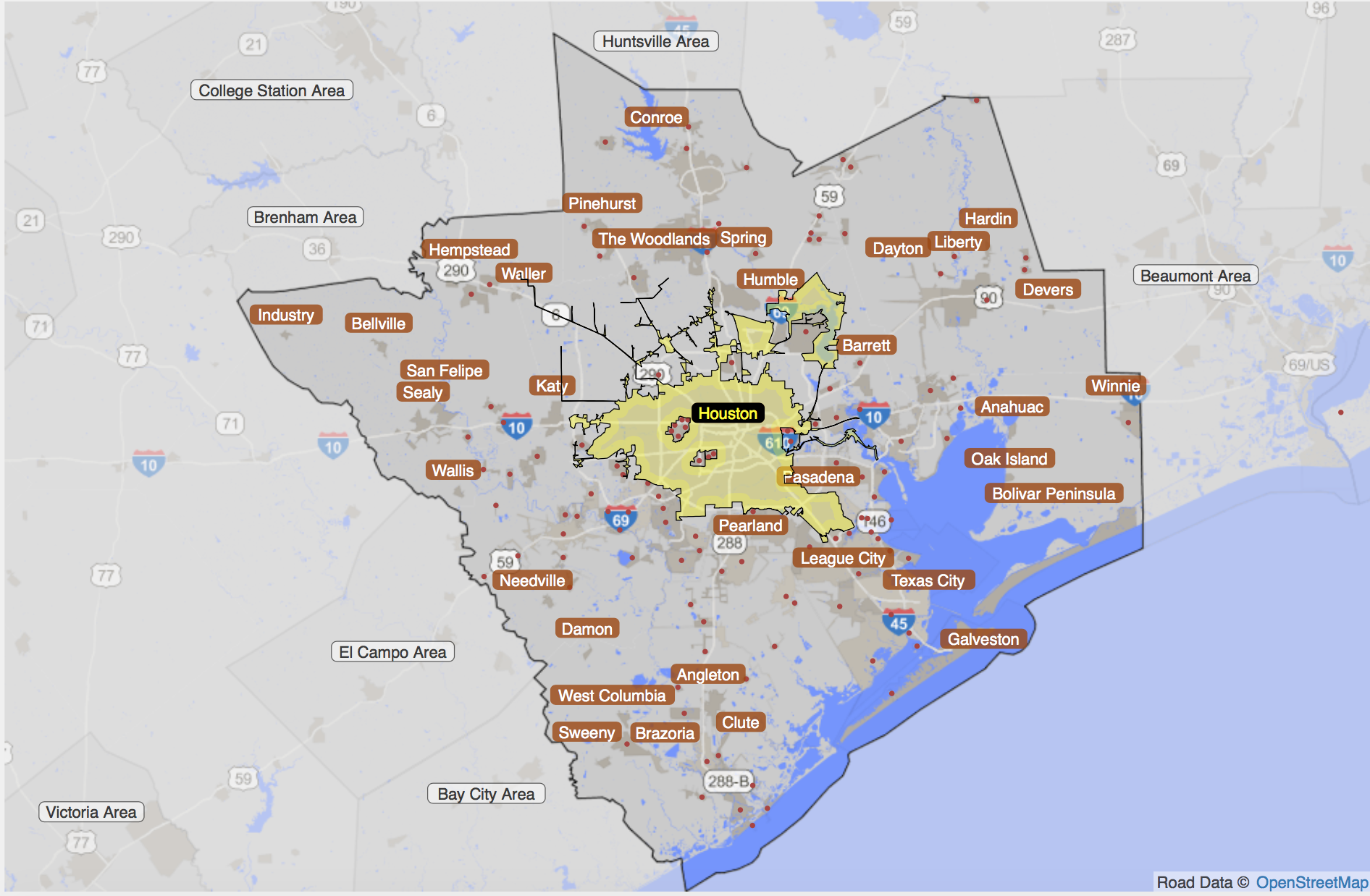
