Assessment v1

Ahmed SOROUR

2024-02-10

The assessment involves analyzing a dataset from two fictional care homes, focusing on physical activity, demographics, and other health-related metrics of the participants. The key tasks for your report include:

Demographics and Baseline Characteristics: Analyze and present demographic and baseline characteristics of individuals in the trial. This includes creating a table, at least one graph, and interpreting any statistical tests conducted.

Association Between Care Home and Physical Activity: Investigate if there's an association between the care home and the type of physical activity undertaken, and whether the amount of moderate activity varies by care home. Explain your analysis choice and interpret the results.

Physical Activity and Longevity: Assess if more physically active individuals live longer and provide a formal recommendation on whether an intervention should be implemented based on these results.

BMI and Moderate Activity: Explore the hypothesis that BMI decreases as the proportion of moderate activity increases. Use visualization and linear regression to test this hypothesis and quantify the relationship.

Load data

Let's proceed by analyzing the provided dataset in R and addressing these questions one by one. I'll start by loading the dataset and conducting preliminary data exploration.

```
pacman::p_load(data.table, rio, here, dplyr, epikit, janitor, lubridate, ggplot2, crosstable, st
ringr, gtsummary, flextable, Hmisc, scales, incidence, tidyverse, kableExtra, knitr)

carehome_data <- import(here("Assessment/Report/carehomedata_assessment2024.csv"))
carehome_data <- carehome_data %>% clean_names()
```

The dataset has been successfully loaded and looked into. In order to go into the questions, the first step is to check the demographic characteristics and also the baseline characteristics of our data.

1. Demographics and Baseline Characteristics

Participant ID: Ranges from 1 to 341, indicating 341 participants in the study. Age at Recording: The average age at recording is approximately 79.85 years, with a standard deviation of about 5.18 years. The age ranges from 66.53 to 93.49 years. Age at Death: The average age at death is around 81.52 years, with a standard deviation of 5.50 years, ranging from 67.39 to 96.87 years. Moderate Activity: On average, participants engaged in moderate activity about 49.73% of the time, with a standard deviation of 9.93%. BMI: The average BMI of participants is 24.92, with a standard deviation of 1.80, ranging from 19.27 to 29.81. Care Home ID: Participants are almost evenly distributed between the two care homes (1 and 2).

and here is the code, and a table to

```
# Descriptive statistics for numerical variables using standardized column names ----
descriptive_stats <- carehome_data %>%
  summarise(
    min_participant_id = min(participant_id),
    max_participant_id = max(participant_id),
    mean_age_at_recording = mean(age_at_recording, na.rm = TRUE),
    sd_age_at_recording = sd(age_at_recording, na.rm = TRUE),
    min_age_at_recording = min(age_at_recording, na.rm = TRUE),
    max_age_at_recording = max(age_at_recording, na.rm = TRUE),
    mean_age_at_death = mean(age_at_death, na.rm = TRUE),
    sd_age_at_death = sd(age_at_death, na.rm = TRUE),
    min_age_at_death = min(age_at_death, na.rm = TRUE),
    max_age_at_death = max(age_at_death, na.rm = TRUE),
    mean_moderate_activity = mean(moderate_activity, na.rm = TRUE),
    sd_moderate_activity = sd(moderate_activity, na.rm = TRUE),
    mean_bmi = mean(bmi, na.rm = TRUE),
    sd_bmi = sd(bmi, na.rm = TRUE),
    min_bmi = min(bmi, na.rm = TRUE),
    max_bmi = max(bmi, na.rm = TRUE)
  )

# Convert descriptive_stats to a more table-friendly format
# Here, we're pivoting the dataframe to have a variable and value format
descriptive_stats_long <- descriptive_stats %>%
  pivot_longer(cols = everything(), names_to = "Statistic", values_to = "Value")

# Create a table using kable and add styling with kableExtra
kable(descriptive_stats_long, format = "html", col.names = c("Statistic", "Value")) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive")) %>%
  column_spec(1, bold = TRUE) %>%
  row_spec(0, bold = TRUE, background = "#D3D3D3")
```

