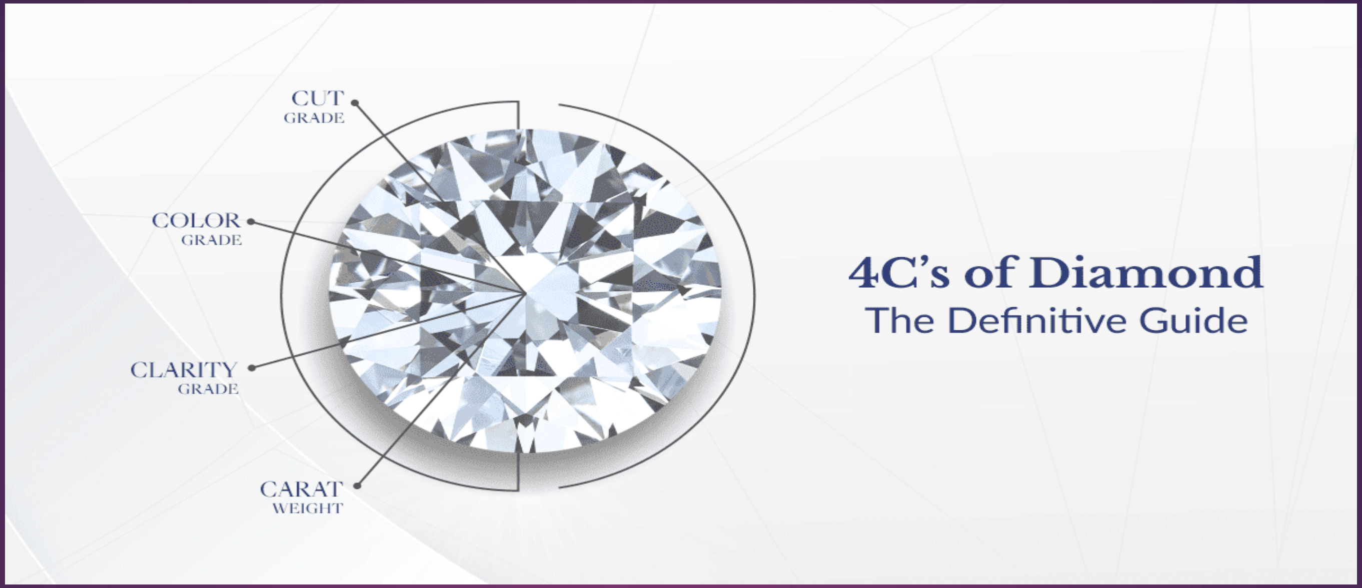


DIAMOND PRICE PREDICTION

By ~ Areeb Shaikh

Data Overview

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

Kaggle (Diamond Price Dataset with 50,000 records).

Features:-

1. Numerical: Carat, Dimensions (x, y, z).
2. Categorical: Cut (Fair, Good, Very Good, Premium, Ideal), Color (J to D), Clarity (I1 to IF).
3. **Target Variable:-** Price in USD.

Project Overview

- **Problem Statement**

In the diamond industry, accurate pricing is a critical challenge due to the variability of diamond features such as carat, cut, color, clarity, and dimensions. The pricing process often involves manual evaluation, which can be subjective and inconsistent. This creates the need for a robust and data-driven solution to estimate diamond prices reliably and transparently.

- **Objective**

The objective of this project is to predict the price of diamonds based on key features such as carat, cut, color, clarity, and dimensions. By leveraging machine learning techniques, the project aims to build an accurate predictive model that can estimate diamond prices, providing insights for the jewelry industry and potential buyers

Methodology

- **Data Preparation :-**

1) Perform exploratory data analysis (EDA) to identify patterns and trends.

- Univariate , Bivariate, Multivariate

2) Handle missing & Duplicate data (if any) and ensure consistent formatting.

3) Retain outliers as they reflect valid variability in diamond characteristics.

4) Feature Engineering

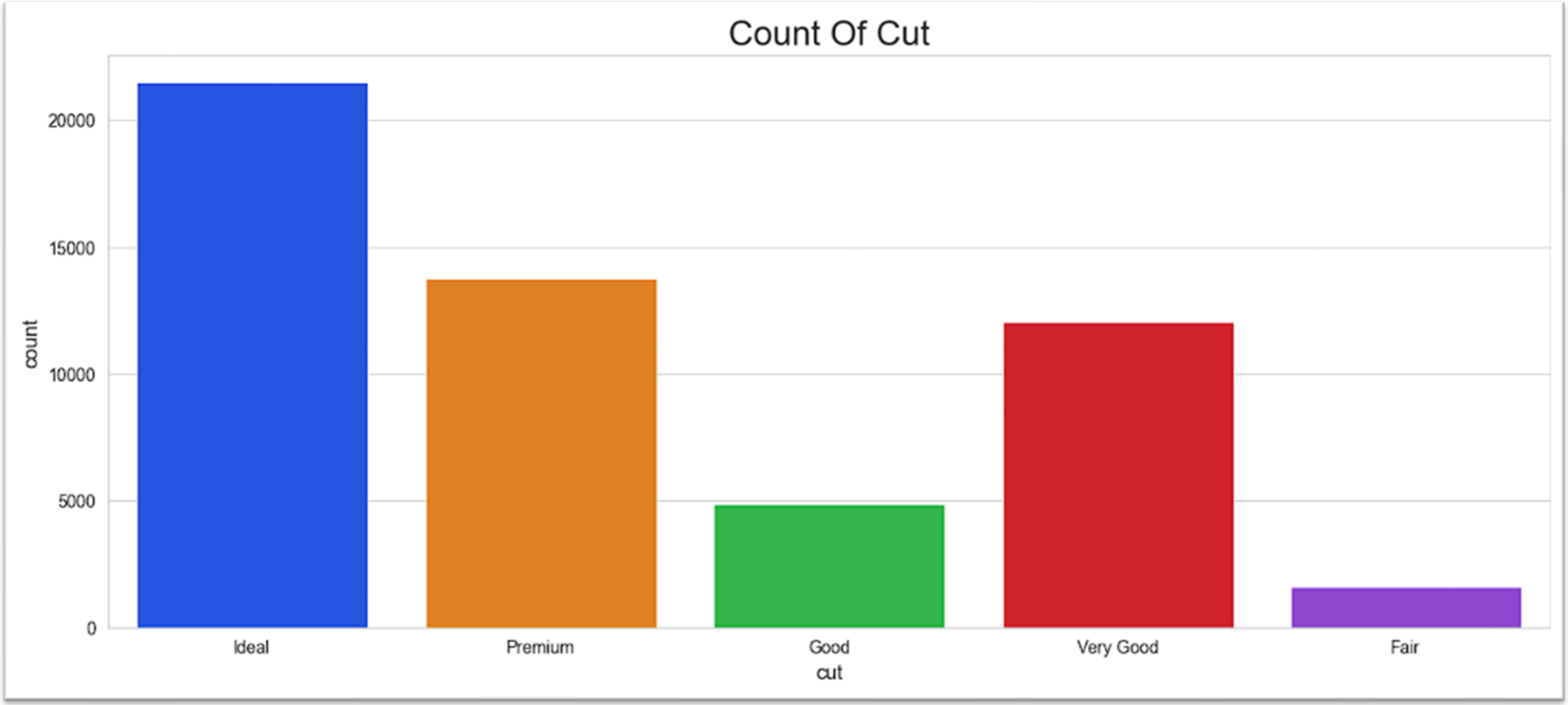
- Created New Attribute (Size) By Combining Existing Attribute (x, y, z)

