DSC520 Week10 Exercise 10.3

Anjani Bonda

February 19th 2022

Project: Impact of AirBnB on rental home prices in Columbus, OH

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v readr   2.1.2

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x plyr::arrange()   masks dplyr::arrange()
## x purrr::compact()  masks plyr::compact()
## x plyr::count()     masks dplyr::count()
## x plyr::failwith()  masks dplyr::failwith()
## x dplyr::filter()   masks stats::filter()
## x plyr::id()        masks dplyr::id()
## x dplyr::lag()      masks stats::lag()
## x plyr::mutate()    masks dplyr::mutate()
## x plyr::rename()    masks dplyr::rename()
## x plyr::summarise() masks dplyr::summarise()
## x plyr::summarize() masks dplyr::summarize()
```

```
library(ggplot2)

setwd("/Users/anjanibonda/DSC520/dsc520")

# Above data set contains information across US cities
# Filtering the data based on Columbus city
library(readr)
airbnb_columbus_df <- readr::read_csv('data/airbnb-listings.csv')
```

```
## Rows: 1562 Columns: 74

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (24): listing_url, name, description, neighborhood_overview, picture_ur...
## dbl  (37): id, scrape_id, host_id, host_listings_count, host_total_listings_...
## lgl   (8): host_is_superhost, host_has_profile_pic, host_identity_verified, ...
## date  (5): last_scraped, host_since, calendar_last_scraped, first_review, la...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(airbnb_columbus_df)
```

```
## spec_tbl_df [1,562 x 74] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id                                          : num [1:1562] 90676 543140 591101 923248 927867 ...
##  $ listing_url                                 : chr [1:1562] "https://www.airbnb.com/rooms/90676" "
##  $ scrape_id                                   : num [1:1562] 2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.0
##  $ last_scraped                                : Date[1:1562], format: "2021-12-24" "2021-12-24" ...
##  $ name                                        : chr [1:1562] "Short North - Italianate Cottage" "Pri
##  $ description                                 : chr [1:1562] "Just steps from High Street and all th
##  $ neighborhood_overview                       : chr [1:1562] "The Short North Italianate Cottage is
##  $ picture_url                                 : chr [1:1562] "https://a0.muscache.com/pictures/950e4
##  $ host_id                                     : num [1:1562] 483306 2350409 2889677 4965048 4965048
```

```
##  $ host_url                                : chr [1:1562] "https://www.airbnb.com/users/show/4833
##  $ host_name                               : chr [1:1562] "Audra & Lacey" "Edward" "Gail" "Mathew
##  $ host_since                              : Date[1:1562], format: "2011-04-04" "2012-05-11" ...
##  $ host_location                           : chr [1:1562] "Columbus, Ohio, United States" "Columb
##  $ host_about                              : chr [1:1562] "Active, young professionals who love t
##  $ host_response_time                      : chr [1:1562] "within an hour" "within an hour" "with
##  $ host_response_rate                      : chr [1:1562] "100%" "100%" "100%" "100%" ...
##  $ host_acceptance_rate                    : chr [1:1562] "97%" "98%" "96%" "96%" ...
##  $ host_is_superhost                       : logi [1:1562] TRUE TRUE TRUE FALSE FALSE FALSE ...
##  $ host_thumbnail_url                      : chr [1:1562] "https://a0.muscache.com/im/users/48330
##  $ host_picture_url                        : chr [1:1562] "https://a0.muscache.com/im/users/48330
##  $ host_neighbourhood                      : chr [1:1562] NA NA NA "Hilo" ...
##  $ host_listings_count                     : num [1:1562] 2 3 1 16 16 2 2 2 2 16 ...
##  $ host_total_listings_count               : num [1:1562] 2 3 1 16 16 2 2 2 2 16 ...
##  $ host_verifications                      : chr [1:1562] "['email', 'phone', 'reviews', 'kba',
##  $ host_has_profile_pic                    : logi [1:1562] TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ host_identity_verified                  : logi [1:1562] TRUE FALSE FALSE TRUE TRUE TRUE ...
##  $ neighbourhood                           : chr [1:1562] "Columbus, Ohio, United States" "Columb
##  $ neighbourhood_cleansed                  : chr [1:1562] "Near North/University" "Near North/Uni
##  $ neighbourhood_group_cleansed            : logi [1:1562] NA NA NA NA NA NA ...
##  $ latitude                                : num [1:1562] 40 40 40 40 40 ...
##  $ longitude                               : num [1:1562] -83 -83 -83 -83 -83 ...
##  $ property_type                           : chr [1:1562] "Entire residential home" "Private roo
##  $ room_type                               : chr [1:1562] "Entire home/apt" "Private room" "Priva
##  $ accommodates                            : num [1:1562] 6 1 2 1 2 6 2 4 2 1 ...
##  $ bathrooms                               : logi [1:1562] NA NA NA NA NA NA ...
##  $ bathrooms_text                          : chr [1:1562] "2 baths" "1 shared bath" "1 private ba
##  $ bedrooms                                : num [1:1562] 3 1 1 1 1 3 1 2 1 1 ...
##  $ beds                                    : num [1:1562] 3 1 1 5 1 3 1 2 1 3 ...
##  $ amenities                               : chr [1:1562] "[\"Heating\", \"TV\", \"Dishes and sil
##  $ price                                   : chr [1:1562] "$122.00" "$25.00" "$100.00" "$35.00"
##  $ minimum_nights                          : num [1:1562] 1 3 2 1 1 1 1 3 1 1 ...
##  $ maximum_nights                          : num [1:1562] 365 1125 30 365 365 ...
##  $ minimum_minimum_nights                  : num [1:1562] 1 3 2 1 1 1 1 3 1 1 ...
##  $ maximum_minimum_nights                  : num [1:1562] 1 3 2 1 1 1 1 3 1 1 ...
##  $ minimum_maximum_nights                  : num [1:1562] 365 1125 1125 30 30 ...
##  $ maximum_maximum_nights                  : num [1:1562] 365 1125 1125 30 30 ...
##  $ minimum_nights_avg_ntm                  : num [1:1562] 1 3 2 1 1 1 1 3 1 1 ...
##  $ maximum_nights_avg_ntm                  : num [1:1562] 365 1125 1125 30 30 ...
##  $ calendar_updated                        : logi [1:1562] NA NA NA NA NA NA ...
##  $ has_availability                        : logi [1:1562] TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ availability_30                         : num [1:1562] 19 13 0 22 17 22 21 7 24 0 ...
##  $ availability_60                         : num [1:1562] 47 13 27 52 45 52 45 7 48 0 ...
##  $ availability_90                         : num [1:1562] 74 24 42 80 75 76 68 28 71 22 ...
##  $ availability_365                        : num [1:1562] 162 197 317 355 348 164 343 303 346 297
##  $ calendar_last_scraped                   : Date[1:1562], format: "2021-12-24" "2021-12-24" ...
##  $ number_of_reviews                       : num [1:1562] 523 113 251 254 60 11 185 153 147 45 .
##  $ number_of_reviews_ltm                   : num [1:1562] 110 14 26 31 15 11 19 17 16 0 ...
##  $ number_of_reviews_l30d                  : num [1:1562] 5 2 1 1 0 4 2 1 0 0 ...
##  $ first_review                            : Date[1:1562], format: "2011-10-11" "2012-07-31" ...
##  $ last_review                             : Date[1:1562], format: "2021-12-19" "2021-12-06" ...
##  $ review_scores_rating                    : num [1:1562] 4.82 4.67 4.92 4.76 4.67 4.73 4.96 4.98
##  $ review_scores_accuracy                  : num [1:1562] 4.85 4.72 4.92 4.81 4.83 4.82 4.98 4.98
##  $ review_scores_cleanliness               : num [1:1562] 4.84 4.3 4.92 4.79 4.63 5 4.98 4.99 4.9
```

