In [8]:
```
import gc
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
import pandas as pd
```

In [9]:
```
pd.set_option('display.max_columns',None)
pd.set_option('display.max_rows',None)
```

In [4]:
```
df1 = pd.read_csv('public_150k_plus_220703.csv')
```

In [5]:
```
df2 = pd.read_csv('public_up_to_150k_12_220703.csv')
```

In [6]:
```
df = pd.concat([df1,df2])
```

In [7]:
```
df.head()
```

Out[7]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress	Bor
0	9547507704	05/01/2020	464	PPP	SUMTER COATINGS, INC.	2410 Highway 15 South	
1	9777677704	05/01/2020	464	PPP	PLEASANT PLACES, INC.	7684 Southrail Road	(
2	5791407702	05/01/2020	1013	PPP	BOYER CHILDREN'S CLINIC	1850 BOYER AVE E	
3	6223567700	05/01/2020	920	PPP	KIRTLEY CONSTRUCTION INC	1661 MARTIN RANCH RD	BER
4	9662437702	05/01/2020	101	PPP	AERO BOX LLC	NaN	

In [8]:
```
df.shape
```

Out[8]:
```
(1568411, 53)
```

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1568411 entries, 0 to 599878
Data columns (total 53 columns):
 #   Column                         Non-Null Count    Dtype  
---  ------                         --------------    -----  
 0   LoanNumber                     1568411 non-null  int64  
 1   DateApproved                   1568411 non-null  object 
 2   SBAOfficeCode                  1568411 non-null  int64  
 3   ProcessingMethod               1568411 non-null  object 
 4   BorrowerName                   1568399 non-null  object 
 5   BorrowerAddress                1568391 non-null  object 
 6   BorrowerCity                   1568394 non-null  object 
 7   BorrowerState                  1568398 non-null  object 
 8   BorrowerZip                    1568393 non-null  object 
 9   LoanStatusDate                 1446434 non-null  object 
 10  LoanStatus                     1568411 non-null  object 
 11  Term                           1568411 non-null  int64  
 12  SBAGuarantyPercentage          1568411 non-null  int64  
 13  InitialApprovalAmount          1568411 non-null  float64
 14  CurrentApprovalAmount          1568411 non-null  float64
 15  UndisbursedAmount              1568171 non-null  float64
 16  FranchiseName                  43639 non-null    object 
 17  ServicingLenderLocationID      1568411 non-null  int64  
 18  ServicingLenderName            1568411 non-null  object 
 19  ServicingLenderAddress         1568411 non-null  object 
 20  ServicingLenderCity            1568411 non-null  object 
 21  ServicingLenderState           1568411 non-null  object 
 22  ServicingLenderZip             1568411 non-null  object 
 23  RuralUrbanIndicator            1568411 non-null  object 
 24  HubzoneIndicator               1568411 non-null  object 
 25  LMIIndicator                   1568411 non-null  object 
 26  BusinessAgeDescription         1568410 non-null  object 
 27  ProjectCity                    1568393 non-null  object 
 28  ProjectCountyName              1568328 non-null  object 
 29  ProjectState                   1568402 non-null  object 
 30  ProjectZip                     1568392 non-null  object 
 31  CD                             1568346 non-null  object 
 32  JobsReported                   1568409 non-null  float64
 33  NAICSCode                      1556478 non-null  float64
 34  Race                           1568411 non-null  object 
 35  Ethnicity                      1568411 non-null  object 
 36  UTILITIES_PROCEED              511409 non-null   float64
 37  PAYROLL_PROCEED                1565948 non-null  float64
 38  MORTGAGE_INTEREST_PROCEED      61004 non-null    float64
 39  RENT_PROCEED                   121152 non-null   float64
 40  REFINANCE_EIDL_PROCEED         26476 non-null    float64
 41  HEALTH_CARE_PROCEED            63363 non-null    float64
 42  DEBT_INTEREST_PROCEED          35514 non-null    float64
 43  BusinessType                   1567626 non-null  object 
 44  OriginatingLenderLocationID    1568411 non-null  int64  
 45  OriginatingLender              1568411 non-null  object 
 46  OriginatingLenderCity          1568411 non-null  object 
 47  OriginatingLenderState         1568411 non-null  object 
 48  Gender                         1568411 non-null  object 
 49  Veteran                        1568411 non-null  object 
 50  NonProfit                      77401 non-null    object 
 51  ForgivenessAmount              1470859 non-null  float64
 52  ForgivenessDate                1470859 non-null  object 
dtypes: float64(13), int64(6), object(34)
memory usage: 646.2+ MB
```

In [10]:
```
df['DateApproved'] = pd.to_datetime(df['DateApproved'])
df['ForgivenessDate'] = pd.to_datetime(df['ForgivenessDate'])
df['LoanStatusDate'] = pd.to_datetime(df['LoanStatusDate'])
print(df['DateApproved'].dtype)
print(df['ForgivenessDate'].dtype)
print(df['LoanStatusDate'].dtype)
```

```
datetime64[ns]
datetime64[ns]
datetime64[ns]
```

In [12]:
```
cat_f = df.select_dtypes('object')
```

In [13]:
```
cat_f.sample(4)
```

Out[13]:

	ProcessingMethod	BborrowerName	BorrowerAddress	BorrowerCity	BorrowerState	BorrowerZip	Lo
304825	PPS	RENOUX FLOORING CO., INC.	3578 Perch Dr SE	Iowa City	IA	52240-8276	P
159796	PPP	PISMO BEACH ATHLETIC CLUB INC.	1751 Price St	Pismo Beach	CA	93449-2230	P
455379	PPP	SIGNATURE RESEARCH, INC.	56905 CALUMET AVE	CALUMET	MI	49913-1972	P
279938	PPS	DANIEL MAZUR	5212 Whiskey Beach Ln SW	Longbranch	WA	98351-9569	P

In [14]:
```
# Data Preprocessing
```

In [15]:
```
for col in cat_f.columns:
    print(col,':',df[col].nunique())
    print('------------------------------')
```

```
ProcessingMethod : 2
------------------------------
BorrowerName : 1363134
------------------------------
BorrowerAddress : 1398584
------------------------------
BorrowerCity : 32587
------------------------------
BorrowerState : 56
------------------------------
BorrowerZip : 794249
------------------------------
LoanStatus : 3
------------------------------
FranchiseName : 2054
------------------------------
ServicingLenderName : 4310
------------------------------
ServicingLenderAddress : 4599
------------------------------
ServicingLenderCity : 2847
------------------------------
ServicingLenderState : 55
------------------------------
ServicingLenderZip : 4792
------------------------------
RuralUrbanIndicator : 2
------------------------------
HubzoneIndicator : 2
------------------------------
LMIIndicator : 2
------------------------------
BusinessAgeDescription : 5
------------------------------
ProjectCity : 32621
------------------------------
ProjectCountyName : 1911
------------------------------
ProjectState : 56
------------------------------
ProjectZip : 800227
------------------------------
CD : 450
------------------------------
Race : 9
------------------------------
Ethnicity : 3
------------------------------
BusinessType : 25
------------------------------
OriginatingLender : 4324
------------------------------
OriginatingLenderCity : 2851
------------------------------
OriginatingLenderState : 55
------------------------------
Gender : 3
------------------------------
Veteran : 3
------------------------------
NonProfit : 1
------------------------------
```

In [16]:
```
x = list(zip(df.InitialApprovalAmount,df.CurrentApprovalAmount))
```

In [17]:
```
x[0]
```

Out[17]:
```
(769358.78, 769358.78)
```

In [18]:
```
for v in x:
    if v[0]==v[1]:
        continue
    else:
        print(v)
```

