Gain Analysis of United Airlines

Vivek Reddy Karra

2023-11-20

```
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")
```

```
UA_flights <- flights %>%
  filter(carrier=="UA")
```

```
UA_flights_new <- na.omit(UA_flights)
```

```
UA_flights <- UA_flights_new %>%
  mutate(net_gain = dep_delay - arr_delay,
         late = dep_delay > 0 ,
    very_late = dep_delay > 30,
    flight_duration = case_when(
    hour < 6 ~ "Short",
    hour >= 6 ~ "Longer"),
    gain_per_hour = net_gain/hour
    )
```

```
summary(UA_flights)
```

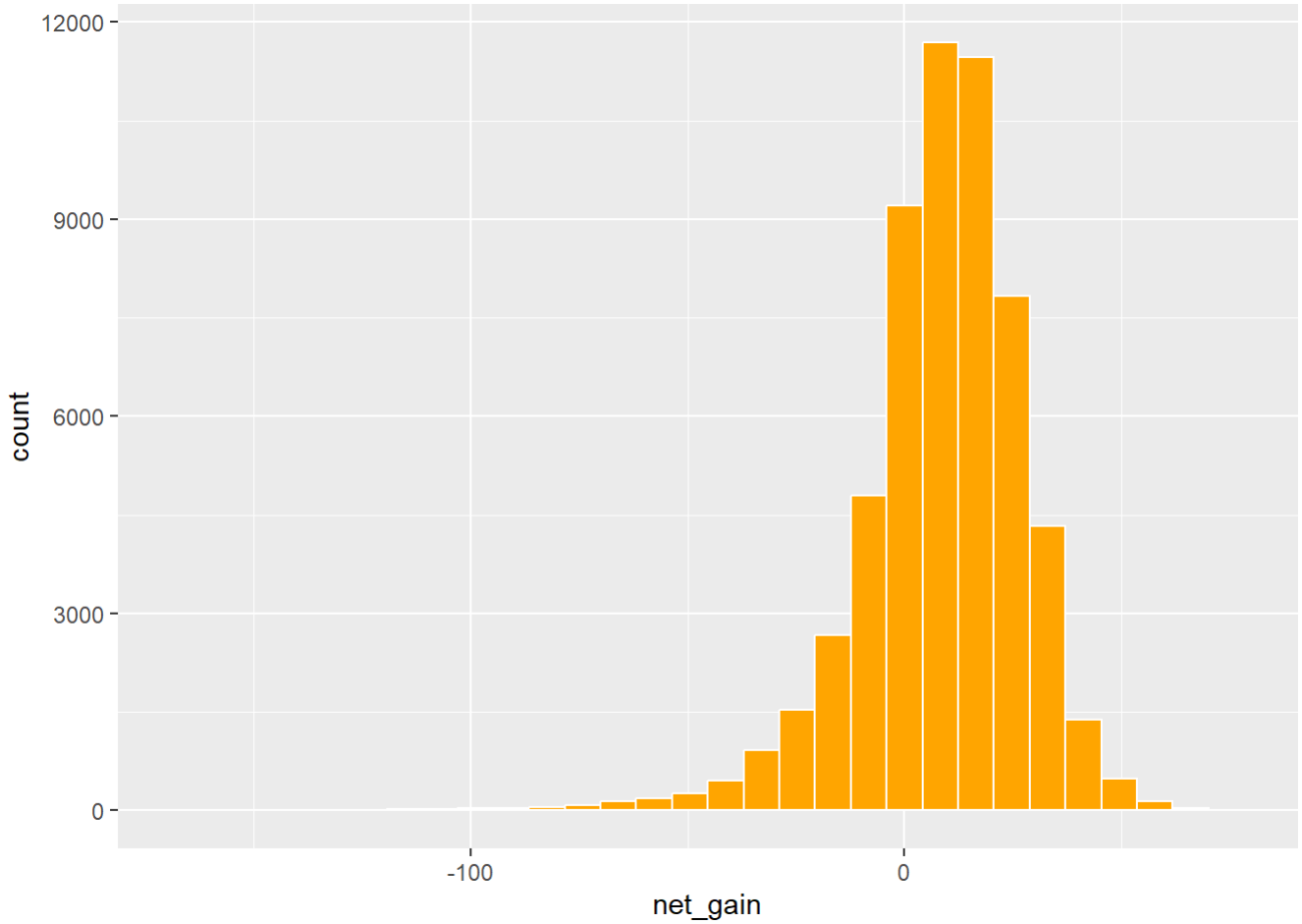
```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 500
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 855   1st Qu.: 855
##  Median :2013   Median : 7.000   Median :16.00   Median :1353   Median :1343
##  Mean   :2013   Mean   : 6.573   Mean   :15.73   Mean   :1327   Mean   :1312
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1733   3rd Qu.:1722
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2358   Max.   :2345
##    dep_delay         arr_time    sched_arr_time   arr_delay
##  Min.   :-20.00   Min.   :   1   Min.   :   1   Min.   :-75.000
##  1st Qu.: -4.00   1st Qu.:1112   1st Qu.:1136   1st Qu.:-18.000
##  Median :  0.00   Median :1547   Median :1607   Median : -6.000
##  Mean   : 12.02   Mean   :1508   Mean   :1543   Mean   :  3.558
##  3rd Qu.: 11.00   3rd Qu.:1944   3rd Qu.:1950   3rd Qu.: 12.000
##  Max.   :483.00   Max.   :2359   Max.   :2359   Max.   :455.000
##    carrier              flight         tailnum             origin
##  Length:57782       Min.   :   1.0   Length:57782       Length:57782
##  Class :character   1st Qu.: 504.0   Class :character   Class :character
##  Mode  :character   Median :1053.0   Mode  :character   Mode  :character
##                     Mean   : 961.8
##                     3rd Qu.:1431.0
##                     Max.   :1744.0
##      dest              air_time        distance         hour
##  Length:57782       Min.   : 23.0   Min.   : 116   Min.   : 5.00
##  Class :character   1st Qu.:135.0   1st Qu.: 937   1st Qu.: 8.00
##  Mode  :character   Median :197.0   Median :1400   Median :13.00
##                     Mean   :211.8   Mean   :1531   Mean   :12.85
##                     3rd Qu.:313.0   3rd Qu.:2425   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4963   Max.   :23.00
##      minute        time_hour                       net_gain
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00.00   Min.   :-165.000
##  1st Qu.: 9.00   1st Qu.:2013-04-05 14:00:00.00   1st Qu.:  -1.000
##  Median :29.00   Median :2013-07-03 20:00:00.00   Median :  10.000
##  Mean   :26.74   Mean   :2013-07-03 22:57:32.52   Mean   :   8.459
##  3rd Qu.:44.00   3rd Qu.:2013-10-02 11:00:00.00   3rd Qu.:  20.000
##  Max.   :59.00   Max.   :2013-12-31 21:00:00.00   Max.   :  74.000
##     late          very_late       flight_duration    gain_per_hour
##  Mode :logical   Mode :logical   Length:57782       Min.   :-23.0000
##  FALSE:30657     FALSE:50232     Class :character   1st Qu.: -0.0625
##  TRUE :27125     TRUE :7550      Mode  :character   Median :  0.8235
##                                                     Mean   :  0.7885
##                                                     3rd Qu.:  1.7143
##                                                     Max.   : 10.8333
```

```
table <- cbind(
summary(UA_flights$dep_delay),
summary(UA_flights$net_gain),
summary(UA_flights$gain_per_hour))
columns <- c("dep_delay", "net_gain", "gain_per_hour")
colnames(table)<- columns
data.frame(table)
```

```
##         dep_delay    net_gain gain_per_hour
## Min.    -20.00000 -165.000000   -23.0000000
## 1st Qu.  -4.00000   -1.000000    -0.0625000
## Median    0.00000   10.000000     0.8235294
## Mean     12.01691    8.458897     0.7885025
## 3rd Qu.  11.00000   20.000000     1.7142857
## Max.    483.00000   74.000000    10.8333333
```

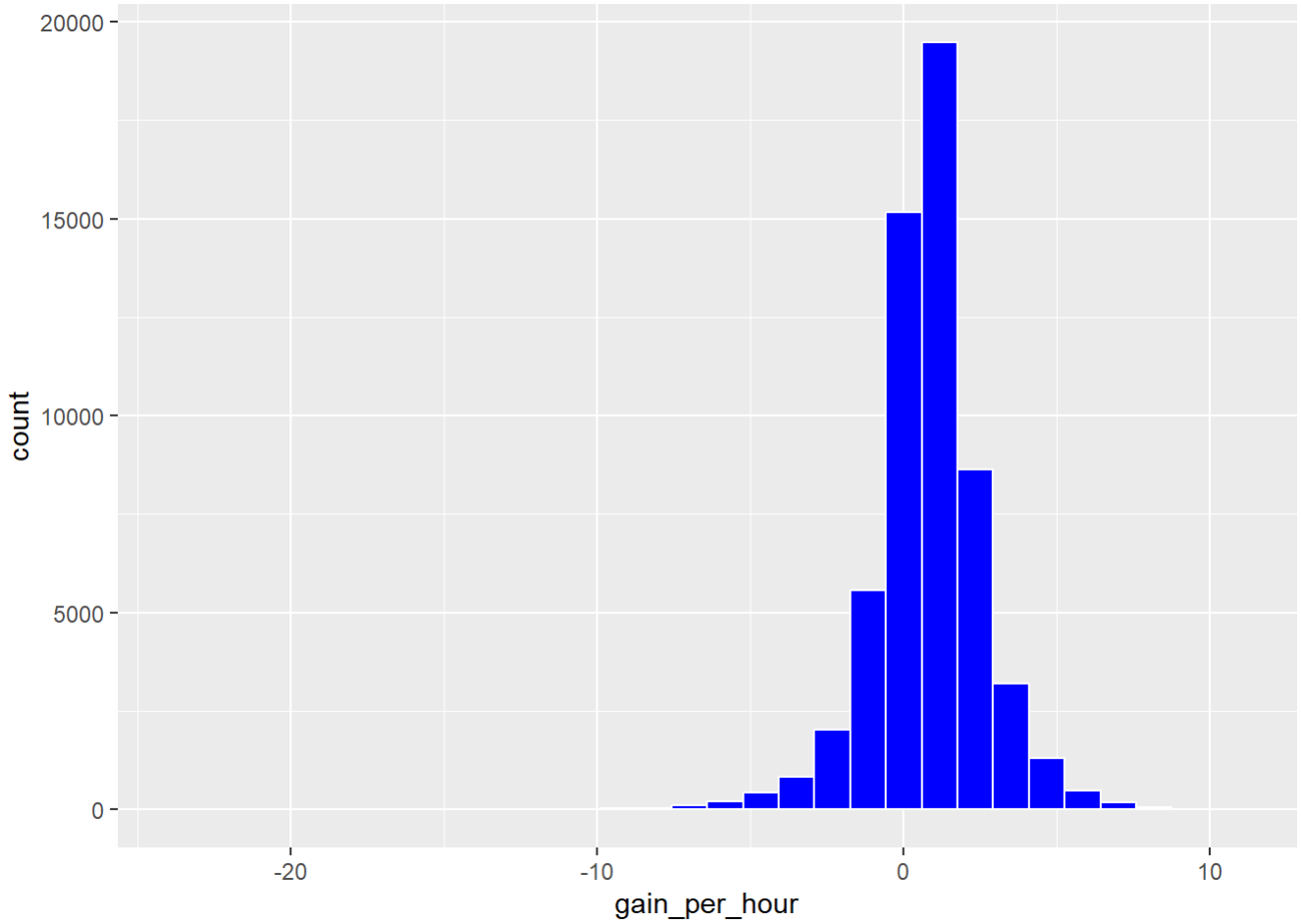
```
ggplot(data = UA_flights, mapping=aes(x=net_gain))+
  geom_histogram(color="white", fill="orange")
```

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

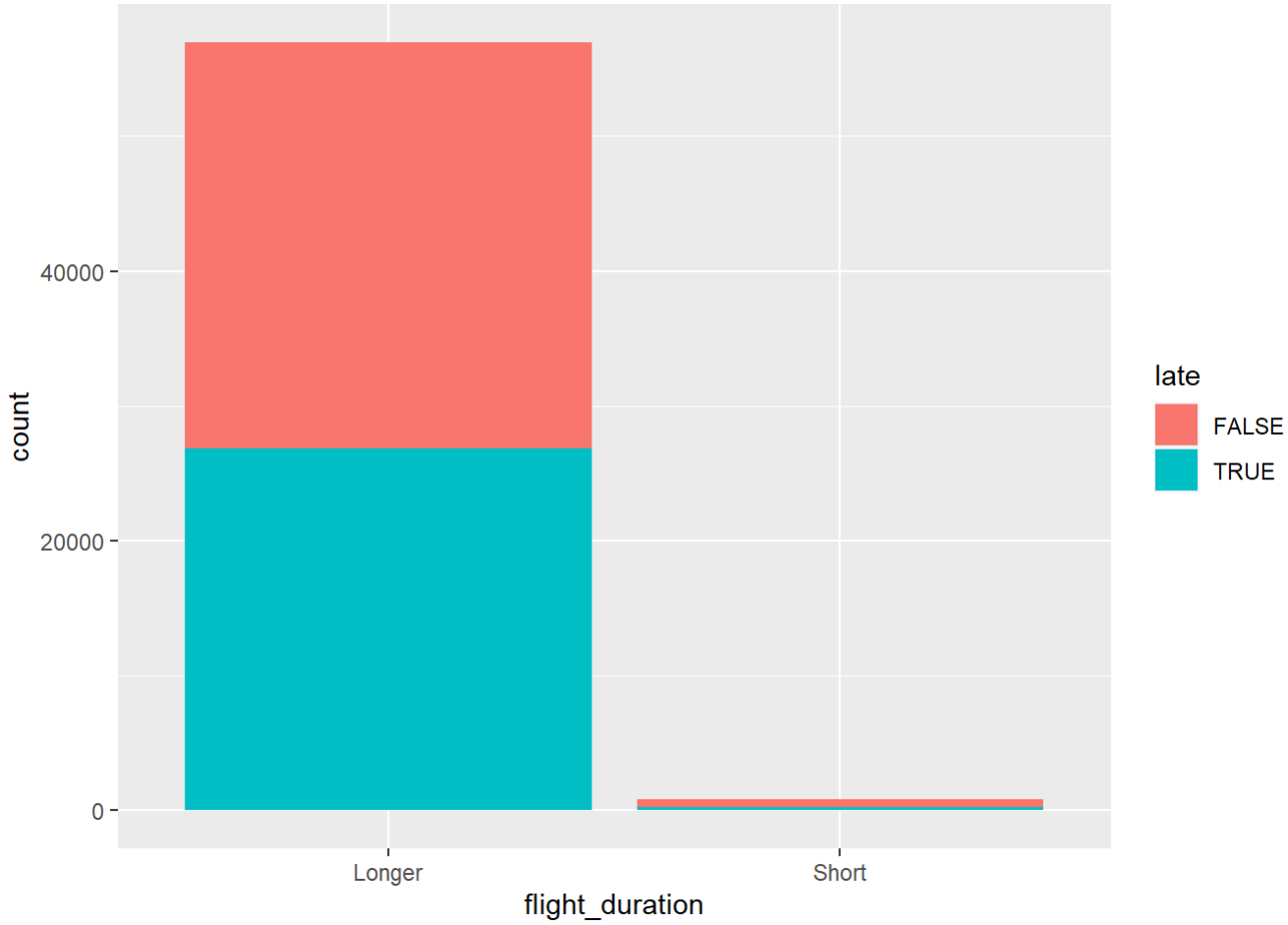



```
ggplot(data = UA_flights, mapping=aes(x=gain_per_hour))+
  geom_histogram(color="white", fill="blue")
```

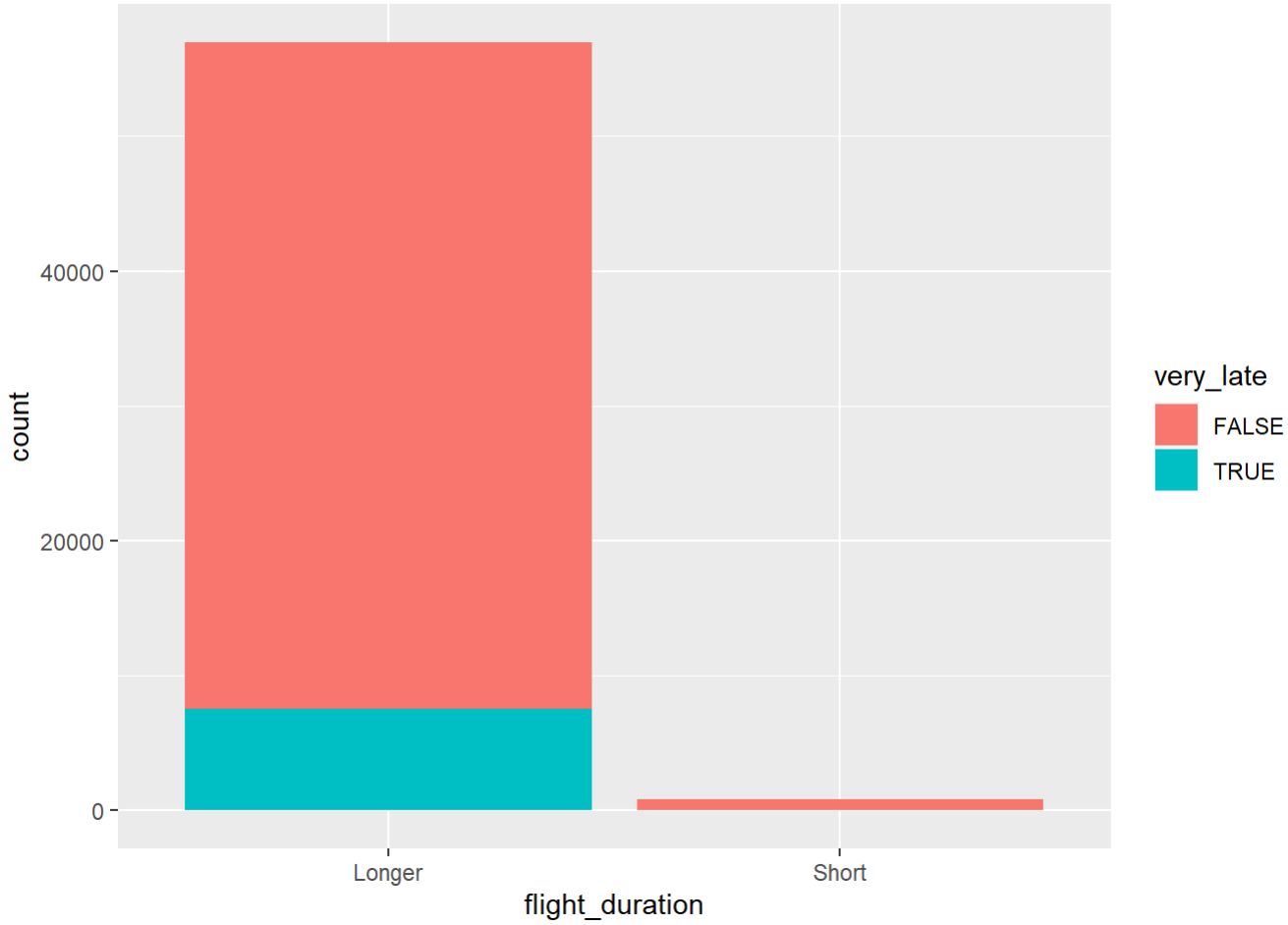
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