```
##  $ review_scores_checkin                       : num [1:1562] 4.94 4.93 4.96 4.9 4.88 5 4.98 4.97 4.9
##  $ review_scores_communication                 : num [1:1562] 4.88 4.88 4.91 4.9 4.88 5 4.98 5 4.97 4
##  $ review_scores_location                      : num [1:1562] 4.93 4.75 4.88 4.71 4.68 4.73 4.96 4.98
##  $ review_scores_value                         : num [1:1562] 4.78 4.77 4.87 4.87 4.69 4.82 4.92 4.89
##  $ license                                     : chr [1:1562] "2019-1568" "2019-1344" "2019-1230" "20
##  $ instant_bookable                            : logi [1:1562] FALSE FALSE FALSE FALSE FALSE FALSE .
##  $ calculated_host_listings_count              : num [1:1562] 3 3 1 8 8 2 2 1 2 8 ...
##  $ calculated_host_listings_count_entire_homes : num [1:1562] 3 0 0 2 2 2 0 1 0 2 ...
##  $ calculated_host_listings_count_private_rooms: num [1:1562] 0 3 1 4 4 0 2 0 2 4 ...
##  $ calculated_host_listings_count_shared_rooms : num [1:1562] 0 0 0 2 2 0 0 0 0 2 ...
##  $ reviews_per_month                           : num [1:1562] 4.21 0.99 2.2 2.36 0.59 4.4 1.79 1.51 1
##  - attr(*, "spec")=
##   .. cols(
##   ..   id = col_double(),
##   ..   listing_url = col_character(),
##   ..   scrape_id = col_double(),
##   ..   last_scraped = col_date(format = ""),
##   ..   name = col_character(),
##   ..   description = col_character(),
##   ..   neighborhood_overview = col_character(),
##   ..   picture_url = col_character(),
##   ..   host_id = col_double(),
##   ..   host_url = col_character(),
##   ..   host_name = col_character(),
##   ..   host_since = col_date(format = ""),
##   ..   host_location = col_character(),
##   ..   host_about = col_character(),
##   ..   host_response_time = col_character(),
##   ..   host_response_rate = col_character(),
##   ..   host_acceptance_rate = col_character(),
##   ..   host_is_superhost = col_logical(),
##   ..   host_thumbnail_url = col_character(),
##   ..   host_picture_url = col_character(),
##   ..   host_neighbourhood = col_character(),
##   ..   host_listings_count = col_double(),
##   ..   host_total_listings_count = col_double(),
##   ..   host_verifications = col_character(),
##   ..   host_has_profile_pic = col_logical(),
##   ..   host_identity_verified = col_logical(),
##   ..   neighbourhood = col_character(),
##   ..   neighbourhood_cleansed = col_character(),
##   ..   neighbourhood_group_cleansed = col_logical(),
##   ..   latitude = col_double(),
##   ..   longitude = col_double(),
##   ..   property_type = col_character(),
##   ..   room_type = col_character(),
##   ..   accommodates = col_double(),
##   ..   bathrooms = col_logical(),
##   ..   bathrooms_text = col_character(),
##   ..   bedrooms = col_double(),
##   ..   beds = col_double(),
##   ..   amenities = col_character(),
##   ..   price = col_character(),
##   ..   minimum_nights = col_double(),
```

```
##   ..   maximum_nights = col_double(),
##   ..   minimum_minimum_nights = col_double(),
##   ..   maximum_minimum_nights = col_double(),
##   ..   minimum_maximum_nights = col_double(),
##   ..   maximum_maximum_nights = col_double(),
##   ..   minimum_nights_avg_ntm = col_double(),
##   ..   maximum_nights_avg_ntm = col_double(),
##   ..   calendar_updated = col_logical(),
##   ..   has_availability = col_logical(),
##   ..   availability_30 = col_double(),
##   ..   availability_60 = col_double(),
##   ..   availability_90 = col_double(),
##   ..   availability_365 = col_double(),
##   ..   calendar_last_scraped = col_date(format = ""),
##   ..   number_of_reviews = col_double(),
##   ..   number_of_reviews_ltm = col_double(),
##   ..   number_of_reviews_l30d = col_double(),
##   ..   first_review = col_date(format = ""),
##   ..   last_review = col_date(format = ""),
##   ..   review_scores_rating = col_double(),
##   ..   review_scores_accuracy = col_double(),
##   ..   review_scores_cleanliness = col_double(),
##   ..   review_scores_checkin = col_double(),
##   ..   review_scores_communication = col_double(),
##   ..   review_scores_location = col_double(),
##   ..   review_scores_value = col_double(),
##   ..   license = col_character(),
##   ..   instant_bookable = col_logical(),
##   ..   calculated_host_listings_count = col_double(),
##   ..   calculated_host_listings_count_entire_homes = col_double(),
##   ..   calculated_host_listings_count_private_rooms = col_double(),
##   ..   calculated_host_listings_count_shared_rooms = col_double(),
##   ..   reviews_per_month = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(airbnb_columbus_df)
```

```
##        id           listing_url          scrape_id          last_scraped
##  Min.   :   90676   Length:1562        Min.   :2.021e+13   Min.   :2021-12-23
##  1st Qu.:34057841   Class :character   1st Qu.:2.021e+13   1st Qu.:2021-12-23
##  Median :45501374   Mode  :character   Median :2.021e+13   Median :2021-12-24
##  Mean   :41011285                      Mean   :2.021e+13   Mean   :2021-12-23
##  3rd Qu.:51479495                      3rd Qu.:2.021e+13   3rd Qu.:2021-12-24
##  Max.   :53966237                      Max.   :2.021e+13   Max.   :2021-12-24
##
##      name           description        neighborhood_overview picture_url
##  Length:1562        Length:1562        Length:1562           Length:1562
##  Class :character   Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character   Mode  :character      Mode  :character
##
##
##
##
```

```
##     host_id            host_url          host_name           host_since
##  Min.   :    78761   Length:1562        Length:1562        Min.   :2010-02-07
##  1st Qu.: 37474030   Class :character   Class :character   1st Qu.:2015-07-03
##  Median :133736445   Mode  :character   Mode  :character   Median :2017-06-06
##  Mean   :163164383                                         Mean   :2017-06-05
##  3rd Qu.:253828606                                         3rd Qu.:2019-04-07
##  Max.   :436156183                                         Max.   :2021-12-15
##
##  host_location       host_about        host_response_time host_response_rate
##  Length:1562        Length:1562        Length:1562        Length:1562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  host_acceptance_rate host_is_superhost host_thumbnail_url host_picture_url
##  Length:1562          Mode :logical     Length:1562        Length:1562
##  Class :character     FALSE:913         Class :character   Class :character
##  Mode  :character     TRUE :649         Mode  :character   Mode  :character
##
##
##
##
##  host_neighbourhood host_listings_count host_total_listings_count
##  Length:1562        Min.   :  0.00      Min.   :  0.00
##  Class :character   1st Qu.:  1.00      1st Qu.:  1.00
##  Mode  :character   Median :  5.00      Median :  5.00
##                     Mean   : 26.34      Mean   : 26.34
##                     3rd Qu.: 24.00      3rd Qu.: 24.00
##                     Max.   :662.00      Max.   :662.00
##
##  host_verifications host_has_profile_pic host_identity_verified
##  Length:1562        Mode :logical        Mode :logical
##  Class :character   FALSE:6              FALSE:238
##  Mode  :character   TRUE :1556           TRUE :1324
##
##
##
##
##  neighbourhood      neighbourhood_cleansed neighbourhood_group_cleansed
##  Length:1562        Length:1562            Mode:logical
##  Class :character   Class :character       NA's:1562
##  Mode  :character   Mode  :character
##
##
##
##
##     latitude       longitude      property_type       room_type
##  Min.   :39.86   Min.   :-83.18   Length:1562        Length:1562
##  1st Qu.:39.96   1st Qu.:-83.01   Class :character   Class :character
##  Median :39.98   Median :-83.00   Mode  :character   Mode  :character
##  Mean   :39.99   Mean   :-82.99
##  3rd Qu.:40.00   3rd Qu.:-82.98
```

```
##  Max.   :40.15   Max.   :-82.78
##
##   accommodates    bathrooms      bathrooms_text        bedrooms
##  Min.   : 1.000   Mode:logical   Length:1562        Min.   : 1.000
##  1st Qu.: 2.000   NA's:1562      Class :character   1st Qu.: 1.000
##  Median : 4.000                  Mode  :character   Median : 2.000
##  Mean   : 4.872                                     Mean   : 1.901
##  3rd Qu.: 6.000                                     3rd Qu.: 3.000
##  Max.   :16.000                                     Max.   :10.000
##                                                     NA's   :62
##       beds          amenities            price           minimum_nights
##  Min.   : 1.000   Length:1562        Length:1562        Min.   :  1.000
##  1st Qu.: 1.000   Class :character   Class :character   1st Qu.:  1.000
##  Median : 2.000   Mode  :character   Mode  :character   Median :  2.000
##  Mean   : 2.401                                        Mean   :  5.569
##  3rd Qu.: 3.000                                        3rd Qu.:  2.000
##  Max.   :13.000                                        Max.   :300.000
##  NA's   :25
##  maximum_nights minimum_minimum_nights maximum_minimum_nights
##  Min.   :   1   Min.   :  1.000        Min.   :  1.000
##  1st Qu.: 100   1st Qu.:  1.000        1st Qu.:  1.000
##  Median :1124   Median :  2.000        Median :  2.000
##  Mean   : 684   Mean   :  5.005        Mean   :  6.541
##  3rd Qu.:1125   3rd Qu.:  2.000        3rd Qu.:  4.000
##  Max.   :1125   Max.   :300.000        Max.   :300.000
##
##  minimum_maximum_nights maximum_maximum_nights minimum_nights_avg_ntm
##  Min.   :   1.0         Min.   :   3.0         Min.   :  1.000
##  1st Qu.:1125.0         1st Qu.:1125.0         1st Qu.:  1.000
##  Median :1125.0         Median :1125.0         Median :  2.000
##  Mean   : 910.7         Mean   : 919.2         Mean   :  6.143
##  3rd Qu.:1125.0         3rd Qu.:1125.0         3rd Qu.:  3.000
##  Max.   :1125.0         Max.   :1125.0         Max.   :300.000
##
##  maximum_nights_avg_ntm calendar_updated has_availability availability_30
##  Min.   :   3.0         Mode:logical     Mode:logical     Min.   : 0.00
##  1st Qu.:1125.0         NA's:1562        TRUE:1562        1st Qu.:10.00
##  Median :1125.0                                           Median :21.00
##  Mean   : 916.5                                           Mean   :17.66
##  3rd Qu.:1125.0                                           3rd Qu.:26.00
##  Max.   :1125.0                                           Max.   :30.00
##
##  availability_60 availability_90 availability_365 calendar_last_scraped
##  Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   :2021-12-23
##  1st Qu.:30.00   1st Qu.:50.00   1st Qu.: 82.25   1st Qu.:2021-12-23
##  Median :48.00   Median :76.00   Median :176.00   Median :2021-12-24
##  Mean   :40.35   Mean   :63.45   Mean   :200.77   Mean   :2021-12-23
##  3rd Qu.:56.00   3rd Qu.:85.00   3rd Qu.:343.00   3rd Qu.:2021-12-24
##  Max.   :60.00   Max.   :90.00   Max.   :365.00   Max.   :2021-12-24
##
##  number_of_reviews number_of_reviews_ltm number_of_reviews_l30d
##  Min.   :  0.00    Min.   :  0.00        Min.   : 0.000
##  1st Qu.:  4.00    1st Qu.:  2.00        1st Qu.: 0.000
##  Median : 24.00    Median : 13.00        Median : 1.000
```

