```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

df = pd.read_csv(r"F:\M.B.A\mod2\int python\INT CA 2\blackfriday.csv")
df.head()
```

```
   User_ID Product_ID Gender   Age  Occupation City_Category  \
0  1000001  P00069042      F  0-17          10             A
1  1000001  P00248942      F  0-17          10             A
2  1000001  P00087842      F  0-17          10             A
3  1000001  P00085442      F  0-17          10             A
4  1000002  P00285442      M   55+          16             C

  Stay_In_Current_City_Years  Marital_Status  Product_Category_1  \
0                          2               0                   3
1                          2               0                   1
2                          2               0                  12
3                          2               0                  12
4                         4+               0                   8

   Product_Category_2  Product_Category_3  Purchase
0                 NaN                 NaN      8370
1                 6.0                14.0     15200
2                 NaN                 NaN      1422
3                14.0                 NaN      1057
4                 NaN                 NaN      7969
```

```
df.describe()
```

```
            User_ID     Occupation  Marital_Status  Product_Category_1
\
count  5.500680e+05  550068.000000   550068.000000       550068.000000

mean   1.003029e+06       8.076707        0.409653            5.404270

std    1.727592e+03       6.522660        0.491770            3.936211

min    1.000001e+06       0.000000        0.000000            1.000000

25%    1.001516e+06       2.000000        0.000000            1.000000

50%    1.003077e+06       7.000000        0.000000            5.000000

75%    1.004478e+06      14.000000        1.000000            8.000000

max    1.006040e+06      20.000000        1.000000           20.000000
```

```
       Product_Category_2  Product_Category_3       Purchase
count       376430.000000       166821.000000  550068.000000
mean             9.842329           12.668243    9263.968713
std              5.086590            4.125338    5023.065394
min              2.000000            3.000000      12.000000
25%              5.000000            9.000000    5823.000000
50%              9.000000           14.000000    8047.000000
75%             15.000000           16.000000   12054.000000
max             18.000000           18.000000   23961.000000
```

```
df.isnull().sum()
```

```
User_ID                           0
Product_ID                        0
Gender                            0
Age                               0
Occupation                        0
City_Category                     0
Stay_In_Current_City_Years        0
Marital_Status                    0
Product_Category_1                0
Product_Category_2           173638
Product_Category_3           383247
Purchase                          0
dtype: int64
```

```
df.isnull().mean()
```

```
User_ID                      0.000000
Product_ID                   0.000000
Gender                       0.000000
Age                          0.000000
Occupation                   0.000000
City_Category                0.000000
Stay_In_Current_City_Years   0.000000
Marital_Status               0.000000
Product_Category_1           0.000000
Product_Category_2           0.315666
Product_Category_3           0.696727
Purchase                     0.000000
dtype: float64
```

```
df.nunique()
```

```
User_ID                      5891
Product_ID                   3631
Gender                          2
Age                             7
Occupation                     21
City_Category                   3
```

