August 2023

Predicting Property Value

Sakava Kiv | Alex Thibeaux
William Jones | Chris Mathew

August 2023

TABLE OF CONTENTS

01	Purpose & Background	→	Purpose of this project and What is property value
02	EDA	→	Initial look into the data, regressions, and KNN
03	Database	→	Relational database design and implementation

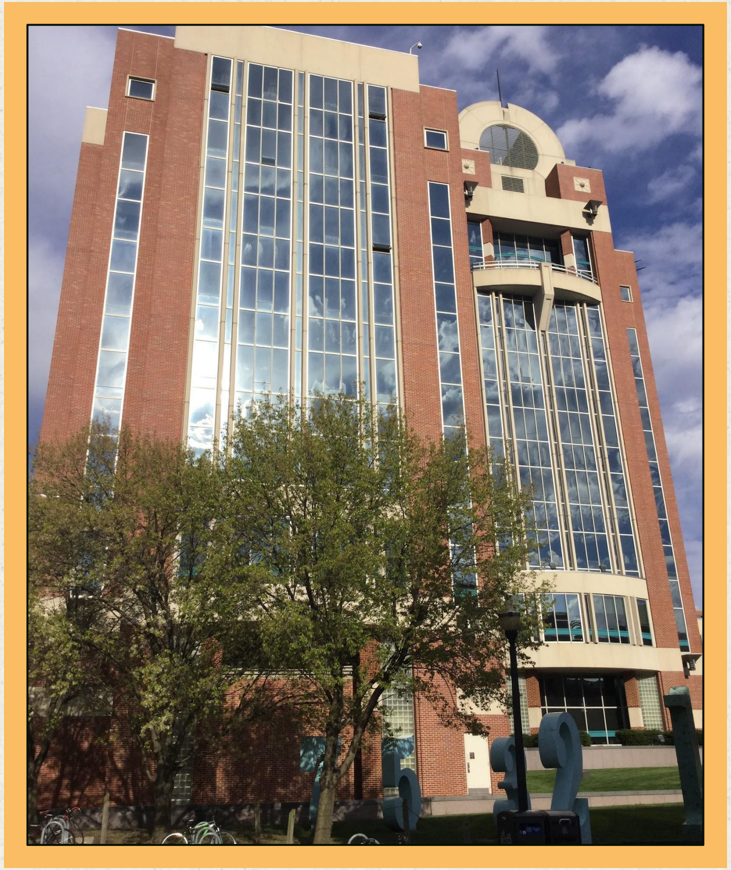
Purpose & Background

01.

August 2023

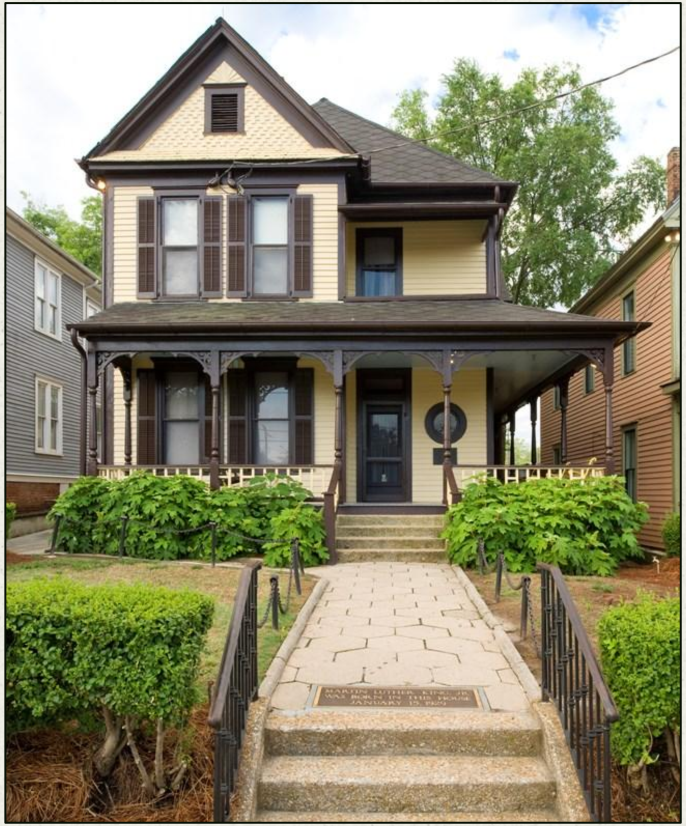
Purpose

- To gain experience applying data science techniques on real-world datasets and implementing a database
- Gain insights in the real estate market
- Predicting the property value



What is property value?

- Probable price of a given property at a given time
- It is important for real estate deals and assessing the property tax
- There are many factors that affect the property value i.e. location, size, condition of building



August 2023

Available Data

MONETARY

taxamount
landtaxvaluedollarcnt
taxvaluedollarcnt
structuretaxvaluedollarcount
saleprice

PROPERTY

bathroomcnt
bathroomcnt
calculatedbathnbr
calculatedfinshedsquarefeet
finishedsquarefeet12
fips
lotsizesquarefeet
roomcnt
yearbuilt
assessmentyear
latitude
longitude
fullbathroomcnt

Identification

parcelid
propertycountylandusecode
propertycountylandusetypeid
rawcensustractandblock
regionidcity
regionidcounty
regionidzip

