Data Crunchers

Used Car Explaratory Data Analysis

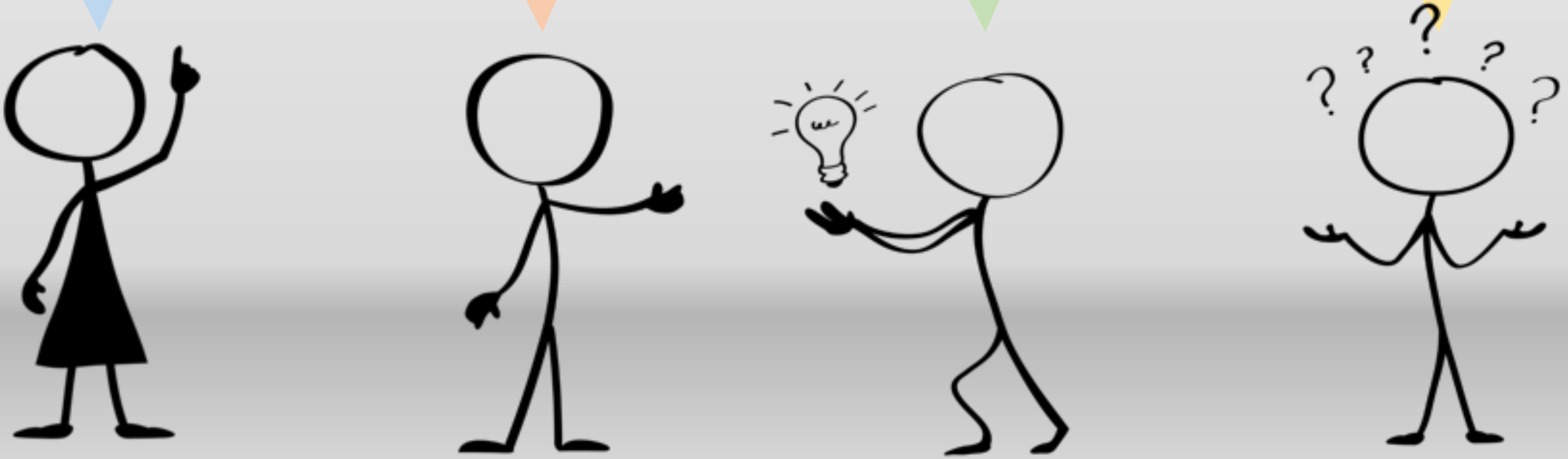
MEF University Big Data Analytics

Gülce Malkarali
Gecenin 4’ü oldu ama yine de bir şey söyleyeceğim….

Selimcan Yilmaz
Ama bu data bozuk….

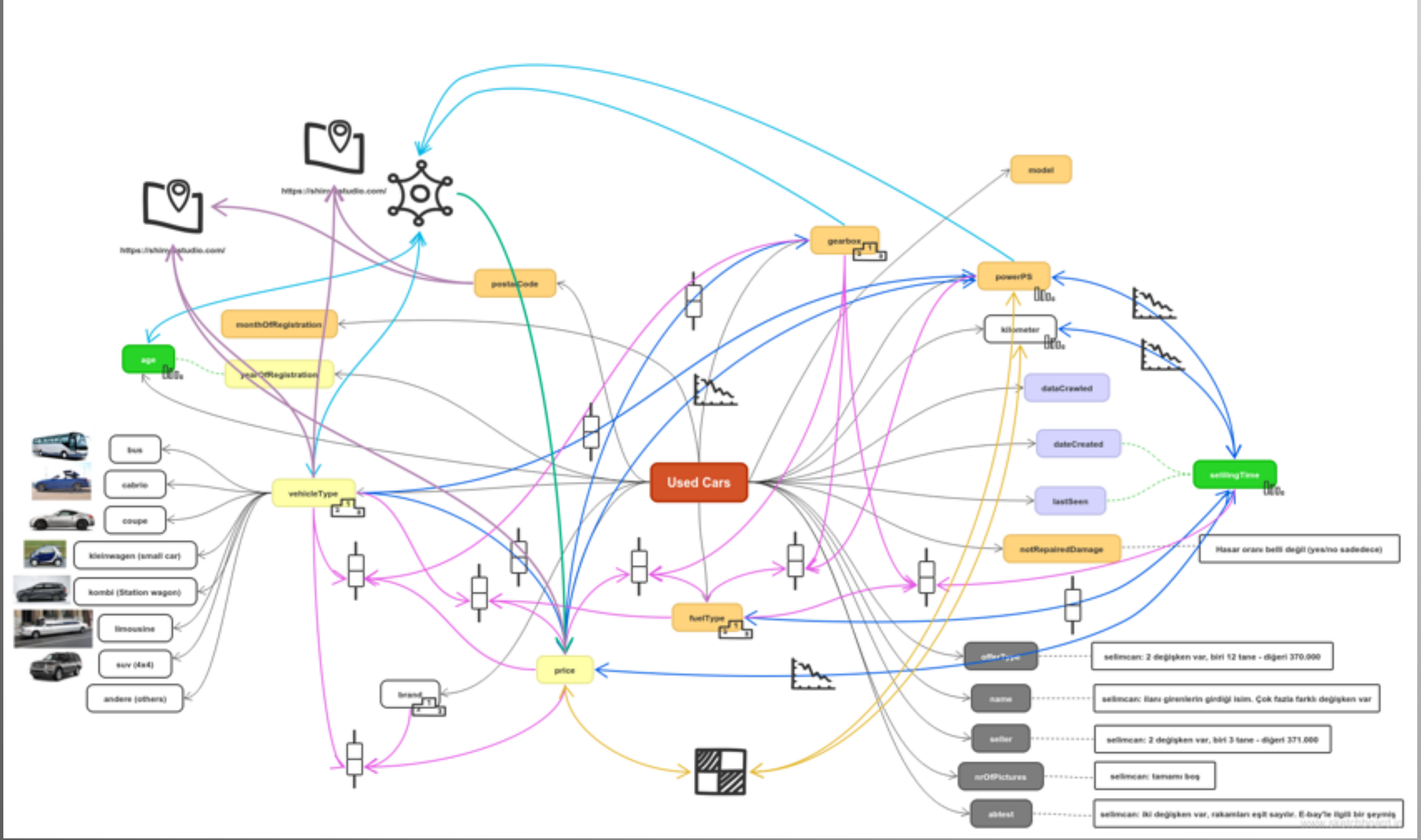
Serkan Ceran
Bu saatte bunu farketmemiz iyi oldu…

Timucin Anuslu
Dur dur şuna da bakmamız lazım….



Data Crunchers Team

Overall View of Data



Hasar oranı belli değil (yes/no sadedece)

selimcan: 2 değişken var, biri 12 tane - diğeri 370.000

selimcan: Ilanı girenlerin girdiği isim. Çok fazla farklı değişken var

selimcan: 2 değişken var, biri 3 tane - diğeri 371.000

selimcan: tamamı boş

selimcan: iki değişken var, rakamları eşit sayılır. E-bay'le ilgili bir şeymiş

Overall View of Data

Used Car Database' scraped from Ebay Kleinanzeigen (in German)

370.000 second-hand cars

40 unique brands.

20 variables

Month of Registration

Year of Registration

Kilometer

Postal Code

~~Seller~~

Power PS

Price

Vehicle Type

Last Seen

~~Offer Type~~

~~AB Test~~

Model

Brand

Gear Box

~~Number of Pictures~~

Fuel Type

Not Repaired Damage

Date Created

Content

🟥 1

🟦 2

🟧 3

⬜ 4

🟩 5

🟦 6

🟩 7

🟧 8

Raw Data Visualization

Check missing data and other mistakes

Create a list of outliers or other anomalies

Data Cleaning

Clean Data Visualization & Mapping

Determine relationships among the variables

Regression Models

Maps

Content

1 • Raw Data Visualization

2 • Check missing data and other mistakes

3 • Create a list of outliers or other anomalies

4 • Data Cleaning

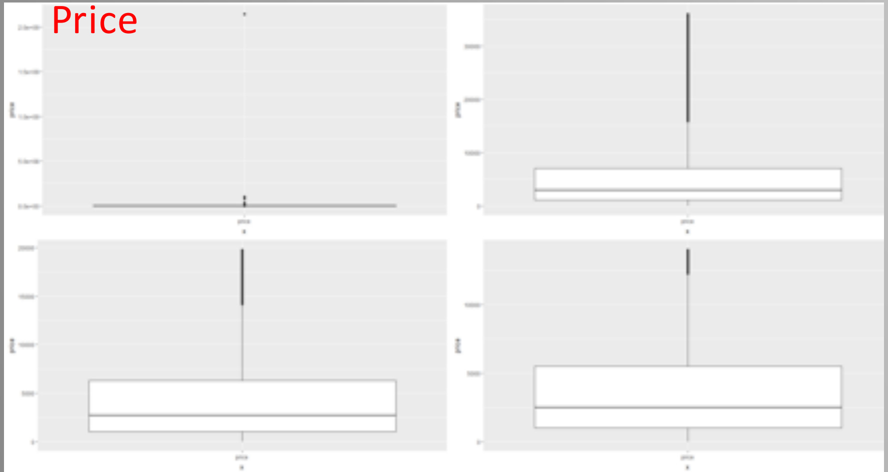
5

6

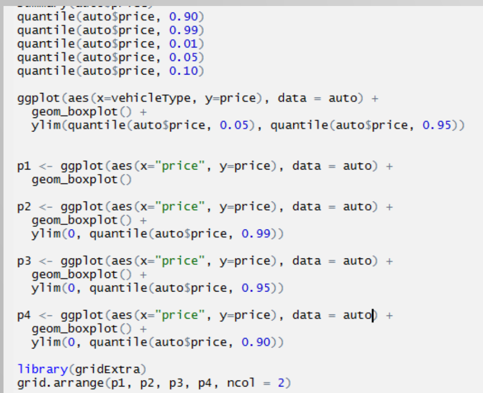
7

8

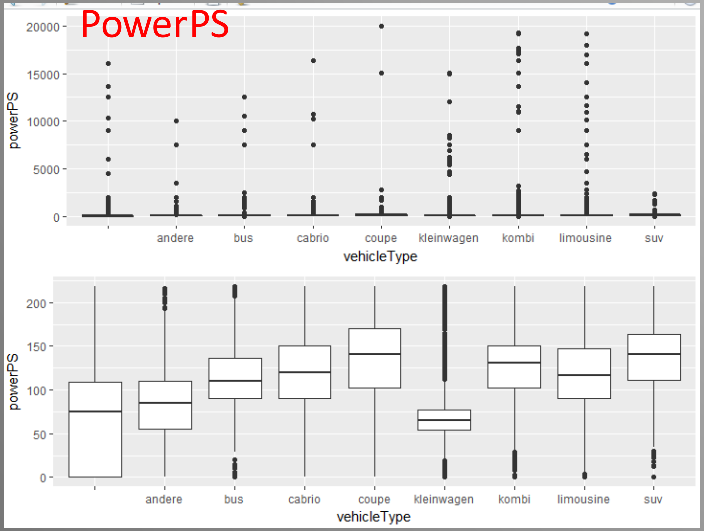
"This is not what I meant when I said 'we need better data cleansing!"



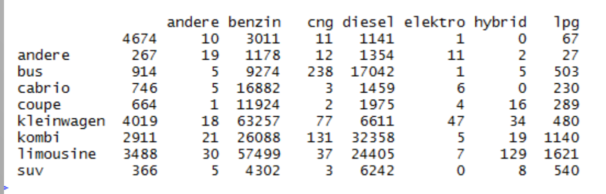
Price



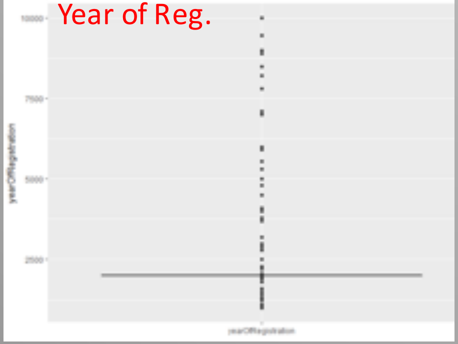
```
quantile(auto$price, 0.90)
quantile(auto$price, 0.99)
quantile(auto$price, 0.01)
quantile(auto$price, 0.05)
quantile(auto$price, 0.10)

ggplot(aes(x=vehicleType, y=price), data = auto) +
  geom_boxplot() +
  ylim(quantile(auto$price, 0.05), quantile(auto$price, 0.95))


p1 <- ggplot(aes(x="price", y=price), data = auto) +
  geom_boxplot()

p2 <- ggplot(aes(x="price", y=price), data = auto) +
  geom_boxplot() +
  ylim(0, quantile(auto$price, 0.99))

p3 <- ggplot(aes(x="price", y=price), data = auto) +
  geom_boxplot() +
  ylim(0, quantile(auto$price, 0.95))

p4 <- ggplot(aes(x="price", y=price), data = auto) +
  geom_boxplot() +
  ylim(0, quantile(auto$price, 0.90))

library(gridExtra)
grid.arrange(p1, p2, p3, p4, ncol = 2)
```



PowerPS



```
                 andere benzin   cng diesel elektro hybrid   lpg
            4674     10   3011    11   1141       1      0    67
andere       267     19   1178    12   1354      11      2    27
bus          914      5   9274   238  17042       1      5   503
cabrio       746      5  16882     3   1459       6      0   230
coupe        664      1  11924     2   1975       4     16   289
kleinwagen  4019     18  63257    77   6611      47     34   480
kombi       2911     21  26088   131  32358       5     19  1140
limousine   3488     30  57499    37  24405       7    129  1621
suv          366      5   4302     3   6242       0      8   540
>
```



Year of Reg.

~~~
```{r}
## clean empty cells for vehicletype
```{r clean empty cells for vehicletype }
summary(auto$vehicleType)
auto <- auto[vehicleType != ""]
summary(auto$vehicleType)