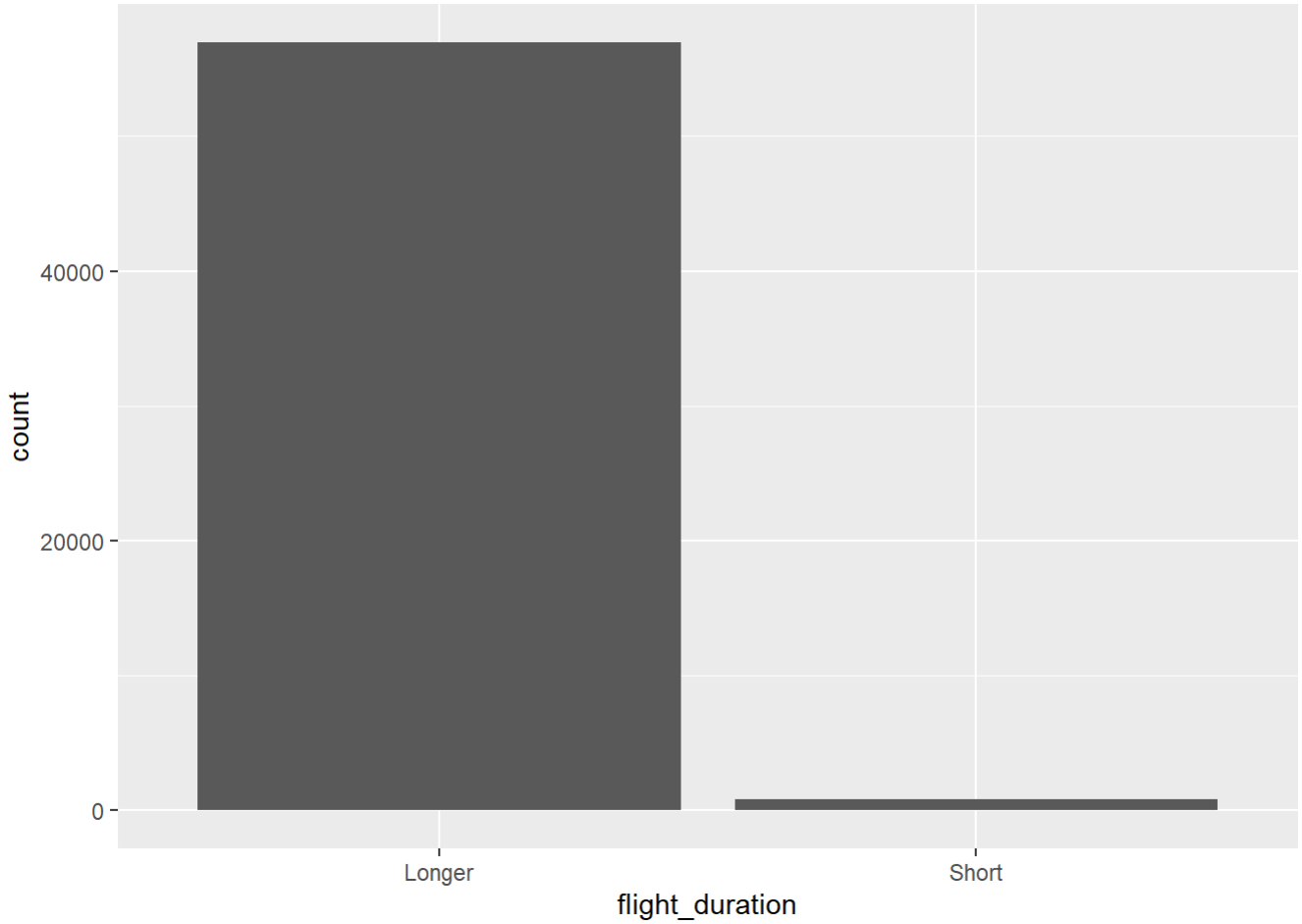

```
ggplot(data = UA_flights, mapping = aes(x = flight_duration, fill = late)) +
  geom_bar()
```



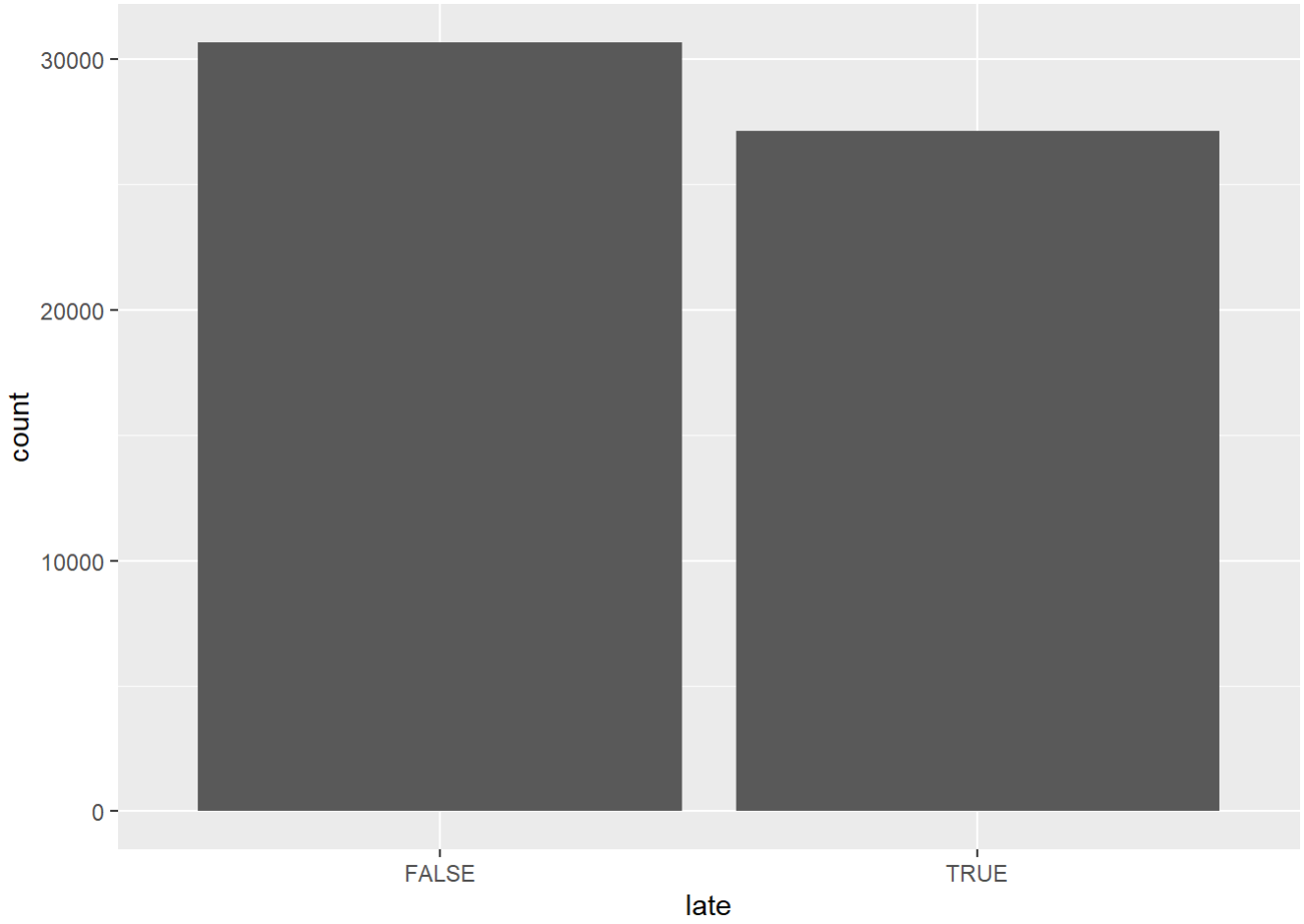

```
ggplot(data = UA_flights, mapping = aes(x = flight_duration, fill = very_late)) +
  geom_bar()
```



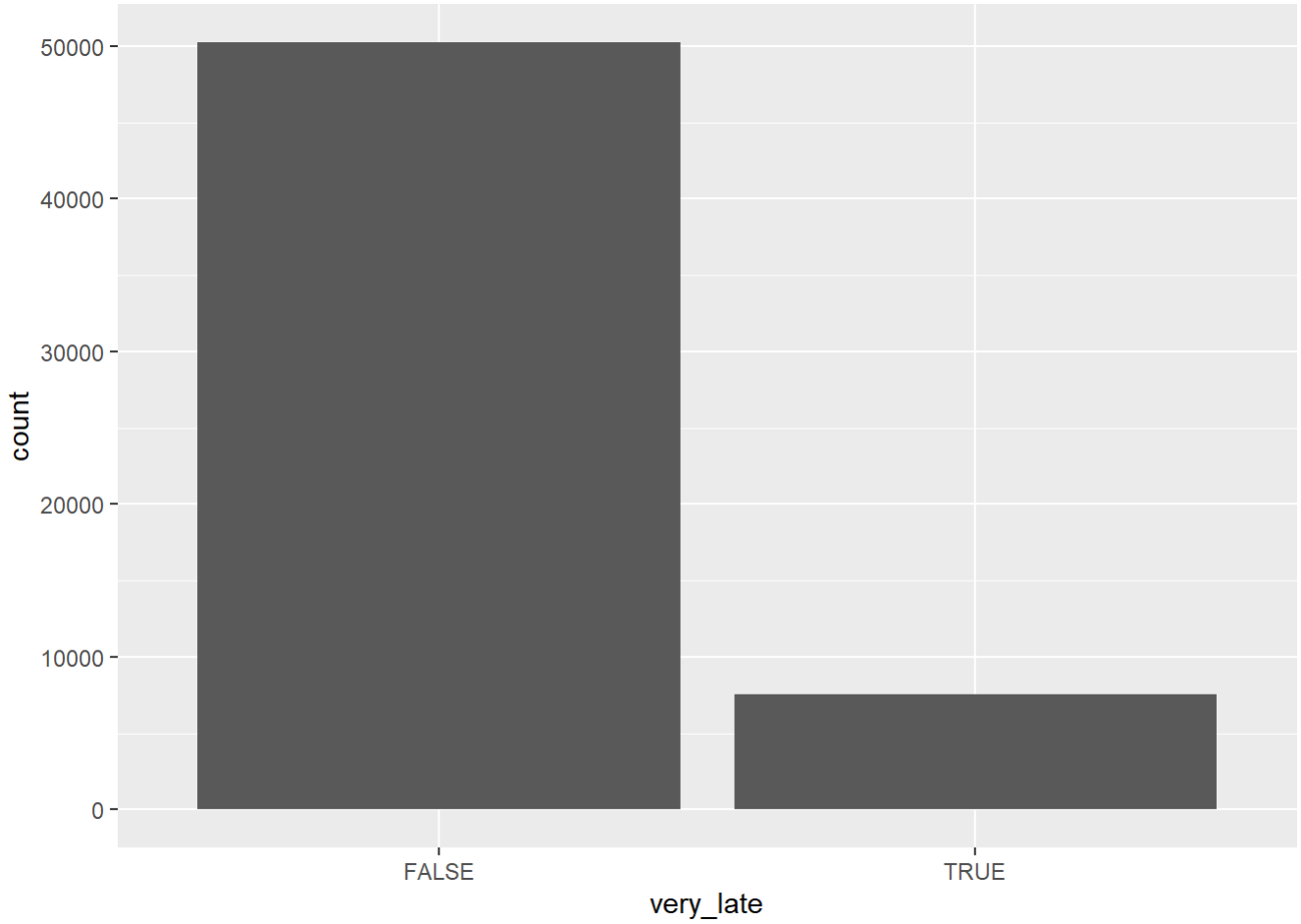

```
ggplot(data = UA_flights, mapping = aes(x = flight_duration)) +
  geom_bar()
```



