NYC Public Wi-Fi Hotspot Analysis and Regression Modeling

Arnav Kucheriya
NJIT CS 301 - Introduction to Data Science

Project Overview

- **Goal:** Analyze Wi-Fi hotspot data and predict latitude placement.
- **Approach:** Data preprocessing, visualization, regression modeling, and evaluation.
- **Real-world Application:** Inform city planners and policymakers to ensure equitable internet access.

Objectives and Impact

- **Objectives:**
  - Explore data and visualize trends.
  - Preprocess data for modeling.
  - Build and evaluate a regression model.
  - Simulate predictions using synthetic input.
- **Impact:**
  - Ensure **equitable urban development**.
  - Guide future hotspot deployment.
  - Improve internet accessibility.

Dataset Overview

Dataset Name: NYC Wi-Fi Hotspot Locations

Source: NYC Open Data Portal

Rows/Records: 4,000+

Features:

- Borough, City, Wi-Fi Type, Provider
- Latitude and Longitude

Key Attributes in Dataset

- **Borough and City:** Geographical details.
- **Wi-Fi Type:** Free or limited.
- **Provider:** Organizations maintaining hotspots.
- **Location Description:** Exact location of the hotspot.
- **Latitude and Longitude:** Spatial coordinates for modeling.

Data Preprocessing Overview

Column Cleaning: Removed redundant columns.

Handling Missing Values: Dropped records with null values.

Encoding Categorical Variables: One-hot encoding was applied.

Data Preprocessing Overview - Column Cleaning

Column Cleaning:

- Dropped irrelevant fields such as:
  - `OBJECTID`, `BBL`, `DOITT_ID` — Unique identifiers that provide no predictive value.
  - `X`, `Y`, and `Location (Lat, Long)` — Duplicates of latitude and longitude data.
- **Reason:** These columns contained administrative or internal IDs that do not contribute to spatial analysis or prediction.

Data Preprocessing Overview - Handling Missing Values

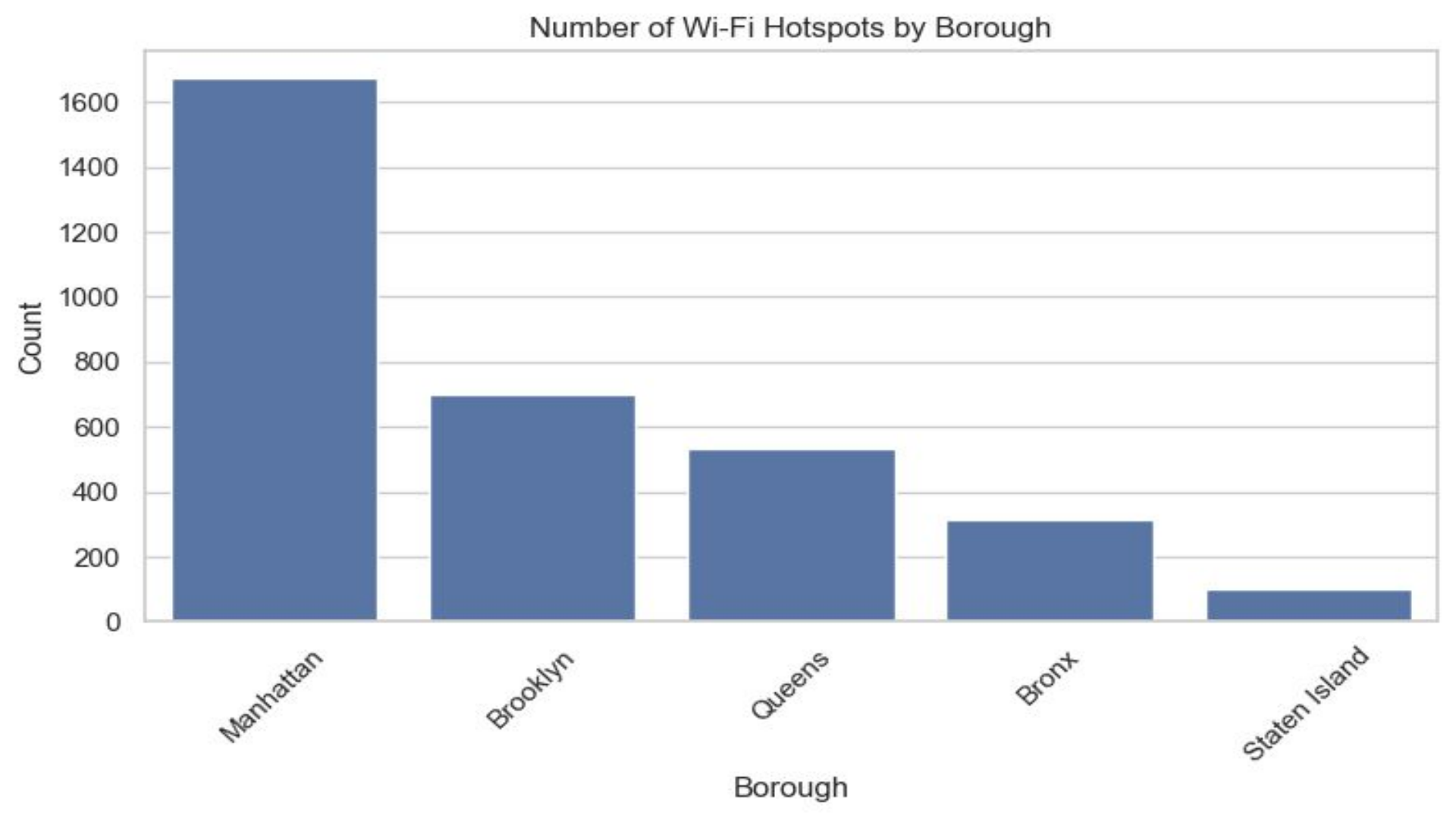
Handling Missing Values:

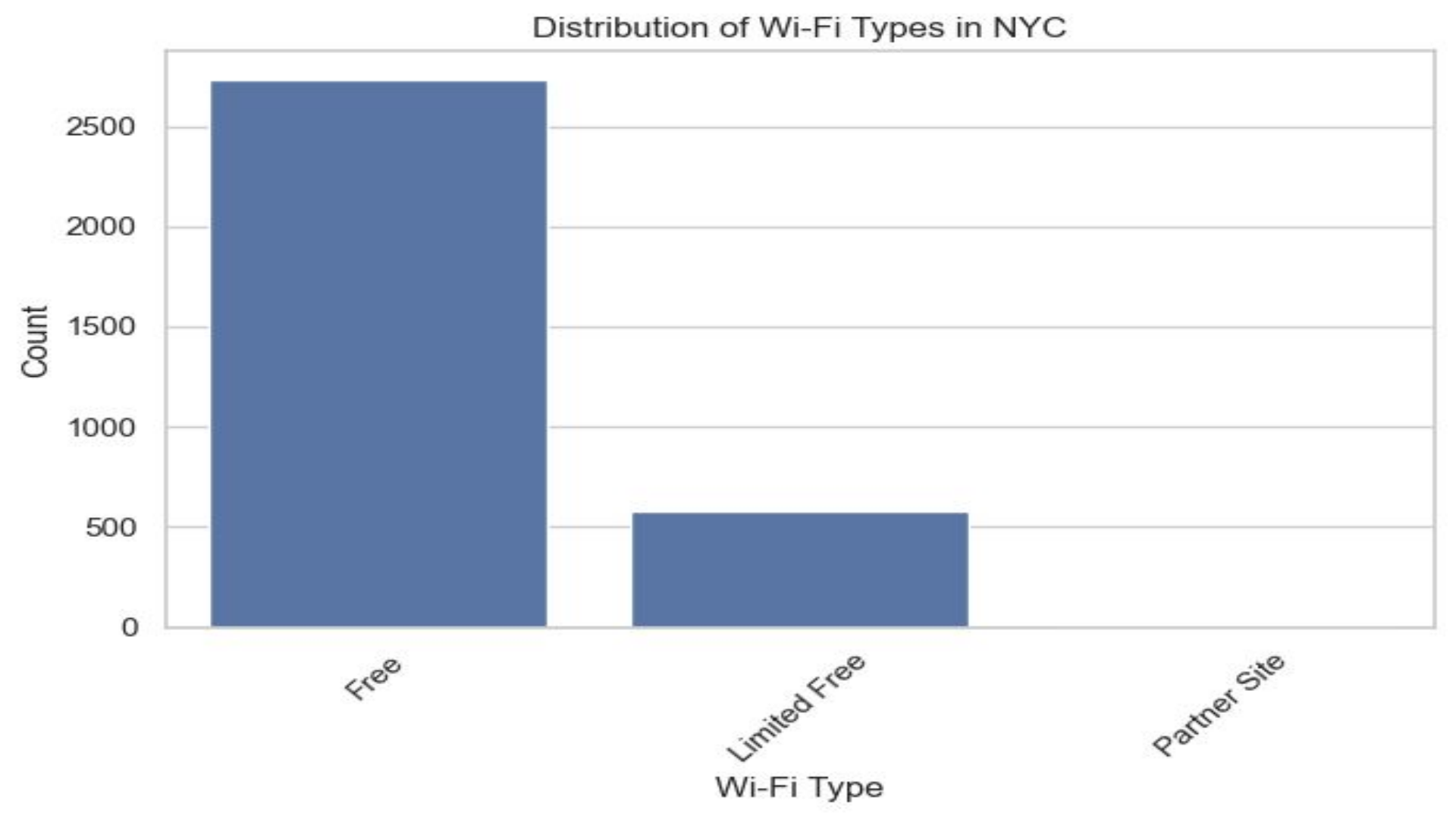
- Dropped rows with null values in:
  - `WiFi_Type`, `Provider`, and `Location` — Core categorical fields required for regression.
- **Reason:** These attributes are essential for one-hot encoding and model training. Retaining incomplete records could distort the model's learning.

