

# Predicting Prices of Used Cars Using R and JMP

# Table of Contents

# STUDENT

02.

Abstract

03.

EDA

04.

Data Preprocessing

05.

Prediction Profiler

06.

Stepwise platform

07.

Multiple Linear Regression

08.

k-NN

09.

Partition Model (R)

10.

Partition Model (JMP)

11.

Conclusion



# ABSTRACT

---

This study utilizes data from ToyotaCorolla.jmp, comprising 1,436 records of used Toyota Corolla cars sold in the Netherlands in 2004. The objective is to predict car prices based on specifications such as Age, Kilometers, Horsepower, and Fuel Type. Three models—Multiple Linear Regression, k-Nearest Neighbors (k-NN), and Regression Tree—were developed and evaluated using training, validation, and test datasets.

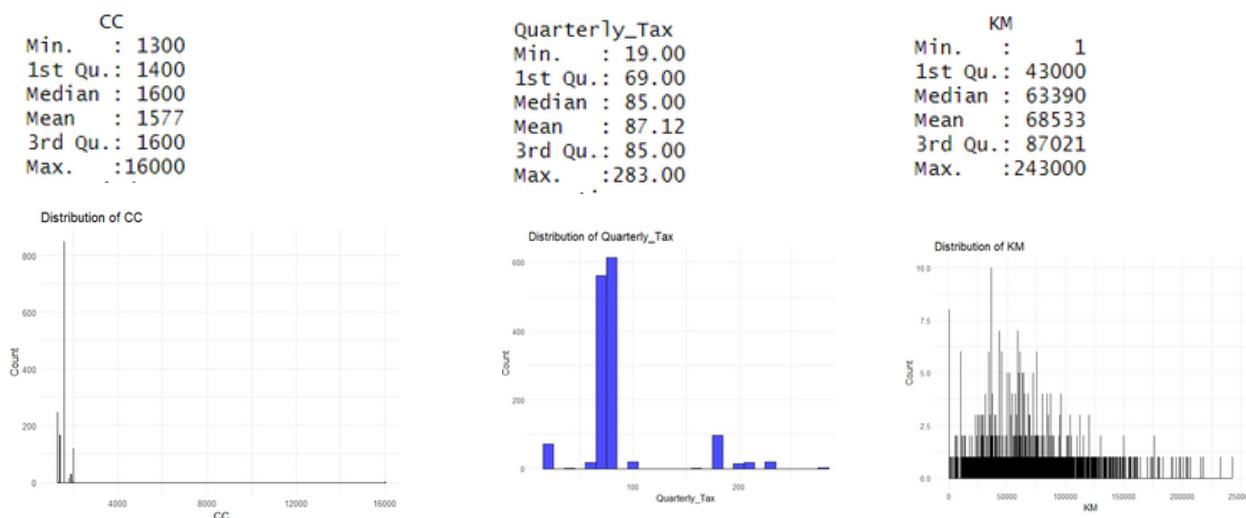
Data preprocessing included splitting into subsets and normalizing variables as needed. Models were constructed and assessed using R for statistical programming and JMP for interactive analysis. Key predictors of price were identified through stepwise regression and feature importance analysis. Model performance was compared using metrics such as RMSE and R-Square to determine the most effective approach.

The results highlight the strengths and limitations of each method, with the reduced regression model offering simplicity and accuracy, while k-NN and regression trees provided complementary insights. This analysis demonstrates the utility of combining R and JMP for predictive modeling and decision-making.

# EDA

In addition to what was asked for in the instructions, we performed some basic EDA of the dataset. We did this for two reasons. Firstly, we want to understand the dataset better including distributions, how variables relate to the target, and more. Secondly, it is important to ensure we have a complete dataset meaning everything is cleaned well and that outliers are accounted for.

