**Lab 1: BBH 505 Fall 2024**

***Introduction to R***

**In this lab, we will practice reading in data, viewing data, running some basic descriptive statistics, and making plots. Please work through each of the steps below, and provide a response when requested. Lines requesting a response will be underlined so you don’t miss them!**

**To submit your assignment, please upload this document with your responses along with your R syntax file.**

1. **If you haven’t already, install the following packages: summarytools, psych, ggplot2. Then load these using the library() function.** (1 point)

**A1.**

**# Installing the packages as instructed**

**install.packages("ggplot2")**

**library(ggplot2)**

**install.packages("psych")**

**library(psych)**

**install.packages("summarytools")**

**library(summarytools)**

1. **Using the setwd function, tell R what folder your data are in. Then use read.csv() to read your csv file containing your data into the program.** (2 points)

**A2. # Which folder my data is in**

**setwd("C:/Users/Aaliya/Desktop/BBH597")**

**# Get the current working directory to verify it's set correctly**

**getwd()**

**# Read the CSV file**

**lab1\_dat <- read.csv("data\_R1.csv")**

1. **Ask for (a) the variable names and (b) the data structure. How many variables and observations do you have**? (3 points)

**A3.**

**#Display variable names(cols)**

**names(lab1\_dat)**

**#Examine the data strcuture**

**str(lab1\_dat)**

**#Number of variables(cols) and observations(rows)**

**ncol(lab1\_dat)**

**nrow(lab1\_dat)**

**I have 3 variables and 10 observations.**

1. **Print the first 6 rows of your data, the last 6 rows of your data, then all of your data.** (3 points)

**A4. #To print the first 6 rows**

**head(lab1\_dat , n=6)**

**#To print the last 6 rows**

**tail(lab1\_dat , n=6)**

**#To print all the data**

**print(lab1\_dat)**

1. **Using the freq() command, ask for frequencies for age and height. (You can do this in two separate freq() commands or in one single freq() command). Which height value is the most frequent? How many observations have this value, and what proportion of the height variable values does it represent?** (5 points)

**A5.**

**freq(lab1\_dat$age)**

**freq(lab1\_dat$height)**

**The output of the command “freq(lab1\_dat$height)” displays frequency of occurence, percentages of each value, cumulative frequencies, and cumulative percentages.**

**Looking at the output we can say that the 45 height value is the most frequent. This value has 3 obervations and the proportion it represents is 30%.**

1. **Using an ifelse() command, recode height into groups. Code everyone shorter than 47 inches as 1, and everyone 47 inches or taller as 2. Call this variable height.grp. Then run a crosstab of between height.grp and height using ctable(). Paste this crosstab below.** (2 points)

**A6. lab1\_dat$height.grp <- ifelse(lab1\_dat$height < 47, 1, 2)**

**print(lab1\_dat)**

**print(lab1\_dat$height.grp)**

**#summary(lab1\_dat$height)**

**#summary(lab1\_dat$height.grp)**

**lab1\_dat$height.grp <- factor(lab1\_dat$height.grp)**

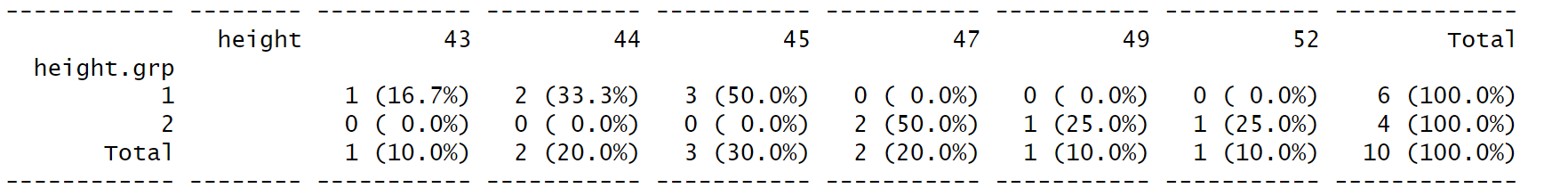
**lab1\_dat$height <- factor(lab1\_dat$height)**

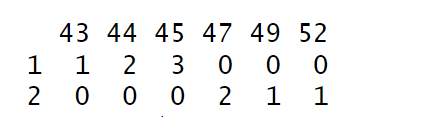
**output<-ctable(lab1\_dat$height.grp , lab1\_dat$height)**

**print(ctable)**

**crosstab <- table(lab1\_dat$height.grp, lab1\_dat$height)**

**print(crosstab)**





1. **Create 4 plots:** **1 bar plot of your height group variable, a histogram of the height variable, a histogram of the height variable with a density curve, and a line plot of the height variable.** **Paste each of these plots below.** (4 points)

