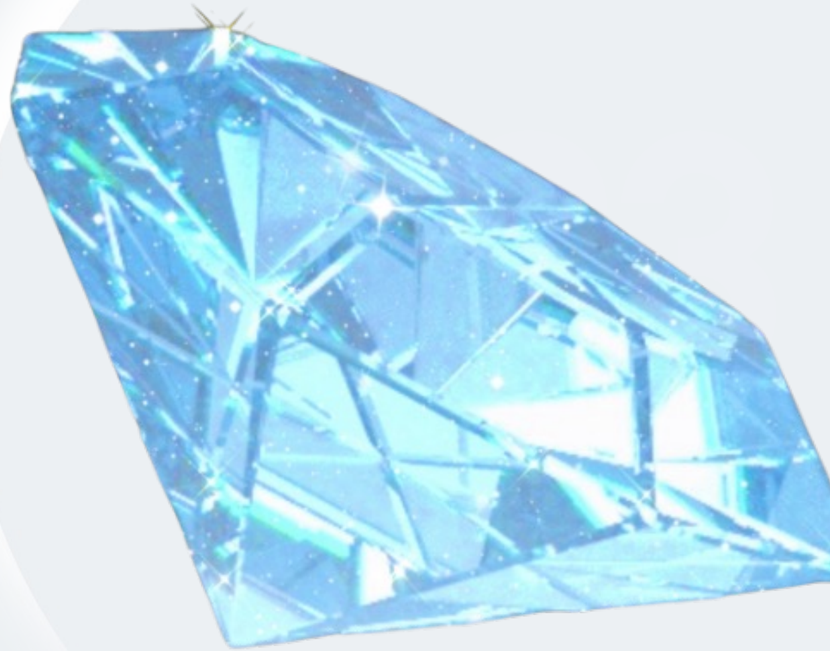
# Diamond Prices Prediction

## using Regression Analysis



Group 7 Xu Ruiqi
Qu Zhemin
Ling Haoyang

Date: 07/22/2023

# CONTENTS



01

# Introduction

Background, problem description, dataset introduction

# Diamond pricing needs higher accuracy

**High influence on profits**

Diamonds are valuable due to rarity and beauty in the jewelry industry, thus having a competitive market.

Historical decisions are often based on intuition and market trends, which is potentially unfair.

**Unfair subjective decisions**

# Diamond pricing needs higher accuracy

**A MODEL THAT CAN PREDICT DIAMOND PRICES ACCURATELY**

# Diamond Dataset

Source: https://www.kaggle.com/datasets/shivam2503/diamonds.

Samples: 53909

02

# Data Preprocessing

Data cleaning, feature engineering, transformations

# Data preprocessing steps

## Data Cleaning

- Observations with non-positive values for the dimensions (x, y, z) and carat weight are removed

## One-hot Encoding for Categorical Variables

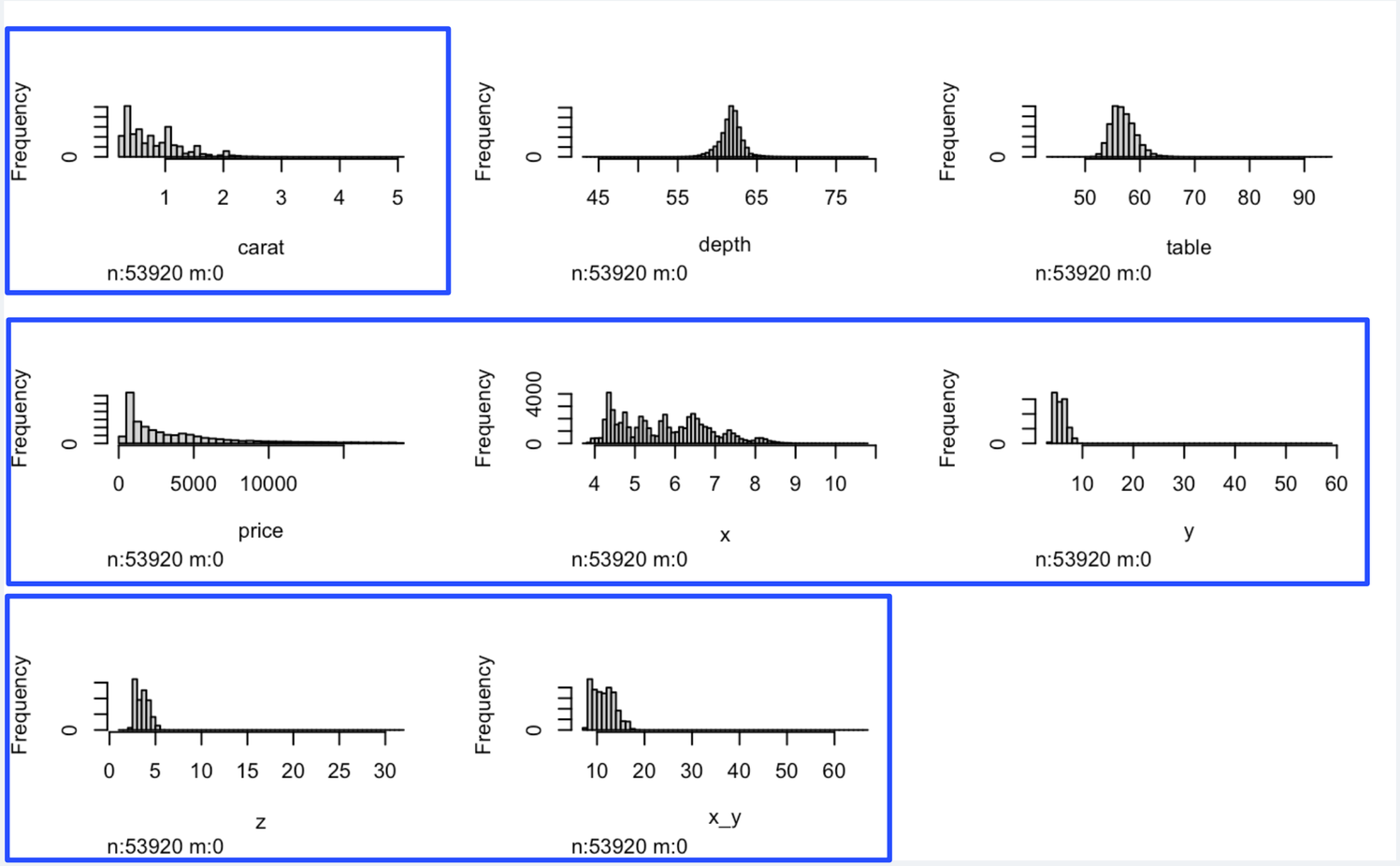
- Each category within variables (“cut”, “color”, “clarity”) is converted into a binary variable

