Instituto Politécnico Nacional

Escuela Superior de Cómputo

Programación para Ciencia de Datos

Práctica 8, parcial 2

De Luna Ocampo Yanina

Galindo Durán Cristal Karina

3MA1

Fecha: 23/11/2021

Ejercicio1.

1. Descargar el archivo de ventas clientes del siguiente enlace.
2. Aplica el Análisis de Componentes Principales.
3. Realiza la interpretación de los resultados.

Procedimiento:

Descargamos los datos del link dado, lo importamos a R con las funciones vistas en clase.

```
datos <- read_csv("C:/Users/Yanina/Desktop/dataSet.csv")
head(datos)
```

y procedemos a visualizarlos para ver que se hayan importado de la forma correcta.

```
View(datos)
```

Filter

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
1	2	3	12669	9656	7561	214	2674	1338
2	2	3	7057	9810	9568	1762	3293	1776
3	2	3	6353	8808	7684	2405	3516	7844
4	1	3	13265	1196	4221	6404	507	1788
5	2	3	22615	5410	7198	3915	1777	5185
6	2	3	9413	8259	5126	666	1795	1451

Obtenemos la media, la varianza y la desviación estándar con el fin de ver que no sean un cero y poder aplicar el método pedido correspondiente, ya que, si alguna llega a ser igual a 0, no tiene sentido aplicar el método a desarrollar.

```
  # MEDIA
apply(X = datos, MARGIN = 2, FUN = mean)
  # VARIANZA
apply(X = datos, MARGIN = 2, FUN = var)
  # DESVIACION ESTANDAR
apply(X = datos, MARGIN = 2, FUN = sd)
```

La media menor pertenece a Channel y más alta pertenece a la categoría Fresh. Podemos observar que Fresh es la 12000 más grande de lo que se obtuvo en Channel. Recordamos que para sacar la media se utiliza la función “mean”.

```
>   # MEDIA
> apply(X = datos, MARGIN = 2, FUN = mean)
        Channel          Region           Fresh            Milk
       1.322727        2.543182    12000.297727     5796.265909
        Grocery          Frozen Detergents_Paper       Delicassen
    7951.277273     3071.931818     2881.493182     1524.870455
```

La varianza con una magnitud superior al resto es la de Grocery y la de menor magnitud pertenece a Fresh, contrario a su media. Recordamos que para sacar la varianza se utiliza la función “var”.

```
>   # VARIANZA
> apply(X = datos, MARGIN = 2, FUN = var)
        Channel          Region           Fresh            Milk
   2.190723e-01    5.994978e-01    1.599549e+08    5.446997e+07
        Grocery          Frozen Detergents_Paper       Delicassen
   9.031010e+07    2.356785e+07    2.273244e+07    7.952997e+06
```

La desviación estándar con una magnitud superior al resto es la de Grocery y la de menor magnitud pertenece a Fresh, contrario a su media. Recordamos que para sacar la desviación estándar se utiliza la función “sd”.

```
>   # DESVIACION ESTANDAR
> apply(X = datos, MARGIN = 2, FUN = sd)
        Channel          Region           Fresh            Milk
   4.680516e-01    7.742724e-01    1.264733e+04    7.380377e+03
        Grocery          Frozen Detergents_Paper       Delicassen
   9.503163e+03    4.854673e+03    4.767854e+03    2.820106e+03
```

La función siguiente [prcomp()] es una de las que podemos utilizar en R para poder obtener nuestro PCA.

Por defecto centran las variables para que tengan media cero. Utilizamos el prcomp como dijimos previamente para poder empezar a implementar nuestro modelo estadistico que nos ayudará a analizar los datos datos.

```
> pca <- prcomp(datos, scale = TRUE)
> names(pca)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

Mandamos llamar individualmente los subgrupos de arriba. Con center y scale obtenemos la media y la desviación típica de las variables, en cuanto al rotation, podemos ver que analiza el vaor de los loading de cada eigenvector.

