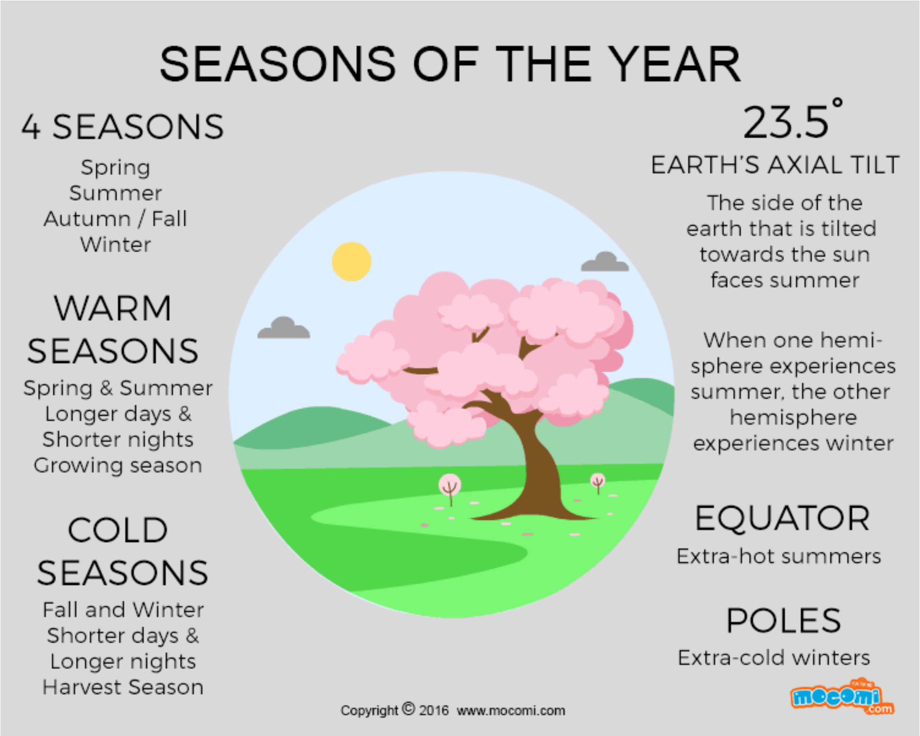
AN EXPLORATION OF THE SEASONS AND FOODBORNE ILLNESSES

TERM PROJECT

Surena Nokham
DSC 530 – T303 DATA EXPLORATION & ANALYSIS
Professor Matthew Metzger

STATISTICAL QUESTION/HYPOTHESIS



Question: Does the prevalence of foodborne illnesses vary with the changing seasons?

Hypothesis: My prediction is that cases will peak during the summer seasons (June, July, August) due to the warmer temperature where bacteria can multiply at a quicker rate, and the frequency of people to cook/leave food outside.

VARIABLES

(describe what the variables mean in the dataset)

