Systems and Methods for Big and Unstructured Data Project

Maestrello Lucrezia
Piotti Eric

Academic Year: 2024-2025

Contents

1 Introduction

The dataset chosen is named Houston housing market 2024, created by Natasha Lekh, using mainly two APIs: Zillow ZIP Code Scraper and Zillow Details Scraper. This dataset contains detailed information on real estate listings in Houston, Texas and provides a comprehensive snapshot of the Houston housing market as of 5th June 2024. This dataset was intentionally chosen for its richness and comprehensiveness, offering researchers the to carry out diverse and sophisticated queries in the field of real estate. This dataset is ideally suited for a MongoDB implementation, given its large data volume and the compatibility of its numerous fields with the document-based format.

2 Datasets

2.1 Dataset

2.1.1 Dataset Description

This dataset, composed of thousands of documents (about 27000), provides a comprehensive view of Houston's real estate landscape. Created by the author through meticulous data collection from specific APIs and web scraping tools, it offers detailed insights into property characteristics, pricing, and market trends specific to Houston, Texas.

Dataset Credits
Author Natasha Lekh
Title Houston housing market 2024
Year 2024
Link Houston housing market 2024
License CC BY-NC-SA 4.0

2.1.2 Dataset Structure

Field Descriptions

In the following, all fields of the dataset are listed and explained.

Field Name	Description
zpid	Unique identifier for the property on Zillow.
id	Another unique identifier.
rawHomeStatusCd	Raw status code indicating the property's listing status.
marketingStatusSimplifiedCd	Simplified status.
imgSrc	URL of the primary image of the property.
hasImage	Boolean indicating if the property listing has an associated image.
detailUrl	URL linking to the listing page on Zillow.
statusType	High-level status of the property.
statusText	Text description of the property's status.
countryCurrency	Symbol of the currency used in the price.
price	Formatted string representation of the property price.
unformattedPrice	Numeric representation of the price.
address	Full address of the property, including street, city, state, and ZIP code.
addressStreet	Street address.
addressCity	City where the property is located.
addressState	State where the property is located.
addressZipcode	ZIP code for the property's location.
isUndisclosedAddress	Boolean indicating whether the full address is undisclosed.
beds	Number of bedrooms in the property.
baths	Number of bathrooms in the property.

area	Total living area in square feet.
latLong	Object containing latitude and longitude coordinates of the property.
latitude	Latitude coordinate of the property.
longitude	Longitude coordinate of the property.
isZillowOwned	Boolean indicating whether Zillow owns the property.
variableData	Additional data related to the listing
type	Type of the additional data.
text	Textual representation of the additional data.
hdpData	Nested object containing additional property details.
homeInfo	Nested object containing detailed property information, including:
zpid	Duplicate of the property ID.
streetAddress	Full street address of the property.
zipcode	ZIP code for the property's location.
city	City where the property is located.
state	State abbreviation (e.g., "TX").
latitude	Latitude coordinate of the property.
longitude	Longitute coordinate of the property
price	Unformatted price of the property in numeric form.
bathrooms	Number of bathrooms.
bedrooms	Number of bedrooms.
livingArea	Total living area in square feet.
homeType	Type of the property.
homeStatus	Status of the property.
daysOnZillow	Number of days the property has been listed on Zillow.
isFeatured	Boolean indicating whether the property is featured.
shouldHighlight	Boolean indicating if the property should be highlighted.
zestimate	An estimated market price calculated by Zillow.
rentZestimate	Estimated monthly rental value of the property calculated by Zillow.
listing_sub_type	Object containing sub-type information about the listing
is_FSBA	Boolean indicating "For Sale by Agent."
isUnmappable	Boolean indicating if the property cannot be mapped geographically.
isPreforeclosureAuction	Boolean indicating if the property is part of a preforeclosure auction.
homeStatusForHDP	Status of the property, specific to the HDP system.
priceForHDP	Price of the property, specific to the HDP system.
timeOnZillow	Timestamp of when the property was listed on Zillow.
isNonOwnerOccupied	Boolean indicating whether the property is not owner-occupied.
isPremierBuilder	Boolean indicating if the listing is from a premier builder.

