PCA Midterm project

Diana Rueda

5/3/2021

Loading the data and displaying the head

```
library(readxl)
data <- read_excel("~/Documents/group5/data.xlsx", range = "A1:P32810", na = "NA")
head(data)
```

```
## # A tibble: 6 x 16
##      id name          host_id host_name  neighbourhood_g~ neighbourhood latitude
##   <dbl> <chr>           <dbl> <chr>      <chr>            <chr>            <dbl>
## 1  2595 Skylit Midto~    2845 Jennifer   Manhattan        Midtown           40.8
## 2  3831 Cozy Entire ~    4869 LisaRoxan~ Brooklyn         Clinton Hill      40.7
## 3  5099 Large Cozy 1~    7322 Chris      Manhattan        Murray Hill       40.7
## 4  5178 Large Furnis~    8967 Shunichi   Manhattan        Hell's Kitch~     40.8
## 5  5238 Cute & Cozy ~    7549 Ben        Manhattan        Chinatown         40.7
## 6  5295 Beautiful 1b~    7702 Lena       Manhattan        Upper West S~     40.8
## # ... with 9 more variables: longitude <dbl>, room_type <chr>, price <dbl>,
## #   minimum_nights <dbl>, number_of_reviews <dbl>, last_review <dttm>,
## #   reviews_per_month <dbl>, calculated_host_listings_count <dbl>,
## #   availability_365 <dbl>
```

Dropping unessesary variables for the purpose of this study

```
#remove useless variables
drop = c( "id", "name", "host_id", "host_name", "latitude", "longitude", "last_review","calculated_host_
data = data[,!(names(data) %in% drop)]
head(data)
```

```
## # A tibble: 6 x 8
##   neighbourhood_g~ neighbourhood room_type price minimum_nights number_of_revie~
##   <chr>            <chr>         <chr>     <dbl>          <dbl>            <dbl>
## 1 Manhattan        Midtown       Entire h~   225              1               45
## 2 Brooklyn         Clinton Hill  Entire h~    89              1              270
## 3 Manhattan        Murray Hill   Entire h~   200              3               74
## 4 Manhattan        Hell's Kitch~ Private ~    79              2              430
## 5 Manhattan        Chinatown     Entire h~   150              1              160
## 6 Manhattan        Upper West S~ Entire h~   135              5               53
## # ... with 2 more variables: reviews_per_month <dbl>, availability_365 <dbl>
```

```
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
#remove dependent variable
#df_no_price <- data[,!(names(data) %in% c("price"))]
```

```
df_no_price <- data
head(df_no_price)
```

```
## # A tibble: 6 x 8
##   neighbourhood_g~ neighbourhood room_type price minimum_nights number_of_revie~
##   <chr>            <chr>         <chr>     <dbl>          <dbl>            <dbl>
## 1 Manhattan        Midtown       Entire h~   225              1               45
## 2 Brooklyn         Clinton Hill  Entire h~    89              1              270
## 3 Manhattan        Murray Hill   Entire h~   200              3               74
## 4 Manhattan        Hell's Kitch~ Private ~    79              2              430
## 5 Manhattan        Chinatown     Entire h~   150              1              160
## 6 Manhattan        Upper West S~ Entire h~   135              5               53
## # ... with 2 more variables: reviews_per_month <dbl>, availability_365 <dbl>
```

