Airbnb Price prediction in Bangkok

Bruno Helmeczy

31/01/2021

Abstract

This report investigates the prices at which apartments in Bangkok can be rented out on Airbnb. Below I discuss a number of features taken into account, compare 6 prediction models of varying complexity, & after choosing & re-estimating my final model, I investigate expected performance across property types, the number of people accommodated, while investigating most important accommodation features. Based on 5-fold cross validation, I found the expanded Random Forest model predicting log-transformed prices to perform best in terms of all metrics. After re-estimation, the final model boasted 54.1% R-squared, & performed with 21 dollar MAE & 51.9 dollar RMSE, ca. 95% of the average price during out-of-sample testing. Finally, model diagnostics showed the model to perform fairly consistently given accommodation size (2-6 people) & property type (apartment or condominium), yet RMSE decreased by ca. 40%, to 28 dollar RMSE, when considering the most frequent neighborhood in the dataset: Khlong Toei.

Data, Cleaning & Feature Engineering

The raw dataset was downloaded from- & is available at AirBnB-s website, from here, & comprises over 19.7K observations & 74 Variables, summarizing all accommodations in the Bangkok area available for rental, as of 23rd-24th December, 2020.

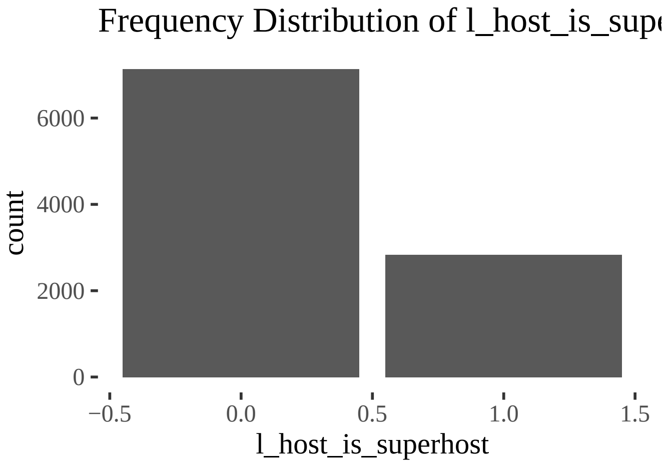
After keeping only apartments-, & condominiums hosting between 2-6 people, 9962 observations remained to clean for analysis. To manage this feature space, i.e. to keep what's important & drop what is not, while looking to maintain & intuition throughout the process, I grouped variables in 7 subjects (excluding ID variables): **Host** information, **Geo-Spatial-**, & **Property** information, **Sales-**, & **Availability**-related data, & **Reviews-**, & **Listings** data.

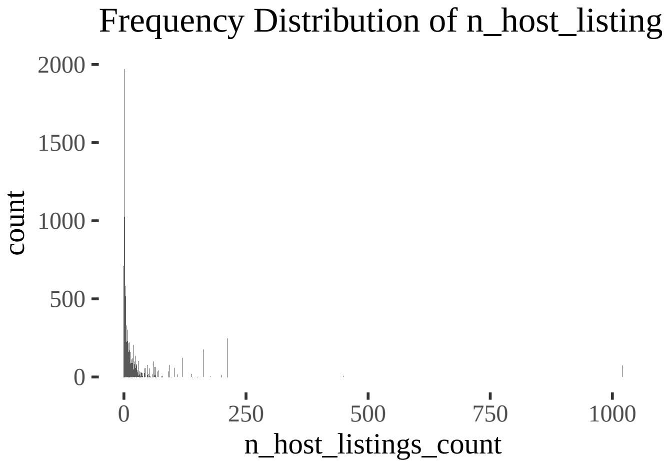
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

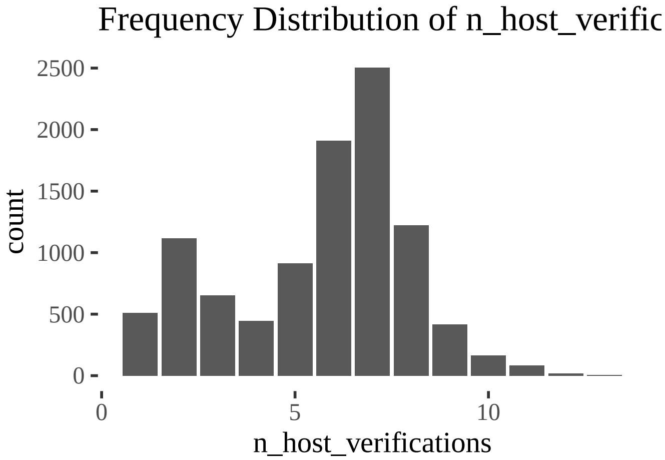
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

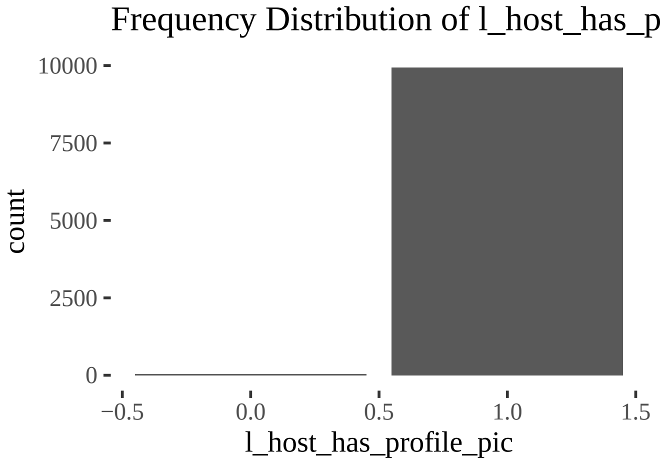
Including Plots

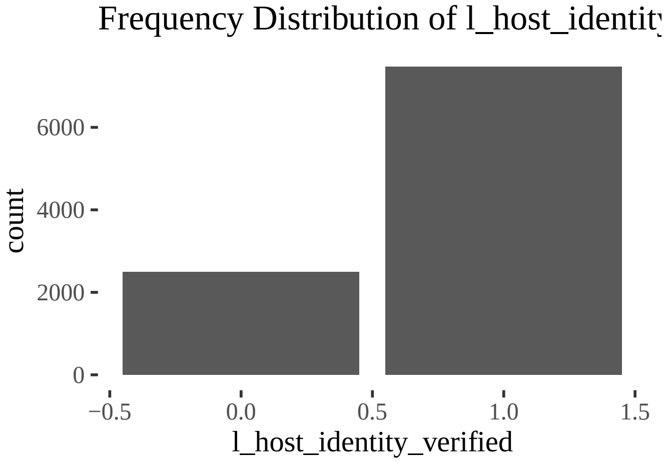
You can also embed plots, for example:

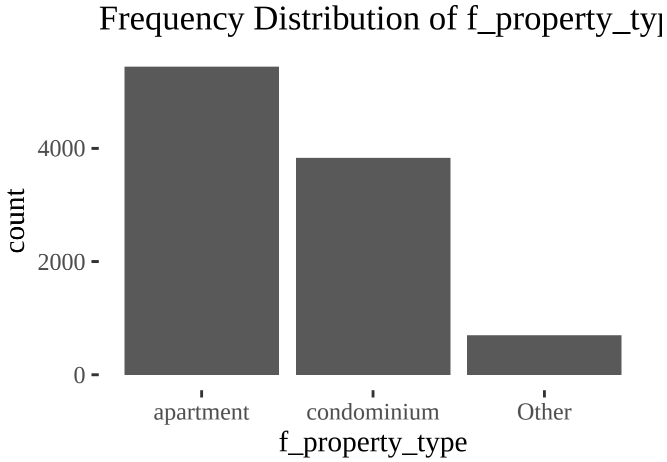

Frequency Distribution of l_host_is_supe

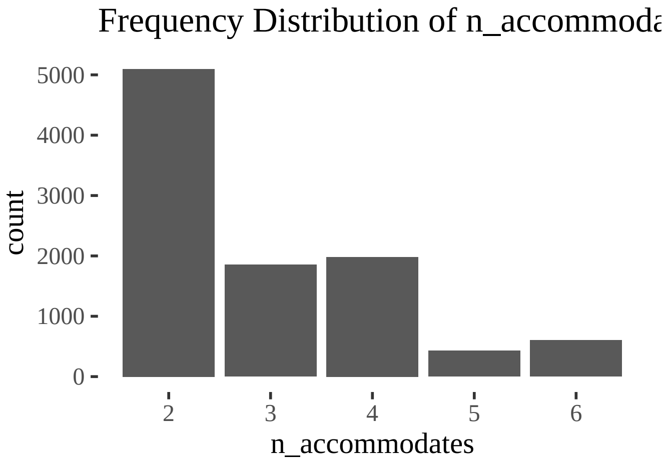

Frequency Distribution of n_host_listing

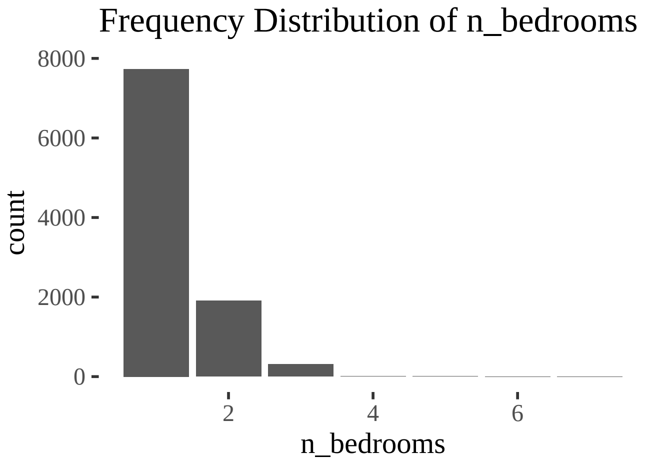

Frequency Distribution of n_host_verific

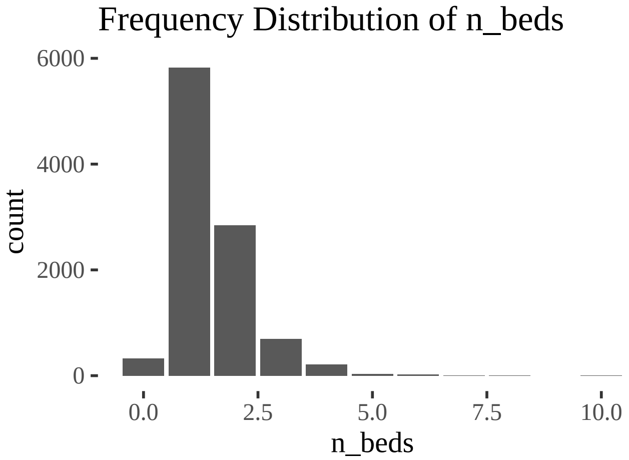

Frequency Distribution of l_host_has_p

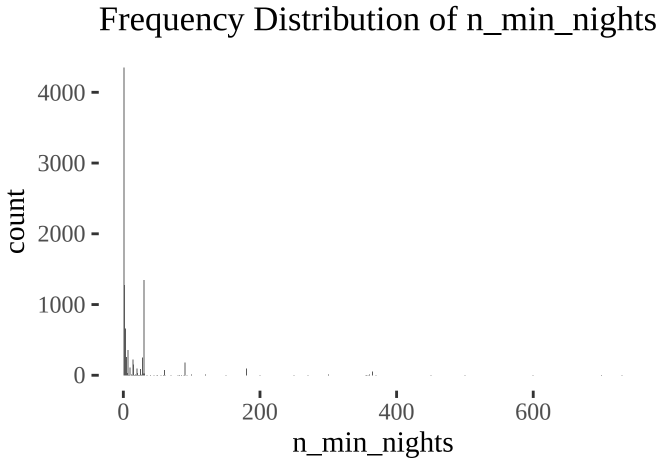

Frequency Distribution of l_host_identity

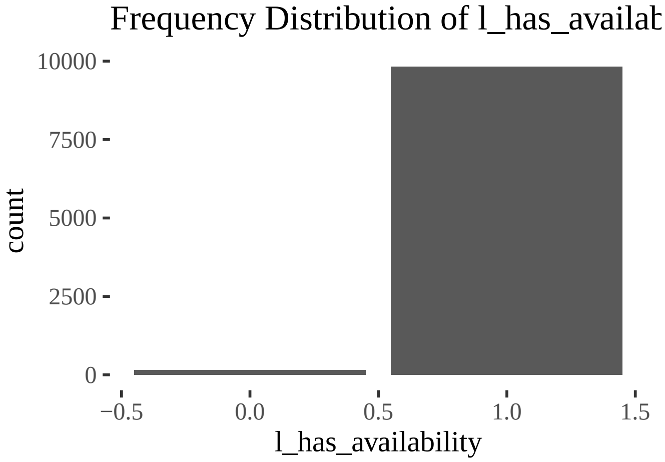

Frequency Distribution of f_property_ty

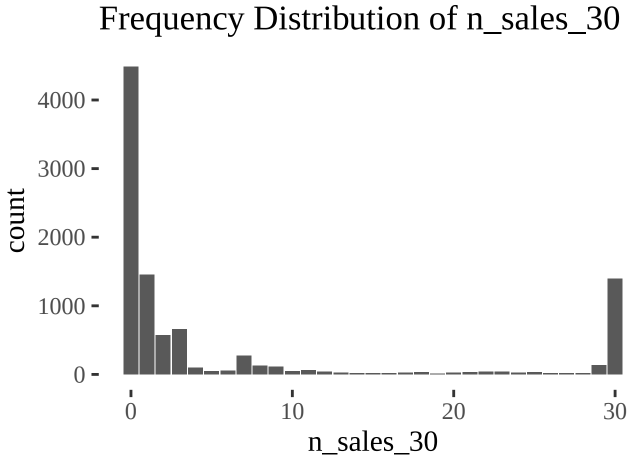

Frequency Distribution of n_accommoda

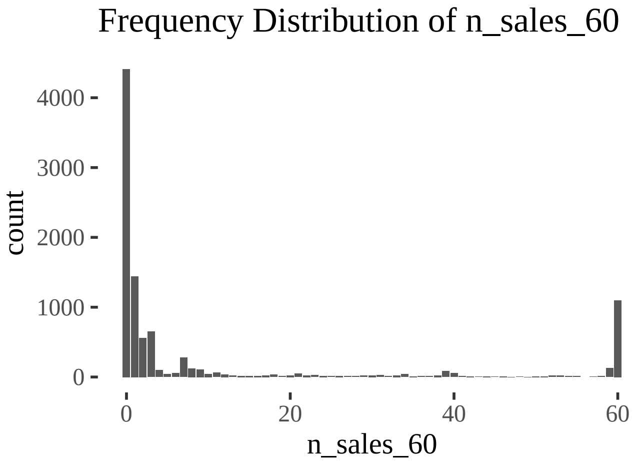

Frequency Distribution of n_bedrooms

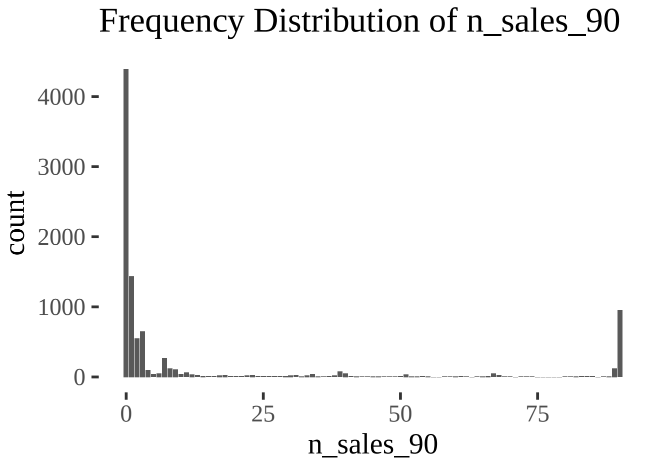

Frequency Distribution of n_beds

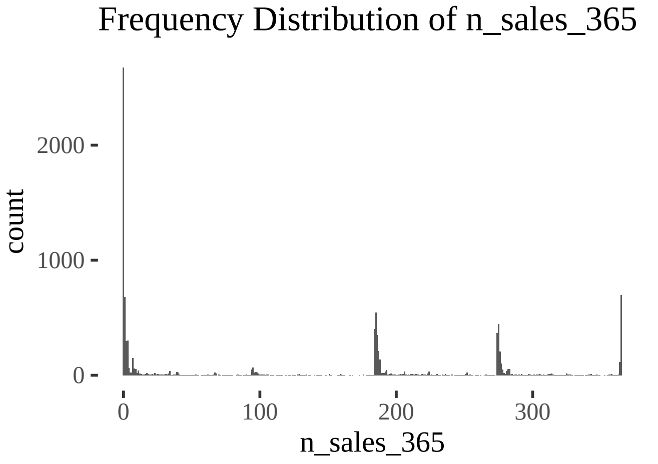

Frequency Distribution of n_min_nights

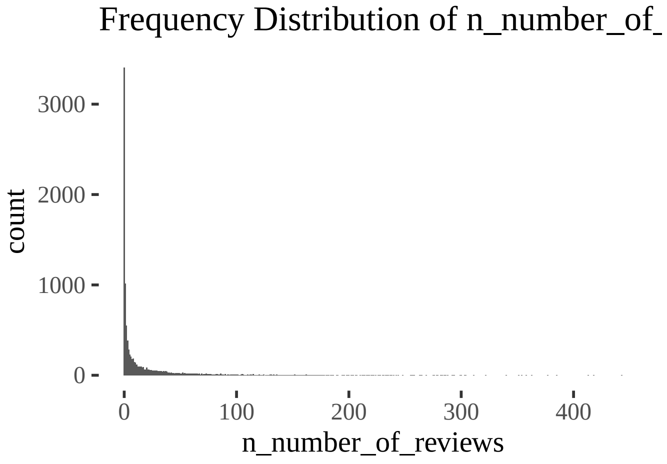

Frequency Distribution of l_has_availab

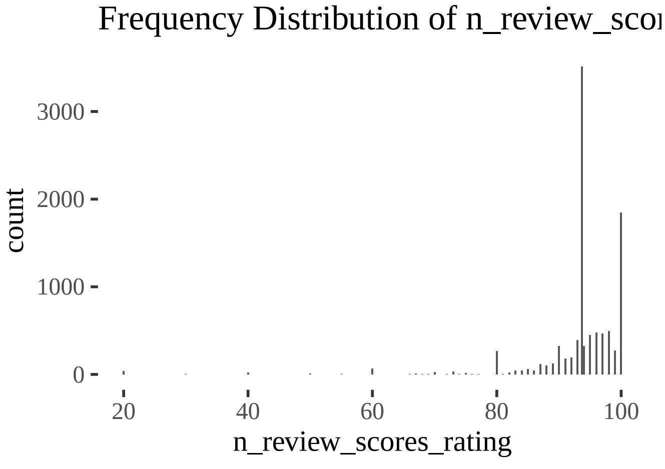

Frequency Distribution of n_sales_30

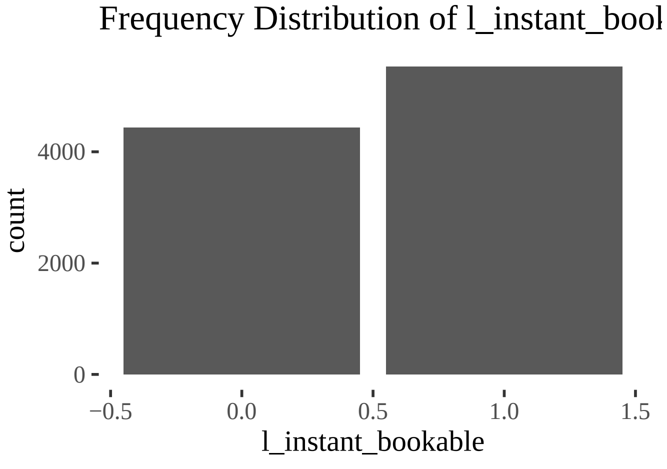

Frequency Distribution of n_sales_60
count
4000
3000
2000
1000
0
0
20
40
60
n_sales_60


Frequency Distribution of n_sales_90
count
4000
3000
2000
1000
0
0
25
50
75
n_sales_90

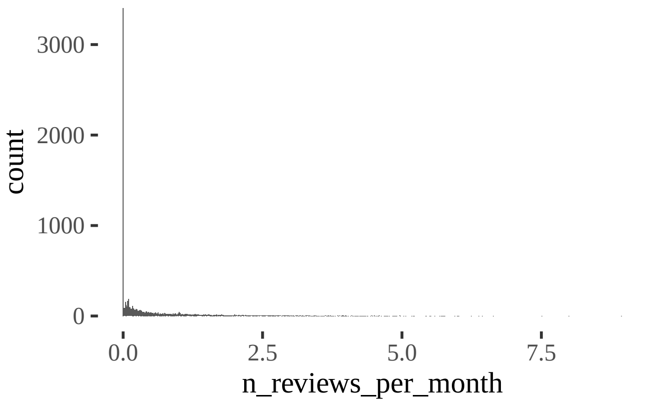

Frequency Distribution of n_sales_365
count
2000
1000
0
0
100
200
300
n_sales_365


Frequency Distribution of n_number_of_
count
3000
2000
1000
0
0
100
200
300
400
n_number_of_reviews

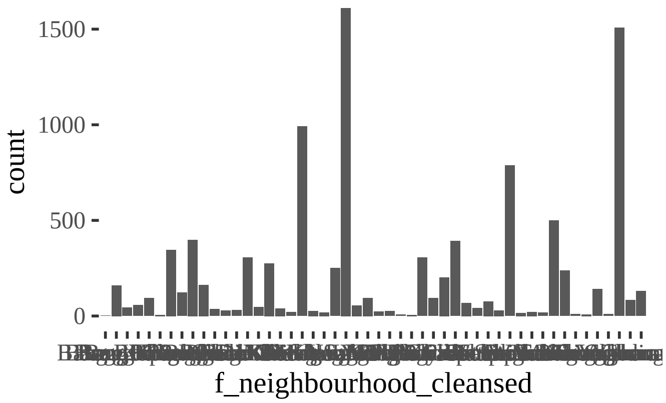

Frequency Distribution of n_review_scor
count
3000
2000
1000
0
20
40
60
80
100
n_review_scores_rating


Frequency Distribution of l_instant_bool
count
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
l_instant_bookable

