Geocoding Truck Stops Documentation

William Co

2025-08-27

This report documents the geocoding of U.S. truck stop data, addressing challenges from inconsistent address formats. Using phone number matching and structured data from Truck Stops and Services, Yelp, Yellow Pages, and iExit, we achieved a 99.19% match rate. A custom interface was developed to support manual verification and ensure data accuracy.

Table of contents

1 Data Cleaning

1.1 Shape and Column Analysis

We begin by checking the structure and columns of the dataset to ensure consistency. The data contains 47,776 rows and 12 columns, including variables such as weight, year, age, education, race, asset_total, asset_housing, debt_total, debt_housing, and wealth. For our analysis, we focus on the variables relevant to wealth and asset calculations, and note that sex and income are not used further.

Column Name	Type	Description
weight	float64	Survey weight
year	int64	Survey year
age	int64	Age of respondent
sex	object	Sex (not used in analysis)
education	object	Education level
race	object	Race/ethnicity
asset_total	float64	Total assets
asset_housing	float64	Housing assets
debt_total	float64	Total debts
debt_housing	float64	Housing debts
income	float64	Income (not used in analysis)
wealth	float64	Calculated wealth

1.2 Missing Values Check

A review of the dataset shows that there are no missing values in any column, so no imputation or removal of rows is necessary.

1.3 Data Types and Ranges

Below are the observed data types and value ranges:

Variable	Type	Min	Max
weight	float64	0.20	31,115.82
year	int64	1989	2016
age	int64	17	95
sex	object	2 unique	
education	object	3 unique	
race	object	4 unique	
asset_total	float64	-22,487,306.62	2,928,346,179.67
asset_housing	float64	0.00	182,642,128.63
debt_total	float64	0.00	293,486,997.64
debt_housing	float64	0.00	44,821,081.33
income	float64	0.00	351,958,858.31
wealth	float64	-221,985,489.24	2,929,687,834.52

1.4 Negative Values Check and Cleaning

We identify that asset_total contains 7 negative values, which is about 0.01% of the data. Since assets cannot logically be negative, we set all negative values in asset_total to zero. This adjustment ensures that all asset values are non-negative, as required by financial logic. After this cleaning step, asset_total has a minimum value of zero, and no negative values remain.

Variable	Negative Values	% of Total
weight	0	0.00%
asset_total	7 (before)	0.01%
asset_total	0 (after)	0.00%
asset_housing	0	0.00%
debt_total	0	0.00%
debt_housing	0	0.00%
income	0	0.00%

A table of the rows with negative asset_total values (before cleaning) is available in the appendix or supplementary materials.

A summary table of asset_total after cleaning:

Statistic	asset_total
Min	0
Max	2,928,346,179.67
Negative Values	0

1.5 Outlier Detection

We also check for outliers using the interquartile range (IQR) method. While some variables have a notable number of outliers, these are retained for analysis unless they are logically impossible (such as negative assets, which have already been addressed).

Variable	Outliers (N)	% of Total	Lower Bound	Upper Bound
weight	330	0.7%	-4,095	12,858
asset_total	8,281	17.3%	-2,215,818	3,831,215
asset_housing	5,405	11.3%	-651,383	1,085,639
debt_total	5,091	10.7%	-236,639	394,398
debt_housing	5,033	10.5%	-167,927	279,879
income	7,542	15.8%	-179,464	385,518

1.6 Categorical Distribution

Race	Count	%
white	37,044	77.5%
black	5,186	10.9%
Hispanic	3,553	7.4%
other	1,993	4.2%

Education	Count	%
college degree	19,444	40.7%
no college	17,820	37.3%
some college	10,512	22.0%

Sex	Count	%
male	37,212	77.9%
female	10,564	22.1%

1.7 Year Distribution