```
unique(df_no_price["neighbourhood"])
```

```
## # A tibble: 218 x 1
##    neighbourhood
##    <chr>
##  1 Midtown
##  2 Clinton Hill
##  3 Murray Hill
##  4 Hell's Kitchen
##  5 Chinatown
##  6 Upper West Side
##  7 South Slope
##  8 Williamsburg
##  9 Fort Greene
## 10 Chelsea
## # ... with 208 more rows
```

```
unique(df_no_price["neighbourhood_group"])
```

```
## # A tibble: 5 x 1
##   neighbourhood_group
##   <chr>
## 1 Manhattan
## 2 Brooklyn
## 3 Queens
## 4 Staten Island
## 5 Bronx
```

```
library(fastDummies)
```

```
df_dummies <- dummy_cols(df_no_price, select_columns = c("neighbourhood_group", "room_type"))
head(df_dummies)
```

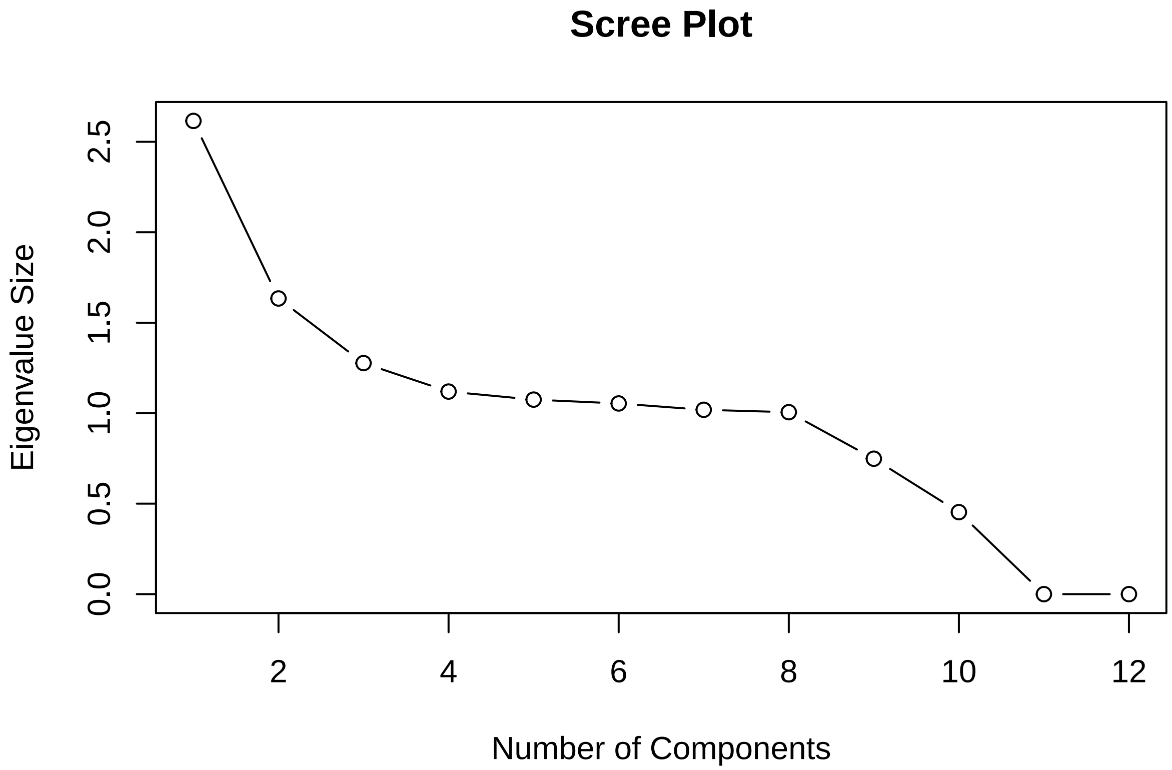
```
## # A tibble: 6 x 16
##   neighbourhood_g~ neighbourhood room_type price minimum_nights number_of_revie~
##   <chr>            <chr>         <chr>     <dbl>          <dbl>            <dbl>
## 1 Manhattan        Midtown       Entire h~   225              1               45
## 2 Brooklyn         Clinton Hill  Entire h~    89              1              270
## 3 Manhattan        Murray Hill   Entire h~   200              3               74
## 4 Manhattan        Hell's Kitch~ Private ~    79              2              430
## 5 Manhattan        Chinatown     Entire h~   150              1              160
## 6 Manhattan        Upper West S~ Entire h~   135              5                53
## # ... with 10 more variables: reviews_per_month <dbl>, availability_365 <dbl>,
## #   neighbourhood_group_Bronx <int>, neighbourhood_group_Brooklyn <int>,
## #   neighbourhood_group_Manhattan <int>, neighbourhood_group_Queens <int>,
## #   neighbourhood_group_Staten Island <int>, room_type_Entire home/apt <int>,
## #   room_type_Private room <int>, room_type_Shared room <int>
```

```
#remove useless variables
drop = c( "neighbourhood_group", "neighbourhood", "room_type", "reviews_per_month")
df_dummies = df_dummies[,!(names(df_dummies) %in% drop)]
head(df_dummies)
```

```
## # A tibble: 6 x 12
##   price minimum_nights number_of_reviews availability_365 neighbourhood_group_B~
##   <dbl>          <dbl>             <dbl>            <dbl>                  <int>
## 1   225              1                45              355                      0
## 2    89              1               270              194                      0
## 3   200              3                74              129                      0
## 4    79              2               430              220                      0
## 5   150              1               160              188                      0
## 6   135              5                53                6                      0
## # ... with 7 more variables: neighbourhood_group_Brooklyn <int>,
## #   neighbourhood_group_Manhattan <int>, neighbourhood_group_Queens <int>,
## #   neighbourhood_group_Staten Island <int>, room_type_Entire home/apt <int>,
## #   room_type_Private room <int>, room_type_Shared room <int>
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.6172 1.2783 1.1300 1.05804 1.03681 1.02653 1.00931
## Proportion of Variance 0.2179 0.1362 0.1064 0.09329 0.08958 0.08781 0.08489
## Cumulative Proportion  0.2179 0.3541 0.4605 0.55380 0.64338 0.73119 0.81609
##                           PC8     PC9    PC10      PC11      PC12
## Standard deviation     1.0028 0.86506 0.67311 4.171e-14 6.135e-15
## Proportion of Variance 0.0838 0.06236 0.03776 0.000e+00 0.000e+00
## Cumulative Proportion  0.8999 0.96224 1.00000 1.000e+00 1.000e+00
```

```
# A scree plot:
plot(1:(length(data_pc$sdev)), (data_pc$sdev)^2, type='b',
     main="Scree Plot", xlab="Number of Components", ylab="Eigenvalue Size")
```



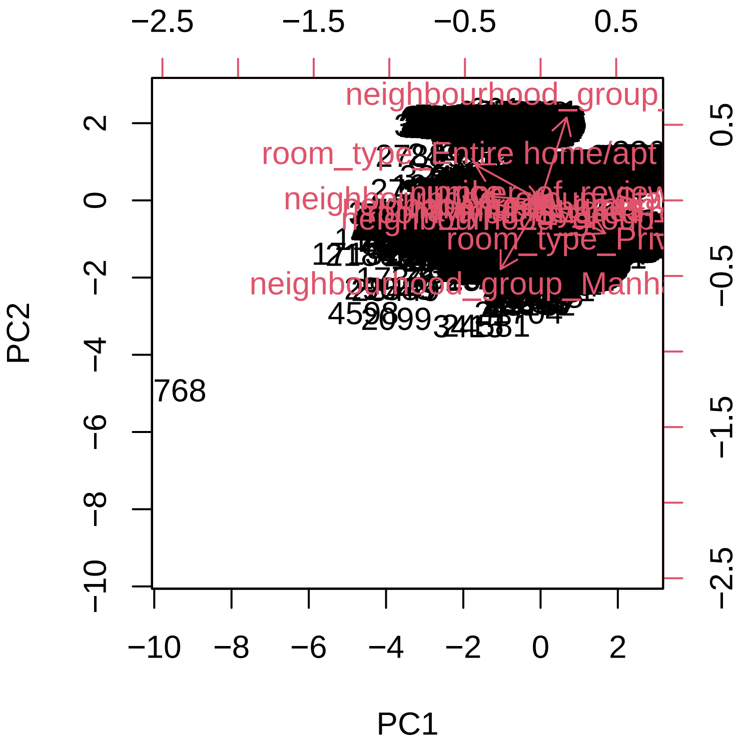
```
data_pc$rotation
```

```
##                                          PC1          PC2          PC3
## price                             -0.47966217  0.036549523 -0.011718544
## minimum_nights                    -0.10185633 -0.066957782 -0.004096455
## number_of_reviews                  0.05068042  0.044129881  0.406144122
## availability_365                  -0.03311190 -0.084708249  0.528064395
## neighbourhood_group_Bronx          0.05719935 -0.048555580  0.166912237
## neighbourhood_group_Brooklyn       0.21500724  0.684914813 -0.189419953
## neighbourhood_group_Manhattan     -0.33021598 -0.569073436 -0.286116743
## neighbourhood_group_Queens         0.14462891 -0.138518220  0.579633077
## neighbourhood_group_Staten Island  0.01325507 -0.004490779  0.174869128
## room_type_Entire home/apt         -0.53875138  0.295982936  0.146587293
## room_type_Private room             0.53433783 -0.271505693 -0.149371424
## room_type_Shared room              0.02188455 -0.096013696  0.009223039
##                                          PC4          PC5          PC6
## price                              0.01819551 -0.054033870  0.111586312
## minimum_nights                     0.40850747  0.726997320  0.025377788
## number_of_reviews                  0.28971710 -0.597112083  0.148670694
## availability_365                   0.52623774  0.105134124  0.006223651
## neighbourhood_group_Bronx          0.13696530 -0.035157401 -0.567366458
## neighbourhood_group_Brooklyn       0.18115778  0.015571562  0.003555969
## neighbourhood_group_Manhattan      0.13511772 -0.173403043  0.070587636
## neighbourhood_group_Queens        -0.57713649  0.258749880  0.155517961
## neighbourhood_group_Staten Island  0.16579528 -0.026565250 -0.013578277
## room_type_Entire home/apt         -0.11212157  0.016984521  0.007875203
## room_type_Private room             0.13838711 -0.007148973  0.191504469
## room_type_Shared room             -0.09898054 -0.037620311 -0.759444936
##                                          PC7         PC8         PC9
## price                             -0.03511387  0.04936867  0.35020298
## minimum_nights                     0.10664732  0.07452865 -0.48483548
## number_of_reviews                  0.17390598  0.16130015 -0.52299194
## availability_365                   0.14933386  0.12618820  0.56278691
## neighbourhood_group_Bronx         -0.73745392  0.16727608 -0.06988434
## neighbourhood_group_Brooklyn       0.13284286  0.05925597  0.09684424
