Data Analysis of Departure Delays for United Airlines

Vivek Reddy Karra

2023-10-23

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(tidyr)
library(nycflights13)
```

```
data("flights")
data("weather")
```

```
flights_weather_joined <- flights %>%
  inner_join(weather, by = c("year", "month", "day", "hour", "origin"))
glimpse(flights_weather_joined)
```

```
## Rows: 335,220
## Columns: 29
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, …
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, …
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1…
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,…
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,…
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1…
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "…
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4…
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394…
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",…
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",…
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1…
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, …
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6…
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0…
## $ time_hour.x    <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
## $ temp           <dbl> 39.02, 39.92, 39.02, 39.02, 39.92, 39.02, 37.94, 39.92,…
## $ dewp           <dbl> 28.04, 24.98, 26.96, 26.96, 24.98, 28.04, 28.04, 24.98,…
## $ humid          <dbl> 64.43, 54.81, 61.63, 61.63, 54.81, 64.43, 67.21, 54.81,…
## $ wind_dir       <dbl> 260, 250, 260, 260, 260, 260, 240, 260, 260, 260, 260, …
## $ wind_speed     <dbl> 12.65858, 14.96014, 14.96014, 14.96014, 16.11092, 12.65…
## $ wind_gust      <dbl> NA, 21.86482, NA, NA, 23.01560, NA, NA, 23.01560, NA, 2…
## $ precip         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ pressure       <dbl> 1011.9, 1011.4, 1012.1, 1012.1, 1011.7, 1011.9, 1012.4,…
## $ visib          <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,…
## $ time_hour.y    <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0…
```

```
UA_flights <- flights_weather_joined %>%
  filter(carrier=="UA")%>%
  filter(!is.na(dep_delay))%>%
  filter(!is.na(temp))%>%
  filter(!is.na(wind_speed))
```

```
table <- cbind(
summary(UA_flights$dep_delay),
summary(UA_flights$hour),
summary(UA_flights$temp),
summary(UA_flights$wind_speed),
summary(UA_flights$precip),
summary(UA_flights$visib))

columns <- c("dep_delay", "hour", "temp", "wind_speed", "precip", "visib")
colnames(table)<- columns
data.frame(table)
```

```
##         dep_delay     hour      temp wind_speed      precip     visib
## Min.    -20.00000  5.00000  10.94000    0.00000 0.000000000  0.000000
## 1st Qu.  -4.00000  8.00000  42.08000    6.90468 0.000000000 10.000000
## Median    0.00000 13.00000  57.92000    9.20624 0.000000000 10.000000
## Mean     12.09353 12.84721  57.35817   10.31214 0.005077694  9.266537
## 3rd Qu.  11.00000 17.00000  73.04000   13.80936 0.000000000 10.000000
## Max.    483.00000 23.00000 100.04000   42.57886 1.210000000 10.000000
```

```
new_df <- UA_flights %>%
  mutate(                          #mutate is used for creating new variables in dataset
    late = dep_delay > 0 ,
    very_late = dep_delay > 30,

    time_of_day = case_when(
      hour>=5 & hour<12 ~ "Morning",
      hour>=12 & hour<17 ~ "Noon",
      hour>=17 & hour<20 ~"Evening",
      TRUE ~ "Night"
    ),

         time_of_year = case_when(
           month %in% c(3, 4, 5) ~ "Spring",
           month %in% c(6, 7, 8) ~ "Summer",
           month %in% c(9, 10, 11) ~ "Fall",
           month %in% c(12, 1, 2) ~ "Winter"
         ),

    temperature = case_when(
    temp < 55 ~ "Cold",
    temp >= 55 & temp < 85 ~ "Mild",
    temp >= 85 ~ "Hot"
  ),

    wind_speeds = case_when(
    wind_speed < 30 ~ "Low",
    wind_speed >= 30 ~ "Strong"
  ),

  precipitation = case_when(
    precip<=0 ~ "Non-rain",
    precip>0 ~ "Raining"
  ),

  visibility = case_when(
    visib< 5.0 ~ "0 to 4m",
    visib>=5 ~ "5 to 10m"
  )

  )
```

```
summary(new_df$dep_delay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -20.00   -4.00    0.00   12.09   11.00  483.00
```

```
summary(new_df$hour)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    8.00   13.00   12.85   17.00   23.00
```

```
summary(new_df$month)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   7.000   6.555  10.000  12.000
```

```
summary(new_df$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.94   42.08   57.92   57.36   73.04  100.04
```

```
summary(new_df$wind_speed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.905   9.206  10.312  13.809  42.579
```

```
summary(new_df$precip)
```

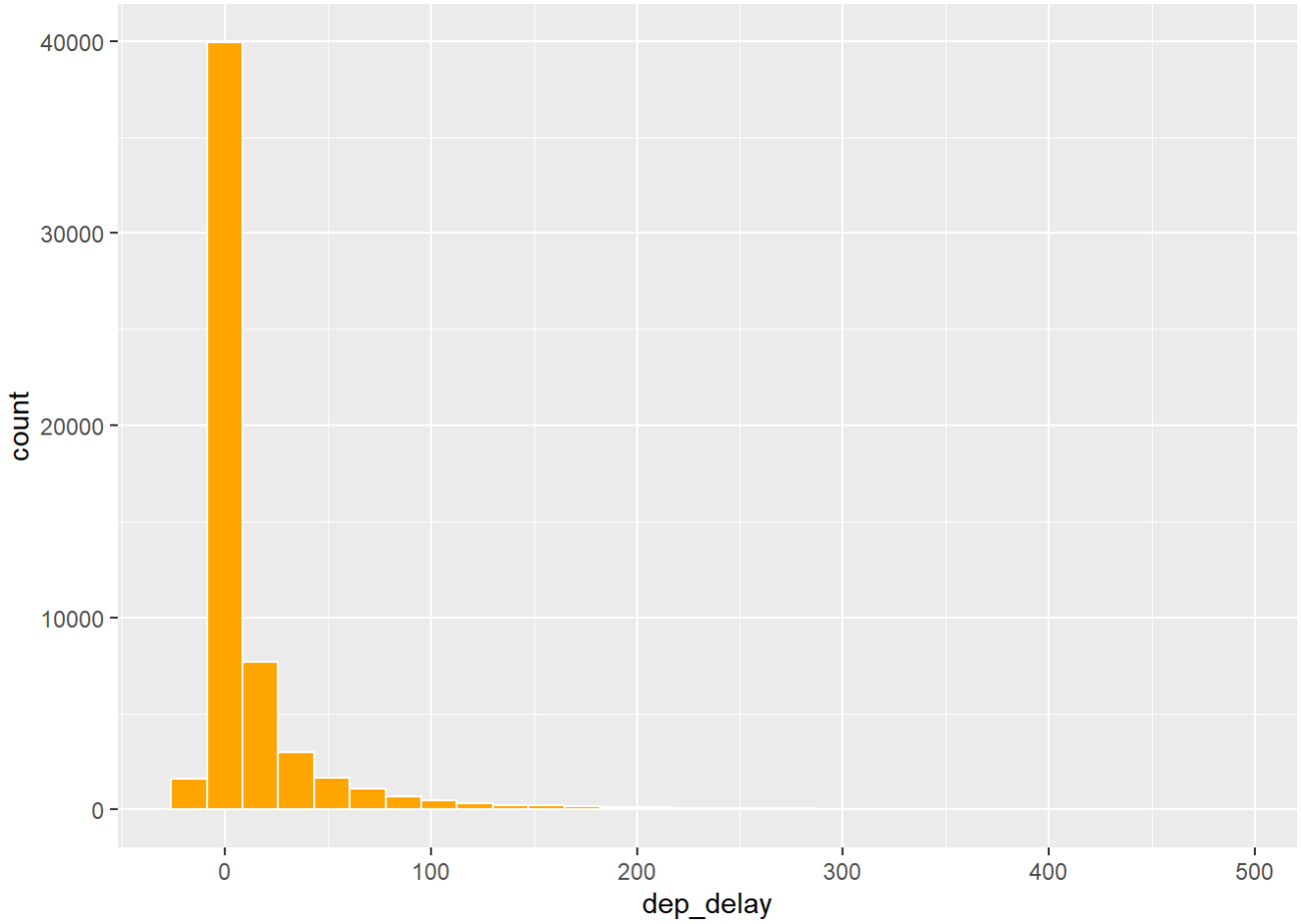
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000000 0.000000 0.005078 0.000000 1.210000
```

```
summary(new_df$visib)
```

