535230080-praktikum07

April 19, 2024

[19]:

```
#Georgia Sugisandhea_535230080_Kelas C

import pandas as pd #Mengimport library pandas dengan variable pd
import numpy as np #Mengimport library numpy dengan variable np
```

[20]:

```
file = "googleplaystore.csv" #Memasukkan nama file yang akan dibaca
df = pd.read_csv(file) #Membaca file dengan library pandas
```

[21]:

```
#Mencetak 10 baris pertama dari dataframe
df.head(10)

#Solution 2
#df[0:10]

#Solution 3
#df.iloc[0:10]
```

[21]:

```
                                                 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
5                         Paper flowers instructions  ART_AND_DESIGN     4.4
6            Smoke Effect Photo Maker - Smoke Editor  ART_AND_DESIGN     3.8
7                                   Infinite Painter  ART_AND_DESIGN     4.1
8                               Garden Coloring Book  ART_AND_DESIGN     4.4
9                      Kids Paint Free - Drawing Fun  ART_AND_DESIGN     4.7

   Reviews  Size     Installs  Type Price Content Rating  \
0      159   19M      10,000+  Free     0       Everyone
1      967   14M     500,000+  Free     0       Everyone
2    87510  8.7M   5,000,000+  Free     0       Everyone
3   215644   25M  50,000,000+  Free     0           Teen
4      967  2.8M     100,000+  Free     0       Everyone
5      167  5.6M      50,000+  Free     0       Everyone
6      178   19M      50,000+  Free     0       Everyone
```

```
7    36815   29M   1,000,000+  Free     0       Everyone
8    13791   33M   1,000,000+  Free     0       Everyone
9      121  3.1M      10,000+  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
5                 Art & Design      March 26, 2017                   1
6                 Art & Design      April 26, 2018                 1.1
7                 Art & Design       June 14, 2018            6.1.61.1
8                 Art & Design  September 20, 2017               2.9.2
9     Art & Design;Creativity        July 3, 2018                 2.8

    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
5    2.3 and up
6  4.0.3 and up
7    4.2 and up
8    3.0 and up
9  4.0.3 and up
```

[22]:

```
#Mencetak 3 baris terakhir dari dataframe
#df.tail(3)

#Solution 2
#df[-3:]

#Solution 3
df.iloc[-3:]
```

[22]:

```
                                                 App           Category  \
10836                         Parkinson Exercices FR            MEDICAL
10837                  The SCP Foundation DB fr nn5n  BOOKS_AND_REFERENCE
10838  iHoroscope - 2018 Daily Horoscope & Astrology          LIFESTYLE

       Rating  Reviews                Size     Installs  Type Price  \
10836     NaN        3                9.5M       1,000+  Free     0
10837     4.5      114  Varies with device       1,000+  Free     0
10838     4.5   398307                 19M  10,000,000+  Free     0
```

```
      Content Rating             Genres      Last Updated         Current Ver  \
10836       Everyone            Medical  January 20, 2017                   1
10837     Mature 17+  Books & Reference  January 19, 2015  Varies with device
10838       Everyone          Lifestyle     July 25, 2018  Varies with device

              Android Ver
10836          2.2 and up
10837  Varies with device
10838  Varies with device
```

[23]:
```
#Mencetak info dari tabel yang sudah dibaca
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10839 entries, 0 to 10838
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10839 non-null  object
 1   Category        10839 non-null  object
 2   Rating          9365 non-null   float64
 3   Reviews         10839 non-null  int64
 4   Size            10839 non-null  object
 5   Installs        10839 non-null  object
 6   Type            10838 non-null  object
 7   Price           10839 non-null  object
 8   Content Rating  10839 non-null  object
 9   Genres          10839 non-null  object
 10  Last Updated    10839 non-null  object
 11  Current Ver     10831 non-null  object
 12  Android Ver     10837 non-null  object
dtypes: float64(1), int64(1), object(11)
memory usage: 1.1+ MB
```

[24]:
```
#mencetak jumlah baris dan kolom dari tabel yang dipakai
df.shape
```

[24]:
```
(10839, 13)
```

[25]:
```
#mencetak nama nama kolom dalam tabel tersebut
df.columns
```

[25]:
```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver',
       'Android Ver'],
      dtype='object')
```

[26]:
```
#Mengubah nama kolom yang disebutkan menjadi nama lain yang dimasukkan
df.rename(index=str, columns={"Content Rating":"ContentRating", "Last Updated":
 ↪"LastUpdated", "Current Ver":"CurrentVersion", "Android Ver":
 ↪"AndroidVersion"}, inplace=True)

#Mencetak nama nama dalam tabel tersebut
df.columns
```

[26]:
```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type',
       'Price', 'ContentRating', 'Genres', 'LastUpdated', 'CurrentVersion',
       'AndroidVersion'],
      dtype='object')
```

[27]:
```
#Menghitung dan mencetak baris baris yang duplikat
duplicatesNum = df.duplicated().sum()
print("There are %d duplicate records"%(duplicatesNum))
```

```
There are 483 duplicate records
```

[28]:
```
#Mencari data data yang ada di df yang nama aplikasinya mengandung nama blood␣
 ↪pressure (terlepas dari upper case atau lower case (case=False))
duplicates = df[(df.duplicated())]
duplicatedWithBloodPressure = duplicates[duplicates['App'].str.contains('Blood␣
 ↪pressure', case=False)]

#Mencetak jumlah data yang ditemukan
print("%d duplicated records have 'blood pressure' in their name"␣
 ↪%(duplicatedWithBloodPressure.shape[0]))

print("%d duplicated records have 'blood perssure' in their name"␣
 ↪%(len(duplicatedWithBloodPressure)))
```

```
9 duplicated records have 'blood pressure' in their name
9 duplicated records have 'blood perssure' in their name
```

[29]:
```
#Mencetak baris baris data yang telah dikumpulkan sebelumnya
duplicatedWithBloodPressure
```

[29]:
```
                                  App Category  Rating  Reviews  Size  \
2502               Free Blood Pressure  MEDICAL     NaN        7  5.7M
2512          JH Blood Pressure Monitor  MEDICAL     3.7        9  2.9M
6584                 iBP Blood Pressure  MEDICAL     4.4      578  704k
6585                     Blood Pressure  MEDICAL     4.2    33033  7.4M
6590           Blood Pressure(BP) Diary  MEDICAL     3.7     3596  8.4M
6593  BP Journal - Blood Pressure Diary  MEDICAL     5.0        6   26M
6594             Blood Pressure Monitor  MEDICAL     4.3       17  6.0M
6595           Blood Pressure Companion  MEDICAL     4.2      178  4.8M
```

```
6598                Free Blood Pressure  MEDICAL     NaN        7  5.7M

        Installs  Type  Price ContentRating   Genres        LastUpdated  \
2502      5,000+  Free      0      Everyone  Medical   October 13, 2016
2512        500+  Free      0      Everyone  Medical      July 21, 2018
6584     10,000+  Paid  $0.99      Everyone  Medical  November 30, 2014
6585  5,000,000+  Free      0      Everyone  Medical      July 24, 2018
6590  1,000,000+  Free      0      Everyone  Medical      July 18, 2018
6593      1,000+  Free      0      Everyone  Medical       May 25, 2018
6594     10,000+  Free      0      Everyone  Medical  February 25, 2017
6595      1,000+  Paid  $0.99      Everyone  Medical      July 22, 2018
6598      5,000+  Free      0      Everyone  Medical   October 13, 2016

        CurrentVersion AndroidVersion
2502             3.0.0   4.0.3 and up
2512               1.1   4.0.3 and up
6584             7.0.1     2.2 and up
6585            3.27.3     4.1 and up
6590             4.0.9   4.0.3 and up
6593            1.0.32     4.4 and up
6594             1.0.1     4.4 and up
6595  4.1.5 (Steglitz)     4.1 and up
6598             3.0.0   4.0.3 and up
```

[30]:

```python
#Mencari jumlah data yang memiliki nama yang duplikat
duplicatedAppsSum = df.duplicated(subset='App').sum()

#Mencetak data yang ditemukan
print("In fact, there are %d duplicate records" %(duplicatedAppsSum))
```

```
In fact, there are 1181 duplicate records
```

[31]:

```python
#Memasukkan data data yang memiliki nama yang duplikat
duplicatedApps = df[df.duplicated(subset='App')]
```

