UNIVERSITAT POLITÈCNICA DE VALÈNCIA

INFORME ESTADÍSTICO SOBRE LOS DIAMANTES

Ismael Fernández Herreruela

Grupo D

Base de datos

Se trata de una base de datos sobre diamantes, incluyendo dos variables cualitativas (corte y color) y 4 continuas (largo, ancho, profundidad y quilates).

He obtenido esta base de datos del siguiente link:
https://www.kaggle.com/datasets/shivam2503/diamonds

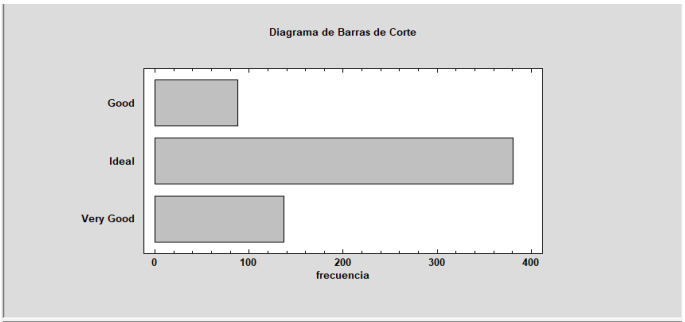
El corte es la calidad del corte del diamante, siguiendo el siguiente orden de calidad: bueno, premium, ideal; el color del diamante indica la calidad del color del diamante desde D (el mejor) hasta I (el peor); el quilate indica el peso del diamante (unos 0,2 gramos) y las características largo, ancho y profundo (en el link x, y, z) miden como de largo, ancho y profundo son los diamantes en milímetros.

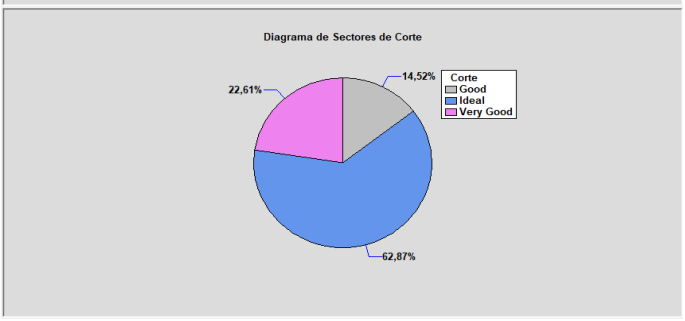
He seleccionado 607 elementos de la base de datos para trabajar con ellos. He juntado las variables de corte “Fair” en “Good” y la de “Premium” en “Ideal”, y por otro lado he juntado sobre la variable de color, desde la “J” hasta la “G” en “G”. Todo esto lo he hecho para poder trabajar mejor con los datos de forma más cómoda y que tengan un tamaño parecido a las otras variables.

Como variable F1 utilizare el Corte y como variable F2 utilizare el Color.

EJERCICIOS

Actividad 5





Es similar entre Good y Very Good, pero como podemos observar, predominan los diamantes con un corte Ideal.

Actividad 6

Tabla de Frecuencia para Color

			Frecuencia	*Frecuencia*	*Frecuencia*
Clase	*Valor*	*Frecuencia*	*Relativa*	*Acumulada*	*Rel. acum.*
1	D	63	0,1040	63	0,1040
2	E	97	0,1601	160	0,2640
3	F	117	0,1931	277	0,4571
4	G	329	0,5429	606	1,0000

Al observar los datos de la tabla, podemos ver que los datos se distribuyen en dos mitades con porcentajes similares. Con mayor porcentaje tenemos G que son todos los colores de calidad media a baja, y el resto de los porcentajes formarían los colores de calidad media a alta.

Actividad 7

Tabla de Frecuencias para Corte por Color

	D	E	F	G	Total por Fila
Good	11	11	14	52	88
	1,82%	1,82%	2,31%	8,58%	14,52%
Ideal	39	58	72	212	381
	6,44%	9,57%	11,88%	34,98%	62,87%
Very Good	13	28	31	65	137
	2,15%	4,62%	5,12%	10,73%	22,61%
Total por Columna	63	97	117	329	606
	10,40%	16,01%	19,31%	54,29%	100,00%

En la tabla podemos observar que la combinación más frecuente de Corte y Color es la G Ideal con un 34,98%. Pero por otro lado la menos común es D Good con solo un 1,82%.

La frecuencia absoluta se trata del conteo de la cantidad exacta de veces que un dato se repite mientras que la relativa consiste en mostrar el porcentaje de la frecuencia absoluta respecto al total.

