- **Problem-** We found out that the categorization system of Google Play does not respect properly similarity of applications**.**No proper way of searching apps according to their ratings and other factors (their reviews, price etc)

- **Our Solution-** In-depth Analysis of the Google Play Store dataset along with enhancing searching according to rating. We will draw a comprehensive picture of current situation of App market in order to help application developers to understand customers' desire and attitude and the trend in the market. Moreover, by employing K-means clustering method, we will propose a new solution to enhance searching based on ratings.

- **Many other Insights are taken along with our main goal.**

- **<u>INSTALLATION AND TOOLS NEEDED</u>**

This covers all the description and installation of all the tools required in this project. All the software and packages used are Open Source software and are freely available.

- ✓ **System Requirements**

Type of Hardware	**Hardware Requirements**
Hardware	Dual core Intel Pentium compatible processor
Disk space	4GB disk space(minimum)
Memory	4GB(recommended)

Type of software	**Software Requirements**
Operating System	Independent
Web browser	Chrome (recommended)
Coding Environment	Jupyter Notebook
Packages needed	Python, Numpy, pandas, matplotlib, seaborn, scipy, sklearn etc

- **Approach-**

✓ A Look at Dataset-

```
In [2]: df = pd.read_csv("googleplaystore.csv")

In [3]: df.head()

Out[3]:
```

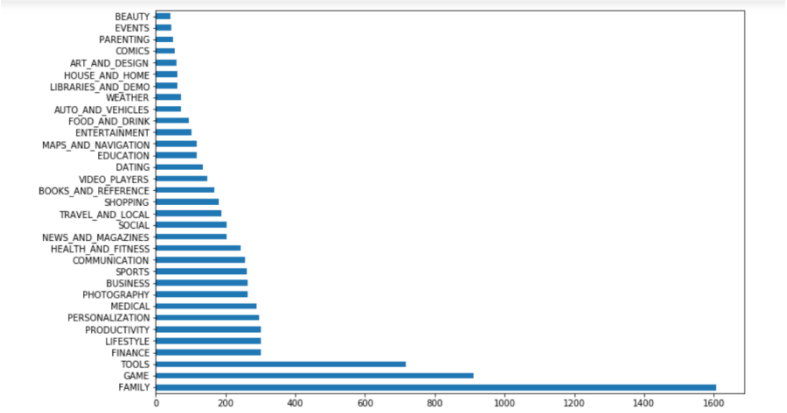
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	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 [4]: df.shape
Out[4]: (10841, 13)
```

✓ Dropping Null values and Duplicate apps

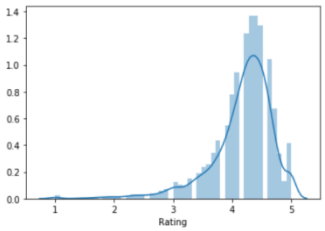
✓ Data Cleaning by removing irrelevant symbols (like + from Installs, M and k from Size etc) and converting string values to int or float as per their need.

✓ Category vs Number of unique apps available.
Insight: Maximum Number of Apps belongs to the Family and Game Category.



✓ **Insight** : Most of the apps, clearly hold a rating above 4 And surprisingly a lot seem to have 5 rating.

```
In [22]: # df.Rating.describe()
         sns.distplot(df.Rating)
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x1b4070c8b88>
```



✓ Even most popular apps like Whatsapp, Facebook, Instagram etc that don't have 5 Rating.

```
In [24]: df[(df.Reviews>25543600)]
```

Out[24]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
335	Messenger – Text and Video Chat for Free	COMMUNICATION	4.0	56642847	NaN	1000000000	Free	0.0	Everyone	Communication	August 1, 2018	Varies with device	Varies with device
336	WhatsApp Messenger	COMMUNICATION	4.4	69119316	NaN	1000000000	Free	0.0	Everyone	Communication	August 3, 2018	Varies with device	Varies with device
1654	Subway Surfers	GAME	4.5	27722264	76.0	1000000000	Free	0.0	Everyone 10+	Arcade	July 12, 2018	1.90.0	4.1 and up
1670	Clash of Clans	GAME	4.6	44891723	98.0	100000000	Free	0.0	Everyone 10+	Strategy	July 15, 2018	10.322.16	4.1 and up
2544	Facebook	SOCIAL	4.1	78158306	NaN	1000000000	Free	0.0	Teen	Social	August 3, 2018	Varies with device	Varies with device
2545	Instagram	SOCIAL	4.5	66577313	NaN	1000000000	Free	0.0	Teen	Social	July 31, 2018	Varies with device	Varies with device
3665	YouTube	VIDEO_PLAYERS	4.3	25655305	NaN	1000000000	Free	0.0	Teen	Video Players & Editors	August 2, 2018	Varies with device	Varies with device
4005	Clean Master- Space Cleaner & Antivirus	TOOLS	4.7	42916526	NaN	500000000	Free	0.0	Everyone	Tools	August 3, 2018	Varies with device	Varies with device

✓ Most of the 5 star Rated apps are Dating apps with low reviews and installs.

