Patrick Page

Project 2 Design Document

H517 Data Visualization

Spring 2023

Introduction

Many cities in the US have been experiencing a housing crisis. Housing prices have been drastically increasing which has made it hard for people to buy a home in the areas where they have grown up. Cities like Oakland and Seattle have experienced intense levels of gentrification which in turn has forced people out of neighborhoods which they can no longer afford and increased populations of homeless.

The intent of the project was to create a data visualization tool in Tableau that could highlight these key cities in the US that have experienced these rapid price increases in home costs, especially in areas where the median income has not been able to experience the same sustained growth.

Data

Data was collected from three different sources. Housing and rental property listings between 2010 and 2017 were collected from Zillow ZTRAX database. This data set consisted of median list prices for every city Zillow operates in (which was way more data than needed for this project). To pair this dataset down to a more manageable number, a dataset from the 2020 US Census was used to determine the top 500 most populated cities in the US. This dataset allowed the Zillow ZTRAX dataset to be filtered to only include those 500 cities. Income data was found by merging the 2010 through 2017 reports of the American Community Survey. This dataset was also filtered to only pertain to the top 500 most populated US cities.

Because the only house and rental listings dataset available is now almost 5 years old, an assumption was made for this project that cities have sustained the same growth in home prices and income from then to the current day.

Housing and Income Prices Dashboard

The first dashboard page is the Housing and Income Prices dashboard. The intention of this page is to provide the most recent story of where the highest housing costs exist. The map at the top contains a color gradient with the darker colors indicating the state has a higher average housing cost. In the example below, the state of California was selected since California appears as the darkest color. It can then be determined what specific cities in California contain the highest median housing prices since the lists below display the list prices are automatically sorted in descending order.

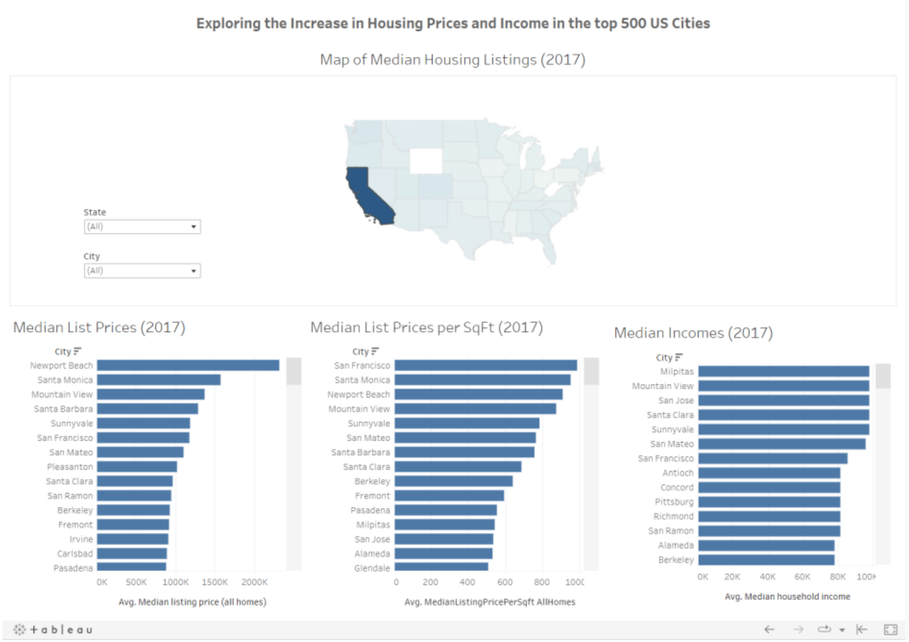
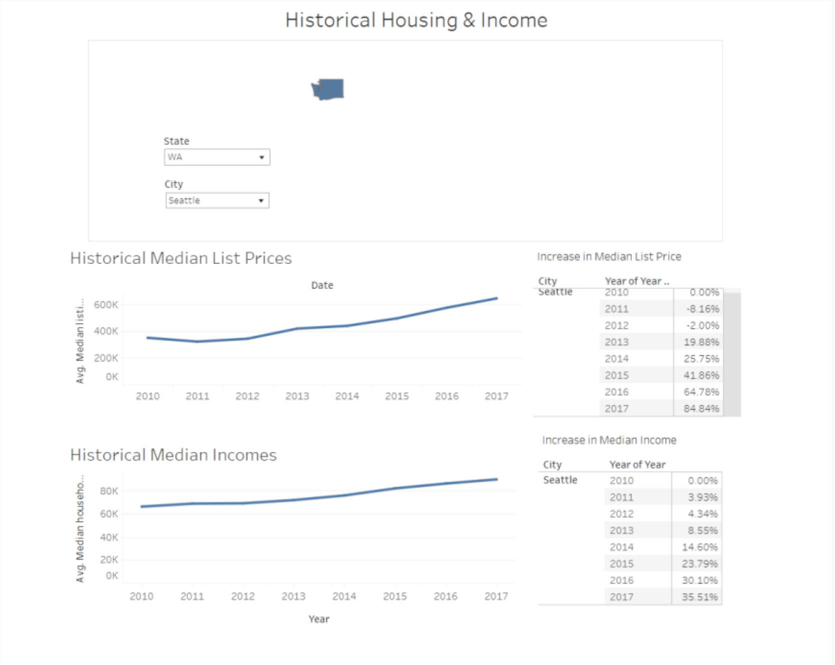


