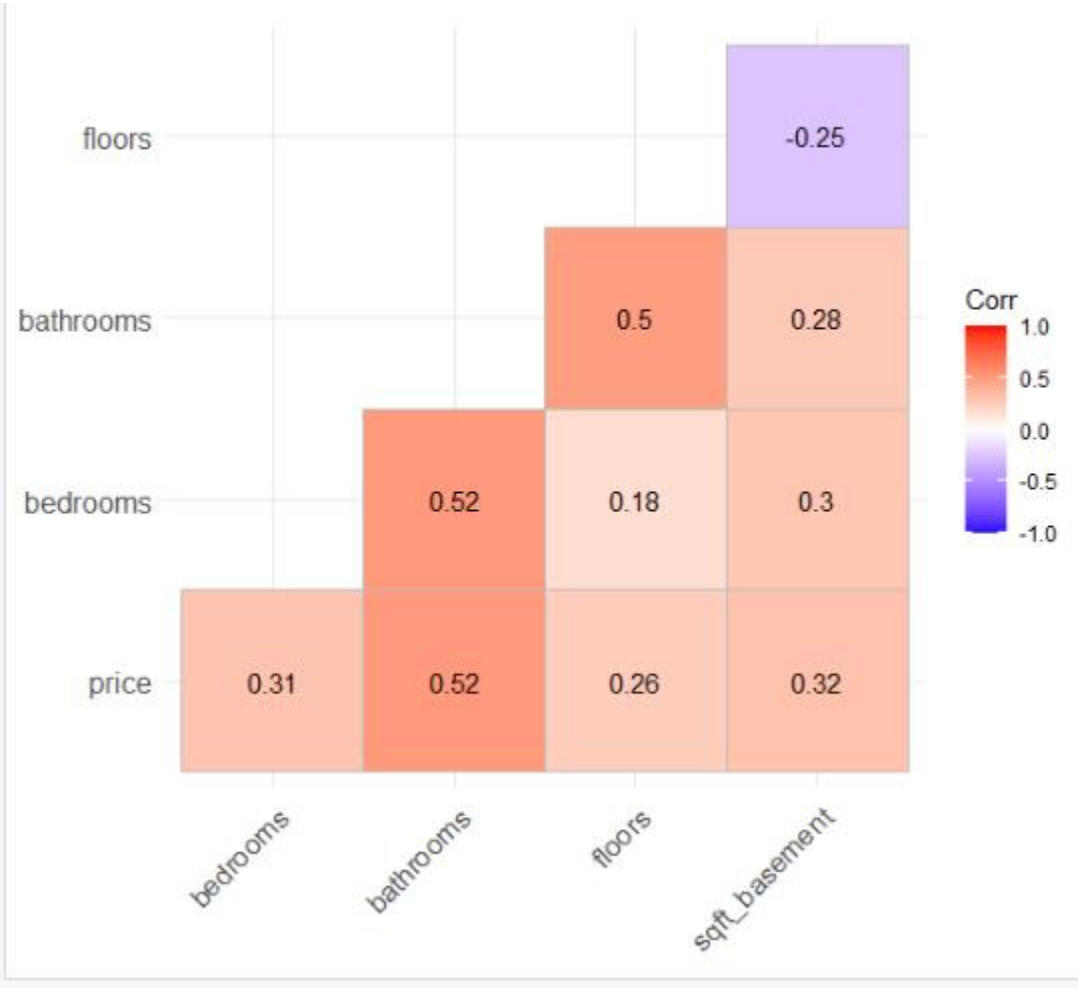
Yijun Wang

Practice 5

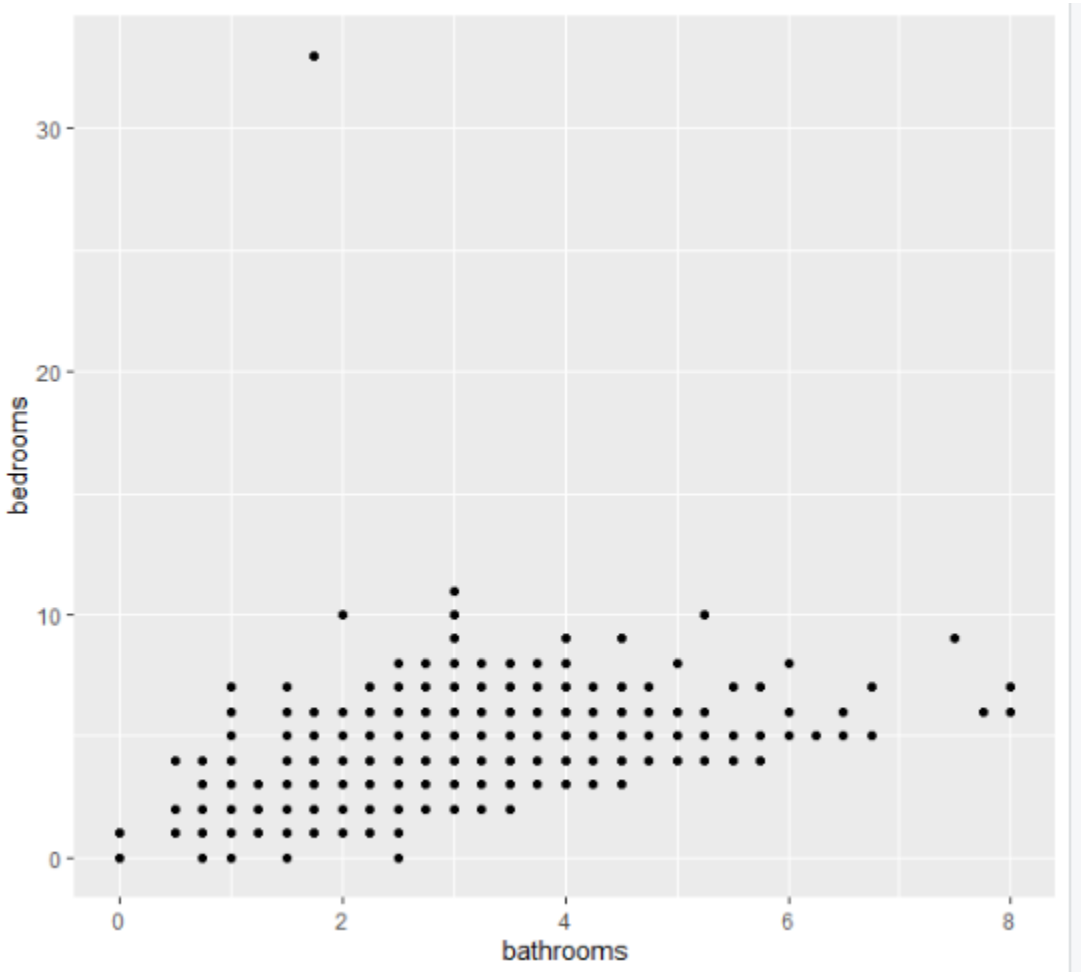
12/09/2021

1. The correlation coefficient reflects the relationship between two variables. I display a plot of two variables with a certain correlation in this manner. We can see the degree of connection between each variable via the color tones.



As we can see from the graph, the relationship between sb and floor is the smallest, because the color is the least, and the correlation of the number is -0.25. Bathroom and bedroom, bathroom

and price are the most closely related, because the color is the darkest, and the correlation of the number is 0.52. Next, the relationship between floor and bathrooms is more related, and the correlation of the number is 0.5. The correlation of the numbers is 0.5.



2.

```
> summary(Lm)

Call:
lm(formula = df, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1111740  -178634   -36945   113458  5620989

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -49009.715   8973.544  -5.462 4.77e-08 ***
bedrooms        5353.610   2640.091   2.028   0.0426 *
bathrooms     187350.320   3863.226  48.496  < 2e-16 ***
floors         78719.905   4996.048  15.756  < 2e-16 ***
sqft_basement    196.287      5.571  35.235  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303500 on 21608 degrees of freedom
Multiple R-squared:  0.3169,    Adjusted R-squared:  0.3168
F-statistic:  2506 on 4 and 21608 DF,  p-value: < 2.2e-16
```

R-squared, often known as R2, describes the extent to which your input variables explain the variation in your output / anticipated variable. As a result, if we are creating Linear regression on many variables, we should always use Adjusted R-squared to estimate the validity of the model.

As we can see, R-square is 0.31, it means 31% of the variation in the output variable is explained by the input variables. Multiple R-squared is bigger than Adjusted R-squared 0.0001. p value is <2.2e-16, which is less than the alpha value 0.05. The Null hypothesis is not related to these, so we can reject the null hypothesis.

Difference：Correlation and regression are the two multivariate distribution-based analyses. A multivariate distribution is a distribution with many variables. Correlation is defined as the analysis that determines the existence or absence of a link between two variables 'x' and 'y'. Regression analysis, on the other hand, predicts the value of the dependent variable based on the known value of the independent variable, assuming that there is an average mathematical connection between two or more variables. In conclusion, correlation measures the strength of link between variables. Regression, on the

other hand, depicts the influence of a unit change in the independent variable on the dependent variable. (S, 2021)

Bibliography

S, S. (2021, 02 26). *Difference Between Correlation and Regression*. Retrieved from keydifferences: https://keydifferences.com/difference-between-correlation-and-regression.html