Excersice Sheet 7

```r
library(tidyverse)
library(forcats)
library(stringr)
library(purrr)
library(modelr)
```

A broker wants to use linear regression to find out which factors have a large influence on the price of a property. For this purpose, the variables described in Table 1 are given for the last 88 sales in the broker's region.

Table 1 House price record

Variabel	Description
`price`	house price ($\times$ 1,000 EUR)
`bdrms`	number bedrooms
`lotsize`	parking area (m^2)
`sqrm`	house area (m^2)
`country`	== 1 when in country house style
`lprice`	`log(price)`
`llotsize`	`log(lotsize)`
`lsqrm`	`log(sqrm)`

1. Create a linear regression model with `price` as dependent variable and `bdrms`, `lotsize`, `sqrm` und `country` as independent variables.
   a) Determine the regression coefficients and p-values of the dependent variable and compare their influence within the model on the predicted value for `price`.
   b) Determine how much variance of the dependent variable is explained.
   c) Check the residuals (graphically) for normal distribution and homoskedasticity.

```r
# Solution for Task 1
hprice <- read_csv(str_c(dirname(getwd()), "/Ex_7/hprice.csv"))

fit <- lm(price ~ bdrms + lotsize + sqrm + country, data = hprice)
sfit <- summary(fit)
sfit
```

```
##
## Call:
## lm(formula = price ~ bdrms + lotsize + sqrm + country, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.268  -38.271   -6.545   28.210  218.040
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.126528  29.603455  -0.815  0.41741
## bdrms        11.004292   9.515260   1.156  0.25080
## lotsize       0.022345   0.006918   3.230  0.00177 **
## sqrm          1.337325   0.143577   9.314 1.53e-14 ***
## country      13.715542  14.637265   0.937  0.35146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.88 on 83 degrees of freedom
## Multiple R-squared:  0.6758, Adjusted R-squared:  0.6602
## F-statistic: 43.25 on 4 and 83 DF,  p-value: < 2.2e-16
```

```r
# a
sfit$coefficients[, 1] #Koeffizienten
```

```
##  (Intercept)        bdrms      lotsize         sqrm      country
## -24.12652827  11.00429220   0.02234481   1.33732481  13.71554214
```

```r
sfit$coefficients[, 4] #p-Werte
```

```
##  (Intercept)        bdrms      lotsize         sqrm      country
## 4.174103e-01 2.507991e-01 1.774189e-03 1.534380e-14 3.514622e-01
```

```r
# Interpretation of the coefficients:
# The increase of the number of bedrooms by 1 leads to an increase of the
# predicted house price by 11000 EUR.
# The increase of the parking space by 1 m^2 leads to an increase of the
# predicted house price by 22 EUR.
# ... Analog for sqrm and country

# Since the independent variables are not scaled, we cannot compare
# the variable weighting using the regression coefficients. Therefore,
# we use the p-values. Influence after p-value: sqrm > lotsize > bdrms > country
```

