Capstone Project Milestone Report

Venkatesh Krishnan

June 20, 2018

Introduction to the problem:

Zillow has a database of approximately 110 million homes in USA. I want to evaluate the factors impacting home prices in order to improve the accuracy of "Zestimates" (current value) and possibly include a projection into future.
(i) Do Household incomes have a direct correlation to home prices?
(ii) What other factors directly and indirectly impact home valuation?
(iii) How can we leverage machine learning to predict home prices for any area?

A deeper dive into the data set:

Important fields and information in the data set – Zip code, City, State, Household Annual Gross Income (AGI) and Zillow Home Value Index (ZHVI).

Limitations of the data set:

1. The available data for both Household Annual Gross Income and Zillow Home Value Index are plain facts and do not talk about the factors influencing the numbers.
2. This data is only from the latest year for Annual Gross Income and the latest home prices but does not have the historical values which is important to study the trend.
3. State-wise data – State Income Tax, Demographics, Crime Data, Education level can all be good factors to consider.
4. Type of usage – Owner occupied or rented is not known.
5. Home type – Condo or Townhome or Single Family is not classified.

Cleaning and Wrangling:

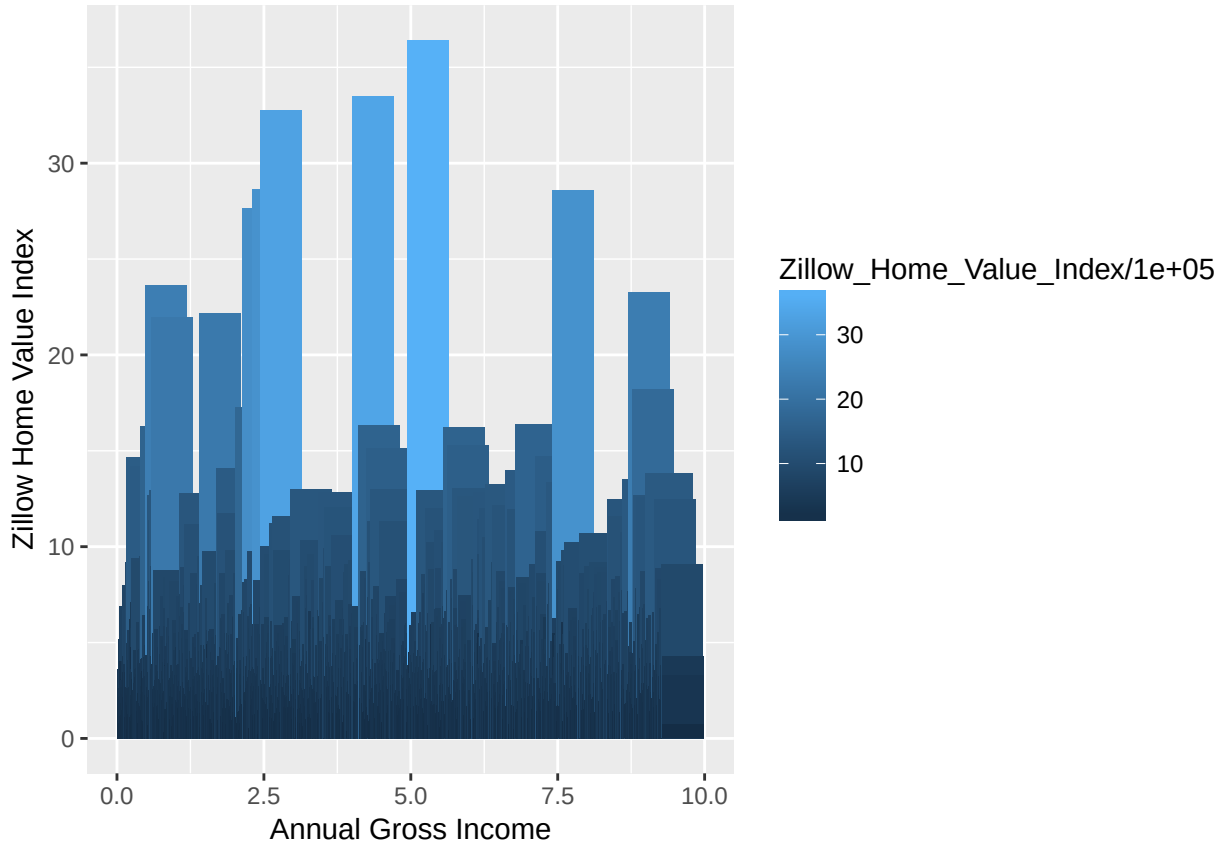
1. Imported data to a data frame using R.
2. In the Zillow Home Value Index data frame, retained RegionID aka "Zip code", City, State and ZHVI columns. Removed other columns that were not applicable.
3. Renamed column "RegionID" to "Zip Code".
4. In the Annual Gross Income data frame, retained Zip code and AGI columns. Removed other columns that were not applicable.
5. Merged Zillow Home Value and Internal Revenue Service Income by Zip code using "left join.
6. Deleted all rows if the that had "NA" or "blank".
7. Retained data with zip codes only in the range "00501 and 99950" as applicable to USA.
8. Created a CSV file post cleanup.

