```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings

warnings.filterwarnings("ignore")

%matplotlib inline



#read the dataset
pd.read_csv("https://raw.githubusercontent.com/sunnysavita10/Statistics-With-Python-TheCompleteGuide/main/clead_google_playstore_data")


#check the dupliocate
df[df.duplicated()]


df=df.drop_duplicates(subset=['App'],keep='first')

#segrate the feaature
numeric_features = [feature for feature in df.columns if df[feature].dtype != 'O']
categorical_features = [feature for feature in df.columns if df[feature].dtype == 'O']


#value counts
df["Type"].value_counts(normalize=True)*100

for col in categorical_features:
    print(f"{col}:{df[col].value_counts(normalize=True)*100}")
    print("============================================")
```

```
#category = [ 'Type', 'Content Rating'] wise count plot

# categorical columns
plt.figure(figsize=(20, 15))
plt.suptitle('Univariate Analysis of Categorical Features', fontsize=20, fontweight='bold',
alpha=0.8, y=1.)
category = [ 'Type', 'Content Rating']
for i in range(0, len(category)):
    plt.subplot(2, 2, i+1)
    sns.countplot(x=df[category[i]],palette="Set2")
    plt.xlabel(category[i])
    plt.xticks(rotation=45)
    plt.tight_layout()


plt.figure(figsize=(15, 15))
plt.suptitle('Univariate Analysis of Numerical Features', fontsize=20, fontweight='bold',
alpha=0.8, y=1.)

for i in range(0, len(numeric_features)):
    plt.subplot(5, 3, i+1)
    sns.kdeplot(x=df[numeric_features[i]],shade=True, color='r')
    plt.xlabel(numeric_features[i])
    plt.tight_layout()


df["Category"].value_counts().plot.pie(y=df["Category"],figsize=(15,15),autopct='%1.1f
%%')

category = pd.DataFrame(df['Category'].value_counts())       #Dataframe of apps on
the basis of category
category.rename(columns = {'Category':'Count'},inplace=True)


plt.figure(figsize=(15,6))
sns.barplot(x=df_cat, y ='Count',data = category[:10],palette='hls')
plt.title('Top 10 App categories')
plt.xticks(rotation=90)
plt.show()


df_cat_installs=df.groupby(['Category'])['Installs'].sum().sort_values(ascending=False
).reset_index()
```

```
plt.figure(figsize = (14,10))
sns.set_context("talk")
sns.set_style("darkgrid")

ax = sns.barplot(x = 'Installs' , y = 'Category' , data = df2 )
ax.set_xlabel('No. of Installations in Billions')
ax.set_ylabel('')
ax.set_title("Most Popular Categories in Play Store", size = 20)




 ### What are the Top 5 most installed Apps in Each popular Categories ??
dfa=df.groupby(['Category','App'])['Installs'].sum().reset_index()

apps = ['GAME', 'COMMUNICATION', 'PRODUCTIVITY', 'SOCIAL' ]
sns.set_context("poster")
sns.set_style("darkgrid")

plt.figure(figsize=(40,30))

for i,app in enumerate(apps):
    df2 = dfa[dfa.Category == app]
    df3 = df2.head(5)
    plt.subplot(4,2,i+1)
    sns.barplot(data= df3,x= 'Installs' ,y='App' )
    plt.xlabel('Installation in Millions')
    plt.ylabel('')
    plt.title(app,size = 20)

plt.tight_layout()
plt.subplots_adjust(hspace= .3)
plt.show()


rating=df.groupby(["App"])["Rating"].sum().sort_values(ascending=False).reset_index
()

rating[rating.Rating==5.0]

rating2 = df.groupby(['Category','Installs',
'App'])['Rating'].sum().sort_values(ascending = False).reset_index()

df.groupby("Category").agg({"Installs":"sum","Reviews":"sum"}).reset_index()

df.groupby(['Category','App'])["Reviews"].sum().reset_index()
df.groupby("Category").agg({"Installs":"sum","Reviews":"sum"}).reset_index()

fig, ax = plt.subplots(1,2,figsize=(20,7))
df.value_counts('Type').plot.pie(y='Type',startangle=90, explode=(0.2,0),
title='Percentage of the Free App and Paid App', legend=False, autopct='%.2f',
ax=ax[0])
ax[0].set(ylabel='Type of Apps')
df.groupby('Type').agg({'Installs':sum}).plot.pie(y='Installs', startangle=90,
explode=(0.2,0), title='Percentage of Installs Number for Free App and Paid App',
legend=False, autopct='%.2f', ax=ax[1])


plt.figure(figsize=(10,10))
sns.boxplot(x="Installs",y="Rating",data=df)
plt.xticks(size=15,rotation=90)
plt.show()


plt.figure(figsize=(10,10))
sns.boxplot(x="Category",y="Rating",data=df)
plt.xticks(size=15,rotation=90)
plt.show()

plt.figure(figsize=(15,15))
plt.xticks(rotation = 90)
sns.barplot('Category','Size',data=df)
plt.show()


plt.figure(figsize=(15,8))
sns.countplot(x='Rating',data = df,palette="Set1_r")
plt.xticks(rotation =90)
plt.title('Countplot for ratings')
plt.show()


plt.subplots(figsize=(20,10))
freq= pd.Series()
freq=df['year'].value_counts()
freq.plot()
plt.xlabel("Dates")
plt.ylabel("Number of updates")
plt.title("Time series plot of Last Updates")
```

https://drive.google.com/drive/folders/1B-LeKjHsew1C6GUciWrr5y5XLIv5X_OG

Dataset:
https://archive.ics.uci.edu/ml/datasets/Phishing+Websites
https://archive.ics.uci.edu/ml/datasets/Census+Income

Do the proper EDA and cleaning it is mandatory
Feature engineering is optional

Submit form:
https://forms.gle/7JoFujsHhNzc6sW18

Before next saturday you need to submit the task