cleaning

June 24, 2023

[1]:
```
import pandas as pd
import numpy as np
```

Importing CSV File

[2]:
```
path= "C:\\Users\\surwa\\Downloads\\Case+Study+Notebook\\Case Study Notebook\\"
```

[3]:
```
inp0= pd.read_csv(path + "googleplaystore_v2.csv")
```

[4]:
```
inp0.head()
```

[4]:

	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	19000.0	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	14000.0	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	8700.0	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	25000.0	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	2800.0	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

[5]:
```
inp0.shape
```

[5]:
```
(10841, 13)
```

[6]:
```
inp0.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  float64
 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(2), object(11)
memory usage: 1.1+ MB
```

[7]:
```
inp0.isnull().sum()
```

[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
```

Handling missing values for Rating - drop the record

[8]:
```
inp1= inp0[~inp0.Rating.isnull()]
inp1.shape
```

[8]:
```
(9367, 13)
```

[9]:
```
inp1.Rating.isnull().sum()
```

[9]:
```
0
```

[10]:
```
inp1.isnull().sum()
```

[10]:
```
App               0
Category          0
Rating            0
Reviews           0
Size              0
Installs          0
Type              0
Price             0
Content Rating    1
Genres            0
Last Updated      0
Current Ver       4
Android Ver       3
dtype: int64
```

Explore/ Understand the null for the Android Ver Column.

[12]:
```
inp1[inp1['Android Ver'].isnull()]
```

[12]:
```
                                          App         Category  Rating  \
4453                    [substratum] Vacuum: P  PERSONALIZATION     4.4
4490                      Pi Dark [substratum]  PERSONALIZATION     4.5
10472  Life Made WI-Fi Touchscreen Photo Frame              1.9    19.0

      Reviews          Size Installs  Type     Price Content Rating  \
4453      230  11000.000000   1,000+  Paid     $1.49       Everyone
4490      189   2100.000000  10,000+  Free         0       Everyone
10472    3.0M  21516.529524     Free     0  Everyone            NaN

                  Genres    Last Updated Current Ver Android Ver
4453     Personalization   July 20, 2018         4.4         NaN
4490     Personalization  March 27, 2018         1.1         NaN
10472  February 11, 2018          1.0.19  4.0 and up         NaN
```

Dropping

[14]:
```
inp1[(inp1['Android Ver'].isnull() & (inp1.Category== '1.9'))]
```

[14]:
```
                                          App Category  Rating Reviews  \
10472  Life Made WI-Fi Touchscreen Photo Frame      1.9    19.0    3.0M
```

```
                Size Installs Type     Price Content Rating             Genres  \
10472  21516.529524     Free    0  Everyone            NaN  February 11, 2018

      Last Updated Current Ver Android Ver
10472       1.0.19  4.0 and up         NaN
```

[15]:
```
inp1=inp1[~(inp1['Android Ver'].isnull() & (inp1.Category== '1.9'))]
```

[16]:
```
inp1[inp1['Android Ver'].isnull()]
```

[16]:
```
                         App         Category  Rating Reviews     Size  \
4453  [substratum] Vacuum: P  PERSONALIZATION     4.4     230  11000.0
4490    Pi Dark [substratum]  PERSONALIZATION     4.5     189   2100.0

     Installs  Type  Price Content Rating           Genres    Last Updated  \
4453   1,000+  Paid  $1.49       Everyone  Personalization   July 20, 2018
4490  10,000+  Free      0       Everyone  Personalization  March 27, 2018

     Current Ver Android Ver
4453         4.4         NaN
4490         1.1         NaN
```

The most common values in Android Ver

[18]:
```
inp1['Android Ver'].value_counts()
```

[18]:
```
4.1 and up            2059
Varies with device    1319
4.0.3 and up          1240
4.0 and up            1131
4.4 and up             875
2.3 and up             582
5.0 and up             535
4.2 and up             338
2.3.3 and up           240
3.0 and up             211
2.2 and up             208
4.3 and up             207
2.1 and up             113
1.6 and up              87
6.0 and up              48
7.0 and up              41
3.2 and up              31
2.0 and up              27
5.1 and up              18
1.5 and up              16
```

```
3.1 and up              8
2.0.1 and up            7
4.4W and up             6
8.0 and up              5
7.1 and up              3
4.0.3 - 7.1.1           2
5.0 - 8.0               2
1.0 and up              2
7.0 - 7.1.1             1
4.1 - 7.1.1             1
5.0 - 6.0               1
Name: Android Ver, dtype: int64
```

[20]:
```
inp1['Android Ver'].mode()[0]
```

[20]:
```
'4.1 and up'
```

Filling the NaNs with this value

[22]:
```
inp1['Android Ver']= inp1['Android Ver'].fillna(inp1['Android Ver'].mode()[0])
```

```
C:\Users\surwa\AppData\Local\Temp\ipykernel_11916\2642485347.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  inp1['Android Ver']= inp1['Android Ver'].fillna(inp1['Android Ver'].mode()[0])
```

[23]:
```
inp1['Android Ver'].value_counts()
```

[23]:
```
4.1 and up            2061
Varies with device    1319
4.0.3 and up          1240
4.0 and up            1131
4.4 and up             875
2.3 and up             582
5.0 and up             535
4.2 and up             338
2.3.3 and up           240
3.0 and up             211
2.2 and up             208
4.3 and up             207
2.1 and up             113
1.6 and up              87
6.0 and up              48
```

```
7.0 and up             41
3.2 and up             31
2.0 and up             27
5.1 and up             18
1.5 and up             16
3.1 and up              8
2.0.1 and up            7
4.4W and up             6
8.0 and up              5
7.1 and up              3
4.0.3 - 7.1.1           2
5.0 - 8.0               2
1.0 and up              2
7.0 - 7.1.1             1
4.1 - 7.1.1             1
5.0 - 6.0               1
Name: Android Ver, dtype: int64
```

[24]:
```
inp1.isnull().sum()
```

[24]:
```
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       4
Android Ver       0
dtype: int64
```

[26]:
```
inp1[inp1['Current Ver'].isnull()]
```

[26]:
```
                                App             Category  Rating Reviews  \
15    Learn To Draw Kawaii Characters       ART_AND_DESIGN     3.2      55
1553             Market Update Helper  LIBRARIES_AND_DEMO     4.1   20145
6322          Virtual DJ Sound Mixer                TOOLS     4.2    4010
7333                      Dots puzzle               FAMILY     4.0     179

         Size    Installs  Type  Price Content Rating            Genres  \
15     2700.0      5,000+  Free      0       Everyone      Art & Design
1553     11.0  1,000,000+  Free      0       Everyone  Libraries & Demo
6322   8700.0    500,000+  Free      0       Everyone             Tools
```

```
7333  14000.0     50,000+  Paid  $0.99       Everyone            Puzzle

          Last Updated Current Ver Android Ver
15         June 6, 2018         NaN  4.2 and up
1553  February 12, 2013         NaN  1.5 and up
6322       May 10, 2017         NaN  4.0 and up
7333     April 18, 2018         NaN  4.0 and up
```

[27]:
```
inp1['Current Ver'].value_counts()
```

[27]:
```
Varies with device    1415
1.0                    458
1.1                    195
1.2                    126
1.3                    120
                      ...
2.9.10                   1
3.18.5                   1
1.3.A.2.9                1
9.9.1.1910               1
0.3.4                    1
Name: Current Ver, Length: 2638, dtype: int64
```

[29]:
```
inp1['Current Ver'].mode()[0]
```

[29]:
```
'Varies with device'
```

[30]:
```
inp1['Current Ver']= inp1['Current Ver'].fillna(inp1['Current Ver'].mode()[0])
```

```
C:\Users\surwa\AppData\Local\Temp\ipykernel_11916\1796284910.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  inp1['Current Ver']= inp1['Current Ver'].fillna(inp1['Current Ver'].mode()[0])
```

[31]:
```
inp1['Current Ver'].value_counts()
```

[31]:
```
Varies with device    1419
1.0                    458
1.1                    195
1.2                    126
1.3                    120
                      ...
2.9.10                   1
```

