# Analysis of the US weekly Nationally Notifiable Disease Surveillance Data

Submitted by:

VINUTHNA AMIREDDY

Marticulation number: 12301006

Email: vinuthna.amireddy@stud.th-deg.de

## ** Overview**

The National Notifiable Disease Surveillance System (NNDSS) is a nationwide collaboration that enables all levels of public health to share health information to monitor, control, and prevent the occurrence and spread of state-reportable and nationally notifiable infectious and some noninfectious diseases and conditions [1]. The level 2 data contains 50 diseases 50 US states and 1284 US cities. From this dataset, we had to pick a disease for analysis. We choose Tuberculosis as this disease was amongst the diseases with highest cases and mortality. We performed a statewise analysis of the Tuberculosis.For the analysis of this disease, we selected top-6 states with highest cases of Tuberculosis. We analyzed long-term transmissions trends of Tuberculosis in each state and how they have been fluctuating in USA over the time period of 1906-1927.Also, to get a better understanding, we collected the population data for all US states between years 1900-1927 and did the analysis of impact of Tuberculosis in each state per 1000 population.

Keywords:epi_week, Diseases,Tuberculosis,US states, Cases,Deaths,Mortality,Population.

## ** Abbrivations**

AL : Alabama
AK : Alaska
AZ : Arizona
AR : Arkansas
CA : California
CO : Colorado
CT : Connecticut
DE : Delaware
FL : Florida
GA : Georgia
HI : Hawaii
ID : Idaho

IL : Illinois
IN : Indiana
IA : Iowa
KS : Kansas
KY : Kentucky
LA : Louisiana
ME : Maine
MD : Maryland
MA : Massachusetts
MI : Michigan
MN : Minnesota
MS : Mississippi
MO : Missouri
MT : Montana
NE : Nebraska
NV : Nevada
NH : New Hampshire NJ : New Jersey
NM : New Mexico
NY : New York
NC : North Carolina ND : North Dakota
OH : Ohio
OK : Oklahoma OR : Oregon
PA : Pennsylvania
RI : Rhod Island
SC : South Carolina SD : South Dakota TN : Tennessee TX : Texas UT : Utah VT : Vermont VA : Virginia
WA : Washington WV : West Virginia
WI : Wisconsin
WY : Wyoming TB : Tuberculosis USA : United States of America EDA : Exploratory Data Analysis

# 1. Introduction

A disease is a health condition that has a specific set of symptoms and traits, which negatively affect the well being of human life. They are often known to be medical conditions that associate with specific signs and symptoms. In our project we are going to analyze a large data set that contains the details of diseases which occurred in all states of the U.S during 1888 to 2014. Source of this data is from Project Tycho who work with national and global health institutes and researchers to make data easier to use to improve global health. In our data set Event is a column which provides info regarding the dataset, if that dataset provides information on number of cases or number of deaths. Disease epidemiology is complex because of environmental influences in each changing season.Tuberculosis is a disease caused by germs that are spread from person to person through the air. By the beginning of the 19th century, tuberculosis, or "consumption," had killed one in seven of all people that had ever lived[4] . TB germs are put into the air when a person with TB disease of the lungs or throat coughs, sneezes, speaks, or sings. These germs can stay in the air for several hours, depending on the environment.BCG is a vaccine for TB disease. BCG is used in many countries, but it is not generally recommended in the United States[2]. This may be the reason that mortality rate for TB in US to be very high.In this report we will examine the trends of TB disease cases in the United States during 1906–1927 also by considering population.

# 2 Problem definition

The aim of this assignment is to analyze the data provided by the "National Notifiable Disease Surveillance System" hosted on Project Tycho's website for the effects in each state from the years 1888-2013. We Look

into the diseases with highest number of cases, deaths and mortality. Further we do statewise and year wise analysis of Tuberculosis and analyse the patterns regarding number of cases, deaths and mortality for Tuberculosis. We also take into account the population for statewise analysis of the Tuberculosis. The population data is created by us by retrieving the individual year population of each state in US from "Macro Trends Website"[3].

# 3. Objectives

To understand disease trends in in US while focusing on:

-Most contagious and deadly diseases -Impact on states -Seasonal trends - population weise impacts of diseases

# 4. Methods

We used following plotting techniques:

Tree Map : In a treemap hierarchical data is displayed as a set of nested rectangles. Each group is represented by a rectangle with size of rectangle proportional to value.

US Geom Map : Used to display the data in the Map of US.

Bar plot : A bar plot is a plot that presents categorical data with rectangular bars with heights proportional to the numerical values that they represent.

Geom point: It is used to plot points on a graph to show relation between multiple categorical and numerical values.

Heat Map : A heat map is a two-dimensional representation of data which represents values as intensities of colors.

# 5. Analysis Protocol

The analysis was divided into subsections. The working steps are introduced below.

Before to start the analysis the required libraries were loaded

## 5.1 Data loading and cleanup

Data loading and Cleanup are fundamental and indispensable part of data analysis and be the first steps in EDA. In this step, data is loaded and processed via identification and modification/removal of incomplete, irrelevant, or missing data.

**Data loading** The full load method was used to load data. The data source was Project Tycho Level 2 Data (Source: https://www.nejm.org/doi/full/10.1056/NEJMms1215400). The data covered 50 diseases 50 US states and 1284 US cities in the period 1888 to 2013. Also a population data was created containing population of US states for years between 1900-1927

```
populationDF <- read_excel("C:/Users/DELL/OneDrive/Desktop/DV/ProjectTycho_Level2_v1.1.0_0/StatewiseUSp
populationDF <-as.data.frame(populationDF)
populationDF <- setNames(cbind(populationDF[1],stack(populationDF[2:29])),c("state","population","year"
tychoDF <- read.csv("C:/Users/DELL/Downloads/ProjectTycho_Level2_v1.1.0_0 (1)/ProjectTycho_Level2_v1.1.
```

### Data cleanup

We modified the names of diseases 'Tuberculosis [Phthisis Pulmonalis]'] to 'Tuberculosis', 'Typhoid Fever [Enteric Fever]'] to 'Typhoid Fever', 'Chickenpox [Varicella]'] to 'Chickenpox', 'Whooping Cough [Pertussis]'] to 'Whooping Cough' for easier visualization without losing any meaningful information

```
tychoDF$disease[tychoDF$disease == 'TUBERCULOSIS [PHTHISIS PULMONALIS]'] <- 'TUBERCULOSIS'

tychoDF$disease[tychoDF$disease == 'TYPHOID FEVER [ENTERIC FEVER]'] <- 'TYPHOID FEVER'

tychoDF$disease[tychoDF$disease == 'CHICKENPOX [VARICELLA]'] <- 'CHICKENPOX'

tychoDF$disease[tychoDF$disease == 'WHOOPING COUGH [PERTUSSIS]'] <- 'WHOOPING COUGH'

tychoDF$disease[tychoDF$disease == 'BRUCELLOSIS [UNDULANT FEVER]'] <- 'BRUCELLOSIS'
```

Year and month was extracted.A temporary datafame for months was created as most of the analysis will be in years.

```
colnames(tychoDF)[9] <- "year"
tychoDF_month<-tychoDF

tychoDF$year<-format(as.POSIXlt(tychoDF$year, format="%Y-%m-%d"),"%Y")

tychoDF_month$year<-format(as.POSIXlt(tychoDF_month$year, format="%Y-%m-%d"),"%m")

colnames(tychoDF_month)[9] <- "month"
```

Irrelevant to the analysis columns were also removed.

```
tychoDF <- tychoDF[ , -11]
tychoDF <- tychoDF[ , -10]
tychoDF <- tychoDF[ , -5]
tychoDF <- tychoDF[ , -4]

tychoDF <- tychoDF[ , -2]
tychoDF <- tychoDF[ , -1]
```

We delete the rows where the number of events are not greater than 0

```
tychoDF <- tychoDF %>% filter(number > 0 )
```

NA values were checked for. Since no NA was present, we proceeded

```
nacount_year<-tychoDF %>% summarise(across(everything(), ~ sum(is.na(.))))
nacount_month<-tychoDF %>% summarise(across(everything(), ~ sum(is.na(.))))
```

Before going to the next step, the structure of the dataset was checked.

```
str(tychoDF)
```

```
## 'data.frame':    2678605 obs. of  5 variables:
##  $ state  : chr  "PA" "PA" "PA" "PA" ...
##  $ disease: chr  "TYPHOID FEVER" "SCARLET FEVER" "DIPHTHERIA" "TYPHOID FEVER" ...
##  $ event  : chr  "DEATHS" "DEATHS" "DEATHS" "DEATHS" ...
##  $ number : int  14 4 4 12 5 7 4 1 1 4 ...
##  $ year   : chr  "1888" "1888" "1888" "1888" ...
```

```
str(populationDF)
```

