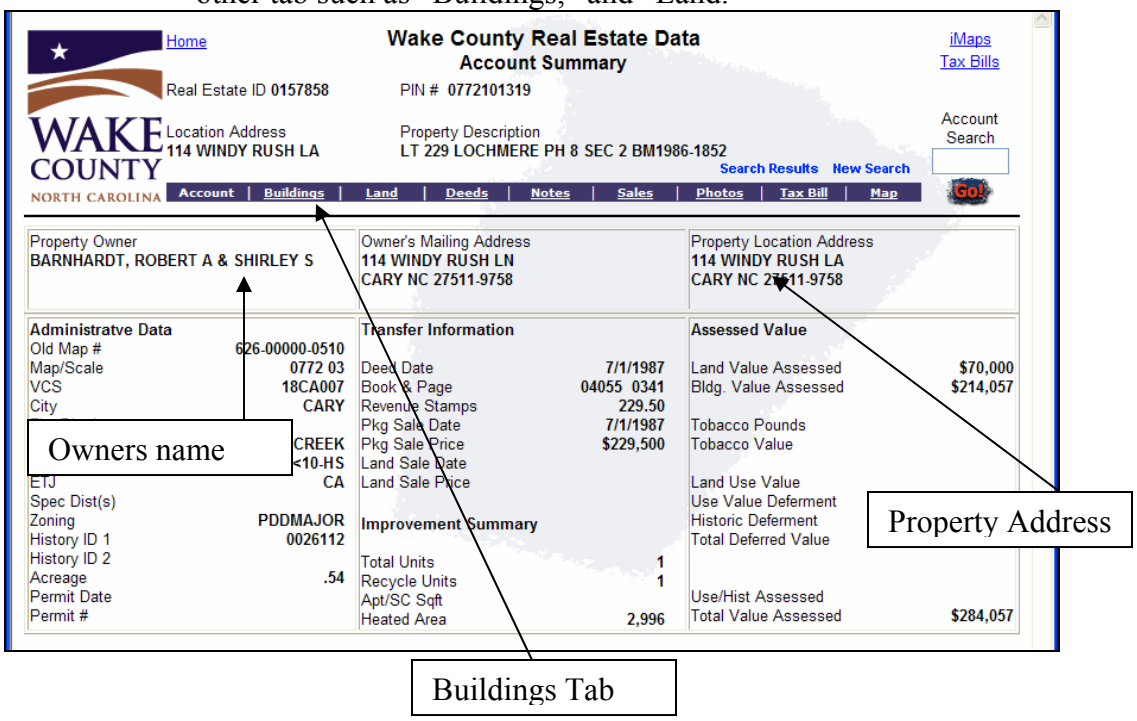
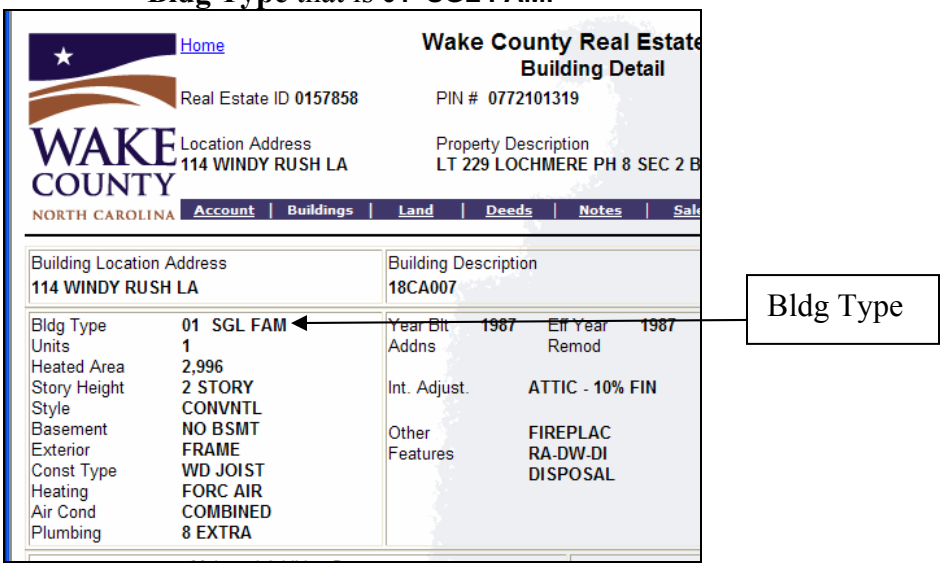
Due Date: _____
Assignment Goal: Familiarize students with data analysis and give them experience determining the margin of error.
Target Audience: Your report should be written so that college freshmen who have not taken a statistics course will understand it.

