Análisis de datos

Apellidos, nombre: Garcia Salinas, Daniel Grupo de prácticas: L3

DNI: 71188618F

Notas:

1) Esta práctica se realizará los días 3, 4, 5 y 6 de octubre de 2023.
2) La fecha límite de entrega de esta práctica, que debe presentarse en el Campus Virtual Uva, son las 14 horas del miércoles 11 de octubre de 2023.

En esta práctica se utilizará el fichero de datos Datos-Valladolid.sgd.

El fichero Datos-Valladolid.sgd es un fichero en formato *Statgraphics* que contiene datos de los 41 municipios de la provincia de Valladolid con más de 1000 habitantes en una determinada fecha F_1. La información que aparece recogida en el fichero proviene de diferentes fuentes oficiales.

Para cada municipio, la información aparece organizada en dos apartados. El primero recogedatos relativos al territorio y la población. El segundo ofrece indicadores sobre el empleo y la contratación laboral (datos referidos a fecha F_2, posterior a F_1).

Así, las variables que aparecen en este fichero son las siguientes:

TERRITORIO Y POBLACIÓN

Municipio	Denominación del municipio
Partido	Partido judicial al que pertenece el municipio
Población	Población oficial, a fecha F_1

ESTRUCTURA PRODUCTIVA

Tasa Paro	Tasa de paro, a fecha F_2
Demandantes Empleo	Número de personas demandantes de empleo, a fecha F_2
Paro	Paro registrado, datos referidos, a fecha F_2
CAg	Número de contratos registrados en el sector de la Agricultura, a fecha F_2
CIn	Número de contratos registrados en el sector de la Industria, a fecha F_2
CCo	Número de contratos registrados en el sector de la Construcción, a fecha F_2
CSe	Número de contratos registrados en el sector Servicios, a fecha F_2

En el anexo se presentan los datos que aparecen en el

fichero. Este fichero de datos puede ser obtenido

1) En \\sanson\Estadistica de la red del laboratorio de Informática.
2) En Campus Virtual UVa

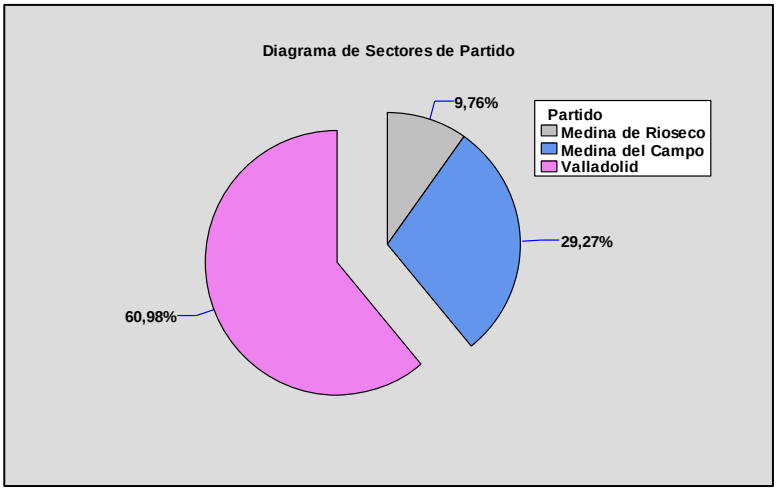
Importar el fichero **Datos-Valladolid.sgd** al programa Statgraphics y almacenarlo en memoria.

EJERCICIO 1

1.1. Clasificar las variables que aparecen en el fichero. Justificar respuesta.

- **Variables cualitativas**: Municipio, Provincia. (Son variables no medibles numéricamente, solo expresan características)
- **Variable cuantitativa continua**: Tasa paro. (Toma valores numéricos pero solo enteros)
- **Variables cuantitativas discretas**: Población, Demandantes de empleo, Paro, Contratos de agricultura/industria/construcción/servicios.

1.2. Dibujar un diagrama de sectores para la variable *Partido*. ¿Cuál es el porcentaje de municipios vallisoletanos, entre los que tienen más de 1000 habitantes, que pertenecen al partido judicial de Valladolid?



El 60,98% de los municipios vallisoletanos que tienen más de 1000 habitantes pertenecen al partido judicial de Valladolid

1.3. Insertar una nueva variable al fichero, *Sector*, para clasificar los municipios en cuatro grupos: (i)Industrial (si más del 14.8 % de los contratos registrados se ha realizado en el sector industrial), (ii) Agrícola (si no es industrial y más del 14 8 % de los contratos registrados se ha realizado en el sector de la agricultura), (iii) Construcción (si no es ni industrial ni agrícola y más del 14.8 % de los contratos registrados se ha realizado en el sector de la construcción) y (iv) Servicios, en caso contrario. Construir la tabla de contingencia de las variables *Partido* y *Sector*. ¿Qué porcentaje de los municipios del partido judicial de Valladolid aparecen clasificados como industriales? ¿Y del partido judicial de Medina del Campo? Razonar la respuesta.

-El 28% de los municipios del partido judicial de Valladolid aparecen como industriales
-El 41,67% de los municipios del partido judicial de Medina del Campo son clasificados como industriales

	Agricola	*Construcción*	*Industriales*	*Servicios*
Medina de Rioseco	1	0	0	3
	2,44%	0,00%	0,00%	7,32%
	25,00%	0,00%	0,00%	75,00%
Medina del Campo	4	0	5	3
	9,76%	0,00%	12,20%	7,32%
	33,33%	0,00%	41,67%	25,00%
Valladolid	4	2	7	12
	9,76%	4,88%	17,07%	29,27%
	16,00%	8,00%	28,00%	48,00%
Total por Columna	9	2	12	18
	21,95%	4,88%	29,27%	43,90%

Explicación: la tabla tiene 3 filas para cada variable cualitativa. Debemos coger la tercera ya que la segunda muestra el porcentaje respecto del total y a nosotros nos interesa el porcentaje condicionado a que el partido judicial sea el de Valladolid o el de Medina del Campo.

EJERCICIO 2

2.1. ¿Qué municipio tiene una tasa de paro superada por el 69% de los municipios? ¿Cuál es su población? ¿En qué sector aparece clasificado?

Percentiles para Tasa Paro	
	Percentiles
1,0%	5,42
5,0%	9,53
10,0%	14,24
25,0%	16,49
31,0%	16,83
50,0%	18,71
75,0%	25,8
90,0%	29,99
95,0%	31,37

Para responder a la pregunta debemos mirar en la tabla de percentiles el P_{31}. Este corresponde a la tasa de paro de 16,83%. Los municipios con esta tasa de paro son Fuensaldaña y Peñafiel

Fuensaldaña	Valladolid	1468	16,83
Peñafiel	Valladolid	5428	16,83

Fuensaldaña pertenece al sector servicios y Peñafiel al sector industrial

2.2. Calcular los tres cuartiles de la variable *Paro*, ¿a qué municipios corresponden?

Percentiles para Paro	
	Percentiles
1,0%	45,0
5,0%	61,0
10,0%	73,0
25,0%	100,0
50,0%	198,0
75,0%	399,0
90,0%	832,0
95,0%	1787,0
99,0%	24926,0

-Q_1 = P25 = 100

- El municipio con 100 personas en paro es Serrada.

-Q_2 = P50 = 198

- El municipio con 198 personas en paro es Mojados.

-Q_3 = P75 = 399

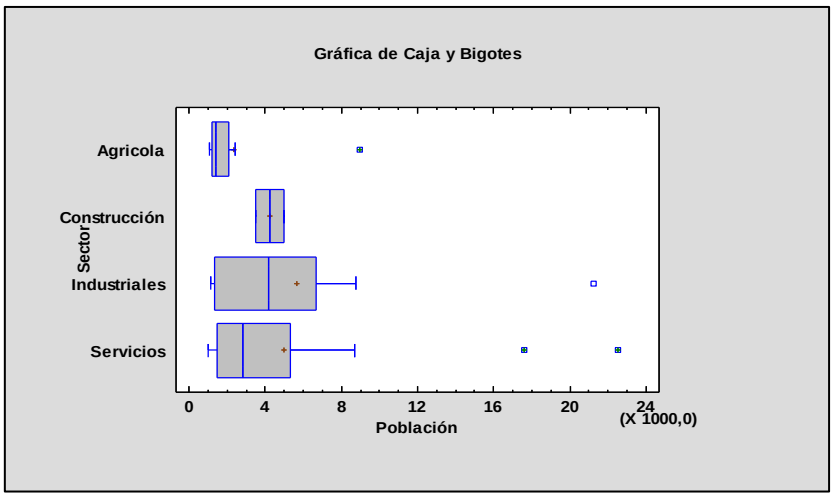
- El municipio con 399 personas en paro es Zaratán.

2.3. Calcular la media de la variable *Paro* e indicar qué municipio tiene la tasa más cercana a ese valor.

Resumen Estadístico para Paro	
Recuento	41
Promedio	952,073

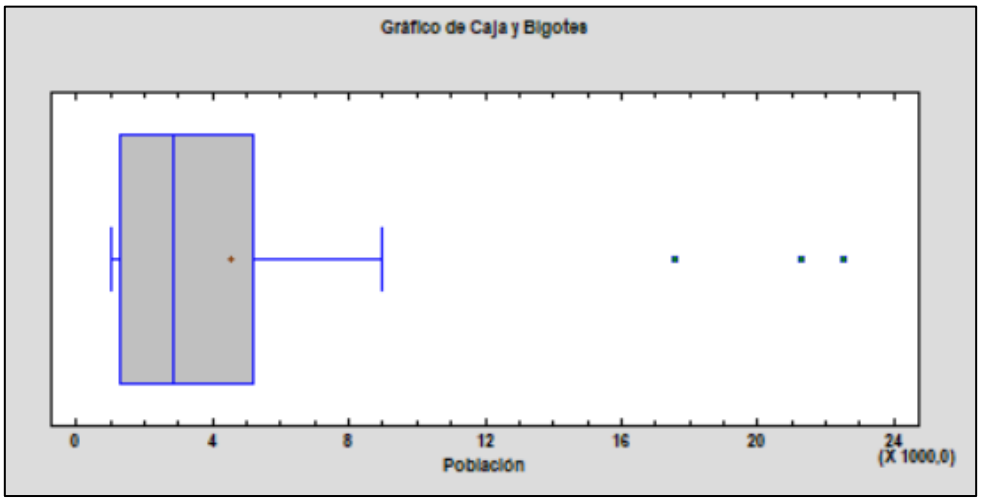
El municipio con la tasa de Paro más cercana a la media es Tordesillas, con un paro de 832 personas.

2.4. Con el fin de comparar la población de los municipios distintos de la capital en función del sector en que fueron clasificados en el ejercicio anterior, construir los correspondientes diagramas de caja. Identificar, si existen, los outliers y los posibles outliers.



Se puede ver que en el sector industrial hay un posible outlier que corresponde al municipio de Medina del Campo. Además, existen 3 outliers; 1 en el sector agrícola que corresponde al municipio de Tordesillas, y 2 más en el sector servicios que corresponden a las poblaciones de Arroyo de la Encomienda y Laguna de Duero

2.5. Realizar ahora un diagrama de caja para la variable *Población* en los municipios distintos de la capital. Identificar, si existen, los outliers y los posibles outliers.



En este caso no existen posibles datos atípicos. Eso si, hay 3 casos atípicos que corresponden a los municipios de Arroyo de la Encomienda (17572 de población), Medina del Campo (21274 de población) y Laguna de Duero (22555 de población)

2.6. Calcular los indicadores de centralización, dispersión y forma más adecuados para la variable *Población* en los municipios distintos de la capital, justificando la elección. ¿Qué se puede decir acerca de la estructura de esta variable?

-Como hay bastantes datos atípicos, debemos utilizar indicadores que no se vean afectados por estos datos (o al menos no en gran cantidad). En otras palabras, vamos a calcular la mediana (centralización), el rango intercuartílico (dispersión) y el coeficiente de Bowley-Yule (forma).