isZillowOwned	Duplicate of the isZillowOwned field in the main dataset.
currency	Currency code used for the property's price (e.g., "USD").
country	Country code where the property is located (e.g., "USA").
taxAssessedValue	Tax-assessed value of the property.
lotAreaValue	Lot area of the property in numeric form.
lotAreaUnit	Unit of measurement for the lot area (e.g., "sqft").
isShowcaseListing	Boolean indicating whether the property is showcased.
isSaved	Boolean indicating if the property is saved by a user.
isUserClaimingOwner	Boolean indicating if the user claims to be the property owner.
isUserConfirmedClaim	Boolean indicating if the user's ownership claim is confirmed.
pgapt	Page category for the listing.
sgapt	Subcategory of the page.
zestimate	Zestimate value, an estimated market price for the property.
shouldShowZestimateAsPrice	Boolean indicating if the Zestimate value should be displayed as the property price.
has3DModel	Boolean indicating if a 3D model is available for the property.
hasVideo	Boolean indicating if a video is available for the property.
isHomeRec	Boolean indicating if the property is recommended to users based on their preferences.
hasAdditionalAttributions	Boolean indicating if the listing has additional attributions or disclaimers.
isFeaturedListing	Boolean indicating if the listing is marked as featured on the platform.
isShowcaseListing	Boolean indicating whether the listing is a showcase listing.
list	Boolean indicating whether the property is part of an active listing.
relaxed	Boolean indicating whether the listing has relaxed criteria or search filters applied.
brokerName	Name of the real estate agent or agency responsible for the listing.
carouselPhotos	Collection of property images, typically used for visualizing the property in a photo carousel.

3 Queries

3.1 Top 10 zip codes with the highest total price of homes

```
db.real_estate.aggregate([
{
    $group: {
        _id: "$addressZipcode",
        totalPrice: { $sum: "$unformattedPrice" },
        totalHomes: { $sum: 1 }
    }
},
{ $sort: { totalPrice: -1 } }
])
```

Figure 1: Top 10 zip codes with the highest total price of homes (partial)

3.2 Average price of properties for each Zip code

```
db.real_estate.aggregate([
{
    $group: {
        _id: "$addressZipcode",
        averagePrice: { $avg: "$unformattedPrice" },
        totalHomes: { $sum: 1 }
    }
},
{
    $sort: { averagePrice: -1 }
```

```
    },
    {
        $project: {
        _id: 0,
        zip_code: "$_id",
        averagePrice: 1,
        totalHomes: 1
        }
    }
])
```

Figure 2: Average price of properties for each Zip code(partial)

3.3 Listings with more than 200 photos

```
    db.real_estate.aggregate([
    {
        $unwind: {path: "$carouselPhotos" }
    },
    {
        $group: {
            _id: "$zpid",
            photoCount: { $sum: 1 }
        }
    },
    { $match: { photoCount: { $gt: 200 } } },
    { $count: "listingsWithMoreThan200Photos" }
])
```

Figure 3: Listings with more than 200 photos

3.4 Listings per broker

```
db.real_estate.aggregate([
    {
        $match:
        { brokerName: { $exists: true, $ne: null } }
    },
    {
        $group: {
            _id: "$brokerName",
            totalListings: { $sum: 1 }
        }
    },
    { $sort: { totalListings: -1 } }
])
```

Figure 4: Listings per broker (partial)

3.5 Properties with price closest to the estimate made by Zillow

```
db.properties.aggregate([
    {
        $project: {
            _id: 0,
            address: 1,
            unformattedPrice: 1,
            zestimate: 1,
            priceDifference: {
              $abs: {
                $subtract: [
```

```
                    "$unformattedPrice",
                    "$zestimate"]
    }
  }
}
    },
    {

    },
    { $group: { _id: null, avgDiff: { $avg: "$diffPercentage" } } }
]);
```

```
unformattedPrice : 309000
address : "15607 Dawnbrook Dr, Houston, TX 77068"
zestimate : 309000
priceDifference : 0

unformattedPrice : 324900
address : "15019 River Park Dr, Houston, TX 77070"
zestimate : 324900
priceDifference : 0

unformattedPrice : 197900
address : "1300 Magnolia St, Baytown, TX 77520"
zestimate : 197900
priceDifference : 0

unformattedPrice : 198500
address : "12203 Northcliffe Manor Dr, Houston, TX
          77066"
zestimate : 198500
priceDifference : 0
```

Figure 5: Properties with price closest to the estimate made by Zillow (partial)

