p8105_hw3_mc5698

2024-10-09

#Problem 1

```
#Access to the dataset
data("ny_noaa")
```

```
#clean the dataset
data_ny = ny_noaa %>%
  janitor::clean_names() |>
  drop_na(tmin,tmax,date)|>
  separate(date, into = c("year", "month", "day"), convert = TRUE)|>
  mutate(
    tmax = as.numeric(tmax) / 10,
    tmin = as.numeric(tmin) / 10,
    prcp = prcp / 10,
    snow = as.numeric(snow)
  )
```

```
snowfall= data_ny %>%
  filter(!is.na(snow)) %>%
  count(snow)%>%
  arrange(desc(n))
head(snowfall,10)
```

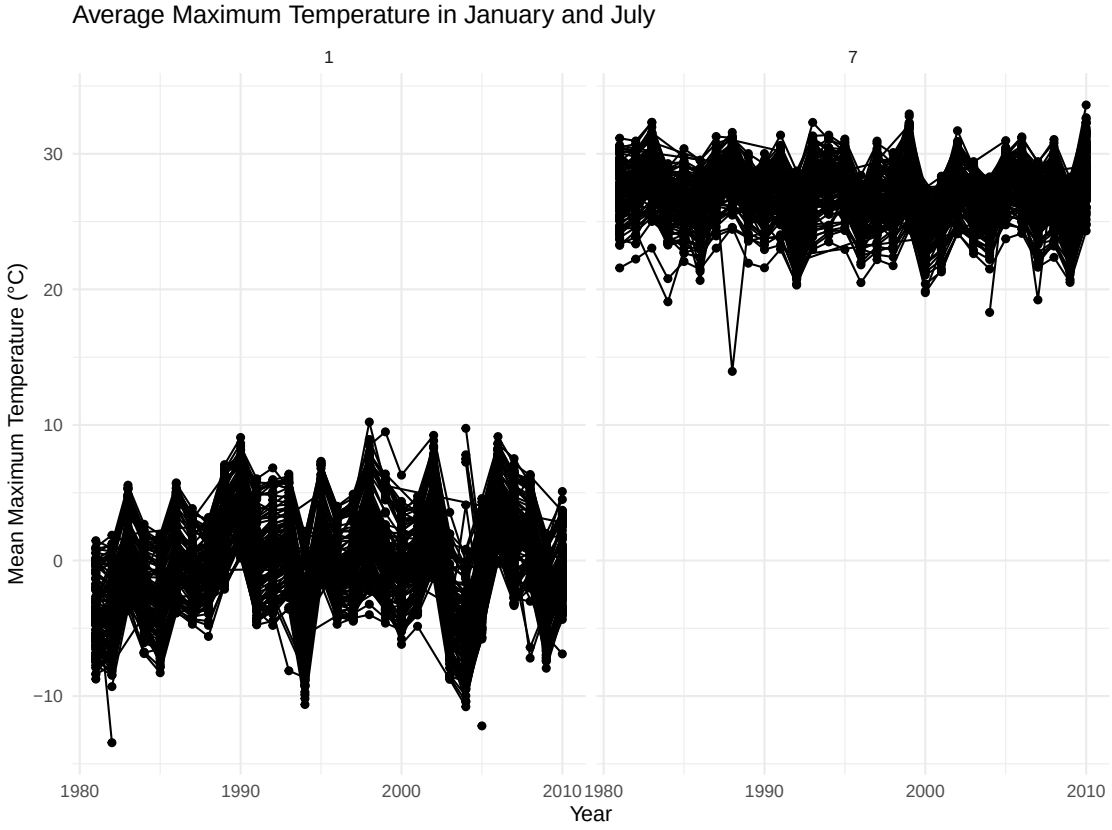
```
## # A tibble: 10 x 2
##     snow       n
##    <dbl>   <int>
##  1     0 1167149
##  2    25   17542
##  3    13   13704
##  4    51   10352
##  5     5    5960
##  6    76    5894
##  7     8    5777
##  8     3    5614
##  9    38    5578
## 10   102    3743
```

For snowball,the most commonly observed values is 0, which means that most days are not snowy.There are some other common values such as 25, 13, 51 in unit of mm.

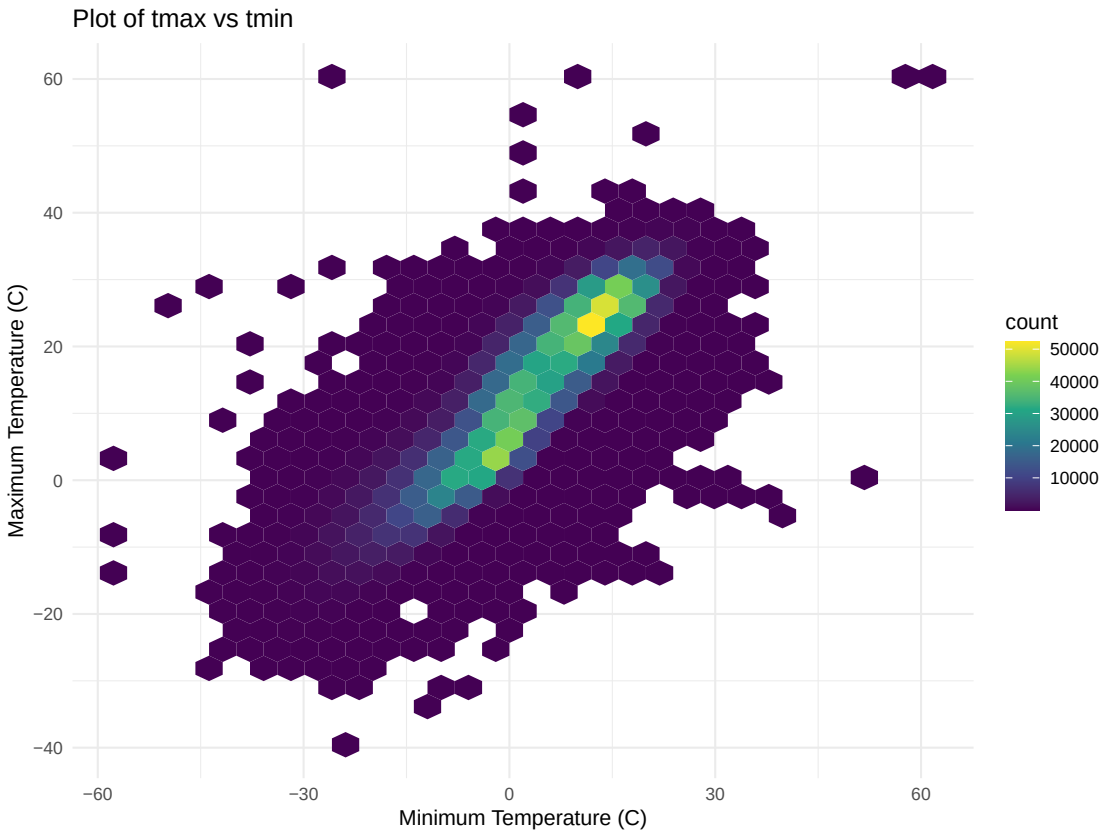
```
#Make a two-panel plot showing the average max temperature in January and in July
data_ny %>%
  filter(month %in% c(1, 7)) %>%
```

```
  group_by(id, year, month) %>%
  summarize(mean_tmax = mean(tmax, na.rm = TRUE, color = id)) %>%
  ggplot(aes(x = year, y = mean_tmax, group = id)) +
  geom_point() + geom_path() +
  facet_grid(~month) +
  labs(title = "Average Maximum Temperature in January and July",
       x = "Year",
       y = "Mean Maximum Temperature (°C)")
```

