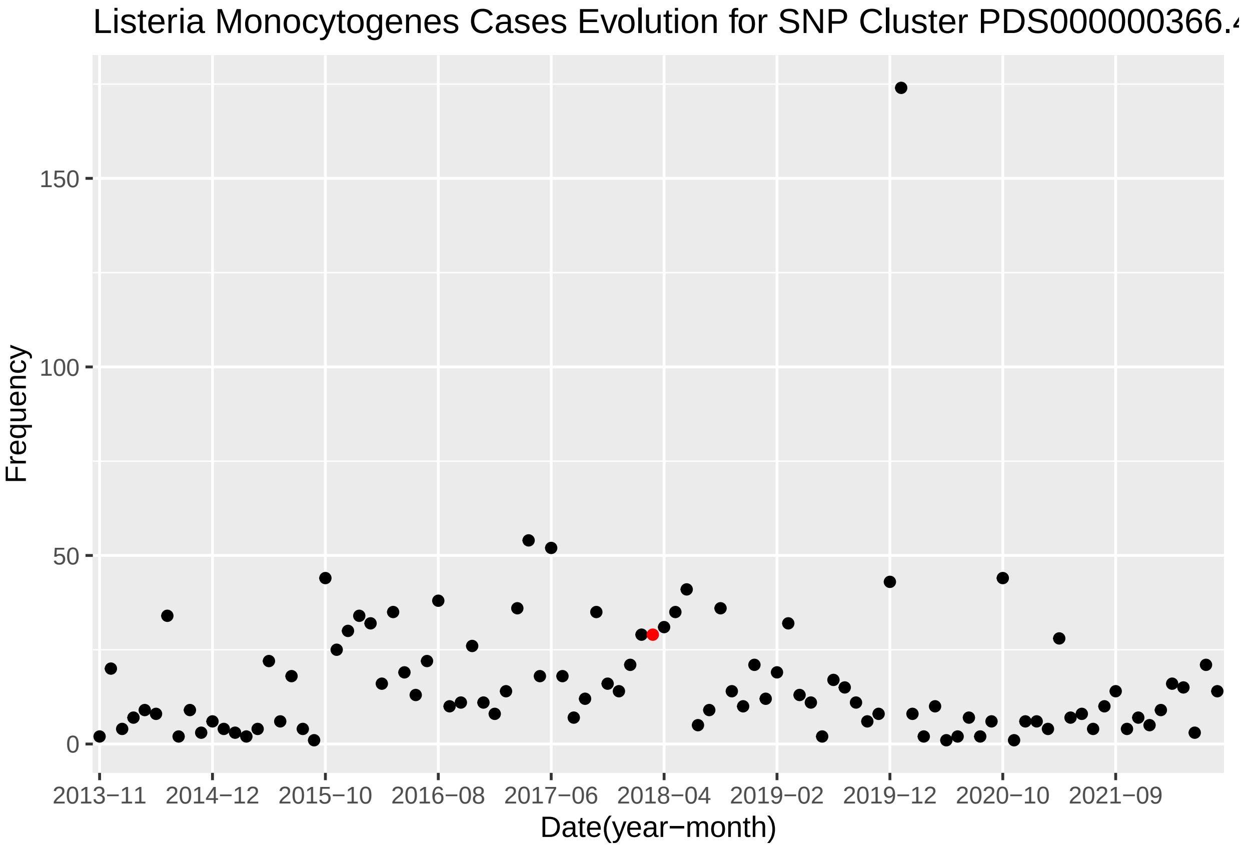
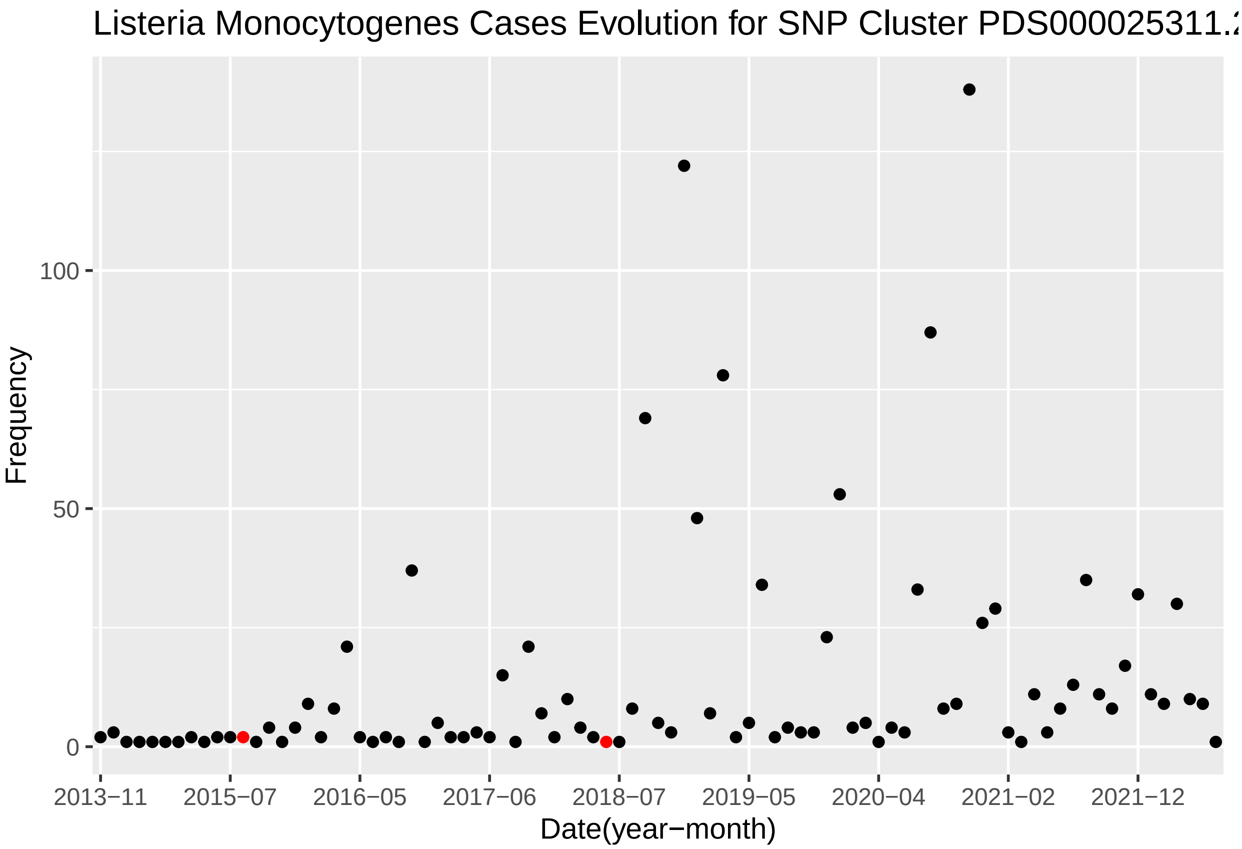
SNP Cluster Analysis by Month and by Week

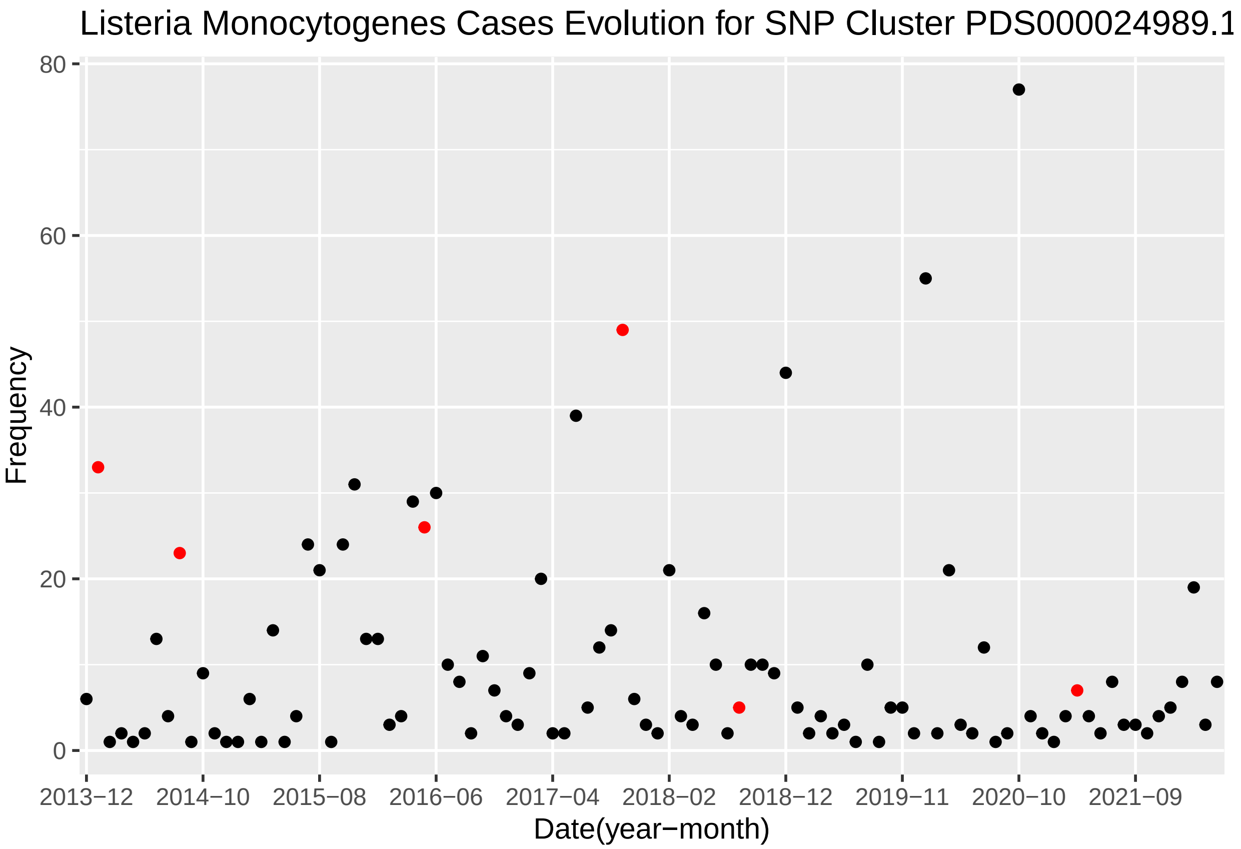
I am going to visualize the evolution of cases within each SNP cluster for Listeria Monocytogenes with month as an interval unit.



