**Foodborne Disease Outbreaks, 1998-2015**

****

**AMOGH MAHESH**

**TANVI GAWADE**

**MS: Information Systems**

**California State University Los Angeles**

**CIS 5270: Business Intelligence**

**Submitted to: Dr. Shilpa Balan**

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
|  | **Topic** | **Pg No** |
| 1 | **Objective of Study** | 3 |
| 2 | **Dataset** | 4 |
| 3 | **Data cleaning** | 5 |
| 4 | **Analysis and Visualizations** | 10 |
| 5 | **Statistics** | 16 |
| 6 | **User Defined Functions** | 18 |
| 7 | **Data cleaning codes** | 24 |
| 8 | **Visualization codes** | 27 |
| 9 | **Statistical/user defined codes** | 28 |
| 10 | **References** | 32 |

**OBJECTIVE OF STUDY**

A foodborne disease outbreak occurs when two or more people get the same illness from the same contaminated food or drink. While most foodborne illnesses are not part of a recognized outbreak, outbreaks provide important information on how germs spread, which foods cause illness, and how to prevent infection. Foodborne diseases due to known pathogens are estimated to cause 9.4 million illnesses each year in the United States. More than 250 pathogens and toxins have been known to cause foodborne illness and almost all of them can cause an outbreak.

CDC conducts surveillance for foodborne disease outbreaks in the United States through the Foodborne Disease Outbreak Surveillance System. Public health agencies in all 50 states, the District of Columbia, U.S. territories, and Freely Associated States have primary responsibility for identifying and investigating outbreaks and use a standard form to report outbreaks voluntarily to CDC. Public health agencies submit reports of outbreaks investigated by their agencies using a web-based reporting platform, the electronic Foodborne Outbreak Reporting System (eFORS).

Although relatively few of these illnesses occur in the setting of a recognized outbreak, data collected during outbreak investigations provide insight into the pathogens and foods that cause illness. Public health officials, regulatory agencies, and the food industry can use these data to inform efforts to prevent foodborne disease

**DATASET**

**Data set URLs:**

<https://www.kaggle.com/cdc/foodborne-diseases/downloads/foodborne-diseases.zip/1>

**Data set file format: CSV file**

About the dataset:

This dataset provides data on foodborne disease outbreaks reported to CDC from 1998 through 2015. Data fields include Year, State, Location, Reported Food, Contaminated Ingredient, Species(etiology), Status, Total Illnesses, Hospitalizations and Fatalities. Datasets consists of the Disease outbreaks reported through the years 1998 to 2015, Datasets consists of 12 columns and 19119 rows.

**Year –** Diseases reported through years 1998 to 2015

**Month –** Month the diseases were reported in.

**State –** Diseaseoutbreaks occurring in specific states. More than one state are listed as "multistate"

**Location-** Locationwhere the food was prepared

**Food –** The type of food that lead to the disease outbreak.

**Ingredient –** Contanimated Ingredient.

**Species -** The pathogen, toxin, or chemical that caused the illnesses.

**Serotype/Genotype -** Group of organisms within a species that have the same type and number of surface antigens

**Status-** Whether the etiology was confirmed or suspected

**Illnesses –** Total reported Illnesses.

**Hospitalizations –** Total reported hospitilizations

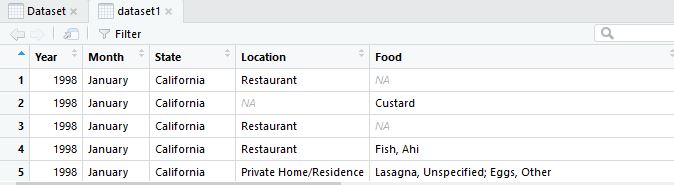
**Fatalities -** Total reported critical condition cases.

**DATA CLEANING**

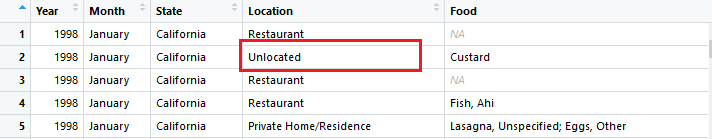
**Data Cleaning 1: Replacing “NA” Values in the Column Location to “Unlocated”**

**Results:** All “NA” values were replaced to “Unlocated”.

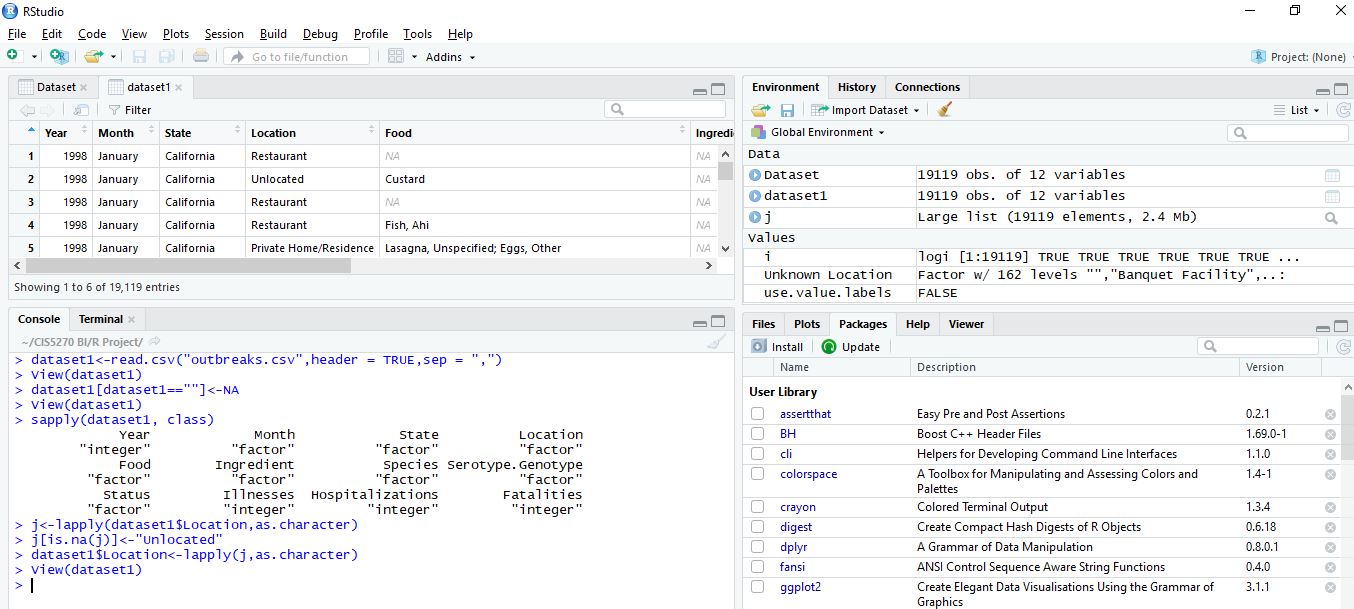
**Before:**

****

**After:**

****

**Code Screenshot:**

****

**Code:**

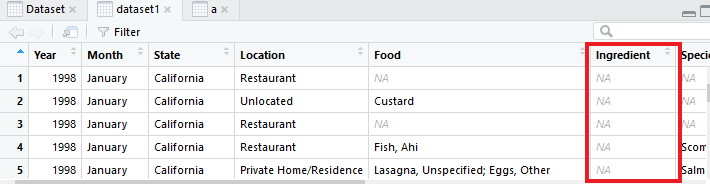
> setwd("~/CIS5270 BI/R Project/")

