Physical Activity in Relation to Health Outcome and Longevity in Care Home in Scotland

Exam Number: B249125

Introduction to statistics in health and social care (2023-2024)

Contents

Introduction

This study explores a dataset from two fictitious Scottish care homes in an effort to learn more about the relationship between senior citizens' health outcomes, demographics, and physical activity. Through a thorough examination of demographic and baseline data, our goal is to identify patterns and trends that may guide focused actions.

We also investigate the relationship between the physical environment of the care home and the physical activities that the residents participate in, carefully examining the possible effects on their longevity and general well-being.

In an effort to add to the growing body of knowledge on healthy aging, special attention is paid to comprehending the relationship between BMI and moderate physical activity. The results of this investigation have the potential to improve the quality of life for residents of care homes by providing practical advice.

Load data

Let's go on to the analysis of the given dataset in R, answering each of these questions individually. I'm going to begin by loading the dataset and doing some initial data investigation.

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

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

participant_id	sex	age_at_recording	age_at_death	physical_activity	moderate_activity	bmi
1	M	82.85775	83.71636	low	47.33063	23.76049
2	M	81.19990	81.19990	low	45.28279	26.31856
3	M	81.46722	83.34632	light	48.88195	22.20059
4	F	72.37810	72.37810	low	51.38753	26.04468
5	M	85.71573	85.71573	light	51.08582	23.19571
6	F	88.01906	91.10457	sitting	38.74970	26.98510

After the dataset was successfully imported into our R environment, a comprehensive preliminary assessment is to be carried out carried out to guarantee data quality and preparation for analysis. The participants' demographic and baseline characteristics will be examined as the next step.

In order to determine the age distribution, gender ratio, BMI ranges, and degree of moderate physical activity among people, this step will involve producing descriptive data. The objective is to lay the groundwork for a deeper analysis of the population being studied, with a focus on the connections between physical activity patterns, health outcomes, care home conditions, and lifespan. This methodical methodology guarantees that the conclusions drawn later are based on a solid examination of the core characteristics of the dataset.

Demographics and Baseline Characteristics (Question 1)

In order to understand the data, some descreptive statistics is done to learn about the demographics and baseline statistics.

```r
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")

descriptive_stats_long
```

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
sd_moderate_activity	9.928704
mean_bmi	24.924828
sd_bmi	1.802885
min_bmi	19.273478
max_bmi	29.809399

There are 341 people in the dataset, and their IDs range from 1 to 341. The age range of the participants during recording is 66.53 to 93.49 years, with an average of 79.85 years and a standard deviation of 5.18 years.

The average age of death is determined to be 81.52 years, with a standard deviation of 5.50 years and a range of 67.39 to 96.87 years.

The average percentage of time that individuals spent engaging in moderate physical activity was 49.73%, with a standard deviation of 9.93%.

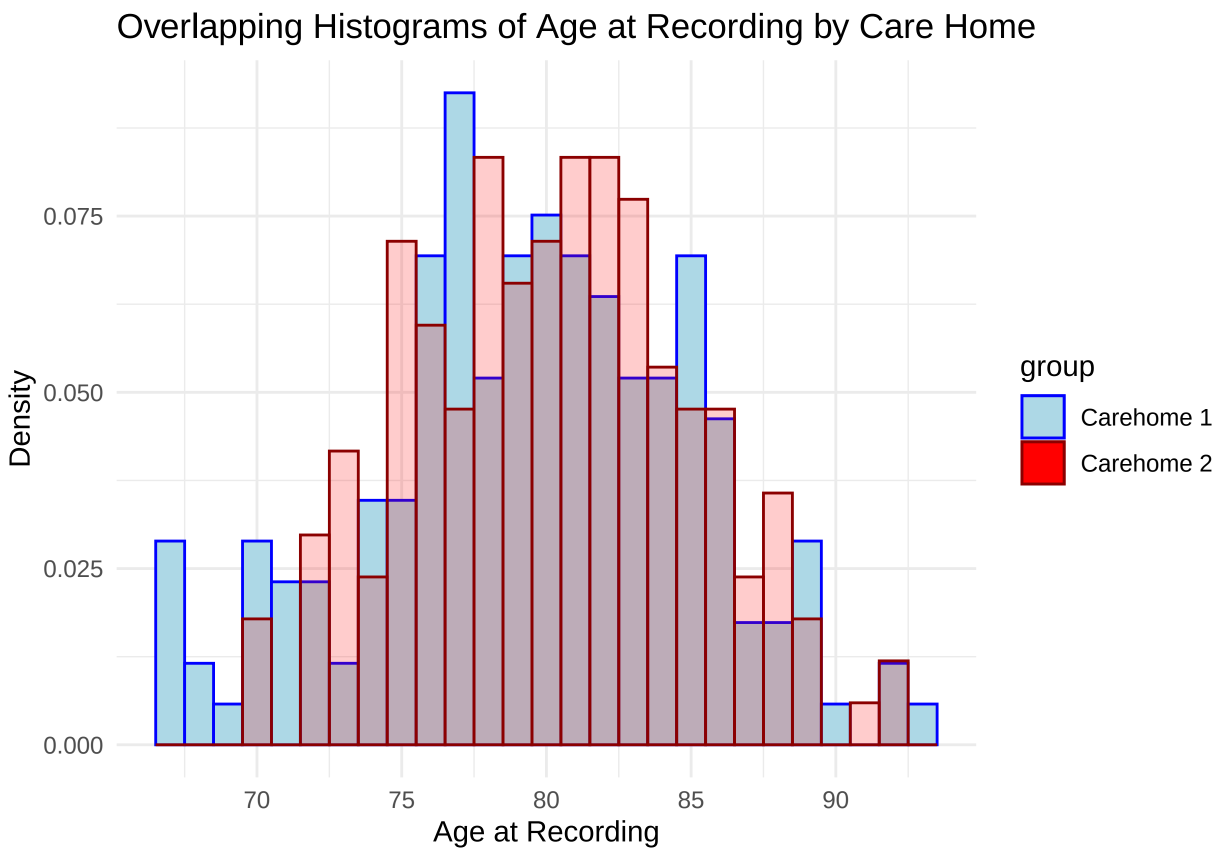
The BMI measurement ranges from 19.27 to 29.81, with a standard deviation of 1.80 and an average of 24.92. Participants are split almost evenly between the two assisted living facilities.

