

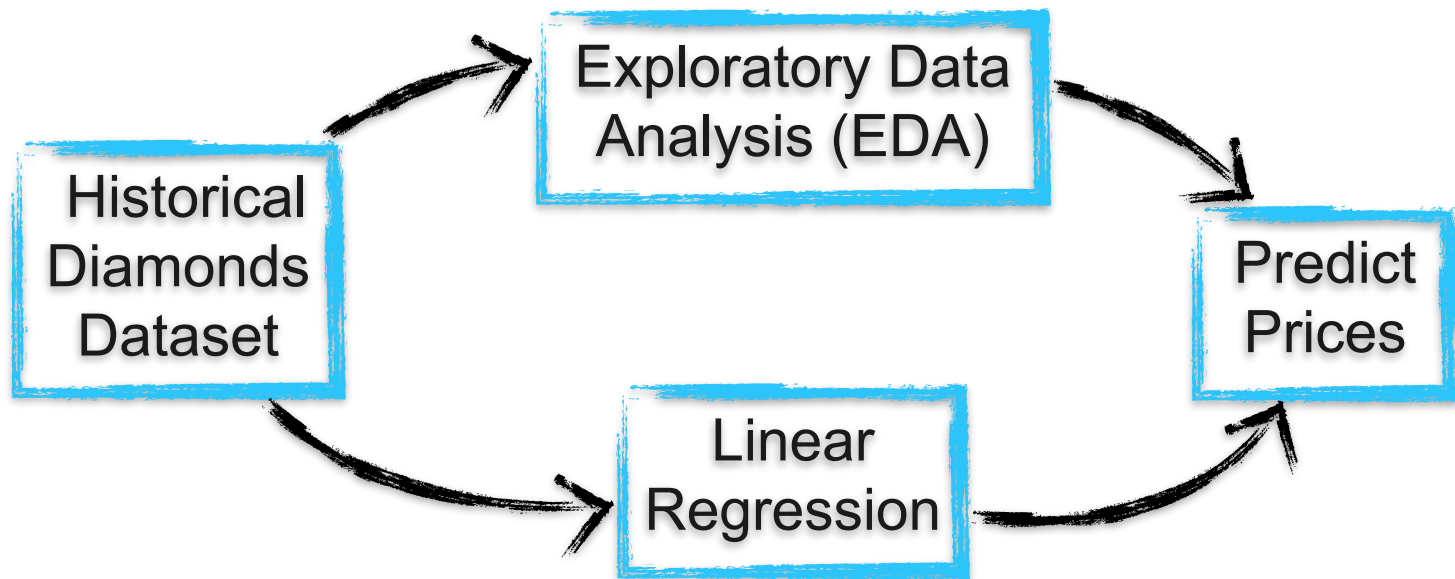


Project - Linear Regression

The world of predictions

Main Objectives

The Linear Regression challenge



Deliverables

- Rick has 5000 diamonds and, in this episode, he's called you to value them.
- Your job is to upload a **.csv file** containing the 5000 diamonds and a new column: **price_predicted** containing the predictions of your linear model.



Deliverables - Example

- Your csv file should be exactly the one you'll be given + the column **price_predicted** containing your predictions of the price of each diamond. Upload it to your GitHub.

	carat	cut	color	clarity	depth	table	x	y	z	price_predicted
0	0.23	Ideal	E	SI2	61.5	55.0	3.95	3.98	2.43	449
1	0.21	Premium	E	SI1	59.8	61.0	3.89	3.84	2.31	393
2	0.23	Good	E	VS1	56.9	65.0	4.05	4.07	2.31	404
3	0.29	Premium	I	VS2	62.4	58.0	4.20	4.23	2.63	418
4	0.31	Good	J	SI2	63.3	58.0	4.34	4.35	2.75	421

Metrics

The metric of success is, of course,

Metrics

The metric of success is, of course, **money**.



Your goal is exclusively to estimate the price of Rick's 5000 diamonds achieving the smallest amount of error, so they can sell it properly.

Metrics

The metric of success is, of course, money.

Specifically, we will measure the **root mean squared error** (RMSE) of your predictions, that is:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_i^N (Y_{\text{predicted}} - Y_{\text{observed}})^2}$$

Remember: ***from sklearn.metrics import mean_squared_error***

Metrics

The metric of success is, of course, money.

