Interview Data Analysis

36322311

2024-09-17

```
library(XML)

#Import data

gono_gender <- read.csv("C:/Users/aneke/Downloads/Surveillance practice/gonoreah diagnosis and rates by

Books_xml <- xmlParse("C:/Users/aneke/Downloads/Surveillance practice/15mb.xml")

#convert xml to dataframe

Books_df <- xmlToDataFrame(nodes = getNodeSet(Books_xml, "// record"))

#Print dataframe

#import data

Books <- read.csv("C:/Users/aneke/Downloads/Surveillance practice/Books.csv")
```

```
# import data

Social_contact <- read.csv("C:/Users/aneke/Downloads/Surveillance practice/Social Contact.csv")

Census <- read.csv("C:/Users/aneke/Downloads/Surveillance practice/Census.csv")
```

```
# Create a table

My_table <- data.frame(
                       ID= c( "mary", "Dan", "John"),
                       Names= c(" school", "church","market"),
                       Gender= c("female", "female", "male"))
# create table
print(My_table)
```

```
##     ID   Names Gender
## 1 mary  school female
## 2  Dan  church female
## 3 John  market   male
```

```
#create a contigency table

Data <- matrix(c(114,115,225, 50,200,250,190,315), nrow=3,byrow=TRUE)
```

```
## Warning in matrix(c(114, 115, 225, 50, 200, 250, 190, 315), nrow = 3, byrow =
## TRUE): data length [8] is not a sub-multiple or multiple of the number of rows
```

[3]

```
#add row and column names

colnames(Data) <- c("diseased","Not diseased","Total")
rownames(Data) <- c("exposed", "Not exposed","Total")



contigency_table <- as.table(Data)

print(contigency_table)
```

```
##             diseased Not diseased Total
## exposed          114          115   225
## Not exposed       50          200   250
## Total            190          315   114
```

```
CT<- matrix(c(114,115,50,200), nrow=2, byrow=TRUE)
col_total <- colSums(CT)
row_total <- rowSums(CT)

grand_total <- sum(CT)

CT_T <- rbind(cbind(CT,row_total),c(col_total,grand_total))


colnames(CT_T) <- c("diseased","Not diseased","Total")
rownames(CT_T) <- c("exposed", "Not exposed","Total")

print(CT_T)
```

```
##             diseased Not diseased Total
## exposed          114          115   229
## Not exposed       50          200   250
## Total            164          315   479
```

#There were 172 patient who were not diagnosed with food poisoning (controls) and 112 patients diagnosed with food poisoning (cases) enrolled into the study. 108 controls had eaten out at a restaurant, while 85 cases had eaten out at a restaurant within the past 2 days (defined as the day of presentation at hospital or the day before).

```
# Create a matrix with the data
Data <- matrix(c(85, 27,    # Food poisoning cases (Ate, Did Not Eat)
                 108, 64),  # No food poisoning (Controls) (Ate, Did Not Eat)
               nrow = 2, byrow = TRUE)

# Add row and column names
rownames(Data) <- c("Food Poisoning (Cases)", "No Food Poisoning (Controls)")
colnames(Data) <- c("Ate at Restaurant", "Did Not Eat at Restaurant")

# Convert to a contingency table
contingency_table <- as.table(Data)

# View the table
print(contingency_table)
```

```
##                              Ate at Restaurant Did Not Eat at Restaurant
## Food Poisoning (Cases)                      85                        27
## No Food Poisoning (Controls)               108                        64
```

```r
# Optional: Add row and column totals
row_totals <- rowSums(Data)
col_totals <- colSums(Data)
Data_with_totals <- rbind(cbind(Data, row_totals), c(col_totals, sum(Data)))

# Add names to the table with totals
rownames(Data_with_totals) <- c("Food Poisoning (Cases)", "No Food Poisoning (Controls)", "Total")
colnames(Data_with_totals) <- c("Ate at Restaurant", "Did Not Eat at Restaurant", "Total")

# Convert to a table with totals
contingency_table_with_totals <- as.table(Data_with_totals)

# View the table with totals
print(contingency_table_with_totals)
```

```
##                              Ate at Restaurant Did Not Eat at Restaurant Total
## Food Poisoning (Cases)                      85                        27   112
## No Food Poisoning (Controls)               108                        64   172
## Total                                      193                        91   284
```

#There were 172 patient who were not diagnosed with food poisoning (controls) and 112 patients diagnosed with food poisoning (cases) enrolled into the study. 108 controls had eaten out at a restaurant, while 85 cases had eaten out at a restaurant within the past 2 days (defined as the day of presentation at hospital or the day before).

```r
# contigency table.

total_control <- 172
total_cases <- 112
Exposed_not_diseased <- 108
Exposed_diseased <- 85
Not_exposed_diseased <- 112-85
Not_exposed_not_diseased<- 172-108

Data_2 <- matrix(c(Exposed_diseased,Exposed_not_diseased,
                   Not_exposed_diseased,Not_exposed_not_diseased), nrow=2, byrow = TRUE)
# ROW AND COLUMN NAMES

rownames(Data_2) <- c("Ate_at_Restraunt","Not_eat_Restraunt")
colnames(Data_2) <- c("Food_poisoining", "No_Food_poisonng")

Data_2_total <- rbind(cbind(Data_2,rowSums(Data_2)),c(colSums(Data_2),248))

#add the names of the total column
rownames(Data_2_total)[3] <- "Total"
colnames(Data_2_total)[3] <- "Total"

contigency_tableT<- Data_2_total

contigency_tableT
```

```
##                   Food_poisoining No_Food_poisonng Total
```

```
## Ate_at_Restraunt              85              108   193
## Not_eat_Restraunt             27               64    91
## Total                        112              172   248
```

#Calculate:

#the risk of a patient having food poisoning #the risk of a patient eating out at a restaurant AND having food poisoning #the risk of a patient not eating out at a restaurant AND having food poisoning

```
#the risk of a patient having food poisoning

total_popn<- 248
total_ate_Restraunt <-193
total_not_Restraunt<-91

attack_rate <- total_cases/total_popn



Risk_disease <- Exposed_diseased/total_ate_Restraunt

Risk_Not__exposed_disease <- Not_exposed_diseased /total_not_Restraunt


RR <- Risk_disease/Risk_Not__exposed_disease

Odds_exposed <- Risk_disease/(1-Risk_disease)

Odds_unexposed <-Risk_Not__exposed_disease/(1-Risk_Not__exposed_disease)

OR<- Odds_exposed/Odds_unexposed

attack_rate
```

```
## [1] 0.4516129
```

```
Risk_disease
```

```
## [1] 0.4404145
```

```
Risk_Not__exposed_disease
```

```
## [1] 0.2967033
```

```
RR
```

```
## [1] 1.48436
```

```
OR
```

```
## [1] 1.865569
```

```
#Import Data

scs <- read.csv("C:/Users/aneke/Downloads/Surveillance practice/Social Contact.csv", na.strings = c("NA
```

```
#explore the data structure

str(scs)
```

```
## 'data.frame':    4217 obs. of  12 variables:
```

```
##  $ id                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ postal            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ unmatched_postcode: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ web               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age               : int  51 62 36 27 35 61 41 73 78 54 ...
##  $ date              : chr  "28/05/2009" "28/05/2009" "28/05/2009" "28/05/2009" ...
##  $ day_of_week       : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ postcode          : chr  "KT11 2JF" NA "GU34 2BG" "SW12 9HJ" ...
##  $ sex               : int  0 1 1 0 0 0 0 0 1 0 ...
##  $ household_size    : int  3 5 3 1 4 1 3 1 2 3 ...
##  $ occupation        : chr  "home" "public" "unknown" "office" ...
##  $ total_contacts    : int  106 20 7 13 44 30 16 1 2 27 ...
```

```
head(scs)
```

```
##   id postal unmatched_postcode web age       date day_of_week postcode sex
## 1  1      1                   0   0  51 28/05/2009           4 KT11 2JF   0
## 2  2      1                   0   0  62 28/05/2009           4     <NA>   1
## 3  3      1                   0   0  36 28/05/2009           4 GU34 2BG   1
## 4  4      1                   0   0  27 28/05/2009           4 SW12 9HJ   0
## 5  5      1                   0   0  35 28/05/2009           4  DT3 6JJ   0
## 6  6      1                   0   0  61 28/05/2009           4 BN10 7PX   0
##   household_size occupation total_contacts
## 1              3       home            106
## 2              5     public             20
## 3              3    unknown              7
## 4              1     office             13
## 5              4   research             44
## 6              1    retired             30
```

```
tail(scs)
```

```
##         id postal unmatched_postcode web age             date day_of_week
## 4212 11481      0                   0   1  33 11/09/2010 19:28           4
## 4213 11482      0                   0   1  29 13/09/2010 08:13           0
## 4214 11483      0                   0   1  57 13/09/2010 08:19           0
## 4215 11484      0                   0   1  25 13/09/2010 23:02           2
## 4216 11486      0                   0   1  34 19/09/2010 09:05           6
## 4217 11487      0                   0   1  37 20/09/2010 08:33           0
##      postcode sex household_size occupation total_contacts
## 4212  HR9 7RG   0              3     health             25
## 4213  CV4 8EA   0              1     office              7
## 4214  WR5 2DE   0              2     office              6
## 4215 OX29 8DH   0              2     health             17
## 4216  OX7 3NE   1              4     office             26
## 4217  CB4 2PX   1              3     office             18
```

```
#Transformation of variables

scs$date <- as.POSIXct(scs$date, tryFormats = c("%d/%m/%Y %H:%M", "%d/%m/%Y", "%Y-%m-%d"))

scs$sex <- factor(scs$sex, levels = c(1,0,-1), labels = c("Male", "Female","Unspecify"))

scs$age <- as.numeric(scs$age)

scs$age_group <- cut(scs$age,
```

```
                breaks = c(0,5,10,15,20,25,30,35,40,
                           45,50,55,60,65,70,75,80,85,
                           90, Inf),
                labels = c("0-4","5-9","10-14","15-19","20-24",
                           "25-29","30-34", "35-39","40-44","45-49",
                           "50-54","55-59","60-64","65-69","70-74","75-79",
                           "80-84","85-89","90+"),right = FALSE)
scs$day_of_week <- factor(scs$day_of_week,
                          levels = c(0:6),
                          labels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sun
 scs$household_size[scs$household_size == -1] <- NA

scs$household_size <- cut(scs$household_size,
                          breaks = c(0, 1,4,6,8,Inf),
                          labels = c("1_Person","2-4persons","5-6persons","7-8persons",">8persons"))

scs$occupation <- as.factor(scs$occupation)

scs$postcode <- as.factor(scs$postcode)

scs$postal <- factor(scs$postal, levels = c(1,0), labels = c("yes", "no"))

scs$web <- factor(scs$web, levels = c(1,0), labels = c("yes", "no"))

scs$survey_type <- ifelse(scs$postal== "yes", "postal",
                          ifelse(scs$web == "yes", "web", "unknown"))


scs$survey_type<- as.factor(scs$survey_type)
```

```
# check for duplicates

duplicate_enteries <- duplicated(scs$id)


sum(duplicate_enteries)
```

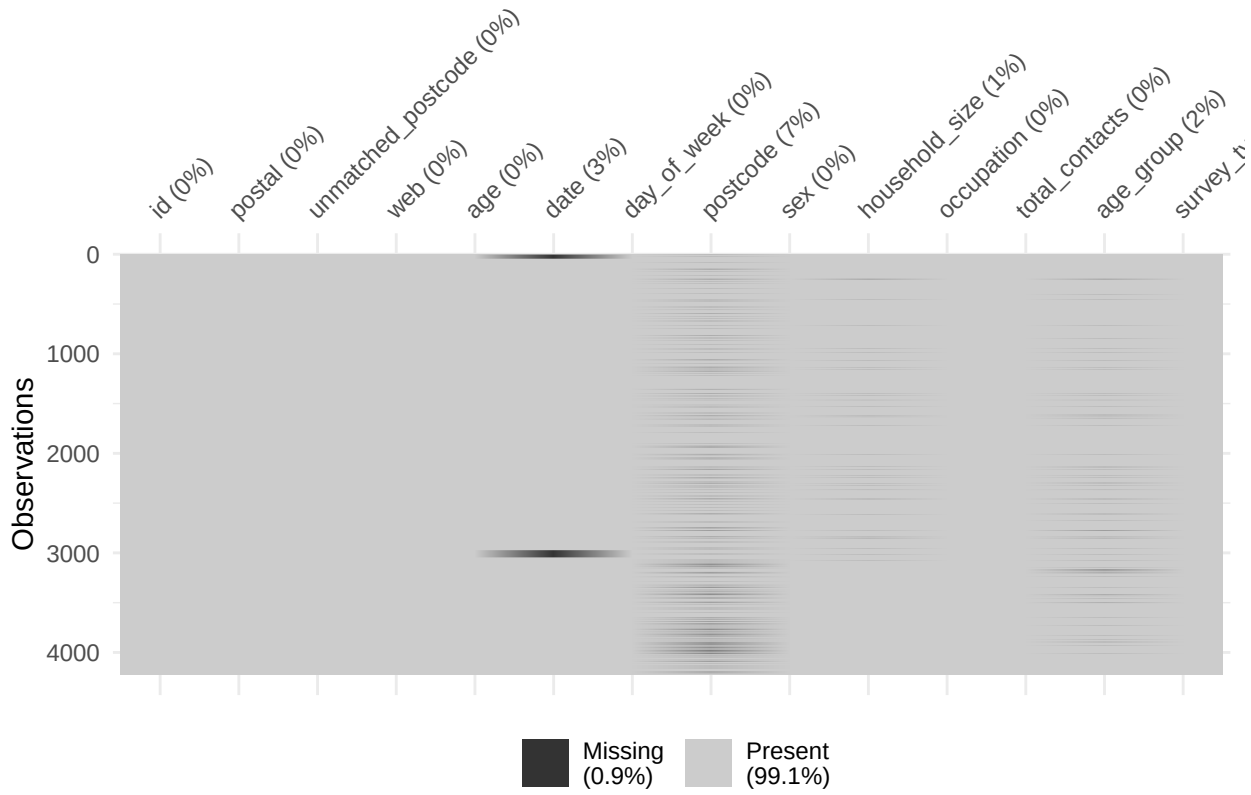
```
## [1] 0
```

```
sum(is.na(scs$date))
```

