Data on housing

The aim was to get broad estimates of tenure at the county level and at the CBSA level.

Two main sources were used :
- the American Community Survey 5-Year Estimates of the U.S Census Bureau
- the Picture of Subsidized Households survey of the US Department of Housing and Urban Development

The American Community Survey

The 2012-2016 version of the American Community Survey (ACS) was used. [1] The variables of interest are the following :
- occupied housing units by tenure
- total population in occupied housing units by tenure

Picture of Subsidized Households

The survey was conducted in 2012. Even though data on multiple programs were available, only data on "Public Housing" were kept.

The variable of interest are :
- Number of units under contract for federal subsidy and available for occupancy
- Occupied units as the % of units available
- Total number of people

Data at the county level

The American Community Survey

Ocupied housing units by tenure

This first data set provides information on tenure. The available variables are :
GEO.id2 Id2
GEO.display-label Geography
HD01_VD01 Estimate; Total:
HD02_VD01 Margin of Error; Total:
HD01_VD02 Estimate; Total: - Owner occupied
HD02_VD02 Margin of Error; Total: - Owner occupied
HD01_VD03 Estimate; Total: - Renter occupied
HD02_VD03 Margin of Error; Total: - Renter occupied

I only used the estimates and ignored the margins of error.

```
data_unit_county <- data_unit_county %>%
  mutate(
    pct_owner_unit = HD01_VD02 / HD01_VD01 * 100, # percentage of owner occupied units
```

---

[1]The 2012-2016 American Community Survey (ACS) data generally reflect the February 2013 Office of Management and Budget (OMB) definitions of metropolitan and micropolitan statistical areas; in certain instances the names, codes, and boundaries of the principal cities shown in ACS tables may differ from the OMB definitions due to differences in the effective dates of the geographic entities. Estimates of urban and rural population, housing units, and characteristics reflect boundaries of urban areas defined based on Census 2010 data. As a result, data for urban and rural areas from the ACS do not necessarily reflect the results of ongoing urbanization.

```r
    pct_renter_unit = HD01_VD03 / HD01_VD01 * 100, # percentage of renter occupied units
    total_unit = pct_renter_unit + pct_owner_unit, # check
    GEOID = str_sub(GEO.id, -5, -1), # for the join
  )
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```r
head(data_unit_county)
```

```
## # A tibble: 6 x 13
##   GEO.id  GEO.id2 `GEO.display-la~ HD01_VD01 HD02_VD01 HD01_VD02 HD02_VD02
##   <chr>     <int> <chr>                <int>     <int>     <int>     <int>
## 1 050000~    1001 Autauga County,~     20800       391     15218       548
## 2 050000~    1003 Baldwin County,~     75149      1285     53905      1142
## 3 050000~    1005 Barbour County,~      9122       286      5829       257
## 4 050000~    1007 Bibb County, Al~      7048       352      5119       303
## 5 050000~    1009 Blount County, ~     20619       403     16254       507
## 6 050000~    1011 Bullock County,~      3556       210      2600       222
## # ... with 6 more variables: HD01_VD03 <int>, HD02_VD03 <int>,
## #   pct_owner_unit <dbl>, pct_renter_unit <dbl>, total_unit <dbl>,
## #   GEOID <chr>
```

Total population in ocupied housing by tenure

This second data set provides information on the total population in occupied housing units by tenure by year householder moved into unit. However, only the broad estimates, which ignored the time dimension, were kept so as to remain consistent with the 2012-2016 data above.

The variables of interest are :
GEO.id2 Id2
GEO.display-label Geography
HD01_VD01 Estimate; Total population in occupied housing units
HD02_VD01 Margin of Error; Total population in occupied housing units
HD01_VD02 Estimate; Owner occupied
HD02_VD02 Margin of Error; Owner occupied
HD01_VD09 Estimate; Renter occupied
HD02_VD09 Margin of Error; Renter occupied

