THE PRICE IS RIGHT

A car pricing exercise *by Agata Gramatyka*

INTRODUCTION

- Selling or buying a car is one of the largest transactions most people make in their lives
- Over 70% of all cars sold in France are second-hand
- In 2016 alone, 5.6 million of them exchanged hands
- Finding the right price can be tricky, despite some guidance from websites like Argus
- Being able to determine the true value is also important for insurers
- Machine Learning can help determine the price in a more accurate and transparent way

THE DATASET

- 371,500 car sale ads scraped from German ebay in 2016
- Available from Kaggle (https://www.kaggle.com/orgesleka/used-cars-database)
- Good level of detail across 13 columns of data including:
  - Car make and model
  - Month and year of registration
  - Mileage
  - Brake horsepower
  - Transmission
  - Body and fuel type
  - Presence of unrepaired damage
  - Sale price

STEP 1: DATA CLEANING

➡ *Missing values: count by variable*

```
name                      0
seller                    0
price                 13320
vehicleType           37869
yearOfRegistration        0
gearbox               20209
powerPS                   0
model                 20484
kilometer                 0
monthOfRegistration       0
fuelType              33386
brand                     0
notRepairedDamage     72060
```

STEP 1: DATA CLEANING

➡ *Missing values: employed techniques*

1. **Price/Unrepaired Damage:** remove all concerned rows
2. **Gearbox/Fuel/Body type:** impute from most common kind by model
3. **Model:** extract from the ad title if possible, remove otherwise

name
Golf_3_1.6
A5_Sportback_2.7_Tdi
Jeep_Grand_Cherokee_"Overland"
GOLF_4_1_4__3T�RER
Skoda_Fabia_1.4_TDI_PD_Classic
BMW_316i___e36_Limousine___Bastlerfahrzeug__Export

STEP 1: DATA CLEANING

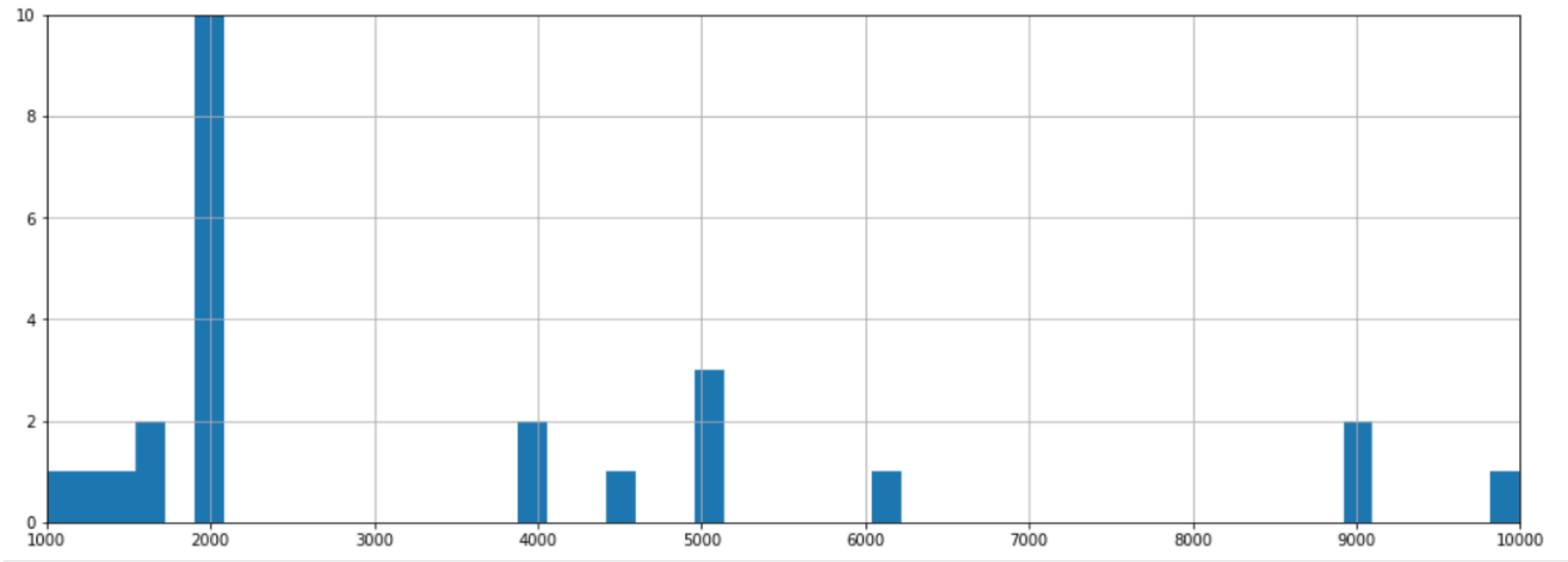
➡ *Outliers: extreme or unrealistic values*

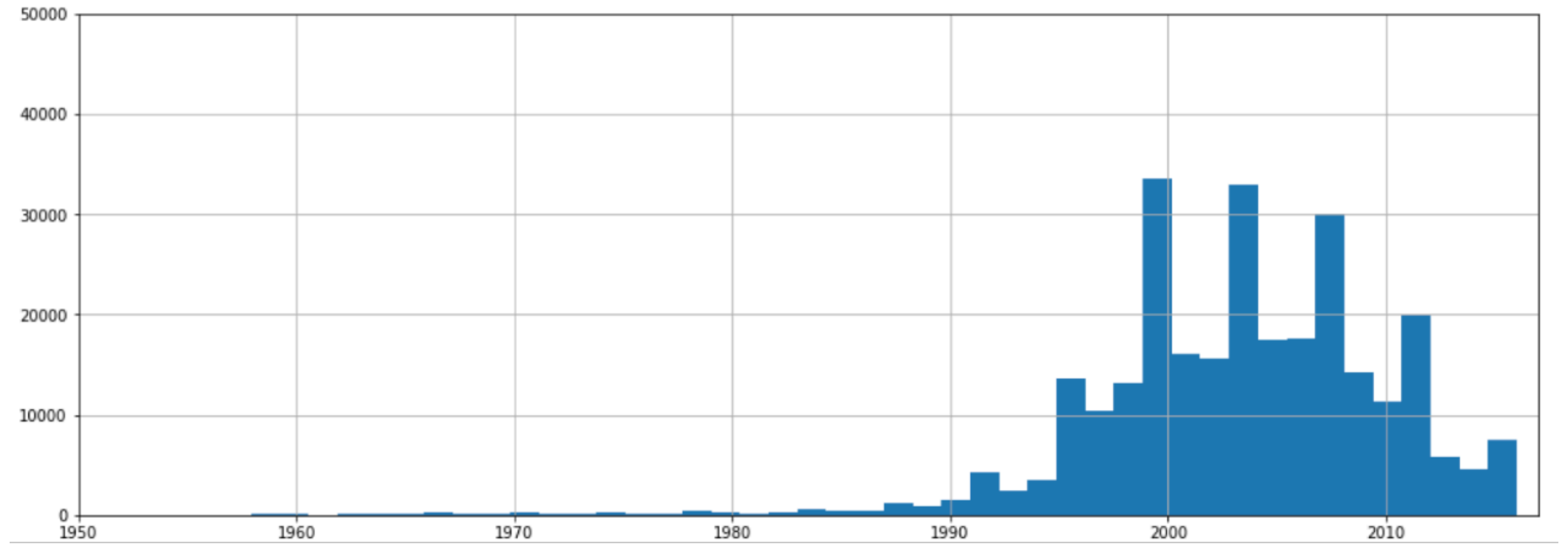
	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration
count	3.715280e+05	371528.000000	371528.000000	371528.000000	371528.000000
mean	1.729514e+04	2004.577997	115.549477	125618.688228	5.734445
std	3.587954e+06	92.866598	192.139578	40112.337051	3.712412
min	$\boxed{0.000000e+00}$	$\boxed{1000.000000}$	$\boxed{0.000000}$	5000.000000	0.000000
25%	1.150000e+03	1999.000000	70.000000	125000.000000	3.000000
50%	2.950000e+03	2003.000000	105.000000	150000.000000	6.000000
75%	7.200000e+03	2008.000000	150.000000	150000.000000	9.000000
max	$\boxed{2.147484e+09}$	$\boxed{9999.000000}$	$\boxed{20000.000000}$	150000.000000	12.000000

