**BÁO CÁO NỘI DUNG THỰC HÀNH**

**Đã hoàn thành :**

**+Trực quan hoá dữ liệu về các sản phẩm danh mục điện thoại trong tiki.vn**

Trong phần xử lý discount e xử lý để chuyển thành dạng số và tính lại giá gốc

#xử lí phần discount  
d = df.discount.tolist()  
for i in range(len(d)):  
 if d[i] == 'Khong giam':  
 d[i]=0  
 else:  
 d[i]=d[i].strip('-%')  
 d[i]=float(d[i])  
df['discount']=d  
df['origin\_price']=df['price']/(100-df['discount'])\*100  
print(df['origin\_price'])  
print(d)

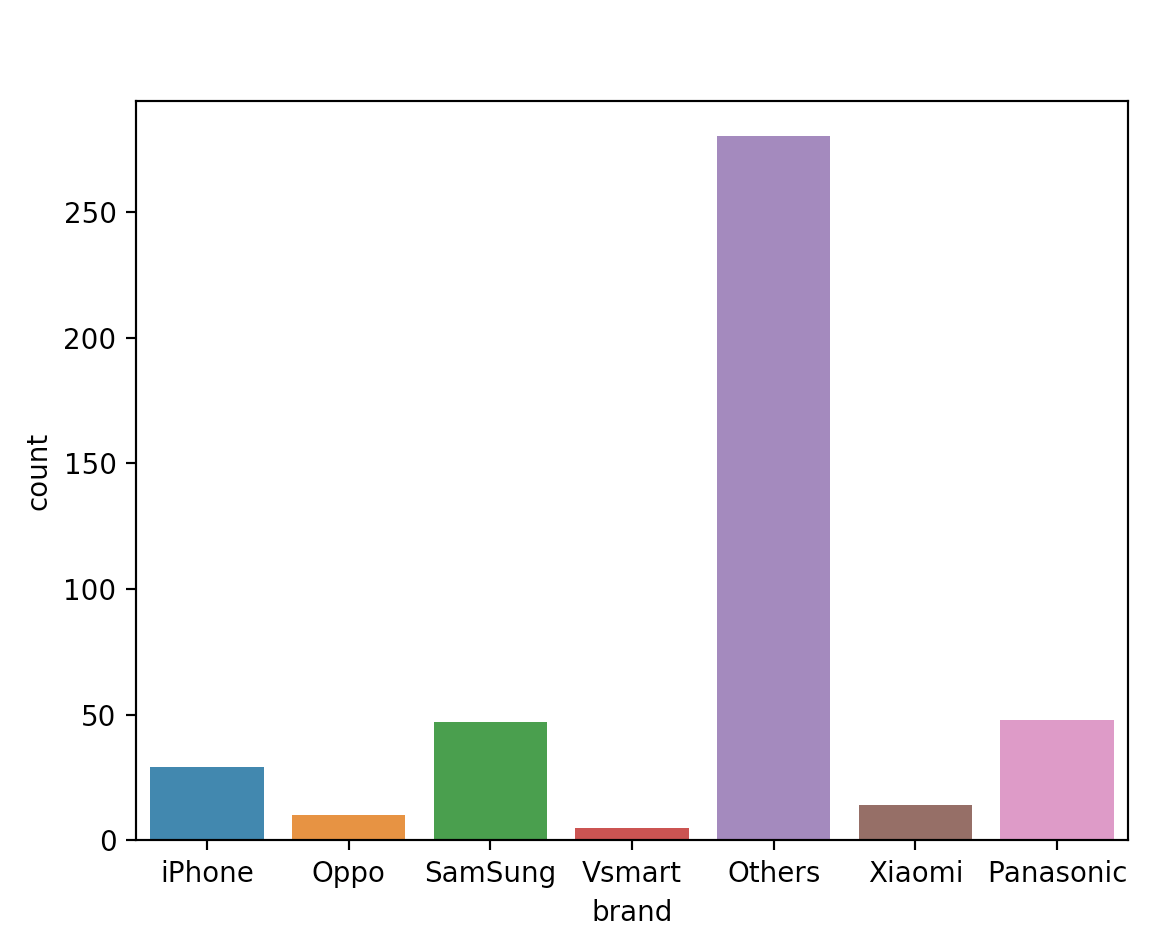
Xử lí phần giá đã giảm (ban đầu crawl được) chuyển thành dạng số

