Seatwork 11.1 Exploratory Data Analysis for Machine Learning

Submitted by: Dela Cruz, Eugene D.G.

Submitted to: Engr. Roman Richard

Import/download libraries

```
!pip install hvplot
```

```
Collecting hvplot
  Downloading hvplot-0.9.2-py2.py3-none-any.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 23.5 MB/s eta 0:00:00
Requirement already satisfied: bokeh>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.3.4)
Requirement already satisfied: colorcet>=2 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.1.0)
Requirement already satisfied: holoviews>=1.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.17.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.0.3)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.25.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from hvplot) (24.0)
Requirement already satisfied: panel>=0.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.3.8)
Requirement already satisfied: param<3.0,>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.1.0)
Requirement already satisfied: Jinja2>=2.9 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (3.1.3)
Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (1.2.1)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (6.0.1)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (6.3.3)
Requirement already satisfied: xyzservices>=2021.09.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (2024.4.0)
Requirement already satisfied: pyviz-comms>=0.7.4 in /usr/local/lib/python3.10/dist-packages (from holoviews>=1.11.0->hvplot) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2024.1)
Requirement already satisfied: markdown in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (3.6)
Requirement already satisfied: markdown-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (3.0.0)
Requirement already satisfied: linkify-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (2.0.3)
Requirement already satisfied: mdit-py-plugins in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (0.4.0)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (2.31.0)
Requirement already satisfied: tqdm>=4.48.0 in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.66.2)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (6.1.0)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.11.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=2.9->bokeh>=1.0.0->hvplot) (2.1.5)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas->hvplot) (1.16.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->panel>=0.11.0->hvplot) (0.5.1)
Requirement already satisfied: uc-micro-py in /usr/local/lib/python3.10/dist-packages (from linkify-it-py->panel>=0.11.0->hvplot) (1.0.3)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-packages (from markdown-it-py->panel>=0.11.0->hvplot) (0.1.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (2024.2.2)
Installing collected packages: hvplot
Successfully installed hvplot-0.9.2
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns
import hvplot.pandas
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
from scipy import stats
from sklearn.linear_model import LinearRegression
%matplotlib inline
```

import dataset

```
pip install ucimlrepo
```

```
Collecting ucimlrepo
  Downloading ucimlrepo-0.0.6-py3-none-any.whl (8.0 kB)
Installing collected packages: ucimlrepo
Successfully installed ucimlrepo-0.0.6
```

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
automobile = fetch_ucirepo(id=10)

# data (as pandas dataframes)
X = automobile.data.features
y = automobile.data.targets

# metadata
print(automobile.metadata)

# variable information
print(automobile.variables)
```

```
{'uci_id': 10, 'name': 'Automobile', 'repository_url': 'https://archive.ics.uci.edu/dataset/10/automobile', 'data_url': 'https://archive.ics.uci.edu/static/public/10/data.csv', 'abstra
                   name     role         type demographic  \
0                 price  Feature   Continuous        None
1           highway-mpg  Feature   Continuous        None
2              city-mpg  Feature   Continuous        None
3              peak-rpm  Feature   Continuous        None
4            horsepower  Feature   Continuous        None
5     compression-ratio  Feature   Continuous        None
6                stroke  Feature   Continuous        None
7                  bore  Feature   Continuous        None
8           fuel-system  Feature  Categorical        None
9           engine-size  Feature   Continuous        None
10    num-of-cylinders  Feature      Integer        None
11         engine-type  Feature  Categorical        None
12         curb-weight  Feature   Continuous        None
13              height  Feature   Continuous        None
14               width  Feature   Continuous        None
15              length  Feature   Continuous        None
16          wheel-base  Feature   Continuous        None
17     engine-location  Feature       Binary        None
18        drive-wheels  Feature  Categorical        None
19          body-style  Feature  Categorical        None
20        num-of-doors  Feature      Integer        None
21          aspiration  Feature       Binary        None
22           fuel-type  Feature       Binary        None
23                make  Feature  Categorical        None
24  normalized-losses  Feature   Continuous        None
25           symboling   Target      Integer        None

