**Term Project**

**BSAN 885 – Business Analytics Capstone**

**Spring 2024**

**(Class Meeting Day and Time: Monday & Wednesday, 4:00 PM)**

**Project Title:**

**Exploring the Travel Expenditure Patterns**

**Assignment Title:**

**Final Project Report**

**Yogananda Theeguru (Q584C295)**

**Table of contents**

**1. Introduction and Background**  2

1.1 Topic and Importance 2

1.2 Background Research 3

1.3 Addressing the Issue: Problem Statement and Research Questions 4

**2. Methodology** 5

2.1 Data Used 5

2.2 Data Processing 6

2.3 Data Distribution and Visualization 8

2.4 Modeling 10

**3. Results**  10

3.1 Model Validity and Significance 10

3.2 Significance of Independent Variables and Relationships 11

3.3 Visualization & Interpretation 12

3.4 Data Storytelling 14

**4. Discussion**  15

4.1 Limitations of the Project 15

4.2 Recommendations for Future Research 16

4.3 Use Cases and Implications, Deployment Recommendation 17

**5. References**  19

**1.Introduction and Background**

**1.1 Topic and Importance**

The travel industry is a pivotal component of the global economy, offering substantial investment opportunities, particularly attractive to venture capitalists. In 2022, direct travel spending in the United States reached $1.2 trillion, underscoring the sector's crucial economic role (Gillian, 2023). This robust economic input, coupled with its capacity for innovation and resilience, positions the travel sector as a prime investment opportunity. Technological advancements such as AI and machine learning are revolutionizing customer experiences and operational efficiencies, demonstrating the sector's potential to adapt and thrive. Forecasts predict a 3.47% annual growth in the global Travel & Tourism market, aiming to reach a market volume of approximately $1063 billion by 2028 (WTTC, 2023). The resurgence in tourism investment is highlighted by a 23% increase in foreign investment projects in 2022, reflecting a strong recovery trajectory (Martijn, 2024). International visitor numbers, recovering to 80% of pre-pandemic levels in early 2023, indicate not only the industry's comeback but also its significant role in job creation and regional economic development (Martijn, 2024; Gillian, 2023). Additionally, the sector enhances global connectivity and cultural exchange, further emphasizing its global importance.

**1.2 Background Research**

The travel industry has demonstrated remarkable resilience and adaptability, responding dynamically to global challenges, notably the COVID-19 pandemic. Findings from Deloitte (2024) reveal a resurgence in travel spending as restrictions ease, with both domestic and international travel experiencing growth. This rebound is crucial as the sector recovers from the pandemic's profound impacts. Furthermore, the integration of digital technologies has reshaped consumer behavior significantly, with platforms like Airbnb transforming the lodging sector by popularizing the sharing economy. This innovation has compelled traditional hotel chains to diversify their offerings, launching home-sharing options to remain competitive (Hall et al. 2022; Deloitte 2024).

Strategic investments and governmental initiatives have played pivotal roles in driving the sector’s growth. For example, as Takahara reports, Japan's "Go To Travel" campaign offers subsidies to encourage domestic tourism, aiding economic recovery. This program is instrumental in revitalizing the local economy by boosting domestic travel (Takahara, 2023). Technology, as McKinsey & Company explains, also plays an important role in how we think about travel. This combination of social media and technology shows how travel is evolving, with new ideas and stories shaping where we go and what we do on our trips (McKinsey 2022). Augmented reality (AR) technology is significantly transforming the hospitality industry by enabling guests to virtually explore accommodations and amenities before booking. This immersive experience not only enhances customer engagement but also improves operational efficiencies, leading to higher satisfaction and increased revenue for service providers (Share et al. 2022). Additionally, the Tourism Investment Report 2023 highlights a 23% increase in foreign investment projects, with companies like JetBlue Ventures and Thayer Ventures leading the way in sustainable travel innovations (Baig, 2023).

The projected growth of the global travel and tourism market, estimated at an annual increase of 3.47%, is expected to reach approximately $1063 billion by 2028, underscoring the industry’s expansive economic role (WTTC, 2023). This growth is fueled by consumer demand for unique and personalized travel experiences. Moreover, academic research continues to shed light on the evolving dynamics of the travel industry. According to the article 'Changing Trends of the Hospitality Industry,' there is an increasing influence of millennials and Gen Z on the hospitality industry, highlighting these demographics' preferences for technology and sustainability, which are reshaping market strategies (Ramgade and Kumar, 2021). The significant rebound in international visitor numbers, reaching 80% of pre-pandemic levels by early 2023, highlights the sector's capability to bounce back and its crucial role in global economic development (WTTC, 2023).

**1.3 Problem Statement and Research Questions**

To address the needs of venture capital investors who are seeking current insights into consumer spending patterns within the travel sector, this study employs a combination of data-driven analysis and sophisticated modeling techniques. Utilizing comprehensive datasets from consumer expenditure surveys, the research focuses on analyzing spending patterns influenced by a range of demographic and economic factors. Advanced statistical tools and machine learning models are applied to uncover trends and predict future spending behaviors. Interactive dashboards and detailed visualizations are created to present these insights in an accessible manner, enabling stakeholders to make well-informed decisions swiftly and effectively.

**Problem Statement**

Investors from a venture capital group seek current insights into consumer spending patterns within the travel sector, aiming to understand how various factors influence these expenditures to guide strategic investment decisions.

**Research Questions**

* How do demographic variables such as age, gender, and family size impact consumer spending within the travel sector?
* How do income, family composition, and the number of earners influence spending behavior on various travel-related activities?

**2. Methodology**

**2.1 Data Used**

The data utilized for analysis was obtained from the Division of Consumer Expenditure Surveys, provided in Excel format, and covering only the first quarter (Q1) of each year from 2019 to 2022. This dataset comprises 4,807 observations detailing consumer behavior and preferences within the travel industry during this specific timeframe. It includes a broad range of variables such as demographic information, financial details, and specific travel expenditures. This dataset facilitates a focused examination of early-year travel spending patterns, assisting in the identification of trends and potential investment opportunities by correlating demographic factors with travel expenditures during this quarter.

**2.2 Data Processing**

Several columns have missing values, with significant missing data in columns such as **RACE2**, **INC\_HRS2**, **HISP2**, **OCCUCOD1**, and **OCCUCOD2**, among others. The **STATE** column has 392 missing values, indicating a need for data imputation or removal strategies for these missing entries. Several duplicate columns have been identified within the dataset, including pairs like "FAM\_SIZE" and "FAM\_SIZE.1", "NO\_EARNR" and "NO\_EARNR.1", "FINATXEM" and "FINATXEM.1", "FINCBTAX" and "FINCBTAX.1", "TOTHRLOC" and "TOTHRLOC.1", "PERSLT18" and "PERSLT18.1", "AS\_COMP5" and "AS\_COMP5.1", and duplicate sequences of 'EDUCA' and 'OCCUCODE' columns referencing 'OCCUCODE' at index 13. The dataset contains a wide range of numerical columns, including 'AGE\_REF', 'FAM\_SIZE', 'APPARPQ', 'NO\_EARNR', 'FINATXEM', 'FINCBTAX', 'TOTEXPPQ', and many others, totaling 112 numerical columns.

To identify outliers, focus will be on key columns that are likely to have a significant impact on the analysis, such as 'FINATXEM', 'TOTEXPPQ', 'TOTXEST'. To address the missing values, primary focus will be on the 'INC\_HRS2' column, which initially had 3280 missing values. Utilizing the Mean of the existing values in the column to fill in the missing entries. This was done using the fillna() method, the Nan values were replaced with the mean value of the column. After this operation, no missing values remained in the 'INC\_HRS2' column.

