WAKE COUNTY HOUSE SALE PRICE ANALYSIS

Project Overview

Made by Lara Ljumovic

SOURCING DATA

Source and type of data:

Data was downloaded directly from official Wake County, North Carolina web site and it contains information about Real Estate Property. This is administrative data. This data was accessed from web site https://www.wake.gov/ on October 29th, 2023. I converted xlsx file to csv file upon downloading.

Content:

Data contains information about ownership, sale information and property details for all Wake County real estate parcels. Original data has 87 columns and 443096 rows and is in xlsx format. I will convert it to csv.

Additional information about codes in this dataset are available in document CodeDescription.pdf

Collection:

Data is most likely collected by tax administration of Wake County and the data files are refreshed daily and reflect property values as of the most recent countywide reappraisal.

Ethics:

It appears there is no collection bias, as this is truthful publicly available information about deeds of every property in Wake County. As they are revised every day, I don't think there is any error present in this data.

However, this data contains Personal Identifiable Information (PII) such as owner names and exact addresses. This will be excluded from data.

Limitations:

From initial exploration of data, it appears that there is no data about the number of bedrooms for residential property. This data only has information about the date when property was sold and for how much, as well as at what amount it was appraised so that property tax can be calculated from this. It would be nice to have information like listed price, how many days property was listed for sale and similar.

Reasons why I chose this data for my final project and my Wishlist to expand it

My husband and I bought a house in this area almost 2 years ago, and we talked about a software or tool or something that will somehow analyze all the houses in this area, rank it per our desires, our budget, calculate distance from work, compare lot sizes and calculate some kind of score that will cover all our "must haves".

We commented back then that buying a house is a 3-dimensional formula – first axis is price, second is lot size (as we wanted at least some square footage of backyard), and third is distance from Research Triangle Park. Funny enough it turns out that in almost every case you can have only 2 of these 3. We "sacrificed" the 3rd one. These were the 3 biggest things important to us. Besides those, we have some more requests: no townhomes, no first-floor bedrooms, and a must-have garage.

So, this project is a homage to that long and difficult search for our dream home and if we decide to move closer to RTP we'll have a new tool to play with – this project. It can be filled with new data whenever pleased.

This project can be valuable to many other people wanting to move to our "Silicone Valley of the East". After all Raleigh, capitol of North Caroline was voted several times as one of the Top 3 cities to live in USA, with biggest Research Park in USA, amazing 3 Universities in area, a lot of job opportunities, diversified communities, low crime rate and a beautiful climate and nature.

In future I would love to combine cities Durham and Chapell Hill in this analysis, that is, Durham county and Orange county, which are adjacent to Wake county. And because RTP always refers to 3 city area (with 3 big universities in each city one): Raleigh, Durham and Chapell Hill. This would widen the area to find prospective dream home. And I would probably scrape some web data from realtor or Zillow to add more information (like number of bedrooms, price listed to compare to price sold, vicinity of good schools and some other interesting data).

DATA CLEANING AND WRANGLING

1. SUBSETTING

In my project I want to analyze just residential properties. That's why the very first step was subsetting using column TYPE_AND_USE.

# of rows before subsetting	Columns used for subsetting	Conditions applied	# of rows after subsetting
443096	TYPE_AND_USE	Only took rows with values 1, 2, 3, 4 , 5 and 6	353866

Codes 1, 2, 3, 4, 5 and 6 will later be replaced with words and phrases.

This step removed 89230 rows from dataset.

2. INITIAL DROP OF COLUMNS – First Wrangling step

As mentioned above, the original data set has 87 columns. First, I will just give a brief description of each, followed by Y or N letter if this column was dropped (Y) or not (N). Next, I will write remaining steps of data cleaning and wrangling.