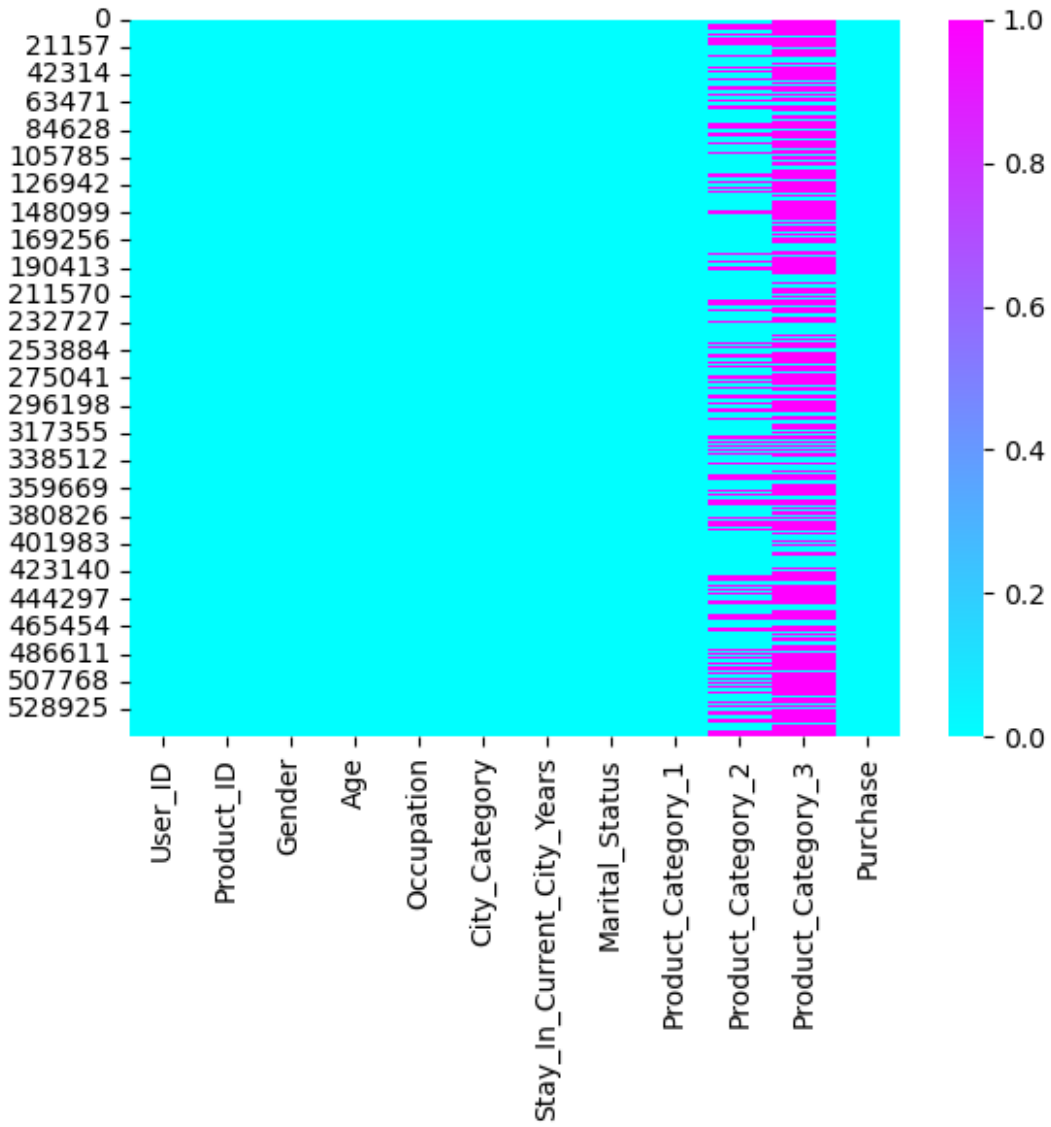
```
Stay_In_Current_City_Years        5
Marital_Status                    2
Product_Category_1               20
Product_Category_2               17
Product_Category_3               15
Purchase                      18105
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 12 columns):
 #   Column                      Non-Null Count   Dtype  
---  ------                      --------------   -----  
 0   User_ID                     550068 non-null  int64  
 1   Product_ID                  550068 non-null  object 
 2   Gender                      550068 non-null  object 
 3   Age                         550068 non-null  object 
 4   Occupation                  550068 non-null  int64  
 5   City_Category               550068 non-null  object 
 6   Stay_In_Current_City_Years  550068 non-null  object 
 7   Marital_Status              550068 non-null  int64  
 8   Product_Category_1          550068 non-null  int64  
 9   Product_Category_2          376430 non-null  float64
 10  Product_Category_3          166821 non-null  float64
 11  Purchase                    550068 non-null  int64  
dtypes: float64(2), int64(5), object(5)
memory usage: 50.4+ MB
```

```
sns.heatmap(df.isnull(),cmap="cool")
```

```
<Axes: >
```




```
df["Product_Category_2"].fillna(0, inplace=True)

df.head()
```

```
   User_ID Product_ID Gender   Age  Occupation City_Category  \
0  1000001  P00069042      F  0-17          10             A
1  1000001  P00248942      F  0-17          10             A
2  1000001  P00087842      F  0-17          10             A
3  1000001  P00085442      F  0-17          10             A
4  1000002  P00285442      M   55+          16             C

  Stay_In_Current_City_Years  Marital_Status  Product_Category_1  \
0                          2               0                   3
1                          2               0                   1
2                          2               0                  12
```

