Machine Learning Assigment-2

Part 1

DATASET-1

PENGUINS

This dataset describes categorizing a penguin based on it's features which are the seven columns namely:

- **species** a factor denoting penguin species (Adélie, Chinstrap and Gentoo)
- **island** a factor denoting island in Palmer Archipelago, Antarctica (Biscoe, Dream or Torgersen)
- **bill_length_mm** a number denoting bill length (millimeters)
- **bill_depth_mm** a number denoting bill depth (millimeters)
- **flipper_length_mm** an integer denoting flipper length (millimeters)
- **body_mass_g** an integer denoting body mass (grams)
- **sex** a factor denoting penguin sex (female, male)

Of the 7 columns, 3 are categorical (species,island,sex) and the rest are numeric.

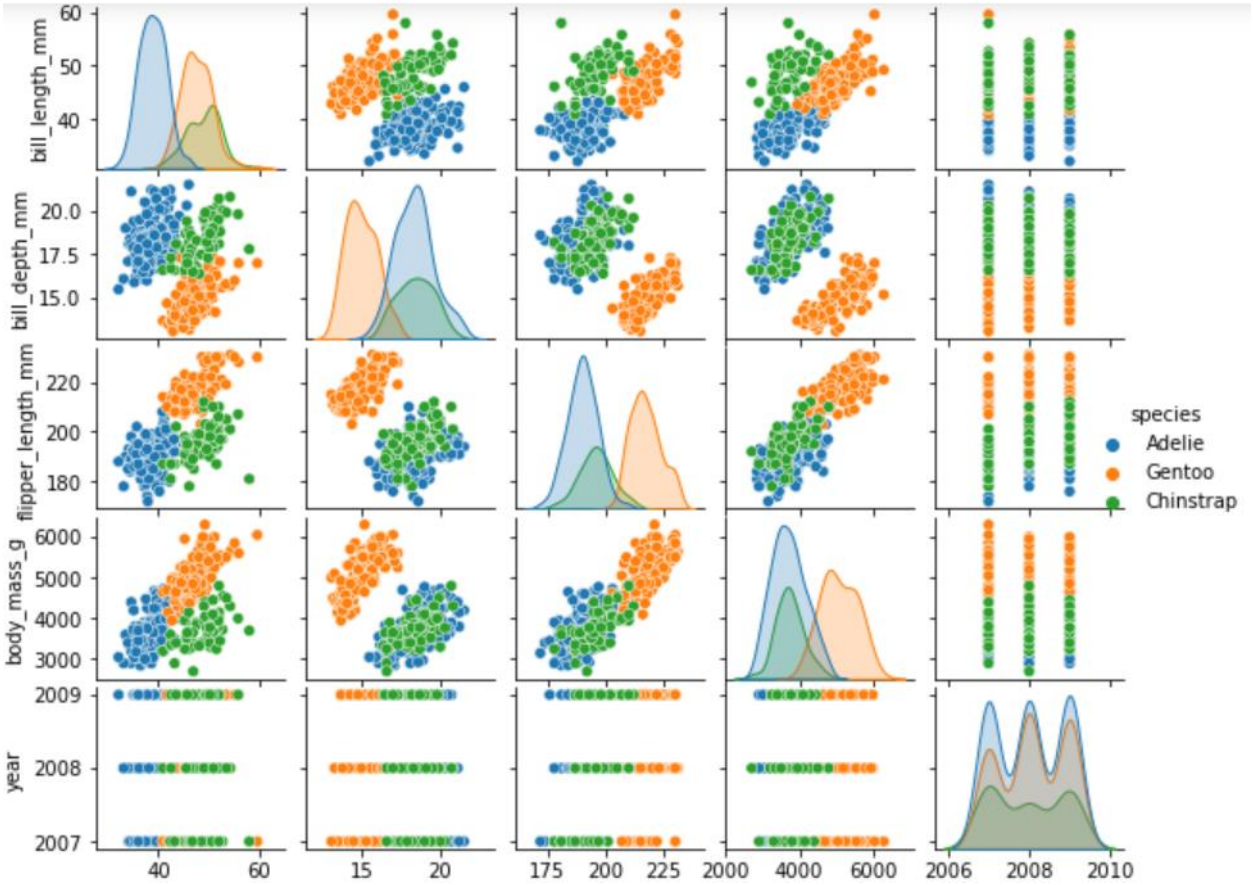
The dataset consists of 344 entries with each of them having different features.

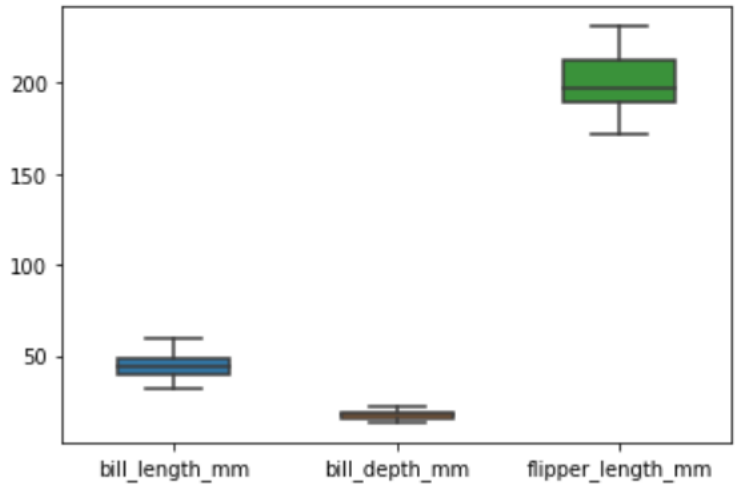
Statistics:

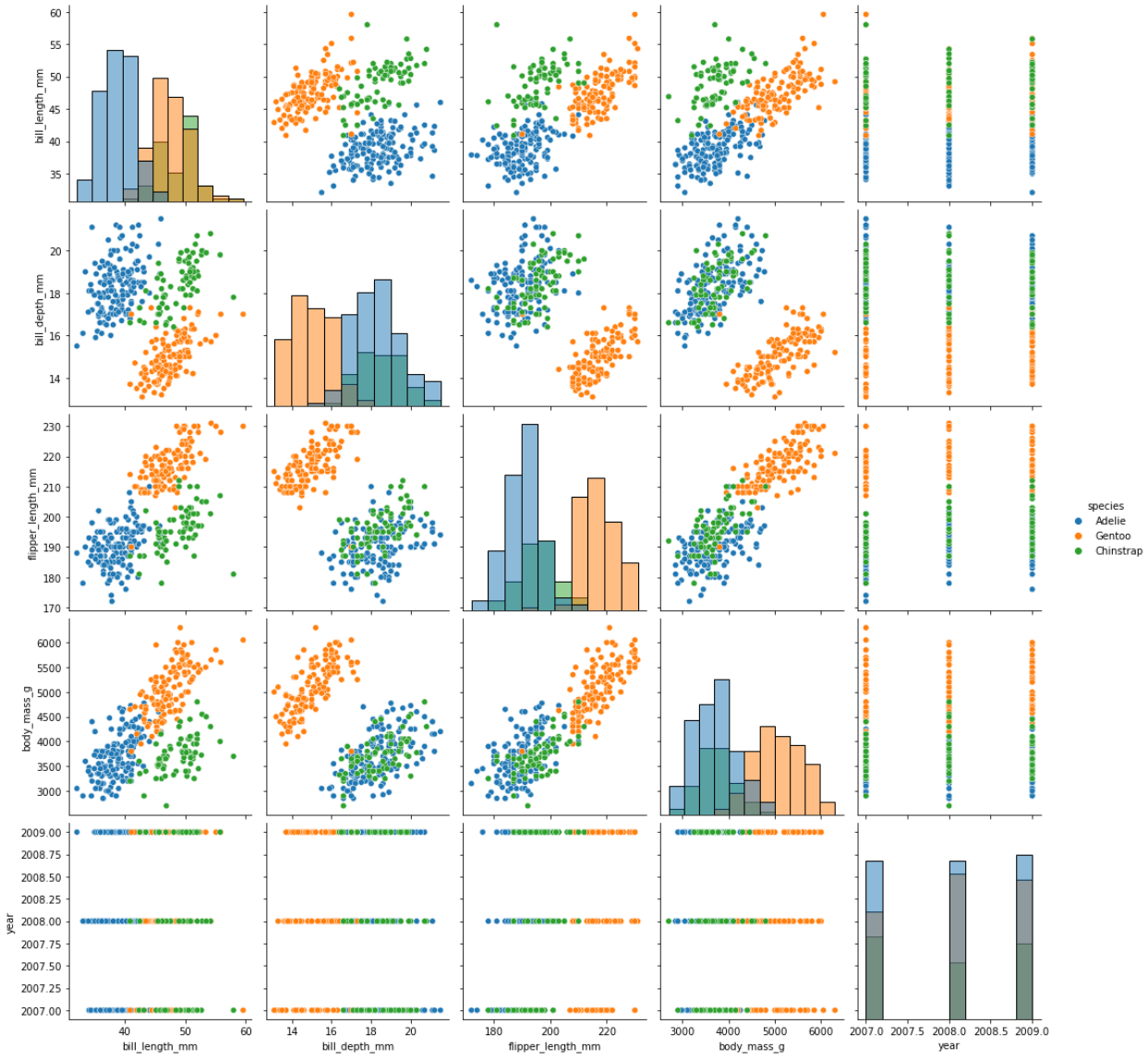
	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	year
count	344.000000	344.000000	344.000000	344.000000	344.000000
mean	43.905523	17.150291	200.851744	4199.418605	2008.029070
std	5.447882	1.969061	14.045266	800.197923	0.818356
min	32.100000	13.100000	172.000000	2700.000000	2007.000000
25%	39.275000	15.600000	190.000000	3550.000000	2007.000000
50%	44.250000	17.300000	197.000000	4025.000000	2008.000000
75%	48.500000	18.700000	213.000000	4750.000000	2009.000000
max	59.600000	21.500000	231.000000	6300.000000	2009.000000

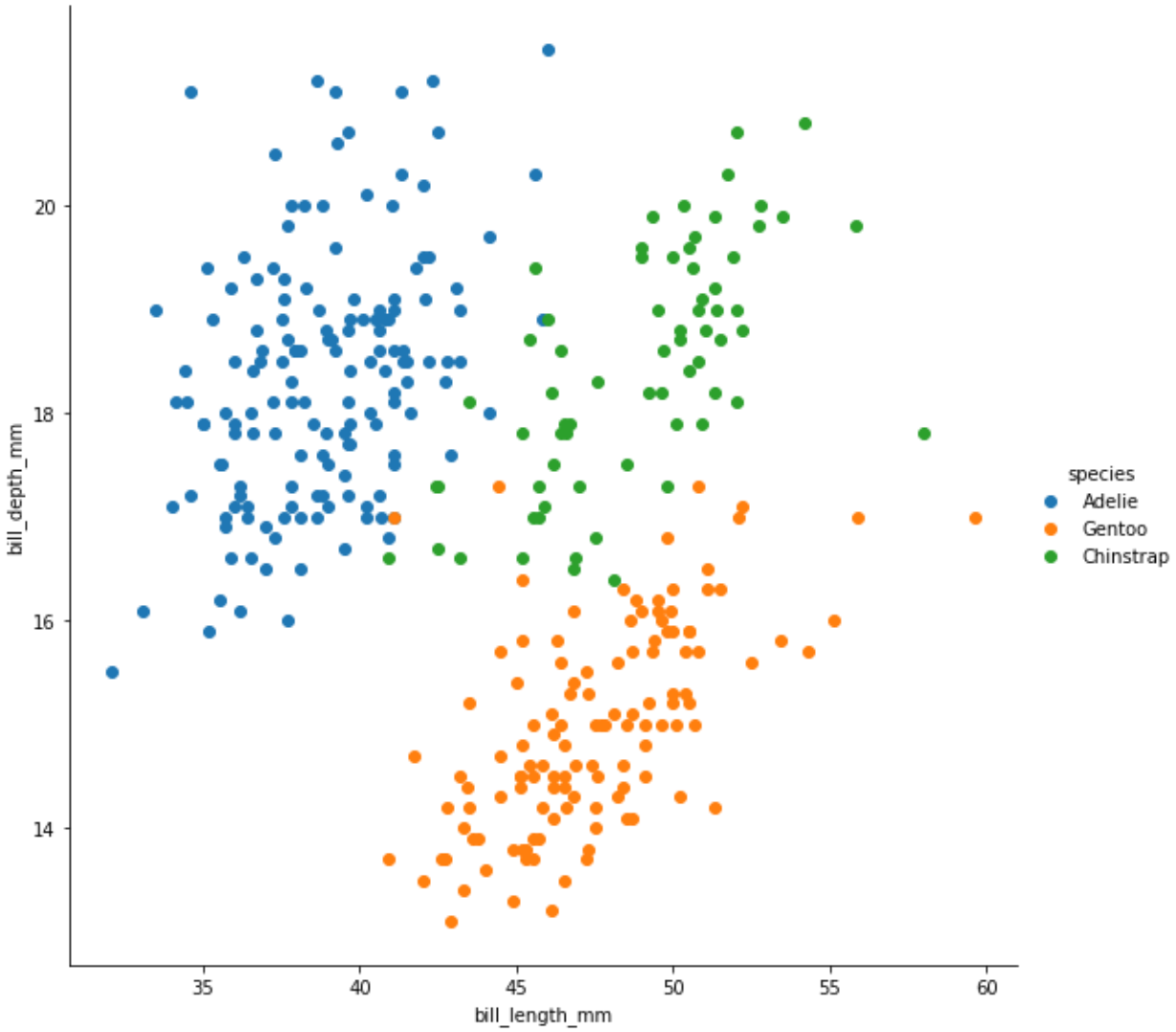
Data visualization:

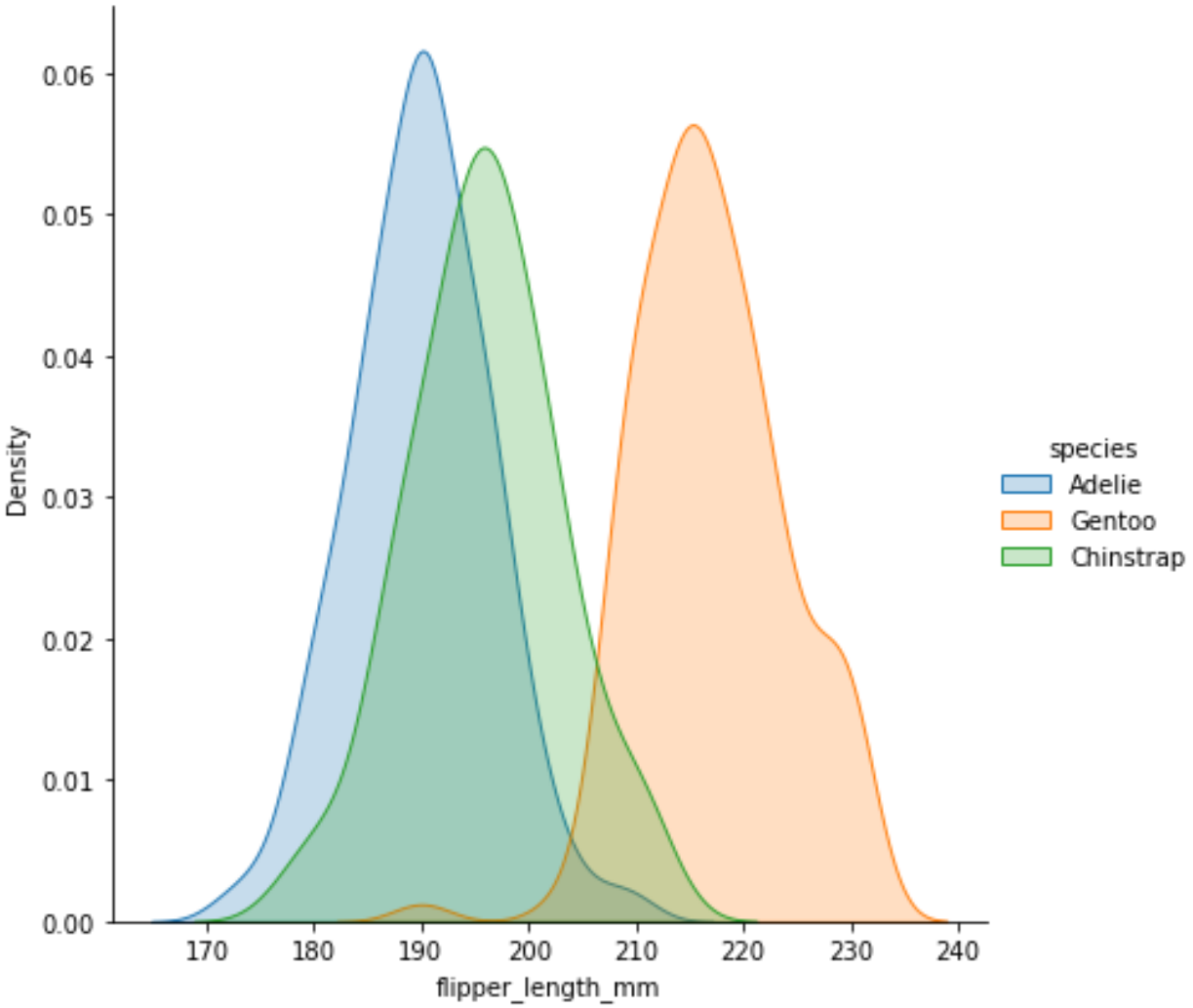
1. Below is a pairplot() – plot pairwise relationships of the penguin dataset, with species as hue that help us identify various species easily.