                                    description units missing_values
0                      continuous from 5118 to 45400  None            yes
1                           continuous from 16 to 54  None             no
2                           continuous from 13 to 49  None             no
3                       continuous from 4150 to 6600  None            yes
4                          continuous from 48 to 288  None            yes
5                            continuous from 7 to 23  None             no
6                       continuous from 2.07 to 4.17  None            yes
7                       continuous from 2.54 to 3.94  None            yes
8        1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi  None             no
9                           continuous from 61 to 326  None             no
```

```
10         eight, five, four, six, three, twelve, two  None            no
11             dohc, dohcv, l, ohc, ohcf, ohcv, rotor  None            no
12                       continuous from 1488 to 4066  None            no
13                        continuous from 47.8 to 59.8  None            no
14                        continuous from 60.3 to 72.3  None            no
15                      continuous from 141.1 to 208.1  None            no
16                          continuous from 86.6 120.9  None            no
17                                       front, rear  None            no
18                                    4wd, fwd, rwd  None            no
19      hardtop, wagon, sedan, hatchback, convertible  None            no
20                                        four, two  None           yes
21                                      std, turbo  None            no
22                                     diesel, gas  None            no
23  alfa-romero, audi, bmw, chevrolet, dodge, hond...  None            no
24                           continuous from 65 to 256  None           yes
25                              -3, -2, -1, 0, 1, 2, 3  None            no
```

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
wine = fetch_ucirepo(id=109)

# data (as pandas dataframes)
Xw = wine.data.features
yw = wine.data.targets

# metadata
print(wine.metadata)

# variable information
print(wine.variables)
```

```
{'uci_id': 109, 'name': 'Wine', 'repository_url': 'https://archive.ics.uci.edu/dataset/109/wine', 'data_url': 'https://archive.ics.uci.edu/static/public/109/data.csv', 'abstract': 'Usi
                        name     role         type demographic  \
0                      class   Target  Categorical        None
1                    Alcohol  Feature   Continuous        None
2                  Malicacid  Feature   Continuous        None
3                        Ash  Feature   Continuous        None
4            Alcalinity_of_ash  Feature   Continuous        None
5                  Magnesium  Feature      Integer        None
6              Total_phenols  Feature   Continuous        None
7                 Flavanoids  Feature   Continuous        None
8          Nonflavanoid_phenols  Feature   Continuous        None
9              Proanthocyanins  Feature   Continuous        None
10              Color_intensity  Feature   Continuous        None
11                          Hue  Feature   Continuous        None
12  0D280_0D315_of_diluted_wines  Feature   Continuous        None
13                      Proline  Feature      Integer        None

   description units missing_values
0         None  None             no
1         None  None             no
2         None  None             no
3         None  None             no
4         None  None             no
5         None  None             no
6         None  None             no
7         None  None             no
8         None  None             no
9         None  None             no
10        None  None             no
11        None  None             no
12        None  None             no
13        None  None             no
```

Data wrangling

```
X.head()
```

	price	highway-mpg	city-mpg	peak-rpm	horsepower	compression-ratio	stroke	bore	fuel-system	engine-size	...	length	wheel-base	engine locatio
0	13495.0	27	21	5000.0	111.0	9.0	2.68	3.47	mpfi	130	...	168.8	88.6	fror
1	16500.0	27	21	5000.0	111.0	9.0	2.68	3.47	mpfi	130	...	168.8	88.6	fror
2	16500.0	26	19	5000.0	154.0	9.0	3.47	2.68	mpfi	152	...	171.2	94.5	fror
3	13950.0	30	24	5500.0	102.0	10.0	3.40	3.19	mpfi	109	...	176.6	99.8	fror
4	17450.0	22	18	5500.0	115.0	8.0	3.40	3.19	mpfi	136	...	176.6	99.4	fror

```
y.head()
```

	symboling
0	3
1	3
2	1
3	2
4	2

Next steps: View recommended plots

```
Xw.head()
```

	Alcohol	Malicacid	Ash	Alcalinity_of_ash	Magnesium	Total_phenols	Flavanoids	Nonflavanoid_phenols	Proanthocyani
0	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.
1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.
2	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.
3	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.
4	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.

