The Lapidarist Problem

noviembre 24

Introduction

Some diamonds have been stolen. We have a huge dataset, containing characteristics and prices for 53930 of diamonds.

With the characteristics of the missing diamonds we want to estimate how much the stolen diamonds are worth.

Objective

Know the value of diamonds with the following characteristics:

Cut	Color	Clarity	Depth	Table	x	y	z	latitude	longitude
Good	I	VVS2	63.1	58	5.64	5.71	3.58	35.02636,	-114.3835
Ideal	G	VS1	62.1	55	6.02	6.05	3.75	35.0035	-109.7896
Ideal	E	VS2	61.5	55	5.11	5.16	3.16	35.10544	-106.6697
Premium	J	VS1	61.6	59	4.67	4.71	2.89	34.9466	-104.6473
Premium	G	VS1	62.1	56	4.43	4.4	2.74	35.18864	-101.986
Good	F	SI2	63.3	57	6.08	6.14	3.87	35.26611	-99.63874
Ideal	D	VS1	60.9	57	5.2	5.17	3.16	35.51572	-97.6708
Ideal	G	VVS2	62.1	54.8	6.64	6.66	4.13	36.163605	-95.7595
Ideal	G	VVS2	62.4	56	4.72	4.74	2.95	37.689186	-92.6473
Premium	I	VS2	62.7	59	4.54	4.58	2.86	38.66303	-90.21808

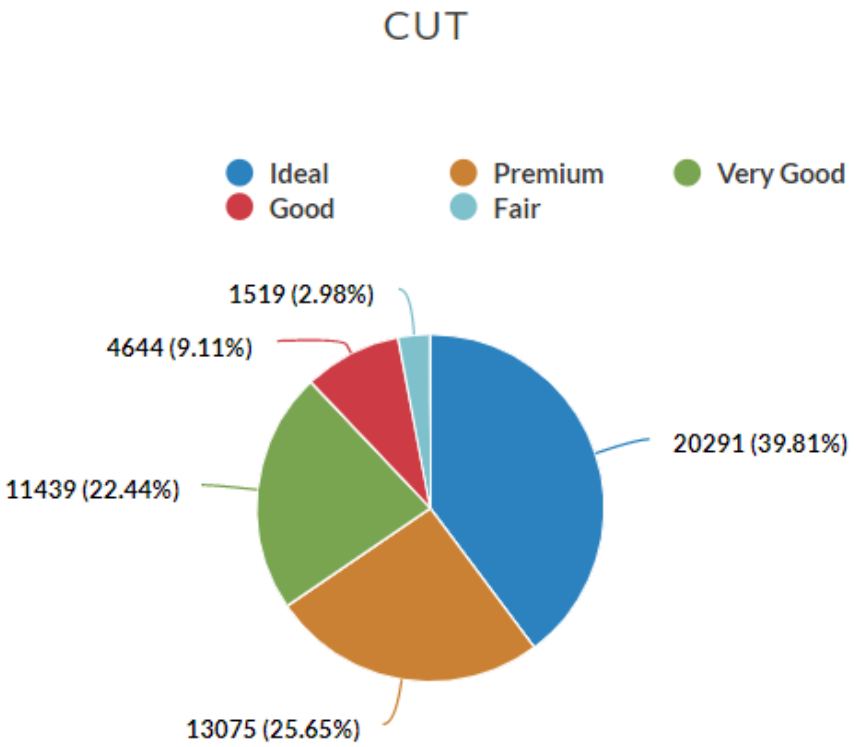
What do we do?

First let's clean the data,removing some special characters and rows containing incomplete data.

