Homework 2

Ques 1: Read in the bodyfat.csv data file and generate a variable "bodycat" to categorize body fat into the three categories above. Make sure all 252 observations are categorized into either athlete, average, or obese.

Ans:

```
> library(readxl)
> bodyfat <- read_excel("C:/Users/Kavit/Downloads/bodyfat.xlsx")
> View(bodyfat)
> summary(bodyfat)
```

```
       id            bodyfat         density           age             weight          height        adiposity    
 Min.   :  1.00   Min.   : 0.00   Min.   :0.995   Min.   :22.00   Min.   :118.5   Min.   :29.50   Min.   :18.10  
 1st Qu.: 63.75   1st Qu.:12.80   1st Qu.:1.041   1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25   1st Qu.:23.10  
 Median :126.50   Median :19.00   Median :1.055   Median :43.00   Median :176.5   Median :70.00   Median :25.05  
 Mean   :126.50   Mean   :18.94   Mean   :1.056   Mean   :44.88   Mean   :178.9   Mean   :70.15   Mean   :25.44  
 3rd Qu.:189.25   3rd Qu.:24.60   3rd Qu.:1.070   3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:27.32  
 Max.   :252.00   Max.   :45.10   Max.   :1.109   Max.   :81.00   Max.   :363.1   Max.   :77.75   Max.   :48.90  
      neck            chest           abdomen             hip             thigh            knee            ankle      
 Min.   :31.10   Min.   : 79.30   Min.   : 69.40   Min.   : 85.0   Min.   :47.20   Min.   :33.00   Min.   :19.1  
 1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0  
 Median :38.00   Median : 99.65   Median : 90.95   Median : 99.3   Median :59.00   Median :38.50   Median :22.8  
 Mean   :37.99   Mean   :100.82   Mean   : 92.56   Mean   : 99.9   Mean   :59.41   Mean   :38.59   Mean   :23.1  
 3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0  
 Max.   :51.20   Max.   :136.20   Max.   :148.10   Max.   :147.7   Max.   :87.30   Max.   :49.10   Max.   :33.9  
     bicep           forearm          wrist      
 Min.   :24.80   Min.   :21.00   Min.   :15.80  
 1st Qu.:30.20   1st Qu.:27.30   1st Qu.:17.60  
 Median :32.05   Median :28.70   Median :18.30  
 Mean   :32.27   Mean   :28.66   Mean   :18.23  
 3rd Qu.:34.33   3rd Qu.:30.00   3rd Qu.:18.80  
 Max.   :45.00   Max.   :34.90   Max.   :21.40  
```

```
> bodyfat$bodycat <- with(bodyfat,
+                         ifelse(bodyfat <= 14, "athlete",
+                                ifelse(bodyfat <= 25, "average",
"obese"))
+ )
> table(bodyfat$bodycat)

athlete average   obese 
     73     120      59 
```

Ques 2: Using summarize to identify the four height quartiles, create a new variable "htcat" to categorize height into "short", "below average", "above average", and "tall".

Ans:

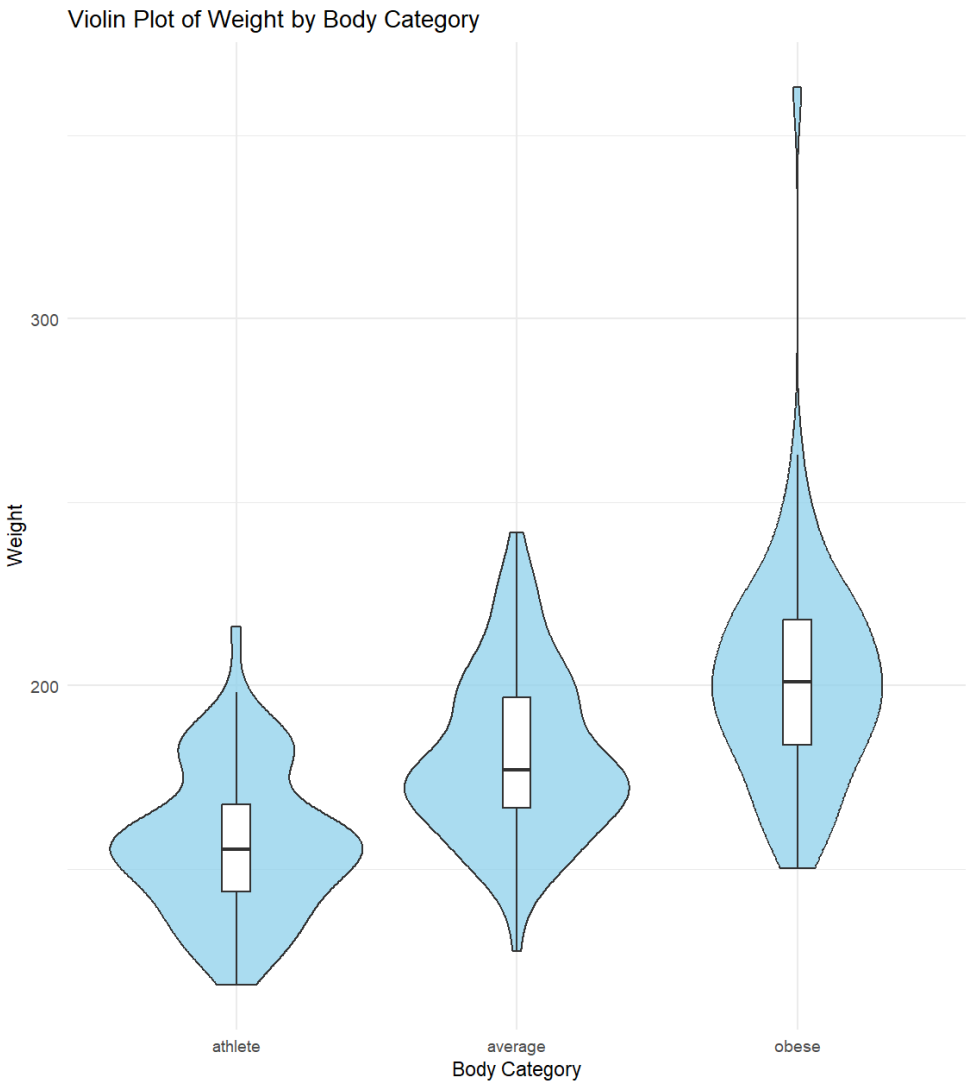
```
> library(dplyr)
> quartiles <- bodyfat %>%
+  summarize(Q1 = quantile(height, 0.25),
+            Q2 = quantile(height, 0.50),
+            Q3 = quantile(height, 0.75))
> bodyfat <- bodyfat %>%
+  mutate(htcat = case_when(
+    height <= quartiles$Q1 ~ "short",
+    height > quartiles$Q1 & height <= quartiles$Q2 ~ "below average",
+    height > quartiles$Q2 & height <= quartiles$Q3 ~ "above average",
+    height > quartiles$Q3 ~ "tall"
+  ))
> head(bodyfat)

> table(bodyfat$htcat)

 above average       below average          short           tall
            70                  62             65             55
```

Ques 3: Create a violin plot of weight separated by bodycat. Make sure your plots show up in some kind of order that makes sense. In complete sentences, summarize what the violin plots tell you. Are the weights evenly distributed within a range for all categories? Do athletes tend to be within a certain weight range? You may use summarize() to help you. Rough estimates are also okay.

