Projected Foreclosures in the Oklahoma County Region

Justin Wolzen (A11546831)

Bill Bryan (A20343350)

Robert Cook (A10073520)

Gandhali Munshi (A20280651)

**Table of Contents**

Summary 3

Statement of Scope 4

*Unit of Analysis* 4

*Variables* 4

Project Schedule 5

*Gantt Chart* 5

Team Member Assignments 6

Data Access 7

*Sources*  7

Data Cleaning 9

Data Transformation 11

Data Reduction 12

Data Consolidation 12

Data Dictionary (Preliminary) 13

**Summary**

Our team will be presenting a list of properties in the State of Oklahoma that expected to be listed for sale on the market soon. Property investors usually must invest in properties that are currently listed on the market which can make their investment process quick, since properties are typically sold shortly after being posted. We are presenting data that would allow for property investors to view properties that have a high chance of being listed for sale on the market. This would give the property investors an idea and more time to decide if a specific property is worth their investment. This project will benefit property investors and their time as well as sellers, as their property could be sold the same day it was listed.

**Statement of Scope**

The scope of this project is to scrape data from multiple public websites about information on real estate properties that are going to be foreclosed soon. After scraping the data, the data is transformed and analyzed to determine properties that will be listed for sale.

**Unit of Analysis**

We are measuring the cost of a foreclosed home that is expected to be listed on the market for sale in the future.

**Variables**

Our team used the address as the primary key when it came to extracting the data from the sources. Next, we provided the city, state, and zip code for where the property was located at. All our data was in the Oklahoma County area. The appraisal value was then listed along with the total square footage of the property. The next set was the bedroom count, bathroom count, and the year that the property was built. If there were any renovations listed, then we included them in this next column about the renovation year. The property was listed for sale, then we provided the number of days on the market it was, the last sale date, and the any other sales dates. We then provided the sale type which included an auction, Sheriff auction, bank owned, and others. The longitude and latitude were extracted to give a more accurate location for the property.

**Project Schedule**

We will complete this project in approximately 18 weeks during the semester. The group had meetings scheduled every Tuesday night at 7:30 PM to discuss tasks and update the team. Our team included a member who was in another time zone from the other 3 members and would adjust their schedule to meet weekly. The group met on occasion throughout the week before assignments were due for the project. Every member of the team was assigned a task to complete and update the group with the findings during the next meeting. Our team used Microsoft Teams to communicate with one another, and to meet through video conferencing.

**Gantt Chart**

Graphical user interface, application

Description automatically generated

**Team Member Assignments**

|  |  |
| --- | --- |
| Team Member: | Task: |
| Robert Cook  (Developer) | * Supervise Code * Data Scraping * Storing Data |
| Bill Bryan  (Project Manager/Developer) | * Data Scraping * Data Transformation |
| Justin Wolzen  (Analyst) | * Project Documentation * Project Requirements |
| Gandhali Munshi  (Analyst/Developer) | * Data Scraping * Data Cleaning |

**Data Access**

Our team used multiple public websites to extract data from for foreclosed homes. These websites are listed below. Our team focused on properties that were in Oklahoma County for this project.

|  |  |
| --- | --- |
| **Sources:** | |
| Oklahoma County Assessor | https://assessor.oklahomacounty.org/203/Search-Property-Records |
| OCSO Sales | https://docs.oklahomacounty.org/sheriff/SheriffSales/ |
| Trulia | https://www.trulia.com/ |
| Zillow | https://www.zillow.com/ |
| Realtor.com | https://www.realtor.com/ |
| Spot Crime | https://spotcrime.com/ |
| US Census API | Welcohttps://geocoding.geo.census.gov/geocoderme to Geocoder (census.gov) |

The Oklahoma County Assessor’s website was used to access additional data on the property that would be listed. This data included the amount of time that the property was owned based off the tax records listed on the site. Accessing this data will give the future investor an idea of how long the home has been owned, and if it has been owned by multiple persons in the past.

The Oklahoma County Sheriff’s Sales website was used to access the future properties that would be listed in the future in Oklahoma County. This site was updated weekly with the following weeks properties listed. These consisted of foreclosed properties that were to be auctioned the next week. This is where the heart of the data was extracted based off the fact that this was where the team discovered if a home was to be listed. The remaining sites were used to add data to the property to give the potential investor more information regarding the listed property.

Trulia, Realtor.com, and Zillow were two websites that were used to extract additional foreclosed properties that were listed in Oklahoma County. These three websites provided more data on the property than the public Oklahoma County websites that listed foreclosed properties. Our team used these sites to extract the foreclosed properties to give the potential investors more properties that may not be auctioned off the following week. All sites list foreclosed properties that are not always auctioned in a timely manner, but they are available for property investors.

The remaining website that was utilized was Spot Crime. The data from this site was about the crime in the area. This data could be used to give the potential investor an idea of the area the property is located in. Investors could use this data to remain away or choose to invest in areas where crime was high or low.

