UNIVERSITI TEKNOLOGI PETRONAS

LAB WEEK 10

SEP 2025 SEMESTER

‘AFFAN NAJIY BIN RUSDI

22010453

BACHELOR OF COMPUTER SCIENCE

DATA SCIENCE

TEB2164

Contents

Code 3

Visualisation 5

Code

Activity 1

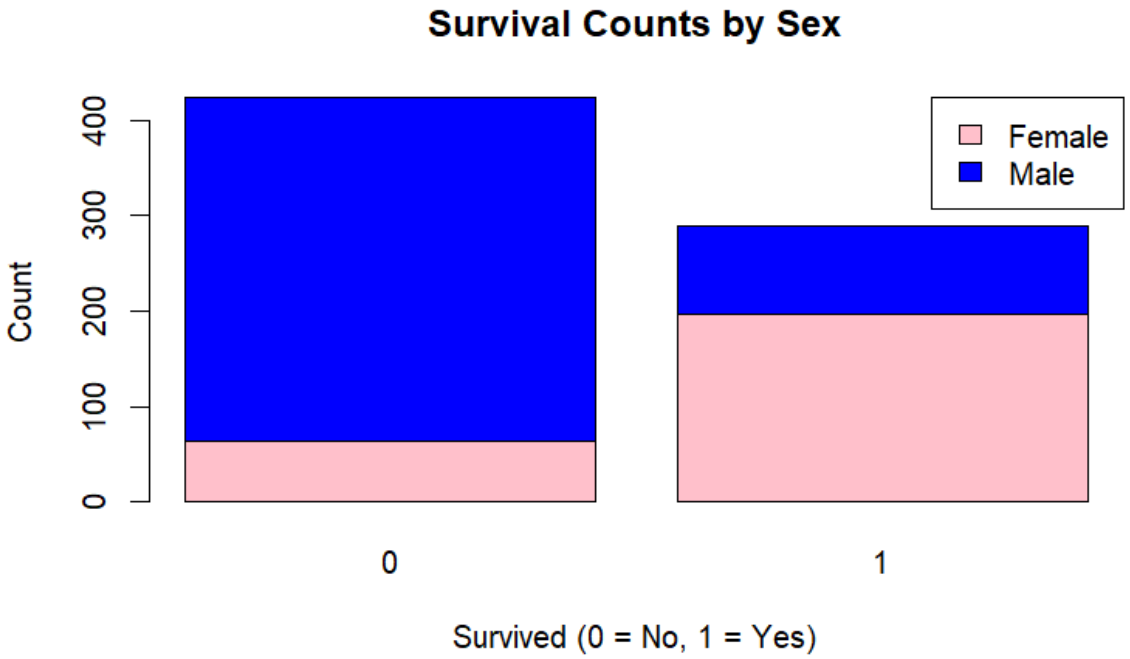
```
#Lab B Activity 1
setwd("C:/Users/AFFAN/Documents/GitHub/DS-academic/DS_Lab8-W10/Lab-B")
list.files()

#Libraries
library(dplyr)

#View Datasets
titanic <- read.csv("titanic.csv")
View(titanic)

cat("Missing Values: ", sum(is.na(titanic)))
which(is.na(titanic))
print(sapply(titanic, function(x) sum(is.na(x))))

#Manage Empty Values
dim(titanic)
titanic_cleaned = na.omit(titanic)
dim(titanic_cleaned)
View(titanic_cleaned)

titanic_sort = arrange(titanic_cleaned, survived)

colnames(titanic_sort)
View(titanic_sort)

#Bar Chart
survived_by_sex <- table(titanic_cleaned$sex, titanic_cleaned$survived)
#table() creates frequency tables that basically counts the data
barplot(survived_by_sex,
        xlab = "Survived (0 = No, 1 = Yes)",
        ylab = "Count",
        col = c("pink", "blue"),
        main = "Survival Counts by Sex",
        border = "black")
legend("topright", c("Female", "Male"), fill = c("pink", "blue"))

#Box Plot
boxplot(age ~ survived, data = titanic_cleaned,
        main = "Age Distribution by Survival",
        xlab = "Survived (0 = No, 1 = Yes)",
        ylab = "Age",
        col = c("red", "green"),
        border = "black")
```

Activity 2

```
#Lab B Activity 2
library(dplyr)
data("starwars")

print(head(starwars, 10))
View(starwars)
str(starwars)

#Summary
summary(starwars$height)
summary(starwars$mass)

#Count
table(starwars$species)
table(starwars$gender)

#Missing Values
cat("Missing Values: ", sum(is.na(starwars))) #105 missing
starwars_clean <- starwars[complete.cases(starwars[, c("height",
"mass")]), ]
#complete.cases handles missing value by identifying it by retuning values
with do not have N/A
View(starwars_clean)

#Bar Chart
species_count <- table(starwars_clean$species)
barplot(species_count,
        main = "Species Distribution",
        ylab = "Count",
        col = rainbow(length(species_count)),
        las = 2, #Rotate X values
        cex.names = 0.6 #shrink the values
)

#Box Plot
boxplot(height ~ gender, data = starwars_clean,
        main = "Height Distribution by Gender",
        xlab = "Gender",
        ylab = "Height(cm)",
        col = c("pink", "lightblue"),
        border = "black",
        notch = TRUE)
```