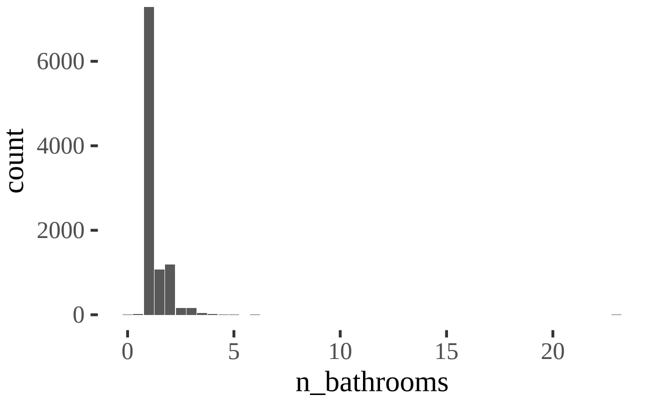
Frequency Distribution of n_reviews_per



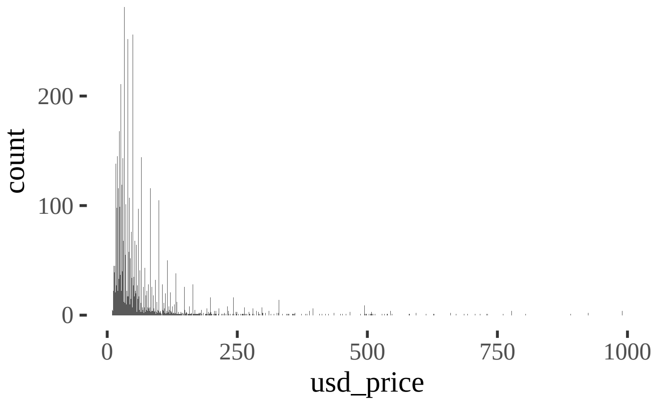
Frequency Distribution of f_neighbourho



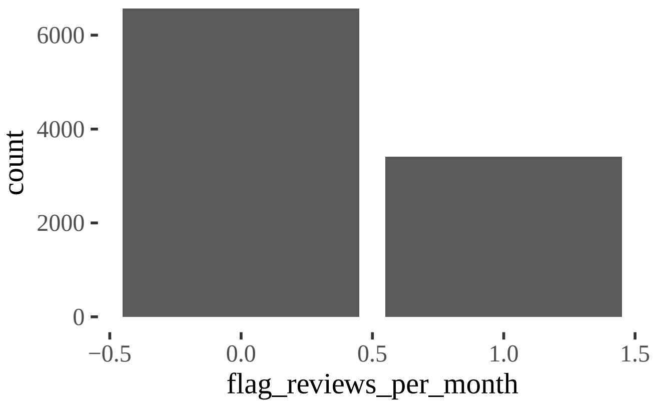
Frequency Distribution of n_bathrooms



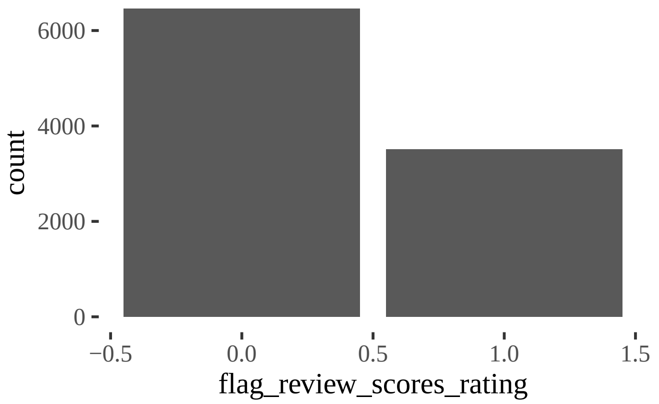
Frequency Distribution of usd_price

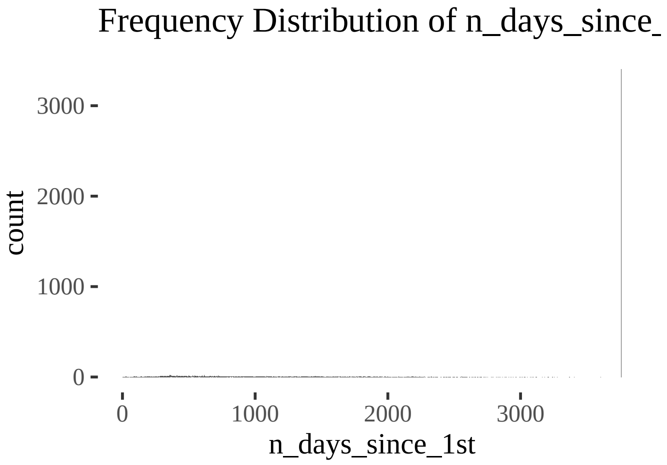


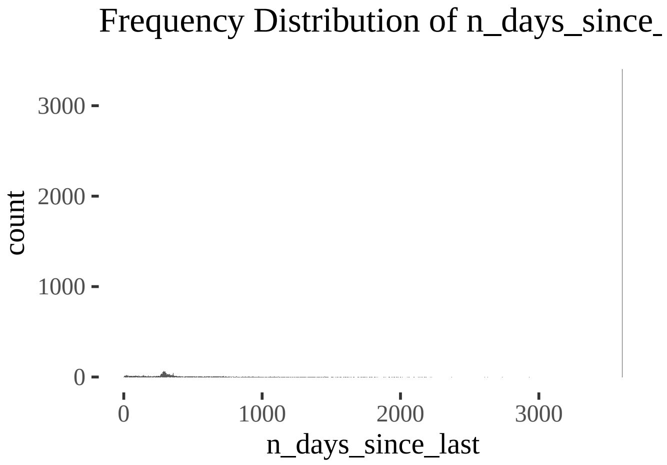
Frequency Distribution of flag_reviews_

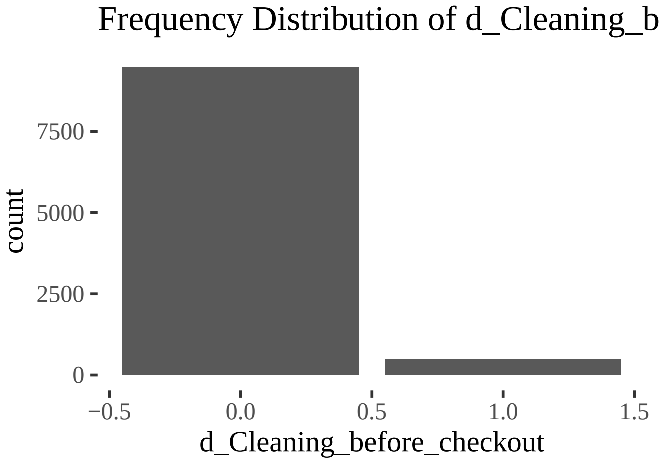


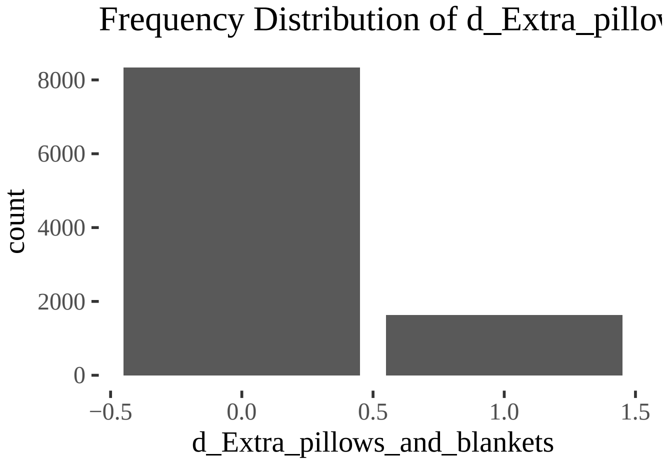
Frequency Distribution of flag_review_s

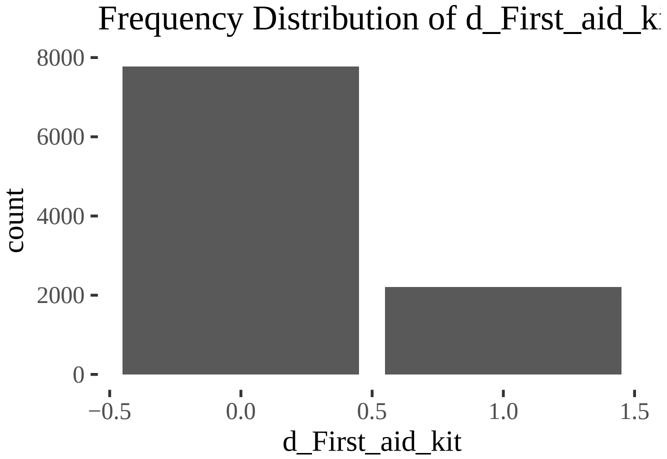


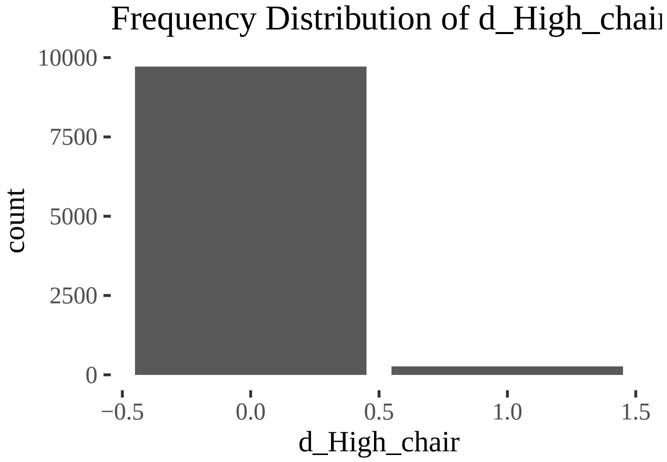

Frequency Distribution of n_days_since_
count
3000
2000
1000
0
0
1000
2000
3000
n_days_since_1st

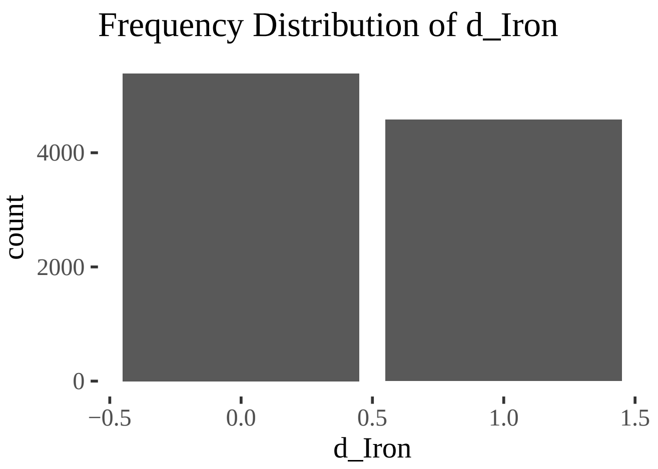

Frequency Distribution of n_days_since_
count
3000
2000
1000
0
0
1000
2000
3000
n_days_since_last

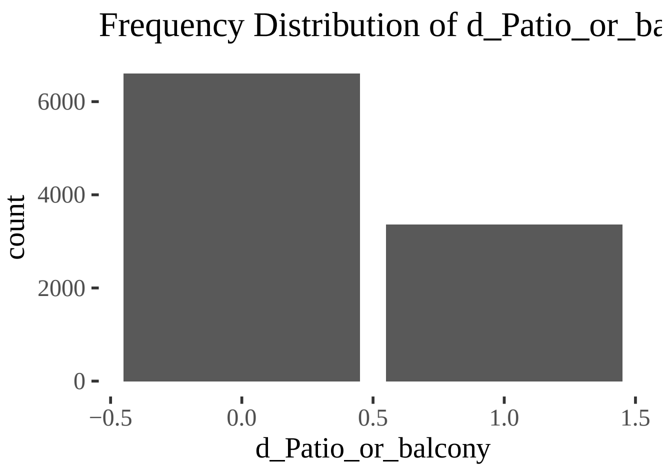

Frequency Distribution of d_Cleaning_b
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Cleaning_before_checkout

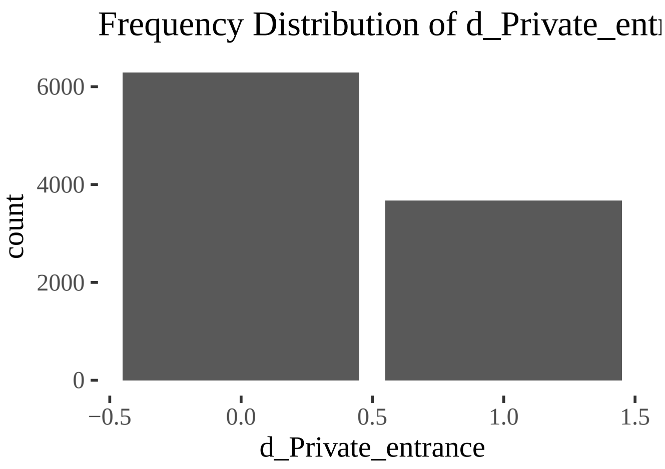

Frequency Distribution of d_Extra_pillov
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Extra_pillows_and_blankets