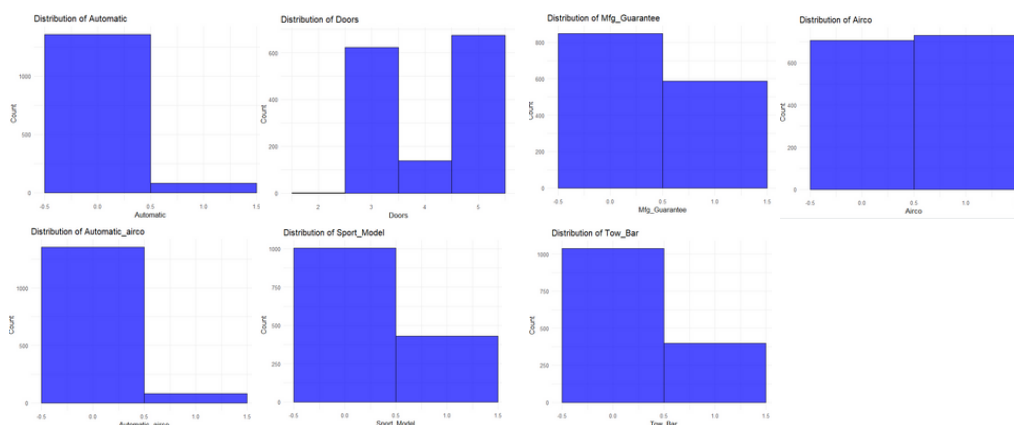
Below are some variables that seem to have odd distributions and some obvious outliers. I will explore this further and possibly take them out if it is a mistake.



CC has one outlier that we will change to the mean. Quarterly Tax has a few outliers I explored but won't change. KM has some but we decided not to change any. Below HP has an outlier we explored and were all the same car, so we won't touch them.. Guarantee Period we will not touch even though it doesn't make much sense.



The three Price outliers were the same car, different ages, and had the highest tax. We won't touch them because it's not obvious they are incorrect. If we were in contact with the owner of this dataset, we would ask about it.



Above are histograms of the remaining 9 features we are taking into account. There aren't any outliers.

# DATA PREPROCESSING

How we split the data

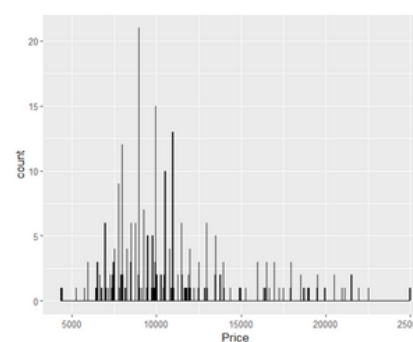
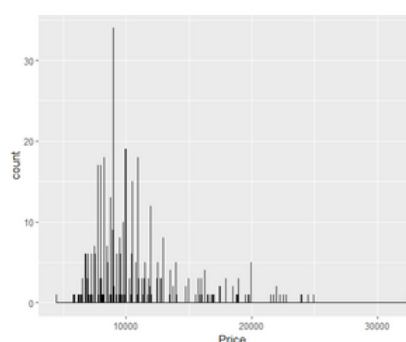
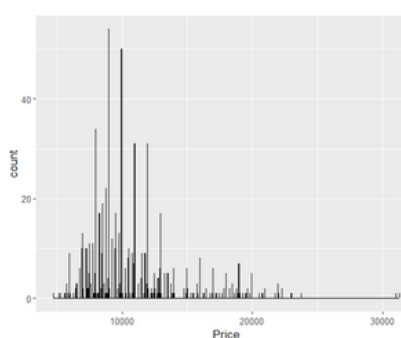
```
> cat("Training set size:", nrow(train_data), "\n")
Training set size: 718
> cat("Validation set size:", nrow(valid_data), "\n")
Validation set size: 430
> cat("Test set size:", nrow(test_data), "\n")
Test set size: 288
```

