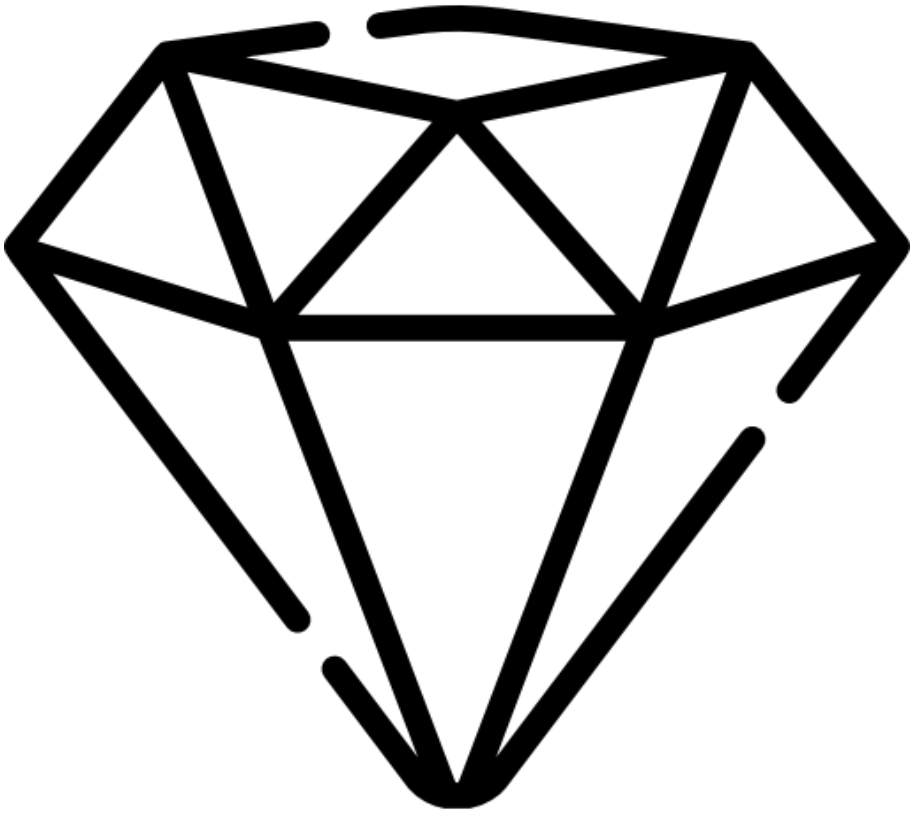
Informe estadístico sobre los Diamantes



Hecho por Ángel Giménez González

En este trabajo he usado los datos de la tabla mencionada en este link: **https://www.kaggle.com/shivam2503/diamonds** al ser una tabla de datos que contenía la cantidad suficiente de datos a analizar, siendo estas; dos variables cualitativas (el corte y el color), y cuatro variables continuas (el largo, el ancho, la profundidad y el carat).

El corte indica la calidad del corte del diamante, siendo en orden de calidad bueno, premium y ideal; el color del diamante indica la calidad del color del diamante, siendo D el mejor y I el peor; el carat indica el peso del diamante, medido este en quilates (unos 0,2 gramos) y las caracteristicas largo, ancho y profundo (en el link x, y, z) miden como de largo, ancho y profundo son los diamantes en milimetros.

Para adaptar más facilmente los datos he seleccionado 200 casos aleatorios de la tabla que son con los que voy a maniobrar, y he juntado las tres variables más malas de corte (fair, good y very good) dado que apenas habia casos particulares de cada una y las he mezclado en good para equilibrarlas en porcentaje con ideal y premium; por el mismo motivo también he mezclado las dos variables que representan los peores colores (I y J) y los he juntado en I para que ambos tuvieran un tamaño comparable al resto de variables.