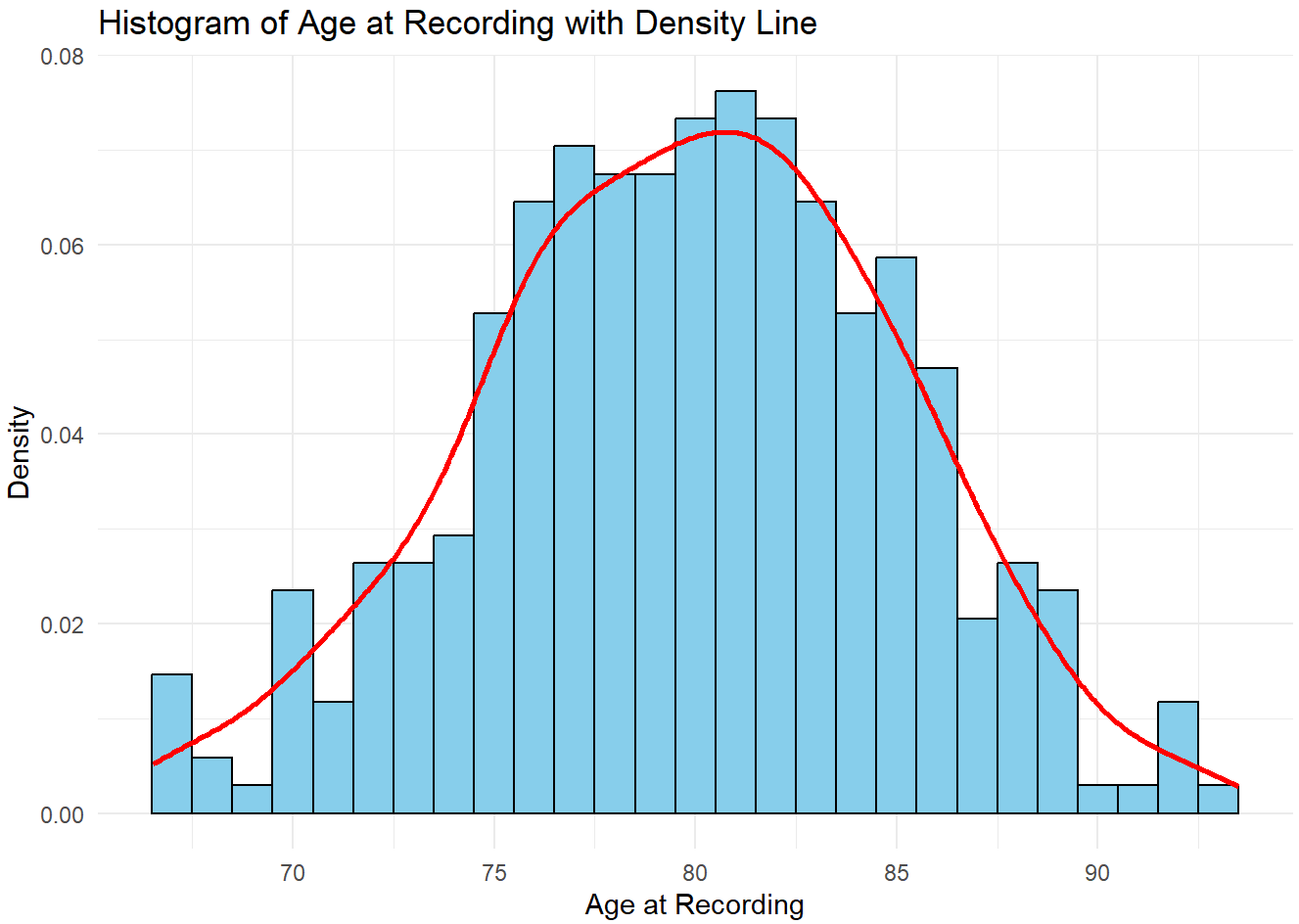
Statistic	Value
min_participant_id	1.000000
max_participant_id	341.000000
mean_age_at_recording	79.846328
sd_age_at_recording	5.176227
min_age_at_recording	66.532592
max_age_at_recording	93.491975
mean_age_at_death	81.516802
sd_age_at_death	5.498229
min_age_at_death	67.393800
max_age_at_death	96.871826
mean_moderate_activity	49.732026

Statistic	Value
sd_moderate_activity	9.928704
mean_bmi	24.924828
sd_bmi	1.802885
min_bmi	19.273478
max_bmi	29.809399

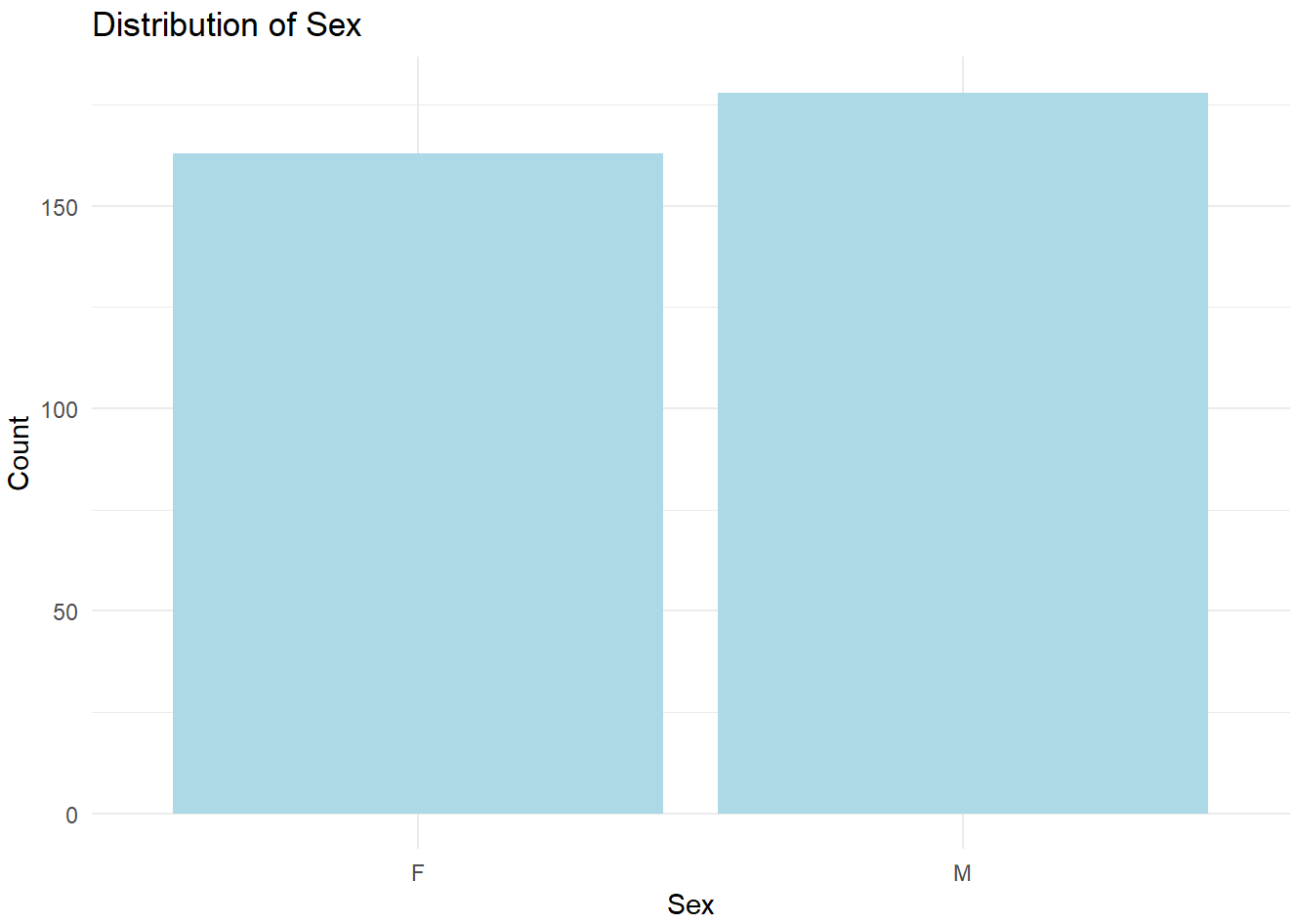
Visualizaiton of the demographics and baseline characteristics

