

Diamonds

Predicting diamond prices using
carat weight



Business problem:

A new jewelry company in the United States wants to develop a reliable approach to pricing its diamonds. The company wants to understand how strongly carat weight relates to price and whether carat weight alone is enough to set appropriate selling prices.

Hypothesis 1:

Diamonds with higher carat weights tend to cost more.

Overview



Research question:

Can a diamond's carat weight predict its price?

Hypothesis 2:

Adding cut, color, and clarity improves price predictions compared with using carat weight alone.



Data and Approach



53,940

diamond records from seaborn

Carat, cut, color, clarity, depth, table, price, x, y, z.

No missing values

Weight: 0.20–5.01 carats

Price: \$326–\$18,823

Predictor: carat weight

Outcome: price in USD

Explore the data

Measure Pearson correlation

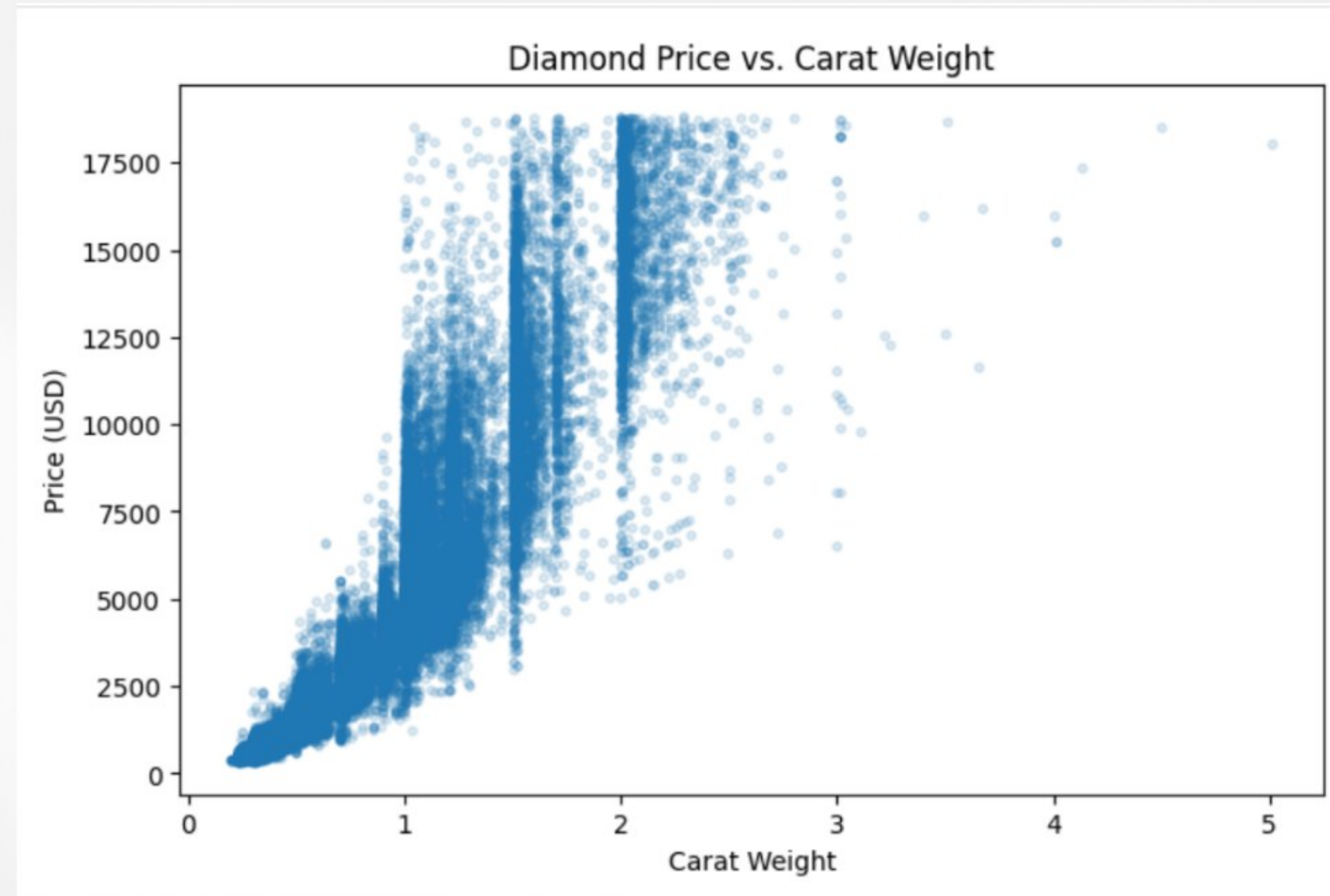
Fit a simple linear regression

Check the residuals

Heavy
diamonds
generally cost
more

R = 0.9216

Meaning: Positive linear association



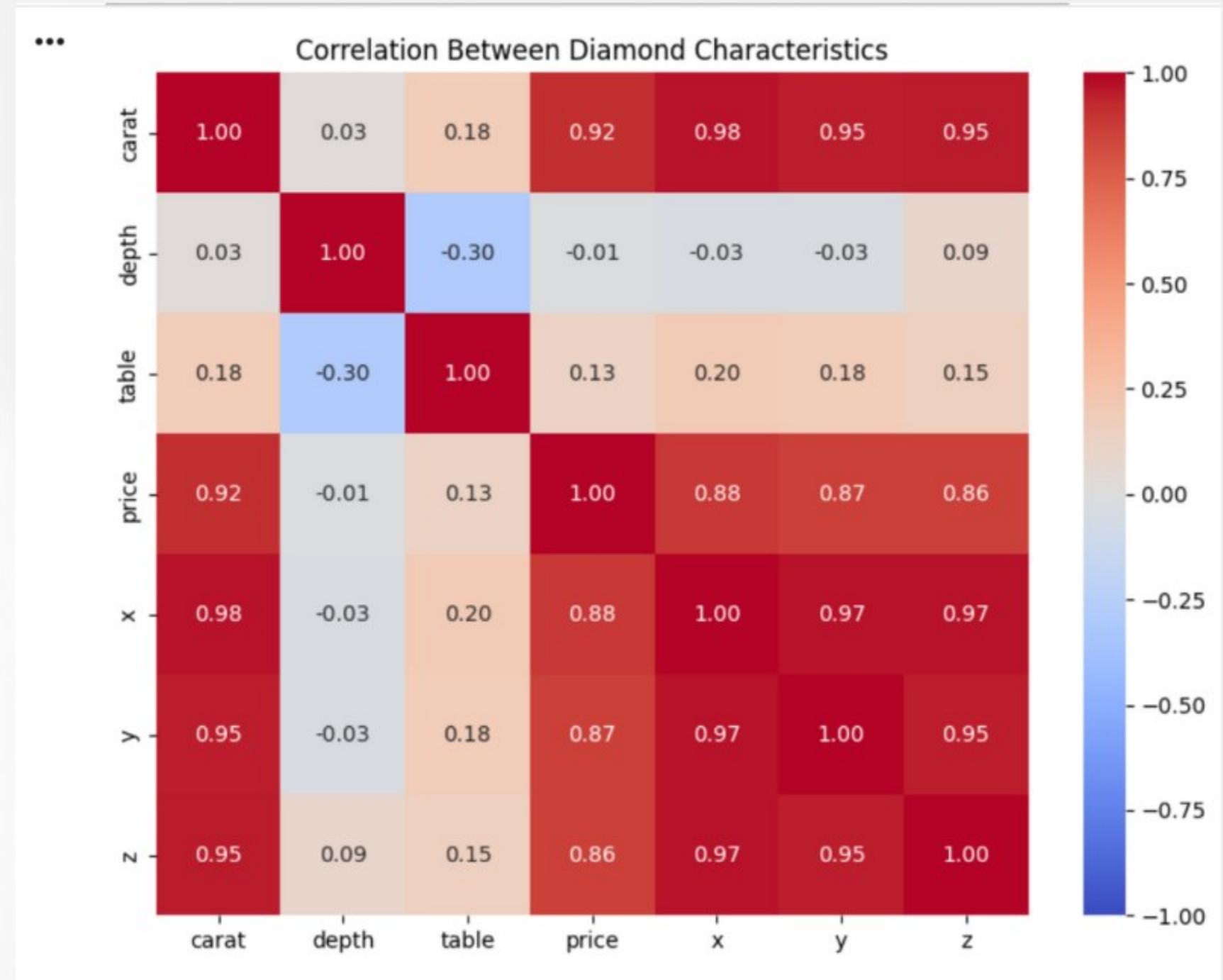
Prices vary even at the same weight



Heatmap

0.92

The heatmap shows a strong relation between price and carat (0.92)