STEP 1: DATA CLEANING

➡ *Outliers: employed techniques*

1. **Year of registration:** remove where <1950 and >2016
   (further narrowed down to 2000-2016)



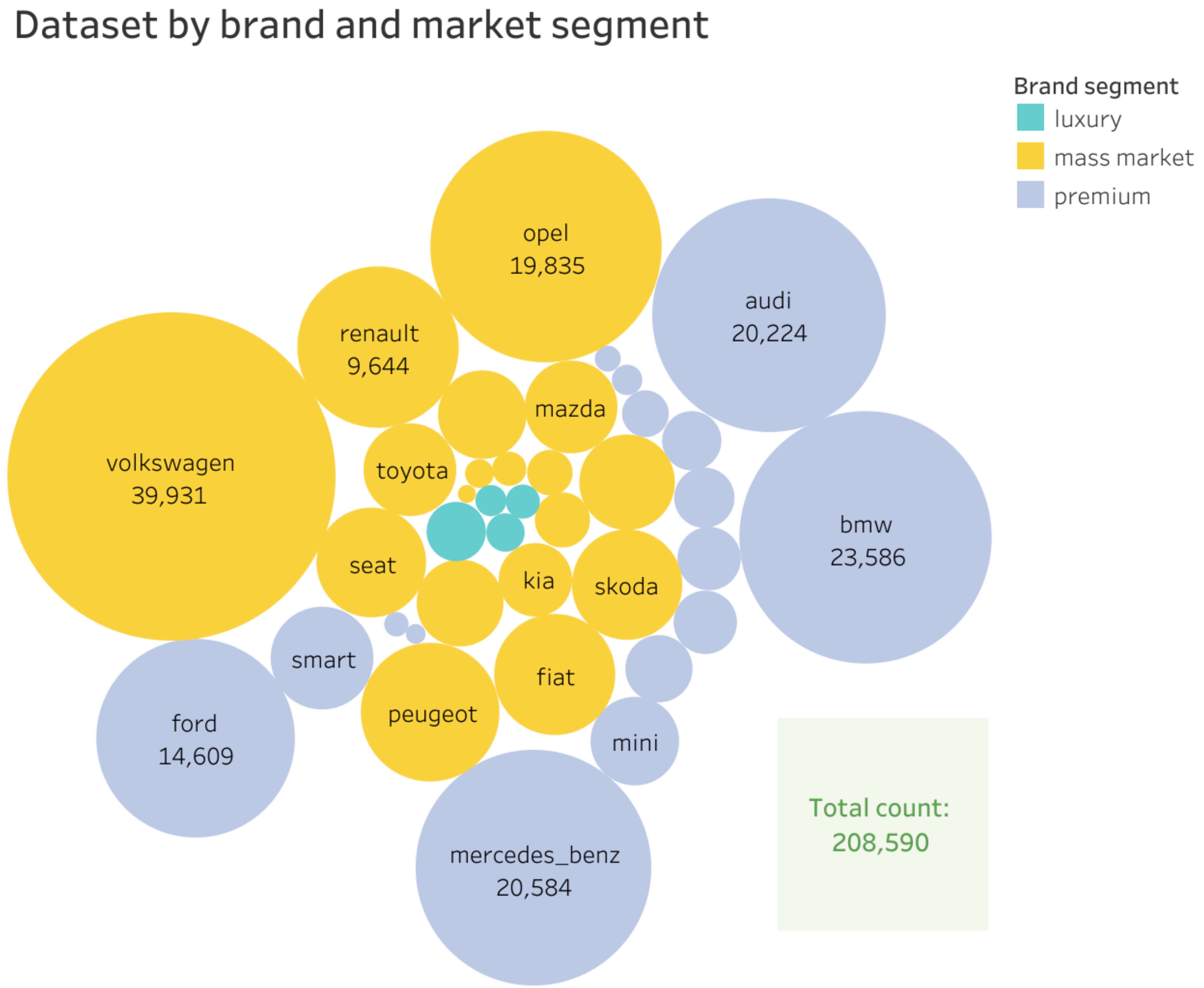


2. **Price:** remove where <100€ and >100,000€

3. **BHP:** replace all zero and >600 observations with a mean for the given model

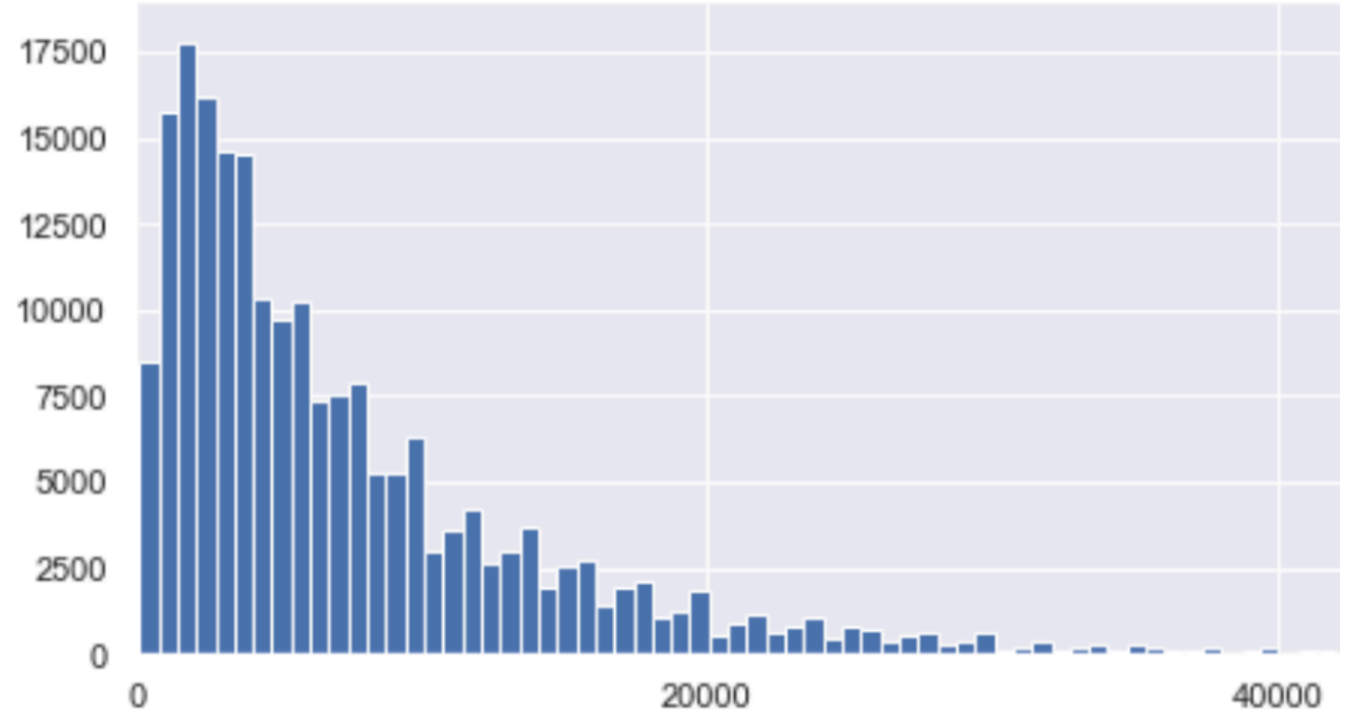
☑ *Size of dataset brought down to 208,590 records*

