Examining Covariations Between Variables

Pauline I. Alvarado

Spring 2018

This coding exercise was from the Program Evaluation and Data Analysis course taught by Dr. Nelson Lim at the University of Pennsylvania. Data was provided by the instructor.

Key Packages

```
library(tidyverse)
library(readr)
library(ggplot2)
library(ggthemes)
```

Bi-Variate or Multivariate Covariations

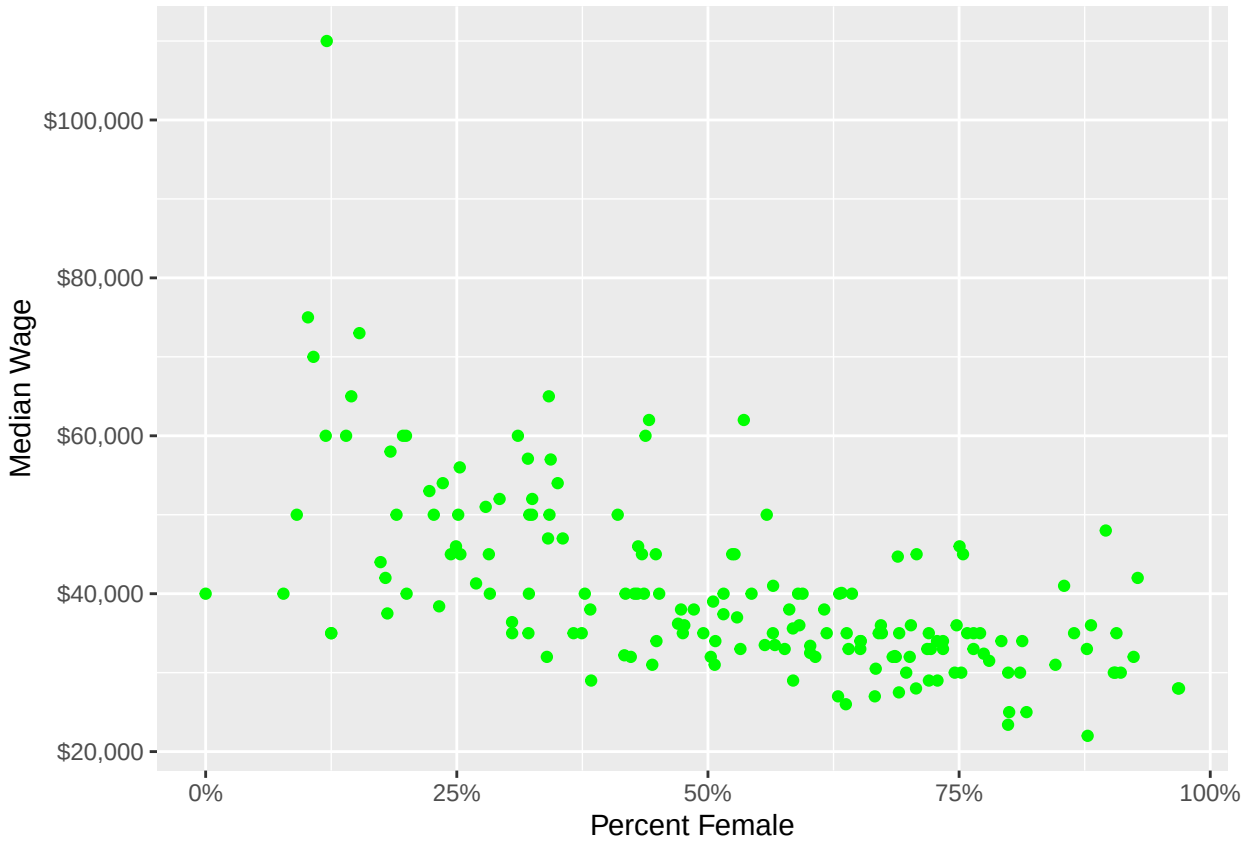
First step toward determine causality. Important to determine input-output, independent-dependent relationship between the variables, and design the visualization and analysis based on that determinant.

