EDA : Analyzing the Prevalence of Foodborne Outbreaks with the Changing Seasons

Surena Nokham

Department of Science & Technology, Bellevue University

DSC 530-T303: Data Exploration and Analysis

Professor Matthew Metzger

June 3, 2023

Summary

For my Exploratory Data Analysis (EDA) term project, I decided to investigate the prevalence of foodborne outbreaks to occur during a specific season: summer. My hypothesis states that cases will peak during the summer season (June, July, August) due to the warmer temperature, and frequency of people to cook and leave food outside.

Overall, the outcome of my EDA project was as expected for my hypothesis, but there was not enough evidence to reject the null hypothesis (see “Term Project - Test Hypothesis” file). There is a strong relationship between the variables as shown in the “Outbreak Case Frequency by Month (2005-2015) histogram, CDF, and analytical distribution calculations. Outbreaks are likely to occur consistently through the spring, fall and winter season but with a particular spike in the summer. The correlation between variables are positive but further analysis is required to establish a causal relationship amongst the variables evaluated.

I realized while doing the analysis,’ my dataset needed more quantitative variables, such as actual counts of confirmed outbreaks in each month or a quantitative number from each of the locations and from what month (whether the cases originated inside or outside). It was challenging working with a dataset full of categorical data that required so much integer conversion and missing values. Even after doing the analysis, I feel like I still do not have a large grasp on understanding the some of the data results, but I do feel like I have proven that foodborne outbreaks peak during the summer season.