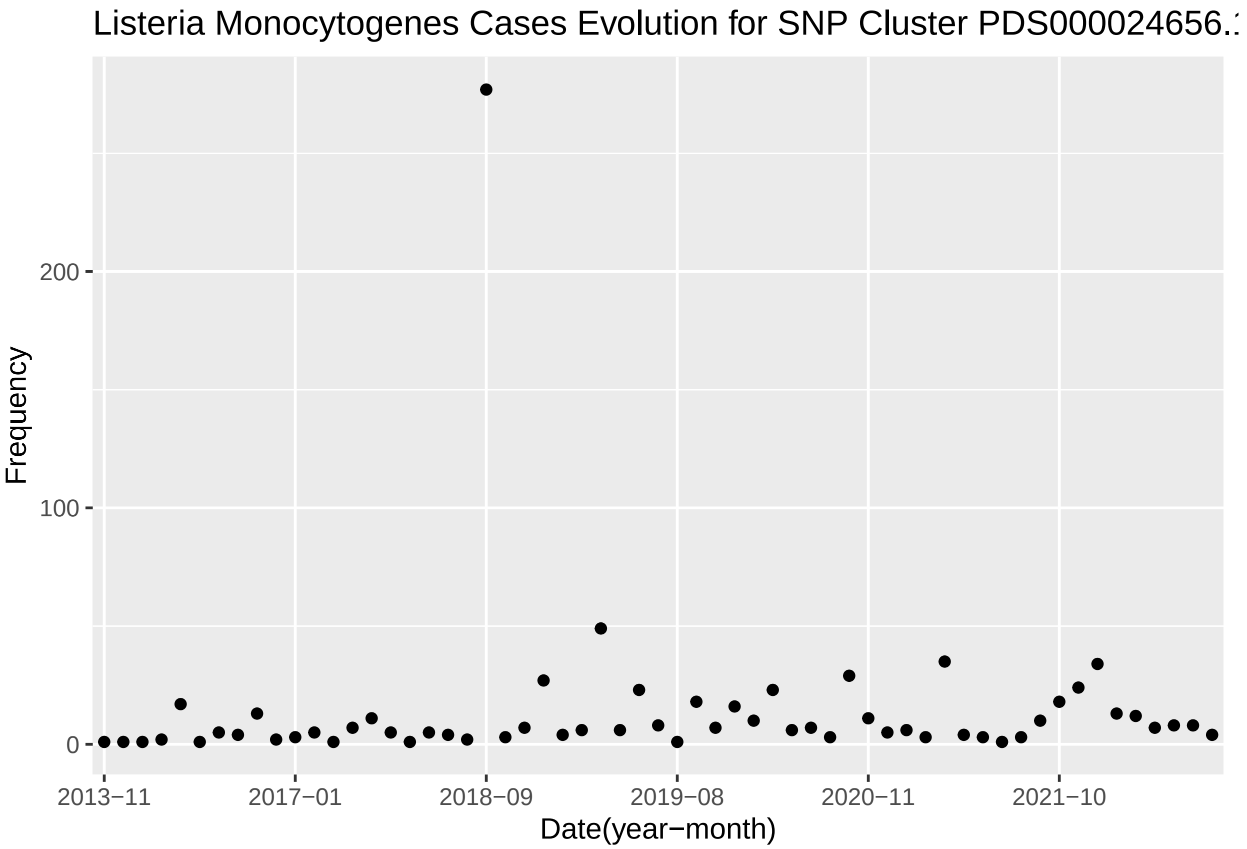
```
## [1] "SNP Cluster PDS000000366.488 has the highest cases of listeria monocytogenes at 2020-01"
```



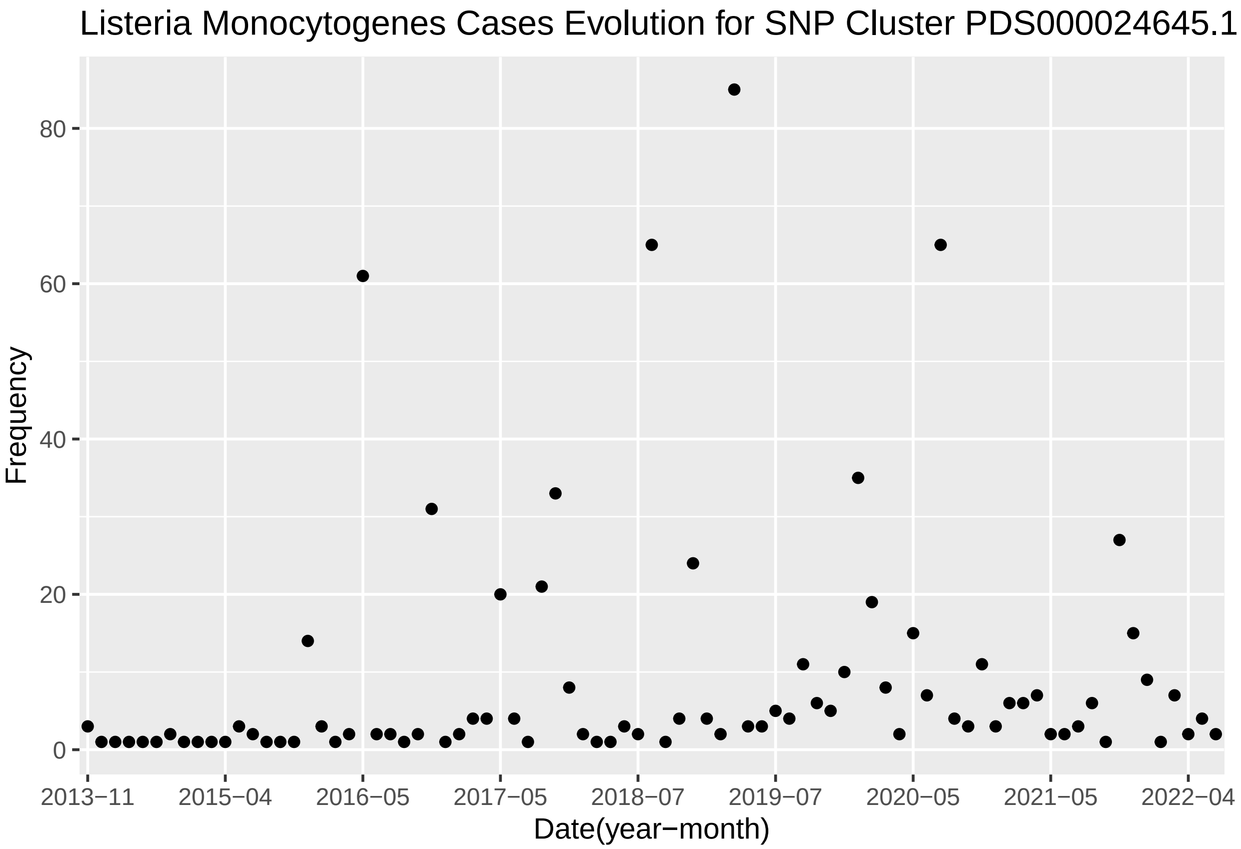
```
## [1] "SNP Cluster PDS000025311.237 has the highest cases of listeria monocytogenes at 2020-11"
```



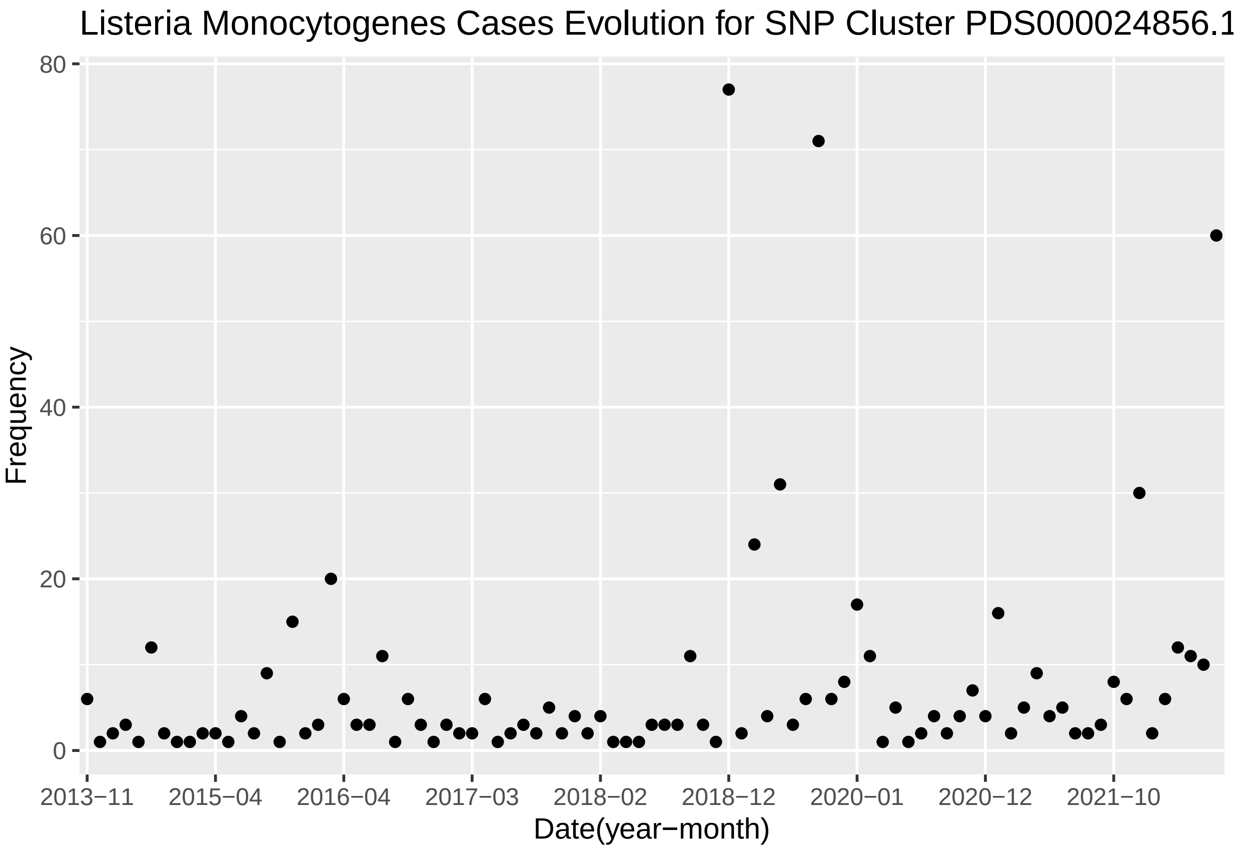
```
## [1] "SNP Cluster PDS000024989.118 has the highest cases of listeria monocytogenes at 2020-10"
```