[32]:

```python
#Mencari data data dengan kolom App yang duplikat
duplicatedAppsSummary = duplicatedApps['App']. value_counts()

#Mengambil baris data teratas dan juga jumlah dari data yang terduplikat
print("The app %s has %d duplicated records" %(duplicatedAppsSummary.index[0],␣
 ↪duplicatedAppsSummary[0]))
```

```
The app ROBLOX has 8 duplicated records

C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\897775467.py:5:
FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a
future version, integer keys will always be treated as labels (consistent with
DataFrame behavior). To access a value by position, use `ser.iloc[pos]`
```

```
  print("The app %s has %d duplicated records" %(duplicatedAppsSummary.index[0],
duplicatedAppsSummary[0]))
```

[33]:

```
#Memasukkan data data yang duplikat dari df kedalam variable duplicatedApps
duplicatedApps=df[df.duplicated(subset='App')]
#Membuat array dupsPerApp
dupsPerApp=[]

#Membuat array dengan nama di kolom App yang unik
duplicatedAppsNames = duplicatedApps['App'].unique()

#Menghitung jumlah duplikasi dari masing masing data duplikat yang ada
for u in duplicatedAppsNames:
    dupsPerApp.append(len(duplicatedApps[duplicatedApps['App']==u]))

#Mencari data dengan duplikasi paling tinggi
dupsPerAppArray = np.array(dupsPerApp)
maxDupRecords = np.max(dupsPerAppArray)
maxDupRecordsApp = duplicatedAppsNames[dupsPerAppArray.argmax()]

#Mencetak nama dan jumlah hasil pencarian duplikasi data terbanyak
print("The app %s has %d duplicated records" %(maxDupRecordsApp, maxDupRecords))
```

```
The app ROBLOX has 8 duplicated records
```

[34]:

```
#Mencari data data yang memiliki nama sesuai dengan variable maxDupRecordsApp␣
 ↪yang telah kita cari sebelumnya (data dengan duplikasi tertinggi)
duplicatedApps[duplicatedApps['App']==maxDupRecordsApp]
```

[34]:

```
         App Category  Rating  Reviews Size      Installs  Type Price  \
1701  ROBLOX     GAME     4.5  4447346  67M  100,000,000+  Free     0
1748  ROBLOX     GAME     4.5  4448791  67M  100,000,000+  Free     0
1841  ROBLOX     GAME     4.5  4449882  67M  100,000,000+  Free     0
1870  ROBLOX     GAME     4.5  4449910  67M  100,000,000+  Free     0
2016  ROBLOX   FAMILY     4.5  4449910  67M  100,000,000+  Free     0
2088  ROBLOX   FAMILY     4.5  4450855  67M  100,000,000+  Free     0
2206  ROBLOX   FAMILY     4.5  4450890  67M  100,000,000+  Free     0
4527  ROBLOX   FAMILY     4.5  4443407  67M  100,000,000+  Free     0

     ContentRating                        Genres    LastUpdated  \
1701  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
1748  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
1841  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
1870  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
2016  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
2088  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
2206  Everyone 10+  Adventure;Action & Adventure  July 31, 2018
```

```
4527  Everyone 10+  Adventure;Action & Adventure  July 31, 2018

     CurrentVersion AndroidVersion
1701   2.347.225742     4.1 and up
1748   2.347.225742     4.1 and up
1841   2.347.225742     4.1 and up
1870   2.347.225742     4.1 and up
2016   2.347.225742     4.1 and up
2088   2.347.225742     4.1 and up
2206   2.347.225742     4.1 and up
4527   2.347.225742     4.1 and up
```

[35]:

```
#Memasukkan data nomor 2976 dan 3007 ked1
d1 = df[2976:2977]._append(df[3007:3008])

#Memasukkan data 3015 dan 3020 ke d2
d2 = df[3015:3016]._append(df[3020:3021])

#Memasukkan data d2 ke d1
d1._append(d2)
```