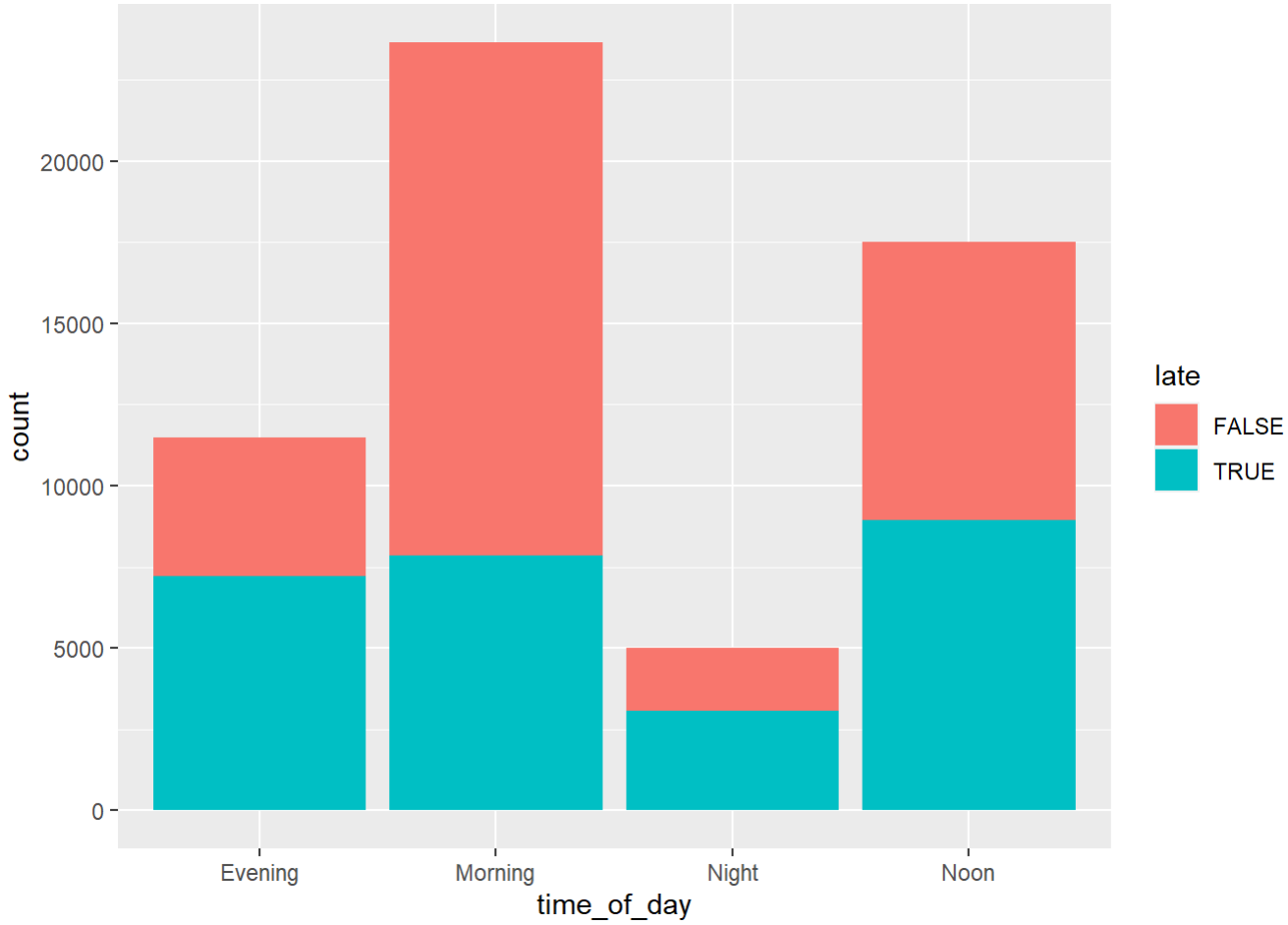
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000  10.000  10.000   9.267  10.000  10.000
```

```
ggplot(data = new_df, mapping=aes(x=dep_delay))+
  geom_histogram(color="white", fill="orange")
```

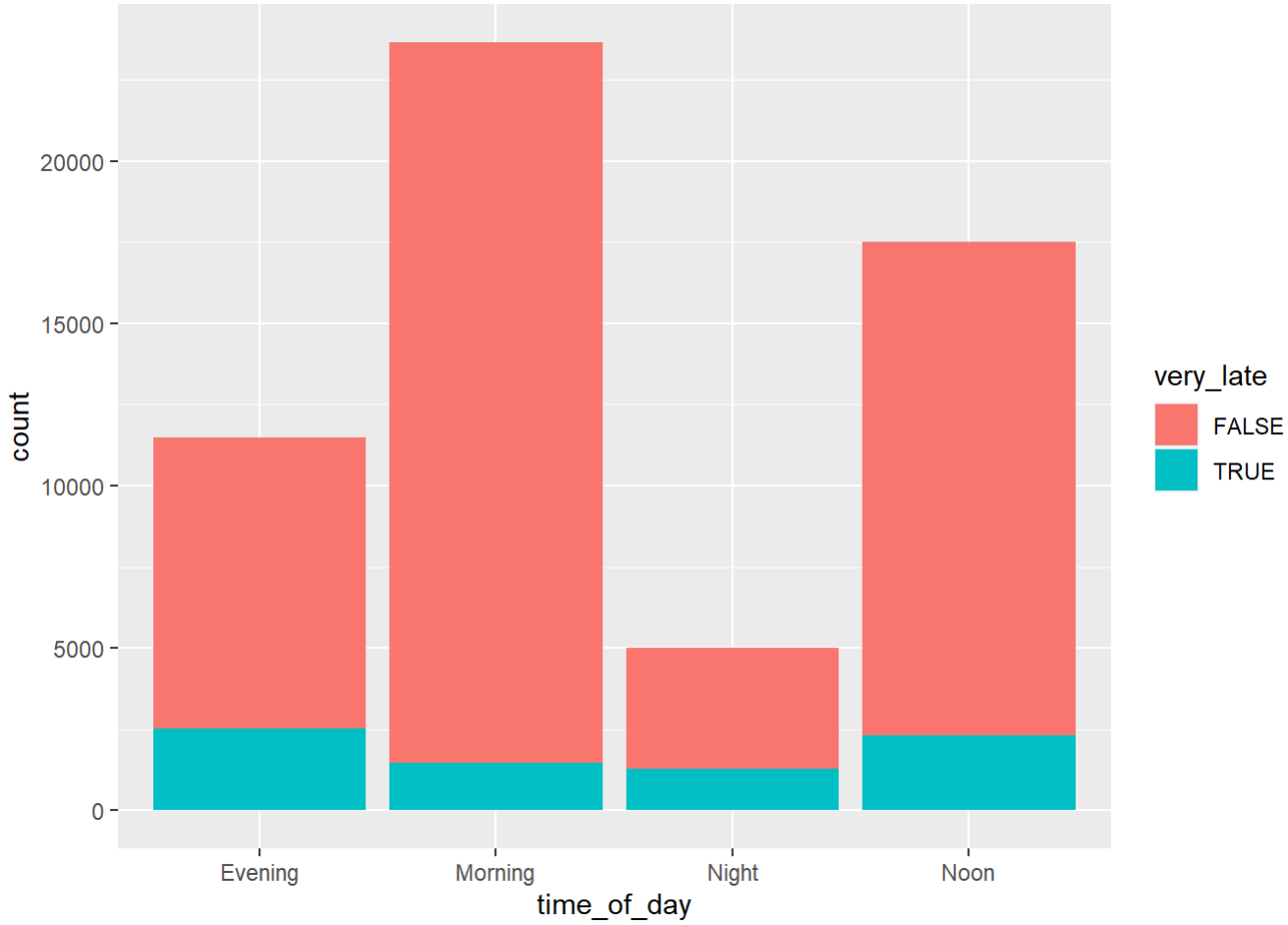
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



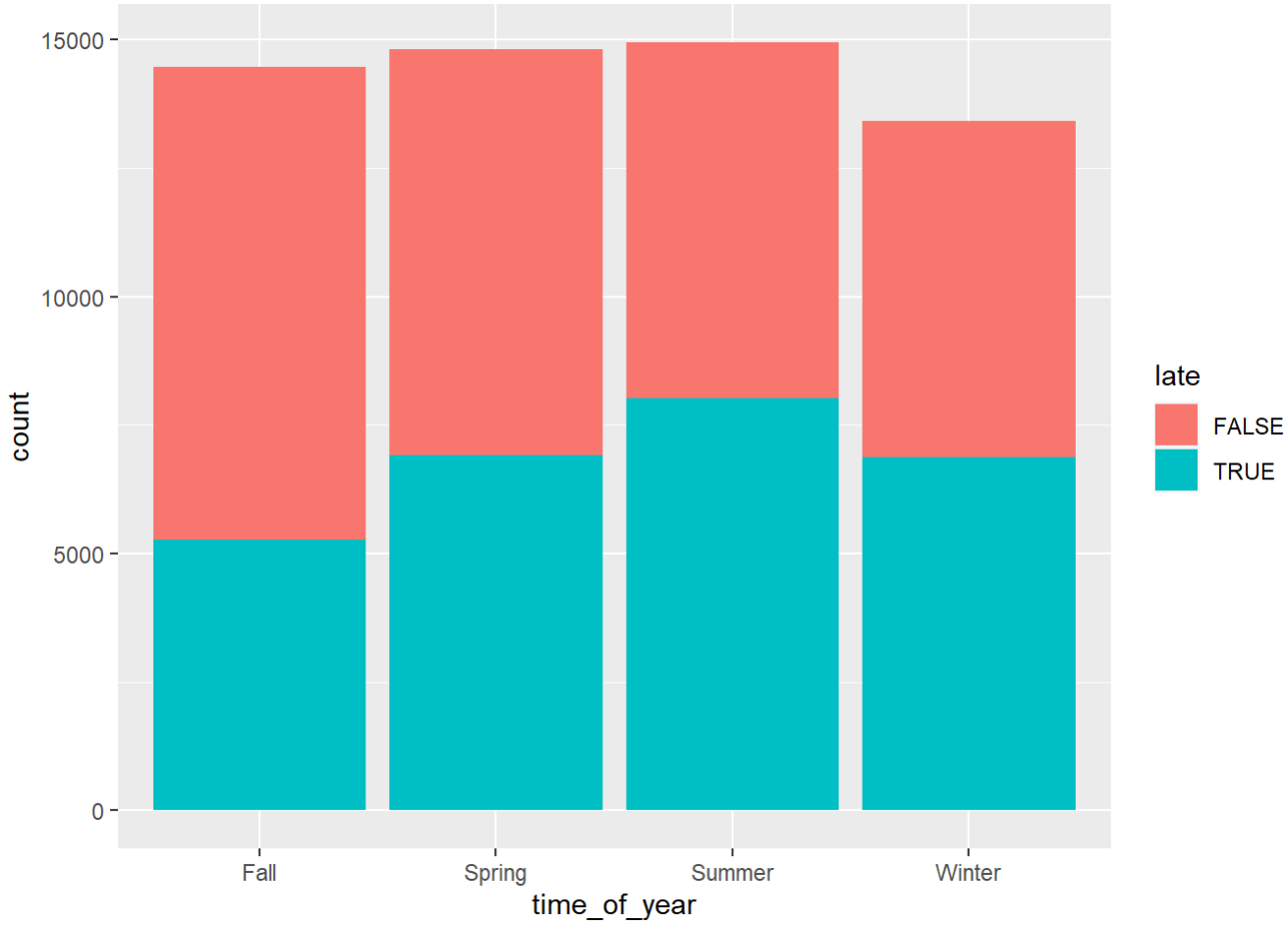

