Data analysis of select U.S. Census data for census year 2010

Melanie A. Fitzgerald

Master of Business Administration Program, Carroll

University BUS 674 – Data Analytics in Practice

John A. Michl, MSBIA, MBA, CBIP

January 26, 2025

EXECUTIVE SUMMARY

Analyzing census data is important to understanding population which is essential for decision making that affects policy, development and resource allocations. “The census tells us who we are and where we are going as a nation, and helps our communities determine where to build everything…It helps the government decide how to distribute funds and assistance to states and localities (US Census Bureau, 2019).” IPUMS is a free database from the University of Minnesota that “provides census and survey data from around the world integrated across time and space…to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community context (Ruggles et al., 2024).” This report uses U.S. Census data from census year 2010 to explore responses from different questions of the census and the relationships between them.

OVERVIEW

Data was extracted from the U.S. Census Data for Social, Economic, and Health Research database available at IPUMS USA. The extracted dataset includes information about households, geographic location, economics, demographic, education, and work-related data (hours, travel time, income) for census years 2020, 2015, and 2010. This report explores select data from census year 2010 for states New York, Delaware, and Pennsylvania using different analytic methods.

METHODS OF ANALYSIS

SPSS is used for the exploratory data analysis using descriptive statistics, correlation analysis and regression analysis of the select dataset variables below:

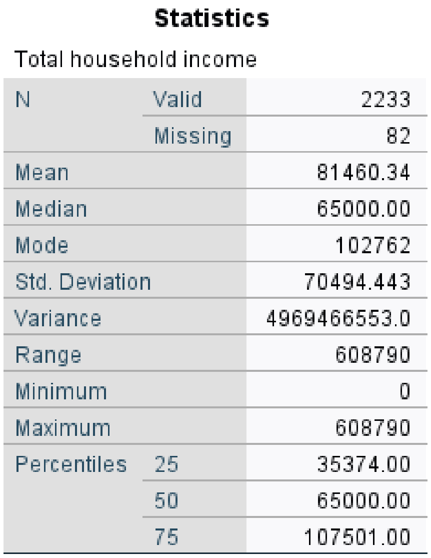
Total household income (HHINCOME)
House value (VALUEH)
Marital status (MARST)
Educational attainment – general (EDUC)
Usual hours worked per week (UHRSWORK)
Means of transportation to work (TRANWORK)

Undefined or missing values for each variable were coded prior to analysis. The data is a mix of both continuous numerical and categorical values.

DATA ANALYSIS AND RESULTS

Descriptive statistics

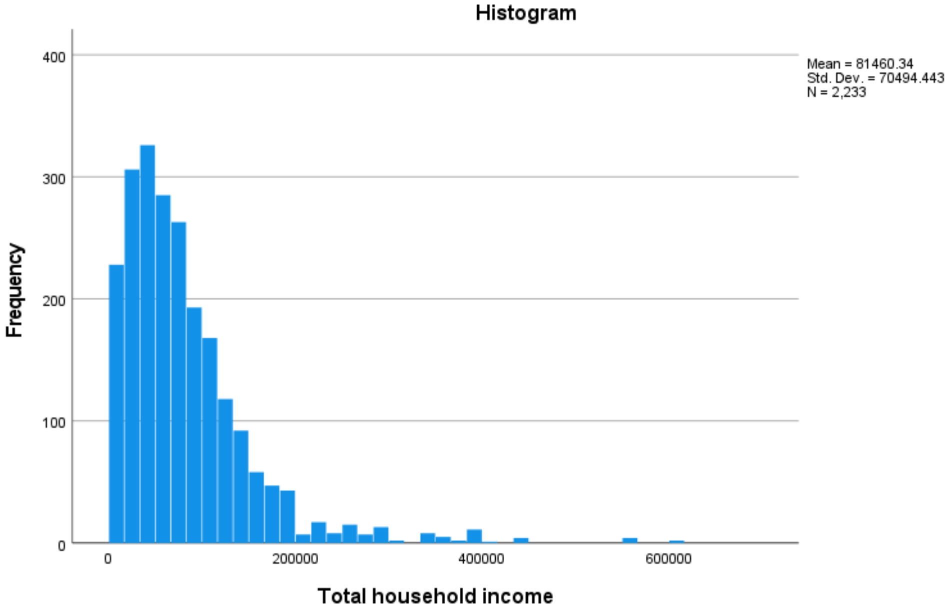
Total household income (HHINCOME) is a continuous numeric variable that reports the “total money income of all household members age 15+ during the previous year (Ruggles et al., 2024).” The descriptive statistics show a mean household income of $81,460 with a range of $608,790. The quartiles show that 75% of household income falls below $107,501. Visualization of the data show a right-skewed histogram with a concentration of household incomes on the left side of the graph and tailing off around $200,000 with outliers to the far right.



