Assignment 3

Anthony Cunningham

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

# Change working dir in RMarkdown cell
knitr::opts_knit$set(root.dir =
'C:/Users/AC069015/kumc_applied_stats/data_824_data_viz_and_acquisition/3_univariate_eda'
)

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(tidyr)
library(ggplot2)

df <- mpg
```

Exercise 1

```
help("mpg")
```

```
str(df)
```

```
## tibble [234 × 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

```
nrow(df)
```

```
## [1] 234
```

```
ncol(df)
```

```
## [1] 11
```

```
char_cols <- colnames(df[sapply(df, is.character)])
print("Character variables in mpg dataset: ")
```

```
## [1] "Character variables in mpg dataset: "
```

```
char_cols
```

```
## [1] "manufacturer" "model"        "trans"        "drv"          "fl"
## [6] "class"
```

```
df[sapply(df, is.character)] <- lapply(df[sapply(df, is.character)], as.factor)
```

```
summary(df)
```

```
##      manufacturer                model         displ            year
##  dodge     :37    caravan 2wd        : 11   Min.   :1.600   Min.   :1999
##  toyota    :34    ram 1500 pickup 4wd: 10   1st Qu.:2.400   1st Qu.:1999
##  volkswagen:27    civic              :  9   Median :3.300   Median :2004
##  ford      :25    dakota pickup 4wd  :  9   Mean   :3.472   Mean   :2004
##  chevrolet :19    jetta              :  9   3rd Qu.:4.600   3rd Qu.:2008
##  audi      :18    mustang            :  9   Max.   :7.000   Max.   :2008
##  (Other)   :74    (Other)            :177
##       cyl               trans    drv          cty             hwy
##  Min.   :4.000   auto(l4)  :83   4:103   Min.   : 9.00   Min.   :12.00
##  1st Qu.:4.000   manual(m5):58   f:106   1st Qu.:14.00   1st Qu.:18.00
##  Median :6.000   auto(l5)  :39   r: 25   Median :17.00   Median :24.00
##  Mean   :5.889   manual(m6):19           Mean   :16.86   Mean   :23.44
##  3rd Qu.:8.000   auto(s6)  :16           3rd Qu.:19.00   3rd Qu.:27.00
##  Max.   :8.000   auto(l6)  : 6           Max.   :35.00   Max.   :44.00
##                  (Other)   :13
##  fl             class
##  c:  1   2seater   : 5
##  d:  5   compact   :47
##  e:  8   midsize   :41
##  p: 52   minivan   :11
##  r:168   pickup    :33
##          subcompact:35
##          suv       :62
```

```
df %>% select(cyl) %>% group_by(cyl) %>% summarise(n_cars = n())
```

```
## # A tibble: 4 × 2
##     cyl n_cars
##   <int>  <int>
## 1     4     81
## 2     5      4
## 3     6     79
## 4     8     70
```

`cyl` only takes on 4 distinct values, so it would make more sense as a factor variable.

```
df <- df %>% mutate(cyl = as.factor(cyl))
summary(select(df, cyl))
```

```
##  cyl
##  4:81
##  5: 4
##  6:79
##  8:70
```

Exercise 2

```
df %>% mutate(cty_null = is.na(cty)) %>% group_by(cty_null) %>% summarise(n = n())
```

