# Trabalho1: Mineração De Dados

*Rodrigo Malta Esteves*

*24/02/2021*



## Introdução

O conjunto de dados a ser analisado corresponde a registros púublicos de vendas de casas feitas de maio de 2014 a maio de 2015 no Condado de King, no estado de Washington,EUA.

## Pacotes

```
#Pacotes
library(ggplot2)
library(readr)
library(tidyr)
library(dplyr)
library(lubridate)
library(corrplot)
library("scatterplot3d")
library(cowplot)
```

## Importando os dados

```r
#Importando os dados do primeiro .csv
housing_1 <- read_csv("C:\\Users\\malta\\Desktop\\Pós Graduação\\Mineração de Dados\\Trabalhos\\Atividad
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   date = col_datetime(format = ""),
##   price = col_double(),
##   bedrooms = col_double(),
##   bathrooms = col_double(),
##   sqft_living = col_double(),
##   sqft_lot = col_double(),
##   floors = col_double(),
##   waterfront = col_double(),
##   view = col_double(),
##   condition = col_double(),
##   date2 = col_character()
## )
```

```r
#Importando os dados do segundo .csv
housing_2 <- read_csv("C:\\Users\\malta\\Desktop\\Pós Graduação\\Mineração de Dados\\Trabalhos\\Atividad
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   date = col_datetime(format = ""),
##   grade = col_double(),
##   sqft_above = col_double(),
##   sqft_basement = col_double(),
##   yr_built = col_double(),
##   yr_renovated = col_double(),
##   zipcode = col_double(),
##   lat = col_double(),
##   long = col_double(),
##   sqft_living15 = col_double(),
##   sqft_lot15 = col_double()
## )
```

## Questão 1

Data frame de todos os dados usando o merge().

```r
#Data frame de todos os dados
house <- merge(housing_1,housing_2,by=c("id","date"))
#Transformando variáveis quantitativas em categóricas
house$id <- as.character(house$id)
house$waterfront <- as.factor(house$waterfront)
house$view <- as.factor(house$view)
house$condition <- as.factor(house$condition)
house$grade <- as.factor(house$grade)
house$zipcode <- as.character(house$zipcode)
```

```r
house$date2 <- parse_date_time((house$date2),orders="%Y-%m")
house$yr_built <- parse_date_time((house$yr_built),orders="%Y")
house$yr_renovated <- parse_date_time((house$yr_renovated),orders="%Y")

str(house)
```

```
## 'data.frame':    21613 obs. of  22 variables:
##  $ id           : chr  "1000102" "1000102" "100100050" "1001200035" ...
##  $ date         : POSIXct, format: "2014-09-16" "2015-04-22" ...
##  $ price        : num  280000 300000 275000 272450 259000 ...
##  $ bedrooms     : num  6 6 3 3 4 3 3 3 3 3 ...
##  $ bathrooms    : num  3 3 1 1 1.5 1 2.25 2.5 2.5 1 ...
##  $ sqft_living  : num  2400 2400 1320 1350 1260 980 1430 1520 1520 1100 ...
##  $ sqft_lot     : num  9373 9373 11090 7973 7248 ...
##  $ floors       : num  2 2 1 1.5 1.5 1 2 2 2 1 ...
##  $ waterfront   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ view         : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ condition    : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 3 5 3 3 3 3 3 ...
##  $ date2        : POSIXct, format: "2014-09-01" "2015-04-01" ...
##  $ grade        : Factor w/ 12 levels "1","3","4","5",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ sqft_above   : num  2400 2400 1320 1350 1260 980 1430 1520 1520 1100 ...
##  $ sqft_basement: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ yr_built     : POSIXct, format: "1991-01-01" "1991-01-01" ...
##  $ yr_renovated : POSIXct, format: NA NA ...
##  $ zipcode      : chr  "98002" "98002" "98155" "98188" ...
##  $ lat          : num  47.3 47.3 47.8 47.4 47.4 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: num  2060 2060 1320 1310 1300 ...
##  $ sqft_lot15   : num  7316 7316 8319 7491 7732 ...
```

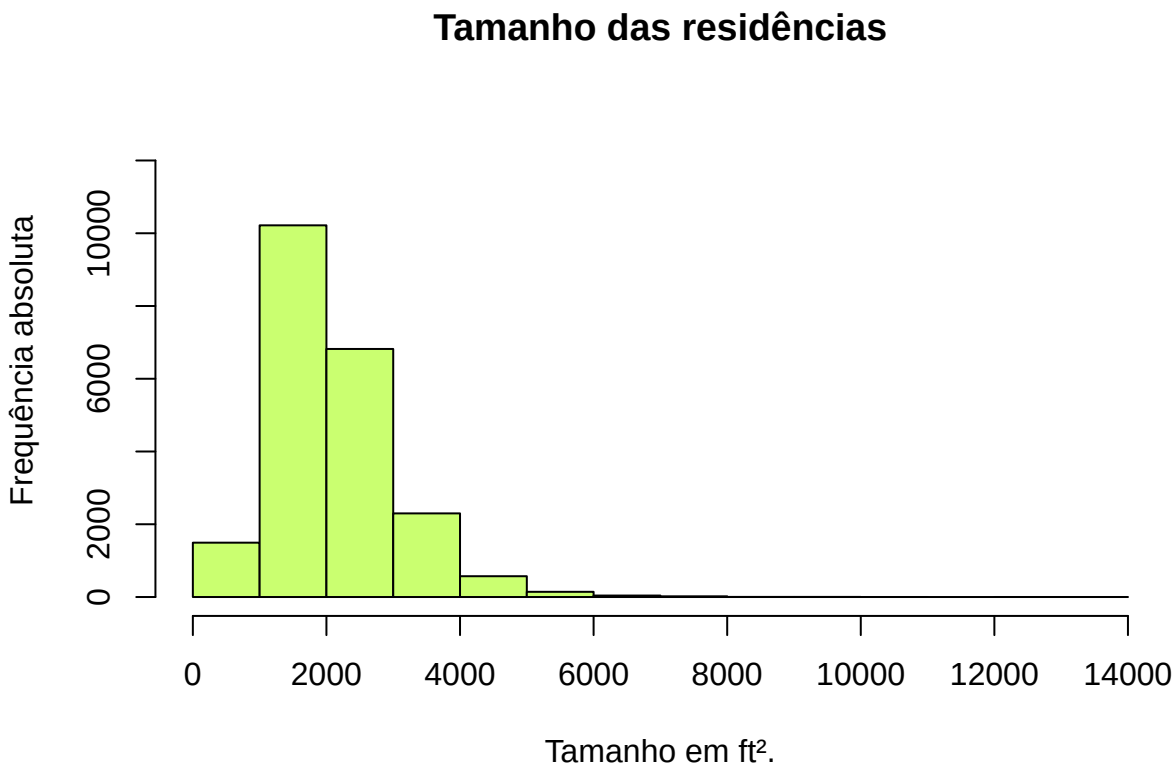
## a) Pés quadrados do interior da casa

A primeira variável analisada foi "sqft_living", referente ao tamanho do interior da casa em pés quadrados. Observamos que em média as casas têm aproximadamente 2018ft² e 95% dos valores se concentram entre 820 e 4270ft². Podemos perceber que existem outliers com valores muito superiores aos valores que concentram a maior quantidade de dados. A maior casa, por exemplo, chega a ter 12540ft².

```r
house %>%
  summarize(media_tamanho = mean(house$sqft_living),
            desvio_padrao_tamanho = sd(house$sqft_living),
            mediana_tamanho = median(house$sqft_living),
            tamanho_min = min(house$sqft_living),
            tamanho_max = max(house$sqft_living),
            quantil_inf = quantile(house$sqft_living,probs=.025),
            quantil_sup = quantile(house$sqft_living,probs=.975))
```

```
##   media_tamanho desvio_padrao_tamanho mediana_tamanho tamanho_min
## 1        2079.9              918.4409            1910         290
##   tamanho_max quantil_inf quantil_sup
## 1       13540         820        4270
```

```
hist(house$sqft_living,xlab="Tamanho em ft².",ylab="Frequência absoluta",main="Tamanho das residências"
```