Covariation between continuous variables

Create scatterplot

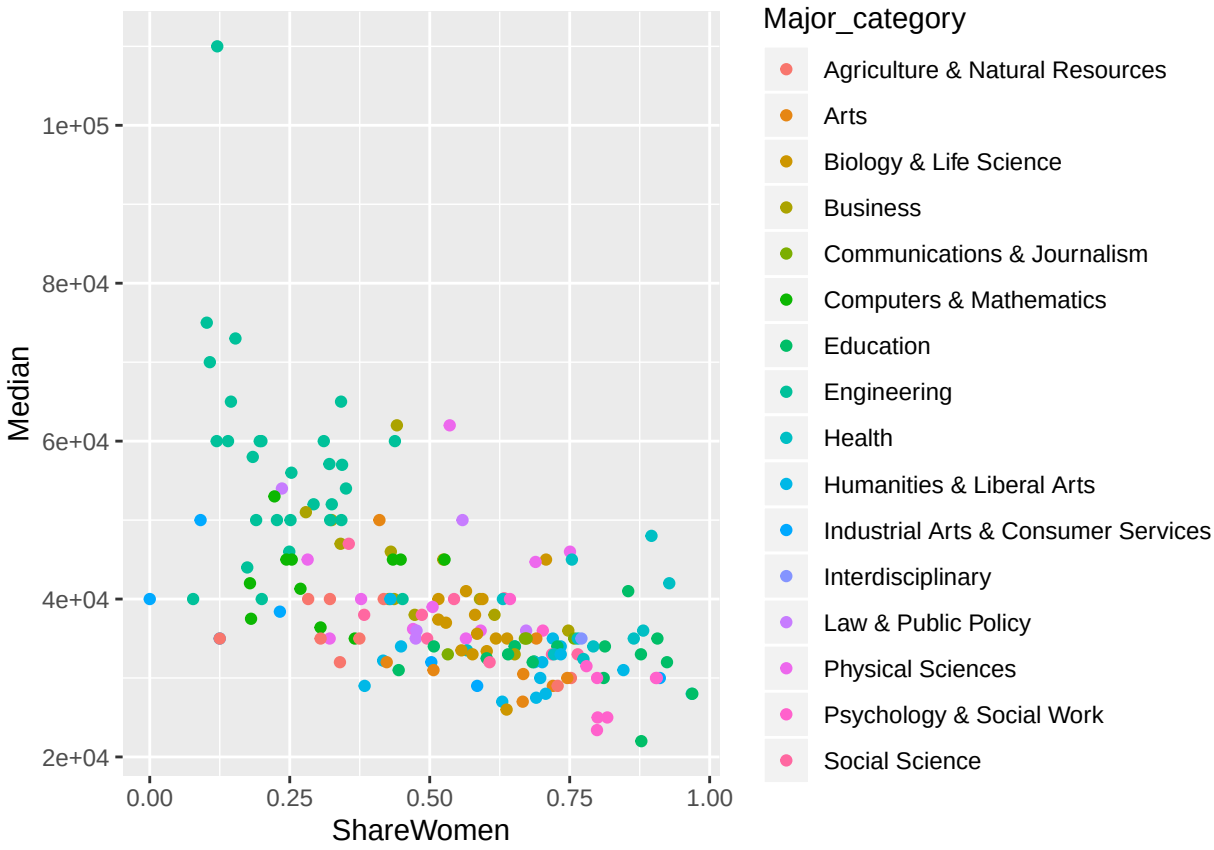
Visualize covariation between the representation of females in a college major and the median income of the the major. Data taken from fivethirtyeight.

```
college_major <- read.csv("recent-grads.csv")

ggplot(data = college_major) + geom_point(mapping = aes(x=ShareWomen, y=Median), color = "green") +

      # Format x and y labels
        scale_x_continuous(name = "Percent Female", labels = scales::percent) +
        scale_y_continuous(name = "Median Wage", labels = scales::dollar)
```



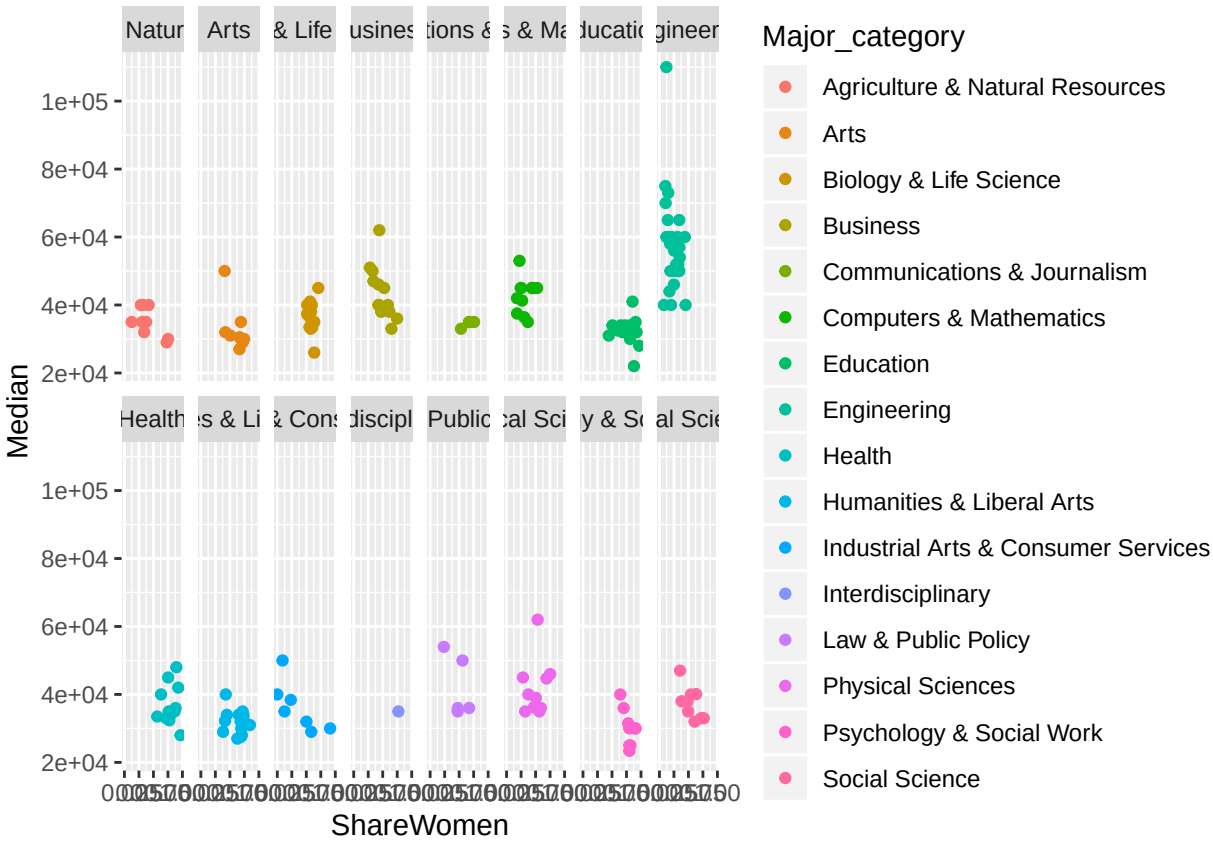
Add third dimension to the plot

```
ggplot(college_major) + geom_point(aes(x = ShareWomen, y = Median, color = Major_category))
```