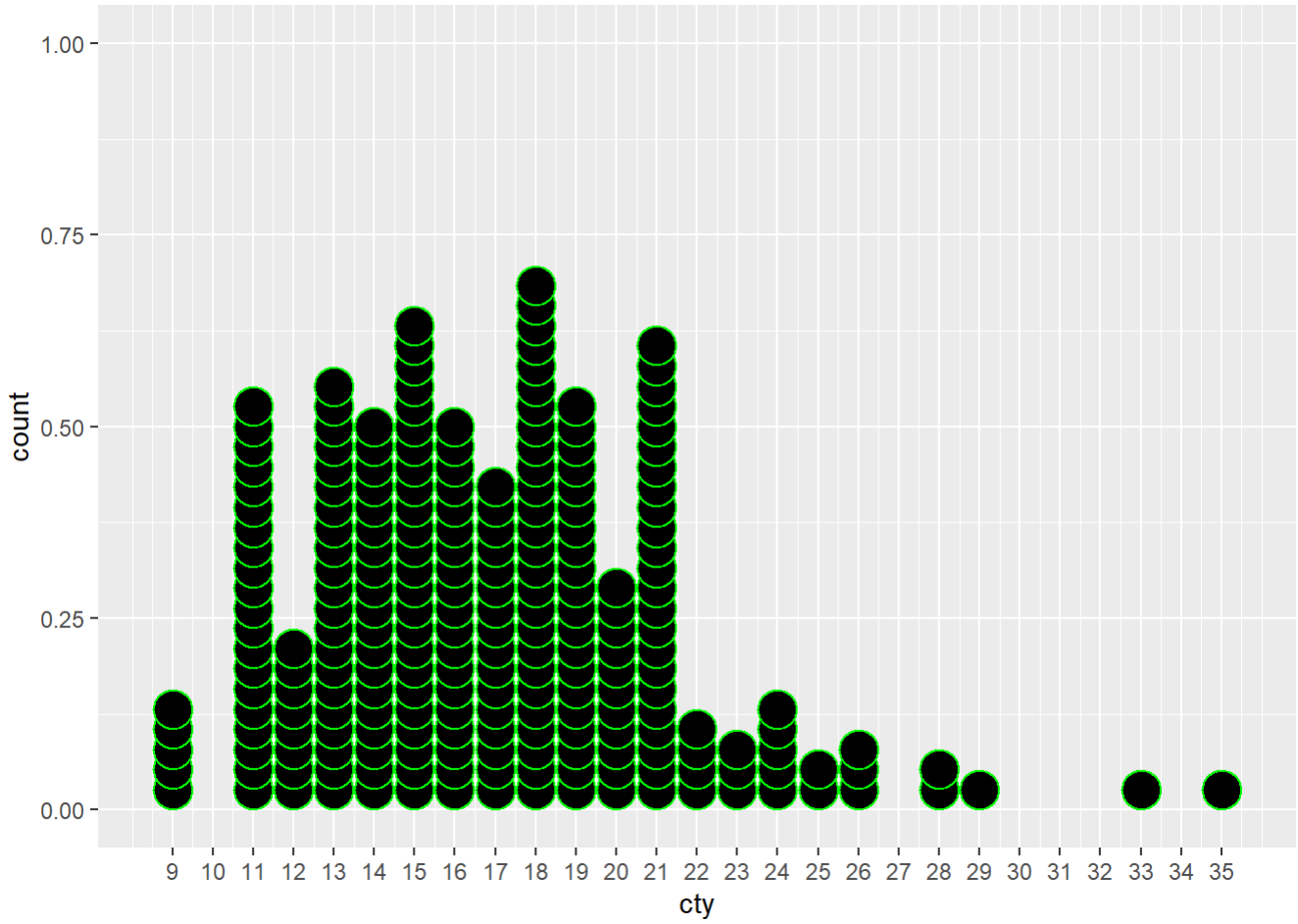
```
## # A tibble: 1 × 2
##   cty_null     n
##   <lgl>    <int>
## 1 FALSE      234
```

No, there are not any missing values in the `cty` variable.

```
df %>% group_by(cty) %>% summarise(n = n()) %>% arrange(desc(n)) %>% head(5)
```

```
## # A tibble: 5 × 2
##     cty     n
##   <int> <int>
## 1    18    26
## 2    15    24
## 3    21    23
## 4    13    21
## 5    11    20
```

```
min_cty <- min(df$cty)
max_cty <- max(df$cty)

df %>%
  ggplot(aes(x = cty)) +
  geom_dotplot(binwidth = 1,
               stackdir = "up",
               color = "green",
               stackratio = 0.5,
               dotsize = 1) +
  scale_x_continuous(breaks = seq(min_cty, max_cty, 1))
```




```
ggsave(
    filename = "images/03_assignment_fig1.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

The peak of `cty` in terms of frequency is at 18 miles per gallon, followed by 15, 21, 13 and 11 mpg.

