

Análisis exploratorio de datos

Harlin Acero Acero

Jeyson Acero Acero

Minería de Datos - Laboratorio 1

```
In [3]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as sp
```

Importar un archivo de datos

```
In [56]: df = pd.read_csv("./Clasificacion.txt", sep="\t")
df.head()
```

Out[56]:

	ID Caso	Clase Vino	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides	Fenole Flavanc
0	1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	
1	2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	
2	3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	
3	4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	
4	5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	

Calculo de estadísticas descriptivas

Previo a conocer el contexto de los datos, de que variables se componen la data y que representan, el siguiente paso sugerido es calcular las estadísticas descriptivas del conjunto de las variables que componen los datos para revisar su frecuencia, distribución completitud y hacerse una idea de ellos.

```
In [57]: # Cálculo de estadísticas descriptivas
print ('El máximo es:' + str(df.Cenizas.max()))
print ('El mínimo es:' + str(df.Cenizas.min()))
print ('La media es:' + str(round(df.Cenizas.mean(),2)))
print ('La desviación estandar es:' + str(round(df.Cenizas.std(),2)))
print ('La varianza es:' + str(round(df.Cenizas.var(),2)))
```

```
El máximo es:3.23
El mínimo es:1.36
La media es:2.37
La desviación estandar es:0.27
La varianza es:0.08
```

```
In [58]: #Obtenemos el promedio de cada columna del dataset
df.mean()
```

```
Out[58]: ID Caso                90.000000
Clase Vino                1.944134
Alcohol                13.006927
Acido Malico                2.346201
Cenizas                2.368603
Alcalinidad Cenizas        19.522905
Magnesio                99.720670
Fenoles Totales            2.293743
Flavanoides              2.022179
Fenoles No Flavanoides      0.362961
Protoantocianinas          1.589553
Intensidad Color            5.081229
Matiz                    0.955508
OD280_OD315 de lso vinos diluidos  2.606034
Prolina                 745.849162
dtype: float64
```

```
In [59]: # Obtiene los estadísticos de todas las columnas
df.describe()
```

Out[59]:

	ID Caso	Clase Vino	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	
count	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179
mean	90.000000	1.944134	13.006927	2.346201	2.368603	19.522905	99.720670	2
std	51.816986	0.776919	0.813932	1.121776	0.274993	3.351116	14.245053	0
min	1.000000	1.000000	11.030000	0.740000	1.360000	10.600000	70.000000	0
25%	45.500000	1.000000	12.365000	1.605000	2.210000	17.200000	88.000000	1
50%	90.000000	2.000000	13.050000	1.870000	2.360000	19.500000	98.000000	2
75%	134.500000	3.000000	13.685000	3.110000	2.560000	21.500000	107.000000	2
max	179.000000	3.000000	14.830000	5.800000	3.230000	30.000000	162.000000	3

```
In [10]: # Obtenemos el estadístico de la columna seleccionada
df["Clase Vino"].describe()
```

```
Out[10]: count    179.000000
mean         1.944134
std          0.776919
min          1.000000
25%          1.000000
50%          2.000000
75%          3.000000
max          3.000000
Name: Clase Vino, dtype: float64
```

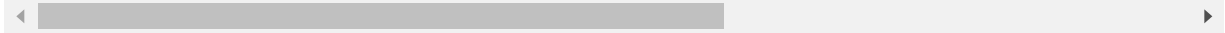
```
In [11]: #Obtiene el tipo de dato, y la cantidad de registros de cada columna
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179 entries, 0 to 178
Data columns (total 15 columns):
ID Caso                179 non-null int64
Clase Vino             179 non-null int64
Alcohol               179 non-null float64
Acido Malico          179 non-null float64
Cenizas               179 non-null float64
Alcalinidad Cenizas   179 non-null float64
Magnesio              179 non-null int64
Fenoles Totales       179 non-null float64
Flavanoides           179 non-null float64
Fenoles No Flavanoides 179 non-null float64
Protoantocianinas     179 non-null float64
Intensidad Color      179 non-null float64
Matiz                 179 non-null float64
OD280_OD315 de lso vinos diluidos 179 non-null float64
Prolina               179 non-null int64
dtypes: float64(11), int64(4)
memory usage: 21.1 KB
```

```
In [8]: # Obtemos los dos primeros registros del dataset
df.head(2)
```

```
Out[8]:
```

	ID Caso	Clase Vino	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides	Fenole Flavanc
0	1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	
1	2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	



```
In [60]: # Obtenemos el número total de registros en el dataset
len(df)
```

```
Out[60]: 179
```

```
In [12]: #Obtenemos los últimos 5 registros del dataset
df.tail(5)
```

Out[12]:

	ID Caso	Clase Vino	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides	Fenoles No Flavanoides
174	175	3	13.40	3.91	2.48	23.0	102	1.80	0.75	
175	176	3	13.27	4.28	2.26	20.0	120	1.59	0.69	
176	177	3	13.17	2.59	2.37	20.0	120	1.65	0.68	
177	178	3	14.13	4.10	2.74	24.5	96	2.05	0.76	
178	179	3	14.13	4.10	2.74	24.5	96	2.05	0.76	

```
In [12]: # Obtiene el Nombre de las columnas
df.columns
```

Out[12]: Index(['ID Caso', 'Clase Vino', 'Alcohol', 'Acido Malico', 'Cenizas',
 'Alcalinidad Cenizas', 'Magnesio', 'Fenoles Totales', 'Flavanoides',
 'Fenoles No Flavanoides', 'Protoantocianinas', 'Intensidad Color',
 'Matiz', 'OD280_OD315 de los vinos diluidos', 'Prolina'],
 dtype='object')

```
In [13]: # Obtiene los datos de la columna seleccionada  
df["Alcohol"]
```

```
Out[13]: 0      14.23
          1      13.20
          2      13.16
          3      14.37
          4      13.24
          5      14.20
          6      14.39
          7      14.06
          8      14.83
          9      13.86
         10      14.10
         11      14.12
         12      13.75
         13      14.75
         14      14.38
         15      13.63
         16      14.30
         17      13.83
         18      14.19
         19      13.64
         20      14.06
         21      12.93
         22      13.71
         23      12.85
         24      13.50
         25      13.05
         26      13.39
         27      13.30
         28      13.87
         29      14.02
          ...
        149      13.08
        150      13.50
        151      12.79
        152      13.11
        153      13.23
        154      12.58
        155      13.17
        156      13.84
        157      12.45
        158      14.34
        159      13.48
        160      12.36
        161      13.69
        162      12.85
        163      12.96
        164      13.78
        165      13.73
        166      13.45
        167      12.82
        168      13.58
        169      13.40
        170      12.20
        171      12.77
        172      14.16
        173      13.71
        174      13.40
```

```
175    13.27
176    13.17
177    14.13
178    14.13
Name: Alcohol, Length: 179, dtype: float64
```

```
In [14]: #Obtiene las 5 primeras filas de la columna seleccionada
df["Alcohol"][:5]
```

```
Out[14]: 0    14.23
         1    13.20
         2    13.16
         3    14.37
         4    13.24
Name: Alcohol, dtype: float64
```

```
In [16]: #Obtiene el dato de la columna de la fila # 5
df["Alcohol"][5]
```

```
Out[16]: 14.2
```

```
In [17]: #Obtiene los 5 primeras filas de las columnas indicadas
df[["Prolina", "Alcohol"][:5]]
```

```
Out[17]:
```

	Prolina	Alcohol
0	1065	14.23
1	1050	13.20
2	1185	13.16
3	1480	14.37
4	735	13.24

```
In [19]: #Cuenta el número de registros agrupados por la columna seleccionada
df["Clase Vino"].value_counts()
```

```
Out[19]: 2    71
         1    59
         3    49
Name: Clase Vino, dtype: int64
```

```
In [20]: #Para la matriz de correlaciones, o el porcentaje de relación entre las columna
s seleccionadas con el algoritmo de pearson:
df.corr(method="pearson")["Magnesio"]["Intensidad Color"]
```

```
Out[20]: 0.1955369254608783
```

```
In [21]: # Obtiene las relaciones de cada variable del dataset con las demás, los porcentajes superiores a 0.7 son los más relacionados
df.loc[:, "Alcohol": "Prolina"].corr()
```

Out[21]:

	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides
Alcohol	1.000000	0.105425	0.219845	-0.295056	0.267250	0.284375	0.224643
Acido Malico	0.105425	1.000000	0.173996	0.297829	-0.056494	-0.336150	-0.417462
Cenizas	0.219845	0.173996	1.000000	0.449652	0.283058	0.125279	0.104342
Alcalinidad Cenizas	-0.295056	0.297829	0.449652	1.000000	-0.084988	-0.322244	-0.358185
Magnesio	0.267250	-0.056494	0.283058	-0.084988	1.000000	0.214844	0.196726
Fenoles Totales	0.284375	-0.336150	0.125279	-0.322244	0.214844	1.000000	0.863080
Flavanoides	0.224643	-0.417462	0.104342	-0.358185	0.196726	0.863080	1.000000
Fenoles No Flavanoides	-0.141707	0.302822	0.195992	0.370354	-0.256766	-0.450050	-0.542939
Protoantocianinas	0.132623	-0.222813	0.006396	-0.199516	0.236897	0.612770	0.652418
Intensidad Color	0.552375	0.260670	0.268745	0.033267	0.195537	-0.058520	-0.182670
Matiz	-0.082642	-0.567126	-0.085295	-0.283130	0.057254	0.434030	0.548286
OD280_OD315 de Iso vinos diluidos	0.060532	-0.376565	-0.006912	-0.285339	0.067703	0.698808	0.789285
Prolina	0.635007	-0.195714	0.217740	-0.442371	0.393759	0.498712	0.495694

Medidas agrupadas

```
In [20]: #Obtiene los estadísticos indicados en la columna seleccionada
#a partir de la agrupación indicada
df.groupby(["Clase Vino"])[ "Alcalinidad Cenizas"].agg(["count", "min", "mean", "max"])
```

Out[20]:

	count	min	mean	max
Clase Vino				
1	59	11.2	17.037288	25.0
2	71	10.6	20.238028	30.0
3	49	17.5	21.479592	27.0


```
In [25]: #Obtenemos el promedio desde y hasta la columna seleccionadas agrupado por la
          columna seleccionada
          df.loc[:, "Clase Vino": "Prolina"].groupby("Clase Vino").mean()
```

Out[25]:

	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides	Fenoles N Flavanoide
Clase Vino								
1	13.744746	2.010678	2.455593	17.037288	106.338983	2.840169	2.982373	0.290000
2	12.278732	1.932676	2.244789	20.238028	94.549296	2.258873	2.080845	0.363660
3	13.173673	3.349388	2.443265	21.479592	99.244898	1.686327	0.781020	0.449790

```
In [27]: #matriz de correlaciones desde y hasta las columnas seleccionadas agrupadas por
          r tipo de vino
          df.loc[:, "Clase Vino": "Cenizas"].groupby("Clase Vino").corr()
```

Out[27]:

		Acido Malico	Alcohol	Cenizas
Clase Vino				
1	Acido Malico	1.000000	-0.040513	0.026221
	Alcohol	-0.040513	1.000000	-0.148595
	Cenizas	0.026221	-0.148595	1.000000
2	Acido Malico	1.000000	-0.021362	0.148708
	Alcohol	-0.021362	1.000000	-0.214851
	Cenizas	0.148708	-0.214851	1.000000
3	Acido Malico	1.000000	0.132197	0.040791
	Alcohol	0.132197	1.000000	0.289670
	Cenizas	0.040791	0.289670	1.000000

```
In [23]: # La matriz de relación muestra el nivel de interacción entre las diferentes variables.
```

Gráficos

Histograma

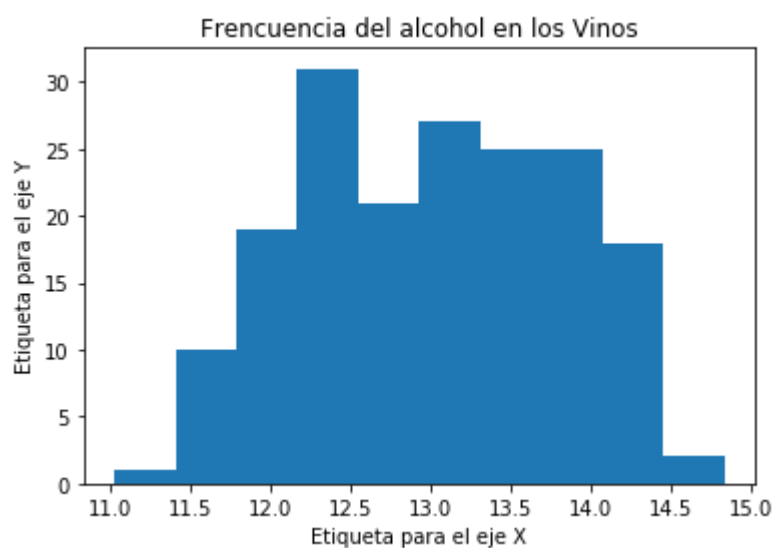
```
In [24]: #muestra la acumulación ó tendencia,
          #la variabilidad o dispersión y la forma de la distribución de una variable dada.
```

Frecuencia de porcentaje de alcohol

```
In [25]: df["Alcohol"].describe()
```

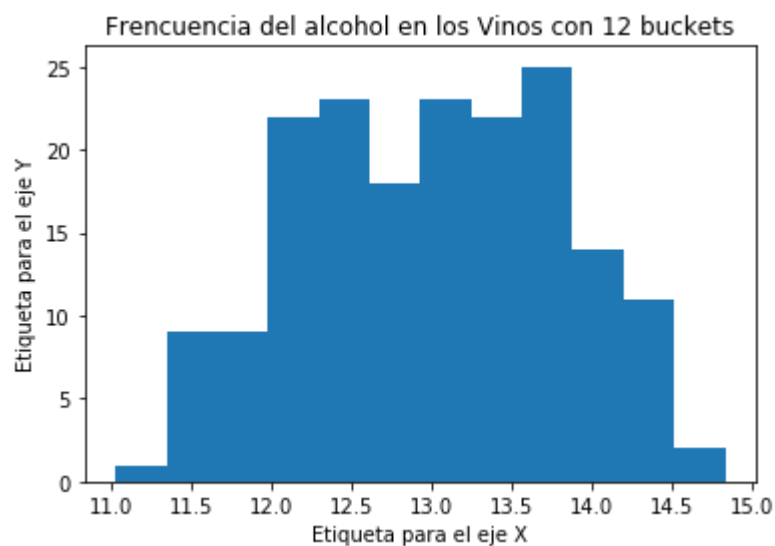
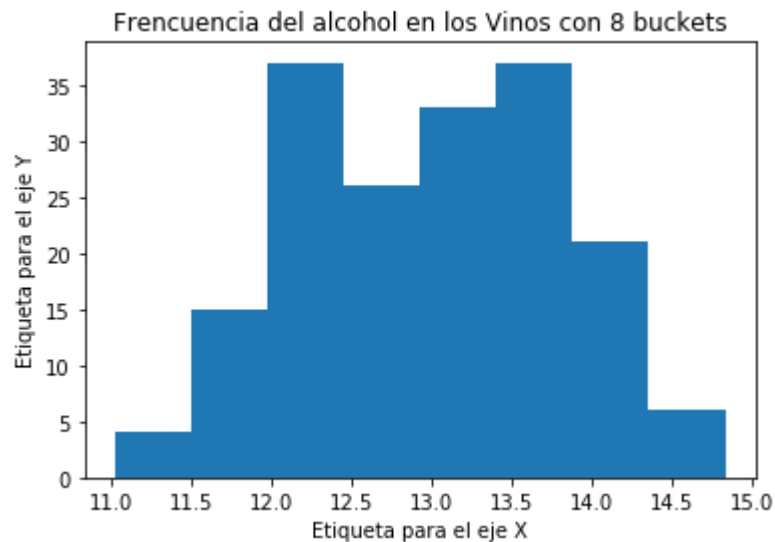
```
Out[25]: count    179.000000  
mean      13.006927  
std       0.813932  
min       11.030000  
25%      12.365000  
50%      13.050000  
75%      13.685000  
max       14.830000  
Name: Alcohol, dtype: float64
```

```
In [26]: plt.hist(df["Alcohol"])  
plt.title("Frecuencia del alcohol en los Vinos")  
plt.xlabel("Etiqueta para el eje X")  
plt.ylabel("Etiqueta para el eje Y")  
plt.show()
```