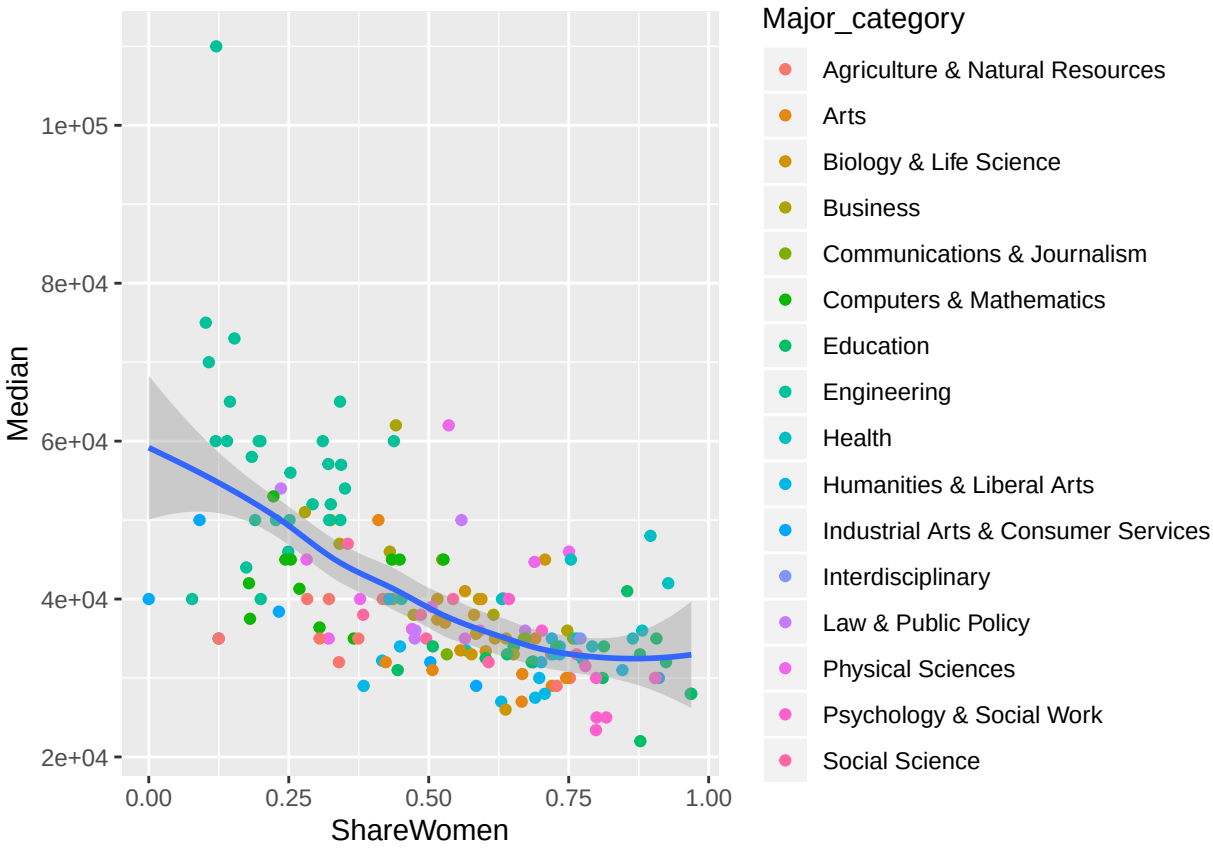
Facets

```
ggplot(data = college_major) +
    geom_point(mapping = aes (x = ShareWomen, y = Median, color = Major_category)) +
    facet_wrap(~Major_category, nrow = 2)
```