Actividad 8

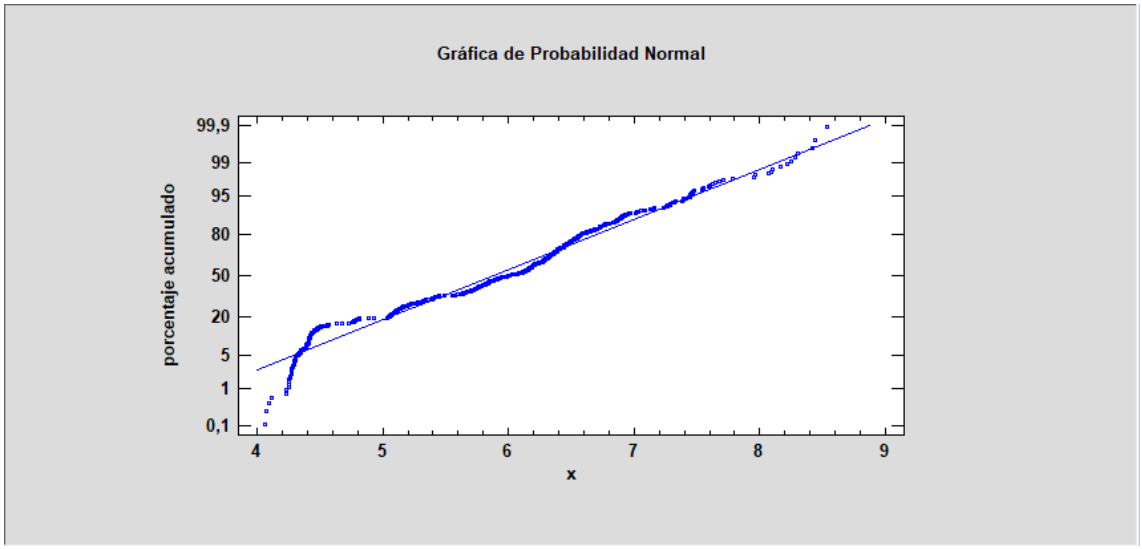
Resumen Estadístico

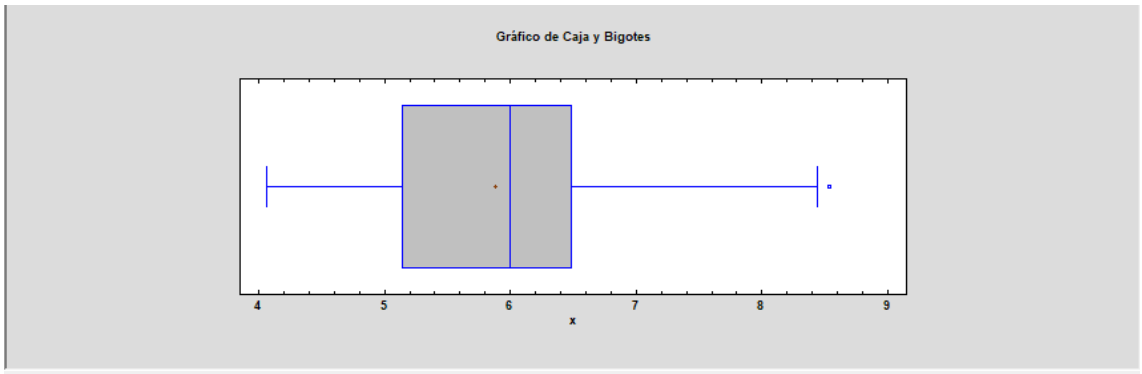
	x	*y*	*z*	*Quilates*
Recuento	606	606	606	606
Promedio	5,87848	5,87568	3,63398	0,832409
Mediana	6,0	6,02	3,72	0,83
Media Geométrica	5,79943	5,79777	3,58574	0,741057
Varianza	0,908811	0,894566	0,341667	0,151347
Desviación Estándar	0,953316	0,945815	0,584523	0,389034
Coeficiente de Variación	16,217%	16,0971%	16,0849%	46,7359%
Mínimo	4,06	4,05	2,55	0,26
Máximo	8,54	8,51	5,29	2,34
Rango	4,48	4,46	2,74	2,08
Rango Intercuartílico	1,34	1,33	0,83	0,51
Sesgo Estandarizado	0,169543	0,0652404	-0,224021	7,83294
Curtosis Estandarizada	-2,32269	-2,35301	-2,5946	4,74727