Regression Results

✦ Predicted price = $-2,256.36 + 7,756.43 \times \text{carat}$

✦ **R-squared = 0.8492**

84.9% of observed price variation explained by the model

✦ **U\$ 5,500.07**

Is the predicted price for a 1.0 carat diamond

ANOVA test $p = 0.001$

```
# Fit a simple linear regression model
model = st.linregress(df["carat"], df["price"])

# Display the regression coefficients and R-squared
print("Slope:", round(model.slope, 4))
print("Intercept:", round(model.intercept, 4))
print("R-squared:", round(model.rvalue ** 2, 4))
```

Slope: 7756.4256
Intercept: -2256.3606
R-squared: 0.8493

```
# Predicting the price of a 1-carat diamond (Chapter 6, section 1.1)

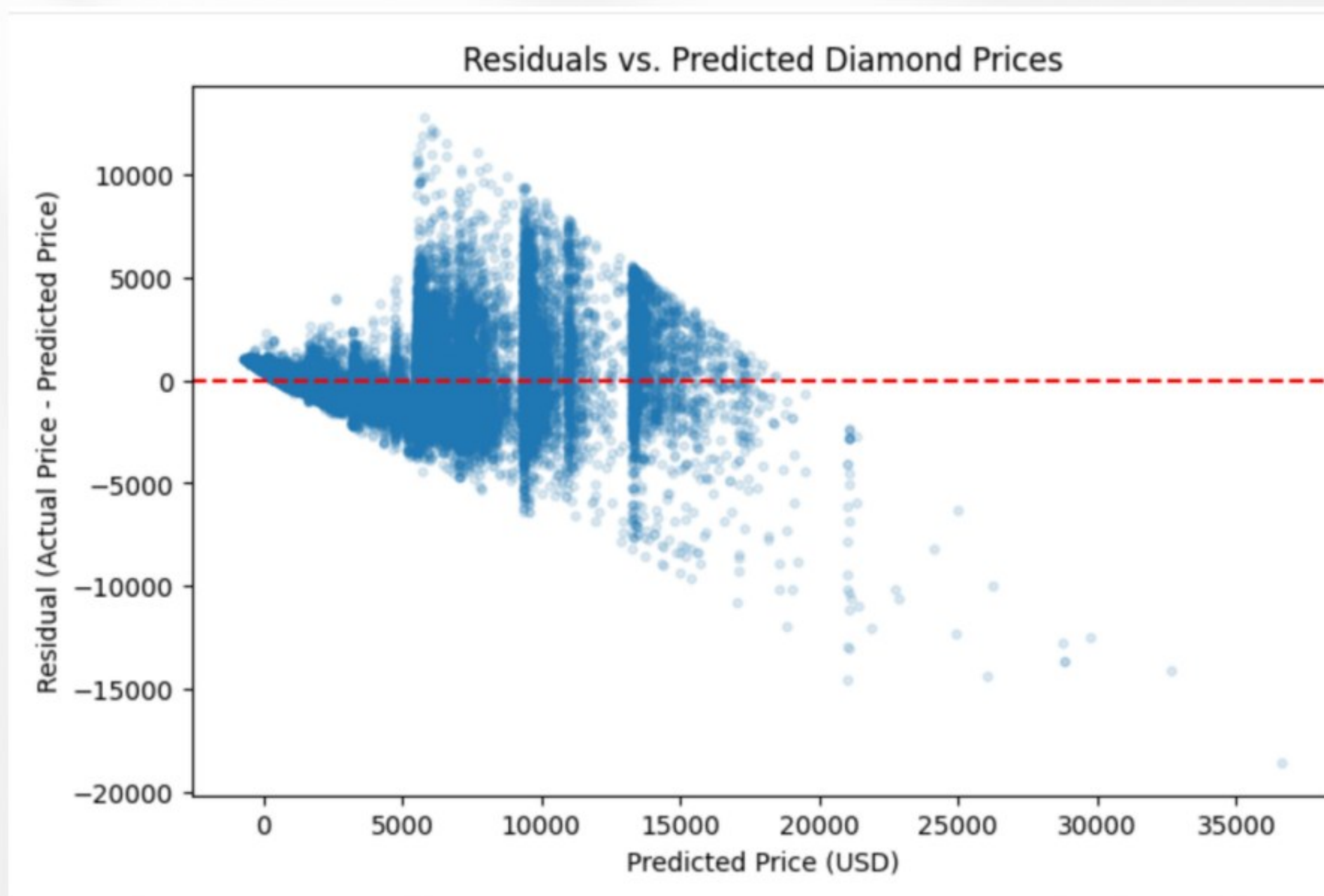
import numpy as np

prediction = np.dot([1, 1], [model.slope, model.intercept])

print("Predicted price for a 1-carat diamond:", round(prediction, 2))
```

Predicted price for a 1-carat diamond: 5500.07





The Model Has Limits

A negative residual means the model predicted too much. ideally, points should be around 0.

