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 also create a variable `review_scores_rating_standardized` with the standardized review score. This helps with the interpretation.

```
# Standardize review_scores_rating
listings_clean_filtered <- listings_clean_filtered %>%
  mutate(review_scores_rating_standardized =
           (review_scores_rating - mean(review_scores_rating, na.rm = TRUE)) /
           sd(review_scores_rating, na.rm = TRUE))
```

We still want to understand what prices depend on. We start with a simple model where we regress the log of the price on the standardized rating.

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

```
##
## Call:
## lm(formula = log(price) ~ review_scores_rating_standardized,
##     data = listings_clean_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45438 -0.29944  0.03945  0.32261  0.95273
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        4.77677    0.02300 207.709   <2e-16 ***
## review_scores_rating_standardized  0.02301    0.02302   0.999    0.318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4719 on 419 degrees of freedom
## Multiple R-squared:  0.002378,   Adjusted R-squared:  -3.252e-06
## F-statistic: 0.9986 on 1 and 419 DF,  p-value: 0.3182
```

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

That says that for a place that has a rating that is one standard deviation higher we can expect to pay a price that is 2.3 percent higher.

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 (we found a coefficient of 0.21 in a regression of log price on 'accommodates'). So maybe that plays a role, too.

To find out, we carry out a multiple regression where we regress the log of the price on both the standardized rating and how many people can stay in it.

```
model2 <- lm(log(price) ~ review_scores_rating_standardized + accommodates, data = listings_clean_filte
summary(model2)
```

```
##
## Call:
## lm(formula = log(price) ~ review_scores_rating_standardized +
##     accommodates, data = listings_clean_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5060 -0.2414 -0.0027  0.2186  0.9949
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.17715    0.04590  90.999   <2e-16 ***
## review_scores_rating_standardized  0.04554    0.01894   2.404   0.0166 *
## accommodates                      0.21285    0.01486  14.327   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3869 on 418 degrees of freedom
## Multiple R-squared:  0.3309, Adjusted R-squared:  0.3277
## F-statistic: 103.4 on 2 and 418 DF,  p-value: < 2.2e-16
```

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

This now shows that the predicted price increases even more when the standardized rating increases by one unit, about 4.6 percent. We also see that if one additional person can stay in a place, then we can expect a price that is about 21 percent higher per night (this is similar to what we found in the Chapter 2 analysis before when we regressed the log price on 'accommodates').

To explain why the coefficient on `review_scores_rating_standardized` changes, we regress `accommodates` on `review_scores_rating_standardized`.

```
model3 <- lm(accommodates ~ review_scores_rating_standardized, data = listings_clean_filtered)
summary(model3)
```

```
##
## Call:
## lm(formula = accommodates ~ review_scores_rating_standardized,
##     data = listings_clean_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1964 -0.8069 -0.7170  1.1532  3.3080
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.81710    0.06201  45.433   <2e-16 ***
## review_scores_rating_standardized -0.10587    0.06208  -1.705   0.0889 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.272 on 419 degrees of freedom
## Multiple R-squared:  0.006894,   Adjusted R-squared:  0.004523
## F-statistic: 2.909 on 1 and 419 DF,  p-value: 0.08885
```

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

This shows that places with a better rating, on average, accommodate less people.

In the lecture, we have learned that from the results in `model2` and `model3`, we can reconstruct what we found in `model1`.

In `model1`, we look at the change in the predicted price when `review_scores_rating_standardized` changes by one unit (and there are no other variables in the model). In `model3`, we see that when `review_scores_rating_standardized` changes by one unit, then `accommodates` is predicted to change by -0.10587 units. But we know from `model2` that when `accommodates` changes by one unit, we predict the log price to be higher by 0.21285 units when we hold the rating fixed (ceteris paribus).

So, we can go from `model2` to `model1` by saying that the predicted change in the log price when the rating changes by one standard deviation is the ceteris paribus effect of a change in `review_scores_rating_standardized`, 0.04554, plus how much we predict `accommodates` to change when `review_scores_rating_standardized` changes by one unit, -0.10587, times the ceteris paribus effect of `accommodates`, 0.21285.

