Bucking the Trend?

Distance from Starbucks as Measure of Population Density

I.Introduction

Commercial market forces and population trends often interact spatially in interesting ways, and the nature of these interactions is a clear topic of interest for contemporary statisticians, geographers, and economists. Knowledge of the exact nature of these relationships can be used to make predictions and estimations, and has the added benefit of allowing us to view ongoing trends from a unique perspective. This is the domain of GIS (geographic information system) science, but looking at these issues from a more statistical perspective can also be rewarding.

With this in mind, the purpose of this research will be to look at the relationship between population trends and the distribution of chain stores. In particular, the question of whether or not the distance from any one municipalities' city of town hall to the nearest Starbucks can be used to predict the population density of that municipality will be examined. Considering that population density plays a major role in urbanization and urban planning, the question has significant potential real-world impact and application.

II. Sampling Methodology

In order to look at the question of whether distance from the nearest Starbucks tells us anything about population density in the United States, a valid representative sample of U.S. municipalities (cities, towns, townships, boroughs, and villages) must be drawn as the first step. To make this sample as representative and least variable as possible, stratification by state was performed. Given that the number of Starbucks locations per unit area in some states may be significantly more or less than in other states due to economic or social considerations, this stratification is valid and in the aggregate likely reduced variability.

Each state was assigned proportional representation in the sample based upon its share of the total U.S. population; for example, California makes up 12.15% of the total U.S population according to the 2010 U.S. Census, so it was assigned 12 spaces in the sample. Then, on a state-by-state basis, the municipalities were numbered and a random number generator was used to randomly select the certain number of municipalities each state was allowed to have in the sample based on population proportions, as discussed previously. Repeats were ignored so that each municipality in the overall sample would be unique.

Once the 98 municipality sample was obtained (because of rounding, the sample did not add to exactly 100), 2010 census data was examined in order to find the population and land area (square miles) for each of the sampled municipalities. Next, the population density (the number of people per square mile) for each municipality was calculated by dividing the population by the land area.

Finally, the “Directions” functionality of Google Maps was used to find the distance, in miles, from each municipalities’ town hall, city hall, town office, or borough office to the nearest Starbucks coffeehouse. While using this functionality only “Driving” directions were enabled so as to remove results that did not rely on roads, like flying or walking.

