ENCUESTA CUATRIMESTRAL DE ESTRUCTURA SALARIAL
TRABAJO FINAL DE INSTRUMENTOS ESTADÍSTICOS PARA EL ANÁLISIS ECONÓMICO
DANIEL MONSALVE

DESCRIPCIÓN DE LA BASE DE DATOS UTILIZADA

La encuesta cuatrimestral de estructura salarial es una operación estadística realizada desde 1995 en el marco de la Unión Europea con criterios comunes de metodología y contenidos, con el fin de obtener unos resultados comparables sobre la estructura y distribución de los salarios entre sus Estados Miembros. La encuesta investiga la distribución de los salarios en función de una gran variedad de variables como son el sexo, la ocupación, la rama de actividad, la antigüedad, o el tamaño de la empresa.

La encuesta cuatrimestral de estructura salarial española, utilizada en este trabajo, se encuentra publicada en el INE.

El objetivo será estudiar la distribución de los salarios en función del sexo, el nivel de estudios y el departamento al que pertenezca el trabajador.

Para más información, revisar la cheat-sheet.

VARIABLES UTILIZADAS

Variable cuantitativa:

- Salarios

Variables cuantitativas:

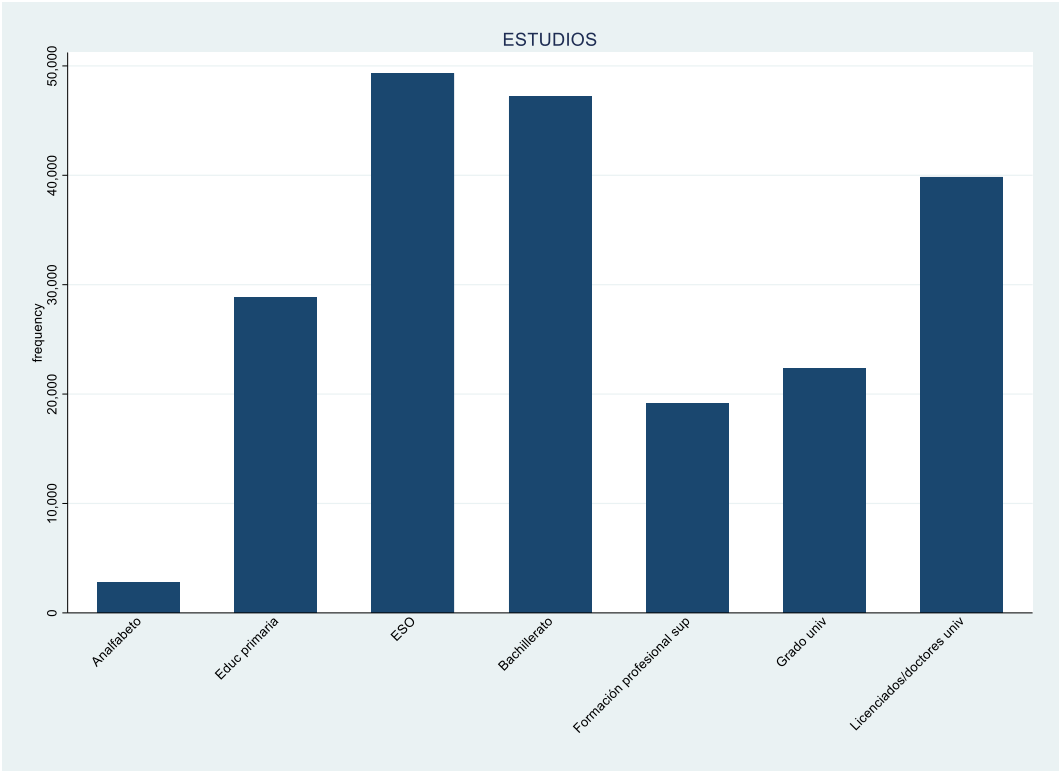
- Nivel de estudios:
  - Analfabeto
  - Educación primaria
  - ESO
  - Bachillerato
  - Formación profesional superior
  - Grado universitario
  - Licenciados/doctores universitarios
- Departamento:
  - Directores y gerentes
  - Técnicos y profesionales científicos e intelectuales
  - Técnicos y profesionales de apoyo
  - Empleados de oficina
  - Restauración y comercio
  - Cuidado de personas
  - Seguridad
  - Trabajadores cualificados del sector primario
  - Trabajadores cualificados de la construcción
  - Trabajadores cualificados de la industria
  - Operadores de maquinaria
  - Trabajadores no cualificados servicios
  - Peones
  - Militares
- Sexo

ANÁLISIS EXPLORATORIO DE LOS DATOS

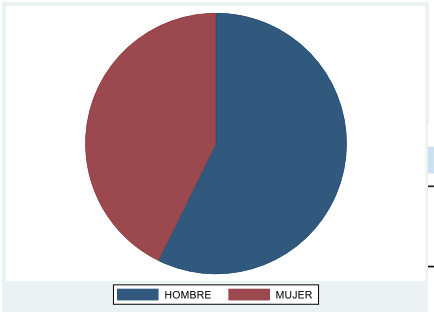
Comenzamos con el análisis estadístico de las variables cualitativas: nivel de estudios, departamento y sexo. En este tipo de variables permite calcular la distribución de frecuencias y la moda, además permit e ser graficada a través de un gráfico de barras o sectores.