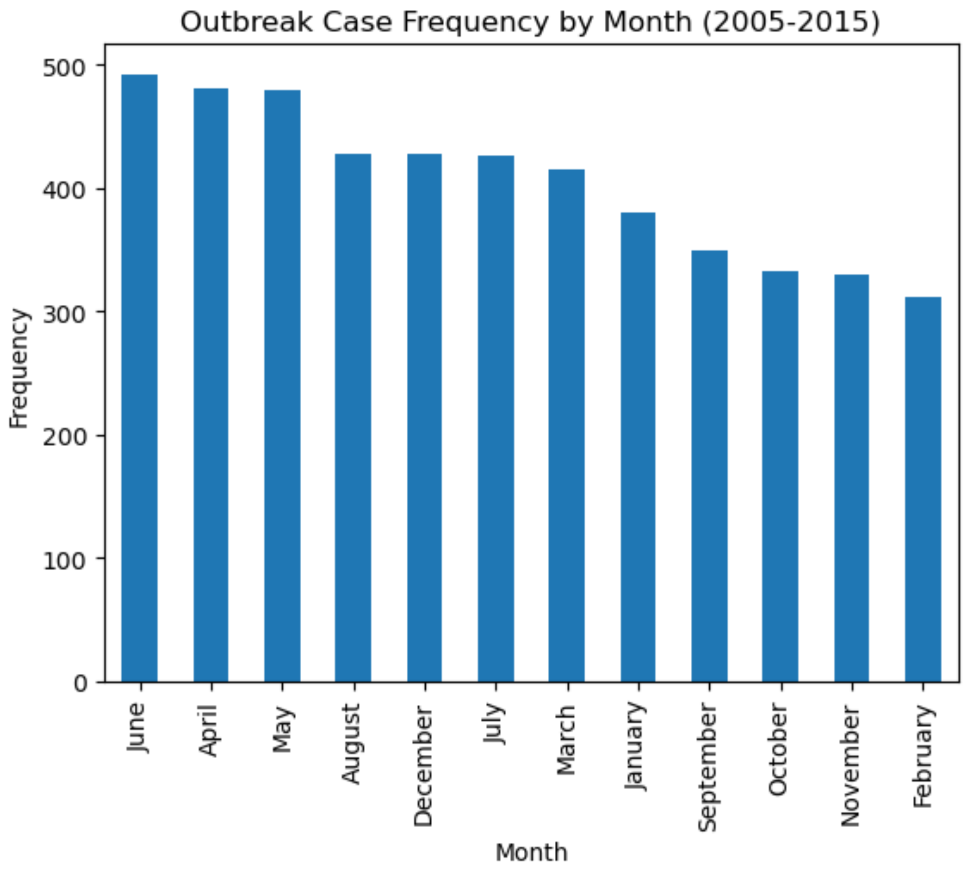
- Year – foodborne disease outbreaks reported to the CDC from 1998 through 2015
- Month – specific month outbreak occurred within the year
- State – specific U.S. state where the outbreak took place (outbreaks occurring in more than one state are listed as "multistate")
- Location - where the food was prepared
- Food - identified source of the reported foodborne illness
- Ingredient – contaminated ingredient
- Species/Genus - etiology (the pathogen, toxin, or chemical that caused the illnesses)
- Status – confirmed etiology (suspected cases not included in this dataset)
- Illnesses – total count of reported illnesses
- Hospitalization – total count of people admitted to hospital from foodborne illness
- Fatalities – total count of resulted death from foodborne illness

VARIABLES CONTINUED...

For this project, I will focus on the following variables in my EDA as they are more closely related to my hypothesis:

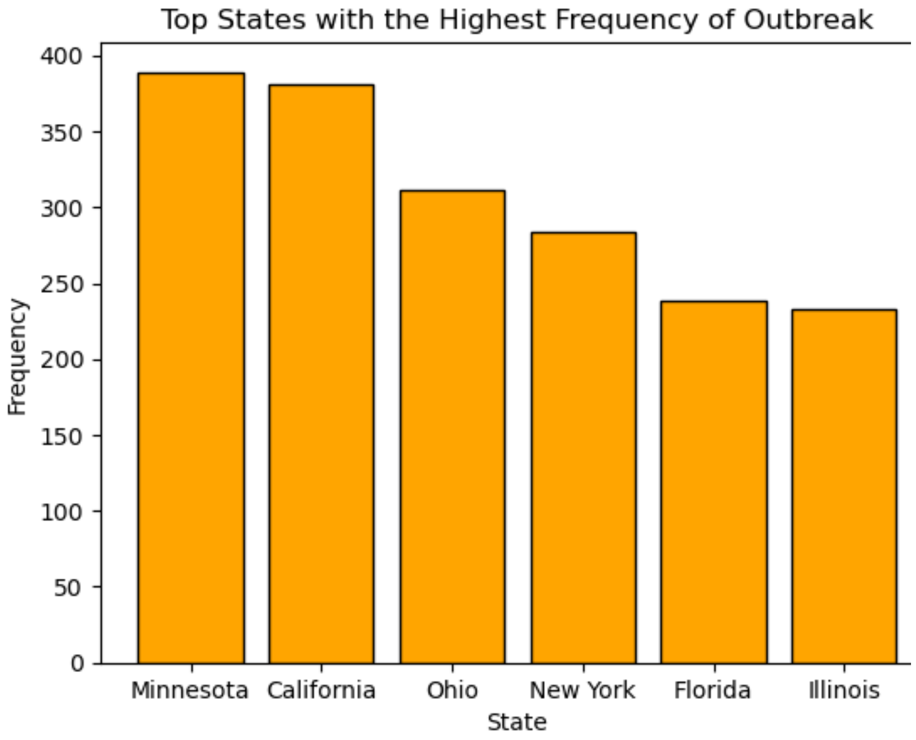
1. Year
2. Month
3. State
4. Ingredients
5. Genus
6. Ilnesses
7. Hospitalizations

