EDA Report

Xiaohan Shi, Zihao Zhang, Suheng Yao

2024-11-15

Read in the dataset

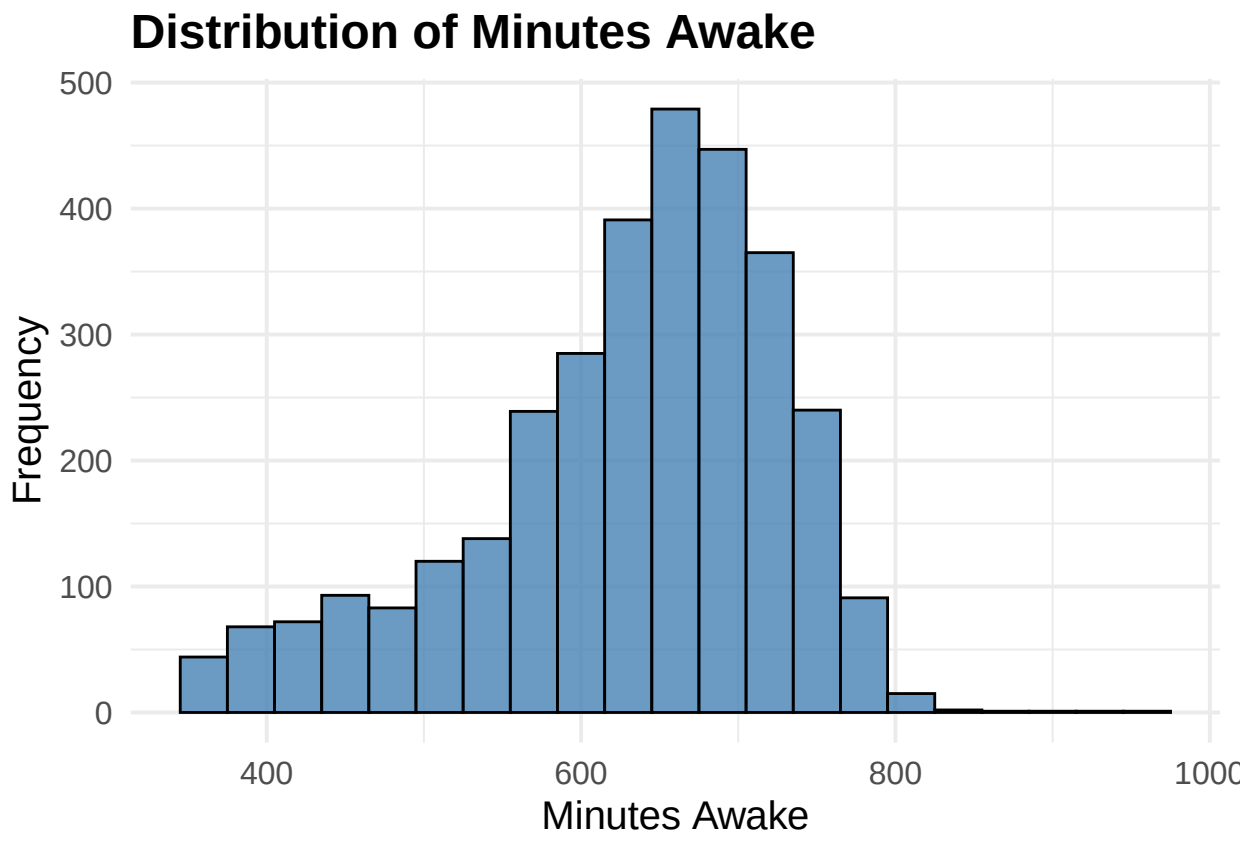
```
library(dplyr)
library(ggplot2)
df_score <- read.csv("diversity_score.csv")
df <- read.csv("final_data.csv")
```

Some Data Description

The client divided all the ages into 7 groups. In the original dataset, if the age is between 0 to 2, then those orangutans are classified as age level 1; if the age is between 2 to 4, then those orangutans are classified as age level 2; if the age is between 4 to 6, then those orangutans are classified as age level 3; if the age is between 6 to 8, then those orangutans are classified as age level 4; if the age is between 8 to 10, then those orangutans are classified as age level 5; if the age is between 10 to 12, then those orangutans are classified as age level 5. For those with ages greater than 12, they go to the age 7 group.

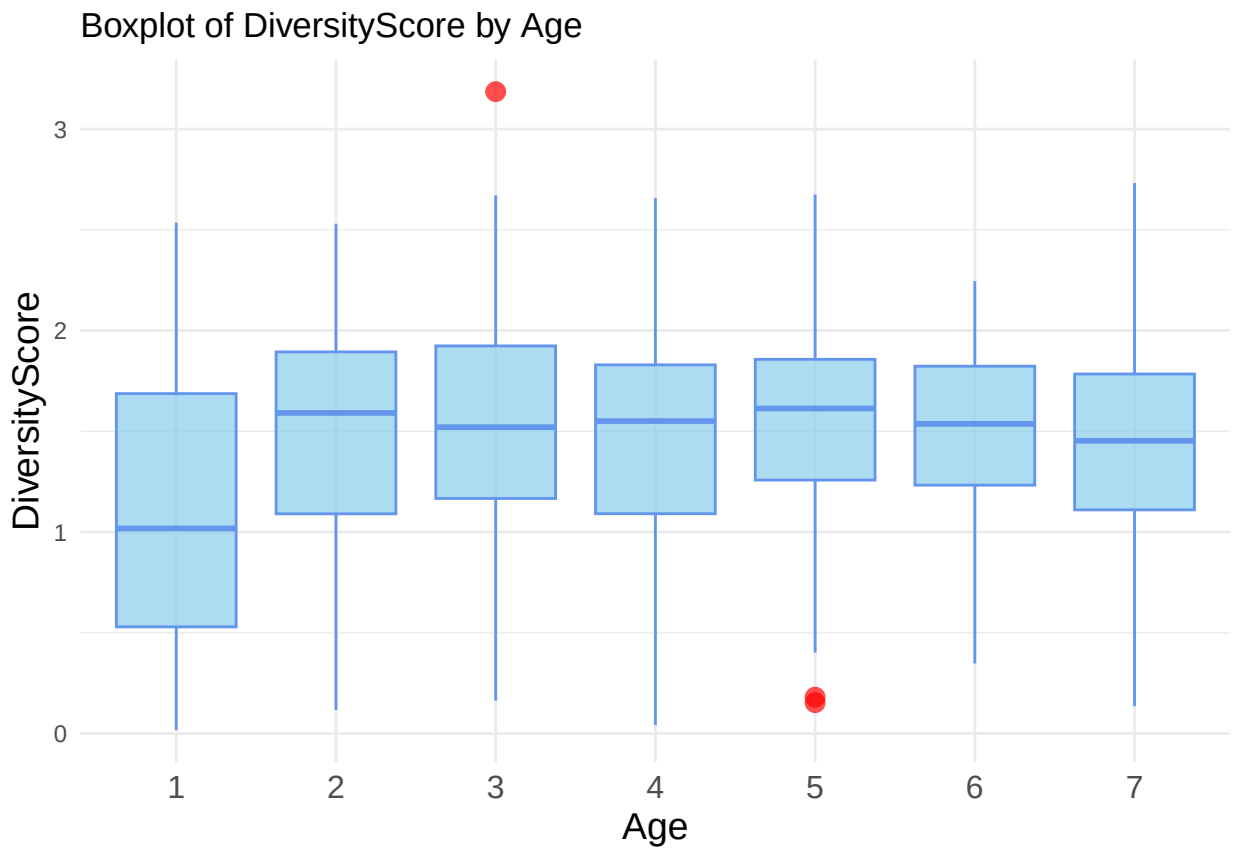
Part 1: The Distribution of Minutes Awake

```
ggplot(df, aes(x = `Minutes.awake`)) +
  geom_histogram(binwidth = 30, fill = "steelblue", color = "black", alpha = 0.8) +
  labs(
    title = "Distribution of Minutes Awake",
    x = "Minutes Awake",
    y = "Frequency"
  ) +
  theme_minimal(base_size = 15) +
  theme(
    axis.text = element_text(size = 12),
    plot.title = element_text(size = 18, face = "bold")
  )
```



Part 2: Visualization of Diversity Score

```r
# Boxplot
ggplot(df_score, aes(x = factor(Age), y = DiversityScore)) +
  geom_boxplot(fill = "skyblue", color = "#6495ED",
               outlier.colour = "red", outlier.size = 3,
               alpha=0.7) +
  labs(
    title = "Boxplot of DiversityScore by Age",
    x = "Age",
    y = "DiversityScore"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 12), axis.title = element_text(size = 14))
```




```r
# Violinplot
ggplot(df_score, aes(x = factor(Age), y = DiversityScore)) +
  geom_violin(fill = "#F3A683", color = "#E67F00", alpha=0.7) +
  labs(
    title = "Violinplot of DiversityScore by Age",
    x = "Age",
    y = "DiversityScore"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 12), axis.title = element_text(size = 14))
```

The boxplot shows that the median of Diversity Score increases slightly with Age, especially from Age group 1 to Age group 2. Then the trend leveled off and the median for all age groups remained at about 1.5. There are two outliers which appeared in Age 3 and Age 5 groups.