Visualizaiton of the demographics and baseline characteristics

In order to visualize the data, and see the current baseline comparing the two care homes against each other, we can check the following in details:

Age at recoding

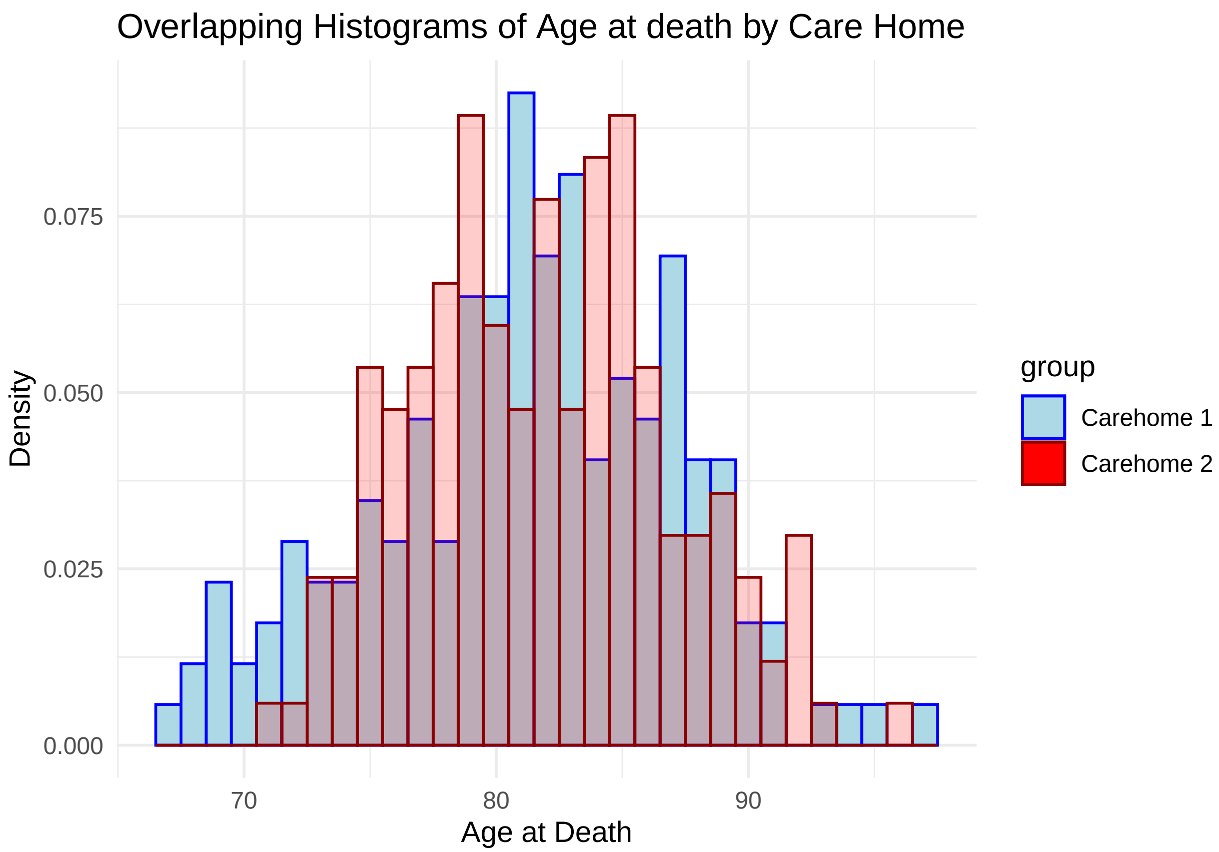
```r
carehome_data$group <- factor(carehome_data$carehome_id, levels = c(1, 2), labels = c("Carehome 1", "Carehome 2"))

ggplot(carehome_data, aes(x = age_at_recording, y = ..density.., fill = group, color = group)) +
  geom_histogram(data = subset(carehome_data, carehome_id == 1),
                 binwidth = 1, alpha = 1.5) +
  geom_histogram(data = subset(carehome_data, carehome_id == 2),
                 binwidth = 1, alpha = 0.2) +
  scale_fill_manual(values = c("Carehome 1" = "lightblue", "Carehome 2" = "red")) +
  scale_color_manual(values = c("Carehome 1" = "blue", "Carehome 2" = "darkred")) +
  theme_minimal() +
  labs(title = "Overlapping Histograms of Age at Recording by Care Home",
       x = "Age at Recording", y = "Density")
```



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.

Age at death

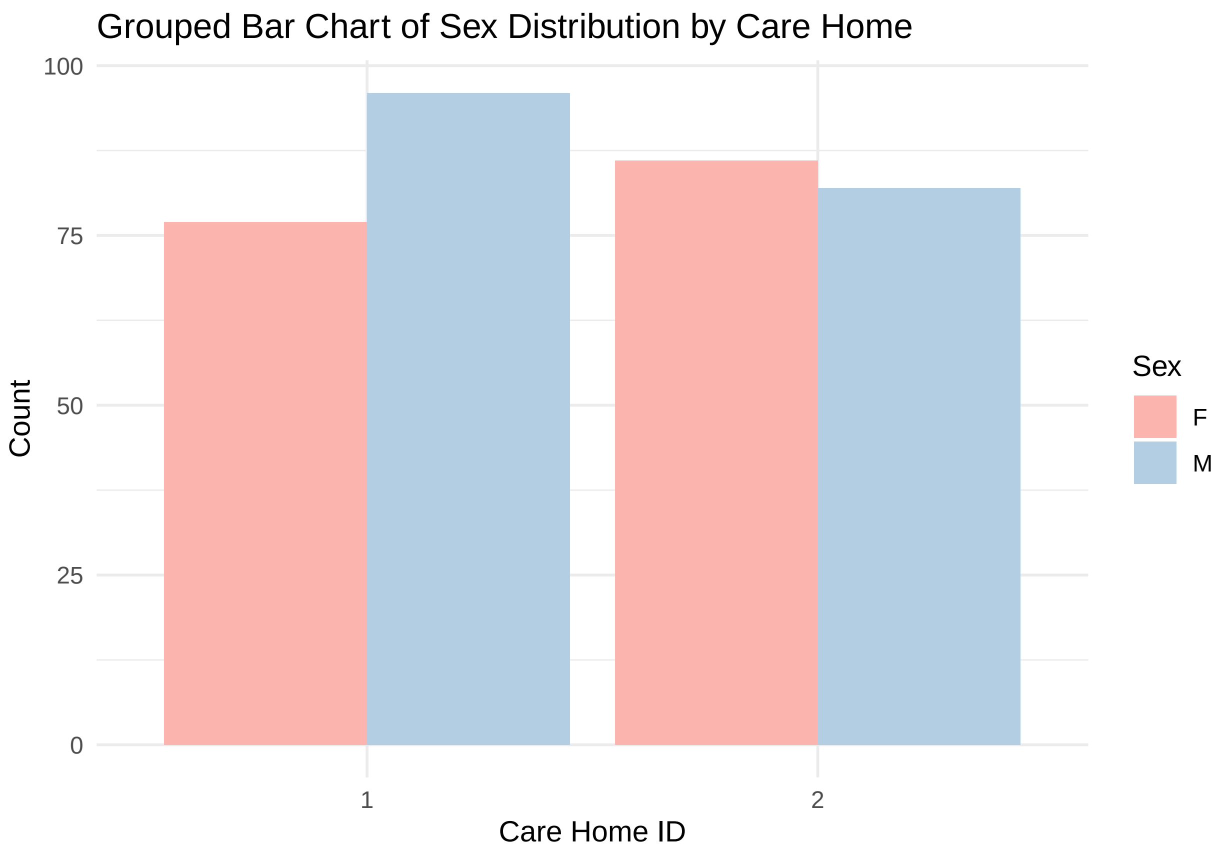
```
carehome_data$group <- factor(carehome_data$carehome_id, levels = c(1, 2), labels = c("Carehome 1", "Carehome 2"))

ggplot(carehome_data, aes(x = age_at_death, y = ..density.., fill = group, color = group)) +
  geom_histogram(data = subset(carehome_data, carehome_id == 1),
                 binwidth = 1, alpha = 1.5) +
  geom_histogram(data = subset(carehome_data, carehome_id == 2),
                 binwidth = 1, alpha = 0.2) +
  scale_fill_manual(values = c("Carehome 1" = "lightblue", "Carehome 2" = "red")) +
  scale_color_manual(values = c("Carehome 1" = "blue", "Carehome 2" = "darkred")) +
  theme_minimal() +
  labs(title = "Overlapping Histograms of Age at death by Care Home",
       x = "Age at Death", y = "Density")
```



