

Project02

Group 08

2024-05-07

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

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

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v dplyr      1.1.4      v readr      2.1.5
```

```
## v forcats    1.0.0      v stringr   1.5.1
```

```
## v ggplot2    3.5.0      v tibble    3.2.1
```

```
## v lubridate  1.9.3      v tidyr     1.3.1
```

```
## v purrr      1.0.2
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Read in data and generate summaries

```
Data<-read.csv("kc_house_data.csv", sep=";", header=TRUE)
summary(Data)
```

```
##           id           date           price           bedrooms
## Min.      :1.000e+06   Length:21613   Min.       : 75000   Min.       : 0.000
## 1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000
## Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000
## Mean     :4.580e+09                                Mean     : 540088   Mean      : 3.371
## 3rd Qu.:7.309e+09                                3rd Qu.: 645000   3rd Qu.: 4.000
## Max.     :9.900e+09                                Max.     :7700000   Max.     :33.000
##   bathrooms   sqft_living   sqft_lot   floors
## Min.       :0.000   Min.      : 290   Min.      : 520   Min.      :1.000
## 1st Qu.:1.750   1st Qu.: 1427   1st Qu.: 5040   1st Qu.:1.000
## Median :2.250   Median : 1910   Median : 7618   Median :1.500
## Mean     :2.115   Mean      :2080   Mean      :15107   Mean     :1.494
## 3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:10688   3rd Qu.:2.000
## Max.     :8.000   Max.     :13540   Max.     :1651359   Max.     :3.500
##   waterfront   view   condition   grade
## Min.       :0.000000   Min.      :0.0000   Min.      :1.000   Min.      : 1.000
```

```
## 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.: 7.000
## Median :0.000000 Median :0.0000 Median :3.000 Median : 7.000
## Mean :0.007542 Mean :0.2343 Mean :3.409 Mean : 7.657
## 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:4.000 3rd Qu.: 8.000
## Max. :1.000000 Max. :4.0000 Max. :5.000 Max. :13.000
## sqft_above sqft_basement yr_built yr_renovated
## Min. : 290 Min. : 0.0 Min. :1900 Min. : 0.0
## 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951 1st Qu.: 0.0
## Median :1560 Median : 0.0 Median :1975 Median : 0.0
## Mean :1788 Mean : 291.5 Mean :1971 Mean : 84.4
## 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997 3rd Qu.: 0.0
## Max. :9410 Max. :4820.0 Max. :2015 Max. :2015.0
## zipcode lat long sqft_living15
## Min. :98001 Min. :47.16 Min. : -122.5 Min. : 399
## 1st Qu.:98033 1st Qu.:47.47 1st Qu.: -122.3 1st Qu.:1490
## Median :98065 Median :47.57 Median : -122.2 Median :1840
## Mean :98078 Mean :47.56 Mean : -122.2 Mean :1987
## 3rd Qu.:98118 3rd Qu.:47.68 3rd Qu.: -122.1 3rd Qu.:2360
## Max. :98199 Max. :47.78 Max. : -121.3 Max. :6210
## sqft_lot15
## Min. : 651
## 1st Qu.: 5100
## Median : 7620
## Mean : 12768
## 3rd Qu.: 10083
## Max. :871200
```

```
Data_cleaned <- na.omit(Data)
```

```
Data_cleaned$equal <- ifelse(Data_cleaned$sqft_above + Data_cleaned$sqft_basement == Data_cleaned$sqft_living15,
summary(Data_cleaned$equal)
```

```
## Length Class Mode
## 21613 character character
```

```
table(Data_cleaned$equal)
```

```
##
## yes
## 21613
```