200
150
100
50
bill_length_mm
bill_depth_mm
flipper_length_mm


bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
year
species
Adelie
Gentoo
Chinstrap


bill_depth_mm
20
18
16
14
35
40
45
50
55
60
bill_length_mm
species
Adelie
Gentoo
Chinstrap


0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
170
180
190
200
210
220
230
240
flipper_length_mm
species
Adelie
Gentoo
Chinstrap

DATASET- 2

AMAZON

This dataset consists of Amazon top selling books categorized based on name, author, user rating, reviews, price, year, and genre.

Genre being the categorical variable, decides whether the book is fiction or non-fiction. Price and user-rating is numeric data.

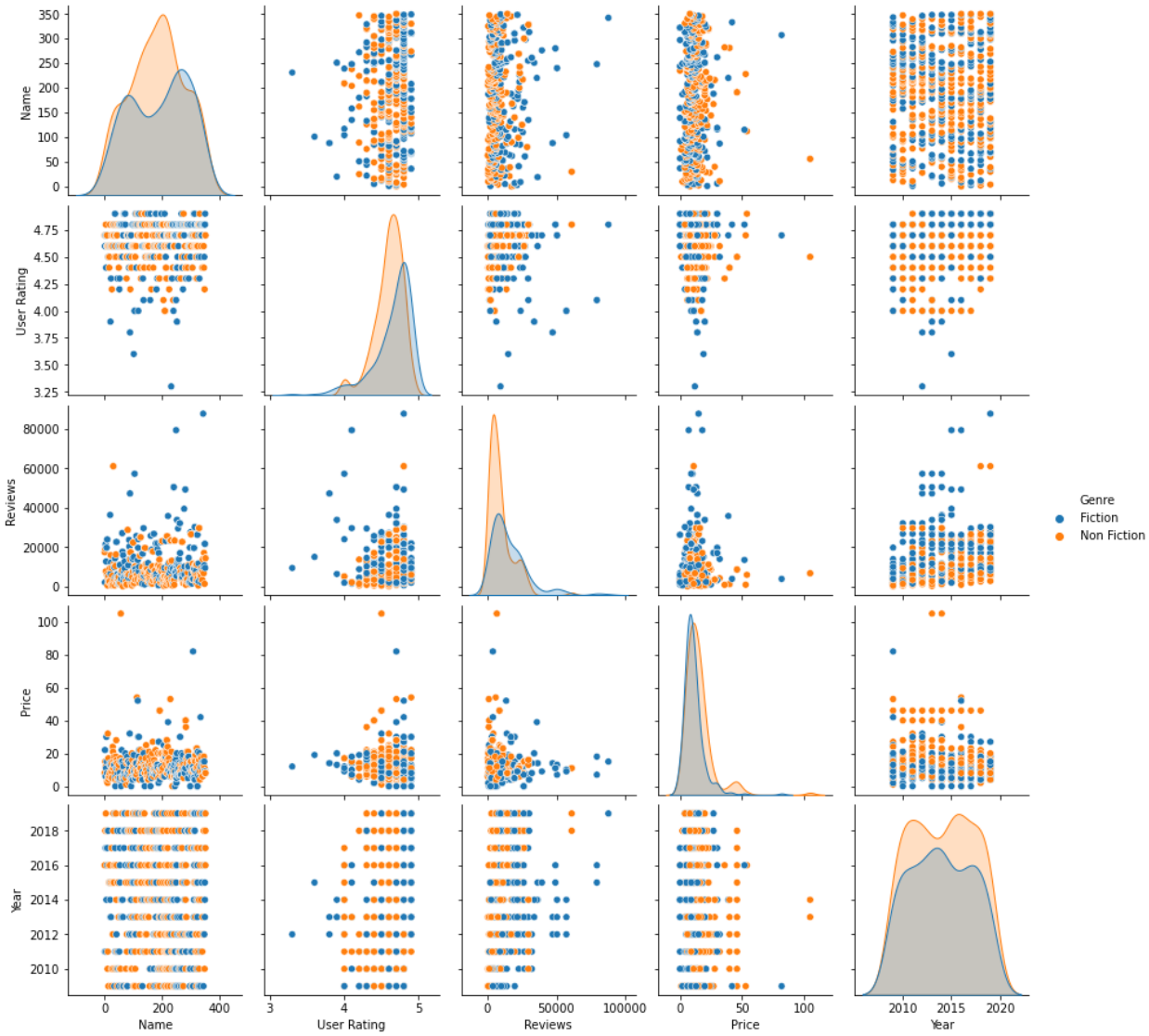
The dataset consists of 550 records with 7 columns as its features.