The average age of death is determined to be 81.52 years, with a standard deviation of 5.50 years and a range of 67.39 to 96.87 years.

Gender distribution

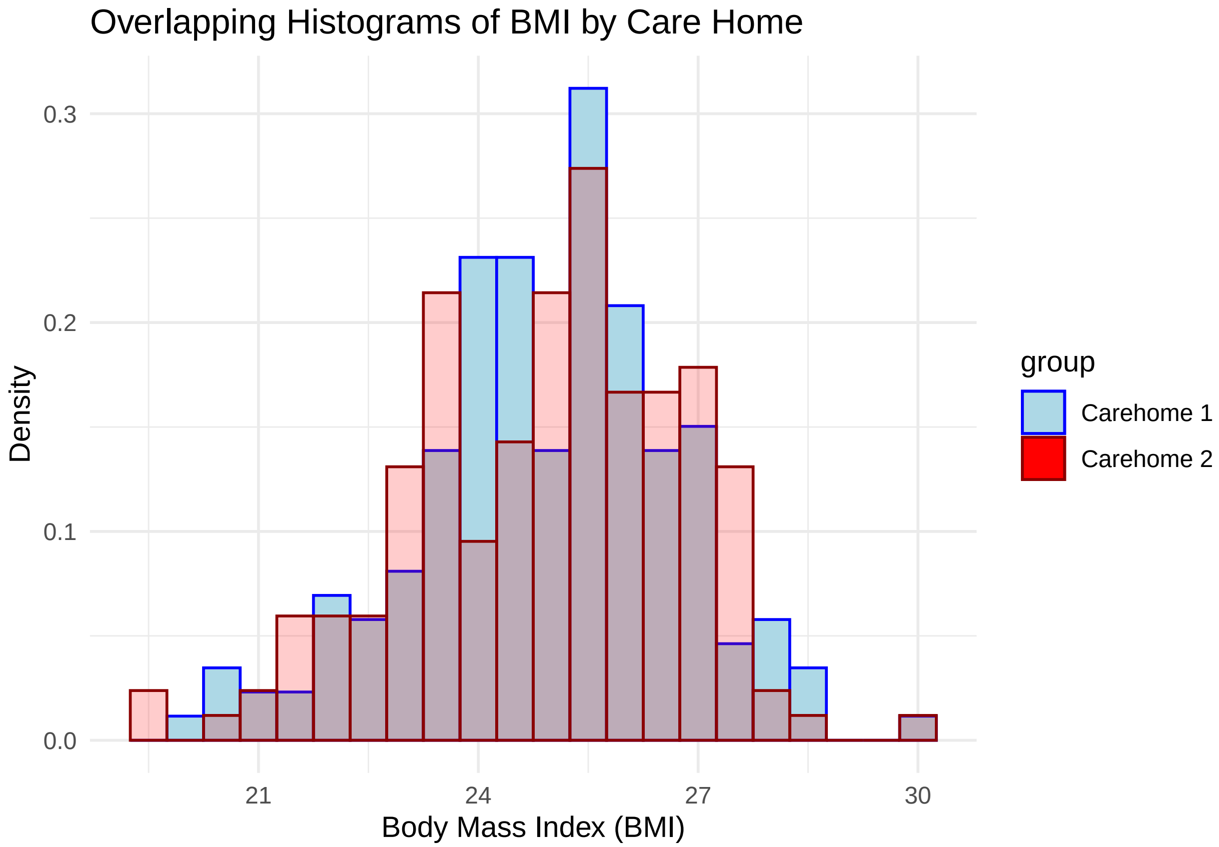
```r
# Grouped bar chart for sex distribution within each care home
ggplot(carehome_data, aes(x = as.factor(carehome_id), fill = sex)) +
  geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Pastel1", name = "Sex") +
  theme_minimal() +
  labs(title = "Grouped Bar Chart of Sex Distribution by Care Home",
       x = "Care Home ID",
       y = "Count")
```



The sex distribution are very close especially in care home 2 and with sligh increase in males in the care home 1.

BMI distibution between the two homes

```
carehome_data$group <- factor(carehome_data$carehome_id, levels = c(1, 2), labels = c("Carehome 1", "Carehome 2"))

ggplot(carehome_data, aes(x = bmi, y = ..density.., fill = group, color = group)) +
  geom_histogram(data = subset(carehome_data, carehome_id == 1),
                 binwidth = 0.5, alpha = 1.5) +
  geom_histogram(data = subset(carehome_data, carehome_id == 2),
                 binwidth = 0.5, alpha = 0.2) +
  scale_fill_manual(values = c("Carehome 1" = "lightblue", "Carehome 2" = "red")) +
  scale_color_manual(values = c("Carehome 1" = "blue", "Carehome 2" = "darkred")) +
  theme_minimal() +
  labs(title = "Overlapping Histograms of BMI by Care Home",
       x = "Body Mass Index (BMI)", y = "Density")
```

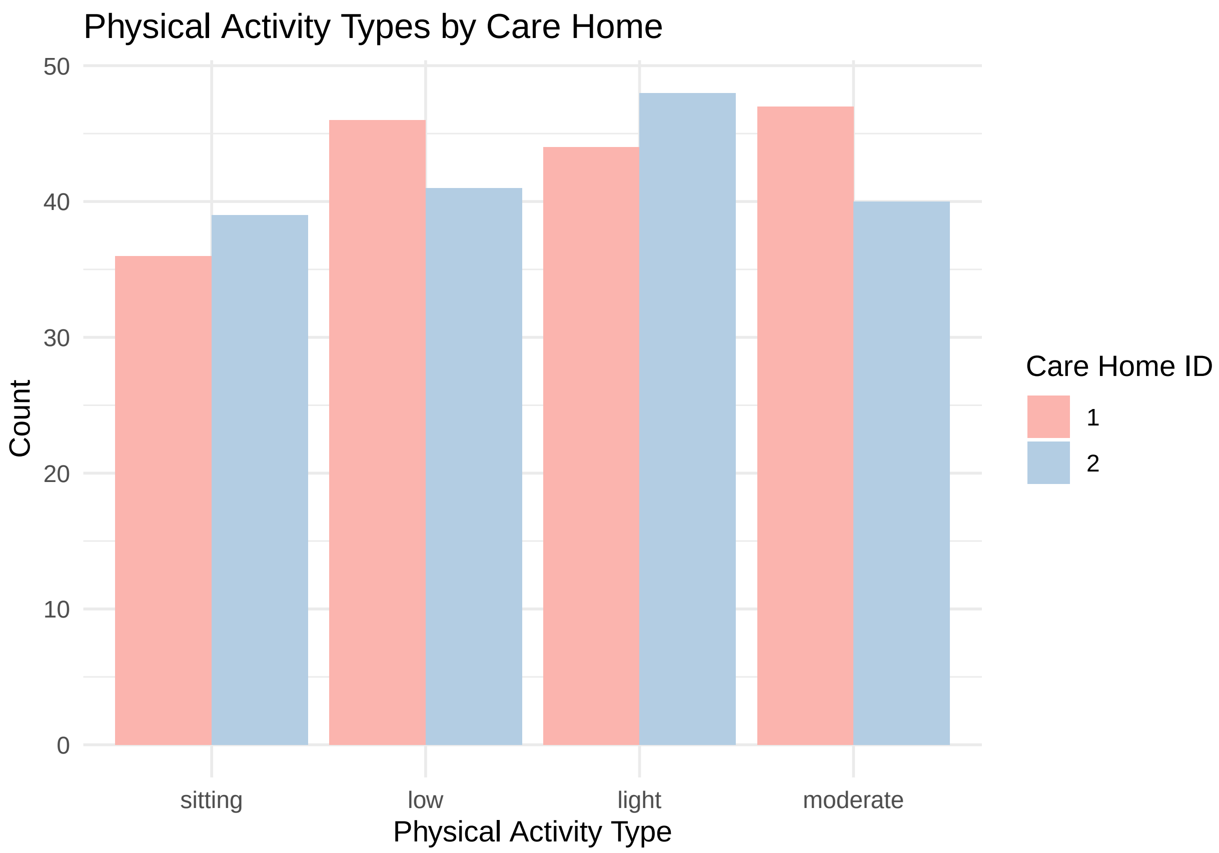


As mentioned before, the participants' Body Mass Index (BMI) ranges from a minimum of 19.27 to a maximum of 29.81 showing a wide range of body weight statuses from normal