Stats & Facts

Group Project

Supervised Learning

Car Rentals Analysis

Alfredo Funicello | Amr Rashad | Shihab Hamati

Introduction

Behavior of daily rental rate

Effect of different features

Full Linear Regression

Other Explorations

Dataset & Exploration Path

- Car Rentals data collected from different websites for major US cities, in July 2020
- There are 5581 observations
- Curious about **how the different features affect car rental prices** (e.g., age of the car, fuel type, ratings, etc.)
- Variables were explored **individually** and **together**
- **Linear regression** was used to generate model

Introduction

Behavior of daily rental rate

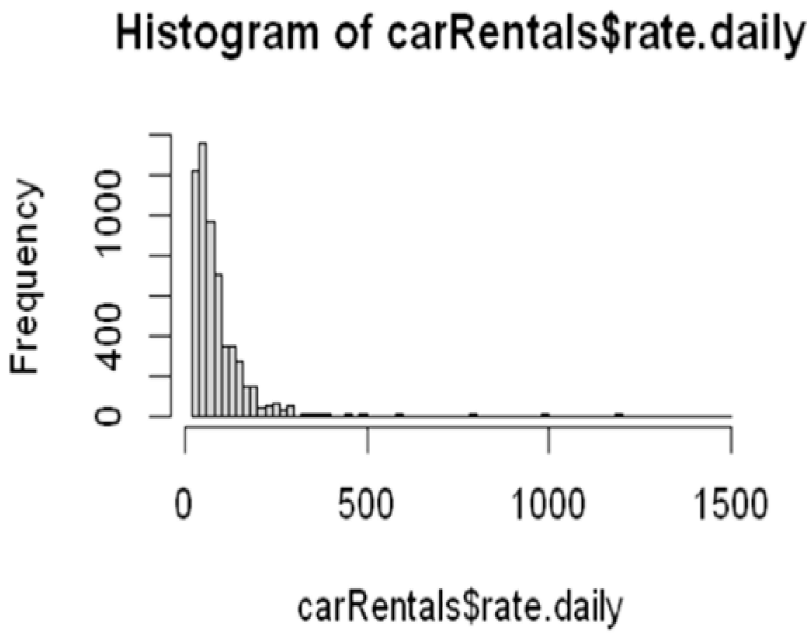
Effect of different features

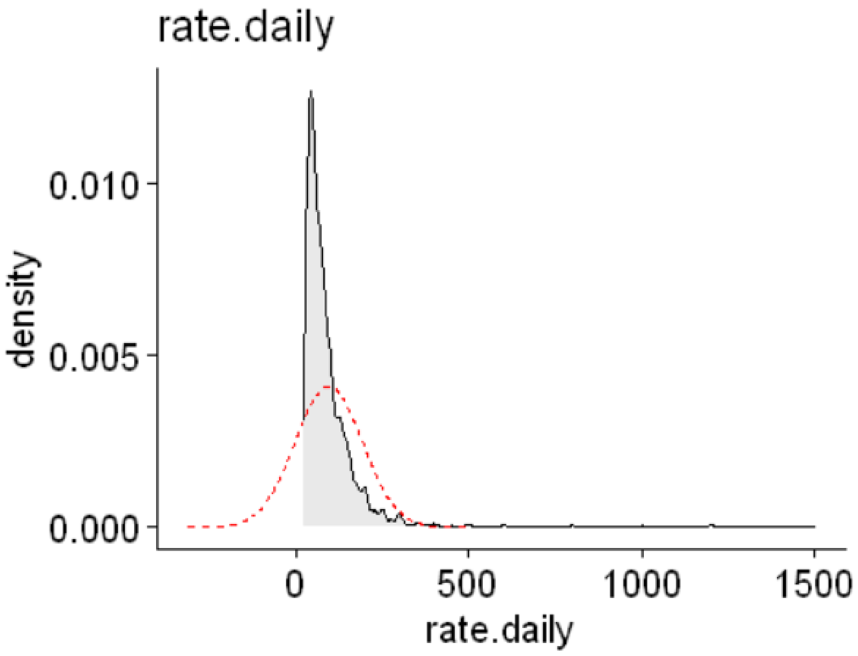
Full Linear Regression

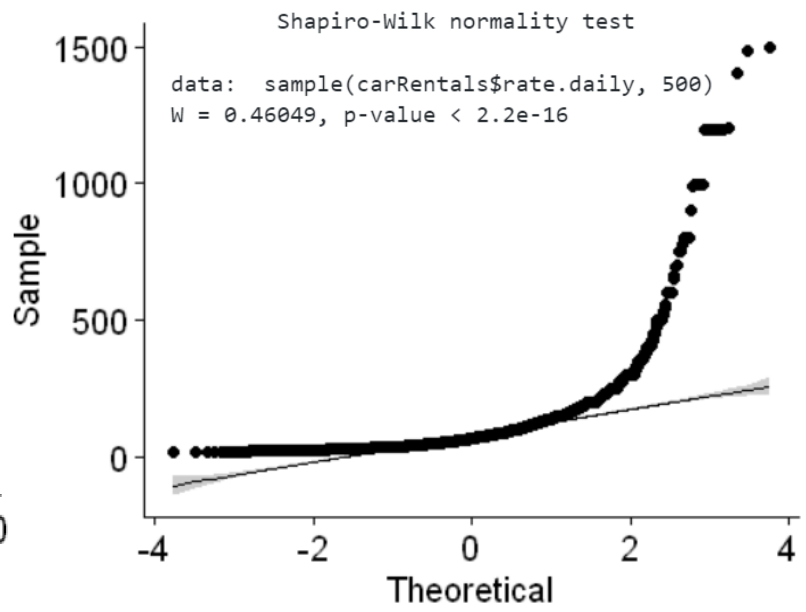
Other Explorations

Understanding the Dependent Variable Distribution

- Variable “rate.daily” suffers from **non-linearity**
- This can be observed visually from the **histogram**, confirmed visually by the **QQ plot**, and tested numerically by the **Shapiro test**