- Nivel de estudios

ESTU	Freq.	Percent	Cum.
Analfabeto	2,767	1.32	1.32
Educ primaria	28,843	13.77	15.09
ESO	49,328	23.55	38.65
Bachillerato	47,170	22.52	61.17
Formación profesional sup	19,170	9.15	70.32
Grado univ	22,318	10.66	80.98
Licenciados/doctores univ	39,840	19.02	100.00
Total	209,436	100.00	



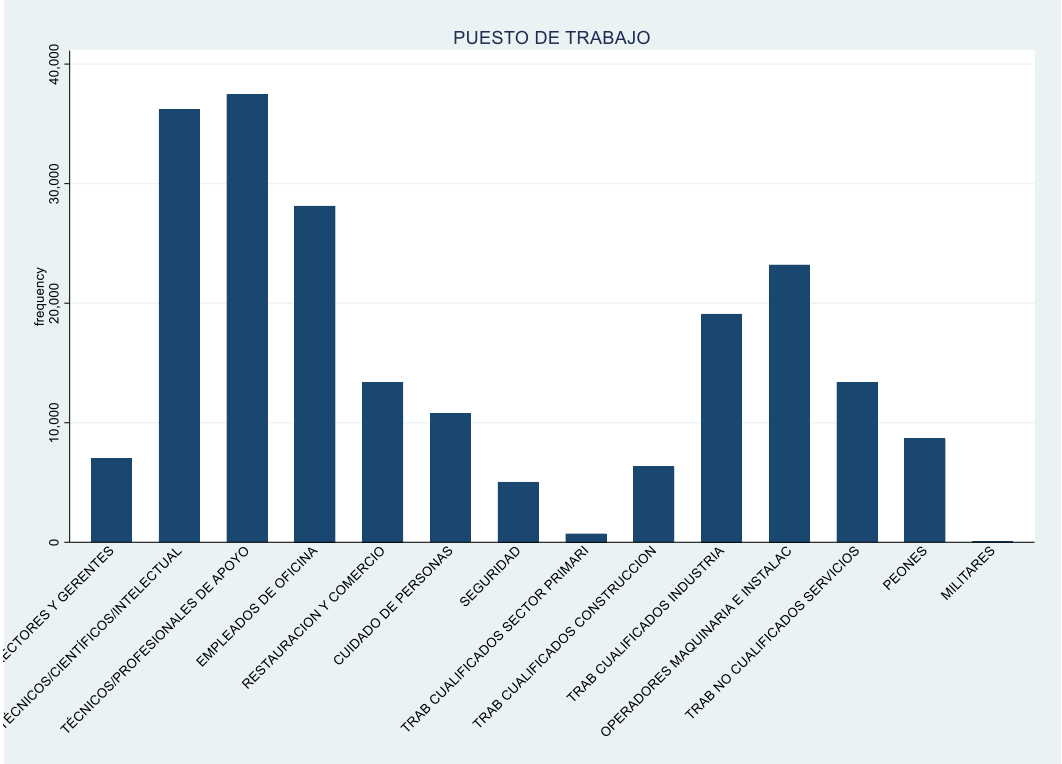
- Sexo



SEXO	Freq.	Percent	Cum.
HOMBRE	119,943	57.27	57.27
MUJER	89,493	42.73	100.00
Total	209,436	100.00	

- Departamento

DEPARTAMENTO	Freq.	Percent	Cum.
DIRECTORES Y GERENTES	6,997	3.34	3.34
TÉCNICOS/CIENTÍFICOS/INTELECTUALES	36,217	17.29	20.63
TÉCNICOS/PROFESIONALES DE APOYO	37,469	17.89	38.52
EMPLEADOS DE OFICINA	28,128	13.43	51.95
RESTAURACION Y COMERCIO	13,387	6.39	58.35
CUIDADO DE PERSONAS	10,755	5.14	63.48
SEGURIDAD	5,036	2.40	65.89
TRAB CUALIFICADOS SECTOR PRIMARIO	712	0.34	66.23
TRAB CUALIFICADOS CONSTRUCCION	6,330	3.02	69.25
TRAB CUALIFICADOS INDUSTRIA	19,094	9.12	78.37
OPERADORES MAQUINARIA E INSTALACIONES	23,206	11.08	89.45
TRAB NO CUALIFICADOS SERVICIOS	13,367	6.38	95.83
PEONES	8,664	4.14	99.96
MILITARES	74	0.04	100.00
Total	209,436	100.00	

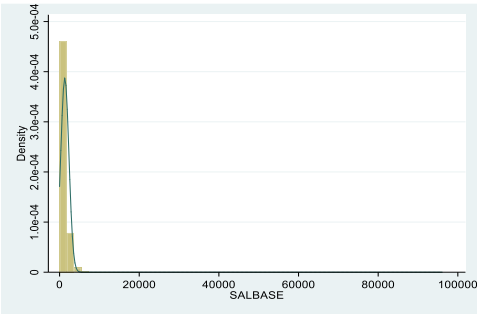


La variable salarios es la única variable cuantitativa utilizada en el análisis. Al tratarse de una variable cuantitativa continua y discreta permite la realización de operaciones algebraicas, y por ello, se pueden realizar numerosos estadísticos descriptivos. Esta variable toma muchos valores por lo que el gráfico recomendado es el histograma.

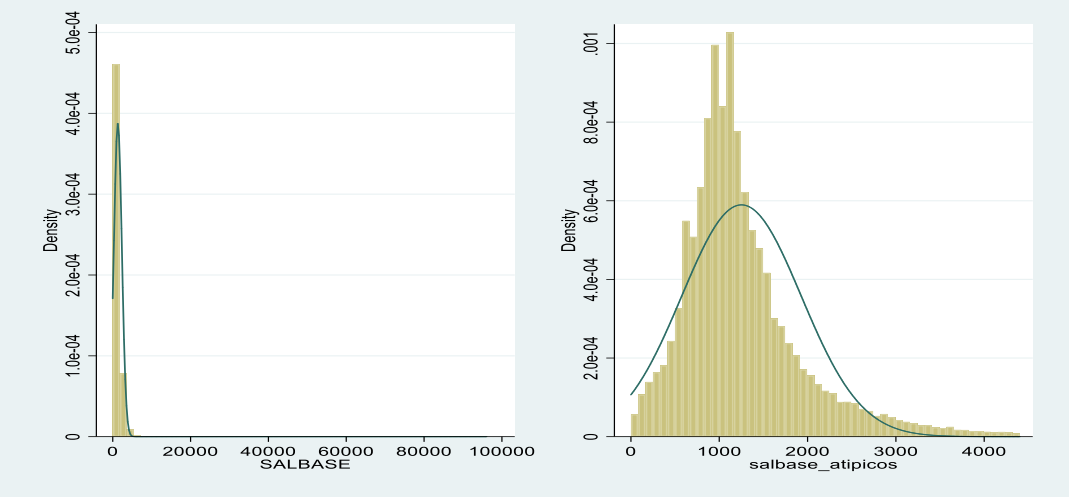
Existen varias tablas descriptivas para la variable cuantitativa, la siguiente es una de las más completas. Con la información de los diferentes percentiles podemos confirmar que la muestra de salarios no se distribuye como una normal, algo que se puede observar en el histograma más claramente.