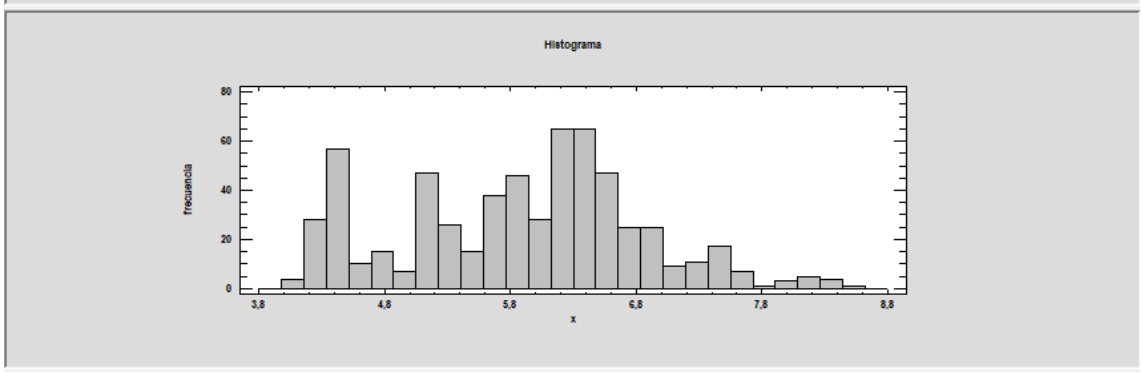
En la tabla tenemos x como Largo, y como Alto y z como Profundidad del diamante.

Tambien hay que tener en cuenta que la mediana de los Quilates es inferior a las demas debido a unos datos anomalos.

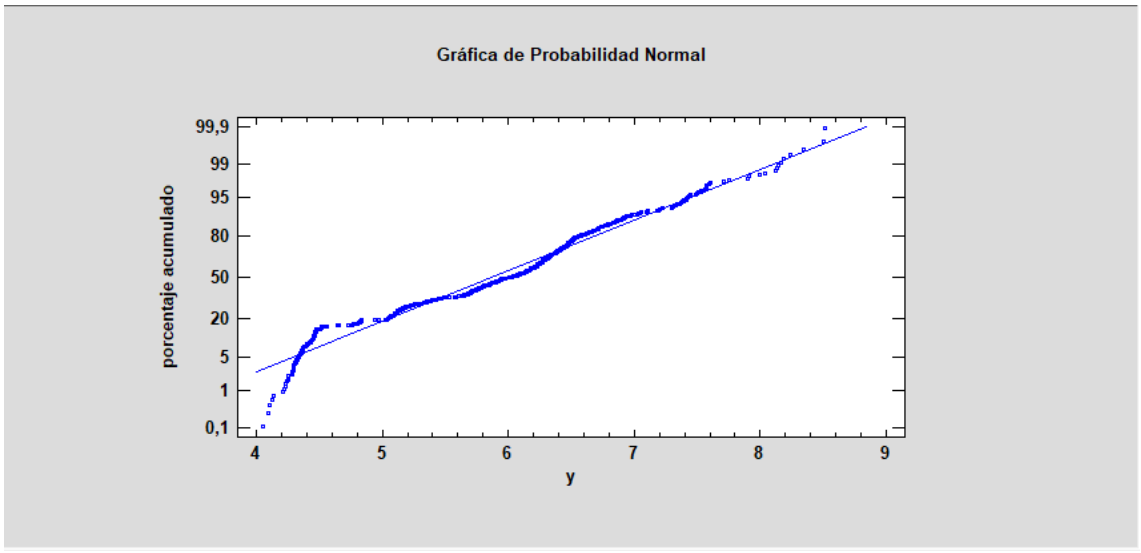
Actividades 9/10/11

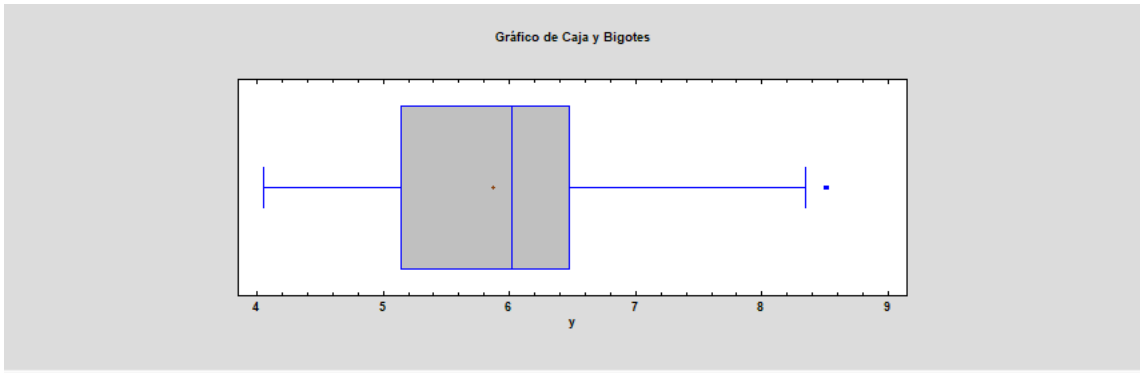


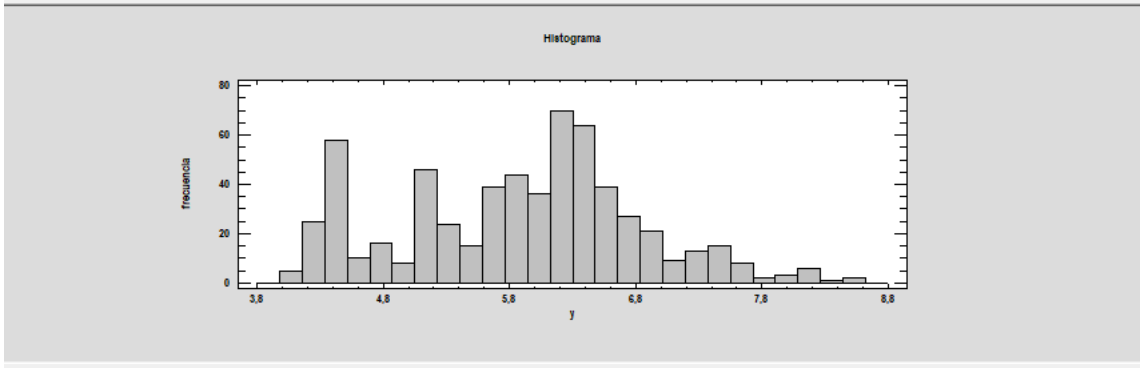




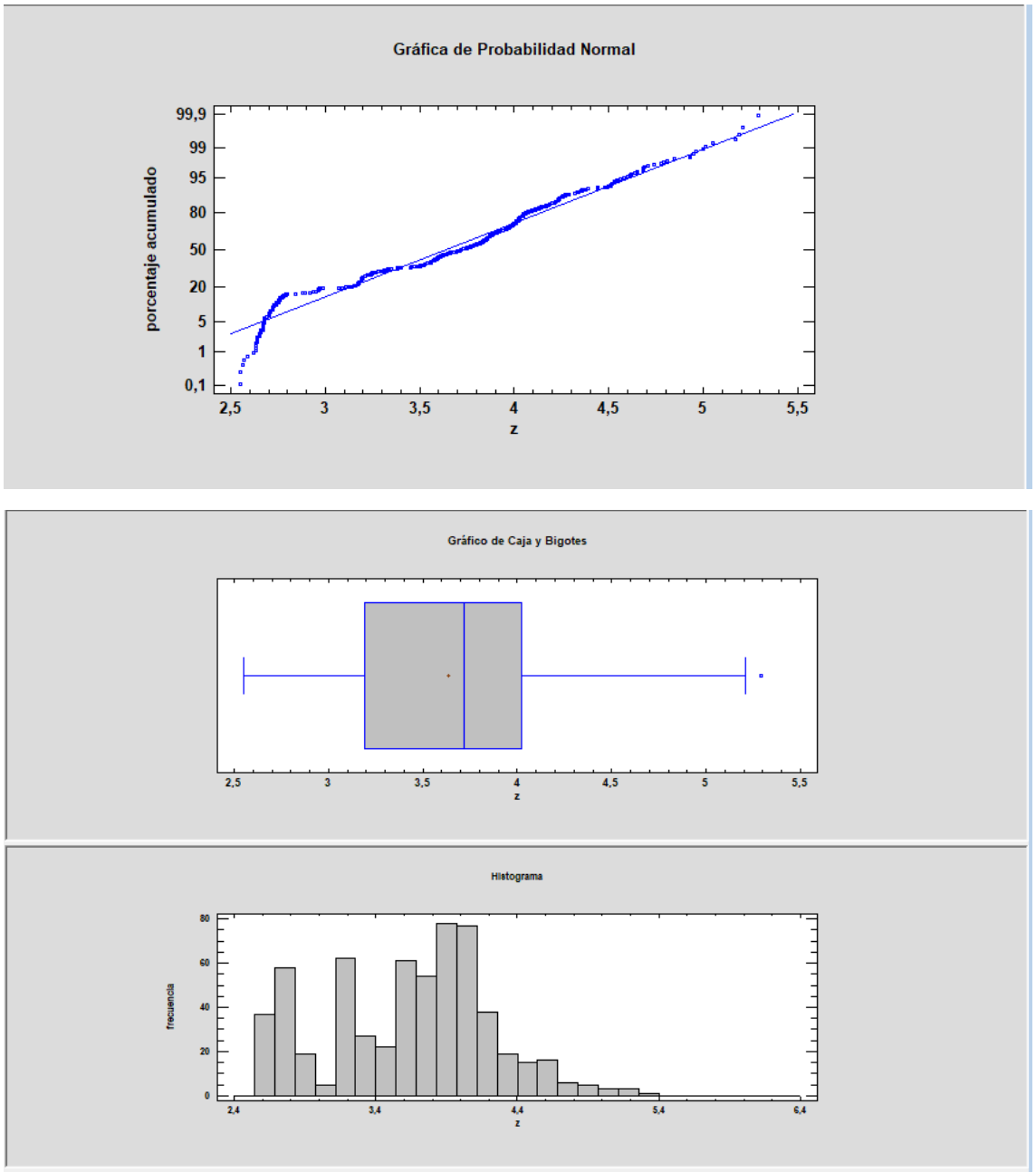
En cuanto a la x obtenemos un coeficiente de asimetria de 0,169543 y un coeficiente de curtosis de -2,32269. El coeficiente de asimetria si que entra entre los limites -2<x<2 pero el coeficiente de curtosis se sale por 0,32 aproximadamente.