Column name	Description	Dropped (Y/N)
OWNER1	Full names of owners or Companies. (PPI and needs to be removed for privacy reasons)	Y
OWNER2		
Mailing_address1	Address of owner or name of the company or PO box with different combinations of these information for these 3 columns. (PPI and need to be removed for privacy reasons)	Y
Mailing_Address2		
Mailing_Address3		
REAL_ESTATE_ID	Real estate unique identifier number.	N
CARD_NUMBER	Card number in deed repository.	Y
NUMBER_OF_CARDS		
Street_Number	Details about address of the real estate.	Y
Street_Prefix		
Street_Name		
Street_Type		
Street_Suffix		
Planning_Jurisdiction	2 letter code for city or town	Y
Street_Misc	Additional codes for street.	Y
Township	Numbers from 01 to 20. Codes for township.	Y
Fire_District	Numbers 23,24,25 or 26 representing fire dep. buildings for district	Y
Land_Sale_Price	Price of land when sold.	N
Land_Sale_Date	Date when land was sold.	N
Zoning	Code for the city/town zone of the land.	Y
Deeded_Acreage	Acres of land as written in Deed.	N
Total_sale_Price	Total price of real estate when sold.	N
Total_Sale_Date	Date when real estate was sold.	N
Assessed_Building_Value	Assessed value of building.	N
Assessed_Land_Value	Assessed value of the land.	N
Parcel_Identification	Unique identifier number for parcel.	Y
Special_District1	Codes for special districts if any.	Y
Special_District2		
Special_District3		
BILLING_CLASS	Codes 1 to 6 for class of billing.	Y
PROPERTY_DESCRIPTION	Address if residential, or name of company or company as owner.	Y
Land_classification	Cde in alphabet letters to denote land classification.	Y
DEED_BOOK	Number of Deed book.	Y
DEED_PAGE	Number of page in Deed book.	Y
Deed_Date	Date when Deed was made.	Y
VCS	Code for deed.	Y
PROPERTY_INDEX	Either VCS or name of company.	Y
Year_Built	Year when real estate was built.	N

NUM_of_Rooms	Number of rooms. All values are 0.	Y
UNITS	Number of units.	N
HEATED_AREA	Square footage of heated area. (In house or building this excludes garage, basement and similar)	N
UTILITIES	Type of basic utilities property has.	N
Street_pavement	If street is paved. All empty cells.	Y
TOPOGRAPHY	Type of topography. All empty cells.	Y
Year_of_Addition	Year when something was added.	N
Effective_year	Year when it became effective property.	N
Remodeled_Year	Year when property was remodeled.	N
Special_Write_In	Short description of type of real estate.	Y
Story_Height	Height of building measured in stories.	N
DESIGN_STYLE	Design style of building.	N
Foundation_Basement		Y
Foundation_Basement_Percent		Y
Exterior_Wall	Type of exterior.	Y
COMMON_WALL	Common wall. All empty cells.	Y
ROOF	Roof. All empty cells.	Y
Roof_Floor_System	Roof Floor system.	Y
Floor_Finish	Type of floor finish.	Y
Interior_Finish	Type of interior finish. All empty cells. Percentages of interior finish 1-99, obsolete.	Y
Interior_Finish1		
Interior_Finish1_percent		
Interior_Finish2		
Interior_Finish2_percent		
HEAT	Type of heating	Y
HEAT_PERCENT	Percent of heating.	Y
AIR	Type of air conditioning.	Y
AIR_PERCENT	Percent of air conditioning.	Y
BATH	Number of bathrooms in codes as letters.	N
BATH_FIXTURES	Codes for bath fixtures.	Y
Built_in1_Description	Additional descriptions and features, such as number of fireplaces, elevators, sprinklers and more.	Y
Built_in2_Description		
Built_in3_Description		
Built_in4_Description		
Built_in5_Description		
CITY	3 letter code for city or town	Y
GRADE	Grade level of property for tax purposes.	Y
Assessed_Grade_Difference	Difference for grade of property for tax purposes.	Y
Accrued_Assessed_Condition_Pct	Percentage of assessed value of property that is taxable.	Y
Land_Deferred_code	Deferred code for land.	Y
Land_Deferred_Amount	Amount for deferred land.	Y
Historic_Deferred_code	Code for deferred property based on historic value.	Y
Historic_Deferred_Amount	Amount for deferred historic property.	Y
RECYCLED_UNITS	Units that are recycled.	Y
Disq_and_Qual_flag	Flag for disqualified or qualified property or land.	Y
Land_Disq_and_Qual_flag		
TYPE_AND_USE	Code for type or usage of real estate. I will only use code 01.	N
PHYSICAL_CITY	City or town name	N

PHYSICAL_ZIP_CODE	Zip code of city or town.	N

Before continuing with data cleaning, I decided to first change name of columns, for easier work later. I changed data types after completing cleaning, but included both this steps in same column for easier presentation of steps.

3. DATA WRANGLING

Column name	New column name (Before cleaning data!)	Changed data types (After cleaning data!)
REAL_ESTATE_ID	real_estate_id	
Land_Sale_Price	land_sale_price	Object changed to int_32
Land_Sale_Date	land_sale_date	
Deeded_Acreage	deeded_acreage	
Total_sale_Price	total_sale_price	Object changed to int_32
Total_Sale_Date	total_sale_date	
Assessed_Building_Value	assessed_building_value	Object changed to int_32
Assessed_Land_Value	assessed_land_value	Object changed to int_32
Year_built	year_built	
UNITS	units	
HEATED_AREA	heated_area	
UTILITIES	utilities	
Year_of_Addition	addition_year	
Effective_year	effective_year	
Remodeled_Year	remodeled_year	
Story_Height	story_height	
DESIGN_STYLE	design_style	
BATH	bath	
TYPE_AND_USE	type_and_use	
PHYSICAL_CITY	city	
PHYSICAL_ZIP_CODE	zip_code	float_64 changed to int_32

4. DATA CLEANING

Some steps of cleaning data were performed before and some after certain steps of wrangling. This was necessary to allow easier work with data in a process of cleaning and wrangling. All steps are written in order of performing them and they are all gathered here for better readability. Of this document.

Issues	How many rows/Action performed	Comment
CHANGING CODES TO REAL VALUES In column ‘story_height’	'A':'1 story', 'B':'1.5 story', 'C':'2 story', 'D':'2.5 story', 'E':'3 story', 'F':'3.5 story', 'G':'4 story', 'H':'Multi story', 'I':'1.75 story', 'J':'1.4 story', 'K':'1.63 story', 'L':'1.88 story', 'M':'2.4 story', 'N':'2.63 story', 'O':'2.75 story'	There was a mistake in CodeDescriptions.pdf document where there was a code R for 2.75 story but it should be O. I check entire raw data set and there are 4 properties with O code. All numbers were changed to decimal number where needed.

Missing values in column ‘story_height’	Imputed 950 rows with value ‘Unknown’.	Removing values was not necessary. It won’t affect analysis.
CHANGING CODES TO REAL VALUES In column ‘design_style’	'A':'Conventional', 'B':'Duplex', 'C':'Townhouse', 'D':'Condo', 'E':'Conversion', 'F':'Colonial', 'G':'Ranch', 'H':'Cape', 'I':'Split level', 'J':'Split foyer', 'K':'Contemporary', 'L':'Log', 'M':'Manuf sngl', 'N':'Manuf multi', 'O':'Modular'	All descriptions for codes were kept as they were.
Missing values in column ‘design_style’	Imputed 951 rows with value ‘Unknown’.	Removing values was not necessary. It won’t affect analysis.
Incorrect values in column ‘design_style’	Imputed 2 values ‘P’ with value ‘Unknown’.	This code didn’t exist in CodeDescription.pdf
CHANGING CODES TO REAL VALUES in column ‘bath’	'A':'1 bath', 'B':'1.5 bath', 'C':'2 bath', 'D':'2.5 bath', 'E':'3 bath', 'F':'3.5 bath', 'G':'Limited plmg', 'H':'No plumbing', 'I':'Adequate', 'J':'NO of fixtures'	Words describing numbers were changed to decimal numbers where needed.
Missing values in column ‘bath’	Imputed 976 rows with value ‘Unknown’.	Removing values was not necessary. It won’t affect analysis.
CHANGING CODES TO REAL VALUES In column ‘type_and_use’	'1':'1 family', '2':'2 family', '3':'3 family', '4':'4 family', '5':'Multi family', '6':'Res. w/busi use'	Words describing numbers were changed to numbers where needed.
Duplicates	Removed 445 rows.	These were full duplicates.
Duplicates in column ‘real_estate_id’	Removed 1493 rows.	Thise were duplicated id-s that should not be in dataset.
Missing values in column ‘city’	Removed 228 rows.	Not suitable for spatial analysis.
Incorrect values in column ‘city’	Removed values CLAYTON, CREEDMOOR, NEW HILL, WILLOW SPRING and YOUNGSVILLE. All together 5563.	These were all municipalities that are not part of Wake County.
Missing values in column ‘zip_code’	Removed 69 rows.	Not suitable for spatial analysis.
‘0’ values in column ‘zip_code’	Removed 7 rows.	Not suitable for spatial analysis. Zip code cannot be 0.
Missing values in column ‘utilities’	Imputed 2344 rows with value ‘Unknown’	Removing values was not necessary. It won’t affect analysis.
Missing values in column ‘land_sale_date’	Imputed 209736 rows with ’00.00.0000’.	These lands were never re-sold.
Missing values in column ‘total_sale_date’	Imputed 20837 rows with ’00.00.0000’.	These properties with land included were never re-sold.
Inconsistent values in column ‘assessed_building_value’	Removed 213 rows with values 0 and 2 rows with values 1.	Unrealistic values.
Inconsistent values in column ‘built_year’	Removed 4 row with values 0 and 11 with 2024.	Unrealistic values.
Inconsistent values in column ‘addition_year’	Removed 23 rows all together.	Unrealistic years > 2023