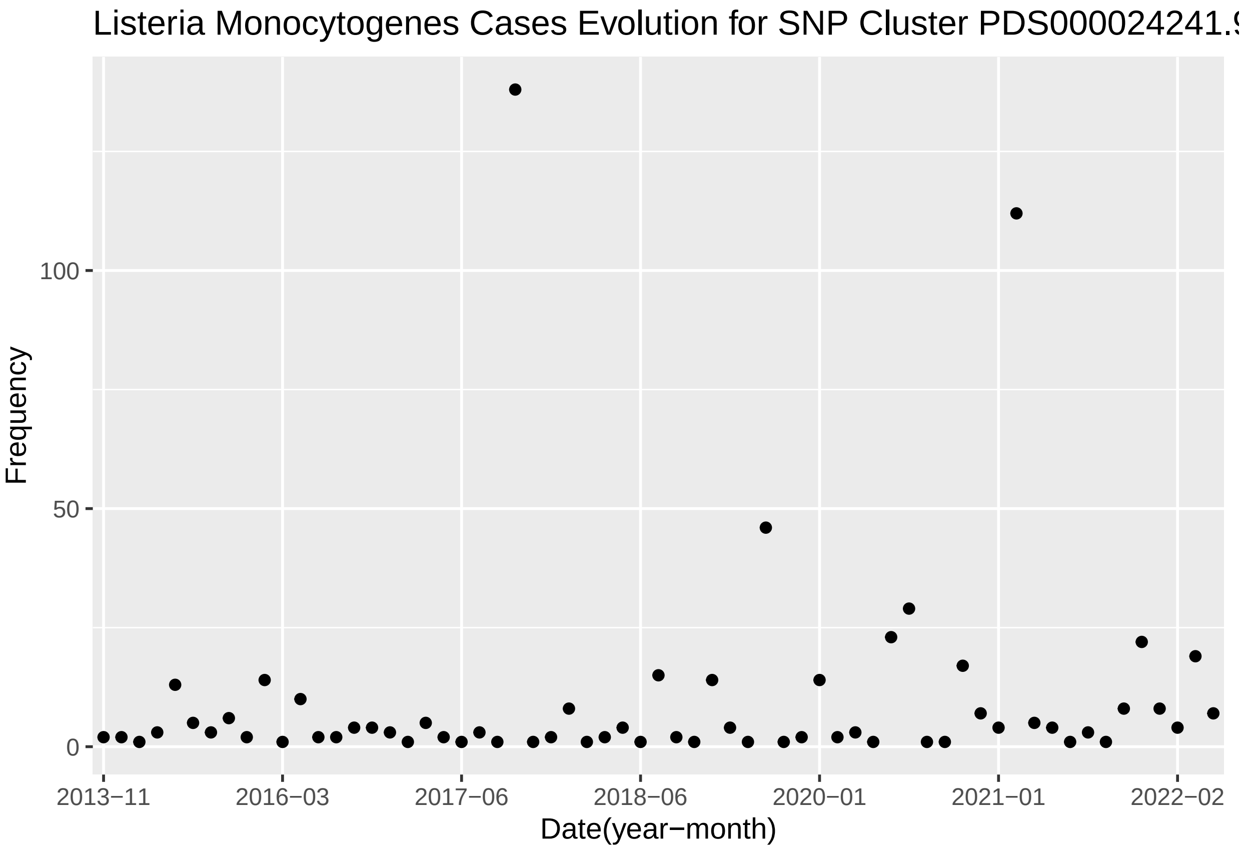
```
## [1] "SNP Cluster PDS000024656.169 has the highest cases of listeria monocytogenes at 2018-09"
```



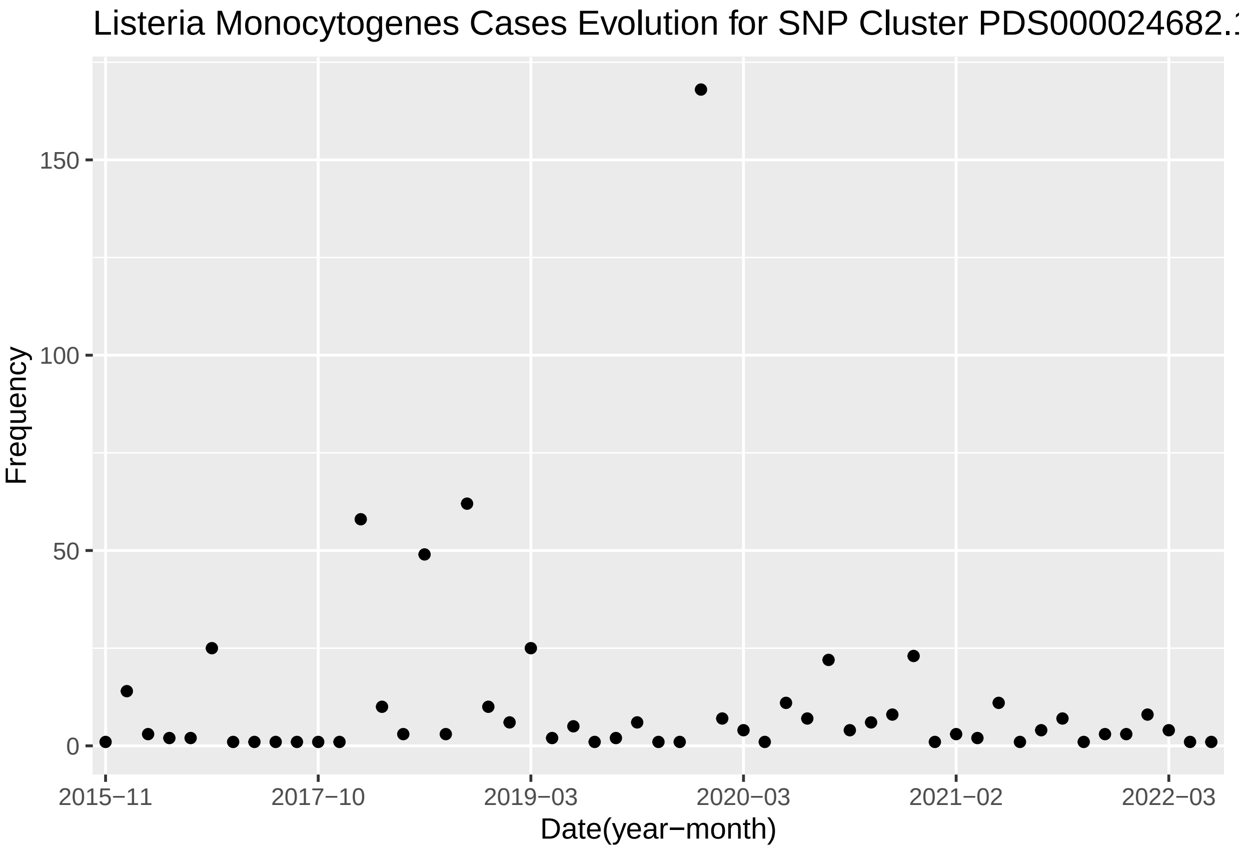
```
## [1] "SNP Cluster PDS000024645.140 has the highest cases of listeria monocytogenes at 2019-03"
```



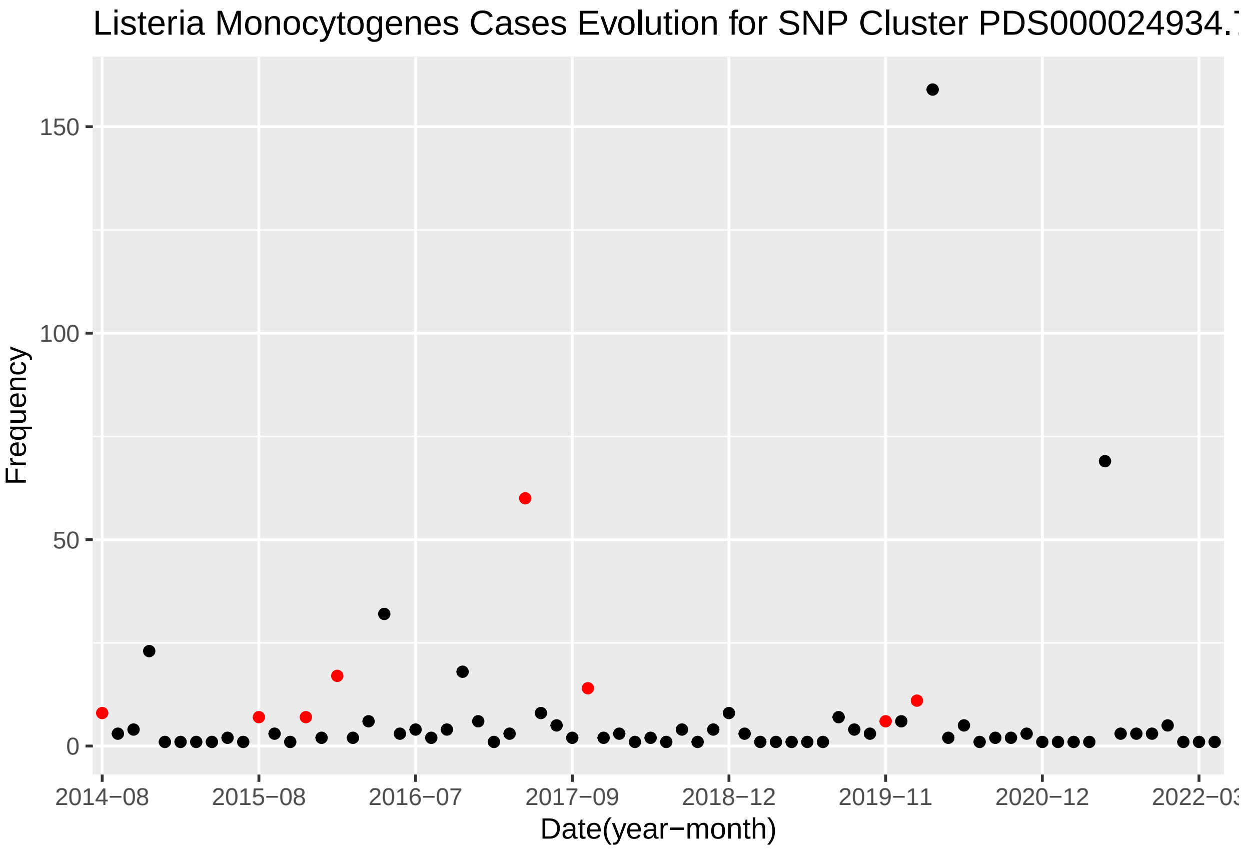
```
## [1] "SNP Cluster PDS000024856.153 has the highest cases of listeria monocytogenes at 2018-12"
```



