Actividad Evaluable

Mapas de calor y boxplots

Equipo 6

- A01245418 Andrés Sarellano Acevedo

- A00829837 Axel Amós Hernández Cárdenas

- A01281371 Izel María Ávila Rodríguez

- A01383422 Melissa Elvia Salazar Carrillo

- A00573134 Macías Romero Jorge Humberto

Carga de Datos

En esta sección cargamos el archivo que vamos a tomar como nuestra base de datos.

In [ ]:

```
import pandas as pd
df = pd.read_csv('wine-clustering.csv')
df
```

Out[ ]:

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins
0	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29
1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28
2	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81
3	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18
4	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82
...	...	...	...	...	...	...	...	...	...
173	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06
174	13.40	3.91	2.48	23.0	102	1.80	0.75	0.43	1.41
175	13.27	4.28	2.26	20.0	120	1.59	0.69	0.43	1.35
176	13.17	2.59	2.37	20.0	120	1.65	0.68	0.53	1.46
177	14.13	4.10	2.74	24.5	96	2.05	0.76	0.56	1.35

178 rows × 13 columns

Análisis de Variables

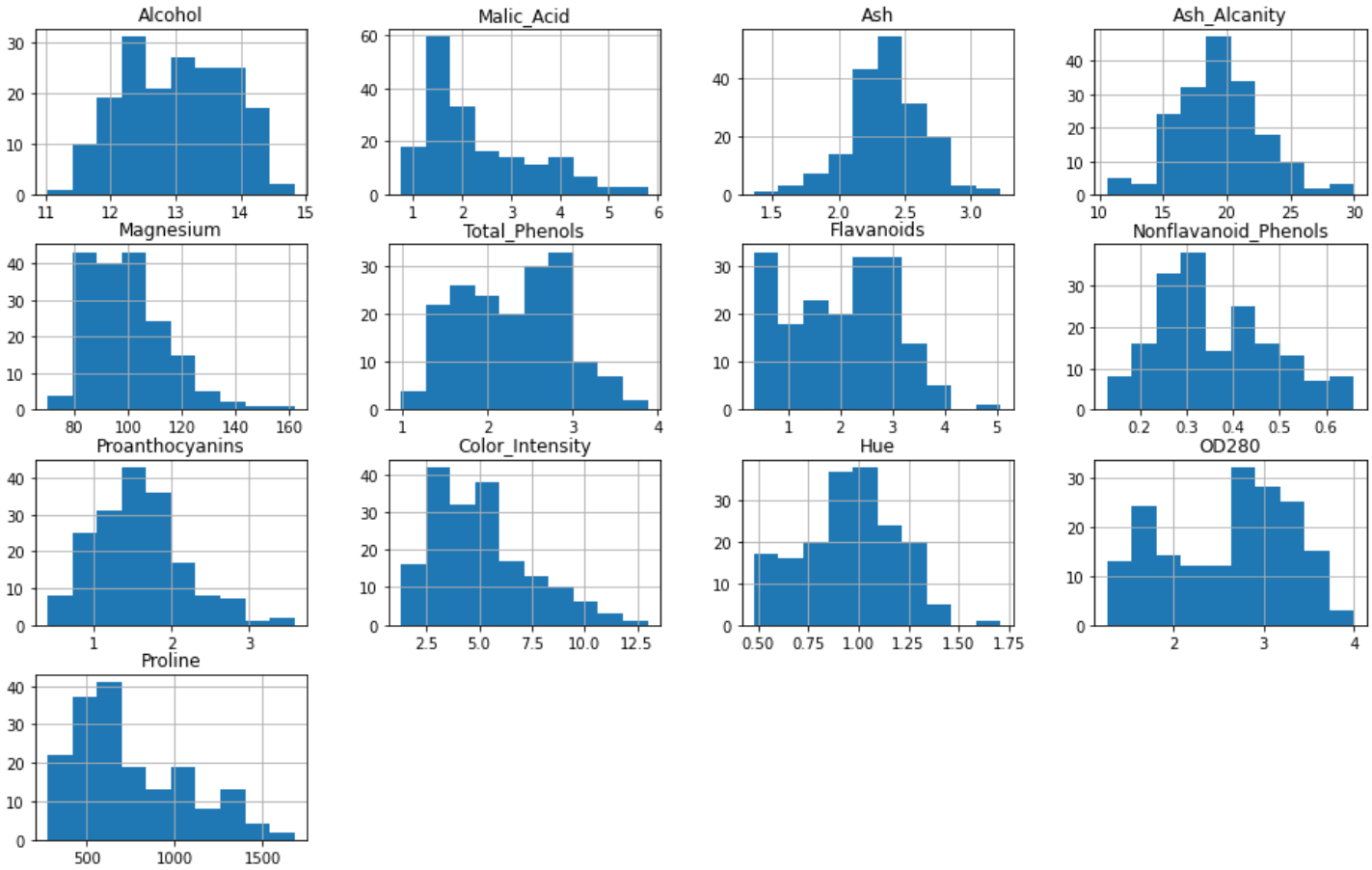
Histograma

In [12]:

```
df.hist(figsize=(16,10))
```

Out[12]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f070097c8d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f07009ea110>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f070093c810>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06fff6ccd0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffe8abd0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffdb5210>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffd6a890>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffd21dd0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffd21e10>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffce6550>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffc4cb90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffc19690>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffbcfc90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffb922d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffbc98d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f06ffb7fed0>]],
      dtype=object)
```



Correlaciones

Se obtiene una matriz de correlaciones

In [13]:

```
Correl = df.corr()  # Se obtiene el coeficiente de Correlación de Pearson
Correl
```

Out[13]:

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phe
Alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.15
Malic_Acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.29
Ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.18

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phe
Ash_Alcanity	0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.36
Magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.25
Total_Phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.44
Flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.53
Nonflavanoid_Phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.00
Proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.36
Color_Intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.13
Hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.26
OD280	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.50
Proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.31

Mapa de calor

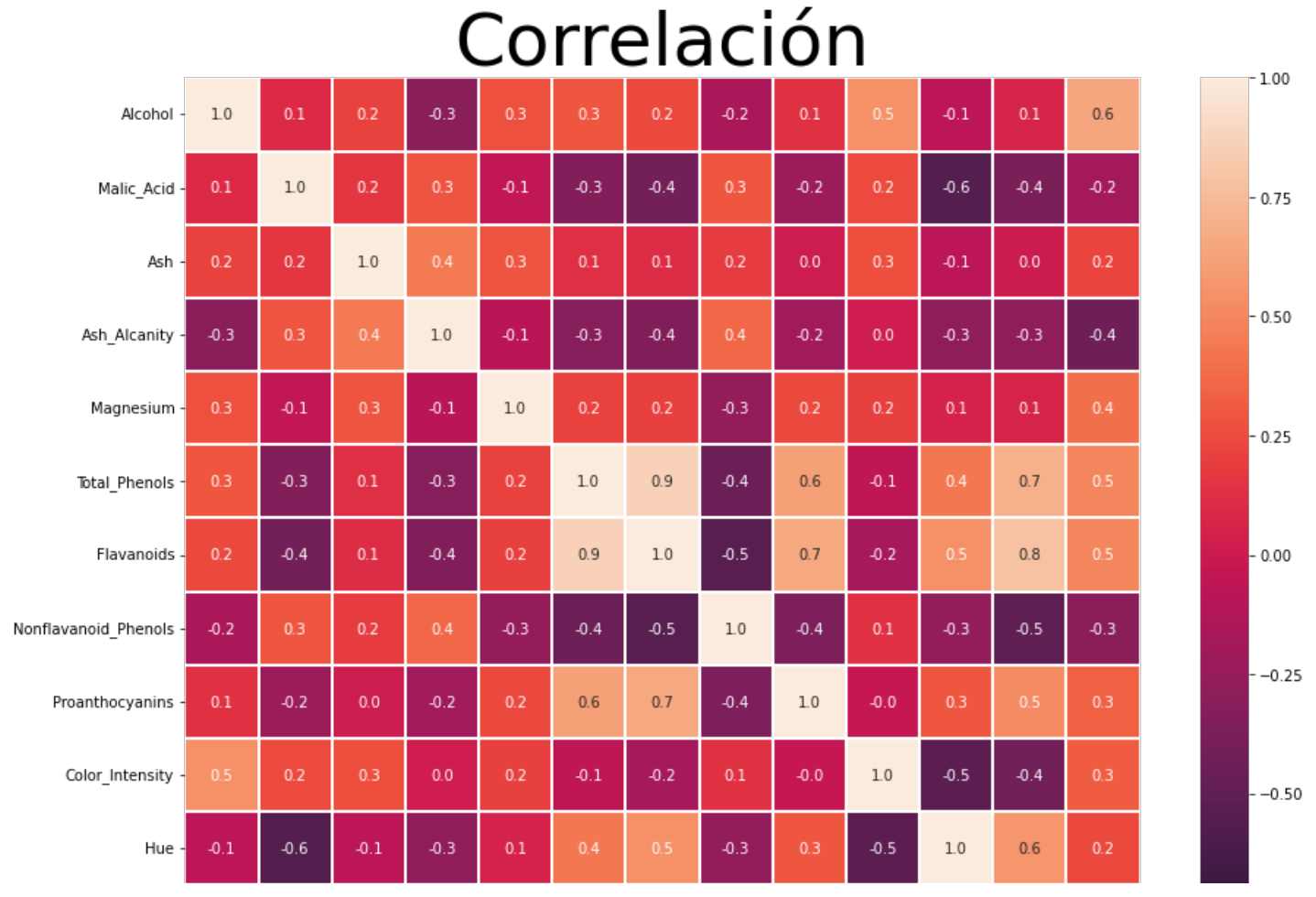
A continuación se muestra el mapa de calor utilizando los datos presentados anteriormente

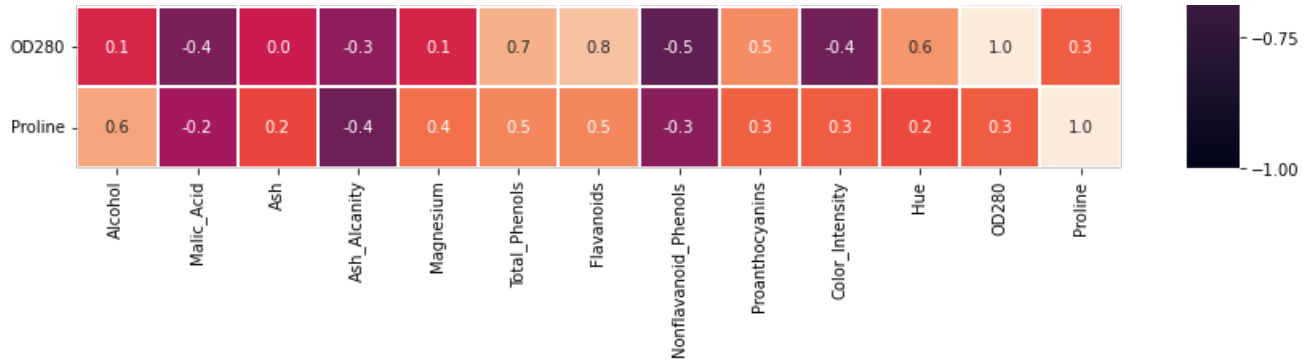
In [14]:

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(15,12))
sns.heatmap(Correl, square = True, annot = True, fmt = '.1f', linewidths= 1, vmin=-1, vm
ax=1).set_title('Correlación', fontsize =50)
#https://seaborn.pydata.org/generated/seaborn.heatmap.html
#https://seaborn.pydata.org/tutorial/color_palettes.html
```

