Univariate and Bivariate

Bibek Sapkota

Univariate Graphs

```
#packages <- c("ggplot2","dplyr2", "scales", "mosaicData", "ggpie", "treemapify", "waffle")
#for (package_name in packages) {
# if (!requireNamespace(package_name, quietly = TRUE)) {
#    install.packages(package_name)
#  }
#}

library(ggplot2)
library(dplyr)
```

Importing library and Dataset

```
##
## 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(scales)
library(mosaicData)
library(ggpie)
library(treemapify)
library(waffle)

data("Marriage", package = "mosaicData")
```

```
#Marriage
```

Displaying the dataset

CATEGORICAL

Bar chart

A bar chart is used because it effectively displays the distribution of a single categorical variable.

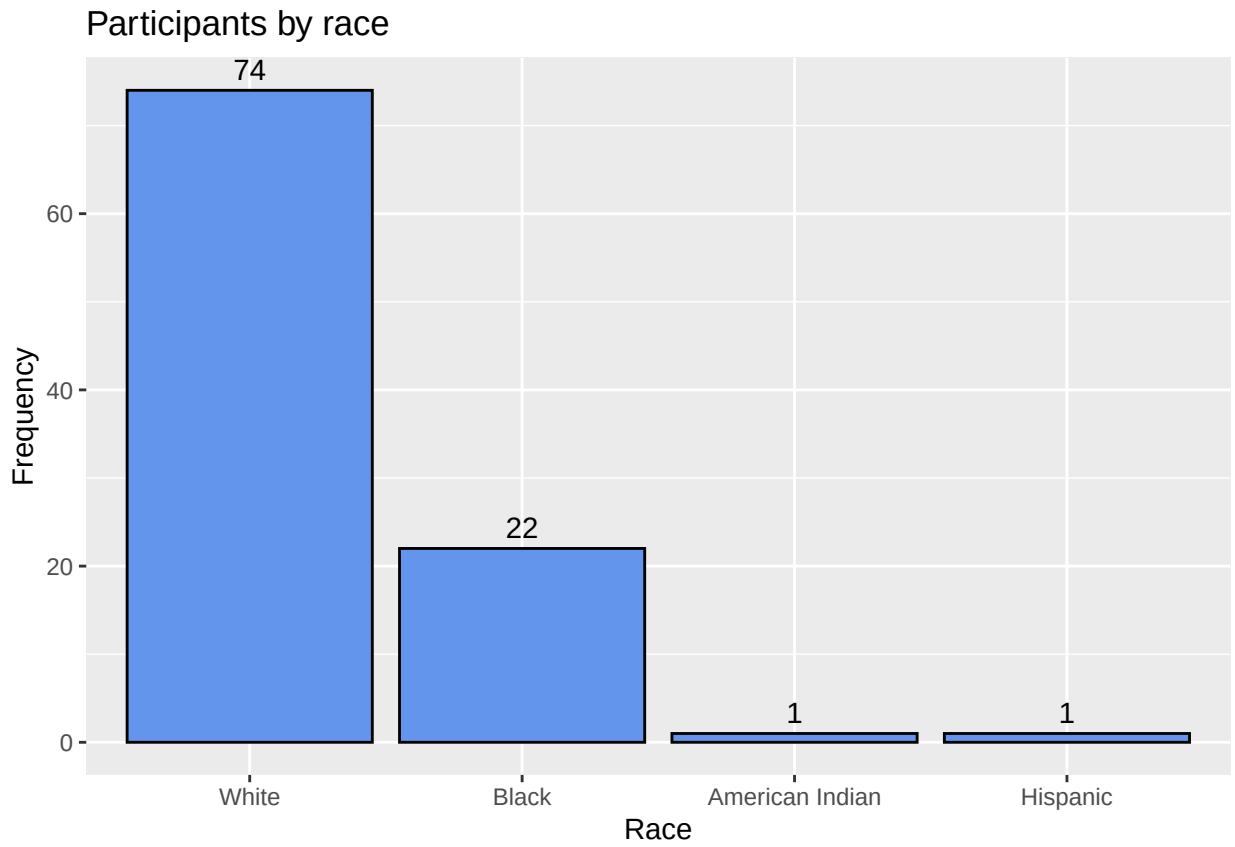
Sorting Categories Calculating the number of participants in each race category

```
plotdata <- Marriage %>%
 count(race)
```

