Kmit

SONET

Data Science

(Level-I)

Ajeet K. Jain, M. Narsimlu

(ML TEAM)- SONET, KMIT, Hyderabad

Session - 24

This session deals with

Data Visualization

Histogram and Scatter plot

seaborn library

Creating various graphs using seaborn

Exercises on Seaborn

Data Preprocessing

Exercise-5

DATA SCIENCE

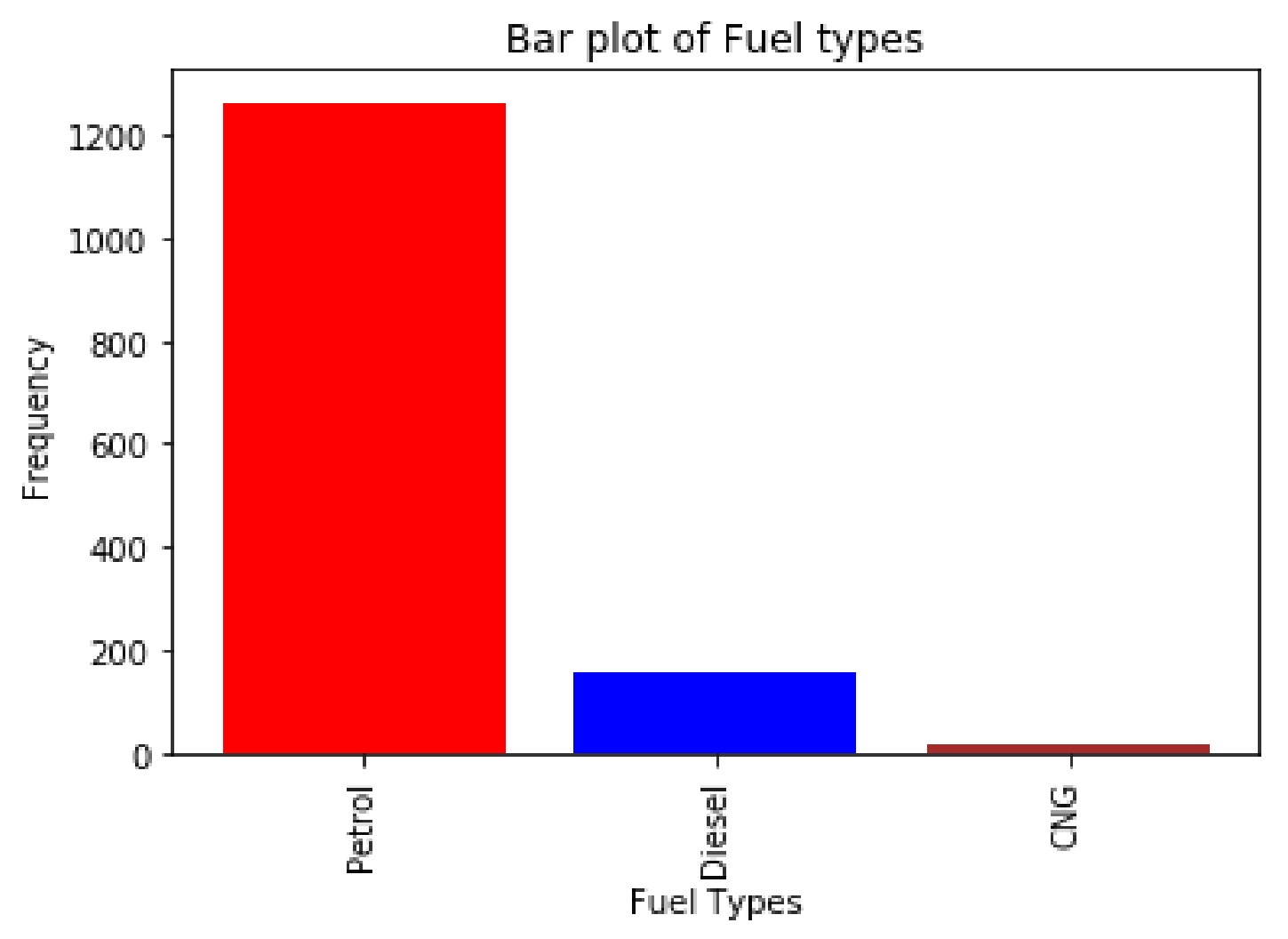
Read a Toyotacars dataset and perform the following tasks

1.Create a one way table on "Fuel_Type" feature
2.create a two list which consists of number of values of categories and categories
3.Find the length of categories list
4.create a bar graph betweeen index and count,add the colors as "red","blue","brown“ to the categories
5.Name of the graph title as "Bar plot of Fuel types
6.Name x label as "Fuel type" and y label as "Frequency"
7.Add sticks to categories as "Petrol","Diesel","CNG" with rotation "90"
8.show the graph

Exercise-5

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
data_cars=pd.read_csv("ToyotaCorolla.csv")
Fuel_tab=pd.crosstab(index=data_cars["Fuel_Type"],columns="counts")
print(Fuel_tab)
counts=[1264,155,17]
fueltype=["Petrol","Diesel","CNG"]
index=np.arange(len(fueltype))
plt.bar(index,counts,color=["red","blue","brown"])
plt.title("Bar plot of Fuel types")
plt.xlabel("Fuel Types")
plt.ylabel("Frequency")
plt.xticks(index,fueltype,rotation=90)
plt.show()
```

Output


Bar plot of Fuel types
Frequency
1200
1000
800
600
400
200
0
Petrol
Diesel
CNG
Fuel Types

Histogram

- Histograms are used to show a distribution whereas a bar chart is used to compare different entities.
- Histograms are useful when you have arrays or a very long list.
- Histograms are graphical representations of a frequency distribution of data.

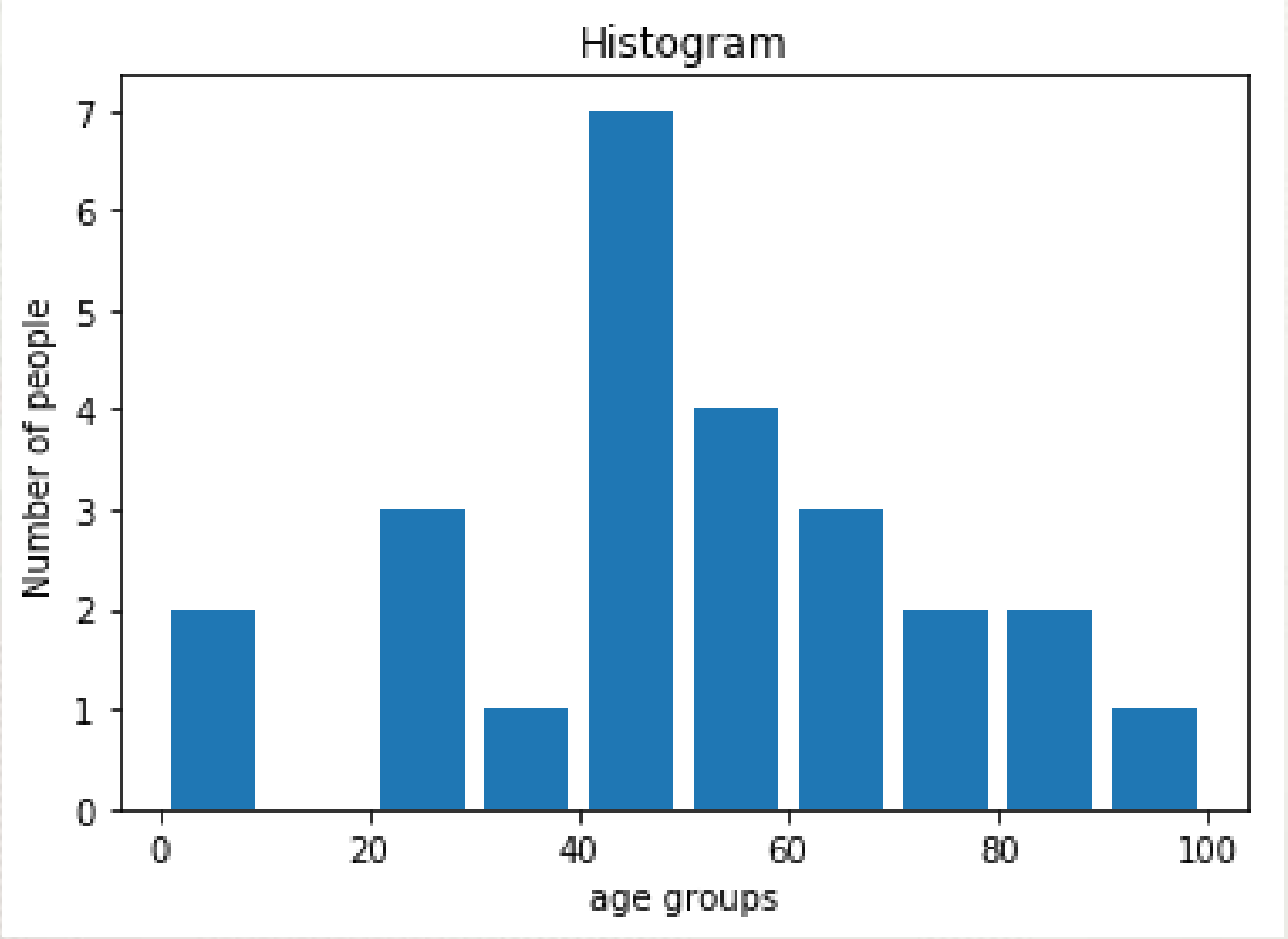
Exercise-5

- Create a histogram on the following data
- population_age =[22,55,62,45,21,22,34,42,42,4,2,102,95,85,55,110,120,
- 70,65,55,111,115,80,75,65,54,44,43,42,48]
- bins = [0,10,20,30,40,50,60,70,80,90,100]

Solution