Out[14]:

```
Text(0.5, 1.0, 'Correlación')
```





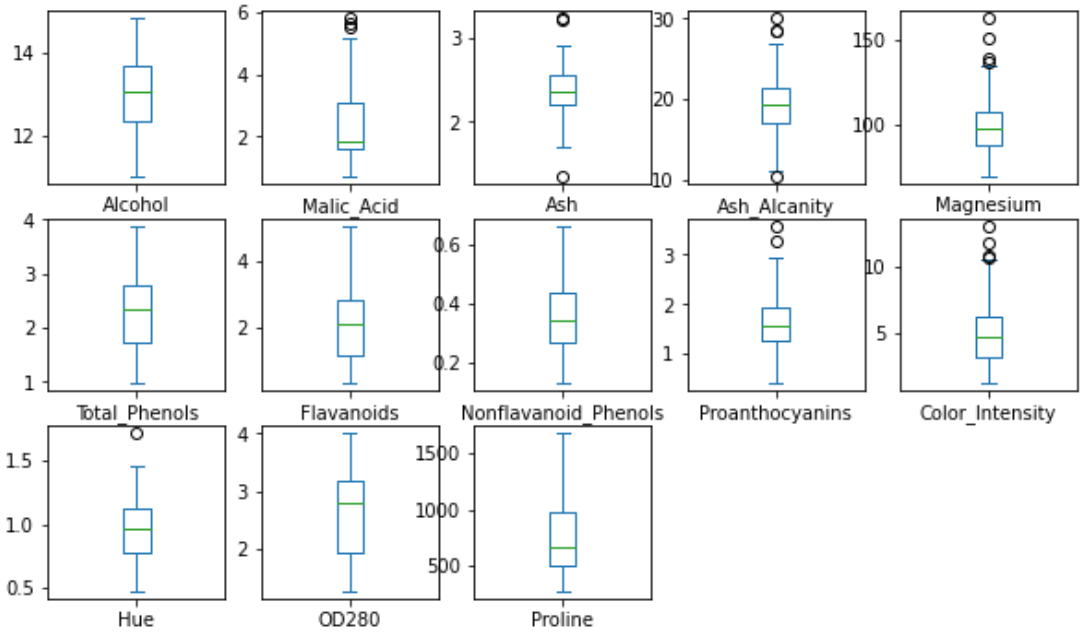
BoxPlot

In [15]:

```
df.plot(subplots=True,layout=(5,5),kind="box",figsize=(10,10),sharex=False)
```

Out[15]:

```
Alcohol                 AxesSubplot(0.125,0.749828;0.133621x0.130172)
Malic_Acid              AxesSubplot(0.285345,0.749828;0.133621x0.130172)
Ash                     AxesSubplot(0.44569,0.749828;0.133621x0.130172)
Ash_Alcanity            AxesSubplot(0.606034,0.749828;0.133621x0.130172)
Magnesium               AxesSubplot(0.766379,0.749828;0.133621x0.130172)
Total_Phenols           AxesSubplot(0.125,0.593621;0.133621x0.130172)
Flavanoids              AxesSubplot(0.285345,0.593621;0.133621x0.130172)
Nonflavanoid_Phenols    AxesSubplot(0.44569,0.593621;0.133621x0.130172)
Proanthocyanins         AxesSubplot(0.606034,0.593621;0.133621x0.130172)
Color_Intensity         AxesSubplot(0.766379,0.593621;0.133621x0.130172)
Hue                     AxesSubplot(0.125,0.437414;0.133621x0.130172)
OD280                   AxesSubplot(0.285345,0.437414;0.133621x0.130172)
Proline                 AxesSubplot(0.44569,0.437414;0.133621x0.130172)
dtype: object
```



Redacte al menos 3 afirmaciones de análisis utilizando la información obtenida por el boxplot.