In the violinplot, the similar width of the violin illustrates that the sample size is relatively balanced across age groups due to, there is no particularly small or large sample group.

It also shows that the Diversity Score for all age groups was roughly distributed between 0 and 3. The distribution of Diversity score in group 2-7 are similar that most observations are concentrated between 1 and 2. The data in Age 1 group are dispersed to a large extent, which is from 0 to 2.5.

In addition, as the age get older, the variability of the data seems to decrease and the distribution becomes more concentrated.

Part 3: Diversity Score Statistics

```
#eda 3
diversity_stats <- df_score %>%
  group_by(Age) %>%
  summarise(
    Min = min(DiversityScore, na.rm = TRUE),
    Max = max(DiversityScore, na.rm = TRUE),
    Median = median(DiversityScore, na.rm = TRUE),
    Mean = mean(DiversityScore, na.rm = TRUE),
    SD = sd(DiversityScore, na.rm = TRUE) # Standard deviation
  )
```

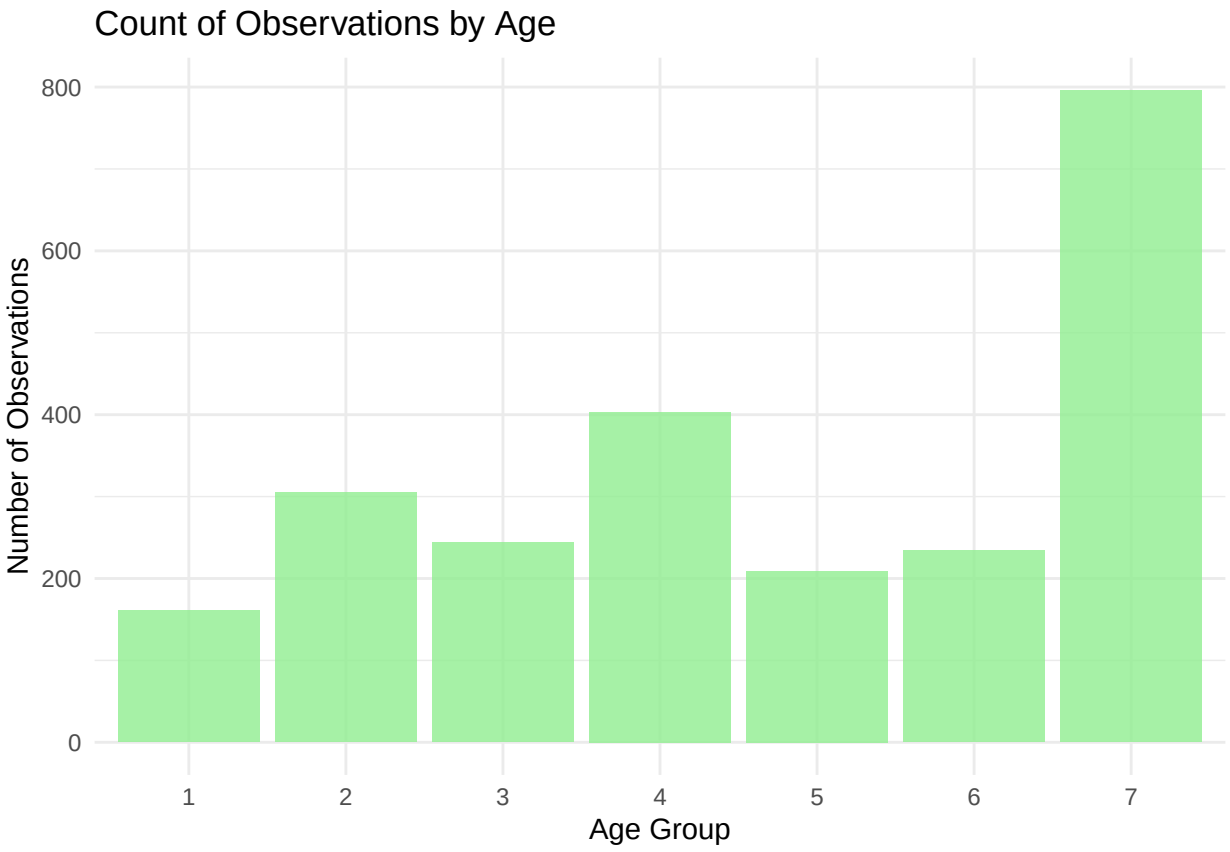
```r
# Print the summary statistics
print(diversity_stats)
```

```
## # A tibble: 7 x 6
##     Age    Min   Max Median  Mean    SD
##   <int>  <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     1 0.0170  2.54   1.02  1.11 0.671
## 2     2 0.116   2.53   1.59  1.50 0.523
## 3     3 0.163   3.19   1.52  1.53 0.521
## 4     4 0.0414  2.66   1.55  1.47 0.507
## 5     5 0.153   2.68   1.61  1.54 0.462
## 6     6 0.347   2.25   1.54  1.50 0.411
## 7     7 0.135   2.73   1.45  1.44 0.499
```

```r
age_counts <- df_score %>%
  group_by(Age) %>%
  summarise(count = n())
# Print the counts to the console
print(age_counts)
```

```
## # A tibble: 7 x 2
##     Age count
##   <int> <int>
## 1     1   161
## 2     2   305
## 3     3   244
## 4     4   403
## 5     5   209
## 6     6   234
## 7     7   796
```

```r
# Create a bar plot for the count of observations by age
ggplot(age_counts, aes(x = as.factor(Age), y = count)) +
  geom_bar(stat = "identity", fill = "lightgreen", alpha = 0.8) + # Bar plot
  labs(
    title = "Count of Observations by Age",
    x = "Age Group",
    y = "Number of Observations"
  ) +
  theme_minimal()
```