### b) Condição do apartamento

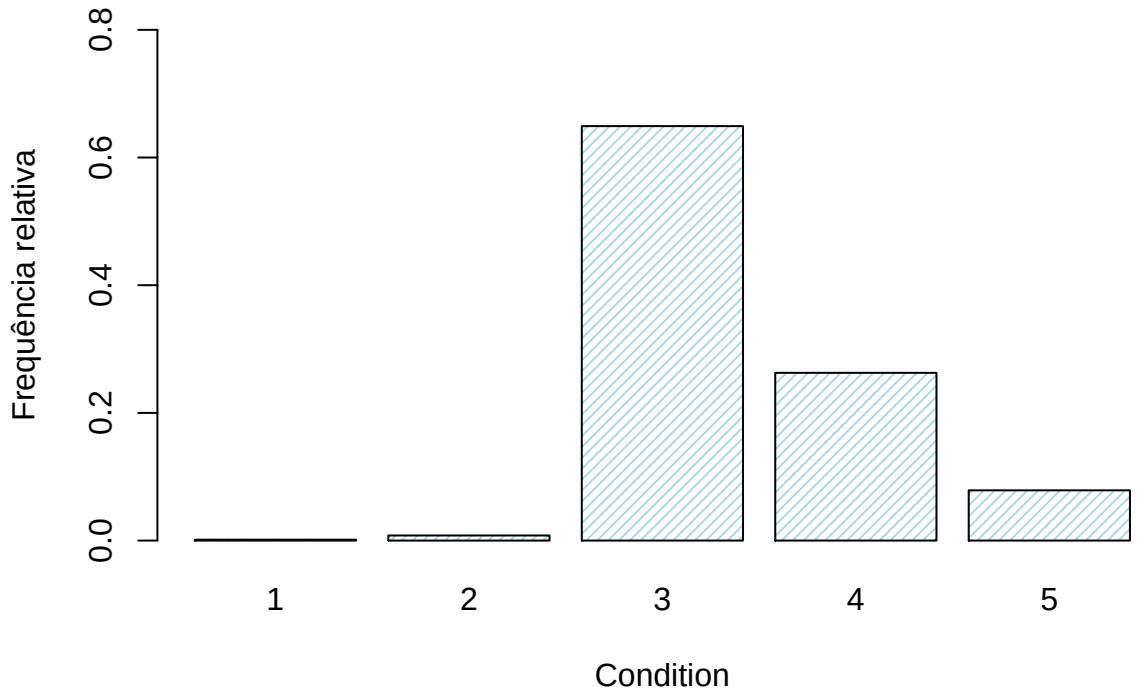
A segunda variável "condition" diz respeito a condição do apartamento. Olhando a tabela de proporções, vemos que a maior parte das casas tem nível 3, 4 ou 5, enquanto apenas 202 unidades apresentam nível interior. Confirmamos essa informação observando os histogramas e o gráfico de setores.

```
house %>%
  count(condition) %>%
  mutate(prop = n/sum(n))
```

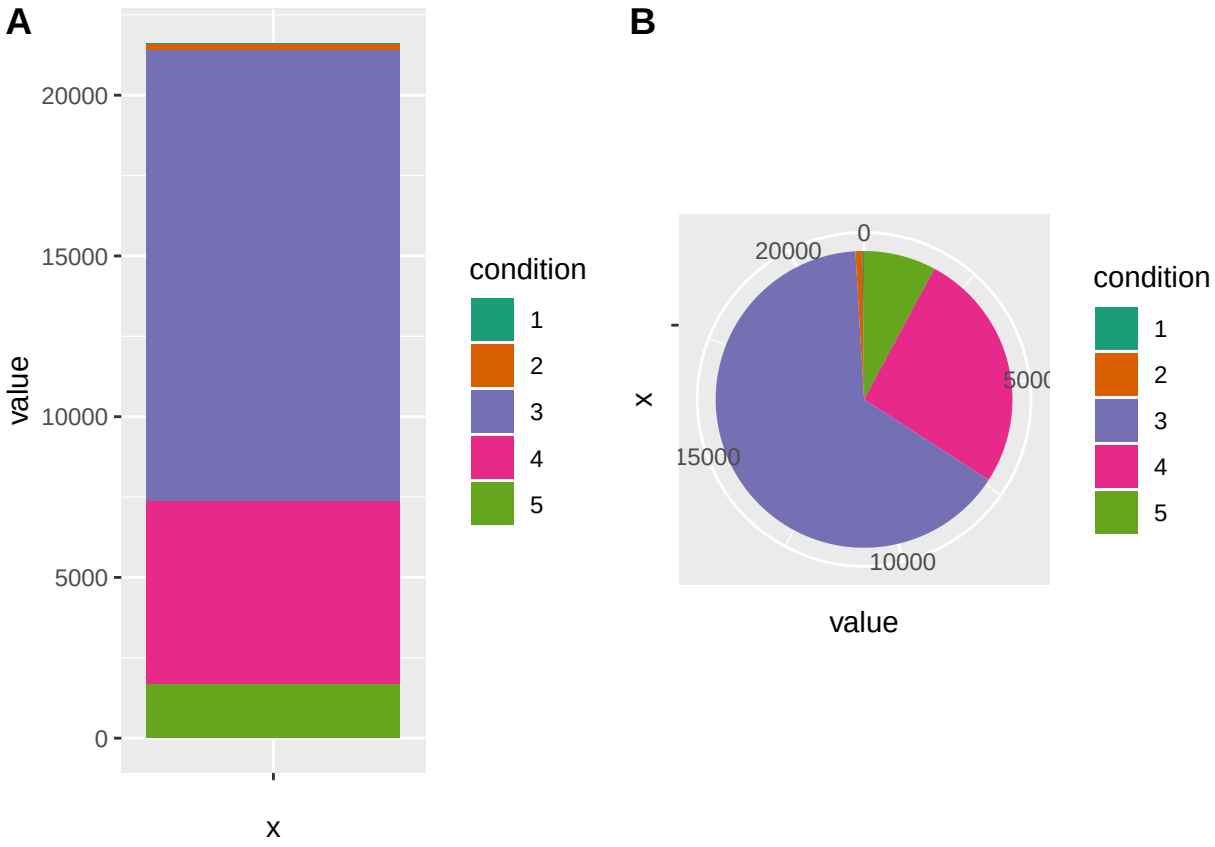
```
## # A tibble: 5 x 3
##   condition     n    prop
##   <fct>     <int>   <dbl>
## 1 1            30 0.00139
## 2 2           172 0.00796
## 3 3         14031 0.649
## 4 4          5679 0.263
## 5 5          1701 0.0787
```

```
#barplot
cont <- table(house$condition)
```

```r
ni <- as.numeric(cont)
fi <- ni/length(house$condition) # tranformando o objeto em um vetor numérico
barplot(fi,names=c("1","2","3","4","5"),ylim=c(0,.8),xlab="Condition",
        ylab="Frequência relativa",col="light blue",density=25)   # gráfico de barras
```




```r
#barplot (empilhado)
condition <- levels(as.factor(house$condition))
value <- table((as.factor(house$condition)))
df <- data.frame(condition,value)
c_bp<- ggplot(df, aes(x="", y=value, fill=condition))+
  geom_bar(width = 1, stat = "identity")+
  scale_fill_brewer(palette="Dark2")

#pieplot
pie <- c_bp + coord_polar("y", start=0)
c_pie <- pie + scale_fill_brewer(palette="Dark2")
plot_grid(c_bp,c_pie, labels = "AUTO")
```




