# Data Wrangling Project Data Proposal

## Group Name and Members

- Group Name: 6
- Members:
  - Member 1: Shuwen Hou
  - Member 2: Bingxiang Peng
  - Member 3: Yihe Ding
  - Member 4: Shiqi Lyu

## Introduction

The proliferation of short-term rental (STR) platforms, such as Airbnb, has profoundly reshaped urban housing markets, introducing both economic opportunities and significant challenges [1]. The impact of STRs on housing affordability and neighborhood composition is a subject of intense debate among urban planners and policymakers [2]. However, rigorous analysis is often hampered by the poor quality and heterogeneity of publicly available data, which frequently suffers from inconsistencies, missing values, and disparate formats [3]. Effectively addressing these data quality issues through a systematic data wrangling process is a critical and necessary prerequisite for any robust empirical analysis [4]. Without a clean and integrated dataset, conclusions drawn about market dynamics remain tentative and potentially misleading. We will analyze data from New York City as a case study to develop a reproducible pipeline for cleaning and integrating multi-source datasets related to the STR market. The complex data landscape of NYC provides an study environment to address these data wrangling challenges, which we specifically aim to resolve in this study.

## Data Description

We will compile three distinct datasets from different sources to create a comprehensive view of the short-term rental market and its demographic context in New York City. The first dataset is the *New York City Airbnb Open Data*, which will be the primary subject of our cleaning procedures. It contains detailed listing information, including unique identifiers for hosts and listings, geographic information such as borough and neighbourhood, latitude and longitude, as well as property characteristics like room type, price, and availability. This dataset is known to contain numerous inconsistencies that make it suitable for a data wrangling project. The second dataset is the *U.S. Airbnb Open Data*. This national-level dataset will serve as a reference for standardizing variables, such as `room_type` categories, to ensure our final dataset is comparable to national benchmarks. For our third dataset, we will source *NYC Housing and Demographic Data* from official sources like the NYC Department of City Planning and the U.S. Census Bureau's American Community Survey. Key variables will include median home value, median rent, population density, and median household income, aggregated at the neighbourhood level. Because each dataset can be linked by a common geographic unit (neighbourhood), we will be able to merge the cleaned Airbnb market data with local socio-economic indicators.

## Aims and Tasks

**Aim 1: Identify determinants of Airbnb pricing in New York City using statistical modeling.** The goal of this aim is to quantify how listing-, host-, and neighborhood-level features contribute to nightly price variation across the NYC short-term rental market. This will establish the foundational statistical relationships that explain pricing dynamics in a highly heterogeneous urban environment.

**1.1. Model key listing characteristics.** We will construct a multiple linear regression model using log-transformed price as the dependent variable. Predictor variables will include property features (room type, number of

bedrooms, number of bathrooms, amenities count), host attributes (Superhost status, response rate, number of listings), and listing history (review count, availability). Borough-level fixed effects will be added to control for unobserved neighborhood differences.

**1.2. Evaluate model performance and feature importance.** We will compare classical and machine learning approaches, including LASSO and Random Forest, to assess predictive accuracy and identify nonlinear effects. Feature importance metrics will be derived to interpret variable influence on pricing. Diagnostic tests such as variance inflation factors (VIF), residual analysis, and tests for heteroskedasticity will be conducted to ensure robustness.

**1.3. Interpret and visualize results.** Model coefficients and feature importances will be summarized and visualized to communicate the magnitude and direction of key effects. We will interpret results within the context of urban market heterogeneity, emphasizing which characteristics most consistently predict high-value listings.

**Aim 2: Characterize spatial heterogeneity and clustering in Airbnb listings across New York City.** This aim focuses on exploring the spatial patterns of Airbnb availability and pricing to identify geographic clusters, tourism hotspots, and location-based disparities across boroughs and neighborhoods.

**2.1. Visualize spatial distributions of listings and pricing.** Using geographic information system (GIS) methods, we will generate geospatial plots—including heatmaps, kernel density surfaces, and choropleth maps—to visualize listing concentration and average nightly price by neighborhood. These visualizations will highlight borough-level differences and the presence of spatial "hot" and "cold" zones.

**2.2. Conduct spatial autocorrelation and clustering analysis.** We will apply spatial statistics, including Global Moran's I and Local Moran's I (LISA), to detect statistically significant clustering of high- and low-price areas. Identified clusters will be further analyzed in relation to landmarks, transportation hubs, and commercial corridors to assess accessibility impacts.

**2.3. Estimate spatial regression models.** To formally test for location-based dependencies, we will estimate Spatial Lag and Spatial Error models that incorporate the influence of neighboring listing prices. These models will reveal whether price patterns are spatially contagious (i.e., influenced by adjacent areas) and quantify the strength of such effects.

**Aim 3: Examine the relationship between Airbnb activity and neighborhood socioeconomic indicators.** The objective of this aim is to integrate Airbnb data with U.S. Census American Community Survey (ACS) datasets to assess how short-term rental activity interacts with broader socioeconomic patterns across New York City's neighborhoods.

**3.1. Integrate Airbnb and socioeconomic datasets.** We will perform a spatial join linking Airbnb listings to ACS data at the ZIP code or census tract level. Socioeconomic variables of interest include median household income, average rent, educational attainment, racial/ethnic composition, and population density. Derived metrics such as Airbnb density (listings per 1,000 residents) and share of entire-home listings will be computed for each area.

**3.2. Analyze correlations and visualize socioeconomic gradients.** Pearson and Spearman correlation analyses will be conducted to examine associations between Airbnb metrics (mean price, density, availability) and socioeconomic indicators. We will create bivariate choropleth maps to visualize spatial alignment between Airbnb intensity and income inequality.

**3.3. Model socioeconomic associations and heterogeneity.** Using multivariable linear regression and geographically weighted regression (GWR), we will estimate how socioeconomic context influences Airbnb activity and whether these relationships vary spatially. The models will identify neighborhoods where Airbnb intensity aligns with higher income and gentrification trends.

**Expected Outcomes:** Completion of these aims will yield a comprehensive understanding of how property characteristics, spatial dynamics, and socioeconomic context jointly shape the Airbnb market in New York City. The project will produce (1) statistical evidence on price determinants, (2) visual and quantitative identification of spatial clusters, and (3) integrated insights linking Airbnb activity to neighborhood-level inequality and urban policy concerns.

## Concluding Remarks

Our primary goal is to execute a transparent and reproducible data wrangling pipeline that transforms raw, disparate sources into a high-quality, analysis-ready dataset. We plan to use a combination of Python (utilizing the pandas and GeoPandas libraries) and R to perform the data cleaning, integration, and subsequent exploratory analysis. The successful completion of this project will provide a valuable data asset and a methodological blueprint for future research into the New York City housing market.

## References


[1] Kyle Barron, Edward Kung, and Davide Proserpio. The effect of home-sharing on house prices and rents: Evidence from airbnb. *Marketing Science*, 40(1):23–47, 2021.

[2] Nicola Camatti, Giacomo di Tollo, Gianni Filograsso, and Sara Ghilardi. Predicting airbnb pricing: A comparative analysis of artificial intelligence and traditional approaches. *Computational Management Science*, 21(1), 2024.

[3] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014.

[4] Foster Provost and Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, 2013.