3.6 Average number of days on Zillow of properties with a price between 750,000$ and 1,250,000$ and at least 3 bedrooms

```
db.properties.aggregate([
    {
        $match: {
          "hdpData.homeInfo.price": {
            $gte: 750000,
            $lte: 1250000
          },
          "hdpData.homeInfo.bedrooms": {$gte: 3}
          }
    },
    { $group: {
```

```
            _id: null,
            averageDaysOnZillow: {
              $avg: "$hdpData.homeInfo.daysOnZillow"
            },
            properitesCount: {
              $sum: 1
            }
    }
    },
    {
        $project: {
            _id: 0,
            averageDaysOnZillow: 1,
            properitesCount: 1
        }
    }
]);
```

```
averageDaysOnZillow : 59.1461061337009
properitesCount : 1451
```

Figure 6: Average number of days on Zillow of properties with a price between 750,000$ and 1,250,000$ and at least 3 bedrooms

3.7 Average price of properties with both 3D model and video

```
db.properties.aggregate([
    {
        $match: {
          has3DModel: true,
          hasVideo: true
        }
},
    {
        $group: {
          _id: null,
          avgPrice: {
            $avg: "$unformattedPrice"
          }
        }
    },
    {
        $project: {
          _id: 0,
          avgPrice: 1
```

```
        }
    }
]);
```

```
avgPrice : 1555850
```

Figure 7: Average price of properties with both 3D model and video

3.8 Most expensive property by home type

```
db.properties.aggregate([
    {
        $match: {
          "hdpData.homeInfo.homeType": {
            $ne: null
          }
        }
    },
    {
        $sort: {
            unformattedPrice: 1
        }
    }
    {
        $group: {
          _id: "$hdpData.homeInfo.homeType",
          mostExpensiveHome: {
            $first: "$$ROOT"
          }
        }
    },
    {
        $project: {
          _id: 0,
          homeType: "$_id",
          price: "$mostExpensiveHome.unformattedPrice"
        }
    }
]);
```

Figure 8: Most expensive property by home type

3.9 Home type distribution

```
db.properties.aggregate([
    {
        $group: {
          _id: "$hdpData.homeInfo.homeType",
          totalCount: {
            $sum: 1
          }
        }
    },
    {
        $project: {
          _id: 0,
          homeType: "$_id",
          totalCount: 1
        }
    },
    {
        $sort: {
          totalCount: -1
        }
    }
]);
```

Figure 9: Home type distribution

3.10 Largest total area by zip code

```
db.properties.aggregate([
    {
        $group: {
          _id: "$addressZipcode",
          avgLotSize: {
            $avg: "$hdpData.homeInfo.livingArea"
          },
          // Calculate the average lot size
          totalLotSize: {
            $sum: "$hdpData.homeInfo.livingArea"
          },
          propertiesCount: {
            $sum: 1
          }
        }
    },
    {
        $project: {
          _id: 0,
          zipCode: "$_id",
          address: "$largestArea.address",
          totalLotSize: 1,
          avgLotSize: 1,
          propertiesCount: 1
        }
    }
]);
```

```
avgLotSize : 2509.551401869159
totalLotSize : 268522
propertiesCount : 135
zipCode : "77581"

avgLotSize : 1408
totalLotSize : 111232
propertiesCount : 80
zipCode : "77036"

avgLotSize : 2241.4535519125684
totalLotSize : 410186
propertiesCount : 222
zipCode : "77338"

avgLotSize : 2034.7884615384614
totalLotSize : 105809
propertiesCount : 54
zipCode : "77014"
```


Figure 10: Largest total area by zip code (partial)

3.11 Homes within 5 miles from downtown Houston ()

```
db.real_estate.aggregate([
"match" : {
  "$expr": {
    "$lt": [
      {
        "$sqrt": {
          "$add": [
            {
              "$pow": [
                { "$subtract": ["$hdpData.homeInfo.latitude", 29.7604] },
                2
              ]
            },
            {
              "$pow": [
                { "$subtract": ["$hdpData.homeInfo.longitude", -95.3698] },
                2
              ]
            }
          ]
        }
      },
      0.0725
    ]
  }
  }
]);
```

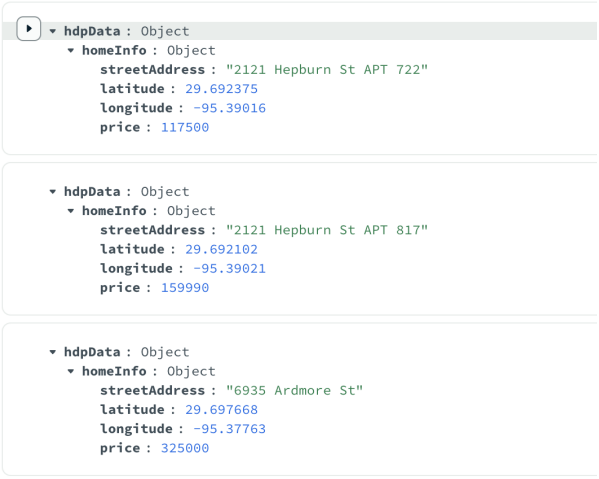

Figure 11: Homes within 5 miles from downtown Houston (partial)