Percentiles para Población	
	Percentiles
25,0%	1267,0
50,0%	2834,0
75,0%	5169,5

MEDIANA

Mediana = $Q_2 = P_{50} = 2834$

RANGO INTERCUARTÍLICO:

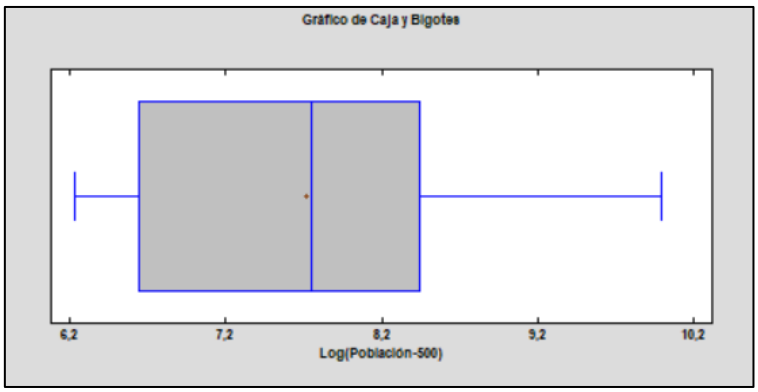
Rango intercuartílico (IQR) = $Q_3 - Q_1 = P_{75} - P_{25}$; IQR = 5331 – 1332; IQR = 3999

COEFICIENTE DE BOWLEY-YULE

$A_B = ((Q_3-Q_2) - (Q_2-Q_1)) / IQR$; $A_B = ((5331 – 3184) – (3184 – 1332)) / 3999$; $A_B = 0.07376$

$A_B > 0$, luego la distribución de la variable población es asimétrica a la derecha.

2.7. Con el fin de eliminar los outliers de la variable anterior (*Población* en los municipios distintos de la capital), se propone realizar la transformación *Ln(Población-500).* Dibujar el diagrama de caja correspondiente, indicando, si existen, los valores "raros". Describir la estructura de esta nueva variable.



Percentiles para Log(Población-500)	
	Percentiles
25,0%	6,63888
50,0%	7,74397
75,0%	8,44821

Resumen Estadístico para Log(Población-500)	
Recuento	40
Promedio	7,7132
Mediana	7,74397
Desviación Estándar	1,07968
Coeficiente de Variación	13,9978%
Mínimo	6,22851
Máximo	10,0013
Rango	3,77278

-Para este caso ya no existen outliers. Pasamos a calcular las nuevas medidas para esta transformación de la variable *Población*:

- Centralización: Media = 7,7132; Mediana = 7,74397
- Dispersión: Desviación Estándar = 1,07968; Coeficiente de Variación (CV) = 1,07968/ 7,7132 = 0,13997; IQR = 8,44821 – 6,63888 = 1,80933
- Forma: A_B = ((8,44821 – 7,74397) – (7,74397–6,63888)) / 1,80933 = -0,2215; Coeficiente de Pearson = (7,7132-7,74397) / 1,07968 = - 0,028

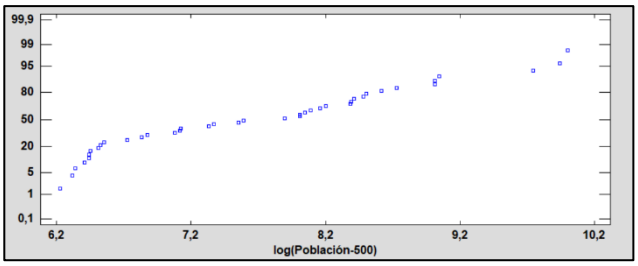
2.8. Calcular los cuartiles y los sextiles para la variable definida en el apartado anterior (es decir, *Ln(Población - 500)* en los municipios distintos de la capital). Identificar los municipios a los que corresponden. Calcular, a partir de los valores anteriores, los correspondientes a la variable *Población*. ¿Coinciden estos valores con los obtenidos directamente para la variable *Población*? Justificar la respuesta.

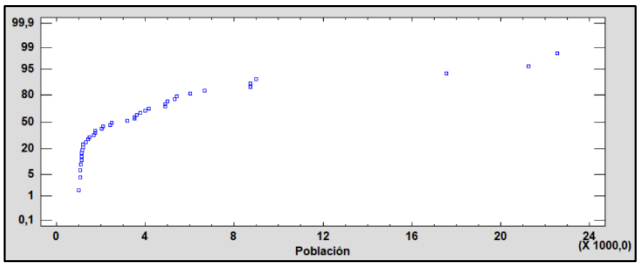
Percentiles para Log(Población-500)	
	Percentiles
16,67%	6,45205
25,0%	6,63888
33,3%	7,07918
50,0%	7,74397
66,66%	8,20385
75,0%	8,44821
83,33%	8,72875

e^y+500
1134
1264
1686
2807
4154
5166
6678

Percentiles para Población	
	Percentiles
16,67%	1134,0
25,0%	1267,0
33,33%	1687,0
50,0%	2834,0
66,67%	4155,0
75,0%	5169,5
83,33%	6678,0

Sabiendo que la variable *y* son los percentiles de la variable Log (Población-500), podemos ver que los valores son valores bastantes similares a los obtenidos.





Como podemos comprobar, la variable log(Población – 500) sigue un recorrido más lineal, mientras que la variable Población sigue un recorrido logarítmico. Esto significa que los valores intermedios de la transformación van a ser más pequeños que los valores entre medios de la variable original (se puede comprobar esto con los valores entre el primer sextil y el quinto sextil)

EJERCICIO 3

3.1. Realizar el diagrama de tallo-hojas para la variable *Tasa Paro*.

Diagrama de Tallo y Hoja para Tasa: unidad = 1,0 1|2 representa 12,0

3	0|599
7	1|3444
(16)	1|5556666777788899
18	2|122444
12	2|55777789
4	3|014
1	3|6

3.2. Construir una tabla de frecuencias adecuada para la variable *Tasa Paro*.

- La variable *Tasa de Paro* es una variable cuantitativa continua. Esto significa que debemos de agrupar sus datos en clases. Para ello debemos calcular el número de clases, el límite inferior y el límite superior adecuado.