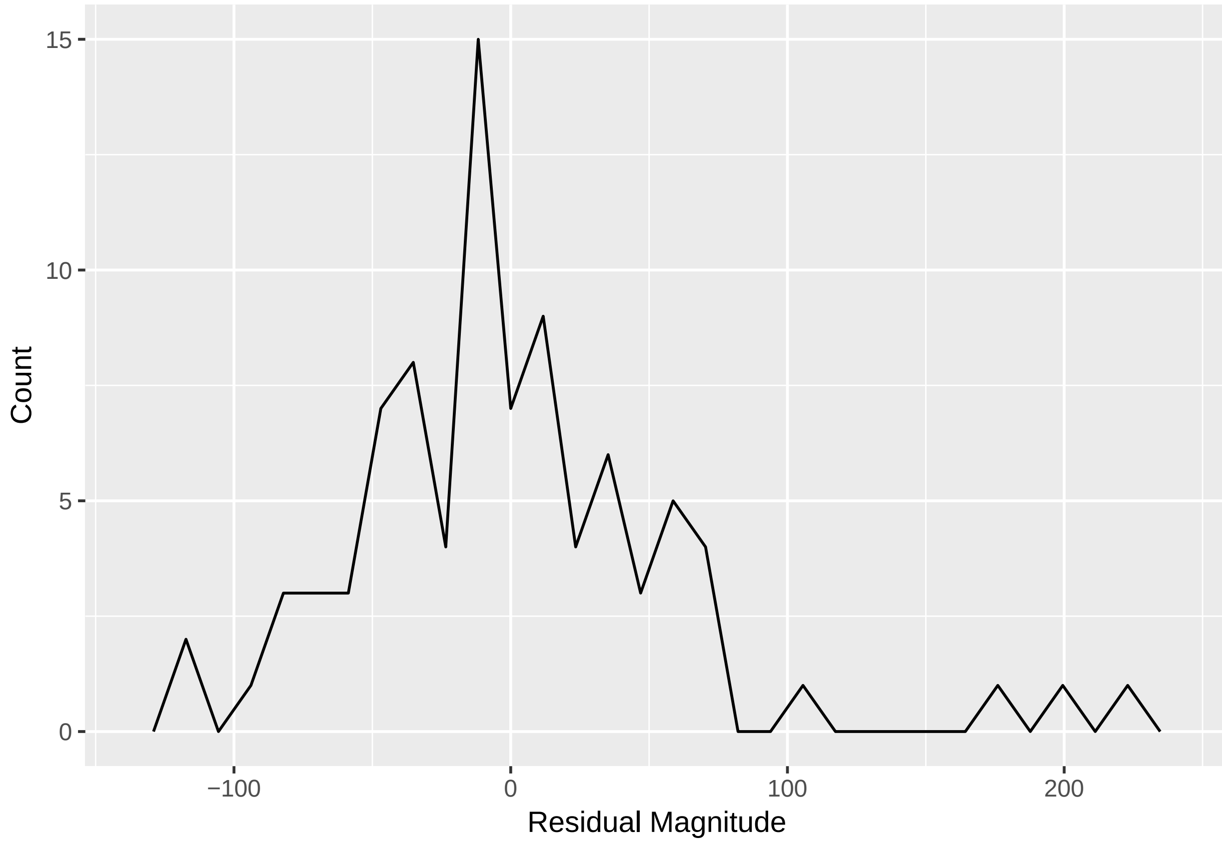
```r
# b
sfit$r.squared
```

```
## [1] 0.6757919
```

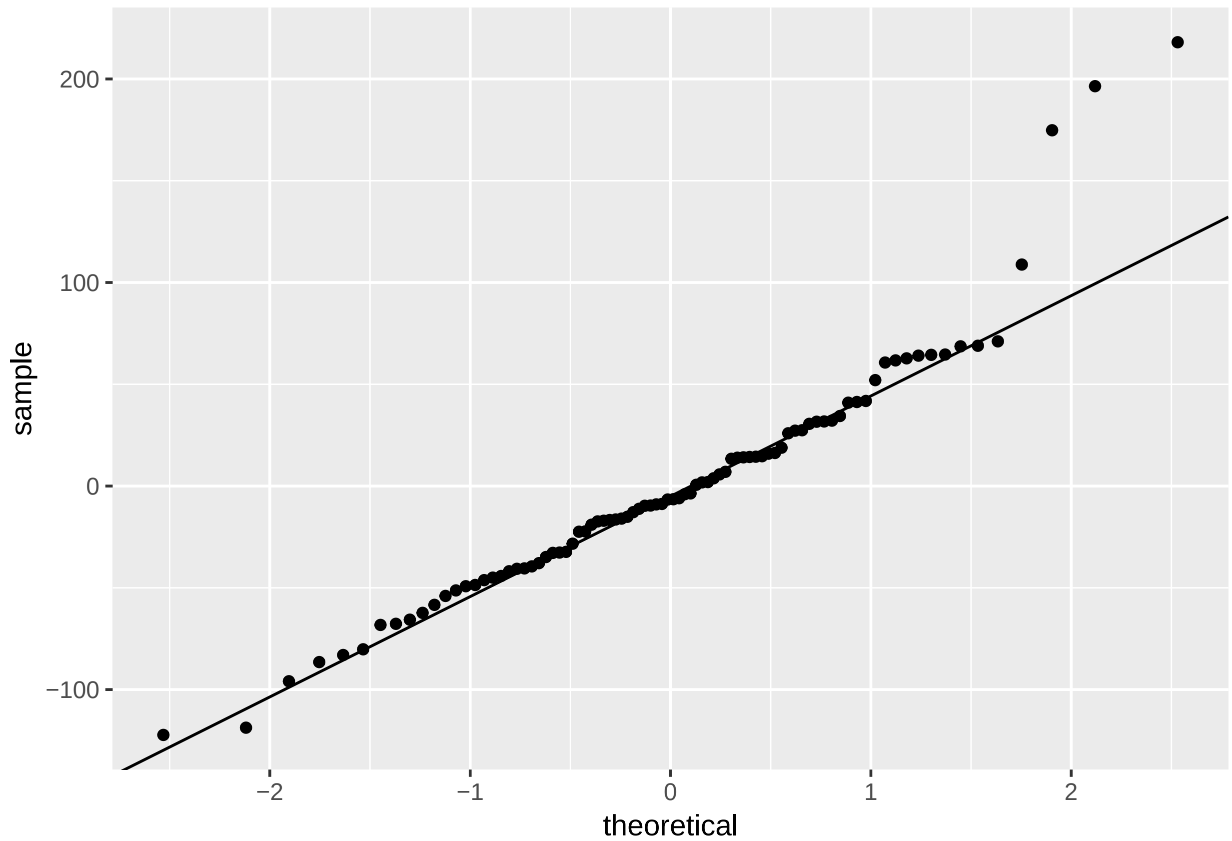
```r
# R^2 expresses the variability of the liniear model w.r.t.
# the price -> 1 - SSE/SSTO
#(Sum of Squared Errors vs. Total Sum of Squares)
```

```r
# c

# y_hat and add residuals to the data frame
hprice_res <- hprice %>%
  select(price, bdrms, lotsize, sqrm, country) %>%
  add_predictions(fit) %>%
  add_residuals(fit)
```

