Case Study 2 EDA

Frances Hung

1/24/2020

R Markdown

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:

```
library(ggplot2)
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(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(tidyr)
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

-delete price=0 (11 observations) - impute availability__365 -make min length of stay categorical (long, med, short) -interaction with room type and neighborhood -add new variable: length of name (optional) -log reviews-per-month and price (log (1+variable)) -impute reviews-per-month as 0 if NA (corresponds to number of reviews is 0) -FOR NOW: linear minimum length of stay (may later change to categorical) -availability-0 check what this means!!

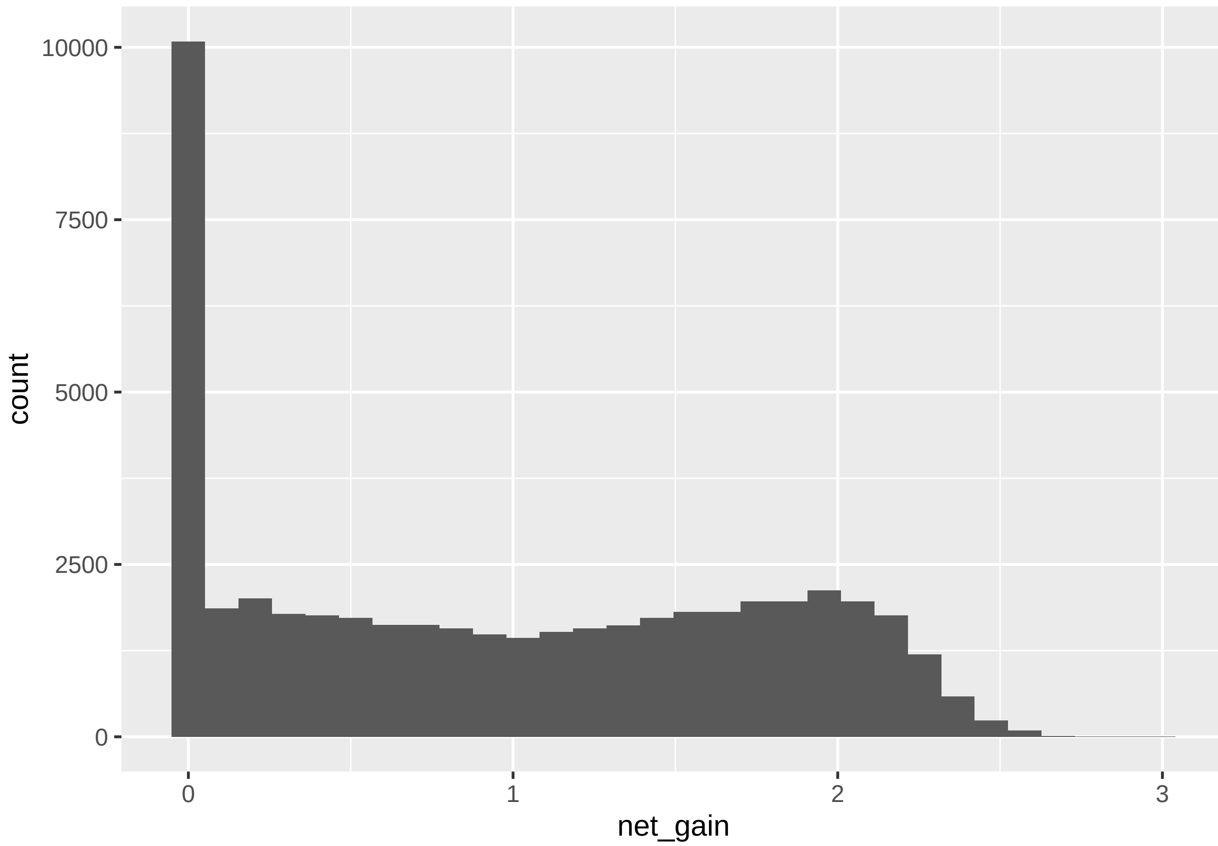
```r
AB<-read.csv("AB_NYC_2019.csv") %>%
  filter(price!=0) %>%
  mutate(reviews_per_month=replace_na(reviews_per_month,0)) %>%
  mutate(reviews_per_month=log(1+reviews_per_month),price=log(1+price))
colMeans(is.na(AB))
```