Inconsistent values in column ‘effective_year’	Removed 5 rows all together.	Unrealistic years > 2023
Inconsistent values in column ‘remodeled_year’	Removed 7 rows all together.	Unrealistic years > 2023
Inconsistent values in column ‘units’	Removed 1314 rows with value 0	Unrealistic value.
Unwanted commas in column ‘land_sale_price’	All rows.	Needed to be removed to be able to convert data type to int.
Unwanted commas in column ‘total_sale_price’	All rows.	Needed to be removed to be able to convert data type to int.
Unwanted commas in column ‘assessed_building_value’	All rows.	Needed to be removed to be able to convert data type to int.
Unwanted commas in column ‘assessed_land_value’	All rows.	Needed to be removed to be able to convert data type to int.

DATA PROFILE

Variable	**Time Variant/ Invariant**	**Structured/ Unstructured**	**Qualitative/ Quantitative**	**Nominal/Ordinal/ Discrete/Continuous**
real_estate_id	Invariant	Structured	Quantitative	Discrete
land_sale_price	Variant	Structured	Quantitative	Continuous
land_sale_date	Invariant	Structured	Quantitative	Discrete, interval
deeded_acreage	Invariant	Structured	Quantitative	Discrete
total_sale_price	Variant	Structured	Quantitative	Continuous
total_sale_date	Variant	Structured	Quantitative	Discrete
assessed_building_value	Variant	Structured	Quantitative	Continuous
assessed_land_value	Variant	Structured	Quantitative	Continuous
year_built	Invariant	Structured	Quantitative	Discrete, interval
units	Invariant	Structured	Quantitative	Discrete
heated_area	Invariant	Structured	Quantitative	Discrete
utilities	Invariant	Unstructured	Qualitative	Nominal
addition_year	Variant	Structured	Quantitative	Discrete, interval
effective_year	Variant	Structured	Quantitative	Discrete, interval
remodeled_year	Variant	Structured	Quantitative	Discrete, interval
story_height	Invariant	Structured	Quantitative	Discrete
design_style	Invariant	Unstructured	Qualitative	Nominal
bath	Invariant	Structured	Quantitative	Discrete
type_and_use	Invariant	Structured	Qualitative	Nominal
city	Invariant	Structured	Qualitative	Nominal
zip_code	Invariant	Structured	Quantitative	Discrete

QUESTIONS I WOULD LIKE TO ANSWER WITH MY ANALYSIS

1. Does age of residential property affects the price? How much?

2. How sold amounts and assessed amounts of property changed over the years?

3. Are properties more expensive in bigger cities? How about cities closer to RTP?

4. Does design style, bats, utilities or something else affects assessed value?

5. Can we forecast price or assessed value of property?

6. What is the average age of property, average land deeded, average heated area, average assessed value? How does this compare in each city?

7. In what month are most houses sold? What was average price? Is this similar in each zip code?

8. How does age of property compare to effective year? How does this compares to zip codes?