Practica 1

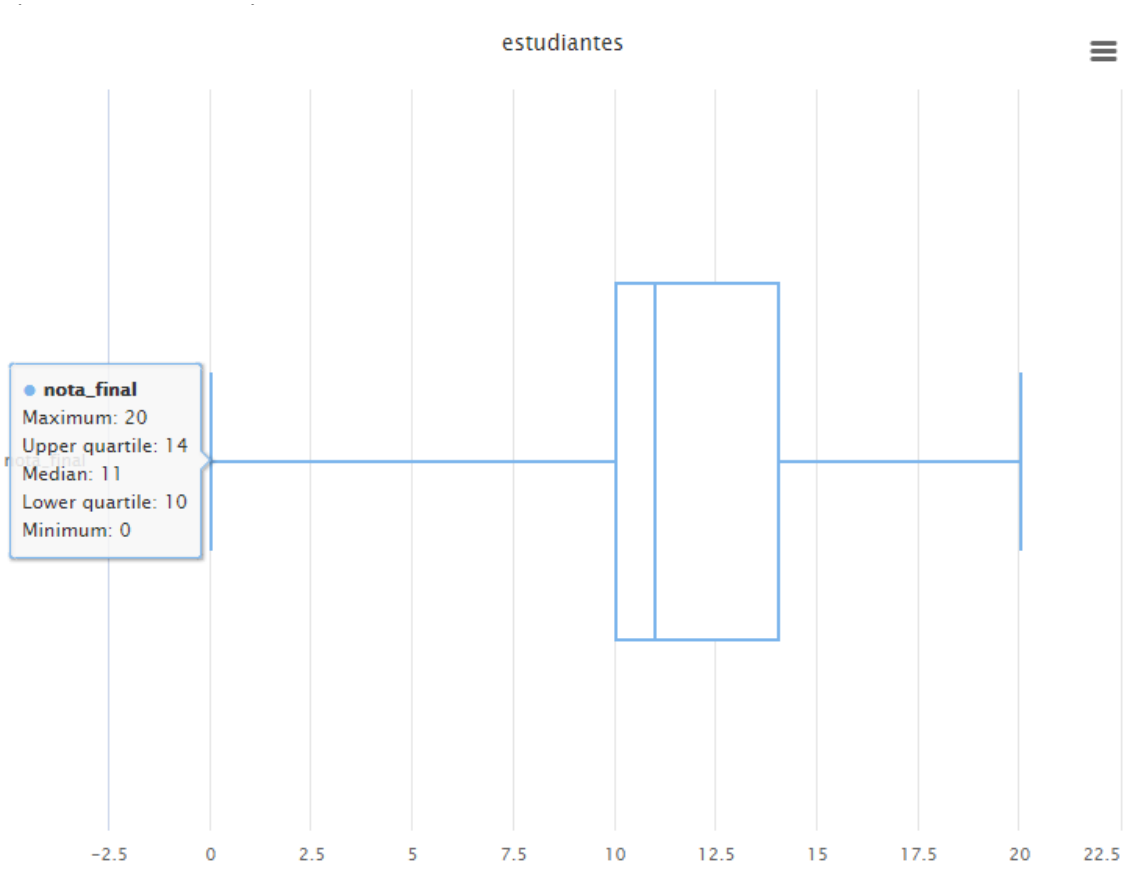
jueves, 11 de marzo de 2021 11:36

1. Indique qué tipo de información brindan las siguientes representaciones gráficas:
   1. Diagrama de dispersión (scatter plot)
   2. Diagrama de caja (box plot)
   3. Histograma
   4. Diagrama de Barras