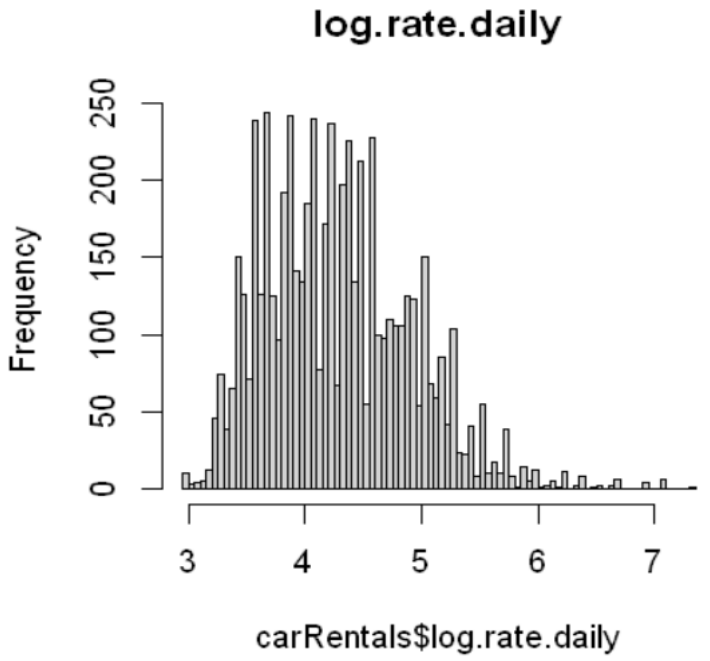


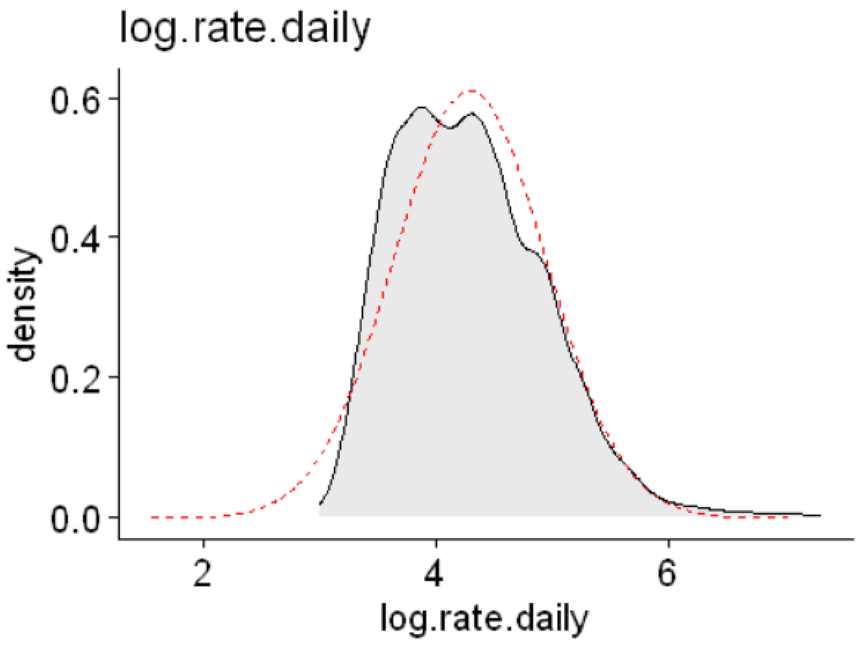


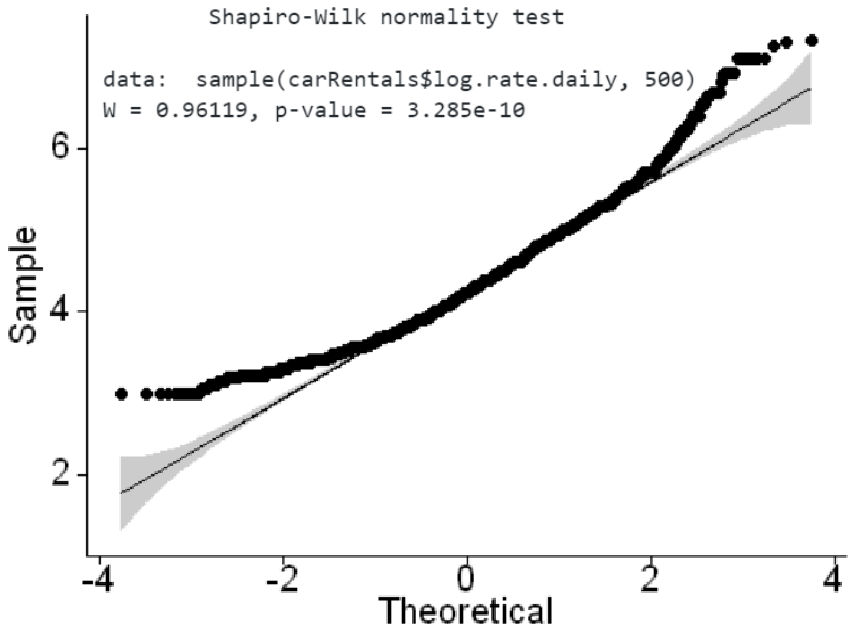


Transformation of the Dependent Variable

- Daily rental rate is non-zero and positive, so a simple log transformation is applied to achieve a **closer-to-normal** distribution