```
#time of day vs dep_delay

ggplot(data = new_df, mapping = aes(x = time_of_day, fill = late)) +
  geom_bar()
```



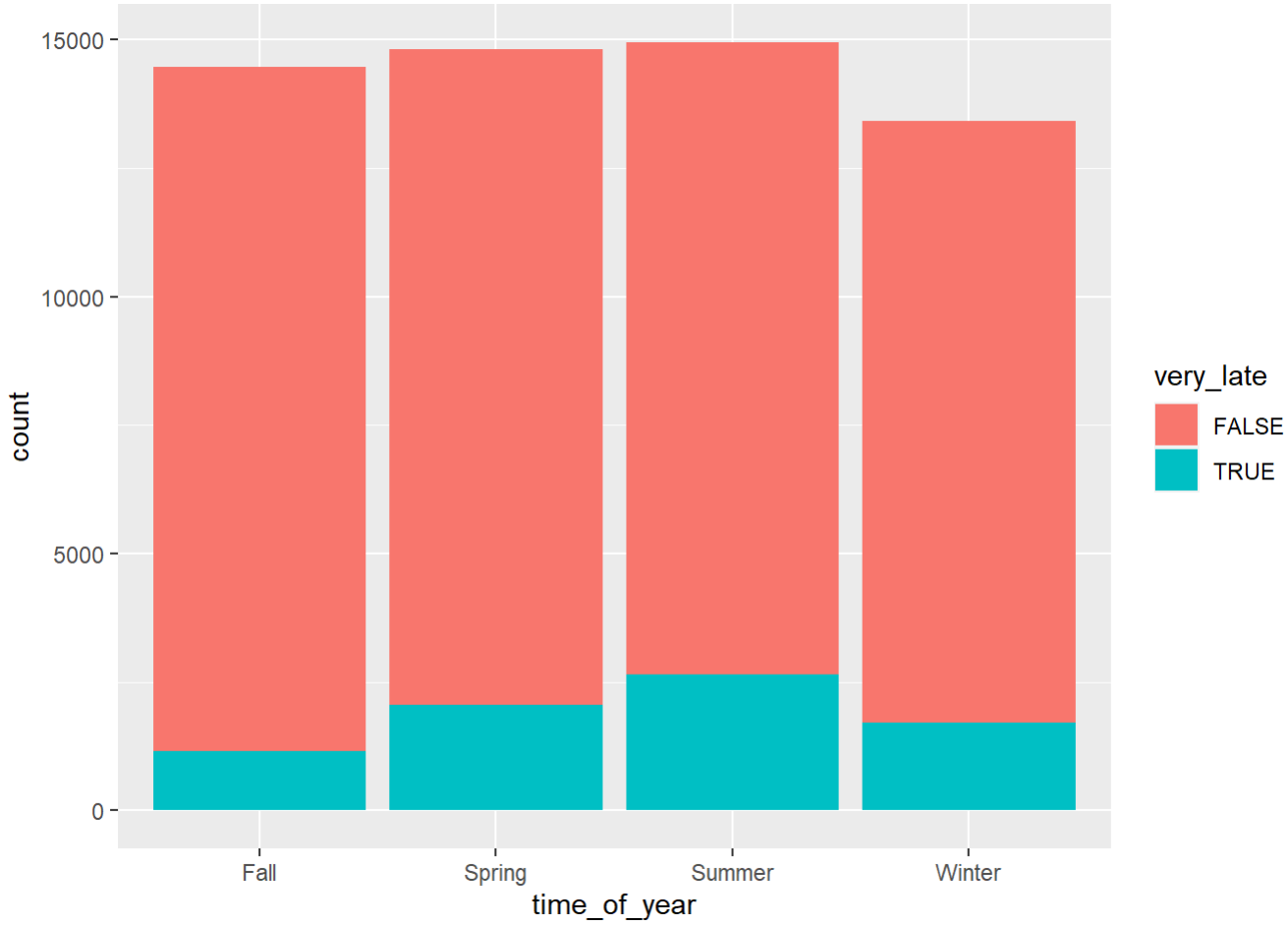

```
ggplot(data = new_df, mapping = aes(x = time_of_day, fill = very_late)) +
  geom_bar()
```



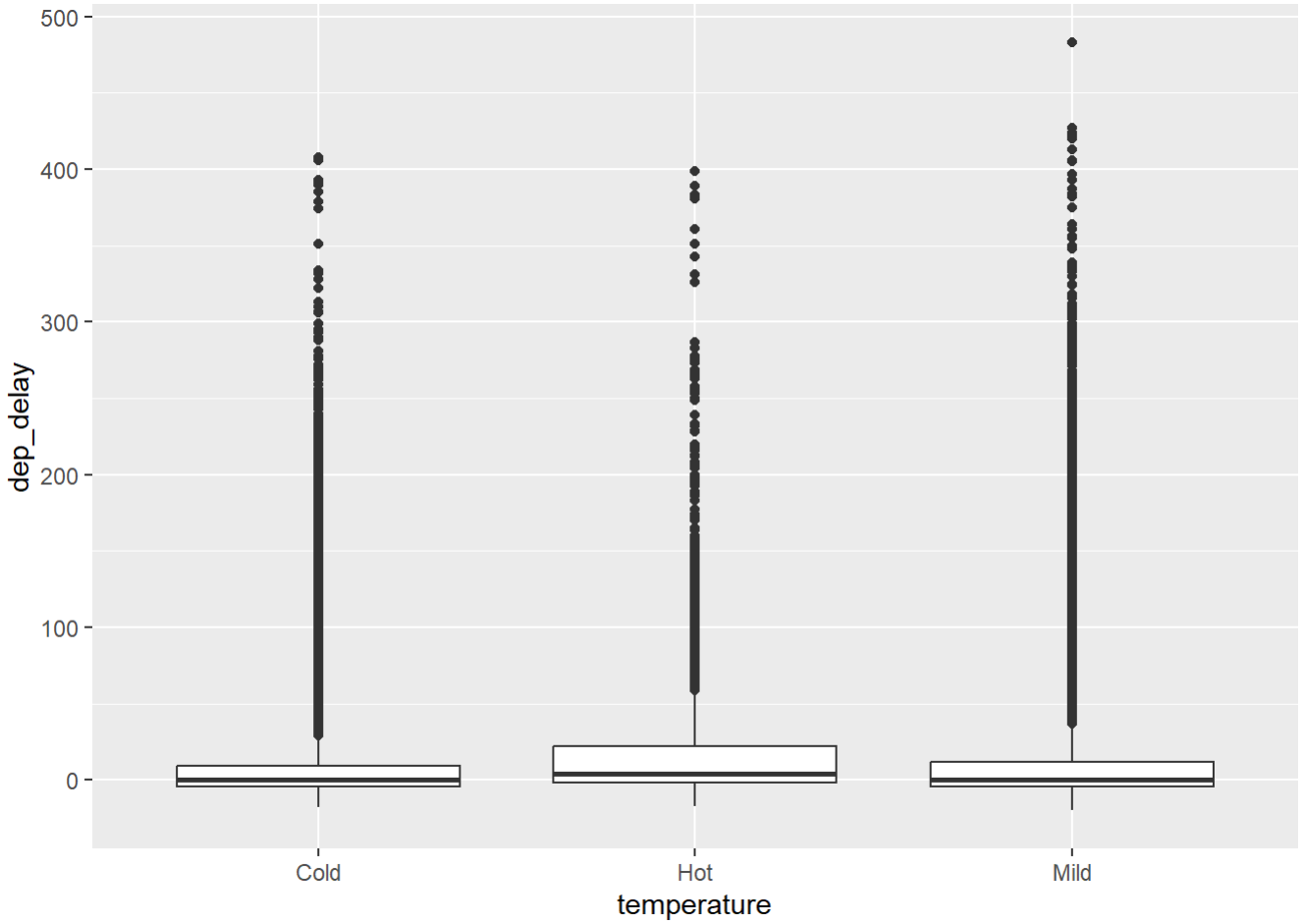