## New Term Adding

- Depth = 2z/(x+y), so a new variable x_y = x+y is added to the dataset

# Transformation for Numerical Variables

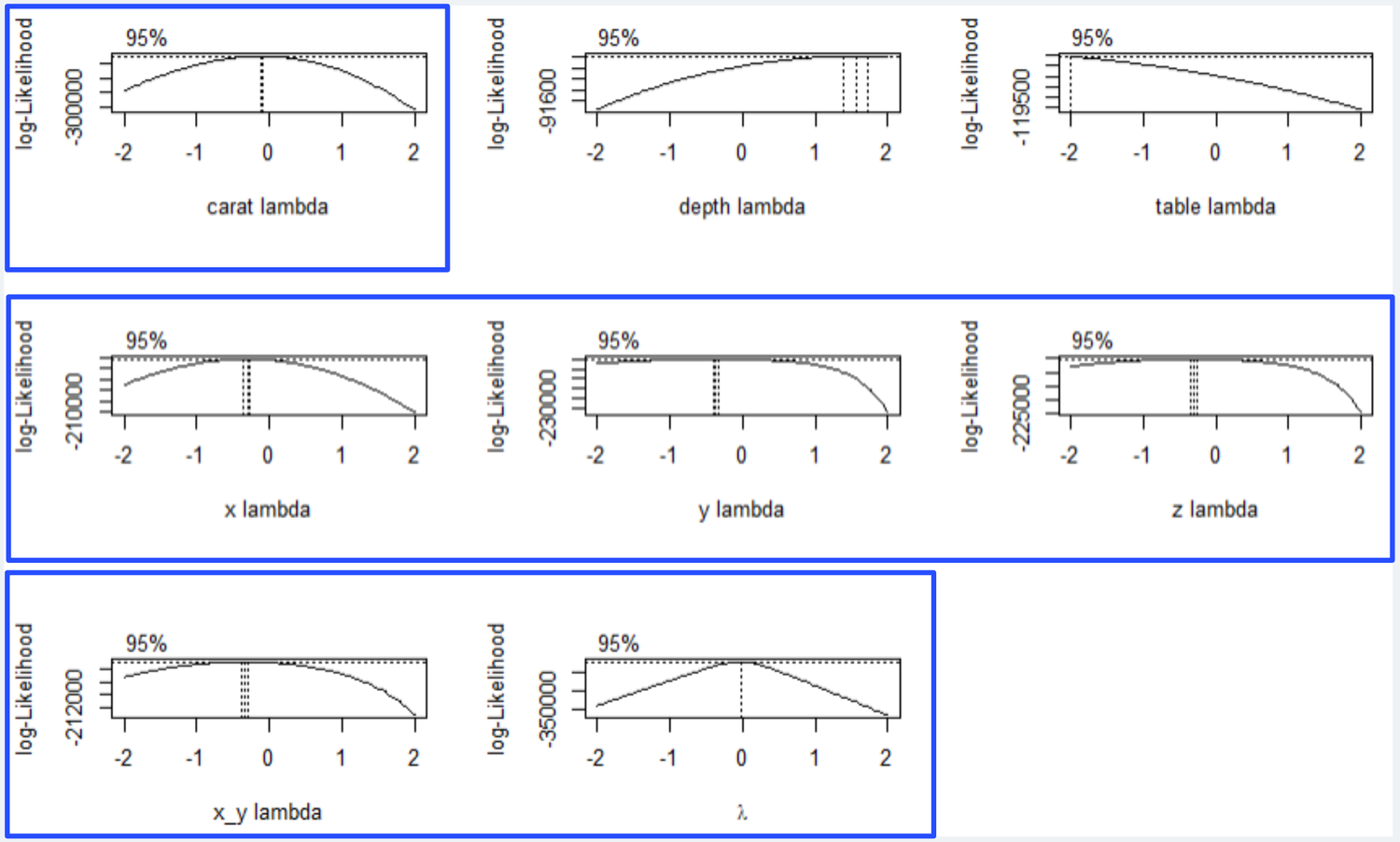
- **Distribution for each variable is plotted**




