Used Car Prices

Lukas Vogt and Philipp Thienel

November 30, 2015

Table of Contents

1. Dataset and Assignment

Dataset:

Messy dataset that contains 1'170 records of a website selling used cars from July 2011. Along with the price there are 21 characteristics (kilometers, inverkehrssetzung, hubraum, etc.) of different VW station wagons (Golf, Passat, Bora, Caddy, Multivan).

1. Dataset and Assignment

Assignment:

1. Import and clean messy dataset
2. Descriptive analysis of the dataset
   What do you observe for prices?
   How have they potentially been sampled?
3. Multivariate analysis of the dataset: Regression Models
4. Most reasonable / best regression model?

2. Cleaning the dataset

Problem 1: The units are included in the data fields and the variable is thus of type character, while it should be of type integer or numeric.

Variables with Units

kilometer	verbrauch	leergewicht	co2.emission
26'200 km	8.3 l/100km	1570 kg	198 g/km
101'500 km	10 l/100km	1790 kg	240 g/km
113'000 km	6.2 l/100km	1613 kg	167 g/km
166'000 km	11 l/100km	1627 kg	266 g/km

2. Cleaning the dataset

Solution 1: To remove the units from the data we will define a function that takes a character vector as input and extracts the first numerical sequence in every element of the input vector.

```
GetValue <- function(x) {
  require(stringr)
  x <- gsub("'","",x)
  x <- str_extract(x, '[0-9]+\\.*[0-9]*')
  return(as.numeric(x))
}
```

2. Cleaning the dataset

Variables without Units

kilometer	verbrauch	leergewicht	co2.emission
26200	8.3	1570	198
101500	10.0	1790	240
113000	6.2	1613	167
166000	11.0	1627	266

2. Cleaning the dataset

Further problems:

1. Extracting the plattfrom from the modell
2. Variables are encoded as integers (hubraum, t?ren, etc.), while they should rather be treated as factors.
3. Calculating the age of the vehicle from the variable 'inverkehrssetzung'.

3. Descriptive Analysis

1. **Number of observations and variables**

There are 1'170 observations of 24 variables

2. **Types of Variables**

There are four types of variables: Integer, Numerical, Factor, and Logical

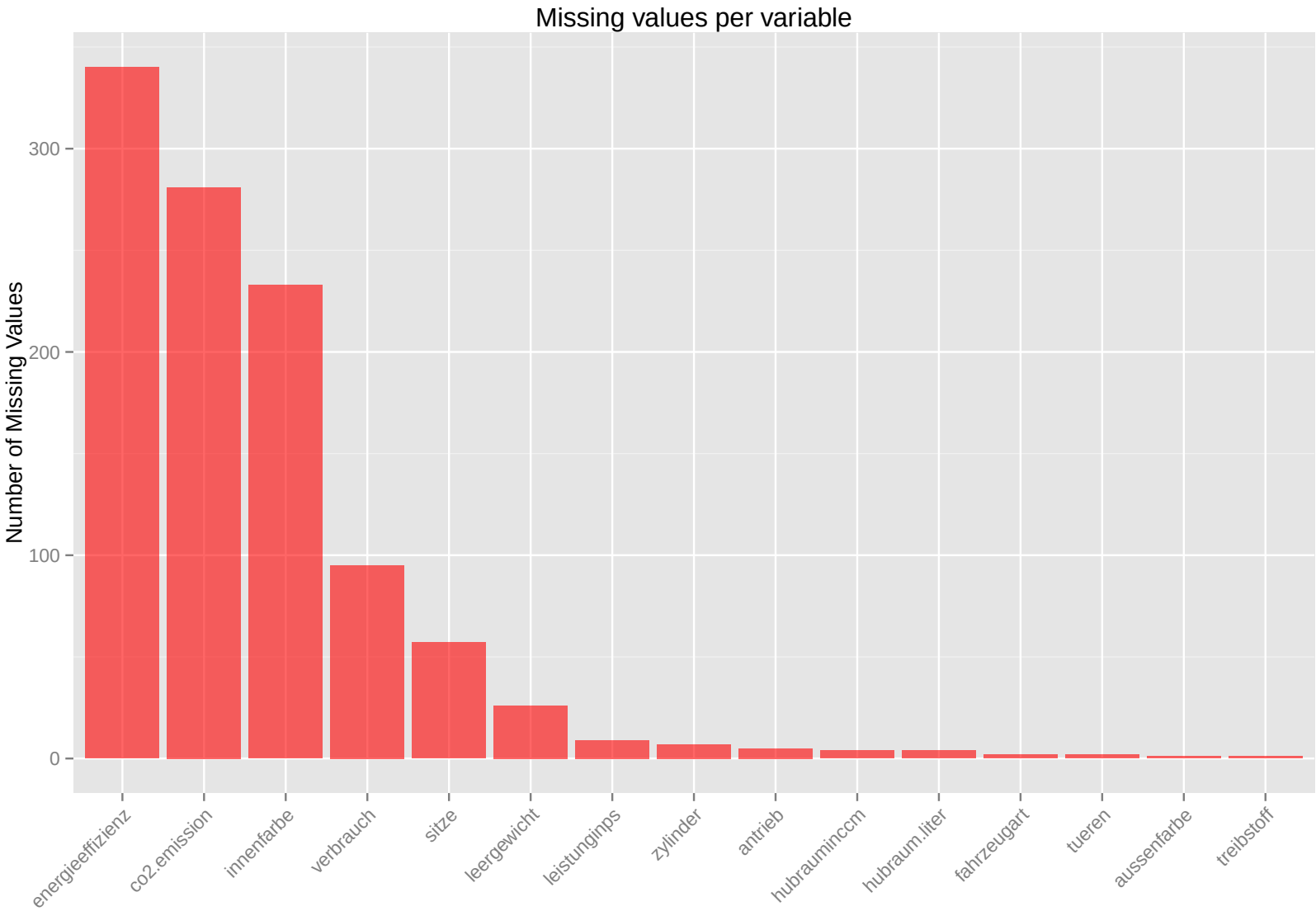
kilometer	numeric
leistunginps	integer
abmfk	logical
plattform	factor

