In [1]:
```
'''
CS 418: Final Project
Title - US Census Demographic Data - DATA PROCESSING -Applying changes on test data
Authors: Anusha Sagi, Fatima Kahack, Lydia Tse
Description: The following is code to analyze the US Census Data and preprocess it.
'''
```

In [2]:
```
# Load libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [4]:
```
# Load US Census Demographic Test Data
census_test_data = pd.read_csv('census_test_data.csv')
census_test_data.head()
```

Out[4]:

	CountyId	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	...	Walk	OtherTransp	WorkAtHome	MeanCommute	Employed	Priva
0	1005	Alabama	Barbour County	26201	13976	12225	4.2	45.7	47.8	0.2	...	2.2	1.7	1.3	23.4	8878	
1	1025	Alabama	Clarke County	24625	11649	12976	0.2	53.0	45.7	0.1	...	0.7	2.7	0.9	23.9	7933	
2	1037	Alabama	Coosa County	10955	5474	5481	0.1	65.3	33.2	0.1	...	1.7	0.6	4.2	29.8	3897	
3	1047	Alabama	Dallas County	40755	18809	21946	1.0	27.7	70.2	0.0	...	1.7	1.4	1.2	22.1	14461	
4	1053	Alabama	Escambia County	37621	19326	18295	2.2	60.2	32.2	3.6	...	0.3	0.5	0.8	22.6	12134	

5 rows × 37 columns

In [5]:
```
#Drop IncomeERR, IncomePerCapErr
census_test_data = census_test_data.drop(['IncomeErr', 'IncomePerCapErr'], axis=1)
census_test_data.head()
```

Out[5]:

	CountyId	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	...	Walk	OtherTransp	WorkAtHome	MeanCommute	Employed	Priva
0	1005	Alabama	Barbour County	26201	13976	12225	4.2	45.7	47.8	0.2	...	2.2	1.7	1.3	23.4	8878	
1	1025	Alabama	Clarke County	24625	11649	12976	0.2	53.0	45.7	0.1	...	0.7	2.7	0.9	23.9	7933	
2	1037	Alabama	Coosa County	10955	5474	5481	0.1	65.3	33.2	0.1	...	1.7	0.6	4.2	29.8	3897	
3	1047	Alabama	Dallas County	40755	18809	21946	1.0	27.7	70.2	0.0	...	1.7	1.4	1.2	22.1	14461	
4	1053	Alabama	Escambia County	37621	19326	18295	2.2	60.2	32.2	3.6	...	0.3	0.5	0.8	22.6	12134	

5 rows × 35 columns

In [6]:
```
#Percent Women - Test data
census_test_data['TotalPop'] = census_test_data['TotalPop'].astype(float)
census_test_data['Women'] = census_test_data['Women'].astype(float)
census_test_data['Percent_Women'] = (census_test_data['Women'] / census_test_data['TotalPop']) *100
census_test_data = census_test_data.drop(['Men', 'Women'], axis=1)
census_test_data.head()
```

Out[6]:

	CountyId	State	County	TotalPop	Hispanic	White	Black	Native	Asian	Pacific	...	OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork
0	1005	Alabama	Barbour County	26201.0	4.2	45.7	47.8	0.2	0.6	0.0	...	1.7	1.3	23.4	8878	74.1
1	1025	Alabama	Clarke County	24625.0	0.2	53.0	45.7	0.1	0.5	0.1	...	2.7	0.9	23.9	7933	83.1
