IDS706 Project3

Author: Yabei Zeng

Loading Packages

In [1]:
```
import polars as pl
import seaborn as sns
import matplotlib.pyplot as plt
```

writing a function to load the dataset

In [2]:
```
def load_data(data_path):
    my_data = pl.read_csv(data_path, separator=";")
    return my_data
```

writing a function to return the data summary

In [3]:
```
def data_summary(data):
    main_sum = data.describe()
    return main_sum
```

writing a function for data visualization

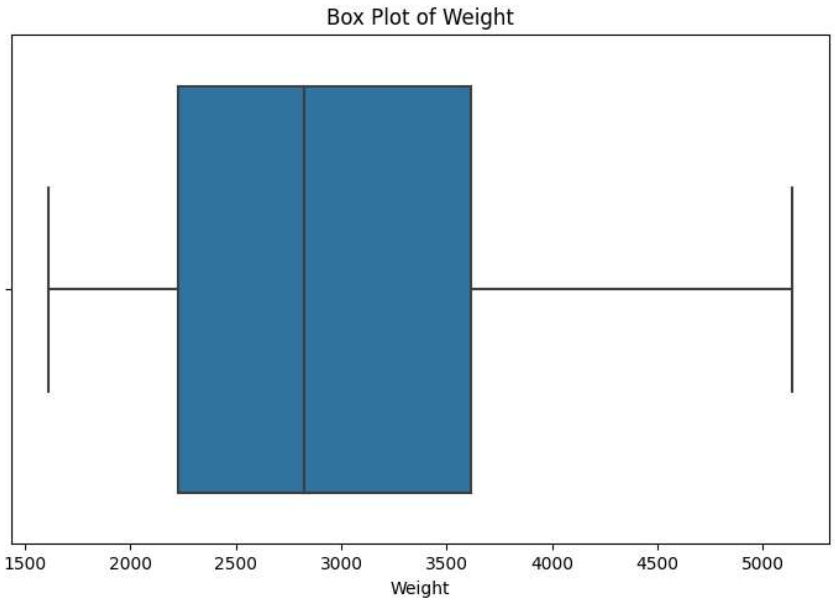
In [4]:
```
def data_visual(data):
    plt.figure(figsize=(15, 10))

    # Box Plot
    plt.subplot(2, 2, 1)
    sns.boxplot(x=data["Weight"])
    plt.title('Box Plot of Weight')

    # Violin Plot
    plt.subplot(2, 2, 2)
    sns.violinplot(x=data["Weight"])
    plt.title('Violin Plot of Weight')

    # CDF Plot
    plt.subplot(2, 2, 3)
    sns.ecdfplot(data=data, x="Weight")
    plt.title('CDF Plot of Weight')

    # KDE Plot
    plt.subplot(2, 2, 4)
    sns.kdeplot(data=data, x="Weight", fill=True)
    plt.title('KDE Plot of Weight')

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

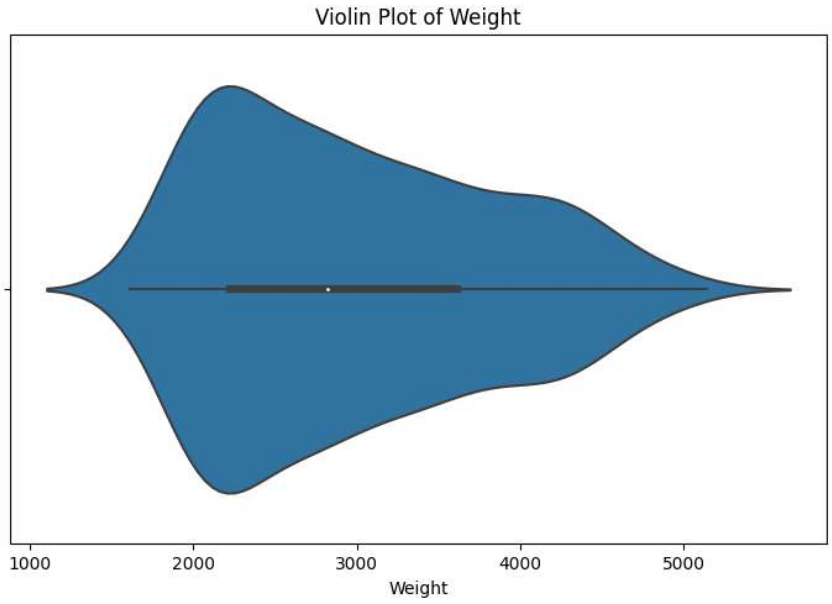
I used a dataset call car.csv, and the following is the results

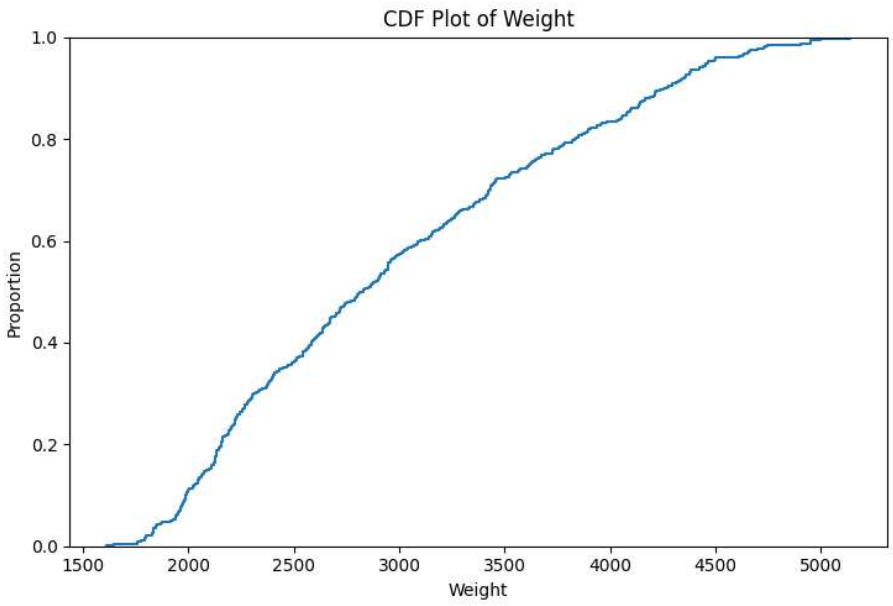
In [5]:
```
def main():
    my_df = load_data("cars.csv")
    print(data_summary(my_df))
    data_visual(my_df)


