Data Cleansing - Google Play Store Data

Source:

https://www.kaggle.com/lava18/google-play-store-apps (https://www.kaggle.com/lava18/google-play-store-apps)

Defining the Problem Statement

This dataset records the attributes of Android mobile applications in the Google Play Store. From this dataset, we would like to be able to find the best clustering results/optimum number of clusters.

Data Cleansing

Before we work on the clustering, we will need to prepare the data:

- First, we will remove all null value rows
- Then, we will drop irrelevant columns
- Also, we will perform encoding of values
- Lastly, we will be removing irrelevant characters such as $.

We will export the clean data to a new CSV file.

In [54]:

```
import pandas as pd
import matplotlib.pyplot as plt
#Read the file
df = pd.read_csv('googleplaystore.csv')
```

In [55]:

```
#Remove all the null value rows
dfclean = df.copy().dropna()
```

In [56]:

```
#drop irrelevant columns
to_drop =['App','Last Updated','Current Ver','Android Ver']
```

In [57]:

```
dfclean.drop(to_drop, inplace=True,axis =1)
```

In [58]:

```
#Categorizing the values of the Content Rating column
dict = {
    'Adults only 18+': 0,
    'Everyone': 1,
    'Everyone 10+': 2,
    'Mature 17+': 3,
    'Teen': 4,
    'Unrated': 5
}

dfclean['Content Rating'] = dfclean['Content Rating'].map(dict)
```

In [59]:

```
#Categorizing the values of the Category column
dict = {'ART_AND_DESIGN':0,
 'AUTO_AND_VEHICLES':1,
 'BEAUTY':2,
 'BOOKS_AND_REFERENCE':3,
 'BUSINESS':4,
 'COMICS':5,
 'COMMUNICATION':6,
 'DATING':7,
 'EDUCATION':8,
 'ENTERTAINMENT':9,
 'EVENTS':10,
 'FAMILY':11,
 'FINANCE':12,
 'FOOD_AND_DRINK':13,
 'GAME':14,
 'HEALTH_AND_FITNESS':15,
 'HOUSE_AND_HOME':16,
 'LIBRARIES_AND_DEMO':17,
 'LIFESTYLE':18,
 'MAPS_AND_NAVIGATION':19,
 'MEDICAL':20,
 'NEWS_AND_MAGAZINES':21,
 'PARENTING':22,
 'PERSONALIZATION':23,
 'PHOTOGRAPHY':24,
 'PRODUCTIVITY':25,
 'SHOPPING':26,
 'SOCIAL':27,
 'SPORTS':28,
 'TOOLS':29,
 'TRAVEL_AND_LOCAL':30,
 'VIDEO_PLAYERS':31,
 'WEATHER':32}

dfclean['Category'] = dfclean['Category'].map(dict)
```

In [60]:

```
#Categorizing the values of the Type column
set(dfclean['Type'])
```

Out[60]:

```
{'Free', 'Paid'}
```

In [61]:

```
dict1 = {'Free':0, 'Paid':1}
dfclean['Type'] = dfclean['Type'].map(dict1)
```

In [62]:

```
#Categorizing the values of the Genre column
dict2 ={'Action':0,
 'Action;Action & Adventure':1,
 'Adventure':2,
 'Adventure;Action & Adventure':3,
 'Adventure;Brain Games':4,
 'Adventure;Education':5,
 'Arcade':6,
 'Arcade;Action & Adventure':7,
 'Arcade;Pretend Play':8,
 'Art & Design':9,
 'Art & Design;Creativity':10,
 'Art & Design;Pretend Play':11,
 'Auto & Vehicles':12,
 'Beauty':13,
 'Board':14,
 'Board;Brain Games':15,
 'Board;Pretend Play':16,
 'Books & Reference':17,
 'Books & Reference;Education':18,
 'Business':19,
 'Card':20,
 'Card;Action & Adventure':21,
 'Card;Brain Games':22,
 'Casino':23,
 'Casual':24,
 'Casual;Action & Adventure':25,
 'Casual;Brain Games':26,
 'Casual;Creativity':27,
 'Casual;Education':28,
 'Casual;Music & Video':29,
 'Casual;Pretend Play':30,
 'Comics':31,
 'Comics;Creativity':32,
 'Communication':33,
 'Communication;Creativity':34,
 'Dating':35,
 'Education':36,
 'Education;Action & Adventure':37,
 'Education;Brain Games':38,
 'Education;Creativity':39,
 'Education;Education':40,
 'Education;Music & Video':41,
 'Education;Pretend Play':42,
 'Educational':43,
 'Educational;Action & Adventure':44,
 'Educational;Brain Games':45,
 'Educational;Creativity':46,
 'Educational;Education':47,
 'Educational;Pretend Play':48,
 'Entertainment':49,
 'Entertainment;Action & Adventure':50,
 'Entertainment;Brain Games':51,
 'Entertainment;Creativity':52,
 'Entertainment;Education':53,
 'Entertainment;Music & Video':54,
 'Entertainment;Pretend Play':55,
 'Events':56,
 'Finance':57,
```

```
 'Food & Drink':58,
 'Health & Fitness':59,
 'Health & Fitness;Action & Adventure':60,
 'Health & Fitness;Education':61,
 'House & Home':62,
 'Libraries & Demo':63,
 'Lifestyle':64,
 'Lifestyle;Education':65,
 'Lifestyle;Pretend Play':66,
 'Maps & Navigation':67,
 'Medical':68,
 'Music':69,
 'Music & Audio;Music & Video':70,
 'Music;Music & Video':71,
 'News & Magazines':72,
 'Parenting':73,
 'Parenting;Brain Games':74,
 'Parenting;Education':75,
 'Parenting;Music & Video':76,
 'Personalization':77,
 'Photography':78,
 'Productivity':79,
 'Puzzle':80,
 'Puzzle;Action & Adventure':81,
 'Puzzle;Brain Games':82,
 'Puzzle;Creativity':83,
 'Puzzle;Education':84,
 'Racing':85,
 'Racing;Action & Adventure':86,
 'Racing;Pretend Play':87,
 'Role Playing':88,
 'Role Playing;Action & Adventure':89,
 'Role Playing;Brain Games':90,
 'Role Playing;Pretend Play':91,
 'Shopping':92,
 'Simulation':93,
 'Simulation;Action & Adventure':94,
 'Simulation;Education':95,
 'Simulation;Pretend Play':96,
 'Social':97,
 'Sports':98,
 'Sports;Action & Adventure':99,
 'Strategy':100,
 'Strategy;Action & Adventure':101,
 'Strategy;Creativity':102,
 'Strategy;Education':103,
 'Tools':104,
 'Tools;Education':105,
 'Travel & Local':106,
 'Travel & Local;Action & Adventure':107,
 'Trivia':108,
 'Video Players & Editors':109,
 'Video Players & Editors;Creativity':110,
 'Video Players & Editors;Music & Video':111,
 'Weather':112,
 'Word':113,
 'Board;Action & Adventure':114}

dfclean['Genres'] = dfclean['Genres'].map(dict2)
```

