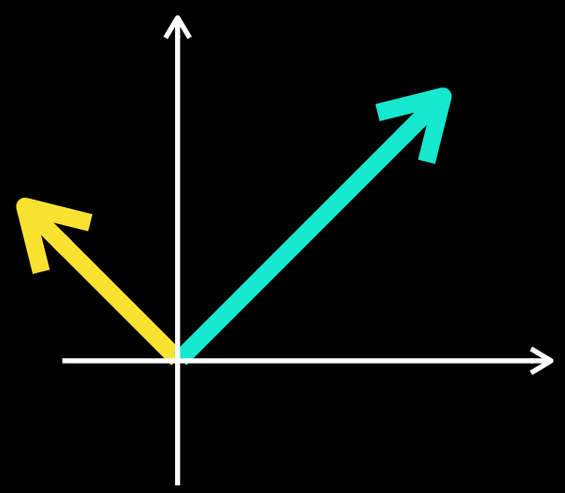
Linear regression: Predicting values of diamonds

Linear Algebra Essentials



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

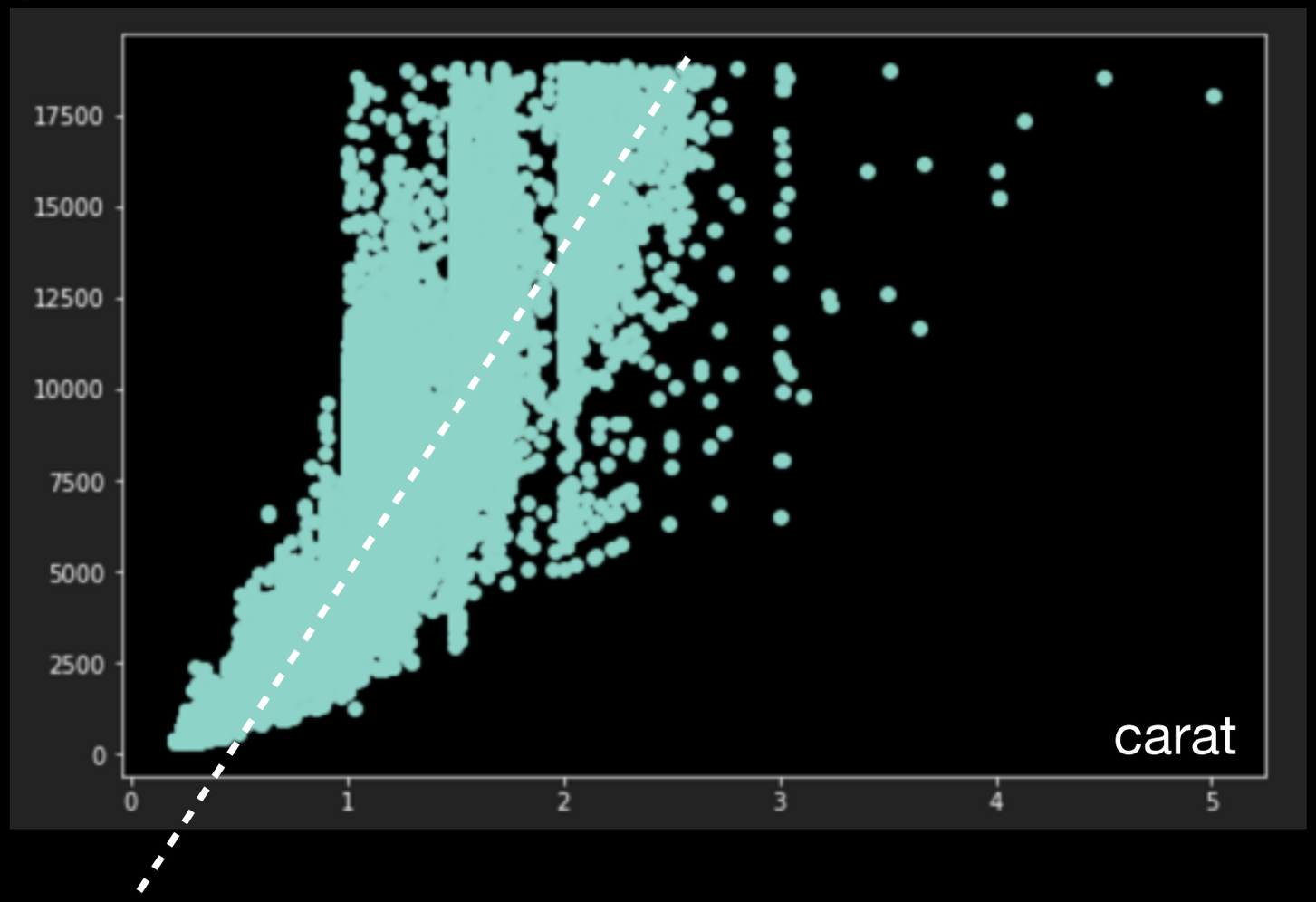
```
data = pd.read_csv("diamonds.csv")
data.head()
```