3. **Missing Values of Variables**

Some variables have missing values

	variable	count.na
18	energieeffizienz	340
17	co2.emission	281
11	innenfarbe	233
16	verbrauch	95
10	sitze	57
15	leergewicht	26

3. Descriptive Analysis



3. Descriptive Analysis

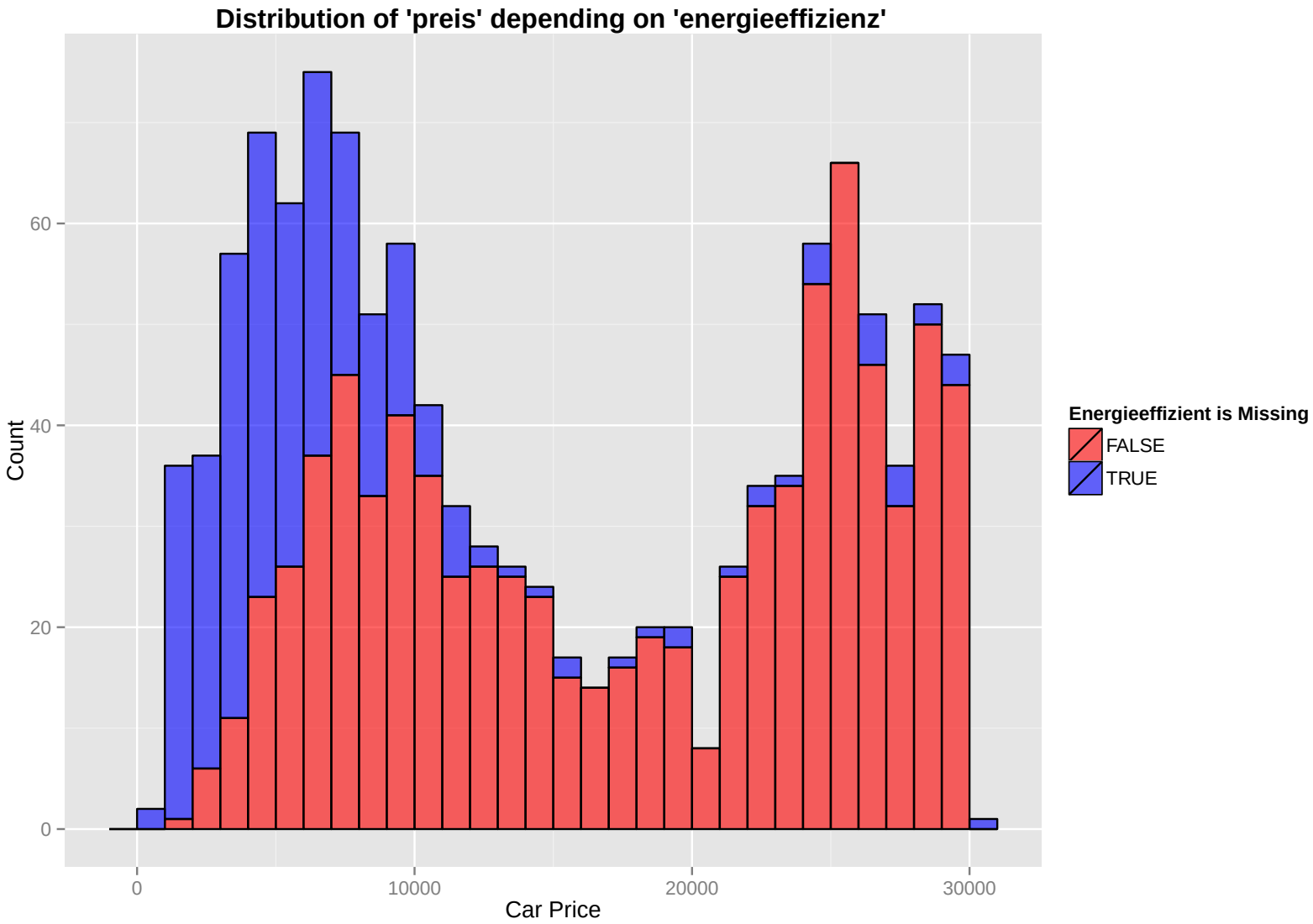
3. **Missing Values of Variables: Selection Bias**

