DEPARTMENT OF
DATA SCIENCE

DS311
Exploratory_Data_Analysis

Introduction to EDA

Lab Activity: Hands-On EDA with the Palmer Penguins Dataset

Lab Overview: In this lab, you will apply what you've learned to a real-world dataset. We will use the **Palmer Penguins** dataset, which contains size measurements for penguins of three different species observed in Antarctica. This dataset is an excellent EDA practice case: it's a modern replacement for the classic iris dataset, with 344 penguins from 3 species collected on 3 islands. The data includes numeric measurements like bill length, flipper length, body mass, as well as categorical attributes like species, island, and sex. (It was explicitly designed as a rich dataset for data exploration and visualization.)

Objective: Follow the steps below to perform exploratory analysis on the penguins dataset. Try to answer each question by writing R code and examining the output or plots. If you're doing this in RMarkdown, you can create code chunks for each task. If you're in the R console or script, run the code and observe the results.

Data preparation: Make sure you have loaded the dataset. If you installed the `palmerpenguins` package as instructed, the data frame `penguins` should be available. (If not, run `library(palmerpenguins)` now.) The `penguins` data frame contains the cleaned dataset we will use.

```
library('palmerpenguins')
```

```
## Warning: package 'palmerpenguins' was built under R version 4.4.3
```

Exercises:

1. **Inspect the dataset structure:** How many rows and columns does the `penguins` dataset have, and what are the data types of each column? Display the first few rows to get a sense of the data. *(Hint: Use functions like `dim()`, `str()`, and `head()`.)* Also, check if there are any missing values in the dataset, and if so, which columns have missing data. *(Hint: one way is using `summary()` or `colSums(is.na(...))`.)*

```r
dim(penguins)
```

```
## [1] 344   8
```

```r
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

```r
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```r
sum(is.na(penguins$body_mass_g))
```

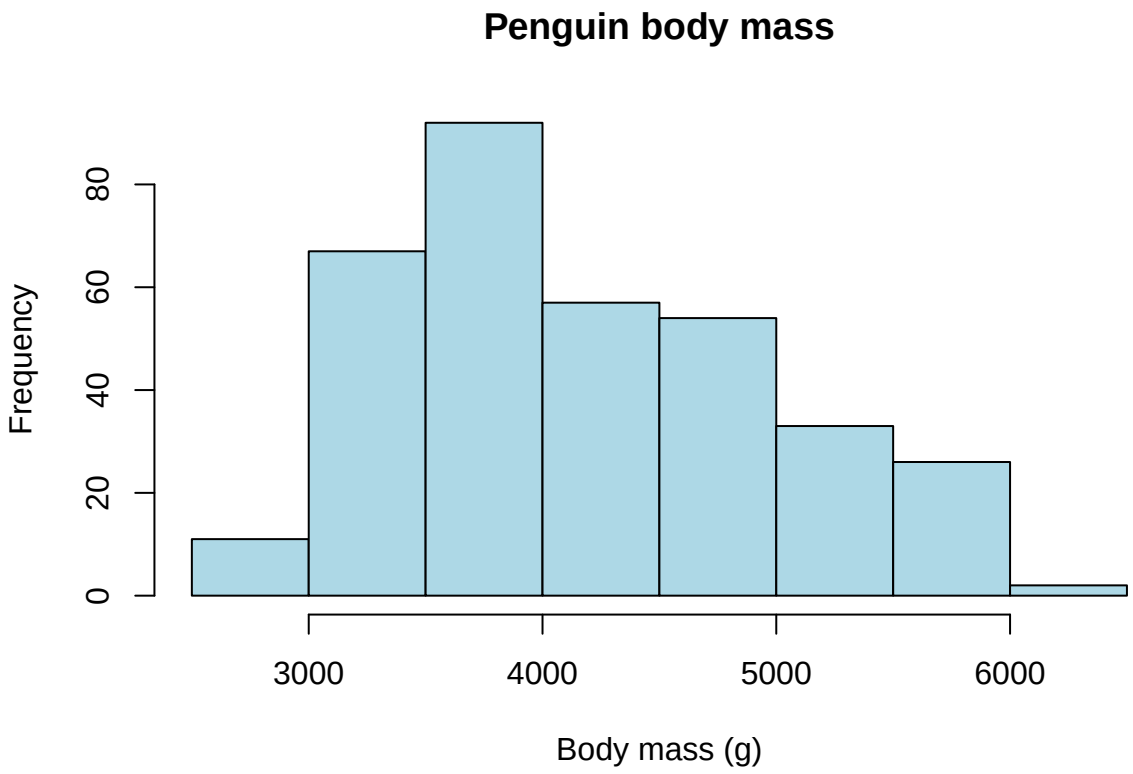
```
## [1] 2
```

2. **Univariate Analysis – Numerical Variable:** Pick a numeric variable in the dataset (for example, **body_mass_g**, which is the body mass of the penguin in grams). Compute the main summary statistics (mean, median, quartiles, etc.) for this variable and then create a histogram to visualize its distribution. Describe the distribution: is it symmetric or skewed? What is roughly the average body mass, and what range does it cover? Are there any noticeable outliers in the histogram?

```r
summary(penguins$body_mass_g)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2700    3550    4050    4202    4750    6300       2
```

```r
hist(penguins$body_mass_g,
     main = "Penguin body mass",
    xlab = 'Body mass (g)',
    col = 'lightblue')
```

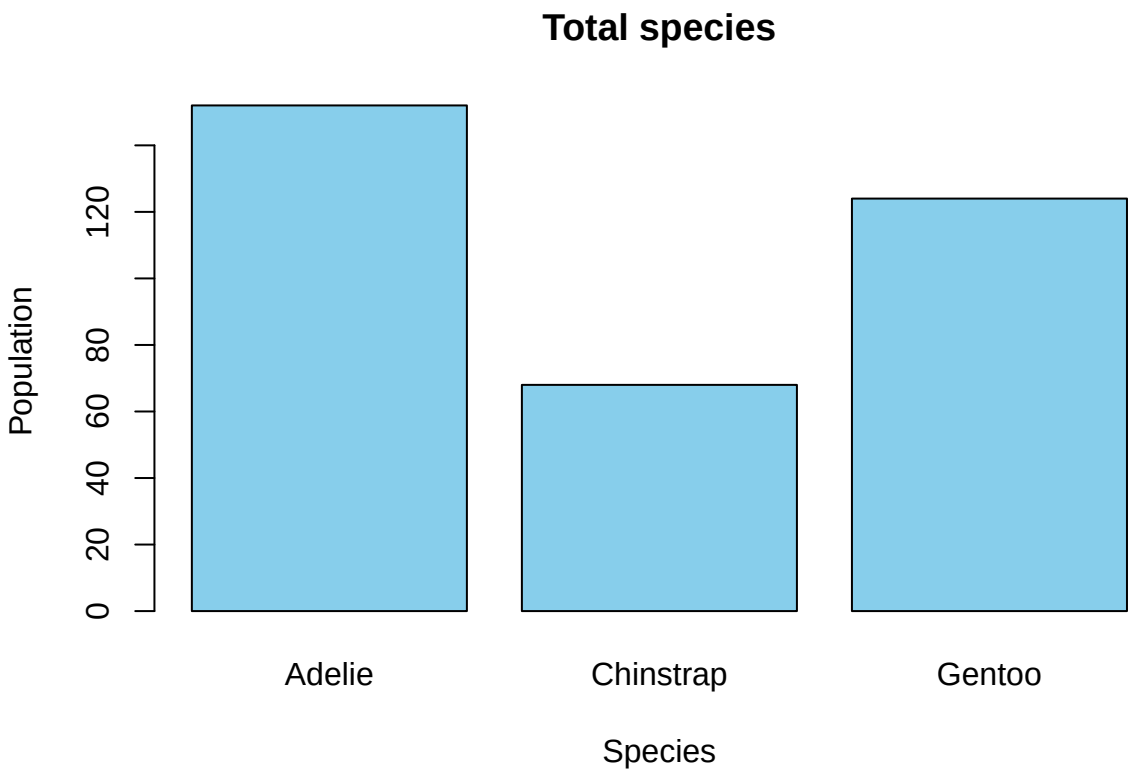


3. **Univariate Analysis – Categorical Variable:** Now consider a categorical variable (for example, **species**). How many penguins of each species are there in the dataset? Create a bar chart to show the count of each species. Which species is most common, and which is least? (The three species are Adelie, Gentoo, and Chinstrap.)