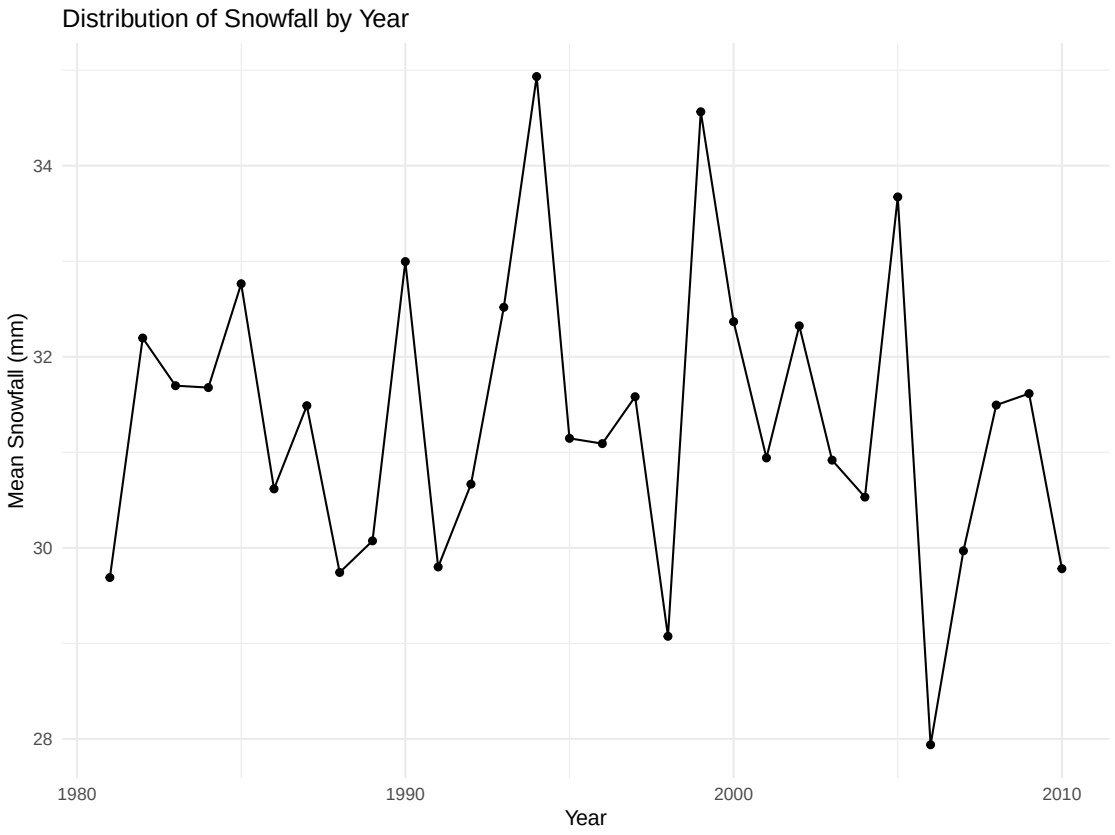
```
## `summarise()` has grouped output by 'id', 'year'. You can override using the
## `.groups` argument.
```




```
#Make a two-panel plot
ggplot(data_ny, aes(x = tmin, y = tmax)) +
  geom_hex() +
  labs(
    title = "Plot of tmax vs tmin",
    x = "Minimum Temperature (C)",
    y = "Maximum Temperature (C)"
  ) +
  theme_minimal()
```




```
#Make a plot showing the distribution of snowfall values
snowfall_data <- data_ny %>%
  filter(snow > 0 & snow < 100) %>%
  group_by(year) %>%
  summarize(mean_snowfall = mean(snow, na.rm = TRUE), .groups = 'drop')

ggplot(snowfall_data, aes(x = year, y = mean_snowfall)) +
  geom_point() + geom_line() +
  labs(
    title = "Distribution of Snowfall by Year",
    x = "Year",
    y = "Mean Snowfall (mm)"
  ) +
  theme_minimal()
```



#Problem 2

```
#Read and clean the datasets
demographics =
  read.csv("data/nhanes_covar.csv") %>%
  janitor::clean_names() %>%
  rename("seqn" = "x") %>%
  filter("age" >= 21) %>%
  drop_na()

accelerometer =
  read.csv("data/nhanes_accel.csv") %>%
  janitor::clean_names()
```

```
#Convert 'seqn' to character
accelerometer <- accelerometer %>%
  mutate(seqn = as.character(seqn))

#Merge two datasets
merged_data =left_join(demographics, accelerometer, by = "seqn") %>%
  janitor::clean_names() %>%
  rename("sex" = "x1_male", "education" = "x1_less_than_high_school", "age" = "x_1") %>%
  drop_na()
```