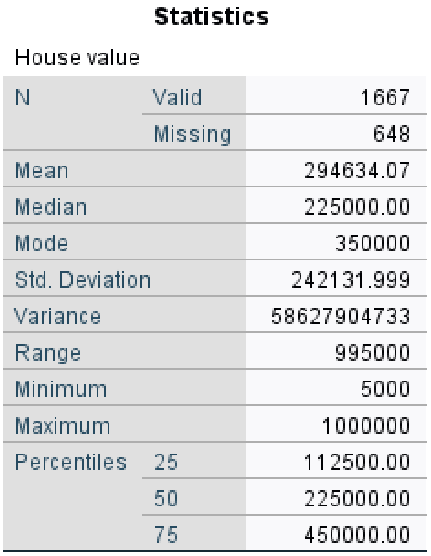
Statistics

Total household income

N	Valid	2233
	Missing	82
Mean		81460.34
Median		65000.00
Mode		102762
Std. Deviation		70494.443
Variance		4969466553.0
Range		608790
Minimum		0
Maximum		608790
Percentiles	25	35374.00
	50	65000.00
	75	107501.00



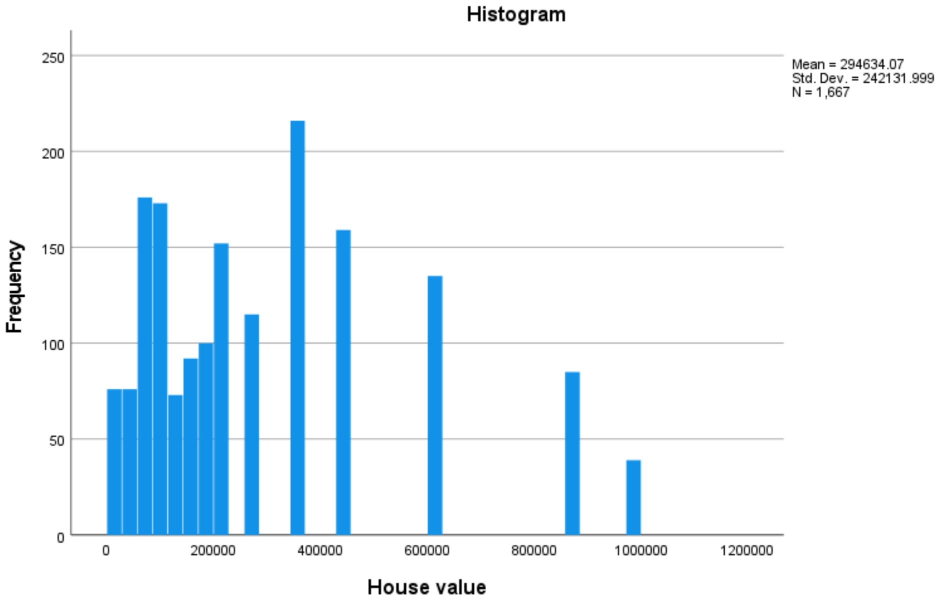
House value (VALUEH) is a continuous numeric variable that “reports the value of housing units (Ruggles et al., 2024).” The descriptive statistics show a mean house value of $294,634 with a range of $995,000. The quartiles show that 75% of house values fall below $450,000. Visualization of the data shows a right-skewed histogram with a concentration of house values on the left side of the graph.