[35]:

```
                                                 App Category  Rating  \
2976  CBS Sports App - Scores, News, Stats & Watch Live   SPORTS     4.3
3007  CBS Sports App - Scores, News, Stats & Watch Live   SPORTS     4.3
3015  CBS Sports App - Scores, News, Stats & Watch Live   SPORTS     4.3
3020  CBS Sports App - Scores, News, Stats & Watch Live   SPORTS     4.3

      Reviews                Size    Installs  Type Price ContentRating  \
2976    91031  Varies with device  5,000,000+  Free     0      Everyone
3007    91031  Varies with device  5,000,000+  Free     0      Everyone
3015    91031  Varies with device  5,000,000+  Free     0      Everyone
3020    91031  Varies with device  5,000,000+  Free     0      Everyone

      Genres     LastUpdated      CurrentVersion AndroidVersion
2976  Sports  August 4, 2018  Varies with device     5.0 and up
3007  Sports  August 4, 2018  Varies with device     5.0 and up
3015  Sports  August 4, 2018  Varies with device     5.0 and up
3020  Sports  August 4, 2018  Varies with device     5.0 and up
```

[36]:

```
#Menggabungkan baris 3020 dengan 3056
df[3020:3021]._append(df[3056:3057])
```

[36]:

```
                                                 App Category  Rating  \
3020  CBS Sports App - Scores, News, Stats & Watch Live   SPORTS     4.3
3056  CBS Sports App - Scores, News, Stats & Watch Live   SPORTS     4.3

      Reviews                Size    Installs  Type Price ContentRating  \
3020    91031  Varies with device  5,000,000+  Free     0      Everyone
3056    91033  Varies with device  5,000,000+  Free     0      Everyone

      Genres      LastUpdated      CurrentVersion AndroidVersion
3020  Sports  August 4, 2018  Varies with device     5.0 and up
3056  Sports  August 4, 2018  Varies with device     5.0 and up
```

[37]:

```
#Menghapus baris baris data yang memiliki nama kolom App sama dan menyimpan␣
  ↪hanya data yang terakhir
dfClean = df.drop_duplicates(subset='App', keep='last')
#Mencetak ukuran kolom dan baris hasil drop
dfClean.shape
```

[37]:

```
(9658, 13)
```

[38]:

```
#Memperbaiki format tanggal kolom LastUpdated di tabel dfClean
#dfClean.loc[:, 'LastUpdated']

dfClean.loc[:, 'LastUpdated'] = pd.to_datetime(dfClean['LastUpdated'],␣
  ↪format='%B %d, %Y')
```

[39]:

```
#Mencari aplikasi yang paling terakhir di update menggunakan sort values dari␣
  ↪kolom LastUpdated
sortedLastUpdated = dfClean['LastUpdated'].sort_values()

print("%s is the least recently updated application and was updated on %s"␣
  ↪%(dfClean.loc[sortedLastUpdated.index[0]]['App'], str(sortedLastUpdated[0])))
```

```
CJ Poker Odds Calculator is the least recently updated application and was
updated on 2011-01-30 00:00:00
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\465773211.py:4:
FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a
future version, integer keys will always be treated as labels (consistent with
DataFrame behavior). To access a value by position, use `ser.iloc[pos]`
  print("%s is the least recently updated application and was updated on %s"
%(dfClean.loc[sortedLastUpdated.index[0]]['App'], str(sortedLastUpdated[0])))
```

[40]:

```
#Mencari jumlah angka harga yang unik
uniquePricesNum = dfClean['Price'].nunique()

#Mencari angka angka harga yang unik dan memasukkannya ke variable baru
uniquePrices = dfClean['Price'].unique()

#Mencetak hasil pencarian
print("There are %d unique price values. The first five are %s"␣
  ↪%(uniquePricesNum, uniquePrices[:5]))
```

```
print("There are %d unique price values. The first five are %s"␣
 ↪%(len(uniquePrices), uniquePrices[:5]))
```

```
There are 92 unique price values. The first five are ['0' '$4.99' '$3.99'
'$1.49' '$2.99']
There are 92 unique price values. The first five are ['0' '$4.99' '$3.99'
'$1.49' '$2.99']
```

[41]:

```
#Membuat fungsi untuk mengganti value kolom harga menjadi float tanpa simbol $
def moneyWithoutCurrencySymbol(v):
    if type(v) is not float:
        return float(v.replace("$", ""))
    else:
        return v
```

[42]:

```
#Memakai fungsi yang telah dibuat sebelumnya pada kolom Price di tabel dfClean␣
 ↪dan mencetaknya
dfClean.loc[:, 'Price']=dfClean['Price'].apply(moneyWithoutCurrencySymbol)
dfClean
```

