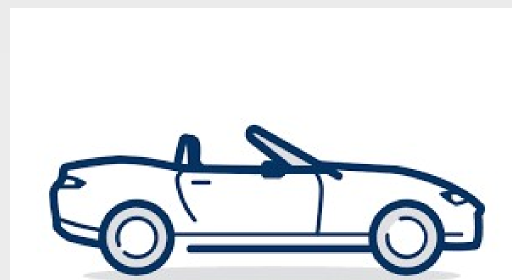
Data Science and Analytics DSA 6000 Final Project

Predicting Used Car Prices using Statistical Learning Methods



Dataset

- Over 370,000 used car data scraped from Ebay-Germany
- Data was scraped in March-April 2016
- Content of the dataset is in German
- It contains a total of 20 Numerical and Categorical Variables
- Response Variable is Price of the used car

Objectives

- Predict price of a Used Car
- Identify various factors driving the selling price for Vehicles
- Calculate the accuracy of the predicted price
- Identify correlation between various vehicle characteristics
- Identify the most popular car brands in German Market

Project Flow

Step 1

Data Cleaning

Step 2

Data Visualization

Step 3

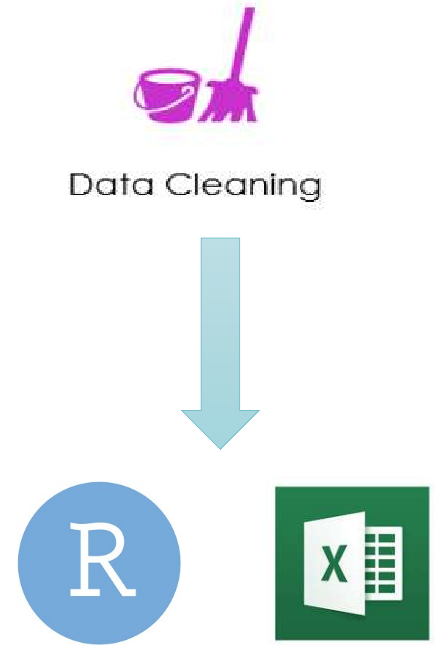
Model Building

Step 4

Calculate error and accuracy rate