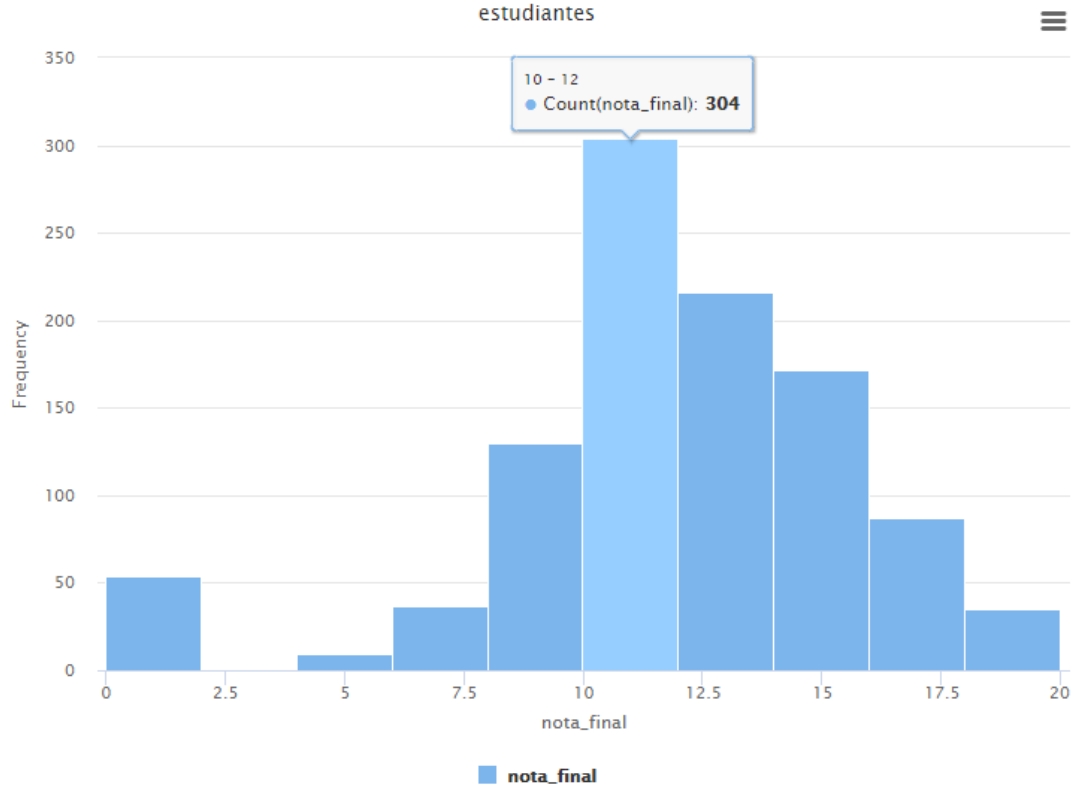
   Realice al menos una de cada una de las representaciones anteriores utilizando la información del archivo ***estudiantes.csv*** y explique cómo interpretarlas.

1) Consiste en dibujar pares de valores (xi , yj) medidos de la v.a. (X,Y) en un sistema de coordenadas. Ayuda a ver el tipo de relacion entre dos variables numéricos.

