Affordable Housing in Worcester, Part 1: A Preliminary View of the American Community Survey and US Census Data

By
Joshua Oliver, Thomas E. Conroy, Ph.D., and Mary Fowler, Ph.D.

Fall 2019

WORCESTER STATE UNIVERSITY
Aisiku STEM Center

Department of Urban Studies
CITYLAB
Worcester State University

INTRODUCTION

Across the country, housing, one of our most basic needs, generates discussion, debate, and dissension in America's cities. Deliberations about how to approach the housing of a population often exclude affected groups, follow political agenda not data, or are made for short-term rationales without taking longer views. All of which can lead to often unnecessary raised temperatures and ill-feelings.

The City of Worcester is experiencing its own growing pains around issues of affordable housing. As economic development decisions are made and new structures are built (or old ones are rehabilitated), concerns about skyrocketing rent, public housing access, affordability, relocation, geographic bias, perseverance/burgeoning of cultural traditions, appropriateness of new buildings and preservation of older ones, feasibility of public transportation systems, and a variety of gentrification-related issues enter the public square for consideration.

The purpose of this report and a series of reports that will follow is to provide facts culled from public data sets and repositories meant to inform public discussion about these weighty topics for all interested parties to use as reference. More specifically, this particular report provides tables and data visualizations of a number of affordable housing-related datasets.

This research team would like to express gratitude to the Department of Mathematics and the Department of Urban Studies for supporting this work; and the Aisiku STEM Center for its generous funding of the research and the WSU CityLab for preparing it for and publishing the report. Appreciation is also extended to our university research colleagues and community partners, especially Forrest Hangen and Professor John Holbrook for their insights and tips during research team meetings.

Finally, a couple of quick notes:

- As decimals are rounded, numbers and percentages may not add up as neatly as expected.
- Some percentages are formed by calculation of data in the US Census or American Community Survey (ACS) tables. These are noted throughout.
- Lastly, "Worcester County (Not City)" or variants mean counts from the City of Worcester (not all cities) have been removed from county totals.

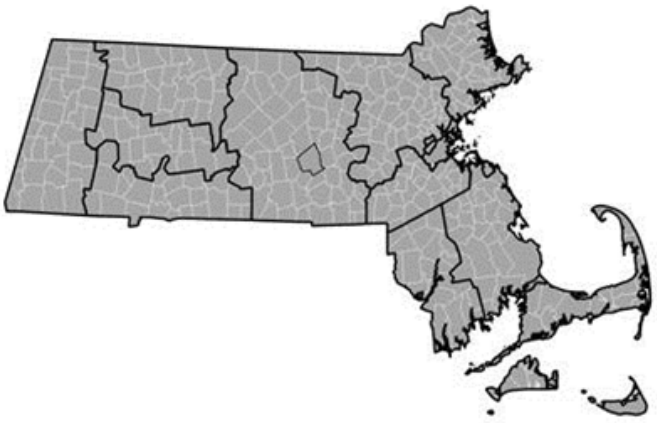
Joshua Oliver, Tom Conroy, and Mary Fowler

Table of Contents and Figures

Unless otherwise noted, all tables are from US Census, American Community Survey (ACS) 2017, 5-year estimates, or calculated using those data tables.

Modeling Percent Rent Burdened of Census Block Groups by Median Income and Percent in Poverty.

By Joshua Oliver



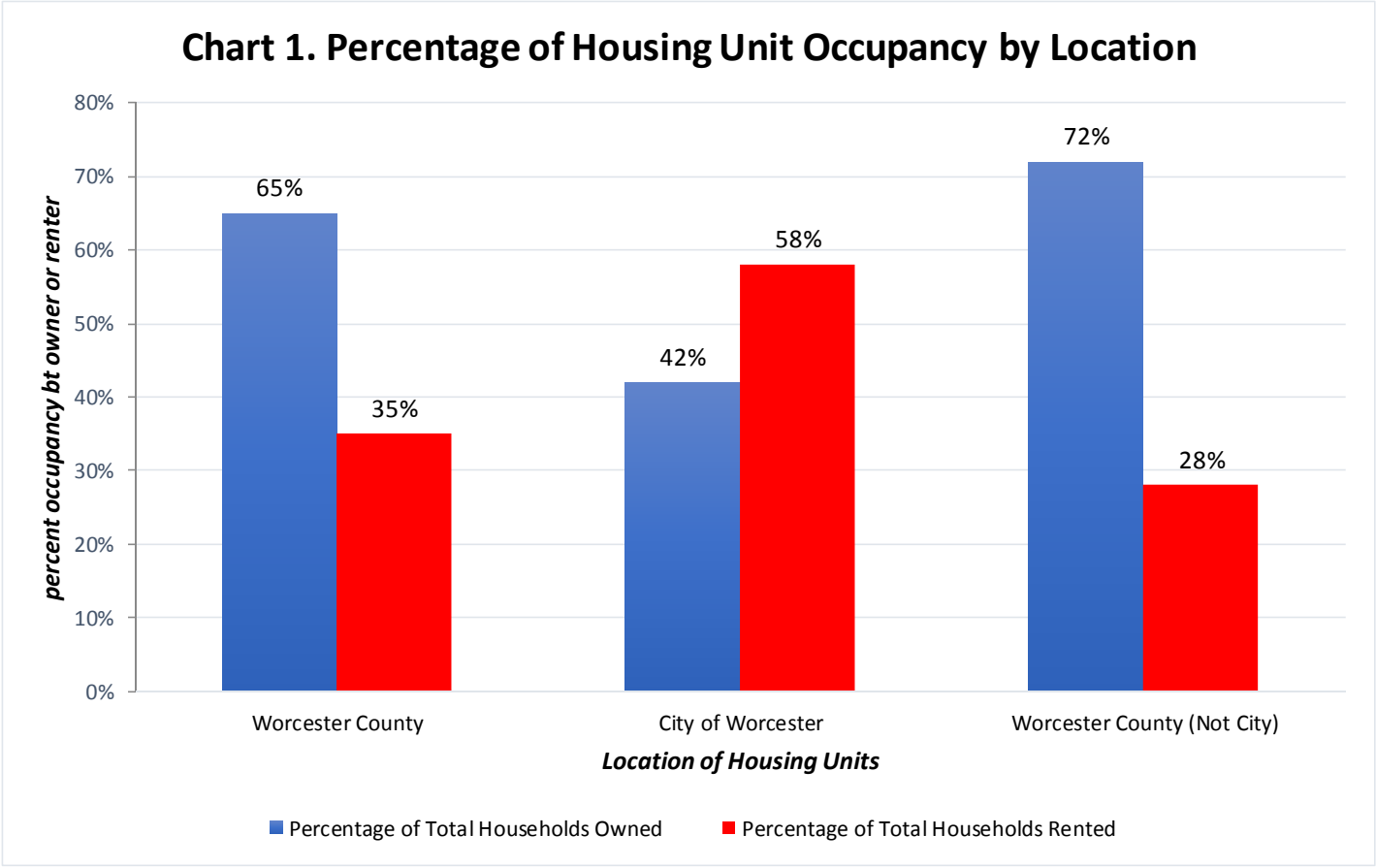
References

TENURE (OWNING/RENTING)

This data set looks at all Worcester area housing units, which, according to the US Census, include a house, an apartment, a mobile home, a group of rooms, a single room occupied as a separate living quarter, or vacant units intended for occupancy.

Table 1: Number of Housing Units by Location			
	Worcester County	City of Worcester	Worcester County, Not City
Owned Housing Units	198010	29825	168185
Percentage of Total Households Owned	65%	42%	72%
Rental Housing Units	107965	40967	66998
Percentage of Total Households Rented	35%	58%	28%

Note that while 23.1% of Worcester County's total housing is located in the City of Worcester, the city's land mass comprises only 2.5% of the county's geographic area. This suggests the importance of studying the City of Worcester with respect to housing and affordable housing. So, too, does the high renter-to-owner ratio illustrated below.



VACANT HOUSING

This data set looks at housing units that are classified as vacant in three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1: Vacant Housing Units: Worcester County		
Housing	Number of Housing Units	Percentage of Total Vacant Housing Units (%)
For Rent Only	5843	22%
Rented, Not Occupied	1282	5%
For Sale Only	2695	10%
Sold, Not Occupied	1620	6%
Seasonal, Recreational, Occasional Use	3716	14%
Other	11401	43%
Total	26557	100%

Table 2: Vacant Housing Units: City of Worcester		
Housing	Number of Housing Units	Percentage of Total Vacant Housing Units (%)
For Rent Only	2086	29%
Rented, Not Occupied	650	9%
For Sale Only	541	7%
Sold, Not Occupied	322	4%
Seasonal, Recreational, Occasional Use	518	7%
Other	3127	43%
Total	7244	100%

Table 3: Vacant Housing Units: Worcester County (Not City)		
Housing	Number of Housing Units	Percentage of Total Vacant Housing Units (%)
For Rent Only	3757	19%
Rented, Not Occupied	632	3%
For Sale Only	2154	11%
Sold, Not Occupied	1298	7%
Seasonal, Recreational, Occasional Use	3198	17%
Other	8274	43%
Total	19313	100%

Overall, Worcester County has over 300,000 housing units, 7.99% of which are vacant. In the City of Worcester, 9.28% of its almost 80,000 housing units are vacant; and in Worcester County outside the city, there are over 250,000 housing units with 7.59% that are vacant.

Chart 1. Vacant Housing Units by Type, Ownership Status, and Location

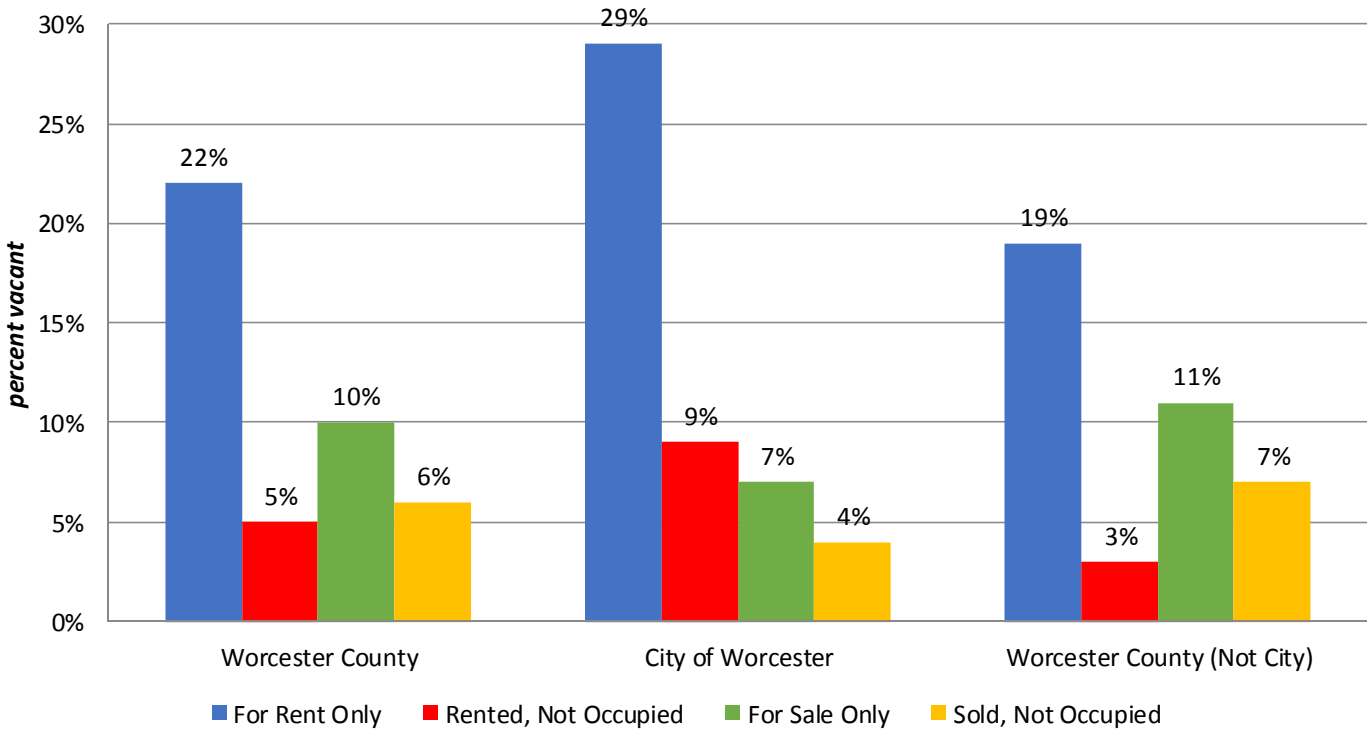
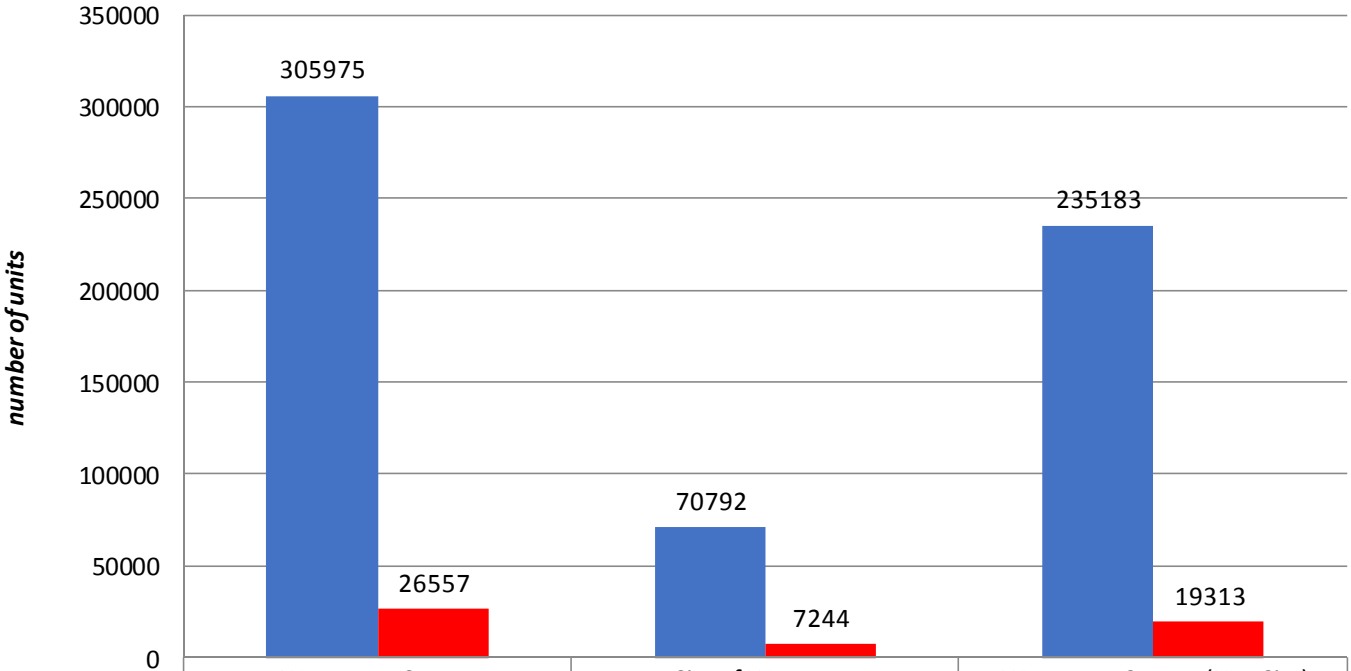


Chart 2. Total Vacant Housing Units By Location



	Worcester County	City of Worcester	Worcester County (Not City)
Total Units	305975	70792	235183
Vacant Units	26557	7244	19313

GENERAL INCOME

This data set looks at total household income of all occupied housing units in three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1: Household Income of Occupied Units

	Worcester County		City of Worcester		Worcester County (Not City)	
Income Level	Number of Households	Percentage of Households (%)	Number of Households	Percentage of Households (%)	Number of Households	Percentage of Households (%)
Less than $10,000	16523	5%	7646	11%	8877	4%
$10,000-$14,999	14687	5%	5522	8%	9165	4%
$15,000-$24,999	25396	8%	8212	12%	17184	7%
$25,000-$34,999	25090	8%	7079	10%	18011	8%
$35,000-$49,999	31821	10%	8920	13%	22901	10%
$50,000-$74,999	49262	16%	11610	16%	37652	16%
$75,000-$99,999	39471	13%	7504	11%	31967	14%
$100,000-149,999	54770	18%	8707	12%	46063	20%
$150,000-$199,999	24784	8%	2973	4%	21811	9%
$200,000 or more	24172	8%	2549	4%	21623	9%
Total	305975	100%	70792	100%	235183	100%

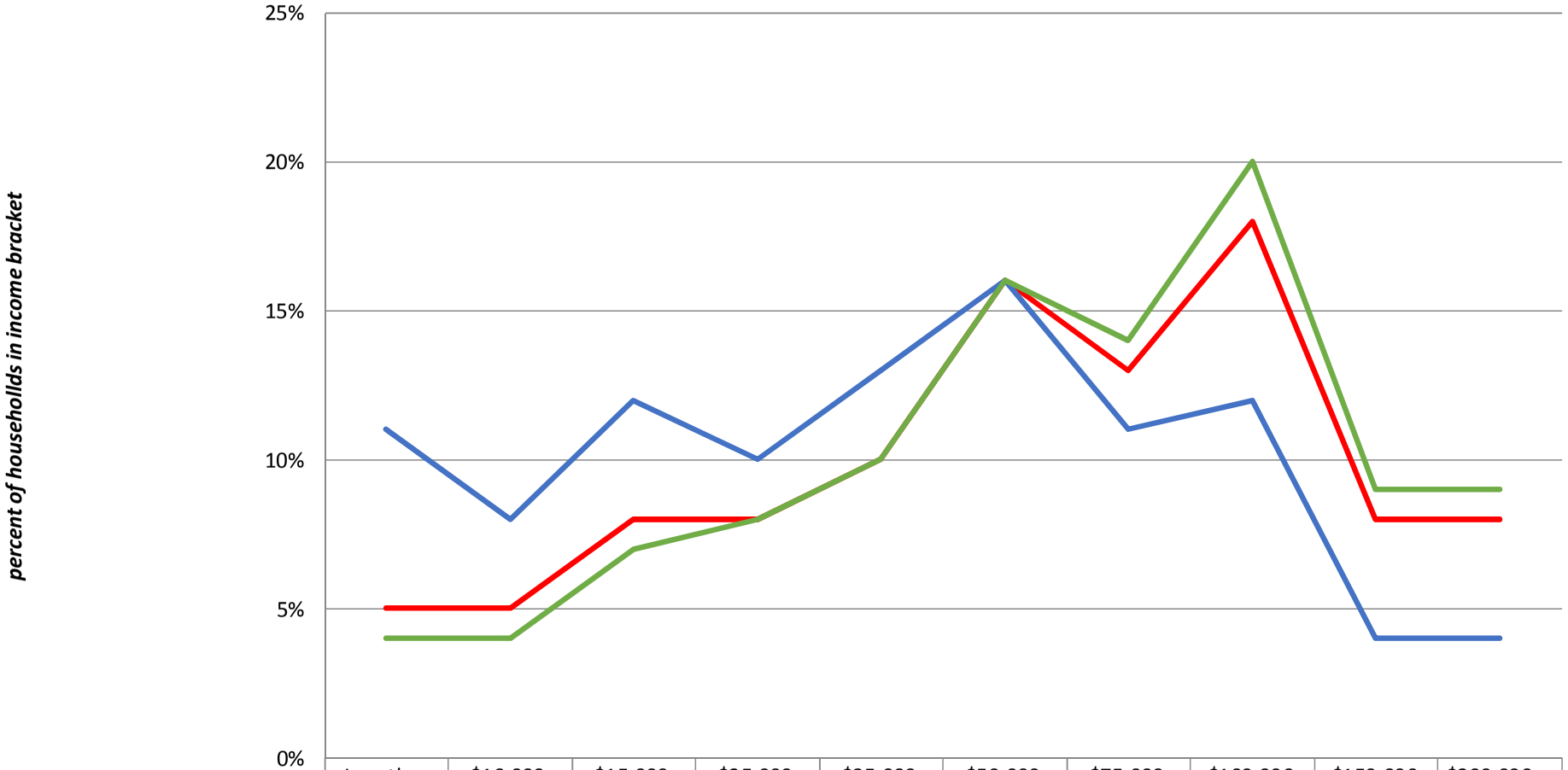
The median income for households in Worcester County is $69,313, and the median income for households in the city of Worcester is $45,869. (Median household incomes are from *Table S1903: 2013 - 2017 American Community Survey, 5 – Year Estimates of the median income in a 12-month period in 2017 inflation-adjusted dollars.*)

The estimated median income for households in Worcester County outside the city was calculated to be $76,786. The median household income here was calculated by 1) estimating which income level 50% of the households fall under and then 2) determining the point within that range in which exactly 50% have most of that income.

