# Kingston County Housing Data Wrangling

## This is a Noteboook designed to clean, wrangle and explore the Kingston Housing Dataset

### Loading Packages

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------ tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

### Loading dataset

```
df_kc <- read.csv("~/DataBEL/kc_house_data.csv")
```

### Checking out dataset

```
nrow(df_kc)
```

```
## [1] 21613
```

```
colnames(df_kc)
```

```
##  [1] "date"          "price"         "bedrooms"      "bathrooms"
##  [5] "sqft_living"   "sqft_lot"      "floors"        "waterfront"
##  [9] "view"          "condition"     "grade"         "sqft_above"
## [13] "sqft_basement" "yr_built"      "yr_renovated"  "zipcode"
## [17] "lat"           "long"          "sqft_living15" "sqft_lot15"
```

```r
head(df_kc, 20)
```

```
##              date   price bedrooms bathrooms sqft_living sqft_lot floors
## 1  20141013T000000  221900        3      1.00        1180     5650    1.0
## 2  20141209T000000  538000        3      2.25        2570     7242    2.0
## 3  20150225T000000  180000        2      1.00         770    10000    1.0
## 4  20141209T000000  604000        4      3.00        1960     5000    1.0
## 5  20150218T000000  510000        3      2.00        1680     8080    1.0
## 6  20140512T000000 1230000        4      4.50        5420   101930    1.0
## 7  20140627T000000  257500        3      2.25        1715     6819    2.0
## 8  20150115T000000  291850        3      1.50        1060     9711    1.0
## 9  20150415T000000  229500        3      1.00        1780     7470    1.0
## 10 20150312T000000  323000        3      2.50        1890     6560    2.0
## 11 20150403T000000  662500        3      2.50        3560     9796    1.0
## 12 20140527T000000  468000        2      1.00        1160     6000    1.0
## 13 20140528T000000  310000        3      1.00        1430    19901    1.5
## 14 20141007T000000  400000        3      1.75        1370     9680    1.0
## 15 20150312T000000  530000        5      2.00        1810     4850    1.5
## 16 20150124T000000  650000        4      3.00        2950     5000    2.0
## 17 20140731T000000  395000        3      2.00        1890    14040    2.0
## 18 20140529T000000  485000        4      1.00        1600     4300    1.5
## 19 20141205T000000  189000        2      1.00        1200     9850    1.0
## 20 20150424T000000  230000        3      1.00        1250     9774    1.0
##    waterfront view condition grade sqft_above sqft_basement yr_built
## 1           0    0         3     7       1180             0     1955
## 2           0    0         3     7       2170           400     1951
## 3           0    0         3     6        770             0     1933
## 4           0    0         5     7       1050           910     1965
## 5           0    0         3     8       1680             0     1987
## 6           0    0         3    11       3890          1530     2001
## 7           0    0         3     7       1715             0     1995
## 8           0    0         3     7       1060             0     1963
## 9           0    0         3     7       1050           730     1960
## 10          0    0         3     7       1890             0     2003
## 11          0    0         3     8       1860          1700     1965
## 12          0    0         4     7        860           300     1942
## 13          0    0         4     7       1430             0     1927
## 14          0    0         4     7       1370             0     1977
## 15          0    0         3     7       1810             0     1900
## 16          0    3         3     9       1980           970     1979
## 17          0    0         3     7       1890             0     1994
## 18          0    0         4     7       1600             0     1916
## 19          0    0         4     7       1200             0     1921
## 20          0    0         4     7       1250             0     1969
##    yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1             0   98178 47.5112 -122.257          1340       5650
## 2          1991   98125 47.7210 -122.319          1690       7639
## 3             0   98028 47.7379 -122.233          2720       8062
## 4             0   98136 47.5208 -122.393          1360       5000
## 5             0   98074 47.6168 -122.045          1800       7503
## 6             0   98053 47.6561 -122.005          4760     101930
## 7             0   98003 47.3097 -122.327          2238       6819
## 8             0   98198 47.4095 -122.315          1650       9711
```

```
## 9             0   98146 47.5123 -122.337          1780       8113
## 10            0   98038 47.3684 -122.031          2390       7570
## 11            0   98007 47.6007 -122.145          2210       8925
## 12            0   98115 47.6900 -122.292          1330       6000
## 13            0   98028 47.7558 -122.229          1780      12697
## 14            0   98074 47.6127 -122.045          1370      10208
## 15            0   98107 47.6700 -122.394          1360       4850
## 16            0   98126 47.5714 -122.375          2140       4000
## 17            0   98019 47.7277 -121.962          1890      14018
## 18            0   98103 47.6648 -122.343          1610       4300
## 19            0   98002 47.3089 -122.210          1060       5095
## 20            0   98003 47.3343 -122.306          1280       8850
```

