# Engage-Integrate

**Kristi Wu**

# Background on the data set

a. **Where did you find the data set? Provide the precise URL where we can find the data file.**

I found this data set on the kaggle site. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset?resource=download (https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset?resource=download)

b. **Write a brief description providing the context of the data set. Use your own words, but if there are parts that would be better expressed by the data source description itself, put those parts in quotes with clear attribution, including a reference to where you found that text. Be sure to describe what an observation in this data set represents and what variables are included and their meaning.**

An observation in this data set represents a single gym member. Each row contains information about the gym member, such as their gender, weight (kg), height (m), max BPM (beats per minute), average BPM, resting BPM (before the workout), duration of workout (hours), calories burnt, workout type, body fat percentage, water intake in liters during workout, how often they workout in a week, experience level (1-3; beginner to expert), and their BMI (body mass index).

Overall, this data set encompasses “physical attributes,” “key performance indicators,” “experience level,” and more of gym members.

c. **Load the data. Once the data is loaded, use the `glimpse` function to provide an initial look at the data frame.**

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.2     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
gym_members <- read_csv("gym_members_exercise_tracking.csv")
```

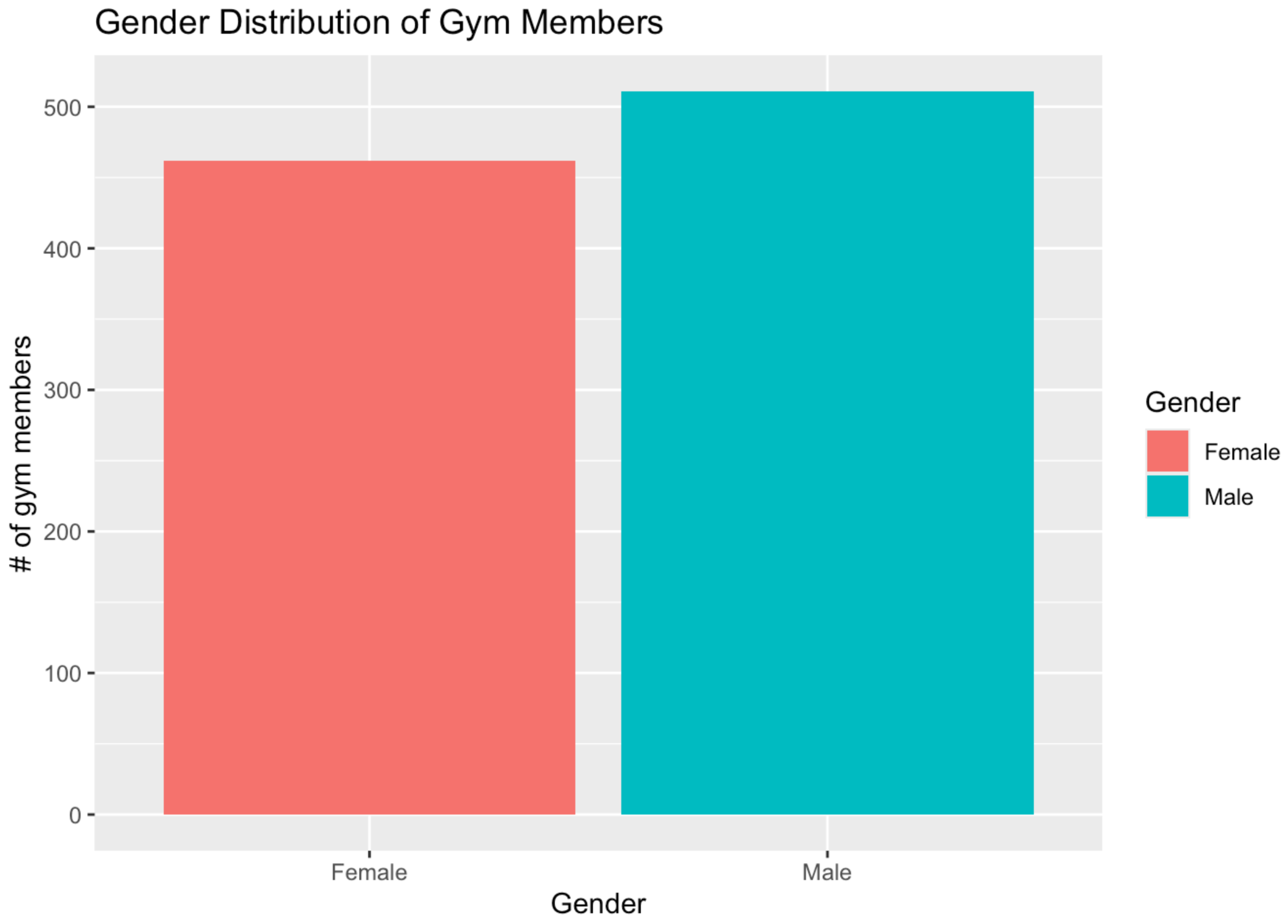
```
## Rows: 973 Columns: 15
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (2): Gender, Workout_Type
## dbl (13): Age, Weight (kg), Height (m), Max_BPM, Avg_BPM, Resting_BPM, Sessi...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(gym_members)
```

```
## Rows: 973
## Columns: 15
## $ Age                             <dbl> 56, 46, 32, 25, 38, 56, 36, 40, 28, 28…
## $ Gender                          <chr> "Male", "Female", "Female", "Male", "M…
## $ `Weight (kg)`                   <dbl> 88.3, 74.9, 68.1, 53.2, 46.1, 58.0, 70…
## $ `Height (m)`                    <dbl> 1.71, 1.53, 1.66, 1.70, 1.79, 1.68, 1.…
## $ Max_BPM                         <dbl> 180, 179, 167, 190, 188, 168, 174, 189…
## $ Avg_BPM                         <dbl> 157, 151, 122, 164, 158, 156, 169, 141…
## $ Resting_BPM                     <dbl> 60, 66, 54, 56, 68, 74, 73, 64, 52, 64…
## $ `Session_Duration (hours)`      <dbl> 1.69, 1.30, 1.11, 0.59, 0.64, 1.59, 1.…
## $ Calories_Burned                 <dbl> 1313, 883, 677, 532, 556, 1116, 1385, …
## $ Workout_Type                    <chr> "Yoga", "HIIT", "Cardio", "Strength", …
## $ Fat_Percentage                  <dbl> 12.6, 33.9, 33.4, 28.8, 29.2, 15.5, 21…
## $ `Water_Intake (liters)`         <dbl> 3.5, 2.1, 2.3, 2.1, 2.8, 2.7, 2.3, 1.9…
## $ `Workout_Frequency (days/week)` <dbl> 4, 4, 4, 3, 3, 5, 3, 3, 4, 3, 2, 3, 3,…
## $ Experience_Level                <dbl> 3, 2, 2, 1, 1, 3, 2, 2, 2, 1, 1, 2, 2,…
## $ BMI                             <dbl> 30.20, 32.00, 24.71, 18.41, 14.39, 20.…
```

d. Make a separate plot for each variable in your data set. Each plot should involve just one variable at a time. If your data set has a very large number of variables, choose the three that you think are most interesting to you. Think about what type of variable each is, and use this to decide on what kind of plot would be most effective.