Would we introduce a selection bias if we exclude the observations where the variable 'energieeffizienz' is missing?

Yes, we would.

'Energieefizienz' is missing disproportionally in lower ranges of the dependent variable 'preis'.

3. Descriptive Analysis



3. Descriptive Analysis

4. **Price: What do we observe?**

Summary Statistic

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
700	6800	11900	14790	24740	30000

3. Descriptive Analysis

4. **Histogram of Price: What do we observe?**

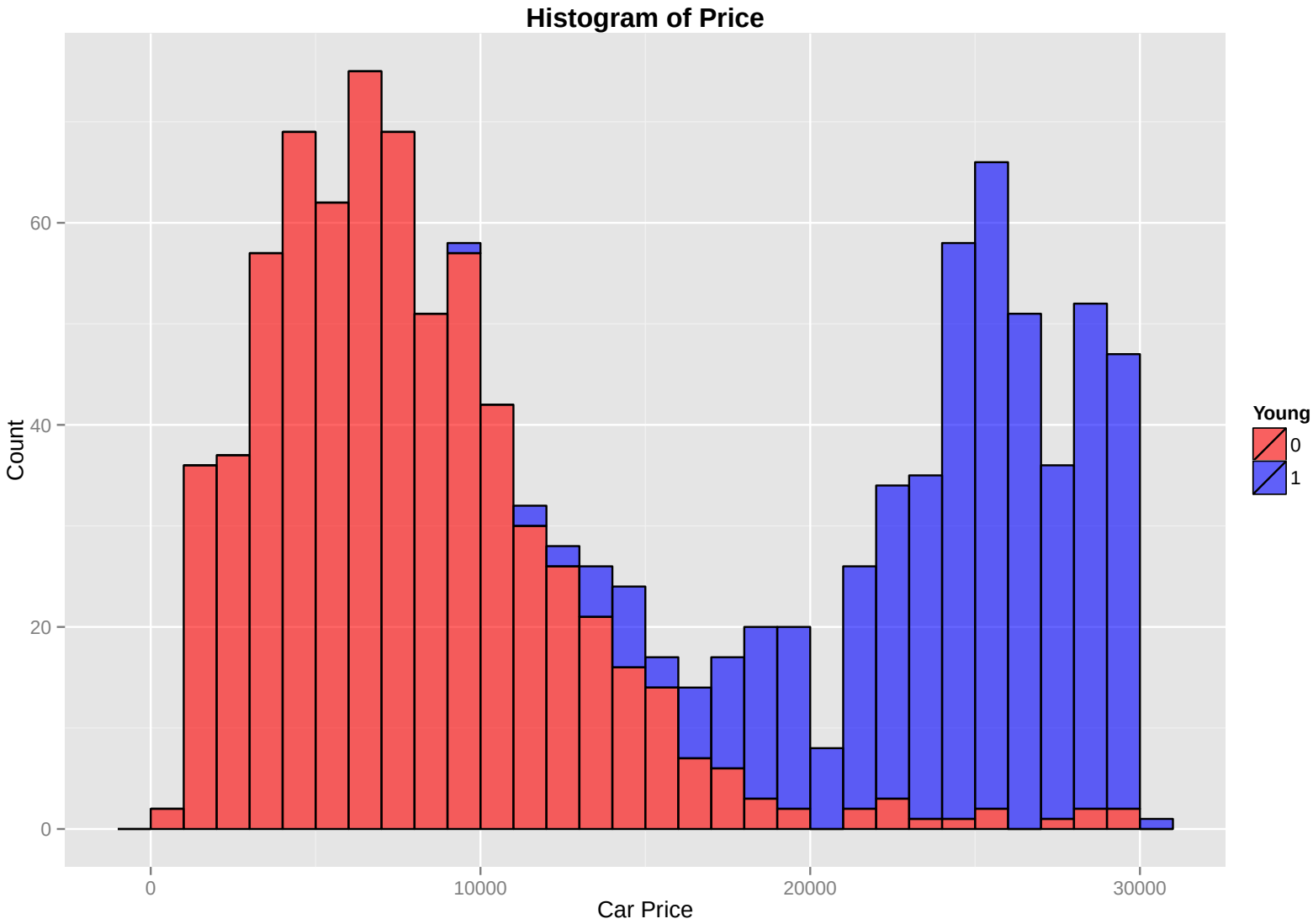
We observe a binominal distribution with two distinct peaks around CHF 5'000 and CHF 25'000