HISTOGRAM OF VARIABLES - **MONTH**



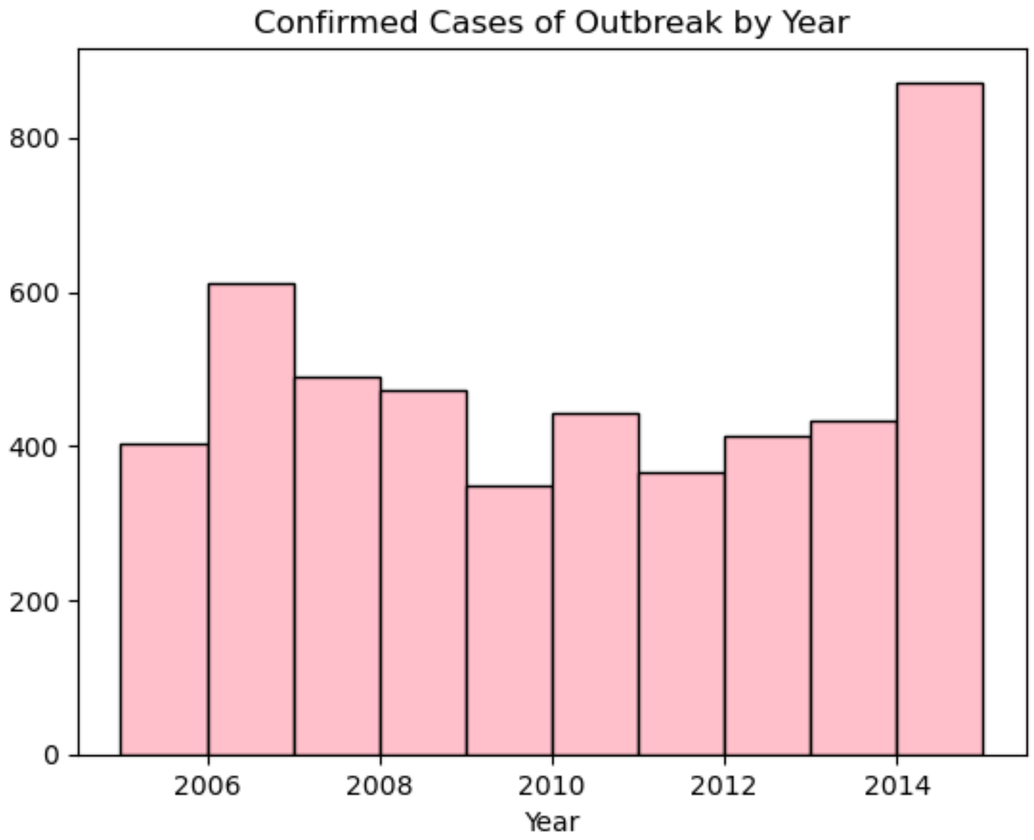
- Mean: 404.25
- Mode: [311 330 332 349 380 415 426 427 428 480 481 492]
- Spread: 181
- Tail (Min): 492
- Tail (Max): 181
- No outliers detected

