Data Visualization Project

THD
TECHNISCHE
HOCHSCHULE
DEGGENDORF

University Logo

TOPIC: POLIOMYTHESIS

SUBJECT : Data Visualization

Professor: Professsor Javier Valdes

Submitted by Team of

Hassan Rafique

Overview

Short Summary

The following study is done to visualize “Poliomythesis” and its impact over the states of US over a certain period of time.

Keywords

Following are some of the keywords:

- Poliomythesis
- epidemic
- vaccine

Figures

1. Introduction

US citizens are affected by a variety of diseases like COVID, Mumps, Influneza, Hepatitis, Dengue,etc. Among st which there is another common disease called Polio or Poliomyelitis. Polio is a life threatening disease that spreads from human to human targeting the spinal cord and also causing paralysis. Other few symptoms of polio include: sore throat, nausea, fever, tiredness, headache and stomach pain.

The first epidemic outbreak occurred in 1950 in the US, and it affected people ranging from infants to the age of 9 and also some individuals at the age of 15. According to a recent study, Polio cases have decreases drastically but still some people remain affected by it. This analysis is performed to visualize the overall ratio of affected people by Polio in the states of US and to see how many cases occur in each state. Although the epidemic seems to be stopped as many people get vaccinated at early stages, but the threat still remains as polio is incurable, though it is preventable.

2. Problem definition

The following analysis aims to study the impact of Polio on the states of US. Which state was most effected by this virus and which state had the lowest cases, whether majority of cases fall into paralysis or meningitis, all these queries will be visualized in this analysis.

3. Objectives

Following are the objectives behind our study:
- to visualize a disease amongst the 50 diseases of the US

- display the impact of the disease on each state
- Over given period of time how many cases were discovered of that specific disease
- compare that specific disease to other common diseases of the dataset
- display the impact of polio vaccine on total number of cases of each state

4. Methods

During our study, we used the following methodology to derive and display our analysis:
- Methods from descriptive statistics: Descriptive Statistics
- Statistics for scientific work: Literate Statistical Analysis

- Advanced-Data visualization techniques: ggplot2

5. Analysis Protocol

The analysis is divided into sub-parts. The working steps are introduced below. Before to start the analysis the required libraries were loaded

Loading required libraries

The following libraries were loaded for the visualization of the study:

```
library(tidyverse) #a collection of open source packages for the R language
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1     
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
library(tinytex) # Helper functions to install and maintain the 'LaTeX' distribution named 'TinyTe
X'
```

```
## Warning: package 'tinytex' was built under R version 4.3.1
```

```
library(dplyr) #  a data manipulation package, used for making tabular data manipulation faster, s
impler and easier
library(ggplot2) # plotting package that provides helpful commands to create complex plots from da
ta
library(hrbrthemes) # provides typography-centric themes and theme components for ggplot2.
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
```

```
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
library(ggrepel) # provides geoms for ggplot2 to repel overlapping text labels
library(usmap) # Obtain United States map data frames of varying region types
```

```
## Warning: package 'usmap' was built under R version 4.3.1
```

```r
#library(sf)
library(plotly) # graphing library makes interactive, publication-quality graphs
```

```
## Warning: package 'plotly' was built under R version 4.3.1
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

5.1 Data loading and cleanup

Data loading and Cleanup is the first and one of the most important steps of any data analysis. Data cleaning is considered a foundation element of basic data science. During this step, data was loaded into R and the data was processed in order to identify the incomplete, irrelevant, or missing parts of the data and then modifying or deleting them according to the necessity.

Loading Data

The following data consists of 50 diseases, 50 US states and 1284 US cities. We import the dataset into R studio:

```r
library(readr) #importing dataset
ProjectTycho_Level2_v1_1_0 <- read_csv("ProjectTycho_Level2_v1.1.0.csv")
```

```
## Rows: 3659360 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): country, state, loc, loc_type, disease, event, url
## dbl  (2): epi_week, number
## date (2): from_date, to_date
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
View(ProjectTycho_Level2_v1_1_0)

mydata <- ProjectTycho_Level2_v1_1_0
str(mydata)
```

```
## spc_tbl_ [3,659,360 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ epi_week : num [1:3659360] 188824 188824 188824 188826 188826 ...
##  $ country  : chr [1:3659360] "US" "US" "US" "US" ...
##  $ state    : chr [1:3659360] "PA" "PA" "PA" "PA" ...
##  $ loc      : chr [1:3659360] "PHILADELPHIA" "PHILADELPHIA" "PHILADELPHIA" "PHILADELPHIA" ...
##  $ loc_type : chr [1:3659360] "CITY" "CITY" "CITY" "CITY" ...
##  $ disease  : chr [1:3659360] "TYPHOID FEVER [ENTERIC FEVER]" "SCARLET FEVER" "DIPHTHERIA" "TYP
HOID FEVER [ENTERIC FEVER]" ...
##  $ event    : chr [1:3659360] "DEATHS" "DEATHS" "DEATHS" "DEATHS" ...
##  $ number   : num [1:3659360] 14 4 4 12 5 7 4 1 1 4 ...
##  $ from_date: Date[1:3659360], format: "1888-06-10" "1888-06-10" ...
##  $ to_date  : Date[1:3659360], format: "1888-06-16" "1888-06-16" ...
##  $ url      : chr [1:3659360] "http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2084171" "http://ww
w.ncbi.nlm.nih.gov/pmc/articles/PMC2084171" "http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2084171"
"http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2084171" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   epi_week = col_double(),
##   ..   country = col_character(),
##   ..   state = col_character(),
##   ..   loc = col_character(),
##   ..   loc_type = col_character(),
##   ..   disease = col_character(),
##   ..   event = col_character(),
##   ..   number = col_double(),
##   ..   from_date = col_date(format = ""),
##   ..   to_date = col_date(format = ""),
##   ..   url = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data cleanup

The dataset contains 50 diseases in total. But for this analysis we will be working only on one disease, so we clean the dataset. 'Polio' is our target disease in this analysis so we only keep columns that contain relevant data for further analysis.

```
Polio_data = mydata %>% filter (disease == "POLIOMYELITIS") #here we are filtering the dataset and
keeping only that part of the dataset that we require for our study

head(Polio_data) #generic function, returns first few rows of the dataset
```

epi_week <dbl>	country <chr>	state <chr>	loc <chr>	loc_type <chr>	disease <chr>	ev… <chr>	num… <dbl>	from_date <date>	to_date <date>
191212	US	MA	BOSTON	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	OH	CINCINNATI	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	MN	DULUTH	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	MI	KALAMAZOO	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	NY	NEW YORK	CITY	POLIOMYELITIS	CASES	12	1912-03-17	1912-03-23
191212	US	NE	OMAHA	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23

6 rows | 1-10 of 11 columns

Irrelevant to the analysis columns were also removed.

```
Polio_data<-Polio_data[,-c(11)] # here, we are removing irrelevant columns so that we can do a better and accurate analysis
head(Polio_data)
```

epi_week	country	state	loc	loc_type	disease	event	num…	from_date	to_date
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
191212	US	MA	BOSTON	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	OH	CINCINNATI	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	MN	DULUTH	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	MI	KALAMAZOO	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	NY	NEW YORK	CITY	POLIOMYELITIS	CASES	12	1912-03-17	1912-03-23
191212	US	NE	OMAHA	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23

6 rows

‘NA’ values are missing values in the dataset and they should be removed for better results. So, NA values were removed

```
Polio_data <- na.omit(Polio_data) # removing NA values so that we can do a better and accurate analysis
head(Polio_data)
```

epi_week	country	state	loc	loc_type	disease	event	num…	from_date	to_date
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
191212	US	MA	BOSTON	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	OH	CINCINNATI	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	MN	DULUTH	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	MI	KALAMAZOO	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23
191212	US	NY	NEW YORK	CITY	POLIOMYELITIS	CASES	12	1912-03-17	1912-03-23
191212	US	NE	OMAHA	CITY	POLIOMYELITIS	CASES	1	1912-03-17	1912-03-23

6 rows

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

```
str(Polio_data) #str function shows structure of dataset
```

```
## tibble [140,774 × 10] (S3: tbl_df/tbl/data.frame)
##  $ epi_week : num [1:140774] 191212 191212 191212 191212 191212 ...
##  $ country  : chr [1:140774] "US" "US" "US" "US" ...
##  $ state    : chr [1:140774] "MA" "OH" "MN" "MI" ...
##  $ loc      : chr [1:140774] "BOSTON" "CINCINNATI" "DULUTH" "KALAMAZOO" ...
##  $ loc_type : chr [1:140774] "CITY" "CITY" "CITY" "CITY" ...
##  $ disease  : chr [1:140774] "POLIOMYELITIS" "POLIOMYELITIS" "POLIOMYELITIS" "POLIOMYELITIS"
...
##  $ event    : chr [1:140774] "CASES" "CASES" "CASES" "CASES" ...
##  $ number   : num [1:140774] 1 1 1 1 12 1 1 1 1 2 ...
##  $ from_date: Date[1:140774], format: "1912-03-17" "1912-03-17" ...
##  $ to_date  : Date[1:140774], format: "1912-03-23" "1912-03-23" ...
```

For Comparing diseases

As done above, we will now filter two more diseases out of the 50 diseases.

```
#filter mumps#
Mumps_data = mydata %>% filter (disease == "MUMPS")
Mumps_data<-Mumps_data[,-c(11)]
Mumps_data <- na.omit(Mumps_data)
head(Mumps_data)
```

epi_week <dbl>	country <chr>	state <chr>	loc <chr>	loc_type <chr>	disease <chr>	event <chr>	num... <dbl>	from_date <date>	to_date <date>
192401	US	ME	PORTLAND	CITY	MUMPS	CASES	4	1923-12-30	1924-01-05
192401	US	MA	BOSTON	CITY	MUMPS	CASES	12	1923-12-30	1924-01-05
192401	US	MA	FALL RIVER	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05
192401	US	MA	SPRINGFIELD	CITY	MUMPS	CASES	6	1923-12-30	1924-01-05
192401	US	MA	WORCESTER	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05
192401	US	RI	PAWTUCKET	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05

6 rows

```
#filter influenza
Influenza_data = mydata %>% filter (disease == "INFLUENZA")
Influenza_data<-Influenza_data[,-c(11)]
Influenza_data <- na.omit(Influenza_data)
head(Influenza_data)
```

epi_week <dbl>	country <chr>	state <chr>	loc <chr>	loc_type <chr>	disease <chr>	event <chr>	num... <dbl>	from_date <date>	to_date <date>
191945	US	AL	ALABAMA	STATE	INFLUENZA	CASES	4	1919-11-02	1919-11-08
191945	US	AR	ARKANSAS	STATE	INFLUENZA	CASES	24	1919-11-02	1919-11-08
191945	US	CA	CALIFORNIA	STATE	INFLUENZA	CASES	31	1919-11-02	1919-11-08
191945	US	CT	CONNECTICUT	STATE	INFLUENZA	CASES	5	1919-11-02	1919-11-08
191945	US	FL	FLORIDA	STATE	INFLUENZA	CASES	31	1919-11-02	1919-11-08

epi_week	country	state	loc	loc_type	disease	event	num…	from_date	to_date
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
191945	US	GA	GEORGIA	STATE	INFLUENZA	CASES	37	1919-11-02	1919-11-08

6 rows

Group Data

Here we are grouping all the filtered polio cases, by state and Total_number of cases. The summarise() function in R, summarises the data frame into a value or vector. The '%>%' pipe symbol takes output of one function and passes it onto another function.

```
group_data <- Polio_data %>% group_by(state) %>% summarise(Total_number = sum(number)) #grouping P
olio Data by state and total number of cases

head(group_data)
```

state	Total_number
<chr>	<dbl>
AK	524
AL	6968
AR	5457
AZ	3097
CA	57512
CO	7182

6 rows

Bar Plot

The main purpose of a bar plot is to divide the data in separate individual bars and convey relational information in a visual track able manner.

```
# using ggplot library to plot a bar plot between Total no of cases and state of US
ggplot(Polio_data, aes(x = state)) +
  geom_bar(stat = "count",fill = "darkblue", bins = 30) +
  xlab("US STATES") + ylab("Total number of CASES") + ggtitle("Total cases of Polio in US States")
+
  theme(axis.text.x = element_text(angle = 90, vjust =0.2, hjust=1,size = 9))
```

```
## Warning in geom_bar(stat = "count", fill = "darkblue", bins = 30): Ignoring
## unknown parameters: `bins`
```

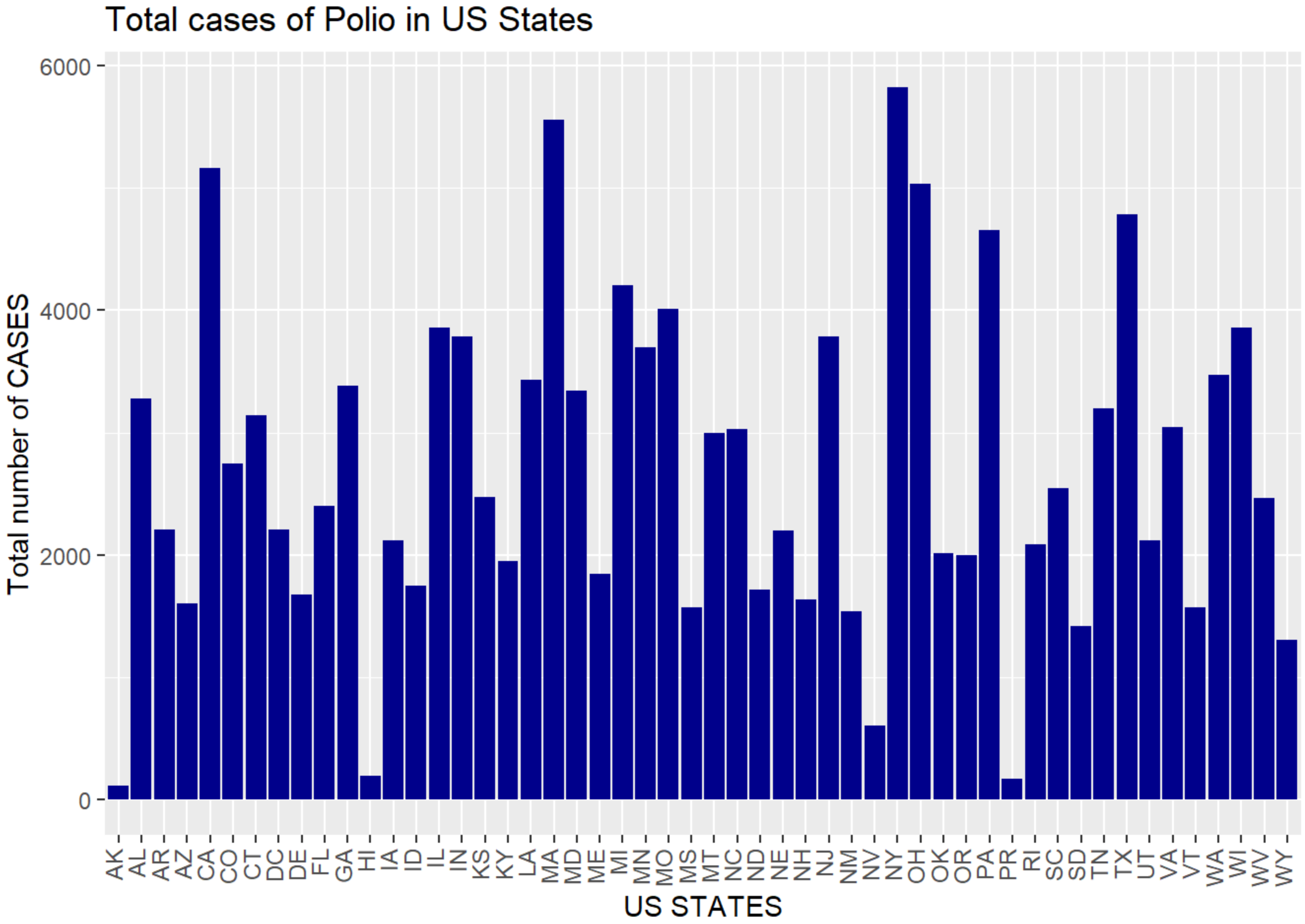


Fig.1: Bar Plot for Total Number of Polio Cases

Here, Fig.1 displays the total number of Polio cases in the states of US. The states ‘NY’, ‘MA’, ‘CA’ have the highest number of cases amongst all the US states, whereas ‘AK’ and ‘PR’ states have the lowest polio cases.