```
In [25]: df[df.Rating == 5]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
329	Hojiboy Tojiboyev Life Hacks	COMICS	5.0	15	37.00000	1000	Free	0.00	Everyone	Comics	June 26, 2018	2.0	4.0.3 and up
612	American Girls Mobile Numbers	DATING	5.0	5	4.40000	1000	Free	0.00	Mature 17+	Dating	July 17, 2018	3.0	4.0.3 and up
615	Awake Dating	DATING	5.0	2	70.00000	100	Free	0.00	Mature 17+	Dating	July 24, 2018	2.2.9	4.4 and up
633	Spine- The dating app	DATING	5.0	5	9.30000	500	Free	0.00	Teen	Dating	July 14, 2018	4.0	4.0.3 and up
636	Girls Live Talk - Free Text and Video Chat	DATING	5.0	6	5.00000	100	Free	0.00	Mature 17+	Dating	August 1, 2018	8.2	4.0.3 and up
640	Online Girls Chat Group	DATING	5.0	5	5.00000	100	Free	0.00	Mature 17+	Dating	August 2, 2018	8.2	4.0.3 and up

```
In [26]: df[(df.Rating == 5) & (df.Reviews >= 30) & (df.Installs > 100) ].App.count()
Out[26]: 12
```

✓ Only 12 apps with 5 Star Rating with sufficient reviews and installs

```
In [27]: #take a look at the apps priced more than 100
expensive_apps = df[df["Price"]>100]
expensive_apps["Installs"].groupby(expensive_apps["App"]).sum()
```

```
Out[27]: App
I AM RICH PRO PLUS                1000
I Am Rich                        10000
I Am Rich Premium                50000
I Am Rich Pro                     5000
I am Rich                         5000
I am Rich Plus                   10000
I am Rich!                        1000
I am extremely Rich               1000
I am rich                       100000
I am rich (Most expensive app)    1000
I am rich VIP                    10000
I am rich(premium)                5000
I'm Rich - Trump Edition         10000
most expensive app (H)             100
💎 I'm rich                      10000
Name: Installs, dtype: int32
```

✓ Applied K-Means Clustering for improving searching based on rating ranges:

```
In [29]: kmeans = KMeans(n_clusters=5, random_state=0).fit(df_tr_std)
         labels = kmeans.labels_
         df_tr['clusters'] = labels
         #Add the column into our list
         clmns.extend(['clusters'])
         clmns

Out[29]: ['Rating', 'Installs', 'clusters']

In [30]: #Lets analyze the clusters
         df_tr[clmns].groupby(['clusters']).mean()
```

Out[30]:

clusters	Rating	Installs
0	4.529281	7.172939e+06
1	4.215000	1.000000e+09
2	2.894142	5.565420e+05
3	3.973677	3.783617e+06
4	4.375000	5.000000e+08

✓ Analysis of Clusters formed:

```
In [32]: df_tr["clusters"].unique()
Out[32]: array([3, 0, 2, 1, 4], dtype=int64)

In [33]: df_tr["clusters"].value_counts()
Out[33]: 0    4310
         3    3119
         2     717
         4      24
         1      20
         Name: clusters, dtype: int64

In [34]: df_tr[(df_tr['clusters']==4)].max()
Out[34]: App                imo free video calls and chat
         Category                        VIDEO_PLAYERS
         Rating                                    4.7
         Reviews                              42916526
         Size                                       74
         Installs                            500000000
         Type                                     Free
         Price                                       0
         Content Rating                           Teen
         Genres                 Video Players & Editors
         Last Updated                      May 25, 2018
         Current Ver                 Varies with device
         Android Ver                 Varies with device
         clusters                                    4
         dtype: object
```

- **Survey-**

✓ Link to research papers: http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/11407/15101108%2C15101020%2C15101109%2C15141002_CSE.pdf?sequence=1&isAllowed=y

✓ https://www.researchgate.net/publication/290102532_Mining_and_analysis_of_apps_in_google_play

- **CONCLUSION** :

By visualizing and analyzing the Google play dataset we will cluster different categories of apps based on their features like number of installs and reviews. This will help to make the search easy and efficient. It will also tell the best categories of app according to given features.