~~~
```
```

Content

1

2

3

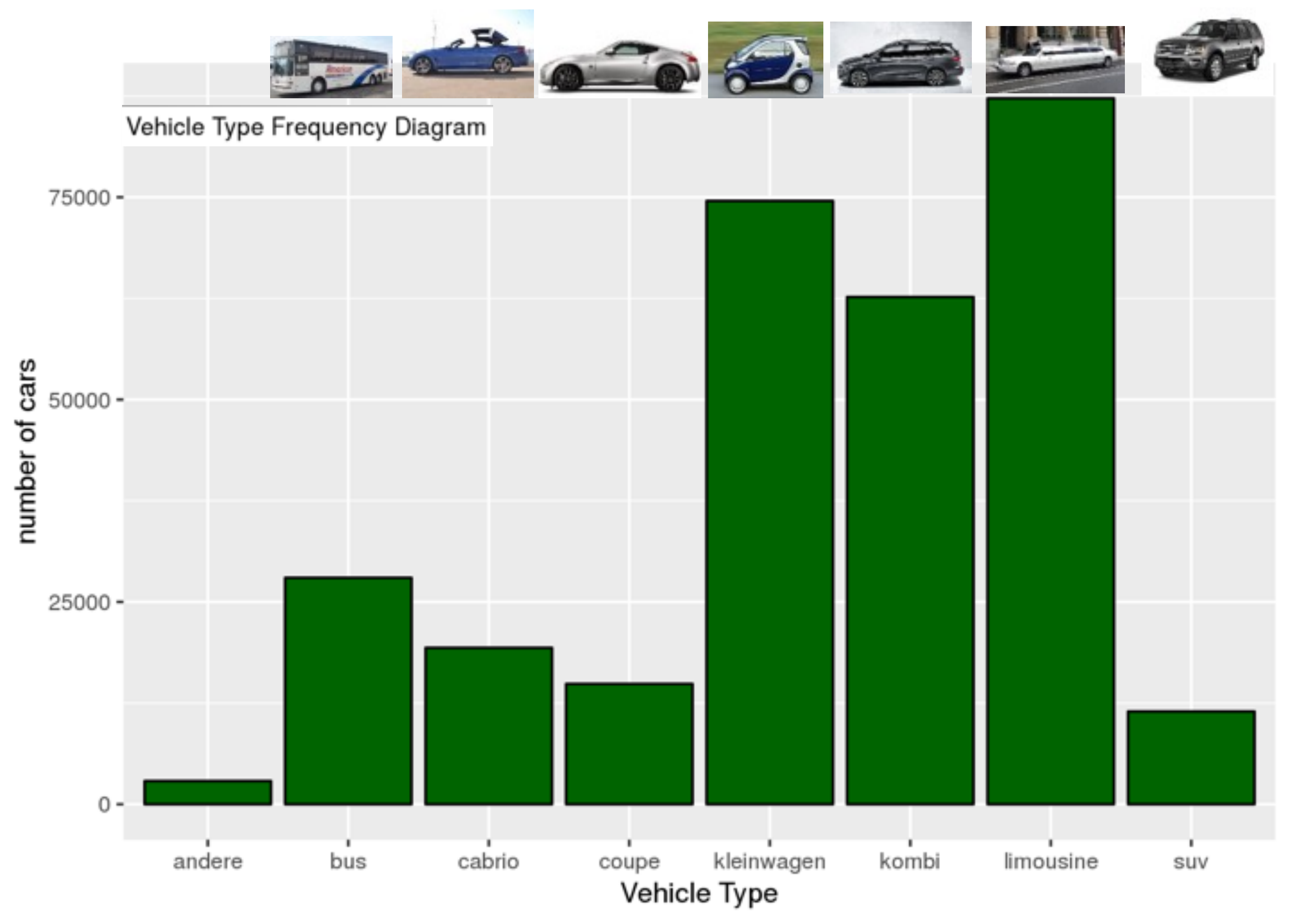
4

5 Clean Data Visualization

6

7

8



Content

1

2

3

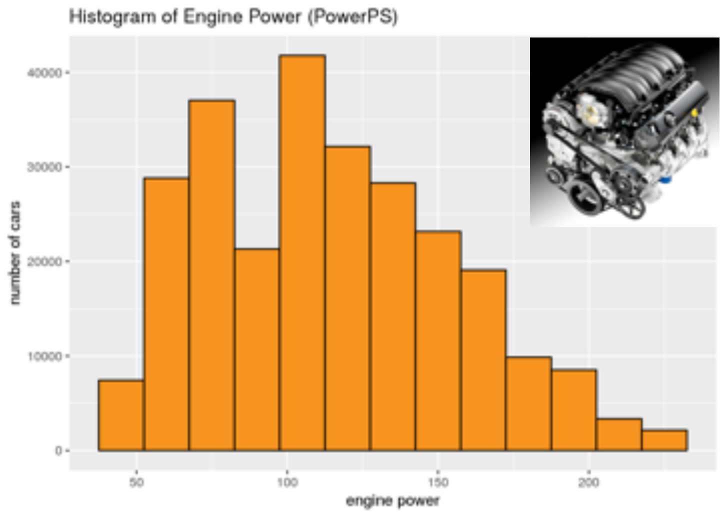
4

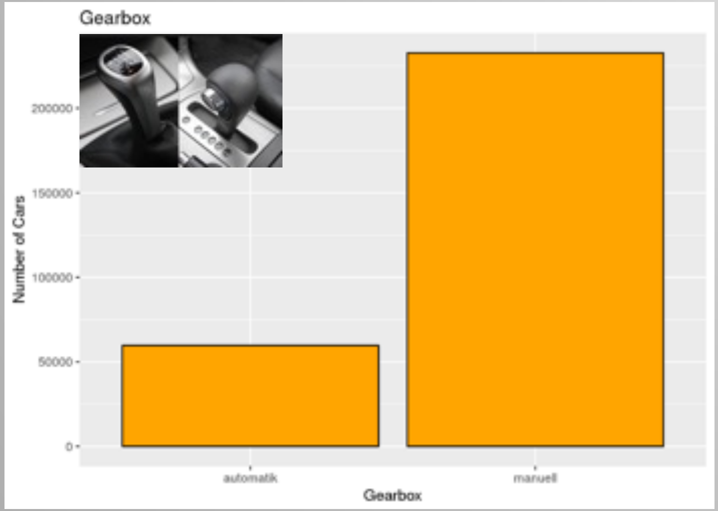
5 Clean Data Visualization

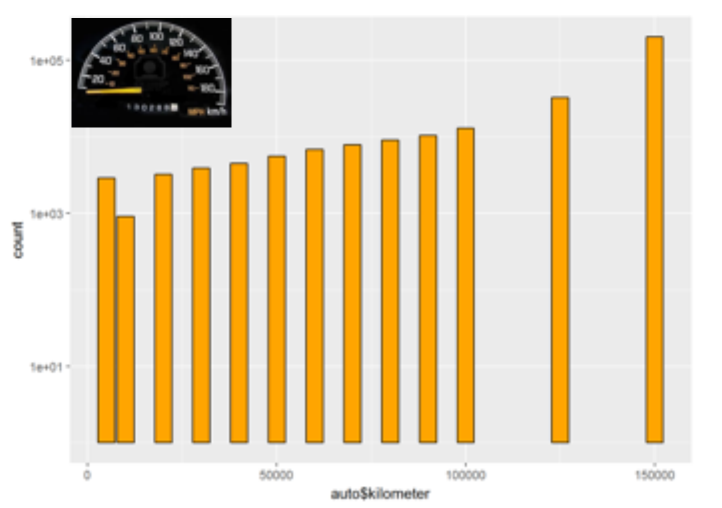
6

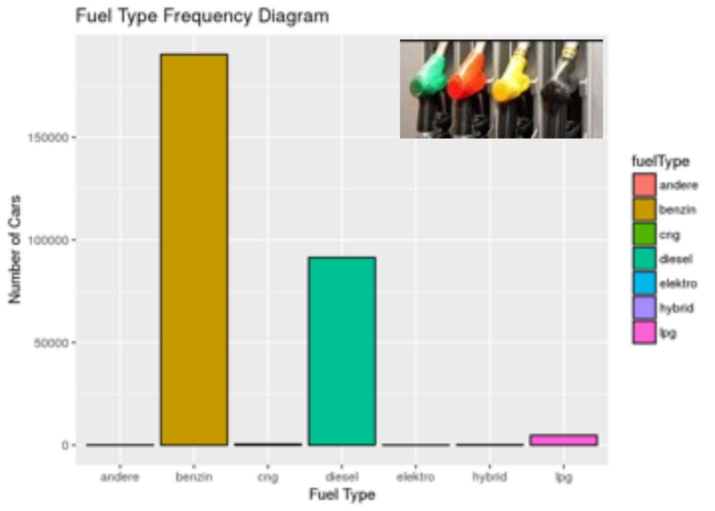
7

8









Content

🟥 1

🟦 2

🟧 3

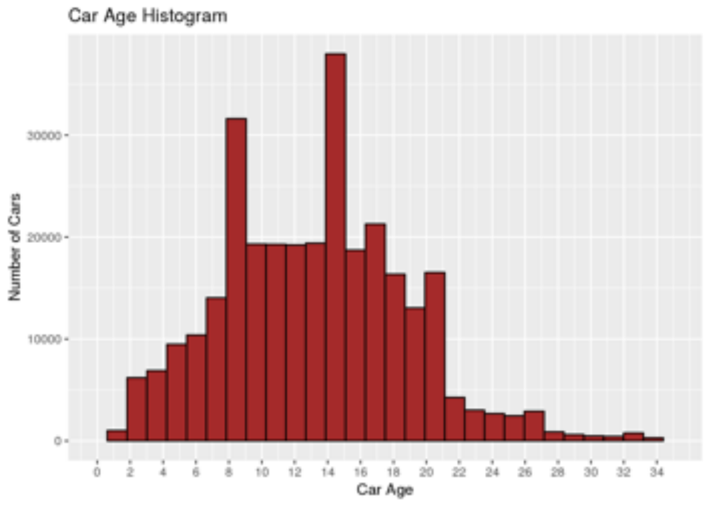
⬜ 4

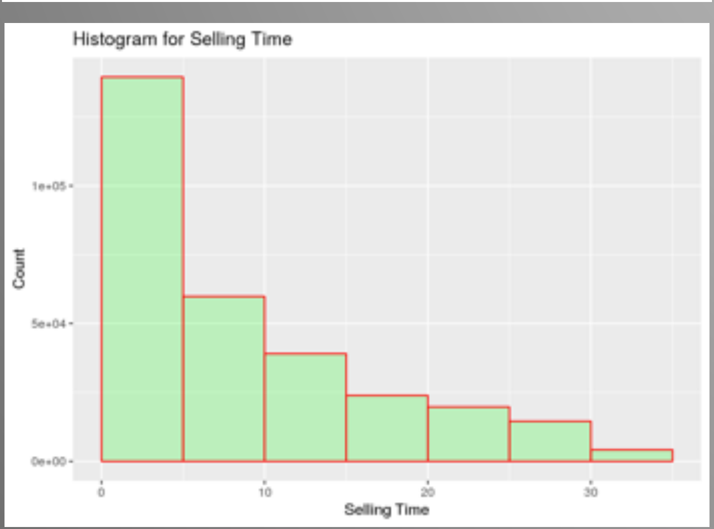
🟩 5 Clean Data Visualization

🟦 6

🟩 7

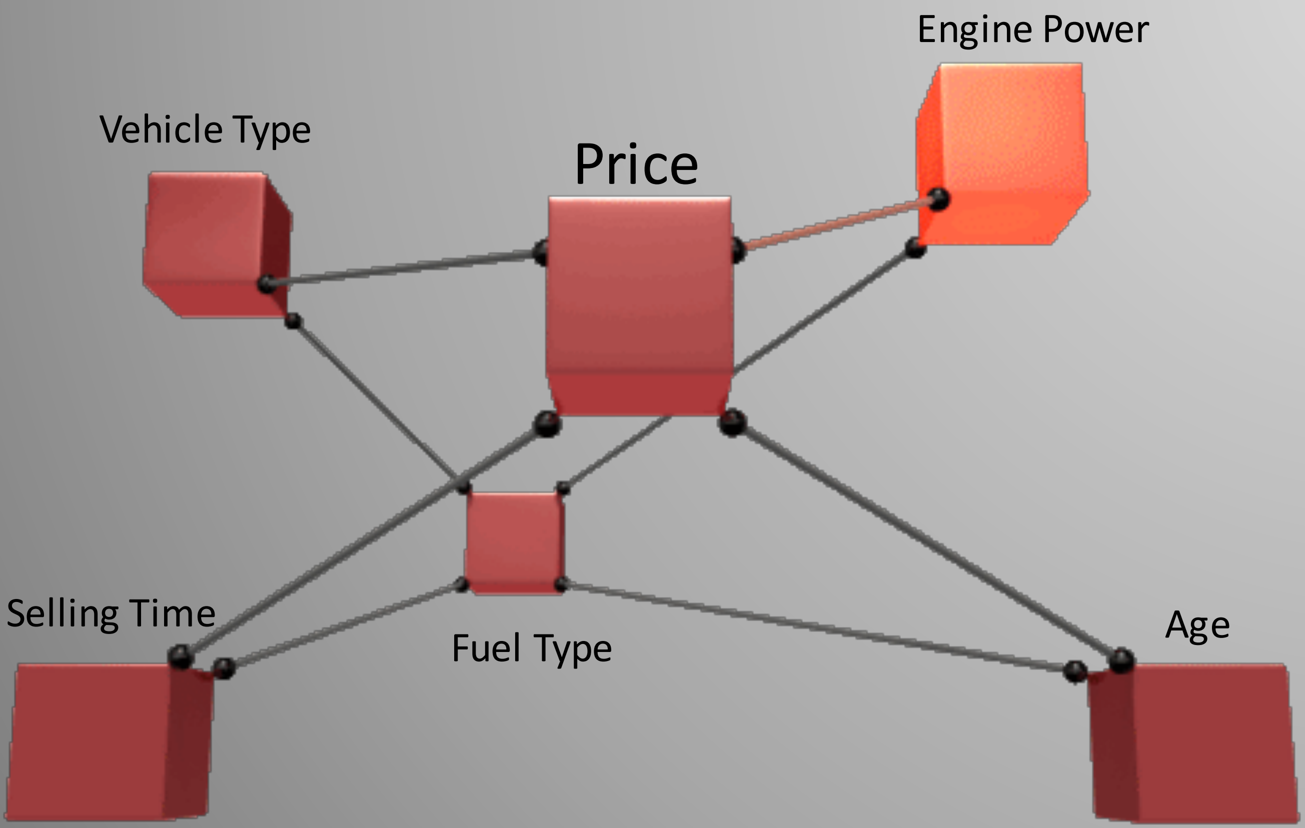
🟥 8





Content

- 🟥 1
- 🟦 2
- 🟧 3
- ⬜ 4
- 🟩 5
- 🟦 6 Determine relationships among the variables
- 🟩 7
- 🟥 8



