DATA ANALYST

DATASET: GOOGLE PLAY STORE

PREPARED BY: JOE MANDE

TOOL USED: PYTHON FOR DATA ANALYSIS

TABLE OF CONTENTS

1. IMPORT PYTHON LIBRARIES

Numpy Libraries

- For data consolidation.

Pandas Libraries

- To load the dataset.
- For data cleaning and data analysis.

Matplotlib Libraries

- For data visualization.

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

2. LOADING THE DATASET

- The dataset was loaded for each course.
- The dataset was read to check if it has headers.

In [2]:

```
data = pd.read_csv('/Users/joemande/Downloads/archive/googleplaystore.csv')
data.head(3)
```

Out[2]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play

2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design

3. DATA CLEANING

- To delete any blank cells

In [3]:
```
data = data.dropna()
```

- To remove any duplicates

In [4]:
```
data = data.drop_duplicates()
```

- To ensure that the data is consistent
  1. Check for data type
  2. Text Handling
  3. Change data type

In [5]:
```
data.dtypes
```

Out[5]:
```
App                object
Category           object
Rating            float64
Reviews            object
Size               object
Installs           object
Type               object
Price              object
Content Rating     object
Genres             object
Last Updated       object
Current Ver        object
Android Ver        object
dtype: object
```

In [6]:
```
data['Installs']=data['Installs'].str.replace(',','')
data['Installs']=data['Installs'].str.replace('+','')
data['Price']=data['Price'].str.replace('$','')
data.head(3)
```

```
/var/folders/0v/2ppkchyx7wl34m49kh987wj80000gp/T/ipykernel_15588/2299274380.py:2: Future
Warning: The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as literal strin
gs when regex=True.
  data['Installs']=data['Installs'].str.replace('+','')
/var/folders/0v/2ppkchyx7wl34m49kh987wj80000gp/T/ipykernel_15588/2299274380.py:3: Future
Warning: The default value of regex will change from True to False in a future version.
In addition, single character regular expressions will *not* be treated as literal strin
gs when regex=True.
  data['Price']=data['Price'].str.replace('$','')
```

Out[6]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
0	Photo	ART_AND_DESIGN	4.1	159	19M	10000	Free	0	Everyone	Art & Design

	Editor & Candy Camera & Grid & ScrapBook									
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500000	Free	0	Everyone	Art & Design;Pretend Play
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5000000	Free	0	Everyone	Art & Design

In [7]:
```
data['Installs']=data['Installs'].astype('int')
data['Price']=data['Price'].astype('float')
data.dtypes
```

Out[7]:
```
App                object
Category           object
Rating            float64
Reviews            object
Size               object
Installs            int64
Type               object
Price             float64
Content Rating     object
Genres             object
Last Updated       object
Current Ver        object
Android Ver        object
dtype: object
```

4. DATA ANALYSIS AND VISUALIZATION

- Total number of Installs for each genres
    1. Table
    2. Chart

In [8]:
```
first_analysis = data.groupby(['Category'])['Installs'].sum()
pd.DataFrame(first_analysis)
```

