Table of Content:-

1. Import Libraries.
2. Read dataset.
3. Show abreif Show a brief of the dataframe.
4. Data Exploration.
5. Data Cleaning.
6. EDA (Exploratory Data Analysis).
7. Conclusion.

1. Import Libraries:-

In [4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as misno
```

2. Read dataset:-

In [5]:
```
## dataset (Google App Store) will be known as GAS
GAS=pd.read_csv(r'C:\Users\dell\Desktop\DATA_ SCIENCE\data_sets\kaggleDS\google app store\google.csv')
```

3. Show abreif Show a brief of the dataframe:-

In [6]:
```
## Dataset Called GAS (Google App Store)
## show first 5 rows
GAS.head()
```

Out[6]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
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	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

In [7]:
```
## Show last 5 rows
GAS.tail()
```

Out[7]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	And
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Va de
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Va de

4. Data Exploration:-

In [8]:
```
## Get shape of data
GAS.shape            # (Row, Column)
```

Out[8]:
```
(10841, 13)
```

In [9]:
```
GAS.columns  # columns in dataset
```

Out[9]:
```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

In [10]:
```
## Get More information about data
GAS.info()
```

```
<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
```

There are 10841 rows , Datatypes are object except Rating is Float64

In [11]:
```
## More Statiscal inforamtion
GAS.describe()
```

Out[11]:

	Rating
count	9367.000000
mean	4.193338
std	0.537431
min	1.000000
25%	4.000000
50%	4.300000
75%	4.500000
max	19.000000

In [10]:
```
## Check null values
GAS.isna().any()
```

Out[10]:
```
App               False
Category          False
Rating             True
Reviews           False
Size              False
Installs          False
Type               True
Price             False
Content Rating     True
Genres            False
Last Updated      False
Current Ver        True
Android Ver        True
dtype: bool
```

That mean (Rating, Type, Content Rating, current Ver, Andoried Ver) have missing values

In [11]:
```
## check the number of the missing values
GAS.isna().sum()
```

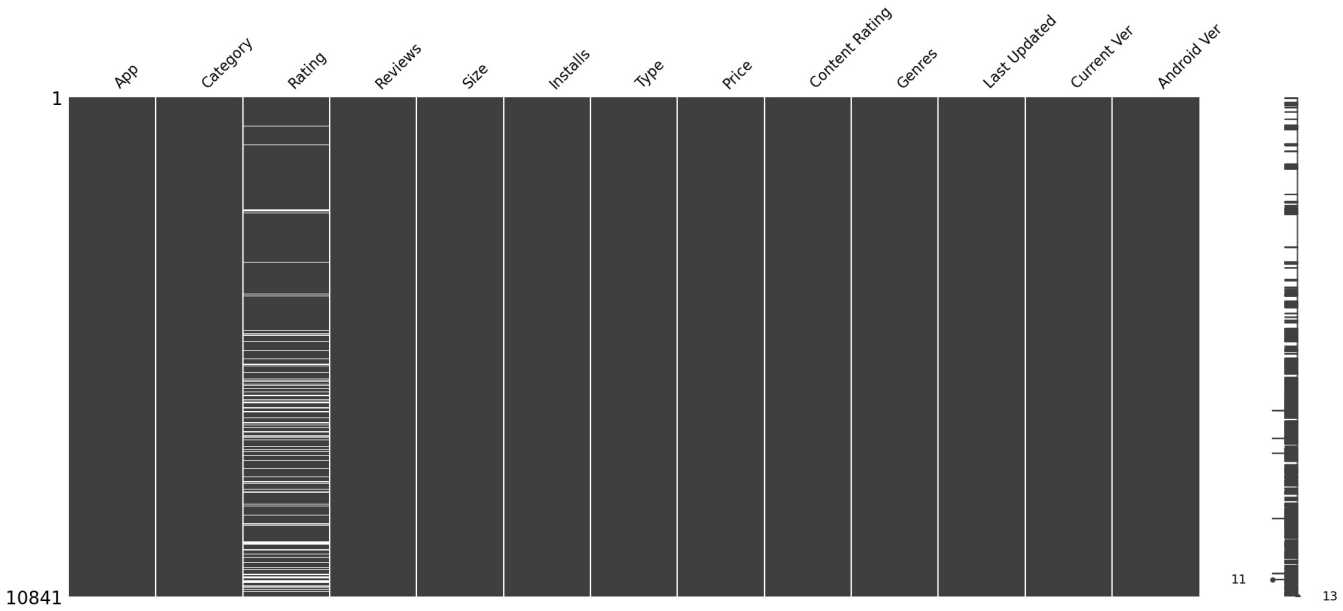
Out[11]:
```
App                0
Category           0
Rating          1474
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     1
Genres             0
Last Updated       0
Current Ver        8
Android Ver        3
dtype: int64
```

