**c.** Provide a brief description of the dataset's variables

The data frame consist of 8 variables and 344 observations

**species:** it is species of the penguin. It has three categories: Adelie, Chinstrap, and Gentoo.

**island:** Indicates the island where the penguin was seen. There are three islands in the dataset: Biscoe, Dream, and Torgersen.

**bill\_length\_mm:** This variable represents the length of the penguin's. It is a continuous numerical variable.

**bill\_depth\_mm:** This variable represents the depth of the penguin's. it is a continuous numerical variable.

**flipper\_length\_mm:** it is the length of the penguin's flipper in millimeters. It is a continuous numerical variable.

**body\_mass\_g**: This is the body mass of the penguin in grams. It is another continuous numerical variable.

**sex:** Indicates the gender of the penguin. It has three categories: Male, Female, and NA (for missing or unknown values).

**year:** Represents the year when the data was recorded. It is a categorical variable with two levels: 2007 and 2008.

**3.c**

. Discuss the insights gained from the summarization process. Note any interesting patterns, anomalies, missing, etc.

**1.Average Physical Characteristics by Species:** You'll see the mean bill length, bill depth, flipper length, and body mass for each penguin species. This will give you insights into the typical physical characteristics of each species.

**2.Comparative Analysis:** we can compare the mean values across species to identify any significant differences or similarities in their physical characteristics.

3. **d.** Apply the skimr5 package to the data and discuss the output. How does it compare to what you did in the previous steps? What additional information is provided?How is this useful

The skimr5 package provides a concise and informative summary of the dataset, including:

1. Variable Details: It provides a list of the variable names, data types, and the quantity of non-missing values.

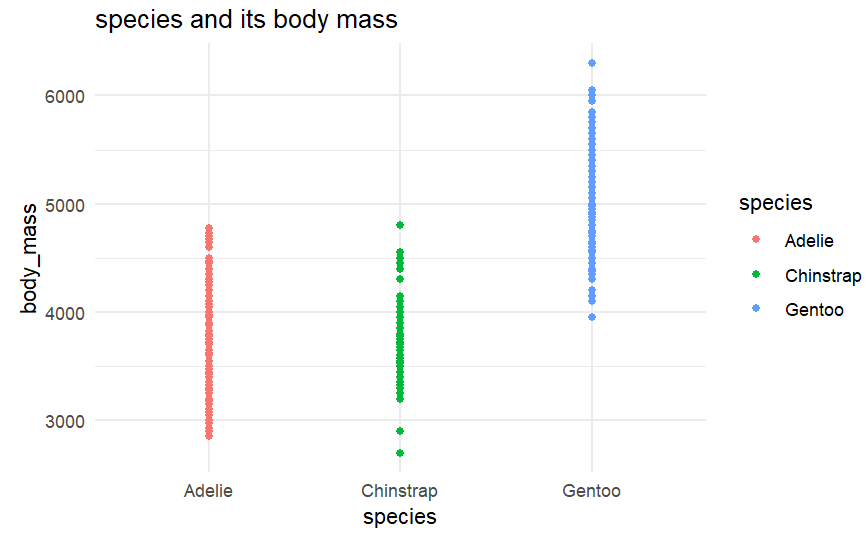
2. Quantiles and Statistics: You will get a summary of statistics for each variable, including the minimum, first quartile, median, mean, third quartile, and maximum values.

3.Missing Values: It provides information on each variable's percentage of missing values.

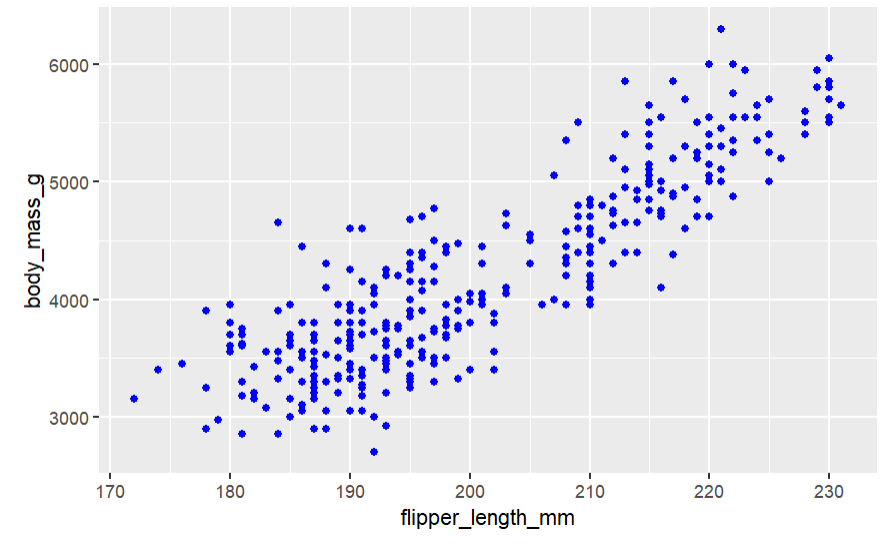
4.Histograms: It produces histograms for numeric variables to show you how the data is distributed visually.

Details relevant to the type: Skimr5 offers pertinent information based on the type of data (for example, category or numeric). The most prevalent values and their frequencies are displayed for categorical variables.

