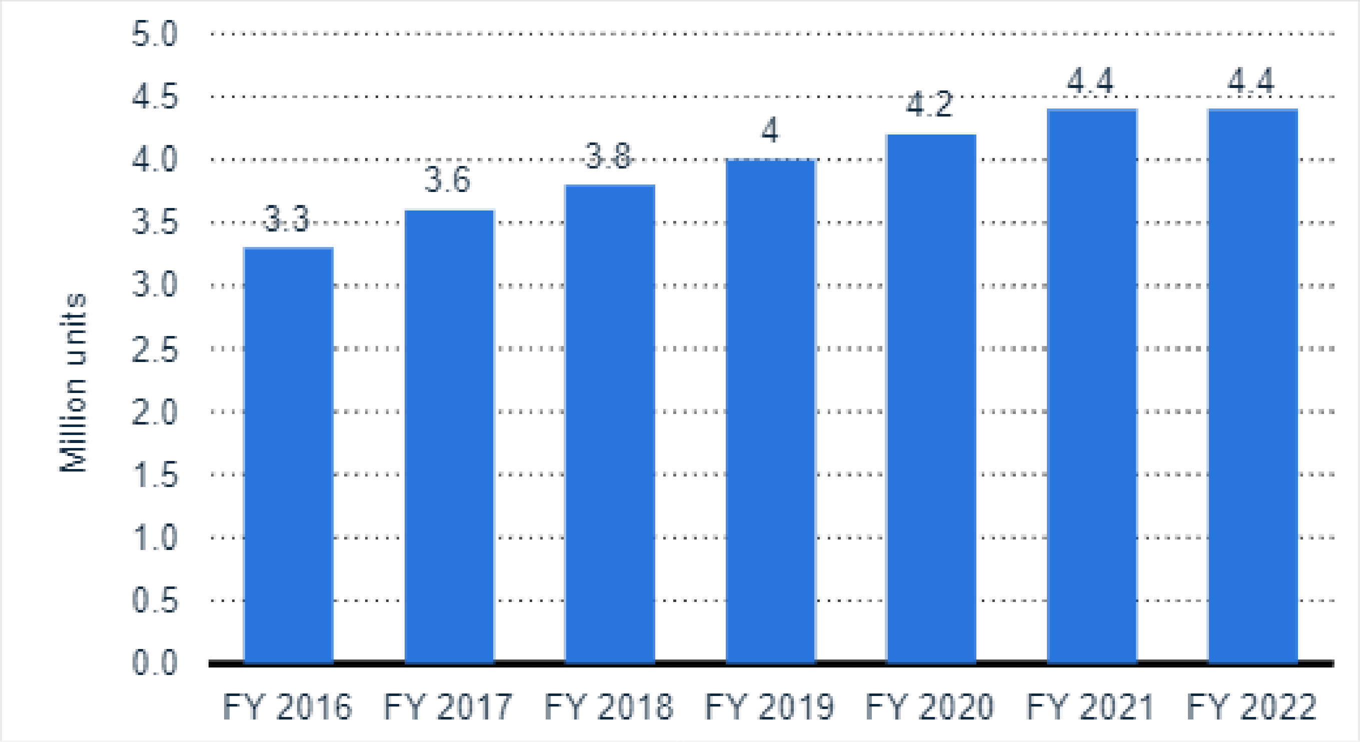
PREDICTING PRICE OF PRE-OWNED CARS IN INDIA

BUSINESS CASE + QUESTION

Market size of pre-owned car India FY 2016-2022



statista https://www-statista-com.eu1.proxy.openathens.net/statistics/1044753/india-pre-owned-car-market-size/

Facts

- ✓ Pre-Owned Car Market In India: $32.14 billion in 2021
- ✓ Anticipated to grow to $74.7 billion by 2027 (CAGR 15% for 2022-27)
- ✓ Sales of 4.4 million units of pre-owned in 2020 as compared to 2.8 million of new cars

Business Question:

Can historical data of past pre-owned car sales in India be used to predict the sale of a pre-owned car?

Business Case:

- ✓ Unorganized market
- ✓ Increasing demand since COVID-19 pandemic in 2020
- ✓ No standardized model to predict pricing
- ✓ No benchmarking standards

ANALYTICAL QUESTION AND GOALS?

Analytical Goals

- Predictive Accuracy

Variable Categories

Quantitative

Categorical

Predictors

New Car Price, Car Model, Kilometers Driver, Age of Vehicle, Owner Type, Transmission Type, Car Condition, Fuel Type

Outcome Variables

Selling Price of a Pre-Owned Car

Question: What are the main factors affecting the selling price of a pre-owned car in India?