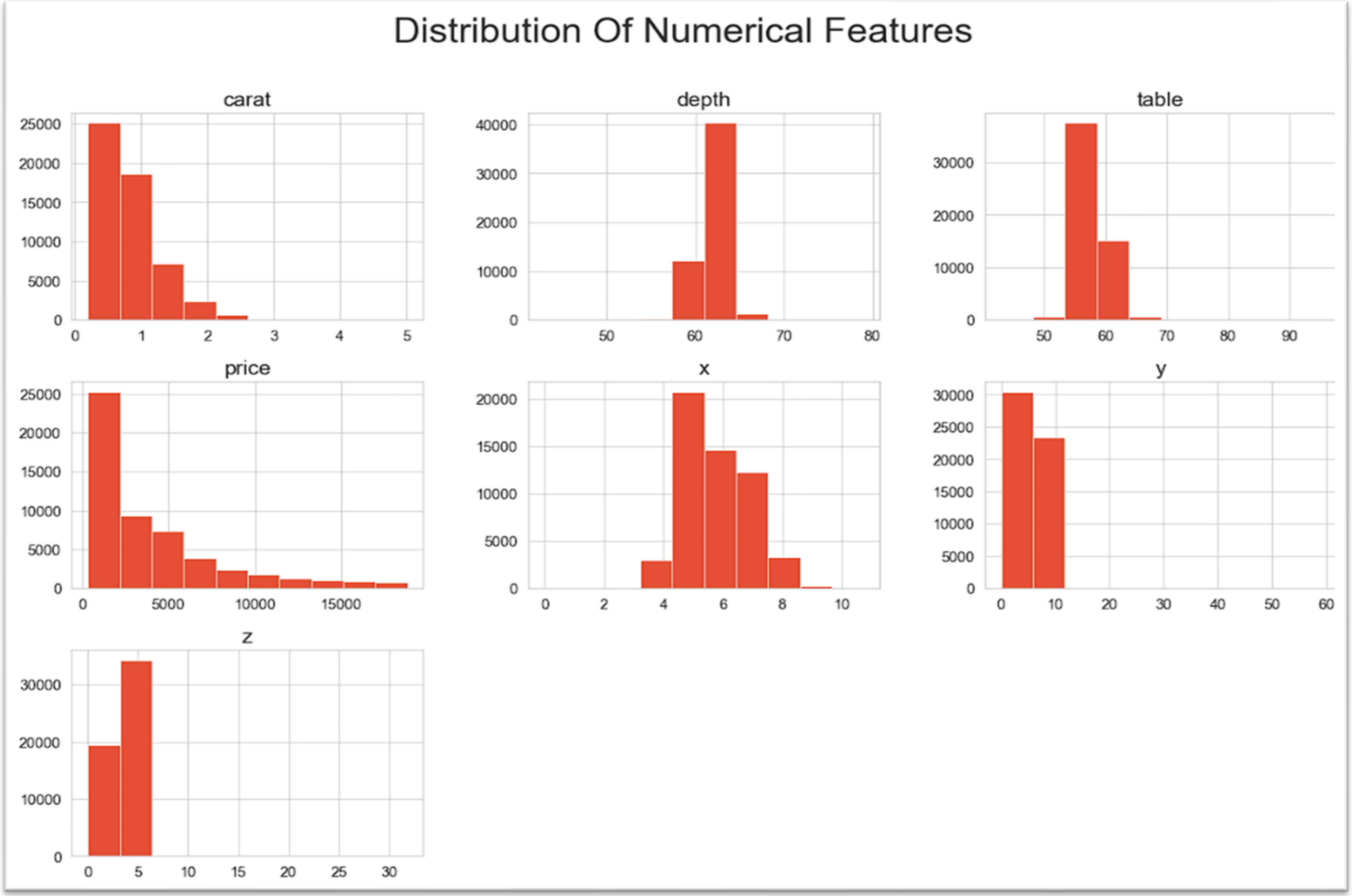
5) Label Encoding To Categorical Features

Insights & Visuals

- **Used visualizations**

1) Bar Plot (Categorical Distribution)
2) Histogram (Numerical Distribution)
3) Scatter Plot (Carat vs Price)
4) Scatter Plot (Size vs Price)
5) Box Plot
6) Scatter + Regression plot
7) Heatmap (Correlation Matrix)
8) Feature Importance Plot


Count Of Cut
count
20000
15000
10000
5000
0
Ideal
Premium
Good
Very Good
Fair
cut



