CAPSTONE

Northeastern University

ALY6980, SPRING 2021

MODULE 3 PROJECT ASSIGNMENT

WEEK 3: ANALYSIS OF COLLECTED DATA

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

SUBMITTED TO: DR. ERIC (YANG) LIU

DATE: 05/27/2021

Introduction

The assignment aims at performing Exploratory Data Analysis for the At Home Group data. We have used the previously prepared home furnishing data, with additional appended variables for performing initial analysis, by demonstrating the important factors of data in the form of data visualizations. This data analysis will help the At Home Group's management in effective decision making for improving the business for the home décor industry. Moreover, considering some of the existing parameters, we have performed initial data analysis that will help in understanding the parameters which will enhance the business for the company. In this assignment we are focusing on ways which will help the home décor industry in improving profit margin for the company.

The industries are very engaged in making profits through their own traditional methods, and therefore, the important business parameters are neglected by these industries. It is important for industries to emphasize upon the customer audience and current market scenario based on the demographic characteristics of every location, which will benefit in relevant advertising and marketing for the business. We have worked on extracting the zip code level data based on the demographic, regional, tax and employment parameters for the At Home Group stores, and further used this data for performing exploratory data analysis, thereby pointing out some of the important parameters for the business.

Methodology

The methodology implies to the systematic approach in working on a particular problem towards attainment of a solution. In order to prepare our data for data analysis, we collated all the zip code level, location based data pertaining to, demographic, regional, tax and employment parameters from authenticated sources for various store locations, into an excel file. Further, this data was appended with additional location based parameters like, designated market area (DMA), Simulated Sales, Address, City, States and Year for every store location. We appended and collated the entire data in the Microsoft Excel file using Vlookup formulas which enabled us to match values for every record effectively.

Initially, we performed data cleaning for the collated data in order to eliminate data inconsistencies, data duplicates and errors from the data, to make the dataset compatible for data analysis. The data cleaning process in very important in performing data analysis as it helps in improving the quality of data which helps in yielding accurate results later. The data cleaning process involved the following :

- DMA: After performing the VLOOKUP for the data, it was observed that, all the DMA values were not present in the Store Sales excel sheet. Therefore, I performed web scraping on a website that provided the DMA data for every store location. Therefore, all the missing values were filled with DMA values for every store location.

- Simulated Sales: The sales provided in the Store Sales dataset were not sufficient for all the store locations and therefore the data had some missing values, after performing

VLOOKUP. Therefore, we filled up remaining sales by calculating the mean values of all the sales and filling the null values with the mean values.

- Year: The years for all the store locations were added based on matching of the VLOOKUP values. However, the missing values were added by performing VLOOKUP with the data extracted from online sources earlier, and the subsequent values were filled up.

- City, State and Address: The segregation of address into City, State and Address was performed in R and Excel, by splitting the textual data into city, state and address.

The data cleaning process was performed in R Studio, a statistical analysis tool and Excel. During the process of performing exploratory data analysis, I worked on R Studio, Excel, Python, used heavily for data analysis implementations and Tableau. However, Excel and Tableau seemed more easier and efficient for data analysis as the data required basic cleaning which was feasible in Excel, as the extracted data from various online sources was in the csv format. Moreover, Tableau is a great business intelligence tool for implementing data visualizations (Singh, D. 2020, June 24), therefore, I selected the tool, Tableau for exploratory data analysis. In the further part of data analysis for the At Home Group data, I intend to work on Python and R for major data analysis, as these analytical tools will be more efficient in implementing data analysis algorithms.

After analyzing the data, it was observed that, regression analysis would be appropriate for further data analysis on the data. This analysis will help in understanding the dependent and independent variables which would help in predicting the right variables for enhancing the business for the store. In addition, Cluster analysis or K Means is recommended for the data, as this approach will help in predicting the customer buying patterns which would enable effective data driven decision making through market segmentation. Therefore, I intend to work on Regression and Cluster analysis for the further data analysis.

Analysis

The Exploratory Data Analysis was performed using Excel and Tableau tools for the data. These tools seemed effective for performing preliminary data analysis, as the analysis involved initial analytics on data.