## neighbourhood_group_Manhattan      0.07584188  0.02833185 -0.04340898
## neighbourhood_group_Queens         0.05630430  0.05693692 -0.03507658
## neighbourhood_group_Staten Island -0.08801012 -0.95486132 -0.03886844
## room_type_Entire home/apt         -0.08503286 -0.01249027 -0.12974097
## room_type_Private room            -0.06891200  0.02432750  0.12733436
## room_type_Shared room              0.58712568 -0.04496640  0.01038809
##                                          PC10          PC11          PC12
## price                              0.79147735  3.575840e-14  1.265504e-15
## minimum_nights                     0.19257223 -1.321335e-15  4.681163e-16
## number_of_reviews                  0.19536529 -4.332497e-16 -2.202910e-16
## availability_365                  -0.26439528  8.080386e-16  1.241154e-15
## neighbourhood_group_Bronx          0.10158995 -1.352020e-02  1.934010e-01
## neighbourhood_group_Brooklyn       0.05161264 -4.313073e-02  6.169676e-01
## neighbourhood_group_Manhattan     -0.18217080 -4.356868e-02  6.232323e-01
## neighbourhood_group_Queens         0.12610242 -2.926917e-02  4.186836e-01
## neighbourhood_group_Staten Island  0.07997197 -8.083098e-03  1.156254e-01
## room_type_Entire home/apt         -0.28095984 -6.944185e-01 -4.854514e-02
## room_type_Private room             0.24256839 -6.927053e-01 -4.842538e-02
## room_type_Shared room              0.14886676 -1.818761e-01 -1.271452e-02
```

```
data_pc
```

```
## Standard deviations (1, .., p=12):
##  [1] 1.617167e+00 1.278297e+00 1.129992e+00 1.058038e+00 1.036809e+00
##  [6] 1.026533e+00 1.009306e+00 1.002770e+00 8.650626e-01 6.731136e-01
## [11] 4.170666e-14 6.135359e-15
## 
## Rotation (n x k) = (12 x 12):
##                                           PC1          PC2          PC3
## price                             -0.47966217  0.036549523 -0.011718544
## minimum_nights                    -0.10185633 -0.066957782 -0.004096455
## number_of_reviews                  0.05068042  0.044129881  0.406144122
## availability_365                  -0.03311190 -0.084708249  0.528064395
## neighbourhood_group_Bronx          0.05719935 -0.048555580  0.166912237
## neighbourhood_group_Brooklyn       0.21500724  0.684914813 -0.189419953
## neighbourhood_group_Manhattan     -0.33021598 -0.569073436 -0.286116743
## neighbourhood_group_Queens         0.14462891 -0.138518220  0.579633077
## neighbourhood_group_Staten Island  0.01325507 -0.004490779  0.174869128
## room_type_Entire home/apt         -0.53875138  0.295982936  0.146587293
## room_type_Private room             0.53433783 -0.271505693 -0.149371424
## room_type_Shared room              0.02188455 -0.096013696  0.009223039
##                                           PC4          PC5          PC6
## price                              0.01819551 -0.054033870  0.111586312
## minimum_nights                     0.40850747  0.726997320  0.025377788
## number_of_reviews                  0.28971710 -0.597112083  0.148670694
## availability_365                   0.52623774  0.105134124  0.006223651
## neighbourhood_group_Bronx          0.13696530 -0.035157401 -0.567366458
## neighbourhood_group_Brooklyn       0.18115778  0.015571562  0.003555969
## neighbourhood_group_Manhattan      0.13511772 -0.173403043  0.070587636
## neighbourhood_group_Queens        -0.57713649  0.258749880  0.155517961
## neighbourhood_group_Staten Island  0.16579528 -0.026565250 -0.013578277
## room_type_Entire home/apt         -0.11212157  0.016984521  0.007875203
## room_type_Private room             0.13838711 -0.007148973  0.191504469
## room_type_Shared room             -0.09898054 -0.037620311 -0.759444936
##                                           PC7         PC8         PC9
## price                             -0.03511387  0.04936867  0.35020298
## minimum_nights                     0.10664732  0.07452865 -0.48483548
## number_of_reviews                  0.17390598  0.16130015 -0.52299194
## availability_365                   0.14933386  0.12618820  0.56278691
## neighbourhood_group_Bronx         -0.73745392  0.16727608 -0.06988434
## neighbourhood_group_Brooklyn       0.13284286  0.05925597  0.09684424
## neighbourhood_group_Manhattan      0.07584188  0.02833185 -0.04340898
## neighbourhood_group_Queens         0.05630430  0.05693692 -0.03507658
## neighbourhood_group_Staten Island -0.08801012 -0.95486132 -0.03886844
## room_type_Entire home/apt         -0.08503286 -0.01249027 -0.12974097
## room_type_Private room            -0.06891200  0.02432750  0.12733436
## room_type_Shared room              0.58712568 -0.04496640  0.01038809
##                                          PC10          PC11          PC12
## price                              0.79147735  3.575840e-14  1.265504e-15
## minimum_nights                     0.19257223 -1.321335e-15  4.681163e-16
## number_of_reviews                  0.19536529 -4.332497e-16 -2.202910e-16
## availability_365                  -0.26439528  8.080386e-16  1.241154e-15
## neighbourhood_group_Bronx          0.10158995 -1.352020e-02  1.934010e-01
## neighbourhood_group_Brooklyn       0.05161264 -4.313073e-02  6.169676e-01
## neighbourhood_group_Manhattan     -0.18217080 -4.356868e-02  6.232323e-01
## neighbourhood_group_Queens         0.12610242 -2.926917e-02  4.186836e-01
## neighbourhood_group_Staten Island  0.07997197 -8.083098e-03  1.156254e-01
## room_type_Entire home/apt         -0.28095984 -6.944185e-01 -4.854514e-02
## room_type_Private room             0.24256839 -6.927053e-01 -4.842538e-02
## room_type_Shared room              0.14886676 -1.818761e-01 -1.271452e-02
```

```r
biplot(data_pc, scale = 0)
```



In the next few cells I experiment with modeling only over data in one of the neighbothood groups. Results were less significant so we disregard this part in our analysis.

```
df_bronx <- data[data$neighbourhood_group == "Bronx",]
head(df_bronx)
```

```