- Número de clases: √(Total de datos (N)) = √(41) = 6,403 => 6 clases

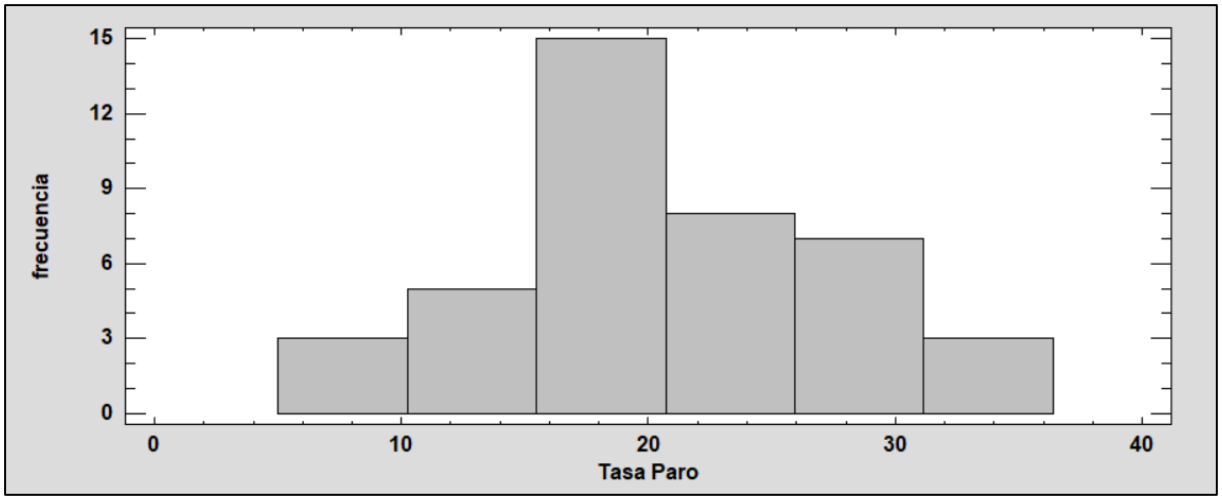
- Límite inferior: 5

- Límite superior: 36,4

Tabla de Frecuencias para Tasa Paro

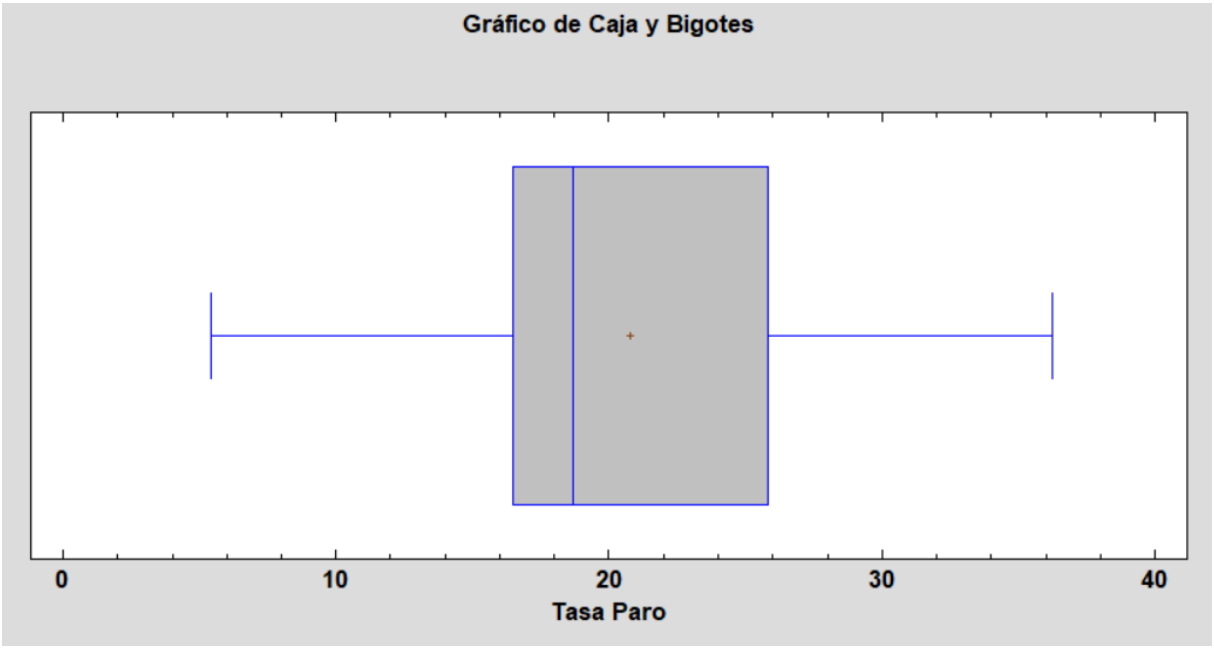
Clase	*Límite Inferior*	*Límite Superior*	*Punto Medio*	*Frecuencia*	*Frecuencia Relativa*
	menor o igual	5		0	0,0000
1	5	10,2333	7,61667	3	0,0732
2	10,2333	15,4667	12,85	5	0,1220
3	15,4667	20,7	18,0833	15	0,3659
4	20,7	25,9333	23,3167	8	0,1951
5	25,9333	31,1667	28,55	7	0,1707
6	31,1667	36,4	33,7833	3	0,0732
	mayor de	36,4		0	0,0000

3.3. Dibujar el histograma con el agrupamiento del apartado anterior.



Como el número de clases es 6, la amplitud del intervalo sería (36,4 – 5) / 6 = 5,23

3.4. Realizar un diagrama de caja para la variable *Tasa Paro*. Identificar, si existen, los outliers y posibles outliers. ¿Qué se puede decir acerca de la estructura de esta variable?



- En esta variable no existen posibles outliers ni outliers. La distribución es asimétrica a la derecha.

3.5. Calcular los indicadores de centralización, dispersión y forma más adecuados, justificando la elección.

Resumen Estadístico para Tasa Paro	
Recuento	41
Promedio	20,7671
Mediana	18,71
Desviación Estándar	6,94589
Coeficiente de Variación	33,4467%
Mínimo	5,42
Máximo	36,22
Rango	30,8

Percentiles para Tasa Paro	
	Percentiles
25,0%	16,49
50,0%	18,71
75,0%	25,8

No hay valores atípicos, luego se pueden calcular el conjunto total de medidas

- Centralización: Media = 20,7671; Mediana = 18,71
- Dispersión: Desviación Estándar = 6,94589; Coeficiente de Variación (CV) = 6,94589/ 20,7671 = 0,334467; IQR = 25,8 – 16,49 = 9,31
- Forma: A_B = ((25,8 – 18,7) – (18,7– 16,49)) / 9,31 = 0,5230; Coeficiente de Pearson = (20,7671-18,71) / 6,94589 = 0,29616