**Right-skewed!**

# Transformation for Numerical Variables

- **Box-Cox transformation is applied to address skewness**




**Log transformation for variables whose $\lambda$ is 0**

03

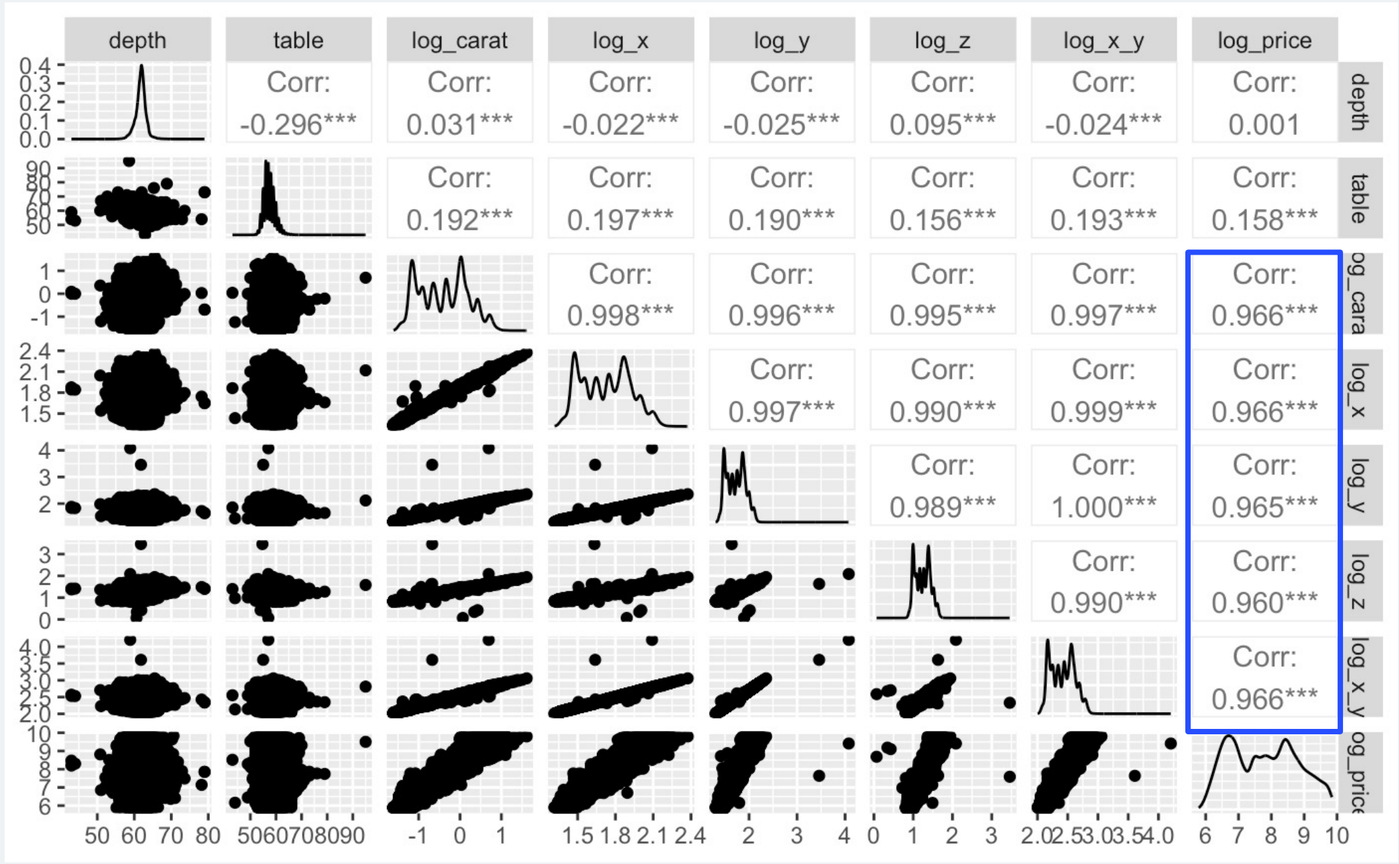
# Model Design And Analysis

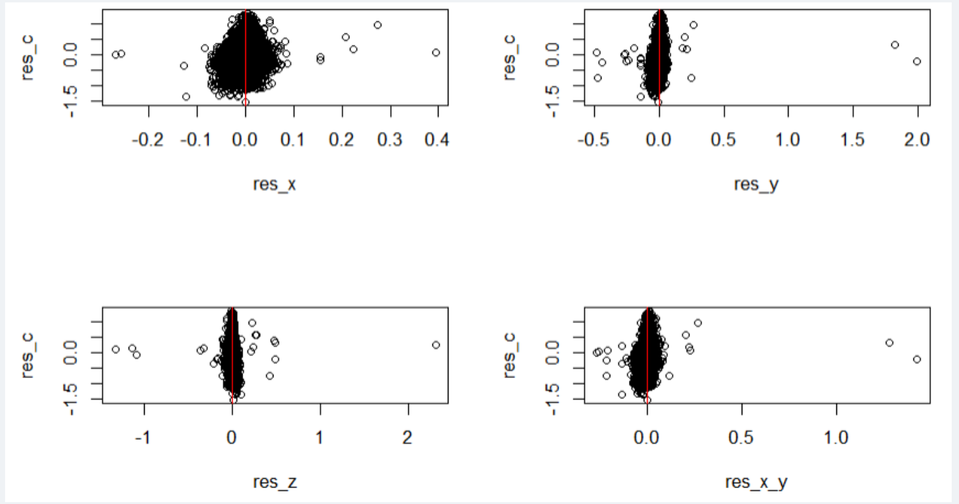
Feature selection, categorical variables, interaction terms, model fitting and evaluation

# Feature Selection

## based on multicollinearity

Variables x, y, z, and x_y can be dropped




| Variable | VIF |
|---|---|
| log_carat | 515.8976 |
| log_x | 395.6284 |
| log_y | 164.8624 |
| log_z | 111.1384 |