Ans:

```
> library(ggplot2)
> library(dplyr)
> weight_summary <- bodyfat %>%
+  group_by(bodycat) %>%
+  summarize(
+    Min_Weight = min(weight),
+    Q1_Weight = quantile(weight, 0.25),
+    Median_Weight = median(weight),
+    Q3_Weight = quantile(weight, 0.75),
+    Max_Weight = max(weight),
```

```
+    Mean_Weight = mean(weight)
+  )
> print(weight_summary)

# A tibble: 3 × 7
  bodycat Min_Weight Q1_Weight Median_Weight Q3_Weight Max_Weight Mean_Weight
  <chr>        <dbl>     <dbl>          <dbl>     <dbl>      <dbl>       <dbl>
1 athlete       118.      144.           155.      168.       216         157.
2 average       128.      167.           177       197.       242.        181.
3 obese         150.      184.           201       218.       363.        203.

> ggplot(bodyfat, aes(x = factor(bodycat, levels = c("athlete", "average",
"obese")), y = weight)) +
+  geom_violin(fill = "skyblue", alpha = 0.7) +
+  geom_boxplot(width = 0.1, fill = "white", outlier.shape = NA) + # Overlay
boxplot for better insights
+  labs(
+    title = "Violin Plot of Weight by Body Category",
+    x = "Body Category",
+    y = "Weight"
+  ) +
+  theme_minimal()
```



Observations:

- Weights are not evenly distributed within each category.
- Athletes tend to fall within a lower weight range.
- Obese individuals have the largest variation in weight, while average individuals fall in the middle.

Ques 4: Create a stem-and-leaf plot for weight. Be sure to find an appropriate scale for the data.

Ans:

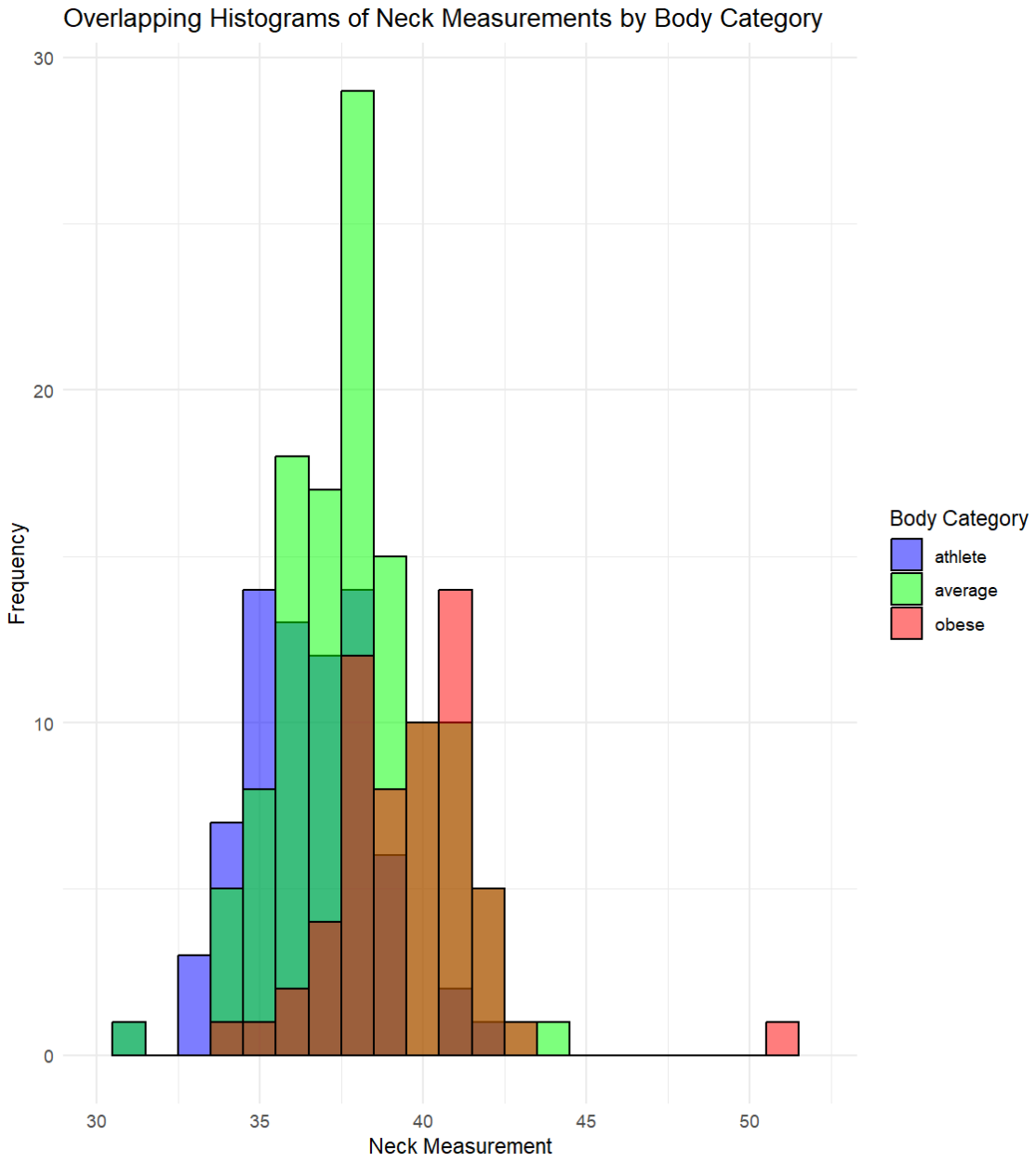
```
> stem(bodyfat$weight, scale = 2)

  The decimal point is 1 digit(s) to the right of the |

  11 | 9

  12 | 556788

  13 | 234466779

  14 | 0012345667888999

  15 | 00111222233344555556667777888899

  16 | 000001111122333334455666777788888888899

  17 | 011111122333334556666777777788888999

  18 | 000011222334444444555667788889

  19 | 00111222334455667777 88999

  20 | 0001111223335566778999

  21 | 0122366667799999

  22 | 335555788

  23 | 0345

  24 | 1247

  25 |

  26 | 3

  27 |

  28 |

  29 |

  30 |

  31 |
```

```
  32 |

  33 |

  34 |

  35 |

  36 | 3
```