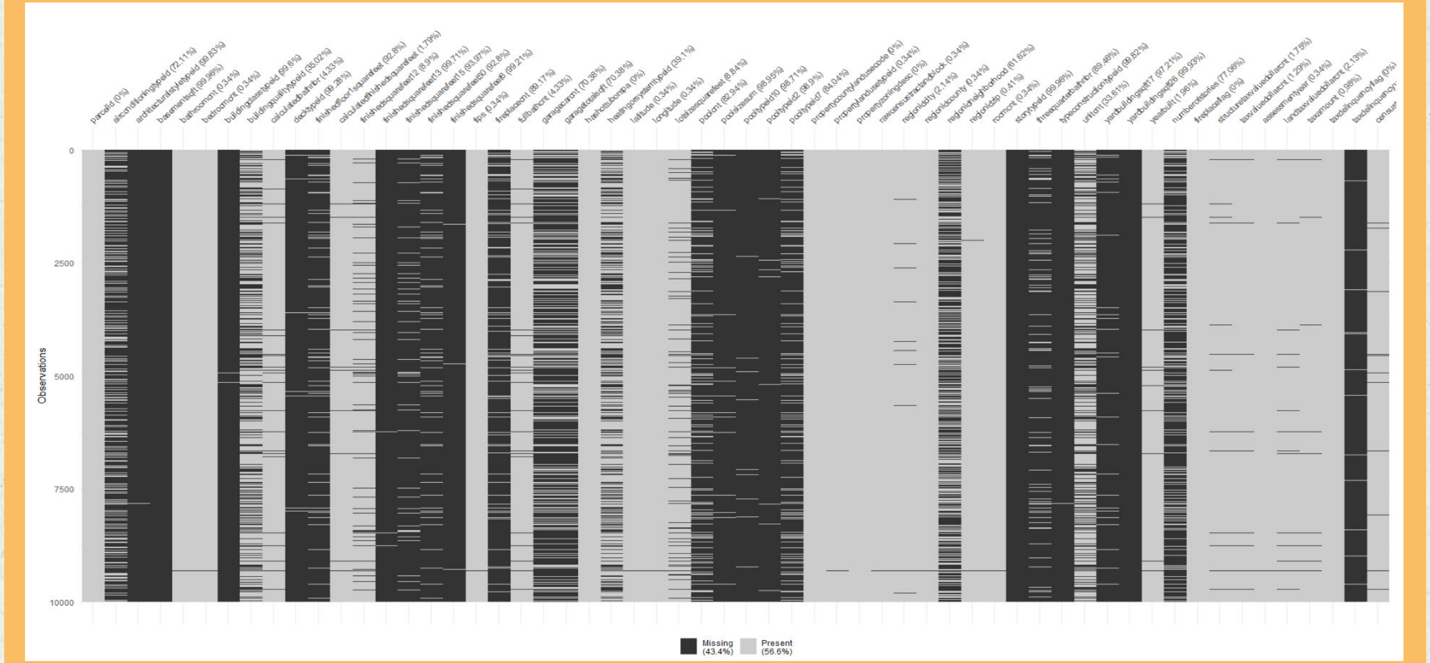
August 2023

Exploratory Data Analysis

02.

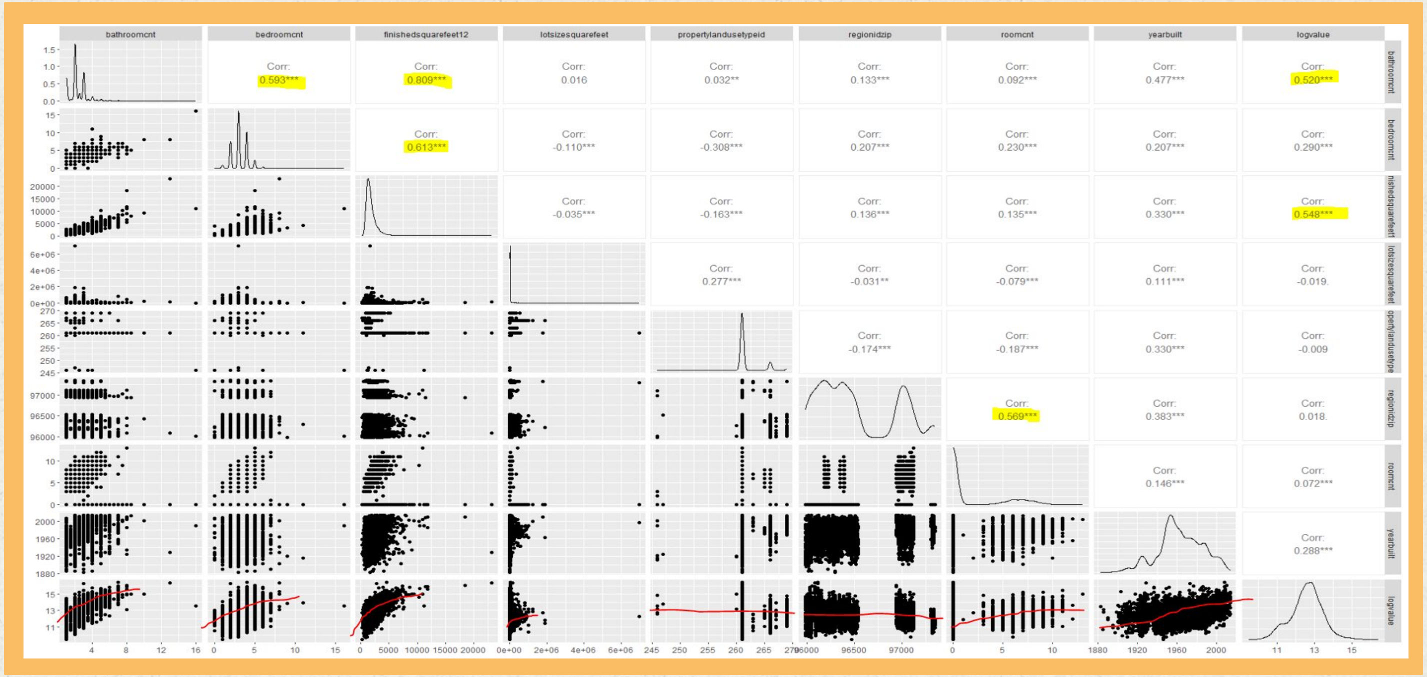
Missing Values

From a random sample of 10,000 observations, it was determined that there were 30 parameters that could not be used in the analysis due to missing data. Imputation is not helpful for parameters that are missing more than 10% of data.

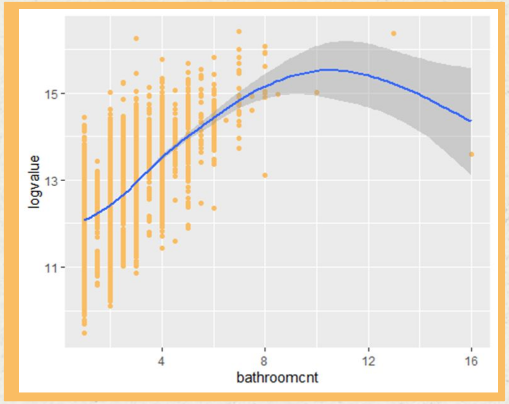


Assumptions: Multi-Linear Regression

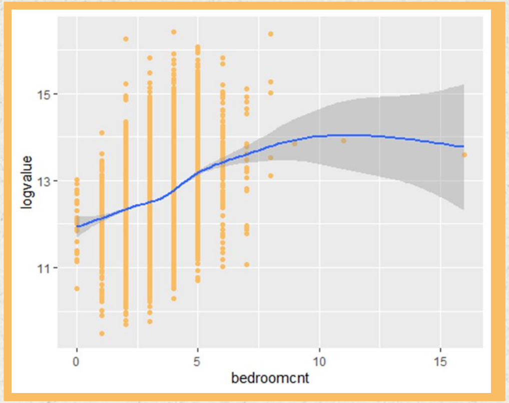
1) linearity
2) homoskedasticity
3) independence of errors
4) normality
5) independence of independent variables