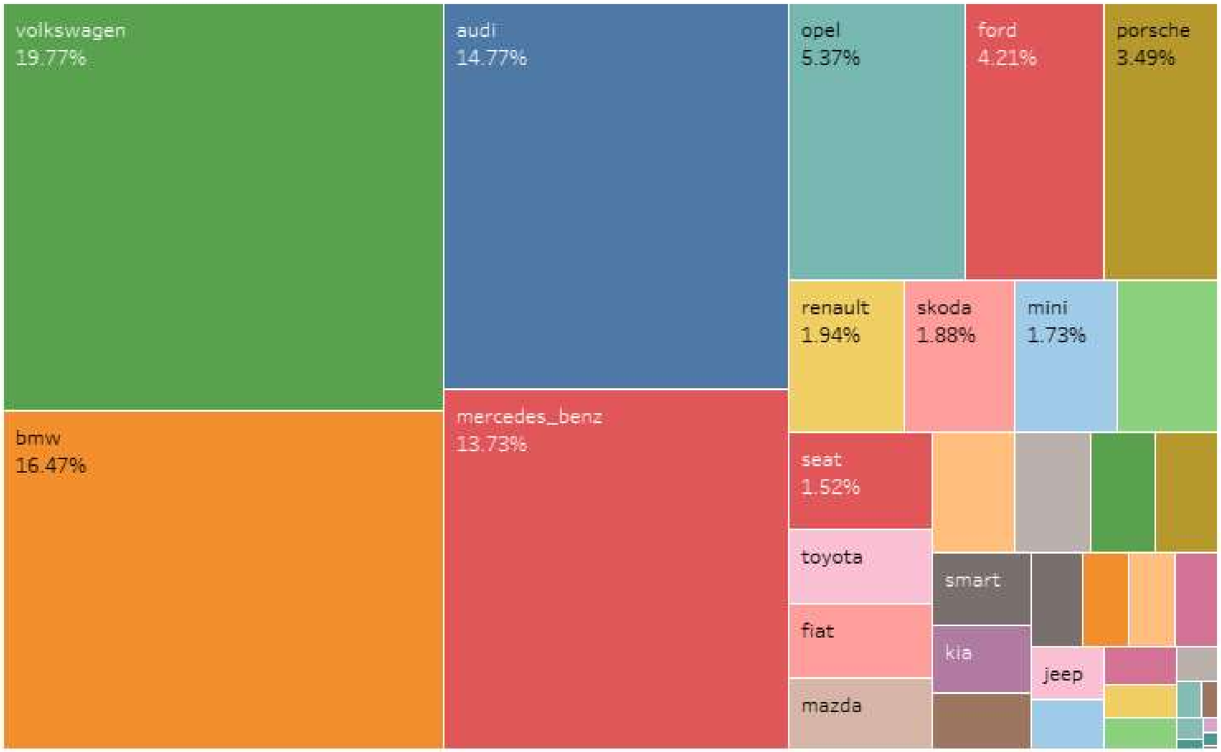
Step 1: Data Cleaning

- Loading Dataset in R
- Removing unwanted Columns (Like Seller Type, Offer Type, Ab test)
  `autodata$seller<-NULL`
- Removing Outliers (Like Price=0, Power>600, yearofRegistration> 2016)
  `autodata<-subset(autodata, autodata$price>200)`
- German Translation to English
  `autodata$fuelType<-gsub("andere","others",autodata$fuelType)`
- Handling NAs
  Assigning the NAs to “others” or “unknown”
  `colSums(is.na(autodata))`
- Calculating Age of Car using Year of Registration



Step 2 : Data Visualization (Using Tableau)

TOP BRANDS WITH TOP PRICES

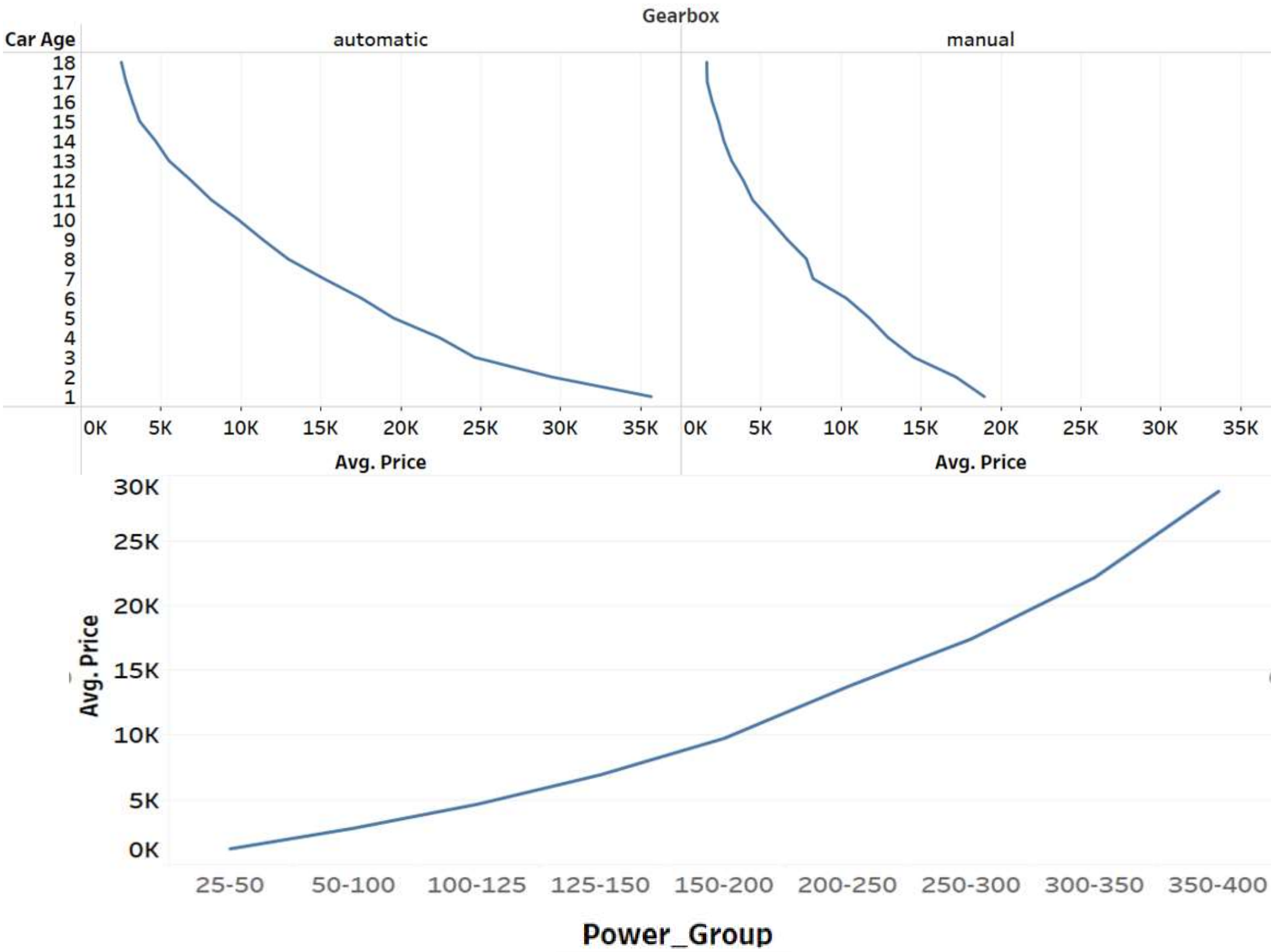


- Using Tableau, we were able to determine some key observations about the data.
- We are seeing that sales volume in the German Market is dominated by 5 major brands