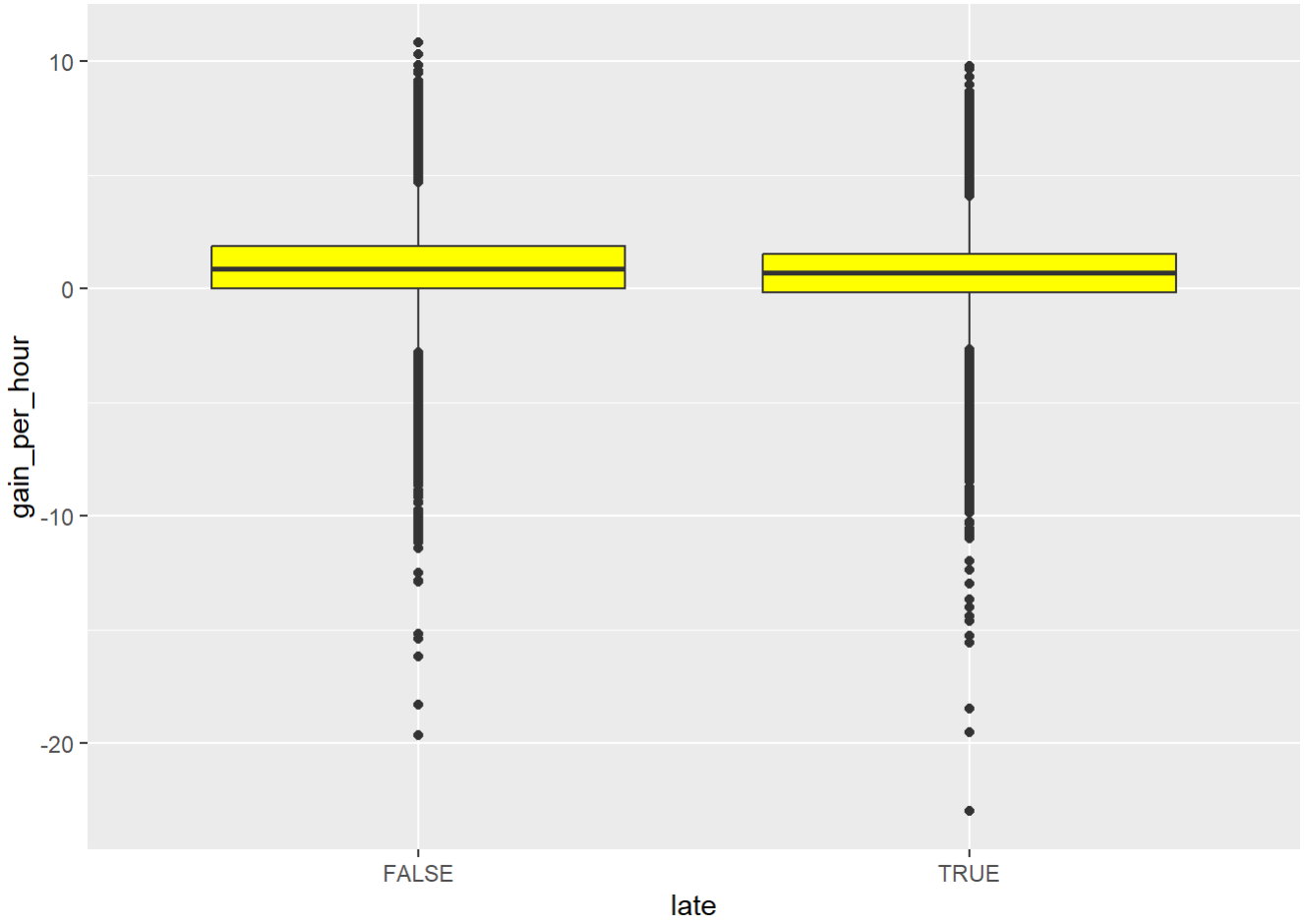

```
ggplot(data = UA_flights, mapping = aes(x = late)) +
  geom_bar()
```



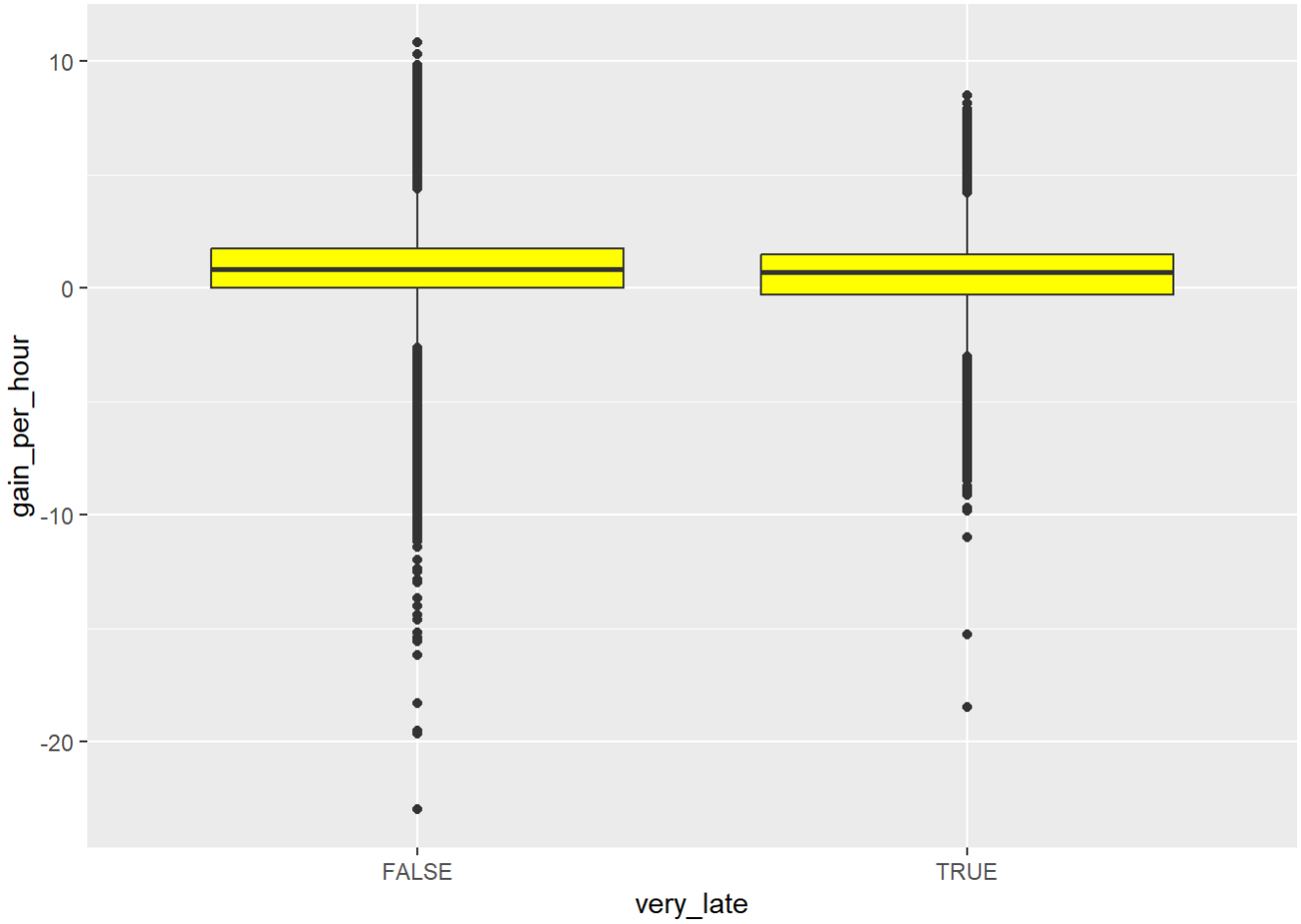

```
ggplot(data = UA_flights, mapping = aes(x = very_late)) +
  geom_bar()
```



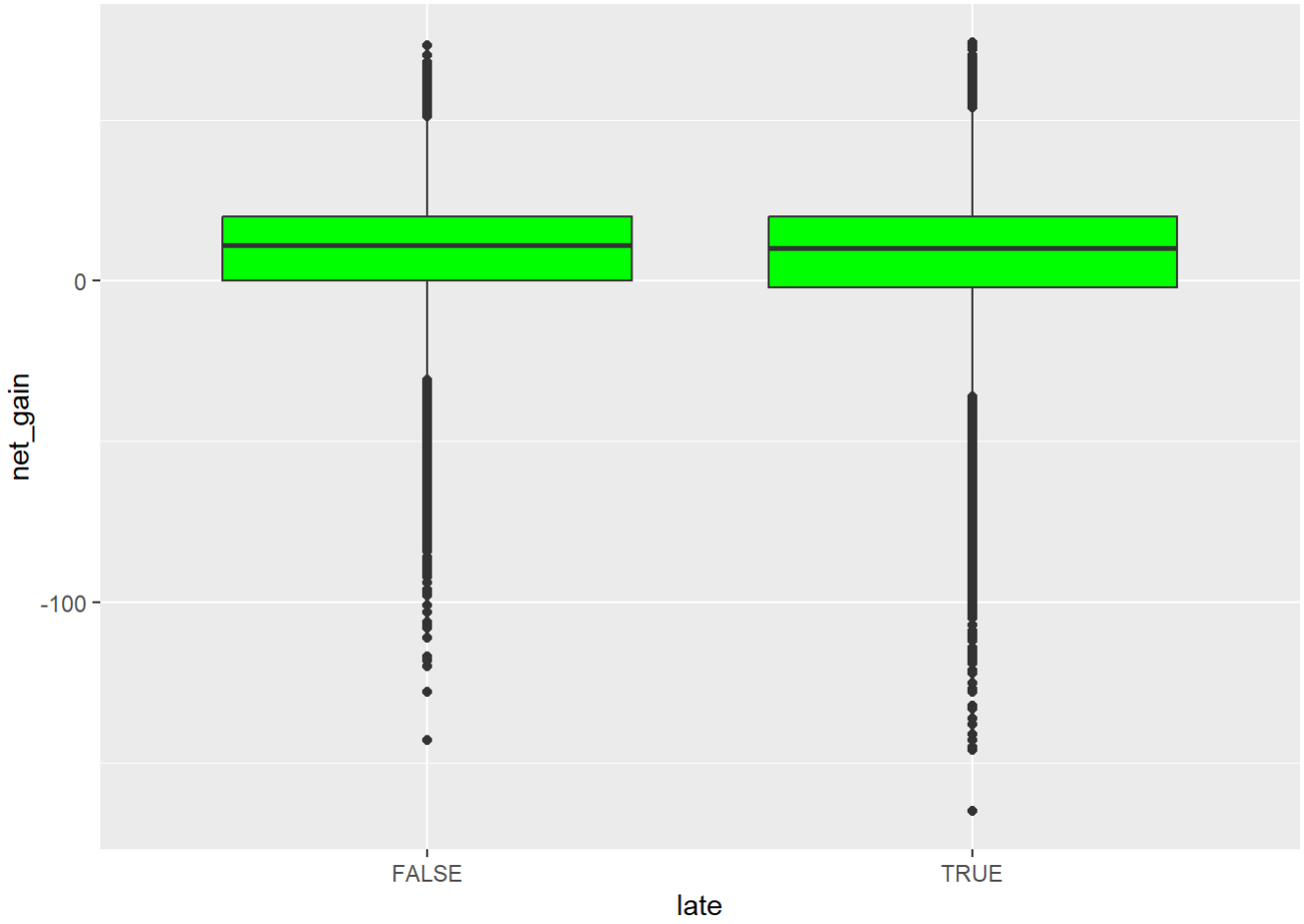

```
ggplot(data = UA_flights, mapping = aes(x = late, y = gain_per_hour)) +
  geom_boxplot(fill="yellow")
```