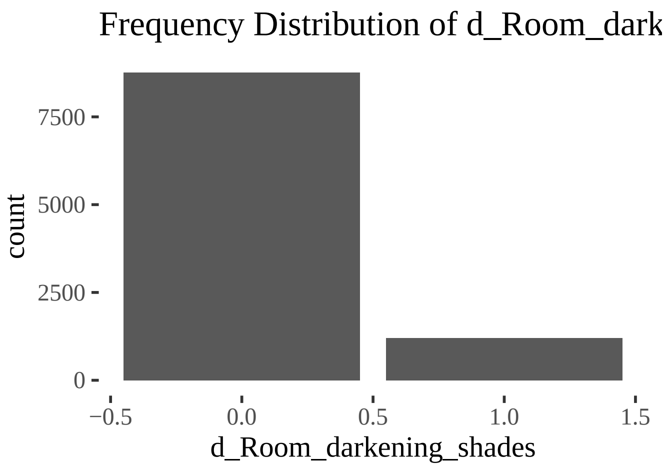

Frequency Distribution of d_First_aid_ki
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_First_aid_kit

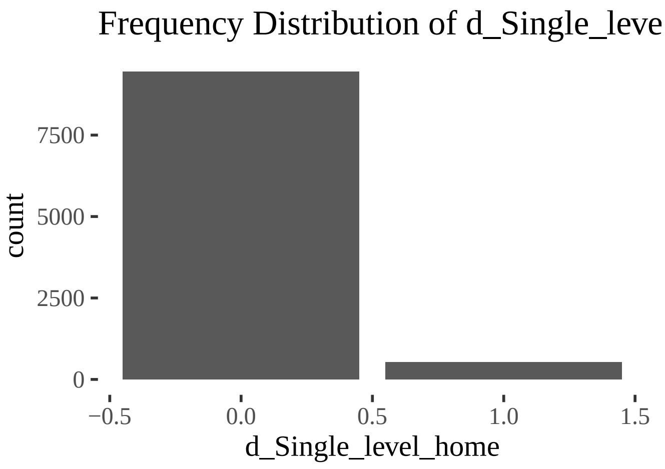

Frequency Distribution of d_High_chai
count
10000
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_High_chair

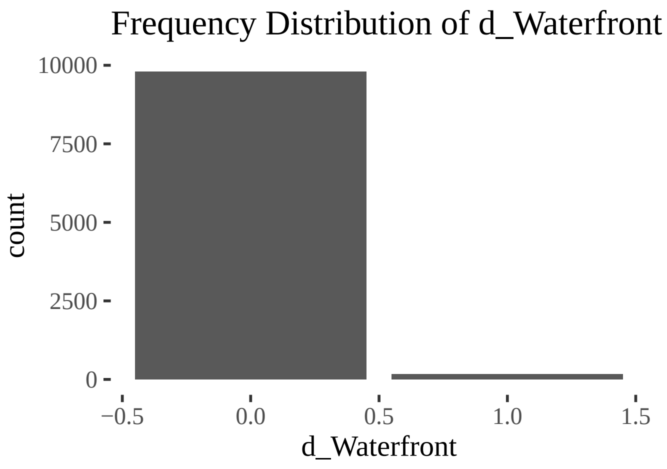

Frequency Distribution of d_Iron

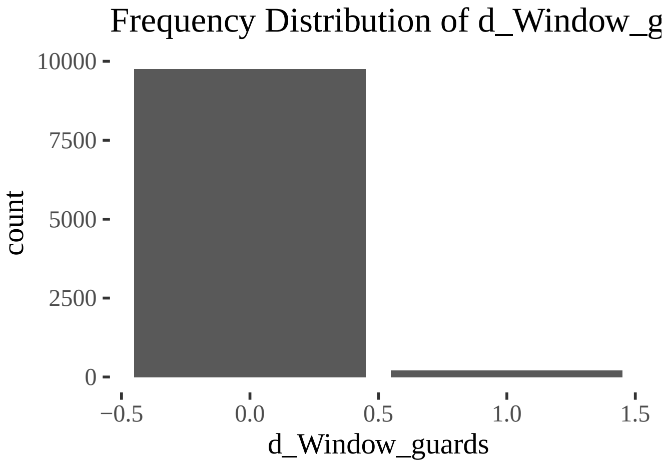

Frequency Distribution of d_Patio_or_ba

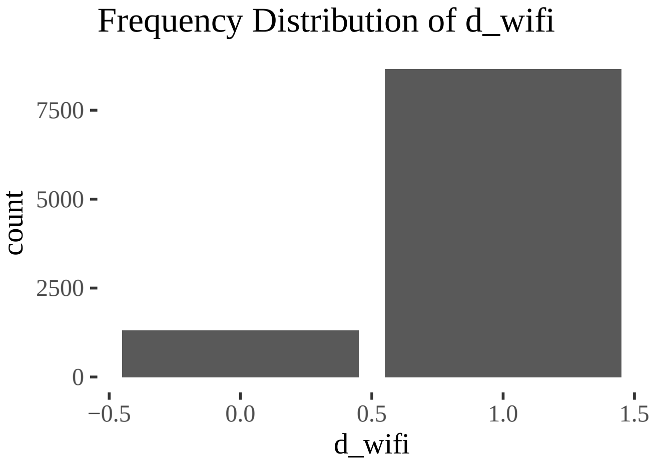

Frequency Distribution of d_Private_ent

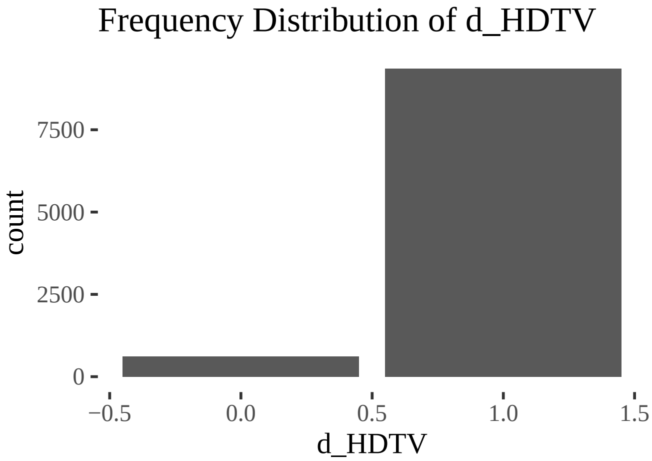

Frequency Distribution of d_Room_dark

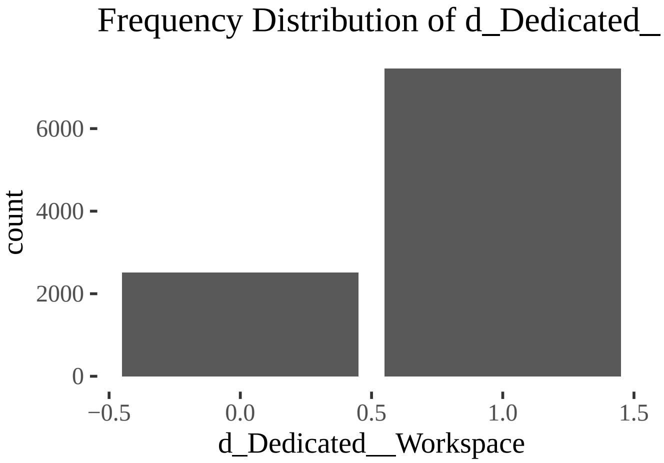

Frequency Distribution of d_Single_leve

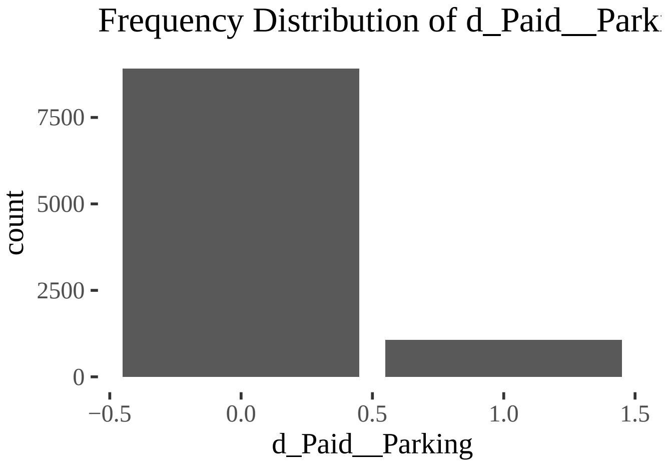

Frequency Distribution of d_Waterfront

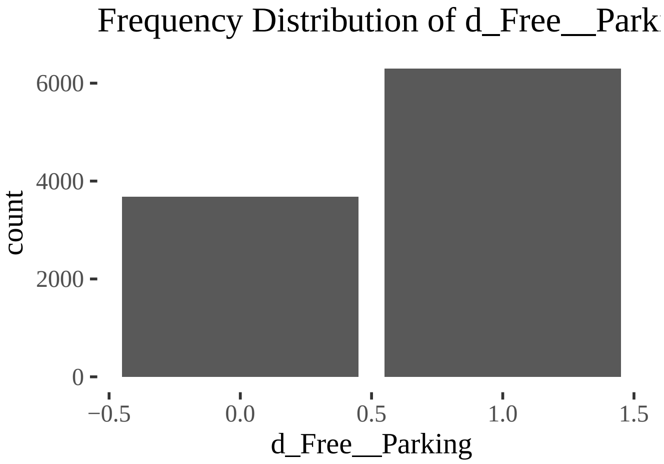

Frequency Distribution of d_Window_g
count
10000
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Window_guards

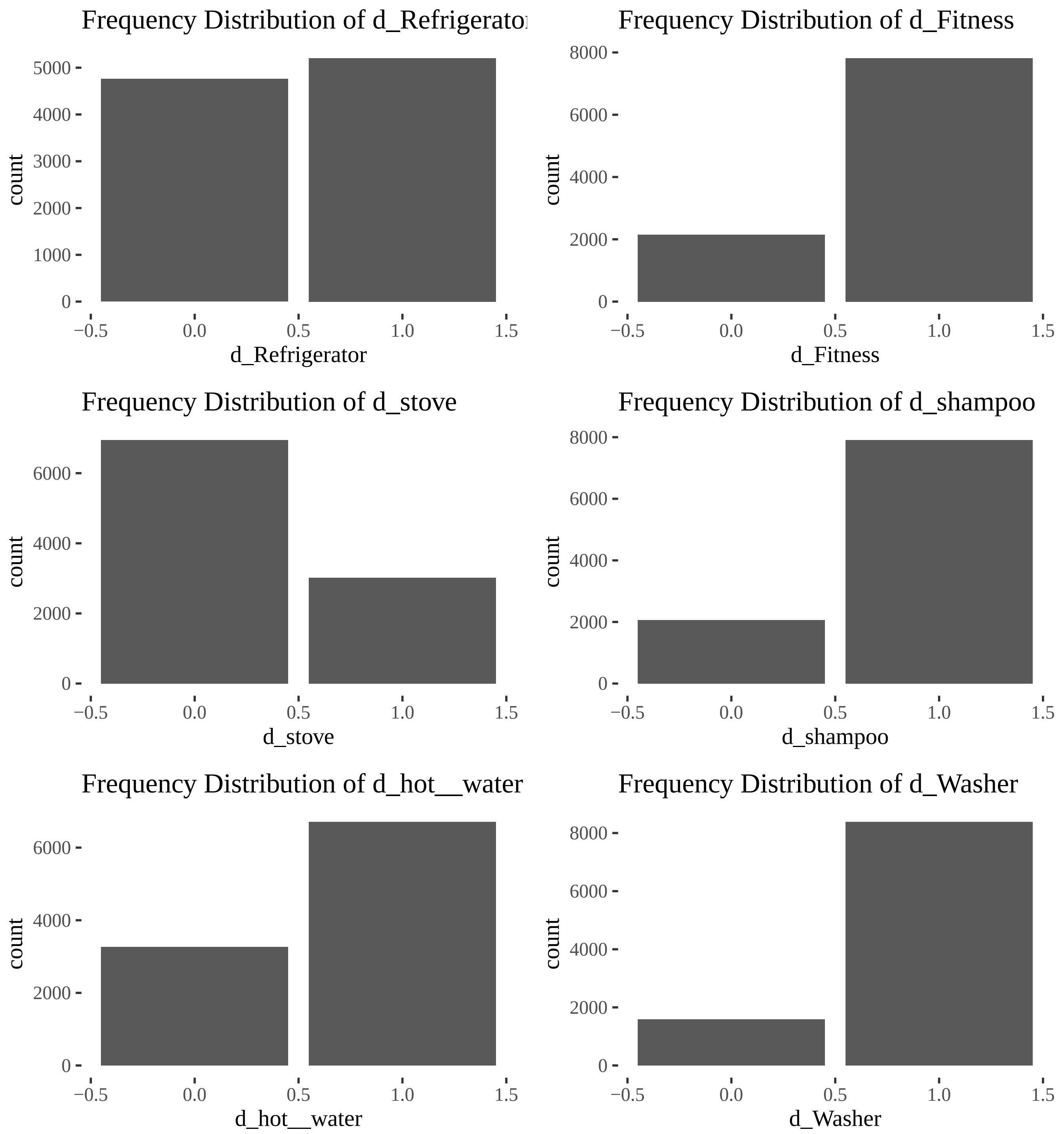

Frequency Distribution of d_wifi
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_wifi


