EDA Report

Xiaohan Shi, Zihao Zhang, Suheng Yao

2024-11-15

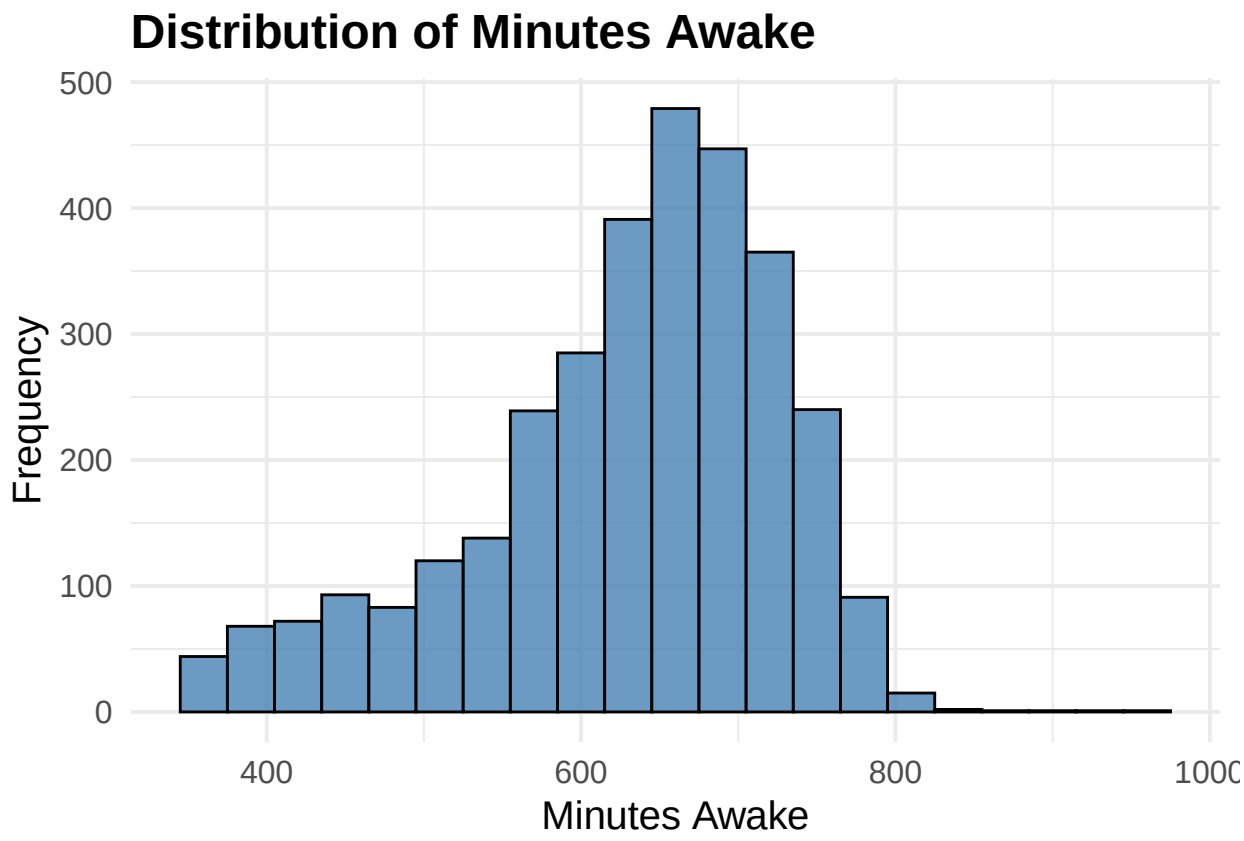
Read in the dataset

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

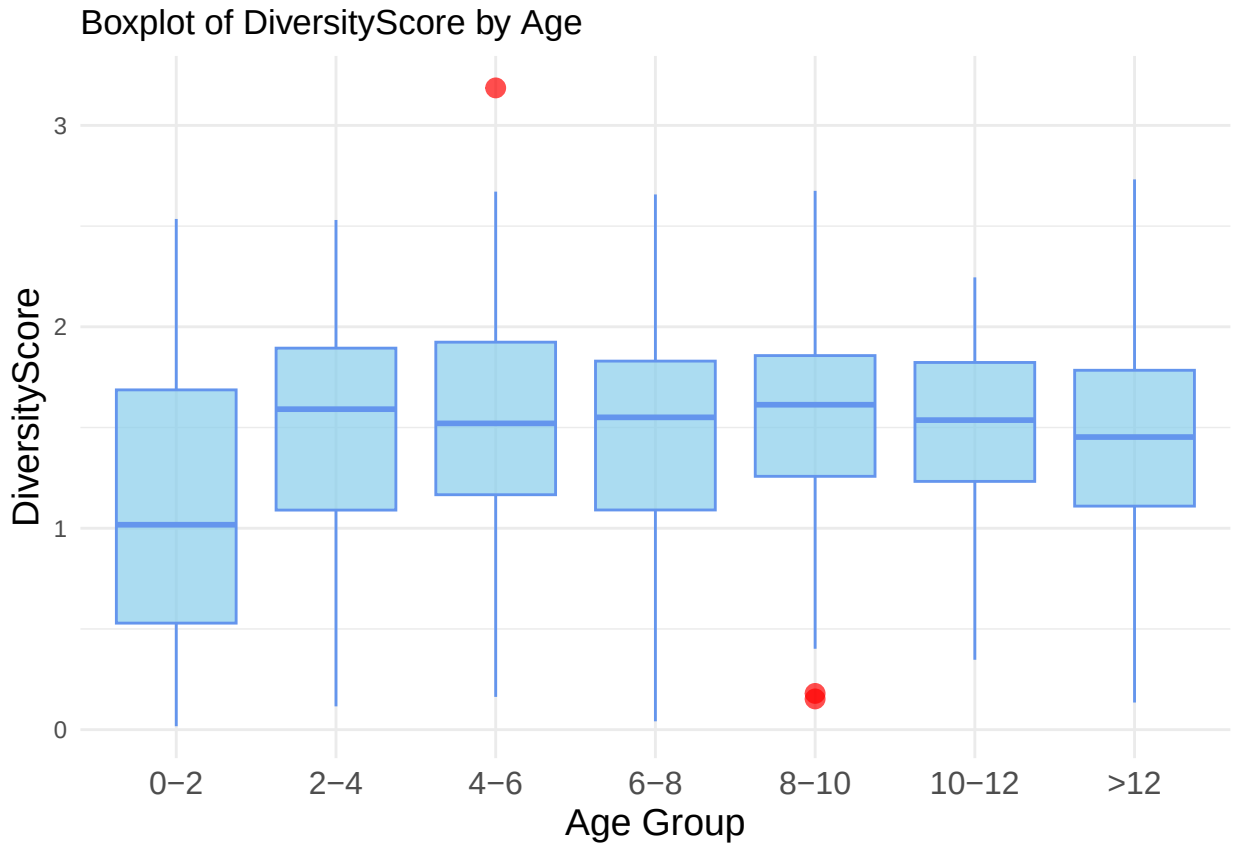
Some Data Description

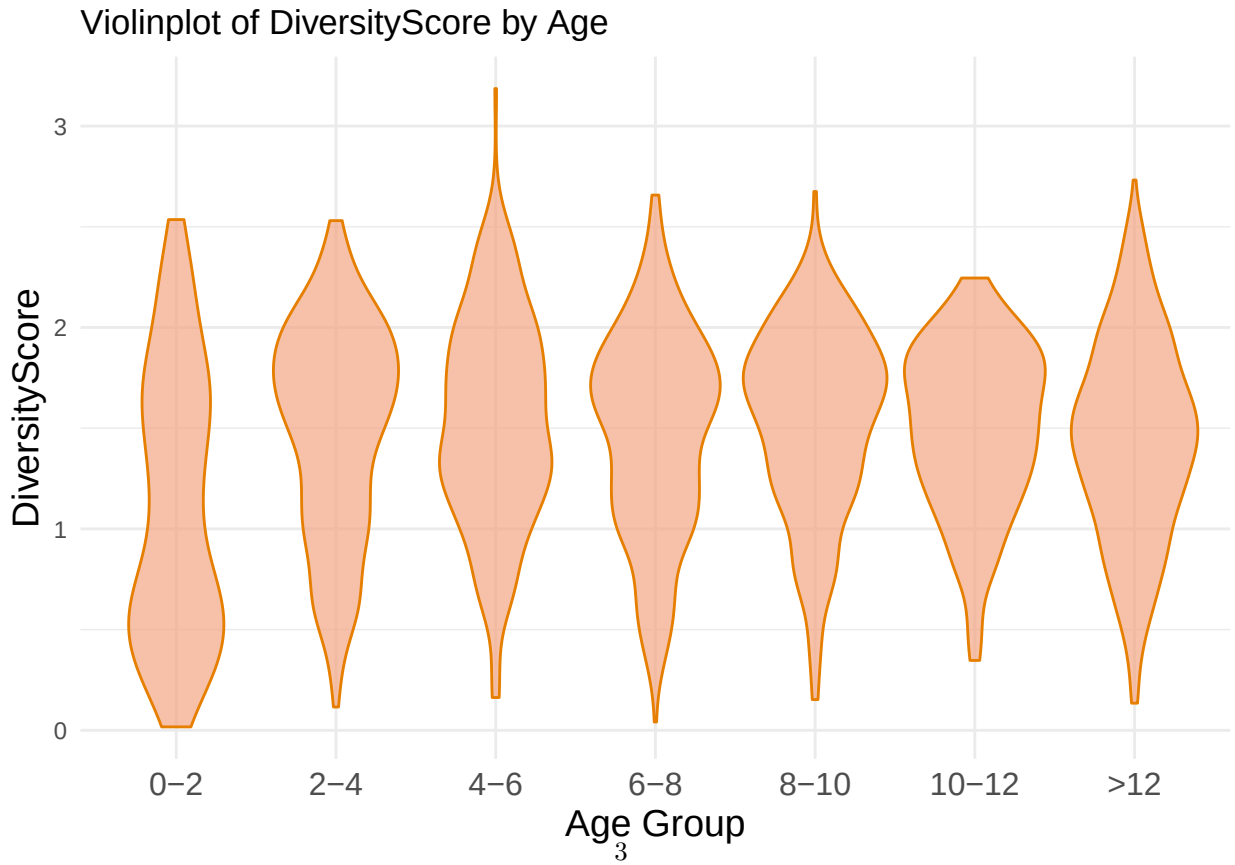
The client divided all the ages into 7 groups. In the original dataset, if the age is between 0 to 2, then those orangutans are classified as age level 1; if the age is between 2 to 4, then those orangutans are classified as age level 2; if the age is between 4 to 6, then those orangutans are classified 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. Also, all the minutes awake are greater than 360 minutes.

Part 1: The Distribution of Minutes Awake



Part 2: Visualization of Diversity Score





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.