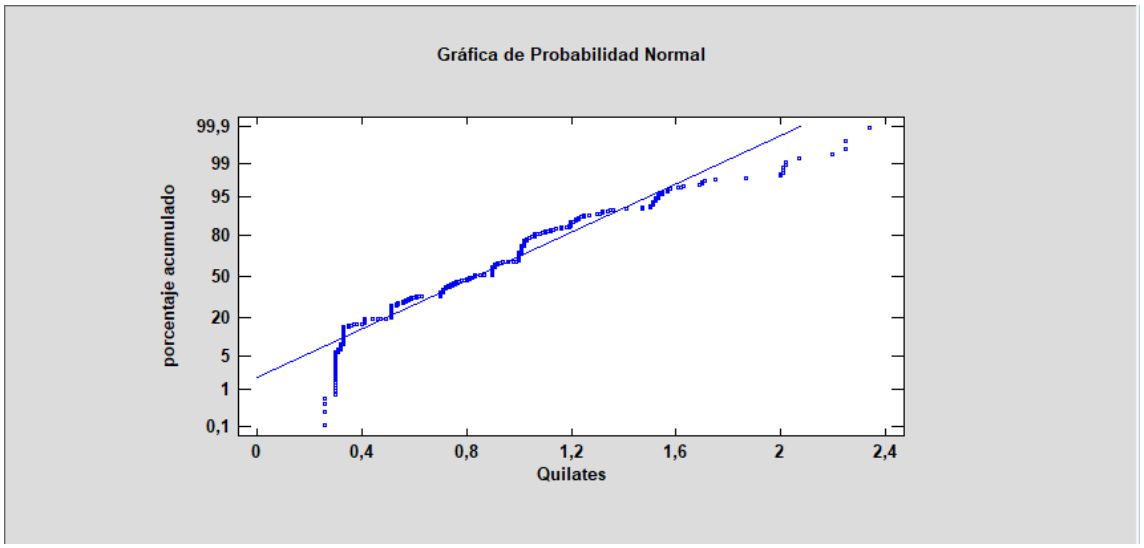


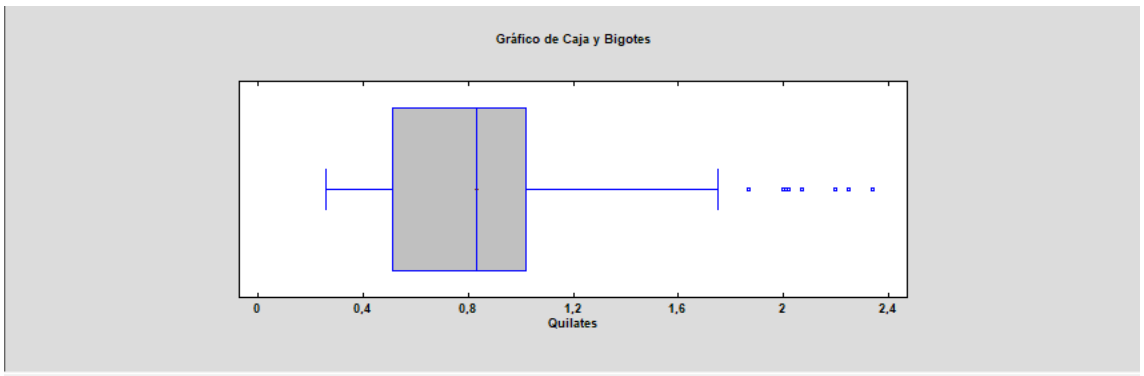


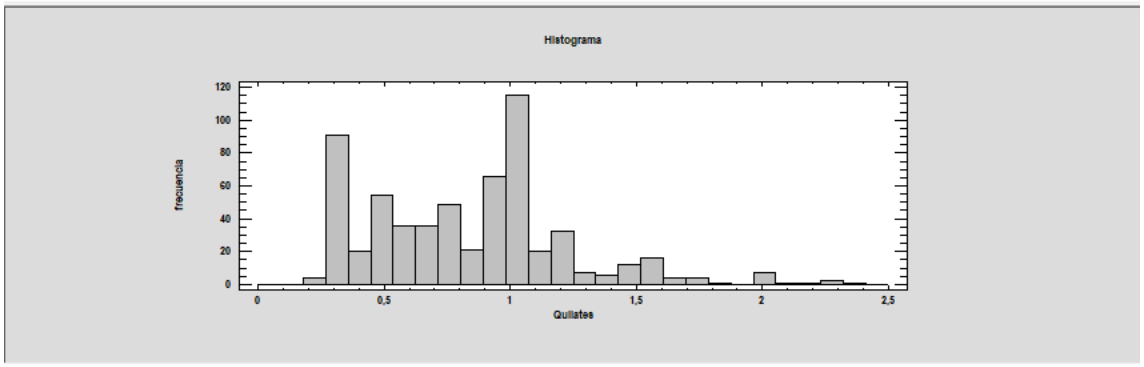
En cuanto a la y obtenemos un coeficiente de asimetria de 0,0652404 el cual entra dentro del requisito -2<x<-2 pero el coeficiente de curtosis es -2,35301 y se sale por 0,35 de ese limite.



En z obtenemos un coeficiente de asimetria de -0,224021 y un coeficiente de curtosis de -2,5946. Nos vuelve a ocurrir lo mismo, el coeficiente de asimetria entra en los limites pero el coeficiente de curtosis no.







Y en ultimo lugar, la variable Quilates nos da un coeficiente de asimetria de 7,83294 y un coeficiente de curtosis de 4,74727. Ambos coeficientes sobrepasan los limites mencionados anteriormente.

Tras realizar todos estos analisis sobre las variables x,y,z y Quilates podemos observar que los histogramas nos dan una informacion mas clara sobre si hay datos anomalos pero por otro lado los histogramas nos aportan unos datos mucho mas detallados. Tras haber analizado todos los graficos, desde mi punto de vista Quilates es el que mas informacion nos puede aportar. Podemos observar que su punto de algidez es el “1” y que tiene diversos datos anomalos en el grafico de caja y bigotes. Tambien podemos observar que en las variables x,y,z, en sus respectivos graficos de cajas y bigotes, son todos muy parecidos con un dato anomalo en la derecha. Hay que resaltar tambien que tienen unas graficas de probabilidad muy similares.

Finalmente, esta sera mi asignación de variables:

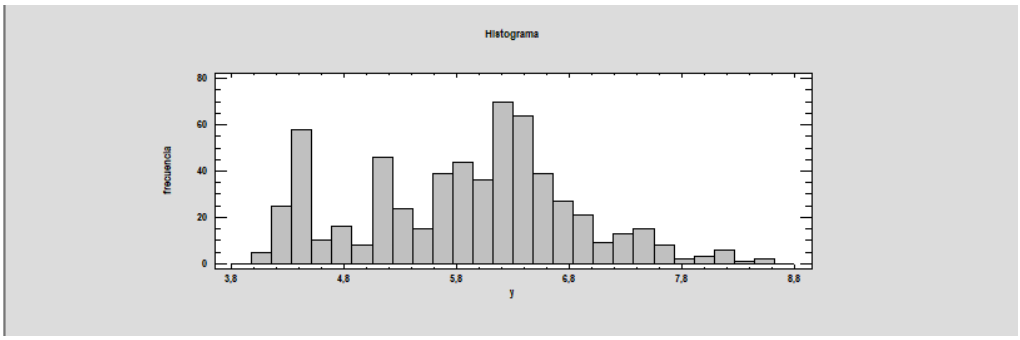
X1 -> x

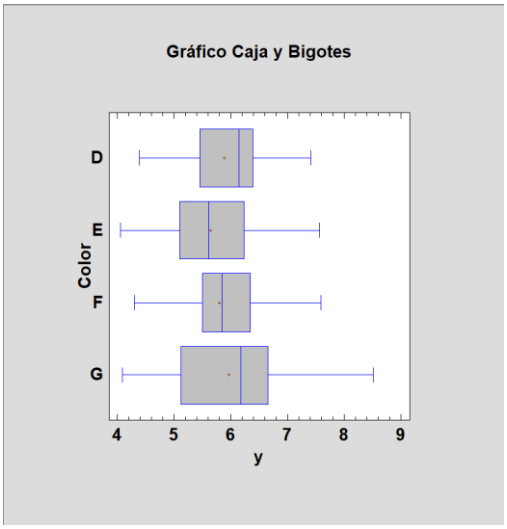
X2 -> y

X3 -> z

X4 -> Quilates

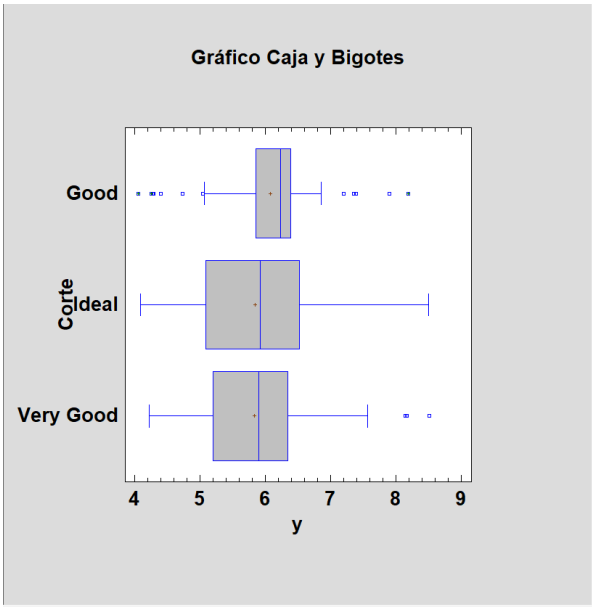
Actividad 12





El grafico de caja y bigotes multiple consiste en realizar un grafico de una variable x sobre otra variable y. Dependiendo de cuantas opciones existan de la variable y, saldran mas o menos graficos. En este caso hacemos el grafico de la variable y sobre la variable Color, y al haber 4 tipos de color nos salen 4 graficos resultantes. No hay datos anomalos.

Actividad 13



Resumen Estadístico para y

Corte	*Recuento*	*Promedio*	*Mediana*	*Desviación Estándar*	*Mínimo*	*Máximo*
Good	88	6,08114	6,235	0,710463	4,05	8,19
Ideal	381	5,8415	5,92	1,00902	4,09	8,5
Very Good	137	5,83876	5,9	0,883128	4,22	8,51
Total	606	5,87568	6,02	0,945815	4,05	8,51

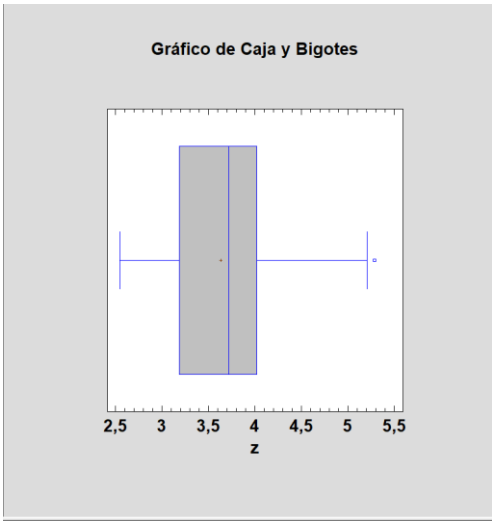
Corte	*Rango*	*Cuartil Inferior*	*Cuartil Superior*	*Rango Intercuartílico*
Good	4,14	5,855	6,385	0,53
Ideal	4,41	5,09	6,52	1,43
Very Good	4,29	5,2	6,34	1,14
Total	4,46	5,14	6,47	1,33

Observando los datos de la tabla, podemos ver que Good tiene una mayor mediana, Ideal tiene un mayor Rango Intercuartilico. Good tiene una asimetria negativa fuerte, Ideal y Very Good tienen una asimetria negativa moderada.

