Screenshots illustrating the regression analysis process carried out

The first stage of the development of each regression model was the data preprocessing, where I checked for duplicate records and missing values, encoded data numerically, located outliers and chose whether to impute or delete appropriately.

Locating outliers:

```
Checking for outliers using modified Z-scores for column Mileage in km...
Outlier. Value: 420550     Modified Z-Score: 3.60758624075992 Row: 29
Outlier. Value: 719847     Modified Z-Score: 7.21212551333786 Row: 65
Outlier. Value: 433811     Modified Z-Score: 3.76729313823519 Row: 90
Outlier. Value: 777777     Modified Z-Score: 7.90979692354391 Row: 167
Outlier. Value: 573249     Modified Z-Score: 5.44659412384387 Row: 247
```

Data checks and modifications:

```
No null/missing values.
3512 Duplicate records removed.
Engine volume column has been split.
Mileage column has been modified.
Category column has been modified.
Color column has been modified.
Drive wheels column has been modified.
Fuel type column has been modified.
```

The next stage was feature selection and involved data visualisation and rankings of features based on their correlation coefficients.

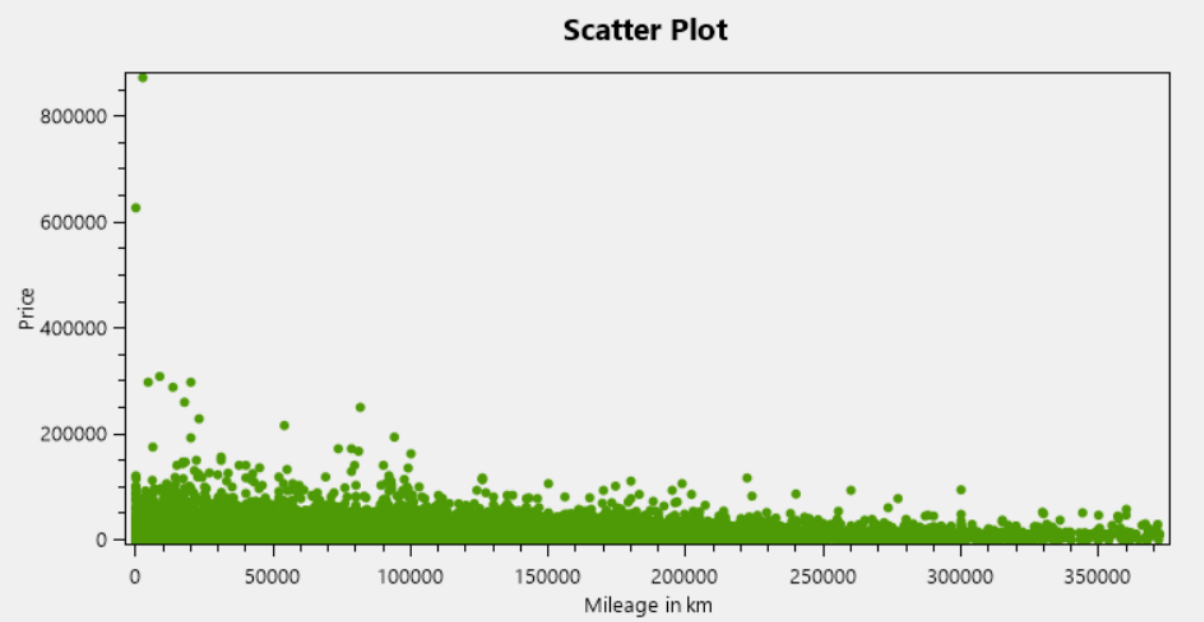
Continuous predictor ranking:

```
Predictors ranked by their pearson correlation coefficient:
1. Prod. year scaled    Coeff: 0.3314583847220 46
2. Mileage in km scaled Coeff: -0.214210566926586
3. Engine displacement scaled Coeff: 0.197196765004989
4. Cylinders scaled     Coeff: 0.139493151955371
5. Levy scaled          Coeff: 0.0902454303900141
6. Airbags scaled       Coeff: 0.0167298529587539
```