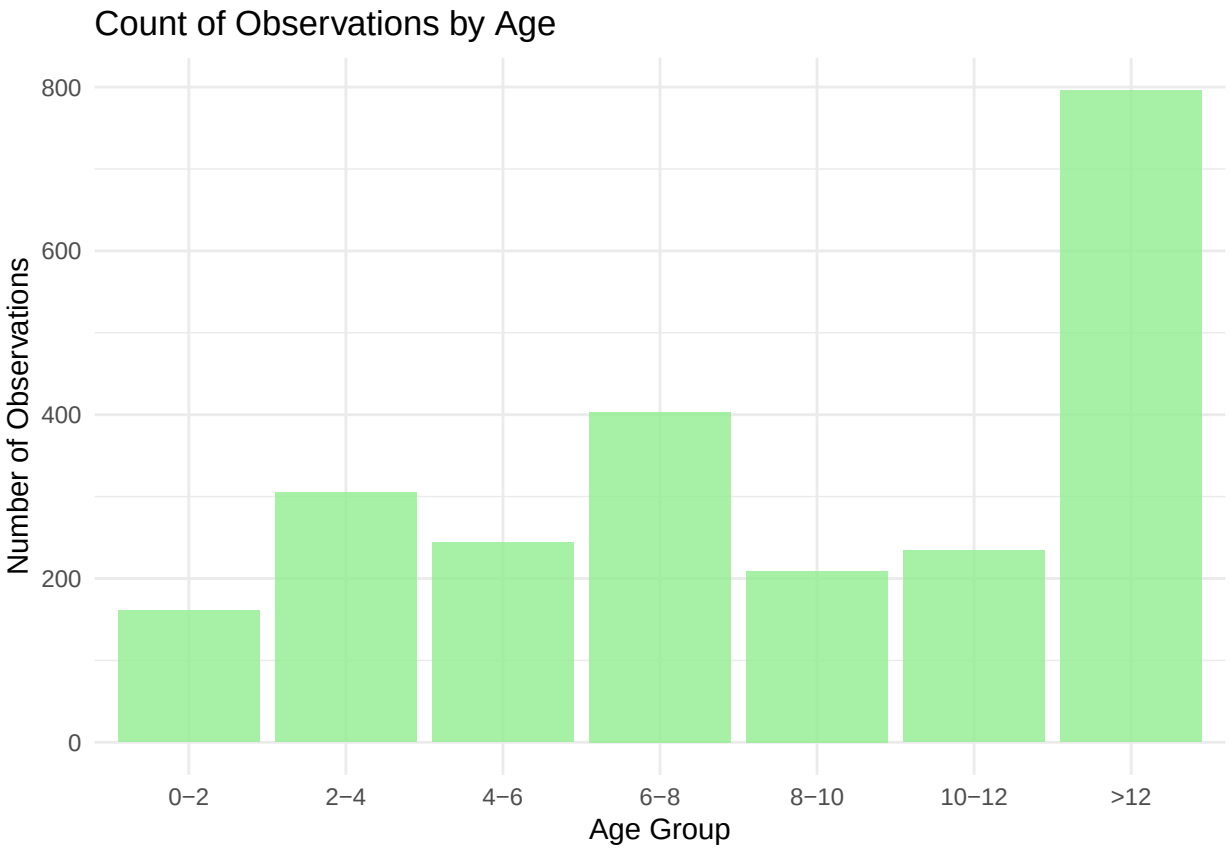
In the violinplot, the similar width of the violin illustrates that the sample size is relatively balanced across age groups due to, there is no particularly small or large sample group.

