Cost of Living Analysis

Ankit Mistry, Saish Malluri, and Shivam Mistry

I. OBJECTIVE

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 perform in-depth data analysis on Kaggle data in order to understand and quantify these variations among cities in the USA. 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. We employed advanced data analysis to understand the factors influencing cost of living disparities across the country.

The main goal for this project is to analyze the distribution of living expenses by utilizing Python and prominent data analysis libraries for visualization. Our goal is to gain insights into the interrelationships between key factors and the cost of living. We aim to address specific research questions, such as the influence of geographical areas on various features and the overall cost of living. Additionally, we will investigate how family size impacts different aspects and the resulting cost of living. Through our analysis, we intend to identify the most and least affordable counties in both the United States and Tennessee, providing valuable information for individuals and policymakers alike.

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 using data analysis 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 dataset from Kaggle incorporated 31430 entries of community-specific estimates for ten family types, including one or two adults with zero to four children, in all 3143 counties and metro areas across the United States.

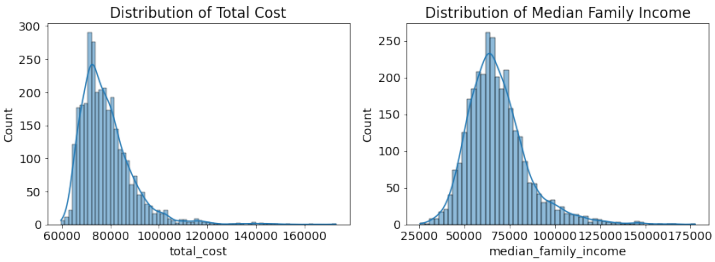

Fig. 1. This figure shows the distribution for total cost and median family income

The dataset composed of these features:

- State, County, Metropolitan area
- Family Size
- Food Cost
- Healthcare Cost
- Transportation Cost
- Childcare Cost
- Housing Cost
- Other Necessities Cost
- Taxes
- Median Family Income
- Total Cost

A. *Exploratory Data Analysis (EDA)*

In exploring the relationship between income distribution and the cost of living in our data analysis project, we observed a diverse range of income levels, spanning from $25,000 to $175,000. Notably, the most prevalent income bracket fell within the range of $50,000 to $80,000, indicating a concentration of individuals or households within this middle-income segment as illustrated in figure 1.

Correspondingly, the distribution of total cost of living exhibited a wide spectrum, ranging from $60,000 to well over $160,000. Strikingly, the most common total cost range was found to hover around $60,000 to $80,000, suggesting that a significant portion of the population incurs living expenses within this bracket. This insight into the interplay between income and cost of living sheds light on the economic dynamics within our dataset, highlighting the prevalence of a moderate income range alongside a corresponding concentration of typical living expenses. Further analysis will be crucial to uncover nuanced patterns and potential implications for individuals' financial well-being in relation to the cost of living.

The correlation matrix (Figure 2) in our data analysis project offers valuable insights into the relationships between total cost (representative of the cost of living) and various

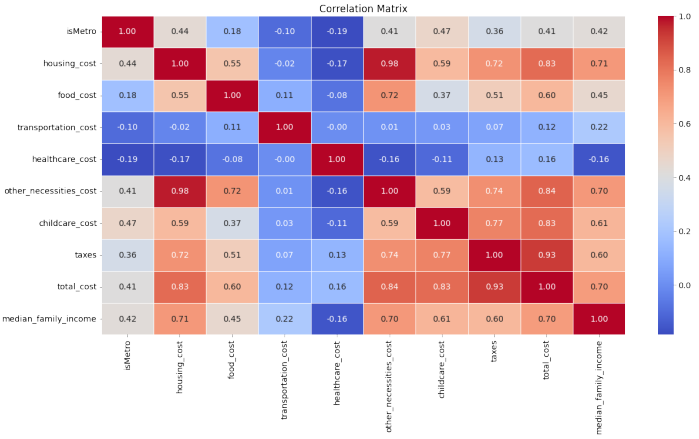

Fig. 2. This figure shows how the total cost (cost of living) is correlated with all the other factors