Data Exploration and Findings:

There was a positive correlation of Household Incomes to Home Values.

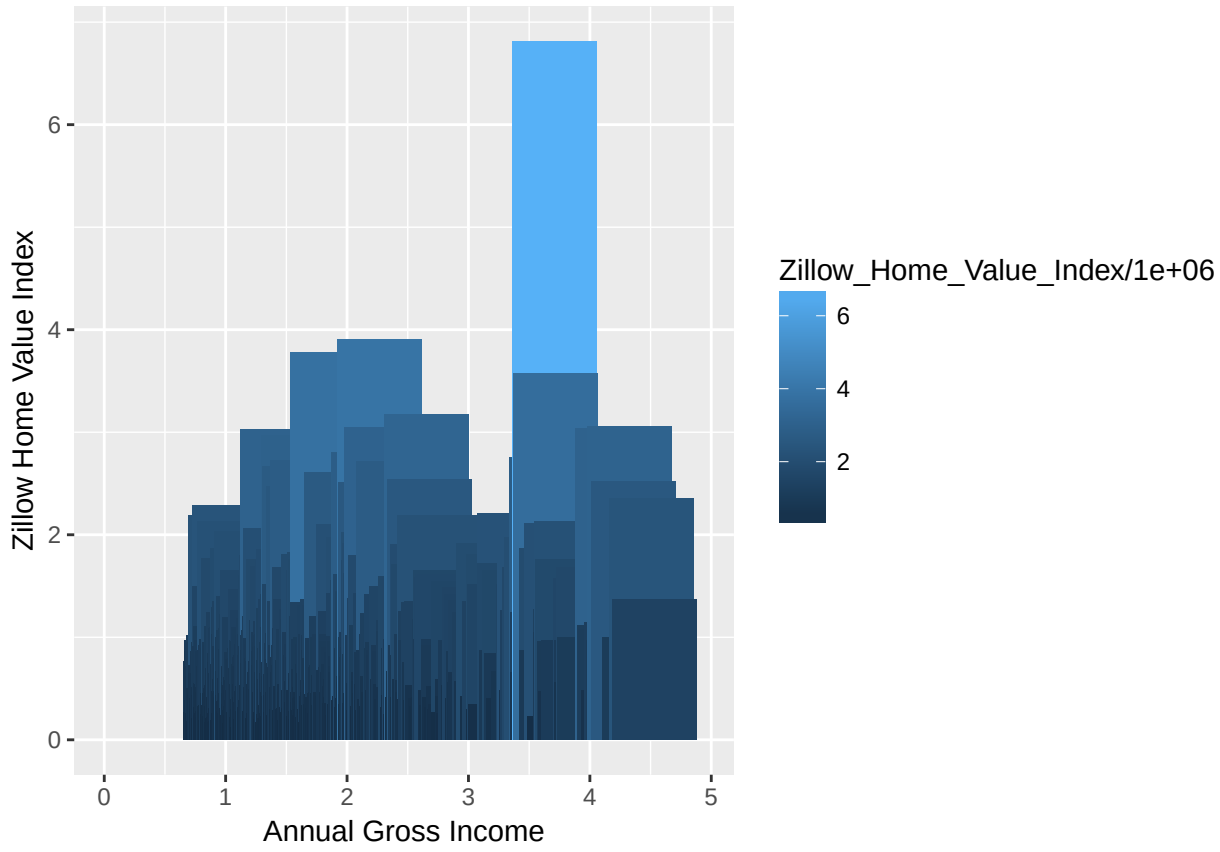
Percentage of households with AGI under $ 1 million.