Table 2: Median Household Incomes For Each Area

Location	Median Household Income
Worcester County	$69,313
City of Worcester	$45,869
Worcester County (Not City)	$76,786

Chart 1. Household Income Distribution of Occupied Units by Location



	Less than $10,000	$10,000-$14,999	$15,000-$24,999	$25,000-$34,999	$35,000-$49,999	$50,000-$74,999	$75,000-$99,999	$100,000-149,999	$150,000-$199,999	$200,000 or more
Worcester County	5%	5%	8%	8%	10%	16%	13%	18%	8%	8%
City of Worcester	11%	8%	12%	10%	13%	16%	11%	12%	4%	4%
Worcester County, Not City	4%	4%	7%	8%	10%	16%	14%	20%	9%	9%

POVERTY

The US Census Bureau uses a set of income thresholds that vary by family size and composition to determine who is in poverty. If a family's total income is less than the family's threshold, that family and every individual in that household, is considered to be in poverty. To determine the pverty threshold for families/households with more than 8 persons, add $4,180 per each additional person. (Note: The two-column table immediately below lists the 2017 ASPE poverty guidelines for the United States, excluding Hawaii and Alaska.)

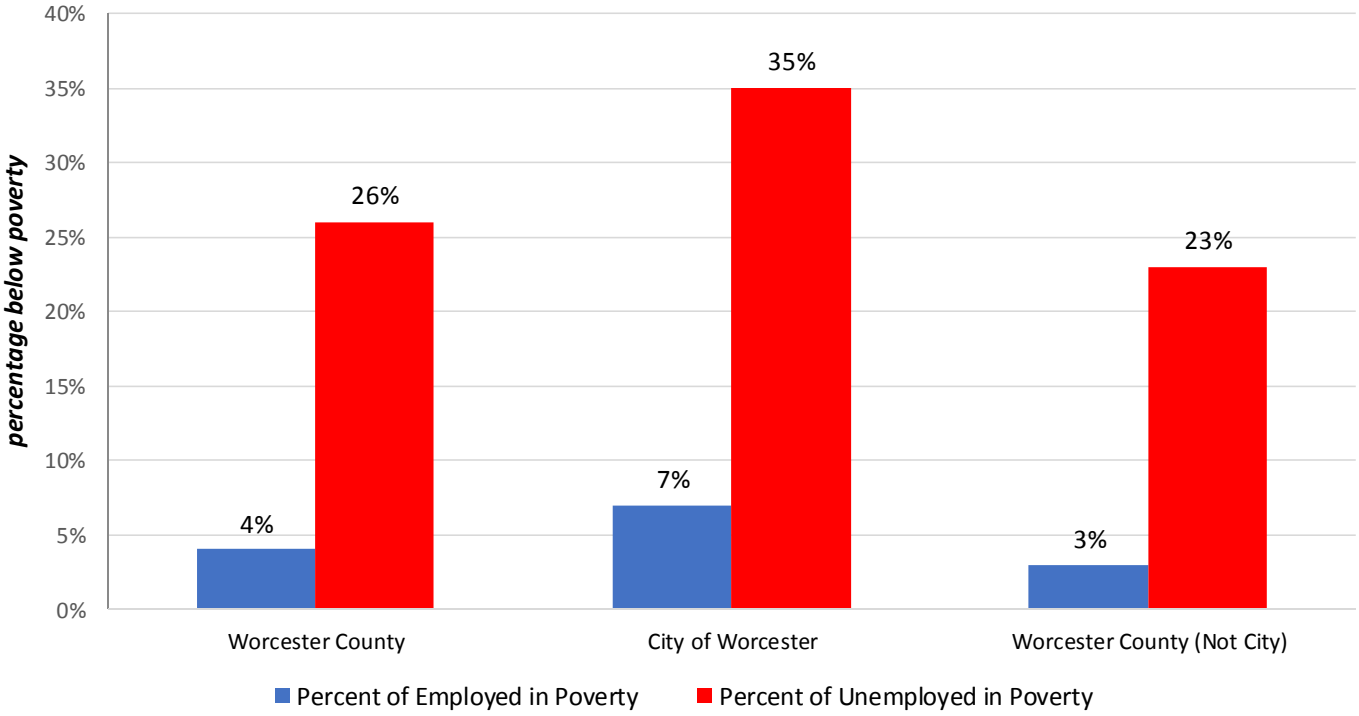
Persons in Family/Household	Poverty Threshold
1	$12,060
2	$16,240
3	$20,420
4	$24,600
5	$28,780
6	$32,960
7	$37,140
8	$41,320

This data set looks at the population for whom poverty status can be determined, which according to the Census Bureau, is the total population excluding individuals living in institutional group quarters (e.g. prisons or nursing homes), college dormitories, military barracks, those in living situations without conventional housing (and who are not in shelters), and unrelated individuals under age 15 (e.g. foster children).

Table 1. Incidence of Poverty by Workforce Status and Location

	Worcester County	City of Worcester	Worcester County (Not City)
Total Employed	408977	82526	326451
Number of Employed in Poverty	16196	6168	10028
Percent of Employed in Poverty	4%	7%	3%
Total Unemployed	26304	6760	19544
Number of Unemployed in Poverty	6868	2347	4521
Percent of Unemployed in Poverty	26%	35%	23%

Chart 1. Percentage In Poverty by Work Status and Location



COST BURDEN

According to the US Census, an individual household is cost burdened when its occupants are spending more than 30% of their yearly income on paying for the household, regardless if they are renting or owning. This data set looks at all occupied housing units.in three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1. Percentage of Owner Households Cost Burdened at Income Level by Location			
Income Bracket	Worcester County	City of Worcester	Worcester County (Not City)
Less than $20,000	91%	92%	91%
$20,000 - $34,999	65%	73%	93%
$35,000 - $49,999	49%	56%	47%
$50,000 - $74,999	38%	34%	39%
More than $75,000	8%	8%	8%

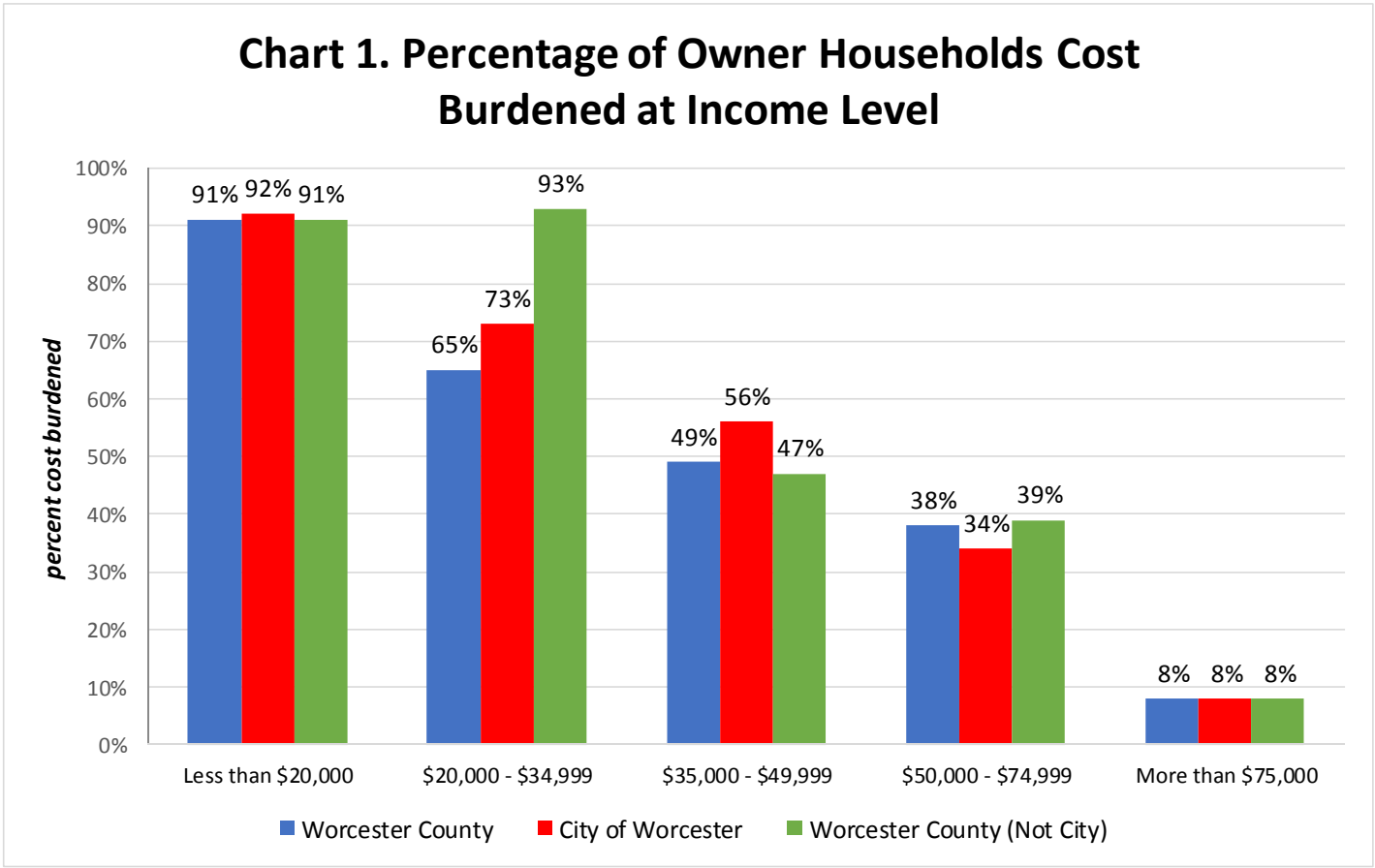
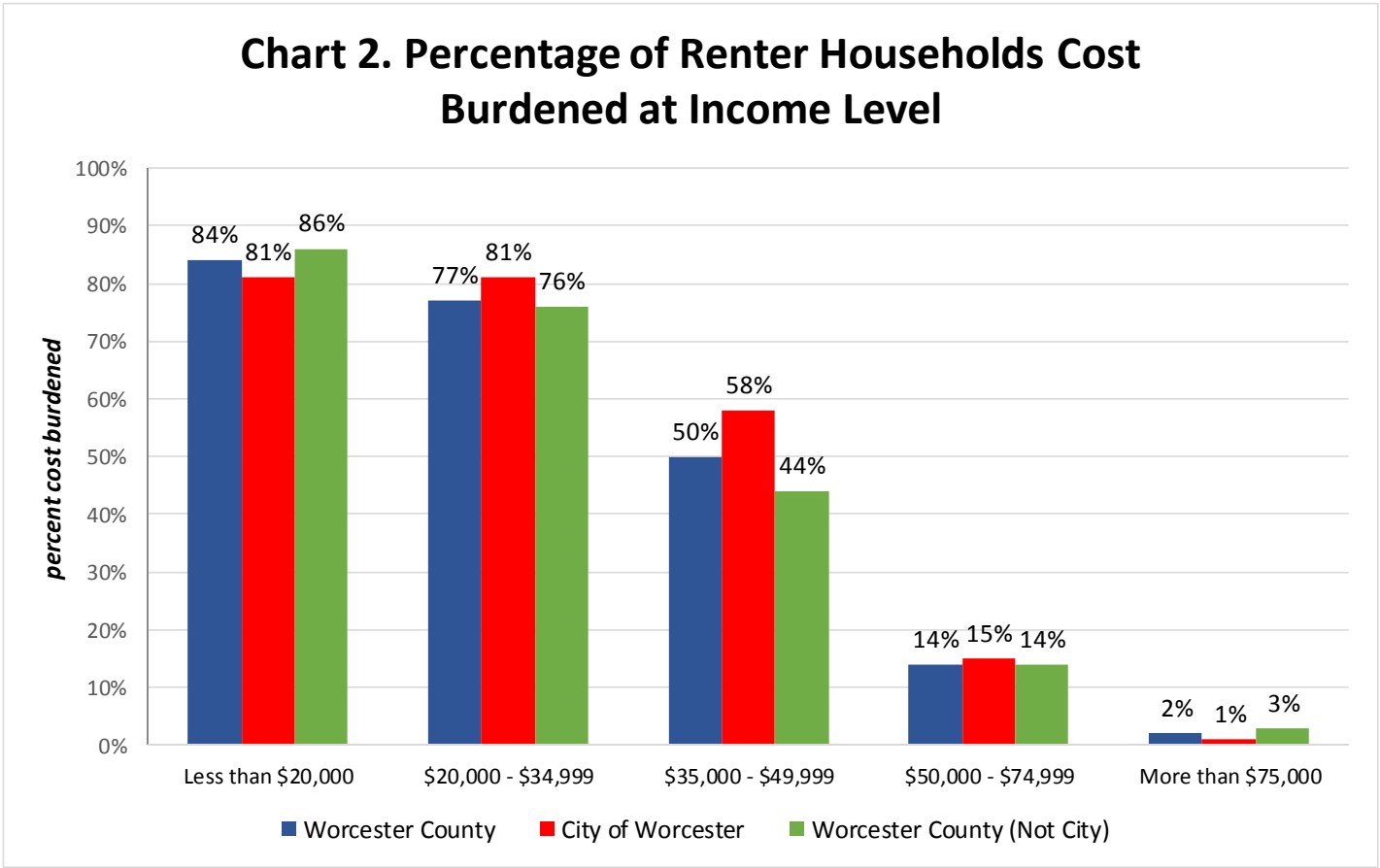


Table 2: Percentage of Renter Households Cost Burdened at Income Level by Location			
Income Bracket	Worcester County	City of Worcester	Worcester County (Not City)
Less than $20,000	84%	81%	86%
$20,000 - $34,999	77%	81%	76%
$35,000 - $49,999	50%	58%	44%
$50,000 - $74,999	14%	15%	14%
More than $75,000	2%	1%	3%



POPULATION DEMOGRAPHICS

This data set looks at the demography of the total population in three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1. Racial and Ethnic Demographics: Worcester County						
	Non-Hispanic	% of Total Population	Hispanic	% of Total Population	Total	% of Total Population
White	637093	77.9%	53823	6.6%	690916	84.4%
Black	35449	4.3%	4109	0.5%	39558	4.8%
Asian	38306	4.7%	300	0.0%	38606	4.7%
Other	18926	2.3%	30243	3.7%	49169	6.0%
Total	729774	89.2%	88475	10.8%	818249	100.0%

Table 2. Racial and Ethnic Demographics: City of Worcester						
	Non-Hispanic	% of Total Population	Hispanic	% of Total Population	Total	% of Total Population
White	105507	57.1%	22701	12.3%	128208	69.4%
Black	21799	11.8%	2637	1.4%	24436	13.2%
Asian	13466	7.3%	31	0.0%	13497	7.3%
Other	5284	2.9%	13318	7.2%	18602	10.1%
Total	146056	79.1%	38687	20.9%	184743	100.0%

Table 3. Racial and Ethnic Demographics: Worcester County (Not City)						
	Non-Hispanic	% of Total Population	Hispanic	% of Total Population	Total	% of Total Population
White	531586	83.9%	31122	4.9%	562708	88.8%
Black	13650	2.2%	1472	0.2%	15122	2.4%
Asian	24840	3.9%	269	0.0%	25109	4.0%
Other	13642	2.2%	16925	2.7%	30567	4.8%
Total	583718	92.1%	49788	7.9%	633506	100.0%

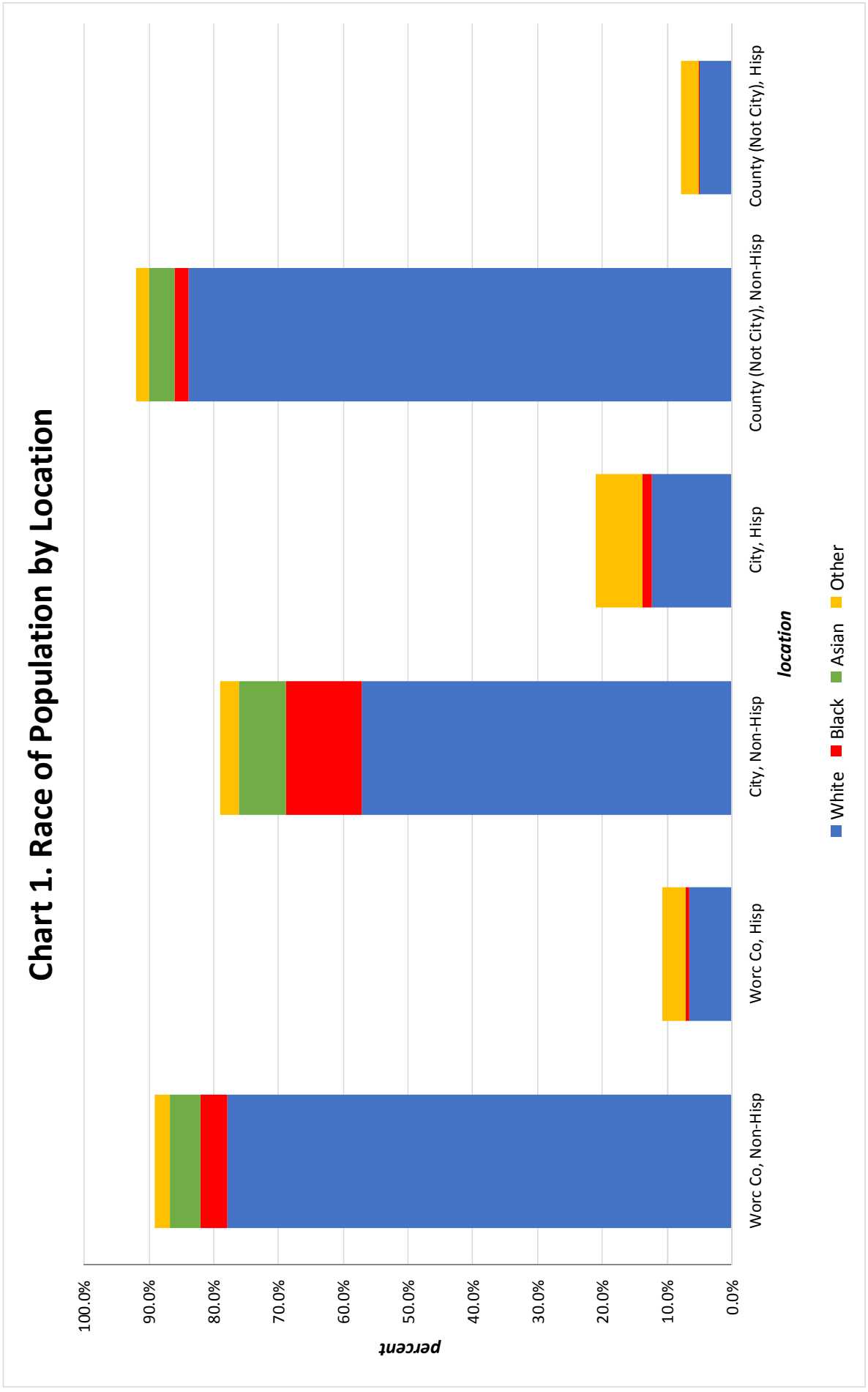


Chart 1. Race of Population by Location
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
percent
Worc Co, Non-Hisp
Worc Co, Hisp
City, Non-Hisp
City, Hisp
County (Not City), Non-Hisp
County (Not City), Hisp
location
White
Black
Asian
Other