```
# 1) Gender (categorical)
ggplot(gym_members, aes(x=Gender, fill=Gender)) +
  geom_bar() +
  labs(x="Gender",
      y = "# of gym members",
      title="Gender Distribution of Gym Members")
```



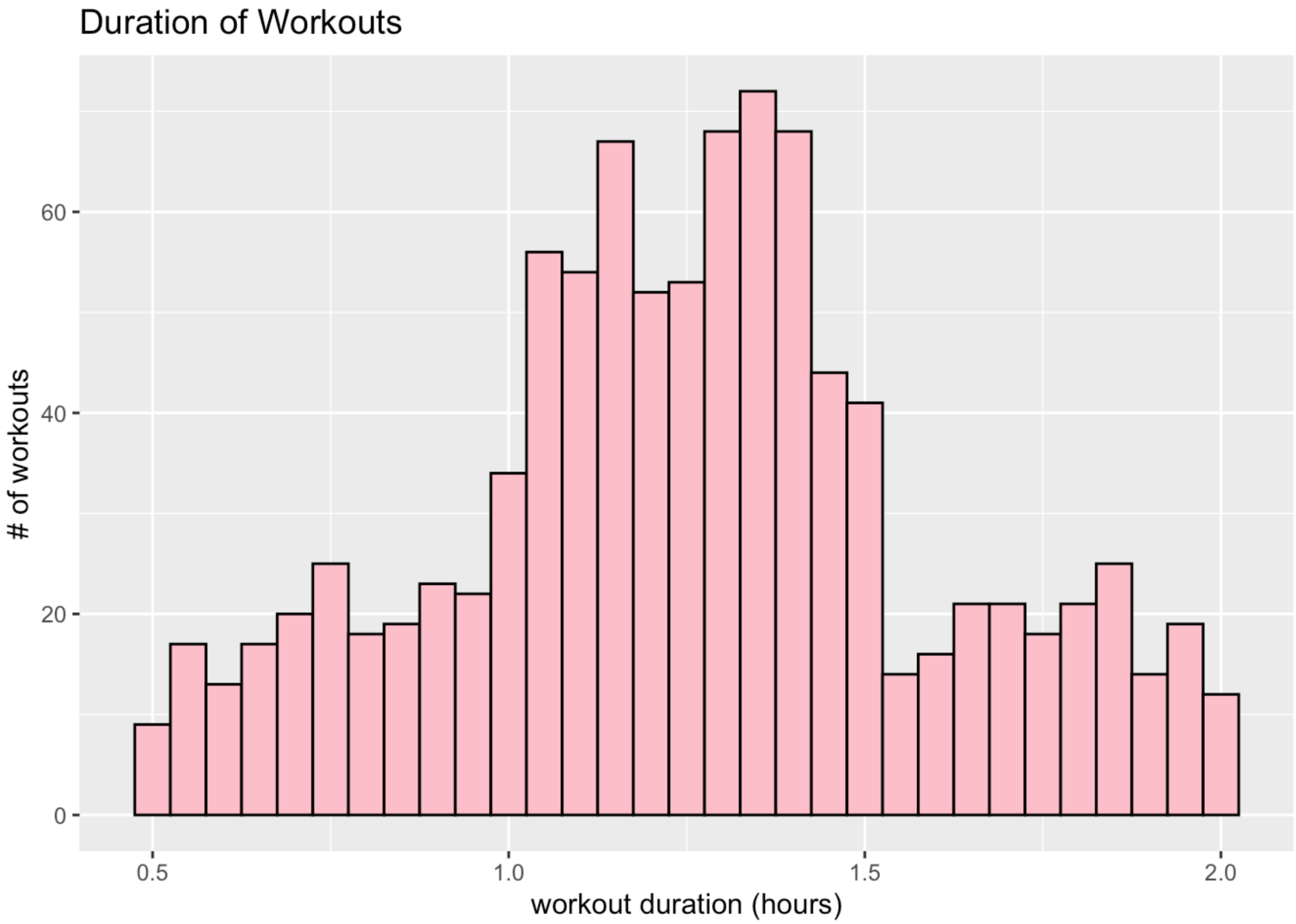
```
# verify how many more males there are
gym_members %>%
  group_by(Gender) %>%
  summarize(count = n())
```

