Chapter 3: multiple regression

We start again by loading the data.

```
# Load the datasets from the RData file
load("../dataCreated/listings_clean.RData")

# Display the first few rows of the cleaned dataset
head(listings_clean)
```

```
##   neighbourhood_cleansed           id
## 1          's-Gravenland 1.012115e+18
## 2          's-Gravenland 1.032655e+18
## 3           Agniesebuurt 8.217094e+17
## 4           Agniesebuurt 9.846371e+17
## 5           Agniesebuurt 7.665157e+17
## 6           Agniesebuurt 6.455105e+17
##                                      listing_url    scrape_id last_scraped
## 1 https://www.airbnb.com/rooms/1012114818168132512 2.024063e+13   2024-06-25
## 2 https://www.airbnb.com/rooms/1032655139418809797 2.024063e+13   2024-06-25
## 3  https://www.airbnb.com/rooms/821709375116932602 2.024063e+13   2024-06-26
## 4  https://www.airbnb.com/rooms/984637143310796841 2.024063e+13   2024-06-25
## 5  https://www.airbnb.com/rooms/766515651027141909 2.024063e+13   2024-06-25
## 6  https://www.airbnb.com/rooms/645510482091510278 2.024063e+13   2024-06-25
##        source                                                name
## 1 city scrape    High-end Serviced Apartment with two bedrooms
## 2 city scrape     High-end Serviced Apartment with one bedroom
## 3 city scrape                                         King Room
## 4 city scrape            Charming Urban Retreat - with parking
## 5 city scrape                       Bed in Mixed Dormitory Room
## 6 city scrape Studio floor with own kitchen & toilet - central
##
## 1
## 2
## 3
## 4              Indulge in the allure of our charming centrally located urban retreat. Enjoy the sunl
## 5                                                                 Sleeping in a dorm room is not or
## 6 10MIN WALK FROM CENTRAL.<br />We have 2 floor apartment and we are renting out our upstairs which i
##
## 1
## 2
## 3
## 4
## 5 Within walking distance to Rotterdam Centraal Station and just a short drive from Rotterdam The Hag
## 6
##
## 1 https://a0.muscache.com/pictures/miso/Hosting-1012114818168132512/original/12df4128-3a61-42a2-a47a-
## 2 https://a0.muscache.com/pictures/miso/Hosting-1032655139418809797/original/69761ba0-3d06-4760-a1e3-
## 3  https://a0.muscache.com/pictures/miso/Hosting-821709375116932602/original/3e333f5e-b739-4f17-873f-
```

```
## 4  https://a0.muscache.com/pictures/miso/Hosting-984637143310796841/original/b16e8b3a-e987-4119-b146-
## 5  https://a0.muscache.com/pictures/miso/Hosting-766515651027141909/original/7b624fb0-bb29-4055-a801-
## 6                                                https://a0.muscache.com/pictures/b3738de7-f406-438a-ab2
##     host_id                                    host_url           host_name
## 1 543381697 https://www.airbnb.com/users/show/543381697 BirdsEye Short Stay
## 2 543381697 https://www.airbnb.com/users/show/543381697 BirdsEye Short Stay
## 3 365217950 https://www.airbnb.com/users/show/365217950              Bianca
## 4 116041691 https://www.airbnb.com/users/show/116041691            Bastiaan
## 5 365217950 https://www.airbnb.com/users/show/365217950              Bianca
## 6  16080777  https://www.airbnb.com/users/show/16080777               Sevim
##   host_since          host_location
## 1 2023-10-25                   <NA>
## 2 2023-10-25                   <NA>
## 3 2020-08-31                   <NA>
## 4 2017-02-12 Rotterdam, Netherlands
## 5 2020-08-31                   <NA>
## 6 2014-05-28 Rotterdam, Netherlands
##
## 1
## 2
## 3
## 4
## 5
## 6 We are a young couple living in Rotterdam both working in corporate. I am Turkish and my boyfriend
##   host_response_time host_response_rate host_acceptance_rate host_is_superhost
## 1     within an hour                96%                  82%              TRUE
## 2     within an hour                96%                  82%              TRUE
## 3     within an hour                85%                  99%             FALSE
## 4     within an hour               100%                  94%              TRUE
## 5     within an hour                85%                  99%             FALSE
## 6 within a few hours               100%                  21%             FALSE
##
## 1 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 2 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 3                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 4                          https://a0.muscache.com/im/pictures/user/db5760d4-54b8-467b-949a-45f95150a0
## 5                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 6                          https://a0.muscache.com/im/pictures/user/0d800f6b-1061-4a87-b4d3-cd0d70671a
##
## 1 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 2 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 3                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 4                          https://a0.muscache.com/im/pictures/user/db5760d4-54b8-467b-949a-45f95150a0
## 5                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 6                          https://a0.muscache.com/im/pictures/user/0d800f6b-1061-4a87-b4d3-cd0d70671a
##   host_neighbourhood host_listings_count host_total_listings_count
## 1               <NA>                   2                         2
## 2               <NA>                   2                         2
## 3               <NA>                  13                        13
## 4               <NA>                   1                         1
## 5               <NA>                  13                        13
## 6               <NA>                   2                         2
##                host_verifications host_has_profile_pic host_identity_verified
## 1               ['email', 'phone']                 TRUE                   TRUE
```

```
## 2               ['email', 'phone']                 TRUE                   TRUE
## 3                        ['phone']                 TRUE                   TRUE
## 4 ['email', 'phone', 'work_email']                 TRUE                   TRUE
## 5                        ['phone']                 TRUE                   TRUE
## 6 ['email', 'phone', 'work_email']                 TRUE                   TRUE
##                          neighbourhood neighbourhood_group_cleansed latitude
## 1 Rotterdam, Zuid-Holland, Netherlands              Prins Alexander 51.92200
## 2 Rotterdam, Zuid-Holland, Netherlands              Prins Alexander 51.92044
## 3                                 <NA>                        Noord 51.92851
## 4                                 <NA>                        Noord 51.92876
## 5 Rotterdam, Zuid-Holland, Netherlands                        Noord 51.92811
## 6 Rotterdam, Zuid-Holland, Netherlands                        Noord 51.93126
##   longitude        property_type       room_type accommodates bathrooms
## 1  4.553830   Entire rental unit Entire home/apt            4       1.0
## 2  4.554210   