Data visualization
Part I

GITAA
Transforming careers

In this lecture

We will learn how to create basic plots using *matplotlib* library

- Scatter plot
- Histogram
- Bar plot

Data Visualization

- Data visualization allows us to quickly interpret the data and adjust different variables to see their effect
- Technology is increasingly making it easier for us to do so

Why visualize data?

- Observe the patterns
- Identify extreme values that could be anomalies
- Easy interpretation

Popular plotting libraries in Python

Python offers multiple graphing libraries that offers diverse features

• ***matplotlib***	• to create 2D graphs and plots
• ***pandas visualization***	• easy to use interface, built on Matplotlib
• ***seaborn***	• provides a high-level interface for drawing attractive and informative statistical graphics
• ***ggplot***	• based on R's ggplot2, uses Grammar of Graphics
• ***plotly***	• can create interactive plots

Matplotlib

- Matplotlib is a 2D plotting library which produces good quality figures

- Although it has its origins in emulating the MATLAB graphics commands, it is independent of MATLAB

- It makes heavy use of NumPy and other extension code to provide good performance even for large arrays

Scatter plot

Scatter Plot

What is a scatter plot?

- A scatter plot is a set of points that represents the values obtained for two different variables plotted on a horizontal and vertical axes

When to use scatter plots?

- Scatter plots are used to convey the relationship between two numerical variables
- Scatter plots are sometimes called correlation plots because they show how two variables are correlated

Importing data into Spyder

- Importing necessary libraries

```
import pandas as pd
```

‘pandas’ library to work with dataframes

```
import numpy as np
```

‘numpy’ library to do numerical operations

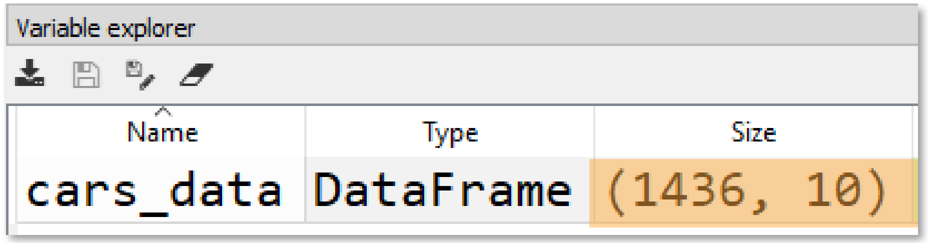
```
import matplotlib.pyplot as plt
```

‘matplotlib’ library to do visualization

Importing data into Spyder

- Importing data

```
cars_data = pd.read_csv('Toyota.csv',index_col=0,
                        na_values=["??","????"])
```

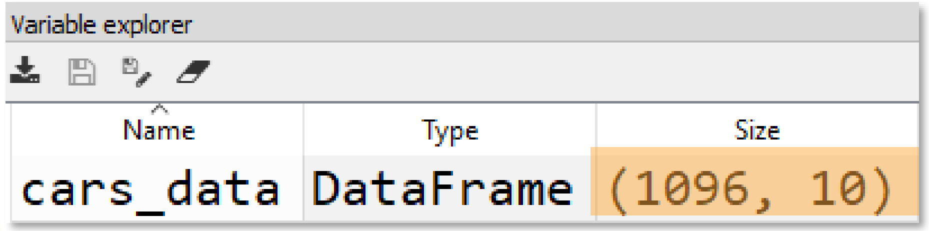


Variable explorer

Name	Type	Size
cars_data	DataFrame	(1436, 10)

- Removing missing values from the dataframe

```
cars_data.dropna(axis = 0, inplace=True)
```



Variable explorer

Name	Type	Size
cars_data	DataFrame	(1096, 10)

Scatter plot

```
                  x                       y
plt.scatter(cars_data['Age'], cars_data['Price'], c='red')
```

```
plt.title('Scatter plot of Price vs Age of the cars')
```