Act 1)

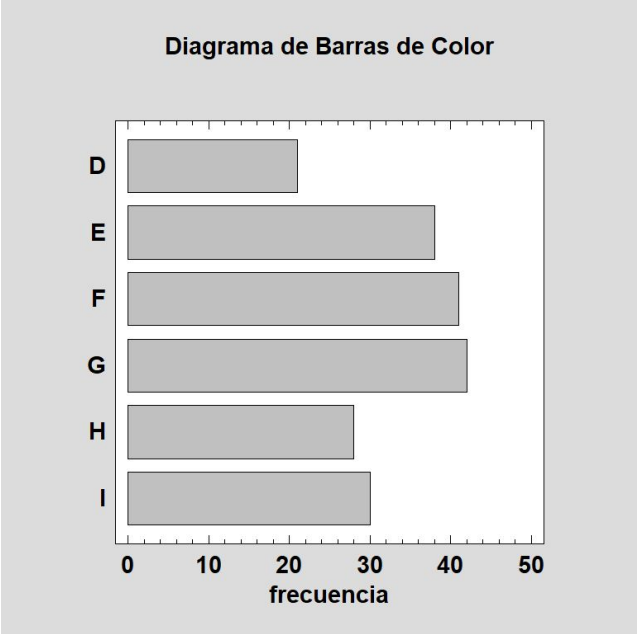
Tabla de Frecuencia para Color

Clase	*Valor*	*Frecuencia*	*Frecuencia Relativa*	*Frecuencia Acumulada*	*Frecuencia Rel. acum.*
1	D	21	0,1050	21	0,1050
2	E	38	0,1900	59	0,2950
3	F	41	0,2050	100	0,5000
4	G	42	0,2100	142	0,7100
5	H	28	0,1400	170	0,8500
6	I	30	0,1500	200	1,0000

(--Tabla 1

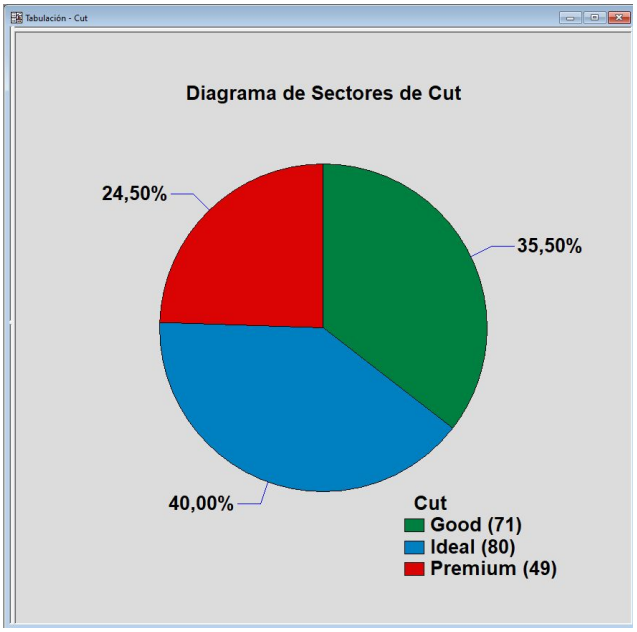
Al ver estos resultados, podemos ver claramente que la frecuencia de colores se divide perfectamente entre la primera y la segunda mitad, indicando que de un diamante aleatorio las posibilidades de elegir un diamante de calidad de color bueno son las mismas que de que salga un color de calidad mala, (Siendo, en orden descendiente de calidad, las buenas: D, E y F y las malas: G, H e I).

Act 2) Fig.1) Gráfico de Barras de la variable F_2, Color.



Al ver el gráfico de barras queda claro que las calidades más comunes de la tabla son la E, F y G, y por ello al cojer un diamante cualquiera hay mas posibilidades de cojer un diamante con una de esas 3 calidades que de cojer una con cualquiera de las otras 3 (Siendo estas calidades D, H e I).

Act 3) Fig.2) Gráfico de tartas de la variable F_1, Cut.



Como se puede apreciar el este gráfico de tartas, las posibilidades de que eligiendo un diamante aleatorio y su corte sea definido “good”, son practicamente de un tercio, lo esperable teniendo 3 variables, sin embargo, la constante “ideal” es muy superior a un tercio, por mas de 6 puntos, a costa de la variable ”premium”, que es significativamente inferior al tercio; por lo cual la más común es Ideal, y la menos común es Premium.