Entire rental unit Entire home/apt            2       1.0
## 3  4.474124        Room in hotel    Private room            2       1.0
## 4  4.474194   Entire rental unit Entire home/apt            4       1.0
## 5  4.475240 Shared room in hotel     Shared room            1       1.0
## 6  4.474320         Entire condo Entire home/apt            2       1.5
##   bathrooms_text bedrooms beds
## 1         1 bath        2    2
## 2         1 bath        1    1
## 3 1 private bath        1    1
## 4         1 bath        2    3
## 5  1 shared bath        1    2
## 6      1.5 baths        1    1
##
## 1 ["Private patio or balcony", "Hot water kettle", "Central heating", "Paid street parking off premis
## 2 ["Private patio or balcony", "Hot water kettle", "Central heating", "Paid street parking off premis
## 3
## 4
## 5
## 6
##   price minimum_nights maximum_nights minimum_minimum_nights
## 1   259             3            20                      2
## 2   204             3            20                      2
## 3   198             1           365                      1
## 4   253             2           365                      2
## 5    75             1           365                      1
## 6   116             7            60                      2
##   maximum_minimum_nights minimum_maximum_nights maximum_maximum_nights
## 1                      3                    20                    20
## 2                      3                    20                    20
## 3                      1                  1000                  1000
## 4                     27                   365                   500
## 5                      1                  1000                  1000
## 6                      7                    60                    60
##   minimum_nights_avg_ntm maximum_nights_avg_ntm calendar_updated
## 1                    2.2                  20.0               NA
## 2                    2.2                  20.0               NA
## 3                    1.0                1000.0               NA
## 4                    5.9                 385.8               NA
## 5                    1.0                1000.0               NA
## 6                    6.8                  60.0               NA
```

```
##   has_availability availability_30 availability_60 availability_90
## 1             TRUE              17              32              59
## 2             TRUE              26              56              86
## 3             TRUE              28              58              88
## 4             TRUE              15              20              20
## 5             TRUE              23              51              81
## 6             TRUE              11              24              54
##   availability_365 calendar_last_scraped number_of_reviews
## 1              334           2024-06-25                89
## 2              361           2024-06-25                87
## 3              363           2024-06-26                12
## 4               20           2024-06-25                32
## 5              356           2024-06-25                55
## 6              144           2024-06-25                16
##   number_of_reviews_ltm number_of_reviews_l30d first_review last_review
## 1                    89                    19   2023-12-29  2024-06-20
## 2                    87                     8   2023-12-31  2024-06-16
## 3                     9                     1   2023-04-23  2024-05-30
## 4                    32                     2   2023-10-14  2024-06-02
## 5                    34                     0   2022-12-12  2024-05-18
## 6                     3                     0   2022-06-18  2024-04-03
##   review_scores_rating review_scores_accuracy review_scores_cleanliness
## 1                 4.84                  4.81                       4.89
## 2                 4.86                  4.86                       4.90
## 3                 4.42                  4.67                       4.33
## 4                 4.81                  4.81                       4.72
## 5                 4.33                  4.47                       4.47
## 6                 4.94                  4.88                       4.88
##   review_scores_checkin review_scores_communication review_scores_location
## 1                  4.76                       4.85                    4.60
## 2                  4.69                       4.83                    4.69
## 3                  4.67                       4.00                    4.08
## 4                  4.91                       4.84                    4.84
## 5                  4.76                       4.44                    4.64
## 6                  4.94                       5.00                    4.81
##   review_scores_value                 license instant_bookable
## 1                4.71 0599 F23E 3F4E 6EF1 B897           FALSE
## 2                4.72 0599 9BD6 36FC CB31 F331            TRUE
## 3                3.83                  Exempt            TRUE
## 4                4.69 0599 8A16 99F9 742E C8F8           FALSE
## 5                4.36                  Exempt            TRUE
## 6                4.63 0599 B71E 378D 56FF 1829           FALSE
##   calculated_host_listings_count calculated_host_listings_count_entire_homes
## 1                              2                                           2
## 2                              2                                           2
## 3                             13                                           0
## 4                              1                                           1
## 5                             13                                           0
## 6                              2                                           2
##   calculated_host_listings_count_private_rooms
## 1                                            0
## 2                                            0
## 3                                           12
## 4                                            0
```

```
## 5                                          12
## 6                                           0
##   calculated_host_listings_count_shared_rooms reviews_per_month Coolness
## 1                                           0             14.83        6
## 2                                           0             14.66        6
## 3                                           1              0.84        6
## 4                                           0              3.75        6
## 5                                           1              2.94        6
## 6                                           0              0.65        6
##   Centrality Quietness Fanciness
## 1          5         6         5
## 2          5         6         5
## 3          6         7         5
## 4          6         7         5
## 5          6         7         5
## 6          6         7         5
```

