DS5110 – FINAL PROJECT

RENTAL PREDICTION

-ESHA ACHARYA, VIVEK DANTU

AGENDA

- Aim of our Project
- Literature Review
- Methodology – Design Diagram
- Tools Used
- Data Collection
- Cleaning
- Removing outliers
- Data Transformation
- Data Visualization
- Statistical Analysis
- Data Training
- Prediction using Linear Regression
- Discussion
- Conclusion
- Future Work/Limitations
- References

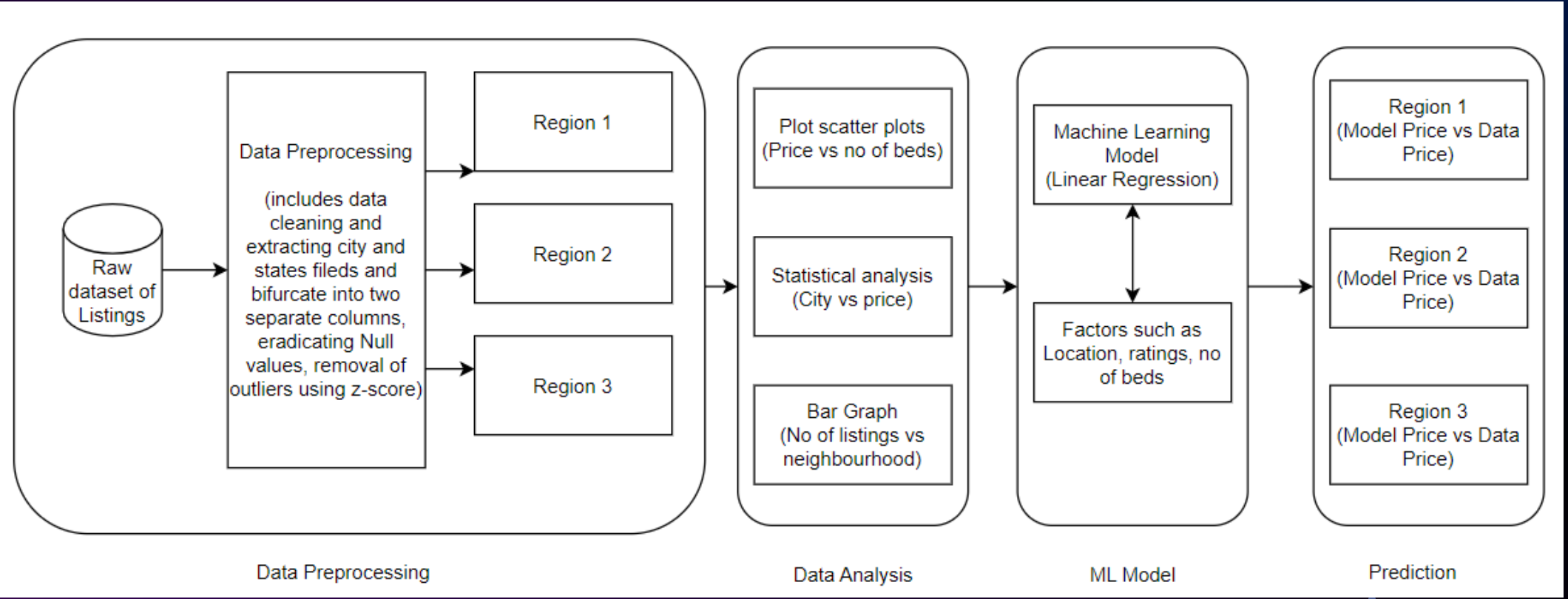
AIM

- This project aims to understand the dynamics of the rental ecosystem across the United States by exploring factors influencing demand, pricing strategies, and predictive modeling.
- The plan is to clean the large dataset, preprocess and analyse the data as per the requirements of our project, take into consideration only the data that matters and remove outliers, and finally, predict prices using machine learning models like linear regression.
- Key findings highlight model accuracy, price distribution, the correlation between price and rating, and provide actionable insights for hosts and travelers.

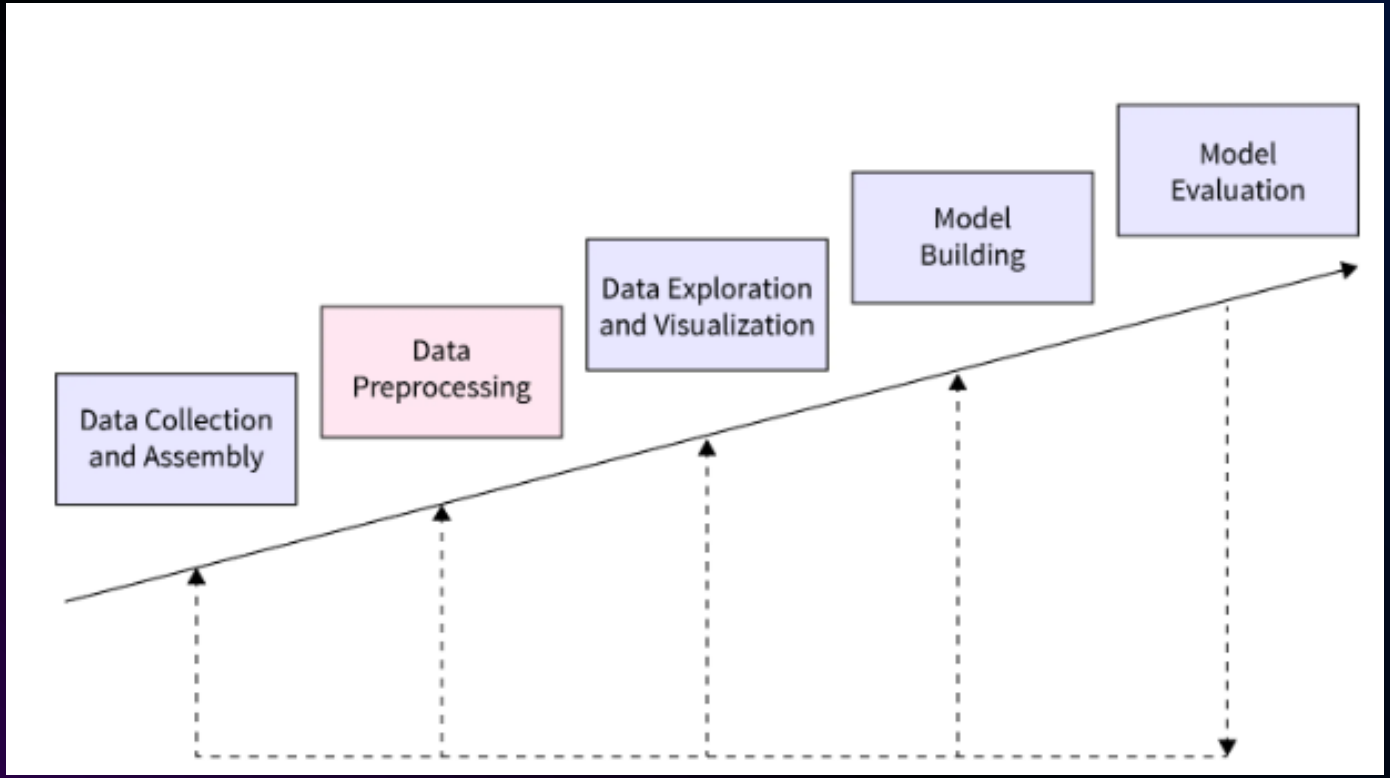
LITERATURE REVIEW

- **Economic Factors**: Studies emphasize the role of interest rates, inflation, and employment in shaping real estate prices and rentals.
- **Demographic Influences**: Urban population growth and age distribution significantly impact property demand and rental costs.
- **Geographic Determinants**: Location, proximity to amenities, and neighborhood quality are crucial for property valuation, supported by hedonic pricing models (Rosen, 1974; Kok, Monkkonen, & Quigley, 2014).
- **Technological Advancements**: Big data and machine learning enhance real estate analysis, improving predictive accuracy with models like regression and neural networks (Choi & Varian, 2012; Wang & Zhou, 2018).

