ITERATION 3

RENTAL PREDICTION

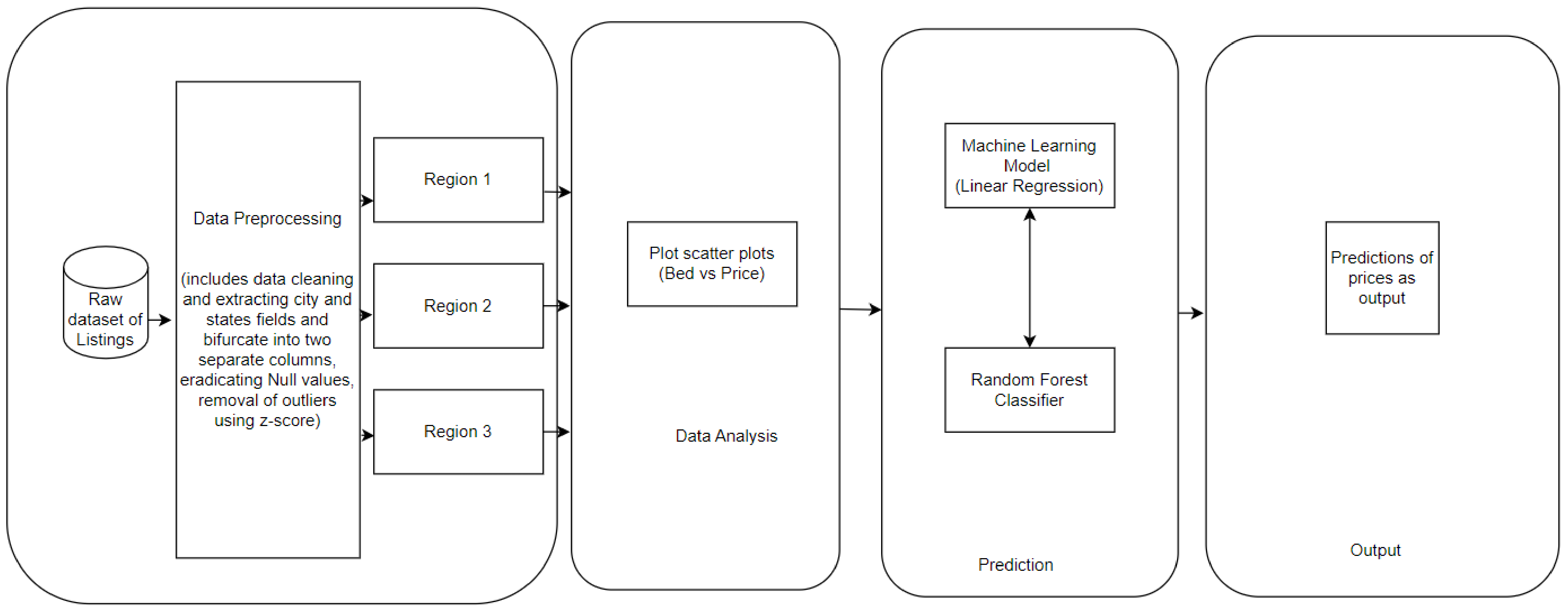
AGENDA

- Aim of our Project
- Design Diagram
- Removing outliers
- Data Analysis
- Data Visualisation
- Future Aim – Prediction using ML models

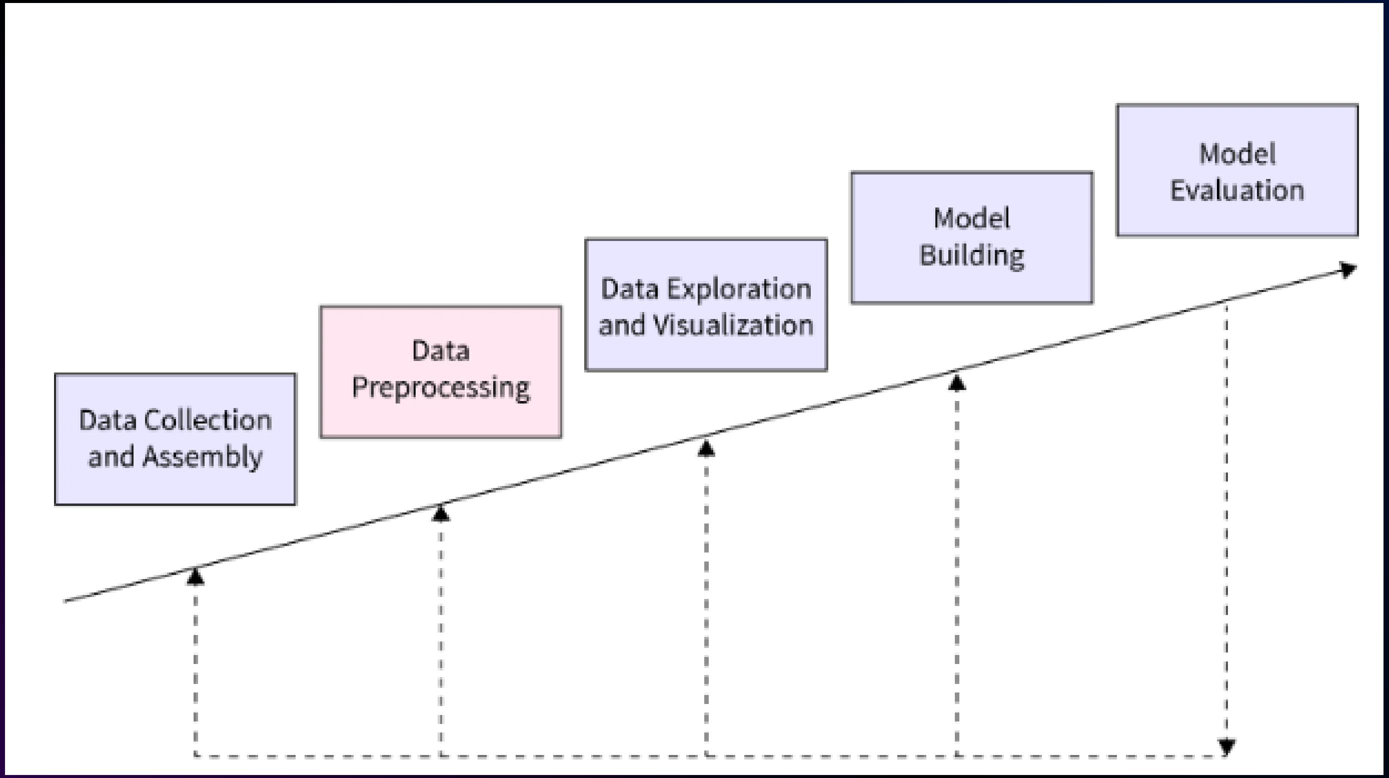
AIM

- 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.

DESIGN DIAGRAM FOR RENTAL PREDICTION

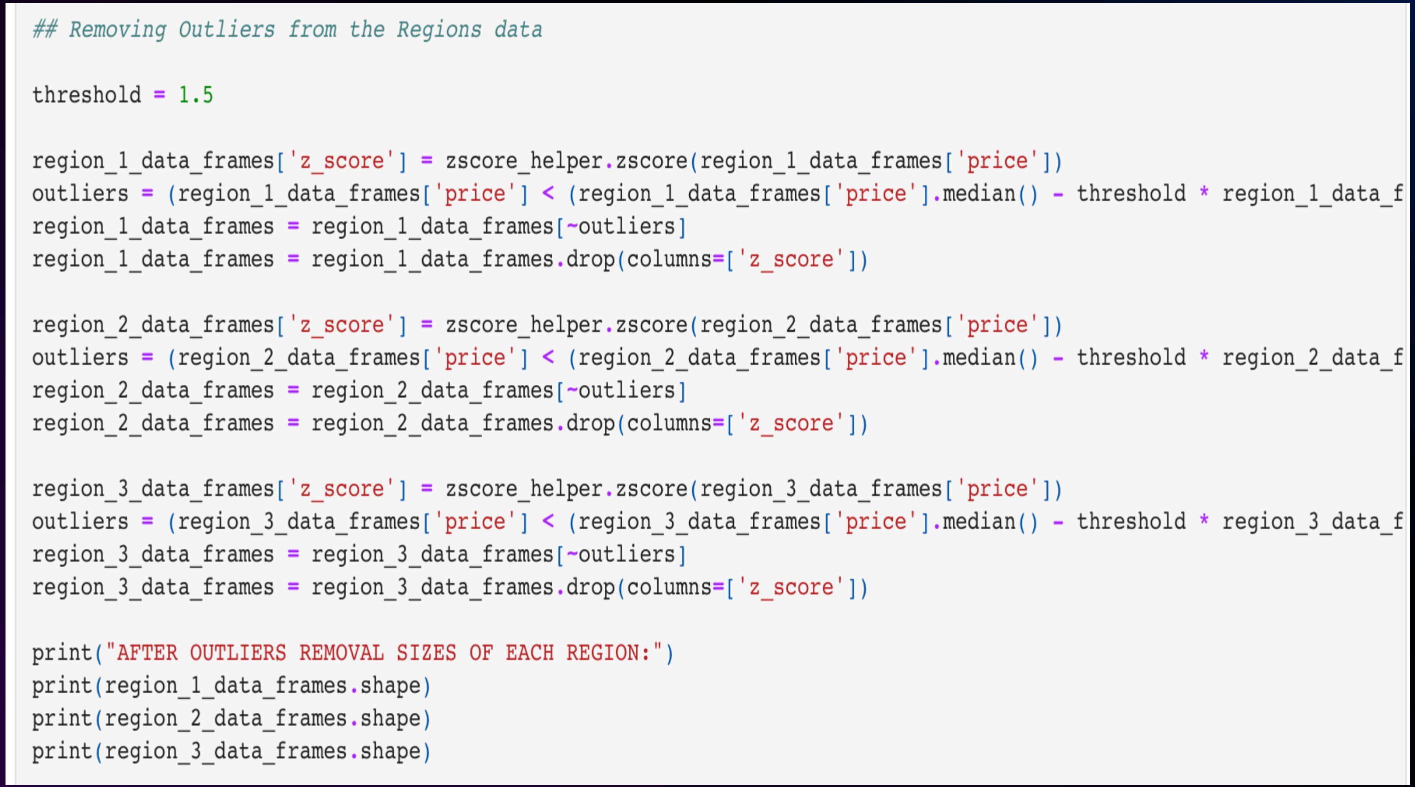

Raw dataset of Listings
Data Preprocessing
(includes data cleaning and extracting city and states fields and bifurcate into two separate columns, eradicating Null values, removal of outliers using z-score)
Region 1
Region 2
Region 3
Plot scatter plots (Bed vs Price)
Data Analysis
Machine Learning Model (Linear Regression)
Random Forest Classifier
Prediction
Predictions of prices as output
Output

BUILDING BLOCKS


Data Collection and Assembly
Data Preprocessing
Data Exploration and Visualization
Model Building
Model Evaluation

REMOVING OUTLIERS

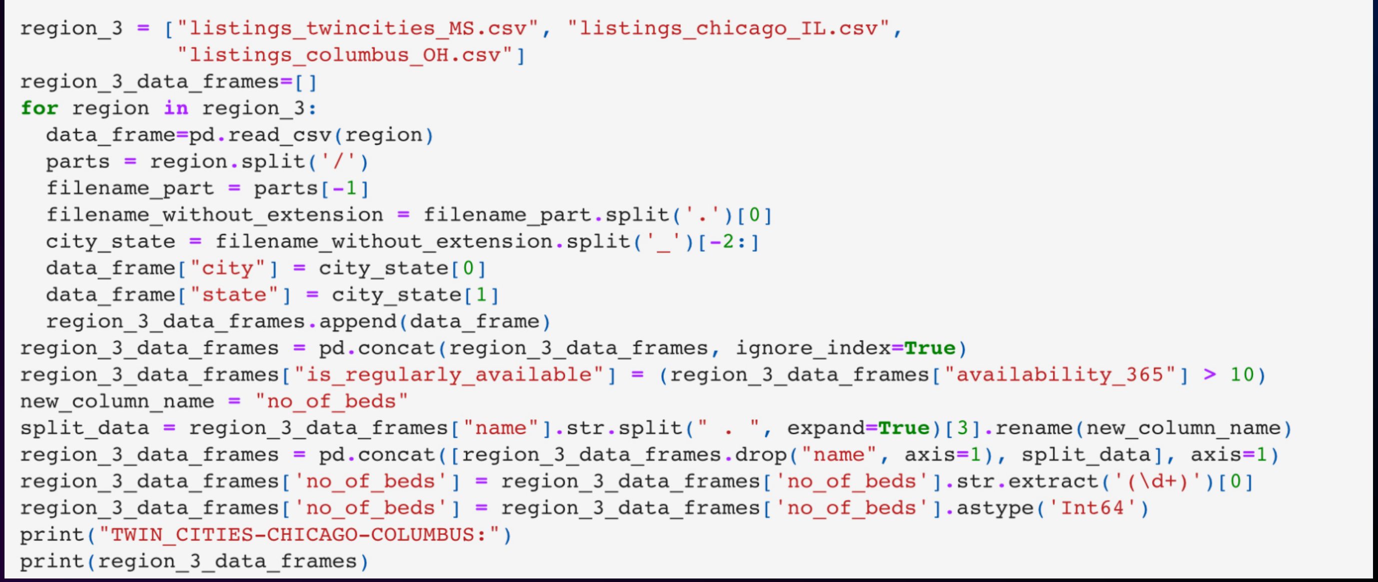
1. We removed approximately 1000 rows from our csv files which consisted of outliers. A threshold of value 1.5 was set to detect the outliers in prices.
2. Standard deviation and median were implemented to calculate the outliers.



```
## 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)
```

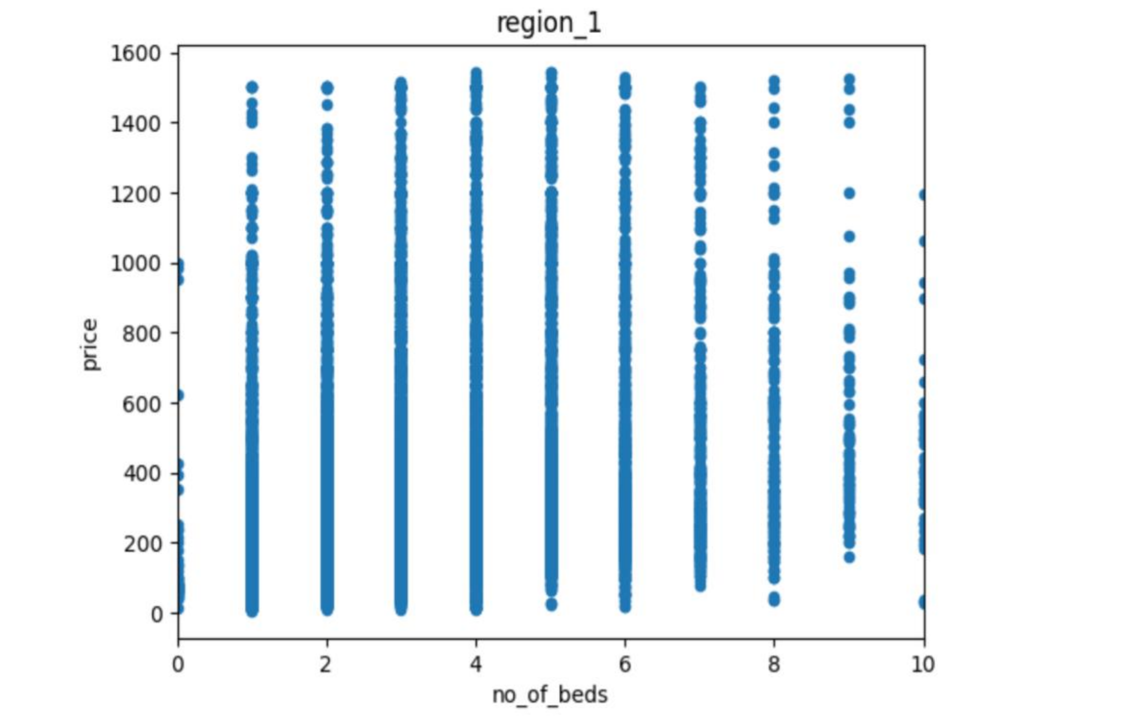
DATA ANALYSIS

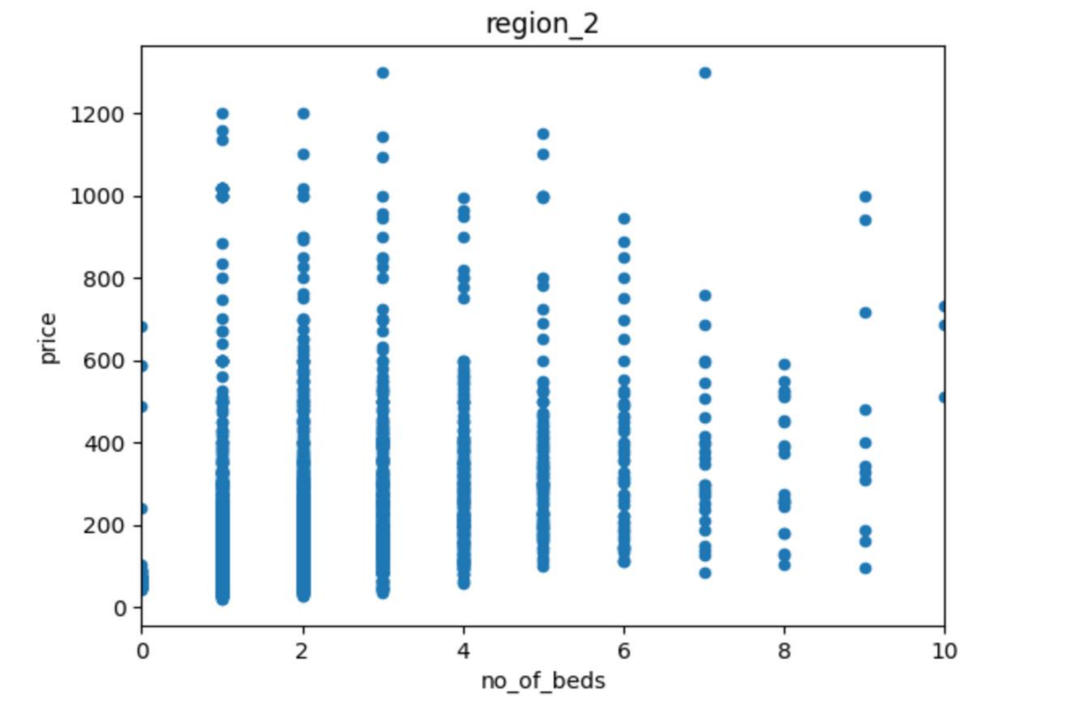
1. After cleaning the data to include the necessary columns and rows, we divided our dataset into three main regions: California, Boston, and Minnesota.
2. From the listings tables, we extracted the number of beds and utilized the number of beds to analyse the rental property prices based on the number of bedrooms as shown in the visualization graphs ahead.



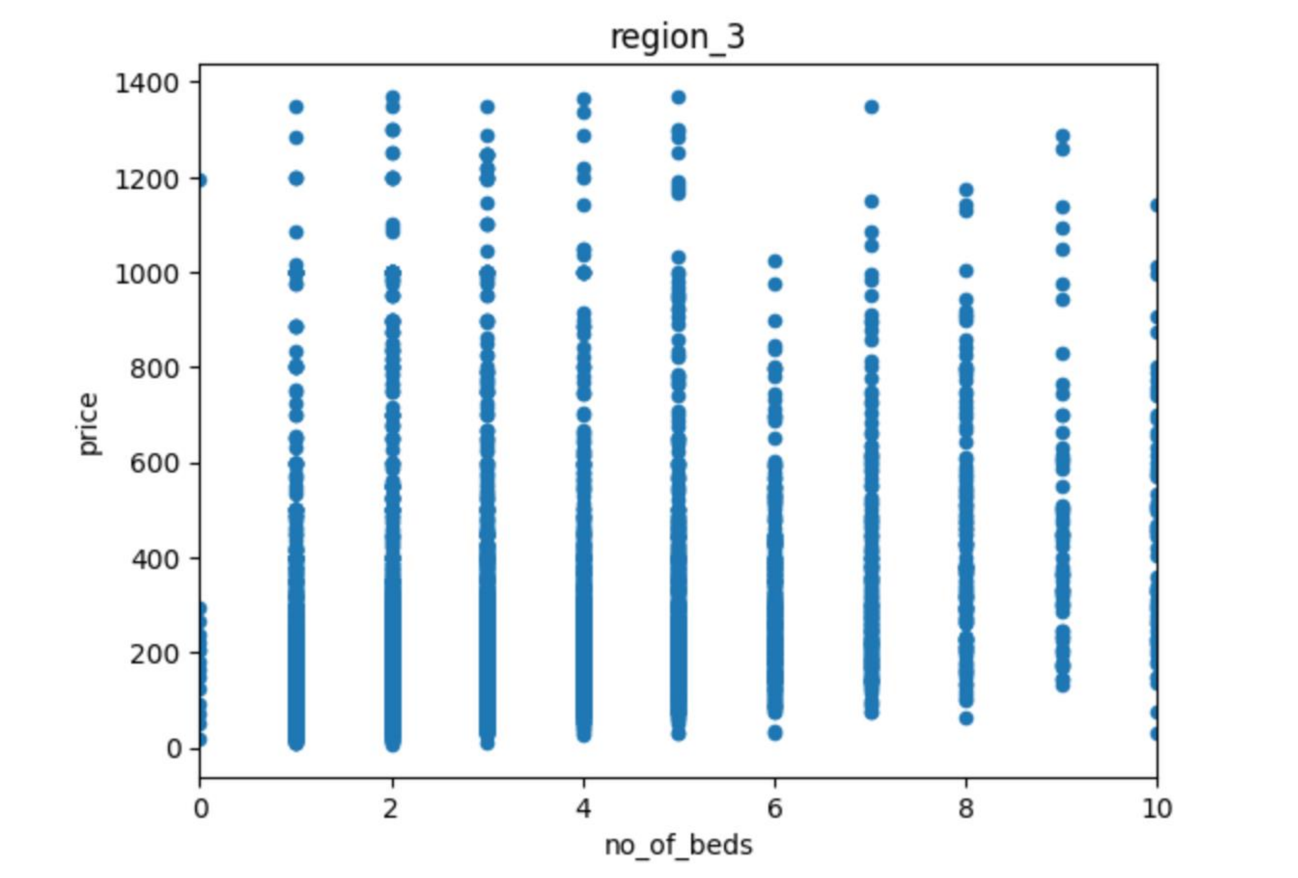
```
region_3 = ["listings_twincities_MS.csv", "listings_chicago_IL.csv",
            "listings_columbus_OH.csv"]
region_3_data_frames=[]
for region in region_3:
  data_frame=pd.read_csv(region)
  parts = region.split('/')
  filename_part = parts[-1]
  filename_without_extension = filename_part.split('.')[0]
  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)
```

DATA VISUALISATION





DATA VISUALISATION



FUTURE AIM

- Implement a machine learning model for price prediction using linear regression, and utilize a random forest as a classifier.
- Enhance data visualization techniques, with a focus on feature contributions to price variations.
- Improve accuracy and error handling to prevent fraudulent behavior.

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