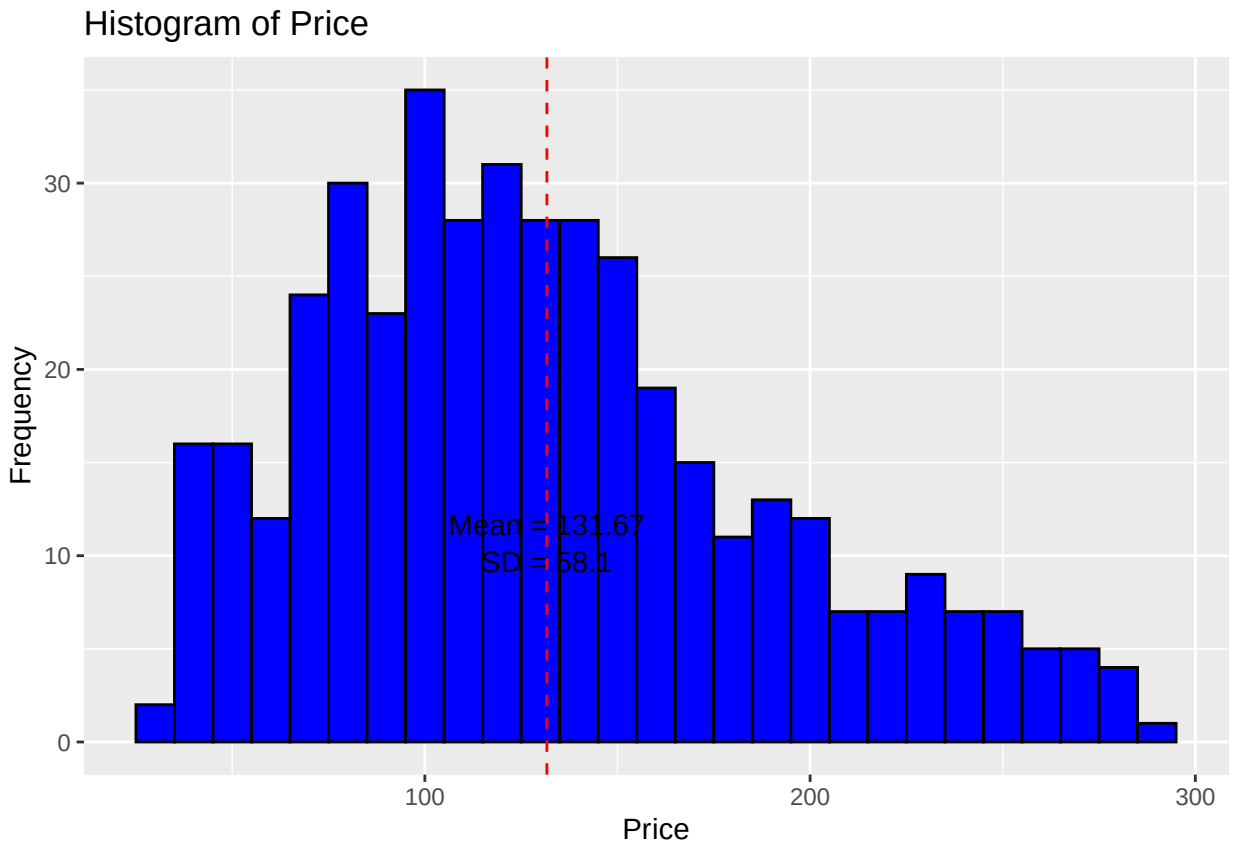
Filter the data so that we use only listings for at most 6 people. From now on use this

```
listings_clean_filtered <- listings_clean %>%
  filter(accommodates <= 6)
```