```
## # A tibble: 2 × 2
##   Gender count
##   <chr>  <int>
## 1 Female   462
## 2 Male     511
```

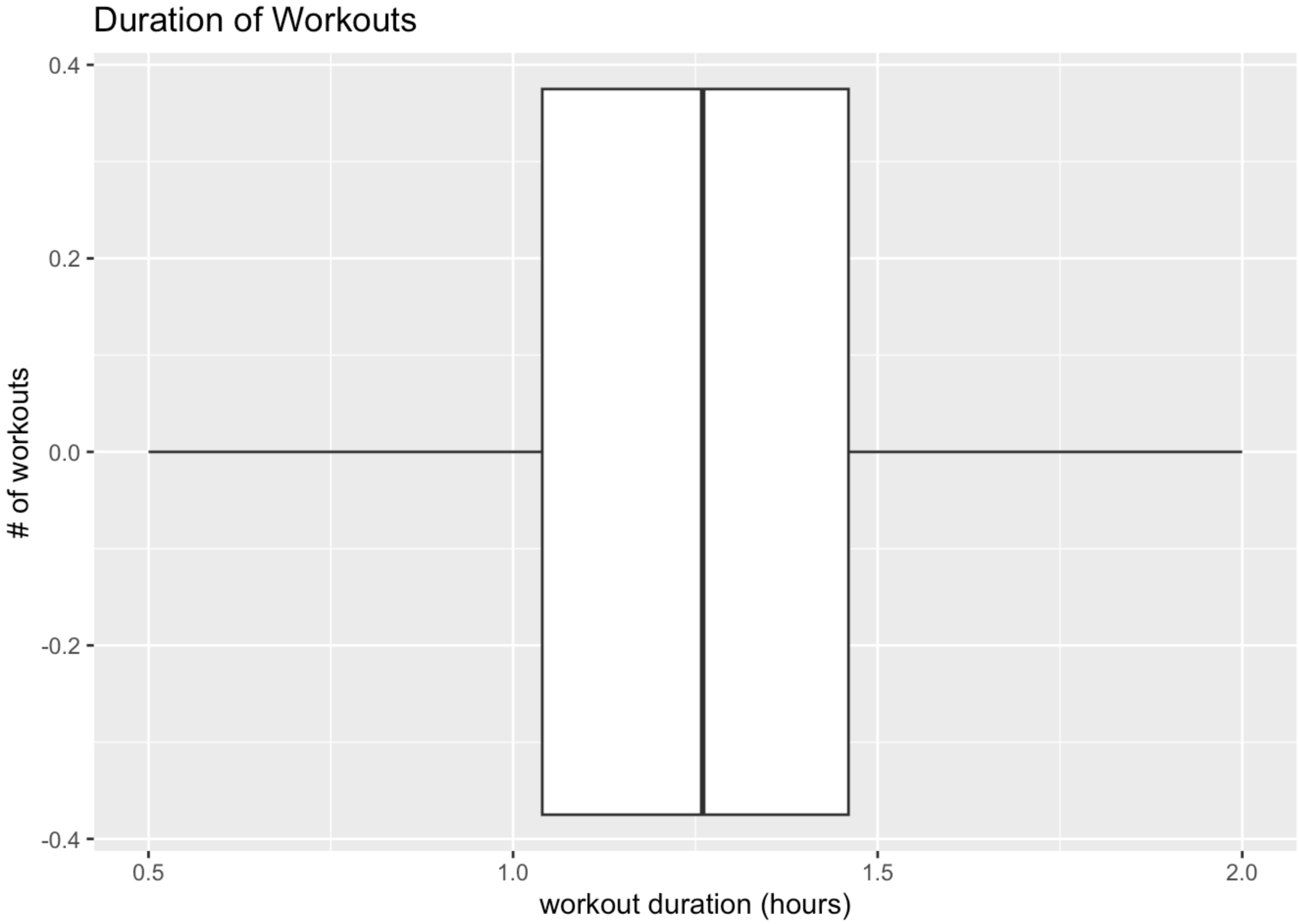
```
# check to see if there is missing data for Gender
gym_members %>%
  mutate(missing_data = is.na(Gender)) %>%
  summarize( num_na = sum(missing_data))
```

```
## # A tibble: 1 × 1
##   num_na
##    <int>
## 1      0
```

```
#2 Session_Duration (hours) (numerical)
# each bin represents 3 minutes
ggplot(gym_members, aes(x=`Session_Duration (hours)`)) +
  geom_histogram(binwidth=0.05, fill="pink", color="black") +
  labs(x="workout duration (hours)",
       y="# of workouts",
       title="Duration of Workouts")
```



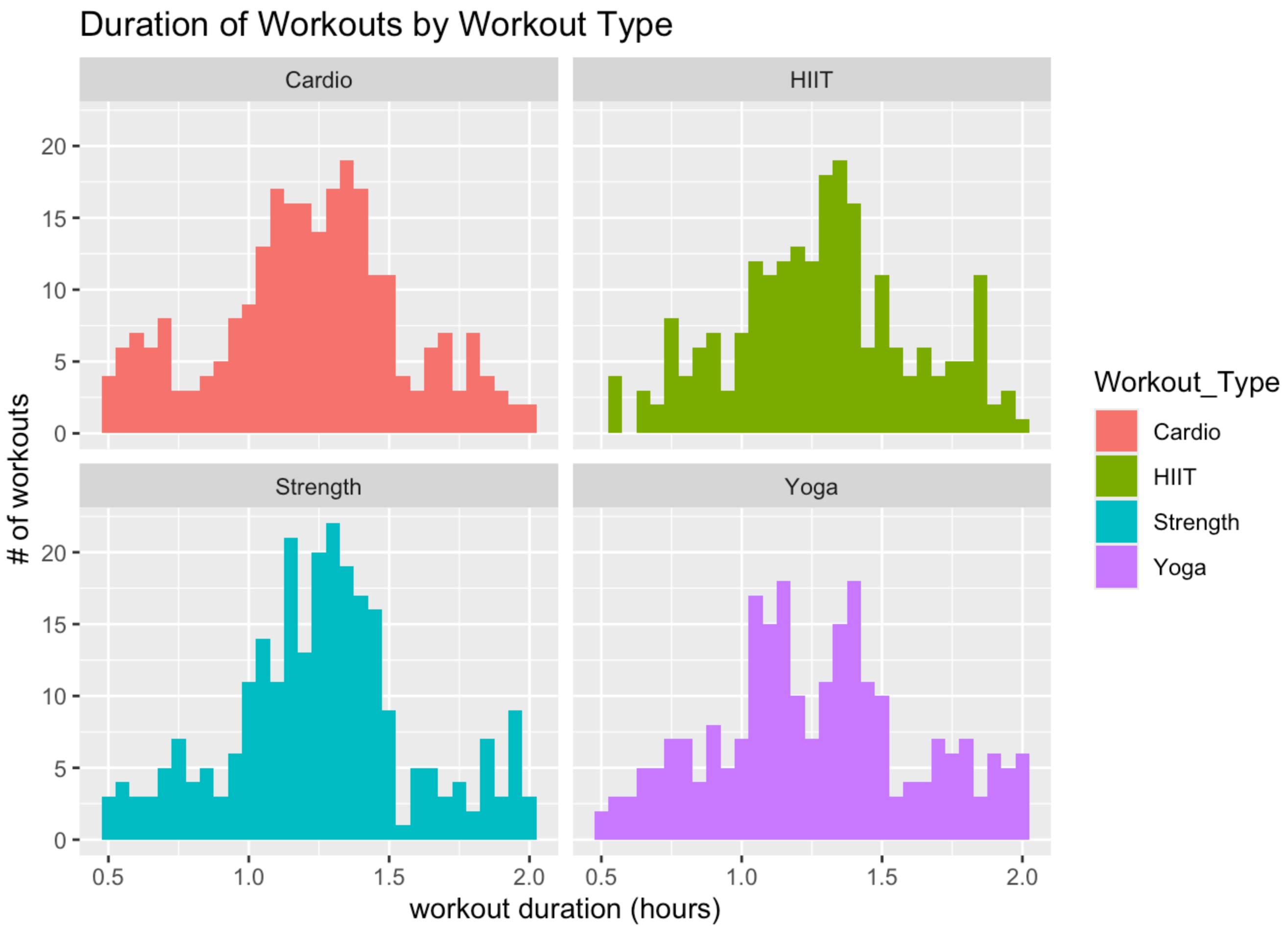
```
# boxplot
ggplot(gym_members, aes(x=`Session_Duration (hours)`)) +
  geom_boxplot() +
  labs(x="workout duration (hours)",
       y="# of workouts",
       title="Duration of Workouts")
```