## Graphs

```
unique(df_kc$floors)
```

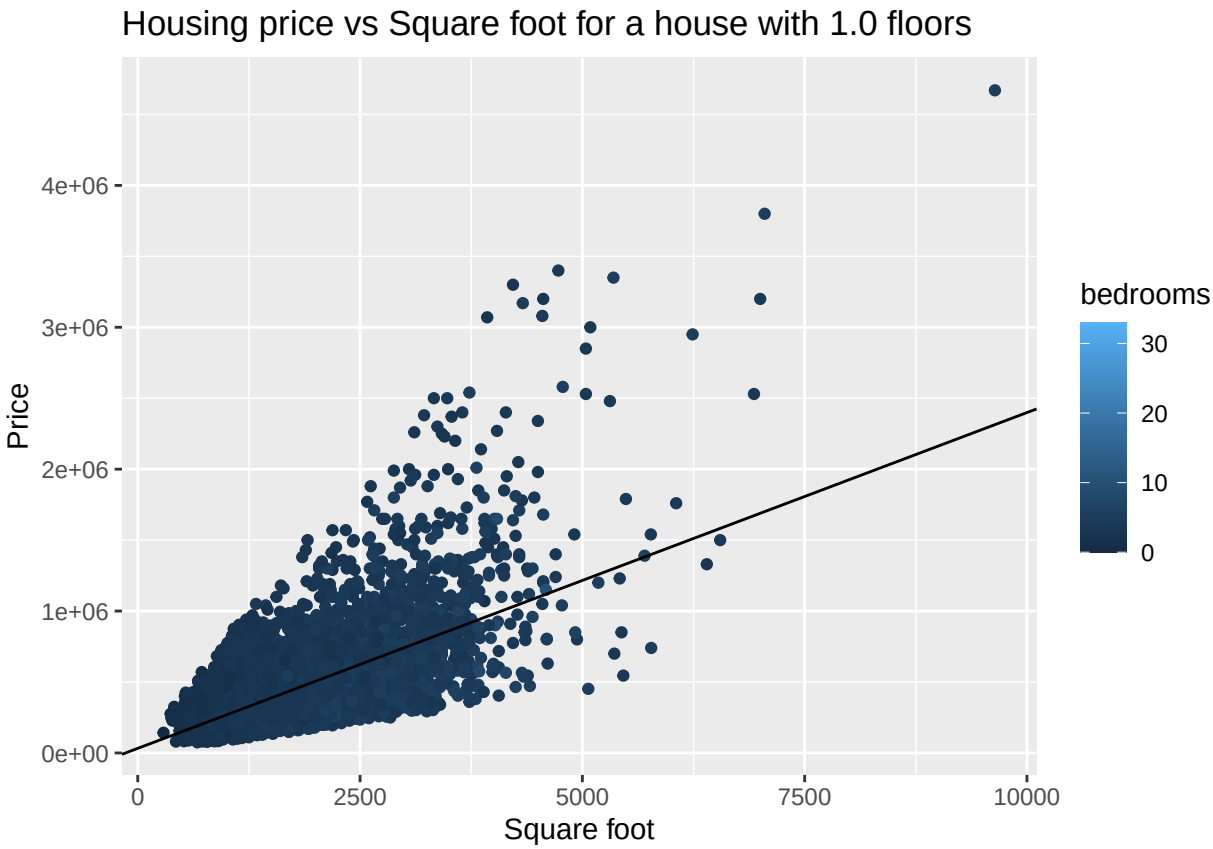
```
## [1] 1.0 2.0 1.5 3.0 2.5 3.5
```

### Housing Price vs Square foot

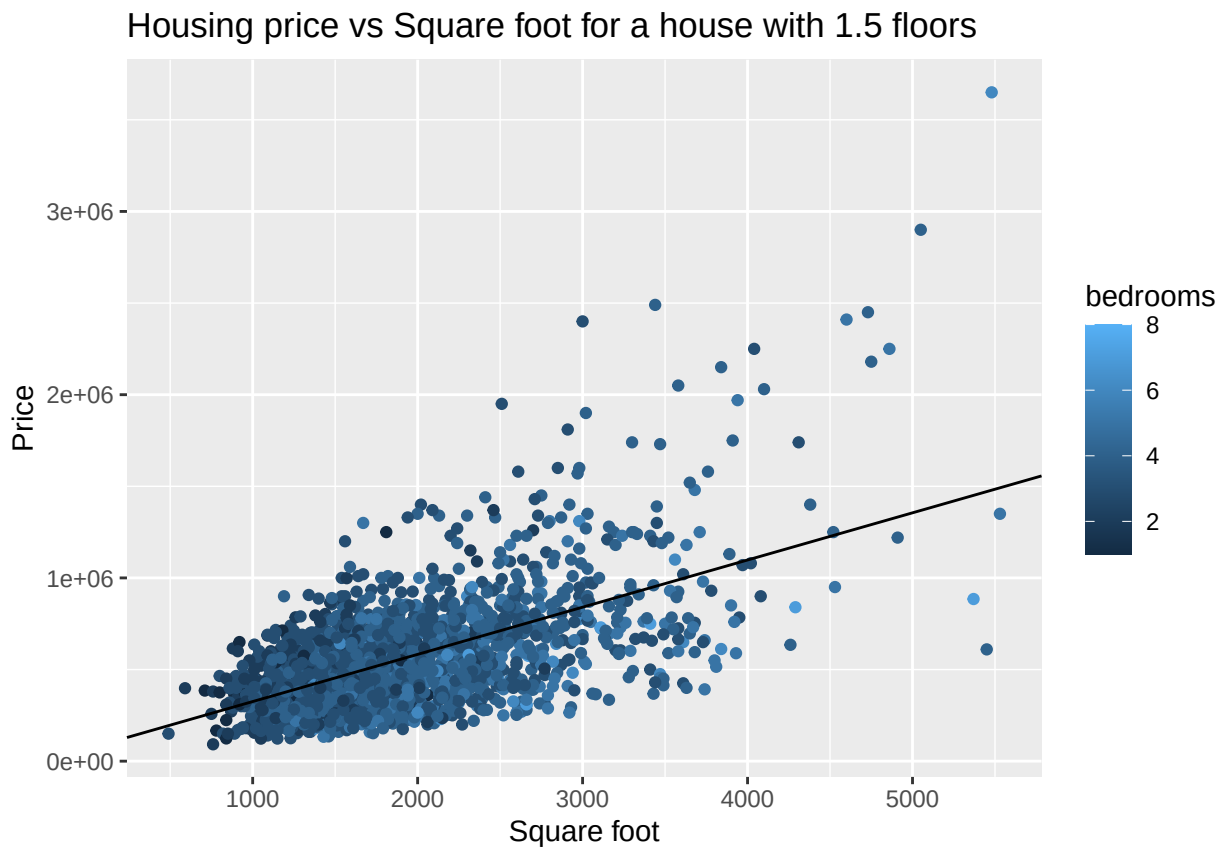
```
print_out <- function(df, x) {
  df_kc2 <- df_kc %>%
    filter(floors == x)
  mod <- lm(price ~ sqft_living, data = df_kc2)
  coeff <- coef(mod)
  ggplot(df_kc2, aes(x = sqft_living,y = price)) +