2	1037	Alabama	Coosa County	10955.0	0.1	65.3	33.2	0.1	0.0	0.0	...	0.6	4.2	29.8	3897	78.8
3	1047	Alabama	Dallas County	40755.0	1.0	27.7	70.2	0.0	0.4	0.0	...	1.4	1.2	22.1	14461	78.8
4	1053	Alabama	Escambia County	37621.0	2.2	60.2	32.2	3.6	0.4	0.1	...	0.5	0.8	22.6	12134	75.6

5 rows × 34 columns

In [7]:
```
#drop % Pacific - lot of zeros(Other related columns have necessary data)
census_test_data = census_test_data.drop(['Pacific'], axis=1)
census_test_data.head()
```

Out[7]:

	CountyId	State	County	TotalPop	Hispanic	White	Black	Native	Asian	VotingAgeCitizen	...	OtherTransp	WorkAtHome	MeanCommute	Employed	Pr
0	1005	Alabama	Barbour County	26201.0	4.2	45.7	47.8	0.2	0.6	20269	...	1.7	1.3	23.4	8878	
1	1025	Alabama	Clarke County	24625.0	0.2	53.0	45.7	0.1	0.5	19087	...	2.7	0.9	23.9	7933	
2	1037	Alabama	Coosa County	10955.0	0.1	65.3	33.2	0.1	0.0	8989	...	0.6	4.2	29.8	3897	
3	1047	Alabama	Dallas County	40755.0	1.0	27.7	70.2	0.0	0.4	30303	...	1.4	1.2	22.1	14461	
4	1053	Alabama	Escambia County	37621.0	2.2	60.2	32.2	3.6	0.4	29129	...	0.5	0.8	22.6	12134	

5 rows × 33 columns

In [8]:
```
#drop Transit & FamilyWork - lot of zeros(Other related columns have necessary data)
census_test_data = census_test_data.drop(['Transit', 'FamilyWork'], axis=1)
census_test_data.head()
```

Out[8]:

	CountyId	State	County	TotalPop	Hispanic	White	Black	Native	Asian	VotingAgeCitizen	...	Walk	OtherTransp	WorkAtHome	MeanCommute	Employ
0	1005	Alabama	Barbour County	26201.0	4.2	45.7	47.8	0.2	0.6	20269	...	2.2	1.7	1.3	23.4	8
1	1025	Alabama	Clarke County	24625.0	0.2	53.0	45.7	0.1	0.5	19087	...	0.7	2.7	0.9	23.9	7
2	1037	Alabama	Coosa County	10955.0	0.1	65.3	33.2	0.1	0.0	8989	...	1.7	0.6	4.2	29.8	3
3	1047	Alabama	Dallas County	40755.0	1.0	27.7	70.2	0.0	0.4	30303	...	1.7	1.4	1.2	22.1	14
4	1053	Alabama	Escambia County	37621.0	2.2	60.2	32.2	3.6	0.4	29129	...	0.3	0.5	0.8	22.6	12

5 rows × 31 columns

In [9]:
```
#drop Employed column - redundant (We have unemployed %)
census_test_data = census_test_data.drop(['Employed'], axis=1)
census_test_data.head()
```

Out[9]:

	CountyId	State	County	TotalPop	Hispanic	White	Black	Native	Asian	VotingAgeCitizen	...	Carpool	Walk	OtherTransp	WorkAtHome	MeanCommut
0	1005	Alabama	Barbour County	26201.0	4.2	45.7	47.8	0.2	0.6	20269	...	11.1	2.2	1.7	1.3	23.
1	1025	Alabama	Clarke County	24625.0	0.2	53.0	45.7	0.1	0.5	19087	...	11.9	0.7	2.7	0.9	23.
2	1037	Alabama	Coosa County	10955.0	0.1	65.3	33.2	0.1	0.0	8989	...	9.7	1.7	0.6	4.2	29.
3	1047	Alabama	Dallas County	40755.0	1.0	27.7	70.2	0.0	0.4	30303	...	8.9	1.7	1.4	1.2	22.
4	1053	Alabama	Escambia County	37621.0	2.2	60.2	32.2	3.6	0.4	29129	...	7.0	0.3	0.5	0.8	22.