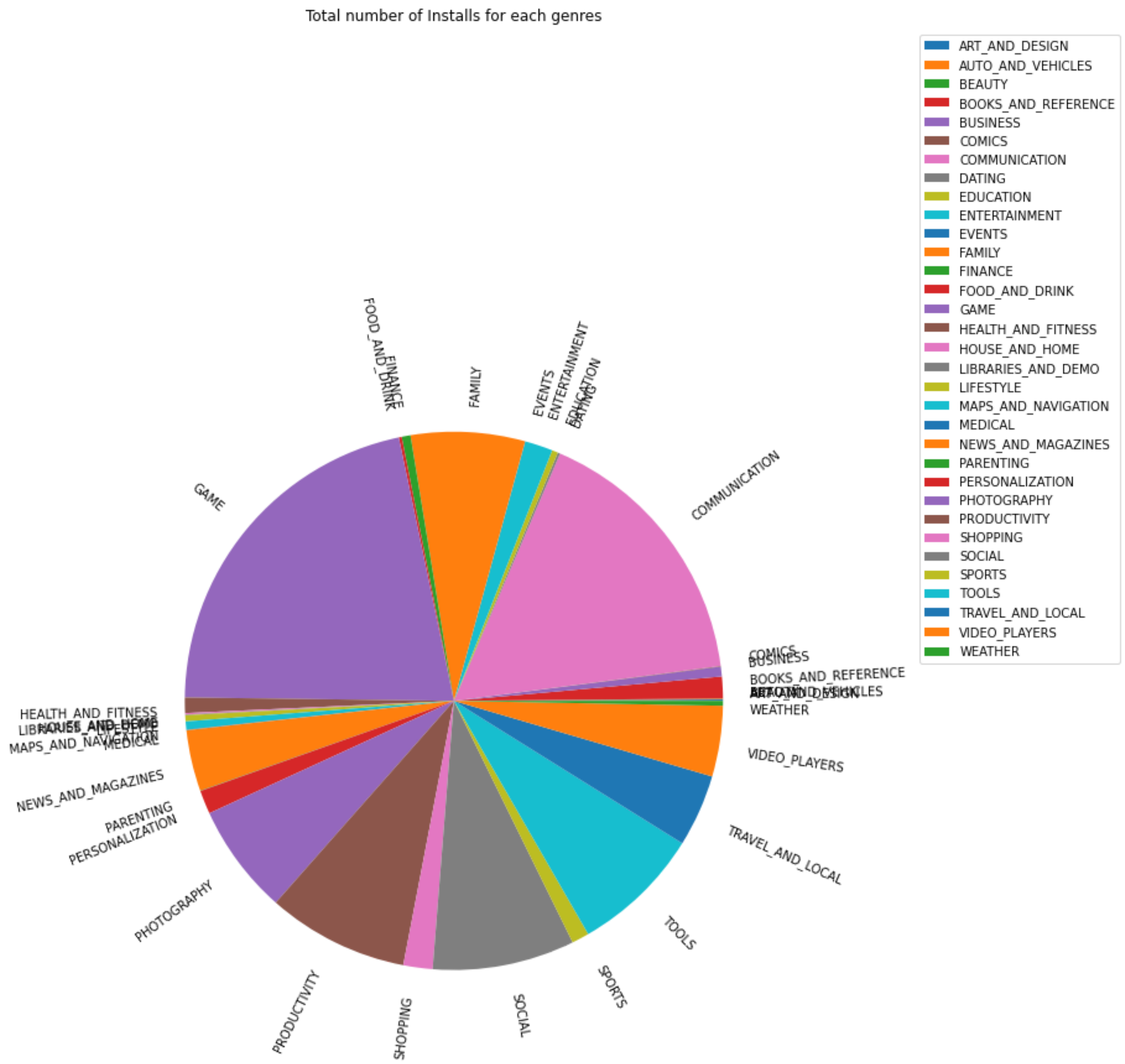
Out[8]:

	Installs
Category	
ART_AND_DESIGN	124228100
AUTO_AND_VEHICLES	53129800
BEAUTY	26916200
BOOKS_AND_REFERENCE	1916291655
BUSINESS	863518120
COMICS	56036100
COMMUNICATION	24152241530
DATING	206522410
EDUCATION	533852000

ENTERTAINMENT	2455660000
EVENTS	15949410
FAMILY	10041080590
FINANCE	770312400
FOOD_AND_DRINK	257777750
GAME	31543862717
HEALTH_AND_FITNESS	1361006220
HOUSE_AND_HOME	125082000
LIBRARIES_AND_DEMO	61083000
LIFESTYLE	534741120
MAPS_AND_NAVIGATION	724267560
MEDICAL	42162676
NEWS_AND_MAGAZINES	5393110650
PARENTING	31116110
PERSONALIZATION	2074341930
PHOTOGRAPHY	9721243130
PRODUCTIVITY	12463070180
SHOPPING	2573331540
SOCIAL	12513841475
SPORTS	1528531465
TOOLS	11450224500
TRAVEL_AND_LOCAL	6361859300
VIDEO_PLAYERS	6221897200
WEATHER	426096500