The purpose of this course is to understand how statistical inference works. In this assignment, you will see how the process of inference works by taking a random sample of single family homes in Wake County, North Carolina. Rather than going from house to house in the county, you will make use of an online database that lists key characteristics of these homes. Specifically you will:

1. **Open the website http://msweb01.co.wake.nc.us/realestate/search.asp .**
    a. This website allows the user to search on real estate ID number as well many other indicators. For this assignment you will search based on a randomly generated Real Estate ID (REID). To get to know the site an example parcel will be used to see how the search works. Enter ID 0157858 into the appropriate box on the search page to see an example. The resulting screen should appear as below. This page provides us some important information. It specifies the name of the owner (in this case Robert and Shirley Barnhardt) and the address of the property. It also provides links to additional information about the property and links to other tab such as "Buildings," and "Land."



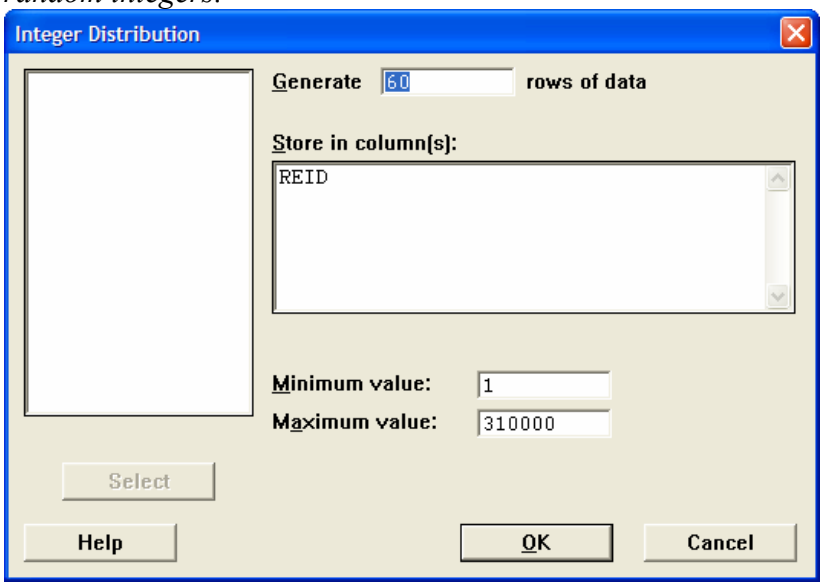
b. In this survey only single family residential properties are of interest. You can determine if the property is a single family residence by looking at the "Buildings" tab. You should look up the properties in the order they are listed on your sample list but only record information on properties with **Bldg Type** that is **01 SGL FAM.**



2. **Select your random sample of 30 single family homes.** Although you can calculate summary statistics and margins of error for your data by hand it is much more efficient to use a statistical software package. To achieve these calculations you can use a program called Minitab. If you are familiar with spreadsheets you will find Minitab is similar and very easy to use.

   a. The first task that Minitab will assist with is randomly selecting our sample. There are approximately 310,000 properties in the Wake County database. Each has a unique REID numbered from 1 to 310000. If we want to take an unbiased sample we can simply select REIDs at random. Minitab can generate random integers buy going to **Calc>Random Data> Integer…**

The dialog window that opens requires the number of random numbers, name of the column, and the range of the values to select from. You will need a random sample of 30 homes. Unfortunately, since this list will be randomly selected, it includes both residential and commercial (stores, warehouses, empty lots etc.) properties. Therefore we will need more than 30 random numbers to account for those which may need to be discarded. *Generate 60 random integers.*

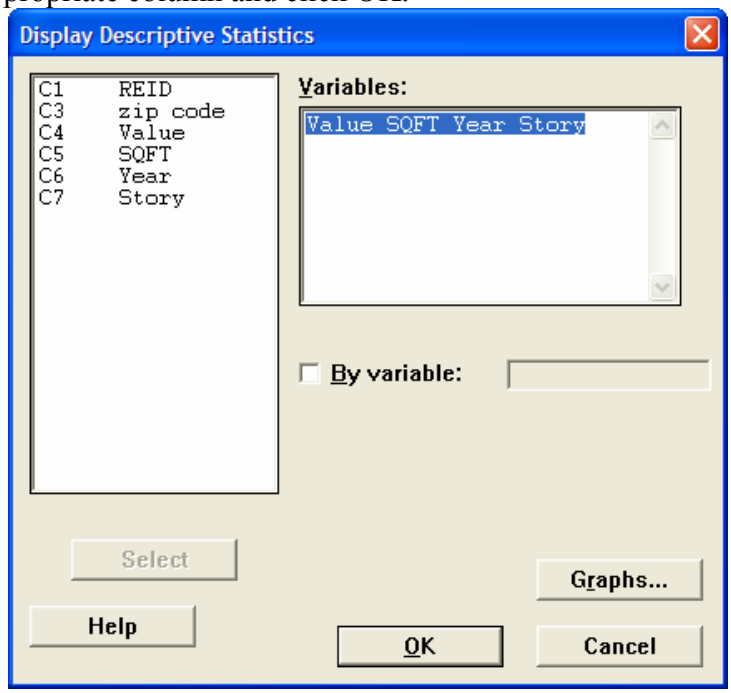


3. **Record the data for your sample.** From the website you can record the data of interest into the Minitab worksheet. Each REID should be put on its own row of the worksheet and each variable has its own column. Most of the data you will need about each home is on the Buildings tab, although you will also need the address and zip code from the Account tab. Record all the variables listed for each of the homes (remember you need only 30 homes.) The variables to be collected are given below along with their abbreviations:
    a. Address (Address) – record the street address
    b. Zip Code (ZIP) – record on the first 5 digits.
    c. Total Value Assessed (Value) – record the amount
    d. Heated Area (SQFT) – record the number
    e. Year built (Year) – record the year
    f. Story height (Story) – record the number
    g. Basement (BSMT) – record yes or no
    h. Air Conditioning (AC) – record yes or no
    i. Wood burning fireplace (fire) – record yes or no

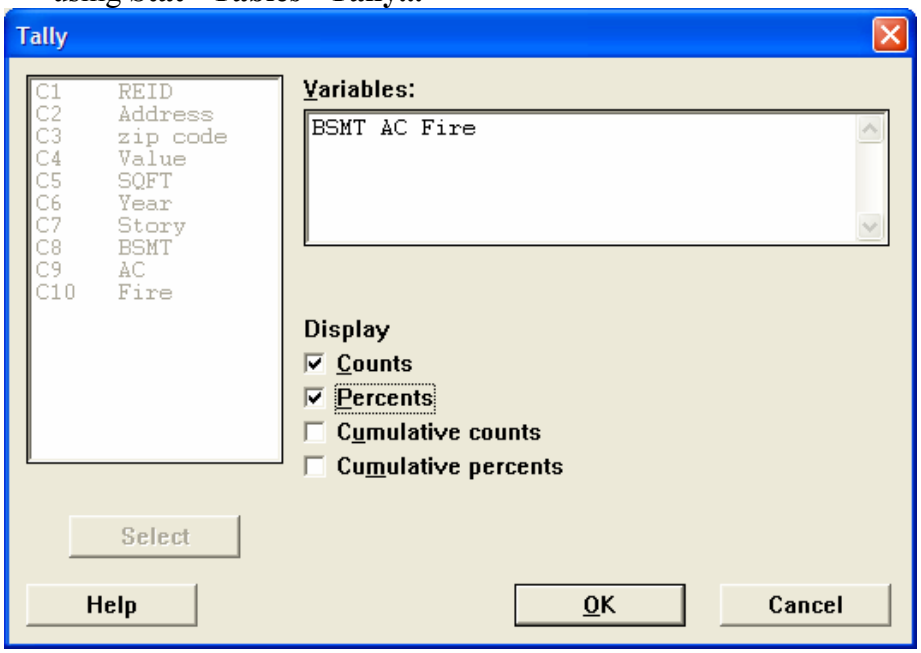
Do not enter data from real estate IDs that are not valid single family residences into your spreadsheet. Save your results often to insure that you do not loose them during the process.

4. **Submit your data.** When you have recorded data on all thirty properties, email the resulting spreadsheet file to the professor.

5. **Use the software to summarize your data.** Calculate summary statistics for your numeric variables (Total Value Assessed, Heated Area, Year built) by using **Stat> Basic Statistic> Display Descriptive Statistics**
Select the appropriate column and click OK.

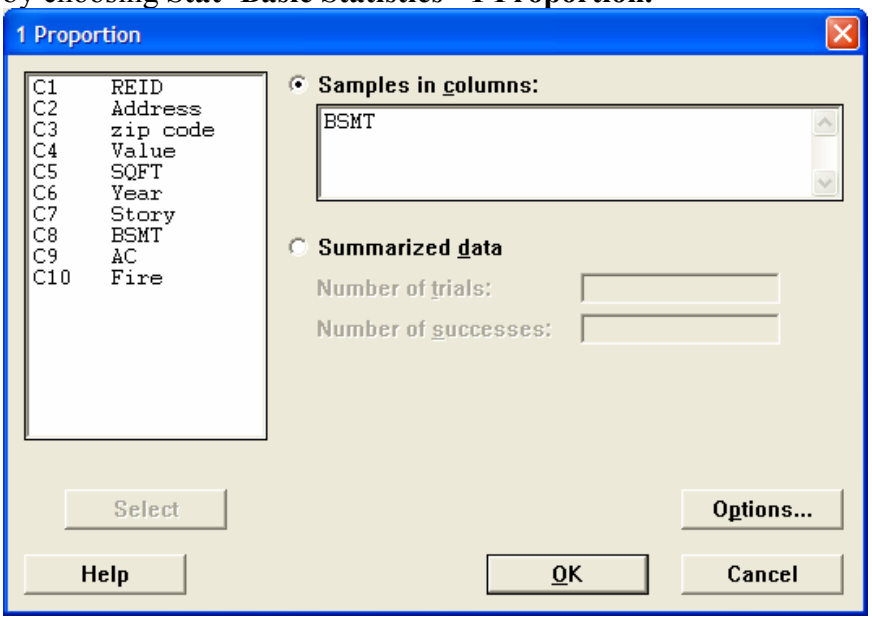


For categorical variables (Basement, AC and Fireplace) calculate the frequencies using **Stat> Tables> Tally...**

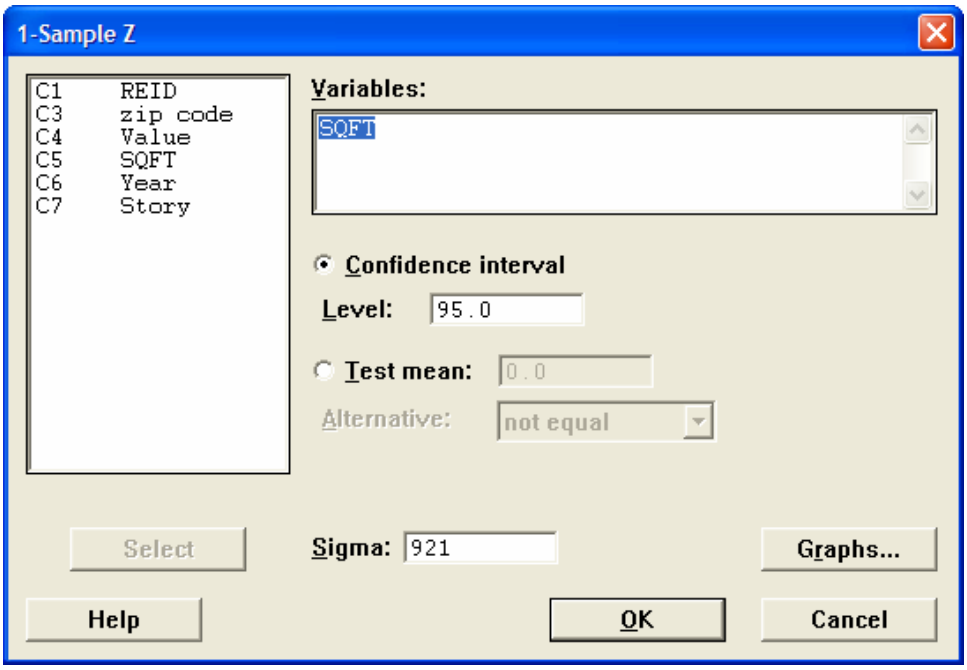


6. **Check your data entry.** Review the summary statistics you produced and insure that the values make sense. For instance, if you have a single family home that has 15 stories you may have made a data entry error, or if your yes/no variables have more than two levels you may need to correct them.