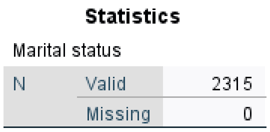
Statistics

House value

N	Valid	1667
	Missing	648
Mean		294634.07
Median		225000.00
Mode		350000
Std. Deviation		242131.999
Variance		58627904733
Range		995000
Minimum		5000
Maximum		1000000
Percentiles	25	112500.00
	50	225000.00
	75	450000.00



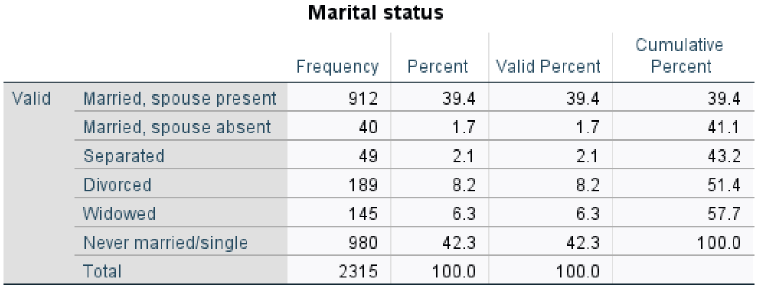
Marital status (MARST) is a categorical value of the different marital statuses defined in the census questionnaire. Unlike household income and house values, the data values fit into six specific categories as shown in the frequency table below. The most frequent response is *Never married/single* with 980 responses followed by *Married, spouse present* with 912 responses. *Married, spouse absent* had the least response with 40 responses. The bar chart shows the frequency distribution of the six marital status categories.



Statistics

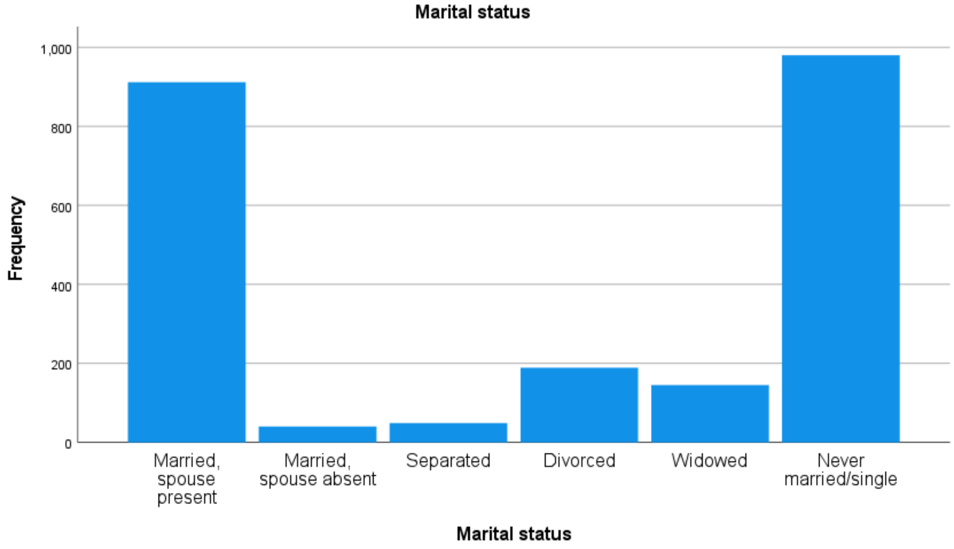
Marital status

N	Valid	2315
	Missing	0

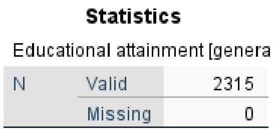


Marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Married, spouse present	912	39.4	39.4	39.4
	Married, spouse absent	40	1.7	1.7	41.1
	Separated	49	2.1	2.1	43.2
	Divorced	189	8.2	8.2	51.4
	Widowed	145	6.3	6.3	57.7
	Never married/single	980	42.3	42.3	100.0
	Total	2315	100.0	100.0	