Identified and removed duplicate columns based on pairs found in the dataset. Variables: 'STATE', 'FAM\_SIZE', 'NO\_EARNR', 'FINATXEM', 'TOTHRLOC', 'PERSLT18', 'AS\_COMP5', and others. Removal of identified duplicate columns using the drop () method. Cleaned up messy column names and removed unnecessary and duplicate columns based on the identified duplicates. Identified outliers in selected numeric columns and then treated them using a capping method. The outliers in specific numeric columns such as 'FINATXEM', 'TOTEXPPQ', 'TOTXEST', among others were focused.

Calculated the Interquartile Range (IQR) for each selected numeric column. Based on the IQR, outliers ware identified using upper and lower bounds. Based on the IQR, lower and upper bounds were determined to identify outliers beyond these bounds were adjusted by replacing them with the respective bound values. Iterated through each selected numeric column and applied the capping method individually.

**Feature Engineering**

The code assigns the identical data from df2 to a new data frame called df4. Preserving the original data while applying adjustments or alterations to a new data frame is a common approach in data analysis. The rename () function is used to alter column names to more comprehensible and descriptive names. As a result, the columns are easier to understand and analyze. 1 denotes a male and 0 denotes a female. The categorical variable "sex" is transformed into a binary format. To prepare categorical variables for machine learning algorithms that need numerical inputs, this is an essential preprocessing step. By dividing the 'family size' and 'earners' columns into separate groups according to specified borders, the variables are converted into category variables. By combining comparable data points, this lessens the impact of outliers and streamlines the analysis.

Using the Scikit-learn MinMaxScaler, min-max scaling is performed to the numerical column's income\_after\_txs, est\_tot\_txs, and total\_exp. To ensure that all characteristics contribute equally to the analysis and to keep features with greater magnitudes from dominating the model, this scaling approach converts the numerical values into a standard range between 0 and 1. Plotting histograms allows one to see the distributions of the scaled variables visually. Understanding the data distribution and spotting any possible trends or outliers is much easier with the aid of this graphic depiction.

**2.3 Data Distribution and Visualization**

**Dependent variable**

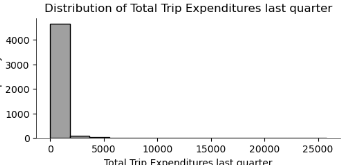


Fig.1

The histogram for "Total Trip Expenditures last quarter" (TTOTALC) indicates a right-skewed distribution. Most data points cluster toward the lower end of the expenditure range, with fewer households exhibiting much higher trip expenditures. This suggests a common trend of lower trip expenditures among households, with some outliers having significantly higher expenses.

**Independent variables**

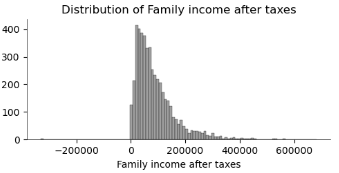


Fig.2

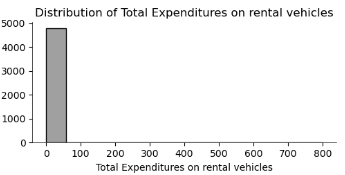


Fig.3

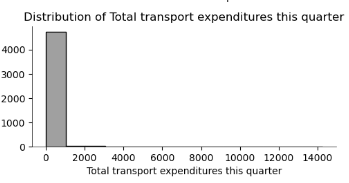


Fig.4

Independent Attributes: Variables like family income after taxes, total expenditures on rental vehicles, and total transportation expenditures this quarter are right-skewed. This suggests that most households have lower financial assets and spend less on certain categories, with a few having exceptionally high values.

**2.4 Modeling**

The study explored numerous variables as potential predictors of total trip expenditures. These included demographic factors such as age, family size, number of earners, marital status, and vehicle ownership. Financial attributes such as rental vehicle expenses, lodging expenses, estimated total taxes, and income after taxes were also analyzed to determine their impact on travel spending. These fields were chosen because they likely influence travel spending patterns. Demographic variables provide insights into household composition and age distribution, while financial attributes offer information on income levels and expenditure habits.

