Final Report

Nick Peters

2024-12-01

Topic

Housing costs across Canada and the United States

Motivation

Rent, mortgage payments, and utilities are the main costs that most people have to plan their budgeting and life around since they are usually the largest and most important bills that people face. Everyone knows that certain geographic areas have higher housing costs than others - such as New York City compared to Wilsall MT - but providing information on what other predictors are correlated with housing costs could prove to be very useful to normal people trying to live a better life and policy makers trying to evaluate the living conditions of their constituents. While people who work in expensive cities may not be able to move to a cheaper area, knowing what other variables effect housing costs could lead them to make positive lifestyle changes. As for policymakers, this issue is incredibly important because it could provide insight on potential discrimination on how people are being charged for housing. For example, if variables like citizenship, race, or gender are correlated with higher/lower housing costs, then it could provide the framework for governmental bodies to investigate the cause of those relationships.

Research Question

The research question of interest is how a few choice variables affect housing costs. These variables include income, marital_status, age, education level, ownership, and age. In order to answer this question I am using a few different linear regression models that will be shown in the findings section of this report.

Data

Canadian Data

The data set used in this study to analyze Canadian housing costs is the public use 2021 Canadian Census micro data. The Census provides a wealth of information on the Canadian population and housing costs with access to over 140 different variables which allows for the study of many different relationships.

American Data

The data set used in this study to analyze housing costs in the USA is a 2021 American Community Survey sample from IPUMS. The ACS is a household level survey that focuses on the American labor force and housing.

Data Processing

The Canadian Census data was downloaded as a CSV from Statistique Canada and the variables in the initial data set are what I used to define the IPUMS ACS extract so that I could carry as many variables as possible through the analysis. While a lot of the variables in the Canadian data are not recorded in the ACS, all of the variables that should have a strong relationship with housing costs were consistent across both data sets. It is unfortunate to lose a lot of data points on potential edge case relationships, but it should not have a significant effect on the overall results of the analysis.

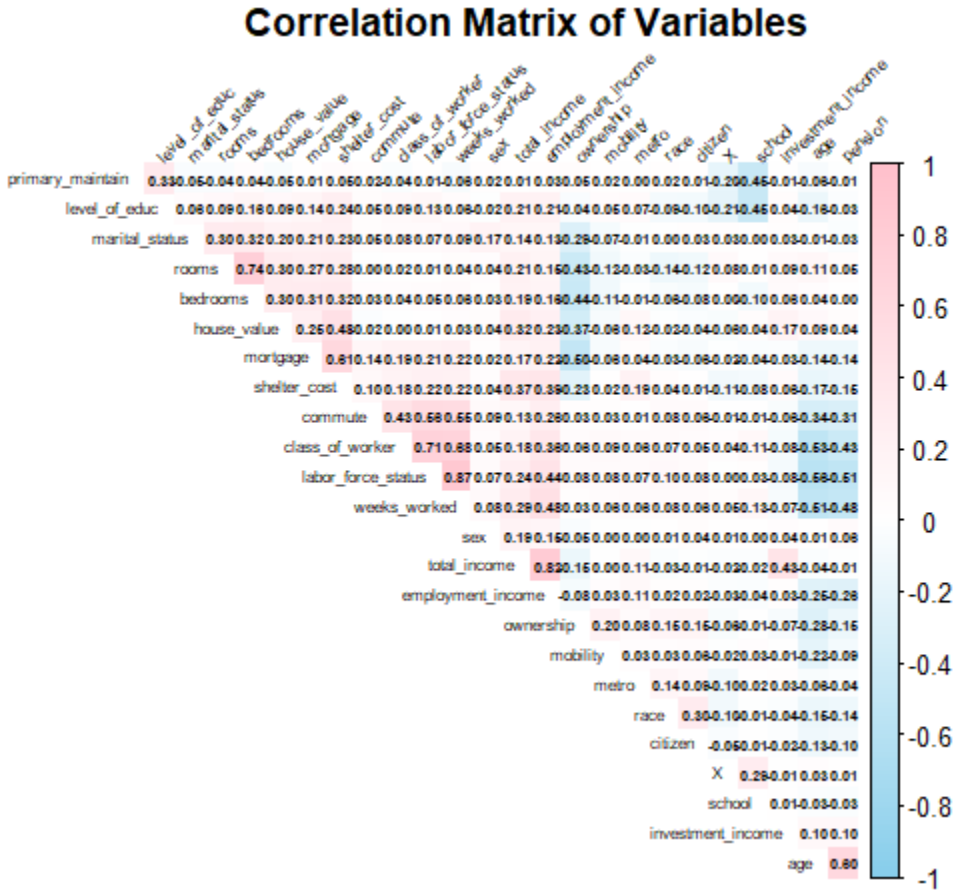
For monetary values, I transformed the Canadian dollar values to American dollars using the average 2021 conversion rate. In order to create a comparable housing cost metric in the ACS, I combined the variables that align with what makes up the value for Shelter cost in the Canadian Census. During this process it became clear that there were values that were generated outside of the normal data gathering process for mortgage payments and owncost in the ACS. In order to combat this, I used best judgement for different cases where the true value should either be 0 or NA.

