

# P2: análisis del voto

Valentina Díaz Torres

## Introducción

El objetivo principal de esta práctica es analizar la base de datos de votos, que recoge 5 variables categóricas de los votantes de los diferentes partidos. Estas son trabaja, doméstico, parado, jubilado y estudiante, por lo que analizan la situación del votante. Por otro lado, los partidos recogidos son PP, PSOE, Unidas Podemos, Ciudadanos y una tercera columna que recoge otros partidos en general.

Se pretende por tanto, conocer cuál es la relación entre el voto a los diferentes partidos, en función de las variables nombradas anteriormente, es decir de su situación laboral. No obstante, hay que tener claro que el objetivo principal es una reducción de las dimensiones. Para ellos, mediante diferentes tipos de análisis se medirán estas relaciones y se intentará reducir las dimensiones para mayor nivel explicativo.

En este caso, no aplicaría realizar análisis de valores nulos, missing values, entre otros, ya que la base de datos es pequeña y se puede observar con claridad. No obstante sería conveniente hacer una exploración previa sobre cómo se distribuyen las variables entre ellas. Además, en esta práctica, no son solo relevantes las columnas, sino que también las filas, cómo se distribuyen filas y columnas en un mapa dos dimensiones y cada una por separado. Es por ello, que por un lado se analizarán filas, por otro columnas y después quedarán representadas y estudiadas las relaciones.


























## Carga y trata de los datos

La matriz de los datos, formando una tabla de contingencia quedaría así:

##	PP	PSOE	UP	Cs	Resto
## Trabaja	462	441	471	576	369
## Doméstico	502	857	83	606	274
## Parado	383	544	551	616	230
## Estudiante	316	376	388	478	762
## Jubilado	846	639	172	499	169

El siguiente gráfico muestra la representación de la tabla anteriormente creada. Como se puede observar, por un lado se encuentra el estado y por otro el partido, haciéndose el punto más grande en aquellos casos en los que hay más votos y más pequeño en los que menos. Se podría destacar que al PP lo votan más los jubilados, al PSOE el estado doméstico, UP parado, Cs queda bastante repartido y el resto más por estudiantes.

## Partidos

		Partido				
		PP	PSOE	UP	Cs	Resto
Estado	Trabaja					
	Doméstico					
	Parado					
	Estudiante					
	Jubilado					

## Contraste de independencia de Chi-cuadrado

Esta prueba se realiza al principio del estudio con el fin de testear la independencia de las variables, para así hacer un análisis que tenga sentido. Basándonos en la p-value obtenida, muy próxima a 0, podríamos decir que la hipótesis de independencia del test de chi-cuadrado se rechaza. Por tanto, existiría relación entre las distintas variables de filas y columnas. Además, los grados de libertad son 16, esta cifra es el resultado del número de filas - 1 por el número de columnas - 1.

```
##
## Pearson's Chi-squared test
##
## data:  partidos
## X-squared = 1704.3, df = 16, p-value < 2.2e-16

## [1] 16
```