HISTOGRAM OF VARIABLES - **STATE**



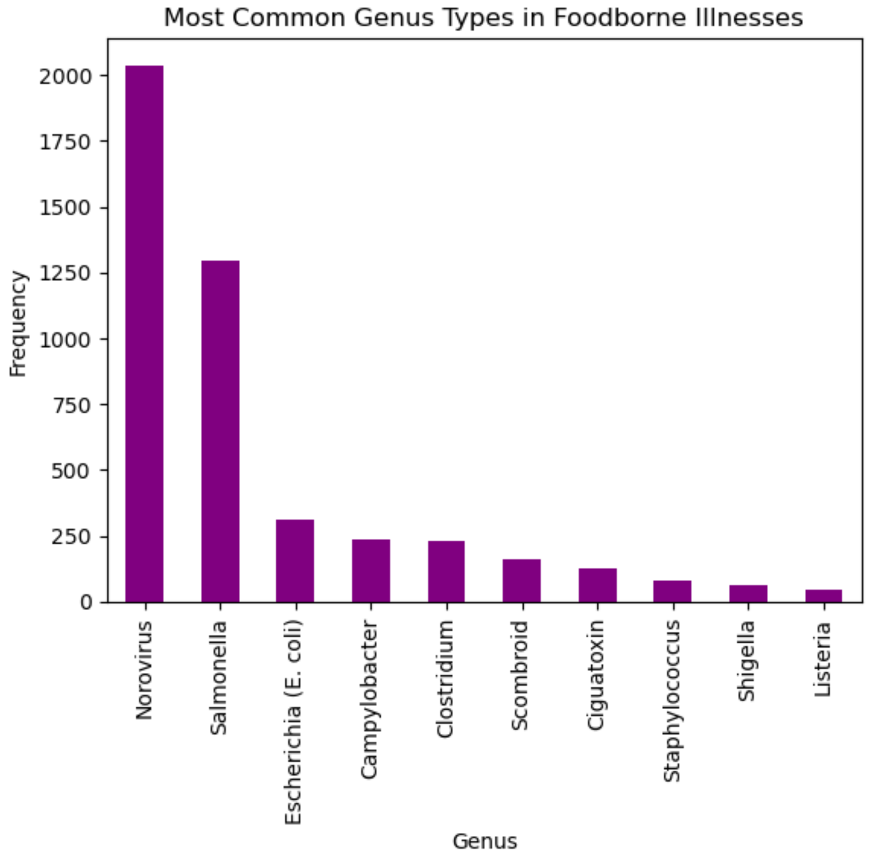
- Mean: 306
- Mode: [233 238 284 311 381 389]
- Spread: 156
- Tail (Min): 233
- Tail (Max): 389
- No outliers detected

HISTOGRAM OF VARIABLES - **YEAR**



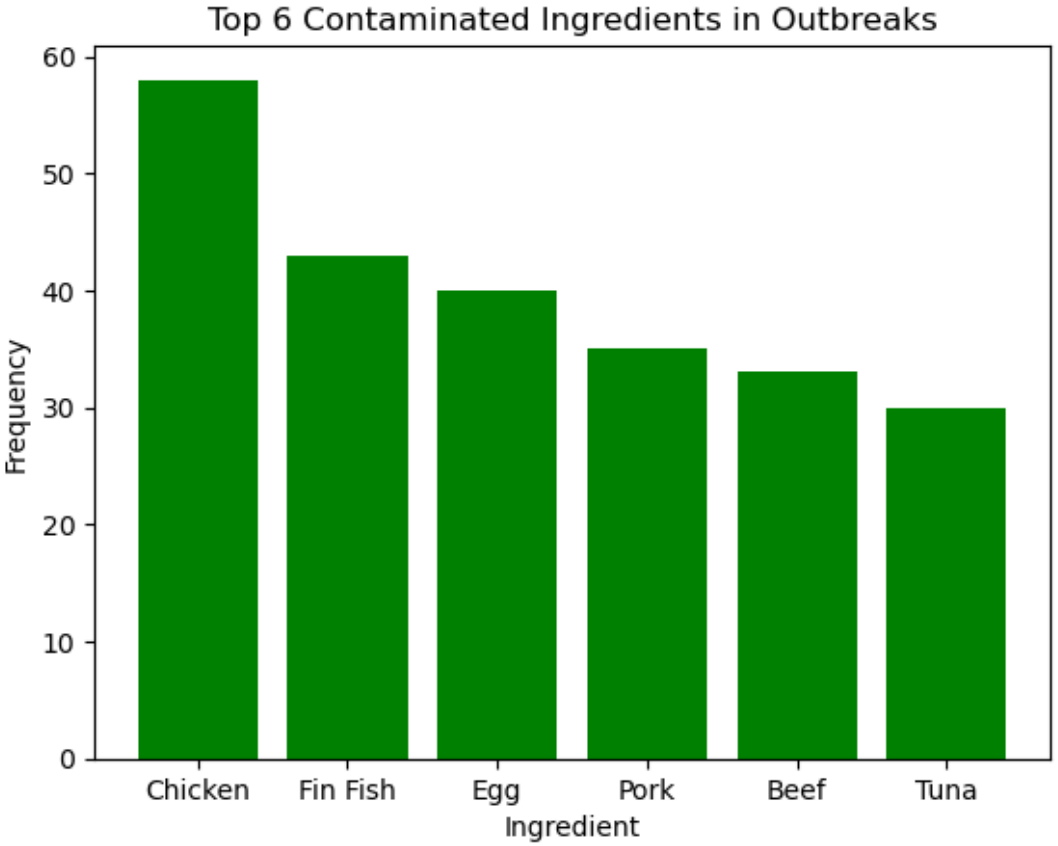
- Mean: 2009
- Mode: 2006
- Spread: 10
- Tail (Min): 2005
- Tail (Max): 2015
- No outliers detected