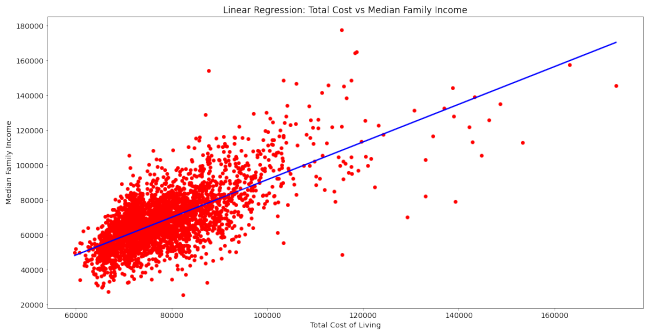

Fig. 3. regression analysis

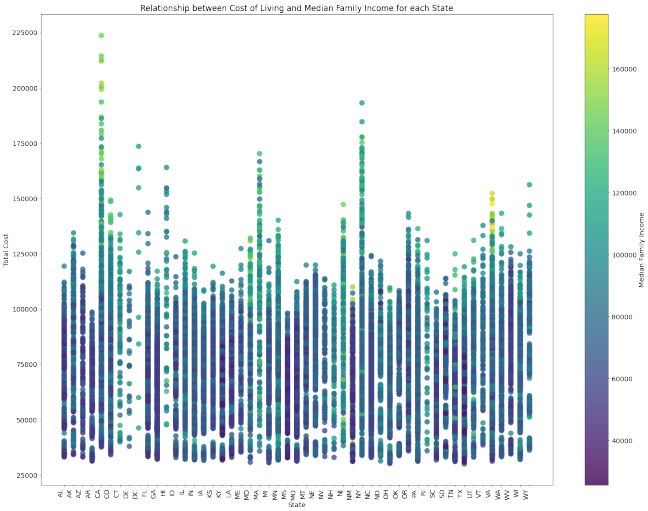

Fig. 4. scatter plot depicting cost of living and income

factors, providing a comprehensive view of the strength of associations between pairs of variables. A noteworthy positive correlation emerges between housing cost and taxes, collectively contributing to the total cost. This observation implies that regions characterized by elevated housing expenses and a higher tax burden tend to exhibit an overall higher cost of living. Conversely, the correlation analysis indicates a less pronounced relationship between transportation cost, healthcare cost, and the total cost. This finding suggests that increased transportation and healthcare expenditures may not significantly contribute to the determination of the overall cost of living, highlighting potential nuances in the factors influencing living expenses. Further exploration of these correlations will be crucial for a nuanced understanding of the complex interplay between specific cost components and the overall cost of living.

IV. Machine Learning Model: Regression Model

In our machine learning approach, we applied logistic regression to model the relationship between income and the cost of living. The analysis underscores a substantial positive correlation between the cost of living and median family income, visually represented by the upward-sloping regression line as represented by figure 3. This trend suggests that as median family income increases, there is a concurrent and notable escalation in the cost of living. The logistic regression provides a quantitative foundation for understanding the interdependence of these two variables, allowing us to make predictive assessments and glean insights into how changes in income levels may impact the cost of living.

V. Results

A. Income vs. Cost of Living

For income and cost of living, the graphical representation of individual family incomes with the corresponding cost of living across counties provides a comprehensive overview of regional economic dynamics. Notably, the analysis reveals distinct patterns across different states. States such as California, Massachusetts, New York, and Virginia emerge as areas characterized by a high cost of living, suggesting that residents in these regions face elevated living expenses in comparison to others. On the contrary, states like Arkansas, Mississippi, and Indiana exhibit a contrasting trend with a notably lower cost of living. These findings underscore the geographical variability in the economic landscape, shedding light on disparities in the affordability of living across different parts of the country. This insight is pivotal for policymakers, researchers, and the general public alike in understanding and addressing the diverse economic challenges faced by individuals and families in various regions.

B. Feature Visualizations: US Map

A comprehensive analysis of the cost of living across various regions reveals distinctive trends in key expenditure categories. Housing costs exhibit a consistent pattern, with coastal areas bearing significantly higher expenses, aligning with the demand for prime locations. Healthcare expenses, in contrast, exhibit a scattered landscape, lacking a discernible trend across regions. The northeastern states and California notably present elevated food costs, while childcare expenses generally follow a similar trend, save for a few exceptions in the northeastern US. Transportation expenses are notably higher in the northwestern US, contributing to regional disparities. Other necessities also exhibit varying costs, with