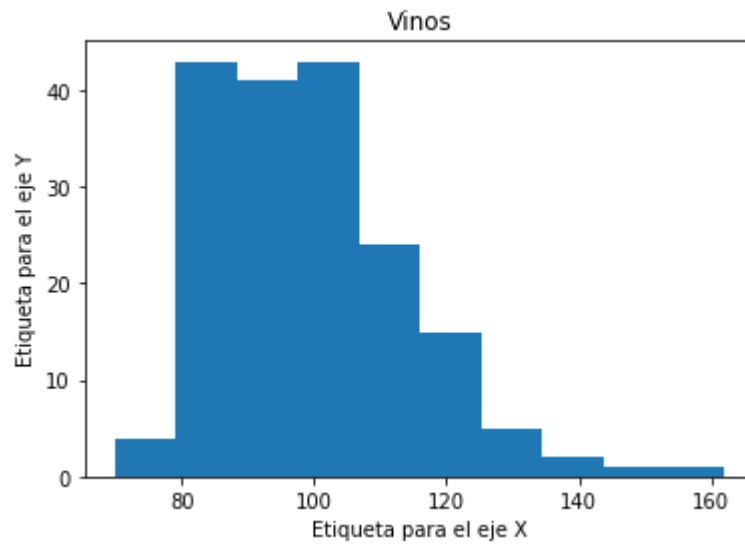
```
In [27]: # Cambiando el número de barras o contenedores
plt.hist(df["Alcohol"], bins=8)
plt.title("Frecuencia del alcohol en los Vinos con 8 buckets")
plt.xlabel("Etiqueta para el eje X")
plt.ylabel("Etiqueta para el eje Y")
plt.show()

plt.hist(df["Alcohol"], bins=12)
plt.title("Frecuencia del alcohol en los Vinos con 12 buckets")
plt.xlabel("Etiqueta para el eje X")
plt.ylabel("Etiqueta para el eje Y")
plt.show()
```

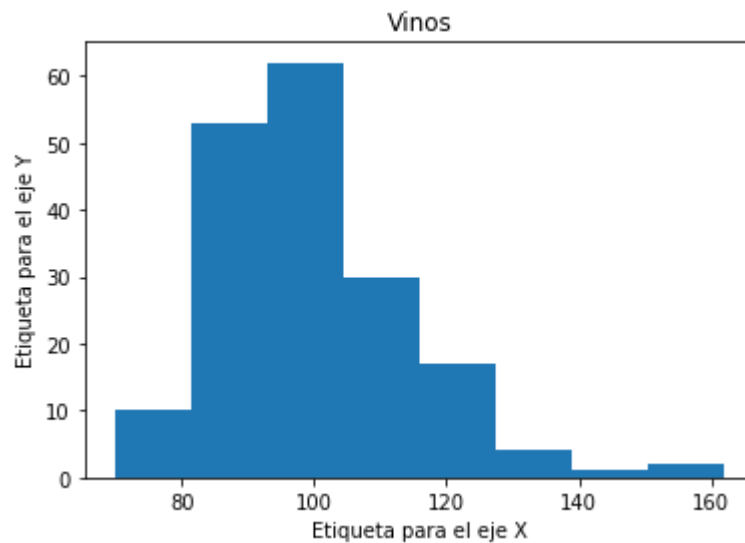


Frecuencia de porcentaje de Magnesio

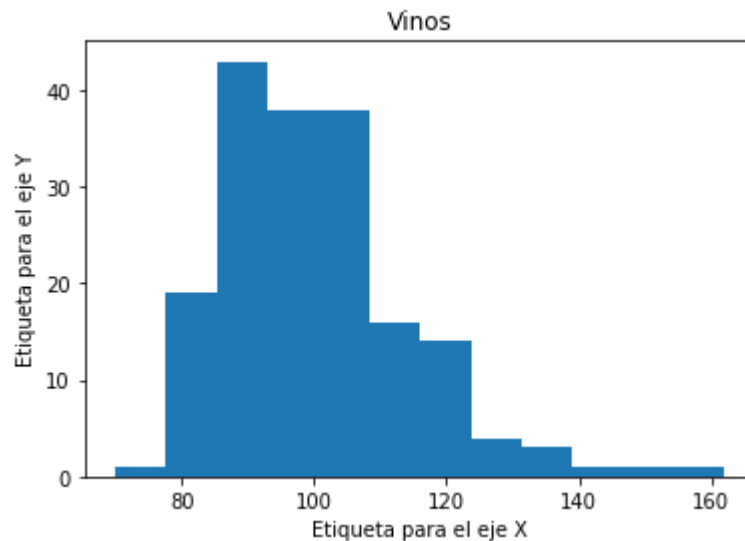
```
In [28]: plt.hist(df["Magnesio"])
plt.title("Vinos")
plt.xlabel("Etiqueta para el eje X")
plt.ylabel("Etiqueta para el eje Y")
plt.show()
```



```
In [29]: plt.hist(df["Magnesio"], bins = 8)
plt.title("Vinos")
plt.xlabel("Etiqueta para el eje X")
plt.ylabel("Etiqueta para el eje Y")
plt.show()
```



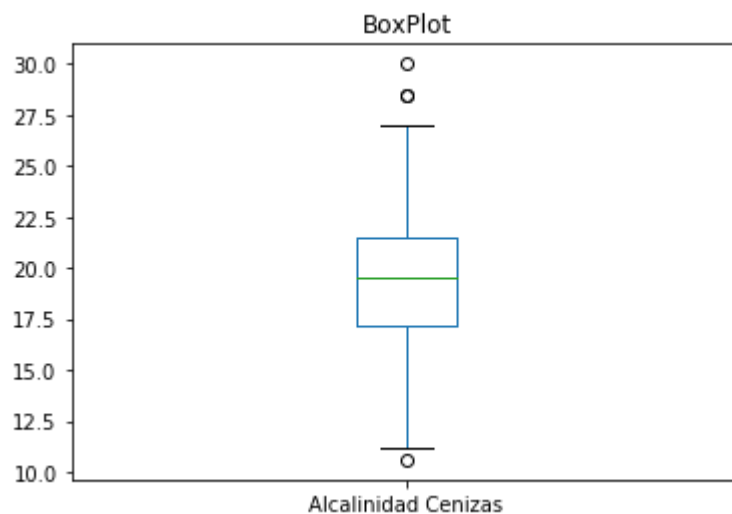
```
In [30]: plt.hist(df["Magnesio"], bins = 12)
plt.title("Vinos")
plt.xlabel("Etiqueta para el eje X")
plt.ylabel("Etiqueta para el eje Y")
plt.show()
```



Boxplot

```
In [31]: # Analizar medidas de tendencia y dispersión
#representar gráficamente una serie de datos numéricos a través de sus cuartiles
S.
#De esta manera, el diagrama de caja muestra a simple vista la mediana y los cuartiles de los datos
#pudiendo también representar los valores atípicos de estos.
df["Alcalinidad Cenizas"].plot(kind="box", title="BoxPlot")
```

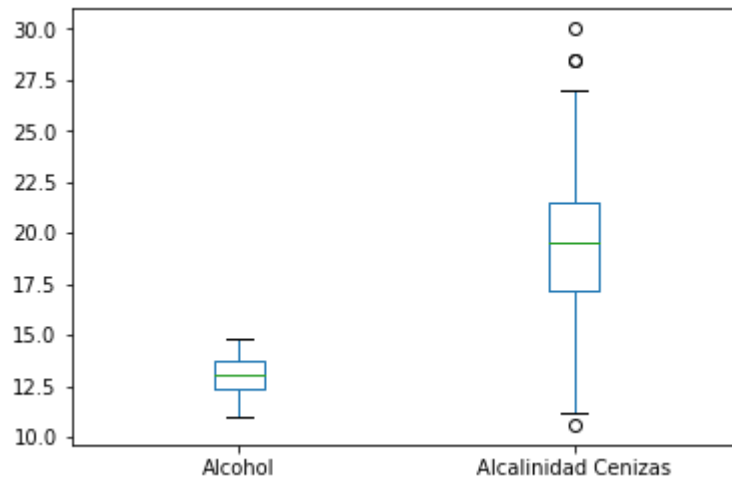
```
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35c9f8ac8>
```



```
In [32]: # hay tres valores atípicos, dos por encima del 4 cuartil y uno por debajo del 1 cuartil
```

```
In [33]: # Graficar para varias variables simultáneamente  
df.loc[:,["Alcohol","Alcalinidad Cenizas"]].plot.box()  
# El alcohol tiene una desviación mucho menor hacia el promedio y no tiene valores atípicos
```

```
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35ca840b8>
```



```
In [34]: # En efecto, las variables se encuentran más relacinadas entre 11 y 15
plt.scatter(df["Alcohol"],df["ID Caso"])
plt.title("Alcohol")
plt.show()
# Las variables se encuentran más dispersas, y hay dos valores atípicos
plt.scatter(df["Alcalinidad Cenizas"],df["ID Caso"])
plt.title("Alcalinidad Cenizas")
plt.show()
```

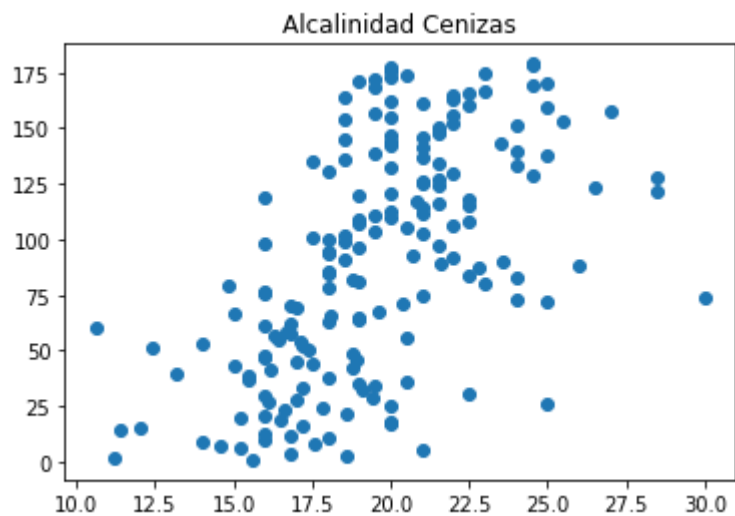
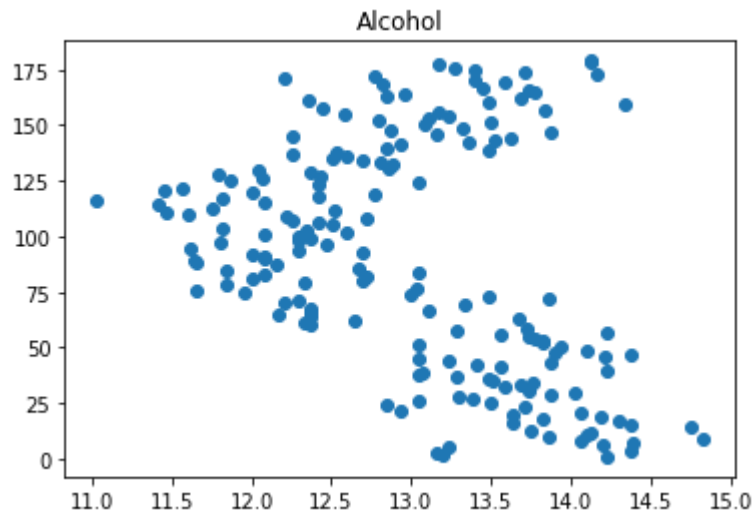


Diagrama de dispersión

```
In [35]: #Permite ver la relación simultánea de variables continuas
plt.scatter(df["Alcohol"],df["Matiz"])
plt.show()
```

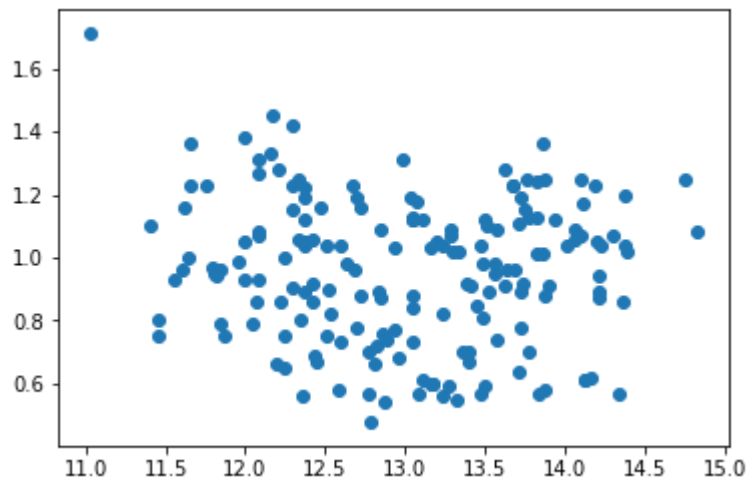
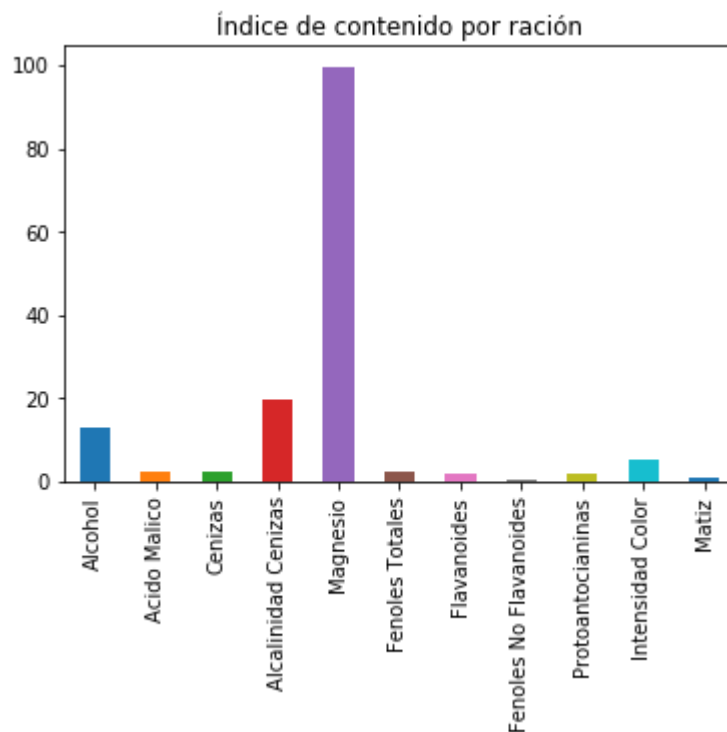


Gráfico de barras

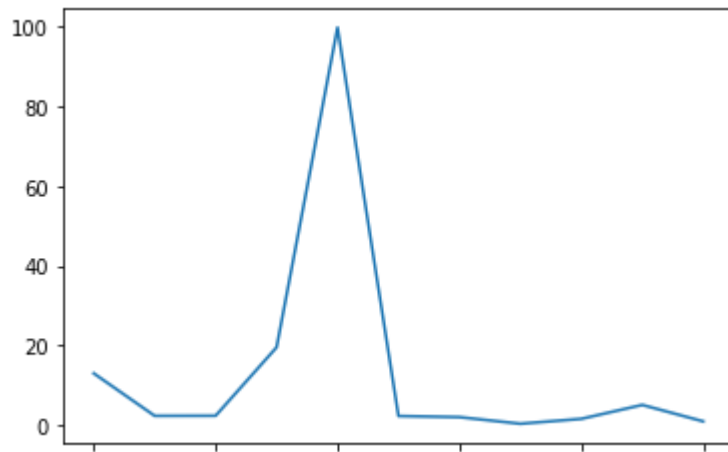
```
In [36]: # Permite comparar la relación de valores entre variables,
#en este caso, el promedio desde la columna "alcohol" hasta la columna "Matiz"
df.loc[:, "Alcohol": "Matiz"].mean().plot(kind="bar", title="Índice de contenido por ración")
```

Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35cc37eb8>



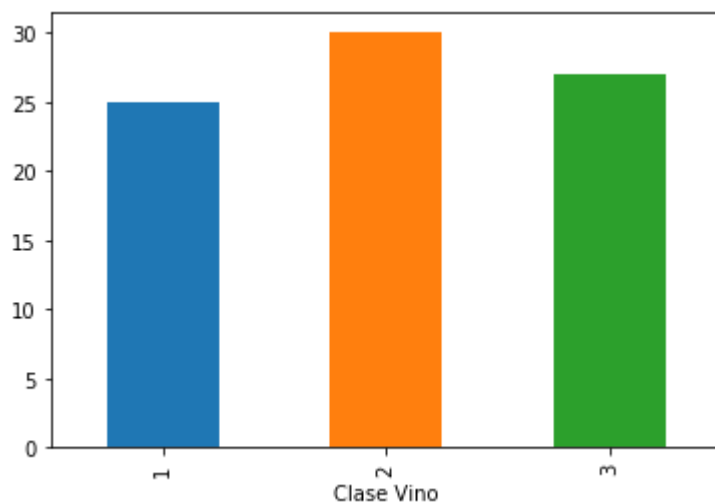

```
In [37]: # Con líneas
df.loc[:, "Alcohol": "Matiz"].mean().plot(kind="line")
```

Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35cca45c0>



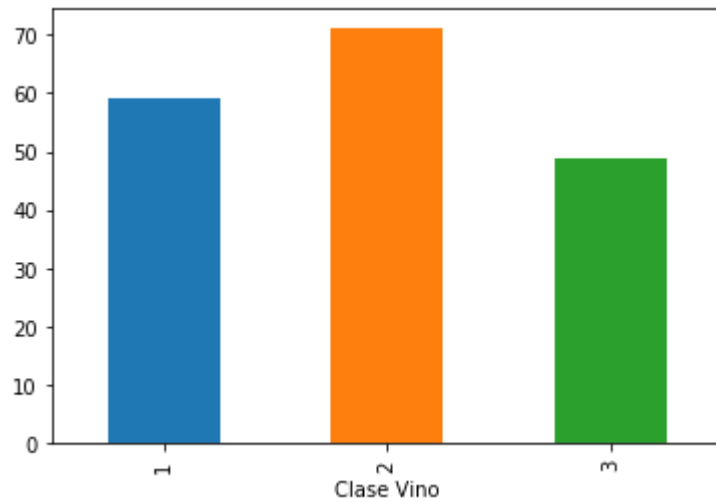
```
In [38]: #valor máximo para una variable entre las distintas clases de vino
df.groupby("Clase Vino")["Alcalinidad", "Cenizas"].max().plot(kind="bar")
```

Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35ca3fa90>



```
In [39]: #Distribución de la clase de vino en el dataset
df.groupby("Clase Vino")["Clase Vino"].count().plot(kind="bar")
```

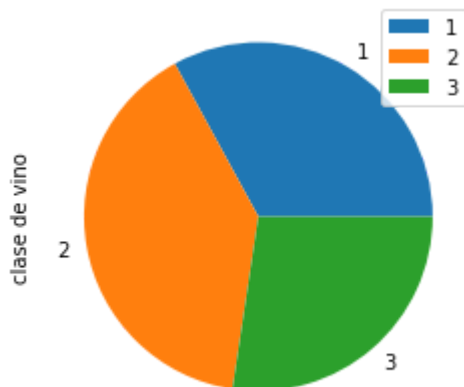
```
Out[39]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35c9e59e8>
```



Diagramas de pie

```
In [40]: #Visualizar los datos en forma circular
df.groupby("Clase Vino")["Clase Vino"].count().plot(kind="pie", label="clase d
e vino", legend="true")
```

```
Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35cb9c2e8>
```



```
In [41]: df.groupby("Clase Vino")["Clase Vino"].count()
```

```
Out[41]: Clase Vino
1      59
2      71
3      49
Name: Clase Vino, dtype: int64
```

Tablas de confusión

```
In [42]: #Resume el numero de ocurrencias de cada valor para una variable dada.  
pd.crosstab(index=df["Clase Vino"], columns="count")
```

```
Out[42]:
```

col_0	count
Clase Vino	
1	59
2	71
3	49

```
In [43]: for co in df.columns:  
         print(co)
```

```
ID Caso  
Clase Vino  
Alcohol  
Acido Malico  
Cenizas  
Alcalinidad Cenizas  
Magnesio  
Fenoles Totales  
Flavanoides  
Fenoles No Flavanoides  
Protoantocianinas  
Intensidad Color  
Matiz  
OD280_OD315 de lso vinos diluidos  
Prolina
```

```
In [30]: #Muestra el número de ocurrencias del dato en cada valor  
tcruzada = pd.crosstab(index=df["Clase Vino"], columns=df["Alcohol"])  
tcruzada.index= ["Vino 1","Vino 2","Vino 3"] #Cambia el nombre de los índices  
de agrupación  
tcruzada
```

```
Out[30]:
```

Alcohol	11.03	11.41	11.45	11.46	11.56	11.61	11.62	11.64	11.65	11.66	...	14.21	14.22	14
Vino 1	0	0	0	0	0	0	0	0	0	0	...	1	2	
Vino 2	1	1	1	1	1	1	1	1	1	1	...	0	0	
Vino 3	0	0	0	0	0	0	0	0	0	0	...	0	0	

3 rows × 126 columns



Análisis, Comprensión y preprocesamiento de los Datos

Parte 1

Seleccionamos 5 variables y de acuerdo a los histogramas y las medidas vistas en clase (media, varianza, skew, kurtosis) indique si se trata de frecuencias unimodales, bimodales simétricas o asimétricas(desequilibradas), y su apreciación de la distribución de los datos de esa variable en el contexto vinos.

```
In [31]: # Hacemos una separación de Los dataset por clase de vino
dfc1 = df[df["Clase Vino"] == 1]
dfc2 = df[df["Clase Vino"] == 2]
dfc3 = df[df["Clase Vino"] == 3]
```

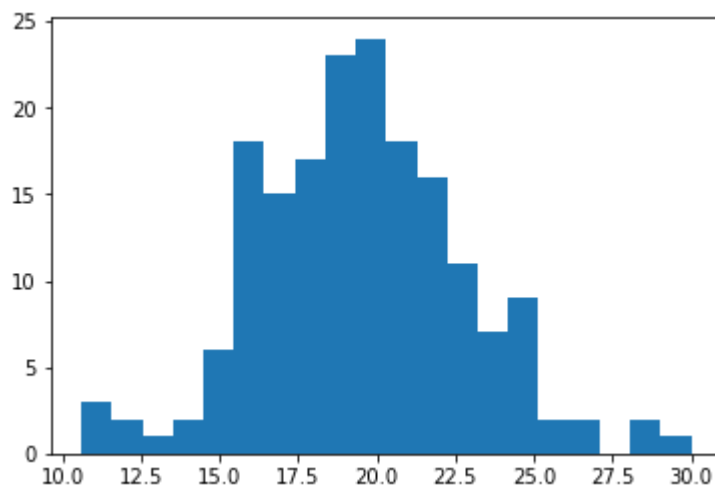
Basados en el índice de correlación de los puntos anteriores, se optó por escoger las variables más correlacionadas en alguna de las clases de vino, para compararla se toman las tres clases de vino y la medida general

Alcalinidad Cenizas

La alcalinidad de las cenizas trata de la suma de los cationes de amonio que se encuentran mezclados en los ácidos orgánicos del vino. Es una variable de tipo Cuantitativo continuo, lo que podría interpretarse como el porcenje.

```
In [52]: print('***** Alcalinidad Cenizas *****')
print("Media = ", df["Alcalinidad Cenizas"].mean())
print("varianza = ", np.var(df["Alcalinidad Cenizas"]))
print("Desviación = ", np.std(df["Alcalinidad Cenizas"]))
print("sknew = ", sp.skew(df["Alcalinidad Cenizas"]))
print("kurtosis = ", sp.kurtosis(df["Alcalinidad Cenizas"]))
print(plt.hist(df["Alcalinidad Cenizas"],bins=20))

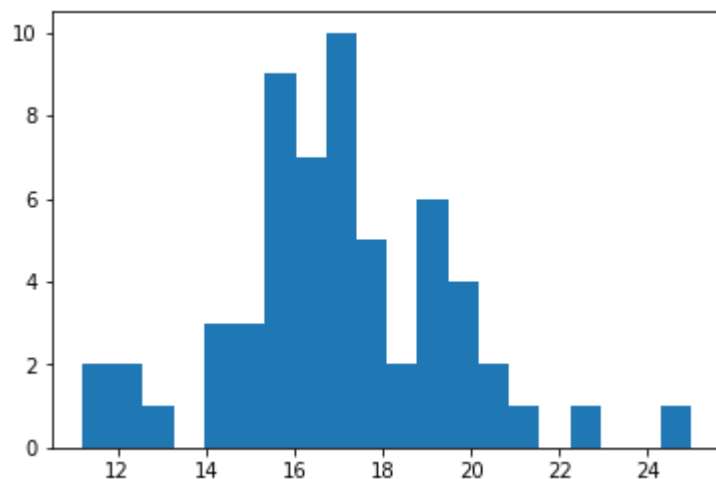
***** Alcalinidad Cenizas *****
Media = 19.52290502793296
varianza = 11.167240722823875
Desviación = 3.341742168813129
sknew = 0.20155959494723427
kurtosis = 0.9993166570687904
(array([ 3.,  2.,  1.,  2.,  6., 18., 15., 17., 23., 24., 18., 16., 11.,
        7.,  9.,  2.,  2.,  0.,  2.,  1.]), array([10.6 , 11.57, 12.54, 13.5
1, 14.48, 15.45, 16.42, 17.39, 18.36,
        19.33, 20.3 , 21.27, 22.24, 23.21, 24.18, 25.15, 26.12, 27.09,
        28.06, 29.03, 30. ]), <a list of 20 Patch objects>)
```



Como se observa en el histograma, la función unimodal positiva (0,2) muestra una alta concentración de Alcalinidad de Cenizas fe forma general en los vinos, el promedio es 19.52 con una desviación de 3.3%.

```
In [43]: print('***** Alcalinidad Cenizas Vino 1 *****')
print("Media = ", dfc1["Alcalinidad Cenizas"].mean())
print("varianza = ", np.var(dfc1["Alcalinidad Cenizas"]))
print("Desviación = ", np.std(dfc1["Alcalinidad Cenizas"]))
print("sknew = ", sp.skew(dfc1["Alcalinidad Cenizas"]))
print("kurtosis = ", sp.kurtosis(dfc1["Alcalinidad Cenizas"]))
print(plt.hist(dfc1["Alcalinidad Cenizas"],bins=20))

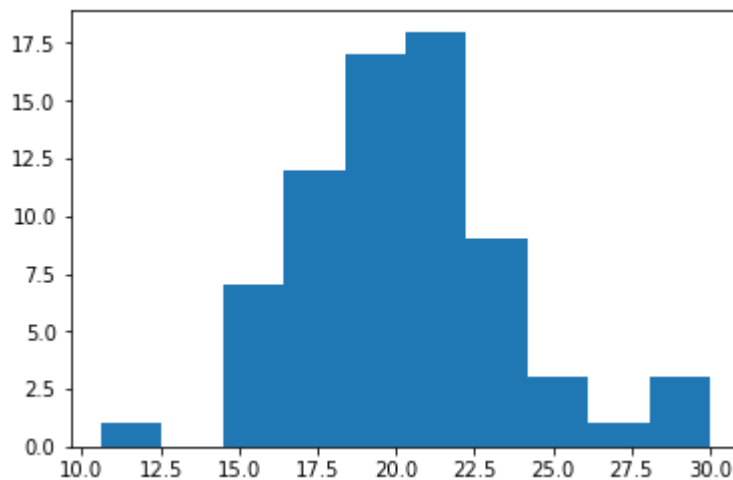
***** Alcalinidad Cenizas Vino 1 *****
Media = 17.037288135593222
varianza = 6.373863832232119
Desviación = 2.524651229820095
sknew = 0.20061132209769655
kurtosis = 0.9993166570687904
(array([ 2.,  2.,  1.,  0.,  3.,  3.,  9.,  7., 10.,  5.,  2.,  6.,  4.,
        2.,  1.,  0.,  1.,  0.,  0.,  1.]), array([11.2 , 11.89, 12.58, 13.2
        7, 13.96, 14.65, 15.34, 16.03, 16.72,
        17.41, 18.1 , 18.79, 19.48, 20.17, 20.86, 21.55, 22.24, 22.93,
        23.62, 24.31, 25. ]), <a list of 20 Patch objects>)
```



En el Vino clase 1 se observa una función unimodal positiva (0,2) con una kurtosis o grosor de curva de 0.99. Se observa una acumulación de Alcalinidad hacia el 17%, como lo marca el promedio, con una desviación estandar de 2,52%.

```
In [47]: print('***** Alcalinidad Cenizas Vino 2 *****')
print("Media = ", dfc2["Alcalinidad Cenizas"].mean())
print("varianza = ", np.var(dfc2["Alcalinidad Cenizas"]))
print("Desviación = ", np.std(dfc2["Alcalinidad Cenizas"]))
print("sknew = ", sp.skew(dfc2["Alcalinidad Cenizas"]))
print("kurtosis = ", sp.kurtosis(dfc2["Alcalinidad Cenizas"]))
print(plt.hist(dfc2["Alcalinidad Cenizas"]))

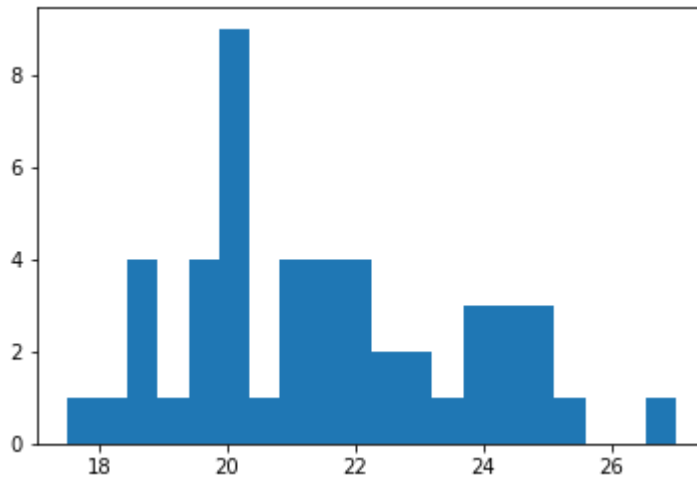
***** Alcalinidad Cenizas Vino 2 *****
Media = 20.238028169014086
varianza = 11.062920055544541
Desviación = 3.326096819929411
sknew = 0.421628487093137
kurtosis = 1.0535213097801348
(array([ 1.,  0.,  7., 12., 17., 18.,  9.,  3.,  1.,  3.]), array([10.6 , 12.
54, 14.48, 16.42, 18.36, 20.3 , 22.24, 24.18, 26.12,
28.06, 30.  ]), <a list of 10 Patch objects>)
```



En el Vino clase 2 se observa una función unimodal positiva (0,42) con una kurtosis o grosor de curva de 1.5. Se observa el promedio en el pico más alto, hacia los 20,2 con una desviación estandar de 3,3.

```
In [47]: print('***** Alcalinidad Cenizas Vino 3 *****')
print("Media = ", dfc3["Alcalinidad Cenizas"].mean())
print("varianza = ", np.var(dfc3["Alcalinidad Cenizas"]))
print("Desviación = ", np.std(dfc3["Alcalinidad Cenizas"]))
print("sknew = ", sp.skew(dfc3["Alcalinidad Cenizas"]))
print("kurtosis = ", sp.kurtosis(dfc3["Alcalinidad Cenizas"]))
print(plt.hist(dfc3["Alcalinidad Cenizas"], bins=20))

***** Alcalinidad Cenizas Vino 3 *****
Media = 21.479591836734695
varianza = 5.081216159933362
Desviación = 2.2541553096300535
sknew = 0.40088838747080563
kurtosis = -0.7018016678984154
(array([1., 1., 4., 1., 4., 9., 1., 4., 4., 4., 2., 2., 1., 3., 3., 3., 1.,
        0., 0., 1.]), array([17.5, 17.975, 18.45, 18.925, 19.4, 19.875, 20.35,
        20.825, 21.3, 21.775, 22.25, 22.725, 23.2, 23.675, 24.15, 24.625,
        25.1, 25.575, 26.05, 26.525, 27. ]), <a list of 20 Patch objects
>)
```



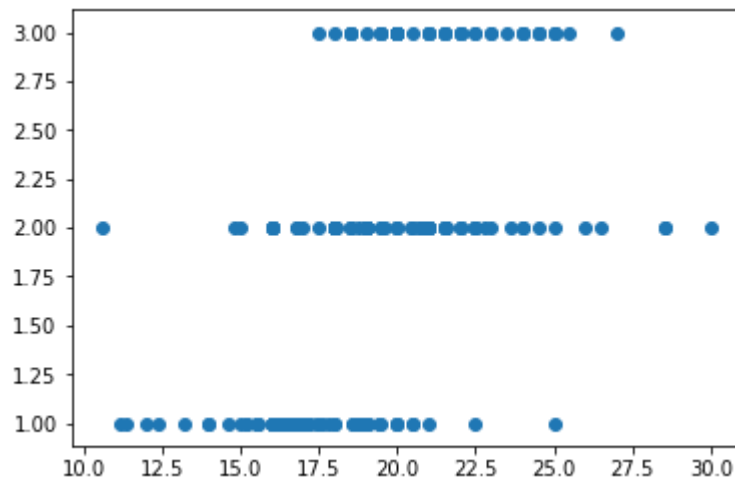
En el Vino clase 3 se observa una función unimodal positiva (0.4). El promedio de Alcalinidad se encuentra hacia los 21.47 con una desviación estandar de 2.25.