3.12 Total property prices in the 5 most expensive neighborhoods

```
db.real_estate.aggregate([
  {
    "$group": {
      "_id": "$addressZipcode",
      "totalPrice": { "$sum": "$unformattedPrice" },
      "propertyCount": { "$sum": 1 }
    }
  },
  {
    "$sort": { "totalPrice": -1 }
  },
  {
    "$limit": 5
  }
]);
```

Figure 12: Total property prices in the 5 most expensive neighborhoods

3.13 The 10 real estate agents with the most listings

```
db.real_esate.aggregate([
  {
    "$group": {
      "_id": "$brokerName",
      "totalListings": { "$sum": 1 }
    }
  },
  {
    "$sort": { "totalListings": -1 }
  },
  {
    "$limit": 10
  }
]);
```

```
_id: "eXp Realty LLC"
totalListings : 1127

_id: "Nan & Company Properties"
totalListings : 385

_id: "Compass RE Texas, LLC - Houston"
totalListings : 328

_id: "HomeSmart"
totalListings : 322

_id: "Keller Williams Realty Metropolitan"
totalListings : 291

_id: "Martha Turner Sotheby's International Realty"
totalListings : 277

_id: "JLA Realty"
totalListings : 269

_id: "Walzel Properties - Corporate Office"
totalListings : 263

_id: "Realty Associates"
totalListings : 251
```


Figure 13: The 10 real estate agents with the most listings

3.14 Find homes with a price reduced by 25% from the original value

```
db.properties.aggregate({
  "$expr": {
    "$lt": [
      "$unformattedPrice",
      { "$multiply": ["$hdpData.homeInfo.originalPrice", 0.75] }
    ]
  }
});
```

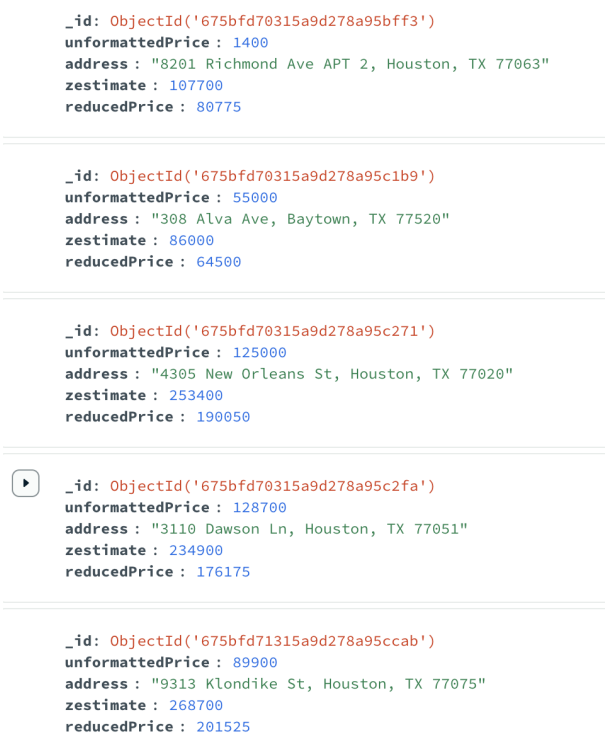

Figure 14: Find homes with a price reduced by 25% from the original value (partial)

3.15 Analyze the distribution of homes with more than 5 bedrooms in each zip code

```
db.real_estate.aggregate([
  {
    "$match": { "beds": { "$gt": 5 } }
  },
  {
    "$group": {
      "_id": "$addressZipcode",
      "count": { "$sum": 1 }
    }
  },
  {
    "$sort": { "count": -1 }
  }
]);
```

Figure 15: Analyze the distribution of homes with more than 5 bedrooms in each zip code (partial)

3.16 Calculate the average price for luxury homes ($price greater than $1,000,000) by property type

```
db.real_estate.aggregate([
  {
    "$match": { "unformattedPrice": { "$gt": 1000000 } }
  },
  {
    "$group": {
      "_id": "$hdpData.homeInfo.homeType",
      "averageLuxuryPrice": { "$avg": "$unformattedPrice" },
      "luxuryCount": { "$sum": 1 }
    }
  },
  {
    "$sort": { "averageLuxuryPrice": -1 }
  }
]);
```

