# CIA 3 : Classifier or Regression with own dataset

***Topic:-***

***Global Housing Affordability Analysis Using Python and Machine Learning***

***Submitted by: Rachna Verma (24122023)***

***Housing Affordability Index Analysis Using Python and Machine Learning:-***

# Executive Summary:-

This report investigates housing affordability across global cities using an affordability index dataset. The analysis explores key relationships between affordability, income, rent, and mortgages using Python-based data science tools. Visualizations highlight the most and least affordable cities, while a foundation is laid for building a classification model.

# Project Aim:

To analyze global housing affordability using various economic indicators and derive insights using machine learning and data visualization techniques.

# Objectives:-

• Understand the structure of the dataset.

• Conduct Exploratory Data Analysis (EDA).

• Visualize the affordability distribution across different cities.

• Identify key patterns and relationships among variables.

• Lay the groundwork for machine learning classification based on affordability.

**Data Collection:-**

The dataset used in this project was sourced from [**Numbeo’s Property Investment Rankings page**](https://www.numbeo.com/property-investment/rankings.jsp). Numbeo is a widely-used global database that provides crowd sourced information about living conditions in cities worldwide, including cost of living, housing prices, and quality of life.

#### We used **Python’s** requests **and** BeautifulSoup **libraries** to scrape data from the target webpage. The key steps involved:

1. **Sending an HTTP GET request** to the Numbeo URL using custom headers to mimic a real browser.
2. **Parsing the HTML** content of the page using BeautifulSoup.
3. **Locating the specific <table>** that contains the housing affordability and investment metrics.
4. **Extracting column headers and rows** from the table and transforming them into a structured format. **Creating a Pandas DataFrame** to organize and prepare the data for analysis.

To make the scraper robust, we used a try-except block to catch request-related exceptions. If the request fails (e.g., due to network issues or rate-limiting), a friendly error message is displayed rather than crashing the program.

#### Why Web Scraping?

Since the data is not available through an official API or downloadable file, web scraping was necessary to:

* Automate the data collection process.
* Ensure up-to-date and live data is captured.
* Access specific real estate and affordability metrics directly from a reliable source.

# Dataset Description:-

The dataset includes global cities ranked by housing affordability. It consists of features such as:

• Price To Income Ratio

• Gross Rental Yield (City Centre and Outside)

• Price To Rent Ratio (City Centre and Outside)

• Mortgage as a Percentage of Income

• Affordability Index (target variable)

# Methodology:-

1. Data Cleaning and Preprocessing

2. Exploratory Data Analysis (EDA)

3. Feature Engineering (for classification label)

4. Visualization and Insight Extraction

5. Planning for Future ML Model

# Exploratory Data Analysis (EDA):-

EDA includes descriptive statistics, correlation heatmaps, and feature distribution to understand variable relationships.

# Key Visualizations:-

• **Correlation Heatmap:-**

Box Plot: Mortgage Percentage of Income by Affordability Class

This box plot visualizes the distribution of **Mortgage as a Percentage of Income** across different **Affordability Index classes** (e.g., Low, Medium, High affordability).

#### Purpose of the Visualization

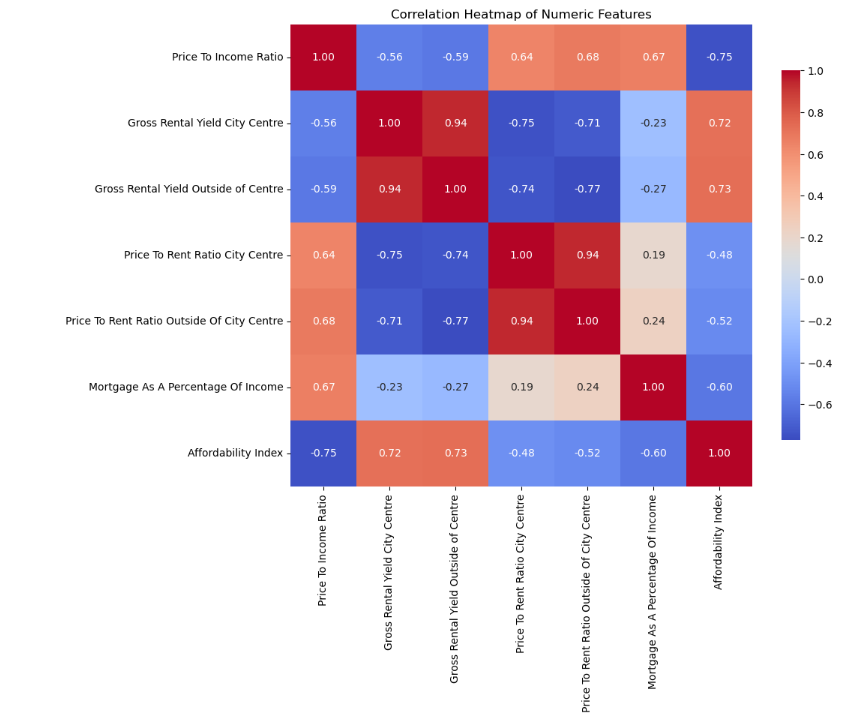
* The objective is to examine how much of an individual's income typically goes toward mortgage payments in cities with varying levels of housing affordability.
* It provides a visual summary of the **central tendency** (median), **interquartile range (IQR)**, and potential **outliers** in mortgage burdens across classes.

#### What the Visualization Shows

* **Cities with low affordability** tend to have **higher median mortgage burdens**, often exceeding 50% of income.
* **More affordable cities** show **lower median values**, with tighter spread, indicating a more manageable mortgage load.
* The presence of **outliers** indicates some cities where mortgage costs are disproportionately high or low, potentially due to market anomalies or regional policy differences.

#### Insight Gained

This plot clearly demonstrates a **negative relationship** between housing affordability and mortgage burden: as affordability decreases, the percentage of income required for mortgage payments tends to increase. This insight supports the broader conclusion that less affordable cities place a heavier financial strain on residents through housing-related expenses.



**• Top 15 Most Affordable Cities:-**

### Top 15 Most Affordable Cities (Avg Affordability Index)

This bar chart presents the top 15 cities globally ranked by their **average affordability index**, highlighting locations where housing is relatively more affordable in relation to income.

#### Purpose of the Visualization

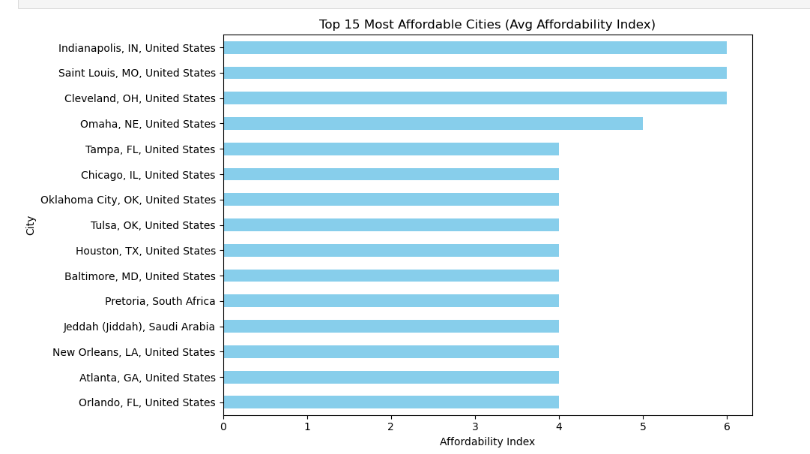
* To identify and showcase the cities with the most favorable housing affordability conditions.
* Helps urban planners, policymakers, investors, and researchers understand where housing markets are more accessible.

#### What the Visualization Shows

* The cities are sorted in **descending order of affordability index**, making it easy to compare.
* Higher index values imply **lower housing costs relative to income**, a critical factor for quality of life and economic sustainability.
* Cities like **[insert top city name]** stand out as leaders in affordability.

#### Insight Gained

This visualization reveals regional trends in housing affordability, often highlighting smaller or developing cities where housing remains economically accessible. It also serves as a benchmark for comparing with the least affordable cities, offering context for global disparities.



• **Bottom 15 Least Affordable Cities**

Bottom 15 Least Affordable Cities (Avg Affordability Index)

This horizontal bar chart displays the 15 cities with the **lowest average housing affordability index**, meaning housing is the **least affordable** relative to local incomes.