to overweight. This would reflect somehow the health status and be later used as a measure of higher activity levels especially if there is positive relationship detected between BMI and activity level, which we will examin later in this study.

As the histogram shows there is big overlap between the two care homes in the ranges with a normally distributed shape for the BMI in each one of them.

Activity type distribution between the two care homes

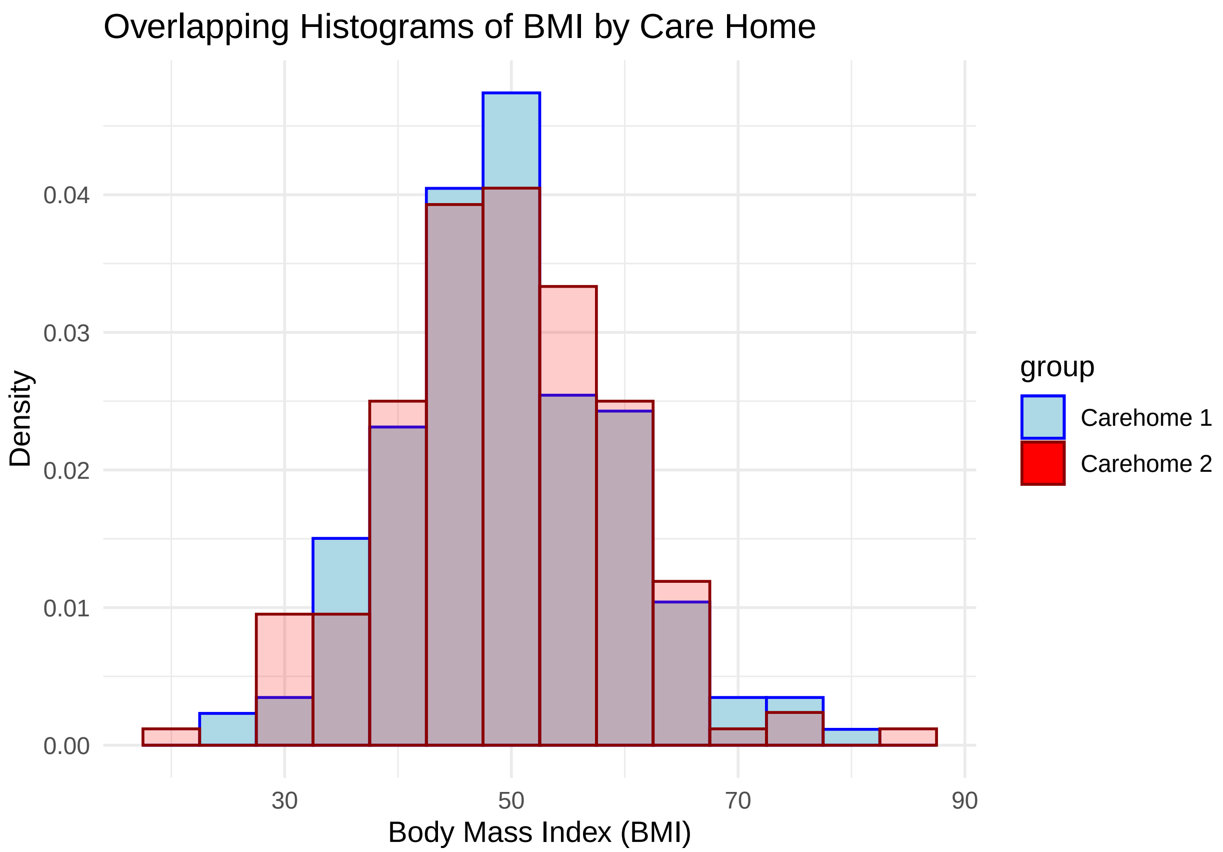
```r
# Bar chart for physical activity types by care home
ggplot(carehome_data, aes(x = factor(physical_activity, levels = c("sitting", "low", "light", "moderate")), fill =
  geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Pastel1", name = "Care Home ID") +
  theme_minimal() +
  labs(title = "Physical Activity Types by Care Home",
       x = "Physical Activity Type",
       y = "Count")
```



Also this is an almost even distribution of the four types of physical activity ranging between low, light, moderate and sitting and almost very close levels when comparing the two care homes as seen above.