```
##  Mean   : 56.19    Mean   : 22.69        Mean   : 1.746
##  3rd Qu.: 75.75    3rd Qu.: 35.00        3rd Qu.: 3.000
##  Max.   :635.00    Max.   :152.00        Max.   :15.000
##
##   first_review         last_review         review_scores_rating
##  Min.   :2011-10-11   Min.   :2015-12-20   Min.   :1.000
##  1st Qu.:2019-03-31   1st Qu.:2021-11-14   1st Qu.:4.660
##  Median :2020-08-16   Median :2021-12-05   Median :4.830
##  Mean   :2020-02-21   Mean   :2021-10-24   Mean   :4.727
##  3rd Qu.:2021-08-03   3rd Qu.:2021-12-18   3rd Qu.:4.960
##  Max.   :2021-12-21   Max.   :2021-12-23   Max.   :5.000
##  NA's   :200          NA's   :200          NA's   :200
##  review_scores_accuracy review_scores_cleanliness review_scores_checkin
##  Min.   :1.000          Min.   :1.000             Min.   :1.000
##  1st Qu.:4.760          1st Qu.:4.643             1st Qu.:4.850
##  Median :4.900          Median :4.860             Median :4.940
##  Mean   :4.789          Mean   :4.729             Mean   :4.839
##  3rd Qu.:4.980          3rd Qu.:4.980             3rd Qu.:5.000
##  Max.   :5.000          Max.   :5.000             Max.   :5.000
##  NA's   :200            NA's   :200               NA's   :200
##  review_scores_communication review_scores_location review_scores_value
##  Min.   :1.00                Min.   :1.000          Min.   :1.000
##  1st Qu.:4.84                1st Qu.:4.680          1st Qu.:4.650
##  Median :4.95                Median :4.880          Median :4.800
##  Mean   :4.84                Mean   :4.741          Mean   :4.704
##  3rd Qu.:5.00                3rd Qu.:4.970          3rd Qu.:4.920
##  Max.   :5.00                Max.   :5.000          Max.   :5.000
##  NA's   :200                 NA's   :200            NA's   :200
##    license          instant_bookable calculated_host_listings_count
##  Length:1562        Mode :logical    Min.   : 1.00
##  Class :character   FALSE:890        1st Qu.: 2.00
##  Mode  :character   TRUE :672        Median : 5.00
##                                      Mean   :16.64
##                                      3rd Qu.:25.75
##                                      Max.   :96.00
##
##  calculated_host_listings_count_entire_homes
##  Min.   : 0.00
##  1st Qu.: 1.00
##  Median : 3.00
##  Mean   :15.47
##  3rd Qu.:24.50
##  Max.   :96.00
##
##  calculated_host_listings_count_private_rooms
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 1.153
##  3rd Qu.: 1.000
##  Max.   :21.000
##
##  calculated_host_listings_count_shared_rooms reviews_per_month
##  Min.   :0.00000                             Min.   : 0.040
```

```
##  1st Qu.:0.00000                          1st Qu.: 1.270
##  Median :0.00000                          Median : 2.565
##  Mean   :0.01665                          Mean   : 2.977
##  3rd Qu.:0.00000                          3rd Qu.: 4.130
##  Max.   :2.00000                          Max.   :21.530
##                                           NA's   :200
```

```
## Load the Affordable rental housing dataset
# housing_df=read.csv("Affordable_Rental_Housing_Developments.csv")
# glimpse(housing_df)

## Load the Average rent Columbus neighborhood dataset
avg_rent_df <- read_excel("data/Avg_rent_columbus_neighbourhood.xlsx")
glimpse(avg_rent_df)
```

```
## Rows: 114
## Columns: 2
## $ neighbourhood_cleansed <chr> "South Campus", "Downtown Columbus", "Indianola~
## $ `Average Rent`         <dbl> 769, 1225, 700, 1455, 770, 649, 700, 1200, 1510~
```

```
#Merge the airbnb df with rental housing df based on neighbourhood
#merged_df <- left_join(airbnb_columbus_df,avg_rent_df,by="neighbourhood_cleansed" )
#glimpse(merged_df)
#head(merged_df)

#Merge the above df with Average rent df based on neighbourhood
merged_df <- inner_join(x=airbnb_columbus_df,y=avg_rent_df,by=c("neighbourhood_cleansed") )
glimpse(merged_df)
```

```
## Rows: 22
## Columns: 75
## $ id                                    <dbl> 15028474, 19928533, 25033~
## $ listing_url                           <chr> "https://www.airbnb.com/r~
## $ scrape_id                             <dbl> 2.021122e+13, 2.021122e+1~
## $ last_scraped                          <date> 2021-12-23, 2021-12-23, ~
## $ name                                  <chr> "Housepitality ~ The Trav~
## $ description                           <chr> "Welcome to the Travelers~
## $ neighborhood_overview                 <chr> "Our neighborhood North L~
## $ picture_url                           <chr> "https://a0.muscache.com/~
## $ host_id                               <dbl> 26958698, 26958698, 40592~
## $ host_url                              <chr> "https://www.airbnb.com/u~
## $ host_name                             <chr> "Benjamin", "Benjamin", "~
## $ host_since                            <date> 2015-01-30, 2015-01-30, ~
## $ host_location                         <chr> "Columbus, Ohio, United S~
## $ host_about                            <chr> "I continue to be excited~
## $ host_response_time                    <chr> "within an hour", "within~
## $ host_response_rate                    <chr> "99%", "99%", "N/A", "90%~
## $ host_acceptance_rate                  <chr> "92%", "92%", "N/A", "83%~
## $ host_is_superhost                     <lgl> FALSE, FALSE, FALSE, FALS~
## $ host_thumbnail_url                    <chr> "https://a0.muscache.com/~
## $ host_picture_url                      <chr> "https://a0.muscache.com/~
## $ host_neighbourhood                    <chr> "Franklinton", "Franklint~
## $ host_listings_count                   <dbl> 108, 108, 1, 1, 2, 9, 0, ~
```