```
##                             id                           name
##                              0                              0
##                        host_id                      host_name
##                              0                              0
##            neighbourhood_group                  neighbourhood
##                              0                              0
##                       latitude                      longitude
##                              0                              0
##                      room_type                          price
##                              0                              0
##                 minimum_nights              number_of_reviews
##                              0                              0
##                    last_review              reviews_per_month
##                              0                              0
## calculated_host_listings_count               availability_365
##                              0                              0
```

```r
AB_net_gain<-AB %>%
  mutate(net_gain=log(1+price*reviews_per_month))

ggplot(AB_net_gain,aes(x=net_gain))+geom_histogram()
```

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



Linear Models)

```
m1<-lmer(price ~ (1|neighbourhood_group)+minimum_nights+reviews_per_month+room_type*neighbourhood_group
summary(m1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## price ~ (1 | neighbourhood_group) + minimum_nights + reviews_per_month +
##     room_type * neighbourhood_group
##    Data: AB
##
## REML criterion at convergence: 73051.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.7305 -0.6186 -0.0811  0.4739 10.0948
##
## Random effects:
##  Groups              Name        Variance Std.Dev.
##  neighbourhood_group (Intercept) 0.0366   0.1913
##  Residual                        0.2604   0.5103
## Number of obs: 48884, groups:  neighbourhood_group, 5
##
## Fixed effects:
##                                                           Estimate Std. Error
```

```
## (Intercept)                                          4.7388206  0.1931241
## minimum_nights                                      -0.0010405  0.0001143
## reviews_per_month                                   -0.0348447  0.0040771
## room_typePrivate room                               -0.6650759  0.0329790
## room_typeShared room                                -0.9355682  0.0709135
## neighbourhood_groupBrooklyn                          0.3020587  0.2718641
## neighbourhood_groupManhattan                         0.5978891  0.2718528
## neighbourhood_groupQueens                            0.1372216  0.2720408
## neighbourhood_groupStaten Island                     0.0895348  0.2745202
## room_typePrivate room:neighbourhood_groupBrooklyn   -0.1505027  0.0337668
## room_typeShared room:neighbourhood_groupBrooklyn    -0.3362166  0.0754221
## room_typePrivate room:neighbourhood_groupManhattan  -0.0946107  0.0337728
## room_typeShared room:neighbourhood_groupManhattan   -0.0981656  0.0747860
## room_typePrivate room:neighbourhood_groupQueens     -0.0463896  0.0359001
## room_typeShared room:neighbourhood_groupQueens      -0.1231206  0.0804164
## room_typePrivate room:neighbourhood_groupStaten Island -0.0986759  0.0628627
## room_typeShared room:neighbourhood_groupStaten Island  -0.0799930  0.1882560
##                                                     t value
## (Intercept)                                          24.538
## minimum_nights                                       -9.106
## reviews_per_month                                    -8.546
## room_typePrivate room                               -20.167
## room_typeShared room                                -13.193
## neighbourhood_groupBrooklyn                           1.111
## neighbourhood_groupManhattan                          2.199
## neighbourhood_groupQueens                             0.504
## neighbourhood_groupStaten Island                      0.326
## room_typePrivate room:neighbourhood_groupBrooklyn    -4.457
## room_typeShared room:neighbourhood_groupBrooklyn     -4.458
## room_typePrivate room:neighbourhood_groupManhattan   -2.801
## room_typeShared room:neighbourhood_groupManhattan    -1.313
## room_typePrivate room:neighbourhood_groupQueens      -1.292
## room_typeShared room:neighbourhood_groupQueens       -1.531
## room_typePrivate room:neighbourhood_groupStaten Island  -1.570
## room_typeShared room:neighbourhood_groupStaten Island   -0.425
```

```
##
## Correlation matrix not shown by default, as p = 17 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```r
AIC(m1)
```

```
## [1] 73089.23
```

From this model, we can infer that price is negatively correlated with reviews per month, whether the room is private or shared (especially in Brooklyn), and increased minimum nights. This fits the narrative that the most expensive rentals tend to be whole-apartment/house rentals catering to wealthy short-term vacationers. The neighbourhood groups/neighbourhoods don't seem to have that much variance in base price as an intercept (probably only a very select few do).