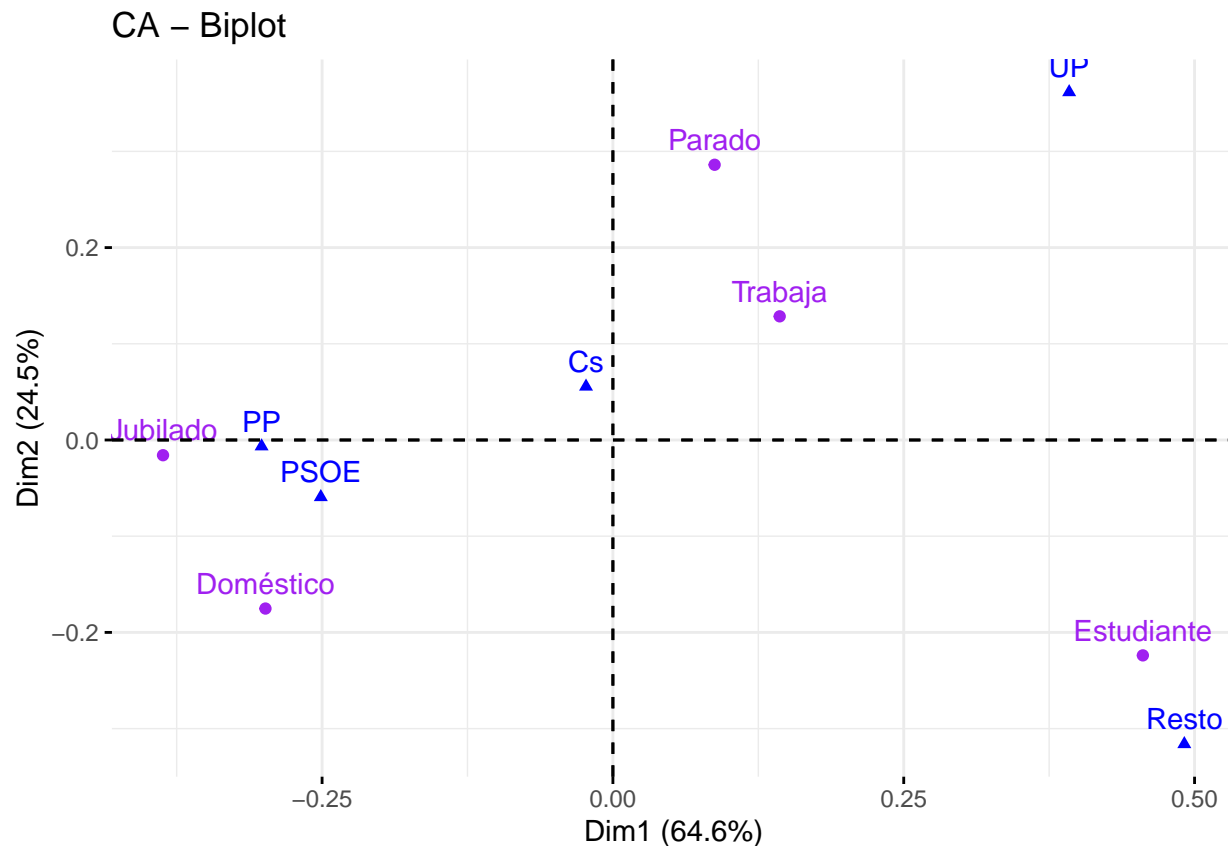
## Análisis de correspondencias

Una vez que se ha demostrado que existe relación entre las variables, se pretende estudiar cuál es esta relación, tanto por separado, como juntas, en un mapa de dos dimensiones.

Mediante la función CA se realiza un análisis de correspondencias entre las filas y columnas. Para ello se tienen en cuenta la tabla de datos partidos, y el número de componentes principales, que es 5.

El gráfico anterior muestra la asociación entre las filas y las columnas, por ejemplo estudiantes y el resto, doméstico y PSOE, son casos claros, donde se encuentran muy cerca en la representación. Dividiendo en dos dimensiones. La primera recogería el 64,65% y la segunda el 24,25%, por lo que entre ambas se recogería un total del 89,91%.

Del análisis de correspondencias se obtiene la siguiente información:



Como la representación mostraba, en cuanto a la elección del número de dimensiones, no tendría sentido añadir una tercera dimensión. Con dos dimensiones ya queda representado el 89,10% y añadiendo una tercer aumentaría a 99,89 %, no obstante, pierde esta tercera dimensión en calidad de representación, basándonos en Cos2. Es por eso, que se ha decidido seguir adelante con tan solo 2 dimensiones. Esta tercera, sería necesario añadirla en el caso de que las dos anteriores no explicasen suficiente, pero con un 89,10% se ha considerado suficiente. También, cabe destacar, que donde más sentido quizá tendría aplicar esta tercera dimensión es en la explicación de los votos de ciudadanos, ya que el valor de cos " es de 0,603.

```
##
## Call:
## CA(X = partidos, graph = FALSE)
##
## The chi square of independence between the two variables is equal to 1704.298 (p-value = 0 ).
##
## Eigenvalues
##
```

	Dim.1	Dim.2	Dim.3	Dim.4
Variance	0.09	0.04	0.02	0.00
% of var.	64.65	24.45	10.79	0.11
Cumulative % of var.	64.65	89.10	99.89	100.00

```
##
```

```
## Rows
##          Iner*1000  Dim.1  ctr  cos2  Dim.2  ctr  cos2
## Trabaja   |      7.86 |  0.14  4.33  0.52 |  0.13  9.20  0.42 |
## Doméstico |     30.66 | -0.30 18.80  0.58 | -0.18 17.11  0.20 |
## Parado    |     19.47 |  0.09  1.61  0.08 |  0.29 45.65  0.84 |
## Estudiante |     51.85 |  0.46 43.70  0.80 | -0.22 27.90  0.19 |
## Jubilado  |     36.95 | -0.39 31.56  0.81 | -0.02  0.14  0.00 |
##
## Columns
##          Iner*1000  Dim.1  ctr  cos2  Dim.2  ctr  cos2
## PP         |     28.84 | -0.30 20.77  0.68 | -0.01  0.03  0.00 |
## PSOE       |     21.35 | -0.25 16.35  0.73 | -0.06  2.41  0.04 |
## UP         |     40.90 |  0.39 23.25  0.54 |  0.36 52.25  0.46 |
## Cs         |      2.45 | -0.02  0.13  0.05 |  0.06  2.05  0.30 |
## Resto      |     53.25 |  0.49 39.51  0.70 | -0.32 43.26  0.29 |
```