In [12]:
```
## check the percentage of the missing values
GAS.isna().sum()/100
```

Out[12]:
```
App               0.00
Category          0.00
Rating           14.74
Reviews           0.00
Size              0.00
Installs          0.00
Type              0.01
Price             0.00
Content Rating    0.01
Genres            0.00
Last Updated      0.00
Current Ver       0.08
Android Ver       0.03
dtype: float64
```

In [13]:
```
misno.matrix(GAS)  ## in the column of Rating there are white lines that mean missing values
```

Out[13]:
```
<AxesSubplot:>
```



It shows the missing values in Rating

4. Data Cleaning:-

In [14]:
```
## show the number of null values of (Rating)
```

```
GAS['Rating'].isna().sum()
```

Out[14]:
```
1474
```

In [15]:
```
## show the dataframe which is have missing values
GAS[GAS['Rating'].isna()]
```

Out[15]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated
23	Mcqueen Coloring pages	ART_AND_DESIGN	NaN	61	7.0M	100,000+	Free	0	Everyone	Art & Design;Action & Adventure	March 7, 2018
113	Wrinkles and rejuvenation	BEAUTY	NaN	182	5.7M	100,000+	Free	0	Everyone 10+	Beauty	September 20, 2017
123	Manicure - nail design	BEAUTY	NaN	119	3.7M	50,000+	Free	0	Everyone	Beauty	July 23, 2018
126	Skin Care and Natural Beauty	BEAUTY	NaN	654	7.4M	100,000+	Free	0	Teen	Beauty	July 17, 2018
129	Secrets of beauty, youth and health	BEAUTY	NaN	77	2.9M	10,000+	Free	0	Mature 17+	Beauty	August 8, 2017
...	...	...	...	...	...	...	...	...	...	...	...
10824	Cardio-FR	MEDICAL	NaN	67	82M	10,000+	Free	0	Everyone	Medical	July 31, 2018
10825	Naruto & Boruto FR	SOCIAL	NaN	7	7.7M	100+	Free	0	Teen	Social	February 2, 2018
10831	payermonstationnement.fr	MAPS_AND_NAVIGATION	NaN	38	9.8M	5,000+	Free	0	Everyone	Maps & Navigation	June 13, 2018
10835	FR Forms	BUSINESS	NaN	0	9.6M	10+	Free	0	Everyone	Business	September 29, 2016
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	January 20, 2017

1474 rows × 13 columns

In [12]:
```
# Delete all rows which is have missing values of Rating column
GAS.dropna(subset=['Rating'] , axis=0, inplace = True)
```

In [13]:
```
## check missng values after deleting
GAS['Rating'].isna().sum()
```

Out[13]:
```
0
```

In [14]:
```
# Delete all rows which is have missing values of Type column
GAS.dropna(subset=['Type'] , axis=0, inplace = True)
```

In [15]:
```
## check missng values after deleting
GAS['Type'].isna().sum()
```

Out[15]:
```
0
```

That mean no missing values , But there are columns we donot need so i will delete them.

In [16]:
```
# Delete columns which is not helpful for us

x = GAS[['Content Rating', 'Current Ver', 'Android Ver']]
GAS.drop(x.columns.tolist(), axis=1, inplace=True)
```

In [17]:
```
## show the first 5 rows after deleting columns
GAS.head()
```

Out[17]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Genres	Last Updated
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Art & Design	January 7, 2018
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Art & Design;Pretend Play	January 15, 2018
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Art & Design	August 1, 2018
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Art & Design	June 8, 2018
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Art & Design;Creativity	June 20, 2018

5. EDA (Exploratory Data Analysis):

5. EDA (Exploratory Data Analysis):-

5.1- What is the categories in dataset?