HISTOGRAM OF VARIABLES - **GENUS**



- Mean: 457.9
- Mode: [47 62 80 125 163 230 234 309 1293 2036]
- Spread: 1989
- Tail (Min): n/a
- Tail (Max): n/a
  - Tail not computable but extends to the right in the graph
- No outliers detected

HISTOGRAM OF VARIABLES - **INGREDIENT**

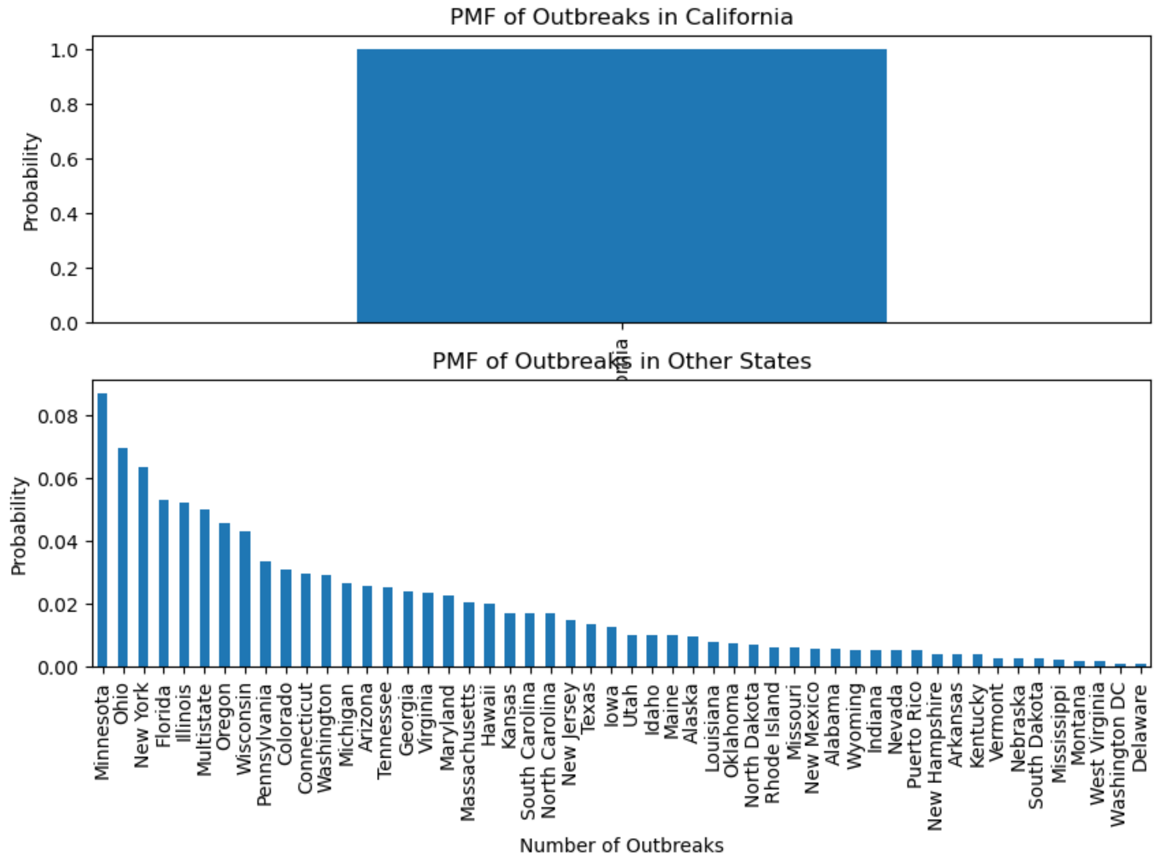


- Mean: 39.83
