Final Project

Khyati Naik / Tyler Brown

2022-12-03

For this project, we are following OSEMN data science workflow. OSEMN (Rhymes with possum) was first described in 2010. It has five phases for a data science project: Obtain, Scrub, Explore, Model, and interpret.

We read geographic designation csv file from FHFA website and we read another csv file from Zillow to get the house prices. Furthermore, we read census data from the webpage as an API service.

Research questions

1. What is the relationship between house prices and income for minority vs non-minority counties?

2. What is the geographical distribution of affordability (price to income ratio) for minority counties vs non minority counties?

3. What is the geographical distribution of top 100 most and least affordable counties?

Load packages

```
library(tidyverse)
library(usmap)
library(jsonlite)
```

Read the csv files

```
#provide the github data path for fhfa geographic designations
dt_geog_path <- "https://raw.githubusercontent.com/Naik-Khyati/data607_final_proj/main/data/lya2022.csv

raw_dt_geog <- read.csv(dt_geog_path, sep=",", stringsAsFactors=FALSE)
#glimpse(raw_dt_geog)
```

```
#provide the github data path for zillow home prices data
dt_zillow_path <- "https://raw.githubusercontent.com/Naik-Khyati/data607_final_proj/main/data/County_zh

raw_dt_zill <- read.csv(dt_zillow_path, sep=",", stringsAsFactors=FALSE)
#glimpse(raw_dt_zill)
```

Read the census data from webpage using the api service

```
api <- "https://api.census.gov/data/2020/acs/acs5/profile?get=group(DP03)&for=county:*&in=state:*&key="

census = fromJSON(api) %>%  data.frame()

census = census %>%
  purrr::set_names(as.character(slice(., 1))) %>%
  slice(-1)

raw_dt_cens_inc = census %>%
  select(1097,1098,681,683,684,682)

raw_dt_cens_inc = raw_dt_cens_inc %>%
  rename("ESTIMATE" = "DP03_0086E",
         "MARGIN ERROR" = "DP03_0086M")
```

Scrub FHFA geographic designation data

Add leading zeroes to state and county code to join with geog designation data.

```
raw_dt_zill$STATE <- sprintf("%02d", raw_dt_zill$StateCodeFIPS)
raw_dt_zill$CNTY <-  sprintf("%03d", raw_dt_zill$MunicipalCodeFIPS)
```

Combine state and county columns to than merge with the geog designations data

```
raw_dt_zill$st_cnty <- paste0(raw_dt_zill$STATE,raw_dt_zill$CNTY,sep='')
raw_dt_zill$st_cnty_n <- paste(raw_dt_zill$State,raw_dt_zill$RegionName,sep='-')
```

Remove unwanted columns from the data

```
raw_dt_zill <- raw_dt_zill %>%
  select(-RegionID, -SizeRank,  -RegionType, -RegionName,   -State, -Metro, -STATE, -CNTY, -StateCodeFI
```

Bring the last two columns to the start in the r dataframe

```
raw_dt_zill <- raw_dt_zill %>% relocate(StateName, st_cnty, st_cnty_n)
```

Convert data from wide to long for further data manipulation

```
zhv_long <- raw_dt_zill %>% gather('period','home_val',4:ncol(.))
glimpse(zhv_long)
```

```
## Rows: 778,160
## Columns: 5
## $ StateName <chr> "CA", "IL", "TX", "AZ", "CA", "CA", "FL", "TX", "NY", "CA", ~
## $ st_cnty   <chr> "06037", "17031", "48201", "04013", "06073", "06059", "12086~
## $ st_cnty_n <chr> "CA-Los Angeles County", "IL-Cook County", "TX-Harris County~
## $ period    <chr> "X2000.01.31", "X2000.01.31", "X2000.01.31", "X2000.01.31", ~
## $ home_val  <dbl> 216805, 175598, 115672, 143126, 224478, 271452, 130901, 1143~
```

