Chicago Airbnb Listings

Background

Our dataset is made up of a random sample of 500 Airbnb listings from Chicago taken from all listings in Chicago from August 2008 to May 2017. The dataset contains attributes for each of the listings such as review scores, features of the listing (eg number of rooms, bathrooms, etc.), prices, location (coordinates, neighborhood), services provided (heat, wifi), rules (whether smoking or pets are allowed). The data was scraped by Professor Laura Ziegler of the ISU statistics department.

Cleaning

```
##
## 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

## 'data.frame':    500 obs. of  63 variables:
##  $ id                          : int  14843458 722570 8153968 18327990 12673204 10344726 5737004 1682
##  $ listing_url                 : Factor w/ 500 levels "https://www.airbnb.com/rooms/10043106",..: 131
##  $ name                        : Factor w/ 500 levels "#61  King Arthurs Court",..: 177 302 191 410 1
##  $ summary                     : Factor w/ 492 levels "","- Townhome 150 yds (2 min walk) from McCorm
##  $ space                       : Factor w/ 366 levels "","- This is a great, two bedroom, one bathroo
##  $ description                 : Factor w/ 499 levels "- Townhome 150 yds (2 min walk) from McCormick
##  $ neighborhood_overview       : Factor w/ 331 levels "","-----------------------------------------
##  $ notes                       : Factor w/ 225 levels "","- Any short-term stays of 3 or lesser night
##  $ transit                     : Factor w/ 346 levels "","- 1 min walk to the Lawrence Red Line stati
##  $ access                      : Factor w/ 315 levels "","#94 is the door code. How to receive keys w
##  $ interaction                 : Factor w/ 313 levels "","24/7 access to the host via phone, text, or
##  $ house_rules                 : Factor w/ 318 levels "","-- Not handicap accessible (there are 59 st
##  $ host_id                     : int  20653807 3731751 16500117 6903096 34473759 45766549 29751294 11
##  $ host_url                    : Factor w/ 461 levels "https://www.airbnb.com/users/show/100027760",.
##  $ host_name                   : Factor w/ 365 levels "Aama","Aamir",..: 170 211 314 117 206 327 111
##  $ host_since                  : Factor w/ 413 levels "01/02/2016","01/02/2017",..: 281 319 170 181 1
##  $ host_location               : Factor w/ 31 levels "Barrington, Illinois, United States",..: 16 7 7
##  $ host_about                  : Factor w/ 319 levels "","\"Depth and breadth are crucial to creativi
##  $ host_response_rate          : Factor w/ 32 levels "0%","100%","33%",..: 2 2 2 2 2 2 32 2 2 10 ...
##  $ host_is_superhost           : Factor w/ 2 levels "f","t": 1 1 2 1 2 1 1 1 2 1 ...
##  $ host_neighbourhood          : Factor w/ 61 levels "","Albany Park",..: 1 28 1 1 28 28 28 28 1 21 .
##  $ host_verifications          : Factor w/ 82 levels "['email', 'phone', 'amex', 'reviews', 'kba', 'w
##  $ host_has_profile_pic        : Factor w/ 2 levels "f","t": 2 2 2 2 2 2 2 2 2 2 ...
##  $ host_identity_verified      : Factor w/ 2 levels "f","t": 2 2 2 1 2 2 2 1 2 1 ...
##  $ street                      : Factor w/ 126 levels "Albany Park, Chicago, IL 60625, United States"
##  $ neighbourhood               : Factor w/ 45 levels "Albany Park",..: 20 20 20 20 20 20 20 20 14 14
##  $ latitude                    : num  42 42 42 42 42 ...
##  $ longitude                  : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ is_location_exact          : Factor w/ 2 levels "f","t": 1 2 1 1 2 2 2 2 1 2 ...
##  $ room_type                  : Factor w/ 3 levels "Entire home/apt",..: 1 1 2 1 1 1 1 1 2 1 ...
##  $ accommodates               : int  3 4 4 2 3 3 2 6 2 4 ...
##  $ bathrooms                  : num  1 2 1 1 1 1 1 2 1 1 ...
##  $ bedrooms                   : int  1 2 1 1 2 0 1 3 1 1 ...
##  $ beds                       : int  1 2 2 1 2 2 1 4 1 2 ...
##  $ bed_type                   : Factor w/ 5 levels "Airbed","Couch",..: 5 5 5 5 5 5 5 5 5 1 ...
##  $ amenities                  : Factor w/ 495 levels "{\"Air conditioning\",Kitchen,\"Pets allowed\"
##  $ price                      : int  69 139 65 80 150 83 85 85 59 120 ...
##  $ monthly_price              : int  NA NA 1600 NA NA 695 NA NA 1470 NA ...
##  $ security_deposit           : int  NA 300 NA NA NA NA NA 150 NA NA ...
##  $ cleaning_fee               : int  20 80 15 NA 35 NA NA 59 20 NA ...
##  $ guests_included            : int  1 3 2 1 1 2 1 2 2 4 ...
##  $ price_extra_people         : int  48 20 10 0 0 10 0 15 0 25 ...
##  $ maximum_nights             : int  14 21 1125 1125 1125 1125 1125 1125 31 1125 ...
##  $ calendar_updated           : Factor w/ 22 levels "1 week ago","12 months ago",..: 19 5 5 5 5 19 1
##  $ availability_30            : int  6 1 16 0 0 0 0 9 8 11 ...
##  $ availability_60            : int  7 1 43 0 8 3 0 15 36 23 ...
##  $ availability_90            : int  13 5 73 0 19 16 0 15 61 30 ...
##  $ availability_365           : int  13 280 73 129 19 73 0 46 332 305 ...
##  $ review_scores_cleanliness  : int  9 10 10 NA 10 10 10 8 10 9 ...
##  $ review_scores_communication: int  10 10 10 NA 10 10 10 10 10 10 ...
##  $ review_scores_value        : int  10 10 10 NA 10 10 10 10 10 9 ...
##  $ instant_bookable           : Factor w/ 2 levels "f","t": 1 1 1 2 1 1 1 1 1 1 ...
##  $ cancellation_policy        : Factor w/ 3 levels "flexible","moderate",..: 1 2 2 1 2 2 2 3 2 1 ...
##  $ reviews_per_month          : num  0.96 0.29 2.61 NA 0.43 2.28 0.16 6.32 3.47 1.58 ...
##  $ cable_tv                   : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 2 2 1 ...
##  $ wireless_internet          : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 2 2 ...
##  $ kitchen                    : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 2 2 ...
##  $ pets_allowed               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ breakfast                  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
##  $ heating                    : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 1 2 2 2 ...
##  $ X24.hour_checkin           : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 2 2 ...
##  $ smoking_allowed            : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 1 ...
##  $ smoking_allowed.2          : int  0 0 0 0 0 0 1 0 0 0 ...
```

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

Initial EDA

Now, we can start with some exploration of individual variables. First off, we'll look at distributions of some simple numerical variables.