```
(9571397.0, 9538531.0)
(3009400.0, 6382400.0)
(5286350.0, 5682490.0)
(3793617.0, 3793618.0)
(3894547.0, 3711148.0)
(3600000.0, 3554703.0)
(2599900.0, 3425709.0)
(1625000.0, 3403381.0)
(2045300.0, 2489263.0)
(2750000.0, 2210033.0)
(2191507.0, 2191500.0)
(2068630.0, 1981787.0)
(1212400.0, 1920138.25)
(789957.0, 1883281.94)
(2564000.0, 1863017.0)
(5385310.0, 1850000.0)
(1231872.72, 1647198.03)
(1341798.0, 1576525.0)
(816200.0, 1481939.0)
```

In [19]:
```
gc.collect()
```

Out[19]:
```
0
```

In [20]:
```
df.isnull().sum()[df.isnull().sum() != 0].sort_values()
```

Out[20]:
```
BusinessAgeDescription          1
JobsReported                    2
ProjectState                    9
BorrowerName                   12
BorrowerState                  13
BorrowerCity                   17
BorrowerZip                    18
ProjectCity                    18
ProjectZip                     19
BorrowerAddress                20
CD                             65
ProjectCountyName              83
UndisbursedAmount             240
BusinessType                  785
PAYROLL_PROCEED              2463
NAICSCode                   11933
ForgivenessAmount           97552
ForgivenessDate             97552
LoanStatusDate             121977
```

In [21]:
```
df.JobsReported.sample(2)
```

Out[21]:
```
369162    43.0
592065    11.0
Name: JobsReported, dtype: float64
```

In [22]:
```
df['JobsReported'].fillna(0,inplace=True)
```

In [23]:
```
df.BorrowerCity.sample(2)
```

Out[23]:
```
15180     Bentonville
318603       Rockford
Name: BorrowerCity, dtype: object
```

In [24]:
```
df['JobsReported'].fillna('Other',inplace=True)
```

In [25]:
```
df.BorrowerZip.sample(2)
```

Out[25]:
```
251811    34120-1648
430824         53205
Name: BorrowerZip, dtype: object
```

In [26]:
```
df.ProjectZip.sample()
```

Out[26]:
```
70593    95008-2050
Name: ProjectZip, dtype: object
```

In [27]:
```
df.dropna(subset=['BorrowerZip'],inplace=True)
```

In [28]:
```
df.dropna(subset=['ProjectZip'],inplace=True)
```

In [29]:
```
df.BusinessAgeDescription.unique()
```

Out[29]:
```
array(['Existing or more than 2 years old',
       'New Business or 2 years or less', 'Unanswered',
       'Change of Ownership', 'Startup, Loan Funds will Open Business',
       nan], dtype=object)
```

In [30]:
```
df['BusinessAgeDescription'].fillna('Unanswered',inplace=True)
```

In [31]:
```
df['BorrowerName'].fillna('Unnamed',inplace=True)
```

In [32]:
```
df['ProjectState'].fillna('Other',inplace=True)
```

In [33]:
```
df.isnull().sum()[df.isna().sum() != 0].sort_values()
```

Out[33]:
```
BorrowerAddress                   3
BorrowerState                     5
CD                               46
ProjectCountyName                64
UndisbursedAmount               240
BusinessType                    782
PAYROLL_PROCEED                2462
NAICSCode                     11932
ForgivenessAmount             97550
ForgivenessDate               97550
LoanStatusDate               121975
UTILITIES_PROCEED           1056989
RENT_PROCEED                1447246
NonProfit                   1490992
HEALTH_CARE_PROCEED         1505033
MORTGAGE_INTEREST_PROCEED   1507389
FranchiseName               1524753
DEBT_INTEREST_PROCEED       1532882
REFINANCE_EIDL_PROCEED      1541916
dtype: int64
```

In [34]:
```
df['BorrowerAddress'].fillna('Not_Given',inplace=True)
```

In [35]:
```
df['BorrowerState'].fillna('Other',inplace=True)
```

In [36]:
```
df.CD.nunique()
```

Out[36]:
```
450
```

In [37]:
```
x=df['CD'].mode()
x
```

Out[37]:
```
0    WA-07
dtype: object
```

In [38]:
```
df['CD'].fillna('WA-07',inplace=True)
```

In [39]:
```
df['ProjectCountyName'].nunique()
```

Out[39]:
```
1911
```

In [40]:
```
df['ProjectCountyName'].fillna('Other',inplace=True)
```

In [41]:
```
df.UndisbursedAmount.head(3)
```

Out[41]:
```
0    0.0
1    0.0
2    0.0
Name: UndisbursedAmount, dtype: float64
```

In [42]:
```
df['UndisbursedAmount'].fillna(0,inplace=True)
```

In [43]:
```
df.BusinessType.unique()
```

Out[43]:
```
array(['Corporation', 'Sole Proprietorship', 'Non-Profit Organization',
       'Limited  Liability Company(LLC)', '501(c)3 - Non Profit',
       'Subchapter S Corporation', 'Cooperative', nan, 'Partnership',
       'Professional Association', 'Employee Stock Ownership Plan(ESOP)',
       'Limited Liability Partnership', 'Non-Profit Childcare Center',
       'Trust', 'Joint Venture', '501(c)6 - Non Profit Membership',
       'Self-Employed Individuals', 'Single Member LLC',
       'Independent Contractors', 'Tribal Concerns', 'Tenant in Common',
       'Housing Co-op', 'Rollover as Business Start-Ups (ROB',
       '501(c) - Non Profit except 3,4,6,',
       '501(c)19 - Non Profit Veterans',
       'Qualified Joint-Venture (spouses)'], dtype=object)
```

In [44]:
```
df['BusinessType'].fillna('Other',inplace=True)
```

In [45]:
```
df.PAYROLL_PROCEED.sample(3)
```

Out[45]:
```
115195      11831.0
13911     1348500.0
333100       9955.0
Name: PAYROLL_PROCEED, dtype: float64
```

In [46]:
```
df['PAYROLL_PROCEED'].fillna(0,inplace=True)
```

In [47]:
```
df.NAICSCode.sample(5)
```

Out[47]:
```
80677     621991.0
575100    522310.0
583557    453998.0
313555    423110.0
715378    211120.0
Name: NAICSCode, dtype: float64
```

In [48]:
```
df['NAICSCode'].fillna(0,inplace=True) # 0 means empty
```

In [49]:
```
((df.isnull().sum()[df.isna().sum() != 0]*100)/len(df)).sort_values()
```

Out[49]:
```
ForgivenessAmount             6.219746
ForgivenessDate               6.219746
LoanStatusDate                7.777074
UTILITIES_PROCEED            67.393164
RENT_PROCEED                 92.275783
NonProfit                    95.065009
HEALTH_CARE_PROCEED          95.960257
MORTGAGE_INTEREST_PROCEED    96.110475
FranchiseName                97.217596
DEBT_INTEREST_PROCEED        97.735898
REFINANCE_EIDL_PROCEED       98.311902
dtype: float64
```

In [50]:
```
df.ForgivenessAmount.sample(4)
```

Out[50]:
```
277133      21641.39
206743    2653179.06
192610     261877.44
343629       7610.30
Name: ForgivenessAmount, dtype: float64
```

In [51]:
```
df['ForgivenessAmount'].fillna(0,inplace=True)
```

In [52]:
```
df.ForgivenessDate.nunique()
```

Out[52]:
```
424
```

In [53]:
```
df.ForgivenessDate.mode()
```

Out[53]:
```
0   2020-11-03
dtype: datetime64[ns]
```

In [54]:
```
df.sample(6)
```

Out[54]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddres
293249	9726937209	2020-04-28	1013	PPP	STILLWATERS ENVIRONMENTAL EDUCATION CENTER	26059 BARBE CUT OFF R
467659	8178438800	2021-04-22	563	PPS	JACQUELINE BABLER	1730 96th Av
465725	4550929005	2021-05-20	563	PPP	SHURONE GOODLOW	7463 N 39th
145591	2711077707	2020-05-01	920	PPP	BOSHART AUTOMOTIVE TESTING SERVICES, INC.	13195 SANT ANA Av
603382	3505068410	2021-02-05	944	PPS	JM FURNITURE INCORPORATED	3333 N Carson
40487	6519907204	2020-04-28	912	PPP	ALL SEASONS ROOFING AND WATERPROOFING, INC	1720 Smith Av