STEP 2: EXPLORATORY DATA ANALYSIS

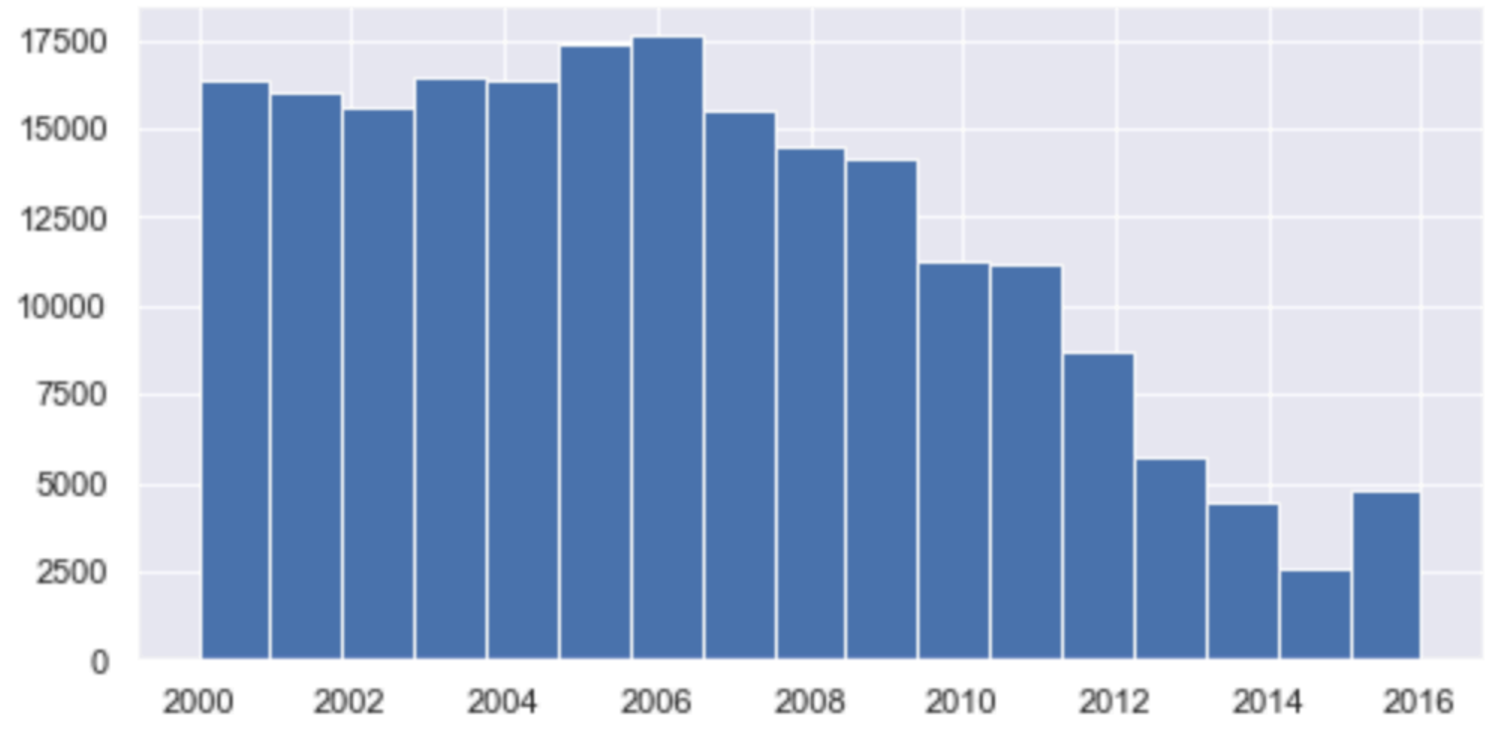


STEP 2: EXPLORATORY DATA ANALYSIS

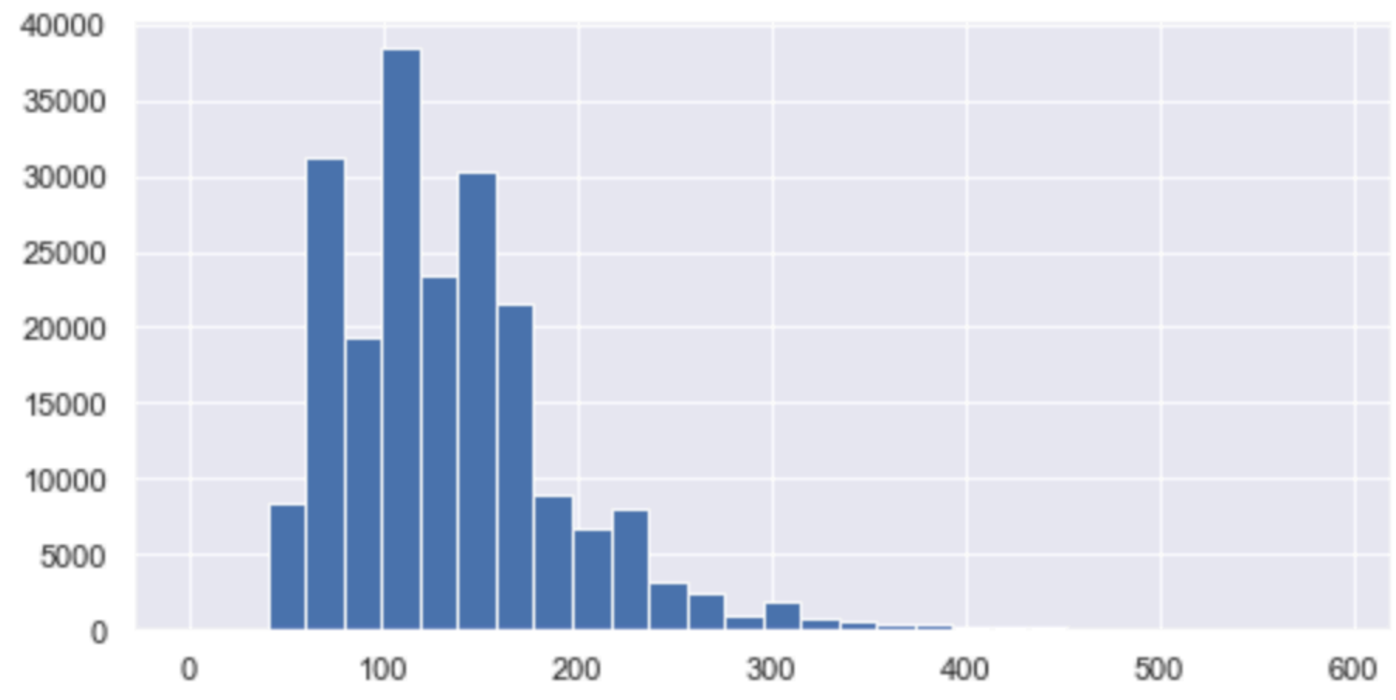
Price



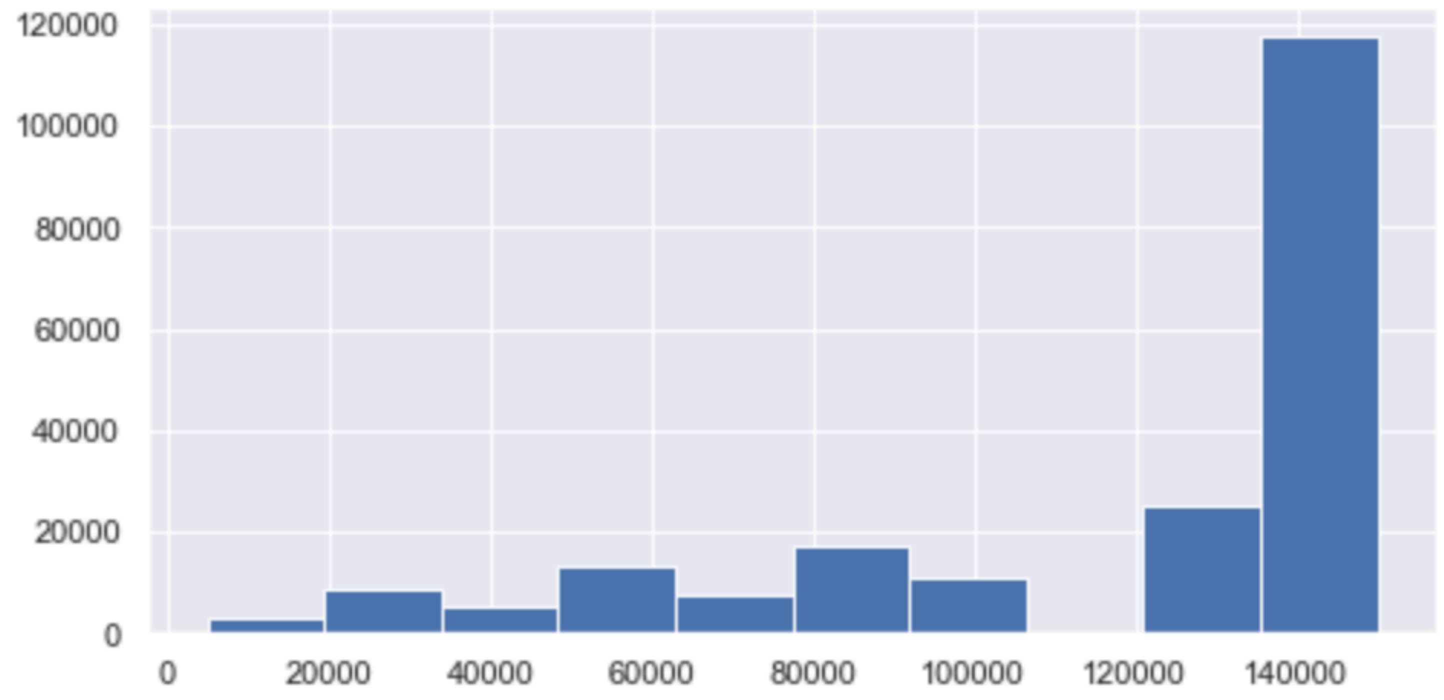
Year of registration



BHP

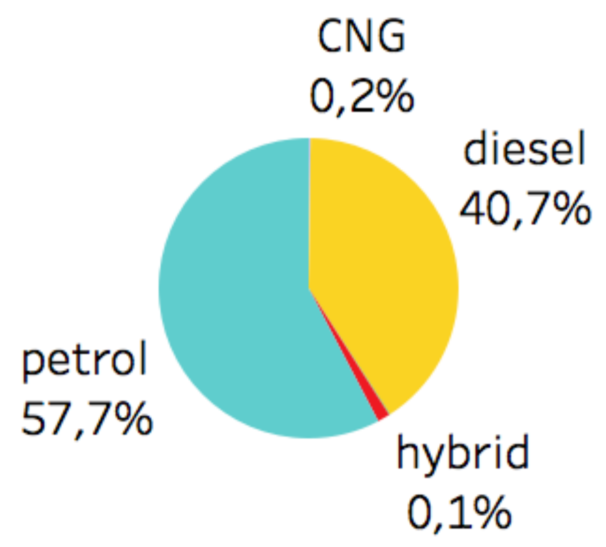


Mileage

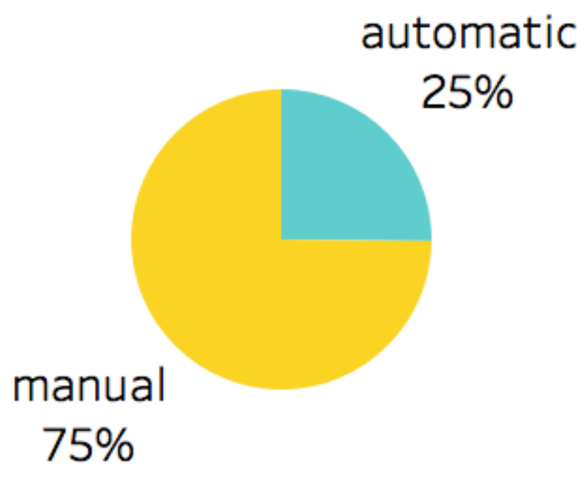