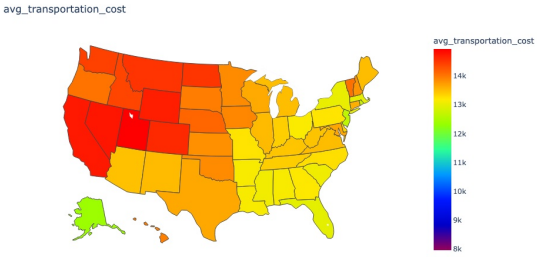

Fig. 5. Average Transportation Cost

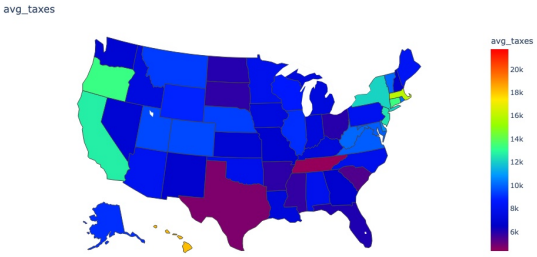

Fig. 6. Average Taxes

some eastern coastal states and California standing out for higher expenses. Average taxes weigh heavily on the west coast and select northeastern states, impacting the overall cost of living. Furthermore, median family incomes tend to be higher along the western coast and in specific northeastern states, influencing affordability dynamics across these regions. Understanding these regional nuances is critical for a comprehensive cost of living analysis, shedding light on the diverse economic landscapes that impact residents' financial realities.

C. Impact of Family Size

The impact of family size on income and cost of living is evident in the box and whisker plots, which effectively illustrate the distribution of these variables in both rural and metro areas. Generally, the data reveals that median income and the cost of living in metro areas tend to be consistently higher, and their respective distributions are comparable.