```
## $ host_total_listings_count                    <dbl> 108, 108, 1, 1, 2, 9, 0, ~
## $ host_verifications                           <chr> "['phone', 'reviews', 'ju~
## $ host_has_profile_pic                         <lgl> TRUE, TRUE, TRUE, TRUE, T~
## $ host_identity_verified                       <lgl> TRUE, TRUE, TRUE, TRUE, T~
## $ neighbourhood                                <chr> "Columbus, Ohio, United S~
## $ neighbourhood_cleansed                       <chr> "North Linden", "North Li~
## $ neighbourhood_group_cleansed                 <lgl> NA, NA, NA, NA, NA, NA, N~
## $ latitude                                     <dbl> 40.01799, 40.02038, 40.04~
## $ longitude                                    <dbl> -82.99314, -82.99268, -82~
## $ property_type                                <chr> "Entire bungalow", "Entir~
## $ room_type                                    <chr> "Entire home/apt", "Entir~
## $ accommodates                                 <dbl> 5, 6, 7, 14, 2, 12, 7, 6,~
## $ bathrooms                                    <lgl> NA, NA, NA, NA, NA, NA, N~
## $ bathrooms_text                               <chr> "1 bath", "1 bath", "2 ba~
## $ bedrooms                                     <dbl> 2, 2, 2, 3, 1, 5, 2, 3, 3~
## $ beds                                         <dbl> 2, 2, 3, 7, 1, 8, 5, 3, 3~
## $ amenities                                    <chr> "[\"Heating\", \"Patio or~
## $ price                                        <chr> "$109.00", "$90.00", "$12~
## $ minimum_nights                               <dbl> 30, 30, 1, 2, 5, 1, 1, 1,~
## $ maximum_nights                               <dbl> 1124, 1124, 1125, 1125, 1~
## $ minimum_minimum_nights                       <dbl> 2, 2, 1, 1, 5, 2, 1, 1, 2~
## $ maximum_minimum_nights                       <dbl> 60, 30, 2, 2, 5, 4, 1, 1,~
## $ minimum_maximum_nights                       <dbl> 1125, 1125, 1125, 1125, 1~
## $ maximum_maximum_nights                       <dbl> 1125, 1125, 1125, 1125, 1~
## $ minimum_nights_avg_ntm                       <dbl> 55.2, 29.9, 1.3, 1.3, 5.0~
## $ maximum_nights_avg_ntm                       <dbl> 1125, 1125, 1125, 1125, 1~
## $ calendar_updated                             <lgl> NA, NA, NA, NA, NA, NA, N~
## $ has_availability                             <lgl> TRUE, TRUE, TRUE, TRUE, T~
## $ availability_30                              <dbl> 0, 11, 14, 26, 8, 12, 25,~
## $ availability_60                              <dbl> 0, 41, 44, 56, 8, 37, 55,~
## $ availability_90                              <dbl> 25, 71, 70, 86, 31, 60, 8~
## $ availability_365                             <dbl> 300, 346, 286, 361, 120, ~
## $ calendar_last_scraped                        <date> 2021-12-23, 2021-12-23, ~
## $ number_of_reviews                            <dbl> 247, 194, 114, 92, 51, 15~
## $ number_of_reviews_ltm                        <dbl> 24, 24, 23, 20, 6, 45, 39~
## $ number_of_reviews_l30d                       <dbl> 0, 0, 3, 0, 0, 1, 2, 0, 1~
## $ first_review                                 <date> 2016-09-22, 2017-09-10, ~
## $ last_review                                  <date> 2021-10-31, 2021-09-26, ~
## $ review_scores_rating                         <dbl> 4.70, 4.76, 4.93, 4.78, 4~
## $ review_scores_accuracy                       <dbl> 4.78, 4.81, 4.93, 4.87, 4~
## $ review_scores_cleanliness                    <dbl> 4.68, 4.71, 4.82, 4.84, 4~
## $ review_scores_checkin                        <dbl> 4.86, 4.85, 4.85, 4.91, 5~
## $ review_scores_communication                  <dbl> 4.84, 4.82, 4.87, 4.89, 4~
## $ review_scores_location                       <dbl> 4.53, 4.59, 4.94, 4.59, 4~
## $ review_scores_value                          <dbl> 4.70, 4.69, 4.89, 4.82, 4~
## $ license                                      <chr> "2019-1140", "2019-1139",~
## $ instant_bookable                             <lgl> FALSE, FALSE, FALSE, FALS~
## $ calculated_host_listings_count               <dbl> 96, 96, 1, 1, 2, 8, 1, 1,~
## $ calculated_host_listings_count_entire_homes  <dbl> 96, 96, 1, 1, 2, 8, 1, 1,~
## $ calculated_host_listings_count_private_rooms <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ calculated_host_listings_count_shared_rooms  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ reviews_per_month                            <dbl> 3.86, 3.72, 2.64, 2.37, 1~
## $ `Average Rent`                               <dbl> 715, 715, 715, 715, 715, ~
```

```r
# By looking at the data we can conclude that:
# Variable id is just an identifier which can be ignored.
# We can factor the field room.type - Private room,Entire home/apt,Hotel room, Shared room
# We can drop the host.id and host.name,neighbourhood.group,name fields from the dataset
# We can drop fields like last.review,number.of.reviews, reviews.per.month,calculated.host.listings.cou

#Average rent Columbus neighborhood data
# Rename the Average Rent to Average_Rent

# Apply above transformation to the dataframe
final_df <- subset(merged_df, select = c("neighbourhood_cleansed",
                                         "latitude",
                                         "longitude",
                                         "room_type",
                                         "price","minimum_nights",
                                         "availability_365",
                                         "property_type",
                                         "Average Rent") )
glimpse(final_df)
```

```
## Rows: 22
## Columns: 9
## $ neighbourhood_cleansed <chr> "North Linden", "North Linden", "North Linden",~
## $ latitude               <dbl> 40.01799, 40.02038, 40.04448, 40.01694, 40.0296~
## $ longitude              <dbl> -82.99314, -82.99268, -82.99409, -82.98537, -82~
## $ room_type              <chr> "Entire home/apt", "Entire home/apt", "Entire h~
## $ price                  <chr> "$109.00", "$90.00", "$121.00", "$165.00", "$48~
## $ minimum_nights         <dbl> 30, 30, 1, 2, 5, 1, 1, 1, 30, 1, 30, 5, 1, 1, 2~
## $ availability_365       <dbl> 300, 346, 286, 361, 120, 320, 160, 330, 344, 89~
## $ property_type          <chr> "Entire bungalow", "Entire bungalow", "Entire r~
## $ `Average Rent`         <dbl> 715, 715, 715, 715, 715, 715, 715, 715, 715, 71~
```

```r
#Rename Average Rent to Average_Rent
colnames(final_df)[9] <- "Average_Rent"

# Checking the summary of data set to gauge the value range of each numerical variable
summary(final_df)
```

```
##  neighbourhood_cleansed    latitude       longitude       room_type
##  Length:22              Min.   :40.02   Min.   :-83.00   Length:22
##  Class :character       1st Qu.:40.02   1st Qu.:-82.99   Class :character
##  Mode  :character       Median :40.02   Median :-82.99   Mode  :character
##                         Mean   :40.03   Mean   :-82.98
##                         3rd Qu.:40.03   3rd Qu.:-82.98
##                         Max.   :40.06   Max.   :-82.94
##     price           minimum_nights   availability_365 property_type
##  Length:22          Min.   : 1.000   Min.   : 56      Length:22
##  Class :character   1st Qu.: 1.000   1st Qu.:114      Class :character
##  Mode  :character   Median : 2.000   Median :164      Mode  :character
##                     Mean   : 9.545   Mean   :205
##                     3rd Qu.:23.750   3rd Qu.:315
##                     Max.   :30.000   Max.   :365
##   Average_Rent
```

```
##  Min.   :715
##  1st Qu.:715
##  Median :715
##  Mean   :715
##  3rd Qu.:715
##  Max.   :715
```

```
# Range of values prices are varies from 0 to 10000.
# It looks like there are outliers in the field.
# Range of values minimum_nights varies from 1 to 365.
# It looks like there are outliers in the field.
# Range of values for availability_365 varies from 0 to 365.


#Calculate the 30 days price for airbnb property.
final_df$airbnb_30_days_price=as.numeric(final_df$price) * 30
```

```
## Warning: NAs introduced by coercion
```

```
summary(final_df)
```

```
##  neighbourhood_cleansed    latitude       longitude       room_type
##  Length:22              Min.   :40.02   Min.   :-83.00   Length:22
##  Class :character       1st Qu.:40.02   1st Qu.:-82.99   Class :character
##  Mode  :character       Median :40.02   Median :-82.99   Mode  :character
##                         Mean   :40.03   Mean   :-82.98
##                         3rd Qu.:40.03   3rd Qu.:-82.98
##                         Max.   :40.06   Max.   :-82.94
##
##     price           minimum_nights   availability_365 property_type
##  Length:22          Min.   : 1.000   Min.   : 56      Length:22
##  Class :character   1st Qu.: 1.000   1st Qu.:114      Class :character
##  Mode  :character   Median : 2.000   Median :164      Mode  :character
##                     Mean   : 9.545   Mean   :205
##                     3rd Qu.:23.750   3rd Qu.:315
##                     Max.   :30.000   Max.   :365
##
##   Average_Rent airbnb_30_days_price
##  Min.   :715   Min.   : NA
##  1st Qu.:715   1st Qu.: NA
##  Median :715   Median : NA
##  Mean   :715   Mean   :NaN
##  3rd Qu.:715   3rd Qu.: NA
##  Max.   :715   Max.   : NA
##                NA's   :22
```

```
#Check missing values
apply(final_df, 2, function(x) any(is.na(x)))
```

```
## neighbourhood_cleansed              latitude             longitude
##                  FALSE                 FALSE                 FALSE
##             room_type                 price        minimum_nights
```

```
##                   FALSE                   FALSE                   FALSE
##       availability_365           property_type           Average_Rent
##                   FALSE                   FALSE                   FALSE
##   airbnb_30_days_price
##                    TRUE
```

```
## 2.What does the final data set look like?
glimpse(final_df)
```

```
## Rows: 22
## Columns: 10
## $ neighbourhood_cleansed <chr> "North Linden", "North Linden", "North Linden",~
## $ latitude               <dbl> 40.01799, 40.02038, 40.04448, 40.01694, 40.0296~
## $ longitude              <dbl> -82.99314, -82.99268, -82.99409, -82.98537, -82~
## $ room_type              <chr> "Entire home/apt", "Entire home/apt", "Entire h~
## $ price                  <chr> "$109.00", "$90.00", "$121.00", "$165.00", "$48~
## $ minimum_nights         <dbl> 30, 30, 1, 2, 5, 1, 1, 1, 30, 1, 30, 5, 1, 1, 2~
## $ availability_365       <dbl> 300, 346, 286, 361, 120, 320, 160, 330, 344, 89~
## $ property_type          <chr> "Entire bungalow", "Entire bungalow", "Entire r~
## $ Average_Rent           <dbl> 715, 715, 715, 715, 715, 715, 715, 715, 715, 71~
## $ airbnb_30_days_price   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
## 3. Questions for future steps.
# a) Need to learn how to visualize more than two variables.
# b) Need to learn application of variable scaling and techniques.
# c) Need to learn how lm() function takes care of variable scaling.
# d) Need to learn correlation between different variables.