I do have to disclose that I used AI in order to parse the documentation for the Canadian data so that I could find the not available/applicable codes and transform them into the applicable values (NA or 0)

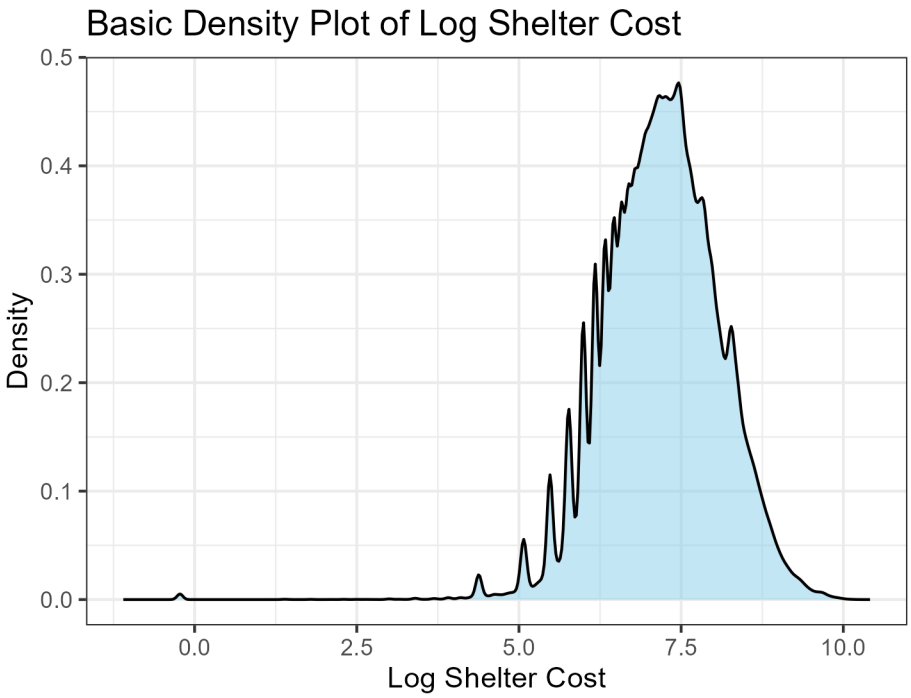
Transformations

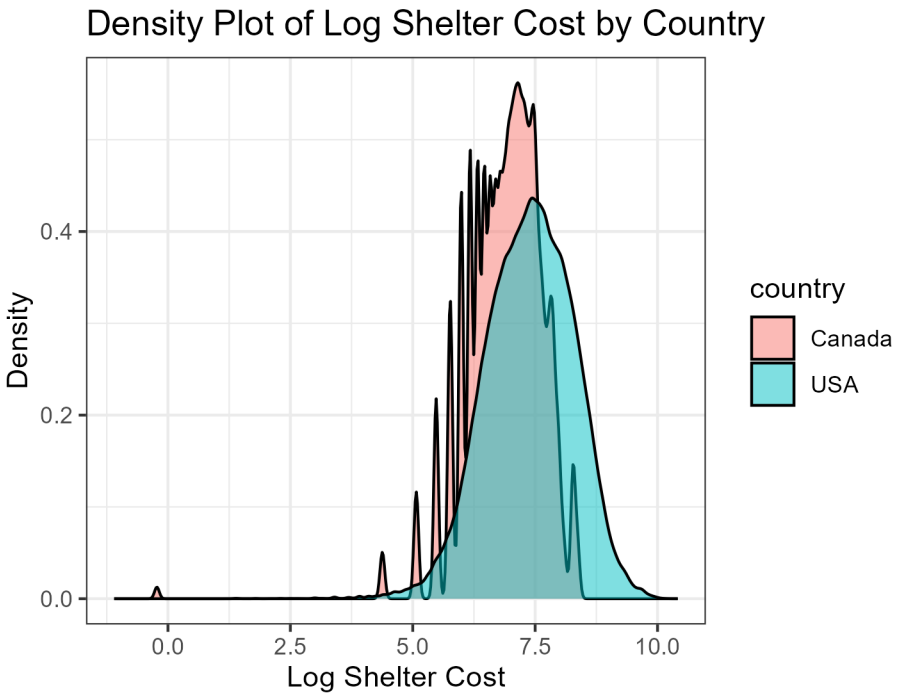
All of the monetary values that are not binned have been log transformed for exploratory and econometric analysis. I transformed them because they all have clear positive skew which can be dealt with using the log transformation.