At the highest predicted prices, the residuals are strongly negative, so the model overestimates those diamonds

A straight line does not fully describe the relationship



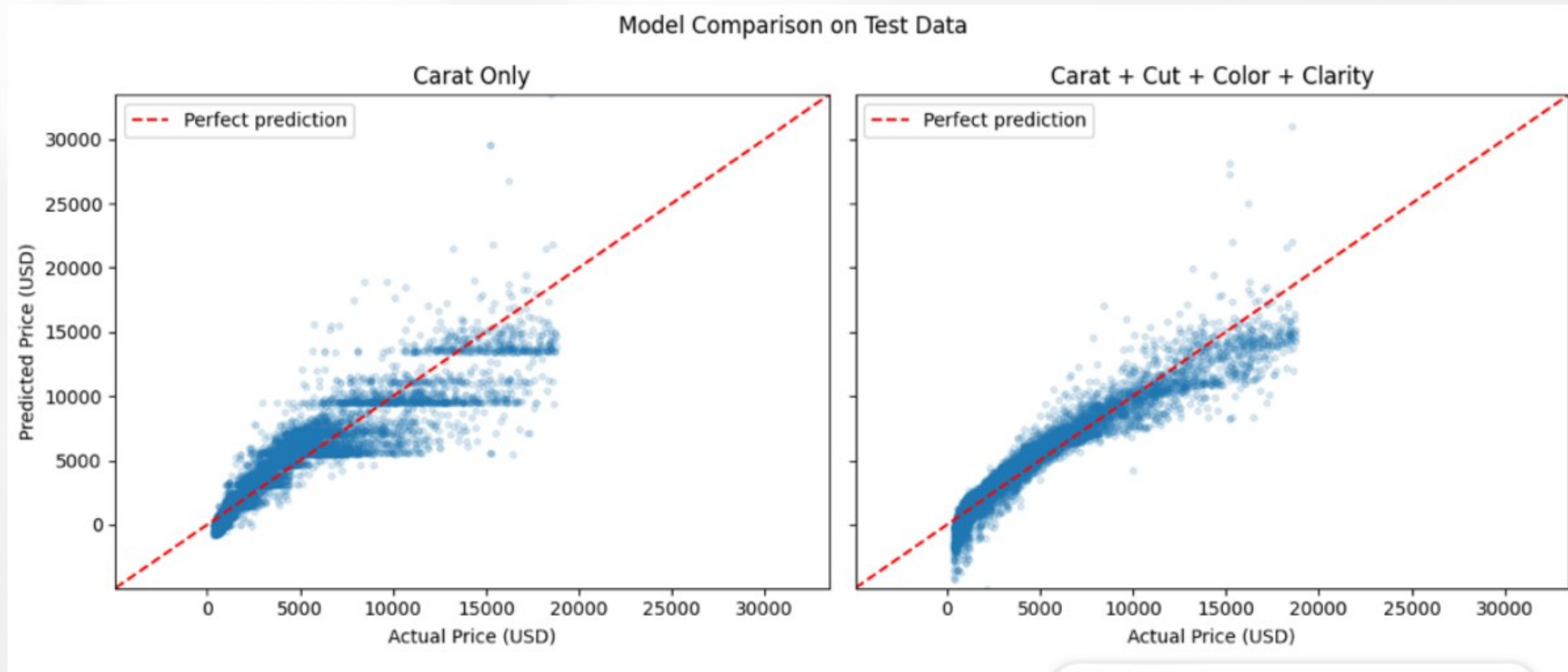
Hypothesis 2

When we added cut, color, and clarity, the average prediction error dropped from USD1,154.76 to USD872.09

The jewelry company should consider these characteristics along with carat weight when pricing diamonds.

	Model	MAE (USD)	RMSE (USD)
0	Carat only	1154.76	1702.35
1	Carat + cut + color + clarity	872.09	1242.03

Model Comparison on Test Data



The model using carat, cut, color, and clarity generally makes predictions closer to the actual prices.

Business Recommendations



Recommendation 1

Use carat weight as a starting point for pricing.

Recommendation 2

Include cut, color, and clarity in the next pricing model.

Recommendation 3

Compare estimates with current U.S. market prices.