Act 4)

Tabla de Frecuencias para Cut por Color

	D	E	F	G	H	I	Total por Fila
Good	5	18	16	14	9	9	71
	2,50%	9,00%	8,00%	7,00%	4,50%	4,50%	35,50%
Ideal	9	10	19	20	11	11	80
	4,50%	5,00%	9,50%	10,00%	5,50%	5,50%	40,00%
Premium	7	10	6	8	8	10	49
	3,50%	5,00%	3,00%	4,00%	4,00%	5,00%	24,50%
Total por Columna	21	38	41	42	28	30	200
	10,50%	19,00%	20,50%	21,00%	14,00%	15,00%	100,00%

Contenido de las celdas:
Frecuencia Observada
Porcentaje de la Tabla

(--Tabla 2

En la tabla se muestran datos importantes, aparte de datos que ya hemos desarrollado en las tres primeras actividades, un ejemplo es que la combinación de calidad de corte y color más común en la tabla es el G-Ideal, situado próximo al centro de la tabla en ambos ejes, mientras que el menos común se situa en el D-Good, situado al extremo de la tabla en ambos ejes, caso que se da por toda la tabla, en la que conforme te alejas del centro de la tabla en cualquier dirrección, tiendes a tener menos casos, dando a entender que las frecuencias tienen una distribución normal.

La diferencia entre frecuencia absoluta y relativa es que la absoluta mide la cantidad de veces exacta que un dato se repite, mientras que la relativa dice el porcentaje de la frecuencia absoluta respecto al total. La frecuencia marginal mide todos los datos de una fila o columna respecto al total y la condicional mide los datos de una intersección concreta.

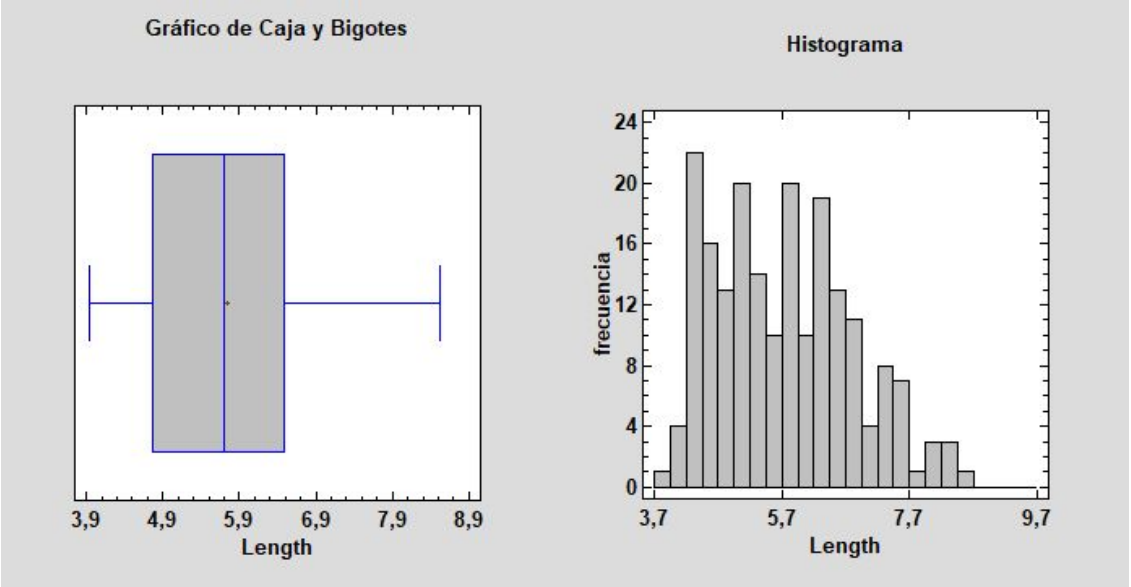
Act 5)

Resumen Estadístico para	Length	Width	Depth	Carat
Recuento	200	200	200	200
Promedio	5,74245	5,7449	3,5588	0,8014
Mediana	5,71	5,74	3,535	0,71
Varianza	1,16716	1,14794	0,463421	0,214941
Desviación Estándar	1,08035	1,07142	0,68075	0,463617
Mínimo	3,94	3,9	2,47	0,23
Máximo	8,53	8,46	5,47	2,3
Rango	4,59	4,56	3,0	2,07
Rango Intercuartílico	1,735	1,7	1,065	0,61
Sesgo Estandarizado	2,38355	2,28392	2,49143	6,44407
Curtosis Estandarizada	-1,79468	-1,80154	-1,50323	2,69082

(--Tabla 3

La mediana de la variable Carat es significativamente menor que la media (mas de un 10% de diferencia) debido a que hay unos cuantos datos anomalos que son demasiado altos respecto al resto dentro de la variable.