While performing exploratory data analysis, It is important to address the business questions that an organization can encounter in future. Some of the business questions, that were analyzed are:

- <u>Understanding Customer buying patterns: What is the likelihood of customer buying specific item from the store?</u>
  It is very important for the businesses to understand the buying patterns of their potential customers. This will help them to sell their products effectively and market their items accordingly. I have performed analysis on the consumer goods by the top 10 states based

on the median household income of people. I used the consumer goods data for every state based on the median income to plot a bar graph for the data.

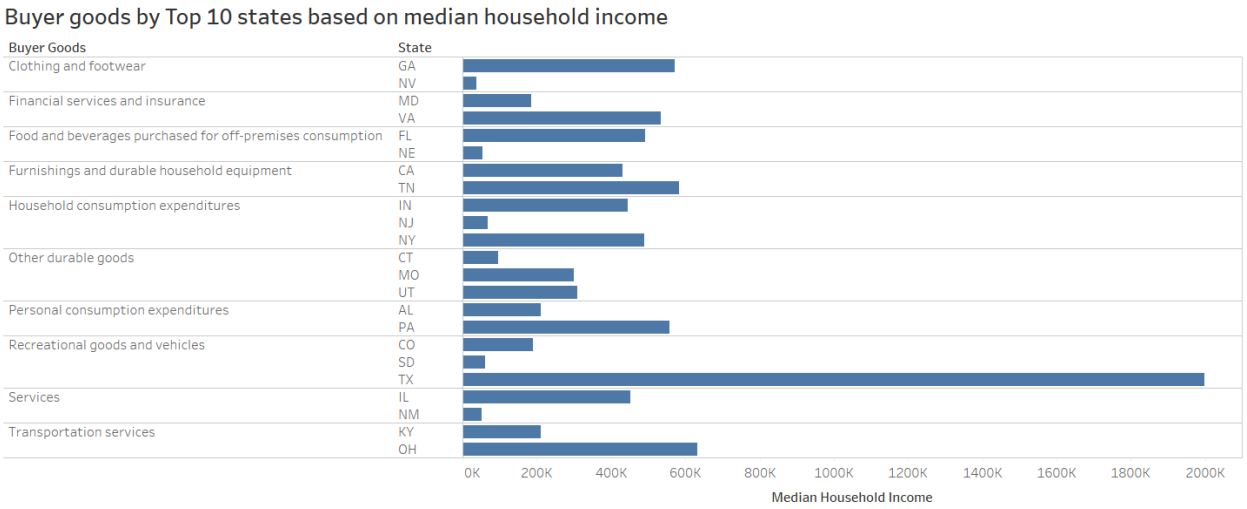

Fig.1: Buyer goods by top 10 states based on median household income

The above bar graph shows the buying patterns of consumers in the top 10 states based on the median household income of people in the respective states. As seen, the people in Texas, with median household income of $19,98,988, seem to spend more on Recreational goods and vehicles, whereas people in Nevada, with median household income of $37,014 seem to spend the least on Clothing and Footwear. Moreover, people in California and Tennessee tend to spend more on furnishings and household equipment. This analysis has helped in understanding the buying patterns of customers at various states, which would help businesses in retaining potential customers.

- <u>Importance of sales at various locations: What has been the sales at a particular state over a period of time?</u>
  This issue is crucial for the businesses in understanding the sales made at the store's locations during a period of time. Also, this will provide details about the designated market areas in every state that are likely to make higher profits, based on the sales made in that area at a particular period of time. The sales at a given location can be determined by the population, buying patterns of people at the location, ethnicity, employment etc.