# xử lí phần price  
l= df.price.tolist()  
for i in range(len(l)):  
 l[i]=l[i].replace('đ','')  
 l[i]=l[i].replace('.','')  
 l[i]=int(l[i])  
print(l)  
df['price']=l  
print(df)

Xử lí phần brand để phân loại các hãng điện thoại

def classify(x):  
 if "iPhone" in x :  
 return "iPhone"  
 elif "Samsung" in x :  
 return "SamSung"  
 elif "Vsmart" in x :  
 return "Vsmart"  
 elif "Oppo" in x :  
 return "Oppo"  
 elif "Xiaomi" in x:  
 return "Xiaomi"  
 elif "Panasonic" in x:  
 return "Panasonic"  
 else:  
 return "Others"

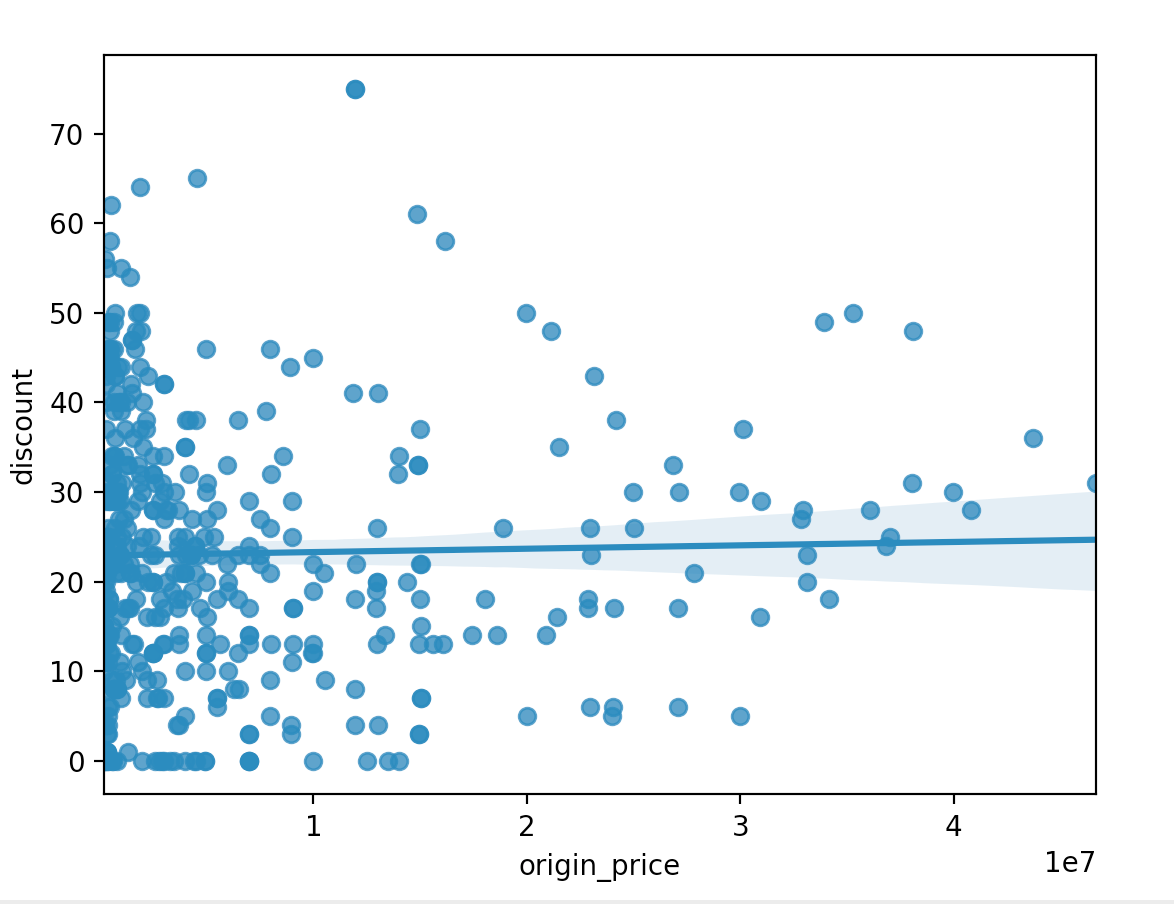
#xu li phan brand  
df['brand']=df['name'].apply(classify)  
print(df['brand'])  
brand\_count = df['brand'].value\_counts()  
print(brand\_count)



Số lượng các sản phẩm được bán theo hãng trên tiki

Sử dụng countplot thể hiện lượng sản phẩm bày bán trên tiki về các hãng phổ biến

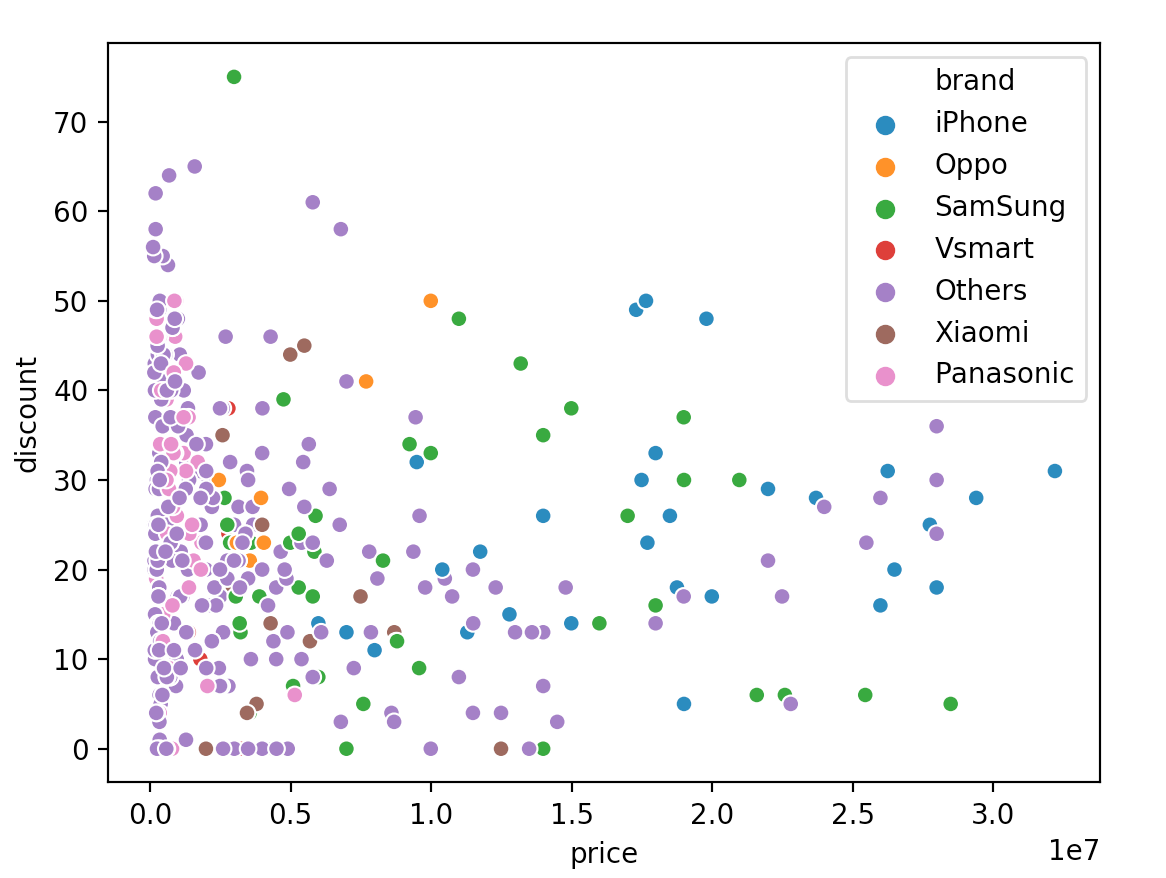
#count plot  
plt.figure(4)  
count=sns.countplot('brand',data=df)