HOUSEHOLDER DEMOGRAPHICS

This data set looks at all householders, which according to the US Census, are the people (usually one per housing unit) in whose name the housing unit is owned or rented (maintained) or, if there is no such person, any adult member, excluding roomers, boarders, or paid employees. It focuses on the populations of three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1. Racial and Ethnic Demographics: Householders in Worcester County

	Non-Hispanic	% of Total Population	Hispanic	% of Total Population	Total	% of Total Population
White	257649	85.0%	11077	3.7%	268726	88.7%
Black	9476	3.1%	1196	0.4%	10672	3.5%
Asian	9284	3.1%	54	0.0%	9338	3.1%
Other	4736	1.6%	9608	3.2%	14344	4.7%
Total	281145	92.8%	21935	7.2%	303080	100.0%

Table 2. Racial and Ethnic Demographics: Householders in the City of Worcester

	Non-Hispanic	% of Total Population	Hispanic	% of Total Population	Total	% of Total Population
White	45891	66.9%	5426	7.9%	51317	74.8%
Black	6304	9.2%	744	1.1%	7048	10.3%
Asian	3169	4.6%	31	0.1%	3200	4.7%
Other	1679	2.5%	5369	7.8%	7048	10.3%
Total	57043	83.1%	11570	16.9%	68613	100.0%

Table 3. Racial and Ethnic Demographics: Householders in Worcester County (Not City)

	Non-Hispanic	% of Total Population	Hispanic	% of Total Population	Total	% of Total Population
White	211758	91.3%	5651	2.4%	217409	92.7%
Black	3172	1.4%	452	0.2%	3624	1.6%
Asian	6115	2.6%	23	0.0%	6138	2.6%
Other	3057	1.3%	4239	1.8%	7296	3.1%
Total	224102	95.6%	10365	4.4%	234467	100.0%

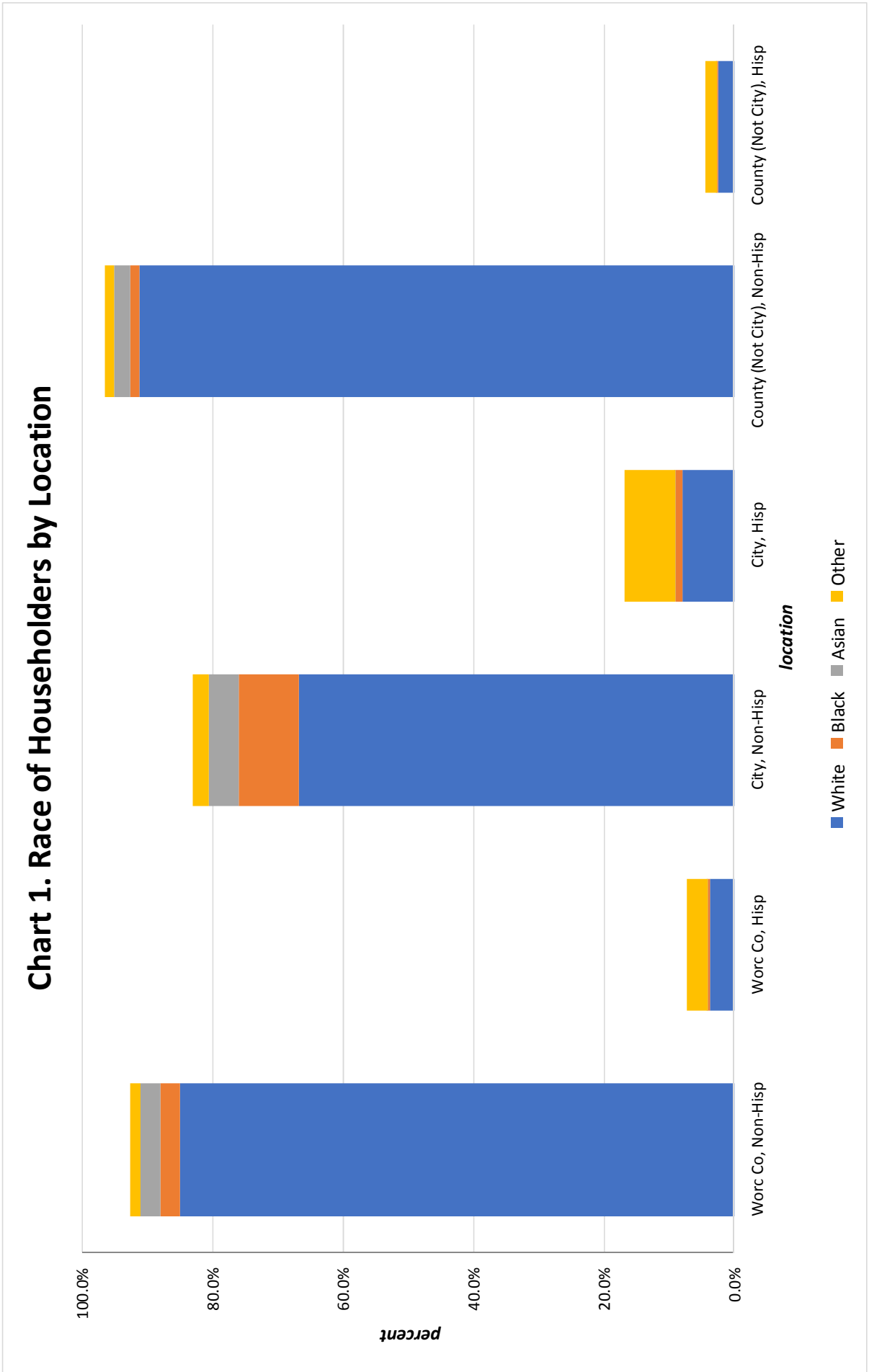


Chart 1. Race of Householders by Location
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
percent
Worc Co, Non-Hisp
Worc Co, Hisp
City, Non-Hisp
City, Hisp
County (Not City), Non-Hisp
County (Not City), Hisp
location
White
Black
Asian
Other

INCOME BY RACE AND ETHNICITY

This data set looks at income and race among the total population in three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1. Household Income by Race: Worcester County								
Income Level	*# White House-holds*	*% White House-holds*	*# Black House-holds*	*% Black House-holds*	*# Asian House-holds*	*% Asian House-holds*	*# Other Race House-holds*	*% Other Race House-holds*
Less than $10,000	13193	5%	1314	10%	690	6%	1369	10%
$10,000 - $14,999	12768	5%	607	5%	273	2%	1165	9%
$15,000 - $19,999	11361	4%	638	5%	193	2%	786	6%
$20,000 - $24,999	10844	4%	461	4%	497	4%	649	5%
$25,000 - $29,999	9757	4%	748	6%	159	1%	1008	7%
$30,000 - $34,999	11130	4%	1011	8%	321	3%	904	7%
$35,000 - $39,999	9691	4%	586	5%	215	2%	646	5%
$40,000 - $44,999	9078	3%	951	8%	279	2%	615	5%
$45,000 - $49,999	8453	3%	346	3%	322	3%	567	4%
$50,000 - $59,999	17661	7%	1187	9%	627	6%	1124	8%
$60,000 - $74,999	25252	9%	1281	10%	1082	10%	1097	8%
$75,000 - $99,999	35247	13%	1424	11%	1446	13%	1235	9%
$100,000 - $124,999	29765	11%	895	7%	1018	9%	1113	8%
$125,000 - $149,999	20447	8%	365	3%	880	8%	372	3%
$150,000 - $199,999	22776	8%	270	2%	1296	12%	438	3%
$200,000 or more	21266	8%	445	4%	1940	17%	464	3%
Total	268689	100%	12529	100%	11238	100%	13519	100%

Table 2. Household Income by Race: City of Worcester								
Income Level	*# White House-holds*	*% White House-holds*	*# Black House-holds*	*% Black House-holds*	*# Asian House-holds*	*% Asian House-holds*	*# Other Race House-holds*	*% Other Race House-holds*
Less than $10,000	5196	10%	1118	13%	467	12%	888	15%
$10,000 - $14,999	4342	8%	419	5%	128	3%	658	11%
$15,000 - $19,999	3465	7%	385	5%	86	2%	373	6%
$20,000 - $24,999	3092	6%	350	4%	178	5%	307	5%
$25,000 - $29,999	2245	4%	509	6%	134	3%	306	5%
$30,000 - $34,999	2521	5%	845	10%	162	4%	364	6%
$35,000 - $39,999	2374	5%	485	6%	105	3%	331	6%
$40,000 - $44,999	1902	4%	718	8%	190	5%	259	4%
$45,000 - $49,999	1900	4%	199	2%	223	6%	236	4%
$50,000 - $59,999	3554	7%	890	11%	277	7%	329	6%
$60,000 - $74,999	4873	9%	759	9%	441	11%	509	9%
$75,000 - $99,999	5709	11%	824	10%	605	16%	341	6%
$100,000 - $124,999	4112	8%	426	5%	316	8%	553	9%
$125,000 - $149,999	2813	5%	214	3%	128	3%	125	2%
$150,000 - $199,999	2500	5%	132	2%	208	5%	166	3%
$200,000 or more	1966	4%	201	2%	228	6%	132	2%
Total	52564	100%	8474	100%	3876	100%	5878	100%

Table 3. Household Income by Race: Worcester County (Not City)								
Income Level	*# White House-holds*	*% White House-holds*	*# Black House-holds*	*% Black House-holds*	*# Asian House-holds*	*% Asian House-holds*	*# Other Race House-holds*	*% Other Race House-holds*
Less than $10,000	7997	4%	196	5%	223	3%	481	6%
$10,000 - $14,999	8426	4%	188	5%	145	2%	507	7%
$15,000 - $19,999	7896	4%	253	6%	107	1%	413	5%
$20,000 - $24,999	7752	4%	111	3%	319	4%	342	4%
$25,000 - $29,999	7512	3%	239	6%	25	0%	702	9%
$30,000 - $34,999	8609	4%	166	4%	159	2%	540	7%
$35,000 - $39,999	7317	3%	101	2%	110	1%	315	4%
$40,000 - $44,999	7176	3%	233	6%	89	1%	356	5%
$45,000 - $49,999	6553	3%	147	4%	99	1%	331	4%
$50,000 - $59,999	14107	7%	297	7%	350	5%	795	10%
$60,000 - $74,999	20379	9%	522	13%	641	9%	588	8%
$75,000 - $99,999	29538	14%	600	15%	841	11%	894	12%
$100,000 - $124,999	25653	12%	469	12%	702	10%	560	7%
$125,000 - $149,999	17634	8%	151	4%	752	10%	247	3%
$150,000 - $199,999	20276	9%	138	3%	1088	15%	272	4%
$200,000 or more	19300	9%	244	6%	1712	23%	332	4%
Total	216125	100%	4055	100%	7362	100%	7641	100%

Table 4. Household Income by Race: Worcester County				
Income Level	*# Hispanic Households*	*% Hispanic Households*	*# Non-Hispanic Households*	*% Non-Hispanic Households*
Less than $10,000	3527	13%	10921	4%
$10,000 - $14,999	2989	11%	10856	4%
$15,000 - $19,999	1638	6%	10352	4%
$20,000 - $24,999	1671	6%	9729	4%
$25,000 - $29,999	1891	7%	8880	4%
$30,000 - $34,999	1614	6%	10266	4%
$35,000 - $39,999	1029	4%	9191	4%
$40,000 - $44,999	1080	4%	8581	3%
$45,000 - $49,999	866	3%	7998	3%
$50,000 - $59,999	1841	7%	16632	7%
$60,000 - $74,999	2302	9%	23843	9%
$75,000 - $99,999	2281	9%	33712	13%
$100,000 - $124,999	1845	7%	28512	11%
$125,000 - $149,999	778	3%	19855	8%
$150,000 - $199,999	590	2%	22400	9%
$200,000 or more	596	2%	20896	8%
Total	26538	100%	252624	100%

Table 5. Household Income by Race: City of Worcester				
Income Level	*# Hispanic Households*	*% Hispanic Households*	*# Non-Hispanic Households*	*% Non-Hispanic House-holds*
Less than $10,000	2392	18%	3711	8%
$10,000 - $14,999	1710	13%	3235	7%
$15,000 - $19,999	882	7%	2888	6%
$20,000 - $24,999	1129	9%	2281	5%
$25,000 - $29,999	767	6%	1838	4%
$30,000 - $34,999	756	6%	2109	5%
$35,000 - $39,999	572	4%	2138	5%
$40,000 - $44,999	449	3%	1726	4%
$45,000 - $49,999	309	2%	1754	4%
$50,000 - $59,999	825	6%	3033	7%
$60,000 - $74,999	876	7%	4438	10%
$75,000 - $99,999	895	7%	5050	11%
$100,000 - $124,999	756	6%	3736	8%
$125,000 - $149,999	238	2%	2641	6%
$150,000 - $199,999	172	1%	2455	5%
$200,000 or more	217	2%	1838	4%
Total	12945	100%	44871	100%

Table 6. Household Income by Race: Worcester County (Not City)				
Income Level	*# Hispanic Households*	*% Hispanic Households*	*# Non-Hispanic Households*	*% Non-Hispanic Households*
Less than $10,000	1135	8%	7210	3%
$10,000 - $14,999	1279	9%	7621	4%
$15,000 - $19,999	756	6%	7464	4%
$20,000 - $24,999	542	4%	7448	4%
$25,000 - $29,999	1124	8%	7042	3%
$30,000 - $34,999	858	6%	8157	4%
$35,000 - $39,999	457	3%	7053	3%
$40,000 - $44,999	631	5%	6855	3%
$45,000 - $49,999	557	4%	6244	3%
$50,000 - $59,999	1016	7%	13599	7%
$60,000 - $74,999	1426	10%	19405	9%
$75,000 - $99,999	1386	10%	28662	14%
$100,000 - $124,999	1089	8%	24776	12%
$125,000 - $149,999	540	4%	17214	8%
$150,000 - $199,999	418	3%	19945	10%
$200,000 or more	379	3%	19058	9%
Total	13593	100	207753	100%

Section Data Notes

Median incomes on tables with County and City data come from Table S1903: ***2013 - 2017 American Community Survey 5 – Year Estimates of the median income in a 12-month period (in 2017 inflation-adjusted dollars)***

Median incomes for tables with County (Not City) data were calculated by estimating which income level 50% of the households fall under and then finding the location within the range in which exactly 50% have at most that income.

Median income for those classified under "other" for race was found by taking a weighted average using the percent of each subsection of race under the "other" category and the median income of that race. The non-Hispanic median income is that of non-Hispanic, white individuals. Also note that those classified as "White", "Black", "Asian", or "Other" in this data set can be either Hispanic or non-Hispanic.