METHODOLOGY - DESIGN DIAGRAM



BUILDING BLOCKS



TOOLS USED

- Python
- Pandas
- NumPy
- Matplotlib and Seaborn
- Scikit-learn
- Jupyter Notebook

DATA COLLECTION

- **Data Sources:** Listings from Seattle, Los Angeles, San Francisco (Region 1); New York City, Boston, Jersey City (Region 2); Twin Cities, Chicago, Columbus (Region 3)
- **Number of Listings:** Initial counts for each region before outlier removal:
  - Region 1: 60,838
  - Region 2: 45,072
  - Region 3: 17,081

CLEANING

- Extraction of city and state from filenames
  - Eg: listings_boston_MA
- Identification of key attributes (e.g., number of beds, ratings, `is_regularly_available`)
  - By using regex like - Rating: (\d+\.\d+) and (\d+) beds
  - `is_regularly_available` determined based on availability > 10 days
- Handling missing values
  - Null values and blanks

REMOVING OUTLIERS

1. We removed approximately 1000 rows from our csv files which consisted of outliers. A threshold of value 1.5 was set to detect the outliers in prices.
2. Standard deviation and median were implemented to calculate the outliers.

```
## Removing Outliers from the Regions data

threshold = 1.5

region_1_data_frames['z_score'] = zscore_helper.zscore(region_1_data_frames['price'])
outliers = (region_1_data_frames['price'] < (region_1_data_frames['price'].median() - threshold * region_1_data_f
region_1_data_frames = region_1_data_frames[~outliers]
region_1_data_frames = region_1_data_frames.drop(columns=['z_score'])

region_2_data_frames['z_score'] = zscore_helper.zscore(region_2_data_frames['price'])
outliers = (region_2_data_frames['price'] < (region_2_data_frames['price'].median() - threshold * region_2_data_f
region_2_data_frames = region_2_data_frames[~outliers]
region_2_data_frames = region_2_data_frames.drop(columns=['z_score'])

region_3_data_frames['z_score'] = zscore_helper.zscore(region_3_data_frames['price'])
outliers = (region_3_data_frames['price'] < (region_3_data_frames['price'].median() - threshold * region_3_data_f
region_3_data_frames = region_3_data_frames[~outliers]
region_3_data_frames = region_3_data_frames.drop(columns=['z_score'])

print("AFTER OUTLIERS REMOVAL SIZES OF EACH REGION:")
print(region_1_data_frames.shape)
print(region_2_data_frames.shape)
print(region_3_data_frames.shape)
```

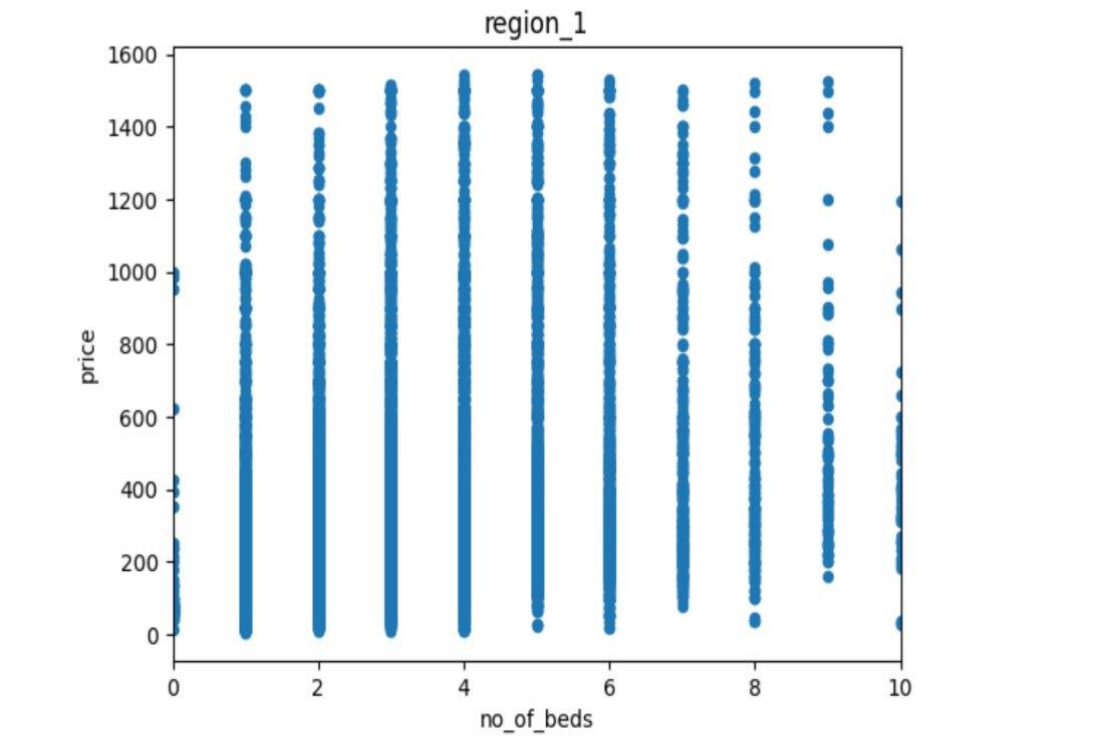
DATA TRANSFORMATION

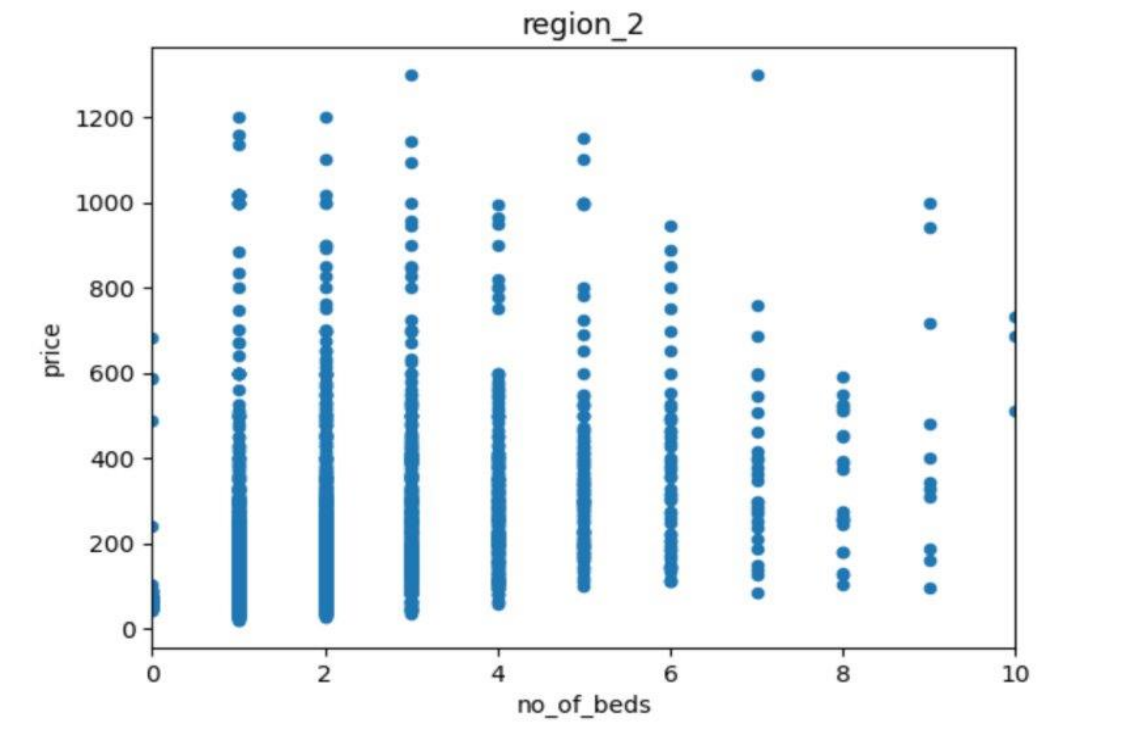
1. After cleaning the data to include the necessary columns and rows, we divided our dataset into three main regions: California, Boston, and Minnesota.
2. **Post-Cleaning Dataset Sizes:**
   - Region 1: 59,805
   - Region 2: 44,874
   - Region 3: 16,948

```
region_3 = ["listings_twincities_MS.csv", "listings_chicago_IL.csv",
            "listings_columbus_OH.csv"]
region_3_data_frames=[]
for region in region_3:
  data_frame=pd.read_csv(region)
  parts = region.split('/')
  filename_part = parts[-1]
  filename_without_extension = filename_part.split('.')[0]
  city_state = filename_without_extension.split('_')[-2:]
  data_frame["city"] = city_state[0]
  data_frame["state"] = city_state[1]
  region_3_data_frames.append(data_frame)
region_3_data_frames = pd.concat(region_3_data_frames, ignore_index=True)
region_3_data_frames["is_regularly_available"] = (region_3_data_frames["availability_365"] > 10)
new_column_name = "no_of_beds"
split_data = region_3_data_frames["name"].str.split(" . ", expand=True)[3].rename(new_column_name)
region_3_data_frames = pd.concat([region_3_data_frames.drop("name", axis=1), split_data], axis=1)
region_3_data_frames['no_of_beds'] = region_3_data_frames['no_of_beds'].str.extract('(\d+)')[0]
region_3_data_frames['no_of_beds'] = region_3_data_frames['no_of_beds'].astype('Int64')
print("TWIN_CITIES-CHICAGO-COLUMBUS:")
print(region_3_data_frames)
```

3. From the listings tables, we extracted and transformed by adding the number of beds and utilized the number of beds to analyse the rental property prices based on the number of bedrooms as shown in the visualization graphs ahead.

DATA VISUALIZATION





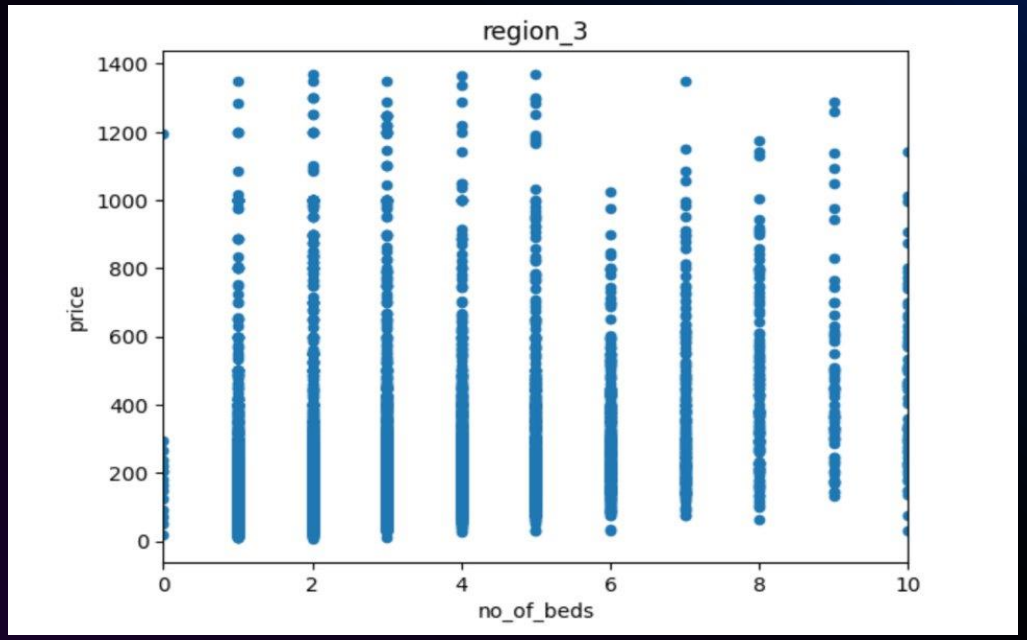
DATA VISUALISATION

•**Region 1:** Prices generally increase with the number of beds, though not perfectly linear; other factors like location and reviews could also influence pricing.

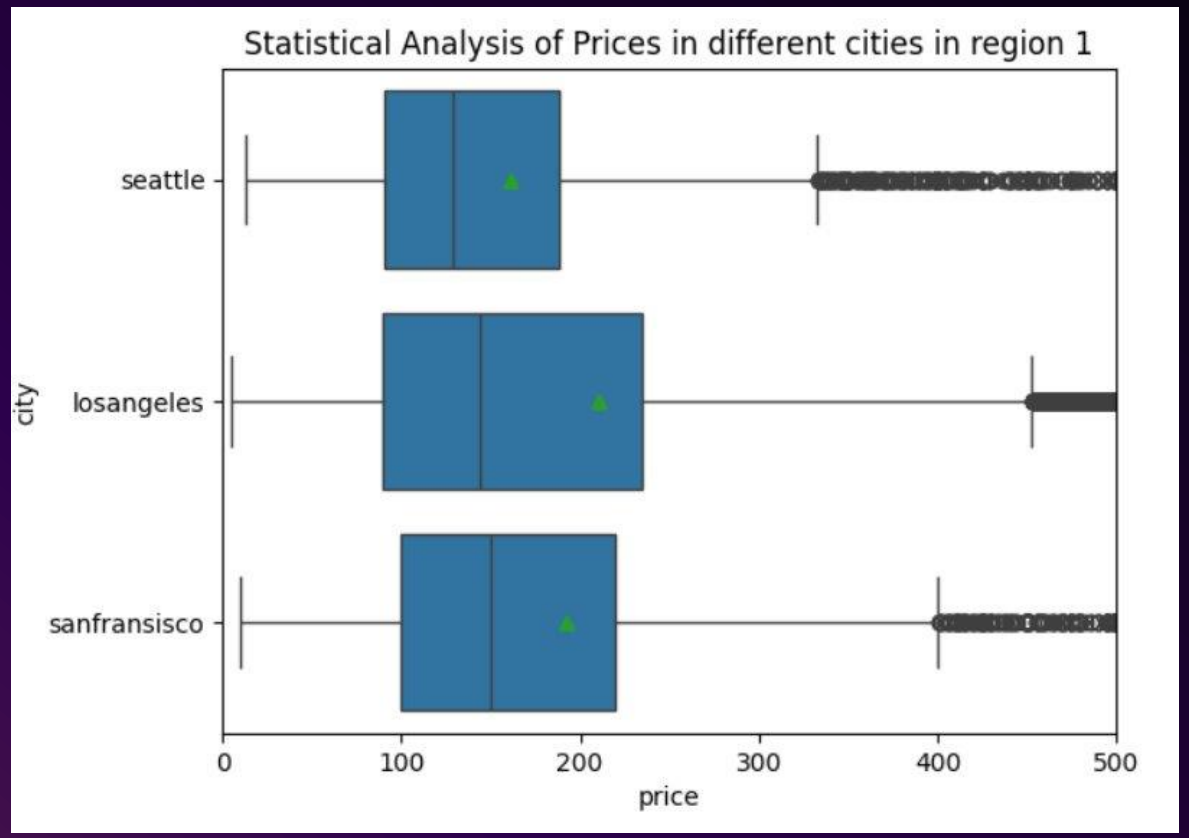
•**Region 2:** Outliers show high prices with fewer beds; scatter suggests less consistency in price correlation with the number of beds.