Content

1

2

3

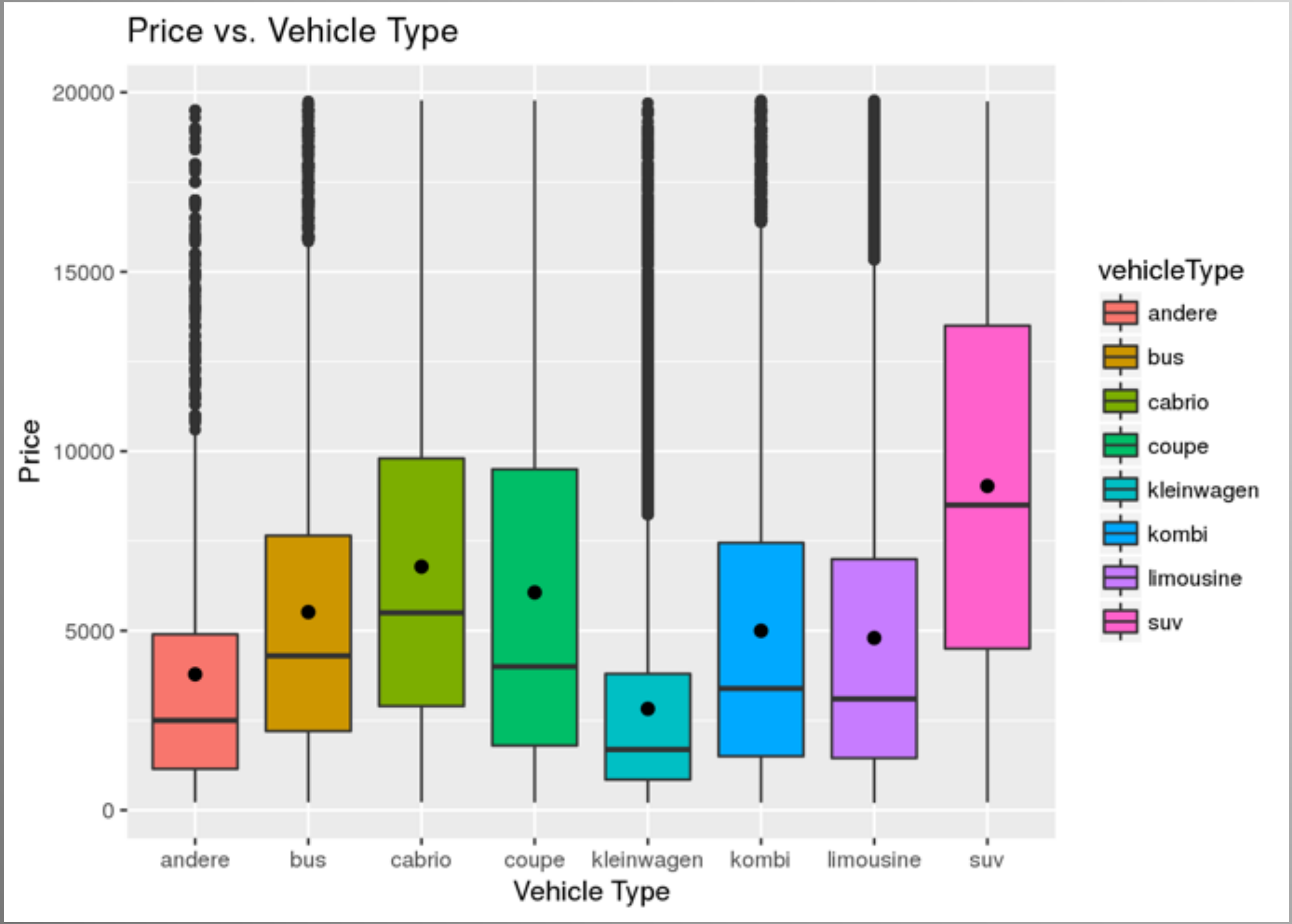
4

5

6 Determine relationships among the variables

7

8



Content

🟥 1

🟦 2

🟧 3

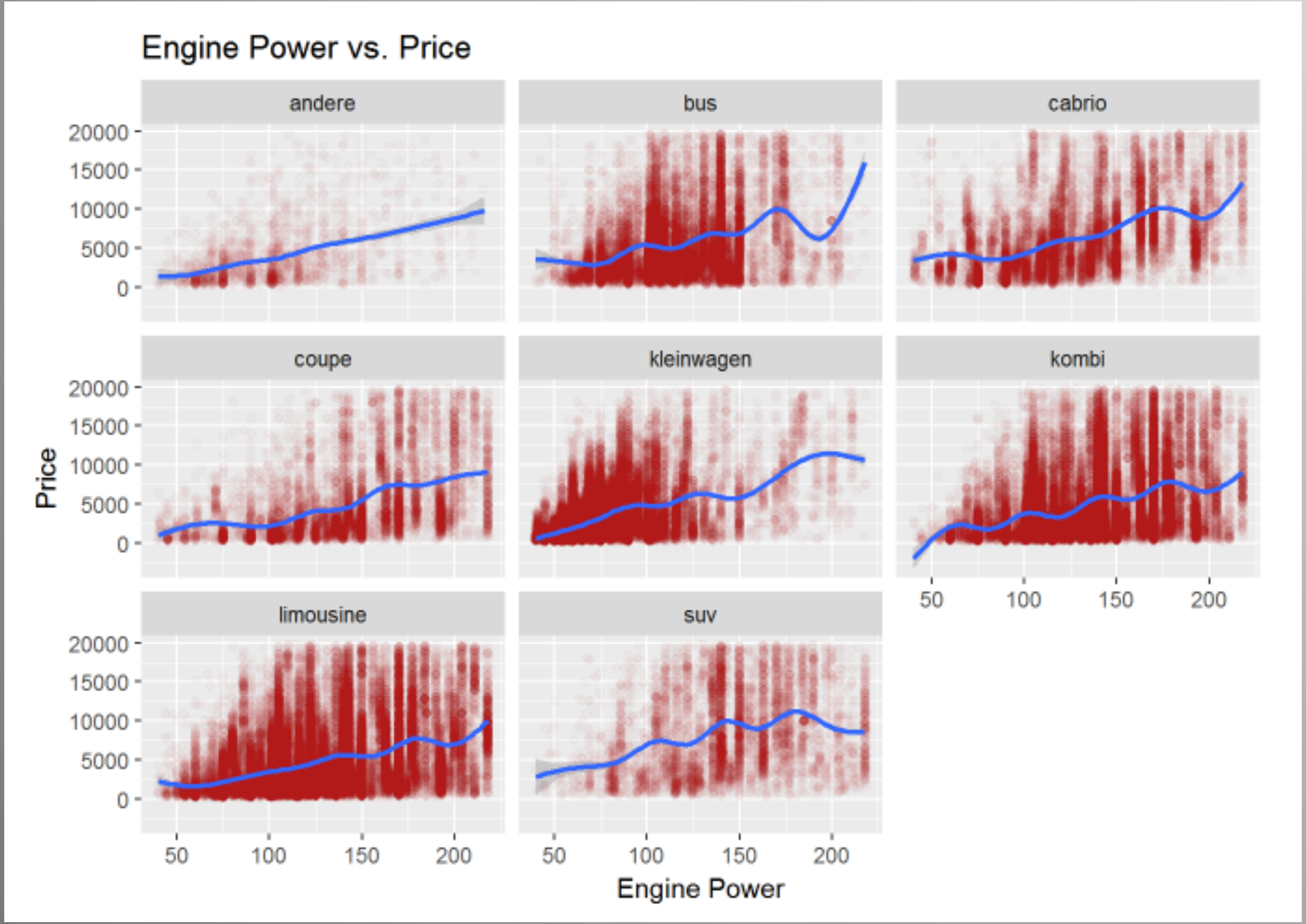
⬜ 4

🟩 5

🟦 6 Determine relationships among the variables

🟩 7

🟥 8



Content

1

2

3

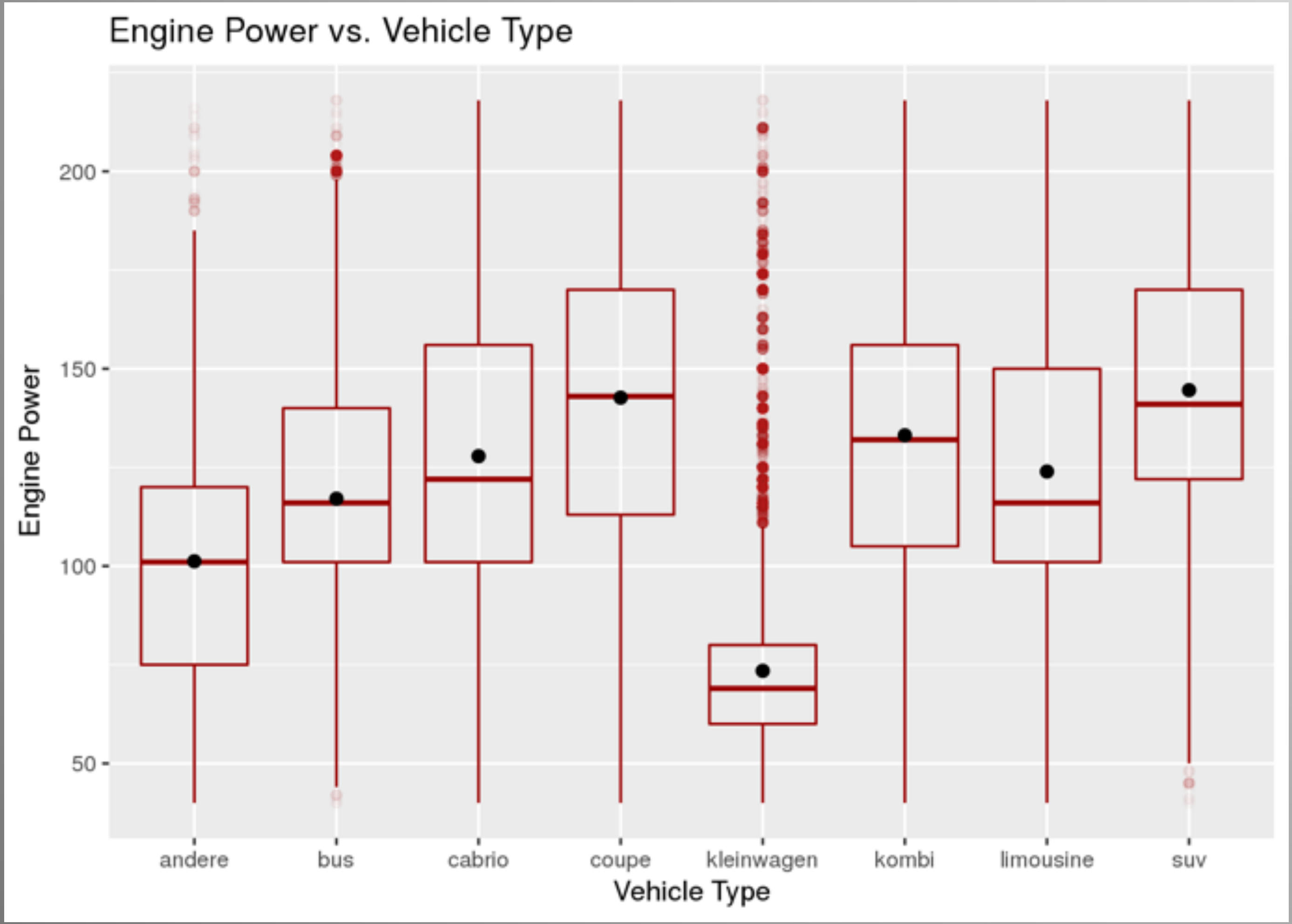
4

5

6 Determine relationships among the variables

7

8



Content

🟥 1

🟦 2

🟧 3

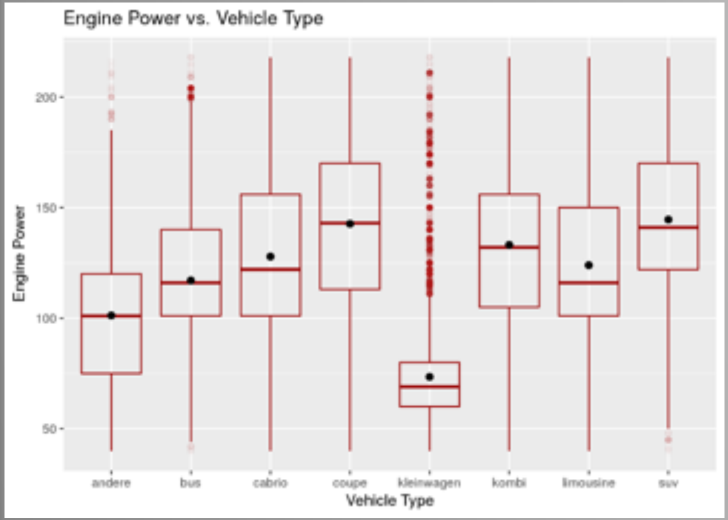
⬜ 4

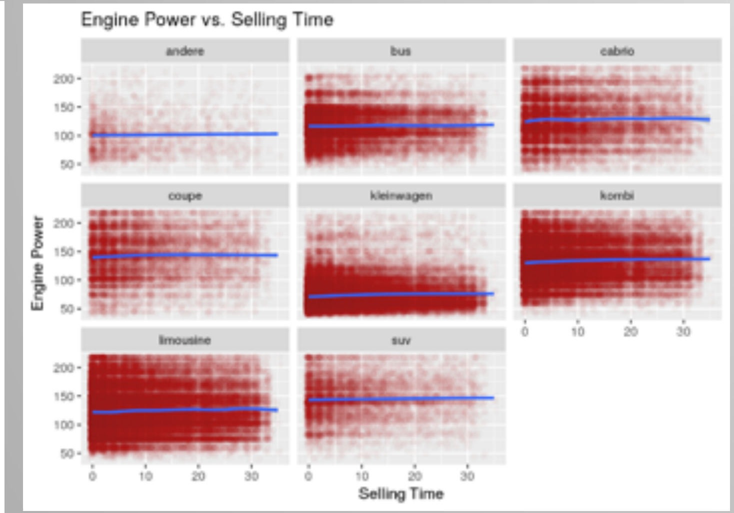
🟩 5

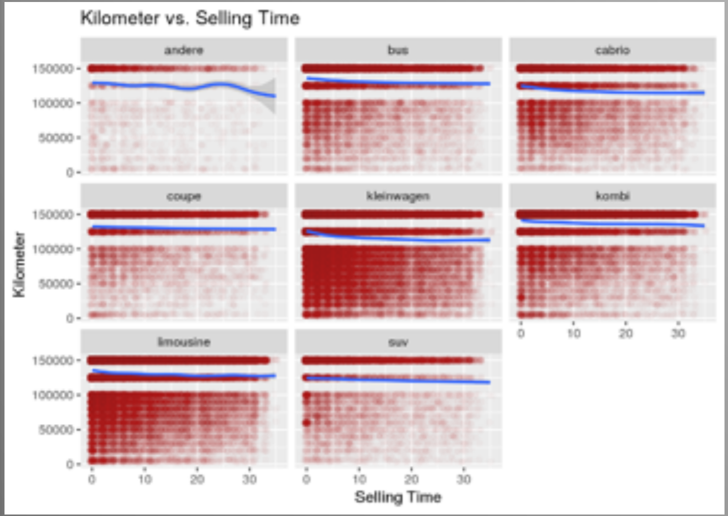
🟦 6 Determine relationships among the variables