Conclusión sobre la alcalinidad

De las tres clases de vino, el dos es quien posee una más amplia concentración de Alcalinidad de Cenizas, que es la suma de cationes presentes en los ácidos del vino. El Vino 1 y el vino 3, por el contrario, tienen de forma más reducida el grado de alcalinidad. A continuación el el siguiente diagrama, se observa con mayor detalle la acumulación de alcalinidad por clase de vino.


```
In [51]: plt.scatter(df["Alcalinidad Cenizas"],df["Clase Vino"])
```

```
Out[51]: <matplotlib.collections.PathCollection at 0x2401c842b70>
```

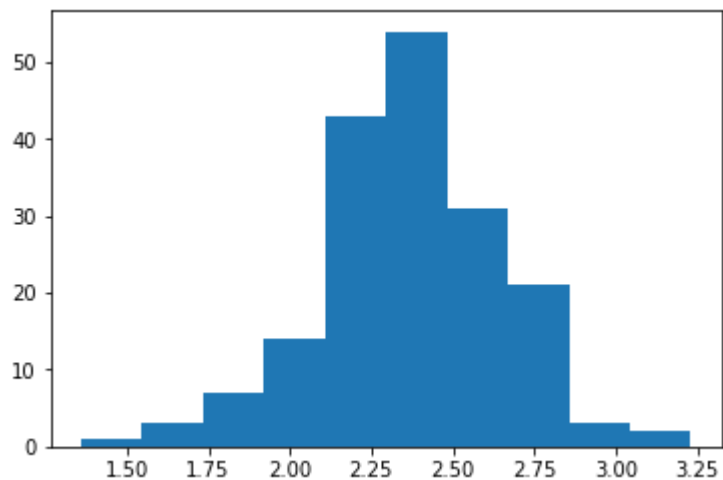


Cenizas

Las cenizas son un conjunto de productos obtenidos como el resultado de la incineración de residuos de evaporación del vino, llevada a cabo esta incineración para la obtención de la totalidad de los cationes. Esta es una variable cuantitativa continua, lo que podría interpretarse como una medida de porcentaje.

```
In [49]: print('***** Cenizas *****')
print("Media = ", df["Cenizas"].mean())
print("varianza = ", np.var(df["Cenizas"]))
print("Desviación = ", np.std(df["Cenizas"]))
print("sknew = ", sp.skew(df["Cenizas"]))
print("kurtosis = ", sp.kurtosis(df["Cenizas"]))
print(plt.hist(df["Cenizas"]))

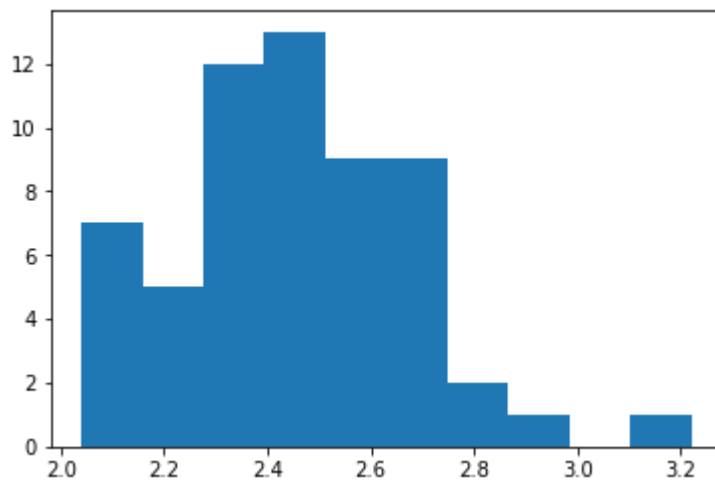
***** Cenizas *****
Media = 2.3686033519553082
varianza = 0.07519860803345718
Desviación = 0.2742236460144478
sknew = -0.18170224181554073
kurtosis = 1.0417989982637899
(array([ 1.,  3.,  7., 14., 43., 54., 31., 21.,  3.,  2.]), array([1.36 , 1.5
47, 1.734, 1.921, 2.108, 2.295, 2.482, 2.669, 2.856,
3.043, 3.23 ]), <a list of 10 Patch objects>)
```



Por el contrario, en cuanto a la cantidad de Cenizas, se observa una función unimodal negativa (0,18) y un grosor de columna de (1.4). El promedio de cenizas en las tres clases de vino es de 2,37 con una desviación de 0.2. Lo que muestra una concentración mayor de cenizas.

```
In [50]: print('***** Cenizas Clase 1 *****')
print("Media = ", dfc1["Cenizas"].mean())
print("varianza = ", np.var(dfc1["Cenizas"]))
print("Desviación = ", np.std(dfc1["Cenizas"]))
print("sknew = ", sp.skew(dfc1["Cenizas"]))
print("kurtosis = ", sp.kurtosis(dfc1["Cenizas"]))
print(plt.hist(dfc1["Cenizas"]))

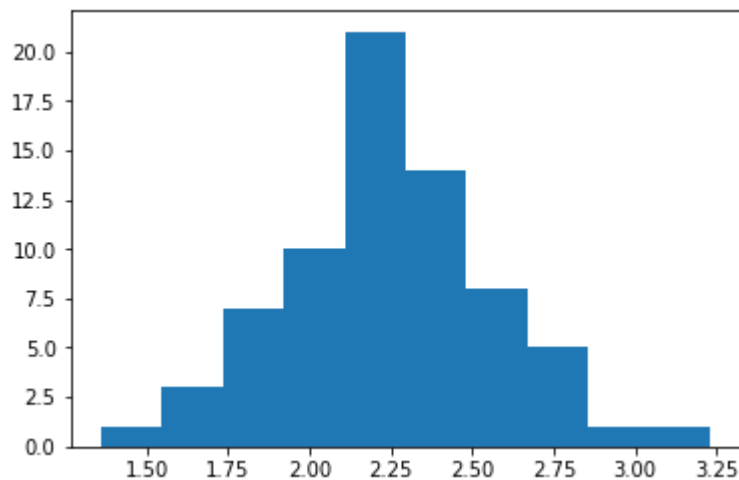
***** Cenizas Clase 1 *****
Media = 2.455593220338984
varianza = 0.05072973283539215
Desviación = 0.22523261938580777
sknew = 0.5298835756421636
kurtosis = 0.7043188473937296
(array([ 7.,  5., 12., 13.,  9.,  9.,  2.,  1.,  0.,  1.]), array([2.04 , 2.158, 2.276, 2.394, 2.512, 2.63 , 2.748, 2.866, 2.984, 3.102, 3.22 ]), <a list of 10 Patch objects>)
```



Para el Vino de clase 1, se muestra una función unimodal positiva (0.5). El promedio para la clase 1 se encuentra sobre el 2.45c on una desviación de 0.22.

```
In [51]: print('***** Cenizas Clase 2 *****')
print("Media = ", dfc2["Cenizas"].mean())
print("varianza = ", np.var(dfc2["Cenizas"]))
print("Desviación = ", np.std(dfc2["Cenizas"]))
print("sknew = ", sp.skew(dfc2["Cenizas"]))
print("kurtosis = ", sp.kurtosis(dfc2["Cenizas"]))
print(plt.hist(dfc2["Cenizas"]))

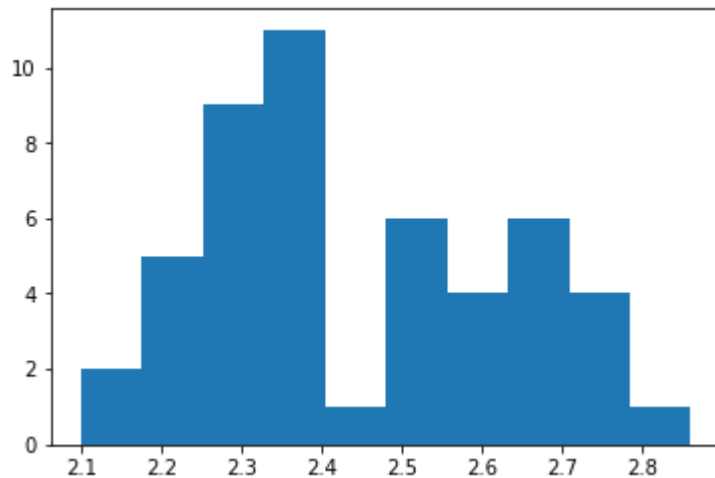
***** Cenizas Clase 2 *****
Media = 2.244788732394365
varianza = 0.09811791311247767
Desviación = 0.3132377900453227
sknew = 0.17184850315027558
kurtosis = 0.7800797821458438
(array([ 1.,  3.,  7., 10., 21., 14.,  8.,  5.,  1.,  1.]), array([1.36 , 1.5
47, 1.734, 1.921, 2.108, 2.295, 2.482, 2.669, 2.856,
3.043, 3.23 ]), <a list of 10 Patch objects>)
```



Para el Vino de clase 2, se muestra una función unimodal positiva (0.17). El promedio para la clase 2 se encuentra sobre el 2.24 on una desviación de 0.31. Hasta el momento, se puede decir que hay más cocentración de Cenizas en esta clase que en la anterior.

```
In [52]: print('***** Cenizas Clase 3 *****')
print("Media = ", dfc3["Cenizas"].mean())
print("varianza = ", np.var(dfc3["Cenizas"]))
print("Desviación = ", np.std(dfc3["Cenizas"]))
print("sknew = ", sp.skew(dfc3["Cenizas"]))
print("kurtosis = ", sp.kurtosis(dfc3["Cenizas"]))
print(plt.hist(dfc3["Cenizas"]))

***** Cenizas Clase 3 *****
Media = 2.4432653061224485
varianza = 0.03455260308204915
Desviación = 0.1858833050116367
sknew = 0.3174907199419924
kurtosis = -0.9207055461949838
(array([ 2.,  5.,  9., 11.,  1.,  6.,  4.,  6.,  4.,  1.]), array([2.1, 2.176, 2.252, 2.328, 2.404, 2.48, 2.556, 2.632, 2.708, 2.784, 2.86 ]), <a list of 10 Patch objects>)
```



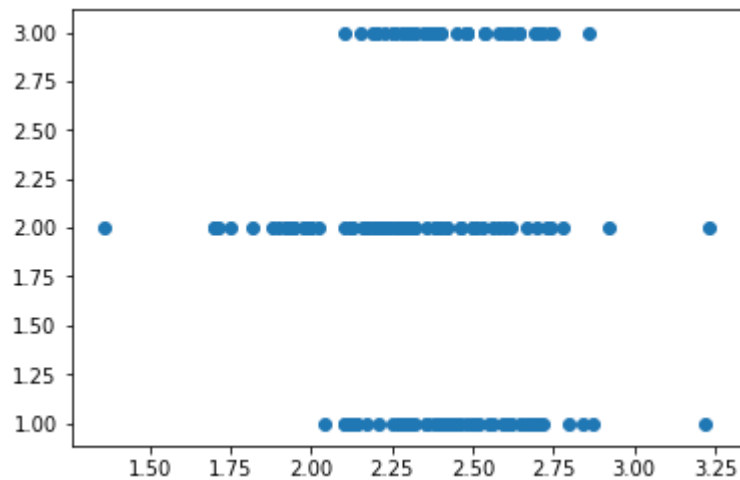
Para el Vino de clase 3, se muestra una función unimodal positiva (0.3). El promedio para la clase 3 se encuentra sobre el 2.44 on una desviación de 0.18.

Conclusión sobre la ceniza

Con respecto a la variable anterior, se puede decir que la cantidad de cenizas es más relevante en el sabor del vino que la alcalinidad, definiendo esta última como la acumulación de cationes resultantes de la mezcla, mientras que la ceniza en si muestra el grado de evaporación de los ingredientes en el proceso de preparación. Lo que se observa con más detalle en la siguiente gráfica:

```
In [54]: plt.scatter(df["Cenizas"], df["Clase Vino"])
```

```
Out[54]: <matplotlib.collections.PathCollection at 0x2401ca87278>
```

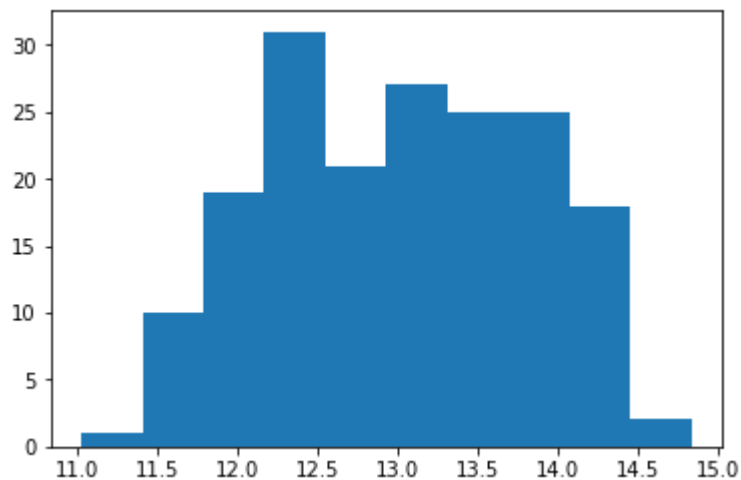


Alcohol

El alcohol del vino depende del que produzca su fermentación. El máximo grado de alcohol que suelen tener los vinos es de 15 grados, a excepción de los vinos de Jerez y de Oporto, que como vinos licorosos, generosos o fortalecidos tienen una graduación más elevada.

```
In [53]: print('***** Alcohol *****')
print("Media = ", df["Alcohol"].mean())
print("varianza = ", np.var(df["Alcohol"]))
print("Desviación = ", np.std(df["Alcohol"]))
print("sknew = ", sp.skew(df["Alcohol"]))
print("kurtosis = ", sp.kurtosis(df["Alcohol"]))
print(plt.hist(df["Alcohol"]))

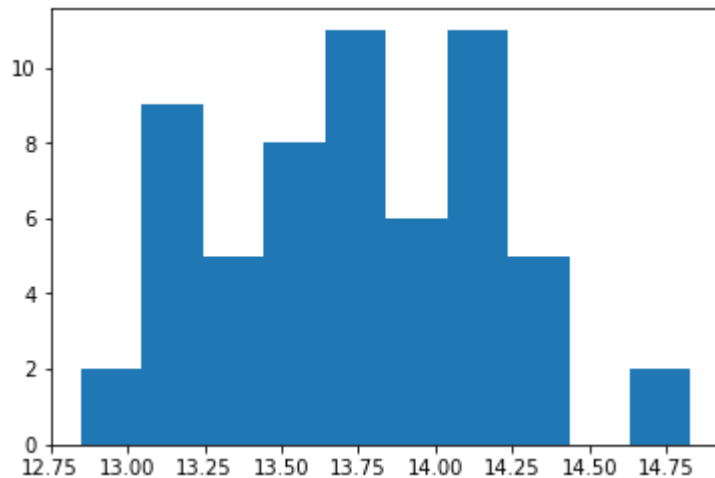
***** Alcohol *****
Media = 13.00692737430167
varianza = 0.6587844137199215
Desviación = 0.8116553540265237
sknew = -0.058637293968574894
kurtosis = -0.8738442242026108
(array([ 1., 10., 19., 31., 21., 27., 25., 25., 18., 2.]), array([11.03, 11.41, 11.79, 12.17, 12.55, 12.93, 13.31, 13.69, 14.07, 14.45, 14.83]), <a list of 10 Patch objects>)
```