- **Los flavonoides, OD280 y los fenoles totales tienen una relación de cuartiles muy similar, así como ninguno de estos cuenta con valores atípicos. Haciendo que tengan una relación de balance entre la cantidad de estas sustancias en el vino.**
- **El ácido málico, el magnesio, la intensidad de color y la alcanidad de ceniza cuentan con una cantidad elevada de valores atípicos, mientras que sus los cuartiles del 25% al 75% esten más cercanos unos de los otros representando sus valores reducidos o más específicos; a excepción de el ácido málico. O bien entré más chico sea el rango de valores de los cuartiles más valores atípicos tendrán.**
- **El alcohol y el total de fenoles cuentan con una alta cantidad de muestras entre sus cuartiles 1 y 3, lo que**

significa que la mayoría de las muestras cuentan con un balance concentrado de estas sustancias en el vino. Esto es importante porque si hay un desbalance o una considerable cantidad de valores atípicos como en el caso del ácido málico puede hacer que hayan muchos vinos ácidos. Lo que significa que la mayoría de las muestras cuentan con un equilibrio concentrado de estás sustancias en el vino.

Preguntas

I. ¿Qué variables no aportan información? Explica por qué dichas variables no son relevantes.

Dado que cada variable es independiente de alguna manera, basándose en el BoxPlot, las que tienen más valores atípicos son las variables Malic_Acid, Magnesium y Color_Intensity, esto estando respaldado por el mismo diagrama al ver que los datos dejan de ser relevantes para el estudio a partir de uno de sus "bigotes".

II. Si tuvieras que eliminar variables, ¿Cuáles quitarías y por qué?

Con base a la pregunta anterior, serían esas 3 sin duda alguna, esto porque son las que más difieren con el resto en cuanto a los valores proporcionados, aspirando a un estudio mejor con la ausencia de las mismas, esto apoyado con las desviaciones, Histogramas y con el mismo BoxPlot.

III. ¿Existen variables que tengan datos extraños o atípicos?

Sí, existen variables que presentan datos atípicos los cuales son sencillos de identificar utilizando el gráfico de boxplot. A continuación se enlistan dichas variables:

- **Malic_Acid: Cuenta con datos atípicos después del extremo máximo del boxplot.**
- **Ash: Cuenta con datos atípicos en ambos extremos del boxplot.**
- **Ash_Alcanity: Cuenta con datos atípicos en ambos extremos del boxplot.**
- **Magnesium: Cuenta con datos atípicos después del extremo máximo del boxplot.**
- **Proanthocyanins: Cuenta con datos atípicos después del extremo máximo del boxplot.**
- **Color_Intensity: Cuenta con datos atípicos después del extremo máximo del boxplot.**
- **Hue: Cuenta con datos atípicos antes del extremo mínimo del boxplot.**

IV. Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?

Algunas de las variables sí se encuentran en rangos similares al momento de comparar el diagrama boxplot en cuanto a los ejes delimitados. De cierta manera, esto puede llegar a afectar, debido a que pueden existir ciertas variables que cuentan con una gran diferencia de valores entre sí, lo cual puede modificar el resultado de un análisis estadístico.

V. ¿Puedes encontrar variables que se parezcan? ¿Cuáles son estas?

De acuerdo con el boxplot, las variables que se parecen son los siguientes: Flavanoids, ,OD280 y el total de los Total_Phenols ya que se son muy parecidos en sus cuartiles, su máximo y mínimo. La mediana es más parecida entre el Total_Phenols y los Flavanoids; en cuanto al bigote se parecen más el OD280 y el Total_Phenols .

VI. Describe con argumentos las razones por las que consideras que el análisis estadístico que has realizado puede usarse para tomar decisiones mejores o mejorar una problemática / producto con el uso de los datos asignados.

Consideramos que el análisis estadístico que hemos realizado puede usarse para tomar mejores decisiones en cuanto al vino ya que se puede observar de una mejor manera la correlación que existe entre diferentes componentes y así observar qué componentes son más importantes que otros y en cuales se puede invertir más. Con ayuda de herramientas tales como el box plot que nos permite identificar valores atípicos y comparar distribuciones, el análisis es más sencillo.