```
(freq_cty <- 18)
```

```
## [1] 18
```

```
(mean_cty <- mean(df$cty))
```

```
## [1] 16.85897
```

```
(med_cty <- median(df$cty))
```

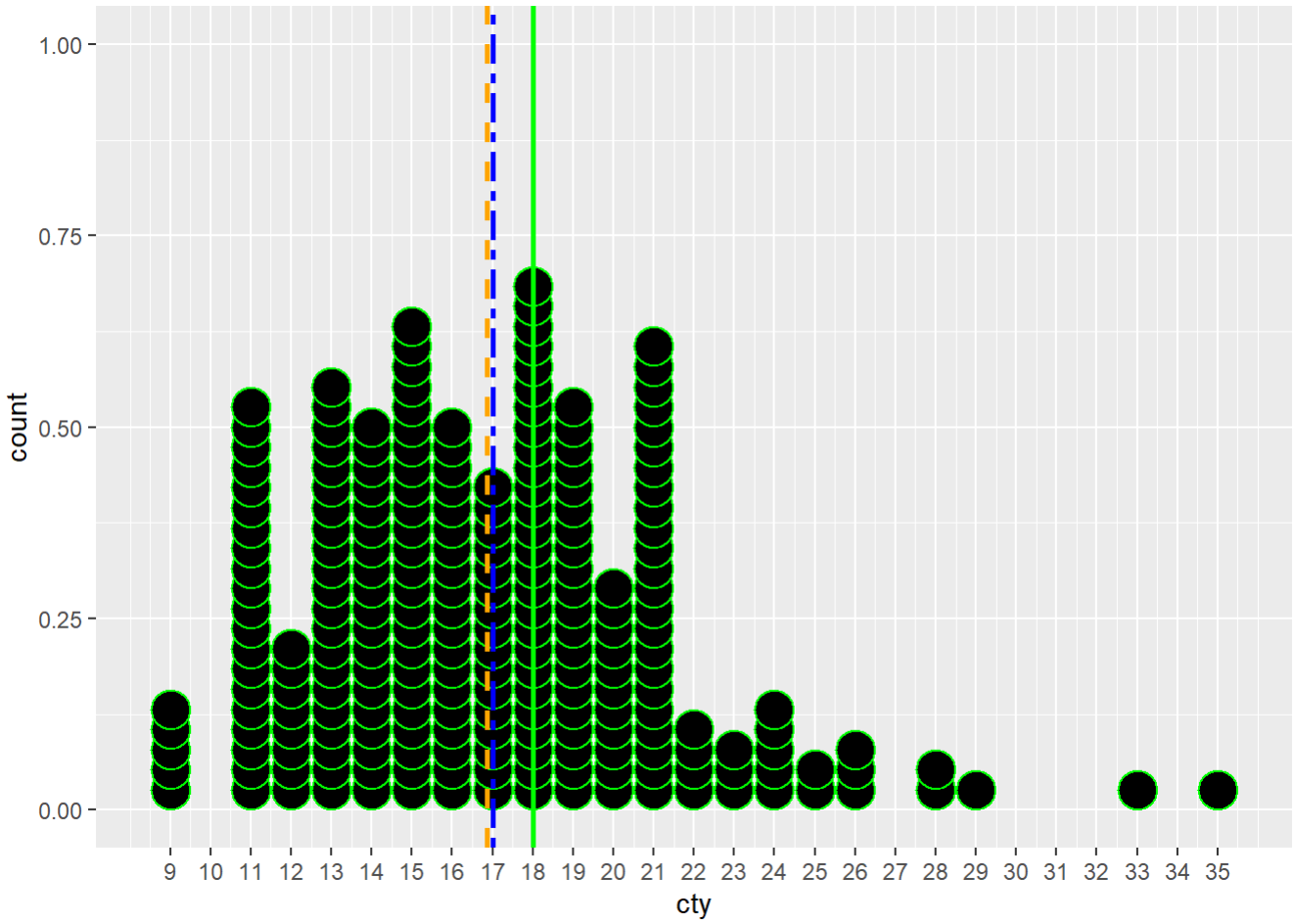
```
## [1] 17
```

```
ifelse(
    (freq_cty - mean_cty) > (freq_cty - med_cty),
    print("Median value is closer to most frequent value"),
    print("Mean value is closer to most frequent value")
    )
```

```
## [1] "Median value is closer to most frequent value"
```

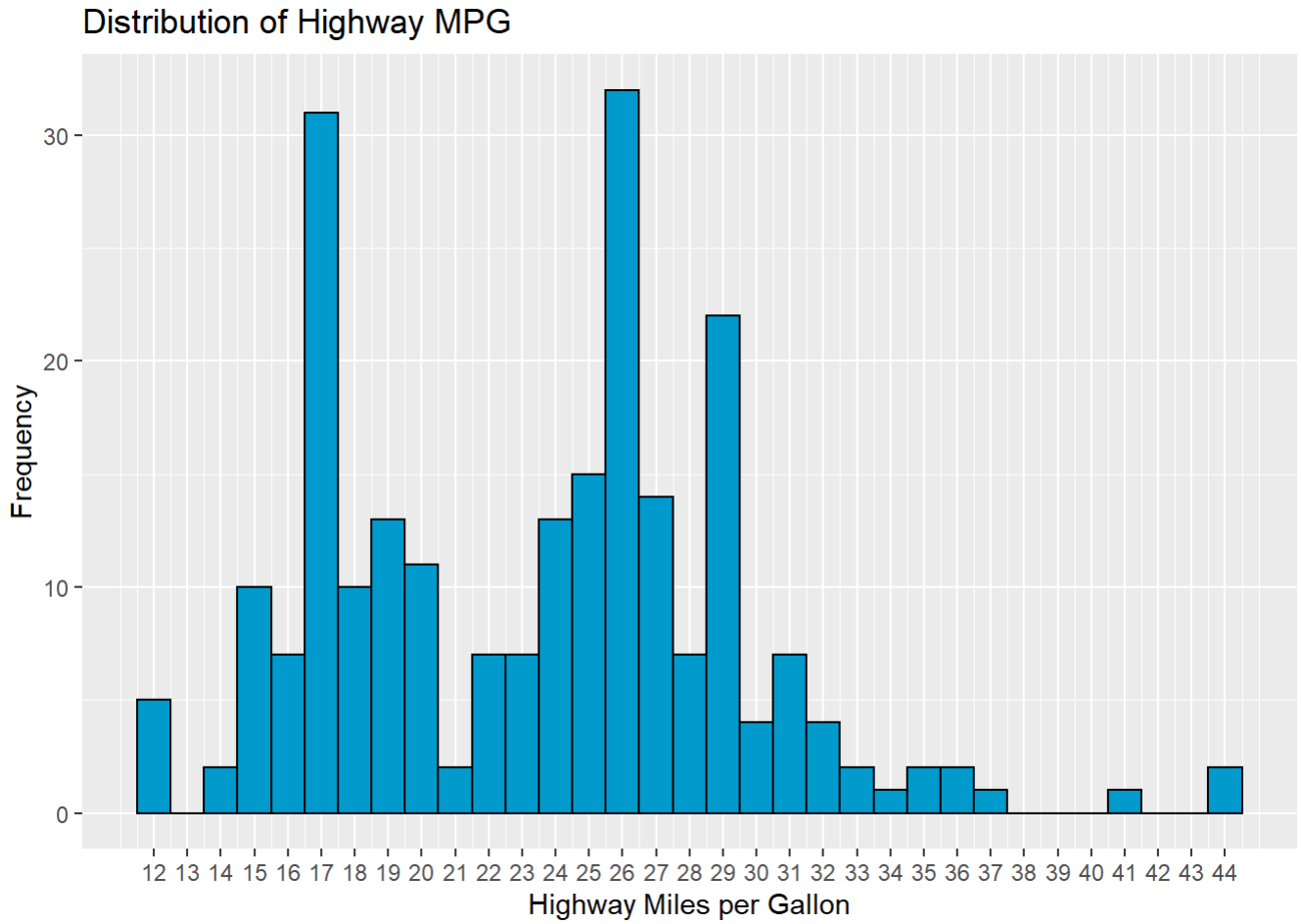
```
## [1] "Median value is closer to most frequent value"
```