De forma general, en las tres clases de vino se observa un contenido de alcohol muy variante sin ser un factor diferenciador entre uno y otro. Aunque se muestra un promedio de 13, hay una gran desviación de 0.8 indicando que el grado de alcohol es muy diverso y poco decisivo entre uno y otro. La función unimodal es negativa con una inclinación hacia la izquierda de -0.05.

```
In [54]: print('***** Alcohol Clase 1 *****')
print("Media = ", dfc1["Alcohol"].mean())
print("varianza = ", np.var(dfc1["Alcohol"]))
print("Desviación = ", np.std(dfc1["Alcohol"]))
print("sknew = ", sp.skew(dfc1["Alcohol"]))
print("kurtosis = ", sp.kurtosis(dfc1["Alcohol"]))
print(plt.hist(dfc1["Alcohol"]))

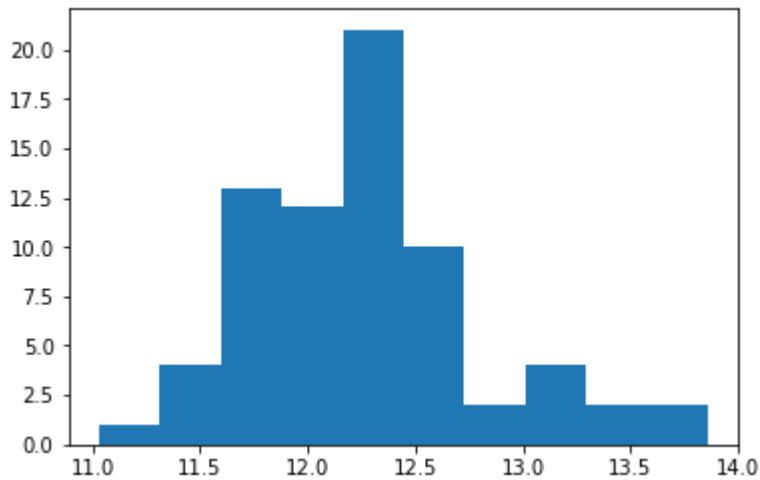
***** Alcohol Clase 1 *****
Media = 13.744745762711865
varianza = 0.20994018960068944
Desviación = 0.4581923063525723
sknew = 0.06923898186233954
kurtosis = -0.5967909626181376
(array([ 2.,  9.,  5.,  8., 11.,  6., 11.,  5.,  0.,  2.]), array([12.85, 13.048, 13.246, 13.444, 13.642, 13.84, 14.038, 14.236, 14.434, 14.632, 14.83 ]), <a list of 10 Patch objects>)
```



De la misma forma que la descripción general, en el Vino clase uno no se percibe una importancia relevante de alcohol en su preparación pues los registros sobre esta clase y en esta variable tienen una dispersión de 0.45. La función descrita es casi bimodal, con una inclinación positiva de 0.07


```
In [55]: print('***** Alcohol Clase 2 *****')
print("Media = ", dfc2["Alcohol"].mean())
print("varianza = ", np.var(dfc2["Alcohol"]))
print("Desviación = ", np.std(dfc2["Alcohol"]))
print("sknew = ", sp.skew(dfc2["Alcohol"]))
print("kurtosis = ", sp.kurtosis(dfc2["Alcohol"]))
print(plt.hist(dfc2["Alcohol"]))

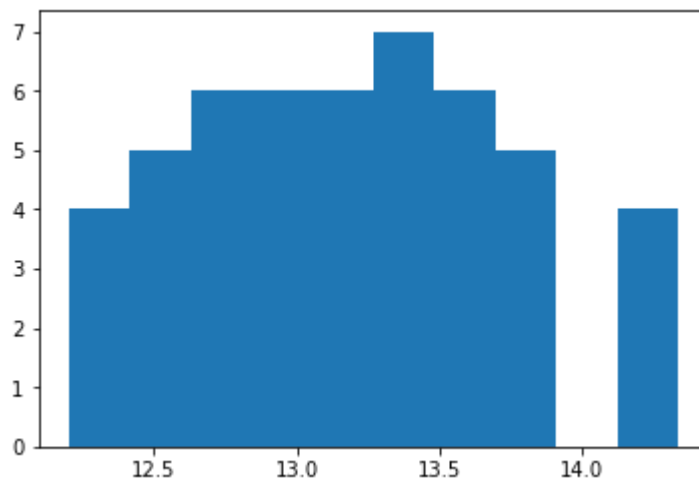
***** Alcohol Clase 2 *****
Media = 12.278732394366198
varianza = 0.2853293790914501
Desviación = 0.534162315304487
sknew = 0.567868111960227
kurtosis = 0.6141811846879026
(array([ 1.,  4., 13., 12., 21., 10.,  2.,  4.,  2.,  2.]), array([11.03 , 11.313, 11.596, 11.879, 12.162, 12.445, 12.728, 13.011, 13.294, 13.577, 13.86 ]), <a list of 10 Patch objects>)
```



En la clase 2, por ejemplo, si se observa un promedio más recogido de Alcohol en 12.27 con una desviación de 0.5, menor que en la anterior. Por tanto, se puede decir que en el vino de clase 2 el nivel de alcohol está mas concentrado en los diferentes registros. La función unimodal es positiva con un desequilibrio de 0.56.

```
In [56]: print('***** Alcohol Clase 3 *****')
print("Media = ", dfc3["Alcohol"].mean())
print("varianza = ", np.var(dfc3["Alcohol"]))
print("Desviación = ", np.std(dfc3["Alcohol"]))
print("sknew = ", sp.skew(dfc3["Alcohol"]))
print("kurtosis = ", sp.kurtosis(dfc3["Alcohol"]))
print(plt.hist(dfc3["Alcohol"]))

***** Alcohol Clase 3 *****
Media = 13.173673469387753
varianza = 0.2887334443981676
Desviación = 0.5373392265582028
sknew = 0.14507160407994793
kurtosis = -0.7290966099985621
(array([4., 5., 6., 6., 6., 7., 6., 5., 0., 4.]), array([12.2 , 12.414, 12.6
28, 12.842, 13.056, 13.27 , 13.484, 13.698,
13.912, 14.126, 14.34 ]), <a list of 10 Patch objects>)
```



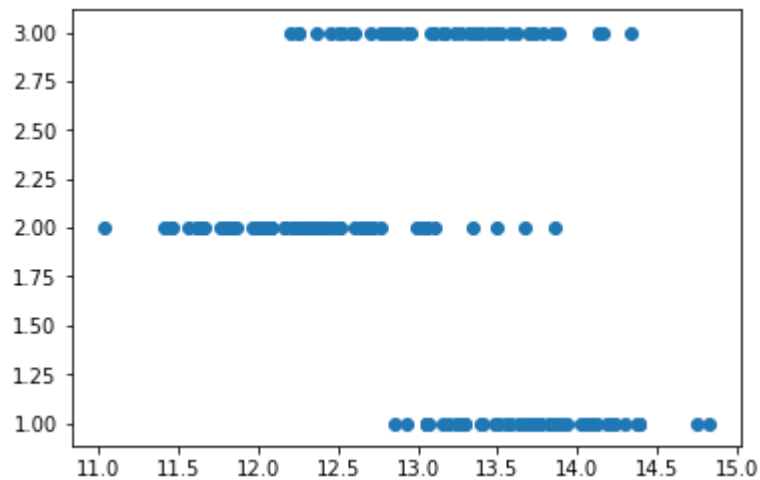
En el vino de clase 3, al igual que en el uno, no se observa una concentración de registros en cuanto al nivel de alcohol, los datos están mas dispersos (0.53) con respecto a la media que es de 13.17.

Conclusión sobre el alcohol

En general, se observa una dispersión mayor en la variable del alcohol según los registros. En promedio, el nivel de alcohol en cada vino no es una variable revelante.

```
In [55]: plt.scatter(df["Alcohol"], df["Clase Vino"])
```

```
Out[55]: <matplotlib.collections.PathCollection at 0x2401b0201d0>
```



```
In [57]: print('***** Magnesio *****')
print("Media = ", df["Magnesio"].mean())
print("varianza = ", np.var(df["Magnesio"]))
print("Desviación = ", np.std(df["Magnesio"]))
print("sknew = ", sp.skew(df["Magnesio"]))
print("kurtosis = ", sp.kurtosis(df["Magnesio"]))
print(plt.hist(df["Magnesio"]))
```

***** Magnesio *****

Media = 99.72067039106145

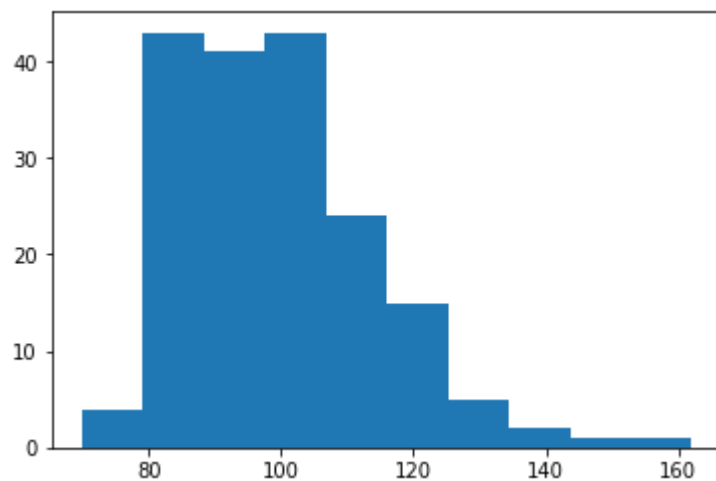
varianza = 201.78789675727955

Desviación = 14.20520667773896

sknew = 1.0956504732726644

kurtosis = 2.0435456692033416

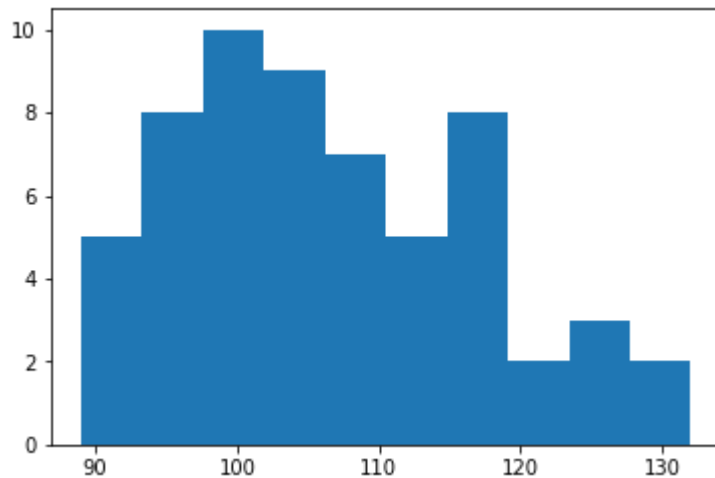
(array([4., 43., 41., 43., 24., 15., 5., 2., 1., 1.]), array([70. , 79.2, 88.4, 97.6, 106.8, 116. , 125.2, 134.4, 143.6, 152.8, 162.]), <a list of 10 Patch objects>)



En forma general, la cantidad de magnesio en la muestra, tiene un promedio de 99,7 con una desviación de 14.2. La función mostrada para la muestra es unimodal positiva con una inclinación de 1.09 y un grosor de 2.04.

```
In [58]: print('***** Magnesio Clase 1*****')
print("Media = ", dfc1["Magnesio"].mean())
print("varianza = ", np.var(dfc1["Magnesio"]))
print("Desviación = ", np.std(dfc1["Magnesio"]))
print("sknew = ", sp.skew(dfc1["Magnesio"]))
print("kurtosis = ", sp.kurtosis(dfc1["Magnesio"]))
print(plt.hist(dfc1["Magnesio"]))

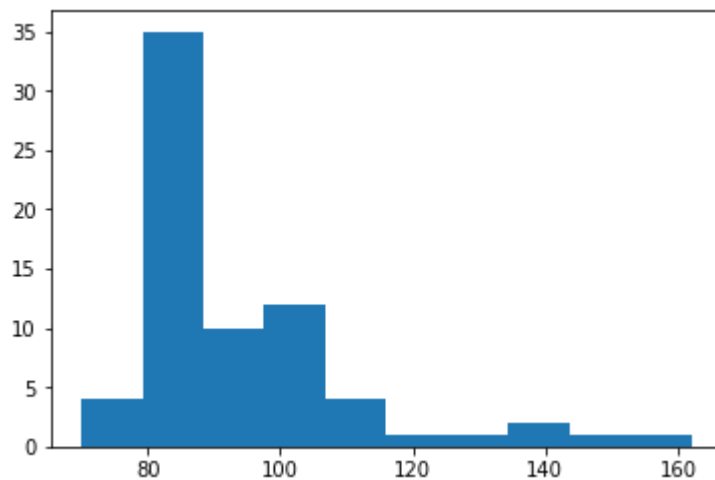
***** Magnesio Clase 1*****
Media = 106.33898305084746
varianza = 108.35966676242458
Desviación = 10.409594937480737
sknew = 0.48083727648064734
kurtosis = -0.5433460015698639
(array([ 5.,  8., 10.,  9.,  7.,  5.,  8.,  2.,  3.,  2.]), array([ 89. ,  93.3,  97.6, 101.9, 106.2, 110.5, 114.8, 119.1, 123.4, 127.7, 132. ]), <a list of 10 Patch objects>)
```



Para el vino de clase 1, se muestra una dispersión mayor en el contenido de magnesio, los vinos de esta clase cuentan entre 90 y 131 mg de magnesio, resaltando una media de 106 con una desviación de 10.4. La función es unimodal con una desviación de 0.5.

```
In [59]: print('***** Magnesio Clase 2*****')
print("Media = ", dfc2["Magnesio"].mean())
print("varianza = ", np.var(dfc2["Magnesio"]))
print("Desviación = ", np.std(dfc2["Magnesio"]))
print("sknew = ", sp.skew(dfc2["Magnesio"]))
print("kurtosis = ", sp.kurtosis(dfc2["Magnesio"]))
print(plt.hist(dfc2["Magnesio"]))

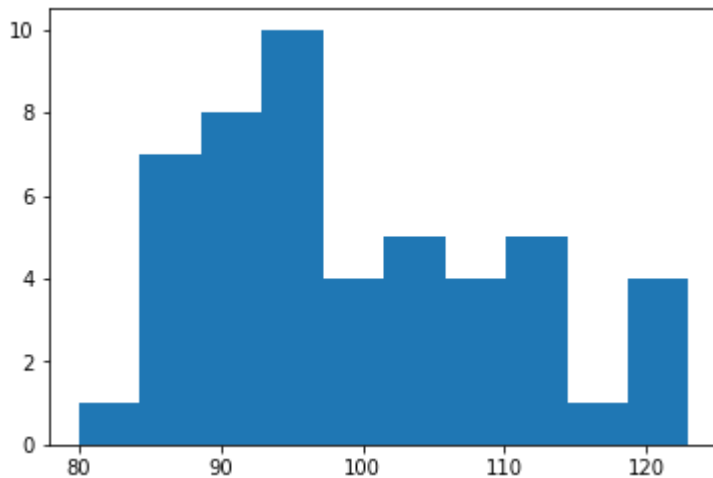
***** Magnesio Clase 2*****
Media = 94.54929577464789
varianza = 276.7264431660386
Desviación = 16.635096728484587
sknew = 2.068766308330599
kurtosis = 4.670458404091595
(array([ 4., 35., 10., 12., 4., 1., 1., 2., 1., 1.]), array([ 70. , 79.2, 88.4, 97.6, 106.8, 116. , 125.2, 134.4, 143.6, 152.8, 162. ]), <a list of 10 Patch objects>)
```



A diferencia de la clase anterior, en los vinos de clase 2, existe una concentración mayor de mg de magnesio, aunque se encuentran desde los 70 hasta los 160, sobresaliendo una media de 94.55 y una desviación de 16. Claramente es una desviación positiva de 2.06

```
In [60]: print('***** Magnesio Clase 3*****')
print("Media = ", dfc3["Magnesio"].mean())
print("varianza = ", np.var(dfc3["Magnesio"]))
print("Desviación = ", np.std(dfc3["Magnesio"]))
print("sknew = ", sp.skew(dfc3["Magnesio"]))
print("kurtosis = ", sp.kurtosis(dfc3["Magnesio"]))
print(plt.hist(dfc3["Magnesio"]))

***** Magnesio Clase 3*****
Media = 99.24489795918367
varianza = 113.98084131611832
Desviación = 10.676181026758506
sknew = 0.5464957464973637
kurtosis = -0.5633341011787278
(array([ 1.,  7.,  8., 10.,  4.,  5.,  4.,  5.,  1.,  4.]), array([ 80. ,  84.3,  88.6,  92.9,  97.2, 101.5, 105.8, 110.1, 114.4, 118.7, 123. ]), <a list of 10 Patch objects>)
```



