FOOD-BORNE ILLNESS ANALYSIS

Authors:

Srijit Ravishankar

Surupa Tushar Chatterjee

What is considered a Food Borne Illness outbreak?

- A foodborne disease outbreak occurs when two or more people get the same illness from the same contaminated food or drink.

- The **Center for Disease Control and Prevention (CDC)** estimates roughly 1 in 6 Americans (48 million people) get sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases each year.

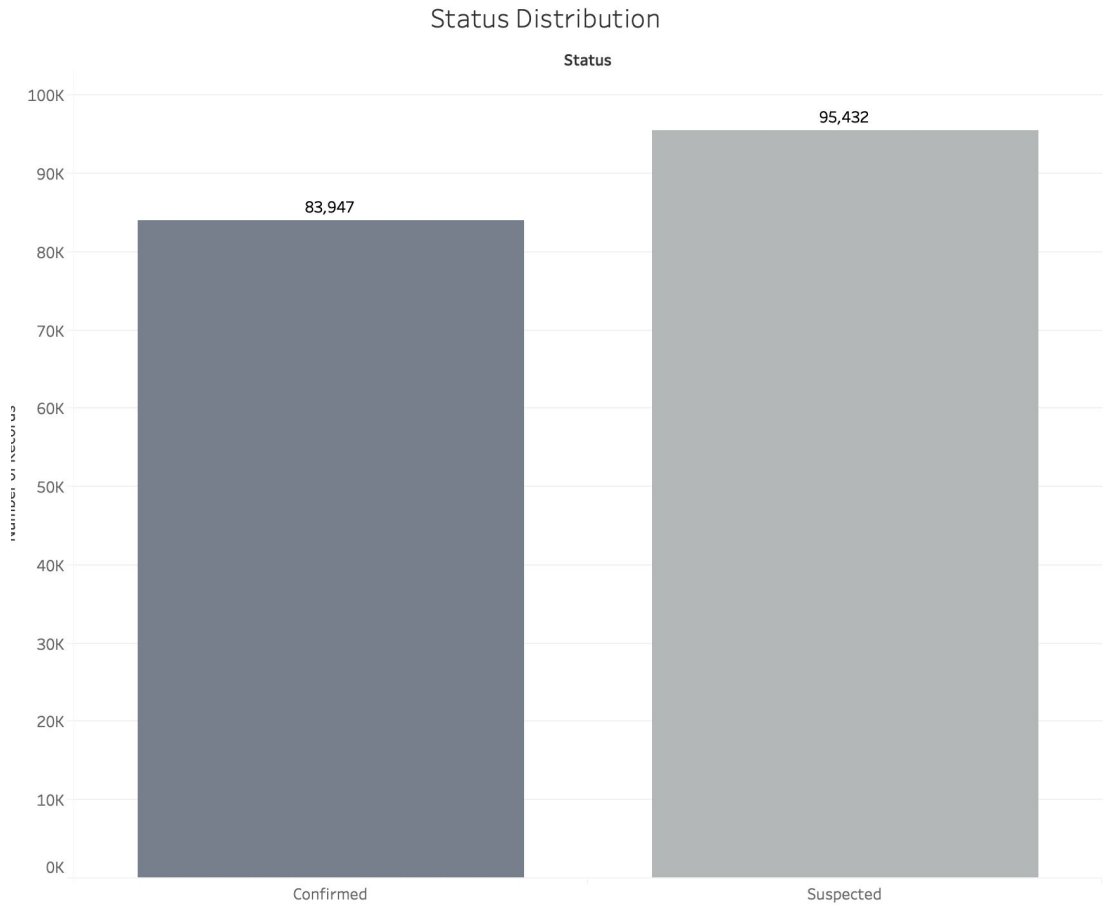
What do we intend to do?

- We aim to apply data mining techniques that can help discover interesting patterns in this dataset
- Apply algorithms to predict the status of a food borne illness (confirmed/suspected)
- Identify the key ingredients/sources in food that lead to infections
- Identify specific regions/areas across U.S where the risk of foodborne diseases is high/low
- Which location is a red zone, where the patient consumed the food(Restaurants/Home etc.)
- Identify the specific period of the year when food borne illnesses are more likely to occur.