```
pairs = df[["LoanStatus", "LoanStatusDate"]].dropna().to_dict()
fill_values = dict(zip(pairs['LoanStatus'].values(), pairs['LoanStatusDate'].values()))
df.LoanStatusDate = df.apply(lambda df: fill_values[df.LoanStatus] if df.LoanStatusDate
is np.nan else df.LoanStatusDate, axis=1)
```

In [55]:
```
df['UTILITIES_PROCEED'].value_counts().head()
```

Out[55]:
```
1.0       362797
0.0         8192
1000.0      3062
5000.0      2685
2000.0      2596
Name: UTILITIES_PROCEED, dtype: int64
```

In [56]:
```
df['RENT_PROCEED'].value_counts().head()
```

Out[56]:
```
0.0        7403
20000.0    1982
10000.0    1938
30000.0    1568
5000.0     1208
Name: RENT_PROCEED, dtype: int64
```

In [57]:
```
df['UTILITIES_PROCEED'].fillna(1,inplace=True)
df['RENT_PROCEED'].fillna(0,inplace=True)
```

In [58]:
```
((df.isnull().sum()[df.isna().sum() != 0]*100)/len(df)).sort_values()
```

Out[58]:
```
ForgivenessDate               6.219746
LoanStatusDate                7.777074
NonProfit                    95.065009
HEALTH_CARE_PROCEED          95.960257
MORTGAGE_INTEREST_PROCEED    96.110475
FranchiseName                97.217596
DEBT_INTEREST_PROCEED        97.735898
REFINANCE_EIDL_PROCEED       98.311902
dtype: float64
```

In [59]:
```
df['BusinessType'].head()
```

Out[59]:
```
0                Corporation
1        Sole Proprietorship
2    Non-Profit Organization
3                Corporation
6    Non-Profit Organization
Name: BusinessType, dtype: object
```

In [60]:
```
df['NonProfit'].head()
```

Out[60]:
```
0    NaN
1    NaN
2      Y
3    NaN
6      Y
Name: NonProfit, dtype: object
```

In [61]:
```
# 'Yes' if Business Type = Non-Profit Organization or Non-Profit Childcare Center or 501
```

In [62]:
```
df['NonProfit'].fillna('N',inplace = True)
```

In [63]:
```
df.HEALTH_CARE_PROCEED.value_counts().head()
```

Out[63]:
```
0.0        7714
10000.0     860
20000.0     850
15000.0     672
30000.0     610
Name: HEALTH_CARE_PROCEED, dtype: int64
```

In [64]:
```
df['HEALTH_CARE_PROCEED'].fillna(0,inplace = True)
```

In [65]:
```
df['NonProfit'].fillna(0,inplace = True)
```

In [66]:
```
df['MORTGAGE_INTEREST_PROCEED'].value_counts().head()
```

Out[66]:
```
0.0        19946
1000.0       635
10000.0      544
5000.0       531
2000.0       493
Name: MORTGAGE_INTEREST_PROCEED, dtype: int64
```

In [67]:
```
df['MORTGAGE_INTEREST_PROCEED'].fillna(0,inplace = True)
```

In [68]:
```
df.shape
```

Out[68]:
```
(1568392, 53)
```

In [69]:
```
df.FranchiseName.isnull().sum()
```

Out[69]:
```
1524753
```

In [70]:
```
df.FranchiseName.value_counts().head()
```

Out[70]:
```
McDonalds
1813
General Motors, LLC (Chevrolet, Buick, GM, Cadillac) Dealer Sales and Service Agreement
1727
Subway
1538
Ford Motor Company Dealer Sales and Service Agreement
1267
IHOP
1068
Name: FranchiseName, dtype: int64
```

In [71]:
```
df['FranchiseName'].fillna('Other',inplace = True)
```

In [72]:
```
df.DEBT_INTEREST_PROCEED.value_counts().head()
```

Out[72]:
```
0.0        17393
10000.0      461
1000.0       452
5000.0       449
2000.0       434
Name: DEBT_INTEREST_PROCEED, dtype: int64
```

In [73]:
```
df['DEBT_INTEREST_PROCEED'].fillna(0,inplace = True)
```

In [74]:
```
df.REFINANCE_EIDL_PROCEED.value_counts().head()
```

Out[74]:
```
0.0         23956
10000.0       996
1000.0        131
2000.0         48
150000.0       41
Name: REFINANCE_EIDL_PROCEED, dtype: int64
```

In [75]:
```
df['REFINANCE_EIDL_PROCEED'].fillna(0,inplace = True)
```

In [76]:
```
df.isnull().sum()[df.isnull().sum() != 0]
```

Out[76]:
```
LoanStatusDate     121975
ForgivenessDate     97550
dtype: int64
```

In [77]:
```
df.ForgivenessDate.mode()
```

Out[77]:
```
0   2020-11-03
dtype: datetime64[ns]
```

In [79]:
```
df.LoanStatusDate.mode()
```

Out[79]:
```
0   2021-01-21
dtype: datetime64[ns]
```

In [80]:
```
df['ForgivenessDate'].fillna('2020-11-03',inplace = True)
df['LoanStatusDate'].fillna('2021-01-21',inplace = True)
```

In [81]:
```
df.isnull().sum()
```

Out[81]:
```
LoanNumber                    0
DateApproved                  0
SBAOfficeCode                 0
ProcessingMethod              0
BorrowerName                  0
BorrowerAddress               0
BorrowerCity                  0
BorrowerState                 0
BorrowerZip                   0
LoanStatusDate                0
LoanStatus                    0
Term                          0
SBAGuarantyPercentage         0
InitialApprovalAmount         0
CurrentApprovalAmount         0
UndisbursedAmount             0
FranchiseName                 0
ServicingLenderLocationID     0
ServicingLenderName           0
ServicingLenderAddress        0
ServicingLenderCity           0
ServicingLenderState          0
ServicingLenderZip            0
RuralUrbanIndicator           0
HubzoneIndicator              0
LMIIndicator                  0
BusinessAgeDescription        0
ProjectCity                   0
ProjectCountyName             0
ProjectState                  0
ProjectZip                    0
CD                            0
JobsReported                  0
NAICSCode                     0
Race                          0
Ethnicity                     0
UTILITIES_PROCEED             0
PAYROLL_PROCEED               0
MORTGAGE_INTEREST_PROCEED     0
RENT_PROCEED                  0
REFINANCE_EIDL_PROCEED        0
HEALTH_CARE_PROCEED           0
DEBT_INTEREST_PROCEED         0
BusinessType                  0
OriginatingLenderLocationID   0
OriginatingLender             0
OriginatingLenderCity         0
OriginatingLenderState        0
Gender                        0
Veteran                       0
NonProfit                     0
ForgivenessAmount             0
ForgivenessDate               0
dtype: int64
```

In [82]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1568392 entries, 0 to 599878
Data columns (total 53 columns):
 #   Column                         Non-Null Count    Dtype
---  ------                         --------------    -----
 0   LoanNumber                     1568392 non-null  int64
 1   DateApproved                   1568392 non-null  datetime64[ns]
 2   SBAOfficeCode                  1568392 non-null  int64
 3   ProcessingMethod               1568392 non-null  object
 4   BorrowerName                   1568392 non-null  object
 5   BorrowerAddress                1568392 non-null  object
 6   BorrowerCity                   1568392 non-null  object
 7   BorrowerState                  1568392 non-null  object
 8   BorrowerZip                    1568392 non-null  object
 9   LoanStatusDate                 1568392 non-null  datetime64[ns]
 10  LoanStatus                     1568392 non-null  object
 11  Term                           1568392 non-null  int64
 12  SBAGuarantyPercentage          1568392 non-null  int64
 13  InitialApprovalAmount          1568392 non-null  float64
 14  CurrentApprovalAmount          1568392 non-null  float64
 15  UndisbursedAmount              1568392 non-null  float64
 16  FranchiseName                  1568392 non-null  object
 17  ServicingLenderLocationID      1568392 non-null  int64
 18  ServicingLenderName            1568392 non-null  object
 19  ServicingLenderAddress         1568392 non-null  object
 20  ServicingLenderCity            1568392 non-null  object
 21  ServicingLenderState           1568392 non-null  object
 22  ServicingLenderZip             1568392 non-null  object
 23  RuralUrbanIndicator            1568392 non-null  object
 24  HubzoneIndicator               1568392 non-null  object
 25  LMIIndicator                   1568392 non-null  object
 26  BusinessAgeDescription         1568392 non-null  object
 27  ProjectCity                    1568392 non-null  object
 28  ProjectCountyName              1568392 non-null  object
 29  ProjectState                   1568392 non-null  object
 30  ProjectZip                     1568392 non-null  object
 31  CD                             1568392 non-null  object
 32  JobsReported                   1568392 non-null  float64
 33  NAICSCode                      1568392 non-null  float64
 34  Race                           1568392 non-null  object
 35  Ethnicity                      1568392 non-null  object
 36  UTILITIES_PROCEED              1568392 non-null  float64
 37  PAYROLL_PROCEED                1568392 non-null  float64
 38  MORTGAGE_INTEREST_PROCEED      1568392 non-null  float64
 39  RENT_PROCEED                   1568392 non-null  float64
 40  REFINANCE_EIDL_PROCEED         1568392 non-null  float64
 41  HEALTH_CARE_PROCEED            1568392 non-null  float64
 42  DEBT_INTEREST_PROCEED          1568392 non-null  float64
 43  BusinessType                   1568392 non-null  object
 44  OriginatingLenderLocationID    1568392 non-null  int64
 45  OriginatingLender              1568392 non-null  object
 46  OriginatingLenderCity          1568392 non-null  object
 47  OriginatingLenderState         1568392 non-null  object
 48  Gender                         1568392 non-null  object
 49  Veteran                        1568392 non-null  object
 50  NonProfit                      1568392 non-null  object
 51  ForgivenessAmount              1568392 non-null  float64
 52  ForgivenessDate                1568392 non-null  datetime64[ns]
dtypes: datetime64[ns](3), float64(13), int64(6), object(31)
memory usage: 646.2+ MB
```