STEP 2: EXPLORATORY DATA ANALYSIS

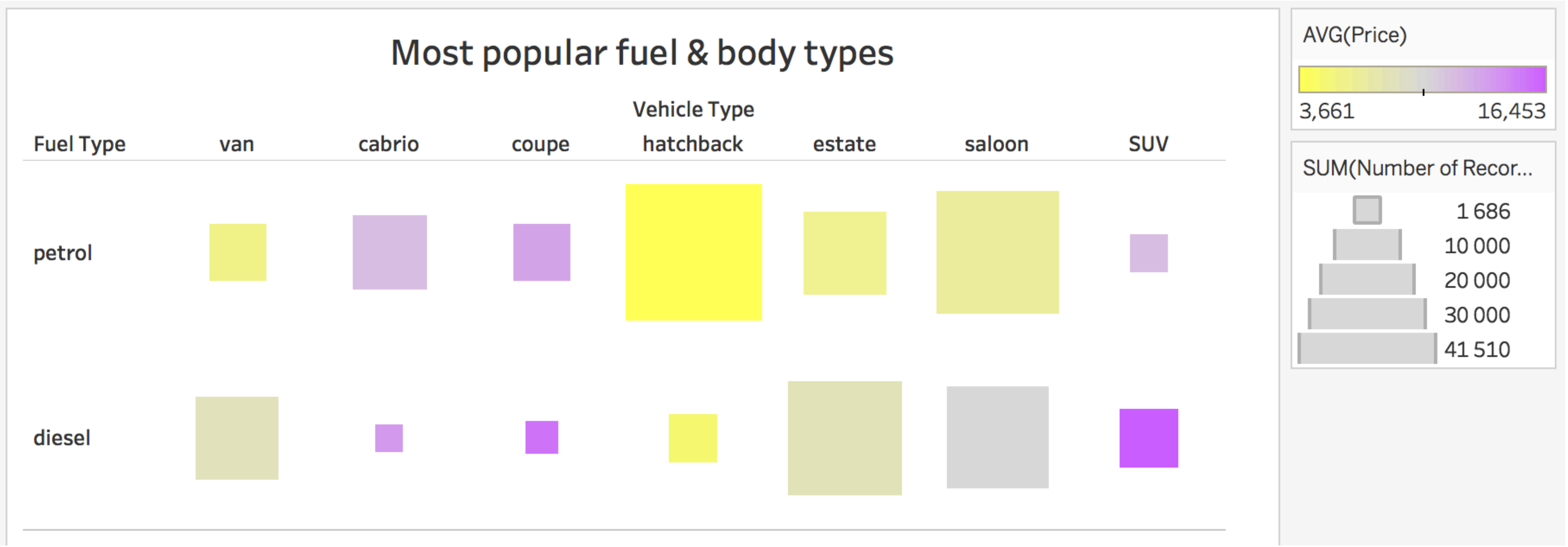
Fuel type



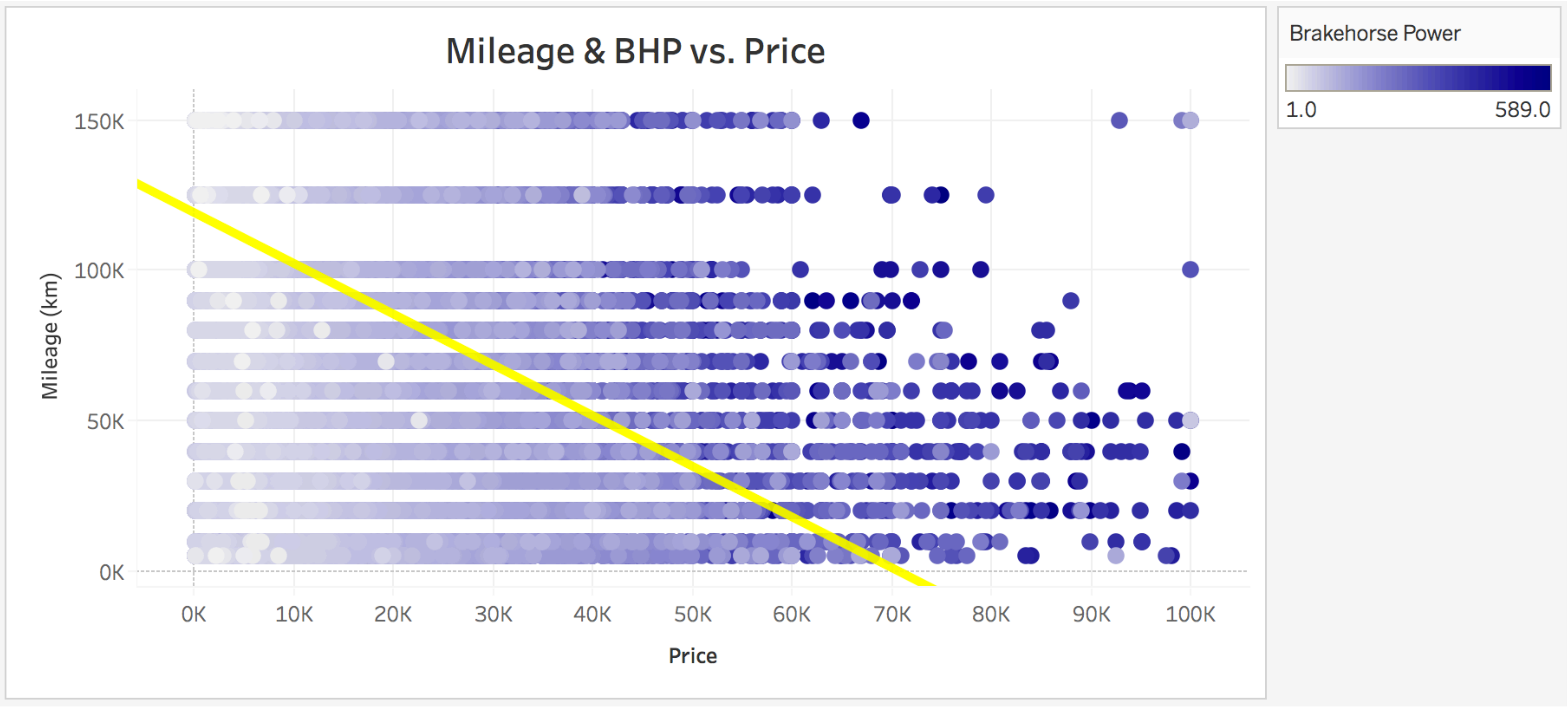
Transmission



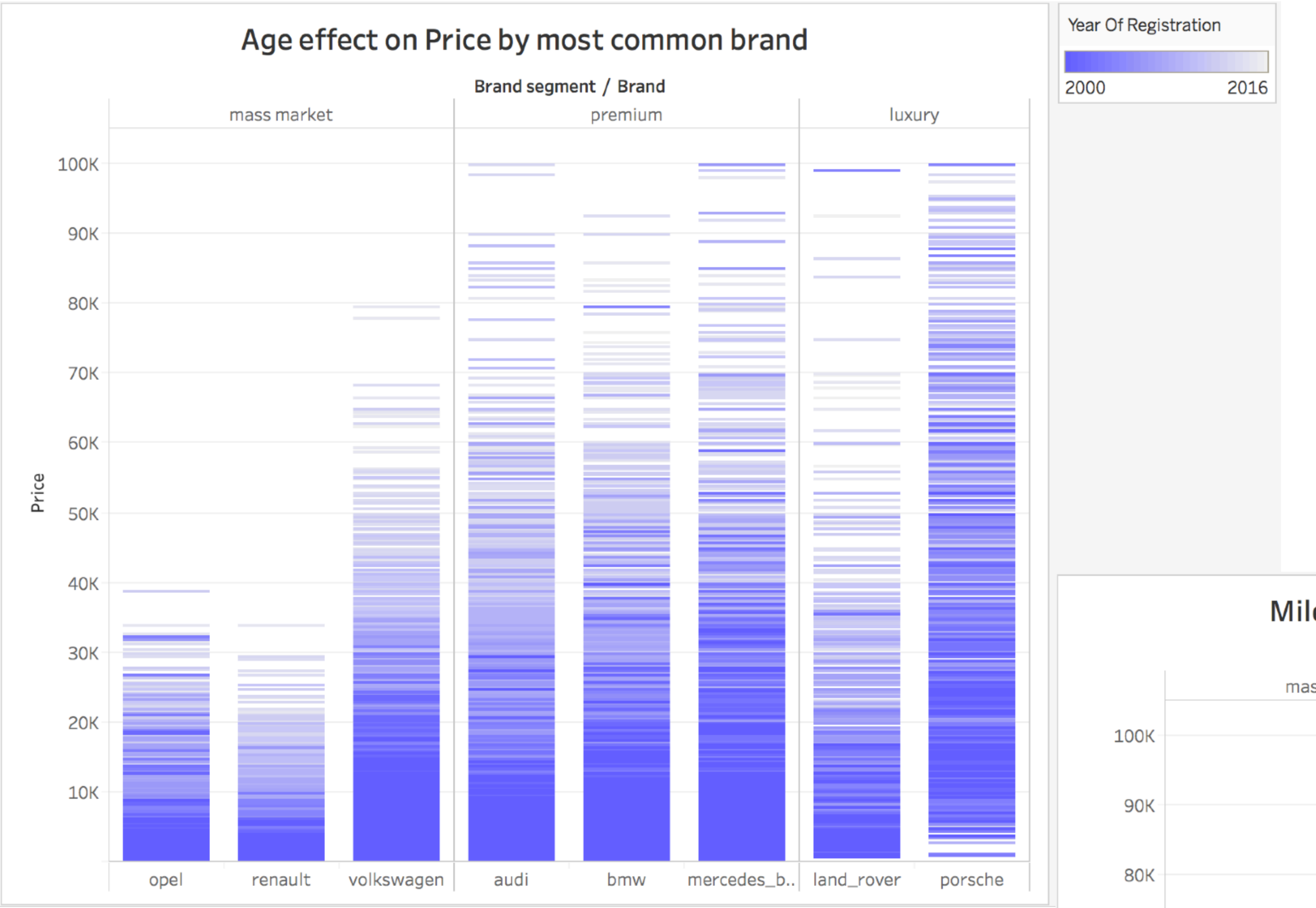
Unrepaired damage

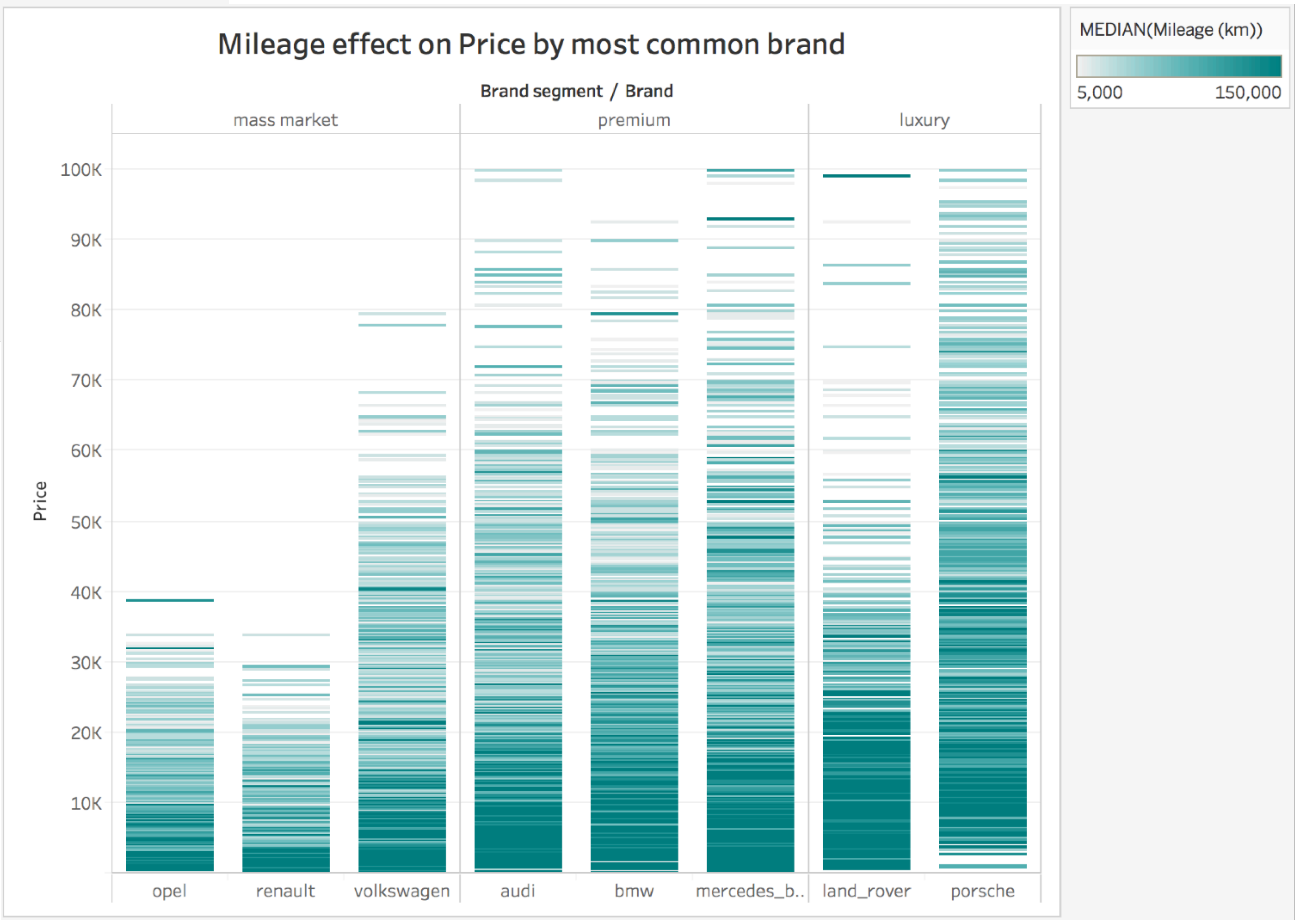


STEP 2: EXPLORATORY DATA ANALYSIS