[42]:

```
                                                     App            Category  \
0          Photo Editor & Candy Camera & Grid & ScrapBook        ART_AND_DESIGN
2      U Launcher Lite – FREE Live Cool Themes, Hide …        ART_AND_DESIGN
3                                    Sketch - Draw & Paint        ART_AND_DESIGN
4                    Pixel Draw - Number Art Coloring Book        ART_AND_DESIGN
5                               Paper flowers instructions        ART_AND_DESIGN
…                                                       …                   …
10834                                    Sya9a Maroc - FR              FAMILY
10835                     Fr. Mike Schmitz Audio Teachings              FAMILY
10836                               Parkinson Exercices FR             MEDICAL
10837                        The SCP Foundation DB fr nn5n  BOOKS_AND_REFERENCE
10838        iHoroscope - 2018 Daily Horoscope & Astrology           LIFESTYLE

       Rating  Reviews                Size     Installs  Type Price  \
0         4.1      159                 19M      10,000+  Free   0.0
2         4.7    87510                8.7M   5,000,000+  Free   0.0
3         4.5   215644                 25M  50,000,000+  Free   0.0
4         4.3      967                2.8M     100,000+  Free   0.0
5         4.4      167                5.6M      50,000+  Free   0.0
…         …       …                   …           …     …    …
10834     4.5       38                 53M       5,000+  Free   0.0
10835     5.0        4                3.6M         100+  Free   0.0
10836     NaN        3                9.5M       1,000+  Free   0.0
10837     4.5      114  Varies with device       1,000+  Free   0.0
10838     4.5   398307                 19M  10,000,000+  Free   0.0

      ContentRating                   Genres          LastUpdated  \
0          Everyone             Art & Design  2018-01-07 00:00:00
2          Everyone             Art & Design  2018-08-01 00:00:00
3              Teen             Art & Design  2018-06-08 00:00:00
4          Everyone  Art & Design;Creativity  2018-06-20 00:00:00
5          Everyone             Art & Design  2017-03-26 00:00:00
…             …                      …                  …
10834      Everyone                Education  2017-07-25 00:00:00
10835      Everyone                Education  2018-07-06 00:00:00
10836      Everyone                  Medical  2017-01-20 00:00:00
10837    Mature 17+        Books & Reference  2015-01-19 00:00:00
10838      Everyone                Lifestyle  2018-07-25 00:00:00

           CurrentVersion      AndroidVersion
0                   1.0.0        4.0.3 and up
2                   1.2.4        4.0.3 and up
3      Varies with device          4.2 and up
4                     1.1          4.4 and up
5                       1          2.3 and up
…                   …                 …
10834                1.48          4.1 and up
10835                   1          4.1 and up
10836                   1          2.2 and up
10837  Varies with device  Varies with device
10838  Varies with device  Varies with device

[9658 rows x 13 columns]
```

[43]:

```
#Solution 2 untuk menghapus simbol $ dari kolom Price
#dfClean["Price"] = pd.to_numeric(dfClean["Price"].str.strip("$"))
```

[44]:

```
#Menghitung jumlah baris kolom yang Not a Number (NaN) dari masing masing kolom
dfClean.isna().sum()
```

[44]:

```
App                  0
Category             0
Rating            1463
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
ContentRating        0
Genres               0
LastUpdated          0
CurrentVersion       8
AndroidVersion       2
dtype: int64
```

[45]:

```
#Menghitung jumlah baris yang isinya null dari masing masing kolom
dfClean.isnull().sum()
```

[45]:

```
App                  0
Category             0
Rating            1463
Reviews              0
Size                 0
Installs             0
Type                 1
Price                0
ContentRating        0
Genres               0
LastUpdated          0
CurrentVersion       8
AndroidVersion       2
dtype: int64
```

[46]:

```
#Mengganti data data yang hilang, yaitu NaN di kolom Type menjadi kolom tersebut

#dfClean[dfClean['Type'].isna()]
dfClean['Type'] = dfClean['Type'].fillna(dfClean['Type'].mode()[0])
#dfClean[dfClean['App] == 'Command & Conquer: Rivals']
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\3239869390.py:4:
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
  dfClean['Type'] = dfClean['Type'].fillna(dfClean['Type'].mode()[0])
```