Ques 5 A: Create overlapping histograms of neck for the three body categories. For this exercise, do not use the default breaks. Use breaks that you think make sense. Remember to make sure that the first histogram is an appropriate window size so when you "add" the other graphs, those histograms aren't cut-off. Also, remember to use the same break widths for overlapped histograms.

Ans:

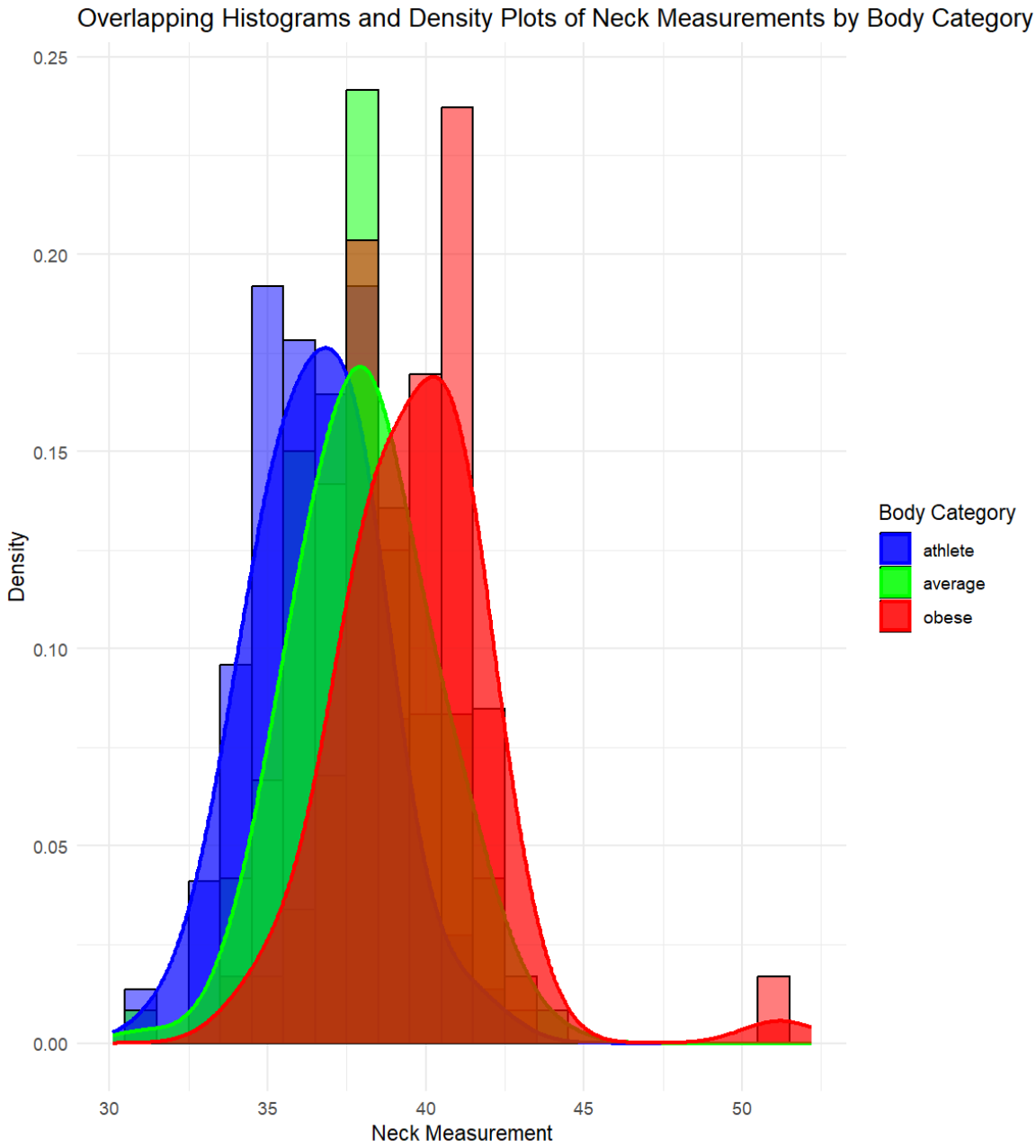
```
> ggplot(bodyfat, aes(x = neck, fill = bodycat)) +
+  geom_histogram(binwidth = 1, alpha = 0.5, position = "identity", color =
"black") +
+  scale_fill_manual(values = c("athlete" = "blue", "average" = "green",
"obese" = "red")) +
+  labs(
+    title = "Overlapping Histograms of Neck Measurements by Body Category",
+    x = "Neck Measurement",
+    y = "Frequency",
+    fill = "Body Category"
+  ) +
+  theme_minimal() +
+  xlim(min(bodyfat$neck) - 1, max(bodyfat$neck) + 1)
```


Overlapping Histograms of Neck Measurements by Body Category
Frequency
30
20
10
0
30
35
40
45
50
Neck Measurement
Body Category
athlete
average
obese

Ques 5 B: In the same window, add 3 density plots—1 for each body category. Do not use the default bandwidth. Use a bandwidth that you think makes sense. (Note, you will need to have used freq = F in your histograms.)

Ans:

```
> ggplot(bodyfat, aes(x = neck, fill = bodycat)) +
+  geom_histogram(aes(y = ..density..), binwidth = 1, alpha = 0.5, position =
"identity", color = "black") +
+  # Add density plots with custom bandwidth
+  geom_density(aes(color = bodycat), alpha = 0.7, size = 1, adjust = 1.5) +
+  scale_fill_manual(values = c("athlete" = "blue", "average" = "green",
"obese" = "red")) +
+  scale_color_manual(values = c("athlete" = "blue", "average" = "green",
"obese" = "red")) +
+  labs(
+    title = "Overlapping Histograms and Density Plots of Neck Measurements by
Body Category",
+    x = "Neck Measurement",
+    y = "Density",
+    fill = "Body Category",
+    color = "Body Category"
+  ) +
+  theme_minimal() +
+  xlim(min(bodyfat$neck) - 1, max(bodyfat$neck) + 1)
```



Ques 5 C: In complete sentences, compare neck circumference across the three body categories using your histograms and density plots.