Actividad 14

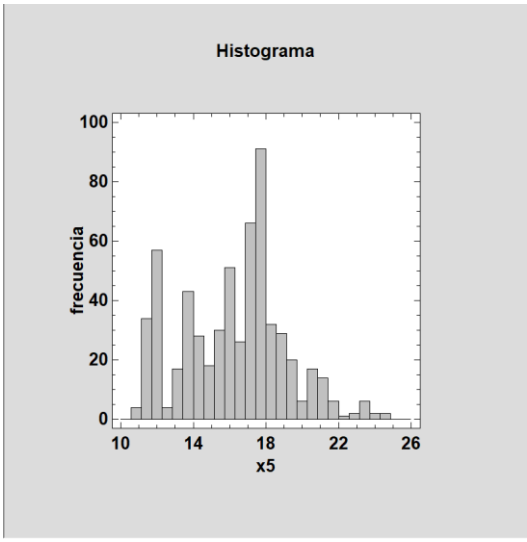
En el caso de x2, al tener una distribución sesgada, en cuanto a la posicion deberiamos usar la mediana ya que nos dara unos mejores resultados ya que lo que queremos obtener es la ubicación central de los datos. Y en el caso de la dispersión, queremos obtener la variabilidad de los datos por lo que al tener una distribución sesgada, seria mejor utilizar el rango intercuartílico. Hay que resaltar que dependiendo de las caracteristicas de los datos, unas veces es mejor utilizar es

Actividad 15



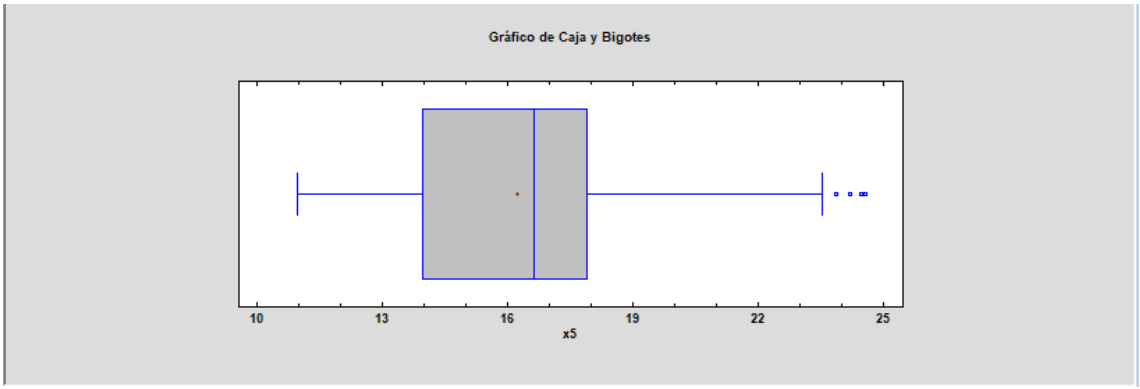
El diagrama de caja y bigotes es mejor que el histograma o que el papel probabilistico. El histograma por ejemplo es un diagrama poco eficiente para detectar datos anomalos, en cambio el diagrama de cajas y bigotes se ven perfectamente ya que son puntitos que estan separados de los bitotes. En este caso se ver perfectamente el dato anomalo siendo el punto ubicado a la derecha del grafico. Tambien podemos obtener datos como la mediana(punto central en el interior de la caja) y la media(linea del interior de la caja).

Actividad 16



Resumen Estadístico para x5

Recuento	606
Promedio	16,2205
Desviación Estándar	2,8591
Coeficiente de Variación	17,6264%
Mínimo	10,97
Máximo	24,58
Rango	13,61
Sesgo Estandarizado	0,936743
Curtosis Estandarizada	-1,82657



Se trata de una distribucion normal ya que al realizar el coeficiente de curtosis y el coeficiente de asimetria, ambos valores se encuentran dentro del limite -2<x<2. Tambien se trata de una asimetria negativa fuerte ya que si vemos su diagrama de cajas y bigotes vemos que la caja se encuentra mas hacia la izquierda, y la parte derecha de la caja es significativamente mas pequeña que la izquierda.