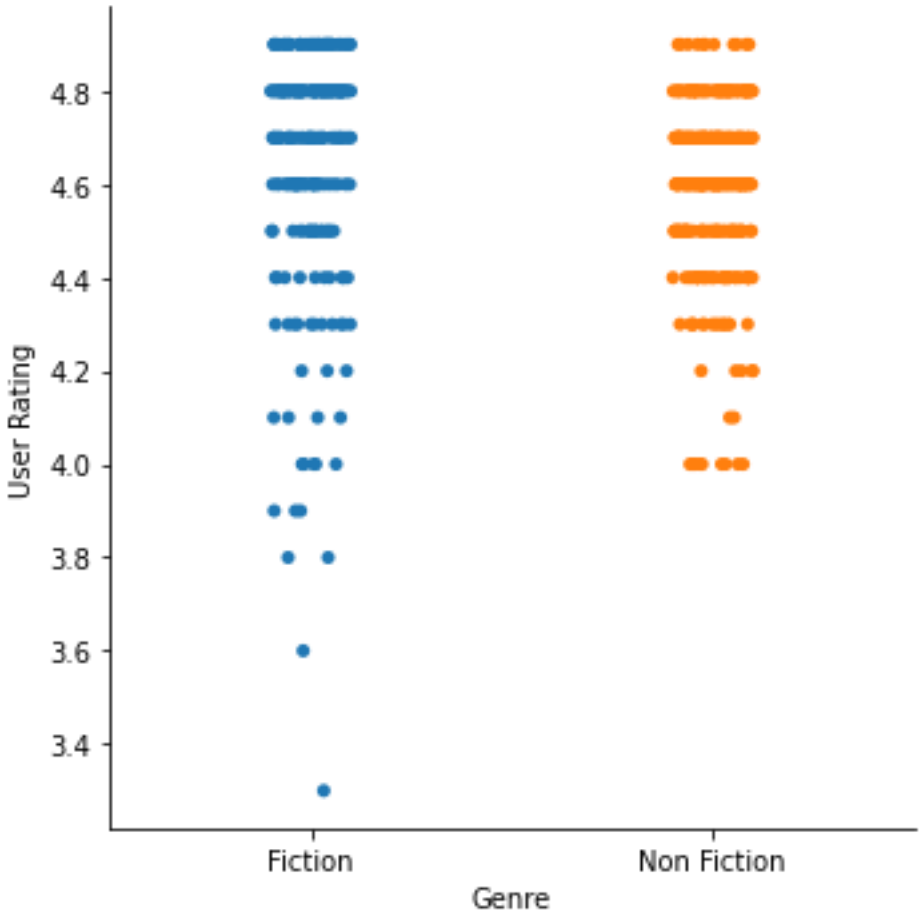
Statistics:

	User Rating	Reviews	Price	Year
count	550.000000	550.000000	550.000000	550.000000
mean	4.618364	11953.281818	13.100000	2014.000000
std	0.226980	11731.132017	10.842262	3.165156
min	3.300000	37.000000	0.000000	2009.000000
25%	4.500000	4058.000000	7.000000	2011.000000
50%	4.700000	8580.000000	11.000000	2014.000000
75%	4.800000	17253.250000	16.000000	2017.000000
max	4.900000	87841.000000	105.000000	2019.000000

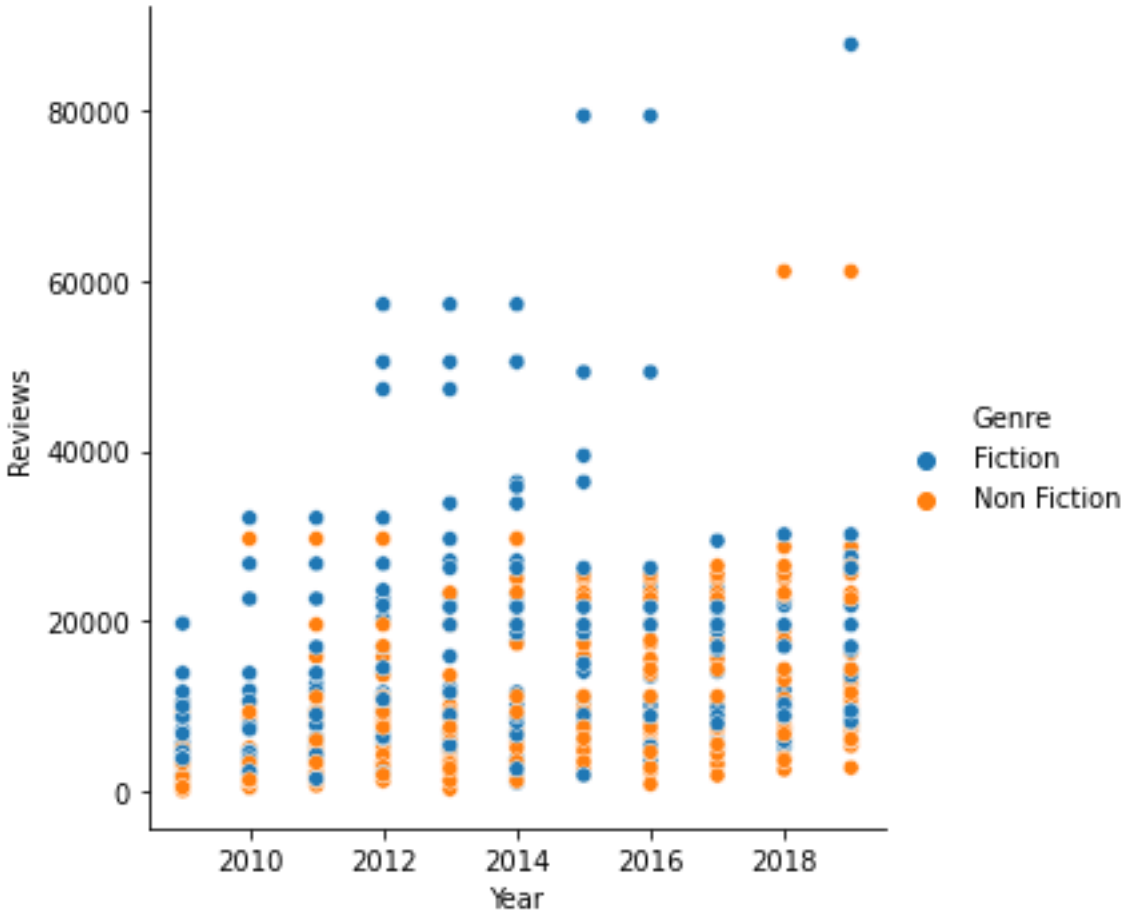
Data Visualization:

Categorizing into fiction or non-fiction based on other features such as price, user rating, name, reviews.