## 4.What information is not self-evident?
# To uncover new information in the data that is not self-evident -
# 1. visualize data to uncover patterns and trends
# 2. correlation among variables
# 3. Check data distribution of variables
# 4. detect outliers and influencial cases
# 5.What are different ways you could look at this data?

# Checking relation between airbnb_30_days_price and Average_Rent using
ggplot()
```

```
library(ggplot2)
ggplot(data = final_df, aes(x = airbnb_30_days_price, y = Average_Rent)) +
  geom_point() + geom_smooth(fill=NA)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

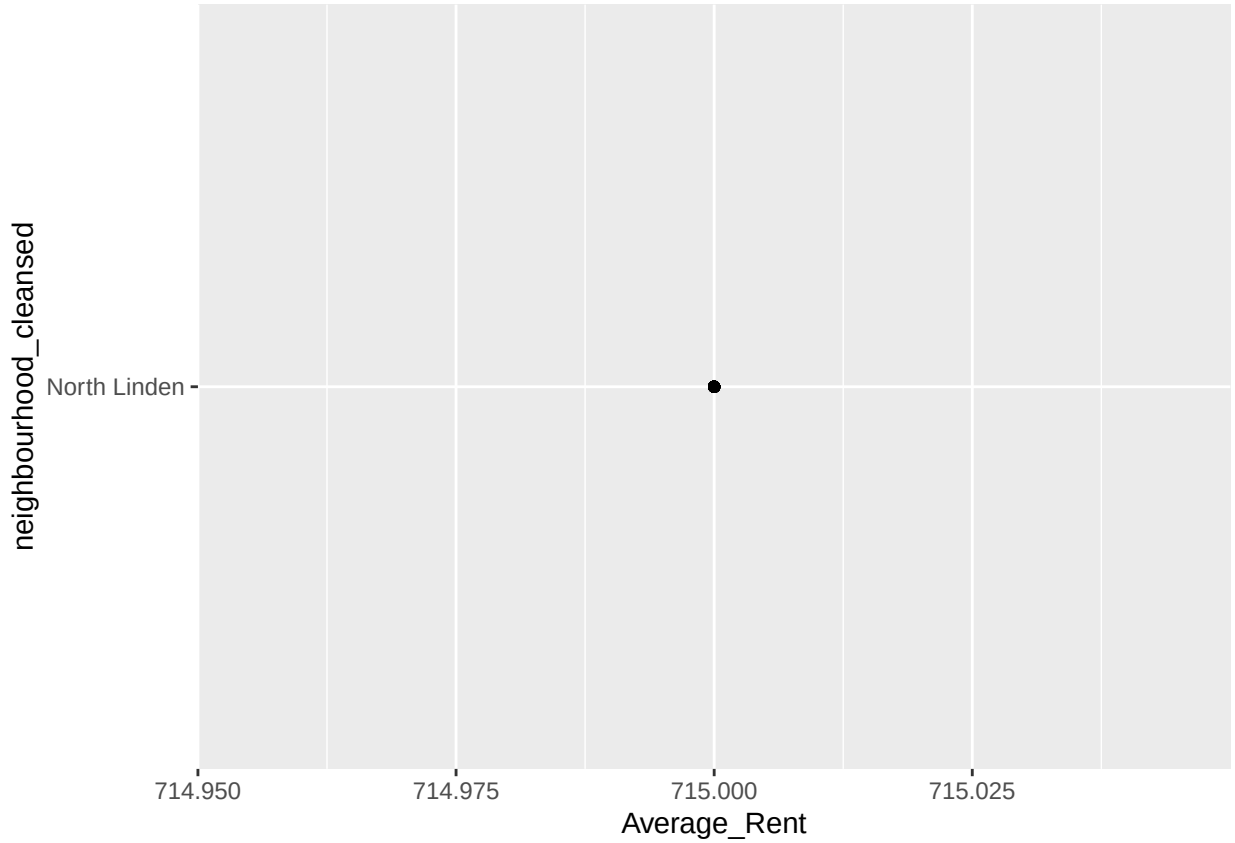
```
## Warning: Removed 22 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 22 rows containing missing values (geom_point).
```


```
# Checking relation between neighbourhood_cleansed and Average_Rent using
ggplot()
```

```
library(ggplot2)
ggplot(data = final_df, aes(y = neighbourhood_cleansed, x = Average_Rent)) +
  geom_point() + geom_smooth(fill=NA)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
# Checking relation between neighbourhood_cleansed and airbnb_30_days_price using
ggplot()
```

```
library(ggplot2)
ggplot(data = final_df, aes(y = neighbourhood_cleansed, x = airbnb_30_days_price)) +
  geom_point() + geom_smooth(fill=NA)
```

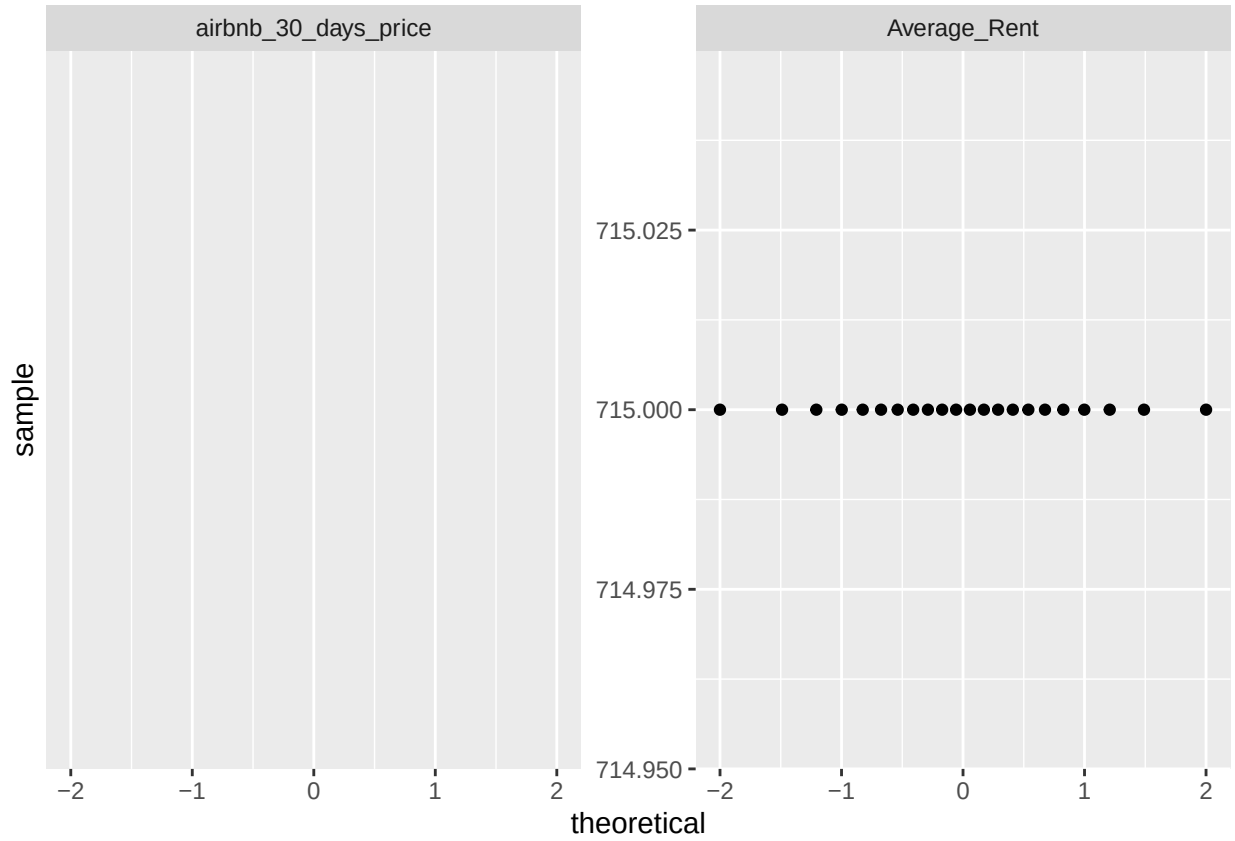
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 22 rows containing non-finite values (stat_smooth).