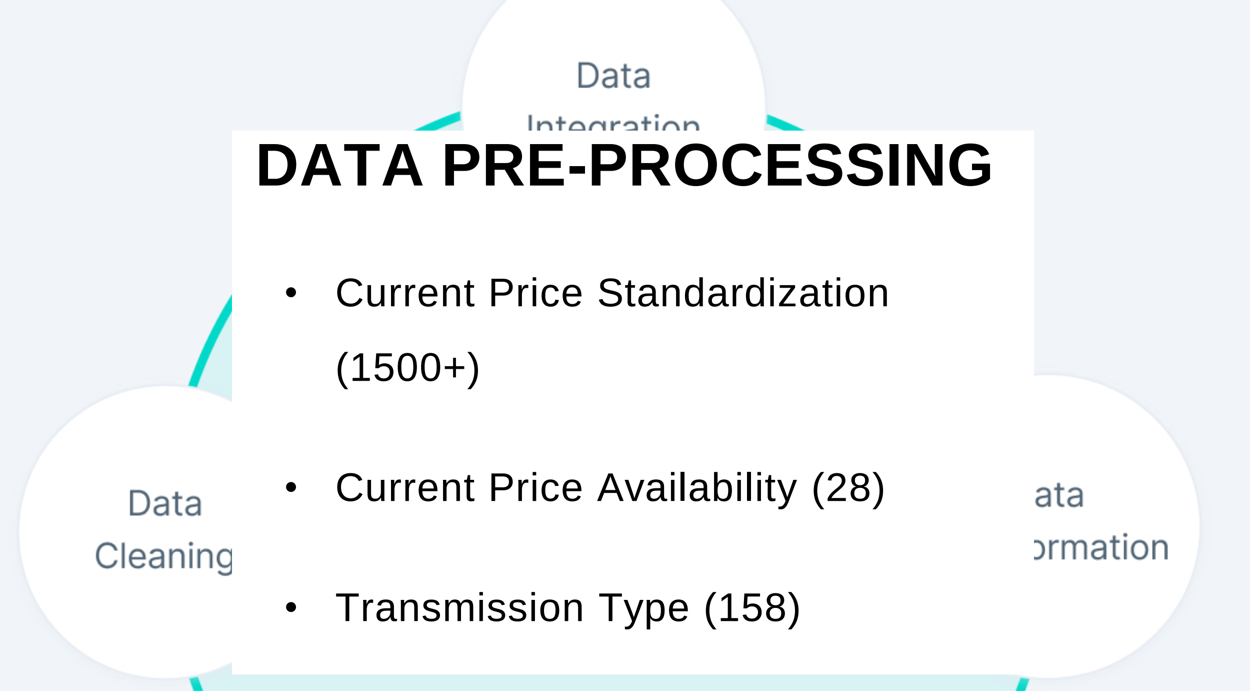
Developing a model to predict the price of a pre-owned car with an acceptable level of accuracy (70%+).

PROJECT DATA SET

Pre-Owned Car Data Set from Kaggle.com

NUMBER OF OBSERVATIONS

2,237



DATA PRE-PROCESSING

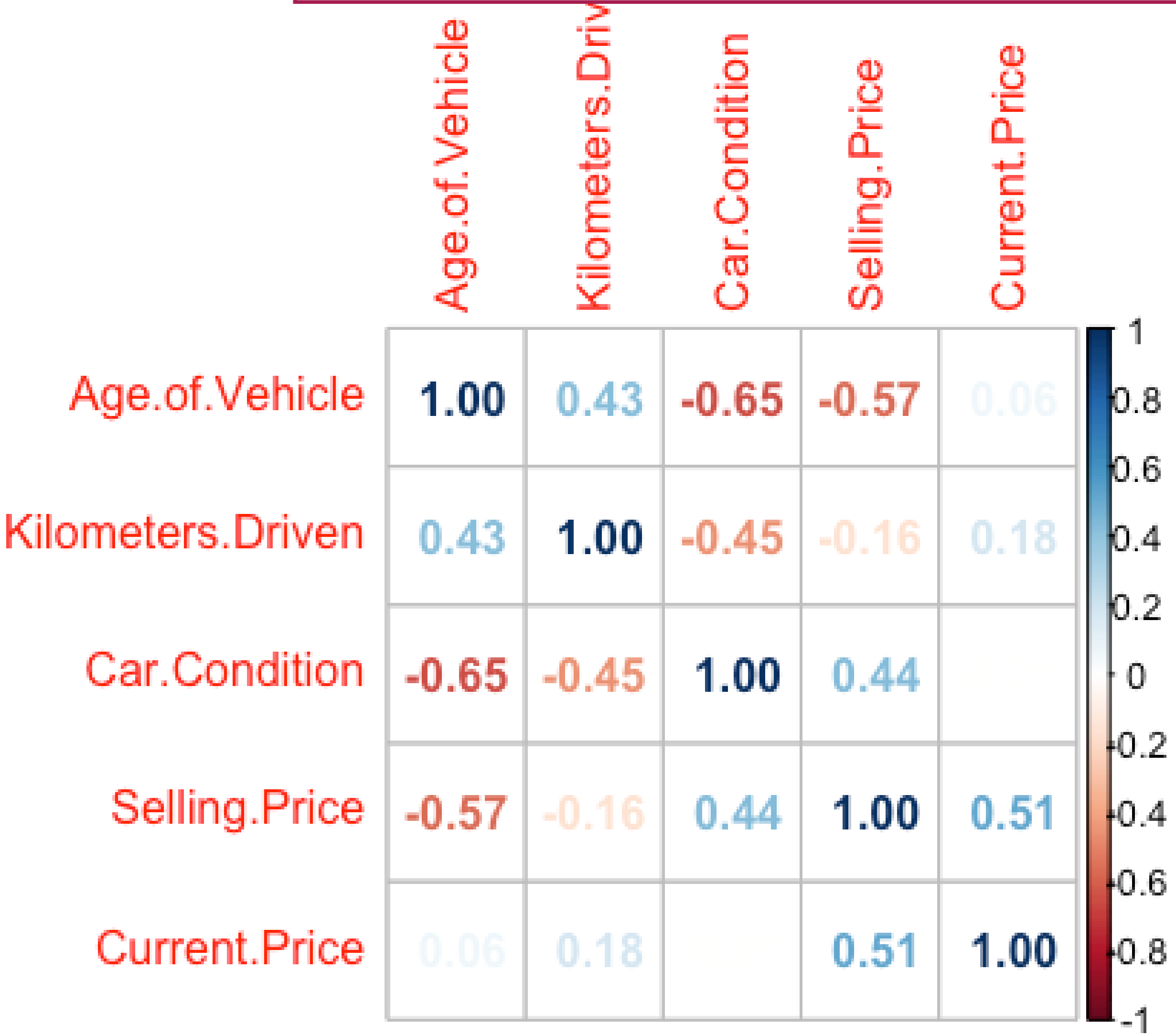
- Current Price Standardization (1500+)
- Current Price Availability (28)
- Transmission Type (158)