Again, the "margins of error" were not taken into account.

```r
data_pop_county <- data_pop_county %>%
  select(c(1:7), c(20:21)) %>%
  mutate(
    pct_owner_pop = HD01_VD02 / HD01_VD01 * 100, # percentage of owner population
    pct_renter_pop = HD01_VD09 / HD01_VD01 * 100, # percentage of renter population
    total_pop = pct_renter_pop + pct_owner_pop, # check
    GEOID = str_sub(GEO.id, -5, -1), # for the join
  )
```

```r
head(data_pop_county)
```

```
## # A tibble: 6 x 13
##   GEO.id  GEO.id2 `GEO.display-la~ HD01_VD01 HD02_VD01 HD01_VD02 HD02_VD02
##   <chr>     <int> <chr>                <int> <chr>         <int>     <int>
## 1 050000~    1001 Autauga County,~     54559 163           40133      1559
## 2 050000~    1003 Baldwin County,~    196599 443          141084      2408
```

```
## 3 050000~     1005 Barbour County,~     23682 244            14939         617
## 4 050000~     1007 Bibb County, Al~     20571 190            14450         824
## 5 050000~     1009 Blount County, ~     57152 131            46037        1169
## 6 050000~     1011 Bullock County,~     10111 139             7258         486
## # ... with 6 more variables: HD01_VD09 <int>, HD02_VD09 <int>,
## #   pct_owner_pop <dbl>, pct_renter_pop <dbl>, total_pop <dbl>,
## #   GEOID <chr>
```

```r
data_ACS_county <- left_join(data_pop_county, data_unit_county, by = c("GEOID", "GEO.id2", "GEO.id", "GE

head(data_ACS_county)
```

```
## # A tibble: 6 x 22
##   GEO.id    GEO.id2 `GEO.display-labe~ HD01_VD01.x HD02_VD01.x HD01_VD02.x
##   <chr>       <int> <chr>                    <int> <chr>             <int>
## 1 0500000U~    1001 Autauga County, A~       54559 163               40133
## 2 0500000U~    1003 Baldwin County, A~      196599 443              141084
## 3 0500000U~    1005 Barbour County, A~       23682 244               14939
## 4 0500000U~    1007 Bibb County, Alab~       20571 190               14450
## 5 0500000U~    1009 Blount County, Al~       57152 131               46037
## 6 0500000U~    1011 Bullock County, A~       10111 139                7258
## # ... with 16 more variables: HD02_VD02.x <int>, HD01_VD09 <int>,
## #   HD02_VD09 <int>, pct_owner_pop <dbl>, pct_renter_pop <dbl>,
## #   total_pop <dbl>, GEOID <chr>, HD01_VD01.y <int>, HD02_VD01.y <int>,
## #   HD01_VD02.y <int>, HD02_VD02.y <int>, HD01_VD03 <int>,
## #   HD02_VD03 <int>, pct_owner_unit <dbl>, pct_renter_unit <dbl>,
## #   total_unit <dbl>
```

There are 3,144 counties in the US (including Alaska and Puerto Rico). In the data, there are 3,142 observations.

Picture of Subsidized Households

```r
data_county <- data_county %>%
  filter(program_label == "Public Housing" & !(is.na(total_units))) %>%
  select(states, entities, total_units, pct_occupied, people_per_unit, people_total, state) %>%
  mutate(
    GEOID = str_sub(entities, -5, -1)
  ) %>%
  filter(!(GEOID == "78999") & !(GEOID == "66999")) # removes Guam and VI

head(data_county)
```

```
## # A tibble: 6 x 8
##   states  entities   total_units pct_occupied people_per_unit people_total
##   <chr>   <chr>            <int>        <int>           <dbl>        <dbl>
## 1 AK Ala~ AK Anchor~         508           95            2.54        1218.
## 2 AK Ala~ AK Bethel~         117           97            4.04         456.
## 3 AK Ala~ AK Fairba~         165           98            2.88         462.
## 4 AK Ala~ AK Juneau~         207           97            2.35         462.
## 5 AK Ala~ AK Ketchi~          73          100            1.81         132.
## 6 AK Ala~ AK Kodiak~          40          100            2.72         109.
## # ... with 2 more variables: state <chr>, GEOID <chr>
```

The data set is only made of 2,010 observations.