Percentage of households with AGI under $ 1M is approximately 83.99% and the corresponding home value range is $ 11351 to $ 3641564.



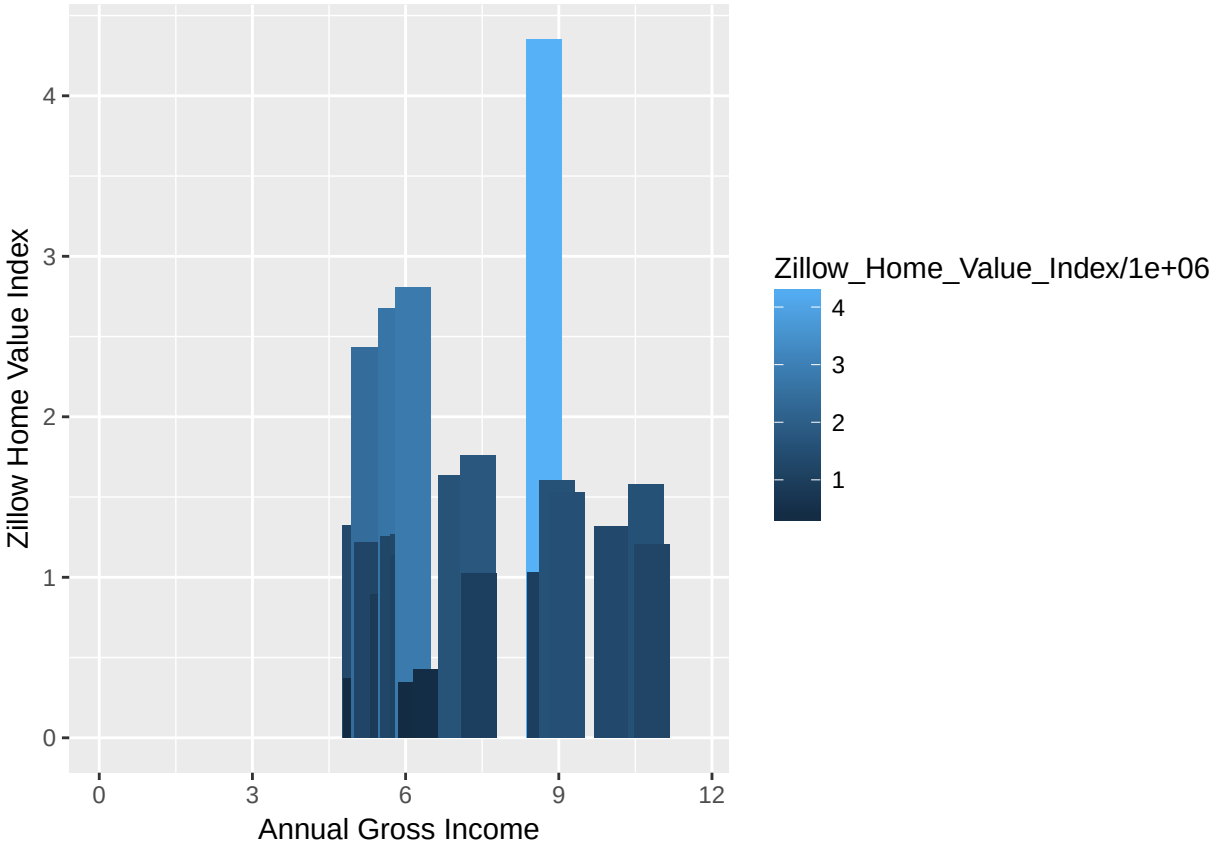
Percentage of households with AGI between $ 1 and $ 5 million.

Percentage of households with AGI between $ 1M and $ 5M is approximately 15.85% and the corresponding home value range is $ 64053 to $ 6810939.