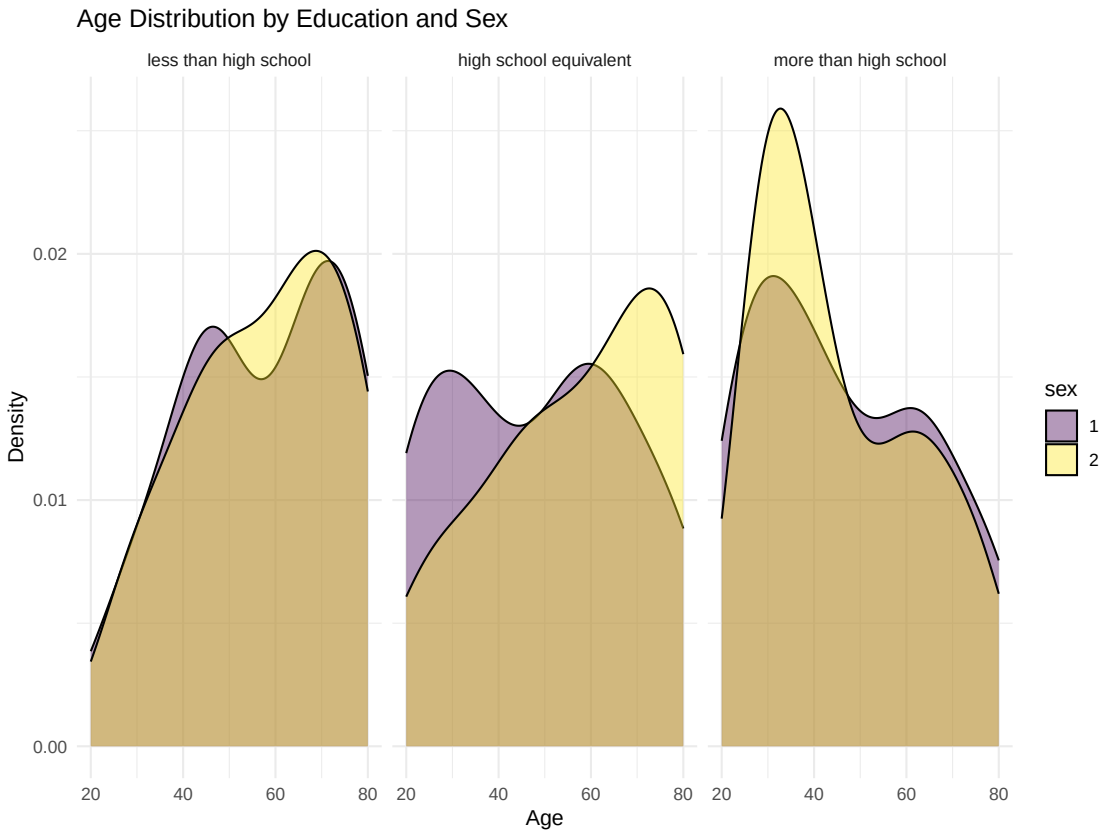
```
#Produce a table for the number of men and women in each education category
education_table = merged_data %>%
```

```
  group_by(sex, education) %>%
  summarize(count = n(), .groups = 'drop') %>%
  pivot_wider(names_from = sex, values_from = count, values_fill = 0) %>%
  rename("male" = 2, "female" = 3) %>%
  mutate(education = factor(education, levels = c(1, 2, 3),
  labels = c("less than high school", "high school equivalent",
                                     "more than high school")))
education_table %>%
  knitr::kable(
    caption = "Education Level and Sex Table",
  )
```

Table 1: Education Level and Sex Table

education	male	female
less than high school	27	28
high school equivalent	36	23
more than high school	56	59

```
# Convert age to numeric
merged_data <- merged_data %>%
  mutate(age = as.numeric(as.character(age)))

new_data = merged_data %>%
  mutate(age_group = cut(age, breaks = seq(20, 80, by = 5), right = FALSE)) %>%
  mutate(education = factor(education, levels = c(1, 2, 3),
 labels = c("less than high school", "high school equivalent",
                                     "more than high school")))

#Create a visualization of the age distributions for men and women in each education category
ggplot(new_data, aes(x = age, fill = sex)) +
  geom_density(alpha = 0.4) +
  facet_wrap(~education) +
  labs(
    title = "Age Distribution by Education and Sex",
    x = "Age",
    y = "Density"
  ) +
  theme_minimal()
```



From the plot, it shows that the number of women would be much more than men to obtain education more than high school, especially for younger women from 20-40 years old. While for the people received education of high school or equivalent, younger women from 20-40 years old is more than men, but for men in 60-80 years old, there are more men than women. For people only received education less than high school, the number of women and men are mostly equivalent across all ages, except for women in 50 to 70 years old.