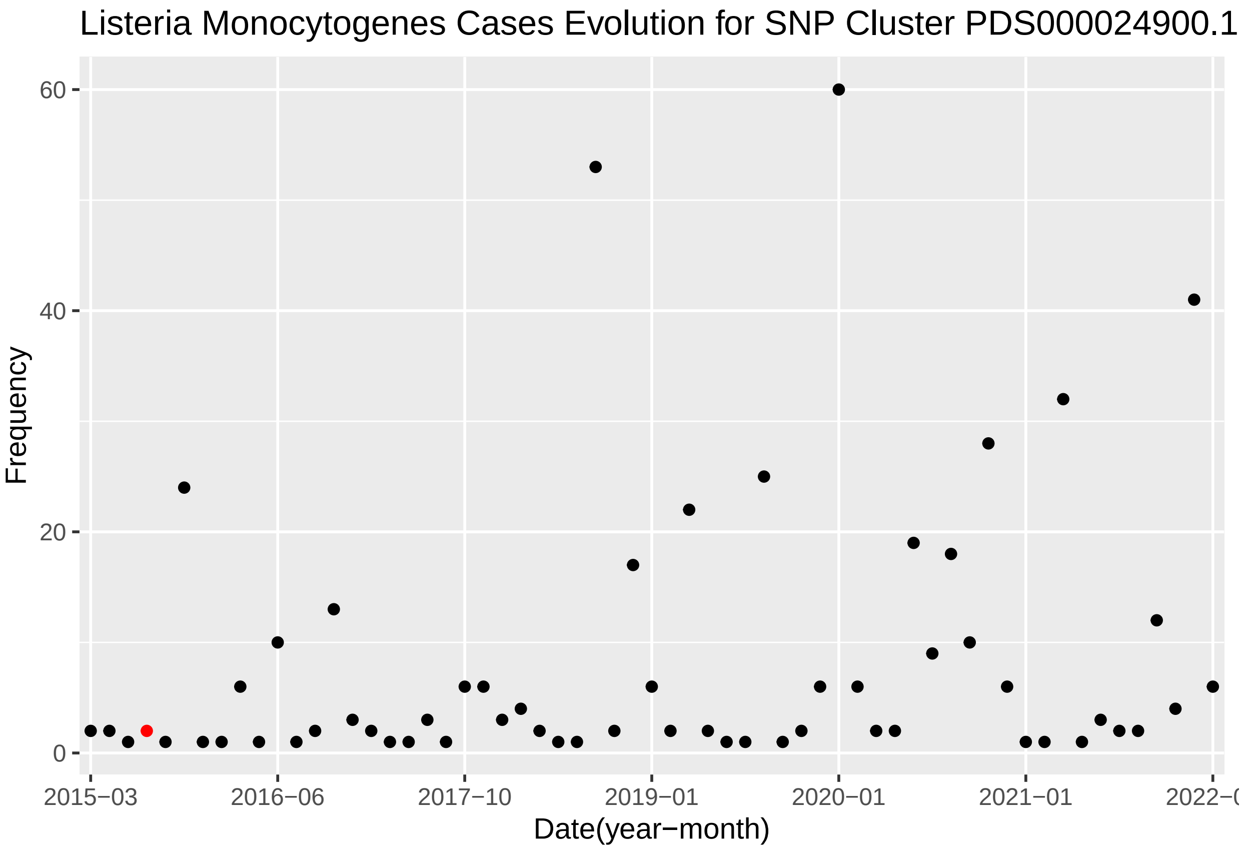
```
## [1] "SNP Cluster PDS000024241.94 has the highest cases of listeria monocytogenes at 2017-10"
```



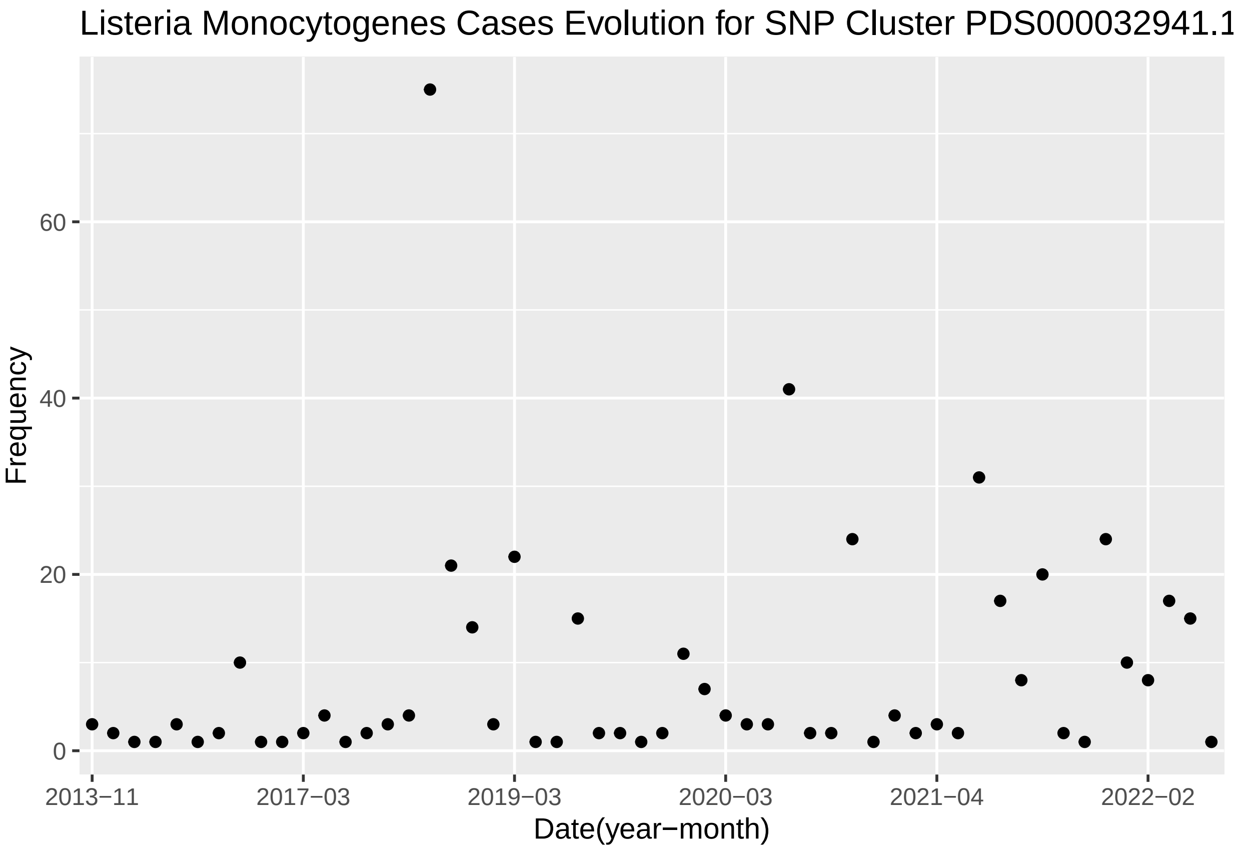
```
## [1] "SNP Cluster PDS000024682.133 has the highest cases of listeria monocytogenes at 2020-01"
```



```
## [1] "SNP Cluster PDS000024934.77 has the highest cases of listeria monocytogenes at 2020-02"
```



```
## [1] "SNP Cluster PDS000024900.112 has the highest cases of listeria monocytogenes at 2020-01"
```

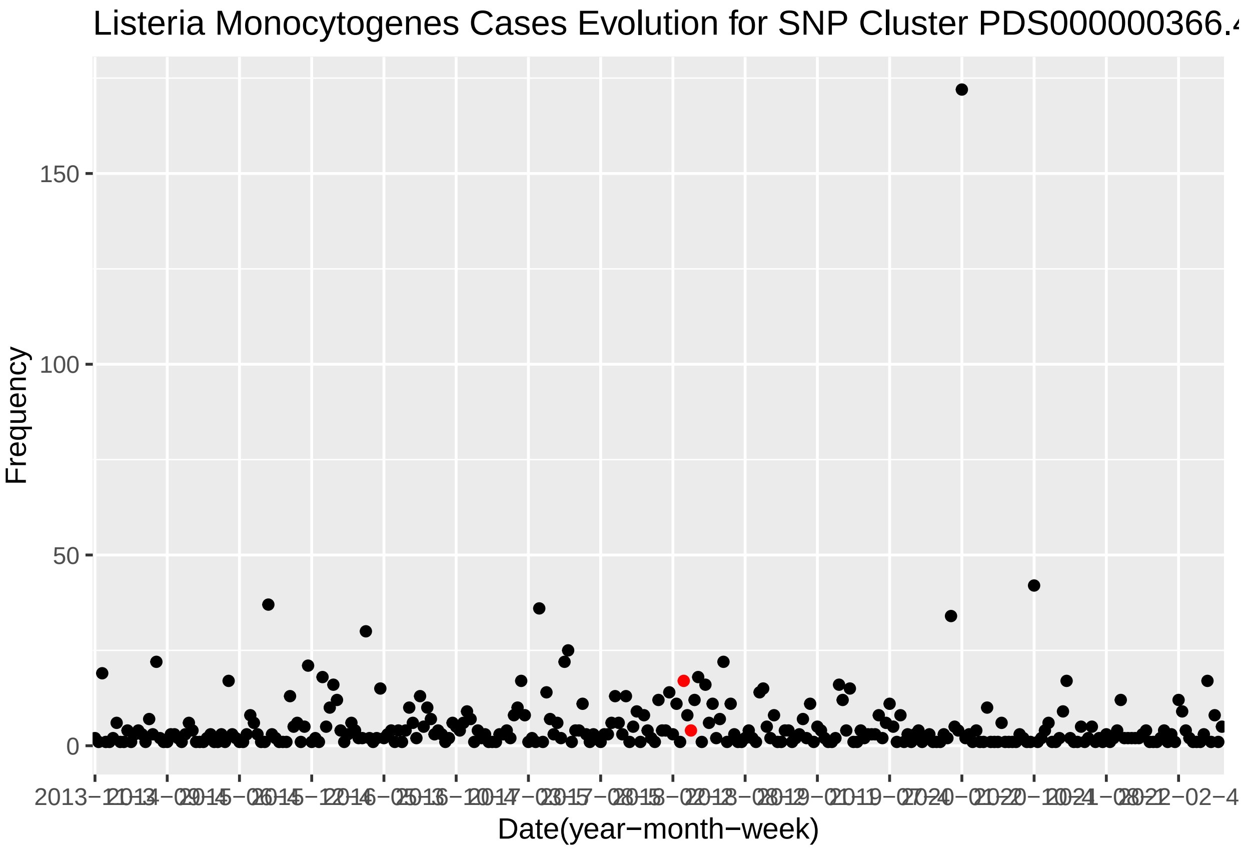


```
## [1] "SNP Cluster PDS000032941.132 has the highest cases of listeria monocytogenes at 2018−09"
```

