# PREDICTIVE MODEL OF MEDICAL RECORDS USING DATA SCIENCE

By

**Muhammad Atta Khan (19k-0932)**

**Abbas Ali Rizvi (18k-1312)**


**Report submitted for the**

**Degree of MS (Data Science)**

**National University of Computer & Emerging Sciences**

Karachi-Pakistan

**2019**

# PREDICTIVE MODEL OF MEDICAL RECORDS USING DATA SCIENCE

# MS Course Project


By

**Muhammad Atta Khan (19k-0932)**

**Abbas Ali Rizvi (18k-1312)**

Course Project Supervisor

**Dr. Muhammad Wasim**


**2019**


**National University of Computer & Emerging Sciences**

Karachi-Pakistan

Pakistan

# Table of Contents

## 1. Introduction

Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data.

The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution. Key components of data science are given in figure 1.

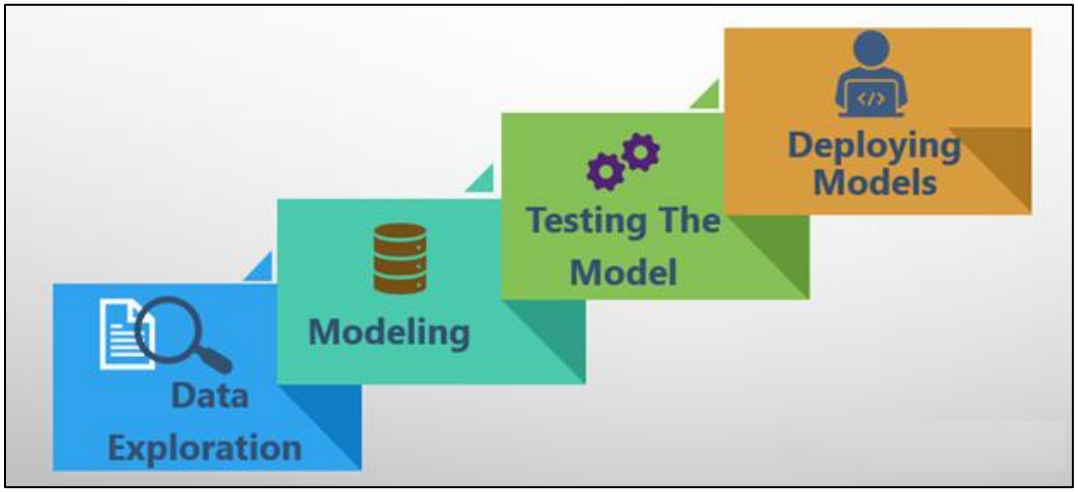



*Figure1. Key Components of Data Science*

**Data Science Process**

Data science process is consists of six key modules — discovery, data preparation, model planning, model building, operation and communicate results. These processes are given in figure 2.

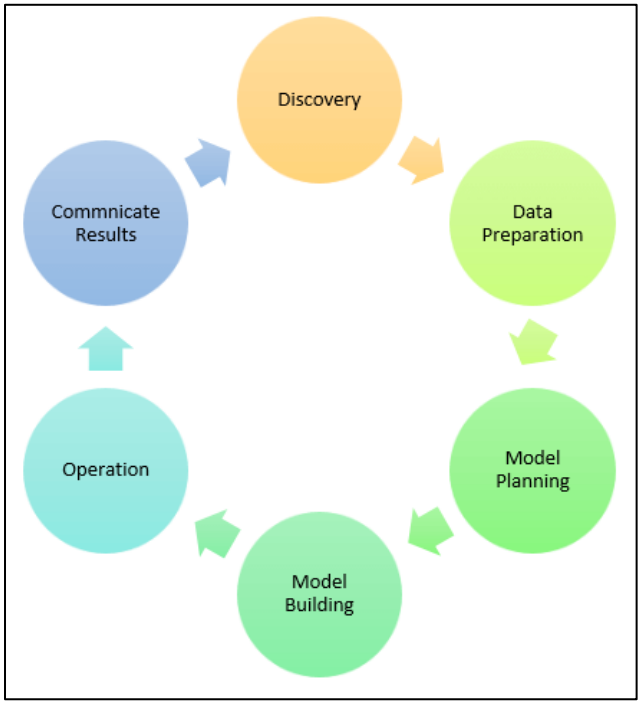



*Figure2. Data Science Processes Model*

**i. Discovery**

Discovery step involves acquiring data from all the identified internal & external sources which helps you to answer the business question.

The data can be:

- Logs from webservers
- Data gathered from social media
- Census datasets
- Data streamed from online sources using APIs

**ii. Data Preparation**

Data can have lots of inconsistencies like missing value, blank columns, incorrect data format which needs to be cleaned. You need to process, explore, and condition data before modeling. The cleaner your data, the better are your predictions.

**iii. Model Planning**

In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

**iv. Model Building**

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.

### v. Operationalize

In this stage, you deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing.

### vi. Communicate Results

In this stage, the key findings are communicated to all stakeholders. This helps you to decide if the results of the project are a success or a failure based on the inputs from the model.

## 1.1 Problem Statement and Domain:

The aim of the research work is to explore the medical data of the reputed hospital and to provide the insights of data in the form of visualization. The visualization help understand the hidden trends in data and help us identify the notable disease among gender and different age groups.

## 1.2 List of Requirements:

Following are the queries or trends identify in this research work.

- Relationship between Age and Disease.
- Visualize Gender vs Disease trend.
- Identify the most common Disease in female.
- Identify the most common Disease in male.
- In which year female having most common Disease.
- In which year male having most common Disease.
- Male vs Female disease pattern.
- Identify the most common Disease in male and female.

- Which year is most crucial for female?
- Which year is most crucial for male?
- Visualization of all results.

### 1.3 Introduction to Data Set:

The Data set is the medical record of patients suffering from different diseases during the period of 4 years (2014 - 2018). The data set contain 4 features i.e. Age, Year, Disease and Gender. The minimum age recorded in the data is 21 years where as the maximum age is recorded as 71 years. All the data is recorded between years 2014 to 2018. The total rows in dataset is 6907. The data is already in its clean form as there is no empty, null or any other ambiguity appear in data. There are 3466 male and 3441 female record in dataset.

### 1.4 Overview of Statistical and Mathematical Methods for Data Science

#### Linear Regression:

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between Age and Gender.

## 2. Data Analytics

We are using different type of statistical analysis between Medical data features (the list of requirements mentions in above section 1.2) and make graphs for visualize the statistical results.

### 2.1 Age vs Disease

In the dataset the age is ranges from 21 to 76. The Age feature is numerical whereas disease feature is categorical and contain name of diseases. In order to find the relation between age and disease we have categorized the age factor into bins of 10. The disease feature is than group on the bases of age by using the formula “value_counts”. There are 20 different diseases recorded in all age group. For visualization we have used bar charts and box plot. By visualizing the data, it is found that Pneumonia is the most occurring disease in all age group as number of patients recorded in Pneumonia is quite high for all ages. A notable number of cases also reported for viral fever. By plotting disease with age bins, it is found that cases of Pneumonia, Dengue, Viral Fever and Typhoid fever is occurring in all age group peoples. But Pneumonia is spreading at alarming phase with more than 3000 cases reported.

**Code Snippet:**

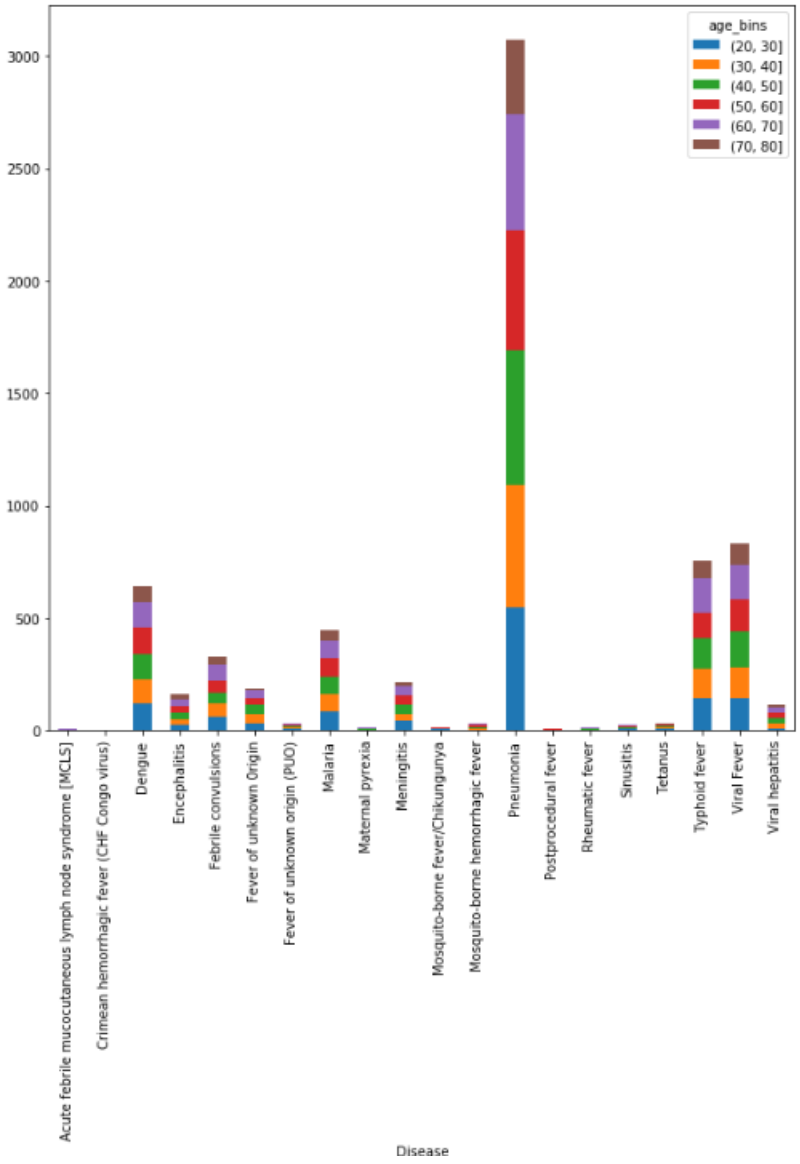
```
import pandas as pd
import os
import seaborn as sns
import matplotlib.pyplot as plt

path = os.path.abspath('C:\\Users\\Muhammad.Atta\\Desktop\\Lab2_data\\Lab2_data')
df = pd.read_csv(path + "\\MedicalData.csv")
df['age_bins'] = pd.cut(x=df['Age'], bins=[20, 30, 40, 50,60,70,80])
age_disease = pd.DataFrame(df.groupby('age_bins')['Disease'].value_counts().unstack())
age_disease.fillna(0, inplace=True)
age_disease['others'] =  age_disease[['Acute febrile mucocutaneous lymph node syndrome [MCLS]' ,
'Crimean hemorrhagic fever (CHF Congo virus)' , 'Maternal pyrexia' ,'Postprocedural fever','Rheumatic
```

```
fever','Sinusitis' ,'Mosquito-borne fever/Chikungunya' , 'Fever of unknown origin (PUO)',
'Tetanus']].sum(axis=1)
age_disease.drop(labels = ['Acute febrile mucocutaneous lymph node syndrome [MCLS]' , 'Crimean
hemorrhagic fever (CHF Congo virus)' , 'Maternal pyrexia' ,'Postprocedural fever','Rheumatic fever',
'Sinusitis' , 'Mosquito-borne fever/Chikungunya','Fever of unknown origin PUO)', 'Tetanus'],
axis="columns", inplace=True)
age_disease[list(age_disease.columns)].plot(kind='bar',stacked=True, figsize=(10,10) )
disease_age = pd.DataFrame(df.groupby('Disease')['age_bins'].value_counts().unstack())
disease_age.fillna(0, inplace=True)
disease_age[list(disease_age.columns)].plot(kind='bar', stacked = True , figsize=(10,10))
age_per_disease = pd.DataFrame(df.groupby('Age')['Disease'].value_counts().unstack())
age_per_disease.fillna(0, inplace=True)
fig = plt.figure(figsize=(10,10))
plt.xticks(rotation=90)
ax = sns.boxplot( data=age_per_disease , palette="Set3")
pd_crosstab1 = pd.crosstab(df["Disease"], df["Age"])
fig = plt.figure(figsize=(10,10))
sns.heatmap(pd_crosstab1, cbar=True, cmap="BuGn", linewidths=0.3)

# Rotate tick marks for visibility
plt.yticks(rotation=0)
plt.xticks(rotation=90)
plt.show()
```

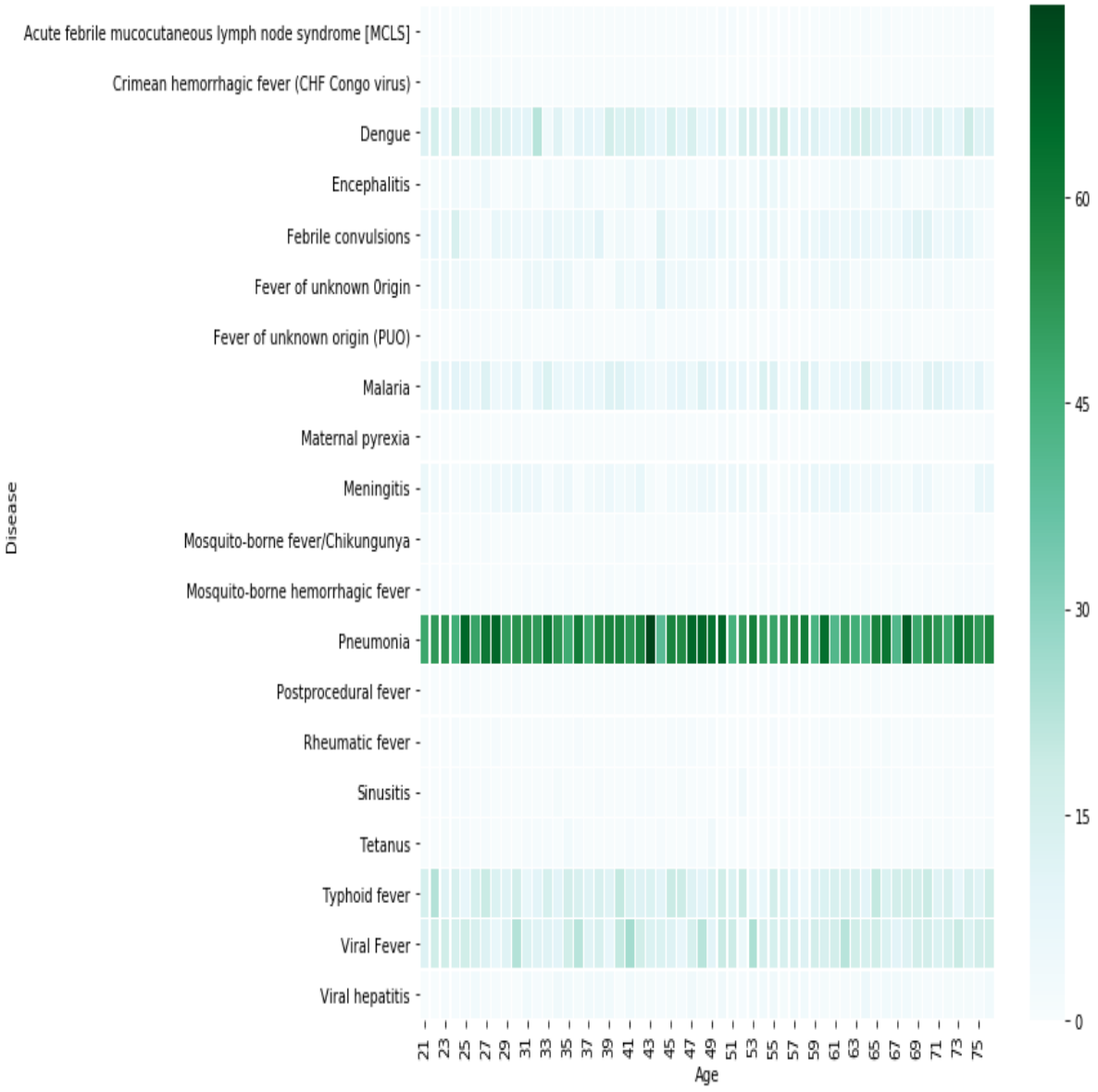
## Table below shows the Age wise bin Vs Disease:

| age_bins | (20, 30] | (30, 40] | (40, 50] | (50, 60] | (60, 70] | (70, 80] |
|---|---|---|---|---|---|---|
| Disease | | | | | | |
| Acute febrile mucocutaneous lymph node syndrome [MCLS] | 0.0 | 0.0 | 1.0 | 1.0 | 2.0 | 0.0 |
| Crimean hemorrhagic fever (CHF Congo virus) | 3.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Dengue | 119.0 | 107.0 | 113.0 | 118.0 | 113.0 | 71.0 |
| Encephalitis | 23.0 | 22.0 | 30.0 | 33.0 | 30.0 | 24.0 |
| Febrile convulsions | 62.0 | 58.0 | 49.0 | 51.0 | 74.0 | 30.0 |
| Fever of unknown 0rigin | 30.0 | 40.0 | 42.0 | 33.0 | 32.0 | 9.0 |
| Fever of unknown origin (PUO) | 7.0 | 4.0 | 9.0 | 4.0 | 6.0 | 2.0 |
| Malaria | 81.0 | 82.0 | 77.0 | 79.0 | 79.0 | 45.0 |
| Maternal pyrexia | 1.0 | 1.0 | 2.0 | 5.0 | 2.0 | 1.0 |
| Meningitis | 39.0 | 35.0 | 37.0 | 41.0 | 45.0 | 19.0 |
| Mosquito-borne fever/Chikungunya | 4.0 | 1.0 | 1.0 | 7.0 | 1.0 | 1.0 |
| Mosquito-borne hemorrhagic fever | 3.0 | 5.0 | 4.0 | 12.0 | 4.0 | 3.0 |
| Pneumonia | 546.0 | 548.0 | 600.0 | 531.0 | 518.0 | 329.0 |
| Postprocedural fever | 2.0 | 0.0 | 1.0 | 2.0 | 1.0 | 0.0 |
| Rheumatic fever | 2.0 | 1.0 | 4.0 | 2.0 | 3.0 | 0.0 |
| Sinusitis | 4.0 | 3.0 | 4.0 | 5.0 | 6.0 | 2.0 |
| Tetanus | 6.0 | 7.0 | 7.0 | 2.0 | 4.0 | 3.0 |
| Typhoid fever | 145.0 | 128.0 | 137.0 | 115.0 | 153.0 | 75.0 |
| Viral Fever | 145.0 | 137.0 | 155.0 | 147.0 | 154.0 | 93.0 |
| Viral hepatitis | 8.0 | 24.0 | 23.0 | 24.0 | 21.0 | 11.0 |

**Graphs below shows the Age wise bin Vs Disease:**



**Graphs below shows the Age Vs Disease:**




## 2.2 Gender vs Disease

Gender and Disease feature are both categorical. The gender feature contains values “Male” and “Female” whereas disease feature contains the name of diseases. For extracting insights between both features the data set is grouped in two way. In first approach the overall data set is grouped

on the gender factor. The second approach is quite instructive as it reveals the count of diseases in each age group.

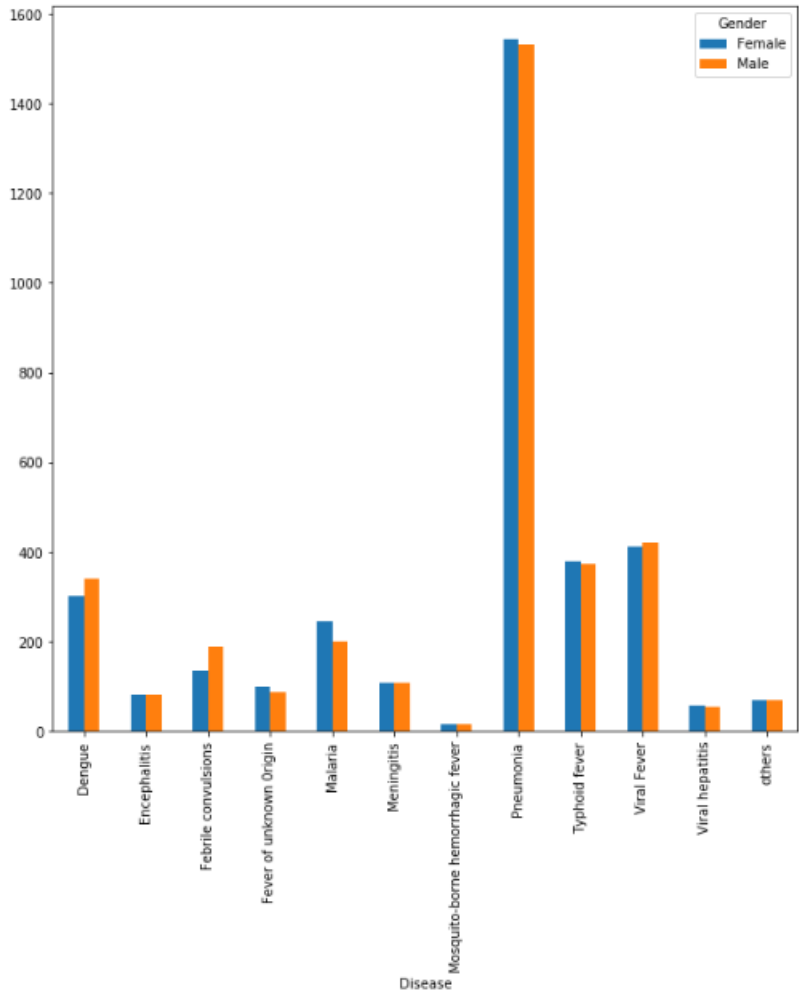
**Code Snippet:**

```
df['Gender'].unique()
gender_disease = pd.DataFrame(df.groupby('Disease')['Gender'].value_counts().unstack())
gender_disease[list(gender_disease.columns)].plot(kind='bar' ,figsize = (10,10))
```