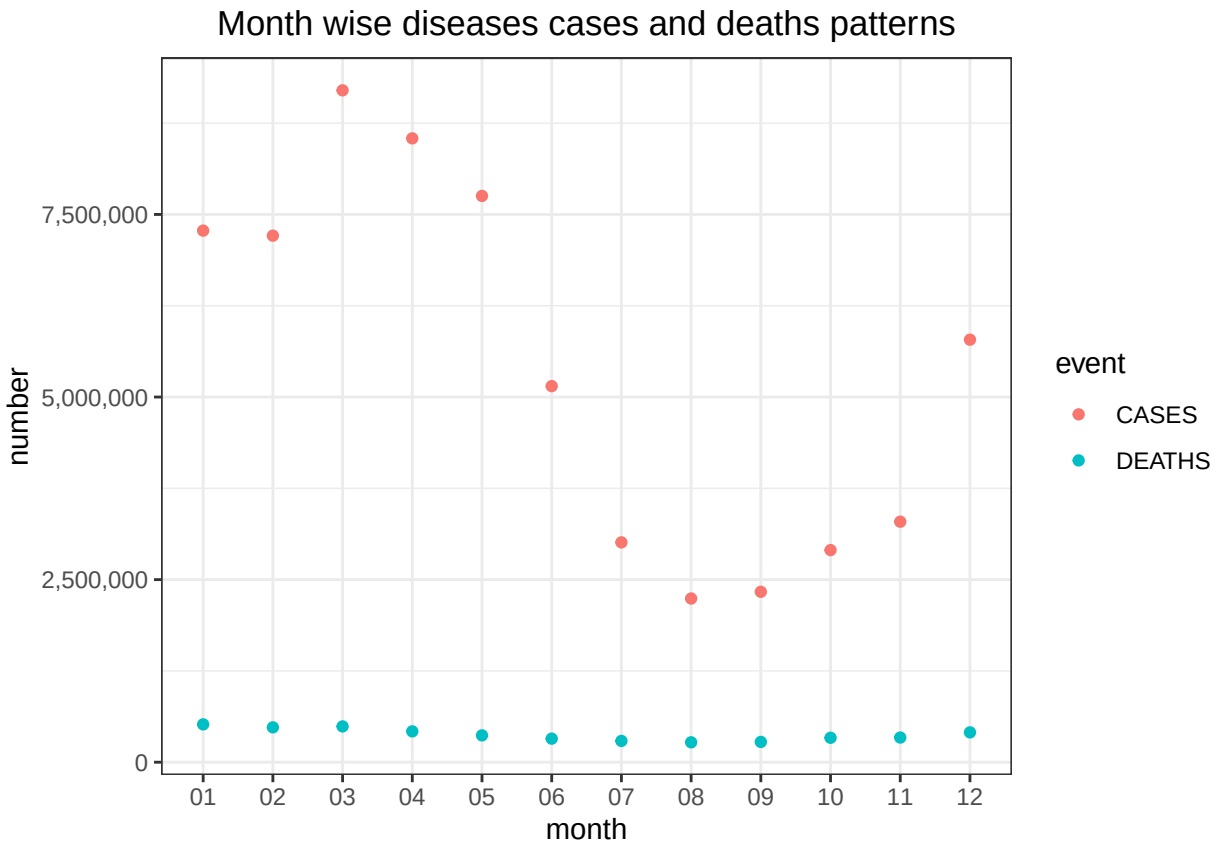
```
## 'data.frame':    1288 obs. of  3 variables:
##  $ state     : chr  "CA" "TX" "FL" "NY" ...
##  $ population: num  1490000 3055000 530000 7283000 6313000 ...
##  $ year      : Factor w/ 28 levels "1900","1901",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The Tycho dataset contains 3659360 observations of 5 variables, representing state, disease, event(Death/Cases),Number of events and year. Population dataset contains 1288 observations of 3 variables,representing state, population,year.

## 5.2 Analyzing the impact of time of the year on the diseases

The graph shows a high spread of the diseases from the month of September till March. The increase in number of cases is accompanied by increase in number of deaths. From March, we witness a decline in number of cases and deaths which goes in till August. The higher number of cases may be attributed to winter as the period from Sep-Oct to Feb-Mar have lower temperatures. In summers diseases tend to be less contagious.

```
tychoDF_month<-tychoDF_month%>%
aggregate(number~month+event,sum)

tychoDF_month%>%
ggplot(aes(month,number))+
geom_point(aes(colour=event))+
scale_y_continuous(labels = scales::comma)+
  theme_bw()+
  ggtitle("Month wise diseases cases and deaths patterns")+
  theme(plot.title = element_text(hjust = 0.5))
```




### 5.3 Visualizing states with highest events

Before visualizing which states have highest cases and deaths we created a data frame named States to see sum for all cases and sum of all deaths in each state in a tabular form, and displayed top 5 rows using head() function as shown below:

```
State <- tychoDF %>% count(state,event, wt = number)
State <- State %>% pivot_wider(
  names_from = event,
  values_from = n,
  values_fill = 0,
  values_fn = list(breaks = mean))
head(State)
```

```
## # A tibble: 6 x 3
##   state   CASES DEATHS
##   <chr>   <int>  <int>
## 1 AK      66768      0
## 2 AL     939144  58830
## 3 AR     721176  14110
## 4 AS        231      0
## 5 AZ     560585  29298
## 6 CA    4306937 311252
```

code below filters the data where only cases and deaths are reported respectively in each separate data frames. An then it counts the total number of cases and deaths caused by all diseases in respective states in U.S in order to visualize it in a heatmap.

```
State_cases_count <- filter (tychoDF, event == "CASES")
State_cases_count <- State_cases_count %>%  group_by(state, disease) %>% summarise(number = sum(number))
```

```
## `summarise()` has grouped output by 'state'. You can override using the
## `.groups` argument.
```

```
head(State_cases_count)
```

```
## # A tibble: 6 x 3
## # Groups:   state [1]
##   state disease           number
##   <chr> <chr>              <int>
## 1 AK    BRUCELLOSIS           23
## 2 AK    CHICKENPOX          4503
## 3 AK    CHLAMYDIA          16122
## 4 AK    CRYPTOSPORIDIOSIS     20
## 5 AK    DIPHTHERIA            97
## 6 AK    GIARDIASIS           442
```

```
State_death_count <- filter (tychoDF, event == "DEATHS")
State_death_count <- State_death_count %>%  group_by(state, disease) %>% summarise(number = sum(number))
```

```
## `summarise()` has grouped output by 'state'. You can override using the
## `.groups` argument.
```

```
head(State_death_count)
```

```
## # A tibble: 6 x 3
## # Groups:   state [1]
##   state disease     number
##   <chr> <chr>        <int>
## 1 AL    BRUCELLOSIS    125
## 2 AL    CHICKENPOX      47
## 3 AL    DIPHTHERIA     222
## 4 AL    INFLUENZA     3640
## 5 AL    MENINGITIS     102
## 6 AL    PELLAGRA       330
```

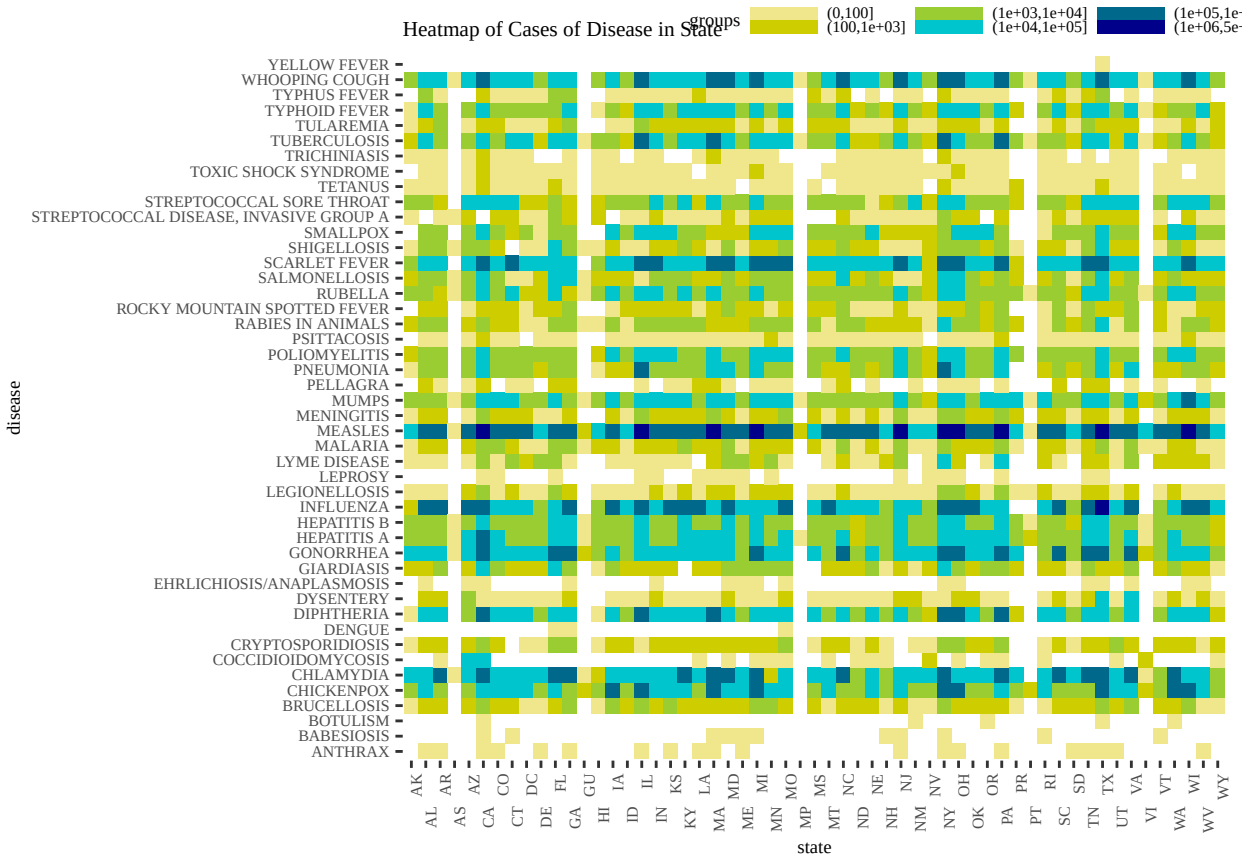
adjust the data so that it groups the data according to certain interval so that it can be represented as a heat map, one for deaths and other for cases respectively.

