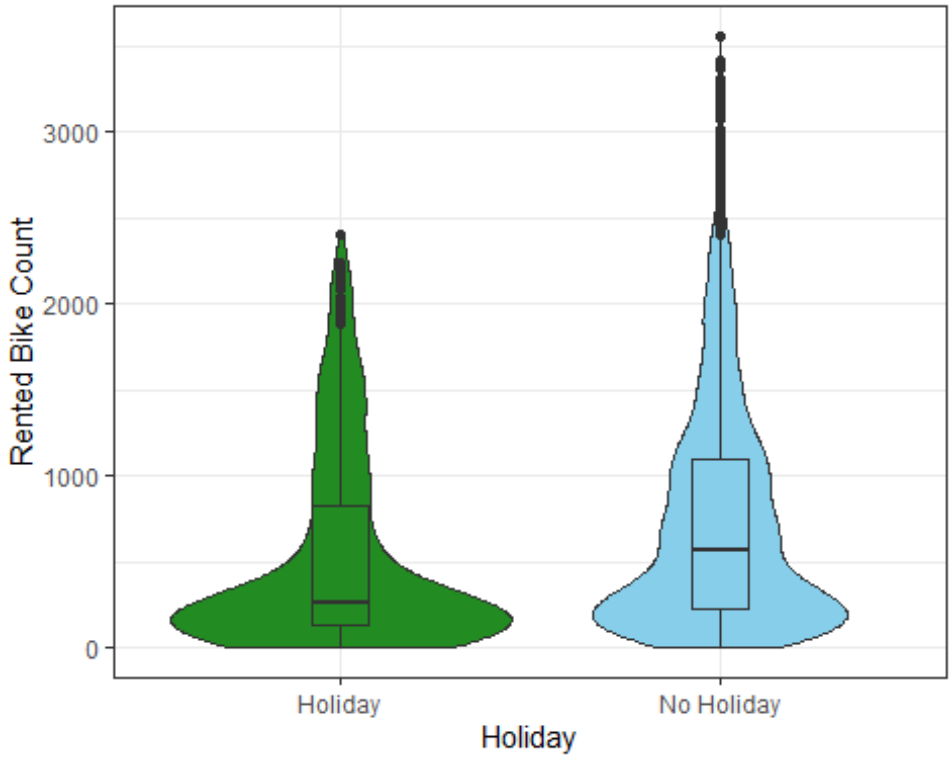
5. A/B Testing

5.1 Phân loại dữ liệu thành hai nhóm: Holiday và No Holiday

```
ggplot(data, aes(x = holiday, y = rented_bike_count, fill = holiday)) +
  geom_violin() +
  geom_boxplot(width = 0.15) +
  scale_fill_manual(breaks = c("Holiday", "No Holiday"),
                    values = c("forestgreen", "skyblue")) +
  labs(x = "Holiday", y = "Rented Bike Count") +
  theme_bw() +
  theme(legend.position = "none")
```



Nhóm A: Dữ liệu vào các ngày không phải là ngày nghỉ lễ.

Nhóm B: Dữ liệu vào các ngày nghỉ lễ.

Thông qua bảng tổng hợp và biểu đồ violin, một giả định có thể là "Số lượng sử dụng xe đạp vào những ngày No Holiday là nhiều hơn những ngày Holiday".

Do đó, ta cần kiểm chứng giả thuyết và đối thuyết sau:

- Giả thuyết: $\mu A = \mu B$
- Đối thuyết: $\mu A < \mu B$

```r
summary_data <- data %>%
  group_by(holiday) %>%
  summarise(
    n = n(),
    mean_rented_bike_count = mean(rented_bike_count, na.rm = TRUE),
    sd_rented_bike_count = sd(rented_bike_count, na.rm = TRUE)
  )
print(summary_data)
```

```
## # A tibble: 2 × 4
##   holiday        n mean_rented_bike_count sd_rented_bike_count
##   <fct>      <int>                  <dbl>                <dbl>
## 1 Holiday      408                   529.                 574.
## 2 No Holiday  8057                   739.                 644.
```

```r
perm_fun <- function(x, nA, nB, R) {
n <- nA + nB
mean_diff <- numeric(R)
for (i in 1:R){
idx_a <- sample(x = 1:n, size = nA)
idx_b <- setdiff(x = 1:n, y = idx_a)
mean_diff[i] <- mean(x[idx_a]) - mean(x[idx_b])
}
return(mean_diff)
}
set.seed(42)
diff_mean_perm <- perm_fun(data$rented_bike_count, nA = 8328 , nB = 432  , R
= 10000)
mean_a <- mean(data$rented_bike_count[data$holiday == 'Holiday'])
mean_b <- mean(data$rented_bike_count [data$holiday == 'No Holiday'])
mean(diff_mean_perm < (mean_a - mean_b))
```

```
## [1] NA
```

Với mức ý nghĩa α = 0, kết quả cho thấy Giả thuyết là không thể bị bác bỏ. Do đó, số lượng sử dụng xe đối với những ngày No Holiday nhiều hơn so với nhứng ngày Holiday là không có ý nghĩa thống kê, hay chỉ là kết quả của sự ngẫu nhiên.