Host response rate:

```
## Total Count:  500
## N:  480
## Minimum:  0
## Maximum:  100
## Mean:  96.23125
## Standard Deviation:  10.81784
## Median:  100
## Quantiles:
##   0%  25%  50%  75% 100%
##    0  100  100  100  100
## IQR:  0
```

Given that we have a median equal to our maximum, 100, along with such a high mean and low IQR, we clearly have a majority of hosts having a 100% response rate.

Number of guests accommodated:

```
## Total Count:  500
## N:  500
## Minimum:  1
## Maximum:  16
## Mean:  3.872
## Standard Deviation:  2.600219
## Median:  3
## Quantiles:
##   0%  25%  50%  75% 100%
##    1    2    3    5   16
## IQR:  3
```

Price:

```
## Total Count:  500
## N:  500
## Minimum:  10
## Maximum:  950
## Mean:  135.416
## Standard Deviation:  125.9147
## Median:  100
## Quantiles:
##   0%  25%  50%  75% 100%
##   10   60  100  155  950
## IQR:  95
```

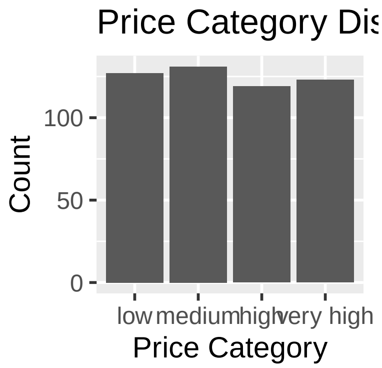
We will get four price categories based on our quantiles, which we'll call very high, high, medium, and low, for later categorization use.

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```

```
## [1] TRUE
```



Maximum nights:

```
## Total Count:  500
## N:  500
## Minimum:  3
## Maximum:  1125
## Mean:  736.986
## Standard Deviation:  518.5947
## Median:  1125
## Quantiles:
##   0%  25%  50%  75% 100% 
##    3   30 1125 1125 1125 
## IQR:  1095
```

Note the high proportion of 1125s, probably the maximum allowed for a listing by the website.

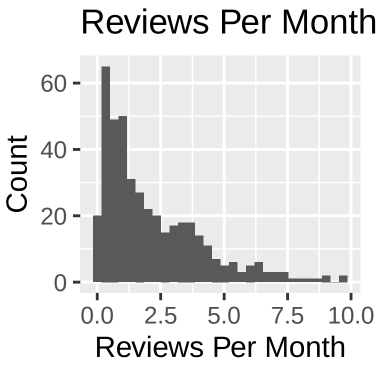
Reviews per month:

```
## Total Count:  500
## N:  426
## Minimum:  0.05
## Maximum:  9.75
## Mean:  2.135563
## Standard Deviation:  1.960553
## Median:  1.485
## Quantiles:
##    0%   25%   50%   75%  100% 
## 0.050 0.620 1.485 3.245 9.750 
## IQR:  2.625
```

This gives an interesting variety, let's look at a histogram.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 74 rows containing non-finite values (stat_bin).
```



We can see a right skew with a floor at 0, so we know most locations aren't reviewed many times per month.

Cleanliness review score (out of 10):

```
## Total Count:  500
## N:  425
## Minimum:  4
## Maximum:  10
## Mean:  9.477647
## Standard Deviation:  0.8467998
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100% 
##    4    9   10   10   10 
## IQR:  1
```

Communication review score (out of 10):

```
## Total Count:  500
## N:  424
## Minimum:  8
## Maximum:  10
## Mean:  9.856132
## Standard Deviation:  0.4016045
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100% 
##    8   10   10   10   10 
## IQR:  0
```

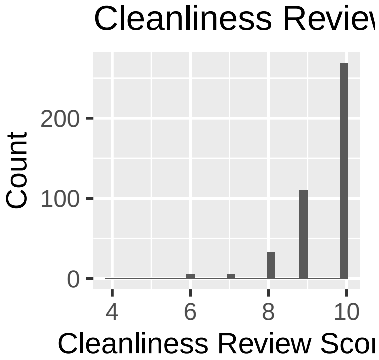
Value review score (out of 10):

```
## Total Count:  500
## N:  422
## Minimum:  6
## Maximum:  10
## Mean:  9.56872
## Standard Deviation:  0.6673718
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100%
##    6    9   10   10   10
## IQR:  1
```

Seems like all three review categories seem to tend toward the high end. A left skew is likely. Let's check:

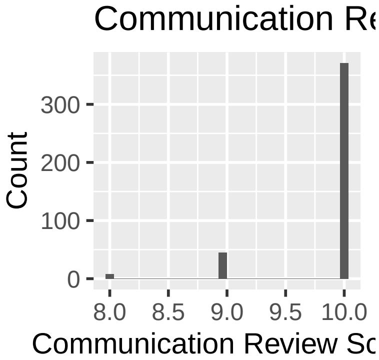
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 75 rows containing non-finite values (stat_bin).
```



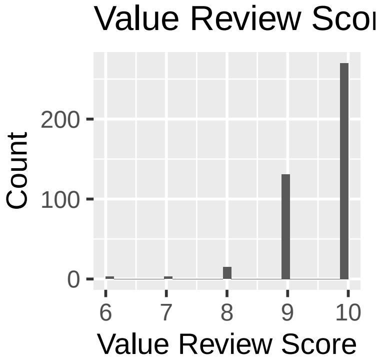
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 76 rows containing non-finite values (stat_bin).
```



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 78 rows containing non-finite values (stat_bin).
```



Unsurprising. People tend to give high reviews in the dataset.

Let's see how easily plottable some variables are:

```
## [1] 365
```

```
## [1] 31
```