SALBASE

	Percentiles	Smallest		
1%	138.48	0		
5%	400.02	0		
10%	581.83	0	Obs	209,436
25%	845.685	0	Sum of Wgt.	209,436
50%	1115.38		Mean	1317.694
		Largest	Std. Dev.	1029.458
75%	1525.44	64190.53		
90%	2214.29	93677.78	Variance	1059783
95%	2836.5	94957.23	Skewness	17.92898
99%	4684.85	96090.44	Kurtosis	1197.937



La variable salarios contiene valores extremos que pueden llegar a complicar el análisis. Sabemos que esos valores extremos no son atípicos, pero puede ser interesante utilizar el criterio egen zi para la eliminación de atípicos y extraer esos valores extremos en algunos momentos del análisis.



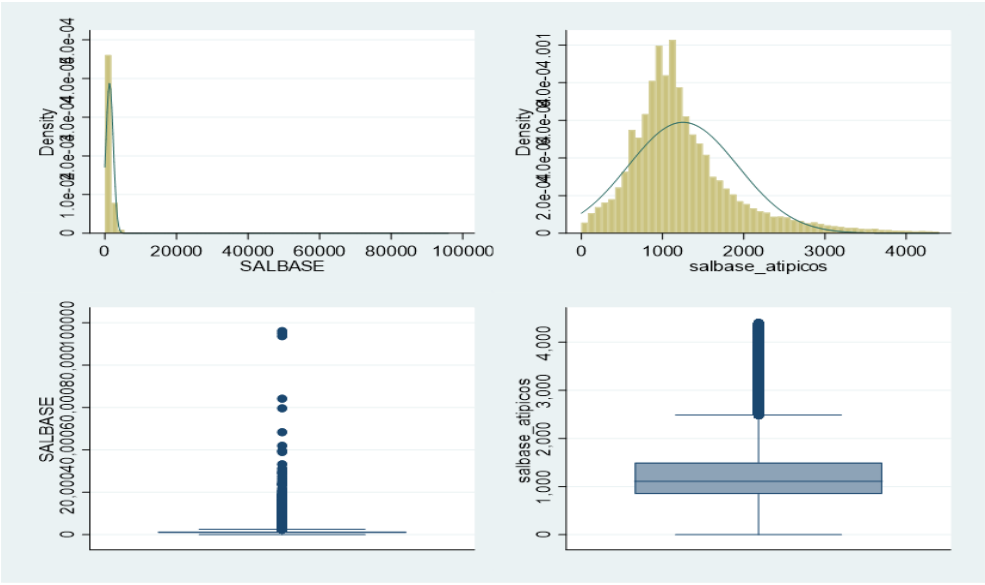
De esta forma, eliminando los sueldos extremos (mediante el criterio egen zi), la muestra queda de la siguiente manera:

variable	N	mean	max	min	p50	p25	p75	sd
salbase	209436	1317.694	96090.44	0	1115.38	845.685	1525.44	1029.458

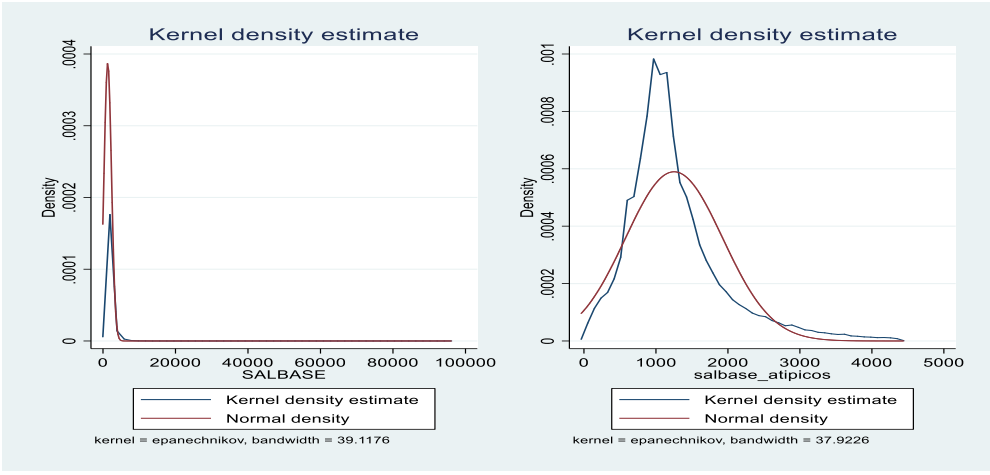
El sueldo máximo pasa de ser 96.090 a 4.404, eliminándose 2.559 valores considerados “atípicos”.

El objetivo del trabajo no extraer una muestra de salarios que se comporte como una normal, sabemos que los salarios no se distribuyen de esa manera. Aun así, en el documento do. he realizado diferentes contrastes de normalidad sobre la nueva variable sin atípicos, donde se puede confirmar que eliminando los “valores atípicos” las muestra se asemeja más a una normal.