En cuanto a los autovalores, estos miden si hay una asociación entre filas y columnas. El resultado obtenido ha sido del 0.383, lo cual se considera como que existe una asociación. El umbral usado es del 0.2, por lo que un 0.38 representa una asociación, aunque no excesivamente fuerte.

```
## [1] 0.3831393
```

También se puede calcular el estadístico chi-cuadrado, para testear la hipótesis de independencia y si por tanto, tiene sentido o no seguir con el análisis. El resultado en este caso sería de 1704.298, con un p-value de 0, lo que nos permitiría seguir con el estudio y rechazar la hipótesis de independencia.

```
## [1] 1704.298
```

```
## [1] 0
```

## Autovalores y gráficos de sedimentación

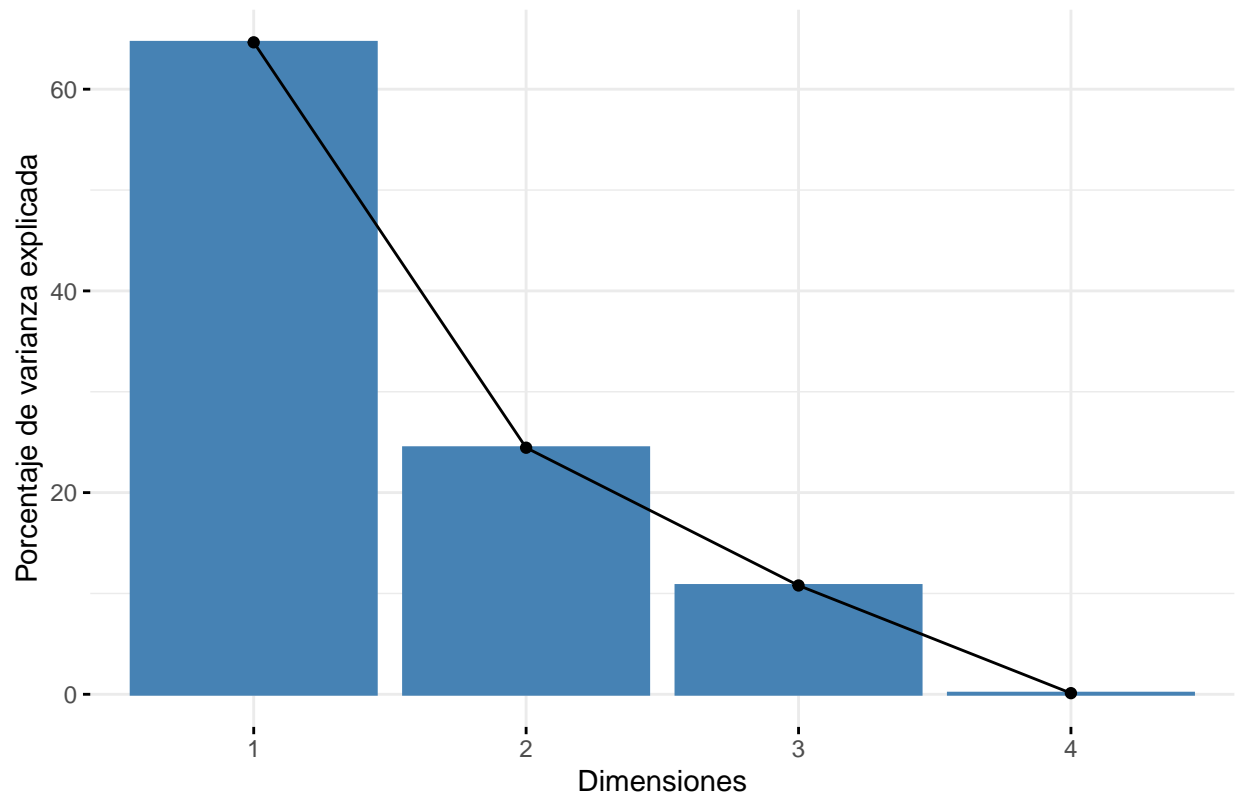
Sería necesario la realización de un examen de autovalores, con el fin de asegurarnos de que las dimensiones elegidas son las apropiadas.