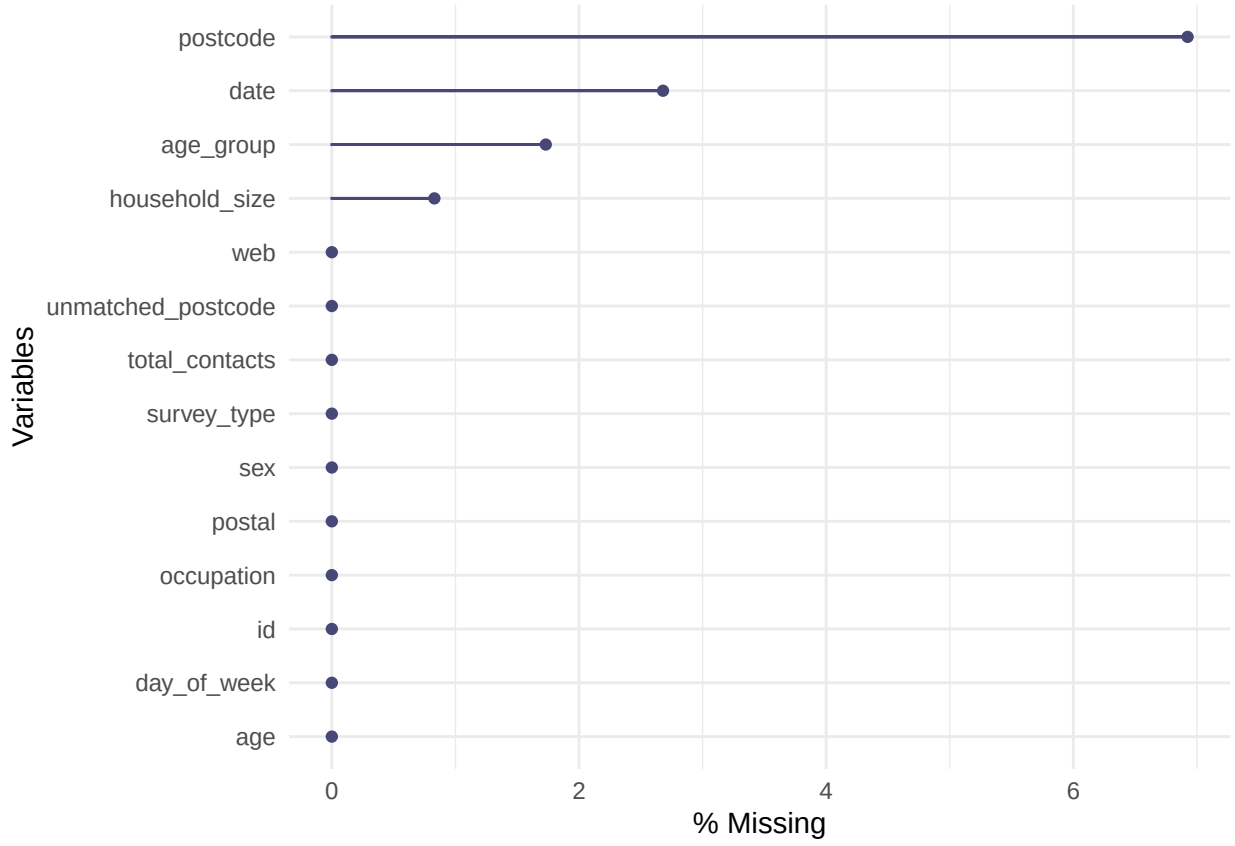
```
## [1] 113
```

#Handling missing values

```
library(naniar)

vis_miss(scs)
```




```
gg_miss_var(scs, show_pct = TRUE)
```




```
n_miss(scs)
```

```
## [1] 513
```

```
n_complete(scs)
```

```
## [1] 58525
```

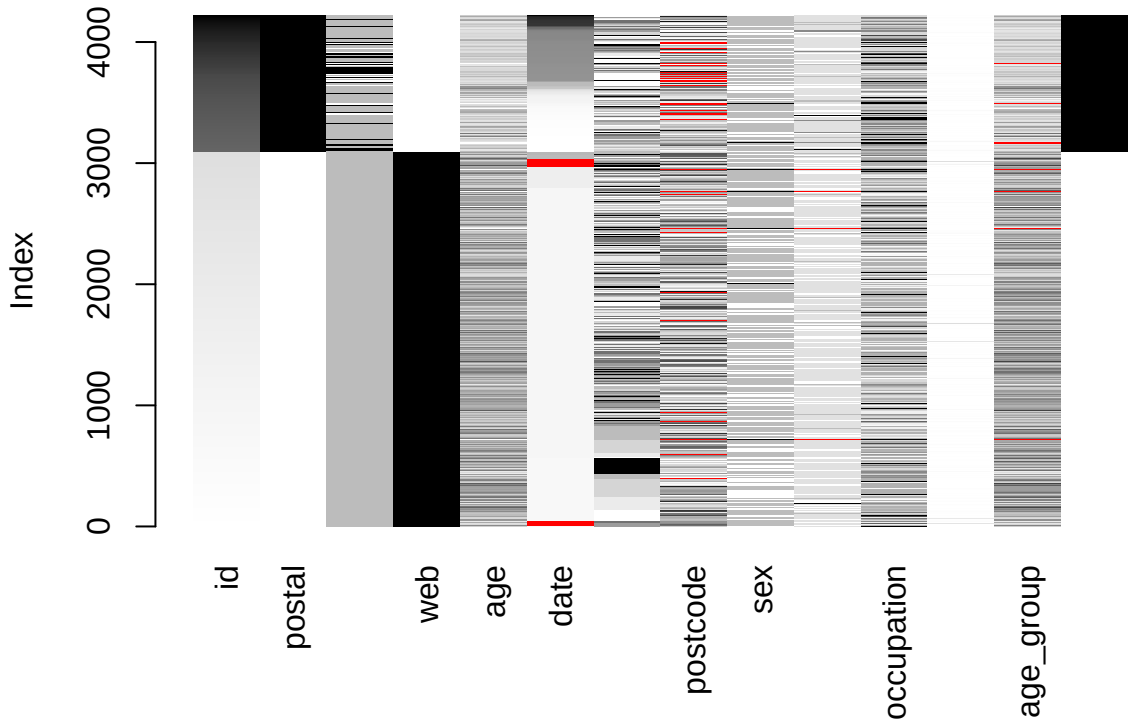
```
miss_var_summary(scs)
```

```
## # A tibble: 14 x 3
##    variable           n_miss pct_miss
##    <chr>               <int>    <dbl>
##  1 postcode              292    6.92
##  2 date                  113    2.68
##  3 age_group              73    1.73
##  4 household_size         35    0.830
##  5 id                      0    0
##  6 postal                  0    0
##  7 unmatched_postcode      0    0
##  8 web                     0    0
##  9 age                     0    0
## 10 day_of_week             0    0
## 11 sex                     0    0
## 12 occupation              0    0
## 13 total_contacts          0    0
## 14 survey_type             0    0
```

```
# pattern of missingness

library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
matrixplot(scs)
```




```
#Descriptive statistics


table(scs$sex)
```

```
##
##      Male    Female Unspecify
##      1393      2757        67
```

```r
summary(scs$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.00   35.00   52.00   49.94   64.00   97.00
```

```r
# age distribution

Age_Dist<- ggplot(scs,aes(x=age))+
  geom_histogram(binwidth = 5, fill= "steelblue", color="white", aes(y= ..density..))+
  geom_density(color="red",size= 1.2)
```

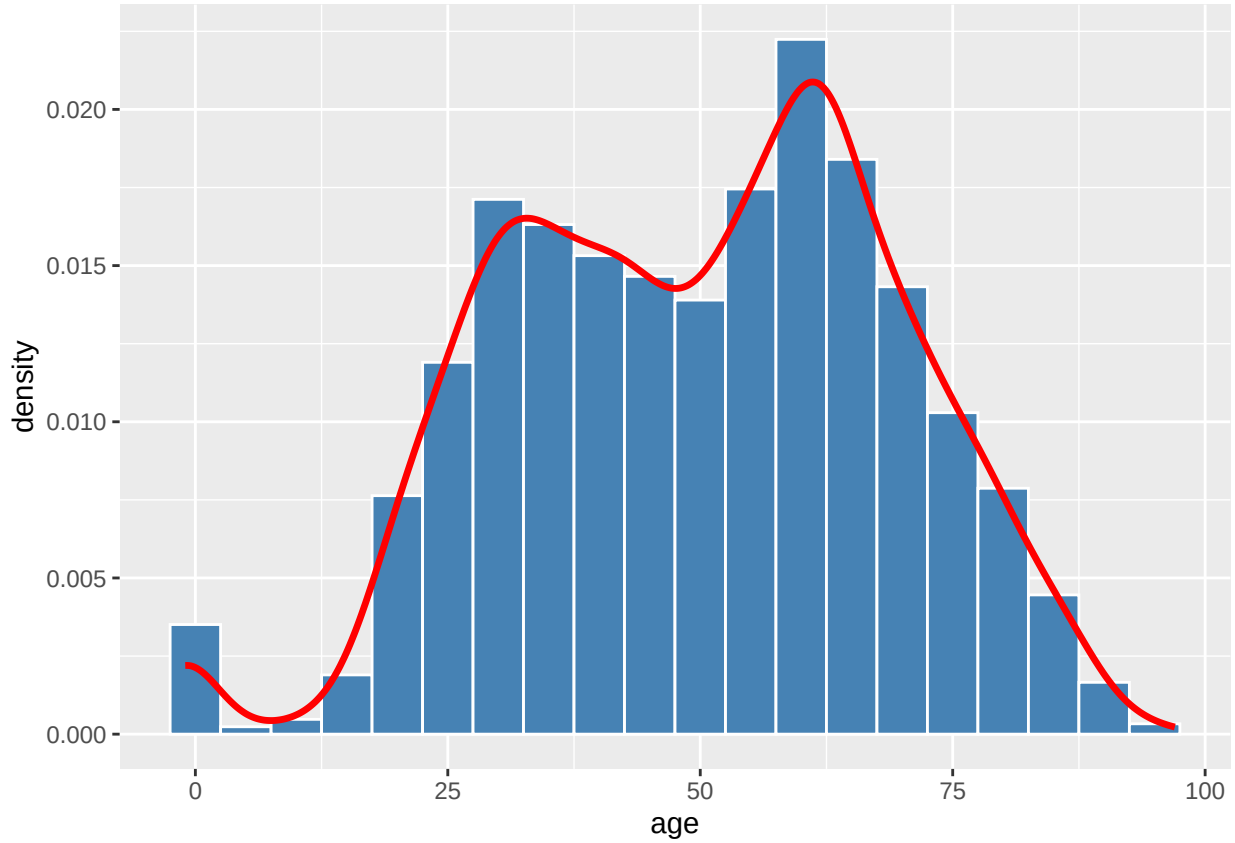
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
  labs(title = "Age Dstribution",
       x= "Age",
       y = "density")
```

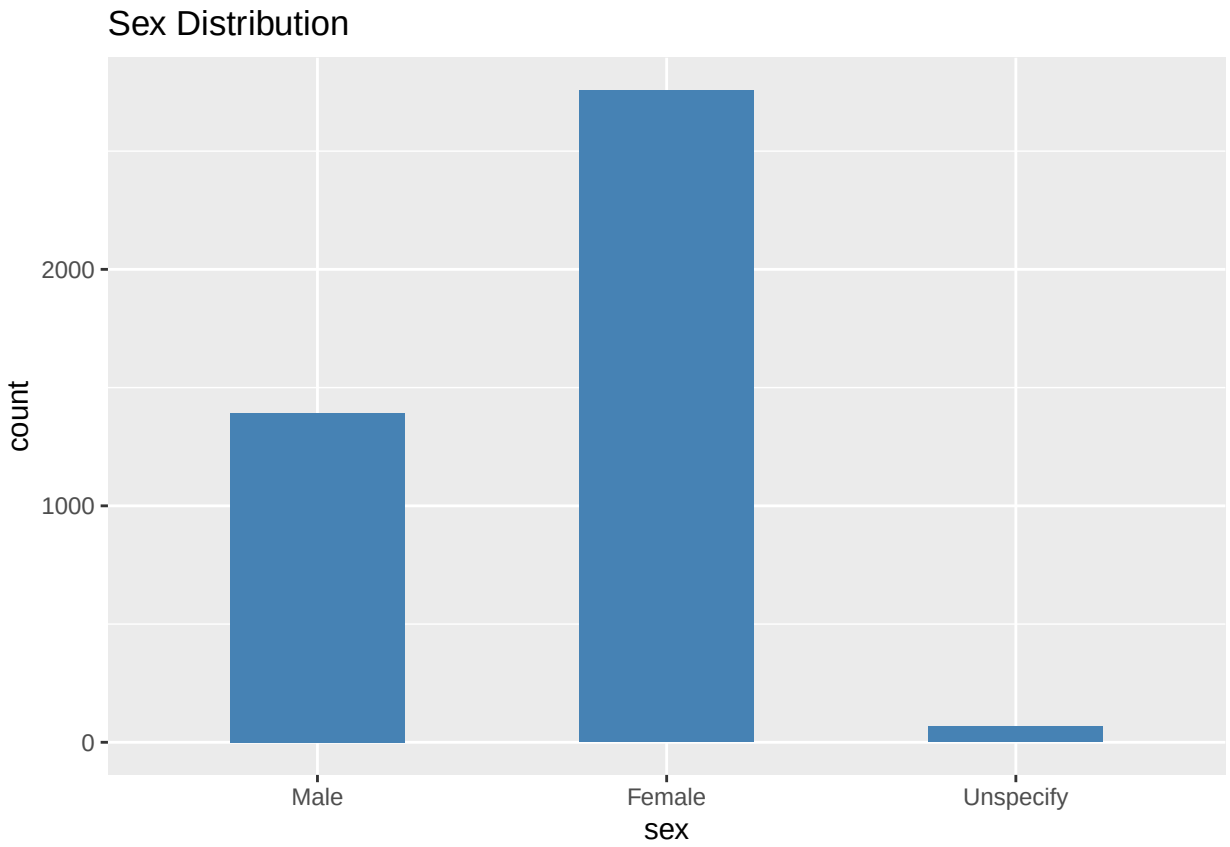
```
## $x
## [1] "Age"
##
## $y
## [1] "density"
##
## $title
## [1] "Age Dstribution"
##
## attr(,"class")
## [1] "labels"
```

```r
# barplot of sex distribution

Gender <- ggplot(scs,aes(x=sex)) +
                  geom_bar(fill= "steelblue", width = 0.5)+
                  labs(title = "Sex Distribution", x= "sex", y="count")

age_group <- ggplot(scs, aes(x= age_group))+
   geom_bar(fill= "steelblue", width = 0.5)+
                  labs(title = "Age_group Distribution", x= "age_group", y="count")

Age_Dist
```

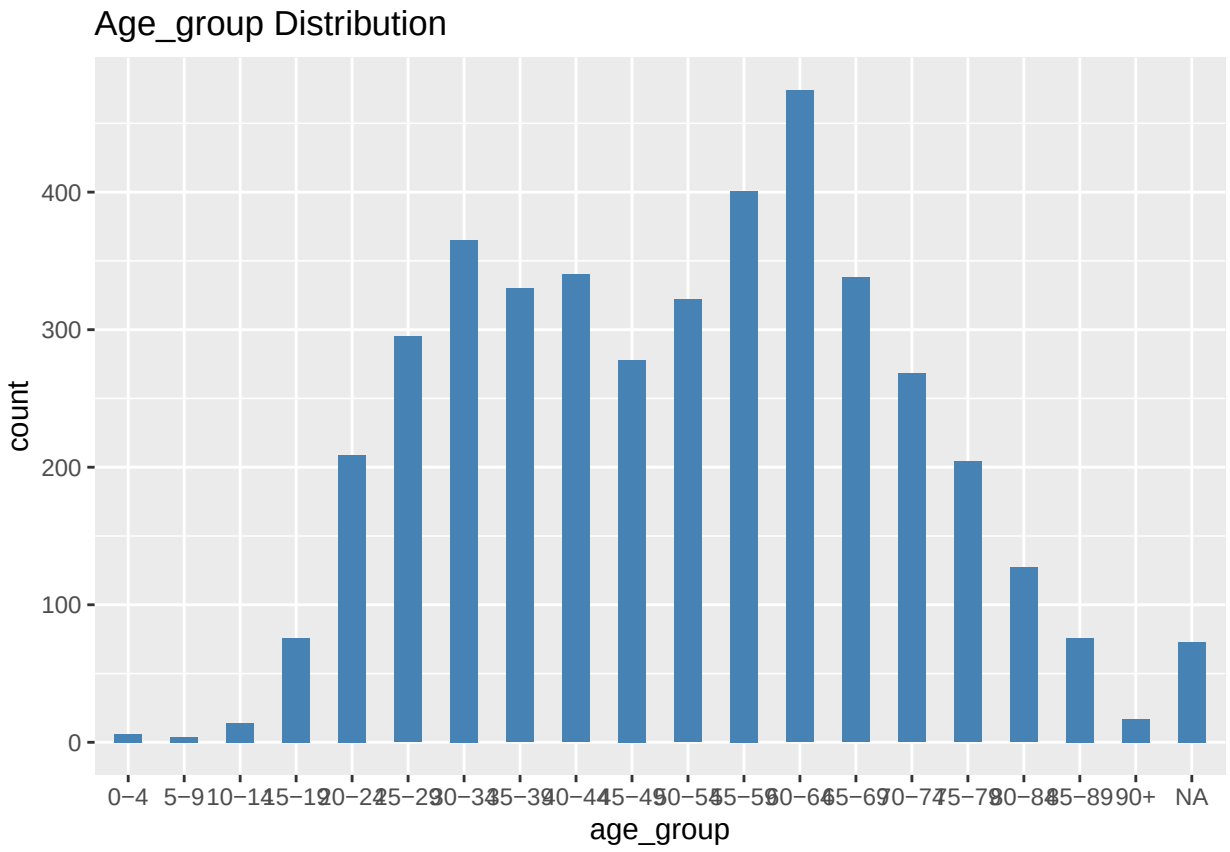
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
Gender
```


Sex Distribution
2000
1000
0
count
Male
Female
Unspecify
sex

```
age_group
```




```
age <- scs %>%
  group_by(age_group, sex) %>%
  summarise(count = n())%>%
  #group_by(age_group) %>%
  mutate(proportion = round((count / sum(count))*100,2))
```

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

```
age
```

```
## # A tibble: 52 x 4
## # Groups:   age_group [20]
##    age_group sex       count proportion
##    <fct>     <fct>     <int>      <dbl>
##  1 0-4       Female        6     100
##  2 5-9       Male          1      25
##  3 5-9       Female        3      75
##  4 10-14     Male          6      42.9
##  5 10-14     Female        8      57.1
##  6 15-19     Male         18      23.7
##  7 15-19     Female       57      75
##  8 15-19     Unspecify     1       1.32
##  9 20-24     Male         67      32.1
## 10 20-24     Female      141      67.5
## # i 42 more rows
```

```
knitr::kable(age, caption = "age group and sex proportion")
```

Table 1: age group and sex proportion

age_group	sex	count	proportion
0-4	Female	6	100.00
5-9	Male	1	25.00
5-9	Female	3	75.00
10-14	Male	6	42.86
10-14	Female	8	57.14
15-19	Male	18	23.68
15-19	Female	57	75.00
15-19	Unspecify	1	1.32
20-24	Male	67	32.06
20-24	Female	141	67.46
20-24	Unspecify	1	0.48
25-29	Male	51	17.29
25-29	Female	243	82.37
25-29	Unspecify	1	0.34
30-34	Male	100	27.40
30-34	Female	265	72.60
35-39	Male	105	31.82
35-39	Female	224	67.88
35-39	Unspecify	1	0.30
40-44	Male	88	25.88
40-44	Female	250	73.53
40-44	Unspecify	2	0.59
45-49	Male	73	26.26
45-49	Female	203	73.02
45-49	Unspecify	2	0.72
50-54	Male	99	30.75
50-54	Female	222	68.94
50-54	Unspecify	1	0.31
55-59	Male	120	29.93
55-59	Female	279	69.58
55-59	Unspecify	2	0.50
60-64	Male	187	39.45
60-64	Female	286	60.34
60-64	Unspecify	1	0.21
65-69	Male	145	42.90
65-69	Female	191	56.51
65-69	Unspecify	2	0.59
70-74	Male	131	48.88
70-74	Female	136	50.75
70-74	Unspecify	1	0.37
75-79	Male	102	50.00
75-79	Female	101	49.51
75-79	Unspecify	1	0.49
80-84	Male	52	40.94
80-84	Female	75	59.06
85-89	Male	40	52.63
85-89	Female	36	47.37
90+	Male	4	23.53

age_group	sex	count	proportion
90+	Female	13	76.47
NA	Male	4	5.48
NA	Female	18	24.66
NA	Unspecify	51	69.86

```
Gender <- scs %>%
  group_by(sex) %>%
  summarise(count = n())%>%
  mutate(proportion = round((count / sum(count))*100, 2))

Gender
```

```
## # A tibble: 3 x 3
##   sex       count proportion
##   <fct>     <int>      <dbl>
## 1 Male       1393      33.0
## 2 Female     2757      65.4
## 3 Unspecify    67       1.59
```

```
knitr::kable(Gender, caption = "Gender proportion")
```

Table 2: Gender proportion

sex	count	proportion
Male	1393	33.03
Female	2757	65.38
Unspecify	67	1.59

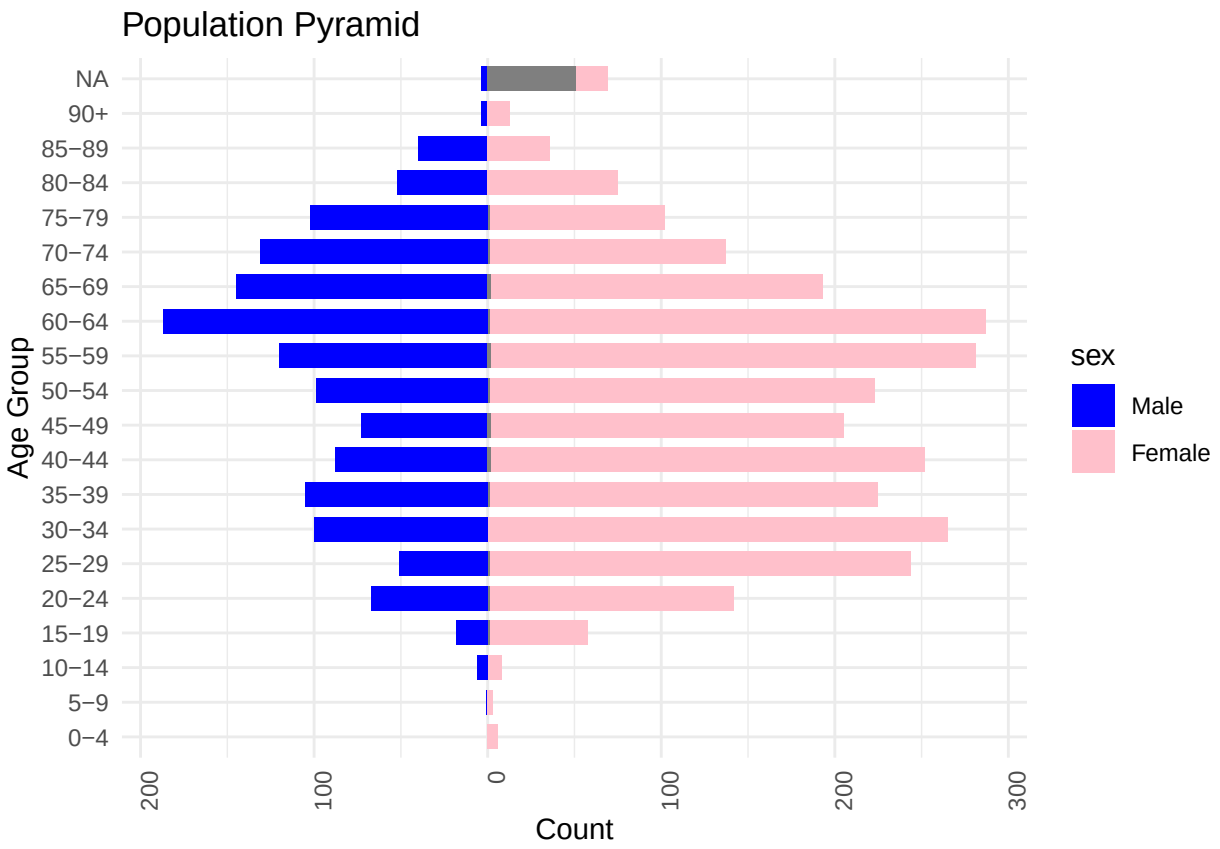
```
#Data representative nature

Census<- read.csv("C:/Users/aneke/Downloads/Surveillance practice/Census.csv")

Census$Age[2]<- "5-9"
```

```
# Create a summary table of counts by age group and sex
age_sex_summary <- scs %>%
  group_by(age_group, sex) %>%
  summarise(count = n())
```

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

```
# Reverse the counts for males to make a pyramid
age_sex_summary <- age_sex_summary %>%
  mutate(count = ifelse(sex == "Male", -count, count))

# Plot the population pyramid
ggplot(age_sex_summary, aes(x = age_group, y = count, fill = sex)) +
  geom_bar(stat = "identity", width = 0.7) +
  coord_flip() +  # Flip coordinates to create a horizontal pyramid
  scale_y_continuous(labels = abs) +  # Show positive values on both sides
  labs(title = "Population Pyramid", x = "Age Group", y = "Count") +
```

```
  theme_minimal() +
  scale_fill_manual(values = c("Male" = "blue", "Female" = "pink")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
# Make sure Males have negative values for the population pyramid
census_data <- Census %>%
  mutate(Males = -Males)

# Assuming `scs` is your dataset with 'age_group' and 'sex' columns
scs$sex <- factor(scs$sex, levels = c(0, 1), labels = c("Male", "Female"))

# Summarize your data by age group and sex
scs_summary <- scs %>%
  group_by(age_group, sex) %>%
  summarise(count = n()) %>%
  ungroup()
```

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

```
# Make Male counts negative for population pyramid
scs_summary <- scs_summary %>%
  mutate(count = ifelse(sex == "Male", -count, count))
```