Create a date variable to plot time series trend

```
zhv_long_dt <- zhv_long %>%
  separate(period, c("yr", "mo" , "day"),"\\.")

zhv_long_dt$yr <-  as.numeric(gsub('X', '', zhv_long_dt$yr))

zhv_long_dt <- zhv_long_dt %>% select(-mo,-day)

head(zhv_long_dt)
```

```
##   StateName st_cnty             st_cnty_n   yr home_val
## 1        CA   06037 CA-Los Angeles County 2000   216805
## 2        IL   17031        IL-Cook County 2000   175598
## 3        TX   48201      TX-Harris County 2000   115672
## 4        AZ   04013    AZ-Maricopa County 2000   143126
## 5        CA   06073   CA-San Diego County 2000   224478
## 6        CA   06059      CA-Orange County 2000   271452
```

Group data by county and year

```
zhv_long_dt_join <- zhv_long_dt %>%
  group_by (StateName, st_cnty, st_cnty_n, yr) %>%
  summarise(mean_hv = mean(home_val))
```

```
## `summarise()` has grouped output by 'StateName', 'st_cnty', 'st_cnty_n'. You
## can override using the `.groups` argument.
```

Scrub geographic data

Add leading zeroes to state and county code to join with geog designation data.

```
raw_dt_geog$STATE <- sprintf("%02d", raw_dt_geog$STATE)
raw_dt_geog$CNTY <-  sprintf("%03d", raw_dt_geog$CNTY)
```

Combine state and county columns to than merge with the geog designations data

```
raw_dt_geog$st_cnty <- paste0(raw_dt_geog$STATE,raw_dt_geog$CNTY,sep='')
```

Bring the last two columns to the start in the r dataframe

```
raw_dt_geog <- raw_dt_geog %>% relocate(st_cnty)
```

Replace 9 with zero in the r dataframe for LYA column

```
raw_dt_geog$LYA[raw_dt_geog$LYA == 9] <- 0
```

Flag tracts with minority percent population greater than 50%

```
raw_dt_geog$flag_min <- ifelse(raw_dt_geog$PCTMIN<=50,0,1)
```

Group data by county

```
raw_dt_geog_manip <- raw_dt_geog %>%  group_by (st_cnty) %>%
  summarise(count_lya_tracts = sum(LYA),
            count_min_tracts = sum(flag_min),
            count_total_tracts = n())
```

Add minority tract share column

```
raw_dt_geog_manip <- raw_dt_geog_manip %>% mutate (
  lya_tract_share = count_lya_tracts/count_total_tracts,
  min_tract_share = count_min_tracts/count_total_tracts
  )
```

Explore the data

```
summary(raw_dt_geog_manip)
```

```
##    st_cnty          count_lya_tracts  count_min_tracts   count_total_tracts
##  Length:3221        Min.   :  0.000   Min.   :   0.000   Min.   :   1.00
##  Class :character   1st Qu.:  0.000   1st Qu.:   0.000   1st Qu.:   4.00
##  Mode  :character   Median :  2.000   Median :   0.000   Median :   8.00
##                     Mean   :  7.887   Mean   :   9.281   Mean   :  26.51
##                     3rd Qu.:  5.000   3rd Qu.:   3.000   3rd Qu.:  19.00
##                     Max.   :989.000   Max.   :1949.000   Max.   :2498.00
##  lya_tract_share  min_tract_share
```

```
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.2000   Median :0.0000
##  Mean   :0.2370   Mean   :0.1725
##  3rd Qu.:0.3333   3rd Qu.:0.2500
##  Max.   :1.0000   Max.   :1.0000
```

If more than 25% of tracts are minority than flag the county as minority.

```
raw_dt_geog_manip$flag_min_cnty <- ifelse(raw_dt_geog_manip$min_tract_share <=0.25,0,1)
raw_dt_geog_manip$flag_lya_cnty <- ifelse(raw_dt_geog_manip$lya_tract_share <=0.25,0,1)
```