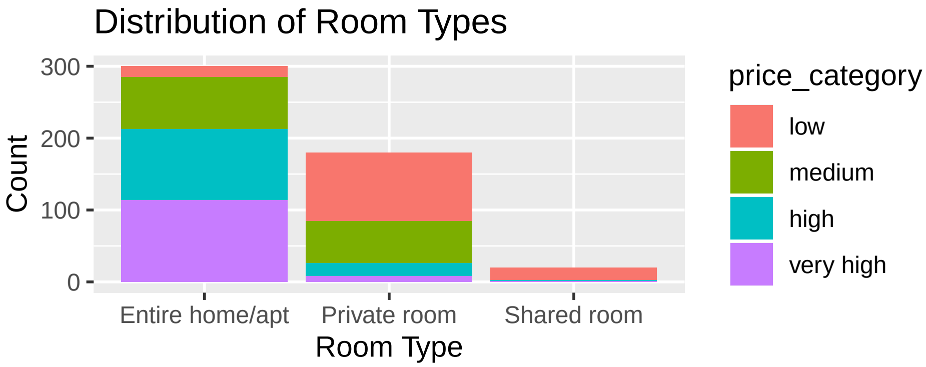
```
## [1] 61
```

```
## [1] 45
```

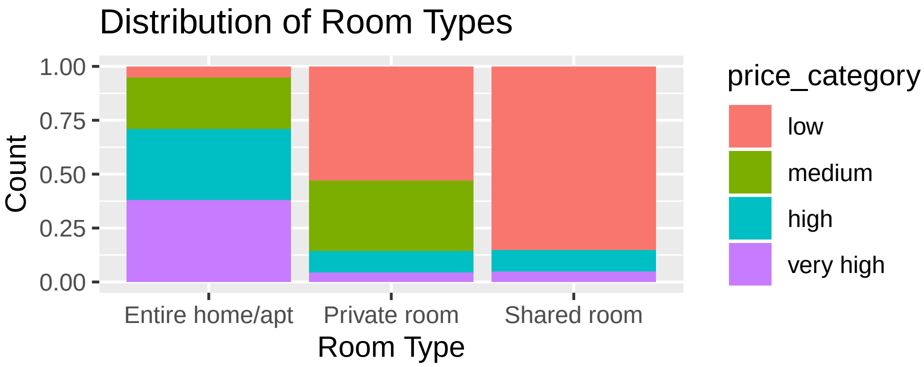
```
## [1] 3
```

```
## [1] 5
```

We can see that some variables have more values than are comfortably put into a bar graph, but we easily look at distributions for room and bed types.

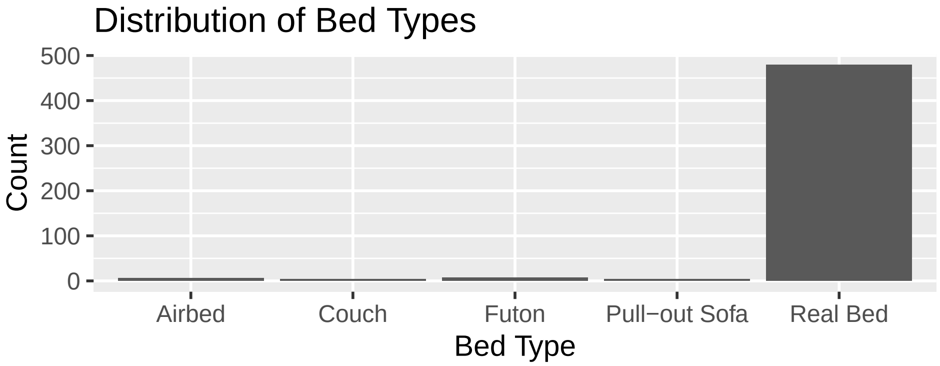


So, we can see that most rooms are an entire home or apartment, many are private rooms, and very few are shared rooms. Note that the majority of high and very high prices appear in the entire home category, unsurprisingly. Most low prices are in the private room category. While private rooms are not the least luxurious, they are much more common than shared rooms, which have a much higher proportion of low prices within its room category. In summary, higher price means more private space.



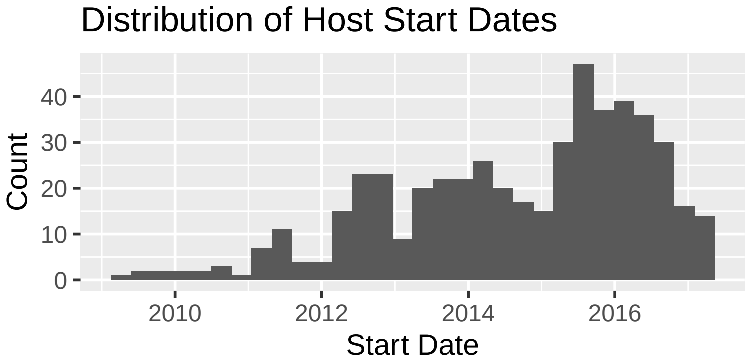
We can make this a bit clearer:

With a clearer view of proportions, we can see that very high and high are fairly similar between private and shared rooms, though low price dominates shared rooms. Between entire homes and private rooms, medium price grows with less space, while high and very high prices become less common.



The overwhelming majority of beds are real beds, with very few out of the 500 total listings in the data set being airbeds, couches, futons, or pull-out sofas.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We have a left skewed distribution, meaning in this context that more currently active hosts started more recently than not. Ie, there has likely either been a sharp rise in hosts in the area over time or hosts tend to host for a short amount of time before quitting, leading to the active hosts beginning more recently. The former is more obvious, but we would need data about inactive hosts to rule out the latter.

Host Data

Let's look at some data involving individual hosts.

```
## [1] 461
```

```
## [1] 365
```

So, it seems at least one host (almost certainly more) has multiple listings available, given that our dataset has 500 listings. Also, hosts share names.

What are some common names?

```
##
##           Joe        Sonder          John          Paul        Justin
##            10             9             7             7             6
##          Nick       Michael         Chris         David       Jessica
##             6             5             4             4             4
##         Laura          Lisa           Liz          Mark          Matt
##             4             4             4             4             4
##          Mike     The Flats        Amanda       Charles           Dan
##             4             4             3             3             3
##        Daniel      Freehand         James      Jennifer         Maria
##             3             3             3             3             3
##         Mario        Nicole         Sarah        Sharon     Stephanie
##             3             3             3             3             3
##           Tom           Ami        Andrew        Ashley   At Home Inn
##             3             2             2             2             2
##     Catherine   Christopher          Dana         Emily         Frank
##             2             2             2             2             2
##          Jane          Jeff         Jenny         Jimmy          Kari
##             2             2             2             2             2
##          Kate           Kim        Kristi          Leon       Liliana
##             2             2             2             2             2
##          Lori          Mary         Megan     Mejai Kai Melanie & Joe
##             2             2             2             2             2
##        Monica       Natalia        Pamela         Peter       Rebecca
##             2             2             2             2             2
##          Ross         Steve         Terry        Thomas        Trevor
##             2             2             2             2             2
```

Some of these are obviously common names, eg Joe, John, Paul. Some seem to be a single business with multiple properties, like “At Home Inn” and “The Flats”. Similarly, we find that our second most common host name, “Sonder”, is also such a business with a bit of research.

Let’s look at ratings by name

```
## # A tibble: 15 x 2
##    host_name     avg
##    <fct>       <dbl>
##  1 Adi            30
##  2 Alan           30
##  3 Alex           30
##  4 Alexander      30
##  5 Ali            30
##  6 Alissa         30
##  7 Amber          30
##  8 Amrit Rania    30
##  9 Amy            30
## 10 Andy           30
## 11 Anjli          30
## 12 Anna-Lisa      30
## 13 Anne           30
## 14 Anne-Marie     30
## 15 April          30
```

We can see have a lot of 30/30 for total scores, meaning many people got very good ratings as mentioned earlier. Unique, however, are low scores. Let’s look at the bottom three.

```
## # A tibble: 3 x 2
##   host_name   avg
##   <fct>     <dbl>
## 1 Wilson       22
## 2 Niki         20
## 3 Sam          18
```

Since these three scores are unique, we can use them as keys for finding more info on the users. Had we used host_id, we could have used that as well, but here we used names for the sake of readability.

We can get a quick data frame of our three hosts of interest with host relevant data.