Applied statistical techniques such as linear regression and decision tree regression to analyse these predictors effectively. Linear regression, chosen for its simplicity and interpretability, quantifies the linear relationship between predictor variables and total trip expenditures, offering coefficient estimates for each variable. Decision tree regression, on the other hand, captures nonlinear relationships and interactions among variables, revealing complex patterns and variable importance. By leveraging the strengths of both methods, aimed to improve the models' predictive accuracy and gain a comprehensive understanding of the factors driving travel expenditures.

**3. Results**

**3.1 Model Validity and Significance**

Based on the tests performed, including Linear Regression and Random Forest models, the results showed moderate predictive performance in estimating total trip expenditures. The linear regression model yielded a mean squared error (MSE) of approximately 5.27 and a root mean squared error (RMSE) of about 2.29. In comparison, the Random Forest model produced an MSE of 5.61 for testing and 5.50 for validation, with corresponding RMSE values of 2.36 and 2.34, respectively. Cross-validation using K-fold improved model performance slightly, with an average MSE of approximately 4.55 and an RMSE of around 2.13. However potential issues with multicollinearity were noted due to high variance inflation factor (VIF) values for certain predictors, indicating possible correlations between them. Moreover, the reliability of certain predictors was challenged by the presence of higher p-values for some variables. Addressing these concerns may enhance the model's accuracy and reliability in predicting total trip expenditures.

**3.2 Significant Independent Variables**

After finding out the p-values and coefficients of each independent variable, significant variables were found if their p-values were less than the standard 0.05 and their relationship was based on the coefficient sign (positive or negative).

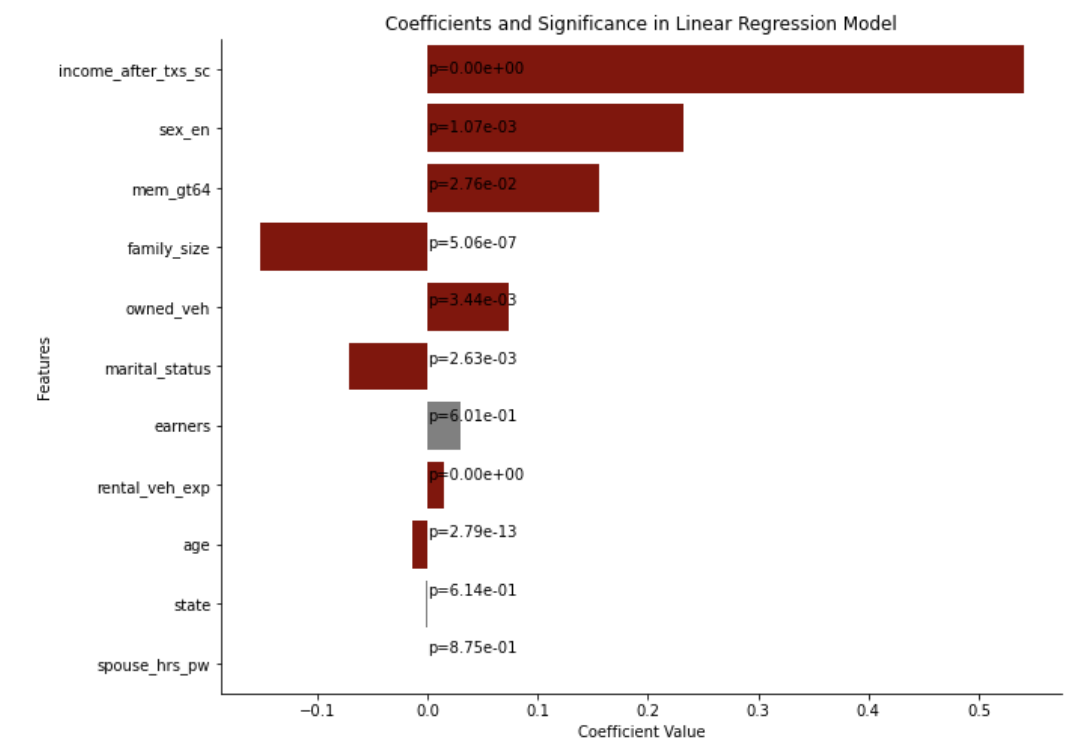
**age:** Age and the dependent variable have a significant negative connection, with a coefficient of -0.0142 and a p-value of 0.0000. This implies that as age increases, the dependent variable declines.

**family\_size:** There is a significant negative relationship between family size and the dependent variable, as indicated by a negative coefficient of -0.1525 and a p-value of 0.0000. This suggests that larger family sizes are associated with lower values of the dependent variable.

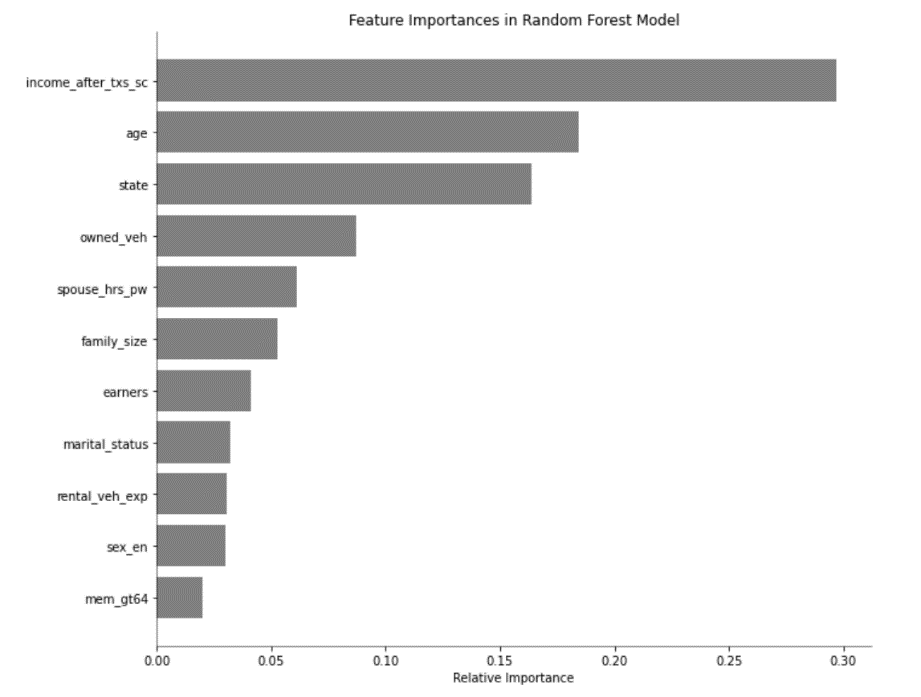
**mem\_gt64:** Mem\_gt64 shows a strong positive relationship with the dependent variable, with a coefficient of 0.1562 and a p-value of 0.0276. This suggests that households with members older than 64 tend to have higher values of the dependent variable.

**income\_after\_txs\_sc:** Income After Taxes has a positive coefficient of 0.5409 and a statistically significant p-value of 0.0000, showing a strong impact on trip expenditures.