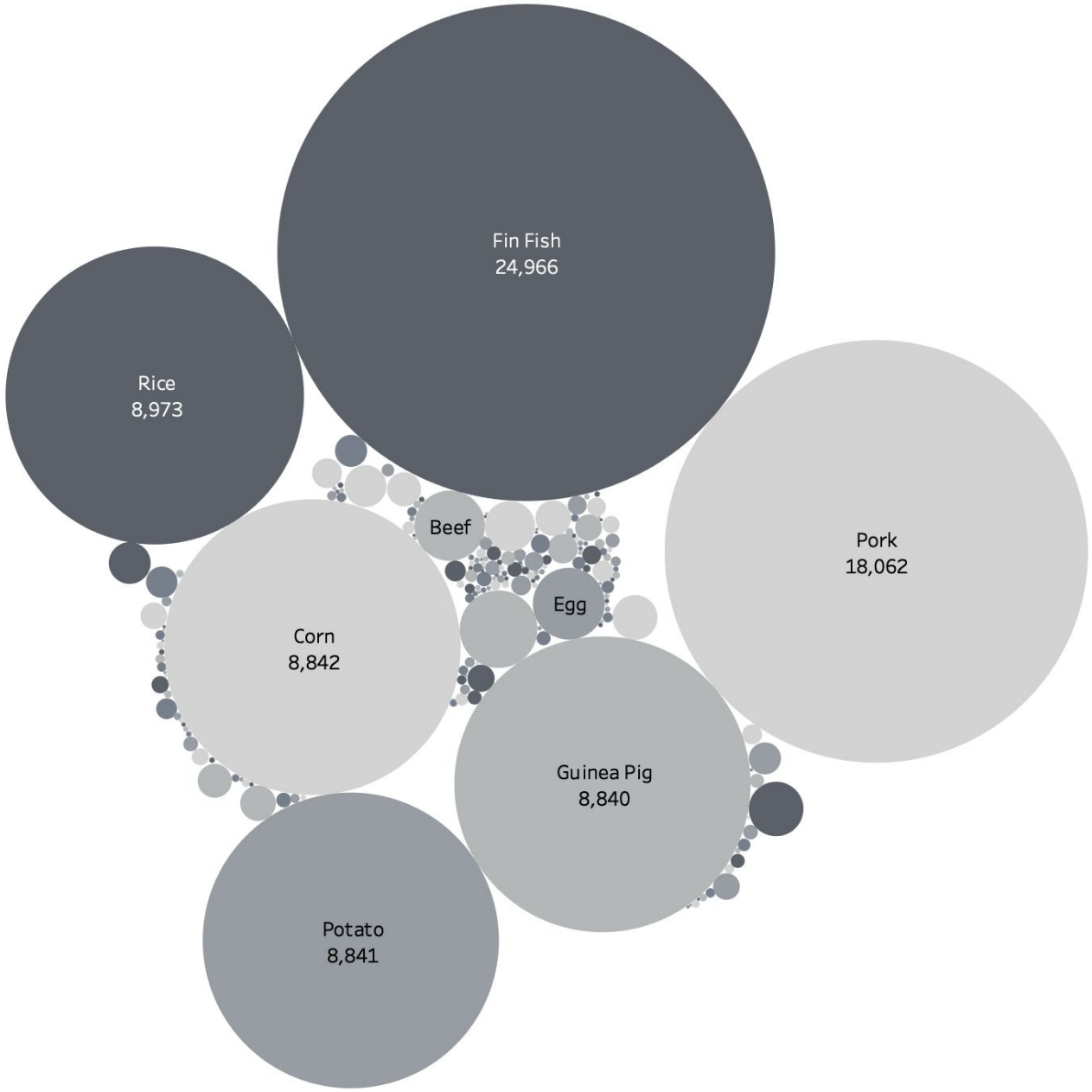
Exploratory Data Analysis of the Food Borne Illness Dataset