Final data set for county level

```
data_combined_county <- full_join(data_ACS_county, data_county, by = ("GEOID"))
head(data_combined_county)
```

```
## # A tibble: 6 x 29
##   GEO.id    GEO.id2 `GEO.display-labe~ HD01_VD01.x HD02_VD01.x HD01_VD02.x
##   <chr>       <int> <chr>                    <int> <chr>             <int>
## 1 0500000U~    1001 Autauga County, A~       54559 163               40133
## 2 0500000U~    1003 Baldwin County, A~      196599 443              141084
## 3 0500000U~    1005 Barbour County, A~       23682 244               14939
## 4 0500000U~    1007 Bibb County, Alab~       20571 190               14450
## 5 0500000U~    1009 Blount County, Al~       57152 131               46037
## 6 0500000U~    1011 Bullock County, A~       10111 139                7258
## # ... with 23 more variables: HD02_VD02.x <int>, HD01_VD09 <int>,
## #   HD02_VD09 <int>, pct_owner_pop <dbl>, pct_renter_pop <dbl>,
## #   total_pop <dbl>, GEOID <chr>, HD01_VD01.y <int>, HD02_VD01.y <int>,
## #   HD01_VD02.y <int>, HD02_VD02.y <int>, HD01_VD03 <int>,
## #   HD02_VD03 <int>, pct_owner_unit <dbl>, pct_renter_unit <dbl>,
## #   total_unit <dbl>, states <chr>, entities <chr>, total_units <int>,
## #   pct_occupied <int>, people_per_unit <dbl>, people_total <dbl>,
## #   state <chr>
```

Data at the Core-base statistical areas level

The American Community Survey

Ocupied housing units by tenure

```
data_unit_CBSA <- data_unit_CBSA %>%
  mutate(
    pct_owner_unit = HD01_VD02 / HD01_VD01 * 100, # percentage of owner occupied units
    pct_renter_unit = HD01_VD03 / HD01_VD01 * 100, # percentage of renter occupied units
    total_unit = pct_renter_unit + pct_owner_unit, # check
    GEOID = str_sub(GEO.id, -5, -1), # for the join
  )
```

```
head(data_unit_CBSA)
```

```
## # A tibble: 6 x 13
##   GEO.id   GEO.id2 `GEO.display-la~ HD01_VD01 HD02_VD01 HD01_VD02 HD02_VD02
##   <chr>      <int> <chr>                <int>     <int>     <int>     <int>
## 1 310M30~    10100 Aberdeen, SD Mi~     17721       332     12171       361
## 2 310M30~    10140 Aberdeen, WA Mi~     27472       501     18430       526
## 3 310M30~    10180 Abilene, TX Met~     60308       718     37194       684
## 4 310M30~    10220 Ada, OK Micro A~     14625       190      9389       255
## 5 310M30~    10300 Adrian, MI Micr~     37856       517     29398       537
## 6 310M30~    10420 Akron, OH Metro~    283472      1265    187951      1630
## # ... with 6 more variables: HD01_VD03 <int>, HD02_VD03 <int>,
## #   pct_owner_unit <dbl>, pct_renter_unit <dbl>, total_unit <dbl>,
## #   GEOID <chr>
```

Total population in ocupied housing by tenure

```
data_pop_CBSA <- data_pop_CBSA %>%
  select(c(1:7), c(20:21)) %>%
  mutate(
    pct_owner_pop = HD01_VD02 / HD01_VD01 * 100, # percentage of owner population
    pct_renter_pop = HD01_VD09 / HD01_VD01 * 100, # percentage of renter population
    total_pop = pct_renter_pop + pct_owner_pop, # check
    GEOID = str_sub(GEO.id, -5, -1), # for the join
  )
```

```
head(data_pop_CBSA)
```

```
## # A tibble: 6 x 13
##   GEO.id  GEO.id2 `GEO.display-la~ HD01_VD01 HD02_VD01 HD01_VD02 HD02_VD02
##   <chr>    <int> <chr>                <int>     <int>     <int>     <int>
## 1 310M30~  10100 Aberdeen, SD Mi~     40749       243     30087       799
## 2 310M30~  10140 Aberdeen, WA Mi~     68170       284     44798      1541
## 3 310M30~  10180 Abilene, TX Met~    155109       686     96299      1908
## 4 310M30~  10220 Ada, OK Micro A~     36811       312     24560       671
## 5 310M30~  10300 Adrian, MI Micr~     93397       412     72719      1342
## 6 310M30~  10420 Akron, OH Metro~    687154       994    475135      4469
## # ... with 6 more variables: HD01_VD09 <int>, HD02_VD09 <int>,
## #   pct_owner_pop <dbl>, pct_renter_pop <dbl>, total_pop <dbl>,
## #   GEOID <chr>
```

```
data_ACS_CBSA <- left_join(data_pop_CBSA, data_unit_CBSA, by = c("GEOID", "GEO.id2", "GEO.id", "GEO.disp
```

```
head(data_ACS_CBSA)
```

```
## # A tibble: 6 x 22
##   GEO.id  GEO.id2 `GEO.display-labe~ HD01_VD01.x HD02_VD01.x HD01_VD02.x
##   <chr>     <int> <chr>                    <int>       <int>       <int>
## 1 310M300U~ 10100 Aberdeen, SD Micr~       40749         243       30087
## 2 310M300U~ 10140 Aberdeen, WA Micr~       68170         284       44798
## 3 310M300U~ 10180 Abilene, TX Metro~      155109         686       96299
## 4 310M300U~ 10220 Ada, OK Micro Area        36811         312       24560
## 5 310M300U~ 10300 Adrian, MI Micro ~       93397         412       72719
## 6 310M300U~ 10420 Akron, OH Metro A~      687154         994      475135
## # ... with 16 more variables: HD02_VD02.x <int>, HD01_VD09 <int>,
## #   HD02_VD09 <int>, pct_owner_pop <dbl>, pct_renter_pop <dbl>,
## #   total_pop <dbl>, GEOID <chr>, HD01_VD01.y <int>, HD02_VD01.y <int>,
## #   HD01_VD02.y <int>, HD02_VD02.y <int>, HD01_VD03 <int>,
## #   HD02_VD03 <int>, pct_owner_unit <dbl>, pct_renter_unit <dbl>,
## #   total_unit <dbl>
```