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.

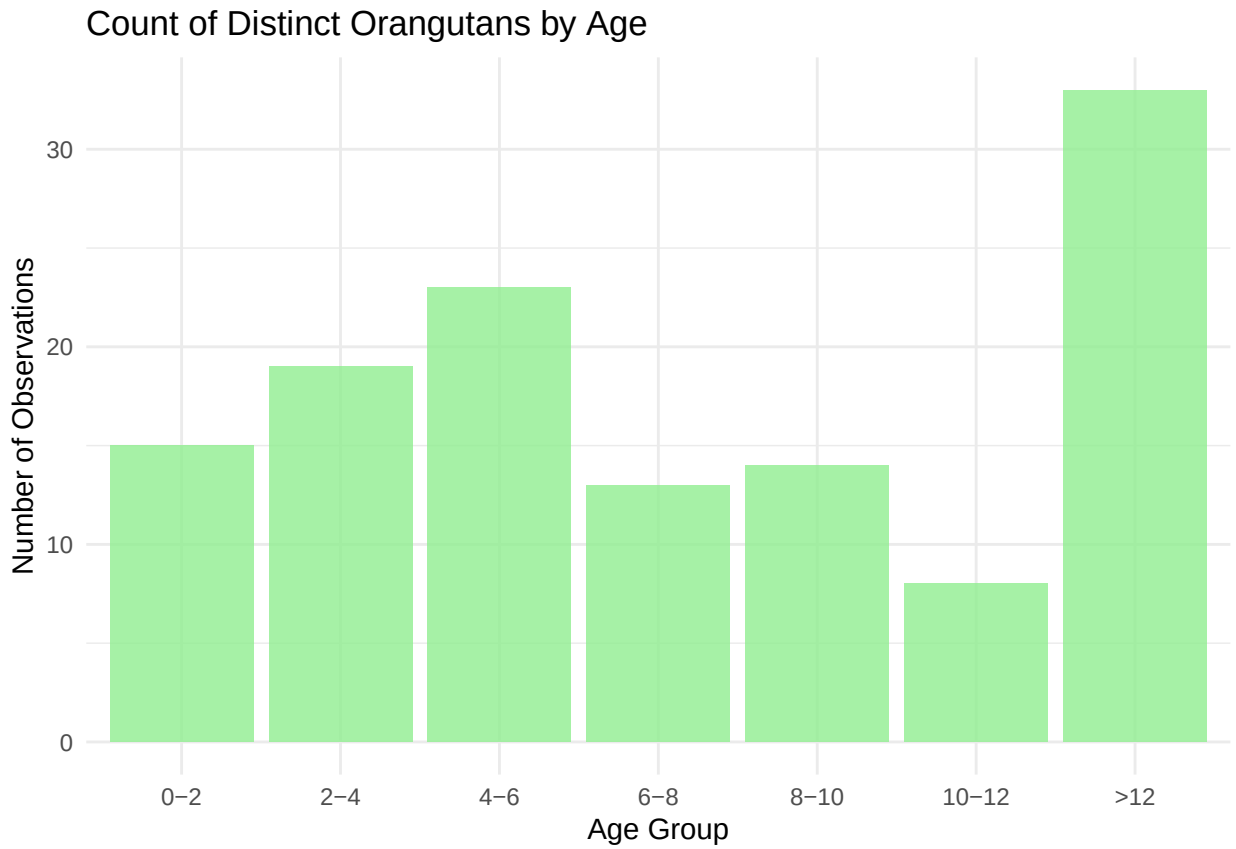
In addition, as the age get older, the variability of the data seems to decrease and the distribution becomes more concentrated.

Part 3: Diversity Score Statistics

```
## # 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
```



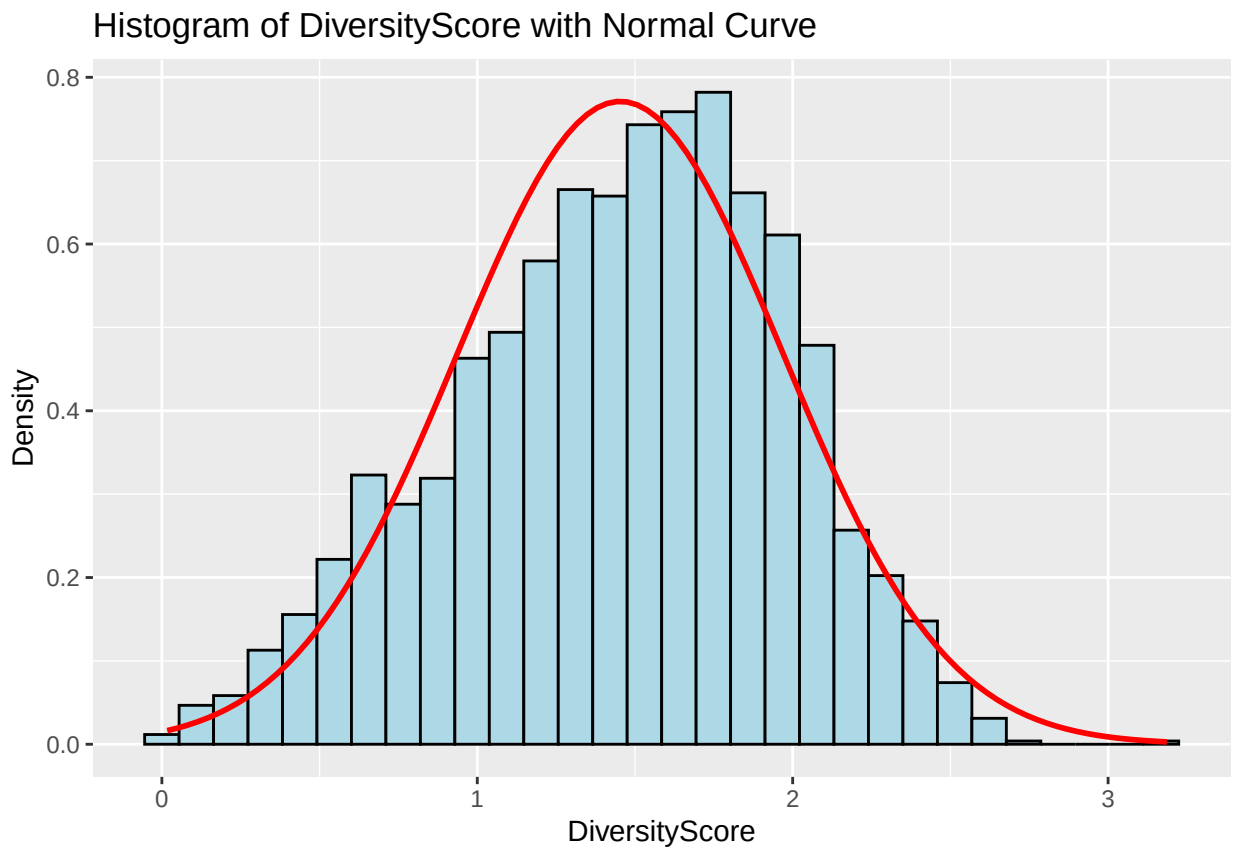
```
## # A tibble: 7 x 2
##     Age count
##   <int> <int>
## 1     1    15
## 2     2    19
## 3     3    23
## 4     4    13
## 5     5    14
## 6     6     8
## 7     7    33
```