```r
summary(penguins$species)
```

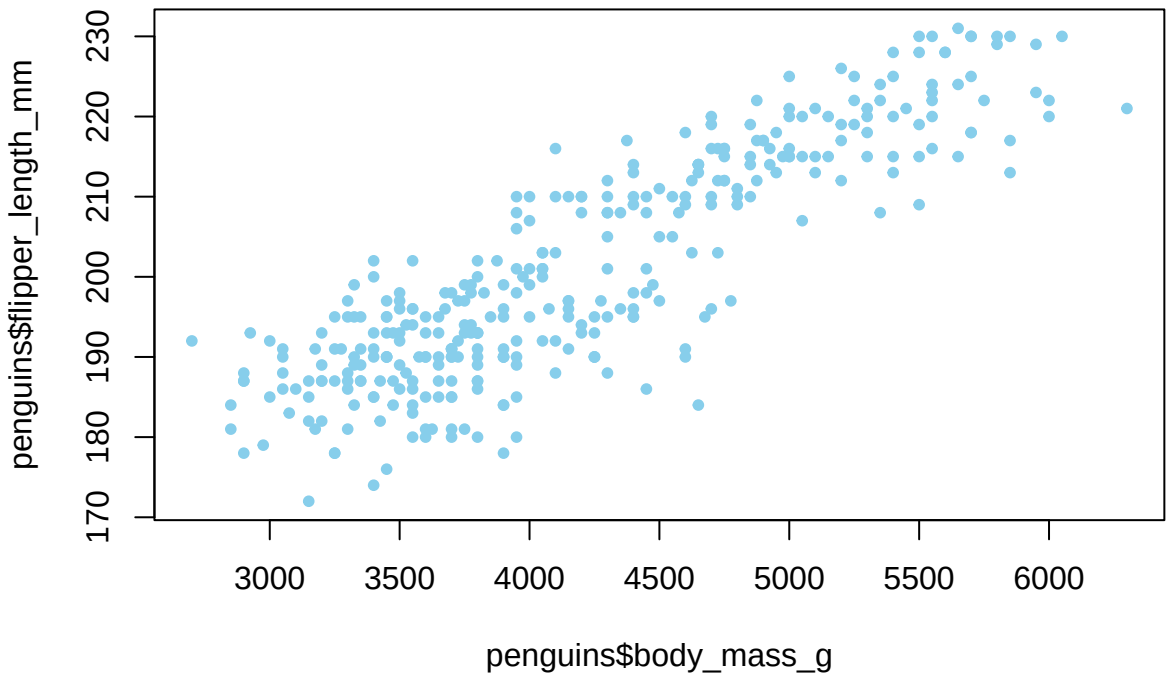
```
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

```r
plot(penguins$species,
     main='Total species',
     xlab='Species',
     ylab='Population',
     col='skyblue')
```



4. **Bivariate Analysis – Two Numeric Variables:** Investigate the relationship between **flipper_length_mm** and **body_mass_g**. Create a scatter plot with flipper length on one axis and body mass on the other. Do you observe a correlation (for instance, do heavier penguins tend to have longer flippers)? Calculate the correlation coefficient between these two variables to quantify the strength of any linear relationship.

```
plot(penguins$body_mass_g, penguins$flipper_length_mm,
     col = 'skyblue',
     pch = 20)
```



5. **Bivariate Analysis – Numeric vs Categorical:** Examine how one of the numeric measurements varies across species. For example, compare **body_mass_g** across the three species. Use a boxplot (or violin plot) to visualize body mass for each species side by side. What differences do you notice? Which species tends to be heaviest on average, and which tends to be lightest? Are the distributions for each species tightly clustered or quite spread out?

(Optional extension for extra practice: If you have time, you could also explore a **categorical vs categorical** relationship in this dataset. For instance, the dataset has an **island** variable (Biscoe, Dream, Torgersen islands). You might ask: Does species distribution differ by island? A contingency table of species by island, and perhaps a stacked bar chart, would reveal that each species is primarily found on certain islands. This kind of analysis can be insightful, but we'll focus on the tasks above for the core lab.)*

Take your time to code each step, and write down observations based on the output or plots. In the next section, we will discuss the solutions and what insights we gain from each part of the analysis.

```
sd(penguins$body_mass_g, na.rm=TRUE)
```

```
## [1] 801.9545
```