ASSIGNMENT 2 QUANT
ARON LESSER
9/14/20

Instructions

For each continuous variable in your dataset, calculate the:
1. sample mean
2. sample standard deviation
3. the 95-percent confidence interval for the population mean
4. the interquartile range.

Create a histogram to illustrate the distribution of each variable and describe the distribution in a sentence or two.

For each categorical variable in your dataset calculate the 95-percent confidence interval for the proportion of the population in each category.

Variables for Renters in Georgia

Continuous

1. Income
2. Number of persons in this household
3. Monthly rent

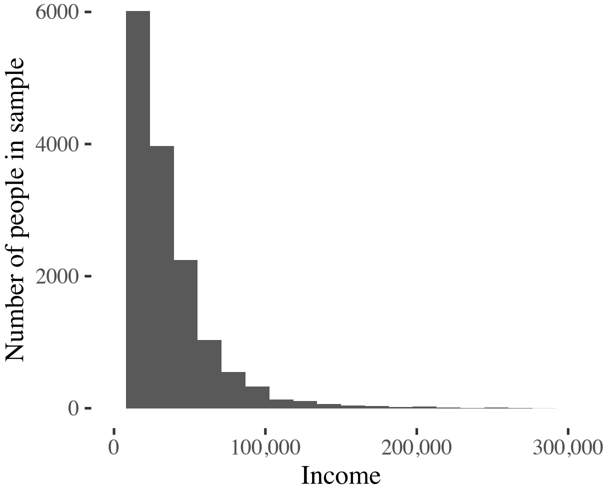
Categorical

1. Recoded detailed Hispanic Origin
2. Mobility status (lived here 1 year ago)

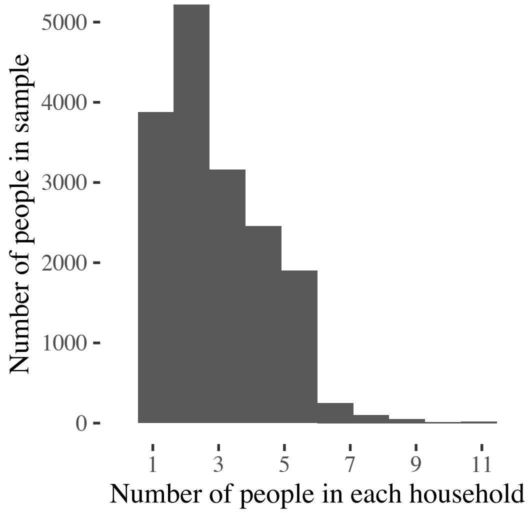
Distribution of Continuous Variables

	Monthly Rent ($)	Income ($)	Household Size (persons)
Sample Mean	882	32,758	2.79
Standard Deviation	488.2142	40,216	1.711492
95-percent confidence interval for the population mean	874.9043 to 889.5501	32155.30 to 33361.72	2.763519 to 2.814861
Interquartile Range	540 to 1,100	11,600 to 40,000	2 to 4

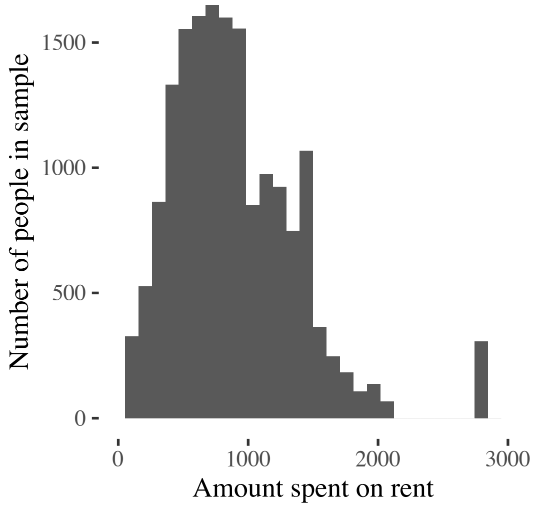
To clean the data, I filtered for incomes above 0. In my original dataset, around 25% of individuals had negative incomes. I was not sure if this was a colection discrepancy or a phenomenon about debt.



In this histogram, I plot the income of individuals from the sample. I altered the bin size in order to better display the trends, and I changed the limits of the X axis to end at $300,000, since there were few outliers that detracted from the visual trends. Interestingly, the standard deviation is greater than $40,000 while the mean is just above $32,000. This demonstrates that in the sample, some individuals have singnificantly larger salaries, though not enough to greatly skew the mean.



In this histogram, I adjusted the bins and the scale to 12 to better display the data. This helps to demonstrate the household size trends. Notably, the interquartile range is from 2-4 and the mean is ~2.8.



While the mean of this sample is ~$880 for individual rent payments, there is a small group that pays between two and three times as much for rent. In a next stage, it will be intersting to consider the relationship between amount spent on rent and income. It may make sense to get the rent/income proportion and add it to my analysis.

Distribution of Categorical Variables

Mobility Status: If the individual moved within the last year

Mobility Status	Proportions of Sample (%)	95-percent confidence interval (%)
Moved within the U.S.	0.257247	0.250690 to 0.26380
Did not move	0.733970	0.727342 to 0.74060
Moved outside of the U.S.	0.008784	0.007384 to 0.01018

Hispanic/Latino Ethnic Origin: Most represented Nationalities

Ethnicity	Proportions of Sample (%)	95-percent confidence interval (%)
Non-Hispanic	89.992387	89.5422409 to 90.44253
Mexican	5.281958	4.9464527 to 5.61746
Puerto Rican	1.165310	1.0043340 to 1.32629
Guatemalan	0.837384	0.7006982 to 0.97407
Honduran	0.392341	0.2985706 to 0.48611

Hispanic/Latino/Latinx Ethnicities with less than 0.39%: Dominican, Ecuadorian, Colombian, Panamanian, Cuban, Salvadoran, Peruvian, Spaniard, Venezuelan, Uruguayan, Nicaraguan, Costa Rican, Argentinina, Chilean, Paraguayan, Other South American, Other Central American, All Other Latino.

I decided to include all groups represented rather than simply mutating this into Hispanic and Non-Hispanic. However, I recognize that this may make further analyses difficult and would consdier making this alteration in the future. These decisions raise important questions about indentitary erasue in data analytics, and the costs and benefits of groupings.