    geom_point(aes(color = bedrooms)) +
    geom_abline(intercept = coeff[1], slope = coeff[2]) +
    ##facet_wrap(~floors, ncol = 2) +
    labs(title = sprintf("Housing price vs Square foot for a house with %.1f floors", x), x = "Square fo
         y = "Price")
}
```

### Applying the code above

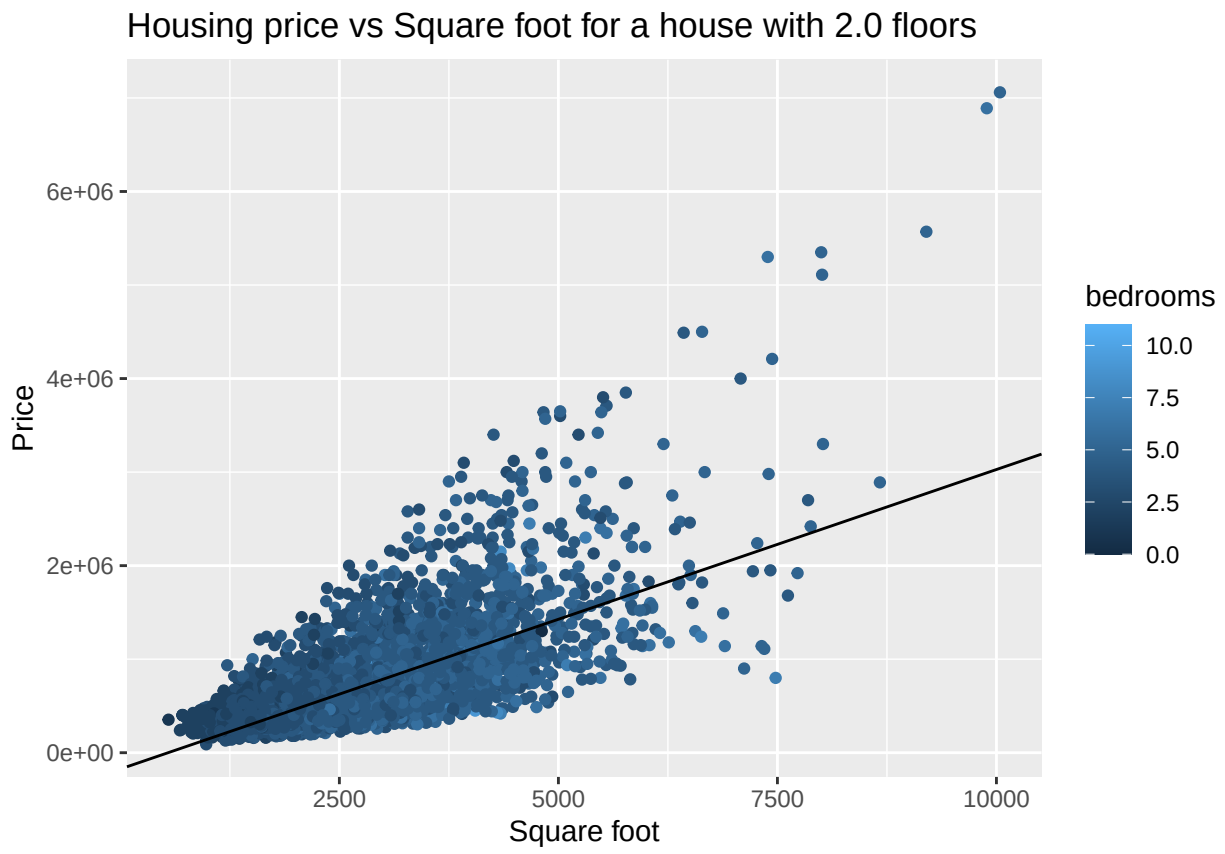
```
par(mfrow = c(3,2))
print_out(df_kc, 1)
```



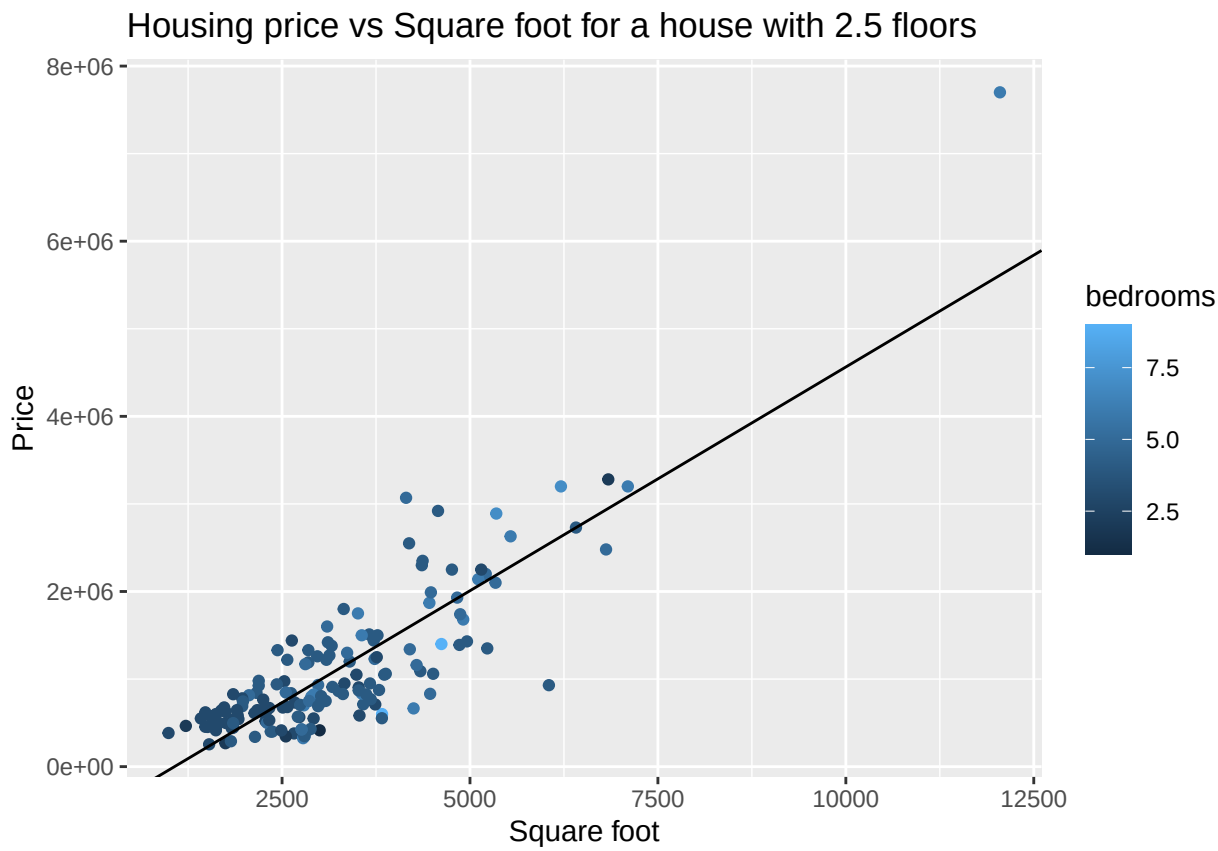

```
print_out(df_kc, 1.5)
```



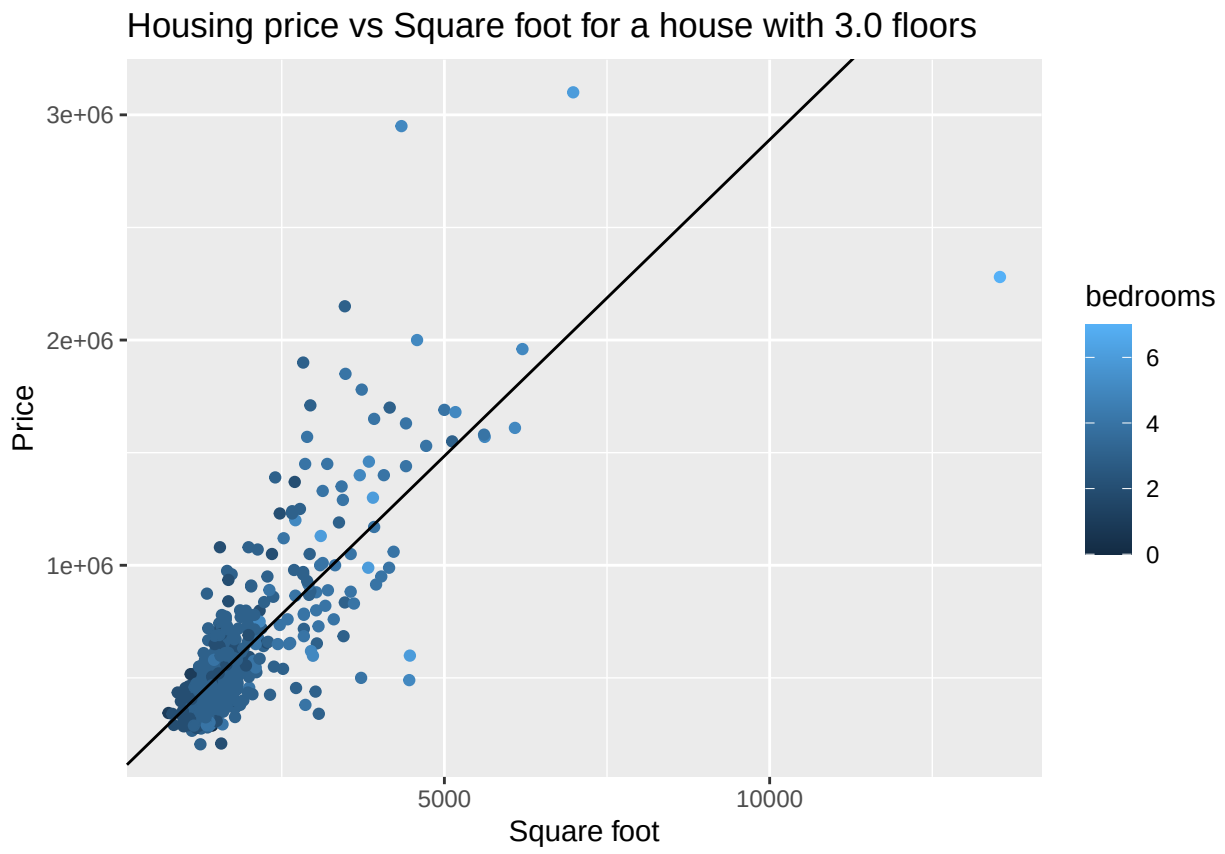

```
print_out(df_kc, 2.0)
```



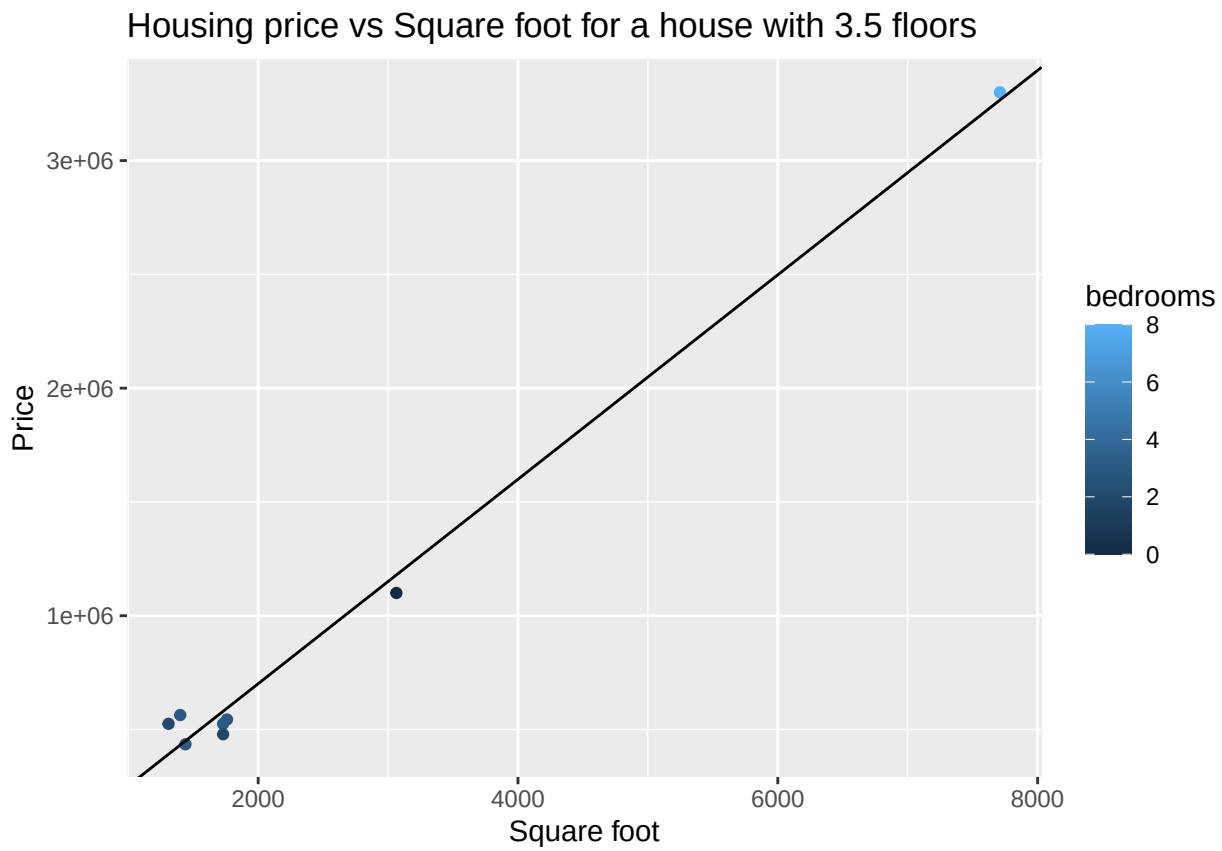

```
print_out(df_kc, 2.5)
```




```
print_out(df_kc, 3.0)