<u>Next Steps:-</u>

- Since we see that 5 brands are dominating the German Car Market, it was considered sufficient to use only these to fit our model moving forward

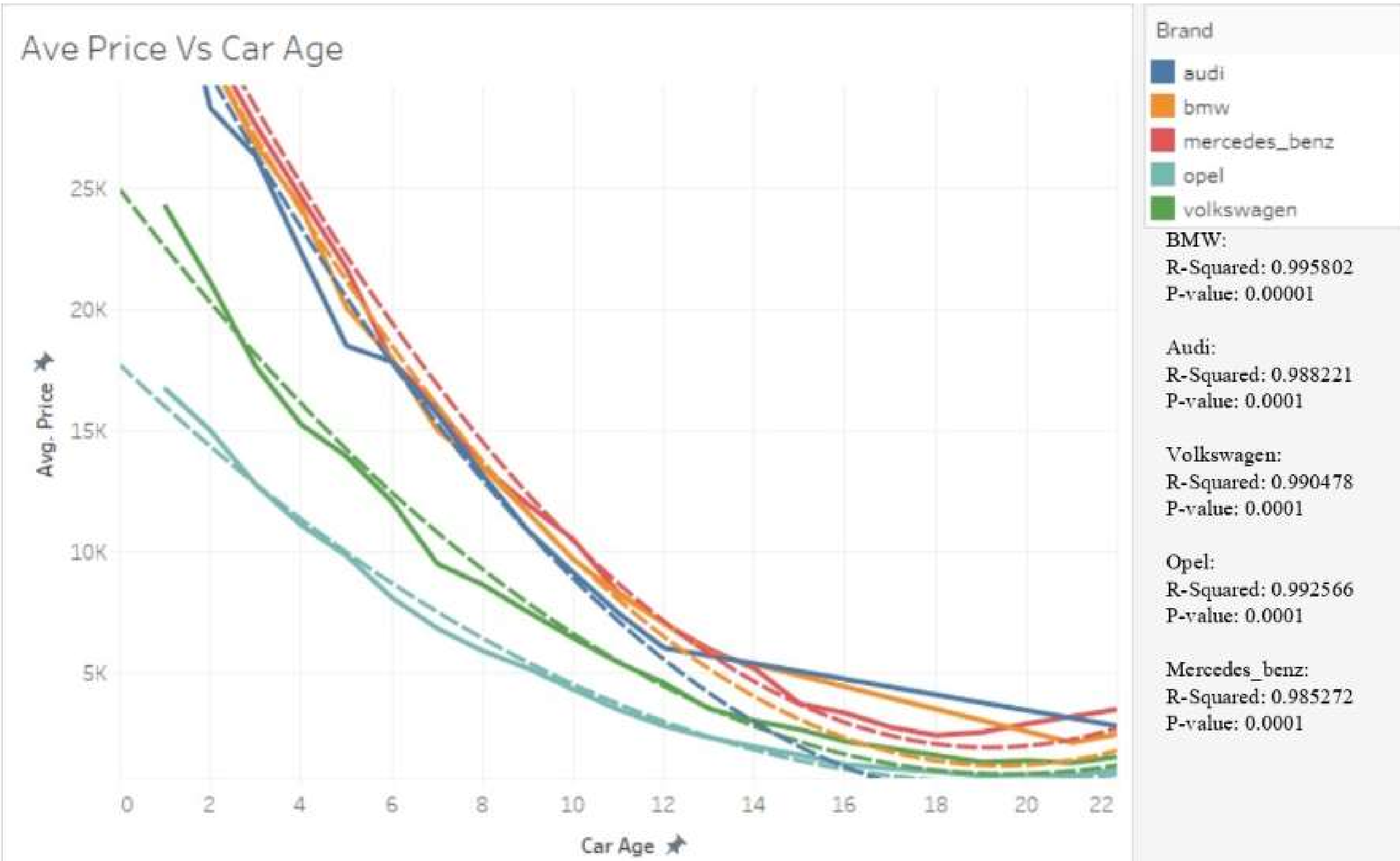
Data Visualization (Cont…)



- We can see relationship between Price vs Car age and Price vs Power
- We can see an inverse relation between Car Age vs Average Price and direct relation between Power vs Price

This helped us determine correlation between response and predicted variables.