Training Set	0.5
Validation Set	0.3
Test Set	0.2

**Training Set**

**Validation Set**

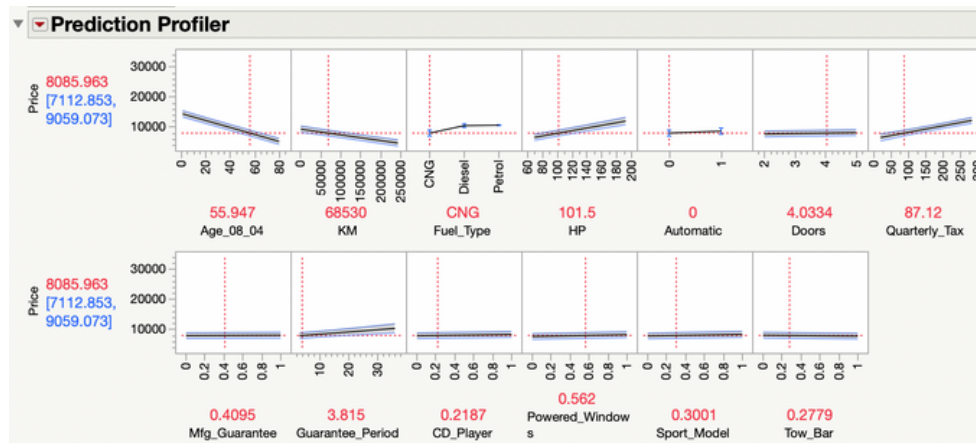
**Test Set**



The splits were done pretty well. All the distributions are consistent which is what you want for accurate results. It is noticeable that there is a right tail. For many models it assumes a normal distribution for the target variable. Meaning if we can make the target normal the more accurate model we can create. A solution to this is using a log or box-cox to transform the target variable.

# PREDICTION PROFILER

What appear to be the three or four most important car specifications for predicting the car's price? Use the Prediction Profiler to explore the relationship between Price and these variables.



## Our Discussion:

The variables that appear to have a large impact in determining price are Age, Guarantee Period, Quarterly Tax, HP, Automatic airco and KM according to the prediction profiler. The correlation matrices backs up all those variables, but also shows high correlation among others like airco, but airco is not relevant in the prediction profiler.

We would conclude the 4 most relevant ones are Age, HP, Quarterly Tax, and KM.

## Multicollinearity:

There is a lot of multicollinearity which affects how well we can understand the significance of each feature. These are the features with multicollinearity.

Fuel type and HP. 0.4

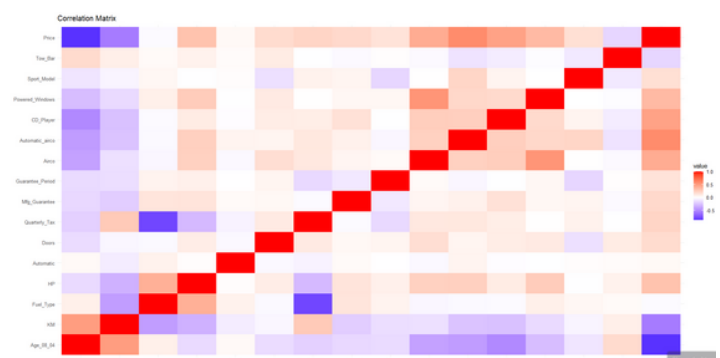
Quarter tax and fuel type -0.79

Age and KM 0.5

air co and powered windows 0.54

cd player and age -0.51

We will take out 4 of these and compare the model with all 15 features, fuel type, cd player, KM, Powered windows

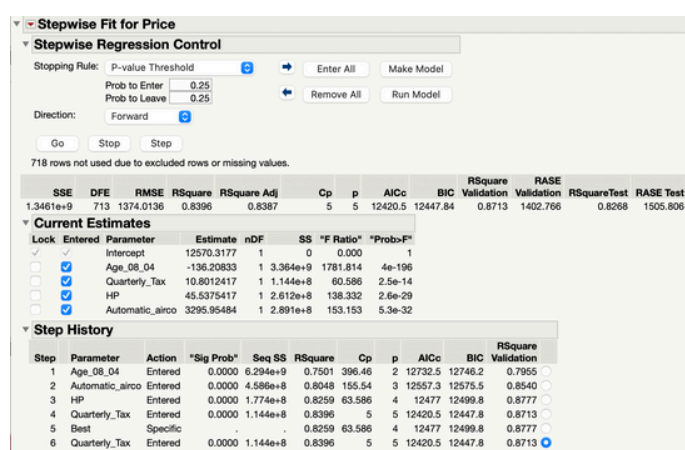


# STEPWISE PLATFORM:

Use the Stepwise platform to develop a reduced predictive model for Price. What are the predictors in your reduced model?