## c) Nível de construção e design

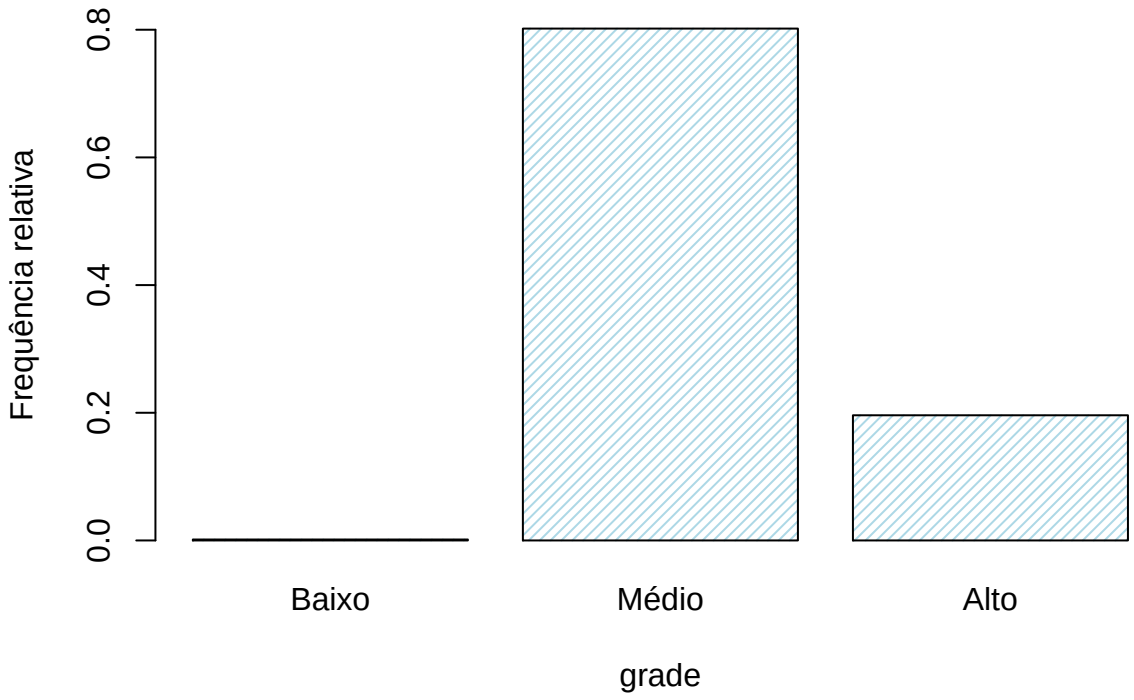
A terceira variável "grade" é um índice de 1 a 13 sobre o nível de construção e design. No entanto, alteramos essa gradação para os níveis categóricos descritos no enunciado. Sendo assim, consideramos os níveis "Baixo" para "grade" de 1 a 3, "Médio" para valores entre 3 e 11 e "Alto" para valores de 11 a 13. Observamos graficamente que grande parte das moradias tem nível "Médio" e "Alto", correspondendo a 99% de todas as unidades. Apenas 32 das 21613 casas foram avaliadas como "Baixo".

```
grade_factors <- cut(as.numeric(house$grade), c(1, 3, 7,11), labels=c("Baixo","Médio","Alto"))
house2 <- cbind(house,grade_factors)
house2 %>%
  count(grade_factors) %>%
  mutate(prop = n/sum(n))
```

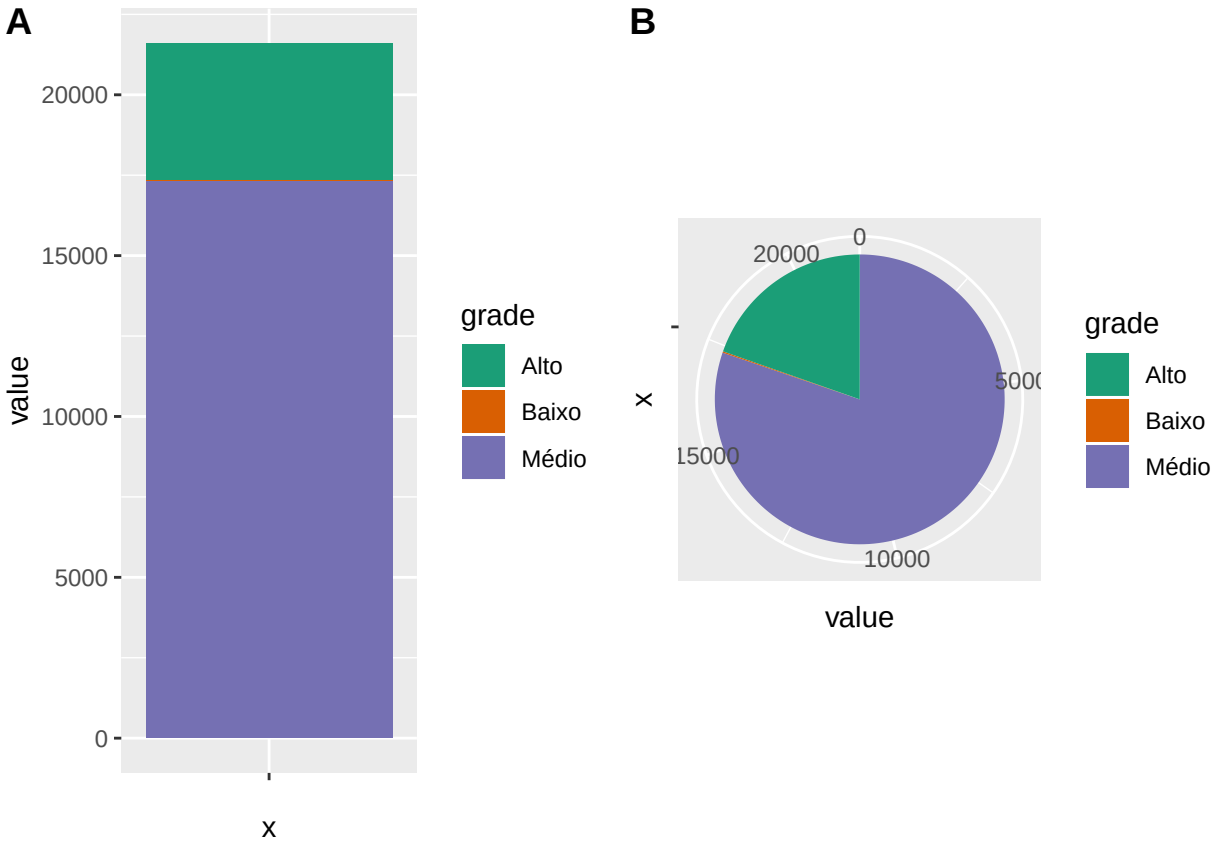
```
## # A tibble: 4 x 3
##   grade_factors     n     prop
##   <fct>         <int>    <dbl>
## 1 Baixo            32 0.00148
## 2 Médio         17329 0.802
## 3 Alto           4238 0.196
## 4 <NA>             14 0.000648
```

```
#barplot
cont <- table(house2$grade_factors)
```

```r
ni <- as.numeric(cont)
fi <- ni/length(house2$grade_factors) # tranformando o objeto em um vetor numérico

barplot(fi,names=c("Baixo","Médio","Alto"),ylim=c(0,.8),xlab="grade",
        ylab="Frequência relativa",col="light blue",density=25)   # gráfico de barras
```




```r
#barplot (empilhado)
grade <- levels(house2$grade_factors)
value <- table(house2$grade_factors)
df <- data.frame(grade,value)
v_bp<- ggplot(df, aes(x="", y=value, fill=grade))+
  geom_bar(width = 1, stat = "identity")+
  scale_fill_brewer(palette="Dark2")

#pieplot
pie <- v_bp + coord_polar("y", start=0)
v_pie <- pie + scale_fill_brewer(palette="Dark2")

plot_grid(v_bp,v_pie, labels = "AUTO")
```