Car Price Depreciation



- We can see a strong relationship between the Car Age and the Depreciation of Price

- Looking at the model, we can see that R-Squared and P-value indicate a strong relationship

Step 3: Model Building

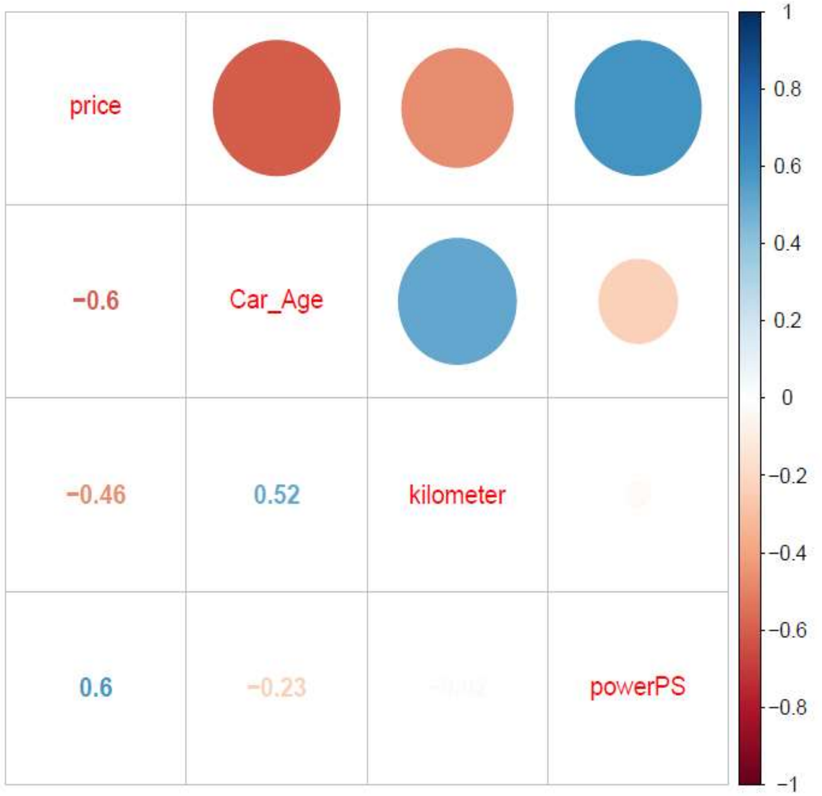
- **Installing required Packages**

- **Finding Correlation in the data**

```
install.packages("corrplot")
library(corrplot)
au<-autodata[,c("price","Car_Age","kilometer","powerPS")]
corrplot.mixed(cor(au))
```

- **Reviewing Dataset**

✓ The new dataset contains 280,081 rows and 10 variables
✓ 1 dependent variables and 9 independent variables
✓ No Missing Values



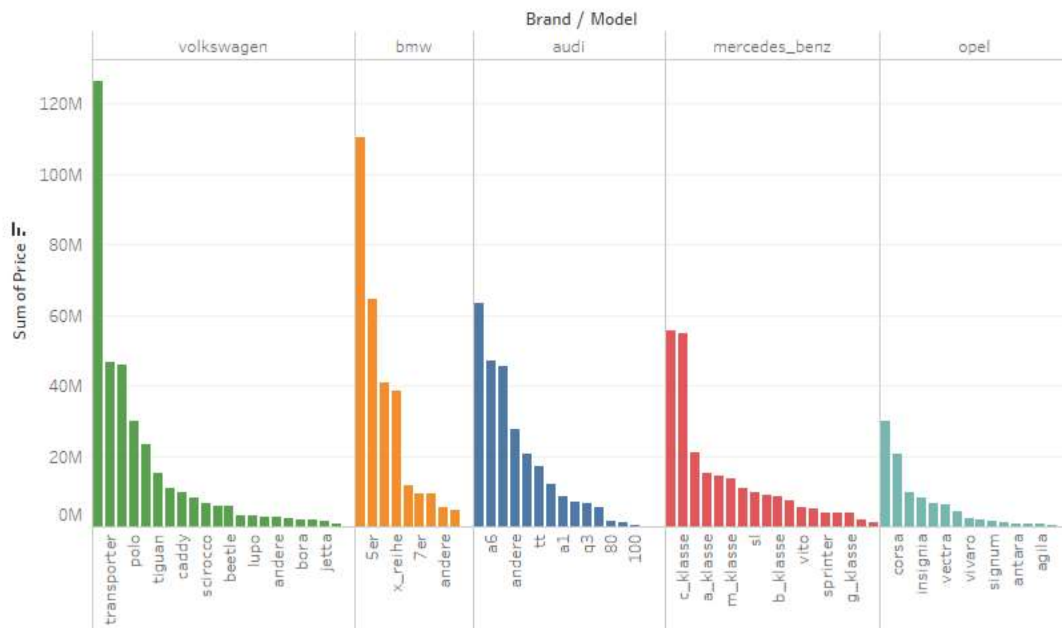
Model Building (Cont..)