Handling outliers

- There are 32 observations that lie beyond 2*IQR away from the Upper Quartile of the log-transformed daily rate
- Upon further exploration, we identify 2 main groups of outliers:
    - classical cars, even if log(rate) is not outlier (16 observations)

1968 · 1965 · 1976 · 1979 · 1980 · 1961 · 1983 · 1995 · 1966 · 1957 · 1966 · 1986 · 1955 · 1965 · 1972 · 1969

- mostly prestigious brands with high rates (29 observations)

Tesla · Lamborghini · BMW · Porsche · Ford · Mercedes-Benz · McLaren · Audi · Ferrari · Rolls Royce · Aston Martin · Chevrolet

- Classical cars were **dropped** (as they have a different market model and insufficient sample size to explore it)
- Outliers of current models were **retained**

Introduction

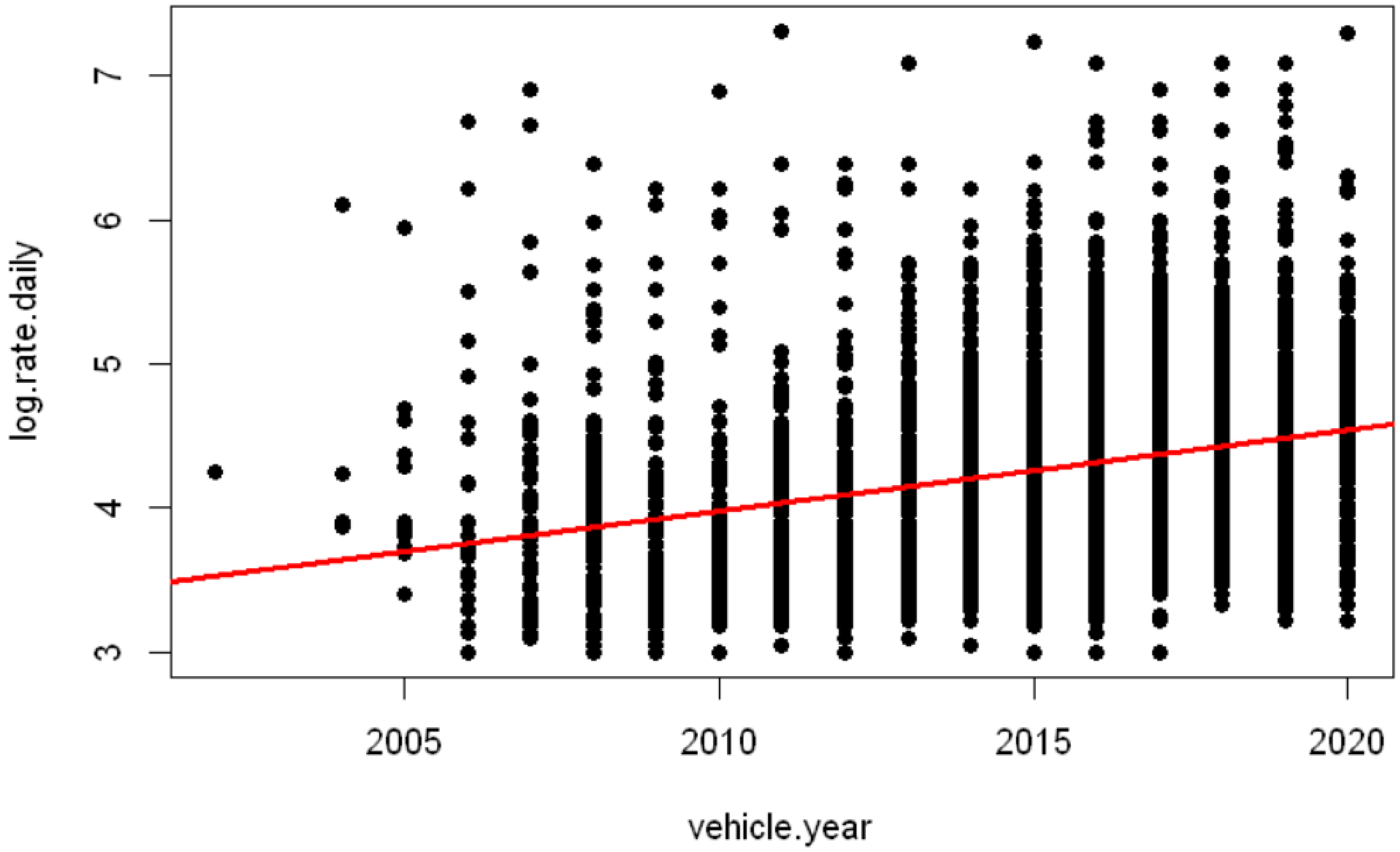
Behavior of daily rental rate

Effect of different features

Full Linear Regression

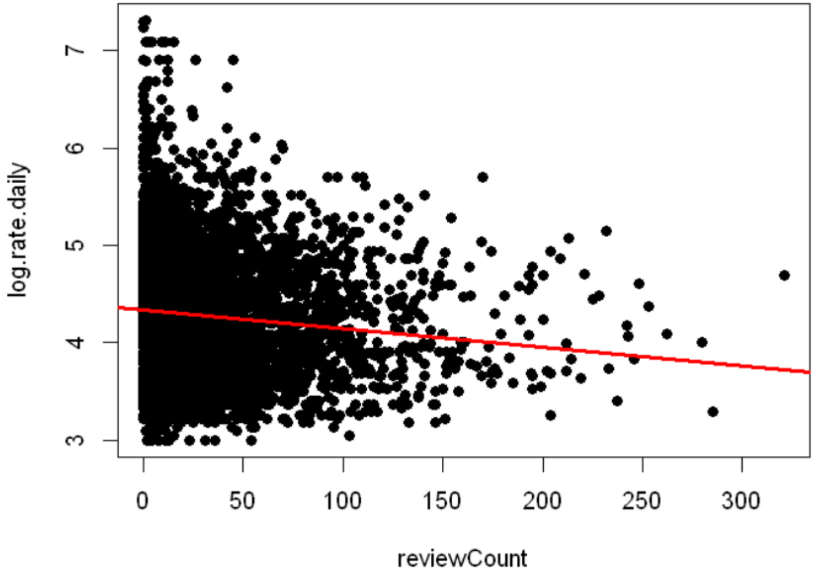