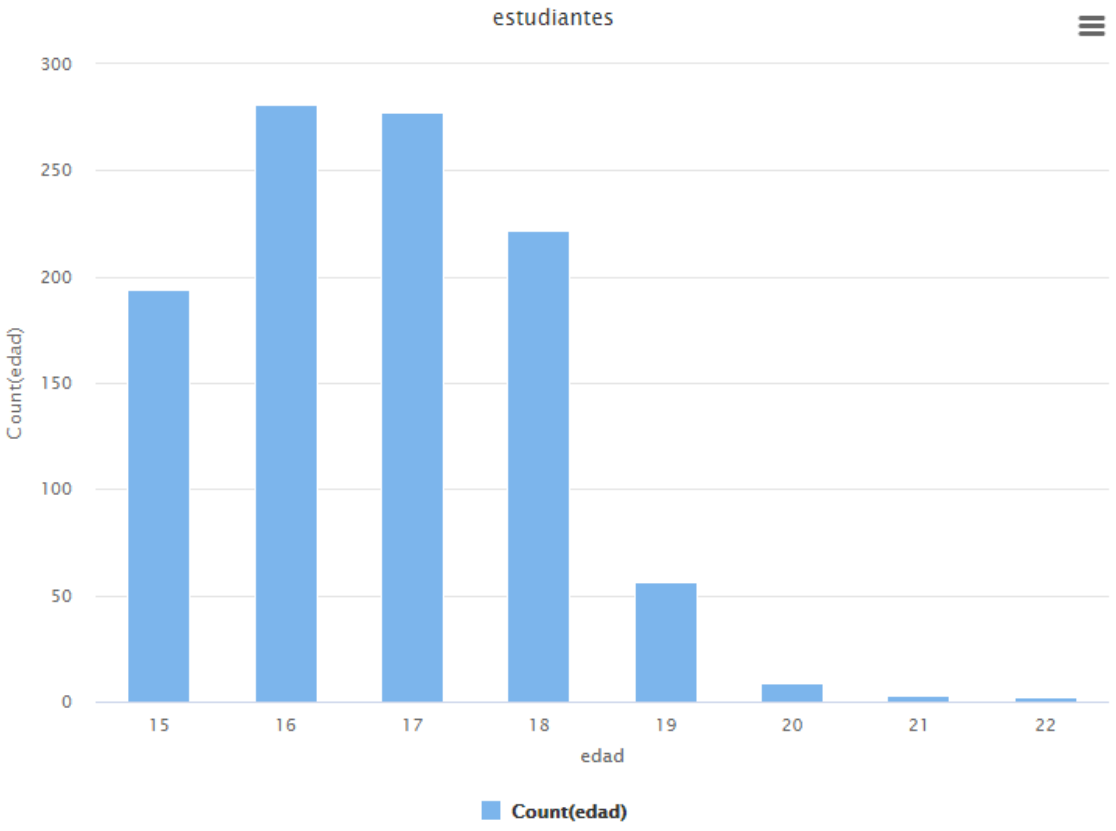
2) El diagrama de caja muestra el tamaño de los cuartiles. El de Tukey muestra los valores atípicos y los bigotes, que definen el límite en donde los valores empiezan a ser atípicos (1.5 x RIC). El de caja simple el valor mínimo y máximo.



3) El histograma es un grafico de barras que divide el rango de valores en sub-rangos, y muestra la frecuencia de los elementos en cada sub-rango.



4) El diagrama de barras muestra la frecuencia de cada elemento, con una barra por cada elemento.



2. En la figura más abajo pueden verse los Diagramas de Caja de Tukey correspondientes al atributo **Asistencias del archivo estudiantes.csv** separando los ejemplos por el atributo *sexo*. Calcule la mediana, los cuartiles Q1 y Q3, el rango intercuartil, y los intervalos en donde hay valores atípicos leves y extremos. Marcar estos valores en el gráfico.
   Luego indique el valor de verdad de las siguientes afirmaciones. En caso de no poder hacerlo, justifique.
   a) Es atípico que un estudiante tenga más de 20 ausencias.
   b) Los cuartiles del atributo *AUSENCIAS* son los mismos para ambos sexos por lo que puede afirmarse que la cantidad de mujeres y varones con más de 6 ausencias coinciden.
   c) Al menos el 25% de las mujeres tiene asistencia perfecta.
   d) Es atípico encontrar un varón que no haya faltado nunca.
   e) La cantidad de mujeres con valores atípicos leves en el atributo AUSENCIAS es mayor que la de varones.