-Might be sampling problem: Selection bias

-Might be other economic reasons in the composition of population: Evidence that 'preis' depends strongly on 'age'

We add binary variable 'young' if age $<$ 60 months (5years) to show the dependency of 'preis' on 'age'

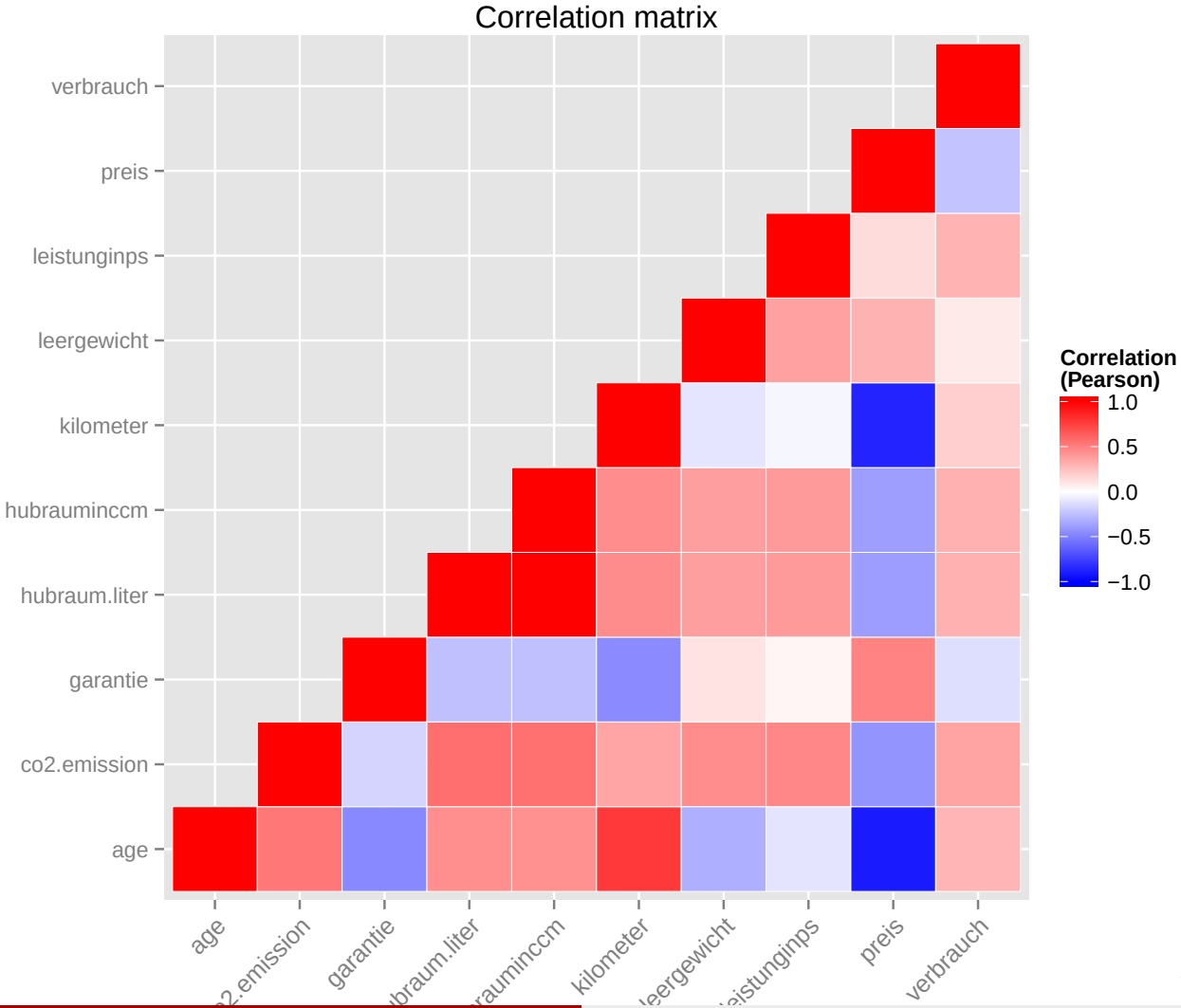
3. Descriptive Analysis



3. Descriptive Analysis

5. **Correlation Matrix**

Strong negative cor. of 'price' and 'age' is visibile in the cor. matrix.



4. Multivariate Analysis

Economic Rational of the Regression Variables:

-Age/kilometer/...: Independent variables expected to affect the dependent variable 'preis'. E.g. the older a car, the less its value.

-Square Root Age: Model the (expected) decreasing impact of age on the value of the car.

-Dummy Variables: Expected to affect the dependent variable 'preis'. E.g. cars with or without automatic transmission are not an identical product.

4. Multivariate Analysis

Three Regression Models:

1. **A linear model with level effects in age:**

preis $\sim$ sqrt(age) + kilometer

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32674.6508666	224.1265363	145.78662	0
sqrt(age)	-1632.5775884	40.9327372	-39.88440	0
kilometer	-0.0385959	0.0020389	-18.92934	0

4. Multivariate Analysis

2. **A linear model with level effects in age and dummies:**

preis $\sim$ sqrt(age) + kilometer + diesel + by4 + automatic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32044.5096598	238.9506062	134.105162	0e+00