Ans:

Athletes (Blue):

- The histogram shows that neck circumferences for athletes are tightly clustered within a narrow range, primarily between 35 cm and 40 cm.
- The density plot has a sharp peak, indicating that most athletes have neck measurements concentrated around 37-38 cm. This reflects the lean and consistent physique typical of athletes.

Average (Green):

- The histogram for the "average" category is wider than the athlete group, with neck circumferences ranging from approximately 36 cm to 42 cm.
- The density plot has a broader peak compared to the athlete category, showing a more evenly distributed range. The majority of average individuals have neck measurements between 37 cm and 41 cm.

Obese (Red):

- The histogram for the obese category spans the widest range, with neck circumferences extending from 38 cm to over 50 cm (outlier).
- The density plot shows a flatter, broader peak, indicating significant variability in neck circumferences. The majority of obese individuals have neck measurements between 40 cm and 45 cm, with a visible tail toward larger measurements.

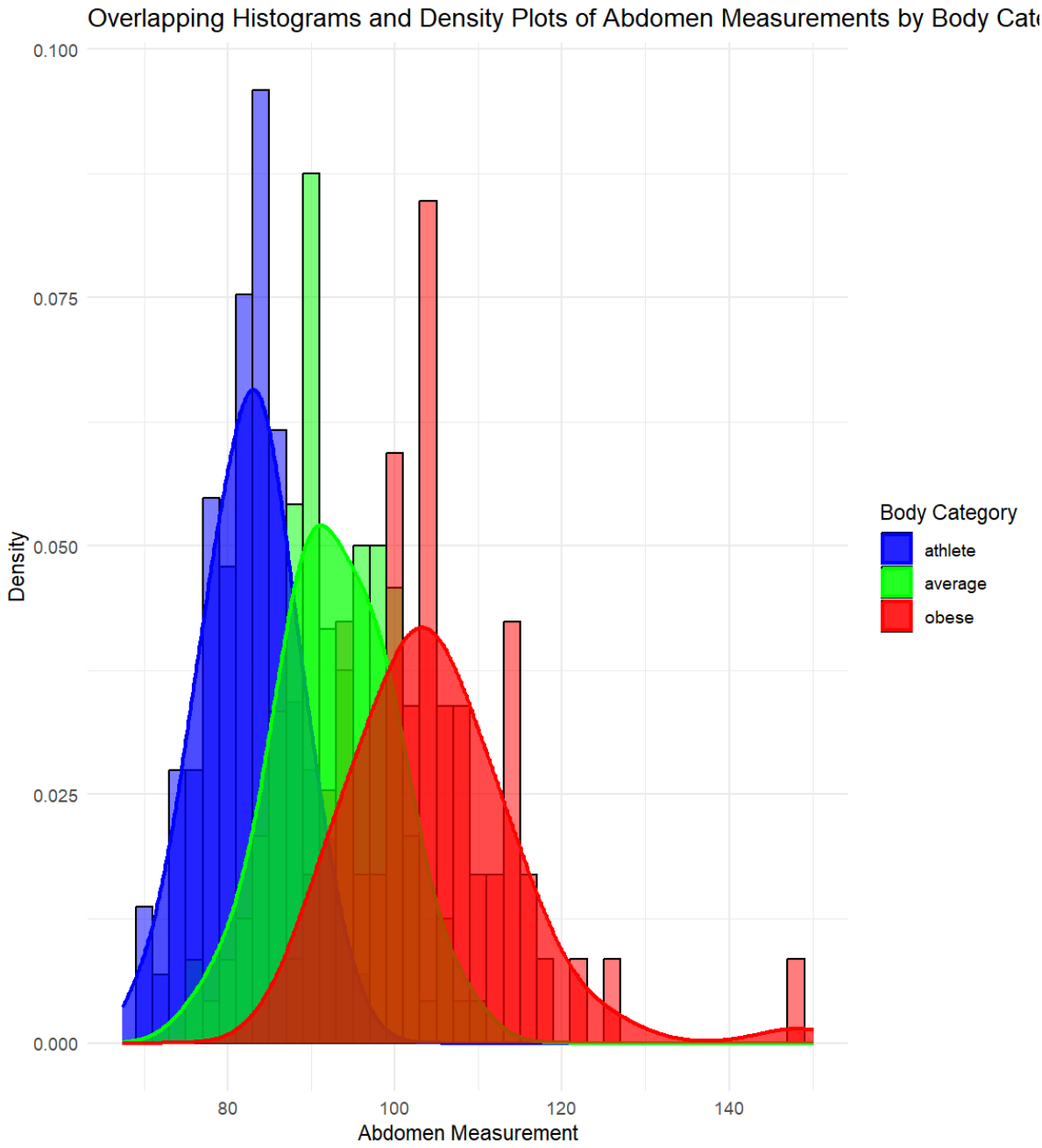
Ques 6 A: Create overlapping histograms of abdomen for the three body categories. For this exercise, do not use the default breaks. Use breaks that you think make sense. Remember to make sure that the first histogram is an appropriate window size so when you "add" the other graphs, those histograms aren't cut-off. Also, remember to use the same break widths for overlapped histograms.