- Los valores de la muestra que pertenezcan a alguno de estos intervalos

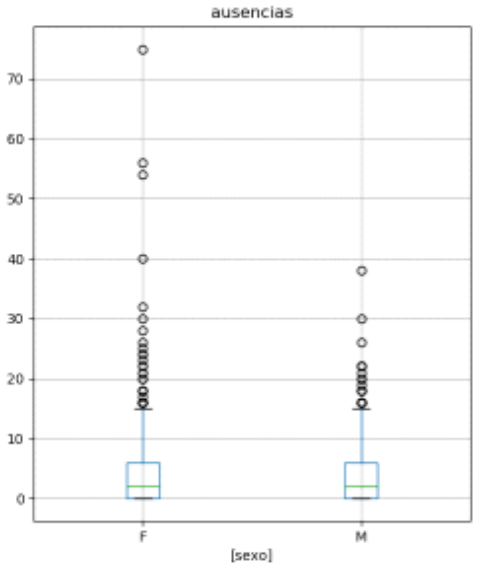
  [Q1– 3*RIC ; Q1 – 1.5*RIC) o (Q3 + 1.5*RIC ; Q3 + 3*RIC]

  serán considerados **valores fuera de rango leves.**

- Los valores de la muestra inferiores a **Q1 – 3*RIC** o superiores a **Q3 + 3*RIC** serán considerados **valores fuera de rango extremos.**

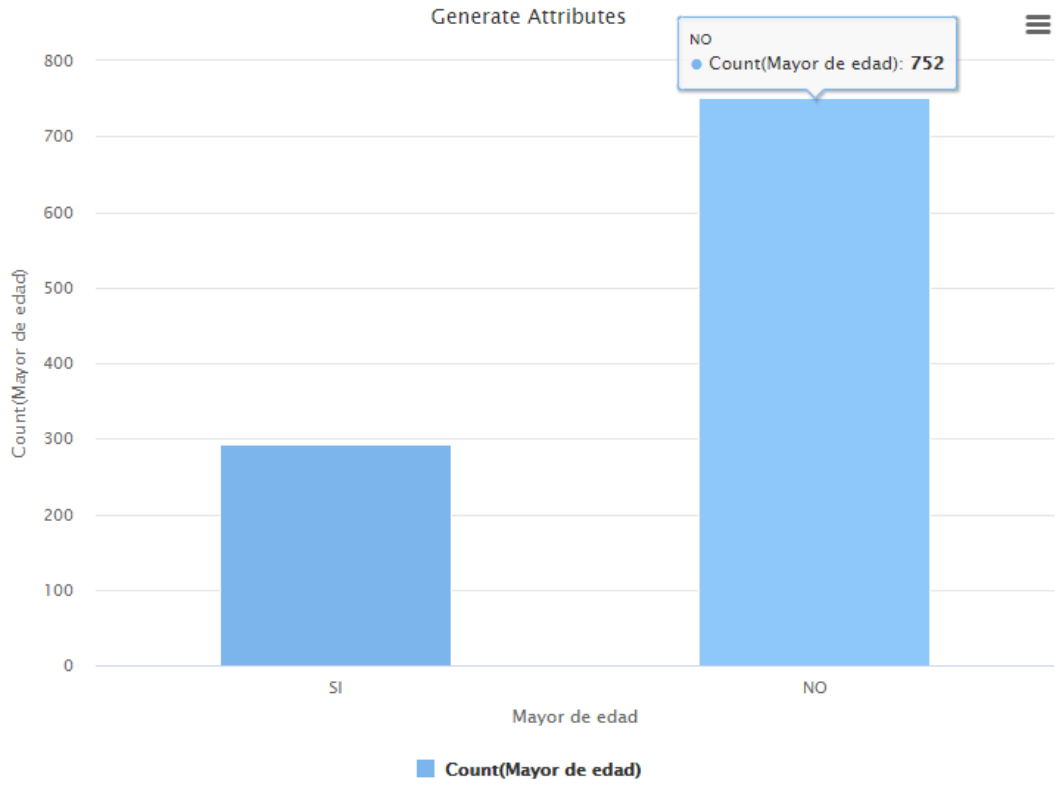
	F	M
Máximo	75	38
3er.cuartil	6	6
2do. cuartil	2	2
1er. cuartil	0	0
Mínimo	0	0
Rango Intercuartil	6	6
Atípicos leves		
Atípicos extremos		