## Questão 2

### a) Gráficos de correlação

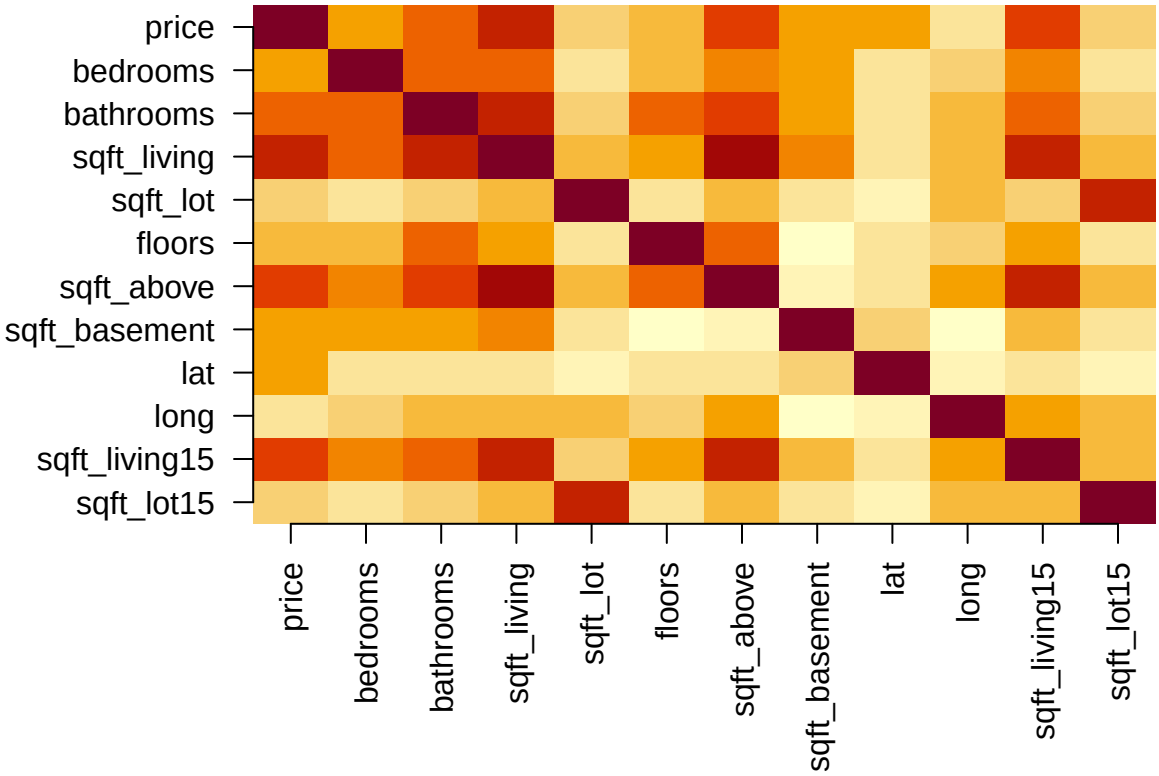
Podemos ver os valores das correlações na tabela e sua distribuição nos gráficos a seguir. Os dois pares mais correlacionados entre as variáveis quantitativas foram "sqft_above" (pés quadrados do interior da casa a cima do nível do solo) e "sqft_living" (pés quadrados do interior da casa), e "sqft_living" e "sqft_living15" (pés quadrados do terreno das 15 casas mais próximas).

```
house_num <- select_if(house, is.numeric)
#Correlação
corre <- cor(house_num)
corre
```

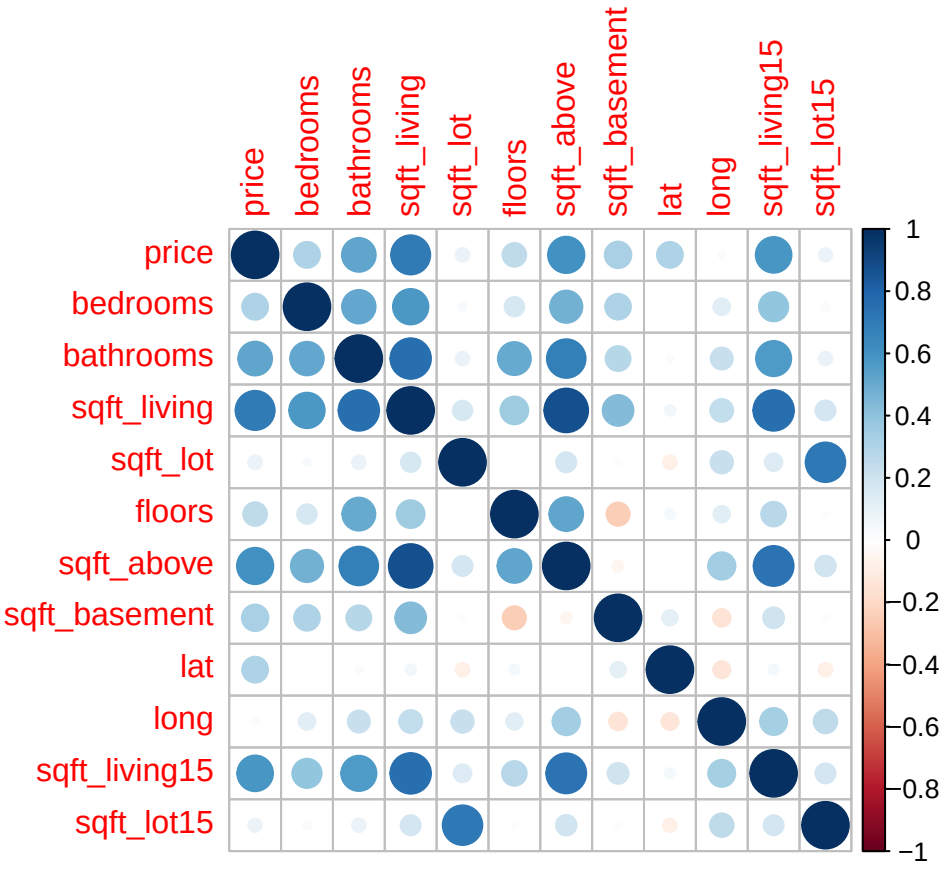
```
##                    price    bedrooms  bathrooms sqft_living     sqft_lot
## price          1.00000000  0.30834960 0.52513751  0.70203505  0.089660861
## bedrooms       0.30834960  1.00000000 0.51588364  0.57667069  0.031703243
## bathrooms      0.52513751  0.51588364 1.00000000  0.75466528  0.087739662
## sqft_living    0.70203505  0.57667069 0.75466528  1.00000000  0.172825661
## sqft_lot       0.08966086  0.03170324 0.08773966  0.17282566  1.000000000
## floors         0.25679389  0.17542894 0.50065317  0.35394929 -0.005200991
## sqft_above     0.60556730  0.47760016 0.68534248  0.87659660  0.183512281
## sqft_basement  0.32381602  0.30309338 0.28377003  0.43504297  0.015286202
## lat            0.30700348 -0.00893101 0.02457295  0.05252946 -0.085682788
```

```
## long          0.02162624  0.12947298 0.22304184  0.24022330  0.229520859
## sqft_living15 0.58537890  0.39163752 0.56863429  0.75642026  0.144608174
## sqft_lot15    0.08244715  0.02924422 0.08717536  0.18328555  0.718556752
##                     floors    sqft_above sqft_basement           lat
## price          0.256793888  0.6055672984    0.32381602  0.3070034800
## bedrooms       0.175428935  0.4776001614    0.30309338 -0.0089310097
## bathrooms      0.500653173  0.6853424759    0.28377003  0.0245729528
## sqft_living    0.353949290  0.8765965987    0.43504297  0.0525294622
## sqft_lot      -0.005200991  0.1835122809    0.01528620 -0.0856827882
## floors         1.000000000  0.5238847103   -0.24570454  0.0496141310
## sqft_above     0.523884710  1.0000000000   -0.05194331 -0.0008164986
## sqft_basement -0.245704542 -0.0519433068    1.00000000  0.1105379580
## lat            0.049614131 -0.0008164986    0.11053796  1.0000000000
## long           0.125419028  0.3438030175   -0.14476477 -0.1355117836
## sqft_living15  0.279885265  0.7318702924    0.20035498  0.0488579321
## sqft_lot15    -0.011269187  0.1940498619    0.01727618 -0.0864188072
##                      long sqft_living15  sqft_lot15
## price          0.02162624    0.58537890  0.08244715
## bedrooms       0.12947298    0.39163752  0.02924422
## bathrooms      0.22304184    0.56863429  0.08717536
## sqft_living    0.24022330    0.75642026  0.18328555
## sqft_lot       0.22952086    0.14460817  0.71855675
## floors         0.12541903    0.27988527 -0.01126919
## sqft_above     0.34380302    0.73187029  0.19404986
## sqft_basement -0.14476477    0.20035498  0.01727618
## lat           -0.13551178    0.04885793 -0.08641881
## long           1.00000000    0.33460498  0.25445129
## sqft_living15  0.33460498    1.00000000  0.18319175
## sqft_lot15     0.25445129    0.18319175  1.00000000
```

```
par(mar=c(8,8,1,1))
image(1:12,1:12,corre[,12:1],axes=FALSE,xlab="",ylab="")
axis(1,at=1:12,labels=names(house_num),las=2)
axis(2,at=1:12,labels=names(house_num)[12:1],las=2)
```