Distribution Of Numerical Features

carat

depth

table

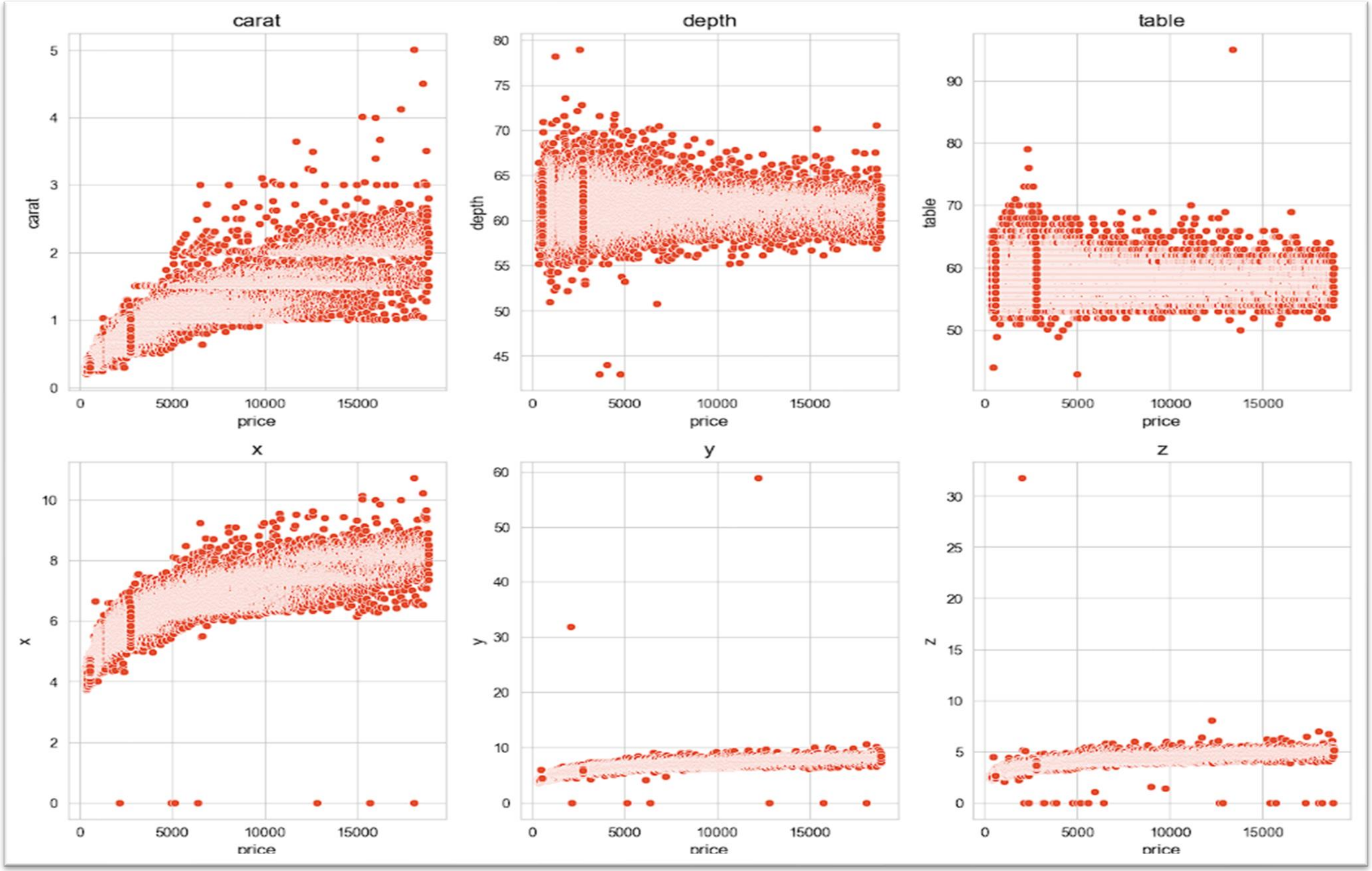
price

