CSCI225 Final Report

Code

Amit Yadav

Devon Eaton

Kennedy Semple

2020-12-04

The Location and Fatality Rates of Foodborne Illnesses in the United States 1998-2015

Abstract

This report has the intentions to summarize the work that has been accomplished on this project. Through this project we have found a better understanding of how to deal with missing values and numeric evidence, along with Several plots, as well as an information map, and a heatmap have been created with the dataset using the R program. We transformed dataset to get to know about the location, most dangerous species among all mentioned in the dataset, and the rate at which patients got hospitalized after getting ill due to food poisoning. We have generated a model to show a relationship between the number of patients hospitalized, and the number of people got sick.

Introduction

Every year in the United States, there are over 48 million people who get sick from a foodborne disease. The illnesses vary regarding the seriousness of the symptoms. Some may have upset stomach, and more flu like symptoms. Others will unfortunately struggle with their symptoms, having underlying health issues, or belonging to a more susceptible age group. These factors contribute to the data and the motivation behind our project, that people die each year simply from eating food that is contaminated or develops a foodborne disease through its process from farm to your fork. It became very interesting to discover what areas in the United States that foodborne illnesses were the biggest issue in, and what illnesses had the highest fatality rates involved. A Foodborne illness is a very broad term used for any disease that come from the mishandling, storage, or issues cooking with the food, such as meat not being cooked all the way. This means there are many types of illnesses, symptoms, and places that they occur.

Dataset

Our dataset is from Kaggle, who were provided the dataset by the Center for Disease Control and Prevention. Logistics wise the dataset has 19,119 observations, which were made on 12 variables. The variables are Year, Month, State, Location, Food, Ingredients, Species, Genotype, Status, Illnesses, Hospitalizations, and Fatalities, and variables that we are looking closer into are:

1. Location – Where did the illness occur? Or, where the food came from. (for example: restaurant, grocery store, catered, nursing home, university cafeteria)
2. The species of the illness – (for example: Salmonella, Norovirus, Hepatitis A)
3. The numeric values for the illnesses caused, the hospitalizations that then occurred, and the fatalities that then resulted from contracting an illness.
4. Missing Values in several categories.

CDC has conducted surveillance for foodborne disease outbreaks in the United States through the Foodborne Disease Outbreak Surveillance System (FDOSS) since 1973. Initially, FDOSS was paper-based surveillance system but in 1998 it became electronic. these outbreaks are identified by PulseNet, the national molecular subtyping network. The information collected in the outbreak report includes etiology (causative bacteria, chemicals or toxins, viruses, and/or parasites), month, year, and state in which the outbreak occurred, number of illnesses, hospitalizations, deaths that resulted from becoming ill during the outbreak, implicated food(s), locations where food was prepared and consumed, and factors that contributed to the outbreaks. Outbreaks are classified by year and season based on the illness onset date for the first case. Winter was considered December, January, and February; Spring was March to May; Summer June to August; and Fall September to November. Below is a sample of our dataset

Year <dbl>	Month <chr>	State <chr>	Location <chr>
1998	January	California	Restaurant
1998	January	California	*NA*
1998	January	California	Restaurant
1998	January	California	Restaurant
1998	January	California	Private Home/Residence
1998	January	California	Restaurant
1998	January	California	Restaurant
1998	January	California	Restaurant
1998	January	Colorado	Restaurant
1998	January	Colorado	Restaurant

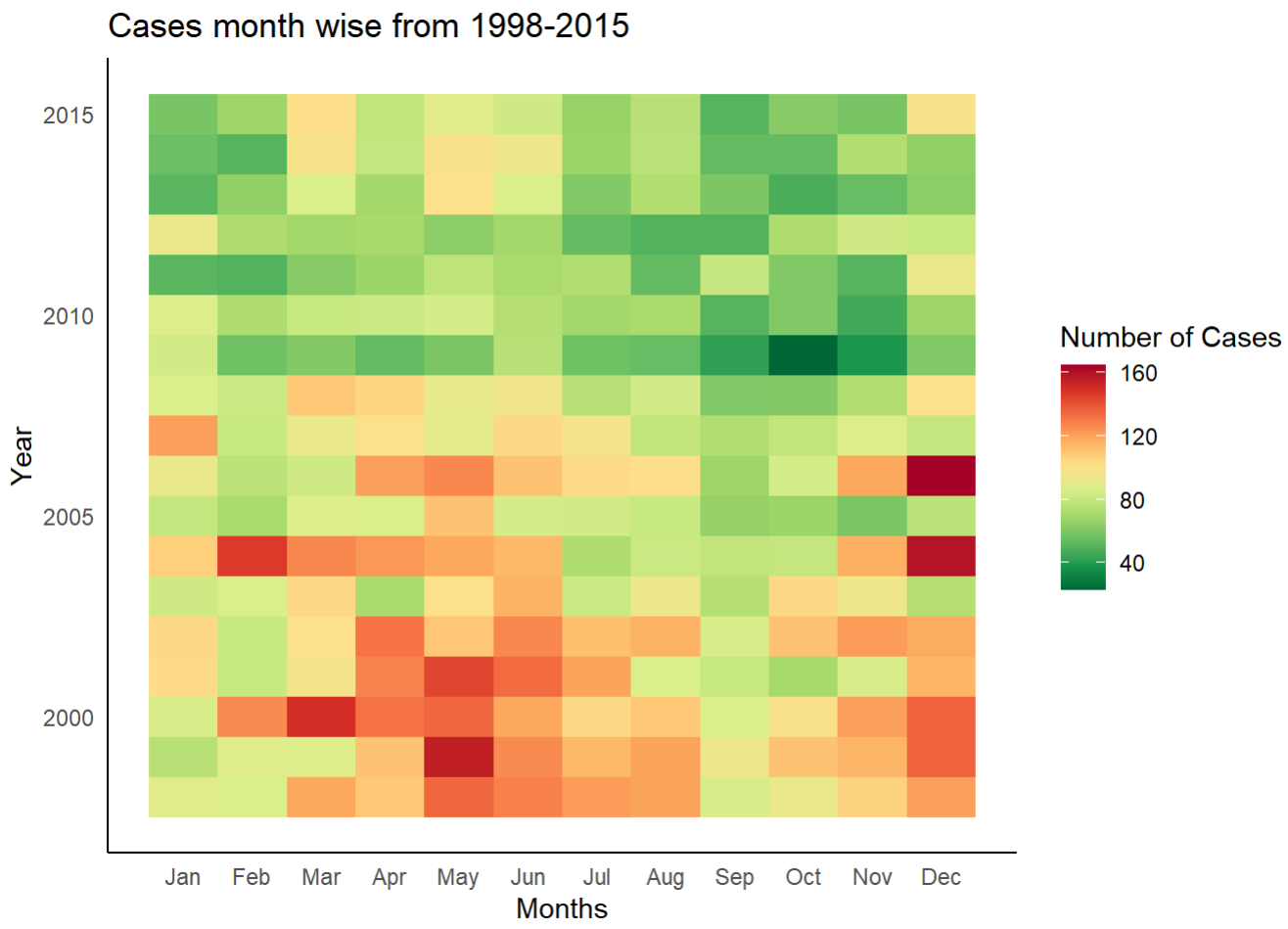
1-10 of 10,000 rows | 1-4 of 12 columns Previous **1** 2 3 4 5 6 … 1000 Next