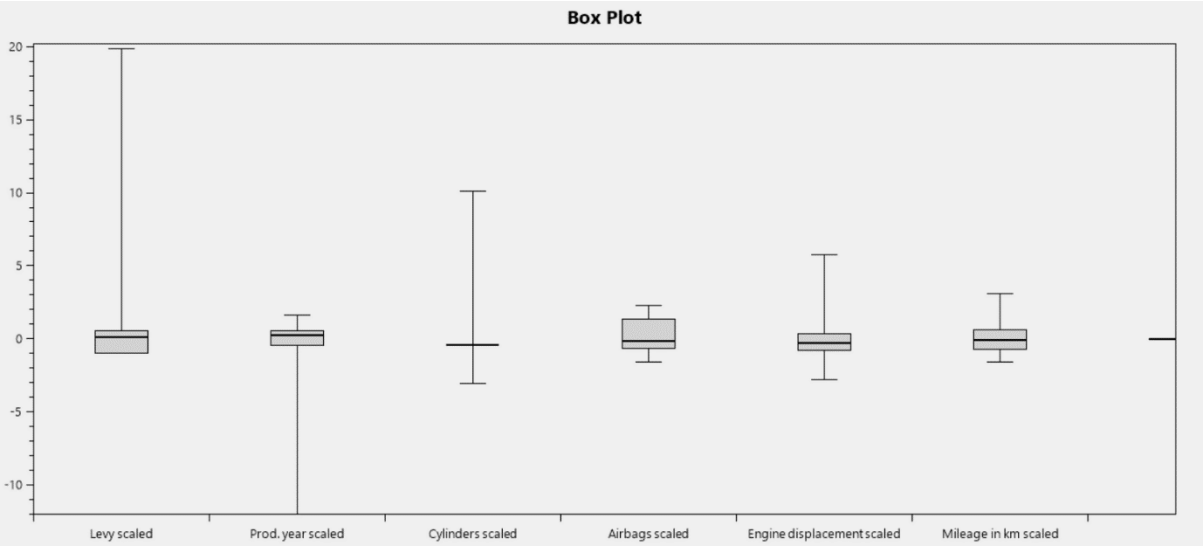
Binary predictor ranking:

```
Predictors ranked by their point biserial correlation coefficient:
1. Is jeep              Coeff: 0.254824307595062
2. Is diesel            Coeff: 0.210156459601851
3. Is leather interior  Coeff: 0.193324998176985
4. Is engine turbo      Coeff: 0.182780027219332
5. Is tiptronic         Coeff: 0.181082901345874
6. Is wheel left        Coeff: 0.162916124772152
7. Is hatchback         Coeff: -0.153628933432364
8. Is manual            Coeff: -0.128743120441147
9. Is sedan             Coeff: -0.125248148990822
10. Is hybrid            Coeff: -0.104790302355864
```

Scatter plot:



Box plots:



The next part of the regression analysis process was the training of the regression model. After training, each model was validated in order to view how well the model performs; the model is validated many times with different features in order to determine the optimal combination of features. I have shown below the result of this process for some of the regression models.

Multi linear regression model using gradient descent:

Table showing Multi Linear Gradient Descent regression model performance for a number of features:

Feature no.	Added feature	MAE	RMSE	R-Squared	Adj R-Squared	Epochs
1	Prod. year scaled	11050.94936	19318.38471	0.102078	0.101754	50
2	Mileage in km scaled	11123.19414	19154.361158	0.117876	0.117239	55
3	Engine displacement scaled	10736.716655	18564.623044	0.171868	0.17097	70
4	Is jeep	10412.130881	18279.770501	0.198022	0.196862	250
5	Is diesel	10184.938905	18009.466295	0.22264	0.221234	430
6	Is tiptronic	10290.087654	17713.973004	0.248435	0.246803	442
7	Is leather interior	10283.346624	17712.128003	0.248568	0.246664	484
8	Is engine turbo	10297.577634	17534.330573	0.264077	0.261945	525
9	Cylinders scaled	10297.430244	17537.042283	0.263773	0.261374	519

Reduction in error decreased so 'Is Tiptronic' is the last useful feature as features after cause only small reductions in error, meaning they will likely hinder the model and lead to overfitting.

KNN regression model:

Table showing KNN regression model performance for a number of features:

No. of features	Added feature	MAE	RMSE	R-Squared	Adj R-Squared
1	Prod. year scaled	11291.188479	19900.413586	0.06672	0.066381
2	Mileage in km scaled	11027.371609	19818.867862	0.072982	0.072308
3	Engine displacement scaled	9519.128551	17700.831473	0.260826	0.26002
4	Is jeep	9180.313353	17415.780656	0.287678	0.286642
5	Is diesel	8765.921052	16986.519285	0.323428	0.322197
6	Is tiptronic	7953.922113	15288.401675	0.452728	0.451533
7	Is leather interior	7675.484665	15132.567964	0.46412	0.462754
8	Is engine turbo	7520.801537	15051.208001	0.470417	0.468873
9	Cylinders scaled	7387.435534	14912.702366	0.480845	0.479142

Clearly shown by the Adjusted R-Squared, 'Is tiptronic' is the last feature to have a large positive impact on the model.

Fast tree regression model:

Table showing Fast tree regression trainer performance for a number of features:

Data	No. of features	Added feature	MAE	RMSE	R-Sqaured	Adj R-squared
Validation	1	ProdYear	10710.35758	19032.027534	0.143991	0.143679
Training			10676.724523	19794.381769	0.138913	0.138835
Validation	2	MileageInKm	10586.797108	19245.991968	0.122687	0.122048
Training			9966.637881	18363.671838	0.258891	0.258756
Validation	3	EngineDisplacement	9362.174997	17730.215065	0.253692	0.252877
Training			8283.922199	13345.75035	0.608575	0.608468
Validation	4	ManufacturerEncoded	8579.876617	16928.170038	0.322649	0.321662
Training			7489.268075	12696.001506	0.645761	0.645632
Validation	5	ModelEncoded	8027.426054	16362.578031	0.368759	0.367608
Training			6966.739544	12391.616201	0.662543	0.662389
Validation	6	IsJeep	8051.292038	16502.370522	0.35451	0.353098
Training			6905.54894	12101.246457	0.678172	0.677997
Validation	7	IsDiesel	8010.555561	16160.206998	0.383322	0.381748
Training			6848.586545	11748.604591	0.696656	0.696463
Validation	8	IsTiptronic	7182.979216	13957.617218	0.53971	0.538366
Training			6238.188271	10508.572866	0.757311	0.757134
Validation	9	IsLeatherInterior	7019.334852	13822.877299	0.548521	0.547038
Training			6081.288689	10378.291527	0.763291	0.763097
Validation	10	IsEngineTurbo	6826.868134	13477.450167	0.571192	0.569626
Training			5907.03789	9896.642544	0.784752	0.784556
Validation	11	IsWheelLeft	6800.120997	13507.815848	0.568699	0.566966
Training			5956.52396	10129.057993	0.774524	0.774298

The model clearly performs better on the training data than the validation data which indicates there is some overfitting (which is to be

Clearly, adding features after 'IsEngineTurbo' is detrimental to the model's performance.

After determining the optimal feature combination for the model, the model can be fine-tuned by adjusting significant hyperparameters. I have shown two examples of this below.

Gradient descent algorithm: (Multi linear regression model)

```
Fine-tuning the learning rate hyperparameter:
Learning rate      Number of epochs
1                  Not converged
0.75               Not converged
0.5                98
0.25               189
0.1                444
```

A learning rate of 0.5 is optimal because it leads to convergence in only 98 epochs, meaning the

KNN regression model:

```
Fine-tuning the k value hyperparameter:
K value    MAE                RMSE               R-Squared          Adj R-Squared
5          8056.071924        15336.452826       0.447814           0.446003
10         7953.922113        15288.401675       0.452728           0.450933
20         7994.922738        15632.817457       0.429304           0.427432
30         8058.8553          15846.168037       0.413382           0.411458
40         8121.079134        16014.779398       0.400899           0.398934
50         8181.749344        16114.332668       0.39349            0.391501
```

When k=10, the model performs best with least error.

After all the models had been fine-tuned, the summaries of each were written to the regression-analysis-conclusion.txt file at run-time, in order to help conclude which model is the best and the one to be integrated into the final system.

Here is the contents:

```
Mean template model:
Errors: MAE = 14258.16907, RMSE = 105403.00715, R-Squared = -0.007762
Hyperparameters: None
Features: None

Simple linear regression model:
Errors: MAE = 11021.46628, RMSE = 19091.115517, R-Squared = 0.11079
Hyperparameters: Z-score threshold for Prod. Year outliers = 4
Features: Prod. year scaled

Multi linear regression model using gradient descent:
Errors: MAE = 10290.016309, RMSE = 17713.968782, R-Squared = 0.248435, Adjusted R-Squared = 0.246804
Hyperparameters: Learning rate = 0.5, Max no. of epochs = 1000, Convergence threshold = 0.1
Features: Prod. year scaled, Mileage in km scaled, Engine displacement scaled, Is jeep, Is diesel, Is
tiptronic

K-nearest neighbours regression model:
Errors: MAE = 7953.922113, RMSE = 15288.401675, R-Squared = 0.452728, Adjusted R-Squared = 0.450933
Hyperparameters: K = 10
Features: Prod. year scaled, Mileage in km scaled, Engine displacement scaled, Is jeep, Is diesel, Is
tiptronic

Microsft.ML Fast tree regression trainer:
Errors: MAE = 6803.006273, RMSE = 13244.066399, R-Squared = 0.586265, Adjusted R-Squared = 0.584755
Hyperparameters: Minimum sample count per leaf = 8 ***Inconclusive value - check with holdout data***
Features: ProdYear, MileageInKm, EngineDisplacement, ManufacturerEncoded, ModelEncoded, IsJeep, IsDiesel,
IsTiptronic, IsLeatherInterior, IsEngineTurbo
```

Lastly, the models are tested with holdout data to gauge whether the model is able to generalise and predict new, unseen data. Below is a table that shows the model performance with the holdout data.

```
Final model testing with holdout data:
Mean teplate model      MAE: 12544.704653    RMSE: 17837.458163    R-Squared: -0.010948       Adj R-Squared: -0.010948
Simple linear model     MAE: 11406.393216    RMSE: 18626.113938    R-Squared: 0.114816        Adj R-Squared: 0.114237
Multi linear model      MAE: 10810.404292    RMSE: 17778.626318    R-Squared: 0.255535        Adj R-Squared: 0.252621
KNN model               MAE: 8470.599935     RMSE: 15294.411082    R-Squared: 0.403165        Adj R-Squared: 0.400813
Fast tree model         MAE: 6708.727058     RMSE: 12119.236952    R-Squared: 0.634855        Adj R-Squared: 0.632446
```

Clearly, the fast tree regression model performs the best, so this was the model I integrated into the final car price estimator system.