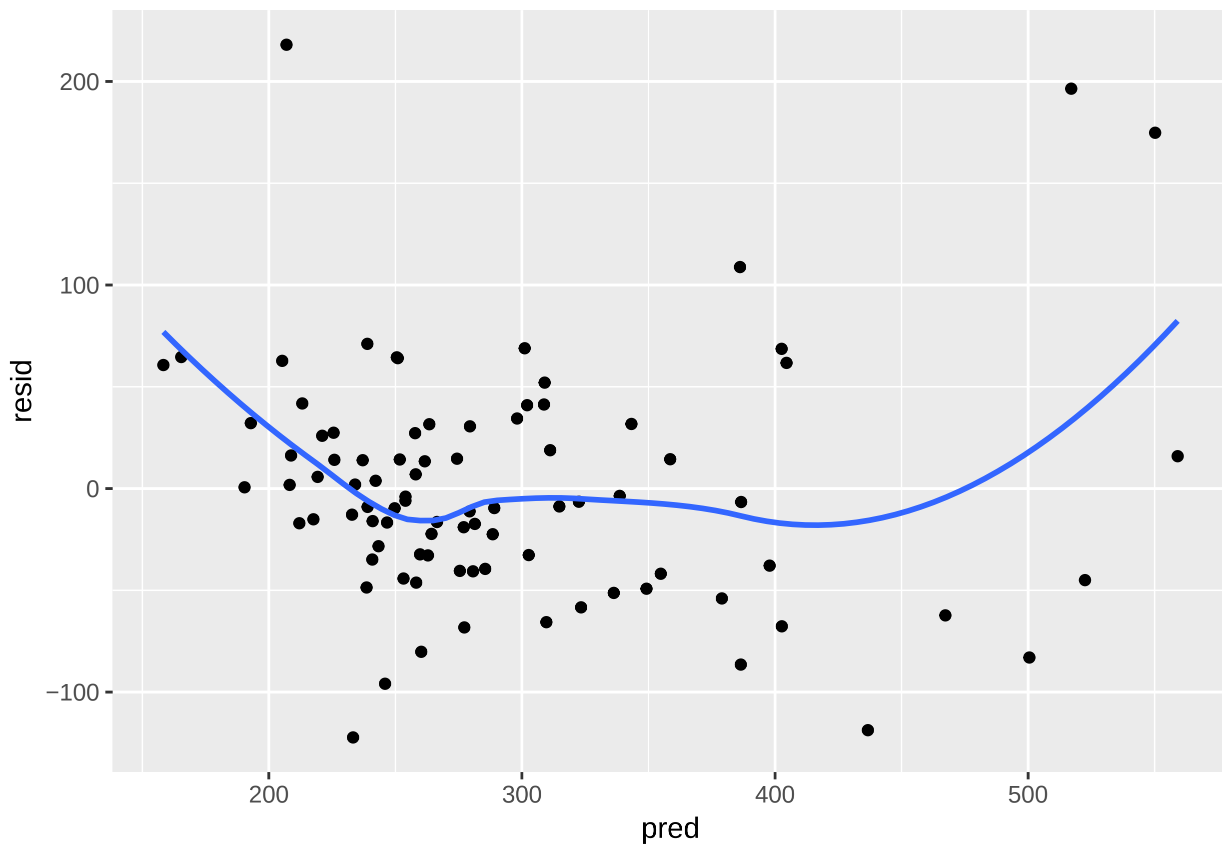
```r
# Frequency Chart of the distribution of the residuals
hprice_res %>%
  ggplot(aes(resid)) +
  geom_freqpoly() +
  labs(x = "Residual Magnitude", y = "Count")
```




```r
# qq-Plot to check the normal distribution
y <- quantile(hprice_res$resid, c(0.25, 0.75))
x <- qnorm(c(0.25,0.75))
slope <- diff(y) / diff(x)
int <- y[1] - slope * x[1]
ggplot(hprice_res, aes(sample = resid)) + stat_qq() +
  geom_abline(aes(slope = slope, intercept = int))
```




```r
# check homoskedasticity
hprice_res %>%
  ggplot(aes(pred, resid)) +
  geom_point() +
  geom_smooth(se = F)
```




```
# homoskedasticity (Increase/decrease in the variance of residuals) is not
# clearly identifiable
# For high predicted values, the variance seems to be greater, but the sample
# size is also very small.
```