Methods and Materials

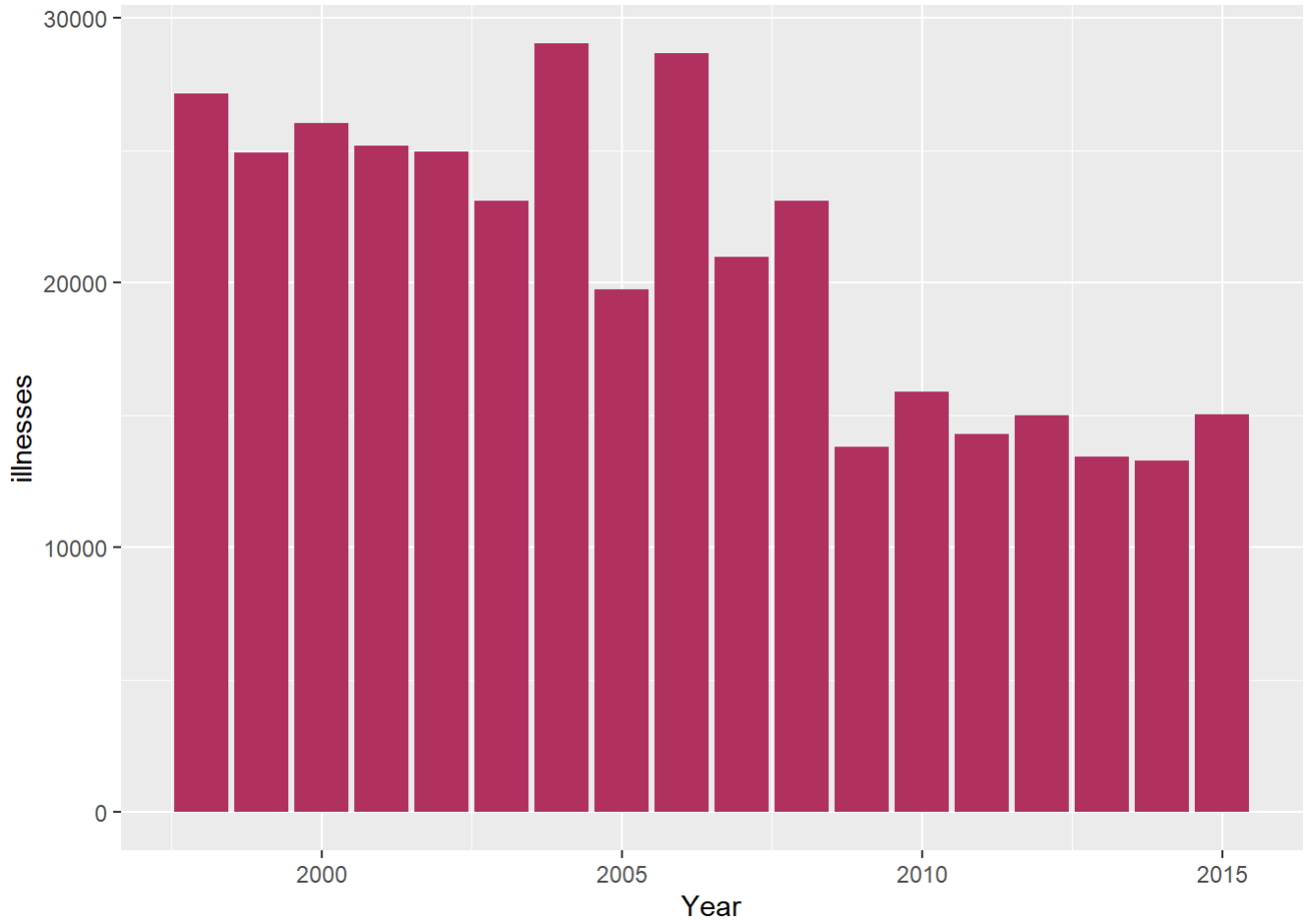
Through the course, and the required text (R for Data Science by Wickham & Grolemund) the basics of the R language were obtained. Here we learned the importance of the Data Science model, in importing, visualizing, transforming, and tidying our data. The newest version of RStudio was used. The data set itself was sourced off of Kaggle. The data was provided to Kaggle through the Centers for Disease Control and Prevention. Important packages installed into RStudio that allowed the completion of the project include tidyverse, dplyr, ggplot2, usmap, and R Markdown to name a few. Using these R packages allowed us to take a deeper look into some variables (like location, illness, fatality) and help visualize our problems. Some different visualization techniques used were scatterplots, histograms, heatmaps, and even the US map.