- For the participants age, in numerical terms, one might say, “The majority of our participants are clustered around their late 70s to early 80s, with fewer individuals below 70 or above 90. This suggests our study predominantly involves older adults, with a peak concentration of ages around the late 70s.”
- Age distribution is almost equal with a slight higher female portion of the populaiton.
- Physical activity categories, are close to each other a little above 75 person in each of the four categories of; light, low, moderate, and sitting.
- As mentioned above the histogram here shows that participants engaged in moderate activity about 49.73% of the time, with a standard deviation of 9.93%.

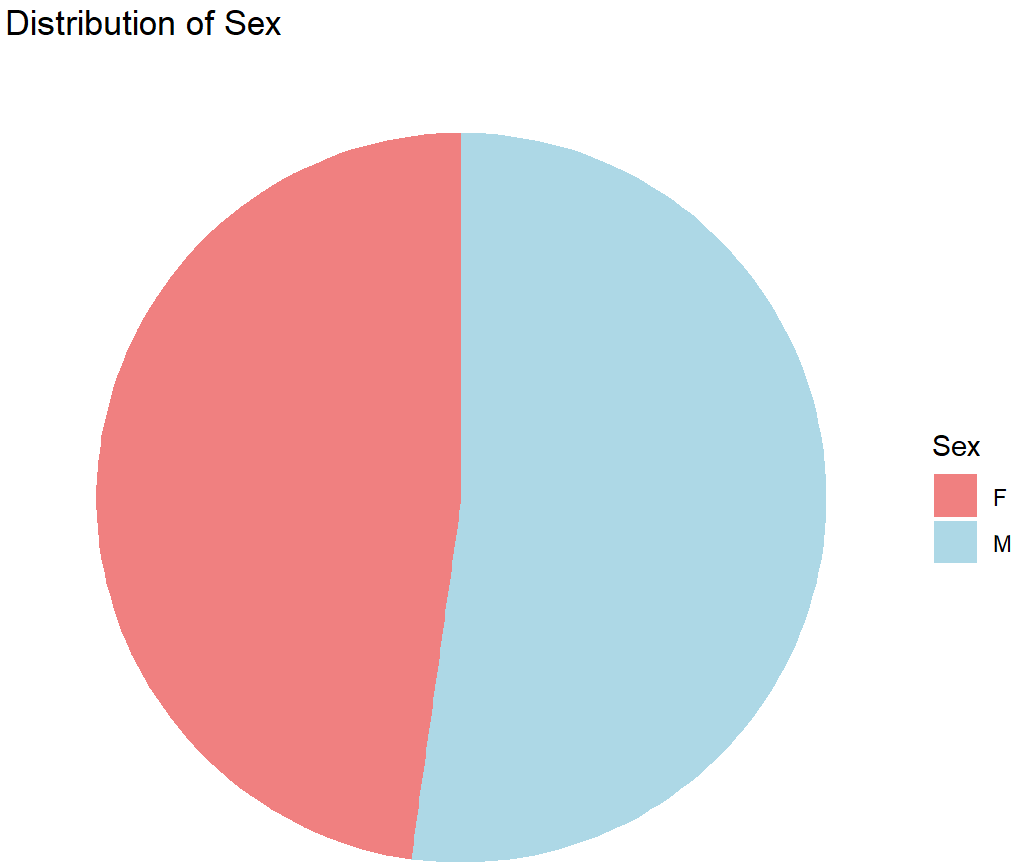
```
# # Histogram for Age at Recording
# ggplot(carehome_data, aes(x = age_at_recording)) +
#   geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
#   theme_minimal() +
#   labs(title = "Histogram of Age at Recording", x = "Age at Recording", y = "Count")

# Histogram for Age at Recording with Density Line using updated syntax
ggplot(carehome_data, aes(x = age_at_recording)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "skyblue", color = "black")
+  # Updated to use after_stat(density)
  geom_density(color = "red", size = 1) +  # Add the density line in red
  theme_minimal() +
  labs(title = "Histogram of Age at Recording with Density Line", x = "Age at Recording", y = "D
ensity")
```