In [9]:

```
plt.subplots(figsize=(20,20))
plt.title('Total number of Installs for each genres')
plt.pie(first_analysis, radius=0.5, rotatelabels=60, labels=['ART_AND_DESIGN', 'AUTO_AND
                                                 'BOOKS_AND_REFERENCE','BUSINESS','COMIC
                                                 'DATING', 'EDUCATION', 'ENTERTAINMENT',
                                                 'FOOD_AND_DRINK','GAME','HEALTH_AND_FIT
                                                 'LIBRARIES_AND_DEMO','LIFESTYLE','MAPS_
                                                 'MEDICAL','NEWS_AND_MAGAZINES','PARENTI
                                                 'PHOTOGRAPHY','PRODUCTIVITY','SHOPPING'
                                                 'TRAVEL_AND_LOCAL','VIDEO_PLAYERS','WEA
plt.legend()
plt.show()
```



- Total number of App for each genres
    1. Table
    2. Chart

In [10]:
```
second_analysis = data.groupby(['Category'])['App'].count()
pd.DataFrame(second_analysis)
```

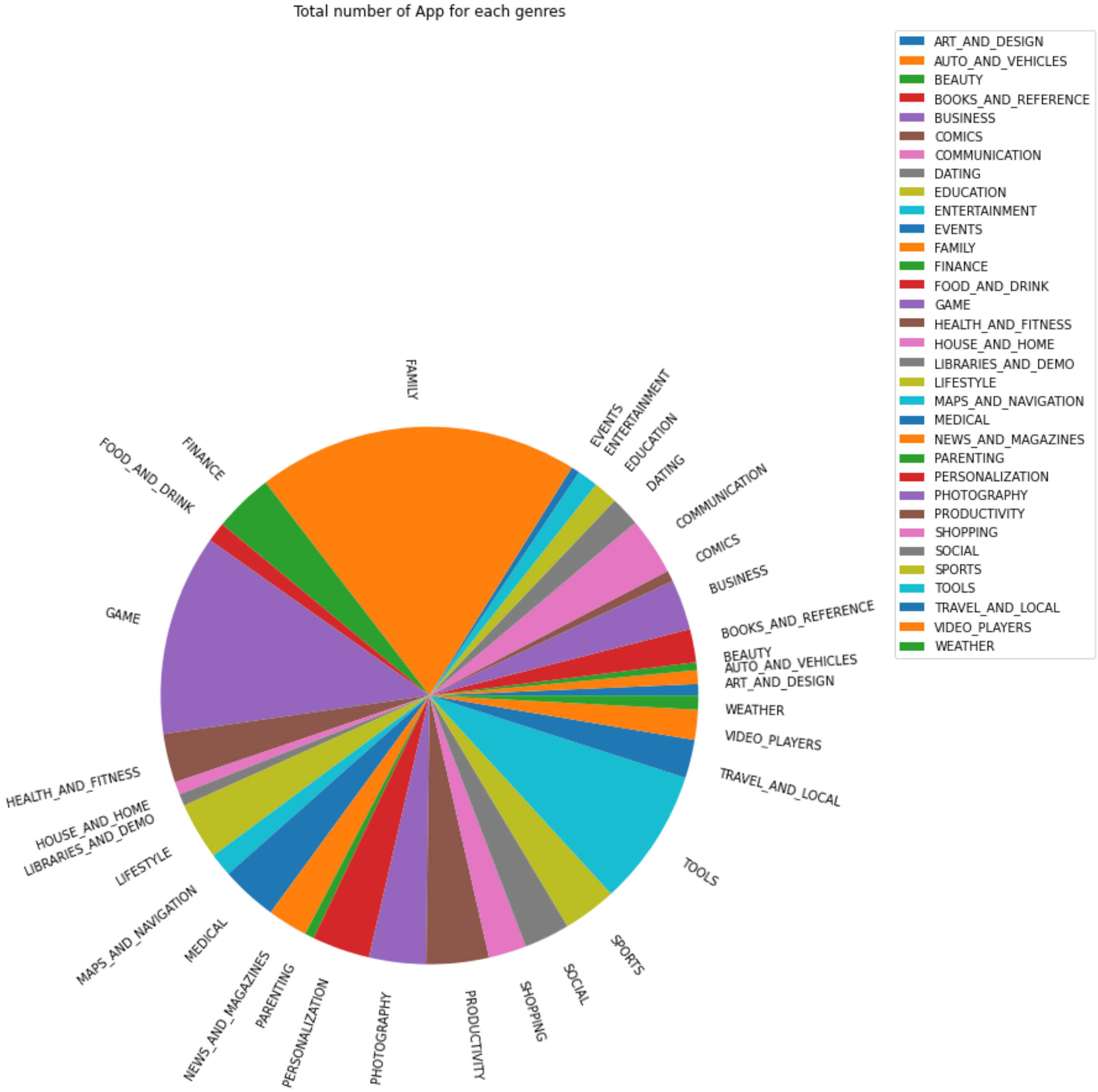
Out[10]:

	App
Category	
ART_AND_DESIGN	61
AUTO_AND_VEHICLES	73
BEAUTY	42
BOOKS_AND_REFERENCE	177
BUSINESS	270

COMICS	58
COMMUNICATION	307
DATING	159
EDUCATION	129
ENTERTAINMENT	111
EVENTS	45
FAMILY	1717
FINANCE	317
FOOD_AND_DRINK	106
GAME	1074
HEALTH_AND_FITNESS	262
HOUSE_AND_HOME	68
LIBRARIES_AND_DEMO	64
LIFESTYLE	305
MAPS_AND_NAVIGATION	124
MEDICAL	302
NEWS_AND_MAGAZINES	214
PARENTING	50
PERSONALIZATION	308
PHOTOGRAPHY	304
PRODUCTIVITY	334
SHOPPING	202
SOCIAL	244
SPORTS	286
TOOLS	733
TRAVEL_AND_LOCAL	205
VIDEO_PLAYERS	160
WEATHER	75

In [11]:

```
plt.subplots(figsize=(20,20))
plt.title('Total number of App for each genres')
plt.pie(second_analysis, radius=0.5, rotatelabels=60, labels=['ART_AND_DESIGN', 'AUTO_AN
                                                 'BOOKS_AND_REFERENCE','BUSINESS','COMIC
                                                 'DATING', 'EDUCATION', 'ENTERTAINMENT',
                                                 'FOOD_AND_DRINK','GAME','HEALTH_AND_FIT
                                                 'LIBRARIES_AND_DEMO','LIFESTYLE','MAPS_
                                                 'MEDICAL','NEWS_AND_MAGAZINES','PARENTI
                                                 'PHOTOGRAPHY','PRODUCTIVITY','SHOPPING'
                                                 'TRAVEL_AND_LOCAL','VIDEO_PLAYERS','WEA
plt.legend()
plt.show()
```



- Total number of App for each Content Rating
    1. Table
    2. Chart

```
In [12]: third_analysis = data.groupby(['Content Rating'])['App'].count()
         pd.DataFrame(third_analysis)
```