## Using JMP

We utilize the Stepwise Regression platform in JMP to develop a reduced predictive model for Price by selecting the 4 predictors based on the statistical significance at p-value threshold of .25.



**Stepwise Fit for Price**

Stopping Rule: P-value Threshold 0.25

Direction: Forward

718 rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RASE	RSquareTest	RASE Test
1.3461e+9	713	1374.0136	0.8396	0.8387	5	5	12420.5	12447.84	0.8713	1402.766	0.8268	1505.806

**Current Estimates**

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	12570.3177	1	0	0.000	1
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Age_08_04	-136.20833	1	3.364e+9	1781.814	4e-196
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Quarterly_Tax	10.8012417	1	1.144e+8	60.586	2.5e-14
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	HP	45.5375417	1	2.612e+8	138.332	2.6e-29
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Automatic_airco	3295.95484	1	2.891e+8	153.153	5.3e-32

**Step History**

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	RSquare Validation
1	Age_08_04	Entered	0.0000	6.294e+9	0.7501	396.46	2	12732.5	12746.2	0.7955
2	Automatic_airco	Entered	0.0000	4.586e+8	0.8048	155.54	3	12557.3	12575.5	0.8540
3	HP	Entered	0.0000	1.774e+8	0.8259	63.586	4	12477	12499.8	0.8777
4	Quarterly_Tax	Entered	0.0000	1.144e+8	0.8396	5	5	12420.5	12447.8	0.8713
5	Best	Specific	.	.	0.8259	63.586	4	12477	12499.8	0.8777
6	Quarterly_Tax	Entered	0.0000	1.144e+8	0.8396	5	5	12420.5	12447.8	0.8713

## Using R

Using the stepAIC model from library MASS in R gives us a nice summary of the importance of each feature on predicting the target, Price.

```
Residuals:
    Min       1Q   Median       3Q      Max
-4228.5  -759.7   -36.7    668.8   7460.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.831e+03  6.184e+02  15.897  < 2e-16 ***
Age_08_04    -1.112e+02  2.757e+00  -40.329  < 2e-16 ***
KM           -1.714e-02  1.209e-03  -14.179  < 2e-16 ***
Fuel_Type     6.947e+02  1.644e+02   4.226  2.53e-05 ***
HP            2.882e+01  2.587e+00   11.139  < 2e-16 ***
Automatic     5.153e+02  1.420e+02   3.629  0.000295 ***
Doors         1.573e+02  3.548e+01   4.433  1.00e-05 ***
Quarterly_Tax 1.968e+01  1.409e+00  13.967  < 2e-16 ***
Mfg_Guarantee 1.700e+02  6.877e+01   2.472  0.013551 *
Guarantee_Period 8.545e+01  1.154e+01   7.406  2.22e-13 ***
Airco         1.705e+02  8.274e+01   2.060  0.039555 *
Automatic_airco 3.189e+03  1.638e+02  19.465  < 2e-16 ***
CD_Player     2.747e+02  9.169e+01   2.996  0.002780 **
Powered_windows 4.390e+02  7.902e+01   5.555  3.31e-08 ***
Sport_Model   3.389e+02  7.516e+01   4.509  7.04e-06 ***
Tow_Bar       -2.579e+02  7.475e+01  -3.450  0.000577 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1221 on 1420 degrees of freedom
Multiple R-squared:  0.8879,    Adjusted R-squared:  0.8867
F-statistic: 749.8 on 15 and 1420 DF,  p-value: < 2.2e-16
```

## Our Discussion:

The reduced model includes four predictors: Age, Quarterly Tax, HP, and Automatic Airco, all of which were selected for their significant contribution to explaining variations in Price. The model achieves an adjusted Rsquared value of .8387, indicating a strong explanatory power. Each predictor has a highly significant p-value (less than .05), confirming their importance in the model. Additionally, metrics like AICc (124290.5) and BIC (12447.84) suggest a well optimized model. Validation metric (Rsquared Validation = .8713) further support the robustness of the model.