In [83]:
```
gc.collect()
```

Out[83]: 0

In [84]:
```
df.to_csv('Data_Merged_cleaned.csv')
```

In [2]:
```
df = pd.read_pickle('Data_Merged_cleaned.csv')
```

In [4]:
```
df.to_pickle('Data_Merged_cleaned_.pkl')
```

##

In [2]:
```
import gc
import warnings
warnings.filterwarnings('ignore')
```

In [3]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
pd.set_option('display.max_columns',None)
pd.set_option('display.max_rows',None)
```

In [5]:
```
df = pd.read_pickle('Data_Merged_cleaned_.pkl')
```

In [5]:
```
df.shape
```

Out[5]:
```
(1568392, 54)
```

In [6]:
```
df.sample(5)
```

Out[6]:

	Unnamed: 0	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	Borrov
1192283	223765	5290998408	2021-02-08	1013	PPS	TENPOINT7 INC.	583 Ba
681443	681454	8493157008	2020-04-08	549	PPP	WALSH JESUIT HIGH SCHOOL	4550 \
654199	654208	1554947110	2020-04-10	202	PPP	STUTTERING ASSOCIATION FOR THE YOUNG	247
99964	99972	4651577308	2020-04-30	914	PPP	ELITE SALES AND SERVICES, LLC	WI
1547989	579476	9221807003	2020-04-09	897	PPP	SCHILLING ICE, INC	19

In [6]:
```
df.drop('Unnamed: 0',axis = 1, inplace = True)
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1568392 entries, 0 to 1568391
Data columns (total 53 columns):
 #   Column                         Non-Null Count    Dtype  
---  ------                         --------------    -----  
 0   LoanNumber                     1568392 non-null  int64  
 1   DateApproved                   1568392 non-null  object 
 2   SBAOfficeCode                  1568392 non-null  int64  
 3   ProcessingMethod               1568392 non-null  object 
 4   BorrowerName                   1568392 non-null  object 
 5   BorrowerAddress                1568392 non-null  object 
 6   BorrowerCity                   1568392 non-null  object 
 7   BorrowerState                  1568392 non-null  object 
 8   BorrowerZip                    1568392 non-null  object 
 9   LoanStatusDate                 1568392 non-null  object 
 10  LoanStatus                     1568392 non-null  object 
 11  Term                           1568392 non-null  int64  
 12  SBAGuarantyPercentage          1568392 non-null  int64  
 13  InitialApprovalAmount          1568392 non-null  float64
 14  CurrentApprovalAmount          1568392 non-null  float64
 15  UndisbursedAmount              1568392 non-null  float64
 16  FranchiseName                  1568392 non-null  object 
 17  ServicingLenderLocationID      1568392 non-null  int64  
 18  ServicingLenderName            1568392 non-null  object 
 19  ServicingLenderAddress         1568392 non-null  object 
 20  ServicingLenderCity            1568392 non-null  object 
 21  ServicingLenderState           1568392 non-null  object 
 22  ServicingLenderZip             1568392 non-null  object 
 23  RuralUrbanIndicator            1568392 non-null  object 
 24  HubzoneIndicator               1568392 non-null  object 
 25  LMIIndicator                   1568392 non-null  object 
 26  BusinessAgeDescription         1568392 non-null  object 
 27  ProjectCity                    1568392 non-null  object 
 28  ProjectCountyName              1568392 non-null  object 
 29  ProjectState                   1568392 non-null  object 
 30  ProjectZip                     1568392 non-null  object 
 31  CD                             1568392 non-null  object 
 32  JobsReported                   1568392 non-null  float64
 33  NAICSCode                      1568392 non-null  float64
 34  Race                           1568392 non-null  object 
 35  Ethnicity                      1568392 non-null  object 
 36  UTILITIES_PROCEED              1568392 non-null  float64
 37  PAYROLL_PROCEED                1568392 non-null  float64
 38  MORTGAGE_INTEREST_PROCEED      1568392 non-null  float64
 39  RENT_PROCEED                   1568392 non-null  float64
 40  REFINANCE_EIDL_PROCEED         1568392 non-null  float64
 41  HEALTH_CARE_PROCEED            1568392 non-null  float64
 42  DEBT_INTEREST_PROCEED          1568392 non-null  float64
 43  BusinessType                   1568392 non-null  object 
 44  OriginatingLenderLocationID    1568392 non-null  int64  
 45  OriginatingLender              1568392 non-null  object 
 46  OriginatingLenderCity          1568392 non-null  object 
 47  OriginatingLenderState         1568392 non-null  object 
 48  Gender                         1568392 non-null  object 
 49  Veteran                        1568392 non-null  object 
 50  NonProfit                      1568392 non-null  object 
 51  ForgivenessAmount              1568392 non-null  float64
 52  ForgivenessDate                1568392 non-null  object 
dtypes: float64(13), int64(6), object(34)
memory usage: 634.2+ MB
```

In [8]:
```
df.isnull().sum()
```

Out[8]:
```
LoanNumber                    0
DateApproved                  0
SBAOfficeCode                 0
ProcessingMethod              0
BorrowerName                  0
BorrowerAddress               0
BorrowerCity                  0
BorrowerState                 0
BorrowerZip                   0
LoanStatusDate                0
LoanStatus                    0
Term                          0
SBAGuarantyPercentage         0
InitialApprovalAmount         0
CurrentApprovalAmount         0
UndisbursedAmount             0
FranchiseName                 0
ServicingLenderLocationID     0
ServicingLenderName           0
ServicingLenderAddress        0
ServicingLenderCity           0
ServicingLenderState          0
ServicingLenderZip            0
RuralUrbanIndicator           0
HubzoneIndicator              0
LMIIndicator                  0
BusinessAgeDescription        0
ProjectCity                   0
ProjectCountyName             0
ProjectState                  0
ProjectZip                    0
CD                            0
JobsReported                  0
NAICSCode                     0
Race                          0
Ethnicity                     0
UTILITIES_PROCEED             0
PAYROLL_PROCEED               0
MORTGAGE_INTEREST_PROCEED     0
RENT_PROCEED                  0
REFINANCE_EIDL_PROCEED        0
HEALTH_CARE_PROCEED           0
DEBT_INTEREST_PROCEED         0
BusinessType                  0
OriginatingLenderLocationID   0
OriginatingLender             0
OriginatingLenderCity         0
OriginatingLenderState        0
Gender                        0
Veteran                       0
NonProfit                     0
ForgivenessAmount             0
ForgivenessDate               0
dtype: int64
```

In [9]:
```
df.describe()
```

Out[9]:

	LoanNumber	SBAOfficeCode	Term	SBAGuarantyPercentage	InitialApprovalAmount	CurrentA
count	1.568392e+06	1.568392e+06	1.568392e+06	1568392.0	1.568392e+06	
mean	5.451237e+09	5.872471e+02	3.952417e+01	100.0	3.391809e+05	
std	2.554000e+09	2.793495e+02	1.796623e+01	0.0	6.346320e+05	
min	1.000007e+09	1.010000e+02	0.000000e+00	100.0	0.000000e+00	
25%	3.291588e+09	3.530000e+02	2.400000e+01	100.0	2.213969e+04	
50%	5.426673e+09	5.630000e+02	2.400000e+01	100.0	1.858482e+05	
75%	7.586270e+09	8.970000e+02	6.000000e+01	100.0	3.570320e+05	
max	9.999009e+09	1.094000e+03	1.200000e+02	100.0	1.000000e+07	

In [13]:
```
df.sample(4)
```

Out[13]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress
348202	5321067005	2020-04-05	507	PPP	MUNIE TRENCHING AND EXCAVATING INCORPORATED	1818 Pine St
1146588	7876577302	2020-04-30	304	PPP	QUALITY COMFORT	3040 Bethel rc
1199120	8462868307	2021-01-29	1013	PPS	BAE FAMILY DENTISTRY INC.	11410 NE 19th St
856715	4588848309	2021-01-23	610	PPS	CRYSTAL CREEK CATTLE COMPANY	2459 Southwell Rc

Loan Approval Amount(at origination)

In [25]:
```
#1. Loan Approval Amount less than 150000
```

In [16]:
```
l1 = df[df['InitialApprovalAmount']<150000]
l1.head()
```

Out[16]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress	Bc
1181	2698217205	2020-04-16	1084	PPP	FAM ALASKA INC	1404 4TH AVE	
1190	3082727310	2020-04-29	1084	PPP	KINGFISHER CHARTERS, LLC	302 Islander Drive	
1252	6971107109	2020-04-14	1084	PPP	CAPTAIN PATTIES FISH HOUSE INC	4241 HOMER SPIT RD HOMER	
1259	7402437108	2020-04-14	1084	PPP	MILLER'S LANDING INC	13890 BEACH DR	
1274	9243077108	2020-04-15	1084	PPP	MAHAY'S RIVERBOAT SERVICE, INC.	22333 S. Talkeetna Spur Rd	T

In [24]:
```
#2. Loan Approval Amount greater than 150000
```

In [17]:
```
l2 =df[df['InitialApprovalAmount']>=150000]
l2.head()
```

Out[17]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress	Bor
0	9547507704	2020-05-01	464	PPP	SUMTER COATINGS, INC.	2410 Highway 15 South	
1	9777677704	2020-05-01	464	PPP	PLEASANT PLACES, INC.	7684 Southrail Road	
2	5791407702	2020-05-01	1013	PPP	BOYER CHILDREN'S CLINIC	1850 BOYER AVE E	
3	6223567700	2020-05-01	920	PPP	KIRTLEY CONSTRUCTION INC	1661 MARTIN RANCH RD	BER
4	9794577700	2020-05-01	491	PPP	FRUIT COVE BAPTIST CHURCH OF JACKSONVILLE FL INC	501 State Road 13	S

 # Statewise Loan Average

In [20]:
```
statewise_loan = df.groupby('BorrowerState',as_index = False)['InitialApprovalAmount'].m
```

In [22]:
```
statewise_loan.sample(5)
```

Out[22]:

	BorrowerState	InitialApprovalAmount
5	CA	534487.008182
28	MS	487215.236441
22	MD	534849.415216
7	CT	518156.160538
26	MO	536779.131301

In [55]:
```
max = statewise_loan.InitialApprovalAmount.max()
```

In [56]:
```
min = statewise_loan.InitialApprovalAmount.min()
```

In [59]:
```
statewise_loan.loc[statewise_loan['InitialApprovalAmount'] == max, 'BorrowerState'].iloc
```

Out[59]:
```
'Other'
```

In [60]:
```
statewise_loan.loc[statewise_loan['InitialApprovalAmount'] == min, 'BorrowerState'].iloc
```

Out[60]:
```
'VI'
```

VI state have the lowest loan amount.

 # countrywise loan average

In [26]:
```
countrywise_loan = df.groupby('ProjectCountyName',as_index = False)['InitialApprovalAmou
```

In [28]:
```
countrywise_loan.head()
```

Out[28]:

	ProjectCountyName	InitialApprovalAmount
0	ABBEVILLE	773451.627500
1	ACADIA	494803.006036
2	ACCOMACK	47394.925424
3	ADA	521445.673490
4	ADAIR	627428.526935

In [49]:
```
# filter the data whose loan amount average is maximum.
```

In [41]:
```
countrywise_loan.InitialApprovalAmount.max()
```

Out[41]:
```
3270927.0
```

In [48]:
```
countrywise_loan.loc[countrywise_loan['InitialApprovalAmount'] == 3270927.0, 'ProjectCou
```

Out[48]:
```
'TWIGGS'
```

TWIGGS country have highest loan Amount

In [ ]:
```
# filter the data whose loan amount average is minimum.
```

In [51]:
```
countrywise_loan.InitialApprovalAmount.min()
```

Out[51]:
```
23503.77064777328
```

In [52]:
```
countrywise_loan.loc[countrywise_loan['InitialApprovalAmount'] == 23503.77064777328, 'Pr
```

Out[52]:
```
'NIOBRARA'
```

NIOBRARA country have lowest loan Amount.

 # Average Loan for perticula city

In [61]:
```
df.head(1)
```

Out[61]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress	Borro
0	9547507704	2020-05-01	464	PPP	SUMTER COATINGS, INC.	2410 Highway 15 South	