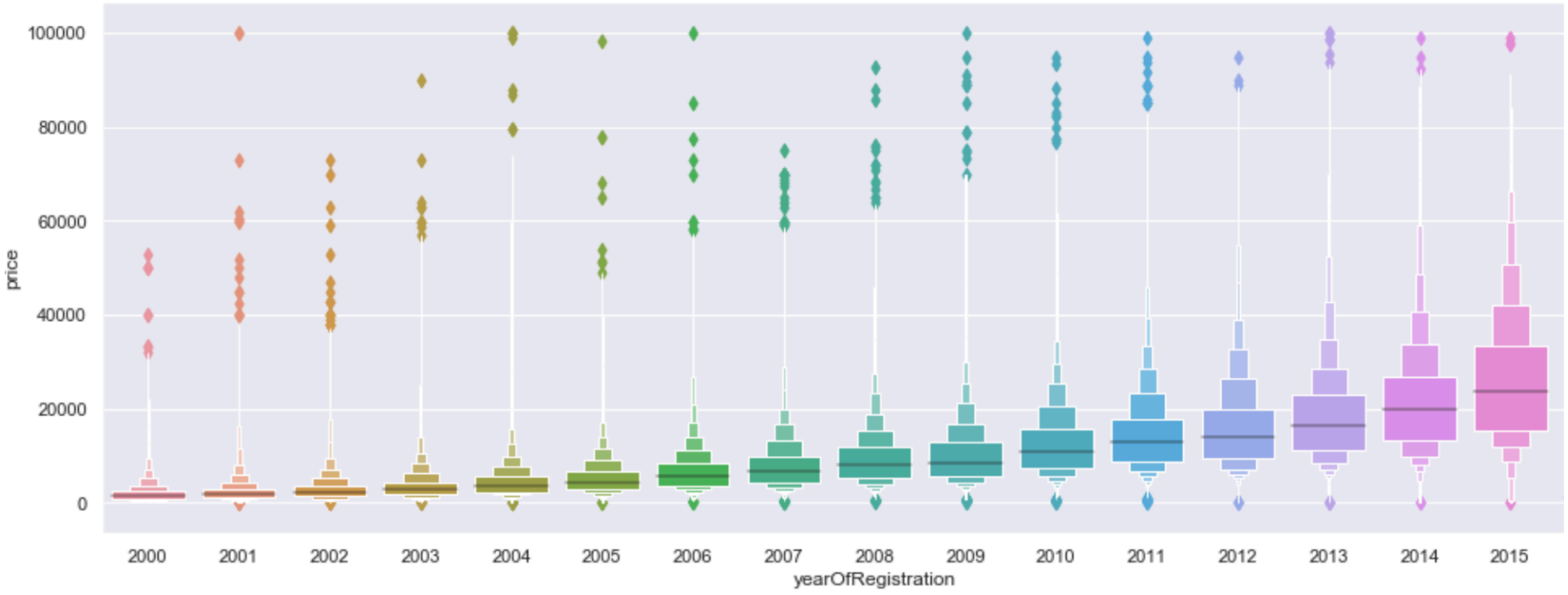
STEP 2: EXPLORATORY DATA ANALYSIS





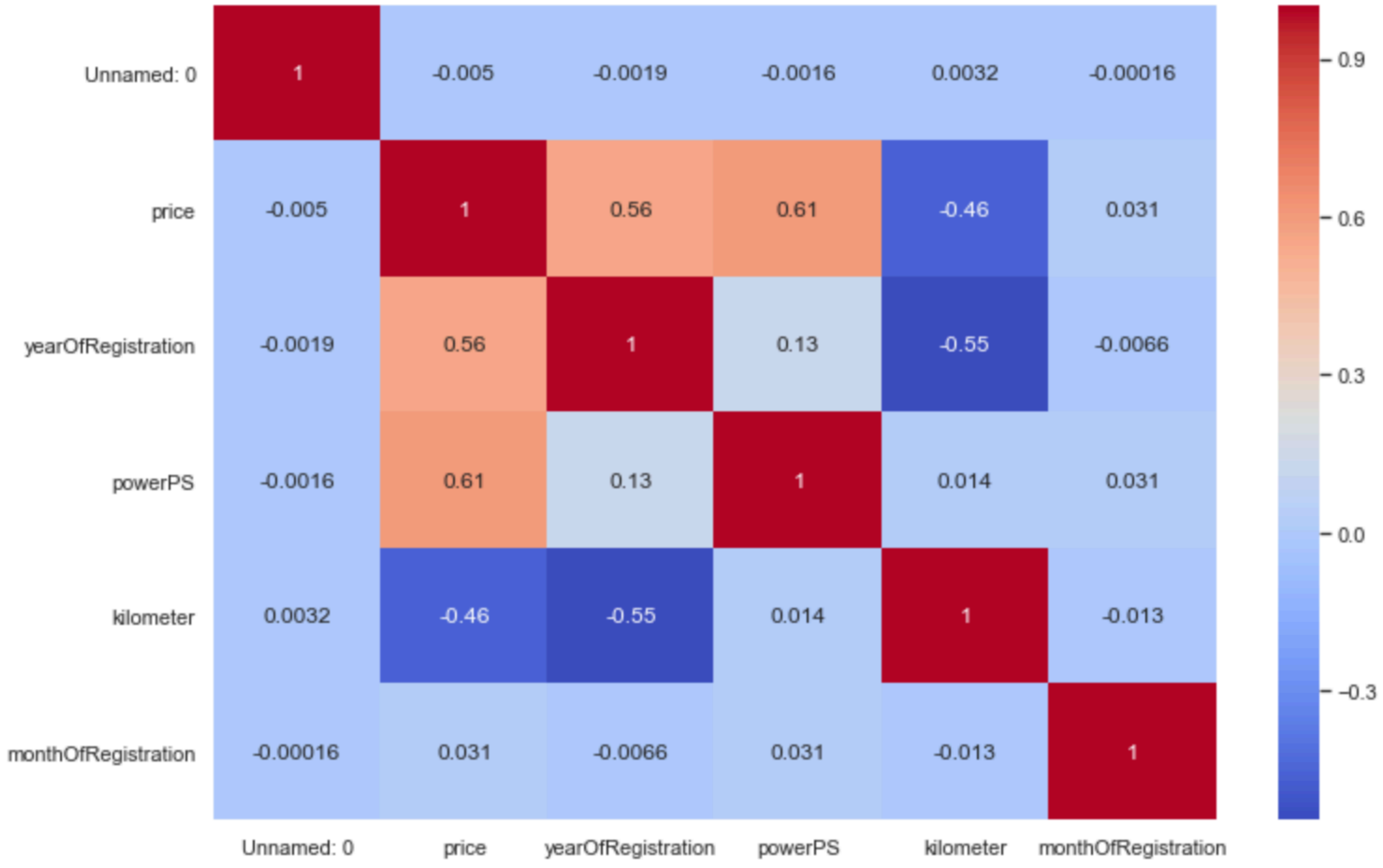
STEP 2: EXPLORATORY DATA ANALYSIS

Distribution of price by year of registration



STEP 2: EXPLORATORY DATA ANALYSIS

Correlations between variables



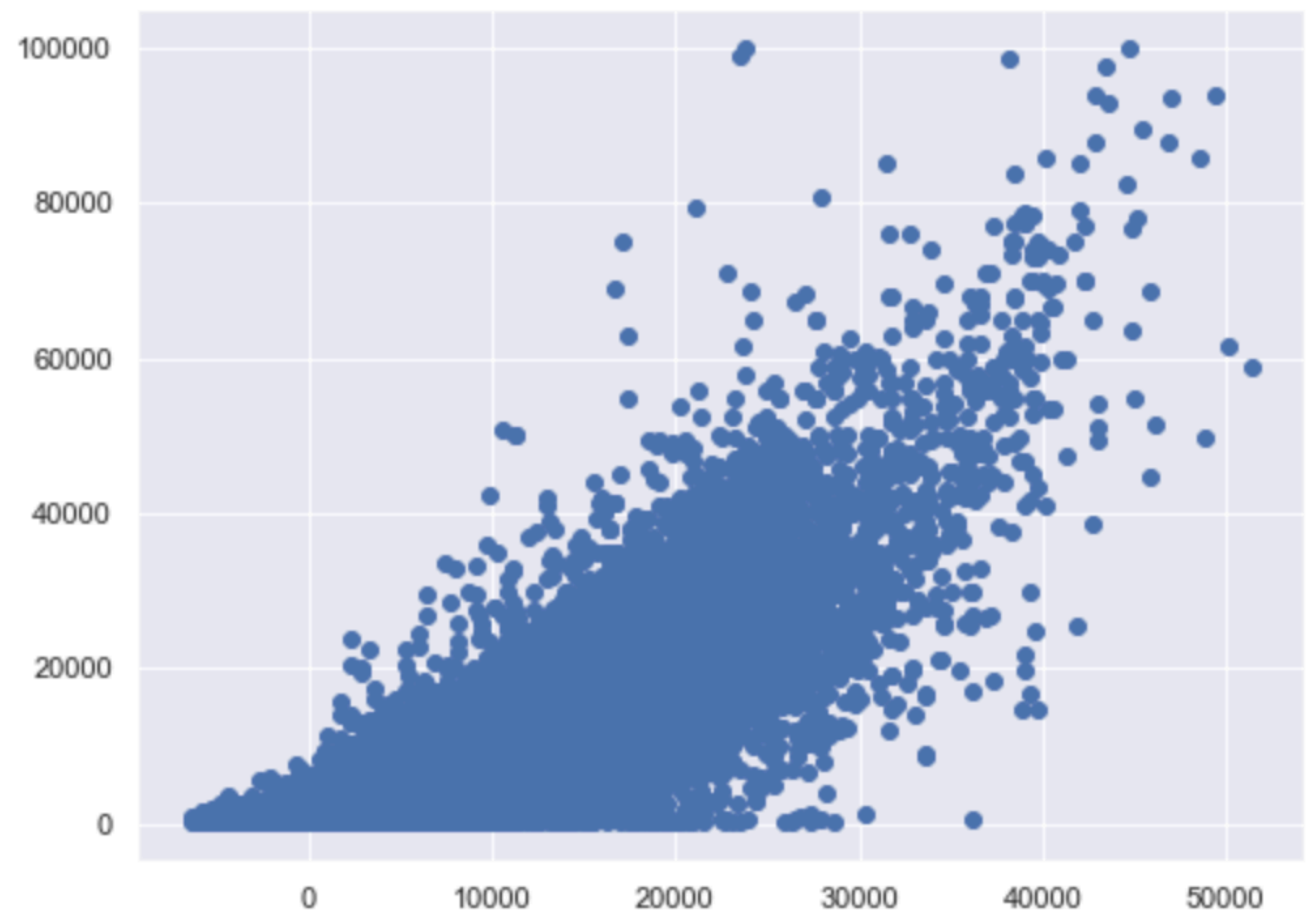