# MULTIPLE LINEAR REGRESSION:

## Using JMP

**Stepwise Fit for Price**

**Stepwise Regression Control**

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
2.1635e+9	1420	1234.3332	0.8854	0.8842	16	16	24537.25	24626.4

**Current Estimates**

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	12147.2205	1	0	0.000	1
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Age_08_04	-113.93333	1	2.625e+9	1722.845	3e-247
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Automatic_Airco	2990.38799	1	4.911e+8	322.321	4.1e-65
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Airco	130.821058	1	3723589	2.444	0.1182
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Mfg_Guarantee	145.728518	1	6706500	4.402	0.03608
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Quarterly_Tax	14.1482615	1	1.233e+8	80.909	7.4e-19
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Tow_Bar	-225.3835	1	13543966	8.890	0.00292
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Sport_Model	295.617115	1	23294154	15.289	0.0001
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Powered_Windows	460.604442	1	50569952	33.192	1.02e-8
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Doors	173.336991	1	35577662	23.351	1.5e-6
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Automatic[0-1]	-281.88087	1	23404194	15.361	9.3e-5
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	HP	35.5596264	1	2.29e+8	150.307	6.4e-33
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	CD_Player	247.972883	1	10923251	7.169	0.0075
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Fuel_Type[CNG&Petrol-Diesel]	-544.0018	2	51553168	16.918	5.48e-8
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Fuel_Type[CNG-Petrol]	-829.51508	1	34232239	22.468	2.35e-6
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	KM	-0.0183945	1	3.364e+8	220.819	1.6e-46

**Step History**

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Age_08_04	Entered	0.0000	1.45e+10	0.7684	1437.4	2	25519	25534.8
2	Automatic_Airco	Entered	0.0000	1.062e+9	0.8247	742.06	3	25121.2	25142.2
3	HP	Entered	0.0000	3.457e+8	0.8430	517.18	4	24964.8	24991.1
4	KM	Entered	0.0000	1.838e+8	0.8527	398.57	5	24874.9	24906.4
5	Quarterly_Tax	Entered	0.0000	3.659e+8	0.8721	160.42	6	24674.2	24711
6	Powered_Windows	Entered	0.0000	87714916	0.8768	104.85	7	24623.1	24665.2
7	Fuel_Type[CNG-Petrol]	Entered	0.0000	53903553	0.8796	73.471	9	24593.5	24646
8	Doors	Entered	0.0001	25085607	0.8810	59.006	10	24579.6	24637.4
9	Sport_Model	Entered	0.0001	24502890	0.8823	44.924	11	24565.9	24628.9
10	Automatic[0-1]	Entered	0.0001	22826830	0.8835	31.942	12	24553.1	24621.3
11	Tow_Bar	Entered	0.0034	13215367	0.8842	25.268	13	24546.5	24620
12	CD_Player	Entered	0.0037	12925978	0.8848	18.784	14	24540	24618.7
13	Mfg_Guarantee	Entered	0.0375	6611904	0.8852	16.444	15	24537.7	24621.6
14	Airco	Entered	0.1182	3723589	0.8854	16	16	24537.3	24626.4

We wanted to try the JMP stepwise regression with all 15 variables as well. Using the Stepwise Regression in JMP again, a reduced predictive model for Price was developed, selecting 15 significant predictors. The final model includes Age, Automatic Airco, Airco, Quakerly Tax, and KM, along with others. each predictor has a statistically significant contribution to the model (p-values < .05), ensuring their relevance in explaining price variability.