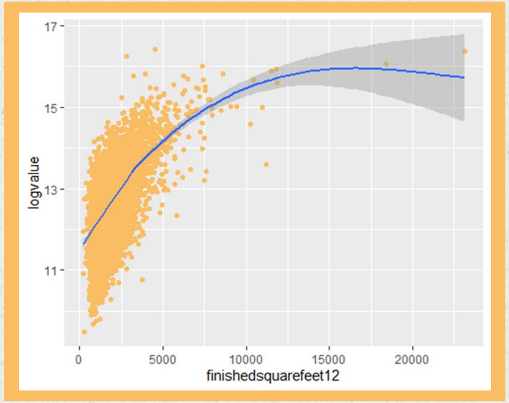
Simple Linear Regression



# Bathrooms

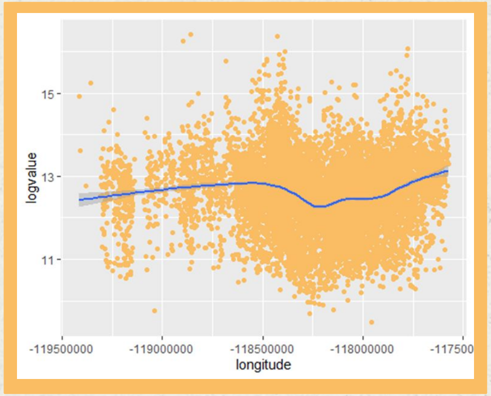


Bedrooms

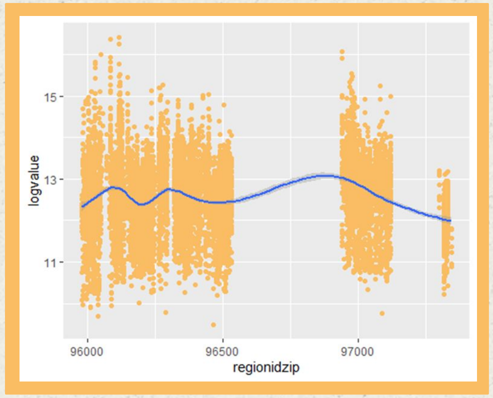


Finished Sq. Ft.

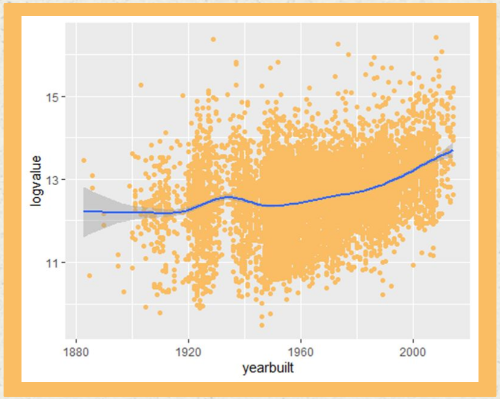
Simple Linear Regression



Longitude



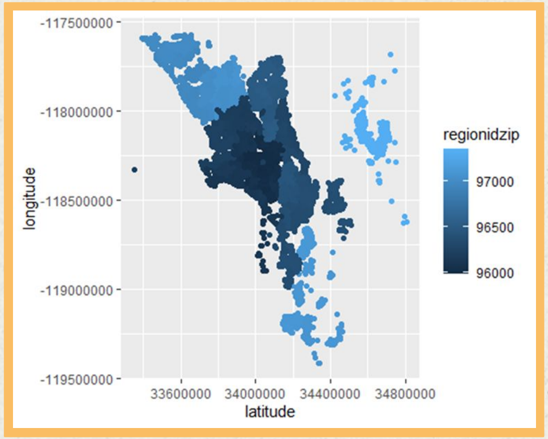
Zip Code



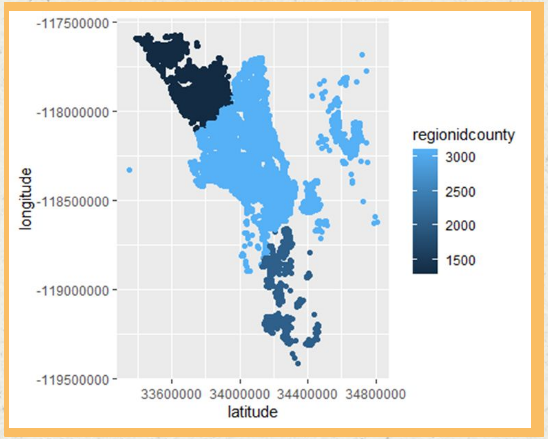
Finished Sq. Ft.

