write up

Nick Peters

2024-11-05

Topic

My broad topic involves observing the differences in shelter/housing costs between the United States and Canada. Housing cost is my outcome variable and I'll be looking at how a multitude of factors effect it across different regions. Some of the key independent variables include income, home ownership, education, household mobility, citizenship, etc.

Motivation

I am interested in discovering some of the causes in disparate housing costs between similar geographic areas that can be easily observed across the United States and Canada. While I won't be able to gather the entire picture during this term project, I think I will be able to draw some interesting conclusions.

Data

For my Canadian data, I am using public use 2021 Canadian census data. One reason I chose this source is that it gives a great representation of the Canadian population while still being comparable to the American data I had available. For America I used a 2021 ACS sample from IPUMS. The ACS (American Community Survey) is a household level survey that focuses on American housing and workforce.

Data Processing

Before I started processing the data I figured out which variables from the Canadian census I should keep since I had no control over limiting which ones were imported at the start. In order to determine which variables I should keep I first looked for anything that could have an effect on housing costs. Secondly, I then compared the remaining variables to the the ACS so that I could have comparable variables when I merged the two data sets. The next step was to make both data sets have the same level of observation since the Canadian census is recorded at the individual level and the ACS at the household level. In order to make the data compatible I decided to make this study about single person households so that both data sets could capture values in similar ways. After I narrowed down the Canadian census data to the variables that are usable, I renamed them to be easier to reference (I later do this with the ACS data as well).

Since both the data sets have different codes for missing values I had to change them to NA or 0 depending on what the code implied for each variable before I merged the data sets. Almost all of the missing value codes in the Canadian data are simply not in universe on the questionnaire which means I can only set them to NA's and work with a smaller sample size. While in the ACS data a few missing variable codes implied that the values were included in another column. For example the cost of electricity was often included in rental cost so I could simply set it to 0 and not lose that observation.

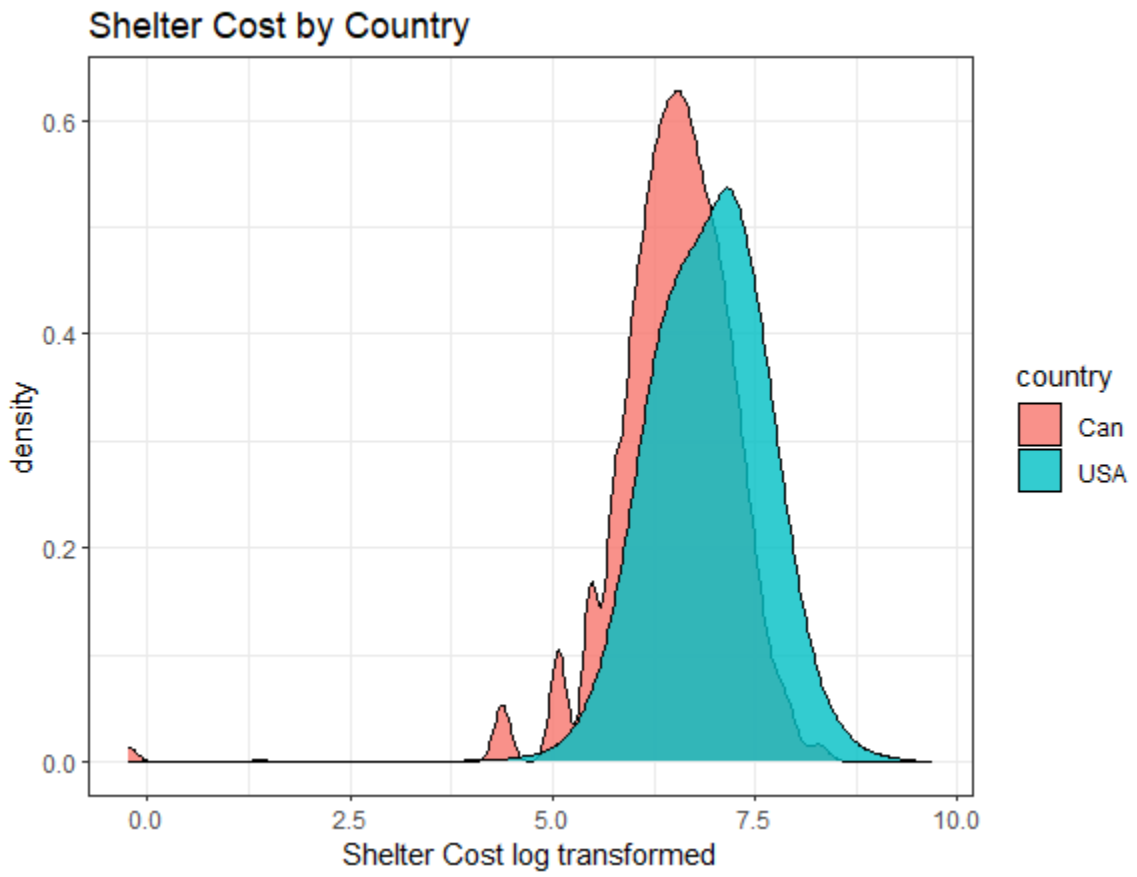
For extreme values I didn't observe anything that appeared to exist outside of the normal data collection process for the rest of the data set so I did not outright remove anything. To make the monetary values compatible I converted the Canadian dollars to 2021 US dollars via the average exchange rate from 2021.