```
> pca$center
        Channel          Region           Fresh            Milk
       1.322727        2.543182    12000.297727     5796.265909
        Grocery          Frozen Detergents_Paper       Delicassen
    7951.277273     3071.931818     2881.493182     1524.870455
> pca$scale
        Channel          Region           Fresh            Milk
   4.680516e-01    7.742724e-01    1.264733e+04    7.380377e+03
        Grocery          Frozen Detergents_Paper       Delicassen
   9.503163e+03    4.854673e+03    4.767854e+03    2.820106e+03
> pca$rotation
                       PC1         PC2           PC3         PC4         PC5
Channel          -0.42829156  0.20469886  0.0829798863 -0.02964416  0.03620585
Region           -0.02472603 -0.04312964  0.9825008891 -0.07784462 -0.13250892
Fresh             0.02531946 -0.51344468  0.0889509074  0.79847592  0.25811686
Milk             -0.47440995 -0.20554061 -0.0257510842 -0.05402202  0.07208576
Grocery          -0.53632914  0.00871762 -0.0453143572  0.12158624 -0.11172990
Frozen            0.02997456 -0.59274525 -0.1221565222 -0.16131688 -0.75421244
Detergents_Paper -0.52390630  0.12108309 -0.0474814388  0.15101211 -0.17650264
Delicassen       -0.16499653 -0.53318082  0.0009301994 -0.53755767  0.54482721
                       PC6          PC7          PC8
Channel          -0.86350670  0.139899044  0.019335373
Region            0.08976479 -0.023279938 -0.001545045
Fresh            -0.14747474 -0.027173693 -0.033851114
Milk              0.31593256  0.789020414 -0.039291347
Grocery           0.21369889 -0.353064294  0.715984124
Frozen           -0.19435993 -0.005336793 -0.012983225
Detergents_Paper  0.19575356 -0.371374310 -0.691672189
Delicassen       -0.05453289 -0.306582655 -0.075642587
```

Asimismo, calcula autimáticamente el valor de los componentes principales para cada observación, multiplicando los datos por los vectores de loadings. El resultado se almacena en la matriz x.

```
> head(pca$x)
            PC1        PC2       PC3         PC4        PC5        PC6        PC7
[1,] -0.8429794  0.5147648 0.7667594  0.04416453  0.4457267 -0.9383731  0.6540173
[2,] -1.0614682  0.4840503 0.6722101 -0.40091542  0.1303098 -0.8662408  0.5104414
[3,] -1.2676975 -0.6812791 0.6633395 -1.63309380  1.1924556 -1.0772155 -0.2029209
[4,]  1.0555808 -0.6101270 0.5050795 -0.19578209 -0.4573340  0.1168258 -0.3134471
[5,] -0.6333096 -0.9730912 0.7703319 -0.18616222  0.8129520 -1.5036608 -0.1602166
[6,] -0.5295082  0.5847458 0.7577620 -0.24672790  0.3784085 -1.0713344  0.6578136
```