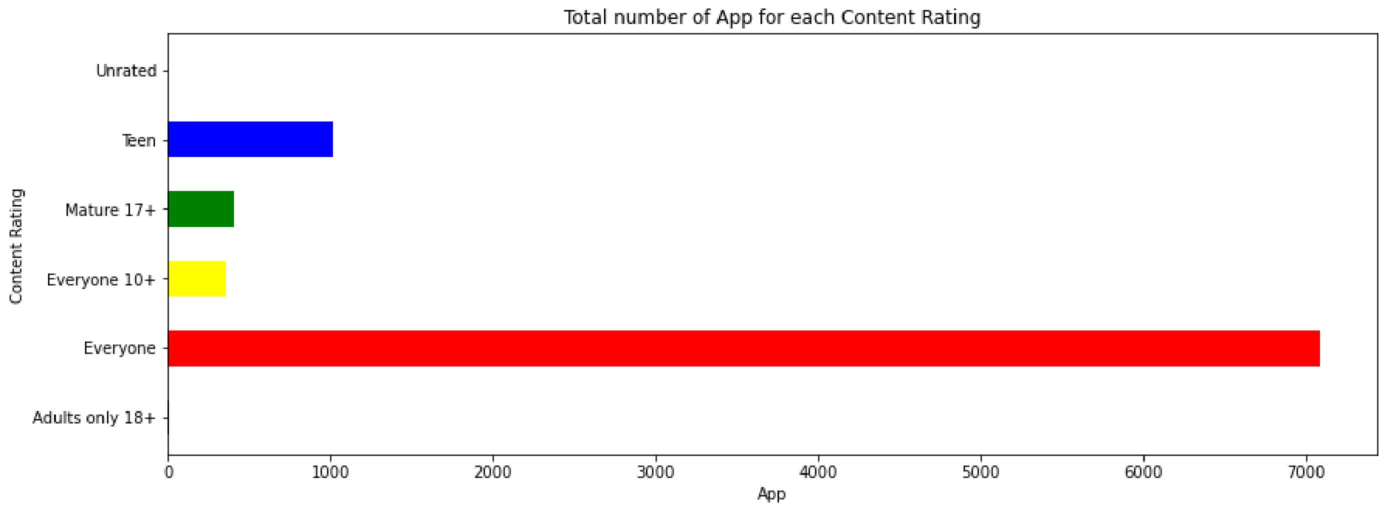
Out[12]:

	App
Content Rating	
Adults only 18+	3
Everyone	7089
Everyone 10+	360
Mature 17+	411
Teen	1022

Unrated	1

In [13]:

```
plt.subplots(figsize=(14,5))
plt.title('Total number of App for each Content Rating')
plt.barh(np.arange(len(third_analysis)), third_analysis, height=0.5, tick_label=['Adults
                                                                                 'Everyo
                                                                                 'Teen',
         color=['black', 'red', 'yellow', 'green', 'blue', 'violet'])
plt.ylabel('Content Rating')
plt.xlabel('App')
plt.show()
```



- Total number of App for each Type
    1. Table
    2. Chart

In [14]:

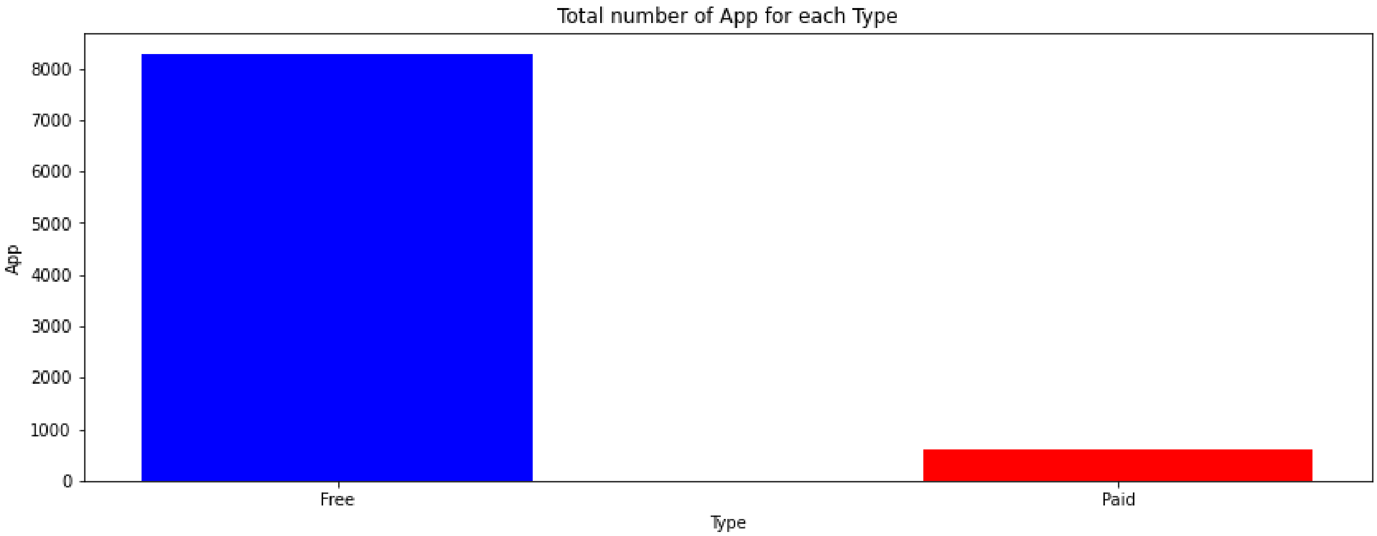
```
fourth_analysis = data.groupby(['Type'])['App'].count()
pd.DataFrame(fourth_analysis)
```