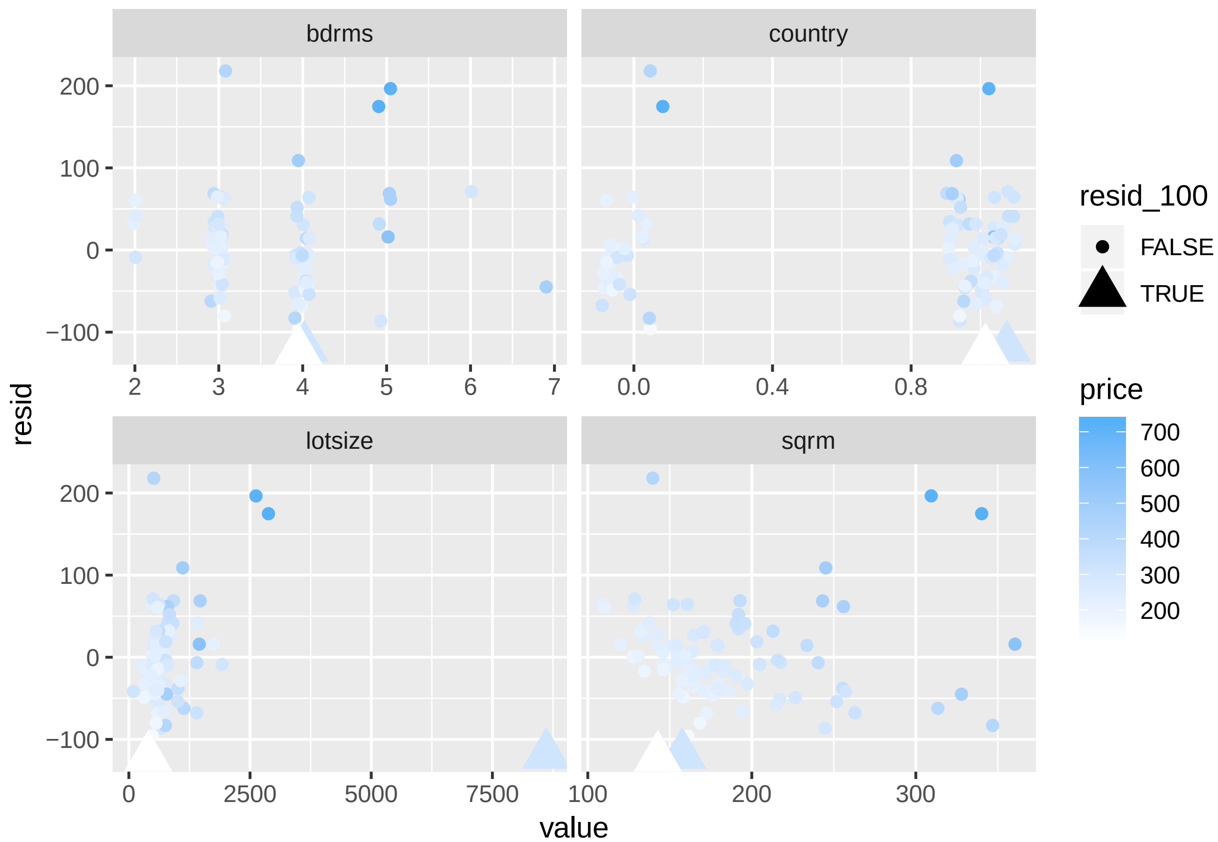
2. Given be the linear regression model from task 1.
   a) Create a scatterplot to display the relationship between the predicted value for `price` and the residual size.
   b) For some houses, the price forecast of the broker model is more than EUR 100,000 off. Highlight houses with a residual size of more than 100 or less than 100. What could be the reasons for high model inaccuracies?
   c) Can the R^2-value be increased by using a linear transformation of one of the independent variables?

```
# Solution for Task 2
# Houses with a residue size greater than 100
hprice_res %>% filter(abs(resid) > 100)
```

```
## # A tibble: 6 x 7
##   price bdrms lotsize  sqrm country  pred resid
##   <dbl> <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl>
## 1  714.     5   2623.  309.       1  517.  196.
## 2  495      4   1112.  245.       1  386.  109.
## 3  725      5   2880.  340.       0  550.  175.
```

```
## 4  425       3    513.  140.        0  207.  218.
## 5  318       4   8610.  158.        1  437. -119.
## 6  111       4    401.  143.        1  233. -122.
```

```r
# Representation of residue size by value per independent variable
hprice_res %>%
  mutate(resid_100 = ifelse(resid < -100, T, F)) %>%
  gather(variable, value, bdrms:country) %>%
  ggplot(aes(value, resid)) +
  geom_jitter(aes(color = price, shape = resid_100, size = resid_100), width = .1) +
  facet_wrap("variable", scales = "free_x") +
  scale_color_continuous(low = "white", high = "#56B1F7")
```