Figure 1: Example #1 (Housing & Income Prices Dashboard)

Based on the data above, it appears that although San Francisco is the priciest California city in terms of square feet, it is not the median cost per house is less than cities like Newport Beach and Santa Monica. This may be because San Francisco is a much more urban environment compared to Newport Beach and Santa Monica so the homes in San Francisco are probably smaller than the homes elsewhere. It can also be determined that just because a city has the highest median list price in a particular state, it doesn't necessarily mean that city will have the highest median income as is the case with Newport Beach. It is important to remember that median income is not the same as mean income and that the median is not influenced by the potential few uber wealthy that could drastically increase the mean income.

Historical Housing & Income Dashboard

The second dashboard is the Historical Housing and Income dashboard. The purpose of this dashboard is to compare trends between median housing prices and median income levels. The state and city can be selected from the drop-down lists. Based on the city selection, the trendlines for housing and income are generated with the year-by-year percent differences appearing on the sides of their respective graphs.



Historical Housing & Income

State: WA

City: Seattle

Increase in Median List Price

City	Year of Year ..	
Seattle	2010	0.00%
	2011	-8.16%
	2012	-2.00%
	2013	19.88%
	2014	25.75%
	2015	41.86%
	2016	64.78%
	2017	84.84%

Increase in Median Income

City	Year of Year	
Seattle	2010	0.00%
	2011	3.93%
	2012	4.34%
	2013	8.55%
	2014	14.60%
	2015	23.79%
	2016	30.10%
	2017	35.51%

Figure 2: Example #2 (Historical Housing & Income Dashboard)

In this example, Seattle was selected in order to better understand how housing prices and income levels have increased from 2010 to 2017. Seattle is a fast-growing city that has a high cost of living, so it is a good example of extreme rates of housing costs. According to the data, in 2017 the median list price has increased nearly 85 percent since 2010, meanwhile, the median income for 2017 was only a 35 percent increase from the number in 2010. This discrepancy in income and housing growth rates seems unsustainable for a city as big as Seattle. Ideally, the growth of income would follow closely the growth of housing costs, but that is not the case here. As the housing costs continue to increase at this rate, more and more people who are not experiencing similar rates of increasing income may soon be displaced from their homes.