```r
m2<-lm(reviews_per_month ~ minimum_nights+price+room_type+neighbourhood_group,data=AB)
summary(m2)
```

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

```
## lm(formula = reviews_per_month ~ minimum_nights + price + room_type +
##     neighbourhood_group, data = AB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8196 -0.4659 -0.2080  0.3861  4.4745
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.9320080  0.0297648  31.312  < 2e-16 ***
## minimum_nights                  -0.0039843  0.0001258 -31.684  < 2e-16 ***
## price                           -0.0414355  0.0050205  -8.253  < 2e-16 ***
## room_typePrivate room           -0.0372180  0.0065739  -5.662 1.51e-08 ***
## room_typeShared room            -0.0685971  0.0179858  -3.814 0.000137 ***
## neighbourhood_groupBrooklyn     -0.1610111  0.0176953  -9.099  < 2e-16 ***
## neighbourhood_groupManhattan    -0.1810977  0.0178768 -10.130  < 2e-16 ***
## neighbourhood_groupQueens       -0.0039847  0.0187751  -0.212 0.831925
## neighbourhood_groupStaten Island  0.0568551  0.0340424   1.670 0.094901 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5674 on 48875 degrees of freedom
## Multiple R-squared:  0.0358, Adjusted R-squared:  0.03564
## F-statistic: 226.8 on 8 and 48875 DF,  p-value: < 2.2e-16
```

```r
AIC(m2)
```

```
## [1] 83325.12
```

```r
confint(m2)
```

```
##                                        2.5 %       97.5 %
## (Intercept)                      0.873668551  0.990347379
## minimum_nights                  -0.004230815 -0.003737867
## price                           -0.051275629 -0.031595302
## room_typePrivate room           -0.050102876 -0.024333144
## room_typeShared room            -0.103849575 -0.033344607
## neighbourhood_groupBrooklyn     -0.195694131 -0.126328099
## neighbourhood_groupManhattan    -0.216136450 -0.146058998
## neighbourhood_groupQueens       -0.040784135  0.032814711
## neighbourhood_groupStaten Island -0.009868434  0.123578715
```

From a naive lm model with no random effects, we infer that popularity (reviews per month) is negatively correlated with minimum nights and price. The other variables are not well estimated, probably because the popularity between districts varies enough to make a difference.

```r
m4<-lmer(reviews_per_month ~ minimum_nights+price+(1|neighbourhood)+room_type,data=AB,REML=FALSE)
summary(m4)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: reviews_per_month ~ minimum_nights + price + (1 | neighbourhood) +
##     room_type
##    Data: AB
##
##      AIC      BIC   logLik deviance df.resid
##  81815.8  81877.4 -40900.9  81801.8    48877
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4865 -0.7867 -0.3507  0.6690  7.6080
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  neighbourhood (Intercept) 0.03404  0.1845
##  Residual                  0.30937  0.5562
## Number of obs: 48884, groups:  neighbourhood, 221
##
## Fixed effects:
##                        Estimate Std. Error t value
## (Intercept)            0.763044   0.029954  25.473
## minimum_nights        -0.003750   0.000124 -30.244
## price                 -0.018578   0.005238  -3.547
## room_typePrivate room -0.037878   0.006521  -5.809
## room_typeShared room  -0.076617   0.017838  -4.295
##
## Correlation of Fixed Effects:
##             (Intr) mnmm_n price  rm_tPr
## minmm_nghts -0.074
## price       -0.865  0.052
## rm_typPrvtr -0.566  0.074  0.567
## rm_typShrdr -0.297  0.026  0.306  0.298
```

```r
AIC(m4)
```

```
## [1] 81815.83
```

This one-level hierarchical model is better as it takes into account differences across neighborhoods. I tried doing a 2 level with neighbourhood and neighborhood group but the model did not converge.

```r
m3<-lmer(reviews_per_month ~ minimum_nights+price+(1|neighbourhood)+(1|room_type),data=AB,REML=FALSE)
summary(m3)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: reviews_per_month ~ minimum_nights + price + (1 | neighbourhood) +
##     (1 | room_type)
##    Data: AB
##
##      AIC      BIC   logLik deviance df.resid
##  81823.4  81876.2 -40905.7  81811.4    48878
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4855 -0.7866 -0.3505  0.6690  7.6037
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  neighbourhood (Intercept) 0.034205 0.18495
##  room_type     (Intercept) 0.000905 0.03008
##  Residual                  0.309385 0.55622
## Number of obs: 48884, groups:  neighbourhood, 221; room_type, 3
##
## Fixed effects:
##                 Estimate Std. Error t value
```

```
## (Intercept)     0.723185   0.032625  22.167
## minimum_nights -0.003748   0.000124 -30.227
## price          -0.017515   0.005198  -3.369
##
## Correlation of Fixed Effects:
##             (Intr) mnmm_n
## minmm_nghts -0.058
## price       -0.701  0.051
```

```r
AIC(m3)
```

```
## [1] 81823.43
```

Linear models predicting net gain

```r
m4<-lmer(net_gain ~ minimum_nights+price+reviews_per_month+(1|neighbourhood)+(1|room_type),data=AB_net_g
summary(m4)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: net_gain ~ minimum_nights + price + reviews_per_month + (1 |
##     neighbourhood) + (1 | room_type)
##    Data: AB_net_gain
##
##      AIC      BIC   logLik deviance df.resid
##  -1892.7  -1831.2    953.4  -1906.7    48877
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -10.0897  -0.8384   0.1053   0.8931   3.0823
##
## Random effects:
##  Groups        Name        Variance  Std.Dev.
##  neighbourhood (Intercept) 0.0024565 0.04956
##  room_type     (Intercept) 0.0006614 0.02572
##  Residual                  0.0559575 0.23655
## Number of obs: 48884, groups:  neighbourhood, 221; room_type, 3
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)       -9.614e-02  1.845e-02  -5.210
## minimum_nights    -4.886e-04  5.319e-05  -9.185
## price              7.190e-02  2.211e-03  32.520
## reviews_per_month  1.291e+00  1.920e-03 672.522
##
## Correlation of Fixed Effects:
##             (Intr) mnmm_n price
## minmm_nghts -0.053
## price       -0.527  0.052
## rvws_pr_mnt -0.076  0.136  0.019
```

```r
AIC(m4)
```

```
## [1] -1892.744
```

We have a very small AIC for this model; it shows that the net gain (price times popularity) is negatively correlated to minimum nights and positively correlated to reviews per month. We have to keep in mind the dependent variable is only an accurate representation for fitting short-term rental profit (a long-term Airbnb may make a lot off of one person staying there for a year, and who only posts 1 review).

```
m4<-lmer(net_gain ~ minimum_nights+price+reviews_per_month+(1|neighbourhood)+neighbourhood_group+room_ty
summary(m4)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: net_gain ~ minimum_nights + price + reviews_per_month + (1 |
##     neighbourhood) + neighbourhood_group + room_type
##    Data: AB_net_gain
##
##      AIC      BIC   logLik deviance df.resid
##  -1919.8  -1814.2    971.9  -1943.8    48872
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -10.1025  -0.8390   0.1058   0.8930   3.0899
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  neighbourhood (Intercept) 0.001991 0.04463
##  Residual                  0.055957 0.23655
## Number of obs: 48884, groups:  neighbourhood, 221
##
## Fixed effects:
##                                    Estimate Std. Error t value
## (Intercept)                      -0.0666599  0.0149092  -4.471
## minimum_nights                   -0.0004907  0.0000532  -9.223
## price                             0.0714018  0.0022264  32.071
## reviews_per_month                 1.2912737  0.0019213 672.071
## neighbourhood_groupBrooklyn       0.0316945  0.0127844   2.479
## neighbourhood_groupManhattan      0.0071828  0.0134109   0.536
## neighbourhood_groupQueens        -0.0227330  0.0131527  -1.728
## neighbourhood_groupStaten Island  0.0185647  0.0185264   1.002
## room_typePrivate room            -0.0405487  0.0027725 -14.626
## room_typeShared room             -0.0624836  0.0075734  -8.250
##
## Correlation of Fixed Effects:
##             (Intr) mnmm_n price  rvws__ nghb_B nghb_M nghb_Q ngh_SI rm_tPr
## minmm_nghts -0.073
## price       -0.710  0.054
## rvws_pr_mnt -0.106  0.136  0.015
## nghbrhd_grB -0.535 -0.002 -0.037  0.022
## nghbrhd_grM -0.467 -0.015 -0.098  0.032  0.624
## nghbrhd_grQ -0.528  0.001 -0.016 -0.008  0.630  0.602
## nghbrhd_gSI -0.382  0.001 -0.004 -0.005  0.448  0.427  0.435
## rm_typPrvtr -0.484  0.077  0.568  0.026  0.004 -0.021 -0.006  0.010
## rm_typShrdr -0.259  0.028  0.307  0.019  0.007 -0.004  0.005  0.013  0.299
```

```
AIC(m4)
```

```
## [1] -1919.792
```

The only category with missing data is the reviews per month variable. There doesn't seem to be an obvious pattern to the missingness; the neighborhoods with more missing data are the neighborhoods which have more listings AND there is no missing data if the date of the last review is recent (i.e. more than 2018). We're interested in current trends anyways, so we can get rid of data where the last review is before 2018.

```
# AB %>% filter(reviews_per_month %>% is.na()) %>% group_by(neighbourhood) %>% summarise(count=n(),med_

AB %>% group_by(neighbourhood) %>% summarise(count=n(),med_price=median(price)) %>% arrange(desc(count)
```