Frequency Distribution of d_HDTV
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_HDTV

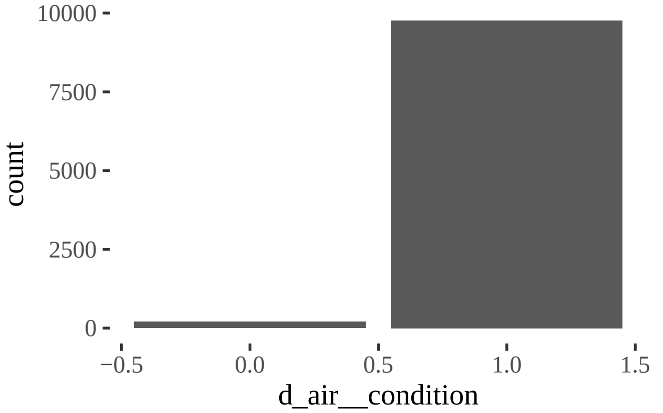

Frequency Distribution of d_Dedicated_
count
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Dedicated__Workspace


Frequency Distribution of d_Paid__Park
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Paid__Parking

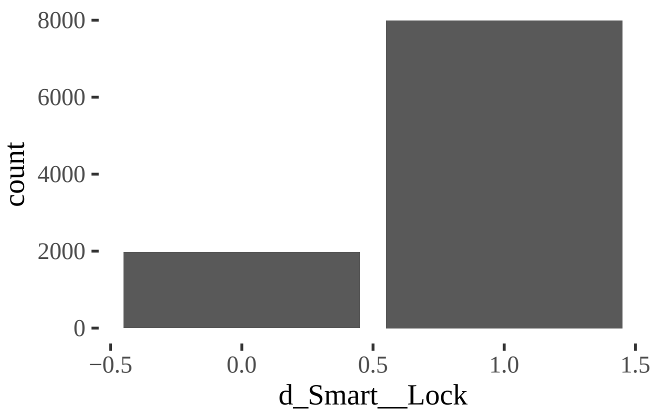

Frequency Distribution of d_Free__Park
count
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Free__Parking


Frequency Distribution of d_Refrigerator
count
5000
4000
3000
2000
1000
0
−0.5
0.0
0.5
1.0
1.5
d_Refrigerator
Frequency Distribution of d_Fitness
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Fitness
Frequency Distribution of d_stove
count
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_stove
Frequency Distribution of d_shampoo
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_shampoo
Frequency Distribution of d_hot__water
count
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_hot__water
Frequency Distribution of d_Washer
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Washer