Other Explorations

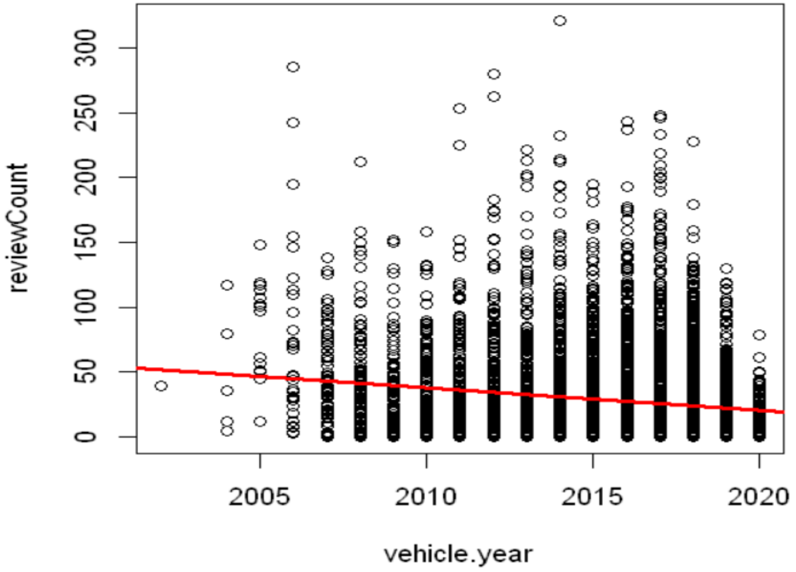
Year: Newer cars are more expensive, as expected



Number of reviews: negative correlation (counterintuitive)

- Counterintuitively, the more reviews the lower the rate
- This is because more reviews are more **likely to be older cars**
- A **quadratic** model could better explain the effect of more reviews vs age
- Higher order polynomials are significant up to 6^{th} order, but loose **interpretability**






```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             4.287548   0.008364 512.633  < 2e-16 ***
poly(reviewCount, 2)1 -5.142679   0.638776  -8.051 9.89e-16 ***
poly(reviewCount, 2)2  3.048197   0.638776   4.772 1.87e-06 ***
```

Car brand and model required data cleaning

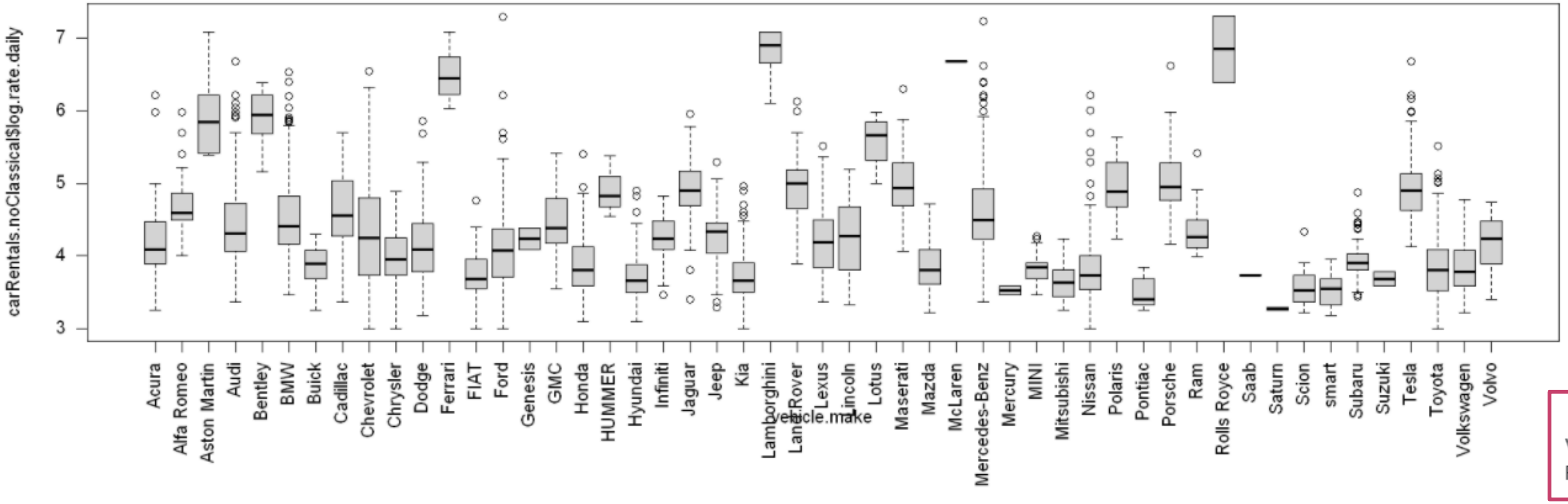
- Since data sources are different, slight variations in spellings of car brands or types existed, and they were reconciled