Chart 1. Household Income by Race, Worcester County

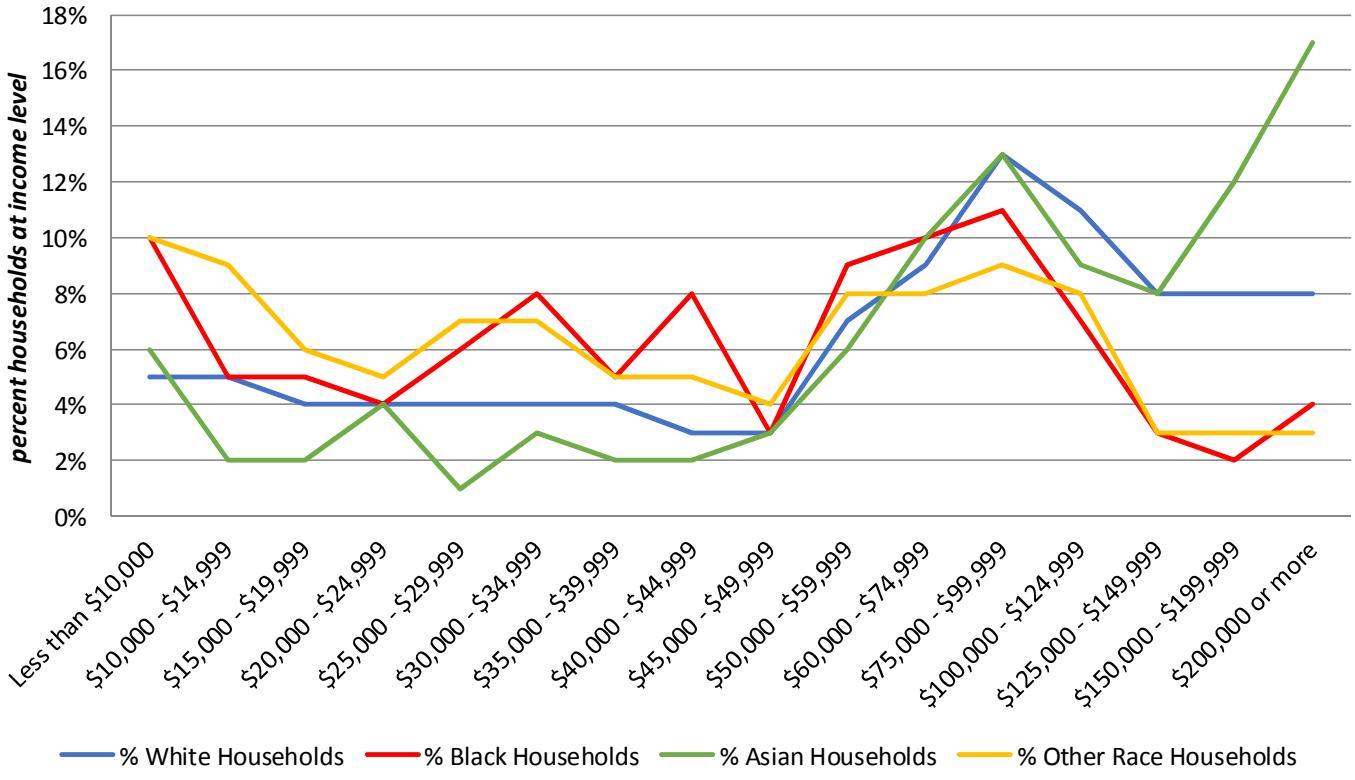


Chart 2. Household Income by Race, City of Worcester

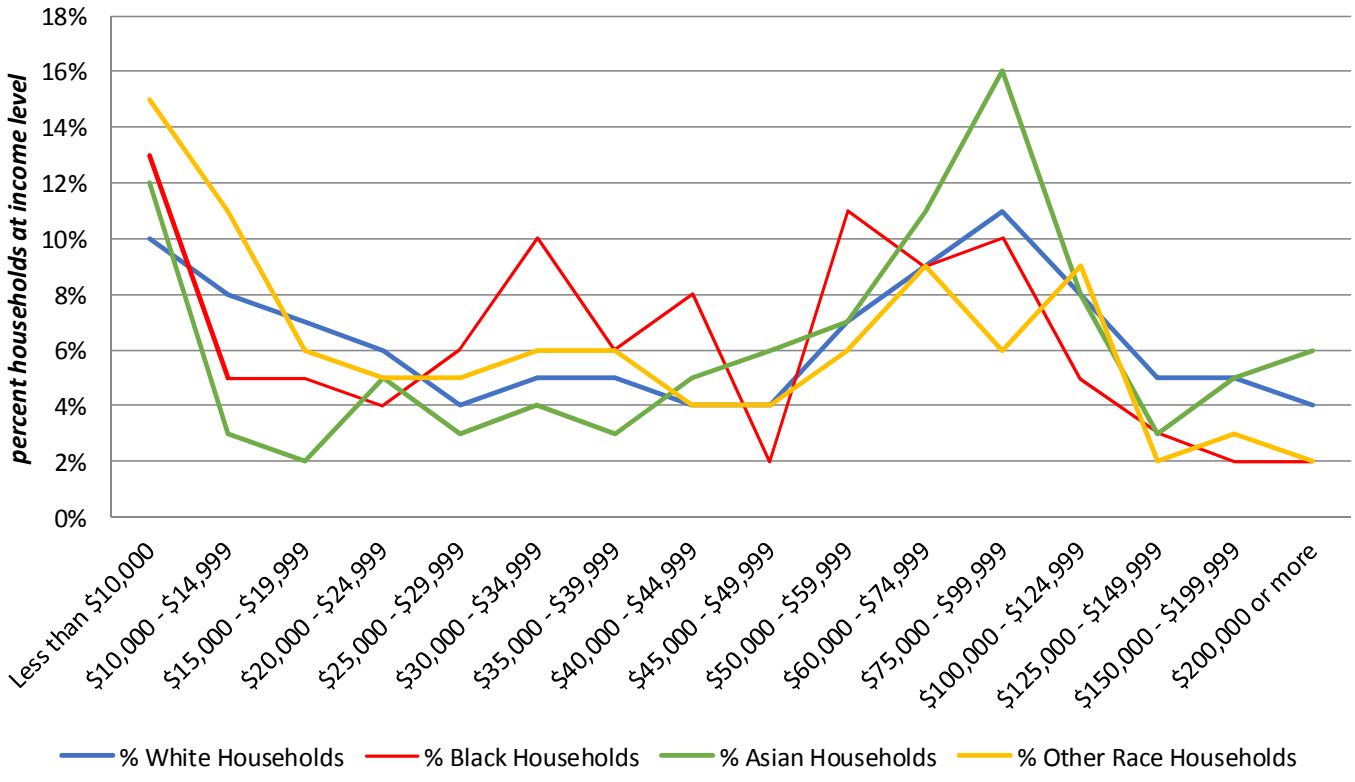


Chart 3. Household Income by Race, County (Not City)

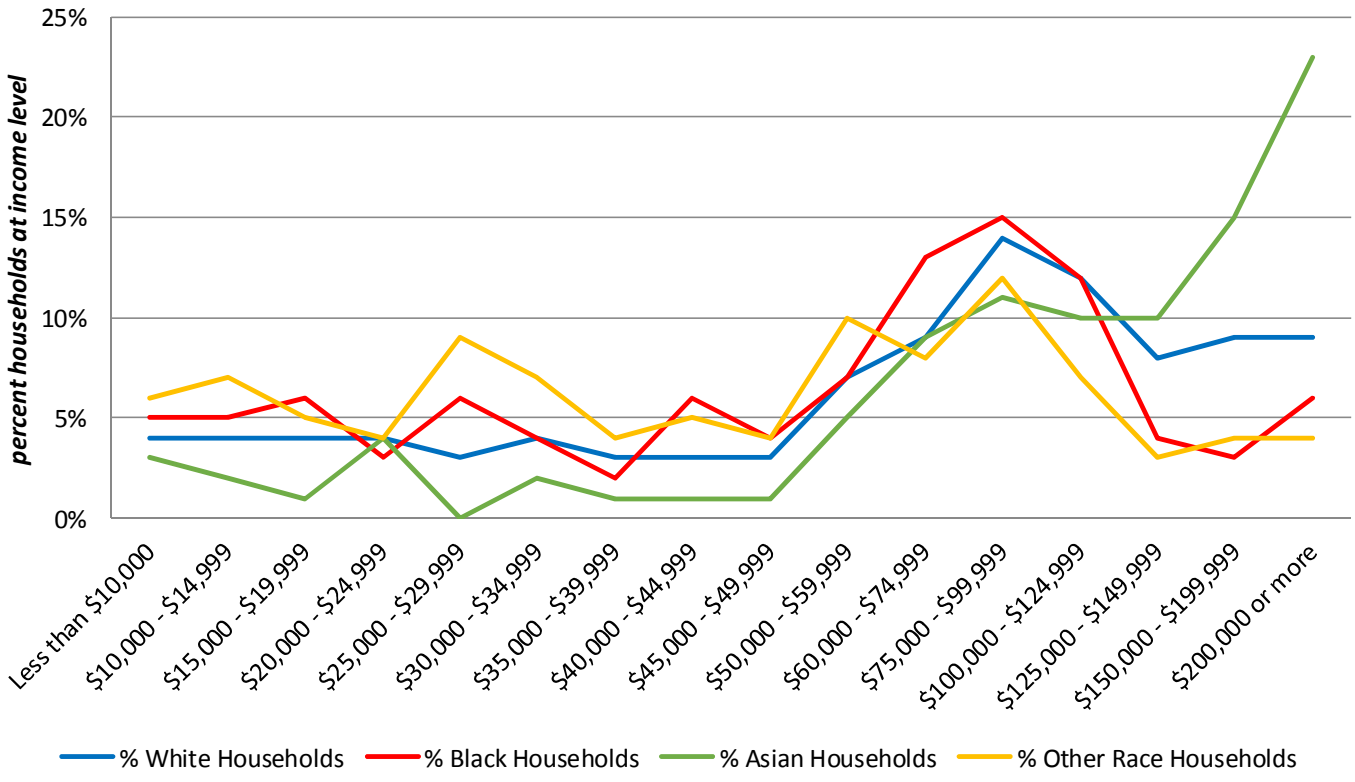
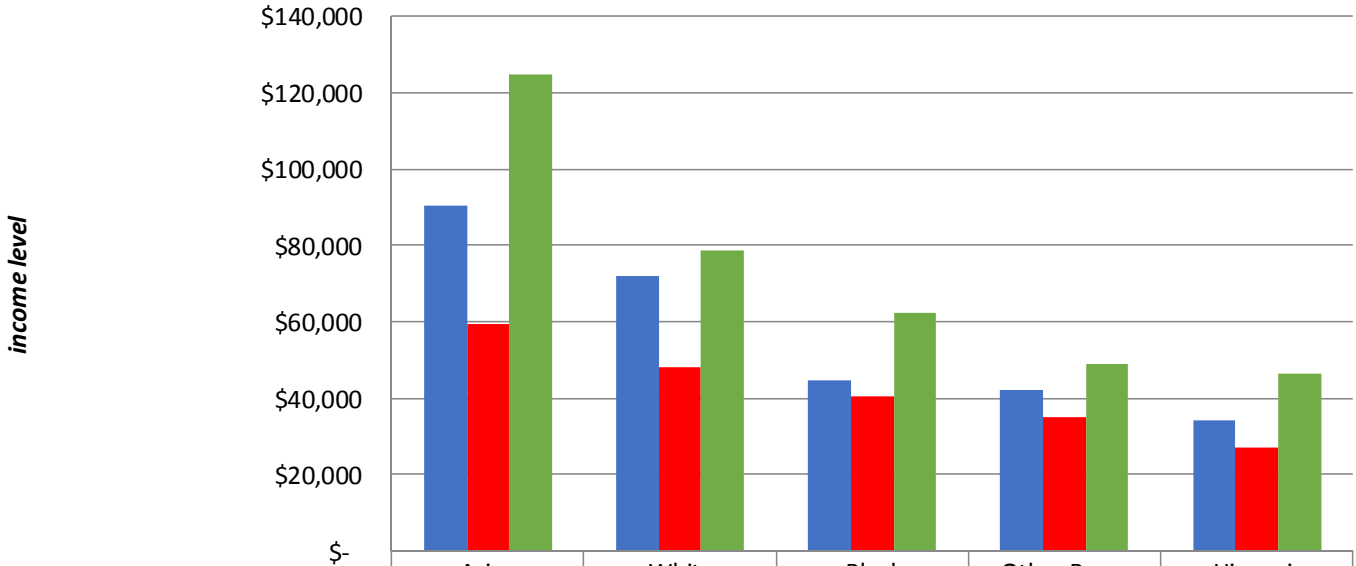


Chart 4. Household Median Income by Race and Location



	Asian	White	Black	Other Races	Hispanic
County Householders	$90,349	$72,095	$44,672	$42,302	$34,373
City Householders	$59,302	$47,903	$40,648	$35,269	$27,057
County (Not City) Householders	$125,000	$78,571	$62,308	$48,750	$46,250

Chart 5. Household Income by Ethnicity, Worcester County

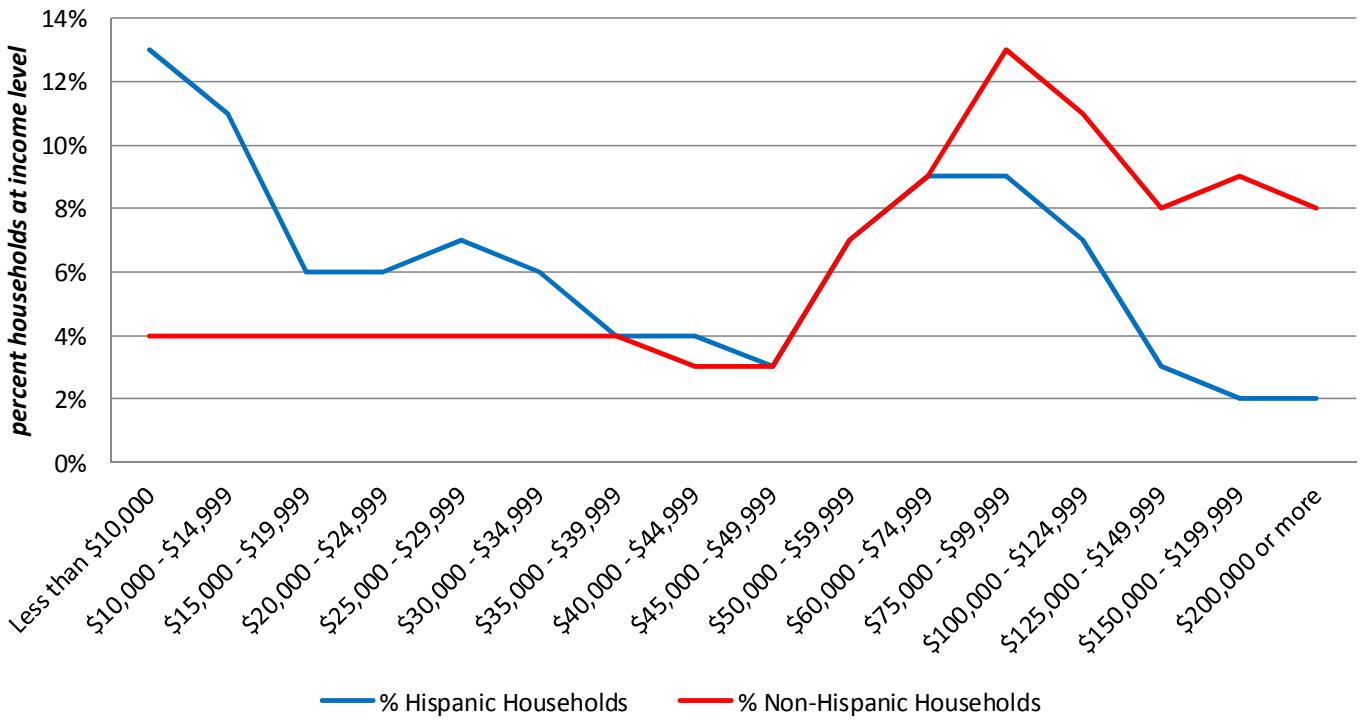


Chart 6. Household Income by Ethnicity, City of Worcester

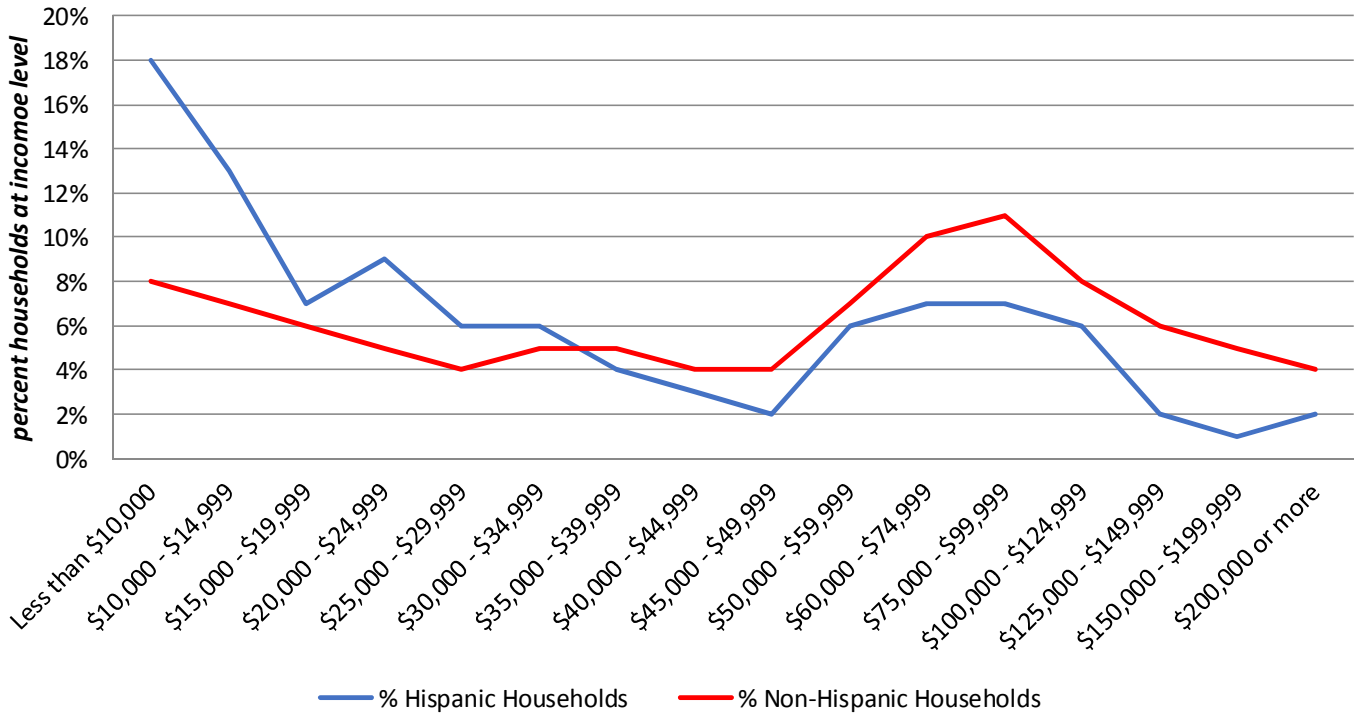
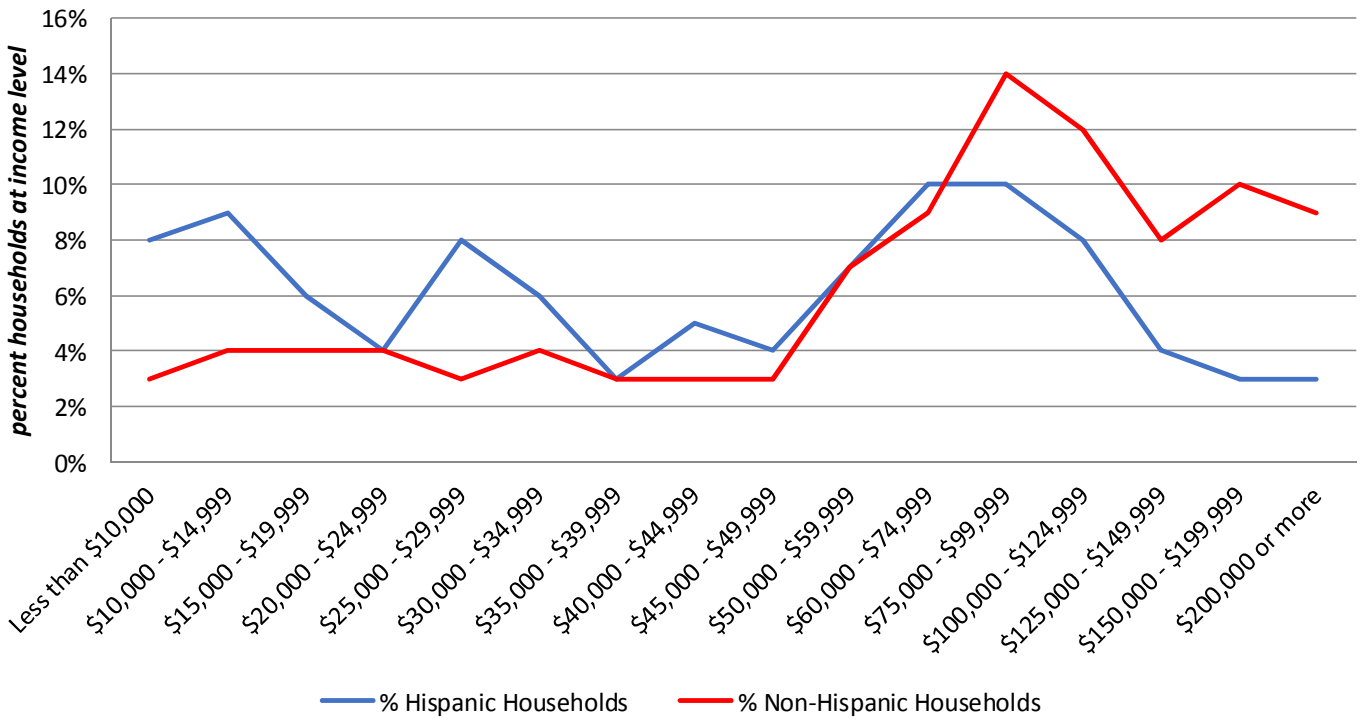


Chart 7. Household Income by Ethnicity, Worcester County (Not City)



TENURE BY RACE AND ETHNICITY

This data set looks at all occupied housing units by race and ownership status among in the three locations: Worcester County, the City of Worcester, and Worcester County (Not City).