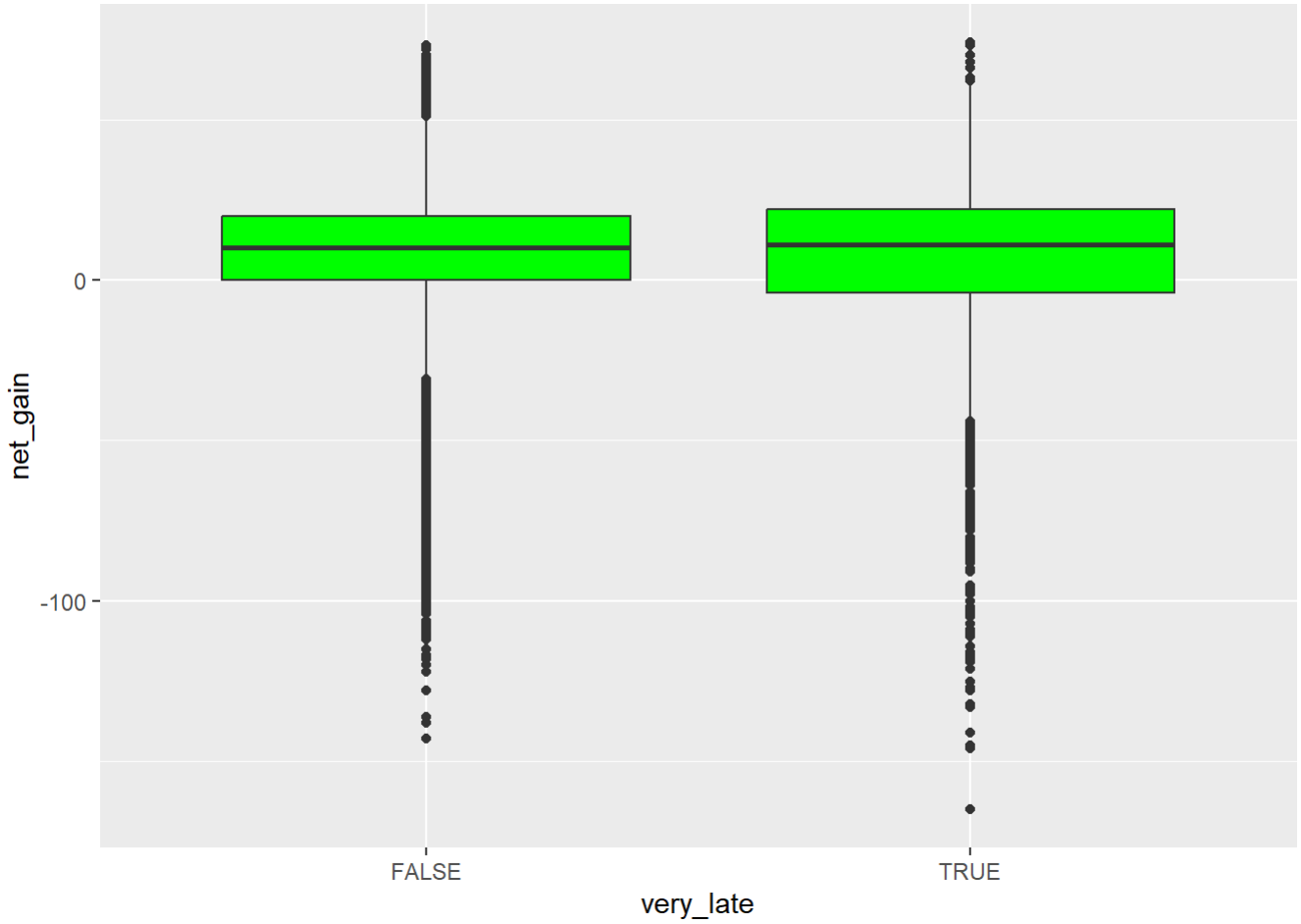

```
ggplot(data = UA_flights, mapping = aes(x = very_late, y = gain_per_hour)) +
  geom_boxplot(fill="yellow")
```




```
ggplot(data = UA_flights, mapping = aes(x = late, y = net_gain)) +
  geom_boxplot(fill="green")
```




```
ggplot(data = UA_flights, mapping = aes(x = very_late, y = net_gain)) +
  geom_boxplot(fill="green")
```

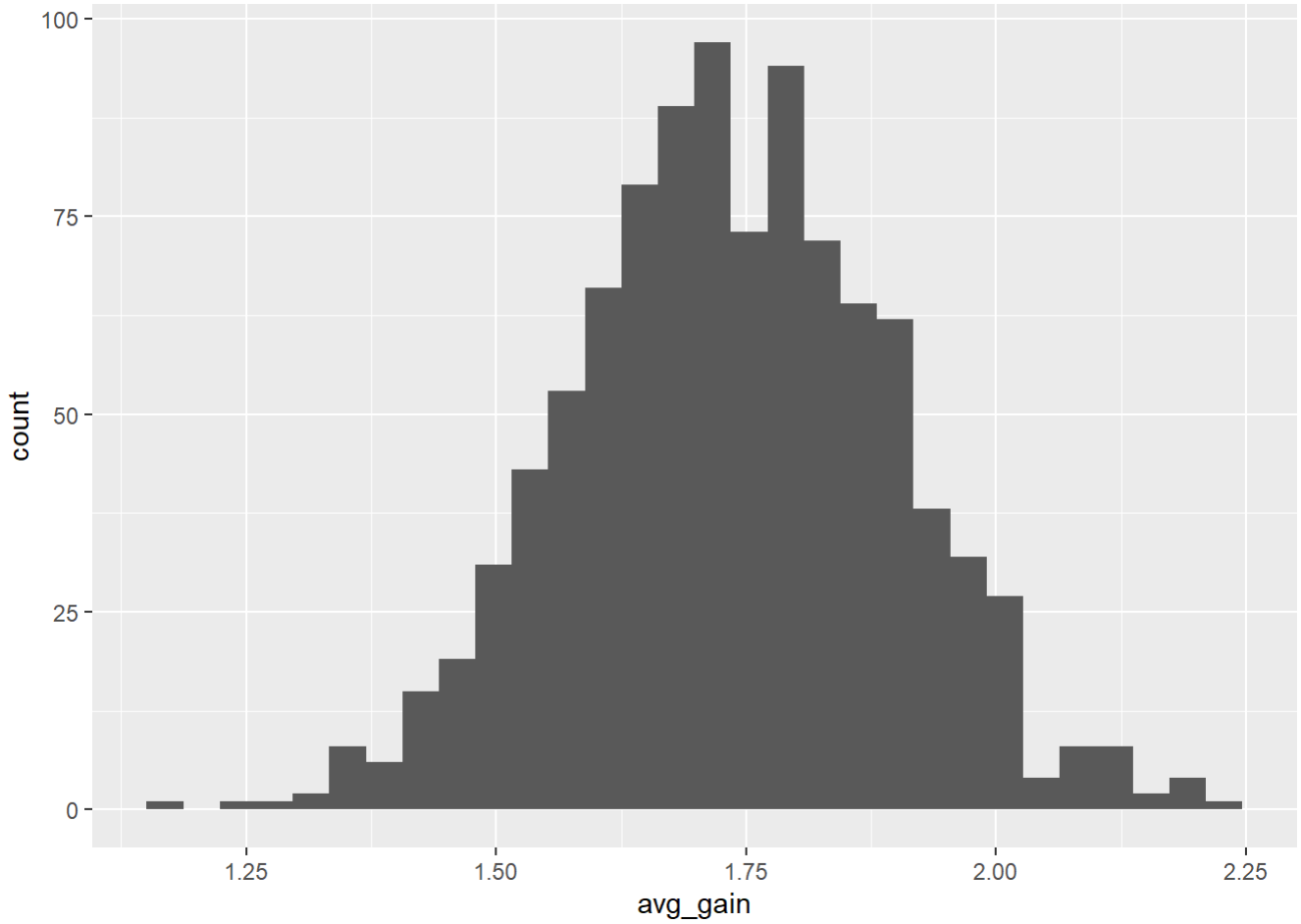



```
UA_flight.not_late <- UA_flights$net_gain[UA_flights$late == "FALSE"]
UA_flight.late <- UA_flights$net_gain[UA_flights$late == "TRUE"]

n.late <- length(UA_flight.late)
n.not_late <- length(UA_flight.not_late)

avg_gain <- numeric(1000)
for(i in 1:1000)
{
  sample.late <- sample(UA_flight.late, size = n.late, replace = TRUE)
  sample.not_late <- sample(UA_flight.not_late, size = n.not_late, replace = TRUE)
  avg_gain[i] <- mean(sample.not_late) - mean(sample.late)
}
ggplot(data=tibble(avg_gain), mapping = aes(x = avg_gain)) +
  geom_histogram()
```