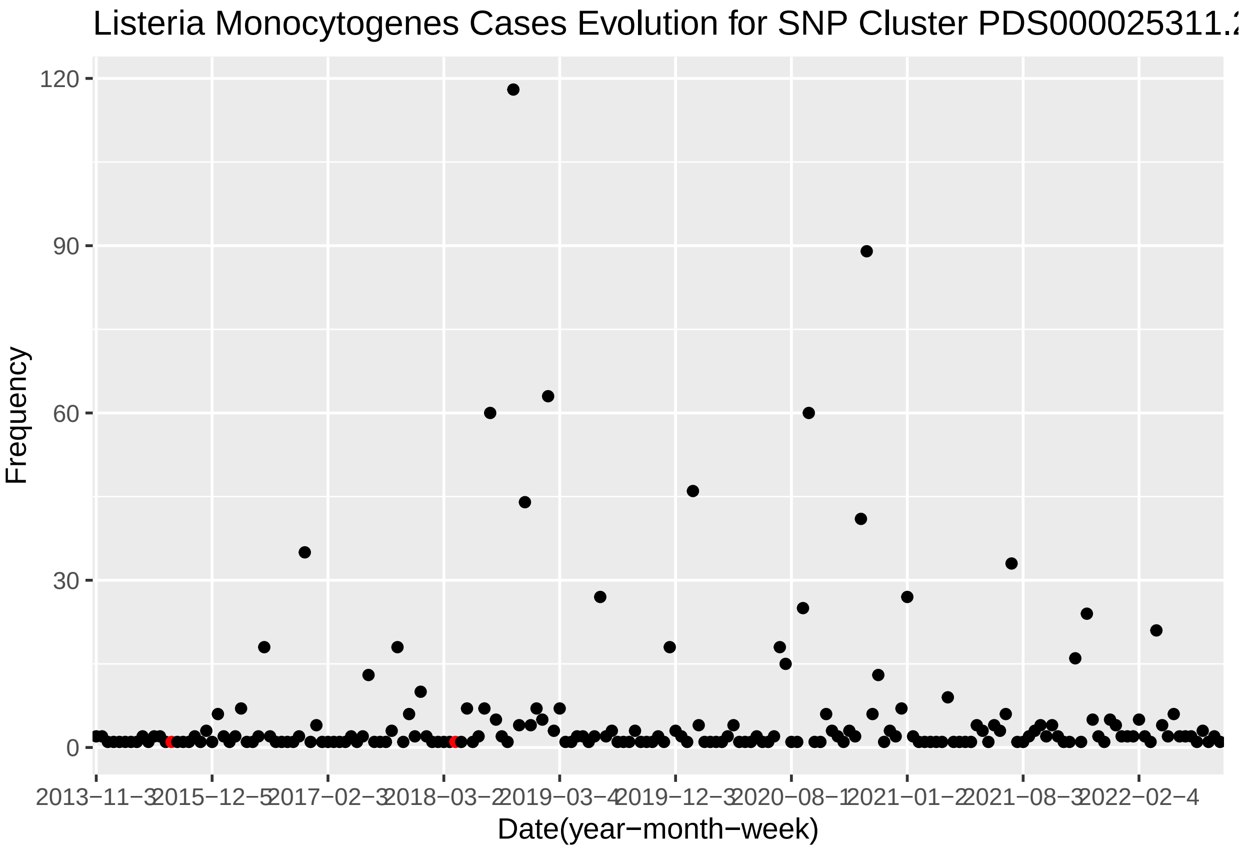
All the red dots in the scatter plot indicating that there is an outbreak for the listeria monocytogenes.

For clusters PDS000000366.488(1), PDS000024682.133(8), and PDS000024900.112(10), they all had highest cases of listeria monocytogenes at 2020-01. Clusters PDS000024656.169(4) and PDS000032941.132(11) had highest cases of listeria monocytogenes at 2018-09. All the other clusters had the highest cases of the disease at a separate date(year-month). Dates 2020-01 and 2018-09 need further investigation since there are several clusters who had the highest cases of the disease in these two months.

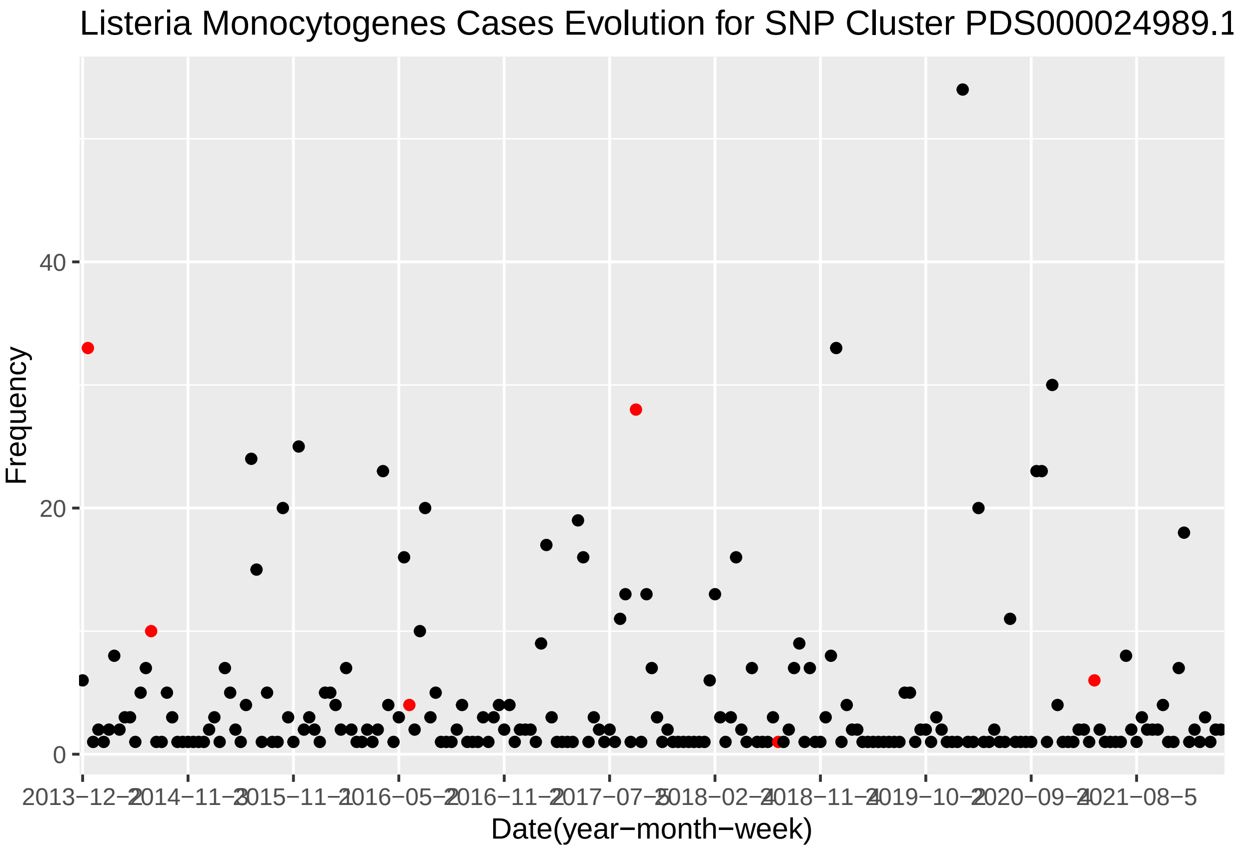
Next, I am going to visualize the evolution of cases within each SNP cluster for Listeria Monocytogenes with week as an interval unit. For each month, I coded date 1 to date 7 as the first week; date 8 to date 14 as the second week; date 15 to date 21 as the third week; date 22 to date 28 as the fourth week; and the rest of the day within each month as the fifth week.



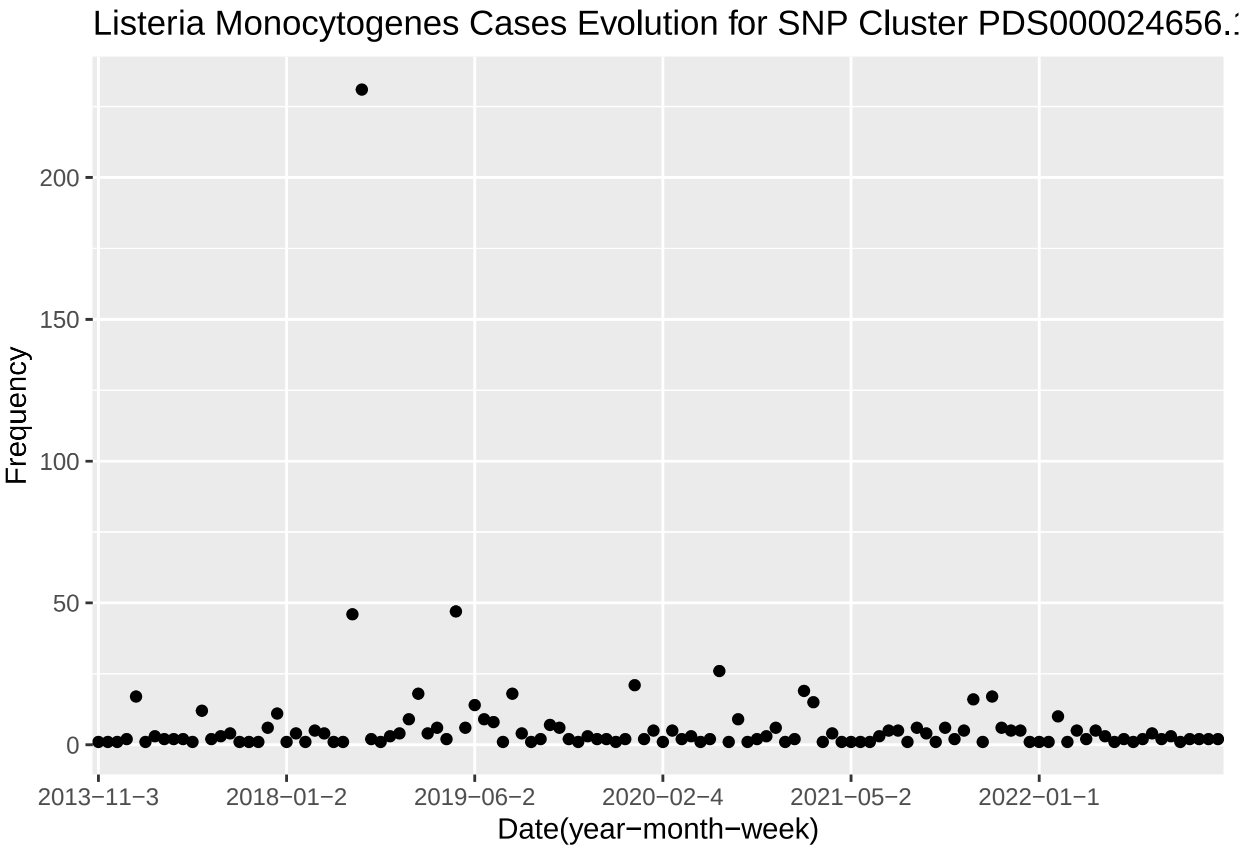
```
## [1] "SNP Cluster PDS000000366.488 has the highest cases of listeria monocytogenes at 2020-01-2"
```