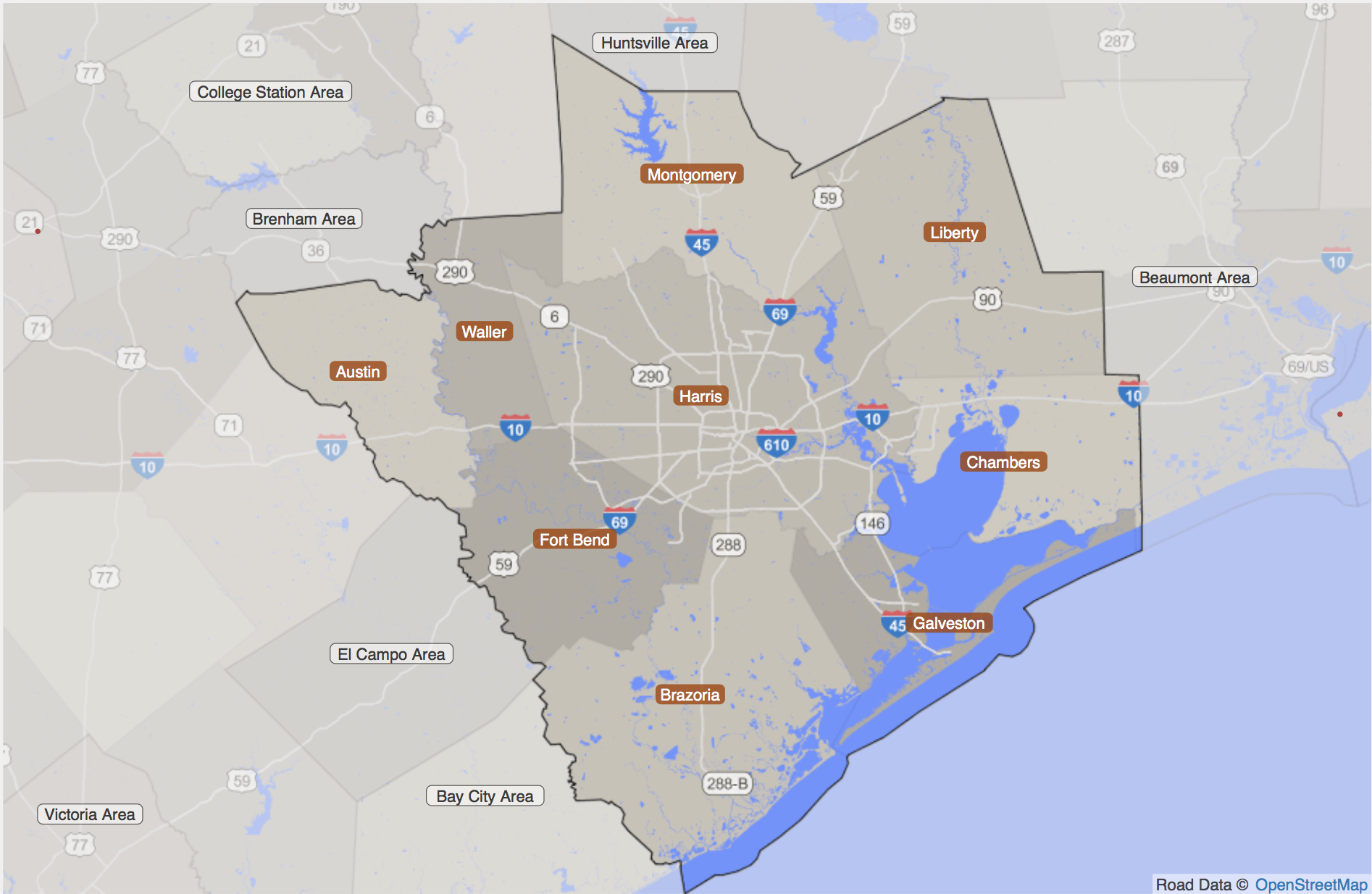
Map of Houston