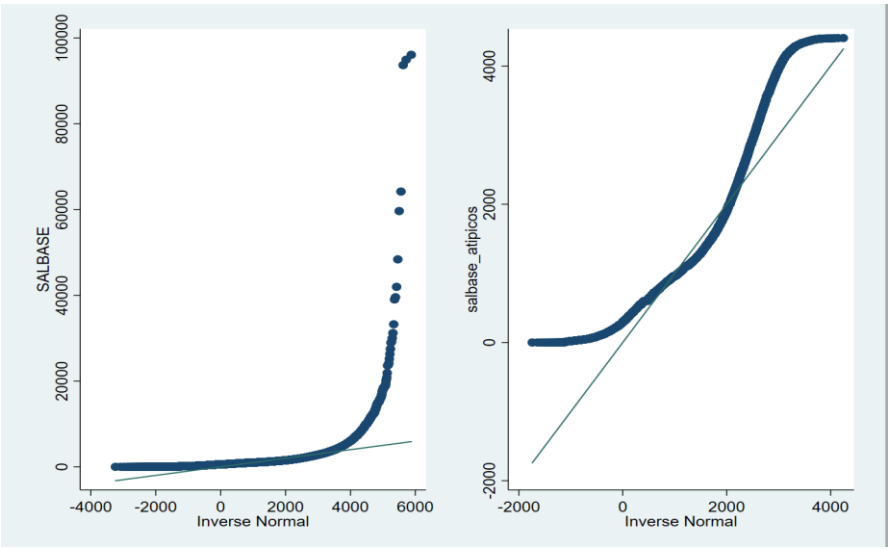
- Gráfico de cajas



- Criterio de densidad de kernel



- Criterio cuartiles de distribución normal



La utilización de gráficos no es la única manera de estudiar la normalidad, fijándonos en la asimetría y la curtosis llegamos a la misma conclusión. Se considera que una variable se distribuye de forma normal cuando el valor de la asimetría es 0 y la curtosis 3, en el caso de nuestra nueva variable los resultados son más próximos.

salbase_atipicos

	Percentiles	Smallest		
1%	137.62	0		
5%	397.92	0		
10%	577.8	0	Obs	206,877
25%	842.63	0	Sum of Wgt.	206,877
50%	1109.05		Mean	1252.086
		Largest	Std. Dev.	676.6101
75%	1500	4400		
90%	2126	4402.17	Variance	457801.2
95%	2637.55	4403.51	Skewness	1.431308
99%	3642.37	4404.35	Kurtosis	5.892025

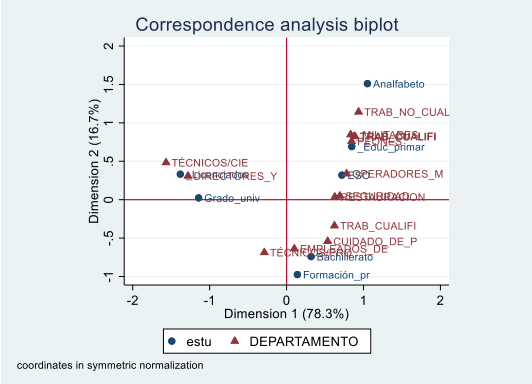
RELACIÓN ENTRE VARIABLES Y ANÁLISIS DE CORRESPONDENCIAS

Una vez analizado las distintas variables por separado el siguiente paso es contrastar la relación entre variables. Primero analizaré las relaciones entre las distintas variables cualitativas, luego analizaré los salarios (variable cuantitativa) con el resto de las variables cualitativas, y por último intentaré transformar la variable salario en una variable cualitativa, para seguir comparando.

- **Relación entre nivel de estudios y departamento**:
  con el contraste de Pearson chi2 se puede contrastar la hipótesis nula de independencia, y con el de V de Cramér's se obtiene la medida de asociación.

  En este caso la V de Cramer nos indica una correlación de 0,3515 siendo 1 correlación absoluta y 0 ninguna relación. Las tablas también nos proporcionan mucha información de interés, por ejemplo: el 11,24% de los licenciados o doctores universitarios son directores o gerentes de sus empresas.

  Una vez rechazada la hipótesis nula de independencia entre ambas variables continuamos con el análisis de correspondencias y sus gráficos descriptivos.



- **Relación entre sexo y departamento:**
  La V de Cramer nos indica una correlación de 0,4324. Destacar que el 95,95% de los militares son hombres, mientras que el 76,82% de los trabajadores dedicados al cuidado de personas son mujeres.
  En este caso los gráficos descriptivos no son tan claros.

- **Relación entre nivel de estudios y sexo**: En este caso el nivel de correlación no es suficientemente significativo.

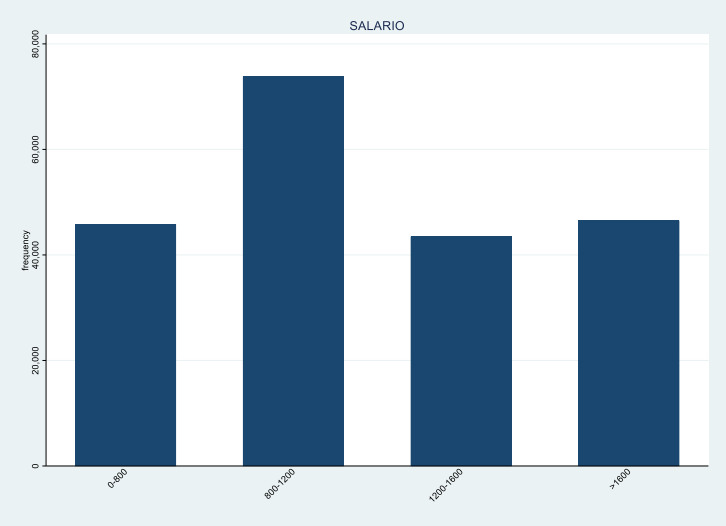
- **RELACIÓN ENTRE SALARIO Y NIVEL DE ESTUDIOS**

ESTU	Summary of SALBASE Mean	Std. Dev.	Freq.
Analfabet	955.65079	447.34598	2,767
Educ pri	1019.447	930.67974	28,843
ESO	1039.057	505.6449	49,328
Bachiller	1210.5624	806.73824	47,170
Formación	1360.0526	698.0218	19,170
Grado uni	1523.8225	914.78404	22,318
Licencia	1894.7472	1612.1745	39,840
Total	1317.6943	1029.4576	209,436