**3.3 Visualization and Interpretation**

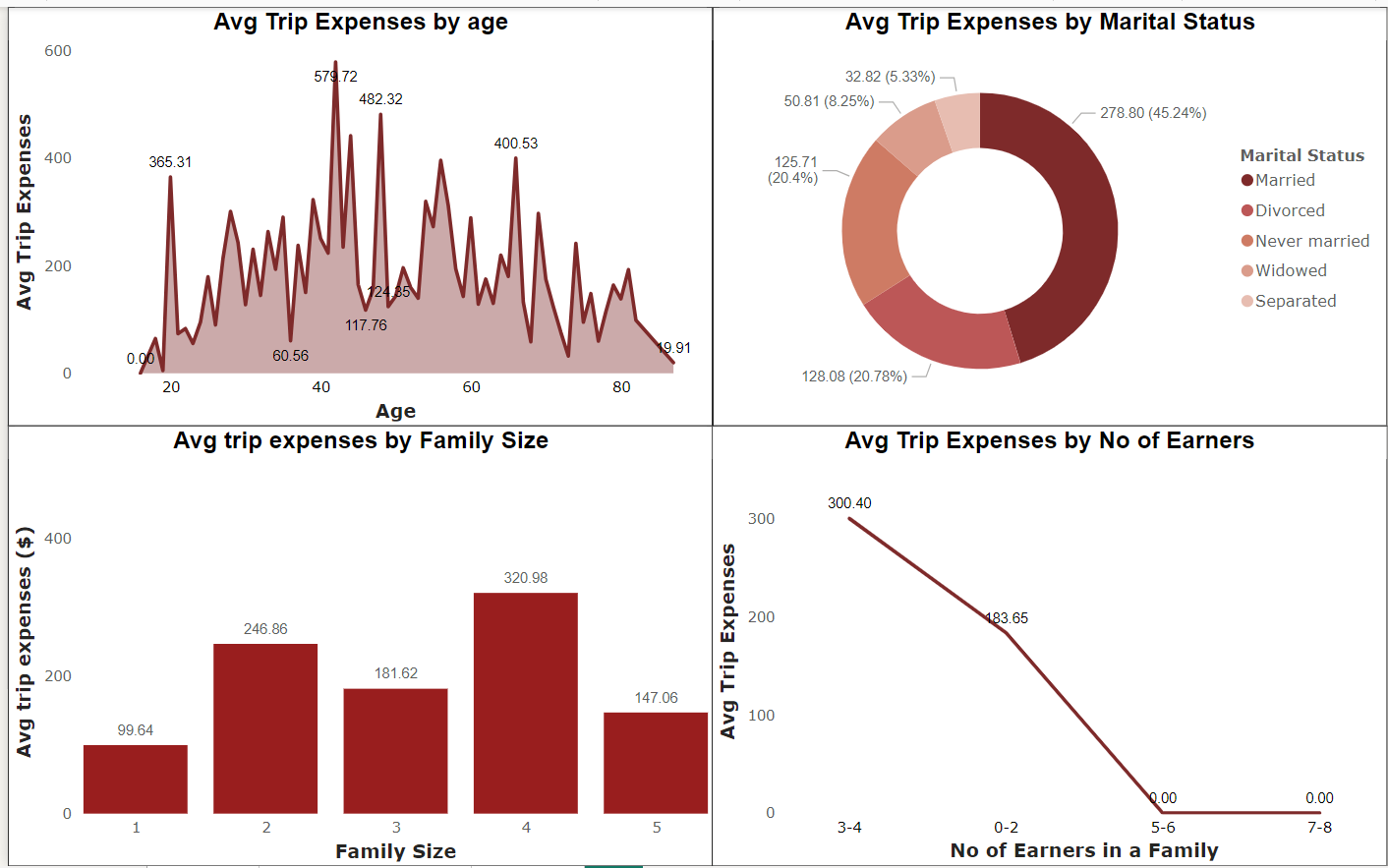


The graph represents the coefficients and associated p-values for various variables in a model predicting travel spending. The length and direction of each bar represent the size and direction of the variable's effect, with positive values suggesting increased travel spending as the variable increases. Variables with small p-values (typically less than 0.05) are considered statistically significant. The variable "Income After Taxes" has a strong positive coefficient and extremely low p-value, indicating that higher income after taxes is associated with higher travel spending. Similarly, "Rental Vehicle Expenditure" shows a positive and significant impact, suggesting that greater rental vehicle expenses correspond to increased overall travel spending. The "Age" variable has a negative coefficient and extremely low p-value, implying that older age groups tend to spend less on travel compared to younger groups. The variables "Earners" and "State" have p-values greater than 0.05, suggesting these factors do not significantly influence travel spending in this model.