1. Houston city
2. Houston area (Metro Area)

<http://statisticalatlas.com/metro-area/Texas/Houston/Overview>





The total population in Houston area is 6,063,540, and there are 2,068,518 households.

Houston area (Metro area) has 9 counties, and they are:

Montgomery (339), Liberty (291), Chambers (071), Galveston (167), Harris (201), Waller (473), Austin (015), Fort Bend (157), Brazoria (039)

**SAMPLE DATA FROM PUMS**

Map census tract to puma, filename: “2010\_Census\_Tract\_to\_2010\_PUMA.txt”

We will get PUMA which is Houston metro area:

'05000' '04801' '04802' '04803' '04400' '04901' '04902' '04903' '04904'

'04905' '04701' '04702' '04601' '04602' '04603' '04604' '04605' '04606'

'04607' '04608' '04609' '04610' '04611' '04612' '04613' '04614' '04615'

'04616' '04617' '04618' '04619' '04620' '04621' '04622' '04623' '04624'

'04625' '04626' '04627' '04628' '04629' '04630' '04631' '04632' '04633'

'04634' '04635' '04636' '04637' '04638' '04501' '04502' '04503' '04504'

All related variables in PUMS person data:

1. SERIALNO: Housing unit/GQ person serial number
2. SPORDER: Person number
3. PUMA
4. AGEP: AGE
5. COW: Class of worker
6. ESR: Employment status recode
7. JWMNP: Travel time to work
8. JWTR: Means of transportation to work
9. MAR: Marital status
10. RELP: Relationship
11. SEX
12. SCHL: Education attainment
13. ESR: Employment status
14. PINCP: Total person’s income
15. ADJINC: Adjustment factor for income and earnings dollar amounts

All related variable in PUMS household data:

1. SERIALNO: serial number
2. PUMA
3. NP: number of person records
4. VEH: vehicles available
5. HHT: household/family type
6. HINCP: household income (more complete than family income)
7. HUPAC: HH presence and age of children
8. WIF: Workers in family during the past 12 months

**AGGREGATE DATA IN COUNTY LEVEL**

<https://factfinder.census.gov>

Choose “Advanced Search”, using option “Geographies” -> choose “County”

Marginal distribution available in County level:

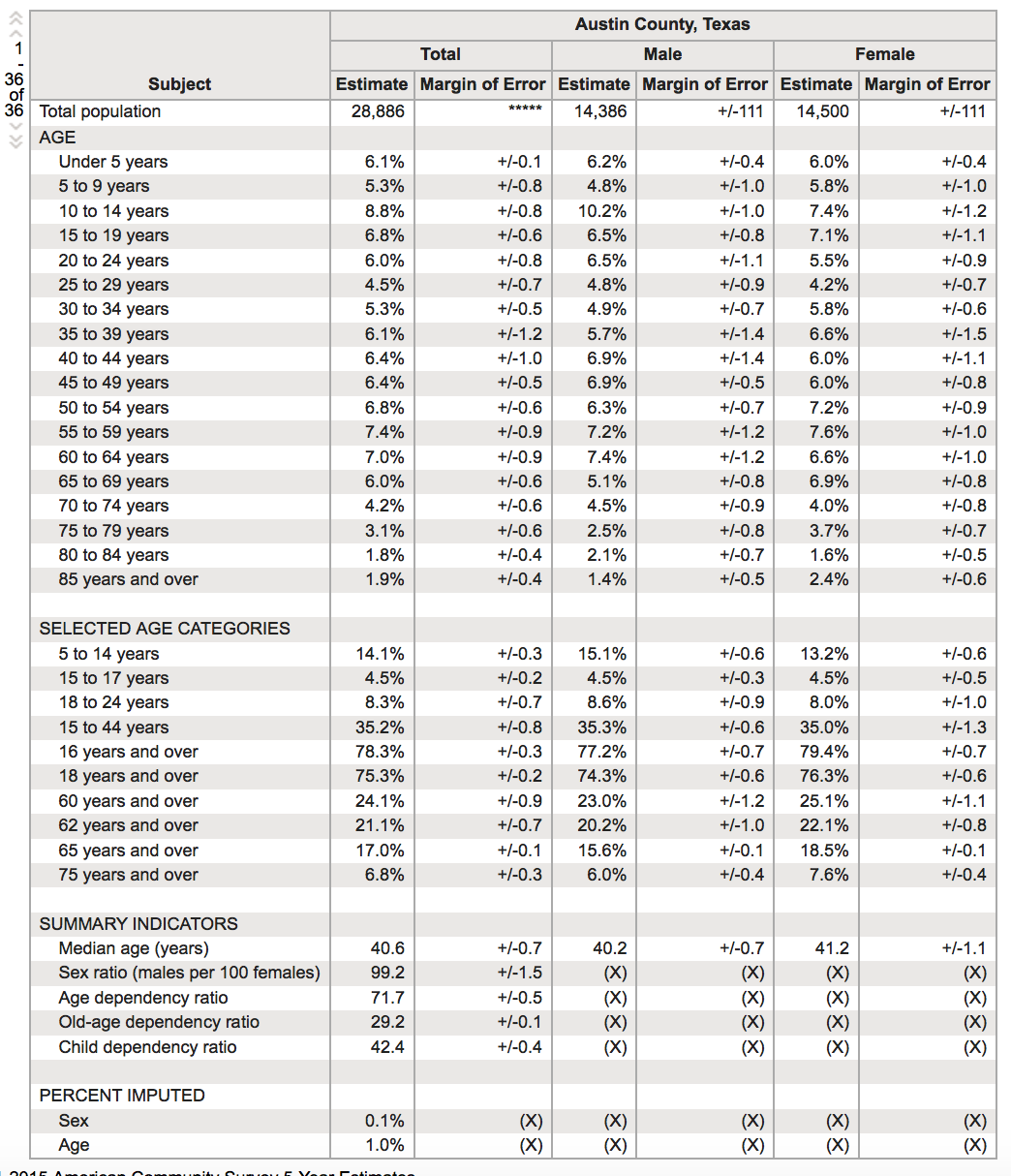
1. Age
2. Sex
3. Marital status
4. Education attainment
5. Individual income in the past 12 month
6. Commuting characteristics by sex: means of transportation to work; place of work; time leaving home to go to work; travel time to work; vehicle available;