```
ggplot(data = new_df, mapping = aes(x = time_of_year, fill = late)) +
  geom_bar()
```



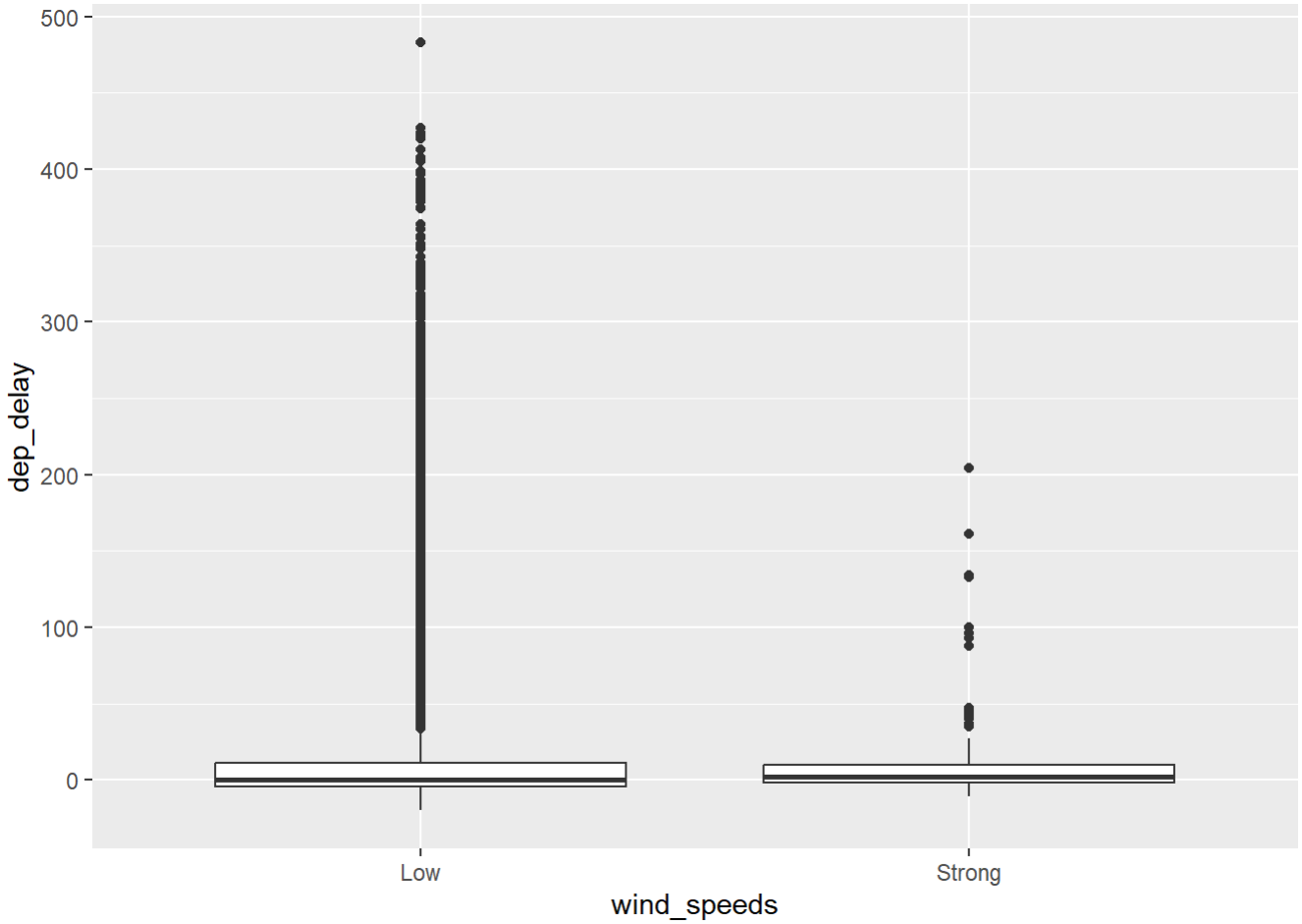

```
ggplot(data = new_df, mapping = aes(x = time_of_year, fill = very_late)) +
  geom_bar()
```



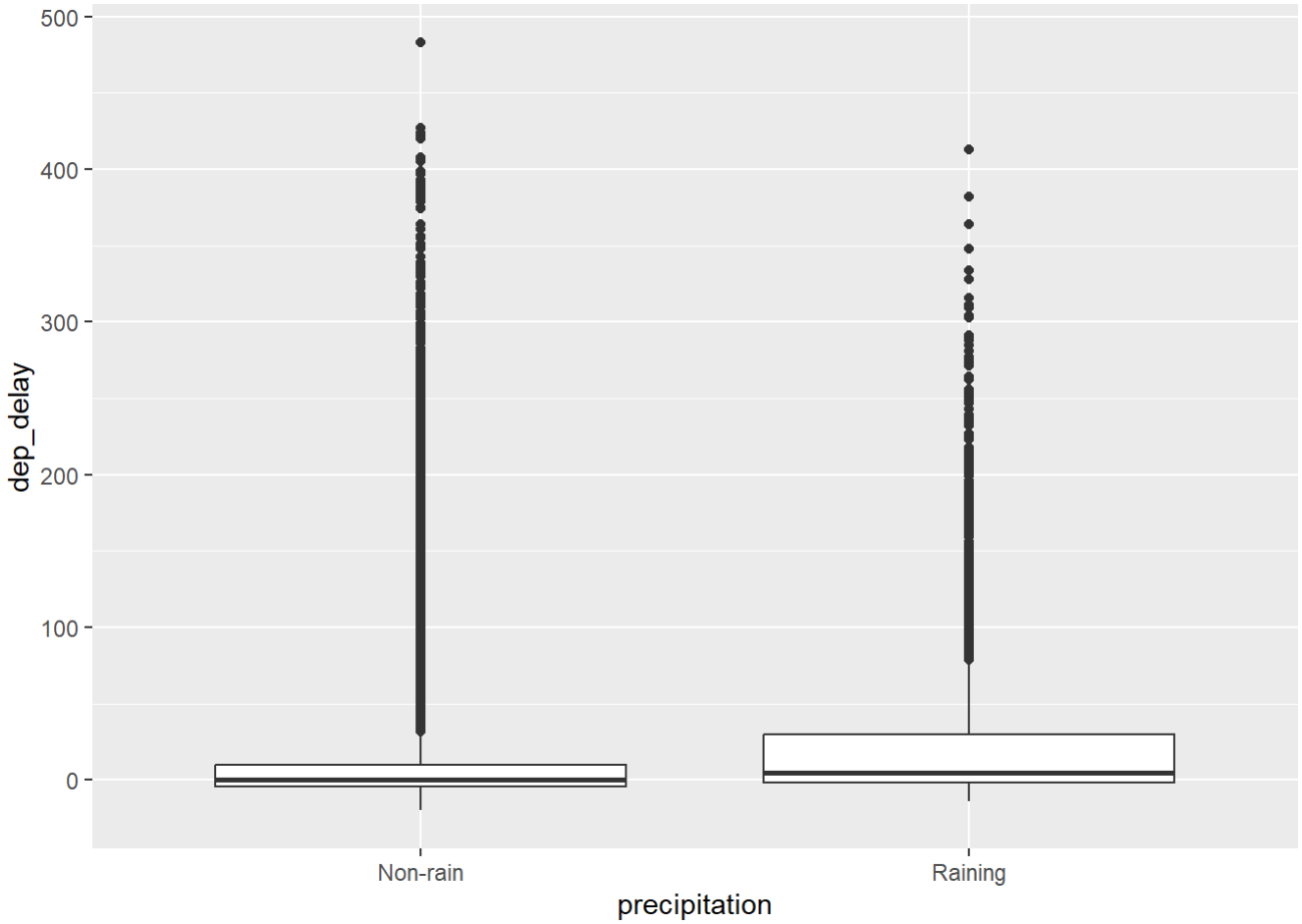

```
ggplot(data = new_df, mapping = aes(x = temperature, y = dep_delay)) +
  geom_boxplot()
```



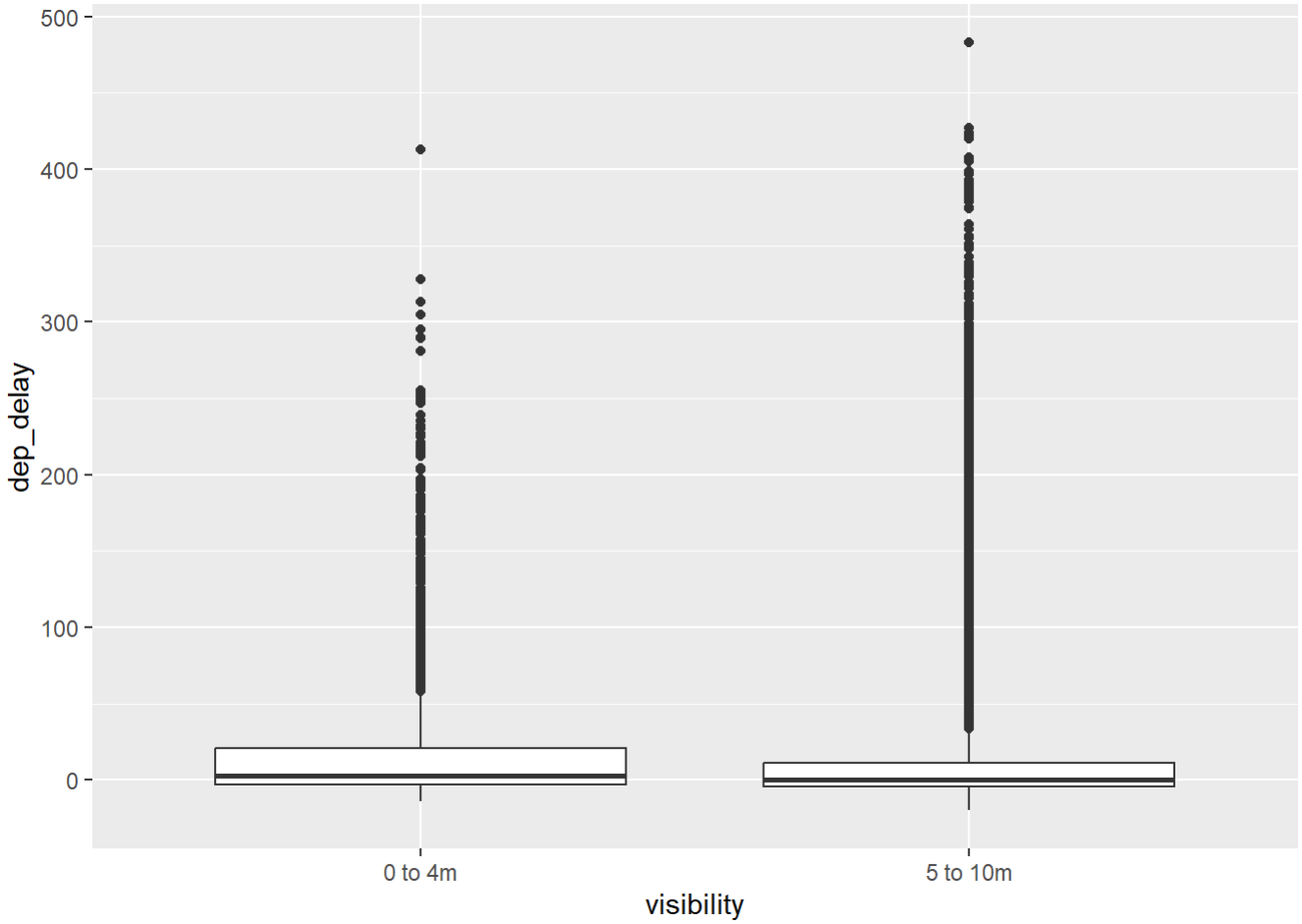