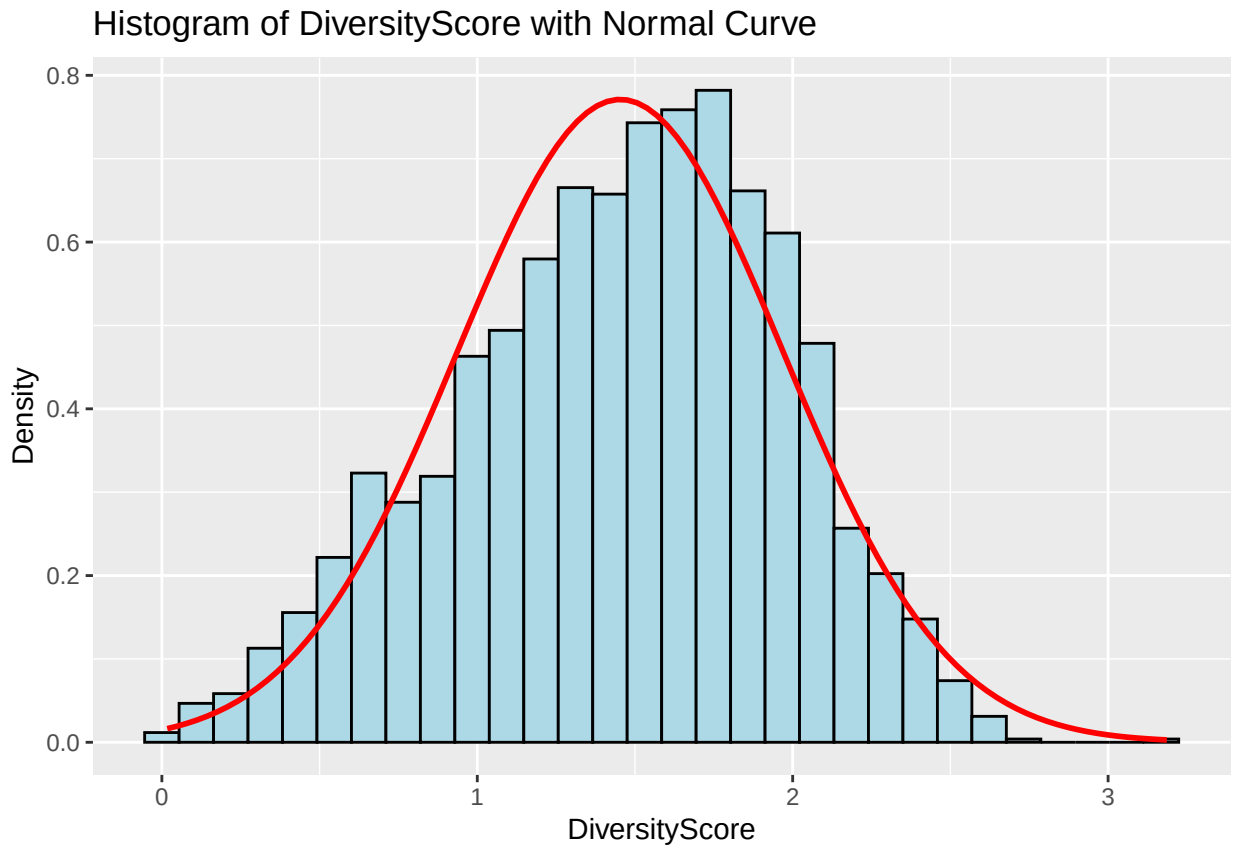


The number of observations varies significantly across age groups. Age group 7 has the largest number of observations(about 800), which might introduce a sampling bias in the data analysis. Age group 1 has the smallest count(about 180), which could affect the reliability of statistical summaries or models for that group.

Part 4: ANOVA Table Analysis

Step 1: Check for Normality of Diversity Score

```
ggplot(df_score, aes(x = DiversityScore)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30, fill = "lightblue", color = "black") +
  stat_function(fun = dnorm, args = list(mean = mean(df_score$DiversityScore, na.rm = TRUE),
                                         sd = sd(df_score$DiversityScore, na.rm = TRUE)),
                color = "red", size = 1) +
  labs(title = "Histogram of DiversityScore with Normal Curve",
       x = "DiversityScore",
       y = "Density")
```



Since as shown in the histogram above, the diversity score follows a close to normal distribution, we can proceed to use ANOVA table to analyze the relationship between diversity score and age.

```
anova_result <- aov(DiversityScore ~ Age, data = df_score)
summary(anova_result)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Age             1    1.6  1.5942   5.969 0.0146 *
## Residuals    2350  627.7  0.2671
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table result, since p-value of f-test is less than 0.05, we can reject the null hypothesis, saying that there is statistically significant difference for diversity score between different age groups.

Part 5: Distribution of Total Minutes Awake

```
#eda 5
final_data_df<-read.csv("final_data.csv")
total_minutes_awake <- final_data_df %>%
  group_by(Age) %>%
  summarise(total_minutes_awake = sum(Minutes.awake, na.rm = TRUE)) # Sum of minutes awake
```

```r
# Print the total minutes awake for each age
print(total_minutes_awake)
```

```
## # A tibble: 7 x 2
##     Age total_minutes_awake
##   <int>               <dbl>
## 1     1             145955.
## 2     2             239800.
## 3     3             205251.
## 4     4             272257.
## 5     5             196419.
## 6     6             224928.
## 7     7             717616.
```

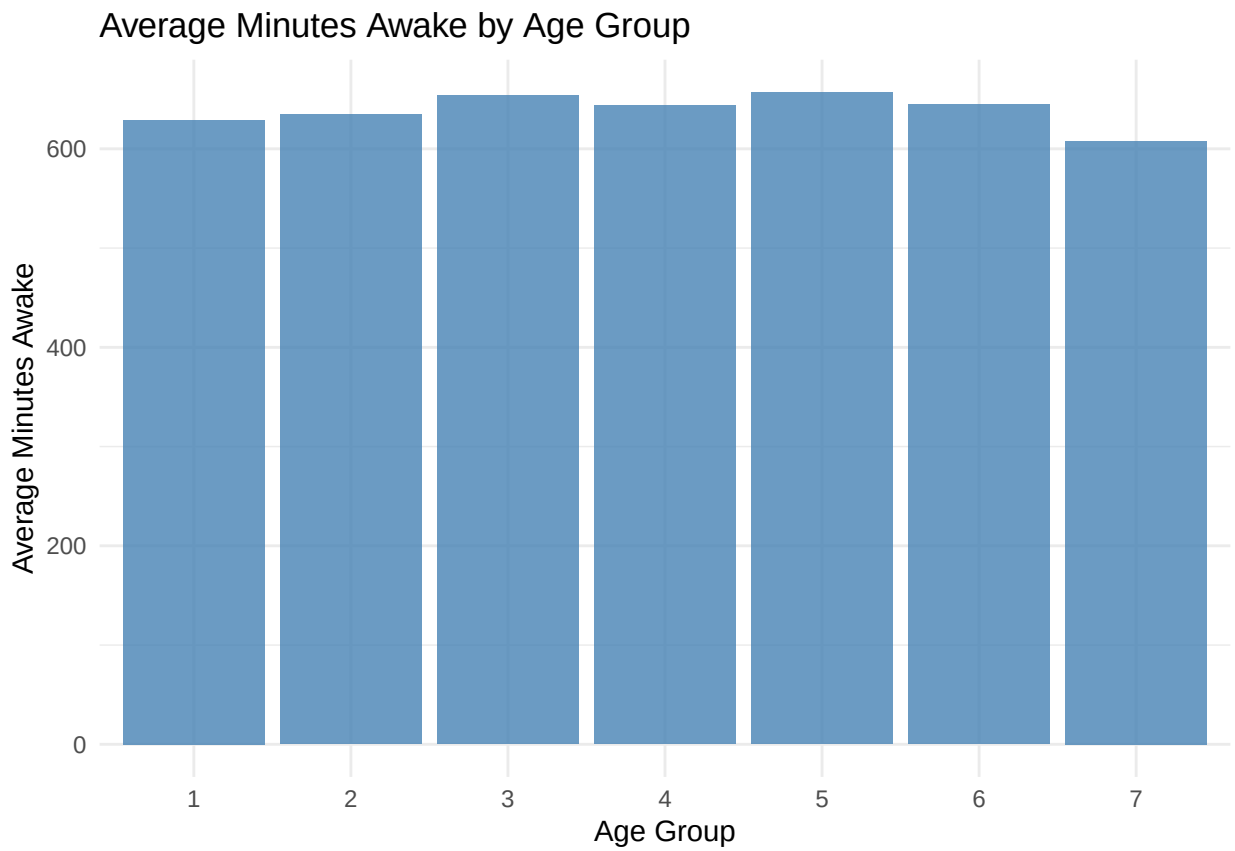
```r
average_minutes_awake <- final_data_df %>%
  group_by(Age) %>%
  summarise(
    total_minutes_awake = sum(Minutes.awake, na.rm = TRUE), # Total minutes awake
    age_count = n() # Count of observations per age
  ) %>%
  mutate(avg_minutes_awake = total_minutes_awake / age_count)
print(average_minutes_awake)
```

```
## # A tibble: 7 x 4
##     Age total_minutes_awake age_count avg_minutes_awake
##   <int>               <dbl>     <int>             <dbl>
## 1     1             145955.       232              629.
## 2     2             239800.       378              634.
## 3     3             205251.       314              654.
## 4     4             272257.       423              644.
## 5     5             196419.       299              657.
## 6     6             224928.       349              644.
## 7     7             717616.      1181              608.
```

```r
str(average_minutes_awake)
```

```
## tibble [7 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Age                : int [1:7] 1 2 3 4 5 6 7
##  $ total_minutes_awake: num [1:7] 145955 239800 205251 272257 196419 ...
##  $ age_count          : int [1:7] 232 378 314 423 299 349 1181
##  $ avg_minutes_awake  : num [1:7] 629 634 654 644 657 ...
```