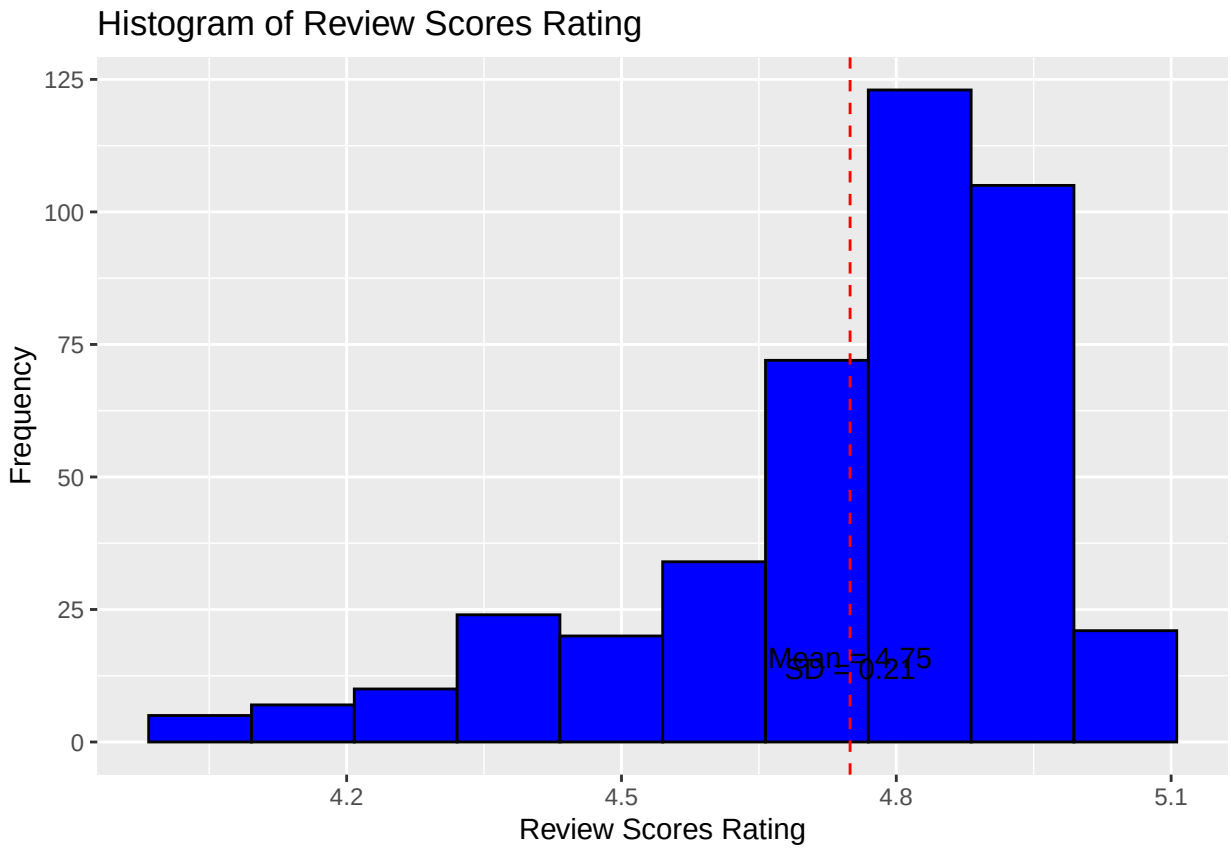
We want to understand what leads to good ratings. Our idea is that better places are also more expensive, so a high price should predict better ratings.