```
ggplot(data = new_df, mapping = aes(x = wind_speeds, y = dep_delay)) +
  geom_boxplot()
```




```
ggplot(data = new_df, mapping = aes(x = precipitation, y = dep_delay)) +
  geom_boxplot()
```




```
ggplot(data = new_df, mapping = aes(x = visibility, y = dep_delay)) +
  geom_boxplot()
```



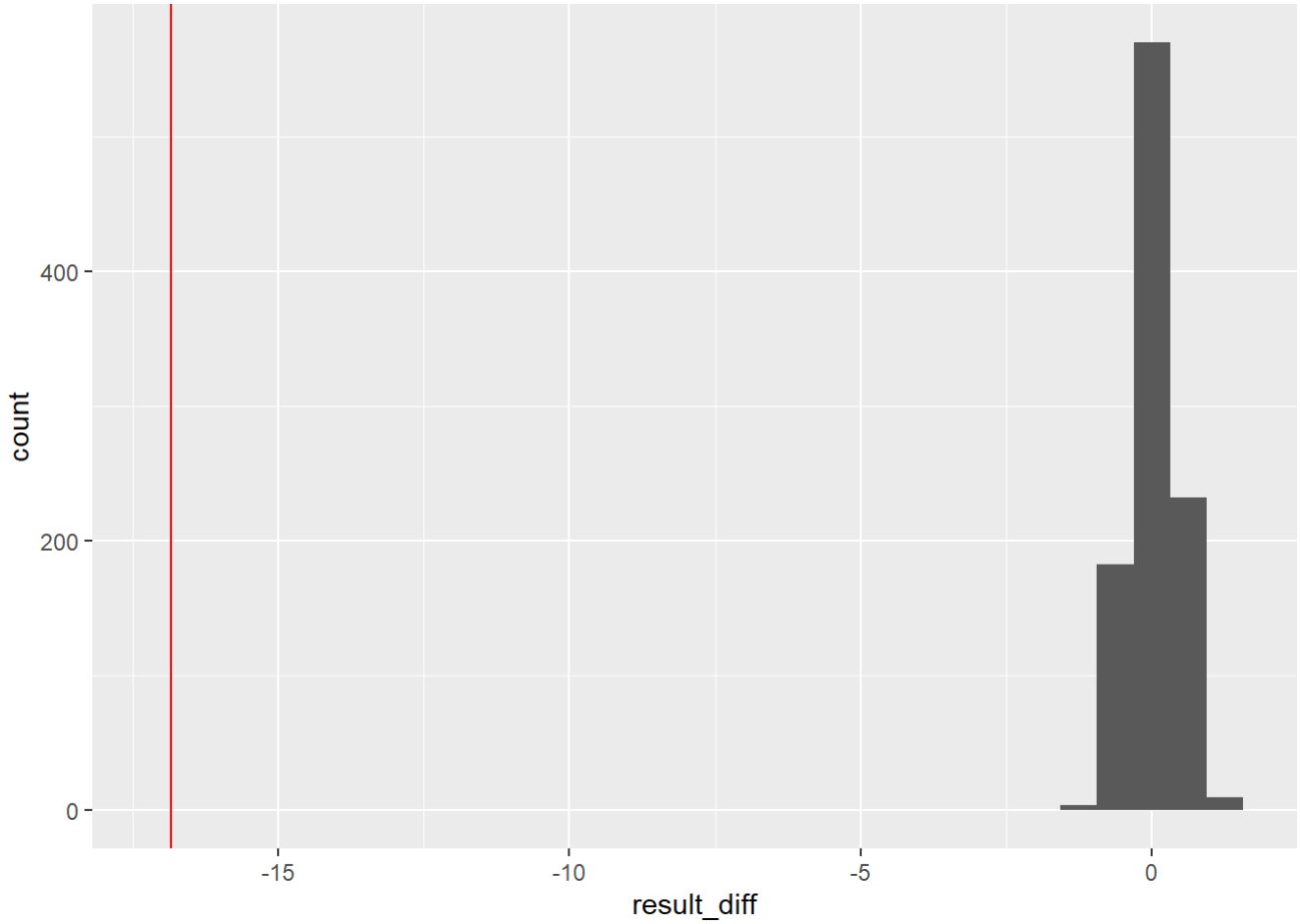
PERMUTATION TEST

```
hour <- new_df%>%
  filter(time_of_day=="Morning" | time_of_day=="Evening")


observed_mean_hour <- mean(hour$dep_delay[hour$time_of_day=="Morning"]) -mean(hour$dep_delay[hou
r$time_of_day=="Evening"])
observed_mean_hour
```

```
## [1] -16.84259
```

```
N <- 10^3-1
sample.size = nrow(hour)
group.1.size = nrow(hour[hour$time_of_day=="Morning",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(hour$dep_delay[index]) - mean(hour$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_hour, color = "red")
```

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




```
#p-value
(sum(result_diff <= observed_mean_hour) + 1) / (N + 1)
```

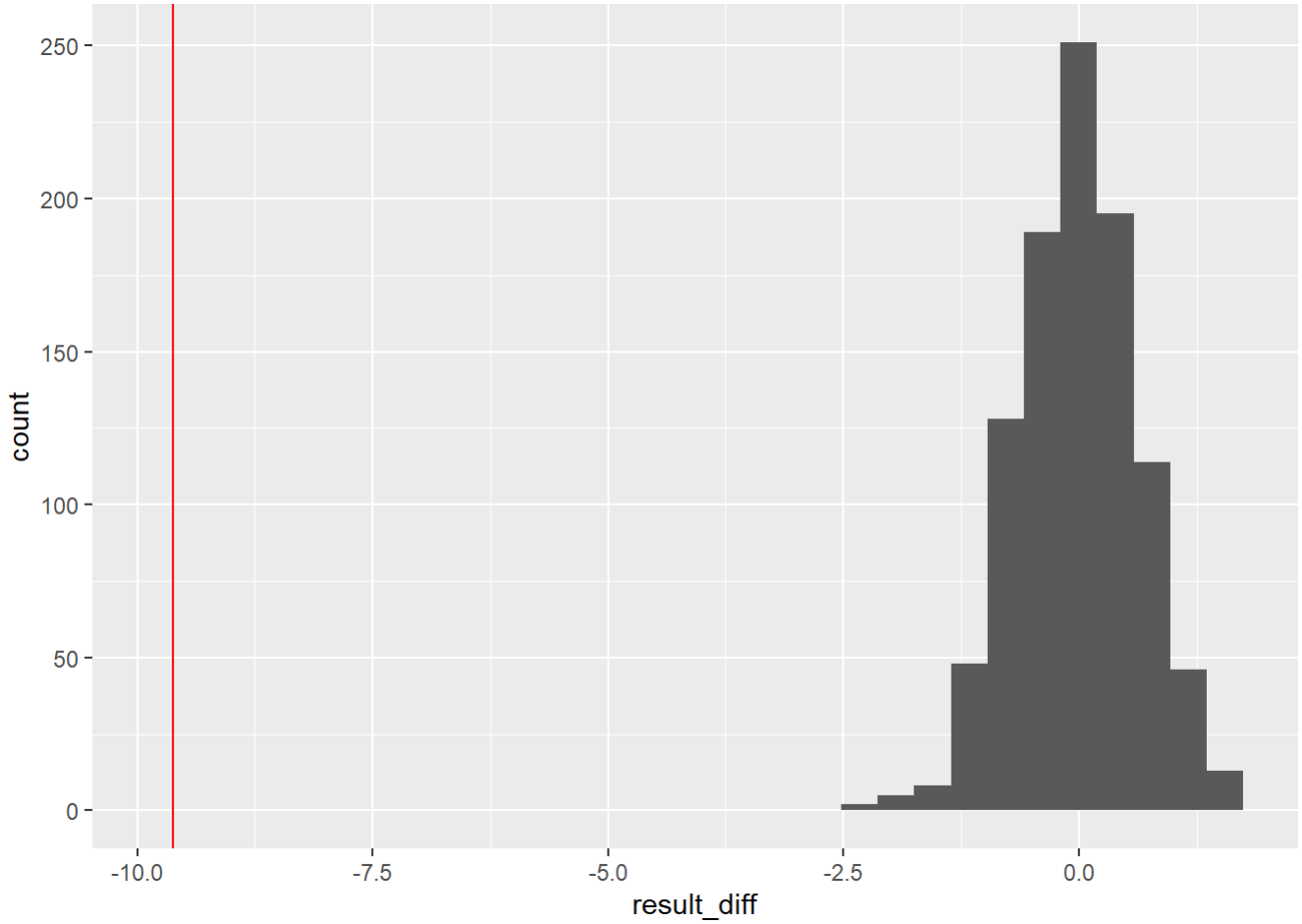
```
## [1] 0.001
```

