Module 1

Noah Anderson

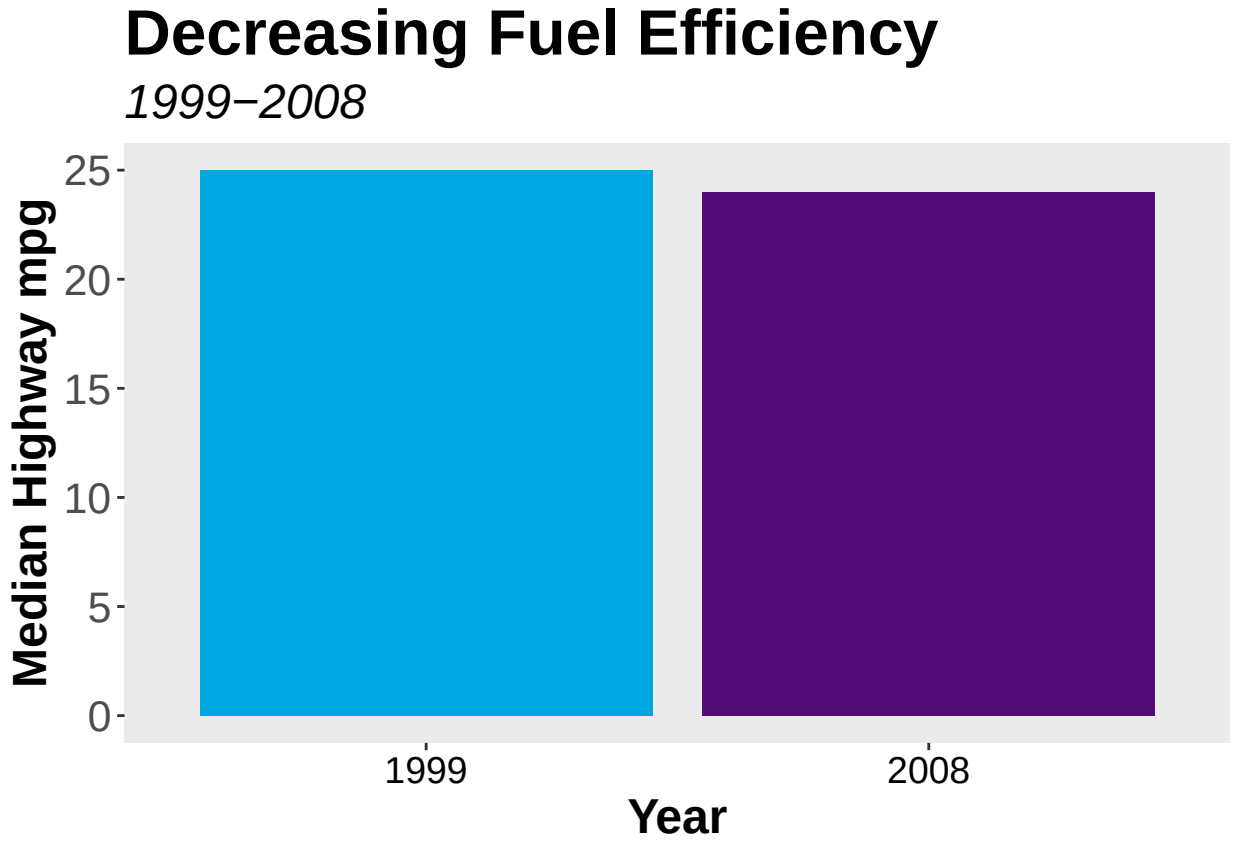
2024-10-26

Read Libraries

```
library(tidyverse)
library(scales)
library(ggpubr)


# Define mpg dataset so it shows up in the global environment
mpg <- mpg %>%

  # Convert year to factor
  mutate(year = as.factor(year))
```