Exploratory analysis

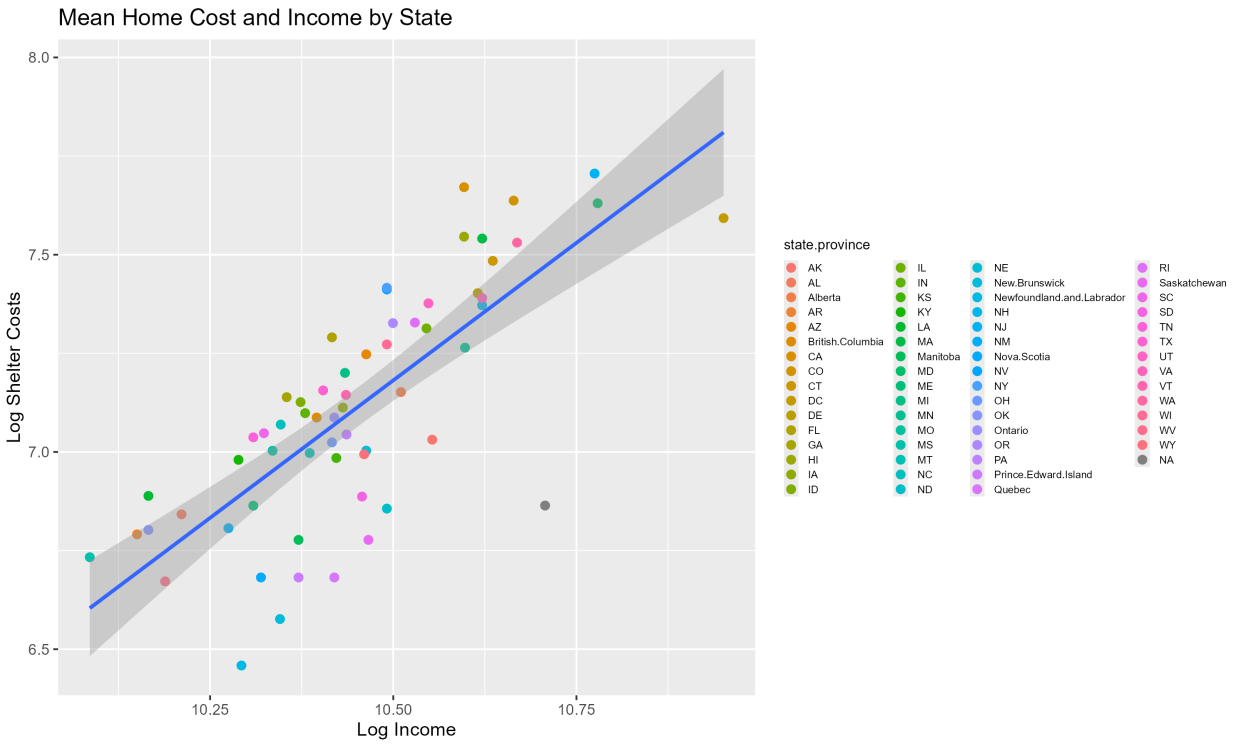


This correlation plot shows that housing costs are relatively positively correlated with income, education, rooms, and marital status. They are also negatively correlated with ownership. Other than those suggestive relationships, there is no strong correlation between shelter cost and other variables.