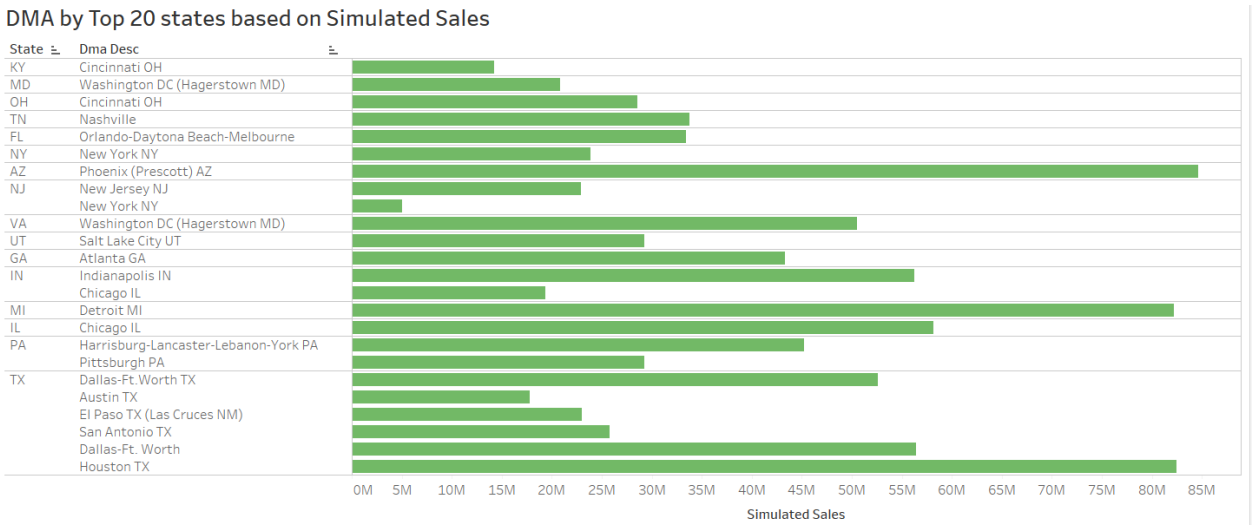

Fig.2: Designated Market Area by top 20 states based on the simulated sales

The above bar graph shows the Designated Market Area by the top 20 states based on the simulated sales. This helps in understanding the sales made in various designated market areas at various states. Every state has DMA where the maximum sales are observed. As observed, Phoenix in Arizona has observed highest sales and is the DMA for Arizona, whereas, New Jersey and New York in the New York state have the highest sales. These designated areas help the businesses in understanding the likelihood of sales at target locations within a state.

- <u>Understanding the employment at various locations: What is the employment ratio of citizens at a given location?</u>
  As we know, higher the employment, higher would be the chances of people visiting stores and buying goods. Therefore, it is important to understand the employment rates of people living at the store locations (State / County) which will enable the store management in deciding the areas to target for opening new stores.