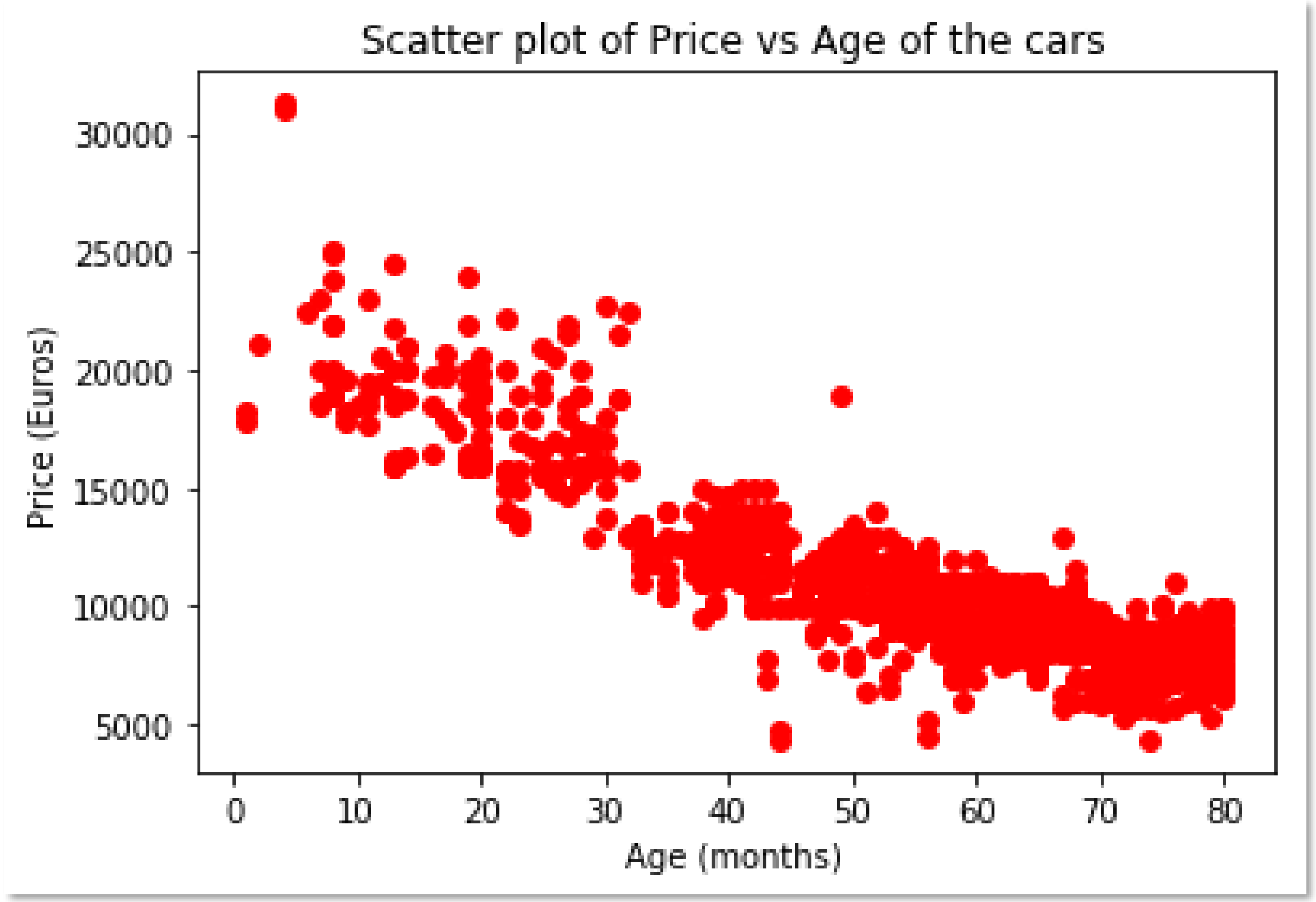
```
plt.xlabel('Age (months)')
```

```
plt.ylabel('Price (Euros)')
```

```
plt.show()
```

Scatter plot

- The price of the car decreases as age of the car increases



Histogram

Histogram

What is a histogram?

