Airbnb Pricing Analysis

Ethan Chu

2025-05-06

Import dataset

```
library(ggplot2)
library(readr)
airbnb_sf <- read_csv("/Users/ethanchu/Desktop/airbnb_sf.csv")
```

```
## New names:
## Rows: 7831 Columns: 69
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (20): last_scraped, source, host_name, host_since, host_location, host_r... dbl
## (42): id, host_id, host_listings_count, host_total_listings_count, latit... lgl
## (7): host_is_superhost, host_has_profile_pic, host_identity_verified, n...
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...31‘
```

Remove outliers from price column

```
filtered_data <- subset(airbnb_sf, price <= 1000)
```

Finding trends using three linear regression models

```
ols <- lm(price ~ bathrooms, data=filtered_data)
summary(ols)
```

```
##
## Call:
## lm(formula = price ~ bathrooms, data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -689.80  -89.91  -36.91   51.09  881.30
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.696      4.017   29.55   <2e-16 ***
## bathrooms     61.211      2.685   22.80   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 150.2 on 5923 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.08067,    Adjusted R-squared:  0.08052
## F-statistic: 519.7 on 1 and 5923 DF,  p-value: < 2.2e-16
```

```r
ols2 <- lm(price ~ bathrooms + bedrooms, data=filtered_data)
summary(ols2)
```

```
##
## Call:
## lm(formula = price ~ bathrooms + bedrooms, data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -741.02  -90.29  -31.29   50.90  850.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.016      3.809  19.169  < 2e-16 ***
## bathrooms     16.080      2.693   5.971  2.5e-09 ***
## bedrooms      76.192      2.019  37.730  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 134.9 on 5906 degrees of freedom
##   (21 observations deleted due to missingness)
## Multiple R-squared:  0.2593, Adjusted R-squared:  0.2591
## F-statistic:  1034 on 2 and 5906 DF,  p-value: < 2.2e-16
```

```r
ols3 <- lm(price ~ bathrooms + bedrooms + accommodates, data=filtered_data)
summary(ols3)
```

```
##
## Call:
## lm(formula = price ~ bathrooms + bedrooms + accommodates, data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -534.66  -71.84  -27.28   41.87  868.65
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.670      3.777   8.913  < 2e-16 ***
## bathrooms      10.245      2.514   4.075 4.66e-05 ***
## bedrooms       22.465      2.589   8.679  < 2e-16 ***
## accommodates   37.609      1.246  30.189  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 125.6 on 5905 degrees of freedom
##   (21 observations deleted due to missingness)
## Multiple R-squared:  0.3584, Adjusted R-squared:  0.358
## F-statistic:  1099 on 3 and 5905 DF,  p-value: < 2.2e-16
```

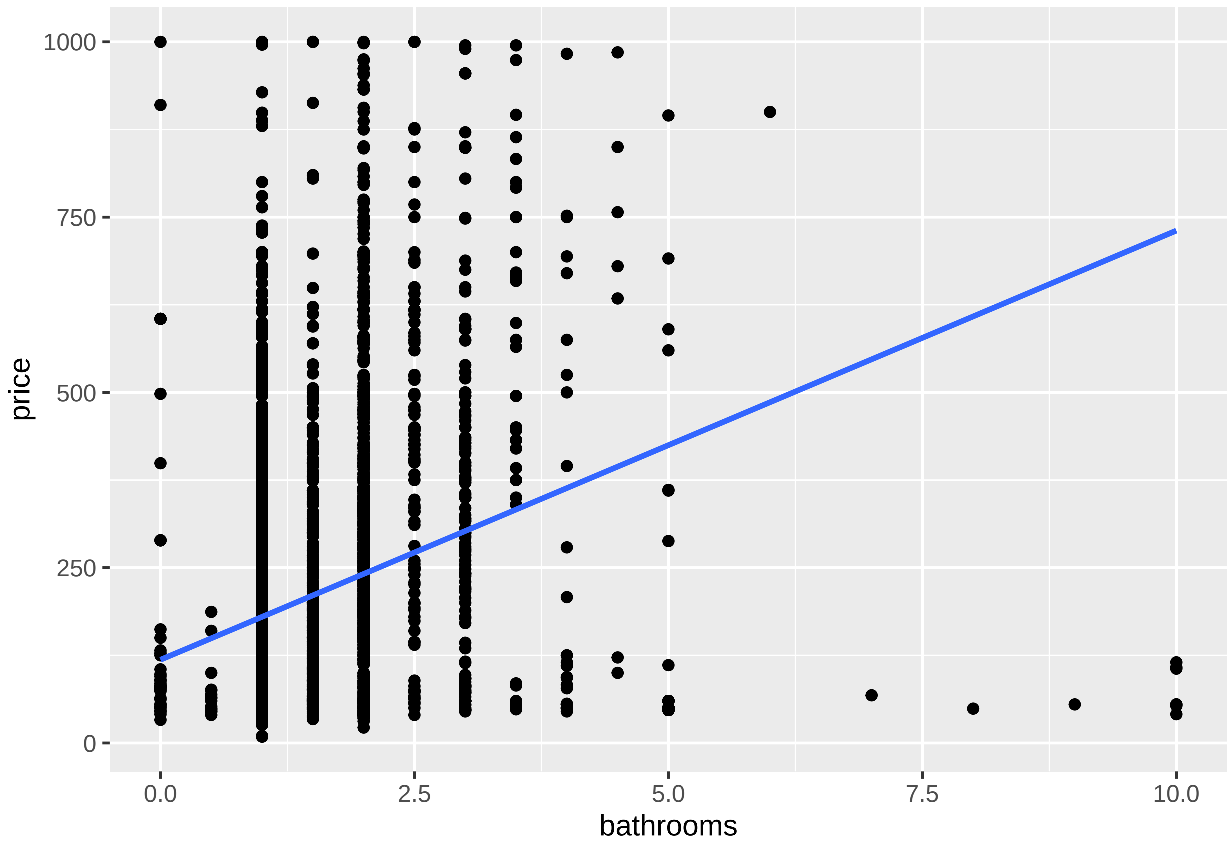
Plotting data and regression line

```
ggplot(filtered_data, aes(x = bathrooms, y = price)) + geom_point() + geom_smooth(method = "lm", se = F
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).
```

```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```


price
bathrooms