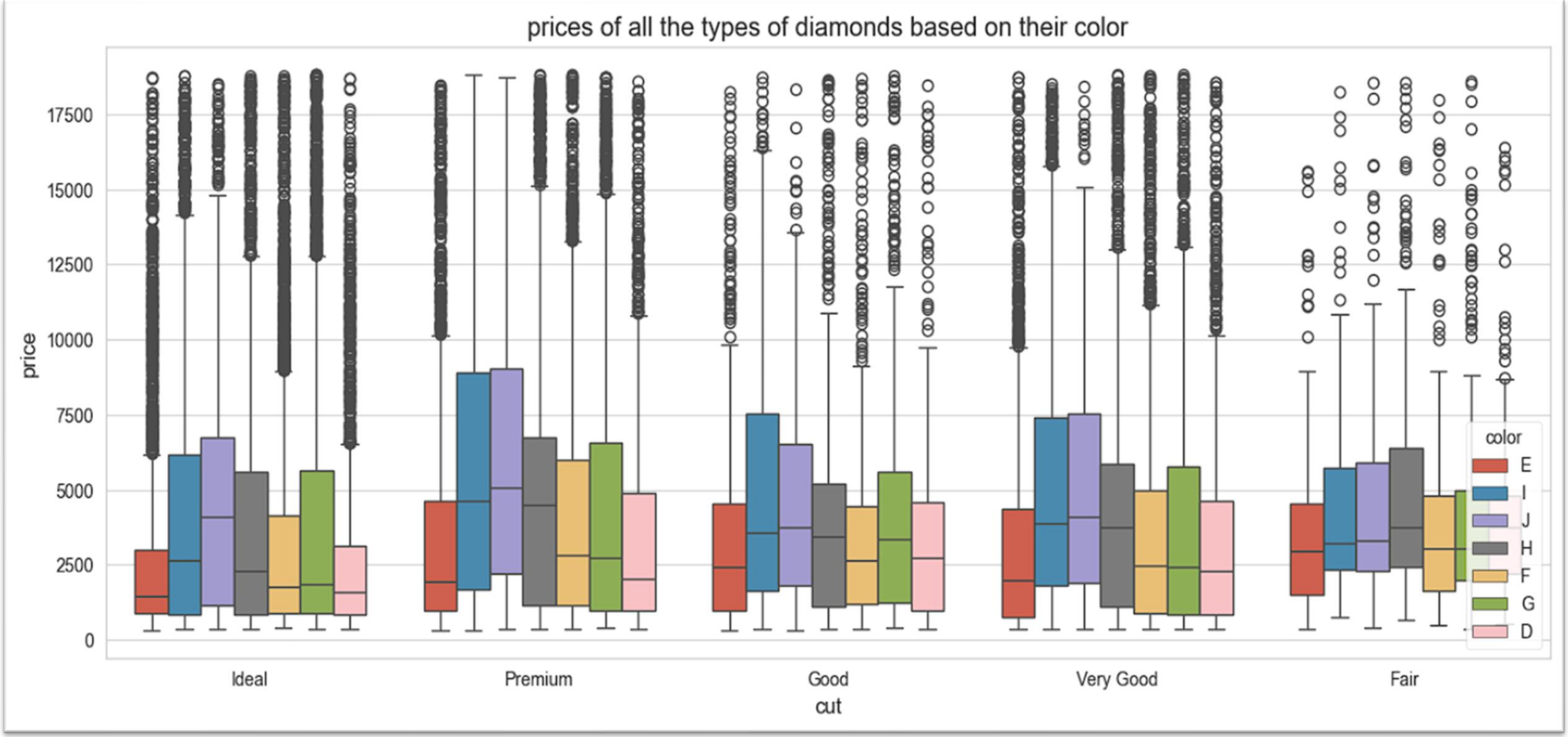
x

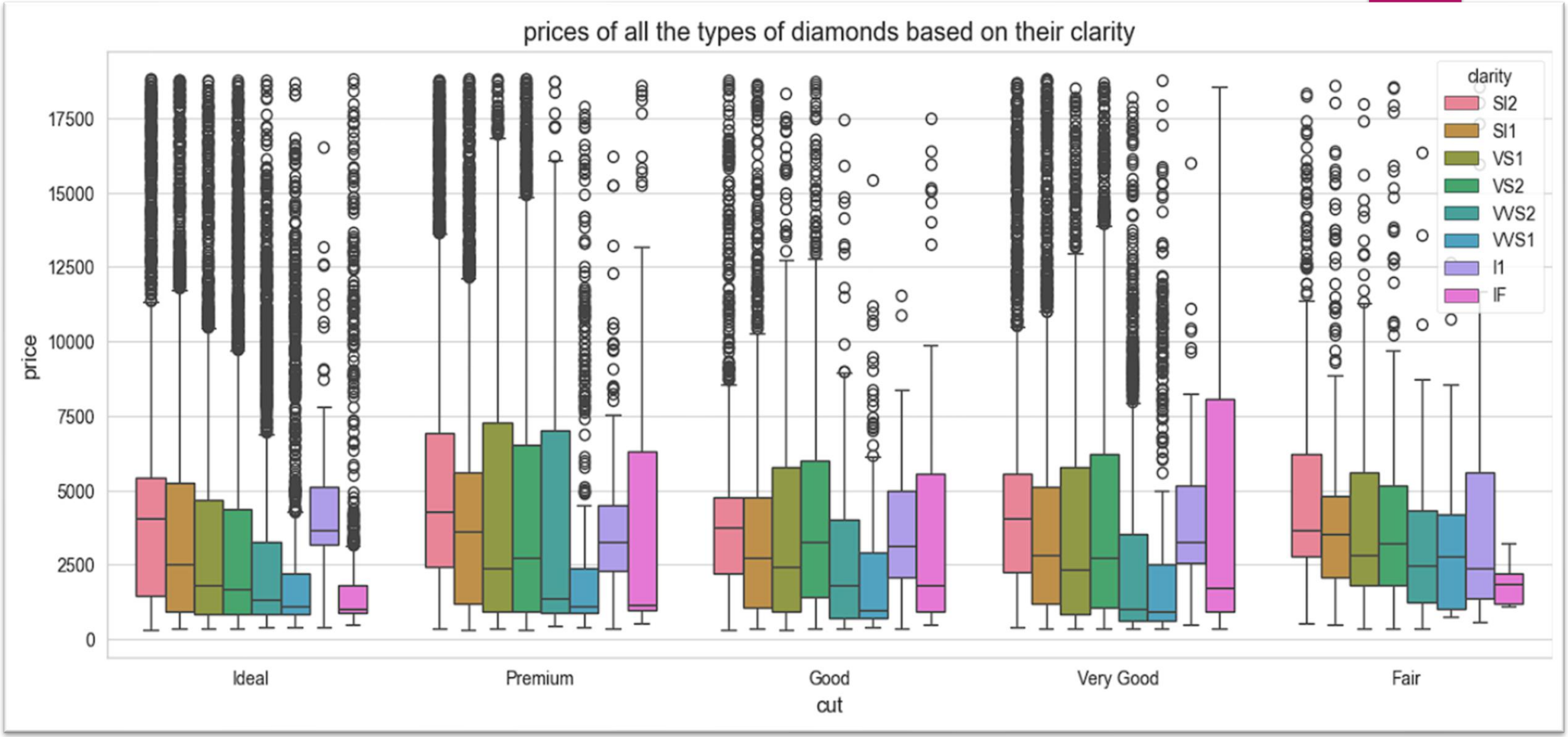
y

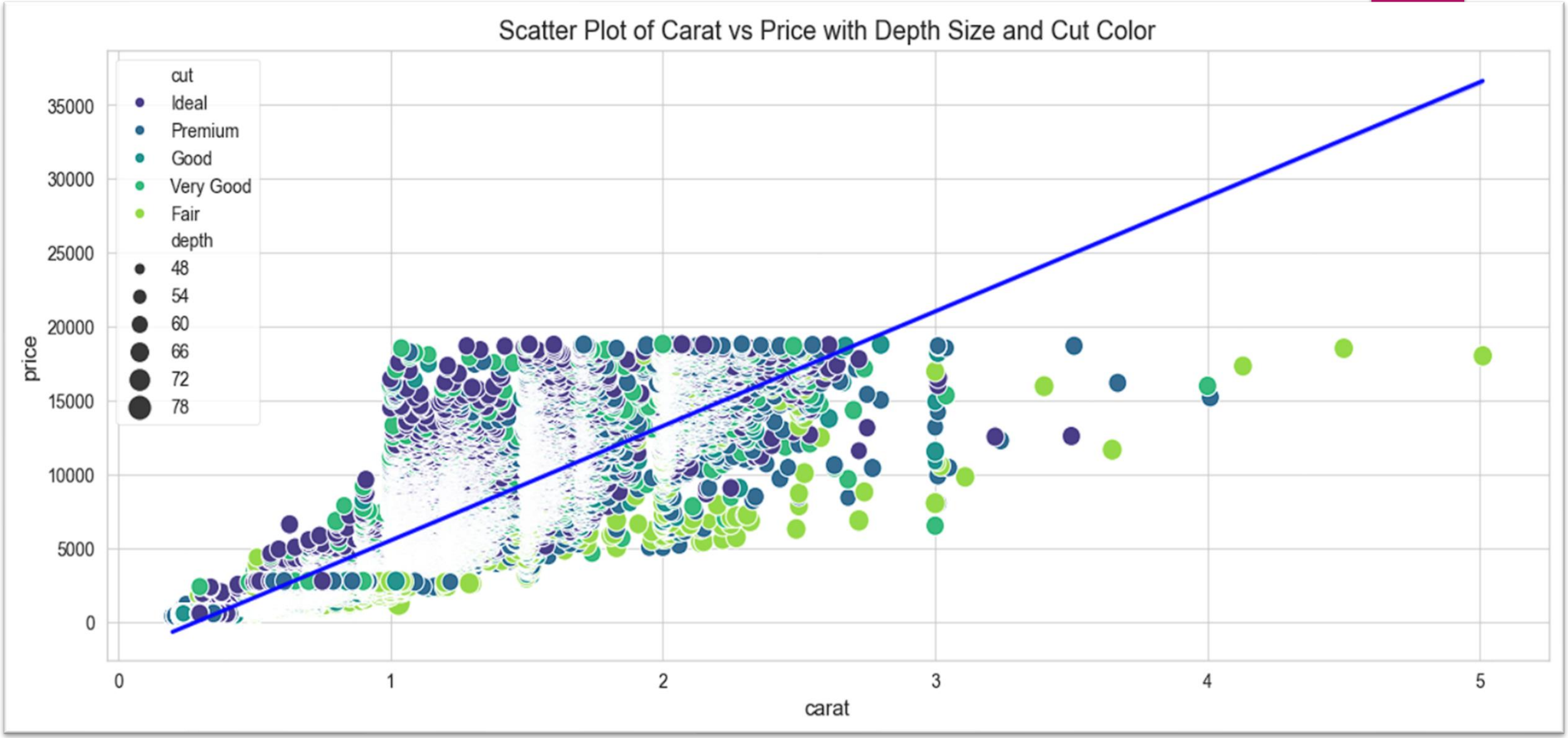