Bar Plot Descending Order

We can sort the data of the visuals as well. So , here we are creating a bar plot with data sorted in the form of descending order.

```
library(highcharter) # is a wrapper for the Highcharts library including shortcut functions to plot R objects.
```

```
## Warning: package 'highcharter' was built under R version 4.3.1
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(dplyr)

temp <- Polio_data %>%    #grouping state and total number of cases in a descending order using pi
pe operator
  group_by(state) %>%
  summarise(total_cases = sum(number)) %>%
  arrange(desc(total_cases))

bar1 <- highchart() %>%
  hc_title(text = sprintf("Polio Cases in The %s's", unique(temp$decade)), align = "center") %>%
  hc_xAxis(categories = unique(temp$state), title = list(text = "State"), labels = list(rotation =
-90, align = "right", style = list(fontSize = "10px"))) %>%
  hc_yAxis(title = list(text = "Number of Cases")) %>%
  hc_add_series(name = "Total Number of Cases", data = temp$total_cases, color = "darkblue") %>%
  hc_chart(type = "column", options3d = list(enabled = TRUE, beta = 2, alpha = 2, width = 800)) %
>%
  hc_add_theme(hc_theme_538())
```

```
## Warning: Unknown or uninitialised column: `decade`.
```

```
bar1
```

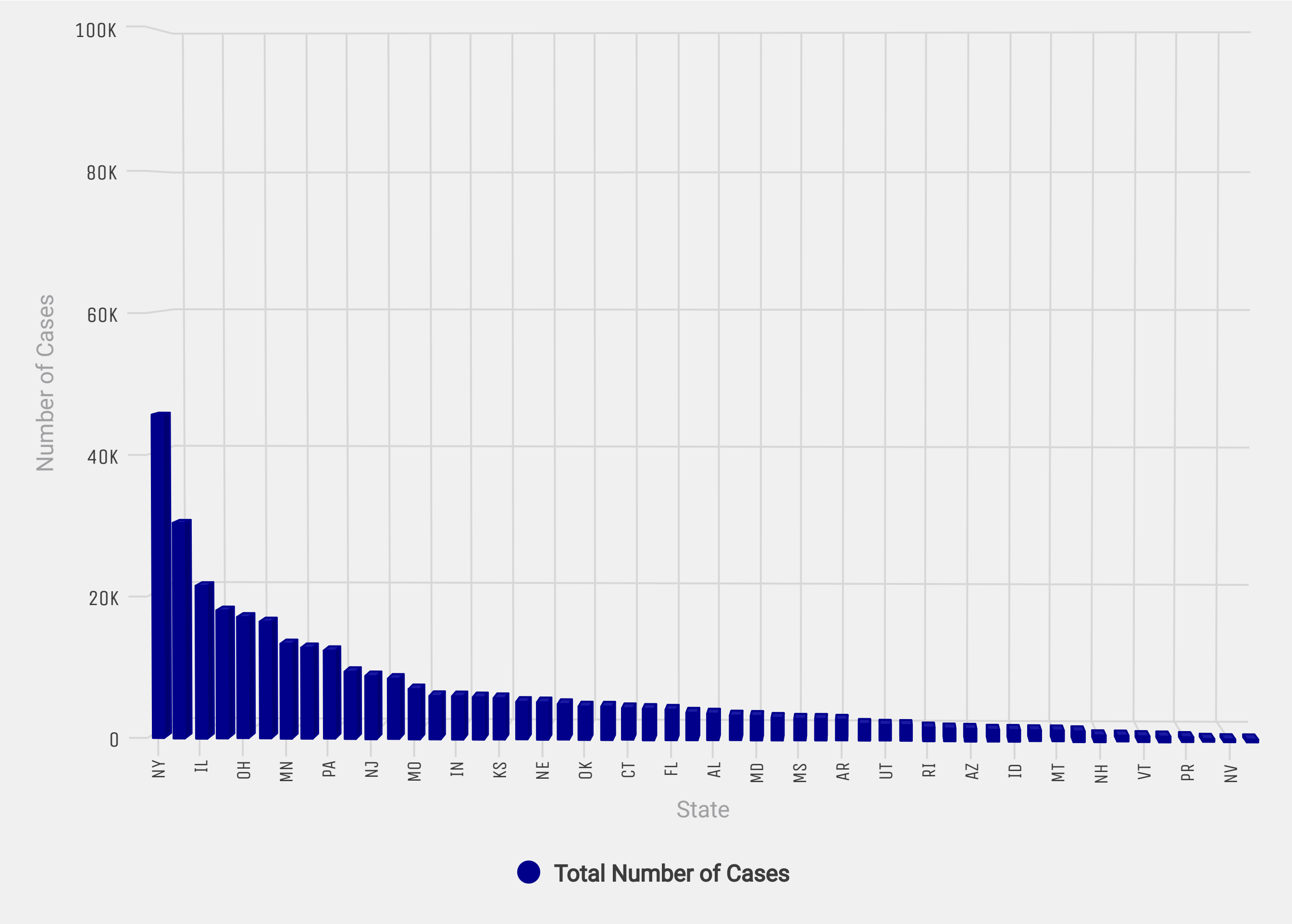


**Fig.2: ** Bar Plot for Total Number of Polio Cases in a Descending Order

Map Chart

The main purpose of a map chart is when geographic context in your data is present and you want to visualize it over a world map or map of a certain country.

```
library(ggiraph) # tool that allows you to create dynamic ggplot graphs.
```

```
## Warning: package 'ggiraph' was built under R version 4.3.1
```

```
library(maps) #  best source of geospatial data in R
```

```
## Warning: package 'maps' was built under R version 4.3.1
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(mapdata) # provide the boundaries of the most common world regions
```

```
## Warning: package 'mapdata' was built under R version 4.3.1
```

```
library(usmap)
library(RColorBrewer) # tool to manage colors with R

df_1_1 <- Polio_data %>% group_by(state) %>% summarise(cases = sum(number)) #grouping data by state and total number of cases

# Create a map plot for number of polio cases in every state

p<- plot_usmap(data = df_1_1, values = "cases", color = "white" , labels = TRUE, label_color = "black") +
  scale_fill_continuous(
    low = "lightblue", high = "darkblue", name = "No. of Polio Cases by State", label = scales::comma)
  theme(legend.position = "left")+
  labs(title= "Distribution of Polio Cases")
```

```
## List of 2
##  $ legend.position: chr "left"
##  $ title          : chr "Distribution of Polio Cases"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
  #here we are writing code that basically does that when you hover your pointer on a US state on the map, it will display number of cases of that certain state
  p$layers[[2]]$aes_params$size <- 2
  p <- p %>%
  plotly::highlight("plotly_hover") %>%
  plotly::style(hoverlabel = list(font = list(size = 13)))
  config(p, displayModeBar = FALSE)
```