- Mode: [30 33 35 40 43 58]
- Spread: 28
- Tail (Min): 30
- Tail (Max): 58
- No outliers detected

PROBABILITY MASS FUNCTION (PMF)

The PMFs represent the distribution of the number of outbreaks in 2 different scenarios:

Scenario 1: California
Scenario 2: Other States

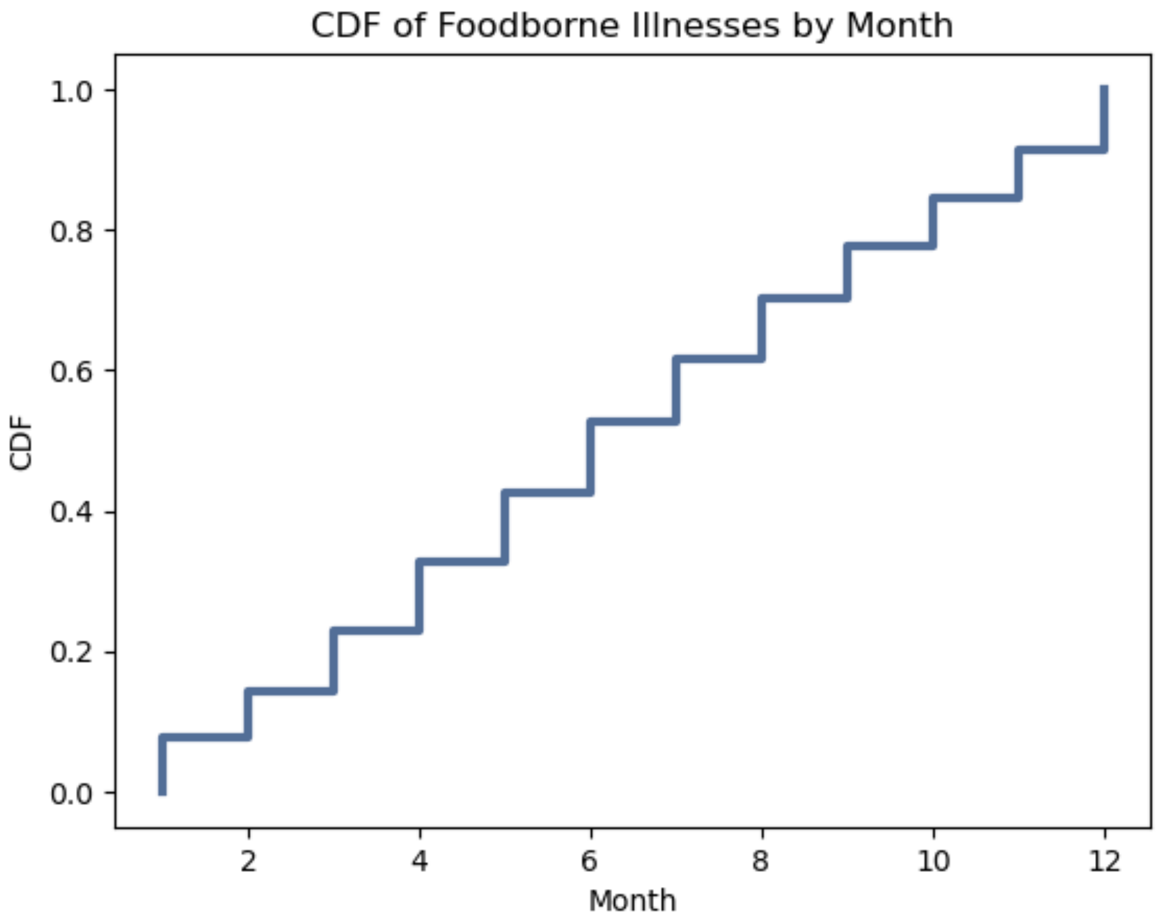


CUMULATIVE DISTRIBUTION FUNCTION (CDF)