## # A tibble: 6 x 8
##   neighbourhood_g~ neighbourhood room_type price minimum_nights number_of_revie~
##   <chr>            <chr>         <chr>     <dbl>          <dbl>            <dbl>
## 1 Bronx            Highbridge    Private ~    40              1              219
## 2 Bronx            Highbridge    Private ~    45              1              138
## 3 Bronx            Clason Point  Private ~    90              2                0
## 4 Bronx            Kingsbridge   Entire h~    90             30                4
## 5 Bronx            Woodlawn      Entire h~    77              1              197
## 6 Bronx            University H~ Private ~    37              4              117
## # ... with 2 more variables: reviews_per_month <dbl>, availability_365 <dbl>
```

```
df_bronx_dummies <- dummy_cols(df_bronx, select_columns = c("neighbourhood_group", "room_type"))
head(df_bronx_dummies)
```

```
## # A tibble: 6 x 12
##   neighbourhood_g~ neighbourhood room_type price minimum_nights number_of_revie~
##   <chr>            <chr>         <chr>     <dbl>          <dbl>            <dbl>
## 1 Bronx            Highbridge    Private ~    40              1              219
## 2 Bronx            Highbridge    Private ~    45              1              138
## 3 Bronx            Clason Point  Private ~    90              2                0
## 4 Bronx            Kingsbridge   Entire h~    90             30                4
## 5 Bronx            Woodlawn      Entire h~    77              1              197
## 6 Bronx            University H~ Private ~    37              4              117
## # ... with 6 more variables: reviews_per_month <dbl>, availability_365 <dbl>,
## #   neighbourhood_group_Bronx <int>, room_type_Entire home/apt <int>,
## #   room_type_Private room <int>, room_type_Shared room <int>
```

```
#remove useless variables
drop = c( "neighbourhood_group", "neighbourhood", "room_type", "reviews_per_month")
df_bronx_dummies = df_bronx_dummies[,!(names(df_bronx_dummies) %in% drop)]
head(df_dummies)
```

```
## # A tibble: 6 x 12
##   price minimum_nights number_of_reviews availability_365 neighbourhood_group_B~
##   <dbl>          <dbl>             <dbl>            <dbl>                  <int>
## 1   225              1                45              355                      0
## 2    89              1               270              194                      0
## 3   200              3                74              129                      0
## 4    79              2               430              220                      0
## 5   150              1               160              188                      0
## 6   135              5                53                6                      0
## # ... with 7 more variables: neighbourhood_group_Brooklyn <int>,
## #   neighbourhood_group_Manhattan <int>, neighbourhood_group_Queens <int>,
## #   neighbourhood_group_Staten Island <int>, room_type_Entire home/apt <int>,
## #   room_type_Private room <int>, room_type_Shared room <int>
```

```
fit_bronx <- lm(price ~ ., data = df_bronx_dummies)
summary(fit_bronx)
```

```