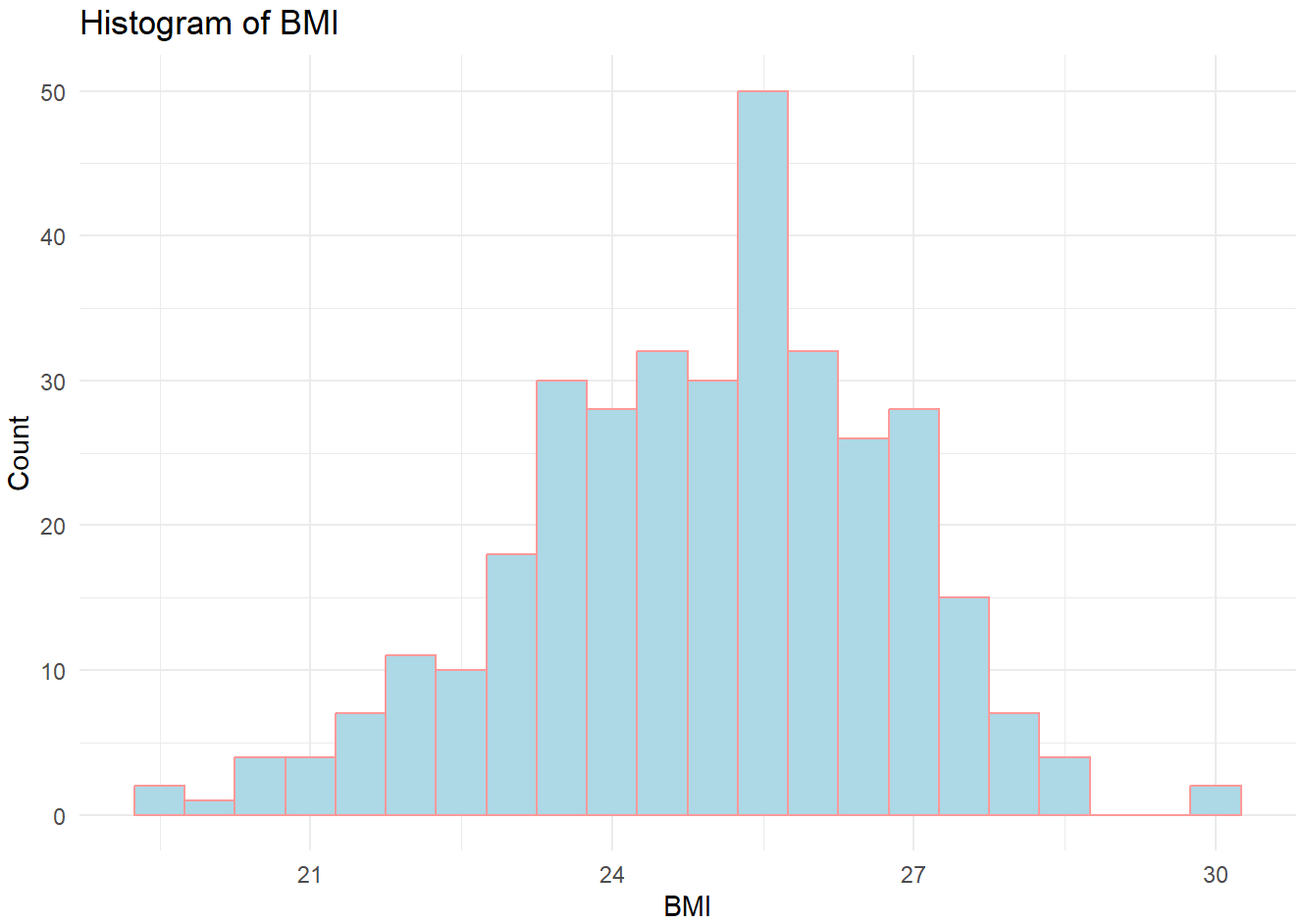

```
# Bar Chart for Sex
ggplot(carehome_data, aes(x = sex)) +
  geom_bar(fill = "lightblue") +
  theme_minimal() +
  labs(title = "Distribution of Sex", x = "Sex", y = "Count")
```