```
df %>%
    ggplot(aes(x = cty)) +
    geom_dotplot(binwidth = 1,
                 stackdir = "up",
                 color = "green",
                 stackratio = 0.5,
                 dotsize = 1) +
    scale_x_continuous(breaks = seq(min_cty, max_cty, 1)) +
    geom_vline(xintercept = freq_cty, lwd = 1, linetype = "solid", color = "green") +
    geom_vline(xintercept = mean_cty, lwd = 1, linetype = "dashed", color = "orange") +
    geom_vline(xintercept = med_cty, lwd = 1, linetype = "twodash", color = "blue")
```



```
ggsave(
    filename = "images/03_assignment_fig2.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

```
min_hwy <- min(df$hwy)
max_hwy <- max(df$hwy)

df %>%
  ggplot(aes(x = hwy)) +
  geom_histogram(binwidth = 1, color = "black", fill = "deepskyblue3") +
  scale_x_continuous(breaks = seq(min_hwy, max_hwy, 1)) +
  xlab("Highway Miles per Gallon") +
  ylab("Frequency") +
  ggtitle("Distribution of Highway MPG")
```



```
ggsave(
    filename = "images/03_assignment_fig3.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

Yes, there is more than one peak value for vehicles' highway miles per gallon. Peaks are at 17 and 26 highway MPG, respectively, both with over 30 vehicles at that value.

There could be multiple peaks because those could be average values for specific types of vehicles (e.g. trucks/suvs vs compact/midsize vehicles).

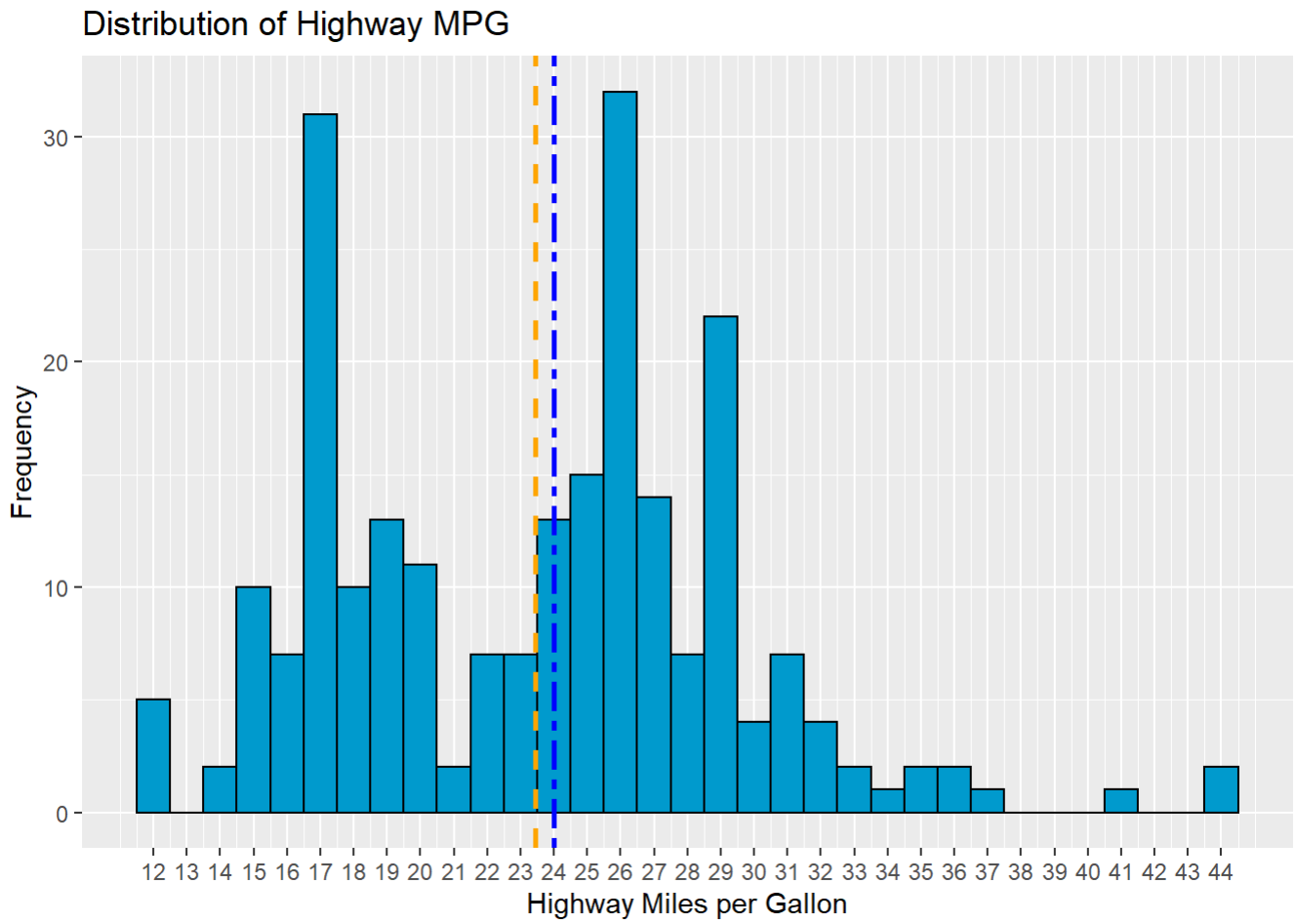
```
mean_hwy <- mean(df$hwy)
med_hwy <- median(df$hwy)

paste0("Mean Hwy MPG: ", mean_hwy)
```

```
## [1] "Mean Hwy MPG: 23.4401709401709"
```

```
paste0("Median Hwy MPG: ", med_hwy)
```

```
## [1] "Median Hwy MPG: 24"
```

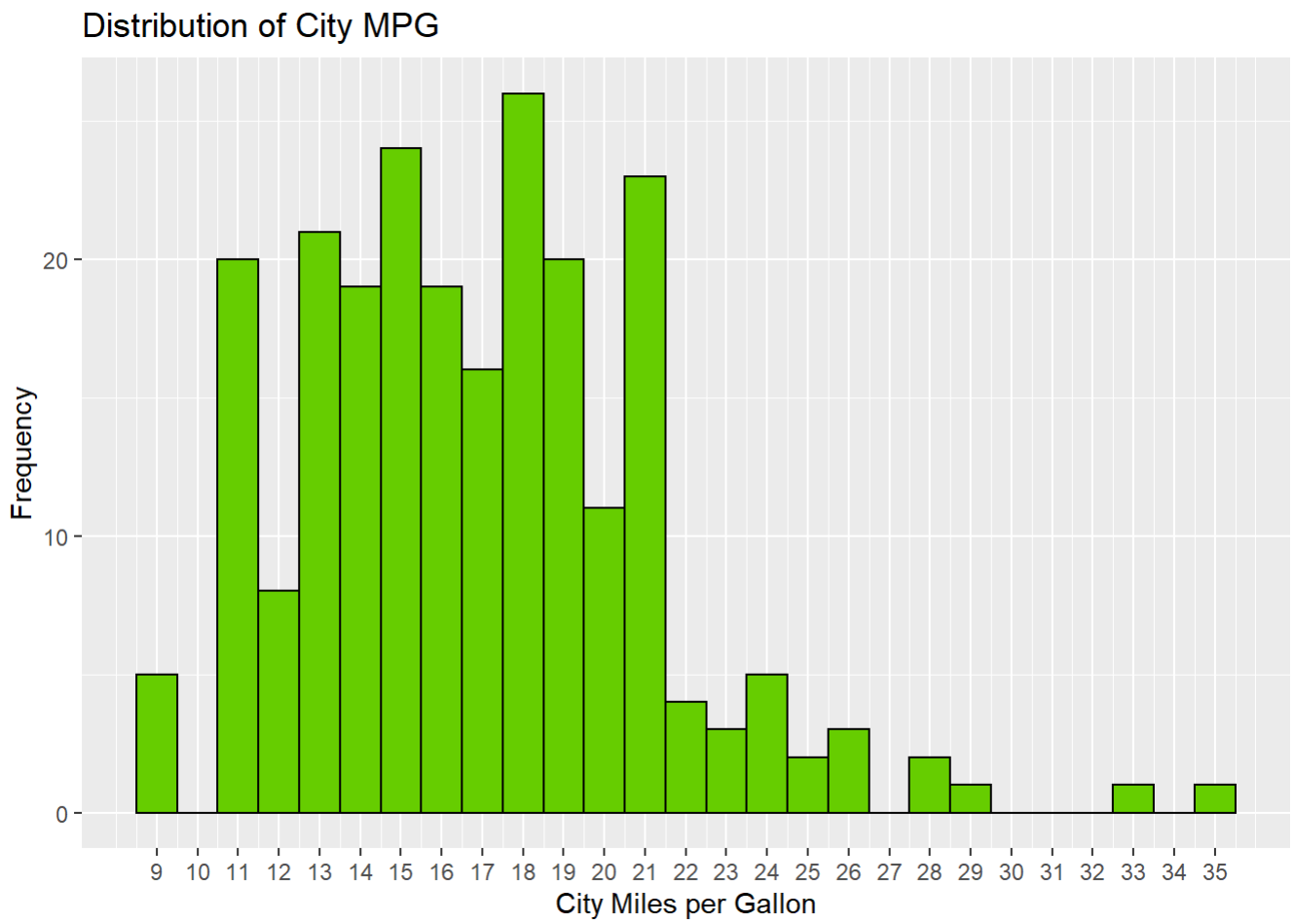
```
df %>%
  ggplot(aes(x = hwy)) +
  geom_histogram(binwidth = 1, color = "black", fill = "deepskyblue3") +
  scale_x_continuous(breaks = seq(min_hwy, max_hwy, 1)) +
  xlab("Highway Miles per Gallon") +
  ylab("Frequency") +
  ggtitle("Distribution of Highway MPG") +
  geom_vline(xintercept = mean_hwy, lwd = 1, linetype = "dashed", color = "orange") +
  geom_vline(xintercept = med_hwy, lwd = 1, linetype = "twodash", color = "blue")
```