Out[14]:

	App
Type	
Free	8275
Paid	611

In [15]:

```
plt.subplots(figsize=(14,5))
plt.title('Total number of App for each Type')
plt.bar(range(len(fourth_analysis)), fourth_analysis, width=0.5, tick_label=['Free','Pai
plt.ylabel('App')
plt.xlabel('Type')
plt.show()
```



- Total average price for each Content Rating
  1. Table
  2. Chart

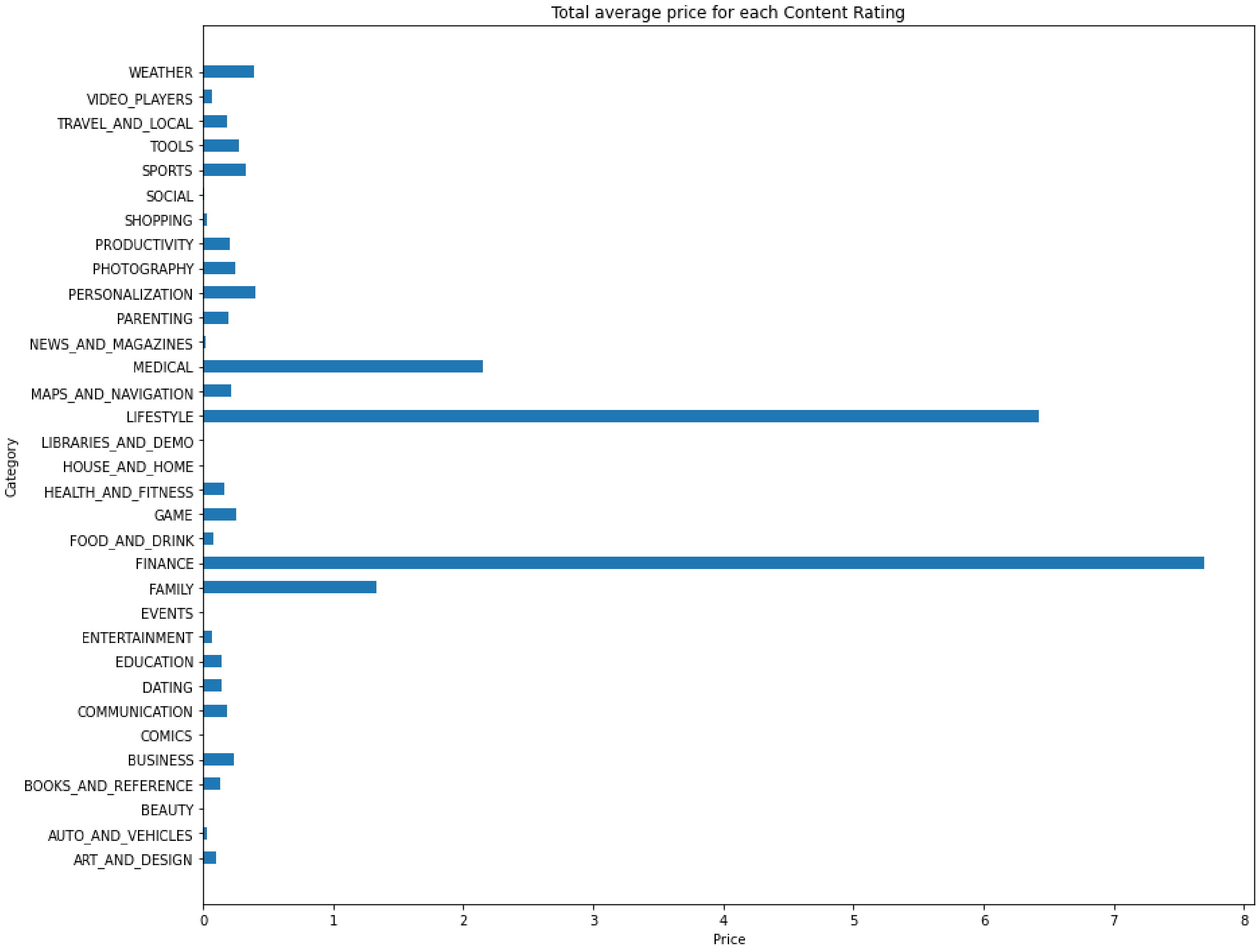
```
In [16]: fifth_analysis = data.groupby(['Category'])['Price'].mean().round(2)
         pd.DataFrame(fifth_analysis)
```