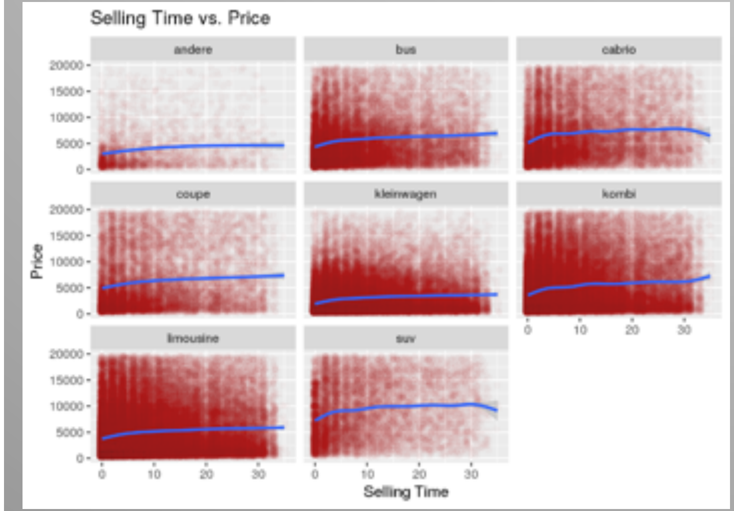
🟩 7

🟥 8









Content

1

2

3

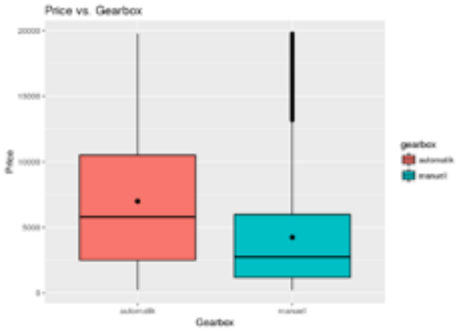
4

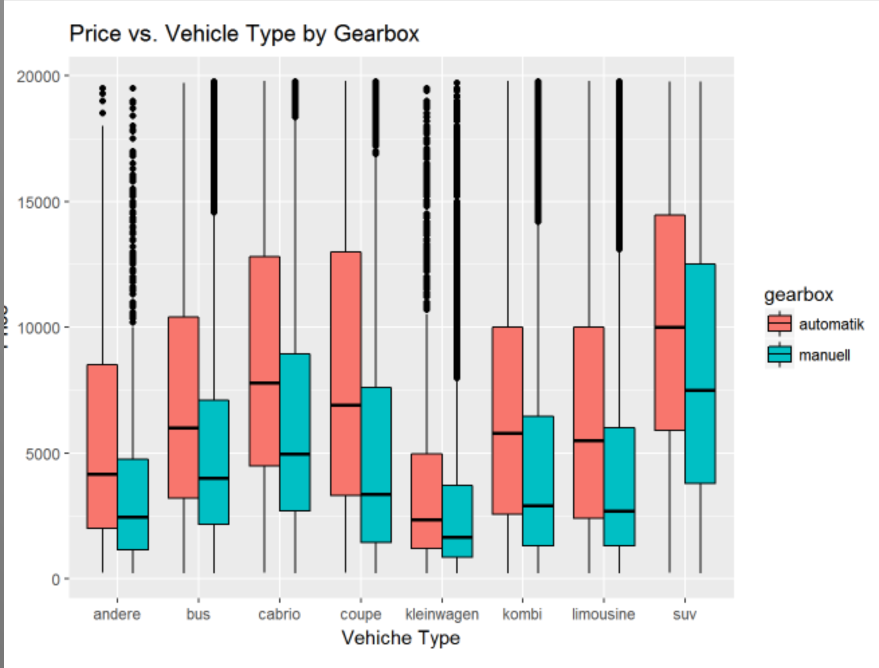
5

6 Determine relationships among the variables

7

8





Automatic gearbox cars are more expensive than the manual ones but the difference becomes less significant in cheap cars, i.e. Kleinwagen.

Content

1

2

3

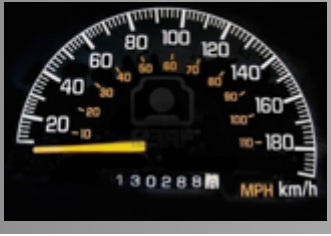
4

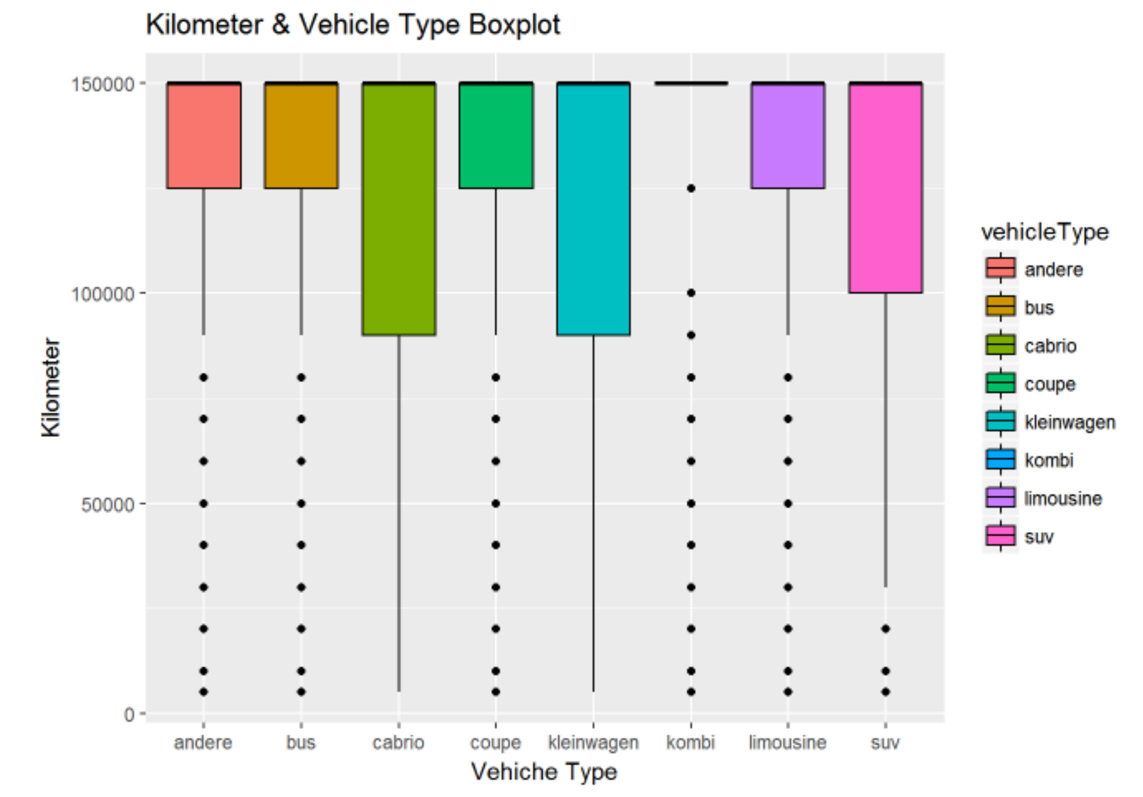
5

6 Determine relationships among the variables

7

8





Content

1

2

3

4

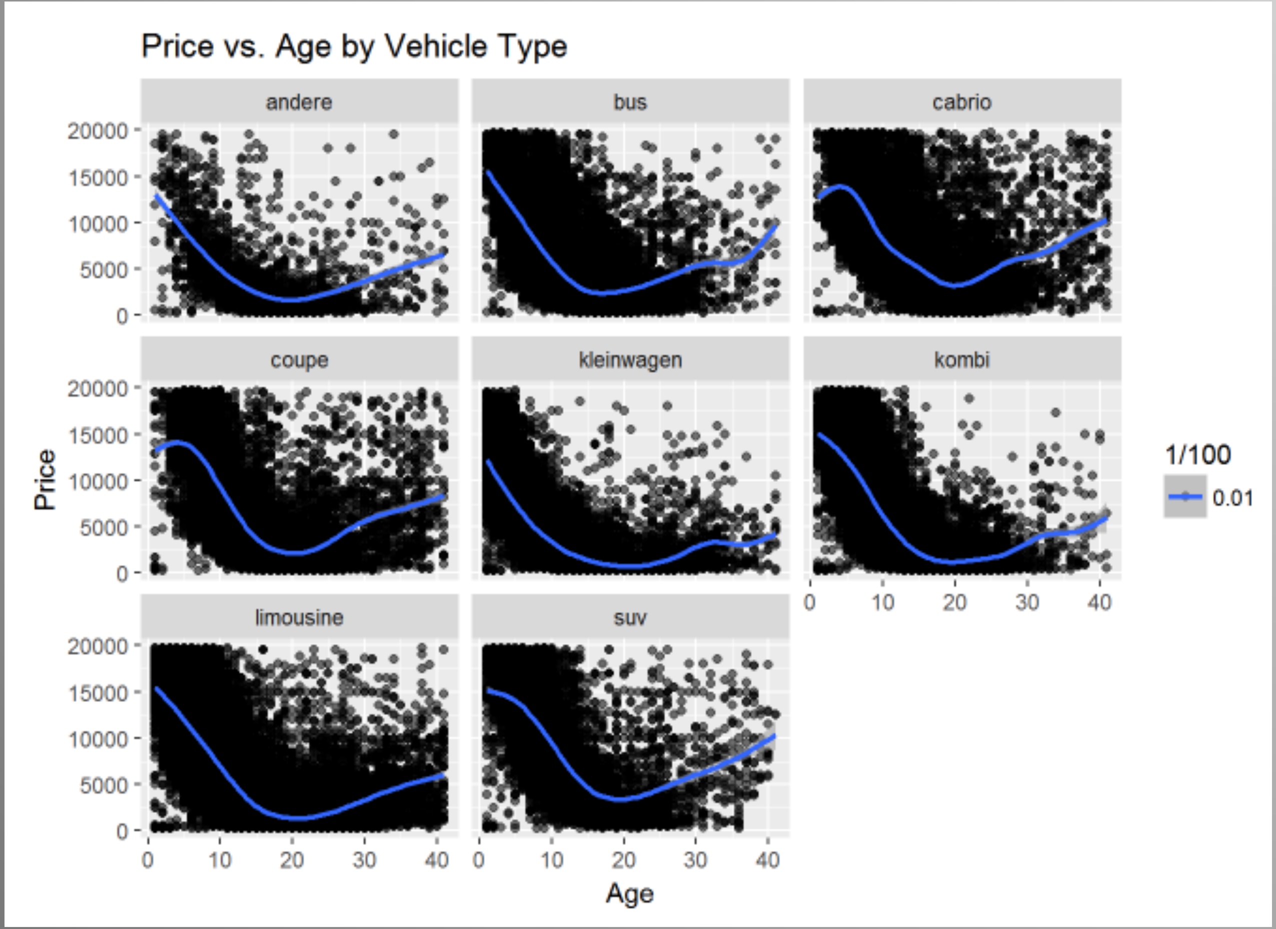
5

6 Determine relationships among the variables

7

8

In all vehicle types, the price continues to decrease between 0-20 years (20 years is the lowest point) but starts increase after between 20-30 years.