```
import matplotlib.pyplot as plt
population_age =[22,55,62,45,21,22,34,42,42,4,2,102,95,85,55,110,120,
                 70,65,55,111,115,80,75,65,54,44,43,42,48]
bins = [0,10,20,30,40,50,60,70,80,90,100]
plt.hist(population_age, bins, histtype='bar', rwidth=0.8)
plt.xlabel('age groups')
plt.ylabel('Number of people')
plt.title('Histogram')
plt.show()
```

Output


Histogram
Number of people
0
1
2
3
4
5
6
7
0
20
40
60
80
100
age groups

Exercise-6

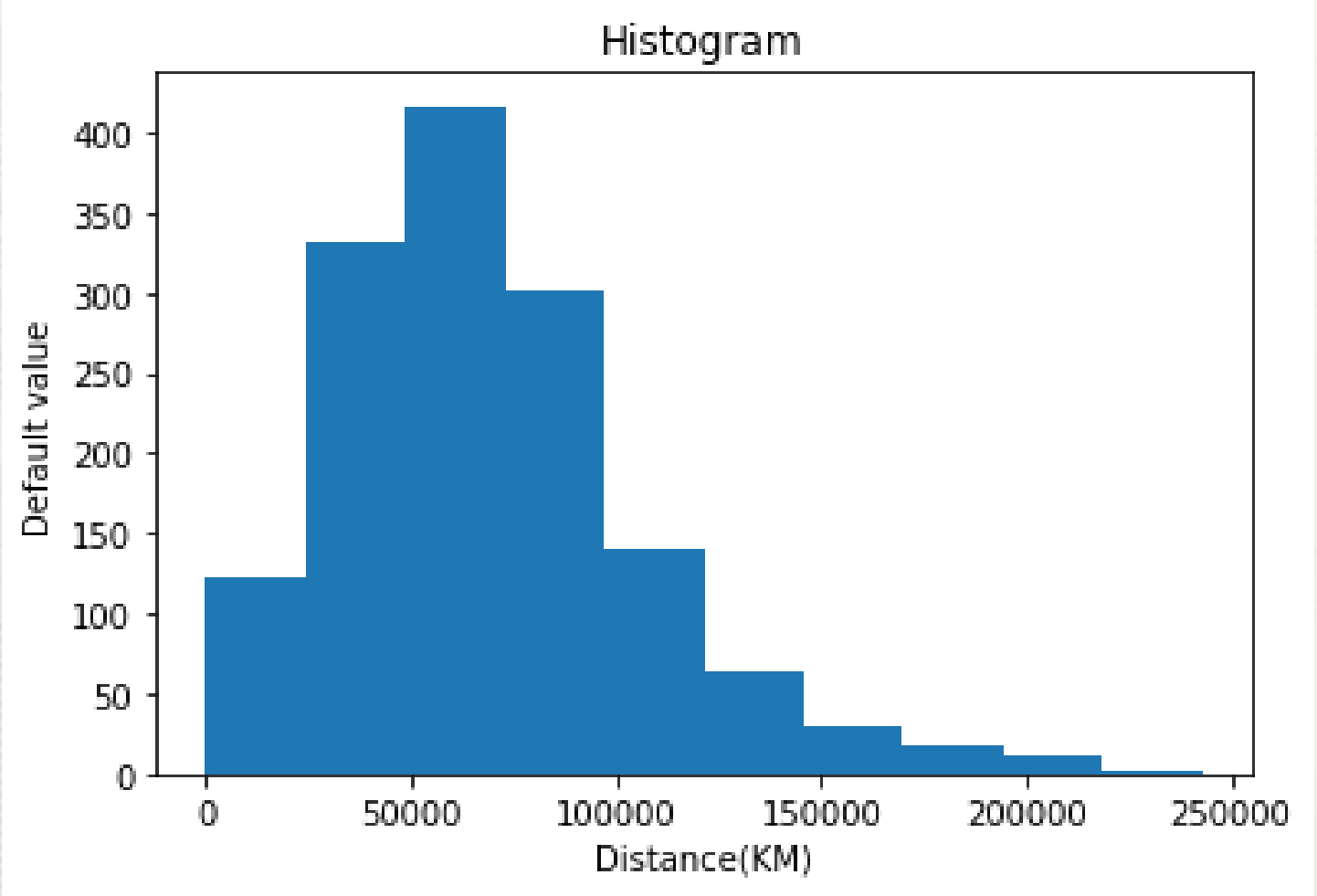
Read a Toyotacars dataset and perform the following tasks

1.Create a histogram with default argument. The car which travelled no of KM.
2.create a histogram with fixed number of bins=“10” and color of the bar and also color of the edge

1.Solution

```
import pandas as pd
import matplotlib.pyplot as plt
data_cars=pd.read_csv("ToyotaCorolla.csv")
plt.hist(data_cars["KM"])
plt.title("Histogram")
plt.xlabel("Distance(KM)")
plt.ylabel("Default value")
plt.show()
```

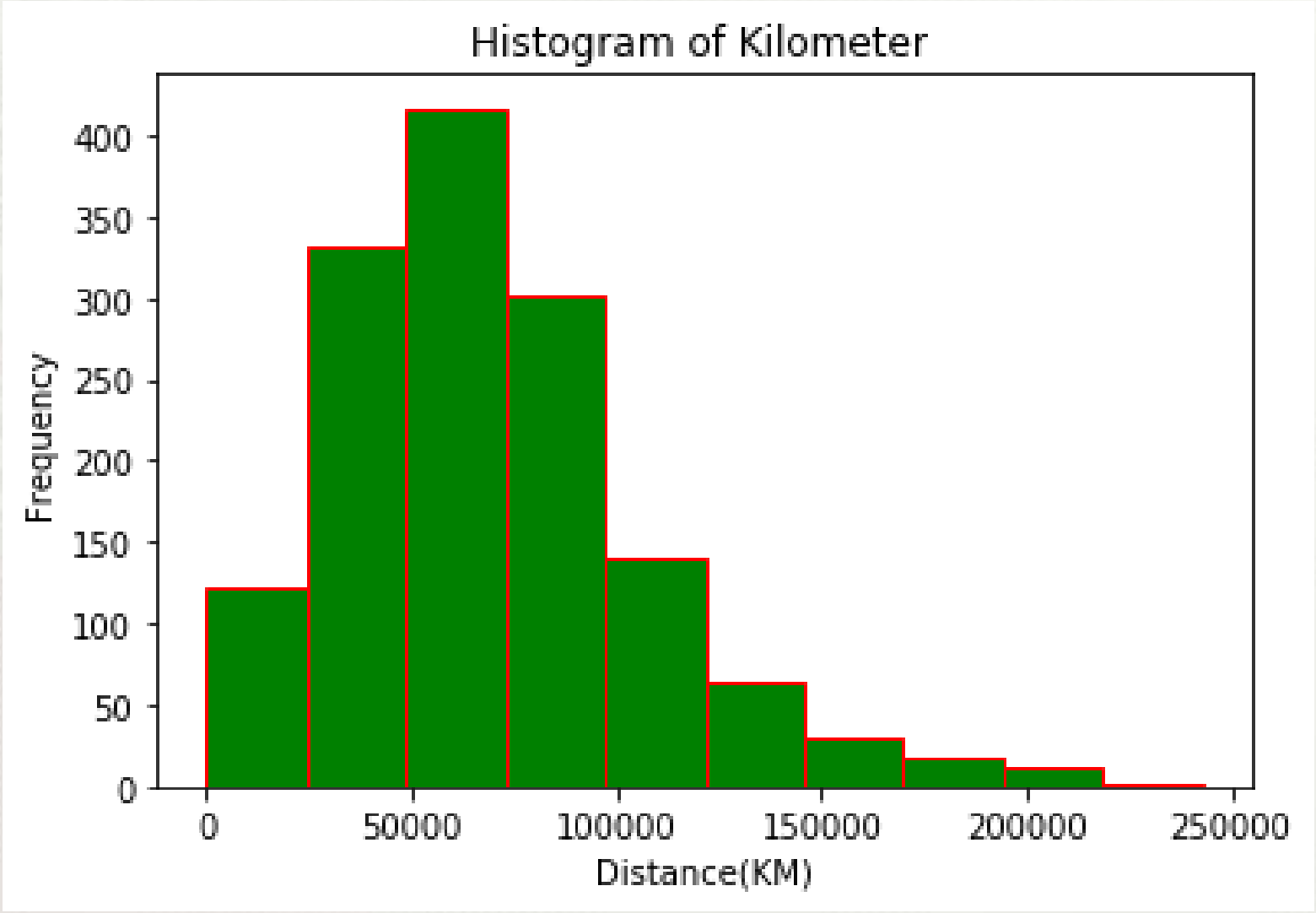
Output


Histogram
Default value
400
350
300
250
200
150
100
50
0
0
50000
100000
150000
200000
250000
Distance(KM)

2.Solution

```
import pandas as pd
import matplotlib.pyplot as plt
data_cars=pd.read_csv("ToyotaCorolla.csv")
plt.hist(data_cars["KM"],bins=10,color="green",edgecolor="red")
plt.title("Histogram of Kilometer ")
plt.xlabel("Distance(KM)")
plt.ylabel("Frequency")
plt.show()
```

Output


Histogram of Kilometer
Frequency
400
350
300
250
200
150
100
50
0
0
50000
100000
150000
200000
250000
Distance(KM)

Scatter Plot

- Usually we need scatter plots in order to compare variables.

- How much one variable is affected by another variable to build a relation out of it.
- The scatter() function makes a scatter plot with (optional) size and color arguments.
- Scatter plots are used to convey the relationship between numerical variables
- The data is displayed as a collection of points, each having the value of one variable which determines the position on the horizontal axis and the value of other variable determines the position on the vertical axis.

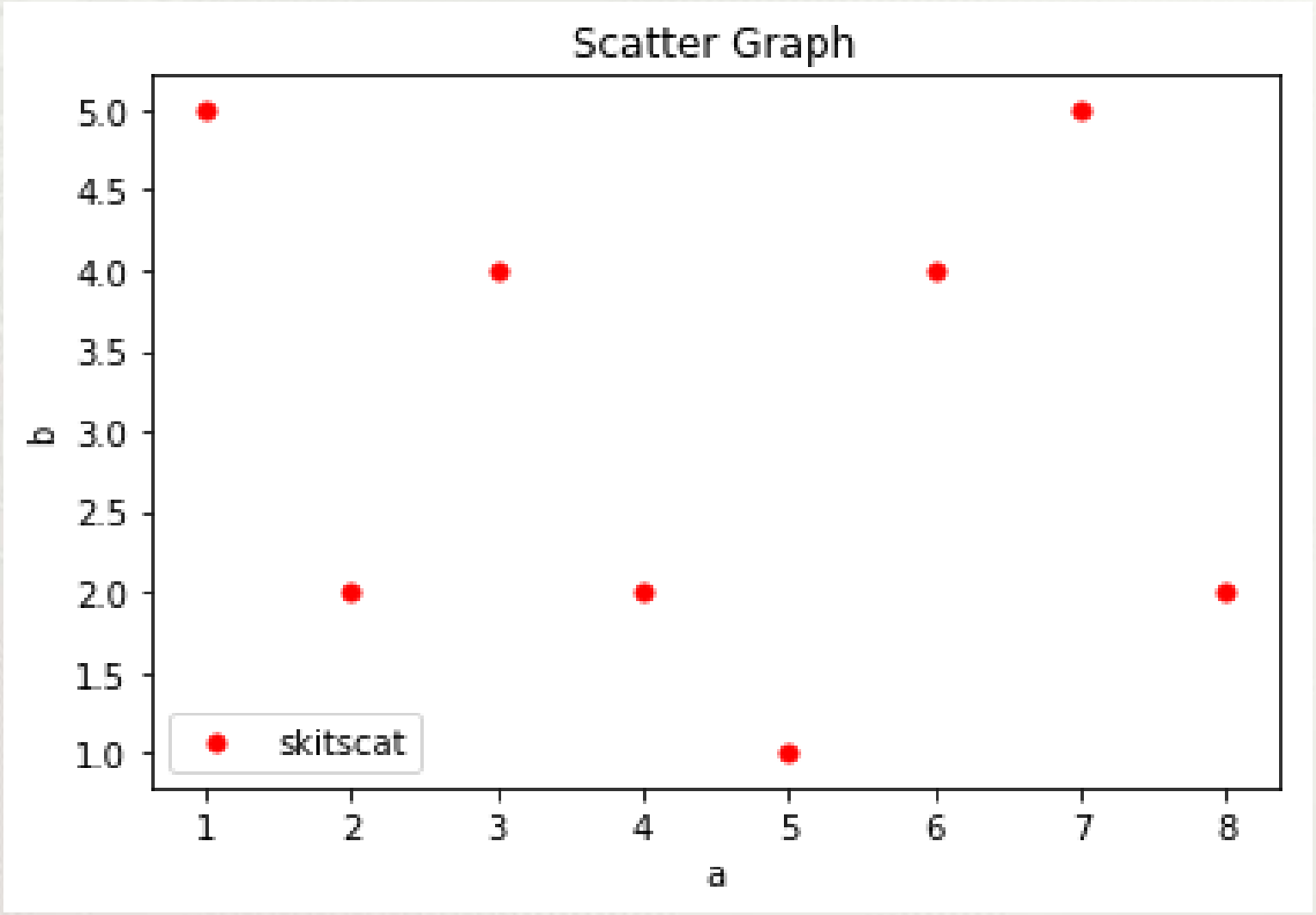
Exercise-6

- Create a scatter graph for the following data:
- a = [1,2,3,4,5,6,7,8]
- b = [5,2,4,2,1,4,5,2]