A data.frame: 54 × 2

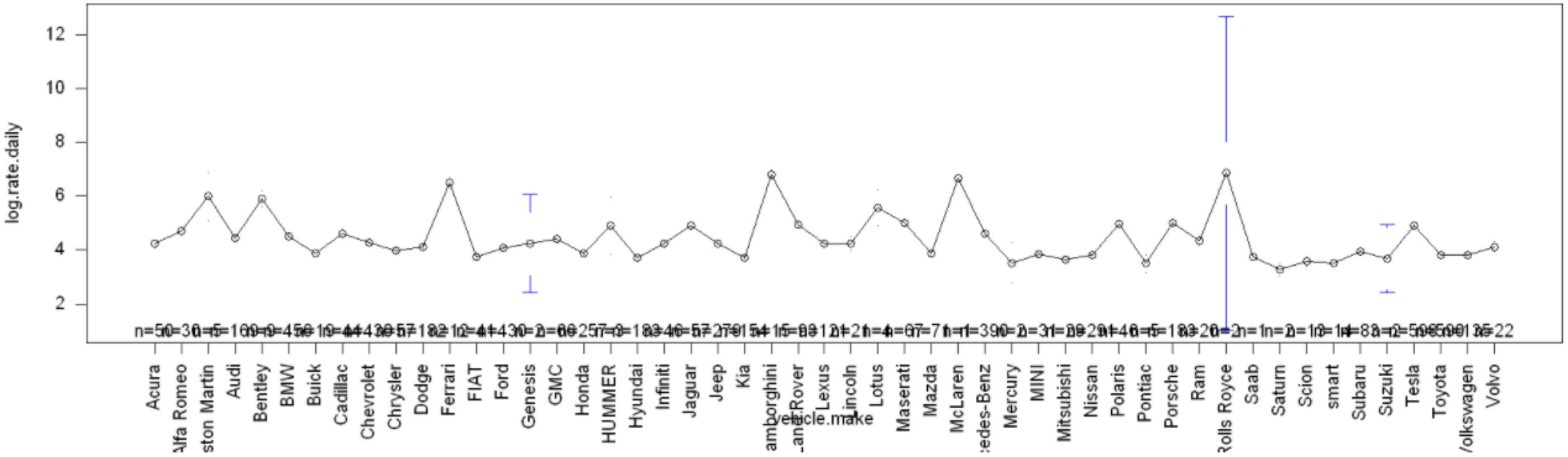
vehicle.make	n
<fct>	<int>
Acura	50
Alfa-romeo	9
Alfa Romeo	21
Aston Martin	5
Audi	169
Bentley	9
BMW	456

	vehicle.model	n
1	1 Series	6
2	124 Spider	8
3	1500	11
4	2	3
5	2-Series	1
6	2 Series	17
7	200	12
8	2500	2
9	3	17
10	3-Series	6
11	3 Series	94
12	3 Series Gran Turismo	2
13	300	11
14	3500	4
15	370Z	3
16	4-Series	6
17	4 Series	43

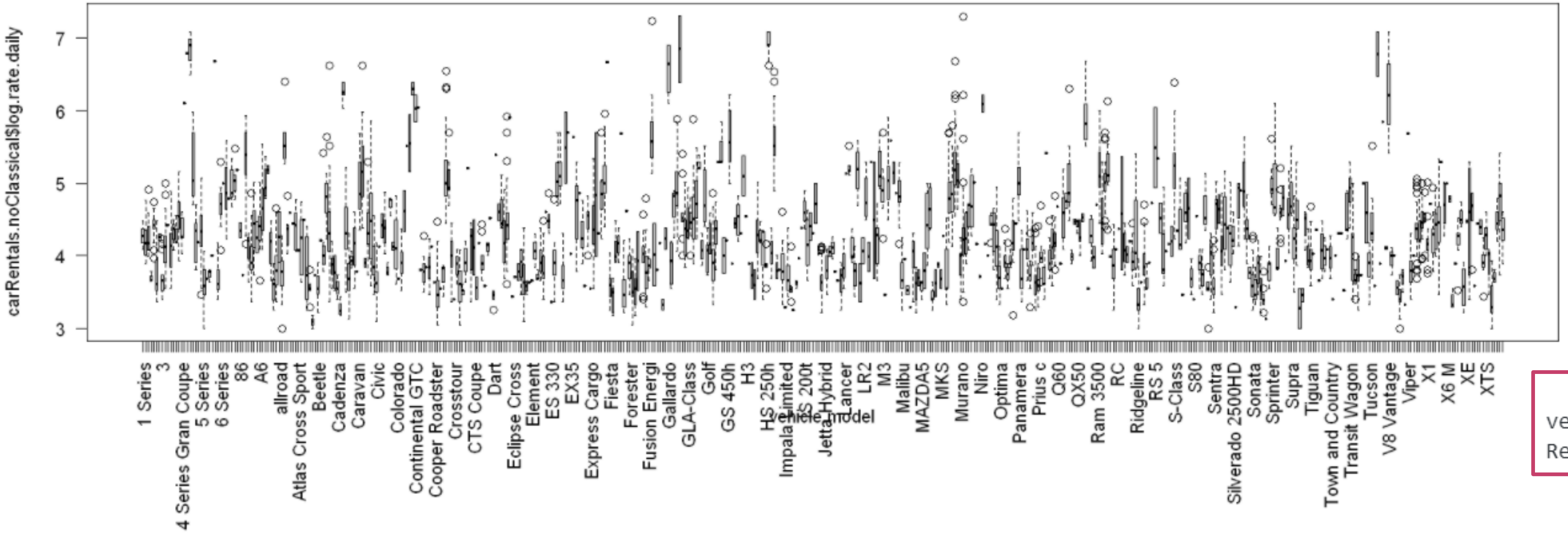
Car brand is a significantly explanative feature for the rental rate