Atípicos leves: [-18,-9) U (15,24]
Atípicos extremos: (-inf, -18) U (24, +inf)

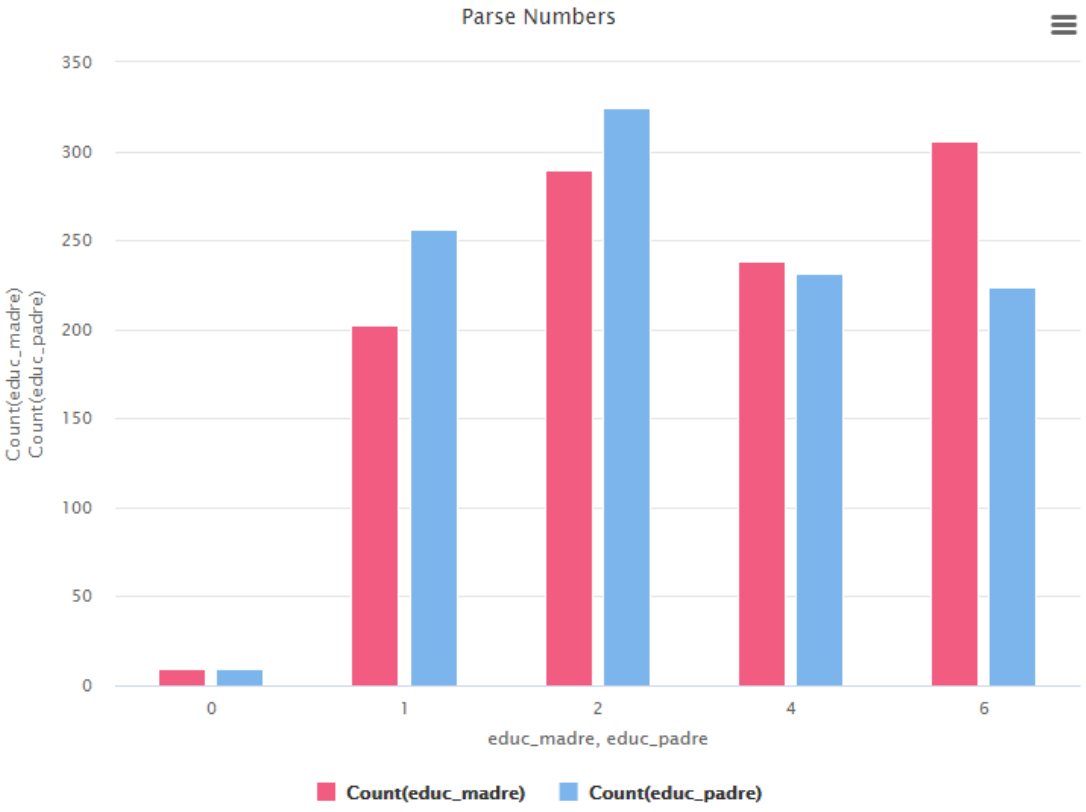


a) Verdadero, el rango de los valores atípicos comienza en las 15 ausencias.
b) No se puede determinar, tienen la misma distribución, pero la cantidad de mujeres y varones pueden ser diferentes.
c) Verdadero.
d) Falso.
e) No se puede determinar, no se lleva una cuenta de los valores atípicos.

3. Se considera que un estudiante es mayor de edad si tiene al menos 18 años. Genere un nuevo atributo que tome el valor "SI" cuando la edad del estudiante sea mayor o igual a 18 años y "NO" en caso contrario. Luego grafique el resultado obtenido mediante un diagrama de barras.



4. Los atributos *EDUC_MADRE* y *EDUC_PADRE* indican la educación recibida por el padre y por la madre de cada estudiante, respectivamente. Los valores posibles son "ninguna", "primaria (hasta 4to)", "primaria (hasta 9no)", "secundaria" y "universitaria". Utilice el operador ***Map*** para mapear estos valores de la siguiente forma: "ninguna" = 0, "primaria (hasta 4to)" = 1, "primaria (hasta 9no)" = 2, "secundaria" = 4 y "universitaria" = 6. Recuerde aplicar el operador ***Parse Number*** para convertir los valores de estos atributos en numéricos. Luego de la numerización, compare el nivel educativo de las madres con el de los padres.