z

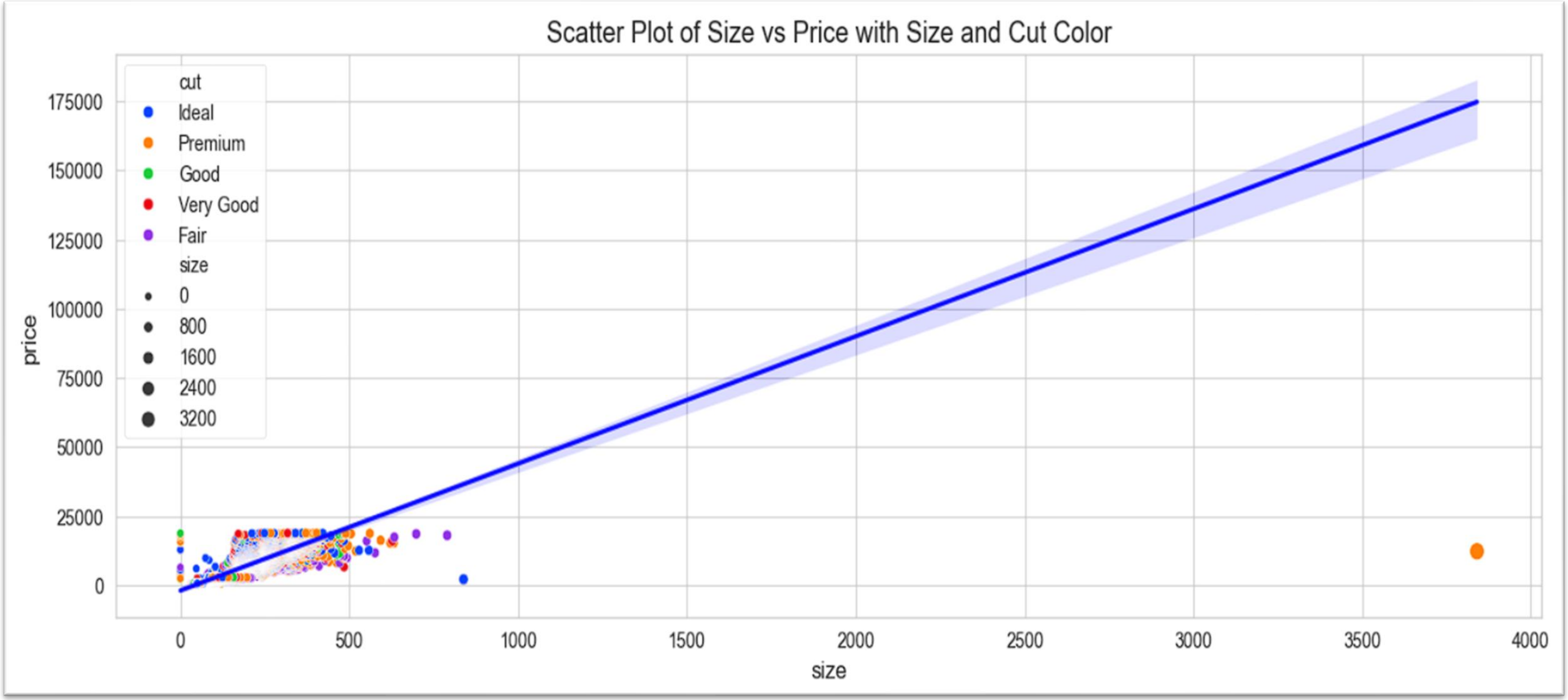
Relationship of all the feature with Target variable