Next steps: View recommended plots

```
yw.head()
```

	class
0	1
1	1
2	1
3	1
4	1

Next steps: View recommended plots

```
atmb = pd.concat([X,y], axis = 1) #atmb for automobile
atmb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   price              201 non-null    float64
 1   highway-mpg        205 non-null    int64  
 2   city-mpg           205 non-null    int64  
 3   peak-rpm           203 non-null    float64
 4   horsepower         203 non-null    float64
 5   compression-ratio  205 non-null    float64
 6   stroke             201 non-null    float64
 7   bore               201 non-null    float64
 8   fuel-system        205 non-null    object 
 9   engine-size        205 non-null    int64  
 10  num-of-cylinders   205 non-null    int64  
 11  engine-type        205 non-null    object 
 12  curb-weight        205 non-null    int64  
 13  height             205 non-null    float64
 14  width              205 non-null    float64
 15  length             205 non-null    float64
 16  wheel-base         205 non-null    float64
 17  engine-location    205 non-null    object 
 18  drive-wheels       205 non-null    object 
 19  body-style         205 non-null    object 
 20  num-of-doors       203 non-null    float64
 21  aspiration         205 non-null    object 
 22  fuel-type          205 non-null    object 
 23  make               205 non-null    object 
 24  normalized-losses  164 non-null    float64
 25  symboling          205 non-null    int64  
dtypes: float64(12), int64(6), object(8)
memory usage: 41.8+ KB
```

```
wine = pd.concat([Xw, yw], axis = 1)
wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Alcohol                       178 non-null    float64
 1   Malicacid                     178 non-null    float64
 2   Ash                           178 non-null    float64
 3   Alcalinity_of_ash             178 non-null    float64
 4   Magnesium                     178 non-null    int64  
 5   Total_phenols                 178 non-null    float64
 6   Flavanoids                    178 non-null    float64
 7   Nonflavanoid_phenols          178 non-null    float64
 8   Proanthocyanins               178 non-null    float64
 9   Color_intensity               178 non-null    float64
 10  Hue                           178 non-null    float64
 11  0D280_0D315_of_diluted_wines  178 non-null    float64
 12  Proline                       178 non-null    int64  
 13  class                         178 non-null    int64  
dtypes: float64(11), int64(3)
memory usage: 19.6 KB
```

```
atmb.head()
```

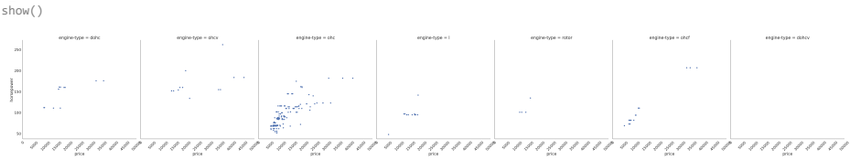
	price	highway-mpg	city-mpg	peak-rpm	horsepower	compression-ratio	stroke	bore	fuel-system	engine-size	...	wheel-base	engine-location	drive wheel
0	13495.0	27	21	5000.0	111.0	9.0	2.68	3.47	mpfi	130	...	88.6	front	rw
1	16500.0	27	21	5000.0	111.0	9.0	2.68	3.47	mpfi	130	...	88.6	front	rw
2	16500.0	26	19	5000.0	154.0	9.0	3.47	2.68	mpfi	152	...	94.5	front	rw
3	13950.0	30	24	5500.0	102.0	10.0	3.40	3.19	mpfi	109	...	99.8	front	fw
4	17450.0	22	18	5500.0	115.0	8.0	3.40	3.19	mpfi	136	...	99.4	front	4w

```
atmb.describe(include='all')
```

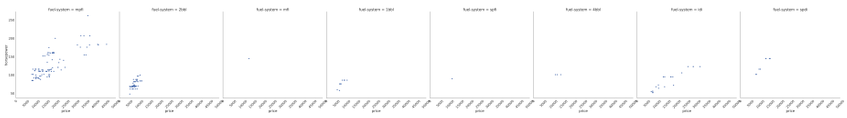
	price	highway-mpg	city-mpg	peak-rpm	horsepower	compression-ratio	stroke	bore	fuel-system	eng
count	201.000000	205.000000	205.000000	203.000000	203.000000	205.000000	201.000000	201.000000	205	205.00
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8	
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	mpfi	
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	94	
mean	13207.129353	30.751220	25.219512	5125.369458	104.256158	10.142537	3.255423	3.329751	NaN	126.90
std	7947.066342	6.886443	6.542142	479.334560	39.714369	3.972040	0.316717	0.273539	NaN	41.64
min	5118.000000	16.000000	13.000000	4150.000000	48.000000	7.000000	2.070000	2.540000	NaN	61.00
25%	7775.000000	25.000000	19.000000	4800.000000	70.000000	8.600000	3.110000	3.150000	NaN	97.00
50%	10295.000000	30.000000	24.000000	5200.000000	95.000000	9.000000	3.290000	3.310000	NaN	120.00
75%	16500.000000	34.000000	30.000000	5500.000000	116.000000	9.400000	3.410000	3.590000	NaN	141.00
max	45400.000000	54.000000	49.000000	6600.000000	288.000000	23.000000	4.170000	3.940000	NaN	326.00