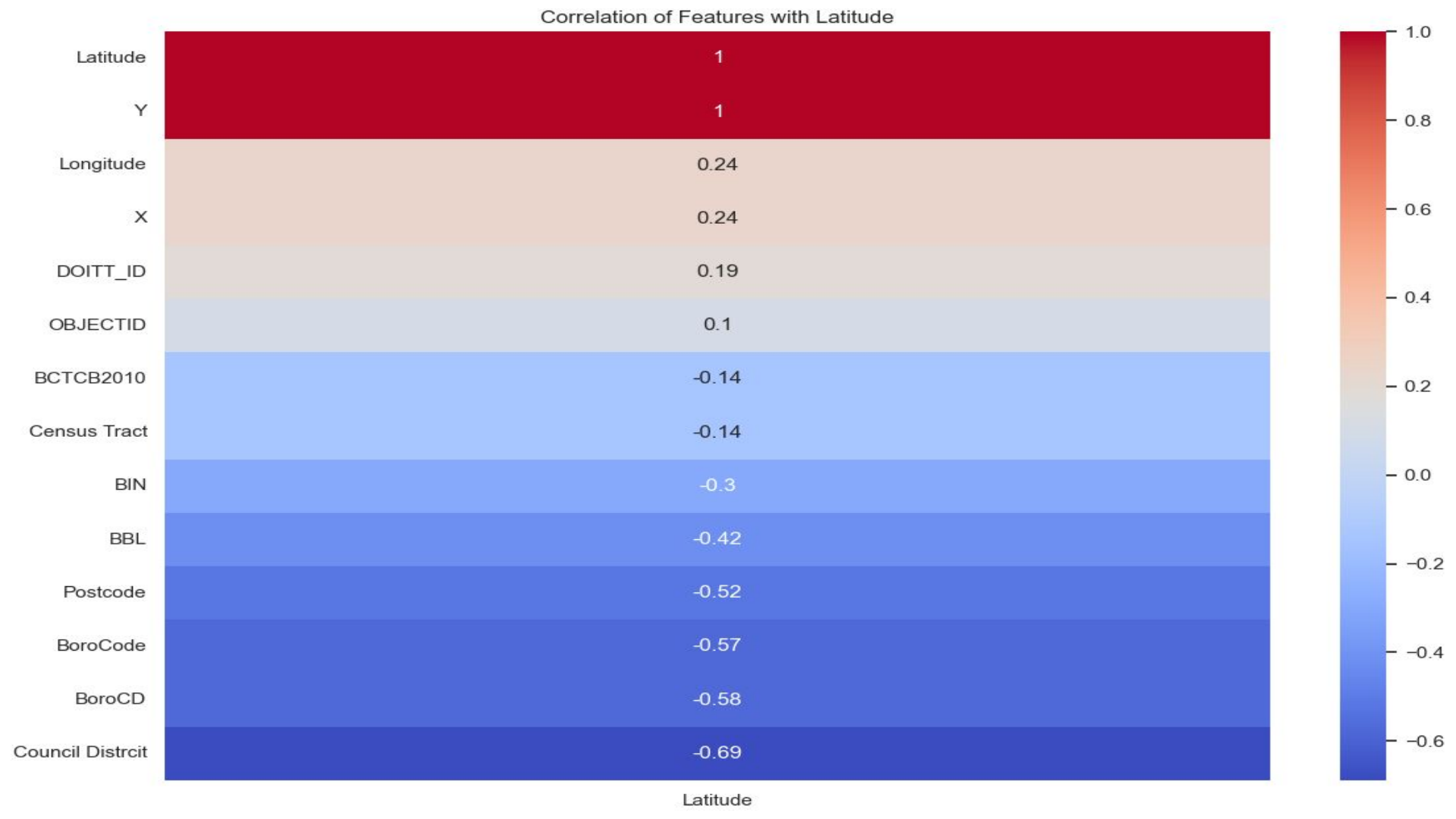
Data Preprocessing Overview Encoding Categorical Var

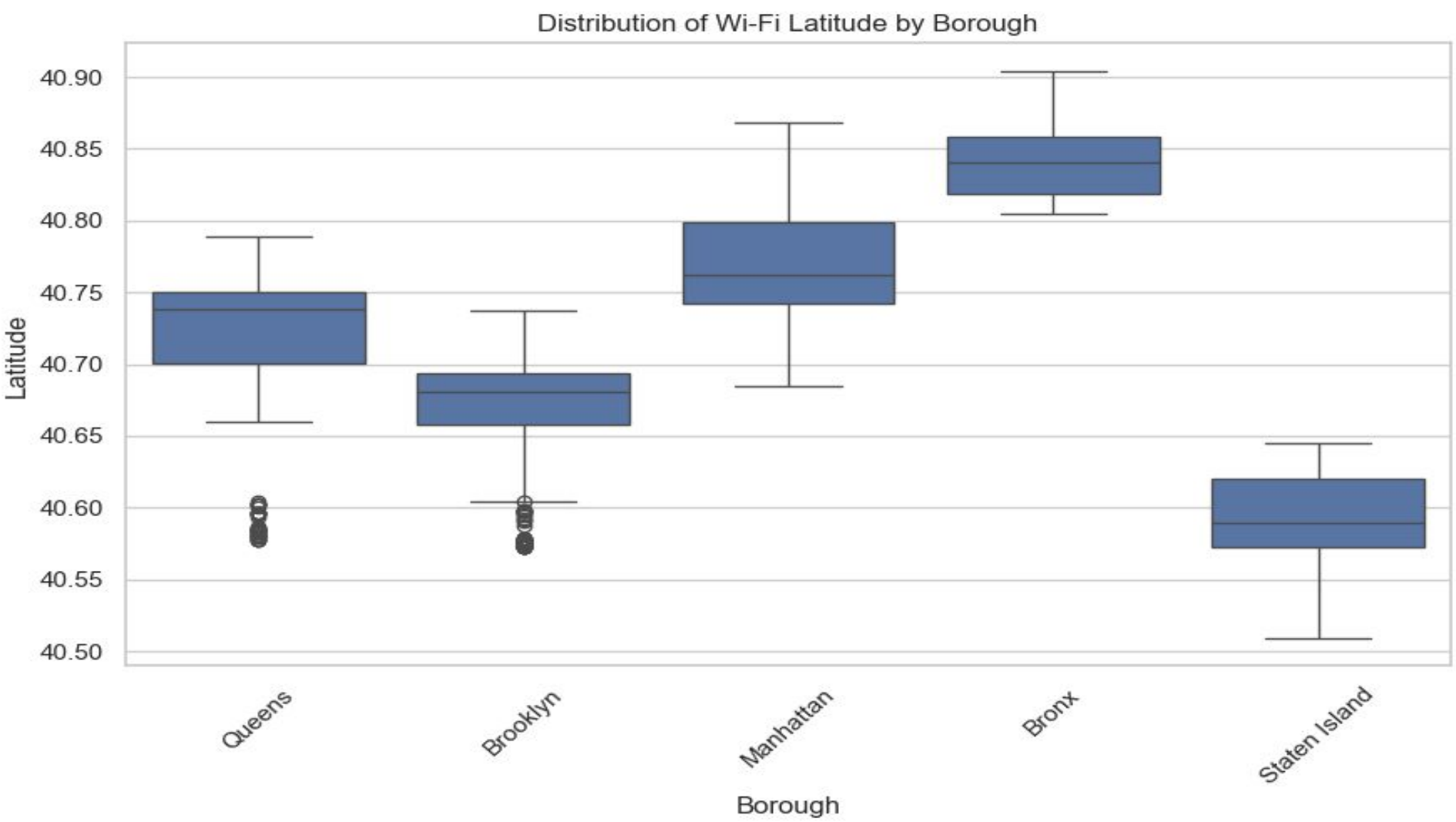
Encoding Categorical Variables:

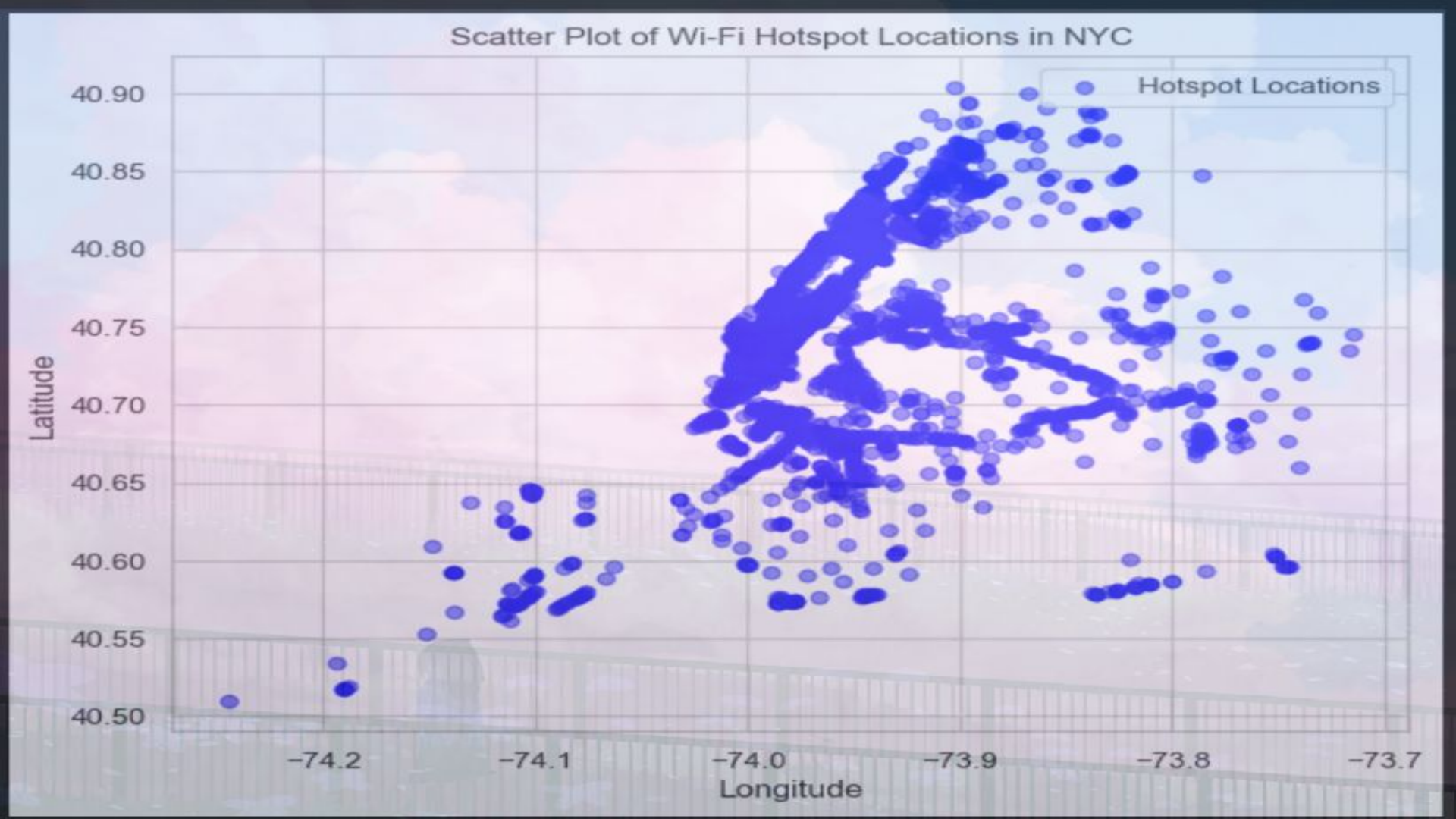
- One-hot encoding was applied to:
  - `Borough`, `City`, `WiFi_Type`, `Provider`, and `Location`
- **Reason:** To convert categorical variables into numerical format for regression.


Number of Wi-Fi Hotspots by Borough
Count
0
200
400
600
800
1000
1200
1400
1600
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Borough


Distribution of Wi-Fi Types in NYC
Count
2500
2000
1500
1000
500
0
Free
Limited Free
Partner Site
Wi-Fi Type


Correlation of Features with Latitude
Latitude
Y
Longitude
X
DOITT_ID
OBJECTID
BCTCB2010
Census Tract
BIN
BBL
Postcode
BoroCode
BoroCD
Council Distrcit
1
1
0.24
0.24
0.19
0.1
-0.14
-0.14
-0.3
-0.42
-0.52
-0.57
-0.58
-0.69
Latitude
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Distribution of Wi-Fi Latitude by Borough
Latitude
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
Queens
Brooklyn
Manhattan
Bronx
Staten Island
Borough


Scatter Plot of Wi-Fi Hotspot Locations in NYC
Hotspot Locations
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
Latitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Longitude

