



**Ajeet K. Jain, M. Narsimlu**  
(ML TEAM)- SONET, KMIT, Hyderabad

# Session - 25

This session deals with

- Data Visualization

- Histogram and Scatter plot
- seaborn library

- Creating various graphs using seaborn

- Exercises on Seaborn

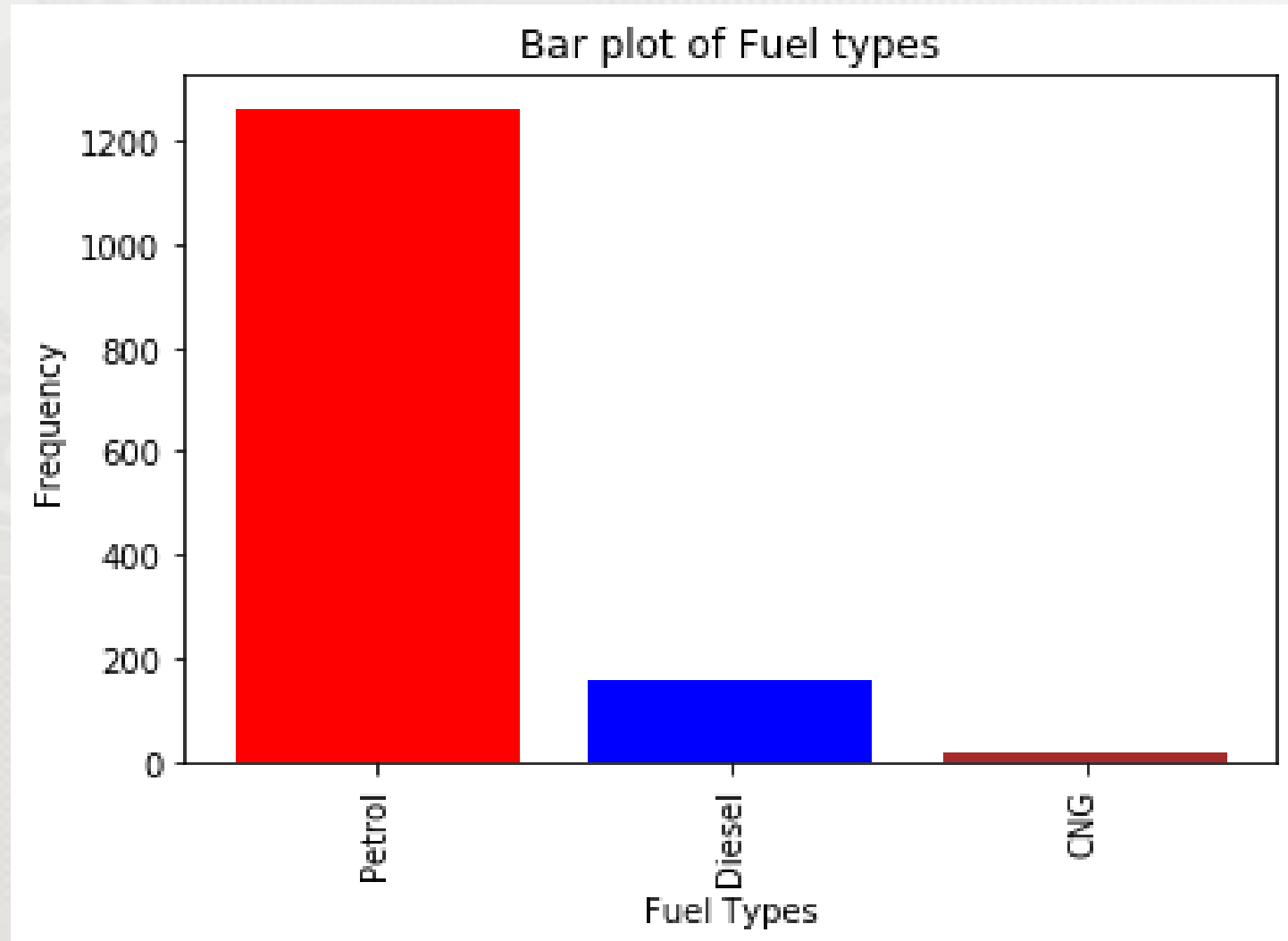
- Data Preprocessing

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 between 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



```
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()
```