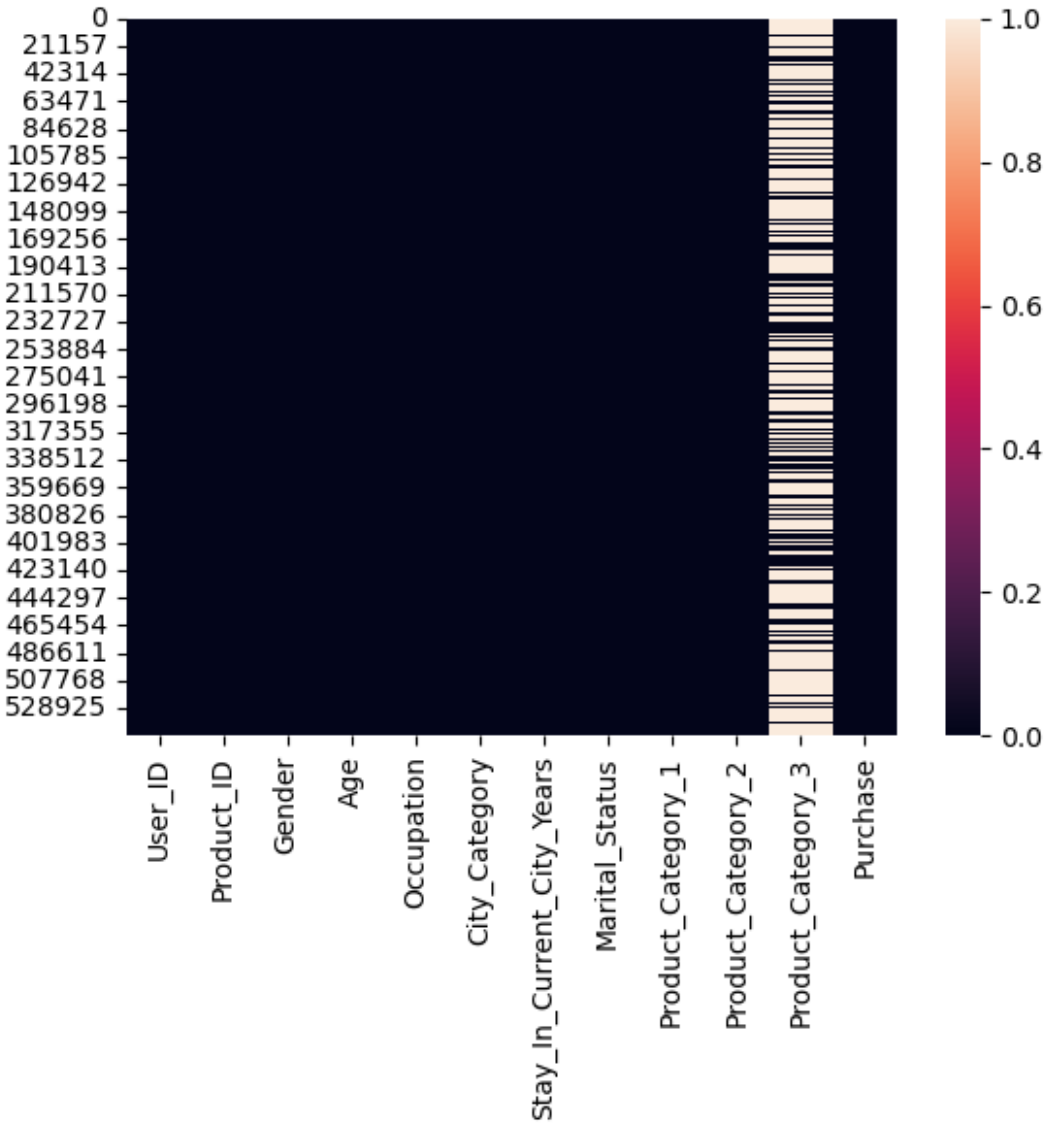
```
3                             2                  0                   12
4                            4+                  0                    8
```

```
   Product_Category_2  Product_Category_3  Purchase
0                 0.0                 NaN      8370
1                 6.0                14.0     15200
2                 0.0                 NaN      1422
3                14.0                 NaN      1057
4                 0.0                 NaN      7969
```

```
df['Product_Category_2'] = df['Product_Category_2'].astype(int)
```

```
sns.heatmap(df.isnull())
```

```
<Axes: >
```



```
df["Product_Category_3"].fillna(-1, inplace=True)

df.head(5)
```