Below table show gender vs disease stats.

| Gender | Female | Male |
|---|---|---|
| Disease | | |
| Acute febrile mucocutaneous lymph node syndrome [MCLS] | 2 | 2 |
| Crimean hemorrhagic fever (CHF Congo virus) | 1 | 2 |
| Dengue | 302 | 339 |
| Encephalitis | 81 | 81 |
| Febrile convulsions | 136 | 188 |
| Fever of unknown 0rigin | 99 | 87 |
| Fever of unknown origin (PUO) | 16 | 16 |
| Malaria | 244 | 199 |
| Maternal pyrexia | 6 | 6 |
| Meningitis | 107 | 109 |
| Mosquito-borne fever/Chikungunya | 9 | 6 |
| Mosquito-borne hemorrhagic fever | 15 | 16 |
| Pneumonia | 1542 | 1530 |
| Postprocedural fever | 1 | 5 |
| Rheumatic fever | 7 | 5 |
| Sinusitis | 14 | 10 |
| Tetanus | 12 | 17 |
| Typhoid fever | 380 | 373 |
| Viral Fever | 411 | 420 |
| Viral hepatitis | 56 | 55 |

Below graph show gender vs disease analyses.




## 2.3 Most Common Disease by Gender:

**Female:**

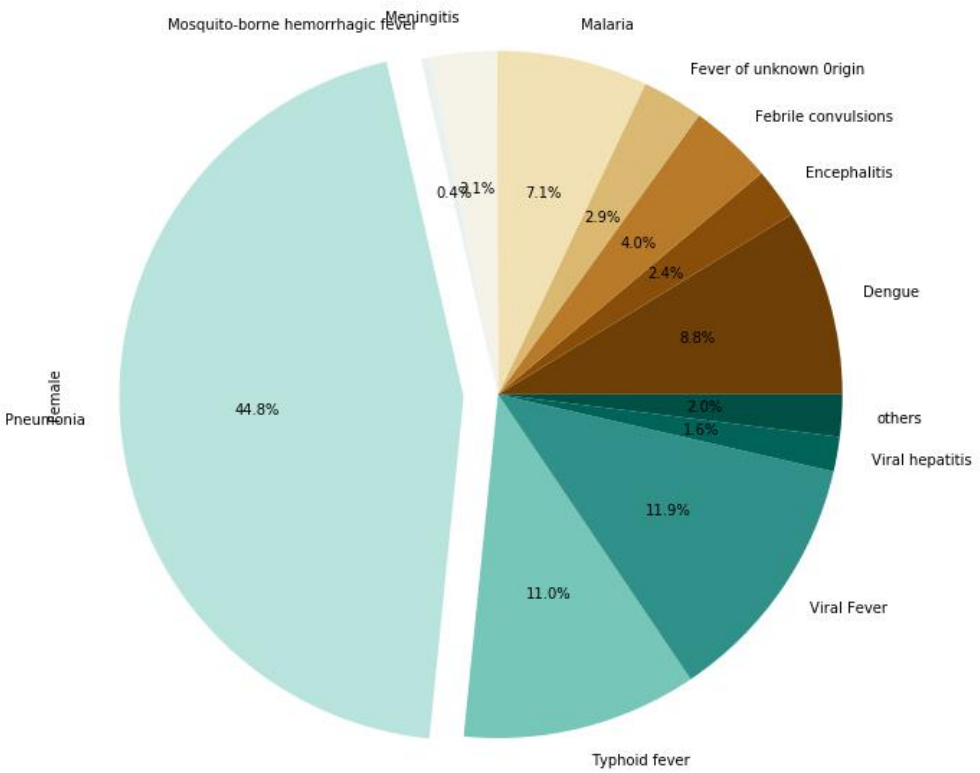
If we categorized disease on the bases of gender and try to closely examine occurrence of diseases in each gender it is reveal that the ratio of cases in female is slightly higher than male gender. For female gender Pneumonia is the most occurring disease with 44.8% of total cases record in the last 4 years. Dengue, Malaria Viral fever and Typhoid fever are the other notable disease. 11 and 12 % of cases are of Typhoid and viral fever whereas 8.8 and 7.1 % of cases recorded for Dengue and Malaria.