11 rows × 26 columns

```
sns.set(rc={'figure.figsize':(20,10)}, style='white', font_scale=2)

g = sns.FacetGrid(atmb, col='engine-type', height=10)
g = g.map(plt.scatter, "price", "horsepower")
g.set_titles(size=25)
g.set_xticklabels(rotation=45)
plt.show()
```



```
sns.set(rc={'figure.figsize':(20,10)}, style='white', font_scale=2)

g = sns.FacetGrid(atmb, col='fuel-system', height=10)
g = g.map(plt.scatter, "price", "horsepower")
g.set_titles(size=25)
g.set_xticklabels(rotation=45)
plt.show()
```



```
wine.head()
```

	Alcohol	Malicacid	Ash	Alcalinity_of_ash	Magnesium	Total_phenols	Flavanoids	Nonflavanoid_phenols	Proanthocyani
0	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.
1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.
2	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.
3	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.
4	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.

Next steps: View recommended plots

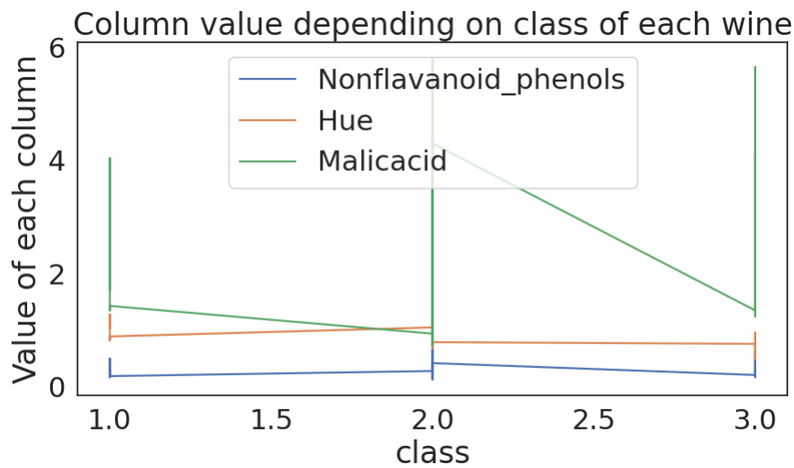
```
wine.describe(include='all')
```

	Alcohol	Malicacid	Ash	Alcalinity_of_ash	Magnesium	Total_phenols	Flavanoids	Nonflavanoid_phenols
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000

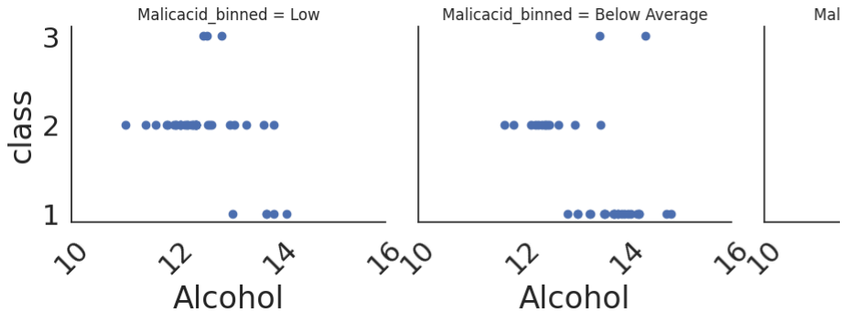
```
wine.tail()
```

	Alcohol	Malicacid	Ash	Alcalinity_of_ash	Magnesium	Total_phenols	Flavanoids	Nonflavanoid_phenols	Proanthocya
173	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	
174	13.40	3.91	2.48	23.0	102	1.80	0.75	0.43	
175	13.27	4.28	2.26	20.0	120	1.59	0.69	0.43	
176	13.17	2.59	2.37	20.0	120	1.65	0.68	0.53	
177	14.13	4.10	2.74	24.5	96	2.05	0.76	0.56	

```
wine.plot(
    kind='line', y=['Nonflavanoid_phenols', 'Hue', 'Malicacid'], x='class',
    title='Column value depending on class of each wine',
    figsize=(10,5)).set_ylabel('Value of each column')
plt.show()
```

