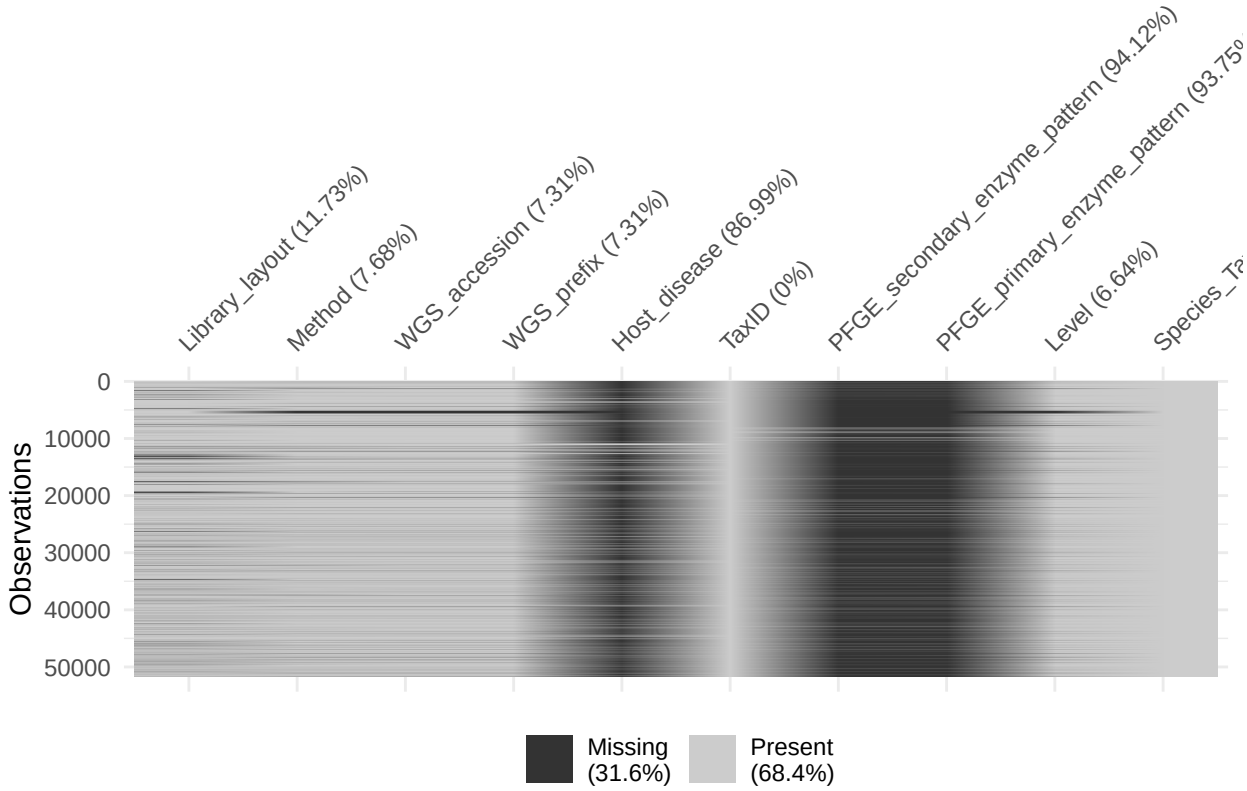
Data Cleaning and EDA

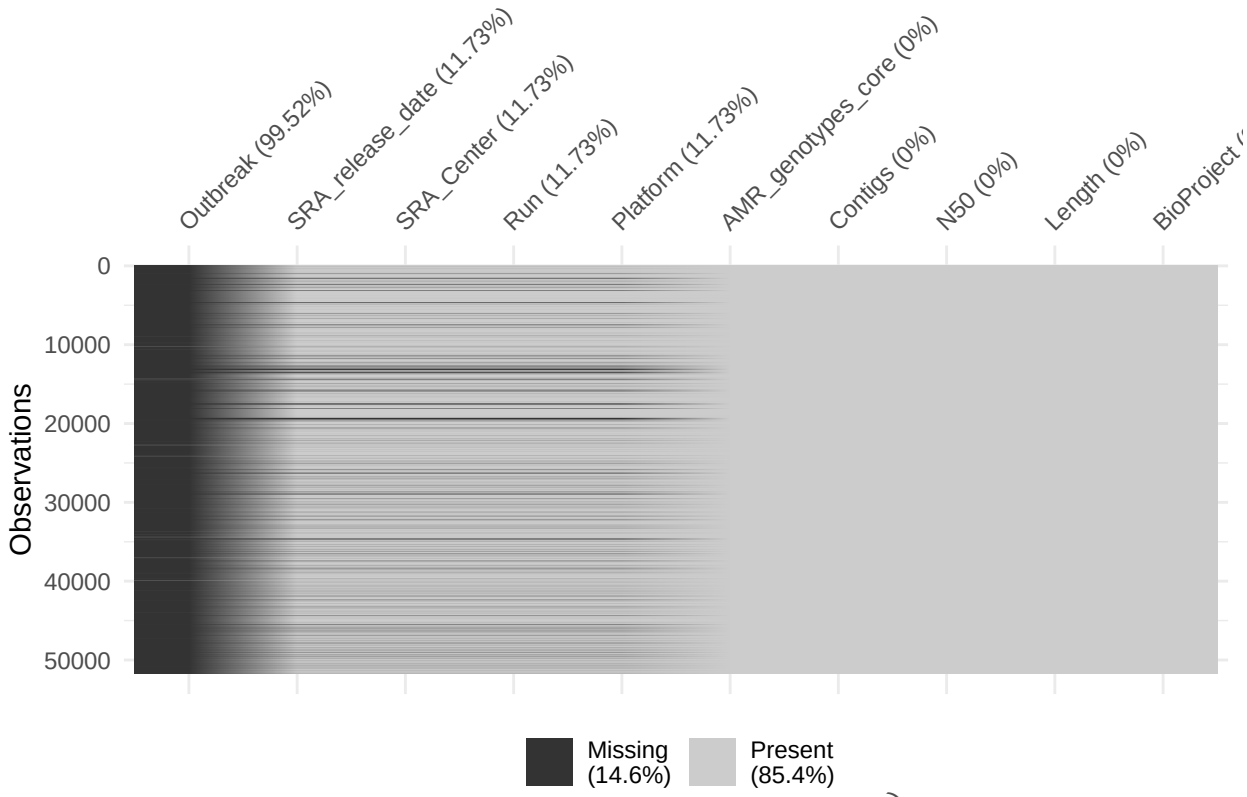
```
## [1] 51738    50
```

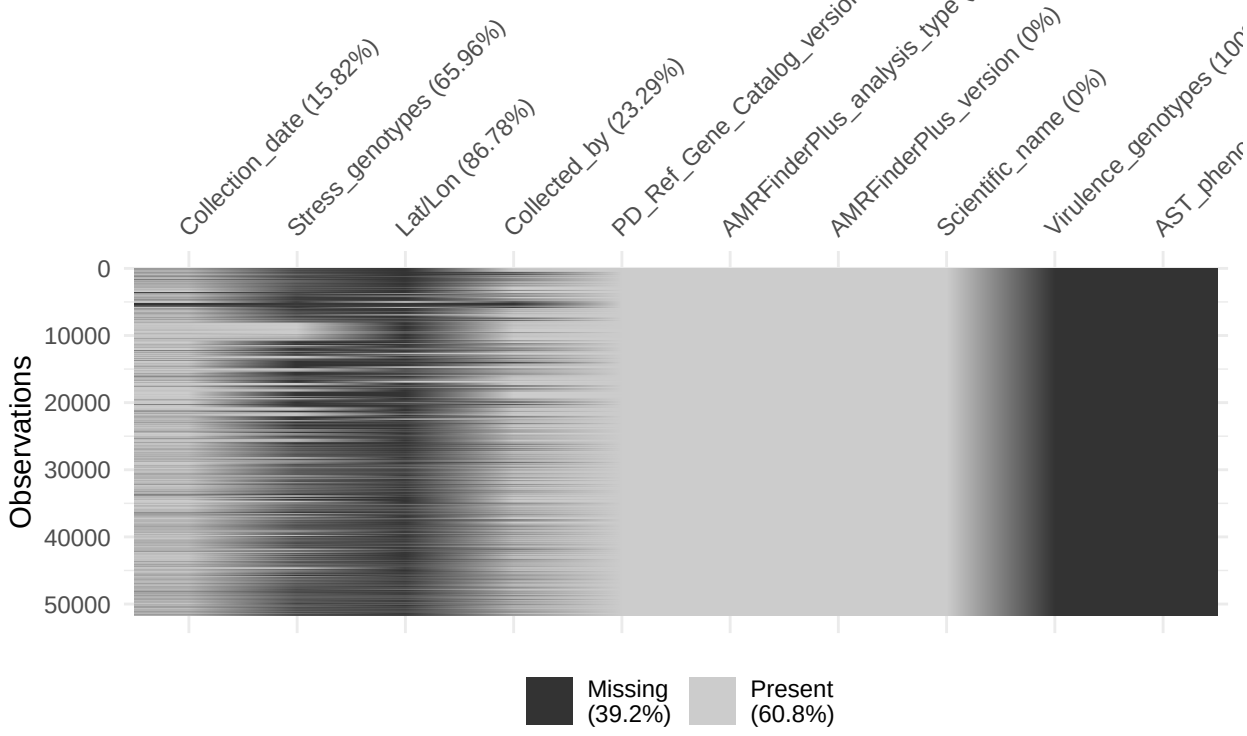
The original dataset without any cleaning has 51738 observations and 50 columns.

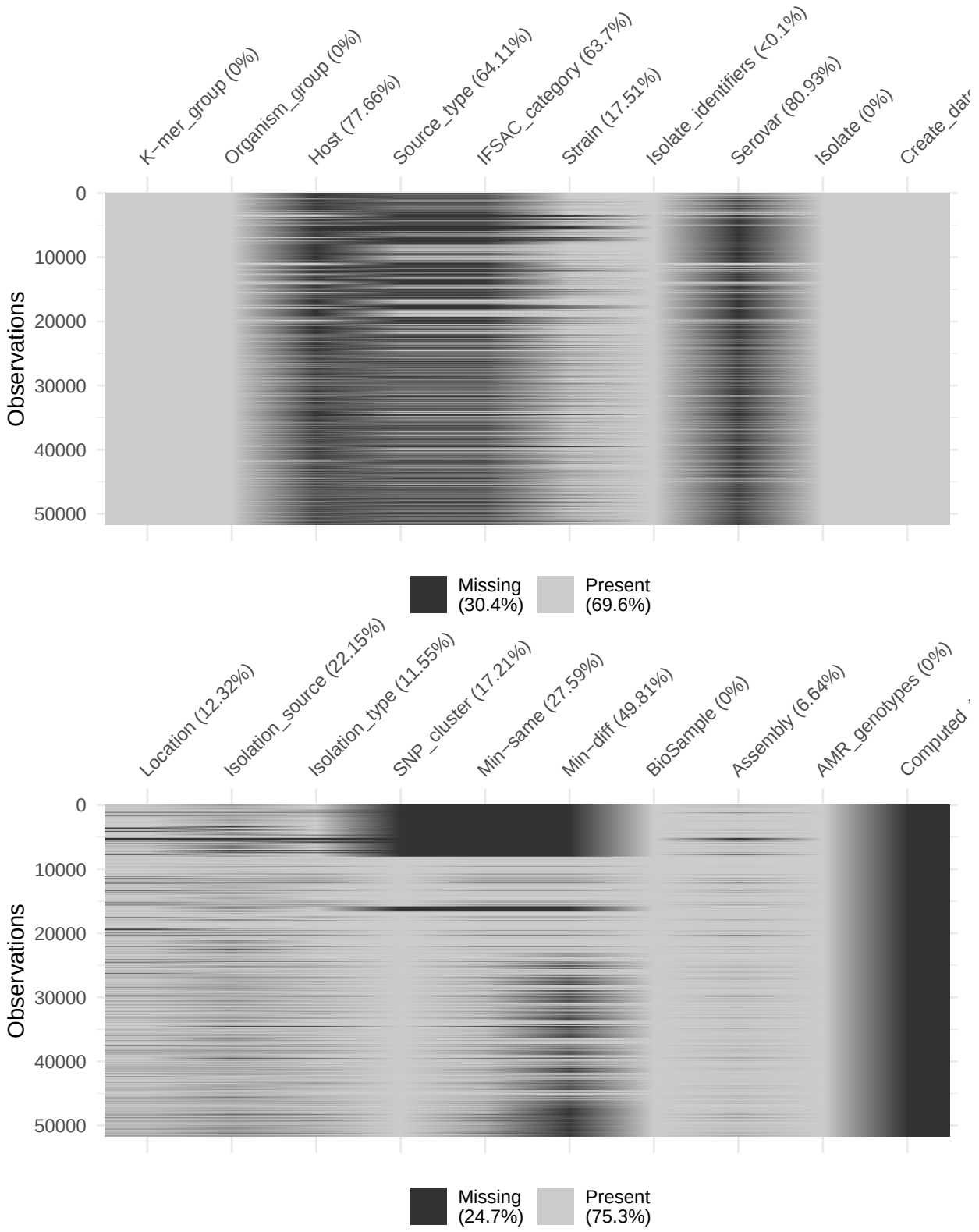
```
##                Library_layout                      Method
##                   0.882658781                 0.923151262
##                WGS_accession                   WGS_prefix
##                   0.926939580                 0.926939580
##                Host_disease                        TaxID
##                   0.130117129                 1.000000000
## PFGE_secondary_enzyme_pattern PFGE_primary_enzyme_pattern
##                   0.058834899                 0.062507248
##                         Level               Species_TaxID
##                   0.933627121                 1.000000000
##                      Outbreak           SRA_release_date
##                   0.004774054                 0.882658781
##                    SRA_Center                         Run
##                   0.882658781                 0.882658781
##                      Platform          AMR_genotypes_core
##                   0.882658781                 1.000000000
##                       Contigs                         N50
##                   1.000000000                 1.000000000
##                        Length                  BioProject
##                   1.000000000                 1.000000000
##               Collection_date            Stress_genotypes
##                   0.841818393                 0.340388109
##                       Lat/Lon               Collected_by
##                   0.132243225                 0.767115080
##   PD_Ref_Gene_Catalog_version AMRFinderPlus_analysis_type
##                   1.000000000                 1.000000000
##         AMRFinderPlus_version             Scientific_name
##                   1.000000000                 1.000000000
##           Virulence_genotypes              AST_phenotypes
##                   0.000000000                 0.000000000
##                   K-mer_group             Organism_group
##                   1.000000000                 1.000000000
##                          Host                 Source_type
##                   0.223414125                 0.358885152
##                IFSAC_category                      Strain
##                   0.362982721                 0.824867602
##           Isolate_identifiers                     Serovar
##                   0.999845375                 0.190691561
##                       Isolate                 Create_date
##                   1.000000000                 1.000000000
##                      Location           Isolation_source
##                   0.876783022                 0.778480034
##                Isolation_type                 SNP_cluster
##                   0.884475627                 0.827882794
```

```
##                  Min-same                  Min-diff
##               0.724071282               0.501913487
##                 BioSample                  Assembly
##               1.000000000               0.933627121
##             AMR_genotypes             Computed_types
##               1.000000000               0.000000000
```




Outbreak (99.52%)
SRA_release_date (11.73%)
SRA_Center (11.73%)
Run (11.73%)
Platform (11.73%)
AMR_genotypes_core (0%)
Contigs (0%)
N50 (0%)
Length (0%)
0
10000
20000
30000
40000
50000
Observations
Missing
(14.6%)
Present
(85.4%)


Collection_date (15.82%)
Stress_genotypes (65.96%)
Lat/Lon (86.78%)
Collected_by (23.29%)
PD_Ref_Gene_Catalog_version (0%)
AMRFinderPlus_analysis_type (0%)
AMRFinderPlus_version (0%)
Scientific_name (0%)
0
10000
20000
30000
40000
50000
Observations
Missing
(39.2%)
Present
(60.8%)



Above table and graphs show the missing pattern of the dataset we are working with.

I drop variables Computed_types, Virulence_genotypes, and AST_phenotypes because these columns are completely empty.

I drop variables Host_disease, PFGE_secondary_enzyme_pattern, PFGE_primary_enzyme_pattern, Stress_genotypes, Lat/Lon, Host, Source_type, IFSAC_category, and Serovar because they contain too many missing values.

I drop variables Species_TaxID, K-mer_group, and Organism_group because every entry of these variables is the same value.

I drop variables WGS_accession, WGS_prefix, Run, Isolate, and Assembly because every level of these variables only contain one observation (there are too many levels for these columns).

I drop variables AMRFinderPlus_version, PD_Ref_Gene_Catalog_version, and Level because they are useless information.

I convert all the categorical variables to factors.

I apply as.Date() function to all the date variables.

I remove all the variables above since they are useless information.

```
## [1] "TaxID: 179"
## [1] "Outbreak: 39"
## [1] "SRA_release_date: 2070"
## [1] "SRA_Center: 104"
## [1] "AMR_genotypes_core: 289"
## [1] "Contigs: 462"
## [1] "N50: 36555"
## [1] "Length: 47717"
## [1] "BioProject: 640"
## [1] "Collection_date: 3690"
## [1] "Collected_by: 379"
## [1] "Scientific_name: 179"
## [1] "Strain: 41053"
## [1] "Create_date: 2065"
## [1] "Location: 364"
## [1] "Isolation_source: 3030"
## [1] "Isolation_type: 3"
## [1] "SNP_cluster: 4378"
## [1] "Min-same: 55"
## [1] "Min-diff: 62"
## [1] "AMR_genotypes: 295"
```

Then, we calculate number of levels for all the categorical variable. I remove Strain from the dataset because it has too many levels and doesn't contain much useful information.

```
table(isolates$Outbreak)
```

```
##
##              0405ml-2          0511MLGX6-1c          0602MLGX6-1c
##                     7                     1                     2
##          0603mlGX6-1c          0610MLGX6-2c          0707MLGX6-1c
##                     1                     2                     1
##         0808MAGX6-1mlc          0811MLGX6-1c           0904MLGX6-1
##                    24                     2                     2
##           0908MLGX6-1           0909MAGX6-2           0909MLGX6-1
##                     6                     2                     4
##           0910MLGX6-3           0910NCGX6-1           0911NJGX6-1
##                     5                     1                    18
##           1005NYGX6-1     1006TXGX6-1 Celery          1008NYCGX6-1
##                    21                     3                    10
```

```
##            1109COGX6-1 1109COGX6-1 Cantaloupe            1110MLGX6-2
##                      4                      24                     15
##            1207PAGX6-1           1208CA2GX6-1            1301MLGX6-1
##                      8                      26                      1
##        1301MLGX6-1,WGS            1302MLGX6-1            1302MLGX6-2
##                      1                      2                       1
##            1307MNGX6-1            1311MLGX6-1         1408MLGX6-3WGS
##                      1                      4                      34
##         1411MLGX6-1WGS            1507MLGX6-2                Cali85
##                      3                      2                       1
##         Carlisle, 1981        cheese outbreak             EON 189848
##                      1                      1                       4
##  North Carolina (2000)                   NS81
##                      1                      1
```

Above table shows different levels for the column 'Outbreak'

```
isolates$Outbreak = ifelse(is.na(isolates$Outbreak), 0, 1)
# if the value is NA, then coded as 0, otherwise, coded as 1

table(isolates$Outbreak)
```

```
##
##     0     1
## 51491   247
```

For the variable 'Outbreak', if the value is missing, I coded it as 0, otherwise, I coded it as 1.

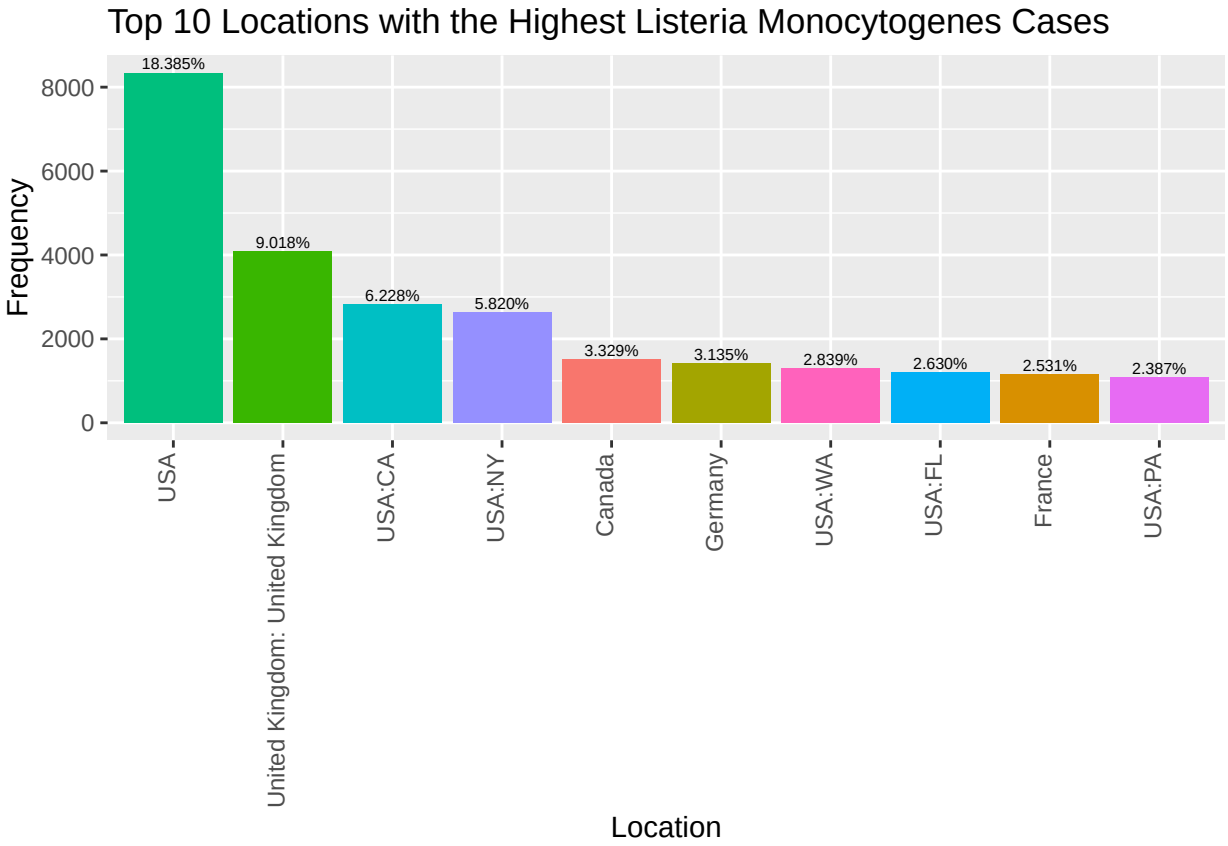
```
##      TaxID          Outbreak        SRA_release_date        SRA_Center
##  1639   :51354   Min.   :0.000000   Min.   :2013-07-30   CFSAN   :15415
##  1906951:   81   1st Qu.:0.000000   1st Qu.:2017-03-27   PULSENET: 4961
##  2291966:   70   Median :0.000000   Median :2019-03-08   EDLB-CDC: 4555
##  2065118:   46   Mean   :0.004774   Mean   :2018-12-13   PHE     : 4537
##  2049008:    4   3rd Qu.:0.000000   3rd Qu.:2020-10-05   ANSES   : 1699
##  882095 :    3   Max.   :1.000000   Max.   :2022-06-24   (Other) :14500
##  (Other):  180                      NA's   :6071         NA's    : 6071
##                                                  AMR_genotypes_core    Contigs
##  fosX=COMPLETE,lin=COMPLETE                                 :37965    Min.   :  1.00
##  abc-f=HMM,fosX=COMPLETE,lin=COMPLETE                       :11626    1st Qu.: 19.00
##  abc-f=HMM,fosX=COMPLETE,lin=COMPLETE,tet(M)=COMPLETE:  639    Median : 26.00
##  fosX=COMPLETE,lin=COMPLETE,tet(M)=COMPLETE                 :  198    Mean   : 42.09
##  abc-f=HMM,fosX=COMPLETE,lin=MISTRANSLATION                 :   82    3rd Qu.: 43.00
##  fosX=COMPLETE                                              :   75    Max.   :935.00
##  (Other)                                                    : 1153
##       N50              Length              BioProject    Collection_date
##  Min.   :   6252   Min.   :2512999   PRJNA215355:16303   2014   : 1483
##  1st Qu.: 201391   1st Qu.:2951914   PRJNA212117: 7780   2016   : 1420
##  Median : 357244   Median :3015076   PRJNA248549: 4581   2015   : 1401
##  Mean   : 384038   Mean   :3023542   PRJNA435747: 1394   2017   : 1260
##  3rd Qu.: 495457   3rd Qu.:3082562   PRJEB38828 : 1378   2018   : 1256
##  Max.   :3100316   Max.   :3631167   PRJNA514286:  974   (Other):36734
##                                      (Other)    :19328   NA's   : 8184
##   Collected_by                                Scientific_name
##  FDA    : 9499   Listeria monocytogenes                    :51354
##  CDC    : 5388   Listeria monocytogenes serotype 1/2a:   81
##  PHE    : 4539   Listeria monocytogenes serotype 1/2b:   70
```

```
##  OSF    : 2132   Listeria monocytogenes serotype 4b  :   46
##  CFIA   : 1381   Listeria monocytogenes serotype 1/2c:    4
##  (Other):16750   Listeria monocytogenes ATCC 19117   :    3
##  NA's   :12049   (Other)                             :  180
##   Create_date                                  Location
##  Min.   :2010-05-12   USA                             : 8340
##  1st Qu.:2017-05-11   United Kingdom: United Kingdom: 4091
##  Median :2019-03-08   USA:CA                          : 2825
##  Mean   :2019-01-21   USA:NY                          : 2640
##  3rd Qu.:2020-11-06   Canada                          : 1510
##  Max.   :2022-06-24   (Other)                         :25957
##                       NA's                            : 6375
##            Isolation_source             Isolation_type
##  food              : 4911   clinical           :16628
##  environmental swab: 4671   environmental/other:29133
##  blood             : 2991   NA's               : 5977
##  environmental     : 2350
##  human             : 1912
##  (Other)           :23442
##  NA's              :11461
##           SNP_cluster       Min-same         Min-diff
##  PDS000000366.488: 1726   Min.   : 0.000   Min.   : 0.00
##  PDS000025311.237: 1219   1st Qu.: 0.000   1st Qu.: 6.00
##  PDS000024989.118:  988   Median : 2.000   Median :16.00
##  PDS000024656.169:  830   Mean   : 5.689   Mean   :16.52
##  PDS000024645.140:  752   3rd Qu.: 7.000   3rd Qu.:26.00
##  (Other)         :37318   Max.   :53.000   Max.   :60.00
##  NA's            : 8905   NA's   :14276    NA's   :25770
##                                                AMR_genotypes
##  fosX=COMPLETE,lin=COMPLETE                           :37948
##  abc-f=HMM,fosX=COMPLETE,lin=COMPLETE                 :11564
##  abc-f=HMM,fosX=COMPLETE,lin=COMPLETE,tet(M)=COMPLETE:  639
##  fosX=COMPLETE,lin=COMPLETE,tet(M)=COMPLETE           :  198
##  abc-f=HMM,fosX=COMPLETE,lin=MISTRANSLATION           :   82
##  fosX=COMPLETE                                        :   75
##  (Other)                                              : 1232
```

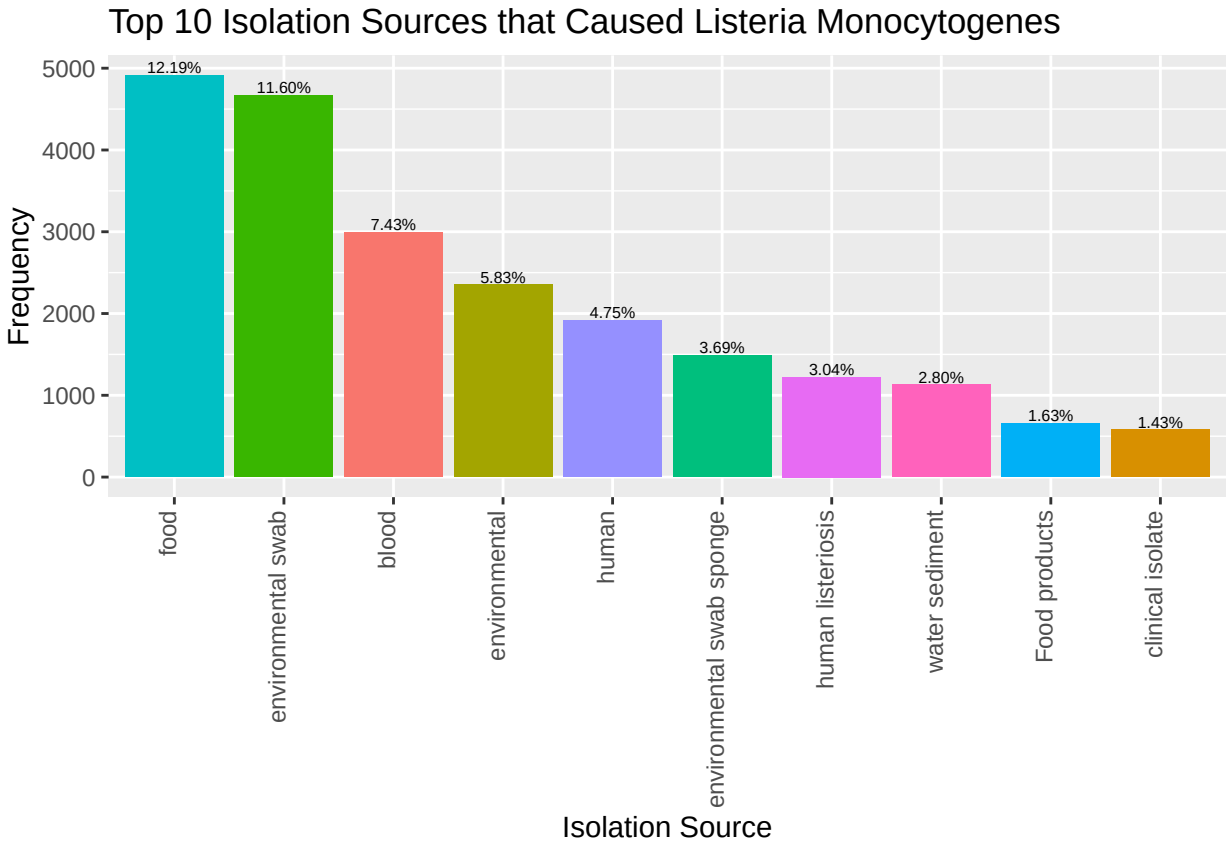
This is the summary table after all the data cleaning procedures.

```
## [1] 51738    20
```

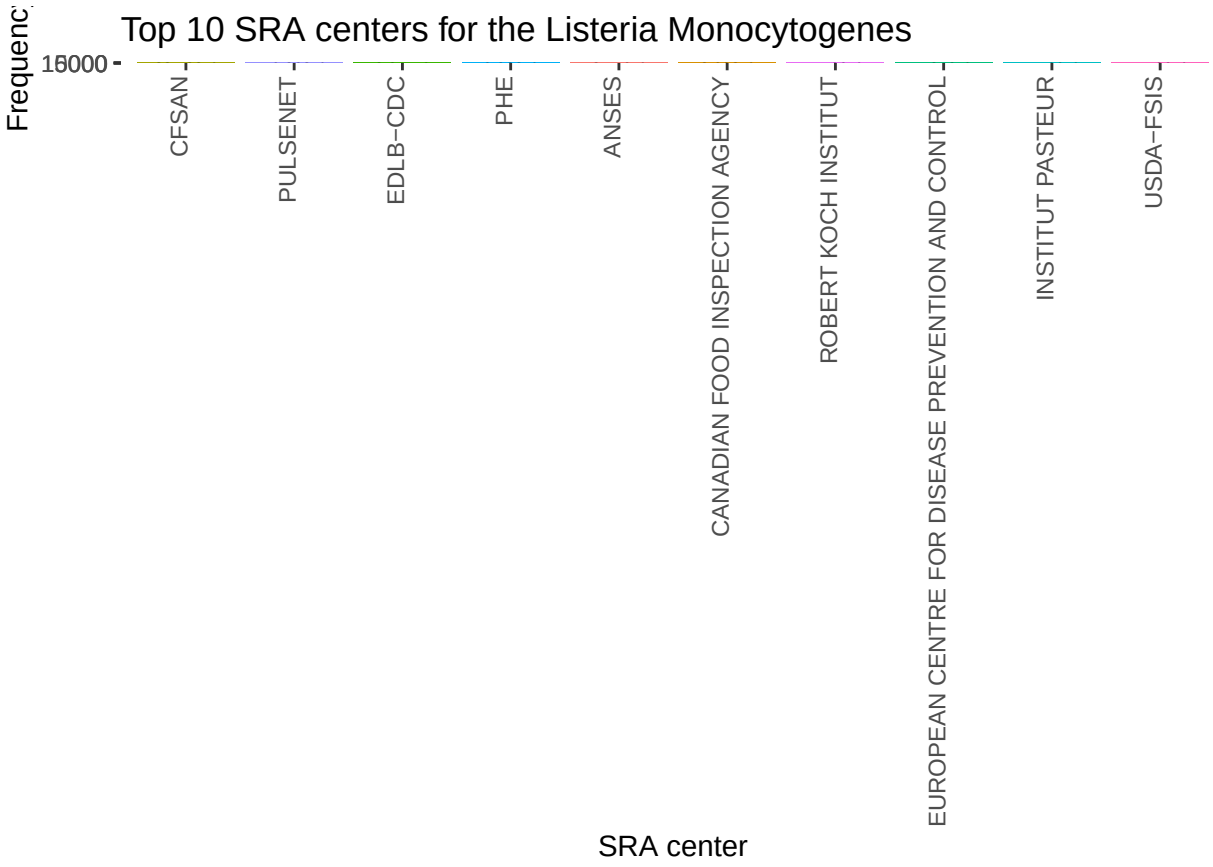
We have 51738 observations and 20 columns in the dataset (we don't conduct complete cases analysis).



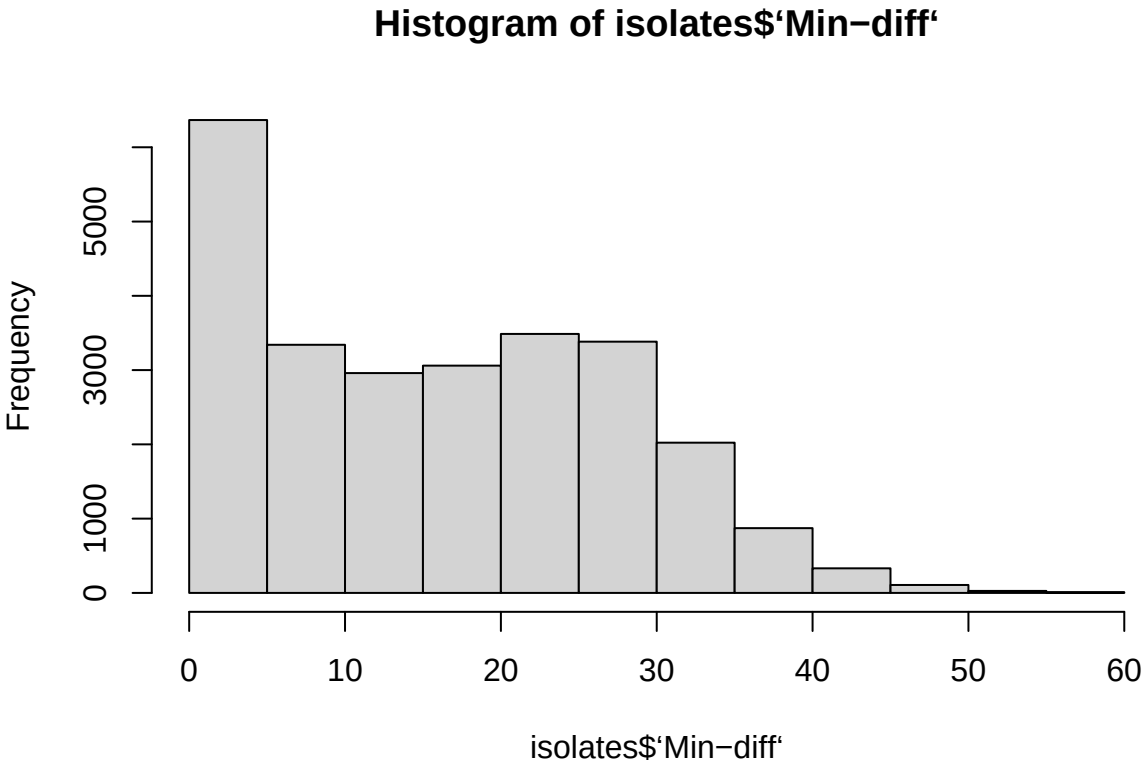
Above graph depicts top 10 location with the highest Listeria Monocytogenes Cases. We can see that USA accounts for almost 19% of all the cases.



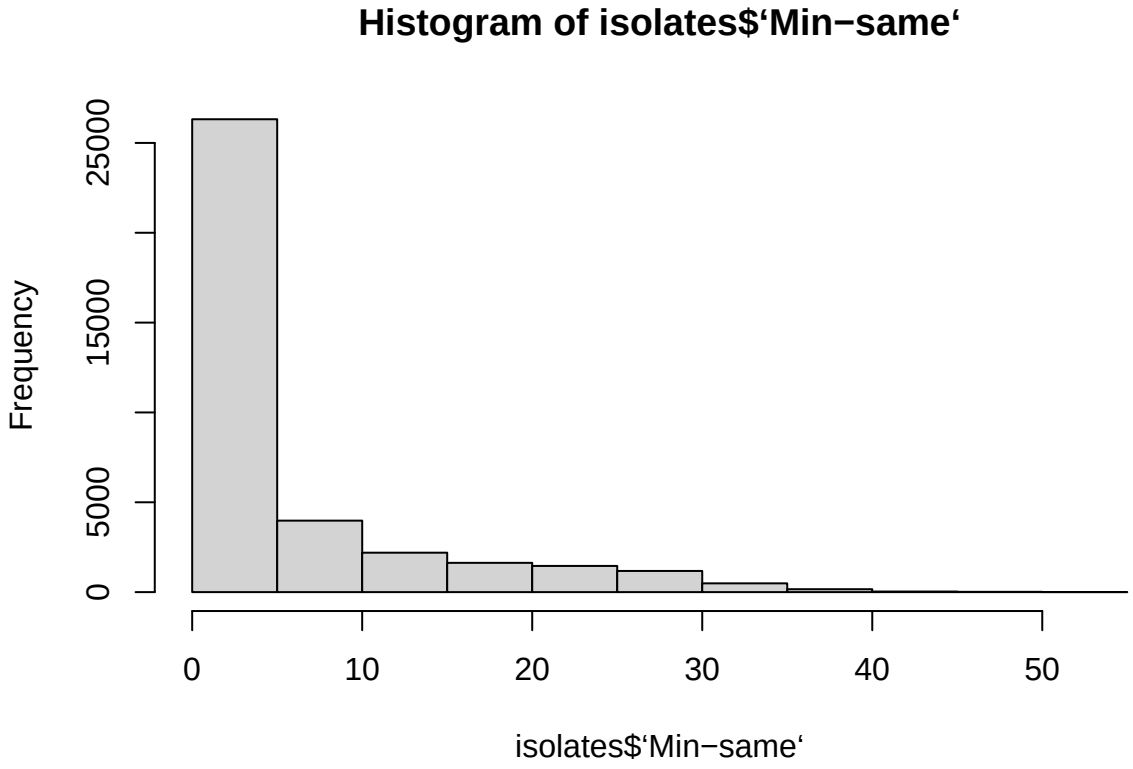
Above graph depicts top 10 isolation sources that caused Listeria Monocytogenes. We can see that both food and environmental swab account for almost 12% of all the cases.



Above graph depicts where the data entry is coming from. We can see that about 34% of observations coming from CFSAN database.



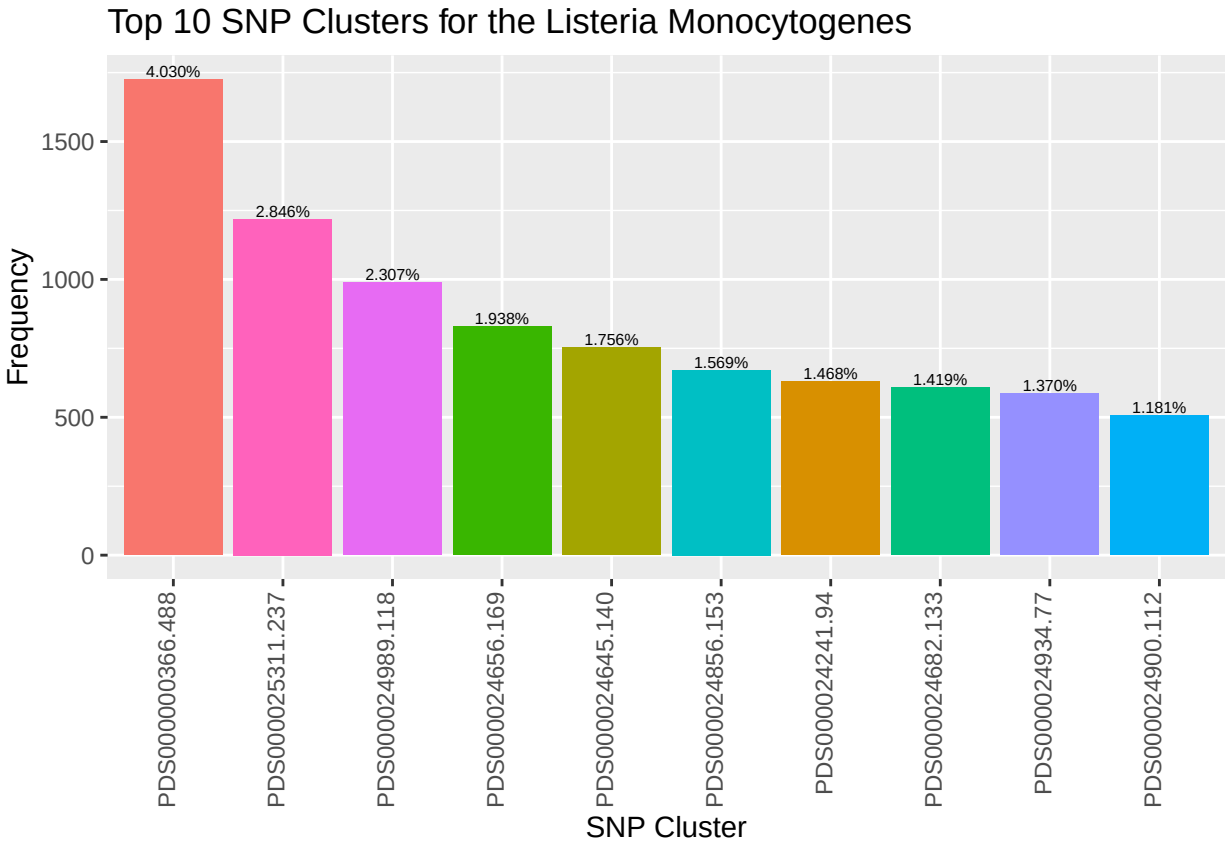
Above graph shows the distribution of variable Min-diff, which means minimum SNP distance from this isolate to one of a different isolation type.



Above graph shows the distribution of variable Min-same, which means Minimum SNP distance from this isolate to one of the same isolation type. Min-same has much lower value compared to Min-diff.

```
## [1] 4378
```

There are total 4378 different SNP clusters in the dataset.



Above graph depicts top 10 SNP clusters for Listeria Monocytogenes cases. We can see that cluster PDS000000366.488 account for almost 4% of all the cases.

```
##           SNP_cluster Frequency SNP_percentage
## 39   PDS000000366.488      1726      4.0296033
## 1246 PDS000025311.237      1219      2.8459365
## 996  PDS000024989.118       988      2.3066327
## 710  PDS000024656.169       830      1.9377583
## 701  PDS000024645.140       752      1.7556557
## 878  PDS000024856.153       672      1.5688838
## 366   PDS000024241.94       629      1.4684939
## 731  PDS000024682.133       608      1.4194663
## 945   PDS000024934.77       587      1.3704387
## 916  PDS000024900.112       506      1.1813322
## 1471 PDS000032941.132       463      1.0809423
## 2734  PDS000058430.33       418      0.9758831
## 423   PDS000024311.15       386      0.9011743
## 59    PDS000003011.70       375      0.8754932
## 1314  PDS000025433.61       334      0.7797726
## 1697  PDS000041947.98       304      0.7097331
## 1206   PDS000025233.7       286      0.6677095
## 703   PDS000024647.63       285      0.6653748
## 209   PDS000005985.16       249      0.5813275
## 1140  PDS000025154.24       237      0.5533117
```

```
## [1] 27.67492
```

Above table shows top 20 clusters with the most observations. We can see that the first 11 clusters all contain

at least 1% of the whole dataset, so that we will further investigate them.

Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library(naniar)
library(readr)
library(dplyr)
library(ggplot2)
library(tableone)
setwd("~/Desktop")
isolates <- read_csv("isolates.csv")
dim(isolates)
apply(isolates, 2,  function(x) sum(complete.cases(x))/nrow(isolates))
isolates1 = isolates[,c(1:10)]
isolates2 = isolates[,c(11:20)]
isolates3 = isolates[,c(21:30)]
isolates4 = isolates[,c(31:40)]
isolates5 = isolates[,c(41:50)]
vis_miss(isolates1)
vis_miss(isolates2)
vis_miss(isolates3)
vis_miss(isolates4)
vis_miss(isolates5)
isolates = isolates %>%
  select(-c(Computed_types, Virulence_genotypes, AST_phenotypes))
isolates = isolates %>%
  select(-c(Host_disease, PFGE_secondary_enzyme_pattern, PFGE_primary_enzyme_pattern, Stress_genotypes,
isolates = isolates %>%
  select(-c(Species_TaxID, `K-mer_group`, Organism_group))
isolates = isolates %>%
  select(-c(WGS_accession, WGS_prefix, Run, Isolate, Assembly))
isolates = isolates %>%
  select(-c(AMRFinderPlus_version, PD_Ref_Gene_Catalog_version, Level))
isolates <- isolates %>%
    mutate(across(.cols=c(Library_layout, Method, SRA_Center, Platform, AMR_genotypes_core, BioProject,
isolates <- isolates %>%
    mutate(across(.cols=c(SRA_release_date, Create_date), .fns = as.Date))
isolates = isolates %>%
  select(-c(Library_layout, Method, Platform, AMRFinderPlus_analysis_type, Isolate_identifiers, BioSampl
for (i in 1:ncol(isolates)){
  number = nrow(unique(isolates[,i]))
  print(sprintf("%s: %s", colnames(isolates)[i], number))
}
isolates = isolates %>%
  select(-Strain)
table(isolates$Outbreak)
isolates$Outbreak = ifelse(is.na(isolates$Outbreak), 0, 1)
# if the value is NA, then coded as 0, otherwise, coded as 1

table(isolates$Outbreak)
summary(isolates)
dim(isolates)
```

```
count_location = as.data.frame(table(isolates$Location))
colnames(count_location)[colnames(count_location) == "Var1"] <- "Location"
colnames(count_location)[colnames(count_location) == "Freq"] <- "Frequency"
count_location =count_location[order(-count_location$Frequency),]
# order by descending
# order() returns indices
count_location_10 = count_location[1:10,]
location_percentage = numeric(10)
for (i in 1:10){
  location_percentage[i] = count_location$Frequency[i]/sum(count_location$Frequency)
}
count_location_10['location_percentage'] <- location_percentage
ggplot(data = count_location_10, aes(x = reorder(Location, -Frequency),
                                     y = Frequency,
                                     label = scales::percent(location_percentage),
                                     fill = Location)) +
  geom_bar(stat = 'identity') +
  ggtitle('Top 10 Locations with the Highest Listeria Monocytogenes Cases') +
  geom_text(vjust = -0.3,
            size = 2) +
  labs(x = 'Location', y = 'Frequency') +
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.1)) +
  theme(legend.position="none")
count_source = as.data.frame(table(isolates$Isolation_source))
colnames(count_source)[colnames(count_source) == "Var1"] <- "Source"
colnames(count_source)[colnames(count_source) == "Freq"] <- "Frequency"
count_source =count_source[order(-count_source$Frequency),]

count_source_10 = count_source[1:10,]
source_percentage = numeric(10)
for (i in 1:10){
  source_percentage[i] = count_source$Frequency[i]/sum(count_source$Frequency)
}
count_source_10['source_percentage'] <- source_percentage
ggplot(data = count_source_10, aes(x = reorder(Source, -Frequency),
                                     y = Frequency,
                                     label = scales::percent(source_percentage),
                                     fill = Source)) +
  geom_bar(stat = 'identity') +
  ggtitle('Top 10 Isolation Sources that Caused Listeria Monocytogenes') +
  geom_text(vjust = -0.2,
            size = 2) +
  labs(x = 'Isolation Source', y = 'Frequency') +
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.1)) +
  theme(legend.position="none")
SRA_center = as.data.frame(table(isolates$SRA_Center))
colnames(SRA_center)[colnames(SRA_center) == "Var1"] <- "SRA_center"
colnames(SRA_center)[colnames(SRA_center) == "Freq"] <- "Frequency"
SRA_center = SRA_center[order(-SRA_center$Frequency),]

count_SRA_center_10 = SRA_center[1:10,]
SRA_center_percentage = numeric(10)
for (i in 1:10){
```

```
  SRA_center_percentage[i] = SRA_center$Frequency[i]/sum(SRA_center$Frequency)
}
count_SRA_center_10['SRA_center_percentage'] <- SRA_center_percentage
ggplot(data = count_SRA_center_10, aes(x = reorder(SRA_center, -Frequency),
                                       y = Frequency,
                                       label = scales::percent(SRA_center_percentage),
                                       fill = SRA_center)) +
  geom_bar(stat = 'identity') +
  ggtitle('Top 10 SRA centers for the Listeria Monocytogenes') +
  geom_text(vjust = -0.1,
            size = 2) +
  labs(x = 'SRA center', y = 'Frequency') +
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.1)) +
  theme(legend.position="none")
hist(isolates$`Min-diff`)
hist(isolates$`Min-same`)
unique_cluster = unique(isolates$SNP_cluster)
length(unique_cluster)
count_SNP = as.data.frame(table(isolates$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_10 = count_SNP[1:10,]
SNP_percentage = numeric(10)
for (i in 1:10){
  SNP_percentage[i] = count_SNP$Frequency[i]/sum(count_SNP$Frequency)
}
count_SNP_10['SNP_percentage'] <- SNP_percentage
ggplot(data = count_SNP_10, aes(x = reorder(SNP_cluster, -Frequency),
                                       y = Frequency,
                                       label = scales::percent(SNP_percentage),
                                       fill = SNP_cluster)) +
  geom_bar(stat = 'identity') +
  ggtitle('Top 10 SNP Clusters for the Listeria Monocytogenes') +
  geom_text(vjust = -0.2,
            size = 2) +
  labs(x = 'SNP Cluster', y = 'Frequency') +
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=0.1)) +
  theme(legend.position="none")
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
count_SNP_20
sum(count_SNP_20$SNP_percentage)
```