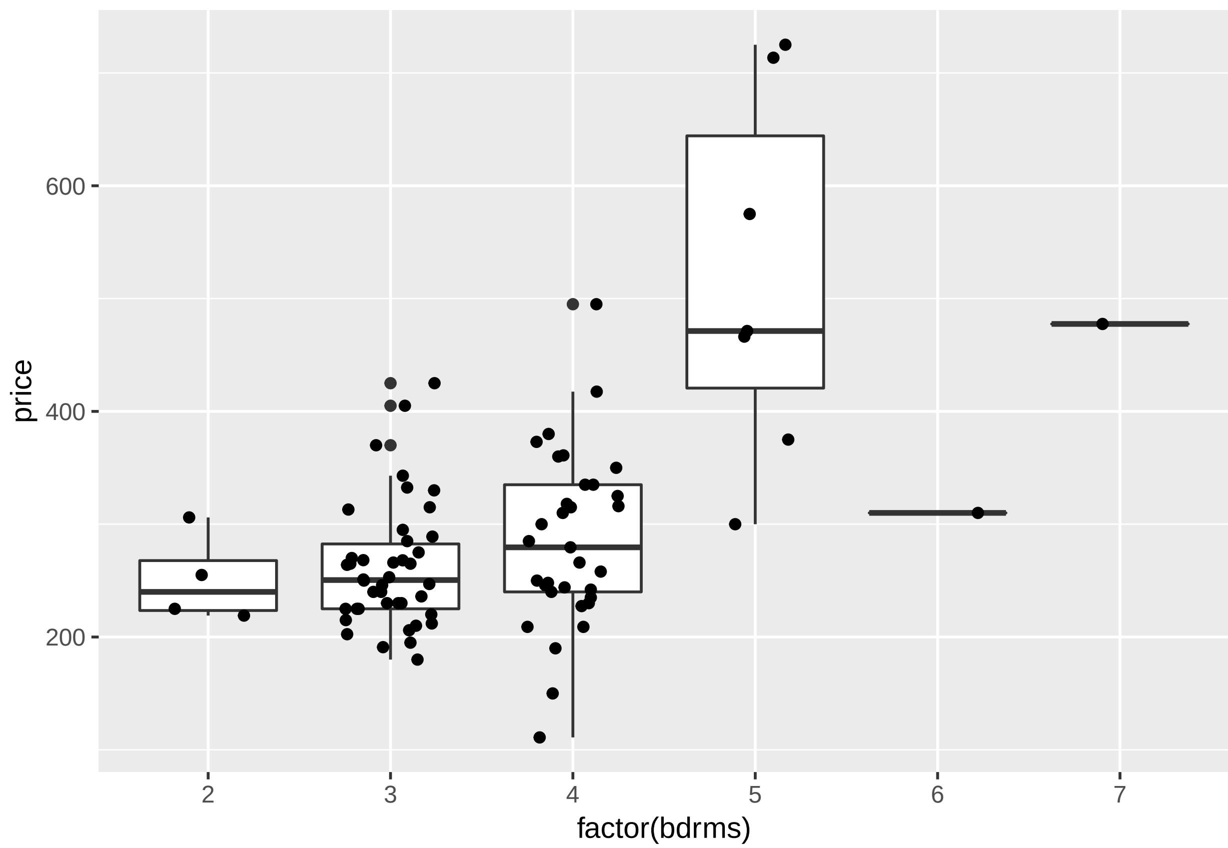

```r
# Outlier for lotsize == 92681 --> distorts the model <- remove outlier
# No connection between the independent variables of the model and the
# other 5 houses recognizable. There seems to be an unmeasured or
# unincluded variable that strongly pushes the price up/down.

# Log-Transformation of parking area (not linear) leads to increase of R^2-values
summary(lm(price ~ bdrms + llotsize  + sqrm + country, data = hprice))
```

```
##
## Call:
## lm(formula = price ~ bdrms + llotsize + sqrm + country, data = hprice)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.578  -35.841   -2.384   25.227  220.179
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -351.5476    75.0435  -4.685 1.08e-05 ***
## bdrms         13.3258     8.9490   1.489    0.140
## llotsize      55.9264    11.8093   4.736 8.89e-06 ***
## sqrm           1.1994     0.1404   8.543 5.36e-13 ***
## country       11.5790    13.7819   0.840    0.403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.37 on 83 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.6988
## F-statistic: 51.47 on 4 and 83 DF,  p-value: < 2.2e-16
```