```



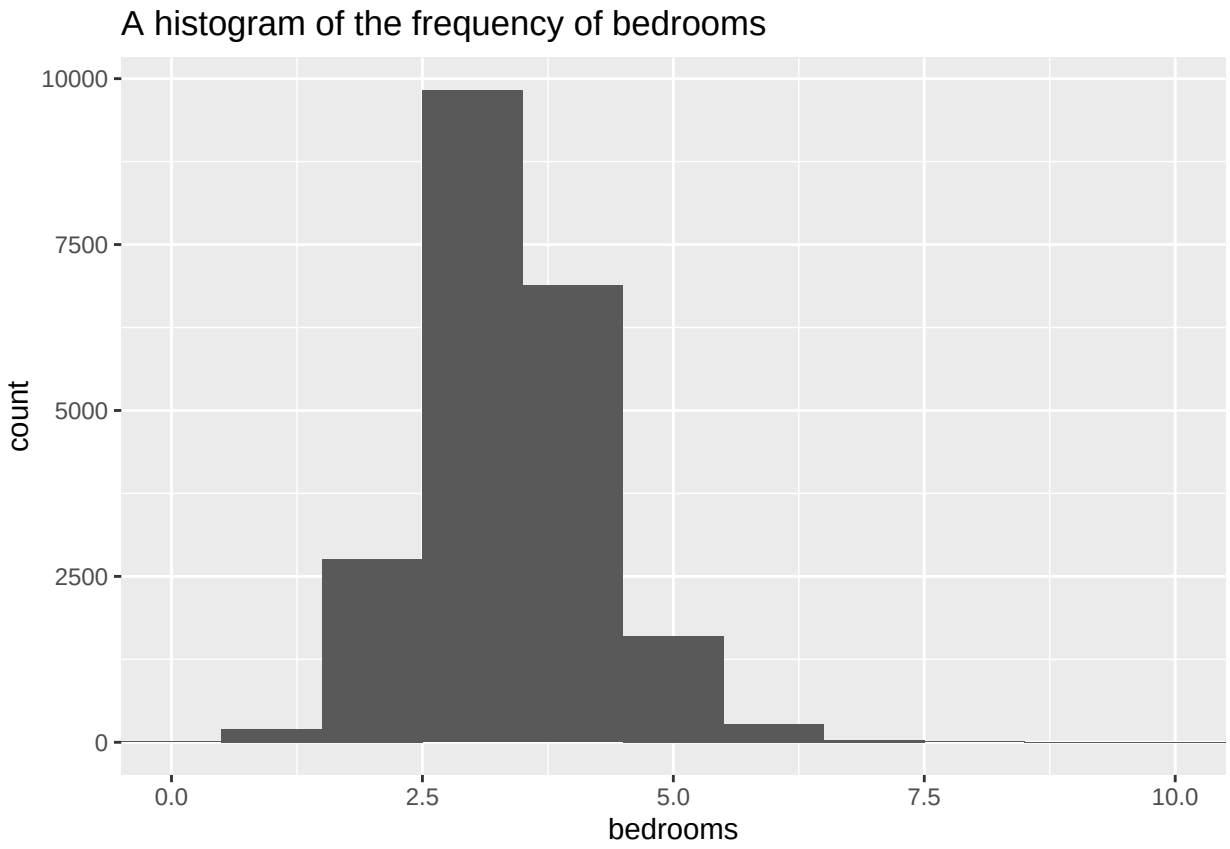

```
print_out(df_kc, 3.5)
```




```
par(mfrow = c(1,1))
```