sqrt(age)	-1619.1538289	41.5575916	-38.961686	0e+00
kilometer	-0.0412678	0.0020368	-20.260745	0e+00
diesel	900.1757104	191.6637011	4.696642	3e-06
by4	1699.2944817	250.4700818	6.784421	0e+00
automatic	1390.4879199	226.4838142	6.139458	0e+00

4. Multivariate Analysis

3. **A linear model with maximum fit:**

preis $\sim$ sqrt(age) + kilometer + leistunginps + garantie + leergewicht + diesel + by4 + automatic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21805.8051531	919.6837199	23.710113	0.0000000
sqrt(age)	-1540.5577248	40.0680327	-38.448549	0.0000000
kilometer	-0.0424727	0.0019018	-22.333204	0.0000000
leistunginps	25.7827023	3.7704325	6.838129	0.0000000
garantie	33.8425052	13.6472498	2.479804	0.0132900
leergewicht	4.4889445	0.6743046	6.657147	0.0000000
diesel	693.6746097	211.2882301	3.283073	0.0010582
by4	1055.4720415	233.8242236	4.513955	0.0000070
automatic	616.1787567	213.5599536	2.885273	0.0039848

4. Multivariate Analysis

Goodness of Fit:

Standard measure to evaluate if a model fits well is the so called R Squared (R^2). It measures how much of the total variation in the dependent variable can be explained by the model.

name	r.squared
Fit1	0.8902183
Fit2	0.9003334
Fit3	0.9210243

4. Multivariate Analysis

Predicitve Quality of a Model:

Common measure for the predictive quality of a model is the Root Mean Squared Error. Our Approach:

-We split the existing dataset randomly into a training and a test dataset

-We optimize (train) our models on the training dataset

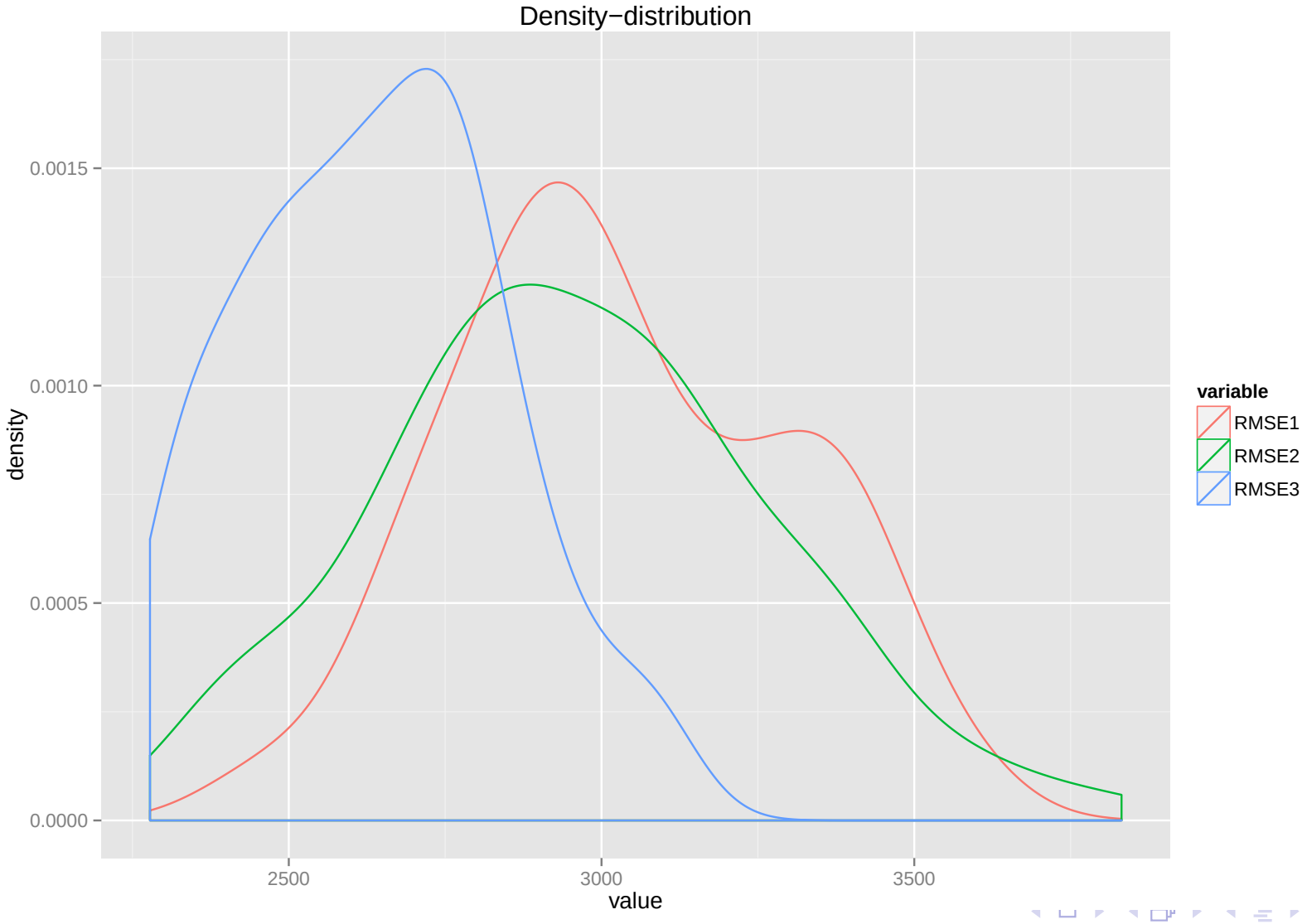
-Then we use the models to predict the 'preis' on the test dataset

-We compare the prediction with the actual values and calculate the Root Mean Squared Error

-Output the densitiy distributions of the Root Mean Squared Error

4. Multivariate Analysis

Predicitve Quality of a Model:



4. Multivariate Analysis

Most reasonable / best regression model?

-As we can see from the density plot, the last model exhibits the best (lowest) RMSE. It has also the highest fit to the data (R^2). Thus, we can conclude that it is the best model.

-Both the first and second regression model (without and with dummy variables) are very reasonable and have a high R^2 (predictive power).

-The fit to the data (R^2) of the 'best' model is only slightly higher than for the other two models.

5. I-Wish-I-Had-Known-Before

Two useful functions:

1. R Markdown: When you generate tables with kable() (libarary(knitr)) the format must be set to “markdown”,i.e. kable(. . . ,format=“markdown”)

2. Regression: Built in function methods(class=“”),
e.g. methods(class=“lm”)