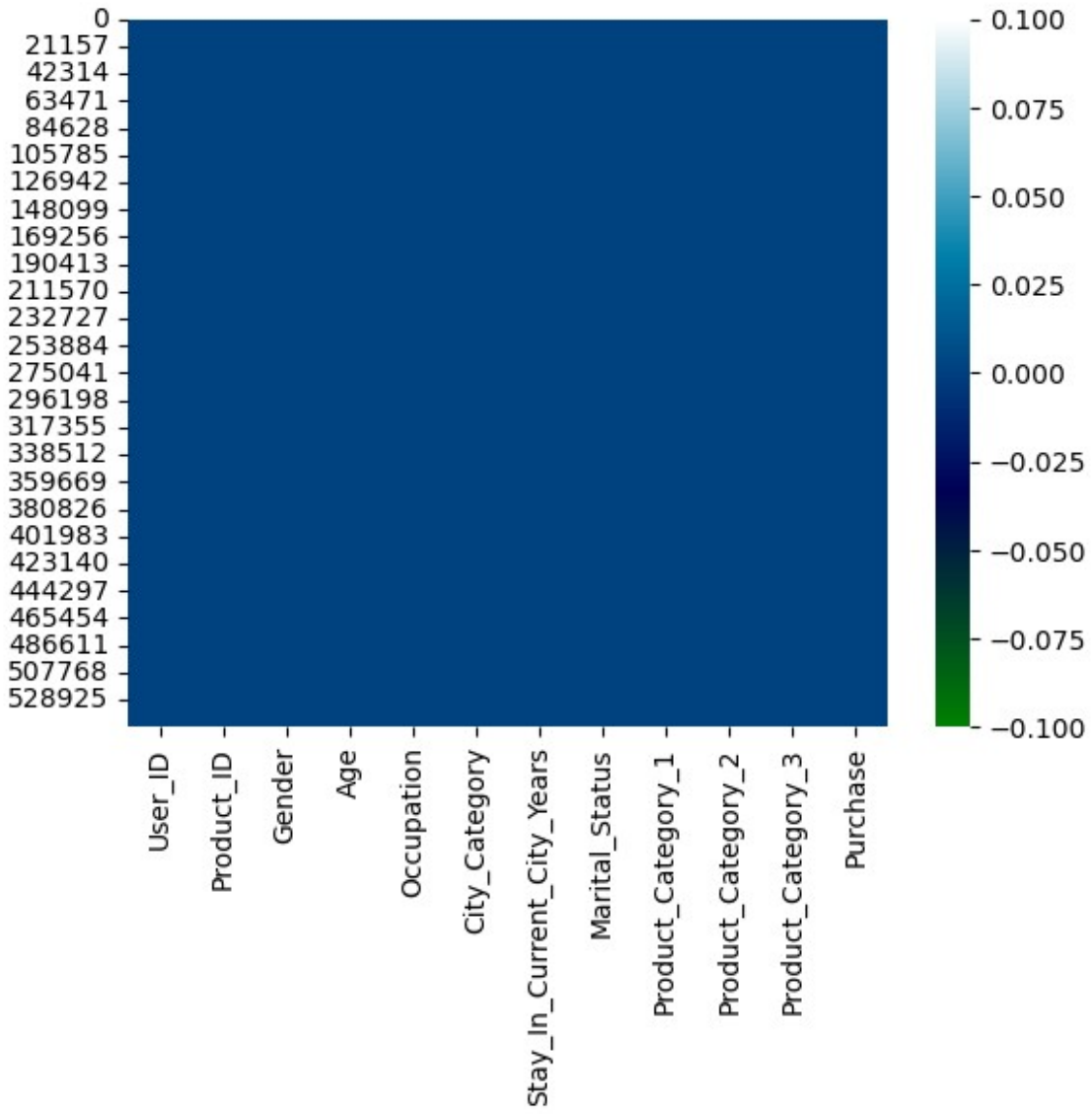
```
   User_ID Product_ID Gender   Age  Occupation City_Category  \
0  1000001  P00069042      F  0-17          10             A
1  1000001  P00248942      F  0-17          10             A
2  1000001  P00087842      F  0-17          10             A
3  1000001  P00085442      F  0-17          10             A
4  1000002  P00285442      M   55+          16             C

  Stay_In_Current_City_Years  Marital_Status  Product_Category_1  \
0                          2               0                   3
1                          2               0                   1
2                          2               0                  12
3                          2               0                  12
4                         4+               0                   8

   Product_Category_2  Product_Category_3  Purchase
0                   0                -1.0      8370
1                   6                14.0     15200
2                   0                -1.0      1422
3                  14                -1.0      1057
4                   0                -1.0      7969
```

```
sns.heatmap(df.isnull(), cmap="ocean")
```

```
<Axes: >
```




```
df['Product_Category_3'] = df['Product_Category_3'].astype(int)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 12 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     550068 non-null  int64
 1   Product_ID                  550068 non-null  object
 2   Gender                      550068 non-null  object
 3   Age                         550068 non-null  object
 4   Occupation                  550068 non-null  int64
 5   City_Category               550068 non-null  object
```

```
 6   Stay_In_Current_City_Years  550068 non-null  object
 7   Marital_Status              550068 non-null  int64
 8   Product_Category_1          550068 non-null  int64
 9   Product_Category_2          550068 non-null  int32
 10  Product_Category_3          550068 non-null  int32
 11  Purchase                    550068 non-null  int64
dtypes: int32(2), int64(5), object(5)
memory usage: 46.2+ MB
```

```
df['Age'].replace(r'^0-17$', 'child', inplace=True, regex=True)
df['Age'].replace(r'^18-25$', 'teenage', inplace=True, regex=True)
df['Age'].replace(r'^26-35$', 'adult', inplace=True, regex=True)
df['Age'].replace(r'^36-45$', 'adult', inplace=True, regex=True)
df['Age'].replace(r'^46-50$', 'adult', inplace=True, regex=True)
df['Age'].replace(r'^51-55$', 'old', inplace=True, regex=True)
df['Age'].replace(r'^55\+$', 'old', inplace=True, regex=True)
df.head(10)
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	\
0	1000001	P00069042	F	child	10	A	
1	1000001	P00248942	F	child	10	A	
2	1000001	P00087842	F	child	10	A	
3	1000001	P00085442	F	child	10	A	
4	1000002	P00285442	M	old	16	C	
5	1000003	P00193542	M	adult	15	A	
6	1000004	P00184942	M	adult	7	B	
7	1000004	P00346142	M	adult	7	B	
8	1000004	P0097242	M	adult	7	B	
9	1000005	P00274942	M	adult	20	A	

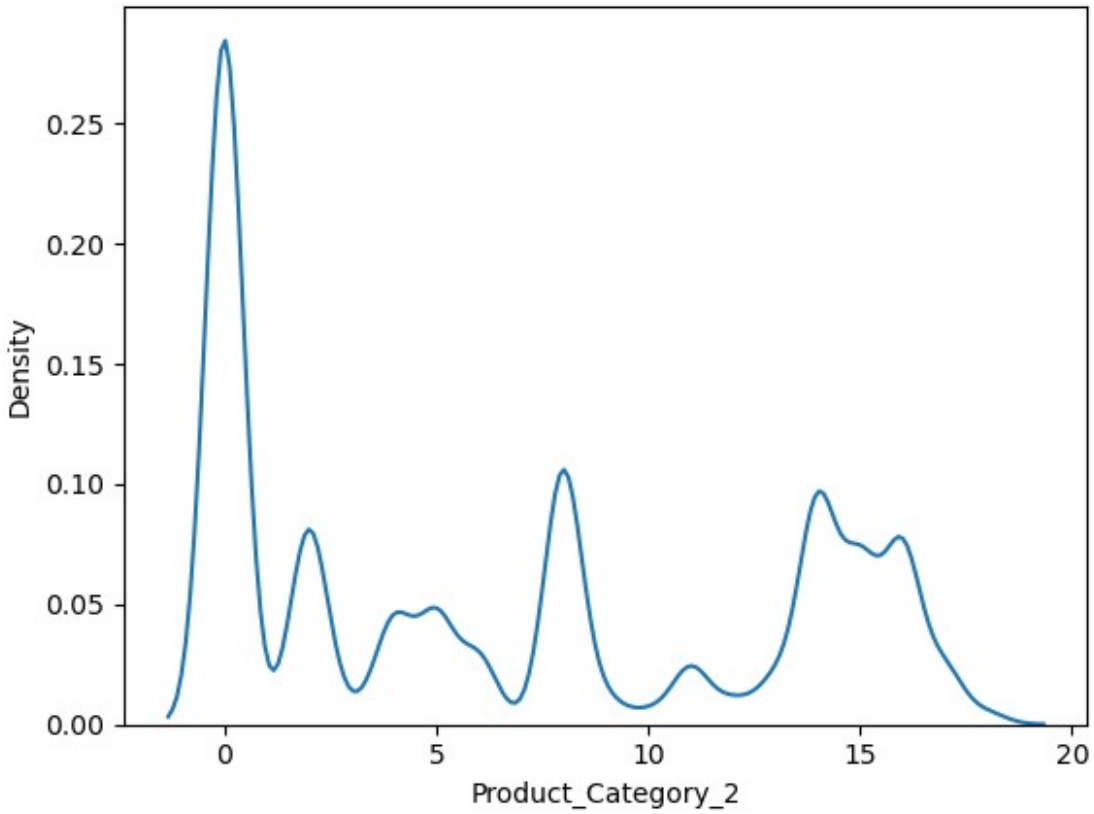
	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	\
0	2	0	3	
1	2	0	1	
2	2	0	12	
3	2	0	12	
4	4+	0	8	
5	3	0	1	
6	2	1	1	
7	2	1	1	
8	2	1	1	
9	1	1	8	

	Product_Category_2	Product_Category_3	Purchase
0	0	-1	8370
1	6	14	15200
2	0	-1	1422
3	14	-1	1057
4	0	-1	7969
5	2	-1	15227
6	8	17	19215