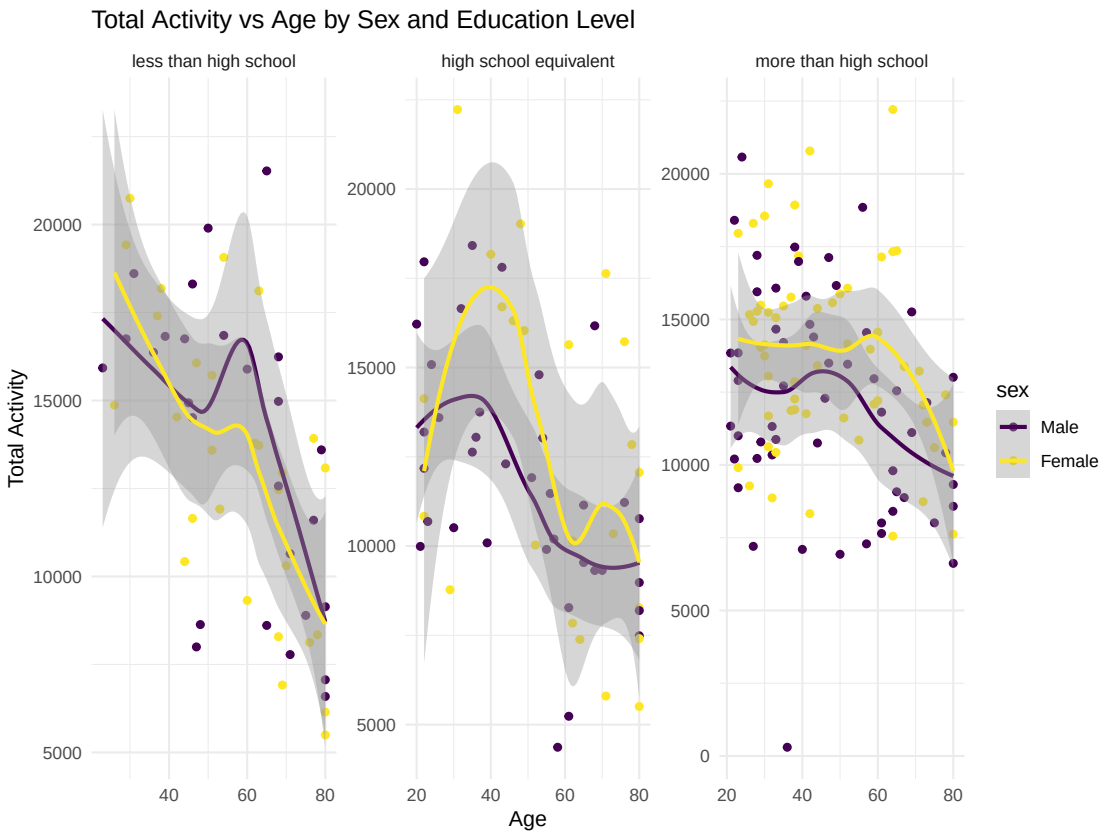
```
#Clean the dataset for total activity
merged_data$sex <- factor(merged_data$sex, levels = c(1, 2), labels = c("Male", "Female"))
merged_data$total_activity = rowSums(select(merged_data, starts_with("min")), na.rm = TRUE)

total_data <- merged_data %>%
  group_by(seqn, age, sex, education) %>%
  mutate(education = factor(education, levels = c(1, 2, 3),
  labels = c("less than high school",
             "high school equivalent",
             "more than high school")))%>%
  summarise(total_activity = sum(total_activity)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'seqn', 'age', 'sex'. You can override
## using the `.groups` argument.
```

```
# Make a plot that shows the 24-hour activity time courses for each education level
ggplot(total_data, aes(x = age, y = total_activity, color = sex)) +
  geom_point() +
  geom_smooth(method = "loess") +
```

```
  facet_wrap(~education, scales = "free") +
  labs(title = "Total Activity vs Age by Sex and Education Level",
       x = "Age",
       y = "Total Activity") +
  theme_minimal()
```