**Male:**

The analyses of male gender is also similar to the female as the ranking of the disease is same with Pneumonia is the outstanding disease following with viral and typhoid fever.

**Code Snippet:**

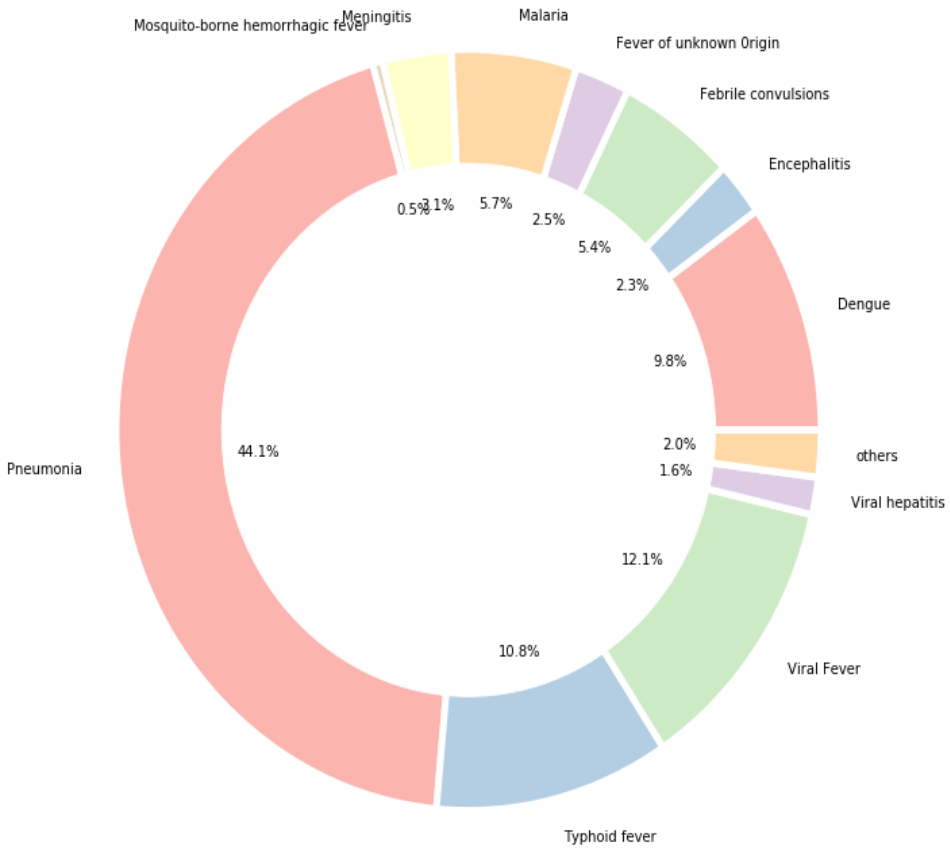
```
gender_disease = gender_disease.T
gender_disease['others'] =  gender_disease[['Acute febrile mucocutaneous lymph node syndrome [MCLS]' , 'Crimean hemorrhagic fever (CHF Congo virus)' , 'Maternal pyrexia' ,'Postprocedural fever','Rheumatic fever','Sinusitis' ,'Mosquito-borne fever/Chikungunya' , 'Fever of unknown origin (PUO)','Tetanus']].sum(axis=1)
gender_disease.drop(labels = ['Acute febrile mucocutaneous lymph node syndrome [MCLS]' , 'Crimean hemorrhagic fever (CHF Congo virus)' , 'Maternal pyrexia' ,'Postprocedural fever','Rheumatic fever','Sinusitis' ,'Mosquito-borne fever/Chikungunya','Fever of unknown origin (PUO)','Tetanus'],axis="columns", inplace=True)
gender_disease = gender_disease.T

explode = ( 0 ,0, 0, 0,0,0,0,0.1,0,0,0,0)
from matplotlib.colors import ListedColormap
import seaborn as sns
gender_disease['Female'].plot(kind='pie',figsize = (11,11),autopct='%1.1f%%' , explode = explode , shadow=False, colormap=ListedColormap(sns.color_palette("BrBG", 20)) )
```



**Code Snippet:**

```
from palettable.colorbrewer.qualitative import Pastel1_7
import matplotlib.pyplot as plt
fig = plt.figure(figsize=(12,12))
my_circle=plt.Circle( (0,0), 0.7, color='white')

plt.pie(gender_disease['Male'], labels=gender_disease['Male'].index, autopct='%1.1f%%',
colors=Pastel1_7.hex_colors , wedgeprops = { 'linewidth' :5, 'edgecolor' : 'white' })
p=plt.gcf()
p.gca().add_artist(my_circle)
plt.show()
```