A large part of my time while processing data went to converting codes to be compatible between the two data sets for categorical variables like race, sex, education, or citizenship. While some of these can convert very cleanly, many were not one to one fits and I had to use my best judgement to not set fundamentally different values equal to each other. One major example of this is in the different process of recording race between the Canadian census and ACS since the census has a much more detailed race code than the ACS. While I am fairly happy with the data I have for this write up, I am still looking to make a few changes or additions to variables in order to capture more of the effects on shelter costs without omitted variable bias.

Transformations

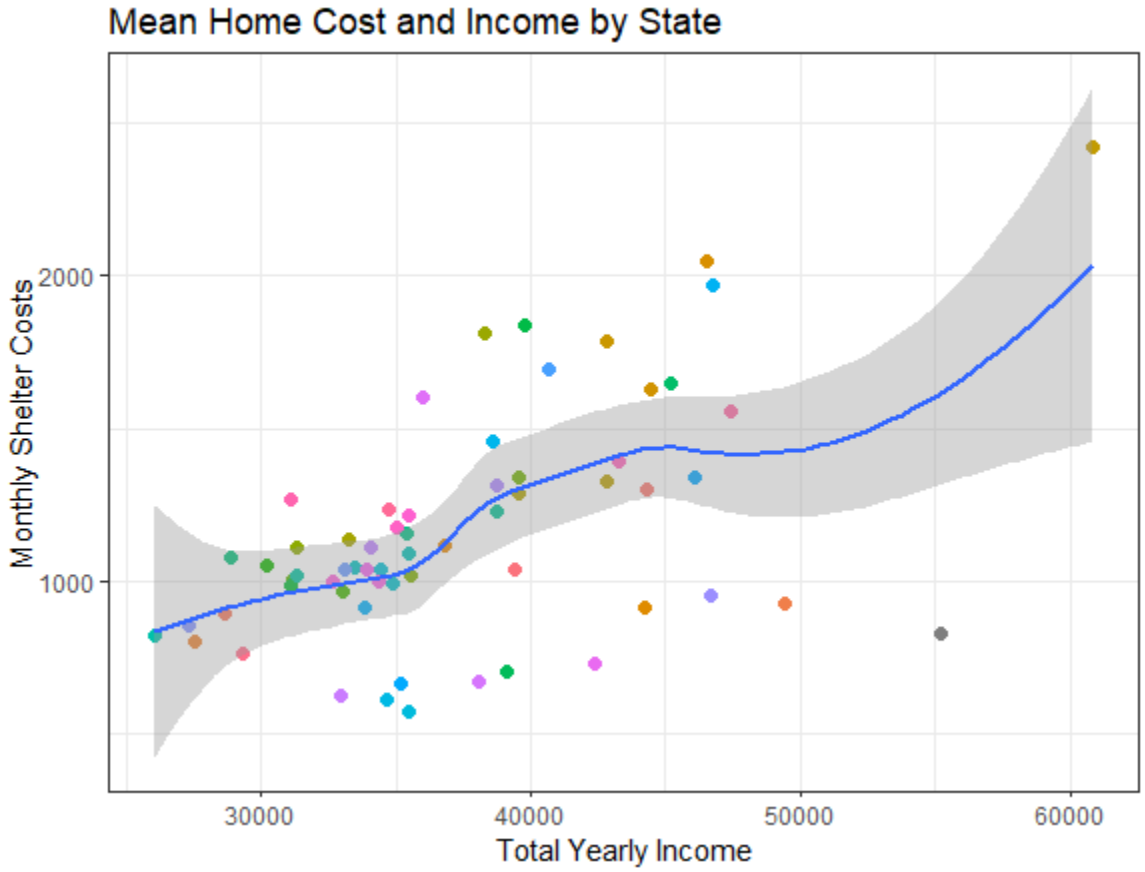
The first variable I looked at to see if it needed a transformation was shelter cost since it is my primary focus. The distribution of shelter cost is fairly positively skewed so I tried a log transformation and it seemed to give it a more normal distribution with only a few low outlines of the central cluster which seemed much better than the non transformed distribution. For total and wage income a log transformation also helps the positive skew by making them shift to be more normal.

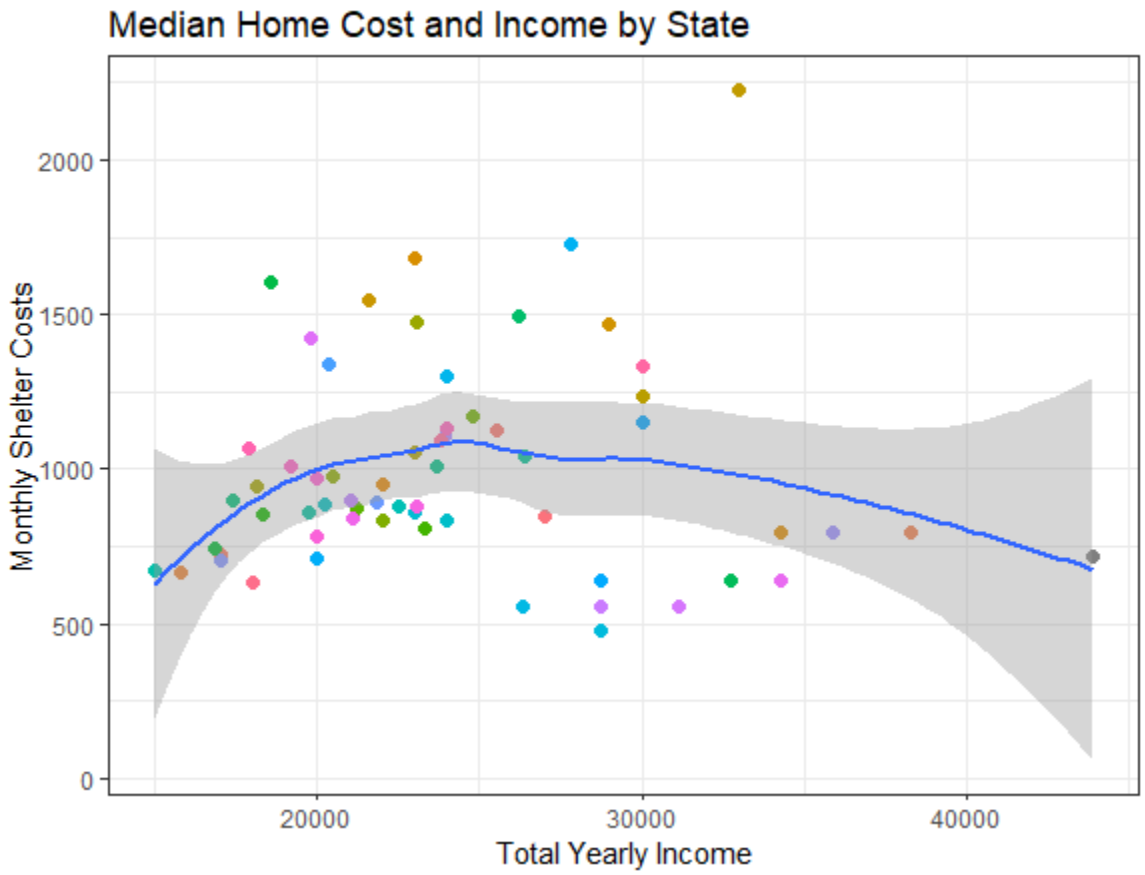
Visualization



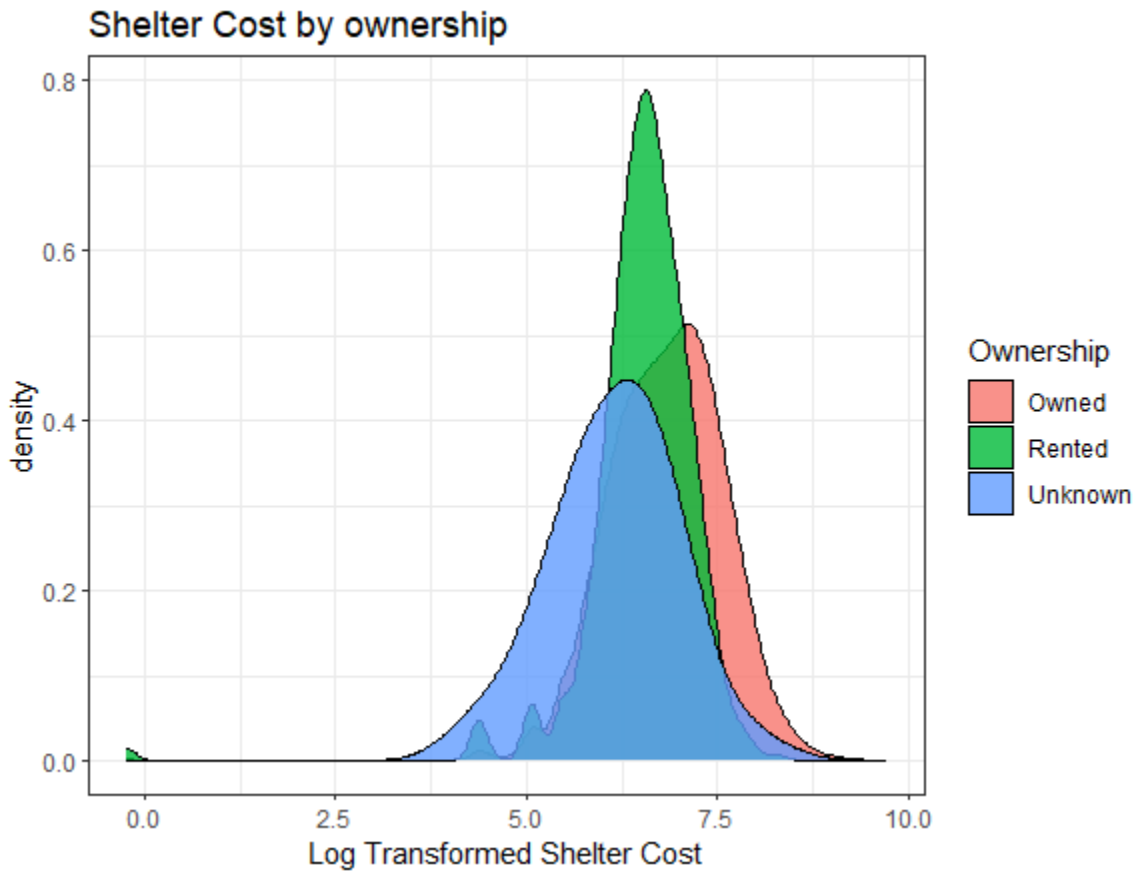
While this density plot of shelter cost is relatively simple, I think it is a necessary place to start for visualization. One reason this is an important graph is because it shows the overall distribution of my primary

outcome variable and suggests a slight difference in costs between the United States and Canada. Based on this density plot it seems that the United States may have a higher upwards bound in the shelter cost data than Canada while the Canadian shelter cost values appear more tightly distributed.





These figures show a scatter plot of the relationship between mean and median Yearly income and monthly shelter costs by state. I initially had these with the legend, but it was still hard to decipher which state was which so I decided to remove it. I think these are interesting because while both are effectively measuring the same thing, the figures display very different correlations between the two variables. The graph that is using mean values from each state has a fairly tight distribution around the trend line suggesting a positive correlation all the way throughout the data set, but the median figure suggests that after around $25,000 in total income the correlation becomes negative. The highly positive correlation at the end of the mean graph is being influenced by one outlying state, but even without it the trend is more positive than the median figure which could be a signal that the means of total income are being affected by large outliers. While the result from using means seems more intuitive, the resistance to outliers from using medians leads me to believe that the relationship could be slightly parabolic. Since this relationship does not seem particularly linear I would have to consider using something more complex than a simple linear regression to truly test it.



This is another simple density plot that is showing Shelter cost by both categories of tenure and unknown values. If I had just made the unknowns NA's, then it would be very easy to just say that this graph suggests a slight increase in shelter cost for owners when compared to renters. The unknown values came from both the ACS and Canadian census seemingly at random, which would suggest it could effect both groups equally pulling the owners costs more in line with renters costs. If the suggestive results from this density plot are to be believed, then it would imply that regions with higher ownership % could have higher shelter costs, but this density plot is far too vague to draw any real causal suggestions from.