The model exhibits strong performance, with an adjusted Rsquared value of .8842, indicating that 88.42% of the variance is price is explained by the selected predictors. The AICc and BIC suggest a well-optimized. So far, this is the second best performing model.

## Using R

Below is the MLR with a non-transformed and with all 15 features.

```
RMSE: 1150.685
> cat("MAE:", mae, "\n")
MAE: 885.2287
> cat("R²:", r_squared, "\n")
R²: 0.9056852
```



Above the RMSE is 1150 MAE 885 and a R squared of 90%. This is relatively good acknowledging that the range of the target is 0 - 30,000. 90% of the values can be predicted essentially. However, when looking at the residuals, they show bias when they should be white noise. There is a large pattern that will be solved using a transformation.

```
RMSE: 1016.784
> cat("MAE:", mae, "\n")
MAE: 806.642
> cat("R²:", r_squared, "\n")
R²: 0.908915
```



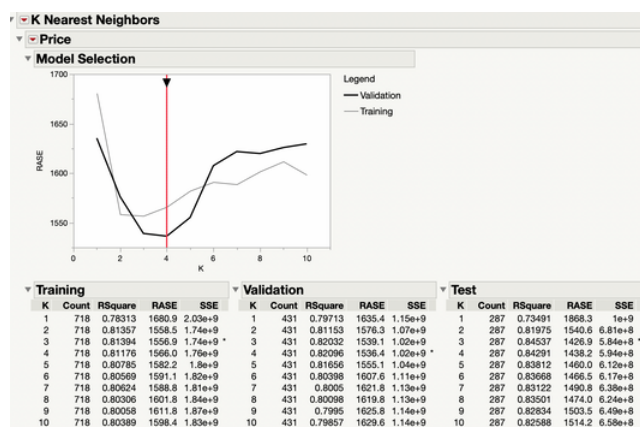
The best MLR was with all 15 variables with the box-cox transformation. We did better in all metrics. This is white noise.

We found a way to do even better in relation to our goal which is find what features predict the target. With only 4 variables we can predict the price with 87.6% accuracy. 2.5% worse than with all 15 but it matches the JMP model with the 4 variables. We used Age, HP, Quarterly Tax, and Automatic Airco. We will focus on these and confirm their importance in the following models.

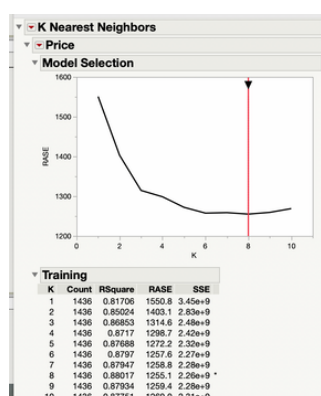
```
RMSE: 1314.221
> cat("MAE:", mae, "\n")
MAE: 975.8747
> cat("R²:", r_squared, "\n")
R²: 0.8769721
```



# K-NN



The K-NN model above using all 15 variables has its optimal K of 4 with a R-squared of 84.4% in the test set. This is a lot better than the K-NN model done in R with 15 variables.



Using the K-NN algorithm in JMP to predict Price, we identify the optimal value for K by analyzing the Root Average Square Error. From the model above which we used just 4 features the optimal K = 8.

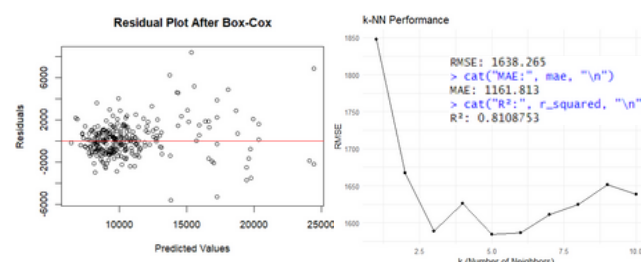
## Our Discussion:

The analysis demonstrates that K-NN is effective for price prediction, achieving a strong balance between accuracy and complexity at K = 8. The results highlight the importance of selecting the appropriate K value to optimize predictive performance. In the R, K-NN the optimal k is 5 although 3 is very similar.