Table 1. Racial and Ethnic Demographics: Owners in Worcester County						
Race, Head of Household	Non-Hispanic	% of County Population	Hispanic	% of County Population	Total	% of County Population
White	184100	91.1%	3404	1.7%	187504	93.6%
Black	3288	1.6%	259	0.1%	3547	1.8%
Asian	5306	2.7%	18	0.0%	5324	2.7%
Other	1995	1.0%	1952	1.0%	3947	2.0%
Total	194689	97.19%	5633	2.81%	200322	100%

Table 2. Racial and Ethnic Demographics: Renters in Worcester County						
Race, Head of Household	Non-Hispanic	% of County Population	Hispanic	% of County Population	Total	% of County Population
White	73549	71.6%	7673	7.5%	81222	79.0%
Black	6188	6.0%	937	0.9%	7125	6.9%
Asian	3978	3.9%	36	0.0%	4014	3.9%
Other	2741	2.7%	7656	7.5%	10397	5.2%
Total	86456	84.1%	16302	15.9%	102758	100%

Table 3. Racial and Ethnic Demographics: Owners in the City of Worcester						
Race, Head of Household	Non-Hispanic	% of City Population	Hispanic	% of City Population	Total	% of City Population
White	24891	81.5%	1105	3.6%	25996	85.2%
Black	1858	6.1%	120	0.4%	1978	6.5%
Asian	1268	4.2%	9	0.0%	1277	4.2%
Other	447	1.5%	832	2.7%	1279	4.2%
Total	28464	93.2%	2066	6.8%	30530	100%

Table 4. Racial and Ethnic Demographics: Renters in the City of Worcester						
Race, Head of Household	Non-Hispanic	% of City Population	Hispanic	% of City Population	Total	% of City Population
White	21000	55.1%	4321	11.4%	25321	66.5%
Black	4446	11.7%	624	1.6%	5070	13.3%
Asian	1901	5.0%	22	0.1%	1923	5.1%
Other	1232	3.2%	4537	11.9%	5769	18.9%
Total	28579	74.0%	9504	25.0%	38083	100%

Table 5. Racial and Ethnic Demographics: Owners in Worcester County (Not City)						
Race, Head of Household	Non-Hispanic	% of County (Not City) Population	Hispanic	% of County (Not City) Population	Total	% of County (Not City) Population
White	159209	93.8%	2299	1.4%	161508	95.1%
Black	1430	0.8%	139	0.1%	1569	0.9%
Asian	4038	2.4%	9	0.0%	4047	2.4%
Other	1548	0.9%	1120	0.7%	2668	1.6%
Total	166225	97.9%	3567	2.1%	169792	100%

Table 6. Racial and Ethnic Demographics: Renters in Worcester County (Not City)						
Race, Head of Household	Non-Hispanic	% of County (Not City) Population	Hispanic	% of County (Not City) Population	Total	% of County (Not City) Population
White	52549	81.3%	3352	5.2%	55901	86.4%
Black	1742	2.7%	313	0.5%	2055	3.2%
Asian	2077	3.2%	14	0.0%	2091	3.2%
Other	1509	2.3%	3119	4.8%	4628	2.7%
Total	57877	89.5%	6798	10.5%	64675	100%

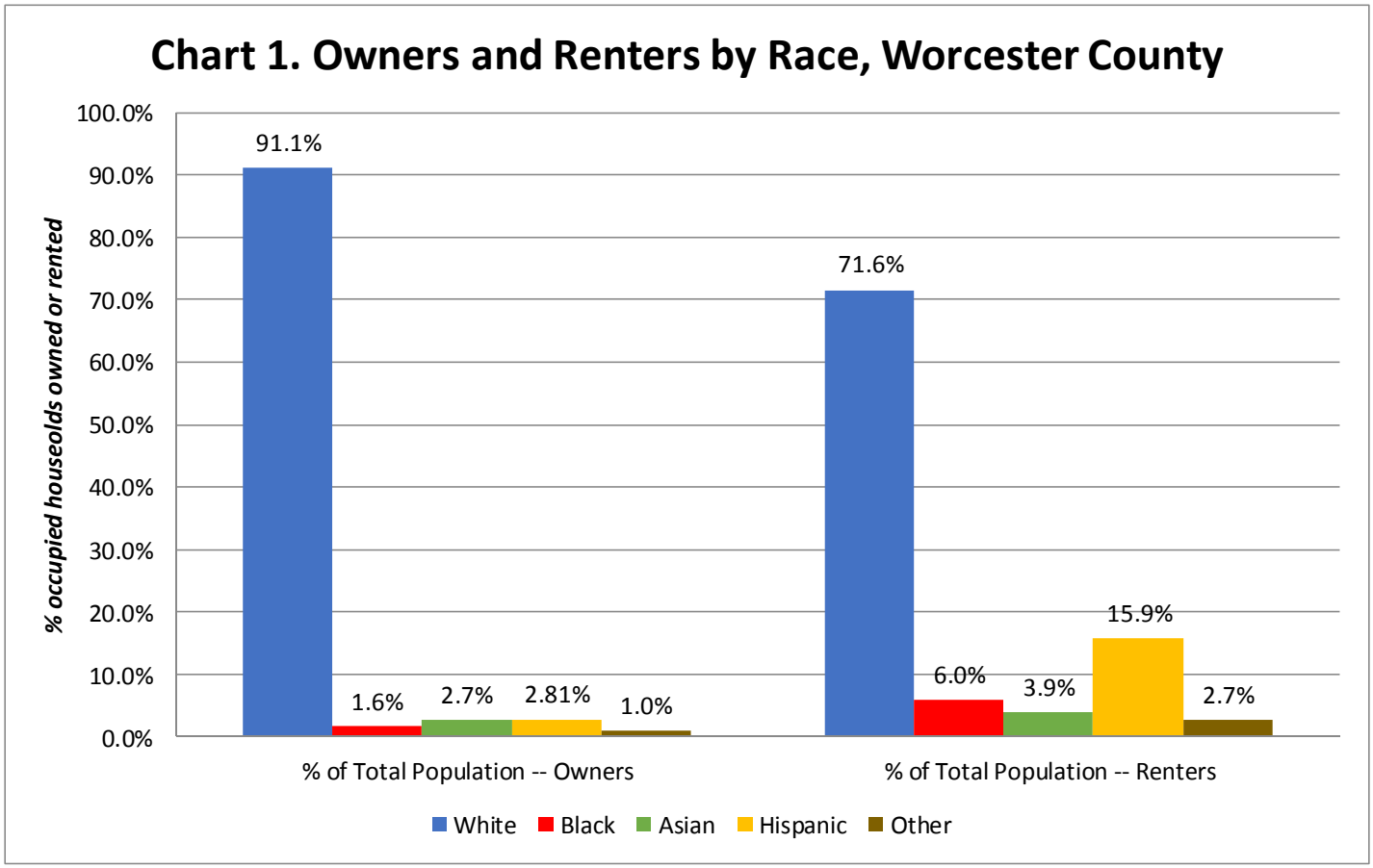


Chart 2. Owners and Renters by Race, City of Worcester

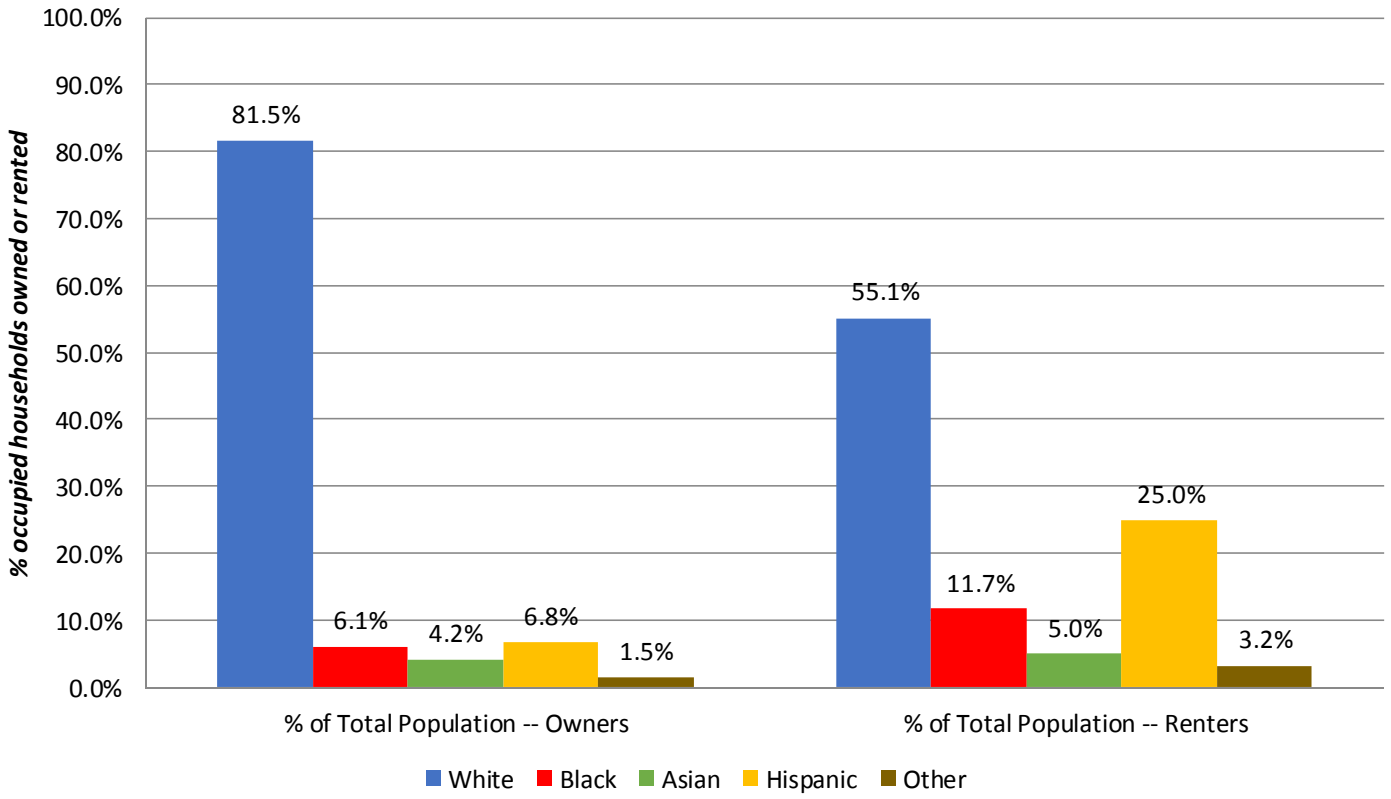
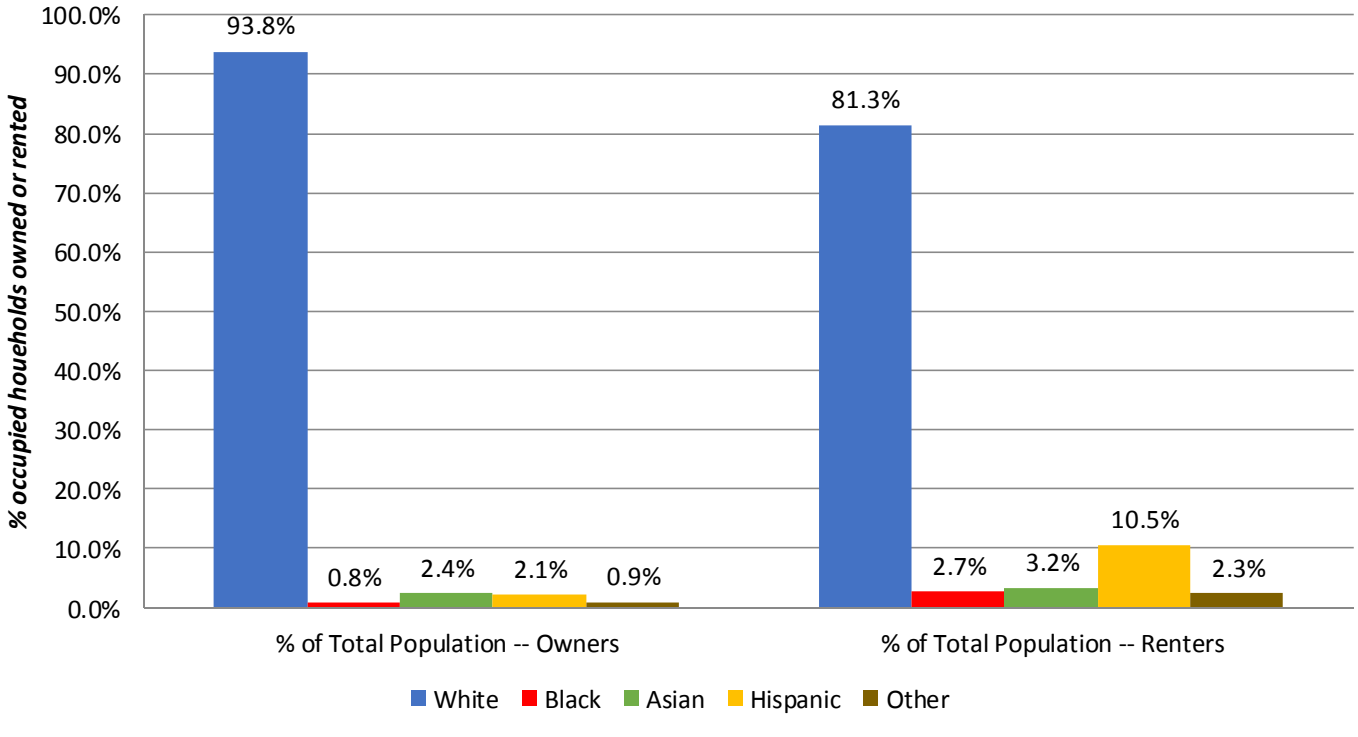


Chart 3. Owners and Renters by Race, Worcester County (Not City)

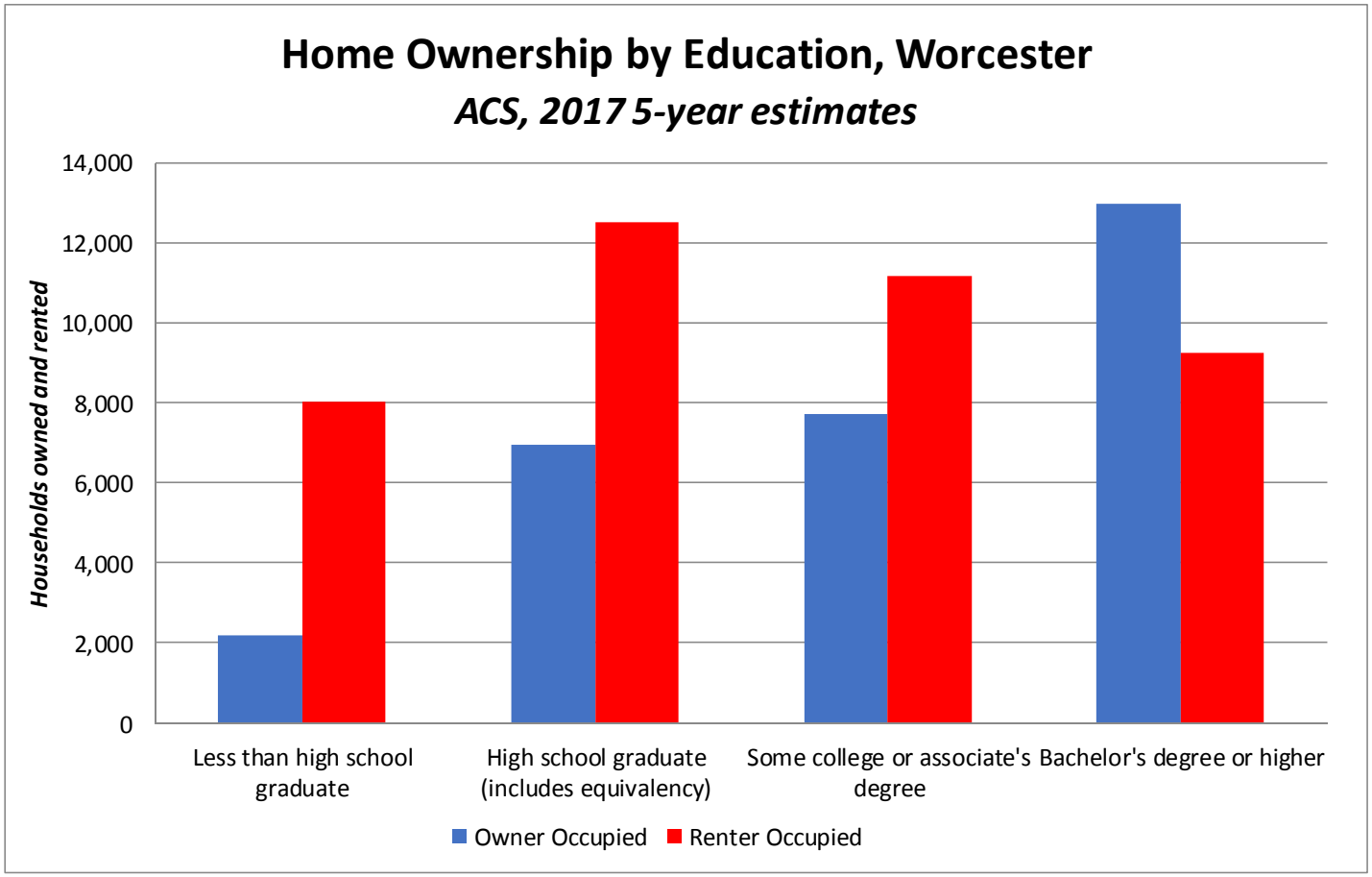


POTENTIAL FUTURE WORK

All of the data included in this report comes from either the 2017 American Community Survey 5-year estimates or the 2010 U.S. Census. This report only uses data from Worcester County and the City of Worcester. It focused on race and ethnicity of different populations, whether it be householders or the entire population in the three locations — Worcester County, the City of Worcester, and Worcester County outside the city. There are numerous other variables on which to focus for future analyses and research. Future statistical work is sorely needed on how the City of Worcester's affordable housing decisions impact the elderly, disabled, veterans, and ethno-racial subgroups. Moreover, qualitative work on these communities that explore how they make ends meet, what "cost burdened" means at a human-scale level in Worcester, and different ways to look at housing that understand the actual needs and lives of the city's residents; GIS work can model where in the city the housing is and where it could be.