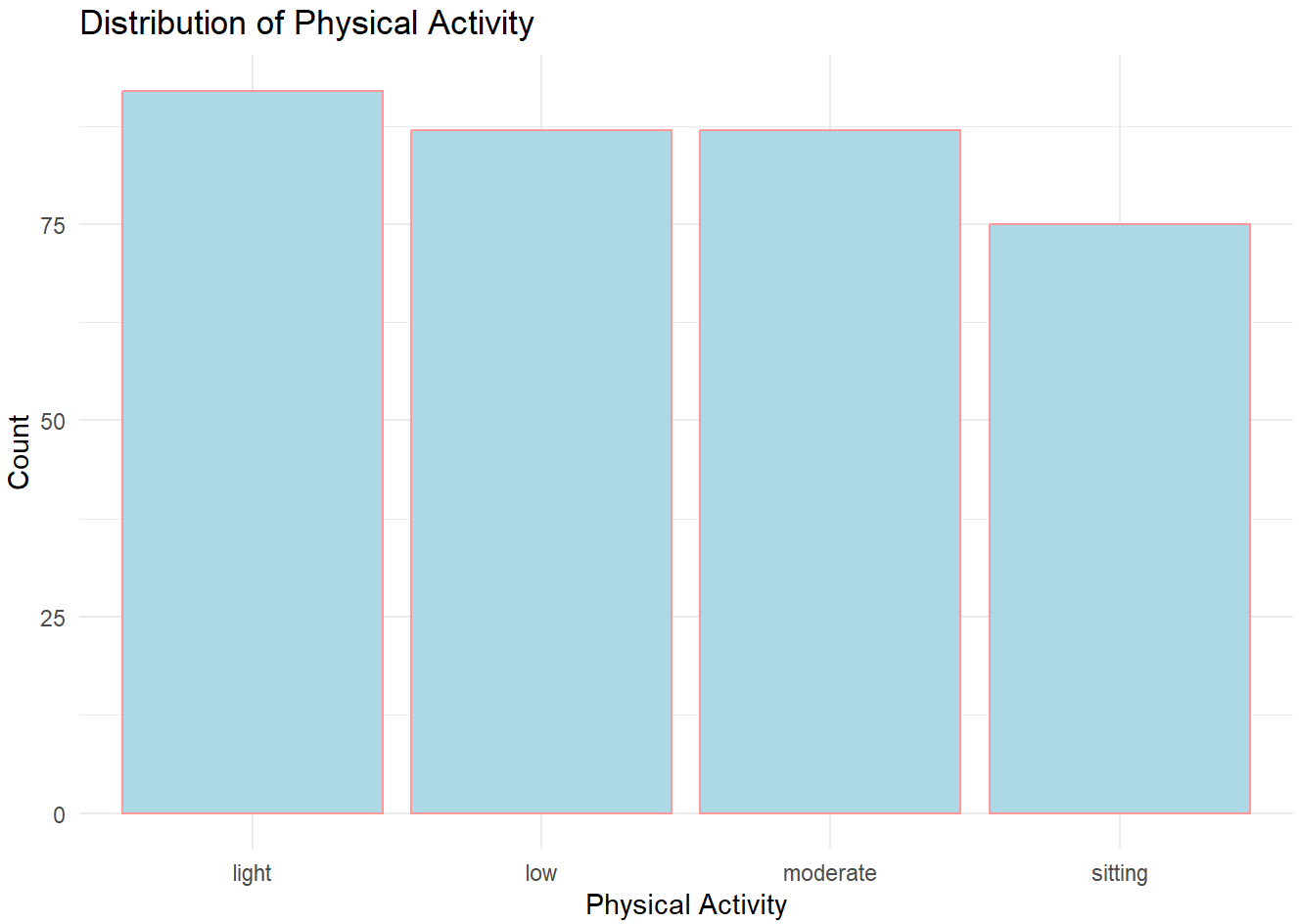

```
# Create a pie chart for the distribution of Sex
ggplot(carehome_data, aes(x = "", fill = sex)) + # Empty x aesthetic and fill by sex
  geom_bar(width = 1) + # Use geom_bar and set width to 1 to create a filled circle
  coord_polar("y") + # Transform the bar chart into a pie chart
  theme_void() + # Use theme_void to minimize extra chart elements
  labs(title = "Distribution of Sex", fill = "Sex") + # Add labels
  scale_fill_manual(values = c("M" = "lightblue", "F" = "lightcoral"))  # Customize colors
```



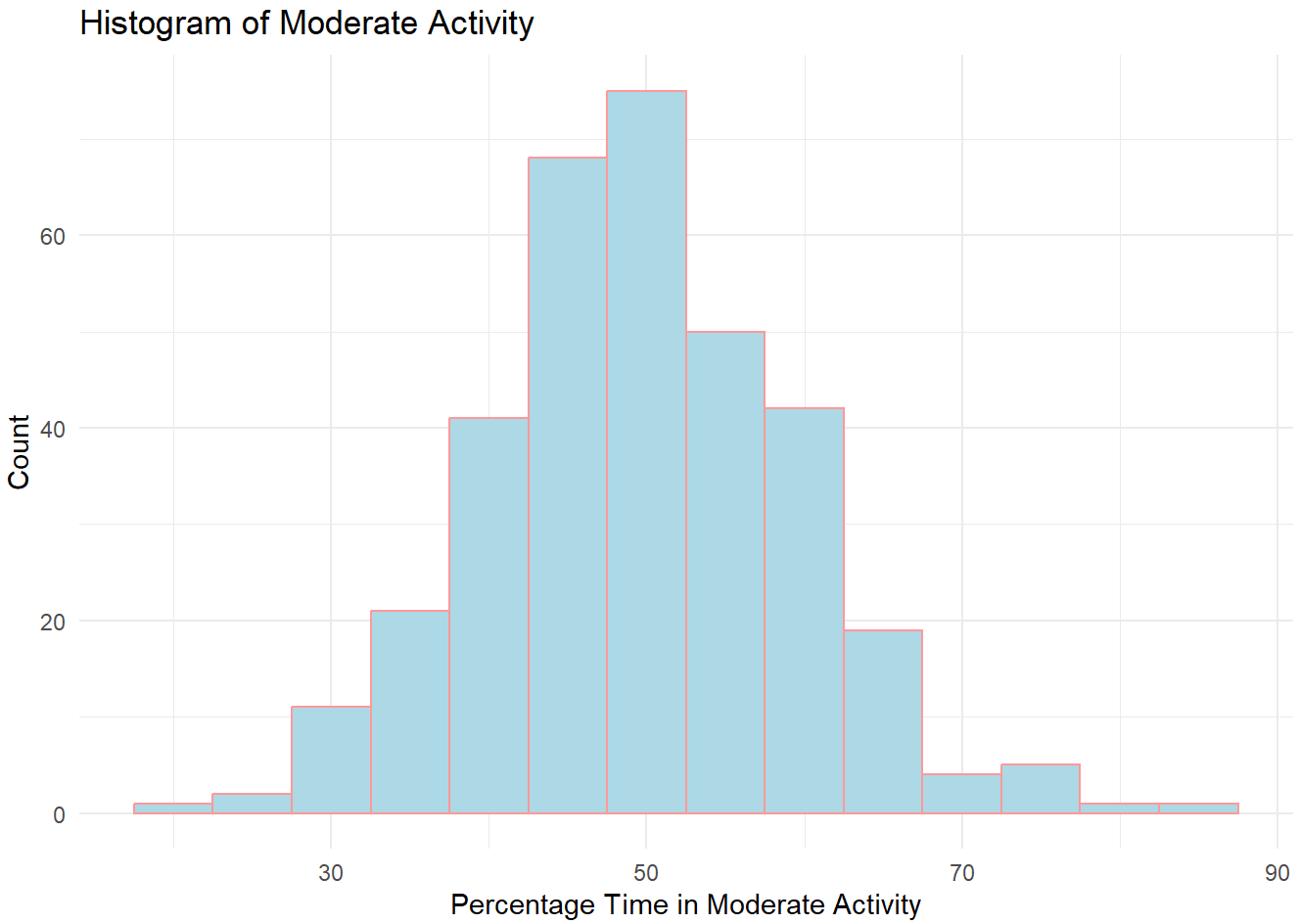

```
# Histogram for BMI
ggplot(carehome_data, aes(x = bmi)) +
  geom_histogram(binwidth = 0.5, fill = "lightblue", color = "#FF9999") +
  theme_minimal() +
  labs(title = "Histogram of BMI", x = "BMI", y = "Count")
```




```
# Bar Chart for Physical Activity Categories
ggplot(carehome_data, aes(x = physical_activity)) +
  geom_bar(color = "#FF9999", fill = "lightblue") +
  theme_minimal() +
  labs(title = "Distribution of Physical Activity", x = "Physical Activity", y = "Count")
```




```
# Histogram for Moderate Activity
ggplot(carehome_data, aes(x = moderate_activity)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "#FF9999") +
  theme_minimal() +
  labs(title = "Histogram of Moderate Activity", x = "Percentage Time in Moderate Activity", y =
"Count")
```