if __name__ == "__main__":
    main()
```

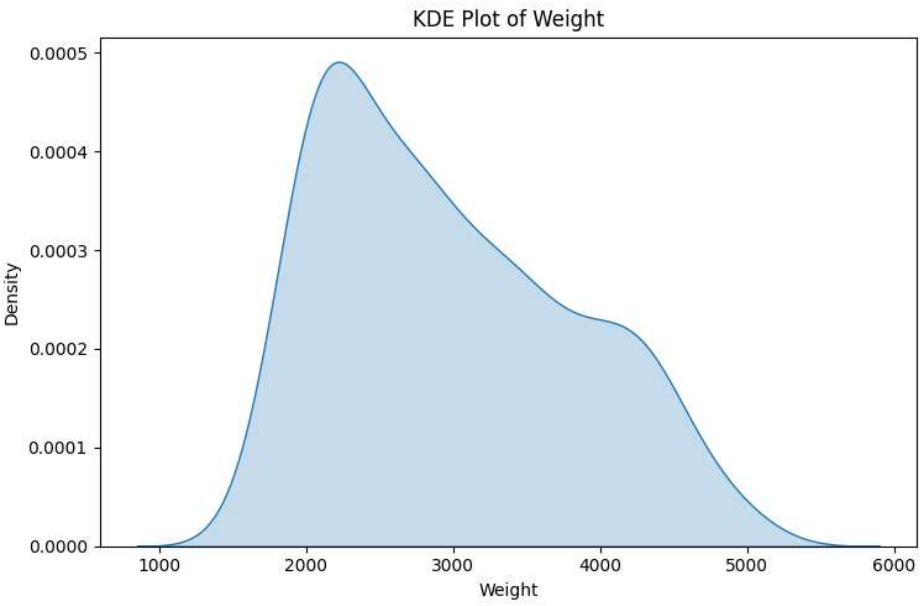
shape: (9, 10)

describe	Car	MPG	Cylinders	…	Weight	Acceleration	Model	Origin
str	str	f64	f64		f64	f64	f64	str
count	406	406.0	406.0	…	406.0	406.0	406.0	406
null_count	0	0.0	0.0	…	0.0	0.0	0.0	0
mean	null	23.051232	5.475369	…	2979.413793	15.519704	75.921182	null
std	null	8.401777	1.71216	…	847.004328	2.803359	3.748737	null
min	AMC Ambassador Brougham	0.0	3.0	…	1613.0	8.0	70.0	Europe
25%	null	17.0	4.0	…	2226.0	13.7	73.0	null
50%	null	22.4	4.0	…	2830.0	15.5	76.0	null
75%	null	29.0	8.0	…	3620.0	17.2	79.0	null
max	Volvo Diesel	46.6	8.0	…	5140.0	24.8	82.0	US









conclusion

based on the summary output, we can tell each variable's count, mean, median, standard deviation, minimum value, maximum value, 25% quantile, 75% quantile
based on the box and kde plot, the weight variable is right-skewed and there is no outlier showing in the plot.