Act 6) Fig.3) Gráfico de caja-bigotes y histograma de la variable X_1, Length.



En el histograma se indica el numero de diamantes cuya longitud coincide con los rangos mostrados en la parte inferior.
En el gráfico de caja y bigotes se pretende representar un rango de longitud y diferenciar entre las zonas en las que hay más datos concentrados respecto a las que los tienen más dispersos.

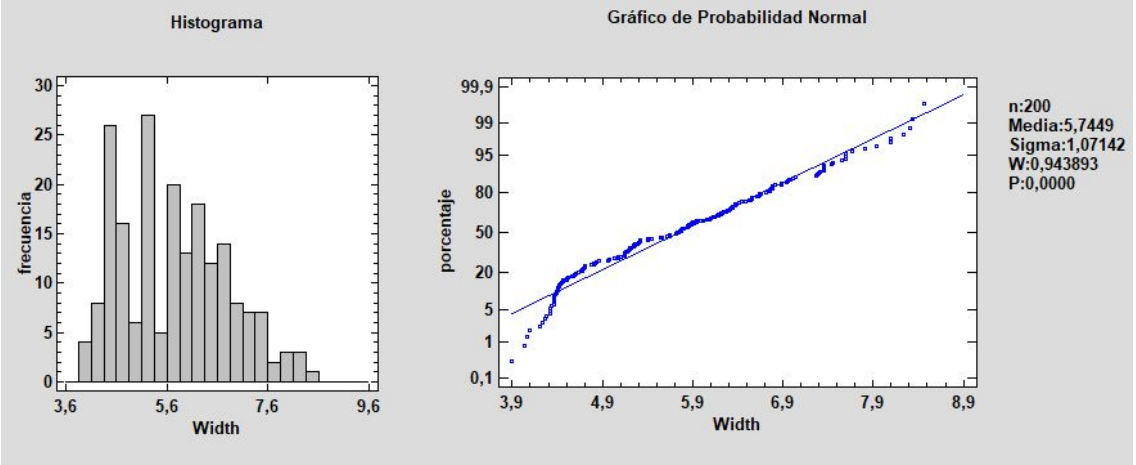
Como se puede facilmente observar en ambas gráficas, la distribución es asimetrica moderada, concretamente asimetrica moderada positiva, ya que el coeficiente de asimetria es positivo (La caja está muy hacia la izquierda).

No se aprecian datos anomalos a simple vista, porque no hay ningún dato que no encaje en el gráfico de caja y bigotes.

Al ver como estan mostrados los datos, y como el coeficiente de asimetria se sale por poco del rango -2<x<2 (siendo este 2,38355) y como el coeficiente de curtosis entra en el rango -2<x<2 (siendo este -1,79468), se podría hacer el argumento de que es razonable considerar que es una distribución normal, aunque yo no lo consideraría como tal al salirse casi 4 decimas del rango.

Con el histograma tienes la desventaja de no poder saber si hay datos anomalos (e este caso ninguno), al contrario que con el de caja y bigotes, pero en el histograma tienes los datos mucho más detallados al contrario que en el de caja y bigotes que solo te indica donde está la concentración de la mayoría de datos, cosa que también puedes ver igualmente clara en el histograma, por lo que el histograma aporta información más útil.

Act 7) Fig.4) Histograma y Papel probabilístico normal de la variable X_2, Width.



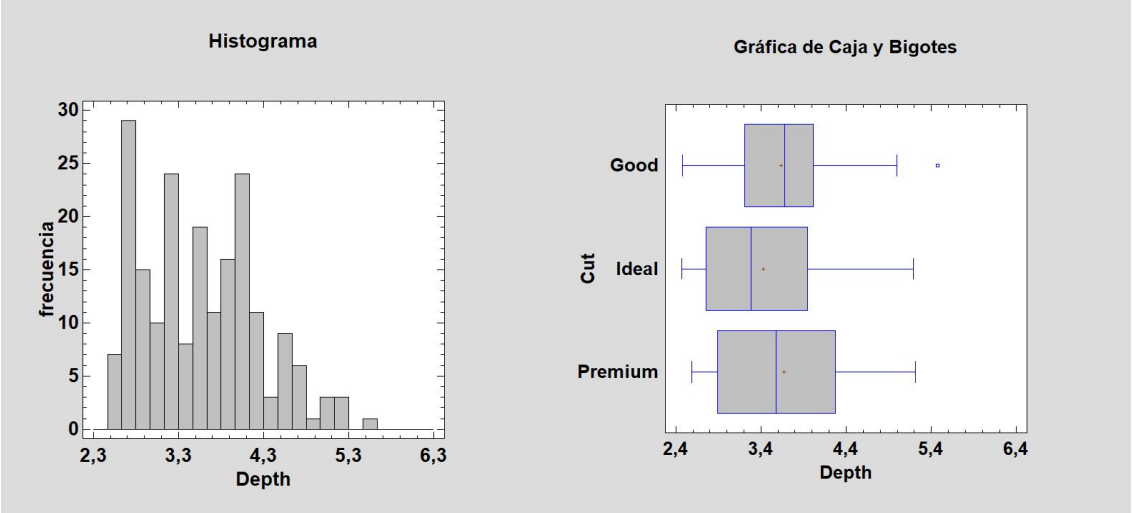
El papel probabilístico normal te indica que distribución sigue tus datos, comparar la distribución con la distribución normal y también te sirve para comprobar si hay datos anomalos.

Como se puede facilmente observar en ambas gráficas, la distribución es asimetrica moderada, concretamente asimetrica moderada positiva, ya que el coeficiente de asimetria es positivo (La aproximación de la campana de gauss del histograma está muy hacia la izquierda).

No se aprecian datos anomalos a simple vista ni en el histograma ni en el papel probabilistico normal.

Al ver como estan mostrados los datos, y como el coeficiente de asimetria se sale por poco del rango -2<x<2 (siendo este 2,28392) y como el coeficiente de curtosis entra en el rango -2<x<2 (siendo este -1,80154), puede ser razonable considerar que es una distribución normal, aunque yo no lo consideraría como tal debido a que al principio del papel probabilístico normal hay muy pocos datos.

Act 8) Fig.5) Histograma y Gráfico de caja-bigotes múltiple de la variable X_3, Depth.



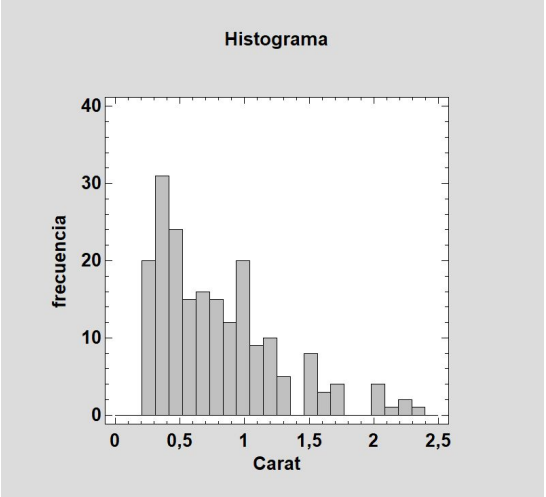
El gráfico de caja y bigotes múltiple es esencialmente hacer varios gráficos de caja y bigotes, separando los datos de la variable X_3 dependiendo de a que categoría de F_1 pertenecen.

Como se puede observar en el gráfico de caja y bigotes múltiple, las variables Ideal y Premium son asimétrica fuerte positiva y asimtrica moderada positiva respectivamente, al estar sus cajas muy infuenciadas hacia la izquierda, mientras que la variable Good es asimétrica leve positiva, estando bastante cerca de ser considerada simétrica, estando su caja solo levemente hacia la izquierda del centro.

Gracias a separar los datos en las tres variables de F_1 se puede observar que en la variable Good hay un dato anomalo que claramente se sale del bigote por el lado derecho, dato que en el histograma también se puede observar al estar separado del resto en la derecha.