```
## # A tibble: 221 x 3
##    neighbourhood      count med_price
##    <fct>              <int>     <dbl>
##  1 Williamsburg        3919      4.66
##  2 Bedford-Stuyvesant  3710      4.39
##  3 Harlem              2658      4.50
##  4 Bushwick            2462      4.19
##  5 Upper West Side     1971      5.02
##  6 Hell's Kitchen      1958      5.13
##  7 East Village        1853      5.02
##  8 Upper East Side     1798      5.01
##  9 Crown Heights       1564      4.45
## 10 Midtown             1545      5.35
## # ... with 211 more rows
```

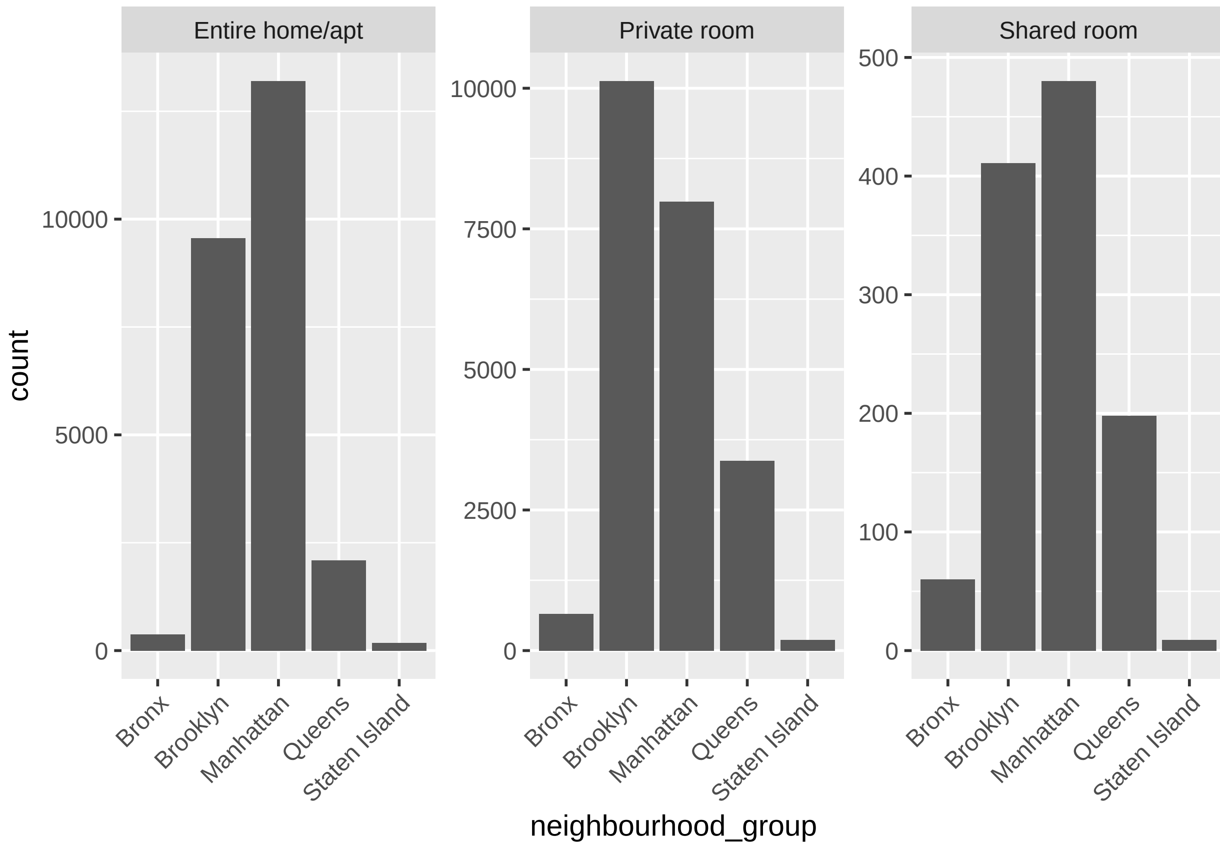
```
# AB %>%
#   group_by(neighbourhood) %>%
#   summarise(n = count(is.na(reviews_per_month))) %>%
#   mutate(freq = n / sum(n))

# AB %>% group_by(neighbourhood) %>% group_by(neighbourhood) %>% summarise(y=count(is.na(reviews_per_mo
```

Including Plots

```
ggplot(AB,aes(x=neighbourhood_group))+
  geom_histogram(stat = "count") +
  facet_wrap(.~room_type,scale="free") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

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




```
AB_w_couples<-AB %>% mutate(couples=ifelse(grepl("and |AND |And | \\& ",host_name),1,0))
```

Leased entire houses/apartments are the most common room type Airbnb offers in Manhattan, while in Brooklyn where living space tends to be larger, private rooms are also common offer. Queens offers mostly private rooms.

The median price of housing listed under couples is about the same as those listed under singles.

```
AB_w_couples %>%
  group_by(neighbourhood) %>%
  summarise(median_price=median(price), q25=quantile(price,.25),q75=quantile(price,.75),count=n()) %>%
  arrange(desc(median_price)) %>%
  filter(count>50)
```