Scrub Census data

Split string to create the geographich ID merge key

```
raw_dt_cens_inc <- raw_dt_cens_inc %>%
  mutate(st_cnty = str_sub(raw_dt_cens_inc$GEO_ID, start= -5),
         MFI = as.numeric(raw_dt_cens_inc$ESTIMATE)) %>%
  select ('st_cnty', MFI)
```

Combine the 3 datasets

```
comb_dt_for_anly <- raw_dt_geog_manip %>%
  left_join(raw_dt_cens_inc, c("st_cnty" = "st_cnty")) %>%
  left_join(zhv_long_dt_join, c("st_cnty" = "st_cnty")) %>%
  select(StateName, st_cnty, st_cnty_n, flag_lya_cnty, flag_min_cnty, MFI, yr, mean_hv)
```

Data Analysis

Correlation analysis

```
comb_dt_for_anly %>% filter(yr==2020) %>% na.omit() %>%
  group_by(flag_min_cnty) %>%
  summarize(cor=cor(MFI, mean_hv))
```

```
## # A tibble: 2 x 2
##   flag_min_cnty   cor
##           <dbl> <dbl>
## 1             0 0.645
## 2             1 0.769
```

Above table shows that there is a higher correlation between income and house prices in minority counties as compared to non minority counties. This is an interesting observation as lower correlation between house prices and affordability could mean higher presence of investors (non owner occupied homes) in non minority county.

Add price to income ratio variable for our analysis

```
comb_dt_for_anly_20 <- comb_dt_for_anly %>% mutate(p_to_i = mean_hv/MFI) %>%
  filter(yr==2020) %>% na.omit()
```

Add affordability rank

```
comb_dt_for_anly_20 <- comb_dt_for_anly_20 %>% arrange(p_to_i) %>%
          mutate(aff_rank = 1:nrow(comb_dt_for_anly_20))
```

Add flag for top 100 counties with highest and lowest p/i ratio

```
comb_dt_for_anly_20 <- comb_dt_for_anly_20 %>% mutate(aff_rank_cat =case_when(
  aff_rank<=100 ~ "Highest affordability",
  aff_rank>(nrow(comb_dt_for_anly_20)-100) ~ "Lowest affordability",
   TRUE ~ "Others"
))
```

Explore the data

```
summary(comb_dt_for_anly_20)
```

```
##   StateName           st_cnty           st_cnty_n         flag_lya_cnty
##  Length:2494        Length:2494        Length:2494        Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000
##                                                           Mean   :0.4018
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
##  flag_min_cnty         MFI               yr          mean_hv
##  Min.   :0.0000   Min.   : 31410   Min.   :2020   Min.   :  32849
##  1st Qu.:0.0000   1st Qu.: 57772   1st Qu.:2020   1st Qu.: 114142
##  Median :0.0000   Median : 66613   Median :2020   Median : 156592
##  Mean   :0.2358   Mean   : 69472   Mean   :2020   Mean   : 189543
##  3rd Qu.:0.0000   3rd Qu.: 77502   3rd Qu.:2020   3rd Qu.: 230363
##  Max.   :1.0000   Max.   :182567   Max.   :2020   Max.   :1751724
##      p_to_i           aff_rank      aff_rank_cat
##  Min.   : 0.5944   Min.   :   1.0   Length:2494
##  1st Qu.: 1.9126   1st Qu.: 624.2   Class :character
##  Median : 2.3677   Median :1247.5   Mode  :character
##  Mean   : 2.6213   Mean   :1247.5
##  3rd Qu.: 2.9758   3rd Qu.:1870.8
##  Max.   :15.0760   Max.   :2494.0
```

Above is the description of the final dataset that will be used for analysis. We have used 2020 data as ACS 5 year estimates for income (sourced from census website using API) is for 2020. Variable mean_hv provides data for home prices from zillow.