Labeling bars

Bar chart with numerical labels Task 1: Ploting the distribution of race in marriage dataset.(The bargraph can be sorted in ascending order using (race ,n) instead of (race, -n))

```
ggplot(plotdata,
       aes(x = reorder(race, -n), y = n)) +
  geom_bar(stat="identity",
           fill="cornflowerblue",
           color="black") +
  geom_text(aes(label = n), vjust=-0.5) +
  labs(x = "Race",
       y = "Frequency",
       title  = "Participants by race")
```

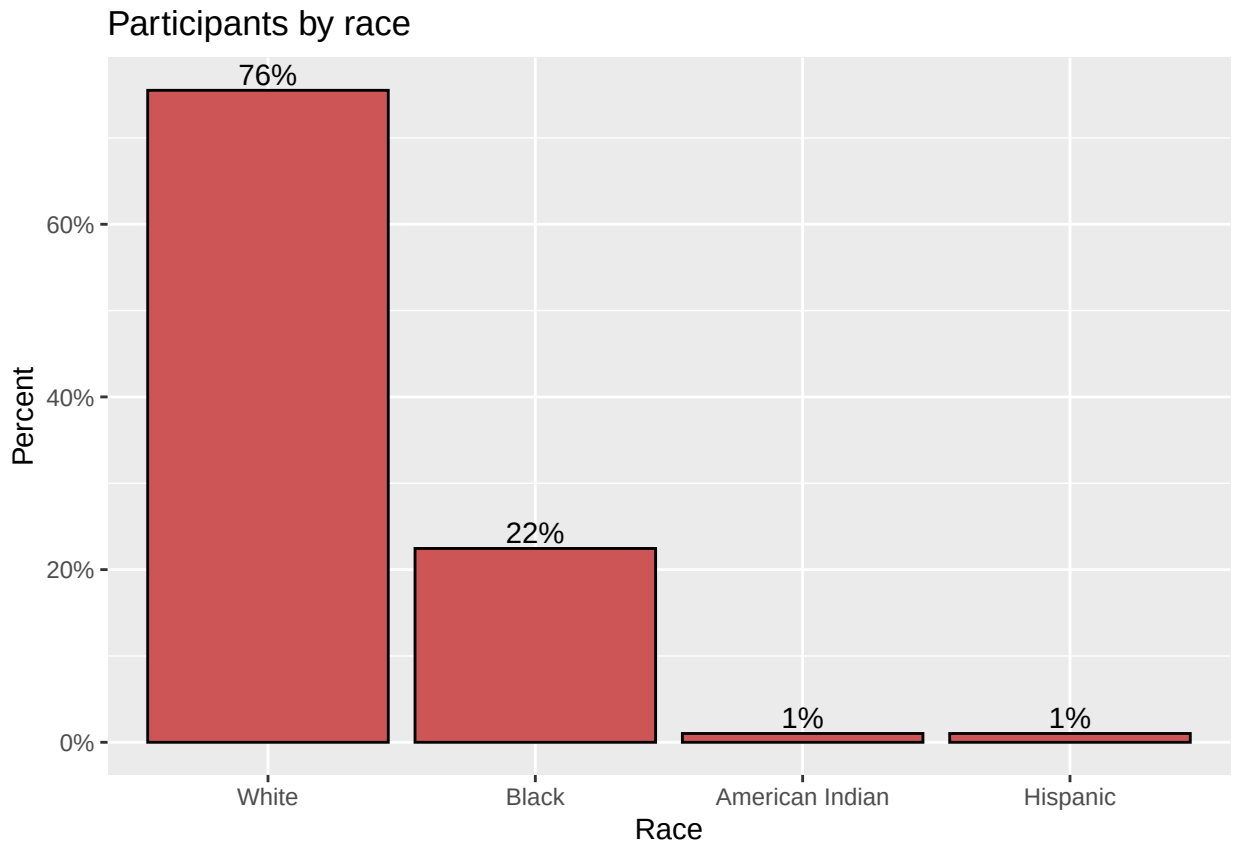


Interpretation: This bar plot shows the frequency of participants by race. Each bar's height represents the number of participants, with the bars ordered in descending order of frequency and labeled with their corresponding counts. The output reveals 74 White participants, 22 Black participants, 1 American Indian participant, and 1 Hispanic participant.

Bar chart with Percents labels Task 2:Visualizing the percentage of marriage participants by race using a bar plot, sorted by frequency with percentage labels.

```
plotdata <- Marriage %>%
  count(race) %>%
  mutate(pct = n / sum(n),
         pctlabel = paste0(round(pct*100), "%"))

ggplot(plotdata,
       aes(x = reorder(race, -pct), y = pct)) +
  geom_bar(stat="identity", fill="indianred3", color="black") +
  geom_text(aes(label = pctlabel), vjust=-0.25) +
  scale_y_continuous(labels = percent) +
  labs(x = "Race",
       y = "Percent",
       title  = "Participants by race")
```