August 2023

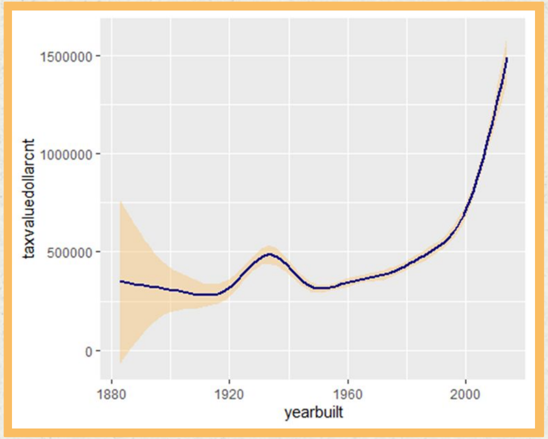
Location & History Data



Zip Code



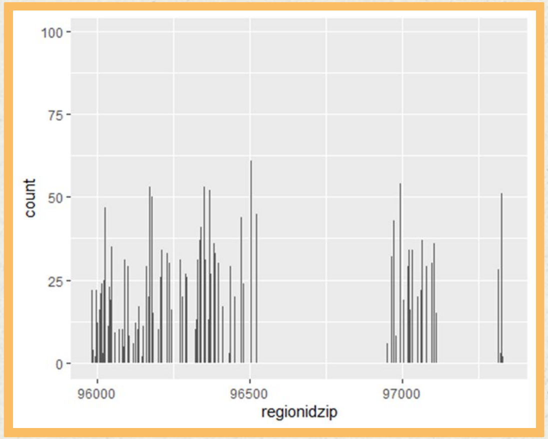
County



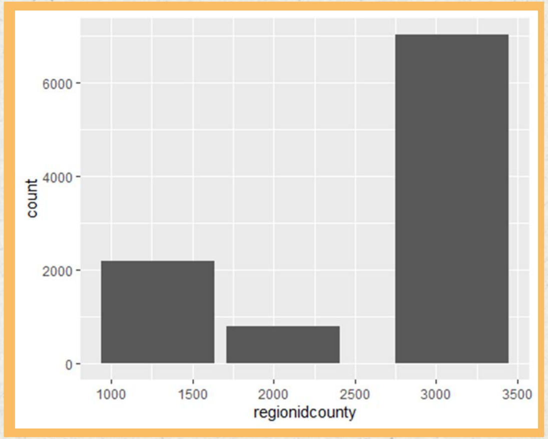
Year Built

August 2023

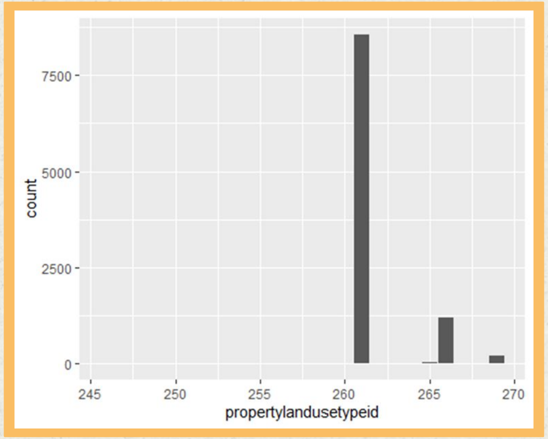
Categorical Parameters



Zip Code

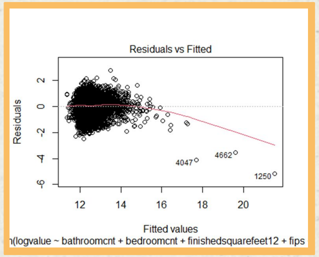


County



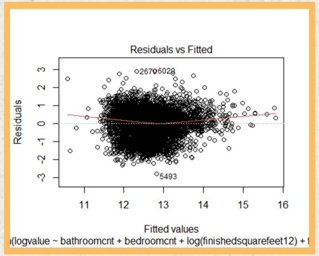
Property Land use Type

Model RMSE



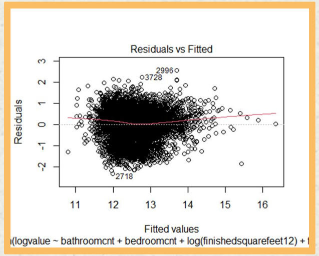
Full Model

RMSE: 0.674



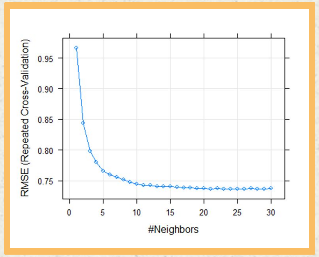
Logged Model, 2016

RMSE: 0.678



Logged Model, 2017

RMSE: 0.669



K-Nearest Neighbors

RMSE: 0.725

Relational Database

03.

SQL Schema Design: Raw vs Cleaned Data

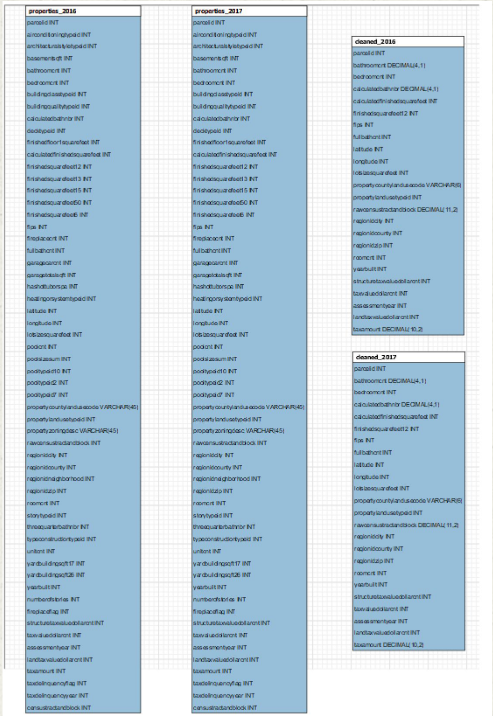
Four Tables

58 Columns

Number of columns in ***raw*** data tables

24 Columns

Number of columns in ***cleaned*** data tables



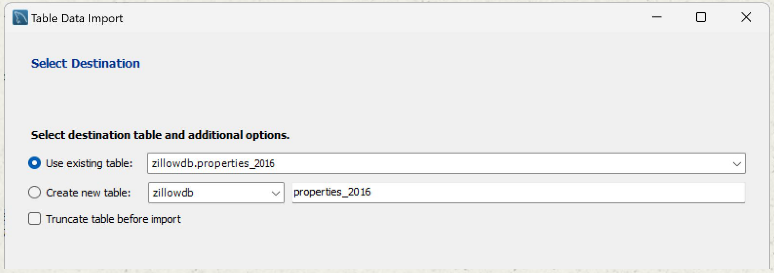
properties_2016
parcelid INT
airconditioningtypeid INT
architecturalstyletypeid INT
basementsqft INT
bathroomcnt INT
bedroomcnt INT
buildingclasstypeid INT
buildingqualitytypeid INT
calculatedbathnbr INT
decktypeid INT
finishedfloor1squarefeet INT
calculatedfinishedsquarefeet INT
finishedsquarefeet12 INT
finishedsquarefeet13 INT
finishedsquarefeet15 INT
finishedsquarefeet50 INT
finishedsquarefeet6 INT
fips INT
fireplacecnt INT
fullbathcnt INT
garagecarcnt INT
garagetotalsqft INT
hashottuborspa INT
heatingorsystemtypeid INT
latitude INT
longitude INT
lotsizesquarefeet INT
poolcnt INT
poolsizesum INT
pooltypeid10 INT
pooltypeid2 INT
pooltypeid7 INT
propertycountylandusecode VARCHAR(45)
propertylandusetypeid INT
propertyzoningdesc VARCHAR(45)
rawcensustractandblock INT
regionidcity INT
regionidcounty INT
regionidneighborhood INT
regionidzip INT
roomcnt INT
storytypeid INT
threequarterbathnbr INT
typeconstructiontypeid INT
unitcnt INT
yardbuildingsqft17 INT
yardbuildingsqft26 INT
yearbuilt INT
numberofstories INT
fireplaceflag INT
structuretaxvaluedollarcnt INT
taxvaluedollarcnt INT
assessmentyear INT
landtaxvaluedollarcnt INT
taxamount INT
taxdelinquencyflag INT
taxdelinquencyyear INT
censustractandblock INT