```
7                   15                   -1     15854
8                   16                   -1     15686
9                    0                   -1      7871
```

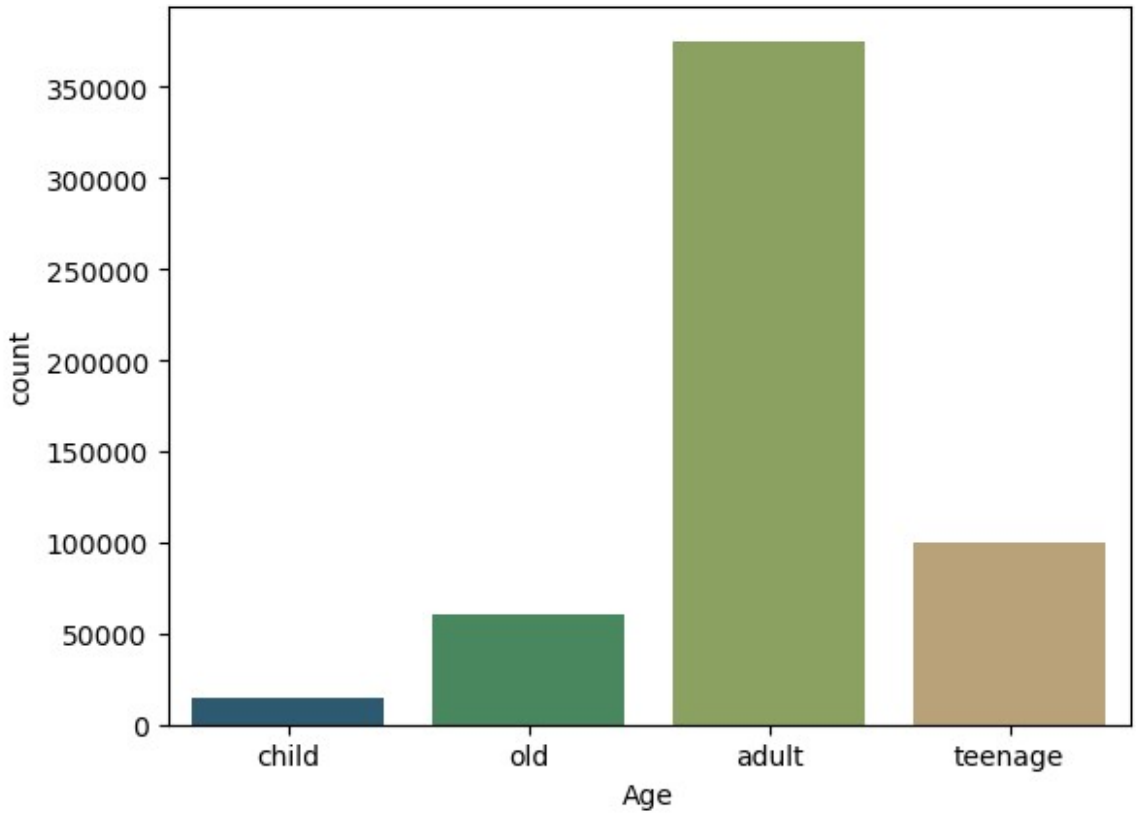
```
sns.kdeplot(df['Product_Category_2'])
```