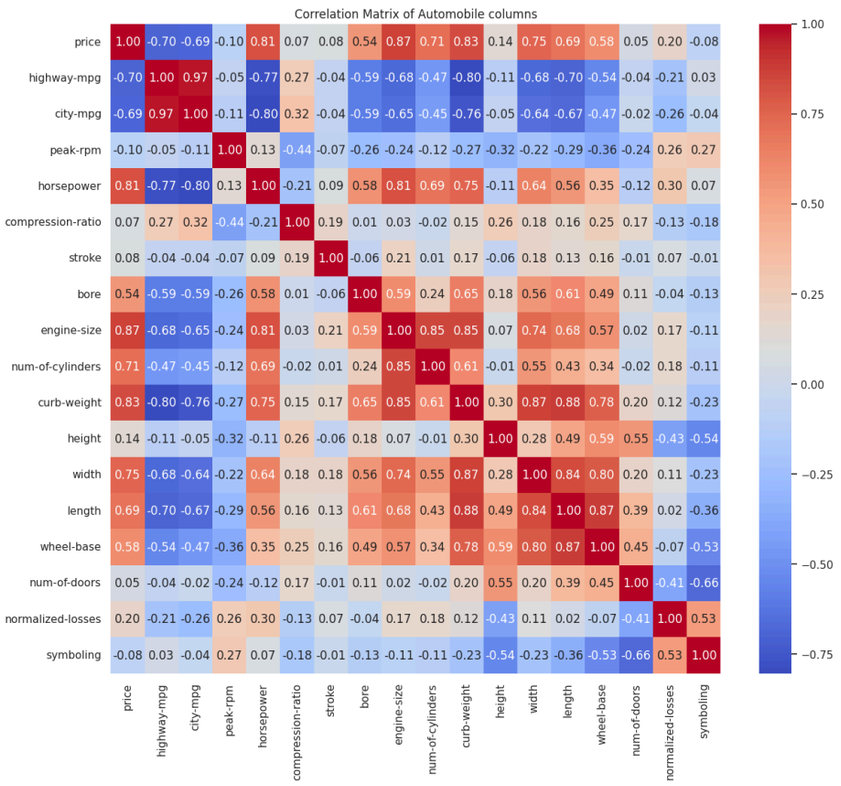



```
# this plot shows the population of Malicacid depending on class and Alcohol
sns.set(rc={'figure.figsize': (20, 10)}, style='white', font_scale=2)
if pd.api.types.is_numeric_dtype(wine['Malicacid']):
    wine['Malicacid_binned'] = pd.qcut(wine['Malicacid'], 5, labels=['Low', 'Below Average', 'Average', 'Above Average', 'High'])
    col_param = 'Malicacid_binned'
else:
    col_param = 'Malicacid'
g = sns.FacetGrid(wine, col=col_param, height=4)
g.map(plt.scatter, 'Alcohol', 'class')
g.set_titles(size=12)
g.set_xticklabels(rotation=45)

plt.show()
```