The graph shows feature importance from a Random Forest model, ranking each variable by its influence on predicting travel spending. The model also shows "Income after taxes" as the most influential variable, indicating that a person's disposable income crucially determines their travel expenditure, with higher income levels potentially increasing travel spending. “Age” ranks as the second most important factor, underscoring varying travel preferences and financial abilities across different age groups. The "State" variable's significance highlights geographical factors like cost of living and travel destination accessibility, which can shape spending patterns. Additionally, the number of vehicles owned ("owned\_veh") mirrors a household's spending power or lifestyle, affecting their travel budget.

**3.4 Data Storytelling**

****

**Family Dynamics:** Larger family sizes like 4 in this case generally incur higher travel costs, possibly due to the need for more accommodation and transportation resources. However, very large families (size 5) might find ways to reduce per-person costs.

**Marital Status Influence:** Married individuals spend the most on trips, likely influenced by family travel or combined household incomes that facilitate more or more expensive travel.

**Age Impact:** Middle-aged individuals tend to spend more on travel, potentially reflecting peak earning years and a higher propensity or ability to travel.

**Income Correlation:** Households with more earners spend significantly more on travel, highlighting the direct relationship between the number of earners and travel expenditure, likely due to increased available income.

These insights can be particularly useful for businesses in the travel industry to tailor their marketing and products to these demographics. For policymakers or social planners, understanding these spending patterns could assist in developing targeted tourism support or incentives.

**4. Discussion**

**4.1 Limitations of the Project**

The data has several notable limitations that impacted the breadth and depth of the analysis. The most important limitation was lack of sufficient data. It is an evident fact that more data provides better results. The data set had 5,000 observations, which was too small for the kind of analyses that should be done to provide insights not just on a segment but of an entire industry. Another key limitation that was identified was the absence of information regarding the mode of travel which could helped in understanding more specific factors influencing travel expenses. Without this data, it is not possible to accurately assess how different modes of transport such as road, train, air, and water affect overall spending. Additionally, the absence of a time dimension, including specific months or seasons, prevents an analysis of seasonal variations in travel costs. Understanding these variations could be crucial for identifying spending patterns throughout the year. Lastly, the data set required extensive cleaning and imputation due to numerous quality issues, including missing values, duplicates, and inconsistencies. These steps, while necessary, could potentially introduce bias or errors into analysis, thereby affecting the reliability of conclusions.

The current travel industry in the United States was developed over the course of decades. People across the world travel here for various purposes like tourism, employment, education, medical treatment etc. So, there were a lot of factors that would have influenced in shaping up the current travel industry. Having the data of only three years (2019-22) in which a major portion of time was one of the least traveling times the country would have seen in history due to the pandemic, it is difficult to predict accurately.

**4.2 Recommendations for Future Research**

A number of suggestions for future data collecting and analysis can be taken into consideration in order to improve the validity and relevance of travel expense study. Above all, a more sophisticated examination of spending patterns would be made possible by the inclusion of multimodal travel data. Through the classification of expenses based on several modes of transportation, including air, land, train, and sea, researchers can discover distinct spending habits and preferences among travelers. This would improve the dataset and offer a more comprehensive understanding of the financial effects of different modes of transportation on total trip expenses.

An added advantage to studying seasonal changes in travel spending would be to include specific temporal data, including months and seasons, in the dataset. These improvements would provide more accurate analysis of the variations in travel expenses between high and low travel seasons, which would support forecasting models and business strategy for travel-related enterprises. Broadening the scope of the research might also provide insightful perspectives on cross-cultural variations and global travel behaviors if the study were expanded to include foreign travel. Furthermore, the use of panel data in longitudinal research could significantly enhance the understanding of the temporal evolution of individual travel expenditure patterns. This strategy would offer a dynamic perspective on travel patterns and allow for a longitudinal analysis that documents the evolution of consumer travel habits and economic impacts.

**4.3 Use Cases and Implications, Deployment Recommendation**

The research on trends in consumer behavior within the travel industry holds significant implications across various sectors and for different stakeholders. For travel companies, such insights are invaluable as they shape marketing strategies, product offerings, and pricing policies. Understanding consumer behavior allows them to tailor their services to match the preferences and needs of their target audience. For instance, if the research indicates a growing preference for eco-friendly travel options, companies could invest in promoting sustainable tourism initiatives or offer packages that align with these values.