- Data Exploration:

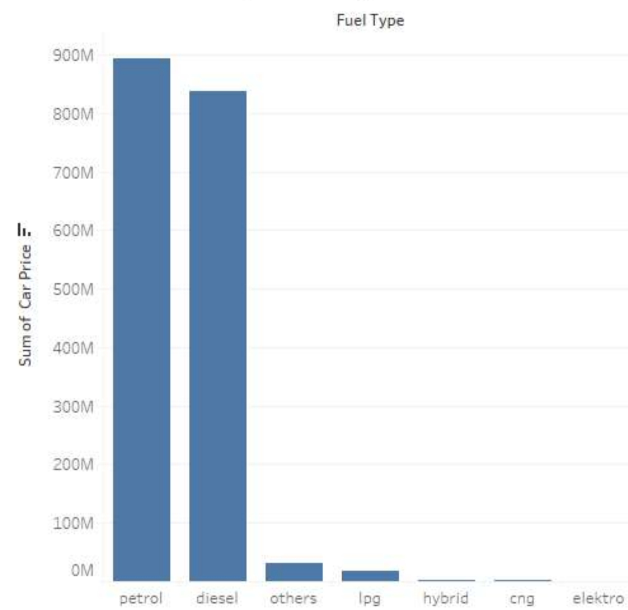
✓ Removing fuel type other than diesel and petrol

✓ Removing outdated Models

Sum of Price -Model-wise for each Brand



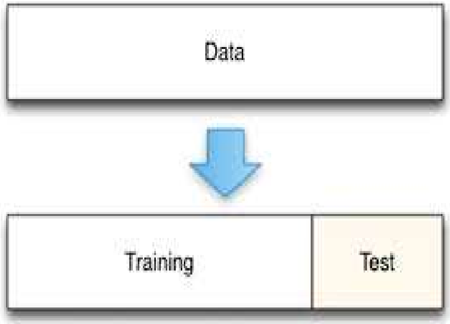
Sum of Car Price as per Fuel Type



Model Building (Cont..)

- Splitting the data into training and test dataset

```
smp_size <- floor(0.7 * nrow(autodat))
train_ind <- sample(seq_len(nrow(autodata)), size = smp_size)
train<- autodata[train_ind,]
test <- autodata[-train_ind, ]
```



- Building and comparing different models

Model 1:-

```
Model_simple<-lm(price~vehicleType+gearbox+powerPS+Kilometer+fuelType+brand+CarAge,
data=train)
summary(Model_simple)
```

```
Residual standard error: 3950 on 113393 degrees of freedom
Multiple R-squared:  0.7314,    Adjusted R-squared:  0.7313
F-statistic: 1.715e+04 on 18 and 113393 DF,  p-value: < 2.2e-16
```

Model Building (Cont..)

Model 2: - Using Log Transformation

```
Model_log<-
lm(log(price)~vehicleType+gearbox+powerPS+Kilometer+fuelType+brand+CarAge, data=train)
```

```
summary(Model_log)
pred_log<-exp(predict(Model_log,train))
RSS<-sum((train$price-pred_log)^2)
TSS<- sum((train$price-mean(train$price))^2)
```

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\mathrm{TSS} = \sum(y_i - \bar{y})^2$$

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}$$

- ✓ **R^2= 0.715**
- ✓ **RSE = 4042.023**

- Test Data Prediction

```
pred_log<-exp(predict(Model_log,test))
```

Model Building (Cont..)

Model 3: - Using Sqrt Transformation

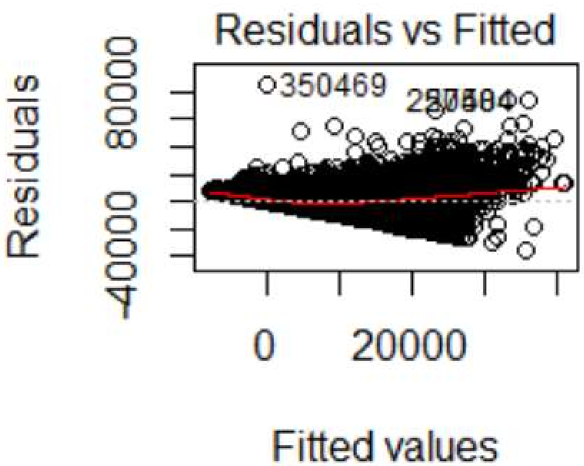
```
Model_Sqrt<-
lm(sqrt(price)~vehicleType+gearbox+powerPS+Kilometer+fuelType+brand+CarAge,
data=train)

summary(Model_Sqrt)
pred_sqrt<-predict(Model_sqrt,train)^2
RSS<-sum((train$price-pred_sqrt)^2)
TSS<- sum((train$price-mean(train$price))^2)
RSE<- sqrt(RSS/(nrow(train)-16-1))
```