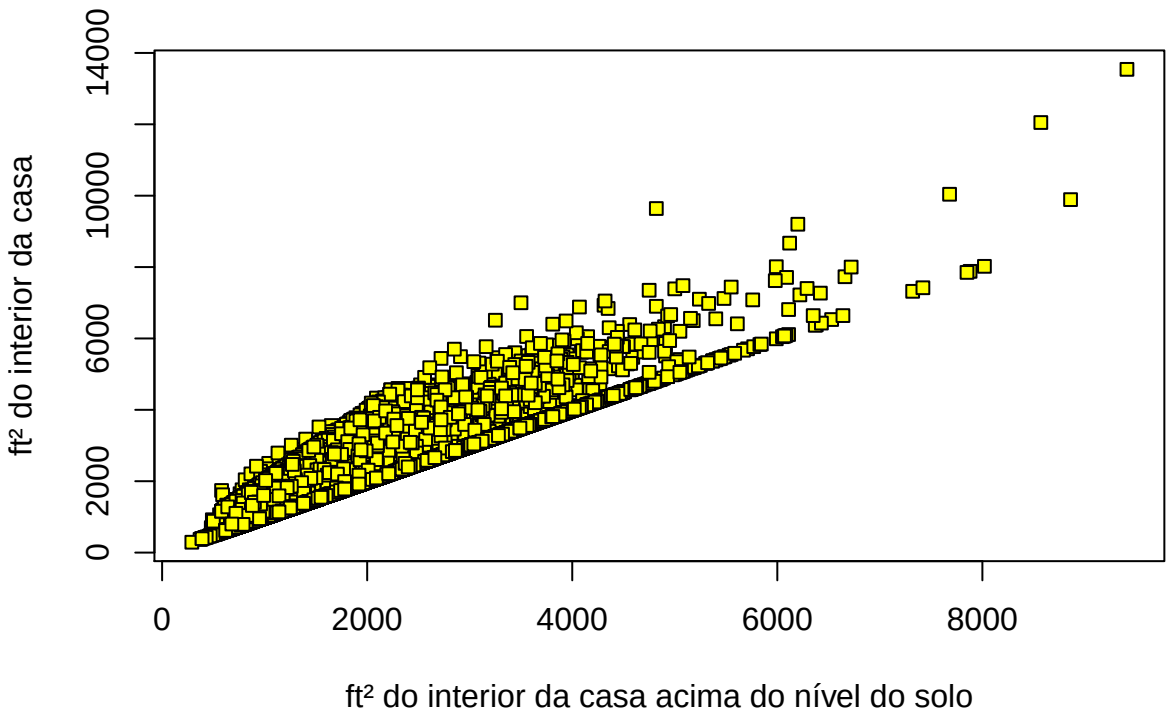

```
#Corre 2
corrplot(corre)
```



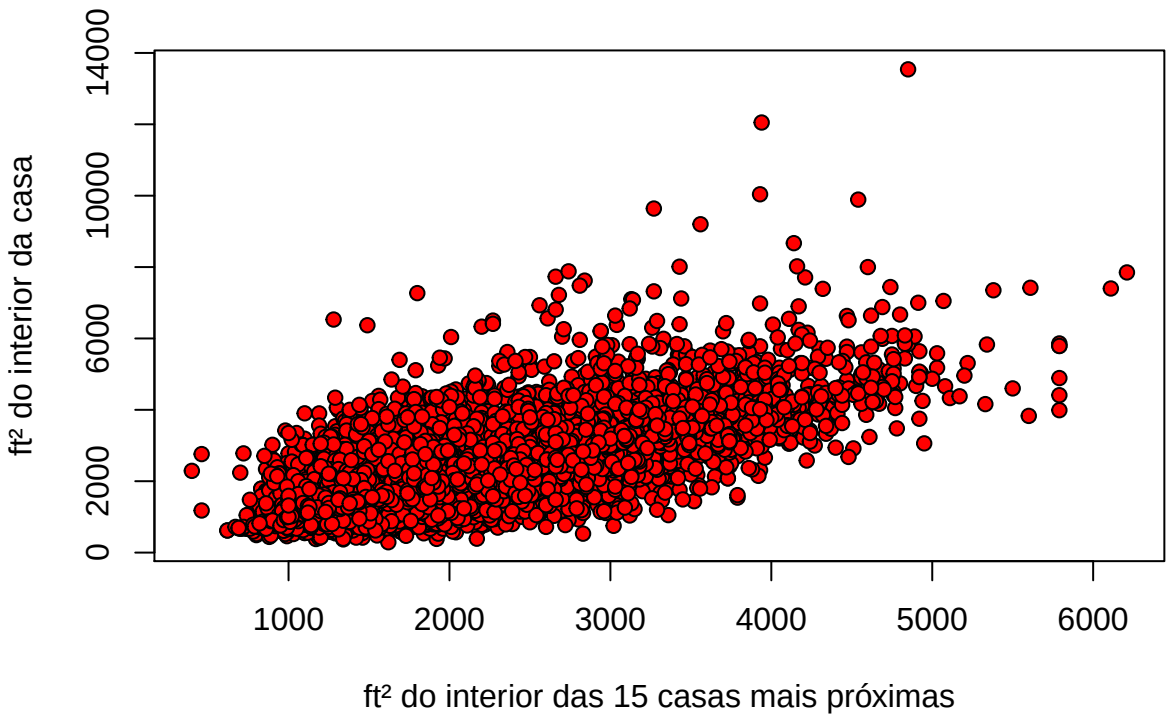

## b) Gráficos de dispersão

Os gráficos de dispersão mostram que há relação de linearidade entre cada par de variável. Assim como indicado pelos valores das correlações, temos uma dependência linear maior entre as variáveis "sqft_above" e "sqft_living".

```
#Dispersão sqft_above x sqft_living
plot(house$sqft_above,house$sqft_living,pch=22,bg="yellow",xlab="ft² do interior da casa acima do nível
```




```
#Dispersão sqft_living x sqft_living15
plot(house$sqft_living15,house$sqft_living,col=1,pch=21,bg=2,xlab="ft² do interior das 15 casas mais pr
```