Distribution of moderate activity comparing the two care homes (Percentage of time each individual spent undertaking activity classified as moderate)

```r
# histograms with density on the y-axis
carehome_data$group <- factor(carehome_data$carehome_id, levels = c(1, 2), labels = c("Carehome 1", "Carehome 2"))

ggplot(carehome_data, aes(x = moderate_activity, y = ..density.., fill = group, color = group)) +
  geom_histogram(data = subset(carehome_data, carehome_id == 1),
                 binwidth = 5, alpha = 1.5) +
  geom_histogram(data = subset(carehome_data, carehome_id == 2),
                 binwidth = 5, alpha = 0.2) +
  scale_fill_manual(values = c("Carehome 1" = "lightblue", "Carehome 2" = "red")) +
  scale_color_manual(values = c("Carehome 1" = "blue", "Carehome 2" = "darkred")) +
  theme_minimal() +
  labs(title = "Overlapping Histograms of BMI by Care Home",
       x = "Body Mass Index (BMI)", y = "Density")
```



On average, participants across both care homes engaged in moderate activity around 49.73% of the time indicating a generally active lifestyle among the residents.

The difference in moderate activity levels between the two care homes is relatively small. Care home 1 in blue may show higher activity especially at different points which suggests slightly more active or engaged resident population when it comes to moderate physical activities.

Association between the care home and type of physical activity first and the with the moderate activity level (Question 2)

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.

مبسطة

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)
contingency_table
```

```
##
##     light low moderate sitting
##   1    44  46       47      36
##   2    48  41       40      39
```

```
# 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.

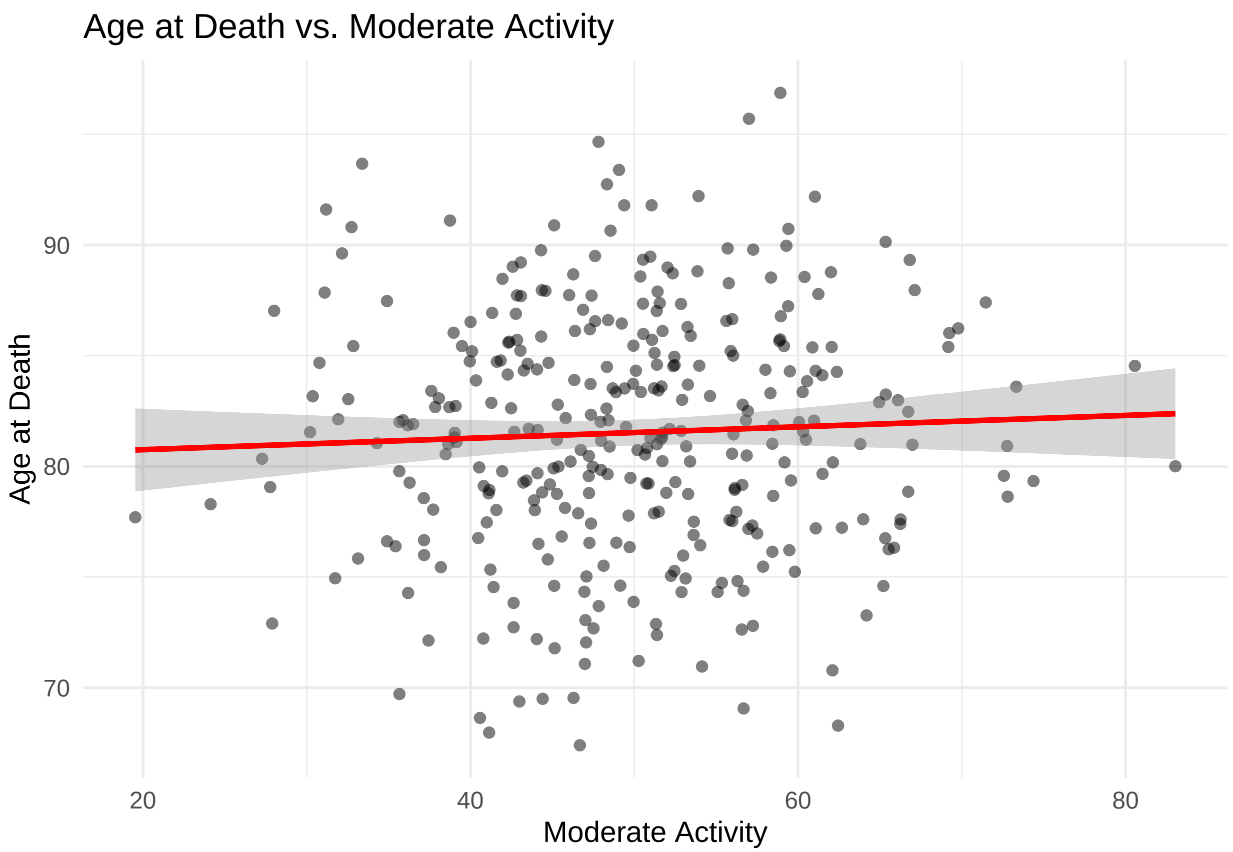
The 95% confidence interval for the difference in means ranges from -1.929110 to 2.307742, which includes zero. This further supports the conclusion that there is no significant difference between the two groups, as the confidence interval suggests that the true difference in means could be as low as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