Let's start with two histograms, so that we understand how our variables look:

```
# Histogram of the price with mean and standard deviation
ggplot(listings_clean_filtered, aes(x = price)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  labs(title = "Histogram of Price", x = "Price", y = "Frequency") +
  geom_vline(aes(xintercept = mean(price, na.rm = TRUE)),
             color = "red", linetype = "dashed") +
  annotate("text", x = mean(listings_clean_filtered$price, na.rm = TRUE),
           y = 10, label = paste("Mean =", round(mean(listings_clean_filtered$price, na.rm = TRUE), 2)),
           vjust = -1) +
  annotate("text", x = mean(listings_clean_filtered$price, na.rm = TRUE),
           y = 8, label = paste("SD =", round(sd(listings_clean_filtered$price, na.rm = TRUE), 2)),
           vjust = -1)
```




```
# Histogram of the review_scores_rating with 10 bins, mean, and standard deviation
ggplot(listings_clean_filtered, aes(x = review_scores_rating)) +
  geom_histogram(bins = 10, fill = "blue", color = "black") +
  labs(title = "Histogram of Review Scores Rating", x = "Review Scores Rating", y = "Frequency") +
  geom_vline(aes(xintercept = mean(review_scores_rating, na.rm = TRUE)),
             color = "red", linetype = "dashed") +
  annotate("text", x = mean(listings_clean_filtered$review_scores_rating, na.rm = TRUE),
           y = 10, label = paste("Mean =", round(mean(listings_clean_filtered$review_scores_rating, na.
           vjust = -1) +
  annotate("text", x = mean(listings_clean_filtered$review_scores_rating, na.rm = TRUE),
           y = 8, label = paste("SD =", round(sd(listings_clean_filtered$review_scores_rating, na.rm =
           vjust = -1)
```