Fuel Economy Decreased from 1999-2008

```
mpg %>%
  group_by(year) %>%

  # Calculate median hwy mpg
  reframe(median_hwy = median(hwy)) %>%
  ggplot(aes(x = year, y = median_hwy, fill = year)) +
  geom_col() +
  scale_fill_manual(values = c( "1999" = "#00A8E2", "2008" = "#510C76")) +
  labs(x = "Year", y = "Median Highway mpg", title = "Decreasing Fuel Efficiency",
       subtitle = "1999-2008") +
  theme(
  legend.position = "none",
  panel.grid.major = element_blank(),  # Remove major gridlines
  panel.grid.minor = element_blank(),  # Remove minor gridlines
  plot.title = element_text(size = 24, face = "bold"),
  plot.subtitle = element_text(size = 18, face = "italic"),
  axis.text.x = element_text(size = 14, color = "black"),
  axis.title = element_text(size = 18, face = "bold"),  # Adjust axis title font size
  axis.text = element_text(size = 16) )
```



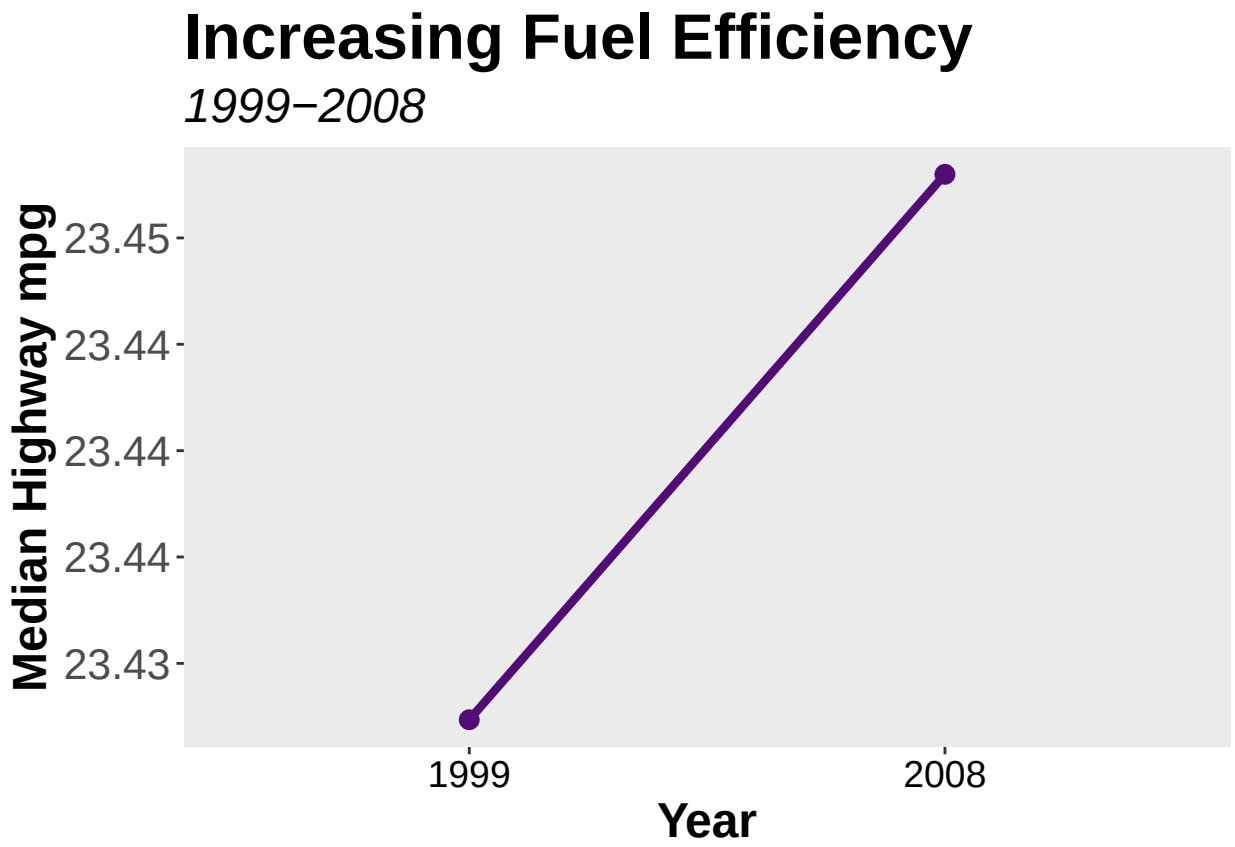
I do not think this necessarily violates principles of good design, but I will say that there is a dishonest data representation present. This aggregates all car classes into one group. Something that I explore in the third plot, the one where I display the graph that tells the most accurate story, is that larger class sizes have become more popular. So when you lump them all together, the median car being driven in 2008 vs 1999 is less efficient, but that is because the cars are bigger.

Fuel Economy Increased from 1999-2008

```
mpg %>%
  group_by(year) %>%

  # Calculate the mean hwy mpg
  reframe(mean_hwy = mean(hwy)) %>%
  ggplot(aes(x = year, y = mean_hwy)) +
  geom_point(color = "#510C76", size = 3) +
  geom_line(group = 1, color = "#510C76", size = 1.5) +

  # scale_fill_manual(values = c( "1999" = "#00A8E2", "2008" = "#510C76")) +
  labs(x = "Year", y = "Median Highway mpg", title = "Increasing Fuel Efficiency",
       subtitle = "1999-2008") +
  scale_y_continuous(labels = label_number(accuracy = 0.01)) +  # Sets y-axis to 2 decimal places

  theme(
  legend.position = "none",
```

```
  panel.grid.major = element_blank(),  # Remove major gridlines
  panel.grid.minor = element_blank(),  # Remove minor gridlines
  plot.title = element_text(size = 24, face = "bold"),
  plot.subtitle = element_text(size = 18, face = "italic"),
  axis.text.x = element_text(size = 14, color = "black"),
  axis.title = element_text(size = 18, face = "bold"),  # Adjust axis title font size
  axis.text = element_text(size = 16) # Adjust axis text font size
  # legend.text = element_text(size = 20),  # Adjust legend text font size
  # legend.title = element_text(size = 22, face = "bold")  # Adjust legend title font size
)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



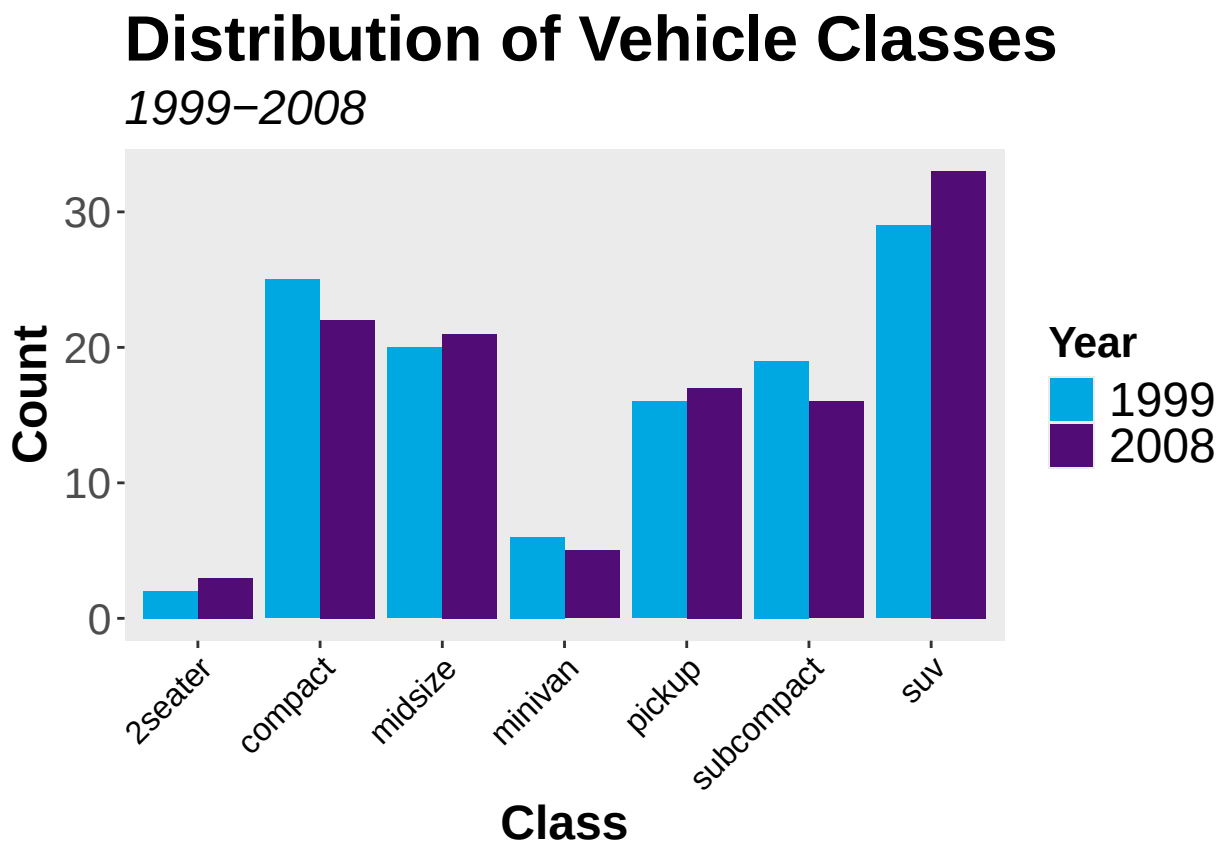
This to me is the most dishonest of the three graphs. For starters, a line plot is not very appropriate for only two data points. Additionally, using the mean here loses a lot of information and allows some outliers in the 2008 dataset to drag up the metric. The most egregious abuse of good data visualization here is the scale. While the distance between the two data points would suggest a large change in mean mpg, in reality it is only a difference in hundredths of a mpg.

What Really Happened?

I think the key factor in change in mpg is that SUVs and pickup trucks have become more and more popular in the U.S. If I had to present this, I would actually do it in two plots. The first one is a distribution of the class types for each year.

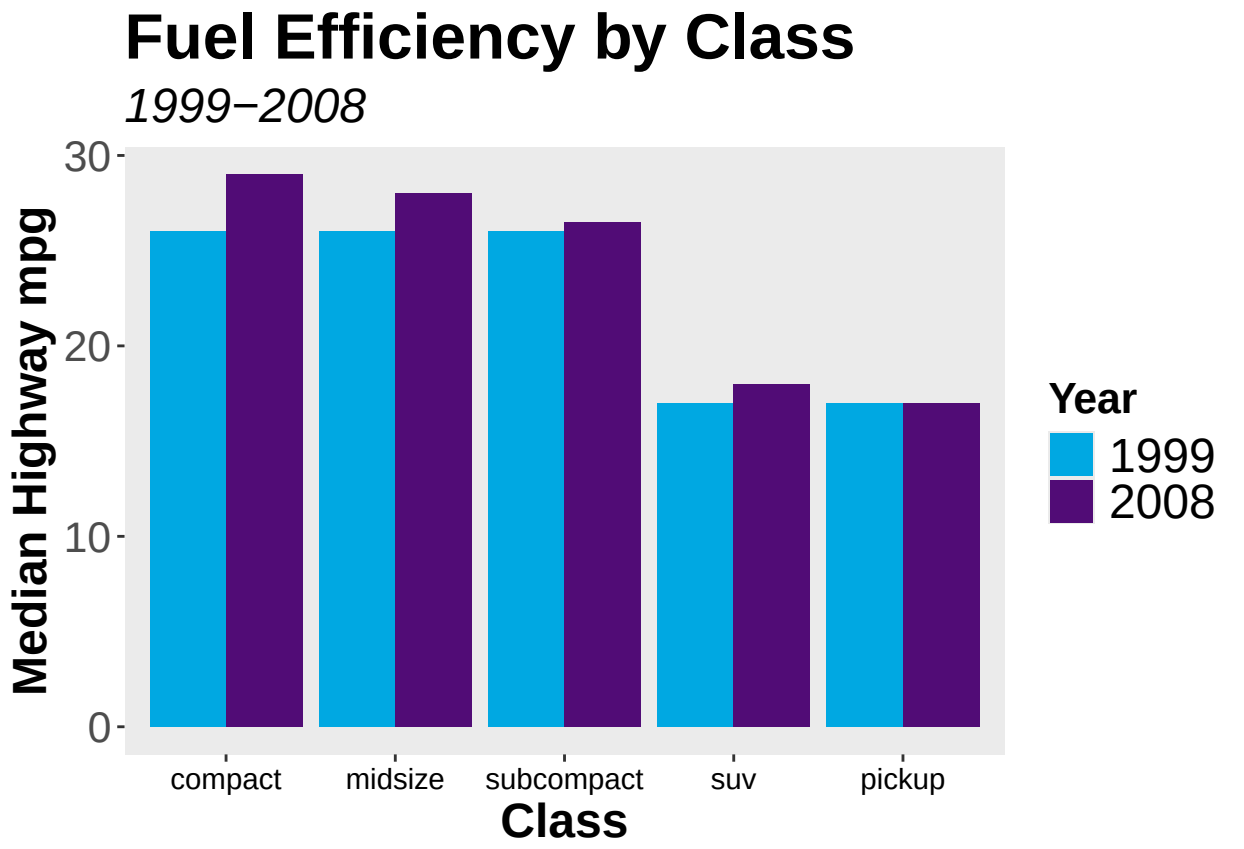
```
mpg %>%
  ggplot(aes(x = class, fill = year)) +
  geom_histogram(stat = "count",position = "dodge") +
  labs(x = "Class", y = "Count", title = "Distribution of Vehicle Classes",
       subtitle = "1999-2008" , fill = "Year") +
    scale_fill_manual(values = c( "1999" = "#00A8E2", "2008" = "#510C76")) +
  theme(
  panel.grid.major = element_blank(),  # Remove major gridlines
  panel.grid.minor = element_blank(),  # Remove minor gridlines
  plot.title = element_text(size = 24, face = "bold"),
  plot.subtitle = element_text(size = 18, face = "italic"),
  axis.text.x = element_text(size = 12, color = "black", angle = 45, hjust = 1),
  axis.title = element_text(size = 18, face = "bold"),  # Adjust axis title font size
  axis.text = element_text(size = 16), # Adjust axis text font size
  legend.text = element_text(size = 18),  # Adjust legend text font size
  legend.title = element_text(size = 16, face = "bold")  # Adjust legend title font size
)
```