Another area of focus that proved beyond the scope of this report because the data is unnecessarily difficult to obtain has to do with the real numbers of affordable housing and what might be called "actually affordable housing." Indeed, because the city meets a numerical threshold for affordable housing does not necessarily mean that housing units so classified are truly affordable. While this is an affordable housing problem, it is also an ethical dilemma for inclusive city, and a Civil Rights issue, one that is simultaneously related to education (e.g. see below).

Finally, what follows in this report is a deep statistical dive exploring one of the most important statistics in affordable housing analysis: percent rent burdened. The model tests what variables most affect being cost burdened, and while the findings may be intuitive, they nevertheless call attention to the necessity for comprehensive, data-driven urban planning and policy making that allows more Worcester residents to share in economic prosperity that will move the city forward.




COMMON

Modeling Percent Rent Burdened of Census Block Groups by Median Income and Percent in Poverty

By Joshua Oliver, Lead Researcher

Abstract

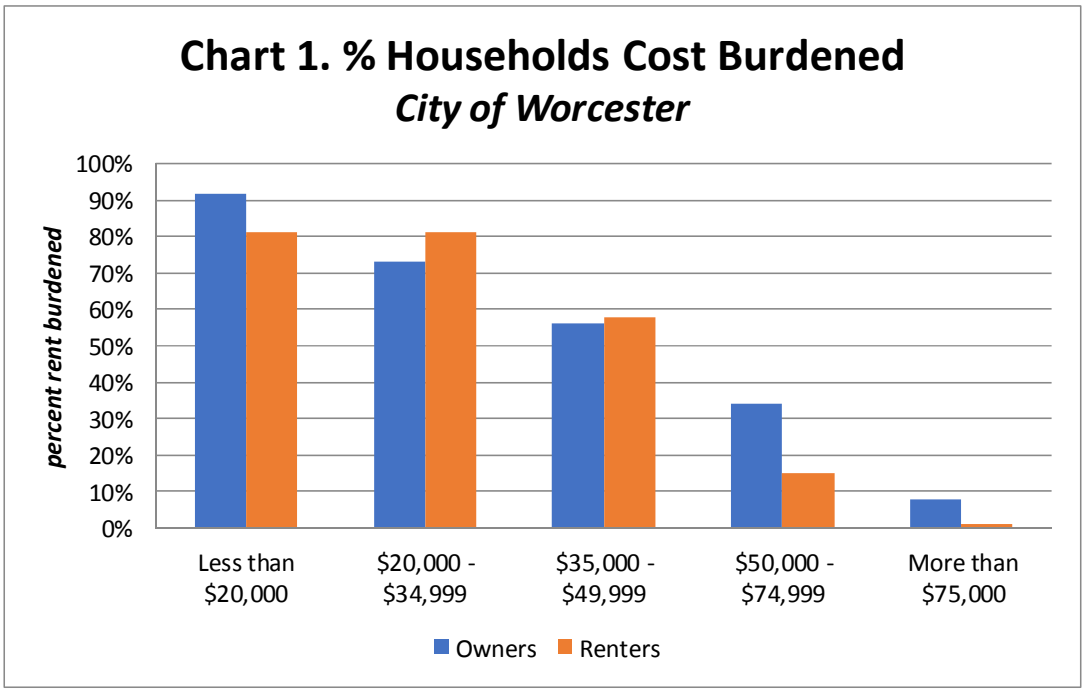
Analyzing data on the City of Worcester and Worcester County (Outside the City) from the American Community Survey shows there are many variables between the two different geographies. This statistical inquiry models the variable that is most important to the overall goal of studying affordable housing: percent rent burden. After forming a table containing variables — percent black, percent Hispanic, percent employed, median income, percent rent burden, percent in poverty, and a variable indicating whether a particular census block is in the city — this inquiry relied on statistical software to form a best-fit, multi-variable linear model that predicts the percent of rent burdened individuals for a particular census block group. The model showed median income and percent in poverty were the most statistically significant variables, but the variables themselves are affected extensively by various socio-economic situations that should be explored more.

Introduction

Massachusetts General Laws, Chap 40B, was created in 1969 to encourage the state's cities and towns to designate at least 10% of their housing stock as affordable, a designation connected to median household income and housing cost burden. The underpinning idea is to allow individuals or families to spend less than 30% of their total income on rent or housing costs.

Affordable housing, though, is a much larger issue. According to the US Department of Housing & Urban Development (HUD), nearly half of all renters in the country are cost burdened. There are national programs to alleviate the burden, such as the federal government's voucher program that provides cost burdened householders some relief. Massachusetts also has programs including Rental Assistance for Housing Rehabilitation, Project-based Section 8 Rental Assistance, Veterans Affairs Supported Housing, Shelter Plus Care, and Family Self Sufficiency initiatives. Yet, these programs nevertheless leave gaps as cost burden problems persist and, importantly, are not associated solely with very low-income families.

The City of Worcester, as it turns out, contains 23.1% of all housing units in county, and its cost burden rates are both high and wide. (See *Chart 1.*) In fact, this research came about less because officials are doing nothing toward closing gaps and more because clear needs are not being met, especially among Worcester's high cost burdened populations. This got researchers thinking about Worcester's affordable housing needs enough to begin exploring the data in New England's second largest city.



Methodology

I first considered what variables would be important in modeling the percentage of rent burdened households within census block groups throughout Worcester County. After determining the variables listed above (see Abstract), I searched through thousands of US Census tables to obtain information on those variables for every census block group in the county. If the desired data percentages were not calculated within the tables, I used RStudio, a data analytical software program, to create them when forming my data structure. Each variable had its own data table which included the census tract number and block group number as shown below.

Exhibit 1 Census Codes

Census Tract	Block Group
700100	7001001
700100	7001002
700100	7001003
700100	7001004
700100	7001005
701100	7011001
701100	7011002

I created a variable that determined whether the census block group was located in the City of Worcester using a for loop in R with the following code to assign a 1 to whether a census block group was in the city and a 0 if it was not. Worcester1$`Census Tract` is a vector containing all census tracts located in the City of Worcester.

Exhibit 2. For Loop

```
for (i in 1:length(Worcester1$`Census Tract`)) {
  if (i == 1) {
    p <- Worcester1$`Census Tract`[i]
    Worcester2 <- as.numeric(Model$`Census Tract` == p)
  }
  else{
    p <- Worcester1$`Census Tract`[i]
    Worcester2 <- Worcester2 + as.numeric(Model$`Census Tract` == p)
  }
}
```

All individual variable tables were merged into a single table. Once the data was organized this way, I used different variable selection techniques to determine which variable or combination of variables best modeled the percentage of rent burdened individuals.

I began data modeling with forward stepwise regression in which 6 single-predictor regression models were formed. By looking at a summary of important information on each model, I chose the single-variable model that best fit the response variable based on which model had the highest R^2 value. R^2, called the coefficient of determination, is the percentage of variation in the response variable that is explained by the regression line. The higher the R^2, the more variation is explained by the model, meaning that the model is a better fit than a model with a lower R^2. After finding which single-variable model fit best, I used that variable to create 5, two-predictor variable models. Using the same steps to choose a best model, I used those two-variables to create 4, three-predictor variable models and so on until I had a full model containing all six explanatory variables predicting my response variable.

When I had my six best models, I used RStudio to compare the models using the **anova( )** function in R to see whether adding another variable was worth the change in the R^2 statistic. (Having a 6 variable model might not be worth the effort when the R^2 value only increases by .000001 between the 6-variable model and the 5-variable model.) After conducting this analysis, I determined which model would be my final model by looking at the p-value produced with the **anova( )** function. If my p-value was greater than .05, I would decide that I should not include the new variable in my model. Once my model was selected, I wanted to use another variable selection technique to see if it came out with the same model.

The **step( )** function in R uses the AIC statistic to calculate the best model. Akaike's Information Criterion (AIC), like R^2, is used to compare different regression models. It does not compare how well the model fits but only tells which model among those compared fits best. Another difference is the lower the AIC, the better the model. After comparing my results from both variable selection techniques, I determined my final model.

From there, I had to make sure that my model met assumptions used for linear regression models. Linear regression models typically have the form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_{p-2} x_{p-2,i} + \beta_{p-1} x_{p-1,i} + \epsilon_i \text{ for } i = 1,2,\ldots,n \text{ where } \epsilon_i \sim N(\mu, \sigma^2)$$

The following are assumptions needing to be met by linear regression models:

- The model parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ and σ are constant;
- Each term in the model is additive;
- The error terms in the regression model are independent and have been sampled from a single population; they also follow a normal probability distribution centered at zero with a fixed variance, σ^2.

Regression assumptions about error terms are checked by looking at residuals from the data. Residuals are found by subtracting the estimated value of the response variable by the model from the actual value in the data.

I checked for heteroscedasticity, which means that the variance of the error term is not constant for all levels of the explanatory variables. To do this, I created graphs of my two explanatory variables vs residuals and then a graph of the fitted values vs the residuals. If heteroskedasticity is shown, transformations are needed to be performed on my variables to reduce heteroskedasticity. After testing different transformations on each of my three variables (two explanatory and one response) and creating graphs looking at my transformed variables vs residuals, I settled on my final transformations. If homoskedasticity (equal variance) was shown, I would conclude that my model assumptions are being met.

I checked for autocorrelation which exists when consecutive error terms are related, meaning the error terms would not be independent. To do this, I made a graph of order vs. residuals. If a distinct pattern is shown, autocorrelation exists and model assumptions are violated.

After all of this, I noticed outliers in my graphs. An outlier is any data value that does not seem to fit the general pattern of the data set. For this study, these were when the percentage of rent burdened households was 0. I removed these points from my data set and created a new linear model using the same variables but using data without the outliers. I then compared the coefficients of my model with the outliers included in the data to those of my model without the outliers included in the data, trying to determine if my outliers were influential. If outliers are determined to be influential, work should be done to verify their accuracy.

Lastly, I tried to verify the model assumption that my residuals were normally distributed. To do so, I created a histogram and a normal probability plot of my residuals. One can assume approximate normality of residuals by determining whether the histogram follows the shape of a bell curve and if the points in the normal probability plot lie relatively close to the line.

It is often beneficial to create new variables that are functions of the existing explanatory variables in a model which are known as interaction terms. I wanted to try including an interaction term by multiplying median income by percent in poverty. Before I came to the conclusion to create the variable, I wanted to see if it was necessary. I created a graph of median income vs. percent rent burdened and noticed that there was a difference between the location of points with a high percentage of poverty and a low percentage of poverty. Thus, I felt the need for the interaction term. I created a model which included the interaction term as a third explanatory variable and then compared it to my model with my two explanatory variables. Using the **anova( )** function, I determined whether adding the third variable, my interaction term, was worth the change in the R^2 value. If I concluded that I should include the interaction term, I would follow the same steps mentioned earlier to check the assumptions for my new model. If not, I would conclude that my model without the interaction term is the best fit model.

Data Results

Exhibit 3. Summaries of Best Models

```
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty`,
##     data = Model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.159 -17.076   0.121  13.442  67.411
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         32.58924    1.38201  23.581  < 2e-16 ***
## `Percent in Poverty`  0.59858    0.07612   7.864 1.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.02 on 558 degrees of freedom
## Multiple R-squared:  0.09977,    Adjusted R-squared:  0.09815
## F-statistic: 61.84 on 1 and 558 DF,  p-value: 1.937e-14
```

```
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income`, data = Model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.802 -17.178  -0.498  12.825  70.071
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.863e+01  3.663e+00  10.546  < 2e-16 ***
## `Percent in Poverty`  4.863e-01  9.874e-02   4.926 1.11e-06 ***
## `Median Income`      -6.593e-05  3.705e-05  -1.780   0.0757 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.97 on 557 degrees of freedom
## Multiple R-squared:  0.1049, Adjusted R-squared:  0.1016
## F-statistic: 32.62 on 2 and 557 DF,  p-value: 4e-14
```

```
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income` + `City Indicator`, data = Model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.626 -17.163  -0.692  12.787  70.031
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.789e+01  3.743e+00  10.122  < 2e-16 ***
## `Percent in Poverty`  4.593e-01  1.027e-01   4.474  9.3e-06 ***
## `Median Income`      -6.013e-05  3.753e-05  -1.602    0.110
## `City Indicator`      2.502e+00  2.594e+00   0.964    0.335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.98 on 556 degrees of freedom
## Multiple R-squared:  0.1064, Adjusted R-squared:  0.1015
## F-statistic: 22.06 on 3 and 556 DF,  p-value: 1.656e-13
```

```
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income` + `City Indicator` + `Percent Employed`,
##     data = Model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.900 -16.913  -0.865  12.746  69.993
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.336e+01  8.404e+00   5.160 3.45e-07 ***
## `Percent in Poverty`  4.248e-01  1.131e-01   3.756 0.000191 ***
## `Median Income`      -5.533e-05  3.813e-05  -1.451 0.147282
## `City Indicator`      2.428e+00  2.597e+00   0.935 0.350188
## `Percent Employed`   -8.620e-02  1.185e-01  -0.728 0.467088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.99 on 555 degrees of freedom
## Multiple R-squared:  0.1072, Adjusted R-squared:  0.1008
## F-statistic: 16.66 on 4 and 555 DF,  p-value: 6.632e-13
```

```
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income` + `City Indicator` + `Percent Employed` +
##     `Percent Hispanic`, data = Model1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57.57 -17.47  -0.86  12.92  69.97
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.341e+01  8.409e+00   5.163  3.4e-07 ***
## `Percent in Poverty`  3.891e-01  1.295e-01   3.005  0.00278 **
## `Median Income`      -5.196e-05  3.861e-05  -1.346  0.17889
## `City Indicator`      2.275e+00  2.613e+00   0.871  0.38430
## `Percent Employed`   -9.340e-02  1.192e-01  -0.784  0.43364
## `Percent Hispanic`    5.374e-02  9.471e-02   0.567  0.57069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24 on 554 degrees of freedom
## Multiple R-squared:  0.1077, Adjusted R-squared:  0.09967
## F-statistic: 13.38 on 5 and 554 DF,  p-value: 2.508e-12
```

```
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income` + `City Indicator` + `Percent Employed` +
##     `Percent Black` + `Percent Hispanic`, data = Model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.769 -17.497  -0.807  13.009  69.947
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.349e+01  8.425e+00   5.162 3.42e-07 ***
## `Percent in Poverty`  3.902e-01  1.297e-01   3.008  0.00275 **
## `Median Income`      -5.236e-05  3.870e-05  -1.353  0.17655
## `City Indicator`      2.568e+00  3.014e+00   0.852  0.39467
## `Percent Employed`   -9.330e-02  1.193e-01  -0.782  0.43455
## `Percent Black`      -3.407e-02  1.744e-01  -0.195  0.84520
## `Percent Hispanic`    5.611e-02  9.557e-02   0.587  0.55736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.02 on 553 degrees of freedom
## Multiple R-squared:  0.1078, Adjusted R-squared:  0.0981
## F-statistic: 11.13 on 6 and 553 DF,  p-value: 9.617e-12
```

Exhibit 4. Analysis of Variance #1

```
## Analysis of Variance Table
##
## Model 1: `Percent Rent Burdened` ~ `Percent in Poverty`
## Model 2: `Percent Rent Burdened` ~ `Percent in Poverty` + `Median Income`
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    558 321975
## 2    557 320154  1    1820.1 3.1667 0.0757 .
```

Exhibit 5. Step Model Variable Selection

```
## Step:  AIC=3561.23
## `Percent Rent Burdened` ~ `Percent in Poverty` + `Median Income`
##
##                         Df Sum of Sq    RSS    AIC
## <none>                                320154 3561.2
## - `Median Income`        1    1820.1 321975 3562.4
## - `Percent in Poverty`   1   13944.6 334099 3583.1

##
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income`, data = Model1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.802 -17.178  -0.498  12.825  70.071
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.863e+01  3.663e+00  10.546  < 2e-16 ***
## `Percent in Poverty`  4.863e-01  9.874e-02   4.926 1.11e-06 ***
## `Median Income`      -6.593e-05  3.705e-05  -1.780   0.0757 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.97 on 557 degrees of freedom
## Multiple R-squared:  0.1049, Adjusted R-squared:  0.1016
## F-statistic: 32.62 on 2 and 557 DF,  p-value: 4e-14
```