## Questão 3

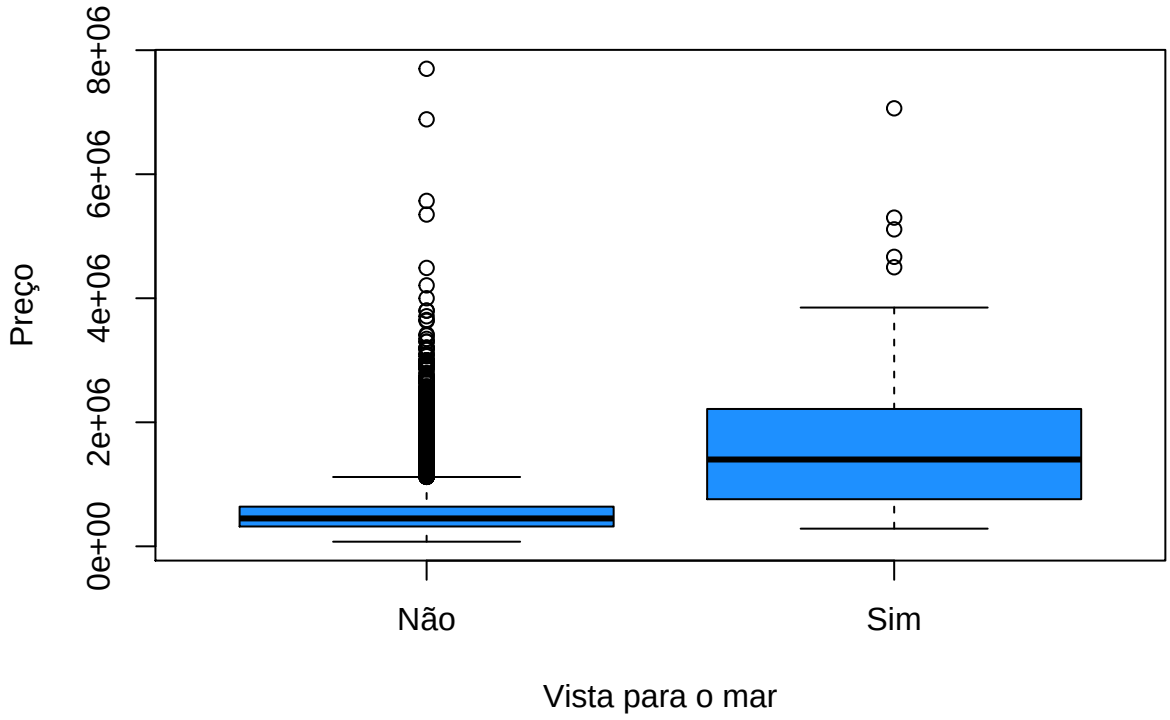
### a) Preço das casas x Vista para o mar

O preço das casas com vista para o mar é em média 3x o valor das casas que não possuem, mas também possuem a maior variação de preço. É interessante observar que o maior valor entre todas as casas não possui vista para o mar.

```
house %>%
  group_by(waterfront) %>%
  summarize(n_casas = n(),
            media_preco = mean(price),
            desvio_padrao_preco = sd(price),
            median_preco = median(price),
            min_preco = min(price),
            max_preco = max(price))
```

```
## # A tibble: 2 x 7
##   waterfront n_casas media_preco desvio_padrao_p~ median_preco min_preco
##   <fct>        <int>       <dbl>            <dbl>        <dbl>     <dbl>
## 1 0            21450     531564.          341600.       450000     75000
## 2 1              163    1661876.         1120372.      1400000    285000
## # ... with 1 more variable: max_preco <dbl>
```

```r
plot(x=house$waterfront,y=house$price,ylab="Preço",xlab="Vista para o mar",col="dodgerblue1" ,xaxt="n")
axis(1, at=1:2, labels=c("Não","Sim"))
```




## b) Preço das casas x Número de quartos

Aqui vemos que há correlação entre o preço da casa e o número de quartos. Em média, quanto maior o número de quartos, maior o preço. Vemos também um número muito grande de outliers para as casas entre 2 e 7 quartos.

```r
#Resumo das características dos quartos
house %>%
  summarize(media_qtd = mean(bedrooms),
            desvio_padrao_qtd = sd(bedrooms),
            mediana_qtd = median(bedrooms),
            qtd_min = min(bedrooms),
            qtd_max = max(bedrooms))
```

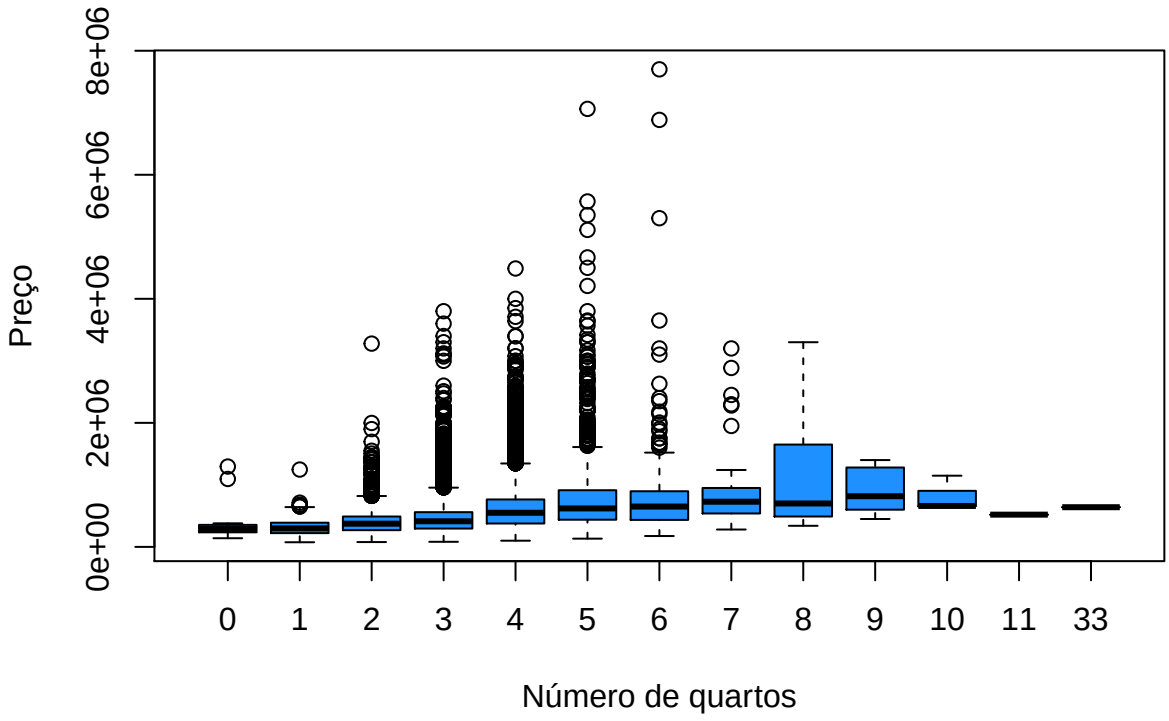
```
##   media_qtd desvio_padrao_qtd mediana_qtd qtd_min qtd_max
## 1  3.370842         0.9300618           3       0      33
```

```r
#Resumo dos preços das casas agrupados pelo número de quartos
house %>%
  group_by(bedrooms) %>%
```

```
  summarize(n_casas = n(),
            media_preco = mean(price),
            desvio_padrao_preco = sd(price),
            median_preco = median(price),
            min_preco = min(price),
            max_preco = max(price))
```

```
## # A tibble: 13 x 7
##    bedrooms n_casas media_preco desvio_padrao_p~ median_preco min_preco
##       <dbl>   <int>       <dbl>            <dbl>        <dbl>     <dbl>
##  1        0      13     409504.          358683.      288000     139950
##  2        1     199     317643.          148865.      299000      75000
##  3        2    2760     401373.          198052.      374000      78000
##  4        3    9824     466232.          262470.      413000      82000
##  5        4    6882     635420.          388594.      549998.    100000
##  6        5    1601     786600.          596204.      620000     133000
##  7        6     272     825521.          799239.      650000     175000
##  8        7      38     951185.          739954.      728580     280000
##  9        8      13    1105077.          897496.      700000     340000
## 10        9       6     894000.          381534.      817000     450000
## 11       10       3     819333.          284678.      660000     650000
## 12       11       1     520000              NaN       520000     520000
## 13       33       1     640000              NaN       640000     640000
## # ... with 1 more variable: max_preco <dbl>
```

```
#Boxplot dos preços pelo número de quartos
boxplot(house$price~house$bedrooms,ylab="Preço",xlab="Número de quartos",col="dodgerblue1")
```



