Jonathan Broadbridge

Dr. Scott

BAN 530

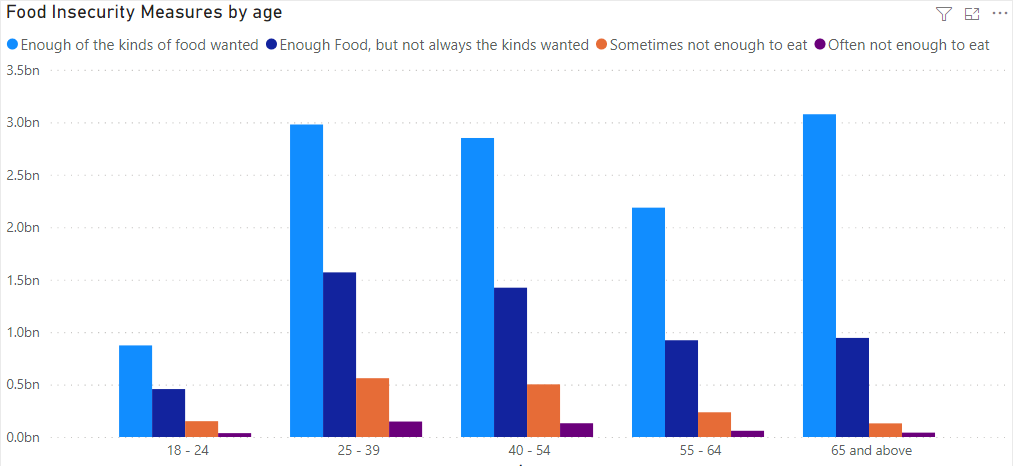
11/8/2021

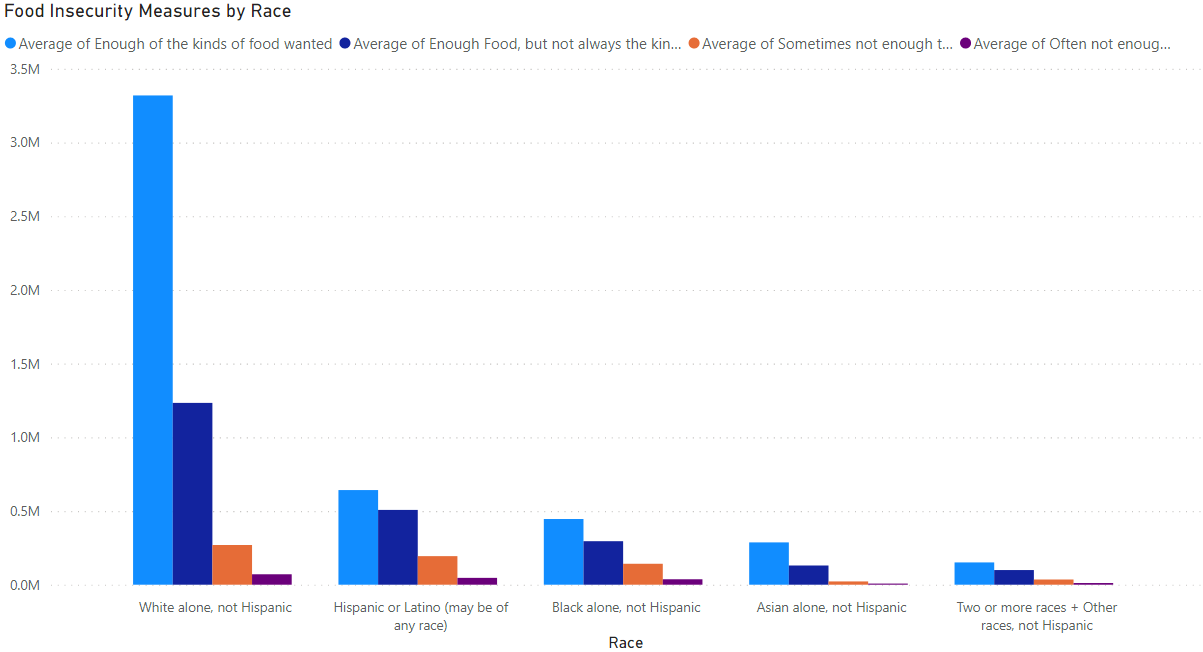
**Data Cleaning & Understanding (Descriptive Analytics: Part 1)**