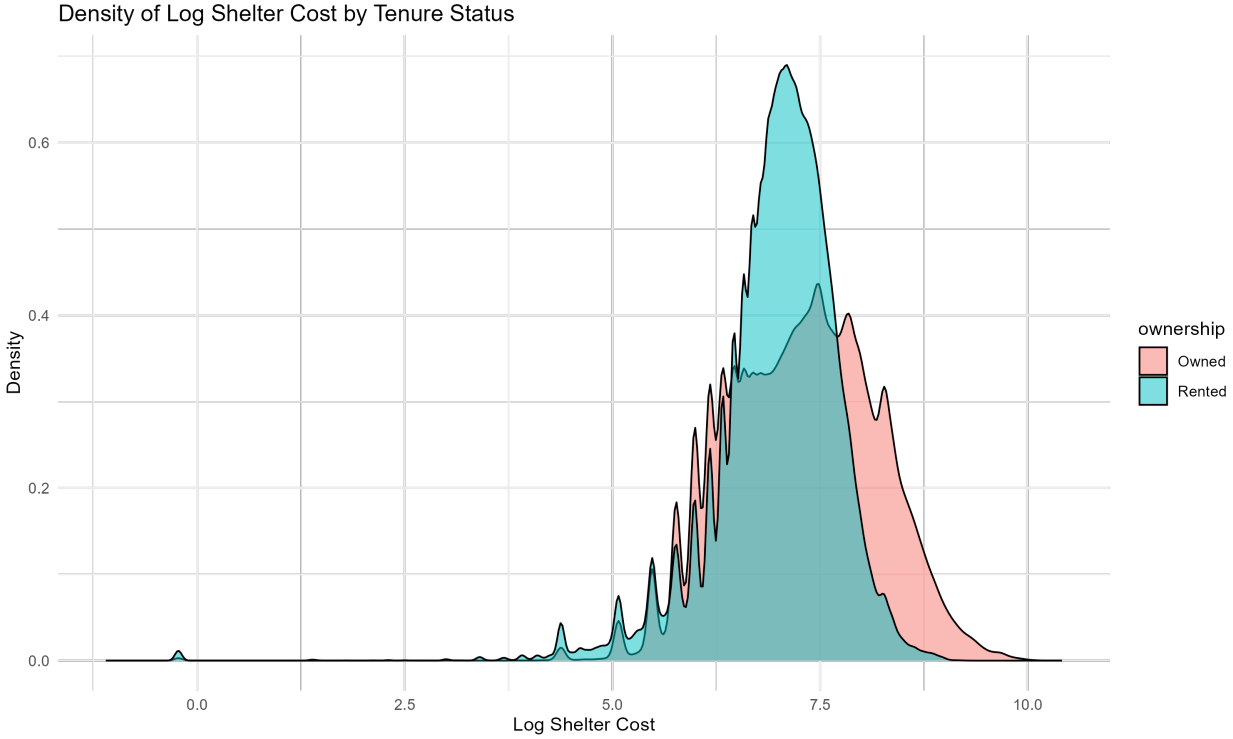


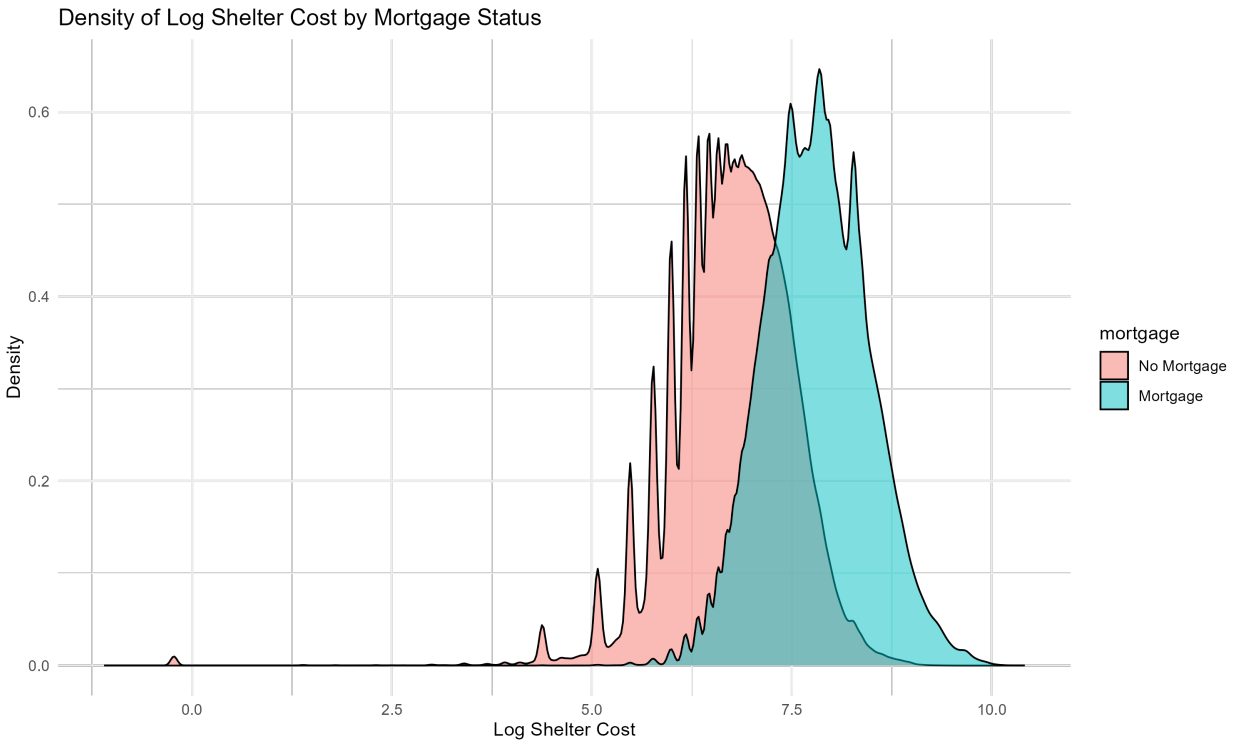


The density plots of log transformed housing costs above show that the transformation gives the data a relatively normal distribution. There is a slight negative skew, but the non transformed version is worse to the positive side. The second plot, which is grouped by country, suggests that the USA has slightly higher housing costs than Canada. One explanation for this suggestive relationship could be that the method used to obtain