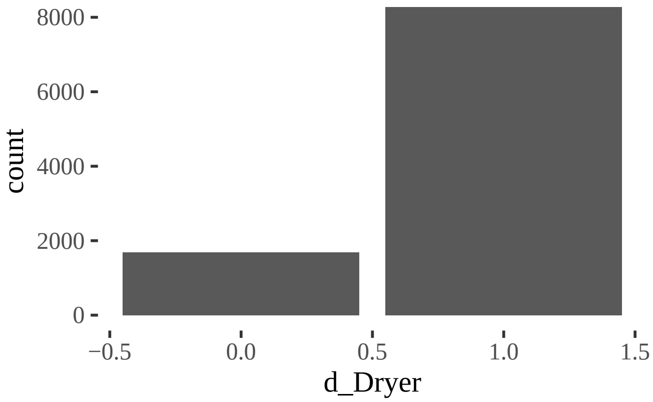
Frequency Distribution of d_air__condi



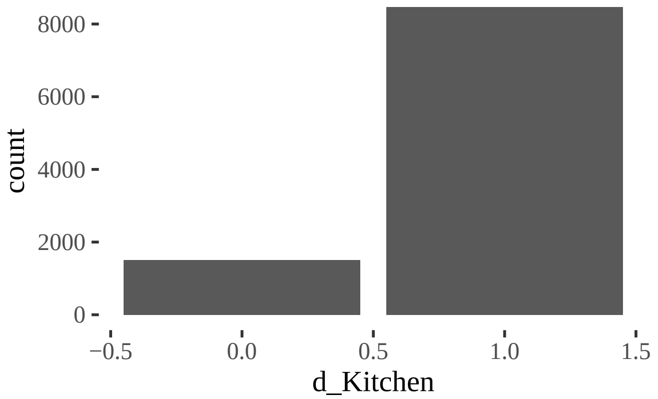
Frequency Distribution of d_Smart__Lo



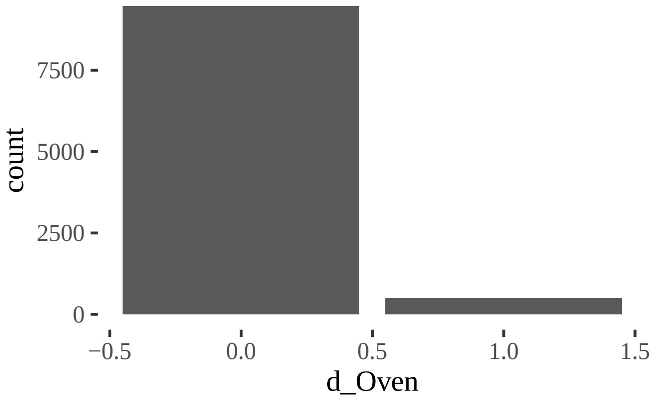
Frequency Distribution of d_Dryer



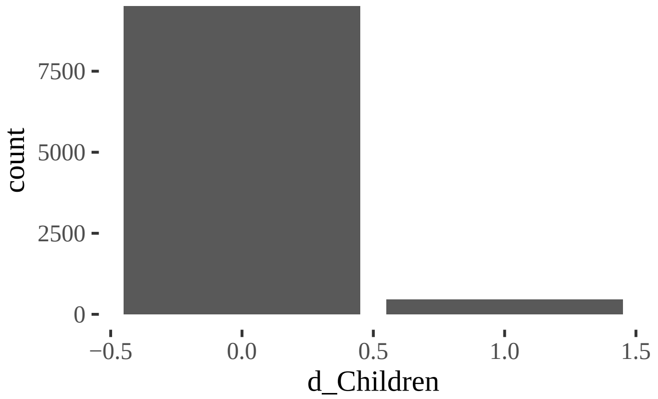
Frequency Distribution of d_Kitchen

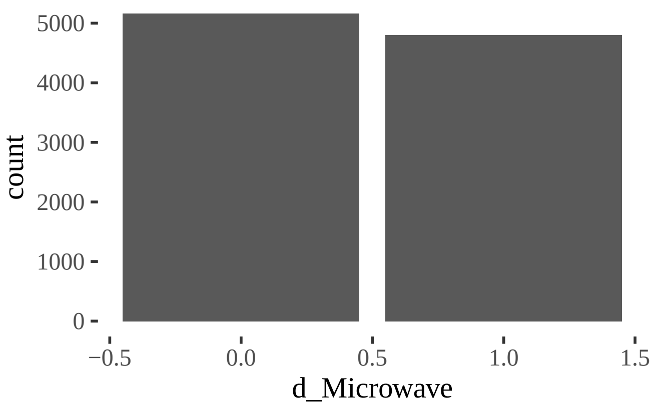


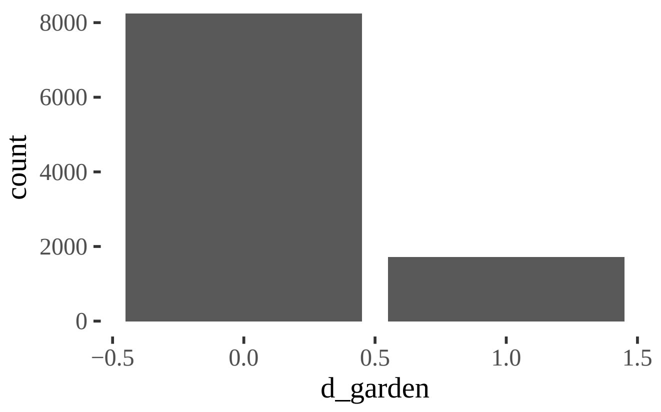
Frequency Distribution of d_Oven

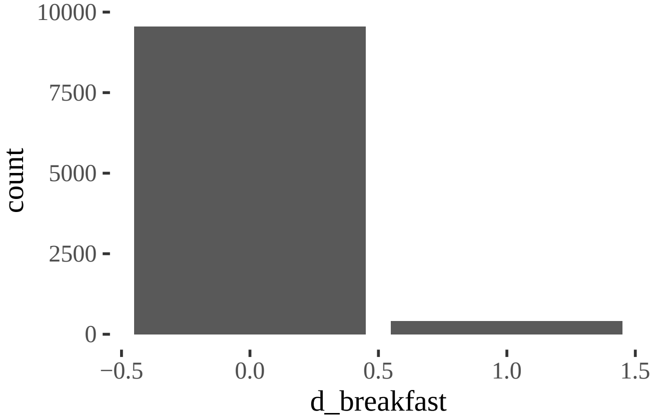


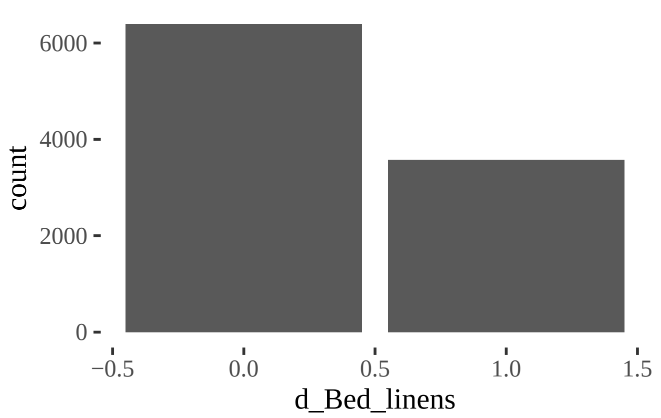
Frequency Distribution of d_Children

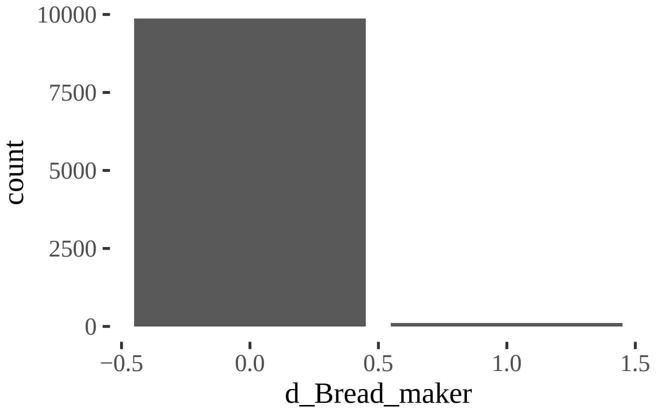


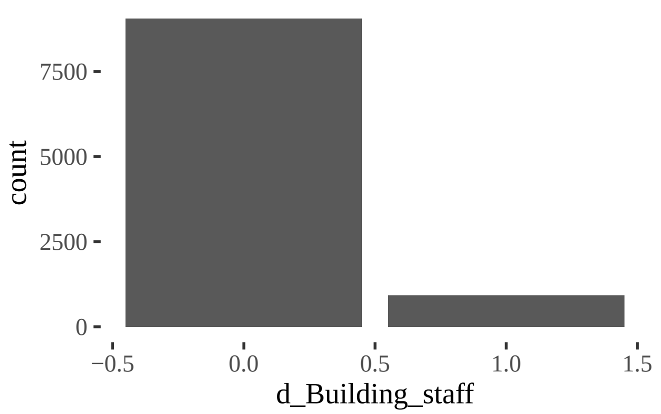

Frequency Distribution of d_Microwave


Frequency Distribution of d_garden

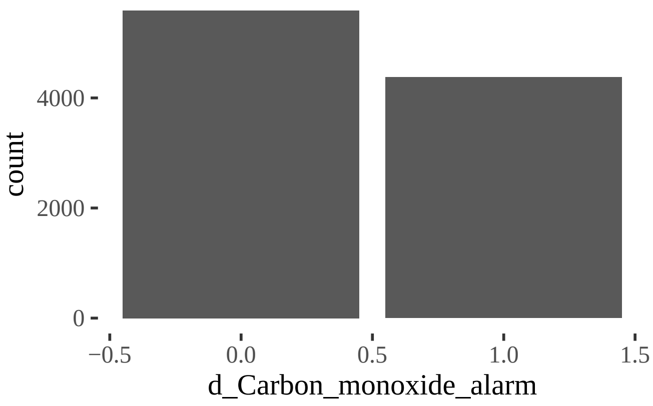

Frequency Distribution of d_breakfast


Frequency Distribution of d_Bed_linens

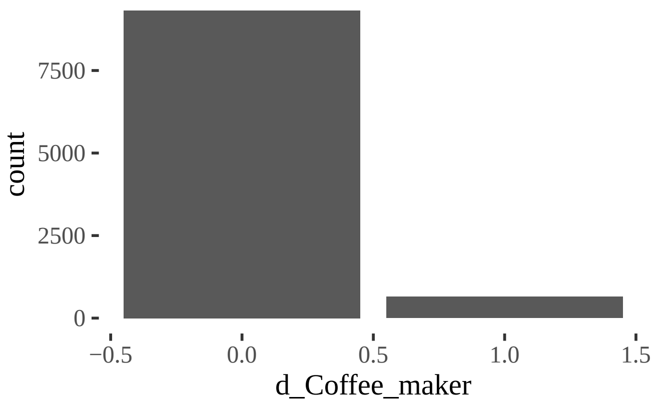

Frequency Distribution of d_Bread_mal


Frequency Distribution of d_Building_st

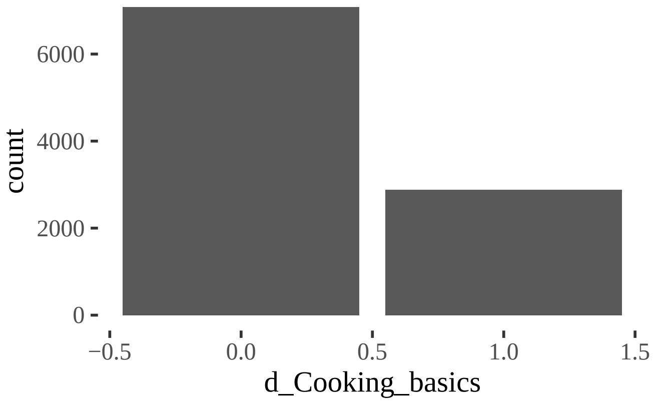
Frequency Distribution of d_Carbon_mo



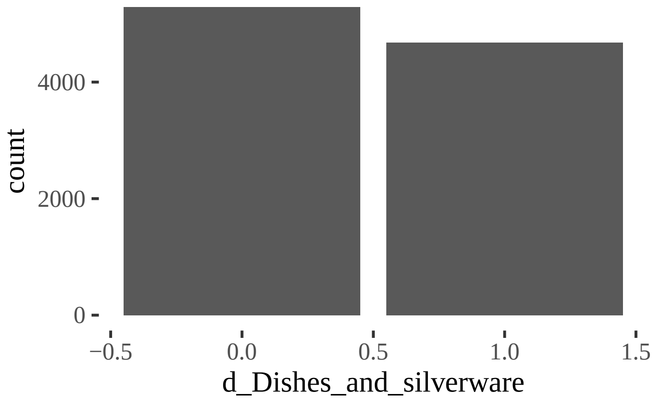
Frequency Distribution of d_Coffee_mal



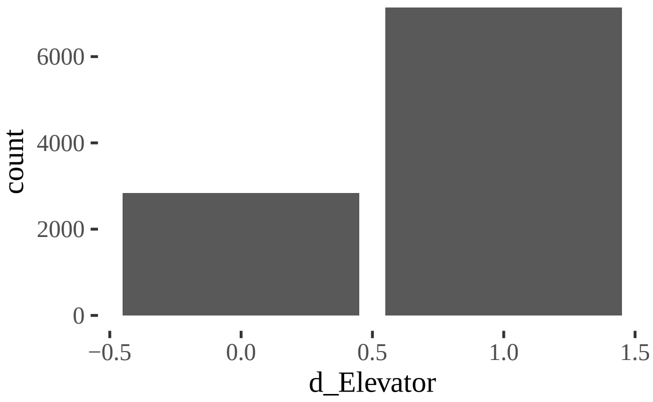
Frequency Distribution of d_Cooking_ba



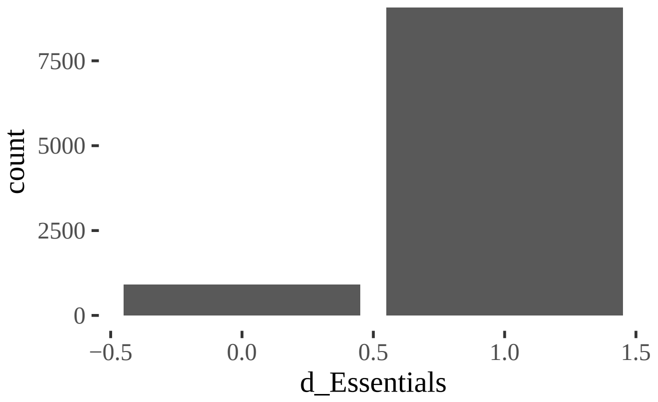
Frequency Distribution of d_Dishes_and

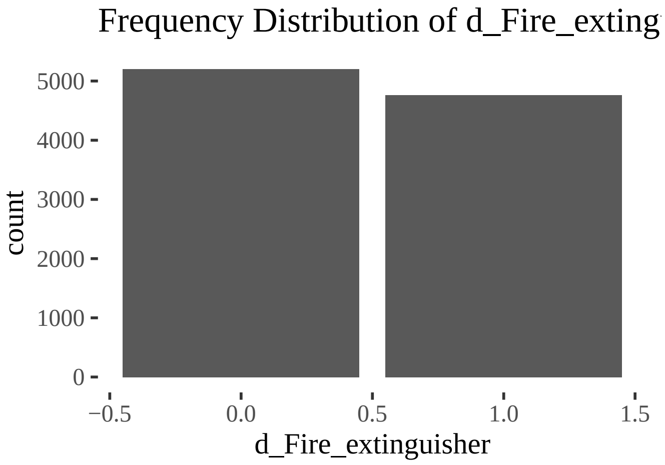


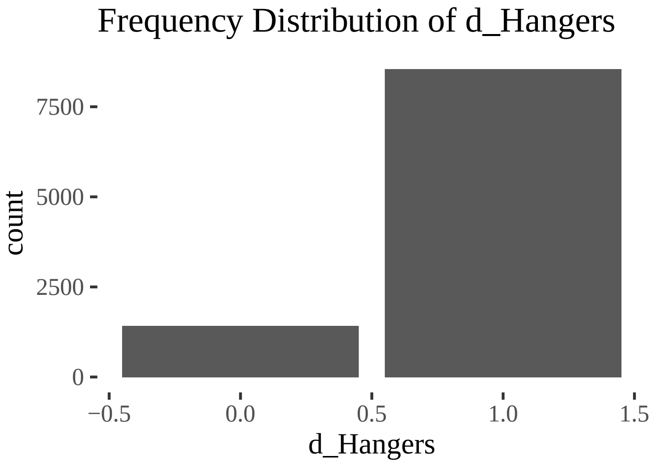
Frequency Distribution of d_Elevator

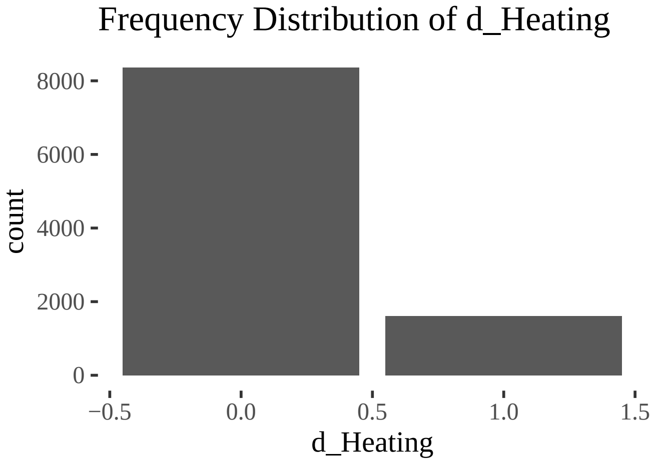


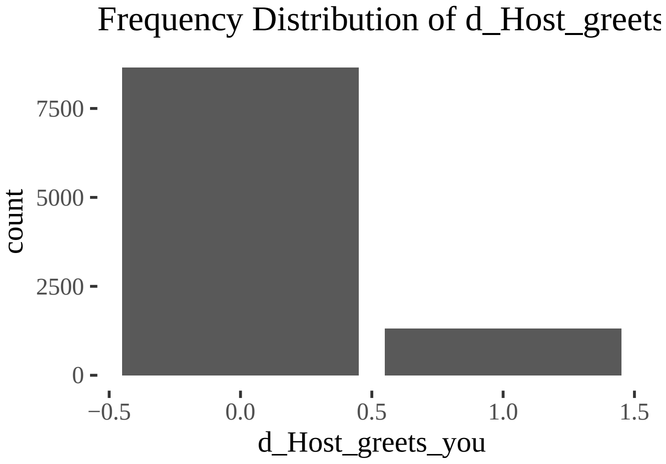
Frequency Distribution of d_Essentials

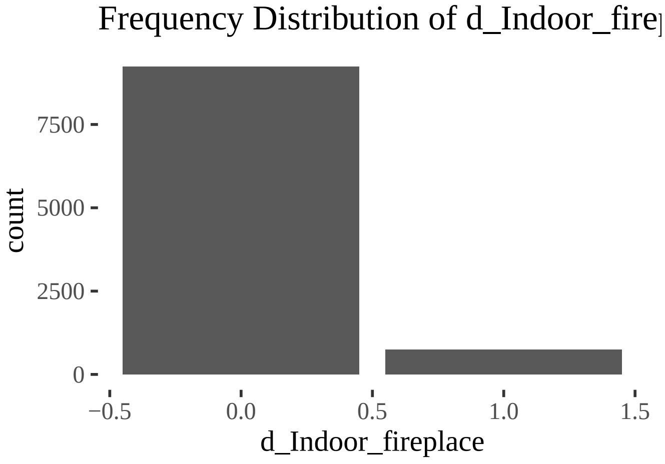


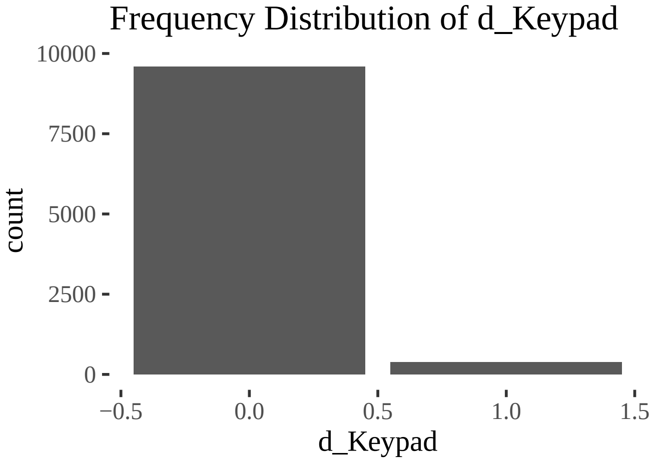

Frequency Distribution of d_Fire_exting
count
5000
4000
3000
2000
1000
0
−0.5
0.0
0.5
1.0
1.5
d_Fire_extinguisher

