Table of Contents

Resources Link – PowerBI File and Scripts

Old Dashboard : https://dms.uom.lk/s/jZfWqsSfT8XdJRX

Updated Dashboard: https://drive.google.com/file/d/1m7TM3RphXM2tX2LS5mBjK4Nzj8K5jv-H/view?usp=sharing

Kaggle Dataset Link

Zillow Prize: Zillow's Home Value Prediction (Zestimate) | Kaggle

Databricks Repository Link –

KavinduUoM20/zillow-bi (github.com)

1. Introduction

In today's data dependent real estate market, home value estimation is one of the significant tasks for both homeowners and investors. A detailed analysis of “Zillow Zestimate Dataset”, obtained from Kaggle which is a rich source with abundant data on US property values has been the main focus during this project. The dataset contains crucial information from 7.5 million machine learning models and each property covers hundreds of data points. Our aim is to unveil meaningful and valuable insights which inform on smart investments in real estate and guide in strategic decision making.

To deal with this large dataset, a combination of several tools; Azure Data Lake for scalable data storage, Databricks for data processing and transformation, SQL Data Warehouse for structured data management, and Power BI for visualization have been effectively employed. The above-mentioned tools were utilized for the effectivity in handling big data, ensuring data integrity and offering insightful visualizations.

The main objective of this comprehensive analysis is to deliver meaningful insights that can benefit in the real estate industry. It is our aim to guide investors in making informed decisions by assisting them in identifying undervalued properties, predicting future price trends, and analyzing growth in neighborhood. These insights could be useful in maximizing returns for investors in a highly competitive market along with the potential in creating novel business opportunities in the real estate sector.

Figure 1: Zillow Official Website

2. Dataset Selection

The selected “Zillow Home Value Prediction (Zestimate)” Dataset exceeds 1.37 GB in size and consists of 2,985,217 rows and 58 columns. This dataset is known to be one of the largest publicly available real estate datasets and it is a source with comprehensive details about the property historical price data, and a variety of attributes like building type, tax assessments, and details on neighborhood. The particular dataset was selected due to its thorough coverage of the United States housing industry and its potential in producing valuable insights for investments in real estate.

This dataset consists of a variety of attributes explaining various aspects of properties, such as.

- parcelid: Unique identifier for parcels (lots)
- airconditioningtypeid: Type of cooling system present in the home (if any)
- architecturalstyletypeid: Architectural style of the home (i.e. ranch, colonial, split-level, etc.)
- basementsqft: Finished living area below or partially below ground level.
- bathroomcnt: Number of bathrooms in the home including fractional bathrooms.
- bedroomcnt: Number of bedrooms in home
- buildingclasstypeid: The building framing type (steel frame, wood frame, concrete/brick)
- buildingqualitytypeid: Overall assessment of condition of the building from best (lowest) to worst (highest)
- calculatedbathnbr: Number of bathrooms in the home including frictional bathroom.
- decktypeid: Type of deck (if any) present on parcel = 66

Moreover, it also covers attributes for financial and structural aspects of the properties such as.

- numberofstories: Number of stories or levels the home has
- Fireplaceflag: Is a fireplace present in this home.
- Structuretaxvaluedollarcnt: The assessed value of the built structure on the parcel
- Taxvaluedollarcnt: The total tax assessed value of the parcel
- Assessmentyear: The year of the property tax assessment
- Landtaxvaluedollarcnt: The assessed value of the land area of the parcel
- Taxamount: The total property tax assessed for that assessment year.
- Taxdelinquencyyear: The total tax assessed value of the parcel
- Censustractandblock: Census tract and block ID combined – also contains blockgroup assignment by extension

The diversity of the attributes enables for an extensive analysis of housing trends, which is highly significant to the objective of generating business value using insights from real estate analysis.

The size of the dataset ensures the adequacy of information to identify trends and create reliable models that has the potential to be applied in various properties and geographical locations, Moreover, an extensive assessment is attainable with the wide range of data available, ranging from identifying undervalued properties to forecasting future trends by employing historical and geographical data.

It has been determined to deliver practical insights that allows for better decision-making in investments in real estate, promoting the business idea of offering specialized consultation services for high-return property investments.

3. Data Analysis

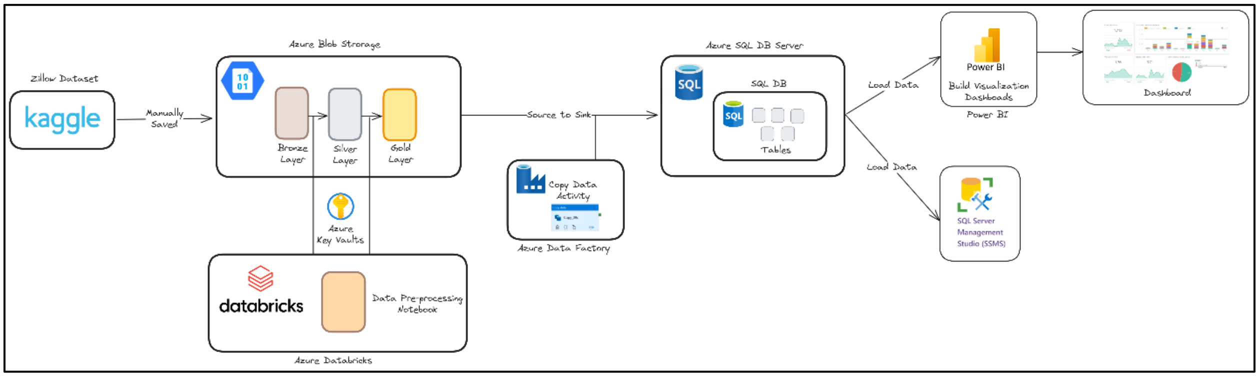


Figure 2: Data Engineering Architecture

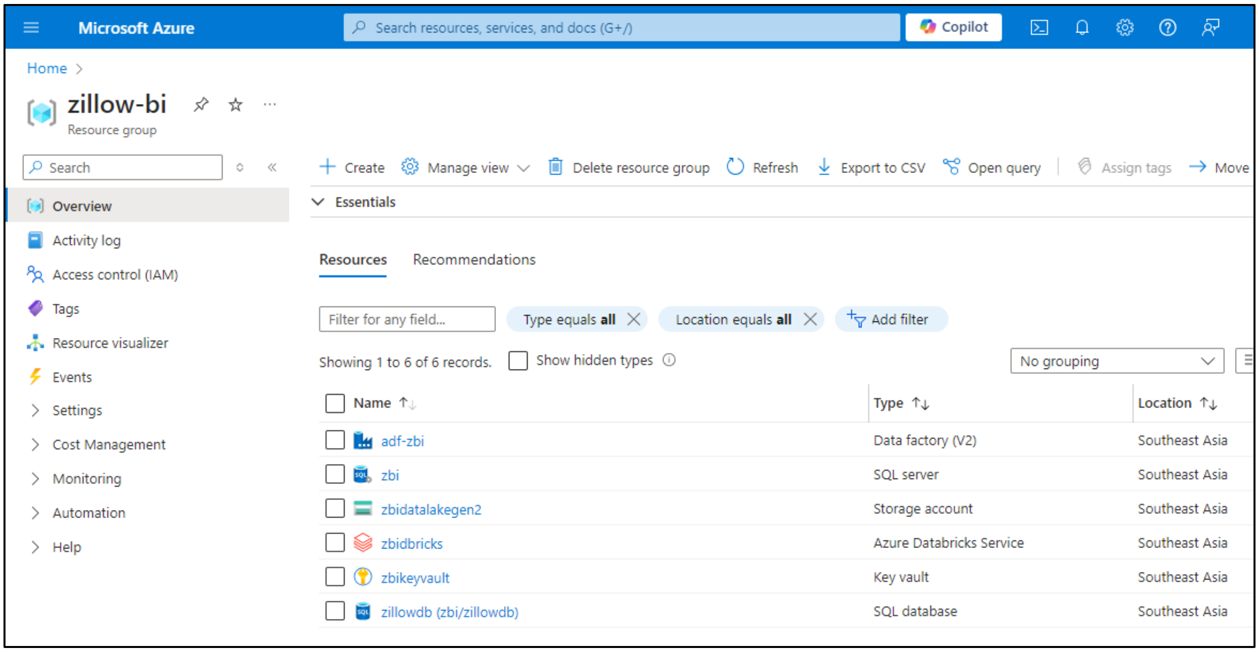
3.1 Data Acquisition

The dataset Zillow Prize: Zillow's Home Value Prediction (Zestimate) was obtained from the Kaggle platform. The dataset size is 1.3 GB and initially comprised of 58 columns.

3.2 Setting Up the Infrastructure

Following infrastructure was utilized for the project.

- Azure Data Lake (Cloud Storage) - For storing the initially obtained raw version of data.
- Azure Data Bricks (Data Processing Platform) - For Data Transformation and analysis
- SQL Warehouse (Data Warehouse) - To store the processed data
- PowerBI (Business Intelligence Tool) - To create visualizations and Dashboards.



Showing 1 to 6 of 6 records.

Name	Type	Location
adf-zbi	Data factory (V2)	Southeast Asia
zbi	SQL server	Southeast Asia
zbidatalakegen2	Storage account	Southeast Asia
zbidbricks	Azure Databricks Service	Southeast Asia
zbikeyvault	Key vault	Southeast Asia
zillowdb (zbi/zillowdb)	SQL database	Southeast Asia

Figure 3: Azure Resource Group

3.3 Data Ingestion and Storage

The dataset was uploaded to Azure Data Lake Storage, which provided a centralized repository for the raw data. Permissions and security measures were configured to ensure that data integrity and controlled access to members.

3.4 Data Cleaning and Transformation (Bronze to Silver)

3.4.1 Data Cleaning Process

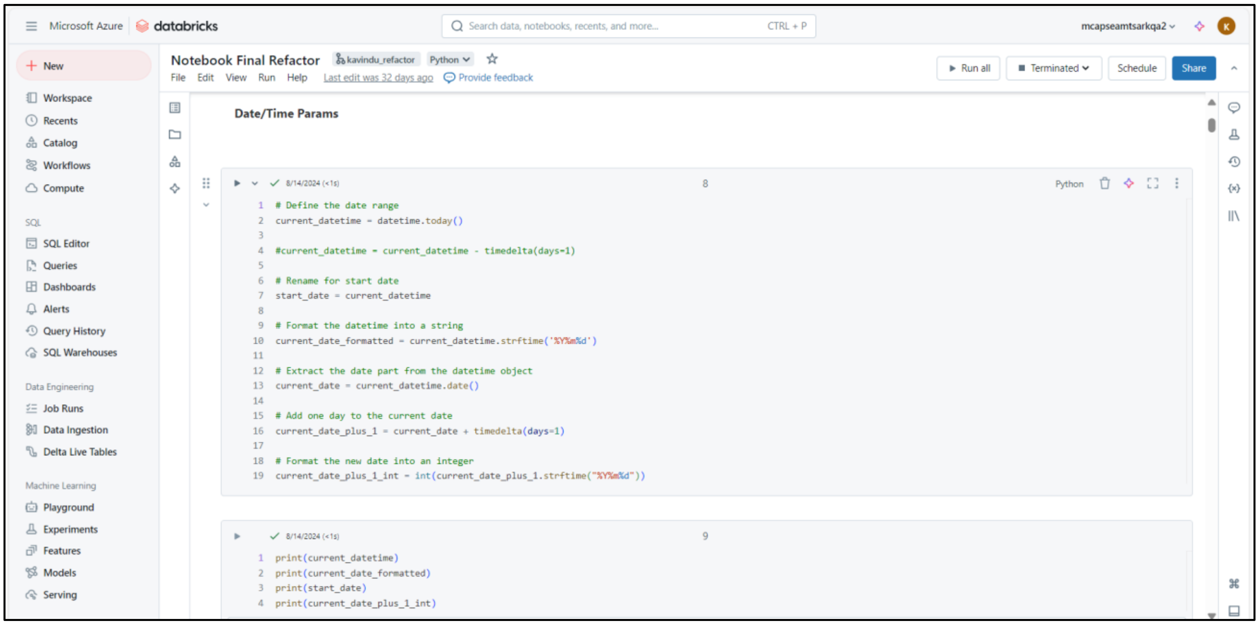
- Loading Data: The dataset was loaded from a CSV file into a Data Frame using Pandas. Removing Unwanted Columns: Specific columns deemed unnecessary for analysis were dropped to streamline the dataset.
- Handling Missing Values:
    - Null values in certain columns were filled with zero or a specific value based on the column's context.
    - For instance, columns related to pool counts and air conditioning types had their nulls filled with logical defaults.
- Inconsistency Checks: The dataset was checked for inconsistencies, such as cases where pool counts were zero while other pool-related columns had non-zero values. Rows with such inconsistencies were removed.

- Final Cleaning: The DataFrame was further refined to ensure no missing values remained, resulting in a cleaned dataset ready for analysis.

3.4.2 Data Division into Tables

The cleaned dataset was then divided into several normalized tables to enhance organization and facilitate analysis:

1. Property Table: Contained information about property attributes such as parcel ID, location, and size.
2. Building Table: Included details on building specifications like bathroom and bedroom counts, and features like air conditioning.
3. Building Quality Table: Defined the quality of buildings using a mapping of building quality IDs to descriptive terms.
4. Air Conditioning Table: Mapped air conditioning type IDs to their respective descriptions.
5. Heating System Table: Similar to the air conditioning table, this mapped heating system type IDs to descriptions.
6. Pool Type Table: Provided a mapping of pool type IDs to their descriptions.
7. Pool Table: Contained information on pool counts and sizes.
8. Region Table: Included geographical identifiers for properties.
9. Tax Table: Captured tax-related information for each property.
10. Yard Table: Focused on yard-related attributes.



```python
# Define the date range
current_datetime = datetime.today()

#current_datetime = current_datetime - timedelta(days=1)

# Rename for start date
start_date = current_datetime

# Format the datetime into a string
current_date_formatted = current_datetime.strftime('%Y%m%d')

# Extract the date part from the datetime object
current_date = current_datetime.date()

# Add one day to the current date
current_date_plus_1 = current_date + timedelta(days=1)

# Format the new date into an integer
current_date_plus_1_int = int(current_date_plus_1.strftime("%Y%m%d"))
```

```python
print(current_datetime)
print(current_date_formatted)
print(start_date)
print(current_date_plus_1_int)
```

Figure 4: Azure Databricks

3.5 Data Modelling and Preparation for Data Warehouse (Silver to Gold)

Here the cleaned data was loaded into the SQL warehouse after mapping relationships among tables and setting primary and foreign keys as applicable.

Data Modelling

- Established relationships between tables by identifying primary keys and foreign keys

Converting data for SQL

- Converting the data to an SQL compatible format using Data Bricks' integrated tools, ensuring seamless integration with SQL Warehouse.

The integrity of relationships and data consistency was validated post-conversion to ensure readiness for analytics

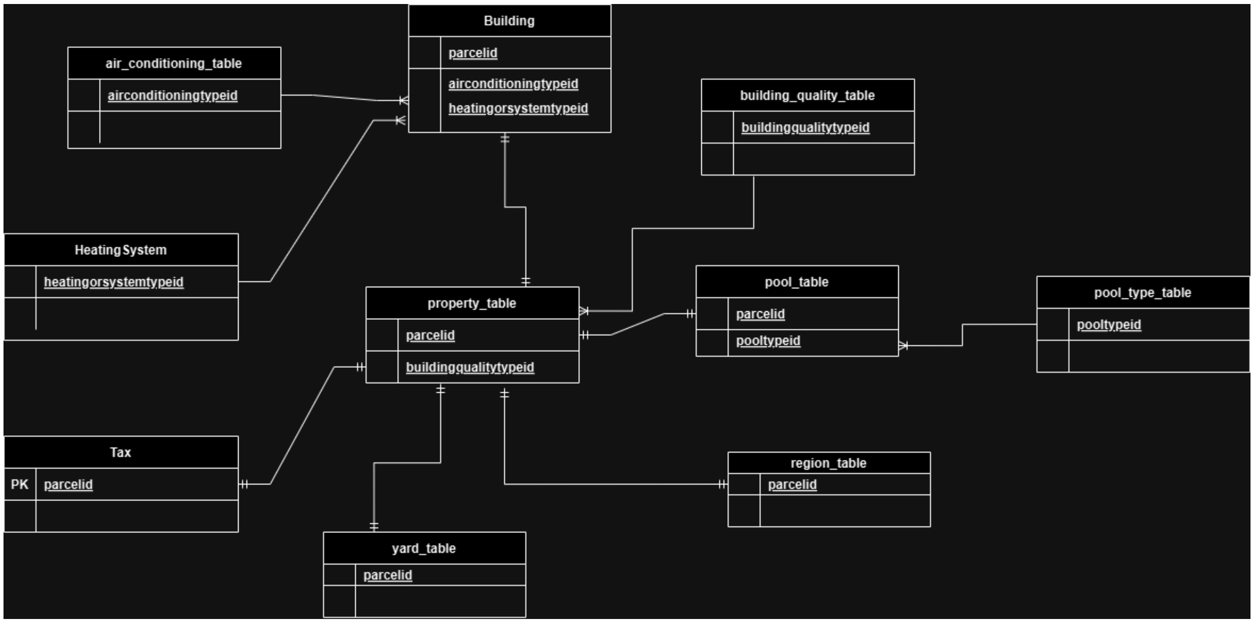


Figure 5: ER Diagram

3.6 Loading the Data into the SQL Data warehouse

The prepared data was loaded into Azure SQL Warehouse using [methods/tools used, e.g., Data Bricks' JDBC connector]. Batch processing was scheduled to handle the large volume of data efficiently, ensuring that the warehouse was optimized for query performance.

3.7 Data Analysis and Visualization

After loading the cleaned and transformed data into the SQL Warehouse, we connected PowerBI to the warehouse to create interactive dashboards showcasing key metrics and insights.

Below outlines the structure and components of the real estate investment dashboard, designed to provide users with an interactive and insightful overview of property characteristics, building features, investment potential, and more. Each page is equipped with KPIs, visualizations, and filters to help users easily navigate through the data and make informed decisions.

Page 1: Home

Purpose: Serves as a central hub for navigation to various sections of the dashboard.

Key Components:

- Navigation Buttons: Direct links to five distinct pages for easy access.

Default Slicers:

- Year Built: Filter data based on the year properties were constructed.
- Region Zip: Filter data by geographical zip codes.

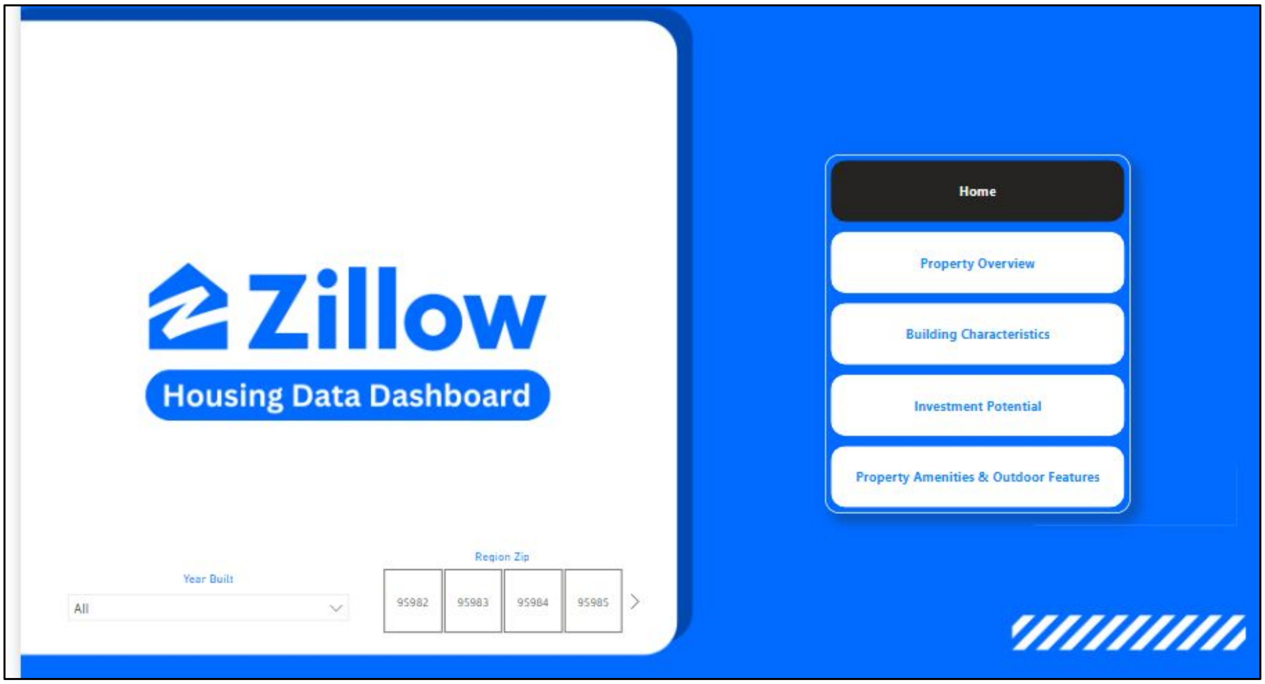


Figure 6: PowerBI Dashboard - Home Page

Page 2: Property Overview

Purpose: Provide insights into property characteristics and performance.

Key Components:

Cards:

- Average Property Value: Calculated as the total property value divided by the number of properties.

```
AveragePropertyValue

= SUM ('9 tax_table' [taxvaluedollarcnt] ) / COUNT ('1 property_table' [parcelid] )
```

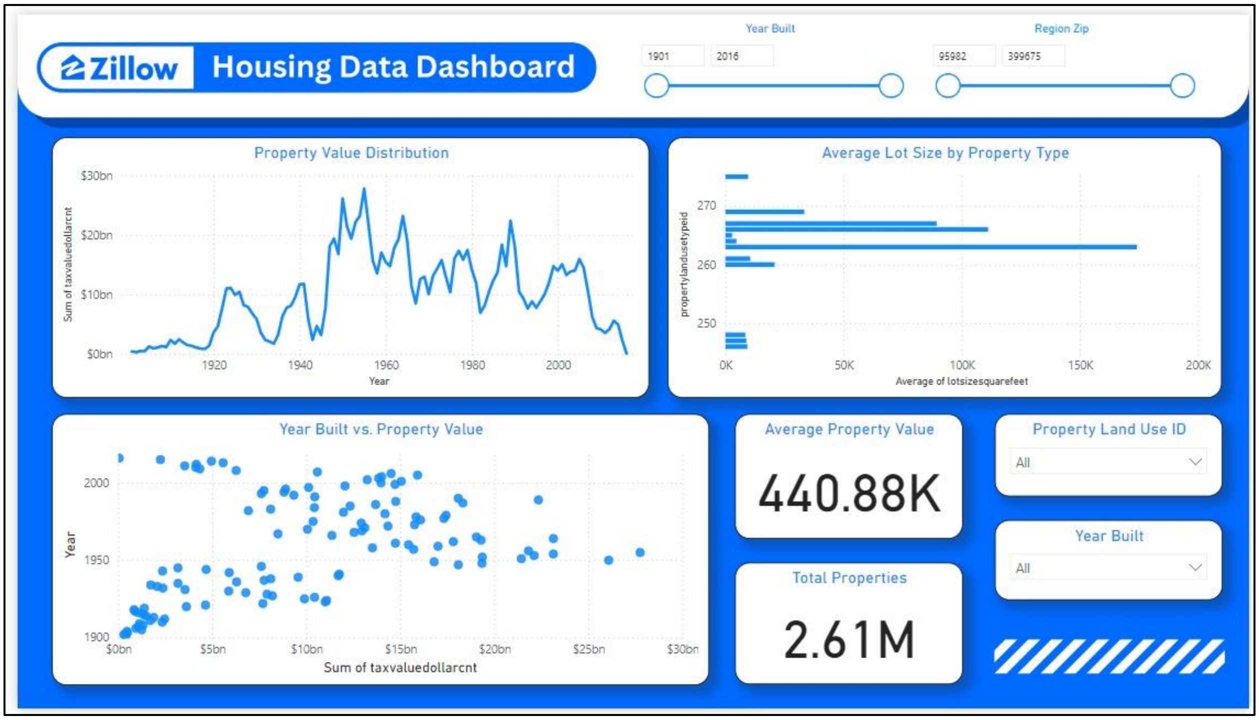
- Total Properties: Displays the total count of properties based on unique parcel IDs.

Charts:

- Property Value Distribution: A line chart illustrating the distribution of property values over time.
- Year Built vs. Property Value: A scatter plot comparing the year built against the property value.
- Average Lot Size by Property Type: A bar chart showing average lot sizes categorized by property type.

Filters:

- Property Type: Filter data based on different property land use types.
- Year Built: Filter data by the year properties were constructed.



Year Built: 1901 – 2016

Region Zip: 95982 – 399675

Average Property Value: 440.88K

Total Properties: 2.61M

Property Land Use ID: All

Year Built: All

Figure 7: PowerBI Dashboard - Properties Overview Page

Page 3: Building Characteristics

Purpose: Analyzes specific attributes related to buildings.

Key Components:

Cards:

- Average Number of Rooms: Calculated as the total number of rooms divided by the number of properties.
- Average Building Quality: Based on building quality ratings.

Charts:

- Building Quality Distribution: An area chart showing the distribution of building quality ratings.
- Bedroom Count vs. Property Value: A line chart illustrating the relationship between bedroom count and property value.
- Bathroom Count vs. Property Value: A scatter plot comparing bathroom count against property value.
- Air Conditioning Type: A pie chart displaying the distribution of air conditioning types.
- Building Quality: A pie chart showing the distribution of building quality descriptions.

Filters:

- Air Conditioning Type: Filter data based on air conditioning types.
- Heating System Type: Filter data based on heating system types.

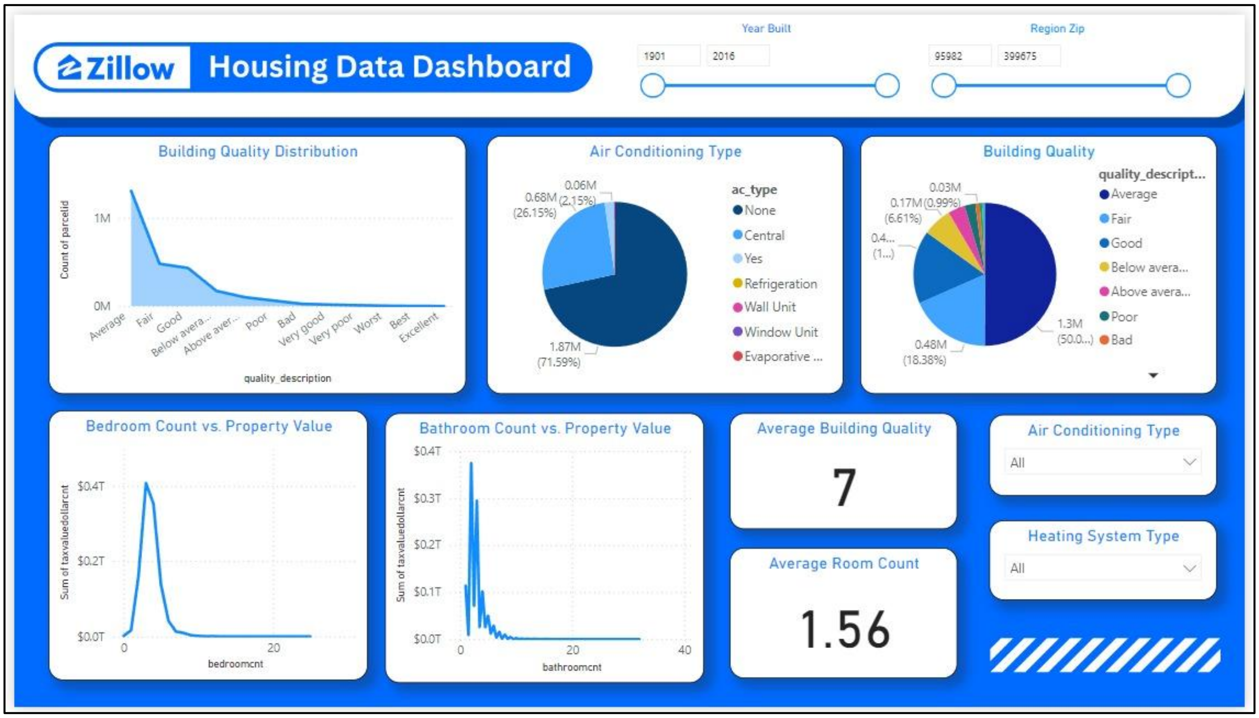


Figure 8: PowerBI Dashboard - Building Characteristics Page

Page 4: Investment Potential

Purpose: Evaluates areas with high investment potential.

Key Components:

Cards:

Average_Growth_Rate = 0.05

Estimated_Future_Value = AVERAGE ('9 tax_table' [taxvaluedollarcnt]) * (1 + [Average_Growth_Rate])

ROI = ([Estimated_Future_Value] – AVERAGE ('9 tax_table' [taxvaluedollarcnt])) / AVERAGE ('9 tax_table' [taxvaluedollarcnt]) * 100

- Highest Potential ROI: Calculated based on historical growth trends and current property values.
- Number of Investment Opportunities: Count of properties with significant appreciation potential.

Charts:

- Structure and Land Tax Breakdown: A stacked bar chart showing the breakdown of structure and land tax values.
- Property Value Based on Pools: A column chart illustrating property values categorized by pool types.
- Property Value Based on Heating System: A column chart showing property values based on heating system types.
- Value Based on Year Built: A line chart displaying property value trends over the years.
- ROI by Region: A bar chart showing potential ROI across different regions.
- Historical Property Value Growth: A line chart depicting property value growth over time.

Filters:

- County ID: Filter data based on county identifiers.
- City ID: Filter data based on city identifiers.

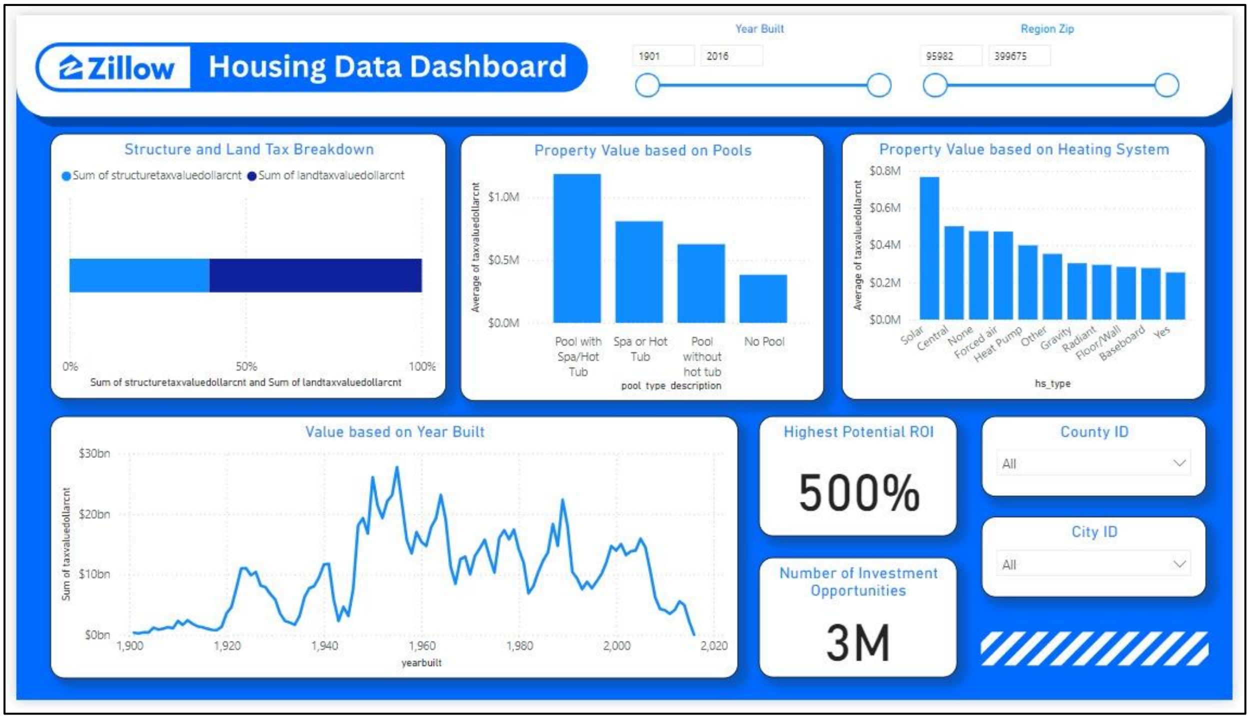


Figure 9: PowerBI Dashboard - Investment Potential Page

Page 5: Location-based Property Insights

Purpose: Evaluates properties and their features based on the location.

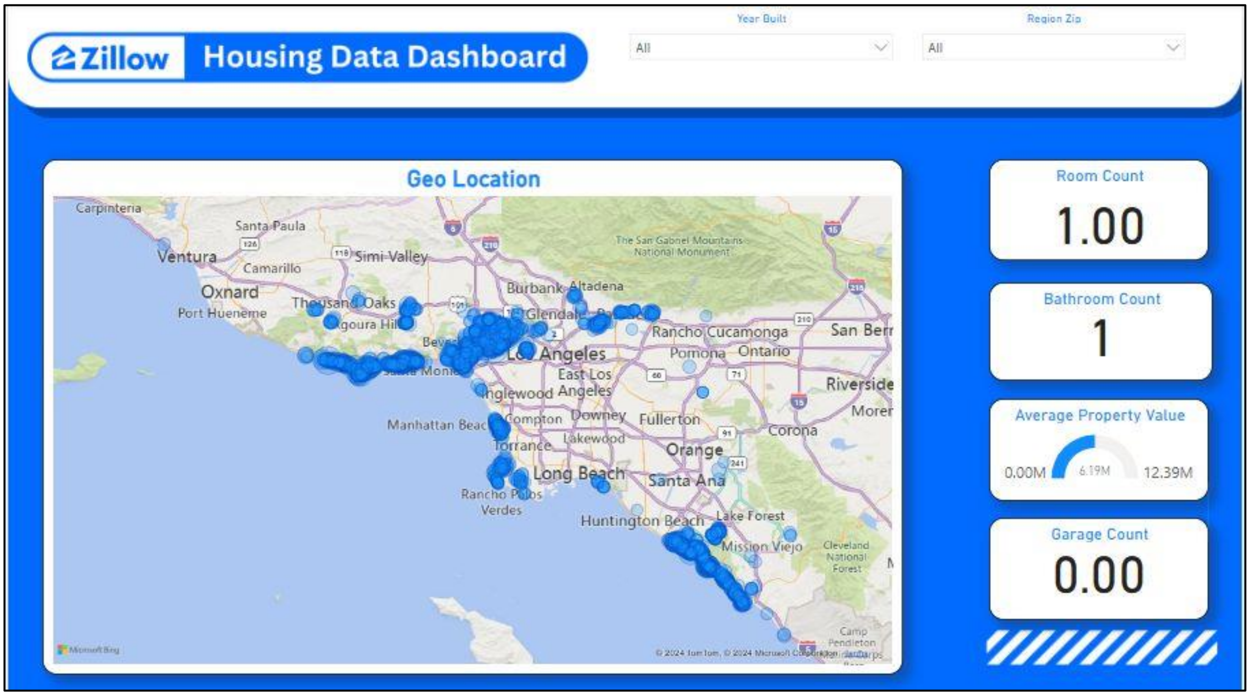
Key Components:

Cards:

- Room Count: Count of Rooms in the selected property on the map.
- Bathroom Count: Count of Bathrooms in the selected property on the map.
- Garage Count: Count of Garages in the selected property on the map.

Charts:

- Map: Shows the Geo locations of each property
- Gauge Chart: Shows the average property value in the selected properties.



Zillow Housing Data Dashboard

Year Built: All

Region Zip: All

Geo Location

Room Count: 1.00

Bathroom Count: 1

Average Property Value: 0.00M 6.19M 12.39M

Garage Count: 0.00

Figure 10: PowerBI Dashboard – Property Features & Location Page

4. Details of Infrastructure

The Zillow BI project makes use of a strong infrastructure that is based on multiple Azure services, guaranteeing smooth data storage, transformation, ingestion, and visualization. **Azure Blob Storage**, which serves as a data lake to store the Zillow dataset at different phases of processing, is the central component of our system. **Bronze, Silver, and Gold** are the three separate levels into which the data is arranged. The unfiltered, raw data that comes from Kaggle is stored in the Bronze layer. The initial cleaning and transformation of the data is stored in the Silver layer, and the final processed dataset, which is prepared for reporting and analysis, is stored in the Gold layer. Data integrity is guaranteed by this methodical approach to data management, which also makes it easier for downstream operations to retrieve data quickly.

One of the main components of the data pipeline automation is **Azure Data Factory (ADF)**. It manages the data transmission to Azure SQL Database and coordinates data migration between various storage tiers. ADF makes sure that data moves through the pipeline without human involvement by using Copy Data Activities, which makes it possible to create a dependable and repeatable procedure. Maintaining a scalable and stable pipeline that can handle different data volumes requires this automation.

Azure Databricks, the engine that powers our infrastructure's data transformation and preprocessing, is a key element. We took advantage of **Apache Spark's** capacity to process massive datasets in parallel in Databricks to analyze the Zillow dataset in a distributed manner. Multiple worker nodes in the **Databricks cluster** employed for the project ensure effective completion of feature engineering, data transformation, and cleaning activities. In addition, we used **Databricks notebooks** to execute our data processing scripts, utilizing **Delta Lake** for effective data version management and storage across the pipeline.

Azure Key Vault was used to store and manage secrets like connection strings and API keys for security and sensitive data management. This lowers the possibility of unwanted access or data breaches by ensuring that sensitive credentials are not hardcoded into scripts but are instead accessible safely.

The project's structured repository, an **Azure SQL Database**, is where the processed data is eventually kept. The final Gold layer data, which is ready for analytics and querying, is stored in this database. We used **SQL Server Management Studio (SSMS)**, which offered an intuitive user interface for managing database schemas, running SQL queries, and guaranteeing data integrity, to administer and query this database.

Power BI, which builds dynamic reports and dashboards by connecting directly to the Azure SQL Database, comprises the last layer of the architecture. With the use of Power BI's visualization features, users may extract valuable insights from the processed Zillow dataset, assisting in the identification of important real estate trends and other information. The data is always current and available for reporting thanks to the connection between Power BI and Azure SQL Database.

All things considered, the architecture employed for the Zillow BI project makes use of **Databricks'** sophisticated data processing capabilities to effectively manage large-scale data transformations, as well as the scalability, security, and flexibility of Azure's cloud services.

5. Challenges Faced during the Project

Several challenges were encountered during the Zillow BI Project that required thorough solutions and adjustments to the workflow.

1. **Data Cleaning and Preprocessing:**

   Managing the inconsistent nature of the unprocessed Zillow dataset was one of the main difficulties. Particularly in important columns like property prices, square footage, and dates, there were a lot of missing values, inaccurate data types, and outliers. Extensive data cleaning, including managing null values and type conversions, was necessary to ensure data quality during the Bronze to Silver layer shift. The magnitude of the dataset made certain conversions very time-consuming to complete, which made the process especially difficult.

2. **Performance and Optimization:**
   Performance became a major challenge as the dataset grew, particularly when executing intricate Databricks transformation processes. Initially, processing the tasks took a long time, which affected overall production. We used Apache Spark and Delta Lake optimizations, such as data segmentation and caching where appropriate, to get around issue. But throughout the project, striking a balance between the necessity of optimizations and accuracy was a constant issue.

3. **Integration Between Azure Services:**
   Thorough configuration was required to enable seamless connectivity across different Azure services, including Azure Data Factory (ADF), Azure Blob Storage, and Azure Databricks. It was challenging to make sure that various services cooperated without human interaction, especially with automated data pipelines. Tiny errors in ADF workflow setups could result in data flow delays, necessitating cross-service troubleshooting to find the problem's root cause.

4. **Scalability of the Databricks Cluster:**

   The Databricks cluster's scalability to handle different workloads created resource management issues. Scaling up was necessary for the cluster to prevent bottlenecks during times of strong demand, particularly when processing bigger chunks of the dataset. But this also meant higher costs, and careful monitoring and decision-making were needed to strike a balance between the budgetary limits and the distribution of resources.

5. **Security and Key Management:**
   Secure key management emerged as a problem to protect sensitive data, like authentication tokens and connection strings. Secrets were stored using Azure Key Vault, however integrating it with Databricks notebooks and other services without a hitch needed some work and a thorough understanding of Azure security procedures. Another element that required careful handling was distributing access control across various team members without jeopardizing security.

6. **Data Transformation Errors:**
   Complicated data conversions led to unforeseen problems throughout the Silver to Gold layer transition stages. These mistakes were frequently caused by mismatched schemas, improperly formatted data, or false presumptions made in the original transformation logic. In order to assure accuracy, debugging these problems meant going back to earlier pipeline stages, which took time.

7. **Power BI Visualization:**

   It was difficult to connect Power BI to the Azure SQL Database and guarantee real-time updates, particularly for sophisticated queries and bigger datasets. It took a lot of work to optimize the Power BI report and the SQL queries in order to make sure that the system could manage the amount of data without experiencing any hiccups when creating interactive and responsive dashboards.

6. Key Findings and Analysis

The analysis conducted on the dataset yielded several key findings that provide valuable insights into property characteristics, investment potential, and overall market trends. The results are summarized as follows:

1. **Average Property Value**: The analysis revealed that the average property value across the dataset is significantly influenced by factors such as location, property type, and building quality. This metric allows stakeholders to gauge the market value of properties in different regions, aiding in pricing strategies and investment decisions.
2. **Investment Potential:** The Investment Potential page highlighted areas with the highest potential return on investment (ROI). Properties with significant appreciation potential were identified, with a calculated ROI based on historical growth trends. This insight enables investors to target high-growth areas, optimizing their investment portfolios.

3. **Building Quality and Property Value Correlation:** The analysis showed a strong correlation between building quality and property value. Properties rated as "Excellent" or "Best" consistently commanded higher prices, suggesting that quality renovations and maintenance can lead to increased market value.
4. **Room Count Impact:** The scatter plots indicated that properties with more bedrooms and bathrooms tend to have higher values. This finding emphasizes the importance of room count in property valuation, guiding both buyers and sellers in their negotiations.
5. **Geographical Trends**: The geographical distribution of property values revealed that certain regions consistently outperform others in terms of property appreciation. The ROI by region chart provided a clear visual representation of these trends, allowing stakeholders to make informed decisions about where to invest.

7. Business Value Generated from the Project

The insights derived from the analysis create significant business value in several ways:

1. **Informed Decision-Making**: The findings empower real estate investors and developers to make data-driven decisions regarding property acquisitions, renovations, and pricing strategies. By understanding which properties have the highest potential for appreciation, businesses can allocate resources more effectively.
2. **Market Positioning:** Real estate agencies can leverage the data to better position themselves in the market. By highlighting properties with strong investment potential and quality features, agencies can attract more clients and close deals faster.
3. **Risk Mitigation:** Understanding the correlation between building quality and property value helps mitigate risks associated with property investments. Investors can prioritize properties that meet quality standards, reducing the likelihood of financial losses due to depreciation.
4. **Strategic Planning:** The analysis provides a foundation for long-term strategic planning. Businesses can use the insights to forecast market trends, identify emerging neighborhoods, and develop targeted marketing strategies that align with consumer demand.
5. **Enhanced Customer Experience:** By utilizing the insights from the analysis, businesses can offer personalized recommendations to customers based on their preferences and needs. This tailored approach enhances customer satisfaction and loyalty.

In conclusion, the results of the analysis not only provide a comprehensive understanding of the property market but also equip stakeholders with the tools necessary to drive business growth and enhance profitability. The integration of data analytics into real estate operations represents a significant step towards achieving competitive advantage in a dynamic market.