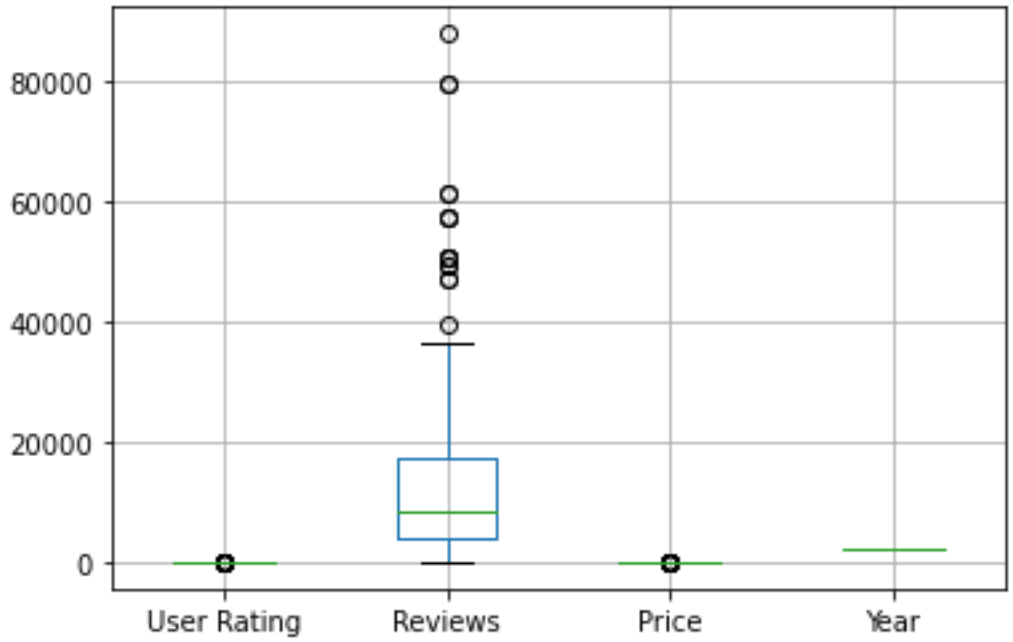


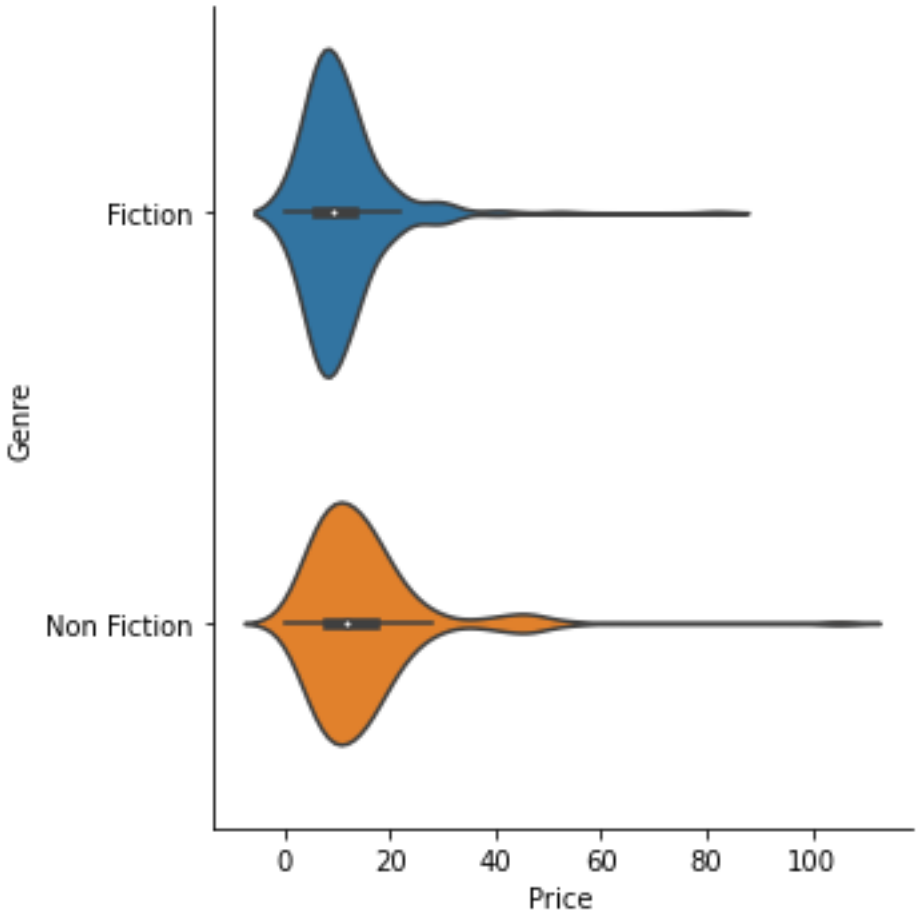


The above plot is classified into genre based on user ratings.



The above graph is plotted between year and reviews posted of various fiction and non-fiction books.




Fiction
Non Fiction
Genre
0
20
40
60
80
100
Price

DATASET - 3

DIAMONDS

The diamonds dataset classifies diamonds based on its cut, carat, dimensions, color, clarity, and depth:

- Carat weight of the diamond.
- Describes cut quality of the diamond. Quality is in increasing order Fair, Good, Very Good, Premium, Ideal.
- Color of the diamond, with D being the best and J the worst.
- Inclusions within the diamond:(in order from best to worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1.
- The height of a diamond, measured from the culet to the table.
- The width of the diamond's table expressed as a percentage of its average diameter.

Cut, color, and clarity are the categorical values.

Carat, depth, table, and dimensions are float values.

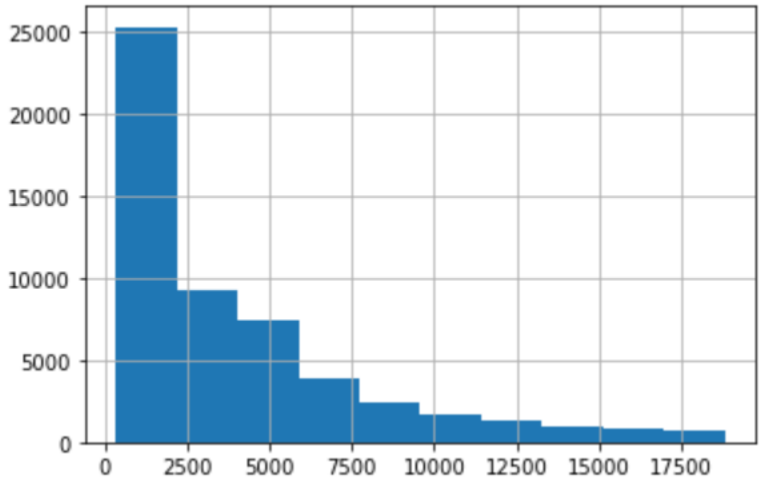
Prices are integers values.

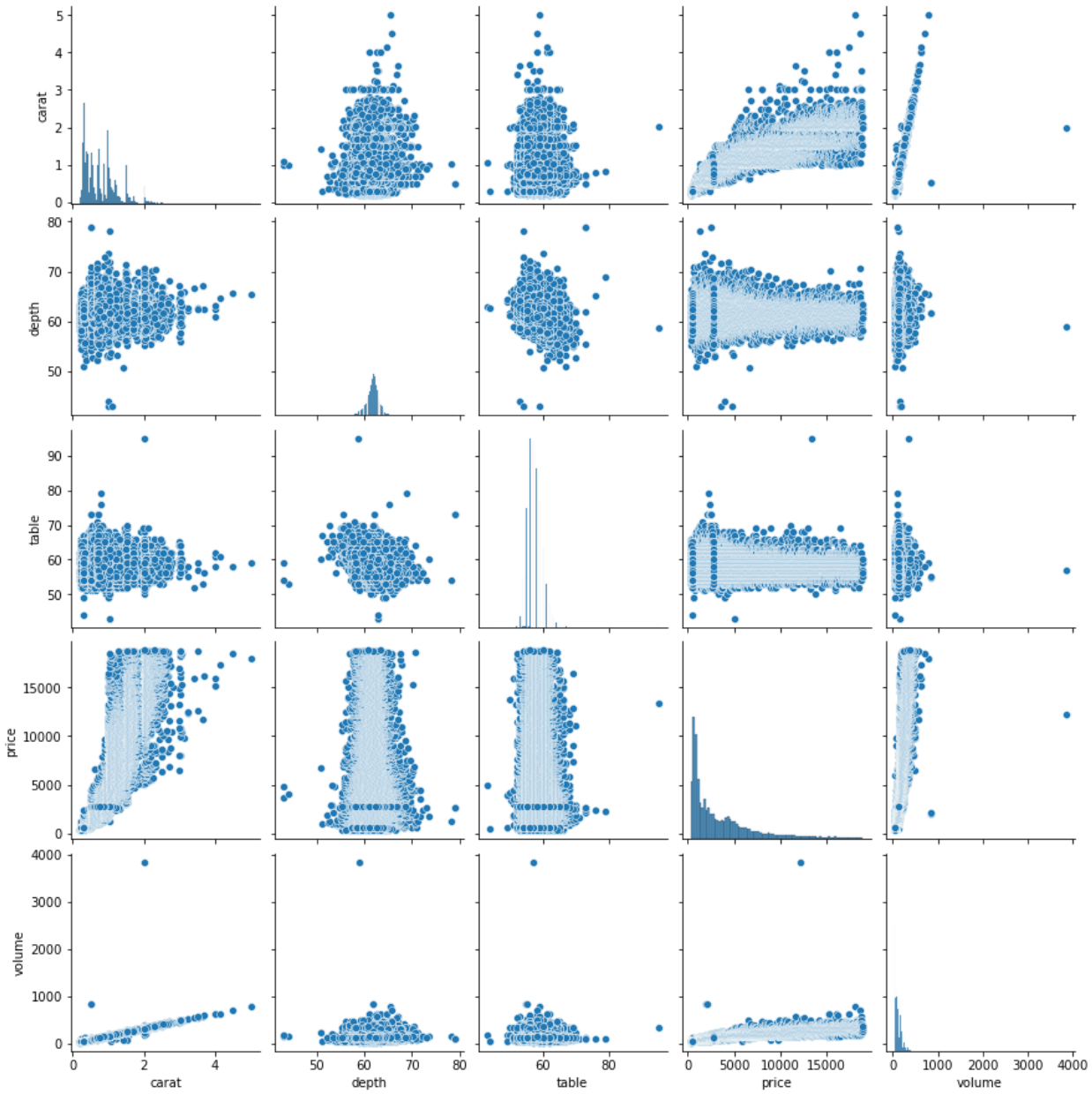
This data set consists of 53940 records and 11 columns as it’s features.