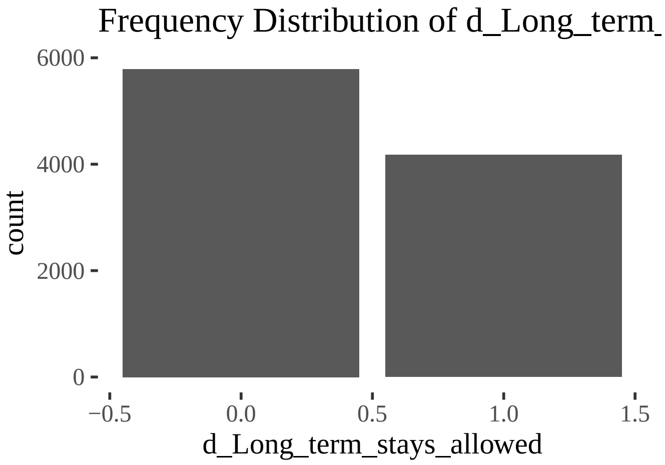

Frequency Distribution of d_Hangers
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Hangers

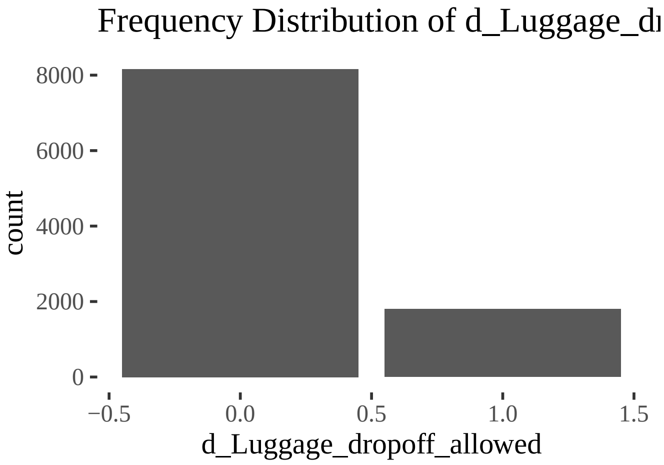

Frequency Distribution of d_Heating
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Heating

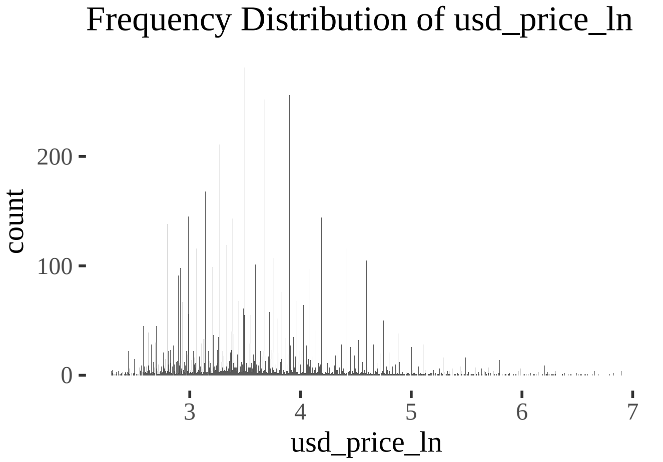

Frequency Distribution of d_Host_greets
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Host_greets_you

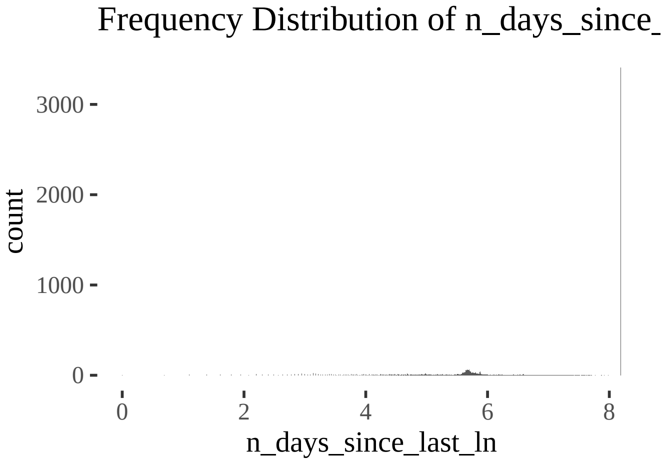

Frequency Distribution of d_Indoor_fire
count
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Indoor_fireplace

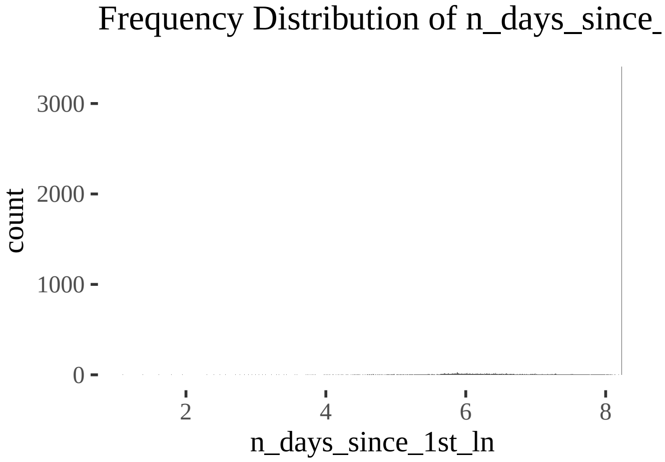

Frequency Distribution of d_Keypad
count
10000
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
d_Keypad

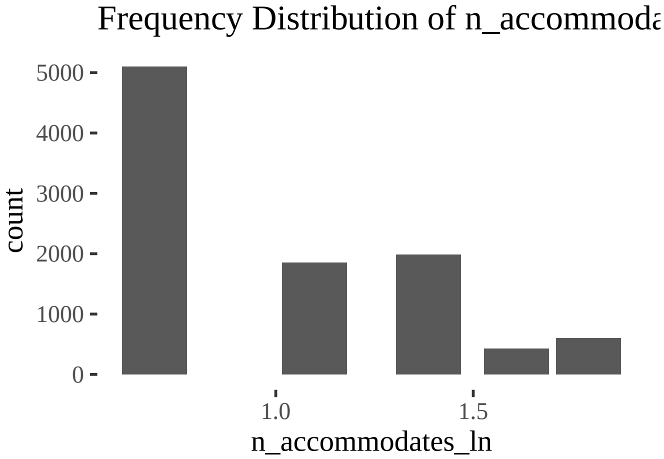

Frequency Distribution of d_Long_term_
count
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Long_term_stays_allowed

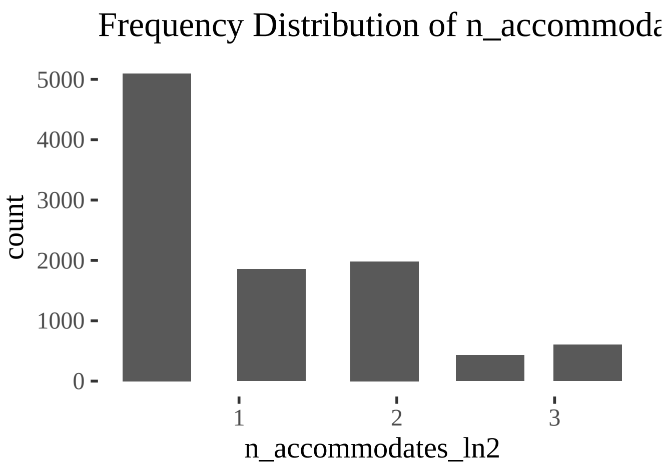

Frequency Distribution of d_Luggage_d
count
8000
6000
4000
2000
0
−0.5
0.0
0.5
1.0
1.5
d_Luggage_dropoff_allowed

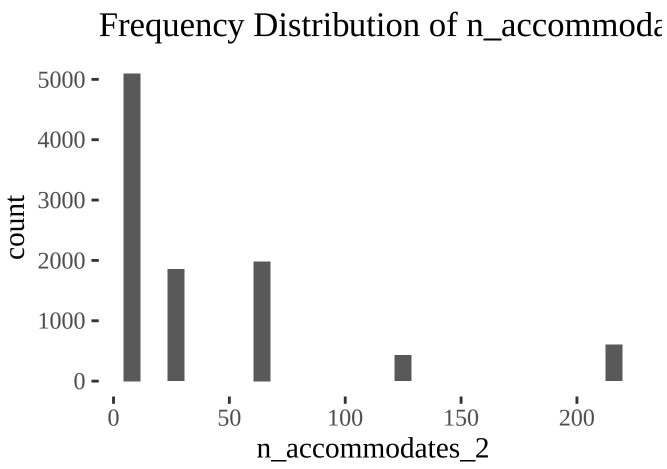

Frequency Distribution of usd_price_ln
count
200
100
0
3
4
5
6
7
usd_price_ln

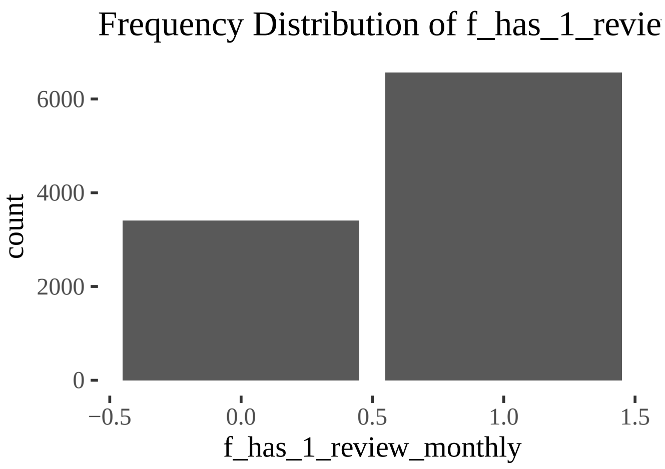

Frequency Distribution of n_days_since_
count
3000
2000
1000
0
0
2
4
6
8
n_days_since_last_ln

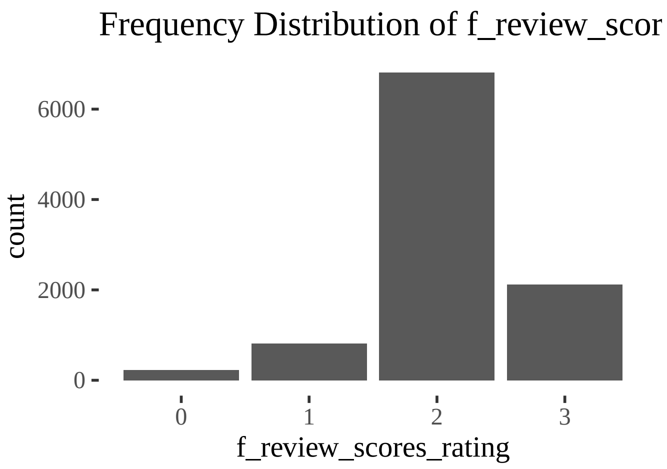

Frequency Distribution of n_days_since_
count
3000
2000
1000
0
2
4
6
8
n_days_since_1st_ln