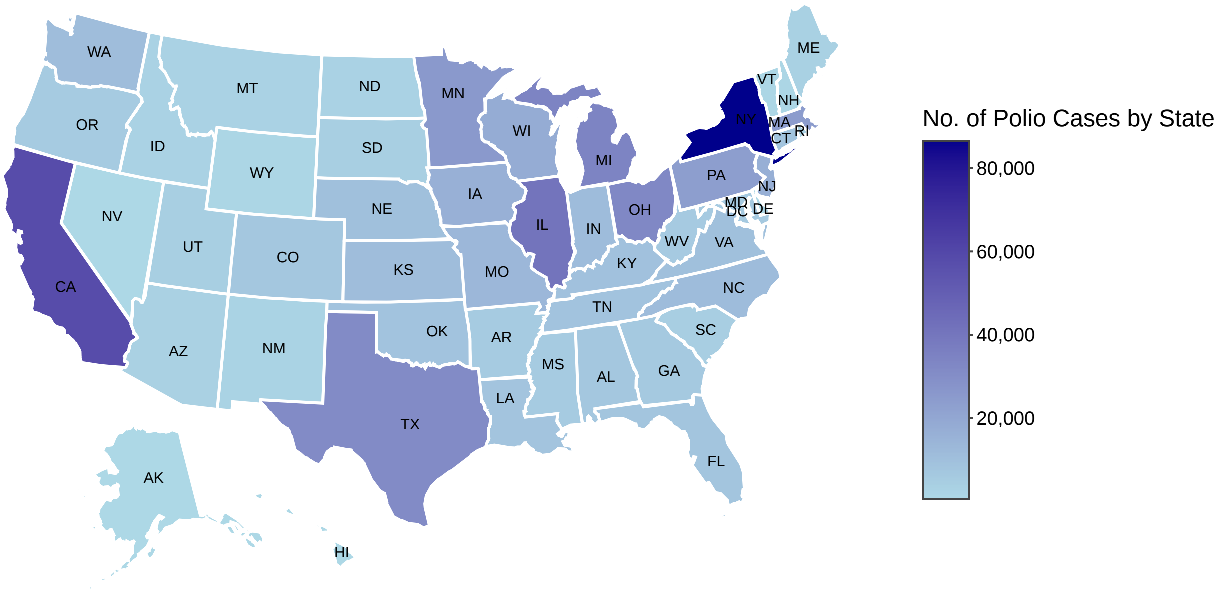


**Fig.3: ** Map Chart displaying total number of polio cases in the US states map

Fig.3 illustrates the complete map of the 50 states of the US. The scale on the right indicates the total number of cases, light blue colored area is the lowest effected area, whereas dark blue area on the map is the most effected area. The state ‘NY’ is the most effected area of Poliomythesis.

LINE CHART

A line chart is a graph that connects a series of points by drawing line segments between them.

```
library(lubridate) # package that makes it easier to work with dates and times

Polio_data_line <- Polio_data %>%
  mutate(year = lubridate:: year(from_date)) %>% #mutate funtion creates new columns that are functions of existing variables.
  group_by(year) %>%
  summarise(total_cases = sum(number))

ggplot(Polio_data_line, aes(x = year, y= total_cases)) +
  geom_line(color = "blue") +
  labs(x = "Year", y= "Total Cases", title = "Total Polio Cases per Year") +
  theme_minimal() +
  geom_point(color = "red") +
  geom_smooth(se = FALSE, color = "green")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

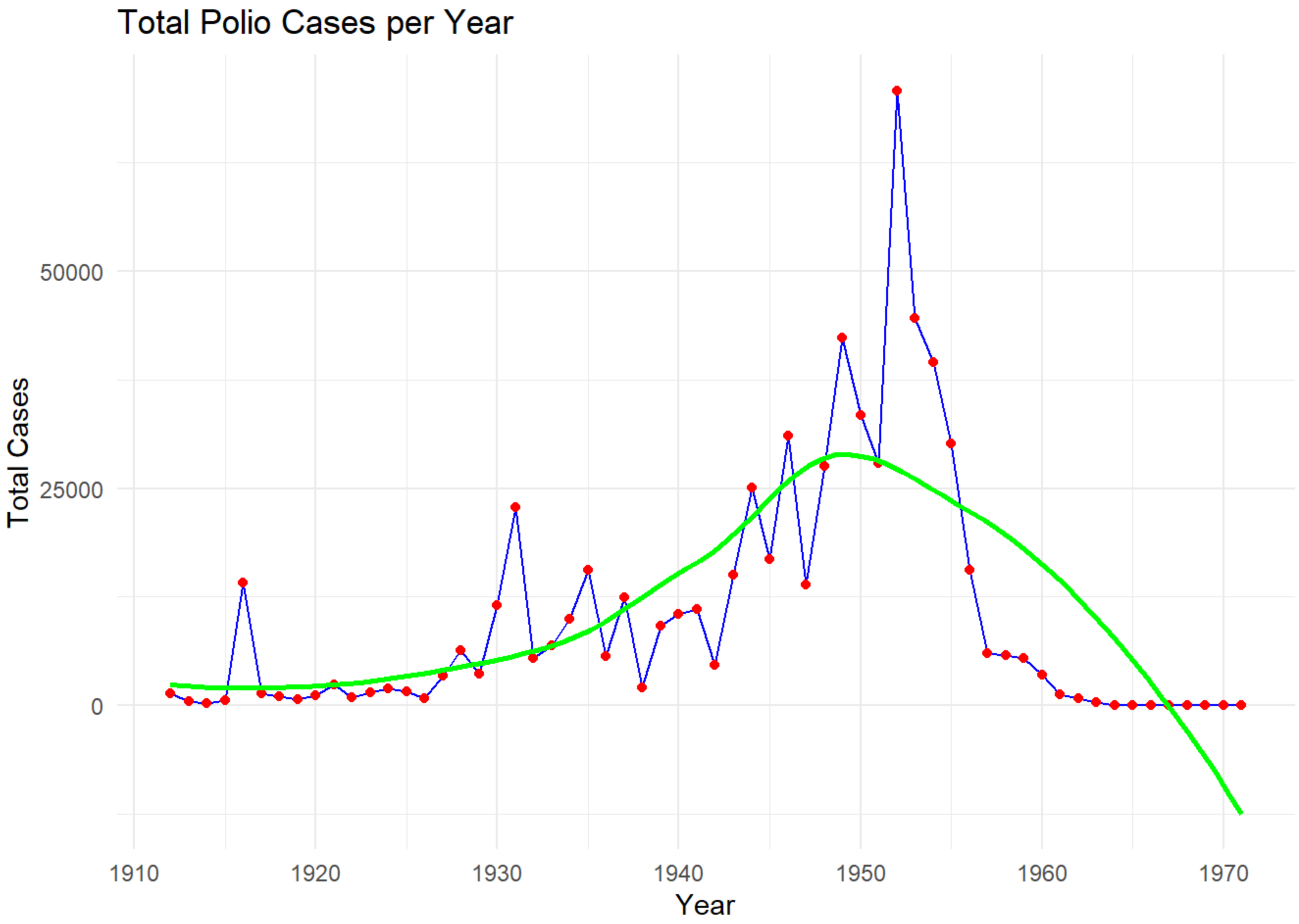


Fig. 4: Line

Chart of Total cases of Polio over the Passing Years

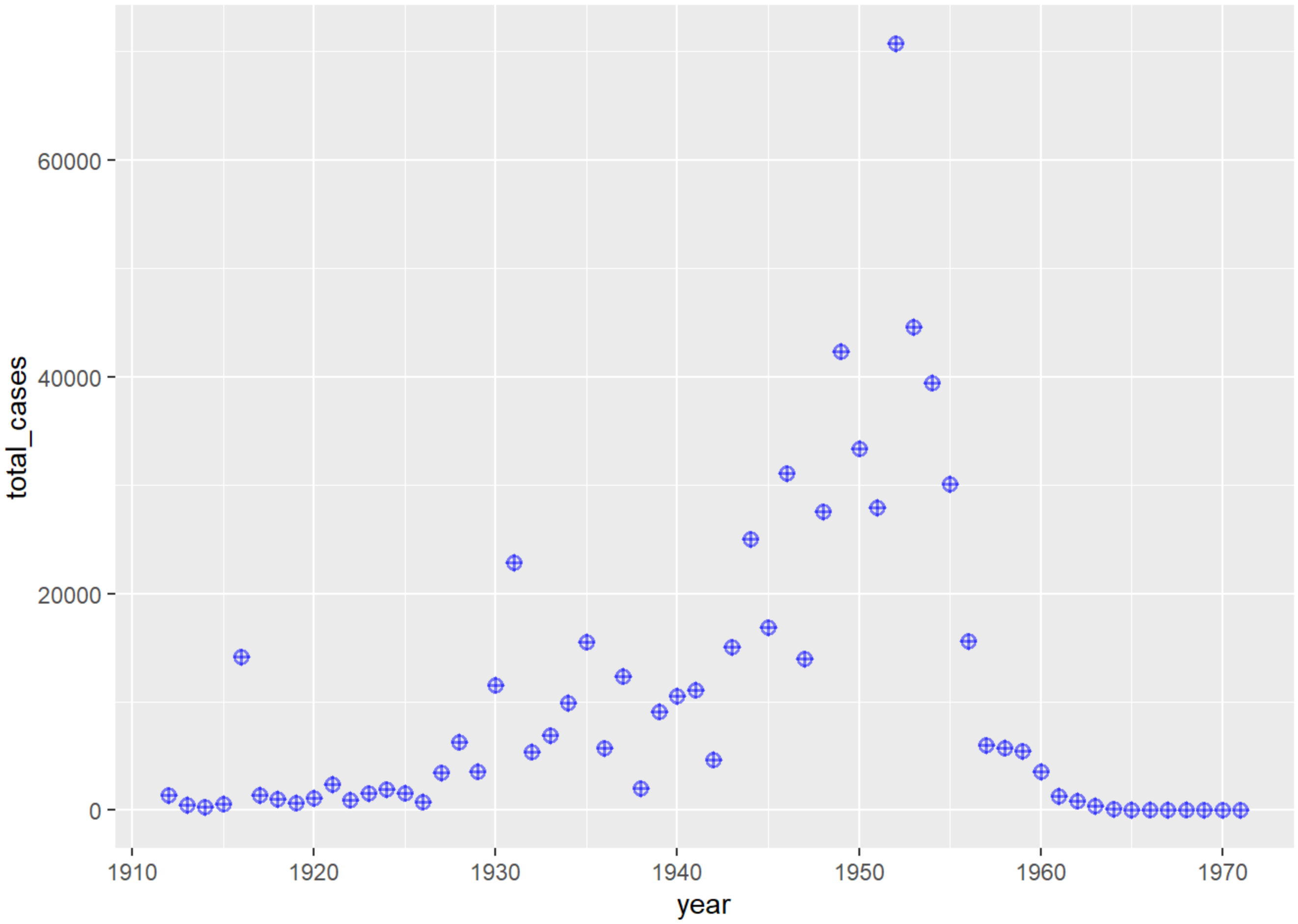
Fig.4 illustrates the total number of polio cases every year. The figure shows that around 1950’s there was a spike in the number of cases but soon after that time, cases started to decline rapidly.

###** SCATTER PLOT **

A Scatterplot displays the relationship between 2 numeric variables. Each dot represents an observation.