|  |
| --- |
| > dataset1<-read.csv("outbreaks.csv",header = TRUE,sep = ",")  > View(dataset1)  > dataset1[dataset1==""]<-NA  > View(dataset1)  > sapply(dataset1, class)  Year Month State Location  "integer" "factor" "factor" "factor"  Food Ingredient Species Serotype.Genotype  "factor" "factor" "factor" "factor"  Status Illnesses Hospitalizations Fatalities  "factor" "integer" "integer" "integer"  > j<-lapply(dataset1$Location,as.character)  > j[is.na(j)]<-"Unlocated"  > dataset1$Location<-lapply(j,as.character) |
|  |
| |  | | --- | |  | |

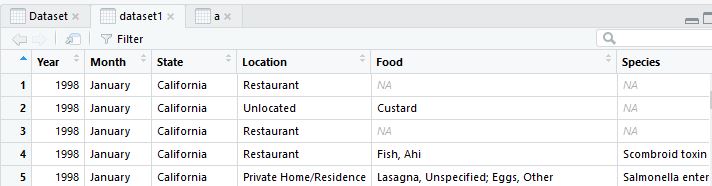
**Data Cleaning 2: Delete the Redundant Column “Ingredient”**

**Results:** Redundant Column Ingredient with all its values are deleted.

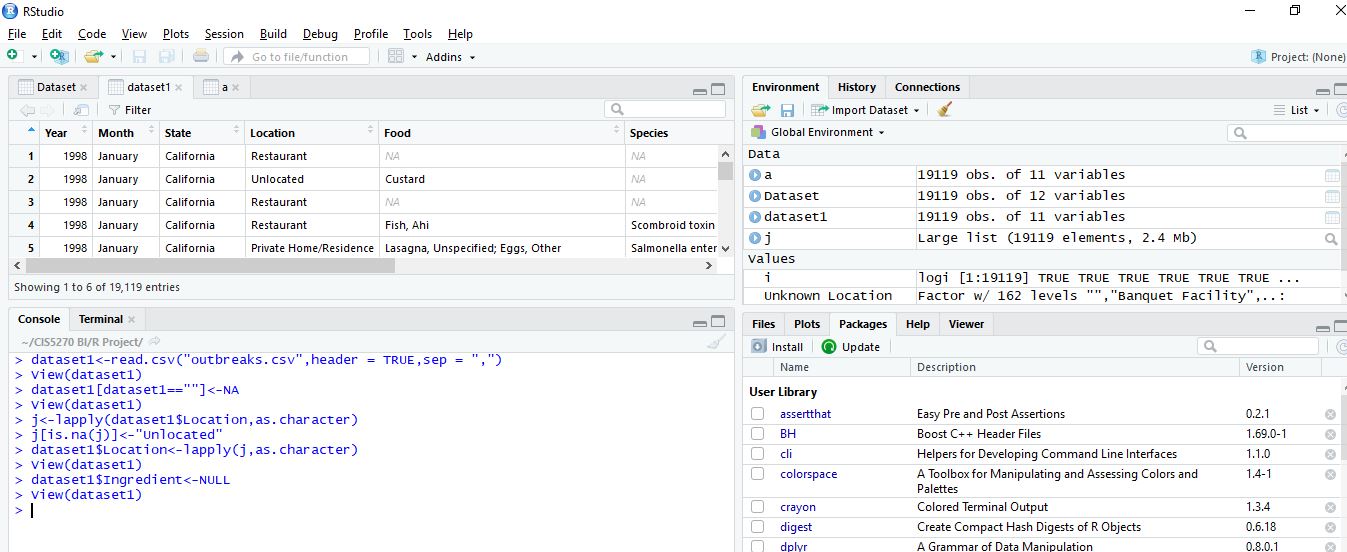
**Before**

****

**After:**

****

**Code Screenshot**

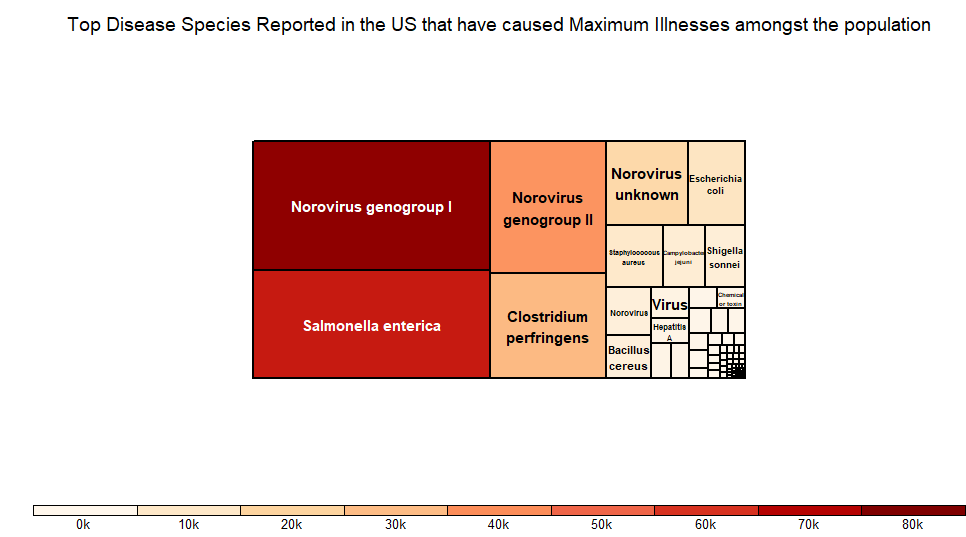
****

**Code:**

> setwd("~/CIS5270 BI/R Project/")

|  |
| --- |
| > dataset1$Ingredient<-NULL  > View(dataset1)  **Data Cleaning 3: Categorizing Values in Column “Food” in a common group.**  **Results:** Values in the column “Food” separated and grouped into Food Categories  **Before:**    **After:**    **Code Screenshot**    **Code:**  > View(Dataset)  > library("tidyr")  Warning message:  package ‘tidyr’ was built under R version 3.5.3  > dataset1<-separate(dataset1,Food,c("Food"),sep=",")  > dataset1<-separate(dataset1,Food,c("Food"),sep=";")  > View(dataset1) |

**ANALYSIS/VISUALIZATIONS**  
**Q.1. What are the Top Diseases Species reported in the United States that have caused maximum Illnesses amongst the population.?**



**Visualization:** Treemap

**Functions Used:** Aggregate**,** Arrange, Colors, Descending, Title

**Packages/library:** Dplyr

**ANALYSIS/DESCRIPTION:**

As per the TreeMap, we can see that the specific diseases species that impacted majority of the population in the United States are represented by the Size of the rectangle and intensity of the color in the decreasing order. With the help of the analysis we understand the top disease species that have caused considerable amount of illnesses amongst the population are NoroVirus Genotype 2 and Salmonella Enterica. It has also been cited by the National Institute of Biotechnology Institute that Noroviruses (NoVs) are the most important cause of non-bacterial gastroenteritis in all ages in both developed and developing countries.1 It has recently been estimated that each year NoVs cause 64,000 episodes of diarrhea requiring hospitalization and 900,000 clinic visits among children in industrialized countries and up to 1.1 million hospitalizations and 200,000 deaths of children < 5 years of age in developing countries.This is a matter of concern for safety issues and awareness should be spread in order to take precautions to avoid certain kinds of food that would cause such diseases from the mentioned species.