STEP 3: MODELLING

- The applied Multiple Linear Regression model takes into account all key variables
- The brands were split into three market segments: mass market, premium and luxury
- The R-squared is 0.70 which indicates good fit
- The weight of the respective coefficients confirms the trends and influences observed through visual analysis of the dataset

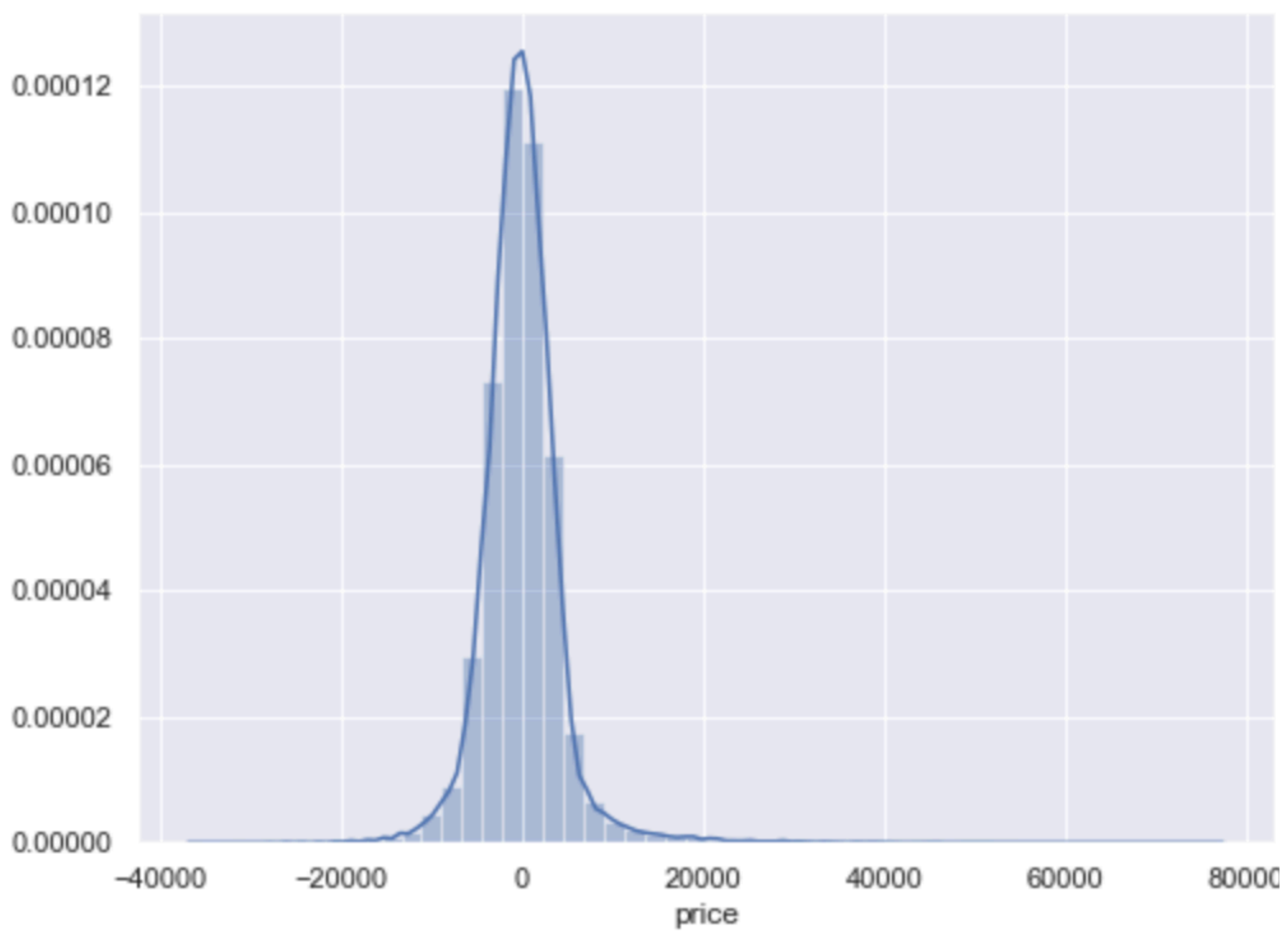
Independent variable	Coefficient
Intercept	11428
brake horsepower	64
mileage (km)	-0,06
Age (years)	-578
unrepaired damage	-2063
Brand segment	
luxury	8396
premium	568
mass market	0
Body type	
cabriolet	1774
coupe	1179
SUV	783
van	493
hatchback	290
saloon	-463
estate	-829
Transmission	
automatic	0
manual	-784
Fuel type	
hybrid	2202
diesel	1990
petrol	0
elektro	-706
lpg	-1118

STEP 3: MODELLING

Predicted vs. observed price



Error distribution



STEP 3: MODELLING

- Now let's sell your car!

THANK YOU