```
import matplotlib.pyplot as plt
a = [1,2,3,4,5,6,7,8]
b = [5,2,4,2,1,4,5,2]
plt.scatter(a,b, label='skitscat', color='red', s=25, marker="o")
plt.xlabel('a')
plt.ylabel('b')
plt.title("Scatter Graph")
plt.legend()
plt.show()
```

Exercise-6

DATA SCIENCE



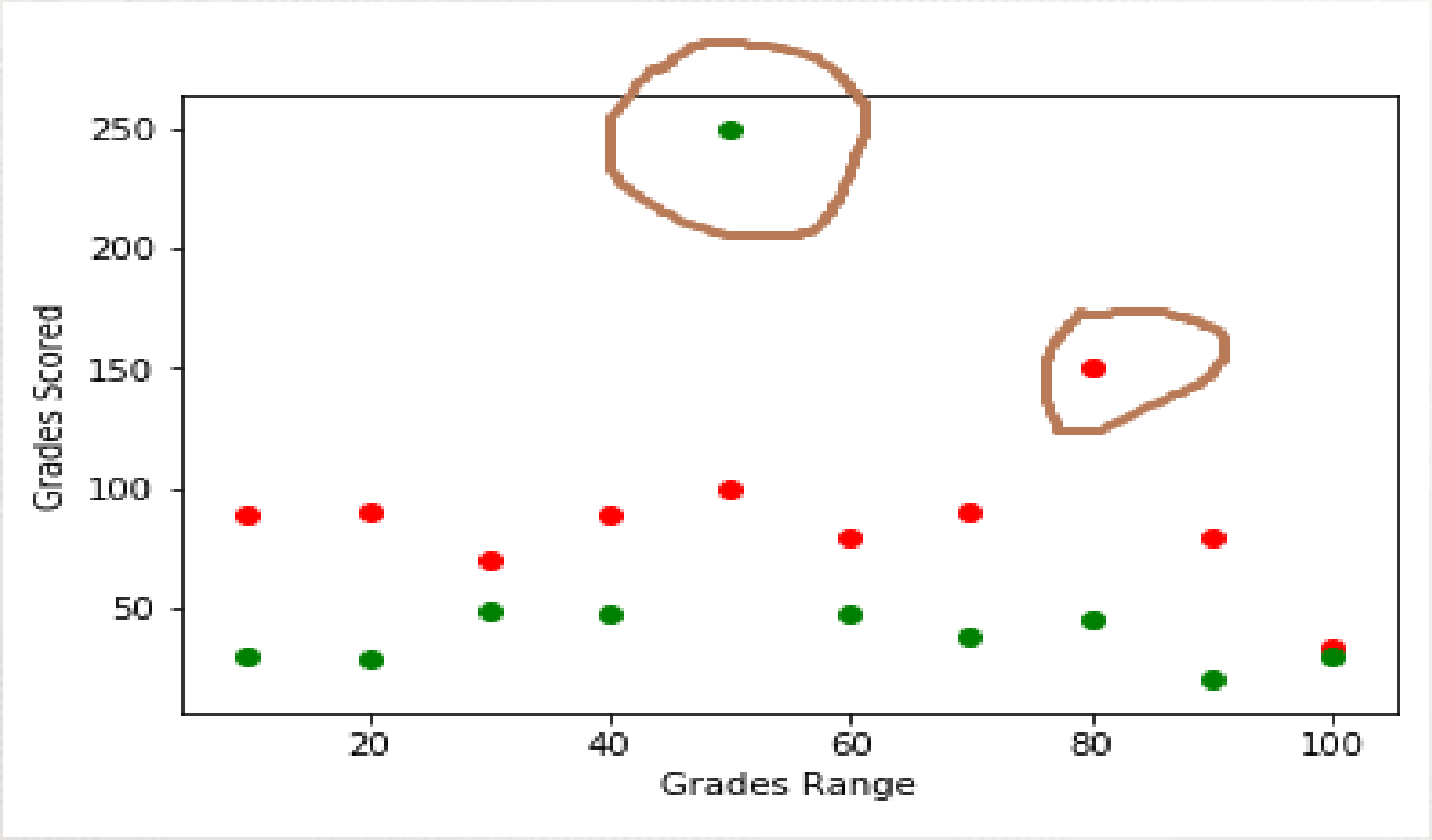
Scatter-plot

- Scatter Plots are usually used to represent the correlation between two or more variables. It also helps to identify **Outliers**, if any.

- Create a scatter plot for the following data
- girls_grades = [89, 90, 70, 89, 100, 80, 90, 150, 80, 34]
- boys_grades = [30, 29, 49, 48, 250, 48, 38, 45, 20, 30]
- grades_range = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
- 1.create a relation between grades ranges and girls grades features
- 2. create a relation between grades ranges and boys grades features

```
import matplotlib.pyplot as plt
import pandas as pd
girls_grades = [89, 90, 70, 89, 100, 80, 90, 150, 80, 34]
boys_grades = [30, 29, 49, 48, 250, 48, 38, 45, 20, 30]
grades_range = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
plt.scatter(grades_range, girls_grades,color='r')
plt.scatter(grades_range, boys_grades,color='g')
plt.xlabel('Grades Range')
plt.ylabel('Grades Scored')
plt.show()
```

Output



Exercise-8

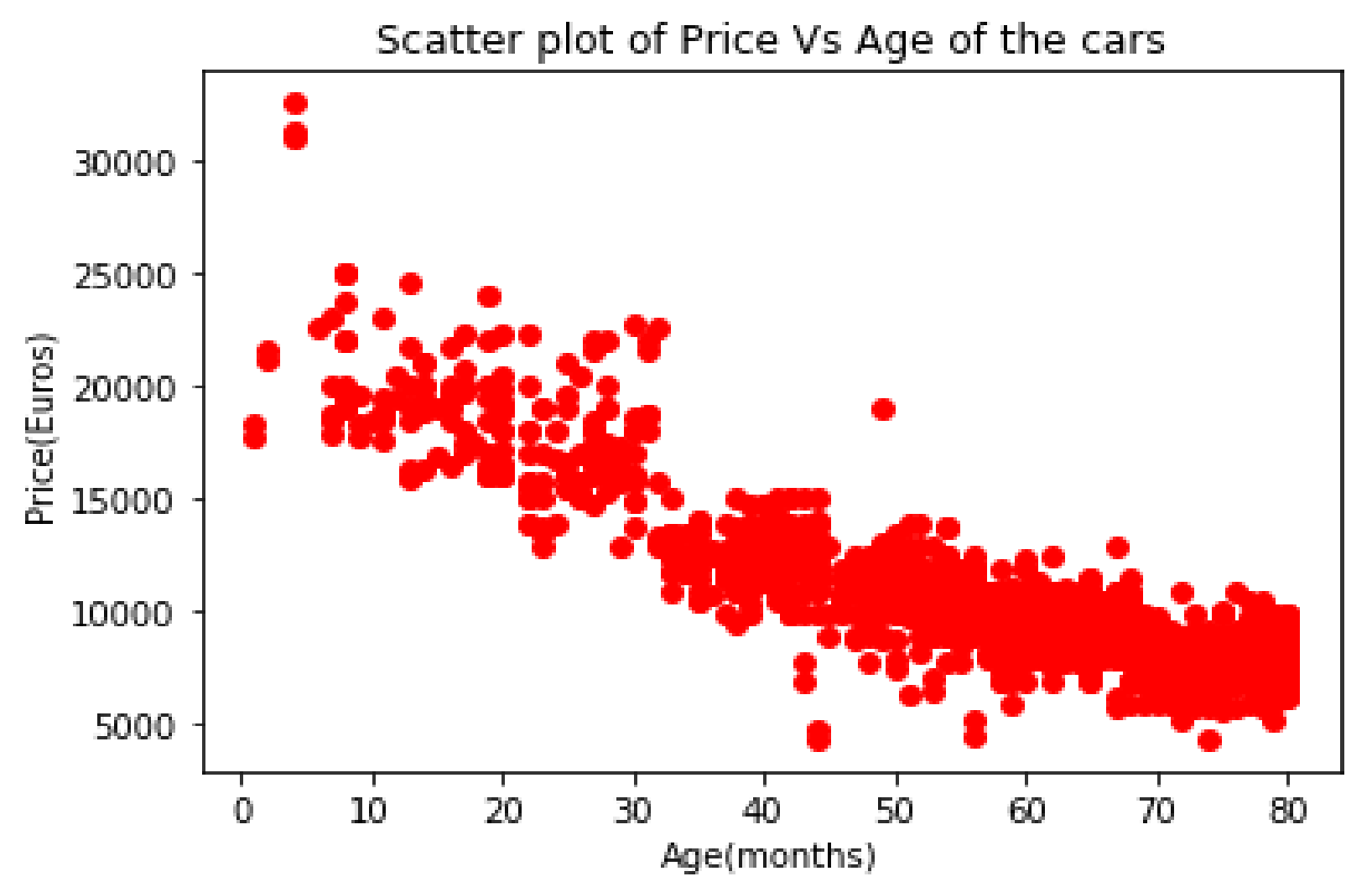
Read a Toyota cars dataset and perform the following tasks

1.Find missing values and display them
2.Remove missing values and display them
3.create a scatter plot between Age vs price

Solution