**Code:**

> library("dplyr")

>aggregate1\_species\_df<-aggregate(dataset1$Illnesses, by=list(Species=dataset1$Species), FUN=sum)

> x = arrange(aggregate1\_species\_df,desc(aggregate1\_species\_df$x) )

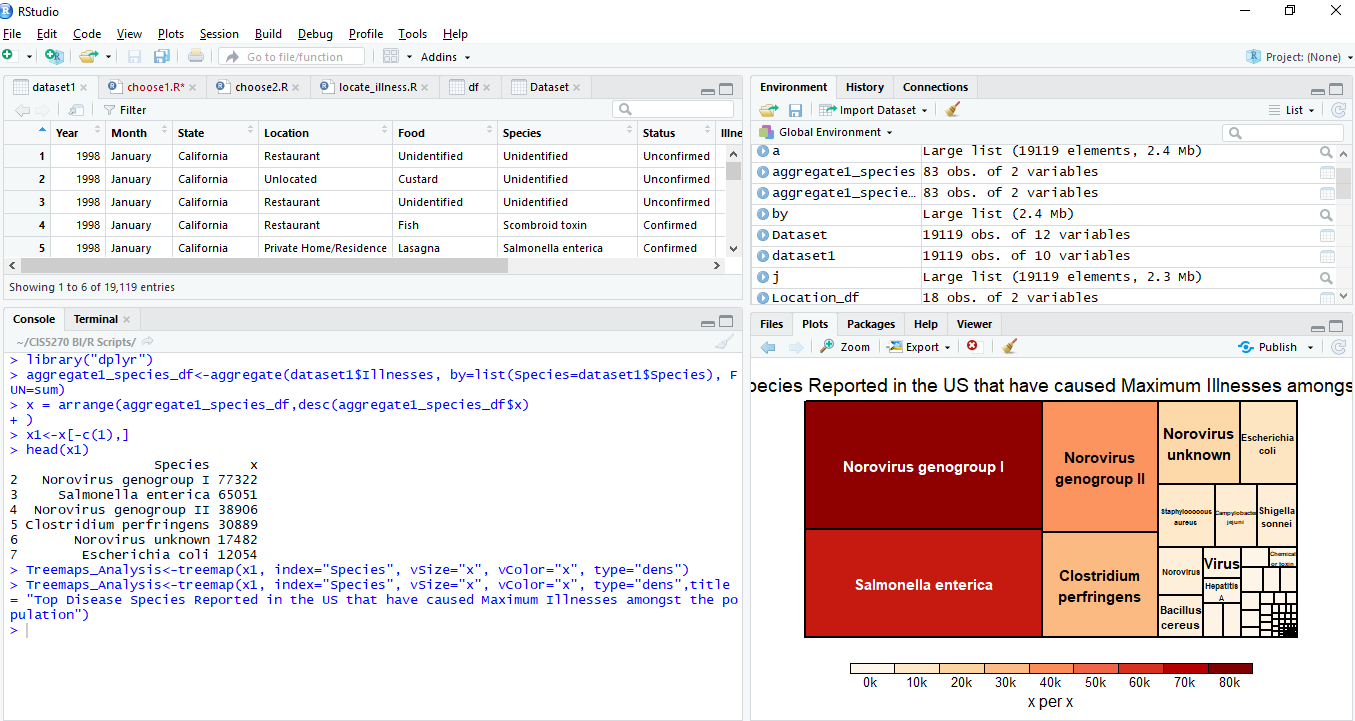
> x1<-x[-c(1),]

> head(x1)

> Treemaps\_Analysis<-treemap(x1, index="Species", vSize="x", vColor="x", type="dens")

> Treemaps\_Analysis<-treemap(x1, index="Species", vSize="x", vColor="x", type="dens",title = "Top Disease Species Reported in the US that have caused Maximum Illnesses amongst the population")

