Assignment 2: Data Preparation

List all the column names of Data Frame

In [1]:

```python
import pandas as pd
file_path = '2000_acs_sample.dta'
df = pd.read_stata(file_path)
print(df.columns)
```

```
Index(['year', 'datanum', 'serial', 'hhwt', 'gq', 'us2000c_serialno', 'pernum',
       'perwt', 'us2000c_pnum', 'us2000c_sex', 'us2000c_age', 'us2000c_hispan',
       'us2000c_race1', 'us2000c_marstat', 'us2000c_educ', 'us2000c_inctot'],
      dtype='object')
```

List all the columns that have unique values: Apply unique method to each column and check whether its length = 1.

In [2]:

```python
dropunique = []
for (colName, colData) in df.iteritems():
    if len(colData.unique()) == 1:
        print(colName)
        dropunique.append(colName)
```

```
year
datanum
hhwt
perwt
```

Drop all columns that have unique values: see: Drop Columns method

In [3]:

```python
df = df.drop(dropunique, axis=1)
```

Additionally drop the following columns: 'us2000c_pnum', 'us2000c_serialno'

In [4]:

```python
df = df.drop(['us2000c_pnum','us2000c_serialno'], axis=1)
```

Replace the column names as suggested below

serial by household, pernum by person, us2000c_sex by sex, us2000c_age by age, us2000c_hispan by hispanic, us2000c_race1 by race, us2000c_marstat by marital_status, us2000c_educ by edu, us2000c_inctot by income

In [5]:

```python
df = df.rename(columns={'serial':'household',
'pernum':'person',
'us2000c_sex':'sex',
'us2000c_age':'age',
'us2000c_hispan':'hispanic',
'us2000c_race1':'race',
'us2000c_marstat':'marital_status',
'us2000c_educ':'edu',
'us2000c_inctot':'income'})
```

Print the information/summary of the columns of the resulting dataframe using info method of the data frame.

In [6]:

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28172 entries, 0 to 28171
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype   
---  ------          --------------  -----   
 0   household       28172 non-null  float64 
 1   gq              28172 non-null  category
 2   person          28172 non-null  int16   
 3   sex             28172 non-null  object  
 4   age             28172 non-null  object  
 5   hispanic        28172 non-null  object  
 6   race            28172 non-null  object  
 7   marital_status  28172 non-null  object  
 8   edu             28172 non-null  object  
 9   income          28172 non-null  object  
dtypes: category(1), float64(1), int16(1), object(7)
memory usage: 2.0+ MB
None
```

Change the type of income column to number: See how to convert from object type to a numeric type. Links to an external site.(Note: you may need errors="coerce" option.

In [7]:

```
df['income'] = pd.to_numeric(df['income'],errors='coerce')
```

Replace the value in columns sex and marital_status by the actual value listed in the associated meta file.

In [8]:

```
df['sex'].replace('1','Male',inplace=True)
df['sex'].replace('2','Female',inplace=True)

df['marital_status'].replace('1','Now married',inplace=True)
df['marital_status'].replace('2','Widowed',inplace=True)
df['marital_status'].replace('3','Divorced',inplace=True)
df['marital_status'].replace('4','Separated',inplace=True)
df['marital_status'].replace('5','Never married (includes under 15 years)',inplace=True)
```

Replace the NA values in the income column by the mode value of the column.

In [9]:

```
df = df.fillna({'income' : df['income'].mode()[0]})
```

Print the resulting data frame.

In [10]:

```
print(df)
```

```
       household                             gq  person     sex age  \
0           37.0  Households under 1970 definition       1  Female  20
1           37.0  Households under 1970 definition       2  Female  19
2           37.0  Households under 1970 definition       3  Female  19
3          241.0  Households under 1970 definition       1  Female  50
4          242.0  Households under 1970 definition       1  Female  29
...          ...                               ...     ...     ...  ..
28167  1236624.0  Households under 1970 definition       1    Male  29
28168  1236624.0  Households under 1970 definition       2  Female  26
28169  1236756.0  Households under 1970 definition       1  Female  58
28170  1236756.0  Households under 1970 definition       2    Male  61
28171  1236779.0  Households under 1970 definition       1    Male  30

      hispanic race                           marital_status edu   income
0           01    1  Never married (includes under 15 years)  11  10000.0
1           01    1  Never married (includes under 15 years)  11   5300.0
2           01    2  Never married (includes under 15 years)  11   4700.0
3           01    1  Never married (includes under 15 years)  14  32500.0
4           01    1  Never married (includes under 15 years)  13  30000.0
...        ...  ...                                      ...  ..      ...
28167       01    1                              Now married  11  50100.0
28168       01    1                              Now married  09  12000.0
28169       01    1                              Now married  14  69800.0
28170       01    1                              Now married  14  40800.0
28171       01    3                                 Divorced  09  22110.0

[28172 rows x 10 columns]
```