Physical activity and longevity (Question no. 3)

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.

```r
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.

BMI realtionship as an indciaiton for health in relation to moderate activity level (Question no. 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.

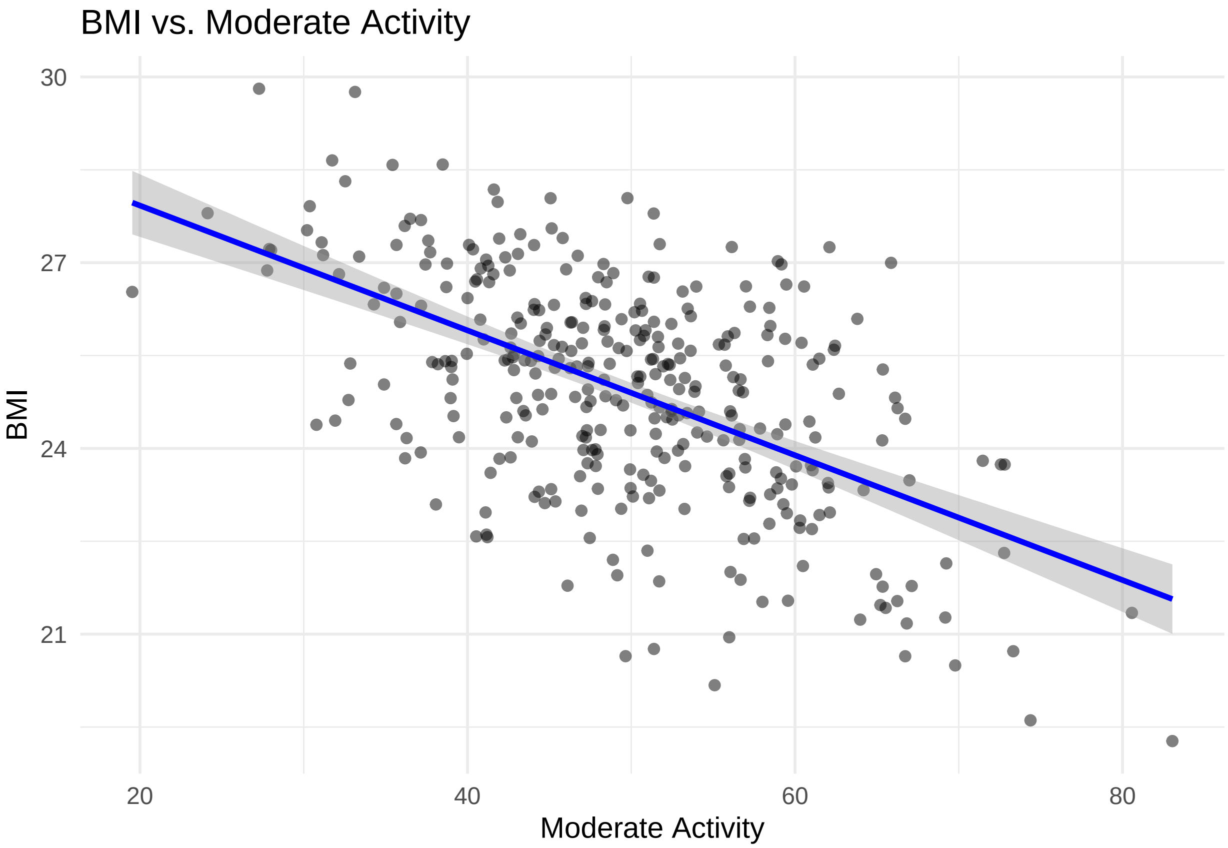
[illegible]

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.