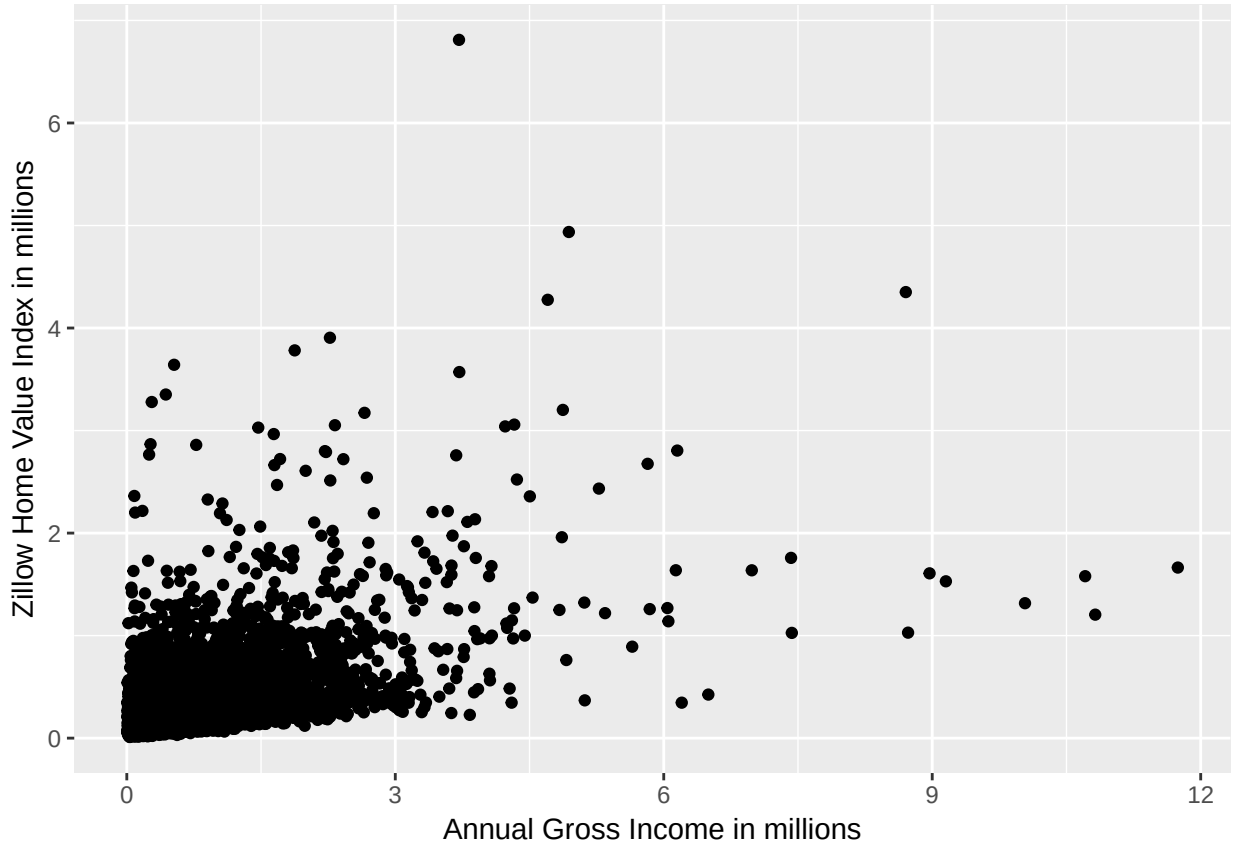


Percentage of households with AGI above $ 5 million.

Percentage of households with AGI over $ 5M is approximately 0.16% and the corresponding home value range is $ 345803 to $ 4350546.


Zillow Home Value Index
4
3
2
1
0
0
3
6
9
12
Annual Gross Income
Zillow_Home_Value_Index/1e+06
4
3
2
1

Scatter Plot:



Recommendations / Leading Questions:

We have only obtained a sample from approx. 14,000 zip codes of USA out of a total of approx. 42,000 zip codes in the country. So, further data is recommended to be obtained to extrapolate this analysis.

1) Can we analyze the data for each state and compare it with the country's data.
2) This is only the latest data from 2016. Can we also obtain historical data for the past 20 years or more to arrive at trend analysis. This will lead us to finding out the driving factors of a particular geography.
3) Currently what we see from Zillow is the actual price paid for a home and it's current value. What is not available now is the projection - 1 year from today, 3 years from today etc., This can provide insight for a customer into a particular market.
4) We can incorporate the crime data of a particular area / zip code that may throw some insight into home valuation.
5) Investments in technology, medicine, infrastructure, real estate (Commercial, Residential and Retail) can act as direct indicators for price prediction in an area.