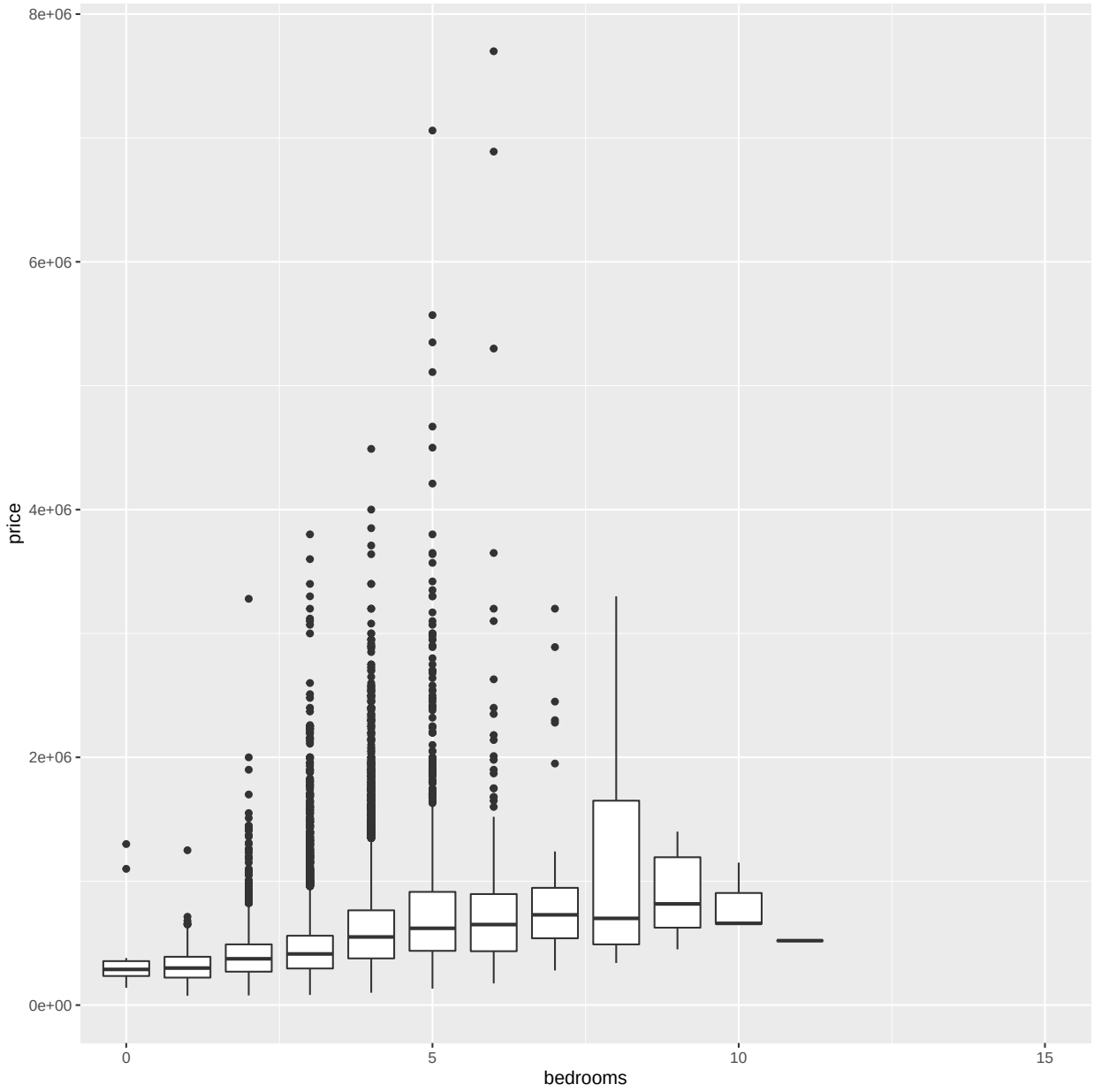
### Histogram of Bedrooms

```
ggplot(df_kc) +
  geom_histogram(aes(x = bedrooms), binwidth = 1) +
  coord_cartesian(xlim = c(0,10)) +
  labs(title = "A histogram of the frequency of bedrooms")
```