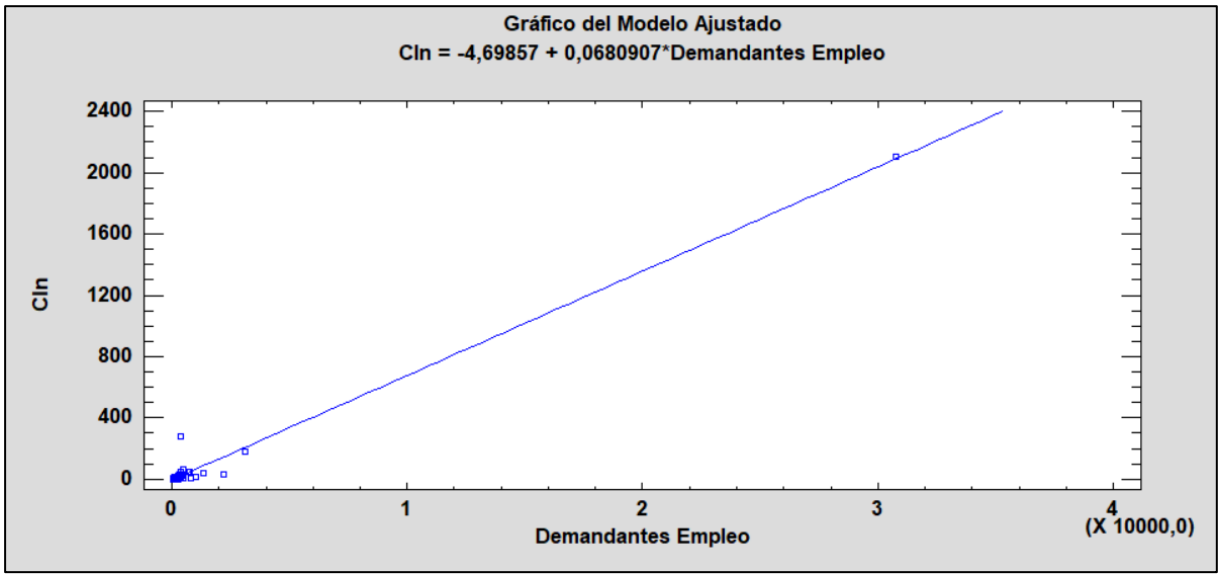
3.6. Se clasifican los municipios en tres grupos *Po, No y M*. El grupo *Po* está constituido por el 15 % de los municipios con menor tasa de paro; el grupo *Mu* por el 15% de los municipios con mayor tasa de paro y el grupo *No* por el resto de los municipios. Identificar, razonando la respuesta, los componentes de los grupos *Po* y *Mu*.

Percentiles para Tasa Paro	
	Percentiles
15,0%	14,85
85,0%	27,97

-Al grupo *Po* pertenecen los municipios con menor tasa de paro que 14,85% (Boecillo, Quintanilla del Onésimo, Villanubla, Olmedo, Campaspero, Medina de Rioseco y Mojados). Al grupo *Mu* pertenecen los municipios con una tasa de paro mayor que el 27,97% (Medina del Campo, Valdestillas, Laguna de Duero, Alaejos, Tudela de Duero, Villanueva de Duero, Cigales).

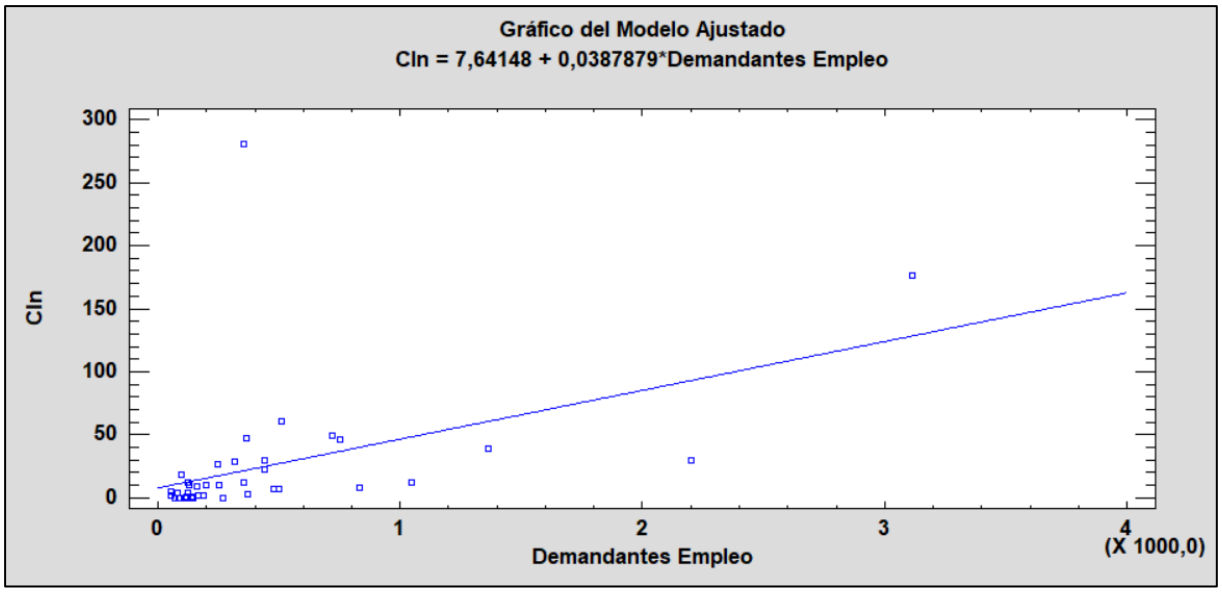
EJERCICIO 4

4.1. Realizar el diagrama de dispersión de las variables *CIn* y *Demandantes Empleo*. ¿Qué tipo de correlación se tiene? ¿Es adecuado realizar un ajuste lineal? Justificar la respuesta.



Coeficiente de Correlación = 0,989005

Ya que el coeficiente de correlación es cercano a 1, se podría decir que hay una gran dependencia entre las variables. Ahora bien, hay un gran outlier (Valladolid), luego se debe excluir este valor para ver si la variación de la recta y del coeficiente de correlación es significante