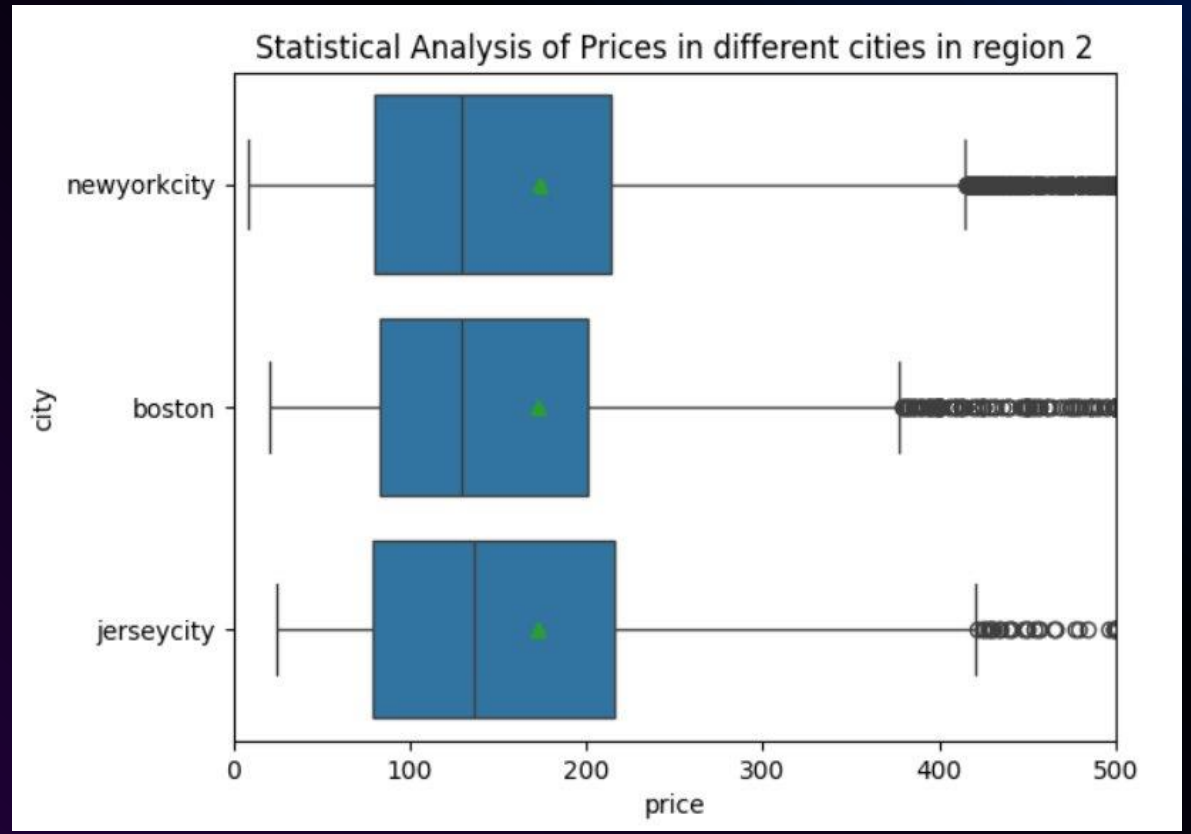
•**Region 3:** Higher prices are associated with fewer beds, indicating a potential premium pricing strategy or more standardized pricing.

•**Overall Trend:** While the number of beds impacts pricing, variability and outliers across regions highlight the importance of additional factors in determining final prices.