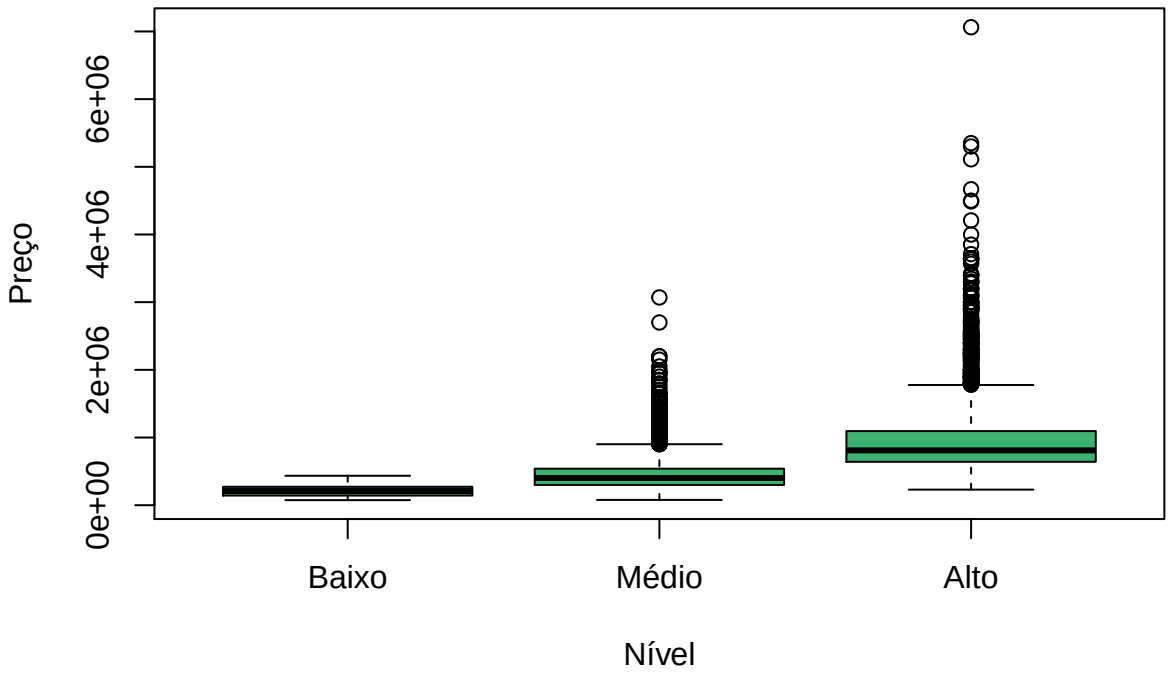

### c) Preço das casas x Nível

Aqui chama a atenção a grande quantidade de outliers nos níveis "Médio" e "Alto".

```
#Resumo dos preços das casas agrupados pelo número de quartos
house2 %>%
  group_by(grade_factors) %>%
  summarize(n_niveis = n(),
            media_preco = mean(price),
            desvio_padrao_preco = sd(price),
            median_preco = median(price),
            min_preco = min(price),
            max_preco = max(price))
```

```
## # A tibble: 4 x 7
##   grade_factors n_niveis media_preco desvio_padrao_p~ median_preco
##   <fct>            <int>       <dbl>            <dbl>        <dbl>
## 1 Baixo               32     213564.           94186.       211000
## 2 Médio            17329     437714.          195708.       402000
## 3 Alto              4238     951528.          526444.       810000
## 4 <NA>                14    3454786.         2025024.      2935500
## # ... with 2 more variables: min_preco <dbl>, max_preco <dbl>
```

```r
boxplot(house2$price~house2$grade_factors,ylab="Preço",xlab="Nível",col="mediumseagreen")
```




## Questão 4

### a) Média de preço da casa x Mês

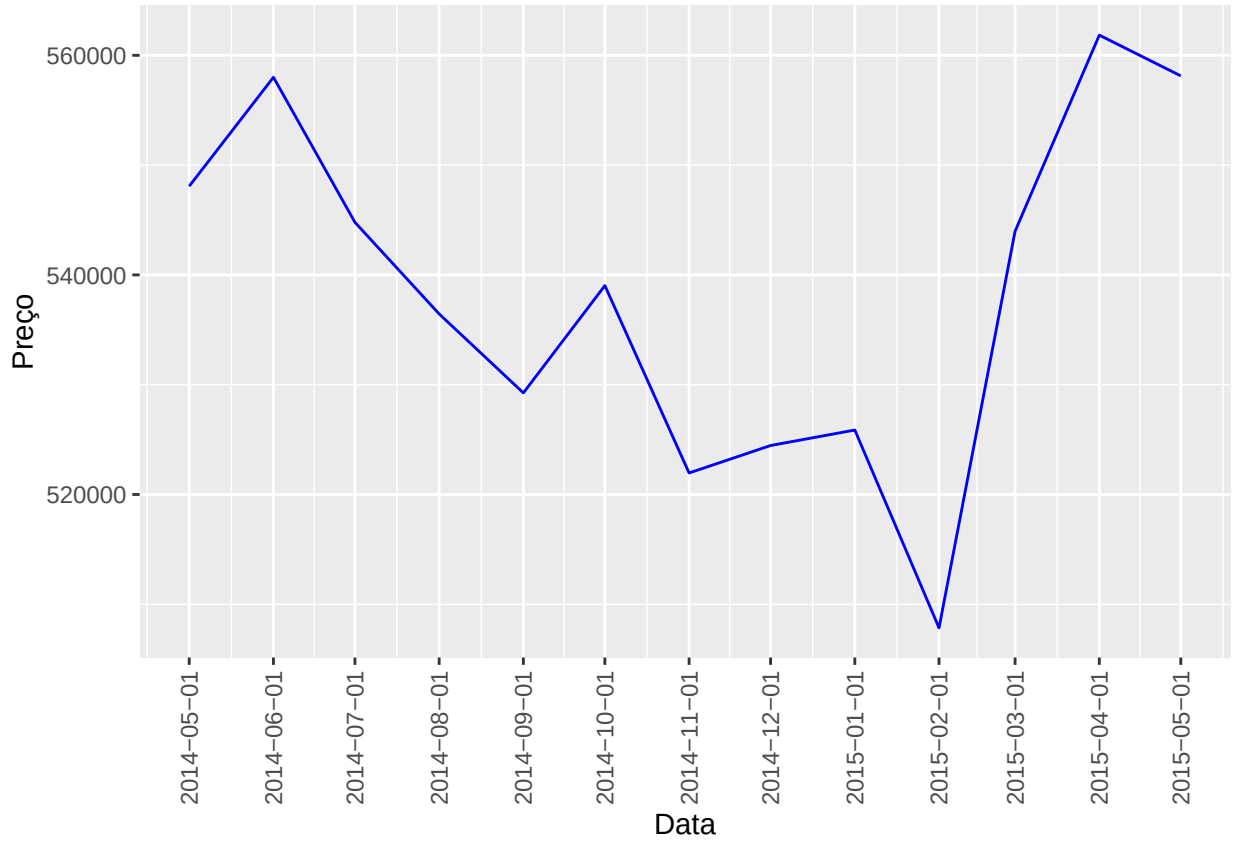
Podemos observar que a média dos preços tem uma tendência de queda de maio de 2014 até fevereiro de 2015, mas dá um grande salto positivo em março e abril de 2015.

```r
#Resumo dos preços das casas por mês
house_date <- house %>%
    group_by(month=floor_date(date2, "month")) %>%
    summarize(summary_variable=mean(price))
house_date
```

```
## # A tibble: 13 x 2
##    month               summary_variable
##    <dttm>                         <dbl>
##  1 2014-05-01 00:00:00          548080.
##  2 2014-06-01 00:00:00          558002.
##  3 2014-07-01 00:00:00          544789.
##  4 2014-08-01 00:00:00          536445.
##  5 2014-09-01 00:00:00          529254.
##  6 2014-10-01 00:00:00          539027.
```

```
##  7 2014-11-01 00:00:00          521961.
##  8 2014-12-01 00:00:00          524462.
##  9 2015-01-01 00:00:00          525871.
## 10 2015-02-01 00:00:00          507851.
## 11 2015-03-01 00:00:00          543977.
## 12 2015-04-01 00:00:00          561838.
## 13 2015-05-01 00:00:00          558127.
```

```
ggplot(house_date,aes(x=month,y=summary_variable))+
  geom_line(color='blue')+
  scale_x_datetime(breaks = house_date$month)+
  xlab("Data")+
  ylab("Preço")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




## b) Média do preço do ft² x Mês

Na primeira tabela podemos observar os preços médios do ft² para cada ID único e na segunda tabela os preços médios do ft² para todas as casas ao longo dos meses. O gráfico mostra a distribuição gráfica dos valores da segunda tabela. Podemos perceber que os preços apresentam poucas mudanças na maioria dos meses. Há uma mudança mais significativa após fevereiro de 2015, o que faz sentido dado que sabemos que nesse mesmo período houve um aumento nos valores totais das casas pelo gráfico anterior.