**A7. # Bar Plot**

**ggplot(lab1\_dat, aes(x = height.grp)) +**

**geom\_bar(fill = "pink", color = "black") +**

**labs(x = "Height Group", y = "Count") +**

**theme\_bw()**

**# Histogram of height**

**hist(table(lab1\_dat$height),**

**main = "Histogram of Height",**

**xlab = "Height",**

**ylab = "Frequency",**

**col = "green",**

**breaks = 10)**

**ggplot(lab1\_dat, aes(x = as.numeric(as.character(height)))) +**

**geom\_histogram(fill = "lightgreen", color = "black", binwidth = 2.5) +**

**labs(x = "Height (inches)", y = "Count") +**

**theme\_bw()**

**#adding density curve to the height variable plot**

**ggplot(lab1\_dat, aes(x = as.numeric(as.character(height)))) +**

**geom\_histogram(aes(y = ..density..), fill = "lightgreen", color = "black", binwidth = 2.5) +**

**geom\_density(color = "blue", size = 1) +**

**labs(x = "Height (inches)", y = "Density") +**

**theme\_bw()**

**# Line Plot**

**ggplot(lab1\_dat, aes(x = seq\_along(height), y = as.numeric(as.character(height)))) +**

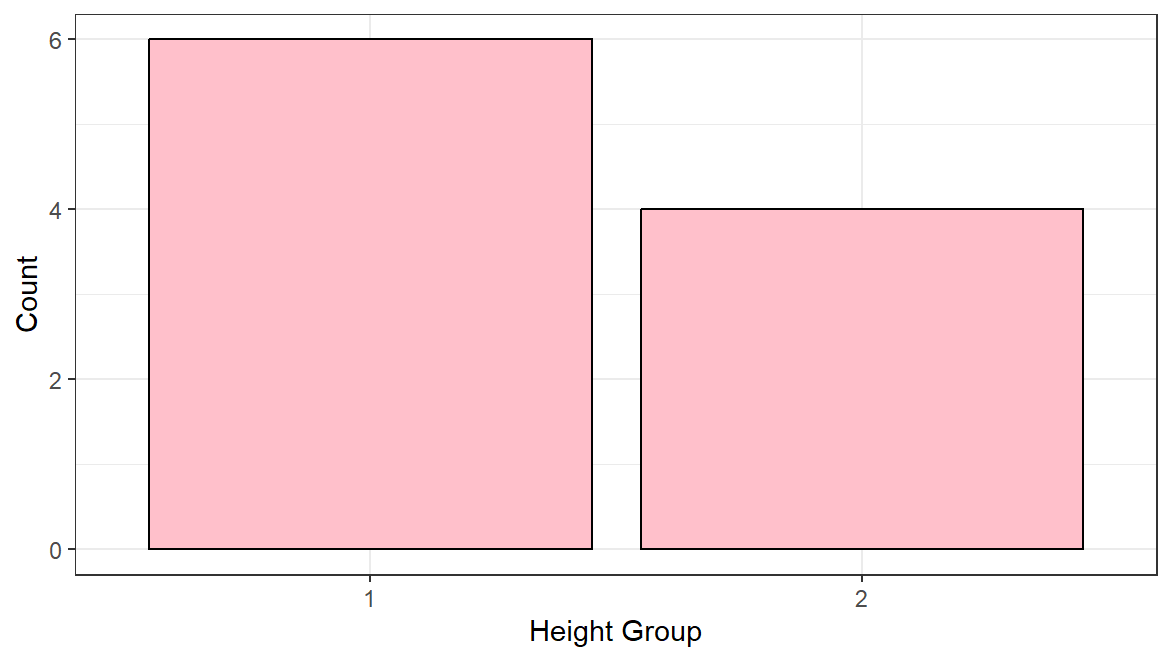
**geom\_line(color = "darkred", size = 1) +**