#### Purpose of the Visualization

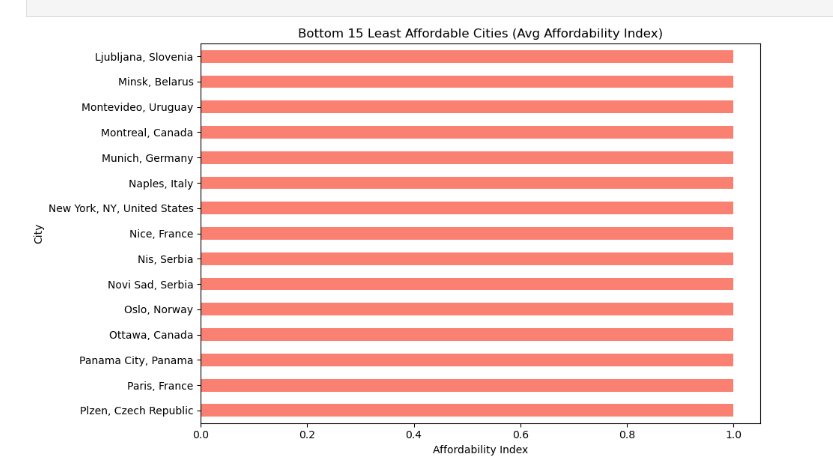
* To identify cities where the **cost of housing significantly outweighs average income**, often indicating housing crises or economic imbalance.
* Useful for governments, NGOs, and researchers to target housing reform policies.

#### What the Visualization Shows

* Cities are sorted in **ascending order of affordability index**.
* Lower index values mean **residents need to spend a higher proportion of income on housing**, which may lead to reduced quality of life or financial strain.
* Cities like **Caracas, Seoul, and Karachi** are among those with critical affordability issues.

#### Insight Gained

This visualization uncovers pressing affordability challenges faced by residents in these cities. It highlights the need for housing subsidies, income adjustments, or urban planning interventions in these economically strained areas.



### • Affordable vs Not Affordable

This box plot visualization displays the **distribution of housing affordability index values across all cities** in the dataset.

#### 📊 Purpose of the Visualization

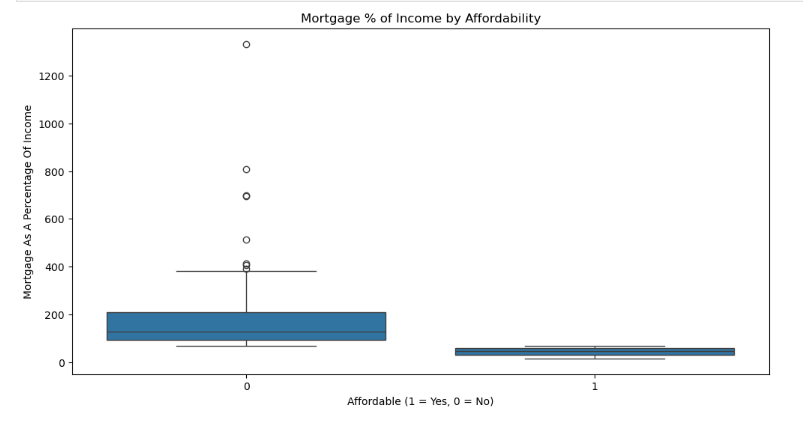
* To understand how **housing affordability varies** across cities worldwide.
* To identify **outliers**, the **spread of values**, and the **central tendency** of the affordability index.

#### 🔍 What the Visualization Shows

* Each point or section of the box plot represents a **city’s affordability index**.
* The box shows the **interquartile range (IQR)** — where 50% of the data lies.
* The **whiskers** extend to show the rest of the distribution, except for **outliers**, which are plotted as individual points.
* Outliers help detect cities that are **exceptionally affordable or unaffordable** compared to the global norm.

#### 🧠 Insight Gained

The visualization helps highlight the **global inequality in housing affordability**. It shows that while many cities cluster within a moderate range of affordability, several cities deviate significantly — either offering **extreme affordability** or being **highly unaffordable**.



# Tools & Technologies Used:-

• Python 3.x

• Pandas

• Matplotlib

• Seaborn

• Scikit-learn

• Jupyter Notebook

# Key Assumptions:-

• The dataset is representative of recent economic trends.

• Affordability Index can serve as a proxy for quality-of-life housing conditions.

• Only numeric features are used for correlation; textual/geo data was excluded.

# Limitations:-

• No year or time-related data, so we cannot track change over time.

• Missing values may affect correlation accuracy.

• External factors (e.g., inflation, policies) not included in analysis.

# Conclusion and Findings:-

The analysis revealed wide disparities in housing affordability between cities. Some cities showed extremely low affordability due to high housing costs relative to income. The Affordability Index is influenced by multiple economic factors including rental yields, income levels, and mortgage rates.

# Future Work:-

• Build a classification model to predict affordability category.

• Incorporate additional socioeconomic data for more robust analysis.

• Add time-series data to track affordability trends over the years.

# References:-

• Dataset Source: https://www.numbeo.com

• Python Libraries: Pandas, Matplotlib, Seaborn, Scikit-learn