**Data Cleaning**

Our team extracted data on the properties from the Oklahoma County Sheriff’s Sales site on the following weeks foreclosed properties that were to be auctioned off. This information was not always complete or accurate. Realtor, Zillow, and Trulia were all utilized to assist with making this data more accurate. The address was used to link the properties up from the Oklahoma County Sheriff’s Sale site to Realtor, Zillow, and Trulia. This data included details like the number of bedrooms in the residence, the number of bathrooms, the square footage of the property, and other identifiers of the property. This data would be used by the investor to determine if the cost of the home was worth their investment, based on the size and details of the property.

Our team utilized the Oklahoma County Assessor’s website to extract the appraisal and compared this value to the estimated appraisal value listed on Realtor, Zillow, and Trulia. This would allow for the most accurate possible price of the home. This would be the most important piece of data to the potential investor as investors would only investigate properties that could afford. Combining the estimated appraisal values would give for a more accurate potential listing value or auction price when the investor goes to attempt to invest in the property. This would allow the investor to have a budget before the auction occurs on foreclosed properties.

Our team used Realtor, Zillow, and Trulia to pull the amount of time that the property has been listed as foreclosed. This data would reference with the Oklahoma Sheriff’s Sales listed properties to disclose the amount of time that the property has been listed as for sale. Some of the properties on the Sheriff’s sales website are not listed as foreclosed properties since they are going to be auctioned. This data would be used by the investor to determine if a property was worth the price that it was listed at based off the amount of time the property was listed on the market. It is another tool that an investor could use to determine if they should invest in a property.

Our team is using the US Census API to pull the location of the property. We will be using a geospatial API to grab the latitude and longitude for the location of the property. This will allow for property investors to have more of an idea about where their property is located that may be researching.

All these properties were matched together using the primary address key. This allowed for the team to gather data on the property size, price history, crime in the area, and others. Not every source had all the material required as different columns would come from another source than the details of the property. Missing data was expected with the data scraping as not every source would contain all the required columns. Some sources would provide the same data but be more thorough. The team analyzed which source was more thorough and elected to choose the data from that source. Missing values were kept as null as not having the value would not be a make or break for a potential investor. Addresses were not to be null as that was the primary key for the data, and the objective was to provide the investors with information on the home with the address listed.

**Data Transformation**

Most of the data was extracted was in a state that could be used to display to investors as it was. The data was retrieved from sources that are already in use by potential investors and home buyers. The data was scraped from these sources and split into a final which displayed the data from each source. After analyzing the data in the raw files, it was determined which data would be retained and which source would be used for each column. A final file containing the data that would be used was then created. The data dictionary listed at the bottom of the document displays the data types for each column of data that was extracted.

**Data Reduction**

This portion of this project consisted of reducing the data that was duplicated from the multiple sources that it was extracted from. This process included retaining the data that was going to primary and used for a column when it was duplicated by another data source. An example of this was the address for each source that was scraped. Almost all the data files extracted from each source had the address column listed as that is how all the data was linked together. This column was reduced to retaining the data from the source that provided the information.

**Data Consolidation**

For the web scraping portion, our team used Python. All the code used to extract the data from each source is provided with this deliverable. All the datafiles that were created from each set of code are also provided. The final datafiles with the finished data is provided in this deliverable.

**Data Dictionary (Preliminary)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field: | Description: | Data Type: | Required: | Source: |
| Address | The street address of the property | char(250) | Y | Trulia  Realtor.com  County Assessor |
| City | The city the property is in | char(50) | N | Trulia  Realtor.com  County Assessor |
| State | The state the property is located | char(2) | N | Trulia  Realtor.com  County Assessor |
| Zip | The zip code the property is in | char(5) | N | Trulia  Realtor.com  County Assessor |
| Value | The last published appraised value in dollars for the property | decimal(10,2) | Y | Trulia  Realtor.com  County Assessor |
| Sqft | The square footage of the property | integer | Y | Trulia  Realtor.com  County Assessor |
| Beds | The number of bedrooms | integer | Y | Trulia  Realtor.com  County Assessor |
| Baths | The number of bathrooms, including fractional | decimal(2,1) | Y | Trulia  Realtor.com  County Assessor |
| BuildYear | The year the property was originally built | integer | Y | Trulia  Realtor.com  County Assessor |
| DataLastRemodel | The date the property last had major remodel | datetime | N | Trulia  Realtor.com  County Assessor |
| DaysOnMarket | Number of days the property has been on the market | integer | N | Trulia  Realtor.com  County Assessor |
| LastSaleDate | Date of the last time the property was sold | datetime | N | Trulia  Realtor.com  County Assessor |
| NextSaleDate | Date the property is due to be sold | datetime | N | Trulia  Realtor.com  County Assessor |
| SourceType | The type of sale of the property | char(30) | Y |  |
| Latitude | Latitude of the property address | decimal(10,2) | N | US Census API |
| Longitude | Longitude of the property address | decimal(10,2) | N | US Census API |