Regression Model Setup

Model: Multiple Linear Regression

Target Variable: Latitude

Train-Test Split: 80/20 ratio

Feature and Target Definition

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

y = wifi_encoded['Latitude']
```

Coefficient Interpretation

- **Positive Coefficients:** Influence northern placement (e.g., College Point, Harlem).

- **Negative Coefficients:** Influence southern placement (e.g., Staten Island, Downtown Manhattan).

Prediction Simulation

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

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

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

Ridge and Decision Tree Model

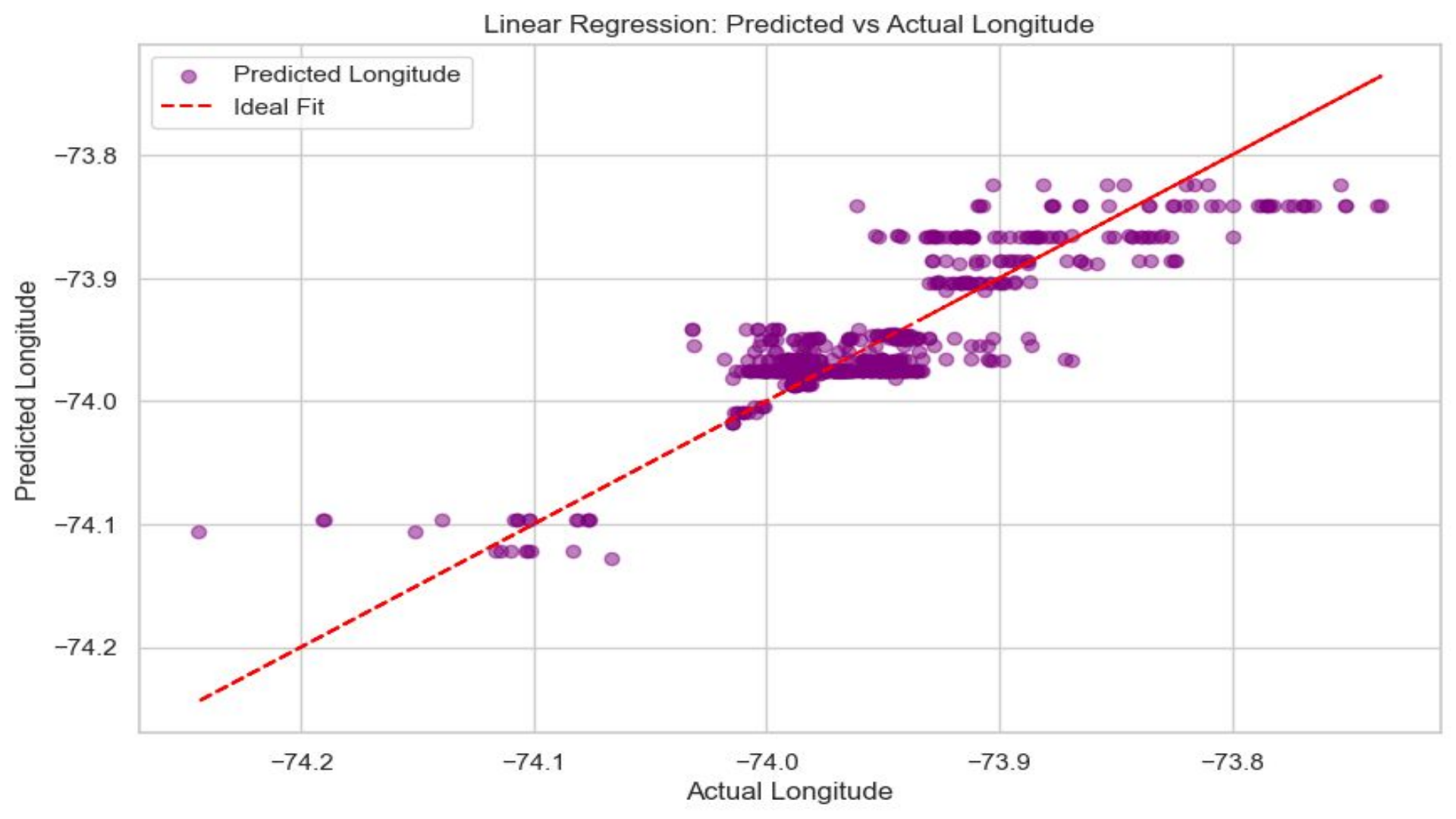