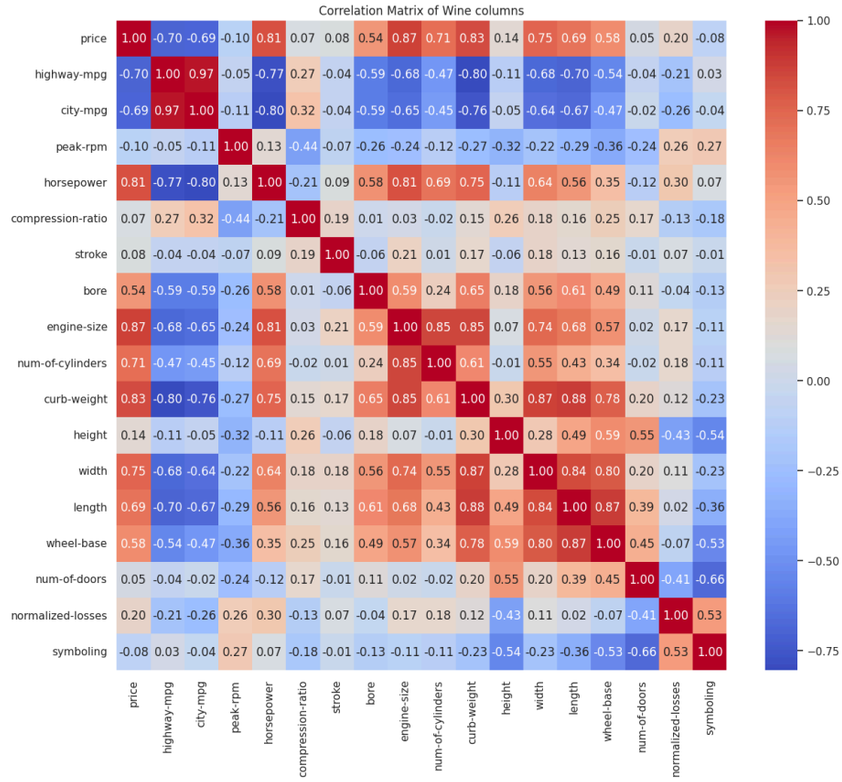
Correlation between the columns of Automobile

```
numeric_val = atmb.select_dtypes(include=[float, int])
correlation_matrix = numeric_val.corr()

sns.set(style='white')
plt.figure(figsize=(14, 12))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix of Automobile columns')
plt.show()
```



Correlation between the columns of Wine

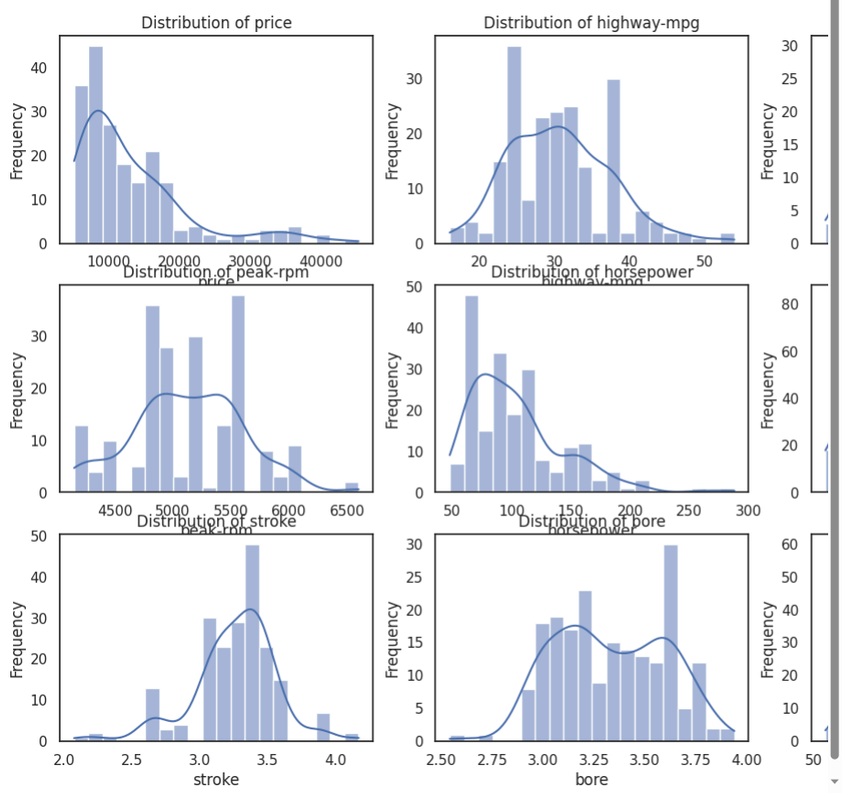
```
correlation_wine = numeric_val.corr()

sns.set(style='white')
plt.figure(figsize=(14, 12))
sns.heatmap(correlation_wine, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix of Wine columns')
plt.show()
```