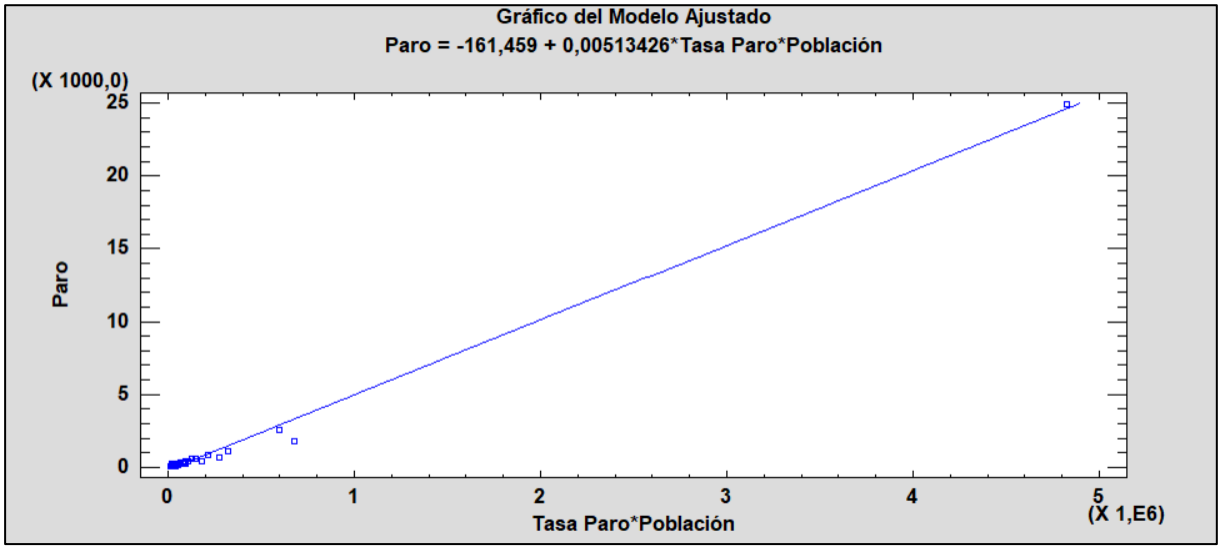


Coeficiente de Correlación = 0,452194

Como podemos ver, el coeficiente de correlación ha variado significativamente, esto ocurre porque al haber un dato atípico, provoca que la recta pase por él, generando un valor del coeficiente de correlación engañoso.

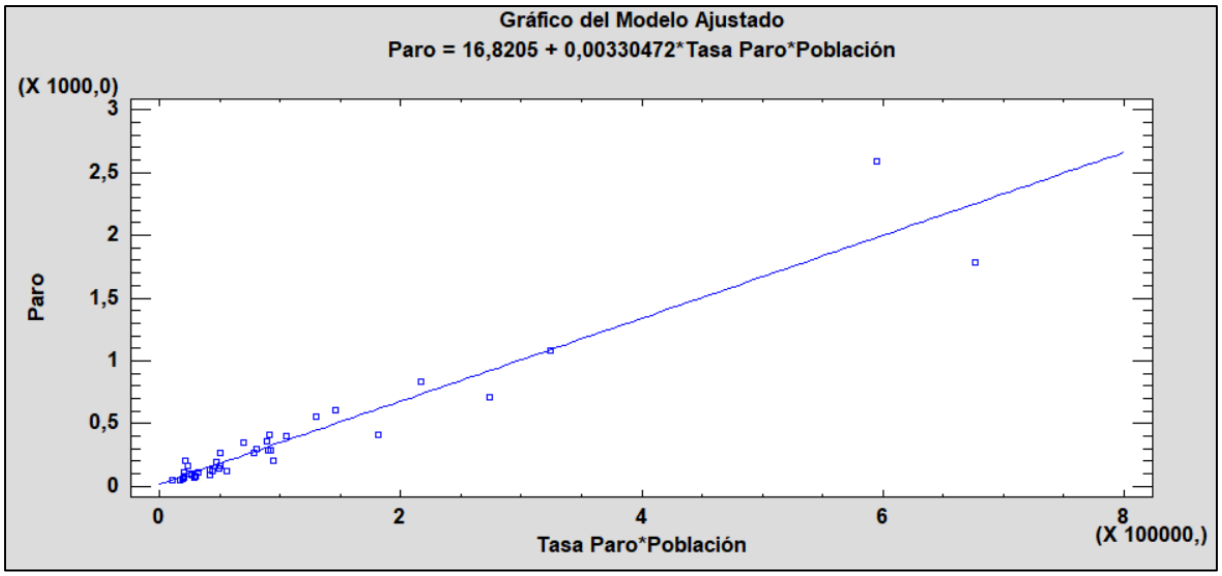
Teniendo esto en cuenta, no es muy adecuado realizar un ajuste lineal en este caso.

4.2. Construir una nueva variable, *Auxiliar Paro*, igual al producto de las variables *Tasa Paro* y *Población*. Repetir el apartado anterior con las variables *Paro* y *Auxiliar Paro.*



Coeficiente de Correlación = 0,997038

Nos ocurre como en el caso anterior, el coeficiente de correlación es muy alto, luego esto sugiere que hay una gran dependencia entre las variables, pero antes de saber si el ajuste es el adecuado vamos a comprobar lo que pasaría sin el dato atípico.



Coeficiente de Correlación = 0,957411

En este caso no ocurre como en el apartado anterior. Al omitir el dato atípico obtenemos un coeficiente de correlación bastante similar al anterior, por lo cuál podemos afirmar que un ajuste lineal es adecuado en este caso.

4.3. Calcular la recta de regresión de *Demandantes Empleo* sobre *Auxiliar Paro.*

Parámetro	*Mínimos Cuadrados Estimado*
Intercepto	-201,431
Pendiente	0,00633384

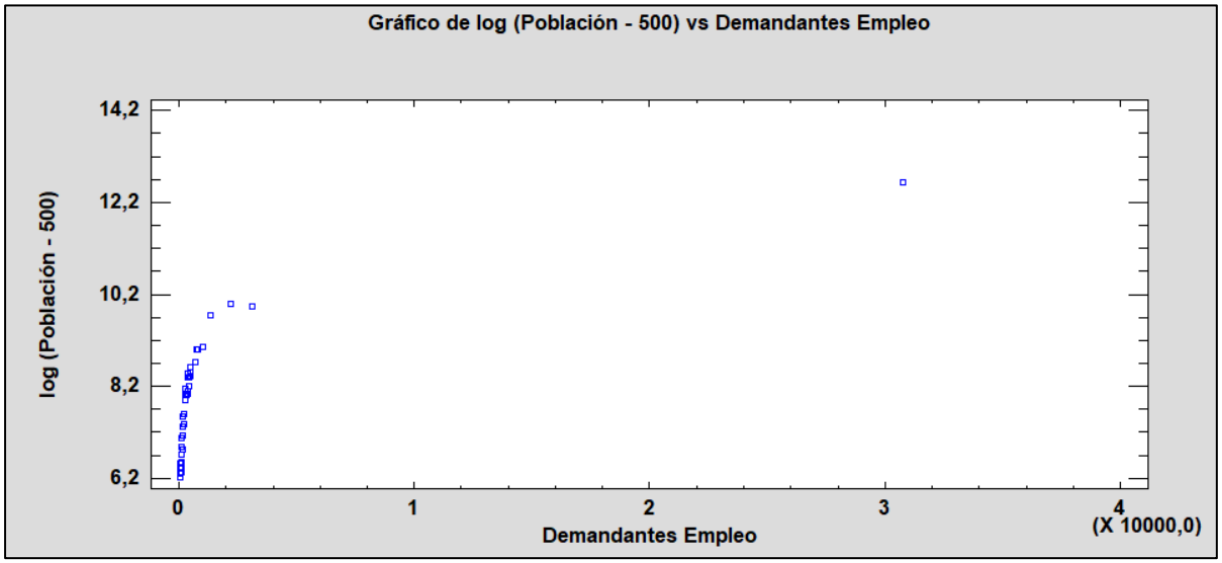
Recta: Demandantes Empleo = -201,431 + 0,00633384*Auxiliar Paro