# 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.

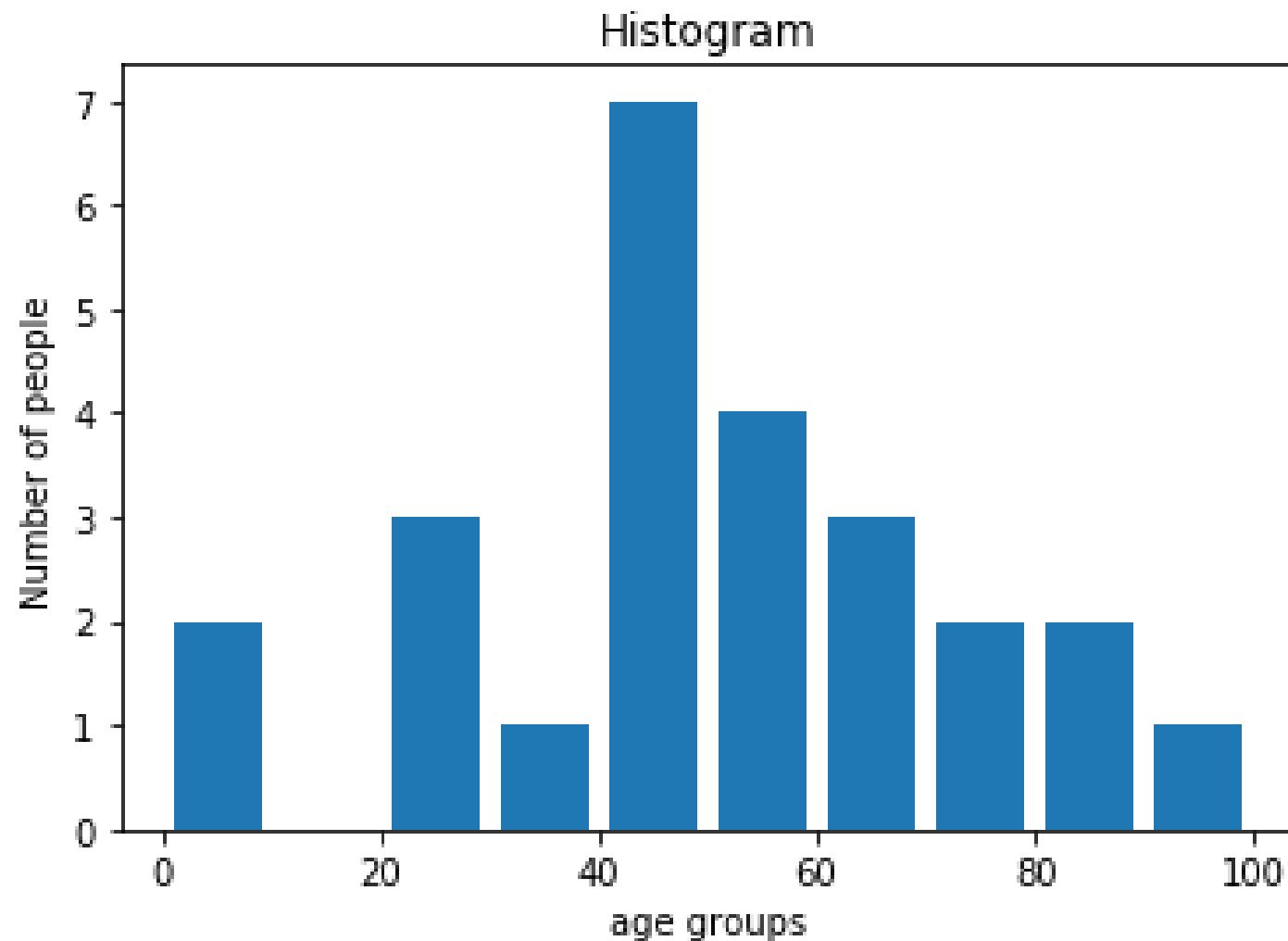
## Exercise-5



- 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]`

```
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()
```