Let's explore the data

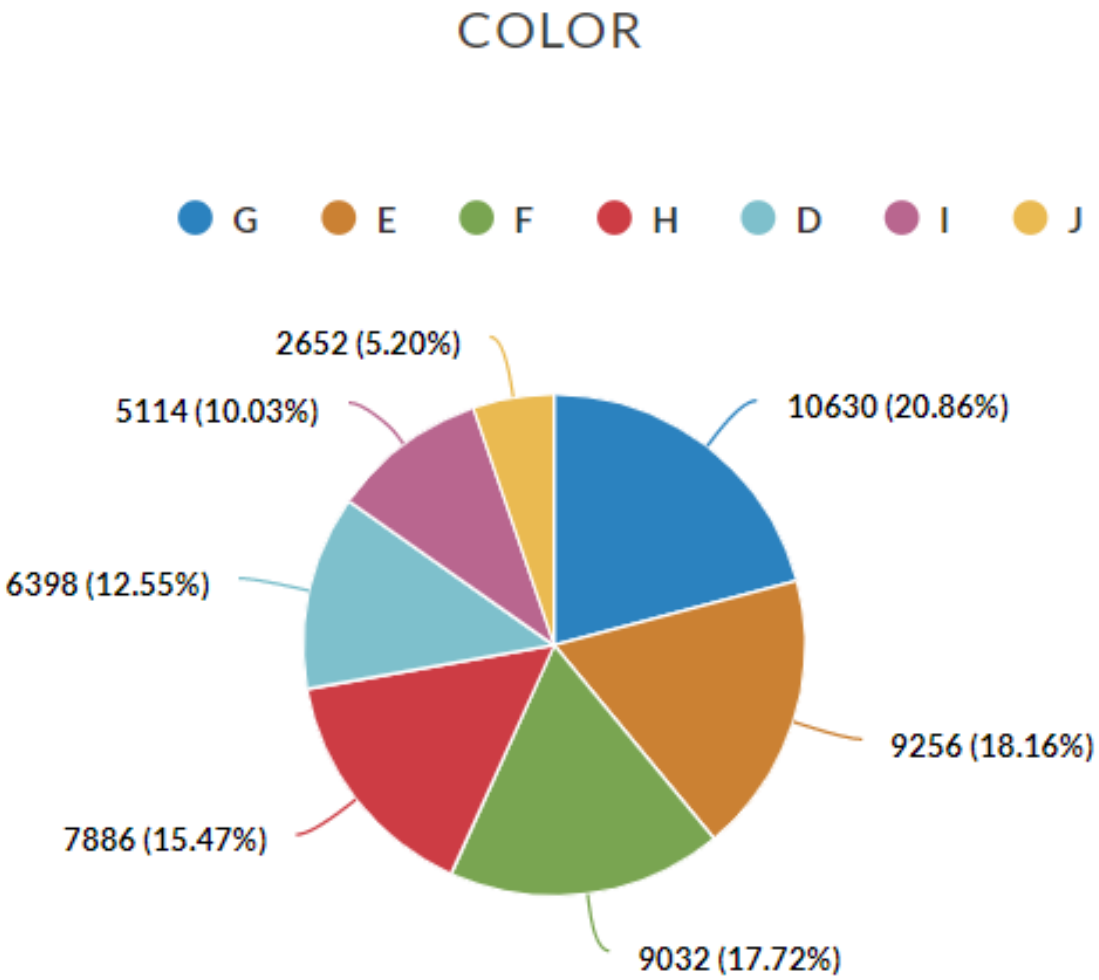
Let's analyze some characteristics:

We have five different cuts with different counts

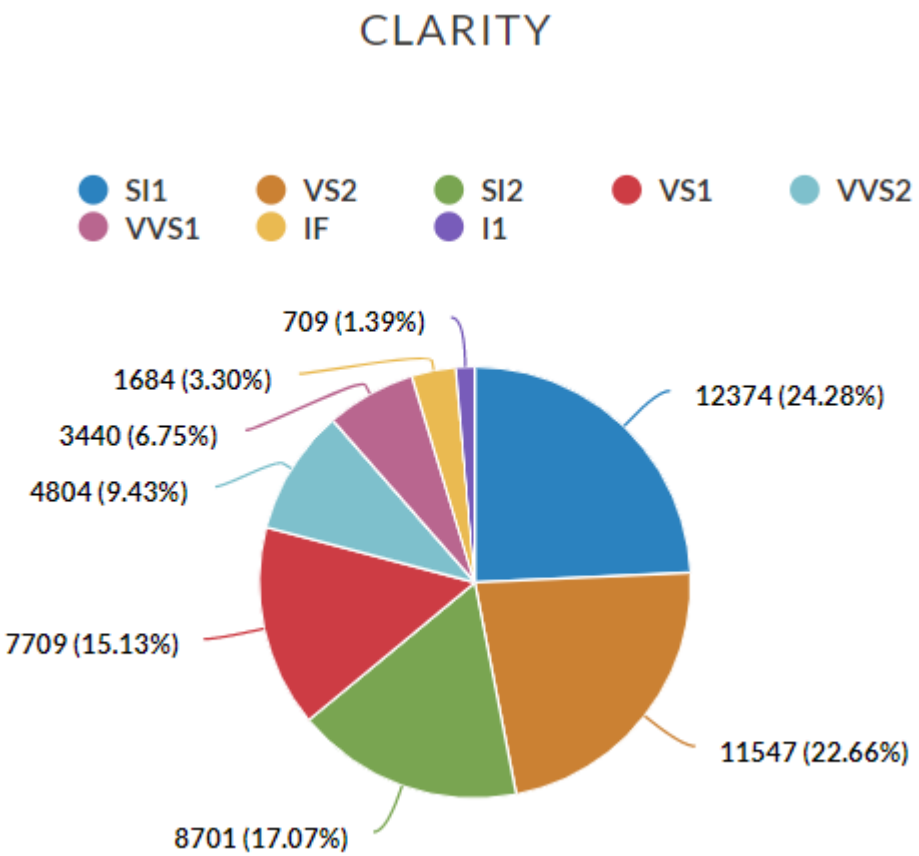


And we have seven different colors.

The most popular is J.



And we have eigth kinds of Clarity.



Algorithm implementation

For this problem, I implemented the K-neighbors algorithm, considering eight neighbors and only some provided features.

I divided the data set into 70% training and 30% testing.

The score on the test set had a precision of 0.83

Calculating the cost of stolen diamonds

Implementing the algorithm with the characteristics provided, we obtain the following respective costs in the order at the beginning of the table:

3899.25
4688.375
1847.
1069.125
784.875
4451.
1782.75
6917.875
928.75
870.625