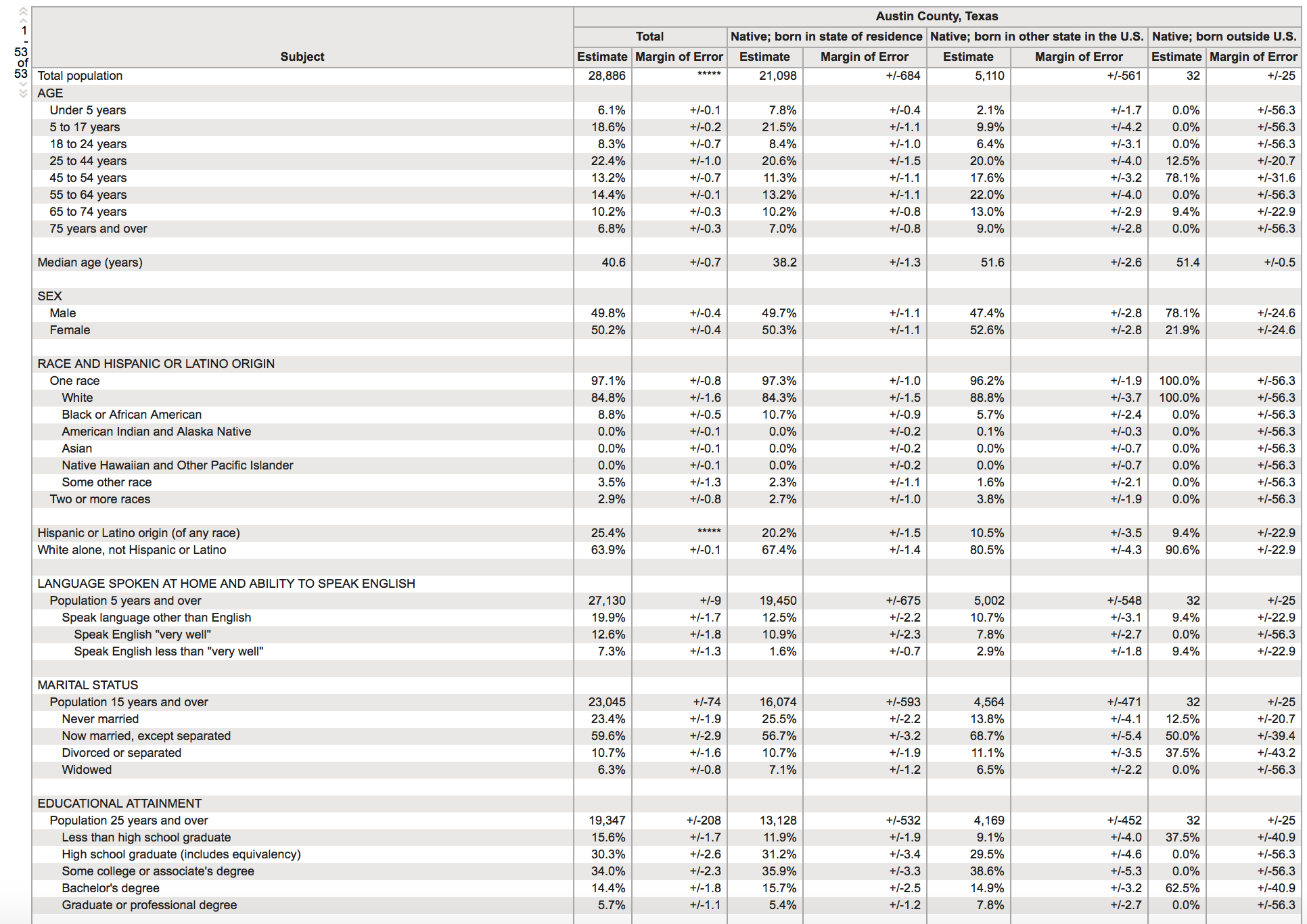
Problem:

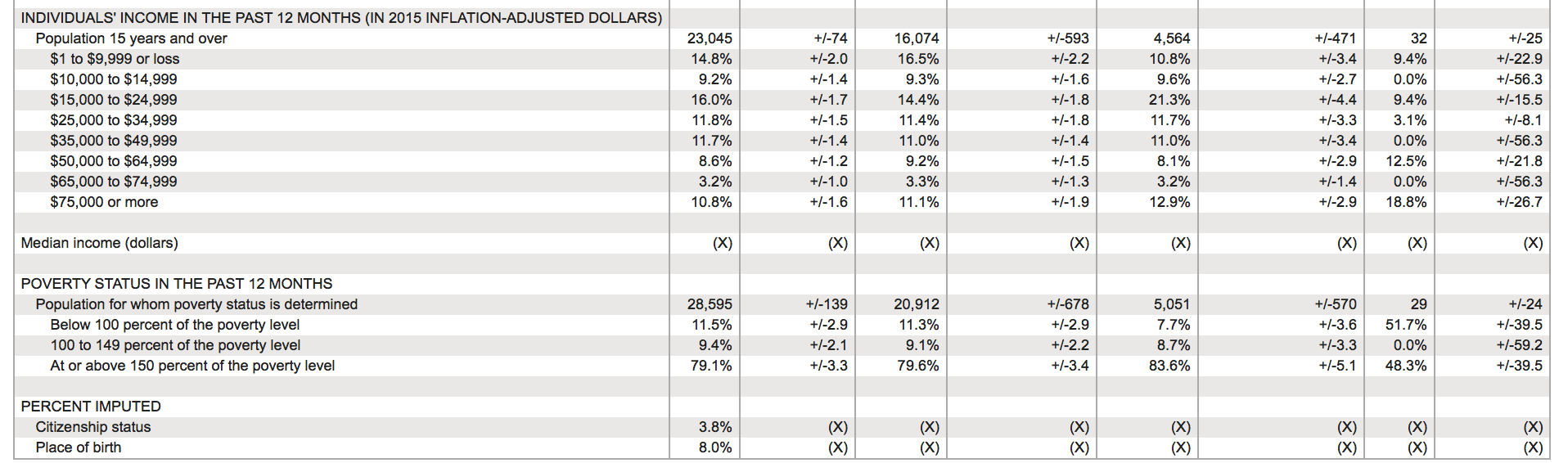
1. In aggregate data, some variables like education attainment, income are only applicable to a subset of population. For example, the income distribution is only for “population 15 years and over”. MAYBE WE CAN CREATE A NEW CATEGORY FOR THE REMAINING
2. To include more characteristics to describe a person than what we specified in “Marginal distribution available in county level”, we will face the problem of time discrepancy. Marginal distribution of different variables will from data in different year.
3. Household aggregate data only in 2010, while person aggregate data is in 2015

Steps in data processing and IPF

1. Get marginal distribution of variables that we are interested in from Census Bureau
2. Processing the data format of marginal distributions
3. Get sample data from PUMS, select sample data in our metro area by mapping census tract to county
4. Select variables in sample data, categorize sample data based on ‘PUMS codebook’ and categories in marginal distributions
5. Generate joint distribution of sample data which means get the frequency of each kind of combination of characteristics in sample data
6. Use joint distribution of sample data and marginal distribution as input to IPF
7. Use output joint distribution of IPF to draw population as synthetic population







**Land Use:**

Aggregate data for population