```
Polio_data_sc <- Polio_data %>%
  mutate(year = lubridate::year(from_date)) %>%
  group_by(year) %>%
  summarise(total_cases = sum(number))

#creating plot for total number of cases over the years in the US states
ggplot(Polio_data_sc, aes(x = year, y = total_cases)) +
  geom_point(
    color = "blue",
    shape = 10,
    alpha = 0.5,
    size = 2,
    stroke = 1)
```



```
  #theme_ipsum()
```

Fig. 5: Scatter Plot of Total cases of Polio over the Passing Years

Fig. 5 displays the total number of polio cases every year. Its data is similar to Fig. 4 so the graphs are also quite similar

COMPARING GRAPHS

Following graphs below are displaying the compared outputs of the three diseases Polio, mumps and influenza.

** MULTI BAR GRAPH**

```
# Assuming you have separate data frames for polio and mumps data named 'polio_data' and 'mumps_data' respectively
# Extract year and total cases for polio data
library(highcharter)
library(dplyr)

# Assuming you have loaded the data into a data frame named 'fltrdData' with columns: state, disease, number

# Filter the data for measles and mumps diseases
# filtered_data <- ProjectTycho %>%
#   filter(disease %in% c("POLIOMYELITIS", "MUMPS","INFLUENZA"))
filtered_data <- ProjectTycho_Level2_v1_1_0 %>%
  filter(disease %in% c("POLIOMYELITIS", "MUMPS", "INFLUENZA") & event == "CASES")

# Group the data by state and disease, and calculate the total cases
total_cases <- filtered_data %>%
  group_by(state, disease) %>%
  summarise(total_cases = sum(number))
```

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

```
# Pivot the data to wide format for plotting
wide_data <- total_cases %>%
  pivot_wider(names_from = disease, values_from = total_cases)

# Create the bar plot
bar_plot <- highchart() %>%
  hc_title(text = "Total Measles,INFLUENZA and Mumps Cases by State") %>%
  hc_xAxis(categories = wide_data$state, title = list(text = "State")) %>%
  hc_yAxis(title = list(text = "Total Cases"), labels = list(format = "{value}")) %>%
  hc_add_series(name = "POLIOMYELITIS", data = wide_data$POLIOMYELITIS, color = "blue") %>%
  hc_add_series(name = "MUMPS", data = wide_data$MUMPS, color = "#FF0000") %>%
  hc_add_series(name = "INFLUENZA", data = wide_data$MUMPS, color = "green") %>%
  hc_chart(type = "column") %>%
 # hc_add_theme(hc_theme_538())%>%
    hc_plotOptions(column = list(pointPadding = 0.2, borderWidth = 0, pointWidth = 7))  # Adjust the pointWidth value as desired

# Print the bar plot
print(bar_plot)
```

Fig. 6: Bar Plot of Compared Diseases

Fig. 6 displays that in all the 50 states, ‘AZ’, ‘MI’ and ‘WI’ had almost equal number of mumps and influenza cases, whereas the highest number of Polio cases has been spotted in the ‘NY’ state of US

** MULTI LINE GRAPH**

```
# Assuming you have separate data frames for polio and mumps data named 'polio_data' and 'mumps_da
ta' respectively

#mean of number:
#filter mumps
Mumps_data = mydata %>% filter (disease == "MUMPS")
str(Mumps_data)
```