Here is some R code that does this calculation:

```
# Extract coefficients from model3 and model2
coef_rating_model2 <- coef(model2)["review_scores_rating_standardized"]
coef_accommodates_model2 <- coef(model2)["accommodates"]
coef_rating_model3 <- coef(model3)["review_scores_rating_standardized"]

# Calculate the desired value
result <- coef_rating_model2 + coef_rating_model3*coef_accommodates_model2

# Print the result
round(result,5)
```

```
## review_scores_rating_standardized
##                           0.02301
```

This is the coefficient on `review_scores_rating_standardized` in `model1`.

Finally, we also make use of data on some neighborhood characteristics. These neighborhood characterstics were generated by chatGPT and are on a scale from 0 to 10. I've merged them with the web-scraped Airbnb data when I read in the data set. See build code for details.

Let's first look at some summary statistics.

```
# Calculate means and standard deviations for the four variables
summary_stats <- listings_clean_filtered %>%
  summarise(
    Centrality_Mean = mean(Centrality, na.rm = TRUE),
```

```
    Centrality_SD = sd(Centrality, na.rm = TRUE),
    Quietness_Mean = mean(Quietness, na.rm = TRUE),
    Quietness_SD = sd(Quietness, na.rm = TRUE),
    Coolness_Mean = mean(Coolness, na.rm = TRUE),
    Coolness_SD = sd(Coolness, na.rm = TRUE),
    Fanciness_Mean = mean(Fanciness, na.rm = TRUE),
    Fanciness_SD = sd(Fanciness, na.rm = TRUE)
  )

# Reshape to have one column for means and another for standard deviations
summary_stats_tidy <- data.frame(
  Variable = c("Centrality", "Quietness", "Coolness", "Fanciness"),
  Mean = c(summary_stats$Centrality_Mean, summary_stats$Quietness_Mean, summary_stats$Coolness_Mean, su
  SD = c(summary_stats$Centrality_SD, summary_stats$Quietness_SD, summary_stats$Coolness_SD, summary_sta
)

# Display the table
summary_stats_tidy
```

```
##     Variable     Mean        SD
## 1 Centrality 6.009501 1.6846379
## 2  Quietness 5.691211 1.0440962
## 3   Coolness 6.536817 1.0266443
## 4  Fanciness 5.494062 0.8718278
```

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

```
model4 <- lm(log(price) ~ review_scores_rating_standardized + accommodates + Centrality + Quietness + C
summary(model4)
```

```
##
## Call:
## lm(formula = log(price) ~ review_scores_rating_standardized +
##     accommodates + Centrality + Quietness + Coolness + Fanciness,
##     data = listings_clean_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48845 -0.24997  0.01026  0.23530  0.94151
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.10015    0.49275   6.292    8e-10 ***
## review_scores_rating_standardized  0.05000    0.01887   2.650  0.00837 **
## accommodates                       0.21641    0.01480  14.625  < 2e-16 ***
## Centrality                         0.08591    0.04406   1.950  0.05186 .
## Quietness                          0.12896    0.04451   2.897  0.00397 **
## Coolness                           0.09606    0.07834   1.226  0.22079
## Fanciness                         -0.14765    0.05382  -2.744  0.00634 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3834 on 414 degrees of freedom
## Multiple R-squared:  0.3494, Adjusted R-squared:   0.34
```

```
## F-statistic: 37.05 on 6 and 414 DF,  p-value: < 2.2e-16
```

This shows that prices are predicted to be higher when chatGPT rates the neighborhood more central, more quiet, more cool, and less fancy. This calls for more analysis.

As a first step, let's look at the list that has the highest value of `Fanciness`:

```
# Get the top 10 unique neighborhoods with the highest Fanciness values
top_neighborhoods_unique <- listings_clean_filtered %>%
  distinct(neighbourhood_cleansed, .keep_all = TRUE) %>%
  arrange(desc(Fanciness)) %>%
  select(neighbourhood_cleansed, Fanciness) %>%
  head(10)

# Print the list
top_neighborhoods_unique
```

```
##    neighbourhood_cleansed Fanciness
## 1            Kop van Zuid         9
## 2          Kralingen Oost         9
## 3                Dijkzigt         8
## 4                Blijdorp         7
## 5             Katendrecht         7
## 6           Stadsdriehoek         7
## 7              Delfshaven         6
## 8             Liskwartier         6
## 9              Middelland         6
## 10          Noordereiland         6
```

At this point, it's unclear why `Fanciness` should lead to lower predicted prices. Feel free to do some analysis and let me know if you find out what could explain this!