additional

2. Question

To answer the question regarding the association between care home and the type of physical activity undertaken, and if a longer amount of moderate activity is observed depending on the care home, a chi-square test of independence can be used for the first part to analyze the association between categorical variables (care home and type of physical activity). For the second part, a t-test or ANOVA can be used to compare the means of moderate activity between the two care homes if the data meets the assumptions for these tests. ### first part – chi squire

The results of the chi-square test will show if there's a statistically significant association between care home and physical activity type. A significant result suggests that the type of activity depends on the care home.

```
# Create a contingency table for care home and activity type
contingency_table <- table(carehome_data$carehome_id, carehome_data$physical_activity)

# Perform chi-square test
chi_square_result <- chisq.test(contingency_table)

# Print the results
print(chi_square_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 1.0714, df = 3, p-value = 0.784
```

The Pearson's Chi-squared test result with a chi-square statistic of 1.0714, degrees of freedom (df) = 3, and a p-value of 0.784 suggests that there is no statistically significant association between the care home and the type of physical activity undertaken by individuals. The high p-value (greater than the typical alpha level of 0.05) indicates that any observed differences in activity types across care homes are likely due to chance rather than a systematic relationship.

Second part

The t-test or ANOVA will indicate if there's a significant difference in moderate activity times between the care homes, with a significant result suggesting that residence in a particular care home might influence the amount of moderate activity undertaken by individuals.

```
# Filter the data for care homes 1 and 2
carehome1_data <- filter(carehome_data, carehome_id == 1)
carehome2_data <- filter(carehome_data, carehome_id == 2)

# Perform an independent two-sample t-test
t_test_result <- t.test(carehome1_data$moderate_activity, carehome2_data$moderate_activity, var.
equal = TRUE)

# Print the results
print(t_test_result)
```

```
##
##  Two Sample t-test
##
## data:  carehome1_data$moderate_activity and carehome2_data$moderate_activity
## t = 0.17578, df = 339, p-value = 0.8606
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.929110  2.307742
## sample estimates:
## mean of x mean of y
##  49.82530  49.63598
```

The results from the two-sample t-test indicate that the t-value is 0.17578 with 339 degrees of freedom, and the p-value is 0.8606. The p-value is much higher than the conventional threshold of 0.05, suggesting that there is no statistically significant difference in the mean moderate activity levels between the two care homes. In other words, the average amount of moderate activity undertaken by individuals does not significantly differ depending on the care home they reside in.

The 95% confidence interval for the difference in means ranges from -1.929110 to 2.307742, which includes zero. This further supports the conclusion that there is no significant difference between the two groups, as the confidence interval suggests that the true difference in means could be as low as approximately -1.93 or as high