Content

1

2

3

4

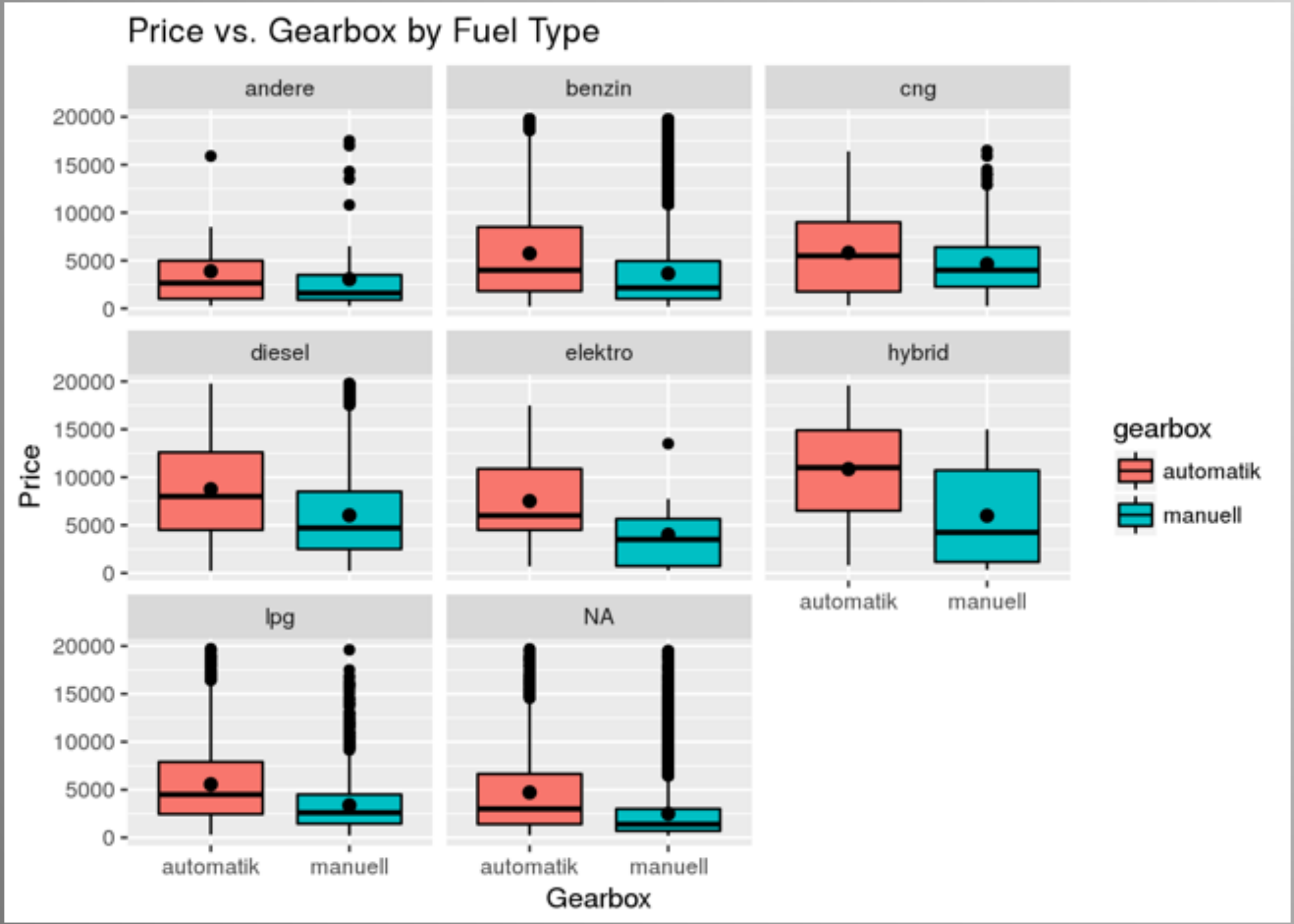
5

6 Determine relationships among the variables

7

8

In general, automatic gearbox cars are more expensive than the manual ones. This is especially significant in hybrid cars.



Content

🟥 1

🟦 2

🟧 3

⬜ 4

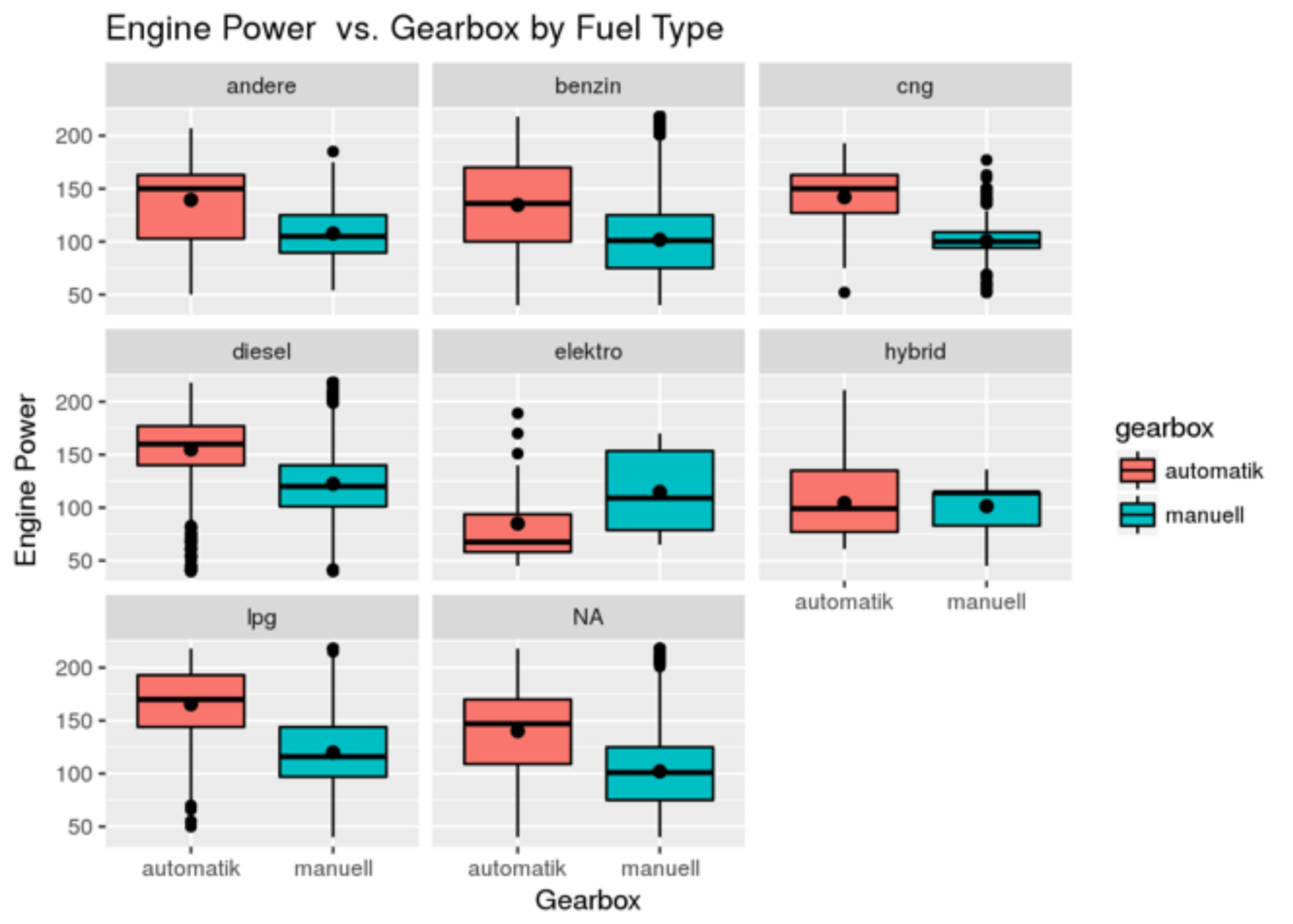
🟩 5

🟦 6 Determine relationships among the variables

🟩 7

🟥 8

Only electric cars with manual gearbox have superior engine power performance to the automatic ones. This divergence is one of the most interesting things we have



Content

1

2

3

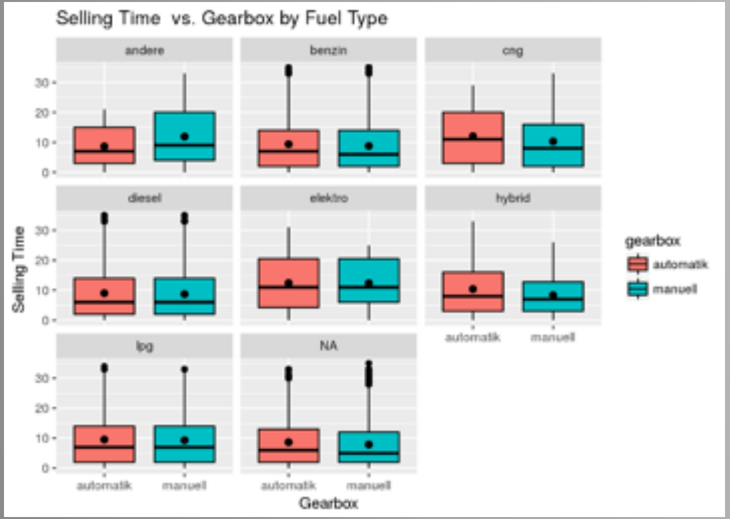
4

5

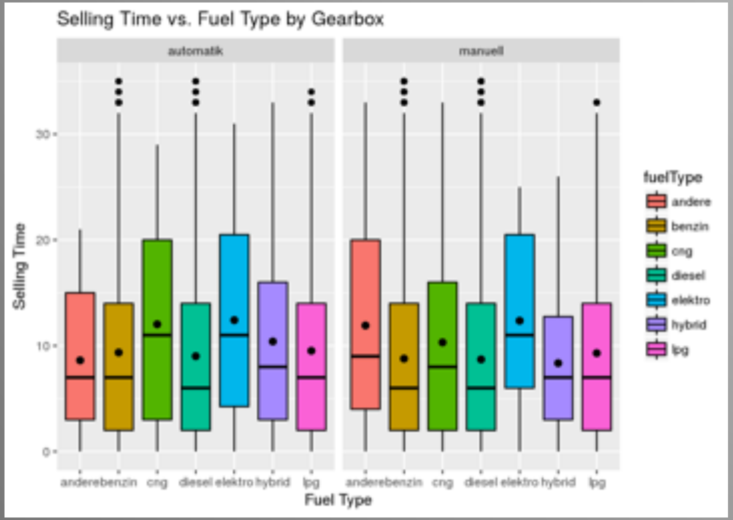
6 Determine relationships among the variables

7

8


Selling Time vs. Gearbox by Fuel Type

Selling time does not seem to be affected by the combinations of gearbox and fuel type as well, with the exception of CNG cars.


Selling Time vs. Fuel Type by Gearbox

The electric and CNG cars show longer selling time trend which may indicate that second-hand car market for hybrid cars have not matured yet.

Content

1

2

3

4

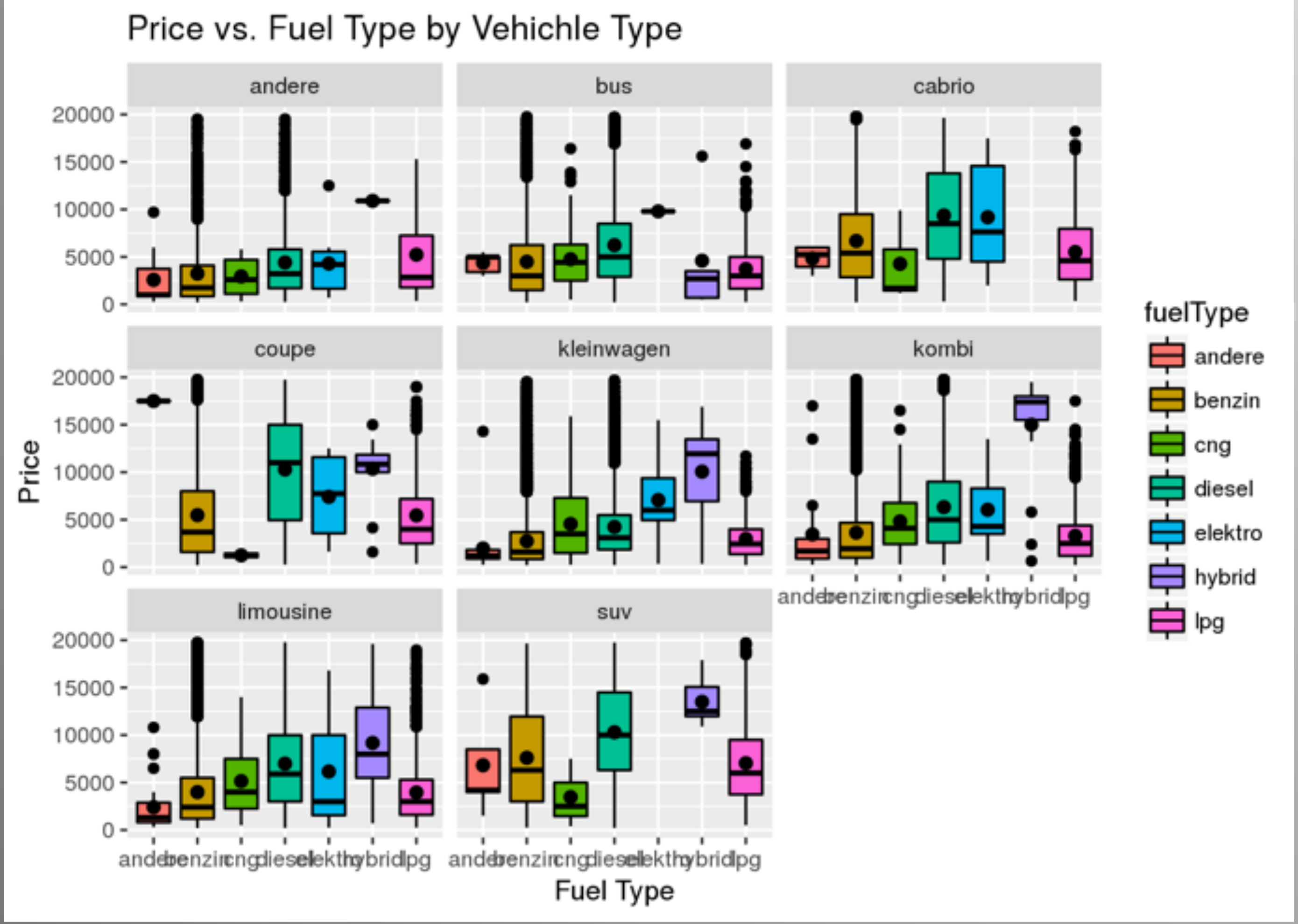
5

6 Determine relationships among the variables

7

8

In cheap cars (Kleinwagen), electric-hybrid fuel type makes a significant increase in price.