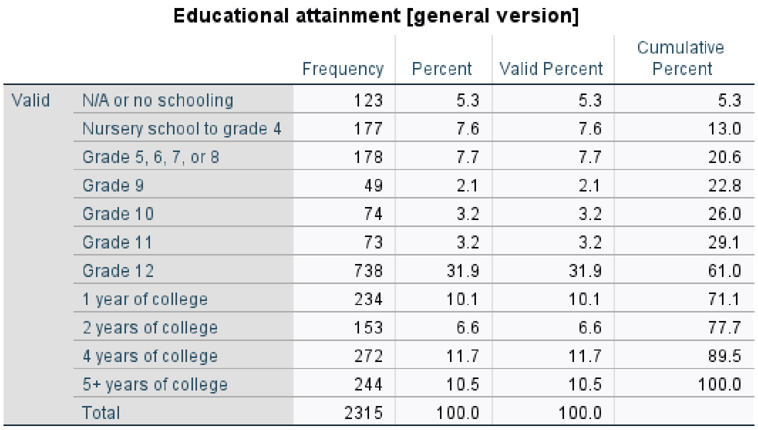
Educational attainment – general (EDUC) is a categorical value assigned to the different levels of education attainment as measured by the highest year of school or degree completed (Ruggles et al., 2024)." The eleven categories are shown in the table below with the frequency of responses for each category. The most frequent response is *Grade 12* with 738 responses followed by *4 years of college* with 272 responses. *Grade 9* had the least response with 49 responses. The bar chart shows the frequency distribution of the eleven educational attainment categories.



Statistics

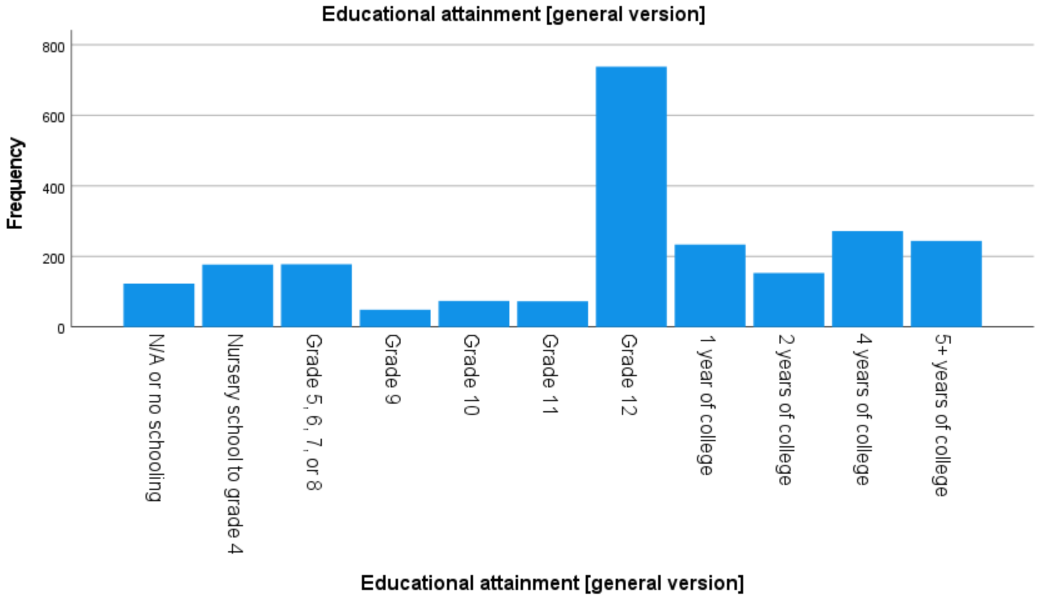
Educational attainment [general

N	Valid	2315
	Missing	0

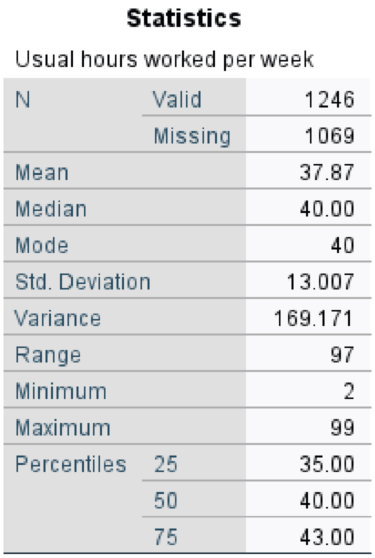


Educational attainment [general version]