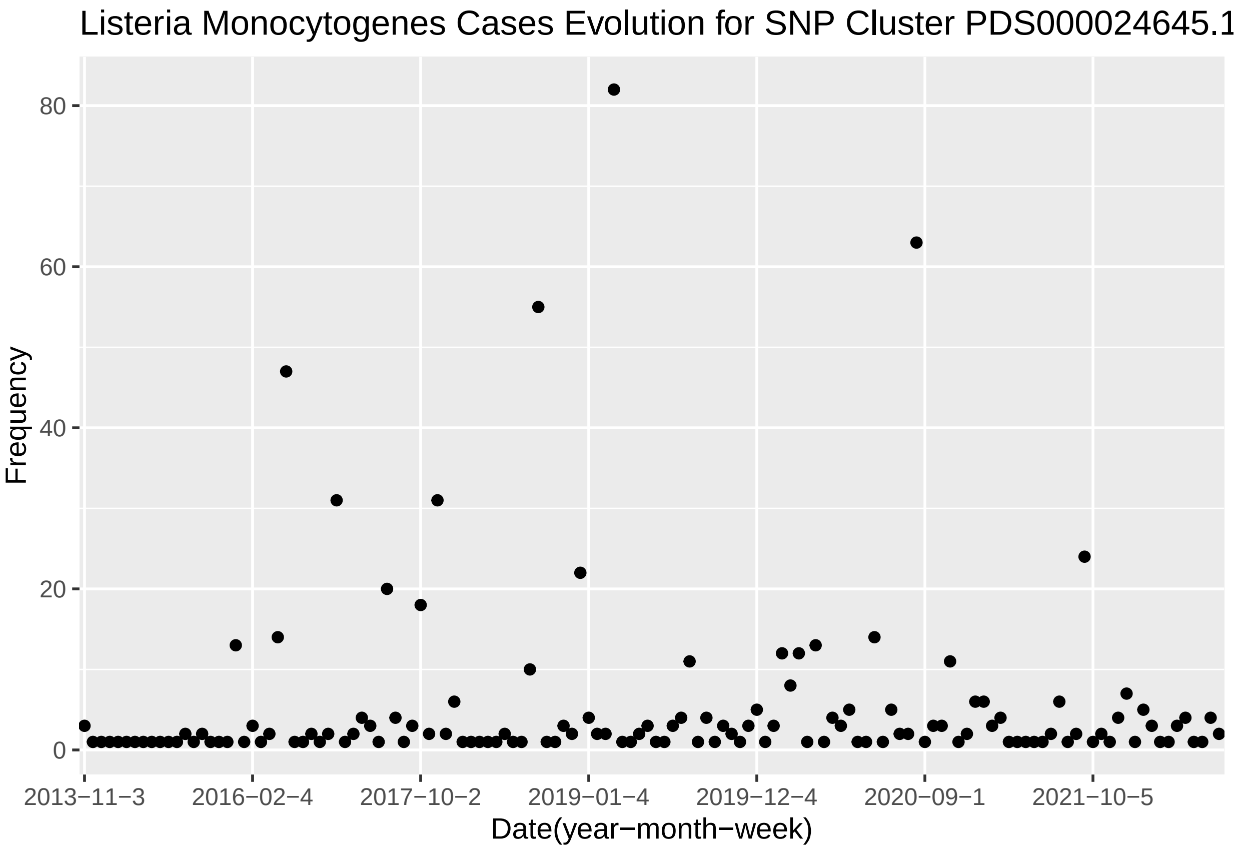
```
## [1] "SNP Cluster PDS000025311.237 has the highest cases of listeria monocytogenes at 2018-12-3"
```



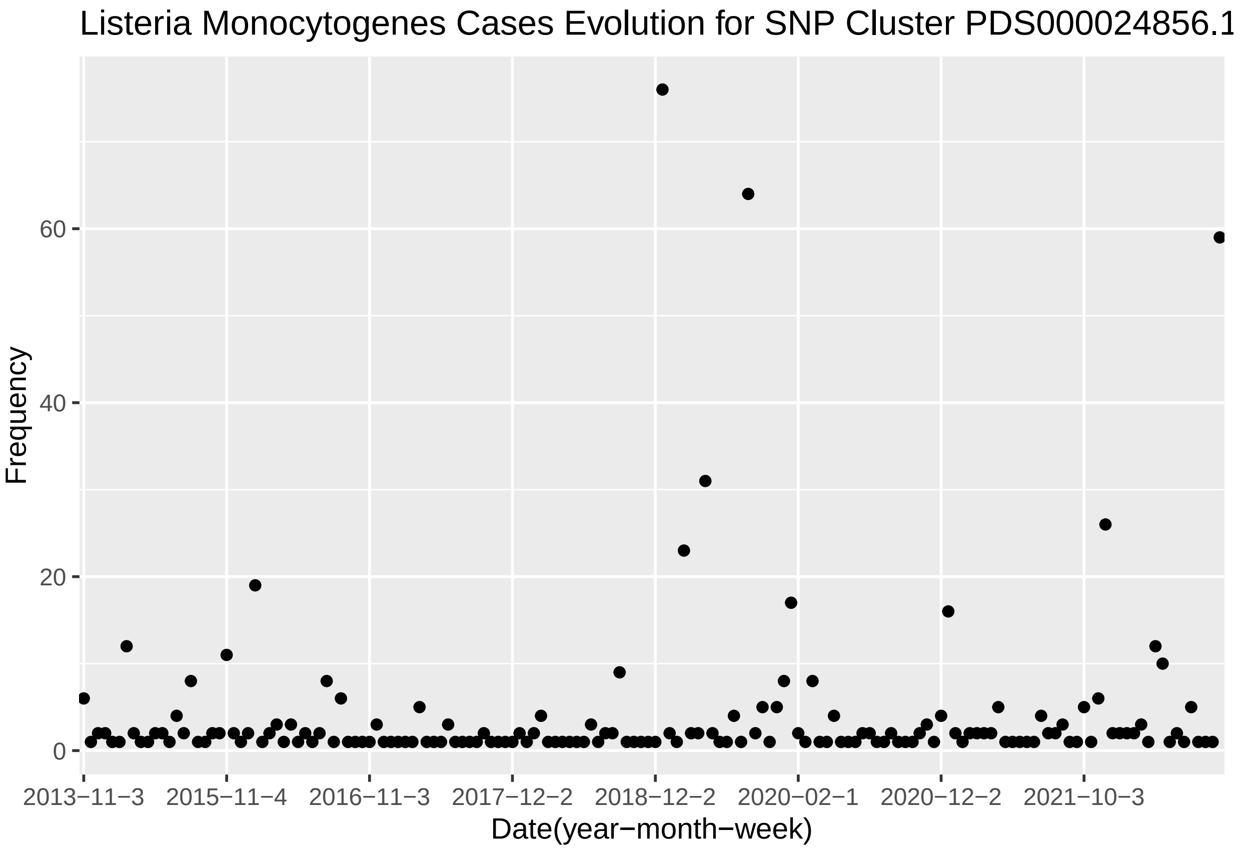
```
## [1] "SNP Cluster PDS000024989.118 has the highest cases of listeria monocytogenes at 2020-01-2"
```



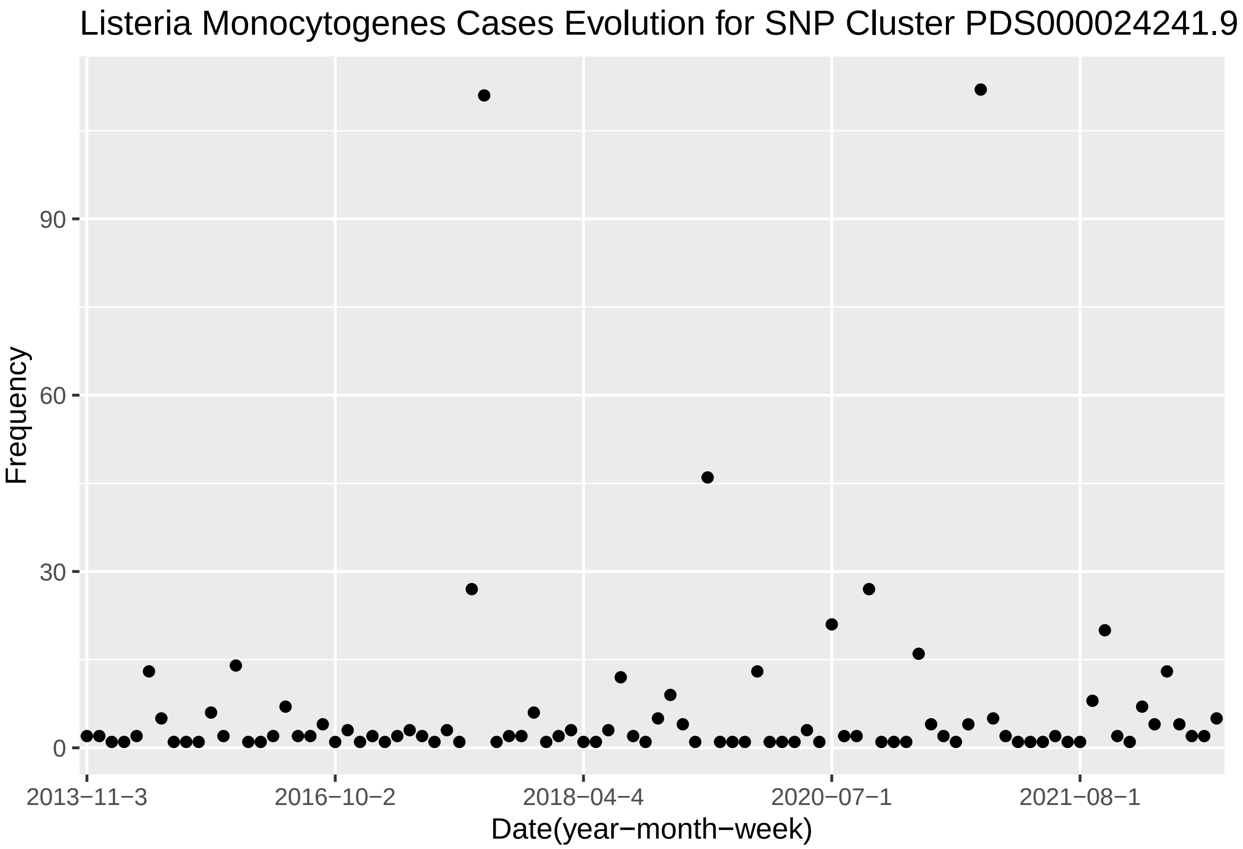
```
## [1] "SNP Cluster PDS000024656.169 has the highest cases of listeria monocytogenes at 2018-09-3"
```