housing costs for the ACS included extra variables that were not recorded for the Canadian data.

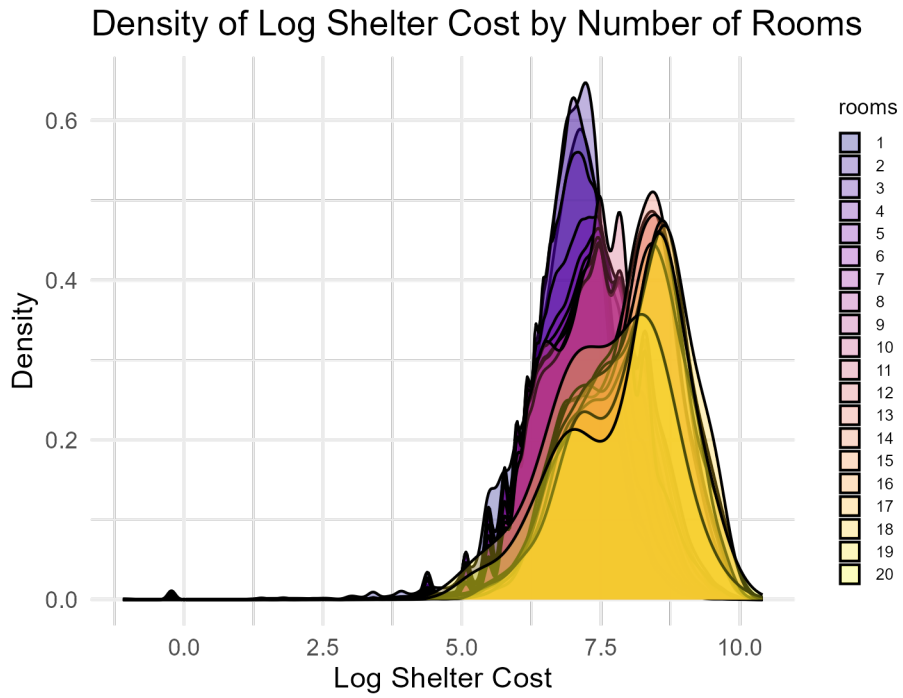


The above graph showing the relationship of log income and log shelter costs seems to support the correlation plot with a positive linear relationship between the two variables. There is a data point for NA states and it is noticeably lower than the other data points with known states. This could suggest that these observations should be removed if they are not randomly distributed.