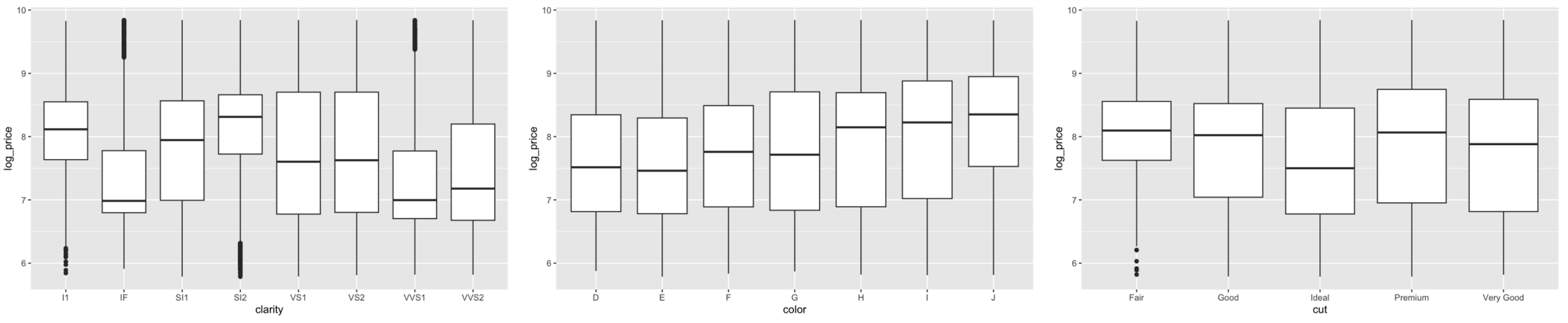
# Feature Selection

## based on Bayesian Information Criterion (BIC)

→ Variables depth and table can be dropped

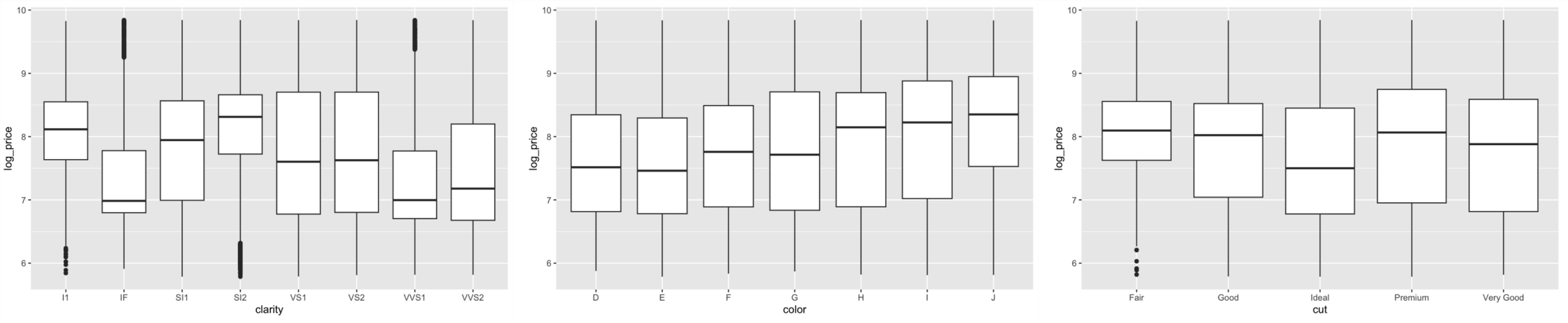
```
Step:  AIC=-216738.6
log_price ~ log_carat + clarity + color + cut + carat

         Df Sum of Sq    RSS     AIC F value  Pr(>F)
<none>                 964.47 -216739
+ depth   1  0.051143 964.42 -216731  2.8583 0.09091 .
+ table   1  0.000848 964.47 -216728  0.0474 0.82765
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Boxplots for Categorical Variables




**Lower color grade**
**but higher price!**

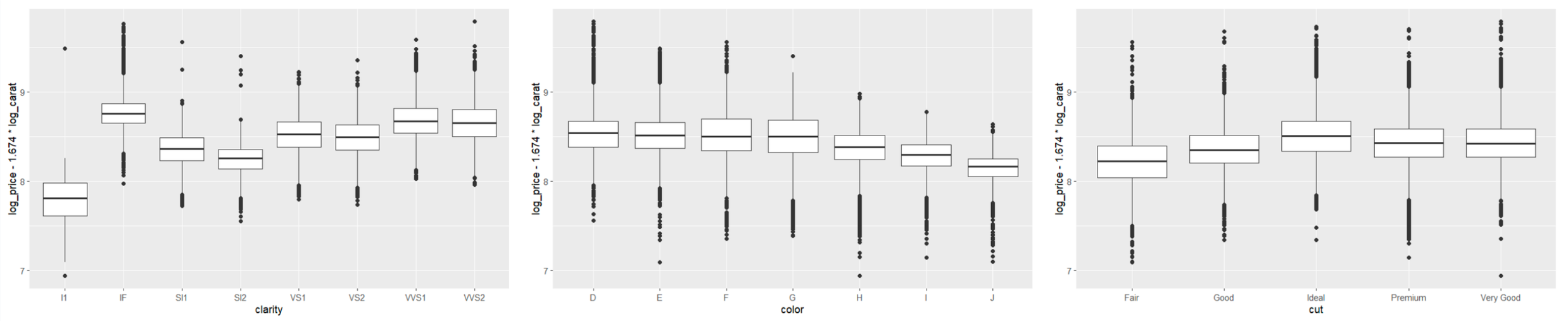