Let's look at their ratings.

```
##   host_name review_scores_cleanliness review_scores_communication
## 1       Sam                         4                           8
## 2      Niki                         6                           8
## 3    Wilson                         6                           8
##   review_scores_value
## 1                   6
## 2                   6
## 3                   8
```

So, it seems that having good communication all around, the hosts have average to good value scores and poor to average cleanliness, their weakest factor in general. Let's see how cleanliness generally compares to other review scores across all hosts.

Now, we can look at summary stats of each rating variable

Cleanliness:

```
## Total Count:  292
## N:  292
## Minimum:  4
## Maximum:  10
## Mean:  9.477645
## Standard Deviation:  0.8622203
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100%
##    4    9   10   10   10
## IQR:  1
```

Communication:

```
## Total Count:  292
## N:  292
## Minimum:  8
## Maximum:  10
## Mean:  9.860826
## Standard Deviation:  0.3800228
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100%
##    8   10   10   10   10
## IQR:  0
```

Value:

```
## Total Count:  292
## N:  292
## Minimum:  6
## Maximum:  10
## Mean:  9.585331
## Standard Deviation:  0.6301351
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100% 
##    6    9   10   10   10 
## IQR:  1
```

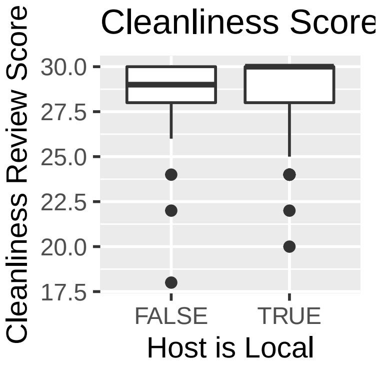
We can see that thought our base review scores don't vary too much in mean, cleanliness has the lowest mean (~9.48), the lowest minimum (4), and the highest standard deviation (~0.862). So, there does seem to be a bit of a lower lump in the tail of our distribution (which we saw in the histograms above). Let's see how many 4s we actually have.

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

Just one it seems. Though we only have 500 listings, it seems Sam's is particularly dirty.

Maybe hosts based in Chicago are more available for cleaning:

```
## Warning: Removed 78 rows containing non-finite values (stat_boxplot).
```



It appears there isn't too big of a different, though we can see our minimum cleanliness (Sam) is not a local host. Let's look at the differences in numbers:

```
## Total Count:  415
## N:  365
## Minimum:  6
## Maximum:  10
## Mean:  9.50137
## Standard Deviation:  0.783149
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100% 
##    6    9   10   10   10 
## IQR:  1
## Total Count:  85
## N:  60
## Minimum:  4
## Maximum:  10
## Mean:  9.333333
## Standard Deviation:  1.159583
## Median:  10
## Quantiles:
##   0%  25%  50%  75% 100%
##    4    9   10   10   10
## IQR:  1
```

We can see that there aren't very many non-local hosts in general, so we're pulling from a smaller sample size of them. Non-local hosts have a slightly lower mean cleanliness score, though not by much. It seems non-local hosts are generally just as good at keeping their listings clean.

Modeling

We can predict some practical values using linear regression models.
str(airbnb_clean)

```
##
## Call:
## lm(formula = price ~ bedrooms + factor(room_type) + accommodates,
##     data = airbnb_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217.57  -44.88  -16.01   18.23  695.30
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     64.461     12.471   5.169 3.43e-07 ***
## bedrooms                        52.662      8.076   6.521 1.73e-10 ***
## factor(room_type)Private room  -53.867     11.312  -4.762 2.53e-06 ***
## factor(room_type)Shared room   -71.000     24.747  -2.869  0.00429 **
## accommodates                     3.880      3.063   1.266  0.20593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.3 on 494 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.334,  Adjusted R-squared:  0.3286
## F-statistic: 61.94 on 4 and 494 DF,  p-value: < 2.2e-16
```

So here our F Statistic is 61.94 on 4 and p-value p-value: < 0.0001 indicating that our model is useful for predicting prices but the R-adjusted and R values are significantly lower 0.334 which is 33.4% of the variability. Also the slopes for the private room and shared room are negative, which makes sense as the price will be lower for those two!

```
##
## Call:
```

```
## lm(formula = price ~ factor(host_identity_verified) + cleaning_fee +
##     review_scores_value, data = airbnb_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -210.37  -39.69  -12.10   27.87  511.77
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         42.7255    59.2546   0.721    0.471
## factor(host_identity_verified)TRUE -10.2205    11.1560  -0.916    0.360
## cleaning_fee                         1.4451     0.1058  13.659   <2e-16
## review_scores_value                  2.2340     6.0749   0.368    0.713
##
## (Intercept)
## factor(host_identity_verified)TRUE
## cleaning_fee                       ***
## review_scores_value
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.76 on 323 degrees of freedom
##   (173 observations deleted due to missingness)
## Multiple R-squared:  0.3665, Adjusted R-squared:  0.3606
## F-statistic: 62.28 on 3 and 323 DF,  p-value: < 2.2e-16
```

So here our F Statistic is 62.28 on 3 and p-value p-value: < 0.0001 indicating that our model is useful for predicting prices but the R-adjusted and R values are significantly lower 0.3606 which is 36.06% of the variability. One surprising thing with this model was the slope for host being verified was negative, which doesnt make sense, but with big datsets the prediction is not always acurate.

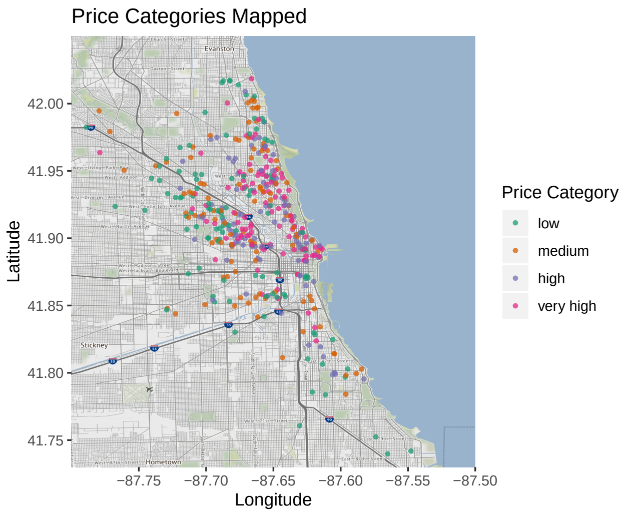
```
##
## Call:
## lm(formula = price ~ +factor(host_is_superhost) + maximum_nights +
##     bathrooms, data = airbnb_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -326.92  -50.58  -17.83   28.17  679.51
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -18.618902  14.022382  -1.328    0.185
## factor(host_is_superhost)TRUE -11.667755  11.812163  -0.988    0.324
## maximum_nights                  0.005245   0.009306   0.564    0.573
## bathrooms                     119.547288   8.766711  13.637   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107.5 on 495 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2765, Adjusted R-squared:  0.2721
## F-statistic: 63.07 on 3 and 495 DF,  p-value: < 2.2e-16
```

So here our F Statistic is 63.07 on 3 and p-value p-value: < 0.0001 indicating that our model is useful for predicting prices but the R-adjusted and R values are significantly lower 0.2765 which is 27.65% of the variability. This was the model with least Rshquare adjusted value, and again the slope fro being a superhost was negative which was a big surprise, ususally superhosts are more reliable so the price is expected to be higher.