[47]:

```
#Mengganti data data yang hilang, yaitu NaN di kolom AndroidVersion menjadi␣
  ↪kolom tersebut
dfClean['AndroidVersion'] = dfClean['AndroidVersion'].
  ↪fillna(dfClean['AndroidVersion'].mode()[0])
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\3334635015.py:2:
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
  dfClean['AndroidVersion'] =
dfClean['AndroidVersion'].fillna(dfClean['AndroidVersion'].mode()[0])
```

[48]:

```
#Mengganti data data yang hilang, yaitu NaN di kolom CurrentVersion menjadi␣
 ↪kolom tersebut
dfClean['CurrentVersion'] = dfClean['CurrentVersion'].
 ↪fillna(dfClean['CurrentVersion'].mode()[0])
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\727592784.py:2:
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
  dfClean['CurrentVersion'] =
dfClean['CurrentVersion'].fillna(dfClean['CurrentVersion'].mode()[0])
```

[49]:

```
#Solution 2 dari mengganti data data yang hilang, yaitu NaN di kolom␣
 ↪AndroidVersion menjadi kolom tersebut

#dfClean[dfClean['AndroidVersion'].isna()]
#dfClean['AndroidVersion'].mode()
dfClean.loc[dfClean['AndroidVersion'].isna(), 'AndroidVersion'] =␣
 ↪dfClean['AndroidVersion'].mode()[0]
#dfClean[dfClean['App'] == '[substratum] Vacuum: P']
#dfClean[dfClean['App'] == 'Pi Dark [substratum]']
```

[50]:

```
#Solution 2 dari mengganti data data yang hilang, yaitu NaN di kolom Type,␣
 ↪AndroidVersion, dan CurrentVersion menjadi kolom tersebut
dfClean.loc[dfClean['Type'].isna(), 'Type'] = dfClean['Type'].mode()[0]

dfClean.loc[dfClean['AndroidVersion'].isna(), 'AndroidVersion'] =␣
 ↪dfClean['AndroidVersion'].mode()[0]

dfClean.loc[dfClean['CurrentVersion'].isna(), 'CurrentVersion'] =␣
 ↪dfClean['CurrentVersion'].mode()[0]
```

[51]:

```
#Mengganti data data yang hilang, yaitu NaN di kolom Rating menjadi kolom␣
 ↪tersebut

#dfClean[dfClean['Rating].isna()]
#Solution1
dfClean['Rating'] = dfClean['Rating'].fillna(0)
dfClean[dfClean['Rating']==0]

#Solution2
#dfClean.loc[dfClean['Rating].isna(), 'Rating'] = 0
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\2425662449.py:5:
```

```
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
  dfClean['Rating'] = dfClean['Rating'].fillna(0)
```