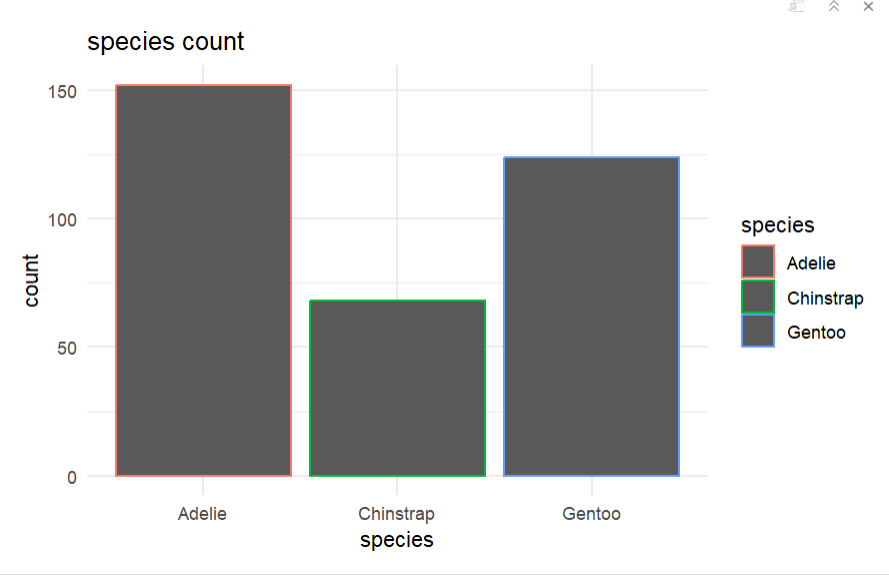
**4.** Data Visualization:  
  
**c.** Interpret the insights obtained from each visualization.



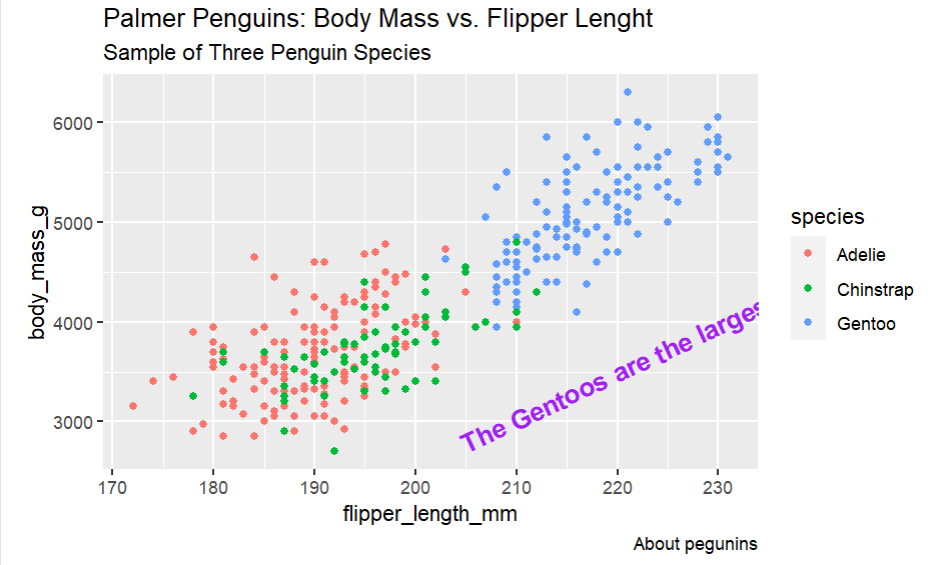
From the above plot we can see that the Gentoo species has highest body\_mass then the other two species(Adelie,chinstrap)



The body\_mass and flipper length are linearly correlated to each other



From this we can see that the number of a Adelie species more and chinstrap species are less in number



By comparing the all the three caterious of peguins with the body\_mass and flipper\_length we got to know that gentoos are the largest

**5.** Project Report and Interpretation:  
  
**b.** Summarize overall patterns, trends, or relationships you discovered. What can you  
say about each penguin?

1.The flipper\_length and body\_mass are related i.e if the body\_mass is more then the flipper\_length is also more vice versa

2.The Adelie species are more in number then the other two species

3.The Gentoos are the largest species with respect to body\_mass and flipper\_length

4.The chinstrap has some outliers with respect to the body\_mass

**5.c.** Reflect on the value of using R and the tidyverse for doing data analysis

**Data Manipulation and Transformation**:

The tidyverse, which contains tools like dplyr and tidyr, offers a potent collection of tools for data transformation and manipulation. Your data can be quickly filtered, arranged, modified, and summarized to increase accessibility and prepare it for analysis.

**Data Visualization:**

You may easily generate high-quality, customisable data visualizations with the help of ggplot2, a foundational tool in the tidyverse. A broad variety of plots and charts can be easily created since it adheres to a standard syntax of graphics.’

**Statistical Analysis:**

R is well known for its powerful statistical abilities. Statisticians and data analysts favor it because it offers a large variety of statistical tests and models.

**Open Source:**

R and the majority of the tidyverse packages are free to use and supported by a large worldwide user and contributor community since they are open-source.