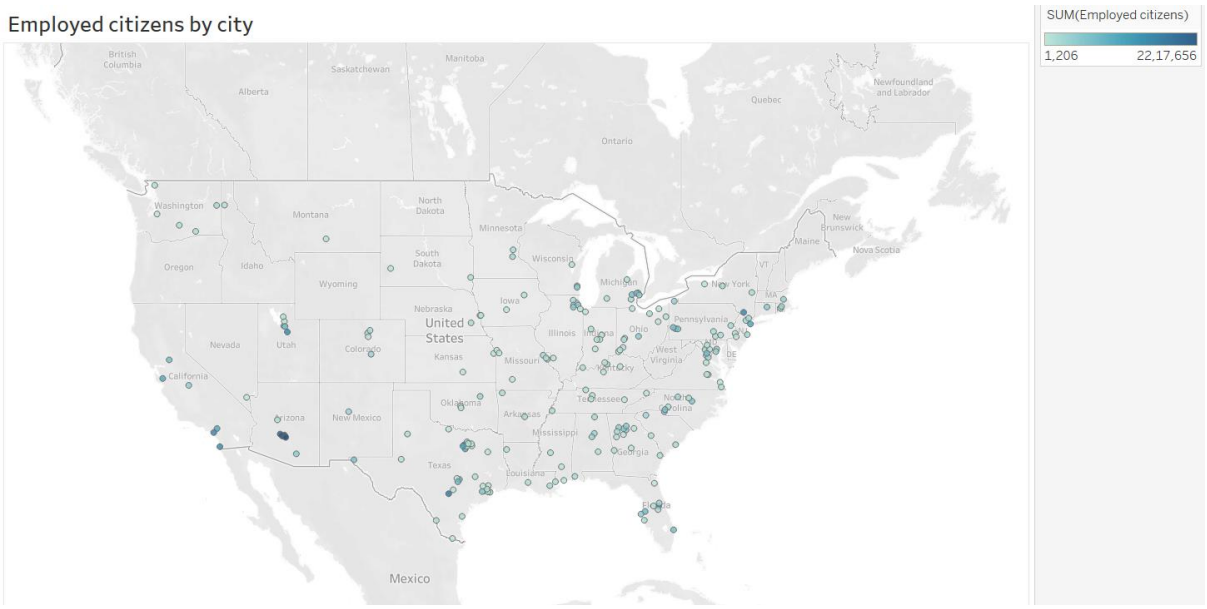

Fig.3: Employed Citizens by City

The above map shows the level of employment at various store counties in the United States. As seen, darker the shade, higher would be the employment, whereas lighter the shade, lower would be the employment. It can be observed that, the cities in Arizona seem to have higher number of employed citizens than, cities like Cincinnati, that have lower employment. This analysis would help in determining higher number of educated individuals in a state.

- <u>Understanding the Sales by every State: What has been the overall sales in a given state?</u>
  It is important for businesses to understand the sales made in the states where their stores are located. This will help them in analyzing the target audience and states that are likely to make more profits.

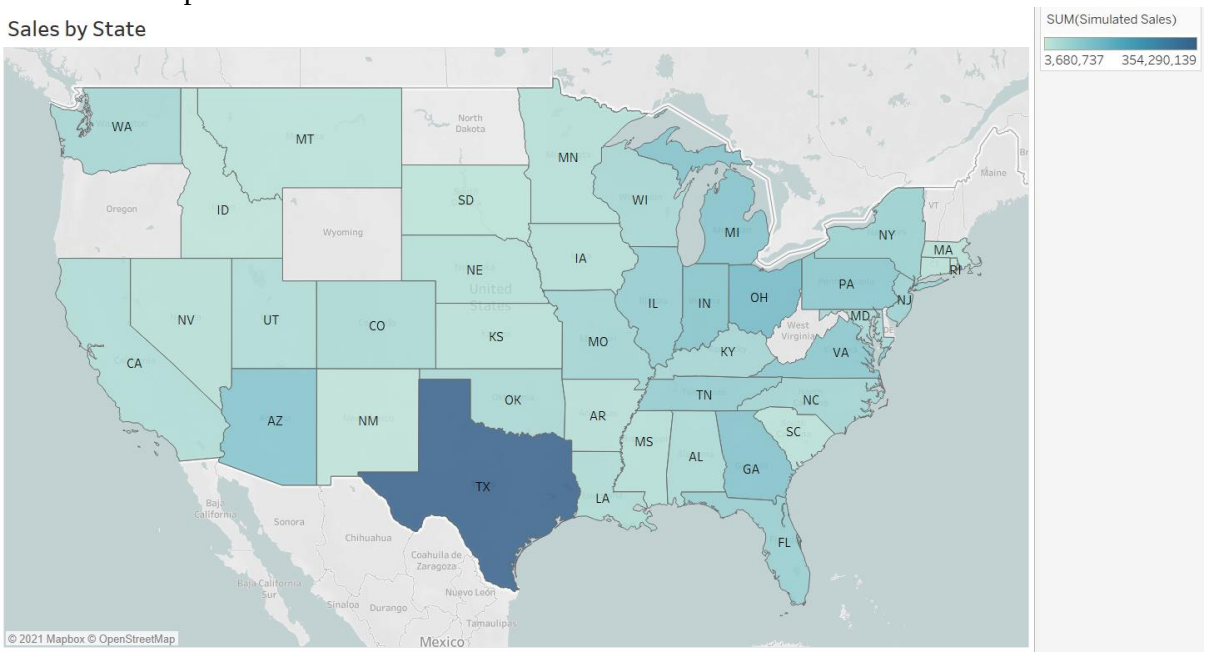

Fig.4: Sales by State

As observed in the above map, Texas shows to have higher amount of sales in the united states whereas Rhode Island shows to have lower amount of sales. The map shows the sales made at various states in the united states, darker the shade, higher is the sales, lighter the shade, lower the sales.