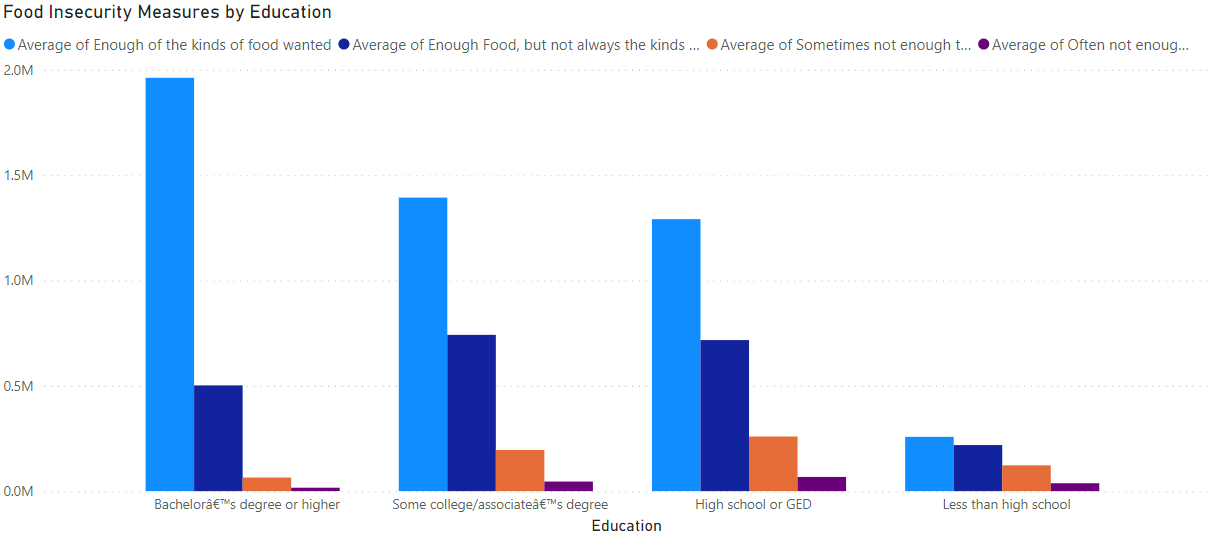
**Introduction:** This week's assignment was toperform an initial EDA of the data and perform data cleaning to make the data analysis process. The tools used to perform the EDA were Python and PowerBI. Graphs were made in PowerBI which looked at rates of food insecurity by the various variables of interest that were highlighted in report 1. Some interesting trends were observed and will be of concern when performing modeling. In Python I filtered the original dataframe down to a reference dataframe from which data frames of tuples were created. These tuples looked at food insecurity measures for each variable of interest indexed by Location and then Week. By doing this I hope to make it simpler to append these various smaller data frames into one based on stacking tuples within each index value.

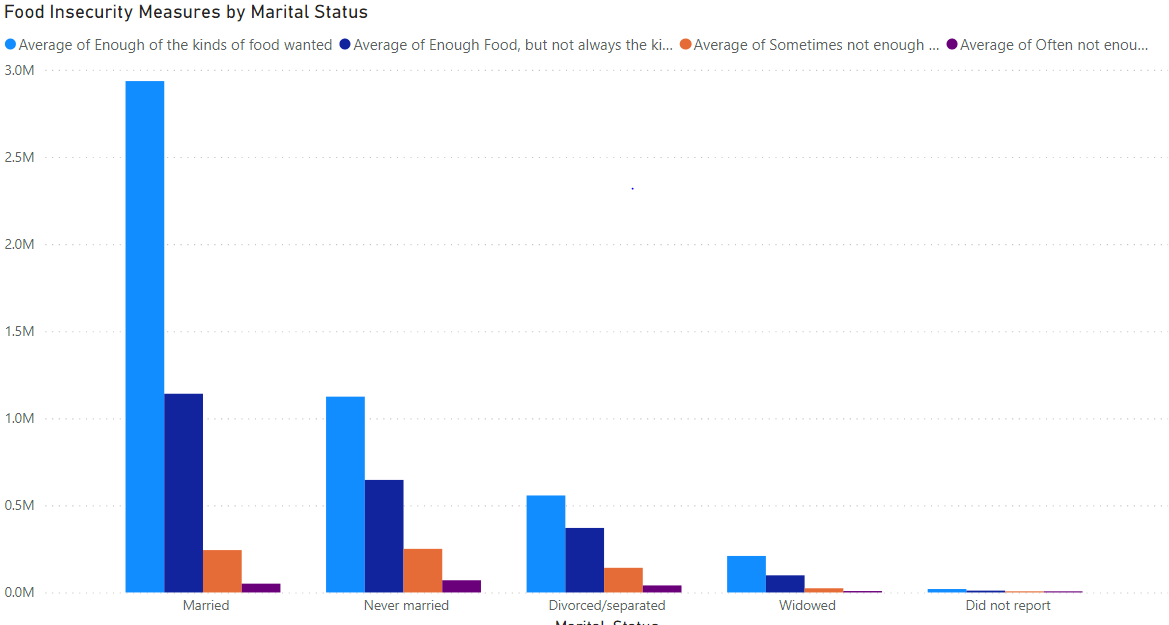
**Variables of Interest and insights:** The key variables of interest in this study are state, age, education, race, sex, marital status and children status. When looking at descriptive statistics and visualizations of these variables with concern to food insecurity we can see some things such as higher levels of education and age leading to more food security and less food insecurity. Marital Status also indicated a trend. Those who were married were more likely to be food secure and individuals who were never married had higher rates of food insecurity than those who were married. Children in the household were more associated with food insecurity than childless households. Location may be another interesting variable. There are several locations that have the highest levels of food security as well as food insecurity by count but the ones that are most interesting are the places like Dallas and Houston Metro areas having high levels of insecurity where they do not appear in the top food secure places. Once the data is run through the modeling process we can see with greater detail the extent of these trends.