```
             PC8
[1,]  0.01808111
[2,]  0.07780623
[3,] -0.25374856
[4,]  0.05431530
[5,]  0.00375915
[6,] -0.02594706
```

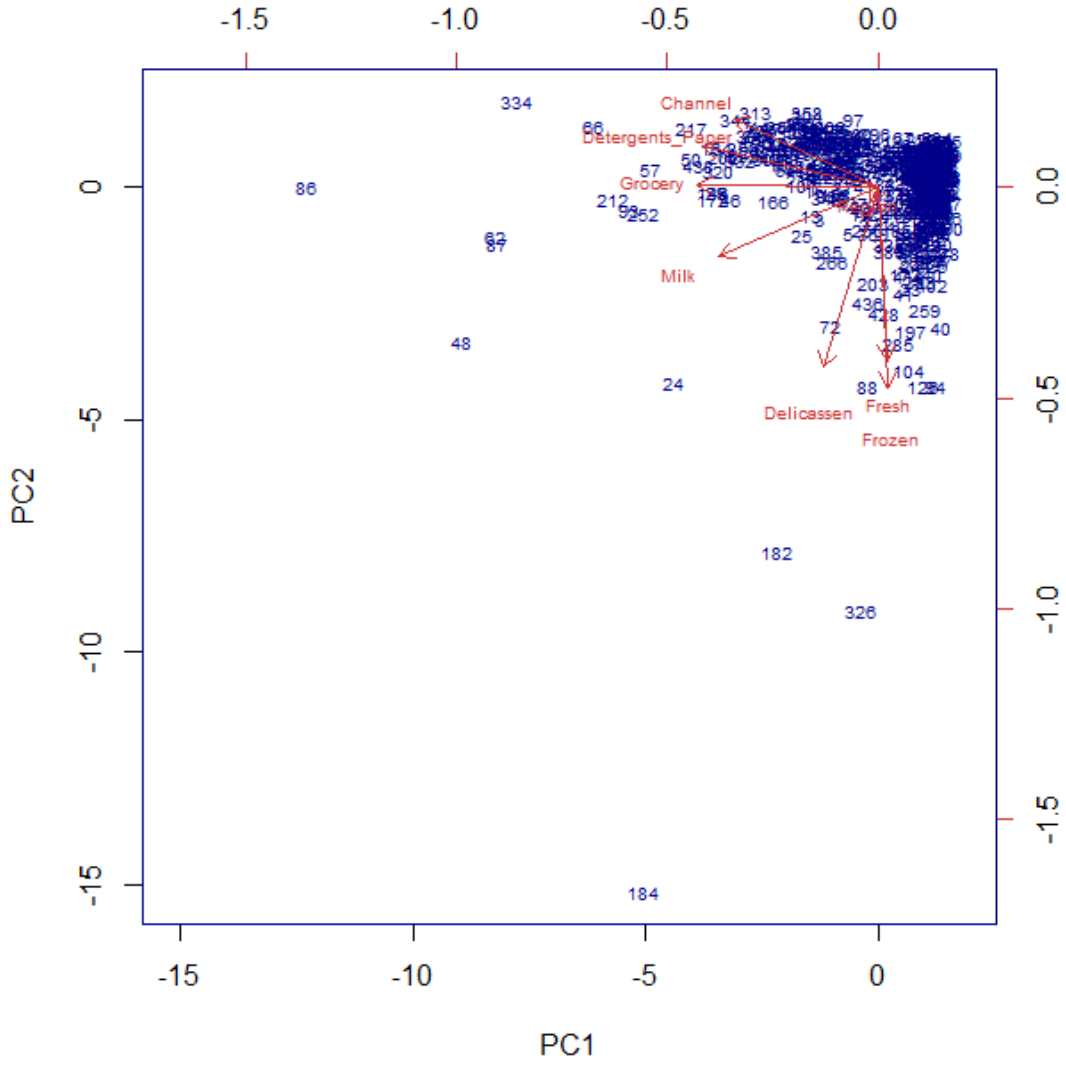
Obtenemos dim de lo que acabamos de sacar.

```
> dim(pca$x)
[1] 440   8
```

Resultado:

Obtenemos como gráfica final la siguiente:

```
biplot(x = pca, scale = 0, cex = 0.6, col = c("blue4", "brown3"))
```



Ejercicio2.

1. Descargar el archivo de tu interés.
2. Aplica el Análisis de Componentes Principales.
3. Realiza la interpretación de los resultados.

Procedimiento:

Descargamos los datos del dataset de nuestro interés, lo importamos a R con las funciones vistas en clase.

```
vino <- read_csv("C:/Users/Yanina/Desktop/wine.csv")
head(vino)
```

y procedemos a visualizarlos para ver que se hayan importado de la forma correcta.