Distribution of the Status Feature

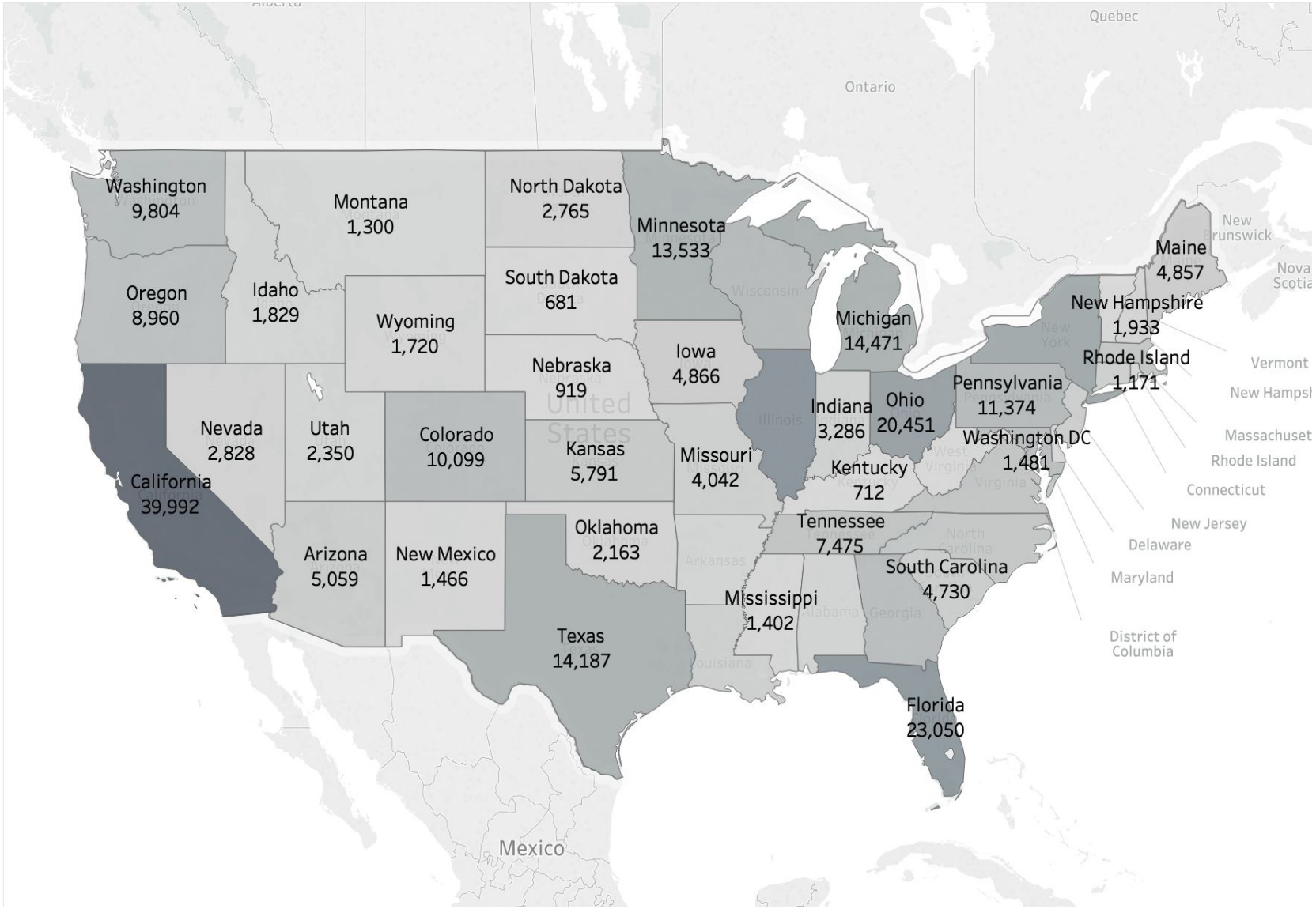
- We can see from the graph that the class is balanced
- Therefore we do not need to under sample or over sample the data.



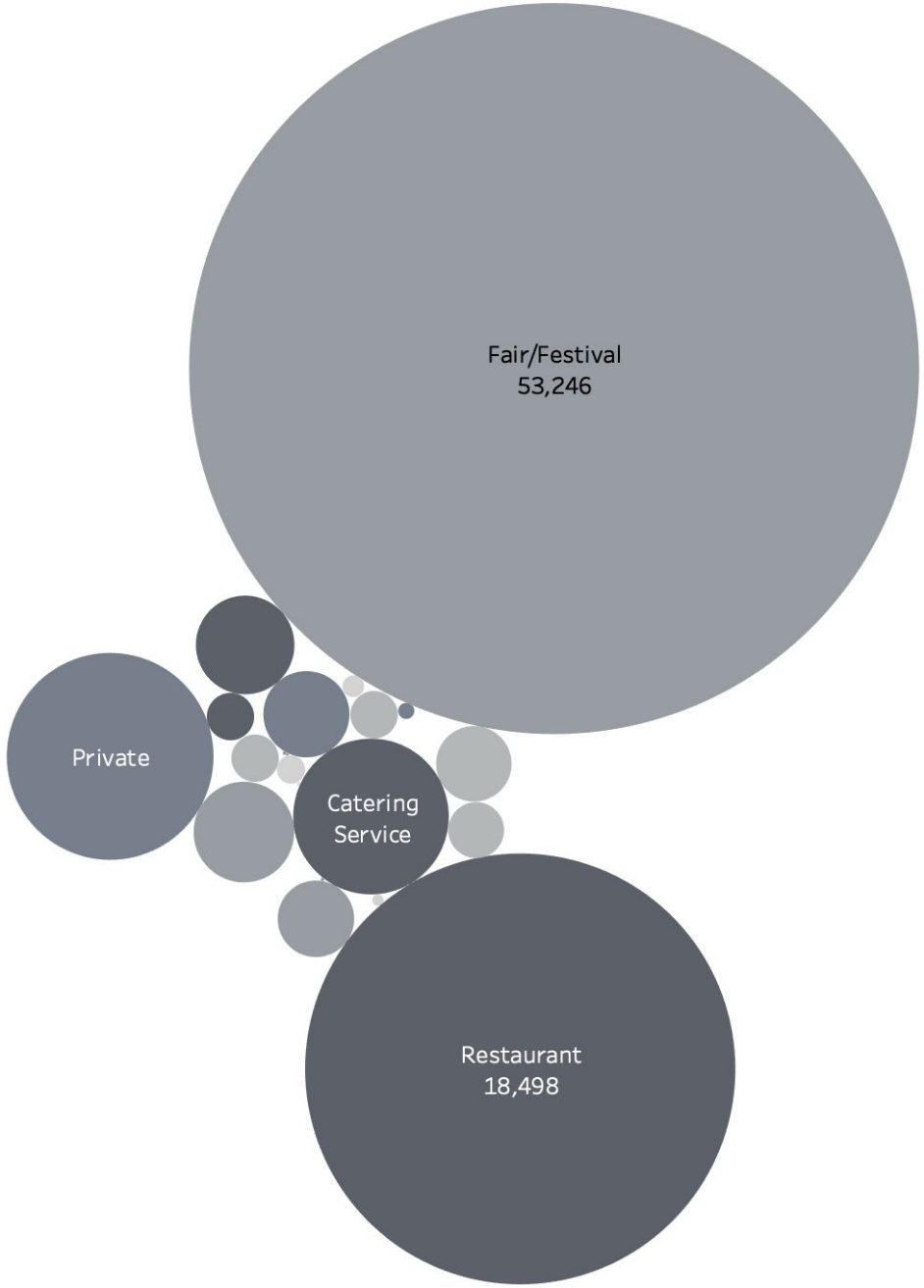
key Ingredients in food that lead to food borne illnesses



