Chapter 3 Visualizing Data

Angelo LaCommare-Soto

2025-02-05

Section 1 - Importing Data

```
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```
#commands header=TRUE in order to treat the first row of the data frame as a header and stringsAsFactor
# importing the uca data
uca <- read.csv("uca.csv",header=TRUE,stringsAsFactors = TRUE)
#View(uca)
#str(uca)

# importing the microbial genomics data
# note .txt file
micro <- read.table("micro.txt",stringsAsFactors = TRUE)
#View(micro)
#str(micro)
```

Question Answers

a. The dataset "uca.csv" contains 73 total observations of 16 different variables.
b. The dataset "micro.txt" contains 296521 total observations of 15 different variables.

Section 2A - Displaying Data for One Variable Using a Bar Graph

```
# load tidyverse to activate ggplot2 package
library(tidyverse)
```

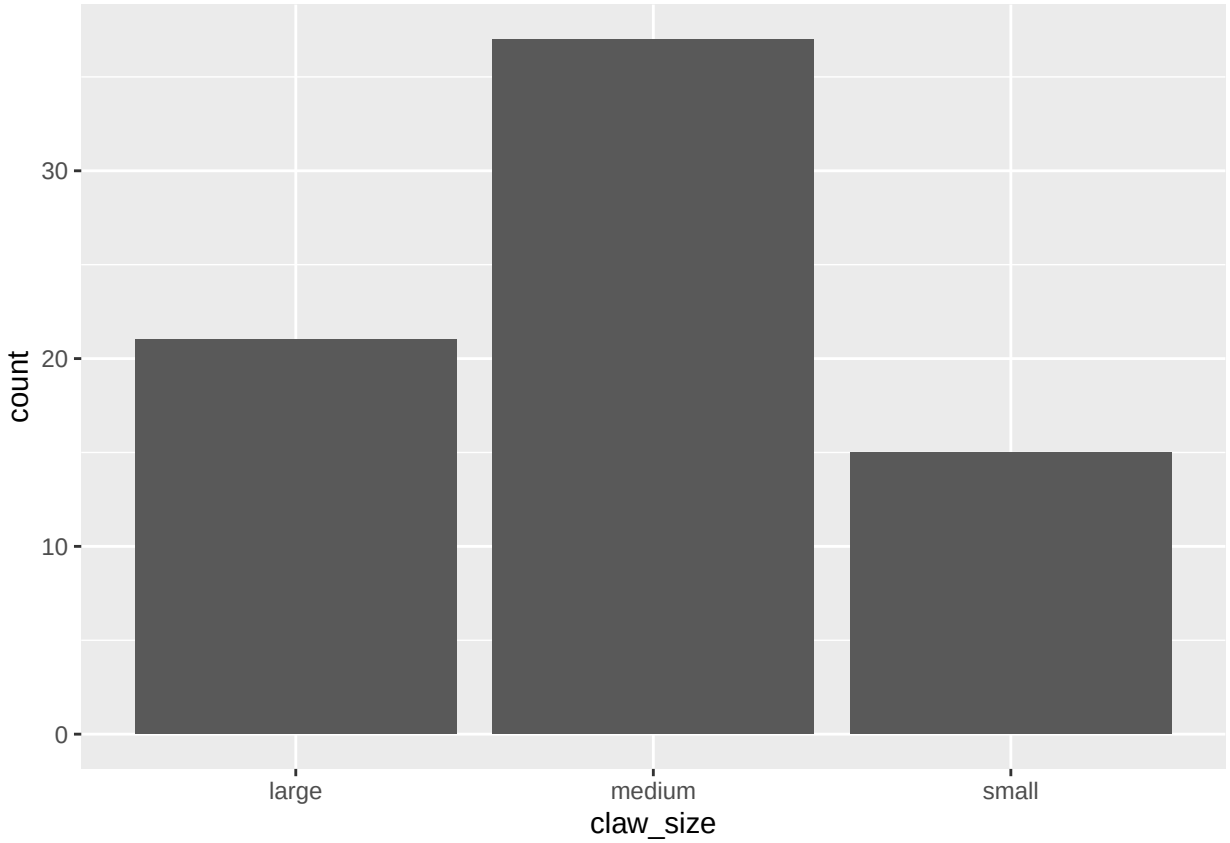
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# checking the first rows of the dataframe
head(uca)
```

```
##   crab car_width car_length claw_length dactyl_length dactyl_height
## 1   68     13.03       8.82       15.11          9.61          2.79
## 2   32     13.30       9.43       17.41         13.03          2.89
## 3   73     13.56       9.50       17.75         12.66          3.03
## 4   10     14.04       9.61       17.64         11.76          3.20
## 5   41     13.96       9.83       17.86         11.77          3.51
## 6   24     13.93       9.67       18.04         12.62          3.47
##   manus_height pollex_length manus_width manus_length apodeme_area car_mass
## 1         6.23          6.67        3.50         8.44        17.06     0.71
## 2         6.77          8.21        4.21         9.20        16.63     0.81
## 3         6.84          9.05        3.94         8.70        19.56     0.82
## 4         7.20          8.08        4.27         9.56        20.01     0.89
## 5         7.13          8.41        4.39         9.45        17.69     0.87
## 6         7.25          8.85        4.06         9.19        15.13     0.93
##   claw_mass crab_mass claw_size mass_class
## 1      0.19      0.90     small        low
## 2      0.26      1.07     small        low
## 3      0.25      1.07     small        low
## 4      0.29      1.18     small        low
## 5      0.31      1.18     small        low
## 6      0.29      1.22     small        low
```

```r
# creating figure: frequency of claw size classes
p1 <- ggplot(data=uca,aes(x=claw_size)) +
  geom_bar()

# show figure
p1
```



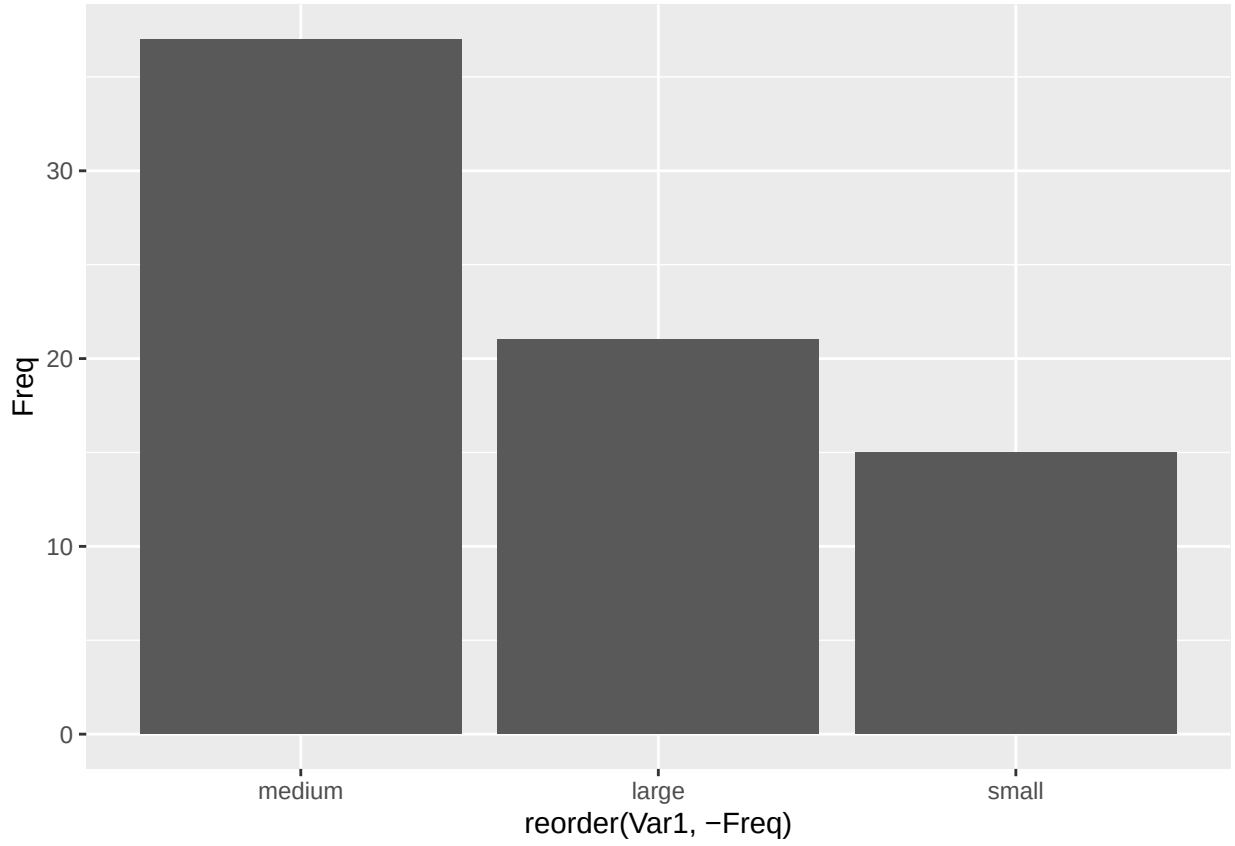

```
# frequency table of claw size classes
freq_t <- table(uca$claw_size)
freq_t
```

```
##
##  large medium  small
##     21     37     15
```

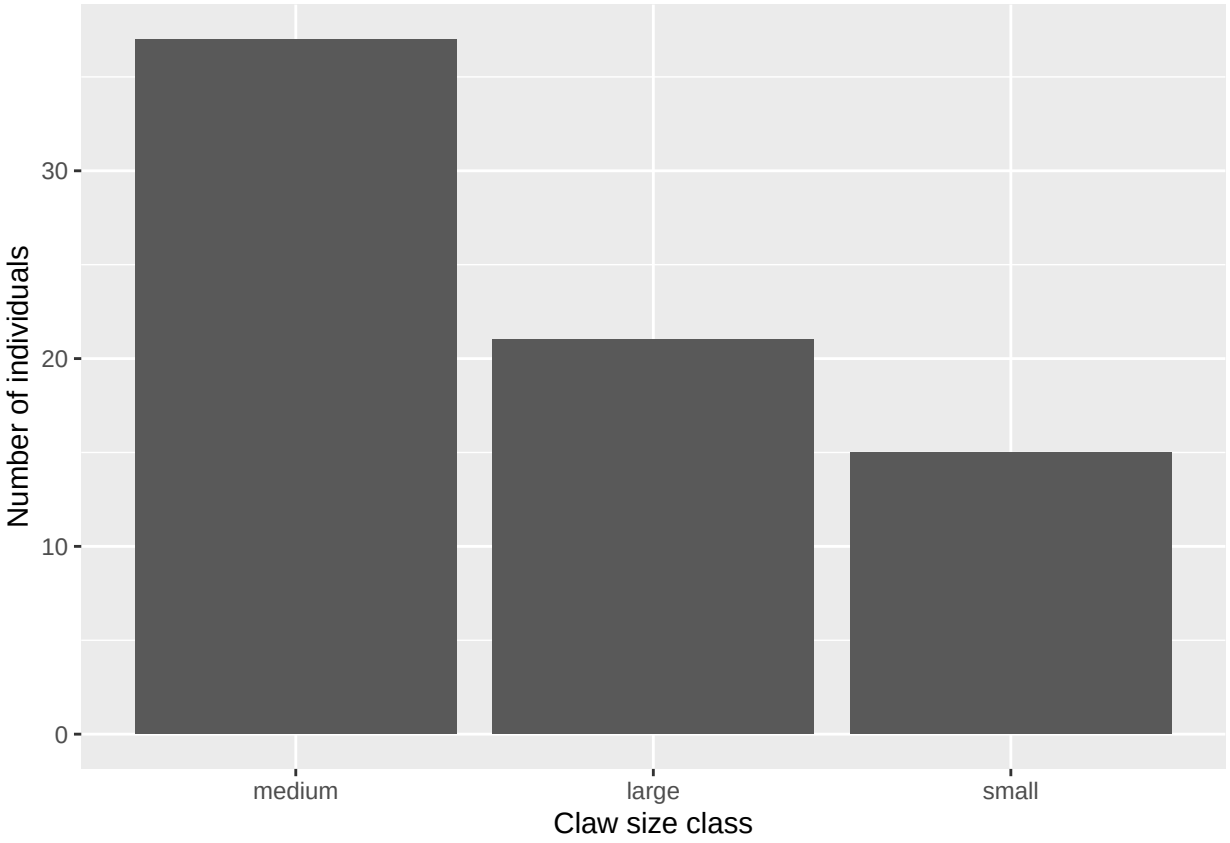
```
# converting freq_t into a data frame to plot it
freq_t <- as.data.frame(freq_t)
freq_t
```

```
##     Var1 Freq
## 1  large   21
## 2 medium   37
## 3  small   15
```

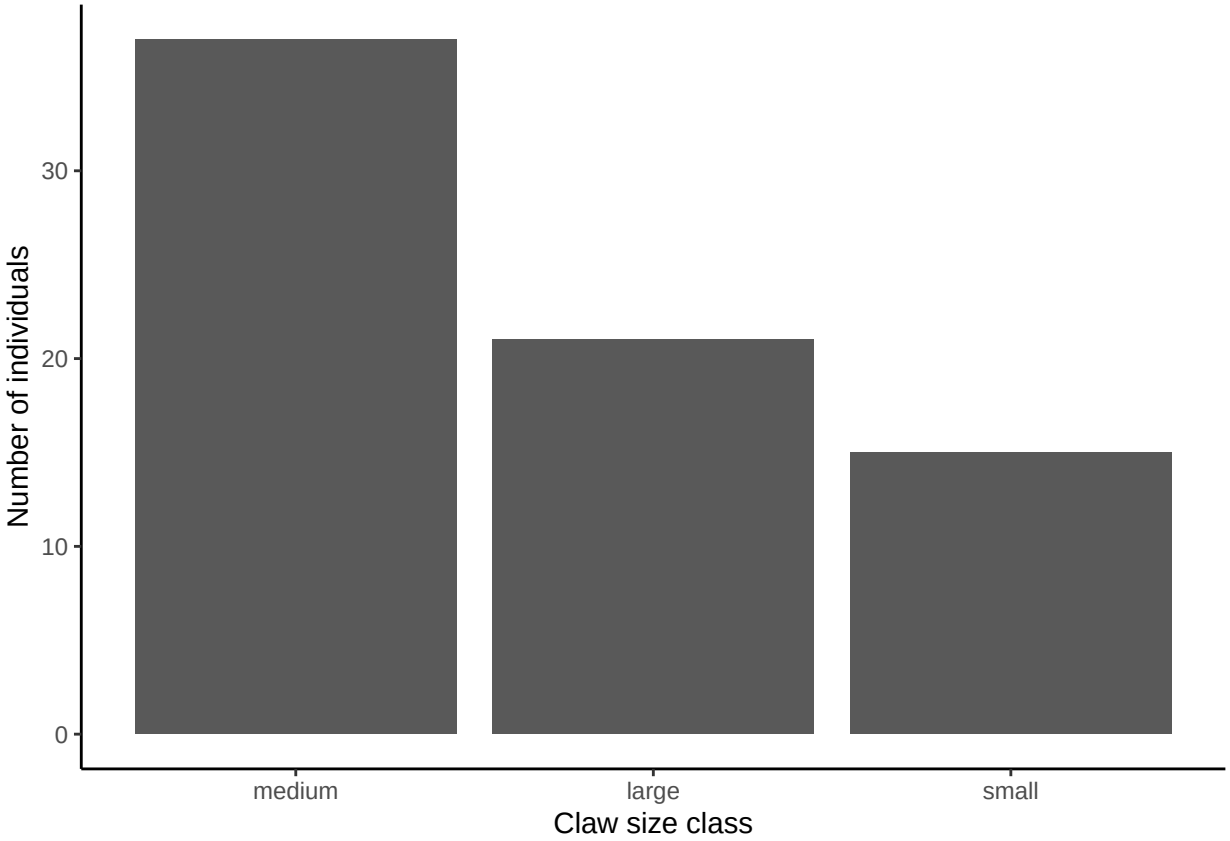
```
# reordering the bars per magnitude using the new dataframe
p1 <- ggplot(data=freq_t,aes(x=reorder(Var1,-Freq),y=Freq)) +
   geom_bar(stat="identity")
p1
```



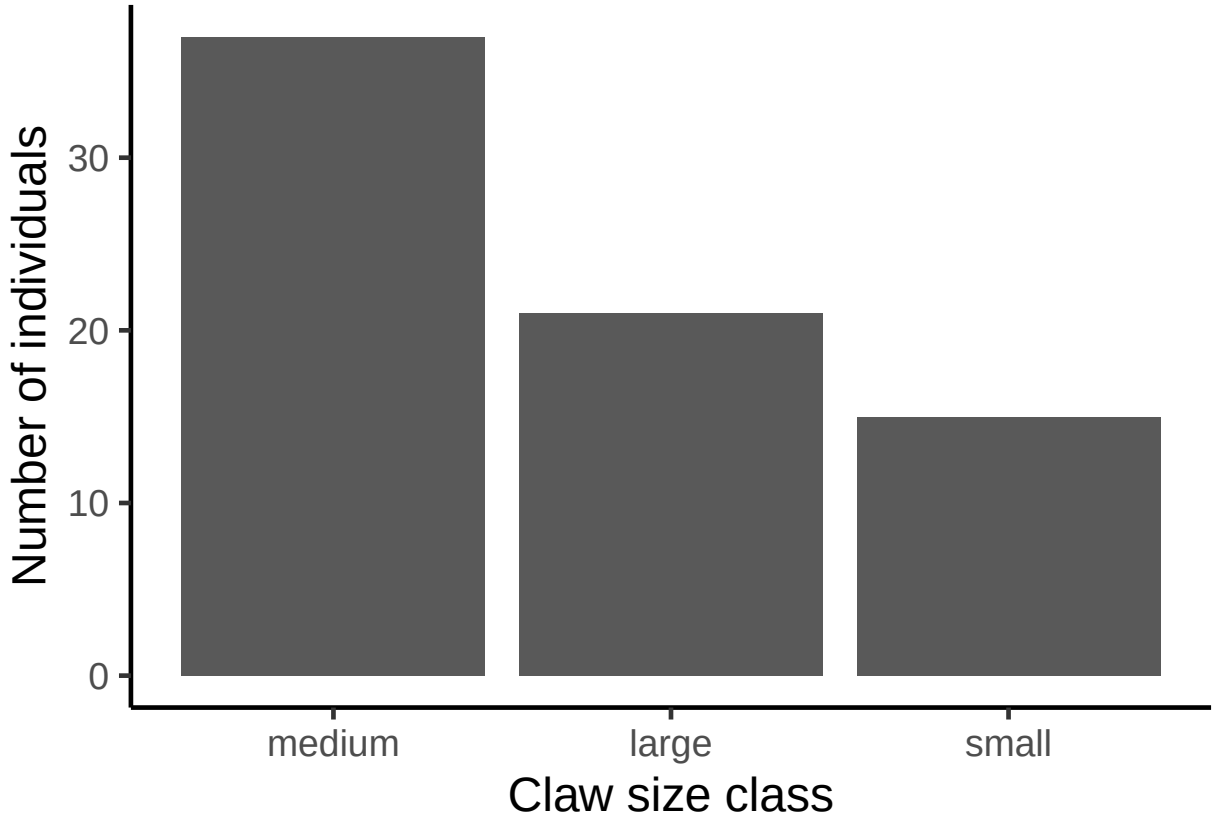

```
# adding y- and x-axis labels using ylab() and xlab()
p1 <- ggplot(data=freq_t,aes(x=reorder(Var1,-Freq),y=Freq)) +
  geom_bar(stat="identity") +
  ylab("Number of individuals") +
  xlab("Claw size class")
p1
```




```
# deleting the background color using theme_classic()
p1 <- ggplot(data=freq_t,aes(x=reorder(Var1,-Freq),y=Freq)) +
  geom_bar(stat="identity") +
  ylab("Number of individuals") +
  xlab("Claw size class") +
  theme_classic()
p1
```




```
# increasing the font size to "18"
p1 <- ggplot(data=freq_t,aes(x=reorder(Var1,-Freq),y=Freq)) +
  geom_bar(stat="identity") +
  ylab("Number of individuals") +
  xlab("Claw size class") +
  theme_classic(18)
p1
```

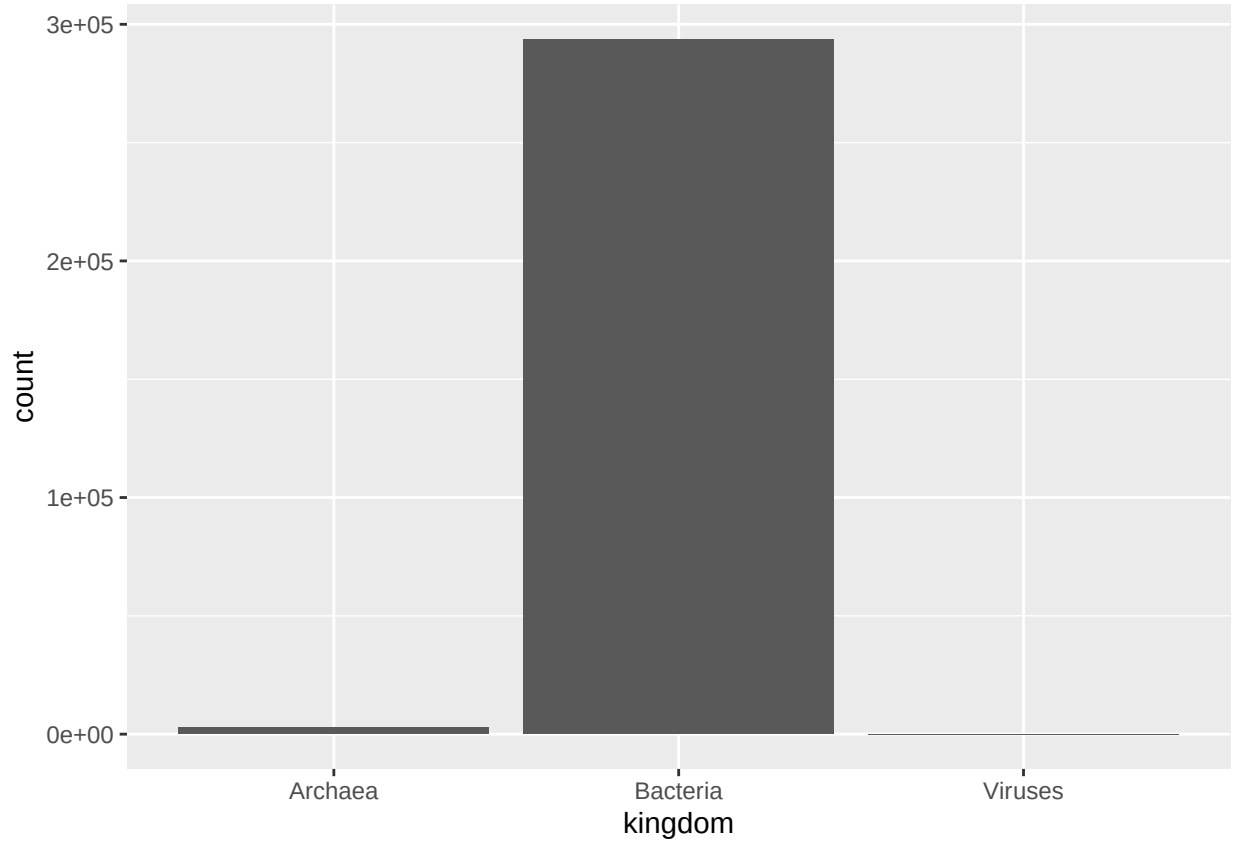


Question Answers

a. The variable "claw_size" is a factor or categorical variable.
b. Claw size classes are not represented equally in the sample data.
c. The claw size class least represented in the sample data is small.

Challenge 1 - Replicate Plot on Lab Manual

```
# name of plot for easy editing further down in the chunks/script: p2
p2 <- ggplot(data=micro,aes(x=kingdom)) +
  geom_bar()
p2
```



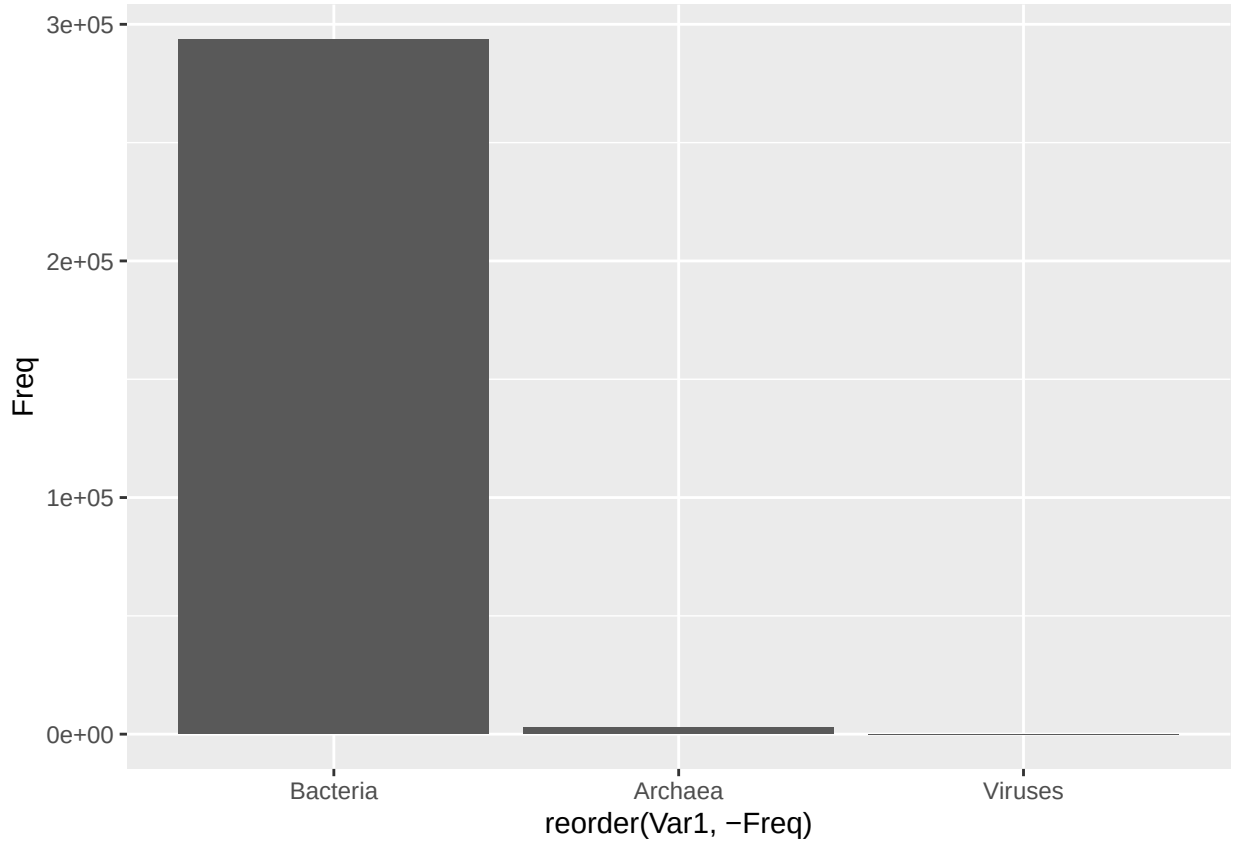

```
#creating a frequency table with only observations of the kingdom variable
freq_tmicro <- table(micro$kingdom)
freq_tmicro
```

```
##
##  Archaea Bacteria  Viruses
##     2854   293650       17
```

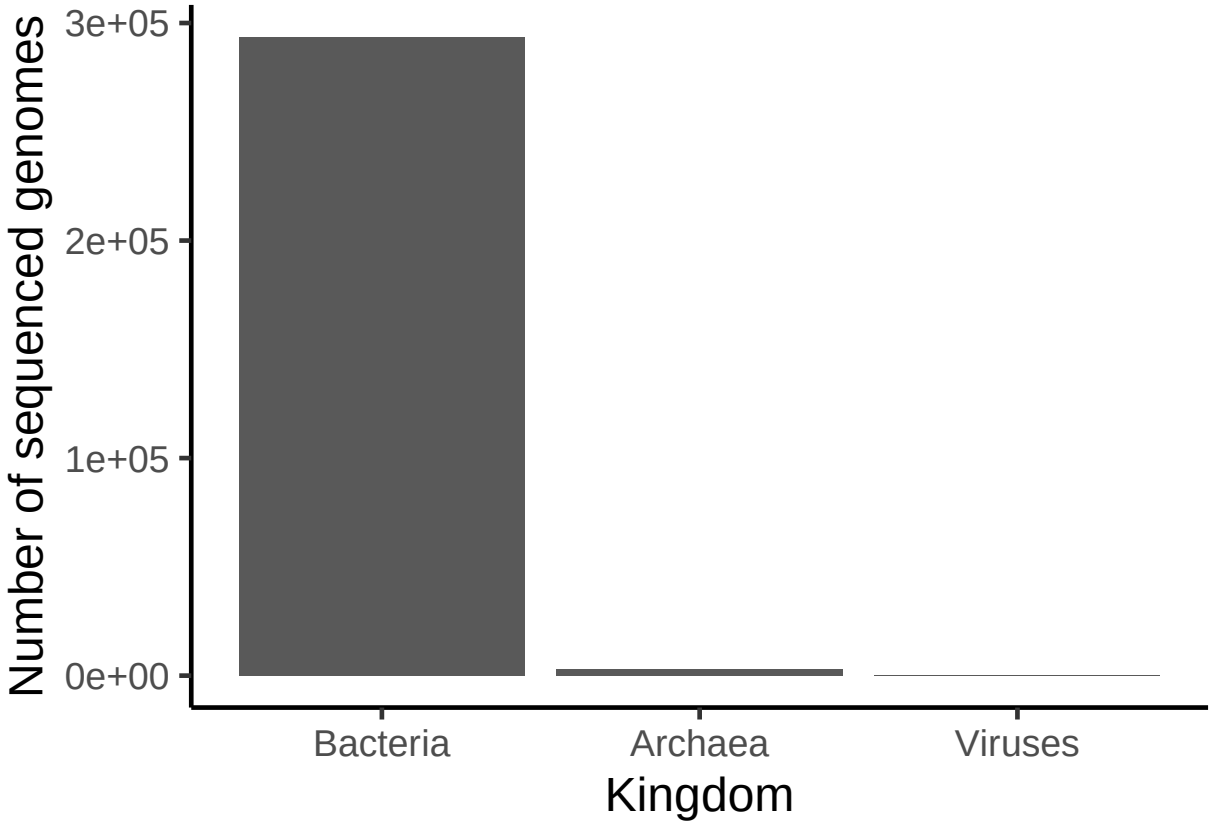
```
# converting freq_tmicro into a data frame to plot it
freq_tmicro <- as.data.frame(freq_tmicro)
freq_tmicro
```

```
##       Var1   Freq
## 1  Archaea   2854
## 2 Bacteria 293650
## 3  Viruses     17
```

```
# reordering the bars per magnitude using the new dataframe
p2 <- ggplot(data=freq_tmicro,aes(x=reorder(Var1,-Freq),y=Freq)) +
  geom_bar(stat="identity")
p2
```



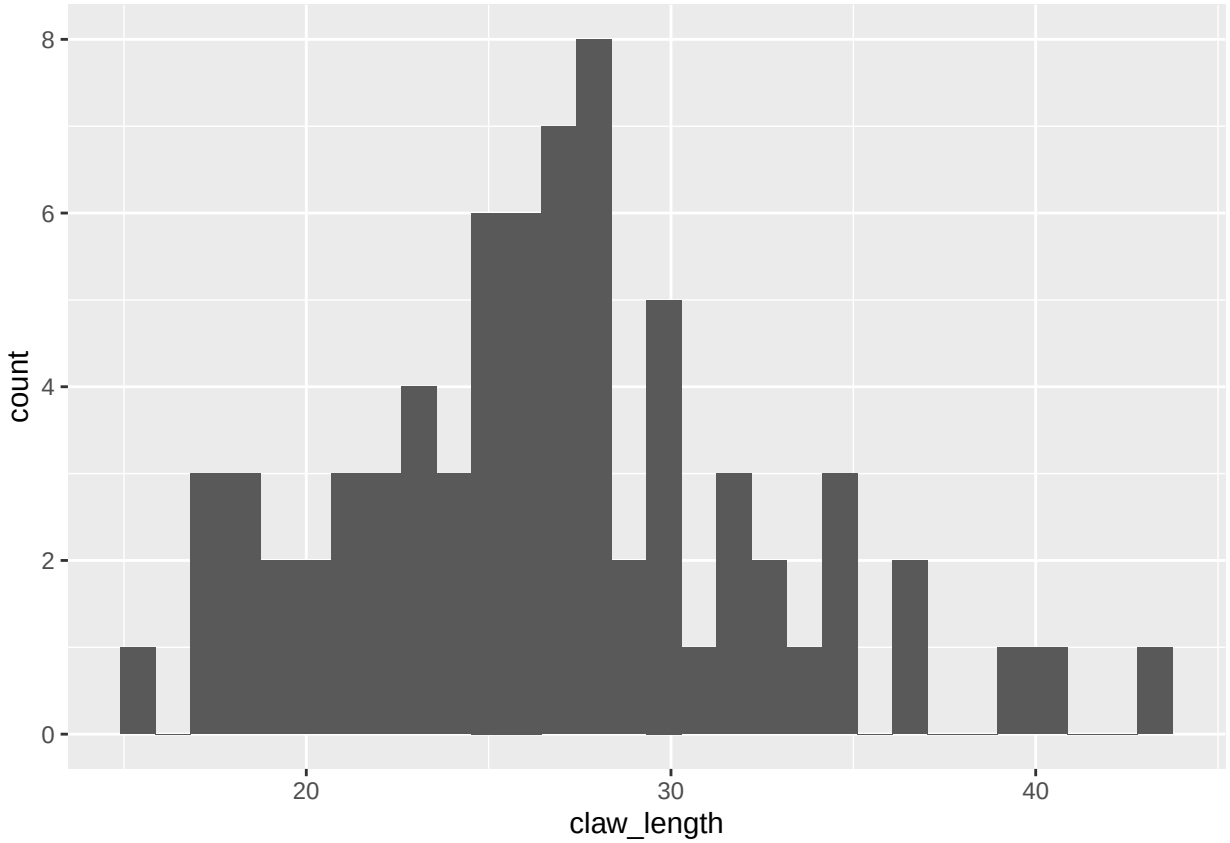

```
# adding labels and increasing font size plus theme_classic()
p2 <- ggplot(data=freq_tmicro,aes(x=reorder(Var1,-Freq),y=Freq)) +
  geom_bar(stat="identity") +
  ylab("Number of sequenced genomes") +
  xlab("Kingdom") +
  theme_classic(18)
p2
```



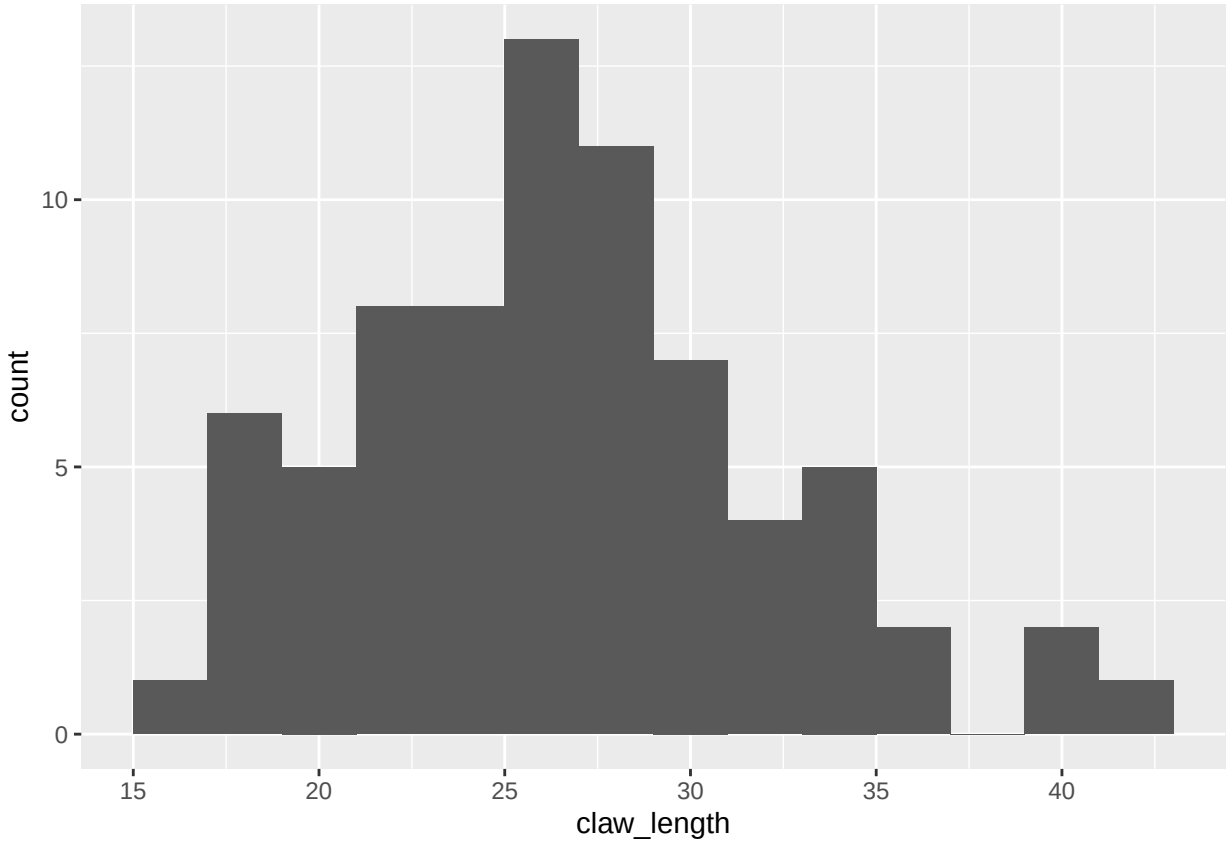
Section 2B - Displaying Data for One Variable Using a HIstogram

```
# histogram for claw length distribution
p3 <- ggplot(uca,aes(x=claw_length)) +
  geom_histogram()
p3
```

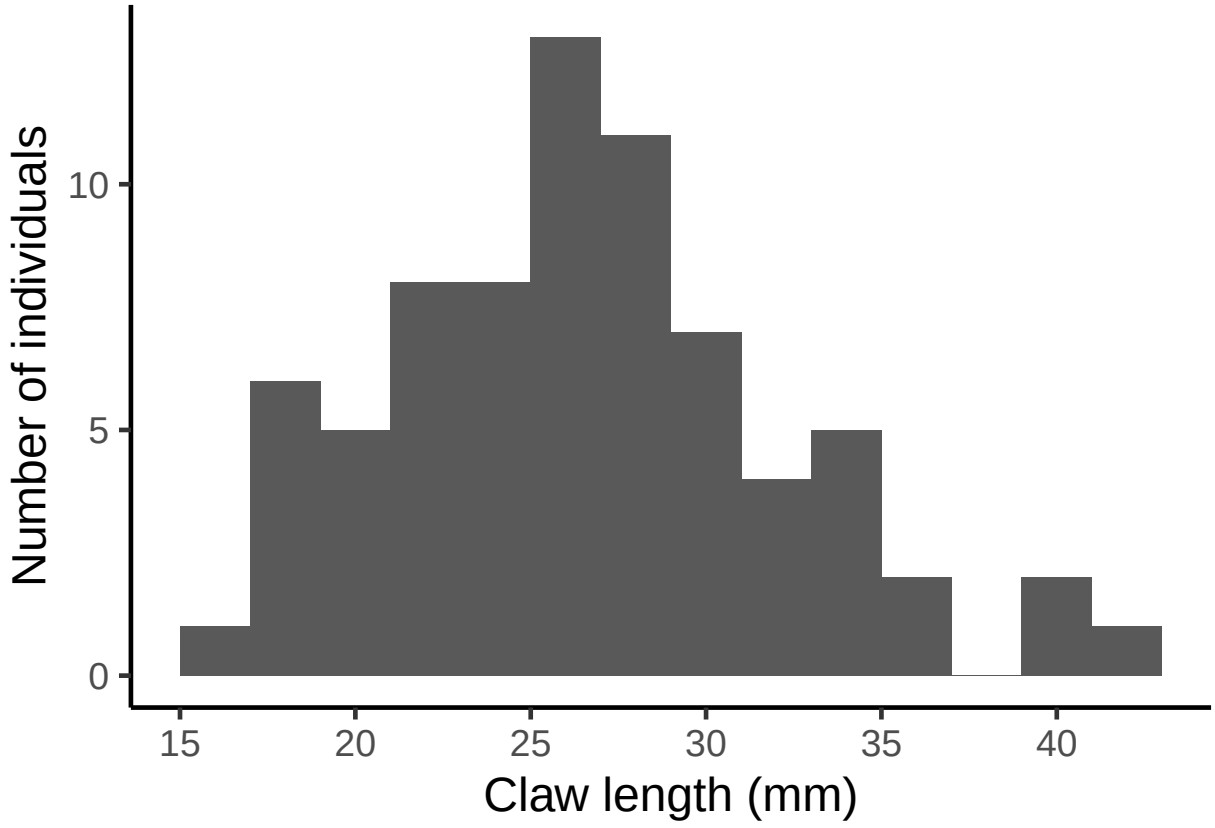
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# changing the bin width
p3 <- ggplot(uca,aes(x=claw_length)) +
  geom_histogram(binwidth = 2)
p3
```




```
# adding aesthetics
p3 <- ggplot(uca,aes(x=claw_length)) +
  geom_histogram(binwidth = 2) +
  ylab("Number of individuals") +
  xlab("Claw length (mm)") +
  theme_classic(18)
p3
```



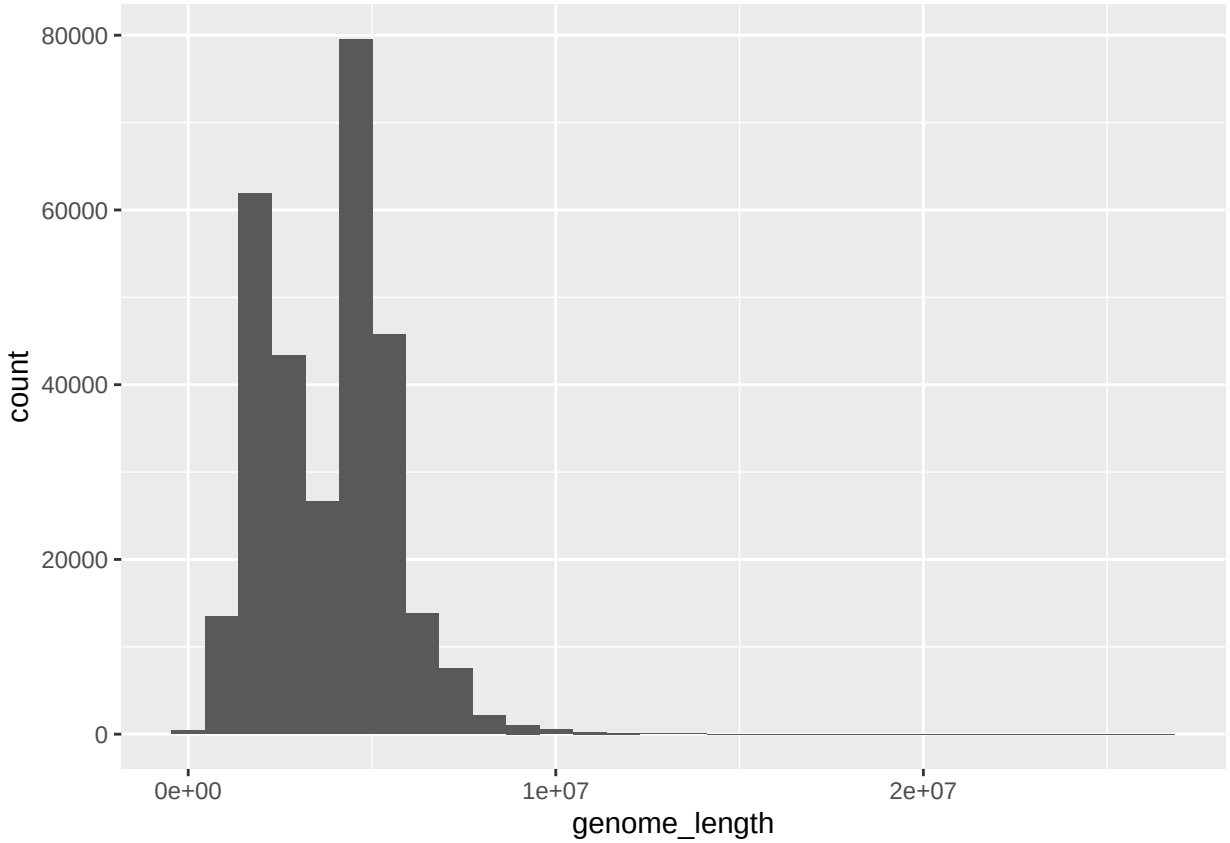
Question Answers

a. The variable 'claw_length' is a numerical one.
b. As one increases bin width, there is a lower resolution of observations on the x-axis. Essentially the size-class ranges of 'claw_length' grow so that a higher frequency of observations are included per bin.
c. The variable 'claw_length' exhibits a is normal frequency distribution.

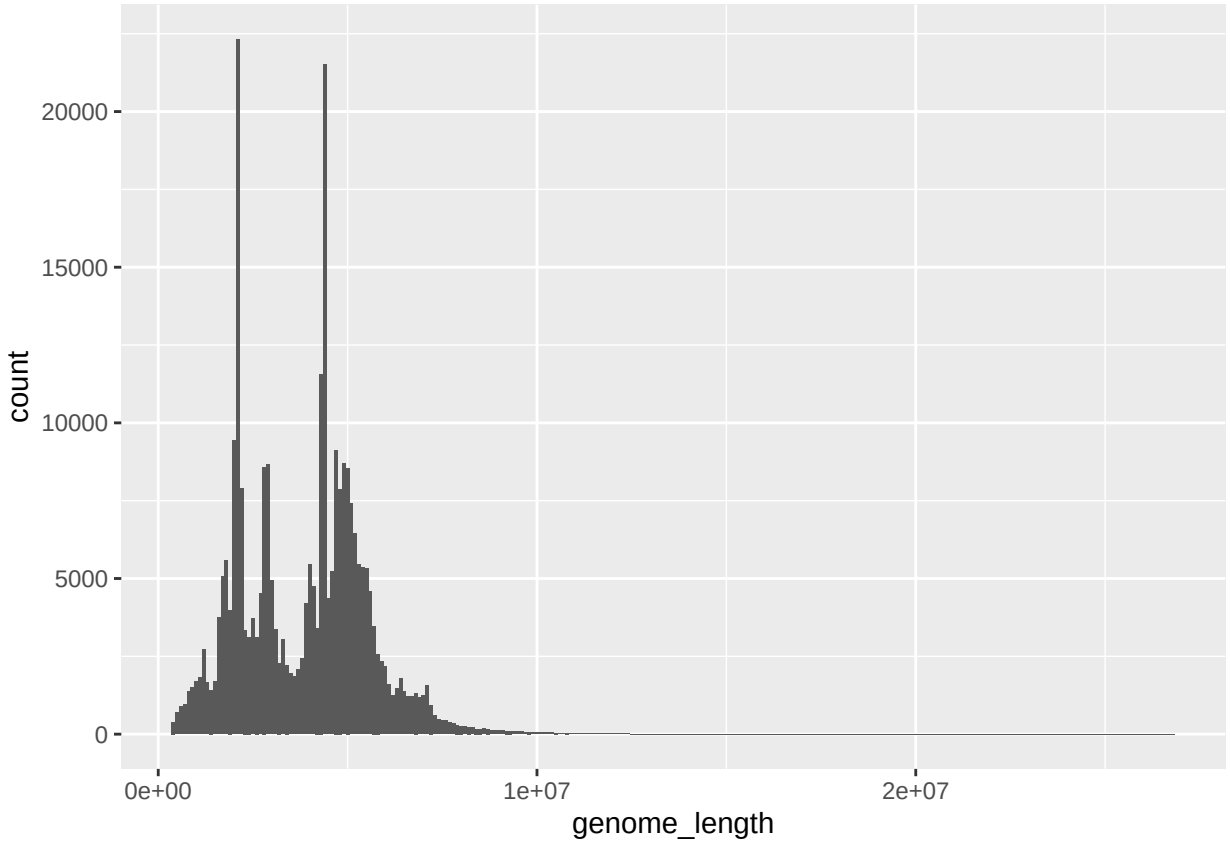
Challenge 2 - Replicate Figure 6 in Lab Manual

```
# name of plot for easy editing further down in the chunks/script: p4
p4 <- ggplot(data=micro,aes(x=genome_length)) +
  geom_histogram()
p4
```

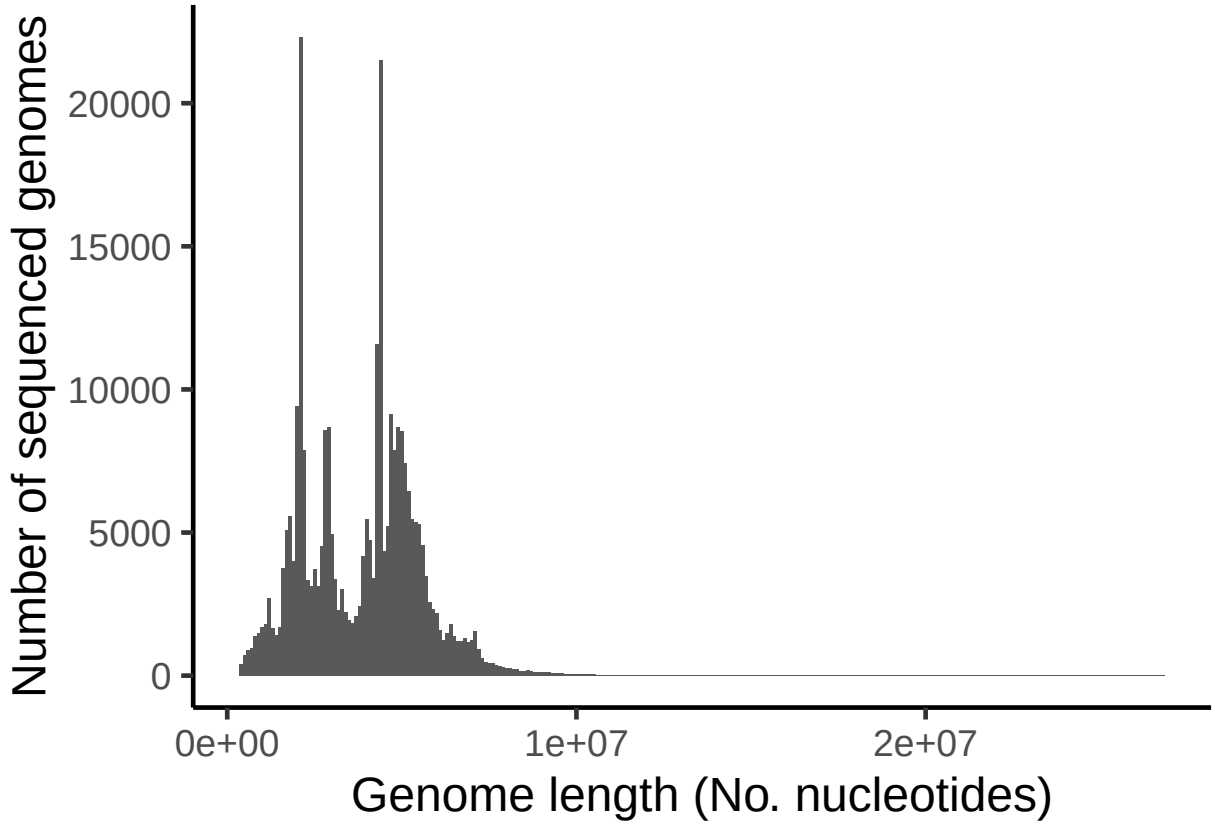
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# can change the number of bins first to get a sense of how many would make sense using (bins=#)
p4 <- ggplot(data=micro,aes(x=genome_length)) +
  geom_histogram(binwidth=100000)
p4
```

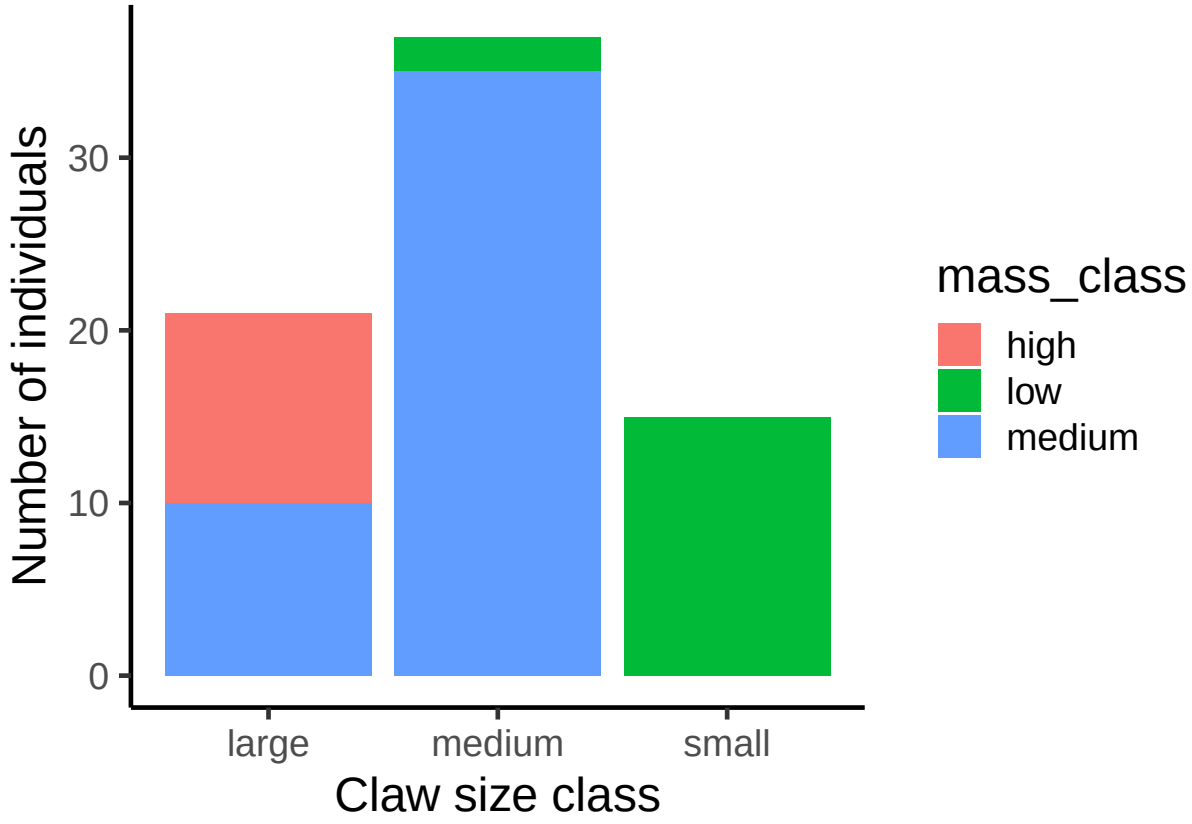



```
# adding aesthetics
p4 <- ggplot(micro,aes(x=genome_length)) +
  geom_histogram(binwidth=100000) +
  ylab("Number of sequenced genomes") +
  xlab("Genome length (No. nucleotides)") +
  theme_classic(18)
p4
```



Section 3A - Displaying Associations Between Two Variables Using a Mosaic Plot

```
# mosaic plot for claw size and crab body mas
p5 <- ggplot(data=uca,aes(x=claw_size,fill=mass_class)) +
  geom_bar() +
  ylab("Number of individuals") +
  xlab("Claw size class") +
  theme_classic(18)
p5
```

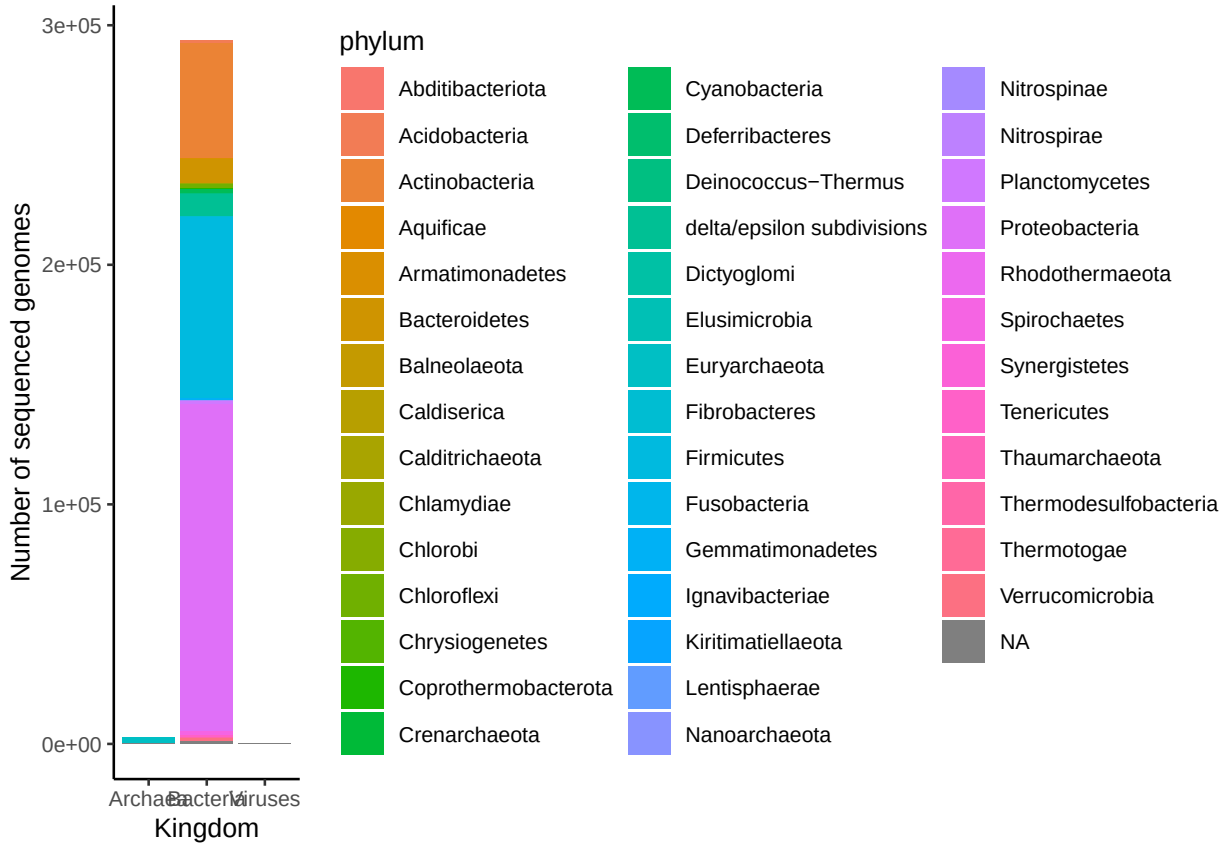


Question Answers

a. The variables 'claw_size' and 'mass_class' are factor or categorical variables.
b. Claw size and crab body mass seem to be positively correlated, with larger claw sizes corresponding to crabs of higher mass classes.
c. Based on the figure, I would expect a normal distribution of mass classes in a population of crabs with large claw sizes.

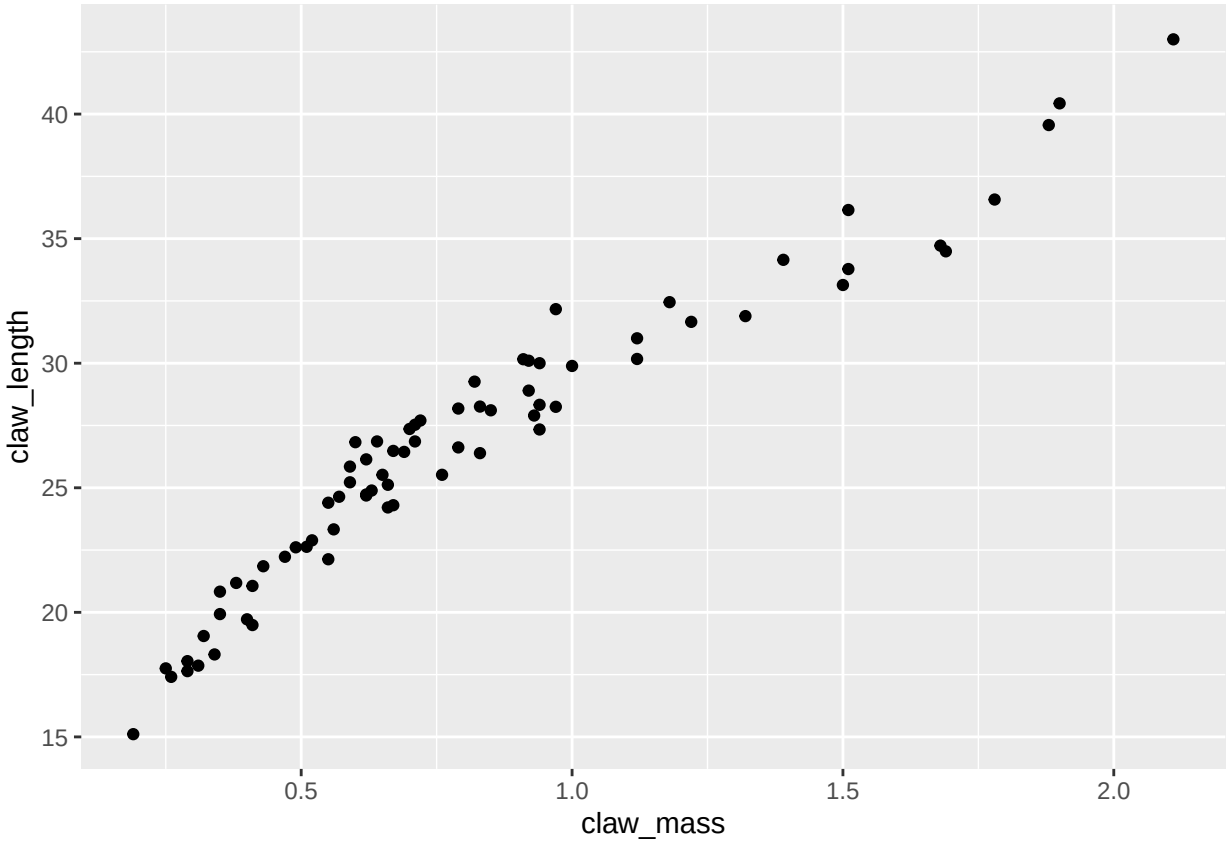
Challenge 3 - Reproduce Figure 7

```
# name of challenge mosaic plot: p6, smaller font to fit large legend
p6 <- ggplot(data=micro,aes(x=kingdom,fill=phylum)) +
  geom_bar() +
  ylab("Number of sequenced genomes") +
  xlab("Kingdom") +
  theme_classic(10)
p6
```

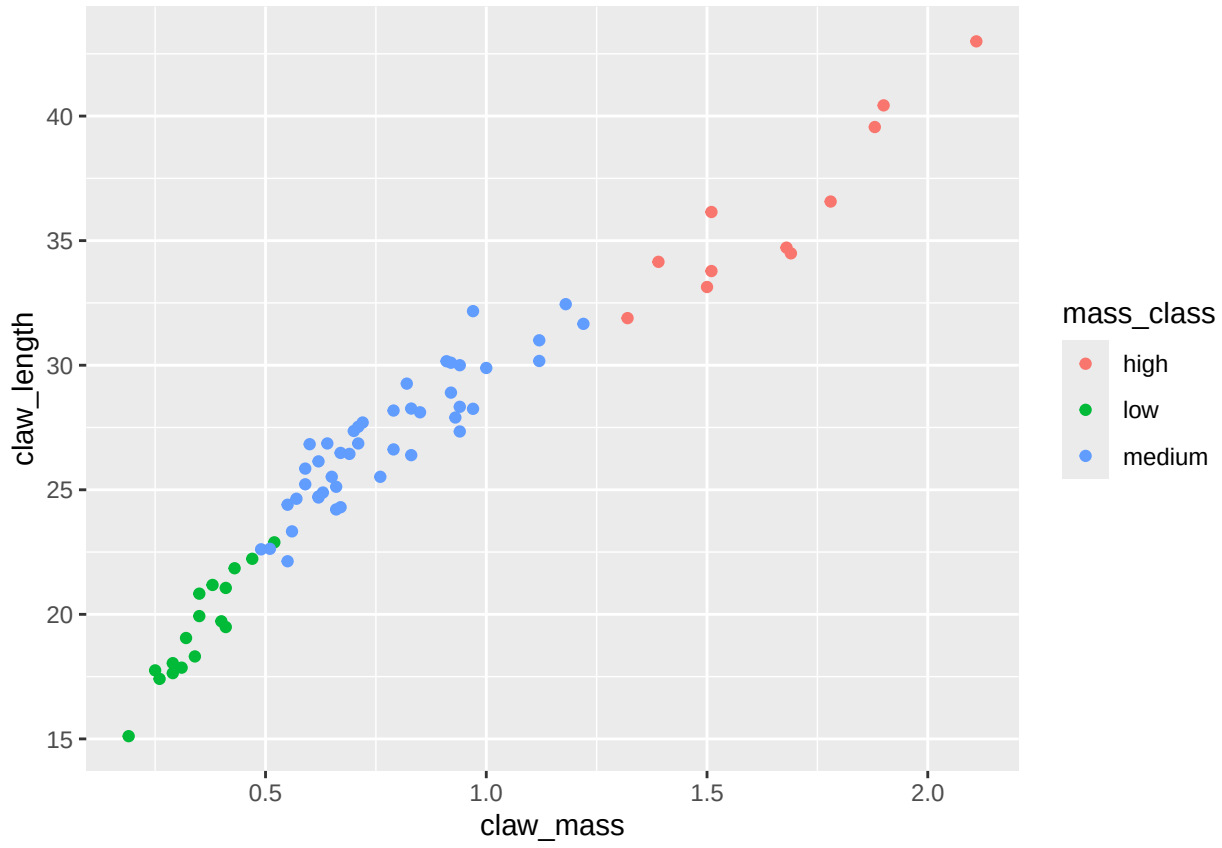


Section 3B - Displaying Associations Between Two Variables Using Scatter Plots

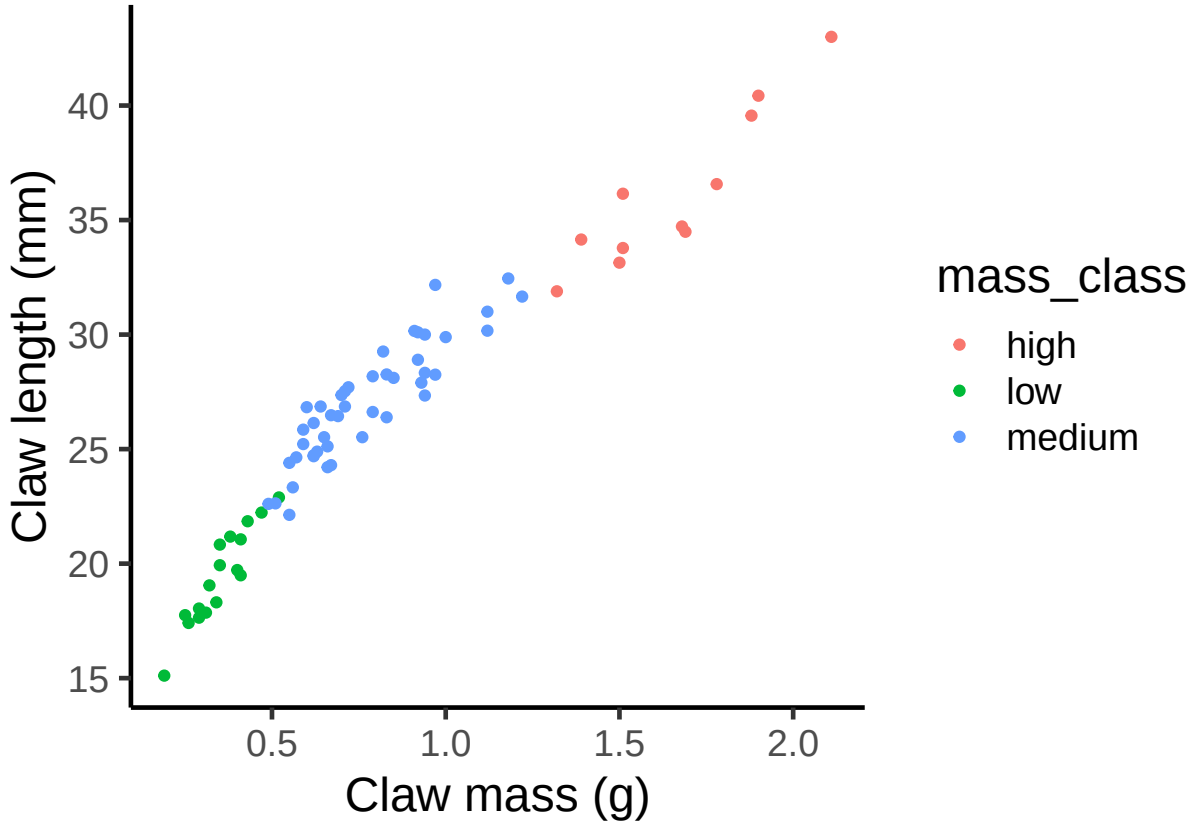
```
# scatter plot for claw length and claw mass
p7 <- ggplot(uca,aes(x=claw_mass,y=claw_length)) +
  geom_point()
p7
```




```
# differentiating by mass class adding the aesthetic of color=mass_class
p7 <- ggplot(uca,aes(x=claw_mass,y=claw_length,color=mass_class)) +
  geom_point()
p7
```




```
# adding all aesthetics
p7 <- ggplot(uca,aes(x=claw_mass,y=claw_length,color=mass_class)) +
  geom_point() +
  ylab("Claw length (mm)") +
  xlab("Claw mass (g)") +
  theme_classic(18)
p7
```

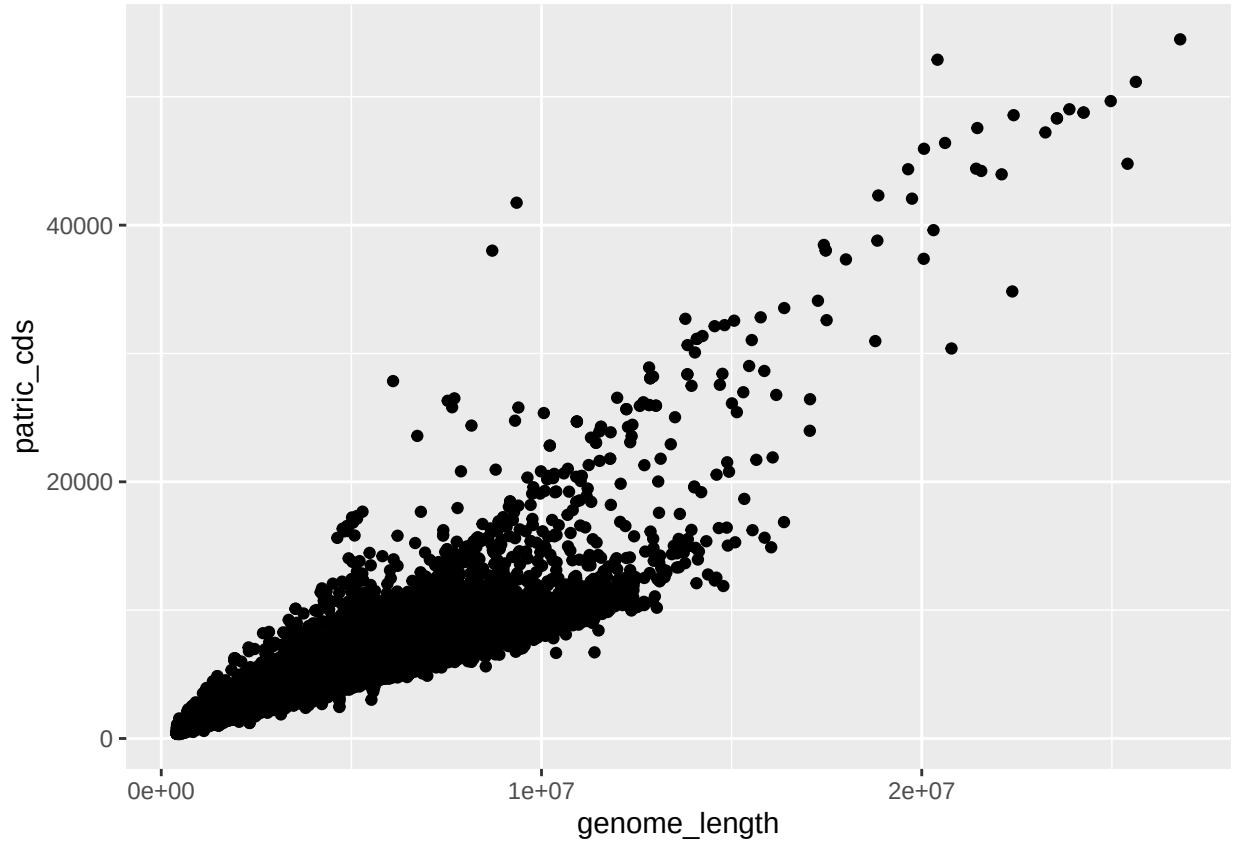


Question Answers

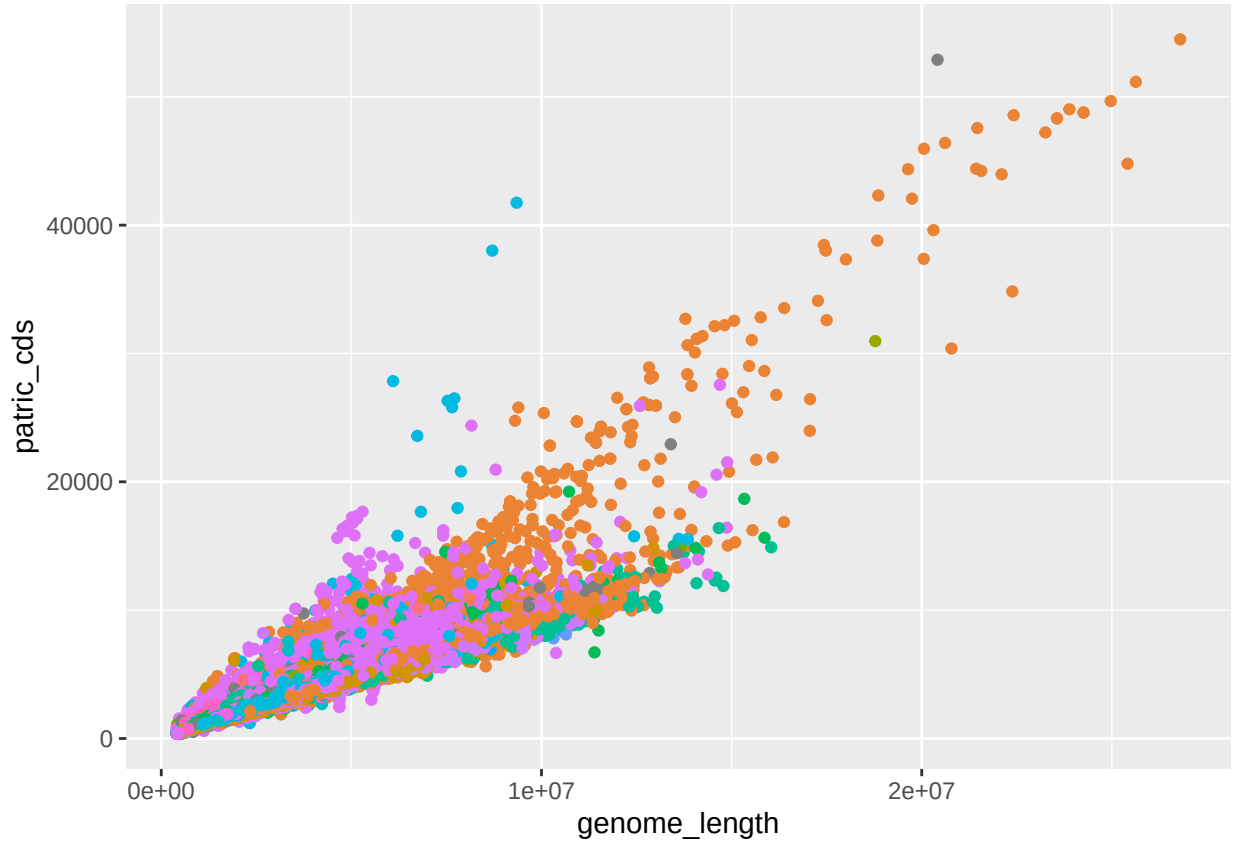
a. The variables 'claw_mass' and 'claw_length' are numerical variables, while the variable 'mass_class' is a factor or categorical variable.
b. In the above figure, the response variable is claw length and the explanatory variable is claw mass.
c. From the figure above, we can infer that claw length is positively correlated with claw mass throughout three size classes of crab body mass.

Challenge 4 - Reproduce Figure 8

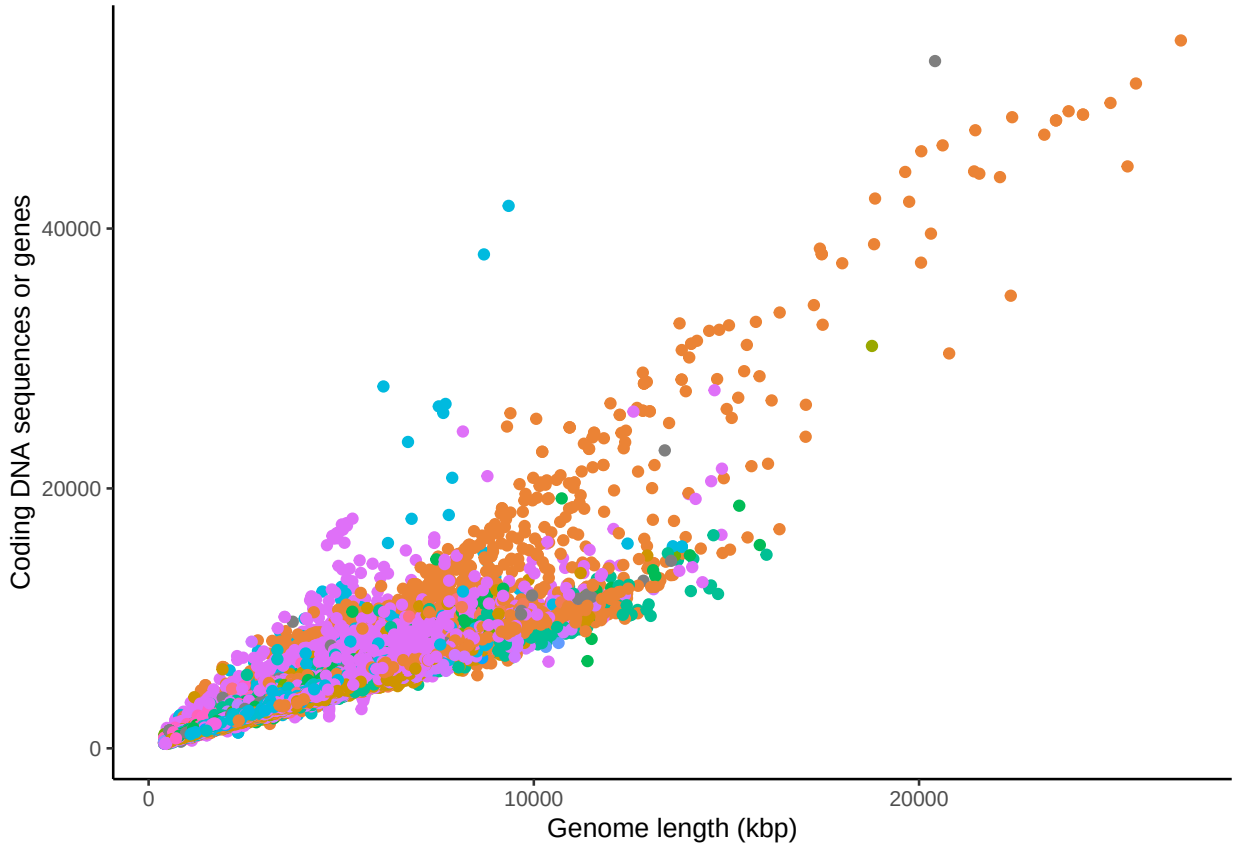
```
# scatter plot for genome length and number of genes
p8 <- ggplot(micro,aes(x=genome_length,y=patric_cds)) +
  geom_point()
p8
```




```
# differentiating by phylum with aesthetic of color=phylum
p8 <- ggplot(micro,aes(x=genome_length,y=patric_cds, color=phylum)) +
  geom_point()+
  theme(legend.position = 'none')
p8
```

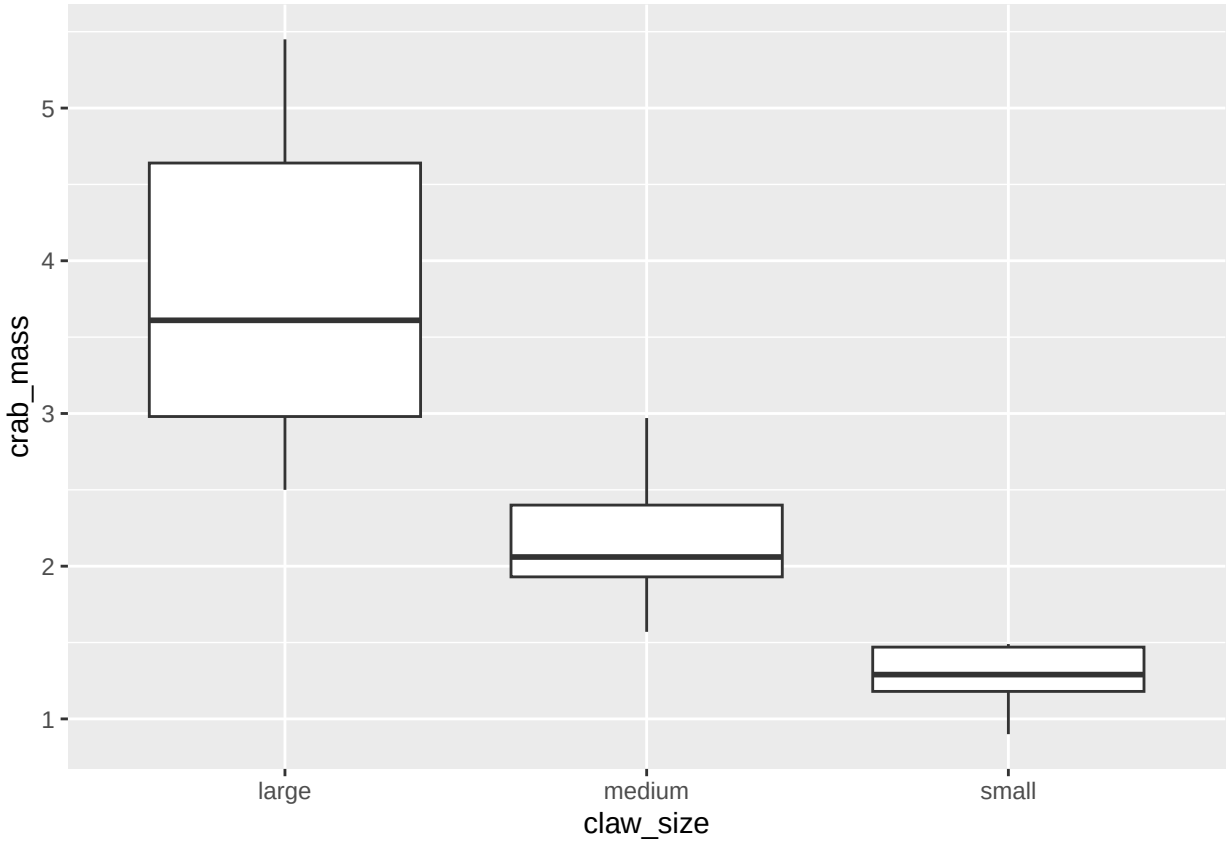



```
# adding all aesthetics, plus the appropriate x-axis scale of kilo-base pairs
p8 <- ggplot(micro,aes(x=genome_length/1000,y=patric_cds, color=phylum)) +
  geom_point() +
  ylab("Coding DNA sequences or genes") +
  xlab("Genome length (kbp)") +
  theme_classic(10)+ theme(legend.position = 'none')
p8
```

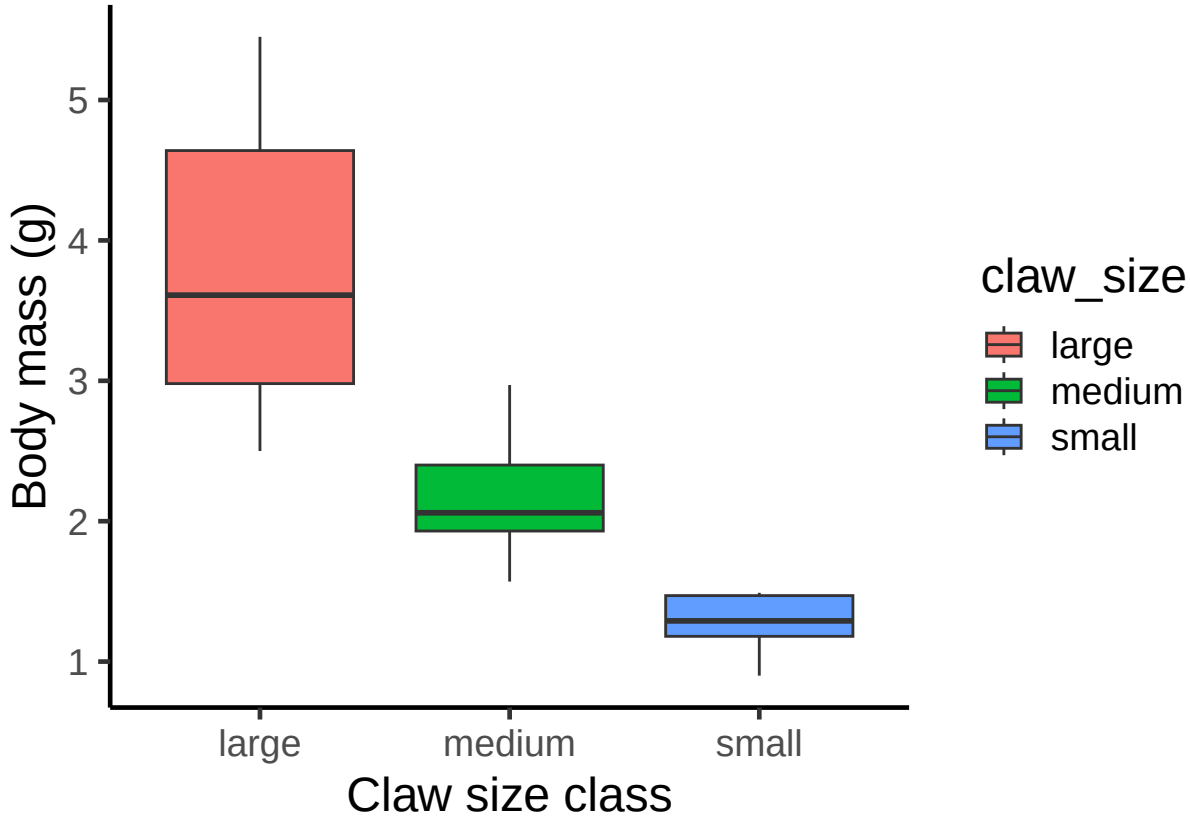


Section 3C - Displaying Associations Between Two Variables (Numerical and Categorical) Using Scatter Box Plots

