Cleaning_file

Bo

2022-10-03

Cleaning InsideAirBnB data

This R script is used to clean the previously downloaded data from InsideAirBnB (through download_file) and prepare this data for further analysis of the research question. Among others, variables are altered, columns are added and datesets are merged within this part of the script.

Step 1) Merge the previously downloaded datasets, such that the "room_type" column will be present in the calender dataset. This can be done by use of a left_join, using the "id" column and "listing_id" column as reference columns:

```
merged_data <- calender_data %>%
  left_join(listing_data, by = c("listing_id" = "id"))
```

Step 2) Remove the NA's, because these values will interfere with further analysis:

```
merged_data_without_na <- na.omit(merged_data)
```

Step 3) Compute the "dates" column and add an extra column which transforms these date notations into days of the week by number, meaning Monday = 1, Tuesday = 2, etc. This will be used to more easily distinguish between weekends and weekdays during analysis:

```
merged_data_without_na$day_num <- format(merged_data_without_na$date,"%u")
merged_data_without_na$day_num <- as.numeric(merged_data_without_na$day_num)
```

Step 4) Create a vector of weekdays or weekends, to create a column that indicates whether a day is in the weekend or on a weekday (weekdays are Sunday, Monday, Tuesday, Wednesday and Thursday. Weekends are Friday and Saturday.):

```
weekdays1 <- c(1, 2, 3, 4, 7)
merged_data_without_na$wDay <- factor(((merged_data_without_na$day_num) %in% weekdays1), levels=c(FALSE
```

Step 5) Alter the price variable, such that it is formulated as a numeric without a dollar sign:

```
merged_data_without_na$price <- parse_number(merged_data_without_na$price)
```

After running this script, the final dataset should be cleaned and prepared for the analysis phase!