Project2_proposalTask

Darreion Bailey

2024-04-13

Proposal

Question 1: How do sqft_living and bathrooms collectively influence the price of a house in King County?

This question is worth exploring for three reasons: 1) it can help to establish a predictive model for forecasting housing prices based on specific features, 2) it can help to validate or establish existing or new theories about the housing market, 3) it can help to explore social issues, such as housing affordability.

Question 2: What influence do house qualities, aside from the size of the house, have on whether a house is categorized as high price or low price per square foot ($/sqft)?

This question is worth exploring for three reasons: 1) it is a challenging way to represent and qualify quantitative and categorical data, which can give way to a new understanding of the housing market, 2) it can reveal market trends, which can affect urban planning, 3) it can influence the decision making process by real estate professionals, buyers, and sellers through understanding what factors can lead a house to be classified as high or low in terms of price per square footage.

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.92 loaded
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(tinytex)
```

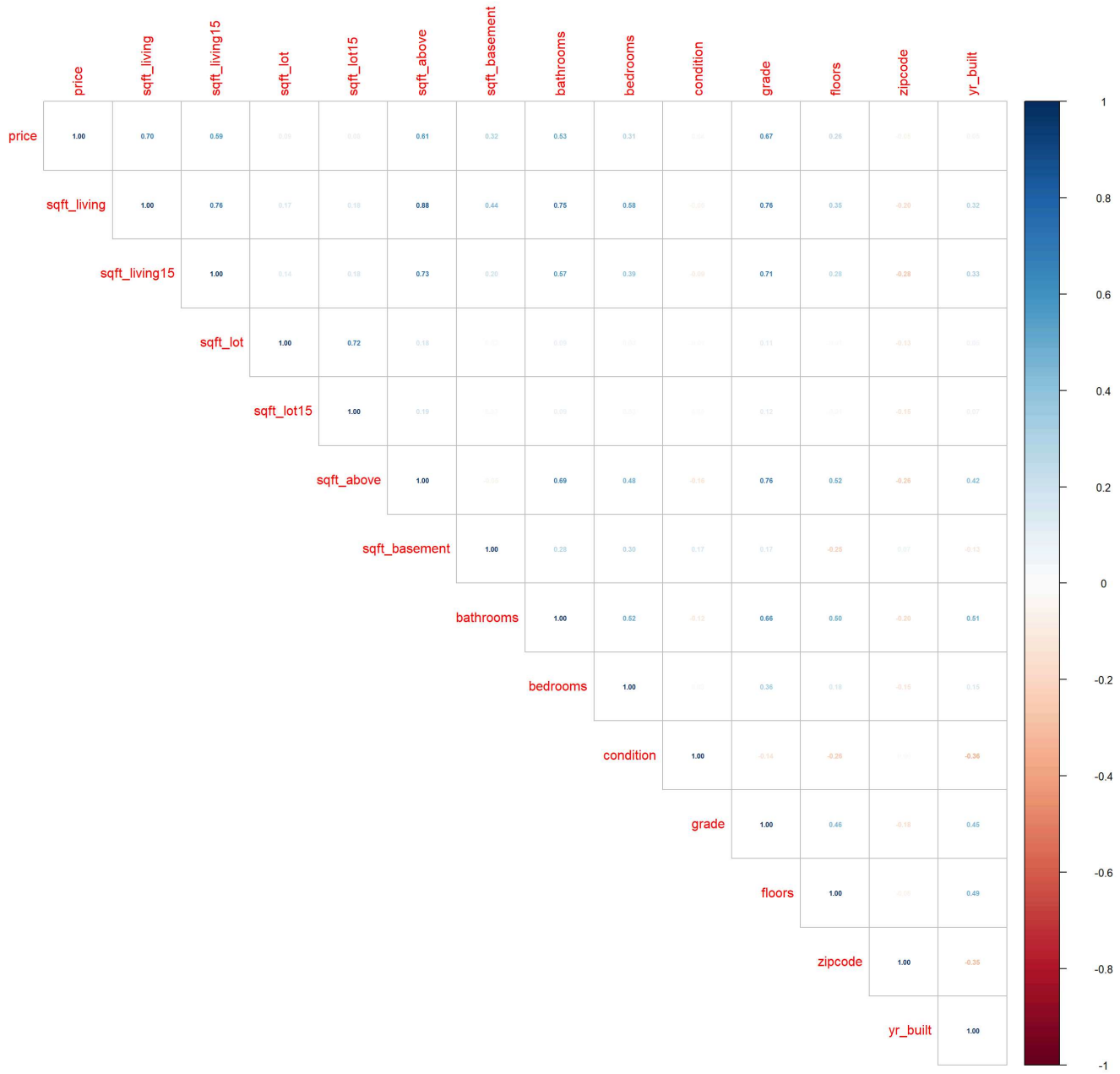
```
## Warning: package 'tinytex' was built under R version 4.3.3
```

```
df<-read.csv("kc_house_data.csv")
str(df)
colnames(df)

df %>%
  filter(waterfront!=0) # 163

df %>%
  filter(yr_renovated!=0) # 914

df %>%
  filter(view != 0) # 2124

# too small for our dataframe, so we can omit these variables; also omit long and lat variables, because such invest
igation is outside the scope of this course
```