Multivariate scatterplot with regression lines

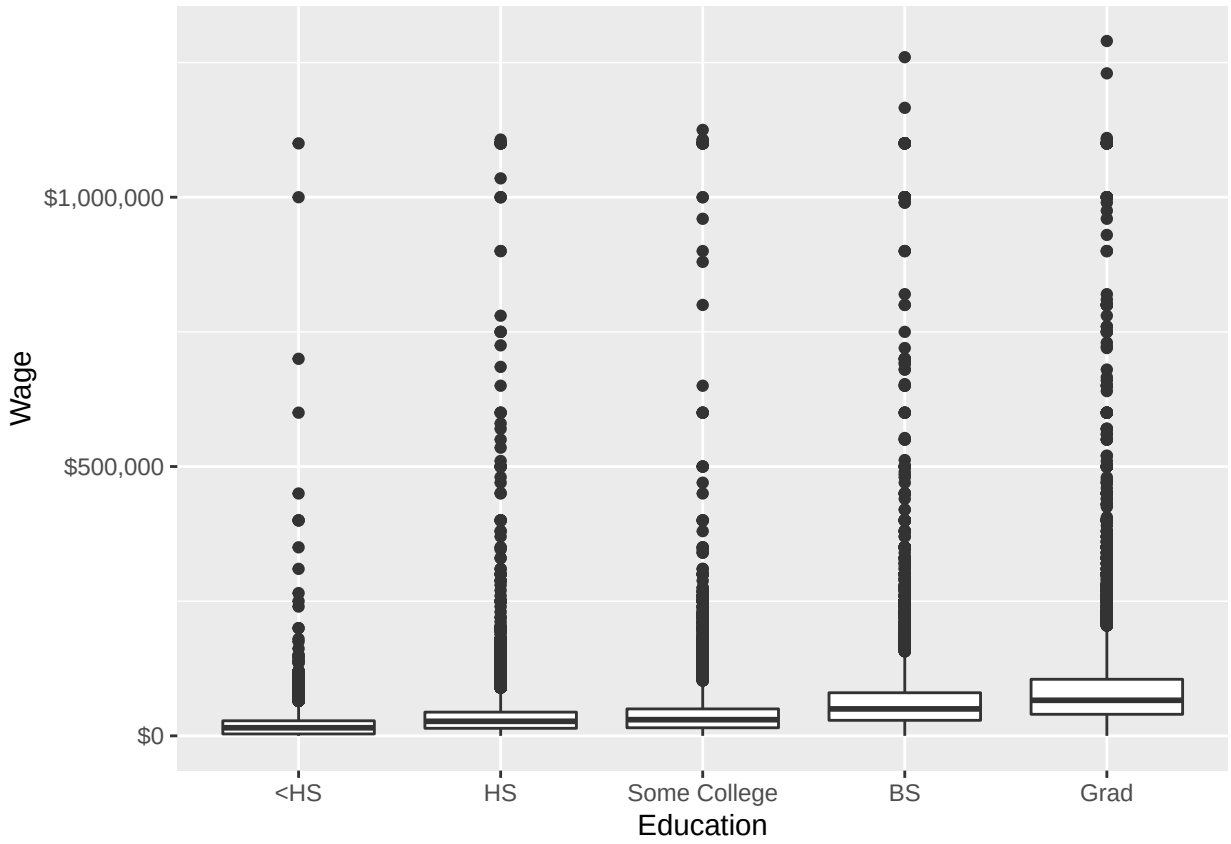
```
ggplot(data = college_major) +
   geom_point(mapping = aes(x=ShareWomen, y=Median, color=Major_category)) +
   geom_smooth(aes(x=ShareWomen, y=Median))
```



Boxplot

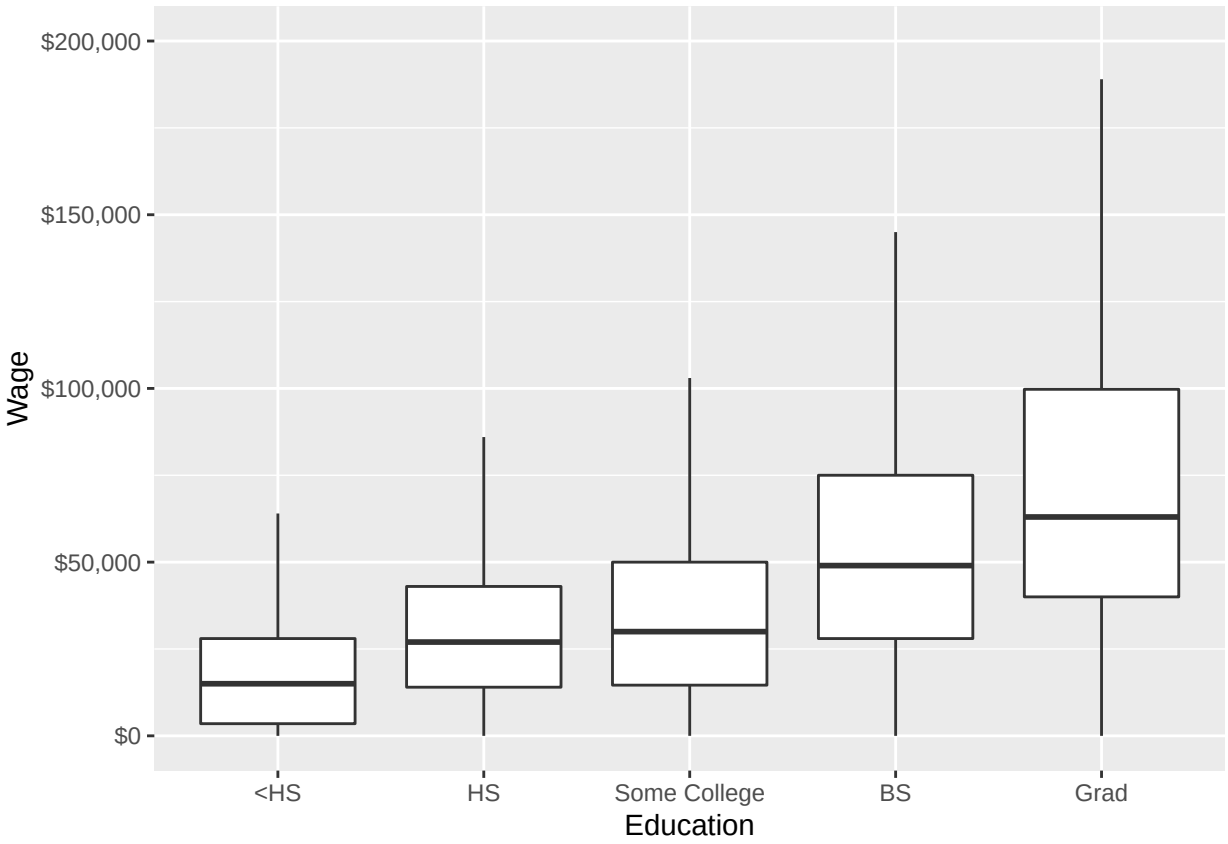
Bin continuous variable to act like a categorical variable. Plot covariation between degree and wage. Data taken from the US Bureau of Labor Statistics' Current Population Survey.

```
cps_small <- readRDS("cps-2017-small.rds")

# Adjust bin for "Yrs_Schooling"
  cps_small <- cps_small %>% drop_na(Yrs_Schooling) %>%
                             mutate(Degree = cut(Yrs_Schooling, breaks=c(-Inf, 11, 12, 15, 16, 20),
                                    labels = c("<HS","HS","Some College", "BS", "Grad")))

# Plot
  ggplot(data = cps_small, mapping = aes(x = Degree, y = Wage)) +
    geom_boxplot() +
    scale_x_discrete(name = "Education") +
    scale_y_continuous(name = "Wage", labels = scales::dollar)
```