Content

1

2

3

4

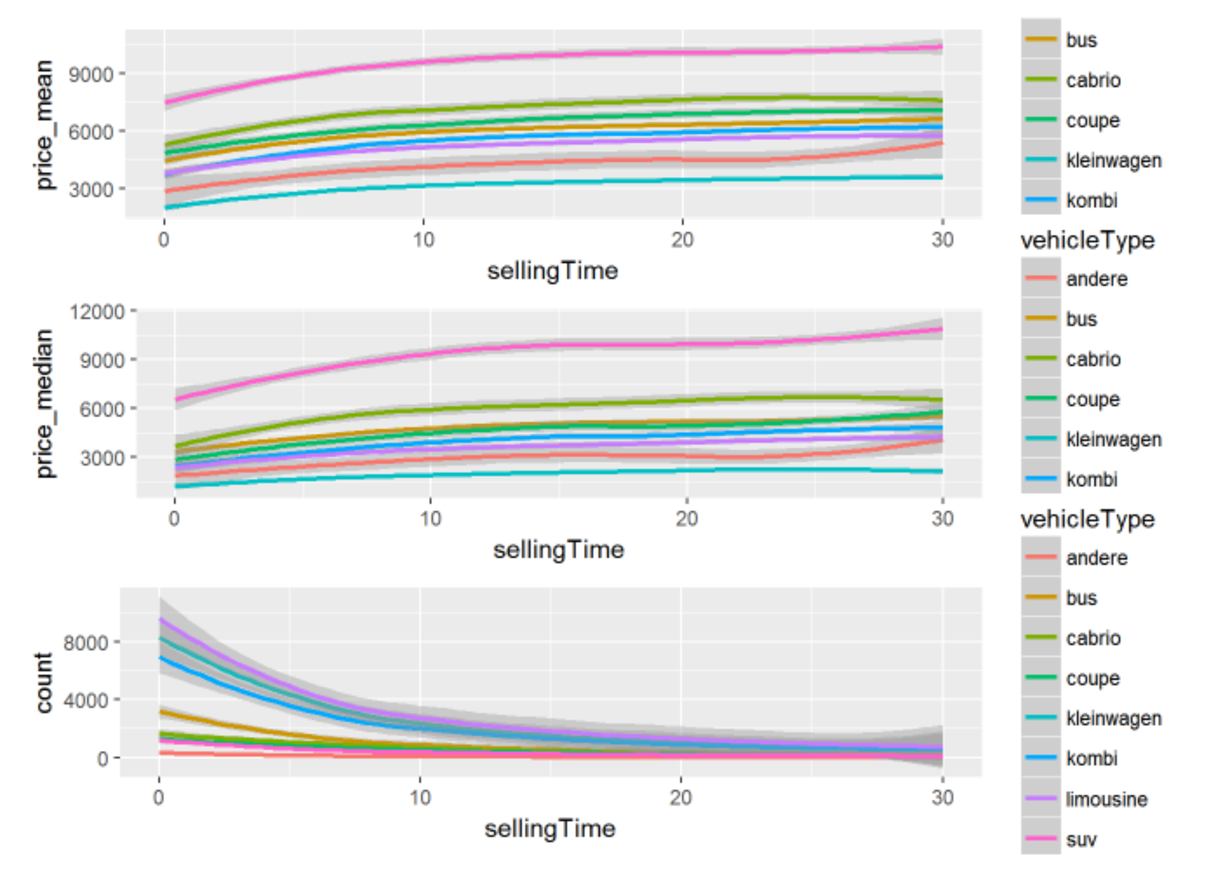
5

6 Determine relationships among the variables

7

8

Majority of the cars are sold only within 10 days in all vehicle types. For the first 20 days, whenever the price goes up, the selling time increases as well but this correlation stabilize in day 20. This trend is especially visible in SUV vehicles.



Content

1

2

3

4

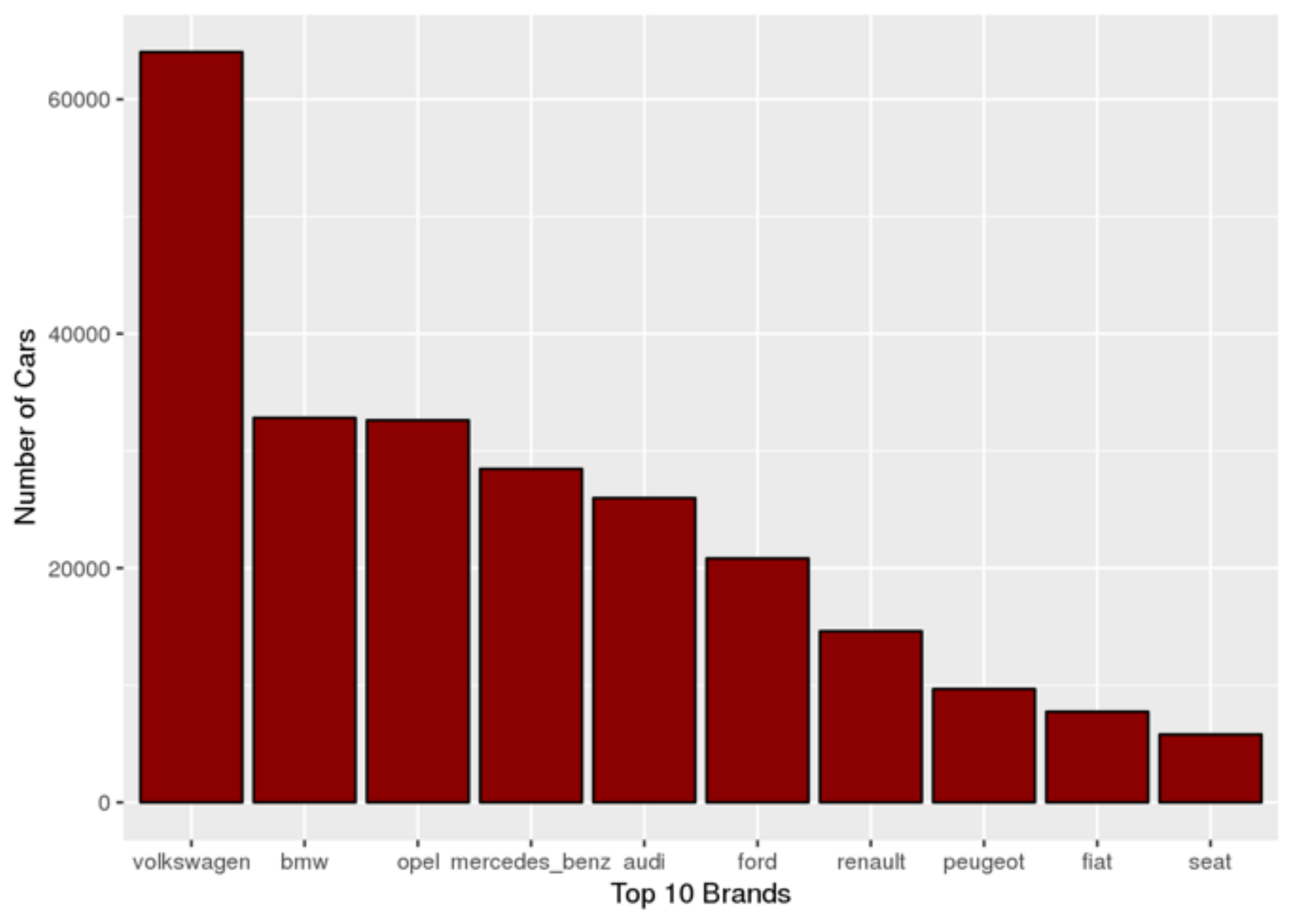
5

6 Determine relationships among the variables

7

8

Top 10 Brands (%80 second hand sales)– Most popular in the second-hand car market