Rick's goal is to obtain an **average error not greater than 900 dollars**. Your goal is to create a prediction that will obtain a root mean squared error less than 1000 dollars. (If you achieve your goal, double your goal 🤪)

Specific Objectives

1. The first deliverable:

- The idea of the first deliverable is to create a fast and simple baseline model.

2. The final deliverable:

- The final deliverable should be your best model.
- All your EDA and Data Cleaning should be focused on enhancing your metric.

Hints

1. There will be a guided path for a structured way to clean and explore your data and create a simple linear model.
2. **Following it is not mandatory**, though (blindly following it is not even advised since it will not get you to the goal required). You can have your own ideas and your own insights during data exploration to achieve a good result.

Hints

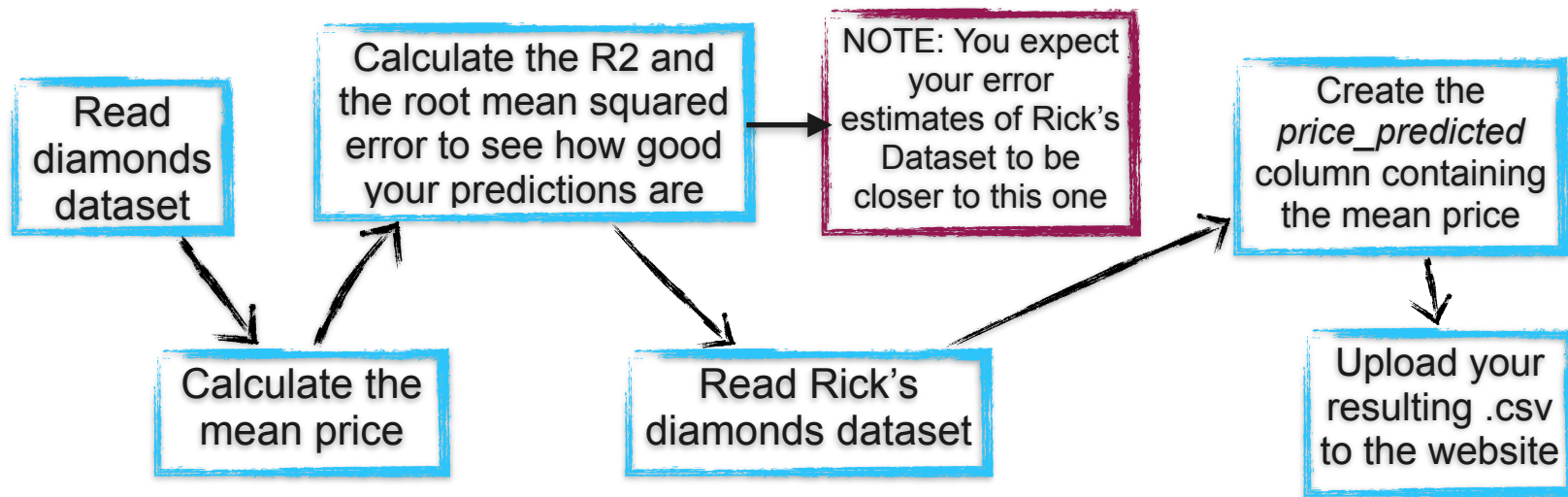
1. The process of building a prediction is iterative - there is no structured path to follow like **EDA, then Data Cleaning, then Regression Model**. You do all of them together.
2. Use your baseline for comparison. As soon as you make a change (for example, remove outliers), compare your new results against your baseline to see whether it had a good effect or not.
3. Understand your dataset. Look for things that impact the price of diamonds to increase your understanding of the problem and, hence, increase your regression model.

Presentation

- You'll make a quick presentation showing your results and which insights you've discovered on the analysis of the dataset that has helped you enhance your results.



Hint: Idea on how to create your first baseline - the mean



- You'll predict all observations to be the same value, the mean historical price of diamonds. This will result in a poor model. The idea here is to look for better estimates to predict price. If your new estimative is worst than this baseline, you should throw it away. As soon as you find a better estimative, this is your new baseline.