- Understanding the state expenditure on consumer goods: How much has the state invested on consumer goods?
  Every state has their own policies and quota that the government invests for the consumer goods. It is important for businesses to understand the investment made by various states on various consumer goods.

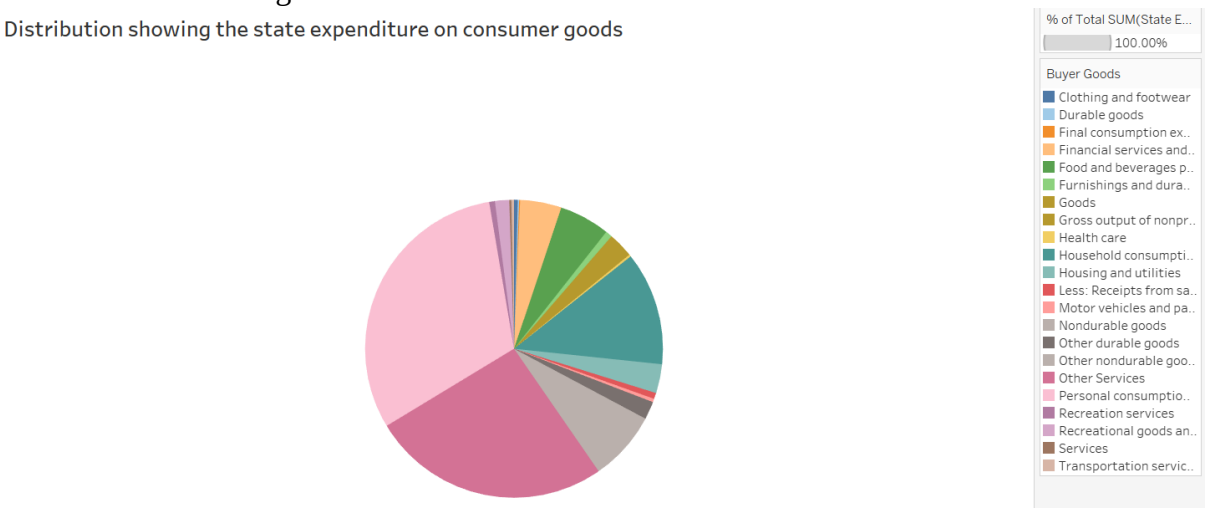

Fig.5: Distribution showing the expenditure on consumer goods

As observed in the above pie chart, which shows the percentage distribution of expenditure on various consumer goods. It can be observed that, the states tend to spend more amount on Personal Consumption Expenditures accounting for 30.99%, whereas, the states spend the least on Recreation Services accounting for 0.61%. Also, around 3.15% is accounts for housing and utilities. This will help to have an understanding of the amount invested for the furnishing services by the states, so that the management can plan its revenue accordingly.

- Understanding the taxes paid by the states: How much has been the tax payment by various states per capita?
  This will enable to know, the tax deductions at various states and amount paid to other states, which will help in deciding the amount to invest on a particular store at a given state.

