendo estas las que estan fuertemente correlacionadas con la segunda y tercera componente.

12 Conclusiones

La primera componente principal (PC1) refleja principalmente la acidez y el pH del vino, mientras que la segunda componente principal (PC2) está relacionada con el alcohol y presenta una relación inversa con ciertos compuestos como los dióxidos de azufre y el azúcar. La tercera componente principal (PC3) destaca por la importancia de la acidez volátil y los dióxidos de azufre. La densidad y el azúcar muestran una menor influencia en las componentes principales.

En general, estos hallazgos sugieren que las propiedades del vino, como la acidez, el alcohol y los compuestos de dióxidos de azufre, juegan un papel crucial en la variabilidad del conjunto de datos, mientras que el azúcar y los sulfatos tienen una influencia menor en las primeras tres componentes principales.

Los valores de comunalidad son cruciales para entender la eficacia de nuestra reducción de dimensiones y para determinar si es necesario ajustar el número de componentes principales utilizadas en el análisis.

Referencias

Joebeachcapital. (2023). Wine Quality Dataset.

Luque-Calvo, P.L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. Disponible en http://destio.us.es/calvo.

Porras, J.C. (2016). Comparacion de pruebas de normalidad multivariada. *Anales Cientificos*, pp. 141-146. Universidad Nacional Agraria La Molina.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, **2**, 117-119.