4.4. ¿Qué número de demandantes de empleo se espera en un municipio con 1650 habitantes y una tasa de paro del 20 %? ¿Es adecuado el modelo? Justificar la respuesta.

- En la ecuación anterior sustituimos la variable Auxiliar Paro por 1650*20
- Demandantes de Empleo = 7,58572

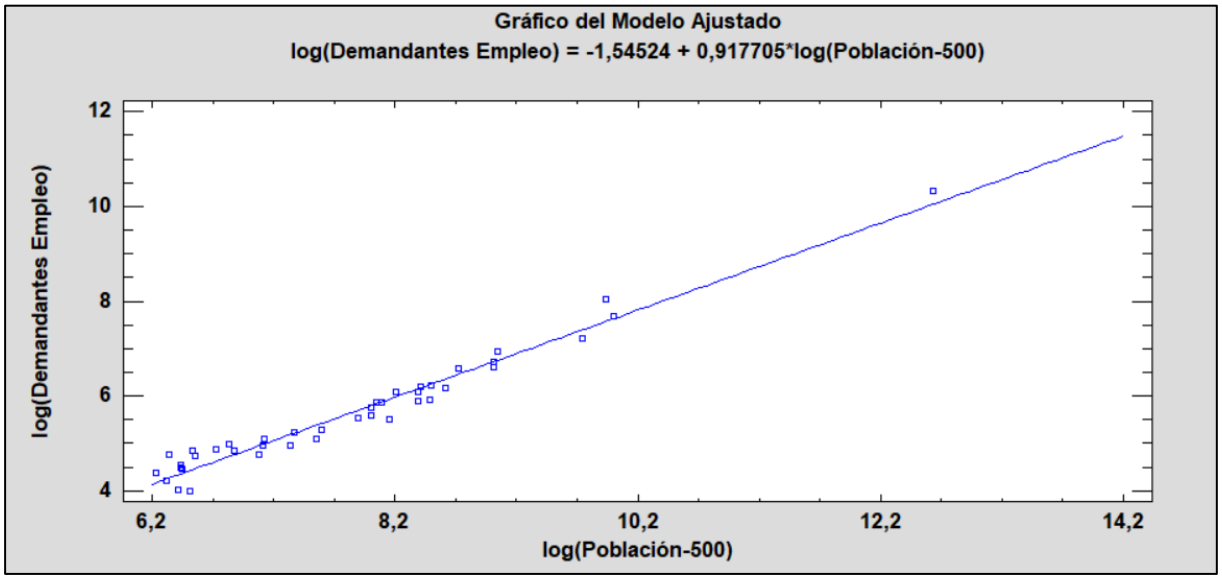
Coeficiente de Correlación = 0,958512
El coeficiente de correlación es alto (ha sido obtenido obviando el dato atípico que es Valladolid), así que el modelo si que es adecuado

4.5. Realizar el diagrama de dispersión de las variables *ln(Población-500)* y *Demandantes Empleo*. ¿Qué tipo de relación entre las dos variables sugiere el gráfico?



La gráfica nos sugiere que la relación entre las variables es de tipo logarítmica

4.6. Calcular la recta de regresión de *ln(Demandantes Empleo)* sobre *ln(Población-500).* ¿Qué número de demandantes de empleo se espera en un municipio con 2100 habitantes? ¿Y en un municipio de 1200 habitantes? Justificar la respuesta.



Recta de regresión: ln (Demandantes Empleo) = -1,54524 + 0,917705*ln (Población – 500)

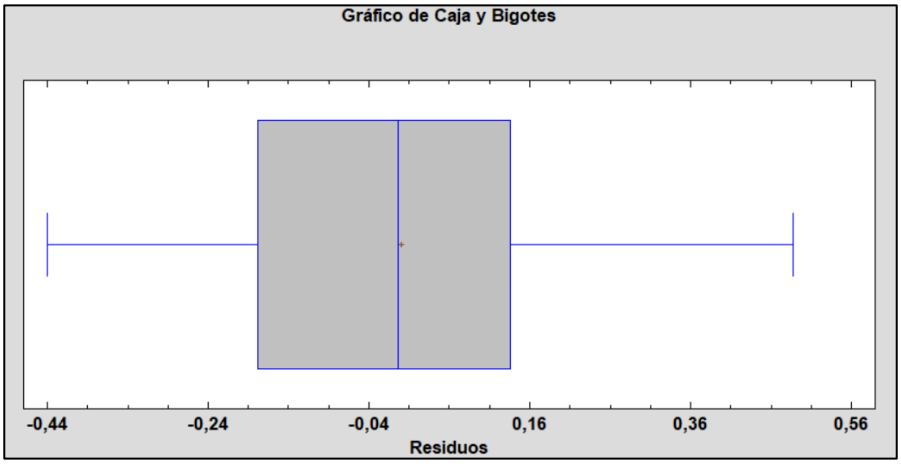
Para calcular el nº de demandantes de empleo que se esperan en un municipio de 2100 habitantes sustituimos el valor en la variable Población:

- Demandantes de Empleo = 186

Para los 1200 habitantes realizamos el mismo cálculo, sustituyendo el valor en la variable Población:

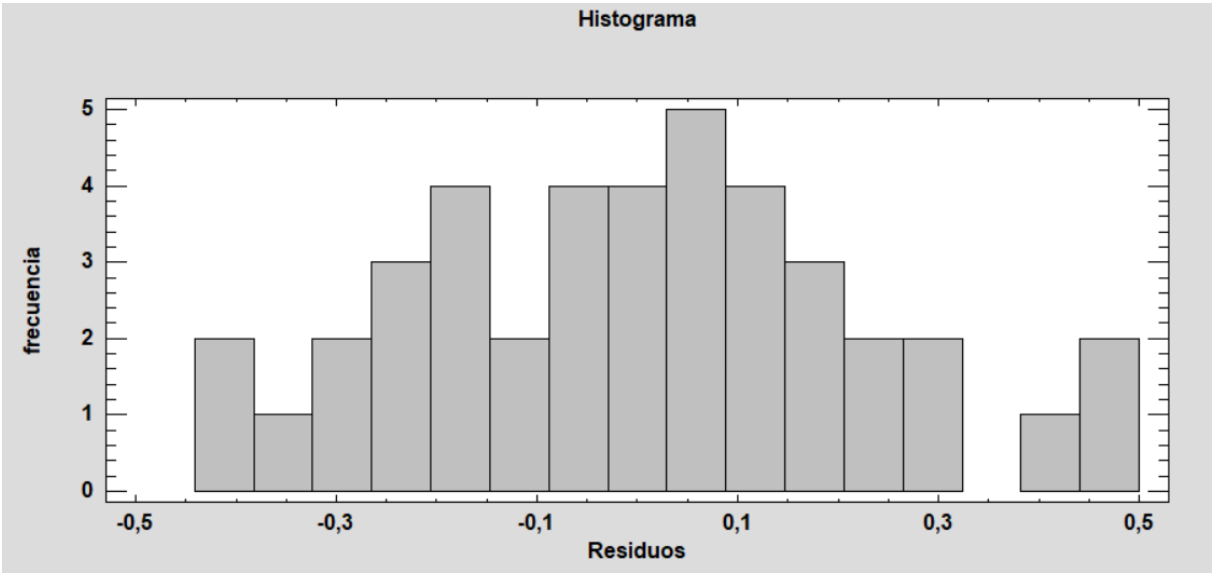
- Demandantes de Empleo = 87

4.7. Guardar en la variable *Residuos* el valor de los errores que se cometen cuando se sustituyen los valores reales de la variable *ln(Demandantes Empleo)* por los correspondientes a la recta de regresión calculada en el apartado anterior. Realizar un diagrama de caja y un histograma con esa variable. Identificar, si existen, los outliers e interpretar los resultados obtenidos.



Resumen Estadístico para Residuos	
Recuento	41
Promedio	0,00000223036
Desviación Estándar	0,22788
Coeficiente de Variación	1,02172E7%
Mínimo	-0,439561
Máximo	0,487217
Rango	0,926779

Como podemos ver por el gráfico, la variable residuos es centrada, su coeficiente de variación es muy bajo, es prácticamente simétrica y no tiene ningún valor atípico



ANEXO

Municipio	Partido	Poblaci6n	Tasa	Demandantes	Paro	CAg	CIn	CCo	CSe
Alaejos	Medina del Campo	1429	30.81	145	126	9	0	1	21
Aldeamayor de San Martin	Valladolid	4891	16.49	366	300	2	47	9	64
Arroyo de la Encomienda	Valladolid	17572	18.43	1366	1083	0	39	21	647
Boecillo	Valladolid	3989	5.42	245	204	42	27	4	278
Cabez6n de Pisuerga	Valladolid	3622	24.91	355	288	4	12	2	52
Campaspero	Valladolid	1174	14.24	55	45	176	2	0	27
Carpio	Medina del Campo	1068	25.58	117	99	30	1	0	16
Cigales	Valladolid	5008	36.22	497	410	3	7	23	82
Cisterniga	Valladolid	8734	16.67	749	606	2	46	18	179
Fuensalda:ia	Valladolid	1468	16.83	126	104	0	4	1	24
fscar	Valladolid	6678	19.53	718	558	53	49	7	126
Laguna de Duero	Valladolid	22555	29.99	2202	1787	12	30	43	220
Matapozuelos	Medina del Campo	1007	19.82	79	66	13	4	0	80
Mayorga	Medina de Rioseco	1687	15.98	116	89	4	0	3	34
Medina del Campo	Medina del Campo	21274	27.97	3115	2588	41	176	43	408
Medina de Rioseco	Medina de Rioseco	4906	14.26	441	351	6	23	2	137
Mojados	Medina del Campo	3184	14.85	252	198	11	10	8	92
Nava del Rey	Medina del Campo	2091	24.12	188	164	5	2	3	23
Olmedo	Medina del Campo	3759	13.52	352	264	13	281	8	206
Pedraja de Portillo, La	Valladolid	1134	25.8	86	73	1	0	0	6
Pedrajas de San Esteban	Valladolid	3503	22.41	316	264	17	29	3	182
Pe:iafiel	Valladolid	5428	16.83	508	414	34	61	5	68
Portillo	Valladolid	2409	17.26	162	130	170	9	4	31
Quintanilla de Onesimo	Valladolid	1109	9.39	56	45	3	5	0	32
Renedo de Esgueva	Valladolid	3507	27.07	267	206	0	0	21	48
Rueda	Medina del Campo	1332	15.31	131	113	14	10	2	10
Santovenia de Pisuerga	Valladolid	4155	21.55	439	359	0	30	14	100
Seca, La	Medina del Campo	1127	17.86	94	80	6	18	0	17
Serrada	Medina del Campo	1184	22.32	126	100	22	12	5	21
Simancas	Valladolid	5331	17.24	368	289	2	3	1	115
Tordesillas	Valladolid	8973	24.12	1045	832	131	12	34	253
Traspinedo	Valladolid	1126	27.08	90	78	8	0	2	38
Tudela de Duero	Valladolid	8717	31.37	831	710	19	8	27	153
Valdestillas	Medina del Campo	1742	28.37	164	143	1	2	4	117
Valladolid	Valladolid	306830	15.73	30755	24926	322	2103	674	8935
Viana de Cega	Valladolid	2031	27.52	142	120	2	1	1	36
Villabragima	Medina de Rioseco	1054	18.21	68	61	1	0	1	10
Villal6n de Campos	Medina de Rioseco	1733	18.71	141	107	4	0	3	20
Villanubla	Valladolid	2484	9.53	197	161	2	10	16	121
Villanueva de Duero	Medina del Campo	1202	34.67	114	95	18	0	1	4
Zaratan	Valladolid	6029	17.46	479	399	3	7	8	210