## Warning: Removed 22 rows containing missing values (geom_point).
```


```
# Identify the relationship between neighbourhood and prices here.
# Checking if data distribution of numeric variables is normal
# combining pipe operator between dplyr transformation and ggplot
final_df %>% select(airbnb_30_days_price, Average_Rent) %>%
  gather() %>%
  ggplot(., aes(sample = value)) +
  stat_qq() +
  facet_wrap(vars(key), scales ='free_y')
```

```
## Warning: Removed 22 rows containing non-finite values (stat_qq).
```

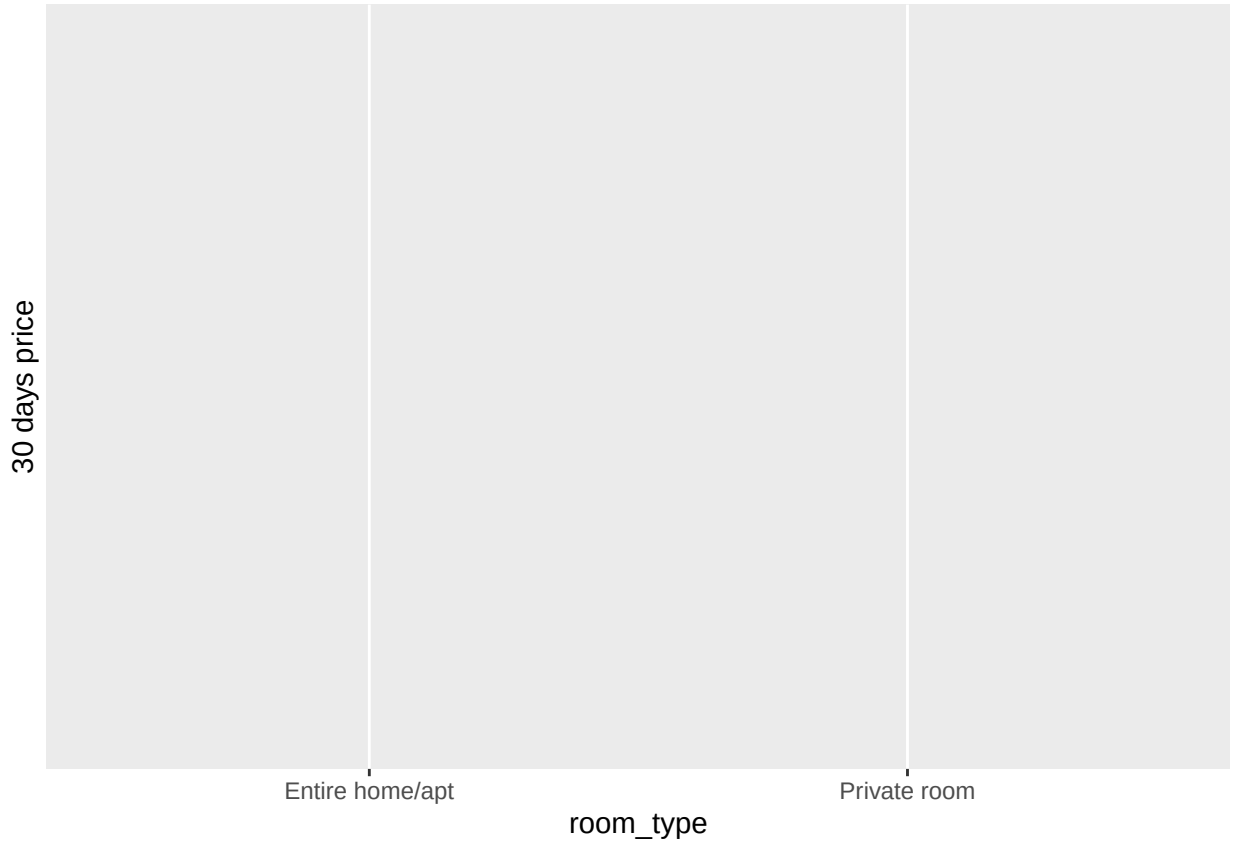



```
#None of the variables looks normally distributed
ggplot(data = final_df, aes(x = neighbourhood_cleansed , y = airbnb_30_days_price)) +
  geom_boxplot() + ylab("airbnb_30_days_price")
```

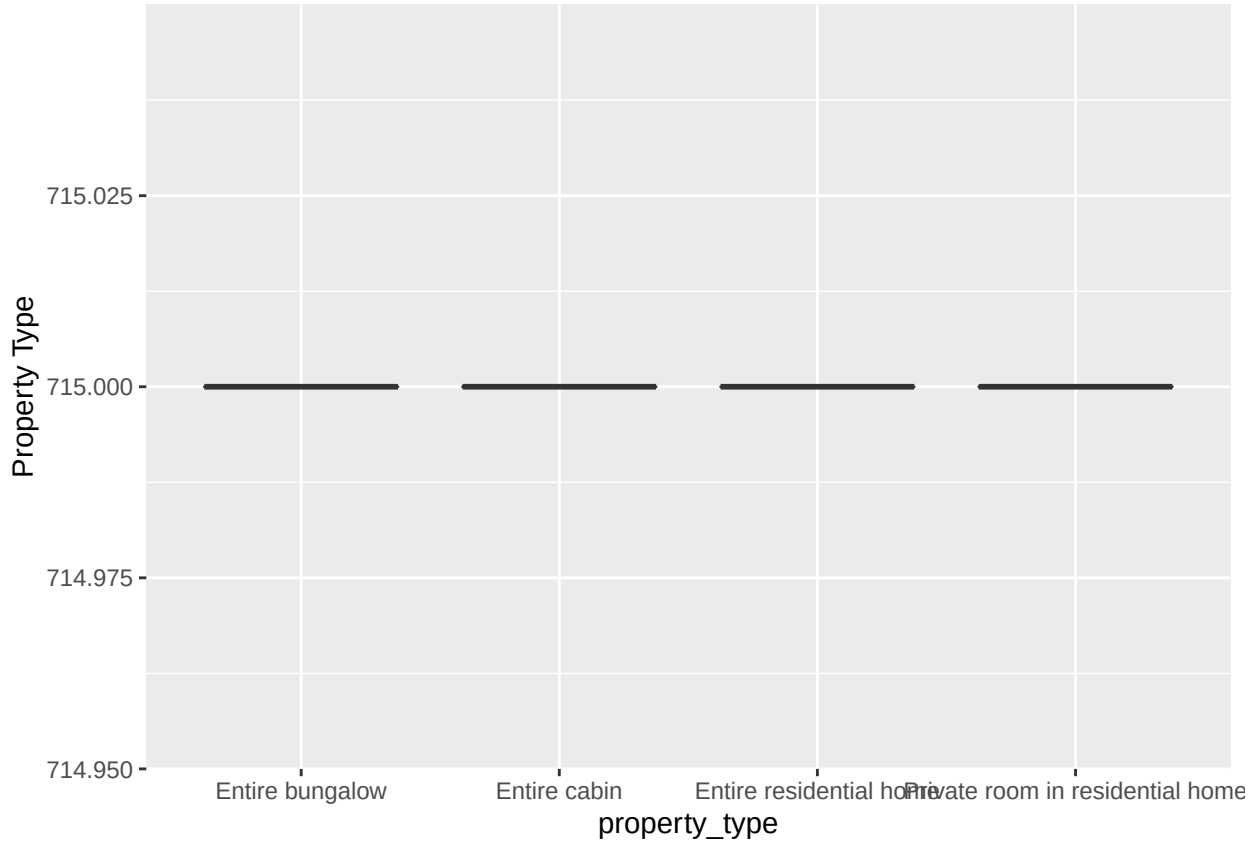
```
## Warning: Removed 22 rows containing non-finite values (stat_boxplot).
```


```
# We can see that there are so many outliers for many neighbourhoods
# thus data is not normally distributed
ggplot(data = final_df, aes(x = room_type , y = airbnb_30_days_price)) +
  geom_boxplot() + ylab("30 days price")
```

```
## Warning: Removed 22 rows containing non-finite values (stat_boxplot).
```




```
# We can see that there are so many outliers for room_type
# thus data is not normally distributed
ggplot(data = final_df, aes(x = property_type , y = Average_Rent)) +
  geom_boxplot() + ylab("Property Type")
```




```
# We can see that there are so many outliers for Property_Type
# thus data is not normally distributed

# 6.How do you plan to slice and dice the data?

unique(final_df[c("neighbourhood_cleansed")])
```

```
## # A tibble: 1 x 1
##   neighbourhood_cleansed
##   <chr>
## 1 North Linden
```

```
# I think need to slice the datasets by zip codes or neighbourhood to analyze
# the data in more granular level