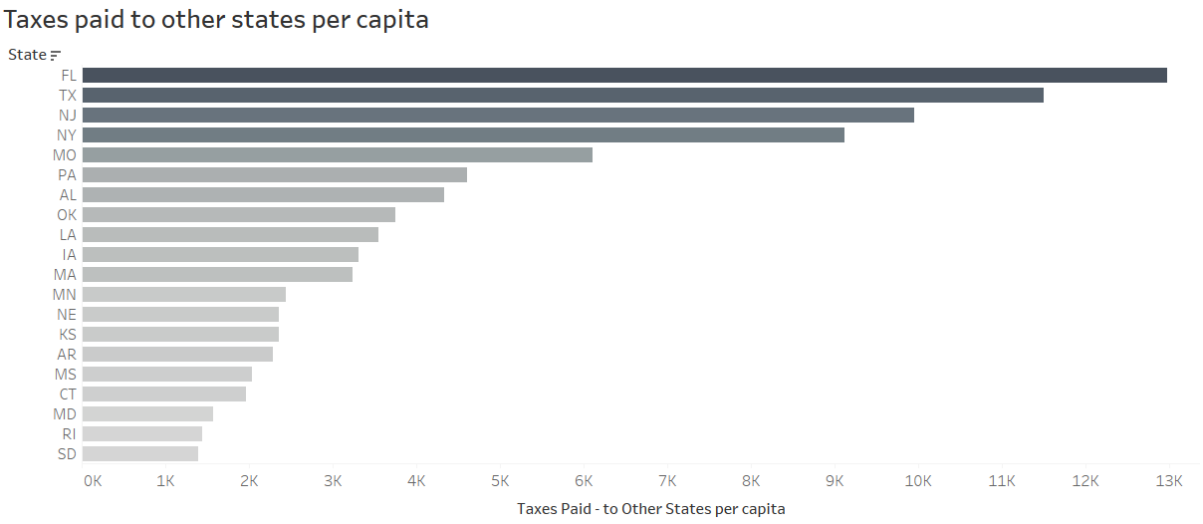

Fig.6: Taxes paid to other states per capita

The above bar graph shows the top 20 states that paid the tax amount to other states. It can be seen, that, Florida paid highest tax amount whereas, South Dakota paid the lowest tax amount.

- <u>Understanding Per Capita Income for top 20 counties: What has been the per capita income for counties?</u>
  The analysis on per capita income at various counties will enable businesses in understanding the income of families present in the counties, so that the sales at their stores can be predicted accordingly.

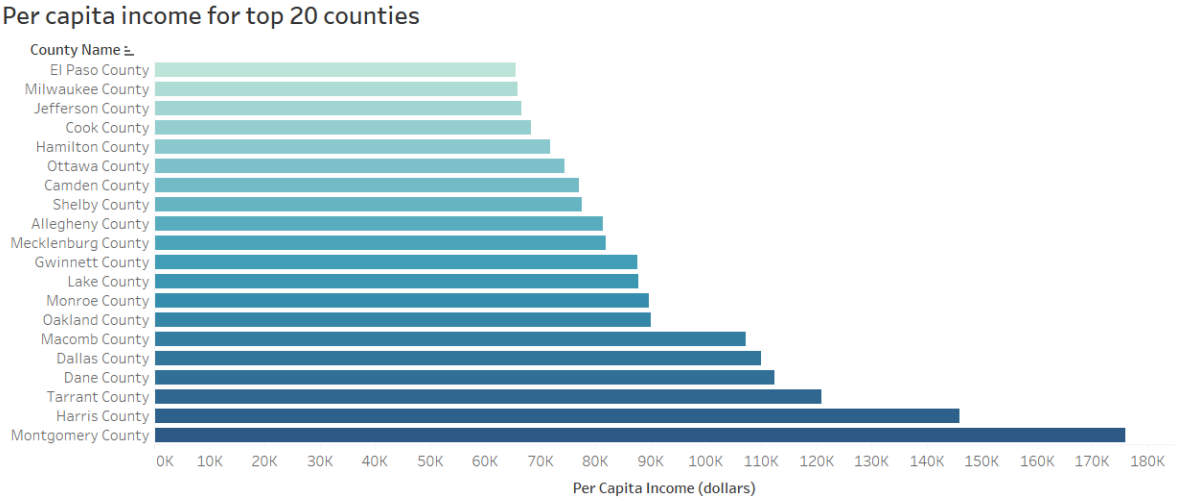

Fig.7: Per Capita Income for top 20 counties

As observed in the bar graph above, that shows the capita income for the top 20 counties in united states, the highest income has been noted in the Montgomery County, whereas, lowest per capita income was noted in the El Paso County.

- <u>Understanding the poverty boundaries at various states: What was the poverty level at a given state?</u>
  In order to predict the likelihood of sales in a given location, it is important to understand the unemployed and people below the poverty line at various locations.