```
import pandas as pd
import matplotlib.pyplot as plt
data_cars=pd.read_csv("ToyotaCorolla.csv")
print(data_cars.isnull().sum())
print(data_cars)
data_cars.dropna(axis=0,inplace=True)
print(data_cars)
plt.scatter(data_cars["Age_08_04"],data_cars["Price"],c="red")
plt.title("Scatter plot of Price Vs Age of the cars")
plt.xlabel("Age(months)")
plt.ylabel("Price(Euros)")
plt.show()
```

Output


Scatter plot of Price Vs Age of the cars
Price(Euros)
30000
25000
20000
15000
10000
5000
0
10
20
30
40
50
60
70
80
Age(months)

Seaborn

Seaborn is a python visualization library based on matplotlib.

It provides a high-level interface for drawing attractive statistical graphics.

It was originally developed at Stanford University and is widely used for plotting and visualizing data.

Seaborn

There are several advantages:

It possesses built-in themes for better visualizations.

It has tools built in statistical functions which reveal hidden patterns in the data set.

It has functions to visualize matrices of data which become very important when visualizing large data sets.

Seaborn

create basic plots using *seaborn* library:

- Scatter plot
- Histogram
- Bar plot
- Box and whiskers plot
- Pairwise plots

Scatter Plot

- Usually we need scatter plots in order to compare variables.

- How much one variable is affected by another variable to build a relation out of it.
- The scatter() function makes a scatter plot with (optional) size and color arguments.
- Scatter plots are used to convey the relationship between numerical variables
- Scatter Plots are usually used to represent the correlation between two or more variables. It also helps it identify **Outliers**, if any.

Exercise-1

DATA SCIENCE

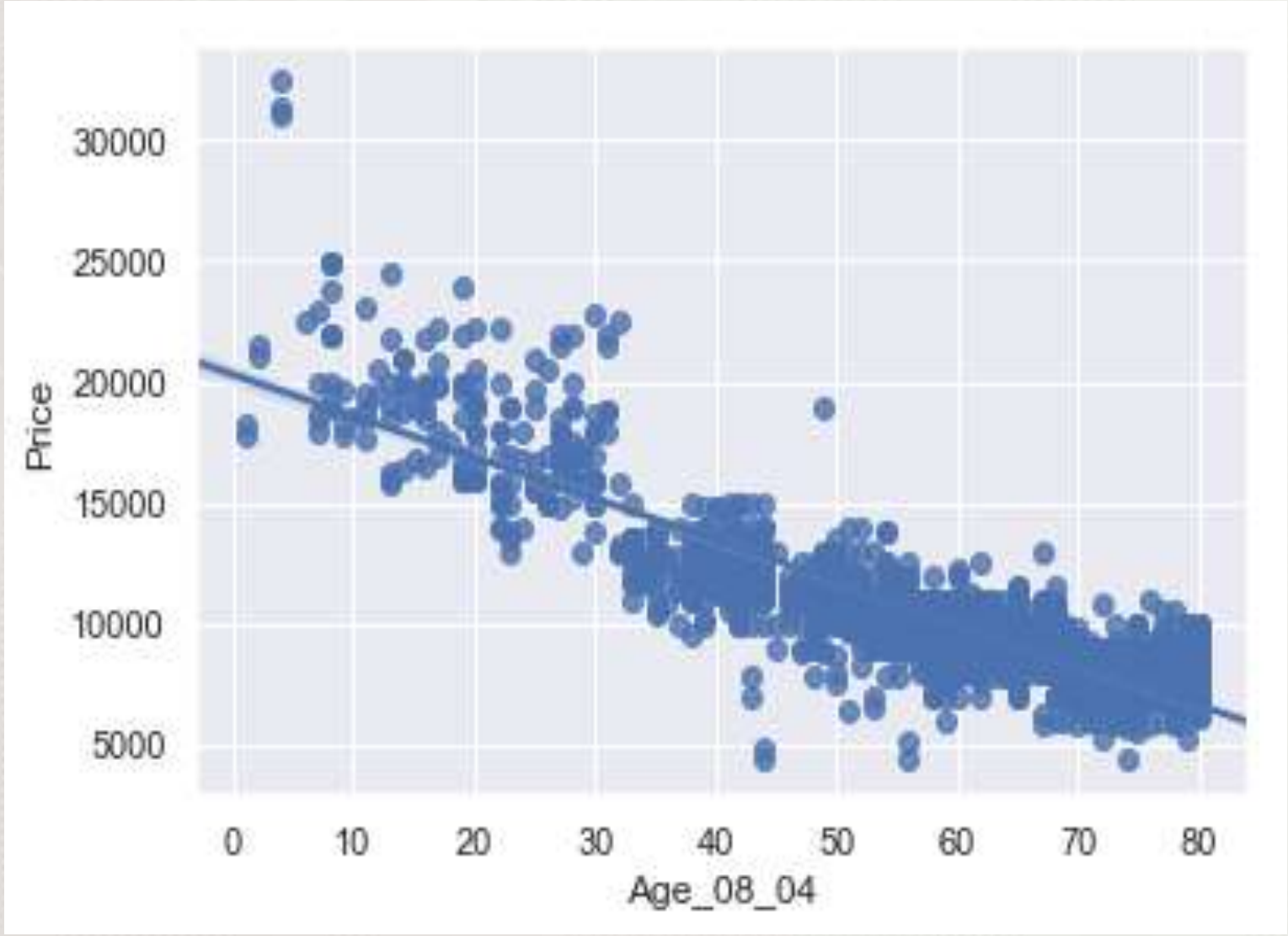
Read a Toyota cars dataset and perform following tasks:
1.create a Scatter plot of Price vs Age with default arguments
Use regplot()-to create a scatter plot(it will plot regression data)

```
import pandas as pd
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.set(style="darkgrid")
sns.regplot(data_cars["Age_08_04"],data_cars["Price"])
```

SONET

Output

DATA SCIENCE



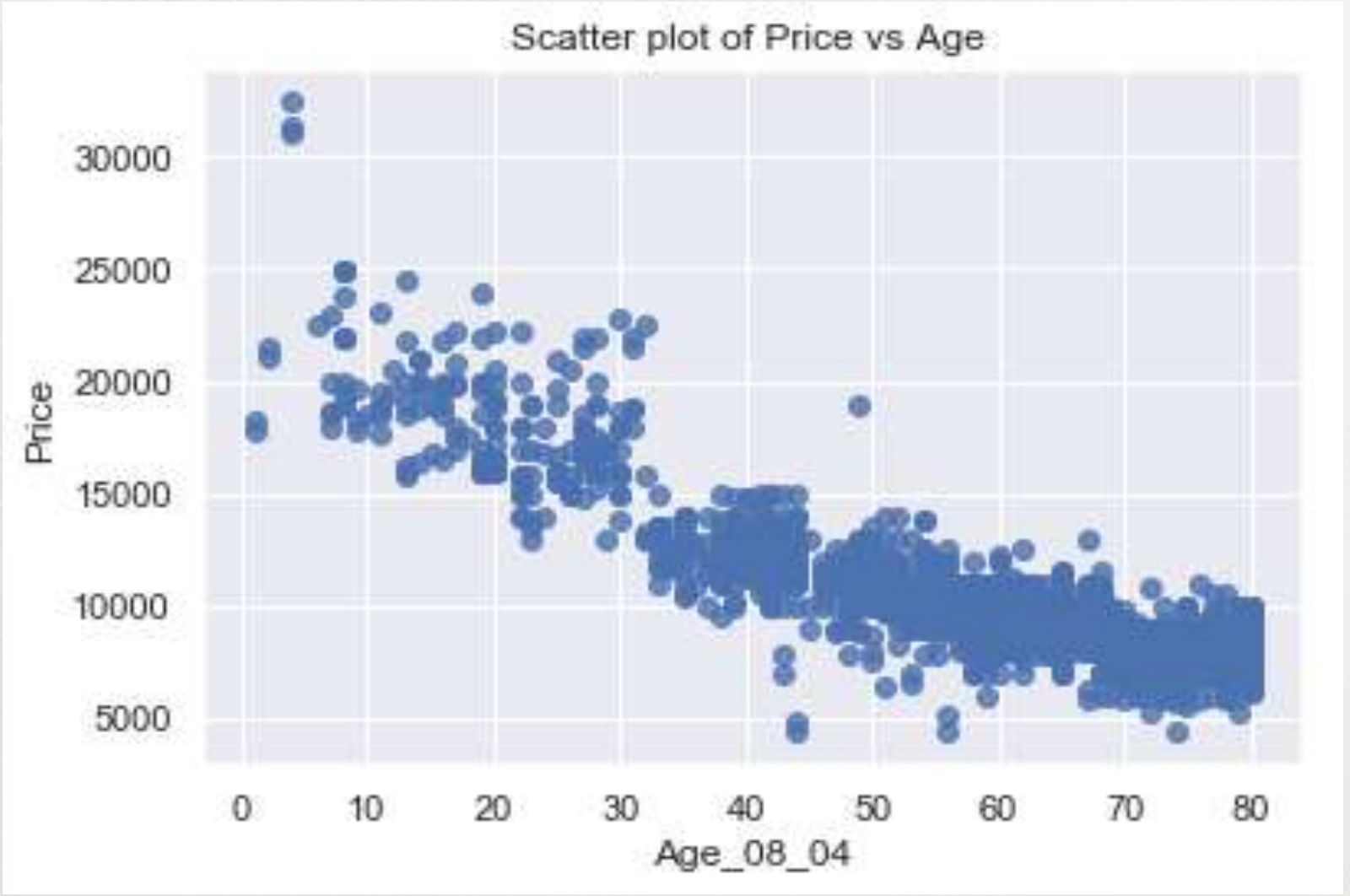
- By default, `fit_reg = True`
- It estimates and plots a regression model relating the x and y variables

Exercise-2

- Scatter plot of *Price vs Age* without the regression fit line

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.regplot(x=data_cars["Age_08_04"],y=data_cars["Price"],
            fit_reg=False)
g=plt.gca()
g.set_title("Scatter plot of Price vs Age ")
```

Output

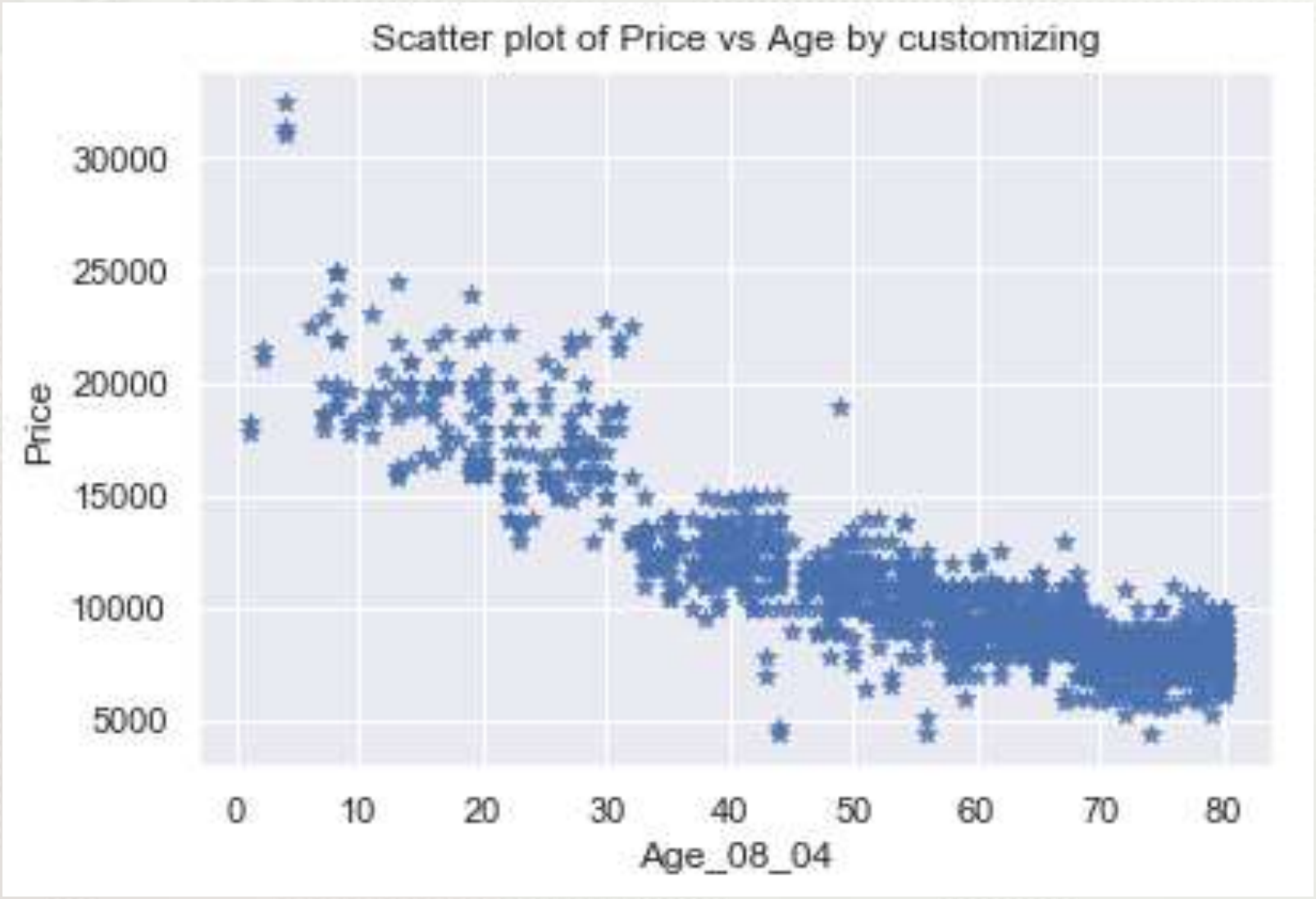

Scatter plot of Price vs Age
Price
30000
25000
20000
15000
10000
5000
0
10
20
30
40
50
60
70
80
Age_08_04

