# The Cost of Popularity: Analyzing Tourism's Effect on Medellín's Housing Market

Mateo Ruiz Alvarez - 23272477

## Introduction

Tourism is a vital driver of economic activity, with significant implications for local housing markets. Medellín, Colombia, has seen a remarkable rise in tourism over the past decade, understanding the relationship between tourism influx and housing market dynamics is essential.This study investigates how tourism, characterized by both domestic and inbound travel, impacts housing sales and rental prices. By analyzing data from 2011 to 2021, this research aims to uncover trends and correlations that provide insight into the economic and social effects of tourism on Medellín's housing sector.

## 1. Question

To what extent do domestic and inbound tourism trends correlate with fluctuations in housing market prices in Medellín, Colombia?

## 2. Data Sources

### 2.1 General Data Information

- **DataSource1:** Monthly entry of foreigners and Colombians at José María Córdova airport from 2008 to 2022: *Metadata URL*, *Data URL*, *Data Type*: CSV
- **DataSource2:** Monthly passenger arrivals by airport of national origin from 2007 to 2022: *Metadata URL*, *Data URL*, *Data Type*: CSV
- **DataSource3:** Monthly passenger arrivals at the airport of international origin from 2007 to 2022: *Metadata URL*, *Data URL*, *Data Type*: CSV
- **DataSource4:** Property offers for sale and rent in Medellín for the year 2021: *Metadata URL*, *Data URL sales*, *Data URL rents*, *Data Type*: KML (for both)

### 2.2 Description

The data for this study was sourced from two reliable platforms: **MEData** - Medellín's official open data portal, providing public access to the city's strategic information, and **the Medellín Real Estate Observatory (OIME)** - an initiative of the Secretariat of Territorial Management and Control, aimed at understanding the city's economic landscape and generating knowledge about urban dynamics. These sources were selected for their relevance and comprehensive coverage of tourism and housing market trends. The datasets from *MEData* (DataSource1, DataSource2, DataSource3) include monthly statistics on traveler arrivals to Medellín's airport, categorized into Colombians and foreigners, along with details such as city of origin for domestic travelers and country of origin for international visitors. The *OIME* data (DataSource4) provides granular information on housing sales and rental offers, including location, price per square meter, property area, type of property, and other key attributes. Together, these datasets offer a robust foundation for analyzing the interplay between tourism influx and housing market dynamics.

### 2.3 Structure and Quality

The data from *MEData* is **structured** in CSV format (tabular structure with a fixed schema), offers significant advantages, including ease of processing, straightforward integration with analysis tools, and consistent formatting across datasets. On the other hand, the *OIME* data is provided in KML format (nested nature), while is highly effective for representing geospatial data, it introduces additional complexity during extraction and transformation.

Regarding **data quality**, the *MEData* datasets excel across all dimensions. They are accurate, reflecting real-world traveler data; complete, with no significant missing information; consistent, using uniform formatting; timely, as the data covers relevant years; and relevant to the study's objectives. In contrast, the *OIME* data faced several issues. Accuracy is generally high but is affected by missing information for certain years. Completeness is compromised by the absence of data points for some attributes. Consistency issues include varying column names to refer the same attribute, different number of attributes across years and differing decimal formats (e.g., commas vs. periods). While the data is timely, spanning the required timeframe, addressing these inconsistencies required extensive preprocessing to ensure the data aligned with the study's goals.

### 2.4 Licenses

All datasets used are licensed under the *CC BY-SA 4.0 (Attribution-ShareAlike 4.0 International)* license. This open-data license permits the use, sharing, and adaptation of the datasets, provided appropriate attribution is given to the original creators and any derivative works are distributed under the same license. To fulfill these obligations, this report will include proper attribution to the data sources in all relevant sections, and any outputs or transformations of the data will be shared under the same *CC BY-SA 4.0* license where applicable. For license verification, in section **2.1 General Data Information**, for each data source the corresponding *Metadata URL* is presented. Once

navigated to the URL, you can access the details about its license by clicking on "Open details" and under the "License" column.

