Cost of Living Analysis

Ankit Mistry, Saish Malluri, and Shivam Mistry

I. INTRODUCTION

The cost of living is a critical economic metric that significantly impacts people's lives. It encompasses various expenses like housing, food, and transportation, which vary greatly across regions. In this paper, we use Kaggle data to create a predictive model for understanding and quantifying these variations among cities in the USA. By employing data science and machine learning techniques, our goal is to provide valuable insights for decision-makers and enhance our understanding of the factors driving cost of living differences in the country. This project's objective is to utilize Kaggle's dataset to construct a predictive model for estimating the cost of living in diverse U.S. cities. Employing advanced data analysis and machine learning techniques, our goal is to understand the factors influencing cost of living disparities across the country.

II. MOTIVATION

The motivation behind this study lies in the practical importance of the cost of living in people's daily lives and its broader implications for economic decision-making. Individuals and families regularly grapple with the challenges of budgeting and managing expenses, while businesses and policymakers require accurate insights into regional cost disparities to make informed choices. By developing a predictive model to estimate the cost of living across U.S. cities, we aim to empower individuals with valuable tools for financial planning, assist businesses in resource allocation, and aid policymakers in crafting effective policies. Furthermore, understanding the key factors influencing cost of living variations can contribute to efforts aimed at improving overall living standards and reducing economic disparities within the United States. This study is motivated by the desire to address real-world challenges and contribute to the well-being of residents across diverse American communities.

III. DISCUSSION OF DATA

The data for this project will be sourced primarily from Kaggle, a renowned platform for sharing and accessing datasets. Kaggle offers a diverse repository of datasets relevant to a wide range of domains, making it an ideal source for our analysis of the cost of living in U.S. cities.

A. Data Source and Attributes

Our project relies on data sourced from Kaggle, a reputable platform for datasets. We will gather a wealth of data attributes, including city information (names, states, coordinates), economic indicators (income, rent, housing prices, unemployment), Consumer Price Index (CPI), demographics (population, age, education), amenities and services data, and historical trends. These factors collectively form the basis for understanding and predicting the cost of living variations across U.S. cities.

B. Data Preparation and Model Development

The collected data will undergo cleaning and preprocessing to ensure its quality and consistency. We'll integrate data from various sources and perform exploratory data analysis (EDA) to unveil insights. Feature engineering will help us refine variables for better cost of living prediction. Subsequently, we'll employ a variety of machine learning algorithms, validate models, and evaluate their performance using metrics like MAE, RMSE, and R2. Our goal is to create an accurate predictive model that aids decision-making and sheds light on cost of living dynamics in American cities.

IV. RESPONSIBILITIES OF MEMBERS

In our collaborative effort to successfully predict the cost of living in various U.S. cities using Kaggle data, each member of our team will contribute their unique skills and expertise to accomplish specific tasks. Our division of responsibilities is as follows:

A. Ankit Mistry

Ankit will take charge of data acquisition, focusing on sourcing and downloading the Kaggle dataset. He will also perform initial data cleaning and preprocessing, ensuring that the dataset is ready for further analysis.

B. Shivam Mistry

Shivam will lead the exploratory data analysis (EDA) phase. This includes conducting statistical analysis, data visualization, and correlation assessments to gain insights into the dataset. Additionally, he will be responsible for feature engineering, identifying and creating new variables that enhance our predictive model's accuracy.

C. Saish Malluri

Saish will be responsible for model development and evaluation. This includes selecting and implementing machine learning algorithms, training predictive models, and testing their performance. Cross-validation and metric evaluation (e.g., MAE, RMSE, R2) will be conducted to ensure the model's reliability.

V. TIMELINE OF MILESTONES

A. Week 1: Project Initiation and Data Collection

- Data collection from Kaggle, initial data assessment.
- Project kick-off meeting to discuss roles, responsibilities, and objectives.
- Complete data collection and preliminary data cleaning.

B. Week 2: Data Preprocessing and Exploration

- In-depth data cleaning and handling missing values.
- Begin exploratory data analysis (EDA) to uncover insights.
- EDA continued, identifying key features.

C. Week 3: Feature Engineering and Model Development

- Start feature engineering to enhance data for modeling.
- Model selection and implementation.
- Initial model development and testing.

D. Week 4: Model Evaluation and Refinement

- Cross-validation and performance evaluation.
- Iterative model refinement based on evaluation results.

E. Week 5: Paper Preparation and Final Model

- Draft project paper, including introduction, methods, and results sections.
- Review and edit project paper, ensure consistency and clarity.
- Finalize the predictive model.

F. Week 6: Finalization, Presentation, and Conclusion

- Interpret model results and gather insights.
- Prepare visuals and charts for the paper.
- Complete and revise the project paper.
- Prepare presentation materials and practice for project presentation.
- Deliver the project presentation and submit the final paper.

VI. EXPECTED OUTCOME

The principal outcome of this project is the creation of a precise predictive model capable of accurately estimating the cost of living in diverse U.S. cities. This model, driven by comprehensive economic, demographic, and social data, is anticipated to uncover key factors influencing cost of living disparities. Furthermore, our geospatial analysis will visually depict cost of living variations across the country, contributing to a deeper understanding of the geographical distribution of affordability within the United States.