DATA TRANSFORMATION

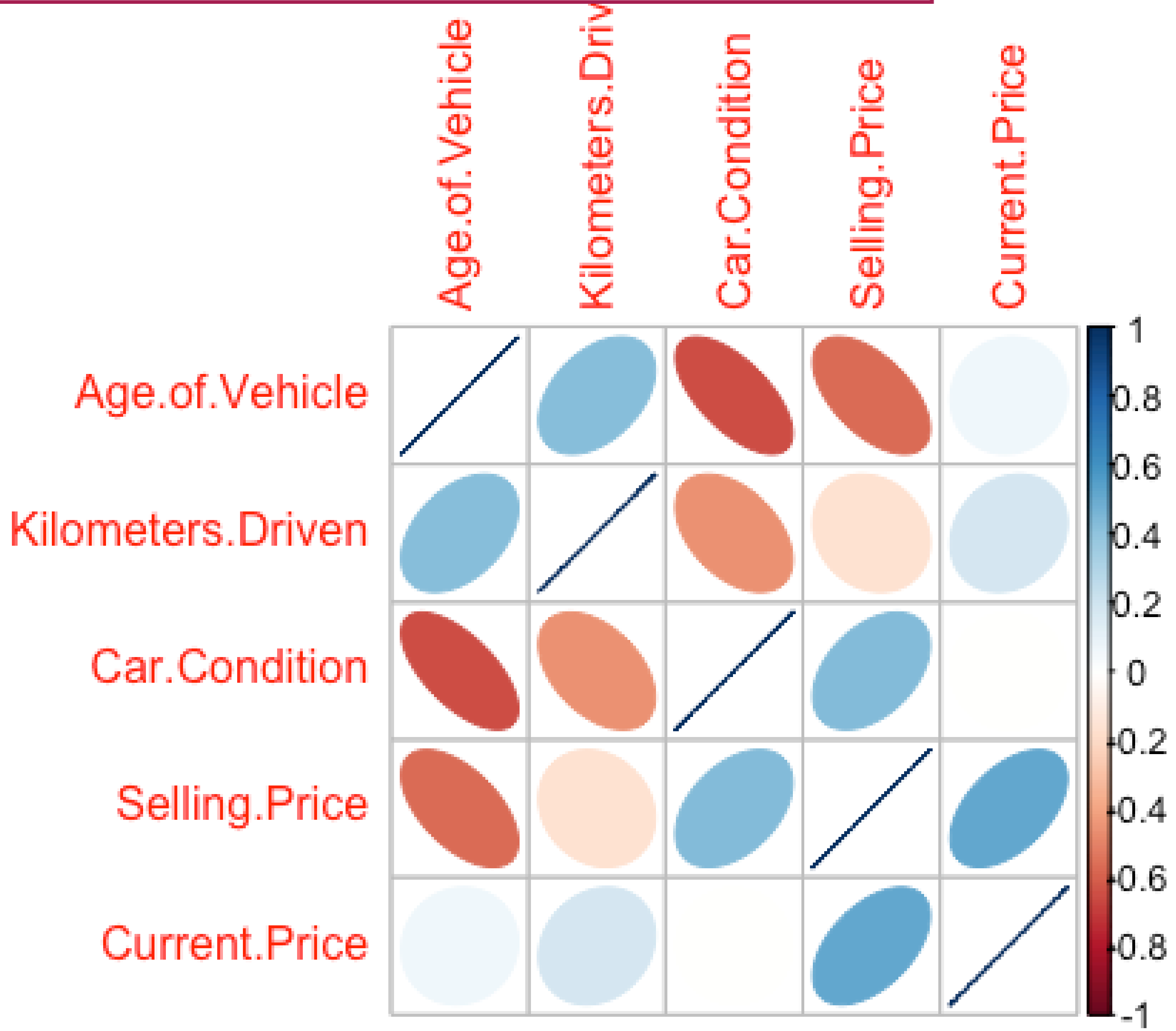
- Standardization of the OLS model

Descriptive Statistics/Analytics

Correlation of different continuous variables

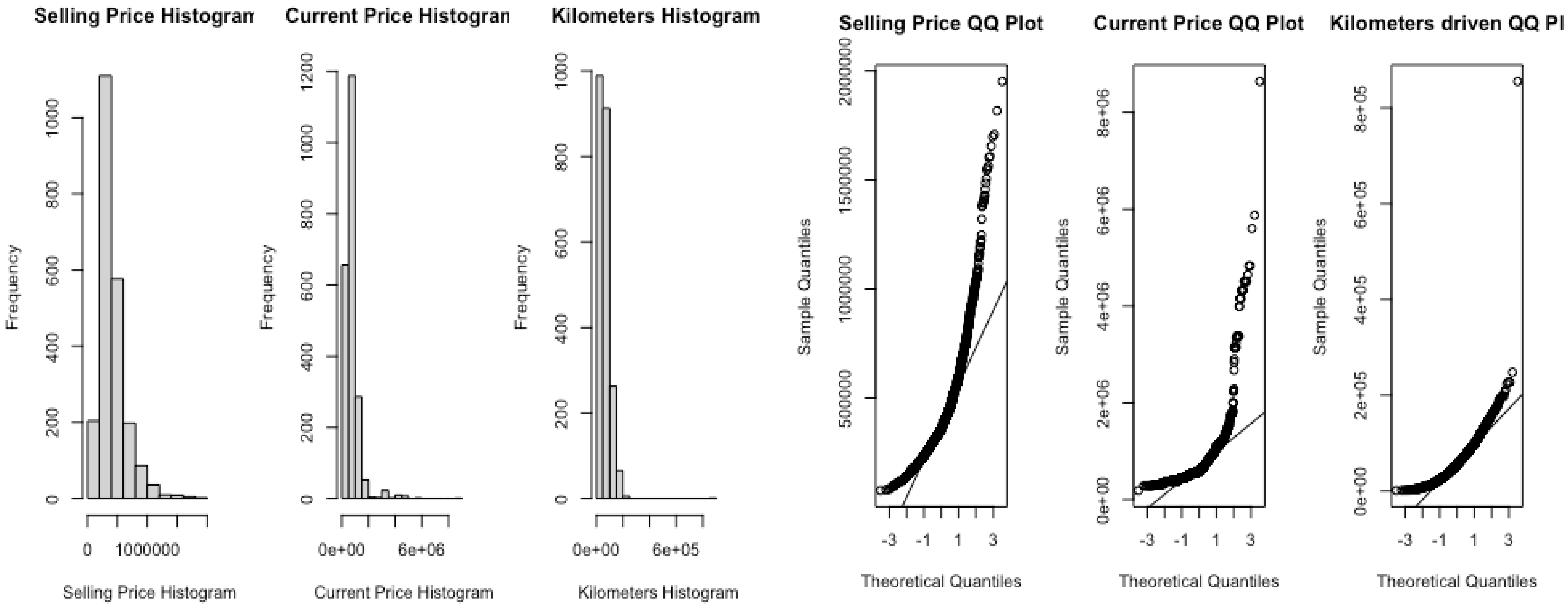


	Age.of.Vehicle	Kilometers.Driv	Car.Condition	Selling.Price	Current.Price
Age.of.Vehicle	1.00	0.43	-0.65	-0.57	0.06
Kilometers.Driven	0.43	1.00	-0.45	-0.16	0.18
Car.Condition	-0.65	-0.45	1.00	0.44	
Selling.Price	-0.57	-0.16	0.44	1.00	0.51
Current.Price	0.06	0.18		0.51	1.00



Descriptive Statistics/Analytics

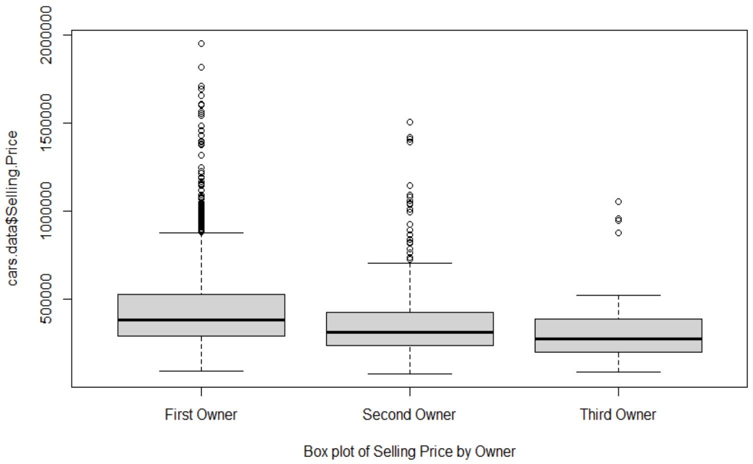
Histogram & qq-plot for Selling Price (Y variable), New Price and Kilometers driven