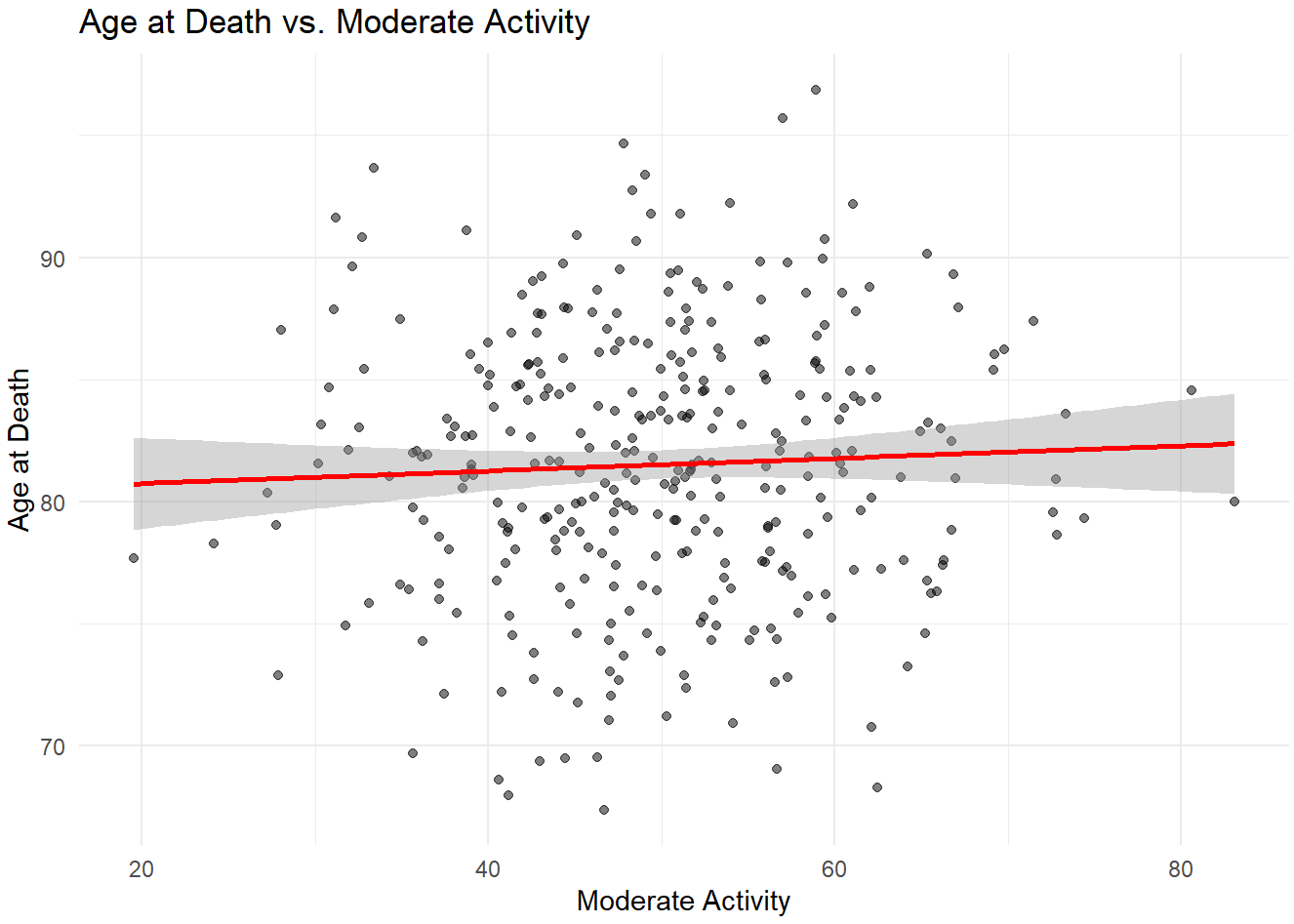
as approximately 2.31, but still includes the possibility of no difference (zero).

The selection of the t-test for this analysis was based on the objective to compare the means of a continuous variable (moderate activity levels) between two independent groups (the two care homes). The t-test is an appropriate statistical tool for this purpose when the data meets the assumptions of normality and equal variances between the two groups. In this context, the test was used to determine if residing in a particular care home has an effect on the amount of moderate activity individuals undertake, and the results suggest that the care home does not have a significant impact on moderate activity levels.

3. Question Physical activity and longevity

This quesiton explores whether the data provide any evidence that those who are more physically active live longer. It requires a formal recommendation on whether an intervention should be provided based on these results. This question aims to explore the relationship between physical activity and longevity, and it's essential to use appropriate statistical methods to analyze this relationship and provide clear understandable recommendations.

```
ggplot(carehome_data, aes(x = moderate_activity, y = age_at_death)) +
  geom_point(alpha = 0.5) +  # Plot the individual data points
  geom_smooth(method = "lm", color = "red") +  # Add a linear regression line
  theme_minimal() +
  labs(title = "Age at Death vs. Moderate Activity",
       x = "Moderate Activity",
       y = "Age at Death")
```



```
# Perform linear regression
model <- lm(age_at_death ~ moderate_activity, data = carehome_data)

# Summary of the regression model
summary(model)
```

```
##
## Call:
## lm(formula = age_at_death ~ moderate_activity, data = carehome_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0442  -3.7428   0.1046   3.9076  15.1179
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       80.23381    1.52355  52.662   <2e-16 ***
## moderate_activity  0.02580    0.03004   0.859    0.391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.5 on 339 degrees of freedom
## Multiple R-squared:  0.00217,    Adjusted R-squared:  -0.0007732
## F-statistic: 0.7373 on 1 and 339 DF,  p-value: 0.3911
```

The linear regression analysis indicates no significant relationship between moderate activity and age at death. The coefficient for moderate activity is 0.02580, meaning for each unit increase in moderate activity, age at death increases by about 0.026 years, but this is not statistically significant (p-value = 0.391). The model explains a very small portion of the variance in age at death (Multiple R-squared: 0.00217), suggesting other factors not included in the model might be more influential in determining longevity.