```
# box plot for body mass across claw size using geom_boxplot
p9 <- ggplot(uca,aes(x=claw_size,y=crab_mass)) +
   geom_boxplot()
p9
```




```
# adding aesthetics
p9 <- ggplot(uca,aes(x=claw_size,y=crab_mass,fill=claw_size)) +
   geom_boxplot() +
  ylab("Body mass (g)") +
  xlab("Claw size class") +
  theme_classic(18)
p9
```



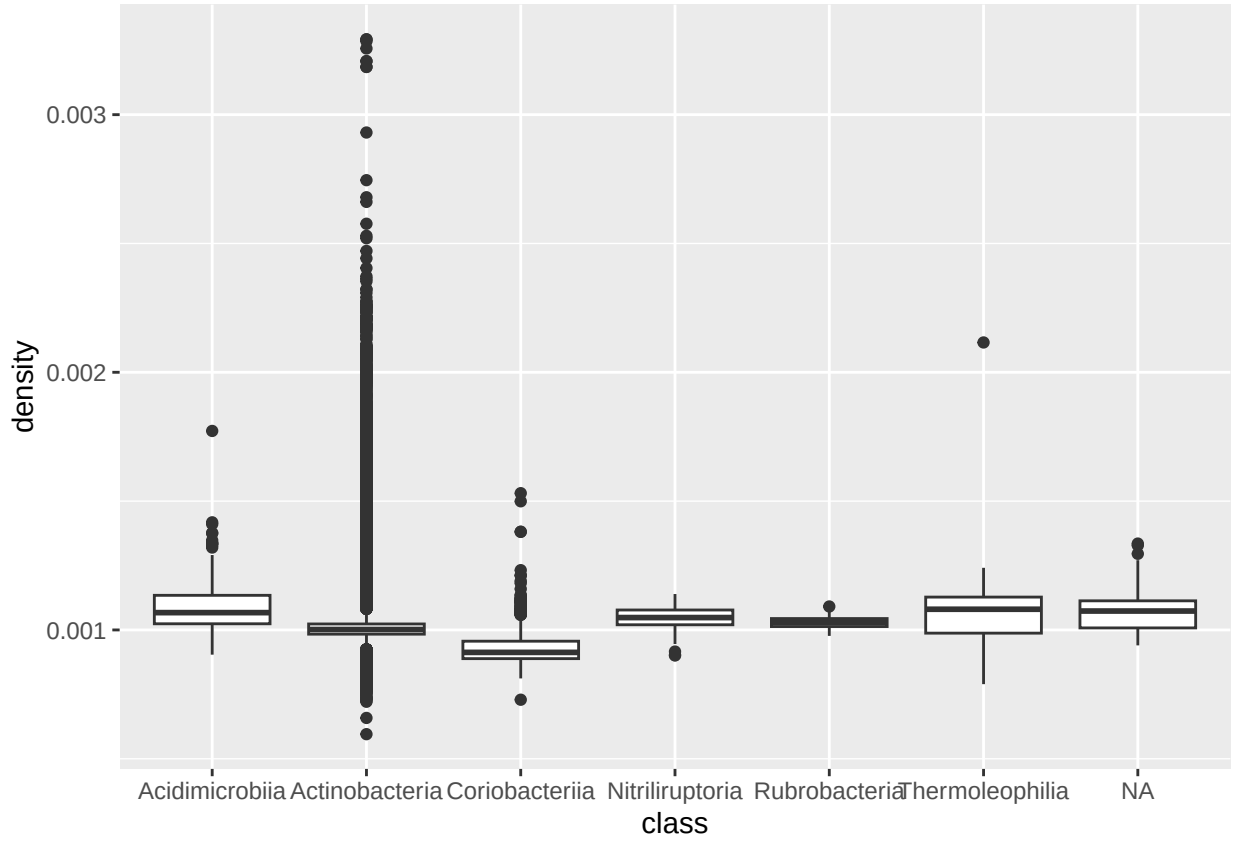
Question Answers

a. The frequency distribution of body mass across claw size classes large and medium exhibit normal or near normal distributions, while the claw size class small is left-skewed or negatively skewed.

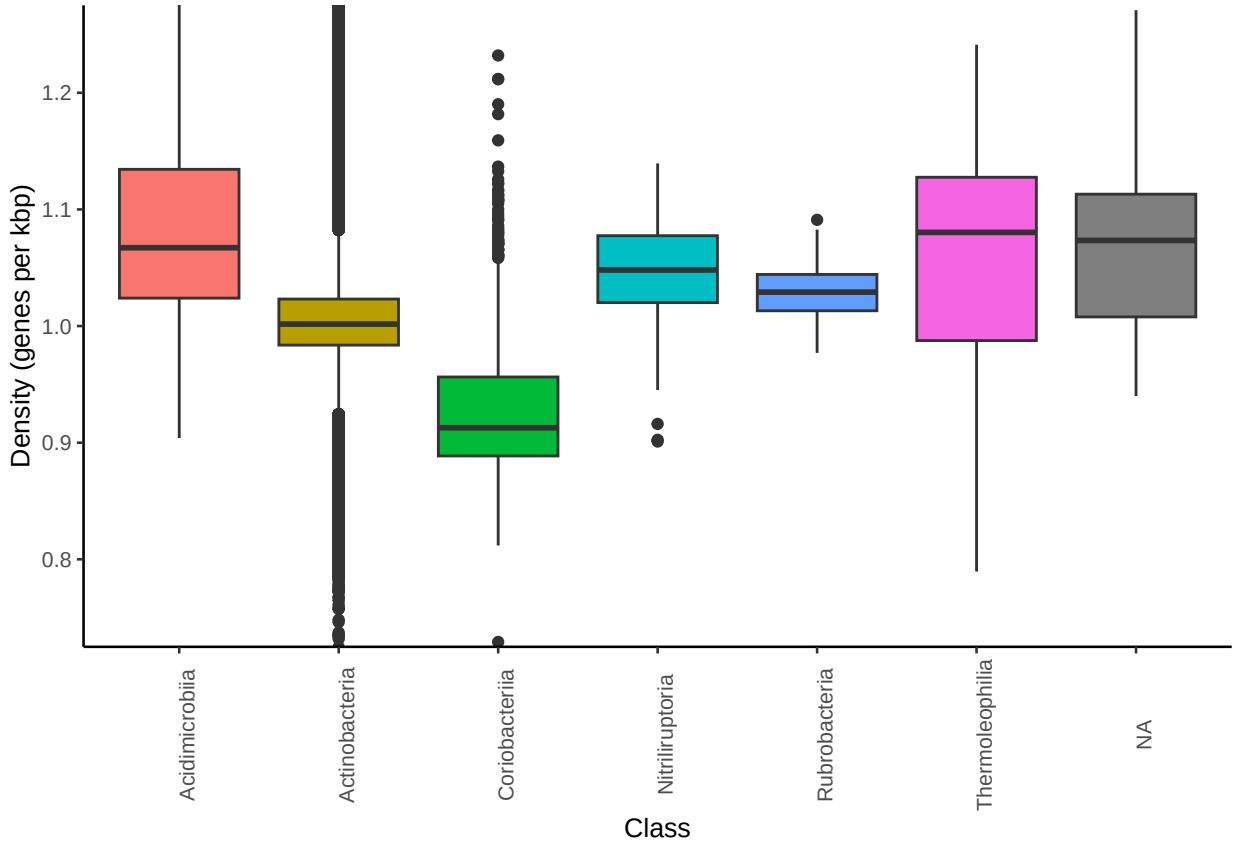
Challenge 5

```
# create a new data frame with only phylum 'Actinobacteria'
Actinobact <- micro %>% filter(phylum == 'Actinobacteria')

# create a plot with new frame
p0 <- ggplot(Actinobact, aes(x=class, y=density))+
  geom_boxplot()
p0
```




```
# add color and axis label and bound adjustments
p0 <- ggplot(Actinobact, aes(x=class, y=density*1000, fill=class))+
  geom_boxplot()+
  ylab("Density (genes per kbp)")+
  coord_cartesian(ylim=c(0.75,1.25))+
  xlab("Class")+
  theme_classic(10)+
  theme(legend.position = 'none', axis.text.x=element_text(angle=90))
p0
```



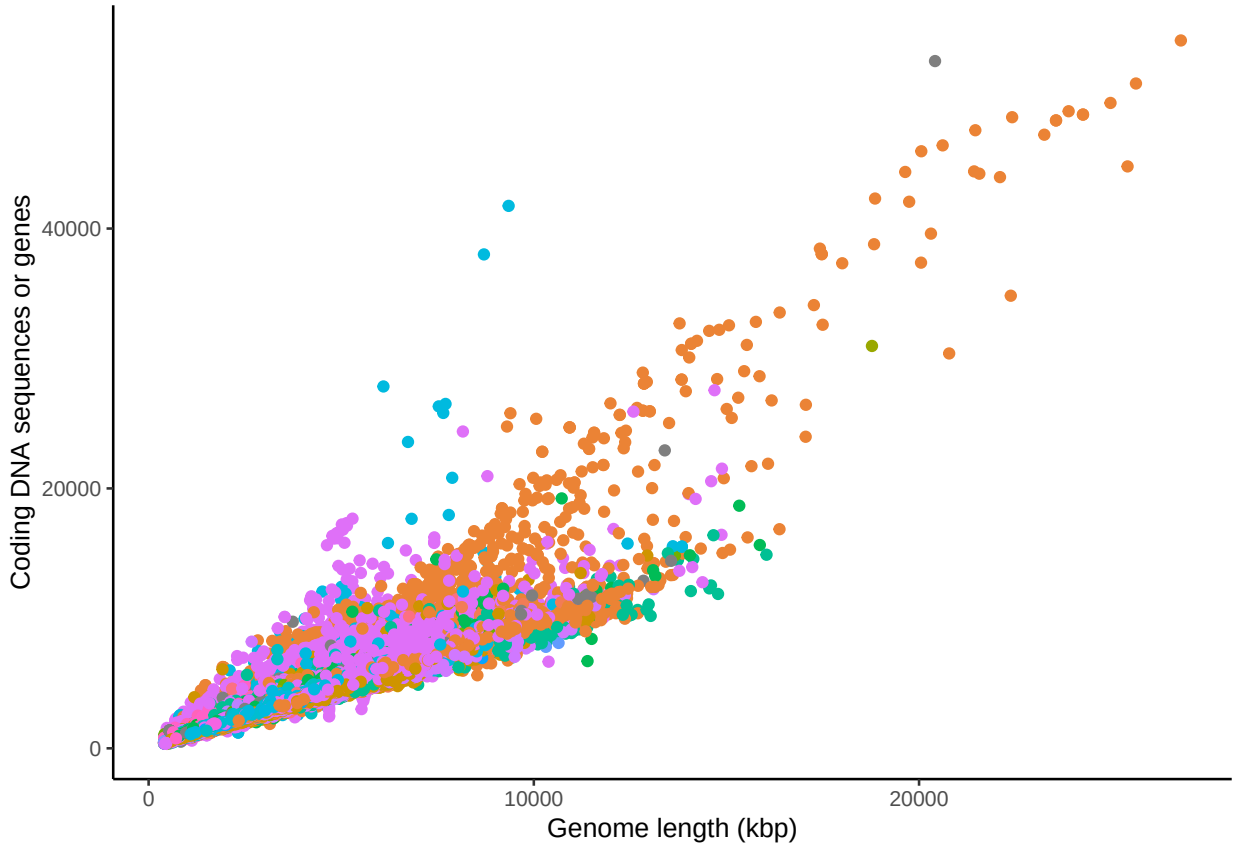
Discussion Question Answers

a. If I am interested in exploring the frequency distribution of one numerical variable, I would use a histogram. Histograms bin frequency of numerical variables in an easy-to-interpret figure that can shed light on the median, modes, and skeweness of a dataset.

b. Boxplots contain a middle meadian value line, which acts as a side to two contiguous rectangles that represent quartiles above and below the median, and whiskers that extend beyond the interquartile range to demonstrate extreme values. Boxplots are a useful visualization because they represent a condensed version of a histogram, which allows viewers to easily understand frequency distribution for a given numerical variable.

c. Figure Caption for the following Figure: 'Figure 8. Relationship between genome length and coding DNA sequences or genes. Data source: PATRICbrc dataset.'

```
p8 <- ggplot(micro,aes(x=genome_length/1000,y=patric_cds, color=phylum)) +
  geom_point() +
  ylab("Coding DNA sequences or genes") +
  xlab("Genome length (kbp)") +
  theme_classic(10)+ theme(legend.position = 'none')
p8
```


Coding DNA sequences or genes
40000
20000
0
0
10000
20000
Genome length (kbp)