7. **Calculate confidence intervals.** As with summary statistics, calculating confidence intervals will differ based on the type of data (categorical vs. numeric).
    a. For categorical data we can create a confidence interval for a proportion by choosing **Stat>Basic Statistics> 1 Proportion.**



    b. For numeric data you can choose **Stat>Basic Statistics> 1-Sample Z.** For the confidence interval based on a Z statistic we will need an estimate of the population standard deviation. This can be found from the summary statistics you found previously (use the sample standard deviation).



8. Write a short (1 to 2 page) report that summarizes the features you have found and what they tell you about the properties in Wake County. Present the confidence intervals you have found and explain their meaning. Explain to the reader how the values of your confidence interval relate to the population of interest and the relevant parameters.

9. Be sure your name is clearly indicated on the first page of the paper. Staple or otherwise permanently attach (no paperclips, folded corners) your report, the data, and the confidence interval output together and have them ready to submit at the beginning of the class period on the due date. (Be sure you have emailed your dataset to the professor.)

Grading Criteria: This paper will be graded on the following criteria:

Professional presentation and clarity (10 points)

The work should be a coherent document that is self contained. It should make points and support those points with evidence from your calculations. It should not be series of unrelated sentences or a long series of numbers connected by words. It should be presented in a manner that shows respect for ones work. Common examples that would show a lack of professional presentation include misspelled words, grammatical errors, handwritten work, items added after the main body is printed, food stains, crumpled paper, and poor print quality.

Content and completeness (15 points)

The majority of the points from this assignment are associated with collection and summary of the data. Be sure to carefully and precisely perform this collection. Be sure to clearly explain the meaning of the confidence intervals you have created in language that a novice will understand.

Hints and Tips

1. Be sure you have someone proofread your paper.

2. Your paper should be a single coherent document, not a series of numbers tied together with words.

3. You paper should be written in plain language that is readily understood by those who have not taken a statistics course.

4. Remember the rules on academic misconduct. Each student must write and run their own program, print their own output, and write their own summary. The professor is very aggressive about these rules.

5. Don't needlessly repeat information that appears in the output. Use numbers when needed in your discussion but don't use them if you are not making a point about them.

Grading Rubric and Instructors Notes

Assignment Goals: Students often pass through statistics courses without a real appreciation of what statistics is about. This assignment can be used to show students that statistics are used in a real setting and have them to think more deeply about the material. Specific outcomes of this assignment are:

1. Exposure to the use of statistical software. This assignment can be given as a first assignment with data analysis software.

2. Expose students to the methods of data collection and data entry.

3. Help students understand the concept of confidence intervals by writing their interpretation.

The data from this assignment can also serve as the basis of future assignments.

Discussion: One advantage of this assignment is that it gives students the opportunity to see the concept of repeated sampling in action. Each student will have a unique sample but all samples are selected from the same population. Therefore, these samples will produce averages that have a predictable sampling distribution. Even with a class of only 30 students, the averages will have visible variability. By illustrating this to the students you will give them a very intuitive feel for the idea of a sampling distribution.

A very informative way of doing this would be to have each student list their values for the average market value of the homes along with their 95% confidence interval. You might create a table such as below and have each student write in their values

Name	**Average Market Value**	**Low end of confidence interval**	**High end of confidence interval**

A major advantage of this dataset is that the "true" parameter for this situation is actually available. Wake County publishes an average for the Market value of all properties listed as single family Residential properties. This can be found on the Wake County web site at http://www.wakegov.com/general/tax/interesting/default.htm.

Once each student has listed their values you can discuss the following:

1. What is the population of interest? What is (are) the samples?
2. What is the parameter of interest? What is the statistic?
3. Why in most real sampling situations will we never really know the parameter, but in this instance we actually have a pretty good idea about the parameter?
4. What would be meant by a sampling distribution in this setting?
5. Which confidence intervals actually contain the true parameter?
6. What is meant by 95% confidence? How is confidence illustrated by this situation?
7. Have the students critique their explanations of confidence (in their papers) based on the discussion of the previous question.

Scope and placement in the course: This assignment should be used after the students have been taught sampling distributions and concurrently with confidence intervals.

Note on uniqueness: Since this is a random sample most students should have a unique sample average. If students have the same exact values they may have "borrowed" from each other. Although it is always good to check their details before accusing them of anything, remember these are pseudo random numbers and occasionally strange things happen.

Time to complete: Students will need to access computers and find data on the internet. Students should be given at least a week to find time to fit this into their schedules. Experience has shown that 2 weeks is ideal, but consideration of other things going on in the course (discussion of the relevant topics) should heavily influence the due date.