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

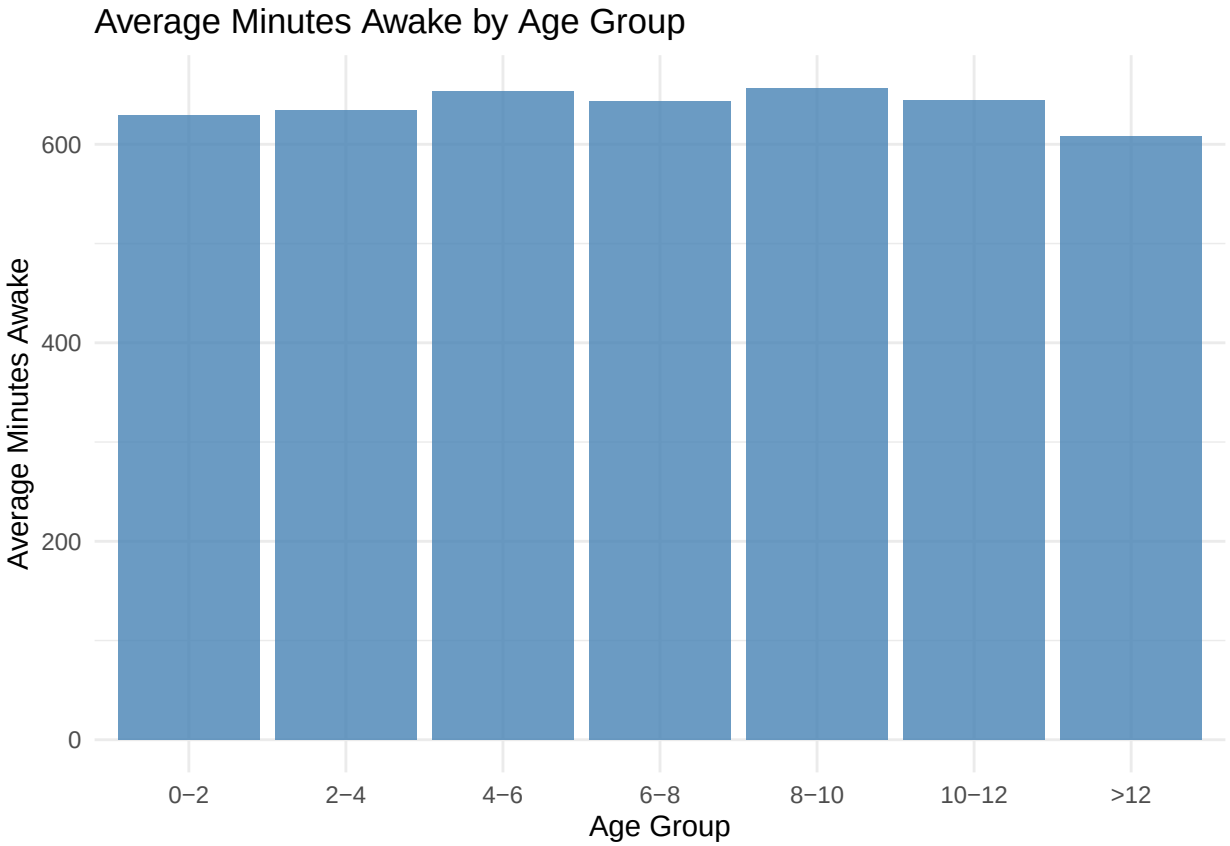


Since as shown in the histogram above, the diversity score follows a close to normal distribution, we can proceed to use ANOVA table to analyze the relationship between diversity score and age.

```
##               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
```