- Ridge Regression: Prevents overfitting by adding a penalty term.
- Decision Tree Regression: Captures non-linear relationships through decision splits.
- Improved model performance compared to Linear Regression.

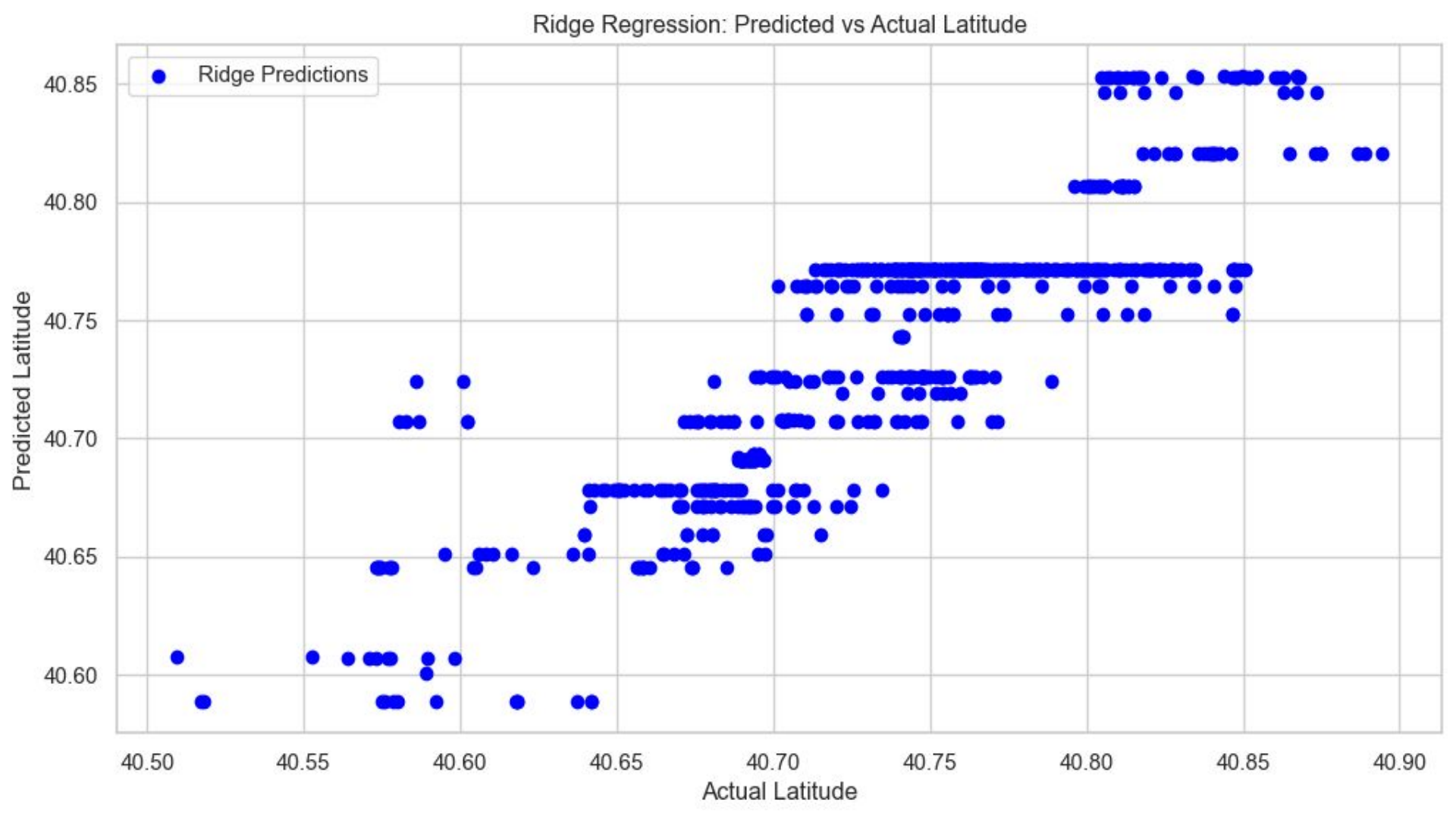
Model Performance Comparison

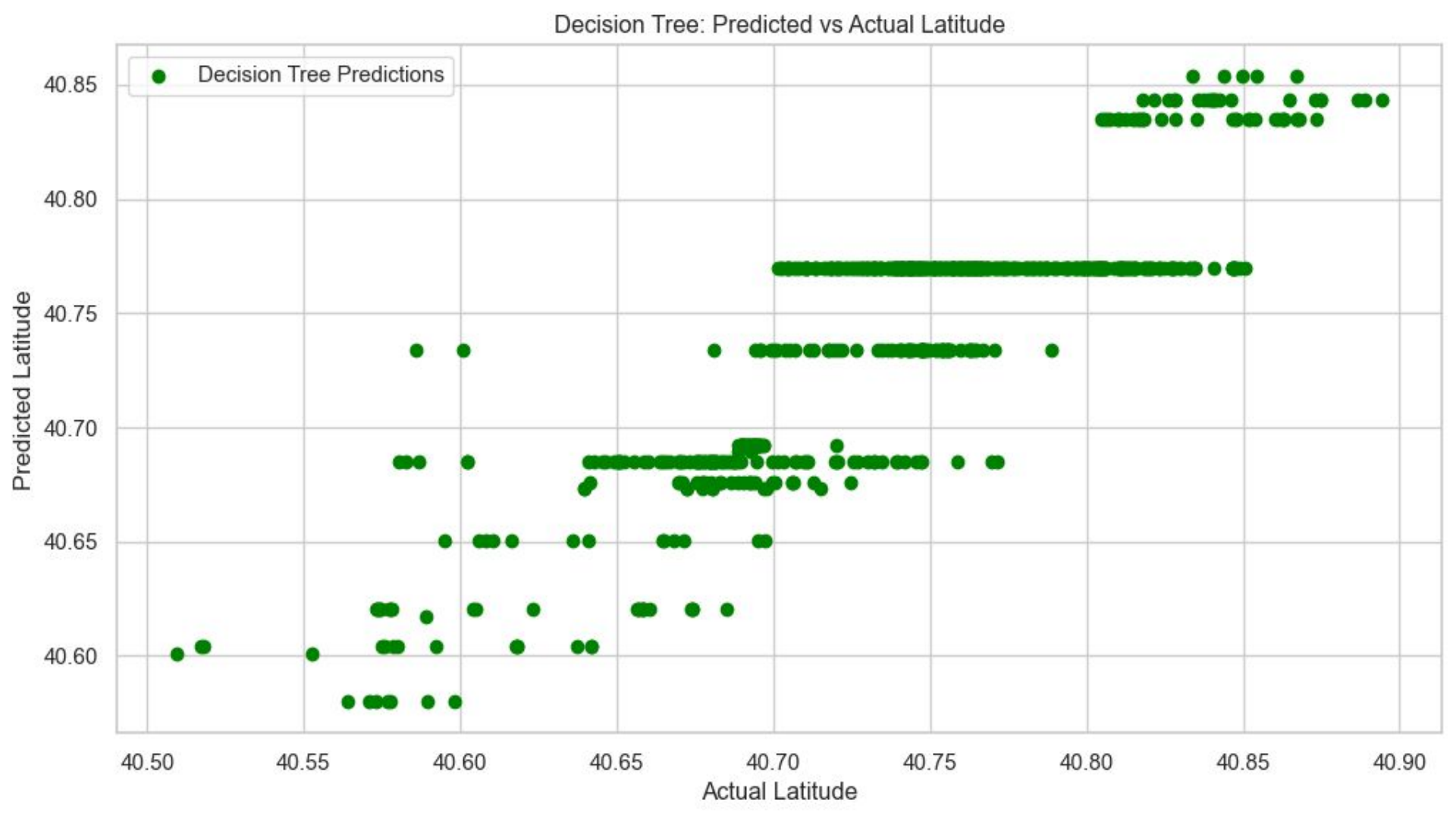
Model	R^2 Score	MSE
Linear Regression	0.51	0.047
Ridge Regression	0.76	0.0011
Decision Tree Regression	0.75	0.00113

Predicted vs. Actual Latitude Visualization

- The scatter plot of predicted vs actual longitude values shows a strong alignment.
- Predicted values closely follow the ideal fit line.
- The model effectively captured longitude coordinates with minimal error.


Linear Regression: Predicted vs Actual Longitude
Predicted Longitude
Ideal Fit
−73.8
−73.9
−74.0
−74.1
−74.2
Predicted Longitude
−74.2
−74.1
−74.0
−73.9
−73.8
Actual Longitude


Ridge Regression: Predicted vs Actual Latitude
Ridge Predictions
Predicted Latitude
40.85
40.80
40.75
40.70
40.65
40.60
40.50
40.55
40.60
40.65
40.70
40.75
40.80
40.85
40.90
Actual Latitude


Decision Tree: Predicted vs Actual Latitude
Decision Tree Predictions
Predicted Latitude
Actual Latitude
40.85
40.80
40.75
40.70
40.65
40.60
40.50
40.55
40.60
40.65
40.70
40.75
40.80
40.85
40.90

Conclusion

Key Takeaways:

- Model captured borough-level patterns.
- Reasonable predictive performance with limited features.