Por ello, en primer lugar, se estudia la varianza explicada por cada una de las dimensiones. En este análisis se observa que el 100% de la variabilidad no se consigue hasta tener en cuenta las 4 dimensiones. No obstante, como ya se ha comentado anteriormente, con la segunda dimensión se obtiene el 89,10%, teniendo, la primera dimensión, un 64.5% y la segunda el 24.45%. La tercera y cuarta dimensión son el 10.79% de la varianza y el 0.11%, por lo que representarían una parte más pequeña de esta.

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1         0.09          64.65          64.65
## Dim.2         0.04          24.45          89.10
## Dim.3         0.02          10.79          99.89
## Dim.4         0.00           0.11         100.00
```

Lo que se ha comentado anteriormente queda representado en la siguiente gráfica. Aquí se entiende, de una forma visual la importancia que tienen cada una de las dimensiones, siendo la 4 casi insignificante.

Gráfico de sedimentación



## Contribución de filas y columnas

El fin de este apartado es conocer qué filas y columnas explican más las dos dimensiones elegidas.

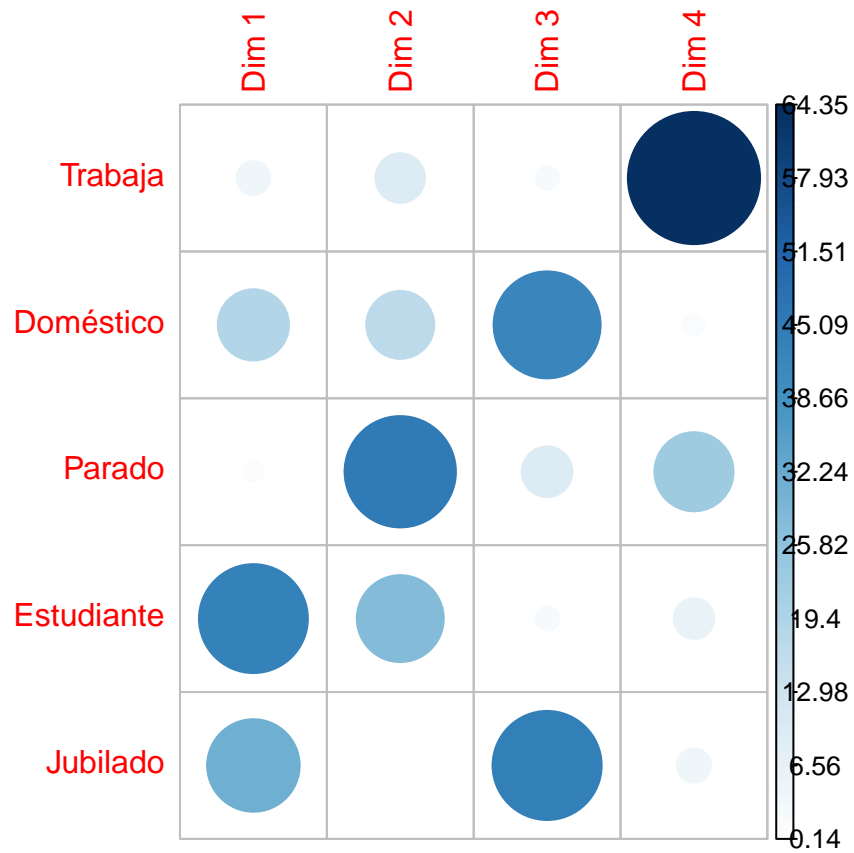
```