Buying vs. Renting Dashboard

Finally, the Buying vs. Renting dashboard was built to analyze the price rates of different types of housing and rental properties. This dashboard helps determine if the increase in cost of living is localized to rental properties or affects both rental properties and home ownership. It can also be determined if small, single-family homes are affected in similar ways to multi-family dwellings.

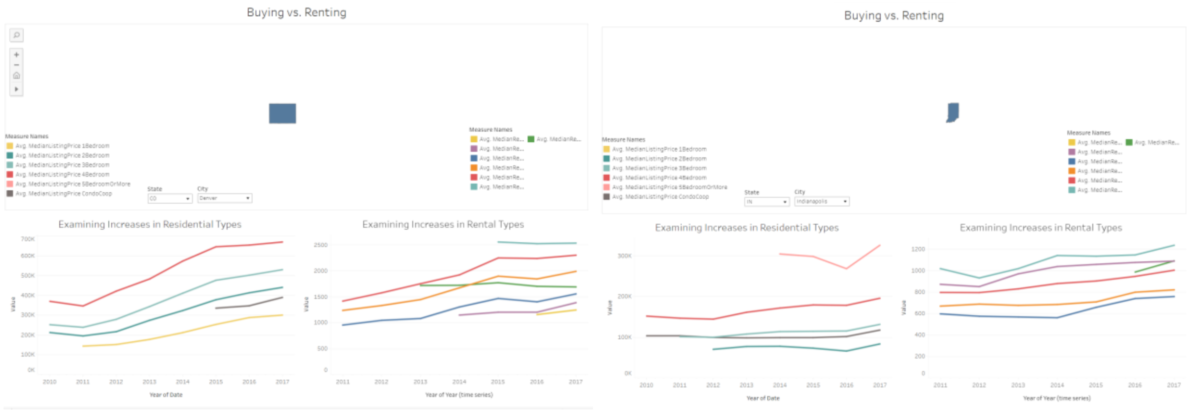


Figure 3: Example #3 (Buying vs. Renting Dashboard)

In the example above, Denver and Indianapolis are compared side-by-side. The data suggests Denver has experienced larger increases in all forms of housing compared to Indianapolis. The increases in Denver seem to have hit all forms of housing equally with each category represented by lines with similar slopes. In Indianapolis, some categories of housing showed steady increases while other categories appeared to stay relatively the same.