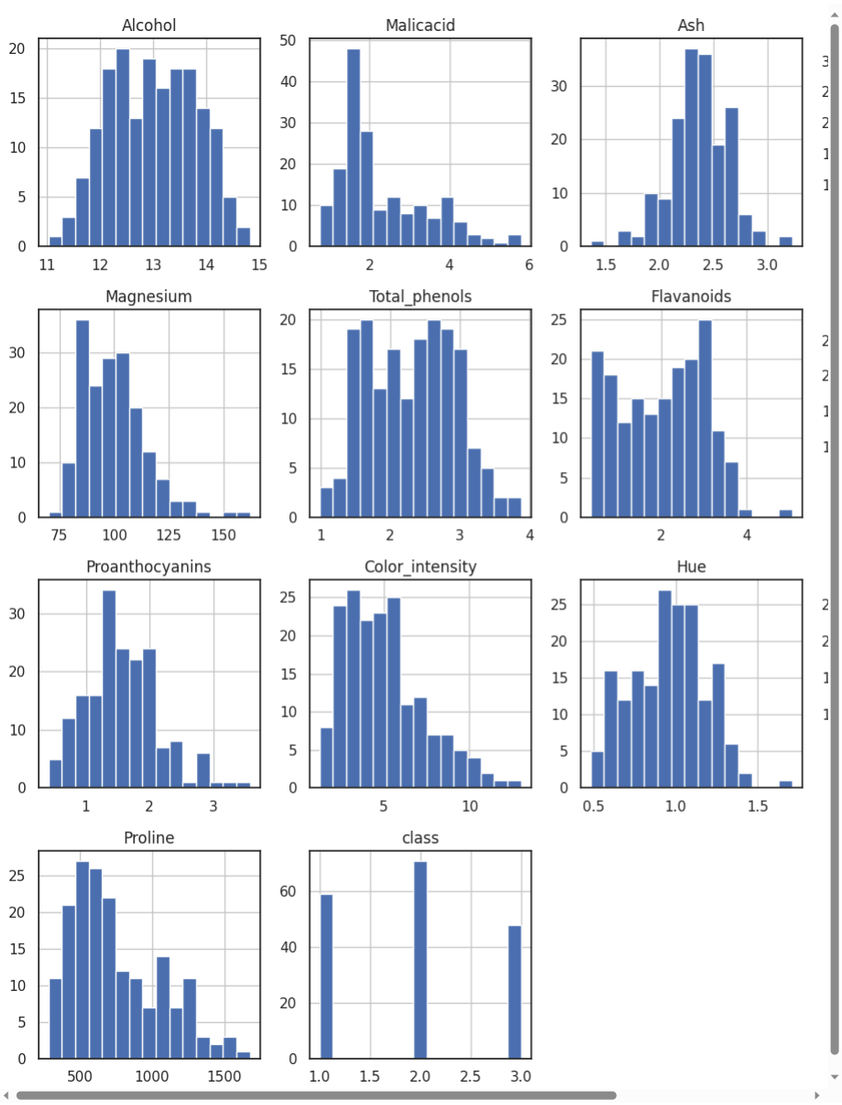
Distribution of each features of automobile

```
plt.figure(figsize=(15, 10))

for i, column in enumerate(numeric_val.columns):
    plt.subplot(3, 3, i + 1)
    sns.histplot(numeric_val[column], kde=True, bins=20)
    plt.title(f'Distribution of {column}')
    plt.xlabel(column)
    plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-92-962808200927> in <cell line: 3>()
      2
      3 for i, column in enumerate(numeric_val.columns):
----> 4     plt.subplot(3, 3, i + 1)
      5     sns.histplot(numeric_val[column], kde=True, bins=20)
      6     plt.title(f'Distribution of {column}')

                                 1 frames
/usr/local/lib/python3.10/dist-packages/matplotlib/gridspec.py in _from_subplot_args(figure, args)
    596         else:
    597             if not isinstance(num, Integral) or num < 1 or num > rows*cols:
--> 598                 raise ValueError(
    599                     f"num must be an integer with 1 <= num <= {rows*cols}, "
    600                     f"not {num!r}"

ValueError: num must be an integer with 1 <= num <= 9, not 10
```



