Project Report

DSCI-6607-002
December 11, 2024
Abdallah Chidjou

California Housing Price Data Analysis

The California Housing Prices dataset was analyzed to uncover insights into housing trends and factors affecting house prices in California. This report documents the steps taken in the data preparation, analysis, and visualization phases.

Objectives

The primary goals of this project were:

- To clean and preprocess the data by addressing missing values and outliers.
- To generate descriptive statistics for a comprehensive understanding of the dataset.
- To explore relationships between features through feature engineering and visualizations.
- To identify key factors influencing median house values.

Data Loading and Preparation

Data Overview
Dataset: California Housing Prices
Features:
- Numeric Columns: Longitude, Latitude, Housing Median Age, Total Rooms, Total Bedrooms, Population, Households, Median Income, Median House Value.
- Categorical Column: Ocean Proximity.

Missing Values: The `total_bedrooms` column had missing values, which were replaced with the median value to preserve data integrity.
Outlier Handling: Outliers were identified and capped using the Interquartile Range (IQR) method to reduce their impact on the analysis.

Descriptive Statistics
A summary of the numeric features was generated to understand central tendencies, spread, and ranges of the data. These statistics provided a foundation for further analysis.

Data Exploration and Visualization

Boxplots: Used to identify and visualize outliers in numeric columns both before and after treatment.
- Scatter Plot: Highlighted the strong positive correlation between median income and median house value.
- Histogram: Visualized the distribution of median house values, showing a cap at $500,000.

- Correlation Heatmap: Illustrated the relationships between numeric features, identifying median income as the strongest predictor of house value.
- Boxplot by Ocean Proximity: Compared median house values across different proximity categories, with coastal properties generally commanding higher prices.

Median Income: The strongest positive correlation with median house value.
Ocean Proximity: Coastal homes ("Near Bay" and "Near Ocean") have significantly higher median values compared to inland properties.
Feature Insights:
- Higher rooms per household positively influenced house value.
- Higher bedrooms per room ratio negatively correlated with house value, indicating inefficiencies in property layouts.
Outliers: Managing outliers ensured robust statistical analysis without undue skewing of results.

I was able to successfully demonstrated the importance of data cleaning, feature engineering, and visualization in analyzing the housing datasets. My findings emphasize the significance of location, income levels, and housing features in determining property values.