## 3. Data Pipeline

### 3.1 Description

The data pipeline was implemented in Python, leveraging libraries such as `pandas` for data manipulation, `xml.etree.ElementTree` for parsing KML files, and `sqlite3` for database storage. The pipeline automates the Extract, Transform, and Load (ETL) process for integrating data from various sources. For *MEData* datasets, the pipeline extracts CSV files directly, whereas for *OIME* datasets, it parses nested KML structures to retrieve property attributes and geospatial coordinates. This pipeline enables seamless integration of diverse datasets into unified SQLite databases for analysis. As shown in Figure 1, the ETL Pipeline automates the process of extracting, transforming, and loading data from diverse sources into unified formats for further use.

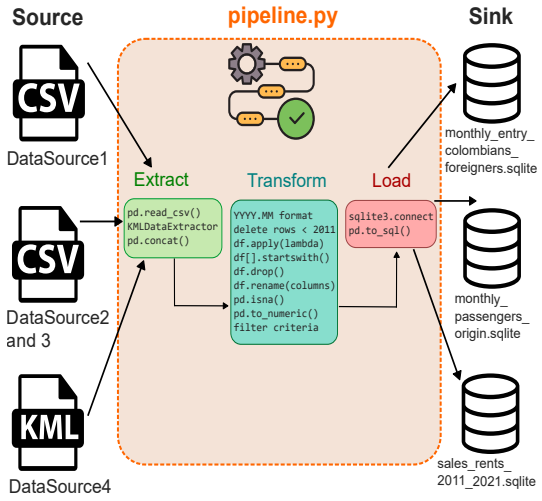


Figure 1: ETL Pipeline diagram

### 3.2 Transformation Tasks

The data transformation steps are outlined in Table 1.

### 3.3 Problems Encountered

The pipeline encountered several challenges:

- **Inconsistent KML Schemas:** The *OIME* datasets varied in schema and formatting across years. This was resolved by implementing year-specific mappings to standardize column names and data structures.
- **Missing Data:** Some years had incomplete records or entirely missing datasets, particularly for certain variables like Neighborhood and Strata. Missing values were addressed by imputing derived values where feasible or limiting analysis to periods with complete data.
- **Encoding Issues:** Special characters in KML files caused occasional parsing errors. These were resolved through preprocessing steps, including encoding normalization and replacing problematic characters.
- **Data Quality Variability:** Numerical inconsistencies, such as mixed decimal formats (e.g., commas vs. periods) in `Valor Comercial`, were corrected using custom cleaning functions to standardize formats across all datasets.

### 3.4 Metaquality Measures

The pipeline incorporated several metaquality measures to ensure reliability and adaptability:

- **Error Handling:** The pipeline handled missing values by calculating them where possible (e.g., deriving `Valor M2`) and flagged records with unresolved issues for review.
- **Schema Validation:** Intermediate outputs were validated against predefined schemas to catch errors early.`KMLDataExtractor` allowed the pipeline to accommodate schema changes.
- **Consistency Monitoring:** Year-specific mappings and column harmonization ensured uniform data structure, particularly for KML data.

## 4. Results and Limitations

### 4.1 Output Description

The pipeline produced three primary databases:

- `monthly_entry_colombians_foreigners.sqlite` [from *DataSource1*]: A dataset summarizing the monthly entry of Colombians and foreigners through Medellín's airport, categorized by nationality.
- `monthly_passengers_origin.sqlite` [from *DataSource2 and 3*]: A dataset combining domestic traveler city origins and international traveler country origins for passengers arriving in Medellín.
- `sales_rents_2011_2021.sqlite` [from *DataSource4*]: A unified dataset containing sales and rental offers extracted from *OIME* data (2011–2021), including property type, price (total and per square meter), neighborhood, area and geospatial details.