```
# boxplot info
gym_members %>%
  summarize(min = min(`Session_Duration (hours)`),
            q1 = quantile(`Session_Duration (hours)`, 0.25),
            median = quantile(`Session_Duration (hours)`, 0.5),
            q3 = quantile(`Session_Duration (hours)`, 0.75),
            max=max(`Session_Duration (hours)`))
```

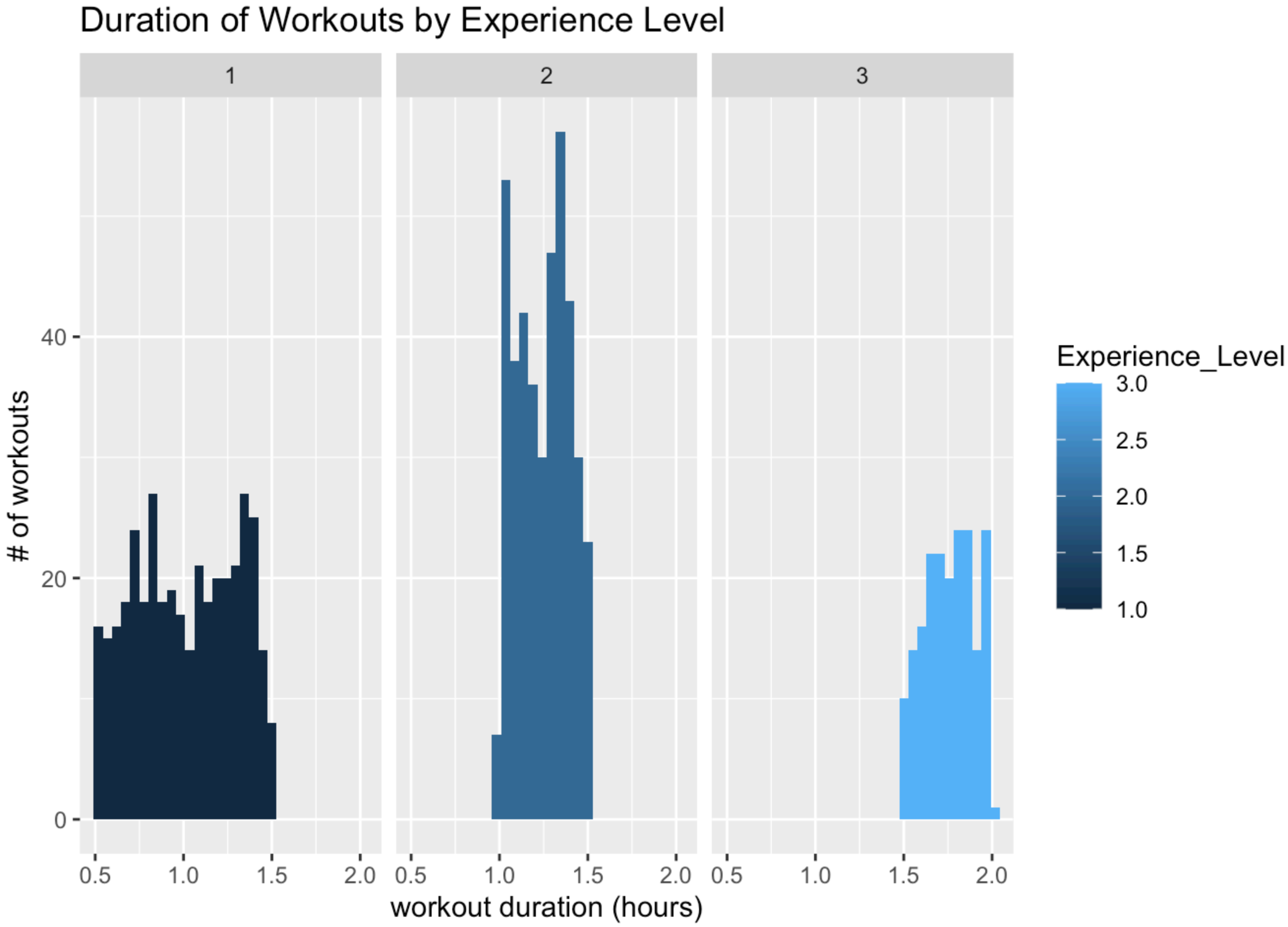
```
## # A tibble: 1 × 5
##     min    q1 median    q3   max
##   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1   0.5  1.04   1.26  1.46     2
```

```
# Distribution of workout duration across workout type
ggplot(gym_members, aes(x=`Session_Duration (hours)`, fill=Workout_Type)) +
  geom_histogram(binwidth=0.05) +
  facet_wrap(~Workout_Type) +
  labs(x="workout duration (hours)",
       y="# of workouts",
       title="Duration of Workouts by Workout Type")
```

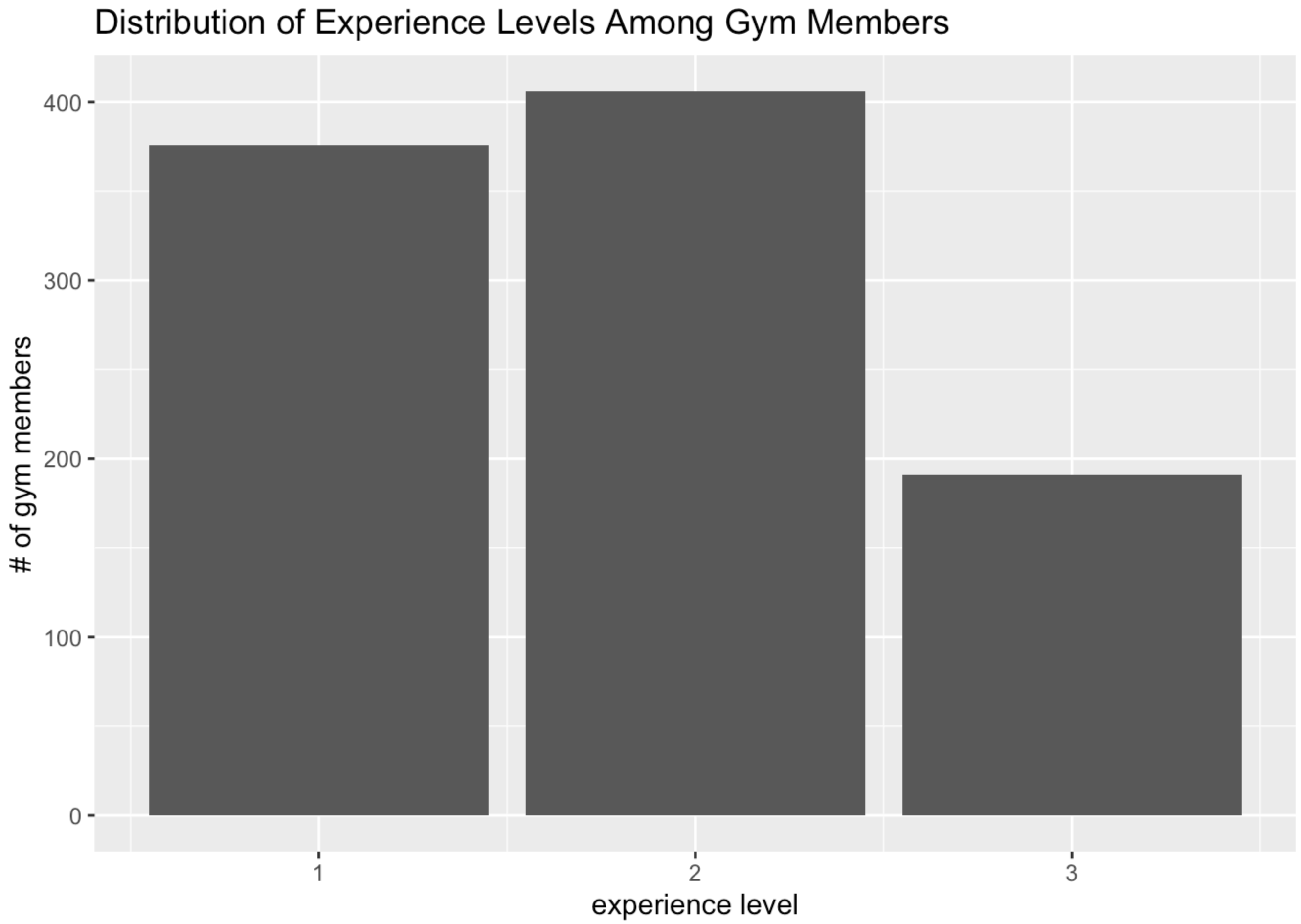