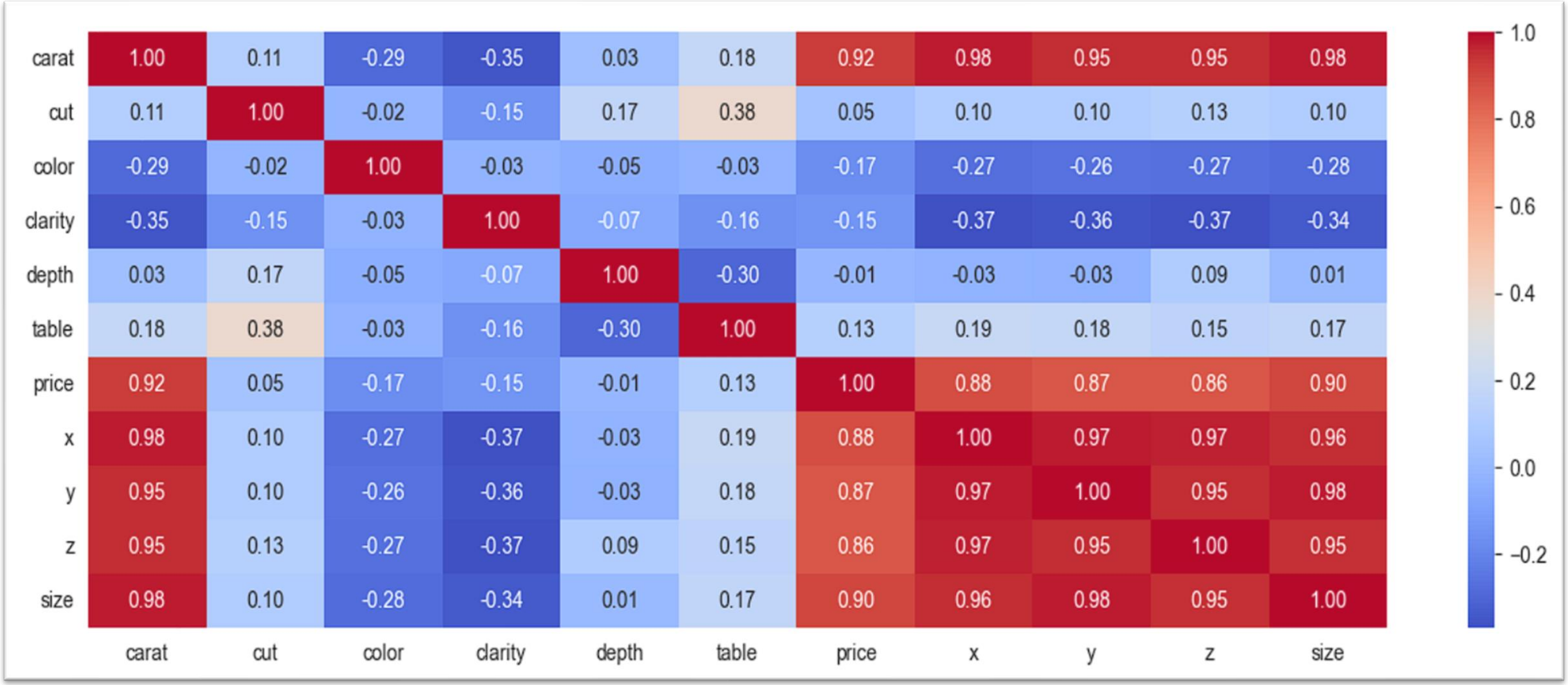

prices of all the types of diamonds based on their color
price
0
2500
5000
7500
10000
12500
15000
17500
cut
Ideal
Premium
Good
Very Good
Fair
color
E
I
J
H
F
G
D


prices of all the types of diamonds based on their clarity
price
17500
15000
12500
10000
7500
5000
2500
0
Ideal
Premium
Good
Very Good
Fair
cut
clarity
SI2
SI1
VS1
VS2
VVS2
VVS1
I1
IF

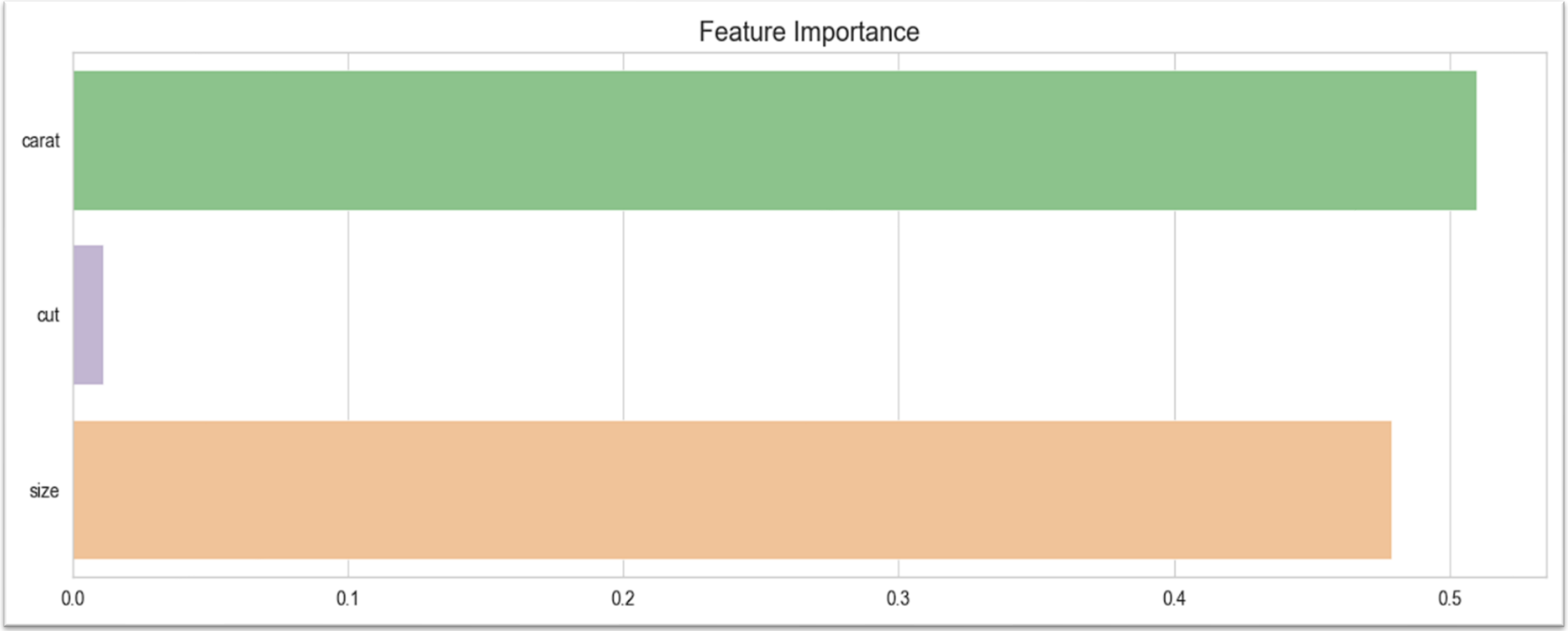

Scatter Plot of Carat vs Price with Depth Size and Cut Color
cut
Ideal
Premium
Good
Very Good
Fair
depth
48
54
60
66
72
78
price
35000
30000
25000
20000
15000
10000
5000
0
carat
0
1
2
3
4
5