```
State_death_count$groups <- cut(State_death_count$number, breaks = c(0,10,100,1000,10000,100000,500000))
```

```
State_cases_count$groups <- cut(State_cases_count$number, breaks = c(0,100,1000,10000,100000,1000000,50
```

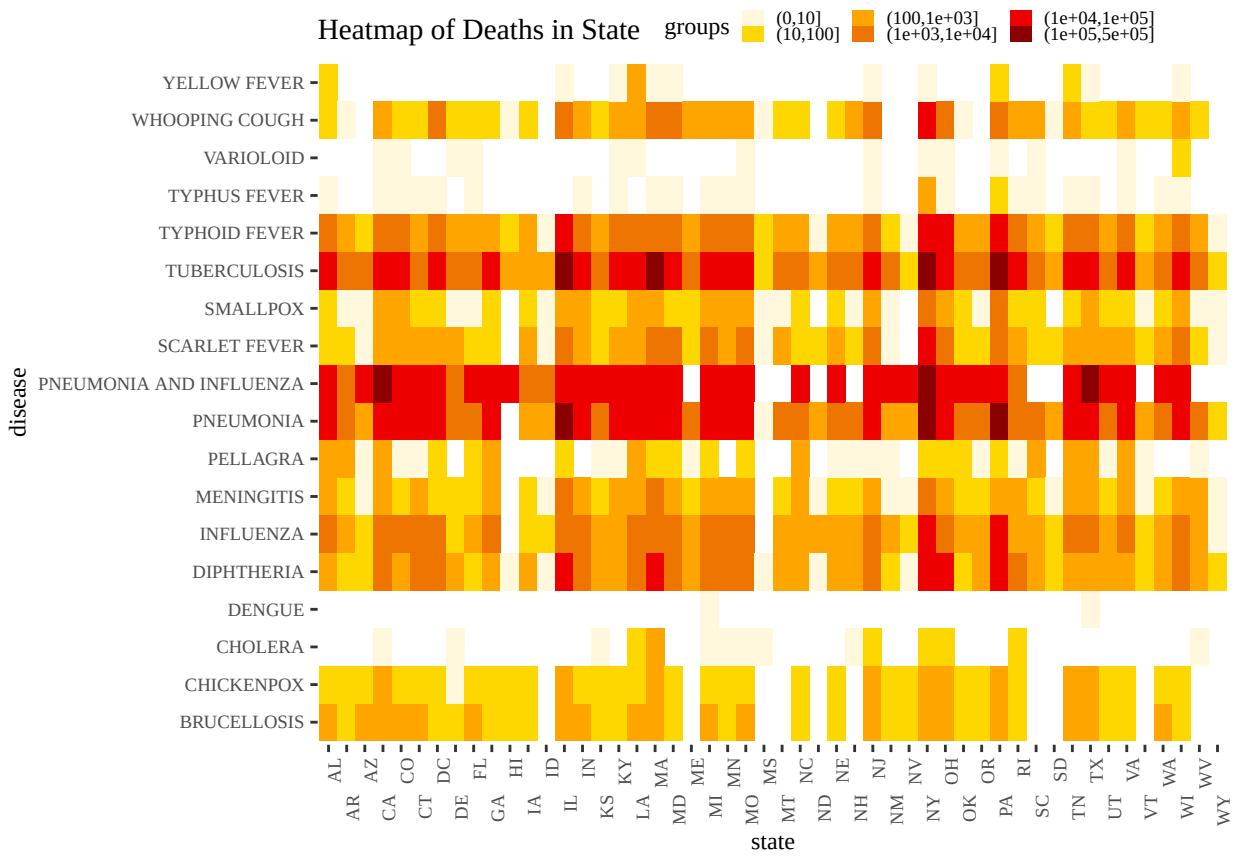
Now, in order to visualize which state has more number of cases and deaths per disease, geom_tile() function in ggplot2 is used to create a heatmap as follows:

```
ggplot(State_cases_count, aes(state, disease )) +
  geom_tile(aes(fill= groups ) ) +
  scale_fill_manual(breaks = levels(State_cases_count$groups),  values = c("khaki","yellow3","olivedrab
  theme_tufte() +
  scale_x_discrete(guide = guide_axis(n.dodge=2))+
  theme(text = element_text(size = 7),axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(legend.key.width = unit(0.9, "cm"), legend.key.height = unit(0.1, "cm"),legend.direction = 'hor
  ggtitle("Heatmap of Cases of Disease in State" )
```



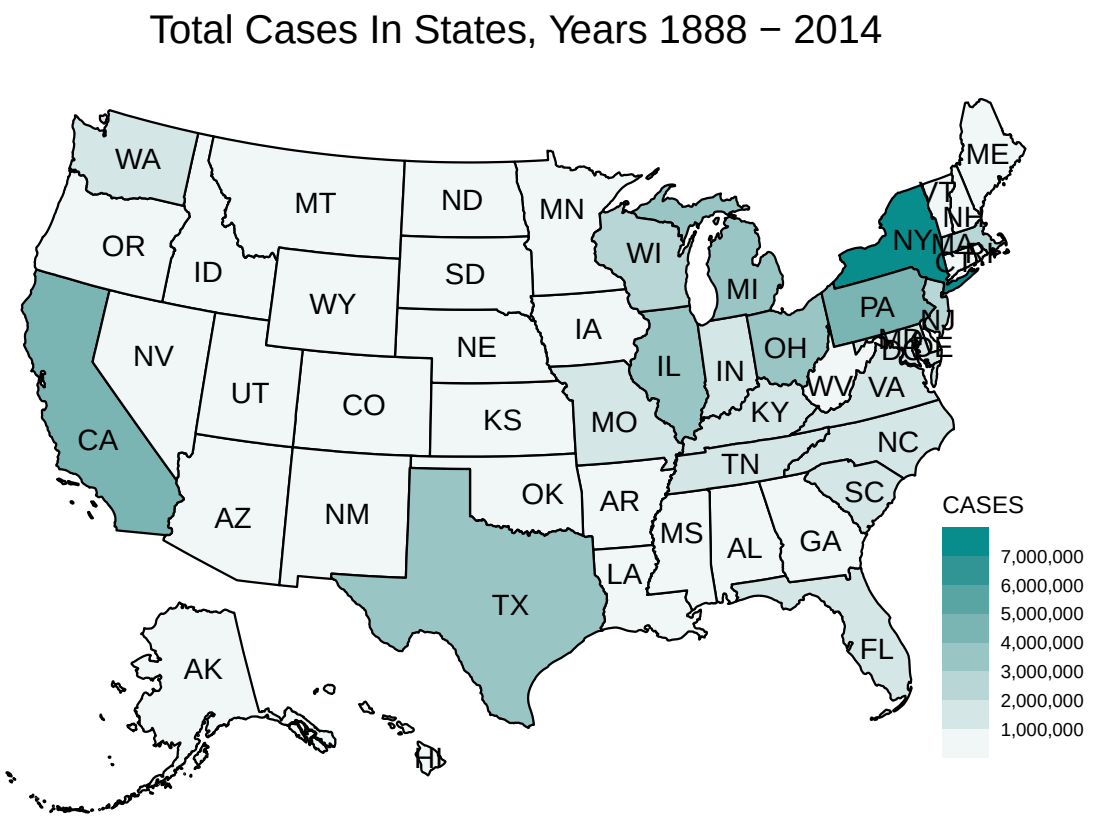

White space shows there are no cases for particular disease for related states.