Content

1

2

3

4

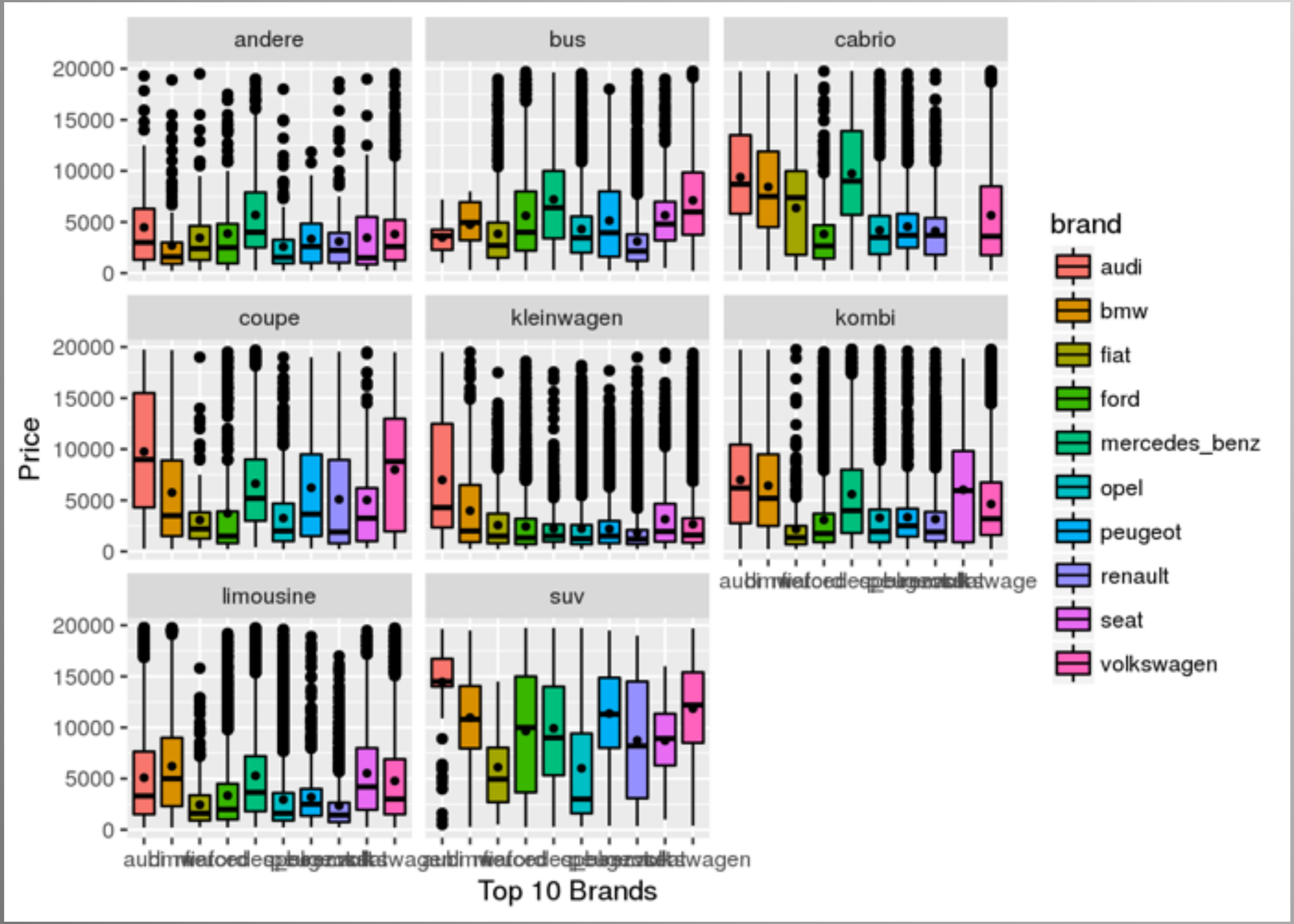
5

6 Determine relationships among the variables

7

8

Top 10 brands



Content

1

2

3

4

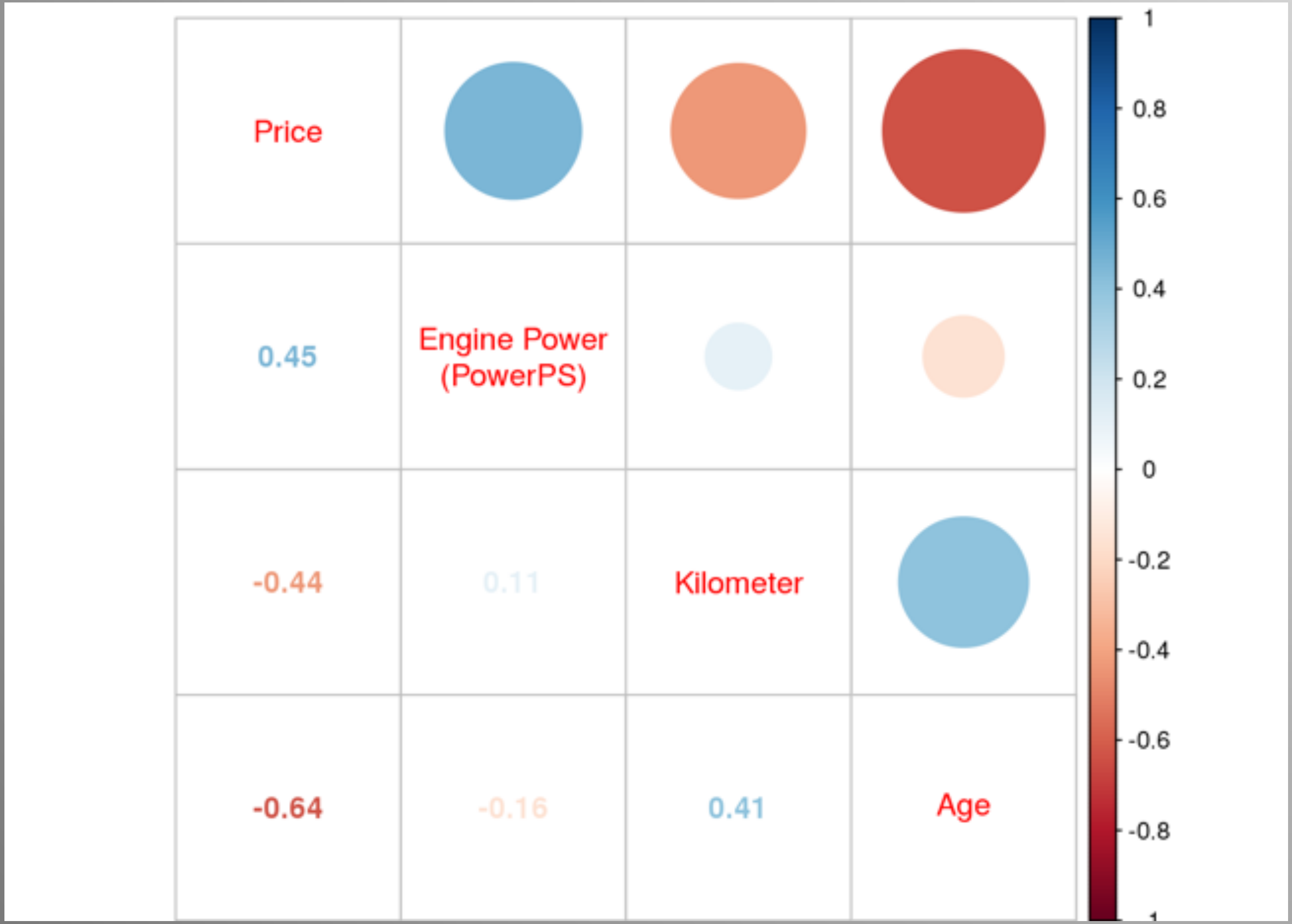
5

6 Determine relationships among the variables

7

8

Correlations



Content

1

2

3

4

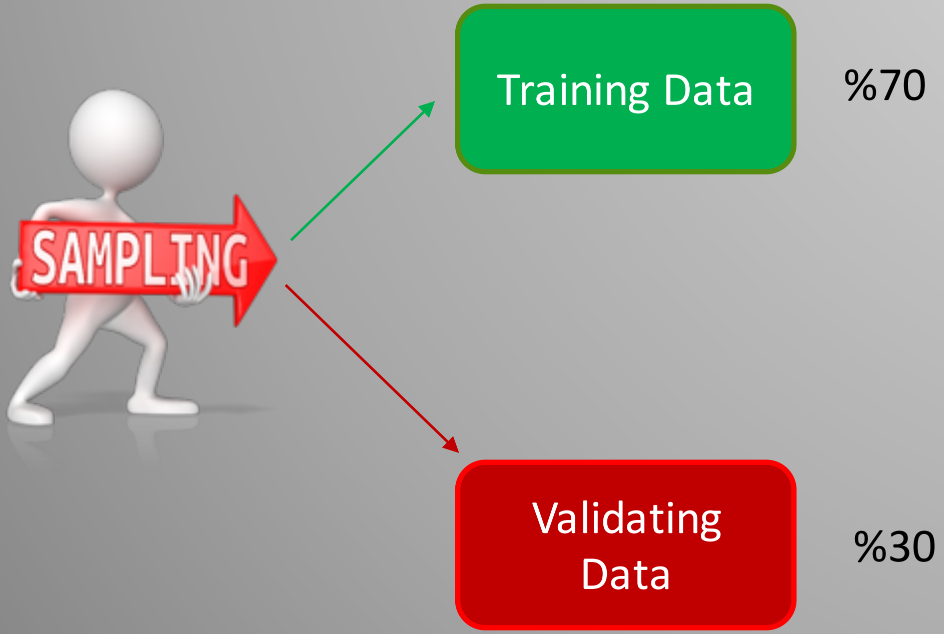
5

6

7 Regression Analysis

8

We split data into 2 parts before regression analysis



Content

🟥 1

🟦 2

🟧 3

⬜ 4

🟩 5

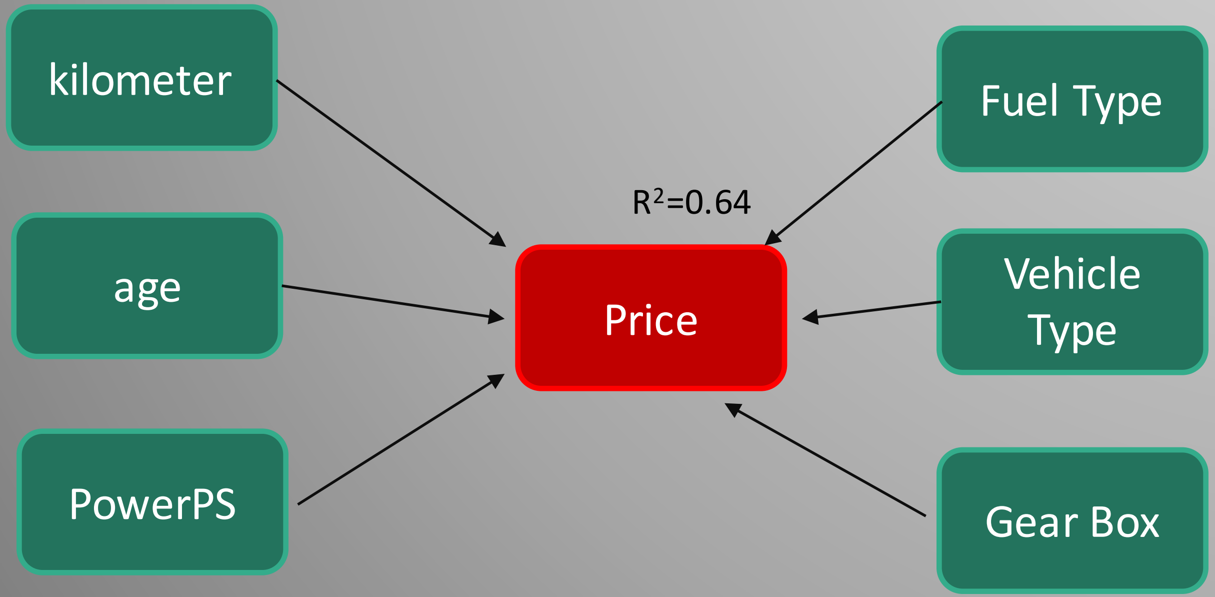
🟦 6

🟩 7 Regression Analysis

🟥 8

Regression Analysis Results for Training Data

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + K \ + \beta_p x_{pi} + \varepsilon_i$$



kilometer

age

PowerPS

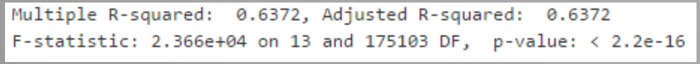
R^2=0.64

Price

Fuel Type

Vehicle Type

Gear Box



```
Multiple R-squared:  0.6372, Adjusted R-squared:  0.6372
F-statistic: 2.366e+04 on 13 and 175103 DF,  p-value: < 2.2e-16
```

Content

🟥 1

🟦 2

🟧 3

⬜ 4

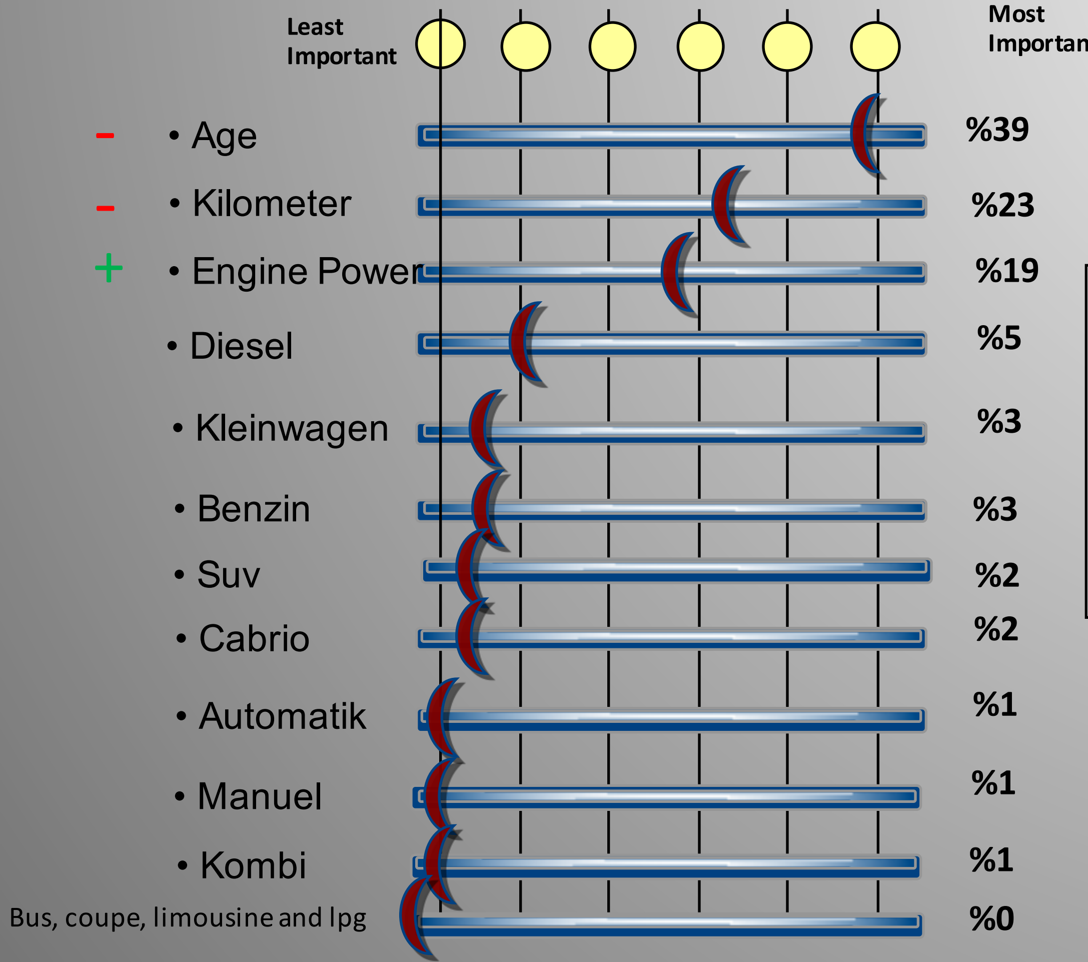
🟩 5

🟦 6

🟩 7 Regression Analysis

🟥 8

Q: What factors are most important for Price?



Age, Kilometer and Engine Power are the most important factors that explains second hand car price

Content

1

2

3

4

5

6

7 Regression Analysis

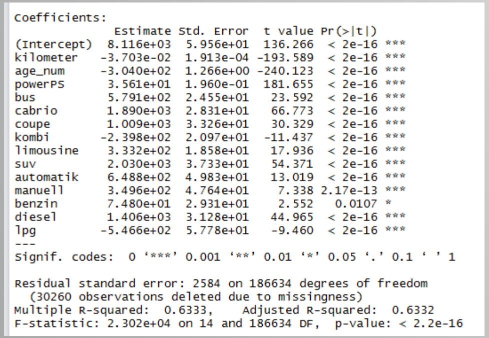
8

Model Diagnostics

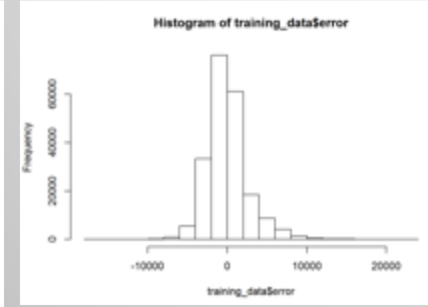
R^2

Training Data

%63

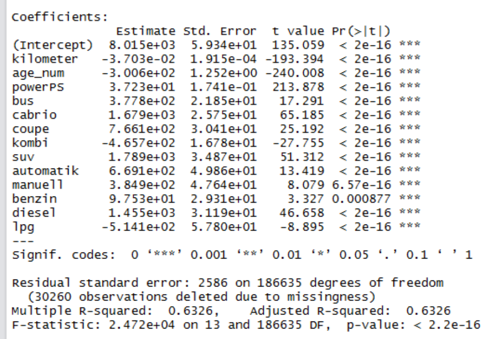


```
Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)  8.116e+03  5.956e+01  136.266  < 2e-16 ***
kilometer   -3.703e-02  1.913e-04 -193.589  < 2e-16 ***
age_num     -3.040e+02  1.266e+00 -240.123  < 2e-16 ***
powerPS      3.561e+01  1.960e-01  181.655  < 2e-16 ***
bus          5.791e+02  2.455e+01   23.592  < 2e-16 ***
cabrio       1.890e+03  2.831e+01   66.773  < 2e-16 ***
coupe        1.009e+03  3.326e+01   30.329  < 2e-16 ***
kombi       -2.398e+02  2.097e+01  -11.437  < 2e-16 ***
limousine    3.332e+02  1.858e+01   17.936  < 2e-16 ***
suv          2.030e+03  3.733e+01   54.371  < 2e-16 ***
automatik    6.488e+02  4.983e+01   13.019  < 2e-16 ***
manuell      3.496e+02  4.764e+01    7.338 2.17e-13 ***
benzin       7.480e+01  2.931e+01    2.552   0.0107 *
diesel       1.406e+03  3.128e+01   44.965  < 2e-16 ***
lpg         -5.466e+02  5.778e+01   -9.460  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2584 on 186634 degrees of freedom
  (30260 observations deleted due to missingness)
Multiple R-squared:  0.6333,    Adjusted R-squared:  0.6332
F-statistic: 2.302e+04 on 14 and 186634 DF,  p-value: < 2.2e-16
```

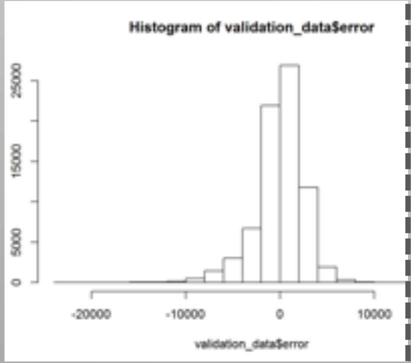


Validating Data

%63



```
Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)  8.015e+03  5.934e+01  135.059  < 2e-16 ***
kilometer   -3.703e-02  1.915e-04 -193.394  < 2e-16 ***
age_num     -3.006e+02  1.252e+00 -240.008  < 2e-16 ***
powerPS      3.723e+01  1.741e-01  213.878  < 2e-16 ***
bus          3.778e+02  2.185e+01   17.291  < 2e-16 ***
cabrio       1.679e+03  2.575e+01   65.185  < 2e-16 ***
coupe        7.661e+02  3.041e+01   25.192  < 2e-16 ***
kombi       -4.657e+02  1.678e+01  -27.755  < 2e-16 ***
suv          1.789e+03  3.487e+01   51.312  < 2e-16 ***
automatik    6.691e+02  4.986e+01   13.419  < 2e-16 ***
manuell      3.849e+02  4.764e+01    8.079 6.57e-16 ***
benzin       9.753e+01  2.931e+01    3.327 0.000877 ***
diesel       1.455e+03  3.119e+01   46.658  < 2e-16 ***
lpg         -5.141e+02  5.780e+01   -8.895  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2586 on 186635 degrees of freedom
  (30260 observations deleted due to missingness)
Multiple R-squared:  0.6326,    Adjusted R-squared:  0.6326
F-statistic: 2.472e+04 on 13 and 186635 DF,  p-value: < 2.2e-16
```