```
hour2 <- new_df%>%
  filter(time_of_day=="Noon" | time_of_day=="Night")


observed_mean_hour2 <- mean(hour2$dep_delay[hour2$time_of_day=="Noon"]) -mean(hour2$dep_delay[ho
ur2$time_of_day=="Night"])
observed_mean_hour2
```

```
## [1] -9.615722
```

```
N <- 10^3-1
sample.size = nrow(hour2)
group.1.size = nrow(hour2[hour2$time_of_day=="Noon",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(hour2$dep_delay[index]) - mean(hour2$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_hour2, color = "red")
```

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




```
#p-value
(sum(result_diff <= observed_mean_hour2) + 1) / (N + 1)
```

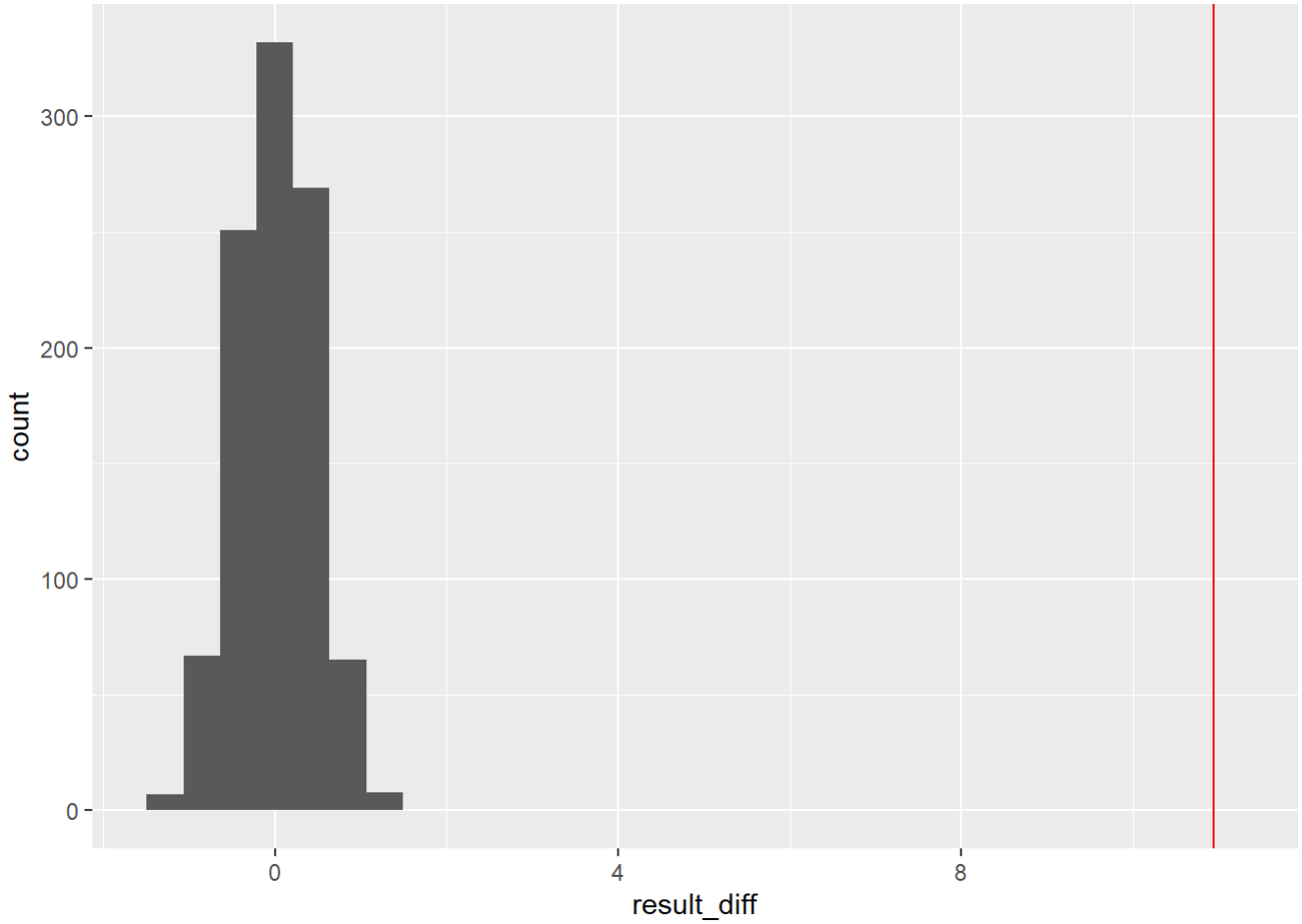
```
## [1] 0.001
```

```
season <- new_df%>%
  filter(time_of_year=="Summer" | time_of_year=="Fall")


observed_mean_year <- mean(season$dep_delay[season$time_of_year=="Summer"]) - mean(season$dep_de
lay[season$time_of_year=="Fall"])
observed_mean_year
```

```
## [1] 10.94356
```

```
N <- 10^3-1
sample.size = nrow(season)
group.1.size = nrow(season[season$time_of_year=="Summer",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(season$dep_delay[index]) - mean(season$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_year, color = "red")
```

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




```
#p-value
(sum(result_diff >= observed_mean_year) + 1) / (N + 1)
```

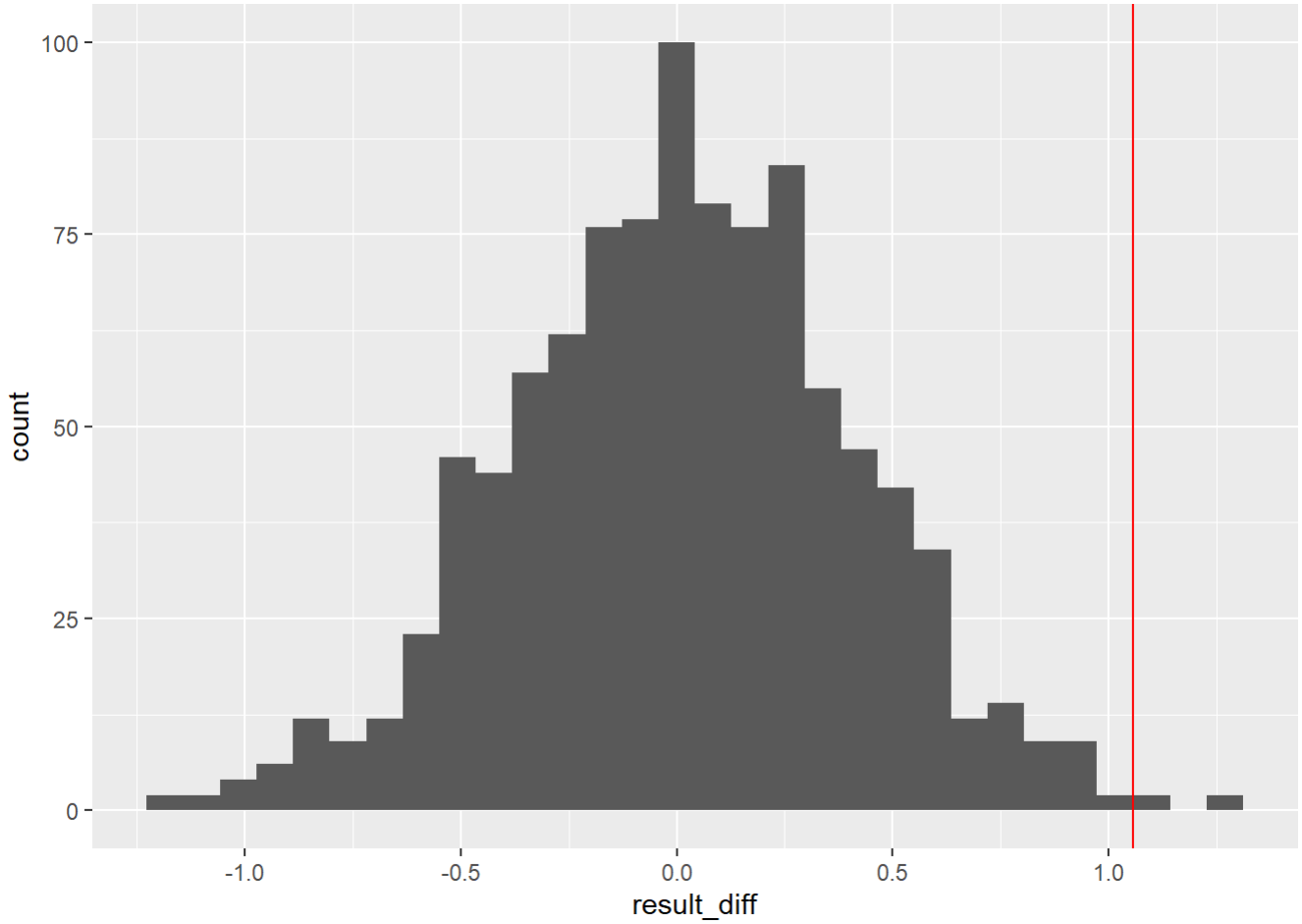
```
## [1] 0.001
```