## Below graph show Male gender vs disease factors.



## 2.4 Year vs Gender:

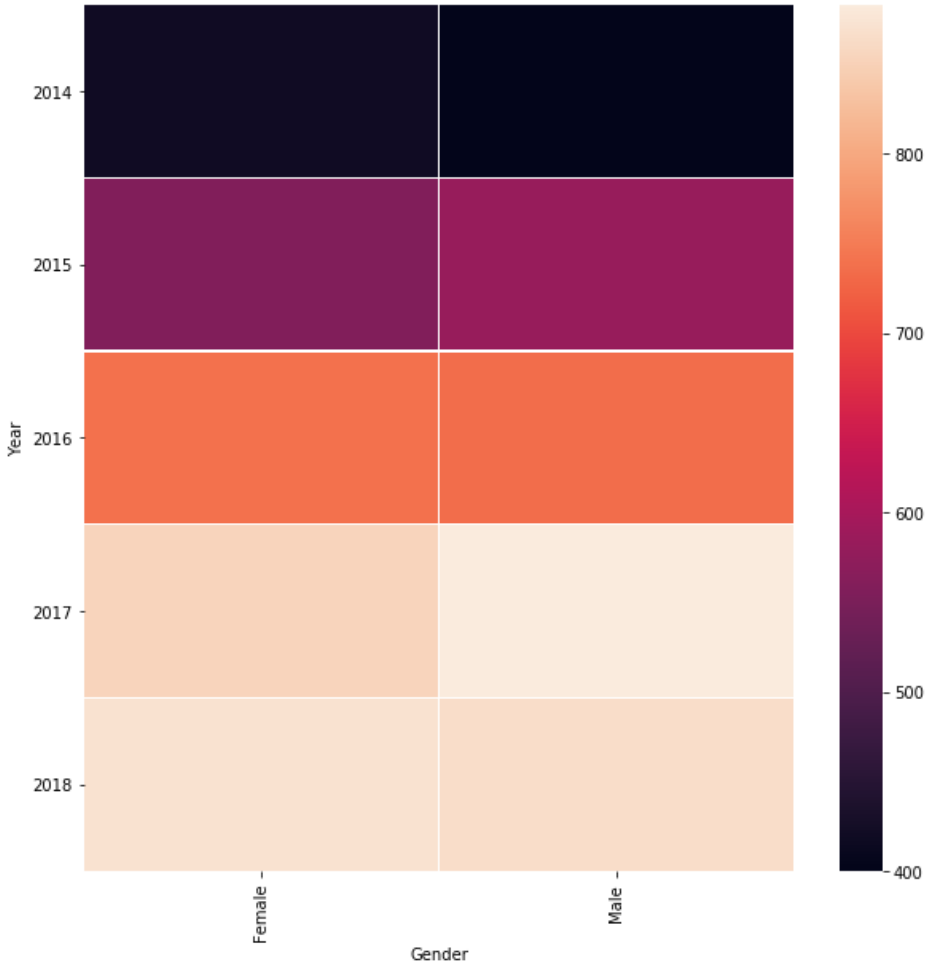
Year and Gender are other features which can be explore. The year feature is in numerical whereas gender is categorical. Crosstab is used to find the relationship between both variable and the results are visualize using heatmap. It is revealing that number of cases in female gender increases with increase with year and it reaches to the highest in the year 2018. Where as in male gender number of cases increases till 2017 but in 2018 there is a slight reduce in the ratio of cases. Year 2018 is the most critical year for female gender and if the ratio remains same then in 2019 the recorded cases for the female gender will be more than the previous years. For male gender the most critical year reported is 2017.

**Code Snippet:**

```
pd_crosstabY = pd.crosstab( df["Year"] , df['Gender'])
pd_crosstabY
fig = plt.figure(figsize=(10,10))
sns.heatmap(pd_crosstabY, cbar=True, linewidths=0.3)

# Rotate tick marks for visibility
plt.yticks(rotation=0)
plt.xticks(rotation=90)

plt.show()
pd_crosstabY[list(pd_crosstabY.columns)].plot(kind='bar')
```

**Below table provide gender vs year stats.**

| Gender | Female | Male |
|---|---|---|
| Year | | |
| **2014** | 420 | 400 |
| **2015** | 558 | 583 |
| **2016** | 739 | 734 |
| **2017** | 854 | 883 |
| **2018** | 870 | 866 |




## 2.5 Year vs Disease:

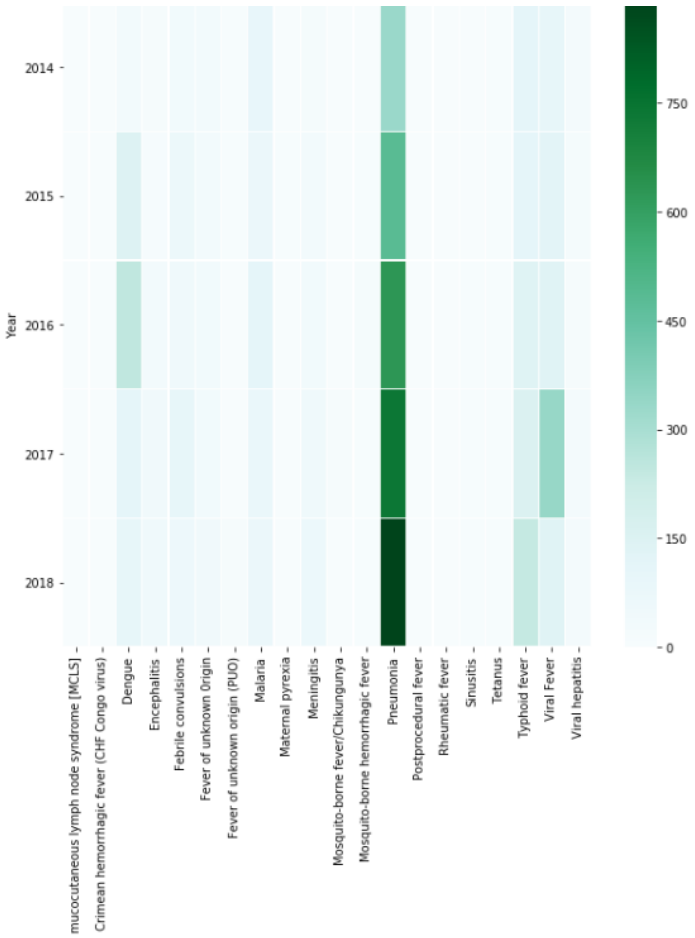
## Code Snippet:

```
pd_crosstabD = pd.crosstab( df["Year"] , df['Disease'])
pd_crosstabD
fig = plt.figure(figsize=(10,10))
sns.heatmap(pd_crosstabD, cbar=True, cmap="BuGn", linewidths=0.3)

# Rotate tick marks for visibility
plt.yticks(rotation=0)
plt.xticks(rotation=90)

plt.show()
```