```
ggplot(State_death_count, aes(state,disease)) +
  geom_tile(aes(fill= groups)) + # Border color
  scale_fill_manual(breaks = levels(State_death_count$groups), values = c("cornsilk","gold","orange","d
  theme_tufte() +
  scale_x_discrete(guide = guide_axis(n.dodge=2))+
  theme(text = element_text(size = 9),axis.text.x = element_text(angle = 90, hjust = 1)) +
 theme(legend.key.width = unit(0.3, "cm"), legend.key.height = unit(0.2, "cm"),legend.direction = 'hori
  ggtitle("Heatmap of Deaths in State" )
```

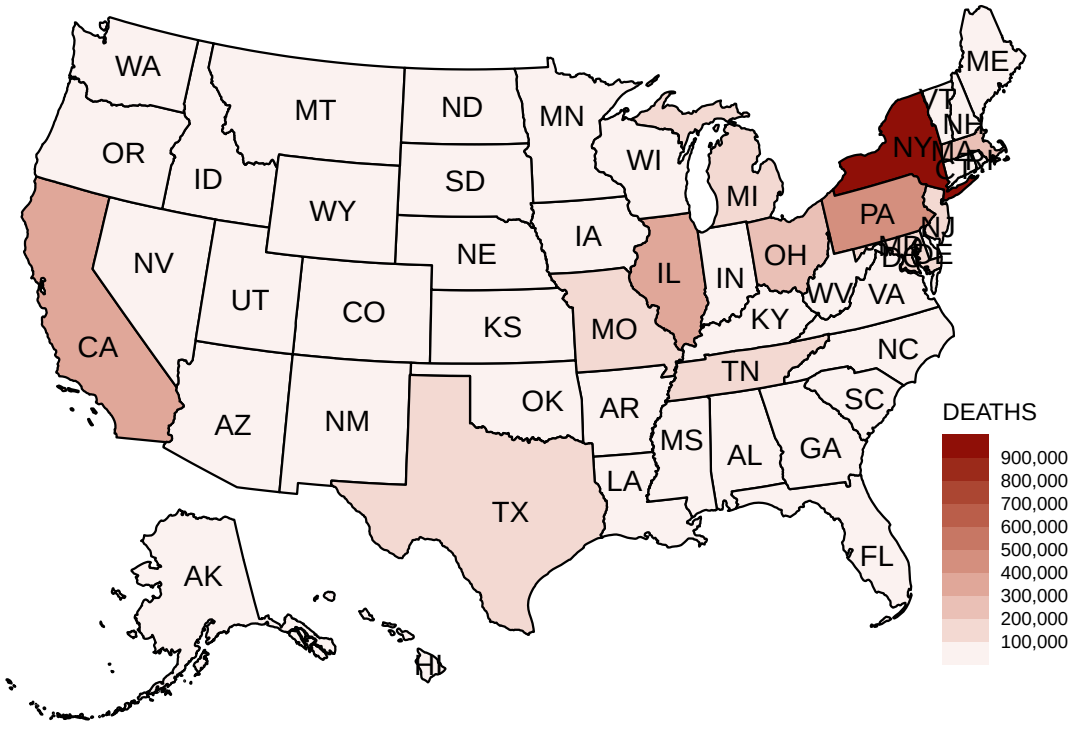



As the heatmap we got above is not clear to understand which states have highest number of cases and deaths respectively, we plotted a geographic map plot of US states to have a better visualization of number of cases per each state using plot_usmap function in ggplot2

```
plot_usmap(data = State,values = "CASES",labels = TRUE) +
  scale_fill_stepsn(n.breaks = 9, label=comma, colors = c("white","cyan4")) +
  ggtitle("Total Cases In States, Years 1888 - 2014") +
  theme(plot.title = element_text(size = 15, hjust = 0.5), legend.position = c(0.87, 0.1))
```




```
plot_usmap(data = State, values = "DEATHS",labels = TRUE) +
  scale_fill_stepsn(n.breaks = 9,label=comma, colors = c("white","darkred")) +
  ggtitle("Total Deaths In States, Years 1888 - 2014") +
  theme(plot.title = element_text(size = 15, hjust = 0.5), legend.position = c(0.87, 0.1))
```

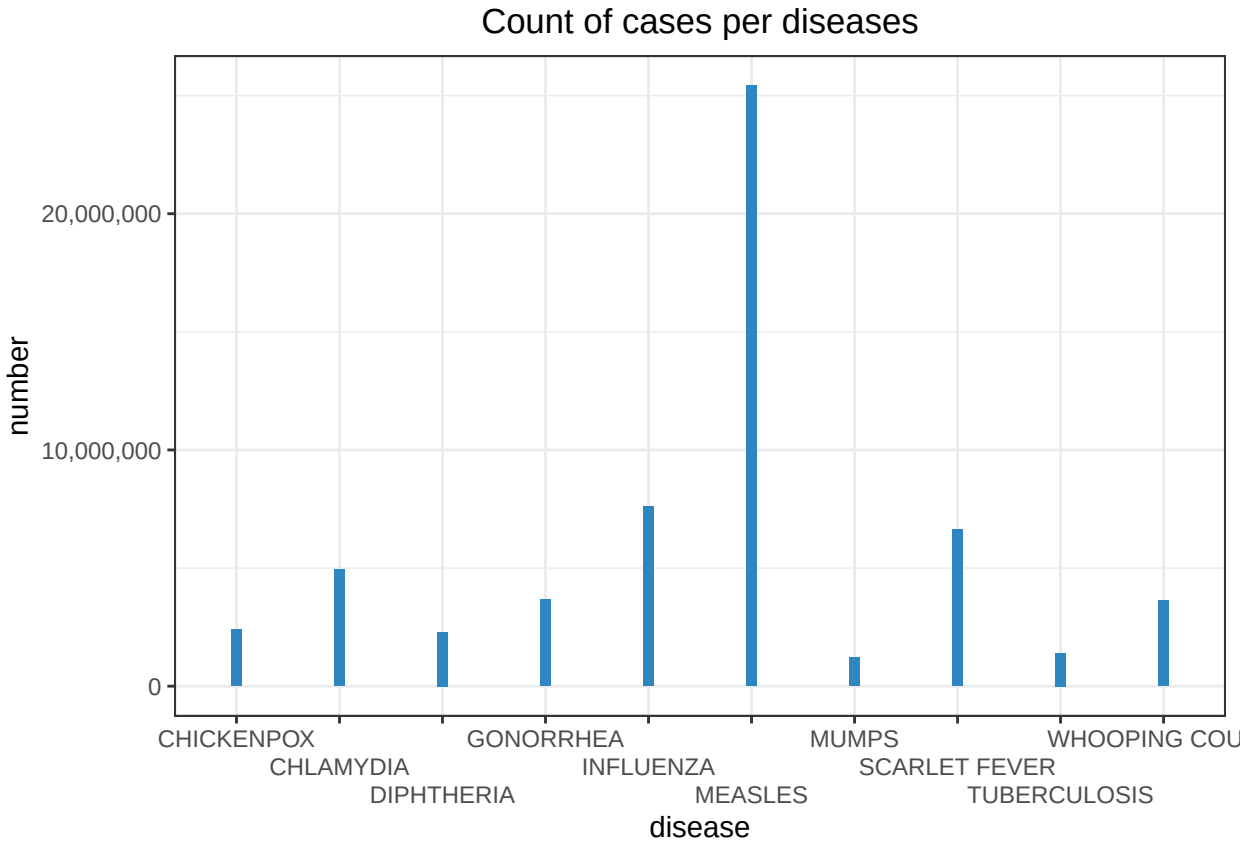



We observed that New York recorded highest number of deaths followed by Illinois and Pennsylvania compared to any other state. This is in line with number of cases as number of deaths is directly proportional to number of cases.

### 5.4 Number of cases by disease

In this section we try to analyse which are the top 10 most contagious diseases. We group data by disease. THe resultant dataframe is printed along with bar graph. As we can see, Measeles had highest number of cases followed by INFLUENZA,SCARLET FEVER,CHLAMYDIA,GONORRHEA,WHOOPING COUGH,CHICKENPOX,DIPHTHERIA,TUBERCULOSIS,MUMPS.

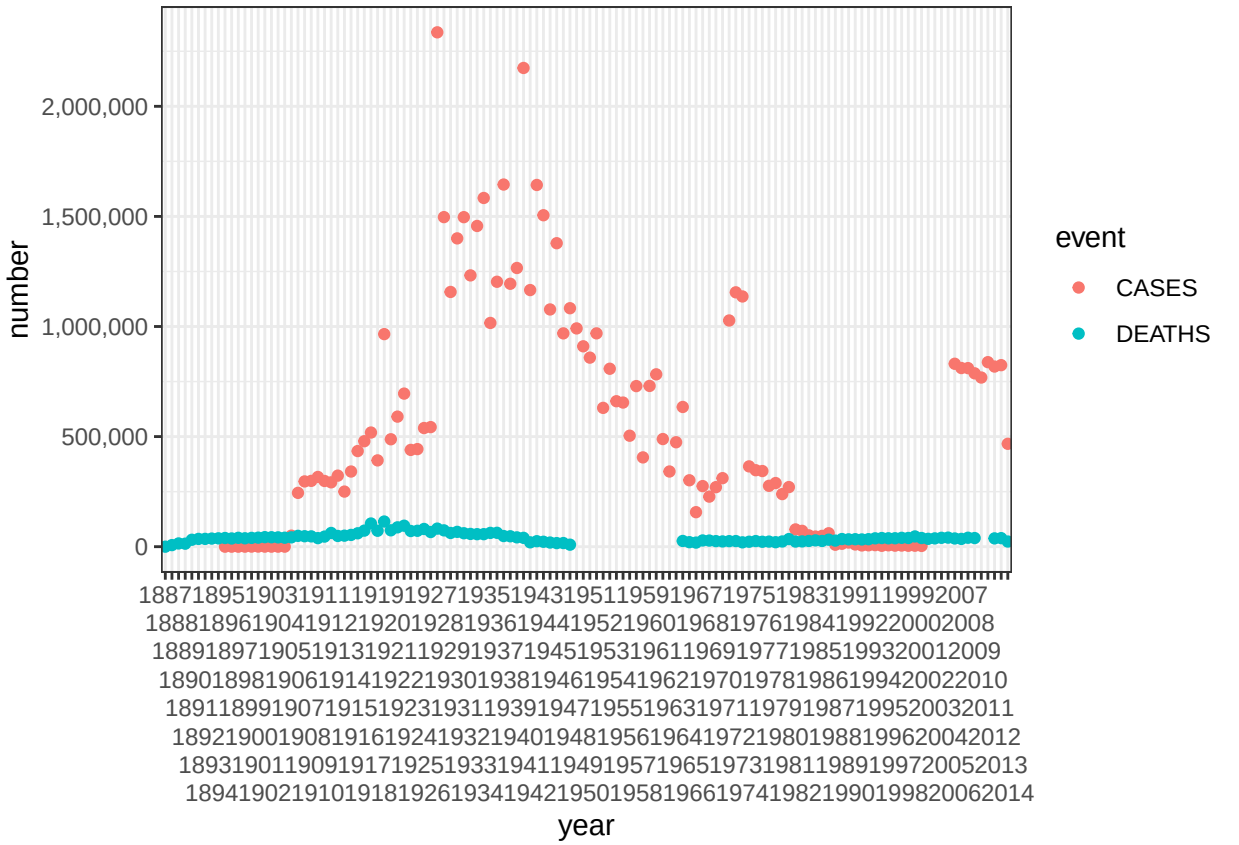
```
tychoDF_cases<-tychoDF%>%filter(grepl('CASES',event,ignore.case=TRUE))%>%aggregate(number~disease,sum)