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

- ✓ **R^2= 0.813**
- ✓ **RSE = 3273.87**

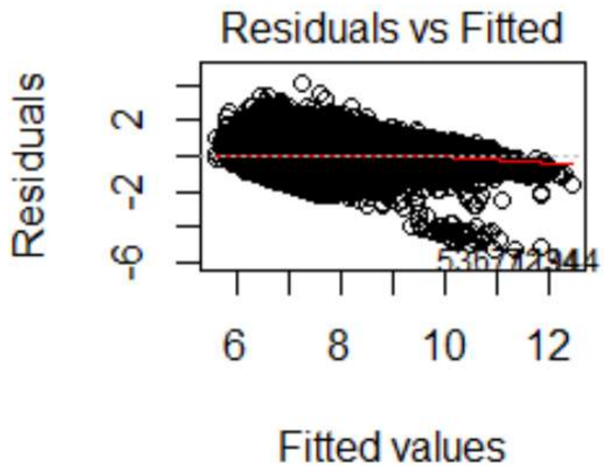
Residual Plots



Simple Model

R^2= 0.731
RSE=3950

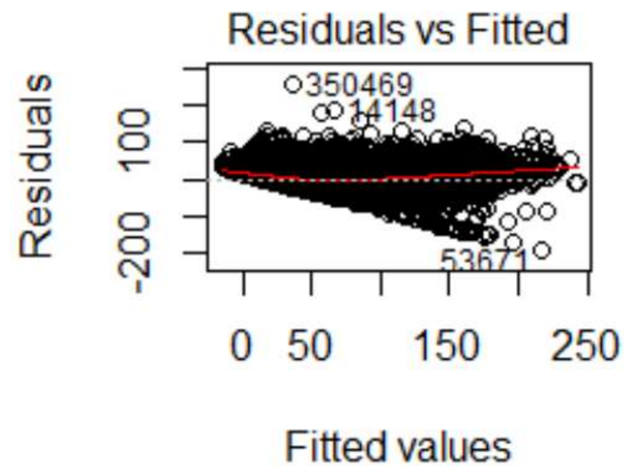
X



Log Model

R^2= 0.715
RSE=4042.023

X



Sqrt Model

R^2= 0.813
RSE=3273.87

✓

✓ Sqrt Model has a better R^2 and **RSE,** hence we have selected this model

Model Building (Cont..)

- No sign of Multicollinearity

```
install.packages("car")
library(carData)
library(car)
```

$$VIF_i = \frac{1}{1 - R_i^2}$$

```
> vif(Model_sqrt)
                GVIF Df GVIF^(1/(2*Df))
vehicleType 1.992173  7        1.050462
gearbox     1.458229  1        1.207571
powerPS     2.157146  1        1.468723
kilometer   1.556937  1        1.247773
fuelType    1.602396  1        1.265858
brand       1.716394  4        1.069860
Car_Age     1.954894  1        1.398175
```

Summary of the SQRT Model

```
Coefficients:
                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)            1.414e+02  3.670e-01  385.179  < 2e-16 ***
vehicleTypeconvertible 6.896e+00  2.708e-01   25.465  < 2e-16 ***
vehicleTypecoupe      -4.761e+00  2.787e-01  -17.086  < 2e-16 ***
vehicleTypekombi      -1.332e+01  2.015e-01  -66.075  < 2e-16 ***
vehicleTypelimousine  -9.758e+00  1.979e-01  -49.318  < 2e-16 ***
vehicleTypeothers     -1.232e+01  4.659e-01  -26.446  < 2e-16 ***
vehicleTypesmall_Car  -1.007e+01  2.231e-01  -45.141  < 2e-16 ***
vehicleTypesuv         3.050e+00  3.575e-01    8.529  < 2e-16 ***
gearboxmanual         -3.274e+00  1.298e-01  -25.214  < 2e-16 ***
powerPS                2.159e-01  1.173e-03  184.026  < 2e-16 ***
kilometer             -2.352e-04  1.627e-06 -144.516  < 2e-16 ***
fuelTypepetrol        -5.854e+00  1.244e-01  -47.076  < 2e-16 ***
brandbmw              -1.315e+00  1.684e-01   -7.810 5.77e-15 ***
brandmercedes_benz    -3.583e+00  1.807e-01  -19.826  < 2e-16 ***
brandopel             -1.471e+01  1.876e-01  -78.413  < 2e-16 ***
brandvolkswagen       -3.404e+00  1.615e-01  -21.079  < 2e-16 ***
Car_Age               -3.840e+00  1.227e-02 -312.917  < 2e-16 ***
```

Model Building (Cont..)

Regression Equation-
sqrt((price) =
141.2 + 6.618*vehicleTypeconvertible -4.833*vehicleTypecoupe -13.5*vehicleTypekombi - 9.743*vehicleTypelimousine -12.05*vehicleTypeothers -10.18*vehicleTypesmall_Car +2.88*vehicleTypesuv - 3.262*gearboxmanual + 0.2177*powerPS -0.0002356*kilometer -5.838*fuelTypepetrol -1.418*brandbmw - 3.464*brandmercedes_benz -14.82*brandopel -3.443*brandvolkswagen -3.832*Car_Age