properties_2017
parcelid INT
airconditioningtypeid INT
architecturalstyletypeid INT
basementsqft INT
bathroomcnt INT
bedroomcnt INT
buildingclasstypeid INT
buildingqualitytypeid INT
calculatedbathnbr INT
decktypeid INT
finishedfloor1squarefeet INT
calculatedfinishedsquarefeet INT
finishedsquarefeet12 INT
finishedsquarefeet13 INT
finishedsquarefeet15 INT
finishedsquarefeet50 INT
finishedsquarefeet6 INT
fips INT
fireplacecnt INT
fullbathcnt INT
garagecarcnt INT
garagetotalsqft INT
hashottuborspa INT
heatingorsystemtypeid INT
latitude INT
longitude INT
lotsizesquarefeet INT
poolcnt INT
poolsizesum INT
pooltypeid10 INT
pooltypeid2 INT
pooltypeid7 INT
propertycountylandusecode VARCHAR(45)
propertylandusetypeid INT
propertyzoningdesc VARCHAR(45)
rawcensustractandblock INT
regionidcity INT
regionidcounty INT
regionidneighborhood INT
regionidzip INT
roomcnt INT
storytypeid INT
threequarterbathnbr INT
typeconstructiontypeid INT
unitcnt INT
yardbuildingsqft17 INT
yardbuildingsqft26 INT
yearbuilt INT
numberofstories INT
fireplaceflag INT
structuretaxvaluedollarcnt INT
taxvaluedollarcnt INT
assessmentyear INT
landtaxvaluedollarcnt INT
taxamount INT
taxdelinquencyflag INT
taxdelinquencyyear INT
censustractandblock INT

cleaned_2016
parcelid INT
bathroomcnt DECIMAL(4,1)
bedroomcnt INT
calculatedbathnbr DECIMAL(4,1)
calculatedfinishedsquarefeet INT
finishedsquarefeet12 INT
fips INT
fullbathcnt INT
latitude INT
longitude INT
lotsizesquarefeet INT
propertycountylandusecode VARCHAR(6)
propertylandusetypeid INT
rawcensustractandblock DECIMAL(11,2)
regionidcity INT
regionidcounty INT
regionidzip INT
roomcnt INT
yearbuilt INT
structuretaxvaluedollarcnt INT
taxvaluedollarcnt INT
assessmentyear INT
landtaxvaluedollarcnt INT
taxamount DECIMAL(10,2)

cleaned_2017
parcelid INT
bathroomcnt DECIMAL(4,1)
bedroomcnt INT
calculatedbathnbr DECIMAL(4,1)
calculatedfinishedsquarefeet INT
finishedsquarefeet12 INT
fips INT
fullbathcnt INT
latitude INT
longitude INT
lotsizesquarefeet INT
propertycountylandusecode VARCHAR(6)
propertylandusetypeid INT
rawcensustractandblock DECIMAL(11,2)
regionidcity INT
regionidcounty INT
regionidzip INT
roomcnt INT
yearbuilt INT
structuretaxvaluedollarcnt INT
taxvaluedollarcnt INT
assessmentyear INT
landtaxvaluedollarcnt INT
taxamount DECIMAL(10,2)

Importing the Data into MySQL Database



Select Destination

Select destination table and additional options.

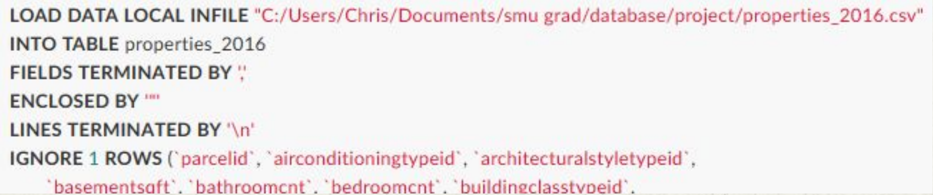
● Use existing table: zillowdb.properties_2016

○ Create new table: zillowdb properties_2016

□ Truncate table before import

■ Table Data Import Wizard

- Easier user-interface
- Slower loading speed



```
LOAD DATA LOCAL INFILE "C:/Users/Chris/Documents/smu grad/database/project/properties_2016.csv"
INTO TABLE properties_2016
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS (`parcelid`, `airconditioningtypeid`, `architecturalstyletypeid`,
    `basementsqft`, `bathroomcnt`, `bedroomcnt`, `buildingclasstypeid`,
```

🟧 LOAD DATA Statement

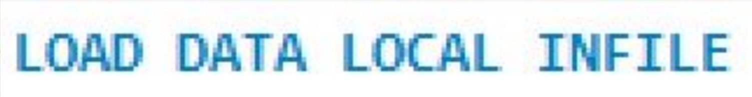
- More configurable
- Faster loading speed