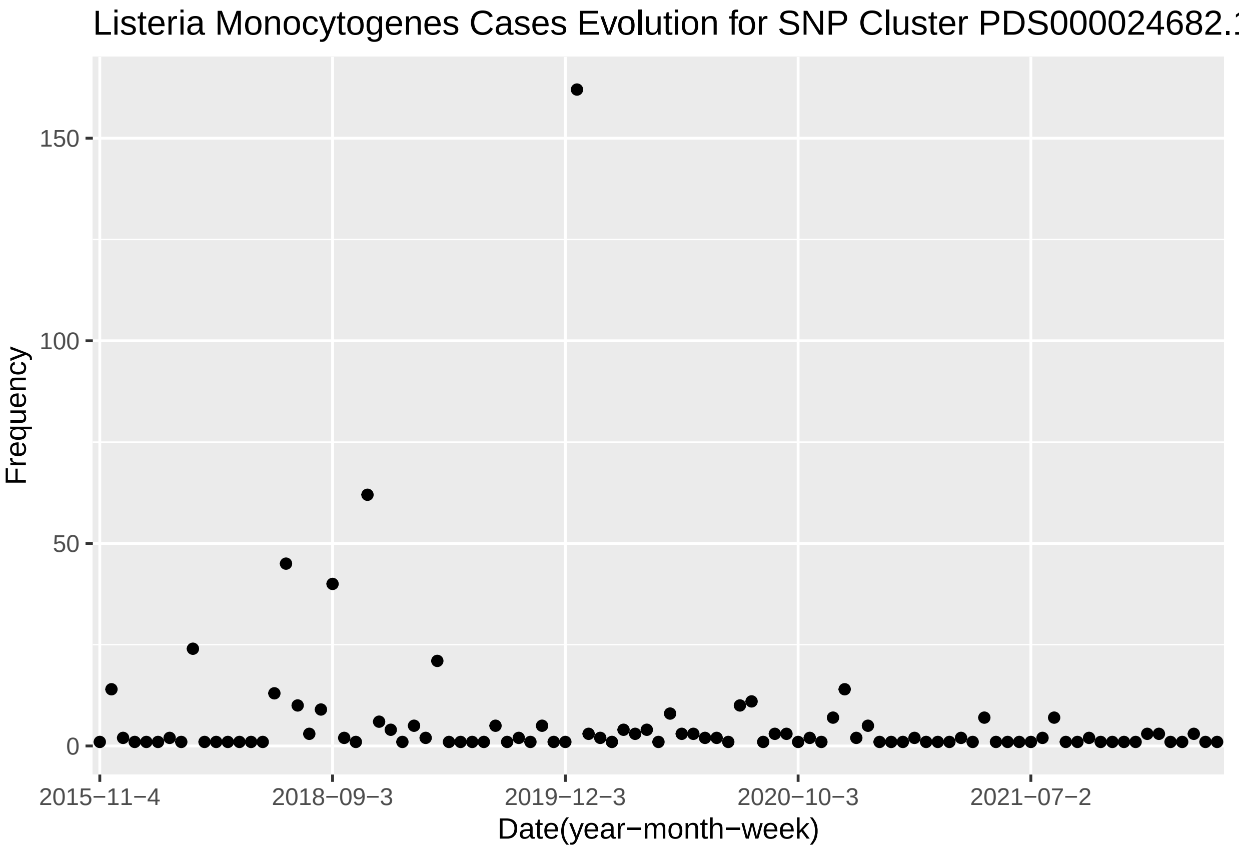
```
## [1] "SNP Cluster PDS000024645.140 has the highest cases of listeria monocytogenes at 2019-03-2"
```



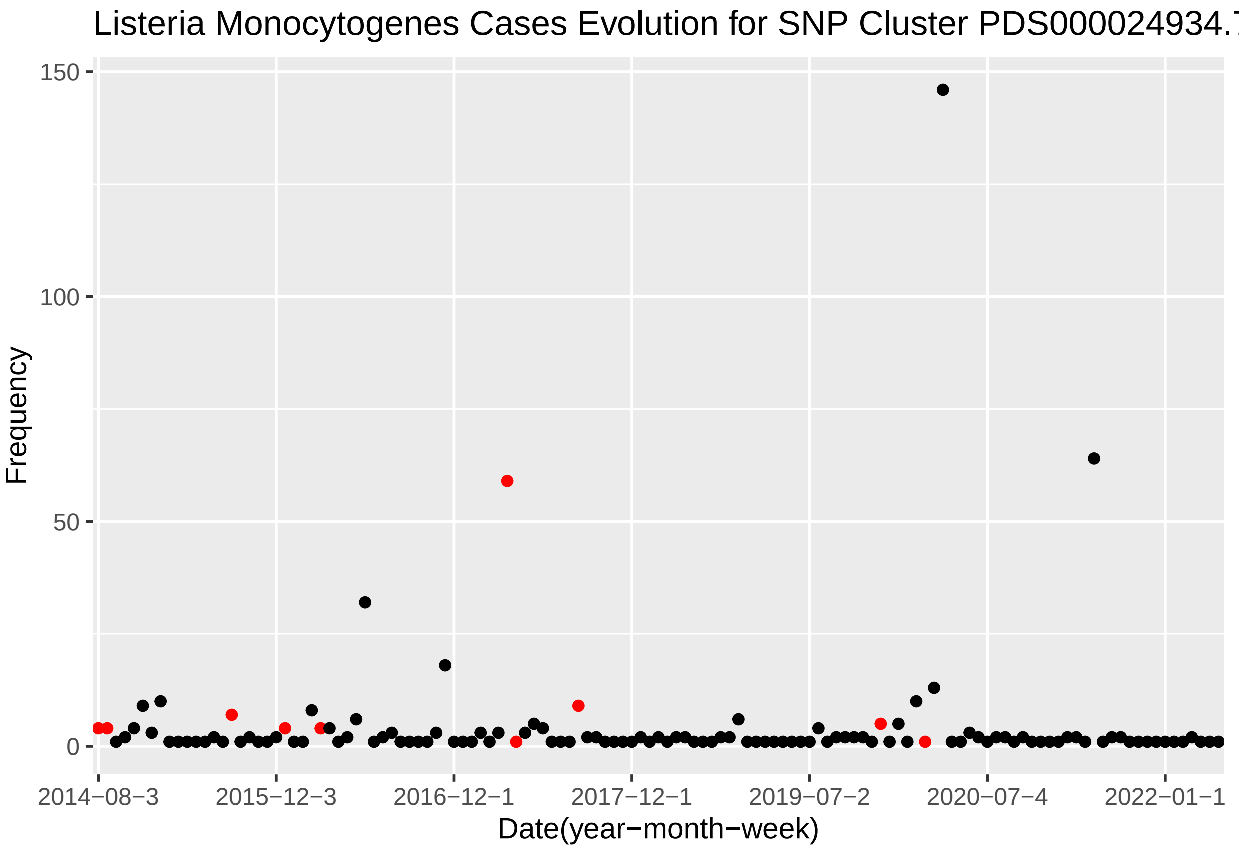
```
## [1] "SNP Cluster PDS000024856.153 has the highest cases of listeria monocytogenes at 2018-12-3"
```



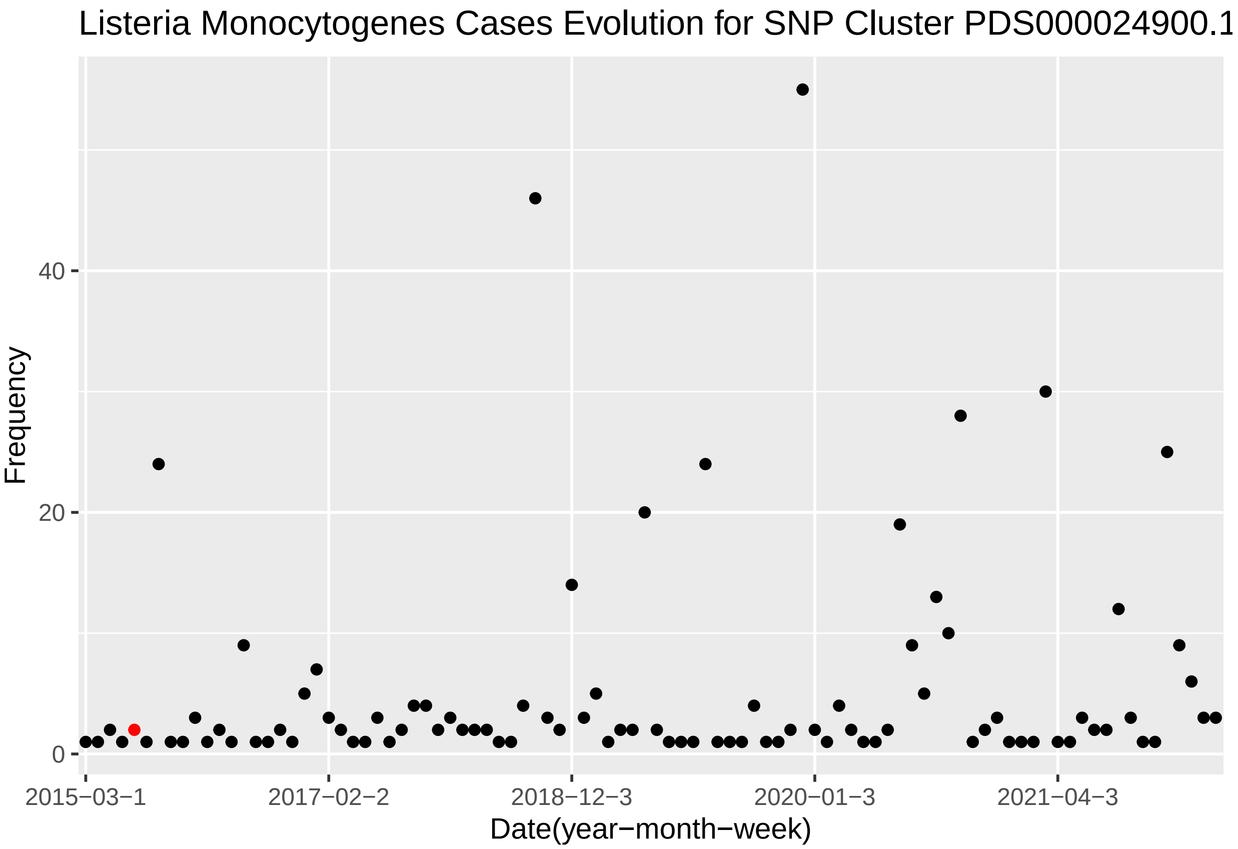
```
## [1] "SNP Cluster PDS000024241.94 has the highest cases of listeria monocytogenes at 2021-02-3"
```