The resulting SQLite databases represent a significant improvement in data quality compared to the originally extracted datasets. Key enhancements include greater consistency through harmonized column names and formats, improved completeness by addressing missing values where possible, and increased relevancy by filtering out irrelevant or noisy data. However, some limitations remain due to deficiencies in the original data. Certain important variables, such as Neighborhood and Strata for earlier years, could not be reconstructed, and the analysis period had to be restricted to ensure sufficient data quality and completeness. Despite these challenges, the processed data provides a robust foundation for analyzing tourism and housing market dynamics in Medellín.

| | | |
|---|---|---|
| **Extract** | [DataSource1]: 1 CSV file | Extracted using `pandas.read_csv()` to capture monthly entry data for Colombians and foreigners from *MEData*. |
| | [DataSource2] + [DataSource3]: 2 CSV files | Extracted using `pandas.read_csv()` to obtain data on domestic and international travelers, including their city or country of origin. |
| | [DataSource4]: 22 KML files (11 + 11 for sale and rent offers) | Processed with the `KMLDataExtractor` class, each file was parsed with `xml.etree.ElementTree` to extract relevant property details based on year-specific mappings. The yearly datasets were then combined into a unified `pandas.DataFrame`, consolidating all sales and rental data. |
| **Transform** | [DataSource1] | Column names were renamed for clarity, dates standardized to `YYYY.MM`, and records prior to 2011 were removed. |
| | [DataSource2] + [DataSource3] | Columns were renamed, unnecessary fields dropped, and the `Period` field reformatted to `YYYY.MM`. Rows with dates earlier than 2011 were excluded. |
| | [DataSource4] | Column inconsistencies resolved (e.g., unifying `Valor M²` and `Valor M2`), missing values calculated for `Valor M2`, and non-residential properties filtered out. |
| **Load** | [DataSource1] | Database connections were managed with `sqlite3.connect`.Stored in an SQLite database using `pandas.to_sql()` for structured querying. |
| | [DataSource2] + [DataSource3] | Combined with `pandas.concat()` and saved to an SQLite database for easy integration and analysis. |
| | [DataSource4] | Loaded into an SQLite database, creating a unified dataset for sales and rental data across 2011–2021. |

Table 1: Tasks performed in the ETL Pipeline

### 4.2 Data Format

The datasets were stored in SQLite databases, offering structured, relational storage that supports efficient querying and integration with analytical tools. The consistent schema design ensures scalability and adaptability, while SQLite's simplicity and portability make it ideal for local processing and broader analytical workflows.

### 4.3 Potential Issues

While the pipeline addresses many inconsistencies in the raw data, several potential issues remain:

- **Data Completeness:** Certain years in the *OIME* data lacked records or were incomplete, which may limit the temporal scope of analyses. Notably, data for the period 2011–2015 lacks information on Neighborhood and Strata (a classification based on physical characteristics of residences), restricting comprehensive analysis to 2016–2021 when these variables are complete.
- **Consistency Challenges:** The original authors of the *OIME* data could have done a better job ensuring consistency. Significant variations in schema and formatting across years made it very time-consuming to configure the KML parsing for each year. Consequently, the implemented code is highly specific to this dataset and cannot be reused elsewhere.
- **Interpretation Limitations:** Variability in the granularity of both tourism and housing data might complicate direct comparisons or trend analyses. Additionally, despite efforts to address missing values, some data gaps remain that could impact interpretations.
- **Future Integration:** A potential improvement involves integrating the three SQLite databases into a single database with multiple interlinked tables for enhanced accessibility and analysis.
- **Data Quality Issues in MEData:** Although the *MEData* datasets were in better condition, some inconsistencies persist, such as rows tagged as "inconsistencies" without additional details and occasional negative passenger values. These anomalies would require additional preprocessing to mitigate their impact.

These limitations will be addressed, where possible, through careful interpretation and transparent reporting in the final analysis.