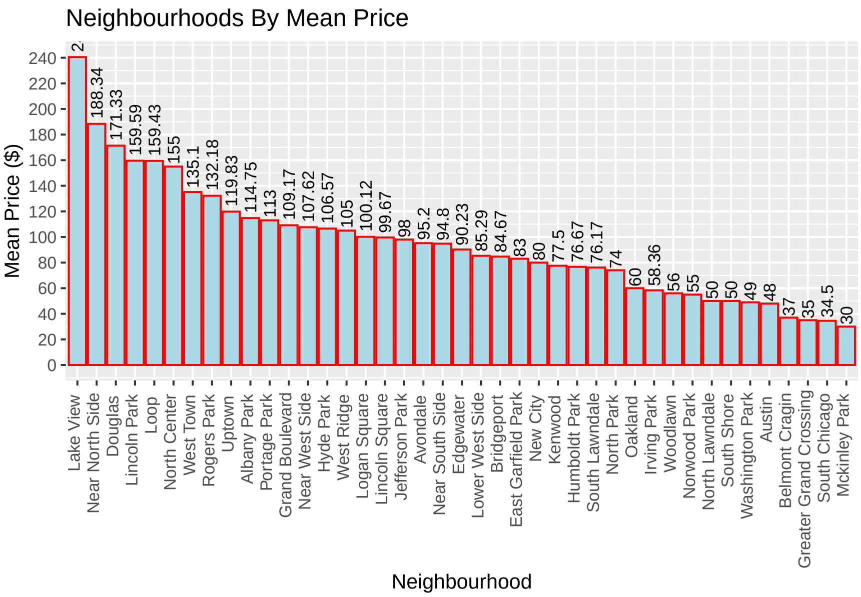
Geographical Data

The dataset provides useful location data for each listing that can provide some insight into different geographical areas.

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.

## Source : http://tile.stamen.com/terrain/12/1049/1519.png

## Source : http://tile.stamen.com/terrain/12/1050/1519.png

## Source : http://tile.stamen.com/terrain/12/1051/1519.png

## Source : http://tile.stamen.com/terrain/12/1052/1519.png

## Source : http://tile.stamen.com/terrain/12/1049/1520.png

## Source : http://tile.stamen.com/terrain/12/1050/1520.png

## Source : http://tile.stamen.com/terrain/12/1051/1520.png

## Source : http://tile.stamen.com/terrain/12/1052/1520.png

## Source : http://tile.stamen.com/terrain/12/1049/1521.png

## Source : http://tile.stamen.com/terrain/12/1050/1521.png

## Source : http://tile.stamen.com/terrain/12/1051/1521.png

## Source : http://tile.stamen.com/terrain/12/1052/1521.png

## Source : http://tile.stamen.com/terrain/12/1049/1522.png

## Source : http://tile.stamen.com/terrain/12/1050/1522.png

## Source : http://tile.stamen.com/terrain/12/1051/1522.png

## Source : http://tile.stamen.com/terrain/12/1052/1522.png

## Source : http://tile.stamen.com/terrain/12/1049/1523.png
## Source : http://tile.stamen.com/terrain/12/1050/1523.png

## Source : http://tile.stamen.com/terrain/12/1051/1523.png

## Source : http://tile.stamen.com/terrain/12/1052/1523.png

## Source : http://tile.stamen.com/terrain/12/1049/1524.png

## Source : http://tile.stamen.com/terrain/12/1050/1524.png

## Source : http://tile.stamen.com/terrain/12/1051/1524.png

## Source : http://tile.stamen.com/terrain/12/1052/1524.png

## Coordinate system already present. Adding new coordinate system, which will replace the existing one

## Warning: Removed 2 rows containing missing values (geom_point).
```

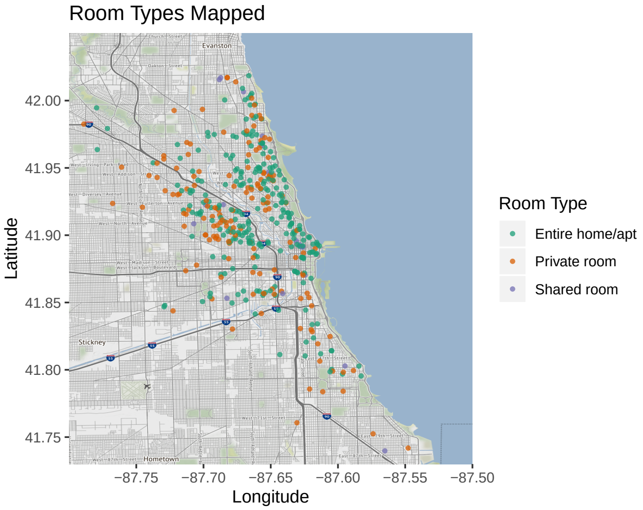


This plot shows us that the listings outside of the cities are usually cheaper than the listing inside the city, due to the high presence of orange and green dots on the outskirts of the graph, with more purple and blue as you get closer to the Loop, the heart of the city. There also seems to be a large number of purple points in the northern part of the city, towards Lake View and Lincoln Park (The area known as "Wrigleyville").

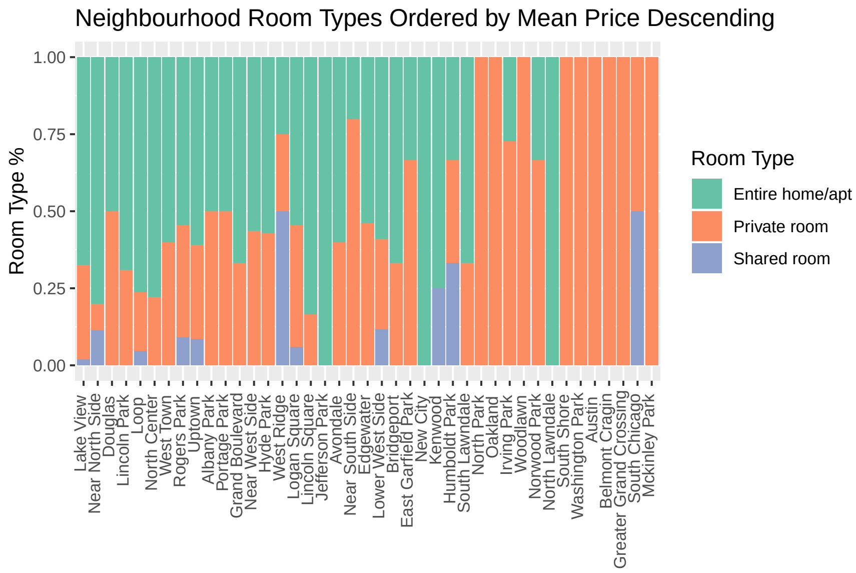


This graph tells us that our assumptions from the map were mostly true, as the neighborhoods that are closer to the center (the Loop, Near North Side,) of the city are more expensive on average, and some northern neighborhoods (Lake View, North Center, Lincoln Park) are also towards the top of this graph.

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one

## Warning: Removed 2 rows containing missing values (geom_point).
```