##
## Call:
## lm(formula = price ~ ., data = df_bronx_dummies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87.60 -26.79 -10.52  10.05 421.10
##
## Coefficients: (2 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 50.44787   11.71433   4.307 1.87e-05 ***
## minimum_nights              -0.26487    0.12539  -2.112    0.035 *
## number_of_reviews           -0.16684    0.04141  -4.029 6.15e-05 ***
## availability_365             0.06873    0.01479   4.646 3.96e-06 ***
## neighbourhood_group_Bronx         NA         NA      NA       NA
## `room_type_Entire home/apt` 67.70292   11.94263   5.669 2.01e-08 ***
## `room_type_Private room`     6.49534   11.81394   0.550    0.583
## `room_type_Shared room`           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.67 on 790 degrees of freedom
## Multiple R-squared:  0.2637, Adjusted R-squared:  0.259
## F-statistic: 56.59 on 5 and 790 DF,  p-value: < 2.2e-16
```

```
# Stepwise Regression
library(MASS)
step <- stepAIC(fit_bronx, direction="both")
```

```
## Start:  AIC=6316.59
## price ~ minimum_nights + number_of_reviews + availability_365 +
##     neighbourhood_group_Bronx + `room_type_Entire home/apt` + 
##     `room_type_Private room` + `room_type_Shared room`
## 
## 
## Step:  AIC=6316.59
## price ~ minimum_nights + number_of_reviews + availability_365 + 
##     neighbourhood_group_Bronx + `room_type_Entire home/apt` + 
##     `room_type_Private room`
## 
## 
## Step:  AIC=6316.59
## price ~ minimum_nights + number_of_reviews + availability_365 + 
##     `room_type_Entire home/apt` + `room_type_Private room`
## 
##                               Df Sum of Sq     RSS    AIC
## - `room_type_Private room`     1       838 2192002 6314.9
## <none>                                     2191164 6316.6
## - minimum_nights               1     12375 2203539 6319.1
## - number_of_reviews            1     45016 2236180 6330.8
## - availability_365             1     59870 2251034 6336.1
## - `room_type_Entire home/apt`  1     89138 2280302 6346.3
## 
## Step:  AIC=6314.9
## price ~ minimum_nights + number_of_reviews + availability_365 + 
##     `room_type_Entire home/apt`
## 
##                               Df Sum of Sq     RSS    AIC
## <none>                                     2192002 6314.9
## + `room_type_Private room`     1       838 2191164 6316.6
## + `room_type_Shared room`      1       838 2191164 6316.6
## - minimum_nights               1     12173 2204176 6317.3
## - number_of_reviews            1     44263 2236266 6328.8
## - availability_365             1     60106 2252109 6334.4
## - `room_type_Entire home/apt`  1    718314 2910317 6538.5
```

```
step$anova # display results
```

```
## Stepwise Model Path 
## Analysis of Deviance Table
## 
## Initial Model:
## price ~ minimum_nights + number_of_reviews + availability_365 + 
##     neighbourhood_group_Bronx + `room_type_Entire home/apt` + 
##     `room_type_Private room` + `room_type_Shared room`
## 
## Final Model:
## price ~ minimum_nights + number_of_reviews + availability_365 + 
##     `room_type_Entire home/apt`
## 
## 
##                           Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                                                790    2191164 6316.594
## 2   - `room_type_Shared room`  0   0.0000       790    2191164 6316.594
## 3 - neighbourhood_group_Bronx  0   0.0000       790    2191164 6316.594
## 4  - `room_type_Private room`  1 838.4199       791    2192002 6314.899
```