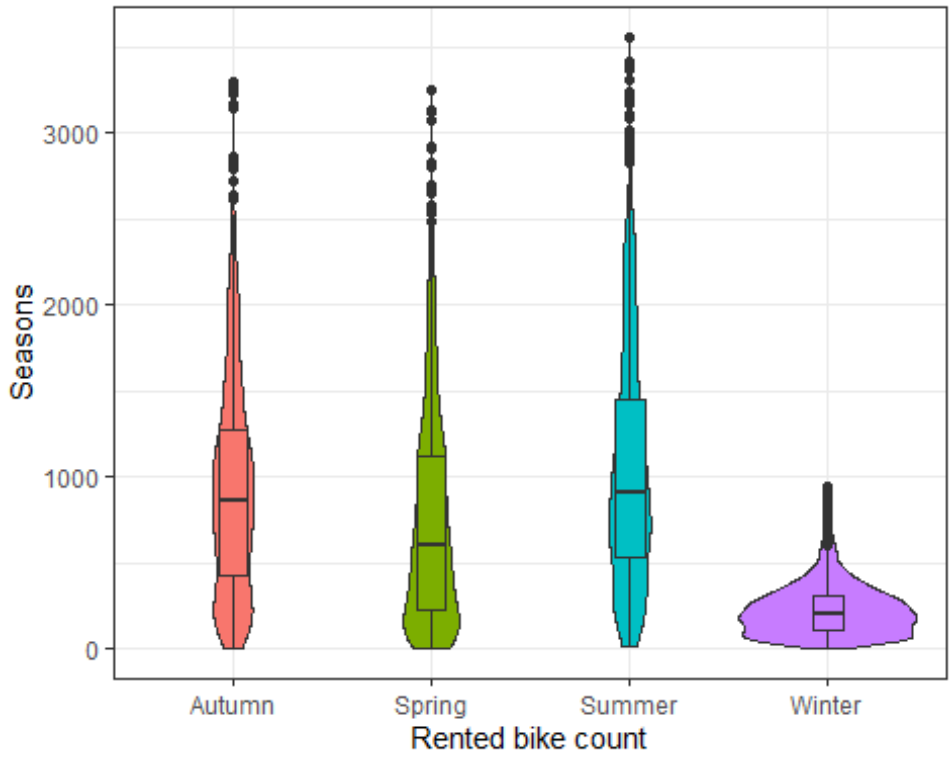
5.2 Season

```r
# Thực hiện nhóm dữ liệu theo 'seasons' và tính toán các giá trị
summary_seasons <- data %>%
  group_by(seasons) %>%
  summarise(
    n = n(),
    mean_rented_bike_count = mean(rented_bike_count, na.rm = TRUE),
    sd_rented_bike_count = sd(rented_bike_count, na.rm = TRUE)
  )
print(summary_seasons)
```

```
## # A tibble: 4 × 4
##   seasons     n mean_rented_bike_count sd_rented_bike_count
```

```
##   <fct>   <int>                 <dbl>                 <dbl>
## 1 Autumn   1937                  924.                  618.
## 2 Spring   2160                  746.                  619.
## 3 Summer   2208                 1034.                  690.
## 4 Winter   2160                  226.                  150.
```

```
ggplot(data, aes(x = seasons, y = rented_bike_count, fill = seasons)) +
  geom_violin() +
  geom_boxplot(width = 0.15) +
  labs(x = "Rented bike count", y = "Seasons") +
  theme_bw() +
  theme(legend.position = "none")
```




```
library(lmPerm)

set.seed(56)
out_aov_1 <- aovp(formula = rented_bike_count ~ seasons,
                  data = data |> mutate(across(seasons, as.factor)))

## [1] "Settings:  unique SS "

summary(out_aov_1)

## Component 1 :
##               Df   R Sum Sq R Mean Sq Iter  Pr(Prob)
## seasons        3  827375618 275791873 5000 < 2.2e-16 ***
## Residuals   8461 2664997916    314974
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

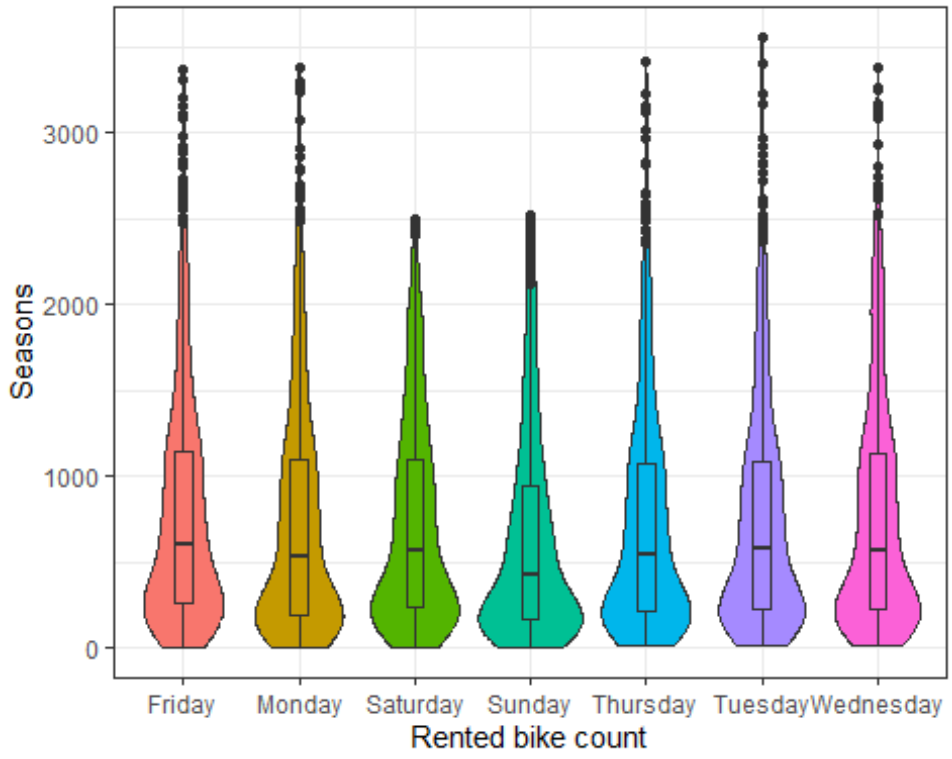
=> Giá trị p-value rất nhỏ (< 2.2e-16) cho thấy sự khác biệt về số lượng xe đạp thuê giữa các mùa là có ý nghĩa thống kê. Điều này có nghĩa là các mùa khác nhau ảnh hưởng đến số lượng xe đạp thuê một cách đáng kể.

5.3 Ngày trong tuần.