Remove the outliers and zoom in and see the relationship better

```
ggplot(data = cps_small, mapping = aes(x = Degree, y = Wage)) +
  geom_boxplot(outlier.shape = NA) +
  scale_x_discrete(name = "Education") +
  scale_y_continuous(name = "Wage", labels = scales::dollar, limits = c(0, 200000))
```

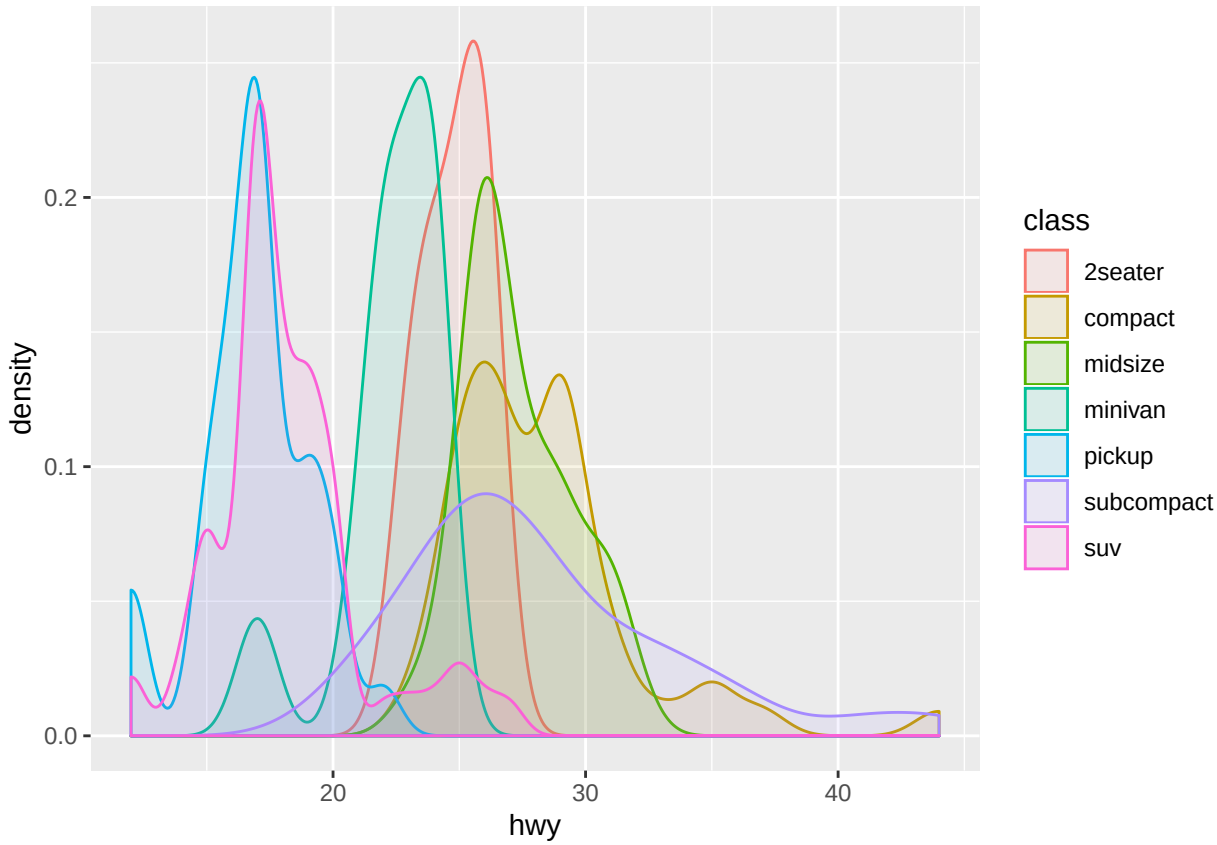


Covariation Between Categorical & Continuous Variable

Explore the distribution of a continuous variable broken down by categorical variable.