```
season2 <- new_df%>%
  filter(time_of_year=="Spring" | time_of_year=="Winter")


observed_mean_year2 <- mean(season2$dep_delay[season2$time_of_year=="Spring"]) - mean(season2$de
p_delay[season2$time_of_year=="Winter"])
observed_mean_year2
```

```
## [1] 1.057979
```

```
N <- 10^3-1
sample.size = nrow(season2)
group.1.size = nrow(season2[season2$time_of_year=="Spring",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(season2$dep_delay[index]) - mean(season2$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_year2, color = "red")
```

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




```
#p-value
(sum(result_diff >= observed_mean_year2) + 1) / (N + 1)
```

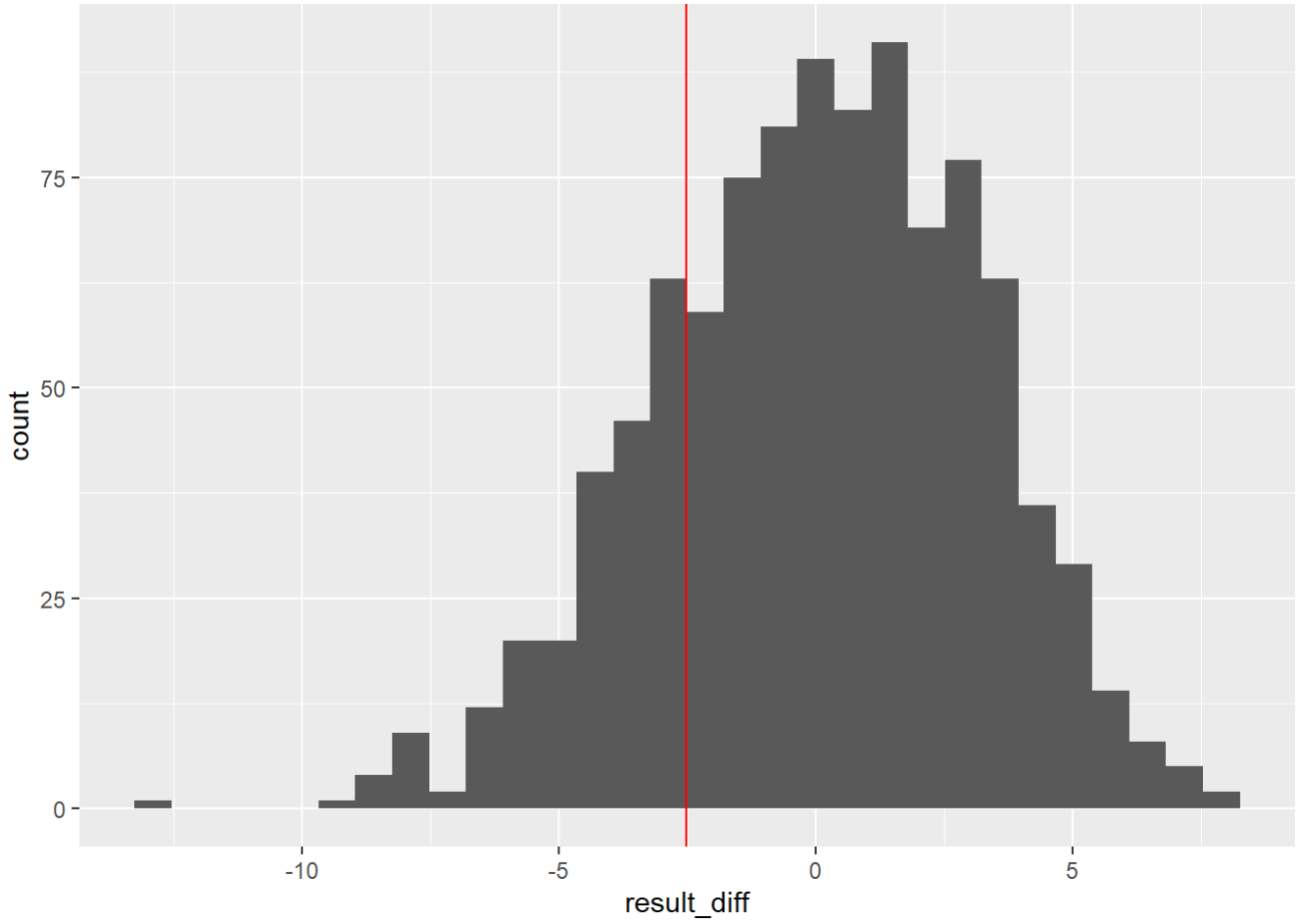
```
## [1] 0.005
```

```
windspeed <- new_df%>%
  filter(wind_speeds=="Low" | wind_speeds=="Strong")


observed_mean_wind <- mean(windspeed$dep_delay[windspeed$wind_speeds=="Low"]) -mean(windspeed$de
p_delay[windspeed$wind_speeds=="Strong"])
observed_mean_wind
```

```
## [1] -2.526209
```

```
N <- 10^3-1
sample.size = nrow(windspeed)
group.1.size = nrow(windspeed[windspeed$wind_speeds=="Low",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(windspeed$dep_delay[index]) - mean(windspeed$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_wind, color = "red")
```

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




```
#p-value
(sum(result_diff <= observed_mean_wind) + 1) / (N + 1)
```

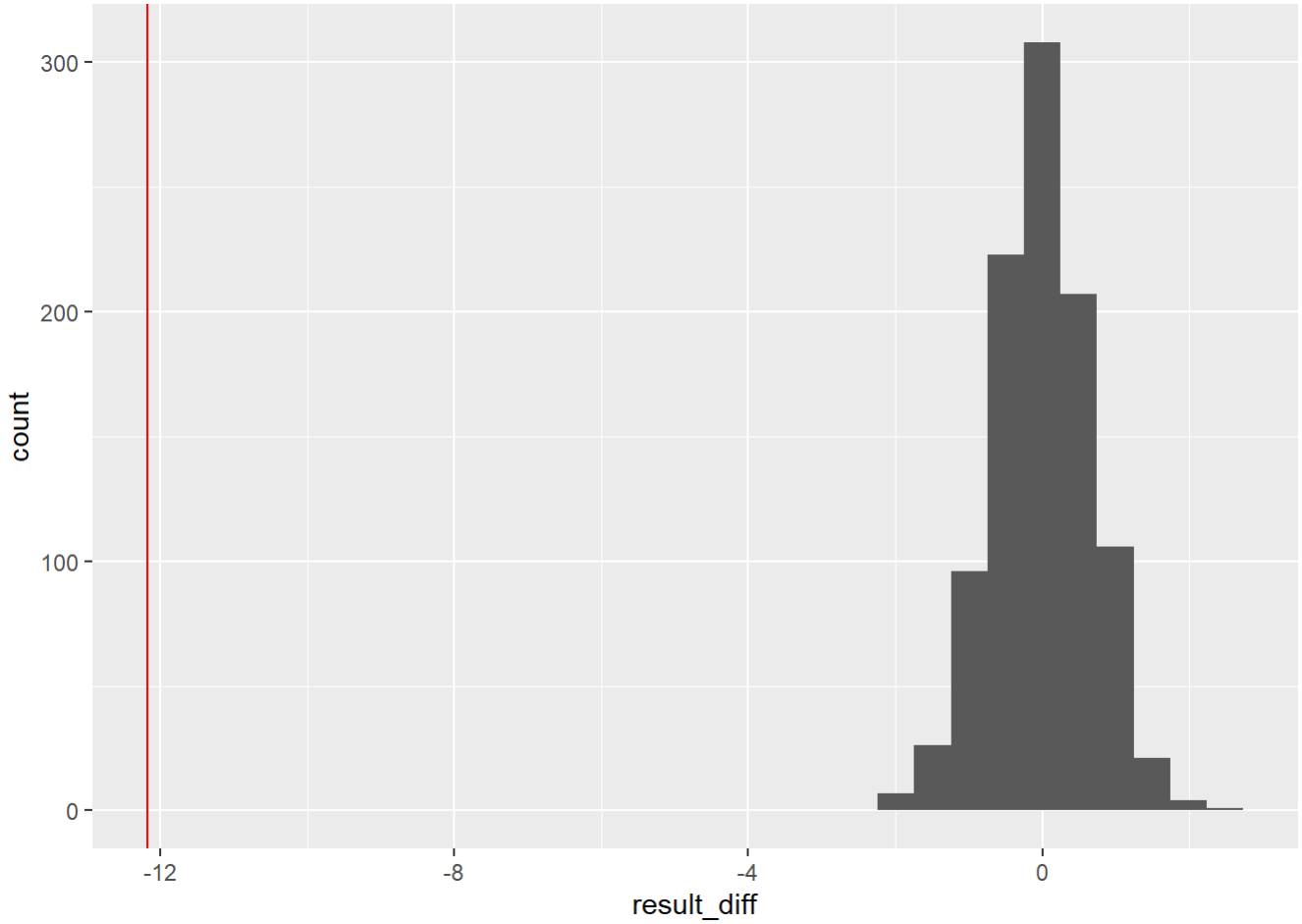
```
## [1] 0.217
```

```
temp <- new_df%>%
  filter(temperature=="Cold" | temperature=="Hot")


observed_mean_temp <- mean(temp$dep_delay[temp$temperature=="Cold"]) - mean(temp$dep_delay[temp
$temperature=="Hot"])
observed_mean_temp
```