```
View(datos)
```

Wine	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	
1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	
1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	
1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	
1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	
1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	
1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	
1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	
1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	
1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	

Obtenemos la media, la varianza y la desviación estándar con el fin de ver que no sean un cero y poder aplicar el método pedido correspondiente, ya que, si alguna llega a ser igual a 0, no tiene sentido aplicar el método a desarrollar.

```
  # MEDIA
apply(X = vino, MARGIN = 2, FUN = mean)
  # VARIANZA
apply(X = vino, MARGIN = 2, FUN = var)
  # DESVIACION ESTANDAR
apply(X = vino, MARGIN = 2, FUN = sd)
```

La media menor pertenece a Nonflavanoid.phenols y más alta pertenece a la categoría Proline. Podemos observar que Proline es al menos 2,068 veces más grande de lo que se obtuvo en Nonflavanoid.phenols. Recordamos que para sacar la media se utiliza la función “mean”.

```
>   # MEDIA
> apply(X = vino, MARGIN = 2, FUN = mean)
                Wine              Alcohol            Malic.acid
           1.9382022           13.0006180            2.3363483
                 Ash                  Acl                   Mg
           2.3665169           19.4949438           99.7415730
             Phenols            Flavanoids Nonflavanoid.phenols
           2.2951124            2.0292697            0.3618539
             Proanth            Color.int                  Hue
           1.5908989            5.0580899            0.9574494
                  OD              Proline
           2.6116854          746.8932584
```

La varianza con una magnitud superior al resto es la de Flavanoids y la de menor magnitud pertenece a Acl. Recordamos que para sacar la varianza se utiliza la función “var”.

```
>   # VARIANZA
> apply(X = vino, MARGIN = 2, FUN = var)
                Wine              Alcohol            Malic.acid
        6.006792e-01         6.590623e-01         1.248015e+00
                 Ash                  Acl                   Mg
        7.526464e-02         1.115269e+01         2.039893e+02
             Phenols            Flavanoids Nonflavanoid.phenols
        3.916895e-01         9.977187e-01         1.548863e-02
             Proanth            Color.int                  Hue
        3.275947e-01         5.374449e+00         5.224496e-02
                  OD              Proline
        5.040864e-01         9.916672e+04
```

La desviación estándar con una magnitud superior al resto es la de Proline y la de menor magnitud pertenece a Nonflavanoid.phenols, al igual que su media. Recordamos que para sacar la desviación estándar se utiliza la función “sd”.

```
>   # DESVIACION ESTANDAR
> apply(X = vino, MARGIN = 2, FUN = sd)
                Wine              Alcohol            Malic.acid
           0.7750350            0.8118265            1.1171461
                 Ash                  Acl                   Mg
           0.2743440            3.3395638           14.2824835
             Phenols            Flavanoids Nonflavanoid.phenols
           0.6258510            0.9988587            0.1244533
             Proanth            Color.int                  Hue
           0.5723589            2.3182859            0.2285716
                  OD              Proline
           0.7099904          314.9074743
```

La función siguiente [prcomp()] es una de las que podemos utilizar en R para poder obtener nuestro PCA.

Por defecto centran las variables para que tengan media cero. Utilizamos el prcomp como dijimos previamente para poder empezar a implementar nuestro modelo estadistico que nos ayudará a analizar los datos datos.

```
> pca <- prcomp(vino, scale = TRUE)
> names(pca)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

