Project2 Pandas Script

Author: Yabei Zeng

First Step Loading Packages

```
In [11]:  import pandas as pd
          import seaborn as sns
          import matplotlib.pyplot as plt
```

Second Step loading data

```
In [12]:  data = pd.read_csv('cars.csv', sep=';')
```

```
In [13]:  #glimpse data
          data.head()
```

Out[13]:

	Car	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model	Origin
0	Chevrolet Chevelle Malibu	18.0	8	307.0	130.0	3504.0	12.0	70	US
1	Buick Skylark 320	15.0	8	350.0	165.0	3693.0	11.5	70	US
2	Plymouth Satellite	18.0	8	318.0	150.0	3436.0	11.0	70	US
3	AMC Rebel SST	16.0	8	304.0	150.0	3433.0	12.0	70	US
4	Ford Torino	17.0	8	302.0	140.0	3449.0	10.5	70	US

Third Step: Data Summary (with mean, median, std, quantiles, min, max for all columns)

```
In [14]:  main_sum = data.describe()
          print(main_sum)

                     MPG   Cylinders  Displacement  Horsepower       Weight  \
          count  406.000000  406.000000    406.000000  406.000000   406.000000
          mean    23.051232    5.475369    194.779557  103.529557  2979.413793
          std      8.401777    1.712160    104.922458   40.520659   847.004328
          min      0.000000    3.000000     68.000000    0.000000  1613.000000
          25%     17.000000    4.000000    105.000000   75.000000  2226.500000
          50%     22.350000    4.000000    151.000000   93.500000  2822.500000
          75%     29.000000    8.000000    302.000000  129.000000  3618.250000
          max     46.600000    8.000000    455.000000  230.000000  5140.000000

                 Acceleration       Model
          count    406.000000  406.000000
          mean      15.519704   75.921182
          std        2.803359    3.748737
          min        8.000000   70.000000
          25%       13.700000   73.000000
          50%       15.500000   76.000000
          75%       17.175000   79.000000
          max       24.800000   82.000000
```

Fourth Step: Data Visualization

```
In [15]:  mean = data['Weight'].mean()
          median = data['Weight'].median()
          quantile25 = data['Weight'].quantile(.25)
          quantile75 = data['Weight'].quantile(.75)

          plot = sns.histplot(data["Weight"], kde=True, color="blue", label="Weight")

          plot.axvline(mean, color="y", linestyle='-', label='Mean')
          plot.axvline(median, color="g", linestyle='-', label='Median')
          plot.axvline(quantile25, color="r", linestyle='-', label='25% quantile')
          plot.axvline(quantile75, color="b", linestyle='-', label='75% quantile')
          # Create Legend
          plot.legend()

          plt.show()
```
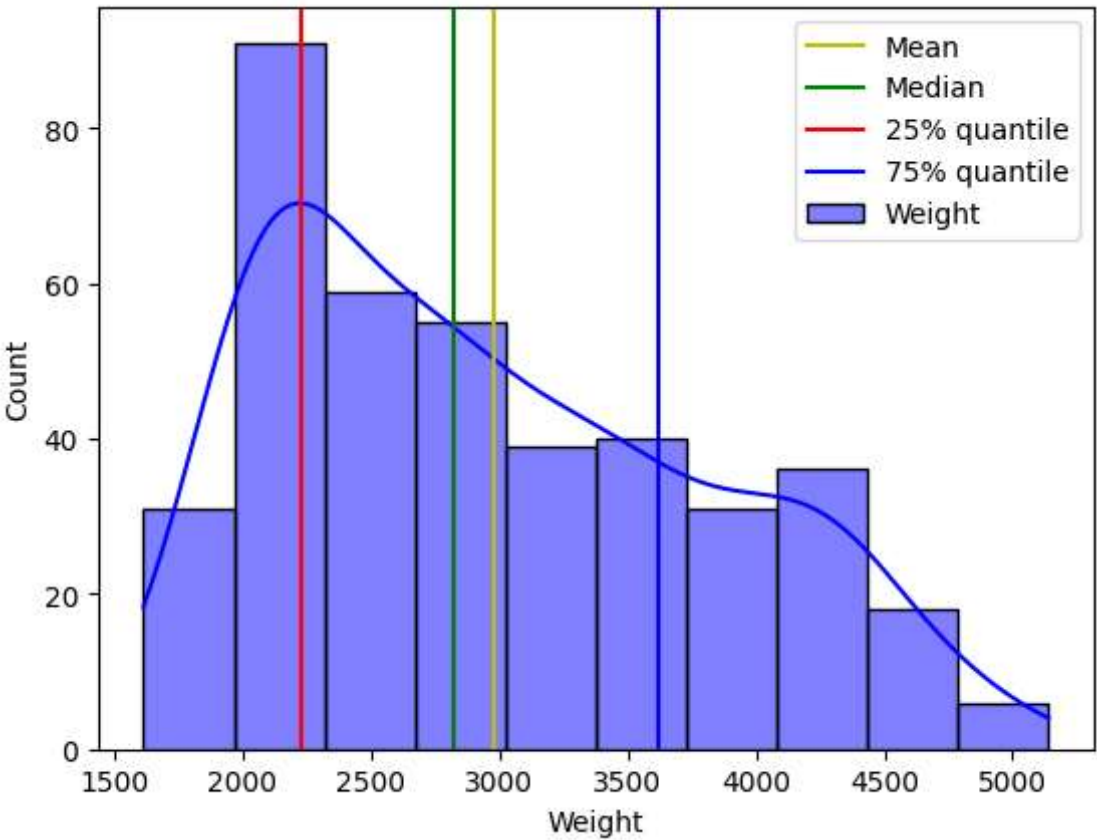
Some Conclusion

In this data summary and visualiztaion, we can tell that the weight(the column/variable we interested) is a right-skewed distribution.

This variable has mean around 3000, median around 2700, 25% quantile around 2400, and 75% quantile around 3600.

In [ ]: