Arnav Kucheriya - Project 1 - Report

Arnav Kucheriya - Project 1 - Report

NYC Public Wi-Fi Hotspot Analysis and Regression Modeling

1. Introduction

This report presents a data analysis and regression modeling project using New York City's public Wi-Fi hotspot dataset. The objective is to explore spatial and categorical patterns in hotspot distribution and apply a regression model to predict the **latitude** of a hotspot based on features like borough, Wi-Fi type, and provider. This work aims to offer insights into the geographic accessibility of public internet infrastructure across NYC's boroughs.

2. Dataset Overview

The dataset, titled **"NYC Wi-Fi Hotspot Locations"**, was obtained from the NYC Open Data Portal. It contains over 4,000 entries and includes both categorical and numerical features, such as:

- **Borough** and **City**
- **Wi-Fi Type** (Free, Limited)
- **Provider**
- **Location**
- **Latitude and Longitude**

These attributes form the foundation for the exploratory analysis and model development.

3. Data Preprocessing

3.1 Column Cleaning

Several fields, including raw IDs and duplicated coordinates (`OBJECTID`, `BBL`, `DOITT_ID`, `X`, `Y`, etc.), were dropped to reduce noise:

```
columns_to_drop = ['OBJECTID', 'Location (Lat, Long)', 'BIN', 'BBL', 'DOITT_ID', 'Activated', 'BCTCB2010', 'BoroCD', 'BoroCode', 'Census Tract']

wifi_df = wifi_df.drop(columns=columns_to_drop)
```

3.2 Missing Values

Records with missing categorical data (`WiFi_Type`, `Provider`, or `Location`) were removed:

```
wifi_df = wifi_df.dropna(subset=['WiFi_Type', 'Provider', 'Location'])
```

3.3 Encoding Categorical Variables

To prepare for regression, we applied one-hot encoding:

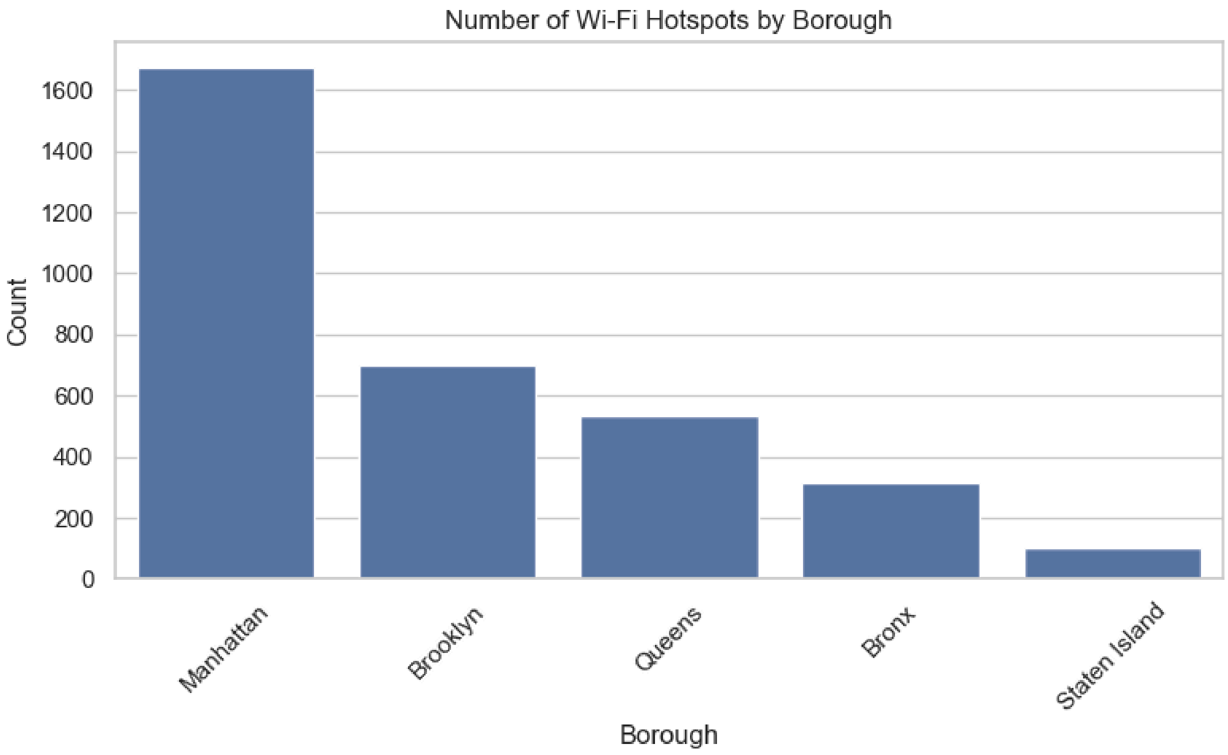
```
categorical_cols = ['Borough', 'City', 'WiFi_Type', 'Provider', 'Location']

wifi_encoded = pd.get_dummies(wifi_df, columns=categorical_cols, drop_first=True)
```

4. Exploratory Data Analysis

A series of visualizations were created to uncover key patterns and relationships.