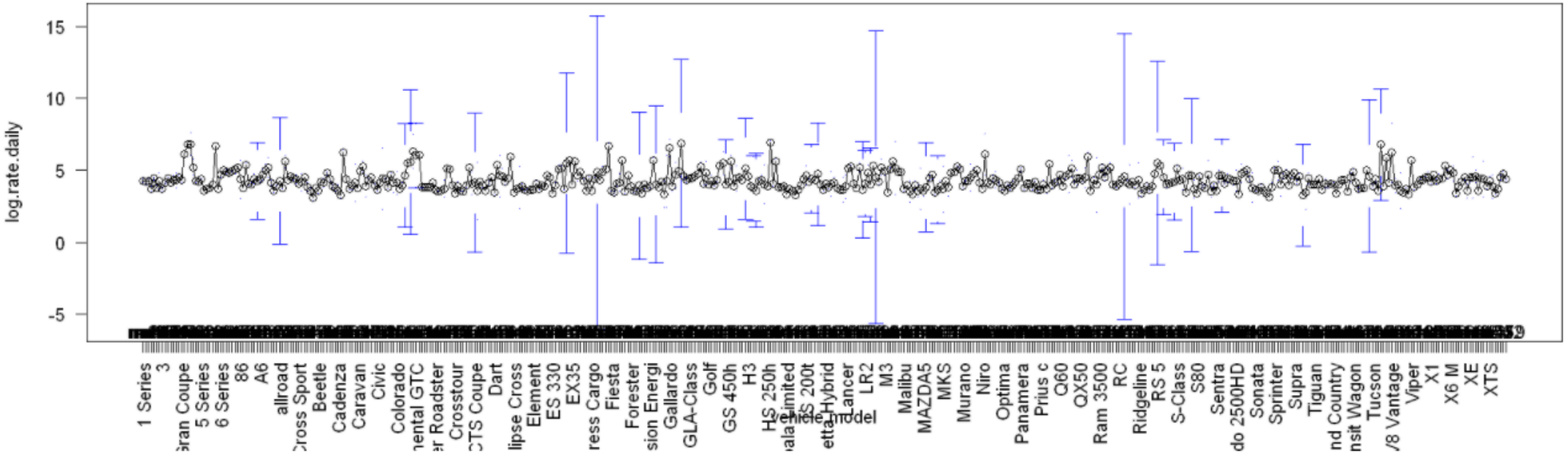
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
vehicle.make	50	1232	24.639	120.5	<2e-16	***
Residuals	5782	1183	0.205			



Car model is also a significantly explanative feature for the rental rate

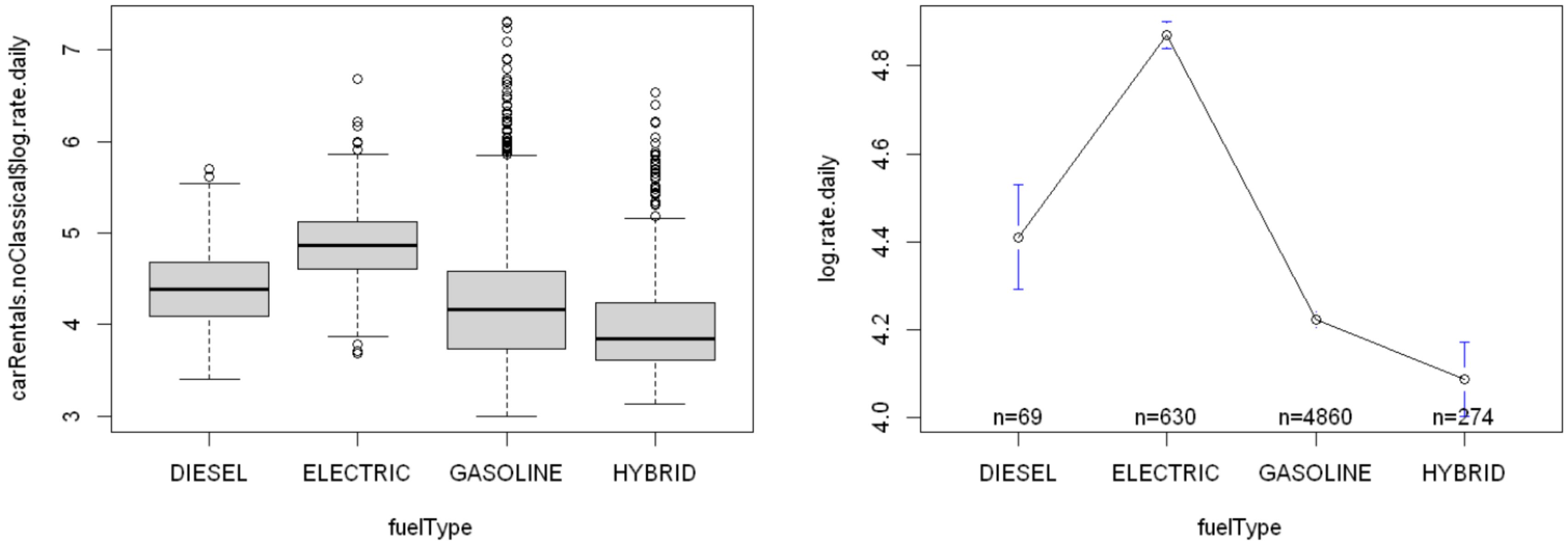


	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
vehicle.model	489	1878.3	3.841	38.27	<2e-16	***
Residuals	5343	536.3	0.100			