Risk of foodborne illness across Regions/areas in the U.S

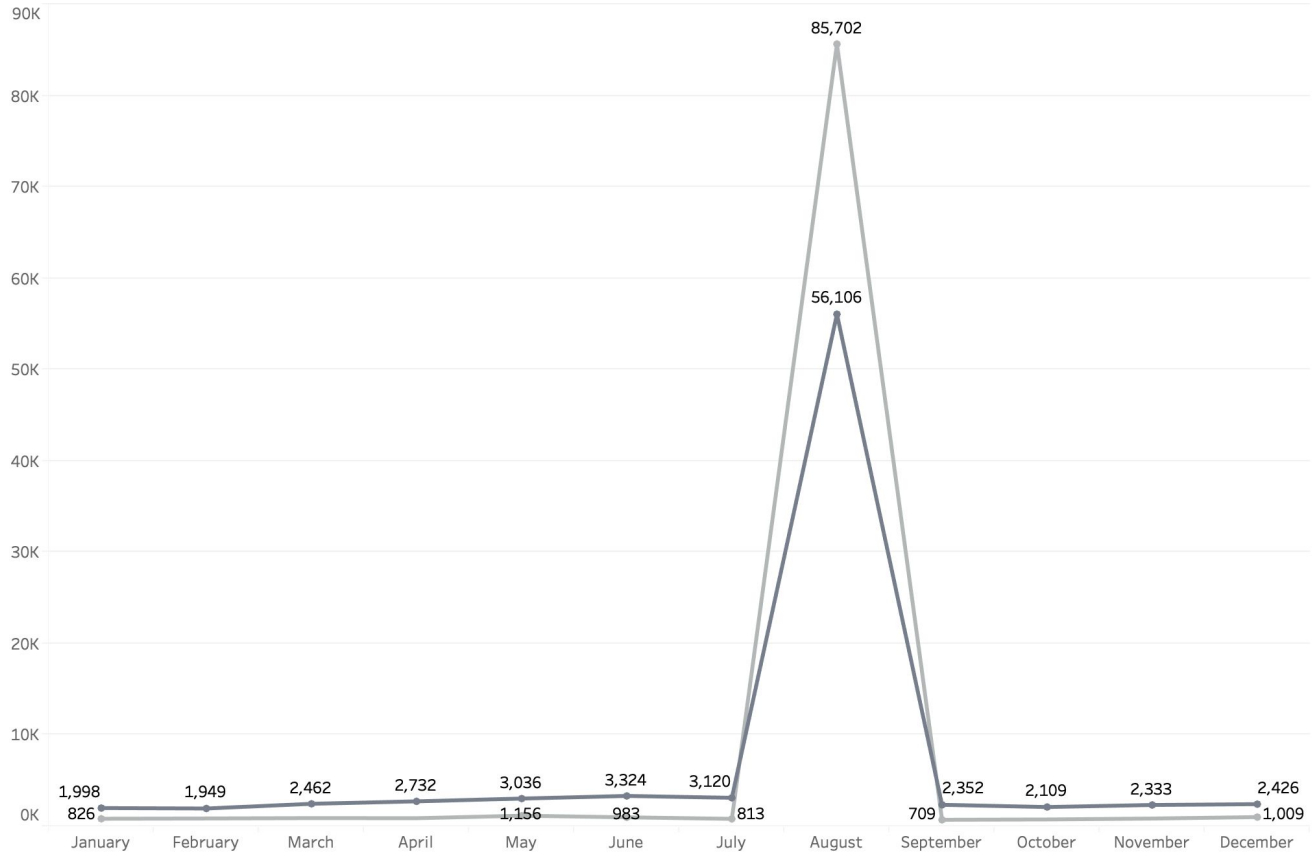


Red Zone locations - where the patient consumed the food and reported food borne illnesses