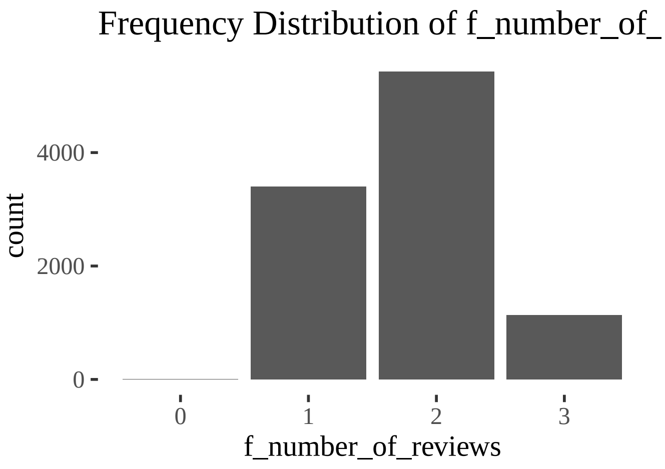

Frequency Distribution of n_accommoda
count
5000
4000
3000
2000
1000
0
1.0
1.5
n_accommodates_ln

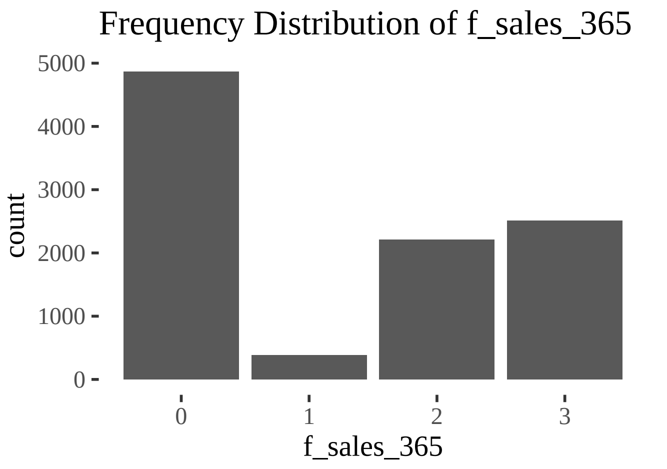

Frequency Distribution of n_accommoda


Frequency Distribution of n_accommoda

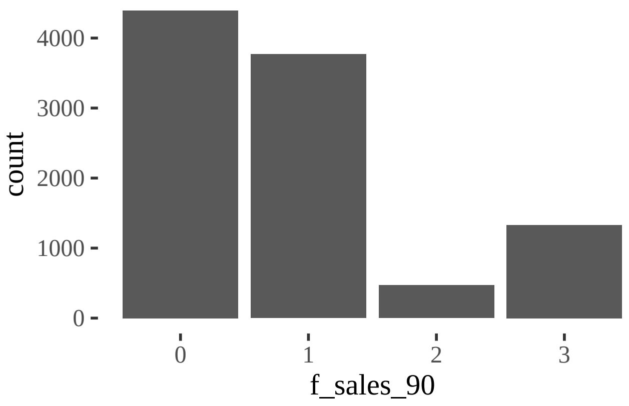

Frequency Distribution of f_has_1_revie


Frequency Distribution of f_review_scor

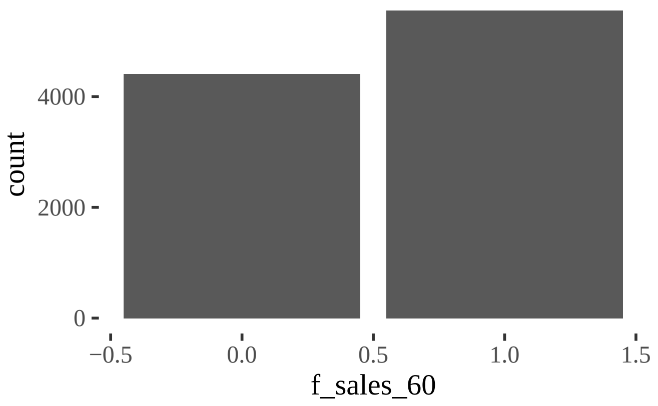

Frequency Distribution of f_number_of_


Frequency Distribution of f_sales_365

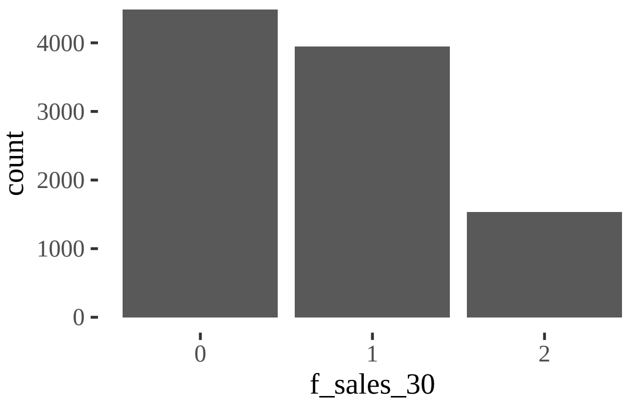
Frequency Distribution of f_sales_90



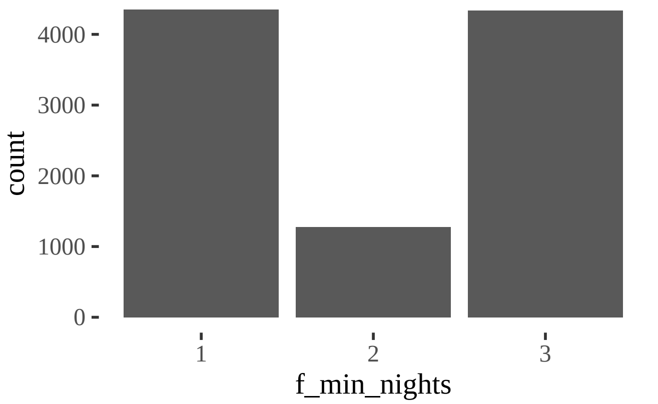
Frequency Distribution of f_sales_60



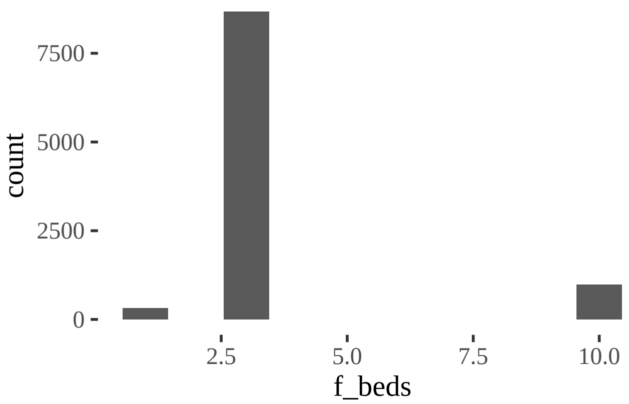
Frequency Distribution of f_sales_30



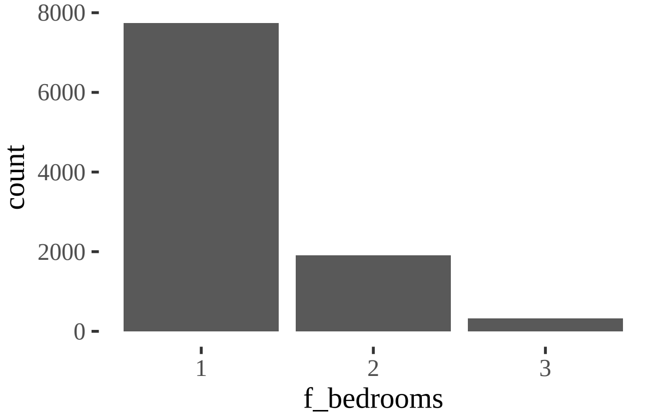
Frequency Distribution of f_min_nights

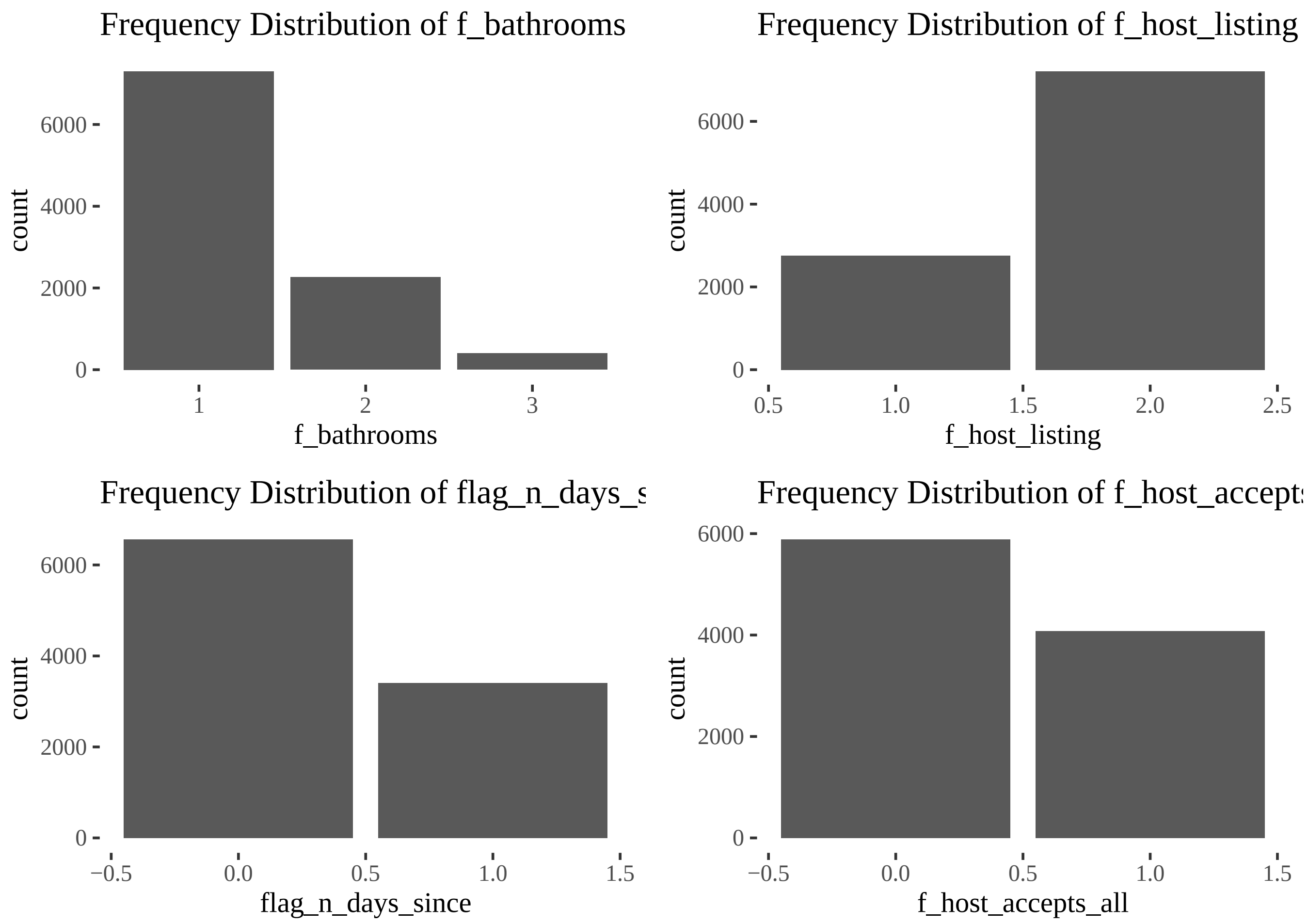


Frequency Distribution of f_beds



Frequency Distribution of f_bedrooms





```
NULL
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.