Problem Set Two: All Things Data (Easier) – Answer Key

Andie Creel

Data Manipulation Problems

Problem 1: Install and Load Necessary Packages

Install and load `dplyr`, `tidyr`, and `ggplot2` if you haven't already.

```
# install.packages("dplyr")
# install.packages("tidyr")
# install.packages("ggplot2")
library(dplyr)
library(tidyr)
library(ggplot2)
```

Problem 2: Data Filtering

Load the `mpg` dataset from `ggplot2` by running `data(mpg)`. Make a new dataset called `toyota_cars` that's filter for cars manufactured by "toyota".

```
# load data
data(mpg)

# filter dataset
toyota_cars <- mpg %>%
  filter(manufacturer == "toyota")
```

Problem 3: Data Transformation

Use `mutate()` to create a new column `hwy_km` in `toyota_cars` converting highway miles per gallon (`hwy`) to kilometers per liter (1 mile = 1.60934 km, 1 gallon = 3.78541 liters). This is a basic unit transformation.

```
# make new variable
toyota_cars <- toyota_cars %>%
  mutate(hwy_km = hwy / 3.78541 * 1.60934)
```

Problem 4: Summarizing Data

Group the `toyota_cars` dataset by the variable `class` and summarize the average highway kilometers per liter for each class.

```r
# summary by vehicle type
average_hwy_km <- toyota_cars %>%
  group_by(class) %>%
  summarize(avg_hwy_km = mean(hwy_km))

# print
average_hwy_km
```

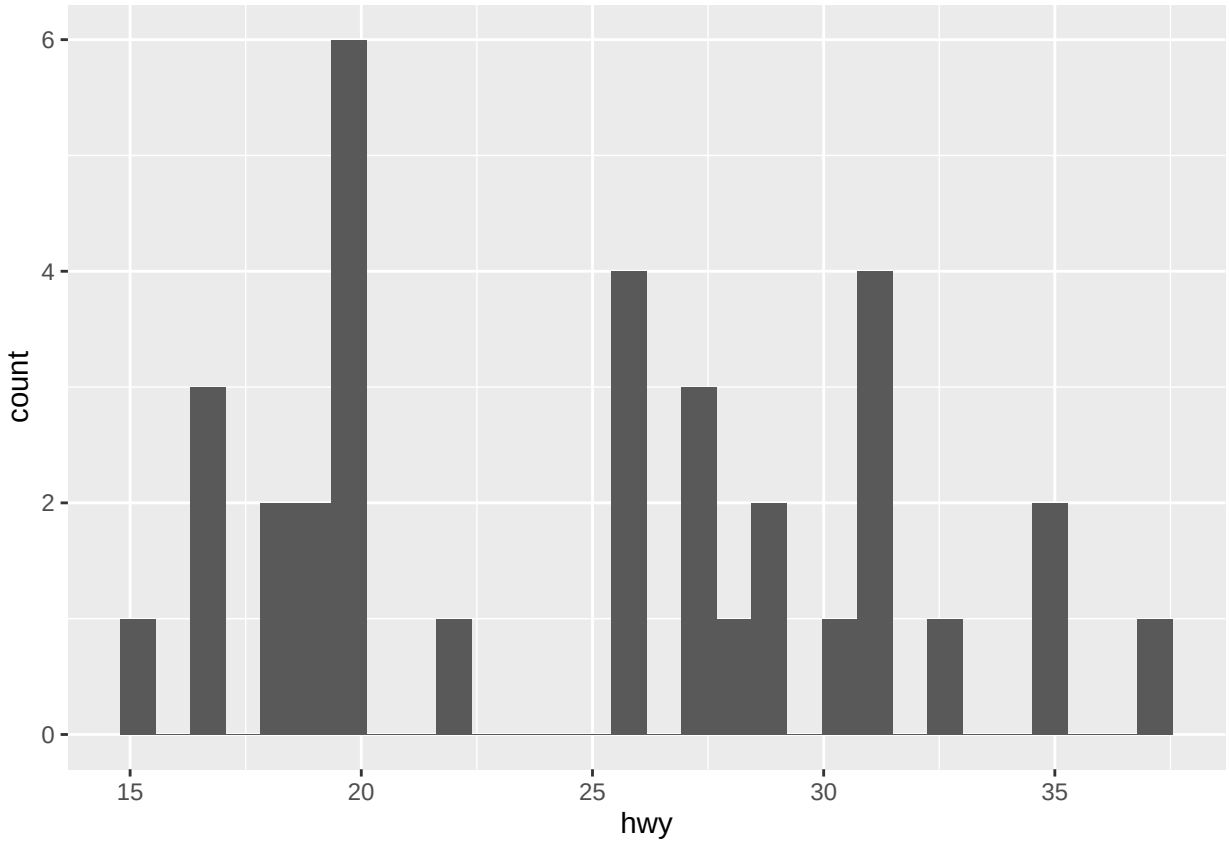
```
## # A tibble: 4 x 2
##   class   avg_hwy_km
##   <chr>        <dbl>
## 1 compact      13.0
## 2 midsize      12.0
## 3 pickup        8.26
## 4 suv           7.76
```