Errors and Obstacles during Database Implementation

1. Managing Load Speeds

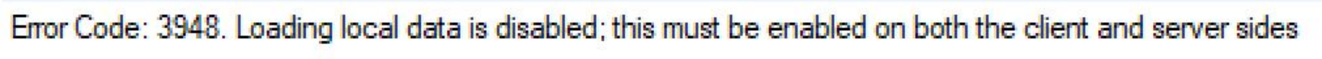
2. Determining Column Data Types

3. Error while loading Data using LOAD DATA statement



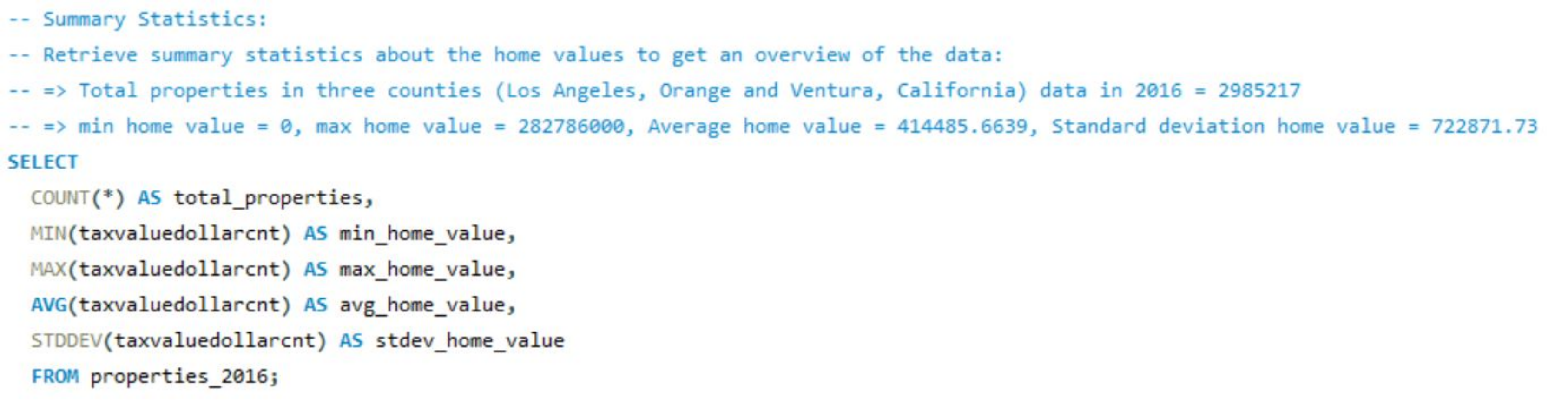
```
LOAD DATA LOCAL INFILE
```

Convert CSV to SQL

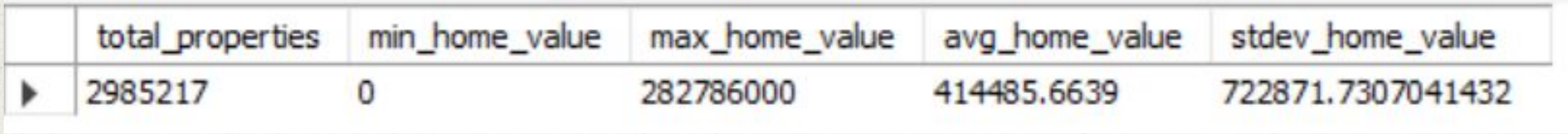


```
Error Code: 3948. Loading local data is disabled; this must be enabled on both the client and server sides
```

Summary Statistics with SQL

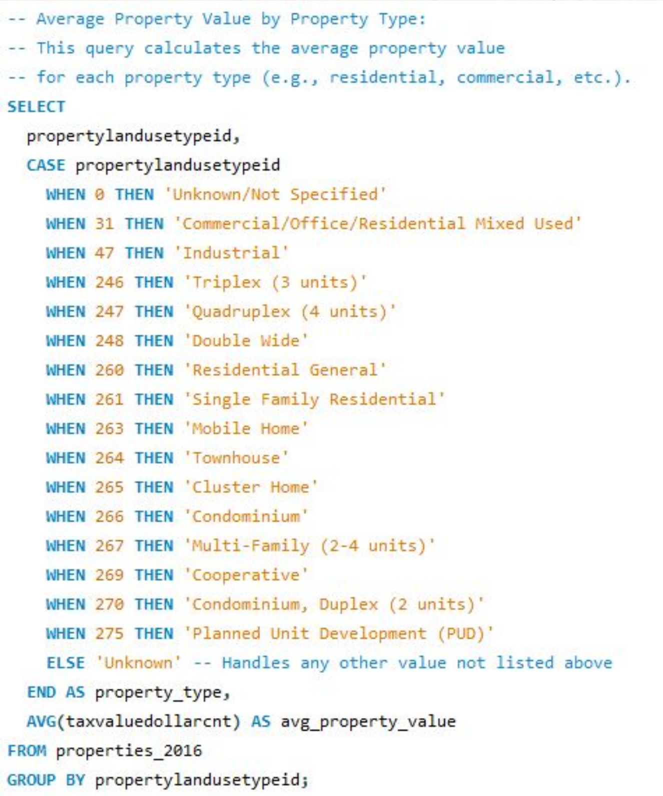


```
-- Summary Statistics:
-- Retrieve summary statistics about the home values to get an overview of the data:
-- => Total properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016 = 2985217
-- => min home value = 0, max home value = 282786000, Average home value = 414485.6639, Standard deviation home value = 722871.73
SELECT
  COUNT(*) AS total_properties,
  MIN(taxvaluedollarcnt) AS min_home_value,
  MAX(taxvaluedollarcnt) AS max_home_value,
  AVG(taxvaluedollarcnt) AS avg_home_value,
  STDDEV(taxvaluedollarcnt) AS stdev_home_value
  FROM properties_2016;
```