Exercise-3

- Scatter plot of *Price vs Age* by customizing the appearance of markers

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.regplot(x=data_cars["Age_08_04"],y=data_cars["Price"],
            fit_reg=False,marker="*")
g=plt.gca()
g.set_title("Scatter plot of Price vs Age by customizing")
```

Output


Scatter plot of Price vs Age by customizing
Price
30000
25000
20000
15000
10000
5000
0
10
20
30
40
50
60
70
80
Age_08_04

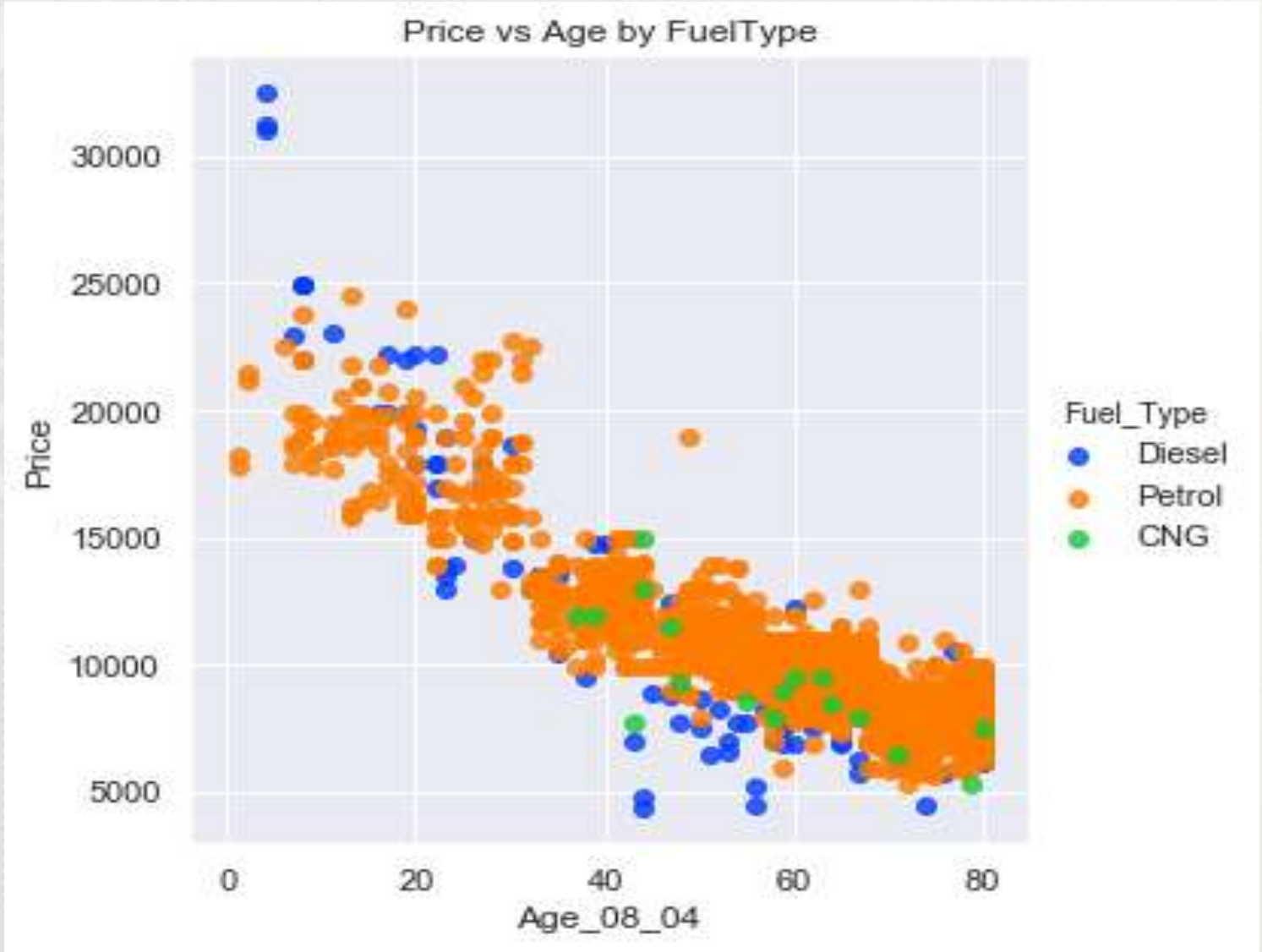
Exercise-4

- Read a Toyota Cars dataset and perform following tasks:
- 1.Set grid background color as darkgrid
- 2.Create a Scatter plot of *Price vs Age by FuelType*
- use Implot() function to create scatter plot
- Implot() methods will take feature attributes and dataset
- 3.Add hue parameter which includes another variable categories(“Fuel_Type” variable) with different colors

Solution

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.set(style="darkgrid")
g=sns.lmplot(x="Age_08_04",y="Price",data=data_cars,fit_reg=False,
             hue="Fuel_Type",legend=True,palette="bright")
#bright- deep,muted,pastel,bright,dark,Set1 and colorblind
ax=plt.gca()
ax.set_title("Price vs Age by FuelType")
```

Output

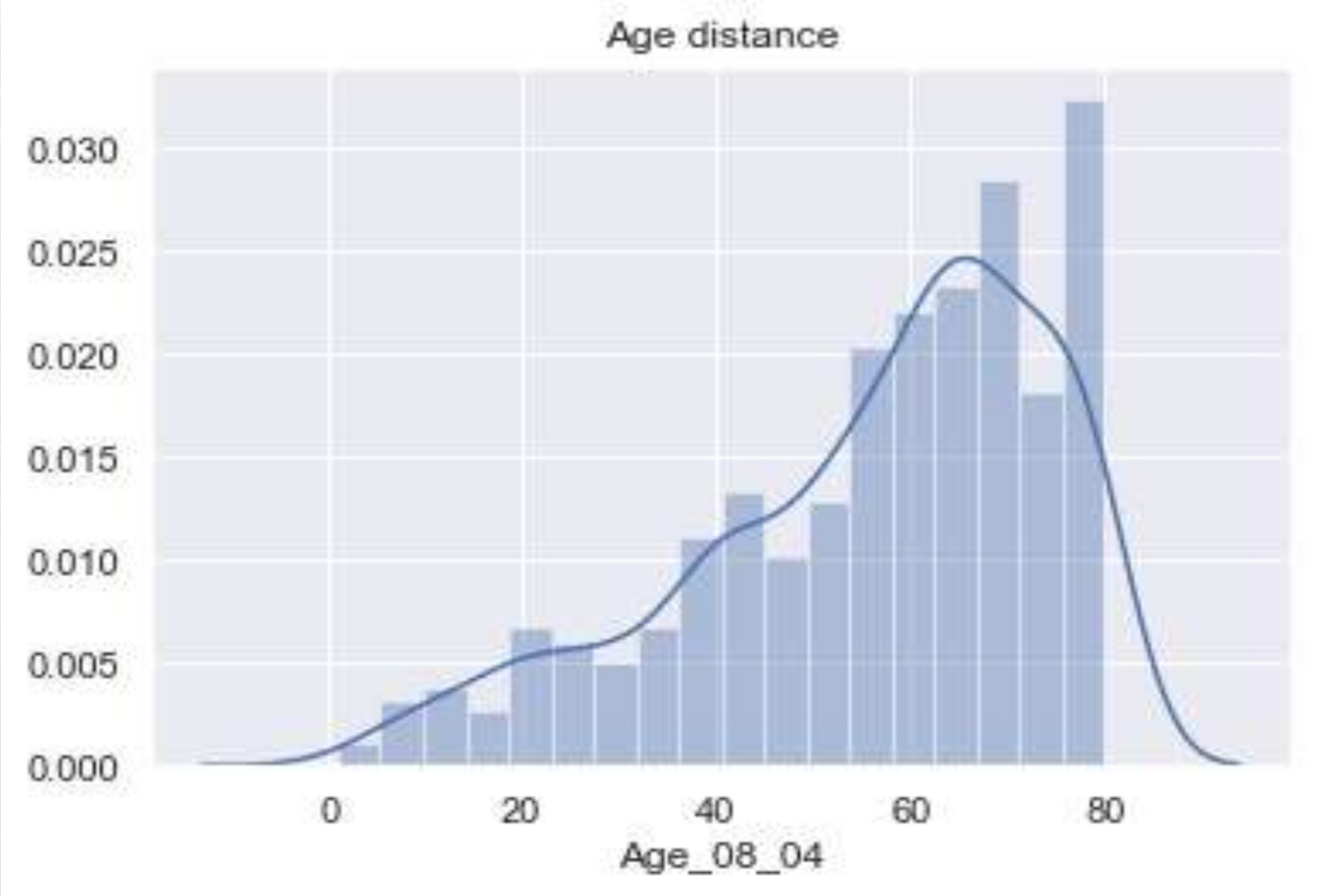

Price vs Age by FuelType
Price
30000
25000
20000
15000
10000
5000
0
20
40
60
80
Age_08_04
Fuel_Type
Diesel
Petrol
CNG

Histogram Exercise-5

- Create a Histogram with default kernel density estimate of Age feature in Toyota cars dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.distplot(data_cars["Age_08_04"])
g=plt.gca()
g.set_title("Age distance")
```

Output

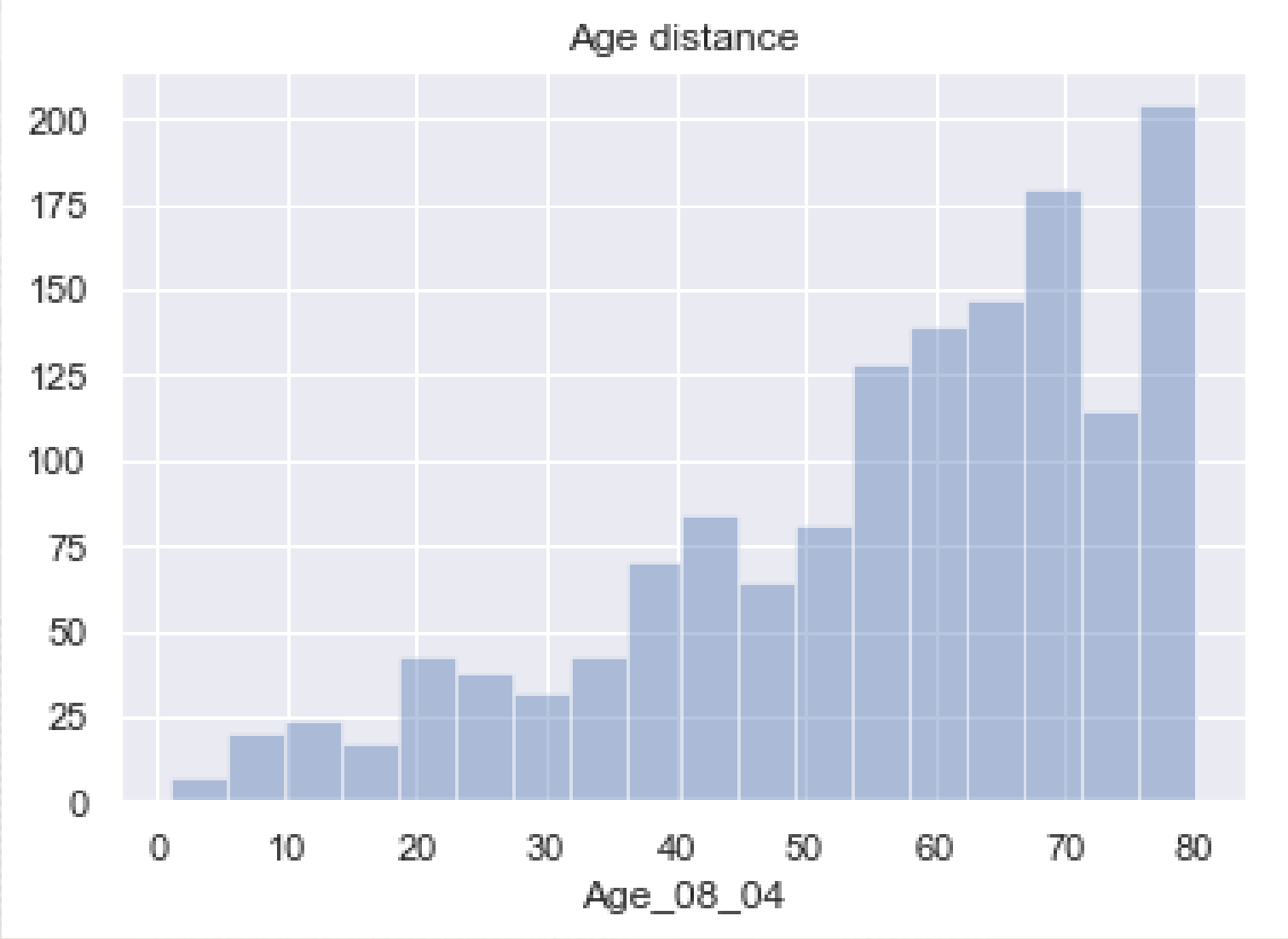

Age distance
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
Age_08_04

Exercise-6

- Create a Histogram without kernel density estimate of Age feature in Toyota cars dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.distplot(data_cars["Age_08_04"],kde=False)
g=plt.gca()
g.set_title("Age distance")
```