Picture of Subsidized Households

```
data_CBSA <- data_CBSA %>%
  filter(program_label == "Public Housing" & !(is.na(total_units))) %>%
  select(states, entities, total_units, pct_occupied, people_per_unit, people_total, state) %>%
  mutate(
    GEOID = str_sub(entities, -5, -1)
  ) %>%
  filter(!(GEOID == "78999") & !(GEOID == "66999")) # removes Guam and VI
```

```
head(data_CBSA)
```

```
## # A tibble: 6 x 8
##   states entities    total_units pct_occupied people_per_unit people_total
##   <chr>  <chr>             <int>        <int>           <dbl>        <dbl>
## 1 <NA>   10020 Abbe~        380           92            2.21         772.
## 2 <NA>   10100 Aber~        100           98            1.72         169.
## 3 <NA>   10140 Aber~        415           95            1.57         621
## 4 <NA>   10180 Abil~        493           96            2.29        1082.
## 5 <NA>   10220 Ada,~        275           98            1.29         347.
## 6 <NA>   10260 Adju~        205          100            2.74         563.
## # ... with 2 more variables: state <chr>, GEOID <chr>
```

There are 929 CBSA in the US (including Puerto Rico). Our data set is made of 798 CBSA.

Final data set for county level

```
data_combined_CBSA <- full_join(data_ACS_CBSA, data_CBSA, by = ("GEOID"))
head(data_combined_CBSA)
```

```
## # A tibble: 6 x 29
##   GEO.id    GEO.id2 `GEO.display-labe~ HD01_VD01.x HD02_VD01.x HD01_VD02.x
##   <chr>       <int> <chr>                    <int>       <int>       <int>
## 1 310M300U~   10100 Aberdeen, SD Micr~       40749         243       30087
## 2 310M300U~   10140 Aberdeen, WA Micr~       68170         284       44798
## 3 310M300U~   10180 Abilene, TX Metro~      155109         686       96299
## 4 310M300U~   10220 Ada, OK Micro Area       36811         312       24560
## 5 310M300U~   10300 Adrian, MI Micro ~       93397         412       72719
## 6 310M300U~   10420 Akron, OH Metro A~      687154         994      475135
## # ... with 23 more variables: HD02_VD02.x <int>, HD01_VD09 <int>,
## #   HD02_VD09 <int>, pct_owner_pop <dbl>, pct_renter_pop <dbl>,
## #   total_pop <dbl>, GEOID <chr>, HD01_VD01.y <int>, HD02_VD01.y <int>,
## #   HD01_VD02.y <int>, HD02_VD02.y <int>, HD01_VD03 <int>,
## #   HD02_VD03 <int>, pct_owner_unit <dbl>, pct_renter_unit <dbl>,
## #   total_unit <dbl>, states <chr>, entities <chr>, total_units <int>,
## #   pct_occupied <int>, people_per_unit <dbl>, people_total <dbl>,
## #   state <chr>
```