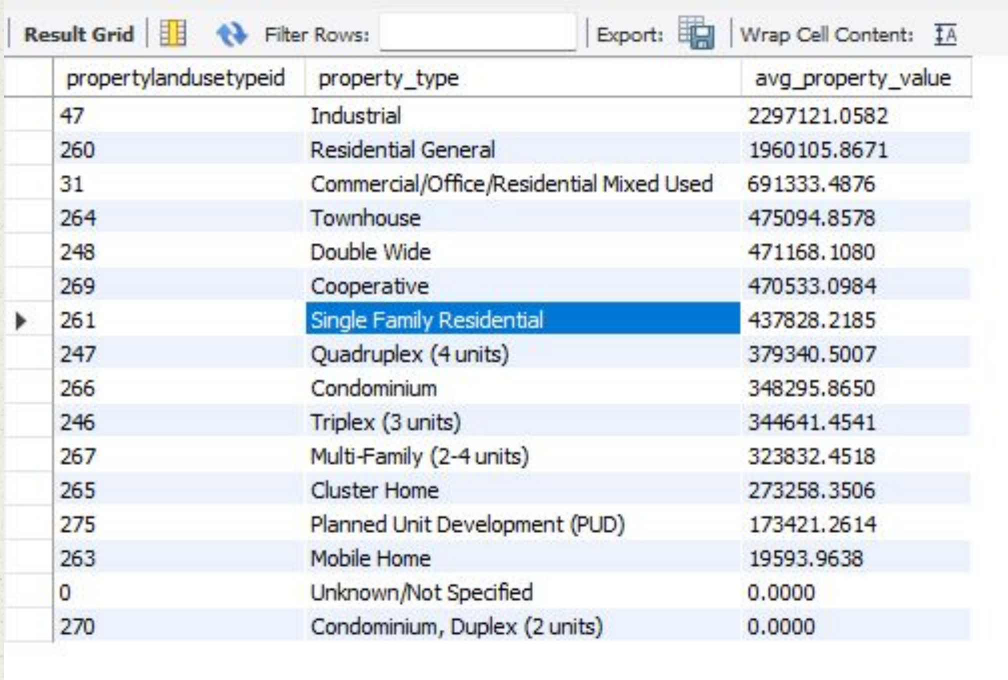


	total_properties	min_home_value	max_home_value	avg_home_value	stdev_home_value
▶	2985217	0	282786000	414485.6639	722871.7307041432

Average Value for Property types SQL

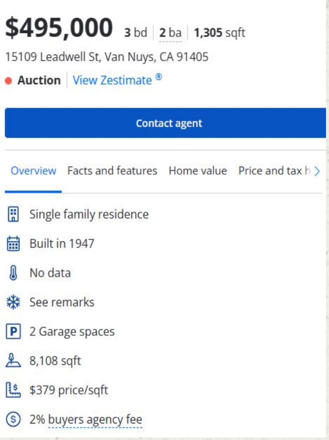


```sql
-- Average Property Value by Property Type:
-- This query calculates the average property value
-- for each property type (e.g., residential, commercial, etc.).
SELECT
  propertylandusetypeid,
  CASE propertylandusetypeid
    WHEN 0 THEN 'Unknown/Not Specified'
    WHEN 31 THEN 'Commercial/Office/Residential Mixed Used'
    WHEN 47 THEN 'Industrial'
    WHEN 246 THEN 'Triplex (3 units)'
    WHEN 247 THEN 'Quadruplex (4 units)'
    WHEN 248 THEN 'Double Wide'
    WHEN 260 THEN 'Residential General'
    WHEN 261 THEN 'Single Family Residential'
    WHEN 263 THEN 'Mobile Home'
    WHEN 264 THEN 'Townhouse'
    WHEN 265 THEN 'Cluster Home'
    WHEN 266 THEN 'Condominium'
    WHEN 267 THEN 'Multi-Family (2-4 units)'
    WHEN 269 THEN 'Cooperative'
    WHEN 270 THEN 'Condominium, Duplex (2 units)'
    WHEN 275 THEN 'Planned Unit Development (PUD)'
    ELSE 'Unknown' -- Handles any other value not listed above
  END AS property_type,
  AVG(taxvaluedollarcnt) AS avg_property_value
FROM properties_2016
GROUP BY propertylandusetypeid;
```



propertylandusetypeid	property_type	avg_property_value
47	Industrial	2297121.0582
260	Residential General	1960105.8671
31	Commercial/Office/Residential Mixed Used	691333.4876
264	Townhouse	475094.8578
248	Double Wide	471168.1080
269	Cooperative	470533.0984
261	Single Family Residential	437828.2185
247	Quadruplex (4 units)	379340.5007
266	Condominium	348295.8650
246	Triplex (3 units)	344641.4541
267	Multi-Family (2-4 units)	323832.4518
265	Cluster Home	273258.3506
275	Planned Unit Development (PUD)	173421.2614
263	Mobile Home	19593.9638
0	Unknown/Not Specifed	0.0000
270	Condominium, Duplex (2 units)	0.0000

Los Angeles, California Vs Dallas, Texas



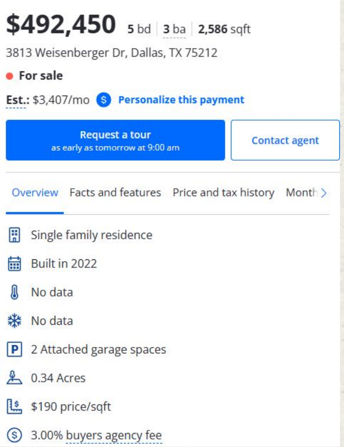
$495,000 3 bd | 2 ba | 1,305 sqft

15109 Leadwell St, Van Nuys, CA 91405

● Auction | View Zestimate ®

- Single family residence
- Built in 1947
- No data
- See remarks
- 2 Garage spaces
- 8,108 sqft
- $379 price/sqft
- 2% buyers agency fee