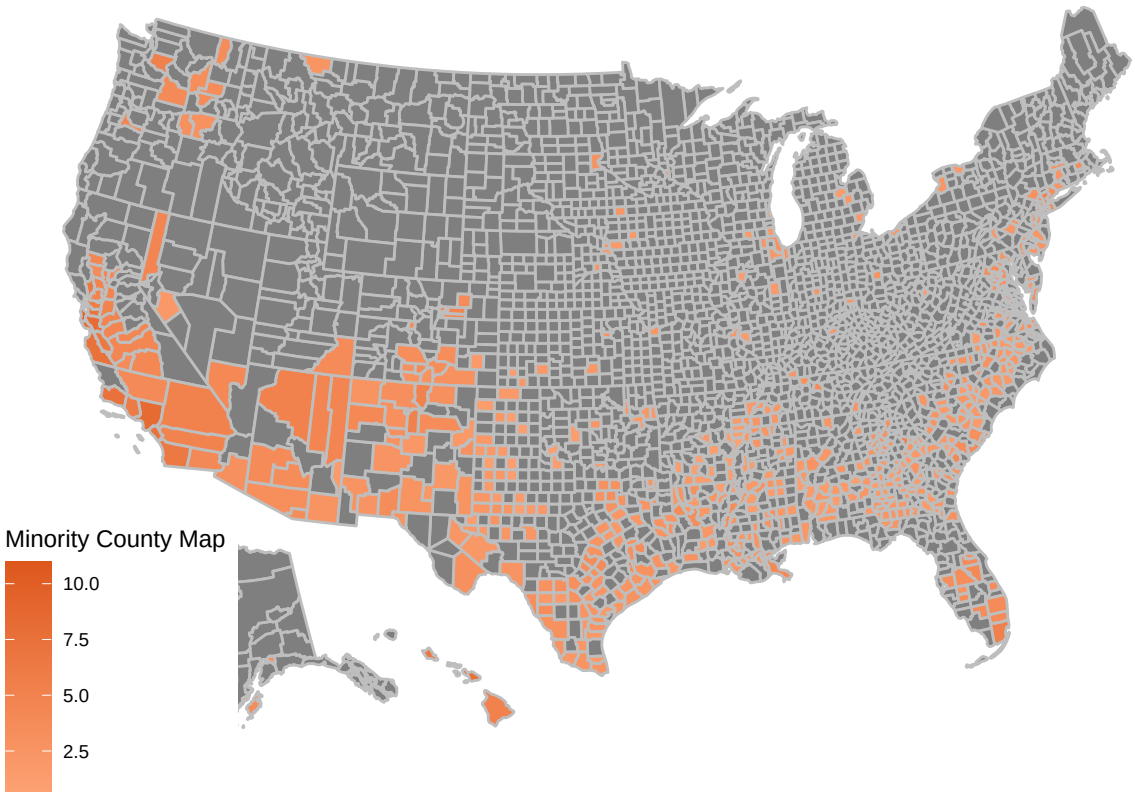
We divide the house prices and income data to create a house price to income ratio (p_to_i) metric, which will give us a sense of house price affordability which is from 0.59 to 15.07, with median value of 2.36.