Results

The months from April to August were more prone to foodborne disease outbreaks, with the period from November to December being less prone to outbreaks. It was interesting to find that the number of outbreaks declined so much after 2007, as there is barely any orange or yellow above that point. It is clearly seen in the heatmap, which has a scale from red to yellow to green, where red shows the highest number of cases and green shows the cases blow 40.

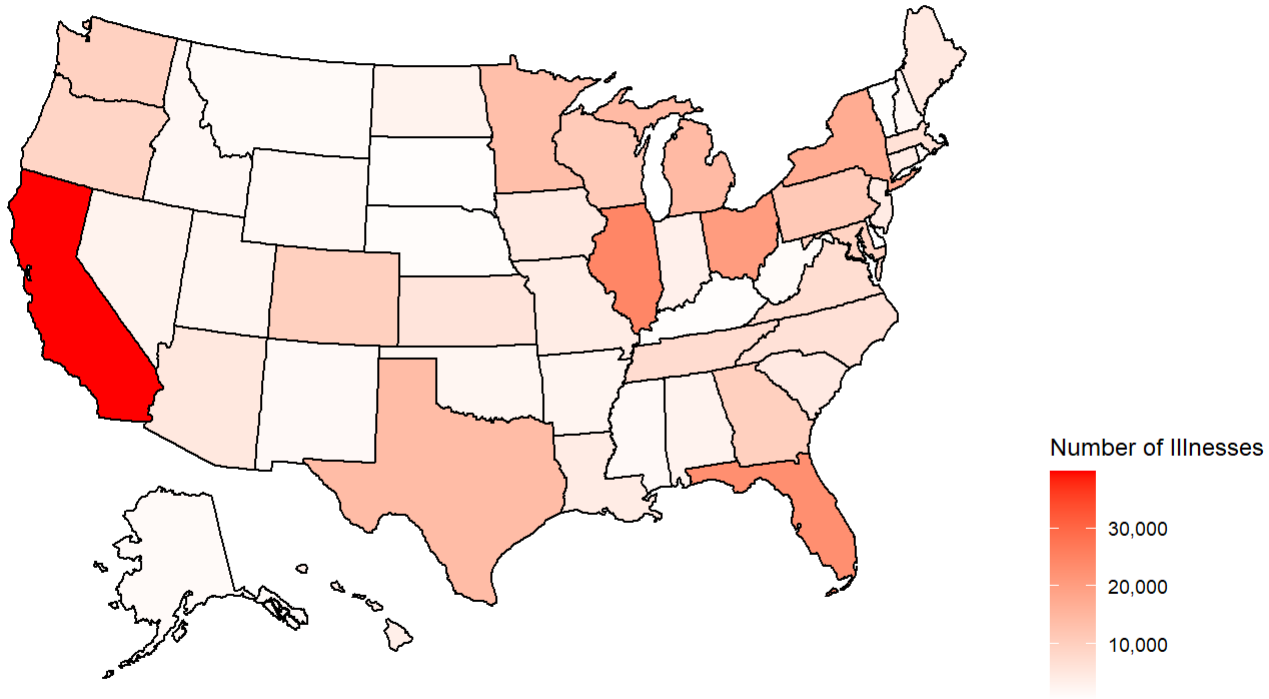


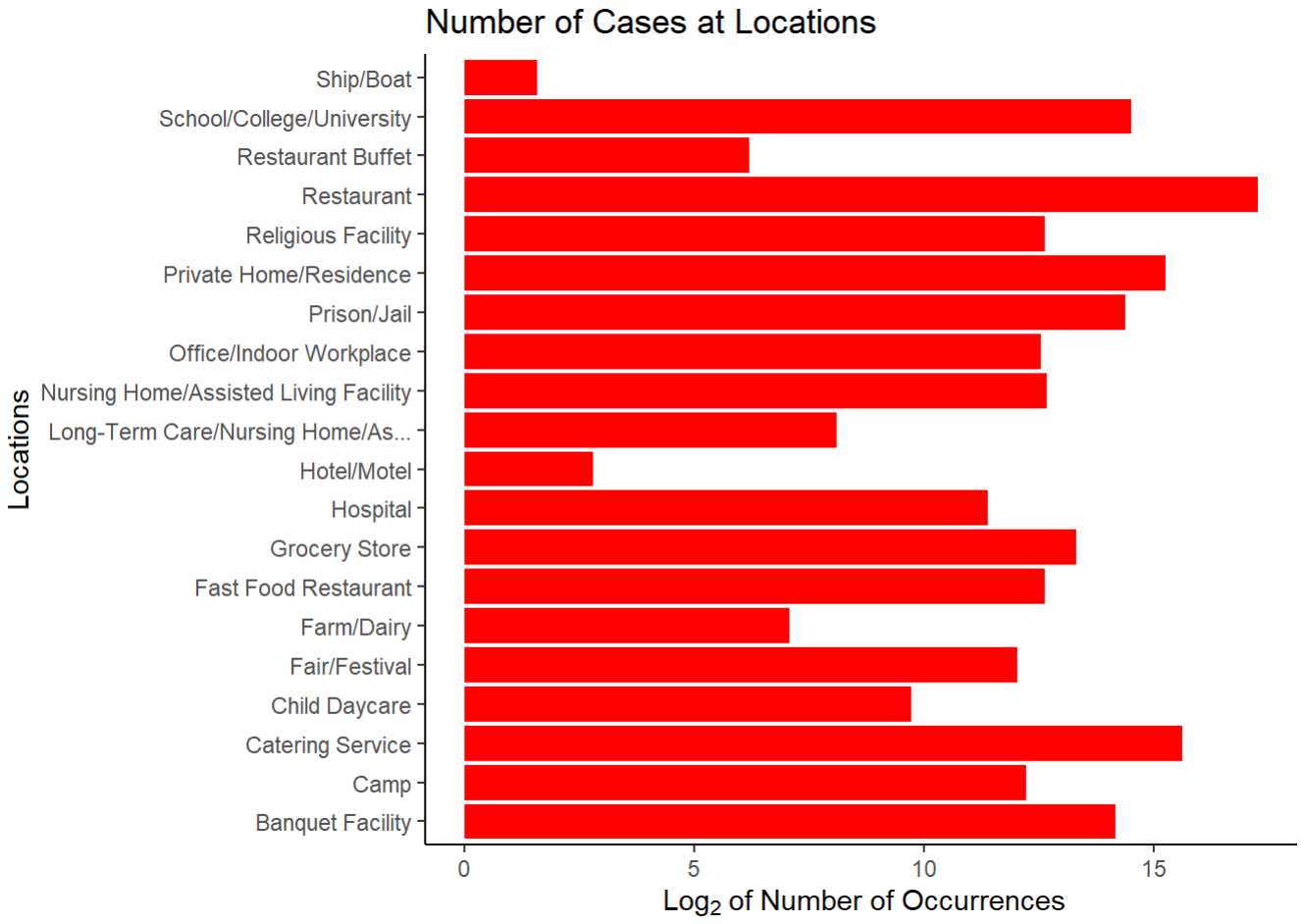
In the bar plot of the number of illnesses each year, it is evident that years 2004 and 2006 had the close to 30000 cases, but after 2006 numbers had dropped significantly around 15000. However, the most important thing while dealing with this dataset is the location where disease outbreak happened.



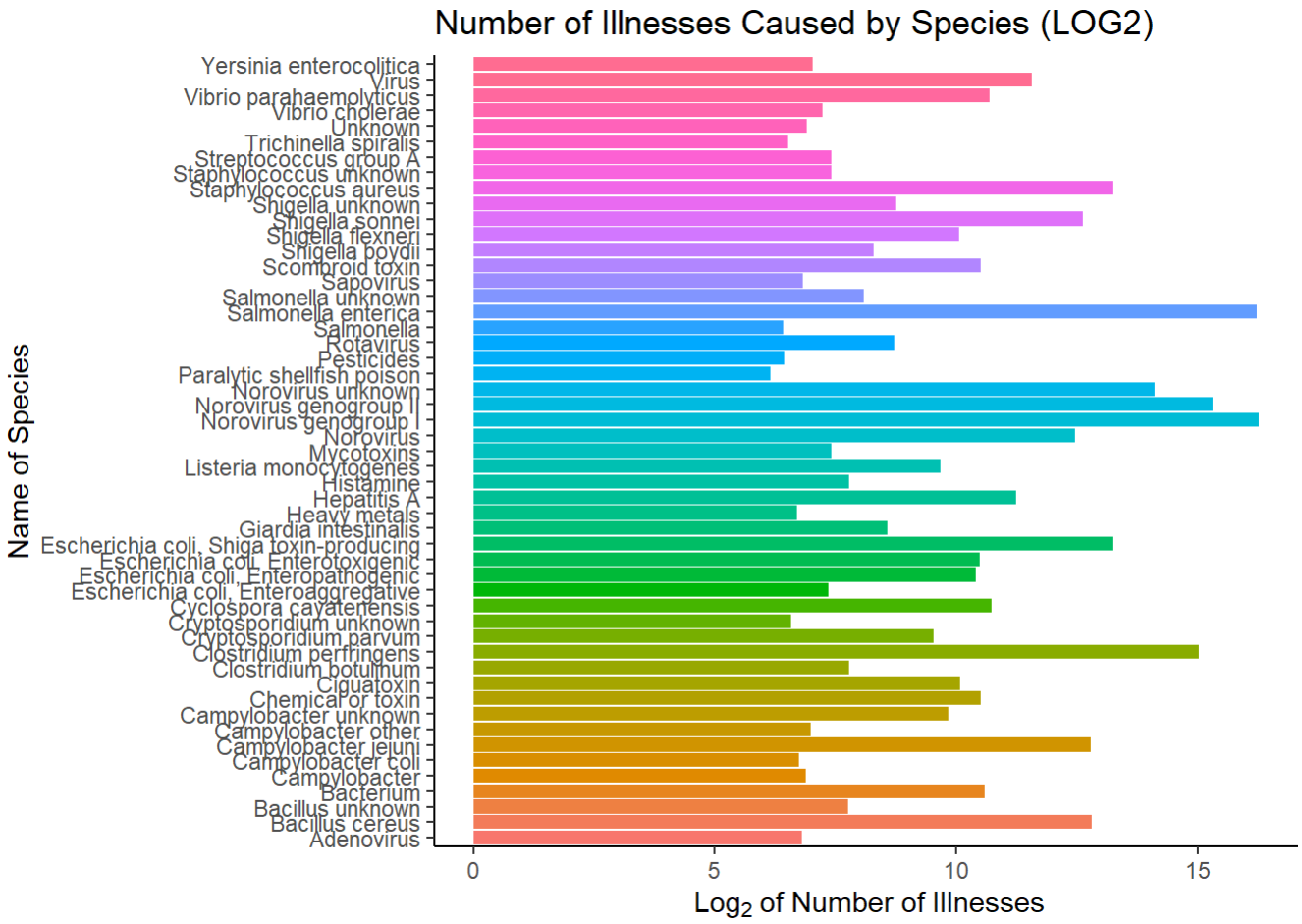
Geographically, US is huge, so it is very important to know which states were severely affected by outbreaks. California had the highest number of cases at over 30000 cases, whereas Illinois, Florida, and Ohio had around 20 000 cases. In the remaining states illnesses reported were not as high. Restaurants, schools, colleges, jail, and residences were the locations that had been the main epicenter of these outbreaks. We developed a bar plot of the locations and number of incidents which occurred at each location. As the occurrences were so varied, we took a $\log_2$ transformation of the number of occurrences to get a better picture. From the plot, it is easiest to compare the number of cases between locations, rather than looking at the actual number of cases at each location. We can see that Restaurants have the highest number of cases, while schools and prisons have almost the same number.