```
# Distribution of workout duration across experience level
ggplot(gym_members, aes(x=`Session_Duration (hours)`, fill=Experience_Level)) +
  geom_histogram()+
  facet_wrap(~ Experience_Level) +
  labs(x="workout duration (hours)",
       y="# of workouts",
       title="Duration of Workouts by Experience Level")
```

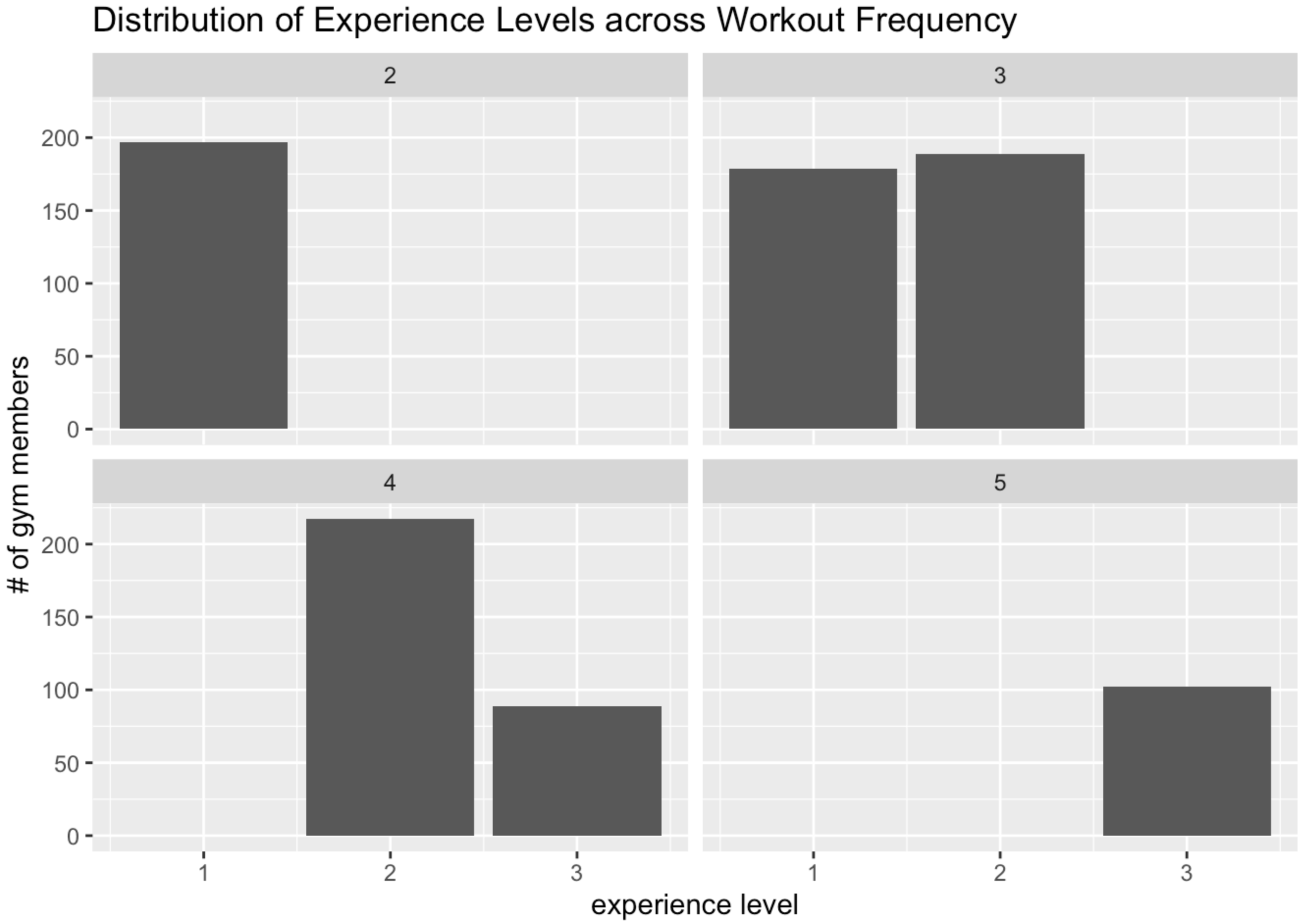
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



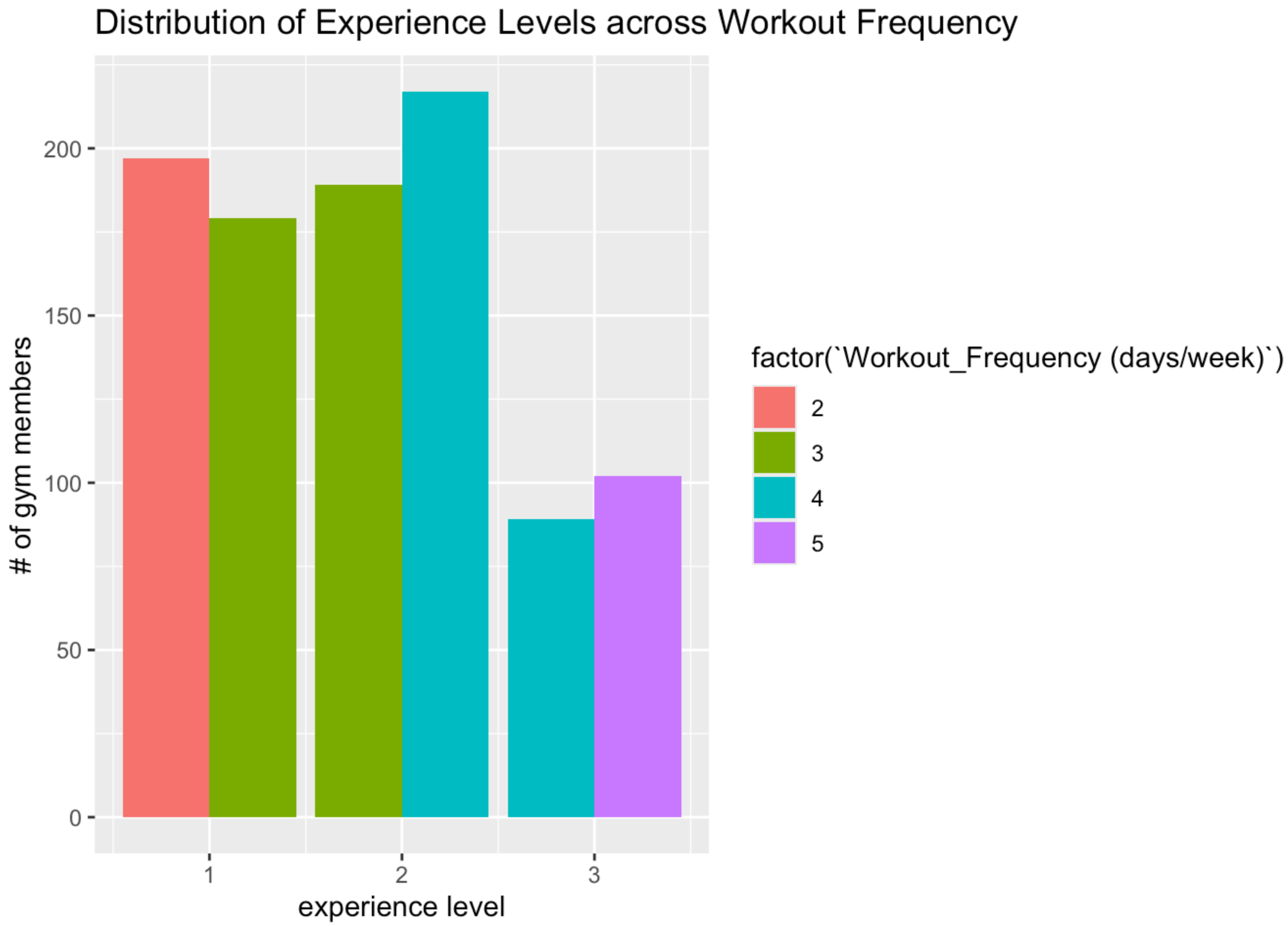
```
#3 Experience_Level (categorical, ordinal)
ggplot(gym_members, aes(x=Experience_Level)) +
  geom_bar() +
  labs(x="experience level",
       y="# of gym members",
       title="Distribution of Experience Levels Among Gym Members")
```