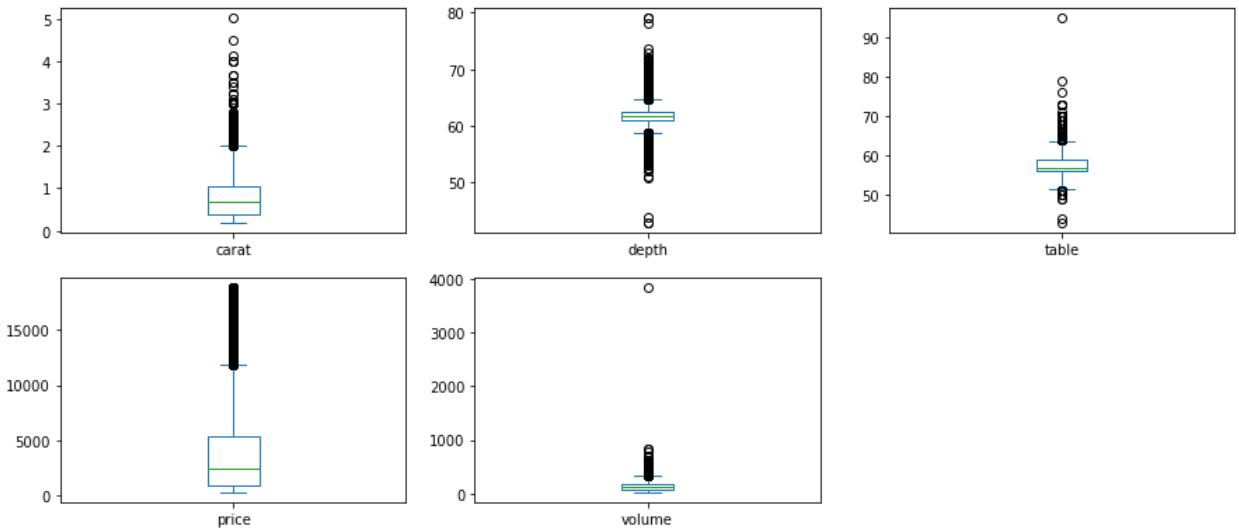
Statistics:

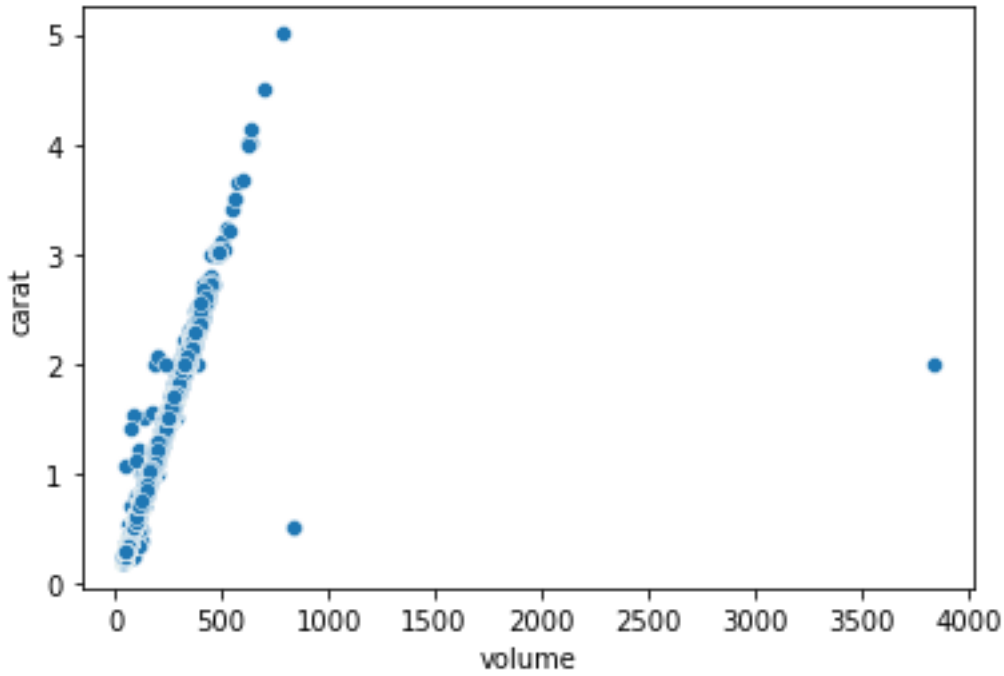
	Unnamed: 0	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	26970.500000	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	15571.281097	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	1.000000	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	13485.750000	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	26970.500000	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	40455.250000	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	53940.000000	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

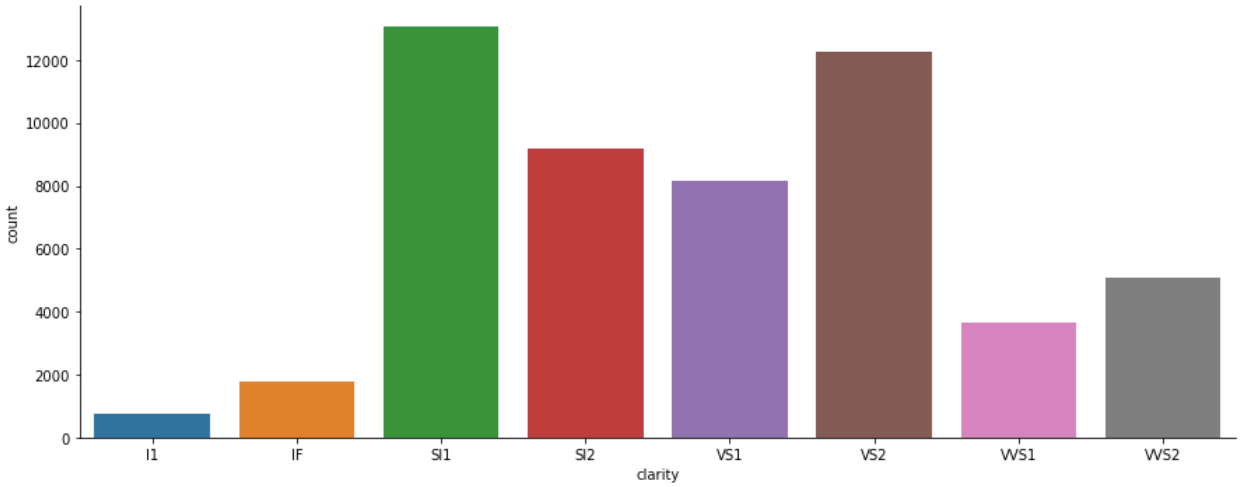
Visualization:




carat
depth
table
price
volume


carat
depth
table
price
volume


carat
volume
0
1
2
3
4
5
0
500
1000
1500
2000
2500
3000
3500
4000


count
clarity
0
2000
4000
6000
8000
10000
12000
I1
IF
SI1
SI2
VS1
VS2
VVS1
VVS2