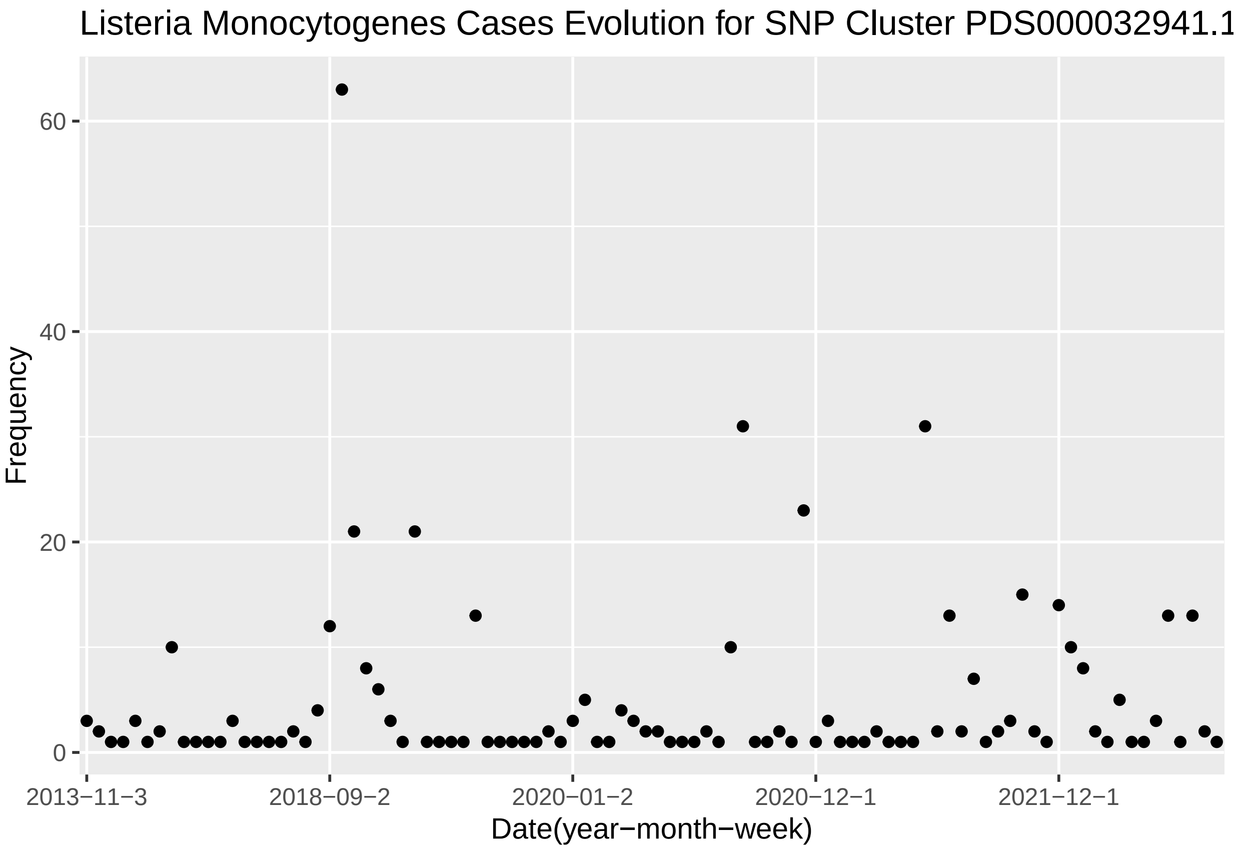
```
## [1] "SNP Cluster PDS000024682.133 has the highest cases of listeria monocytogenes at 2020-01-2"
```



```
## [1] "SNP Cluster PDS000024934.77 has the highest cases of listeria monocytogenes at 2020-02-4"
```



```
## [1] "SNP Cluster PDS000024900.112 has the highest cases of listeria monocytogenes at 2020-01-2"
```



```
## [1] "SNP Cluster PDS000032941.132 has the highest cases of listeria monocytogenes at 2018-09-3"
```

All the red dots in the scatter plot indicating that there is an outbreak for the listeria monocytogenes.

Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library(naniar)
library(readr)
library(dplyr)
library(ggplot2)
library(tableone)
setwd("~/Desktop")
isolates <- read_csv("isolates.csv")
isolates = isolates %>%
  select(-c(Computed_types, Virulence_genotypes, AST_phenotypes))

isolates = isolates %>%
  select(-c(Host_disease, PFGE_secondary_enzyme_pattern, PFGE_primary_enzyme_pattern, Stress_genotypes,

isolates = isolates %>%
  select(-c(Species_TaxID, `K-mer_group`, Organism_group))

isolates = isolates %>%
  select(-c(WGS_accession, WGS_prefix, Run, Isolate, Assembly))
```

```
isolates = isolates %>%
  select(-c(AMRFinderPlus_version, PD_Ref_Gene_Catalog_version, Level))

isolates <- isolates %>%
    mutate(across(.cols=c(Library_layout, Method, SRA_Center, Platform, AMR_genotypes_core, BioProject,

isolates <- isolates %>%
    mutate(across(.cols=c(SRA_release_date, Create_date), .fns = as.Date))


isolates = isolates %>%
  select(-c(Library_layout, Method, Platform, AMRFinderPlus_analysis_type, Isolate_identifiers, BioSampl

isolates = isolates %>%
  select(-Strain)

isolates$Outbreak = ifelse(is.na(isolates$Outbreak), 0, 1)
count_SNP = as.data.frame(table(isolates$SNP_cluster))
colnames(count_SNP)[colnames(count_SNP) == "Var1"] <- "SNP_cluster"
colnames(count_SNP)[colnames(count_SNP) == "Freq"] <- "Frequency"
count_SNP =count_SNP[order(-count_SNP$Frequency),]

count_SNP_20 = count_SNP[1:20,]
SNP_percentage = numeric(20)
for (i in 1:20){
  SNP_percentage[i] = (count_SNP$Frequency[i]/sum(count_SNP$Frequency))*100
}
count_SNP_20['SNP_percentage'] <- SNP_percentage
for (i in 1:11){
  new_cluster = isolates %>%
    filter((SNP_cluster == count_SNP_20[i,1]))

  new_cluster$Create_date_YM = format(as.Date(new_cluster$Create_date), "%Y-%m")

  count_date = as.data.frame(table(new_cluster$Create_date_YM))
  colnames(count_date)[colnames(count_date) == "Var1"] <- "Date"
  colnames(count_date)[colnames(count_date) == "Freq"] <- "Frequency"

  count_date$red = 0
  for (j in 1:dim(new_cluster)[1]){
    if(new_cluster$Outbreak[j] == 1){
      number = which(count_date$Date == noquote(new_cluster$Create_date_YM[j]))
      count_date$red[number] = 1
      }
    }


  cluster_name = count_SNP_20[i,1]

  print(ggplot(data = count_date) +
          geom_point (mapping = aes (x=Date, y=Frequency), color=ifelse(count_date$red == 1, "red", "bla
          scale_x_discrete(breaks = count_date$Date[seq(1, length(count_date$Date), by = 10)]) +
          ggtitle(paste("Listeria Monocytogenes Cases Evolution for SNP Cluster", cluster_name)) +
          labs(x = 'Date(year-month)', y = 'Frequency'))


  count_date = count_date[order(-count_date$Frequency),] # order returns indexes
  print(sprintf("SNP Cluster %s has the highest cases of listeria monocytogenes at %s", cluster_name, c
}
for (i in 1:11){
  new_cluster = isolates %>%
    filter((SNP_cluster == count_SNP_20[i,1]))
  new_cluster$Create_date = format(as.Date(new_cluster$Create_date), "%Y-%m-%d")
  new_cluster$Create_date_YM = format(as.Date(new_cluster$Create_date), "%Y-%m")

  for (j in 1:dim(new_cluster[1])){
  date = as.numeric(format(as.Date(new_cluster$Create_date[j]), "%d"))
  new_cluster$week[j] = if(date >= 1 && date <= 7){
    1
  } else if(date >= 8 && date <= 14){
    2
  } else if(date >= 15 && date <= 21){
    3
  } else if(date >= 22 && date <= 28){
    4
  } else{
    5
  }
  new_cluster$Create_date_YMW[j] = sprintf("%s-%s", new_cluster$Create_date_YM[j], new_cluster$week[j])
  }

  count_date = as.data.frame(table(new_cluster$Create_date_YMW))
  colnames(count_date)[colnames(count_date) == "Var1"] <- "Date"
  colnames(count_date)[colnames(count_date) == "Freq"] <- "Frequency"

  count_date$red = 0
  for (j in 1:dim(new_cluster)[1]){
    if(new_cluster$Outbreak[j] == 1){
      number = which(count_date$Date == noquote(new_cluster$Create_date_YMW[j]))
      count_date$red[number] = 1
      }
    }

  # print(table(count_date$red))

  cluster_name = count_SNP_20[i,1]

  print(ggplot(data = count_date) +
          geom_point (mapping = aes (x=Date, y=Frequency), color=ifelse(count_date$red == 1, "red", "bla
          scale_x_discrete(breaks = count_date$Date[seq(1, length(count_date$Date), by = 20)]) +
          ggtitle(paste("Listeria Monocytogenes Cases Evolution for SNP Cluster", cluster_name)) +
          labs(x = 'Date(year-month-week)', y = 'Frequency'))


  count_date = count_date[order(-count_date$Frequency),]  # order returns indexes
  print(sprintf("SNP Cluster %s has the highest cases of listeria monocytogenes at %s", cluster_name, c
}
```