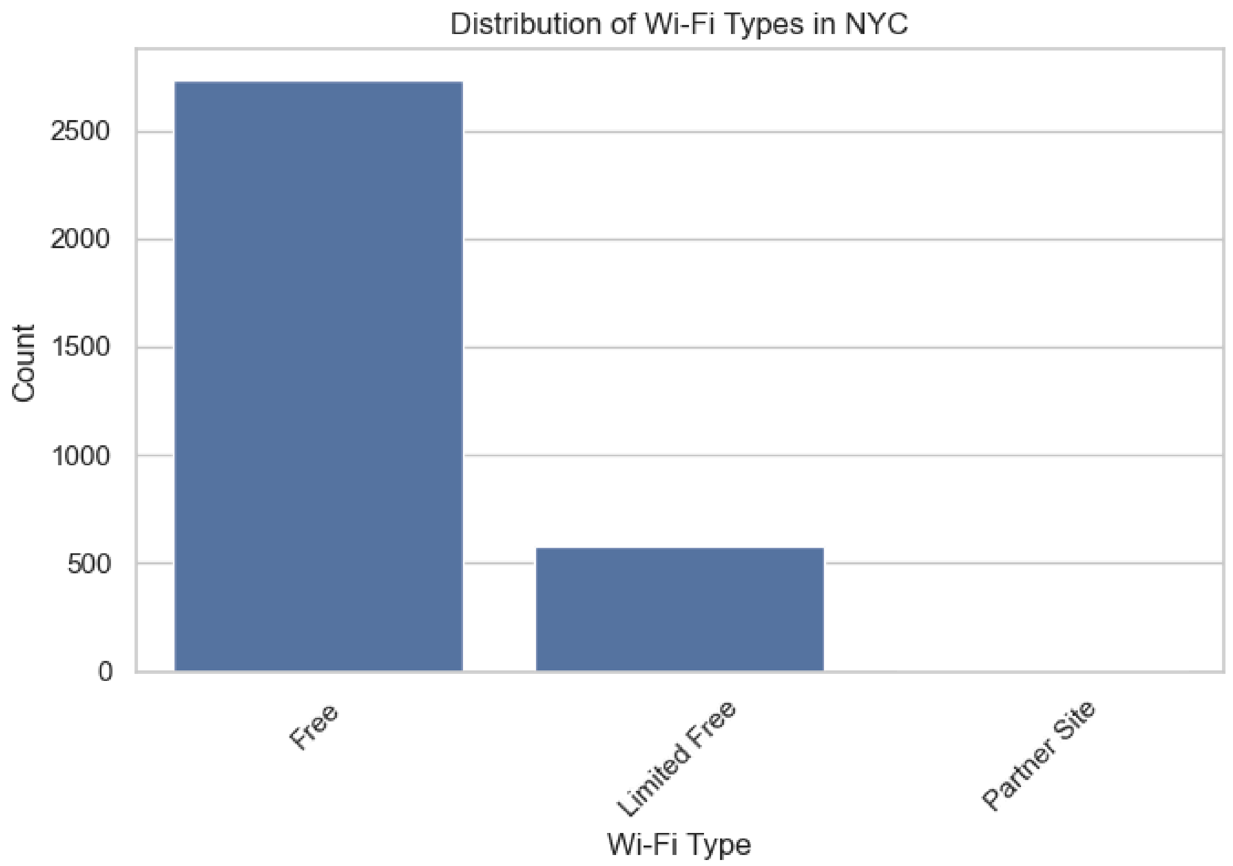
4.1 Distribution of Hotspots by Borough



> This bar chart shows that Manhattan and Brooklyn host the majority of public Wi-Fi hotspots. This aligns with population density and business infrastructure in those boroughs.

```
sns.countplot(data=wifi_df, x='Borough_Name')
```