Number of Illnesses per State





After transforming the data, we found out that Norovirus genogroup I had been responsible for 78096 cases, the most illnesses out of all the cases recorded, and Salmonella enterica had the second most cases recorded at 76122.



Hide

```
df.species_number %>%
  arrange(desc(ill_species))
```

Species <chr>	ill_species <dbl>
Norovirus genogroup I	78096
Salmonella enterica	76122
Norovirus genogroup II	40039
Clostridium perfringens	33283
Norovirus unknown	17564
Escherichia coli, Shiga toxin-producing	9754
Staphylococcus aureus	9649
Bacillus cereus	7143

Species <chr>	ill_species <dbl>
Campylobacter jejuni	6997
Shigella sonnei	6260

Even though Norovirus genogroup I had the highest illnesses recorded, salmonella enterica was responsible for the most hospitalizations, which we did not expect.

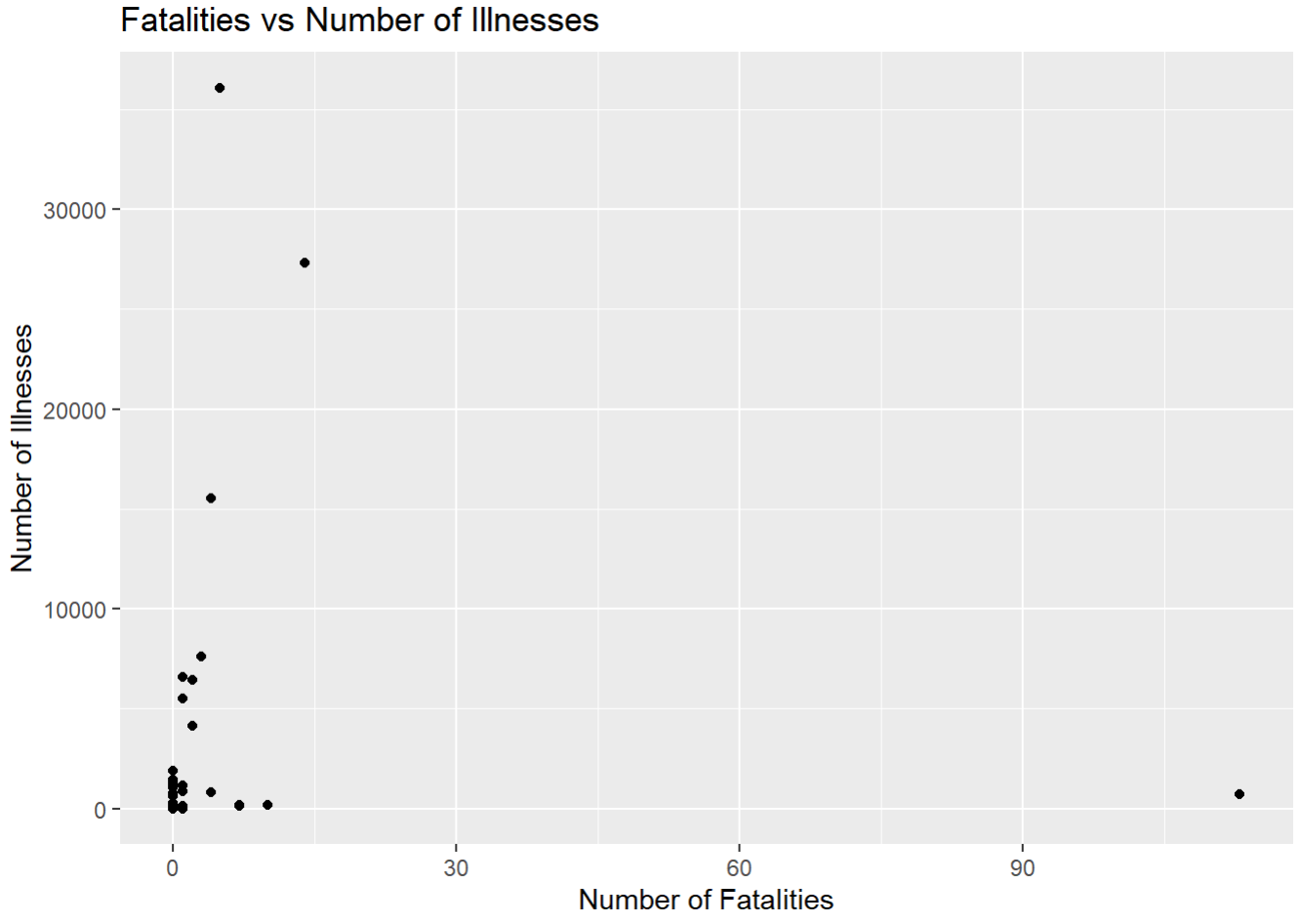
Hide

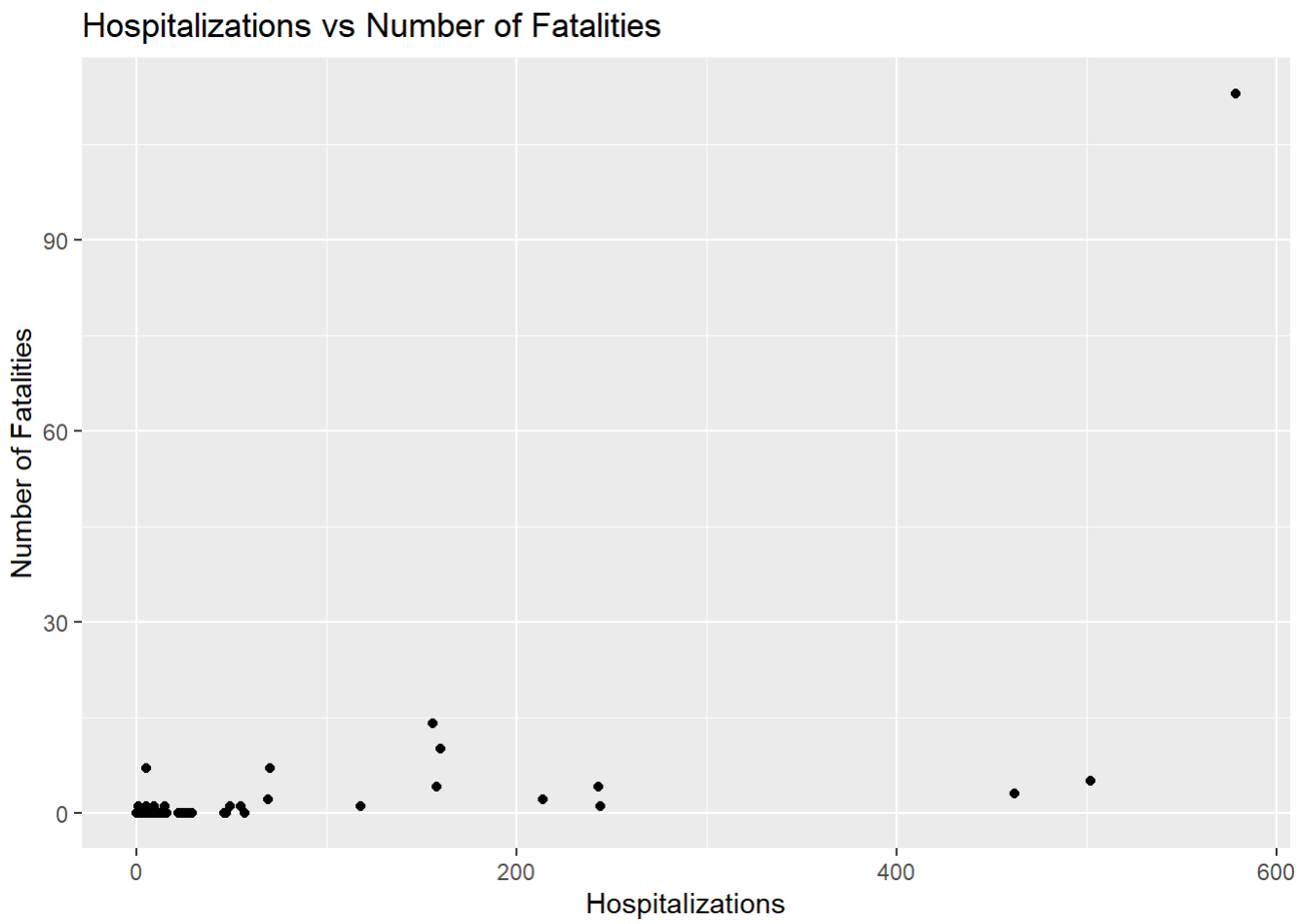
```
separate_rows(df, Species, sep = "; ") %>%
  filter(Species!="NA", Hospitalizations!="NA") %>%
  group_by(Species) %>%
  summarise(hosp_num = sum(Hospitalizations), .groups = "drop") %>%
  arrange(desc(hosp_num))
```

Species <chr>	hosp_num <dbl>
Salmonella enterica	8710
Escherichia coli, Shiga toxin-producing	1992
Norovirus genogroup I	689
Listeria monocytogenes	578
Norovirus genogroup II	534
Staphylococcus aureus	499
Campylobacter jejuni	252
Hepatitis A	244
Shigella sonnei	226
Clostridium botulinum	166

We generated some plots to show the relationships between illnesses, hospitalizations, and fatalities with various scatter plots.