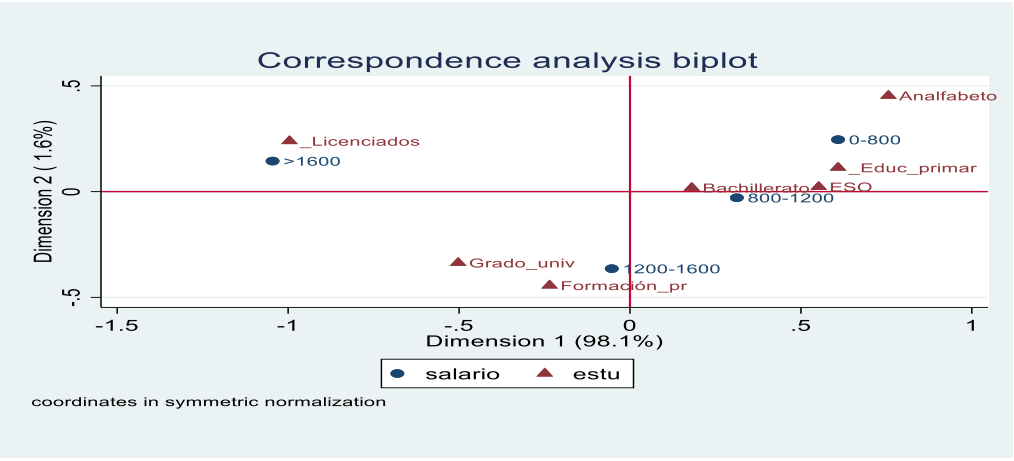
Como es de esperar, los salarios más altos corresponden a un mayor nivel de estudios.

Para continuar con el análisis de correspondencias he transformado la variable cuantitativa salarios en una variable cualitativa, siguiendo el siguiente criterio:

Salario=1	Si salbase>=0
Salario=2	Si salbase>=800
Salario=3	Si salbase>=1200
Salario=4	Si salbase>=1600



De esta manera el grafico descriptivo utilizado para los análisis de correspondencias anteriores queda de la siguiente manera:



- **RELACIÓN SEXO Y SALARIO**

  La correlación de nos indica que existe una pequeña correlación entre el sexo y el salario. Las mujeres cobran 276,422 euros menos de media.

  Esta diferencia se puede deber al techo de cristal según el cual la mayoría de los puestos de dirección son ocupados por hombres, departamentos en los que los sueldos son más altos. Podemos excluir ese efecto utilizando la muestra en la que habíamos eliminado los sueldos extremos anteriormente (salbase_sin atípicos).

  Eliminando los valores extremos, las mujeres cobran 215,194 euros menos de media, luego se confirma el techo de cristal.

- **RELACIÓN ENTRE SEXO Y DEPARTAMENTO**

  La siguiente tabla muestra cómo algunos empleos están ocupados por un sexo o por otro de forma mayoritaria.

DEPARTAMENTO	HOMBRE	MUJER	Total		DEPARTAMENTO	HOMBRE	MUJER	Total
DIRECTORES Y GERENTES	4,705	2,292	6,997		TRAB CUALIFICADOS SEC	619	93	712
	67.24	32.76	100			86.94	13.06	100
	3.92	2.56	3.34			0.52	0.1	0.34
TÉCNICOS/CIENTÍFICOS/	17,265	18,952	36,217		TRAB CUALIFICADOS CON	6,206	124	6,330
	47.67	52.33	100			98.04	1.96	100
	14.39	21.18	17.29			5.17	0.14	3.02
TÉCNICOS/PROFESIONALE	22,311	15,158	37,469		TRAB CUALIFICADOS IND	16,968	2,126	19,094
	59.55	40.45	100			88.87	11.13	100
	18.6	16.94	17.89			14.15	2.38	9.12
EMPLEADOS DE OFICINA	10,370	17,758	28,128		OPERADORES MAQUINAR	19,145	4,061	23,206
	36.87	63.13	100			82.5	17.5	100
	8.65	19.84	13.43			15.96	4.54	11.08
RESTAURACION Y COMERC	4,508	8,879	13,387		TRAB NO CUALIFICADOS	4,412	8,955	13,367
	33.67	66.33	100			33.01	66.99	100
	3.76	9.92	6.39			3.68	10.01	6.38
CUIDADO DE PERSONAS	2,493	8,262	10,755		PEONES	6,575	2,089	8,664
	23.18	76.82	100			75.89	24.11	100
	2.08	9.23	5.14			5.48	2.33	4.14
SEGURIDAD	4,295	741	5,036		MILITARES	71	3	74
	85.29	14.71	100			95.95	4.05	100
	3.58	0.83	2.4			0.06	0	0.04
					Total	119,943	89,493	209,436
						57.27	42.73	100
						100	100	100

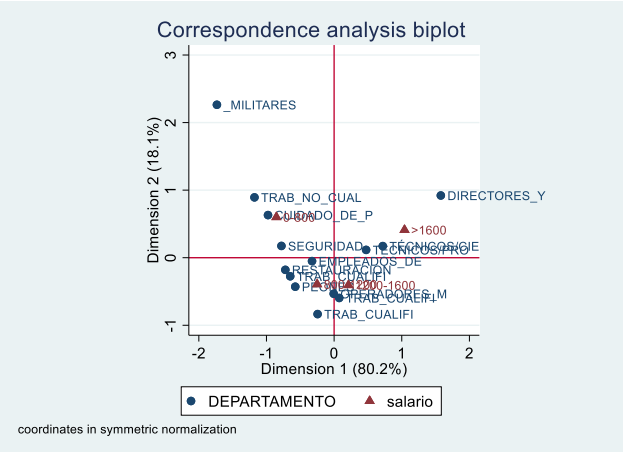
- **RELACIÓN ENTRE SALARIO Y DEPARTAMENTO**

  La siguiente tabla nos indica la diferencia en el salario medio de cada departamento respecto al puesto de director o gerente (departamento de referencia). Como era de esperar todos los trabajadores que no se dedican a la dirección cobran menos de media. Por ejemplo, los técnicos científicos e intelectuales cobran unos 1081,74 euros menos, frente a los 2175,17 euros menos de media de los militares.