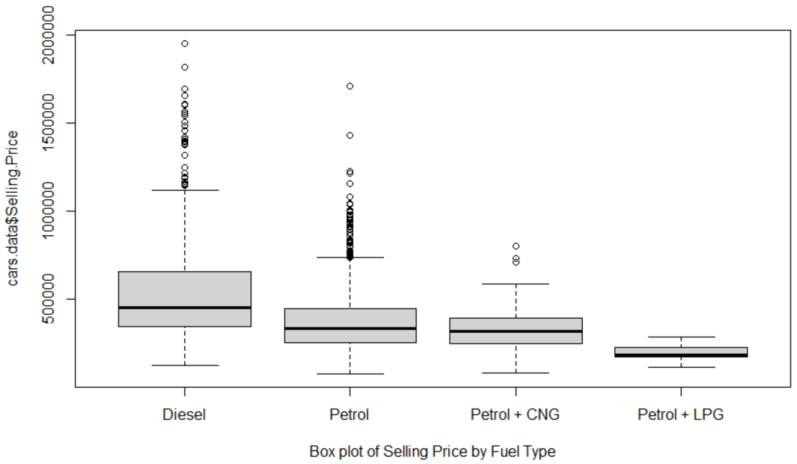
Descriptive Statistics/Analytics

Box Plots

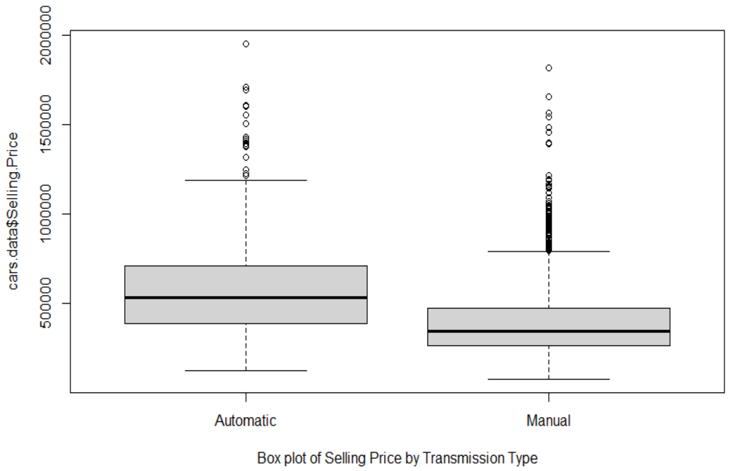
Boxplot for Selling Price by Owner



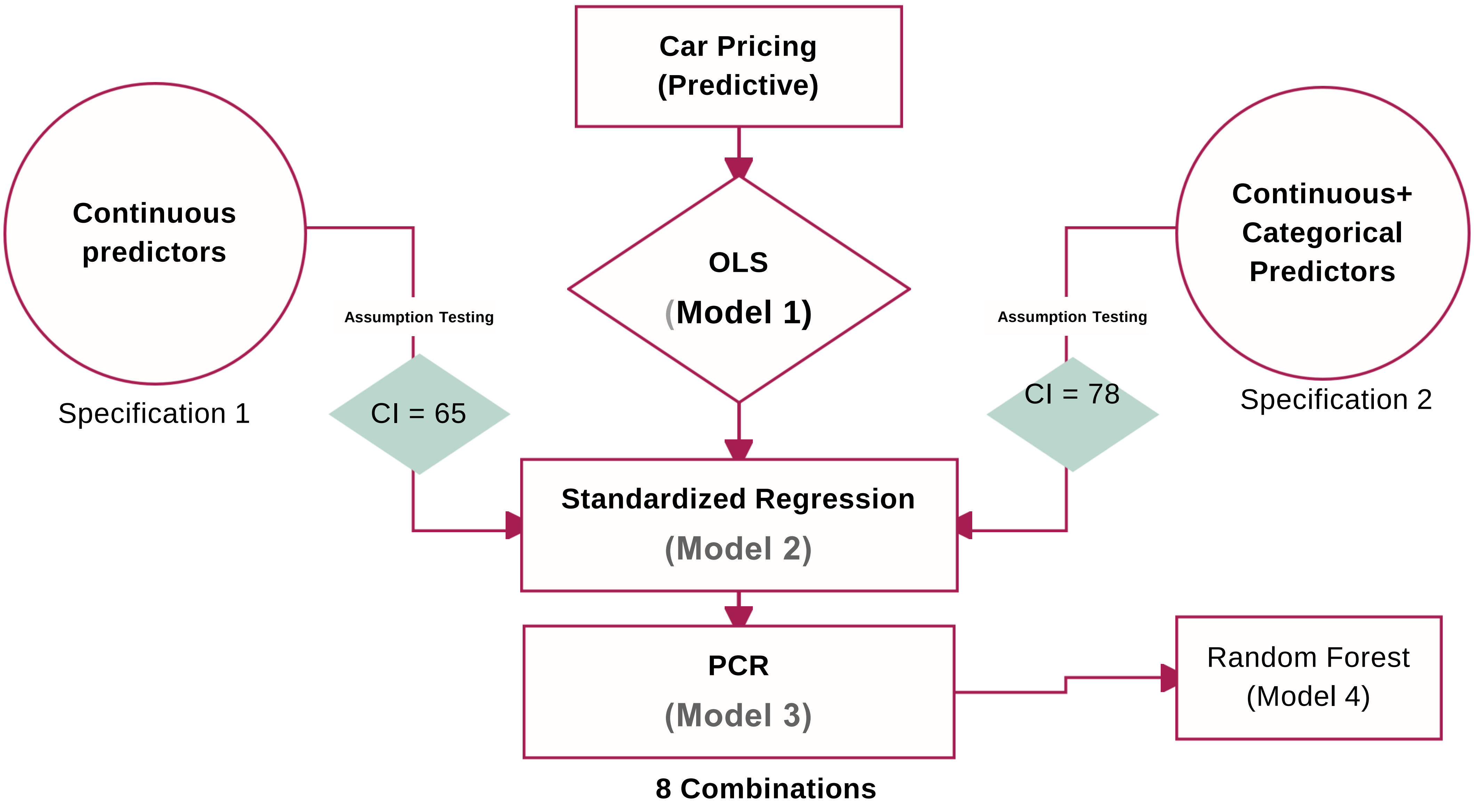
Boxplot for Selling Price by Fuel Type



Boxplot for Selling Price by Transmission Type