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




```
quantile(avg_gain, c(.025, .975))
```

```
##     2.5%    97.5%
## 1.429875 2.041232
```

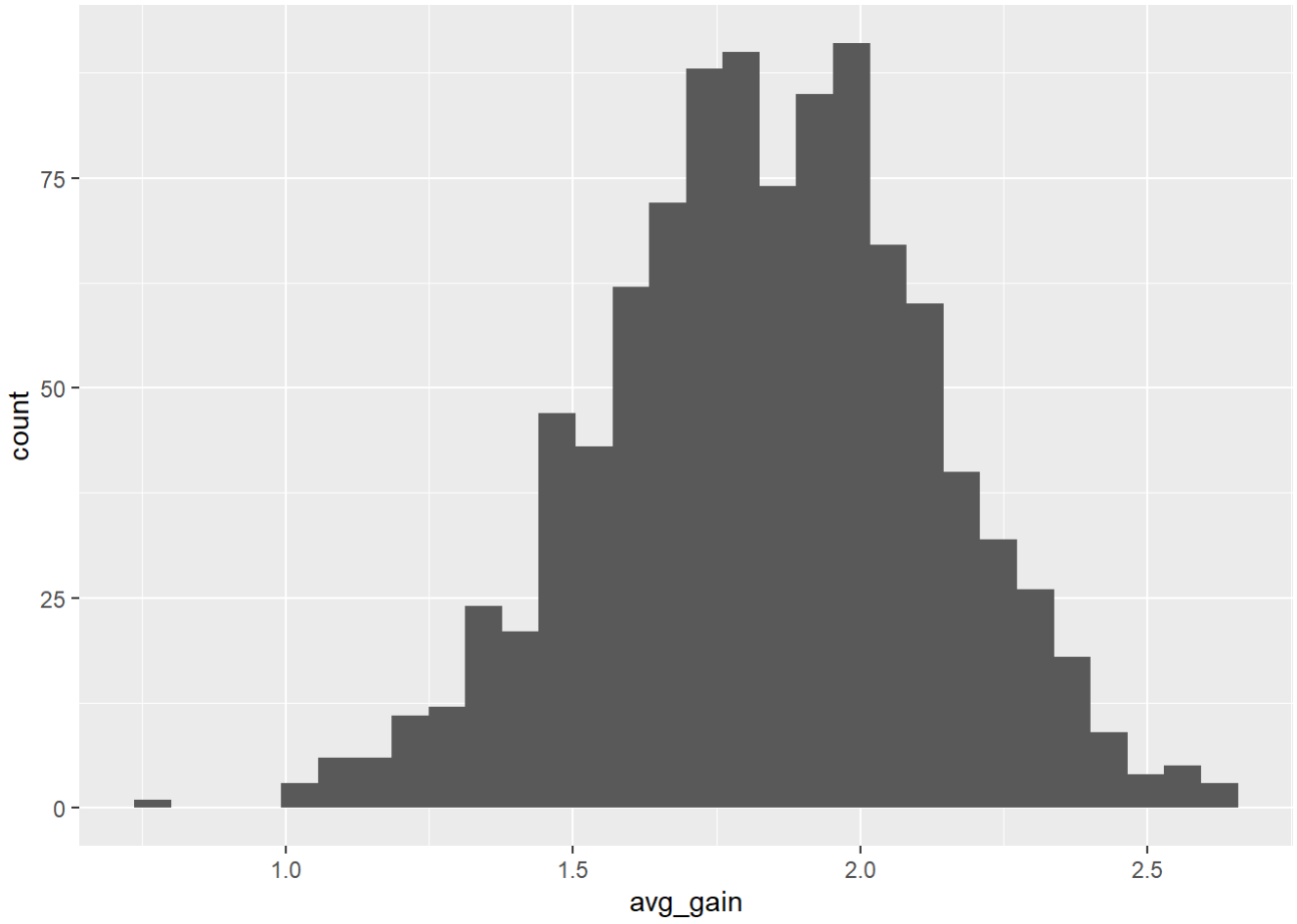
```
UA_flight.not_verylate <- UA_flights$net_gain[UA_flights$very_late == "FALSE"]
UA_flight.very_late <- UA_flights$net_gain[UA_flights$very_late == "TRUE"]


n.very_late <- length(UA_flight.very_late)
n.not_verylate <- length(UA_flight.not_verylate)

avg_gain <- numeric(1000)
for(i in 1:1000)
{
  sample.very_late <- sample(UA_flight.very_late, size = n.very_late, replace = TRUE)
  sample.not_verylate <- sample(UA_flight.not_verylate, size = n.not_verylate, replace = TRUE)
  avg_gain[i] <- mean(sample.not_verylate) - mean(sample.very_late)
}

ggplot(data=tibble(avg_gain), mapping = aes(x = avg_gain)) +
  geom_histogram()
```

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




```
quantile(avg_gain, c(.025, .975))
```

```
##     2.5%    97.5%
## 1.245190 2.390206
```

```
t.test(net_gain~late,data=UA_flights, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  net_gain by late
## t = 10.749, df = 52833, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not eq
ual to 0
## 95 percent confidence interval:
##  1.411308 2.040805
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            9.269172            7.543115
```

```
t.test(net_gain~very_late,data=UA_flights, alternative = "two.sided")
```

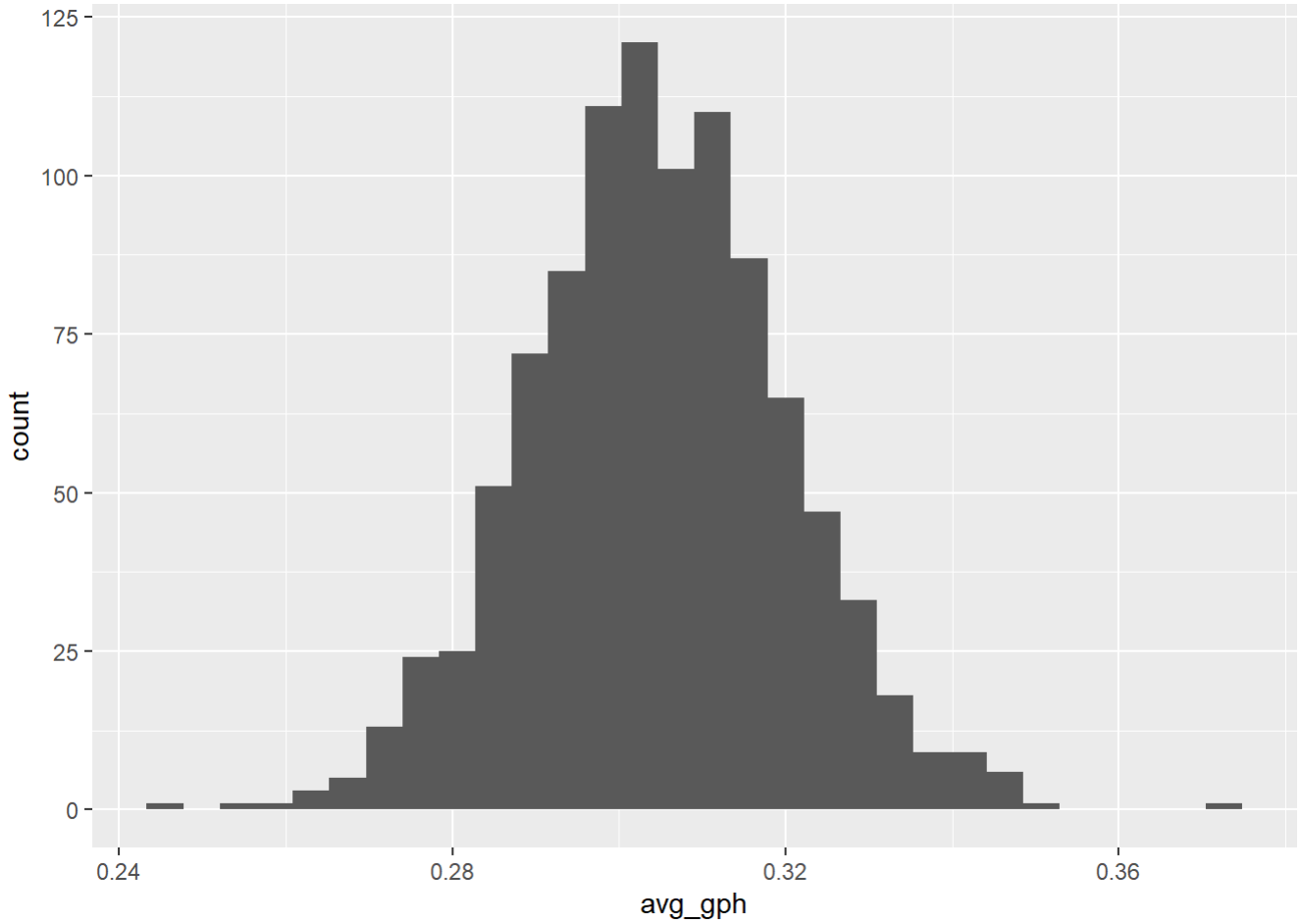
```
##
##  Welch Two Sample t-test
##
## data:  net_gain by very_late
## t = 6.2953, df = 8838.6, p-value = 3.215e-10
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not eq
ual to 0
## 95 percent confidence interval:
##  1.268195 2.415112
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            8.699534            6.857881
```

```
UA_flight.not_late <- UA_flights$gain_per_hour[UA_flights$late == "FALSE"]
UA_flight.late <- UA_flights$gain_per_hour[UA_flights$late == "TRUE"]

n.late <- length(UA_flight.late)
n.not_late <- length(UA_flight.not_late)

avg_gph <- numeric(1000)
for(i in 1:1000)
{
  sample.late <- sample(UA_flight.late, size = n.late, replace = TRUE)
  sample.not_late <- sample(UA_flight.not_late, size = n.not_late, replace = TRUE)
  avg_gph[i] <- mean(sample.not_late) - mean(sample.late)
}
ggplot(data=tibble(avg_gph), mapping = aes(x = avg_gph)) +
  geom_histogram()
```

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




```
quantile(avg_gph, c(.025, .975))
```