In [85]:
```
## show all Categories
GAS['Category'].unique()
```

Out[85]:
```
array(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
       'BOOKS_AND_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATION',
       'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANCE',
       'FOOD_AND_DRINK', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME',
       'LIBRARIES_AND_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MEDICAL',
       'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL_AND_LOCAL',
       'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING', 'WEATHER',
       'VIDEO_PLAYERS', 'NEWS_AND_MAGAZINES', 'MAPS_AND_NAVIGATION'],
      dtype=object)
```

In [86]:
```
## show the number Categories
GAS['Category'].nunique()
```

Out[86]:
```
33
```

5.2- What is the top 5 categories?

In [87]:
```
A = GAS['Category'].value_counts().index
A
```

Out[87]:
```
Index(['FAMILY', 'GAME', 'TOOLS', 'PRODUCTIVITY', 'MEDICAL', 'COMMUNICATION',
       'FINANCE', 'SPORTS', 'PHOTOGRAPHY', 'PERSONALIZATION', 'LIFESTYLE',
       'BUSINESS', 'HEALTH_AND_FITNESS', 'SOCIAL', 'SHOPPING',
       'NEWS_AND_MAGAZINES', 'TRAVEL_AND_LOCAL', 'DATING',
       'BOOKS_AND_REFERENCE', 'VIDEO_PLAYERS', 'EDUCATION', 'ENTERTAINMENT',
       'MAPS_AND_NAVIGATION', 'FOOD_AND_DRINK', 'HOUSE_AND_HOME', 'WEATHER',
       'AUTO_AND_VEHICLES', 'LIBRARIES_AND_DEMO', 'ART_AND_DESIGN', 'COMICS',
       'PARENTING', 'EVENTS', 'BEAUTY'],
      dtype='object', name='Category')
```

In [88]:
```
B = GAS['Category'].value_counts().values
B
```

Out[88]:
```
array([1747, 1097,  734,  351,  350,  328,  323,  319,  317,  314,  314,
        303,  297,  259,  238,  233,  226,  195,  178,  160,  155,  149,
        124,  109,   76,   75,   73,   65,   62,   58,   50,   45,   42],
      dtype=int64)
```

In [89]:
```
## show each Category and it is number
GAS['Category'].value_counts()
```

Out[89]:
```
Category
FAMILY                1747
GAME                  1097
TOOLS                  734
PRODUCTIVITY           351
MEDICAL                350
COMMUNICATION          328
FINANCE                323
SPORTS                 319
PHOTOGRAPHY            317
PERSONALIZATION        314
LIFESTYLE              314
BUSINESS               303
HEALTH_AND_FITNESS     297
SOCIAL                 259
SHOPPING               238
NEWS_AND_MAGAZINES     233
TRAVEL_AND_LOCAL       226
DATING                 195
BOOKS_AND_REFERENCE    178
VIDEO_PLAYERS          160
EDUCATION              155
ENTERTAINMENT          149
MAPS_AND_NAVIGATION    124
FOOD_AND_DRINK         109
HOUSE_AND_HOME          76
WEATHER                 75
AUTO_AND_VEHICLES       73
LIBRARIES_AND_DEMO      65
ART_AND_DESIGN          62
COMICS                  58
PARENTING               50
EVENTS                  45
BEAUTY                  42
Name: count, dtype: int64
```

that mean top 5 categories are (FAMILY, GAME, TOOLS, PRODUCTIVITY, MEDICAL).

5.3- What are the Ratings in dataset?

In [90]:
```
## show the numbers of Ratings
GAS['Rating'].nunique()
```

Out[90]:
```
39
```

In [91]:
```
## show the Ratings
GAS['Rating'].unique()
```

Out[91]:
```
array([4.1, 3.9, 4.7, 4.5, 4.3, 4.4, 3.8, 4.2, 4.6, 3.2, 4. , 4.8, 4.9,
       3.6, 3.7, 3.3, 3.4, 3.5, 3.1, 5. , 2.6, 3. , 1.9, 2.5, 2.8, 2.7,
       1. , 2.9, 2.3, 2.2, 1.7, 2. , 1.8, 2.4, 1.6, 2.1, 1.4, 1.5, 1.2])
```

In [92]:
```
GAS['Rating'].max()     ## maxmum Rating
```

Out[92]:
```
5.0
```

In [93]:
```
GAS['Rating'].min()     ## minimum Rating
```

Out[93]:
```
1.0
```

In [94]:
```
GAS['Rating'].mean()    ## mean Rating
```

Out[94]:
```
4.191757420456972
```

In [95]:
```
GAS['Rating'].median()  ## median Rating
```

Out[95]:
```
4.3
```

In [96]:
```
GAS['Rating'].std()   ## Standered Deviation (spread of data) of Rating
```