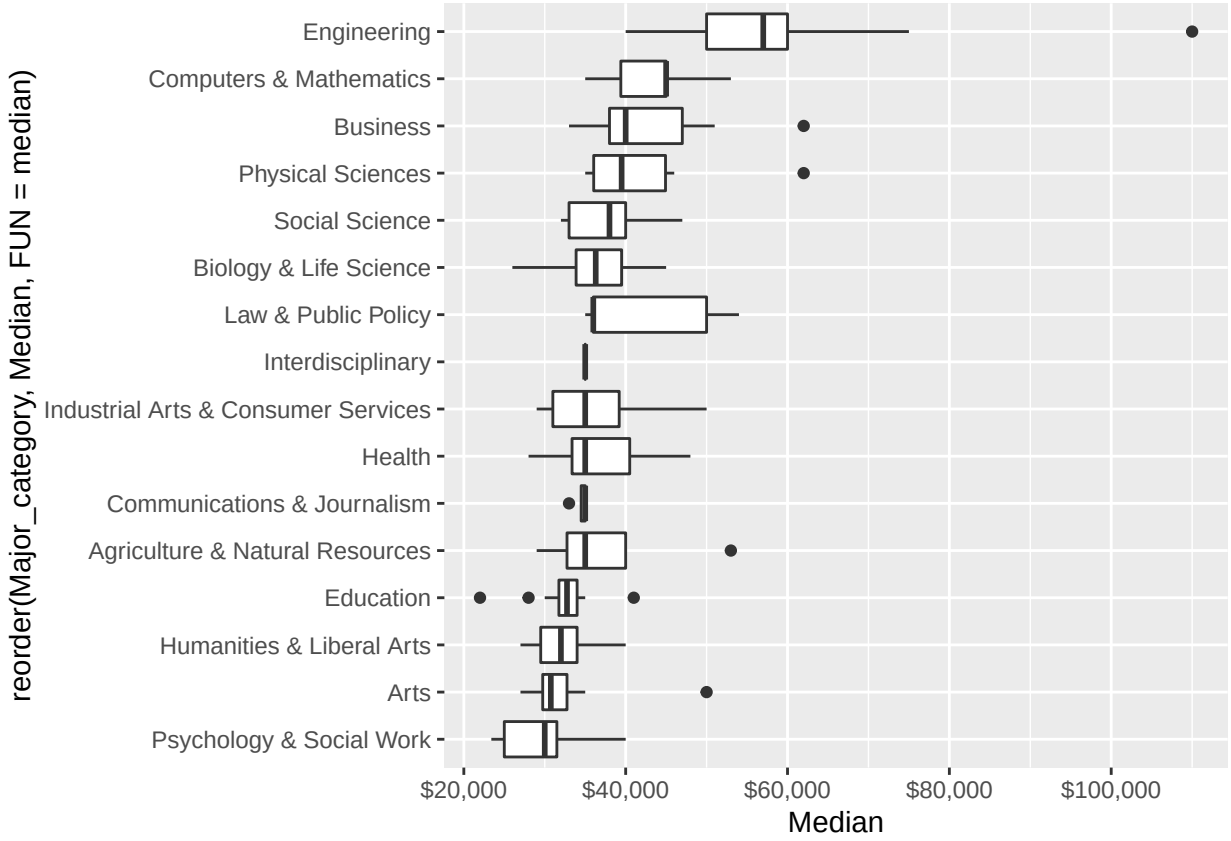
Density plot

```
#categorical = input/x
#continuous = output/y

ggplot(mpg, aes(hwy, fill=class, colour=class)) + geom_density(alpha=0.1)
```