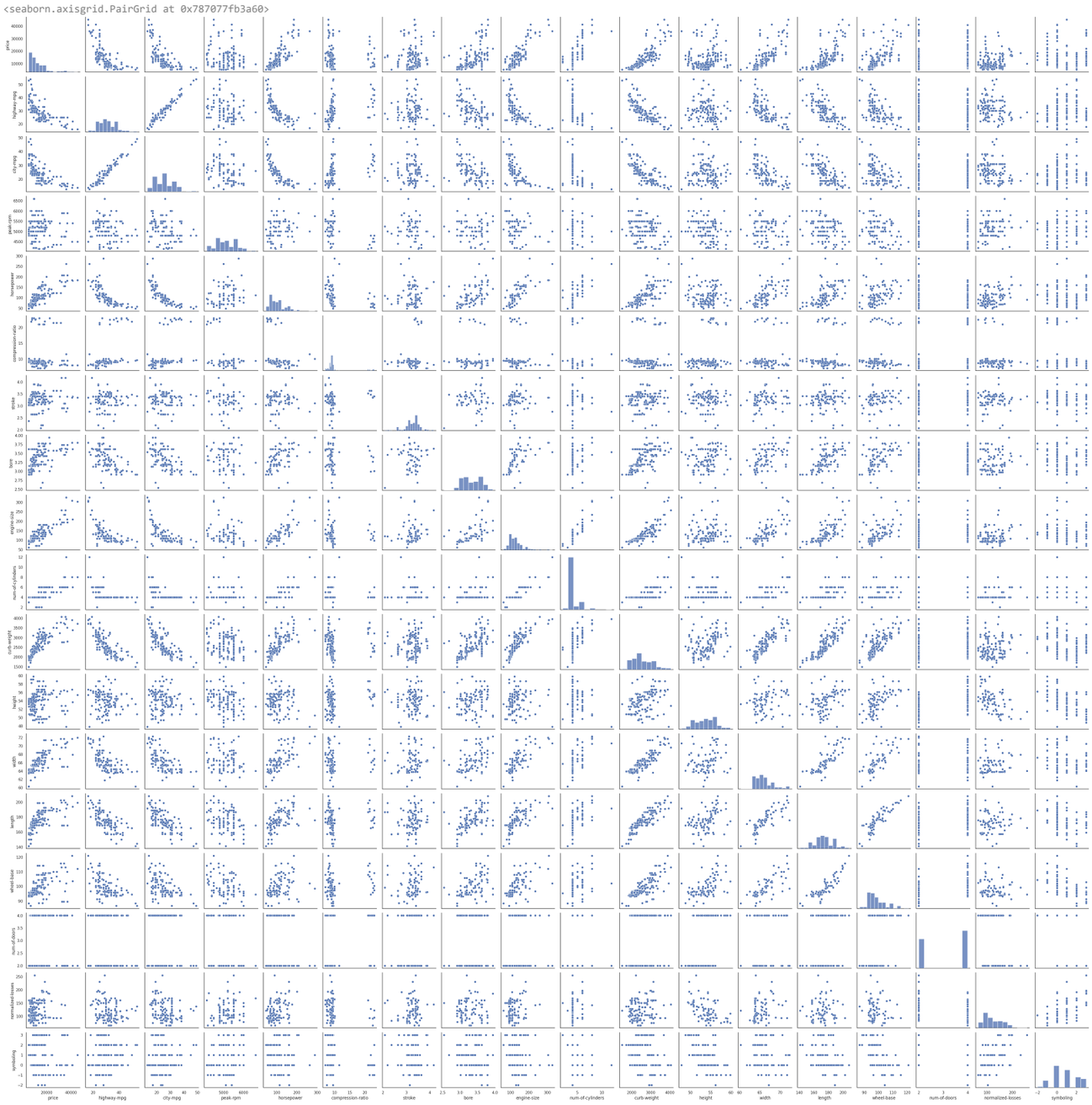

```
wine.hist(bins=15, figsize=(12, 12))
plt.tight_layout()
plt.show()
```



Explore relationships between the features of automobile and wine

```
sns.pairplot(atmb)
```

```
<seaborn.axisgrid.PairGrid at 0x787077fb3a60>
```




```
sns.pairplot(wine)
```

```
<seaborn.axisgrid.PairGrid at 0x78707a5a2830>
```