In [63]:

```
#Categorizing the values of the Installs column
dfclean
dfclean['Installs'].unique()
```

Out[63]:

```
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 [64]:

```
dfclean['Installs'].sort_values().unique()
```

Out[64]:

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

In [65]:

```
dict3 = {'1+':1,
         '5+':2,
         '10+':3,
         '50+':4,
         '100+':5,
         '500+':6,
          '1,000+':7,
         '5,000+':8,
         '10,000+':9,
         '50,000+':10,
         '100,000+':11,
         '500,000+':12,
        '1,000,000+':13,
         '5,000,000+':14,
         '10,000,000+':15,
         '50,000,000+':16,
         '100,000,000+':17,
         '500,000,000+':18,'1,000,000,000+':19
  }
```

In [66]:

```
dfclean['Installs'] = dfclean['Installs'].map(dict3)
```

In [67]:

```
dfclean.loc[df['Rating']<= 2, 'Rating'] = 'Low Rating'
dfclean.loc[(df['Rating'] > 2) & (df['Rating'] < 4), 'Rating'] = 'Average Rating'
dfclean.loc[df['Rating'] >= 4, 'Rating'] = 'High Rating'
```

In [68]:

```
set(dfclean['Price'])
```

Out[68]:

```
{'$0.99',
 '$1.00',
 '$1.20',
 '$1.29',
 '$1.49',
 '$1.50',
 '$1.59',
 '$1.61',
 '$1.70',
 '$1.75',
 '$1.76',
 '$1.97',
 '$1.99',
 '$10.00',
 '$10.99',
 '$11.99',
 '$12.99',
 '$13.99',
 '$14.00',
 '$14.99',
 '$15.46',
 '$15.99',
 '$16.99',
 '$17.99',
 '$18.99',
 '$19.40',
 '$19.99',
 '$2.00',
 '$2.49',
 '$2.50',
 '$2.56',
 '$2.59',
 '$2.90',
 '$2.95',
 '$2.99',
 '$24.99',
 '$29.99',
 '$299.99',
 '$3.02',
 '$3.04',
 '$3.08',
 '$3.28',
 '$3.49',
 '$3.88',
 '$3.90',
 '$3.95',
 '$3.99',
 '$33.99',
 '$37.99',
 '$379.99',
 '$389.99',
 '$39.99',
 '$399.99',
 '$4.29',
 '$4.49',
 '$4.59',
 '$4.60',
 '$4.77',
 '$4.84',
```

```
 '$4.99',
 '$400.00',
 '$5.49',
 '$5.99',
 '$6.49',
 '$6.99',
 '$7.49',
 '$7.99',
 '$79.99',
 '$8.49',
 '$8.99',
 '$9.00',
 '$9.99',
 '0'}
```

In [69]:

```
#Removing $ in front of Price Column
dfclean['Price'] = dfclean['Price'].str.replace('$','')
```

In [70]:

```
#Removing k and M values at the end of the Size Column
dfclean['Size'] = dfclean['Size'].str.rstrip('M')
dfclean['Size'] = dfclean['Size'].str.rstrip('k')
```

In [71]:

```
dfclean = dfclean[dfclean.Size != 'Varies with device']
```

In [72]:

```
dfclean=dfclean.copy()
dfclean['Size']=dfclean['Size'].apply(lambda x:float(x))
dfclean['Reviews']=dfclean['Reviews'].apply(lambda x:float(x))
dfclean['Price']=dfclean['Price'].apply(lambda x:float(x))
```

In [73]:

```
dfclean
```

Out[73]:

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
0	0	High Rating	159.0	19.0	9	0	0.0	1	9
1	0	Average Rating	967.0	14.0	12	0	0.0	1	11
2	0	High Rating	87510.0	8.7	14	0	0.0	1	9
3	0	High Rating	215644.0	25.0	16	0	0.0	4	9
4	0	High Rating	967.0	2.8	11	0	0.0	1	10
5	0	High Rating	167.0	5.6	10	0	0.0	1	9
6	0	Average Rating	178.0	19.0	10	0	0.0	1	9
7	0	High Rating	36815.0	29.0	13	0	0.0	1	9
8	0	High Rating	13791.0	33.0	13	0	0.0	1	9
9	0	High Rating	121.0	3.1	9	0	0.0	1	10
10	0	High Rating	13880.0	28.0	13	0	0.0	1	9
11	0	High Rating	8788.0	12.0	13	0	0.0	1	9
12	0	High Rating	44829.0	20.0	15	0	0.0	4	9
13	0	High Rating	4326.0	21.0	11	0	0.0	1	9
14	0	High Rating	1518.0	37.0	11	0	0.0	1	9
16	0	High Rating	3632.0	5.5	12	0	0.0	1	9
17	0	High Rating	27.0	17.0	9	0	0.0	1	9
18	0	High Rating	194216.0	39.0	14	0	0.0	1	9
19	0	High Rating	224399.0	31.0	15	0	0.0	1	9
20	0	High Rating	450.0	14.0	11	0	0.0	1	9
21	0	High Rating	654.0	12.0	11	0	0.0	1	9
22	0	High Rating	7699.0	4.2	12	0	0.0	2	9
24	0	High Rating	118.0	23.0	10	0	0.0	1	9
25	0	High Rating	192.0	6.0	9	0	0.0	1	9
26	0	High Rating	20260.0	25.0	12	0	0.0	1	10
27	0	High Rating	203.0	6.1	11	0	0.0	1	9
28	0	Average Rating	136.0	4.6	9	0	0.0	1	9
29	0	High Rating	223.0	4.2	11	0	0.0	1	9
30	0	High Rating	1120.0	9.2	11	0	0.0	1	9
31	0	High Rating	227.0	5.2	10	0	0.0	1	9
...	...	...	...	...	...	...	...	...	...
10792	14	High Rating	21661.0	16.0	13	0	0.0	1	108
10793	14	High Rating	28510.0	78.0	12	0	0.0	4	20
10795	29	High Rating	7339.0	4.0	11	0	0.0	1	104
10796	29	High Rating	61445.0	7.8	13	0	0.0	1	104

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
10797	18	High Rating	32433.0	46.0	13	0	0.0	1	64
10799	27	High Rating	2036.0	6.8	11	0	0.0	1	97
10800	29	High Rating	174.0	12.0	8	0	0.0	1	104
10801	11	High Rating	52.0	19.0	7	0	0.0	1	36
10802	11	High Rating	185.0	28.0	9	0	0.0	4	49
10803	14	High Rating	56496.0	81.0	13	0	0.0	4	0
10804	14	High Rating	5442.0	17.0	11	0	0.0	4	20
10805	18	High Rating	3.0	15.0	5	0	0.0	1	64
10809	11	High Rating	376223.0	24.0	13	0	0.0	1	100
10810	4	High Rating	19.0	21.0	5	0	0.0	1	19
10812	11	High Rating	80.0	13.0	7	0	0.0	1	36
10814	11	High Rating	785.0	31.0	10	0	0.0	4	49
10815	3	High Rating	5775.0	4.9	12	0	0.0	1	17
10817	29	High Rating	885.0	8.0	11	0	0.0	1	104
10819	3	Average Rating	52.0	3.6	8	0	0.0	4	17
10820	11	High Rating	22.0	8.6	7	0	0.0	4	36
10827	11	High Rating	117.0	13.0	8	0	0.0	1	36
10828	5	Average Rating	291.0	13.0	9	0	0.0	1	31
10829	3	High Rating	603.0	7.4	9	0	0.0	1	17
10830	21	Average Rating	881.0	2.3	11	0	0.0	1	72
10832	32	Average Rating	1195.0	582.0	11	0	0.0	1	112
10833	3	High Rating	44.0	619.0	7	0	0.0	1	17
10834	11	High Rating	7.0	2.6	6	0	0.0	1	36
10836	11	High Rating	38.0	53.0	8	0	0.0	1	36
10837	11	High Rating	4.0	3.6	5	0	0.0	1	36
10840	18	High Rating	398307.0	19.0	15	0	0.0	1	64

7723 rows × 9 columns

In [74]:

```
dfclean.to_csv("./googleps_cleaned.csv", sep=',',index=False)
```