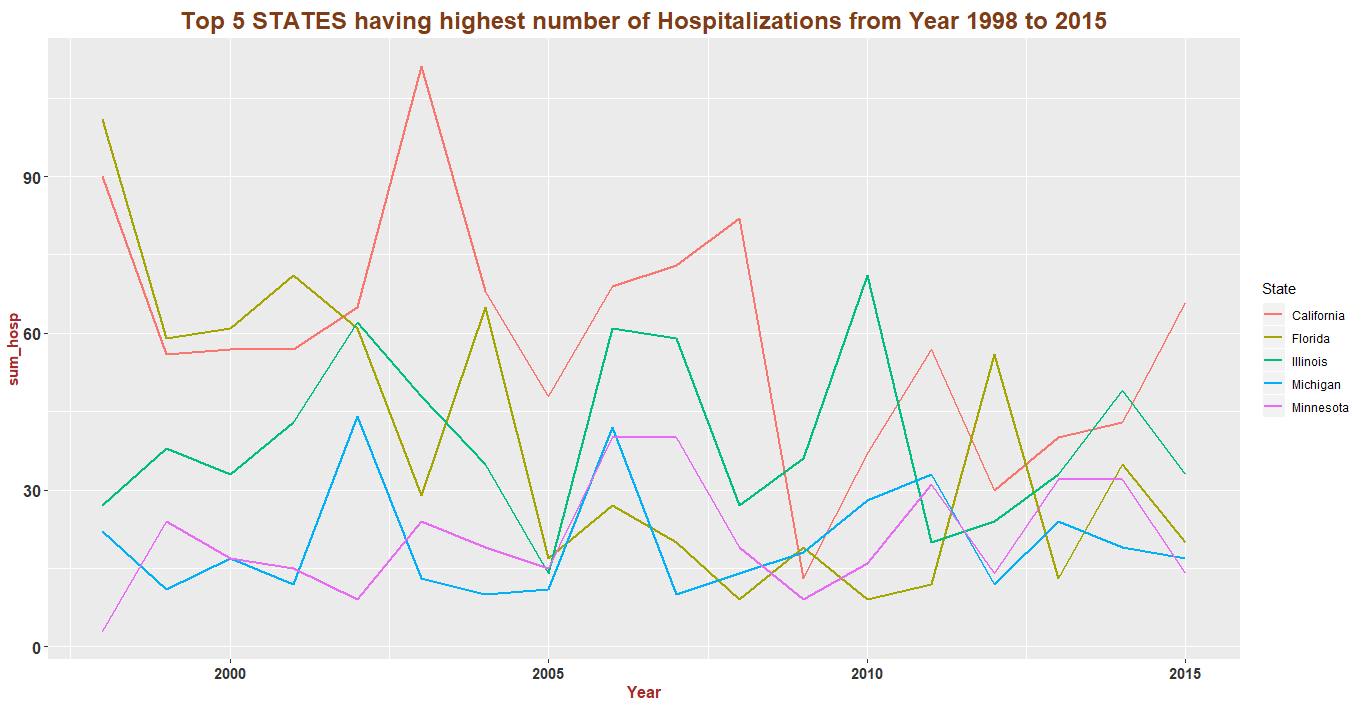
**Screenshot:**



**Complete Code Screenshot:**



**Q.2. Show 5 States having highest number of hospitalizations.**



**Visualizations:** line chart (time series)

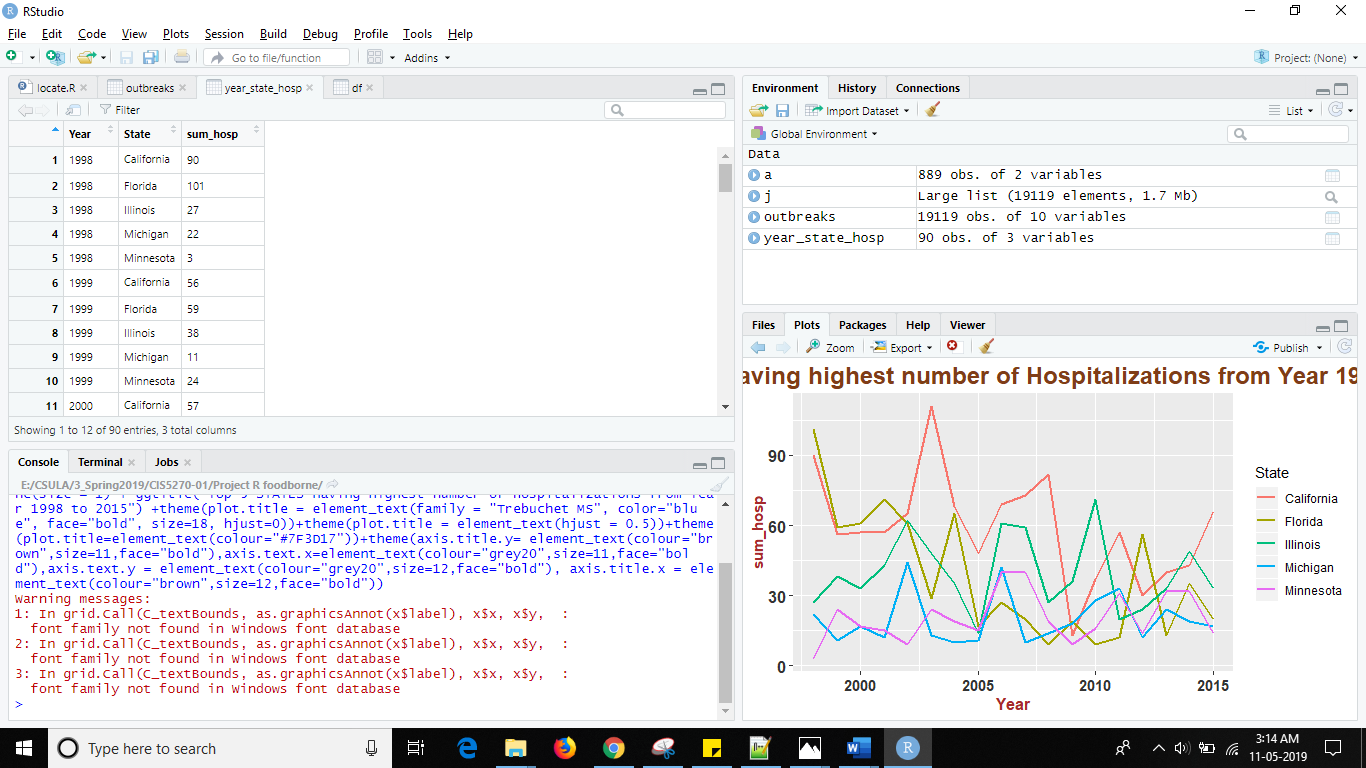
**Functions Used:** labels, ggplot2, geom\_line, geom\_text, theme, color

**Packages/library:** ggplot2

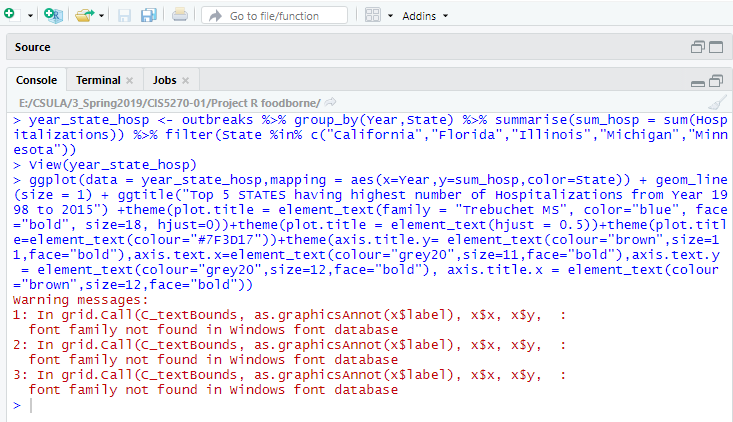
**Analysis/Description:**

The visulizations shows us the number of hospitalizations happened in states like California, Florida, Illinois, Michigan and Minnesota from year 1998 to 2015. We can see that there is lots of variation in lines, but the number of hospitalizations has decreased compared to past years. In 2003, California had highest number of hospitalizations and in 2009 it has the lowest. Hospitalizations increases with increase in number of illness. We can see that illness is caused by the harmful species and can also lead to fatalities.