```
<Axes: xlabel='Product_Category_2', ylabel='Density'>
```

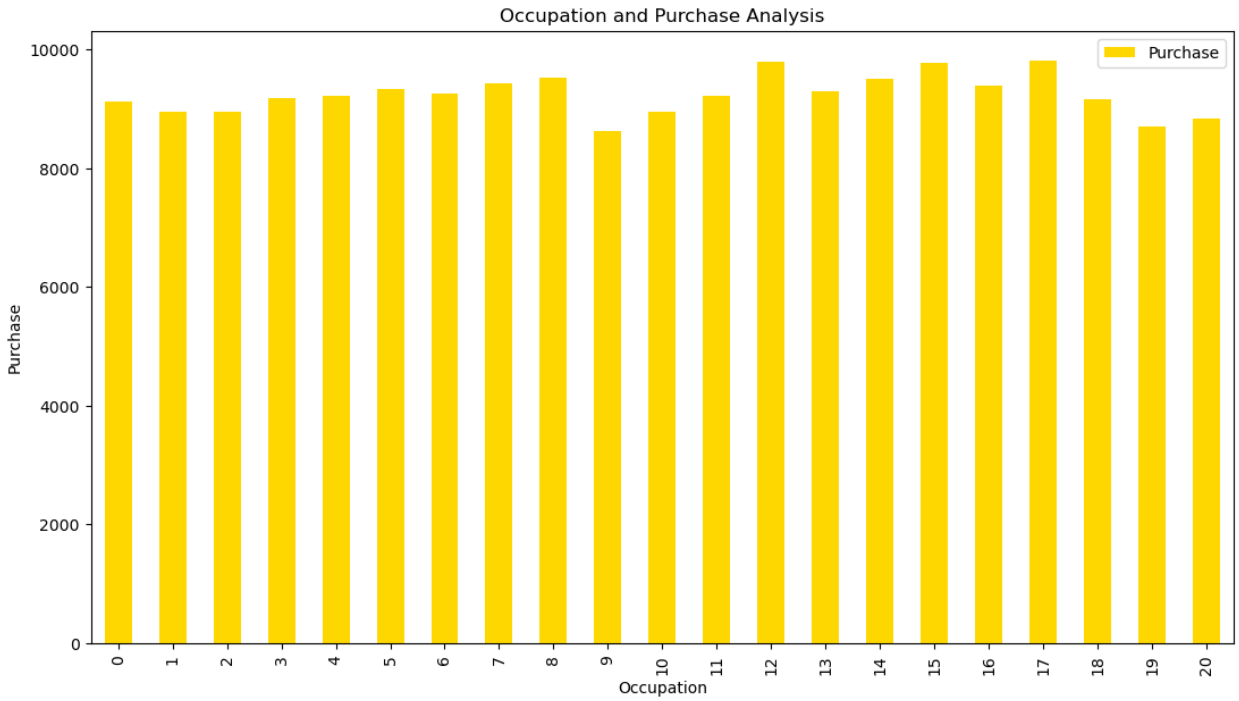