```
## spc_tbl_ [88,897 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ epi_week : num [1:88897] 192401 192401 192401 192401 192401 ...
##  $ country  : chr [1:88897] "US" "US" "US" "US" ...
##  $ state    : chr [1:88897] "ME" "MA" "MA" "MA" ...
##  $ loc      : chr [1:88897] "PORTLAND" "BOSTON" "FALL RIVER" "SPRINGFIELD" ...
##  $ loc_type : chr [1:88897] "CITY" "CITY" "CITY" "CITY" ...
##  $ disease  : chr [1:88897] "MUMPS" "MUMPS" "MUMPS" "MUMPS" ...
##  $ event    : chr [1:88897] "CASES" "CASES" "CASES" "CASES" ...
##  $ number   : num [1:88897] 4 12 0 6 0 0 0 15 0 129 ...
##  $ from_date: Date[1:88897], format: "1923-12-30" "1923-12-30" ...
##  $ to_date  : Date[1:88897], format: "1924-01-05" "1924-01-05" ...
##  $ url      : chr [1:88897] "https://www.tycho.pitt.edu/raw/PDF/1924/04.pdf" "https://www.tych
o.pitt.edu/raw/PDF/1924/04.pdf" "https://www.tycho.pitt.edu/raw/PDF/1924/04.pdf" "https://www.tych
o.pitt.edu/raw/PDF/1924/04.pdf" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   epi_week = col_double(),
##   ..   country = col_character(),
##   ..   state = col_character(),
##   ..   loc = col_character(),
##   ..   loc_type = col_character(),
##   ..   disease = col_character(),
##   ..   event = col_character(),
##   ..   number = col_double(),
##   ..   from_date = col_date(format = ""),
##   ..   to_date = col_date(format = ""),
##   ..   url = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
Mumps_data
```

epi_week	country	state	loc	loc_type	disease	event	num...	from_date	to_date
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
192401	US	ME	PORTLAND	CITY	MUMPS	CASES	4	1923-12-30	1924-01-05
192401	US	MA	BOSTON	CITY	MUMPS	CASES	12	1923-12-30	1924-01-05
192401	US	MA	FALL RIVER	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05
192401	US	MA	SPRINGFIELD	CITY	MUMPS	CASES	6	1923-12-30	1924-01-05
192401	US	MA	WORCESTER	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05
192401	US	RI	PAWTUCKET	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05

epi_week	country	state	loc	loc_type	disease	event	num…	from_date	to_date
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
192401	US	RI	PROVIDENCE	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05
192401	US	CT	NEW HAVEN	CITY	MUMPS	CASES	15	1923-12-30	1924-01-05
192401	US	NY	BUFFALO	CITY	MUMPS	CASES	0	1923-12-30	1924-01-05
192401	US	NY	NEW YORK	CITY	MUMPS	CASES	129	1923-12-30	1924-01-05

1-10 of 10,000 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 ... 1000 Next

```
#filter influenza
Influenza_data = mydata %>% filter (disease == "INFLUENZA")
str(Influenza_data)
```

```
## spc_tbl_ [236,673 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ epi_week : num [1:236673] 191945 191945 191945 191945 191945 ...
##  $ country  : chr [1:236673] "US" "US" "US" "US" ...
##  $ state    : chr [1:236673] "AL" "AR" "CA" "CT" ...
##  $ loc      : chr [1:236673] "ALABAMA" "ARKANSAS" "CALIFORNIA" "CONNECTICUT" ...
##  $ loc_type : chr [1:236673] "STATE" "STATE" "STATE" "STATE" ...
##  $ disease  : chr [1:236673] "INFLUENZA" "INFLUENZA" "INFLUENZA" "INFLUENZA" ...
##  $ event    : chr [1:236673] "CASES" "CASES" "CASES" "CASES" ...
##  $ number   : num [1:236673] 4 24 31 5 31 37 56 29 5 12 ...
##  $ from_date: Date[1:236673], format: "1919-11-02" "1919-11-02" ...
##  $ to_date  : Date[1:236673], format: "1919-11-08" "1919-11-08" ...
##  $ url      : chr [1:236673] "https://www.tycho.pitt.edu/raw/PDF/1919/46.pdf" "https://www.tych
o.pitt.edu/raw/PDF/1919/46.pdf" "https://www.tycho.pitt.edu/raw/PDF/1919/46.pdf" "https://www.tych
o.pitt.edu/raw/PDF/1919/46.pdf" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   epi_week = col_double(),
##   ..   country = col_character(),
##   ..   state = col_character(),
##   ..   loc = col_character(),
##   ..   loc_type = col_character(),
##   ..   disease = col_character(),
##   ..   event = col_character(),
##   ..   number = col_double(),
##   ..   from_date = col_date(format = ""),
##   ..   to_date = col_date(format = ""),
##   ..   url = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
Influenza_data
```

epi_week	country	state	loc	loc_type	disease	ev…	num…	from_date	to_date
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
191945	US	AL	ALABAMA	STATE	INFLUENZA	CASES	4	1919-11-02	1919-11-08
191945	US	AR	ARKANSAS	STATE	INFLUENZA	CASES	24	1919-11-02	1919-11-08

epi_week	country	state	loc	loc_type	disease	ev…	num…	from_date	to_date ▸
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<date>	<date>
191945	US	CA	CALIFORNIA	STATE	INFLUENZA	CASES	31	1919-11-02	1919-11-08
191945	US	CT	CONNECTICUT	STATE	INFLUENZA	CASES	5	1919-11-02	1919-11-08
191945	US	FL	FLORIDA	STATE	INFLUENZA	CASES	31	1919-11-02	1919-11-08
191945	US	GA	GEORGIA	STATE	INFLUENZA	CASES	37	1919-11-02	1919-11-08
191945	US	IL	ILLINOIS	STATE	INFLUENZA	CASES	56	1919-11-02	1919-11-08
191945	US	IN	INDIANA	STATE	INFLUENZA	CASES	29	1919-11-02	1919-11-08
191945	US	IA	IOWA	STATE	INFLUENZA	CASES	5	1919-11-02	1919-11-08
191945	US	LA	LOUISIANA	STATE	INFLUENZA	CASES	12	1919-11-02	1919-11-08

1-10 of 10,000 rows | 1-10 of 11 columns Previous **1** 2 3 4 5 6 ... 1000 Next

```
# Extract year and total cases for polio data
polio_data_line1 <- Polio_data %>%
  mutate(year = lubridate::year(from_date)) %>%
  group_by(year) %>%
  summarise(polio_cases = mean(number))

Mumps_data_line <- Mumps_data %>%
  mutate(year = lubridate::year(from_date)) %>%
  group_by(year) %>%
  summarise(mumps_cases = mean(number))

Influenza_data_line <- Influenza_data %>%
  filter(event == "CASES") %>%
  mutate(year = lubridate::year(from_date)) %>%
  group_by(year) %>%
  summarise(influenza_cases = mean(number))

# Merge polio, mumps, and influenza data
combined_data <- merge(polio_data_line1, Mumps_data_line, by = "year", all = TRUE)
combined_data <- merge(combined_data, Influenza_data_line, by = "year", all = TRUE)


ggplot(combined_data, aes(x = year)) +
  geom_point(aes(y = polio_cases, color = "Polio"), size = 2) +
  geom_point(aes(y = mumps_cases, color = "Mumps"), size = 2) +
  geom_point(aes(y = influenza_cases, color = "Influenza"), size = 2) +
  geom_line(aes(y = polio_cases, color = "Polio"), group = 1) +
  geom_line(aes(y = mumps_cases, color = "Mumps"), group = 1) +
  geom_line(aes(y = influenza_cases, color = "Influenza"), group = 1) +
  labs(x = "Year", y = "Total_Cases", title = "Total Polio, Mumps, and Influenza Cases per Year")
+
  scale_color_manual(values = c("Polio" = "blue", "Mumps" = "red", "Influenza" = "green")) +
  theme_minimal()
```

```
## Warning: Removed 34 rows containing missing values (`geom_point()`).
## Warning: Removed 45 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 61 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 34 rows containing missing values (`geom_line()`).
```