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	N/A or no schooling	123	5.3	5.3	5.3
	Nursery school to grade 4	177	7.6	7.6	13.0
	Grade 5, 6, 7, or 8	178	7.7	7.7	20.6
	Grade 9	49	2.1	2.1	22.8
	Grade 10	74	3.2	3.2	26.0
	Grade 11	73	3.2	3.2	29.1
	Grade 12	738	31.9	31.9	61.0
	1 year of college	234	10.1	10.1	71.1
	2 years of college	153	6.6	6.6	77.7
	4 years of college	272	11.7	11.7	89.5
	5+ years of college	244	10.5	10.5	100.0
	Total	2315	100.0	100.0	



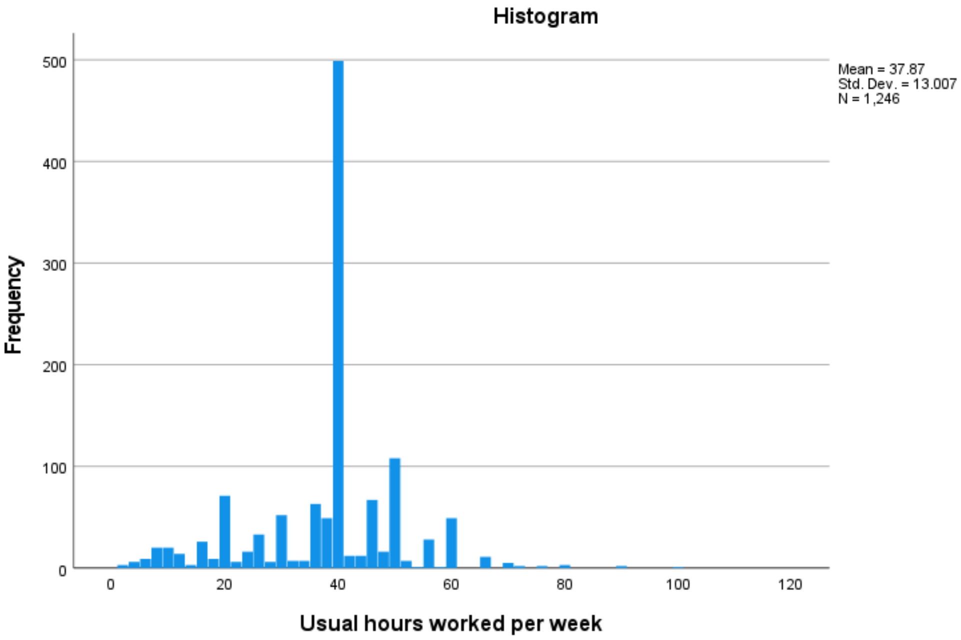
Usual hours worked per week (UHRSWORK) is a continuous numeric value that "reports the number of hours per week that the respondent usually worked, if the person worked during the previous year (Ruggles et al., 2024)." The descriptive statistics show a mean hours worked of 37.87 hours with a range of 97 hours. The histogram appears to be a normal distribution with a median value of 40 hours, which is very close to the mean. In a perfect normal distribution, the mean and median are equal. It is notable that the normal curve is tall and narrow indicating that most of the data points are clustered around the mean resulting in a small standard deviation.



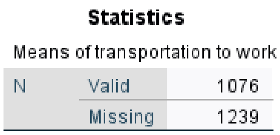
Statistics

Usual hours worked per week

N	Valid	1246
	Missing	1069
Mean		37.87
Median		40.00
Mode		40
Std. Deviation		13.007
Variance		169.171
Range		97
Minimum		2
Maximum		99
Percentiles	25	35.00
	50	40.00
	75	43.00