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

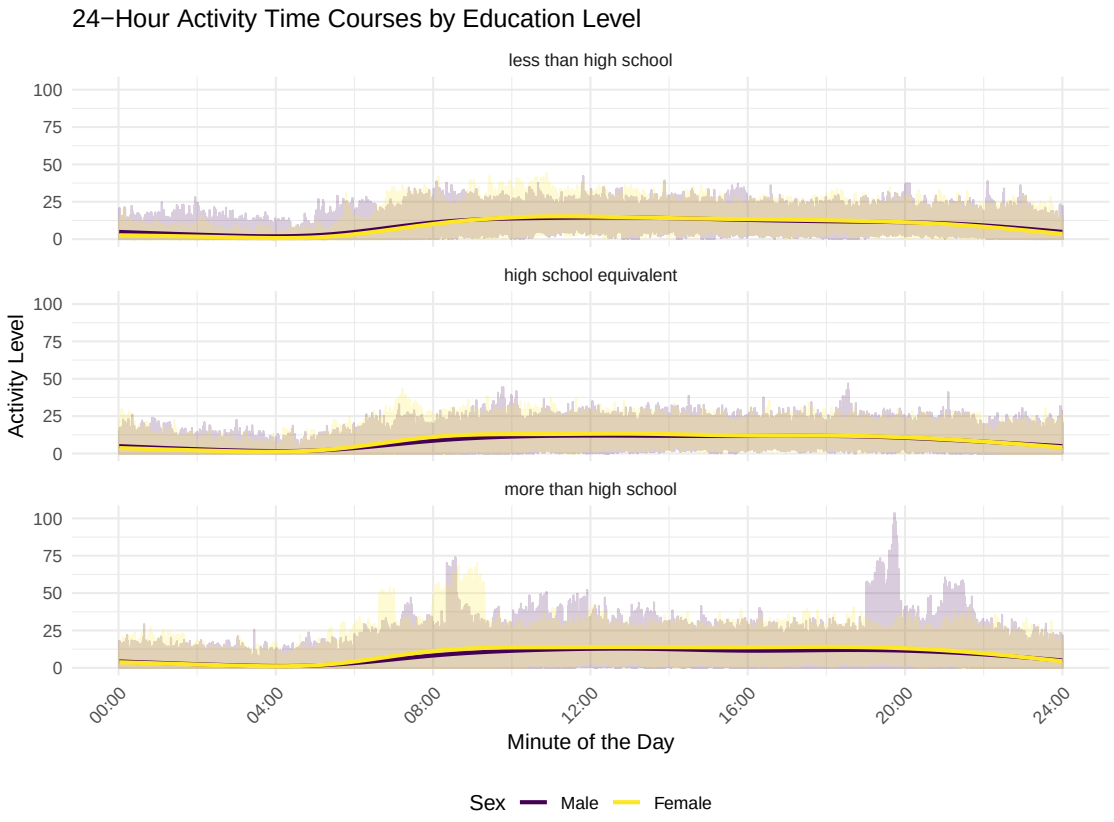


In education level 1, males generally begin with higher total activity levels compared to females. By the age of 60–70, male activity levels drop significantly, often nearing zero. Females in this education category start with lower activity levels, but their decline in activity is more gradual than for males. For participants in education level 2, both males and females exhibit a gradual decline in activity with age, but the trend is less steep than in level 1. Interestingly, females in this group experience a slight increase in activity during age 40–60 years, resulting in more sustained activity levels compared to their male counterparts. In education level 3, both males and females display relatively stable total activity levels over time. For males, the decline in activity with age is less steep compared to the other education levels, indicating that higher education may correlate with more sustained physical activity into older age. Females maintain higher and more stable activity levels compared to the other education levels, with only a slight drop after age 50.

```
## Plot the 24-hour activity time courses for each education level
activity_data <- merged_data %>%
  select(sex, education, starts_with("min")) %>%
  pivot_longer(cols = starts_with("min"), names_to = "minute", values_to = "activity") %>%
  mutate(minute = as.numeric(gsub("min", "", minute)),
         education = factor(education, levels = c(1, 2, 3),
         labels = c("less than high school",
```

```
              "high school equivalent",
              "more than high school")))

ggplot(activity_data, aes(x = minute, y = activity, color = as.factor(sex))) +
  geom_line(aes(group = interaction(sex, education)), alpha = 0.2) +
  geom_smooth(se = FALSE) +
  facet_wrap(~education, ncol = 1) +
  labs(
    title = "24-Hour Activity Time Courses by Education Level",
    x = "Minute of the Day",
    y = "Activity Level",
    color = "Sex"
  ) +
  theme_minimal() +
  scale_x_continuous(breaks = seq(0, 1440, by = 240), labels = c("00:00", "04:00", "08:00", "12:00", "16:00", "20:00", "24:00")) +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



From the plot, higher education levels are associated with more distinct activity peaks, particularly in the group with more than high school education. Females generally show higher overall activity, especially during the morning, across all education levels. Males exhibit lower activity levels in the early morning but tend to catch up and show spikes in activity, especially during the afternoon and evening hours. For all groups, the smooth trends show a typical daily activity pattern, with a clear increase in activity starting in the early morning, peaking around midday to early evening. These patterns likely correspond to typical work or school schedules, combined with some relaxed activities in the afternoon.

#Problem 3

```
#Load and combine the datasets
july_2024_data <- read_csv("data/citibike/July 2024 Citi.csv", show_col_types = FALSE)%>%
  mutate(year = 2024, month = "July")
jan_2020_data <- read_csv("data/citibike/Jan 2020 Citi.csv", show_col_types = FALSE) %>%
  mutate(year = 2020, month = "January")
july_2020_data <- read_csv("data/citibike/July 2020 Citi.csv", show_col_types = FALSE) %>%
  mutate(year = 2020, month = "July")
jan_2024_data <- read_csv("data/citibike/Jan 2024 Citi.csv", show_col_types = FALSE) %>%
  mutate(year = 2024, month = "January")

citibike_data <- bind_rows(jan_2020_data, july_2020_data, jan_2024_data, july_2024_data)
```

```
#Create a summary table
ride_summary <- citibike_data %>%
  group_by(year, month, member_casual) %>%
  summarise(total_rides = n(),.groups = "drop") %>%
  arrange(year, month) %>%
  pivot_wider(
    names_from = member_casual,
    values_from = total_rides,
    names_prefix = "rider_"
  ) %>%
  rename(
    Casual_Riders = rider_casual,
    Member_Riders = rider_member
  )
ride_summary
```

```
## # A tibble: 4 x 4
##    year month   Casual_Riders Member_Riders
##   <dbl> <chr>           <int>         <int>
## 1  2020 January           984         11436
## 2  2020 July             5637         15411
## 3  2024 January          2108         16753
## 4  2024 July            10894         36262
```

```
ride_summary %>%
  knitr::kable(
    caption = "Total Number of Rides by Year, Month, and Membership Status",
    col.names = c("Year", "Month", "Casual Riders", "Member Riders")
  )
```

Table 2: Total Number of Rides by Year, Month, and Membership Status

Year	Month	Casual Riders	Member Riders
2020	January	984	11436
2020	July	5637	15411
2024	January	2108	16753
2024	July	10894	36262

Year	Month	Casual Riders	Member Riders

Compare the riders in January 2020 and January 2024, both casual and member rides increased over time, but the number of members grows more. In 2024, member riders are greatly more than casual riders in both January and July. July sees a significant spike in both casual and member rides compared to January, especially for members. In both years, members consistently take more rides than casual riders, indicating that frequent riders are more likely to become members and take advantage of lower rates.