```
# Thực hiện nhóm dữ liệu theo 'day_of_week' và tính toán các giá trị
summary_seasons <- data %>%
  group_by(day_of_week) %>%
  summarise(
    n = n(),
    mean_rented_bike_count = mean(rented_bike_count, na.rm = TRUE),
    sd_rented_bike_count = sd(rented_bike_count, na.rm = TRUE)
  )
print(summary_seasons)

## # A tibble: 7 × 4
##   day_of_week     n mean_rented_bike_count sd_rented_bike_count
##   <chr>       <int>                  <dbl>                <dbl>
## 1 Friday       1224                   776.                 654.
## 2 Monday       1248                   731.                 665.
## 3 Saturday     1217                   728.                 601.
## 4 Sunday       1224                   637.                 605.
## 5 Thursday     1200                   718.                 639.
## 6 Tuesday      1152                   745.                 647.
## 7 Wednesday    1200                   770.                 674.

ggplot(data, aes(x = day_of_week, y = rented_bike_count, fill = day_of_week))
+
  geom_violin() +
  geom_boxplot(width = 0.15) +
  labs(x = "Rented bike count", y = "Seasons") +
  theme_bw() +
  theme(legend.position = "none")
```




```
set.seed(69)

# Thực hiện phân tích ANOVA với dữ liệu đã được xử lý
out_aov_2<- aovp(formula = rented_bike_count ~ day_of_week, data = data, perm
= "Prob")

## [1] "Settings:  unique SS "

summary(out_aov_2)

## Component 1 :
##                Df   R Sum Sq R Mean Sq Iter  Pr(Prob)
## day_of_week1    6   15480791   2580132 5000 < 2.2e-16 ***
## Residuals    8458 3476892744    411077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

=> Phân tích cho thấy rằng có sự khác biệt đáng kể (p < 2.2e-16) về số lượng xe đạp được thuê giữa các ngày trong tuần (các nhóm của biến day_of_week).