# Boxplots for Categorical Variables




**Lower color grade but higher price!**

Interactions between categorical variables and carat

# Boxplots for Categorical Variables

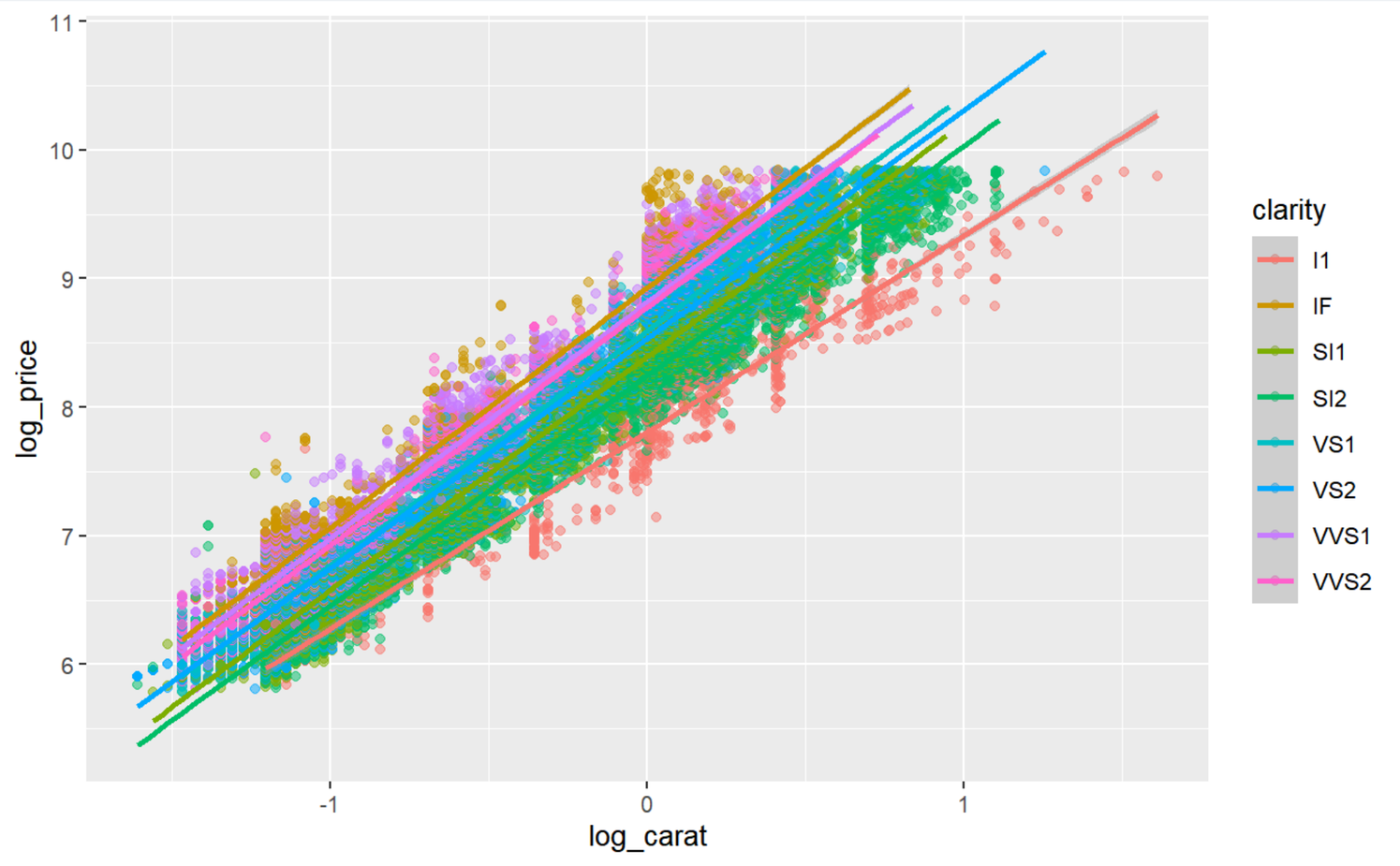



substituting the y-axis of with log_price−1.674*log_carat

fitted linear parameter “log price”~“log carat”

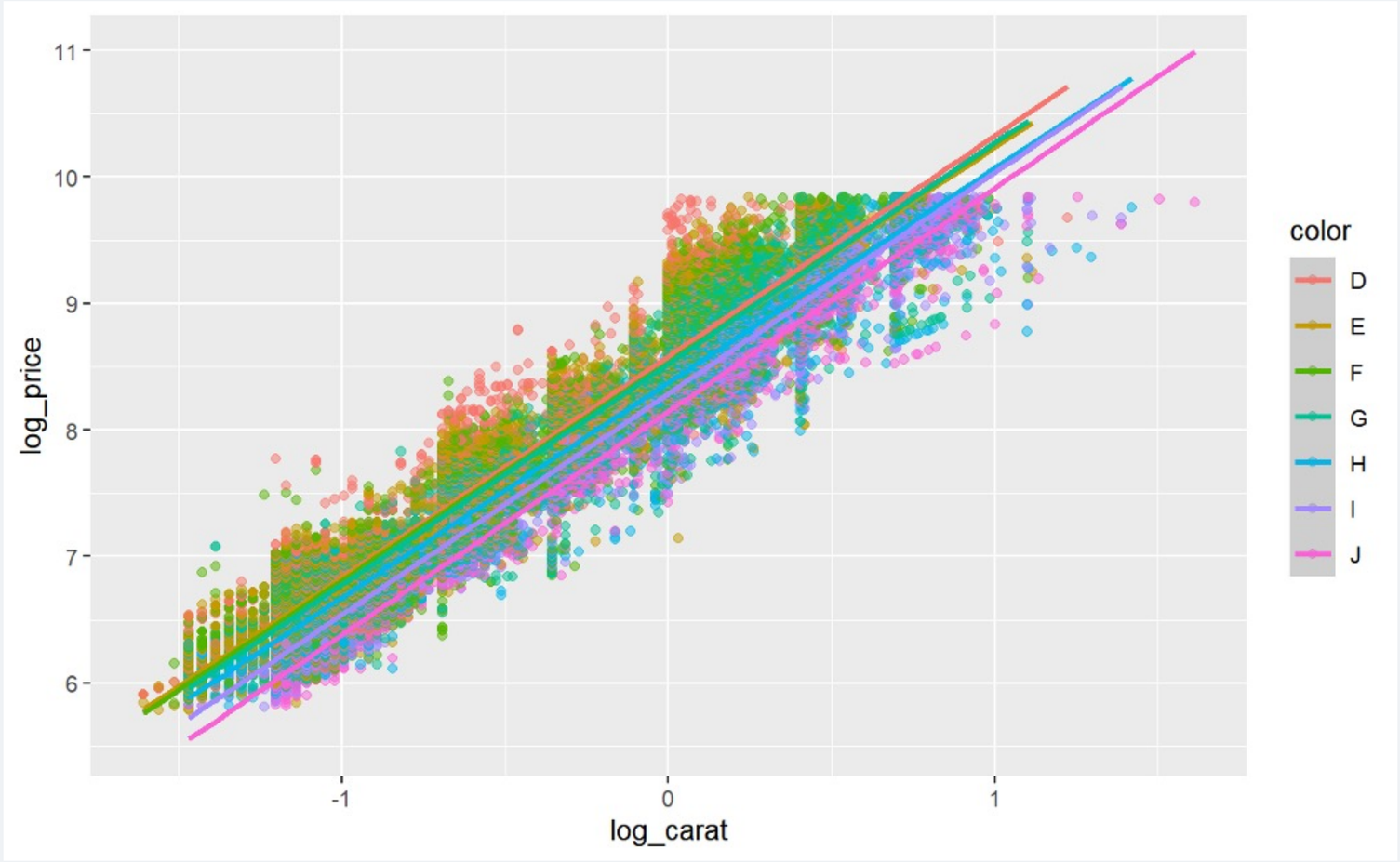
# Interaction Terms

## between Categorical and Numerical Variables



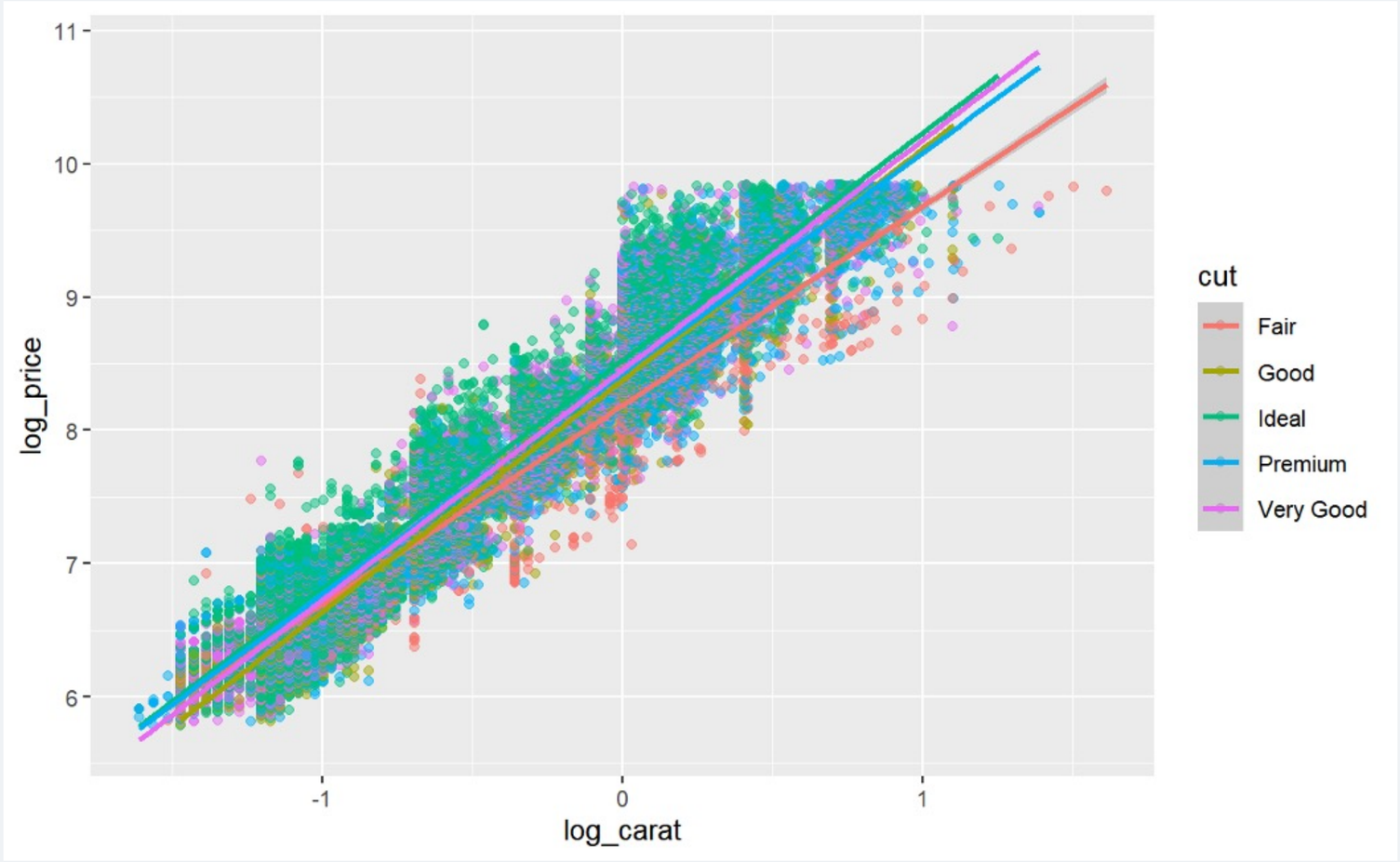
# Interaction Terms

## between Categorical and Numerical Variables



# Interaction Terms

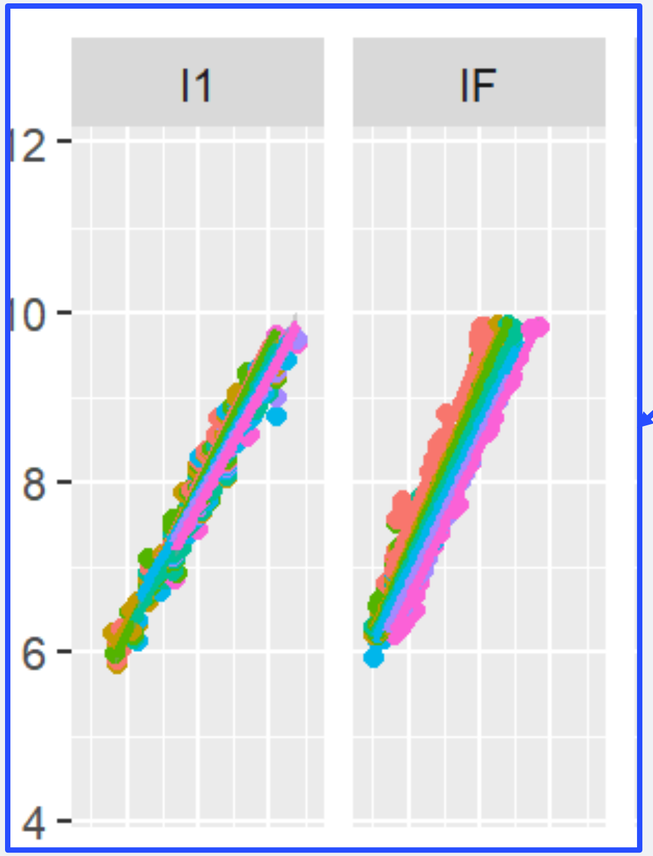
## between Categorical and Numerical Variables

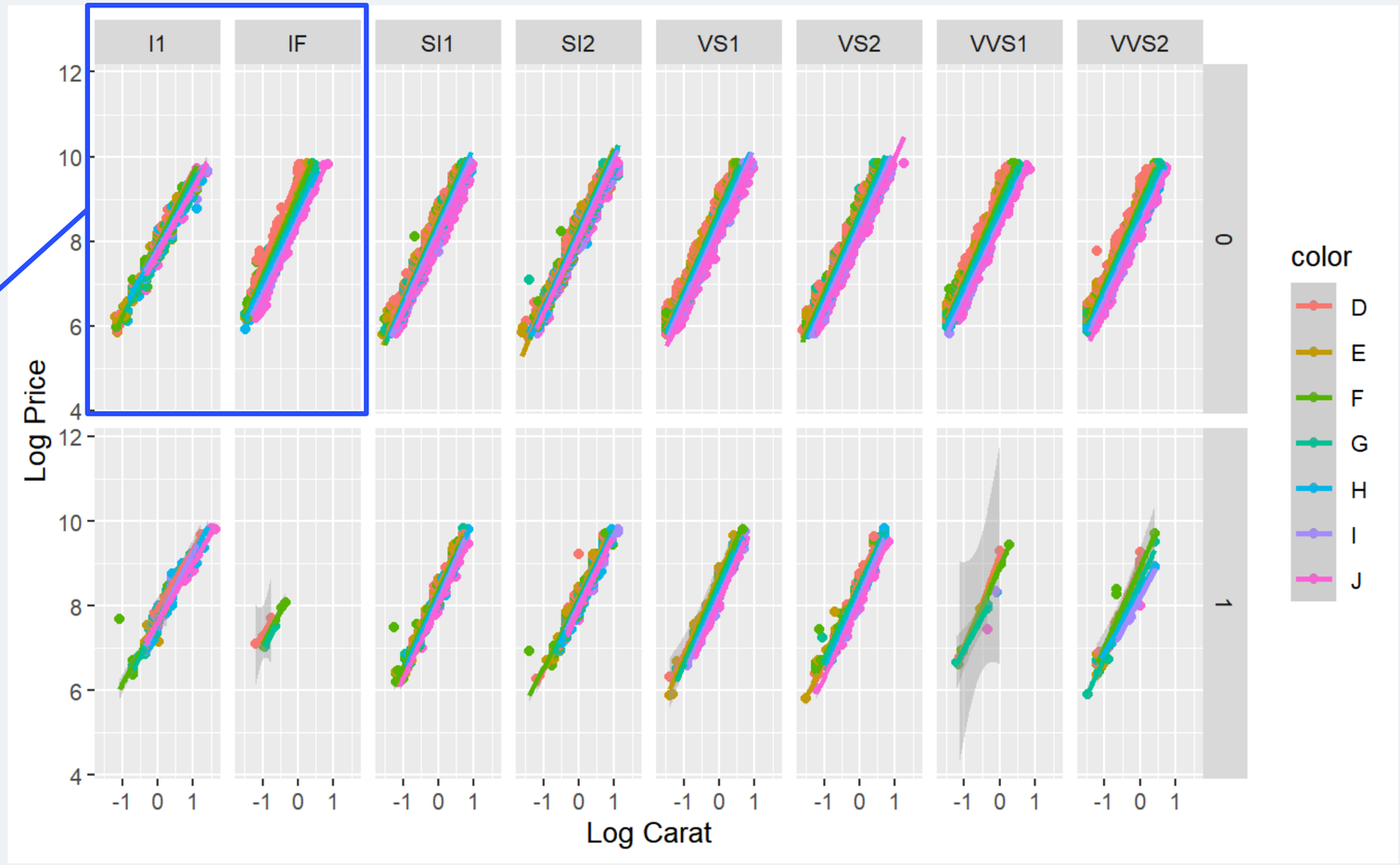