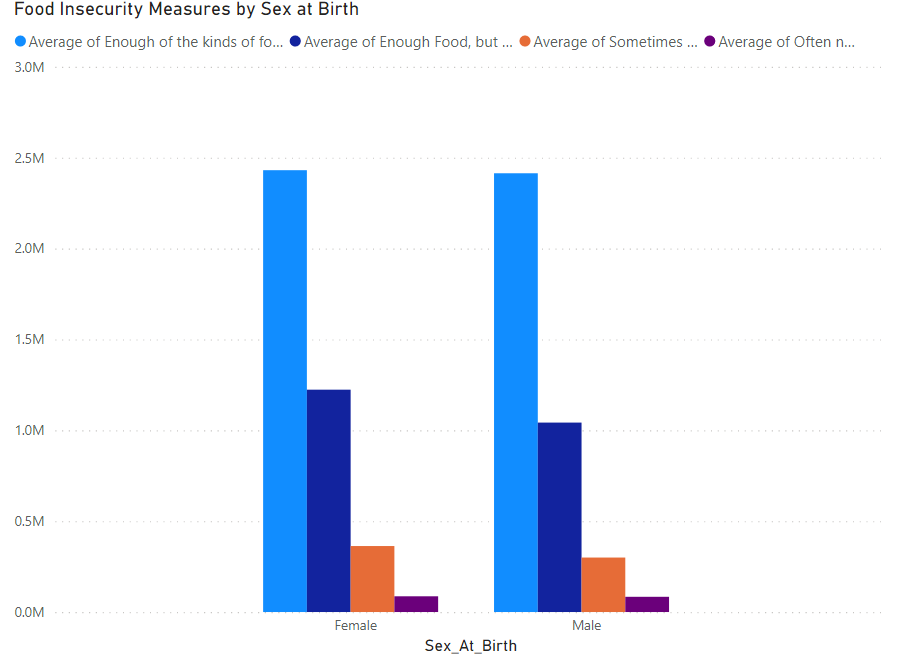
**Visualizations:** Visualizations were done in PowerBI and show rates of food Insecurity measures for each of the variables of interest. Below are the graphs.