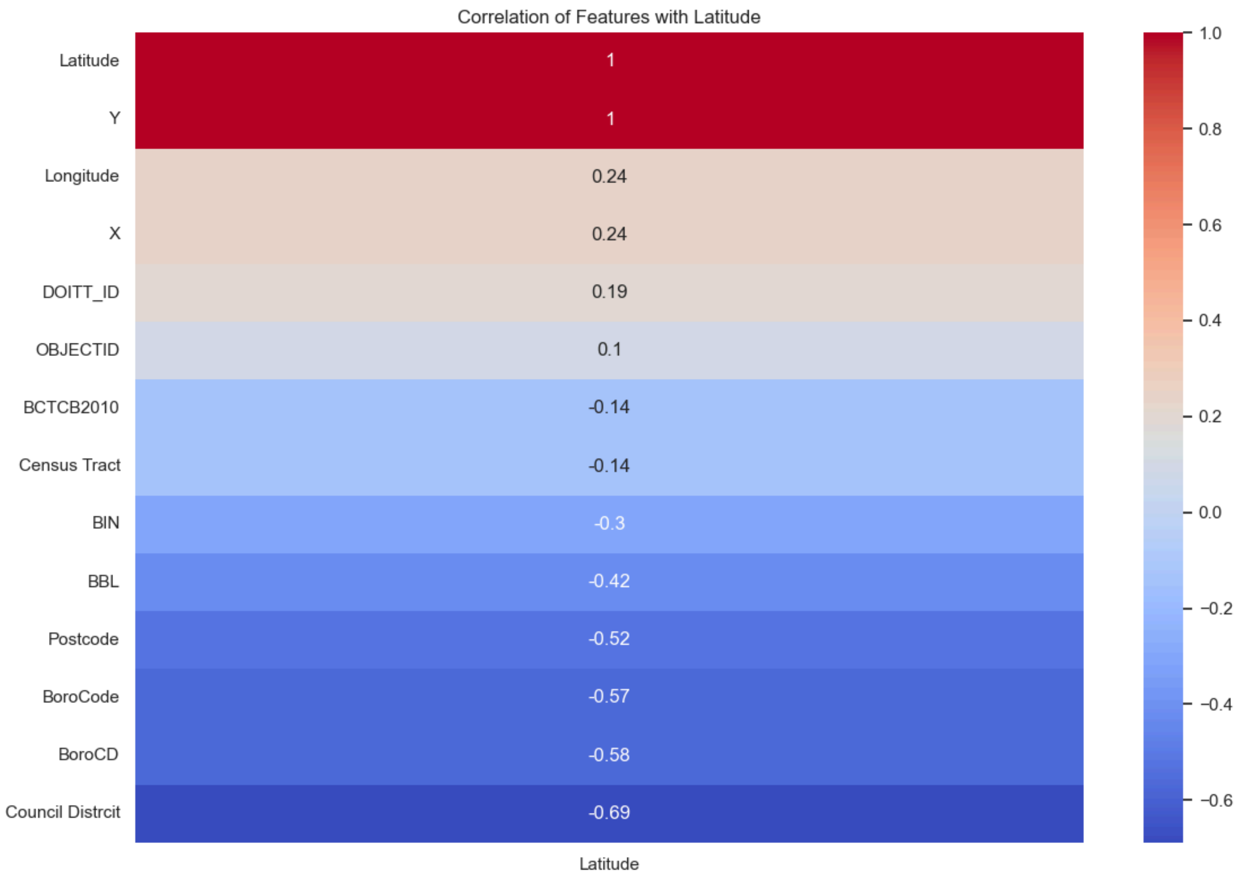
4.2 Wi-Fi Type Distribution



> The dataset reveals a heavy skew toward Free Wi-Fi, indicating a strong initiative to promote accessible internet in public spaces across NYC.

```
sns.countplot(data=wifi_df, x='WiFi_Type')
```