Question 4

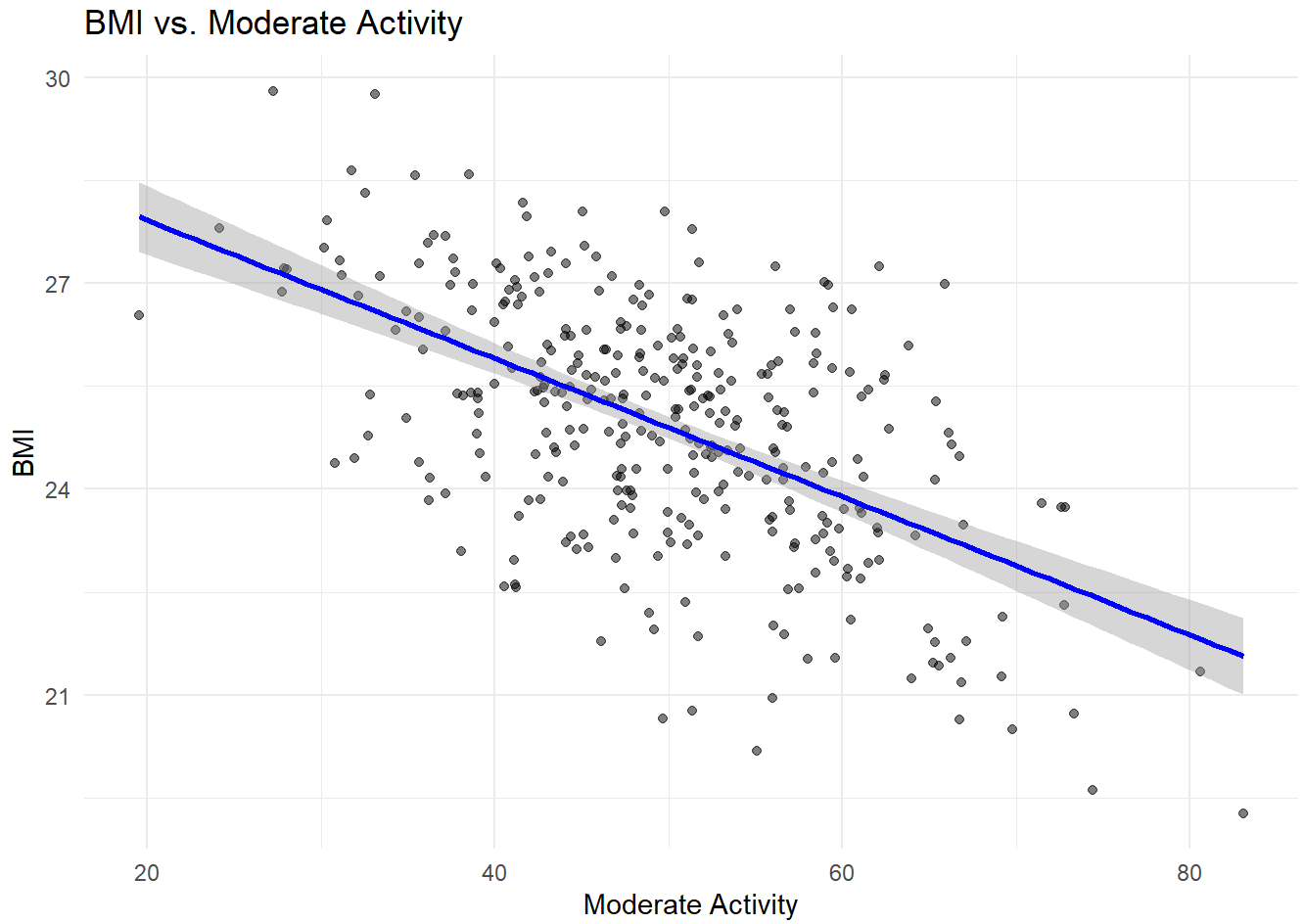
Question four explores the hypothesis that BMI decreases as the proportion of moderate activity increases. It involves using appropriate visualizations and a linear regression model to test this hypothesis and quantify the relationship, providing insights into the impact of moderate activity on BMI.

To perform the linear regression analysis exploring the relationship between BMI and moderate activity, and to visualize this relationship, you can use the following R code:

```
model2 <- lm(bmi ~ moderate_activity, data = carehome_data)

# Display the summary of the regression model
summary(model2)
```

```
##
## Call:
## lm(formula = bmi ~ moderate_activity, data = carehome_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2880 -0.9991  0.0876  1.0663  3.6989
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.938631   0.415956   71.98   <2e-16 ***
## moderate_activity -0.100816   0.008203  -12.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.502 on 339 degrees of freedom
## Multiple R-squared:  0.3083, Adjusted R-squared:  0.3062
## F-statistic: 151.1 on 1 and 339 DF,  p-value: < 2.2e-16
```

```
# Plot BMI vs. Moderate Activity with regression line
ggplot(carehome_data, aes(x = moderate_activity, y = bmi)) +
  geom_point(alpha = 0.5) +  # Plot individual data points
  geom_smooth(method = "lm", color = "blue") +  # Add linear regression line
  theme_minimal() +
  labs(title = "BMI vs. Moderate Activity",
       x = "Moderate Activity",
       y = "BMI")
```



The linear regression analysis results indicate a significant relationship between moderate activity and BMI. The coefficient for moderate activity is -0.100816, with a highly significant p-value (<2e-16), suggesting that for every unit increase in moderate activity, BMI decreases by approximately 0.101 units. The negative sign of the coefficient confirms that the relationship is inverse, aligning with the hypothesis that increased moderate activity is associated with lower BMI.

The intercept, 29.938631, represents the estimated BMI when moderate activity is zero. The t-value for the moderate activity coefficient, -12.29, further emphasizes its statistical significance.

The model's residual standard error is 1.502, indicating the average distance of the data points from the fitted regression line. The R-squared value of 0.3083 suggests that approximately 30.83% of the variability in BMI can be explained by the model, which is a moderate amount of explanatory power.

Overall, the analysis provides strong evidence supporting the hypothesis that higher levels of moderate activity are associated with lower BMI values, making a compelling case for promoting moderate physical activity as part of weight management strategies.