- It is a graphical representation of data using bars of different heights
- Histogram groups numbers into ranges and the height of each bar depicts the frequency of each range or bin

When to use histograms?

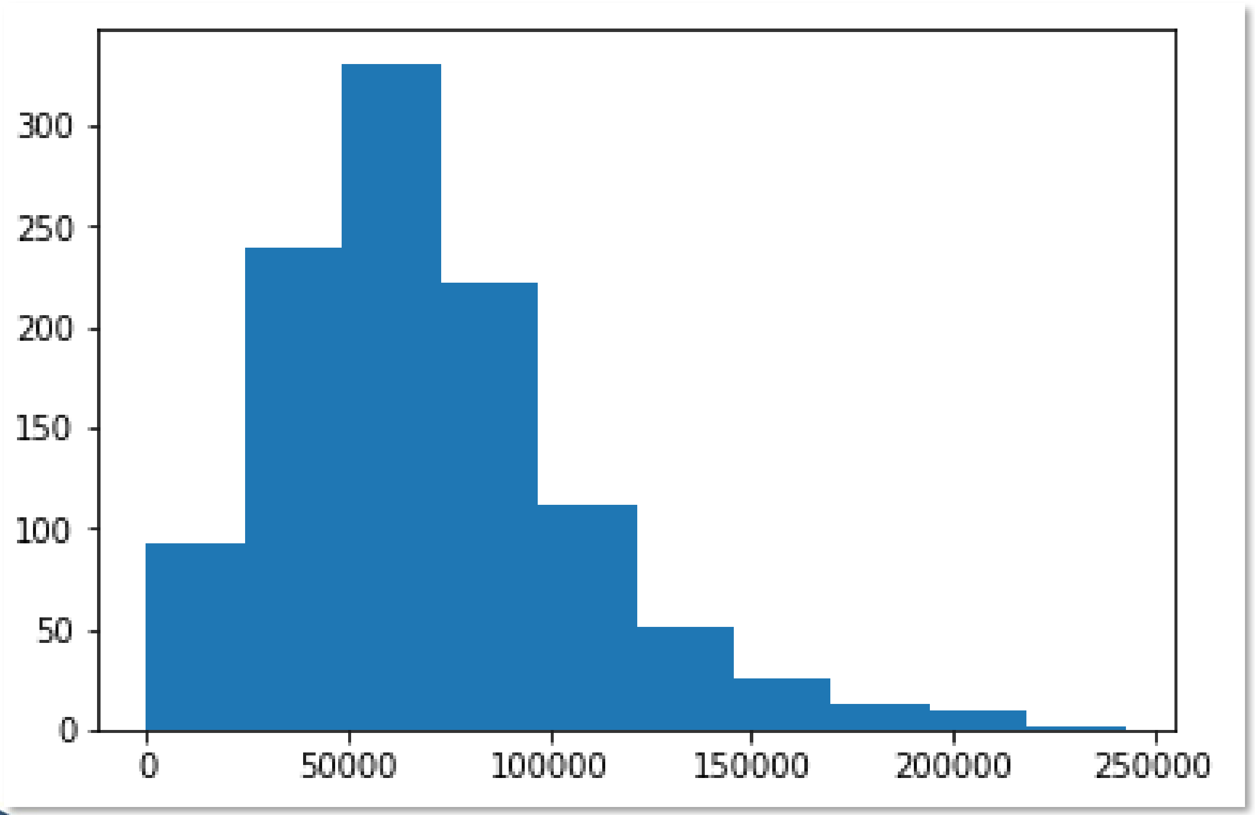
- To represent the frequency distribution of numerical variables