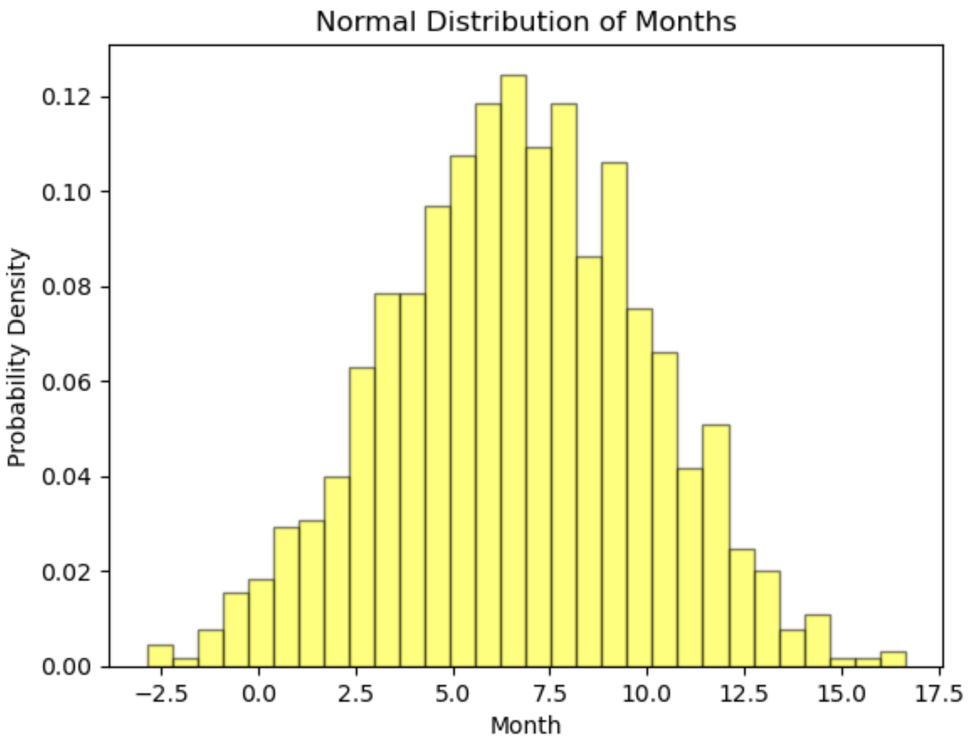
Slightly steep curve during the spring & summer months

Indicates high prevalence of foodborne illnesses during those seasons.

Overall, minimal changes through the months, indicating a consistent prevalence of foodborne illnesses throughout the year.