- `pred_sqrt_test<(predict(Model_sqrt,test)^2)`

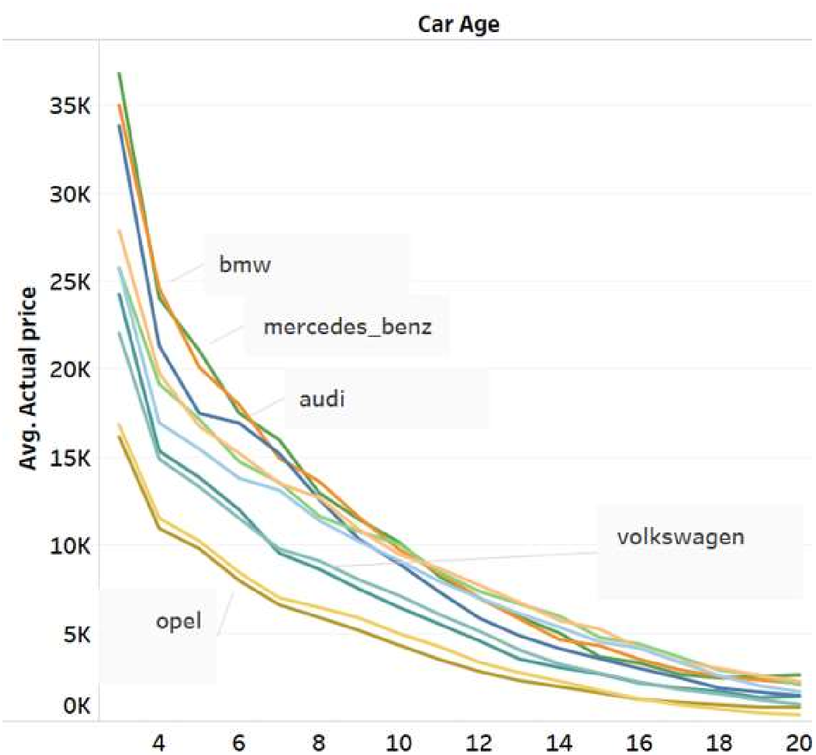
price	Vehicle Type	gearbox	powerPS	kilometer	fuelType	brand	Car_Age
9300	kombi	manual	143	150000	diesel	audi	8

141.2 -13.5*vehicleTypekombi -3.262*gearboxmanual +0.2177*powerPS -0.0002356*kilometer - 5.838*fuelTypepetrol +1.418*brandbmw -3.464*brandmercedes_benz -14.82*brandopel -3.443*brandvolkswagen -3.832*8
=141.2 -13.5 -3.262+ 0.2177*143-0.0002356*150000 -0+0+0+0+0 -3.832*8
= 89.5731

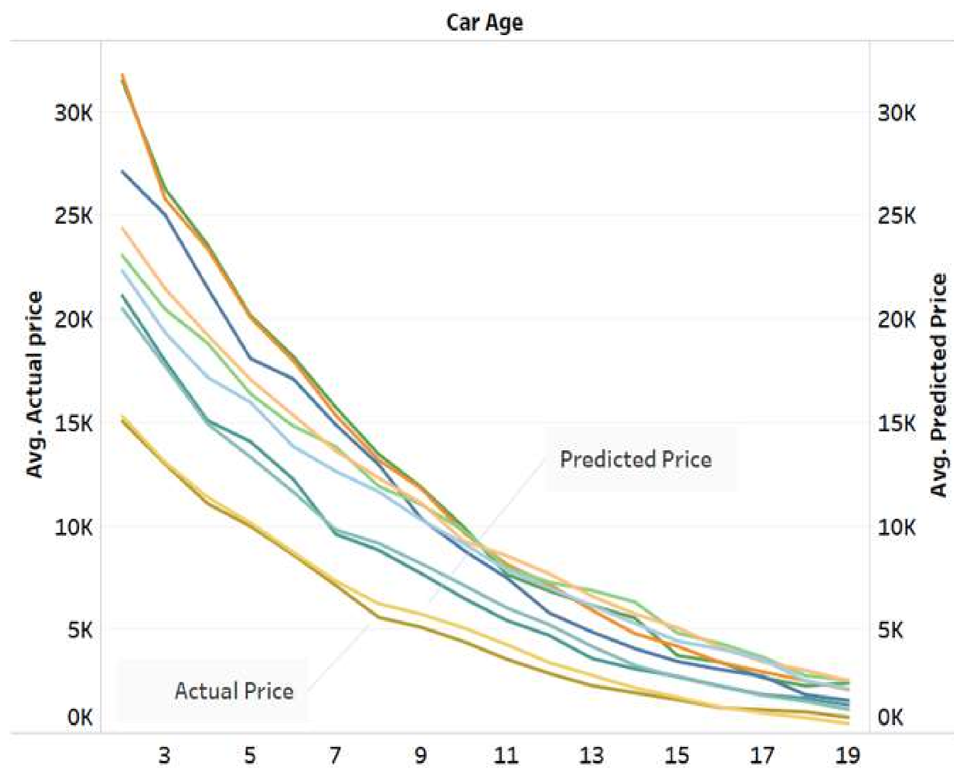
✓ Actual Prediction- (89.5731)^2 = 8023.34 EURO