Modeling Methods and Specifications



Car Pricing
(Predictive)

OLS
(Model 1)

Continuous predictors

Specification 1

Assumption Testing

CI = 65

Continuous+ Categorical Predictors

Specification 2

Assumption Testing

CI = 78

Standardized Regression
(Model 2)

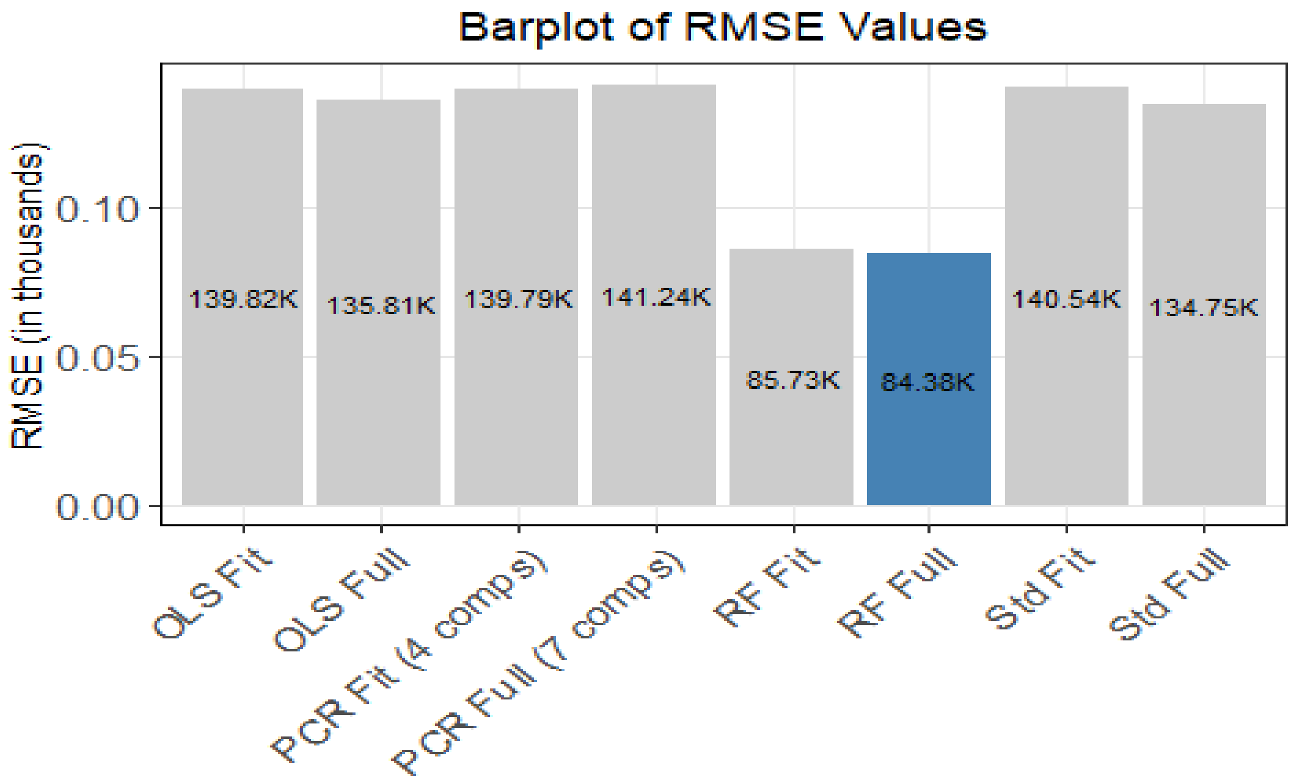
PCR
(Model 3)