[51]:

```
                                       App             Category  Rating  \
113              Wrinkles and rejuvenation               BEAUTY     0.0
123                 Manicure - nail design               BEAUTY     0.0
126           Skin Care and Natural Beauty               BEAUTY     0.0
129    Secrets of beauty, youth and health               BEAUTY     0.0
130     Recipes and tips for losing weight               BEAUTY     0.0
…                                    …                  …     …
10822                            Cardio-FR              MEDICAL     0.0
10823                   Naruto & Boruto FR               SOCIAL     0.0
10829             payermonstationnement.fr  MAPS_AND_NAVIGATION     0.0
10833                             FR Forms             BUSINESS     0.0
10836               Parkinson Exercices FR              MEDICAL     0.0

       Reviews  Size  Installs  Type Price ContentRating             Genres  \
113        182  5.7M  100,000+  Free   0.0  Everyone 10+             Beauty
123        119  3.7M   50,000+  Free   0.0      Everyone             Beauty
126        654  7.4M  100,000+  Free   0.0          Teen             Beauty
129         77  2.9M   10,000+  Free   0.0    Mature 17+             Beauty
130         35  3.1M   10,000+  Free   0.0  Everyone 10+             Beauty
…        …   …       …   …   …           …                …
10822       67   82M   10,000+  Free   0.0      Everyone            Medical
10823        7  7.7M      100+  Free   0.0          Teen             Social
10829       38  9.8M    5,000+  Free   0.0      Everyone  Maps & Navigation
10833        0  9.6M       10+  Free   0.0      Everyone           Business
10836        3  9.5M    1,000+  Free   0.0      Everyone            Medical

               LastUpdated CurrentVersion AndroidVersion
113    2017-09-20 00:00:00              8     3.0 and up
123    2018-07-23 00:00:00            1.3     4.1 and up
126    2018-07-17 00:00:00           1.15     4.1 and up
129    2017-08-08 00:00:00              2     2.3 and up
130    2017-12-11 00:00:00              2     3.0 and up
…                    …            …            …
10822  2018-07-31 00:00:00          2.2.2     4.4 and up
10823  2018-02-02 00:00:00              1     4.0 and up
10829  2018-06-13 00:00:00      2.0.148.0     4.0 and up
10833  2016-09-29 00:00:00          1.1.5     4.0 and up
10836  2017-01-20 00:00:00              1     2.2 and up
```

```
[1463 rows x 13 columns]
```

```
[52]: #Mencetak angka unik dari kolom Installs
      dfClean['Installs'].unique()
```

```
[52]: array(['10,000+', '5,000,000+', '50,000,000+', '100,000+', '50,000+',
             '1,000,000+', '10,000,000+', '5,000+', '500,000+',
             '1,000,000,000+', '100,000,000+', '1,000+', '50+', '100+', '500+',
             '10+', '1+', '5+', '500,000,000+', '0+', '0'], dtype=object)
```

```
[53]: #Mengganti nilai kolom installs yang 0 menjadi 0+
      dfClean.loc[dfClean["Installs"]=="0","Installs"] = "0+"
```

```
[54]: #Mengecek apakah masih ada baris yang nilai Installs nya berupa 0
      dfClean[(dfClean['Installs']=="0")]
```

```
[54]: Empty DataFrame
      Columns: [App, Category, Rating, Reviews, Size, Installs, Type, Price,
      ContentRating, Genres, LastUpdated, CurrentVersion, AndroidVersion]
      Index: []
```

```
[55]: #Membuat fungsi yang mengubah nilai dalam variable yang dimasukkan menjadi␣
        ↪float dan menghilangkan M atau k
      def sizeToFloat (v):
          if "M" in v:
              return float(v.strip("M"))
          elif v[-1] =="k":
              return float(v.strip("k"))/1024
          else:
              return 1.0

      #Menggunakan fungsi tersebut dalam kolom Size di table dfClean
      dfClean['Size'] = dfClean['Size'].apply(sizeToFloat)
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\2873227039.py:11:
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
  dfClean['Size'] = dfClean['Size'].apply(sizeToFloat)
```

```
[56]: #Mengurutkan nilai nilai pada kolom Size
      sortedSize = dfClean['Size'].sort_values()
```

```
#Mencari aplikasi dengan ukuran paling kecil dari hasil pengurutan yang␣
 ↪dilaksanakan
print("%s is the smallest application and its size is %.4fM" % (dfClean.
 ↪loc[sortedSize.index[0]]['App'], sortedSize[0]))
```

```
Essential Resources is the smallest application and its size is 0.0083M
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\2355141338.py:5:
FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a
future version, integer keys will always be treated as labels (consistent with
DataFrame behavior). To access a value by position, use `ser.iloc[pos]`
  print("%s is the smallest application and its size is %.4fM" %
(dfClean.loc[sortedSize.index[0]]['App'], sortedSize[0]))
```

[57]:

```
#Mengambil aplikasi dengan ukuran terbesar dari pengurutan Size yang sudah kita␣
 ↪laksanakan
largestApps = dfClean.loc[sortedSize[sortedSize==sortedSize[-1]].index]

#Mengambil nama data Category yang paling besar
largestAppsCategory = largestApps['Category'].mode()[0]

#Mengambil baris baris data yang termasuk dalam Category yang paling besar
largestAppsCategoryNames =␣
 ↪largestApps[largestApps['Category']==largestAppsCategory]['App']

#Mencetak hasilnya
print("The most common category of the largest applications is␣
 ↪",largestAppsCategory)

print("The applications that belong to this category are \n",␣
 ↪largestAppsCategoryNames.values)
```

```
The most common category of the largest applications is  GAME
The applications that belong to this category are
 ['Car Crash III Beam DH Real Damage Simulator 2018'
 'Mini Golf King - Multiplayer Game' 'The Walking Dead: Our World'
 'Miami crime simulator' 'Stickman Legends: Shadow Wars'
 'Hungry Shark Evolution']
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\956566684.py:2:
FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a
future version, integer keys will always be treated as labels (consistent with
DataFrame behavior). To access a value by position, use `ser.iloc[pos]`
  largestApps = dfClean.loc[sortedSize[sortedSize==sortedSize[-1]].index]
```