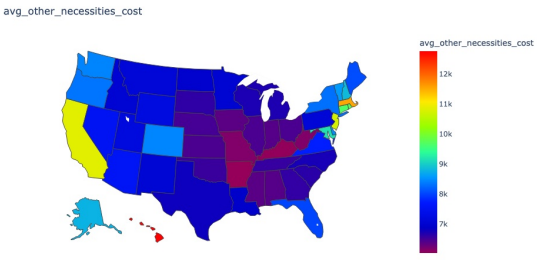

Fig. 7. Average Other Neccessites Cost

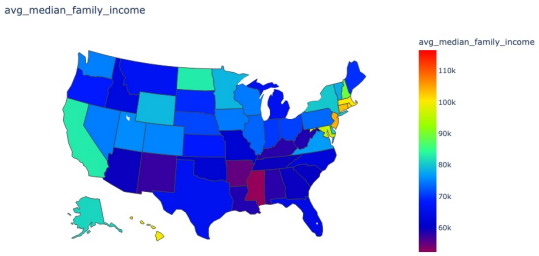

Fig. 8. Average Median Family Income

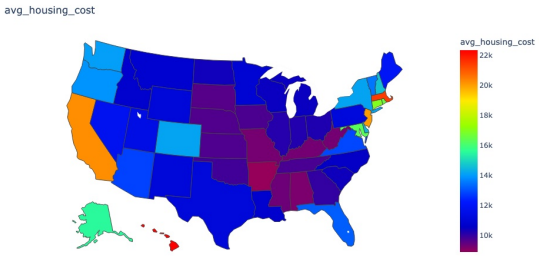

Fig. 9. Average Housing Cost

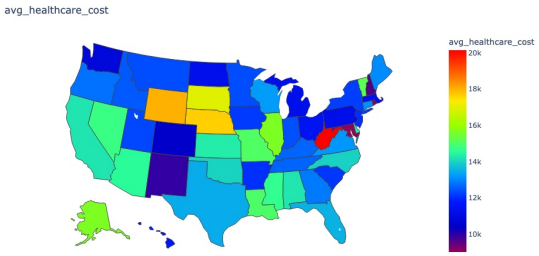

Fig. 10. Average Healthcare Cost

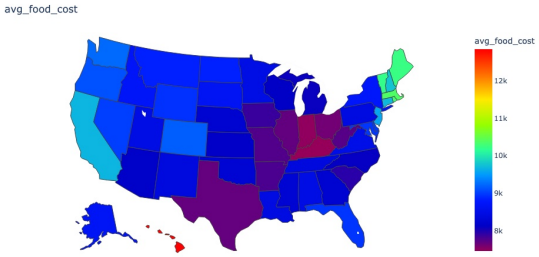

Fig. 11. Average Food Cost

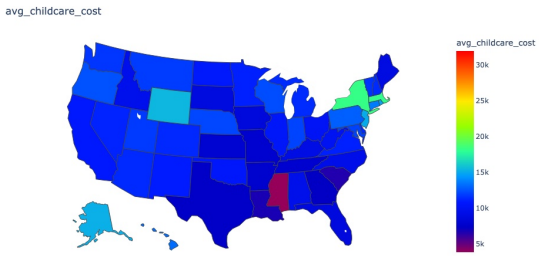

Fig. 12. Average Childcare Cost

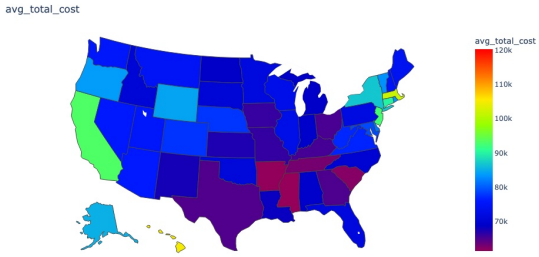

Fig. 13. Average Total Cost of Living

Intriguingly, in rural areas, the cost of living surpasses the median income, indicating a potential economic strain for families in these regions. Furthermore, the analysis highlights a noteworthy trend—across both rural and metro areas, all costs examined demonstrate an increase with the expansion of family size. This suggests that larger families face higher overall living expenses. On a nuanced note, the data indicates that housing costs and other essential expenditures tend to decrease as the number of children in a family decreases. These findings underscore the intricate relationship between family size, income, and the cost of living, offering valuable insights for policymakers and individuals alike in understanding and addressing the economic challenges associated with varying family structures.

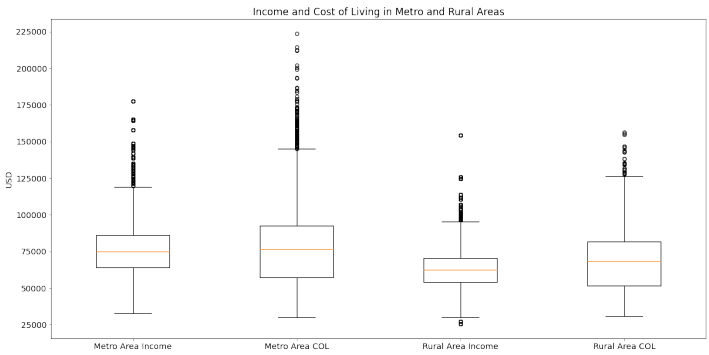

Fig. 14. Cost of Living: Metro and Rural

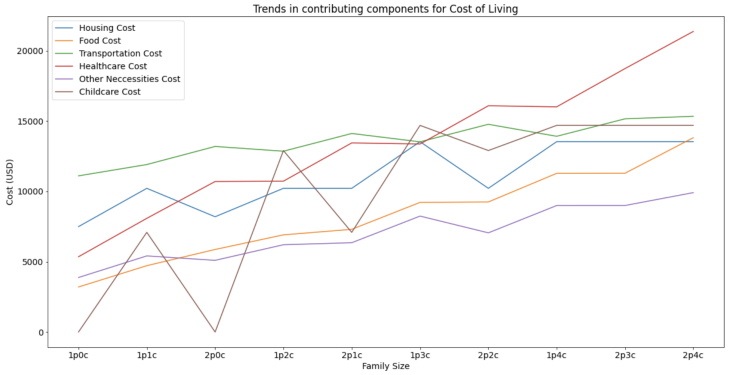

Fig. 15. Trends in Cost of Living

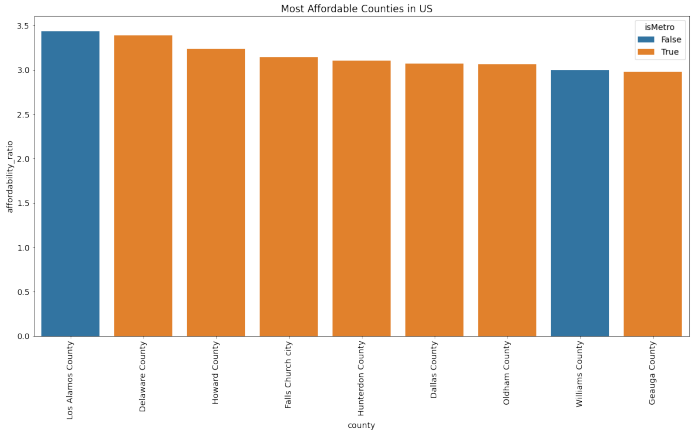

Fig. 16. Most affordable counties in US

D. *Affordability Analysis*

Our affordability analysis, calculated by dividing income by the cost of living for all counties in the U.S., offers valuable insights into regional economic dynamics. The bar chart depicting the most and least affordable counties unveils a compelling trend. The most affordable counties predominantly cluster around major metropolitan areas, with notable examples including Delaware County outside Philadelphia and Falls Church City outside Washington, D.C. These areas, often suburban to major cities, demonstrate a favorable balance between income and living expenses. Conversely, the least affordable counties are primarily located in rural areas, indicating a potential economic strain for residents in these regions. Moreover, our focused analysis on Tennessee corroborates this trend, with Williamson County outside Nashville emerging as the most affordable, followed by Knox County. This consistent pattern underscores the significance of geographical context in understanding affordability disparities and provides crucial information for policymakers and individuals seeking to navigate economic challenges across different locales.