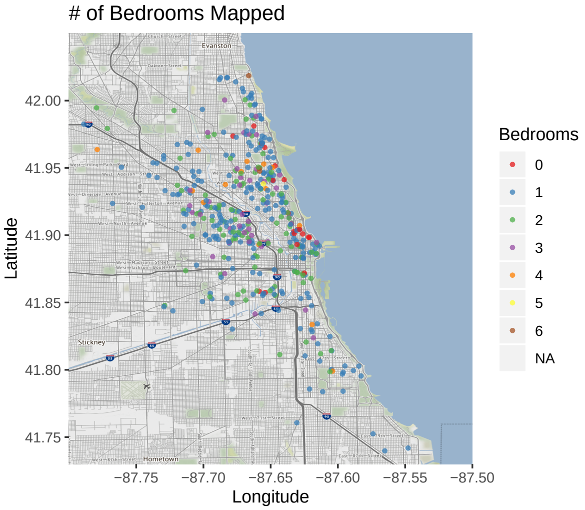
From this graph, we can assume that there are more entire units listed in the middle of the middle of the city due to the overwhelming green. This may be because the apartments in the city are smaller, and it would be harder for a guest to share that room with somebody else. We can see more orange on the northern and western parts of the city, as these are typically where single family homes are built.



This graph tells us some valuable information about the listings in the city. As the prices go up, typically, the odds of the listing being a entire apartment or house go up. This information could be useful to a potential lister in Chicago, if they want to charge more for their listing, they may have to rent out their whole unit.

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

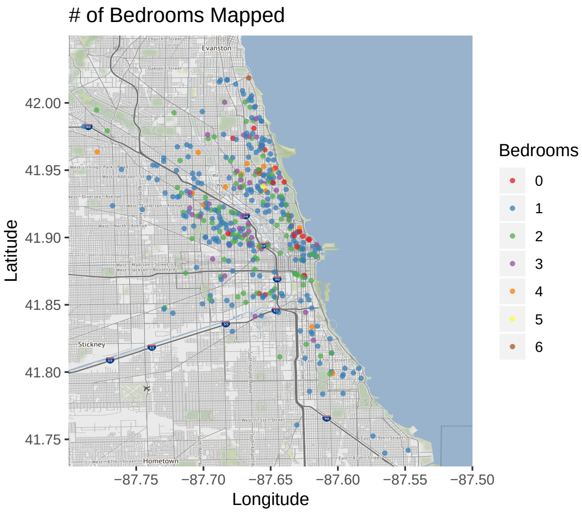


That NA value seems out of place, so we need to figure out how many there are, and if there are many, what does it mean?

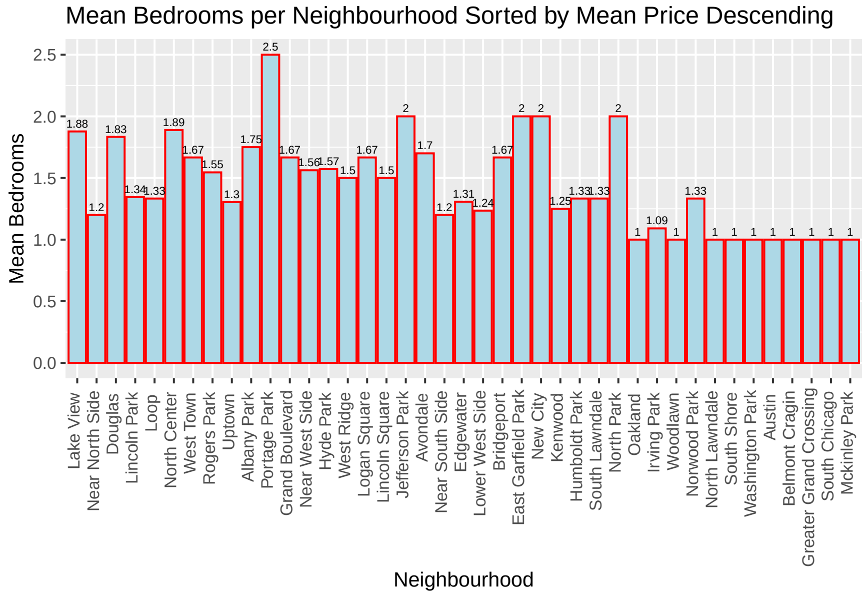
There is only one row with the NA value in bedrooms, and looking at the dataframe in rstudio, it seems that it is a small studio for which the lister put NA instead of 0. That should be switched to a zero.

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



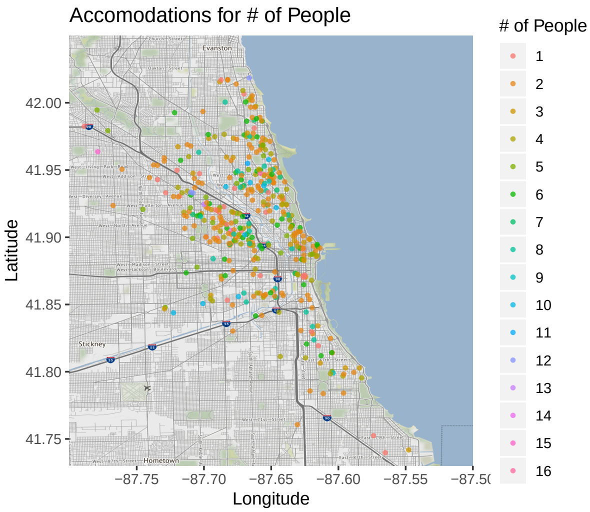
Much better. From this fixed plot we can see that it is rare for a listing in Chicago to have more than one or two bedrooms, based on the fact that there are a large number of blue, red and green points in the middle of the city. The presence of red points tell us that there are more studio apartments in the area, which would make sense considering the red points on this graph are right near the Magnificent Mile, one of the most sought after, and expensive, areas in the city.