**Screenshot:**



**Complete Code Screenshot:**



**Code:**

> library("ggplot2")

> library("dplyr")

> year\_state\_hosp <- outbreaks %>% group\_by(Year,State) %>% summarise(sum\_hosp = sum(Hospitalizations)) %>% filter(State %in% c("California","Florida","Illinois","Michigan","Minnesota"))

> View(year\_state\_hosp)

> ggplot(data = year\_state\_hosp,mapping = aes(x=Year,y=sum\_hosp,color=State)) + geom\_line(size = 1) + ggtitle("Top 5 STATES having highest number of Hospitalizations from Year 1998 to 2015") +theme(plot.title = element\_text(family = "Trebuchet MS", color="blue", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title=element\_text(colour="#7F3D17"))+theme(axis.title.y= element\_text(colour="brown",size=11,face="bold"),axis.text.x=element\_text(colour="grey20",size=11,face="bold"),axis.text.y = element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**Statistics**

**Statistical Summary of Dataset:**

> summary(outbreaks)

Year Month State Location

Min. :1998 May :1898 Florida : 2409 Restaurant :10448

1st Qu.:2001 June :1819 California: 2293 Unlocated : 2166

Median :2005 December:1816 Ohio : 1288 Private Home/Residence: 1681

Mean :2006 April :1725 Illinois : 1133 Catering Service : 1089

3rd Qu.:2010 March :1724 New York : 909 Banquet Facility : 367

Max. :2015 July :1538 Michigan : 882 Fast Food Restaurant : 366

(Other) :8599 (Other) :10205 (Other) : 3002

Food Species Status

Unidentified:8963 Unidentified :6619 Confirmed :7909

Chicken : 713 Norovirus genogroup I :2765 Unconfirmed :6619

Sandwich : 551 Salmonella enterica :2390 Suspected :4068

Fish : 531 Norovirus genogroup II :1445 Suspected; Suspected: 310

Ground Beef : 382 Clostridium perfringens: 826 Confirmed; Confirmed: 133

Pizza : 279 Norovirus unknown : 794 Confirmed; Suspected: 32

(Other) :7700 (Other) :4280 (Other) : 48

Illnesses Hospitalizations Fatalities

Min. : 2.00 Min. : 0.0000 Min. : 0.00000

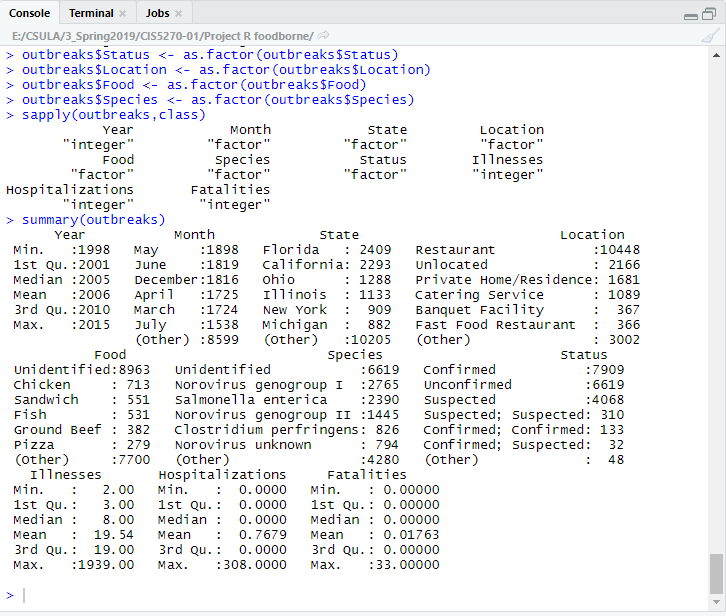
1st Qu.: 3.00 1st Qu.: 0.0000 1st Qu.: 0.00000

Median : 8.00 Median : 0.0000 Median : 0.00000

Mean : 19.54 Mean : 0.7679 Mean : 0.01763

3rd Qu.: 19.00 3rd Qu.: 0.0000 3rd Qu.: 0.00000

Max. :1939.00 Max. :308.0000 Max. :33.00000



**Analysis/Description:**

We observe from the above statistical analysis maximum number of illnesses reported amongst the population in the United States is much higher than the maximums of Hospitalizations and even less for Fatalities. From this analysis we can observe that the illnesses reported die to various species might have been cured/quarantined before reaching to the severe levels of illnesses. We also observe that the 3rd quartile or the upper quartile which consists of the upper 25% of our data. The 3rd quartile consists of the maximum illnesses caused from all states. consists of the maximum reported. Hence due to the advancement in health-related sciences and technology we can observe that there are a smaller number of Hospitilizations and Fatalities reported due to the outbreaks.