Exhibit 6. Heteroskedasticity #1

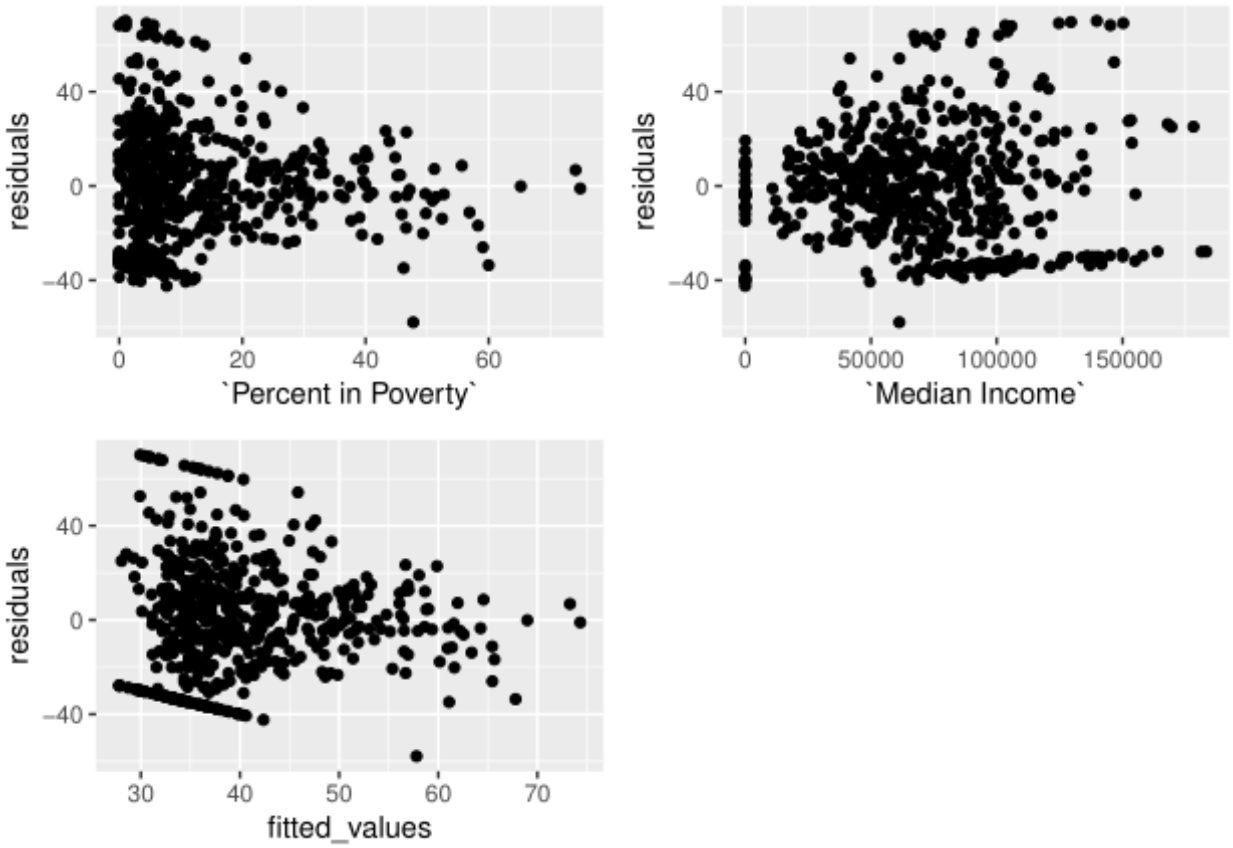


Exhibit 7. Transformations

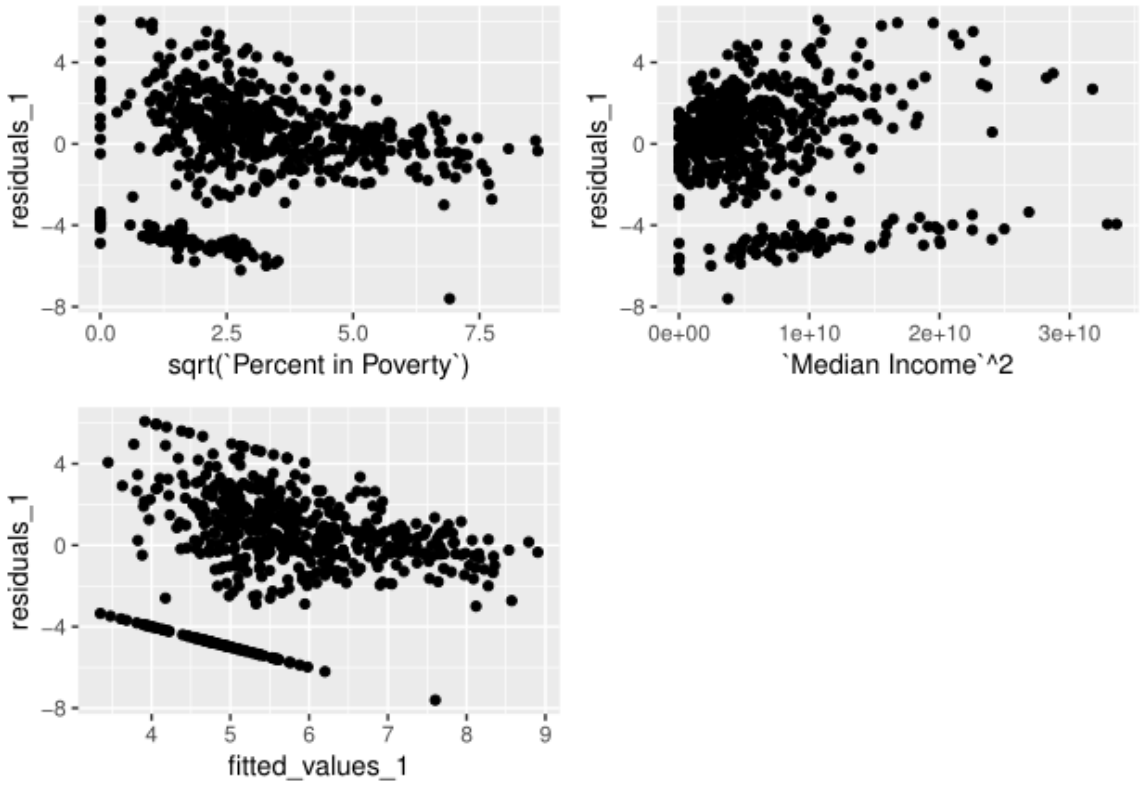


Exhibit 8. Summary of Transformed Model

```
## Call:
## lm(formula = sqrt(`Percent Rent Burdened`) ~ sqrt(`Percent in Poverty`) +
##     `Median Income`^2, data = Model1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.602 -1.104  0.343  1.617  6.086
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.879e+00  5.106e-01   9.555  < 2e-16 ***
## sqrt(`Percent in Poverty`)  4.769e-01  8.393e-02   5.681 2.15e-08 ***
## `Median Income`            -9.333e-06  4.113e-06  -2.269   0.0237 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.579 on 557 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.147
## F-statistic: 49.15 on 2 and 557 DF,  p-value: < 2.2e-16
```

Exhibit 9. Autocorrelation #1

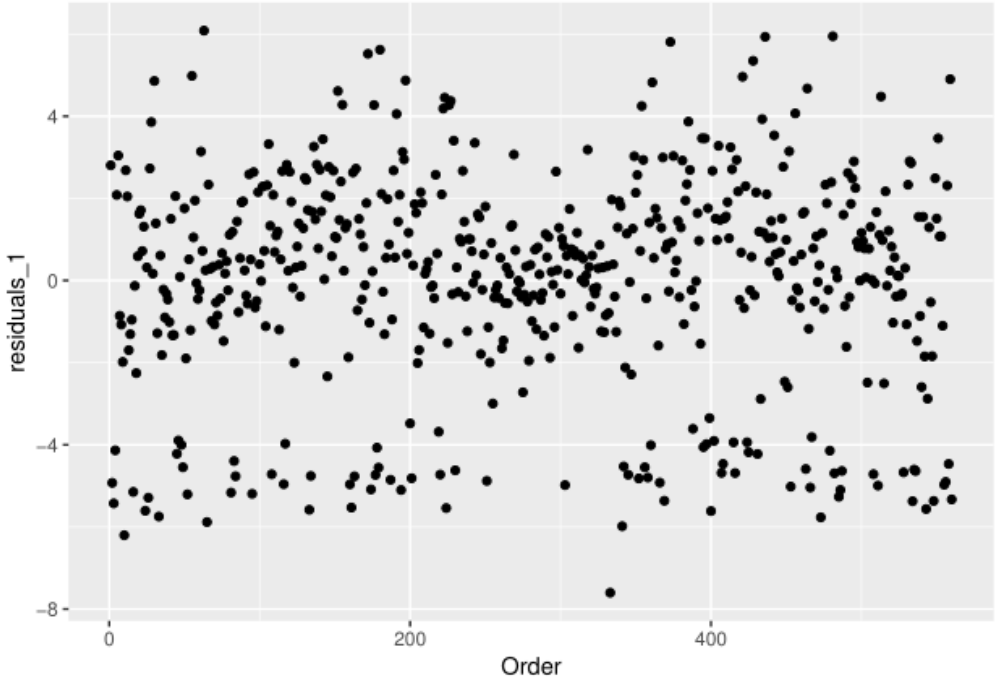


Exhibit 10. Comparison of Coefficients

Without Outliers

```
## Coefficients:
##                              Estimate
## (Intercept)                 5.697e+00
## sqrt(`Percent in Poverty`) 2.353e-01
## `Median Income`             3.320e-06
```

With Outliers

```
## Coefficients:
##                              Estimate
## (Intercept)                 4.879e+00
## sqrt(`Percent in Poverty`)  4.769e-01
## `Median Income`            -9.333e-06
```

Exhibit 11. Normality #1

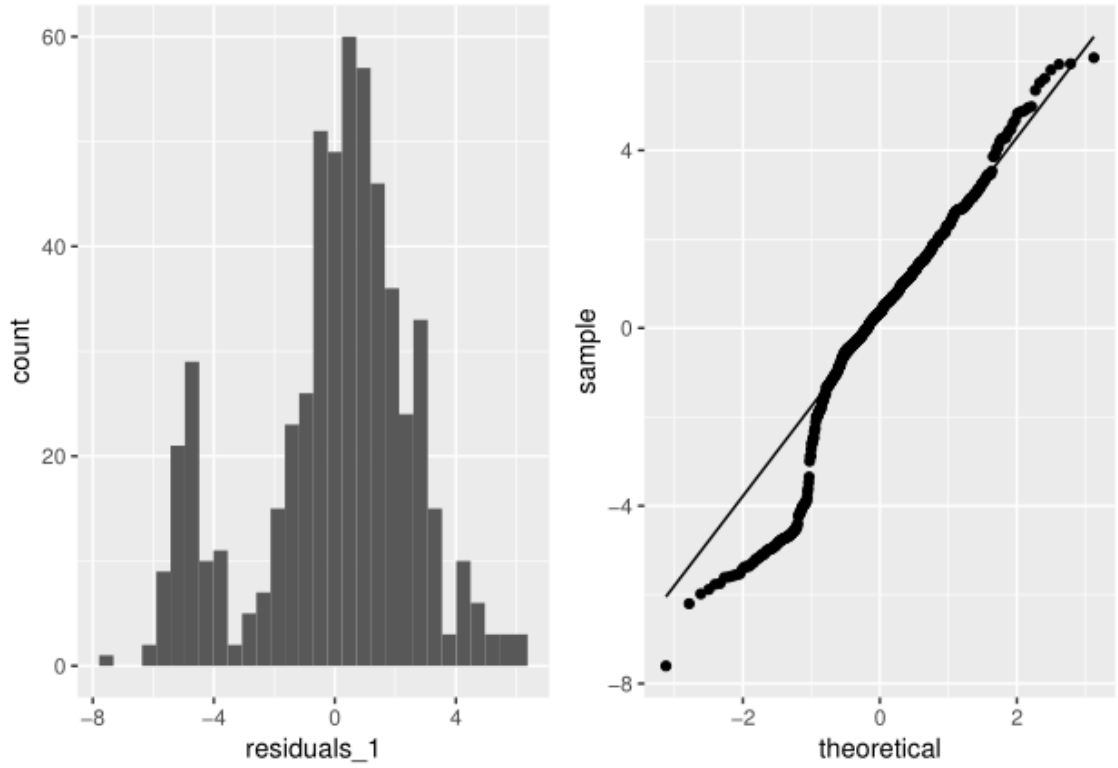


Exhibit 12. Interaction Between Variables

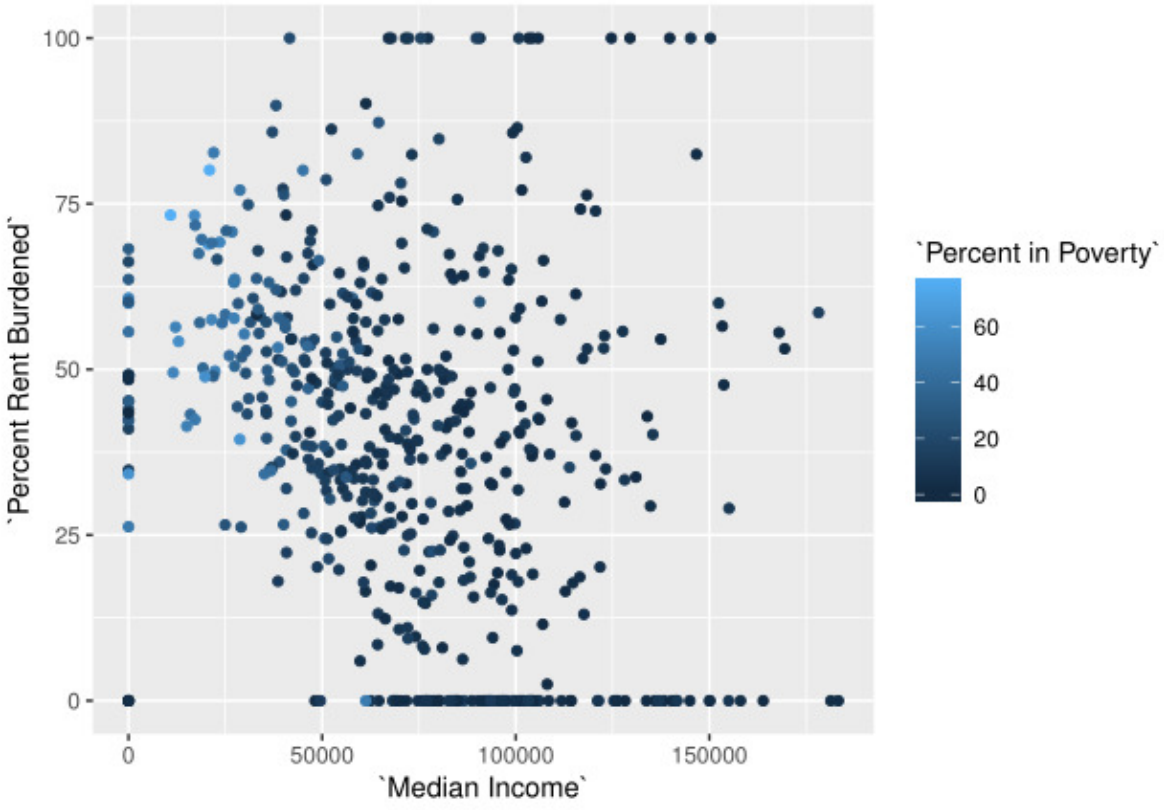


Exhibit 13. Analysis of Variance #2

```
## Analysis of Variance Table
##
## Model 1: sqrt(`Percent Rent Burdened`) ~ sqrt(`Percent in Poverty`) +
##     `Median Income`^2
## Model 2: sqrt(`Percent Rent Burdened`) ~ sqrt(`Percent in Poverty`) +
##     `Median Income`^2 + `Interaction Term`
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    557 3703.7
## 2    556 3701.2  1    2.5438 0.3821 0.5367
```

Exhibit 14. Model with One Set of Outliers Removed

```
## Step:  AIC=2843.42
## `Percent Rent Burdened` ~ `Percent in Poverty` + `Median Income` +
##     `Percent Employed`
##
##                        Df Sum of Sq    RSS    AIC
## <none>                              185835 2843.4
## - `Median Income`       1    1933.3 187768 2846.3
## - `Percent Employed`    1    2558.0 188393 2847.9
## - `Percent in Poverty`  1    3227.0 189062 2849.6

##
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent in Poverty` +
##     `Median Income` + `Percent Employed`, data = Outlier_removed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41.36 -12.65  -0.77  10.40  57.50
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.509e+01  7.855e+00   7.013 8.13e-12 ***
## `Percent in Poverty`  2.869e-01  1.003e-01   2.860  0.00443 **
## `Median Income`       8.367e-05  3.780e-05   2.214  0.02734 *
## `Percent Employed`   -2.828e-01  1.111e-01  -2.546  0.01121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.86 on 471 degrees of freedom
## Multiple R-squared:  0.05957,    Adjusted R-squared:  0.05358
## F-statistic: 9.944 on 3 and 471 DF,  p-value: 2.287e-06
```

Exhibit 15. Heteroskedasticity #2

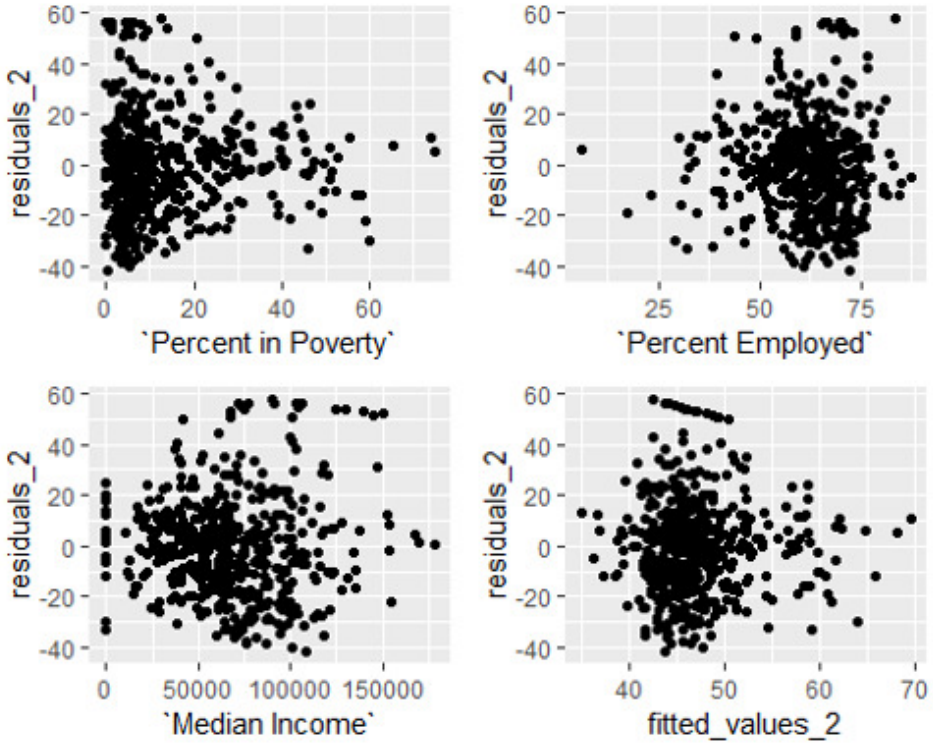


Exhibit 16. Model with Both Sets of Outliers Removed

```
## Step:  AIC=2584.29
## `Percent Rent Burdened` ~ `Percent Employed` + `Percent Hispanic`
##
##                      Df Sum of Sq    RSS    AIC
## <none>                             128857 2584.3
## - `Percent Hispanic`  1    3719.7 132577 2595.3
## - `Percent Employed`  1    6606.3 135464 2605.1

##
## Call:
## lm(formula = `Percent Rent Burdened` ~ `Percent Employed` + `Percent
Hispanic`,
##     data = Outlier_removed1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.069 -12.141   0.394  10.088  48.927
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        65.39880    5.15000  12.699  < 2e-16 ***
## `Percent Employed` -0.37421    0.07756  -4.825 1.92e-06 ***
## `Percent Hispanic`  0.19715    0.05446   3.620 0.000328 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.85 on 454 degrees of freedom
## Multiple R-squared:  0.1131, Adjusted R-squared:  0.1091
## F-statistic: 28.93 on 2 and 454 DF,  p-value: 1.489e-12
```