3. Graphically display the relationship between `bdrms` and `price`. Check whether this relationship is also reflected in the regression model from Task 1. Create a regression model with `bdrms` as the only independent variable. Compare the regression coefficients with those of the model from Task 1 and interpret the differences.

```r
# The regression coefficient in the model from Task 1 is 11 for bdrms
# What does regression coefficient mean?
# If the values of all other independent variables remain the same,
# then another bedroom leads to a price increase of 11 EUR.
# If the house area and parking space do not change, then another
# bedroom has no influence on the house price.
# This means that an existing room is split into two smaller rooms or
# a normal room is converted into a bedroom, while the living space remains the same.
# The causality is not given in the second model despite its significance.
ggplot(hprice, aes(x = factor(bdrms), y = price)) +
  geom_boxplot() +
  geom_jitter(width = .25) +
  geom_smooth(method = "lm", se = F)
```




```
hprice %>%
  group_by(bdrms) %>%
  summarize(mean(price))
```

```
## # A tibble: 6 x 2
##   bdrms `mean(price)`
##   <dbl>         <dbl>
## 1     2          251.
## 2     3          262.
## 3     4          285.
## 4     5          518.
## 5     6          310
## 6     7          478.
```

```
summary(lm(price ~ bdrms, data = hprice))
```

```
##
## Call:
## lm(formula = price ~ bdrms, data = hprice)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -209.33  -52.81   -7.83   32.20  342.65
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    72.23      41.55   1.738   0.0857 .
## bdrms          62.02      11.34   5.470 4.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.98 on 86 degrees of freedom
## Multiple R-squared:  0.2581, Adjusted R-squared:  0.2495
## F-statistic: 29.93 on 1 and 86 DF,  p-value: 4.344e-07
```

Dataset:

- http://isgwww.cs.uni-magdeburg.de/cv/lehre/VisAnalytics/material/exercise/datasets/hprice.csv