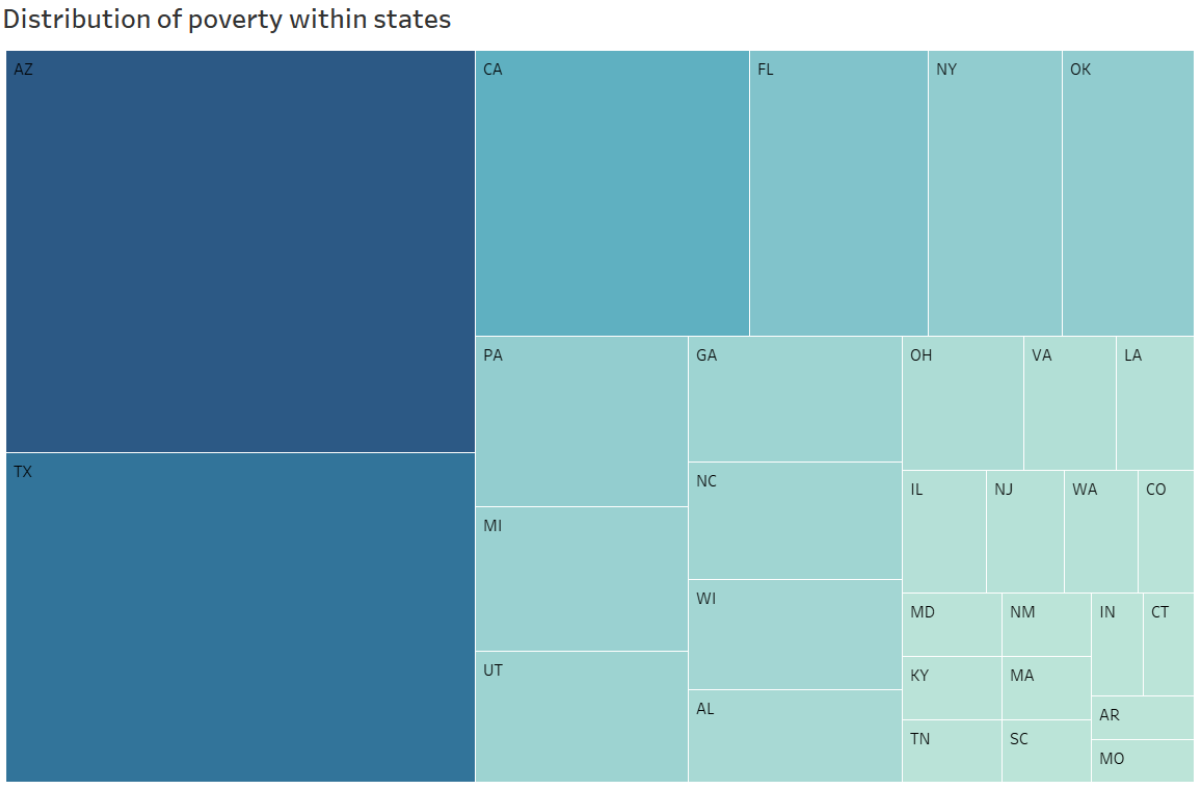

Fig.8: Distribution of poverty within states

The above tree map shows the distribution of citizens who are below the poverty line within various states in the United States. It can be observed that, Arizona has the highest level of poverty, whereas Missouri has the lowest level of poverty. This parameter is also dependent on the population and education level of people within that particular state.

Conclusion

After performing the Exploratory Data Analysis in Tableau and Excel, it can be observed that, for a business to gain profit and revenue, they must consider the consumer buying patterns, employment of citizens within various locations, tax level detail for the states, the median household income per capita at various locations, ethnicity of individuals at various locations and state-wise expenditure on consumer goods. These seem to be important parameters to consider from a business standpoint as this will help the management in data driven decision making to improve the sales of their stores. Moreover, these factors will help the businesses in understanding the market segmentation, thereby enabling them to focus on potential customers. Therefore, I would recommend businesses to focus on market segmentation while marketing their goods. The data analysis shown in this document, gives a high level view of initial data

analysis, however, there are some parameters pertaining to taxation that could have bolstered the analysis, and intend on implementing the tax level details in the follow up analysis with regression algorithms.

References

Grolemund, H. W. and G. (n.d.). *R for Data Science*. 7 Exploratory Data Analysis | R for Data Science. https://r4ds.had.co.nz/exploratory-data-analysis.html.

Singh, D. (2020, June 24). *Deepika Singh*. Pluralsight. https://www.pluralsight.com/guides/exploratory-data-analysis-with-tableau.