```
sns.countplot(x='Age',data=df,palette="gist_earth")
```

```
<Axes: xlabel='Age', ylabel='count'>
```

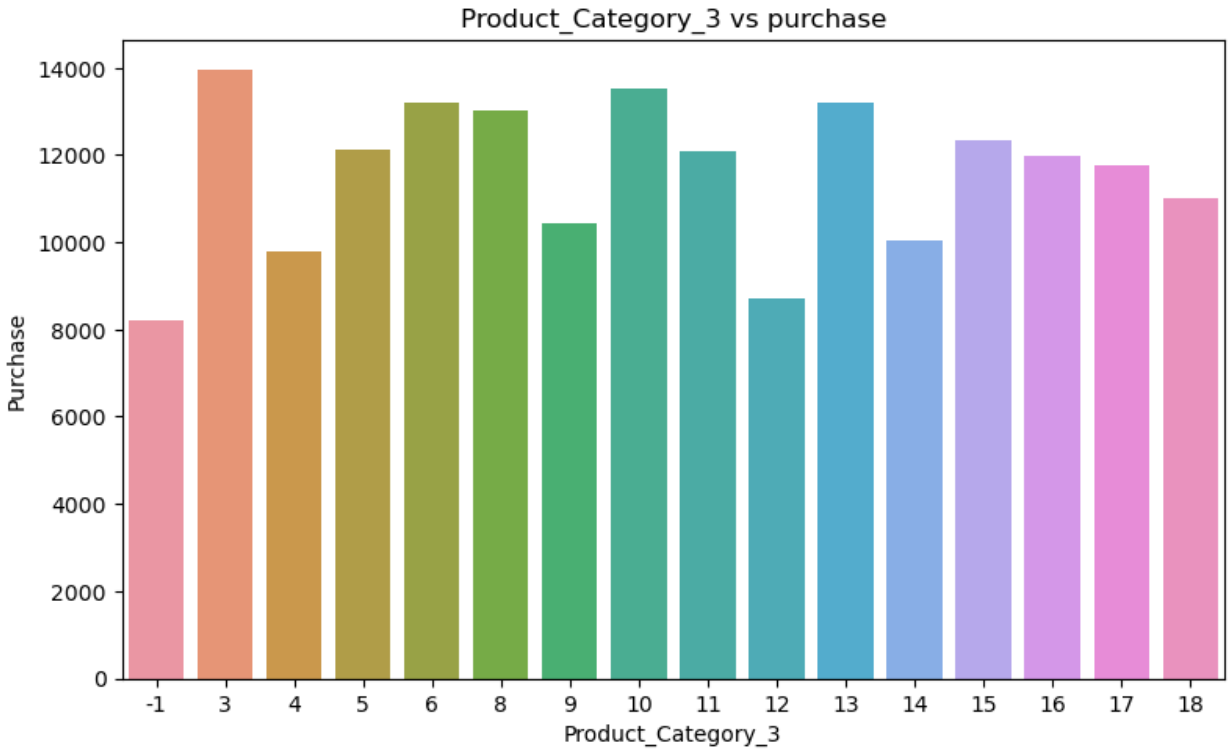



```
occupation_plot = df.pivot_table(index='Occupation',
values='Purchase')
occupation_plot.plot(kind='bar',figsize=(13, 7),color='gold')
plt.xlabel('Occupation')
plt.ylabel("Purchase")
plt.title("Occupation and Purchase Analysis")
plt.show()
```

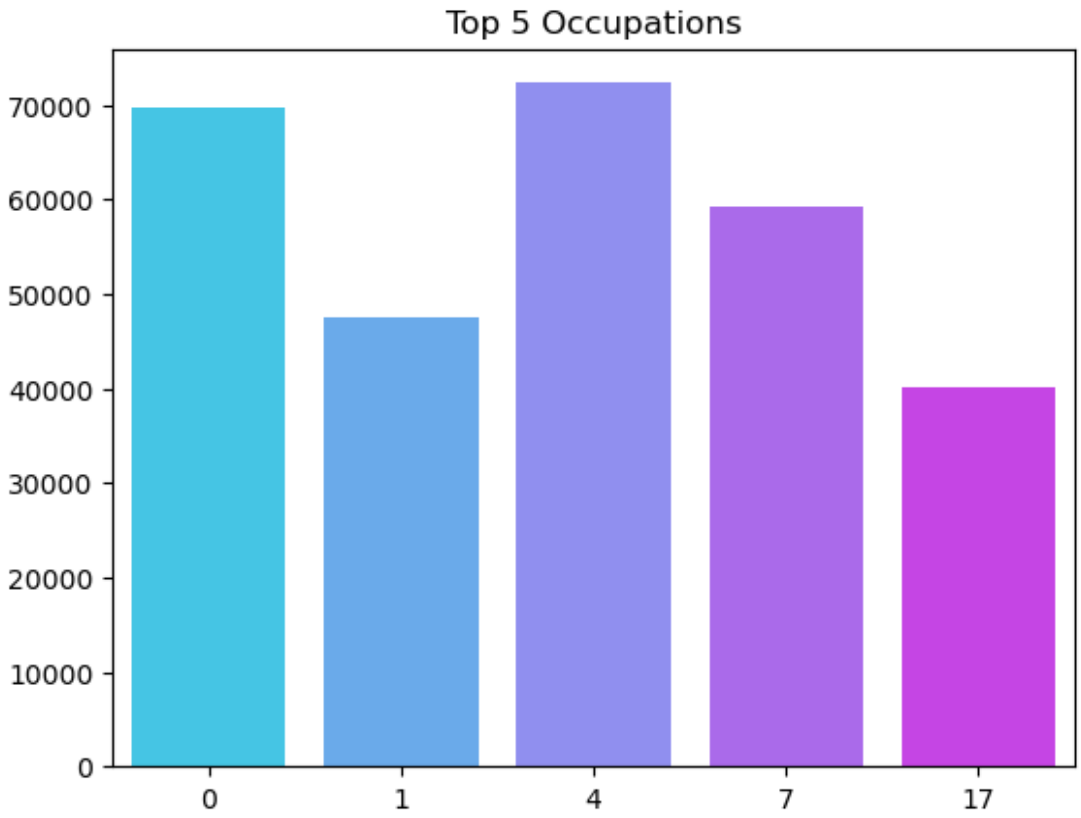