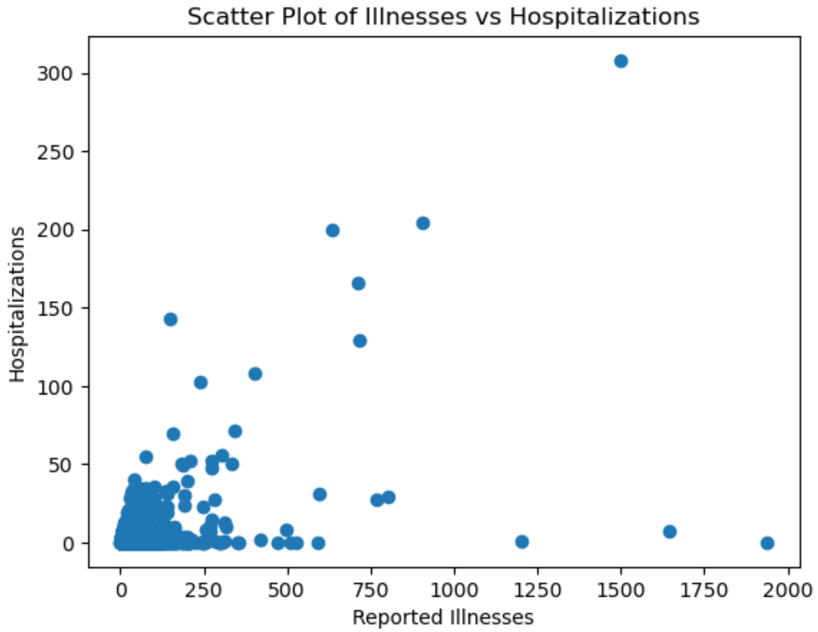


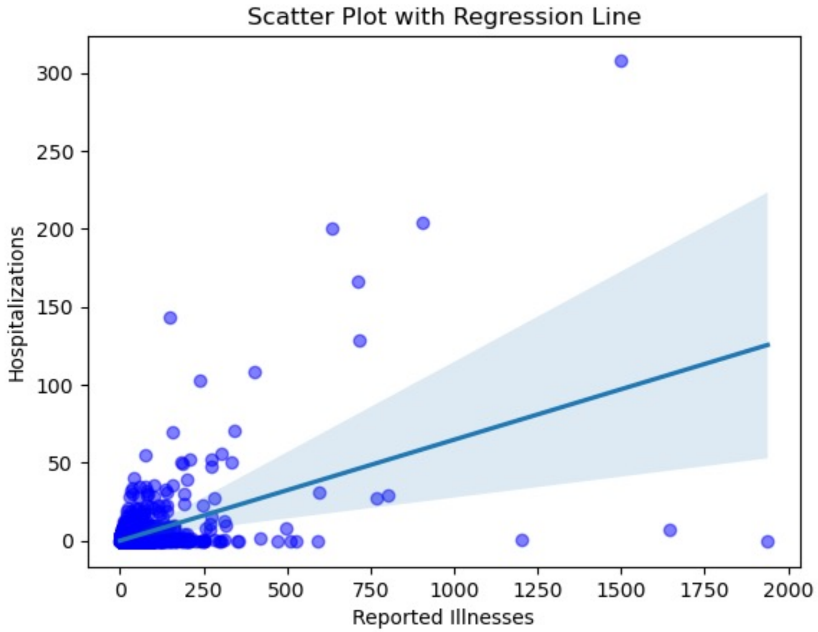
ANALYTICAL DISTRIBUTION



- *The normal distribution peaks between 5.0-7.5 which corresponds to the following months, respectively: May, June, July, mid-August.*
- *Peaking in the summer months = high probability of outbreaks to occur.*
- *Although, this distribution is evenly spread, indicating a relatively consistent occurrence of outbreaks throughout the year.*

SCATTER PLOTS





- Correlation coefficient: 0.50
  - Covariance: 275.72
- Pearson's correlation: 0.50

→ *strong, positive linear relationship between the "Illnesses" and "Hospitalizations"*

REGRESSION ANALYSIS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Illnesses   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     6.175
Date:                Sat, 03 Jun 2023   Prob (F-statistic):             0.0130
Time:                        16:34:42   Log-Likelihood:                -27149.
No. Observations:                4851   AIC:                         5.430e+04
Df Residuals:                    4849   BIC:                         5.432e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         29.4794      2.031     14.513      0.000      25.497      33.462
Month         -0.6977      0.281     -2.485      0.013      -1.248      -0.147
==============================================================================
Omnibus:                     9373.686   Durbin-Watson:                   1.203
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         22317626.950
Skew:                          15.000   Prob(JB):                         0.00
Kurtosis:                     333.930   Cond. No.                         15.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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.

CITATIONS

- Centers For Disease Control and Prevention. (2017, February 15). *Foodborne Disease Outbreaks, 1998-2015*. Kaggle. https://www.kaggle.com/datasets/cdc/foodborne-diseases?resource=download
- Downey, A. B. (2014). *Think stats: Exploratory data analysis in python* (Version 2.2). Green Tea Press.
- Mocomi Kids (Ed.). (2020, July 29). *4 seasons of the Year - gifographic for kids: Mocomi*. Mocomi Kids. https://mocomi.com/seasons-of-the-year/