Assignment 4: Performance Metrics, and Optimisation

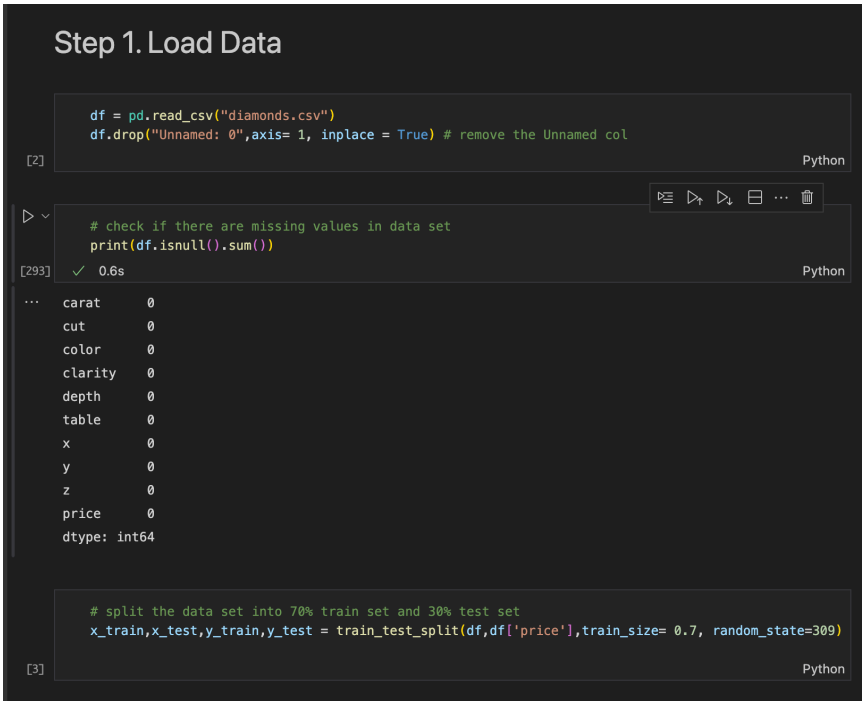
Student ID: 300637212 Student Name: Xieji Li

Part 1: Performance Metrics in Regression [30 marks]

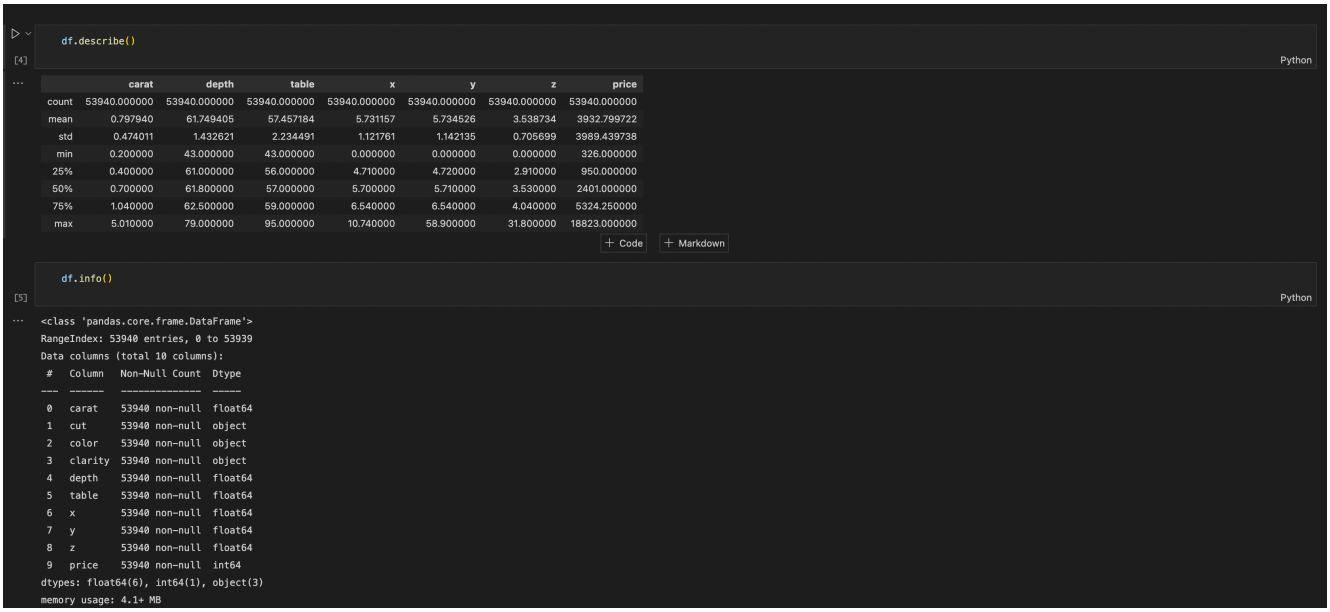
Requirements

Based on exploratory data analysis, discuss what preprocessing that you need to do before regression, and provide evidence and justifications.