****

biểu đồ giữa giá gốc và giảm giá

Sử dụng reg plot để biểu hiện về quan hệ giảm giá và giá

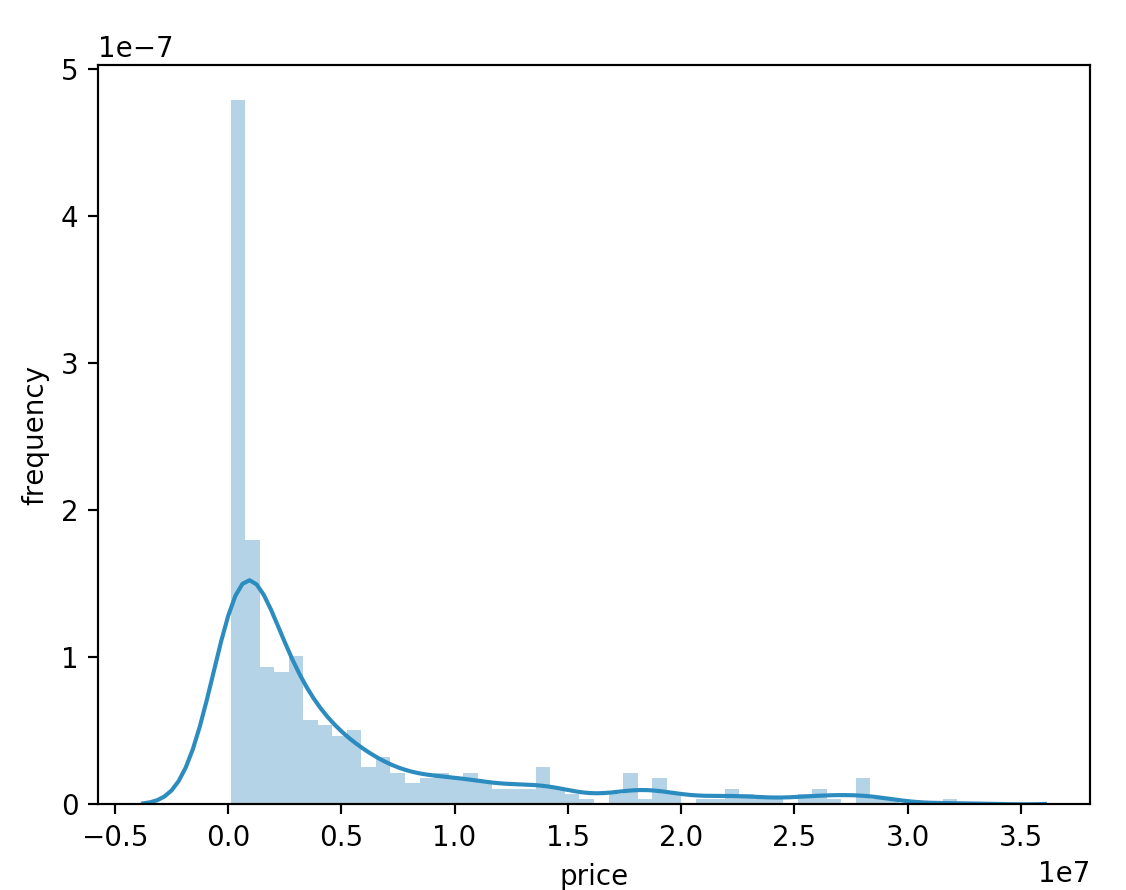
#reg plot  
plt.figure(2)  
scatter = sns.regplot(x='origin\_price',y='discount',data=df)

****

Biểu đồ phân bố giữa giá , discount và brand

Sử dụng scatter plot để thể hiện được biểu đồ trên

#scatter  
plt.figure(3)  
# df['brand\_color']=df['brand'].apply(recode)  
# plt.scatter(x=df['price'],y=df['discount'],c=df['brand\_color'],alpha=0.5)  
sns.scatterplot(x='price',y='discount',hue='brand',data=df)



Biểu đồ Density và Bar về giá và discountfr

Sử dụng distplot để hiện thị được mật độ về mức giá theo tần suất

#histogram plot  
plt.figure(1)  
plt.ylabel('frequency')  
hist=sns.distplot(df['price'],bins=50)

