Exercise 6.1: sourcing open Data

Project Definition: “New York Rental Properties Pricing Market Analysis Dataset”.

This dataset provides a comprehensive snapshot of the rental landscape in New York City, with details of rental properties and pricing spanning various neighbourhoods. The project will aim to dissect and understand the rental market dynamics in New York City by analysing a mock dataset tailored for educational purposes.

Objective:

The objective is to analyse the New York rental properties pricing market to identify trends and insights that can inform both renters and property owners.

Goal:

The goal of this project is to understand the factors affecting rental prices and occupancy rates. Additionally, to provide recommendations for pricing strategies based on location and property type.

Scope:

The analysis will only cover rental properties in New York City as listed in the dataset. It will focus on data from the year 2019 for occupancy trends and availability in 2020. The project will examine variables such as neighbourhood, room type, price, occupancy, reviews, and availability. The project will not be focusing on predictions but will be limited to historical data analysis. Lastly, the project will conclude with a set of actionable insights for both renters looking for properties and owners looking to list their properties.

Why chose this dataset:

I selected the New York Rental Market dataset for its comprehensive yet accessible nature, making it ideal for junior analysts. The dataset offers a range of data points, such as location, price and review, which are quite fundamental in understanding real estate dynamics. Moreover, New York City has a vibrant and diverse market which provides an intriguing case study for urban rental trends.

Data Source:

The data is externally sourced which is available on Kaggle. It is sourced from a community of data enthusiasts and scientists.

Data collection:

This data is an administrative data set and a simulated collection of rental property information in New York City, designed specifically for educational purposes and analytical practices.

Data contents:

The variables in this data set contains Id, neighbourhood, latitude, longitude, room_type, price, days_occupied_in_2019, minimum_nights, number_of_reviews, reviews_per_month, availability_2020.

Data Limitations and ethics

Since the data is purely for educational purposes, there are some potential limitations and biases to consider, e.g., The data might not accurately reflect the complexities and nuances of the real New York rental market. Some important factors that affect rental prices and demand, like current economic conditions, local events, or market trends, may not be represented. The data set may not cover all neighbourhoods or types of rental properties equally, leading to a skewed understanding of the market.

When considering the ethics of using this dataset, since it's a simulated dataset and open source, there are no real individual data involved, which addresses major concerns about privacy and data protection. The data is created to avoid biases or stereotypes, especially regarding neighbourhoods or property types.

Data profile: **NY Rental Properties Data Characteristics**

Variable	Description	Time-Variant/Invariant	Structured/Unstructured	Categorical (Binary/Nominal/Ordinal)	Discrete/Continuous	Qualitative/Quantitative
id	Unique identifier for each rental	Invariant	Structured	Nominal	Discrete	Quantitative
neighborhood	Area in NYC where the property is located	Invariant	Structured	Nominal	Discrete	Qualitative
latitude	North-south position of the property	Invariant	Structured	-	Continuous	Quantitative
longitude	East-west position of the property	Invariant	Structured	-	Continuous	Quantitative
room_type	Type of rental space	Invariant	Structured	Nominal	Discrete	Qualitative
price	Rental price per night	Variant	Structured	-	Continuous	Quantitative
days_occupied_in_2019	Number of days rented in 2019	Variant	Structured	-	Discrete	Quantitative
minimum_nights	Minimum number of nights required	Variant	Structured	-	Discrete	Quantitative

number_of_reviews	Total number of reviews received	Variant	Structured	-	Discrete	Quantitative
reviews_per_month	Average number of reviews per month	Variant	Structured	-	Continuous	Quantitative
availability_2020	Number of days available in 2020	Variant	Structured	-	Discrete	Quantitative

Questions to explore.

With the variables provided in the dataset, I want to explore some analytical questions namely:

Price Analysis:

- What is the average rental price per neighbourhood?
- How do rental prices correlate with the geographical location (latitude and longitude)?

Occupancy Analysis:

- Which neighbourhoods have the highest and lowest days occupied in 2019?
- Is there a seasonal pattern to the occupancy rates?

Types of Room Analysis:

- What type of room is most common in each neighbourhood?
- How does a room type affect rental price and occupancy rates?

Customer Review Analysis:

- How do the number of reviews and reviews per month relate to the price location?
- Are more reviewed properties more likely to be occupied?

Availability Analysis:

- How does availability in 2020 compare across different neighbourhoods?
- Is there a relationship between availability and minimum nights required?

Minimum Nights Analysis:

- What are the average minimum nights stay per neighbourhood, and how does this affect occupancy?

Geospatial Analysis:

Can we identify clusters of rental types, pricing, or occupancy rates geographically within New York City?