Data Visualization Problems

Problem 5: Basic Histogram

Create a histogram of the `hwy` variable for the `toyota_cars`.

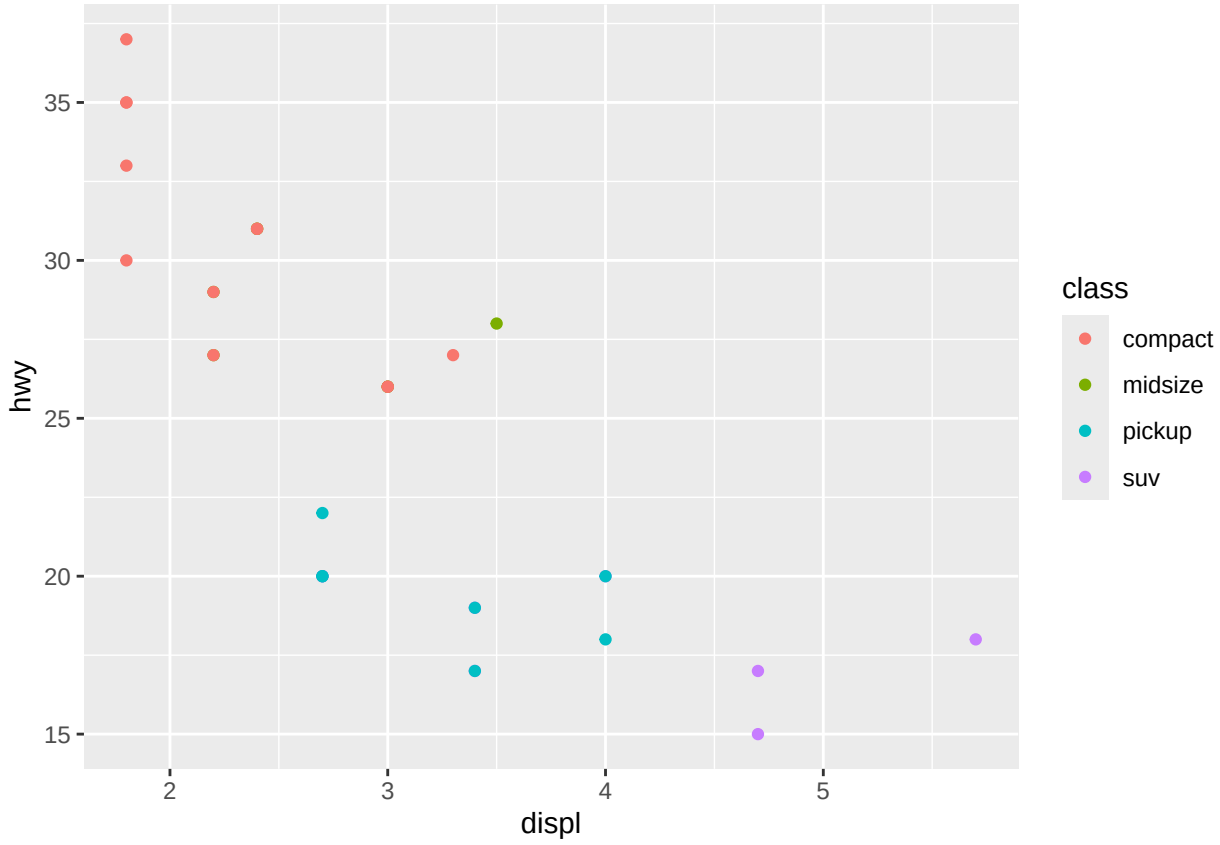
```r
# basic histogram
ggplot(toyota_cars, aes(x = hwy)) +
  geom_histogram()
```