[58]:

```
#Mengecek apakah ada aplikasi yang memiliki Type free tapi memiliki Price yang␣
 ↪tidak 0
free = dfClean[(dfClean['Price']>0) & (dfClean['Type']=="Free")]
```

```
#Mengecek apakah ada aplikasi yang memiliki Type paid tapi memiliki Price yang 0
paid = dfClean[(dfClean['Price']<=0) & (dfClean['Type']=="Paid")]

#Mencetak hasilnya
print("There are %d Free applications for which you have to pay" %(len(free)))
print("There are %d Paid applications which are given for free" %(len(paid)))
```

```
There are 0 Free applications for which you have to pay
There are 0 Paid applications which are given for free
```

[59]:
```
#Mengecek jika ada aplikasi yang tidak mempunyai review apapun namun punya␣
 ↪rating lebih dari 0
dfClean[(dfClean['Reviews']==0) & (dfClean['Rating']>0)]
```

[59]:
```
Empty DataFrame
Columns: [App, Category, Rating, Reviews, Size, Installs, Type, Price,
ContentRating, Genres, LastUpdated, CurrentVersion, AndroidVersion]
Index: []
```

[60]:
```
#Mengambil statistik deskriptif untuk beberapa kolom yang ada, seperti count,␣
 ↪mean (rata rata), standar deviasi, dll
dfClean.describe()
```

[60]:
```
            Rating       Reviews         Size
count  9658.000000  9.658000e+03  9658.000000
mean      3.541054  2.166735e+05    17.935304
std       1.575689  1.830831e+06    21.393800
min       0.000000  0.000000e+00     0.008301
25%       3.600000  2.500000e+01     2.900000
50%       4.200000  9.680000e+02     9.100000
75%       4.500000  2.940800e+04    25.000000
max       5.000000  7.812821e+07   100.000000
```

[61]:
```
from datetime import datetime,date #Mengimport datetime,date dari library␣
 ↪datetime

#Menghitung jangka waktu terakhir update sampai tanggal sekarang sebagai␣
 ↪NotUpdatedFor dan menambahkan jadi kolom baru
dfClean["NotUpdatedFor"] = pd.to_datetime(datetime.today().
 ↪strftime("%m-%d-%Y")) - dfClean["LastUpdated"]

#Mencetak 5 baris pertama dari tabel dfClean
dfClean.head()
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_38292\1545231461.py:4:
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
  dfClean["NotUpdatedFor"] =
pd.to_datetime(datetime.today().strftime("%m-%d-%Y")) - dfClean["LastUpdated"]
```

[61]:

```
                                                 App        Category  Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN     4.1
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
5                         Paper flowers instructions  ART_AND_DESIGN     4.4

   Reviews  Size     Installs  Type Price ContentRating  \
0      159  19.0      10,000+  Free   0.0      Everyone
2    87510   8.7   5,000,000+  Free   0.0      Everyone
3   215644  25.0  50,000,000+  Free   0.0          Teen
4      967   2.8     100,000+  Free   0.0      Everyone
5      167   5.6      50,000+  Free   0.0      Everyone

                    Genres          LastUpdated      CurrentVersion  \
0             Art & Design  2018-01-07 00:00:00               1.0.0
2             Art & Design  2018-08-01 00:00:00               1.2.4
3             Art & Design  2018-06-08 00:00:00  Varies with device
4  Art & Design;Creativity  2018-06-20 00:00:00                 1.1
5             Art & Design  2017-03-26 00:00:00                   1

  AndroidVersion       NotUpdatedFor
0   4.0.3 and up  2294 days 00:00:00
2   4.0.3 and up  2088 days 00:00:00
3     4.2 and up  2142 days 00:00:00
4     4.4 and up  2130 days 00:00:00
5     2.3 and up  2581 days 00:00:00
```