Histogram

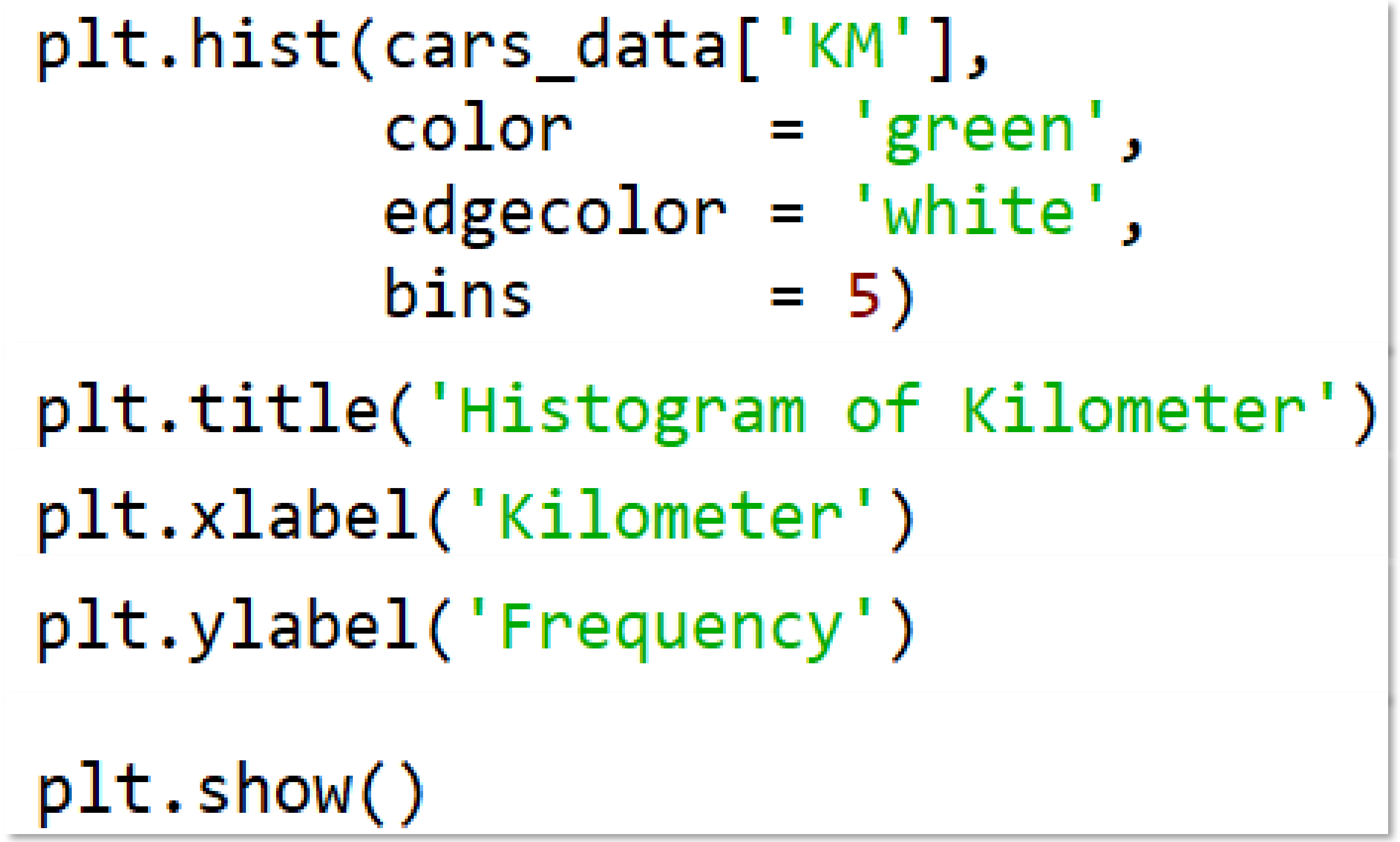
x

```
plt.hist(cars_data['KM'])
```

→ Histogram with default arguments



Histogram

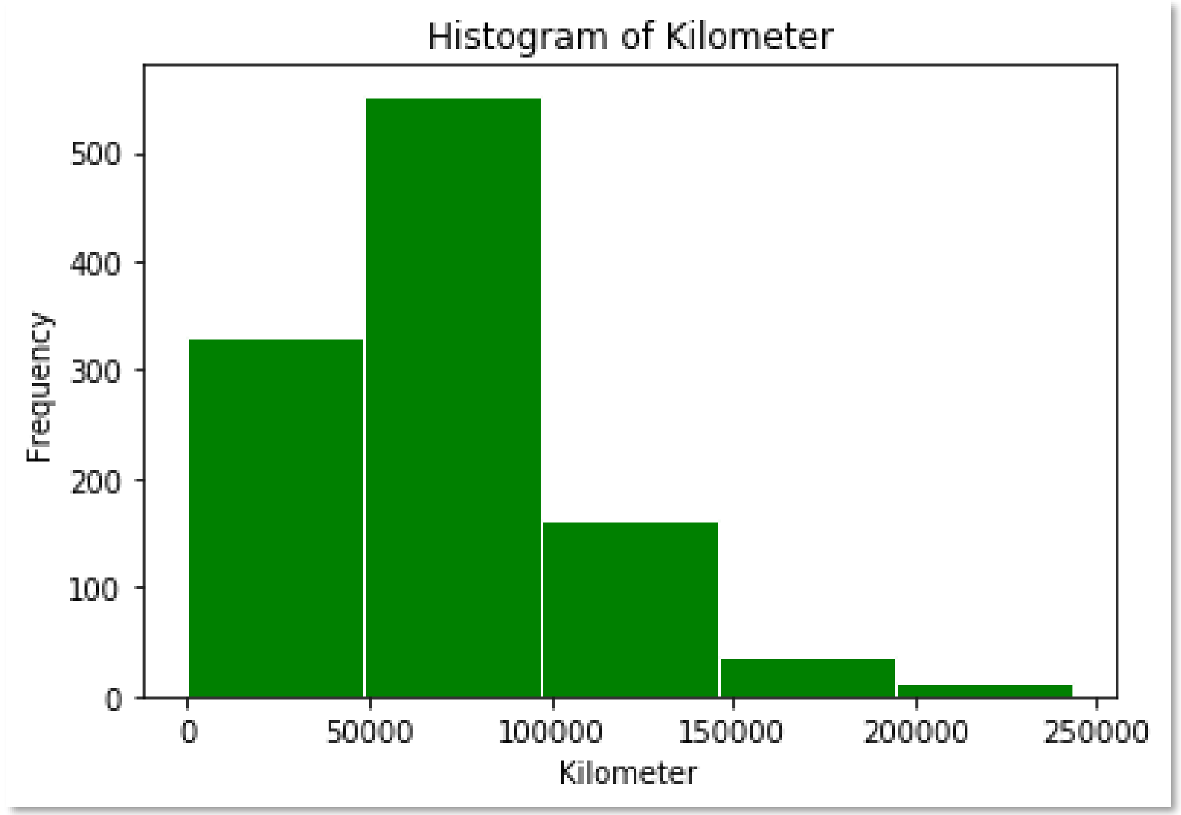


```
plt.hist(cars_data['KM'],
         color     = 'green',
         edgecolor = 'white',
         bins      = 5)
plt.title('Histogram of Kilometer')
plt.xlabel('Kilometer')
plt.ylabel('Frequency')

plt.show()
```

Histogram

- Frequency distribution of kilometre of the cars shows that most of the cars have travelled between 50000 – 100000 km and there are only few cars with more distance travelled



Bar plot

Bar plot

What is a bar plot?

- A bar plot is a plot that presents categorical data with rectangular bars with lengths proportional to the counts that they represent

When to use bar plot?