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

cut: { Fair, Good, Very Good, Premium, Ideal } —> { 0, 1, 2, 3, 4, }

color: { D, E, … , J } —> { 0, 1, … }

price



$$per\text{-}carat = \frac{price}{carat^2}$$

	carat	cut	color	clarity	depth	table	price	x	y	z	per-carat	_cut	_color	_clarity
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43	6162.570888	4	1	1
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31	7392.290249	3	1	2
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31	6181.474480	1	1	4
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63	3971.462545	3	5	3
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75	3485.952133	1	6	1

```python
y = data['per-carat'].values
X = data[['_cut', '_color', '_clarity', 'carat', 'x', 'y', 'z']].values.T
X = np.vstack([np.ones(len(y)), X])
X[:,:5]
```

```
array([[1.  , 1.  , 1.  , 1.  , 1.  ],
       [4.  , 3.  , 1.  , 3.  , 1.  ],
       [1.  , 1.  , 1.  , 5.  , 6.  ],
       [1.  , 2.  , 4.  , 3.  , 1.  ],
       [0.23, 0.21, 0.23, 0.29, 0.31],
       [3.95, 3.89, 4.05, 4.2 , 4.34],
       [3.98, 3.84, 4.07, 4.23, 4.35],
       [2.43, 2.31, 2.31, 2.63, 2.75]])
```

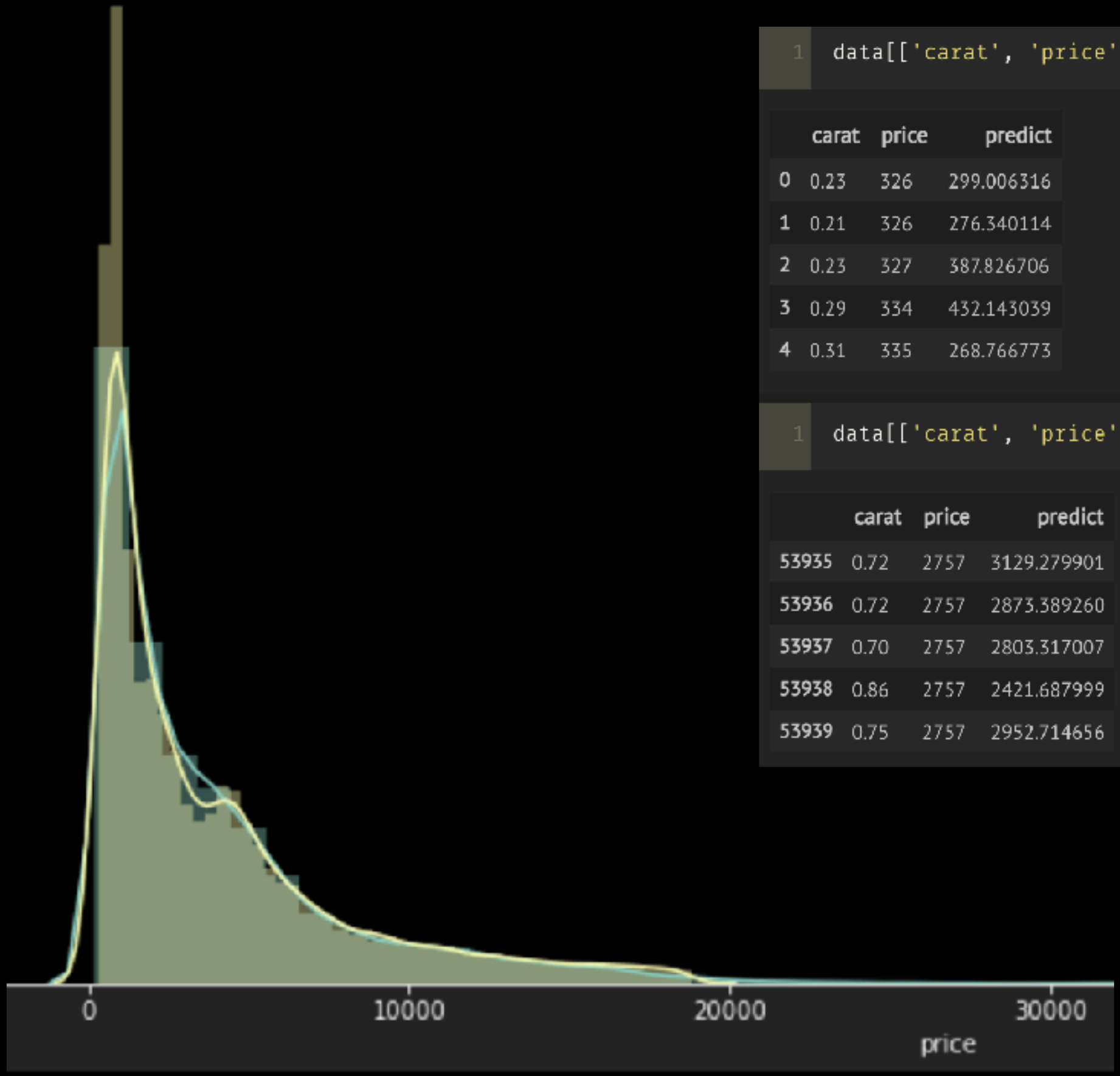
```python
a = np.linalg.inv(X.dot(X.T)).dot(X).dot(y)
a
```

```
array([7284.4611339 ,  170.43665284, -426.58684553,  744.52169235,
       1050.68424425, -579.38546073,  -49.47883071, -159.67421698])
```

```python
data['predict'] = X.T.dot(a) * X[4, :]**2
```

```python
(np.abs(data['predict'] - data['price']) / (data['predict'] + data['price'])).mean()
```

```
0.05982607651939629 6
```




```
data[['carat', 'price', 'predict']].head()
```

	carat	price	predict
0	0.23	326	299.006316
1	0.21	326	276.340114
2	0.23	327	387.826706
3	0.29	334	432.143039
4	0.31	335	268.766773

```
data[['carat', 'price', 'predict']].tail()
```

	carat	price	predict
53935	0.72	2757	3129.279901
53936	0.72	2757	2873.389260
53937	0.70	2757	2803.317007
53938	0.86	2757	2421.687999
53939	0.75	2757	2952.714656