```
#Make a table showing the 5 most popular starting stations for July 2024
top_july_2024 <- july_2024_data %>%
  group_by(start_station_name) %>%
  summarise(total_rides = n()) %>%
  arrange(desc(total_rides)) %>%
  slice_max(order_by = total_rides, n = 5)

top_july_2024 %>%
  knitr::kable(
    caption = "Top 5 Most Popular Starting Stations for July 2024",
    col.names = c("Starting Station", "Total Rides")
    )
```

Table 3: Top 5 Most Popular Starting Stations for July 2024

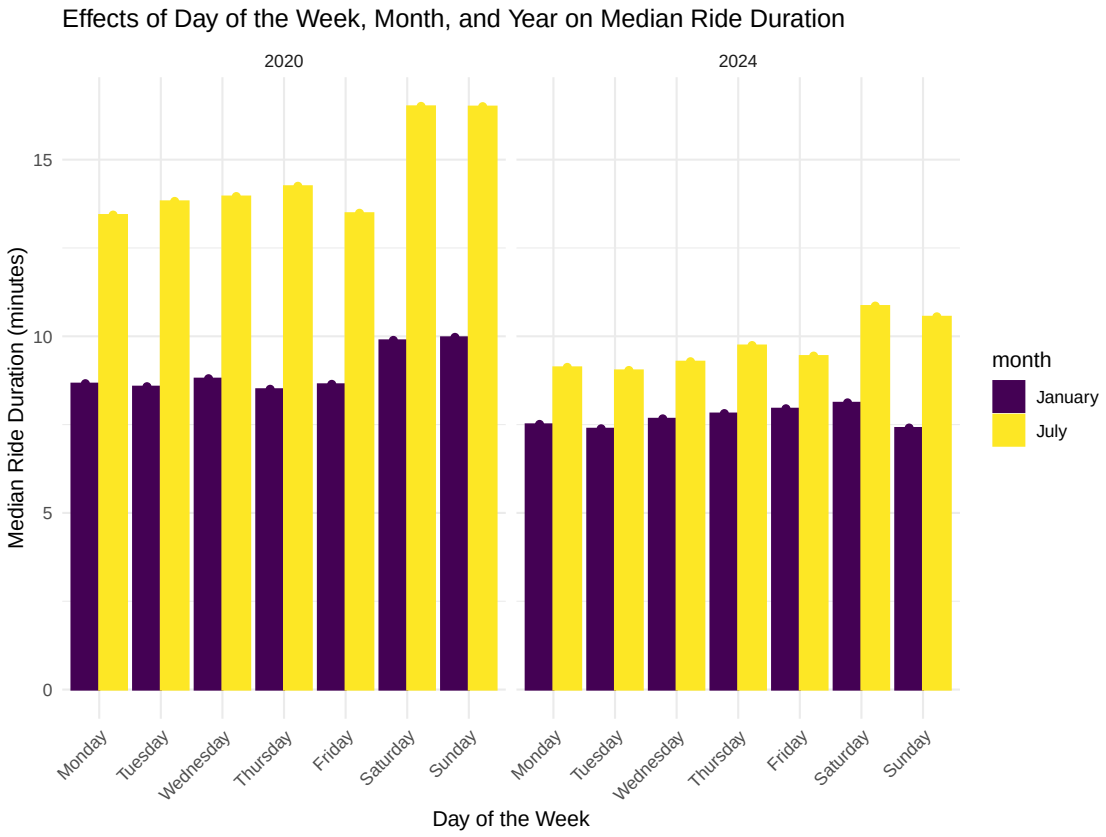
Starting Station	Total Rides
Pier 61 at Chelsea Piers	163
University Pl & E 14 St	155
W 21 St & 6 Ave	152
West St & Chambers St	150
W 31 St & 7 Ave	146