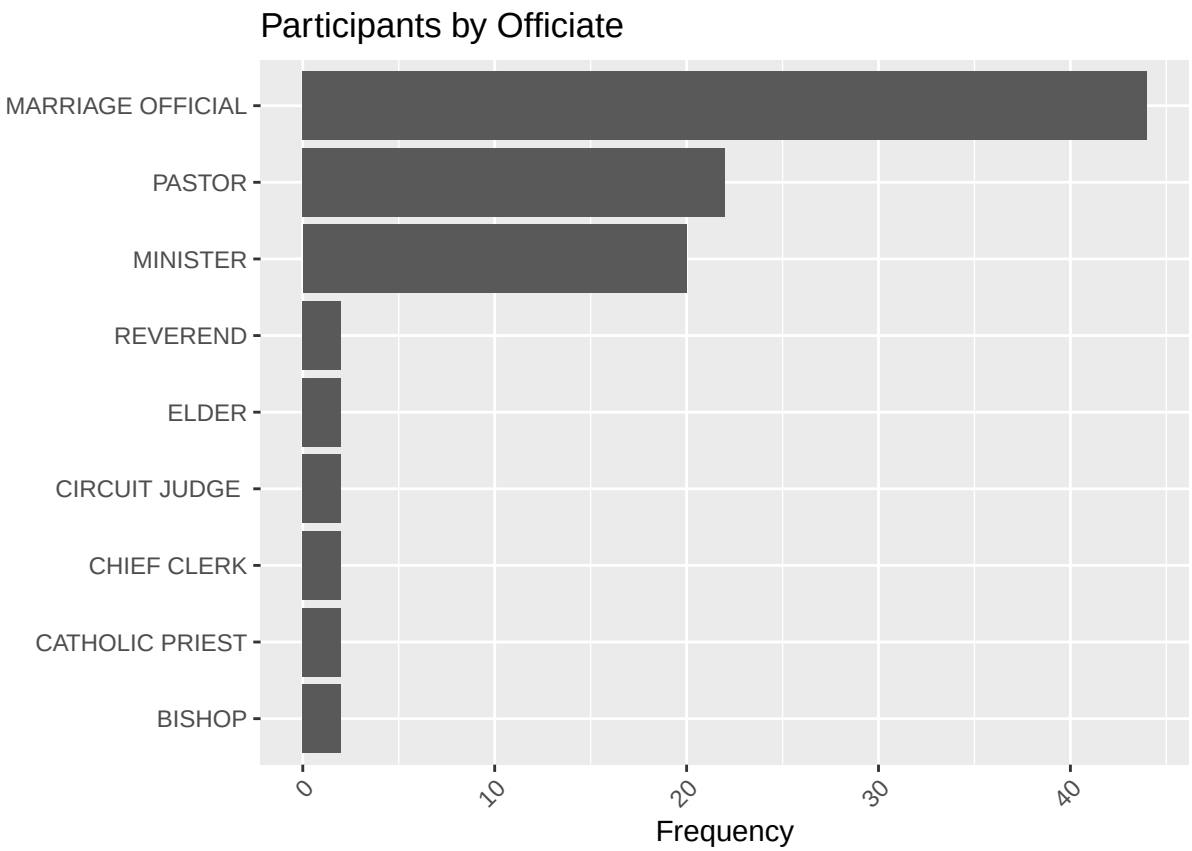


Overlapping labels

Task: Visualizing the Participants by Officiate and fliping the bar plot to plot the label clearly without overlapping.

```
plotdata <- Marriage %>%
 count(officialTitle)

ggplot(plotdata, aes(x = reorder(officialTitle, n), y = n)) +
  geom_bar(stat = "identity") +
  labs(x = " ",
       y = "Frequency",
       title = "Participants by Officiate") +