Minority County Map

```
map_dt <- countypop %>%   left_join(comb_dt_for_anly_20, c("fips" = "st_cnty"))

dt <- map_dt %>%  filter(flag_min_cnty==1)
plot_usmap( data = dt, values = "p_to_i", color="grey") +
scale_fill_continuous( low = "#FDA172", high = "#DD571C", name = "Minority County Map")
```

```
## Warning: Ignoring unknown parameters: linewidth
```




```
summary(comb_dt_for_anly_20 %>% filter(flag_min_cnty==1))
```

```
##   StateName           st_cnty            st_cnty_n         flag_lya_cnty
##  Length:588         Length:588         Length:588         Min.   :0.000
##  Class :character   Class :character   Class :character   1st Qu.:0.000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.000
##                                                           Mean   :0.682
##                                                           3rd Qu.:1.000
##                                                           Max.   :1.000
##  flag_min_cnty      MFI                yr          mean_hv
```

```
##  Min.   :1     Min.   : 31410   Min.   :2020   Min.   :  32849
##  1st Qu.:1     1st Qu.: 52320   1st Qu.:2020   1st Qu.: 106734
##  Median :1     Median : 61979   Median :2020   Median : 152831
##  Mean   :1     Mean   : 66788   Mean   :2020   Mean   : 200221
##  3rd Qu.:1     3rd Qu.: 74385   3rd Qu.:2020   3rd Qu.: 235123
##  Max.   :1     Max.   :165016   Max.   :2020   Max.   :1407372
##      p_to_i           aff_rank      aff_rank_cat
##  Min.   : 0.6997   Min.   :   2.0   Length:588
##  1st Qu.: 2.0028   1st Qu.: 728.2   Class :character
##  Median : 2.4704   Median :1388.0   Mode  :character
##  Mean   : 2.7824   Mean   :1331.2
##  3rd Qu.: 3.1413   3rd Qu.:1964.2
##  Max.   :10.9933   Max.   :2491.0
```

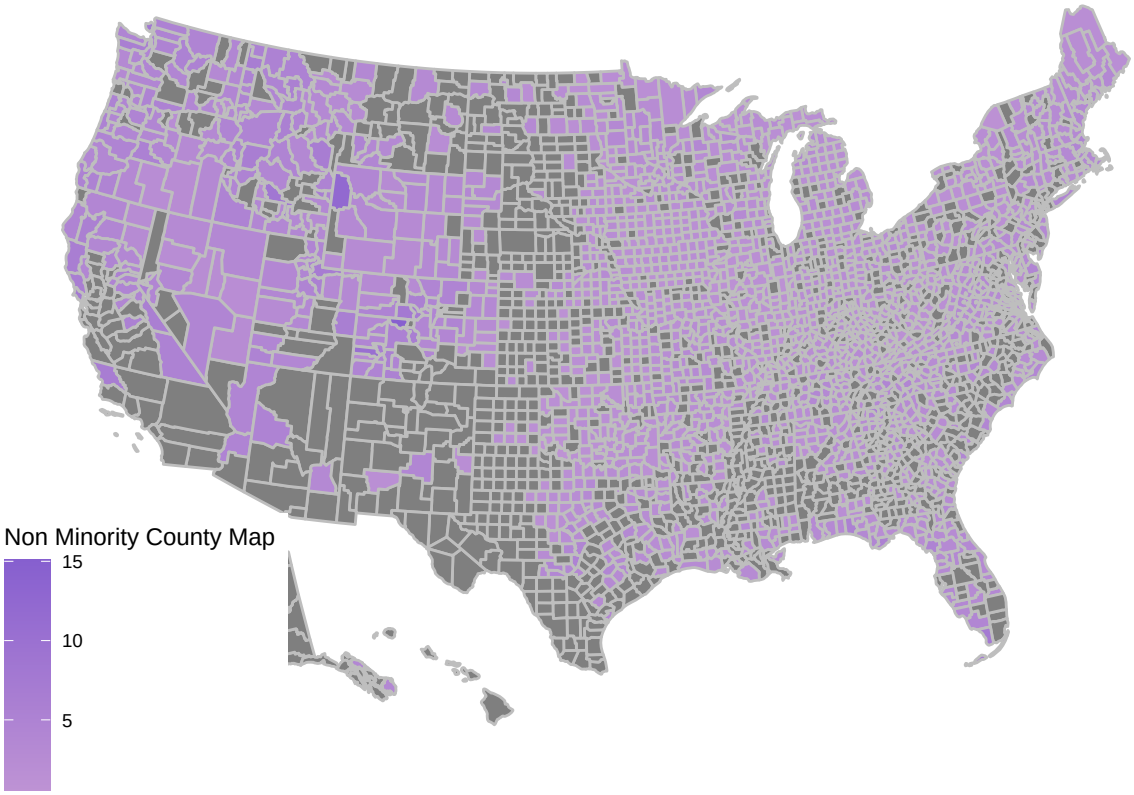
The FHFA minority flag was at tract level, so we converted it to be at county level. We define county level minority flag as counties where more than 25% of tracts are minority tracts. Minority tracts represents tracts where minority population is more than 50%.