**A boxplot of bedrooms' prices**

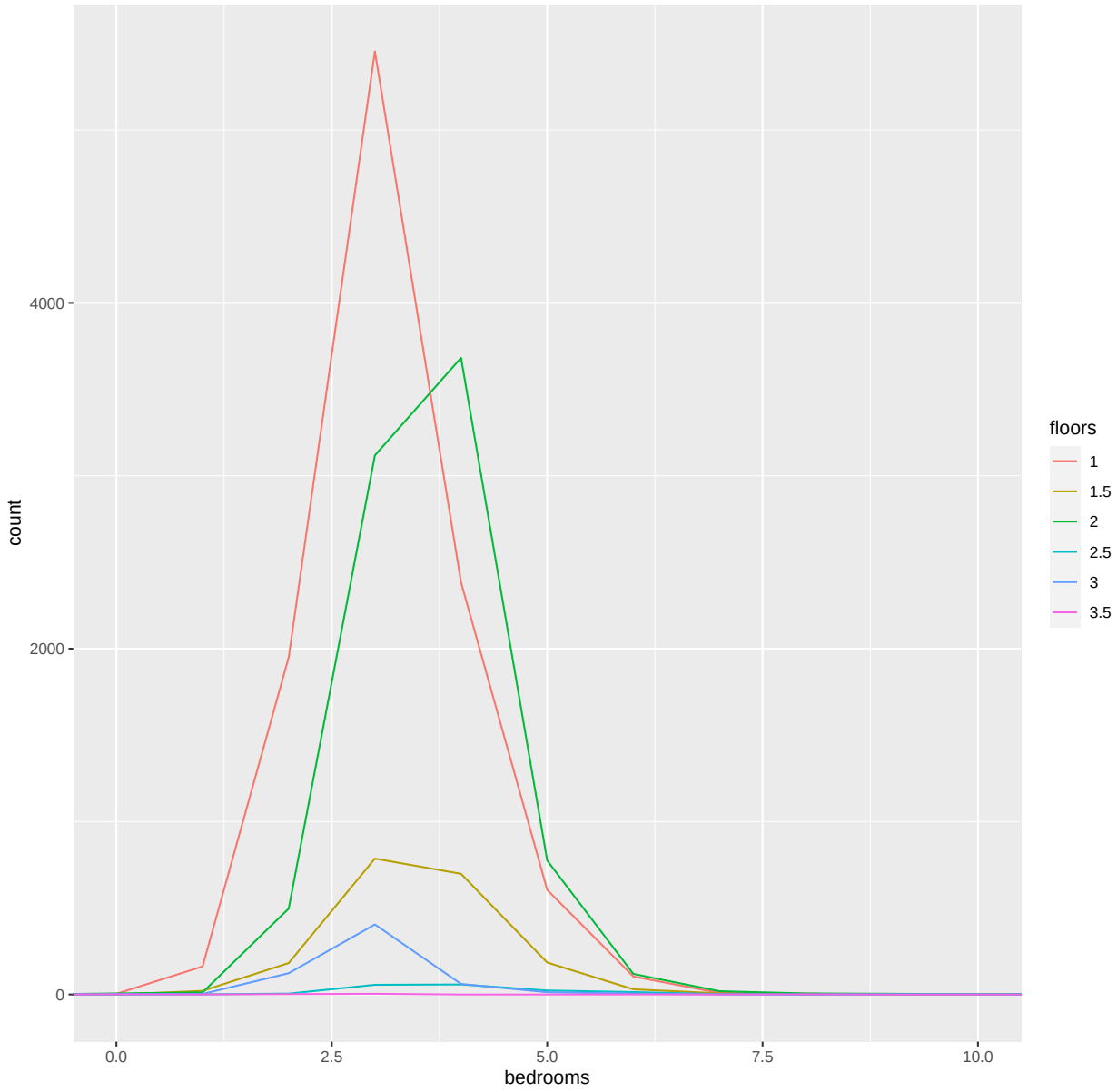
```
ggplot(df_kc) +
  geom_boxplot(aes(x = bedrooms, y = price, group = bedrooms)) +
  coord_cartesian(xlim = c(0, 15))
```




**A frequency polygon plot for bedrooms**

```
df_kc$floors <- as.factor(df_kc$floors)
ggplot(df_kc) +
  geom_freqpoly(aes(x = bedrooms, color = floors), binwidth = 1) +
  coord_cartesian(xlim = c(0,10)) +
  labs(title = "A frequency ploygon plot of bedrooms based on the number of floors")
```

A frequency ploygon plot of bedrooms based on the number of floors




```
sprintf("The house with the most bedrooms has %d bedrooms", max(df_kc$bedrooms))
```

```
## [1] "The house with the most bedrooms has 33 bedrooms"
```

```
sprintf("The most expensive house costs $%d" , max(df_kc$price))
```

```
## [1] "The most expensive house costs $7700000"
```