Interestingly, in R the 15-feature K-NN model performed 5% worse than K-NN with 4 features. And the JMP K-NN model performed even better. With a 88% r-squared compared to 86%.

So far, the JMP K-NN is the second best performing model right behind the step-wise regression with 15 features.

Below is using the KNN model to predict Price with the original 15 features. Again, we normalized the data. The chosen K was 5. We also tried 3 however because it is close.



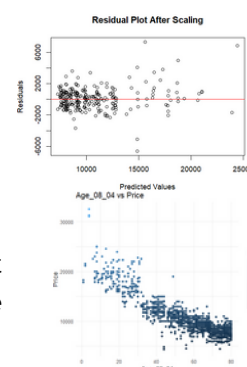
First thing you notice is the circle in the residuals which is really interesting. From research, this pattern can appear from having non-linear dependencies or our features aren't scaled well.

With the 4 features, we get better results than using all 15. We tried two scalers and the center scaler worked the best.

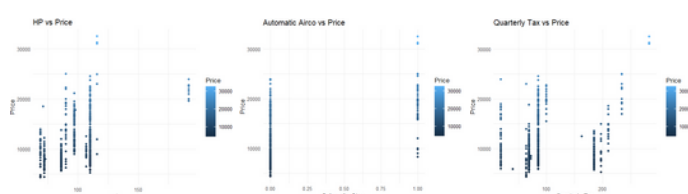
```

RMSE: 1401.045
> cat("MAE:", mae, "\n")
MAE: 971.0127
> cat("R^2:", r_squared, "\n")
R^2: 0.8616801

```

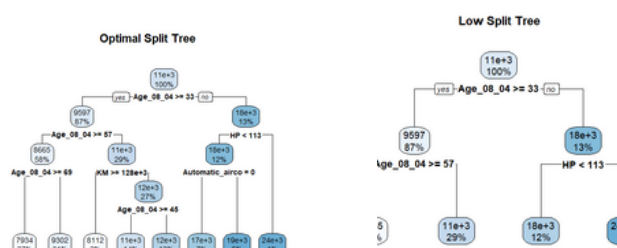


To understand the relation of these 4 important variables to price, we made scatter plots of the data.

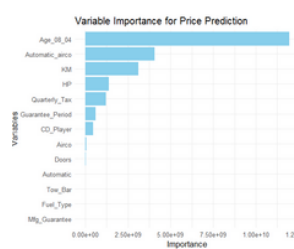


# PARTITION MODEL RSTUDIO

In the first two trees, we used all 15 variables. The first tree is just with 2 splits. We wanted to see how the metrics changed in R based off the splits. With too few splits the accuracy was bad. With an optimal amount of splits the accuracy improved. With too many it will be overfitted.

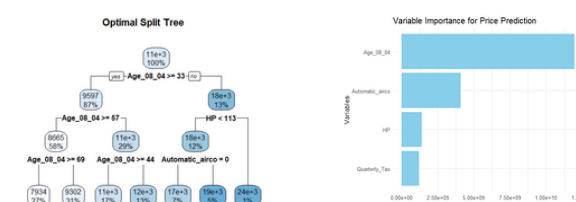


RMSE for Low Split Tree: 0.8071846



RMSE for Optimal Split Tree: 1517.245  
R-squared for Optimal Split Tree: 0.8377848