Problem 6: Scatter Plot

Create a scatter plot with `displ` on the x-axis and `hwy` on the y-axis. Color the points by `class`.

```
# basic scatter plot
ggplot(toyota_cars, aes(x = displ, y = hwy, color = class)) +
  geom_point()
```

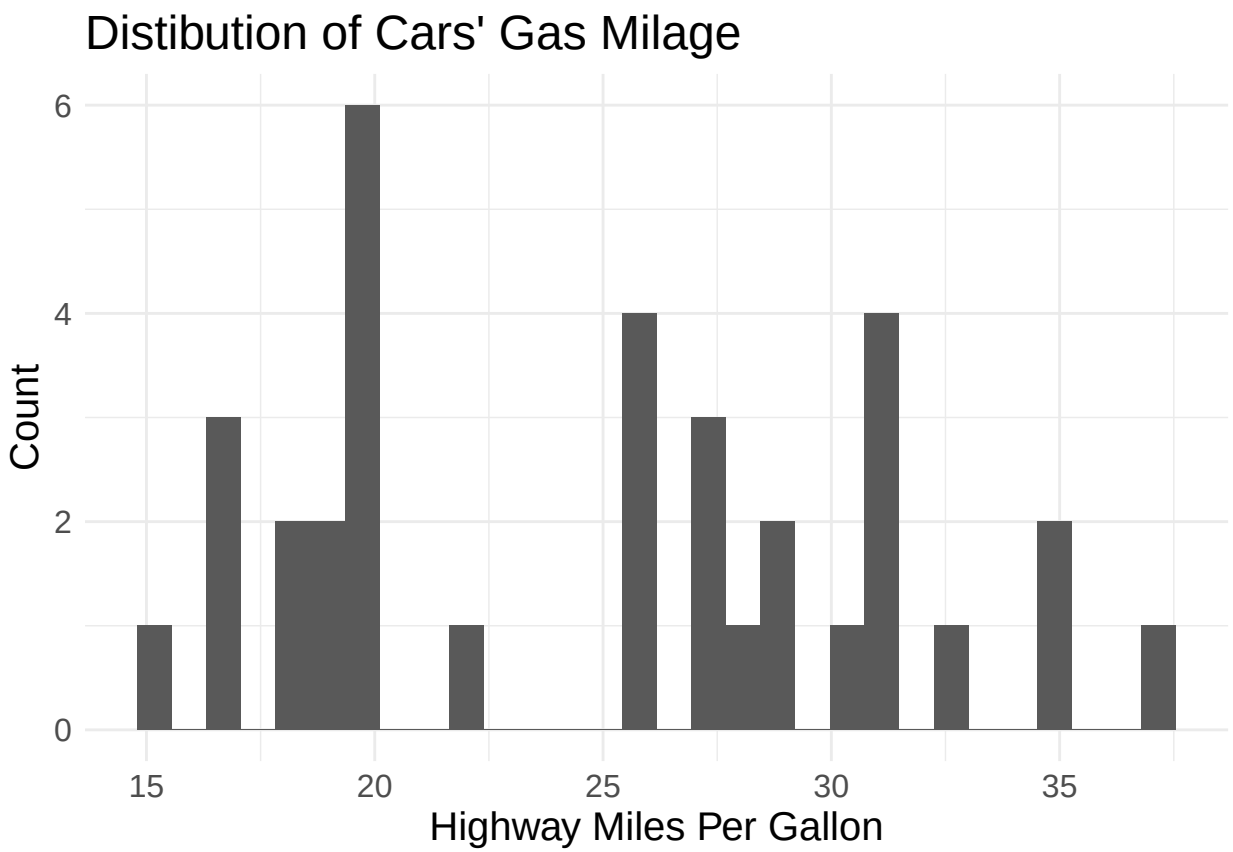


Problem 7: Best Data Visualization Practice

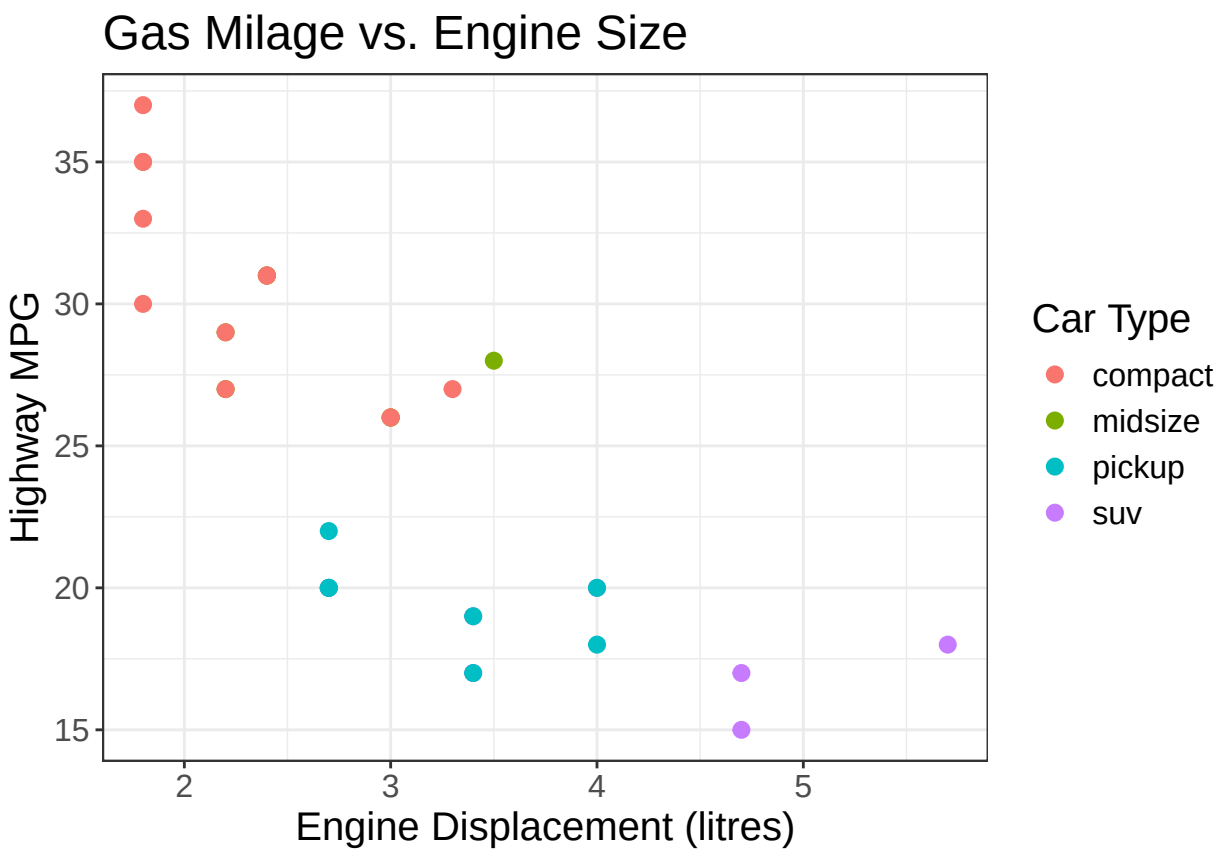
Choose one of the plots above to apply best data visualization practices to. Specifically:

- Write clear labels and titles
- Make it as simple as possible while not becoming reductive
- Make sure all parts of graph are legible
- Consider the colors, if using them

```
# Problem 5:
ggplot(toyota_cars, aes(x = hwy)) +
  geom_histogram() +
  labs(title = "Distibution of Cars' Gas Milage",
       x = "Highway Miles Per Gallon",
       y = "Count") +
  theme_minimal() +
  theme(text = element_text(size=15))
```




```
# Problem 6:
ggplot(toyota_cars, aes(x = displ, y = hwy, color = class)) +
  geom_point(size = 2.5) +
  labs(title = "Gas Milage vs. Engine Size",
     x = "Engine Displacement (litres)",
     y = "Highway MPG",
     color = "Car Type") +
  theme_bw() +
  theme(text = element_text(size=15))
```



Problem 8: Save Your Plot

Save one of the plots you created to your project directory as a PNG file.

```
# save last figure
ggsave("2_my_scatter_plot.png", plot = last_plot(), width = 10, height = 8)
```

Problem 9: Best File Practice

Consider a research project (one of your own or one you made up). Describe the file structure you would use for your project, including what your raw data may look like and what your cleaned data may look like. List the name of the scripts you'd write, and what each would do.

> Files would include `data/raw_data/`, `data/clean_data/`, `scripts/`, `results/figures/`, `results/tables/`, `manuscript/`, `presentation/`. The scripts would be `0_data_clean.R`, `1_explore.R`, `2_descriptive_figures.R`, `3_analysis.R`. They would clean the data and store the clean dataset, do exporatory analysis, generate descriptive figure, and do regression analysis that has final tables, respectively.