```
compare = df.groupby('Product_Category_3').agg({'Purchase':
'mean'}).reset_index()
plt.figure(figsize=(20,18),dpi=100)
plt.subplot(3,2,1)
plt.title('Product_Category_3 vs purchase')
sns.barplot(x='Product_Category_3',y='Purchase',data=compare,)
```

```
<Axes: title={'center': 'Product_Category_3 vs purchase'},
xlabel='Product_Category_3', ylabel='Purchase'>
```

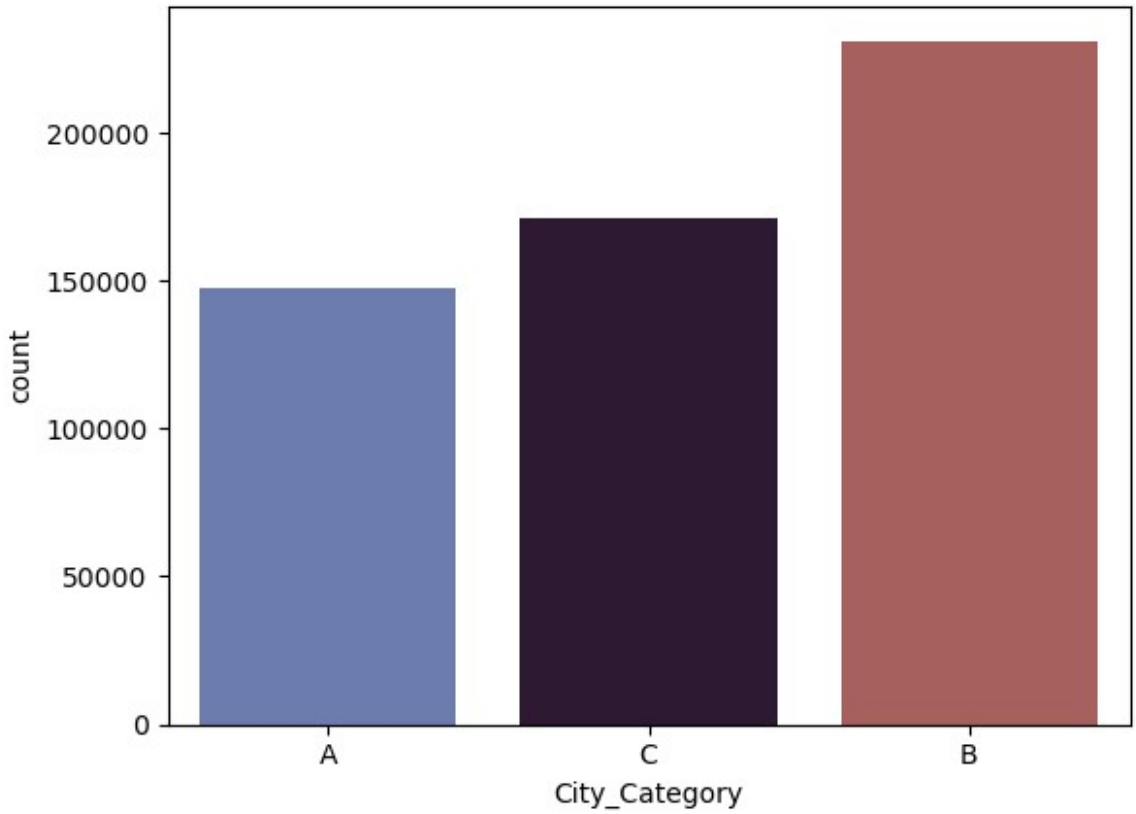



```
a=list(df.Occupation.value_counts().head().index)
b=list(df.Occupation.value_counts().head().values)
sns.barplot(x=a, y=b, palette='cool').set(title='Top 5 Occupations')
plt.show()
```




```
sns.countplot(x='City_Category',data=df,palette=("twilight"))
```

```
<Axes: xlabel='City_Category', ylabel='count'>
```




```
compare = df.groupby('City_Category').agg({'Purchase':
'mean'}).reset_index()
plt.title('City_Category vs purchase')
sns.barplot(x='City_Category',y='Purchase',data=compare,palette="color
blind")
```

```
<Axes: title={'center': 'City_Category vs purchase'},
xlabel='City_Category', ylabel='Purchase'>
```


City_Category vs purchase
10000
8000
6000
4000
2000
0
Purchase
A
B
C
City_Category