# Interaction Terms

## between Categorical Variables

color * clarity



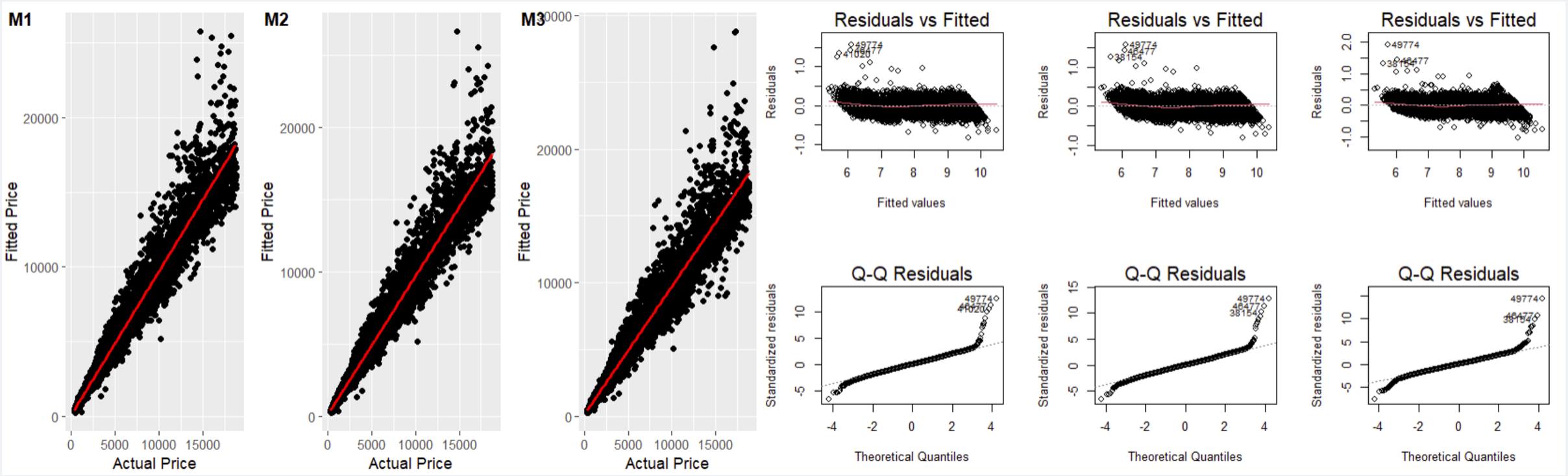



# Model Fitting and Evaluation

M1: log_price ~ color*clarity + log_carat*clarity + log_carat*cut

M2: log_price ~ color*clarity + log_carat*clarity + log_carat*cut + log_carat*color

M3: log_price ~ log_carat + cut + color + clarity + depth + table^(-2) + log_x + log_y + log_z + log_x_y



# Model Fitting and Evaluation

| Model | $R^2$ | $R^2_{adj}$ | Test MSE | Performance Improvement (compared with full model) |
|---|---|---|---|---|
| M1 | 0.9846 | 0.9845 | 487594.9 | 60.28% |
| M2 | 0.9848 | 0.9848 | 449861.2 | 63.37% |
| M3 | 0.9827 | 0.9827 | 625240.4 | 49.07% |