```
ggsave(
    filename = "images/03_assignment_fig4.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

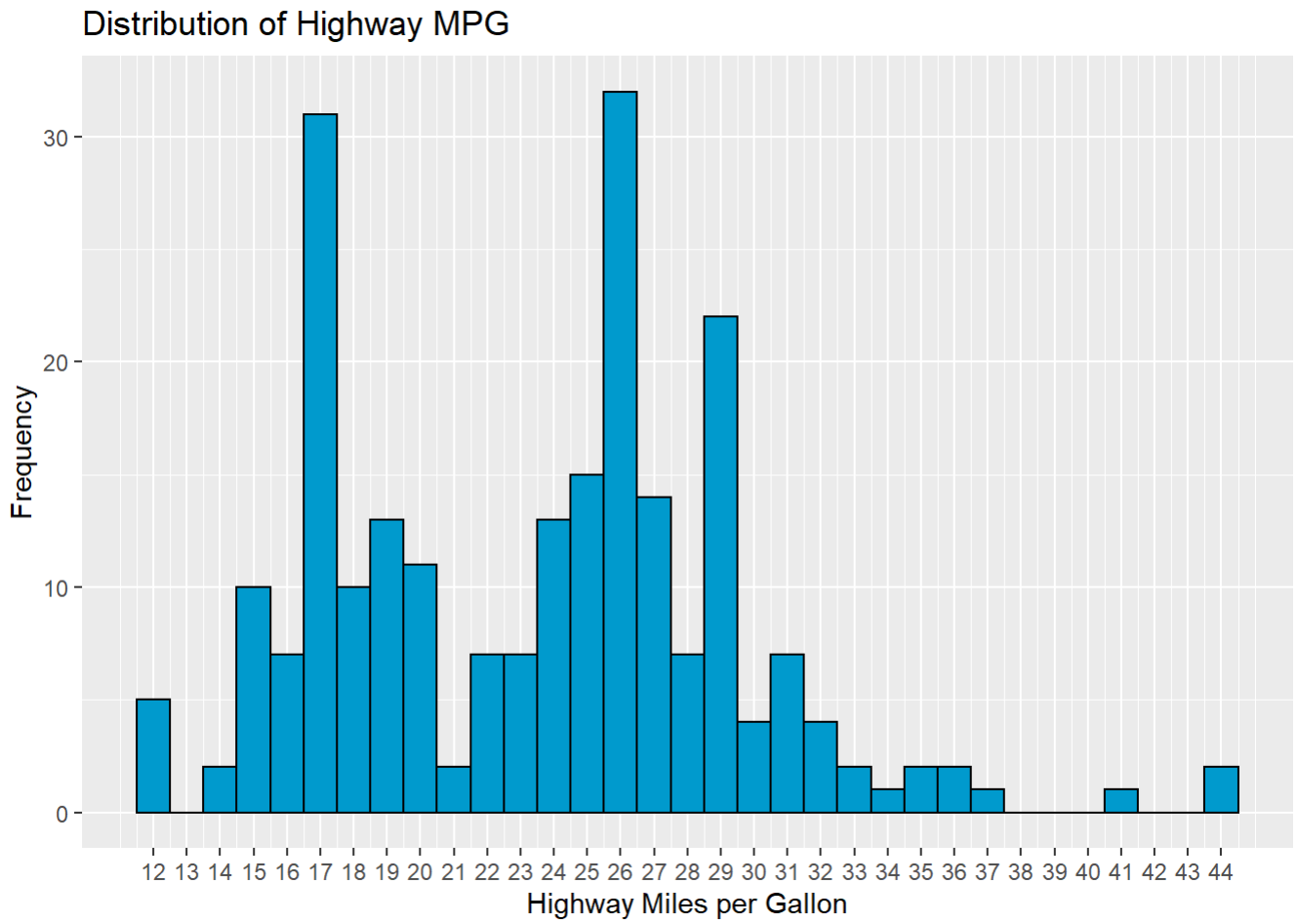
Median highway MPG is slightly larger than mean highway MPG, indicating possible large outliers or a slight right tail. However, since the distribution looks to be bimodal, the overall mean and median are not very informative, as there are likely other attributes of vehicles that segments that would be useful to compare.

```
min_cty <- min(df$cty)
max_cty <- max(df$cty)

df %>%
  ggplot(aes(x = cty)) +
  geom_histogram(binwidth = 1, color = "black", fill = "chartreuse3") +
  scale_x_continuous(breaks = seq(min_cty, max_cty, 1)) +
  xlab("City Miles per Gallon") +
  ylab("Frequency") +
  ggtitle("Distribution of City MPG")
```