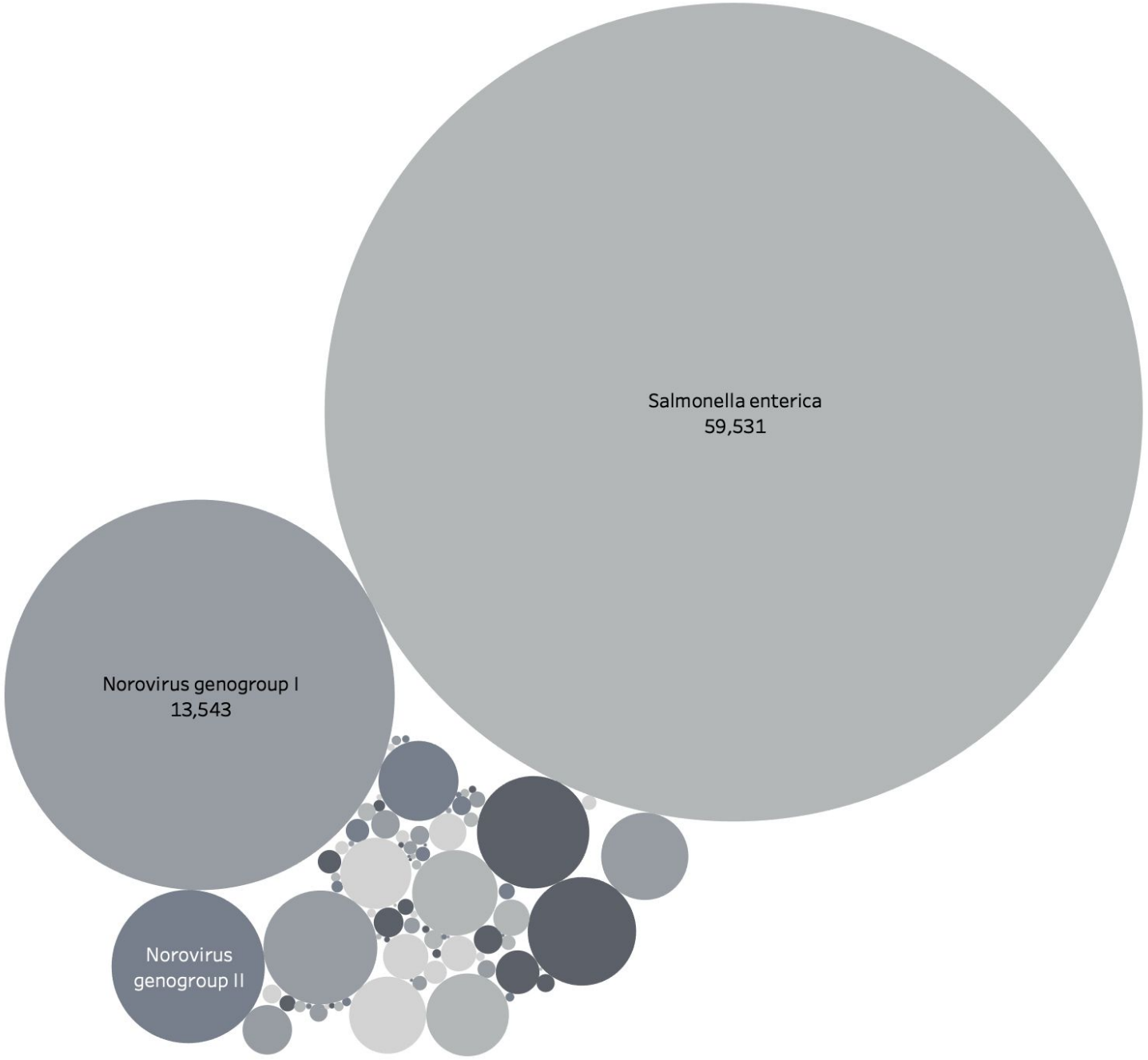


Specific period of the year when food borne illnesses are likely to occur.