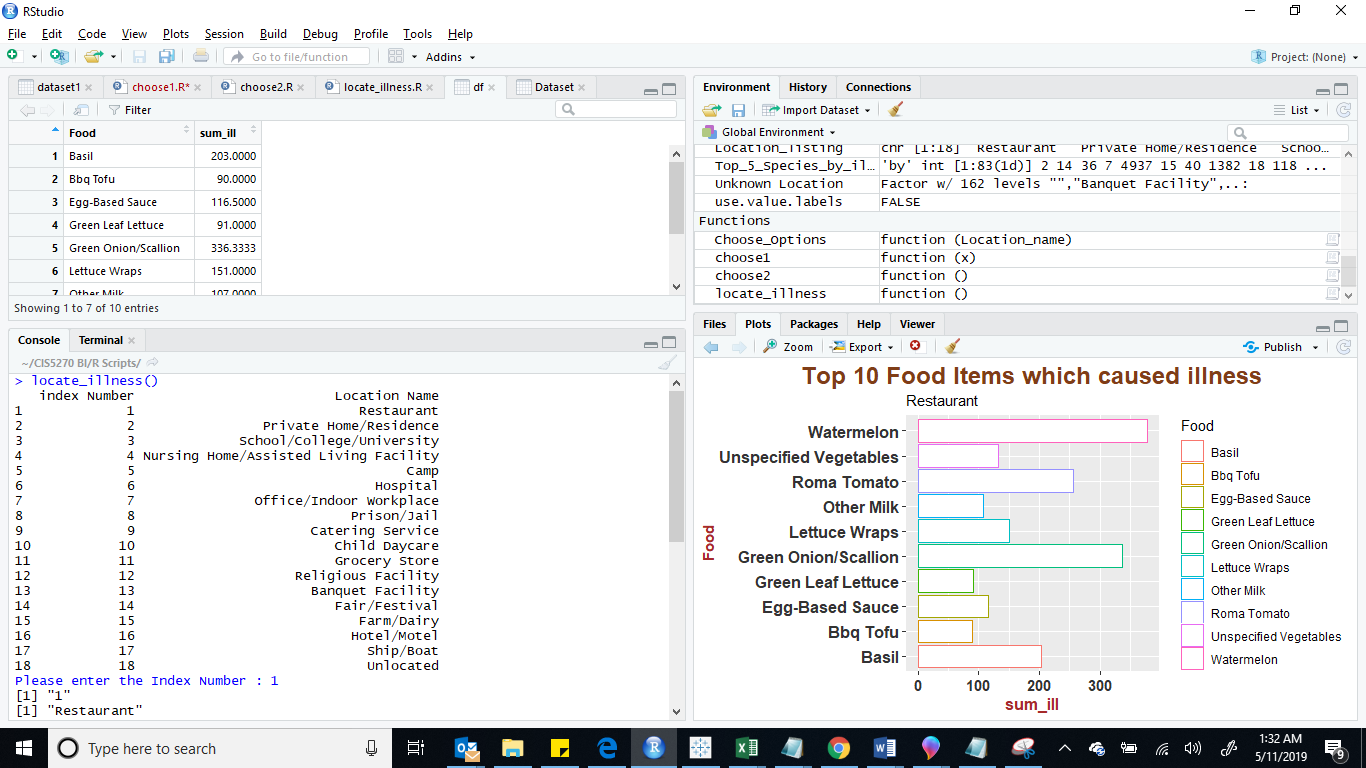
**USER DEFINED FUNCTIONS:**

**Q) Which are the Food Items specific to the locations that have caused major outbreaks in the United States?**

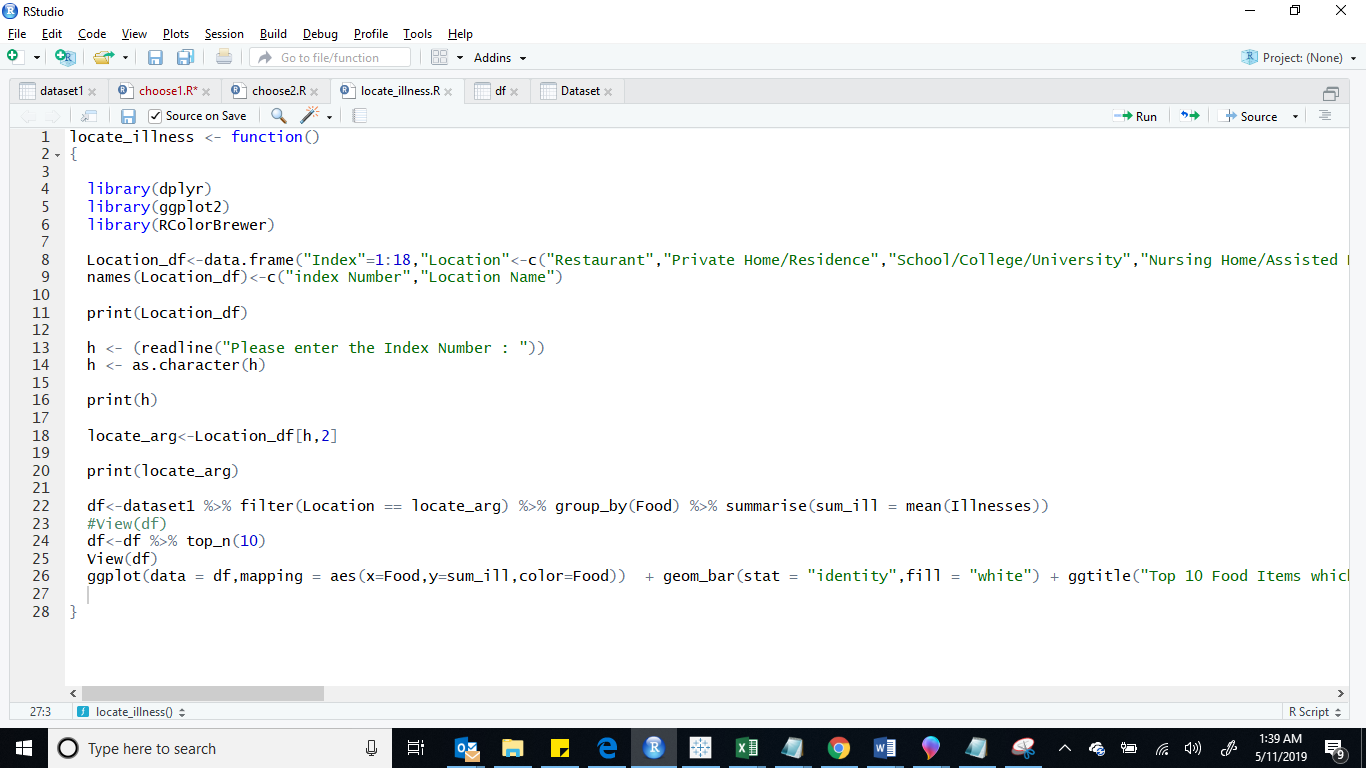
|  |  |
| --- | --- |
| **Index Number** | **Location** |
| 1 | Restaurant |
| 2 | Private Home/Residence |
| 3 | School/College/University |
| 4 | Nursing Home/Assisted Living Facility |
| 5 | Camp |
| 6 | Hospital |
| 7 | Office/Indoor Workplace |
| 8 | Prison/Jail |
| 9 | Catering Service |
| 10 | Child Daycare |
| 11 | Grocery Store |
| 12 | Religious Facility |
| 13 | Banquet Facility |
| 14 | Fair/Festival |
| 15 | Farm/Dairy |
| 16 | Hotel/Motel |
| 17 | Ship/Boat |
| 18 | Unlocated |

Datasets consists of Food Items, Location and Illnesses caused due to the specific food Items based on the locations. Funtion Locate\_Ilnnesses is created that accepts a user definied argument. User is displayed with a list as shown above to choose a specific location to analyze which are the top 10 food items specific to each location choosen by the user that have caused major outbreaks. Function returns the statistic associated to the chosed location.