Source	SS	df	MS		
Model	3.3210e+10	13	2.5546e+09	Number of obs	= 209,436
Residual	1.8875e+11	209,422	901269.475	F(13, 209422)	= 2834.46
				Prob > F	= 0.0000
Total	2.2196e+11	209,435	1059782.93	R-squared	= 0.1496
				Adj R-squared	= 0.1496
				Root MSE	= 949.35

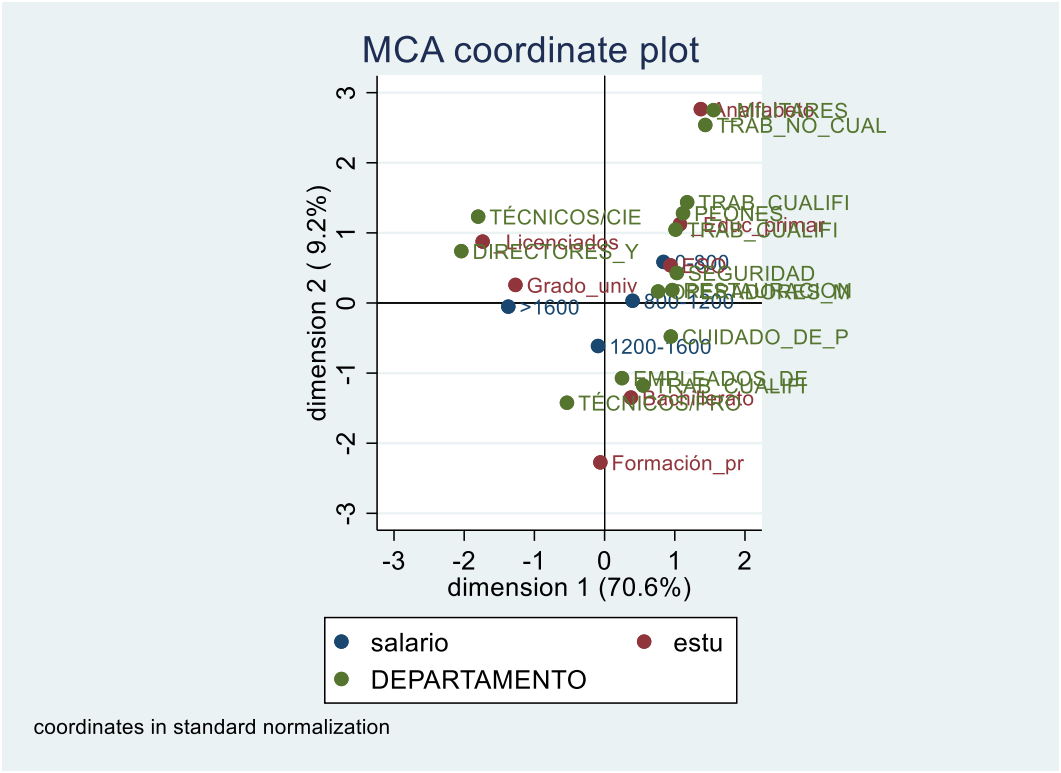
salbase	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
DEPARTAMENTO						
TÉCNICOS/CIENTÍFICOS/INTELECTUALES	-1081.745	12.39731	-87.26	0.000	-1106.043	-1057.446
TÉCNICOS/PROFESIONALES DE APOYO	-1251.591	12.36373	-101.23	0.000	-1275.823	-1227.358
EMPLEADOS DE OFICINA	-1649.884	12.68265	-130.09	0.000	-1674.742	-1625.027
RESTAURACION Y COMERCIO	-1830.609	14.00472	-130.71	0.000	-1858.058	-1803.161
CUIDADO DE PERSONAS	-1928.583	14.58108	-132.27	0.000	-1957.161	-1900.004
SEGURIDAD	-1800.3	17.54347	-102.62	0.000	-1834.685	-1765.916
TRAB CUALIFICADOS SECTOR PRIMARIO	-1770.89	37.34482	-47.42	0.000	-1844.085	-1697.695
TRAB CUALIFICADOS CONSTRUCCION	-1658.066	16.46781	-100.69	0.000	-1690.343	-1625.79
TRAB CUALIFICADOS INDUSTRIA	-1511.546	13.26686	-113.93	0.000	-1537.549	-1485.543
OPERADORES MAQUINARIA E INSTALACIONES	-1555.979	12.94781	-120.17	0.000	-1581.357	-1530.602
TRAB NO CUALIFICADOS SERVICIOS	-2025.148	14.00832	-144.57	0.000	-2052.604	-1997.692
PEONES	-1763.838	15.25885	-115.59	0.000	-1793.745	-1733.931
MILITARES	-2175.179	110.942	-19.61	0.000	-2392.623	-1957.736
_cons	2778.928	11.34936	244.85	0.000	2756.683	2801.172

Utilizando la transformación de la variable salario en una variable cualitativa podemos representar esta relación. A pesar de que la V de Cramer nos indica poca correlación, 0,2778, la representación parece coherente.



Para finalizar con esta parte del análisis de correspondencias realizo un análisis conjunto de todas las variables anteriores, un **análisis de correspondencias múltiples**.

El siguiente gráfico resume todas las relaciones que habíamos analizado por separado en el anterior apartado. Si se quiere información más precisa se puede acudir a las tablas.