We start with a simple model where we regress the rating on the price.

```
model1 <- lm(review_scores_rating ~ price, data = listings_clean_filtered)
summary(model1)
```

```
##
## Call:
## lm(formula = review_scores_rating ~ price, data = listings_clean_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74826 -0.08111  0.04972  0.16237  0.26352
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.7245161  0.0256149 184.444   <2e-16 ***
## price       0.0001899  0.0001780   1.067    0.287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.212 on 419 degrees of freedom
## Multiple R-squared:  0.00271,    Adjusted R-squared:  0.0003294
## F-statistic: 1.138 on 1 and 419 DF,  p-value: 0.2866
```

```
coef_model1 <- coef(model1)  # save coefficients
```

That's already interesting. It says that a place that for a place that costs 58 dollars more per night (that's one standard deviation of the price, see first histogram above) we can expect a rating that is about 0.0001899*58

= 0.011 higher. This is a relatively small effect (recall from the second histogram above that ratings have a standard deviation of 0.21).

However, we also recall that earlier, we have already found out that the price is highly correlated with how many people can stay in an apartment. Recall that we can describe this with a regression as well:

```
model2 <- lm(price ~ accommodates, data = listings_clean_filtered)
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ accommodates, data = listings_clean_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -144.65  -33.46  -10.91   22.35  159.90
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    60.731      5.735   10.59   <2e-16 ***
## accommodates   25.183      1.855   13.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.48 on 419 degrees of freedom
## Multiple R-squared:  0.3055, Adjusted R-squared:  0.3038
## F-statistic: 184.3 on 1 and 419 DF,  p-value: < 2.2e-16
```

```
coef_model2 <- coef(model2)  # save coefficients
```

We see that if one additional person can stay in a place, then we can expect a price that is about 25 dollars higher per night. So the question is whether `accommodates` also has an influence on the rating. Maybe apartments where more people can stay in are a bit less fancy and therefore the rating is lower?

To find out, we estimate a model where we regress `review_scores_rating` on `price` and `accommodates`.

```
model3 <- lm(review_scores_rating ~ price + accommodates, data = listings_clean_filtered)
summary(model3)
```

```
##
## Call:
## lm(formula = review_scores_rating ~ price + accommodates, data = listings_clean_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76796 -0.07488  0.04719  0.15837  0.28402
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7571610  0.0280100 169.838  < 2e-16 ***
## price         0.0005146  0.0002119   2.428  0.01560 *
## accommodates -0.0267614  0.0096558  -2.772  0.00583 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2103 on 418 degrees of freedom
## Multiple R-squared:  0.02071,    Adjusted R-squared:  0.01602
```

```
## F-statistic: 4.419 on 2 and 418 DF,  p-value: 0.01261
```

```
coef_model3 <- coef(model3)  # save coefficients
```

This now shows that the coefficient on price increases from 1.9×10^{-4} to 5.15×10^{-4}.

Finally, we estimate a richer model where we regress `review_scores_rating` on `price`, `accommodates`, and 4 neighborhood characteristics.

```
model <- lm(review_scores_rating ~ price + accommodates + Centrality + Quietness + Coolness + Fanciness
summary(model)
```

```
##
## Call:
## lm(formula = review_scores_rating ~ price + accommodates + Centrality +
##     Quietness + Coolness + Fanciness, data = listings_clean_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77079 -0.07837  0.04187  0.15522  0.30276
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0022506  0.2700367  18.524  < 2e-16 ***
## price         0.0005828  0.0002155   2.705  0.00712 **
## accommodates -0.0298174  0.0097644  -3.054  0.00241 **
## Centrality   -0.0358945  0.0242109  -1.483  0.13895
## Quietness    -0.0452547  0.0245493  -1.843  0.06598 .
## Coolness     -0.0193393  0.0429716  -0.450  0.65291
## Fanciness     0.0644724  0.0296859   2.172  0.03044 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2098 on 414 degrees of freedom
## Multiple R-squared:  0.03446,    Adjusted R-squared:  0.02046
## F-statistic: 2.462 on 6 and 414 DF,  p-value: 0.02373
```