```
#Make a plot to investigate the effects of day of the week, month, and year on median ride duration
citibike_data <- citibike_data %>%
  mutate(weekdays = factor(weekdays, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday",

median_duration <- citibike_data %>%
  group_by(year, month, weekdays) %>%
  summarise(median_duration = median(duration, na.rm = TRUE)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'year', 'month'. You can override using the
## `.groups` argument.
```

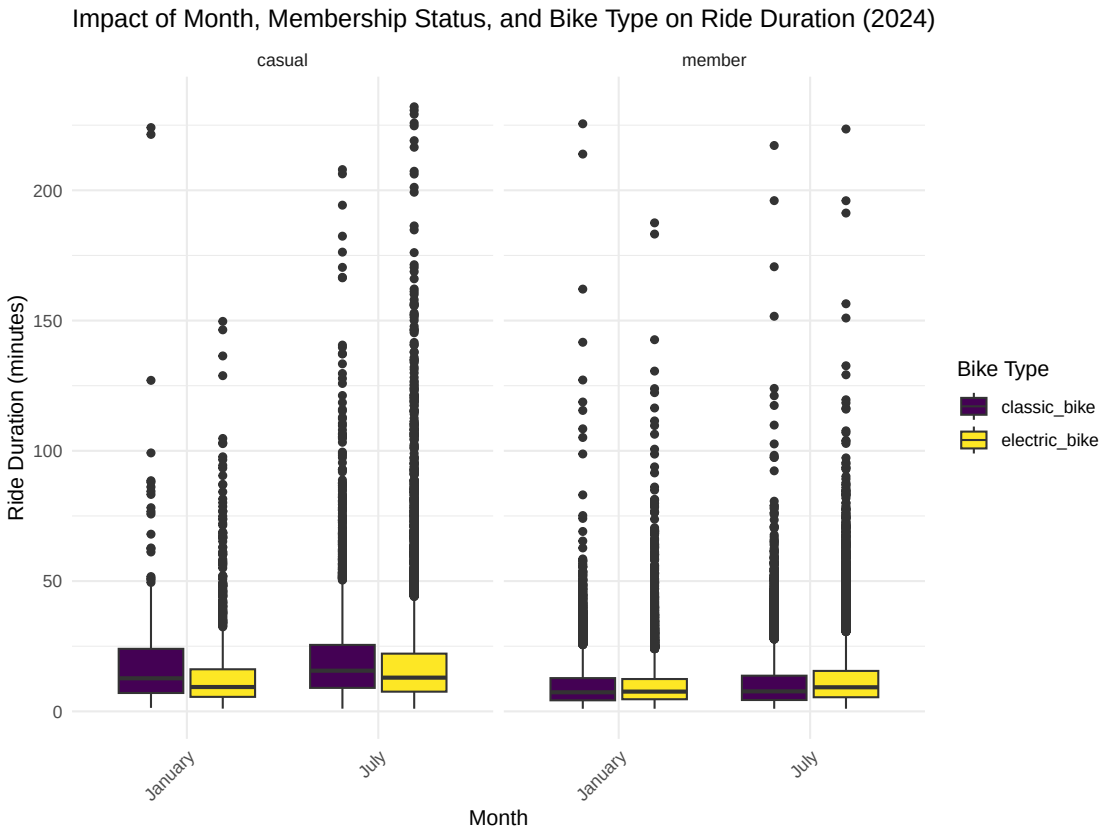
```
ggplot(median_duration, aes(x = weekdays, y = median_duration, color = month, group = month)) +
  geom_col(position = "dodge", aes(fill = month)) +  # Use geom_col() for actual values
  geom_point(position = position_dodge(width = 0.9)) +  # Optional: Use points to show individual value
  facet_wrap(~year) +
  labs(
    title = "Effects of Day of the Week, Month, and Year on Median Ride Duration",
    x = "Day of the Week",
    y = "Median Ride Duration (minutes)"
```

```
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The plot shows that July consistently has longer ride durations compared to January across all days of the week in both 2020 and 2024, with weekends showing the largest differences. In comparing 2020 to 2024, July 2024 shows slightly shorter ride durations than July 2020, especially on weekends. January 2024 shows slightly longer rides than January 2020, although the difference is minimal. Overall, the plot highlights seasonal and weekly patterns in Citi Bike usage, with longer rides during warmer months and weekends.

```
#Make a figure that shows the impact of month, membership status, and bike type on the distribution of
citibike_2024 = citibike_data %>%
  filter(year == 2024)

ggplot(citibike_2024, aes(x = month, y = duration, fill = rideable_type)) +
  geom_boxplot() +
  facet_wrap(~member_casual) +
  labs(
    title = "Impact of Month, Membership Status, and Bike Type on Ride Duration (2024)",
    x = "Month",
    y = "Ride Duration (minutes)",
    fill = "Bike Type"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the figure, people would like to takinf electric Citibikes for shorter rides, such as members who use these bikes for commuting. There is more electric bike usage for casual riders as well, but their ride durations could still be longer, which might be for recreational use. Moreover, members tend to have shorter, more consistent ride durations. Casual riders might have more varied ride times, especially in July.