## Below graph show year vs disease stats:

| Disease<br>Year | Acute febrile mucocutaneous lymph node syndrome [MCLS] | Crimean hemorrhagic fever (CHF Congo virus) | Dengue | Encephalitis | Febrile convulsions | Fever of unknown 0rigin | Fever of unknown origin (PUO) | Malaria | Maternal pyrexia | Meningitis | Mosquito-borne fever/Chikungunya | Mosquito-borne hemorrhagic fever |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2014** | 1 | 1 | 30 | 17 | 36 | 34 | 5 | 93 | 0 | 24 | 0 | 0 |
| **2015** | 0 | 0 | 149 | 11 | 72 | 32 | 8 | 77 | 5 | 34 | 0 | 4 |
| **2016** | 0 | 2 | 249 | 31 | 53 | 35 | 1 | 113 | 3 | 35 | 0 | 19 |
| **2017** | 1 | 0 | 109 | 48 | 101 | 38 | 3 | 83 | 2 | 55 | 14 | 4 |
| **2018** | 2 | 0 | 104 | 55 | 62 | 47 | 15 | 77 | 2 | 68 | 1 | 4 |

## Below graph show year vs disease visualization:



## 3. Model Creation:

After exploring the data different features and their combination is selected for model creation. First linear regression is applying on the year and disease features to predict the future possible count of cases recorded in coming years. Model predicted the possible value which is quite near to the actual values. Beside that combination of year and pneumonia disease data is also provided to the model to predict the criticality of disease in coming years. It is predicted by the model that the number of cases increases in coming years so urgent precautionary steps is required.

**Code Snippet:**

```
import numpy as np
from sklearn.linear_model import LinearRegression
x = np.array(year_disease_m['Year']).reshape((-1, 1))
y = np.array(year_disease_m['Summarize'])

model = LinearRegression()
model.fit(x, y)
model = LinearRegression().fit(x, y)
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
print('intercept:', model.intercept_)
print('slope:', model.coef_)
y_pred = model.predict(x)
print('predicted response:', y_pred, sep='\n')
y_pred_eq = model.intercept_ + model.coef_ * x
print('predicted response:', y_pred_eq, sep='\n')
x_new = np.array([2019,2020,2021,2022,2023]).reshape((-1, 1))
y_new = model.predict(x_new)
print("new value predictions" ,y_new)
```

Model 1: Year Vs Disease count

```
coefficient of determination: 0.9310369241642511
intercept: -488103.4000000002
slope: [243.8]
predicted response:
[2909.8 3153.6 3397.4 3641.2 3885. ]
predicted response:
[[2909.8]
 [3153.6]
 [3397.4]
 [3641.2]
 [3885. ]]
new value predictions [4128.8 4372.6 4616.4 4860.2 5104. ]
```

**Code Snippet:**

```
x = np.array(year_disease_m['Year']).reshape((-1, 1))
y = np.array(year_disease_m['Pneumonia'])

model = LinearRegression()
model.fit(x, y)
model = LinearRegression().fit(x, y)
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
print('intercept:', model.intercept_)
print('slope:', model.coef_)
y_pred = model.predict(x)
print('predicted response:', y_pred, sep='\n')
y_pred_eq = model.intercept_ + model.coef_ * x
print('predicted response:', y_pred_eq, sep='\n')
x_new = np.array([2019,2020,2021,2022,2023]).reshape((-1, 1))
y_new = model.predict(x_new)
print("new data predicted value" , y_new)
```

Model 2: Year Vs Pneumonia

```
coefficient of determination: 0.9972345203790212
intercept: -274569.60000000003
slope: [136.5]
predicted response:
[341.4 477.9 614.4 750.9 887.4]
predicted response:
[[341.4]
 [477.9]
 [614.4]
 [750.9]
 [887.4]]
new data predicted value [1023.9 1160.4 1296.9 1433.4 1569.9]
```

## 4. Conclusion and Results:

Analyzing patient records and finding hidden factors and insights is very important as it can help improve the health issues. Analyses perform on this data set reveals that number of diseases are increasing with respect to age and year. Pneumonia disease is the most critical one having very high ratio. It is suggested to give special attention in order to reduce the cases in coming years. It is also revealing that viral and typhoid fever is also gradually increasing in coming years. So, steps should be taken from now so that it will not become a health threat in coming years.