Above map is for minority counties for price to income ratio. The minimum value for the ratio is 0.69 with median value of close to 2.47. Interestingly, most o0f the minority counties are in the lower half of US states right from California, Texas to Florida.

Non Minority County Map

```
dt <- map_dt %>%  filter(flag_min_cnty==0)
plot_usmap( data = dt, values = "p_to_i", color="grey") +
scale_fill_continuous( low = "#BE93D4", high = "#865FCF", name = "Non Minority County Map")
```

```
## Warning: Ignoring unknown parameters: linewidth
```




```
summary(comb_dt_for_anly_20 %>% filter(flag_min_cnty==0))
```

```
##   StateName           st_cnty           st_cnty_n         flag_lya_cnty
##  Length:1906        Length:1906        Length:1906        Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000
##                                                           Mean   :0.3153
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
##  flag_min_cnty      MFI                yr          mean_hv
##  Min.   :0     Min.   : 35855   Min.   :2020   Min.   :  35817
##  1st Qu.:0     1st Qu.: 59721   1st Qu.:2020   1st Qu.: 115881
##  Median :0     Median : 67785   Median :2020   Median : 157924
##  Mean   :0     Mean   : 70300   Mean   :2020   Mean   : 186249
##  3rd Qu.:0     3rd Qu.: 78011   3rd Qu.:2020   3rd Qu.: 227046
##  Max.   :0     Max.   :182567   Max.   :2020   Max.   :1751724
##      p_to_i           aff_rank      aff_rank_cat
##  Min.   : 0.5944   Min.   :   1.0   Length:1906
##  1st Qu.: 1.8927   1st Qu.: 605.2   Class :character
##  Median : 2.3459   Median :1211.0   Mode  :character
##  Mean   : 2.5716   Mean   :1221.7
##  3rd Qu.: 2.9298   3rd Qu.:1842.5
##  Max.   :15.0760   Max.   :2494.0
```