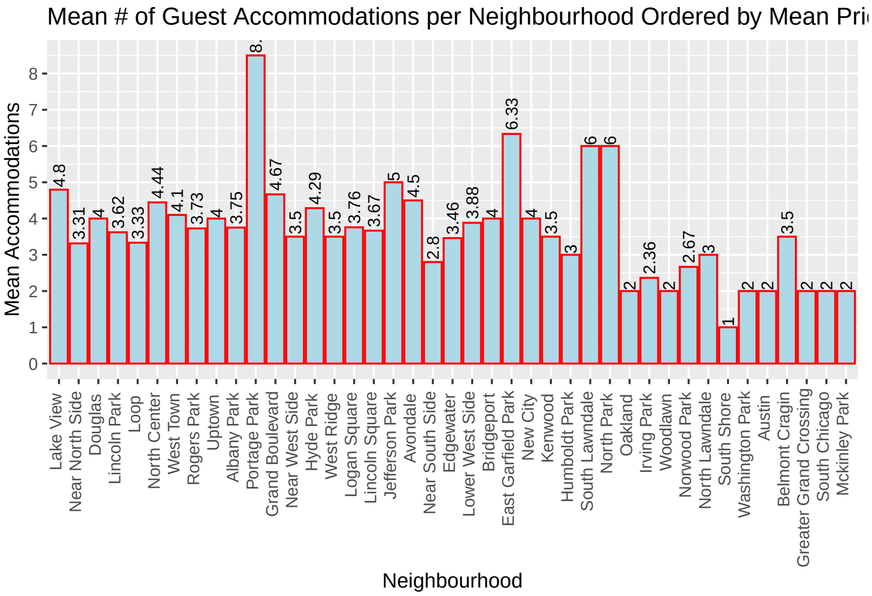
This bar chart tells us that most listings in the city lie between one and two bedrooms. This graph also seems to show less of a relationship than the room type bar graph, as it does not seem to flow in a certain direction like the other one did.

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Here we can see that the normal listing in the middle of Chicago is not meant to accommodate more that ~5 people, given that there are mainly orange, gold and gold/green points in the heart of the city. This tracks with our previous graph of bedrooms, as less bedrooms would mean less guests could stay in that listing.

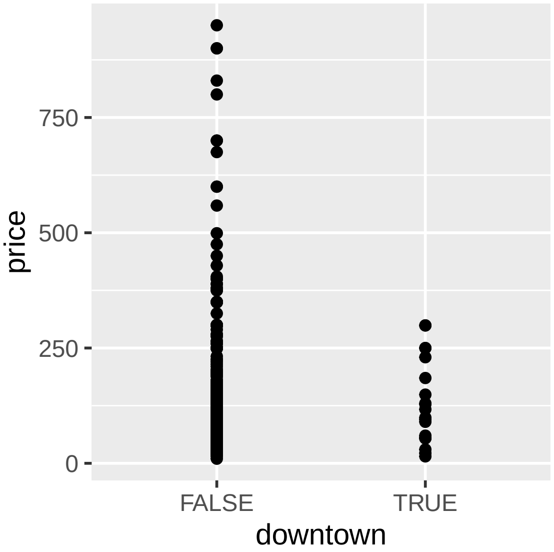


Here, we can see that neighbourhoods with higher priced listings do have some larger accommodation sizes that other less expensive neighbourhoods. This graph also tells us that most listings fit around 3-4 people, which does make sense logically, as most apartments or units typically have a 2 person bed and a couch. Many AirBnB hosts make sure that couch is a sleeper sofa, which would account for 1-2 more people.

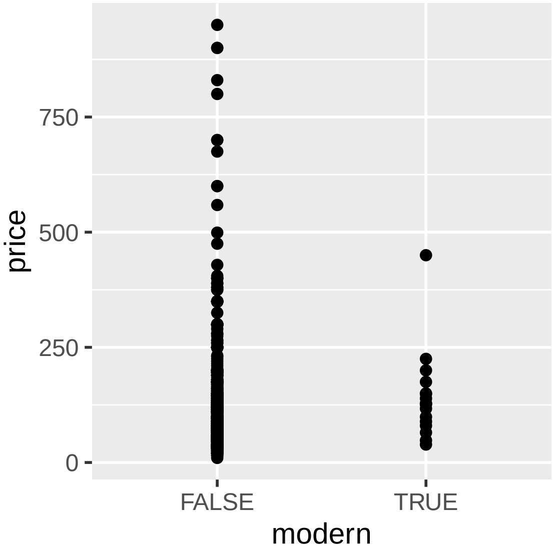
Text Data

Several of the variables here come in the form of long text descriptions from which useful information can be pulled with some effort. First, lets make some variables that can tell us the presence of a certain word in each of the titles of the entries. For most of this work we will be utilizing functinons from the stringi package

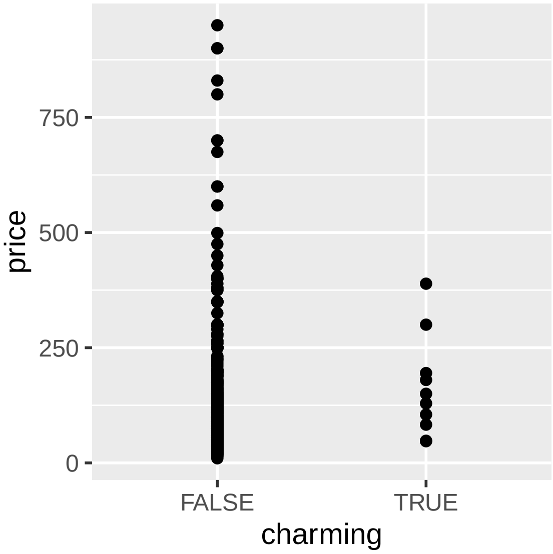
Now we will look to see if there are any cases where the price is generally higher if a certain word is included:



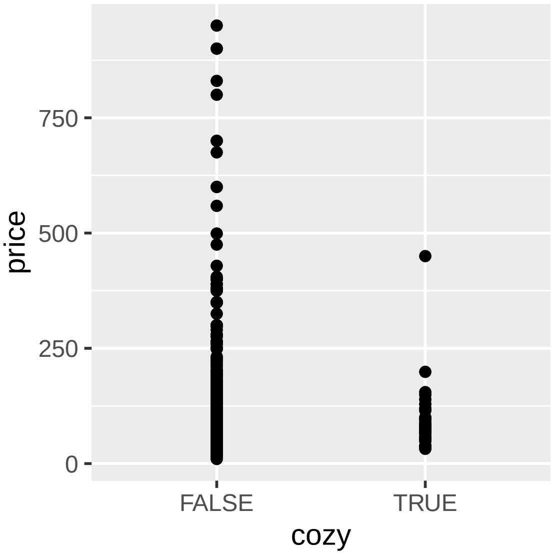
Downtown:



Modern:



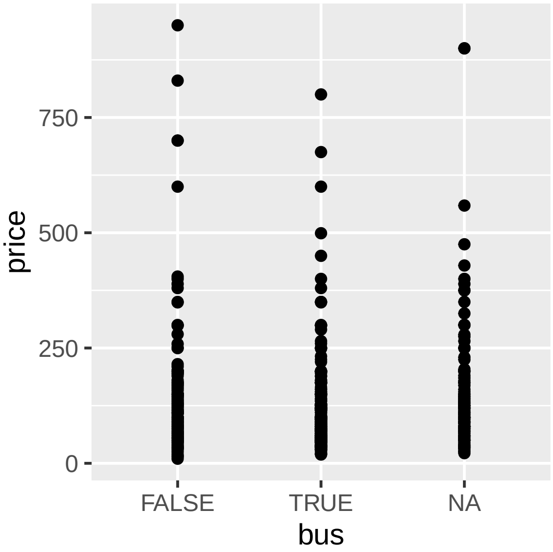
Charming:

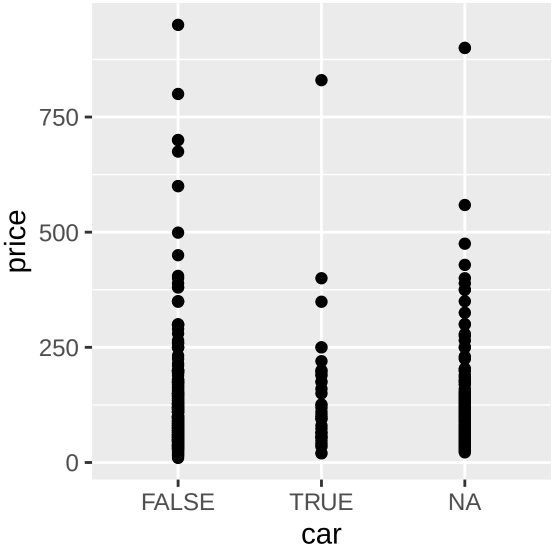


Cozy:

There are a few things we can gather from these visualizations. While it may not show that using a specific word will mean a higher priced listing, all of the data for each specific word seem to be grouped closely together. This apparent relationship may be from some common trait that is shared by all listings that have a specific word in them and we are just seeing that relationship expressed through the use of a specific word in the title. Either way, it is an interesting observation.

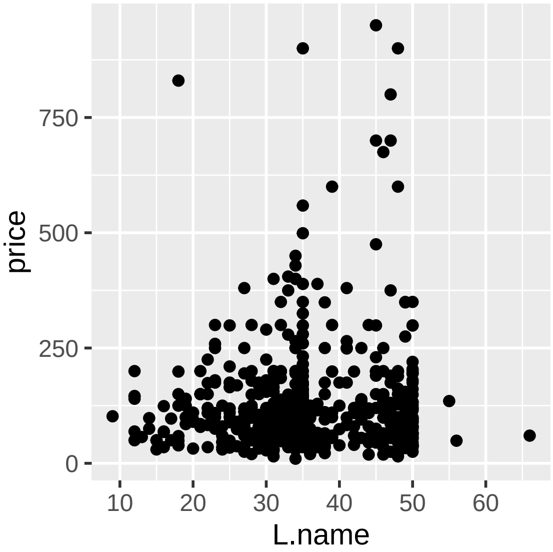
Let's see if the effect of having access to certain types of transportation show up in terms of a price change. To do this, we are going to pull out the words "car" and "bus" from the column titled transit, which briefly describes the different modes of transportation relevant to each listing.





Lets compare that to entries that contain “car” :

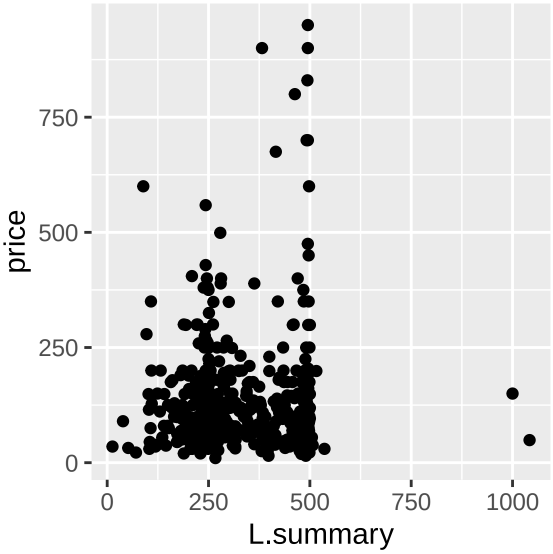
From these two charts we can see that the range of prices in the listings with bus access reach higher than the listings with car access (garage, street parking, etc.) This makes sense because the bus routes are likely closer to the “downtown” area, where we could expect to see higher prices in general.

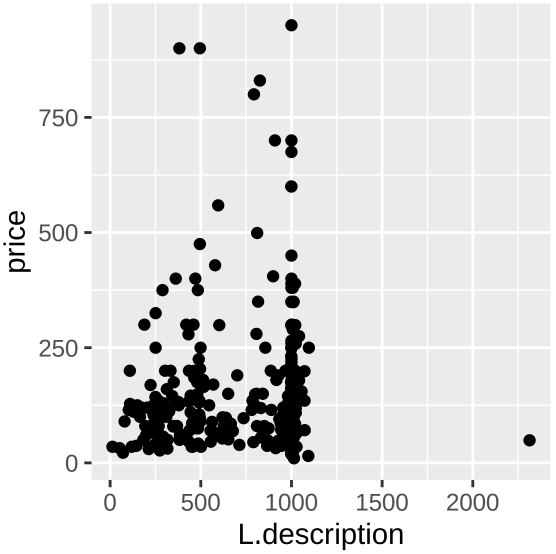


We should take a look at the length of the title and see what that can tell us:

Let's do the same thing with the summary and description:

```
## Warning: Removed 7 rows containing missing values (geom_point).
```





The only thing that can really be gained from these charts is that there seems to be some character limits on some of these entries. It looks like most summaries are around 250 and 500 characters. Most titles are around 25-35 and 45-50 characters long, and finally most descriptions are around 1000 characters long. There doesn't seem to be much of a relationship between any of these character lengths and price. This makes sense intuitively.

Conclusion

So, we were able, using our 63 starting variables about airbnb listings, gauge much valuable information regarding the hosts, the locations of listings, the text attitrbutes of listings (like description), as well as get some models for predicting variables from the data.

Individual Contributions

Kaleb: I attended all the meetings and actively participated. I did not do much with choosing the dataset because it was chosen by two group members from a class that they were both in but I asked questions and did some preliminary research on the dataset / the terms and definitions relating to AirBnB. The main section of the project I was in charge of was looking at the text and character data to gather some meaningful information. I did not do the most work out of all the group members but I think that I did a satisfactory job with the part that I was assigned.

Spencer: I helped with planning overall (ie dataset, structure, etc.), organizing, and acted as coordinator, as in I did a good amount of the github work and arranged the files. I did all the cleaning, initial exploration section, the host data section, and various smaller things here and there. I believe I contributed a fair amount to the project as a whole.

Peter: For the project, I helped come up with ideas for what we could do with the dataset in the planning stage, after we decided what dataset we would work with. My contribution for the project presentation and code was the "Geographical" section of the markdown file. This included making all the graphs, and manipulating the dataframes in that section so that they would work with the visualizations I made. I also wrote all of the blurbs below the graphs, giving insight.

Karthik: I helped the team with selecting the dataset and giving some background information about the datset and some ideas to start with as me and another teammate where using this dataset for another class, even though I didnt contribute to the team as much as the others, I showed up to the team meetings and also did the modeling part of the project, So I believe I did put in less work but sure did support the team with the best I can.