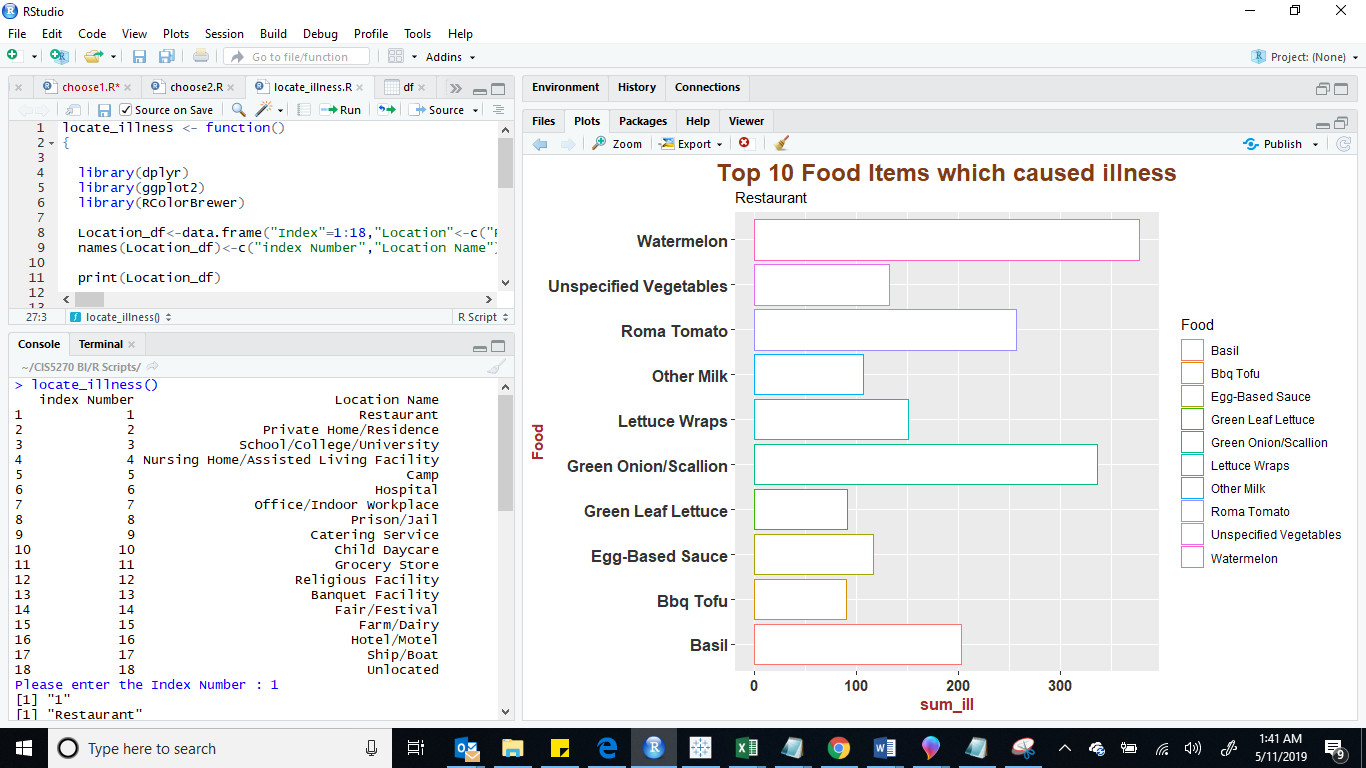
**Screenshot**



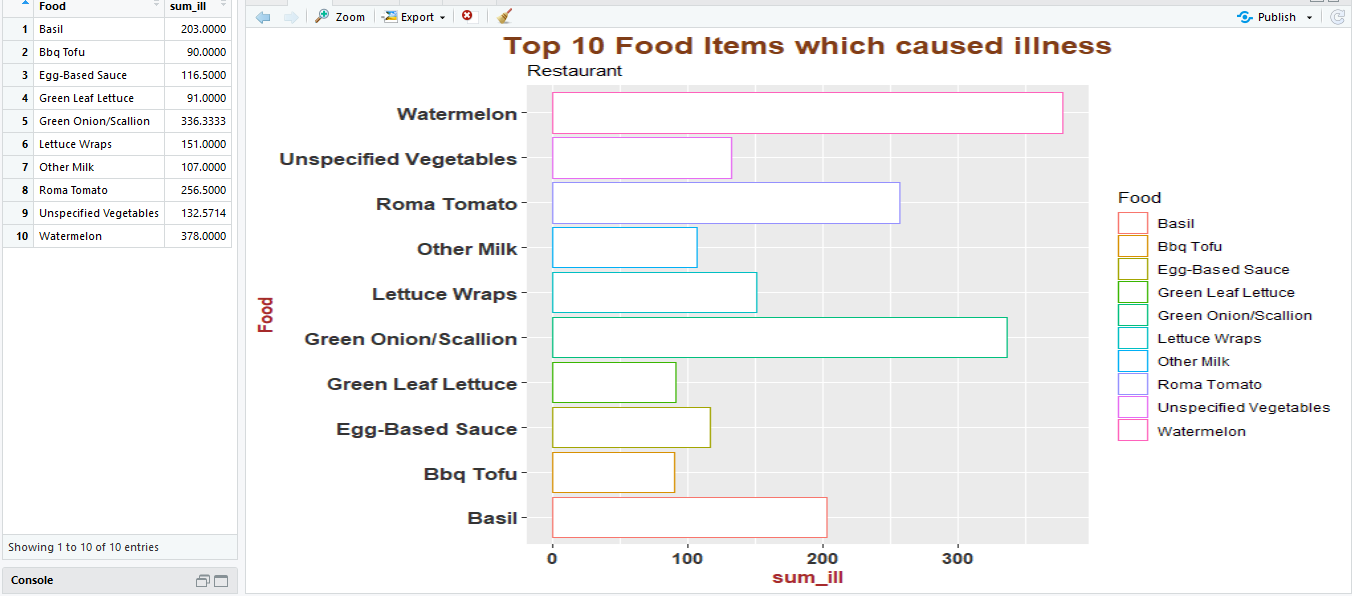
**CODE SCREENSHOT:**



**EXECUTION/OUTPUT SCREENSHOT:**



**EXECUTION WITH VISUALIZATION:**



**Code:**

#Defining function locate\_illness

#That will return the statistics for top 10 food items specific for the location entered by the user, which caused major disease outbreaks

locate\_illness <- function()

{

library(dplyr)

library(ggplot2)

library(RColorBrewer)

Location\_df<-data.frame("Index"=1:18,"Location"<-c("Restaurant","Private Home/Residence","School/College/University","Nursing Home/Assisted Living Facility","Camp","Hospital","Office/Indoor Workplace","Prison/Jail","Catering Service","Child Daycare","Grocery Store","Religious Facility","Banquet Facility","Fair/Festival","Farm/Dairy","Hotel/Motel","Ship/Boat","Unlocated"),stringsAsFactors = FALSE)

names(Location\_df)<-c("index Number","Location Name")

print(Location\_df)

h <- (readline("Please enter the Index Number : "))

h <- as.character(h)

print(h)

locate\_arg<-Location\_df[h,2]

print(locate\_arg)

df<-dataset1 %>% filter(Location == locate\_arg) %>% group\_by(Food) %>% summarise(sum\_ill = mean(Illnesses))

#View(df)

df<-df %>% top\_n(10)

View(df)

ggplot(data = df,mapping = aes(x=Food,y=sum\_ill,color=Food)) + geom\_bar(stat = "identity",fill = "white") + ggtitle("Top 10 Food Items which caused illness",subtitle = Location\_df[,2])+ coord\_flip() +theme(plot.title = element\_text(family = "Trebuchet MS", color="green", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title=element\_text(colour="#7F3D17"))+theme(axis.title.y= element\_text(colour="brown",size=11,face="bold"),axis.text.x=element\_text(color="grey20",size=11,face="bold"),axis.text.y = element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

}

**Execution Result:**

> setwd("~/CIS5270 BI/R Scripts/")

> source('~/CIS5270 BI/R Scripts/locate\_illness.R')

> locate\_illness()

index Number Location Name

1 1 Restaurant

2 2 Private Home/Residence

3 3 School/College/University

4 4 Nursing Home/Assisted Living Facility

5 5 Camp

6 6 Hospital

7 7 Office/Indoor Workplace

8 8 Prison/Jail

9 9 Catering Service

10 10 Child Daycare

11 11 Grocery Store

12 12 Religious Facility

13 13 Banquet Facility

14 14 Fair/Festival

15 15 Farm/Dairy

16 16 Hotel/Motel

17 17 Ship/Boat

18 18 Unlocated

Please enter the Index Number: 1

[1] "1"

[1] "Restaurant"

**NOTE: Visulization with respect to the Location indexed, “Restaurant” will be shown.**

**CODES:**

**DATA CLEANING CODES**

**1) Change Location Column Value from NA to Unlocated:**

setwd("~/CIS5270 BI/R Scripts/")

dataset1<-read.csv("outbreaks.csv",header = TRUE,sep = ",")

View(dataset1)

dataset1[dataset1==""]<-NA

View(dataset1)

sapply(dataset1, class)

Year Month State Location

"integer" "factor" "factor" "factor"

Food Ingredient Species Serotype.Genotype

"factor" "factor" "factor" "factor"