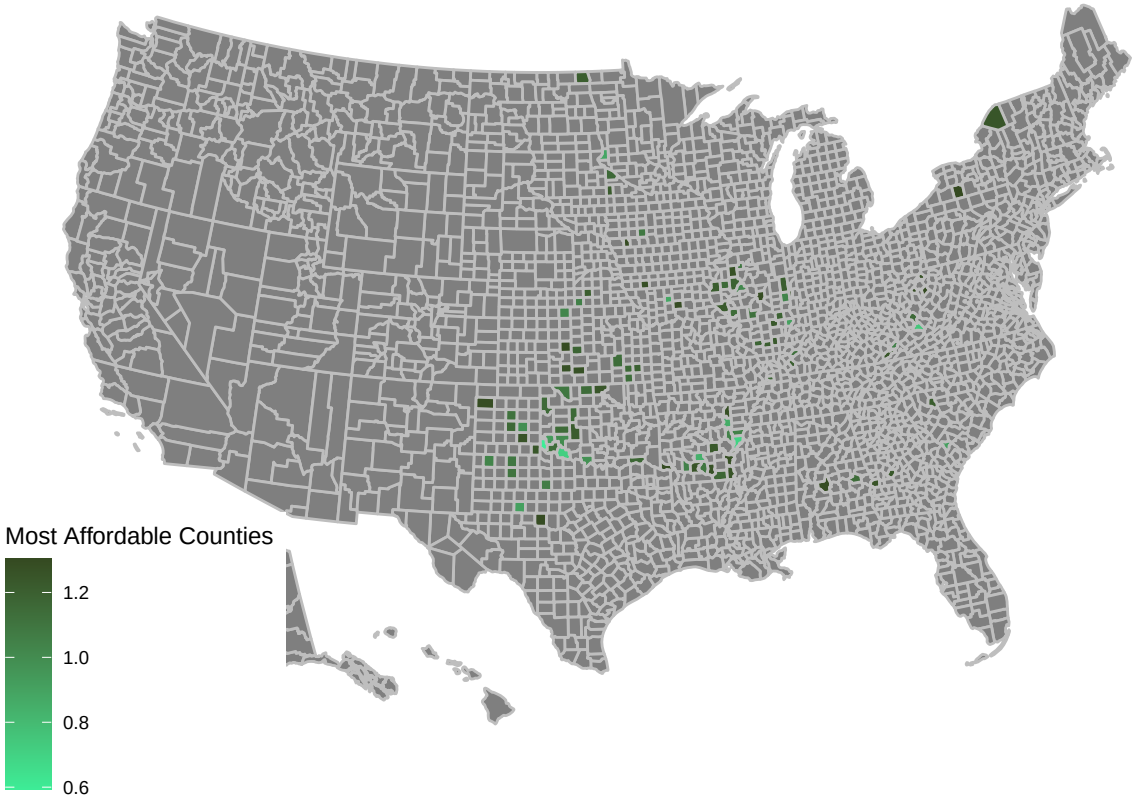
In Non minority counties map, minimum price to income ratio is 0.59. Interestingly, the 1st quartile, median and 3rd quartile ratio is lower than minority counties map. It could be because there are so many counties

from Midwest (which are non minority), where home prices are the lowest of the 4 US regions. However, the max value is higher (15.07) in non minority counties map, due to New York and Massachusetts.

Top 100 Most Affordable Counties map

```
dt <- map_dt %>%  filter(aff_rank_cat=="Highest affordability")
plot_usmap( data = dt, values = "p_to_i", color="grey") +
scale_fill_continuous( low = "#3DeD97", high = "#354A21",  name = "Most Affordable Counties")
```

```
## Warning: Ignoring unknown parameters: linewidth
```