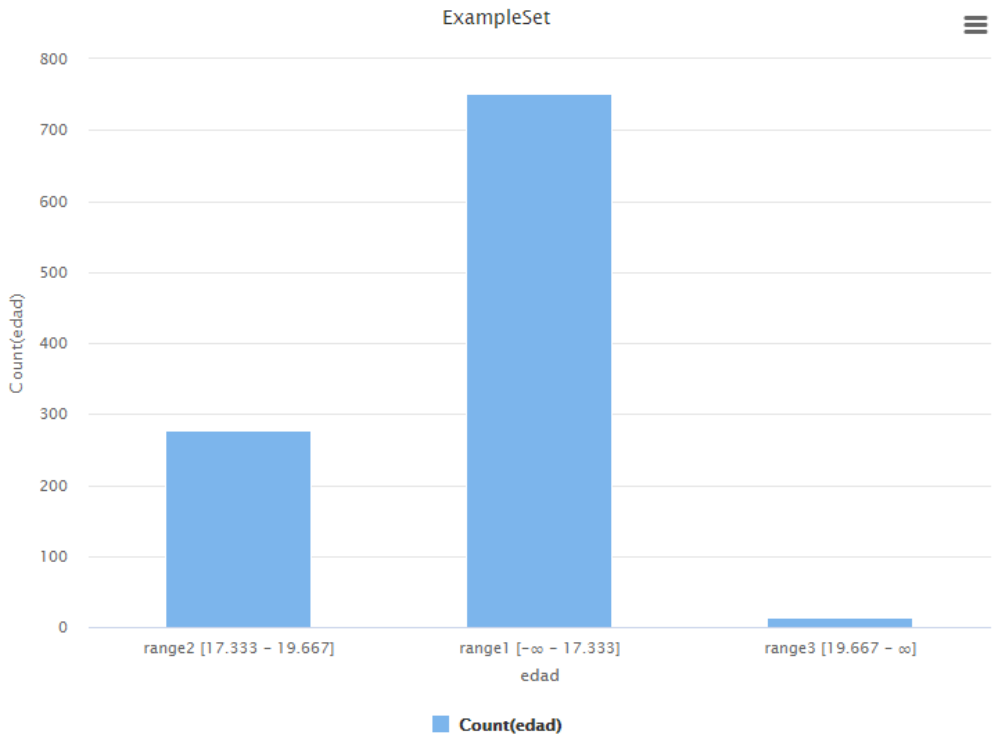


En general, la educación de las madres es más elevada.

5. Discretice el atributo EDAD en tres intervalos de dos formas distintas:
   a. Utilizando una discretización por rango (operador ***DiscretizeByBinning***)
   b. Utilizando discretización por frecuencia (operador ***DiscretizeByFrecuency***)