```
# Distribution of Experience Level across Workout Frequency
# more experienced gym members tend to go more often
ggplot(gym_members, aes(x=Experience_Level)) +
  geom_bar() +
  facet_wrap(~`Workout_Frequency (days/week)`) +
  labs(x="experience level",
       y="# of gym members",
       title="Distribution of Experience Levels across Workout Frequency")
```

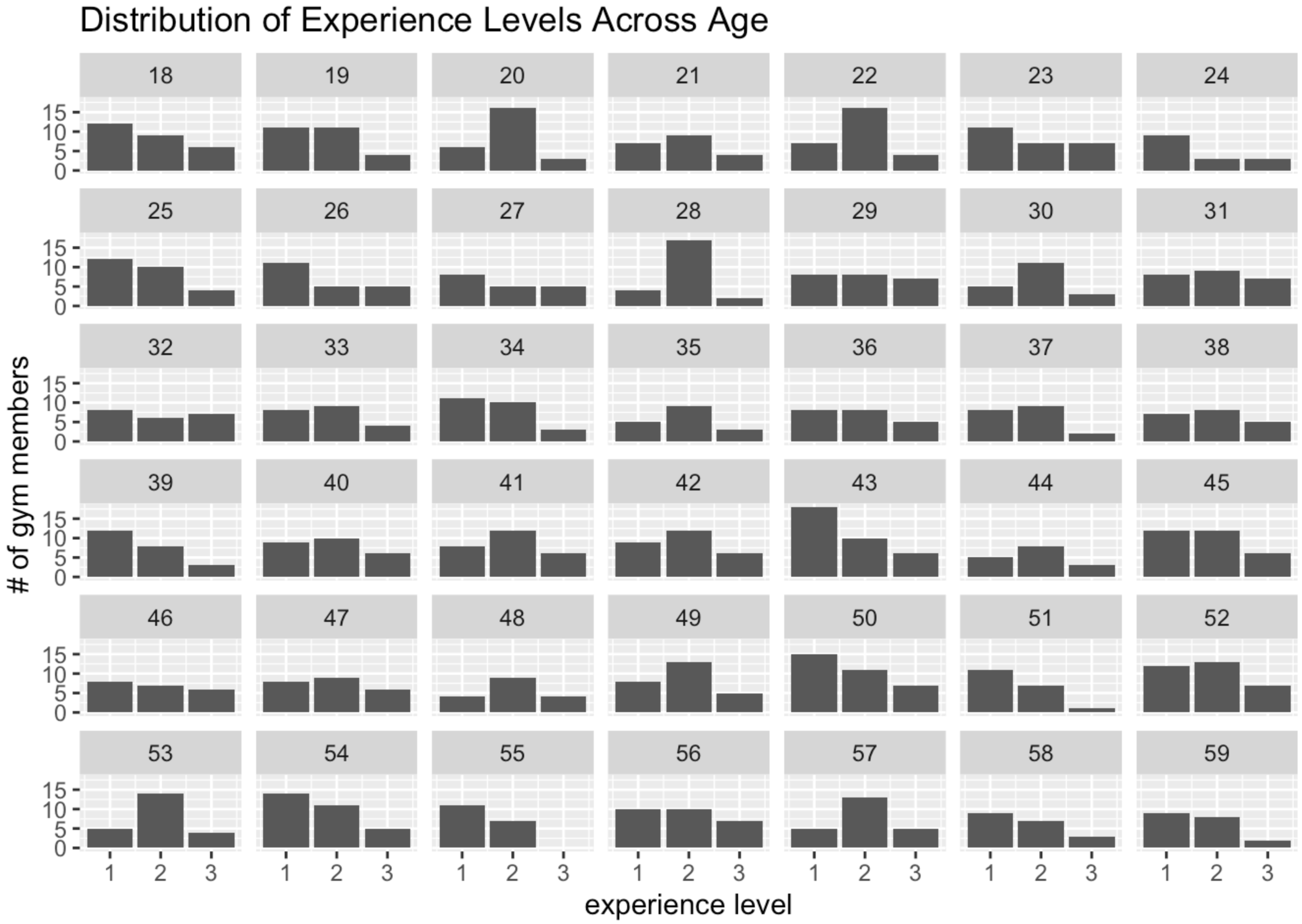