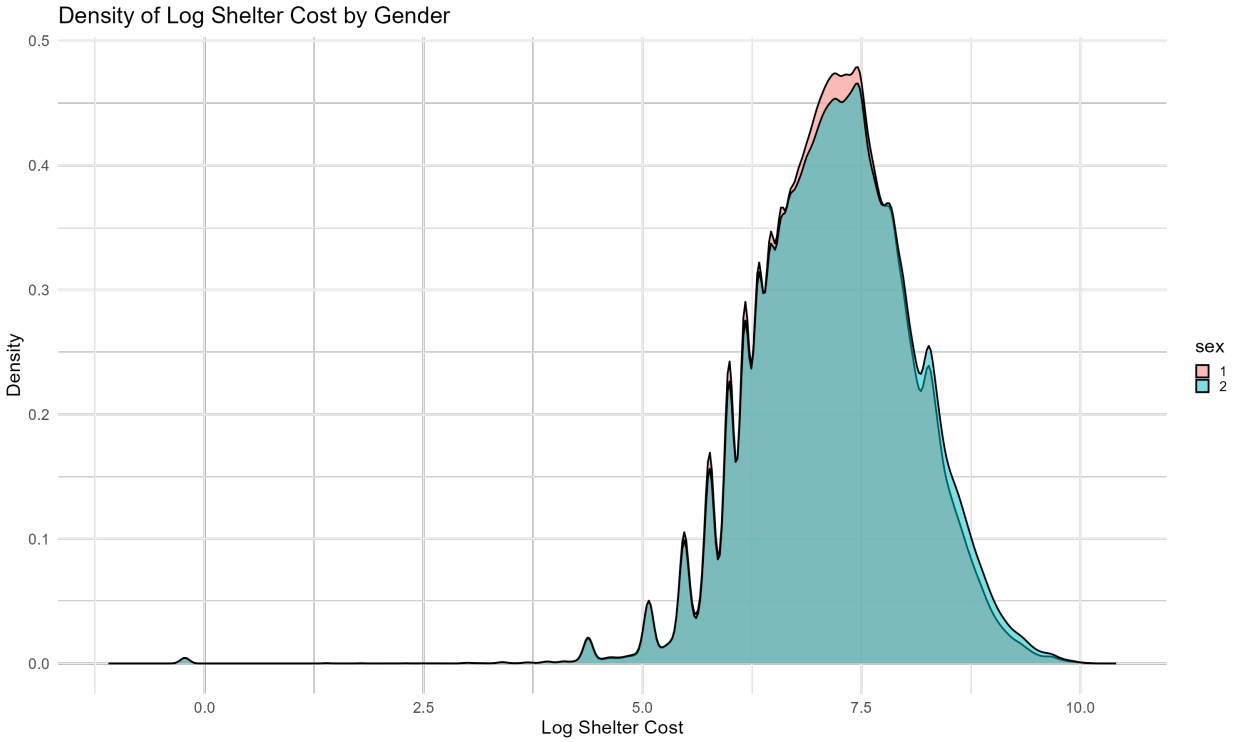


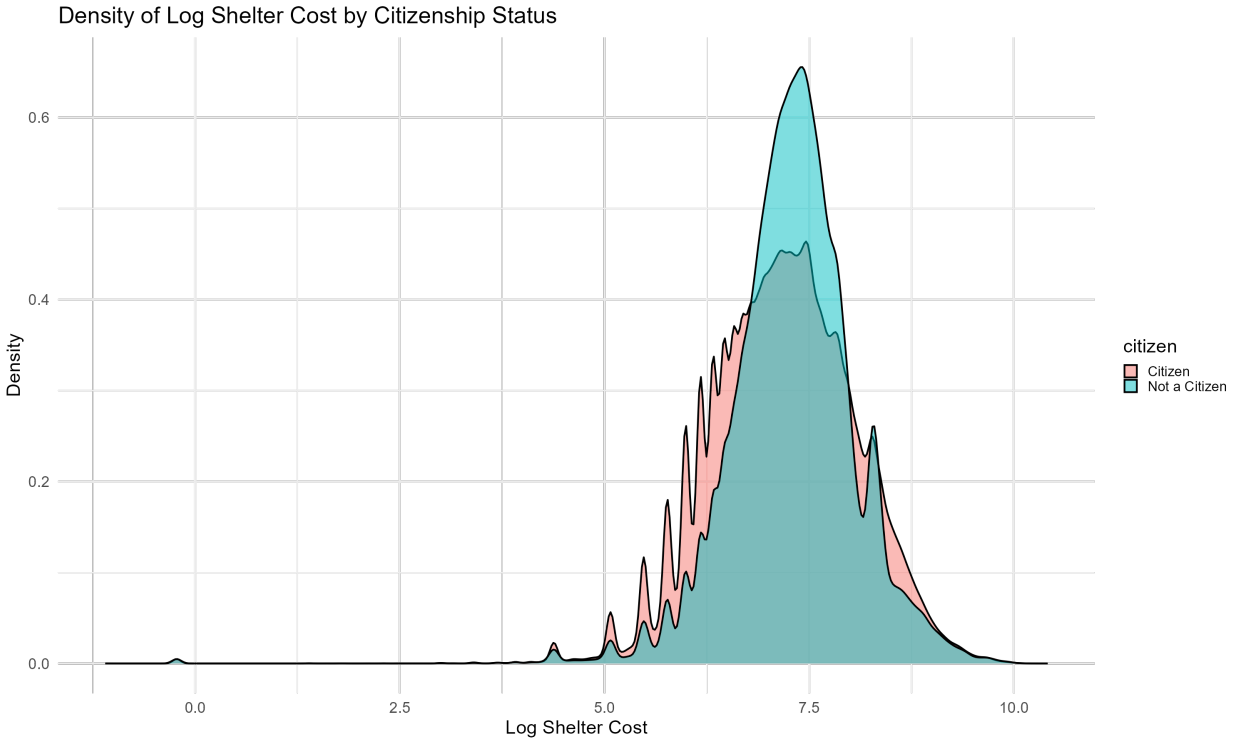


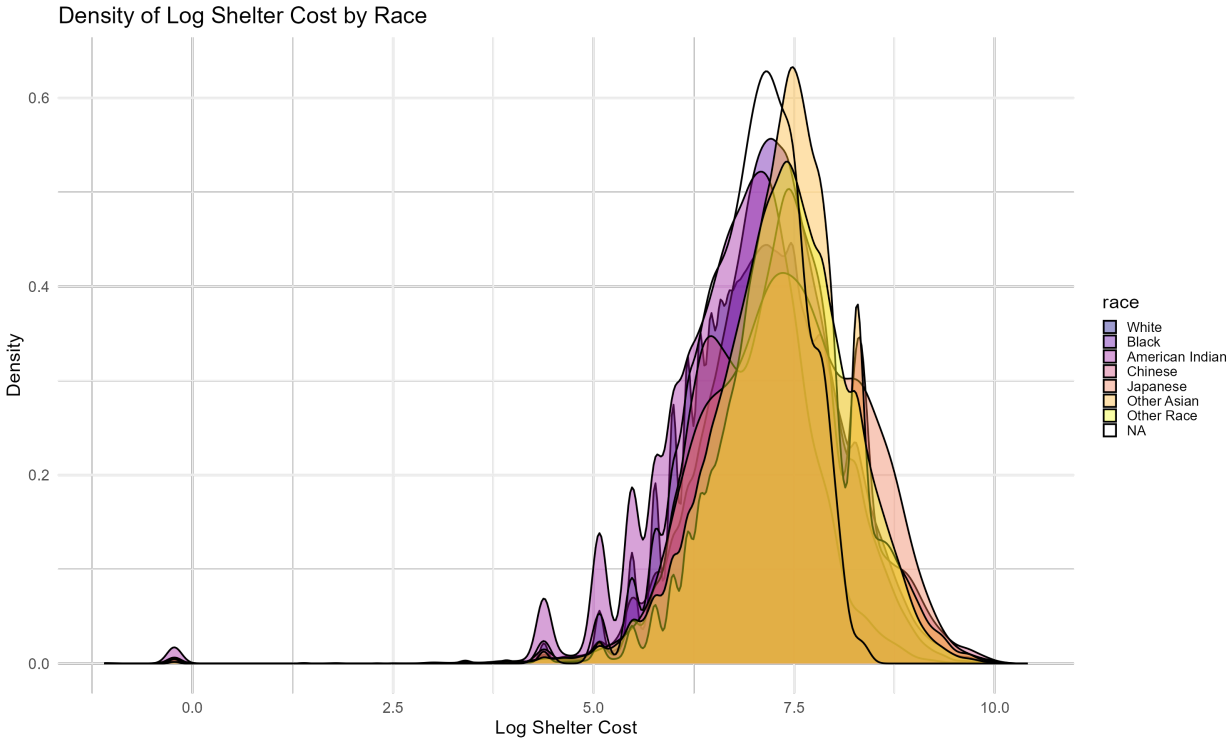
While the density plot that groups shelter costs based on ownership status does not suggest any difference in values, the density plot grouped by mortgage presence implies a strong positive relationship between mortgages and shelter costs. This is backed up by the correlation plot and makes intuitive sense when thinking about mortgages. Since the mortgage amount should be picked up in the ownership group of graph 1, it would imply that once homeowners payoff their mortgage the overall cost of housing goes down when compared to renters.