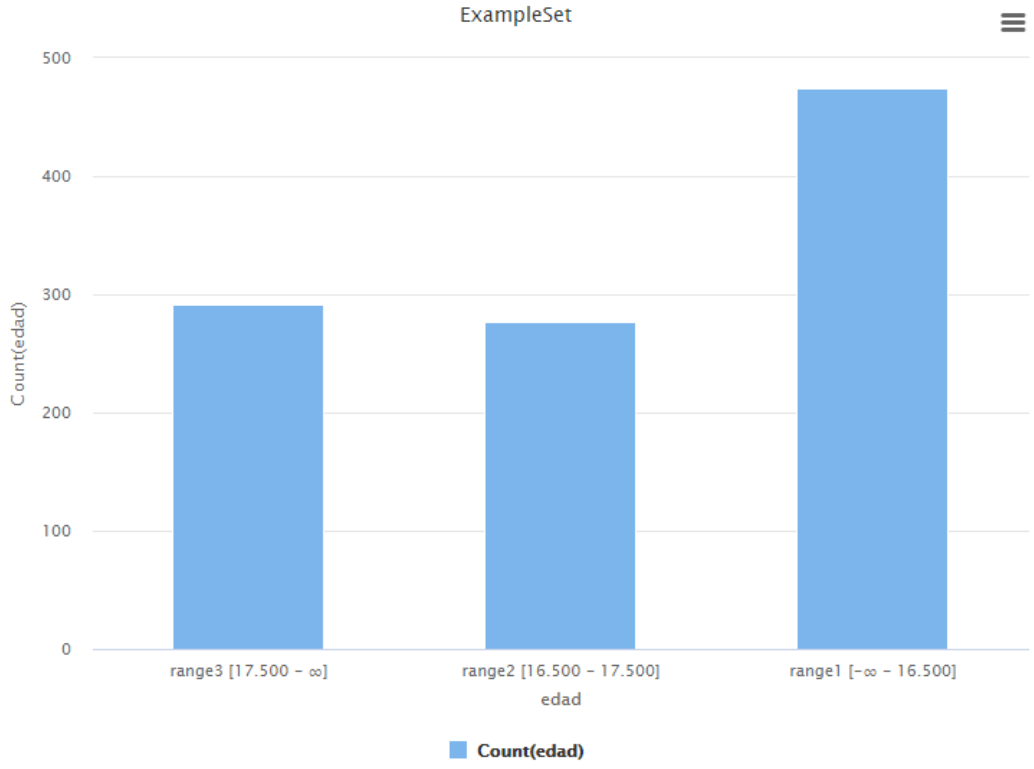
   Analice y compare los resultados obtenidos. Explique cómo se determina en cada caso los intervalos a utilizar.

a)



El objetivo es dividir el rango del atributo (intervalo entre el máximo y el mínimo) en una cierta cantidad k de partes iguales

b)



El objetivo es dividir los valores del atributo numérico en k partes con la misma cantidad de valores en cada una de ellas.

6. Analice la información del archivo **automobile.csv** cuyo contenido se encuentra descripto en "automobile_names.txt"
   a. Indique qué tipo de gráfica puede construir con los atributos. Ejemplifique cada caso.
   b. Utilizando distintas representaciones gráficas, describa la distribución de los atributos, e indique si observa relaciones entre los mismos.
   c. La Minería de Datos permite extraer dos tipos de conocimiento: descriptivo y predictivo. Ejemplifíquelos para el caso de los autos.
   d. Calcule el coeficiente de correlación lineal entre los atributos numéricos. Relacione los valores obtenidos con los diagramas de dispersión de cada par de atributos.

a) Hay muchos gráficos para hacer.
b) Hay relaciones entre los atributos, por ejemplo, a más caro el auto, más ancho y largo el auto, ratio de eficiencia de consumo menor, y tienen más caballo de fuerza. Para ver bien los atributos relacionados se puede usar el operador matriz de correlación (correlation matrix).
c) Descriptivo: el conocimiento descriptivo muestran nuevas relaciones entre las variables, en base a analizar los datos que tenemos, como las relaciones en la b).
Predictivo: el conocimiento predictivo plantea que, en base al modelo construido, es posible predecir hechos futuros. Por ejemplo, si ingreso un auto con nuevos valores pero al que le falta el precio, en base a los demás atributos se puede estimar el valor que le corresponde.
d) Para calcular el coeficiente de correlación lineal utilizamos la correlation matrix. A más cercano a 1 sea el coeficiente, más lineal es la relación entre los atributos, esto significa que tienen una relación cercana, como puede ser el precio de un auto con sus caballos de fuerza, aumentan a la par. Si el valor se acerca a 0, la relación es más distante, y el gráfico tiene los puntos más distribuídos, como puede ser la relación entre precio y altura de un auto, que no tienen una relación directa.