tychoDF_cases%>%
  arrange(desc(number)) %>%
  slice_head(n = 10) %>%
  ggplot(aes(disease,number))+
  geom_bar(stat="identity",width=c(0.1),fill="#2E86C1")+
  scale_y_continuous(labels = scales::comma)+
  scale_x_discrete(guide = guide_axis(n.dodge=3))+
  theme_bw()+
  ggtitle("Count of cases per diseases")+
  theme(plot.title = element_text(hjust = 0.5))
```




### 5.5 Year wise analysis of events

We group data by year and event to find the pattern of number of instances of cases and deaths from 1888 to 2014. As is visible, for many years, there is no data for either death or cases.There is no cases data before year 1906 and no death data between 1949-1964. So we remove the data belonging to that years.

```
tychoDF_year_event<-tychoDF%>%
aggregate(number~year+event,sum)

tychoDF_year_event%>%
ggplot(aes(year,number))+
geom_point(aes(colour=event))+
scale_y_continuous(labels = scales::comma)+
scale_x_discrete(guide = guide_axis(n.dodge=8))+
  theme_bw()
```



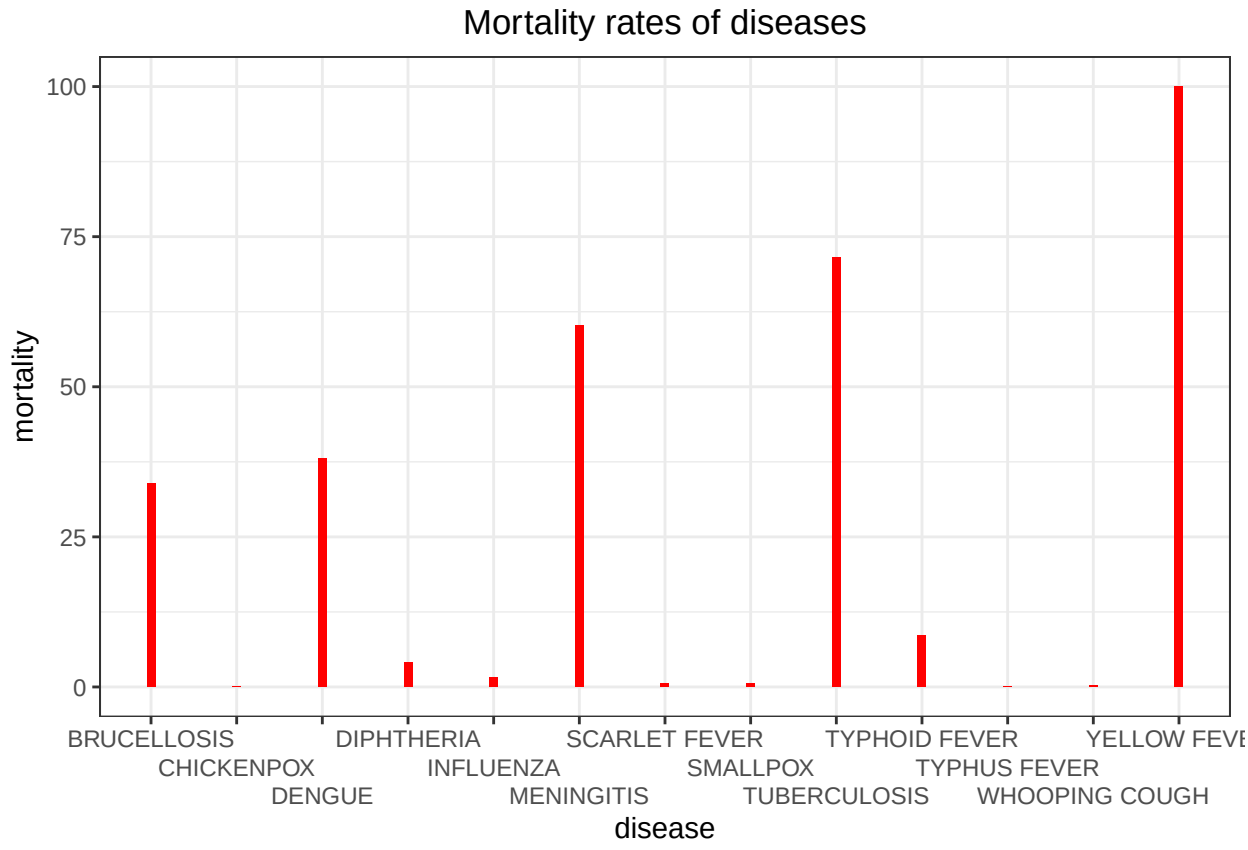

```
tychoDF<-tychoDF%>%
filter(year>1906,year<1949|year>1965)
```

### 5.6 Finding the diseases with the highest mortality

Now we try to find the diseases with the highest mortality (deaths per 100 cases) to find the most deadly diseases. For this we create a new dataset that has diseases and their mortalities. SInce so many disease don't have death data and 1 also has no cases data. We only have 15 diseases with both death and cases data.Also 2 diseases have mortality higher than 100 which is not possible. We filter them as well, We plot the mortality graph for remaining 13 diseases. As it can be observed Yellow fever has unusually high mortality rate of 100%(Yellow fever generally has a mortality between 20% and 50) followed by Tuberculosis, Meningitis and Dengue. Given that Tuberculosis is amongst the top 10 most contagious disease and the disease with second highest mortality amongst the diseases which have data for both cases and deaths, we choose Tuberculosis for further analysis.

```
tychoDF_groupby_event_disease<-tychoDF%>%
aggregate(number~disease+event,sum)

tychoDF_disease_deaths<-tychoDF_groupby_event_disease%>%filter(grepl('DEATHS',event,ignore.case=TRUE))

tychoDF_disease_deaths<-setNames(tychoDF_disease_deaths,c("disease","death_event","death_number"))