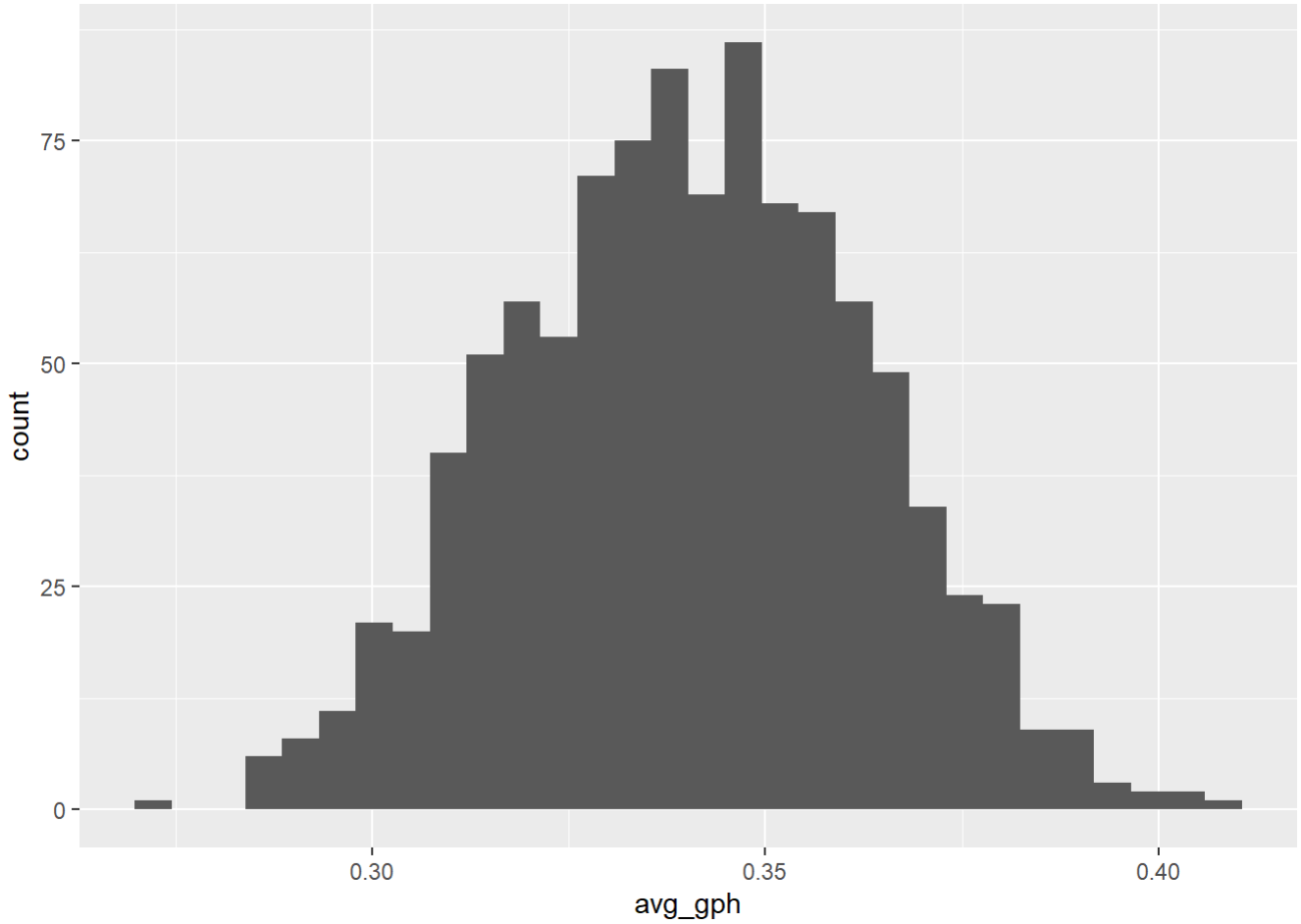
```
##      2.5%     97.5%
## 0.2740455 0.3360296
```

```
UA_flight.not_verylate <- UA_flights$gain_per_hour[UA_flights$very_late == "FALSE"]
UA_flight.very_late <- UA_flights$gain_per_hour[UA_flights$very_late == "TRUE"]

n.very_late <- length(UA_flight.very_late)
n.not_verylate <- length(UA_flight.not_verylate)

avg_gph <- numeric(1000)
for(i in 1:1000)
{
  sample.very_late <- sample(UA_flight.very_late, size = n.very_late, replace = TRUE)
  sample.not_verylate <- sample(UA_flight.not_verylate, size = n.not_verylate, replace = TRUE)
  avg_gph[i] <- mean(sample.not_verylate) - mean(sample.very_late)
}
ggplot(data=tibble(avg_gph), mapping = aes(x = avg_gph)) +
  geom_histogram()
```

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




```
quantile(avg_gph, c(.025, .975))
```

```
##      2.5%     97.5%
## 0.2975973 0.3827264
```

```
t.test(gain_per_hour~late,data=UA_flights, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  gain_per_hour by late
## t = 20.056, df = 57473, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not eq
ual to 0
## 95 percent confidence interval:
##  0.2739012 0.3332350
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           0.9310086           0.6274405
```

```
t.test(gain_per_hour~very_late,data=UA_flights, alternative = "two.sided")
```

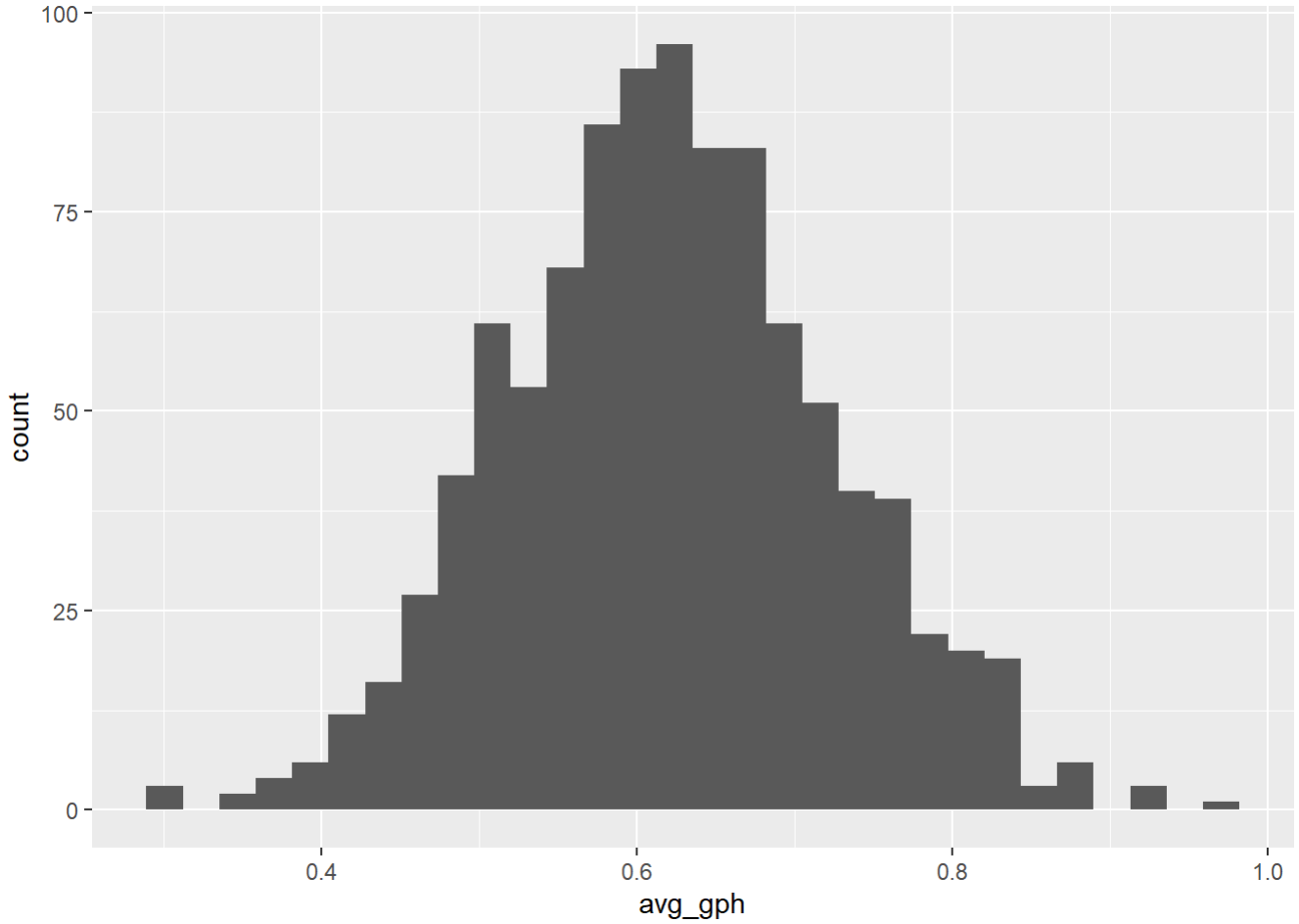
```
##
##  Welch Two Sample t-test
##
## data:  gain_per_hour by very_late
## t = 14.971, df = 9882.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not eq
ual to 0
## 95 percent confidence interval:
##  0.2960713 0.3852843
## sample estimates:
## mean in group FALSE  mean in group TRUE
##           0.8330167           0.4923389
```

```
UA_flight.shorter <- UA_flights$gain_per_hour[UA_flights$flight_duration == "Short"]
UA_flight.longer <- UA_flights$gain_per_hour[UA_flights$flight_duration == "Longer"]

n.shorter <- length(UA_flight.shorter)
n.longer <- length(UA_flight.longer)

avg_gph <- numeric(1000)
for(i in 1:1000)
{
  sample.shorter <- sample(UA_flight.shorter, size = n.shorter, replace = TRUE)
  sample.longer <- sample(UA_flight.longer, size = n.longer, replace = TRUE)
  avg_gph[i] <- mean(sample.shorter) - mean(sample.longer)
}
ggplot(data=tibble(avg_gph), mapping = aes(x = avg_gph)) +
  geom_histogram()
```

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




```
quantile(avg_gph, c(.025, .975))
```

```
##      2.5%     97.5%
## 0.4254994 0.8279272
```

```
t.test(gain_per_hour~flight_duration,data=UA_flights, alternative = "two.sided")
```