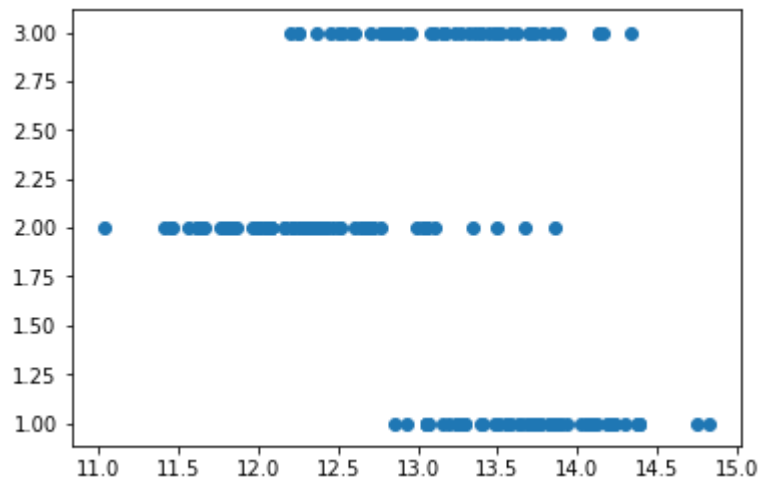
La clase 3, por el contrario, tiene una dispersión mayor, de forma similar a la clase 1. El promedio mostrado en la función es de 99.25. La función unimodal es positiva 0.55

Conclusión

El magnesio, definitivamente, es un factor diferenciador en las tres clases de vino, en particular resalta la cantidad de magnesio mostrada en el vino de clase 2. A continuación se muestra el grado de dispersión sobre esta variable para los tres vinos:

```
In [62]: plt.scatter(df["Alcohol"], df["Clase Vino"])
```

```
Out[62]: <matplotlib.collections.PathCollection at 0x2401c6778d0>
```

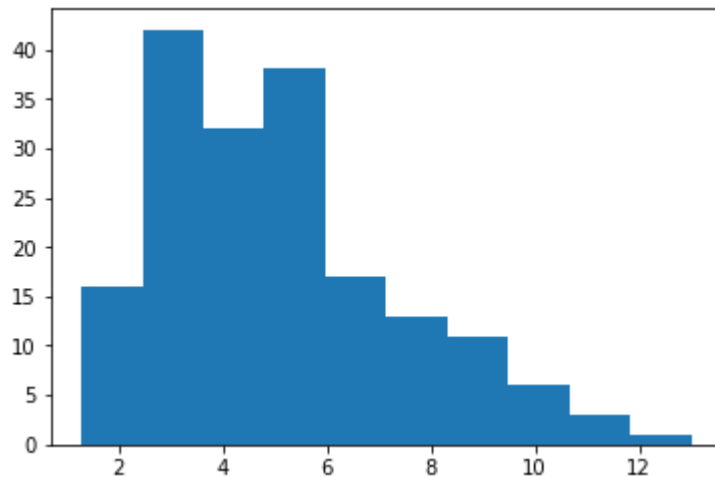


Intensidad Color

La intensidad de color de un vino hace referencia al grado en que la luz lo puede atravesar, al grado de opacidad del vino, cuando se observa el centro de la elipse que forma el vino en una copa inclinada.

```
In [61]: #Intensidad Color
print('***** Intensidad Color *****')
print("Media = ", df["Intensidad Color"].mean())
print("varianza = ", np.var(df["Intensidad Color"]))
print("Desviación = ", np.std(df["Intensidad Color"]))
print("sknew = ", sp.skew(df["Intensidad Color"]))
print("kurtosis = ", sp.kurtosis(df["Intensidad Color"]))
print(plt.hist(df["Intensidad Color"]))

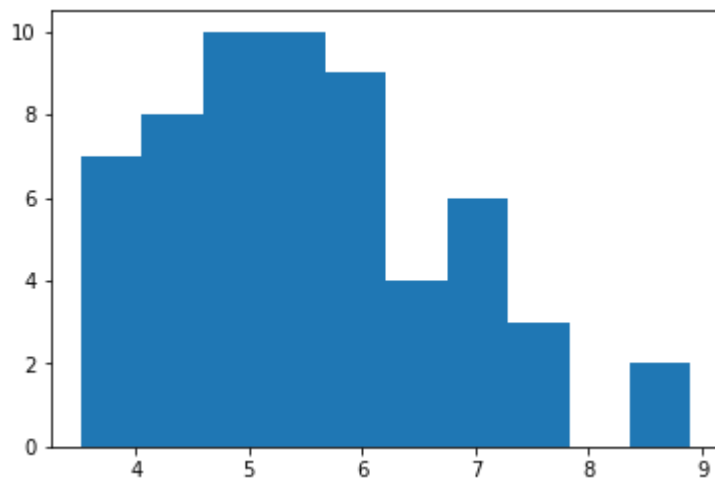
***** Intensidad Color *****
Media = 5.081229044692739
varianza = 5.409704580845796
Desviación = 2.325877163748291
sknew = 0.8426439704843415
kurtosis = 0.2609649250665993
(array([16., 42., 32., 38., 17., 13., 11., 6., 3., 1.]), array([ 1.28 ,
2.452, 3.624, 4.796, 5.968, 7.14 , 8.312, 9.484,
10.656, 11.828, 13. ]), <a list of 10 Patch objects>)
```



Para toda la muestra, el los vinos muestran una intensidad de color presentada desde 1 hasta 13 puntos. La media general está mostrada sobre 5.08 y una desviación de 5.5. La función es unimodal positiva (0.8)


```
In [65]: print('***** Intensidad Color Vino 1*****')
print("Media = ", dfc1["Intensidad Color"].mean())
print("varianza = ", np.var(dfc1["Intensidad Color"]))
print("Desviación = ", np.std(dfc1["Intensidad Color"]))
print("sknew = ", sp.skew(dfc1["Intensidad Color"]))
print("kurtosis = ", sp.kurtosis(dfc1["Intensidad Color"]))
print(plt.hist(dfc1["Intensidad Color"])))

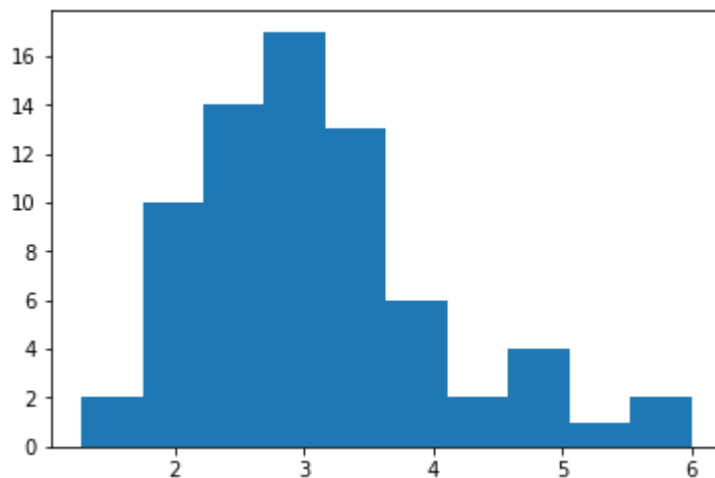
***** Intensidad Color Vino 1*****
Media =  5.528305084745763
varianza =  1.5080615340419419
Desviación =  1.2280315688295402
sknew =  0.5749398979172918
kurtosis =  0.01631563579778339
(array([ 7.,  8., 10., 10.,  9.,  4.,  6.,  3.,  0.,  2.]), array([3.52 , 4.0
58, 4.596, 5.134, 5.672, 6.21 , 6.748, 7.286, 7.824,
      8.362, 8.9  ]), <a list of 10 Patch objects>)
```



Para el vino 1, la intensidad de color está entre 3 y 8, con un promedio de 5.5 y una desviación de 1.2.

```
In [63]: print('***** Intensidad Color Vino 2*****')
print("Media = ", dfc2["Intensidad Color"].mean())
print("varianza = ", np.var(dfc2["Intensidad Color"]))
print("Desviación = ", np.std(dfc2["Intensidad Color"]))
print("sknew = ", sp.skew(dfc2["Intensidad Color"]))
print("kurtosis = ", sp.kurtosis(dfc2["Intensidad Color"]))
print(plt.hist(dfc2["Intensidad Color"])))

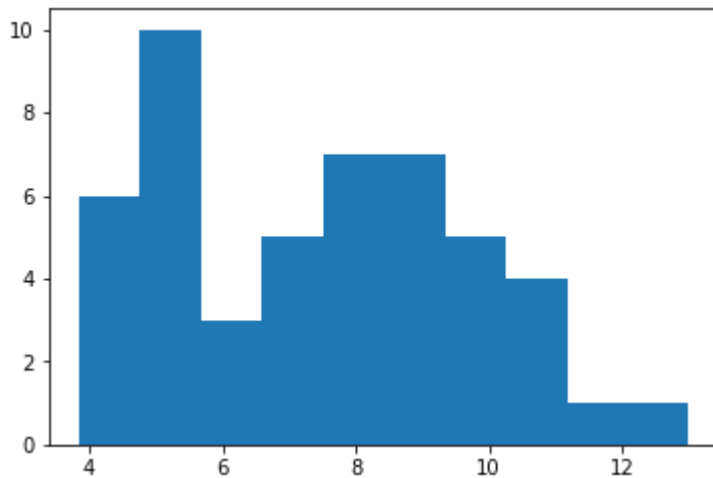
***** Intensidad Color Vino 2*****
Media =  3.08661971830986
varianza =  0.8434449117238643
Desviación =  0.918392569506017
sknew =  1.0177366365997242
kurtosis =  1.142659869202645
(array([ 2., 10., 14., 17., 13.,  6.,  2.,  4.,  1.,  2.]), array([1.28 , 1.7
52, 2.224, 2.696, 3.168, 3.64 , 4.112, 4.584, 5.056,
5.528, 6.   ]), <a list of 10 Patch objects>)
```



Para el vino de clase 2, la intensidad de color está entre 1 y 6. El promedio mostrado por la función es de 3,08 y una desviación de 0.9. La función es unimodal 1.02

```
In [64]: print('***** Intensidad Color Vino 3*****')
print("Media = ", dfc3["Intensidad Color"].mean())
print("varianza = ", np.var(dfc3["Intensidad Color"]))
print("Desviación = ", np.std(dfc3["Intensidad Color"]))
print("sknew = ", sp.skew(dfc3["Intensidad Color"]))
print("kurtosis = ", sp.kurtosis(dfc3["Intensidad Color"]))
print(plt.hist(dfc3["Intensidad Color"]))

***** Intensidad Color Vino 3*****
Media = 7.433061204081633
varianza = 5.187519099641835
Desviación = 2.2776125876983193
sknew = 0.2510725826044246
kurtosis = -0.8476976039002846
(array([ 6., 10., 3., 5., 7., 7., 5., 4., 1., 1.]), array([ 3.85 ,
4.765, 5.68 , 6.595, 7.51 , 8.425, 9.34 , 10.255,
11.17 , 12.085, 13. ])), <a list of 10 Patch objects>)
```



La intensidad de color en el vino de clase 3 es mayor con respecto a los dos anteriores. Para este vino está entre 4 y 13, con un promedio de 7.5 y una desviación de 5. La función es unimodal positiva de 0.2.

Conclusión

De los tres vinos, el 3 es el más oscuro. Su nivel de color es más intenso, lo que puede darse como resultado de la combinación de ingredientes.

Parte 2

A partir de matrices de correlación y scatterplot indique su análisis (correlaciones fuertes principalmente)

```
In [66]: #matriz de correlaciones por tipo de vino  
df.loc[:, "Clase Vino": "Prolina"].groupby("Clase Vino").corr()
```

Out[66]:

		Acido Malico	Alcalinidad Cenizas	Alcohol	Cenizas	Fenoles No Flavanoides	Fenoles Totales	FI
Clase	Vino							
1	Acido Malico	1.000000	0.060031	-0.040513	0.026221	-0.089366	-0.083514	
	Alcalinidad Cenizas	0.060031	1.000000	-0.318367	0.549330	0.302262	-0.222999	
	Alcohol	-0.040513	-0.318367	1.000000	-0.148595	0.015659	0.420687	
	Cenizas	0.026221	0.549330	-0.148595	1.000000	0.465901	0.004802	
	Fenoles No Flavanoides	-0.089366	0.302262	0.015659	0.465901	1.000000	-0.016992	
	Fenoles Totales	-0.083514	-0.222999	0.420687	0.004802	-0.016992	1.000000	
	Flavanoides	-0.191278	-0.287103	0.414904	-0.070454	-0.089538	0.803784	
	Intensidad Color	-0.257587	-0.210951	0.408291	-0.124220	-0.152460	0.650086	
	Magnesio	0.079317	0.238337	0.159361	0.382549	0.237248	0.307241	
	Matiz	-0.419981	0.092980	0.080020	0.239214	0.411831	-0.224330	
	OD280_OD315 de Iso vinos diluidos	0.173244	-0.117704	0.069818	-0.081593	-0.323488	0.053165	
	Prolina	-0.372629	-0.122436	0.360646	-0.029525	-0.015333	0.294994	
	Protoantocianinas	-0.080798	-0.173628	0.307571	-0.145471	-0.144535	0.373601	
2	Acido Malico	1.000000	0.237923	-0.021362	0.148708	0.127605	0.039441	
	Alcalinidad Cenizas	0.237923	1.000000	-0.056282	0.695264	0.182617	0.127942	
	Alcohol	-0.021362	-0.056282	1.000000	-0.214851	-0.068159	-0.046321	
	Cenizas	0.148708	0.695264	-0.214851	1.000000	0.299757	0.112146	
	Fenoles No Flavanoides	0.127605	0.182617	-0.068159	0.299757	1.000000	-0.424746	
	Fenoles Totales	0.039441	0.127942	-0.046321	0.112146	-0.424746	1.000000	
	Flavanoides	0.111932	0.311356	-0.038247	0.314937	-0.235258	0.770999	
	Intensidad Color	-0.203258	-0.085862	0.269789	0.060247	0.018537	0.169072	
	Magnesio	-0.076788	0.003263	-0.029911	0.129130	-0.194138	0.070085	
	Matiz	-0.407986	-0.076825	-0.002038	-0.031244	-0.033717	0.039685	
	OD280_OD315 de Iso vinos diluidos	0.157838	0.382078	-0.130313	0.160590	-0.413119	0.484666	
	Prolina	-0.224214	-0.014507	0.043174	0.041959	-0.152688	0.016927	
	Protoantocianinas	0.210541	0.108838	-0.189617	0.042955	-0.321587	0.382578	
3	Acido Malico	1.000000	0.102580	0.132197	0.040791	0.153949	-0.143099	
	Alcalinidad Cenizas	0.102580	1.000000	0.248752	0.768735	0.008476	0.382379	
	Alcohol	0.132197	0.248752	1.000000	0.289670	0.069891	0.239192	

Clase Vino		Acido Malico	Alcalinidad Cenizas	Alcohol	Cenizas	Fenoles No Flavanoides	Fenoles Totales	FI
	Cenizas	0.040791	0.768735	0.289670	1.000000	0.010237	0.486916	
	Fenoles No Flavanoides	0.153949	0.008476	0.069891	0.010237	1.000000	0.339238	
	Fenoles Totales	-0.143099	0.382379	0.239192	0.486916	0.339238	1.000000	
	Flavanoides	-0.281181	0.264269	0.070818	0.267206	-0.630452	0.234686	
	Intensidad Color	-0.149031	0.178252	0.365255	0.146820	0.040399	0.346686	
	Magnesio	-0.178351	0.147762	-0.092508	0.195295	-0.506631	-0.045668	
	Matiz	0.069727	0.009349	-0.056326	0.153835	0.139320	-0.039277	
	OD280_OD315 de Iso vinos diluidos	0.002683	0.032304	0.116216	0.208914	0.297999	0.191455	
	Prolina	-0.005952	-0.113336	-0.108169	-0.162911	0.188398	0.029321	
	Protoantocianinas	-0.214767	0.271190	0.380828	0.204113	0.178966	0.621897	