5 rows × 30 columns

In [10]:
```
#Arranging columns
census_test_data = census_test_data[['CountyId', 'State', 'County', 'TotalPop', 'Percent_Women', 'Hispanic', 'White', 'Black', 'Native', 'Asian', 'VotingAgeCitizen', 'Income', 'IncomePerCap', 'Professional', 'Service', 'Office', 'Construction', 'Production', 'Drive', 'Carpool', 'Walk', 'OtherTransp', 'WorkAtHome', 'MeanCommute', 'PrivateWork', 'PublicWork', 'SelfEmployed', 'Unemployment']]
census_test_data.head()
```

Out[10]:

	CountyId	State	County	TotalPop	Percent_Women	Hispanic	White	Black	Native	Asian	...	Drive	Carpool	Walk	OtherTransp	WorkAtHome	MeanCo
0	1005	Alabama	Barbour County	26201.0	46.658524	4.2	45.7	47.8	0.2	0.6	...	83.4	11.1	2.2	1.7	1.3	
1	1025	Alabama	Clarke County	24625.0	52.694416	0.2	53.0	45.7	0.1	0.5	...	83.7	11.9	0.7	2.7	0.9	
2	1037	Alabama	Coosa County	10955.0	50.031949	0.1	65.3	33.2	0.1	0.0	...	83.2	9.7	1.7	0.6	4.2	
3	1047	Alabama	Dallas County	40755.0	53.848608	1.0	27.7	70.2	0.0	0.4	...	86.6	8.9	1.7	1.4	1.2	
4	1053	Alabama	Escambia County	37621.0	48.629755	2.2	60.2	32.2	3.6	0.4	...	91.3	7.0	0.3	0.5	0.8	

5 rows × 28 columns

In [11]:
```
#drop Native, Asian - Not important
census_test_data = census_test_data.drop(['Native', 'Asian'], axis=1)
census_test_data.head()
```

Out[11]:

	CountyId	State	County	TotalPop	Percent_Women	Hispanic	White	Black	VotingAgeCitizen	Income	...	Drive	Carpool	Walk	OtherTransp	WorkAtHo
0	1005	Alabama	Barbour County	26201.0	46.658524	4.2	45.7	47.8	20269	33368	...	83.4	11.1	2.2	1.7	
1	1025	Alabama	Clarke County	24625.0	52.694416	0.2	53.0	45.7	19087	33827	...	83.7	11.9	0.7	2.7	
2	1037	Alabama	Coosa County	10955.0	50.031949	0.1	65.3	33.2	8989	34792	...	83.2	9.7	1.7	0.6	
3	1047	Alabama	Dallas County	40755.0	53.848608	1.0	27.7	70.2	30303	30065	...	86.6	8.9	1.7	1.4	
4	1053	Alabama	Escambia County	37621.0	48.629755	2.2	60.2	32.2	29129	35026	...	91.3	7.0	0.3	0.5	

5 rows × 26 columns

In [12]:
```
#drop Office, Construction - Not important
census_test_data = census_test_data.drop(['Office', 'Construction'], axis=1)
census_test_data.head()
```

Out[12]:

	CountyId	State	County	TotalPop	Percent_Women	Hispanic	White	Black	VotingAgeCitizen	Income	...	Drive	Carpool	Walk	OtherTransp	WorkAtHo
0	1005	Alabama	Barbour County	26201.0	46.658524	4.2	45.7	47.8	20269	33368	...	83.4	11.1	2.2	1.7	
1	1025	Alabama	Clarke County	24625.0	52.694416	0.2	53.0	45.7	19087	33827	...	83.7	11.9	0.7	2.7	
2	1037	Alabama	Coosa County	10955.0	50.031949	0.1	65.3	33.2	8989	34792	...	83.2	9.7	1.7	0.6	
3	1047	Alabama	Dallas County	40755.0	53.848608	1.0	27.7	70.2	30303	30065	...	86.6	8.9	1.7	1.4	
4	1053	Alabama	Escambia County	37621.0	48.629755	2.2	60.2	32.2	29129	35026	...	91.3	7.0	0.3	0.5	

5 rows × 24 columns

In [13]:
```
#drop Drive, Walk, OtherTransp - Not important