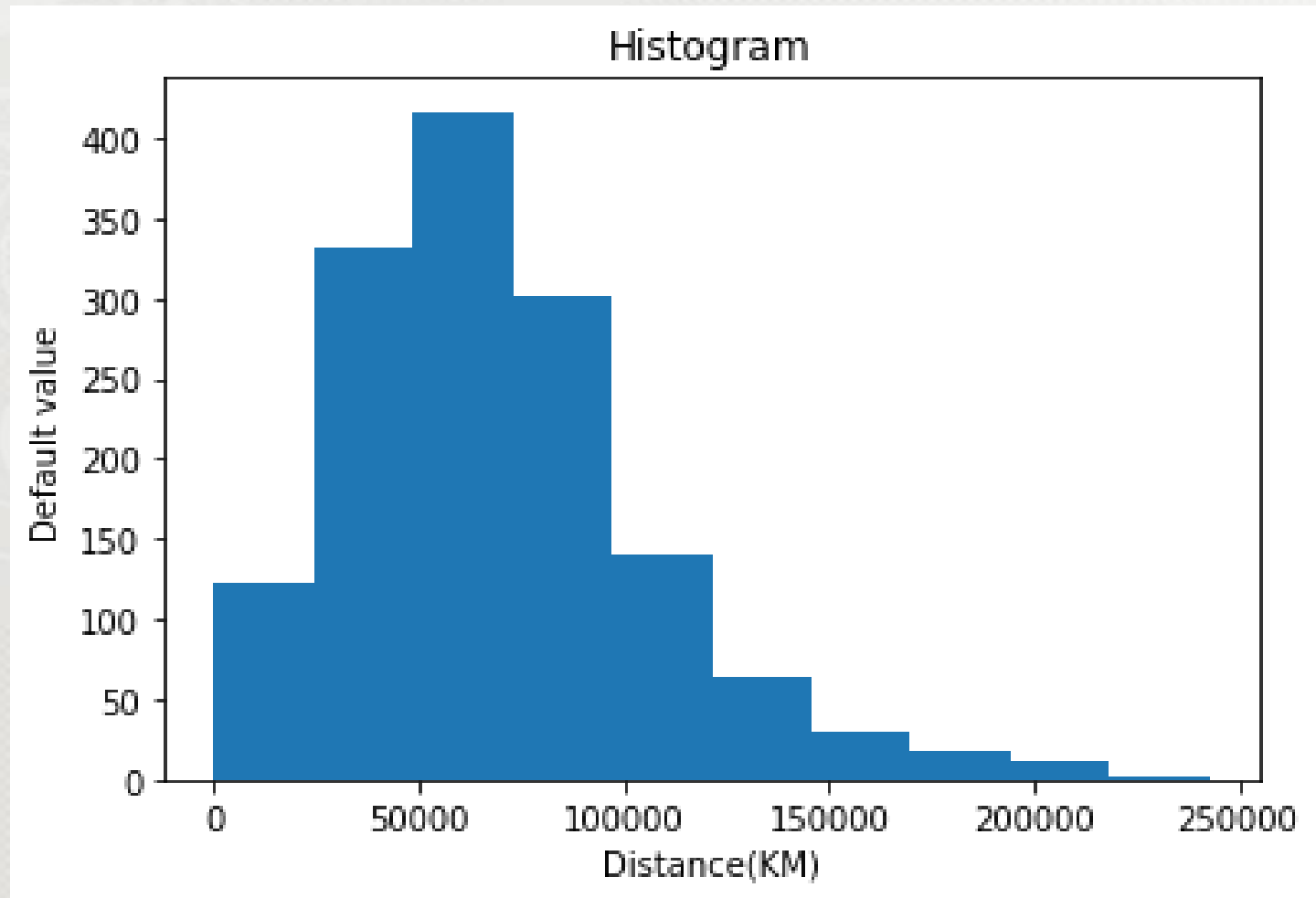




Read a Toyotacars dataset and perform the following tasks

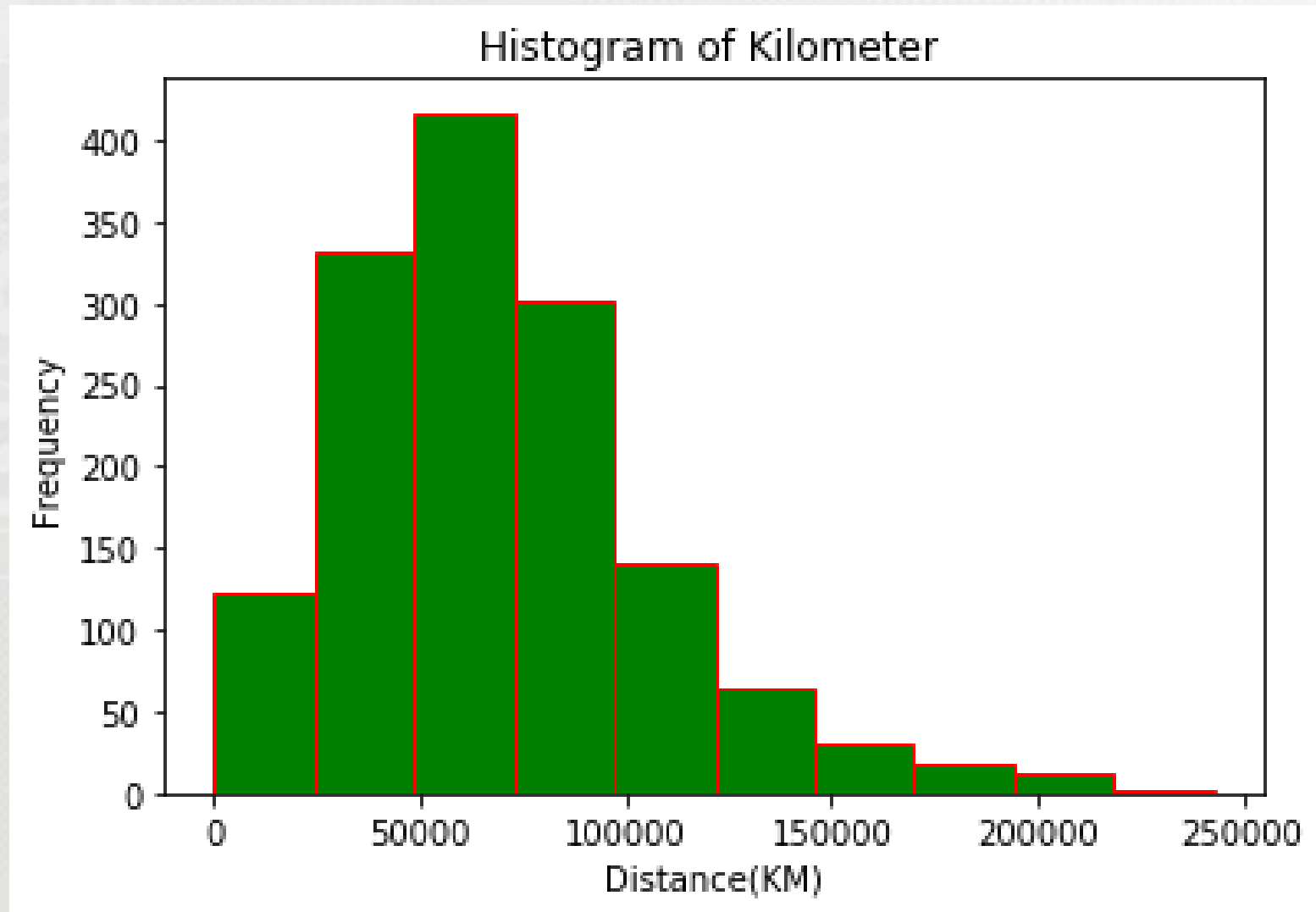
1. Create a histogram with default argument. The car which travelled no of KM.
2. create a histogram with fixed number of bins="10" and color of the bar and also color of the edge

```
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()
```





```
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()
```



- 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

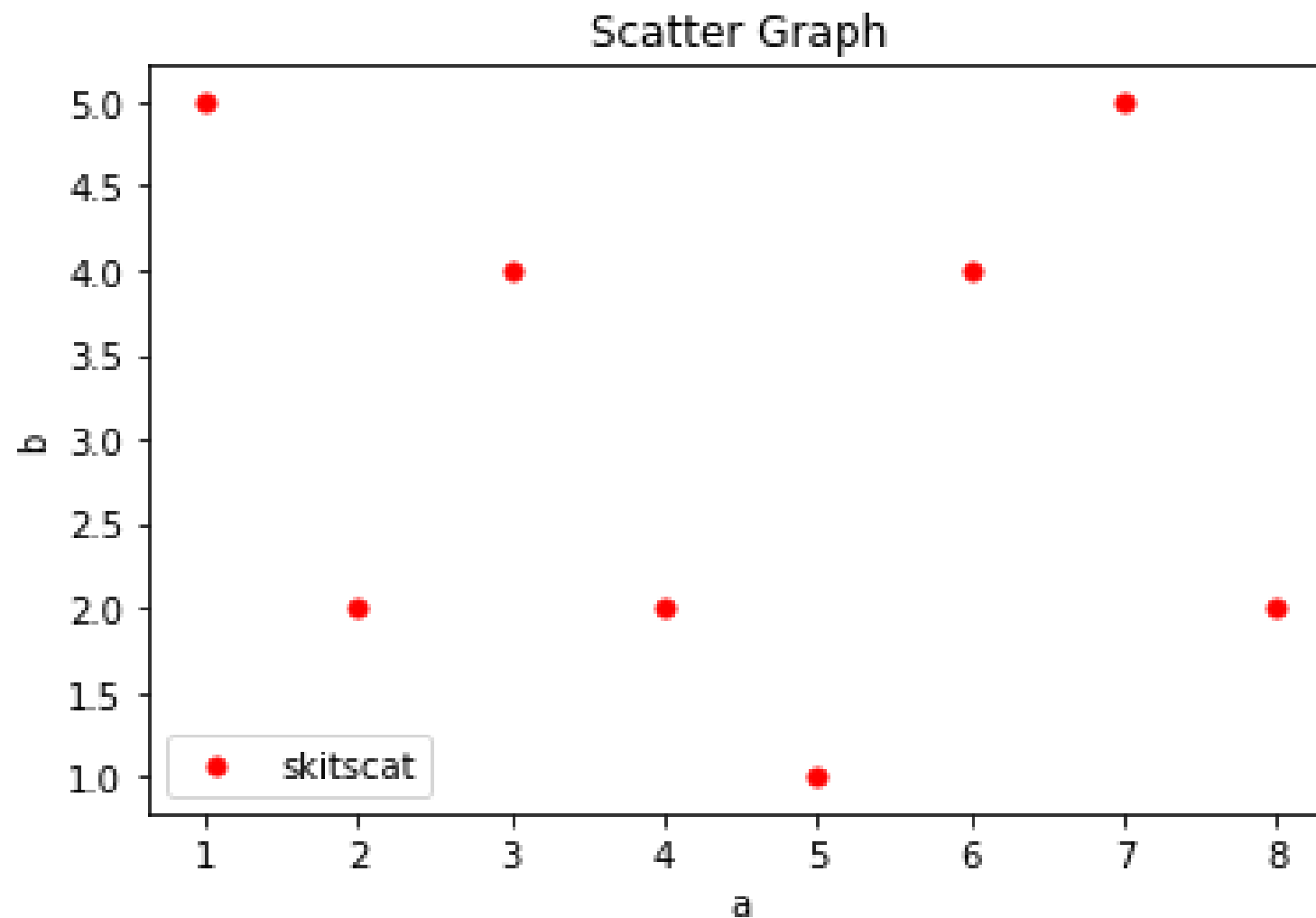


- Create a scatter graph for the following data:
- $a = [1, 2, 3, 4, 5, 6, 7, 8]$
- $b = [5, 2, 4, 2, 1, 4, 5, 2]$

```
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()
```