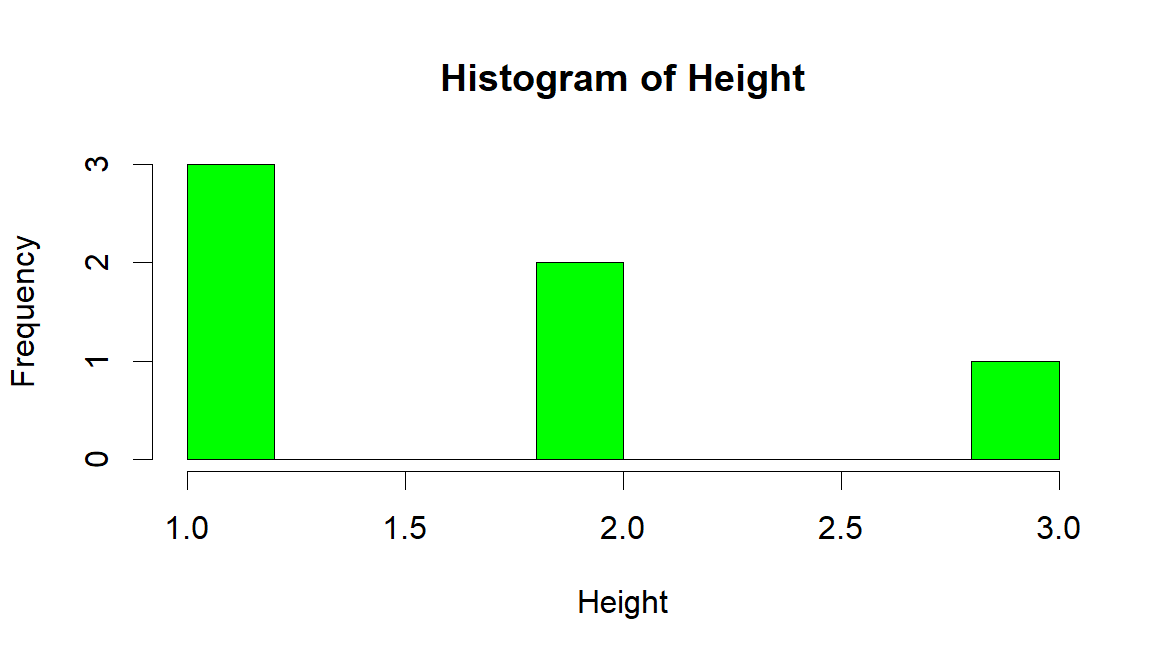
**labs(x = "Index", y = "Height (inches)") +**

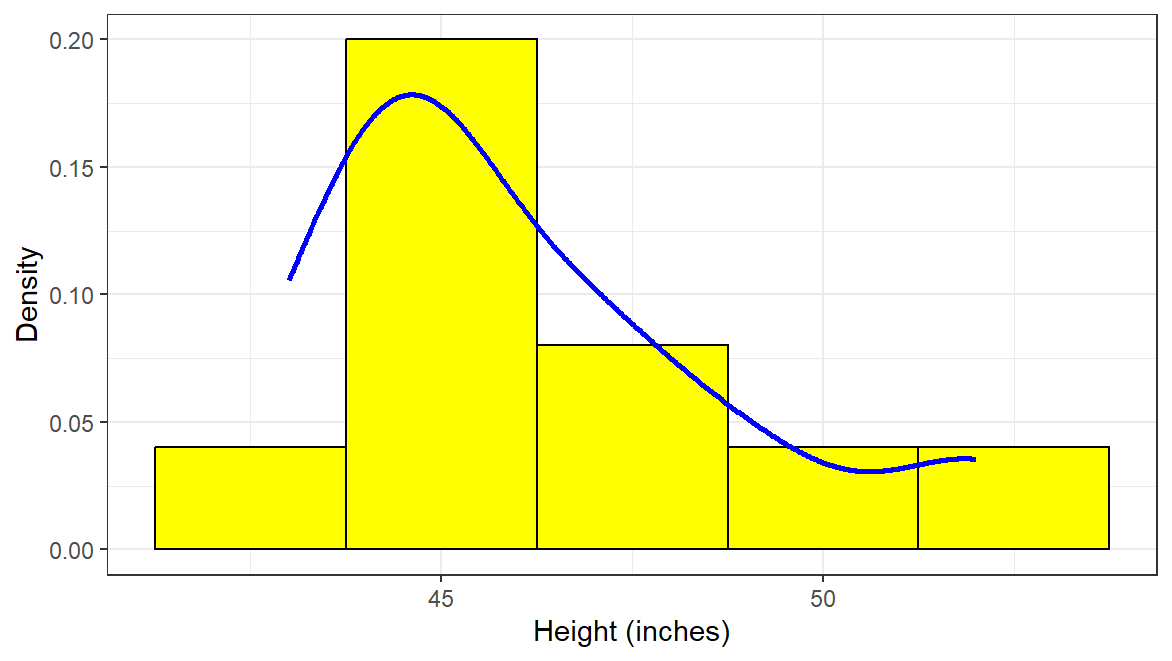
**theme\_bw()**

Images of the plots:

1. Barplot



1. Histogram of height variable
2. Histogram of height variable with density curve



1. Line plot