tychoDF_disease_cases<-tychoDF_groupby_event_disease%>%filter(grepl('CASES',event,ignore.case=TRUE))
```

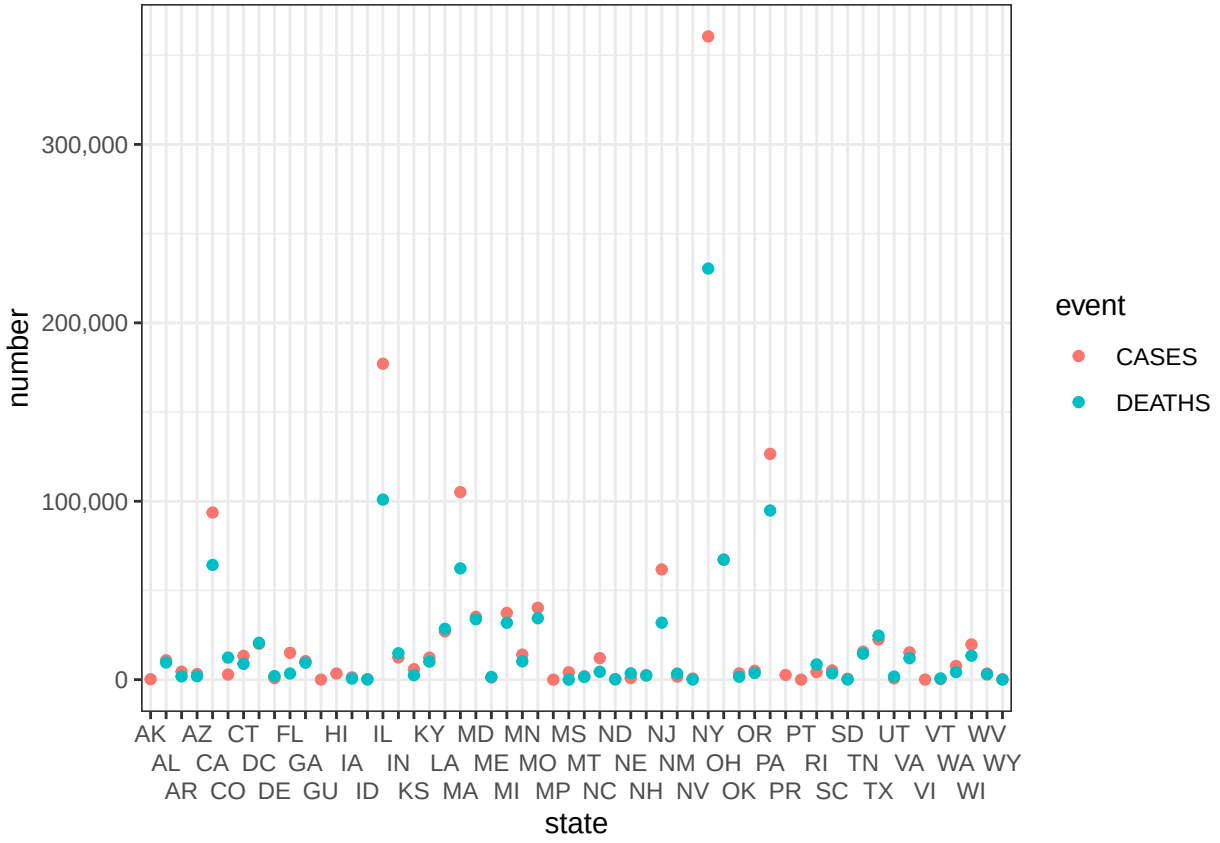
```r
tychoDF_disease_cases<-setNames(tychoDF_disease_cases,c("disease","cases_event","cases_number"))

tychoDF_merge_events<-merge(x = tychoDF_disease_deaths, y = tychoDF_disease_cases, by = "disease")

tychoDF_mortality <- tychoDF_merge_events[c('disease','cases_number')]
tychoDF_mortality<-setNames(tychoDF_mortality,c("disease","mortality"))


tychoDF_mortality$mortality<-(tychoDF_merge_events$death_number/tychoDF_merge_events$cases_number)*100

tychoDF_mortality<-tychoDF_mortality%>%filter(mortality<=100)


tychoDF_mortality%>%
  ggplot(aes(disease,mortality))+
  geom_bar(stat="identity",width=c(0.1),fill="#ff0000")+
  scale_y_continuous(labels = scales::comma)+
  scale_x_discrete(guide = guide_axis(n.dodge=3))+
  theme_bw()+
  ggtitle("Mortality rates of diseases")+
  theme(plot.title = element_text(hjust = 0.5))
```



## 5.7 Understanding the impact of Tuberculosis state wise

We analyse the statewise impact of tuberculosis.As we can see CO has more deaths then cases which is not possible. ALso few states don't have any data either on deaths or on cases for Tuberculosis. So we pick NY,IL,MA, PA, CA, OH for statewise comparison

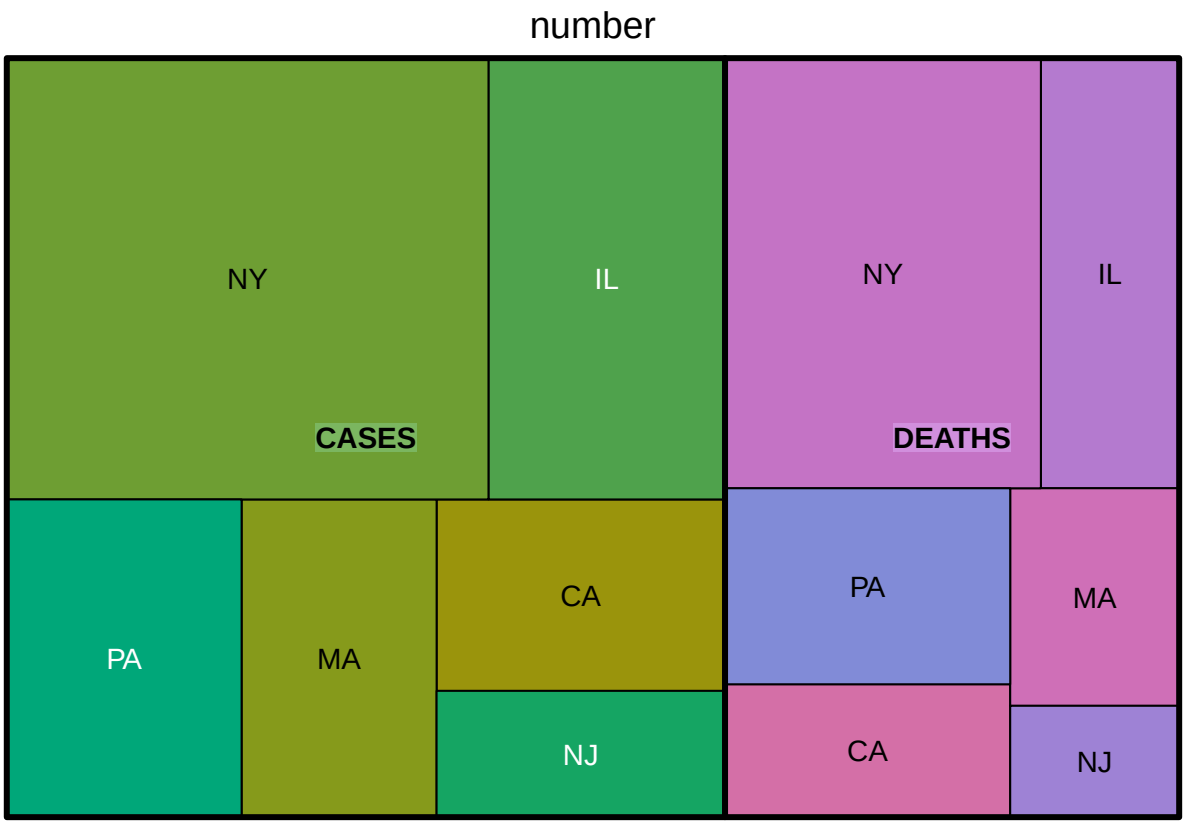
```
tychoDF_TB<-tychoDF%>%
filter(grepl('TUBERCULOSIS',disease,ignore.case=TRUE))%>%
aggregate(number~state+event,sum)

tychoDF_TB%>%
ggplot(aes(state,number))+
geom_point(aes(colour=event))+
scale_y_continuous(labels = scales::comma)+
scale_x_discrete(guide = guide_axis(n.dodge=3))+
  theme_bw()
```




## 5.8 Comparison for Tuberculosis cases between selected states

We group the TB data by state event and year.As We can see, New York has highest number of Cases followed by Illinois, Pennsylvania and Massachusetts. New York had most deaths followed by Illinois, Pennsylvania and California. Further we analyse year wise patterns for these states

```
tychoDF_TB_year<-tychoDF%>%
filter(grepl('TUBERCULOSIS',disease,ignore.case=TRUE))%>%
filter(grepl(('CA|NY|IL|PA|MA|NJ'),state,ignore.case=TRUE))%>%
aggregate(number~state+event+year,sum)