- Step1. Load Data && split the dataset



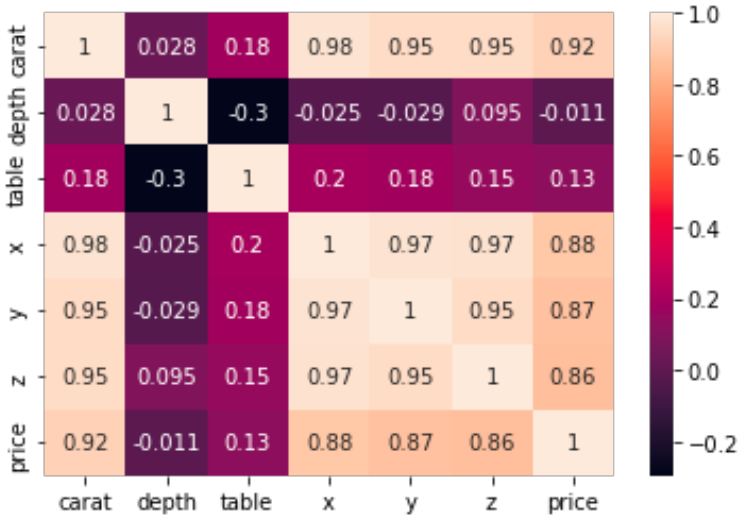
- Step 2. Initial Data Analysis



Conclusion: In this stage we can know there are 10 features in this dataset. We need to predict the value of price based on other 9 features. Also, there is no missing value in this dataset.

- correlation analysis

  Heat map



price	
carat	0.921591
x	0.884435
y	0.865421
z	0.861249
price	1.000000

So the size of diamond have a high correlation with price

- Step 3. Preprocess Data && Step 4. Exploratory Data Analysis
  - First, use histogram to display features, if the feature is numeric type then plot the hist according to the value of feature. If the feature is category type then plot the hist according to the frequency of the value.