Visualisation

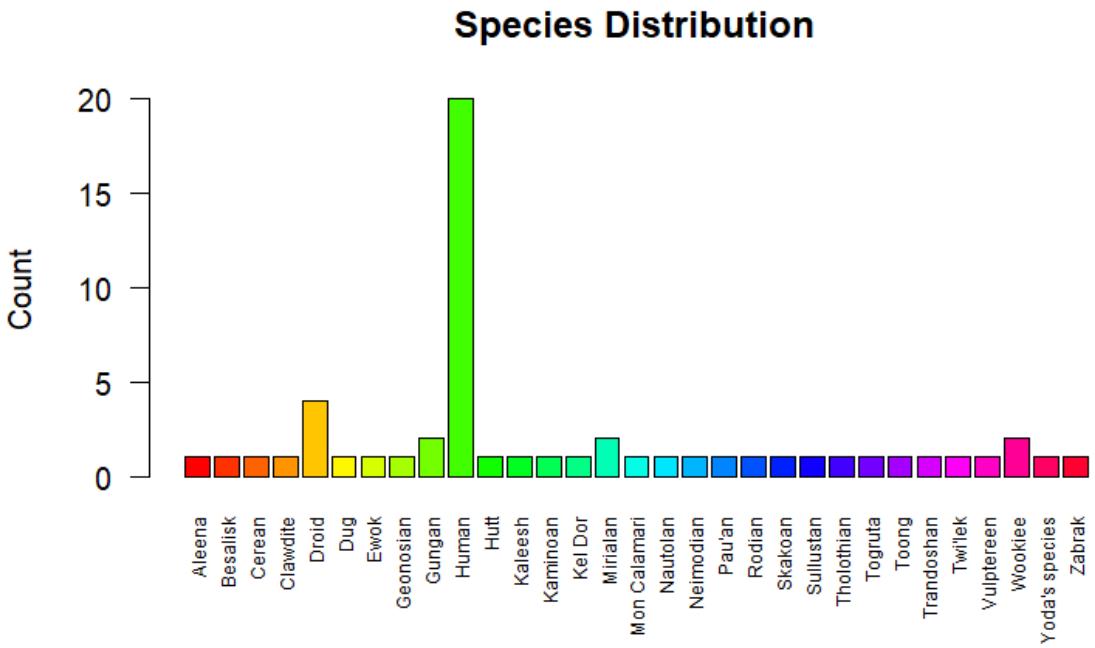
Activity 1



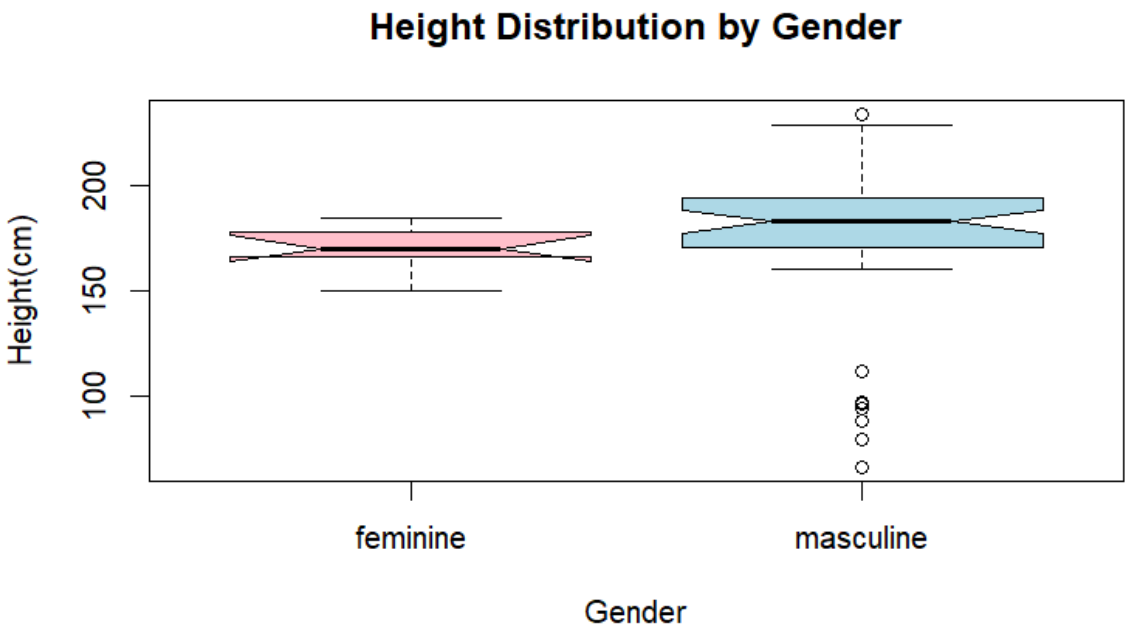
I used the Lab 7 cleaned data to produce this Bar Chart, in the beginning I tried to do some direct implementation into the chart syntax, but the genders (male, female) will the distributed continuously even though I just want the survival based on the numbers 0 and 1 which basically means survived and not survived respectively. So, the solution was to use the table() function which counted the data in the column I intended to use so I got it. The survival count of non survivors have a higher proportion for Males while higher proportions for the survivors are Female.

The age distribution for the non survivors have a similar range with the survivor's distribution. However, there is an outlier for one of the survivors to be around their 80s that survived the incident.

Activity 2



I have used the starwars built in database which showed that Human has the highest species count in the cinematic universe of Star Wars, the Droids came in second place, and Mon Calamari and Wookiees tie the spot in third place. Other species were mentioned in the universe but mainly have a count of 1 in the dataset.



The height distribution of all the fictional characters in Star Wars varies between their genders. The feminine have a lower height median as opposed to the masculine counterparts. However, it must be noted that some masculine characters have a very small height measurements as shown in the number of outliers which are less than 100 cm.