treemap(tychoDF_TB_year,
        index=c("event","state"),
        vSize="number",
        type="index")
```

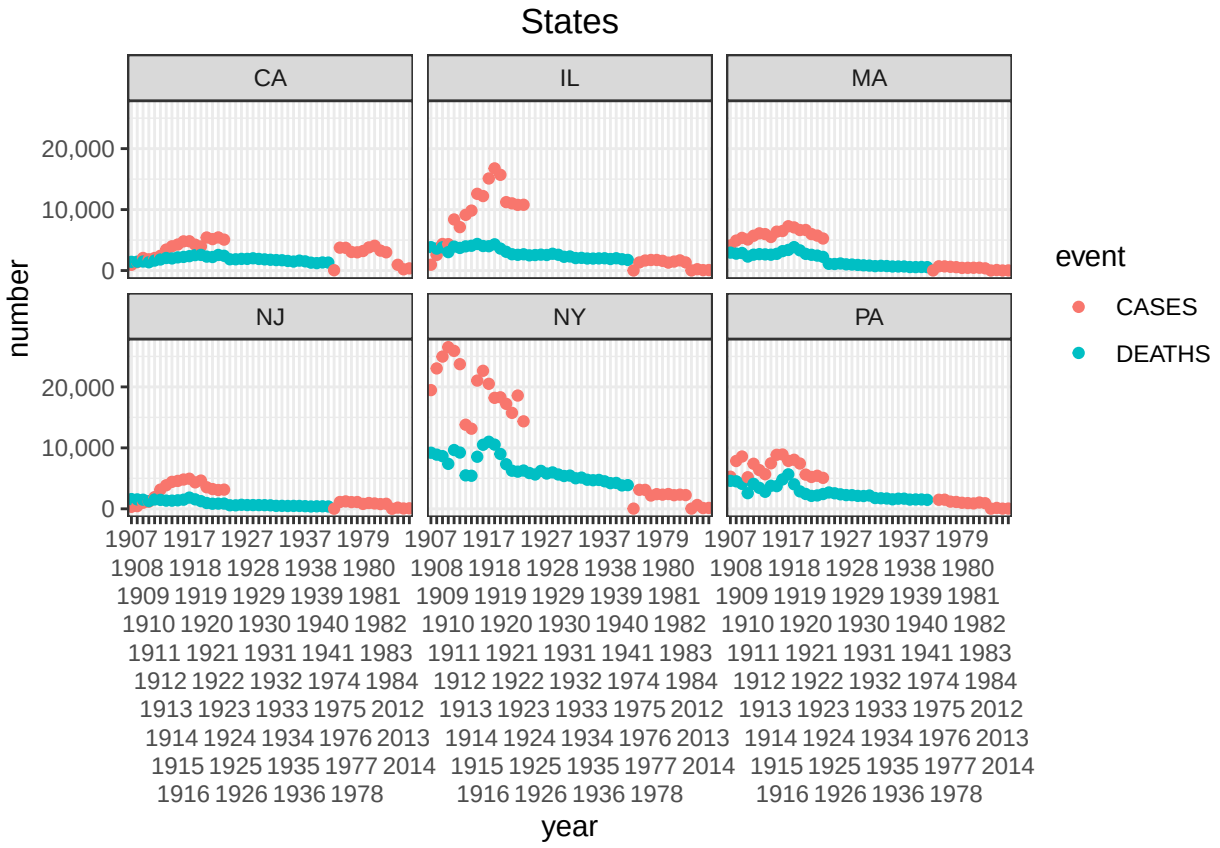



### 5.9 Comparison for Tuberculosis cases between selected states by year

Now we do the analysis year wise.Higher number of cases in followed by higher number of deaths in all states with exception in case of California, New Jersey and Illinois in the years 1907-1911. In these years the number of deaths are unusually high compared to number of cases. Number of cases and deaths tend to be higher for period 1915-1920 with New York showing very high cases and deaths in the years 1907-1911.Post 1920 we see decline in cases. This can be attributed to public clinics and better prevention education campaigns taken during those times. As we can see for each state, data for both cases and deaths is only available till 1923. So for further analysis, we will only consider years 1907-1923

```
tychoDF_TB_year%>%
ggplot(aes(year,number))+
geom_point(aes(colour=event))+
```

```
scale_y_continuous(labels = scales::comma)+
scale_x_discrete(guide = guide_axis(n.dodge=10))+
  facet_wrap(~state)+
  theme_bw()+
  ggtitle("States")+
  theme(plot.title = element_text(hjust = 0.5))
```

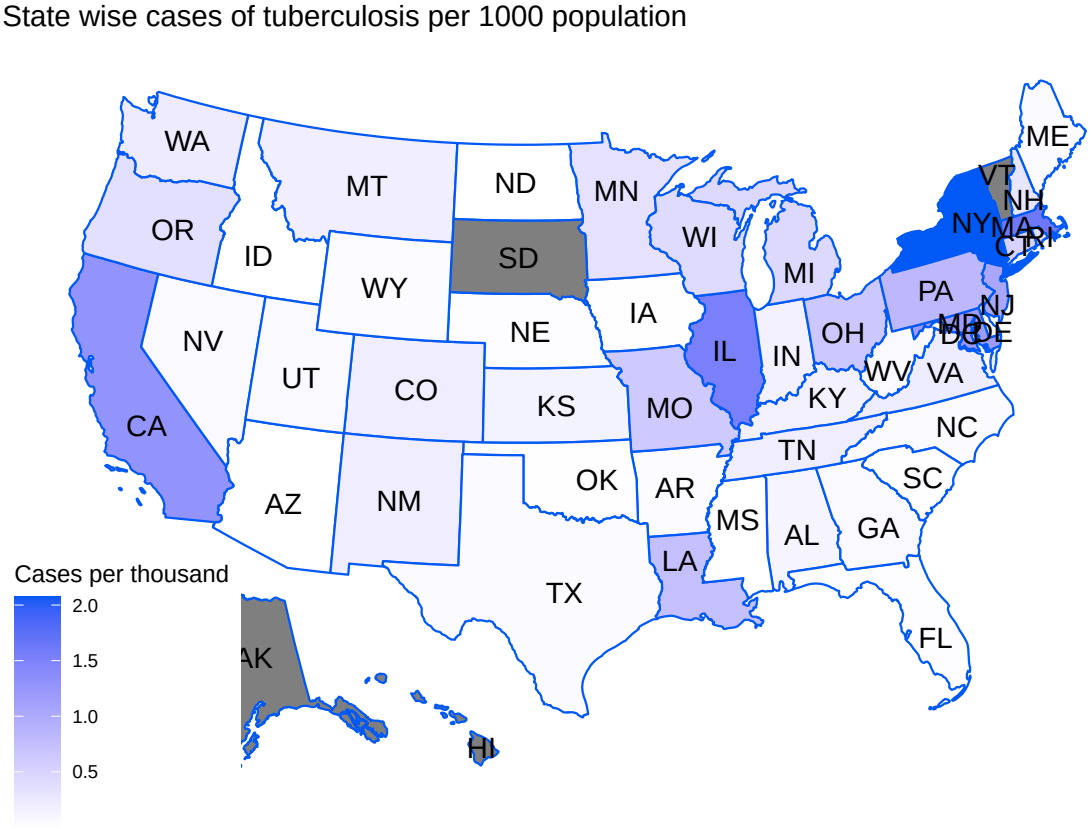



## 5.10 Comparison of states on number of Tuberculosis cases per 1000 people

We also need to take into account the population to properly understand the impact of Tuberculosis. This way we can analyze how Tuberculosis is impacting on per capita basis. Here we plot the graphs on the basis of impact per 1000 people to understand people of which states in United States were impacted more from tuberculosis. We create the required dataset. For this we use a geographic map plot of US.As we can see, in per 1000 cases, New York had highest cases of tuberculosis, followed by Massachusetts, Illinois, California and Maryland.

```
tychoDF_groupby_state_year_event<-tychoDF%>%
filter(grepl('TUBERCULOSIS',disease,ignore.case=TRUE))%>%
aggregate(number~state+year+event,sum)

tychoDF_grouped_population<-merge(x = tychoDF_groupby_state_year_event, y = populationDF, by = c("state

tychoDF_grouped_population$number<-(tychoDF_grouped_population$number/tychoDF_grouped_population$popula
```

```
tychoDF_grouped_population<-setNames(tychoDF_grouped_population,c("state","year","event","instancesPer1

tychoDF_pop_cases<-tychoDF_grouped_population%>%filter(grepl(('CASES'),event,ignore.case=TRUE))

tychoDF_pop_cases<-tychoDF_pop_cases%>%aggregate(instancesPer1000~state,mean)