Out[96]:
```
0.5152188586177868
```

that mean and median between 4.19 : 4.3 that mean Rating 19 is (outlier) extreme value

In [18]:
```
## print all rows whichs is have catagory = 19
GAS[GAS['Rating']==19]
```

Out[18]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Genres	Last Updated
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	February 11, 2018	1.0.19

In [19]:
```
## Delete the the outlier to avoid it is effect on data
GAS.drop(index=10472, inplace=True)
```

5.4- What is top 10 Applications according to Rating?

In [20]:
```
## print the top 10 Apps with Rating in ascending order
GAS.sort_values(by='Rating' , axis = 0 , ascending = False ).head(10)
```

Out[20]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Genres	Last Updated
9056	Santa's Monster Shootout DX	GAME	5.0	4	33M	50+	Paid	$1.99	Action	August 15, 2013
8395	DG TV	NEWS_AND_MAGAZINES	5.0	3	5.7M	100+	Free	0	News & Magazines	May 26, 2018
8493	PK and DK Audio App	FAMILY	5.0	2	3.9M	100+	Free	0	Entertainment	October 25, 2017
6330	HON. B.J. ACS COLLEGE ALE	FAMILY	5.0	3	1.8M	100+	Free	0	Education	December 26, 2016
6342	BJ Foods	BUSINESS	5.0	3	1.5M	10+	Free	0	Business	February 7, 2018
6363	Read it easy for BK	LIFESTYLE	5.0	1	3.2M	50+	Free	0	Lifestyle	July 15, 2018
9766	ER Assist	PRODUCTIVITY	5.0	3	28M	10+	Free	0	Productivity	December 6, 2016
6364	BK Video Status	FAMILY	5.0	13	2.1M	100+	Free	0	Entertainment	July 7, 2018
6372	BK Formula Calculator	TOOLS	5.0	6	11M	100+	Free	0	Tools	August 8, 2015
6375	Dr Bk Sachin bhai	LIFESTYLE	5.0	19	3.1M	1,000+	Free	0	Lifestyle	December 7, 2017

5.5- What is best free Applications according to installs?

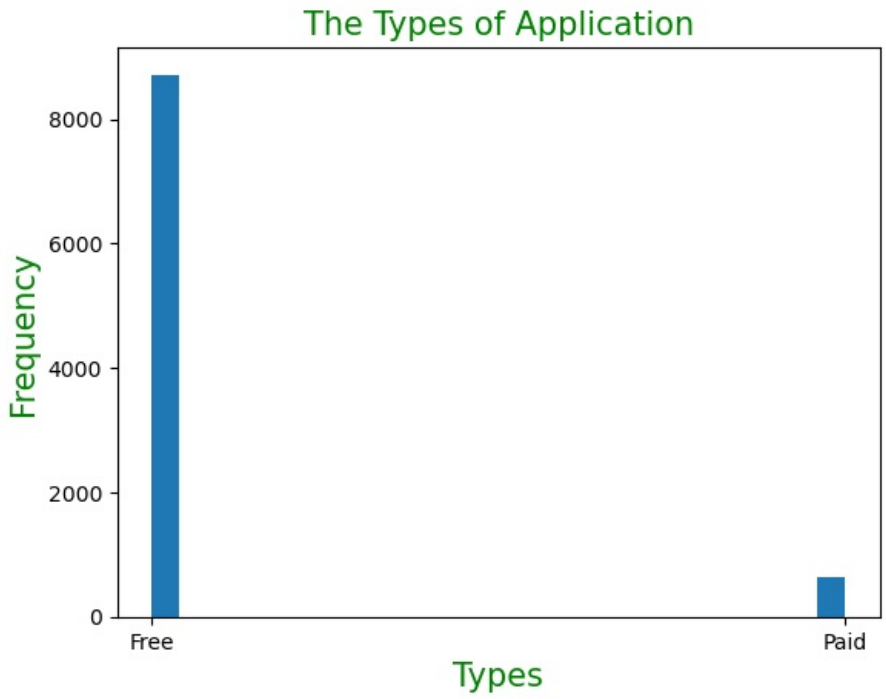
In [23]:
```
## check number of types
GAS['Type'].value_counts()
```