Scatter Plot of Size vs Price with Size and Cut Color
cut
Ideal
Premium
Good
Very Good
Fair
size
0
800
1600
2400
3200
price
175000
150000
125000
100000
75000
50000
25000
0
size
0
500
1000
1500
2000
2500
3000
3500
4000

Heatmap Of Correlation Matrix



	carat	cut	color	clarity	depth	table	price	x	y	z	size
carat	1.00	0.11	-0.29	-0.35	0.03	0.18	0.92	0.98	0.95	0.95	0.98
cut	0.11	1.00	-0.02	-0.15	0.17	0.38	0.05	0.10	0.10	0.13	0.10
color	-0.29	-0.02	1.00	-0.03	-0.05	-0.03	-0.17	-0.27	-0.26	-0.27	-0.28
clarity	-0.35	-0.15	-0.03	1.00	-0.07	-0.16	-0.15	-0.37	-0.36	-0.37	-0.34
depth	0.03	0.17	-0.05	-0.07	1.00	-0.30	-0.01	-0.03	-0.03	0.09	0.01
table	0.18	0.38	-0.03	-0.16	-0.30	1.00	0.13	0.19	0.18	0.15	0.17
price	0.92	0.05	-0.17	-0.15	-0.01	0.13	1.00	0.88	0.87	0.86	0.90
x	0.98	0.10	-0.27	-0.37	-0.03	0.19	0.88	1.00	0.97	0.97	0.96
y	0.95	0.10	-0.26	-0.36	-0.03	0.18	0.87	0.97	1.00	0.95	0.98
z	0.95	0.13	-0.27	-0.37	0.09	0.15	0.86	0.97	0.95	1.00	0.95
size	0.98	0.10	-0.28	-0.34	0.01	0.17	0.90	0.96	0.98	0.95	1.00