  plot_usmap(data = tychoDF_pop_cases, values = "instancesPer1000", color = "#0058F5",labels = TRUE) +
  scale_fill_continuous(low = "white", high = "#0058F5", name = "Cases per thousand", label = scales::c
  labs(title = "State wise cases of tuberculosis per 1000 population") +
  theme(panel.background=element_blank())
```

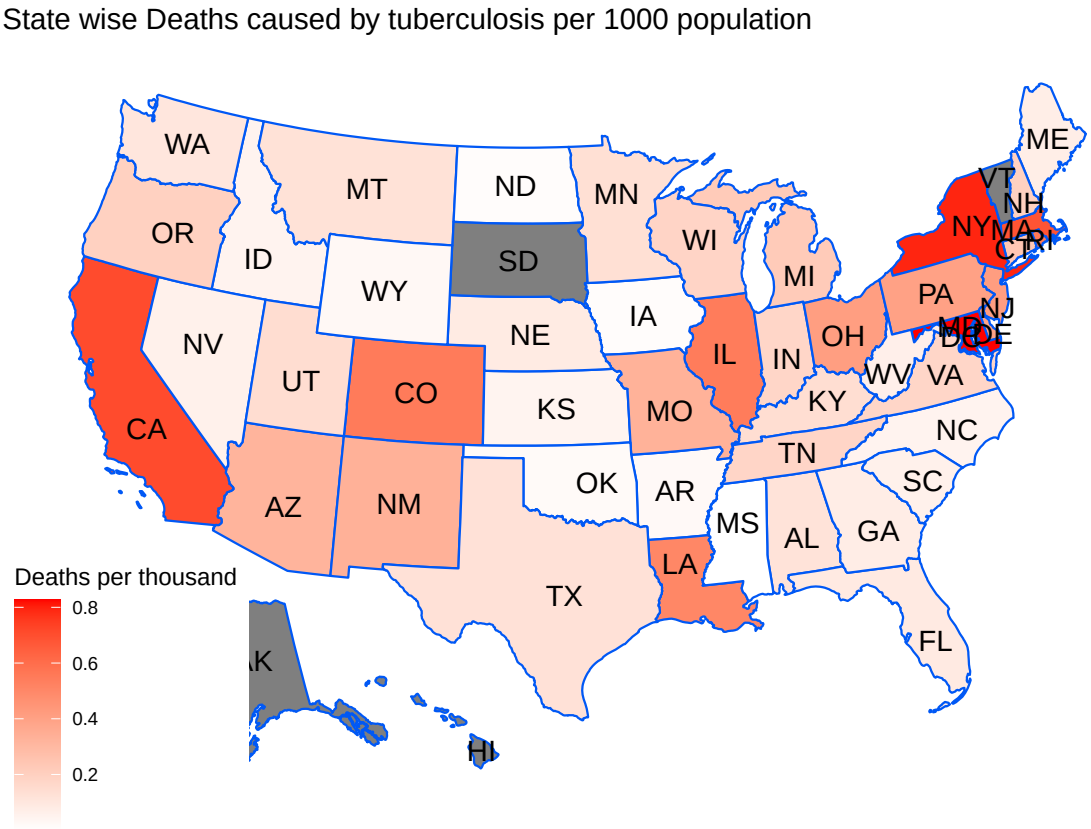



## 5.11 Comparison of states on number of Tuberculosis deaths per 1000 people

Now we draw graph for number of deaths. Here, as we can see Maryland has highest number of deaths per 1000, followed by New York, California, Massachusetts and Rhode Island. Deaths per thousand in Rhode Island is considerably high given that it does not have a lot of cases per 1000

```
tychoDF_pop_deaths<-tychoDF_grouped_population%>%filter(grepl(('DEATHS'),event,ignore.case=TRUE))

tychoDF_pop_deaths<-tychoDF_pop_deaths%>%aggregate(instancesPer1000~state,mean)


  plot_usmap(data = tychoDF_pop_deaths, values = "instancesPer1000", color = "#0058F5",labels = TRUE) +
  scale_fill_continuous(low = "white", high = "#ff0000", name = "Deaths per thousand", label = scales::
```

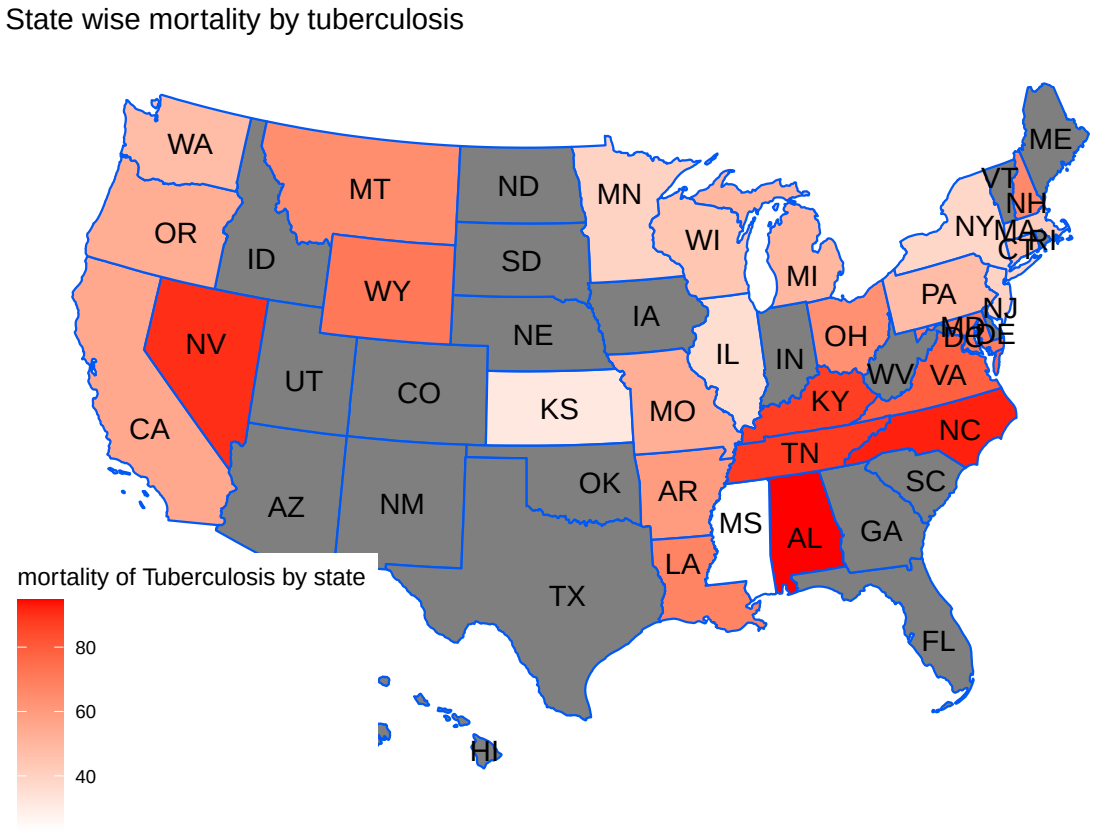
```
  labs(title = "State wise Deaths caused by tuberculosis per 1000 population") +
  theme(panel.background=element_blank())
```




## 5.12 Comparison of states by mortality from Tuberculosis

As we can see a lot of states had mortality above 100 which is not possible. One of the states with mortality higher than 100 is Rhode Island. This combined with higher number of deaths per 1000 population could mean higher reported deaths in Rhode Island or then actual deaths or lower number of total reported deaths.

```
tychoDF_pop_merge<-merge(x = tychoDF_pop_deaths, y = tychoDF_pop_cases, by = c("state"))

tychoDF_pop_mortality<-tychoDF_pop_merge[c('state','instancesPer1000.x')]

tychoDF_pop_mortality<-setNames(tychoDF_pop_mortality,c("state","mortality"))


tychoDF_pop_mortality$mortality<-100*tychoDF_pop_merge$instancesPer1000.x/tychoDF_pop_merge$instancesPe

tychoDF_pop_mortality<-tychoDF_pop_mortality%>%filter(mortality<100)

    plot_usmap(data = tychoDF_pop_mortality, values = "mortality", color = "#0058F5",labels = TRUE) +
  scale_fill_continuous(low = "white", high = "#ff0000", name = "mortality of Tuberculosis by state", la
  labs(title = "State wise mortality by tuberculosis") +
  theme(panel.background=element_blank())
```




## 6. Results

During the analysis the data was processed and studied from the following perspectives:

- Overall cases and deaths per disease in each state
- Year and month wise analysis of events
- Diseases with the highest mortality
- Impact of Tuberculosis in each state and by year
- Impact of population on events of Tuberculosis
- Comparison of states by mortality from Tuberculosis

## 7. Discussion

**Contra** As per the analysis it is observed that from the year 1949-1964 there were only cases and no deaths reported, and before 1906 no cases were reported but only deaths. For some states and diseases there were more death cases which is not possible. The GDP per capita income could be considered for better analysis of why some states have more deaths of a particular disease than others.

**Pro** This data set helps us in analyzing patterns of diseases in different states of US, it helps in identifying the most contagious diseases and which disease has the highest mortality. It also helped us in studying the seasonal trends of diseases, and understand the importance of educating people in preventing diseases as was noticed post 1920. This research can be utilized in better preparations and preventive actions against contagious diseases.

# 8. Conclusion

The overall analysis of the data shows that NY, IL and PA are amongst the states with highest events for cases and deaths. The seasonal trend suggest that the diseases tend to spread more in winter season and may also be impacted by festivals. We observed that Measles, Influenza, and Scarlet Fever are amongst the most contagious diseases. And Yellow Fever, Tuberculosis and Meningitis are amongst the diseases with highest mortality. In state wise analysis of Tuberculosis, NY, CA, IL, MA are among the states with highest number of cases and deaths. We also see increase in cases and deaths of Tuberculosis till 1920, post which we see a decrease in the cases which can be attributed to increase number of clinics and better awareness amongst the general populace. In population wise analysis we saw NY, CA, MA, MD amongst the states with highest number of cases and deaths with MD registering comparatively higher deaths per 1000 population.

# Literature

[1] National Notifiable Diseases Surveillance System (NNDSS) https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/national-notifiable-diseases-surveillance-system-nndss#:~:text=The%20National%20Notifiable%20Disease%20Surveillance,state%2Dreportable%20and%20nationally%20not

[2] Tuberculosis General Information Fact Sheet https://www.cdc.gov/tb/publications/factsheets/general/tb.htm#:~:text=What%20is%20TB%3F,they%20do%20not%20get%20treatment.

[3] Population 1900-2022 https://www.macrotrends.net/states/california/population

[4] THE FORGOTTEN PLAGUE | TB in America: 1895-1954 https://www.pbs.org/wgbh/americanexperience/features/plague-gallery/