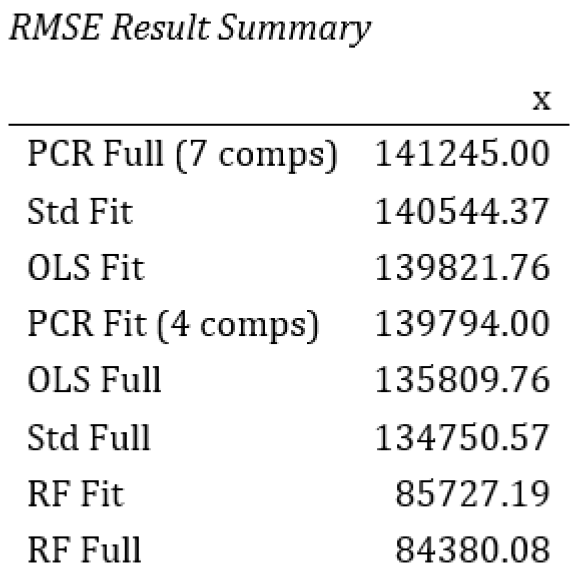
8 Combinations

Random Forest
(Model 4)

MODELING METHODS AND SPECIFICATIONS

Comparison of RMSE among Models (10 FCV)





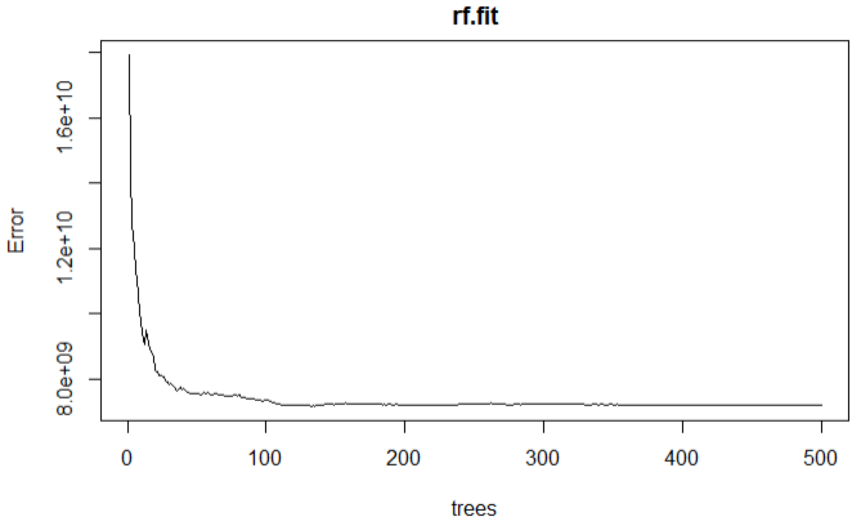
RMSE Result Summary

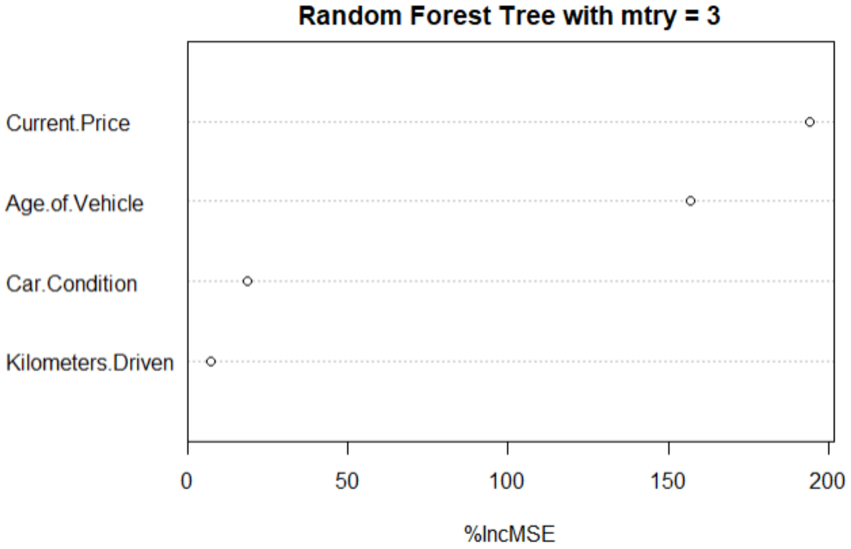
	x
PCR Full (7 comps)	141245.00
Std Fit	140544.37
OLS Fit	139821.76
PCR Fit (4 comps)	139794.00
OLS Full	135809.76
Std Full	134750.57
RF Fit	85727.19
RF Full	84380.08