```r
# Create a bar plot to show the distribution of total minutes awake by age
ggplot(average_minutes_awake, aes(x = factor(Age), y = avg_minutes_awake)) +
  geom_bar(stat = "identity", fill = "steelblue", alpha = 0.8) + # Bar plot
  labs(
    title = "Average Minutes Awake by Age Group",
    x = "Age Group",
    y = "Average Minutes Awake"
  ) +
  theme_minimal() +
  theme( )
```



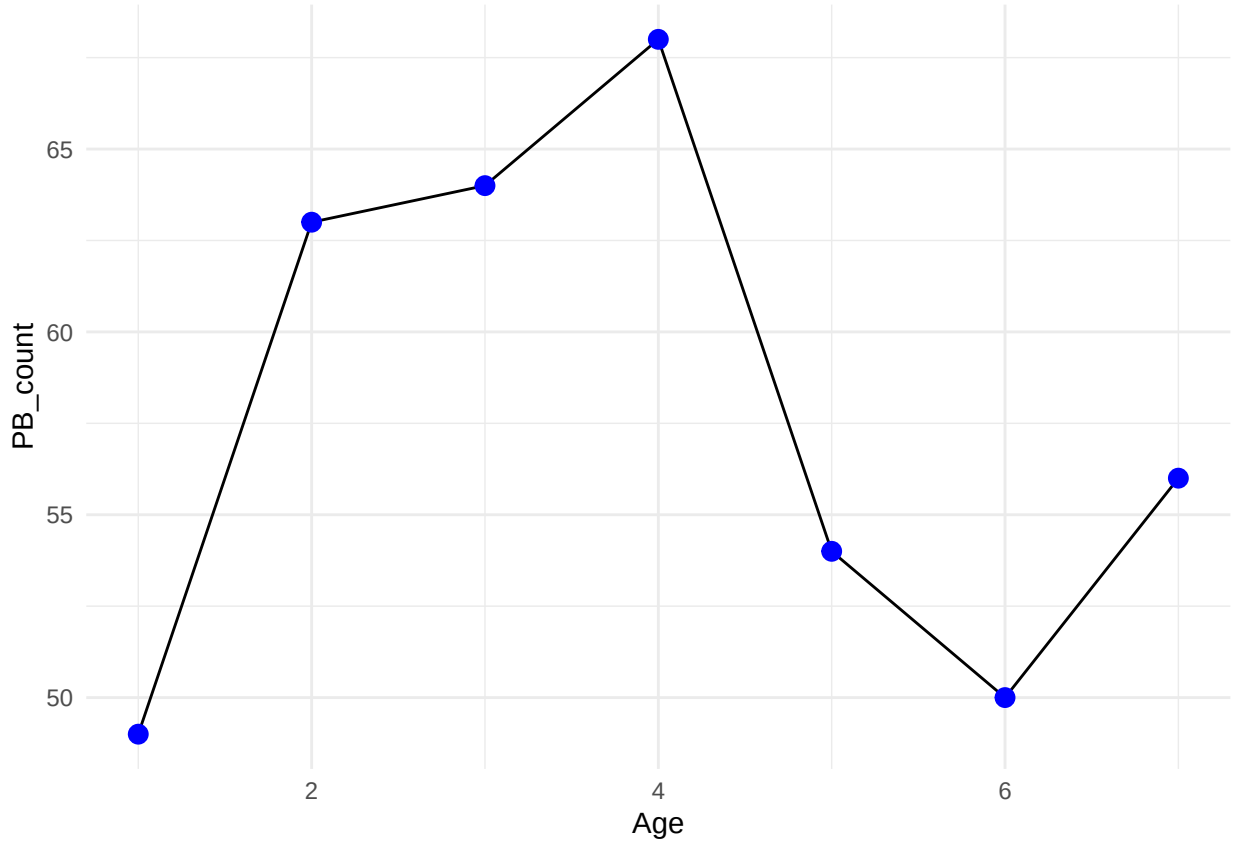
The average minutes awake is relatively consistent across all age groups, hovering around 630–660 minutes. There is no significant increase or decrease in the average minutes awake as age increases, suggesting that wakefulness may not be strongly age-dependent within this dataset.

Part 6: Find the Number of Unique Positional Behaviors for Each Age Group

```
result <- df %>%
  pivot_longer(cols = 7:ncol(df),
               names_to = "Behavior",
               values_to = "Value") %>%
  filter(Value != 0.0)

result <- result %>%
  group_by(Age) %>%
  summarise(PB_count = n_distinct(Behavior)) %>%
  arrange(desc(PB_count))

ggplot(result, aes(x=Age, y=PB_count)) +
  geom_line() +
  geom_point(color = "blue", size = 3) +
  theme_minimal()
```




```
print(result)
```

```
## # A tibble: 7 x 2
##     Age PB_count
##   <int>    <int>
## 1     4       68
## 2     3       64
## 3     2       63
## 4     7       56
## 5     5       54
## 6     6       50
## 7     1       49
```

From the table above, Age 4 group has 68 kinds of positional behaviors, which is the most number of distinct positional behaviors across all age groups. Age 1 group has only 49 kinds of positional behaviors, which is the least number of distinct behaviour among 7 groups.

Also, from the line graph, starting from age 2 to 4, there is a surge in the number of distinct positional behaviors, and after year 4, there is a significant decrease in number of positional behaviors, which means that year 4 is an important time point to focus on.