STATISTICAL ANALYSIS





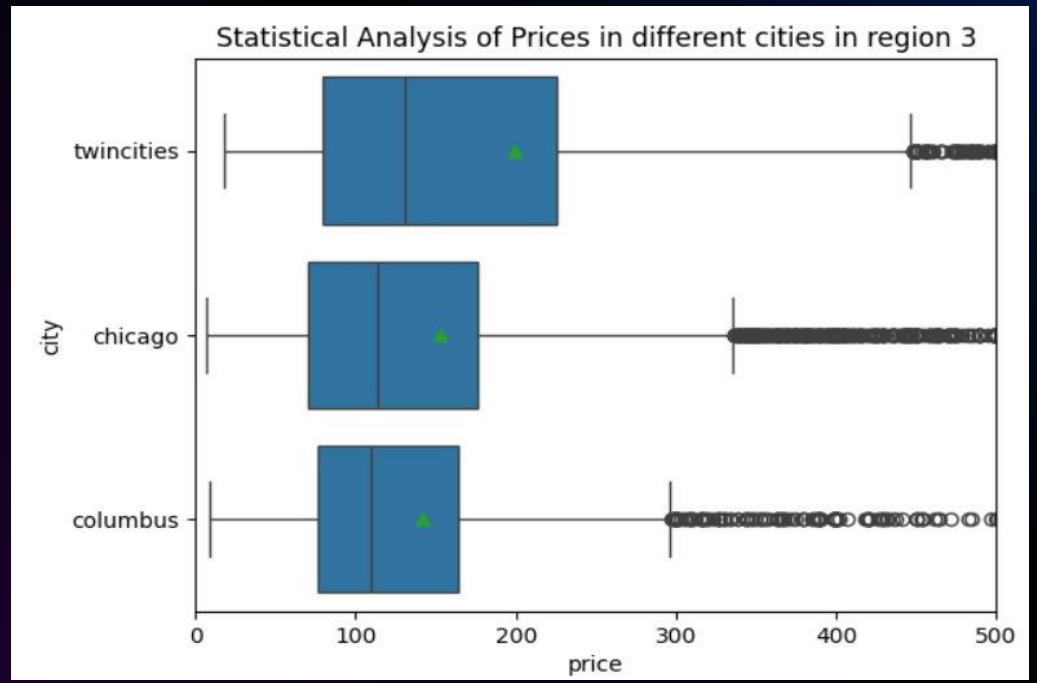
STATISTICAL ANALYSIS

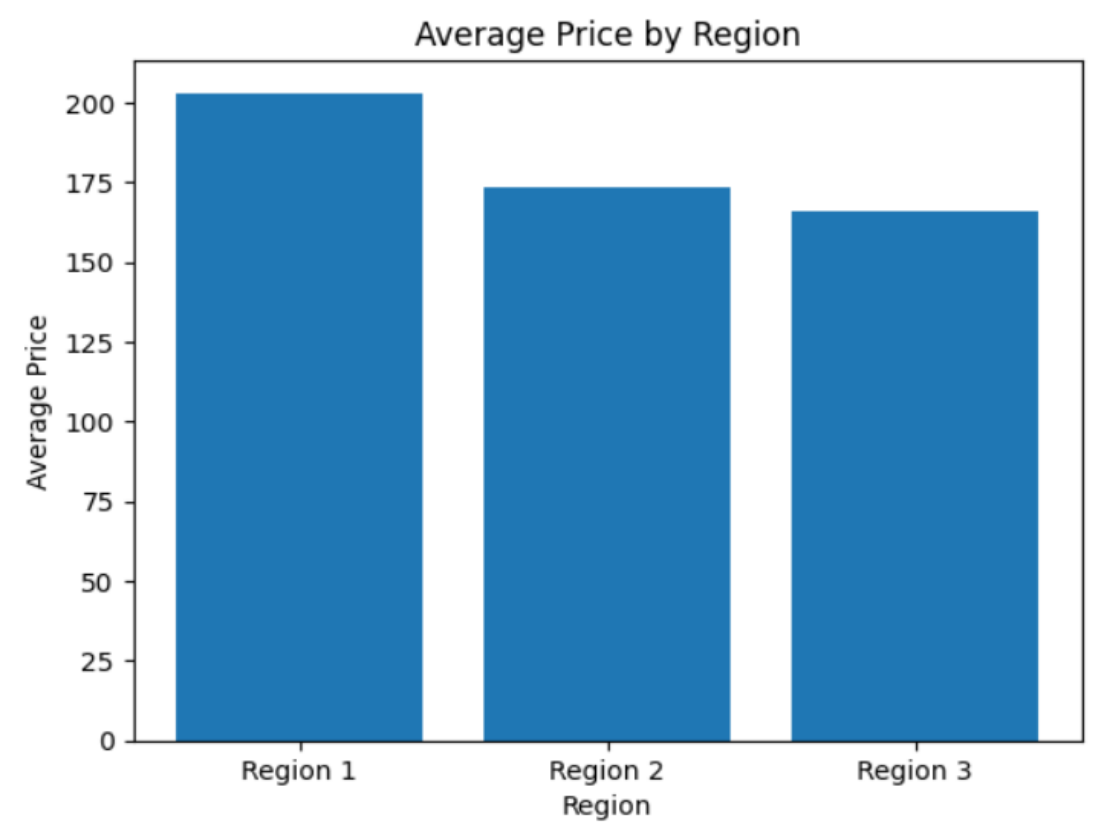
•**Region 1:** LA shows the highest upper quartile and price variability, with Seattle moderately priced and San Francisco having the lowest upper quartile among the three cities.

•**Region 2:** New York has the widest price range and highest variability, while Boston shows lower and less variable prices, and New Jersey falls in between with substantial variability.

•**Region 3:** The Twin Cities exhibit the highest upper quartile and price variability, outpacing Chicago and Columbus, indicating a market with a concentration of higher-end properties.

•**Overall Trends:** Across regions, boxplots reveal significant variability and outliers, highlighting local market conditions and the impact of unique factors on property pricing.

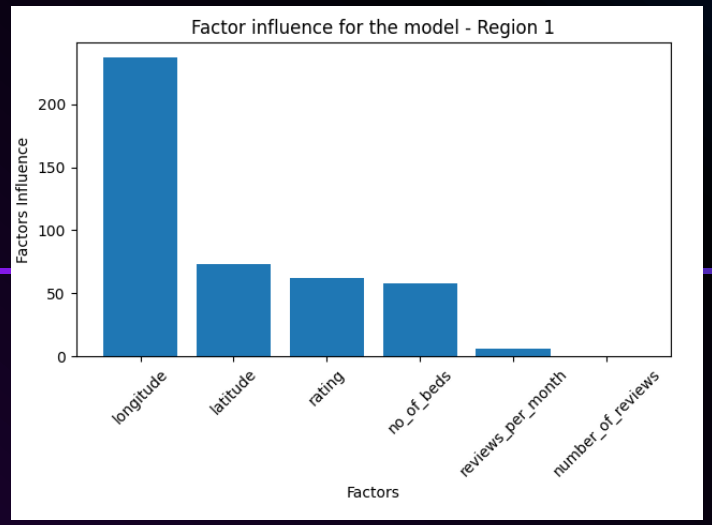


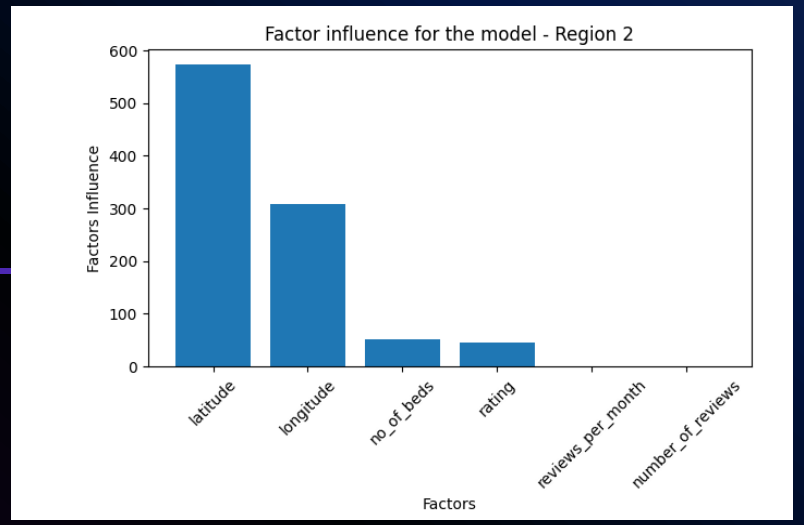


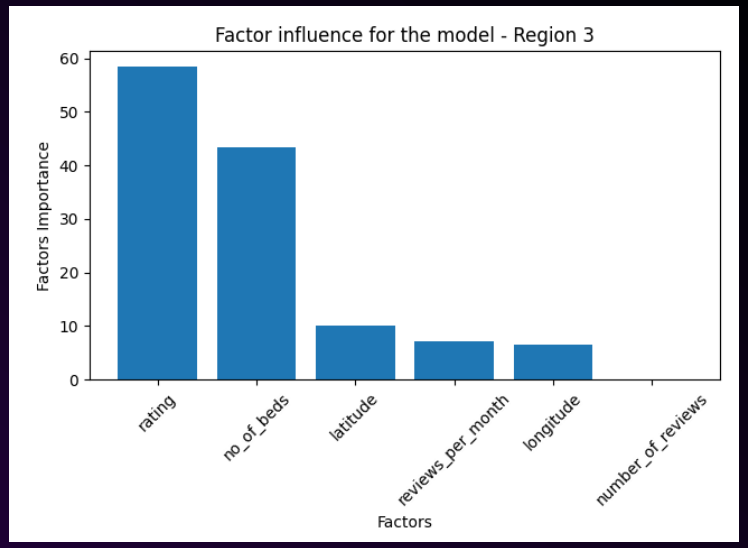
- **Region 1:** Highest average price at $200,000, indicating the most premium and expensive market.
- **Region 2:** Moderate average price at $170,000, suggesting a balanced market with diverse property values.
- **Region 3:** Lowest average price at $165,000, making it the most affordable option for property investments.
- **Overall Trend:** Clear gradient in property values with Region 1 being the most expensive, Region 2 moderately priced, and Region 3 the most affordable.

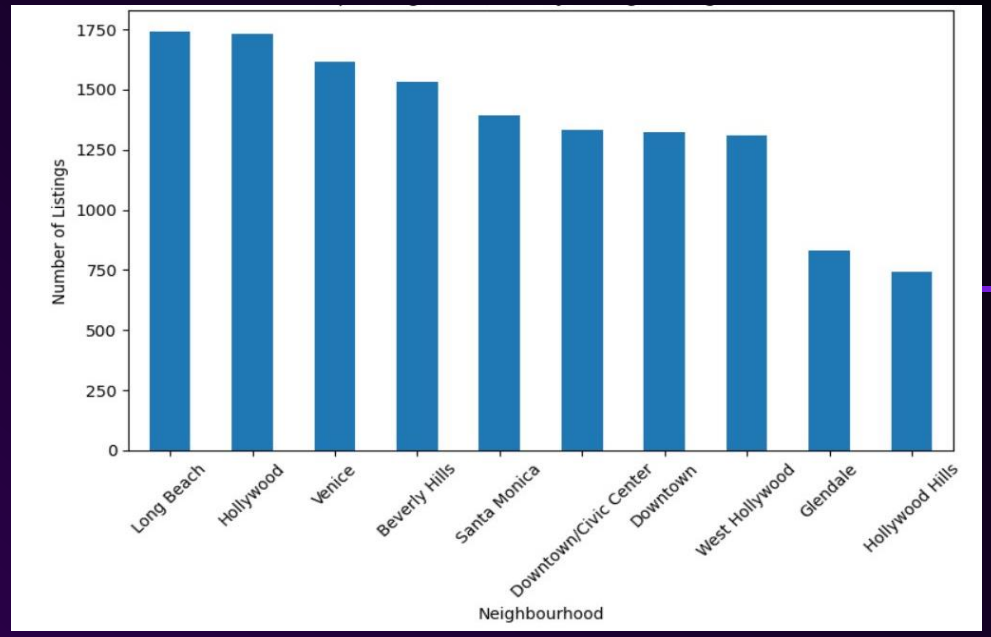
DATA TRAINING

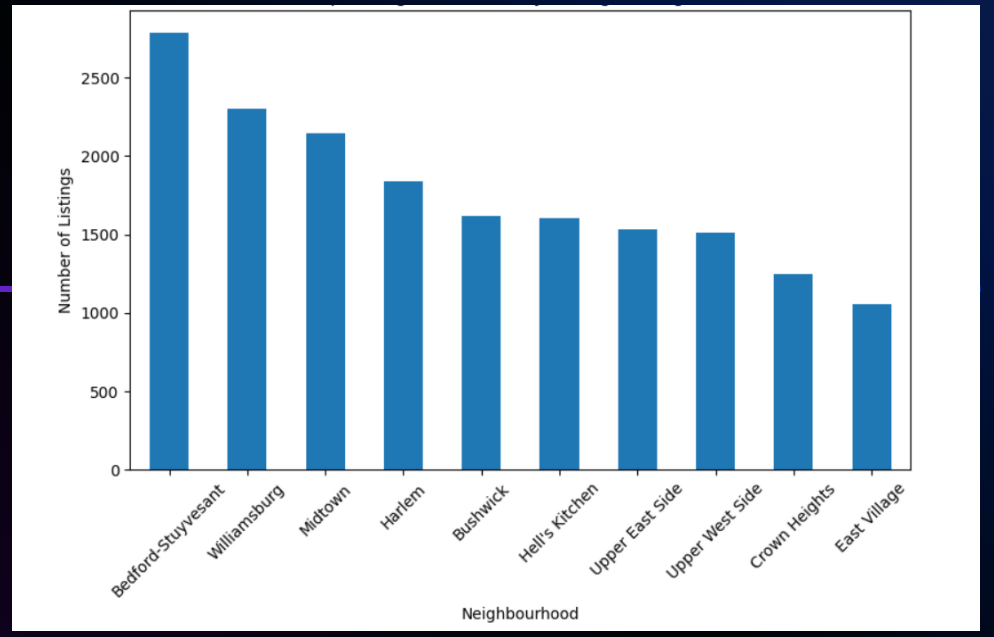
Key Analysis Factors: Price, number of beds, ratings, availability, longitude's impact

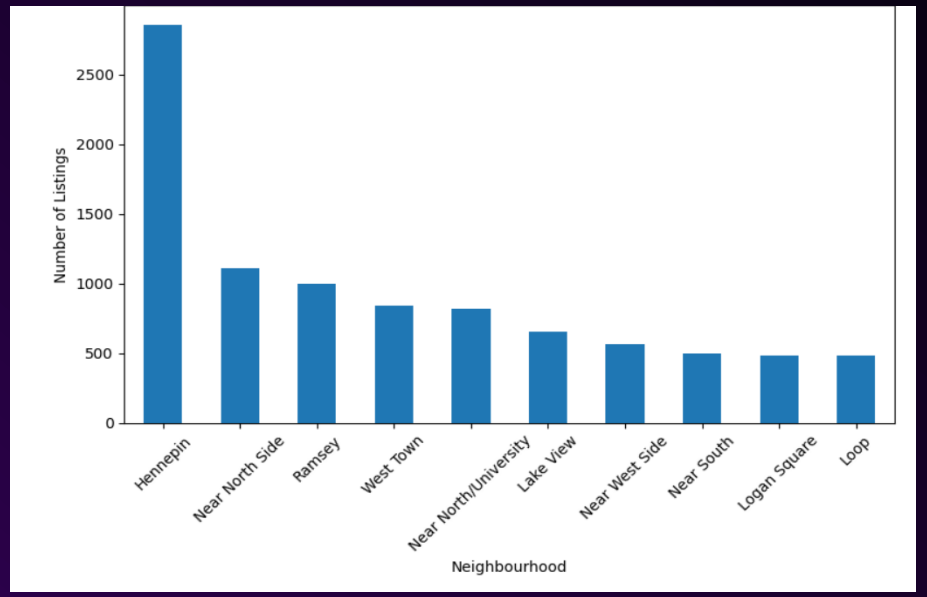






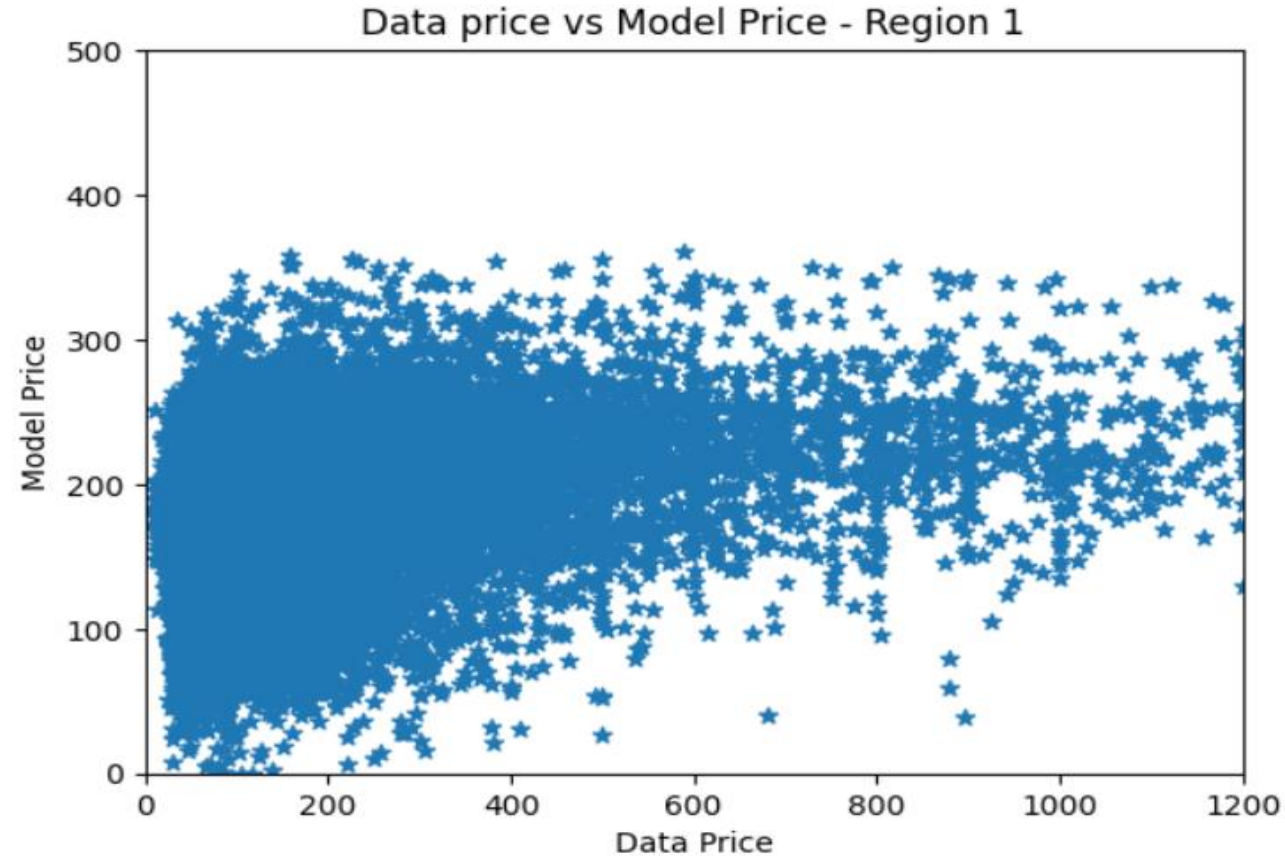

Number of Listings
1750
1500
1250
1000
750
500
250
0
Long Beach
Hollywood
Venice
Beverly Hills
Santa Monica
Downtown/Civic Center
Downtown
West Hollywood
Glendale
Hollywood Hills
Neighbourhood


Number of Listings
2500
2000
1500
1000
500
0
Bedford-Stuyvesant
Williamsburg
Midtown
Harlem
Bushwick
Hell's Kitchen
Upper East Side
Upper West Side
Crown Heights
East Village
Neighbourhood


Number of Listings
2500
2000
1500
1000
500
0
Hennepin
Near North Side
Ramsey
West Town
Near North/University
Lake View
Near West Side
Near South
Logan Square
Loop
Neighbourhood

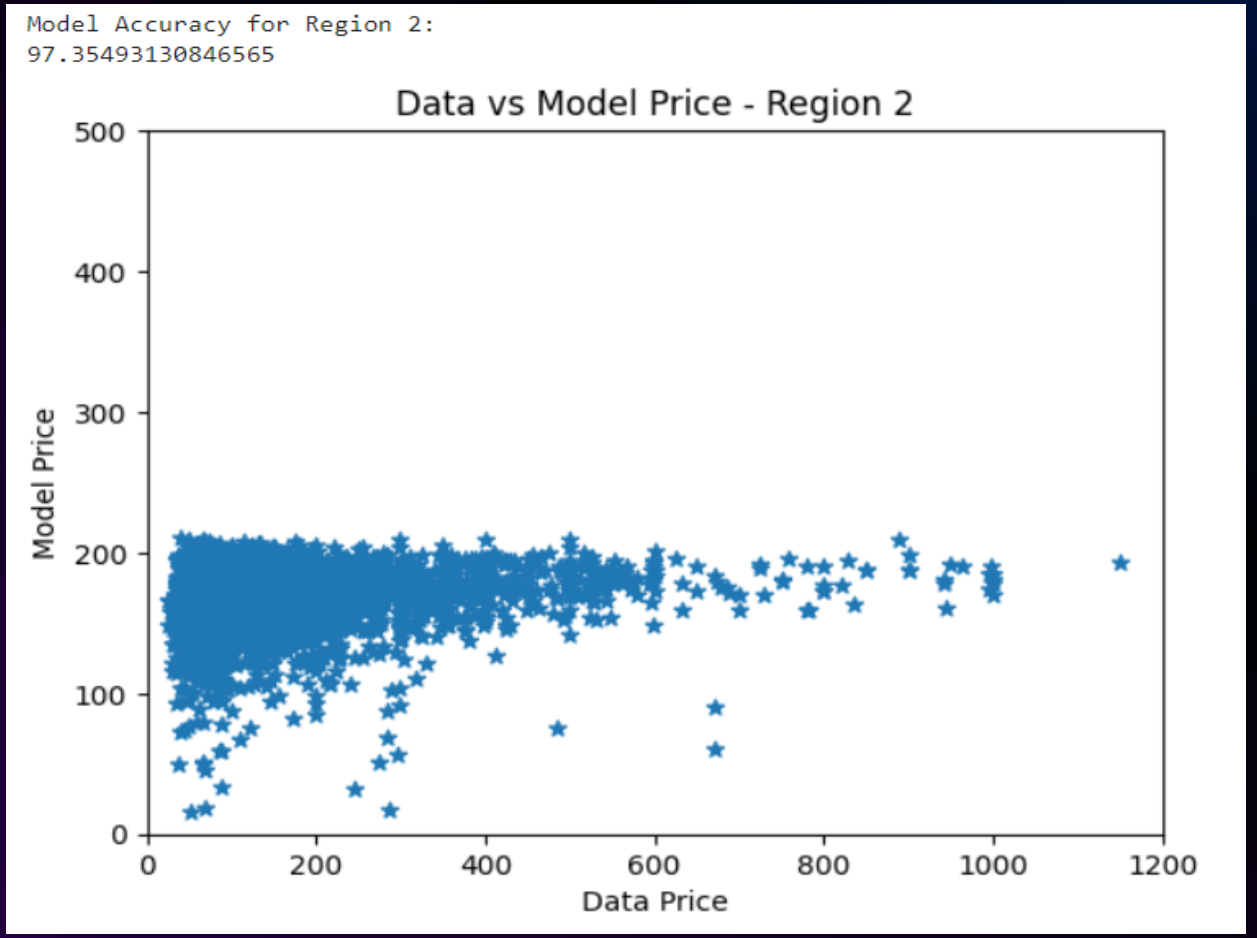
PREDICTION

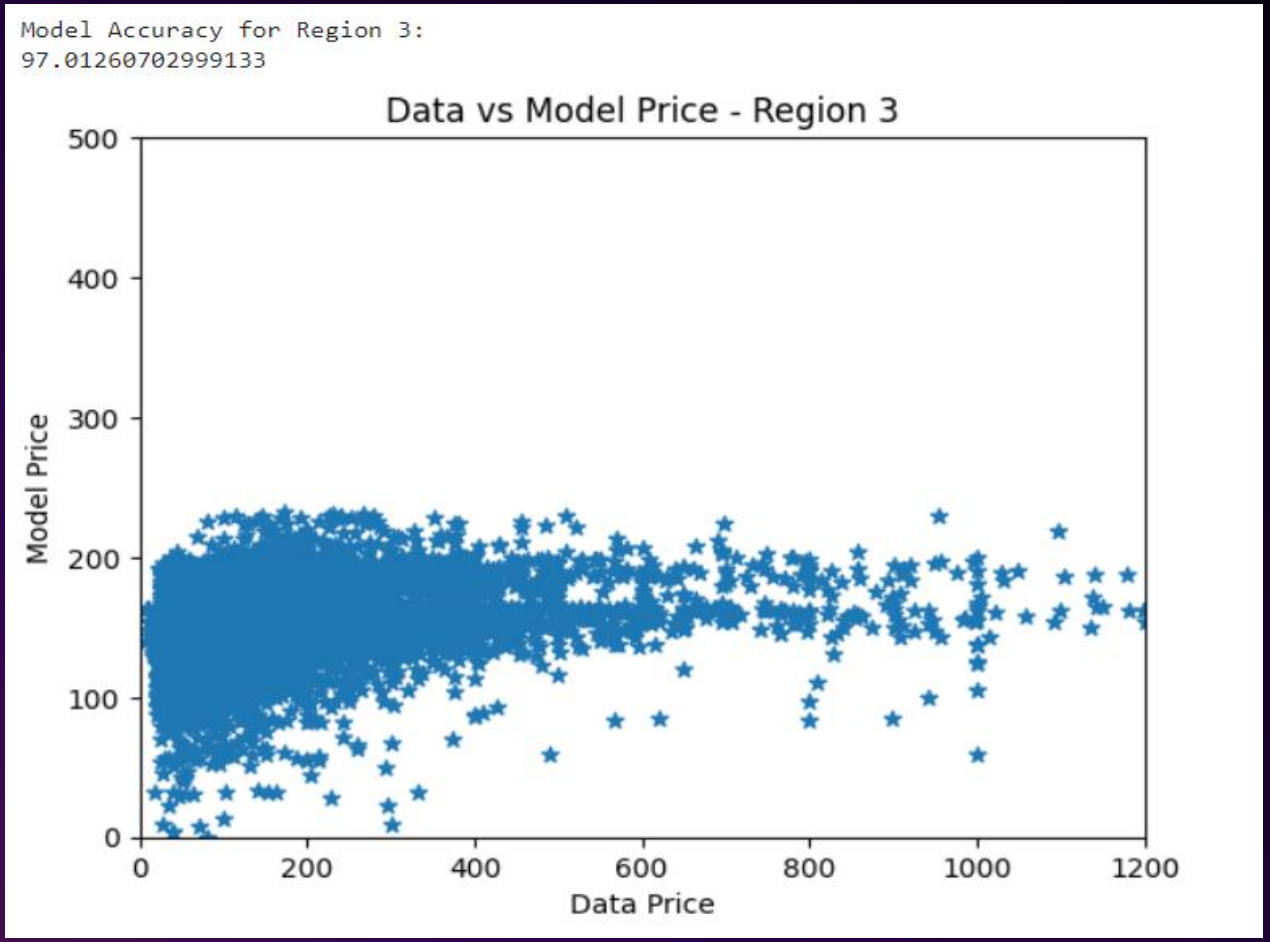
Model Accuracy for Region 1:
93.00692048137142



- The figures and results present the comparison between the actual prices (data price) and the predicted prices (model price) for each region, based on a linear regression model using three chosen factors: rating, latitude, and longitude.
- Scatter plot shows a good alignment between actual and predicted prices with an R^2 score of 93.01%, indicating the model explains 93% of price variability.

Scatter plot reveals a tight clustering around the line of equality with an R^2 score of 97.35%, demonstrating a high level of model accuracy.





- Scatter plot shows strong alignment between actual and predicted prices with an R^2 score of 97.01%, reflecting high model accuracy.
- High accuracy across regions suggests rating, latitude, and longitude are strong predictors

DISCUSSION

Implications of findings:

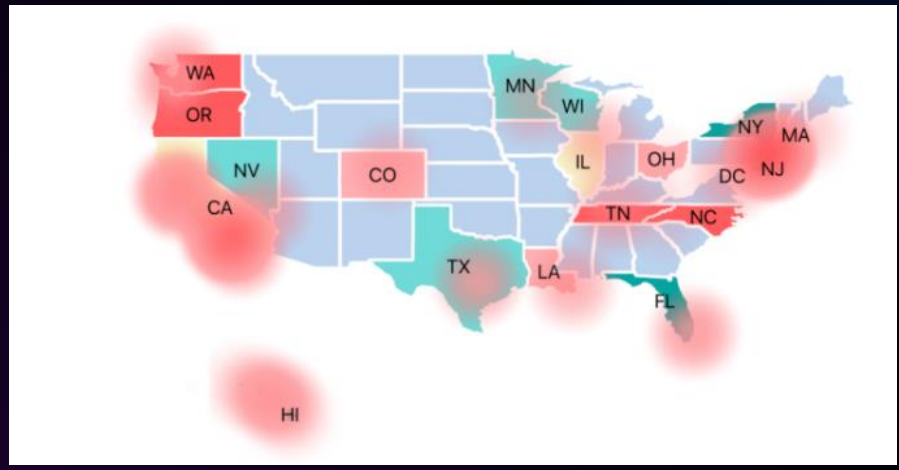
- **Location Significance:** Location is a major factor influencing accommodation choices, with some areas offering diverse price ranges and others focusing on either high-end or budget options.
- **Price Sensitivity:** Travelers often prioritize affordability, leading to a preference for lower-cost listings regardless of rating or location.
- **Accommodation Type Trends:** Entire homes or apartments are more popular among travelers than shared or private rooms, guiding developers and investors in market alignment.

CONCLUSION

Market Variability: LA shows the highest price variability and upper quartile prices, reflecting a diverse and potentially inflated market compared to Seattle and San Francisco.

Longitude Impact: Significant positive correlation between property prices and longitude, indicating geographical location's crucial role in real estate valuation.

Regional Insights: Overall, Seattle and San Francisco exhibit lower price variability and a less inflated market compared to LA.



FUTURE WORK/LIMITATIONS

- **Incorporate Additional Variables:** Include factors like local economic conditions, infrastructure, and demographics for a more comprehensive analysis.
- **Expand Geographic Scope:** Analyze more cities and regions to enhance generalizability and uncover new patterns.
- **Temporal Analysis:** Track property price changes over time to capture recent trends and market dynamics.
- **Micro-Level Study:** Investigate neighborhood or district-level variations to identify localized factors affecting prices.
- **International Comparison:** Compare with global real estate markets to identify unique patterns or commonalities in property valuation.

REFERENCES

- Inside Airbnb: https://insideairbnb.com/
- NY Rental Properties Pricing on Kaggle: https://www.kaggle.com/datasets/ ivanchvez/ny-rental-properties-pricing

THANK YOU

Github Link: https://github.com/viv-dan/ldmp_rental_prediction_project

Esha Acharya

– acharya.e@northeastern.edu

Vivek Dantu

– dantu.vi@northeastern.edu