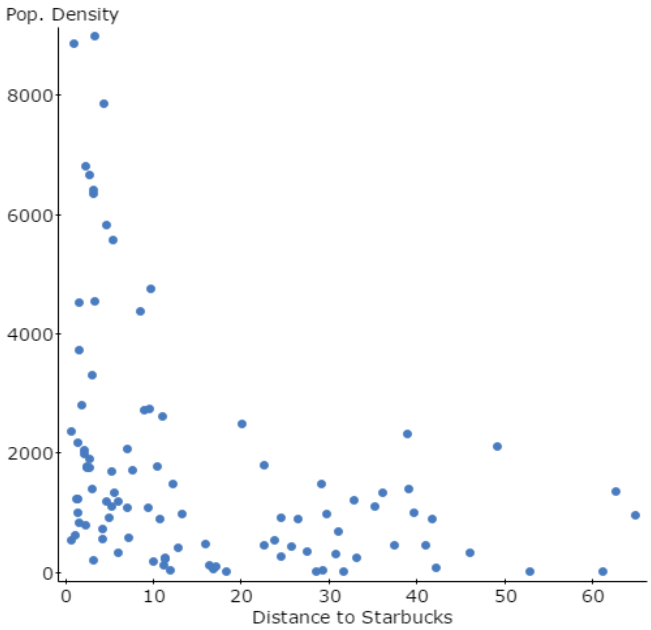

FIGURE 1A

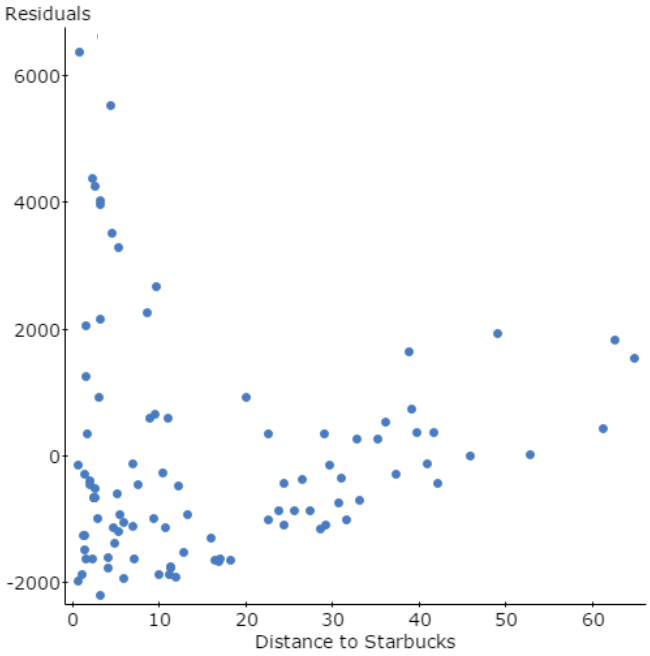

FIGURE 1B

III. Graphical Analysis

The goal of this research is to understand the relationship between the distance from a center of municipal government (town hall, city hall, town office, or borough office) to the nearest Starbucks (this variable will be abbreviated to "*Distance to Starbucks*") and the population density in number of people per square mile (this variable will be abbreviated to "*Pop. Density*"). *Distance to Starbucks* is the explanatory variable (independent variable) whereas *Pop. Density* is the response variable (dependent variable; what we are predicting).

Figure 1A, the scatter plot graph of *Pop. Density* (y) versus *Distance to Starbucks* (x), illustrates the relationship between the explanatory and response variable. However, a closer look at the scatter plot shows that there is reason to doubt the linearity of the relationship. Without approximate linearity, the Straight Enough Condition will fail and we will not be able to use the regression model or run regression slope tests or intervals. Examining the scatter plot of the residuals versus *Distance to Starbucks* after regression is run (Figure 1B) confirms this suspicion: the residual scatter plot appears to have a trend (clumping occurs in that the residuals are much closer together at smaller values of *Distance to Starbucks* than they are at larger values). This means that the Equal Variance Assumption is not met and we cannot proceed with regression inference. In order to address these problems, the data needs to be re-expressed, which we will do by taking the square root of the *Pop. Density* values. Figure 2A is the scatter plot of the re-expressed data, $(Pop.\ Density)^{1/2}$ versus *Distance to Starbucks*, and appears to be more linear than the original scatter plot. Examining the residuals also shows that the re-expressed data is better suited for our purposes: the residual scatter plot (Figure 2B) doesn't appear to show much of a pattern or trend, though caution is still necessary.

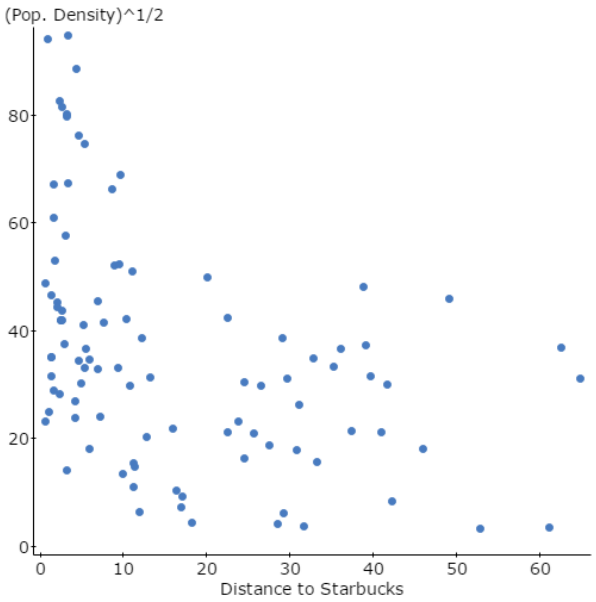

FIGURE 2A

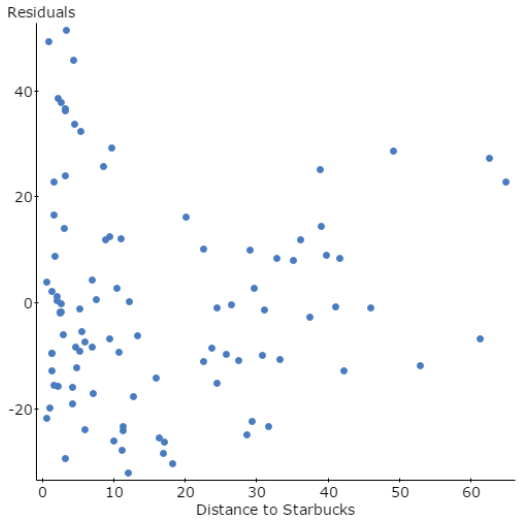

FIGURE 2B

IV. Checking the Conditions for Regression Inference

Before proceeding, the conditions for inference must be met:

1.Linearity Assumption: The scatter plot of $(Pop.\ Density)^{1/2}$ versus *Distance to Starbucks* (Figure 2A) looks to be approximately or roughly linear. This means the Straight Enough Condition is met, which in turn means that the Linearity Assumption is met.

2.Independence Assumption: The residual plot does not show much evidence of patterns, trends, or clumping, so the Random Residual Condition seems to be met. We should proceed cautiously, however, given that residuals at lower Distance to Starbuck values seems to be closer together than residuals at larger Distance to Starbuck values, which may be evidence of slight clumping. Despite this, we certainly do not have any reason to believe that residuals are not independent of each other: municipalities were randomly selected using valid methods, so the sampled municipalities probably aren't related, making their values for the variables independent.

3. Randomization: A stratified random sampling method was used, guaranteeing randomization.

4.Equal Variance Assumption: As discussed above, the residual scatter plot of the re-expressed data generally shows no shape, trend, or form. There is some evidence of thickening and/or clumping but not enough for us for us to fail to meet this condition. The variation is generally constant, so we'll proceed, albeit cautiously.

5.Normal Population Assumption: The histogram of the residuals for the re-expressed data (Figure 3) regression appears to be nearly Normal. Additionally, with a fairly large sample size, we are generally confident that the Central Limit Theorem means that this assumption is met.

Since the conditions for inference are met, a regression model can be used, as can regression slope t-tests and intervals.

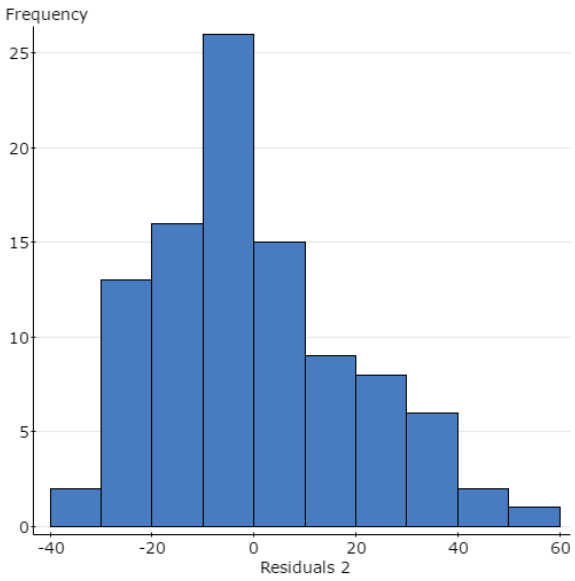

FIGURE 3

V.Hypothesis Test

Hypotheses:

H_0: $\boldsymbol{\beta}_1 = 0$
There is no relationship between *Distance to Starbucks* and *(Pop. Density)$^{1/2}$* **:** *Distance to Starbucks* and *(Pop. Density)$^{1/2}$* are independent**.**

H_a: $\boldsymbol{\beta}_1 \neq 0$
There is a relationship between *Distance to Starbucks* and *(Pop. Density)$^{1/2}$* : *Distance to Starbucks* and *(Pop. Density)$^{1/2}$* are associated.

$$t_{n-2} = t_{96} = \frac{b_1 - \boldsymbol{\beta}_1}{SE(b_1)} = \frac{-0.5713 - 0}{0.1246} = -4.5847$$

Figure 4 shows the scatter plot fitted with the LSRL. Table 1 shows the output for the t-test, including parameter estimates, r and R squared values, and standard errors.

Since the p-value of < 0.0001 is less than the significance level of $\alpha = 0.05$, the null will be rejected, meaning that there is sufficient evidence to conclude that *Distance to Starbucks* and *(Pop. Density)$^{1/2}$* are associated.

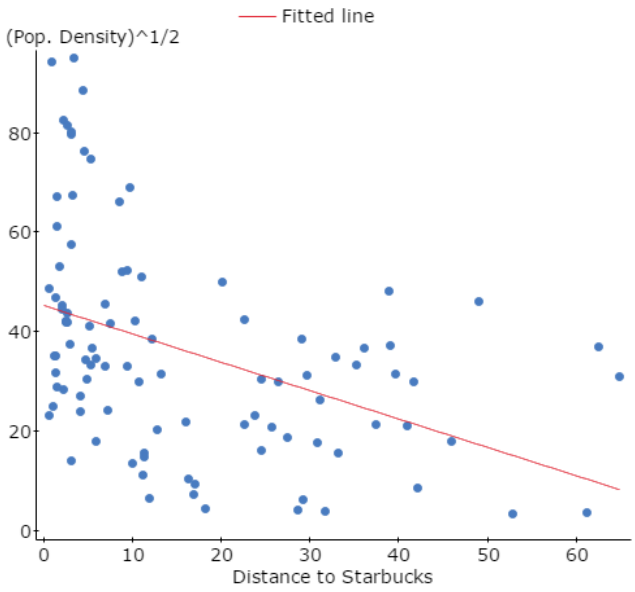

FIGURE 4

Simple linear regression results:
Dependent Variable: Y^1/2
Independent Variable: Distance to Starbucks
Y^1/2 = 45.273733 - 0.57128766 Distance to Starbucks
Sample size: 98
R (correlation coefficient) = -0.42382099
R-sq = 0.17962423
Estimate of error standard deviation: 19.70881

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	45.273733	2.8359561	≠ 0	96	15.964186	<0.0001
Slope	-0.57128766	0.12460721	≠ 0	96	-4.5847078	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	8164.7729	8164.7729	21.019545	<0.0001
Error	96	37289.969	388.43718		
Total	97	45454.742			

Residuals stored in new column: Residuals

TABLE 1

VI. Conclusion

The analysis shows that there is a relationship between *Distance to Starbucks* and $(Pop.\ Density)^{1/2}$. However, it's difficult to know exactly what the nature of this relationship is. It appears to be approximately linear, but we would want to draw at least several more samples of equal if not greater size and test them before reaching any definite conclusions.

However, we can definitely say that as *Distance to Starbucks* increases, $(Pop.\ Density)^{1/2}$ decreases. This makes logical sense; we'd expect smaller, more rural municipalities with few people per square mile to be farther away from a Starbucks because of the various lurking variables that we know impact the relationship. The first one of these would concern the market area of any individual Starbucks and the basic rules of supply and demand. If a Starbucks can't draw in enough customers from the area surrounding it, it can't turn a profit, and this makes places with less people per square miles less likely to have a nearby Starbucks. We also have to consider the differences in infrastructure between more densely populated, urban areas, and more rural areas. Densely populated areas almost always have better transportation infrastructure than less densely populated areas. This means that it costs more for Starbucks to open and ship products to a Starbucks in a rural, sparsely populated area than a urban, heavily populated area, again impacting the bottom line. Finally, we should consider how social and cultural distributions may affect where Starbucks are located. It's possible that the cultural atmosphere in more heavily populated, urbanized municipalities is more conducive to and supportive of coffee drinking and/or going to coffeehouses than the cultural atmosphere in rural areas.

In short, the distance from a center of municipal government to the nearest Starbucks does seem to tell us something about the population density of the area. The extent to which *Distance to Starbucks* is a valid estimator or predictor of $(Pop.\ Density)^{1/2}$ is difficult to estimate from just this one sample, but that doesn't mean that we should ignore the underlying relationship. We should also think about what the relationship means in a real-world context beyond prediction. There are various theories concerning how services and industries are spatial distributed in the social sciences, and our relationship could be interpreted in the context of any one of them with varying degrees of success. In particular, the relationship discussed here seems to fit well with Walter Christaller's Central Place Theory, which (to simplify) postulated that large cities tend to contain most of the services and "draw in" people from nearby settlements, which in turn contain some services that draw in people from more rural settlements, and so on. This means that services tend to be clustered in cities, something that supports our analysis, as we concluded that the services Starbucks provides tend to be much more easily available in densely populated places.

VII. Possible Sources of Error

1. Google Maps may not have every Starbucks in the United States registered in its system, which may have increased the *Distance to Starbucks* for some municipalities artificially.
2. The Google Maps algorithm generally finds the shortest route, but may sometimes fail to do so, which would have also artificially inflated the *Distance to Starbucks*.
3. Google Maps may not always give completely accurate mile readings.
4. The sample size could have been larger to reduce variability, perhaps eliminating our need to re-express the data and making the relationship more clearly linear.
5. Potentially there could be some exclusion of unincorporated communities, leading to undercoverage.
6. Using Google Maps to find the distance may have, in some cases, inflated *Distance to Starbucks* because Google Directions takes into account traffic, time of day, etc. to some degree when calculating optimal routes. This could lead to a faster route being picked over a shorter route, especially near or in large cities during peak traffic times.

VIII.Future Research

1. With more time and funds, we could run a large sample and could use the information Starbucks offers about its store locations to cross-reference with Google Maps, producing more accurate data
2. Other applications: We could compare the distributions of Starbucks across the United States to that of other chains stores, like McDonald's. We could see if the relationship we investigated here still holds true outside the United States, and, if it differs, in what way.

IX. Note on Sources

In order to draw the sample, I utilized U.S. census data from 2010. I used sites like www.factfinder.census.org and www.census.org to compile lists of municipalities by state. Then I followed the sampling methodology described in part II to draw my sample. Next, I used the same sites (www.factfinder.census.org and www.census.org) to find out the land area and population for each municipality. From these two quantitative variables I calculated the population density, my response variable, by dividing the population for each municipality by the land area for each municipality.

Name	Population	Land area	Pop. Density	Distance to Starbucks	(Pop. Density)^1/2
Calimesa, CA	7,879.00	14.85	530.57	0.70	23.03410515
Selma, CA	23,219.00	5.14	4,517.32	1.60	67.211011
Taft, CA	9,327.00	15.11	617.27	1.10	24.844919
Temple City, CA	35,558.00	4.01	8,867.33	0.90	94.16650147
Orange Cove, CA	9,078.00	1.91	4,752.88	9.70	68.94113431
Monterey Park, CA	60,269.00	7.67	7,857.76	4.40	88.64400713
Tustin, CA	75,540.00	11.08	6,817.69	2.30	82.56930422
Whittier, CA	85,331.00	14.65	5,824.64	4.60	76.31932914
Perris, CA	68,386.00	31.39	2,178.59	1.40	46.67536824
Indio, CA	76,036.00	29.18	2,605.76	11.10	51.04664534
Downey, CA	111,772.00	12.41	9,006.61	3.40	94.90316117
Piedmont, CA	10,667.00	1.68	6,349.40	3.20	79.68312243
Cisco TX	3,899.00	4.90	795.71	2.30	28.2083321
Bellevue TX	362.00	0.80	452.50	37.40	21.27204739
Rio Vista TX	873.00	0.80	1,091.25	9.40	33.03407332
New Berlin TX	511.00	4.80	107.00	16.40	10.34408043
Blossom TX	1,439.00	2.50	575.60	7.20	23.99166522
Beeville TX	12,863.00	6.10	2,108.69	49.10	45.92047474
Jourdanton TX	3,871.00	3.50	1,106.00	35.20	33.2565783
Nacogdoches TX	32,996.00	26.91	1,226.16	1.40	35.01656751
Hays TX	217.00	0.20	1,085.00	7.00	32.93933818
Eatonville, FL	2,243.00	1.10	2,039.09	2.10	45.15628417
Highland Beach, FL	3,640.00	1.10	3,309.09	3.10	57.52469035
Miami Gardens, FL	111,378.00	20.00	5,568.90	5.40	74.62506281
Mulberry, FL	3,817.00	3.20	1,192.81	6.00	34.53708152
Medley, FL	842.00	4.30	195.81	3.20	13.99321264
Fort Myers, FL	68,190.00	40.40	1,687.87	5.20	41.08369506
Poughkeepsie, NY	32,736.00	5.10	6,418.80	3.20	80.11741384
Solon, NY	1,079.00	29.70	37.00	29.30	6.08276253
Virgil, NY	2,401.00	47.30	51.00	16.90	7.141428429
North Salem, NY	5,104.00	21.40	238.50	11.30	15.44344521
Rose, NY	2,369.00	33.90	69.90	42.20	8.360621986
Stillwater, NY	7,522.00	41.40	181.90	10.00	13.48703081
Media, PA	5,327.00	0.80	6,658.75	2.70	81.60116421
Spring Grove, PA	2,167.00	0.80	2,708.80	8.90	52.04613338
New Bloomfield, PA	1,247.00	0.50	2,494.00	20.10	49.93996396
Hunker, PA	291.00	0.40	727.50	4.20	26.97220792
Annawan IL	878.00	1.98	444.43	41.00	21.08150848
Cortland, IL	4,270.00	3.63	1,176.30	4.70	34.29723021
Mason City, IL	2,343.00	1.01	2,319.80	38.90	48.16430213
Johnston City, IL	3,543.00	2.06	1,719.90	7.60	41.47167708