  coord_flip()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

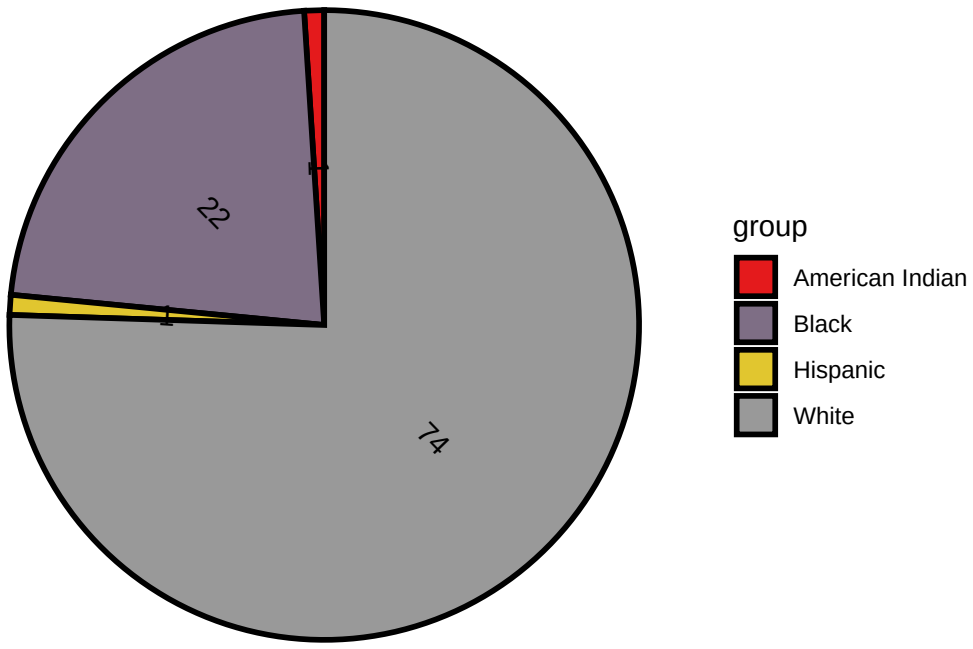


Interpretation: This bar chart creates a horizontal bar plot to visualize the frequency of participants by their officiate titles We can see the Marriage official has the highest frequency followed by pastor and so on where as Bishop has the lowest frequency.

Pie chart

Task: Visualizing the marriage group using pie-chart