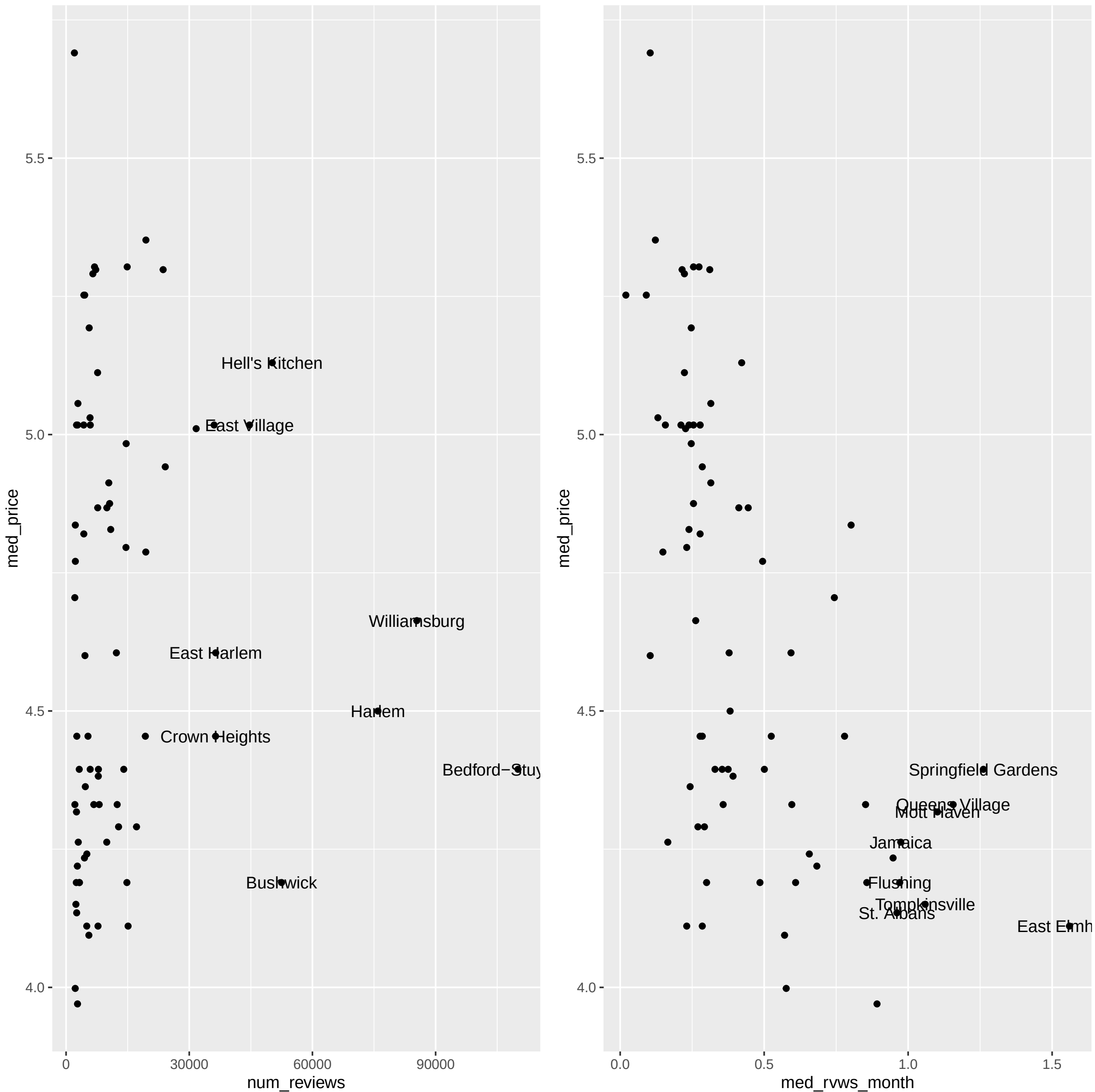
```
## # A tibble: 96 x 5
##    neighbourhood      median_price   q25   q75 count
##    <fct>                     <dbl> <dbl> <dbl> <int>
##  1 Tribeca                    5.69  5.30  6.19   177
##  2 NoHo                       5.53  5.19  5.86    78
##  3 Flatiron District          5.42  5.07  5.94    80
##  4 Midtown                    5.35  4.98  5.83  1545
##  5 Financial District         5.30  4.98  5.52   744
##  6 West Village               5.30  5.04  5.62   768
##  7 Chelsea                    5.30  4.88  5.60  1113
##  8 SoHo                       5.30  4.84  5.83   358
##  9 Greenwich Village          5.29  4.94  5.53   392
## 10 Battery Park City          5.28  4.62  5.56    70
## # ... with 86 more rows
```

```r
most_pop_neighborhoods<-AB %>% drop_na() %>% group_by(neighbourhood) %>%
  summarise(num_reviews=sum(number_of_reviews),med_price=median(price),med_rvws_month=median(reviews_pe
  filter(num_reviews>2000)

total_rvws_plot<-most_pop_neighborhoods %>% ggplot(aes(x=num_reviews,y=med_price))+geom_point()+
  geom_text(data=subset(most_pop_neighborhoods, num_reviews>quantile(num_reviews,.9) | med_price>150),a

per_month_rvws_plot<-most_pop_neighborhoods %>% ggplot(aes(x=med_rvws_month,y=med_price))+geom_point()+
  geom_text(data=subset(most_pop_neighborhoods, med_rvws_month>quantile(med_rvws_month,.9) | med_price>

grid.arrange(total_rvws_plot,per_month_rvws_plot,ncol=2)
```



```r
most_pop_neighborhoods %>% filter(num_reviews>quantile(num_reviews,.9))
```

```
## # A tibble: 8 x 6
##   neighbourhood      num_reviews med_price med_rvws_month district  available
##   <fct>                    <int>     <dbl>          <dbl> <fct>         <dbl>
## 1 Bedford-Stuyvesant      110068      4.39          0.501 Brooklyn       59
## 2 Bushwick                 52491      4.19          0.300 Brooklyn       19
## 3 Crown Heights            36408      4.45          0.285 Brooklyn       20
## 4 East Harlem              36446      4.61          0.593 Manhattan      36
## 5 East Village             44670      5.02          0.255 Manhattan       3
## 6 Harlem                   75962      4.50          0.382 Manhattan      43
## 7 Hell's Kitchen           50227      5.13          0.422 Manhattan      93.5
## 8 Williamsburg             85424      4.66          0.262 Brooklyn        3
```