Como se puede ver al ver los gráficos los coeficientes de asimetría y curtosis, tanto el Good (1,43631 ; 1,77228) como el Premium (0,780678 ; -1,7737) se podría asumir un modelo de distribución normal estando ambas variables en el rango [-2<x<2]; al contrario, la variable Ideal (2,31647 ; -0,948394) no sería razonable considerarla al salirse el coeficiente de asimetría del rango [-2<x<2].

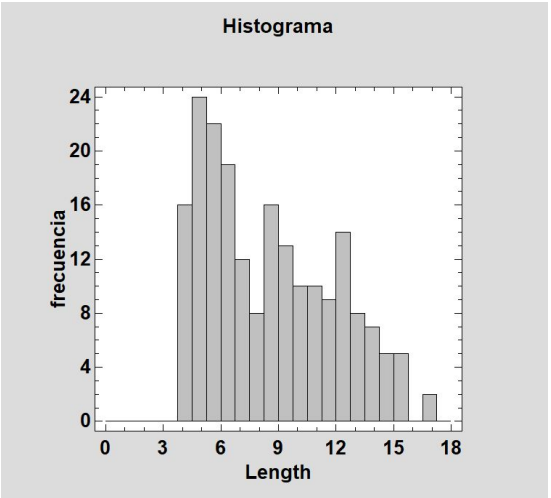
Act 9) Fig.6) Histograma de la variable X_4, Carat.



He elegido este gráfico porque tras hacer los ejercicios 6, 7 y 8 y poder ver los pros y contras de cada gráfico, he determinado que el histograma es el que aporta más información sin sacrificar practicamente nada a cambio.

Lo primero a destacar es que la variable posee una fuerte asimetría positiva ya que crece rapido al principio hasta llegar al punto algido y decrece a partir de ese punto álgido de forma lenta y progresiva, asi que por ello se descarta una distribución normal; otra cosa a destacar es la gran cantidad de variables anomalas que hay, ya que a partir de el valor 1,5 hay dos grupos de numeros separados entre ellos y del resto, influenciando la asimetria.

Act 10) Fig.7) Histograma de la variable N, Length con la mitad de sus números multiplicados por 2.



En el histograma queda claro que la forma del gráfico no ha cambiado mucho (sigue siendo asimetrica), simplemente se han cogido multiples variables del grafico y se han extendido pasado el mismo, y ahora en vez de parar alrededor del 8 continua hasta el 16-17; el mayor cambio es el hecho de que se a creado dos variables anomalas separadas del resto en el 17, al haber coincidido la multiplicación aleatoria con unos de los números más grandes.

Act 11)
Dado que no dispongo de ninguna variable discreta, no me es posible realizar este ejercicio concreto.

Act 12)
Las cuatro distribuciones de mis datos no siguen ni una distribución binomial ni de poisson ya que son continuas, tampoco siguen ni una exponencial ni una uniforme. Carat al no seguir una normal no sigue ninguna distribución que hayamos dado. Las variables Length, Width y Depth se puede asumir que tienen una distribución normal debido a que su curtosis estandarizada esta dentro del rango [-2<x<2] y el coeficiente de asimetría estandarizado está próximo del rango.

Act 13)
Tras aplicar logaritmos tipo Ln en todas las variables los resultados en el coeficiente de asimetría y en el de curtosis estandarizada son practicamente iguales entre si. Por ello y como en el ejercicio anterior, se puede asumir que tienen una distribución normal debido a que su coeficiente de asimetría esta dentro del rango [-2<x<2] y su curtosis estandarizada está próxima del rango por una diferencia de alrededor de media unidad.

Act 14)
Ninguna sigue una distribución normal, si bien las variables Length, Width y Depth se asemejan bastante a una como ya ha sido mencionado en actividades previas.

Act 15)
Tras crear las mencionadas distribuciones normales y sumarlas ordenadas por parejas, el resultado es otra distribución normal (como cabe a esperar por las propiedades de la normal, que implican que cualquier combinación lineal de variables normales independientes se distribuye también de forma normal), salvo que como rasgo identificativo tiene sus datos más distribuidos a lo largo de todo el gráfico; otro dato importante a notar es que la media, la desviación tipica y el rango son la suma de sus contrapartes de las distribuciones N1 y N2, y el coeficiente de variación es la mitad de la suma del coeficiente de variacion de N1 y N2.

Final de la primera entrega