```r
#Resumo do preço médio do sqft^2 para cada ID
house_id <- house %>%
    group_by(id,sqft_living) %>%
    summarize(summary_price = mean(price)) %>%
    summarize(preco_medio_sqft = summary_price/sqft_living)
house_id
```

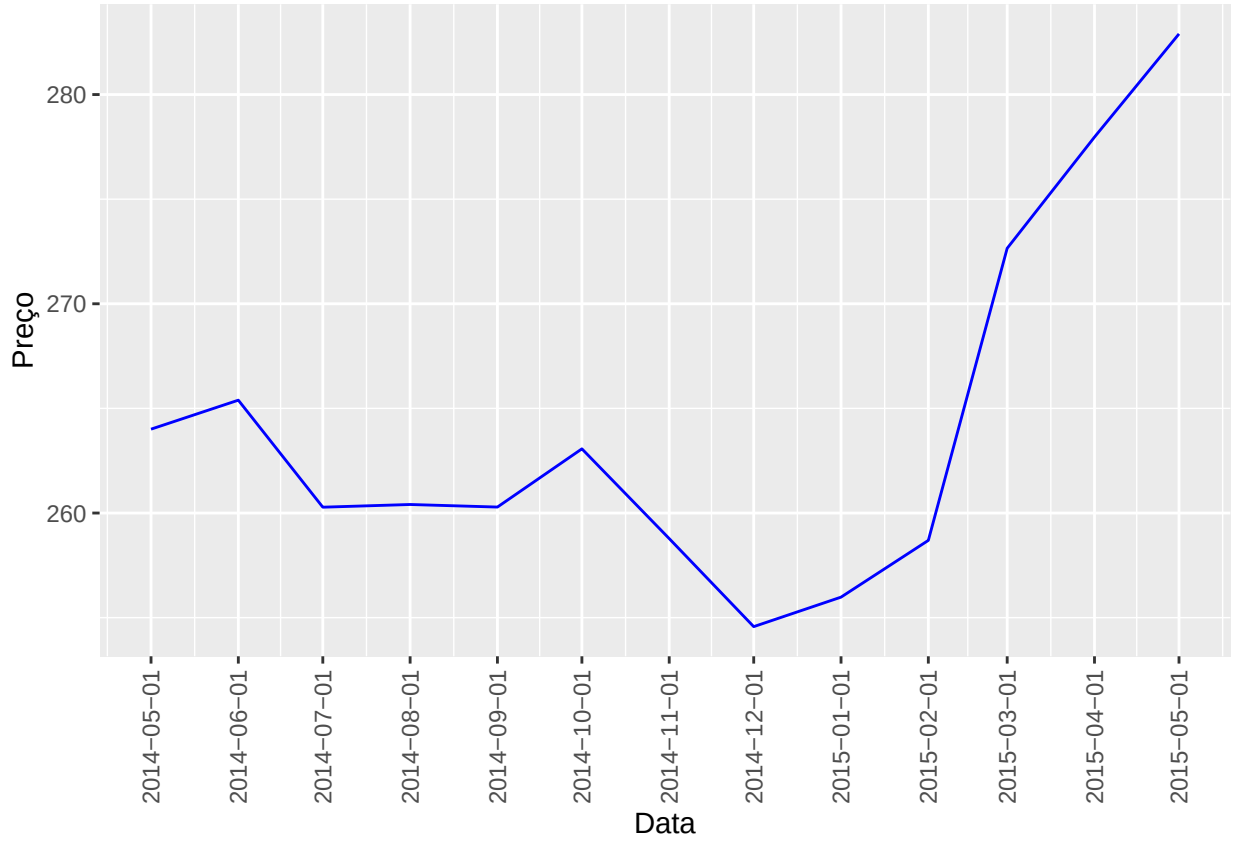
```
## # A tibble: 21,436 x 2
##    id         preco_medio_sqft
##    <chr>                 <dbl>
##  1 1000102                121.
##  2 100100050              208.
##  3 1001200035             202.
##  4 1001200050             206.
##  5 1003000175             226.
##  6 100300280              248.
##  7 100300500              219.
##  8 100300530              217.
##  9 1003400155             212.
## 10 1003400245             159.
## # ... with 21,426 more rows
```

```r
#Resumo do preço médio por sqft^2 ao longo do tempo
preco_medio_tempo_df <- merge(house_id,house2,by="id")
preco_medio_tempo <- preco_medio_tempo_df %>%
    group_by(month=floor_date(date2, "month")) %>%
    summarize(preco_medio_sqft=mean(preco_medio_sqft))

preco_medio_tempo
```

```
## # A tibble: 13 x 2
##    month               preco_medio_sqft
##    <dttm>                         <dbl>
##  1 2014-05-01 00:00:00             264.
##  2 2014-06-01 00:00:00             265.
##  3 2014-07-01 00:00:00             260.
##  4 2014-08-01 00:00:00             260.
##  5 2014-09-01 00:00:00             260.
##  6 2014-10-01 00:00:00             263.
##  7 2014-11-01 00:00:00             259.
##  8 2014-12-01 00:00:00             255.
##  9 2015-01-01 00:00:00             256.
## 10 2015-02-01 00:00:00             259.
## 11 2015-03-01 00:00:00             273.
## 12 2015-04-01 00:00:00             278.
## 13 2015-05-01 00:00:00             283.
```

```r
#Variação da média de preço do sqft^2 x mês
ggplot(preco_medio_tempo,aes(x=month,y=preco_medio_sqft))+
  geom_line(color='blue')+
  scale_x_datetime(breaks = preco_medio_tempo$month)+
  xlab("Data")+
```

```
  ylab("Preço")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




## Questão 5

## a) Quartis e Gráfico 3D

A tabela apresenta os quartis de interesse.

```
#Definindo os quantis
quantile(house2$price)
```

```
##      0%     25%     50%     75%    100%
##   75000  321950  450000  645000 7700000
```

```
quantis_preco_factors <- cut(house2$price, c(75000,321950,450000,645000,7700000), labels=c("Faixa 1","Fa
house3 <- cbind(house2,quantis_preco_factors)

#Associando as cores
nota <- sort(unique(house3$quantis_preco_factors))   # organizando em ordem crescente
m <- length(nota)   # tamanho do vetor
cores1 <- topo.colors(m)   # criando palheta
cores2 <- NULL
```

```
for(i in 1:dim(house3)[1]){
  cores2[i] <- cores1[nota==house3[i,24]]
}
s3d <- scatterplot3d(house3$lat,house3$long,house3$price,color=cores2,pch=16,angle=55,box=TRUE,scale.y=

legend(s3d$xyz.convert(47, -122.4, 16e+06),legend = levels(quantis_preco_factors),col =  unique(cores2)
```