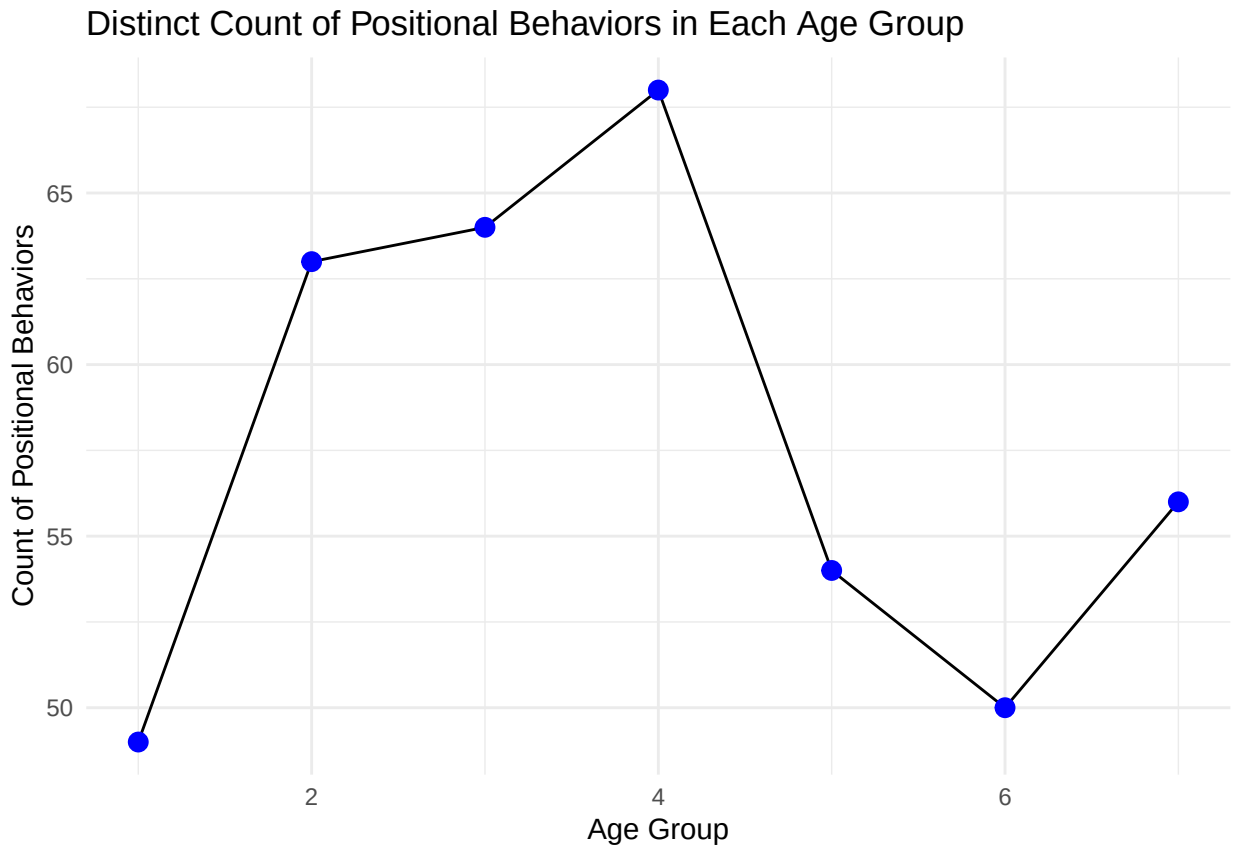
From the ANOVA table result, since p-value of f-test is less than 0.05, we can reject the null hypothesis, saying that there is statistically significant difference for diversity score between different age groups.

Part 5: Distribution of Total Minutes Awake



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




```
## # 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.