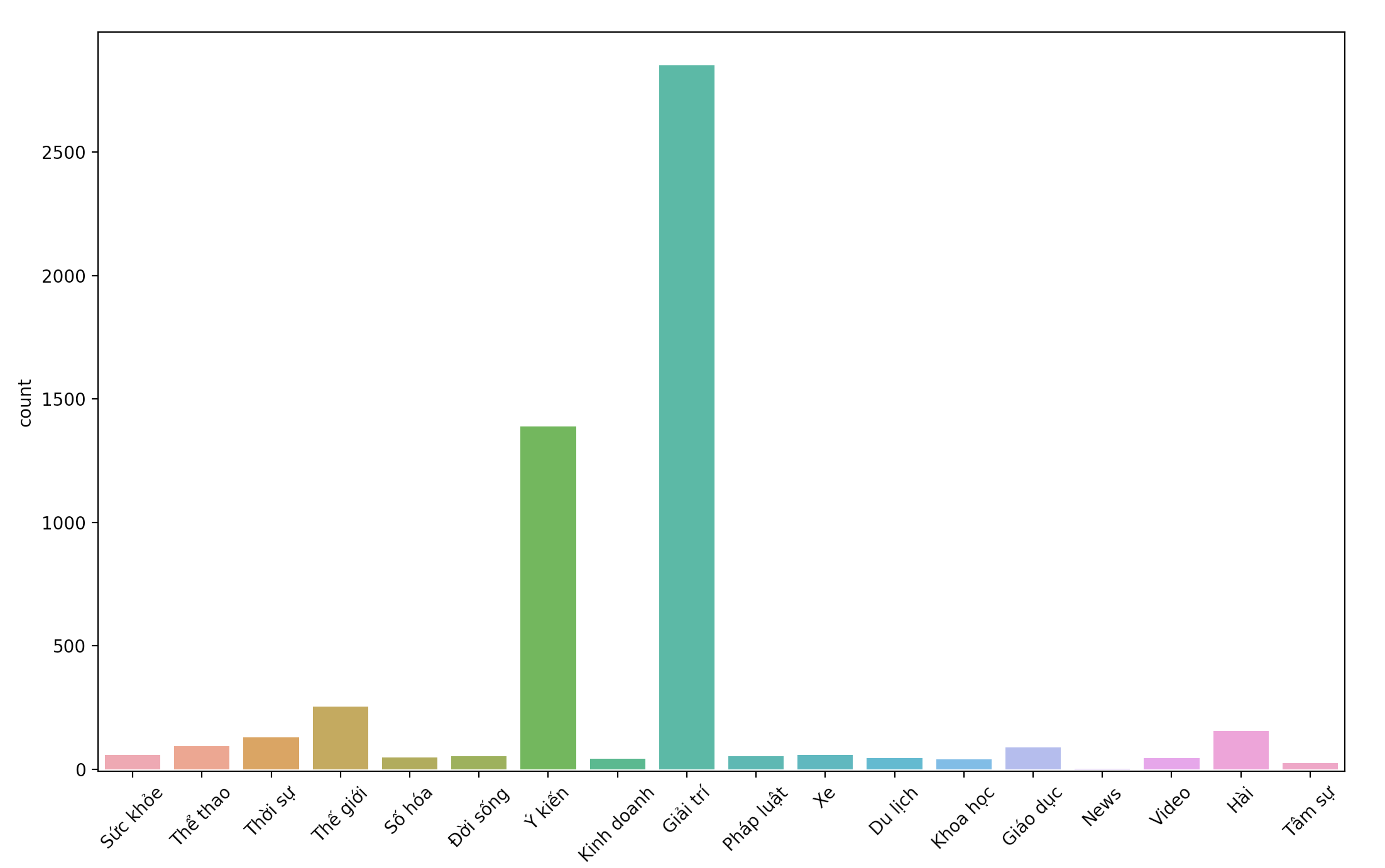
* Ta thấy hầu hết điện thoại được giảm giá nhiều và giá rẻ và bán nhiều nhất là những điện thoại giá rẻ

Sản phẩm với hạng giá cao thì sản phẩm sẽ được disocunt ít

**+ Trực quan hoá số liệu trên báo Vnexpress:**

Xử lý phần tags và category để gom được các nhóm bài phù hợp

#handle tags to have category (dealing with category='')  
df=df.replace({'':'Others'})  
c=df['category'].tolist()  
t=df['tags'].tolist()  
print(c)  
for i in range(len(t)):  
 if c[i]=='Others':  
 c[i]=handle(t[i])  
print(c)  
#df['category']=df.loc[(df.category=='Others'),'tags'].apply(handle)  
  
#aggregated category  
print(df['category'])  
df.loc[(df.category == 'Dịch viêm phổi virus corona' )| (df.category == 'Vaccine'),'category']='Thời sự'  
df.loc[df.category == '\nThượng đỉnh Mỹ - Triều\n','category']='Thế giới'  
df.loc[df.category == 'Kinh tế cho tương lai','category']='Kinh doanh'  
print(df['category'].value\_counts())  
print(df[df['category']=='Dịch viêm phổi virus corona']['category'])