```
# another visualization
ggplot(gym_members, aes(x=Experience_Level, fill=factor(`Workout_Frequency (days/week)
`))) +
  geom_bar(position="dodge") +
  labs(x="experience level",
       y="# of gym members",
       title="Distribution of Experience Levels across Workout Frequency")
```

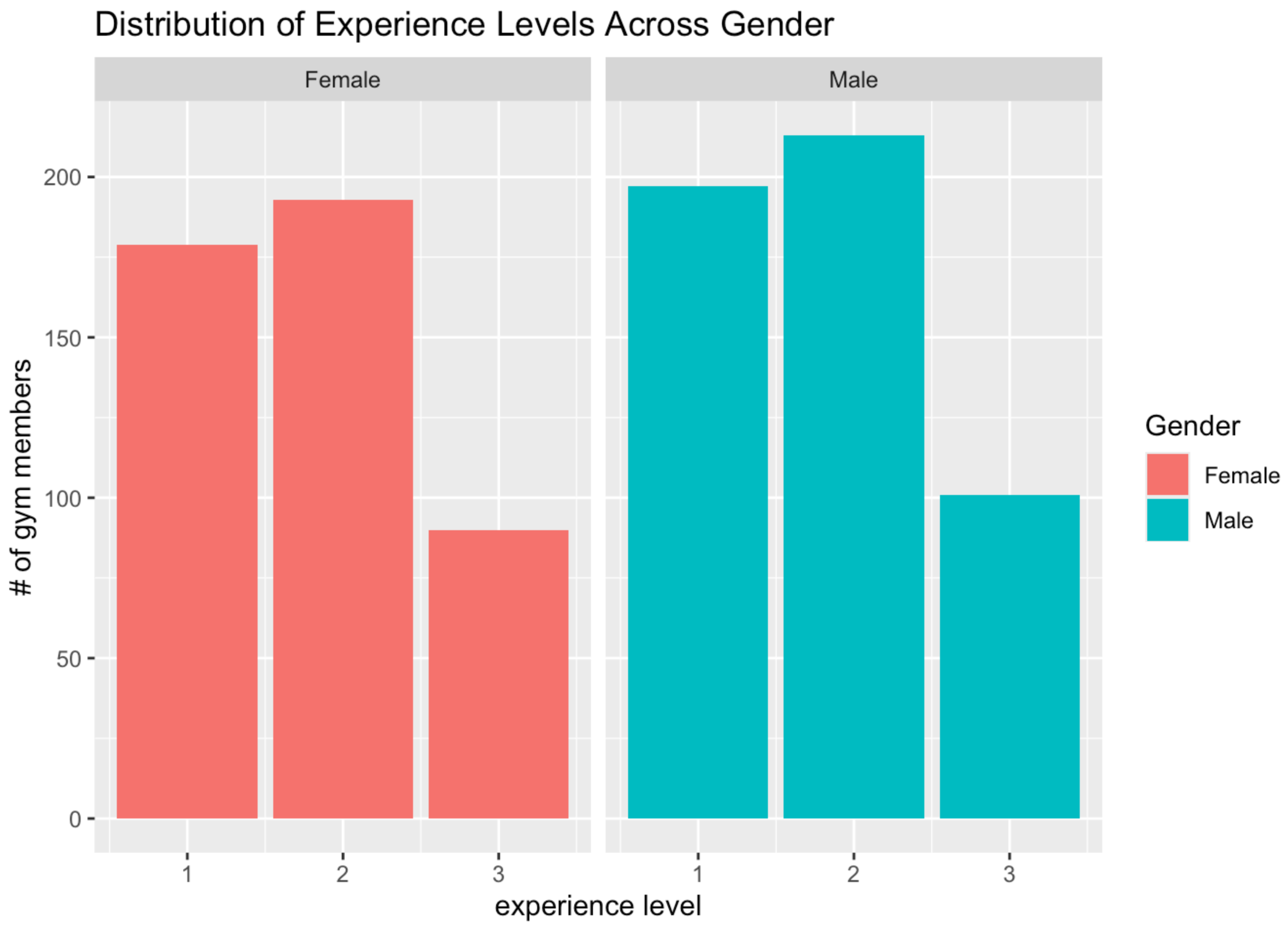


Distribution of Experience Levels across Workout Frequency

```r
# Distribution of Experience Level across Age 18-59
ggplot(gym_members, aes(x=Experience_Level)) +
geom_bar() +
facet_wrap(~Age)+
labs(x="experience level",
     y="# of gym members",
     title="Distribution of Experience Levels Across Age")
```