While this graph does suggest a positive relationship between the number of rooms in a home and shelter cost, it is not as obvious as I thought it would be. One reason for this could be that houses/apartments in big cities are less likely to have more rooms, but would still be more expensive than their larger rural counterparts.







These three density plots do not suggest any significant relationship for gender, race, and citizenship.

Note on Prediction Attempts and Struggles

My initial plan for this project was to conduct a predictive analysis of housing costs, but I was not able to generate a satisfactory model/I underestimated the compute time of the more complex models. I was able to tune simple elasticnet and linear regression models, but they returned a rmse of slightly over 1000. I also attempted to use extreme gradient boosting and random forest methods, but in order for my computer to be able to tune the models I had to create a very small sample which negatively effected the results. I still have a lot of code for what I wanted to do which will be pushed to github with this writeup. In order to adhere to time constraints and deadlines I have ultimately decided to stick with a descriptive regression for my econometric analysis.

Econometric Analysis

Motivation

The descriptive analysis in this section provides more concrete and quantitative evidence for the relationships that were first suggested during the exploratory phase. While there won't be any real causal evidence from my econometric analysis, it will still be useful to expand my exploratory analysis.

Methods

Model The first model used to study the relationships of housing costs is a basic linear regression with log shelter cost as the dependent variable.

ln(shelter_cost_i) = B0 +B1ln(total_income_i + 1) +B2ln(house_value_i + 1) +B3ln(employment_income_i+1) +B4ln(investment_income_i + 1) +B5bedrooms_i +Bumarital_status +B6age_i +Bvlevel_of_educ +Bllabor_force_status +Bsstate_province +Btownership+Busex +e_i

A lot of the log transformations have been offset by one because the large amount of 0's in the raw data affected the results.

Results

	Dependent variable:
	Log Shelter Costs
log_total_income	0.017***
	(0.0003)
log_house_value	0.303***
	(0.001)
log_employment_income	0.013***
	(0.0002)
log_investment_income	-0.012***
	(0.0002)
bedrooms	0.080***
	(0.001)
marital_status2	0.108***
	(0.001)
Observations	1,398,321
R^2	0.360
Adjusted R^2	0.360
Residual Std. Error	0.763 (df = 1398203)
F Statistic	6,731.303*** (df = 117; 1398203)
Note:	*p<0.1; **p<0.05; ***p<0.01

Results

The strongest relationship that the model captures is that each additional bedroom is estimated to correspond with an 8% increase in shelter costs. That result is significant at the 99% level. One of the more interesting results from this regression is that being married is estimated to be associated with 10% higher shelter costs than for those who are not married. One potential reason for this observed relationship is that married individuals are probably more likely to have kids, which could result in the desire to live in a nicer home which is not captured by this model. One other result of note that is not present in the above output is that once people reach around 50-55 years of age, it is estimated that they will spend around 10% less on housing than the omitted group. This pattern continues as people get older too. Almost all of my results are significant at the 99% level which could be the result of over fitting the model.

Conclusion

Even though this study has not provided any information on causal relationships between variables in the Canadian or American data, it has provided a thorough foundation for anyone who wants to study the causes of higher housing costs further. Overall, I am disappointed that my prediction models did not end

with satisfactory results. I have two prediction models that could have lead to interesting results if they can successfully compute. I created an XG boost model for only Canada using all 140 variables in the data set and then wanted to compare it to a prediction model that uses the ACS and Canadian data, but with only around 40 variables. If I can get the models to run before my presentation on Tuesday, I will talk about them then.