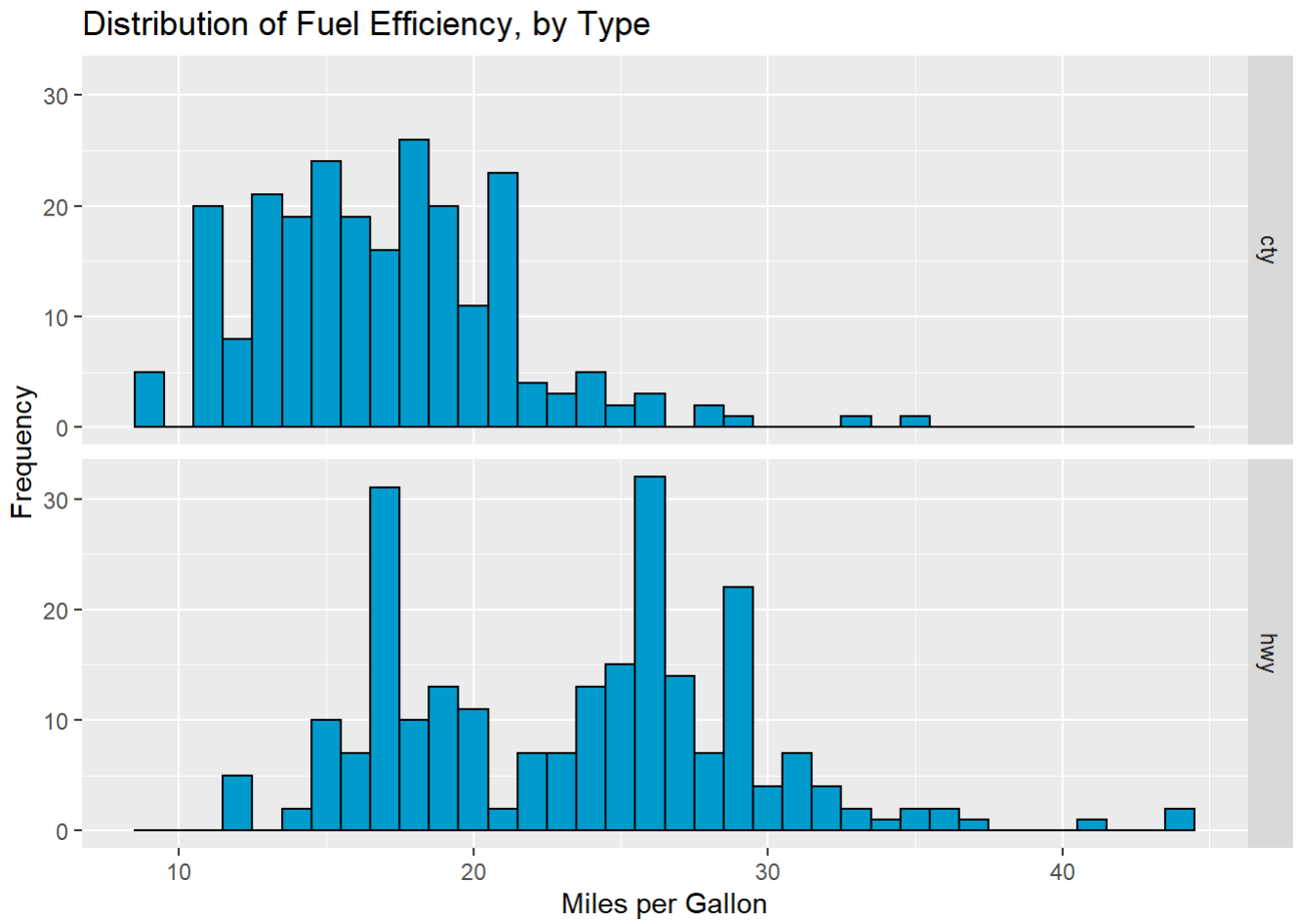

```
ggsave(
    filename = "images/03_assignment_fig5.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