```
# Distribution of Experience Level across Gender
ggplot(gym_members, aes(x=Experience_Level, fill=Gender)) +
geom_bar() +
facet_wrap(~Gender)+
labs(x="experience level",
     y="# of gym members",
     title="Distribution of Experience Levels Across Gender")
```



e. For each plot above, comment on what you see from the plot. Does what you see make sense? Are there any surprises? If there are surprises, explain how you have investigated these surprises (think about how in PS3 we explained all the surprises involving the duration variable).

1. The distribution of gym members by Gender revealed that there more male gym members in this data set, which I later verified that there are 49 more male gym members than female. There wasn't anything too surprising about this. I will just have to keep in mind that there is not an equal number of males and females when calculating summary statistics of average BMI, frequency of workout, etc. Also, I verified that there were no "NA" entries for Gender. In other words, there is no missing data for Gender.

2. The histogram showing the distribution workout duration seemed to have around 3-4 modes, with a majority being between 1 hour and 1.5 hour range. I further made a boxplot and calculated the Q1, median, and Q3 to get a sense of the distribution of the workout durations. 50% of workouts are between 0.5 and 1.26 minutes, and the upper half fall between 1.26 and 2 hours.

I also wondered if the workout duration varied across workout types, and did facet wrap by Workout_Type. All 4 workout types (cardio, HIIT, strength, yoga) seemed to have more than one mode still. Majority of workouts across different workout types still fell between 1 to 1.5 hours. Cardio has a small peak in the lower range between 30-45 min.

When I facet the distribution of workout duration by experience level, this revealed significant information. The more experienced gym members are, the longer their workout is. The distribution for beginners (1) is bimodal and falls within 0.5-1.5 hours Intermediate (2) gym members spend between 1 - 1.5 hours, and lastly advanced (3) gym members spend 1.5-2 hours at the gym.

3. The distribution of experience levels among gym members was a bit surprising because it shows that there are more beginner- and intermediate-level members than experienced members. By faceting by

Workout_Frequency, the plots revealed that more experienced gym members tend to go to the gym more days out of the week than less experienced gym members. More specifically, expert gym members go to the gym 4-5 days per week, while intermediate gym members will go 3-4 times a week. Lastly, beginner gym members only go 2-3 days per week.

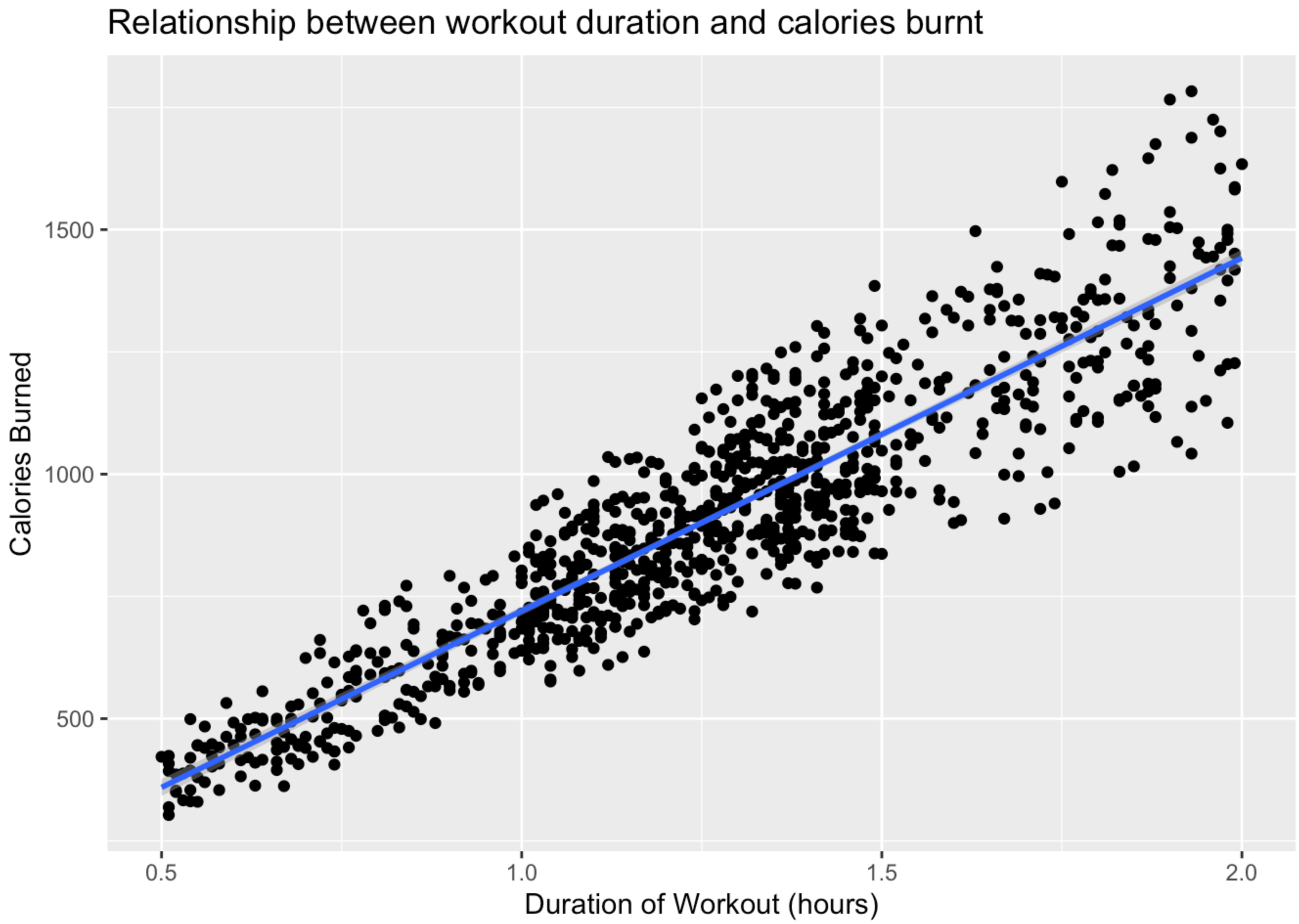
I wanted to see if age affected the experience level, but the distributions didn’t reveal any significant trends. Some things I noted were that most 18-year-olds were beginners and had descending experience level, while people that were 20, 22,28,49,53, and 57 were mostly intermediate level. No age group was significantly dominated by experts.

I did the same process for Gender to see if there were perhaps more Males that were more/less experienced. However, the plots reveal that the experience levels for Males and Females are very similar. Both gender groups have high amounts of beginner and intermediate level members, and their smallest group falls within the expert-level.

f. Make a plot that could help you understand whether there is an association between two or more variables.

```
ggplot(gym_members, aes(x= `Session_Duration (hours)`, y= Calories_Burned)) +
  geom_point()+
  geom_smooth(method=lm)+
  labs( x= "Duration of Workout (hours)",
        y = "Calories Burned",
        title = "Relationship between workout duration and calories burnt")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



g. Choose a numerical variable and take this as the response variable and fit a linear regression model to predict it based on one or more of the other variables.

```
library(moderndive)
fit <- lm(Calories_Burned ~ `Session_Duration (hours)`, data= gym_members)
get_regression_table(fit)
```

```
## # A tibble: 2 × 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                  -1.45      13.9    -0.104   0.917    -28.7     25.8
## 2 `Session_Duration (hou…   722.        13.9    67.6     0        701.     743.
```

h. Provide some interpretation of the model you have fit, by discussing the value of one or more of the coefficients that you estimated in your model in part (g). Are the coefficient values consistent with what you might have expected from your knowledge of the topic covered by the data set? (e.g. if it’s a data set about basketball, does the coefficient value match what you might have expected based on your knowledge of basketball?)

The y-intercept is -1.446, which has no practical interpretation in this context. Its purpose is to just set the baseline. The interpretation of the slope is for every 1 unit increase in Session_Duration (hours), there’s an associated increase of, on average, 721.786 Calories burned. This is pretty surprising because I didn’t expect gym members to burn more than 700 calories during an hour-long workout.