# How could you summarize your data to answer key questions?
library("ggpubr")
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:plyr':
##
##     mutate
```

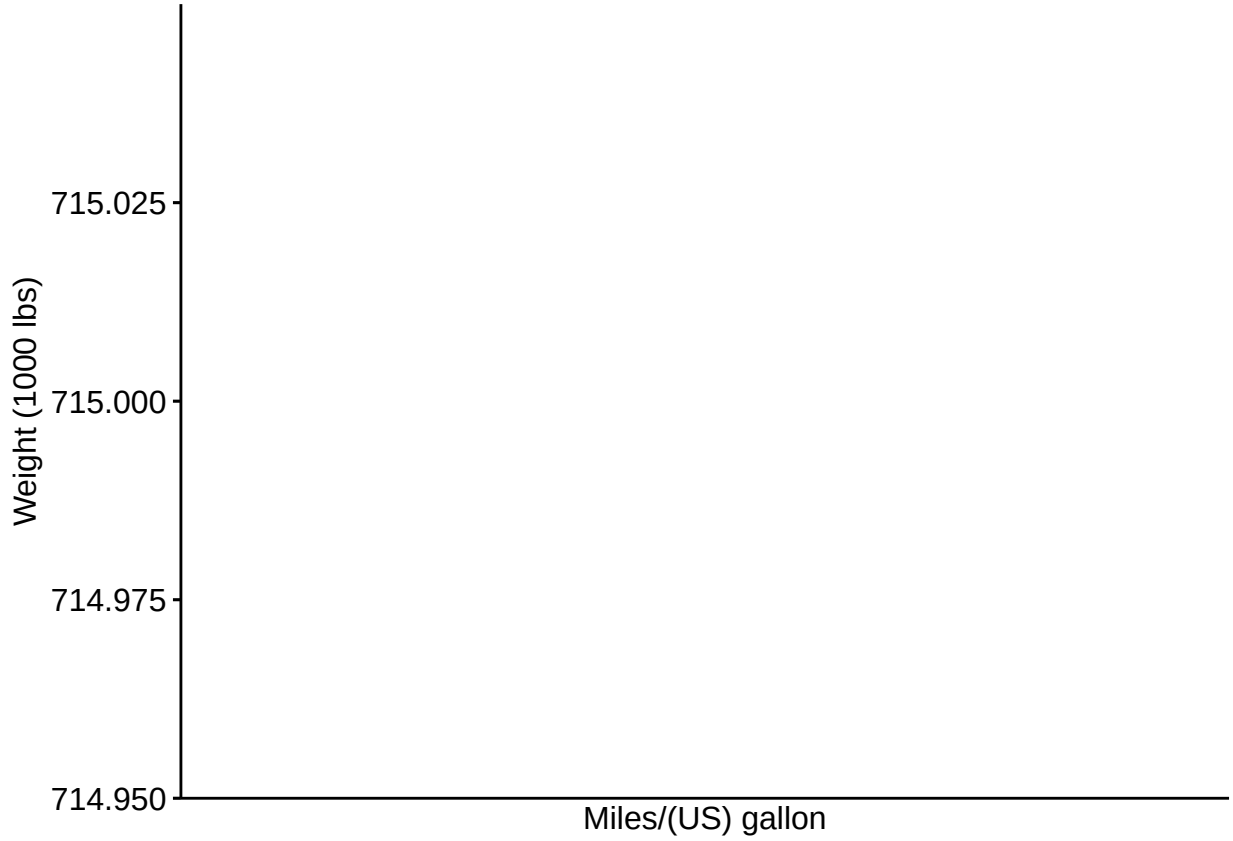
```
ggscatter(final_df, x = "airbnb_30_days_price", y = "Average_Rent",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Miles/(US) gallon", ylab = "Weight (1000 lbs)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 22 rows containing non-finite values (stat_smooth).
```

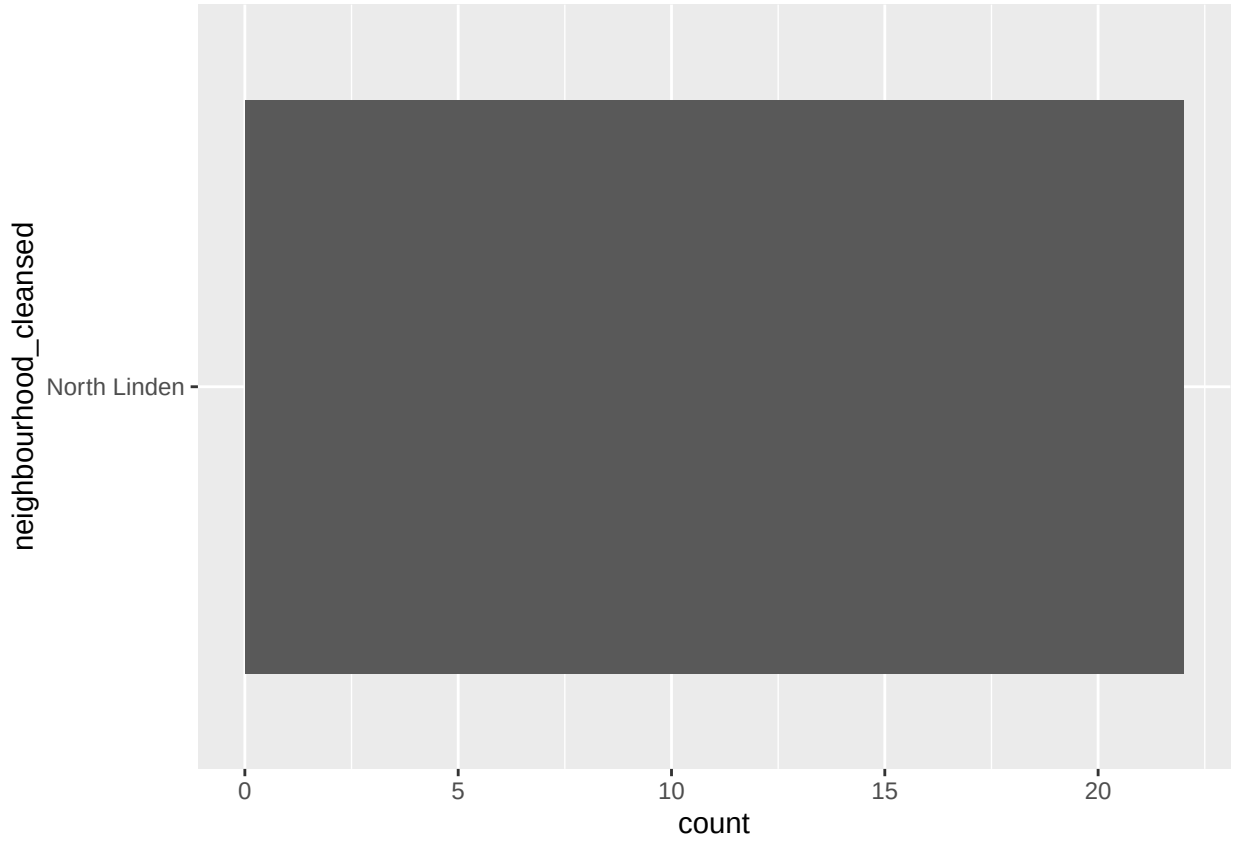
```
## Warning: Removed 22 rows containing non-finite values (stat_cor).
```

```
## Warning: Removed 22 rows containing missing values (geom_point).
```




```
#a) What are the Airbnb rental prices for different areas in Chicago?
ggplot(data=final_df,aes(y=neighbourhood_cleansed)) + geom_histogram(stat = "count")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
ggplot(aes(y=neighbourhood_cleansed,x=airbnb_30_days_price),data=final_df)+
  geom_point()
```

```
## Warning: Removed 22 rows containing missing values (geom_point).
```

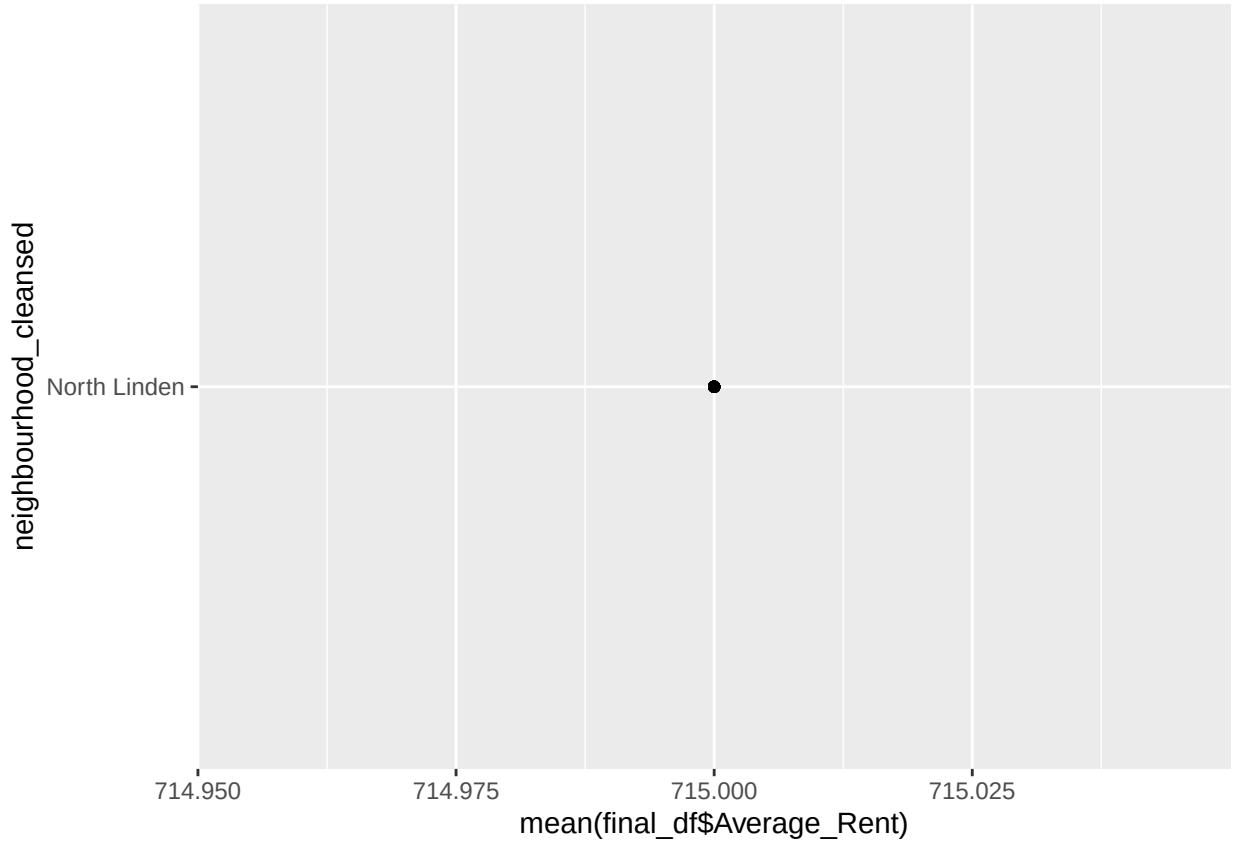