- To represent the frequency distribution of categorical variables
- A bar diagram makes it easy to compare sets of data between different groups

Bar plot

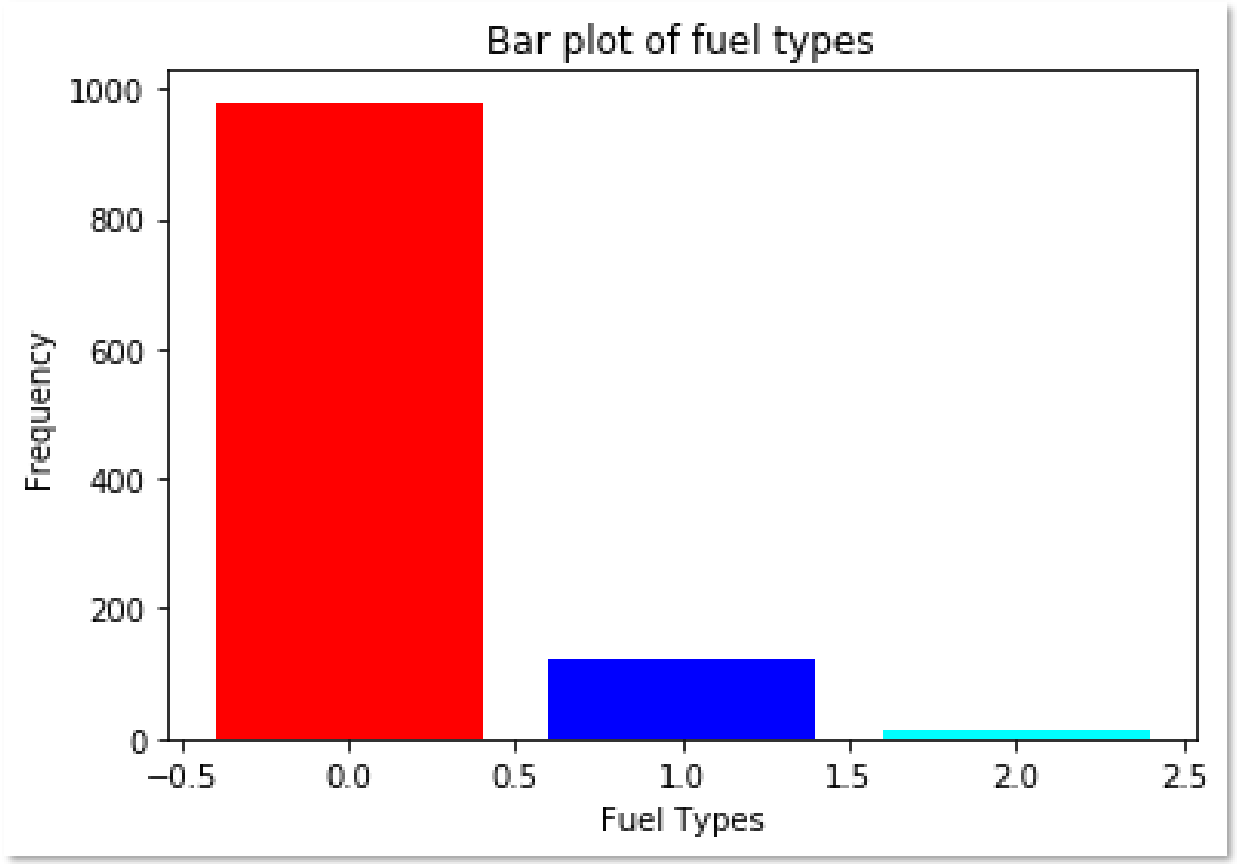
```
counts   = [979, 120, 12]
fuelType = ('Petrol', 'Diesel', 'CNG')
index    = np.arange(len(fuelType))
```

x　　height of the bars

```
plt.bar(index, counts, color=['red', 'blue', 'cyan'])
plt.title('Bar plot of fuel types')
plt.xlabel('Fuel Types')
plt.ylabel('Frequency')
plt.show()
```

