Denton County Housing Values

# Background:

Every Year Denton County in the Dallas Fort worth metroplex appraises values of all the houses within its jurisdiction. A homeowner needs to pay property taxes which are 2.26% of the appraised value of his/her house.

*Example: If my house is appraised at $300,000 then I will pay $300,000\*2.26% = $6,780 in annual property taxes.*

# Problem Statement:

If the appraised value of a house increases by $20,000 from 2018 to 2019 then the homeowner will have to pay an additional $450 in taxes. This is extra money that he may not have planned for in his budget.

Homeowners have the option to either accept the increase in the appraised value of their house or file an appeal. In order to file an appeal and protest, they need to convince the appraiser on why the value of their house should be lower. The best way to do this would be to have data on similar properties in the neighborhood that have been appraised at a lower value. This can help them save money on their property taxes.

Most of the homeowners don’t have this data and they either accept the increase or reach out to their real estate agent for help. The real estate agent may not be able to help the homeowners in a lot of cases if there is not enough data on sales in the neighborhood for the previous 12 months.

The appraisal department would normally look at all the houses in the neighborhood & sometimes don’t consider factors such as age of the property or the location of the home within the neighborhood.

# Target Customers:

Home owners whose appraised value has increased in 2019 and they think that their appraised value should be lower than what the county says. They need supporting data on the value of similar homes nearby to argue their case with the county.

Target customers can also be home owners who want a tool to predict what their home value would be in 2019 so they can plan for it ahead.

**Project Phase 2** - Use Four Square API to build a tool that can give prospective homes to buyers based on criteria such as: living area, total land, nearby venues, prices, distance to schools, school ratings etc. For this capstone we will not add school ratings database to the list.

# Data Sources:

# Denton County open dataset “County Parcels”

# *Link:* [*http://data1-dentontxgis.opendata.arcgis.com/datasets/county-parcels/data*](http://data1-dentontxgis.opendata.arcgis.com/datasets/county-parcels/data)

# Fields selected for our study: land area, living area, address, city, state, zip, price, latitude and longitude, year built.

# Four Square API data: Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers.

# Note: We will be using only the 1st 10,000 records from the Denton County database due to the time it takes to parse through the database. 2 minutes per 10k records.

# Data Wrangling / Pre Processing:

# Data downloaded from the Denton County website in json format. Time taken to convert data from json to a pandas dataframe was approximately 1 minute per 5000 records. For this reason we took only the 1st 10,000 rows as sample data for our analysis.

# If we had time or a faster way to convert the data, this analysis could have been done on the entire dataset and would have been a more robust analysis.

# There were records where we had only land and no living area constructed. Those records were removed since we did not need them for our analysis. Also since we are only interested in Single Family Residential houses, all other records such as Commercial properties, rural land and multifamily homes were removed from the analysis.

# This left us with a total of 8629 records.

# Converted attributes such as land area, living area, year built, price etc. to numeric data types since we will be running regression models on these.

# Feature Engineering:

# New feature Age added to the dataset (2019 – year built).

# Sample of the table after all the changes:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sub divison** | **appraised** | **type** | **city** | **land** | **Subdivison\_2** | **area** | **year** | **zip** | **Latitude** | **Longitude** |
| MORNINGSIDE 1 | 168499 | Real, Residential, Single Family | CARROLLTON | 7548 | MORNINGSIDE 1 BLK 2 LOT 15 | 1684 | 1980 | 75006 | 32.9894 | -96.8705 |
| RIVER OAKS ADDN PH 3 | 182000 | Real, Residential, Single Family | CROSSROADS | 7607 | RIVER OAKS ADDN PH 3 BLK F LOT 11 | 3063 | 2001 | 76227 | 33.1434 | -97.1156 |
| O T JUSTIN | 19500 | Real, Vacant, Commercial Lot | FORT WORTH | 6500 | O T JUSTIN BLK 28 LOT 3 & 10' OF ALLEY | 0 | 0 | 76117 | 33.0866 | -97.2912 |

# Scatter Plots:

# Created scatter plots to look for correlation between price and attributes such as Living Area, Land Area and Age of the house.

* Strong positive relationship between Living Area and Price/value of house.
* Weak positive relationship between Land Area and Price. This could be due to the fact that most houses in the Sub division have similar land area.
* Age of House does not seem to have a correlation with Price

# Sample Sub division for the model:

# I selected Lewisville Valley subdivision as a sample to build the model since it has the most records: 559. Due to the small sample size we did not split the data into training and test sets.

# Predictive Modeling/Machine Learning:

# Multiple Linear Regression:

# I first created a Multiple Regression model with Price as the response variable and age, land area and living area as the independent variables.

# Output of the MLR Model:

# R square = 61%. I then decided to run KNN Regression to see if we can get a better result.

# *Next steps which have not been explored in this project for making the Multiple regression model more robust:* explore residual charts to look for non-linear relationship. In case of non-linear relationship we could have explored polynomial regression also.

# KNN Regression:

# Ran KNN Regression for k = 1 to 10.

# Based on the elbow in the chart we picked 3 as the number of neighbors for the model.

# R square for KNN Regression came at 79% which should be a good model.

# Prediction scenario for the model:

# If a homeowner in the Lewisville Valley area has a 40 year old house with 1800 sqft. of living space

# and 7000 sqft. of total land area, they can use the tool to predict the value of their house.

# Prediction:

# KNN gives a prediction of 156,800 and MLR gives a prediction of 147,600.

# 

# Analyzing the closest neighbors both the models are doing a fair job in predicting the price but KNN

# seems to be doing better. Below is a table of homes similar to our prediction scenario.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Subdivison** | **appraised** | **Land Area** | **Living Area** | **zip** | **Age** |
| LEWISVILLE VALLEY 3 | **$124,478** | 7130 | 1800 | 75077 | 44 |
| LEWISVILLE VALLEY 4 | **$157,265** | 7080 | 1815 | 75067 | 39 |
| LEWISVILLE VALLEY 3 | **$150,549** | 7130 | 1809 | 75077 | 47 |
| LEWISVILLE VALLEY 3 | **$148,596** | 7130 | 1789 | 75077 | 44 |
| LEWISVILLE VALLEY 3 | **$148,975** | 7015 | 1808 | 75077 | 44 |
| LEWISVILLE VALLEY 3 | **$135,905** | 7130 | 1784 | 55364 | 47 |
| LEWISVILLE VALLEY 3 | **$141,200** | 7130 | 1810 | 75077 | 44 |
| LEWISVILLE VALLEY 3 | **$140,900** | 7130 | 1797 | 75077 | 46 |
| LEWISVILLE VALLEY 3 | **$146,447** | 7015 | 1753 | 75024 | 47 |

# Map of Nearest Neighbors

# 

# Nearby Venues:

# Used Four Square API data to look at the 10 most common nearby venues for Lewisville area.

# This is for Phase 2 where we increase the functionality of the tool so it can be used by prospective home buyers. Based on their criteria of buying a house: Living Area, Age, Price, Nearby venues etc. the tool can then select the neighborhoods that will work for them

|  |  |
| --- | --- |
| **Venue Type** | **Count** |
| **Cosmetics Shop** | 336.00 |
| **Park** | 204.00 |
| **Gym / Fitness Center** | 168.00 |
| **American Restaurant** | 152.00 |
| **Disc Golf** | 150.00 |
| **Ice Cream Shop** | 142.00 |
| **Donut Shop** | 130.00 |
| **Health & Beauty Service** | 127.00 |
| **Restaurant** | 115.00 |
| **Coffee Shop** | 110.00 |

# Findings:

# The model can be used to assist home owners with:

# Review the value of their houses. If they feel that the appraised value of their house by the county is too high/low we can use the model to not only assess it but also pull data on nearest neighbors.

# Help the homeowners predict the value of their houses for the coming year so they can plan for the change in house taxes

# Help home buyers decide the neighborhood they want to purchase a house in based on their criteria: Living area, age of the house, total land area and nearby venues

# Conclusion

# We can make this analysis more robust by adding additional information such as:

# 1 – Assigned Elementary/Middle and High schools & their ratings

# 2- Year on Year change in home prices

# 3 – Listings of houses on sale

# 4 – Sale prices of all houses sold in the year

# This will become an effective tool for home buyers, home sellers, home owners who want to dispute the appraised value of their houses & a tool for planning property taxes for the next year.