Chapter 7 - Exercise 3: Visualization with Seaborn - Diamond

Nghịch lý Simpson hay hiệu ứng Yule–Simpson, là một nghịch lý trong xác suất và thống kê, trong đó một xu hướng xuất hiện trong dữ liệu sẽ bị đảo ngược khi được phân tích dưới góc nhìn khác.

Cho file dữ liệu diamonds.csv. Hãy thực hiện các yêu cầu sau, để phát hiện nghịch lý Simpson khi phân tích giá kim cương bằng các công cụ trực quan hóa dữ liệu:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set_style("darkgrid")
```

```
# Câu 1: Đọc dữ liệu diamonds.csv, đưa vào biến diamonds
diamonds = pd.read_csv(r'data/diamonds.csv')
diamonds.head()
```

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

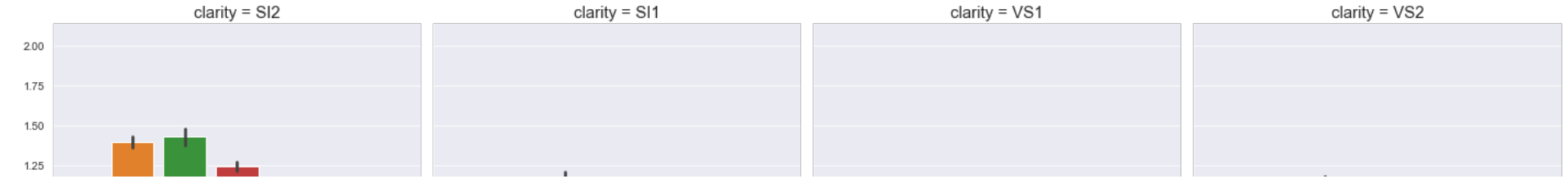
```
fig.suptitle('Price Decreasing with Increasing Quality?',fontsize=15)
```

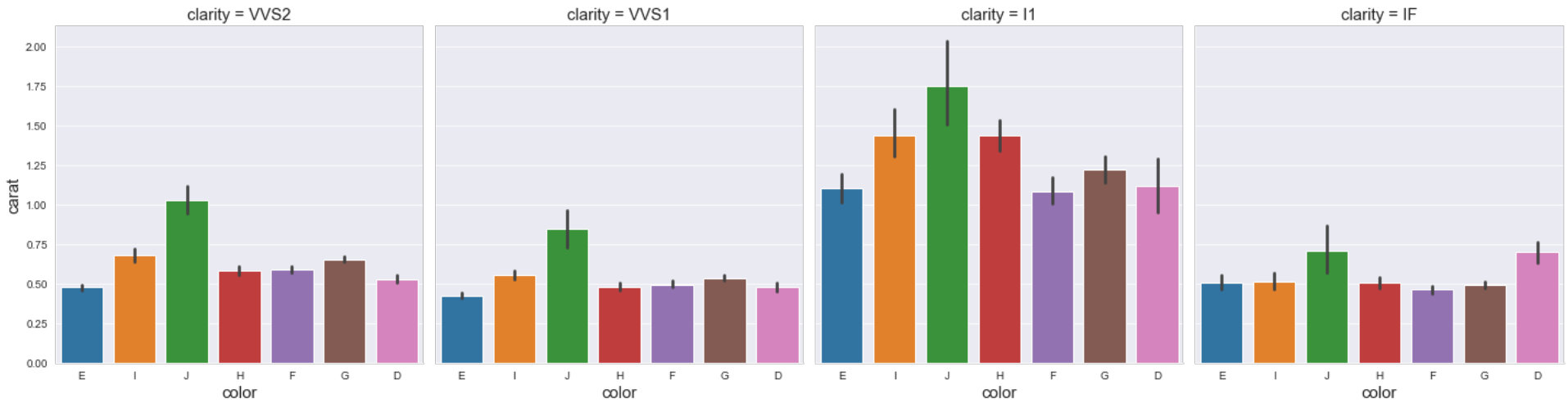
```
Text(0.5, 0.98, 'Price Decreasing with Increasing Quality?')
```

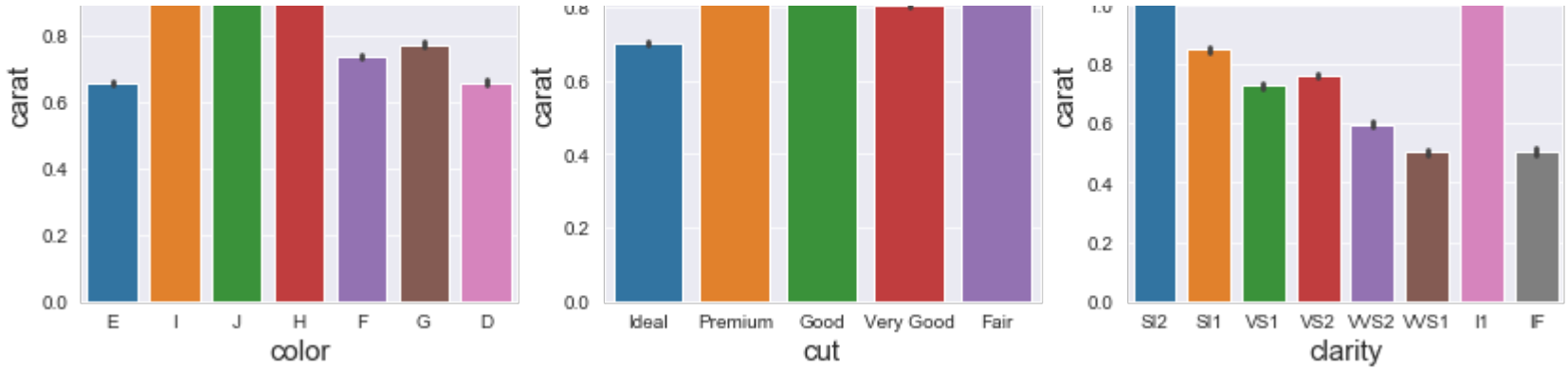
> Có gì đó không ổn ???