```
df2<-select(df, "price", "sqft_living", "sqft_living15", "sqft_lot", "sqft_lot15", "sqft_above", "sqft_basement", "b
athrooms", "bedrooms", "condition", "grade", "floors", "zipcode", "yr_built")
colnames(df2)
```

```
##  [1] "price"         "sqft_living"   "sqft_living15" "sqft_lot"
##  [5] "sqft_lot15"    "sqft_above"    "sqft_basement" "bathrooms"
##  [9] "bedrooms"      "condition"     "grade"         "floors"
## [13] "zipcode"       "yr_built"
```

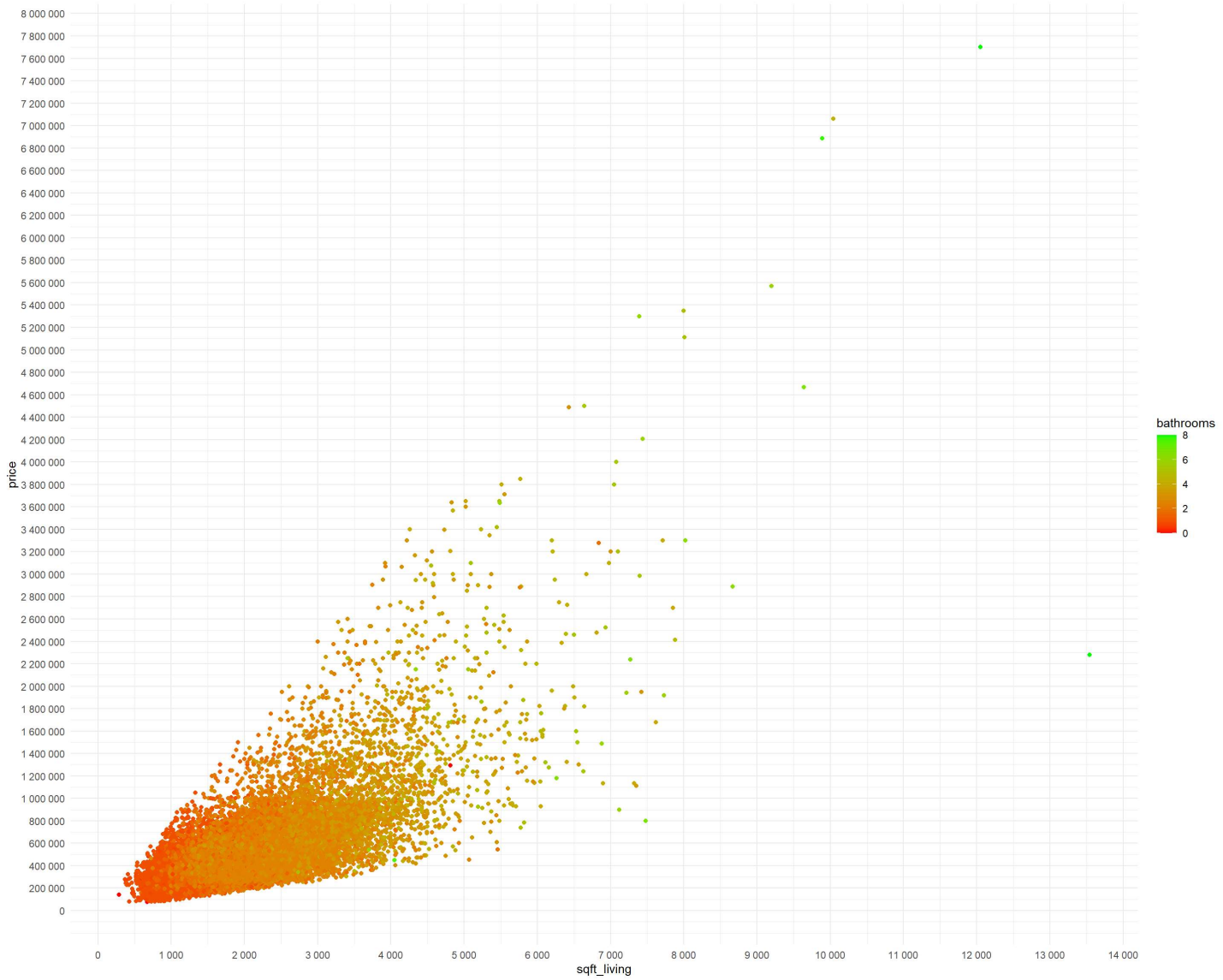
```
corrplot(cor(df2),
         method = "number",
         type = "upper",
         number.cex=0.50)
```




```
# From this correlation matrix, we see that price/sqft_living has a coefficient of 0.70, and an $R^2$ of 0.49
```

From this correlation matrix, we see that price/sqft_living has a coefficient of 0.70, and an R^2 of 0.49, meaning that 49% of the variation in price is explained by the variance in sqft_living.

```
ggplot(data=df2) +
  geom_point(mapping=aes(x=sqft_living, y=price, color=bathrooms)) +
  scale_x_continuous(labels=label_number(), breaks = seq(0, 20000, by = 1000)) +
  scale_y_continuous(labels=label_number(), breaks = seq(0, 8000000, by = 200000)) +
  theme_minimal() +
  scale_color_gradient(low="red", high = "green")
```




```
# bathrooms is the only other variable that makes sense in this context, and is the highest correlation of those tha
t make sense
```