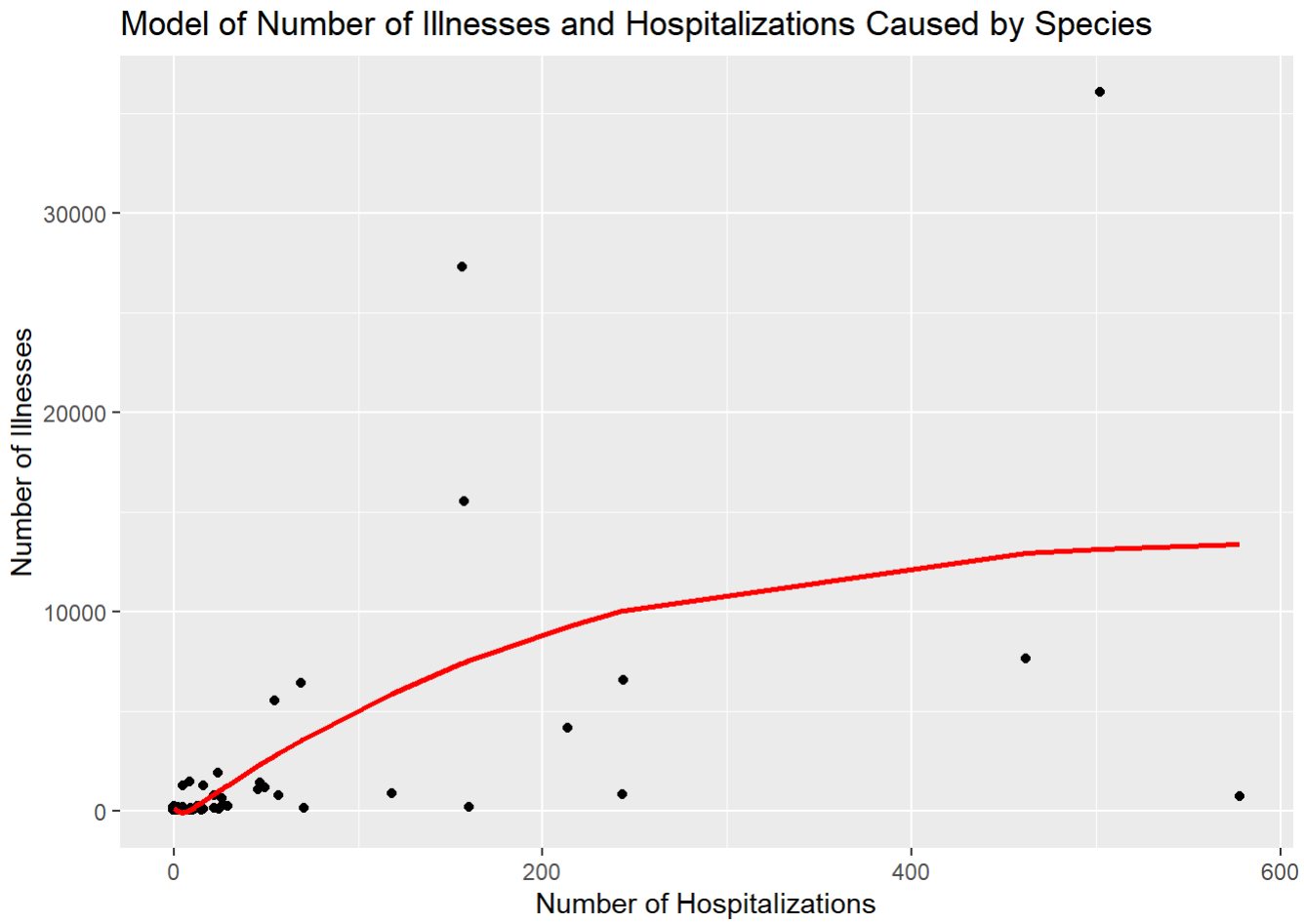

Hospitalizations vs Number of Illnesses
Number of Illnesses
30000
20000
10000
0
0
200
400
600
Number ofHospitalizations


Fatalities vs Number of Illnesses
Number of Illnesses
30000
20000
10000
0
0
30
60
90
Number of Fatalities