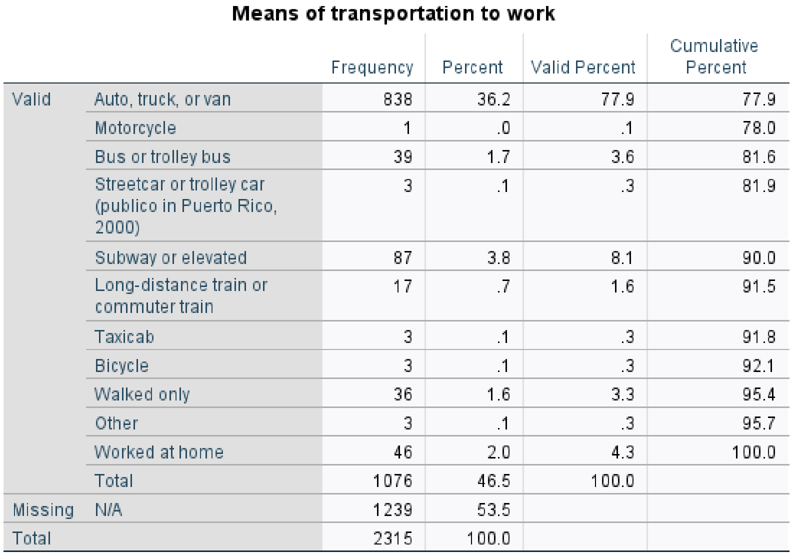
Means of transportation to work (TRANWORK) is a categorical value that "reports the respondent's primary means of transportation to work...over the course of the previous week (Ruggles et al., 2024)." The eleven categories are shown in the table below with the frequency of responses for each category. The most frequent response is *Auto, tuck, or van* with 838 responses followed by *Subway or elevated* with 87 responses. *Motorcycle* had the least response with only 1 response.