B: In the same window, add 3 density plots—1 for each body category. Do not use the default bandwidth. Use a bandwidth that you think makes sense. (Note, you will need to have used freq = F in your histograms.)

Ans:

```
> bodyfat_data_clean <- bodyfat %>%
+   filter(!is.na(abdomen) & is.finite(abdomen))
>
> # Ensure 'bodycat' is a factor with the correct levels
> bodyfat_data_clean$bodycat <- factor(bodyfat_data_clean$bodycat,
```

```
+                                    levels = c("athlete", "average",
"obese"))
>
> # Create overlapping histograms and density plots
> ggplot(bodyfat_data_clean, aes(x = abdomen, fill = bodycat)) +
+  # Overlapping histograms with normalized frequencies
+  geom_histogram(aes(y = ..density..), binwidth = 2, alpha = 0.5, position =
"identity", color = "black") +
+  # Add density plots with custom bandwidth
+  geom_density(aes(color = bodycat), alpha = 0.7, size = 1, adjust = 1.5) +
+  scale_fill_manual(values = c("athlete" = "blue", "average" = "green",
"obese" = "red")) +
+  scale_color_manual(values = c("athlete" = "blue", "average" = "green",
"obese" = "red")) +
+  labs(
+    title = "Overlapping Histograms and Density Plots of Abdomen Measurements
by Body Category",
+    x = "Abdomen Measurement",
+    y = "Density",
+    fill = "Body Category",
+    color = "Body Category"
+  ) +
+  theme_minimal() +
+  xlim(min(bodyfat_data_clean$abdomen) - 2, max(bodyfat_data_clean$abdomen) +
2)
```



C. In complete sentences, compare abdomen circumference across the three body categories using your histograms and density plots.

Ans:

Athletes (Blue):

- The abdomen circumferences for athletes are tightly concentrated, mostly within the range of 70 to 90 cm.
- The density plot shows a sharp and narrow peak around 80 cm, indicating that the majority of athletes have consistent abdomen measurements.
- This narrow range reflects the leaner body composition typical of athletes, with minimal variation.

Average (Green):

- The "average" category displays a broader range of abdomen circumferences, spanning approximately 85 to 110 cm.
- The density plot shows a wider peak centered around 95 cm, suggesting more variability in abdomen size compared to athletes.
- This broader distribution reflects the diverse body types encompassed in the "average" category.

Obese (Red):

- The "obese" category exhibits the largest range of abdomen circumferences, stretching from 100 cm to over 140 cm (outlier).
- The density plot for obese individuals has a flatter and wider shape, with the majority of measurements falling between 110 and 120 cm, but with some measurements extending to higher values.
- This large range indicates substantial variability in body size within the obese group.

==========