Bar plot

- Frequency distribution of fuel type



Bar plot

```
counts   = [979, 120, 12]
fuelType = ('Petrol', 'Diesel', 'CNG')
index    = np.arange(len(fuelType))
```

x   height of the bars

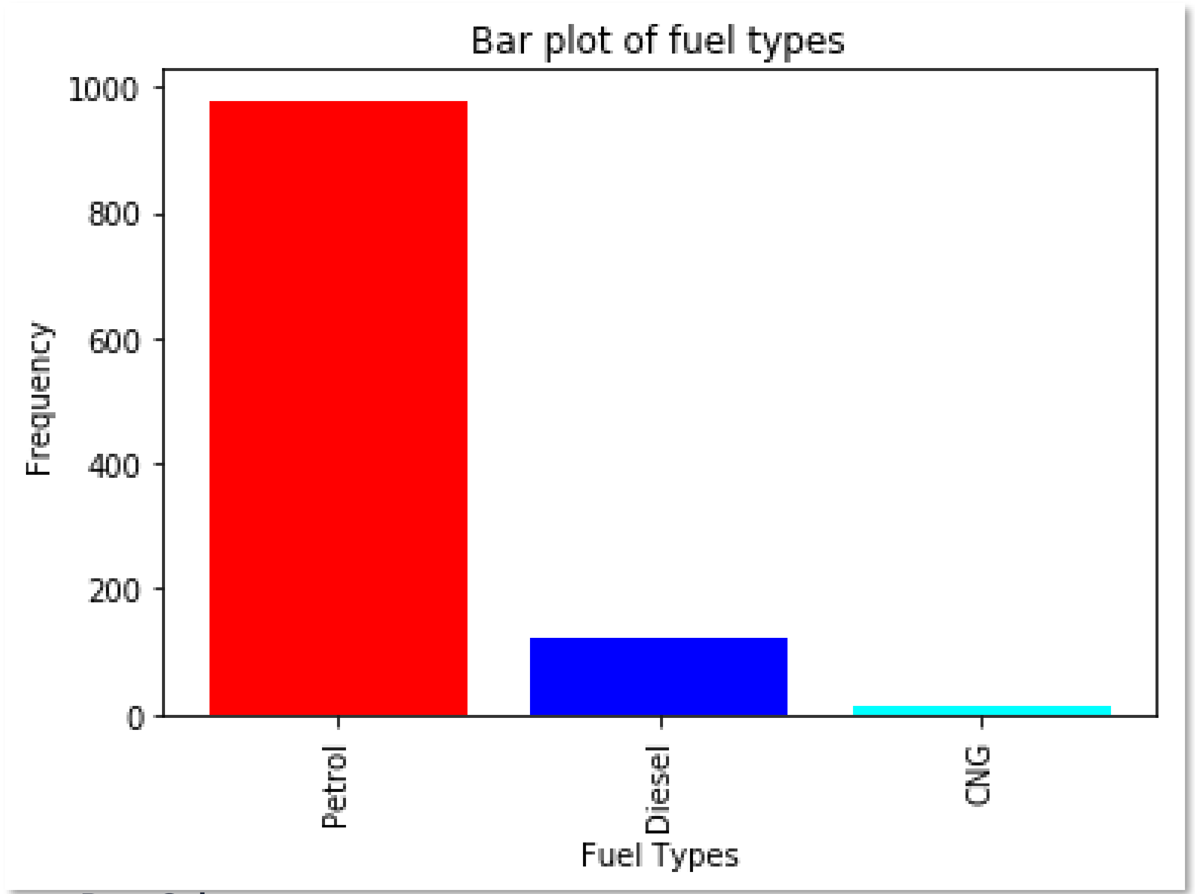
```
plt.bar(index, counts, color=['red', 'blue', 'cyan'])
plt.title('Bar plot of fuel types')
plt.xlabel('Fuel Types')
plt.ylabel('Frequency')
plt.xticks(index, fuelType,rotation = 90)
plt.show()
```

Set the labels of the xticks

Set the location of the xticks

Bar plot

- Bar plot of fuel type shows that most of the cars have petrol as fuel type



Summary

We have learnt how to create basic plots using *matplotlib* library

- Scatter plot
- Histogram
- Bar plot

THANK YOU