```
# Câu 3: Bây giờ, hãy thử Phân tích chi tiết hơn thuộc tính 'carat' theo 'color' và 'clarity' qua biểu đồ catplot - bar
# Bạn nhận xét gì qua biểu đồ này
plt.rcParams["axes.labelsize"] = 15
sns.catplot(x='color', y='carat', col='clarity', col_wrap=4, data=diamonds, kind='bar')
```

```
<seaborn.axisgrid.FacetGrid at 0x5aa11c8e08>
```

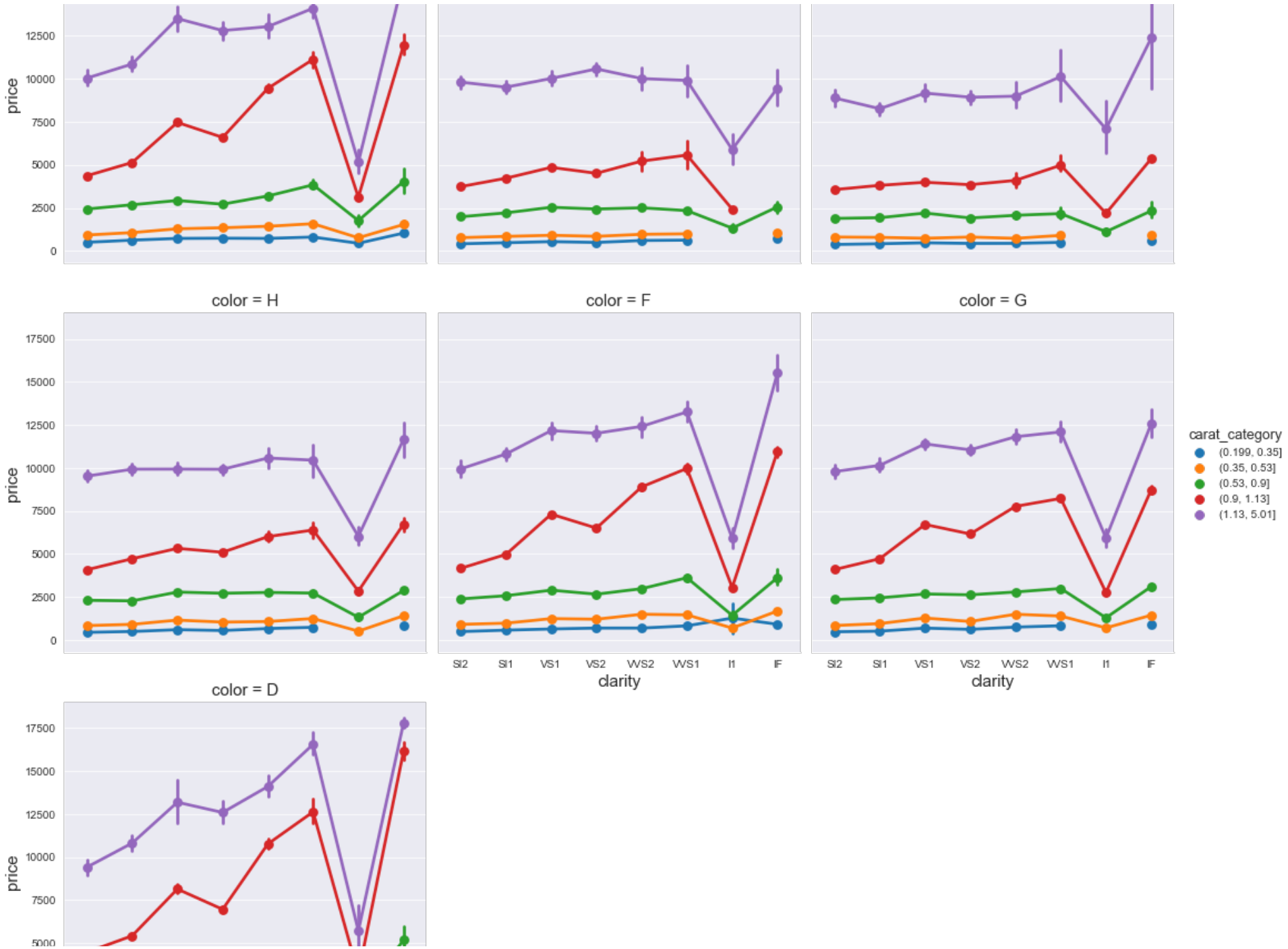



clarity = VVS2
clarity = VVS1
clarity = I1
clarity = IF
carat
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
E I J H F G D
color




```
# Câu 5: Hãy chia carat ra làm 5 khoảng giá trị, tạo cột diamonds['carat_category'] chứa khoảng giá trị tương ứng
# Hướng dẫn: sử dụng hàm pd.qcut
diamonds['carat_category'] = pd.qcut(diamonds.carat, 5)
diamonds.head()
```

	carat	cut	color	clarity	depth	table	price	x	y	z	carat_category
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43	(0.199, 0.35]
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31	(0.199, 0.35]
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31	(0.199, 0.35]
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63	(0.199, 0.35]
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75	(0.199, 0.35]


color = H
color = F
color = G
color = D
price
clarity
SI2
SI1
VS1
VS2
VVS2
VVS1
I1
IF
carat_category
(0.199, 0.35]
(0.35, 0.53]
(0.53, 0.9]
(0.9, 1.13]
(1.13, 5.01]

```
# Câu 7: Kết luận
```