Out[16]:

	Price
Category	
ART_AND_DESIGN	0.10
AUTO_AND_VEHICLES	0.03
BEAUTY	0.00
BOOKS_AND_REFERENCE	0.13
BUSINESS	0.24
COMICS	0.00
COMMUNICATION	0.18
DATING	0.14
EDUCATION	0.14
ENTERTAINMENT	0.07
EVENTS	0.00
FAMILY	1.33
FINANCE	7.70
FOOD_AND_DRINK	0.08
GAME	0.26
HEALTH_AND_FITNESS	0.16
HOUSE_AND_HOME	0.00
LIBRARIES_AND_DEMO	0.00
LIFESTYLE	6.43
MAPS_AND_NAVIGATION	0.22
MEDICAL	2.15

NEWS_AND_MAGAZINES	0.02
PARENTING	0.19
PERSONALIZATION	0.40
PHOTOGRAPHY	0.25
PRODUCTIVITY	0.21
SHOPPING	0.03
SOCIAL	0.01
SPORTS	0.33
TOOLS	0.28
TRAVEL_AND_LOCAL	0.18
VIDEO_PLAYERS	0.07
WEATHER	0.39

```
In [17]: plt.subplots(figsize=(14,12))
         plt.title('Total average price for each Content Rating')
         plt.barh(np.arange(len(fifth_analysis)), fifth_analysis, height=0.5,
                  tick_label=['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
                                                    'BOOKS_AND_REFERENCE','BUSINESS','COMIC
                                                    'DATING', 'EDUCATION', 'ENTERTAINMENT',
                                                    'FOOD_AND_DRINK','GAME','HEALTH_AND_FIT
                                                    'LIBRARIES_AND_DEMO','LIFESTYLE','MAPS_
                                                    'MEDICAL','NEWS_AND_MAGAZINES','PARENTI
                                                    'PHOTOGRAPHY','PRODUCTIVITY','SHOPPING'
                                                    'TRAVEL_AND_LOCAL','VIDEO_PLAYERS','WEA
         plt.xlabel('Price')
         plt.ylabel('Category')
         plt.show()
```


Total average price for each Content Rating
Category
WEATHER
VIDEO_PLAYERS
TRAVEL_AND_LOCAL
TOOLS
SPORTS
SOCIAL
SHOPPING
PRODUCTIVITY
PHOTOGRAPHY
PERSONALIZATION
PARENTING
NEWS_AND_MAGAZINES
MEDICAL
MAPS_AND_NAVIGATION
LIFESTYLE
LIBRARIES_AND_DEMO
HOUSE_AND_HOME
HEALTH_AND_FITNESS
GAME
FOOD_AND_DRINK
FINANCE
FAMILY
EVENTS
ENTERTAINMENT
EDUCATION
DATING
COMMUNICATION
COMICS
BUSINESS
BOOKS_AND_REFERENCE
BEAUTY
AUTO_AND_VEHICLES
ART_AND_DESIGN
0
1
2
3
4
5
6
7
8
Price