In [27]:
```
import pandas as pd
```

In [28]:
```
READ_PATH_PPP = "/Users/jon/Desktop/data-cleaning-project/data/PPP Data up to 15
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  
---  ------            --------------  -----  
 0   2017 NAICS Code   1057 non-null   float64
 1   2017 NAICS Title  1057 non-null   object 
 2   Unnamed: 2        1 non-null      object 
 3   Unnamed: 3        0 non-null      float64
 4   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  
---  ------          --------------  -----  
 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  float64
 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 
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):
 #   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
dtypes: float64(2), int64(1), object(10), string(1)
memory usage: 2.3+ MB
```

In [31]:
```
naics_df.drop(["Unnamed: 2", "Unnamed: 3", "Unnamed: 4"], axis=1, inplace=True)
naics_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1058 entries, 0 to 1057
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   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):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   2017 NAICS Code   1057 non-null   float64
 1   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
---  ------            --------------  -----
 0   2017 NAICS Code   1057 non-null   string
 1   2017 NAICS Title  1057 non-null   string
dtypes: string(2)
memory usage: 24.8 KB
```

In [34]:
```
joined_ppp = ppp_df.merge(
    right=naics_df,
    how='left',
    left_on='NAICSCode',
    right_on='2017 NAICS Code'
)
```

In [35]:
```
joined_ppp.info()
```

```
<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
```

```
 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
 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
dtypes: float64(2), int64(1), object(10), string(1)
memory usage: 2.3+ MB
```

In [37]:
```
joined_ppp
```

Out[37]:

	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	Unansv
2	149800.0	HONOLULU	HI	96816	722511	Corporation	Asian	Male O

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

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	Unansv
2	149800.0	HONOLULU	HI	96816	722511	Corporation	Asian	Male O
3	149800.0	HONOLULU	HI	96815	722511	Corporation	Unanswered	Unansv
4	149700.0	AIEA	HI	96701	621111	Limited Liability Partnership	Unanswered	Unansv
...	...	...	...	...	...	...	...	
21899	200.0	EWA BEACH	HI	96706	531210	Sole Proprietorship	Unanswered	Unansv

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

21904 rows × 14 columns

In [65]:
```
WRITE_PATH = "/Users/jon/Desktop/data-cleaning-project/data/ppp-naics-joined.csv
joined_ppp.to_csv(WRITE_PATH)
```

In [ ]: