Tyler Trocchi

Jasmine Weekes

Athina Osei Kyerematen

Jared Jones

MSA 8040

API Project Final Report

For our API project, we decided to make an API call, build a data pipeline and perform an analysis through SQL queries. Our group researched APIs available through RapidAPI.com which gives access to over 40 million public APIs. We chose to perform our analysis on a publicly available API provided by Zillow, a popular site used to purchase, lease or rent out properties. This API call provided lots of housing data including key features such as estimates on properties purchase value, rent, and property details such as square footage, bedrooms, bathrooms and home type. Due to the nature of the API call and the limited data we could obtain without paying for it, we chose to focus our analysis on the area of Snellville, GA (zip code 30078). We used the API call four times to sort the data by zip code, and made the maximum number of calls before we would be charged, leaving us with 151 rows of listing data.

Data Exploration and Preparation

The group created a Google Cloud Platform account and then downloaded the google.cloud Python package which would give us access to the bigquery sub-module so we can create and access BigQuery tables. To authenticate calls, we installed the Google Cloud Command Line Interface, initialized it, and created local authentication credentials for the user

account. Then we created the dataset to house the BigQuery table and the BigQuery table itself using the Google API Client Library in Python.

To make requests to the endpoint for the Zillow API using the URL provided on rapidapi.com, we imported the requests Python library and made get requests to the URL using the zip code 30078 as the parameter for the request. We had to make multiple calls including the page number of the listings in the parameters because the response didn't return all the listings for the zipcode at once. Then we loaded the request bodies into a data frame and used the Client class to access the table in the database. Some of the columns were in JSON form and had to be normalized to be tabular using Pandas functions. Additionally, null values were replaced with 'False' or the mean of the column where appropriate. Columns that were mostly or entirely null or didn't contain relevant information, such as links to images, were dropped from the data. We then defined a partial schema for the table and loaded the data frame to the BigQuery table with certain job configurations.

Data Analysis

Three different queries were run to further analyze and visualize the data to gain more understanding on interactions between the various attributes of the data set and make inferences.

The first query was run to determine the homes which are overpriced grouped by home type. As shown below,the query computes the average price per square foot for each home type and compares the price per square foot of individual homes to the result according to the home type. Properties which are more than 120% more than the average price for that type of property are considered overpriced or expensive and are shown in descending order.

This query allows Zillow to identify specific homes which may be too expensive as well as their street addresses and enables them to assess if homes are worth the prices being charged for them,based on location,size,hometype and other factors.It also enables clients or home buyers know if the home they are looking to buy is relatively expensive based on concrete data.

```
WITH AvgPricePerSqft AS (
    SELECT
        homeType,
        AVG(price / livingArea) AS avg_price_per_sqft
    FROM ZillowDataset
    GROUP BY homeType
)
SELECT
    z.streetAddress,
    z.homeType,
    z.price,
    z.livingArea,
    (z.price / z.livingArea) AS price_per_sqft,
    a.avg_price_per_sqft
FROM ZillowDataset z
JOIN AvgPricePerSqft a ON z.homeType = a.homeType
WHERE (z.price / z.livingArea) > 1.2 * a.avg_price_per_sqft
ORDER BY price_per_sqft DESC;
```

Fig.1 Query to determine overpriced homes in descending order according to home type.

Our second query was aimed at answering a specific question for property owners on Zillow: should I rent this property or sell it outright? Specifically, a new metric when created based on the estimate given to us for the monthly rent, and divided against the estimate given to us for the sale of the property as a whole. Based on the value of this new metric, they were assigned a label with high, or moderate or low rentability. A highly rentable property means that this property would generate lots of value if you were to rent it rather than sell it. On the flip

side, this new “rentability” category for property owners can be reversed and used as a metric for consumers. For example, a property with “Low Rentability” for the property owner, would have a high rentability property for the consumer, since he would be paying a smaller proportion of the rent with regards to the over price of the house. The results of this query were that only 11 of these properties, 10 of which being lots, were highly rentable, while the majority of the properties were rated as “low rentability”. The overall conclusion we can draw from this query would be that the area of Snellville has more of a renter’s economy than a seller’s economy.

```sql
SELECT
    zpid,
    streetAddress,
    homeType,
    bedrooms,
    bathrooms,
    price AS estimated_price,
    rentZestimate AS estimated_rent,
    ROUND((rentZestimate / price) * 100, 2) AS rent_to_price_ratio,
    CASE
        WHEN (rentZestimate / price) * 100 > 1 THEN 'High Rentability'
        WHEN (rentZestimate / price) * 100 BETWEEN .7 AND 1 THEN 'Moderate Rentability'
        ELSE 'Low Rentability'
    END AS rentability_category
FROM
    zillow_dataset1
WHERE
    price > 0 AND rentZestimate > 0
ORDER BY
    rent_to_price_ratio DESC;
```

Fig 2. Query to determine which properties are profitable to rent out rather than to sell outright.

For our third query, we wanted to analyze the accuracy of Zillow’s home estimates compared to its tax-assessed value. This is an important factor in enhancing the trust of Zillow’s users as an inaccurate estimate could hinder buyers and sellers in their pricing decisions. We created a query that would identify properties in the city whose Zillow estimate (“Zestimate”) is

significantly higher than its tax-assessed value. By doing this, Zillow can identify which listings may have been over-assessed by the company. Upon running our query, we identified five listings with significantly high Zestimate-to-tax-assessed-value ratios: 135.7, 10.2, 7.8, 6.5, and 2.4. These figures stand out compared to the calculated total average ratio of 1.3, indicating notable discrepancies. Addressing these outliers would be a valuable opportunity for Zillow to enhance its reliability by reassessing these homes and improving the accuracy of its estimates.

```sql
SELECT zpid, city, zestimate, taxAssessedValue,
       (zestimate / taxAssessedValue) AS price_to_tax_ratio
FROM real_estate
WHERE taxAssessedValue > 0 AND (zestimate / taxAssessedValue) > (
    SELECT AVG(zestimate / taxAssessedValue)
    FROM real_estate
    WHERE taxAssessedValue > 0
)
ORDER BY price_to_tax_ratio DESC;
```

Fig. 3 Query to determine which property listings have a significantly high Zestimate-to-Tax-Assessed Value ratio

Results

Fig. 4 Expensive homes in the city of Snellville.

From fig. 4 above, there are 9 homes in Snellville which are overpriced. The most expensive home costs $5,799 per square footage which is $5,560.03 per square footage above average within its home category or type.This indicates that 95.9% of the cost of this home is more than the average cost for this home type. Home buyers with this data will be able to avoid these 9 homes if they want to cut cost and get value for their money with no preference for location or other factors that may be affecting cost.

Fig. 5 Home Types with expensive homes in the city of Snellville .

From fig. 5 above, all the 9 expensive homes from fig. 4 are either in the single family or lot category. On average, prices per square foot of overpriced properties which are lots are higher as compared to the average price per square foot of lots than those which are single family homes. With this information, home buyers who would want to cut cost could look into the option of collaborating with other families to purchase a multifamily home or purchase a townhome instead of a single family home.

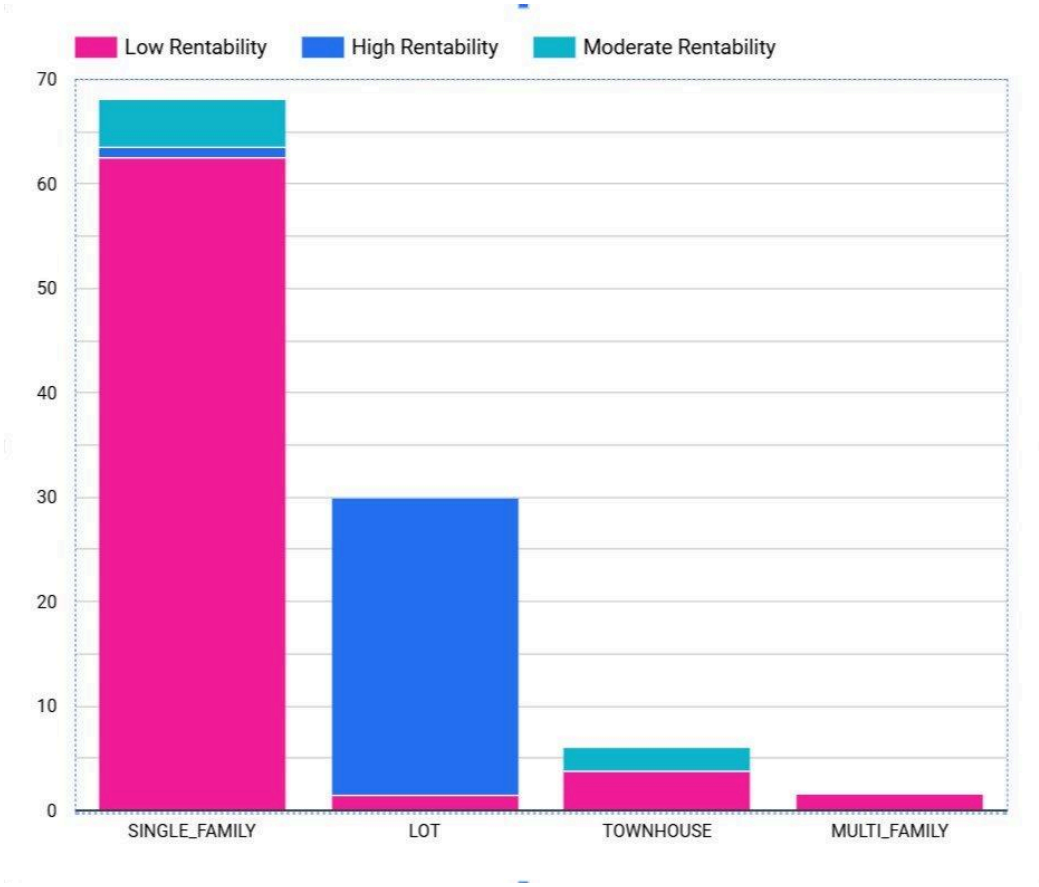

Fig. 6 Results from the second query, grouped by home type.

As we can see in Fig. 6 above, the majority of our data is on single family homes. Single family homes are almost half of the entire dataframe, and the large majority of these homes are categorized as “Low Rentability”. Another result we can see is that with the exception of the home type “Lot”, the large majority of these homes have “Low Rentability” as well. As for the home type “Lot”, these homes as a majority have “High Rentability”. This tells us unless your property is a lot, it is much more profitable to sell than to rent.

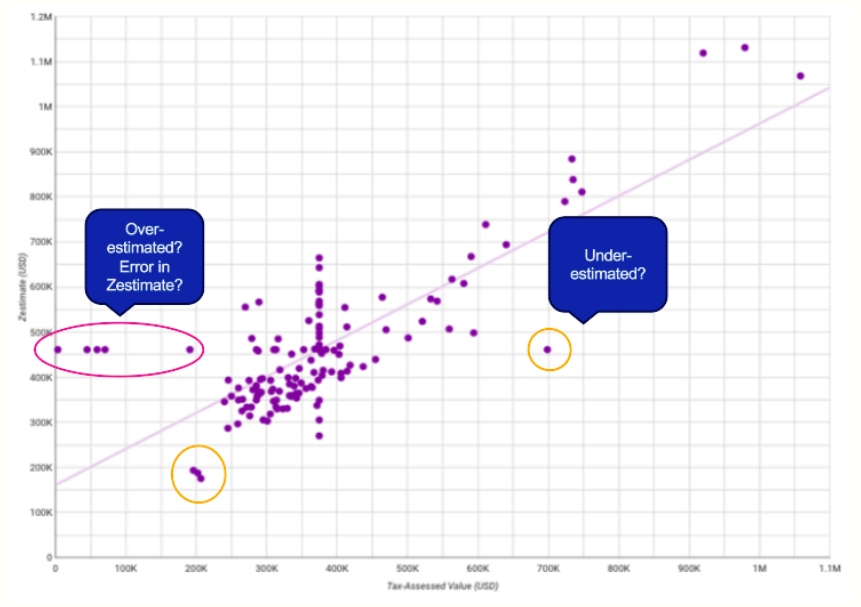

Fig. 7 Tax-Assessed Value vs. Zestimate

Fig. 7 above plots each listing's tax-assessed value against its corresponding Zillow estimate, with a trend line representing the average ratio across all listings. The five listings returned by our query can be found on the left side of the graph, far above the trend line, indicating that these properties have been significantly overestimated by Zillow. These inaccuracies could stem from a system error, miscalculation, issues with the assessed value, or other factors. Flagging these listings is crucial for Zillow to investigate the root causes of these discrepancies. Additionally, the graph reveals homes with underestimated Zestimates, with ratios as low as 0.66. These properties may also warrant further examination to understand why their estimates fall significantly below their tax-assessed values.

Conclusion

Throughout this project, we were able to integrate public APIs, Google Cloud Platform, and analysis using Python and SQL to derive recommendations based on real-world data. By

using the Zillow API, we discovered meaningful insights into property pricing in Snellville, GA . The queries we executed identified property listings that may be overpriced, analyzed home rentability, and assessed the accuracy of Zillow's property estimates. This analysis provides actionable insights for both Zillow and its consumers. By further analyzing this data, there is great opportunity to improve Zillow's platform and build greater trust among its users.