ANALYSIS OF RESULTS

Candidate Model – Random Forest Fit Model with 3 variables at a time

RMSE value : 86052.59

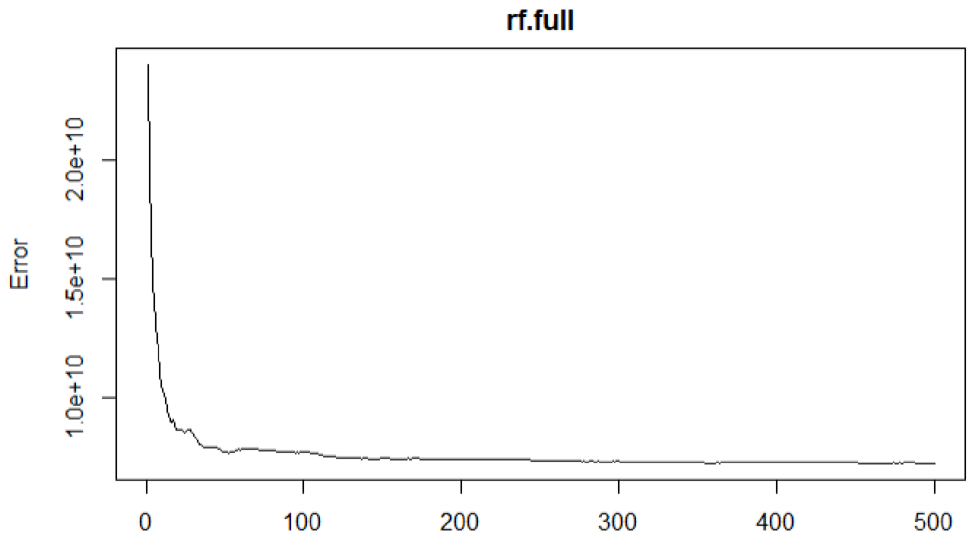


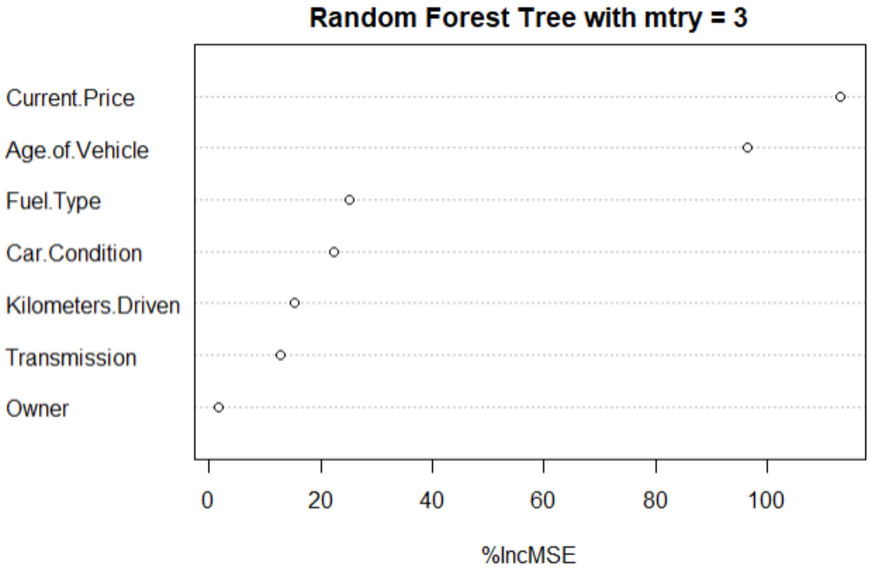


ANALYSIS OF RESULTS

Candidate Model – Random Forest Full Model with 3 mvariables at a time

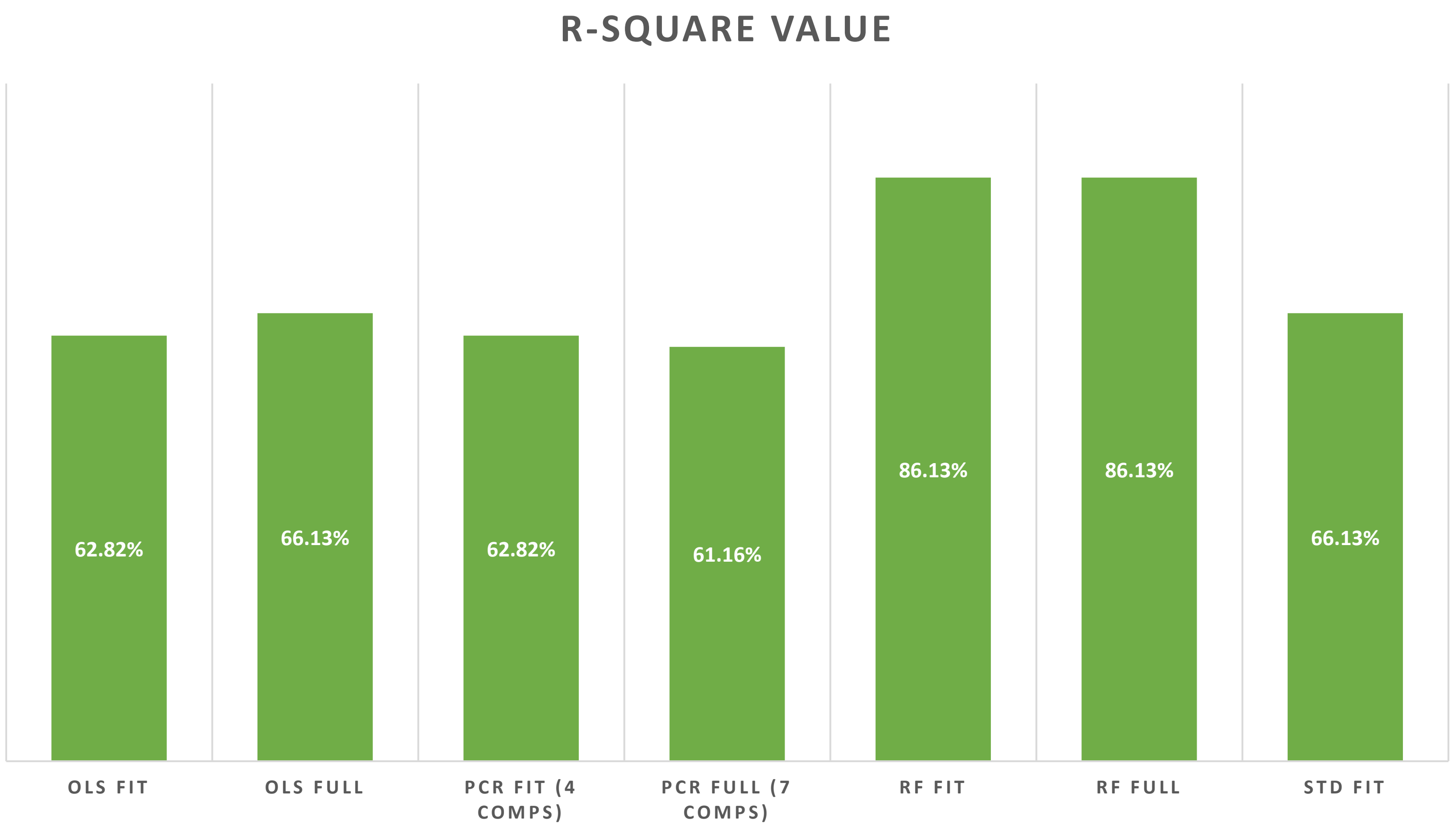
RMSE value : 87122.86





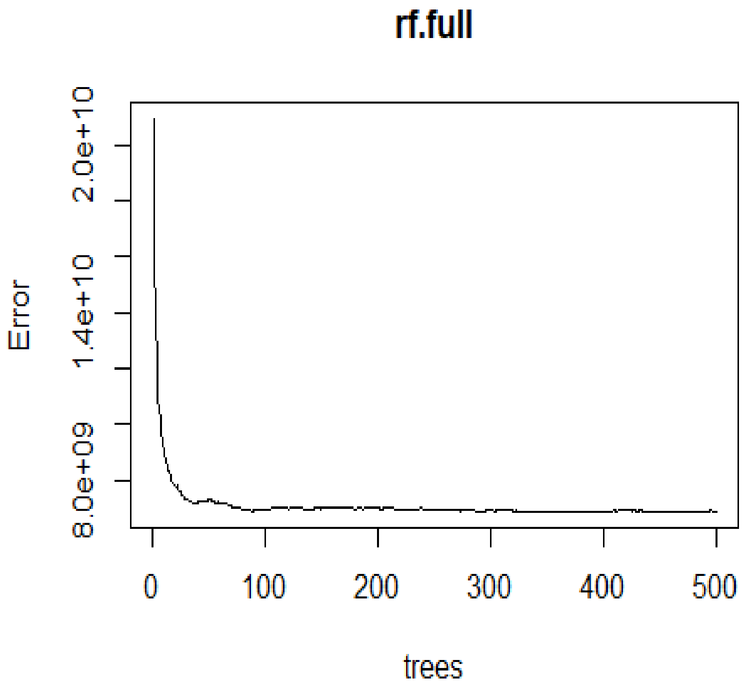
ANALYSIS OF RESULTS

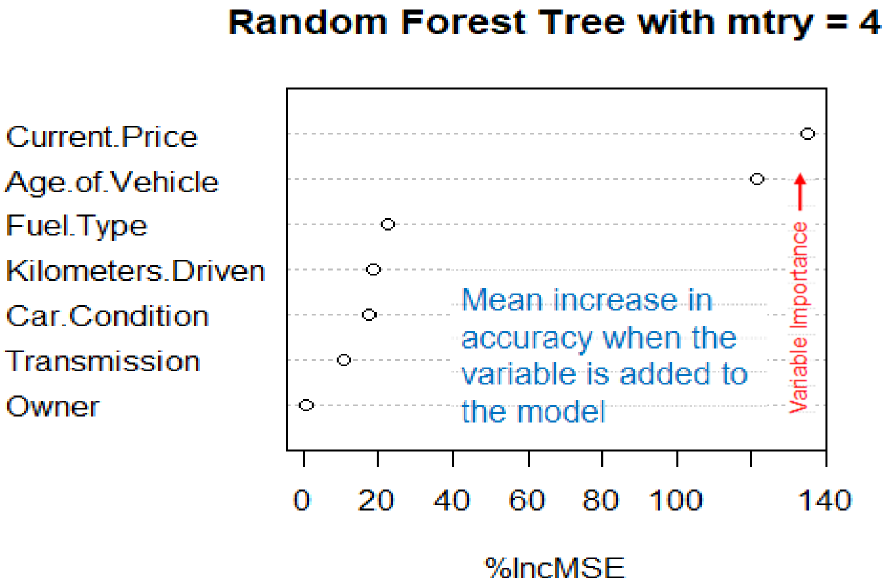
R Square Value of All Models



ANALYSIS OF RESULTS

Final Model – Random Forest Full Model with 4 variables at a time

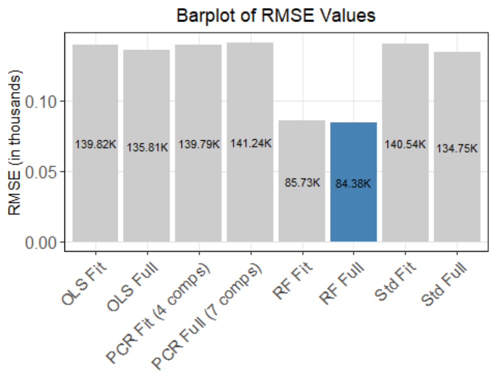


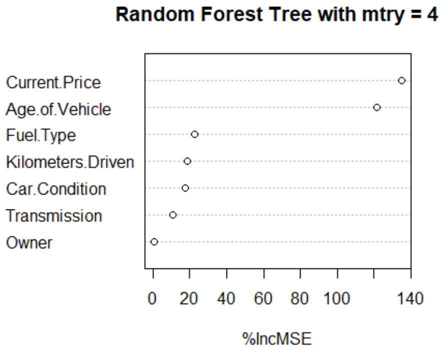


Mean increase in accuracy when the variable is added to the model

CONCLUSIONS

- Random Forest best for predictive accuracy
    - Original Sale Price (Current.Price) and Age of Vehicle most important variables to include in model
    - Low interpretability of model
- Future research should look for larger data sets, both in quantity of vehicles and in predictors
    - Random forest models useful for models with more predictors





CHALLENGES AND LESSONS LEARNED

- ✓ Dealing with data with missing values/ incorrect values during data preprocessing and transformation
- ✓ Selecting the models based on initial assessment of OLS regression model
- ✓ Random forest computing requirements
- ✓ Project management and coordination in and around finals