```
## [1] -12.18133
```

```
N <- 10^3-1
sample.size = nrow(temp)
group.1.size = nrow(temp[temp$temperature=="Cold",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(temp$dep_delay[index]) - mean(temp$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_temp, color = "red")
```

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




```
#p-value
(sum(result_diff <= observed_mean_temp) + 1) / (N + 1)
```

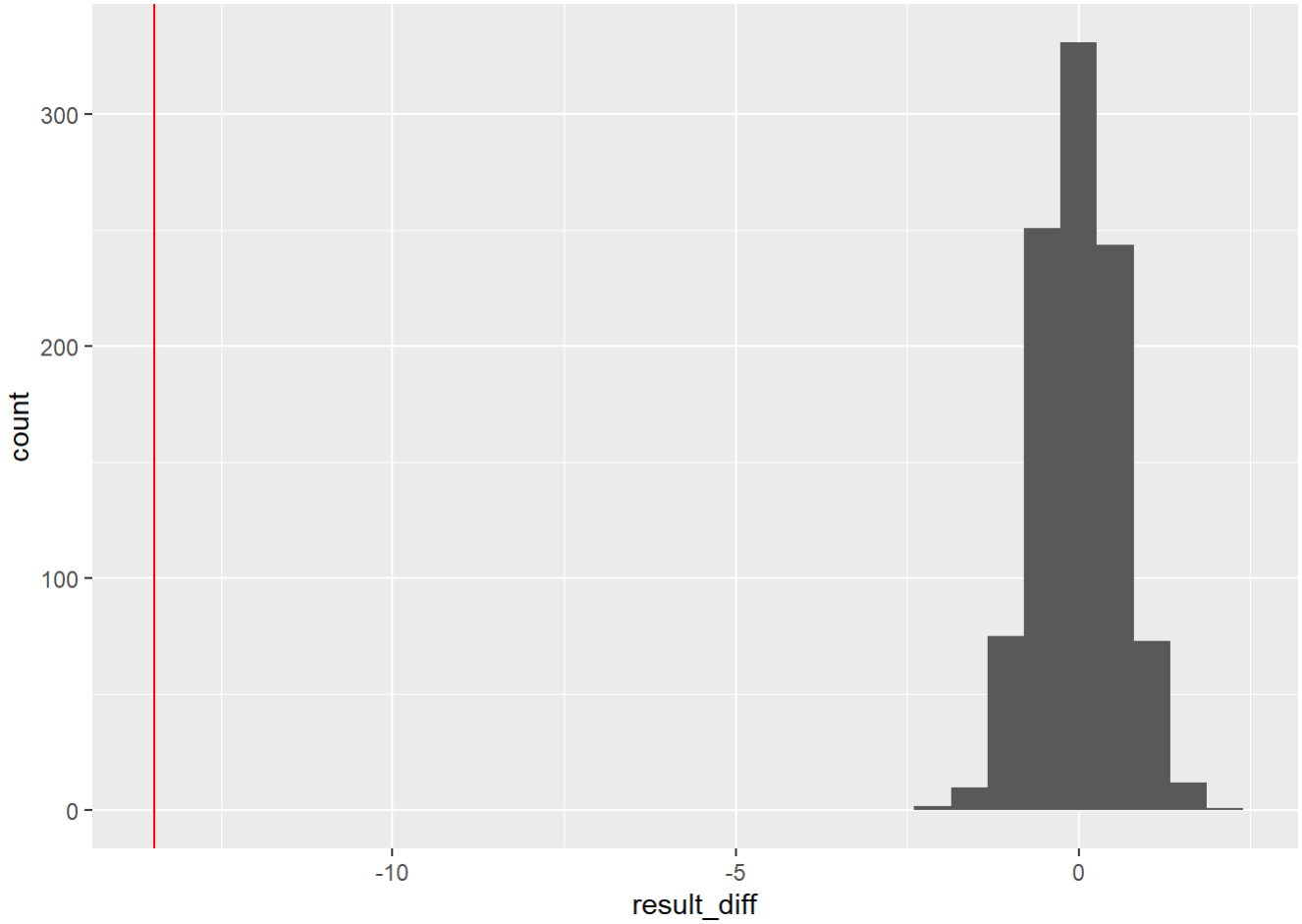
```
## [1] 0.001
```

```
observed_mean_diff <- mean(new_df$dep_delay[new_df$precipitation=="Non-rain"]) - mean(new_df$dep
_delay[new_df$precipitation=="Raining"])
observed_mean_diff
```

```
## [1] -13.46498
```

```
N <- 10^3-1
sample.size = nrow(new_df)
group.1.size = nrow(new_df[new_df$precipitation=="Non-rain",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(new_df$dep_delay[index]) - mean(new_df$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_diff, color = "red")
```

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




```
#p-value
(sum(result_diff <= observed_mean_diff) + 1) / (N + 1)
```

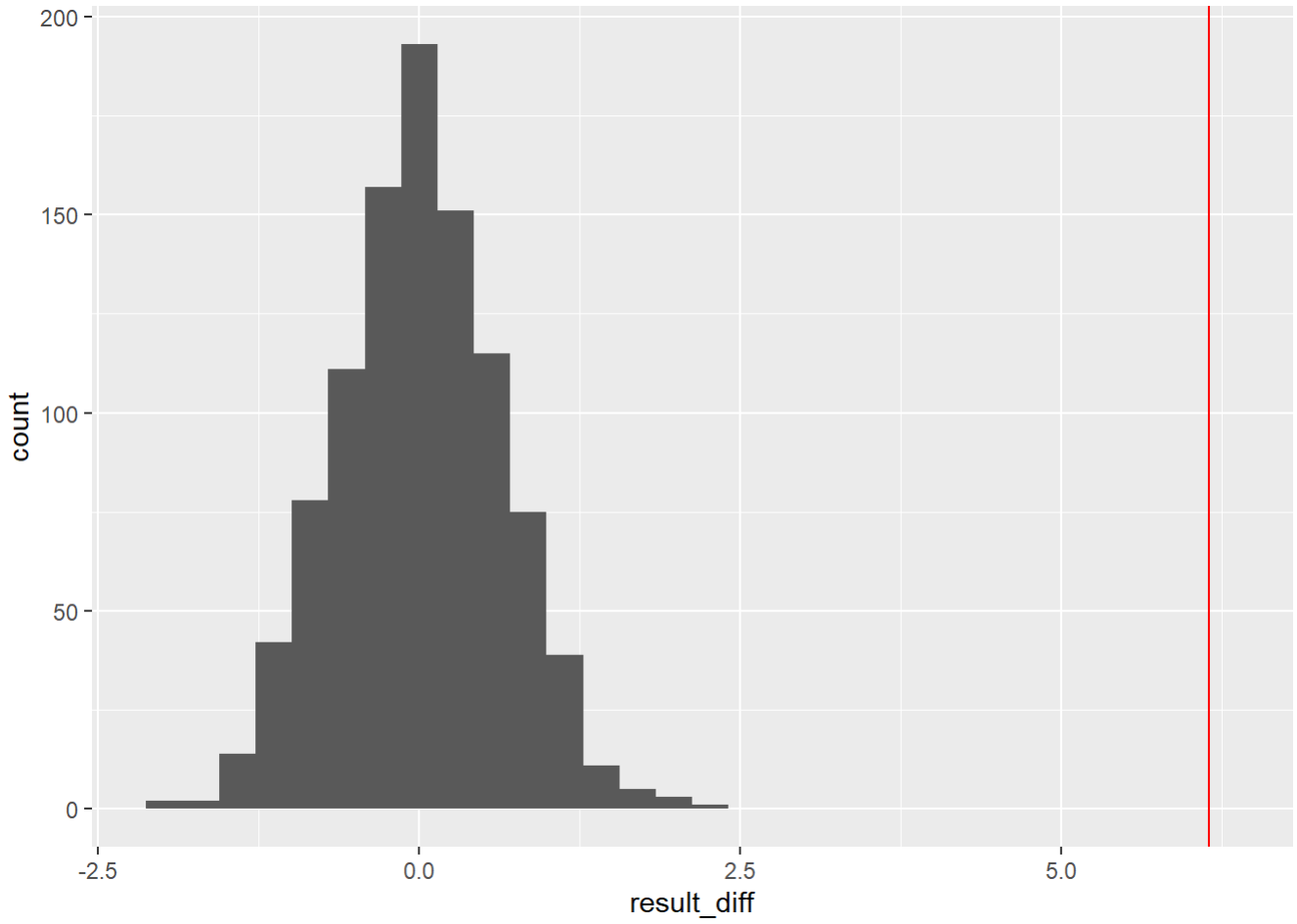
```
## [1] 0.001
```

```
observed_mean_visib <- mean(new_df$dep_delay[new_df$visibility=="0 to 4m"]) - mean(new_df$dep_de
lay[new_df$visibility=="5 to 10m"])
observed_mean_visib
```

```
## [1] 6.155003
```

```
N <- 10^3-1
sample.size = nrow(new_df)
group.1.size = nrow(new_df[new_df$visibility=="0 to 4m",])
result_diff <- numeric(N)
for(i in 1:N)
{
  index = sample(sample.size, size=group.1.size, replace = FALSE)
  result_diff[i] = mean(new_df$dep_delay[index]) - mean(new_df$dep_delay[-index])
}

#plot a histogram of the simulated differences
#add a vertical line at the observed difference
ggplot(data=tibble(result_diff), mapping = aes(x=result_diff)) +
  geom_histogram() +
  geom_vline(xintercept = observed_mean_visib, color = "red")
```

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



```
#p-value
(sum(result_diff >= observed_mean_visib) + 1) / (N + 1)
```

```
## [1] 0.001
```