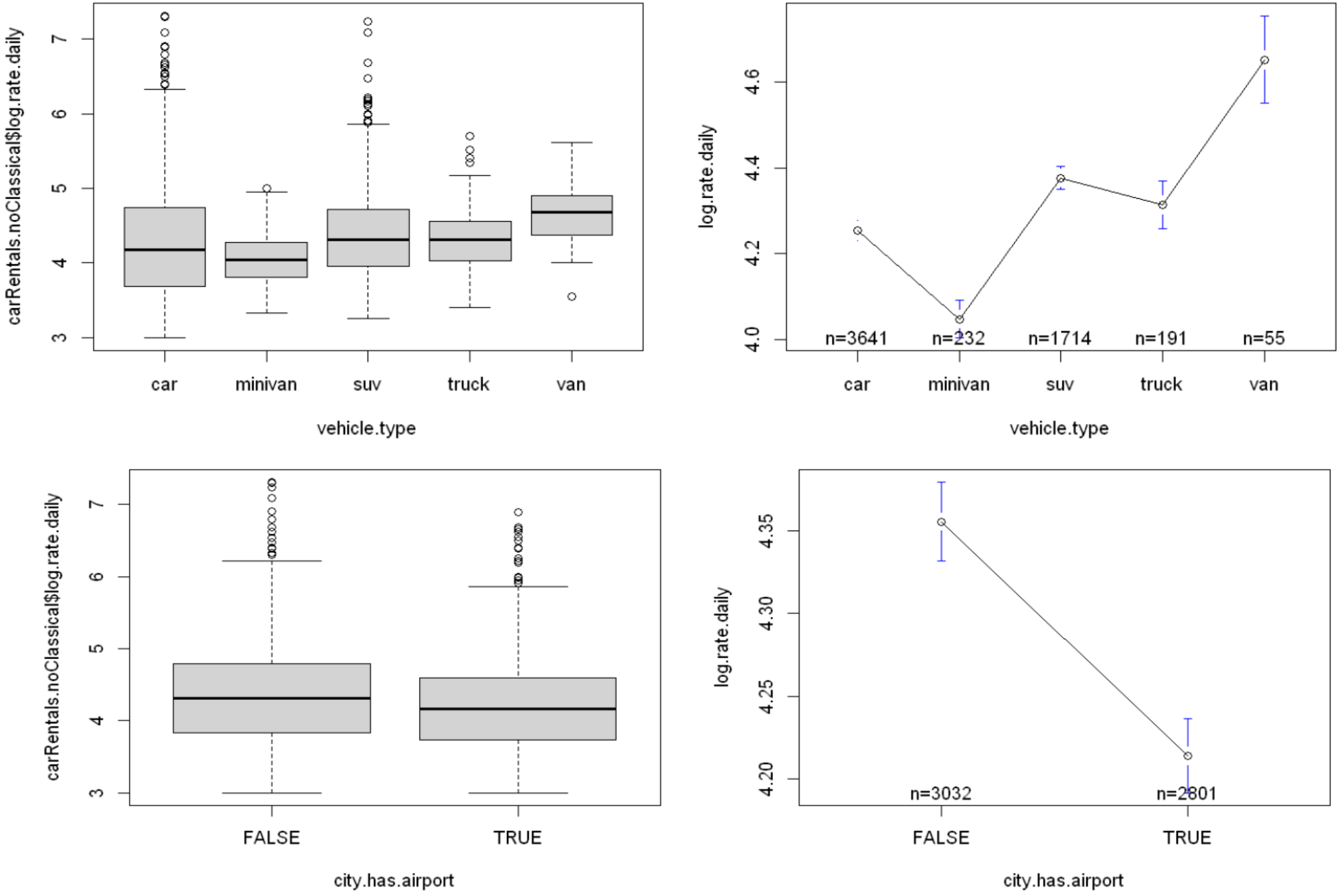


The type of fuel, vehicle type, and proximity to airports have significant effects on price (1/2)

- Electric cars are the most expensive on average
- We expect this to be due to a combination of:
  - purchase price level differences for rental companies
  - customer fuel costs – increased willingness to pay more on rental and less on fuel to reduce overall cost)



The type of fuel, vehicle type, and proximity to airports have significant effects on price (2/2)



Introduction

Behavior of daily rental rate

Effect of different features

Full Linear Regression

Other Explorations

The numerical features are not sufficient to produce the optimal linear regression model

```
Call:
lm(formula = log.rate.daily ~ vehicle.year + rating + reviewCount +
    renterTripsTaken, data = carRentals.noClassical)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3662 -0.4545 -0.0631  0.3684  3.2122

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -97.151973   5.155530 -18.844  < 2e-16 ***
vehicle.year       0.049598   0.002568  19.311  < 2e-16 ***
rating             0.302167   0.045948   6.576 5.28e-11 ***
reviewCount        0.007923   0.002120   3.738 0.000188 ***
renterTripsTaken  -0.007395   0.001781  -4.153 3.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6026 on 5333 degrees of freedom
  (495 observations deleted due to missingness)
Multiple R-squared:  0.0961,	Adjusted R-squared:  0.09542
F-statistic: 141.8 on 4 and 5333 DF,  p-value: < 2.2e-16
```