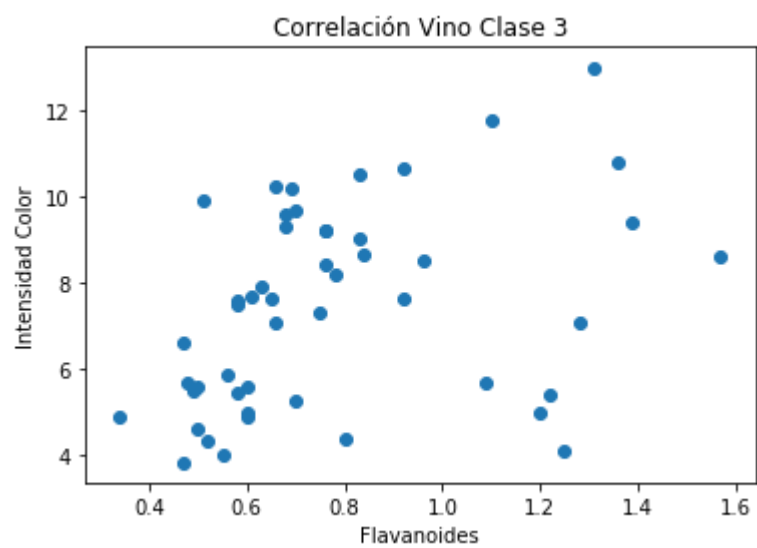
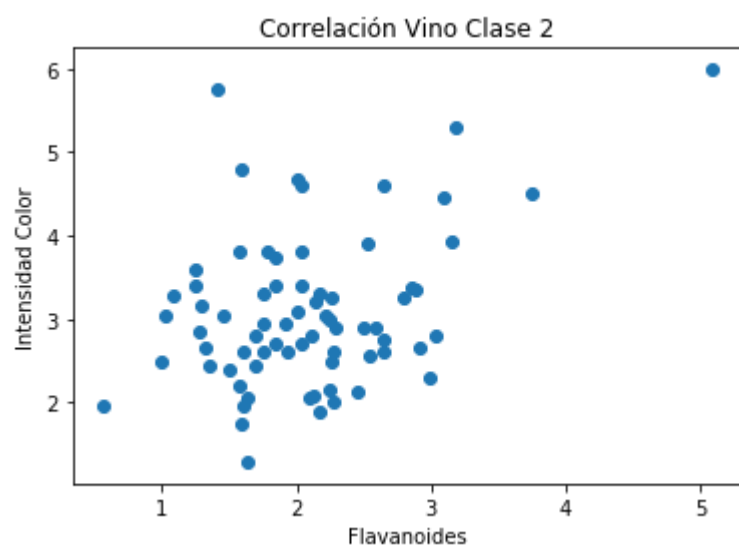
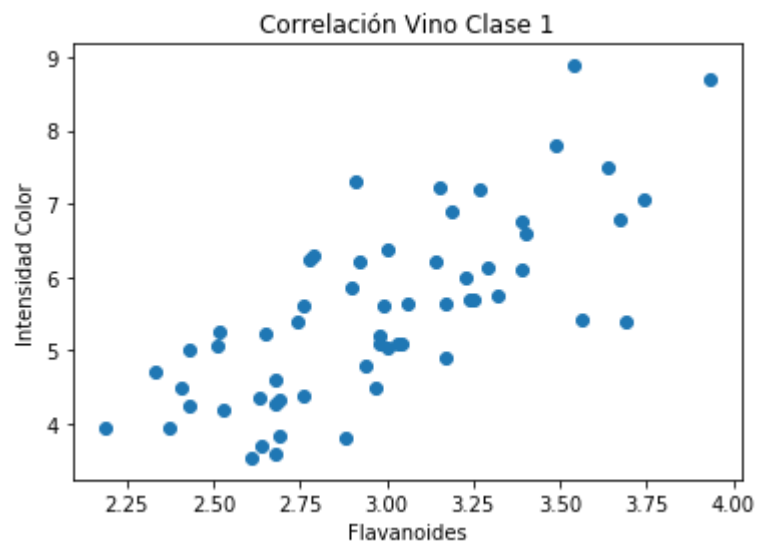
Para el análisis, se toman aquellos ingredientes que muestran una relación superior a los 0.7. De la tabla de correlaciones se observa una relación de Flavanoides con Intensidad de Color, Flavanoides con Fenoles Totales y Cenizas con Intensidad Color. Por tanto, el análisis se realiza sobre estas 3 correlaciones, de forma separada para cada clase de vino.

Relación Flavanoides - Intensidad de Color

```
In [66]: #Obtenemos la dispersión para los vinos de clase 1
plt.scatter(df1["Flavonoides"],df1["Intensidad Color"])
plt.title("Correlación Vino Clase 1")
plt.xlabel("Flavonoides")
plt.ylabel("Intensidad Color")
plt.show()

plt.scatter(df2["Flavonoides"],df2["Intensidad Color"])
plt.title("Correlación Vino Clase 2")
plt.xlabel("Flavonoides")
plt.ylabel("Intensidad Color")
plt.show()

plt.scatter(df3["Flavonoides"],df3["Intensidad Color"])
plt.title("Correlación Vino Clase 3")
plt.xlabel("Flavonoides")
plt.ylabel("Intensidad Color")
plt.show()
```



Conclusión

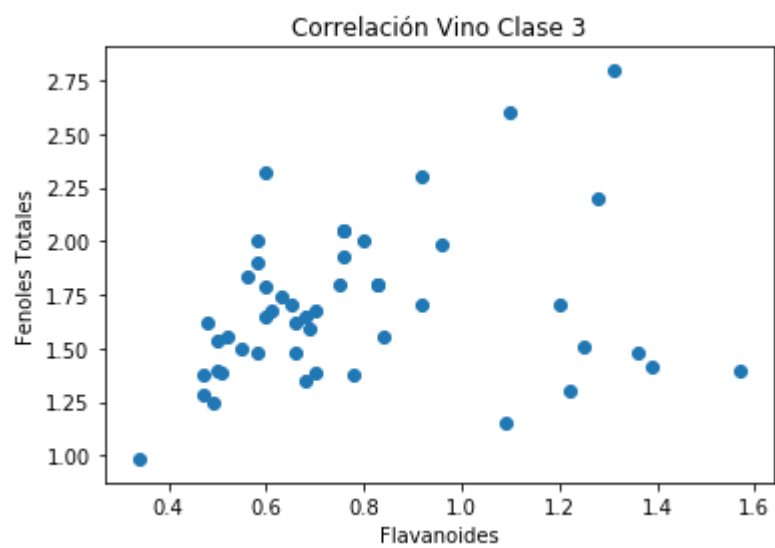
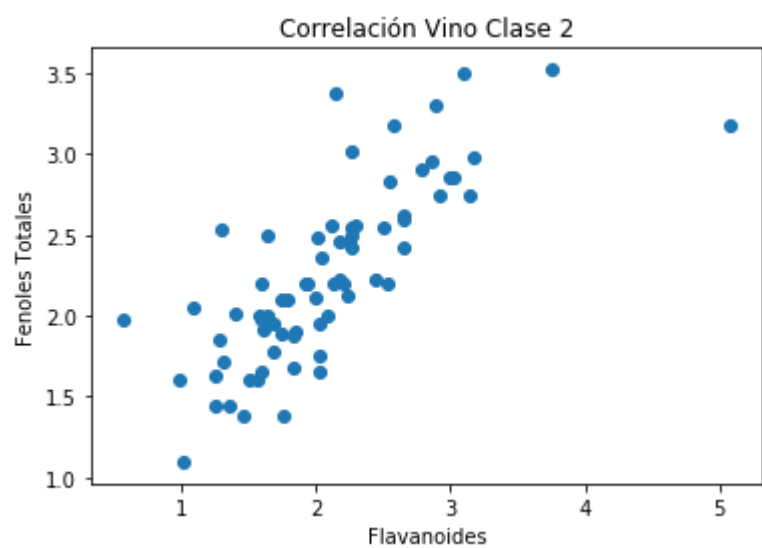
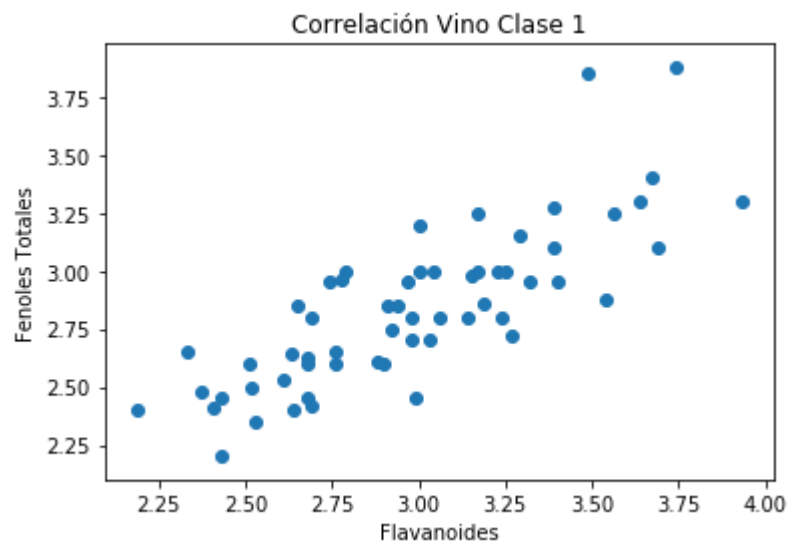
En el vino de clase 1, existe una mayor relación entre los flavanoides y la intensidad de color. El vino 1 y 3 pueden ser los más oscuros y parte de esa intensidad de color puede darse gracias a los flavanoides.

Relación Flavanoides - Fenoles Totales

```
In [67]: plt.scatter(df1["Flavanoides"],df1["Fenoles Totales"])
plt.title("Correlación Vino Clase 1")
plt.xlabel("Flavanoides")
plt.ylabel("Fenoles Totales")
plt.show()

plt.scatter(df2["Flavanoides"],df2["Fenoles Totales"])
plt.title("Correlación Vino Clase 2")
plt.xlabel("Flavanoides")
plt.ylabel("Fenoles Totales")
plt.show()

plt.scatter(df3["Flavanoides"],df3["Fenoles Totales"])
plt.title("Correlación Vino Clase 3")
plt.xlabel("Flavanoides")
plt.ylabel("Fenoles Totales")
plt.show()
```



Conclusión

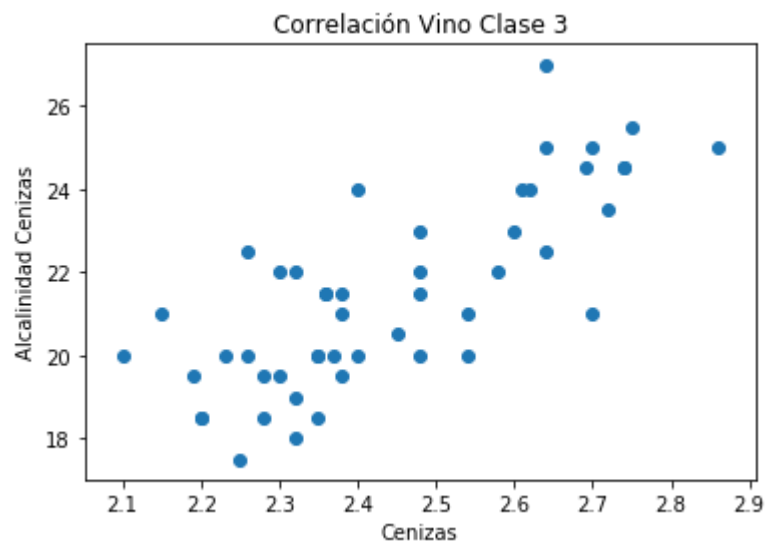
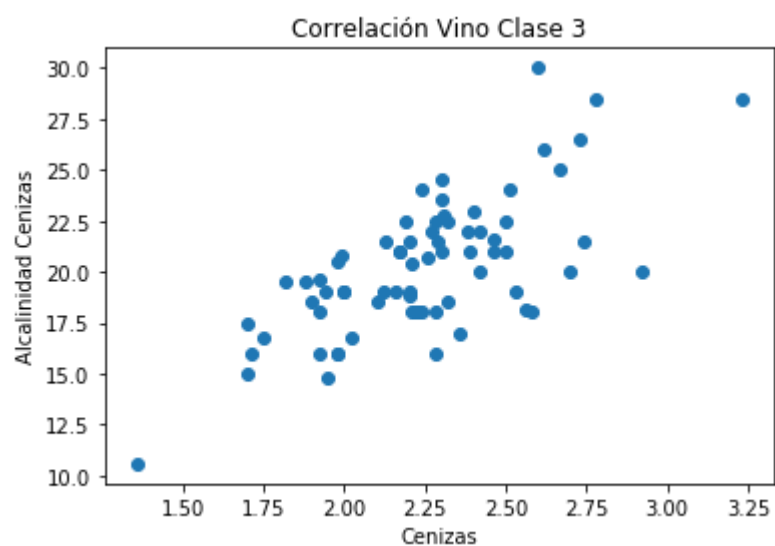
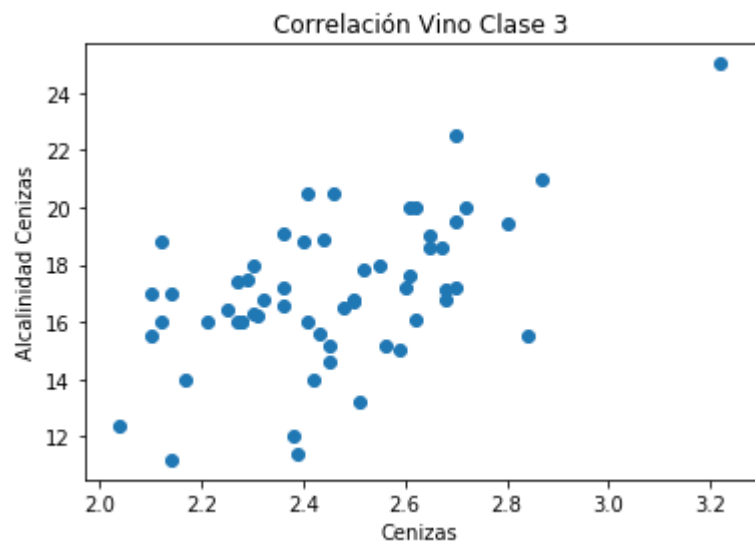
Tanto en el vino de clase 1 y el de clase 2, se observa una mayor relación entre los Flavaniodes y los Fenoles totales.

Relación Cenizas- Alcalinidad Cenizas

```
In [67]: #Obtenemos la dispersión para los vinos de clase 1,
plt.scatter(df1["Cenizas"],df1["Alcalinidad Cenizas"])
plt.title("Correlación Vino Clase 3")
plt.xlabel("Cenizas")
plt.ylabel("Alcalinidad Cenizas")
plt.show()

#Obtenemos la dispersión para los vinos de clase 2
plt.scatter(df2["Cenizas"],df2["Alcalinidad Cenizas"])
plt.title("Correlación Vino Clase 3")
plt.xlabel("Cenizas")
plt.ylabel("Alcalinidad Cenizas")
plt.show()

#Obtenemos la dispersión para los vinos de clase 3
plt.scatter(df3["Cenizas"],df3["Alcalinidad Cenizas"])
plt.title("Correlación Vino Clase 3")
plt.xlabel("Cenizas")
plt.ylabel("Alcalinidad Cenizas")
plt.show()
```



Conclusión

Se muestra una relación mayor de cenizas y alcalinidad cenizas en los vinos de clase 2 y 3.