Status Illnesses Hospitalizations Fatalities

"factor" "integer" "integer" "integer"

j<-lapply(dataset1$Location,as.character)

j[is.na(j)]<-"Unlocated"

dataset1$Location<-lapply(j,as.character)

View(dataset1)

**2) Delete the Redundant Column Ingredient:**

dataset1$Ingredient<-NULL

**3) Change Species Column Value from NA to Unidentified**

j<-lapply(dataset1$Species,as.character)

j[is.na(j)]<-"Unidentified"

dataset1$Species<-lapply(j,as.character)

View(dataset1)

**4) Drop Column Serotype/Genotype**

dataset1$Serotype.Genotype<-NULL

**5) Change Status Column Value from NA to Unconfirmed**

j<-lapply(dataset1$Status,as.character)

j[is.na(j)]<-"Unconfirmed"

dataset1$Status<-lapply(j,as.character)

View(dataset1)

**6) Insert integer Value 0 in place of NA under column Hospitilization**

j<-lapply(dataset1$Hospitalizations,as.character)

j[is.na(j)]<-0

dataset1$Hospitalizations<-lapply(j,as.integer)

View(dataset1)

**7) Insert integer Value 0 in place of NA under column Fatalities**

j<-lapply(dataset1$Fatalities,as.character)

j[is.na(j)]<-0

dataset1$Fatalities<-lapply(j,as.integer)

View(dataset1)

**8) Change Food Column Value from NA to Unspecified**

j<-lapply(dataset1$Food,as.character)

j[is.na(j)]<-"Unspecified"

dataset1$Food<-lapply(j,as.character)

View(dataset1)

**9) Separate Food Column with only 1st word having separater ",":**

library("tidyr")

dataset1<-separate(dataset1,Food,c("Food"),sep=",")

**10) Separate Food Column with only 1st word having separater ";"**

dataset1<-separate(dataset1,Food,c("Food"),sep=";")

**11) Separate Species Column with only 1st word having separater ";"**

dataset1<-separate(dataset1,Species,c("Species"),sep=";")

**12) Separate Species Column with only 1st word having separater ","**

dataset1<-separate(dataset1,Species,c("Species"),sep=",")

**13) Change Status Column Value from NA to Unconfirmed**

> j<-lapply(dataset1$Food,as.character)

> j[is.na(j)]<-"Unidentified"

> dataset1$Food<-lapply(j,as.character)

> View(dataset1)

**Visualization Codes:**

**Visualization 1:**

> setwd("~/CIS5270 BI/R Scripts/")

> Dataset<-read.csv("outbreaks.csv",header=T,sep=",")

> View(Dataset)

> library("dplyr")

>aggregate1\_species\_df<-aggregate(dataset1$Illnesses, by=list(Species=dataset1$Species), FUN=sum)

> x = arrange(aggregate1\_species\_df,desc(aggregate1\_species\_df$x) )

> x1<-x[-c(1),]

> head(x1)

> Treemaps\_Analysis<-treemap(x1, index="Species", vSize="x", vColor="x", type="dens")

> Treemaps\_Analysis<-treemap(x1, index="Species", vSize="x", vColor="x", type="dens",title = "Top Disease Species Reported in the US that have caused Maximum Illnesses amongst the population")

**Visualization 2:**

> library("ggplot2")

> year\_state\_hosp <- outbreaks %>% group\_by(Year,State) %>% summarise(sum\_hosp = sum(Hospitalizations)) %>% filter(State %in% c("California","Florida","Illinois","Michigan","Minnesota"))

> View(year\_state\_hosp)

> ggplot(data = year\_state\_hosp,mapping = aes(x=Year,y=sum\_hosp,color=State)) + geom\_line(size = 1) + ggtitle("Top 5 STATES having highest number of Hospitalizations from Year 1998 to 2015") +theme(plot.title = element\_text(family = "Trebuchet MS", color="blue", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title=element\_text(colour="#7F3D17"))+theme(axis.title.y= element\_text(colour="brown",size=11,face="bold"),axis.text.x=element\_text(colour="grey20",size=11,face="bold"),axis.text.y = element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**USER DEFINED CODES : Visualization 3:**

#Defining function locate\_illness

#That will return the statistics for top 10 food items specific for the location entered by the user , which caused major disease outbreaks

locate\_illness <- function()

{

library(dplyr)

library(ggplot2)

library(RColorBrewer)

Location\_df<-data.frame("Index"=1:18,"Location"<-c("Restaurant","Private Home/Residence","School/College/University","Nursing Home/Assisted Living Facility","Camp","Hospital","Office/Indoor Workplace","Prison/Jail","Catering Service","Child Daycare","Grocery Store","Religious Facility","Banquet Facility","Fair/Festival","Farm/Dairy","Hotel/Motel","Ship/Boat","Unlocated"),stringsAsFactors = FALSE)

names(Location\_df)<-c("index Number","Location Name")

print(Location\_df)

h <- (readline("Please enter the Index Number : "))

h <- as.character(h)

print(h)

locate\_arg<-Location\_df[h,2]

print(locate\_arg)

df<-dataset1 %>% filter(Location == locate\_arg) %>% group\_by(Food) %>% summarise(sum\_ill = mean(Illnesses))

#View(df)

df<-df %>% top\_n(10)

View(df)

ggplot(data = df,mapping = aes(x=Food,y=sum\_ill,color=Food)) + geom\_bar(stat = "identity",fill = "white") + ggtitle("Top 10 Food Items which caused illness",subtitle = Location\_df[,2])+ coord\_flip() +theme(plot.title = element\_text(family = "Trebuchet MS", color="green", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title=element\_text(colour="#7F3D17"))+theme(axis.title.y= element\_text(colour="brown",size=11,face="bold"),axis.text.x=element\_text(color="grey20",size=11,face="bold"),axis.text.y = element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**Execution Result:**

> setwd("~/CIS5270 BI/R Scripts/")

> source('~/CIS5270 BI/R Scripts/locate\_illness.R')

> locate\_illness()

index Number Location Name

1 1 Restaurant

2 2 Private Home/Residence

3 3 School/College/University

4 4 Nursing Home/Assisted Living Facility

5 5 Camp

6 6 Hospital

7 7 Office/Indoor Workplace

8 8 Prison/Jail

9 9 Catering Service

10 10 Child Daycare

11 11 Grocery Store

12 12 Religious Facility

13 13 Banquet Facility

14 14 Fair/Festival

15 15 Farm/Dairy

16 16 Hotel/Motel

17 17 Ship/Boat

18 18 Unlocated

Please enter the Index Number: 1

[1] "1"

[1] "Restaurant"

**REFERENCE**

1. <https://www.cdc.gov/fdoss/pdf/2016_FoodBorneOutbreaks_508.pdf>
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820314/>
3. “Deep Learning in R.” *An R Introduction to Statistics | R Tutorial*, [www.r-tutor.com](http://www.r-tutor.com)
4. CDC Current Outbreak List , <https://www.cdc.gov/outbreaks/>