Aberdeen, OH	1,638.00	1.35	1,213.30	32.90	34.83245613
Lakemore, OH	3,068.00	1.48	2,073.00	7.00	45.53020975
Wayne Lakes, OH	718.00	0.53	1,354.70	62.60	36.80624947
Rayland, OH	417.00	0.47	887.20	10.80	29.78590271
Villa Rica, GA	13,956.00	14.20	982.82	13.30	31.34996013
Newborn, GA	696.00	1.60	435.00	25.70	20.85665361
Lexington, GA	239.00	0.50	478.00	16.00	21.86321111
Mint Hill, NC	23,341.00	21.20	1,100.99	5.30	33.18116936
Misenheimer, NC	728.00	1.62	449.38	22.60	21.19858486
Roxobel, NC	240.00	1.00	240.00	33.20	15.49193338
Readmond Township, MI	493.00	31.00	15.90	28.60	3.987480407
Freedom Township, MI	1,428.00	35.40	40.34	12.00	6.351377803
Masonville Township, MI	1,734.00	167.70	10.34	52.90	3.215587038
Hopewell, NJ	1,922.00	0.703	2,735.20	9.50	52.29913957
Woodland Township, NJ	1,788.00	94.56	18.90	18.30	4.347413024
Manville, NJ	10,344.00	2.36	4,382.00	8.60	66.19667665
Edinburg, VA	1,041.00	0.70	1,487.14	29.10	38.5634542
Chincoteague, VA	2,941.00	9.10	323.19	46.00	17.97748592
Buena Vista, VA	6,650.00	6.70	992.00	39.70	31.4960315
Medical Lake, WA	5,060.00	3.40	1,488.20	12.20	38.57719534
Shoreline, WA	53,007.00	11.67	4,542.20	3.30	67.39584557
Maricopa AZ	43,482.00	47.47	915.99	4.90	30.2653267
Litchfield Park, AZ	5,476.00	3.10	1,766.50	2.50	42.02975137
Stockbridge. MA	1,947.00	22.70	86.00	17.10	9.273618495
Dighton. MA	7,086.00	22.00	322.09	6.00	17.94686602
Sevierville, TN	16,490.00	19.90	829.00	1.60	28.7923601
Rodgersville, TN	4,420.00	3.30	1,339.39	36.10	36.59767752
Terre Haute, IN	60,785.00	34.54	1,759.80	2.50	41.9499702
Gentryville, IN	268.00	0.39	687.20	31.10	26.2144998
Lamar, MO	4,532.00	5.12	885.20	26.50	29.75231083
Maysville, MO	1,114.00	1.15	968.70	29.70	31.12394577
Ridgely, MD	1,639.00	1.78	920.80	24.50	30.34468652
Westminster, MD	18,590.00	6.63	2,803.90	1.80	52.95186493
Highland, WI	797.00	64.70	12.30	61.20	3.507135583
Irving, WI	602.00	43.90	13.70	31.70	3.701351105
Larkspur, CO	183.00	1.50	122.00	11.20	11.04536102
Haxtun, CO	481.00	0.50	960.00	64.80	30.98386677
Chaska, MN	23,770.00	16.97	1,400.70	3.00	37.42592684
International Falls, MN	6,424.00	6.42	1,000.60	1.40	31.63226201
Hickory Grove, SC	337.00	1.30	261.00	24.50	16.15549442
Irmo, SC	11,097.00	6.30	1,761.40	2.70	41.96903621
Cowarts, AL	1,546.00	7.20	214.70	11.40	14.65264481
Roseland, LA	1,123.00	2.10	534.76	23.80	23.12487838
West Point, KY	1,100.00	2.70	408.50	12.80	20.21138293
Veneta, OR	4,561.00	2.57	1,774.70	10.40	42.12718837
Lamont, OK	417.00	0.30	1,390.00	39.10	37.28270376

Plymouth, CT	12,243.00	21.70	564.19	4.20	23.75268406
Rockwell, IA	1,039.00	2.98	348.70	27.50	18.67351065
Santa Clara, UT	6003.00	4.90	1225.10	1.30	35.00142854
Bentonia, MS	440.00	1.40	314.29	30.80	17.72822608
Pocahontas, AR	6608.00	7.40	892.97	41.70	29.88260363
Fallon, NV	8606.00	3.63	2370.80	0.70	48.69086157
Salina, KS	47707.00	25.11	1899.92	2.70	43.58807176
Ruidoso Downs, NM	2815.00	2.10	1340.48	5.50	36.61256615
Gering, NE	8500.00	4.30	1976.74	2.10	44.46054431
Star City, WV	1825.00	0.49	3724.49	1.60	61.02859985
Parma. ID	1983.00	1.10	1802.73	22.60	42.45856804
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0
					0