Identificación de Ruido

```
In [72]: #Utilizamos la siguiente función para identificar los niveles de ruido en el dataset
def detect_outlier(data_1):
    outliers=[]
    threshold=3
    mean_1 = np.mean(data_1)
    std_1 =np.std(data_1)

    for y in data_1:
        z_score= (y - mean_1)/std_1
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers
```

Para conocer las variables con mayor ruido, utilizamos un ciclo e invocamos la función por cada dataset

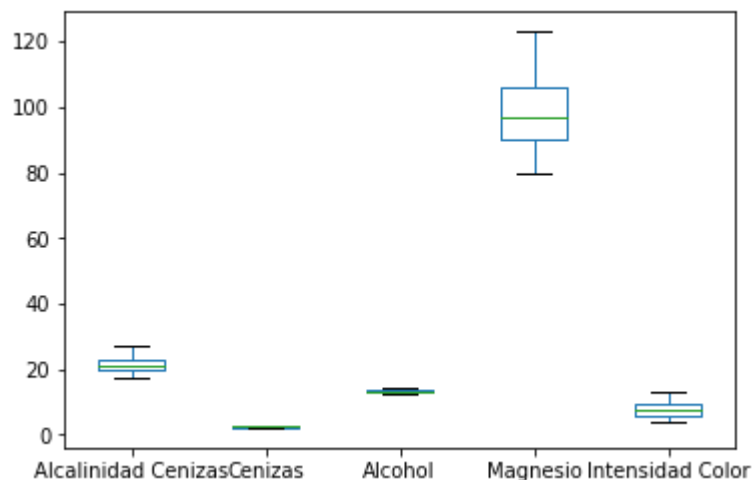
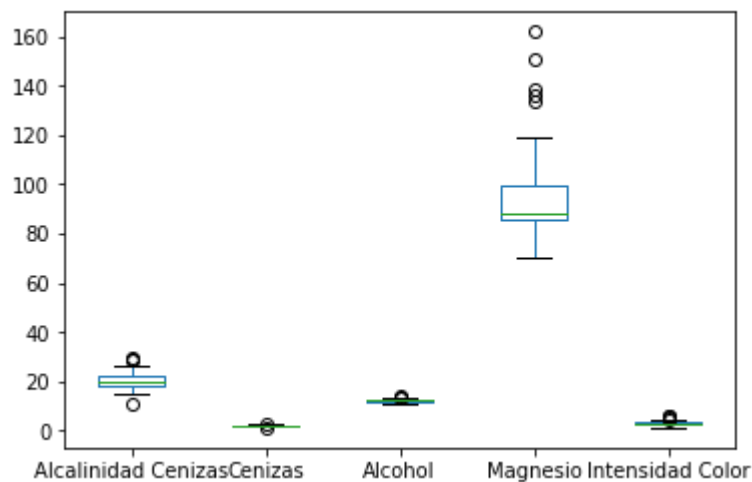
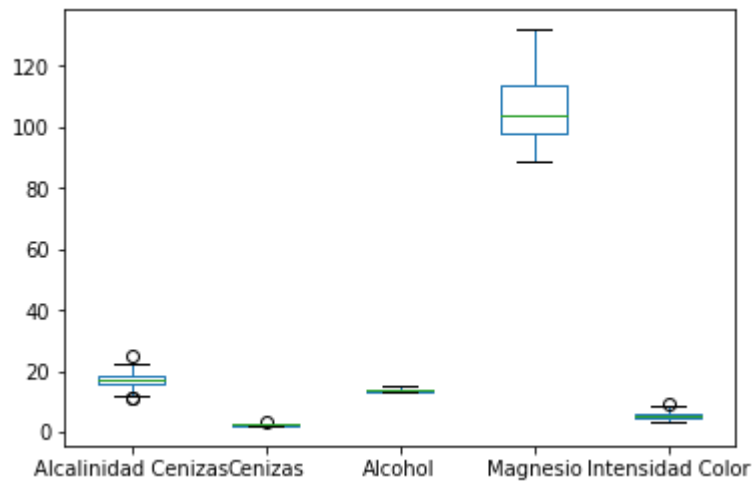
```
In [83]: for column in df.columns:
        print(column + ' -> ' + str(detect_outlier(df[column])))
```

```
ID Caso -> []
Clase Vino -> []
Alcohol -> []
Acido Malico -> [5.8]
Cenizas -> [3.22, 1.36, 3.23]
Alcalinidad Cenizas -> [30.0]
Magnesio -> [151, 162]
Fenoles Totales -> []
Flavanoides -> [5.08]
Fenoles No Flavanoides -> []
Protoantocianinas -> [3.58]
Intensidad Color -> [13.0]
Matiz -> [1.71]
OD280_OD315 de los vinos diluidos -> []
Prolina -> []
```

De acuerdo a lo anterior se elige: Alcalinidad Cenizas, Cenizas, Alcohol, Magnesio e Intensidad Color

```
In [89]: # Con ayuda del boxplot también podemos identificar las variables con mayor ruid
dfc1.loc[:,["Alcalinidad Cenizas", "Cenizas", "Alcohol", "Magnesio", "Intensidad Color"]].plot.box()
dfc2.loc[:,["Alcalinidad Cenizas", "Cenizas", "Alcohol", "Magnesio", "Intensidad Color"]].plot.box()
dfc3.loc[:,["Alcalinidad Cenizas", "Cenizas", "Alcohol", "Magnesio", "Intensidad Color"]].plot.box()
```

Out[89]: <matplotlib.axes._subplots.AxesSubplot at 0x2401d184978>



Conclusión

De los cinco ingredientes elegidos anteriormente, el Magnesio muestra un mayor nivel de ruido seguido por Alcalinidad Cenizas, en cambio, Alcohol no presenta ruido. Usamos la función “detect_outlier(data)” para encontrar el nivel exacto de ruido.

Indique que variables tienen valores faltantes

Recordamos que la cantidad de registros en el dataset es de 179, para buscar los valores faltantes usamos la siguiente función.

```
In [91]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179 entries, 0 to 178
Data columns (total 15 columns):
ID Caso                                179 non-null int64
Clase Vino                             179 non-null int64
Alcohol                                179 non-null float64
Acido Malico                           179 non-null float64
Cenizas                                179 non-null float64
Alcalinidad Cenizas                     179 non-null float64
Magnesio                               179 non-null int64
Fenoles Totales                         179 non-null float64
Flavanoides                            179 non-null float64
Fenoles No Flavanoides                  179 non-null float64
Protoantocianinas                       179 non-null float64
Intensidad Color                        179 non-null float64
Matiz                                   179 non-null float64
OD280_OD315 de lso vinos diluidos       179 non-null float64
Prolina                                 179 non-null int64
dtypes: float64(11), int64(4)
memory usage: 21.1 KB
```

Como se observa, el data set no tiene datos faltantes.

Operaciones básicas de datos

Discretizar variables (binning)

Se eligen las variables: Magnesio, Alcalinidad Cenizas e Intensidad Color pues, como se observó en el punto anterior, presentan mayor nivel de outliers. A continuación se describe cada uno:

```
In [94]: # Discretice 3 variables usando le método cut. Explique porqué discretizó esta
s variables.
df["Magnesio_3bin"] = pd.cut(df.loc[:, "Magnesio"], 3, labels=["good", "medium",
"bad"])
df.groupby("Magnesio_3bin").size()
```

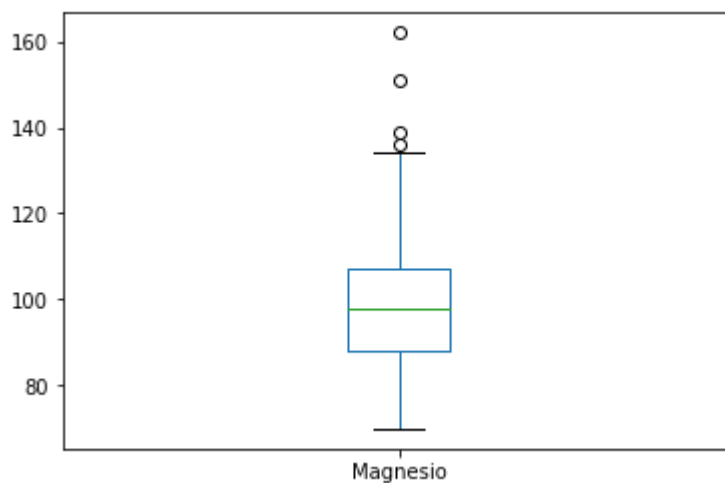
```
Out[94]: Magnesio_3bin
good      101
medium     72
bad         6
dtype: int64
```

```
In [95]: df["Magnesio_cuantiles"] = pd.qcut(df.loc[:, "Magnesio"], [0, .25, .5, .75, 1],
labels=["min", "cu1", "cu2", "cu3"])
df.groupby("Magnesio_cuantiles").size()
```

```
Out[95]: Magnesio_cuantiles
min       47
cu1       50
cu2       38
cu3       44
dtype: int64
```

```
In [74]: df.loc[:, ["Magnesio"]].plot.box()
```

```
Out[74]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35cf8ec50>
```



```
In [75]: df["Alcalinidad_3bin"] = pd.cut(df.loc[:, "Alcalinidad Cenizas"], 3, labels=["g
ood", "medium", "bad"])
df.groupby("Alcalinidad_3bin").size()
```

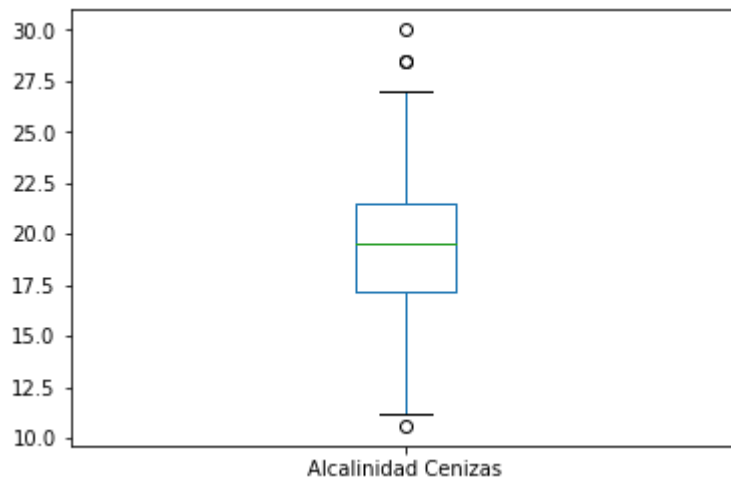
```
Out[75]: Alcalinidad_3bin
good       43
medium    114
bad        22
dtype: int64
```

```
In [76]: df["Alcalinidad_cuantiles"] = pd.qcut(df.loc[:, "Alcalinidad Cenizas"], [0, .25, .5, .75, 1], labels=["min", "cu1", "cu2", "cu3"])
df.groupby("Alcalinidad_cuantiles").size()
```

```
Out[76]: Alcalinidad_cuantiles
min      47
cu1      48
cu2      42
cu3      42
dtype: int64
```

```
In [77]: df.loc[:, ["Alcalinidad Cenizas"]].plot.box()
```

```
Out[77]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35ce3d390>
```



```
In [78]: df["Color_3bin"] = pd.cut(df.loc[:, "Intensidad Color"], 3, labels=["good", "medium", "bad"])
df.groupby("Color_3bin").size()
```

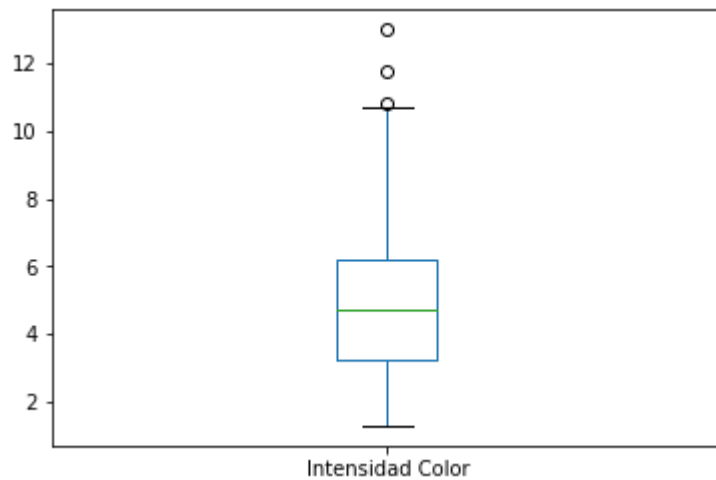
```
Out[78]: Color_3bin
good      103
medium     62
bad       14
dtype: int64
```

```
In [79]: df["Color_cuantiles"] = pd.qcut(df.loc[:, "Intensidad Color"], [0, .25, .5, .75, 1], labels=["min", "cu1", "cu2", "cu3"])
df.groupby("Color_cuantiles").size()
```

```
Out[79]: Color_cuantiles
min      45
cu1      45
cu2      44
cu3      45
dtype: int64
```

```
In [80]: df.loc[:,["Intensidad Color"]].plot.box()
```

```
Out[80]: <matplotlib.axes._subplots.AxesSubplot at 0x2b35ccf3630>
```



Conclusión

Las variables seleccionadas muestran un ruido mayor por fuera de los cuartiles superiores.

Contar los missing

Puede con las funciones `isna` y `notna` identificar y contar los registros nulos en una fila. Si no aplica la función suma es decir solo con `pd.isna(df2['one'])` enlistara los vacios con del dataset para la columna indicada

```
In [81]: for column in df.columns:
          print("El número de vacíos en la columna "+column+" es: " + str(sum(pd.isna(df[column]))))
```

```
El número de vacíos en la columna ID Caso es: 0
El número de vacíos en la columna Clase Vino es: 0
El número de vacíos en la columna Alcohol es: 0
El número de vacíos en la columna Acido Malico es: 0
El número de vacíos en la columna Cenizas es: 0
El número de vacíos en la columna Alcalinidad Cenizas es: 0
El número de vacíos en la columna Magnesio es: 0
El número de vacíos en la columna Fenoles Totales es: 0
El número de vacíos en la columna Flavonoides es: 0
El número de vacíos en la columna Fenoles No Flavonoides es: 0
El número de vacíos en la columna Protoantocianinas es: 0
El número de vacíos en la columna Intensidad Color es: 0
El número de vacíos en la columna Matiz es: 0
El número de vacíos en la columna OD280_OD315 de los vinos diluidos es: 0
El número de vacíos en la columna Prolina es: 0
El número de vacíos en la columna Magnesio_3bin es: 0
El número de vacíos en la columna Magnesio_cuantiles es: 0
El número de vacíos en la columna Alcalinidad_3bin es: 0
El número de vacíos en la columna Alcalinidad_cuantiles es: 0
El número de vacíos en la columna Color_3bin es: 0
El número de vacíos en la columna Color_cuantiles es: 0
```

Conclusión

El data set no tiene valores null

Normalizar una variable

Normalizar una variable continua es redimensionarla en medida de desviaciones estándar conocidas como valores Z por la tabla de puntajes Z de la distribución normal, proceso que se hace con la siguiente fórmula

```
In [82]: df["Alcohol_normalized"] =(df["Alcohol"] - df["Alcohol"].mean())/df["Alcohol"]
          .std()
          df.loc[:,["Alcohol", "Alcohol_normalized"]].head(5)
```

Out[82]:

	Alcohol	Alcohol_normalized
0	14.23	1.502672
1	13.20	0.237210
2	13.16	0.188066
3	14.37	1.674676
4	13.24	0.286354

Identificar valores atípicos

```
In [98]: df["Alcohol_outlier"] = np.where((df["Alcohol"] > (df["Alcohol"].mean() + (2*  
df["Alcohol"].std())) |  
                                           (df["Alcohol"] > (df["Alcohol"].mean() -  
(2*df["Alcohol"].std()))), 'yes', 'no')
```

```
In [99]: df[df["Alcohol_outlier"] == "yes"].head()
```

Out[99]:

	ID Caso	Clase Vino	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides	Fenole Flavanc
0	1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	
1	2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	
2	3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	
3	4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	
4	5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	

```
In [85]: print((df["Alcohol"].mean() - (2*df["Alcohol"].std()))
```

11.379063191082375

Ahora que ya se han valores atípicos, escriba una rutina para suavizar/corregir dichos valores, en una nueva columna, llamada Alcohol_sav.

Para suavizar los valores atípicos reemplazamos estos valores invirtiendo los datos de la encuación que los detecto

```
In [105]: df["Alcohol_sav"] = np.where((df["Alcohol"] > (df["Alcohol"].mean() + (2*df[  
"Alcohol"].std())) |  
                                           (df["Alcohol"] > (df["Alcohol"].mean() -  
(2*df["Alcohol"].std()))),  
                                           ((df["Alcohol"].mean() - (2*df["Alcohol"].st  
d()))),  
                                           ((df["Alcohol"].mean() + (2*df["Alcohol"].std  
()))))
```

```
In [106]: df.head()
```

```
Out[106]:
```

	ID Caso	Clase Vino	Alcohol	Acido Malico	Cenizas	Alcalinidad Cenizas	Magnesio	Fenoles Totales	Flavanoides	Fenole Flavanc
0	1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	
1	2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	
2	3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	
3	4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	
4	5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	

Parte final

Teniendo en cuenta los gráficos generados y las medidas tomadas al dataset Clasificación.csv , podría lanzar alguna hipótesis inicial que discrimine las distintas clase de vino?

Cada clase de vino tiene una preparación diferente. La mezcla de sus ingredientes hace que cada vino posea un sabor y una coloración diferente. De su mezcla, resaltan el magnesio y la Intensidad del Color como principales diferenciadores entre clases, seguidos por el la Alcalinidad de Cenizas, y las cenizas, cuyos promedios de concentración varían por cada mezcla. Por el contrario, el nivel de alcohol en las tres clases de vinos es casi similar.