Out[23]:
```
Type
Free    8719
Paid     647
Name: count, dtype: int64
```

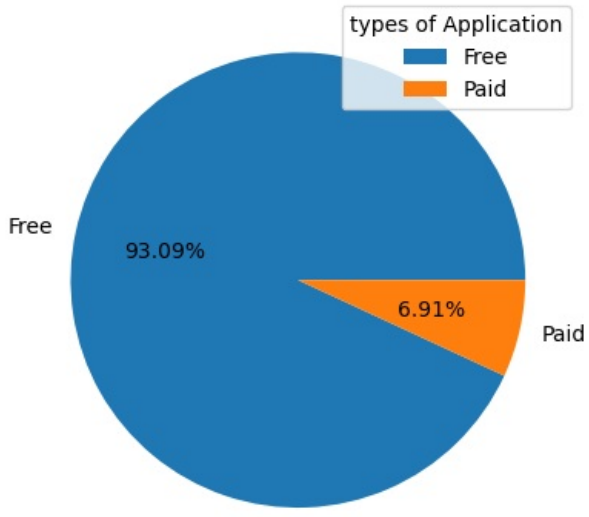
In [26]:
```
## visuallize the number of types

plt.hist(GAS['Type'], bins=25)

plt.title('The Types of Application', color = 'green', fontsize = 15)
plt.xlabel('Types', color = 'green', fontsize = 15)
plt.ylabel('Frequency', color = 'green', fontsize = 15)
plt.show()
```



In [90]:
```
## show a pie chart of types of Application
type_show = ['Free', 'Paid']
Value_count =  [8719,647]
plt.pie(Value_count,labels=type_show,autopct="%2.2f%%")
plt.legend(title='types of Application')
plt.show()
```



In [64]:
```
## print the top 10 free Apps with Installs in ascending order
GAS.sort_values(by=['Installs','Type','Rating'] , axis = 0 , ascending = False ).head(5)
```

Out[64]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Genres	Last Updated
4005	Clean Master- Space Cleaner & Antivirus	TOOLS	4.7	42916526	Varies with device	500,000,000+	Free	0	Tools	August 3, 2018
7536	Security Master - Antivirus, VPN, AppLock, Boo...	TOOLS	4.7	24900999	Varies with device	500,000,000+	Free	0	Tools	August 4, 2018
371	Google Duo - High Quality Video Calls	COMMUNICATION	4.6	2083237	Varies with device	500,000,000+	Free	0	Communication	July 31, 2018
3255	SHAREit - Transfer & Share	TOOLS	4.6	7790693	17M	500,000,000+	Free	0	Tools	July 30, 2018
4039	Google Duo - High Quality Video Calls	COMMUNICATION	4.6	2083237	Varies with device	500,000,000+	Free	0	Communication	July 31, 2018

It mean that the best free App is (Clean Master-Space Cleaner&Antivirus) based on installs,Rating

---

5.6- What is the number of installs?

In [72]:
```
## show the installs numbers
GAS['Installs'].unique()
```

Out[72]:
```
array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,000+',
       '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,000,000+',
       '1,000,000,000+', '1,000+', '500,000,000+', '100+', '500+', '10+',
       '5+', '50+', '1+'], dtype=object)
```

In [73]:
```
GAS['Installs'].value_counts()
```

Out[73]:
```
Installs
1,000,000+        1577
10,000,000+       1252
100,000+          1150
10,000+           1010
5,000,000+         752
1,000+             713
500,000+           538
50,000+            467
5,000+             432
100,000,000+       409
100+               309
50,000,000+        289
500+               201
500,000,000+        72
10+                 69
1,000,000,000+      58
50+                 56
5+                   9
1+                   3
Name: count, dtype: int64
```

---

7. Conclusion:-

-We explore the Google App Store dataset and learn more about data attributes then jump into how to visualize the data with Exploratory Data Analysis.

-We saw some basic and advanced level charts of matplotlib like (Pie-chart , Histogram , missino-matrix)

- The questions were answeres within the Analysis: _

1. What is the categories in dataset?
2. What is the top 5 categories?
3. What are the Ratings in dataset?
4. What is top 10 Applications according to Rating?
5. What is best free Applications according to installs?
6. What is the number of installs?

Thank You, Happy to get any suggestions or feedbacks