carat	cut	Color	Clarity	Depth

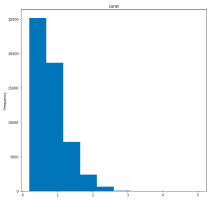
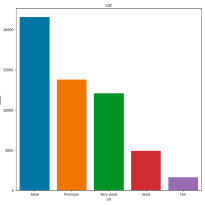
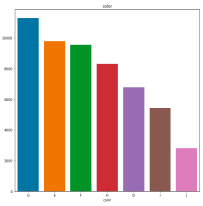
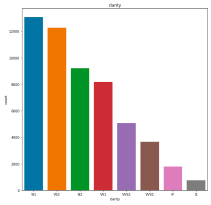
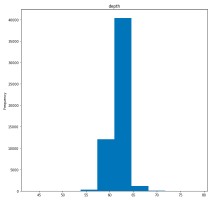
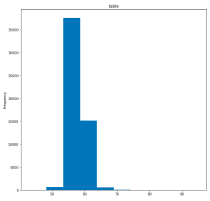
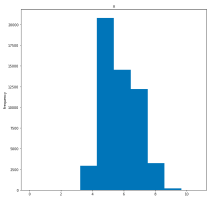
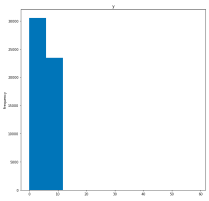
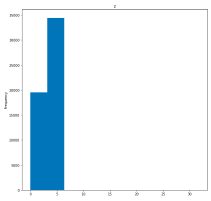
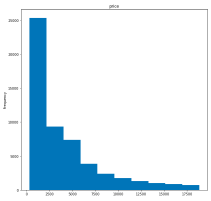
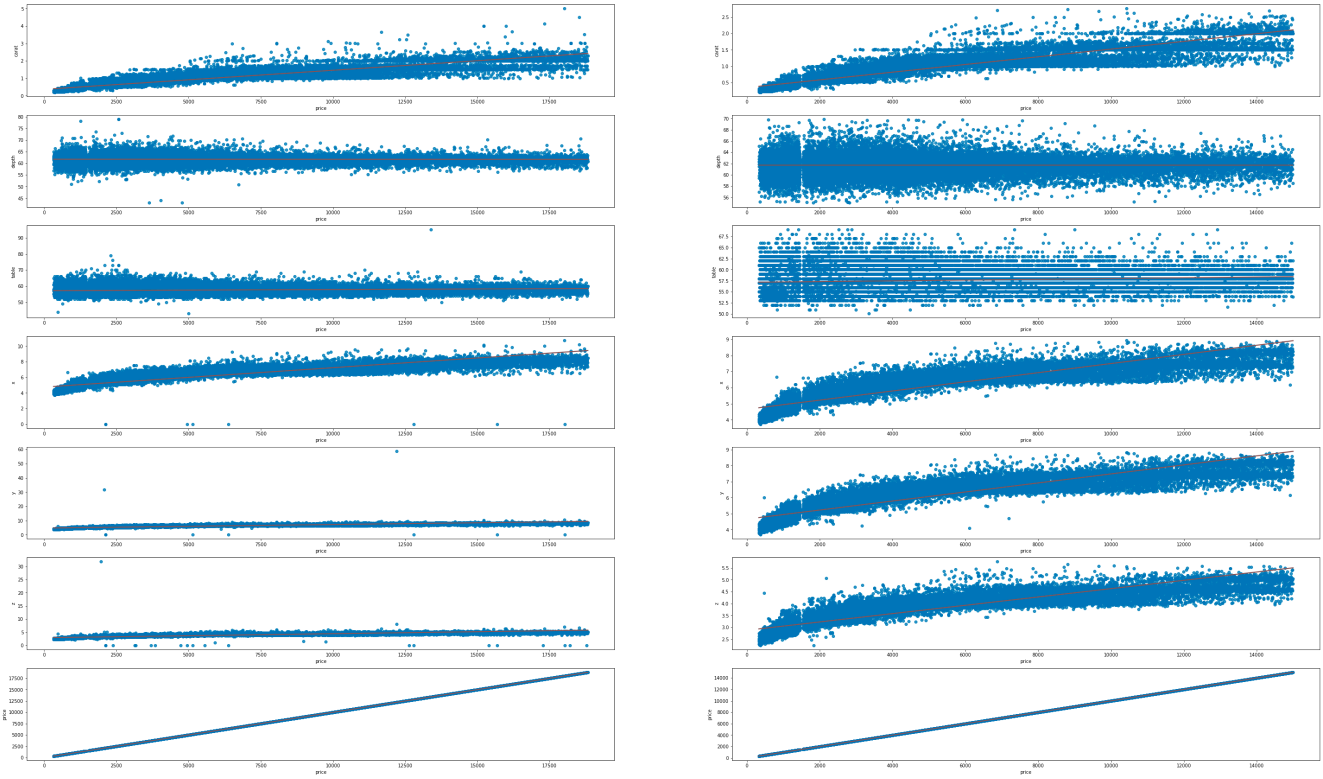

				

table	**x**	**y**	**z**	**price**
				

- Remove outliers

1. In carat plot, remove the points - carat > 2.9
2. In depth plot, remove the points - depth > 70 || depth <= 55
3. In table plot, remove the points - table >= 70 || table <= 50
4. In x plot, remove the points - x >= 9 && price >= 15000
5. In y plot, remove the points - y >= 20 || y == 0
6. in z plot, remove the points - z >= 6 || z <= 1

- Right(origin), Left(after removing outliers)



- Encode categorical features based on diamond documentation
  - cut

Ideal	Predium	Very Good	Good	Fair
100	80	60	40	20

  - color
    - One Hot Encode
  - clarity

I1	SI2	SI1	VS2	VVS2	VVS1	IF
30	40	50	60	70	80	90

- Step 5. Build classification (or regression) models using the training data && Step 7. Assess model on the test data.

Model	Parameters	MSE	RMSE	RSE	MAE	excution time
linear regression	positive = True	1647909.22	1283.71	0.13	816.89	0.02s
k-neighbors regression	Default	1339014.10	1157.16	0.12	554.29	1.49s
Ridge regression	Default	2190847.01	1480.15	0.21	848.85	0.004s
decision tree regression	Max_depth = None	825284.43	908.45	0.06	413.07	0.02s
random forest regression	n_estimators = 1000	632325.04	795.19	0.05	336.00	1m50.00s
gradient Boosting regression	Max_depth = none	791343.44	889.57	0.06	401.06	17.83s
SGD regression	Default	2178494.94	1475.97	0.22	864.34	0.20s
support vector regression (SVR)	C=1500	998458.52	999.23	0.09	524.38	3m6.66s
linear SVR	max_iter=50000, C = 5.0, loss = 'squared_epsilon_insensitive' ,dual = True	2201090.06	1483.61	0.21	848.94	10.78s
multi-layer perceptron regression	max_iter=5000	570093.37	755.05	0.04	391.20	3m22.46s

Discussion