In [ ]:
```

```

In [63]:
```
citywise_loan = df.groupby('BorrowerCity',as_index = False)['InitialApprovalAmount'].mea
```

In [67]:
```
max = citywise_loan['InitialApprovalAmount'].max()
min = citywise_loan['InitialApprovalAmount'].min()
```

In [66]:
```
citywise_loan.sample(5)
```

Out[66]:

	BorrowerCity	InitialApprovalAmount
18021	Mt Clemens	301133.000000
9305	Floodwood	161000.000000
24063	SELLERSVILLE	692731.736842
18285	NATHALIE	13486.817273
22081	Prairieville	543648.933333

In [69]:
```
citywise_loan.loc[citywise_loan['InitialApprovalAmount'] == max,'BorrowerCity'].iloc[0]
```

Out[69]: 'Gardenda'

Gardenda city have maximum loan amount.

In [70]:
```
citywise_loan.loc[citywise_loan['InitialApprovalAmount'] == min,'BorrowerCity'].iloc[0]
```

Out[70]: 'NORTHWEST ARTCENTER'

NORTHWEST ARTCENTER city have minimum loan amount.

 # Loan Amount grouped by Business Type, Race,Gender etc.

In [72]:
```
bt_wise_loan = df.groupby('BusinessType',as_index = False)['InitialApprovalAmount'].mean
```

In [73]:
```
bt_wise_loan.head()
```

Out[73]:

	BusinessType	InitialApprovalAmount
0	501(c) – Non Profit except 3,4,6,	374094.889000
1	501(c)19 – Non Profit Veterans	35269.808095
2	501(c)3 – Non Profit	994268.343158
3	501(c)6 – Non Profit Membership	329806.493616
4	Cooperative	580710.307567

In [82]:
```
genderwise_loan = df.groupby('Gender',as_index = False)['InitialApprovalAmount'].mean()
```

In [83]:
```
genderwise_loan
```

Out[83]:

	Gender	InitialApprovalAmount
0	Female Owned	216248.664373
1	Male Owned	337933.725974
2	Unanswered	358404.747011

In [80]:
```
brg_wise_loan = df.groupby(['BusinessType','Race','Gender'],as_index = False)['InitialAp
```

In [81]:
```
brg_wise_loan.head()
```

Out[81]:

	BusinessType	Race	Gender	InitialApprovalAmount
0	501(c) – Non Profit except 3,4,6,	Unanswered	Female Owned	288200.000000
1	501(c) – Non Profit except 3,4,6,	Unanswered	Male Owned	546789.375000
2	501(c) – Non Profit except 3,4,6,	Unanswered	Unanswered	374523.000000
3	501(c) – Non Profit except 3,4,6,	White	Male Owned	172181.796667
4	501(c)19 – Non Profit Veterans	Unanswered	Male Owned	111253.290000

In [103]:
```
df.sample(4)
```

Out[103]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress
1459926	2465057204	2020-04-16	563	PPP	ROB'S CYCLE WORKS, LLC	141204 ROCKING HORSE RD
1156132	2011507105	2020-04-10	150	PPP	DONNA P JOHNSON PHYSICAL THERAPY PC	28 Fourth St
858906	4579437208	2020-04-27	610	PPP	SPOONER & ASSOCIATES INC	309 BYERS ST, Suite 100
158836	8266667008	2020-04-08	914	PPP	CHRISTOPHER GRATTAN, INC.	141 ECOLORADO BLVD

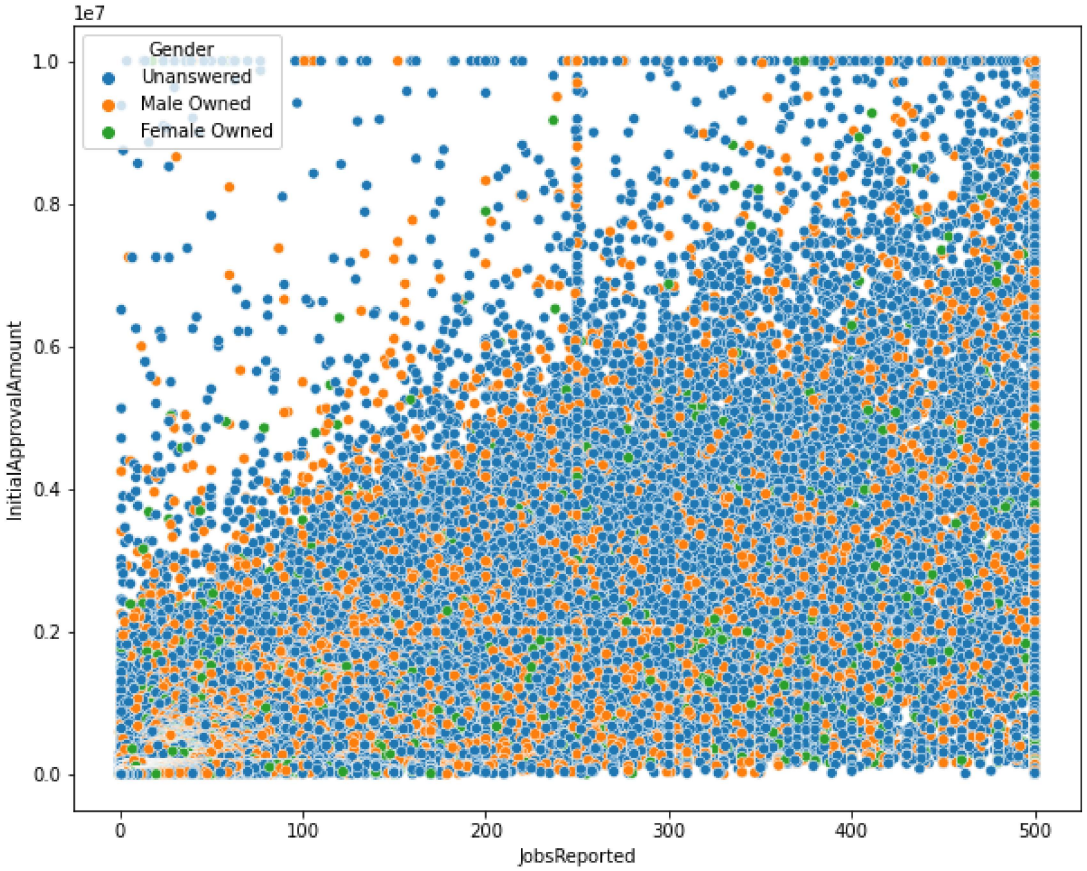
In [107]:
```
df.iloc[:,[13,32]].head().sort_values(by='InitialApprovalAmount')
```

Out[107]:

	InitialApprovalAmount	JobsReported
4	289765.00	89.0
3	499871.00	21.0
2	691355.00	75.0
1	736927.79	73.0
0	769358.78	62.0

In [8]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [9]:
```
plt.figure(figsize=(10,8))
sns.scatterplot(data = df ,y = 'InitialApprovalAmount', x = 'JobsReported', hue = 'Gende
```



From the above plot we can say that as the JobsReported(Number of Employees) increases loan is also increases.

 # Amount of loan given by each lender etc.

In [97]:
```
lender_wise_loan = df.groupby(['ServicingLenderName'],as_index = False)['InitialApproval
```

In [98]:
```
lender_wise_loan.head()
```

Out[98]:

	ServicingLenderName	InitialApprovalAmount
0	\tFarm Credit Services of Western Arkansas, ACA	20833.33
1	\tYankee Farm Credit, ACA	10630436.26
2	121 Financial CU	29051350.00
3	1st Advantage Bank	20623549.93
4	1st Advantage FCU	4220556.69

In [94]:
```
# cross check
s = df[df.ServicingLenderName=='1st Advantage Bank'][['ServicingLenderName','InitialAppr
```

In [95]:
```
s['InitialApprovalAmount'].sum() # matches with the inex 3 of lender wise loan
```

Out[95]: 20623549.93

In [100]:
```
max =lender_wise_loan.InitialApprovalAmount.max()
min =lender_wise_loan.InitialApprovalAmount.min()
```

In [102]:
```
lender_wise_loan.loc[lender_wise_loan['InitialApprovalAmount'] == max,'ServicingLenderNa
```

Out[102]: 'JPMorgan Chase Bank, National Association'

JPMorgan Chase Bank, National Association give the highest loan

In [101]:
```
lender_wise_loan.loc[lender_wise_loan['InitialApprovalAmount'] == min,'ServicingLenderNa
```

Out[101]: '1st Liberty FCU'

1st Liberty FCU give the highest loan

In [10]:
```
df.sample(5)
```

Out[10]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress
1146153	2242447304	2020-04-29	353	PPP	JAMES T. TRAUTZ JR.	20565 Warburtor Bay Square
96766	1628288505	2021-02-19	942	PPS	JM SURLA LLC	431 12th S
395621	2879688604	2021-03-15	679	PPS	REB TRUCKING LLC	6176 Twin Bridges Rc
665041	2003757709	2020-05-01	235	PPP	MAYFAIR POWER SYSTEMS, INC.	347 N MAIN ST
1307826	6513217709	2020-05-01	1013	PPP	CAMPBELL CONSTRUCTION	13867SE 240TH ST

 # Highest loan lender in each city.

In [7]:
```
xx = df.groupby(['BorrowerCity'],as_index=False)['ServicingLenderName','InitialApprovalA
xx.sort_values(by='InitialApprovalAmount',ascending = True).head()
```

Out[7]:

	BorrowerCity	ServicingLenderName	InitialApprovalAmount
18847	NORTHWEST ARTCENTER	Bank of America, National Association	31.0
879	Alkol	Capital Plus Financial, LLC	230.0
25125	SW OLYMPIA	Wells Fargo Bank, National Association	354.0
2877	Berryville, VA	Bank of Clarke County	500.0
16022	MD SHOHID PATOARY	Bank of America, National Association	543.0