Content

🟥 1

🟦 2

🟧 3

⬜ 4

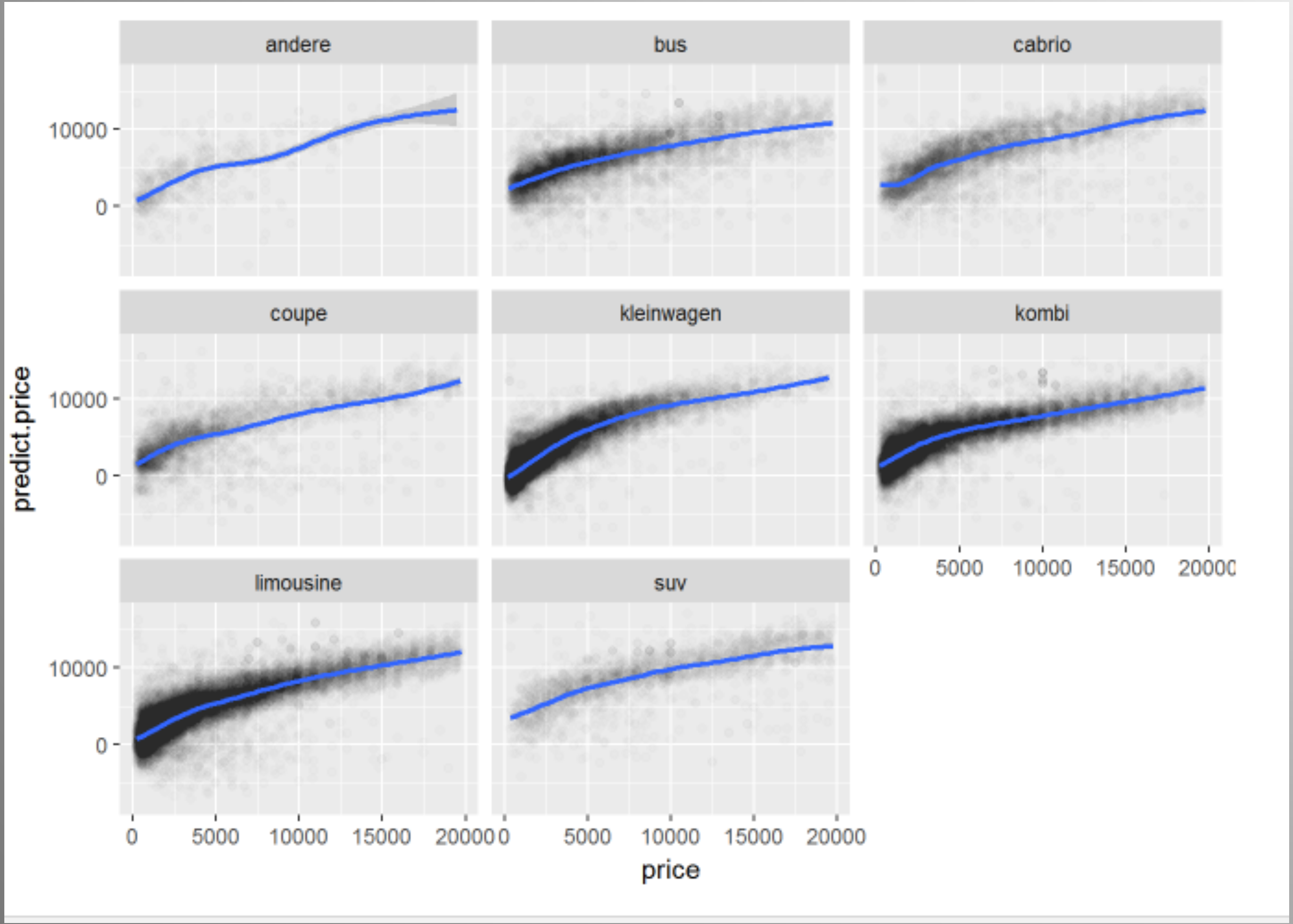
🟩 5

🟦 6

🟩 7 Regression Analysis

🟥 8

Actual vs Predicted Price



Content

🟥 1

🟦 2

🟧 3

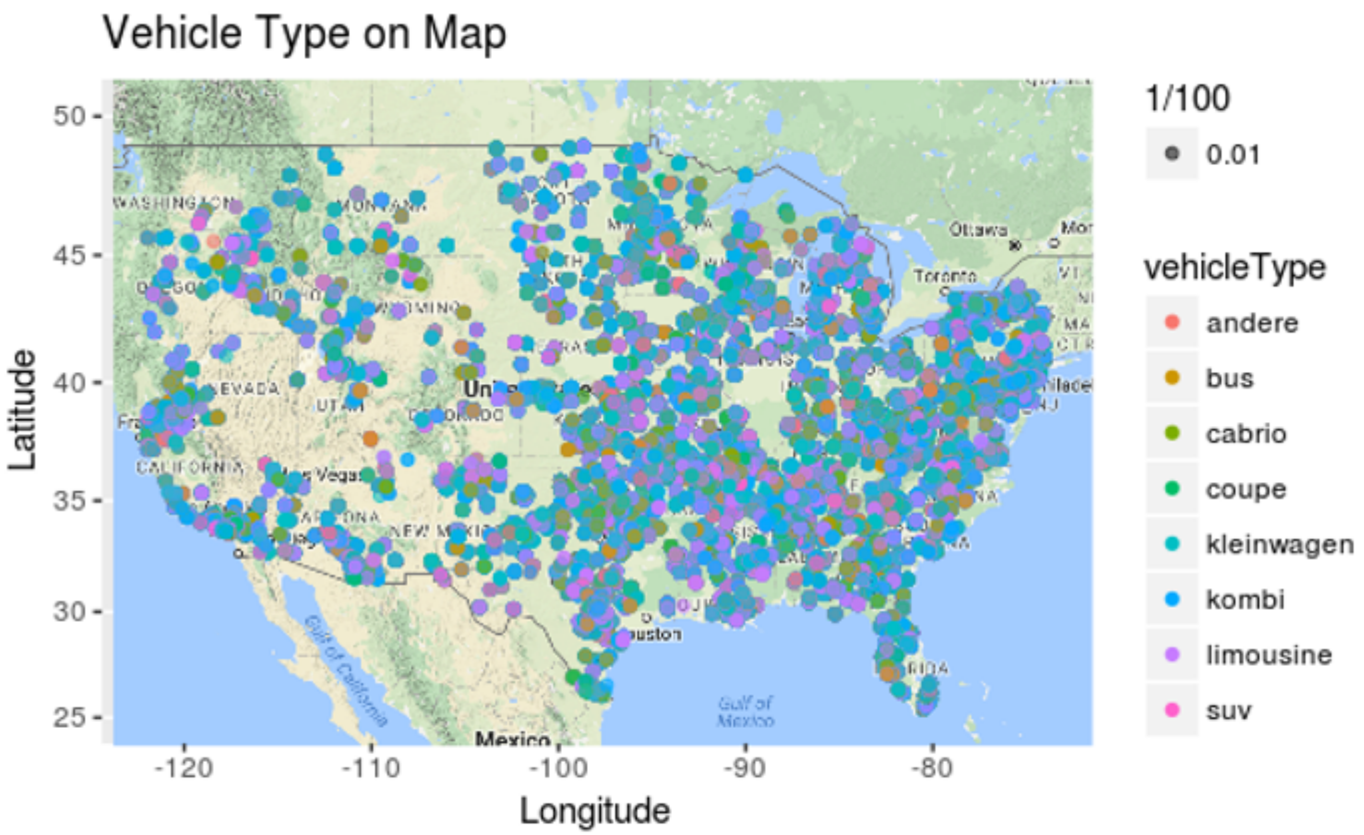
⬜ 4

🟩 5

🟦 6

🟩 7

🟥 8 Maps



Content

1

2

3

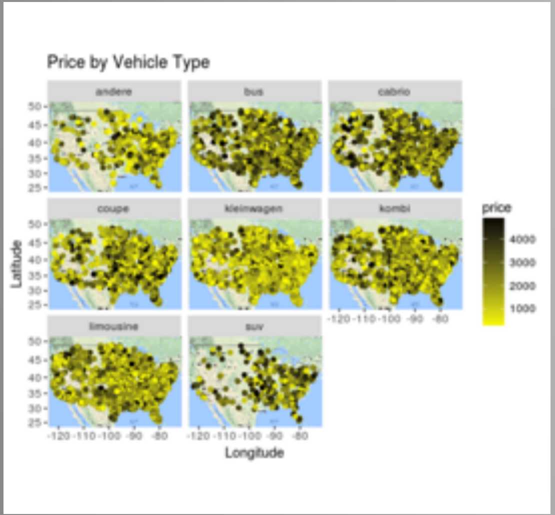
4

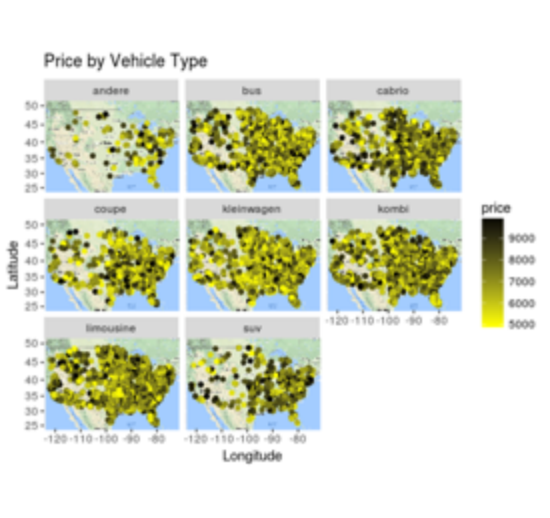
5

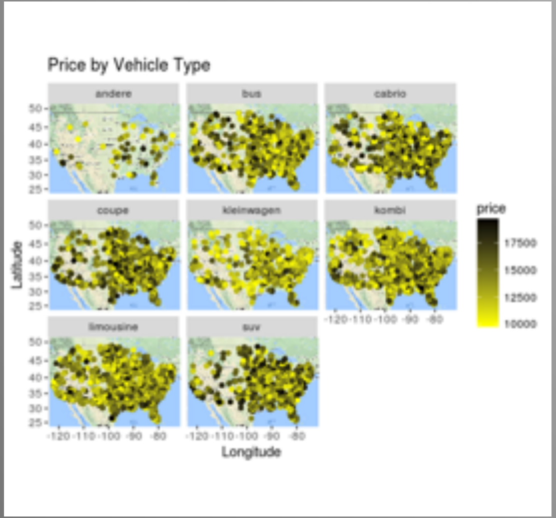
6

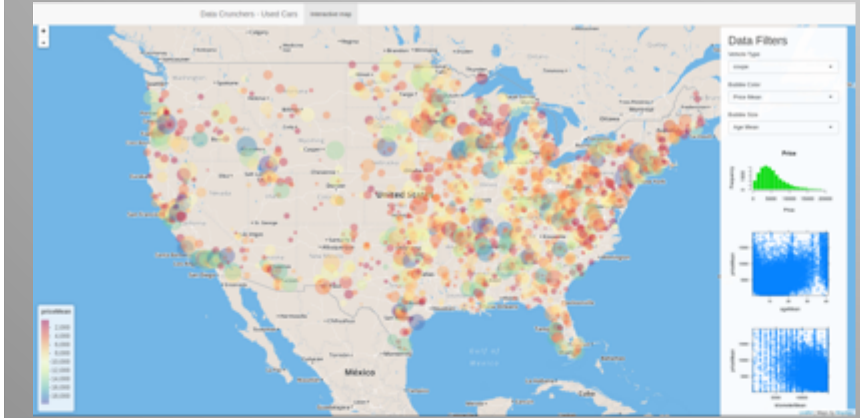
7

8 Maps









Content

🟥 1
🟦 2
🟧 3
⬜ 4
🟩 5
🟦 6
🟩 7
🟥 8
🟦 9 Conclusion

- Limousine, kombi and kleinwagen are the most popular vehicle types in the second-hand market. Most expensive cars are SUV’s while the cheapest ones are kleinwagens.
- On average Kleinwagen vehicle type is the cheapest and has the lowest engine power. But it also shows the most outliers – might be as a result of brand-model diversity.
- The most popular brands are Volkswagen, BMW, Opel, Mercedes, Audi, Ford, Renault, Peugeot, Fiat and Seat. These 10 brand correspond to almost 80% of the cars. (Originally our dataset contains around 40 brands)
- Most of the cars in the second-hand market are above 100.000 km, even 150.000 km. People does not frequently change cars according to our data set.
- Majority of the second-hand cars are sold only within 35 days. The ratio of the first 10 days (day 0 stands for same day sale) is quite high. This shows us that either Ebay-Kleinanzeigen is very successful at targeting customers or the second-hand market is more fluid that we actually thought.

Content

1

2

3

4

5

6

7

8

9 Conclusion

- To our surprise, there is no strong/significant correlation between selling time and vehicle type, kilometer and price. We saw that whenever price goes up the change to be sold in 10-20 days increases especially in SUV vehicles (rather than 0-10 days) but this is not a general trend.
- Hybrid (electro engine, CNG) second-hand car market is emerging but shows longer selling time trend.
- In all vehicle types, the price continues to decrease between 0-20 years (20 years is the lowest point) but starts increase after between 20-30 years. Maybe, 20+ year old second-hand cars can be considered as 'antique' and users' emotional attachment may cause abnormalities.
- Hybrid cars with manual gearbox have superior engine power performance. This is a divergence from all correlations and one of the most interesting things we have found.
- According to our regression analysis, age (39%), kilometer(%23) and engine power(%19) are the most important factors explaining second hand price.
- Zip code analysis shows that kombi is the most popular second hand car when the zip code is provided.
- East Coast second hand car market is bigger.