```
## Warning: Removed 11 rows containing missing values (`geom_line()`).
```

```
## Warning: Removed 61 rows containing missing values (`geom_line()`).
```

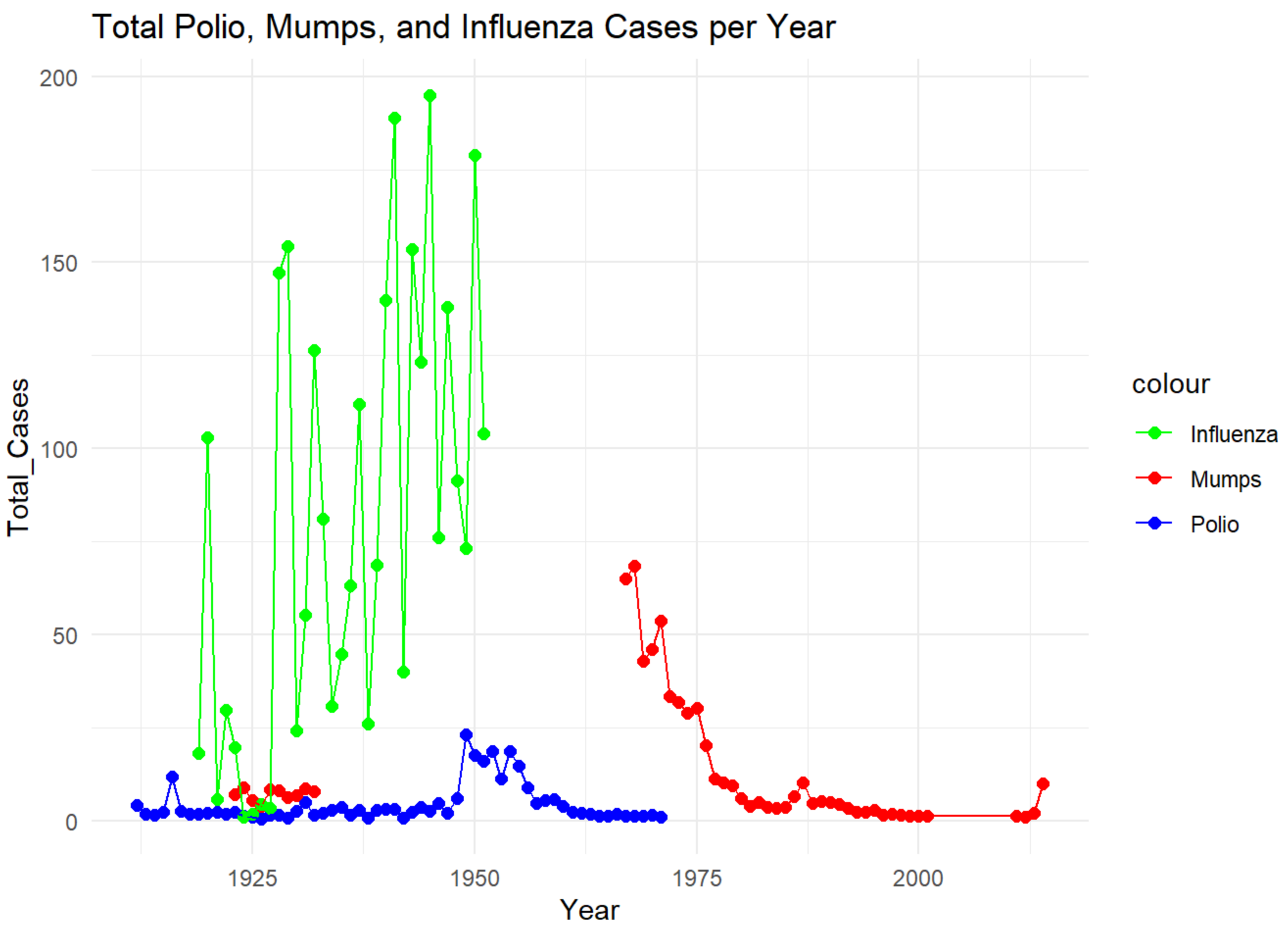


Fig. 7: Line

Graph of Compared Disease

Fig. 7 displays the output when the three diseases: influenza, mumps and polio are compared and tells us about the number of cases over the years.

###** HEATMAP **

Heatmap is basically a graphical representation of data using colors to visualize the value of the matrix.

```
library(ggplot2)
library(dplyr)
library(lubridate)
library(grid) #  contains a set of functions and classes that represent graphical objects, can be
manipulated like any other R object.
library(heatmaply) # provides an interface based around the plotly R package
```

```
## Warning: package 'heatmaply' was built under R version 4.3.1
## Loading required package: viridis
```

```
## Loading required package: viridisLite
```

```
##
## Attaching package: 'viridis'
```

```
## The following object is masked from 'package:maps':
##
##     unemp
```

```
##
## ======================
## Welcome to heatmaply version 1.4.2
##
## Type citation('heatmaply') for how to cite the package.
## Type ?heatmaply for the main documentation.
##
## The github page is: https://github.com/talgalili/heatmaply/
## Please submit your suggestions and bug-reports at: https://github.com/talgalili/heatmaply/issue
s
## You may ask questions at stackoverflow, use the r and heatmaply tags:
##   https://stackoverflow.com/questions/tagged/heatmaply
## ======================
```

```
polio_data_heatmap <- Polio_data %>%
  mutate(year = lubridate::year(from_date)) %>%
  group_by(year, disease, state) %>%
  summarise(total_cases = sum(number), .groups = "drop")

ggplot(polio_data_heatmap, aes(x = year, y = state, fill = total_cases)) +
  geom_tile(width = 0.9, height = 0.9, color = "black") +
  scale_fill_distiller(palette = "Spectral") +
  labs(x = "Year", y = "State", title = "Polio Cases Heatmap by State") +
  theme_minimal() +
  theme(axis.text.y = element_text(size=6))
```

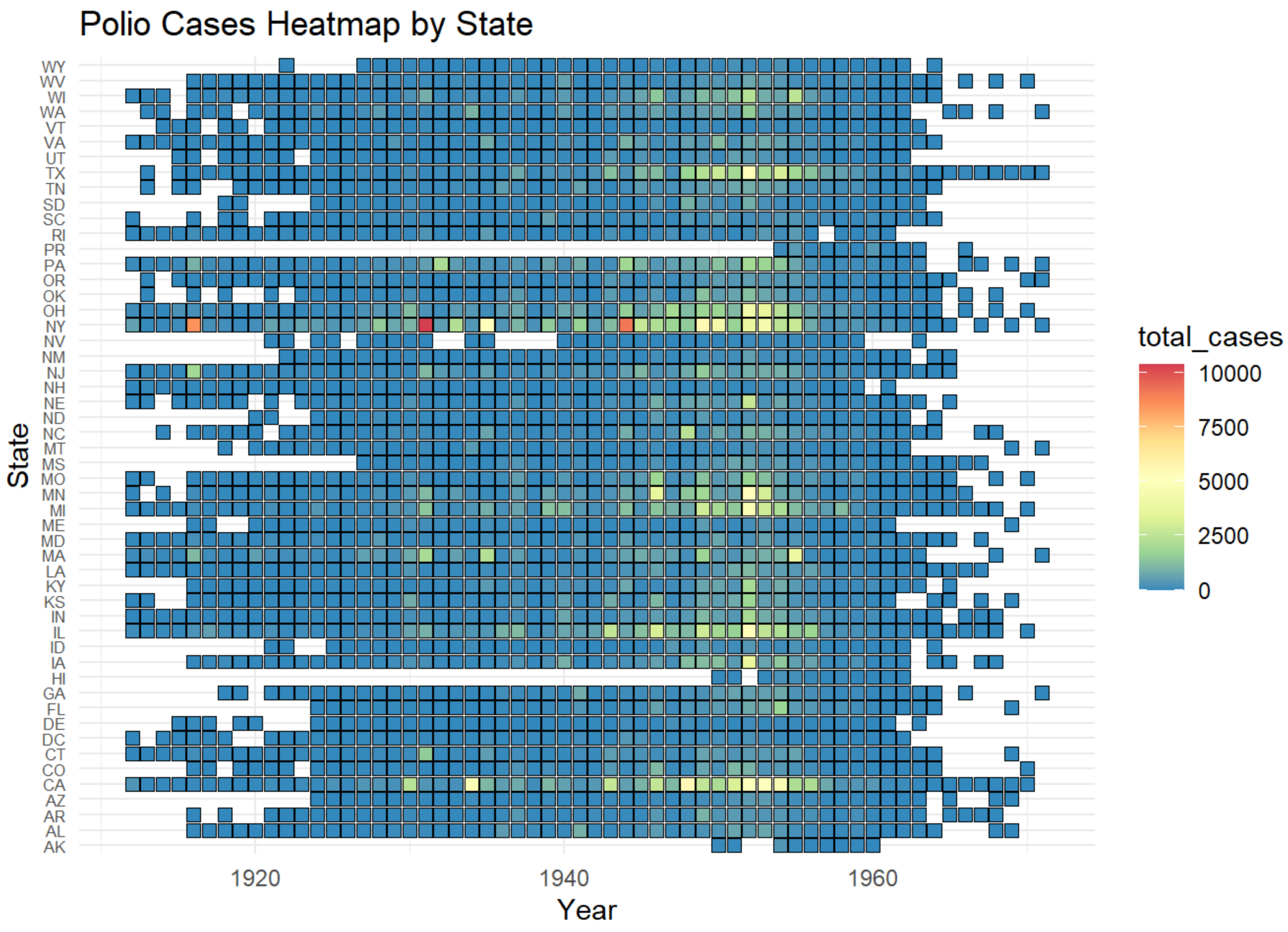


Fig. 8: Heatmap of effected US States over the Passing Years

The heatmap above displays a detailed view of the dataset, regarding in which year which state had the highest number of polio cases.

BAR PLOT AFTER POLIO VACCINE