Results

- Comparing Predicted versus Actual Price for both Training and Test Data

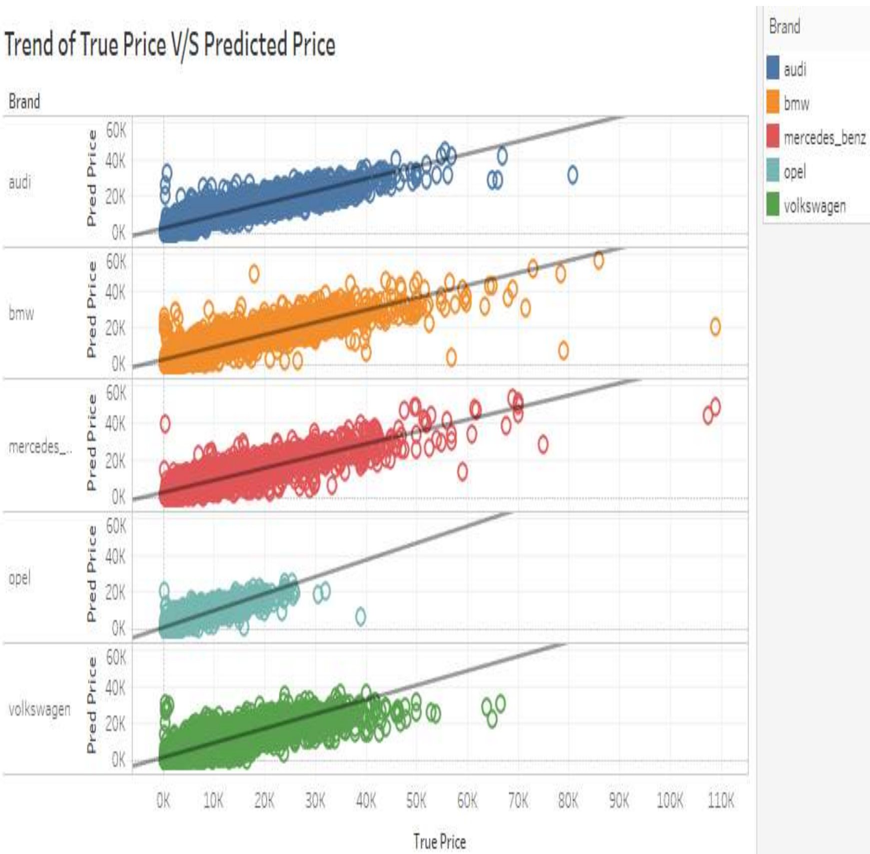


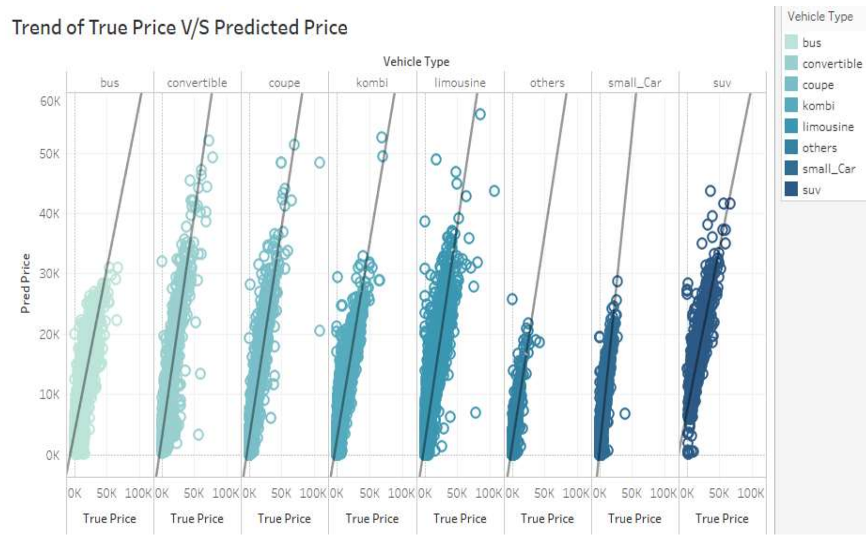
Training Data



Test Data

Results





Sno	brand	Actual_price	Predicted_Price
1	bmw	1,800	2,706
2	volkswagen	3,799	3,665
3	volkswagen	750	767
4	volkswagen	3,950	3,704
5	mercedes_benz	2,990	2,525
6	bmw	29,799	27,448
7	opel	3,330	3,056
8	mercedes_benz	5,500	5,745
9	audi	23,990	22,141

Conclusions

- SQRT model gives the best fit for the dataset
- Volkswagen, BMW, Audi, Mercedes Benz and Opel are most popular brands in the German Market
- Vehicle Type, Gearbox, Power PS, Kilometers driven, Brand, Car Age, Brand are significant variables in the predicting the price of a Used Car
- An R^2 of 0.81 and RSE of 3273.87 is achieved using our prediction model

Questions!