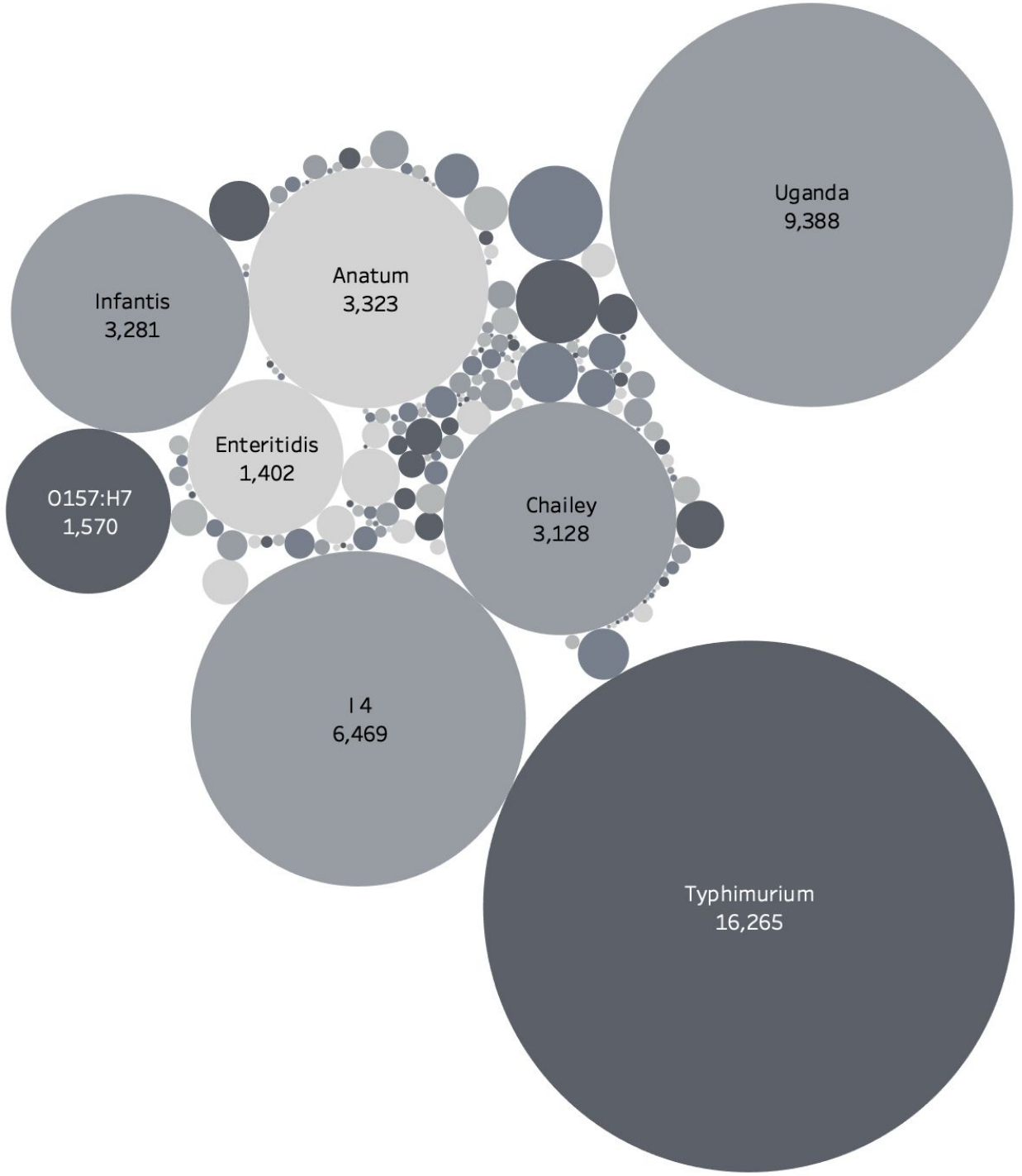
- We can see that cases of confirmed and suspected food borne illnesses are at peak during the month of August.