# Model Fitting and Evaluation

| Model | $R^2$ | $R^2_{adj}$ | Test MSE | Performance Improvement (compared with full model) |
|---|---|---|---|---|
| M1 | 0.9846 | 0.9845 | 487594.9 | 60.28% |
| M2 | 0.9848 | 0.9848 | 449861.2 | 63.37% |
| M3 | 0.9827 | 0.9827 | 625240.4 | 49.07% |

M1: log_price ~ color*clarity + log_carat*clarity + log_carat*cut

M2: log_price ~ color*clarity + log_carat*clarity + log_carat*cut + log_carat*color

# Model Fitting and Evaluation

| Model | $R^2$ | $R^2_{adj}$ | Test MSE | Performance Improvement (compared with full model) |
|---|---|---|---|---|
| M1 | 0.9846 | 0.9845 | 487594.9 | 60.28% |
| M2 | 0.9848 | 0.9848 | 449861.2 | 63.37% |
| M3 | 0.9827 | 0.9827 | 625240.4 | 49.07% |

M2: log_price ~ color*clarity + log_carat*clarity + log_carat*cut + log_carat*color

F-test ($p - value < 2.2e - 16$)
Can't be simplified

04

# Results

Interpretation for model coefficients

# Take-aways from the coefficients

**The "log_carat" coefficient has a positive estimate of 1.53.**

- For every one-unit increase in the logarithm of carat weight, the log price of the diamond is expected to increase by approximately 1.53 units.

**Better categories leads to higher price.**

- Higher clarity levels, cut grades and color grades generally correspond to higher log prices.

**The effect of "log_carat" on "log_price" varies across different levels of clarity and cut.**

# THANK YOU

## Q&A