Output

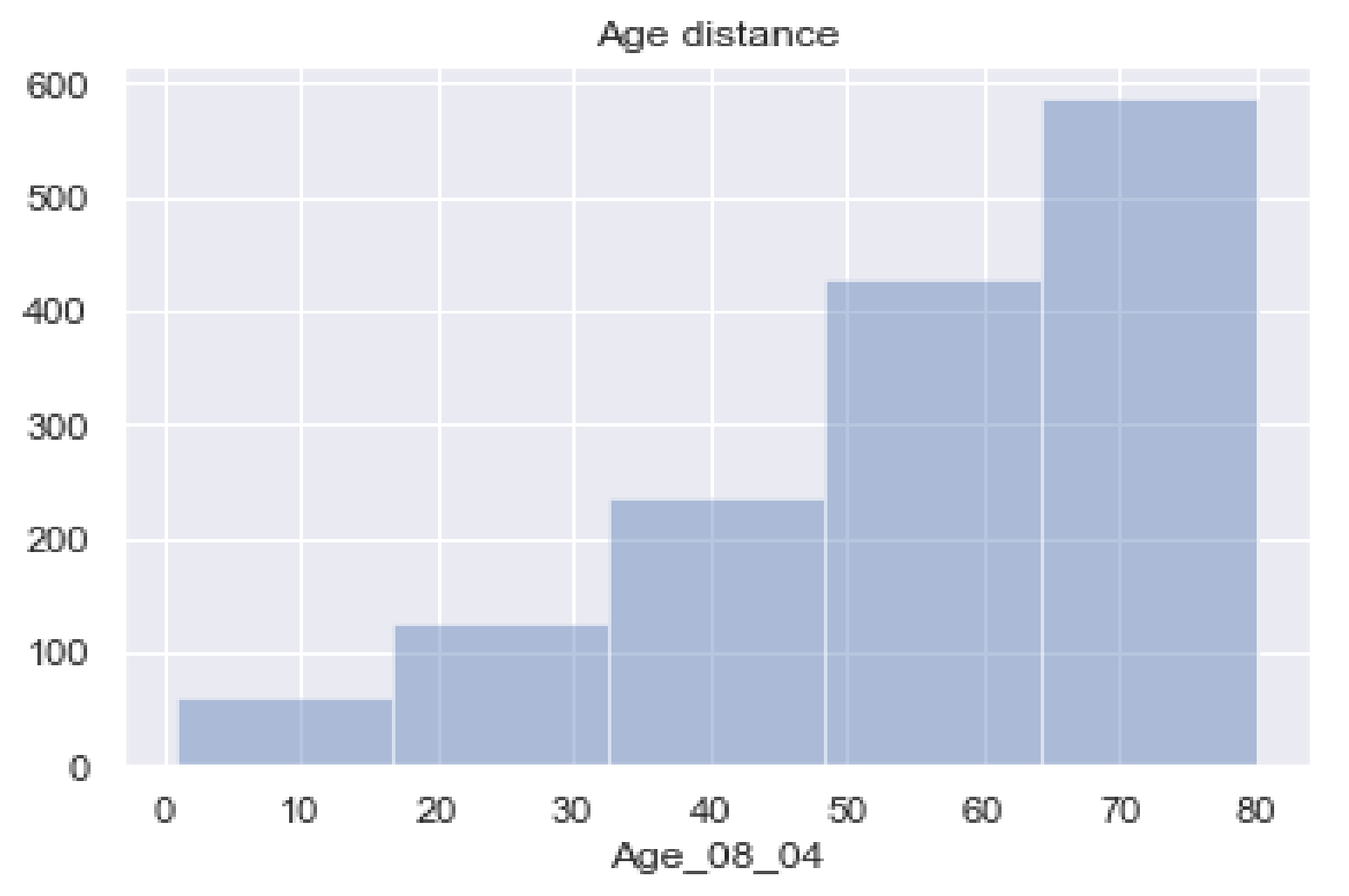

Age distance
200
175
150
125
100
75
50
25
0
0
10
20
30
40
50
60
70
80
Age_08_04

Exercise-7

- Create a Histogram with fixed no. of bins without kernel density estimate of Age feature in Toyota cars dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.distplot(data_cars["Age_08_04"],kde=False,bins=5)
g=plt.gca()
g.set_title("Age distance")
```

Output

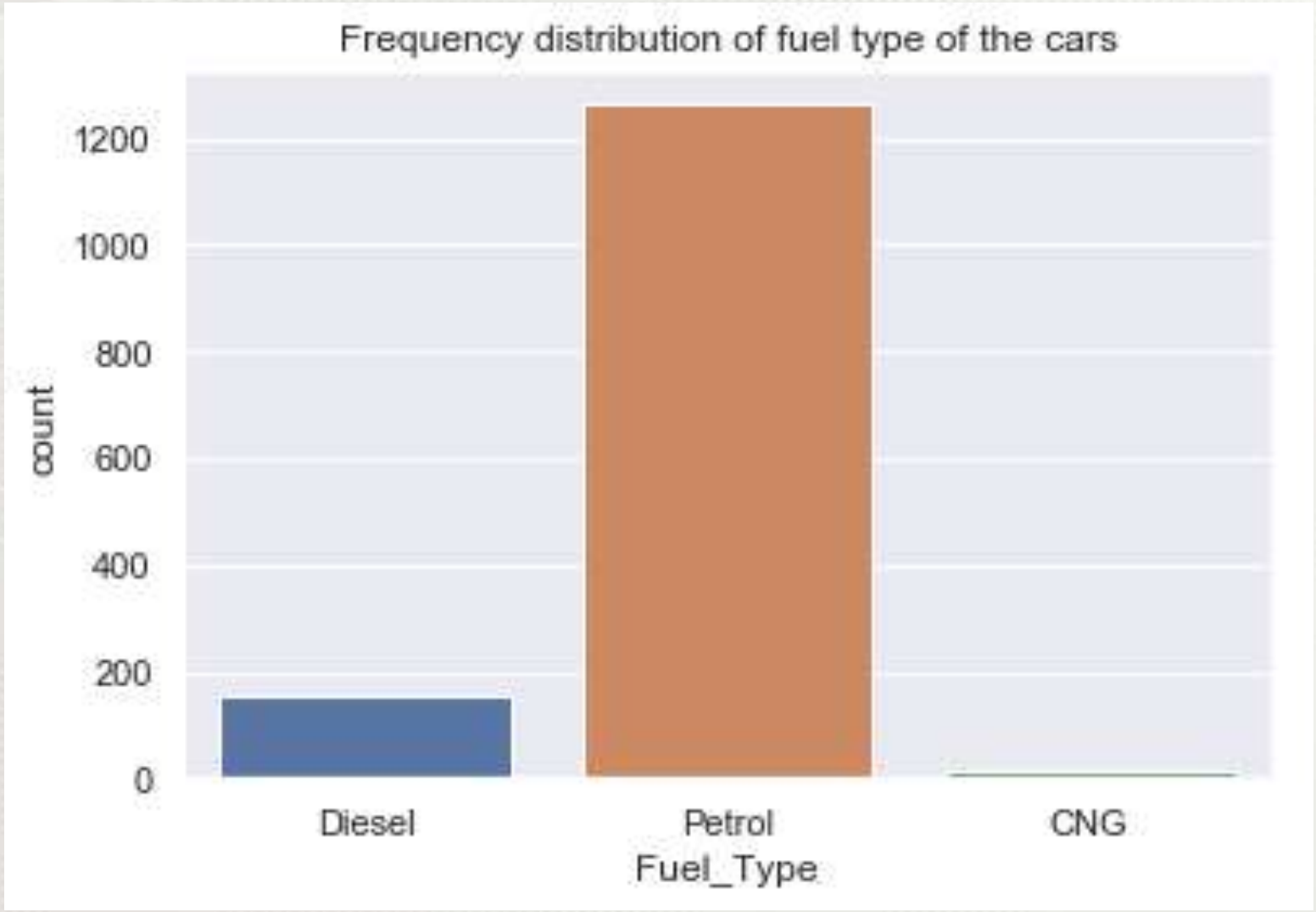

Age distance
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
Age_08_04

Bar plot Ex-8

- Read the Toyota cars dataset and perform the following tasks:
- Create count plot on "Fuel _Type" which says Frequency distribution of fuel type of the cars

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.countplot(x="Fuel_Type",data=data_cars)
g=plt.gca()
g.set_title("Frequency distribution of fuel type of the cars")
```

Output


Frequency distribution of fuel type of the cars
1200
1000
800
600
400
200
0
ount
Diesel
Petrol
CNG
Fuel_Type

Grouped Bar plot

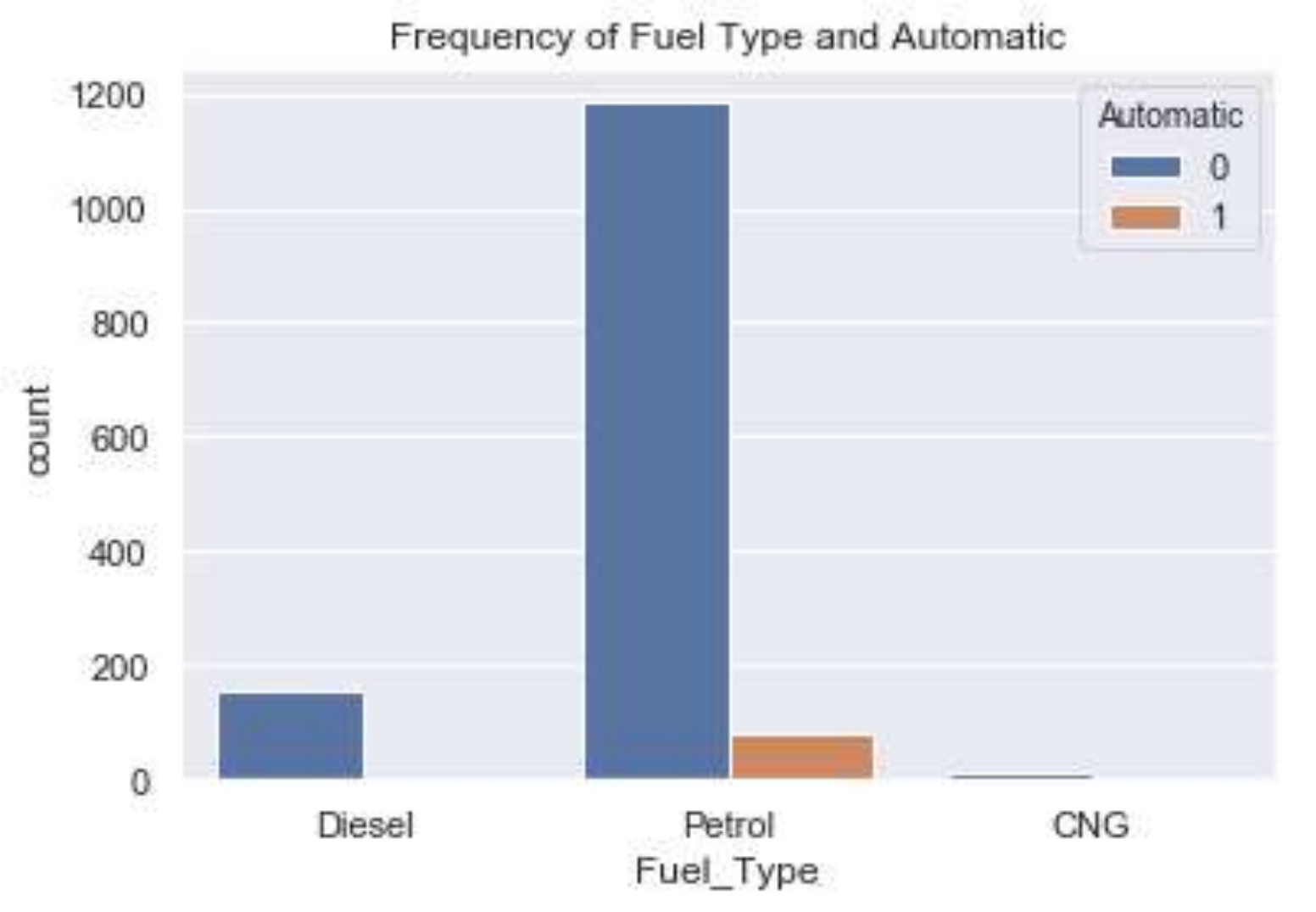
Read a Toyota cars dataset and perform the following tasks