## Correspondence Analysis - Results for rows
## =====
##   Name      Description
## 1 "$coord"   "Coordinates for the rows"
## 2 "$cos2"    "Cos2 for the rows"
## 3 "$contrib" "contributions of the rows"
## 4 "$inertia" "Inertia of the rows"
```

A continuación se muestra la contribución de cada fila con las dimensiones y la calidad de la representación (cos2) de cada una de ellas. De la dimensión 1, la que mayor calidad tiene es jubilado, seguido de estudiante y de la 2 parado, seguido de trabaja. Se podría concluir por tanto, que teniendo en cuenta la calidad de la representación estas son las filas más representativas o que más explican estas dos dimensiones. En cuanto a la contribución, se afirman lo obtenido en la calidad de la representación.

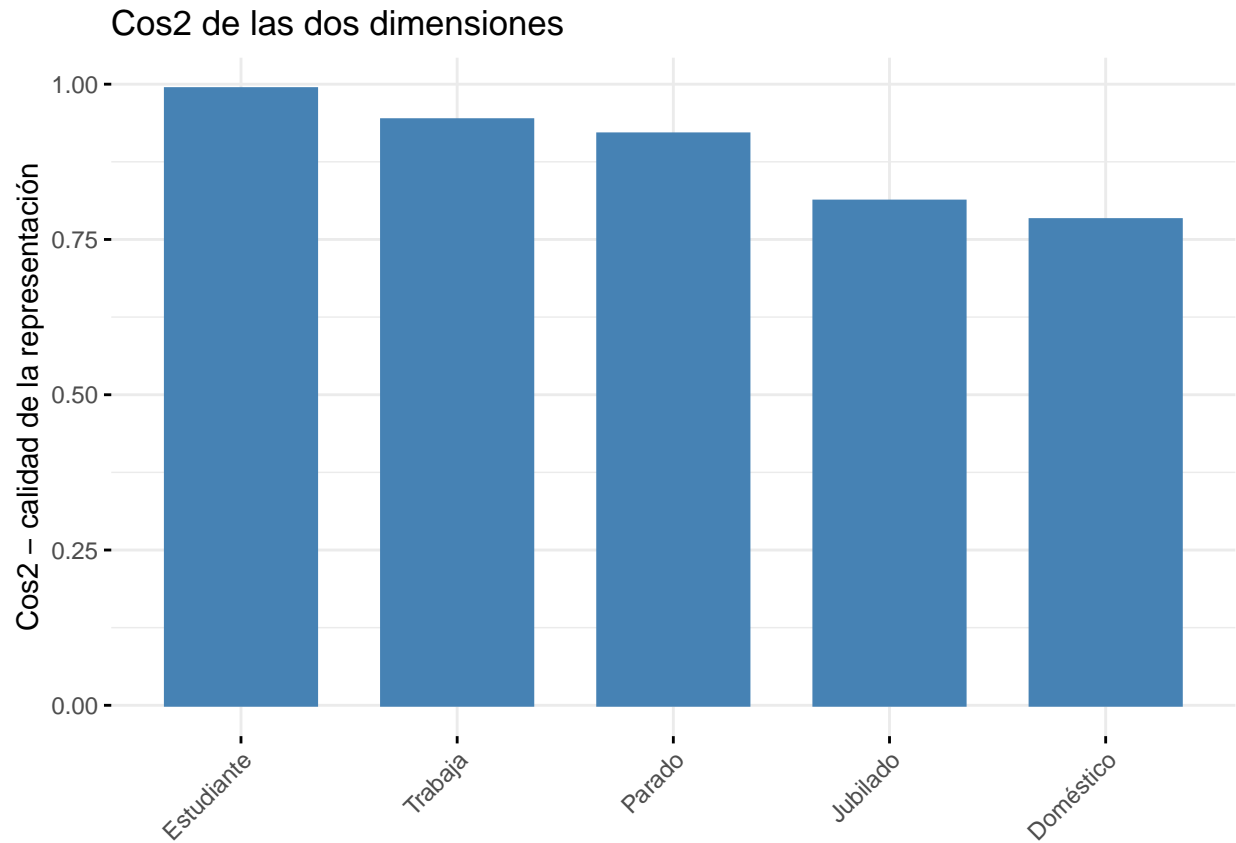
	Dim 1	Dim 2	Dim 3	Dim 4
Trabaja	0.52272535	0.42015307	0.043457197	0.0136643843
Doméstico	0.58178728	0.20023314	0.217875266	0.0001043068
Parado	0.07858872	0.84148021	0.077949469	0.0019816103
Estudiante	0.79978954	0.19314220	0.006870034	0.0001982266
Jubilado	0.81051753	0.00135683	0.187925250	0.0002003867

##	Dim 1	Dim 2	Dim 3	Dim 4
## Trabaja	4.326948	9.1954695	2.155071	64.348351
## Doméstico	18.799106	17.1067366	42.176704	1.917453
## Parado	1.612647	45.6542607	9.582616	23.133250
## Estudiante	43.702314	27.9038500	2.248948	6.162114
## Jubilado	31.558984	