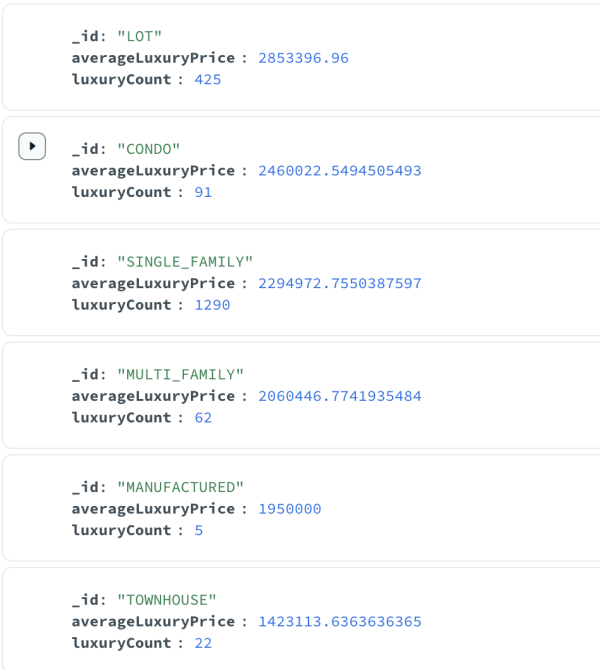

Figure 16: Calculate the average price for luxury homes ($price greater than $1,000,000) by property type

3.17 Find homes with at least 6 bedrooms and more bathrooms than bedrooms.

```
db.properties.aggregate({
```

```
  "$and": [
    { "beds": { "$gte": 6 } },
    {
      "$expr": {
        "$gt": ["$baths", "$beds"]
      }
    }
  ]
});
```

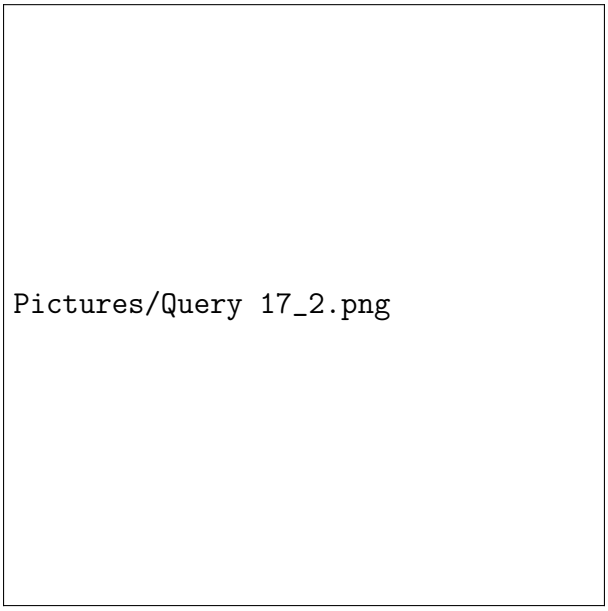

Figure 17: Find homes with at least 6 bedrooms and more bathrooms than bedrooms (partial)

3.18 Calculate the average time properties stay on the market for each agent

```
db.properties.aggregate([
  {
    "$group": {
      "_id": "$brokerName",
      "averageDaysOnMarket": { "$avg": "$hdpData.homeInfo.daysOnZillow" },
      "propertyCount": { "$sum": 1 }
    }
  },
  {
    "$sort": { "averageDaysOnMarket": 1 }
  }
]);
```

Figure 18: Calculate the average time properties stay on the market for each agent

3.19 Calculate the maximum and minimum price in each zip code

```
db.properties.aggregate([
 {
   "$group": {
     "_id": "$addressZipcode",
     "maxPrice": { "$max": "$unformattedPrice" },
     "minPrice": { "$min": "$unformattedPrice" },
     "propertyCount": { "$sum": 1 }
   }
 },
 {
   "$sort": { "maxPrice": -1 }
 }
```

```
]);
```

Figure 19: Calculate the maximum and minimum price in each zip code (partial)

3.20 Find properties with an estimated rent (rentZestimate) greater than 10% of their value

```
db.properties.aggregate([
 {
 "$expr": {
   "$gt": [
     "$hdpData.homeInfo.rentZestimate",
     { "$multiply": ["$unformattedPrice", 0.1] }
   ]
 }
} ]);
```