```r
most_pop_neighborhoods %>% filter(med_rvws_month>quantile(med_rvws_month,.9))
```

```
## # A tibble: 8 x 6
##   neighbourhood      num_reviews med_price med_rvws_month district     available
##   <fct>                    <int>     <dbl>          <dbl> <fct>            <dbl>
## 1 East Elmhurst            15107      4.11          1.56  Queens           150
## 2 Flushing                 14818      4.19          0.971 Queens           136.
## 3 Jamaica                   9910      4.26          0.975 Queens           175
## 4 Mott Haven                2542      4.32          1.10  Bronx             96.5
## 5 Queens Village            2147      4.33          1.16  Queens           138.
## 6 Springfield Garde~        5873      4.39          1.26  Queens           179
## 7 St. Albans                2584      4.13          0.961 Queens           260.
## 8 Tompkinsville             2400      4.15          1.06  Staten Isla~     242
```

There seems to be a correlation between number of reviews per month and number of reviews, but it is not absolute. Perhaps the reviews per month is more indicative of up-and-coming neighborhoods than the total number (which may include Airbnbs which have been on the market for a long time). Looking at the total number of reviews versus median reviews per month, we can see that we have expensive rentals with relatively low numbers of reviews; these also unsurprisingly correspond to low numbers of reviews per month.

Things get interesting when we look at the neighborhoods with most total number of reviews (mostly in Brooklyn and Manhattan) and neighborhoods with the most reviews per month (mostly in Queens).

Manhattan/Brooklyn has quite a few renters who usually have available full-apartment space to rent for two or three months every year; we'd assume that they are likely people who rent out the spaces they live in while they're on vacation. Queens has quite a few renters who are renting private rooms or full apartments for a much larger portion of the year for cheaper; they probably have designated rooms for renting out. now does days available correspond to types of rooms? maybe a better profit metric is dollars per review per day available.

This is a hierarchical model: important metrics seem to be neighborhood_group, possibly the metric described above,