```
3.18.5                  1
1.3.A.2.9               1
9.9.1.1910              1
0.3.4                   1
Name: Current Ver, Length: 2638, dtype: int64
```

[33]:
```
inp1.isnull().sum()
```

[33]:
```
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
```

Changing the variables to correct type.

[37]:
```
inp1.dtypes
```

[37]:
```
App                object
Category           object
Rating            float64
Reviews            object
Size              float64
Installs           object
Type               object
Price              object
Content Rating     object
Genres             object
Last Updated       object
Current Ver        object
Android Ver        object
dtype: object
```

[38]:
```
inp1.head()
```

[38]:
```
                                                 App        Category  Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN     4.1
1                                Coloring book moana  ART_AND_DESIGN     3.9
```

```
2  U Launcher Lite - FREE Live Cool Themes, Hide …  ART_AND_DESIGN     4.7
3                              Sketch - Draw & Paint  ART_AND_DESIGN     4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN     4.3

  Reviews     Size     Installs  Type Price Content Rating  \
0     159  19000.0      10,000+  Free     0       Everyone
1     967  14000.0     500,000+  Free     0       Everyone
2   87510   8700.0   5,000,000+  Free     0       Everyone
3  215644  25000.0  50,000,000+  Free     0           Teen
4     967   2800.0     100,000+  Free     0       Everyone

                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

Price Column

[39]:
```
inp1.Price.value_counts()
```

[39]:
```
0          8719
$2.99       114
$0.99       107
$4.99        70
$1.99        59
           …
$1.29          1
$299.99        1
$379.99        1
$37.99         1
$1.20          1
Name: Price, Length: 73, dtype: int64
```

[40]:
```
inp1.Price= inp1.Price.apply(lambda x: 0 if x == '0' else float(x[1:]))
```

```
C:\Users\surwa\AppData\Local\Temp\ipykernel_11916\1313750918.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  inp1.Price= inp1.Price.apply(lambda x: 0 if x == '0' else float(x[1:]))
```

[41]:
```
inp1.Price.dtype
```

[41]:
```
dtype('float64')
```

[44]:
```
inp1.Price.value_counts()
```

[44]:
```
0.00      8719
2.99       114
0.99       107
4.99        70
1.99        59
           …
1.29         1
299.99       1
379.99       1
37.99        1
1.20         1
Name: Price, Length: 73, dtype: int64
```

Handel the Reviews Column

[45]:
```
inp1.Reviews.value_counts()
```

[45]:
```
2         83
3         78
4         74
5         74
1         67
          ..
49657      1
41420      1
7146       1
44706      1
398307     1
Name: Reviews, Length: 5992, dtype: int64
```

[46]:
```
inp1.Reviews= inp1.Reviews.astype('int32')
```

```
C:\Users\surwa\AppData\Local\Temp\ipykernel_11916\1078745672.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  inp1.Reviews= inp1.Reviews.astype('int32')
```

[47]:
```
inp1.Reviews.dtype
```

[47]:
```
dtype('int32')
```

Handle the Installs column

[51]:
```
inp1.Installs.value_counts()
```

[51]:
```
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: Installs, dtype: int64
```

[52]:
```
inp1['Installs'] = inp1['Installs'].str.replace(',', '')
```

```
C:\Users\surwa\AppData\Local\Temp\ipykernel_11916\1155763510.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  inp1['Installs'] = inp1['Installs'].str.replace(',', '')
```

[53]:
```
inp1.Installs.value_counts()
```

[53]:
```
1000000       1577
10000000      1252
```

```
100000        1150
10000         1010
5000000        752
1000           713
500000         538
50000          467
5000           432
100000000      409
100            309
50000000       289
500            201
500000000       72
10               69
1000000000       58
50               56
5                 9
1                 3
Name: Installs, dtype: int64
```

[54]:
```
inp1.isnull().sum()
```

[54]:
```
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
```