1.Create a Two way table on "Fuel_Type“ and Automatic feature
2.Create a count plot by grouping Automatic by Fuel Type

Solution

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
auto_tab=pd.crosstab(index=data_cars["Automatic"],
                     columns=data_cars["Fuel_Type"],dropna=True)
sns.countplot(x="Fuel_Type",data=data_cars,hue="Automatic")
g=plt.gca()
g.set_title("Frequency of Fuel Type and Automatic")
```

Output

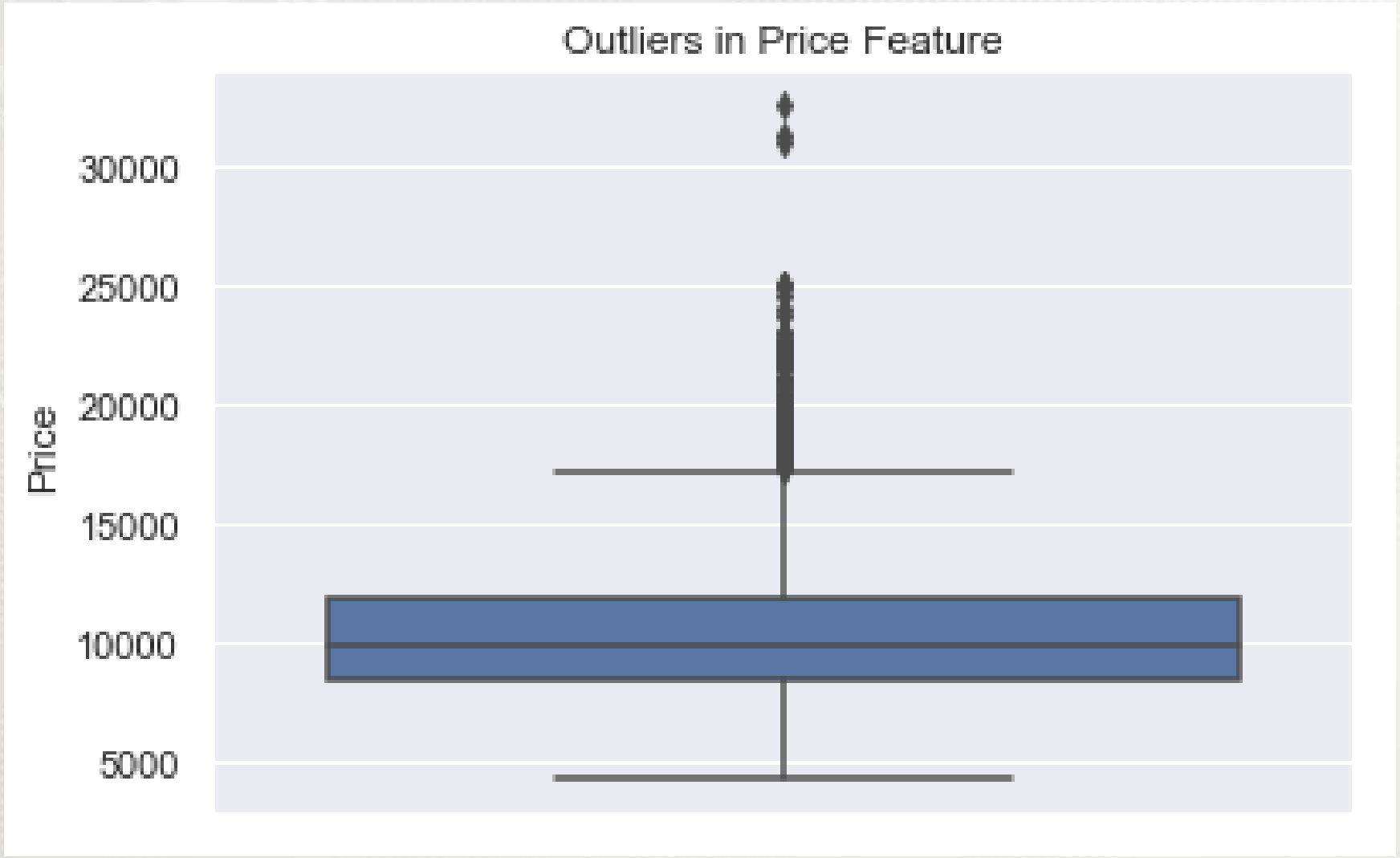

Frequency of Fuel Type and Automatic
1200
1000
800
600
400
200
0
count
Automatic
0
1
Diesel
Petrol
CNG
Fuel_Type

Box and Whiskers plot – numerical variable

- Box and whiskers plot of *Price* to visually interpret the five-number summary

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.boxplot(y=data_cars["Price"])
g=plt.gca()
g.set_title("Outliers in Price Feature")
```

Output

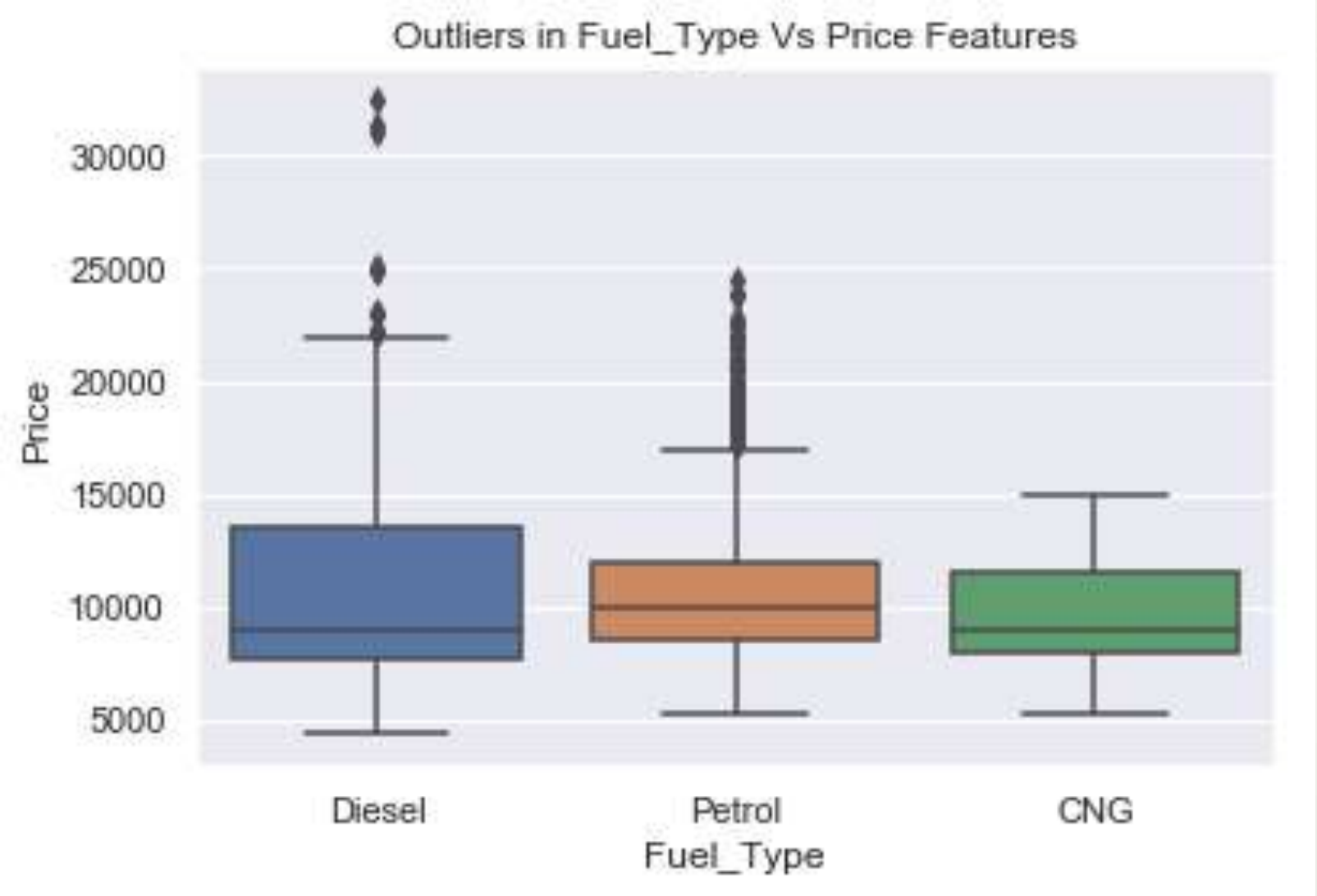

Outliers in Price Feature
Price
30000
25000
20000
15000
10000
5000

Box and Whiskers plot

- Box and whiskers plot for numerical vs categorical variable
- Price of the cars for various fuel types

