Term Project Write-up

Felipe Rodriguez

DSC 540 Data Preparation

Professor Catherine Williams

August 12, 2023

This project involved exploring GDP Data in the United States, House Price Data, and Population Data. Using this data, an analysis was conducted on the relationship between the three and can be looked at by city and state.

This project was created with three data sets and provided some insights on Population, Home Sales, and GDP. The plots that were created explain some of the trends of the data. For example, “Scatter Plot: Total Sales vs Population” shows that there is no clear trend between those two variables. The heatmap that follows, “Correlation Matrix: Population and Total Sales”, confirms this. However, in the Scatter Plot “GDP 2022 vs Total Sales”, a positive trend is seen. When GDP increases, Total Sales increases as well. The heat map “Correlation Matrix: GDP and Total Sales” displays a much higher correlation between the two than Total Sales and Population. Among the plots created, it is clearer that GDP has a bigger effect on Total Sales, however, Population and GDP are correlated as well since most states with high populations have high GDP. This can be seen in the Diverging Bar Plots “States by Population” and “States by GDP”.

This project involved new concepts that I had not explored before, such as reading Website data and APIs. These two were the most challenging. Some of the transformations conducted such as reading the populations from the API and creating a data frame, involved longer functions that needed to access the correct subset in the JSON as well as take into the time in between each request. Reading the website data created some challenges since the data was in a multilevel index, and an index needed to be removed from the data we needed.

Gathering data created some ethical implications, one being the credibility of the sites where the data is being used from. Most sets are verified; however, a validation of data quality

would be beneficial in this project. Another ethical implication is the timing of the different data sources. Population is displayed as a total, however, if population could be displayed over time, this could give more insight to the relationships in the data. This is the same for GDP. GDP is only included for 2022 and 2023 where the sales data has from 2008-2020.

Overall, this project helped with the understanding of GDP Data in the United States, House Price Data, and Population Data. Although there are some ethical implications, this topic can continue to be explored in the future with more data and resources.