Box plot

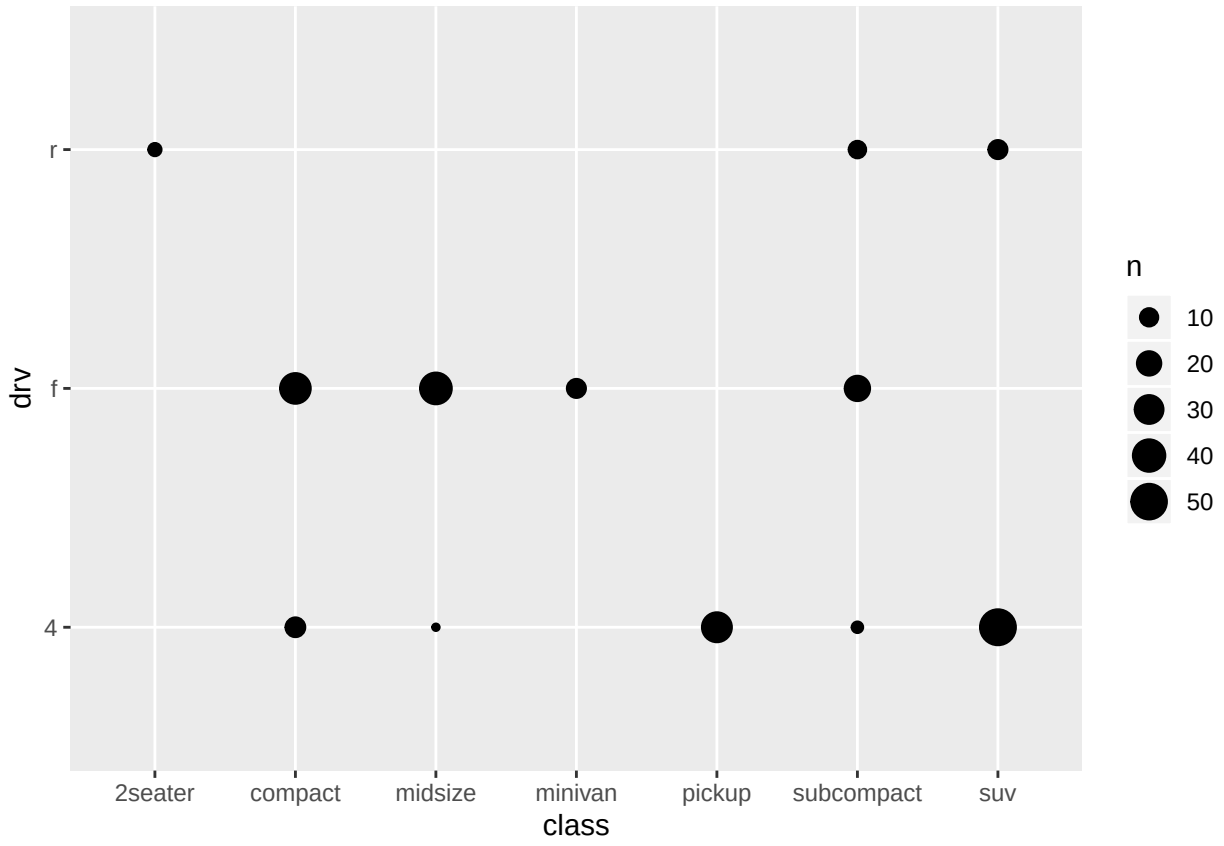
```
ggplot(data = college_major) +

  # sort majors highest to lowest median salary
  geom_boxplot(mapping = aes(x=reorder(Major_category, Median, FUN=median), y=Median)) +

  #flip the axes
  coord_flip() + #flip the axes

  #remove scientific notation
  scale_y_continuous(labels=scales::dollar)
```



Covariation between two categorical variables

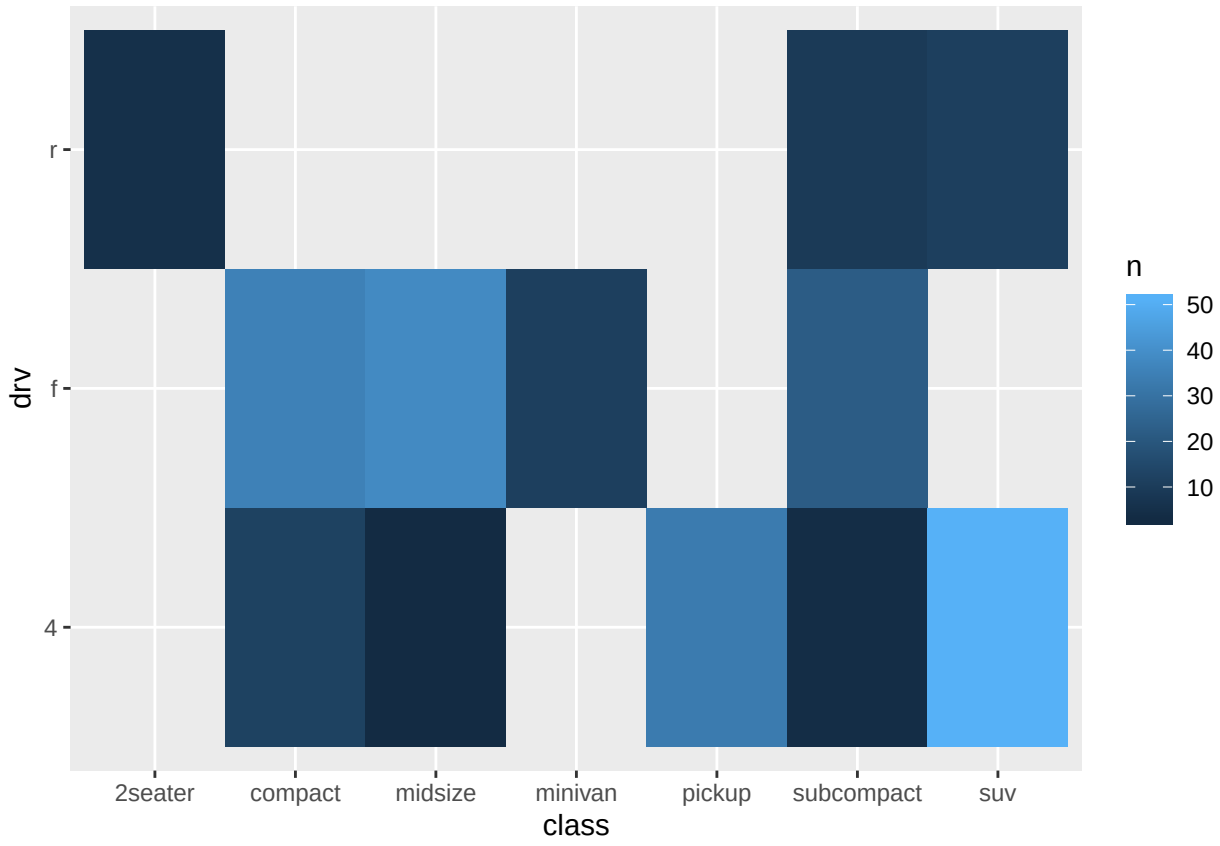
Circles

```
ggplot(data = mpg) + geom_count(mapping = aes(x=class, y=drv))
```