Feature Importance
carat
cut
size
0.0
0.1
0.2
0.3
0.4
0.5

Model Development

- **Tree Based Models** :-

1. Decision Tree
2. Random Forest
3. Ada Boost
4. XGB

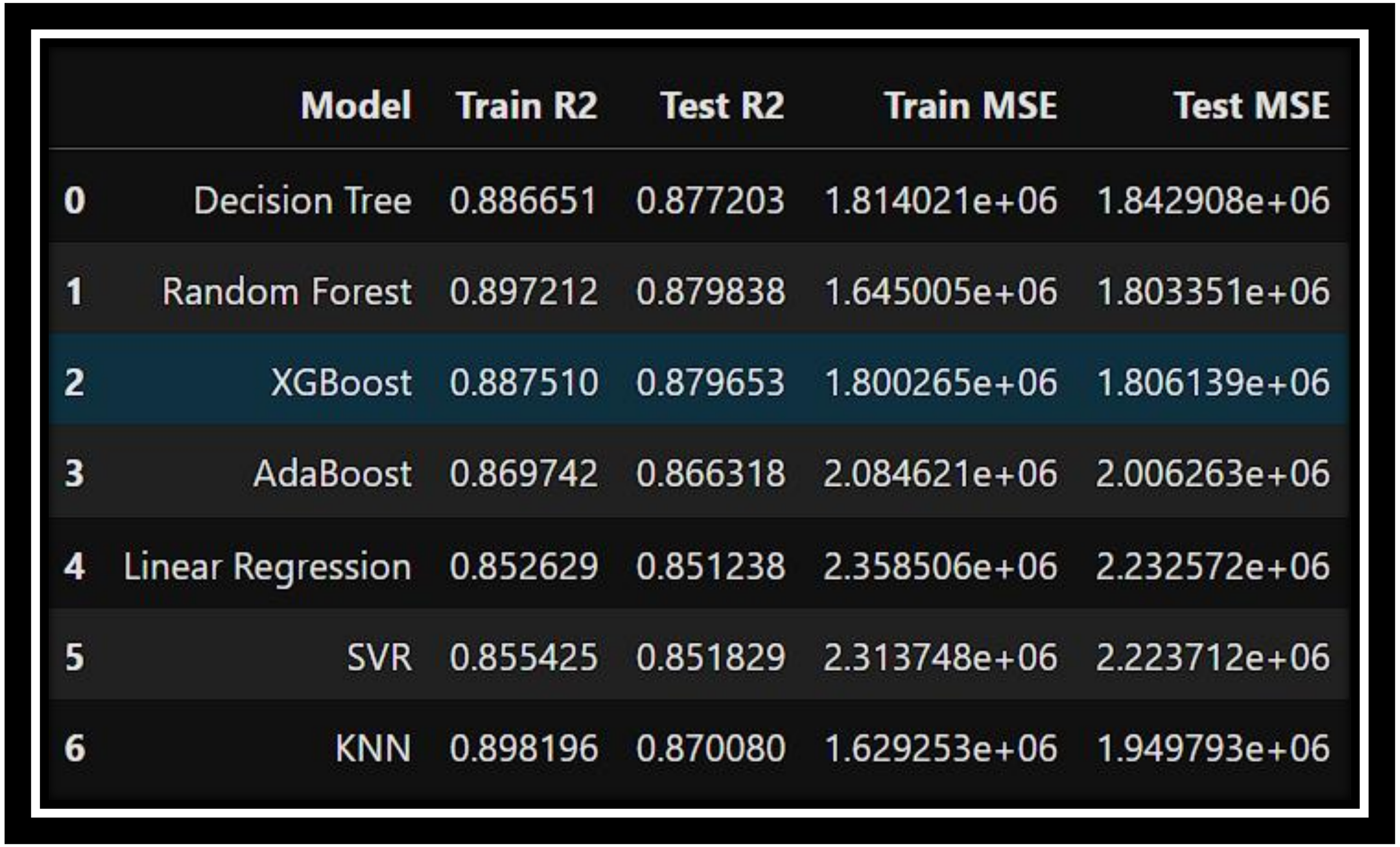
- **Scaling-Based Models**

1. Linear Regression
2. KNN
3. SVM

Challenges

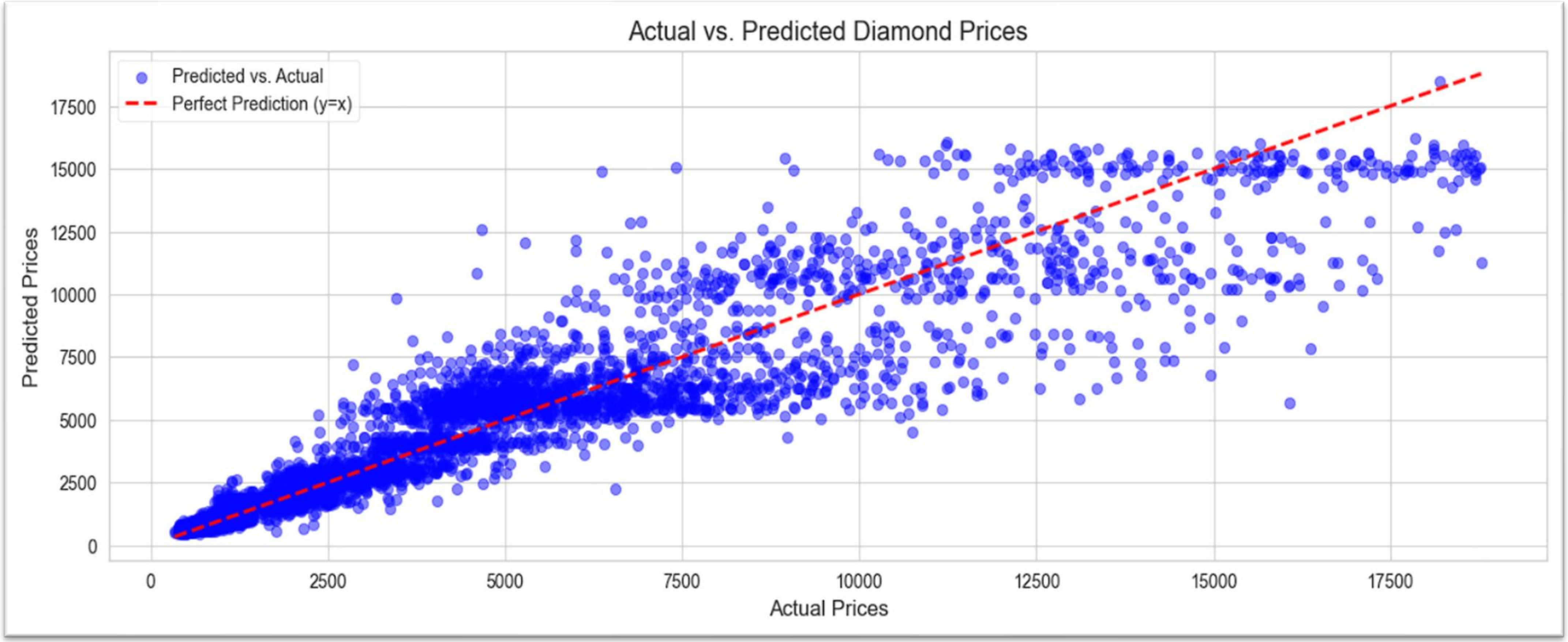
1. **Outliers:-** Managing outliers to ensure they do not overly bias the model while retaining their valid variability.
2. **Categorical Features:-** Encoding qualitative attributes like cut, color, and clarity effectively for both tree-based and scaling-based models.
3. **Model Tuning:-** Balancing computation time during hyperparameter tuning for a large dataset.

Model Performance Comparison



	Model	Train R2	Test R2	Train MSE	Test MSE
0	Decision Tree	0.886651	0.877203	1.814021e+06	1.842908e+06
1	Random Forest	0.897212	0.879838	1.645005e+06	1.803351e+06
2	XGBoost	0.887510	0.879653	1.800265e+06	1.806139e+06
3	AdaBoost	0.869742	0.866318	2.084621e+06	2.006263e+06
4	Linear Regression	0.852629	0.851238	2.358506e+06	2.232572e+06
5	SVR	0.855425	0.851829	2.313748e+06	2.223712e+06
6	KNN	0.898196	0.870080	1.629253e+06	1.949793e+06

Actual VS Predicted Diamond Price



Conclusion & Practical Application

Conclusion

- **Impact of Models**:- Accurate diamond price predictions help jewelers, buyers, and insurers determine fair prices.
- **Best Model**: XG boost, due to its high R2 and low MSE on both training and test sets.
- **Good Alternatives**: Random Forest , Decision Tree and KNN also show strong performance.
- **Baseline Models**: Linear Regression and SVR provide good baselines but are outperformed by ensemble methods.

Practical Application

1. **E-Commerce Platforms :-** Automating pricing recommendations for diamonds listed on online marketplaces.
2. **Inventory Management :-** Helping jewelers value their inventory based on consistent pricing.
3. **Fraud Detection :-** Identifying discrepancies in diamond pricing for authentication and fraud prevention.
4. **Customer Decision Support :-** Assisting customers in understanding and justifying diamond prices during purchases.

Deployment Details

- **Description:** The diamond price prediction model has been deployed as a web application using the Streamlit framework, enabling users to interact with the model seamlessly.
- **Platform**: Streamlit cloud platform.
- **Deployment Process :**

1. Model trained and saved using Pickle.
2. Interactive web app built with Streamlit for a user-friendly interface.
3. Hosted on Streamlit's free cloud service.

- **Accessibility:**

1. Fully functional and accessible from any device via the link.
2. **Link :** https://diamond-price-predictions.streamlit.app/