Correlation Analysis Between Airbnb Prices & Subway Distance

Gangyi Li
12/04/2025

Issues and Objectives

- ***Core issue:***.Are Airbnb prices really higher when you are closer to the subway?

- ***Code:*** Multi-library collaboration (pandas/geopandas/sklearn/folium)

- ***Process:*** data → modeling → visualization

- ***Outcome:*** Python implementation for spatial data processing and regression analysis.

Data Source

Airbnb Listing Data:

Content: 20,000+ NYC records in CSV/GZ format

Python Implementation: **pd.read_csv()** with native compression support

Subway Station Data:

Content: Latitude, Longitude, Station Name

Python Implementation: Custom **read_subway()** function for automatic column name detection

Neighborhood Income Data:

Content: Sourced from Census API (Population, Median Income)

Python Implementation: Data fetched via **requests** and parsed into a **pd.DataFrame()**

Analysis and Models

Data Cleaning

Remove Missing Values:

```
df_airbnb = df_airbnb.dropna(subset=['latitude', 'longitude', 'price'])
```

Standardize Parameters:

```
df_airbnb['price'] = df_airbnb['price'].apply(parse_price)
```

Spatial Format Conversion:

```
gdf_airbnb = gpd.GeoDataFrame(
    df_airbnb,
    geometry=gpd.points_from_xy(df_airbnb.longitude, df_airbnb.latitude),
    crs="EPSG:4326"
)
```

Analysis and Models

Spatial Analysis

Use KMeans to cluster properties with similar geographic locations into 5 categories:

```
coords = pd.DataFrame({'x': gdf_airbnb_m.geometry.x, 'y': gdf_airbnb_m.geometry.y})
gdf_airbnb_with_tract['cluster'] = KMeans(n_clusters=5, random_state=42).fit_predict(coords)
```

Analysis and Models

Regression Modeling

```
X = df_model[['dist_km', 'median_household_income']]
reg = LinearRegression().fit(X, np.log1p(df_model['price']))
```

Why np.log1p(y)? Reduce impact of $1000+ luxury listings outliers

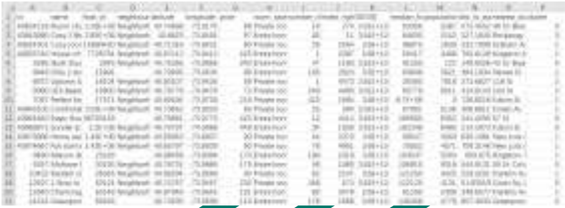
Outputs: Coefficients (impact strength) + R^2 (model fit)

Output files

airbnb_processed.csv:

Content: Cleaned data (including distance, clustering, and community attributes)

Python Implementation: Save using **df.to_csv()** with new added features.



nyc_airbnb_map.html:

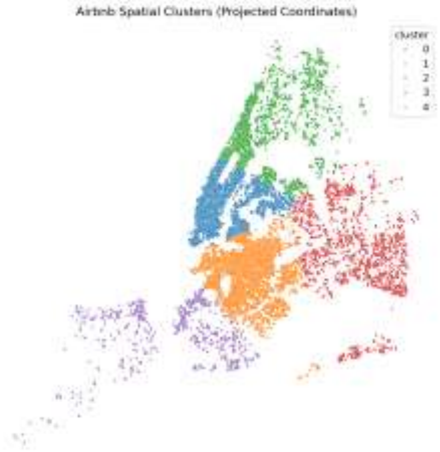
Content: Interactive Maps (Property Listings, Subway Lines, Income Heat Maps)

Python Implementation: Built with **folium.Map()**, supporting layer control.

airbnb_clusters.png:

Content: Spatial cluster results (5 groups)

Python Implementation: Clustered via **sklearn.KMeans**, visualized with seaborn.

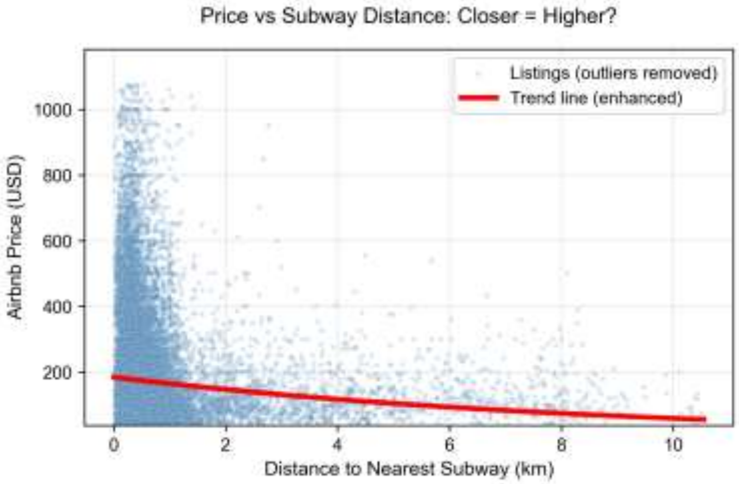


Result

- The spatial autocorrelation index Moran's I is 0.035 (p=0.026) : Weak but significant price clustering.

- In the regression model, the coefficient for subway distance is -0.05: 5% price drop per 1km.

- Model fit $R^2 = 0.45$: 45% price variation explained (distance & income)



- The red trendline illustrates the pattern that 'the farther from the subway, the lower the housing prices,' controlling for confounding factors such as community income.

Challenge

Data Format Inconsistency:

- The data contains numerous mixed string formats (e.g., price information: $100, $200, “350”, etc.).
- def parse_price(x): s = str(x).replace('$', '').replace(',', '').strip()

Outlier Impact on Regression:

- High-priced properties distort the results.
- y = np.log1p(df_model['price'])