VI. Challenges and Future Work

The challenges encountered during our analysis have provided valuable insights into areas for improvement and

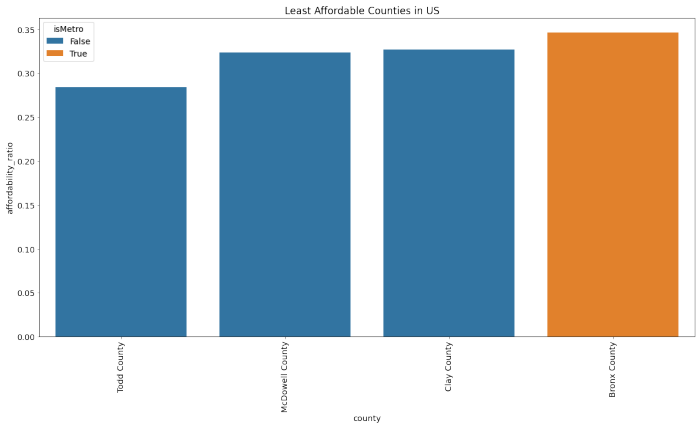

Fig. 17. Least affordable counties in US

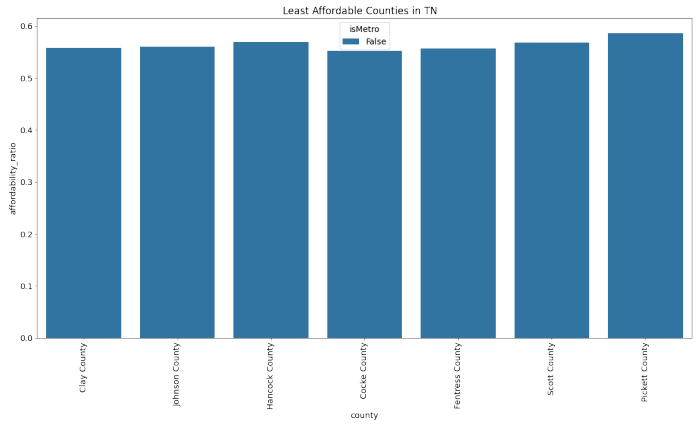

Fig. 18. Least affordable counties in TN

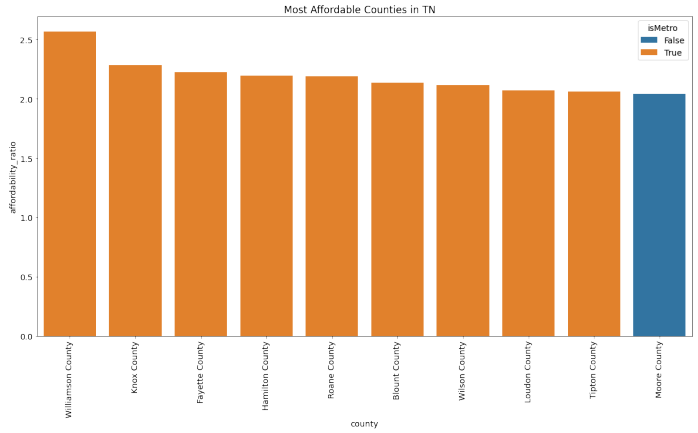

Fig. 19. Most affordable counties in TN

avenues for future research. One notable challenge was the absence of a comprehensive Cost of Living Index in the dataset, necessitating the use of raw costs in USD. This limitation posed difficulties in comparing and visualizing the data effectively. Additionally, the process of installing Plotly and generating map visualizations presented its own set of challenges. To address these issues and enhance the robustness of our analysis, future work could involve calculating a Cost of Living Index, providing a more standardized metric for comparison purposes. Furthermore, implementing machine learning models to predict the cost of living based on various factors could offer predictive capabilities and a deeper understanding of the underlying relationships. Beyond that, extending our analysis to different cultural contexts could shed light on how cost of living factors vary globally, providing a more comprehensive and nuanced perspective. These avenues for future work aim to refine our methodology, broaden the scope of our analysis, and contribute to the ongoing discourse on the complexities of cost of living dynamics.

VII. 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.

VIII. 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.