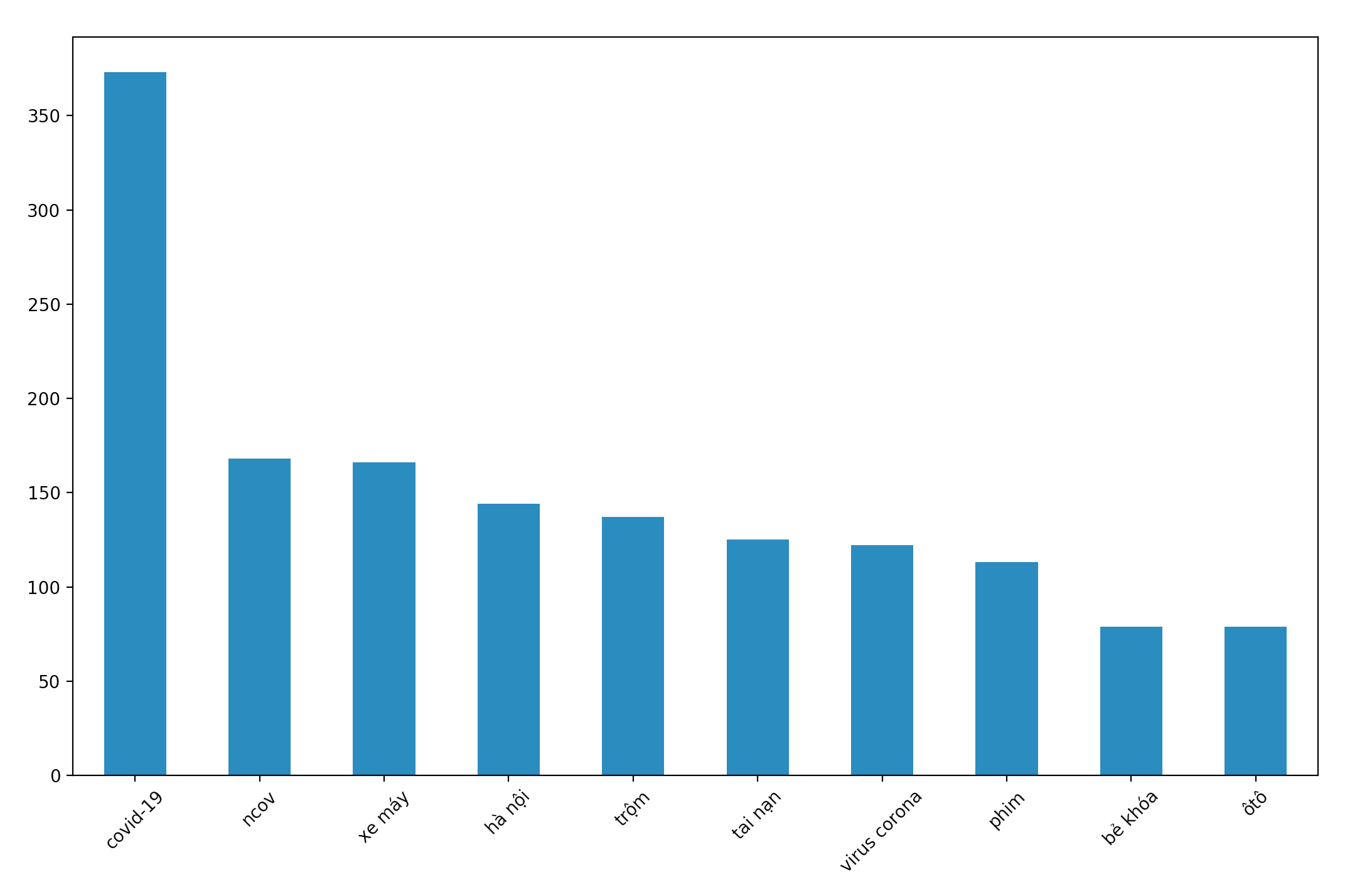
Biểu đồ số bài báo theo chuyên mục

Sử dụng countplot với seaborn để hiện thị biểu đồ trên

#countplot  
plt.figure(1)  
plt.xticks(rotation=45)  
plt.ylabel('NumsOfArticles')  
count=sns.countplot('category',data=df)

Xử lý phần tags

#handle tags  
li\_tags=list()  
for tag in df["tags"]:  
 if '' in tag:  
 tag.remove('')  
 li\_tags+=tag  
Tag=pd.Series(li\_tags).value\_counts().iloc[0:10]



top 10 tags phổ biến trên Vnexpress

Biểu thị qua bar plot

#bar plot  
plt.figure(2)  
Tag.plot.bar(x='tags',y='tags\_num',rot=45)

* Vậy ta thấy những bài báo được viết trên Vnexpress nhiều nhất là chủ đề Giải trí và tags đang phổ biến nhất là covid19

**+ Chương 2,3 Pandas for everyone: đã hoàn thành lưu trong thư mục ipython>Untitled.ipynb**