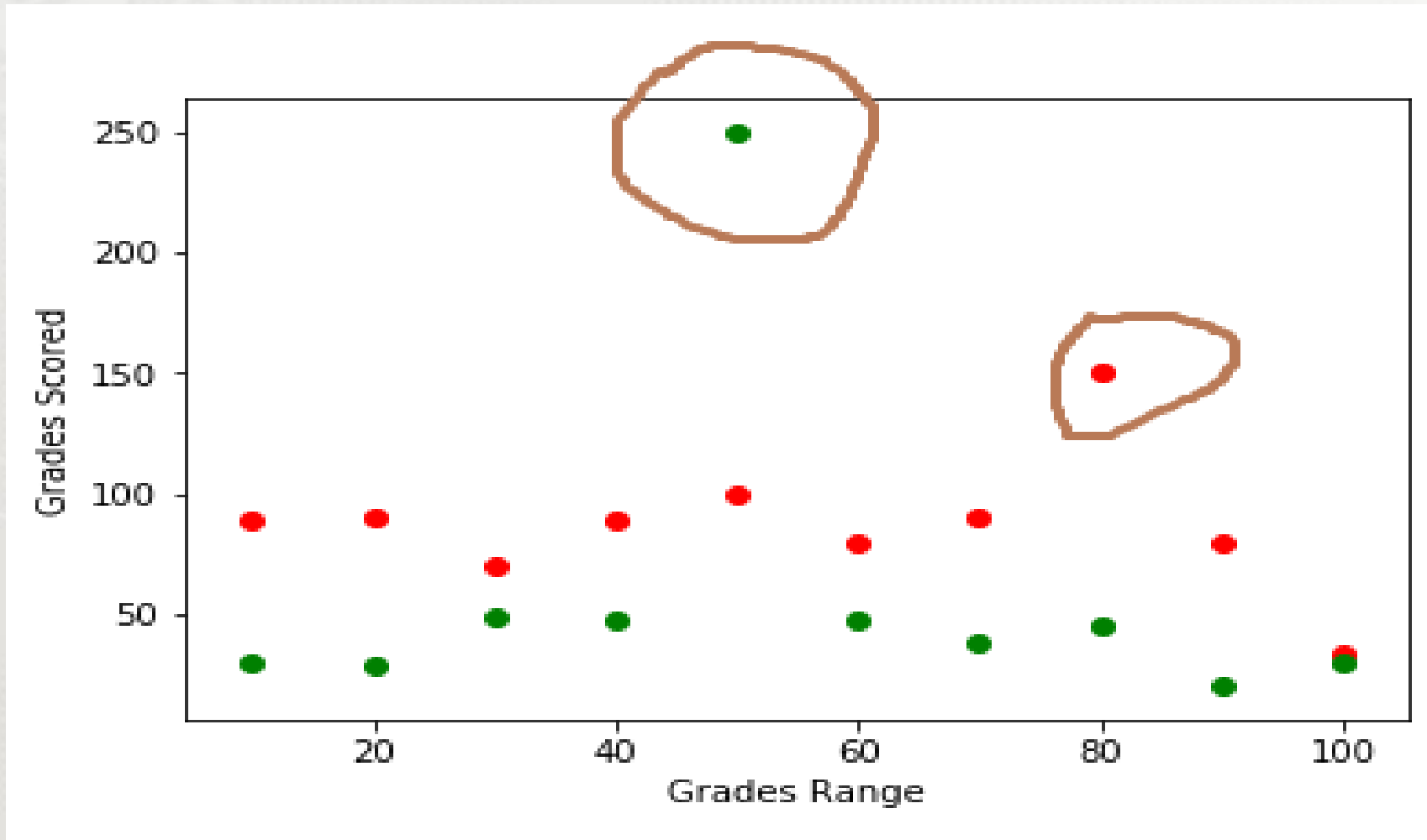


## Exercise-6



- 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()
```



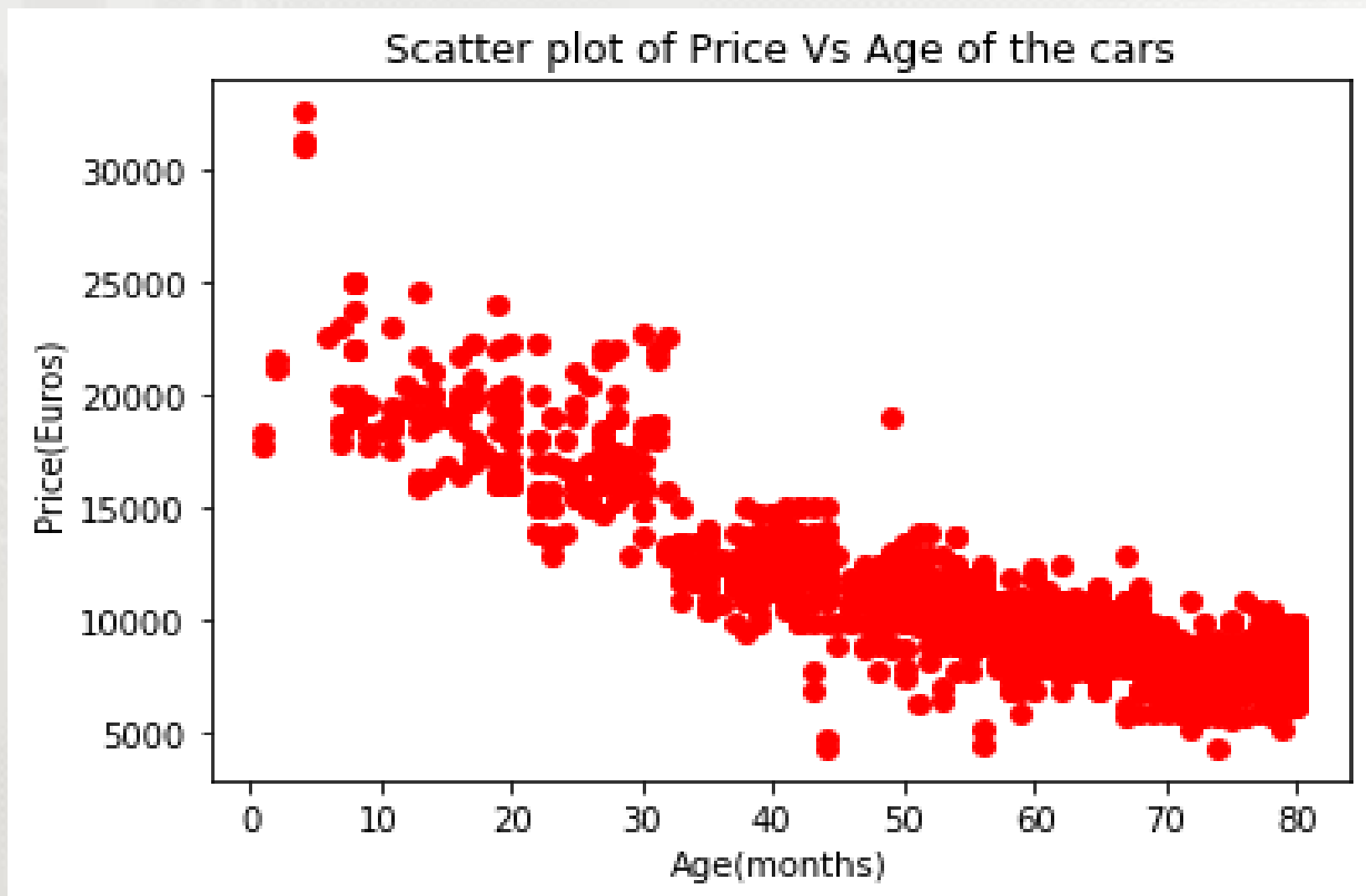


Read a Toyota cars dataset and perform the following tasks

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



```
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()
```



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.



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.

create basic plots using *seaborn* library:

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

- 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

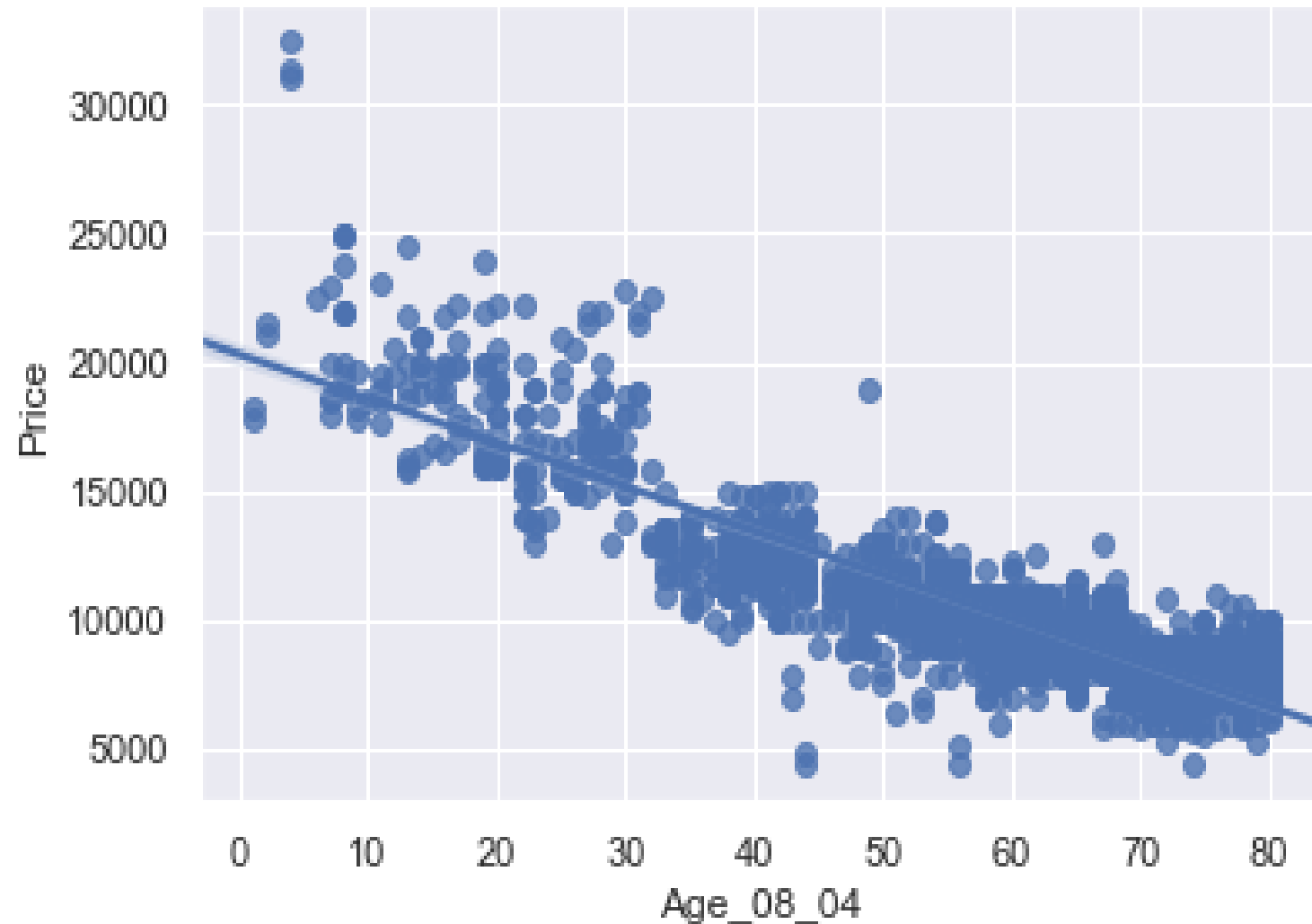
Read a Toyota cars dataset and perform following tasks:

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"])
```



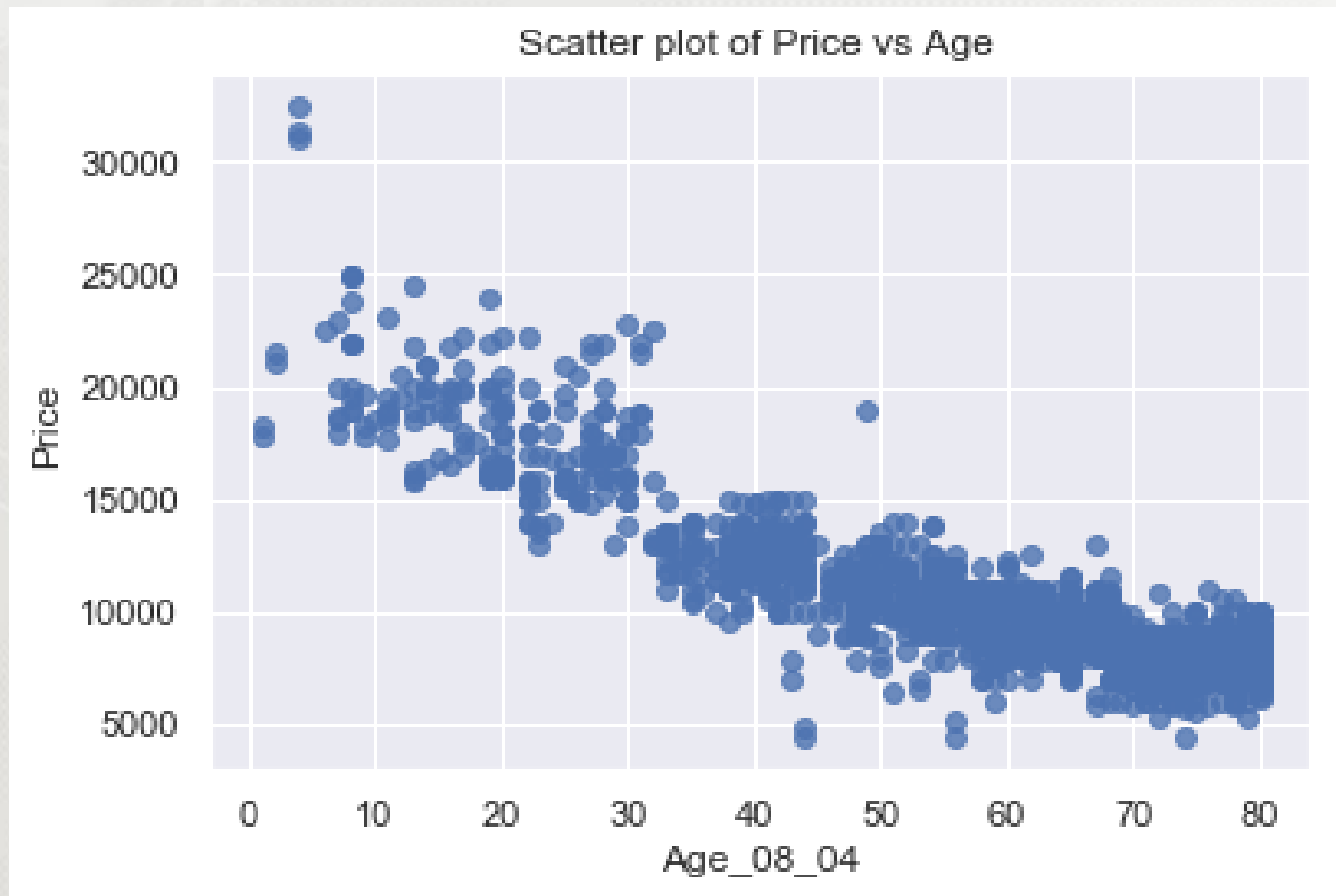


- By default, `fit_reg = True`
- It estimates and plots a regression model relating the x and y variables



- 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 ")
```



## 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")
```



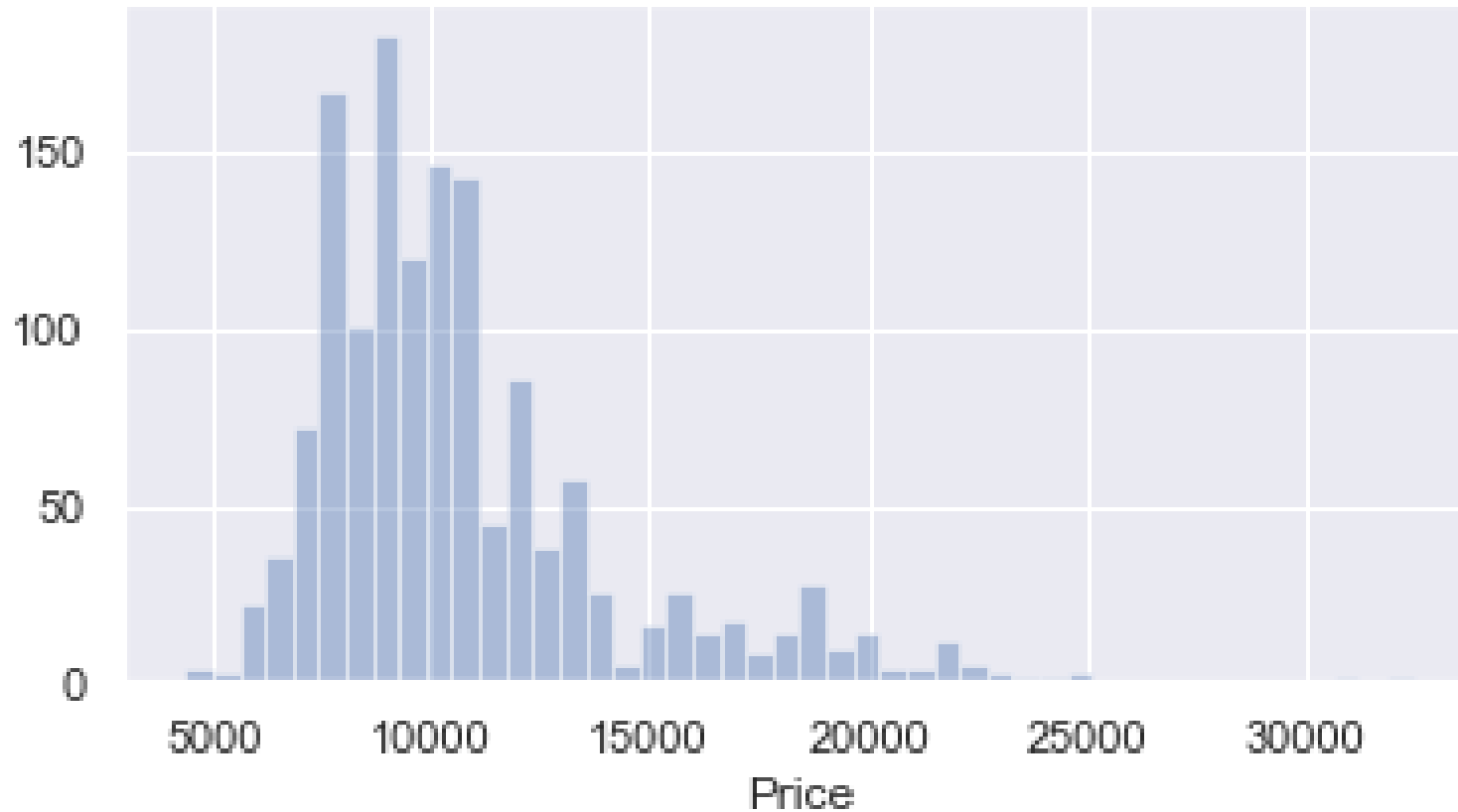
- Let's plot box-whiskers plot and histogram on the same window
- Split the plotting window into 2 parts

```
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")
```





box plot of Price vs FuelType and Automatic



# Conclusion



You are aware of

Data Visualization

Data Interpretation

We will proceed with

Data Preprocessing



THANK  
YOU