Statistics

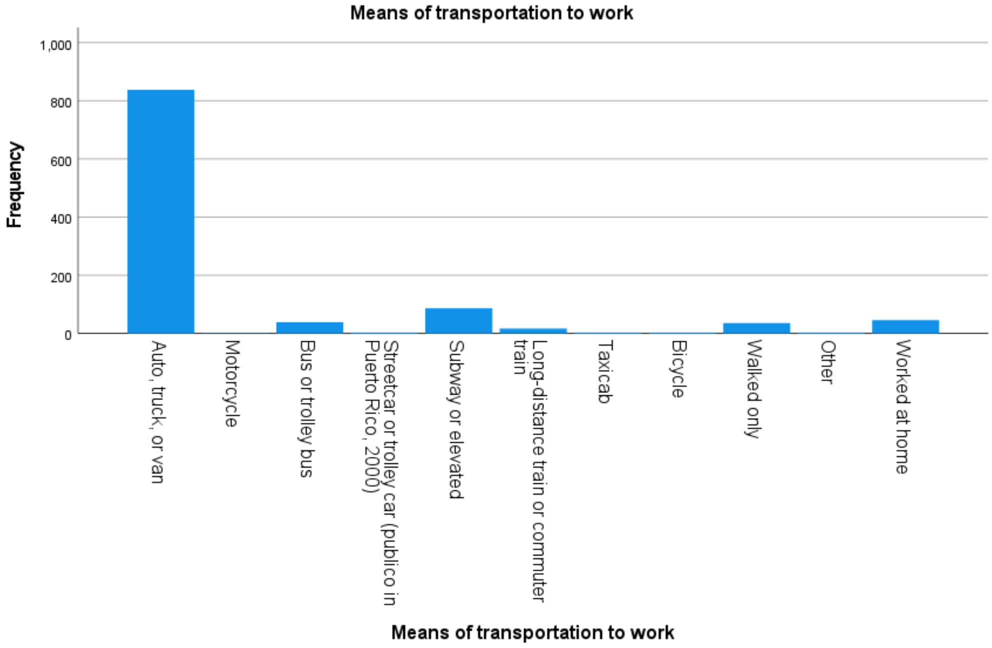
Means of transportation to work

N	Valid	1076
	Missing	1239



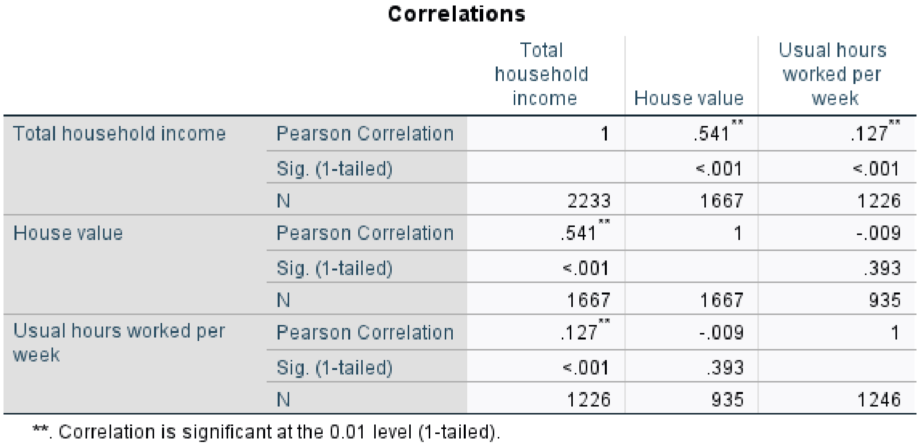
Means of transportation to work

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Auto, truck, or van	838	36.2	77.9	77.9
	Motorcycle	1	.0	.1	78.0
	Bus or trolley bus	39	1.7	3.6	81.6
	Streetcar or trolley car (publico in Puerto Rico, 2000)	3	.1	.3	81.9
	Subway or elevated	87	3.8	8.1	90.0
	Long-distance train or commuter train	17	.7	1.6	91.5
	Taxicab	3	.1	.3	91.8
	Bicycle	3	.1	.3	92.1
	Walked only	36	1.6	3.3	95.4
	Other	3	.1	.3	95.7
	Worked at home	46	2.0	4.3	100.0
	Total	1076	46.5	100.0	
Missing	N/A	1239	53.5		
Total		2315	100.0		



Correlation analysis

A correlation analysis of the continuous variables HHINCOME, VALUEH, and UHRSWORK was done and is shown in the table below. A significance of 0.01 is assumed. The analysis shows that HHINCOME has a positive correlation with both VALUEH and UHRSWORK. However, the correlation of HHINCOME with VALUEH is stronger than the correlation of HHINCOME with UHRSWORK, as indicated by the correlation coefficients 0.541 and 0.127, respectively. A correlation coefficient of 0 indicates no correlation while a coefficient of 1 indicates a perfect positive correlation.



Correlations

		Total household income	House value	Usual hours worked per week
Total household income	Pearson Correlation	1	.541**	.127**
	Sig. (1-tailed)		<.001	<.001
	N	2233	1667	1226
House value	Pearson Correlation	.541**	1	-.009
	Sig. (1-tailed)	<.001		.393
	N	1667	1667	935
Usual hours worked per week	Pearson Correlation	.127**	-.009	1
	Sig. (1-tailed)	<.001	.393	
	N	1226	935	1246

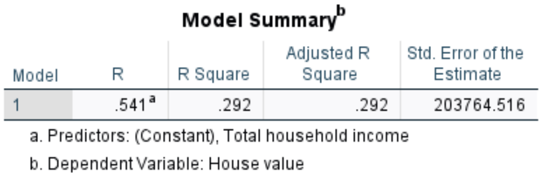
**. Correlation is significant at the 0.01 level (1-tailed).

Regression analysis

Because HHINCOME and VALUEH are shown to have a statistically significant correlation, a simple linear regression analysis can be used to further examine the relationship between the two variables. In particular, does HHINCOME affect VALUEH? To test this question using regression, VALUEH is defined as the dependent variable and HHINCOME as the independent (predictor) variable. The regression analysis, as shown in the tables below, results in a constant coefficient of 1.301x10^5 and a regression coefficient for variable HHINCOME of 1.7. A linear model of the analysis can be written mathematically as

$$y = 1.301 \times 10^5 + 1.774X$$

Where *y* is the predicted variable VALUEH and *X* is the independent variable HHINCOME. However, the regression analysis also shows an R^2 value of 0.292 which suggests that only 29.2% of the variance can be explained by the independent variable despite the significance being less than 0.001. The scatterplot below shows the data with the linear model. A concentration of data points appears to fall below the linear model suggesting that the observed values tend to be lower than the predicted values meaning that the linear model may be overestimating the dependent variable.