```
## Warning in geom_histogram(stat = "count", position = "dodge"): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```



In this plot you can see that compacts and subcompacts decreased while pickups and SUVs increased. That context helps understand why we see the apparent decrease in median mpg from 1999 to 2008 that is shown in the first plot.

```
mpg %>%
  mutate(class = fct_reorder(class, hwy, .fun = median, .desc = TRUE)  ) %>%


  # Dropping 2seater and minivan since they have very low n's compared to the other
  # classes
  filter(!class %in% c("2seater", "minivan")) %>%

  group_by(year, class) %>%
  reframe(median_hwy = median(hwy)) %>%
  ggplot(aes(x = class, y = median_hwy, fill = year)) +
  geom_col(position = "dodge")  +
  scale_fill_manual(values = c( "1999" = "#00A8E2", "2008" = "#510C76")) +
  labs(x = "Class", y = "Median Highway mpg", title = "Fuel Efficiency by Class",
       fill = "Year", subtitle = "1999-2008") +

  theme(
  panel.grid.major = element_blank(),  # Remove major gridlines
  panel.grid.minor = element_blank(),  # Remove minor gridlines
  plot.title = element_text(size = 24, face = "bold"),
  plot.subtitle = element_text(size = 18, face = "italic"),
  axis.text.x = element_text(size = 11, color = "black"),
  axis.title = element_text(size = 18, face = "bold"),  # Adjust axis title font size
  axis.text = element_text(size = 16), # Adjust axis text font size
  legend.text = element_text(size = 18),  # Adjust legend text font size
  legend.title = element_text(size = 16, face = "bold")  # Adjust legend title font size
)
```



I think this plot communicates a more accurate picture of what really has happened with fuel economy over time. We see that there were modest improvements from 1999 to 2008 across all classes except for pickup trucks which have appears to stagnate. Putting together the information we learned from the overall median plot, the distributions of class types in each year, and this plot we get closer to an answer of what is going on. From an engineering point of view, it is certainly less alarming that we are not getting worse at designing cars. From a climate point of view though, the growth in larger cars has the effect of evening out the gains made in fuel efficiency. Considering that we need to dramatically reduce carbon emissions, this is an alarming development.I think this is a great example of how data visualizations, but also framing and narrative can dramatically change the major takeaways from the very same dataset.

As far as design choices, the use of barcharts for the median certainly loses a lot of information, but I found that the boxplots were not super helpful due to some overlapping data points resulting in boxes rendering as just a median line.