The categorical features yield better model's performance*

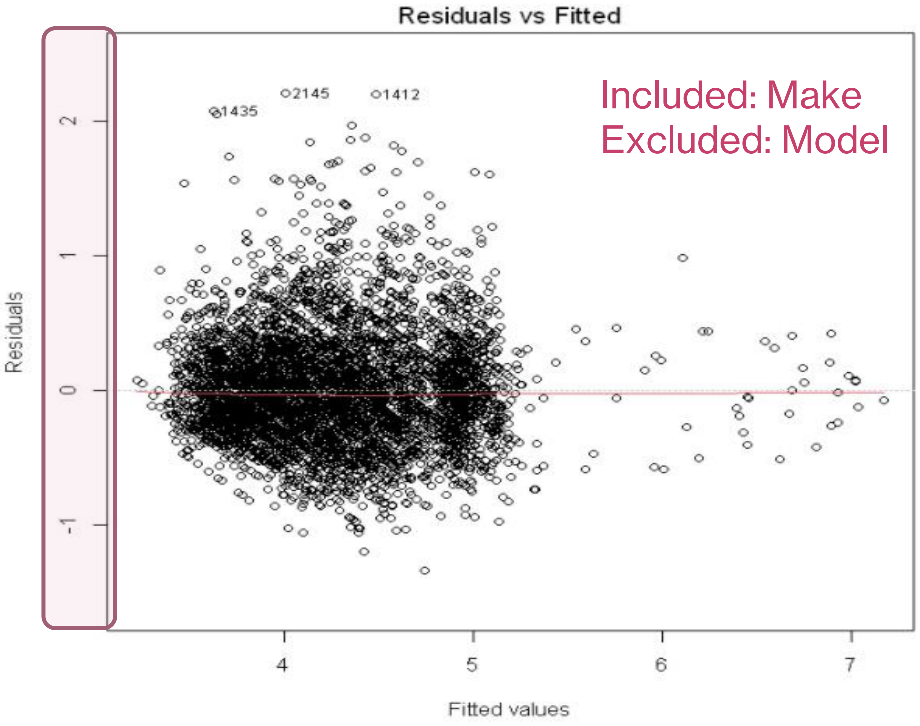
```
Call:
lm(formula = log.rate.daily ~ fuelType + vehicle.type + city.has.airport +
    vehicle.make, data = carRentals.noClassical)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4566 -0.2854 -0.0402  0.2335  3.2805

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.1051163  0.0846072  48.520  < 2e-16 ***
fuelTypeELECTRIC       0.0731256  0.0945423   0.773 0.439276
fuelTypeGASOLINE       0.0713591  0.0560565   1.273 0.203074
fuelTypeHYBRID         0.1542717  0.0631220   2.444 0.014554 *
vehicle.typeminivan    0.1798460  0.0334128   5.383 7.63e-08 ***
vehicle.typesuv        0.1668522  0.0147690  11.297  < 2e-16 ***
vehicle.typetruck      0.3492786  0.0358081   9.754  < 2e-16 ***
vehicle.typevan        0.4805187  0.0616219   7.798 7.42e-15 ***
city.has.airportTRUE  -0.0863223  0.0116565  -7.406 1.49e-13 ***
vehicle.makeAlfa Romeo 0.5144162  0.1018362   5.051 4.52e-07 ***
vehicle.makeAston Martin 1.8679034 0.2069371  9.026  < 2e-16 ***
          ...                ...       ...       ...      ...
```

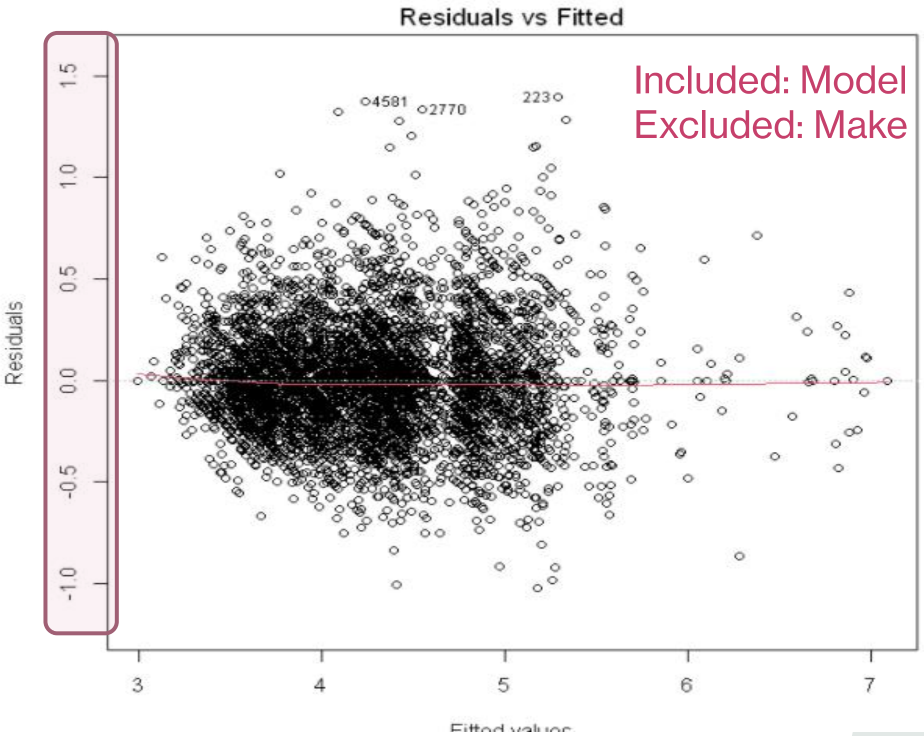
```
Residual standard error: 0.4408 on 5774 degrees of freedom
Multiple R-squared:  0.5353,    Adjusted R-squared:  0.5306
F-statistic: 114.7 on 58 and 5774 DF,  p-value: < 2.2e-16
```

* Excluding model type at this point

Overall model (with numeric features): Including car model yields better in-sample fit

- However, this is the effect of grouping observations into more dummy variables (51 brans vs 490 models ~ almost 10 times less per dummy)
- Further exploration of whether this leads to overfitting should be performed if the model is to be used for prediction purposes later on





Thank you