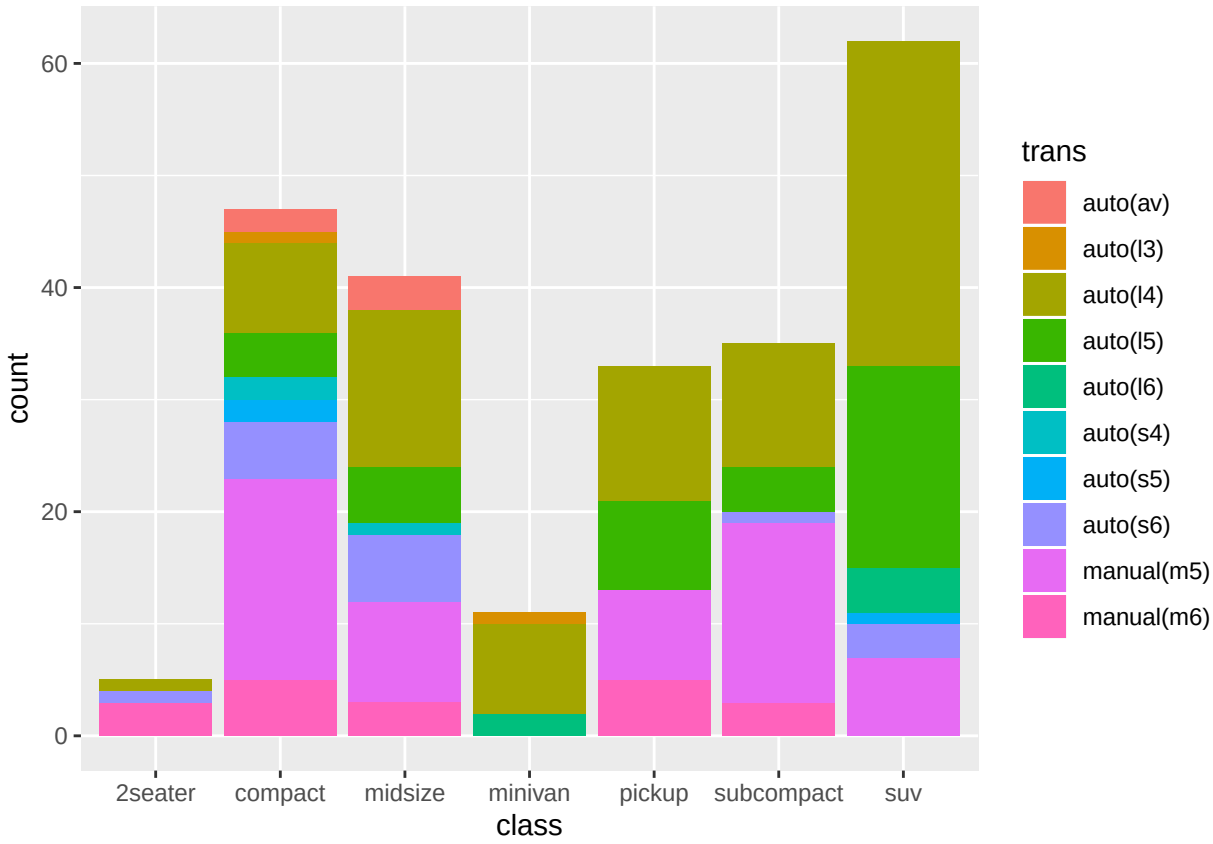
Tiles

```
mpg %>%
  count(class,drv) %>%
  ggplot(mapping=aes(x=class, y=drv)) +
  geom_tile(mapping = aes(fill=n))
```



Stacked Bar Graph

```
ggplot(data = mpg) + geom_bar(mapping = aes(x=class, fill=trans))
```



Formatting Exercise

Difference in wage distributions across states

```
ggplot(data = cps_small) +
    geom_smooth(mapping = aes(x = Yrs_Schooling, y = Wage, colour = Sex)) +

  #add title, subtitle, caption
  labs(
    title = "Gender gap in wages increase with the level of education in 2017",
    subtitle = "On average, women earn less than men across all levels of education",
    caption = "Source: Current Population Survey") +

  #add annotations
  annotate("text", x = 8, y = 55000, label = "Non-linear regression lines with standard errors") +

  #format labels
  scale_x_continuous(name = "Education (Years of Schooling)", breaks = seq(0, 20, 2)) +
  scale_y_continuous(name = "Wage", breaks = seq(0, 150000, 25000), labels = scales::dollar) +

  #format line color
  scale_color_manual(values = c(Male="red", Female="blue"), labels=c("Men", "Women")) +

  #format grid lines
  theme(
```

```
    panel.grid.major = element_line(size=.5, linetype = 'solid', colour = "White"),
    panel.grid.minor = element_line(size = 0.25, linetype = 'solid', colour = "White")
  )
```