Mandamos llamar individualmente los subgrupos de arriba. Con center y scale obtenemos la media y la desviación típica de las variables, en cuanto al rotation, podemos ver que analiza el vaor de los loading de cada eigenvector.

```
> pca$center
          Wine        Alcohol      Malic.acid
     1.9382022     13.0006180      2.3363483
           Ash            Acl             Mg
     2.3665169     19.4949438     99.7415730
       Phenols      Flavanoids Nonflavanoid.phenols
     2.2951124      2.0292697      0.3618539
       Proanth      Color.int             Hue
     1.5908989      5.0580899      0.9574494
            OD        Proline
     2.6116854    746.8932584
> pca$scale
          Wine        Alcohol      Malic.acid
     0.7750350      0.8118265      1.1171461
           Ash            Acl             Mg
     0.2743440      3.3395638     14.2824835
       Phenols      Flavanoids Nonflavanoid.phenols
     0.6258510      0.9988587      0.1244533
       Proanth      Color.int             Hue
     0.5723589      2.3182859      0.2285716
            OD        Proline
     0.7099904    314.9074743
```

```
> pca$rotation
                       PC1         PC2         PC3         PC4
Wine                0.393669533 -0.005690412  0.001217953 -0.12246373
Alcohol            -0.136325011 -0.484160868 -0.207400812  0.08191848
Malic.acid          0.222676383 -0.223590947  0.088796064 -0.46988824
Ash                -0.002257932 -0.315855884  0.626102363  0.24984122
Acl                 0.224298489  0.011615737  0.611989600 -0.07199322
Mg                 -0.124630159 -0.300551432  0.130984580  0.16321412
Phenols            -0.359264042 -0.067119829  0.146507749 -0.19098521
Flavanoids         -0.390711715  0.001313454  0.150962746 -0.14461667
Nonflavanoid.phenols  0.267001203 -0.026988703  0.169975512  0.32801272
Proanth            -0.279062504 -0.041222563  0.149879586 -0.46275771
Color.int           0.089318293 -0.529782740 -0.137266298 -0.07211248
Hue                -0.276822650  0.277907354  0.085328539  0.43466618
OD                 -0.350526181  0.162776250  0.166204360 -0.15672341
Proline            -0.269515252 -0.366058862 -0.126686846  0.25579490
```

	PC5	PC6	PC7	PC8	PC9
Wine	0.15758395	-0.20033864	0.05938234	-0.07179553	-0.162368819
Alcohol	-0.25089415	0.13517139	0.09269887	-0.42154435	-0.450190708
Malic.acid	-0.18860015	0.59841948	-0.37436980	-0.08757556	-0.006025687
Ash	-0.09352360	0.10799983	0.16708856	0.17208034	0.262494455
Acl	0.04656750	-0.08811224	0.26872469	-0.41324857	-0.118633417
Mg	0.77833048	0.14483831	-0.32957951	0.14881189	-0.252536278
Phenols	-0.14466563	-0.14809748	0.03789829	0.36343884	-0.406373544
Flavanoids	-0.11200553	-0.06247252	0.06773223	0.17540500	-0.090919334
Nonflavanoid.phenols	-0.43257916	-0.25868639	-0.61111195	0.23075135	-0.159122818
Proanth	0.09158820	-0.46627764	-0.42292282	-0.34373920	0.265786794
Color.int	-0.04626960	-0.42525454	0.18613617	0.04069617	-0.075264592
Hue	-0.02986657	0.01565089	-0.19204101	-0.48362564	-0.212416815
OD	-0.14419358	0.21770365	0.07850980	0.06865116	-0.084264837
Proline	-0.08440794	0.06656550	-0.05420370	-0.11146671	0.544905394

	PC10	PC11	PC12	PC13	PC14
Wine	0.19899373	-0.01444169	0.01575769	-0.49224318	-0.669045280
Alcohol	-0.31127983	0.22154641	-0.26411262	-0.05610645	-0.090626055
Malic.acid	0.32592413	-0.06839251	0.11921210	0.06675544	0.025225306
Ash	0.12452347	0.49452428	-0.04502305	-0.19201787	0.001635816
Acl	-0.15716811	-0.47461722	-0.06131271	0.20007784	0.095361066
Mg	-0.12773363	-0.07119731	0.06116074	0.05829909	-0.022300745
Phenols	0.30772263	-0.29740957	-0.30087591	-0.35952714	0.253037788
Flavanoids	0.14044000	0.03219187	-0.05001396	0.59834288	-0.601909165
Nonflavanoid.phenols	-0.24054263	-0.12200984	0.04266558	0.06403952	-0.082230935
Proanth	-0.10869629	0.23292405	-0.09334264	-0.11013538	0.058641979
Color.int	0.21704255	-0.01972448	0.59795428	0.15917751	0.178821145
Hue	0.50966073	0.06140493	0.25774292	-0.04923091	0.022582562
OD	-0.45570504	-0.06646166	0.61109218	-0.32941979	-0.135092159
Proline	0.04620802	-0.55130818	-0.07268036	-0.17322892	-0.216043617