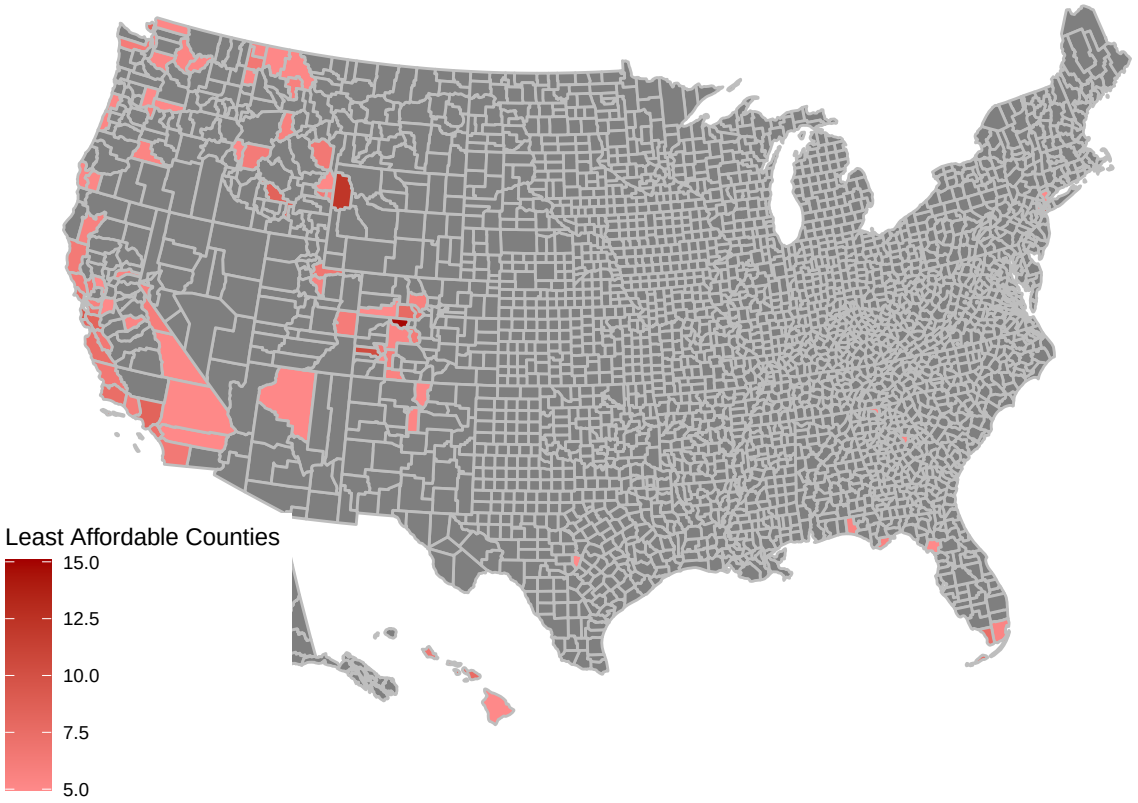


Top 100 most affordable counties are mostly in Midwest and South. There are some counties in Northeast which make it in the top 100 affordable list. There are no counties from West in the top 100 affordable list.

Top 100 Least Affordable Counties map

```
dt <- map_dt %>%  filter(aff_rank_cat=="Lowest affordability")
plot_usmap( data = dt, values = "p_to_i", color="grey") +
scale_fill_continuous( low = "#FF8A8A", high = "#A30000",  name = "Least Affordable Counties")
```

```
## Warning: Ignoring unknown parameters: linewidth
```



The map shows that most of the top 100 least affordable counties are in West namely in California, Washington etc. There are some pockets in Southern US such as Florida and Texas too that have low affordability.

Conclusion

- We observe that there is a high correlation between home prices and income in minority counties compared to non minority counties.
- From the minority maps, we observe that lower half of US has most of the minority counties and in terms of affordability it is mostly similar to non minority counties. However, non minority counties have higher max values due to presence of high cost areas from New York and Massachusetts.
- Furthermore, we also observe that Western part of US is the most expensive where as Midwest has the most counties in top 100 most affordable counties.

Challenges

It was challenging to self learn map plotting. We realized that there are various packages that we can use to plot map. example, urban institute has their package called as urbnmapr. There are some other packages also such as usmap that we used in this case for plotting. Additionally, implementing the API caused a "hiccup" in the process. The available APIs through the Census Bureau were difficult to interpret and load. We also liked leaflet package a lot and we hope to use it in the future for maps, as it has some advanced features as well such as adding multiple layers to the map (example state borders in county level map) or use another variable as a metric and represent it as a bubble etc. It is also highly interactive and we can customize labels and popups which could be extremely useful especially in R shiny apps.

Learnings

We used various data reading techniques such as reading data as csv from github and reading data through API in this project. We performed multiple data manipulation techniques right from string split, omitting missing rows, subsetting data, joining multiple dataframes, formatting data from wide to long, converting data from character to numeric and viceversa, adding leading zeroes to form a join key with other datasets.

It was also interesting that we had to understand the data and think about creating different metrics as suited for the analysis. For example we had to think about how to change the minority flag granularity level from tract to county, as our analysis was at county level. Similarly, we came up with the idea of price to income ratio which made it easier to understand affordability and we only had to look at one variable instead of two different variables.

Lastly, we also though about different ways to make the data more digestable for users and hence created maps instead of tables or other bar/line charts to analyze the data.