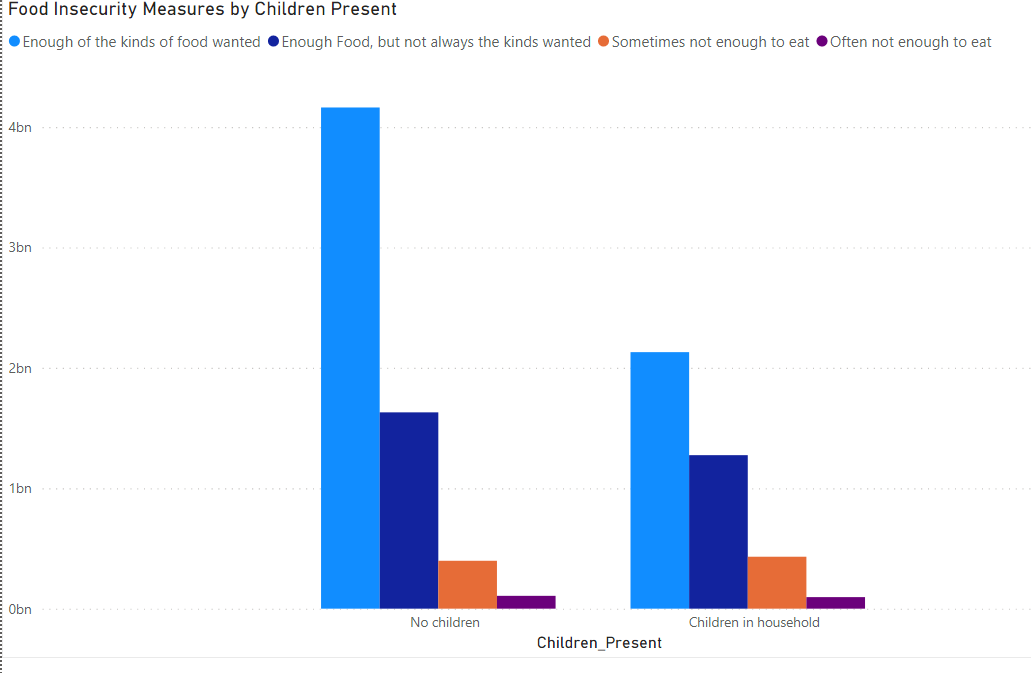


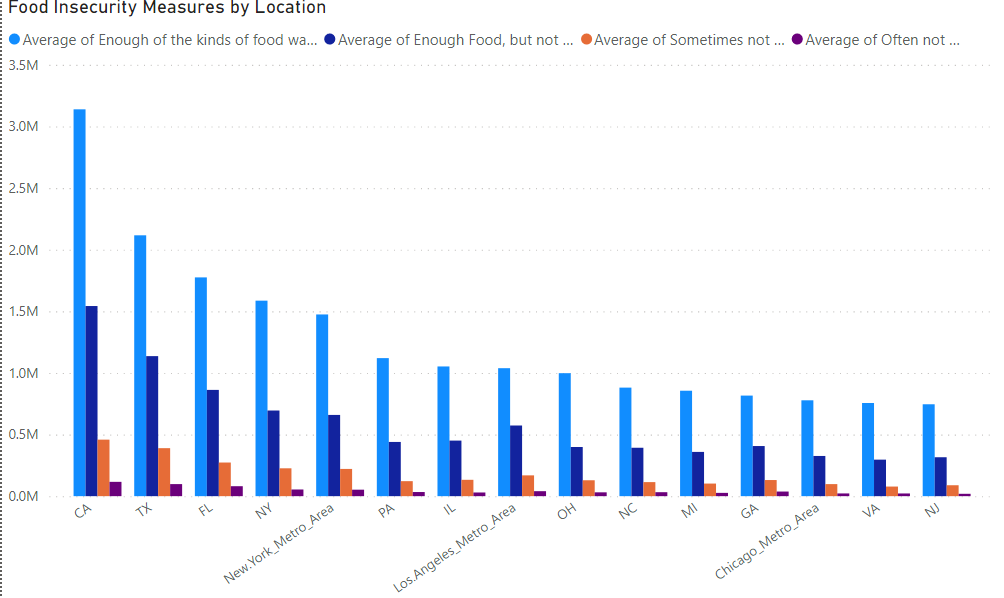


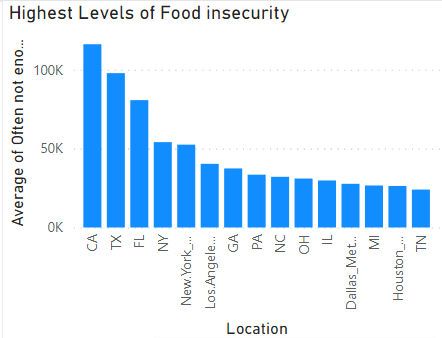












It is important to note that all of the above graphs are using an aggregate of the data for its main measure. Average of each food insecurity measure is used and we may see differing trends from the data over time.

**Conclusion:** Overall we have made some decent progress on this analysis. We have established that some variables may be indicative of food insecurity while others indicate food security, exploratory data visualizations have been constructed giving us some direction for forming a hypothesis and we have broken down the data into more digestible pieces with each data frame containing 5 columns in tuple form. This will make the data much easier to analyze with computational methods. Next week the data will be broken into training and testing splits which I plan on using a non random split serializing on the week variable.