Exhibit 17. Heteroskedasticity #3

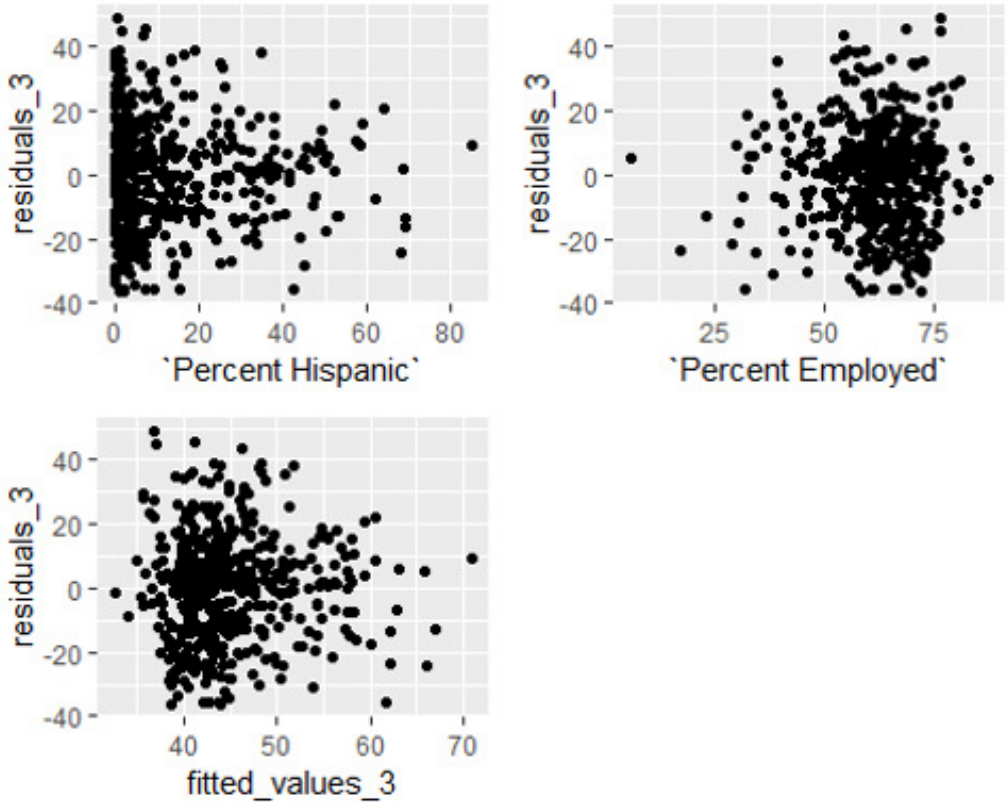


Exhibit 18. Autocorrelation #2

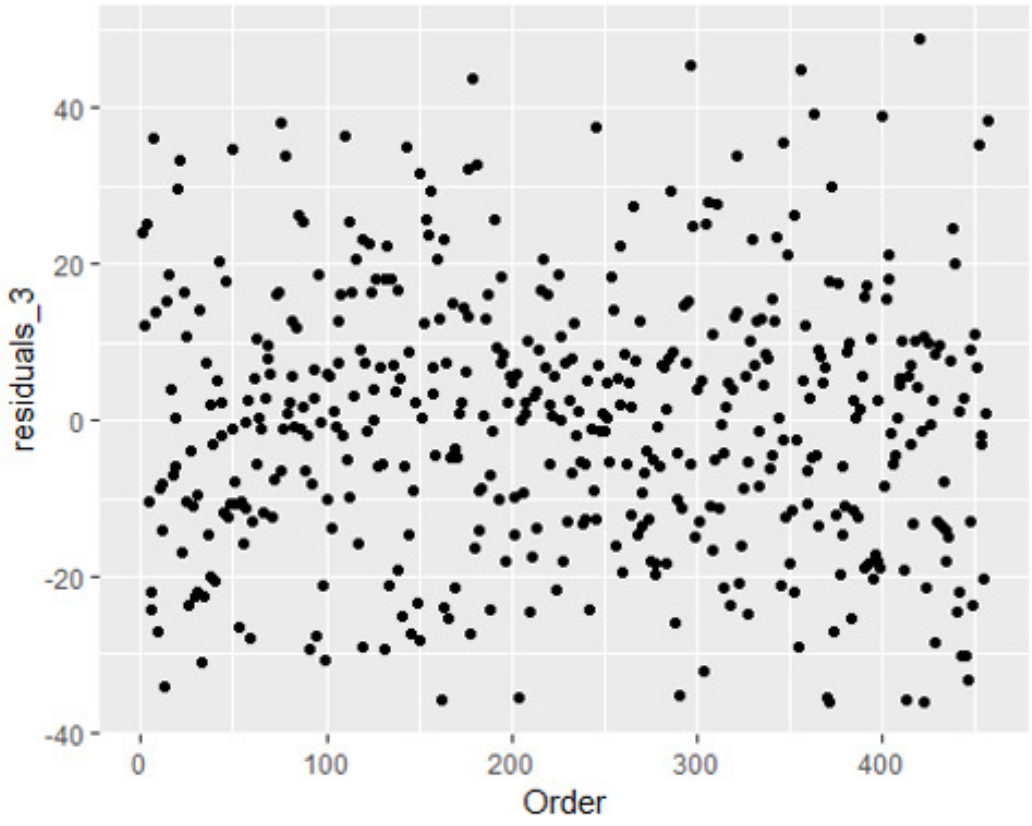


Exhibit 19. Normality #2

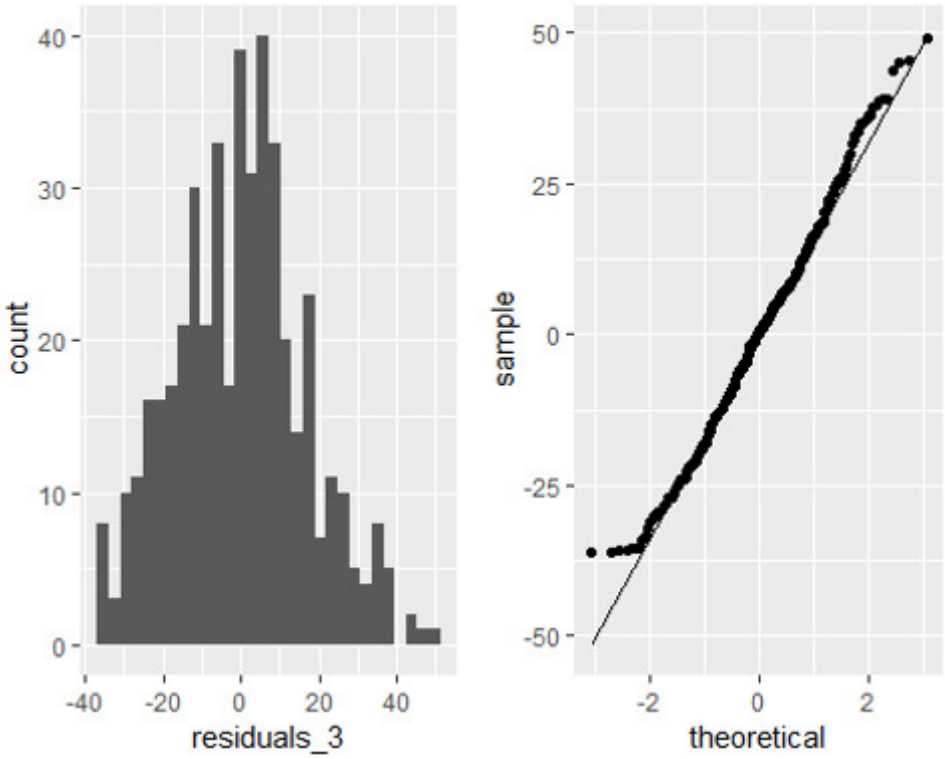


Exhibit 20. Analysis of Variance #3

```
## Analysis of Variance Table
##
## Model 1: `Percent Rent Burdened` ~ `Percent Hispanic` + `Percent Employed`
## Model 2: `Percent Rent Burdened` ~ `Percent Employed` + `Percent Hispanic`
+
##     `Interaction Term`
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    454 128857
## 2    453 128759  1    98.324 0.3459 0.5567
```

Discussion & Explanation

Exhibit 3 shows the summaries of the six best models created. It is clear that when one more variable is added to each model, the R^2 values increased every time with a low of 0.0981 for the six explanatory variable models and a high of 0.1016 for the two explanatory variable models. Note that I am going by the adjusted R^2 value as opposed to the multiple R^2 value. This is because the adjusted R^2 value has been modified from the multiple R^2 value for the number of predictors in the model. Every time a variable is added to a model, the multiple R^2 statistic will increase, but that is not necessarily true with the adjusted R^2 statistic; just because another variable is added to a model does not mean the model is a better fit than one with fewer variables.

Going by the R^2 value alone suggests the best fit model would be the final model, but to confirm findings it was matched against results from other variable selection techniques. Exhibit 4 shows an analysis of variance table that resulted when I compared my one-variable model containing the percentage of a census block group in poverty and the two-variable model containing the percentage in poverty and the median income of a specific census block group. For this output, the p-value indicated whether including the second variable was worth the increase in the R^2 statistic. A P-value of 0.0757 suggests that at the 5% significance level, adding the second variable was not worth it. Yet, this study west against that because it relatively to have two variables to predict a response variable than to only have one.

Knowing that one statistics technique called for a two-variable model and another called for only a one-variable model, it was decided to try one last technique. Exhibit 6 shows the output of a specific function in RStudio that looks at the AIC statistic for certain models and stops when the AIC does not get any lower. What is shown in Exhibit 6 is only part of the longer output done. It shows the last step where the program chose what it deemed the best fitting model — one containing percent in poverty and median income — to model the percentage of rent burdened households in a specific census block group. Therefore, it was concluded that the best fit model was that using what was found in two out of three variable selection techniques (one might say in all three).

After determining the variables, the researcher checked the model assumptions. As mentioned earlier, residuals should show equal variance throughout. The three graphs of Exhibit 6 show the percent in poverty on the x-axis and residuals on the y-axis. The second graph shows median income on the x-axis and residuals on the y-axis, and the third graph shows fitted values, or y-values calculated from the model on the x-axis and residuals on the y-axis. Graph one shows a clear curving pattern in the points, where the variance starts off large with lower percentages of poverty and small with higher percentages of poverty. From this it was concluded that variances are not equal throughout. The second graph shows a coning pattern, where variances start small with lower median incomes and end up large with higher median incomes; the third graph shows what were initially called outliers at the top and bottom of the graph and an uneven pattern of data points in between. It is clear in the third graph that variances are unequal, showing heteroskedasticity along with the first two graphs, all of which suggest transformations on variables needed to be performed to try to get the data to more equal variances or homoskedasticity.

After trying numerous transformations on each variable including taking the square root of a variable, taking the natural log of the variable, and dividing 1 by the variable, it was found that taking the square root of the percentage of rent burdened households was the most accurate and useful variable. This response variable was the result of taking the square root of the percentage in poverty, one of my explanatory variables, and finally squaring median income, my second explanatory variable. Exhibit 7 shows the same three graphs, except the residuals come from a model containing all the transformations: the fitted values come from the model with transformations, and the explanatory variables graphed are the transformed values. Comparing the graphs in Exhibit 7 to those in Exhibit 6, shows a subtle yet more consistent variance throughout in transformed model. While not perfect, compared to the previous model without the transformations. the transformations the data broke out into a couple clusters of data points made the heteroskedasticity better. Also, the transformations seemed to make the outliers

on top of the fitted values vs. residuals graph less pronounced, making them fit the pattern of the data better. Although not textbook-perfect, these transformations adhered to model assumptions better than the model without the transformations.

Accordingly, it was decided make the new model one that included transformations of my two explanatory variables and my response variable. Exhibit 8 shows a summary of the new model. The adjusted R^2 statistic is 0.147, whereas without the transformations, it was 0.1016. Increased adjusted R^2 indicates that the model with the transformations is a better fit than the model without the transformations, keeping the number of explanatory variables constant.*

Exhibit 9 is a graph of the order of data points on the x-axis and residuals from the model containing transformed variables. Note that from here, this report will refer to those residuals as just residuals without explaining they come from the transformed model. This exhibit checks the autocorrelation, or whether one point is related to another, to identify patterns or clusters. Although there seems to be two different clusters of points, they are not completely distinct. While all the points seem to fit an overall (albeit not perfect) pattern, the majority of the data meets the model assumption that the data points are independent from another. The two clusters in the data resulted from outliers in the data which were points where the percentage of rent burdened households was 0.

Exhibit 10 shows a comparison of coefficients from two different models. The models themselves are constant, where they are both two-variable models containing the same transformations done on the variables as above. The only difference is that the coefficients on the left come from a model using data without the outliers and the coefficients on the right come from the model containing all the data points, including outliers. These outliers appeared to have a strong influence on the coefficients, especially when it came to median income. Therefore, the outliers were not removed from the data set and had to have their accuracy verified. The following table compares averages of certain variables for a data set without the outliers and one with only the outliers. In other words, one data set had the percentage of rent burdened households equal to 0 in census block groups and the other had the percentage of rent burdened households not equal to 0 in census block groups.

Table 1. Rent Burdened by Variables

Variable	0% Rent Burdened	>0% Rent Burdened
Median Income	$97,762.15	$65,927.81
Percentage Employed	65.77%	61.67%
Percentage in Poverty	4.13%	13.79%
Percentage of Population Black	2.24%	5.43%
Percentage of Population Hispanic	5.38%	13.02%
Number of Housing Units Being Rented	36	221
Location in the City of Worcester	.12	.29

Starting with the median income, census block groups that contain zero households that are rent burdened

*A perfectly fit model has an R^2 value of 1, which is far off from 0.147. Yet, it is important to note that this is the best fitting model using the available public data for this specific location containing 560 census block groups. By using this model, it is fair to say that with some certainty, that one can take the square root of the percentage of individuals in poverty for a specific census block group and multiply it by 0.4769, subtract the median income squared multiplied by -.000009333 and add that to 4.879 to find the square root of the percentage of rent burdened households. Using the information calculated in the model can help provide a better idea of what is the situations in certain census block groups, the ultimate goal of linear modeling.

have a median income almost $35,000 greater than census block groups where there are households that are rent burdened. Having more money per household obviously provides a household a greater ability to spend less than 30% of their income on rent. More people are employed in census block groups with zero rent burdened households, meaning more people are making money, likely helping to explain why the median household income is higher. The percentage of individuals classified as being in poverty comprises almost 1/3 of the rent burdened for census block groups with rent burdened households. The other variables also show that these outliers are indeed accurate. Even the last variable hints toward the overall problem that began this research. As noted above, a 0 to census block groups located outside the city and a 1 to census block groups located in the city was assigned. The average of that variable is the percentage of census block groups located in the City of Worcester. In census block groups with zero rent burdened households, 12% are located in the City of Worcester, and 29% of the census block groups with at least one rent burdened household are located the City of Worcester.

Table 2. Rent Burdened by Block Groups

	0 Rent Burdened Households	Greater than 0 Rent Burdened Households	All Households Rent Burdened	Total Census Block Groups
City	10 (7%)	135 (91%)	4 (3%)	149 (100%)
Outside the City	75 (18%)	322 (78%)	14 (3%)	411 (100%)
Total Census Block Groups in the County	85 (15%)	457 (82%)	18 (3%)	560 (100%)

Additionally, 93% of all census block groups in the City of Worcester have a percentage of rent burdened households greater than 0 while outside of the city, that percentage is only 82% showing there is a greater need for resources and time spent on affordable housing inside the city as opposed to outside.

Exhibit 11 shows two graphs, one of which is a histogram of the residuals and the other is a normal probability plot. The histogram shows almost two different bell curves in the same graph. Again, the assumption was that the first bell curve was the outlies and the second was the rest of the data, but both bell curves look approximately normal. The normal probability plot was more difficult to analyze. Instead of relying on shape, it was determined whether the points lied close to the linear line shown on the graph. Beyond the first third of data points, the points were relatively close, if not on, the line. The first third on the other hand are relatively far from the line. It is assumed that those points belong to the census block outliers. By looking into these graphs and factoring in outliers, it was concluded the majority of the residuals follow a normal distribution, verifying the model assumptions.

Finally the way all the variables interacted with each other was tested. Exhibit 12 shows a graph of median income vs. the percentage of rent burdened households, colored by the percentage of individuals in poverty, where each point represents 1 of 560 census block groups in Worcester County. There seems to be a clear pattern where the lighter colored points, representing a higher percentage of individuals in poverty, fall at a lower median income. This makes intuitive sense so what statisticians call an interaction term was created in which created a variable that multiplied the two explanatory variables together. After creating a model containing three explanatory variables, one of which was my new interaction term and the other two my original transformed explanatory varia-

bles, researchers created an analysis of variance table to compare the two-variable model to the new three-variable model. The P-value of 0.5367, shown in Exhibit 13, is so large that at any common significance level used in statistics there is no way to include the interaction term in my final model.

After all of analysis of model assumptions, it was concluded that although my transformed model was closer to meeting assumptions, model assumptions were still not being fully met. Therefore, researchers created a data set without a set of outliers in which the percent of rent burdened households in each census block group is 0. Exhibit 14 shows the output of the step function in RStudio, showing that in a data set where all census block groups have at least one rent burdened household, the best fit model contains the percentage of individuals in poverty, the percentage of individuals over the age of 16 that are employed, and the median income as explanatory variables.

As with all previous models before transformations were performed, researchers checked to see if they were necessary. Exhibit 15 shows three graphs of the explanatory variables vs. residuals from the new model and a fourth graph of the fitted values from the new model vs the residuals from the new model. Looking at the fourth graph, a set of points at the top did not fit the general pattern, meaning more outliers corresponding to census block groups where the percentage of rent burdened individuals is 100 had to be removed.

After creating a data set where all census block groups had a percentage of rent burdened individuals between, but not including 0 and 100, I once again used the step function. Results are shown in Exhibit 16. Unlike previous variable selections, this new model was different: it did not include median income or the percentage of individuals in poverty. Now, the coding is telling me to model the percentage of rent burdened individuals by using the percentage of Hispanic individuals in a population for a certain census block group and the percentage of the population over the age of 16 who are employed. Once again, the new modeled was checked to see if it adhered to linear model assumptions.

Exhibit 17 shows the check for heteroskedasticity. When reviewing the graphs (shown), it seemed the variances were at equal throughout which prompted me to skip making any transformations to any of the variables in the model. Exhibits 18-20 show similar analyses as performed before, checking for autocorrelation, normality, and whether an interaction term should be added. It was determined that there was almost no autocorrelation shown. The residuals looked approximately normal in both graphs, more normal than the transformed model containing all of the 560 census blocks throughout Worcester County. Lastly, it was decided an interaction term would not help the model become a better fit.

The analyses were concluded and it was determined that my best-fit model, meeting all necessary linear model assumptions, is: $y_i = 65.39880 + .19715x_{1.i} - .37421x_{2.i}$ where y_i represents the percentage of rent burdened households for the i^{th} census block group, $x_{1.i}$ represents the percentage of Hispanic individuals within a population for the i^{th} census block group, and $x_{2.i}$ represents the percentage of individuals over the age of 16 in a population who are employed for the i^{th} census block group.

This model shows that a census block group that is 0% Hispanic and has no employed individuals will have 65.4% of its population who rent being rent burdened. When looking at the effect the percentage of a population that is Hispanic has on the percentage of rent burdened households, if you hold everything else constant, an increase of 1% will increase the percentage of rent burdened individuals 0.2%. When looking at the effect the percentage of individuals greater than 16 years old in a population who are employed has on the percentage of rent burdened households, if you hold everything else constant, an increase of 1% will decrease the percentage of rent burdened individuals 0.4%.

In short, this model indicates employment is a stronger indicator than ethnicity when it comes to being able to afford rent. However, a variety of socio-economic and educational factors influence employment and further studies are needed that focus on those factors to more fully understand their statistical affect on cost burden.

REFERENCES

2010 U.S. Census, http://www.census.gov/

2017 American Community Survey 5-Year Estimates, http://www.census.gov/

"2017 Poverty Guidelines." *ASPE*, U.S Department of Health and Human Services, 12 Jan. 2018, aspe.hhs.gov/2017-poverty-guidelines#guidelines.

Ashok, Sowmiya. "The Rise of the American 'Others'." *The Atlantic*, Atlantic Media Company, 27 Aug. 2016, www.theatlantic.com/politics/archive/2016/08/the-rise-of-the-others/497690/.

"Housing Choice Vouchers Fact Sheet." *HUD*.gov, U.S. Department of Housing and Urban Development, www.hud.gov/topics/housing_choice_voucher_program_section_8.

"How the Census Bureau Measures Poverty." *United States Census Bureau*, U.S. Department of Commerce, 16 Aug. 2018, www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html.

US Census Bureau. "Current Population Survey." *United States Census Bureau* , U.S. Department of Commerce, 27 Feb. 2019, www.census.gov/programs-surveys/cps/technical-documentation/subject-definitions.html#household.