```
import pandas as pd
In [27]:
         READ PATH PPP = "/Users/jon/Desktop/data-cleaning-project/data/PPP Data up to 15
In [28]:
         READ PATH NAICS = "/Users/jon/Desktop/data-cleaning-project/data/6-digit 2017 Co
          ppp df = pd.read csv(READ PATH PPP)
          naics_df = pd.read_csv(READ_PATH_NAICS)
          naics_df.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 1058 entries, 0 to 1057
         Data columns (total 5 columns):
              Column
                               Non-Null Count Dtype
              2017 NAICS Code 1057 non-null float64
          0
              2017 NAICS Title 1057 non-null
                                              object
          1
              Unnamed: 2
                              1 non-null
                                               object
          2
          3
             Unnamed: 3
                               0 non-null
                                               float64
             Unnamed: 4 0 non-null
                                               float64
         dtypes: float64(3), object(2)
         memory usage: 41.5+ KB
In [29]:
         ppp_df.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 21904 entries, 0 to 21903
         Data columns (total 14 columns):
              Column
                     Non-Null Count Dtype
         ___
                            _____
             LoanAmount 21904 non-null float64
City 21904 non-null object
          0
          1
                          21904 non-null object
          2
              State
                            21904 non-null int64
          3
             Zip
             NAICSCode 21799 non-null float64
BusinessType 21903 non-null object
          4
          5
          6
             RaceEthnicity 21904 non-null object
          7
             Gender 21904 non-null object
             Veteran 21904 non-null object
NonProfit 678 non-null object
          9
          10 JobsReported 19457 non-null float64
          11 DateApproved 21904 non-null object
                            21904 non-null object
          12 Lender
                            21904 non-null object
          13
             CD
         dtypes: float64(3), int64(1), object(10)
         memory usage: 2.3+ MB
In [30]:
         ppp df = ppp df.astype({"NAICSCode": "Int32"})
          ppp df = ppp df.astype({"NAICSCode": "string"})
         ppp_df.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 21904 entries, 0 to 21903
         Data columns (total 14 columns):
          #
                      Non-Null Count Dtype
              Column
                            _____
             LoanAmount
                            21904 non-null float64
          0
                           21904 non-null object
          1
              City
                           21904 non-null object
          2
              State
          3
              Zip
                           21904 non-null int64
              NAICSCode
                           21799 non-null string
             BusinessType 21903 non-null object
          5
              RaceEthnicity 21904 non-null object
          6
                            21904 non-null object
          7
              Gender
                            21904 non-null object
              Veteran
```

```
9 NonProfit
                          678 non-null
                                          object
         10 JobsReported 19457 non-null float64
         11 DateApproved 21904 non-null object
         12 Lender
                          21904 non-null object
         13
                           21904 non-null object
        dtypes: float64(2), int64(1), object(10), string(1)
        memory usage: 2.3+ MB
         naics_df.drop(["Unnamed: 2", "Unnamed: 3", "Unnamed: 4"], axis=1, inplace=True)
In [31]:
         naics df.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 1058 entries, 0 to 1057
        Data columns (total 2 columns):
            Column
                            Non-Null Count Dtype
                             -----
         ____
             2017 NAICS Code 1057 non-null float64
         1
             2017 NAICS Title 1057 non-null object
        dtypes: float64(1), object(1)
        memory usage: 16.7+ KB
In [32]: naics df = naics df.dropna()
         naics_df.info()
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 1057 entries, 1 to 1057
        Data columns (total 2 columns):
                       Non-Null Count Dtype
         # Column
                              _____
         ---
             2017 NAICS Code 1057 non-null float64
             2017 NAICS Title 1057 non-null object
        dtypes: float64(1), object(1)
        memory usage: 24.8+ KB
In [33]: | naics_df = naics_df.astype({"2017 NAICS Code":"int32","2017 NAICS Title":"string
         naics_df = naics_df.astype({"2017 NAICS Code":"string"})
         naics df.info()
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 1057 entries, 1 to 1057
        Data columns (total 2 columns):
         # Column Non-Null Count Dtype
                             _____
         ___
             2017 NAICS Code 1057 non-null string
           2017 NAICS Title 1057 non-null string
        dtypes: string(2)
        memory usage: 24.8 KB
         joined_ppp = ppp_df.merge(
In [34]:
             right=naics df,
             how='left',
             left on='NAICSCode',
             right on='2017 NAICS Code'
         )
        joined ppp.info()
In [35]:
        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 21904 entries, 0 to 21903
        Data columns (total 16 columns):
         # Column
                       Non-Null Count Dtype
         0
             LoanAmount
                             21904 non-null float64
```

```
21904 non-null object
 1
    City
 2
    State
                      21904 non-null object
 3
                      21904 non-null int64
    Zip
                      21799 non-null string
 4
    NAICSCode
 5
    BusinessType
                      21903 non-null object
 6
                      21904 non-null object
    RaceEthnicity
 7
                      21904 non-null
                                      object
    Gender
 8
    Veteran
                      21904 non-null
                                      object
 9
    NonProfit
                      678 non-null
                                      object
                      19457 non-null float64
 10 JobsReported
                      21904 non-null object
 11 DateApproved
 12 Lender
                      21904 non-null object
 13
    CD
                      21904 non-null
                                     object
 14
    2017 NAICS Code
                      21595 non-null
                                      string
 15 2017 NAICS Title 21595 non-null string
dtypes: float64(2), int64(1), object(10), string(3)
memory usage: 2.8+ MB
```