Average Single Family home of $437,828 for California counties combined



$492,450 5 bd | 3 ba | 2,586 sqft

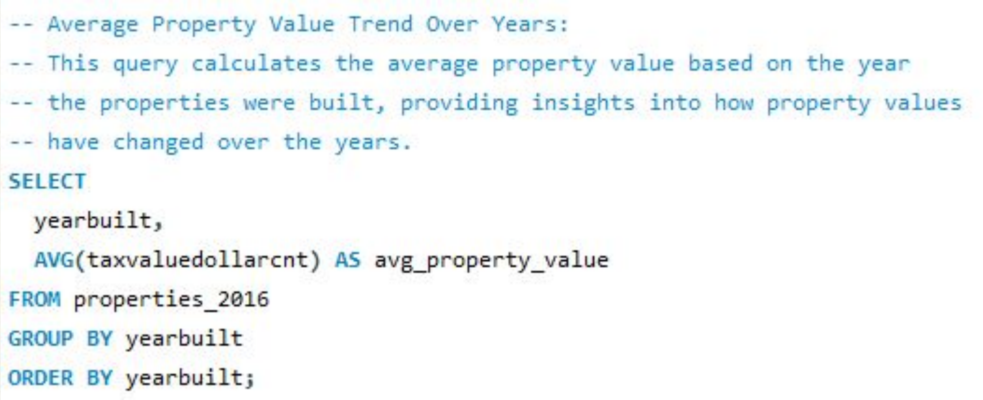
3813 Weisenberger Dr, Dallas, TX 75212

● For sale

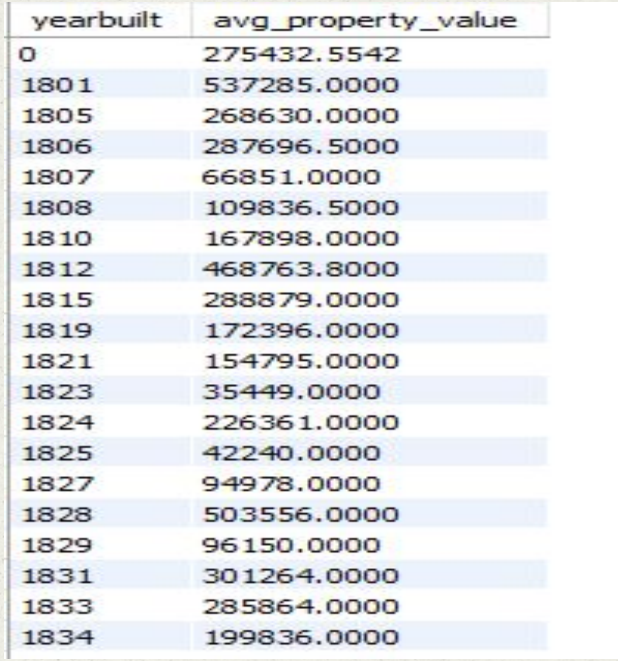
Est.: $3,407/mo Personalize this payment

- Single family residence
- Built in 2022
- No data
- No data
- 2 Attached garage spaces
- 0.34 Acres
- $190 price/sqft
- 3.00% buyers agency fee

Average Value for Property Trends (Year Built) SQL



```
-- Average Property Value Trend Over Years:
-- This query calculates the average property value based on the year
-- the properties were built, providing insights into how property values
-- have changed over the years.
SELECT
  yearbuilt,
  AVG(taxvaluedollarcnt) AS avg_property_value
FROM properties_2016
GROUP BY yearbuilt
ORDER BY yearbuilt;
```

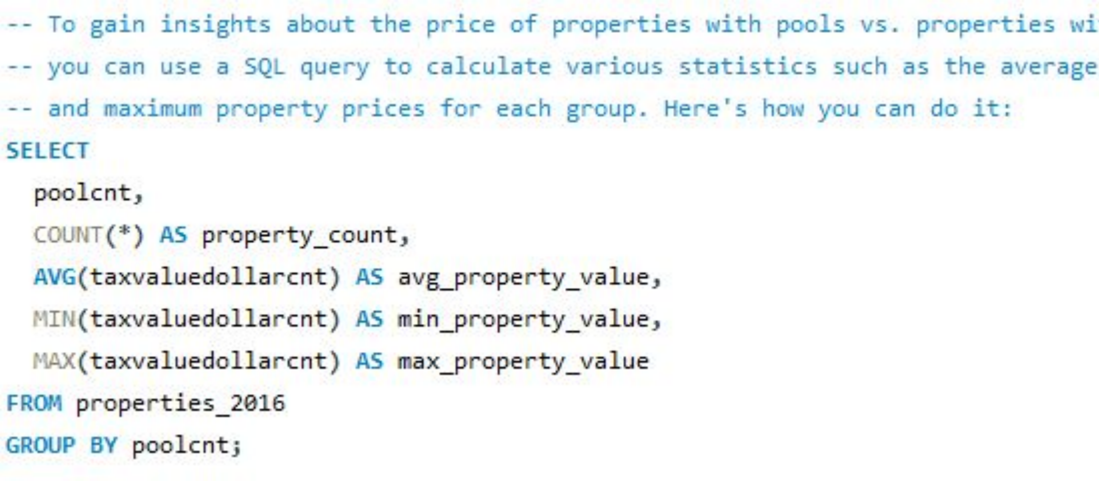


yearbuilt	avg_property_value
0	275432.5542
1801	537285.0000
1805	268630.0000
1806	287696.5000
1807	66851.0000
1808	109836.5000
1810	167898.0000
1812	468763.8000
1815	288879.0000
1819	172396.0000
1821	154795.0000
1823	35449.0000
1824	226361.0000
1825	42240.0000
1827	94978.0000
1828	503556.0000
1829	96150.0000
1831	301264.0000
1833	285864.0000
1834	199836.0000

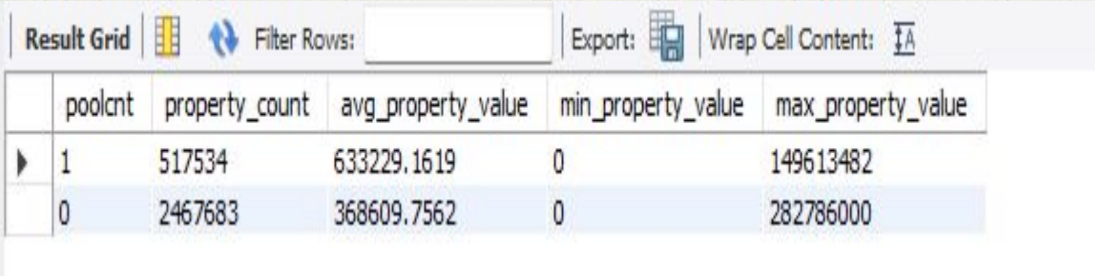
List continues to 2015...

Image of a single family home built in the 1900's
in Los Angeles California

Feature Analysis SQL



```sql
-- To gain insights about the price of properties with pools vs. properties without pools,
-- you can use a SQL query to calculate various statistics such as the average, minimum,
-- and maximum property prices for each group. Here's how you can do it:
SELECT
  poolcnt,
  COUNT(*) AS property_count,
  AVG(taxvaluedollarcnt) AS avg_property_value,
  MIN(taxvaluedollarcnt) AS min_property_value,
  MAX(taxvaluedollarcnt) AS max_property_value
FROM properties_2016
GROUP BY poolcnt;
```



poolcnt	property_count	avg_property_value	min_property_value	max_property_value
1	517534	633229.1619	0	149613482
0	2467683	368609.7562	0	282786000

- Our model was not able to fit the data set well due to missing data in some crucial columns.
- Having a home built closer to current date increases property value
- When building the RDBMS loading the data using the LOAD DATA statement was faster than using the data import wizard.

Future Works

- Inquire to data collectors about missing data.
- Develop a stronger model by integrating different data science techniques together

Conclusion

August 2023

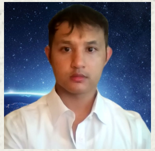
THANKS!

DO YOU HAVE ANY QUESTIONS?

Alex Thibeaux
athibeaux@smu.edu

William Jones
wfjones@smu.edu

Chris Mathew
mathewc@smu.edu



Sakava Kiv
skiv@smu.edu