Investors can leverage these findings to identify promising sectors within the travel industry for potential investment. For example, if the research highlights a surge in demand for experiential travel, investors may consider allocating funds towards businesses that specialize in providing unique and immersive travel experiences. Government agencies responsible for tourism promotion and economic development can utilize this research to design effective policies and programs. By understanding the factors driving travel spending and preferences, policymakers can strategically allocate resources towards infrastructure development, marketing campaigns, and other initiatives aimed at boosting tourism and economic growth in specific regions.

Market research firms can incorporate these findings into their industry reports, providing valuable insights for clients such as travel companies, financial institutions, and government agencies. These reports can inform strategic decision-making and resource allocation, helping businesses stay competitive in a dynamic market landscape. Academic researchers can also benefit from this research by exploring further studies on customer behavior while traveling and the broader impacts of tourism on the economy. Building upon existing findings, scholars can contribute to the body of knowledge in economic sociology and tourism science, enriching the understanding of these complex phenomena. Overall, the research on consumer behavior in the travel industry offers actionable insights for a wide range of stakeholders, from businesses and investors to policymakers and academics. By leveraging these insights, organizations can enhance their competitiveness, mitigate risks, stimulate economic growth, and ultimately, improve customer satisfaction and loyalty.

**5. References**

[1] Wilson, Gillian. “The Travel Industry’s Vital Role in Bolstering Our Economy.” *U.S. Travel Association,* April 26, 2023, www.ustravel.org/news/travel-industrys-vital-role-bolstering-our-economy.

[2] Council, World Travel & Tourism. “Travel & Tourism Economic Impact: World Travel & Tourism Council (WTTC).” *World Travel & Tourism Council, December* 7, 2023, wttc.org/research/economic-impact.

[3] Barten, Martijn. “Travel Industry: An Overview of One of the Largest Service Industries.” *Revfine.Com*, 22 Jan. 2024, [www.revfine.com/travel-industry/](http://www.revfine.com/travel-industry/).

[4] Hall, C. Michael, et al. “Airbnb and the Sharing Economy.” Current Issues in Tourism, vol. 25, no. 19, 22 Sept. 2022, pp. 3057–3067, doi:10.1080/13683500.2022.2122418.

[5] “2024 Travel Industry Outlook.” Deloitte United States, 1 Feb. 2024, www2.deloitte.com/us/en/pages/consumer-business/articles/travel-hospitality-industry-outlook.html.

[6] baig, Rafiqa. “Government’s Role in Promoting Investments.” *LinkedIn*, 19 Oct. 2023, www.linkedin.com/pulse/governments-role-promoting-investments-rafiqa-baig/.

[7] Takahara, Kanako. “Japan’s Domestic Travel Discount Program Is Back. Here’s What You Need to Know.” *The Japan Times*, 10 Jan. 2023, [www.japantimes.co.jp/news/2023/01/10/national/japans-domestic-travel-discount-program-back-heres-need-know/](http://www.japantimes.co.jp/news/2023/01/10/national/japans-domestic-travel-discount-program-back-heres-need-know/)

[8] Share, et al. “The Impact of Augmented Reality on the Hospitality Industry.” Booking.Com for Partners, 20 July 2022, partner.booking.com/en-us/click-magazine/trends-insights/impact-augmented-reality-hospitality-industry.

[9] Ramgade, A., and A. Kumar. "Changing trends of hospitality industry: Emergence of millennials and gen Z as future customers and their influence on the hospitality industry." *Vidyabharati International Interdisciplinary Research Journal* 12.01 (2021): 336-342.

[10] “Covid-19: Implications for Business.” *McKinsey & Company*, McKinsey & Company, 13 Apr. 2022, www.mckinsey.com/industries/travel-logistics-and-transport-infrastructure/our-insights/covid-19-implications-for-business.