Metrics:

1. **Mean Squared Error**
2. **Root Mean Squared Error**

MSE and RMSE were chosen because they penalise large errors, whereas MAE does not.

1. Removing outliers and duplicate data

MSE:

	MODEL	MSE
6	RF Reg	11.509174
1	Polynomial Reg	13.029404
7	GB Reg	13.231229
2	Ridge Reg	14.764694
0	Linear Reg	14.798267
3	Lasso Reg	15.066875
4	SVR	15.068401
5	DT Reg	21.331237

RMSE:

	MODEL	RMSE
6	RF Reg	3.392517
1	Polynomial Reg	3.609627
7	GB Reg	3.637476
2	Ridge Reg	3.842485
0	Linear Reg	3.846852
3	Lasso Reg	3.881607
4	SVR	3.881804
5	DT Reg	4.618575

- Random Forest Regressor has the least MSE and RMSE values

Actual vs Predicted values for Random Forest Regressor:

	Real Values	Predicted Values
377	31.0	35.101000
99	18.0	16.691000
217	30.0	24.927200
129	31.0	28.313000
74	13.0	12.685000
395	32.0	30.579000
52	30.0	23.367000
303	31.8	33.470567
221	17.5	15.310000
228	18.5	15.083000

2. Removing outliers, keeping duplicate data

MSE:

	MODEL	MSE
1	Polynomial Reg	11.719115
3	Lasso Reg	12.530900
7	GB Reg	13.033130
2	Ridge Reg	13.054529
0	Linear Reg	13.070301
6	RF Reg	13.580186
4	SVR	13.659578
5	DT Reg	16.135800

RMSE:

	MODEL	RMSE
1	Polynomial Reg	3.423319
3	Lasso Reg	3.539901
7	GB Reg	3.610143
2	Ridge Reg	3.613105
0	Linear Reg	3.615287
6	RF Reg	3.685130
4	SVR	3.695887
5	DT Reg	4.016939

- Polynomial Regression has the least MSE and RMSE values

Actual vs Predicted values for Polynomial Regression

	Real Values	Predicted Values
148	26.0	26.421344
389	22.0	27.424982
350	34.7	33.764390
52	30.0	27.119942
332	29.8	37.895533
377	31.0	39.504716
150	26.0	23.999457
216	31.5	31.530536
328	30.0	26.026031
395	32.0	33.580601

3. Removing duplicate data, keeping outliers

MSE:

	MODEL	MSE
6	RF Reg	12.155437
1	Polynomial Reg	12.483416
7	GB Reg	14.268807
2	Ridge Reg	14.971310
0	Linear Reg	15.057292
3	Lasso Reg	15.237866
4	SVR	15.346475
5	DT Reg	15.370010

RMSE:

	MODEL	RMSE
6	RF Reg	3.486465
1	Polynomial Reg	3.533188
7	GB Reg	3.777407
2	Ridge Reg	3.869278
0	Linear Reg	3.880373
3	Lasso Reg	3.903571
4	SVR	3.917458
5	DT Reg	3.920460

- Random Forest Regressor has the least MSE and RMSE values

Actual vs Predicted values for Random Forest Regressor

	Real Values	Predicted Values
377	31.0	34.755996
99	18.0	17.888800
217	30.0	24.033891
129	31.0	28.126577
74	13.0	12.724405
395	32.0	30.204226
52	30.0	23.683493
303	31.8	34.051446
221	17.5	15.314475
228	18.5	17.853182

4. Keeping outliers and duplicate data

MSE:

	MODEL	MSE
1	Polynomial Reg	12.407422
3	Lasso Reg	12.912579
7	GB Reg	13.040676
2	Ridge Reg	13.400830
0	Linear Reg	13.416292
4	SVR	14.454738
6	RF Reg	14.991425
5	DT Reg	19.009662

RMSE:

	MODEL	RMSE
1	Polynomial Reg	3.522417
3	Lasso Reg	3.593408
7	GB Reg	3.611188
2	Ridge Reg	3.660714
0	Linear Reg	3.662826
4	SVR	3.801939
6	RF Reg	3.871876
5	DT Reg	4.360007

- Polynomial Regression has the least MSE and RMSE values

Actual vs Predicted values for Polynomial Regression

	Real Values	Predicted Values
148	26.0	26.228459
389	22.0	27.097509
350	34.7	33.585260
52	30.0	27.256115
332	29.8	37.827724
377	31.0	39.687234
150	26.0	24.026982
216	31.5	31.544929
328	30.0	26.986198
395	32.0	34.049650

Conclusion:

- We can see an improvement once outliers and duplicate data are removed.
    - The outliers in horsepower, which are around 200, are smaller in number. So they were removed.

    - On checking for outliers again, values around 130 were spotted as outliers. But they are not small in number, as compared to the initial set of outliers.
    - Even though there were only 4 duplicate rows, it shows that removing them improves performance.
- These outliers can skew the model and the duplicate data can create a bias while training the model.