Asimismo, calcula autimáticamente el valor de los componentes principales para cada observación, multiplicando los datos por los vectores de loadings. El resultado se almacena en la matriz x.

```
> head(pca$x)
          PC1        PC2        PC3         PC4        PC5        PC6
[1,] -3.513024 -1.4490110 -0.1643319  0.01323549  0.7352712  0.2998703
[2,] -2.521745  0.3290909 -2.0210056  0.41597096 -0.2824171  0.8818219
[3,] -2.777195 -1.0340191  0.9804719 -0.66236396 -0.3864748 -0.4675228
[4,] -3.911554 -2.7604234 -0.1744760 -0.56349826 -0.3234473 -0.2618777
[5,] -1.403552 -0.8653321  2.0201309  0.43966556  0.2273080  0.5920919
[6,] -3.278880 -2.1241831 -0.6272230  0.60366902 -0.4084742 -0.2575494
           PC7         PC8         PC9         PC10       PC11         PC12
[1,] -0.57226129  0.05548077 -0.45747458 -1.06257216  0.4193111  0.551372411
[2,]  0.02963289  1.00752977  0.21819060  0.02012526  0.1296539  0.393860128
[3,] -0.48693182 -0.26820049  1.21932990  0.10595025  0.2782891  0.001892654
[4,]  0.39724354  0.61710292 -0.11433580 -0.10736825 -0.7716890 -0.230279641
[5,] -0.44661570  0.43371385 -0.26081202 -0.11211805  0.5364180 -0.226048523
[6,] -0.37751103  0.36579361  0.04478595  0.21735308 -0.4066383 -0.375655765
          PC13        PC14
[1,] -0.30212591 -0.20029245
[2,] -0.14623280 -0.12604679
[3,]  0.02121816  0.05559469
[4,] -0.49986674 -0.01984687
[5,]  0.27333786  0.51604220
[6,] -0.01739357 -0.23888844
```

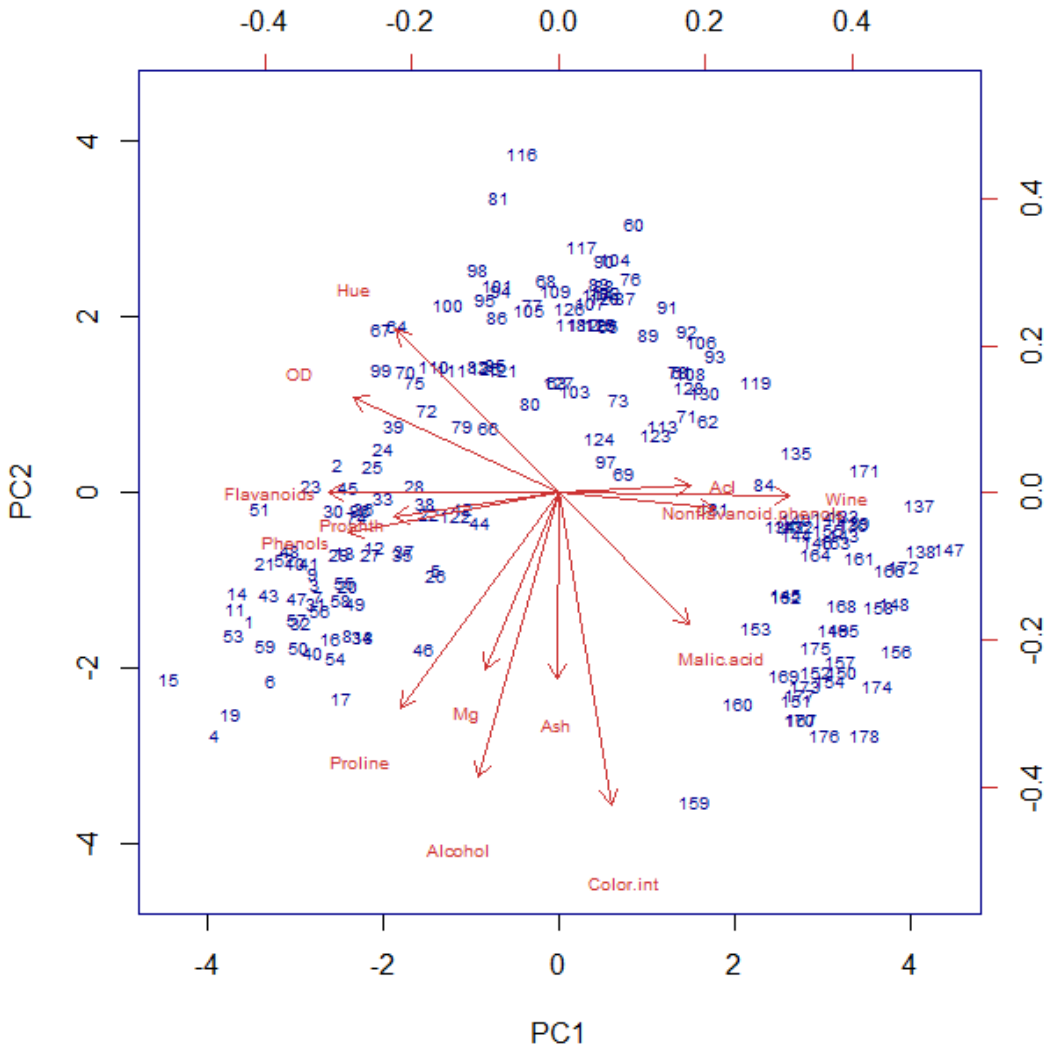
Obtenemos dim de lo que acabamos de sacar.

```
> dim(pca$x)
[1] 178  14
```

Resultado:

Obtenemos como gráfica final la siguiente:

```
biplot(x = pca, scale = 0, cex = 0.6, col = c("blue4", "brown3"))
```



¿Qué aprendí con esta práctica?

Reforcé la importación de datasets en R. Aprendí a implementar un nuevo método estadístico que nos ayuda a simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.

Nos permite condensar la información apartada por múltiples variables en pocos componentes. Esto lo hace un método muy útil de aplicar previa utilización de otras técnicas estadísticas.

Información de los DataSets:

Data1, ventas:

Abstract: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories

Data Set Characteristics:	Multivariate	**Number of Instances:**	440	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	8	**Date Donated**	2014-03-31
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	411768

Attribute Information:

1) FRESH: annual spending (m.u.) on fresh products (Continuous);
2) MILK: annual spending (m.u.) on milk products (Continuous);
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)
5) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
7) CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)
8) REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)
Descriptive Statistics:

Data2, wine:

Fuente: https://archive.ics.uci.edu/ml/datasets/wine

Data Set Characteristics:	Multivariate	**Number of Instances:**	178	**Area:**	Physical
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	13	**Date Donated**	1991-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1792988

Data Set Information:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a list of what the 30 or so variables were, but a.) I lost it, and b.), I would not know which 13 variables are included in the set.

The attributes are (dontated by Riccardo Leardi, riclea '@' anchem.unige.it)
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280/OD315 of diluted wines
13)Proline

In a classification context, this is a well posed problem with "well behaved" class structures. A good data set for first testing of a new classifier, but not very challenging.