Options: Some options to consider:

1. Using software that is available in your particular lab.

2. Reduce the number of variables they collect; you can actually make this much simpler if the students only collect market value. Note that by doing this you loose some flexibility in using the data later in other settings.

3. Ask the students to collect one variable that requires them to look at the photo of the house that is in the database (for example you might want the color of the house). This will reinforce the reality of this assignment.

4. Pick another major metropolitan area that would be more relevant to your students. Most areas now have databases such as this, but not all are as detailed or have a simple index like the REID.

Grading Rubric

Student:_____

1. Presentation and clarity.
    a. Student provided the completed data sheet via email. (3 points)___

    b. Paper is presented without distracting grammatical, spelling or structural flaws. (5 points) ___

    c. Other issues of presentation. (2 points)___

2. Content and completeness.
    a. Student used appropriate data set from the website. (1 point)___

    b. Student produced confidence intervals for each variable. (6 points)___
    (Total Value Assessed, Heated Area, Year built, Basement, AC and Fire.)

    c. Student summarized the features of homes in Wake County. (2 points)___

    d. Student explained confidence intervals clearly, pointing out the population of interest and the interpretation of the parameter. (4 points) ___

    e. Paper is free of obvious errors in statistical terminology and content. (2 points) ___

Total____

Example Data

REID	Address	zip code	Value	SQFT	Year	Story	BSMT	AC	Fire
140438	5412 Round Hill LN	27616	111472.00	1272.00	1985	1	no	yes	Yes
59087	102 Cliffview Dr.	27545	125750.00	1916.00	1973	1	yes	yes	Yes
98191	10837 Wilmore Dr	27614	190870.00	1968.00	1977	1	no	yes	Yes
235821	4300 Triland Wy	27539	211939.00	2201.00	1998	1	no	yes	Yes
70038	2013 Barwell Rd	27601	107639.00	1344.00	1986	1	no	yes	Yes
294052	913 Glenmacie Dr	27526	154546.00	1854.00	2002	2	no	yes	Yes
186904	7915 Hogan Dr	27587	291907.00	3210.00	1994	2	no	yes	Yes
82516	2333 Zebulon Rd	27597	404399.00	4354.00	1886	1	no	yes	Yes
6735	1216 Gunnison Pl	27609	205459.00	2673.00	1962	2	no	yes	No
17380	429 N First AV	27545	81229.00	1446.00	1944	1	no	yes	Yes
255119	1005 Burnt Hickory Ct	27502	181097.00	2092.00	2000	2	no	yes	Yes
294888	2364 Whistling Straits Wy	27520	52535.00	704.00	2002	2	no	yes	No
174941	6101 Battleview Dr	27613	291545.00	2783.00	1991	2	no	yes	Yes
186937	103 Perry Farms Dr	27502	123066.00	1500.00	1992	2	no	yes	Yes
135069	1608 Quiet Oaks Rd	27545	60216.00	1512.00	1985	1	no	yes	Yes
277645	8413 Fawncrest Dr	27603	153475.00	1711.00	2002	2	no	yes	Yes
234383	5452 Grand Traverse Dr	27604	131308.00	1408.00	1998	2	no	yes	Yes
44980	504 Center St	27502	131164.00	1680.00	1981	1	no	yes	Yes
267714	1601 Burgess Hill Ct	27539	281079.00	2592.00	2000	2	no	yes	Yes
140652	307 Trappers Sack Rd	27513	137142.00	1884.00	1985	2	no	yes	Yes
289366	10821 Round Brook Cir	27617	453519.00	3495.00	2003	2	no	yes	yes
21905	3801 Lassiter Mill Rd	27609	251941.00	2080.00	1981	1	no	yes	yes
78052	2836 Claremont Rd	27608	159312.00	990.00	1952	1	no	yes	no
203621	12005 Cadnire Ct	27613	182217.00	2422.00	1995	2	no	yes	yes
228641	5741 Presentation St	27545	91020.00	1026.00	1997	1	no	yes	no
165375	106 Timber View La	27511	212483.00	2462.00	1988	2	no	yes	yes
157838	102 Windy Point La	27511	370715.00	4576.00	1990	2	no	yes	yes
294928	5426 Patuxent Dr	27616	94565.00	1041.00	2002	1	no	yes	yes
200529	106 Starden Brook Ct	27519	192527.00	2450.00	1994	2	no	yes	yes