census_test_data = census_test_data.drop(['Drive', 'Walk', 'OtherTransp'], axis=1)
census_test_data.head()
```

Out[13]:

	CountyId	State	County	TotalPop	Percent_Women	Hispanic	White	Black	VotingAgeCitizen	Income	...	Professional	Service	Production	Carpool	W
0	1005	Alabama	Barbour County	26201.0	46.658524	4.2	45.7	47.8	20269	33368	...	25.0	16.8	24.1	11.1	
1	1025	Alabama	Clarke County	24625.0	52.694416	0.2	53.0	45.7	19087	33827	...	21.6	14.3	25.6	11.9	
2	1037	Alabama	Coosa County	10955.0	50.031949	0.1	65.3	33.2	8989	34792	...	17.6	23.2	20.9	9.7	
3	1047	Alabama	Dallas County	40755.0	53.848608	1.0	27.7	70.2	30303	30065	...	26.7	18.2	25.3	8.9	
4	1053	Alabama	Escambia County	37621.0	48.629755	2.2	60.2	32.2	29129	35026	...	24.6	21.2	18.3	7.0	

5 rows × 21 columns

In [14]:
```
#drop VotingAgeCitizen- Not important
census_test_data = census_test_data.drop(['VotingAgeCitizen'], axis=1)
census_test_data.head()
```

Out[14]:

	CountyId	State	County	TotalPop	Percent_Women	Hispanic	White	Black	Income	IncomePerCap	Professional	Service	Production	Carpool	WorkAt
0	1005	Alabama	Barbour County	26201.0	46.658524	4.2	45.7	47.8	33368	17561	25.0	16.8	24.1	11.1	
1	1025	Alabama	Clarke County	24625.0	52.694416	0.2	53.0	45.7	33827	20765	21.6	14.3	25.6	11.9	
2	1037	Alabama	Coosa County	10955.0	50.031949	0.1	65.3	33.2	34792	20342	17.6	23.2	20.9	9.7	
3	1047	Alabama	Dallas County	40755.0	53.848608	1.0	27.7	70.2	30065	18248	26.7	18.2	25.3	8.9	
4	1053	Alabama	Escambia County	37621.0	48.629755	2.2	60.2	32.2	35026	18164	24.6	21.2	18.3	7.0	

In [15]:
```
census_test_data.to_csv("test_dp_output.csv")
census_test_data
```

Out[15]:

	CountyId	State	County	TotalPop	Percent_Women	Hispanic	White	Black	Income	IncomePerCap	Professional	Service	Production	Carpool	Work/
0	1005	Alabama	Barbour County	26201.0	46.658524	4.2	45.7	47.8	33368	17561	25.0	16.8	24.1	11.1	
1	1025	Alabama	Clarke County	24625.0	52.694416	0.2	53.0	45.7	33827	20765	21.6	14.3	25.6	11.9	
2	1037	Alabama	Coosa County	10955.0	50.031949	0.1	65.3	33.2	34792	20342	17.6	23.2	20.9	9.7	
3	1047	Alabama	Dallas County	40755.0	53.848608	1.0	27.7	70.2	30065	18248	26.7	18.2	25.3	8.9	
4	1053	Alabama	Escambia County	37621.0	48.629755	2.2	60.2	32.2	35026	18164	24.6	21.2	18.3	7.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
788	72107	Puerto Rico	Orocovis Municipio	21906.0	49.826532	99.7	0.3	0.0	14296	7326	26.2	17.1	12.1	6.4	
789	72113	Puerto Rico	Ponce Municipio	148863.0	51.785870	99.3	0.6	0.1	16561	10775	32.0	21.2	11.0	6.6	
790	72129	Puerto Rico	San Lorenzo Municipio	38689.0	51.216108	99.9	0.0	0.0	18589	10497	32.0	20.4	12.1	7.8	
791	72141	Puerto Rico	Utuado Municipio	30209.0	51.385349	99.5	0.4	0.1	15931	8140	24.7	24.4	9.0	8.1	
792	72151	Puerto Rico	Yabucoa Municipio	35025.0	51.508922	99.9	0.1	0.0	15586	8672	27.7	26.0	16.0	9.2	

793 rows × 20 columns