```
sns.histplot(df,x = 'BorrowerCity', y = 'InitialApprovalAmount');
```

In [5]:
```
gc.collect()
```

Out[5]:
```
0
```

 # Number of jobs reported by different business types.

In [6]:
```
jr_by_bt = df.groupby(['BusinessType'],as_index = False)['JobsReported'].sum()
```

In [11]:
```
jr_by_bt.sort_values(by = 'JobsReported', ascending = False)
```

Out[11]:

	BusinessType	JobsReported
5	Corporation	20988909.0
10	Limited Liability Company(LLC)	14478468.0
22	Subchapter S Corporation	9266970.0
13	Non-Profit Organization	4544220.0
15	Partnership	961344.0
21	Sole Proprietorship	783408.0
11	Limited Liability Partnership	638018.0
16	Professional Association	341654.0
2	501(c)3 – Non Profit	247089.0
4	Cooperative	132544.0
6	Employee Stock Ownership Plan(ESOP)	65908.0
12	Non-Profit Childcare Center	61632.0
19	Self-Employed Individuals	50111.0
14	Other	47988.0
8	Independent Contractors	35461.0
3	501(c)6 – Non Profit Membership	25312.0
25	Trust	25210.0
20	Single Member LLC	10642.0
9	Joint Venture	7894.0
24	Tribal Concerns	6738.0
7	Housing Co-op	3634.0
23	Tenant in Common	1230.0
0	501(c) – Non Profit except 3,4,6,	478.0
17	Qualified Joint-Venture (spouses)	282.0
1	501(c)19 – Non Profit Veterans	212.0
18	Rollover as Business Start-Ups (ROB	98.0

 # Average Loan Amount per business type.

In [8]:
```
avg_ln_bt = df.groupby(['BusinessType'],as_index = False)['InitialApprovalAmount'].mean(
```

In [17]:
```
pd.set_option('display.float_format', lambda x: '%.4f' % x)
```

In [18]:
```
avg_ln_bt.sort_values(by = 'InitialApprovalAmount', ascending = False)
```

Out[18]:

	BusinessType	InitialApprovalAmount
6	Employee Stock Ownership Plan(ESOP)	1394017.4327
2	501(c)3 – Non Profit	994268.3432
24	Tribal Concerns	798165.5318
14	Other	704243.5431
25	Trust	627543.8661
4	Cooperative	580710.3076
13	Non-Profit Organization	551321.1128
11	Limited Liability Partnership	448404.1393
15	Partnership	442311.0340
22	Subchapter S Corporation	425407.0438
5	Corporation	418133.9903
16	Professional Association	413029.4534
0	501(c) – Non Profit except 3,4,6,	374094.8890
12	Non-Profit Childcare Center	361404.6805
7	Housing Co-op	349286.4173
3	501(c)6 – Non Profit Membership	329806.4936
9	Joint Venture	325121.9697
10	Limited Liability Company(LLC)	313036.4635
23	Tenant in Common	244213.5012
18	Rollover as Business Start-Ups (ROB	221332.4000
21	Sole Proprietorship	42559.9119
1	501(c)19 – Non Profit Veterans	35269.8081
20	Single Member LLC	18773.4124
19	Self-Employed Individuals	16893.5808
17	Qualified Joint-Venture (spouses)	16492.7379
8	Independent Contractors	12319.1589

In [21]:
```
df.sample(4)
```

Out[21]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress
39998	5200198303	2021-01-25	914	PPS	CLASSIC WIRE CUT COMPANY INC.	28210 Constellation Rd
1008561	5116247208	2020-04-27	304	PPP	SHILOH FARMS, LLC	13610 Shiloh Dr.
1344958	8221427002	2020-04-08	563	PPP	WINGRA LEASING LLC	2975 KAPEC ROAD
888185	8943268305	2021-01-30	883	PPS	DYNAMIC BLENDING SPECIALISTS INC	523 E 1750 N Ste 100

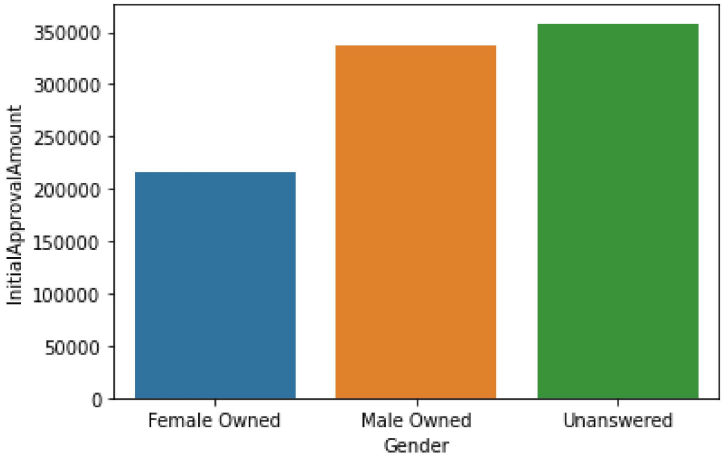
 # loan amounts of Male Owned businesses VS FemaleOwned businesses

In [23]:
```
ln_amount_mf = df.groupby(['Gender'],as_index = False)['InitialApprovalAmount'].mean()
ln_amount_mf
```

Out[23]:

	Gender	InitialApprovalAmount
0	Female Owned	216248.6644
1	Male Owned	337933.7260
2	Unanswered	358404.7470

In [31]:
```
sns.barplot(data = ln_amount_mf ,x='Gender',y='InitialApprovalAmount');
```



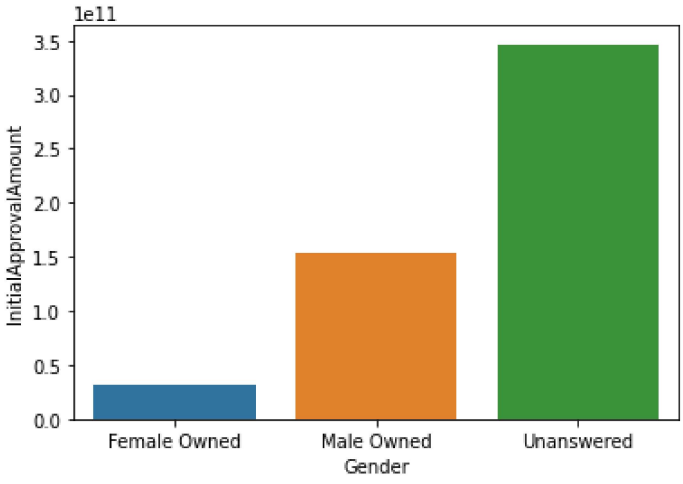
from above graph we can say that Male Owned businesses is higher than the Female Owned businesses

In [34]:
```
ln_amount_mf_sum = df.groupby(['Gender'],as_index = False)['InitialApprovalAmount'].sum(
ln_amount_mf_sum
```

Out[34]:

	Gender	InitialApprovalAmount
0	Female Owned	31707244165.0250
1	Male Owned	153639540911.7350
2	Unanswered	346621832548.6290

In [35]:
```
sns.barplot(data = ln_amount_mf_sum ,x='Gender',y='InitialApprovalAmount');
```



From above plot we can say that the Male Owned businesses have the more loan amount than Female Owned businesses

 # months in which high amount of loans were sanctioned.

In [36]:
```
df.sample(4)
```

Out[36]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress
15185	6612707208	2020-04-28	669	PPP	BRAD HENDRICKS, P.A.	500-C PLEASANT VALLEY DR