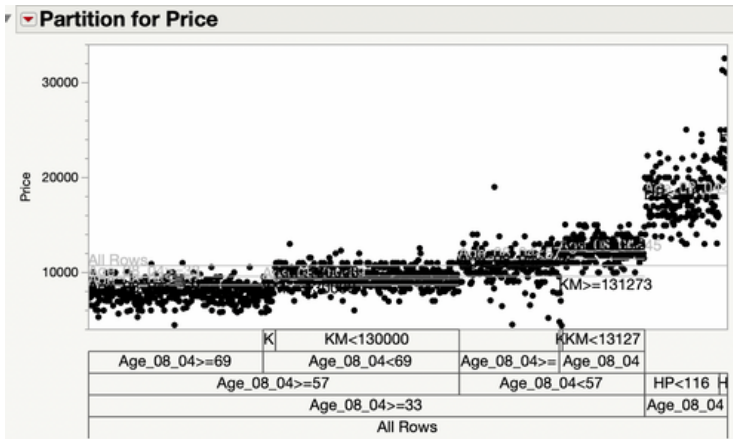
The first 2 graphs just show the logic of the splits. The bottom left graph shows Predicted vs Actual. The bottom right shows the ranked importance of all the features we put into the model. It's interesting the top 5 features it realized are important are included in the 4 features we determined earlier are the most important. The model only uses those features out of all the ones it could choose from.



R-squared: 0.8373166

This is essentially the same splits as the optimal split tree with 15 variables. The r-squared is practically the same.

# PARTITION MODEL JMP

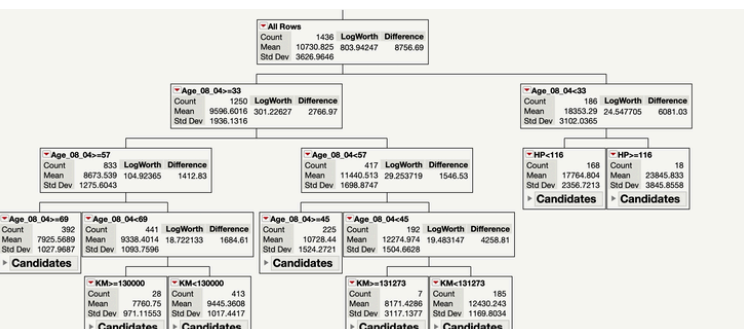


RSquare	RASE	N	Number of Splits	AICc
0.848	1414.8667	1436	7	24929.1

The model achieved an RSquared of .848, indicating that 84.8% of the variance in car prices is explained by the predictors in the training set.

As the tree was grown and additional splits were made, the RSquare value increased, and the RASE decreased. This is expected because the model fits more closely to the training data as splits are made.

While RSqaure increased and RASE decreased as the tree grew, performance plateaued or worsened because it began to overfit This shows that beyond a certain point, the splits began to model noise rather than meaningful patterns.



The test set performance was worse compared to the training set when looking at RSqaure and RASE. This is due to overfitting. The validation set performance was closer to the test set performance, indicating its usefulness with unseen data.

Term	Number of Splits	SS	Portion
Age_08_04	4	1.5205e+10	0.9501
HP	1	601205715	0.0376
KM	2	196750073	0.0123
Fuel_Type	0	0	0.0000
Doors	0	0	0.0000
Quarterly_Tax	0	0	0.0000
Mfg_Guarantee	0	0	0.0000
Automatic	0	0	0.0000
Automatic_airco	0	0	0.0000
Airco	0	0	0.0000
CD_Player	0	0	0.0000
Guarantee_Period	0	0	0.0000
Sport_Model	0	0	0.0000
Powered_Windows	0	0	0.0000
Tow_Bar	0	0	0.0000

Based on the column contributions in the regression tree, Age (95.01%), HP (Horsepower) (37.6%) and KM (1.23%) are the most important car specifications.



# CONCLUSION

---

The overall best-performing model was the MLR model with 15 features, achieving an R-Squared of 90.8% and an RMSE of 1016, indicating strong predictive accuracy and low error. The second-best model was the K-NN model in JMP, which used 4 features and achieved an R-Squared of 88% with an RMSE of 1255.

Based on our analysis, the four most important variables for determining car price are Age, HP, Quarterly Tax, and KM. These variables consistently contributed the most to model performance across different approaches.

Despite the simplicity of models with fewer predictors, we conclude that the MLR model with 15 features is the best for making predictions due to its superior performance metrics, explaining the highest proportion of variance in price while minimizing prediction error.