```
# b) What is the correlation between the Airbnb rental prices and Columbus
# neighborhood rent prices?

cor(final_df$airbnb_30_days_price,final_df$Average_Rent)
```

```
## [1] NA
```

```
# c)What are the average rent prices by the neighborhood?
ggplot(aes(y=neighbourhood_cleansed,x=mean(final_df$Average_Rent)),data=final_df)+
  geom_point()
```

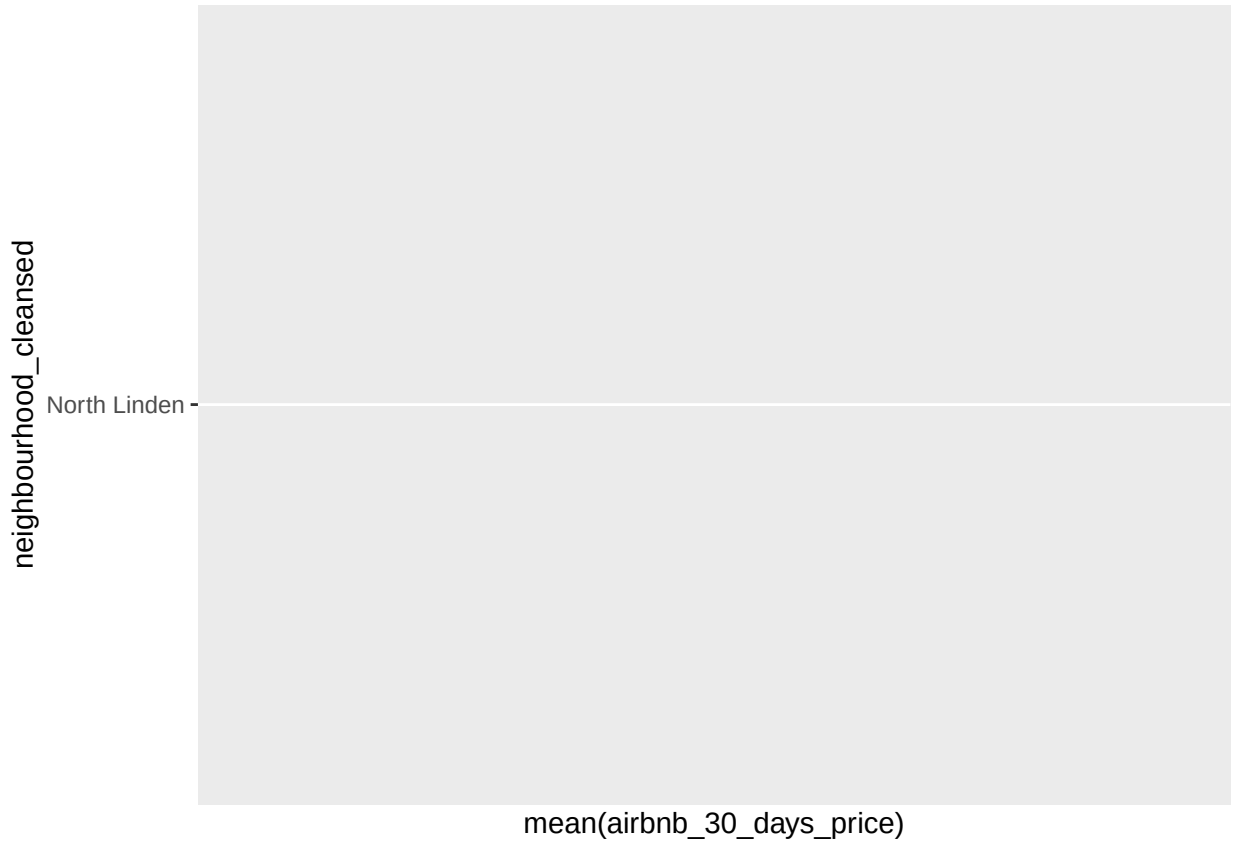
```
## Warning: Use of ‘final_df$Average_Rent‘ is discouraged. Use ‘Average_Rent‘
## instead.
```




```
#The average rent price is xxx per month
# d)What are the average rent prices for Airbnb by the neighborhood?

ggplot(aes(y=neighbourhood_cleansed,x=mean(airbnb_30_days_price)),data=final_df)+
  geom_point()
```

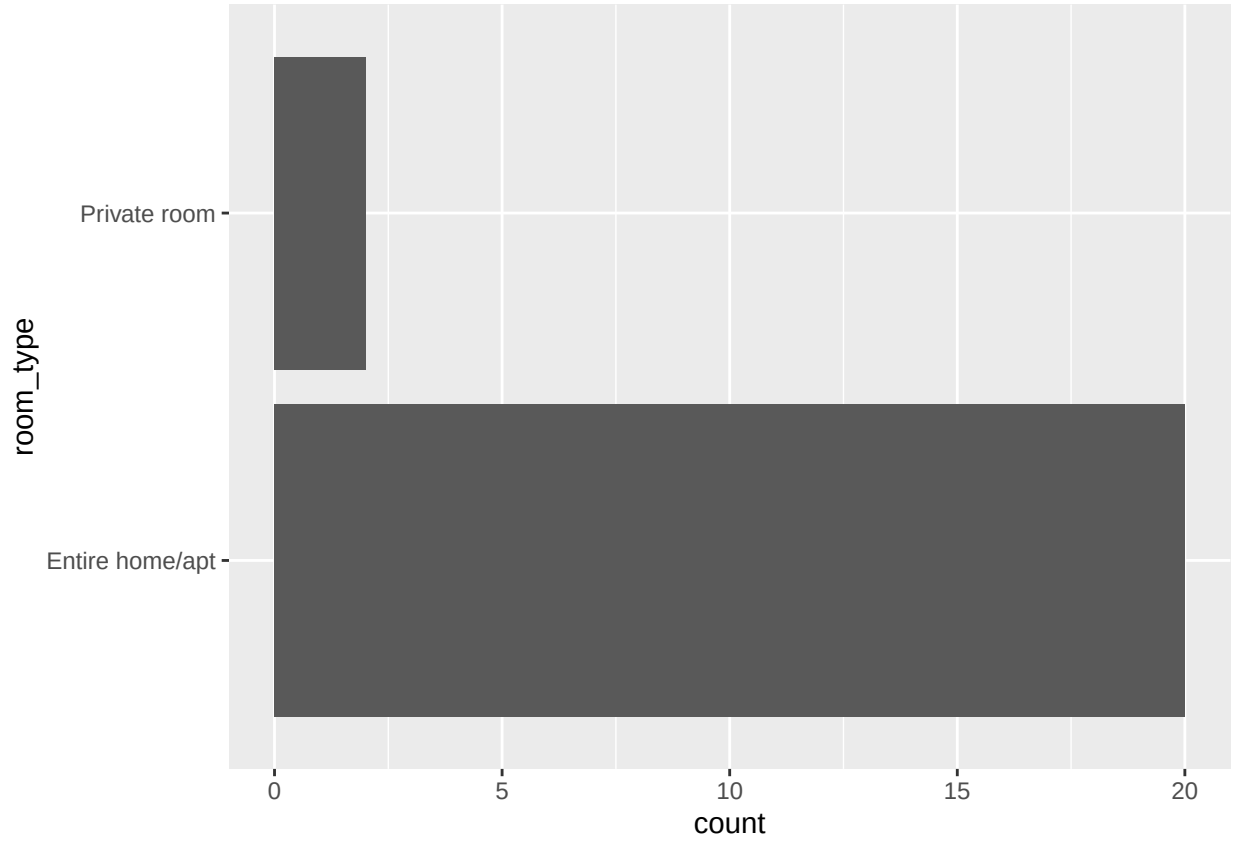
```
## Warning: Removed 22 rows containing missing values (geom_point).
```




```
#The average airbnb price is xxxx per month

# e) What type of houses are most rented on Airbnb?
ggplot(data=final_df,aes(y=room_type)) + geom_histogram(stat ="count")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
# f)What is the monthly rent from the Airbnb properties?
df1 <-final_df%>%select(neighbourhood_cleansed, airbnb_30_days_price, Average_Rent)
df1 %>% group_by(neighbourhood_cleansed) %>% summarize(mean_airbnb_30_days_price =
                                                         mean(airbnb_30_days_price))
```

```
##   mean_airbnb_30_days_price
## 1                        NA
```

```
#Airbnb monrthly average rent is xxxx
# Do you plan on incorporating any machine learning techniques to answer
# your research questions? Explain.
# performing multiple linear regression
# splitting the data into training and test set
library(caTools)
#mymodel_1 <-lm(airbnb_30_days_price ~ neighbourhood_cleansed,data = final_df)
#summary(mymodel_1)

#mymodel_2 <-lm(airbnb_30_days_price ~ Zip.Code,data = final_df)
#summary(mymodel_2)

# Questions for future steps?
# # Iam still working on fixing my data for the datasets being used here.
# Might need to few more variables especially zip codes which I am missing currently in my data.
# I would like to plot the airbnb properties on map
# I think I need to look for more data to determine the correlation and to
# predict prices accurately
```