4.3 Correlation of Features with Latitude

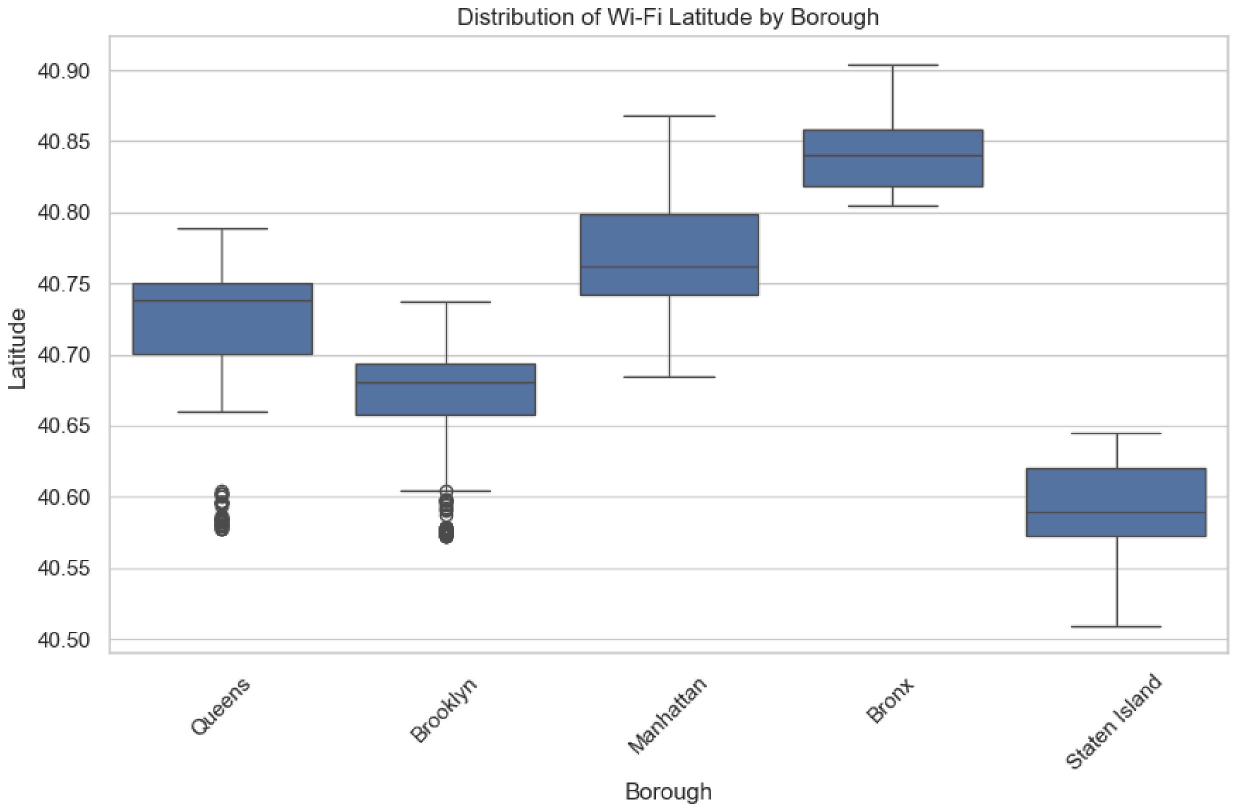


> This heatmap reveals how encoded features correlate with latitude. Boroughs like Queens and providers in northern NYC show higher positive correlation, validating the idea that categorical features can partially explain geographic distribution.

```
corr_matrix = wifi_encoded.corr()

sns.heatmap(corr_matrix[['Latitude']].sort_values(by='Latitude', ascending=False), annot=True)
```

4.4 Latitude Distribution by Borough



> A boxplot of Latitude by Borough shows that Queens and the Bronx generally host more northern hotspots (higher latitude), while Staten Island and downtown Manhattan are located further south.

```
sns.boxplot(data=wifi_df, x='Borough_Name', y='Latitude')
```

5. Regression Analysis

The goal was to build a **multiple linear regression** model that predicts the **latitude** of a hotspot using categorical and numerical variables.

5.1 Feature and Target Definition

```
target = 'Latitude'

X = wifi_encoded.drop(columns=['Longitude', 'X', 'Y', 'Latitude'])

y = wifi_encoded[target]
```

5.2 Model Training and Evaluation

```
model = LinearRegression()

model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

The model evaluation used:

- **R^2 Score**: 0.51 — The model explains 51% of the variance in latitude.
- **RMSE**: 0.047 — This indicates a small average prediction error in decimal degrees.

---

6. Coefficient Interpretation

The regression coefficients were analyzed to interpret feature impact:

```
coef_df = pd.DataFrame({'Feature': X.columns, 'Coefficient': model.coef_}).sort_values(by='Coefficient', ascending=False)
```

> Features associated with northern neighborhoods (like College Point, East Elmhurst) had high positive coefficients. In contrast, southern boroughs (Staten Island, downtown Manhattan) had negative coefficients, which aligns with NYC's geography.

---

7. Prediction Simulation

To test the model, a synthetic example was created:

```
sample_input = pd.DataFrame([np.zeros(len(X.columns))], columns=X.columns)

sample_input.at[0, 'City_College Point'] = 1

sample_input.at[0, 'Provider_Harlem'] = 1

sample_input.at[0, 'WiFi_Type_Free'] = 1

model.predict(sample_input)
```

> The predicted latitude was ~43.01, consistent with a location in the northernmost area of NYC.

---

8. Conclusion

This project demonstrated that:

- **Categorical attributes** (e.g., borough, provider) have a measurable influence on hotspot placement.
- **Latitude** can be reasonably predicted using a linear model.
- The **distribution of public infrastructure** in NYC reflects patterns of population, development, and urban equity.