```
library(ggplot2)
library(hrbrthemes)
library(lubridate)

polio_data_line <- Polio_data %>%
  mutate(year = lubridate::year(from_date)) %>%
  group_by(year) %>%
  summarise(total_cases = sum(number)) %>%
  mutate(vaccine_year = ifelse(year == 1953, total_cases, 0))

ggplot(polio_data_line, aes(x = year, y = total_cases)) +
  geom_col(fill = "blue") +
  geom_col(data = polio_data_line, aes(x = year, y = vaccine_year), fill = "red") +
  labs(x = "Year", y = "Total Cases", title = "Total Polio Cases per Year") +
  theme_minimal() +
  coord_cartesian(ylim = c(0, max(polio_data_line$total_cases, polio_data_line$vaccine_year) * 1.
1))
```

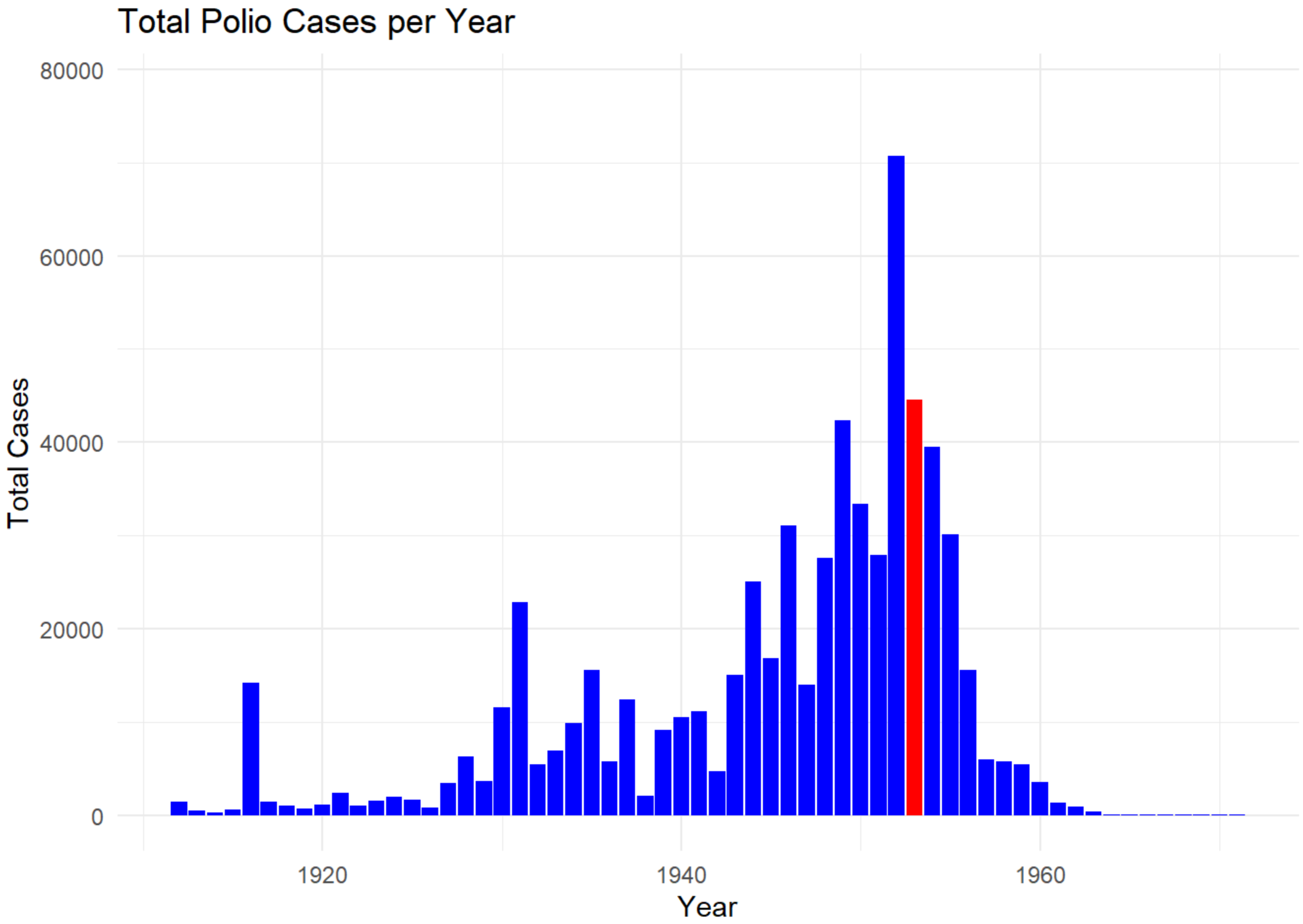


Fig. 9: Bar Plot after Vaccine was Discovered

The above bar chart illustrates that during the 1950’s, the polio virus was at its peak but soon around 1953 the polio vaccine was discovered in the US. After applying the vaccine the total number of cases drastically dropped and kept dropping in the years ahead.

6. Result

The dataset was thoroughly analyzed and during the study our main focus was on the following points:

- visualizing total cases of polio virus in each state
- displaying number of Polio cases over every passing year in the US state starting from 1910 to onward
- comparing multiple common diseases of the US with our specific disease ‘Polio’.
- displaying drastic change after polio vaccine was invented

7. Discussion

Contradictions

The only contradiction that i felt during the study was that in the dataset symptoms of the diseases were not mentioned, if they were present we could further do a better analysis on the dataset.

Pros

1. The dataset is large enough to get a proper visualization of the objectives of the study.
2. With the lockdown restriction timeline data available to us, we can expand this analysis further in order to gain even more insights

8. Conclusion

This analysis gives us the visuals of Polio virus.The analysis sheds light on how many cases were there of Polio virus in the states of US. Our graphs show that during the year 1950, there was a spike in the number of cases of Polio, but later dropped down. After the polio vaccine was discovered in 1953, cases started dropping gradually. It is now advised that polio vaccines be given to children of younger age to avoid the virus in the older age.

** Literature **

1-Poliomyelitis https://www.who.int/news-room/fact-sheets/detail/poliomyelitis (https://www.who.int/news-room/fact-sheets/detail/poliomyelitis)

2- What is Polio? https://www.cdc.gov/polio/what-is-polio/index.htm# (https://www.cdc.gov/polio/what-is-polio/index.htm#):~:text=Polio%2C%20or%20poliomyelitis%2C%20is%20a,move%20parts%20of%20the%20body).

3- Polio by Saloni Dattani, Fiona Spooner, Sophie Ochmann and Max Roser https://ourworldindata.org/polio (https://ourworldindata.org/polio)