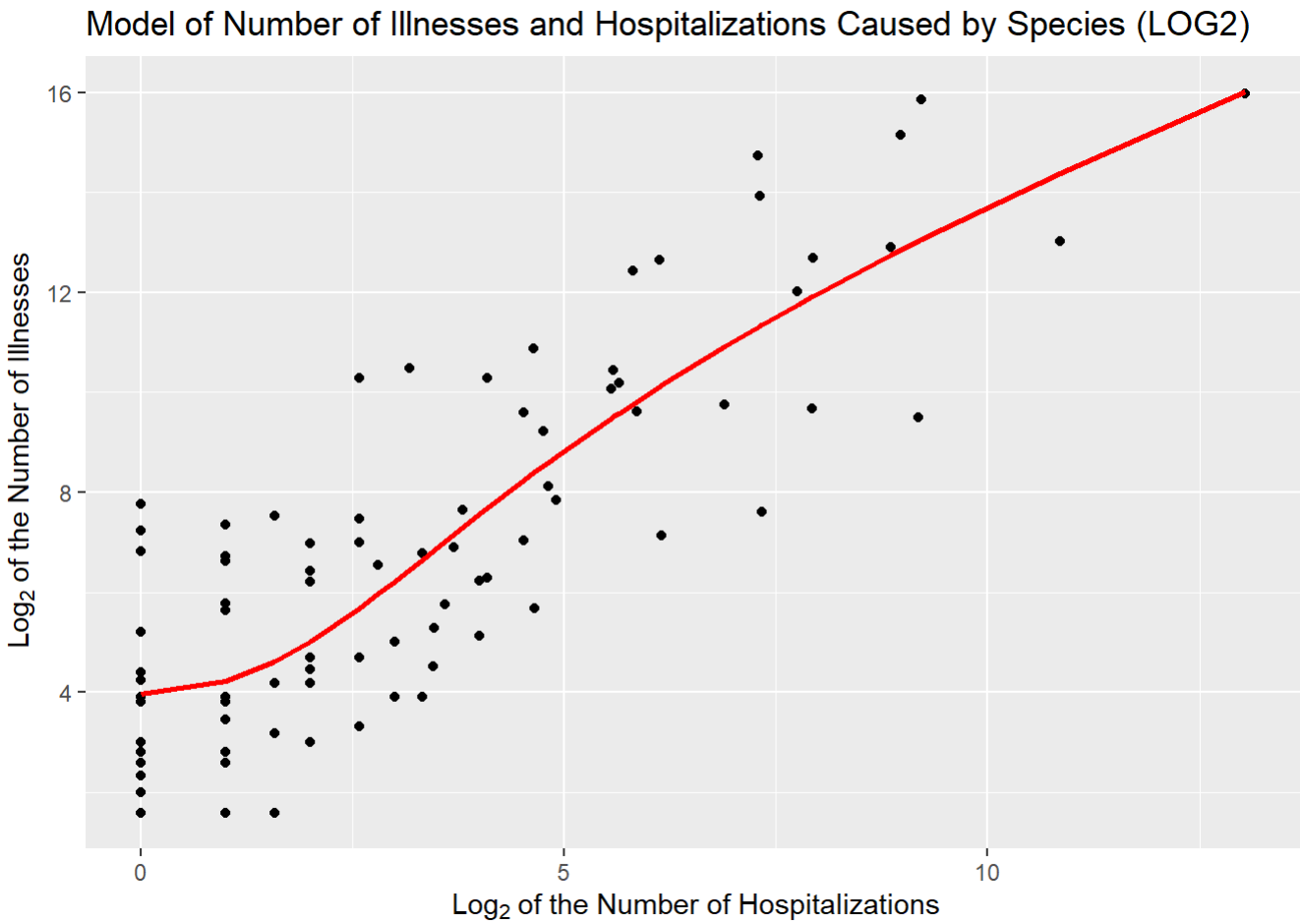


We can see that the plots including fatalities are less scattered than the points in the Number Illnesses vs Hospitalizations plot, which lead to us using that plot to generate a linear model.

The problem in this plot is that most of the points are clustered near origin and very few points are after 200 mark on the x-axis. This lead to us changing to a polynomial model rather than a linear one, which would fit the data better.



In this plot also points are clustered near the origin, but the slope of the line is different. Slope starts decreasing after ~250 hospitalizations cases and decreases further after ~460 cases. The model line is covering far more points than the model with the straight line. In order to get a better picture of the variation of the model within the clustered points, we applied a $\log_2$ transformation to both axes. The generated model appears much smoother and covered almost all scattered points.



Discussion and Conclusion

The main obstacle that we faced while dealing with the dataset was the missing values. To deal with the missing values we used the dplyr package to transform dataset which not only helped us to generate various plots and models, but also helped to reach logical conclusions. We found out that the California had the highest number of illnesses. Question arises what is the main reason behind it? The main reason is that California is the most populous state with a population of 39.5 million[1] and most of the places where outbreaks occurred were common places like schools, colleges, residences, and restaurants where people go daily which lead one of our conclusions being the higher the population, the higher the chances of outbreaks.

Once the species responsible is known, it becomes easier to stop the spread of cases. We found out that Norovirus genogroup I caused the highest number of illnesses, but was not the deadliest. The species causing the most hospitalizations was Salmonella enterica, even though it caused mild symptoms. According to NCBI (National Center for Biotechnology Information)[2] a salmonella outbreak in the US from 2006-2011 was linked to poultry, eggs, pork, beef, and dairy.

One good thing was the low percentage of fatalities from hospitalized cases, which is likely caused by many hospitalized patients recovering after receiving fluids. We focused more on the illnesses and hospitalizations variables rather than fatalities. We generated a model to get an idea of what the relation between these two variables would be after 2015.

References

1. https://en.wikipedia.org/wiki/California (https://en.wikipedia.org/wiki/California)
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310208/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310208/)
3. Centers for Disease Control and Prevention. (2017, February 15). Foodborne Disease Outbreaks, 1998-2015. Retrieved from https://www.kaggle.com/cdc/foodborne-diseases

(https://www.kaggle.com/cdc/foodborne-diseases)
4. Center for Food Safety and Applied Nutrition. (n.d.). What You Need to Know about Foodborne Illnesses. Retrieved from https://www.fda.gov/food/consumers/what-you-need-know-about-foodborne-illnesses (https://www.fda.gov/food/consumers/what-you-need-know-about-foodborne-illnesses)
5. Data Science Model Image Sourced from R for Data Science Text Grolemund, G., & Wickham, H. (n.d.). R for Data Science. Retrieved from https://r4ds.had.co.nz/ (https://r4ds.had.co.nz/)
6. https://www.who.int/health-topics/foodborne-diseases#tab=tab_1 (https://www.who.int/health-topics/foodborne-diseases#tab=tab_1)