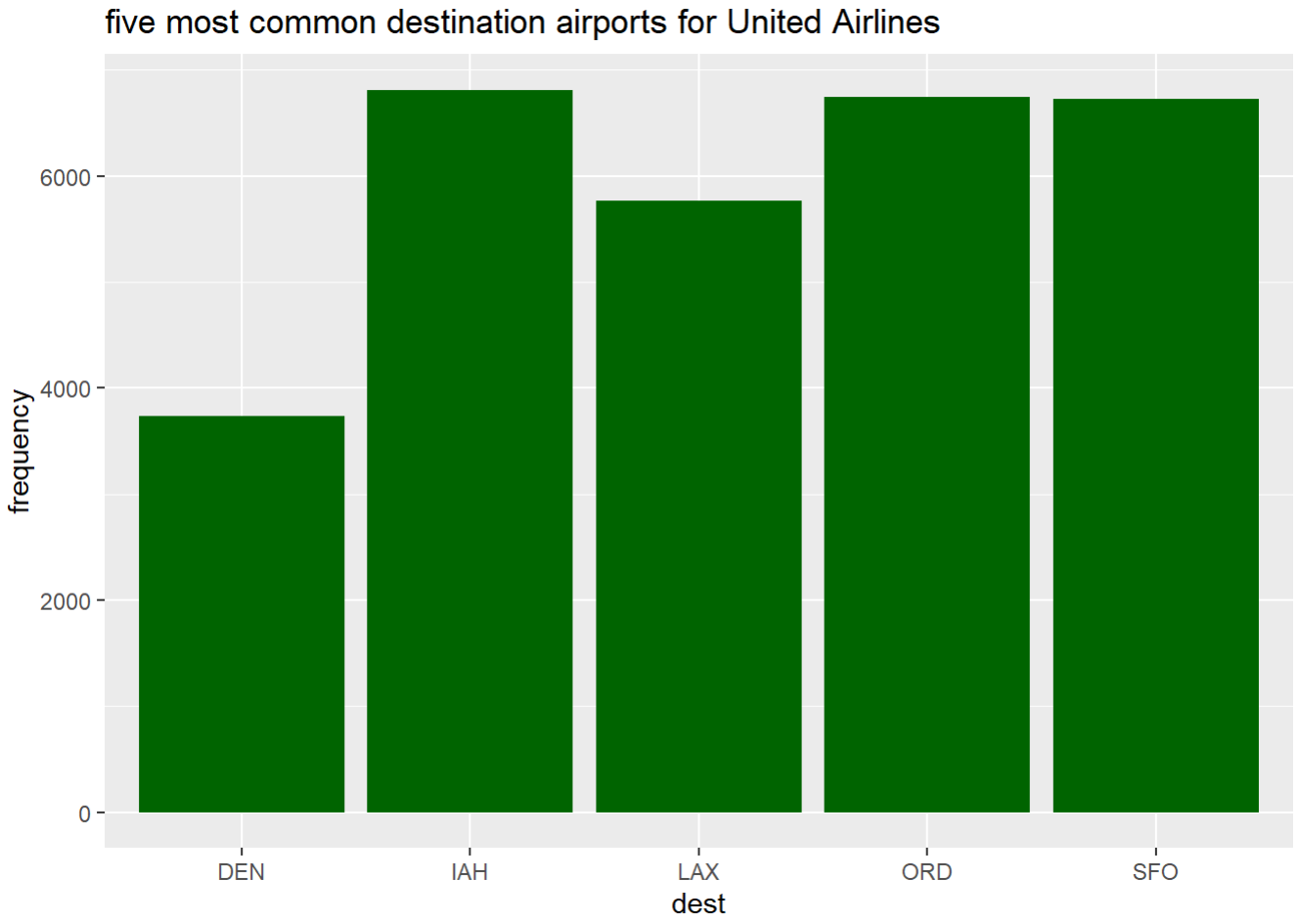
```
##
##  Welch Two Sample t-test
##
## data:  gain_per_hour by flight_duration
## t = -5.887, df = 840.61, p-value = 5.681e-09
## alternative hypothesis: true difference in means between group Longer and group Short is not
equal to 0
## 95 percent confidence interval:
##  -0.8276301 -0.4137410
## sample estimates:
## mean in group Longer  mean in group Short
##            0.7795546            1.4002401
```

```
top_airports <- UA_flights %>%
  group_by(dest) %>%
  summarise(frequency = n()) %>%
  arrange(desc(frequency)) %>%
  head(5)

print(top_airports)
```

```
## # A tibble: 5 × 2
##   dest  frequency
##   <chr>     <int>
## 1 IAH        6814
## 2 ORD        6744
## 3 SFO        6728
## 4 LAX        5770
## 5 DEN        3737
```

```
ggplot(data = top_airports, aes(x = dest, y = frequency)) +
  geom_bar(stat = "identity", fill="darkgreen") +
  ggtitle("five most common destination airports for United Airlines")
```



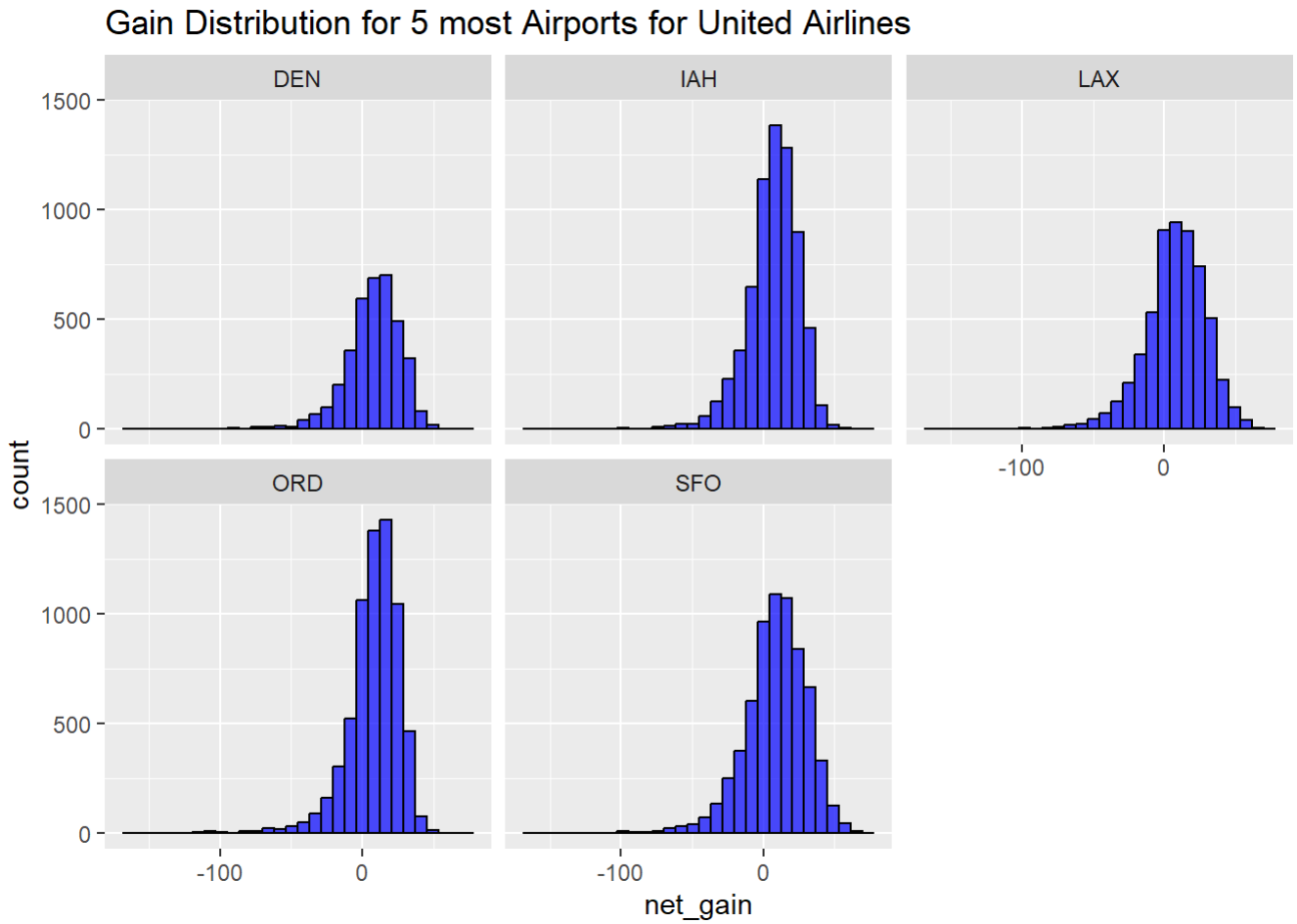
```r
for (airport in top_airports$dest) {
  airport_data <- UA_flights %>%
    filter(dest == airport)
  avg_gain <- mean(airport_data$net_gain)
  conf_int <- t.test(airport_data$net_gain)$conf.int
  print(paste("Average Gain for", airport, ":", avg_gain))
  print(paste("Confidence interval for ", airport, ":", conf_int[1],conf_int[2]))
}
```

```
## [1] "Average Gain for IAH : 6.86175520986205"
## [1] "Confidence interval for  IAH : 6.42381974530183 7.29969067442227"
## [1] "Average Gain for ORD : 7.77743179122183"
## [1] "Confidence interval for  ORD : 7.32013459022188 8.23472899222177"
## [1] "Average Gain for SFO : 8.69500594530321"
## [1] "Confidence interval for  SFO : 8.15947541162006 9.23053647898636"
## [1] "Average Gain for LAX : 7.82530329289428"
## [1] "Confidence interval for  LAX : 7.25968142356738 8.39092516222118"
## [1] "Average Gain for DEN : 7.3023815895103"
## [1] "Confidence interval for  DEN : 6.65934839135574 7.94541478766487"
```

```r
airport_data <- UA_flights %>%
  filter(dest %in% top_airports$dest)

# Create a boxplot with facets for each airport
ggplot(data = airport_data, aes(x = net_gain)) +
  geom_histogram(position = "identity", alpha = 0.7, fill = "blue",color="black") +
  ggtitle("Gain Distribution for 5 most Airports for United Airlines") +
  facet_wrap(~ dest, ncol = 3)
```

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


Gain Distribution for 5 most Airports for United Airlines
DEN
IAH
LAX
ORD
SFO
count
1500
1000
500
0
-100
0
net_gain