52449	7680837208	2020-04-28	942	PPP	MARKON COOPERATIVE	1023 S MAIN ST
352739	8145227009	2020-04-08	562	PPP	SUMMIT HEARTLAND, LLC	3823 W 1800 S
251249	1914017706	2020-05-01	491	PPP	OCEAN BREEZE BAR & GRILL LLC	521 FLAGLER AVE

In [45]:
```
pd.DatetimeIndex(df.DateApproved).month
```

Out[45]:
```
Int64Index([5, 5, 5, 5, 5, 8, 4, 8, 4, 4,
            ...
            3, 4, 1, 5, 4, 5, 3, 5, 3, 3],
           dtype='int64', name='DateApproved', length=1568392)
```

In [50]:
```
x = df.loc[df.InitialApprovalAmount == 10000000.0000,'DateApproved']
```

In [54]:
```
pd.DatetimeIndex(x).month.unique()
```

Out[54]:
```
Int64Index([4, 5, 3, 2, 1, 7, 8, 6], dtype='int64', name='DateApproved')
```

[4, 5, 3, 2, 1, 7, 8, 6] are the months in which high amount of loans were sanctioned.

In [55]:
```
df.LoanNumber.nunique()
```

Out[55]:
```
1568392
```

In [56]:
```
df.shape
```

Out[56]:
```
(1568392, 53)
```

In [57]:
```
df.BorrowerName.nunique()
```

Out[57]:
```
1363120
```

In [72]:
```
pd.DatetimeIndex(df.DateApproved).month.unique().sort_values()
```

Out[72]:
```
Int64Index([1, 2, 3, 4, 5, 6, 7, 8, 12], dtype='int64', name='DateApproved')
```

In [83]:
```
x_1 = df.loc[(pd.DatetimeIndex(df.DateApproved).month == 1) ,['DateApproved','JobsReport
```

In [76]:
```
x_1.shape
```

Out[76]:
```
(153945, 2)
```

In [77]:
```
x_1.sample(3)
```

Out[77]:

	DateApproved	JobsReported
1297113	2021-01-25	3.0000
1270259	2021-01-31	2.0000
1550143	2021-01-29	5.0000

In [88]:
```
pd.DatetimeIndex(x_1.DateApproved).year.unique()
```

Out[88]:
```
Int64Index([2021], dtype='int64', name='DateApproved')
```

In [79]:
```
x_2 = df.loc[pd.DatetimeIndex(df.DateApproved).month == 2 ,['DateApproved','JobsReported
```

In [80]:
```
x_2.shape
```

Out[80]:
```
(178983, 2)
```

In [89]:
```
pd.DatetimeIndex(x_2.DateApproved).year.unique()
```

Out[89]:
```
Int64Index([2021], dtype='int64', name='DateApproved')
```

In [81]:
```
x_2.head()
```

Out[81]:

	DateApproved	JobsReported
75	2021-02-20	92.0000
77	2021-02-05	150.0000
81	2021-02-10	226.0000
89	2021-02-12	221.0000
97	2021-02-07	199.0000

In [7]:
```
gc.collect()
```

Out[7]:
```
0
```

 # number of loans approved in each month

In [90]:
```
df.sample(3)
```

Out[90]:

	LoanNumber	DateApproved	SBAOfficeCode	ProcessingMethod	BorrowerName	BorrowerAddress
1455034	5972068506	2021-03-02	563	PPP	PAUL COX	525 Cornelia St
1495924	8125917203	2020-04-28	563	PPP	JULES PILATES LLC	2160 ATWOOD AVE
16573	5365427008	2020-04-05	669	PPP	WEND-XX INC	2901 ARKANSAS BLVD

In [8]:
```
df.shape
```

Out[8]:
```
(1568392, 53)
```

In [93]:
```
df.LoanNumber.nunique()
```

Out[93]: 1568392

In [95]:
```
# All the loanNumbers are unique
```

In [9]:
```
ln_approved_each_month = pd.DataFrame(df.DateApproved.value_counts()).sort_index()
```

In [10]:
```
ln_approved_each_month.reset_index(level = 0, inplace = True)
```

In [11]:
```
ln_approved_each_month.sample()
```

Out[11]:

	index	DateApproved
210	2021-04-26	3567

In [12]:
```
ln_approved_each_month.rename(columns = {'index':'DateApproved','DateApproved':'loans_ap
```

In [13]:
```
ln_approved_each_month.head(3)
```

Out[13]:

	DateApproved	loans_approved
0	2020-04-03	10548
1	2020-04-04	18432
2	2020-04-05	21795

In [14]:
```
ln_approved_each_month.DateApproved=pd.to_datetime(ln_approved_each_month.DateApproved,e
```

In [15]:
```
ln_approved_each_month.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype         
---  ------          --------------  -----         
 0   DateApproved    244 non-null    datetime64[ns]
 1   loans_approved  244 non-null    int64         
dtypes: datetime64[ns](1), int64(1)
memory usage: 3.9 KB
```

In [21]:
```
ln_approved_each_month['DateApproved'] = ln_approved_each_month['DateApproved'].dt.strft
```

In [22]:
```
ln_approved_each_month.head()
```

Out[22]:

	DateApproved	loans_approved
0	2020-04	10548
1	2020-04	18432
2	2020-04	21795
3	2020-04	31955
4	2020-04	43761

In [25]:
```
data_ts = pd.DataFrame(ln_approved_each_month.groupby('DateApproved')['loans_approved'].
```

In [27]:
```
data_ts
```

Out[27]:

DateApproved	loans_approved
2020-04	723296
2020-05	172086
2020-06	35059
2020-07	15353
2020-08	8392
2020-12	4
2021-01	153945
2021-02	178983
2021-03	146550
2021-04	90212
2021-05	44034
2021-06	477
2021-07	1

In [28]:
```
data_ts.to_csv('data_ts.csv')
```