```
min_hwy <- min(df$hwy)
max_hwy <- max(df$hwy)

df %>%
  ggplot(aes(x = hwy)) +
  geom_histogram(binwidth = 1, color = "black", fill = "deepskyblue3") +
  scale_x_continuous(breaks = seq(min_hwy, max_hwy, 1)) +
  xlab("Highway Miles per Gallon") +
  ylab("Frequency") +
  ggtitle("Distribution of Highway MPG")
```



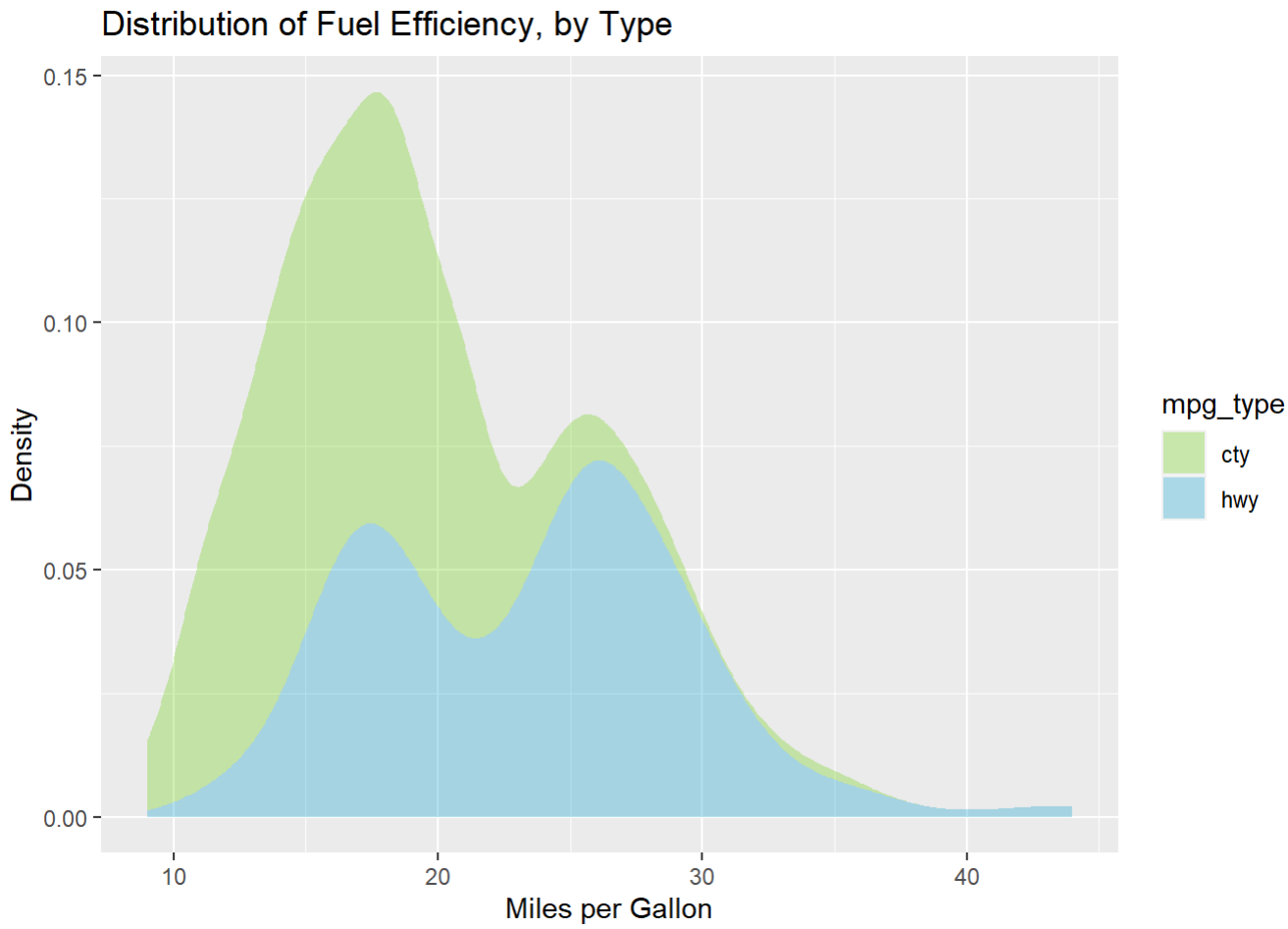

```
ggsave(
    filename = "images/03_assignment_fig6.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

```
df %>%
    select(cty, hwy) %>%
    pivot_longer(cols = c(cty, hwy), names_to = "mpg_type", values_to = "mpg") %>%
    ggplot(aes(x = mpg)) +
    geom_histogram(binwidth = 1, color = "black", fill = "deepskyblue3") +
    xlab("Miles per Gallon") +
    ylab("Frequency") +
    ggtitle("Distribution of Fuel Efficiency, by Type") +
    facet_grid(mpg_type ~ .)
```




```
ggsave(
    filename = "images/03_assignment_fig7.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```

```
df %>%
    select(cty, hwy) %>%
    pivot_longer(cols = c(cty, hwy), names_to = "mpg_type", values_to = "mpg") %>%
    ggplot(aes(x = mpg, fill = mpg_type)) +
    geom_area(
      stat = "density",
      kernel = "gaussian",
      alpha = 0.3
      ) +
    scale_fill_manual(values = c("chartreuse3", "deepskyblue3")) +
    xlab("Miles per Gallon") +
    ylab("Density") +
    ggtitle("Distribution of Fuel Efficiency, by Type")
```




```
ggsave(
    filename = "images/03_assignment_fig8.png",
    units = "cm",
    width = 29.7,
    height = 21,
    dpi = 600
    )
```