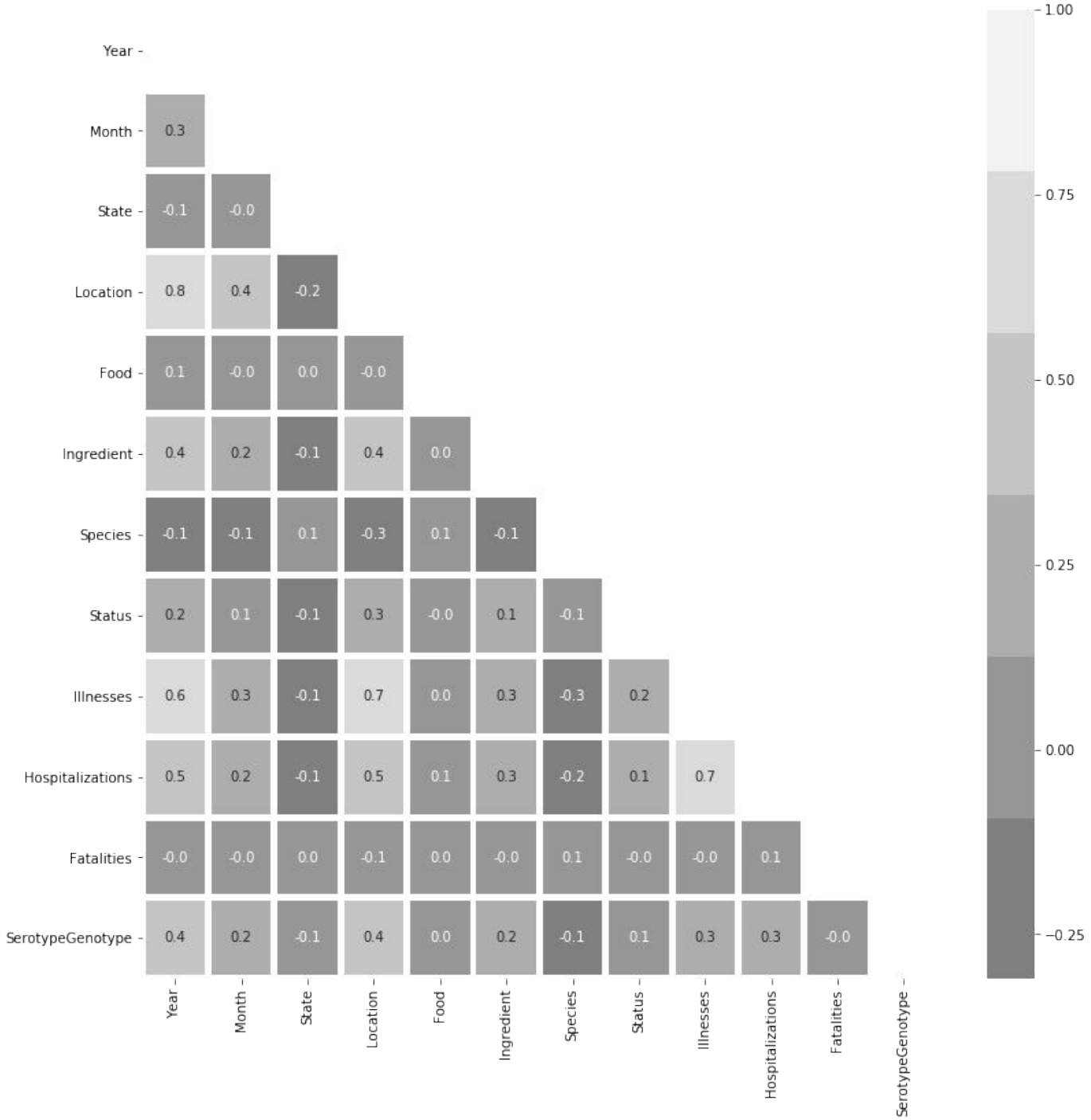
Species predominantly causing Food borne illnesses



Serotype/Genotype of the species predominantly causing Food borne illnesses



Correlation Plot for all the features in the Food Borne Illness Dataset



Predicting Status of the Food Borne Illness as Confirmed or Suspected

Models we used for prediction

- K Nearest Neighbors Classifier
- Logistic Regression Classifier
- Ada boost
- Random Forest
- Gradient Descent Classifier

K Nearest Neighbors Classifier

Model Report
Accuracy : 0.5975
AUC Score: 0.645335

	precision	recall	f1-score	support
0	0.57	0.56	0.57	16796
1	0.62	0.63	0.62	19080
micro avg	0.60	0.60	0.60	35876
macro avg	0.60	0.60	0.60	35876
weighted avg	0.60	0.60	0.60	35876

Logistic Regression

Model Report
Accuracy : 0.6396
AUC Score: 0.624351

	precision	recall	f1-score	support
0	0.76	0.34	0.47	16796
1	0.61	0.90	0.73	19080
micro avg	0.64	0.64	0.64	35876
macro avg	0.68	0.62	0.60	35876
weighted avg	0.68	0.64	0.61	35876

Ada Boost Classifier

Model Report
Accuracy : 0.6638
AUC Score: 0.656718

	precision	recall	f1-score	support
0	0.84	0.35	0.49	16796
1	0.62	0.94	0.75	19080
micro avg	0.66	0.66	0.66	35876
macro avg	0.73	0.64	0.62	35876
weighted avg	0.72	0.66	0.63	35876

Random Forest Classifier

Model Report
Accuracy : 0.6852
AUC Score: 0.665214

	precision	recall	f1-score	support
0	0.92	0.36	0.52	16796
1	0.63	0.97	0.77	19080
micro avg	0.69	0.69	0.69	35876
macro avg	0.78	0.67	0.64	35876
weighted avg	0.77	0.69	0.65	35876

