Google Play Store EDA and Data Visualisation

Loading the dataset

1. Import necessary libraries

In [1]:

2. Read from googleplaystore.csv and display first five rows of data.

In [2]:

Out[2]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone

Understanding the dataset

3. Show all the columns' names.

In [3]:

```
```

Out[3]:

```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

4. Replace the space in the column names with an underscore.

In [4]:

```
```

Out[4]:

```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content_Rating', 'Genres', 'Last_Updated', 'Current_Ver',
       'Android_Ver'],
      dtype='object')
```

5. Let's look at the number of rows and columns in the dataset.

In [5]:

```
```

Out[5]:

```
(10841, 13)
```

6. Show the data types of all columns.

In [6]:

```
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content_Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last_Updated    10841 non-null  object 
 11  Current_Ver     10833 non-null  object 
 12  Android_Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

Handling missing data

7. Let's have a look at number of missing data in each column.

In [7]:

```
```

Out[7]:

```
App                  0
Category             0
Rating            1474
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
Content_Rating       1
Genres               0
Last_Updated         0
Current_Ver          8
Android_Ver          3
dtype: int64
```

8. Impute the missing data in the rating column using median. Check the number of missing data in each column again, to confirm the missing data has been imputed.

In [8]:

```
```

Out[8]:

```
App               0
Category          0
Rating            0
Reviews           0
Size              0
Installs          0
Type              1
Price             0
Content_Rating    1
Genres            0
Last_Updated      0
Current_Ver       8
Android_Ver       3
dtype: int64
```

9. Let's remove the other missing data as it is very little.

In [9]:

```
```

Out[9]:
```
App               0
Category          0
Rating            0
Reviews           0
Size              0
Installs          0
Type              0
Price             0
Content_Rating    0
Genres            0
Last_Updated      0
Current_Ver       0
Android_Ver       0
dtype: int64
```

Data preprocessing

10. Shows first 10 values in genres.

In [10]:

Out[10]:
```
0                  Art & Design
1    Art & Design;Pretend Play
2                  Art & Design
3                  Art & Design
4      Art & Design;Creativity
5                  Art & Design
6                  Art & Design
7                  Art & Design
8                  Art & Design
9      Art & Design;Creativity
Name: Genres, dtype: object
```

11. Some data have 2 values. Let's split them into 2 columns named Genres and SubGenres respectively.

In [11]:

Out[11]:

	Genres	SubGenres
0	Art & Design	None
1	Art & Design	Pretend Play
2	Art & Design	None
3	Art & Design	None
4	Art & Design	Creativity
5	Art & Design	None
6	Art & Design	None
7	Art & Design	None
8	Art & Design	None
9	Art & Design	Creativity

12. Show the top 5 columns with most occurence in Genres

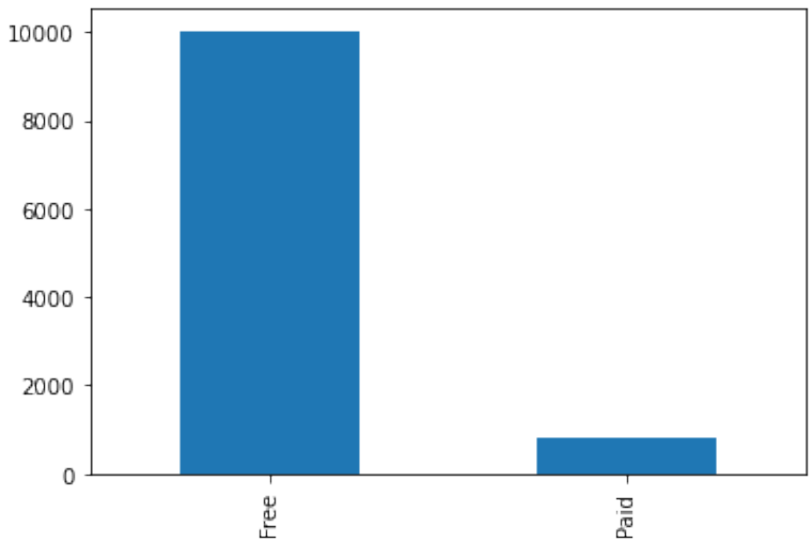
In [12]:

Out[12]:
```
Tools            841
Entertainment    666
Education        644
Medical          463
Business         460
Name: Genres, dtype: int64
```

Data visualization

13. Plot a bar plot for the type column

In [13]:



14. What's your observation?

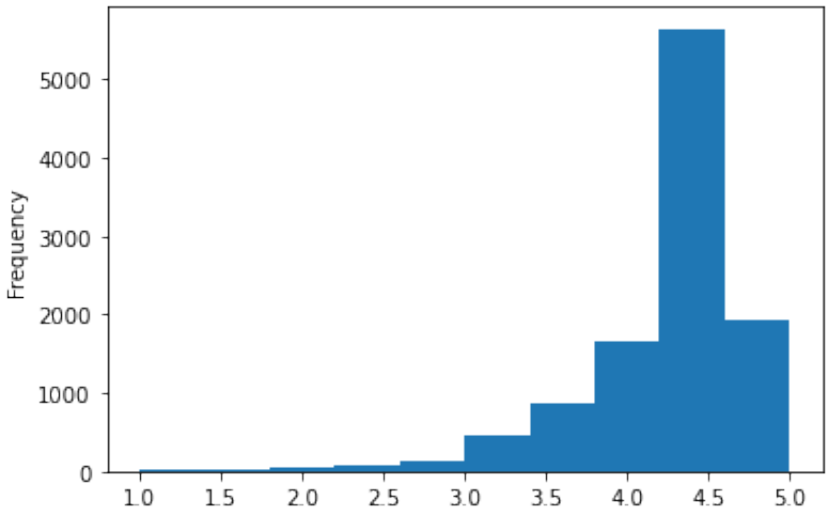
Google play store have more __ apps than ____ apps.

15. Plot a histogram for the rating column

In [14]:

```
```

Out[14]: <AxesSubplot:ylabel='Frequency'>



16. What's your observation?

Most ratings are distributed around value of _.