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.boxplot(x=data_cars["Fuel_Type"],y=data_cars["Price"])
g=plt.gca()
g.set_title("Outliers in Fuel_Type Vs Price Features")
```

Output

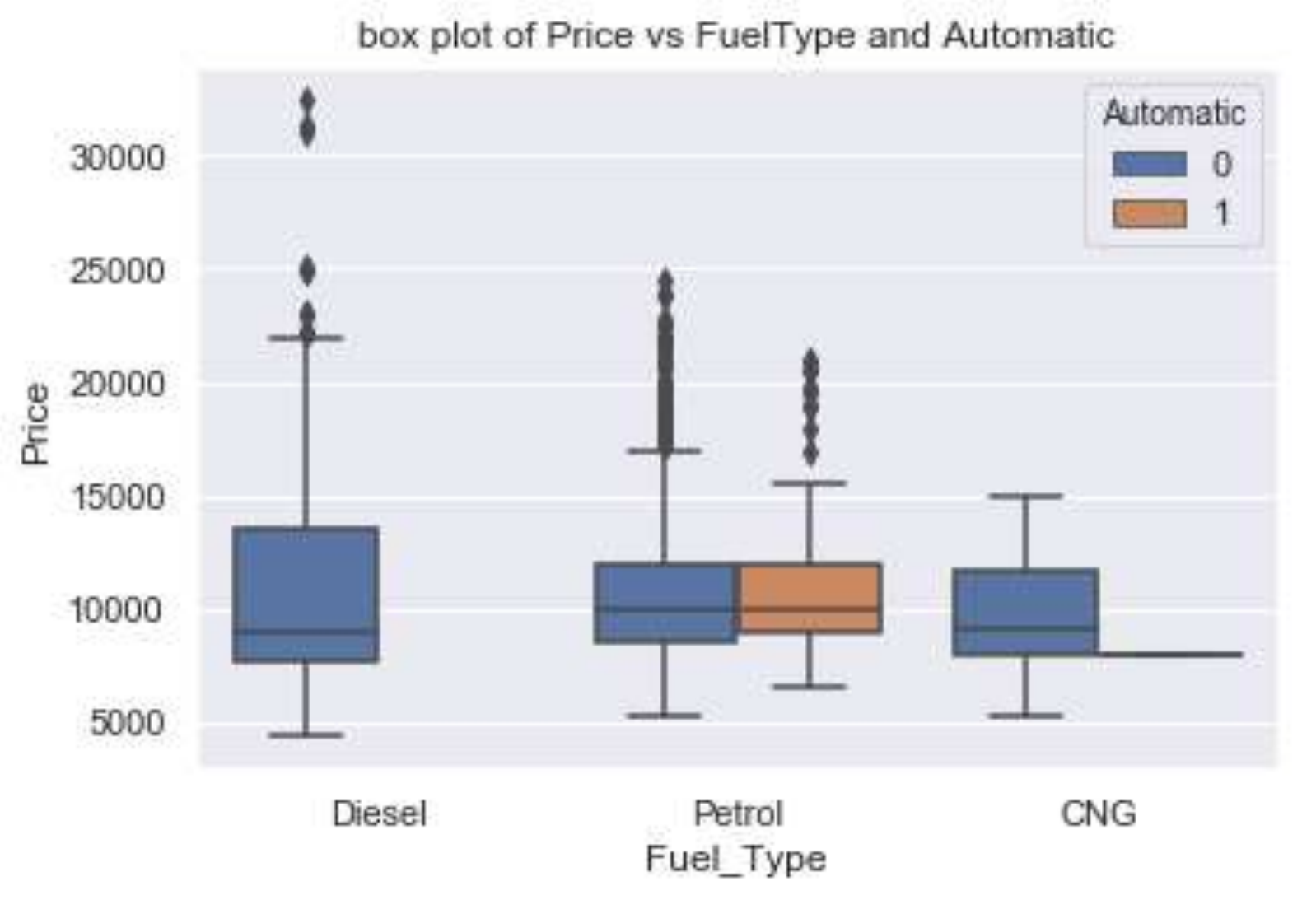

Outliers in Fuel_Type Vs Price Features
Price
30000
25000
20000
15000
10000
5000
Diesel
Petrol
CNG
Fuel_Type

Grouped box and whiskers plot

- Grouped box and whiskers plot of *Price* vs *FuelType* and *Automatic*

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.boxplot(x=data_cars["Fuel_Type"],y=data_cars["Price"],
            data=data_cars,hue="Automatic")
g=plt.gca()
g.set_title("box plot of Price vs FuelType and Automatic")
```

Output

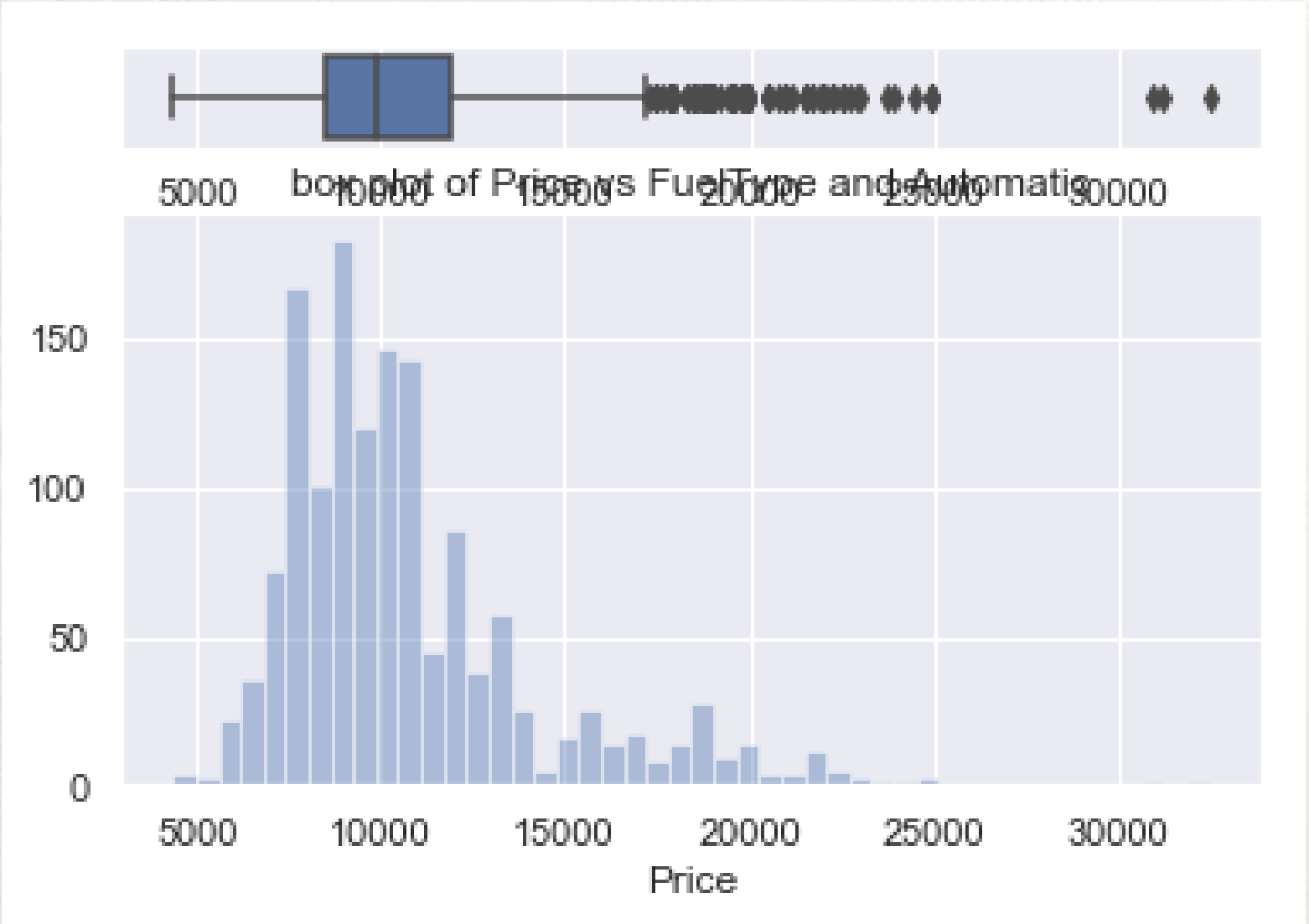

box plot of Price vs FuelType and Automatic
Automatic
0
1
30000
25000
20000
15000
10000
5000
Price
Diesel
Petrol
CNG
Fuel_Type

Box-whiskers plot and Histogram

- Let's plot box-whiskers plot and histogram on the same window
- Split the plotting window into 2 parts

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
f,(ax_box,ax_hist)=plt.subplots(2,gridspec_kw={"height_ratios":[.15,.85]})
sns.boxplot(data_cars["Price"],ax=ax_box)
sns.distplot(data_cars["Price"],ax=ax_hist,kde=False)
g=plt.gca()
g.set_title("box plot of Price vs FuelType and Automatic")
```

Output


boxplot of Price vs FuelType and Automatic
5000
10000
15000
20000
25000
30000
150
100
50
0
5000
10000
15000
20000
25000
30000
Price

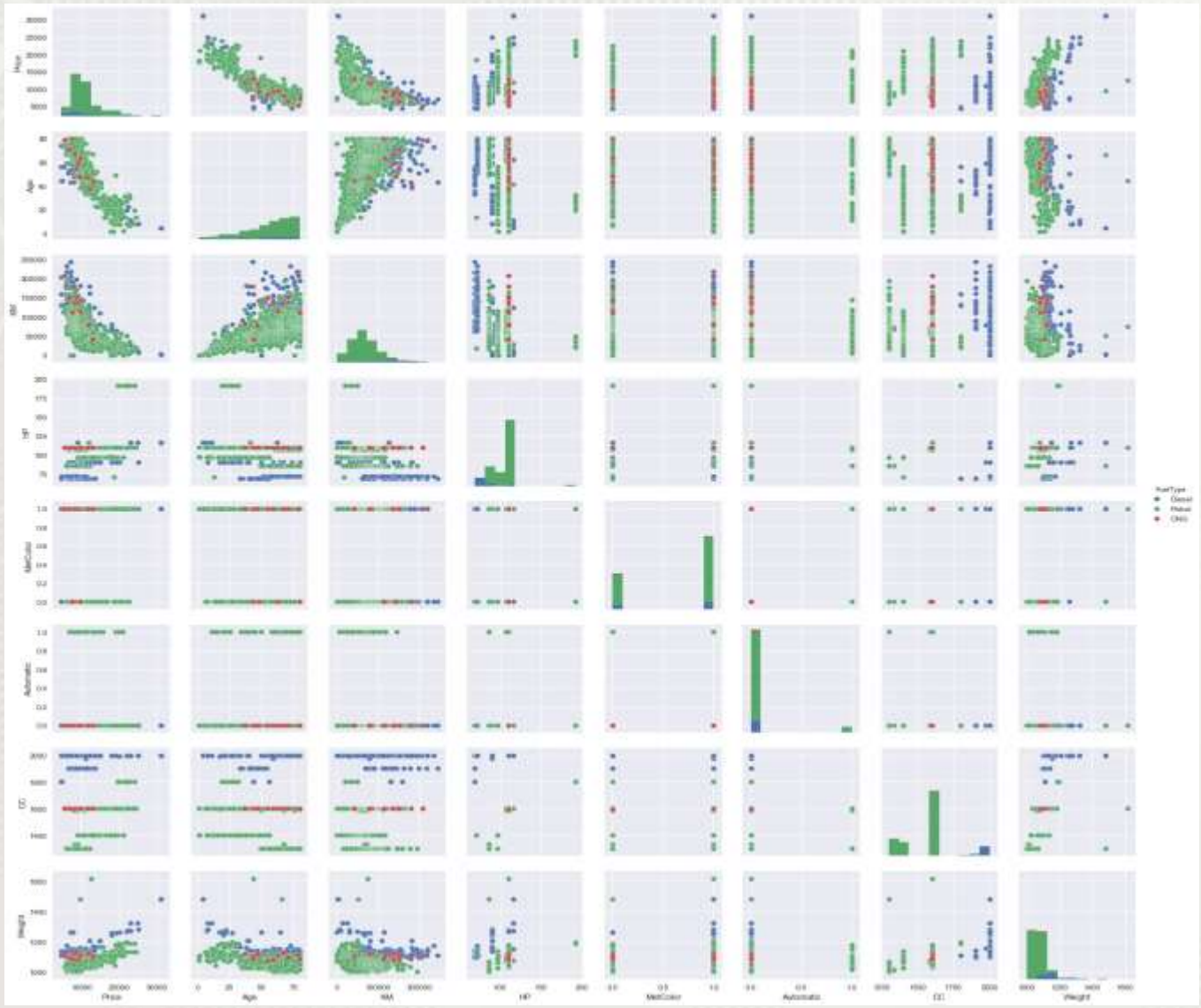
Pairwise plots

- It is used to plot pairwise relationships in a dataset
- Creates scatterplots for joint relationships and histograms for univariate distributions

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data_cars=pd.read_csv("ToyotaCorolla.csv")
sns.pairplot(data_cars, kind="scatter", hue="Fuel_Type")
g=plt.gca()
g.set_title("Pairwise plot")
```

Pairwise plots

Output:



Conclusion

You are aware of

Data Visualization

Data Interpretation

We will proceed with

Data Preprocessing

Kmit

SONET

DATA SCIENCE

THANK YOU