0.1396832	43.836661	4.438832

Para ver de forma más gráfica lo comentado anteriormente, podemos observar este gráfico, donde se representan gráficamente las contribuciones de cada dimensión.



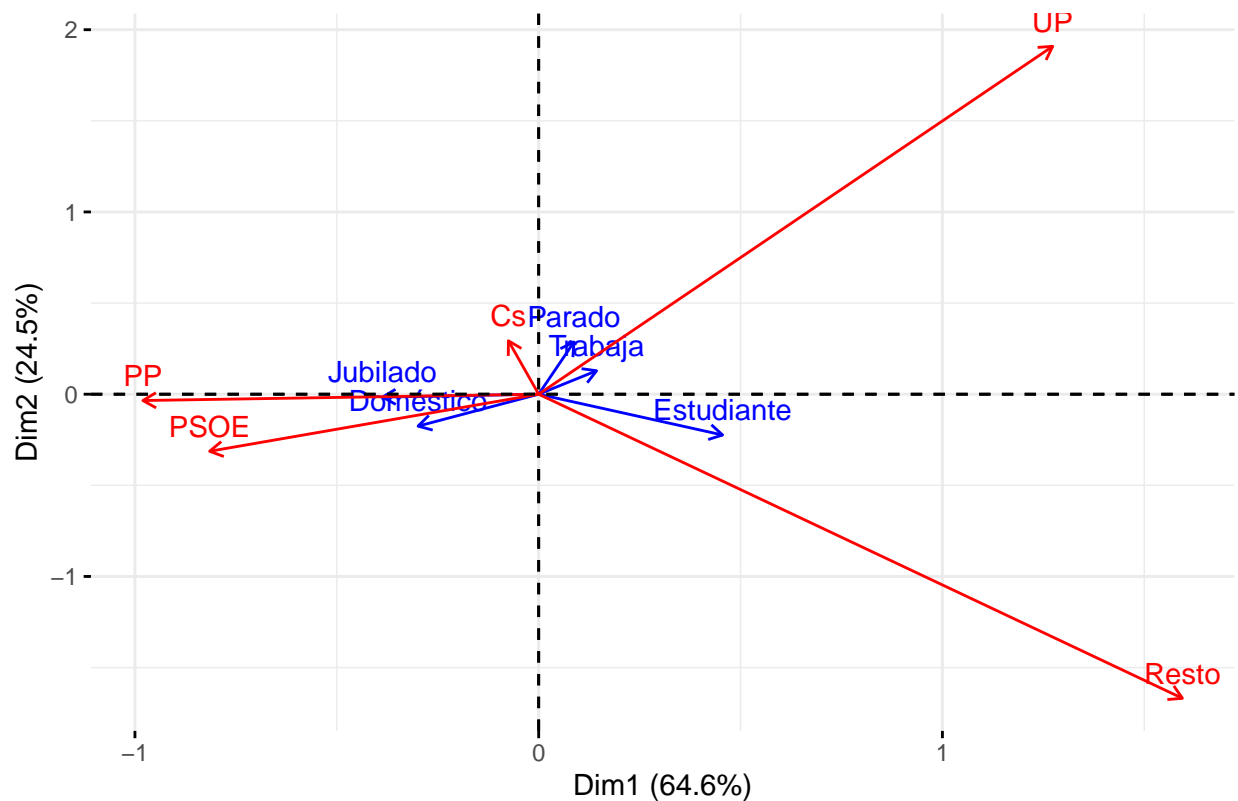
En este gráfico se muestra la calidad de la representación de cada una de las filas, donde se aprecia que estudiante tiene la mayor calidad, seguido de trabaja.



## Representación conjunta de filas y columnas

La siguiente representación, consiste en un gráfico asimétrico de filas y columnas juntas, donde a mayor ángulo, mayor desasociación y mientras más cerrado es este ángulo mayor asociación hay.

### Análisis de correspondencias simples. Gráfico asimétrico.



### Conclusión

Como conclusión se ha decidido reducir las dimensiones a 2, teniendo en cuenta la calidad de la representación y la contribución de las variables. Una tercera dimensión solo tendría sentido en el caso de querer enfocarnos en el partido de Ciudadanos. Por lo tanto la explicación total conseguida ha sido de 89,10%. Se ha podido descubrir de una forma clara y visual qué tipo de votantes están suelen ser los de cada partido. En algunos casos de ha visto de una forma muy clara, como los jubilados al PP o la parte doméstica al PSOE y en otros no tanto.