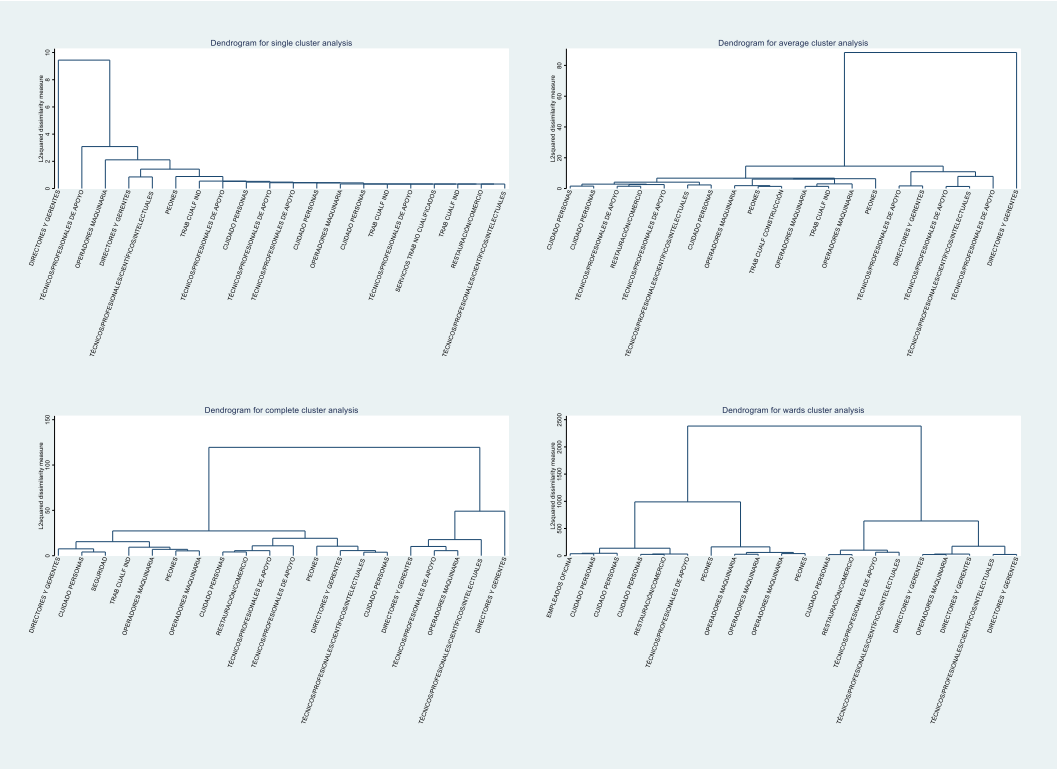
ANÁLISIS CLÚSTER

El Análisis Clústeres una técnica estadística multivariante que busca agrupar elementos (o variables) en grupos, tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

El primer paso antes de realizar este tipo de análisis es estandarizar las variables, ya que los cálculos serán sensibles a las unidades en que estén medidas dichas variables.

Realizaré primero un clúster jerárquico en busca del número de grupos más adecuado. también utilizaré los diferentes criterios de agregación: - Vecino más cercano (singlelinkage) - Promedio (average) - Vecino más lejano (complete) - Ward (wards).

Además, también será necesario seleccionar el tipo de medida que se va a utilizar para calcular la distancia entre las observaciones. En nuestro caso, hemos seleccionado la Euclidia (L2), puesto que era la que mejor se ajustaba. Una vez generado el clúster, el siguiente paso es generar un dendograma con los resultados obtenidos. Es recomendable generar un dendograma para cada clúster calculado, ya que facilitaran la elección del clúster que se utilizara posteriormente.



El criterio que da lugar a una estructura de los grupos más adecuada es el Ward. Sin embargo, por la naturaleza de las variables utilizadas, no utilizaré los grupos propuestos.

Es necesario decidir con cuantos grupos nos quedamos, en este caso la decisión será arbitraria. También, se pueden utilizar el criterio de selección del pseudo-F de Calinski/Harabasz (se selecciona el número de cluster que ofrezca el mayor valor en los primeros clusters) y el criterio de Duda/Hart que ofrece el valor Je(2)/Je(1) y el pseudo T-squared (en este caso se busca el mayor valor para el primero y el menor para el segundo).

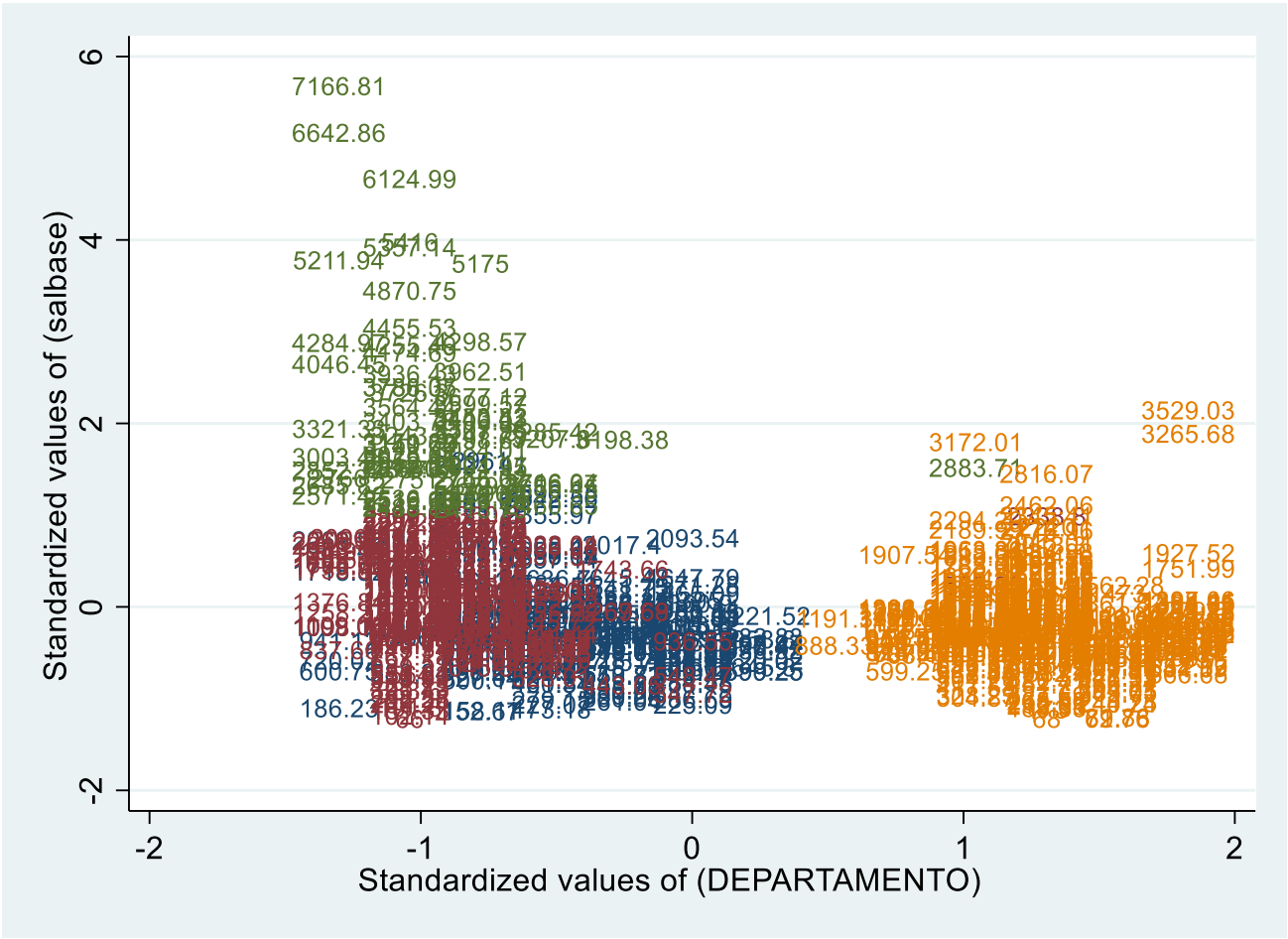
Aplicando las reglas de clúster stop obtenemos los resultados de las tablas contiguas.