Conclusion

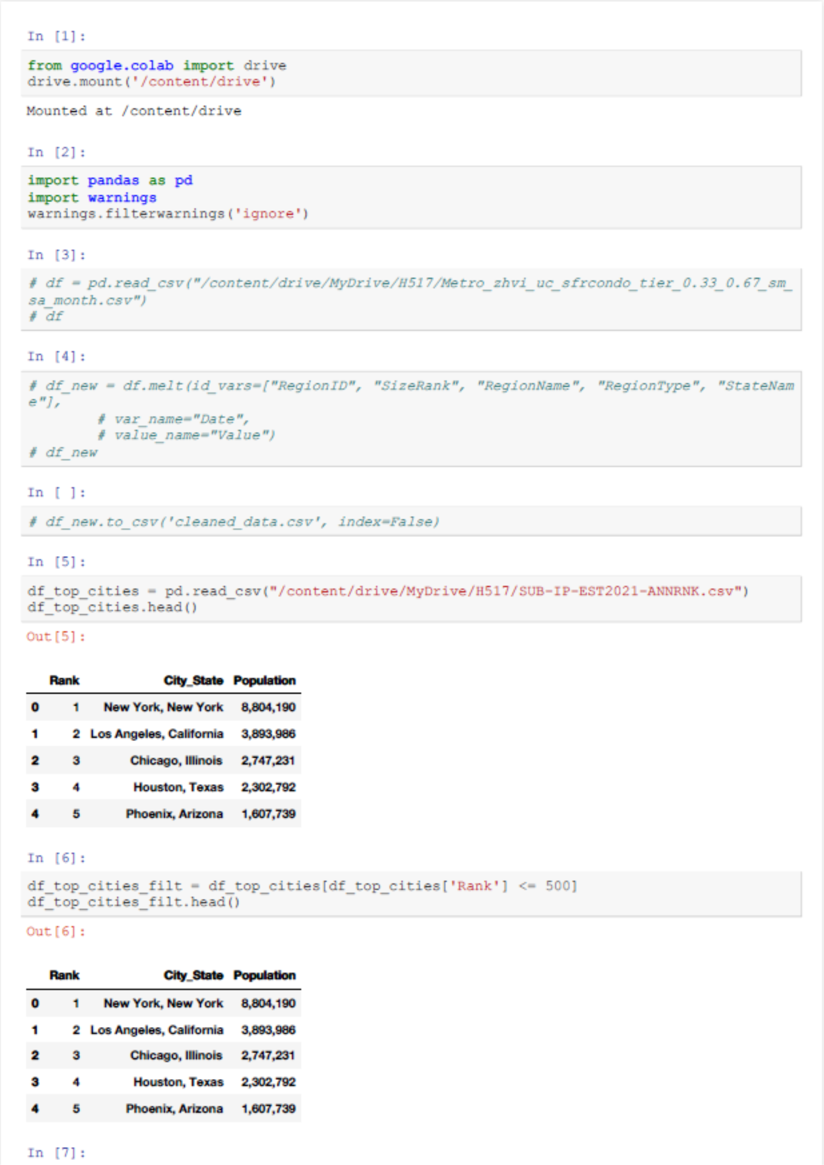
The problem of increasing housing costs and median income levels not experiencing the same levels of increase is an apparent issue for numerous cities in the US. The tool built from this project can serve as a reliable method for indicating cities that are experiencing these sorts of issues. This is a nuanced problem, and not all cities experience increases in housing costs the same way. While this tool can indicate cities that have relatively low median incomes and high median housing prices (potential housing crisis), the tool can also indicate cities with relatively high median incomes and low median housing prices (low cost of living).

Appendix

Data Sources:

- https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html
- https://data.census.gov/table?t=Income+(Households,+Families,+Individuals)&g=010XX00US$0400000&y=2017&tid=ACSST1Y2017.S1901
- https://www.kaggle.com/datasets/zillow/zecon

Python Notebook File



In [1]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [2]:

```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

In [3]:

```
# df = pd.read_csv("/content/drive/MyDrive/H517/Metro_zhvi_uc_sfrcondo_tier_0.33_0.67_sm_
sa_month.csv")
# df
```

In [4]:

```
# df_new = df.melt(id_vars=["RegionID", "SizeRank", "RegionName", "RegionType", "StateNam
e"],
        # var_name="Date",
        # value_name="Value")
# df_new
```

In [ ]:

```
# df_new.to_csv('cleaned_data.csv', index=False)
```

In [5]:

```
df_top_cities = pd.read_csv("/content/drive/MyDrive/H517/SUB-IP-EST2021-ANNRNK.csv")
df_top_cities.head()
```

Out[5]:

	Rank	City_State	Population
0	1	New York, New York	8,804,190
1	2	Los Angeles, California	3,893,986
2	3	Chicago, Illinois	2,747,231
3	4	Houston, Texas	2,302,792
4	5	Phoenix, Arizona	1,607,739

In [6]:

```
df_top_cities_filt = df_top_cities[df_top_cities['Rank'] <= 500]
df_top_cities_filt.head()
```

Out[6]:

	Rank	City_State	Population
0	1	New York, New York	8,804,190
1	2	Los Angeles, California	3,893,986
2	3	Chicago, Illinois	2,747,231
3	4	Houston, Texas	2,302,792
4	5	Phoenix, Arizona	1,607,739

In [7]:

```
df_top_cities_filt_split = pd.DataFrame(df_top_cities_filt['City_State'].str.split(',',1)
.tolist(), columns = ['City','State'])
df_top_cities_filt_split = df_top_cities_filt_split.applymap(lambda x: x.strip() if isin
stance(x, str) else x)
df_top_cities_filt_split.head()
```

Out[7]:

	City	State
0	New York	New York
1	Los Angeles	California
2	Chicago	Illinois
3	Houston	Texas
4	Phoenix	Arizona

In [8]:

```
df_abb = pd.read_csv("/content/drive/MyDrive/H517/us-states-territories.csv", encoding =
"ISO-8859-1")
df_abb = pd.DataFrame(df_abb, columns = ['Name', 'Abbreviation'])
df_abb = df_abb.applymap(lambda x: x.strip() if isinstance(x, str) else x)
df_abb.head()
```

Out[8]:

	Name	Abbreviation
0	Alabama	AL
1	Alaska	AK
2	Arizona	AZ
3	Arkansas	AR
4	California	CA

In [9]:

```
df_top_cities_filt_split_abb = pd.merge(df_top_cities_filt_split, df_abb, left_on='State
', right_on='Name', how='left')
df_top_cities_filt_split_abb['City_State'] = df_top_cities_filt_split_abb['City'] + ', '
+ df_top_cities_filt_split_abb['Abbreviation']
df_top_cities_filt_split_abb.head()
```

Out[9]:

	City	State	Name	Abbreviation	City_State
0	New York	New York	New York	NY	New York, NY
1	Los Angeles	California	California	CA	Los Angeles, CA
2	Chicago	Illinois	Illinois	IL	Chicago, IL
3	Houston	Texas	Texas	TX	Houston, TX
4	Phoenix	Arizona	Arizona	AZ	Phoenix, AZ

In [10]:

```
df_relate = pd.read_csv("/content/drive/MyDrive/H517/cities_crosswalk.csv")
df_relate['City_State'] = df_relate['City'] + ', ' + df_relate['State']
df_relate['County_State'] = df_relate['County'] + ', ' + df_relate['State']
df_relate.head()
```

Out[10]:

	Unique_City_ID	City	County	State	City_State	County_State

	Unique_City_ID	City	County	State	City_State	County_State
0	oak_grovechristianky	Oak Grove	Christian	KY	Oak Grove, KY	Christian, KY
1	jarvisburgcurrituckne	Jarvisburg	Currituck	NC	Jarvisburg, NC	Currituck, NC
2	mcminnvilleyamhillor	McMinnville	Yamhill	OR	McMinnville, OR	Yamhill, OR
3	union_townshiperiepa	Union Township	Erie	PA	Union Township, PA	Erie, PA
4	oshkoshwinnebagowi	Oshkosh	Winnebago	WI	Oshkosh, WI	Winnebago, WI

In [11]:

```
df_cities_time = pd.read_csv("/content/drive/MyDrive/H517/City_time_series.csv")
df_cities_time = df_cities_time[df_cities_time['Date'] >= '2010-1-1']
df_cities_time = pd.merge(df_cities_time, df_relate, left_on='RegionName', right_on='Uni
que_City_ID', how='left')
df_cities_time['City_State'] = df_cities_time["City"].astype(str) + ', ' + df_cities_tim
e["State"].astype(str)
df_cities_time_filt = df_cities_time[df_cities_time['City_State'].isin(df_top_cities_filt
_split_abb['City_State'])]
df_cities_time_filt.head()
```

Out[11]:

	Date	RegionName	InventorySeasonallyAdjusted_AllHomes	InventoryRaw_AllHomes	MedianListingPricePerSqf
13	2010-10-31	abilenetaylortx	869.0	871.0	
84	2010-10-31	akronsummitoh	1485.0	1530.0	
88	2010-10-31	alamedaalamedaca	NaN	NaN	
98	2010-10-31	albanyalbanyny	336.0	368.0	
119	2010-10-31	albuquerquebernalillonm	3494.0	3597.0	

5 rows × 87 columns

In [ ]:

```
df_cities_time_filt.to_csv('/content/drive/MyDrive/H517/City_time_series_filtered.csv', i
ndex=False)
```

In [12]:

```
df_relate_cleaned = df_relate[df_relate['City_State'].isin(df_top_cities_filt_split_abb['
City_State'])]
df_relate_cleaned.head()
```

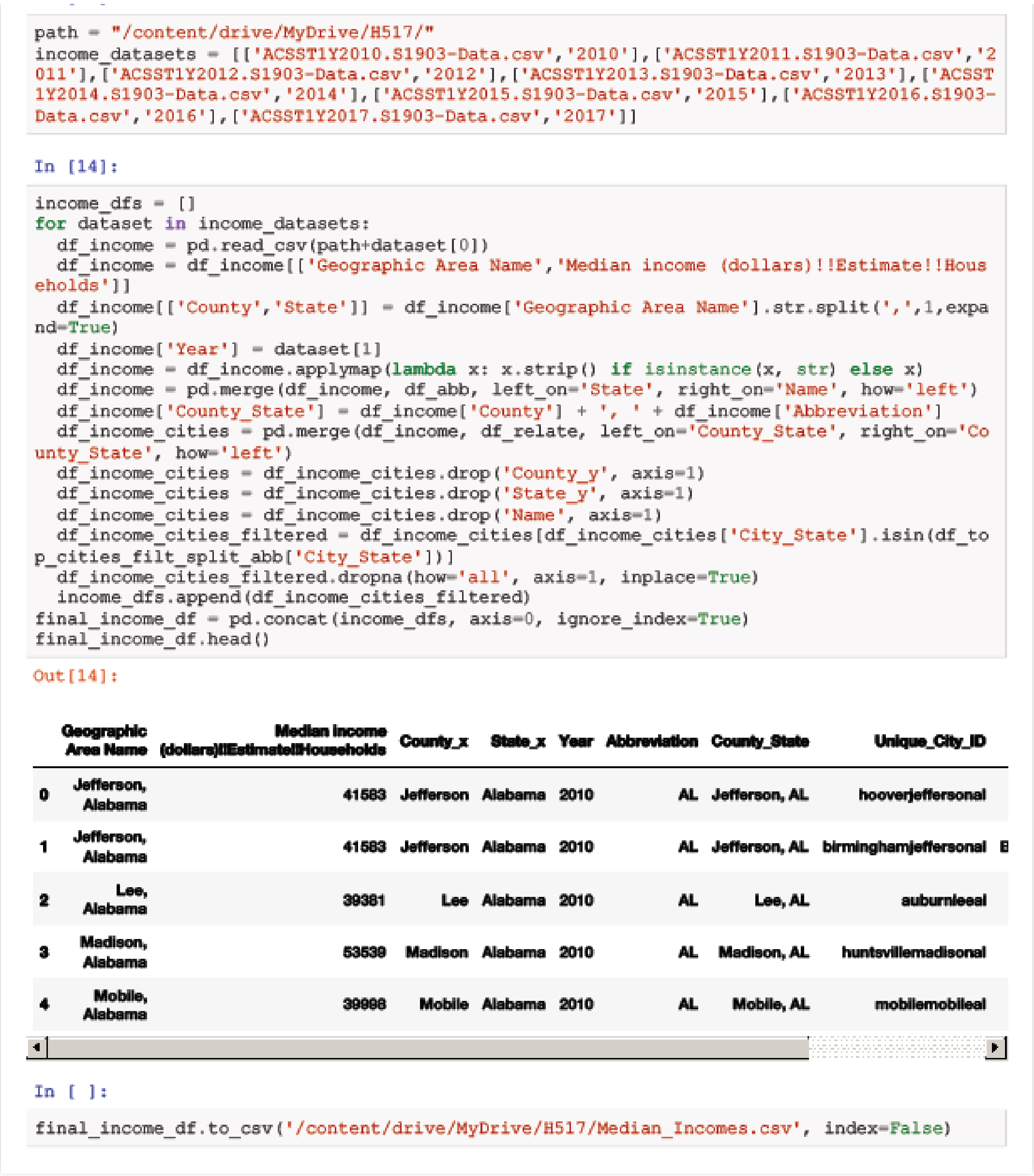
Out[12]:

	Unique_City_ID	City	County	State	City_State	County_State
84	lawtoncomancheok	Lawton	Comanche	OK	Lawton, OK	Comanche, OK
169	jacksonhindsms	Jackson	Hinds	MS	Jackson, MS	Hinds, MS
188	south_gatelos_angelesca	South Gate	Los Angeles	CA	South Gate, CA	Los Angeles, CA
236	melbournebrevardfl	Melbourne	Brevard	FL	Melbourne, FL	Brevard, FL
291	carmelhamiltonin	Carmel	Hamilton	IN	Carmel, IN	Hamilton, IN

In [ ]:

```
df_relate_cleaned.to_csv('/content/drive/MyDrive/H517/cities_crosswalk_filtered.csv', ind
ex=False)
```

In [13]:



```
path = "/content/drive/MyDrive/H517/"
income_datasets = [['ACSST1Y2010.S1903-Data.csv','2010'],['ACSST1Y2011.S1903-Data.csv','2011'],['ACSST1Y2012.S1903-Data.csv','2012'],['ACSST1Y2013.S1903-Data.csv','2013'],['ACSST1Y2014.S1903-Data.csv','2014'],['ACSST1Y2015.S1903-Data.csv','2015'],['ACSST1Y2016.S1903-Data.csv','2016'],['ACSST1Y2017.S1903-Data.csv','2017']]
```

In [14]:

```
income_dfs = []
for dataset in income_datasets:
  df_income = pd.read_csv(path+dataset[0])
  df_income = df_income[['Geographic Area Name','Median income (dollars)!!Estimate!!Households']]
  df_income[['County','State']] = df_income['Geographic Area Name'].str.split(',',1,expand=True)
  df_income['Year'] = dataset[1]
  df_income = df_income.applymap(lambda x: x.strip() if isinstance(x, str) else x)
  df_income = pd.merge(df_income, df_abb, left_on='State', right_on='Name', how='left')
  df_income['County_State'] = df_income['County'] + ', ' + df_income['Abbreviation']
  df_income_cities = pd.merge(df_income, df_relate, left_on='County_State', right_on='County_State', how='left')
  df_income_cities = df_income_cities.drop('County_y', axis=1)
  df_income_cities = df_income_cities.drop('State_y', axis=1)
  df_income_cities = df_income_cities.drop('Name', axis=1)
  df_income_cities_filtered = df_income_cities[df_income_cities['City_State'].isin(df_top_cities_filt_split_abb['City_State'])]
  df_income_cities_filtered.dropna(how='all', axis=1, inplace=True)
  income_dfs.append(df_income_cities_filtered)
final_income_df = pd.concat(income_dfs, axis=0, ignore_index=True)
final_income_df.head()
```

Out[14]:

	Geographic Area Name	Median income (dollars)!!Estimate!!Households	County_x	State_x	Year	Abbreviation	County_State	Unique_City_ID
0	Jefferson, Alabama	41583	Jefferson	Alabama	2010	AL	Jefferson, AL	hooverjeffersonal
1	Jefferson, Alabama	41583	Jefferson	Alabama	2010	AL	Jefferson, AL	birminghamjeffersonal
2	Lee, Alabama	39381	Lee	Alabama	2010	AL	Lee, AL	auburnleeal
3	Madison, Alabama	53539	Madison	Alabama	2010	AL	Madison, AL	huntsvillemadisonal
4	Mobile, Alabama	39996	Mobile	Alabama	2010	AL	Mobile, AL	mobilemobileal

In [ ]:

```
final_income_df.to_csv('/content/drive/MyDrive/H517/Median_Incomes.csv', index=False)
```