Gradient Boosting Classifier

Model Report
Accuracy : 0.67
AUC Score: 0.666600

	precision	recall	f1-score	support
0	0.85	0.36	0.50	16796
1	0.63	0.95	0.75	19080
micro avg	0.67	0.67	0.67	35876
macro avg	0.74	0.65	0.63	35876
weighted avg	0.73	0.67	0.64	35876

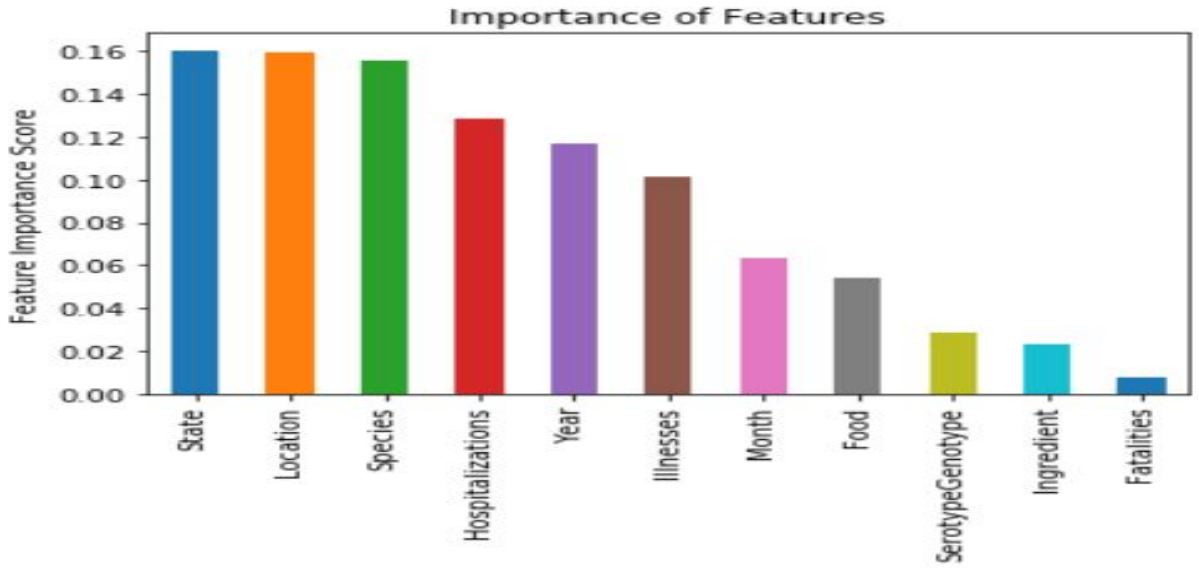
Hyper Parameter tuning for Random Forest Classifier

Parameter	Values Tested	Optimal Value Achieved
n_estimators	5,10,15,20	20
max_depth	20,25,30,35,40,45	25
min_samples_split	2,4,6,8,10,20,40,60,100	2
min_samples_leaf	1,3,5,7,9	1

```
Accuracy of the RF on test set: 0.687
AUC Score: 0.665191
              precision    recall  f1-score   support

           0       0.93      0.36      0.52     16796
           1       0.63      0.98      0.77     19080

   micro avg       0.69      0.69      0.69     35876
   macro avg       0.78      0.67      0.64     35876
weighted avg       0.77      0.69      0.65     35876
```



Predicting Number of people affected by Food Borne Illness

Models we used for prediction

- Linear Regression
- Ridge Regression
- Decision Tree Regression
- Gradient Boosting Regression
- Random Forest Regression

Regression Model Performance

Model	Training Accuracy	Test Accuracy
Linear Regressor	0.6856	0.6619
Ridge Regressor	0.6856	0.6619
Decision Tree Regressor	1.0000	0.7968
Gradient Boosting Regressor	0.8254	0.7831
Random Forest Regressor	0.9822	0.8203

- Random Forest Regressor achieved best Training/Test Accuracy.
- Performed Hyperparameter tuning for Random Forest Regressor to improve prediction accuracy.

Hyper Parameter Tuning : Random Forest Regressor

Parameter	Values Tested	Optimal Value Achieved
n_estimators	20,50,80,100,120,150	120
min_samples_split	2,4,6,8,10,20,40,60,100	2
min_samples_leaf	1,3,5,7,9	1

After model tuning:

```
Model Report
Train Accuracy : 0.9891
Test Accuracy : 0.8342
```

Analysis

- As the data set contains both categorical as well as continuous variables, Tree based Regression Models tend to perform better than Linear Regression Models, since tree can accurately divide data based on categorical variables.

Future Work

- Further Hyper Parameter Tuning to improve Accuracy Scores.
- Create models to answer different queries like food that leads to hospitalizations or fatalities, food combinations and potential infection source etc.