Number of clusters	Calinski/ Harabasz pseudo-F
2	775.29
3	808.99
4	924.87
5	822.55
6	783.40
7	768.59
8	752.54
9	720.16
10	700.21
11	676.76
12	656.15
13	640.87
14	626.02
15	616.14

Number of clusters	Duda/Hart	
	Je(2)/Je(1)	pseudo T-squared
1	0.5628	775.29
2	0.4689	748.61
3	0.4700	377.76
4	0.3896	81.48
5	0.6476	188.31
6	0.6670	156.26
7	0.6369	160.20
8	0.4886	176.88
9	0.7426	91.16
10	0.7381	72.75
11	0.5940	95.00
12	0.6176	77.39
13	0.7007	45.28
14	0.6394	25.38
15	0.6542	64.48

Según el criterio pseudo-F de Calinski/Harabasz el mayor valor se consigue en el grupo 4 del clúster. En cambio, si se observa el criterio de Duda/Hart el mayor valor del Je(2)/Je(1) se consigue en el 9. Sin embargo, el valor que ofrece de pseudo T-squared es relativamente alto.

He decidido seleccionar 4 grupos. El último paso será realizar un nuevo cluster no jerárquico, utilizando los centroides de los grupos obtenidos. Con este último cluster se pueden realizar una serie de gráficos y gráficas descriptivas según el número de grupos realizados.



Por último, las siguientes tablas incluyen información sobre el contenido de cada uno de los grupos seleccionados en el cluster.

GRUPO 1	ESTU								
DEPARTAMENTO	Analfabeto	Educ	ESO	Bachiller	Formación	Total	Salario medio		
							1006.84304		
DIRECTORES Y GERENTES	0	0	2	6	0	8			
TÉCNICOS/CIENTÍFICOS/	0	0	4	6	0	10			
TÉCNICOS/PROFESIONALE	0	8	22	54	11	95			
EMPLEADOS DE OFICINA	0	8	27	47	12	94			
RESTAURACION Y COMERC	1	9	19	26	1	56			
CUIDADO DE PERSONAS	1	10	11	22	7	51			
SEGURIDAD	0	1	5	9	0	15			
Total	2	36	90	170	31	329			

GRUPO 2	ESTU								
DEPARTAMENTO	Formación p	Grado uni	Licenciado	Total	Salario medio				
					1218.19341				
DIRECTORES Y GERENTES	3	2	12	17					
TÉCNICOS/CIENTÍFICOS/	2	36	81	119					
TÉCNICOS/PROFESIONALE	18	17	31	66					
EMPLEADOS DE OFICINA	1	18	7	26					
RESTAURACION Y COMERC	0	1	3	4					
CUIDADO DE PERSONAS	0	2	1	3					
TRAB CUALIFICADOS IND	0	0	1	1					
OPERADORES MAQUINARIA	0	0	1	1					
Total	24	76	137	237					

GRUPO 3	ESTU								
DEPARTAMENTO	ESO	Bachiller	Formación	Grado Univ	Licenciado	Total	Salario medio		
							3304.91401		
DIRECTORES Y GERENTES	0	1	0	0	11	12			
TÉCNICOS/CIENTÍFICOS/	0	1	0	5	24	30			
TÉCNICOS/PROFESIONALE	1	3	1	11	6	22			
EMPLEADOS DE OFICINA	1	1	2	1	1	6			
RESTAURACION Y COMERC	1	0	0	0	0	1			
TRAB CUALIFICADOS IND	0	0	0	1	0	1			
Total	3	6	3	18	42	72			

GRUPO 4	ESTU								
DEPARTAMENTO	Analfabet	Educ	ESO	Bachiller	Formación	Grado	Licenciado	Total	Salario medio
									987.1552883
TRAB CUALIFICADOS SEC	0	2	0	0	0	0	0	2	
TRAB CUALIFICADOS CON	3	6	12	3	2	0	0	26	
TRAB CUALIFICADOS IND	2	21	33	33	16	0	0	105	
OPERADORES MAQUINARIA	0	36	59	26	5	2	0	128	
TRAB NO CUALIFICADOS	2	19	28	5	1	0	1	56	
PEONES	2	9	25	7	1	1	0	45	
Total	9	93	157	74	25	3	1	362	