In [36]: | ppp\_df.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 21904 entries, 0 to 21903 Data columns (total 14 columns):

#	Column	Non-Null Count	Dtype		
0	LoanAmount	21904 non-null	float64		
1	City	21904 non-null	object		
2	State	21904 non-null	object		
3	Zip	21904 non-null	int64		
4	NAICSCode	21799 non-null	string		
5	BusinessType	21903 non-null	object		
6	RaceEthnicity	21904 non-null	object		
7	Gender	21904 non-null	object		
8	Veteran	21904 non-null	object		
9	NonProfit	678 non-null	object		
10	JobsReported	19457 non-null	float64		
11	DateApproved	21904 non-null	object		
12	Lender	21904 non-null	object		
13	CD	21904 non-null	object		
dtype	es: float64(2),	int64(1), object	t(10), string(1)		
memory usage: 2.3+ MB					

## In [37]:

joined ppp

## Out[37]:

G€	RaceEthnicity	BusinessType	NAICSCode	Zip	State	City	LoanAmount	
Unansı	Unanswered	Corporation	238220	96813	НІ	HONOLULU	149957.5	0
Unansv	Unanswered	Non-Profit Organization	541990	96814	НІ	HONOLULU	149900.0	1
Male O	Asian	Corporation	722511	96816	НІ	HONOLULU	149800.0	2

	LoanAmount	City	State	Zip	NAICSCode	BusinessType	RaceEthnicity	G€
3	149800.0	HONOLULU	НІ	96815	722511	Corporation	Unanswered	Unansı
4	149700.0	AIEA	HI	96701	621111	Limited Liability Partnership	Unanswered	Unansı
•••			•••					
21899	200.0	EWA BEACH	НІ	96706	531210	Sole Proprietorship	Unanswered	Unansı
21900	117.0	Honolulu	НІ	96814	541922	Sole Proprietorship	Unanswered	Unansv
21901	104.0	Haiku	НІ	96708	561510	Sole Proprietorship	Unanswered	Unansı
21902	89.0	KIHEI	НІ	96753	721199	Sole Proprietorship	Unanswered	Unansı
21903	77.0	HONOLULU	HI	96817	524210	Sole Proprietorship	Unanswered	Unansı

21904 rows × 16 columns

In [38]:	ppp_df	-							
Out[38]:		LoanAmount	City	State	Zip	NAICSCode	BusinessType	RaceEthnicity	Ge
	0	149957.5	HONOLULU	HI	96813	238220	Corporation	Unanswered	Unansv
	1	149900.0	HONOLULU	HI	96814	541990	Non-Profit Organization	Unanswered	Unansı
	2	149800.0	HONOLULU	HI	96816	722511	Corporation	Asian	Male O
	3	149800.0	HONOLULU	HI	96815	722511	Corporation	Unanswered	Unansı
	4	149700.0	AIEA	НІ	96701	621111	Limited Liability Partnership	Unanswered	Unansı
	•••	•••	•••			•••		•••	
	21899	200.0	EWA BEACH	НІ	96706	531210	Sole Proprietorship	Unanswered	Unansv

Gŧ	RaceEthnicity	BusinessType	NAICSCode	Zip	State	City	LoanAmount	
Unansı	Unanswered	Sole Proprietorship	541922	96814	НІ	Honolulu	117.0	21900
Unansı	Unanswered	Sole Proprietorship	561510	96708	НІ	Haiku	104.0	21901
Unansv	Unanswered	Sole Proprietorship	721199	96753	НІ	KIHEI	89.0	21902
Unansv	Unanswered	Sole Proprietorship	524210	96817	НІ	HONOLULU	77.0	21903

21904 rows × 14 columns