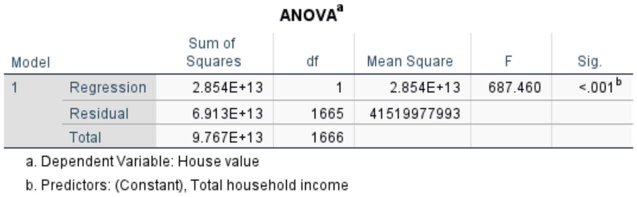


Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.541[a]	.292	.292	203764.516

a. Predictors: (Constant), Total household income

b. Dependent Variable: House value

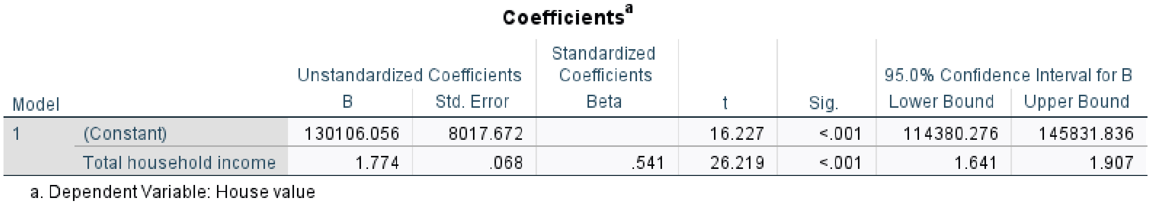


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.854E+13	1	2.854E+13	687.460	<.001[b]
	Residual	6.913E+13	1665	41519977993		
	Total	9.767E+13	1666			

a. Dependent Variable: House value

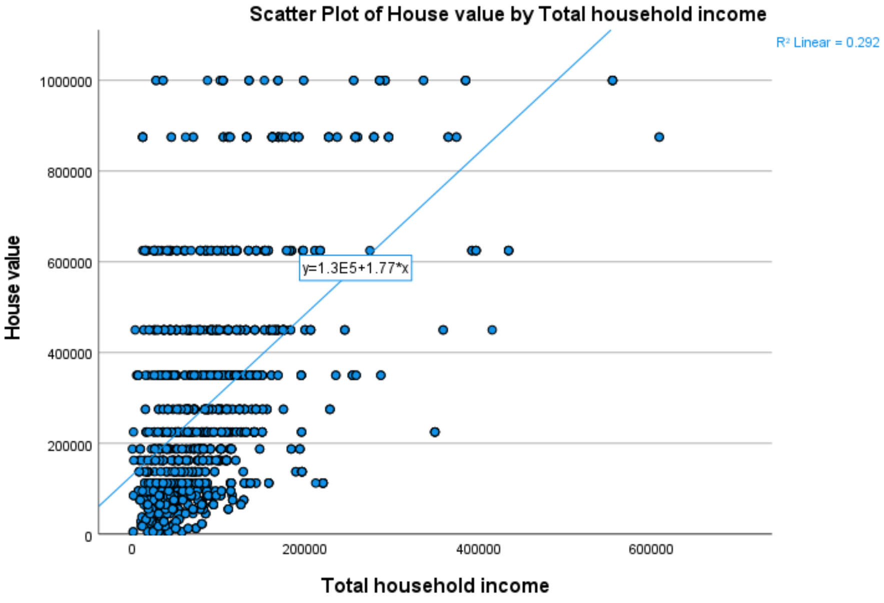
b. Predictors: (Constant), Total household income



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	130106.056	8017.672		16.227	<.001	114380.276	145831.836
	Total household income	1.774	.068	.541	26.219	<.001	1.641	1.907

a. Dependent Variable: House value



DISCUSSION

The linear model using only HHINCOME as the independent variable is not sufficient to predict the dependent variable VALUEH given that it can only explain 29.2% of the variance. To improve the predictive quality of the linear model, the linear model could either consider a different independent variable for a simple linear regression or multiple independent variables for a multilinear regression.

REFERENCES

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS, 2024

US Census Bureau. (2019, May 2). Our Censuses, U.S. Census Bureau Censuses. The United States Census Bureau. https://www.census.gov/programs-surveys/censuses.html