7. En el siguiente link encontrará información referida al uso de bicicletas que el Gobierno de la Ciudad de Buenos Aires pone a disposición de la población en forma gratuita como medio de transporte:

   https://recursos-data.buenosaires.gob.ar/ckan2/bicicletas-publicas/recorrido-bicis-2016.csv

   Estas bicicletas están ubicadas en distintos puntos de la ciudad y se encuentran disponibles las 24 horas del día durante todo el año. En el archivo encontrará información referida a las estaciones de origen y destino, la hora de partida y la duración de los viajes realizados por las bicicletas durante el año 2016.

a. A partir del atributo FECHA_HORA_RETIRO genere un atributo nuevo que contenga únicamente el horario en el cual la bicicleta fue retirada. Para ello puede obtener el substring que contiene la hora con las funciones de texto de **GenerateAttributes**, y luego convertir ese substring a un número entero con el mismo operador. También puede utilizar la función *date_parse* de **GenerateAttributes** para convertir el string a un tipo date, y luego utilizar *date_get* para obtener la hora.
b. Utilizando un histograma donde cada hora represente una barra, informe si hay horarios inusuales de retiro de bicicletas. Justifique su respuesta utilizando la frecuencia relativa de cada hora para decidir qué es un horario inusual (datos) y en qué horas tradicionalmente circula la gente por una ciudad (su conocimiento sobre el dominio)

Los horarios inusuales son desde la 7 a las 23.

c. Indique el valor de verdad de la siguiente proposición: "Se obtendrán los mismos resultados si se discretiza por rango el atributo generado en a) utilizando 4 intervalos que si se lo discretiza por frecuencia utilizando 4 intervalos". Justifique su respuesta.

No, porque al discretizar por rango parte el horario en 4, mientras que para discretizar por frecuencia toma en cuenta la cantidad de ocurrencias de los elementos.

d. A partir del atributo FECHA_HORA_RETIRO genere un segundo atributo con el número de mes en el cual la bicicleta fue retirada. Grafique manualmente el histograma correspondiente a este atributo utilizando 3 intervalos. ¿Hay un componente estacional en el uso de bicicletas?

Si, después de graficarlo notamos que hay un componente estacional, el uso crece a medida que pasa el año.

8. El archivo estrellas.xslx contiene información sobre estrellas de una zona del espacio previamente inexplorada. Utilizando ese archivo:
   a. Discretice por frecuencia el atributo **Temperatura** en dos intervalos llamados **Baja** y **Alta**. Indique los rangos de los dos intervalos resultantes, así como la cantidad de ejemplos que hay en cada intervalo.

	Baja	Alta
Intervalos	[-infinito ; 3225]	[3225 ; infinito]
Valores	5	5

   b. Discretice por rango el atributo **Temperatura** en dos intervalos llamados **Baja** y **Alta**. Indique los rangos de los dos intervalos resultantes, así como la cantidad de ejemplos que hay en cada intervalo.

	Baja	Alta
Intervalos	[-infinito ; 4450]	[4450 ; infinito]
Valores	6	4

   c. Calcule la correlación lineal entre los atributos **Temperatura** y **Luminosidad**. Indique la intensidad de la correlación (no hay correlación/débil/fuerte) y el tipo (positiva/negativa)

Valor	0.555
Intensidad	Débil
Tipo	Positiva

   d. Dibuje un Diagrama de Caja de Tukey de la variable **Luminosidad** e inclúyalo en la respuesta. Indique también los valores del cuadro:

Mediana	4
Q1	0,75
Q3	8,5
RI	7,75
Bigote superior:	16
Bigote inferior:	0,2
Intervalos de valores atípicos leves	[-22,5 ; -10.875) U (20,125 ; 31,75]
Valores atípicos leves	22
Intervalos de valores atípicos extremos	(-infinito ; -22,5) U (31,75 ; infinito)
Valores atípicos extremos	Ninguno