```
ggpie(Marriage, group_key = "race", count_type="full", label_info = "count")
```

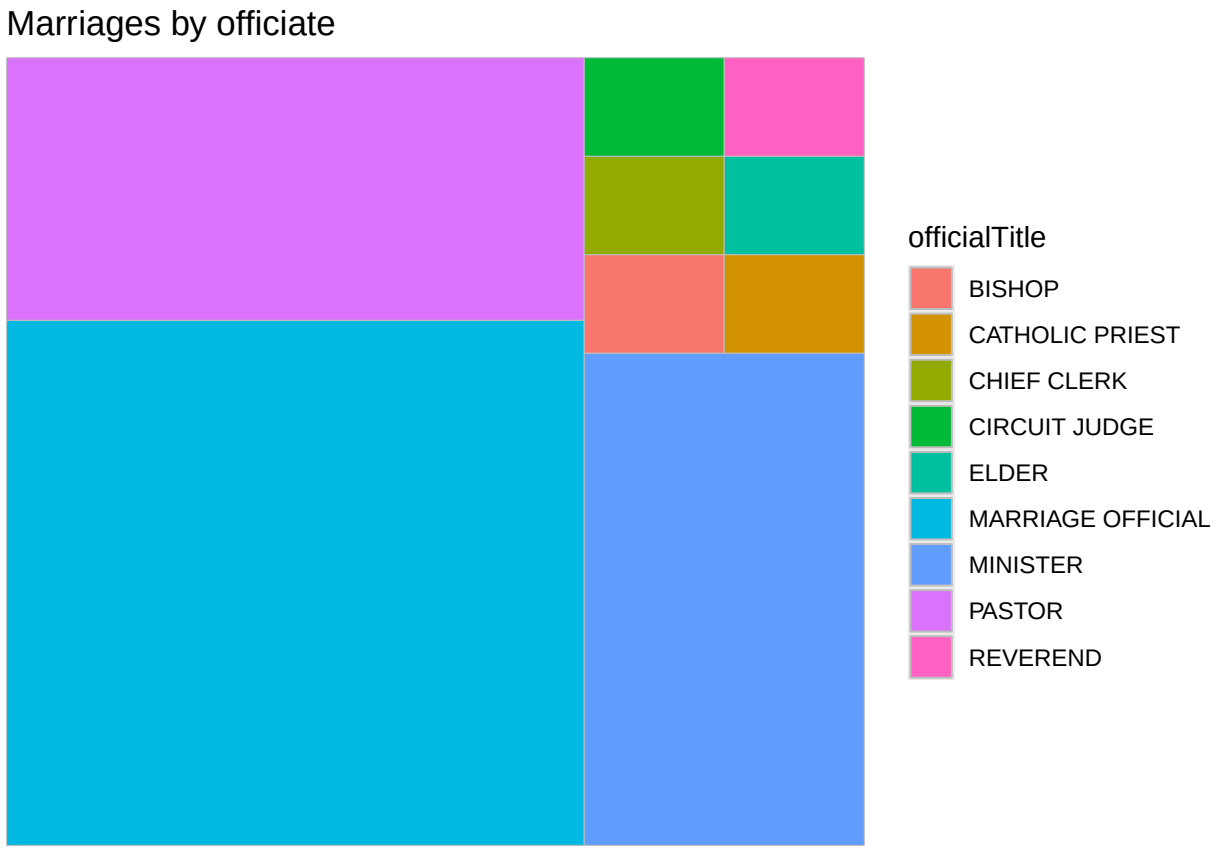


Interpretation: This pie chart visualizes the distribution of marriage participants by race. The largest segment represents White participants, followed by Black and Hispanic participants, with American Indian participants forming the smallest segment.

Tree Map

```
plotdata <- Marriage %>%
  count(officialTitle)

ggplot(plotdata,
       aes(fill = officialTitle, area = n)) +
  geom_treemap() +
  labs(title = "Marriages by officiate")
```

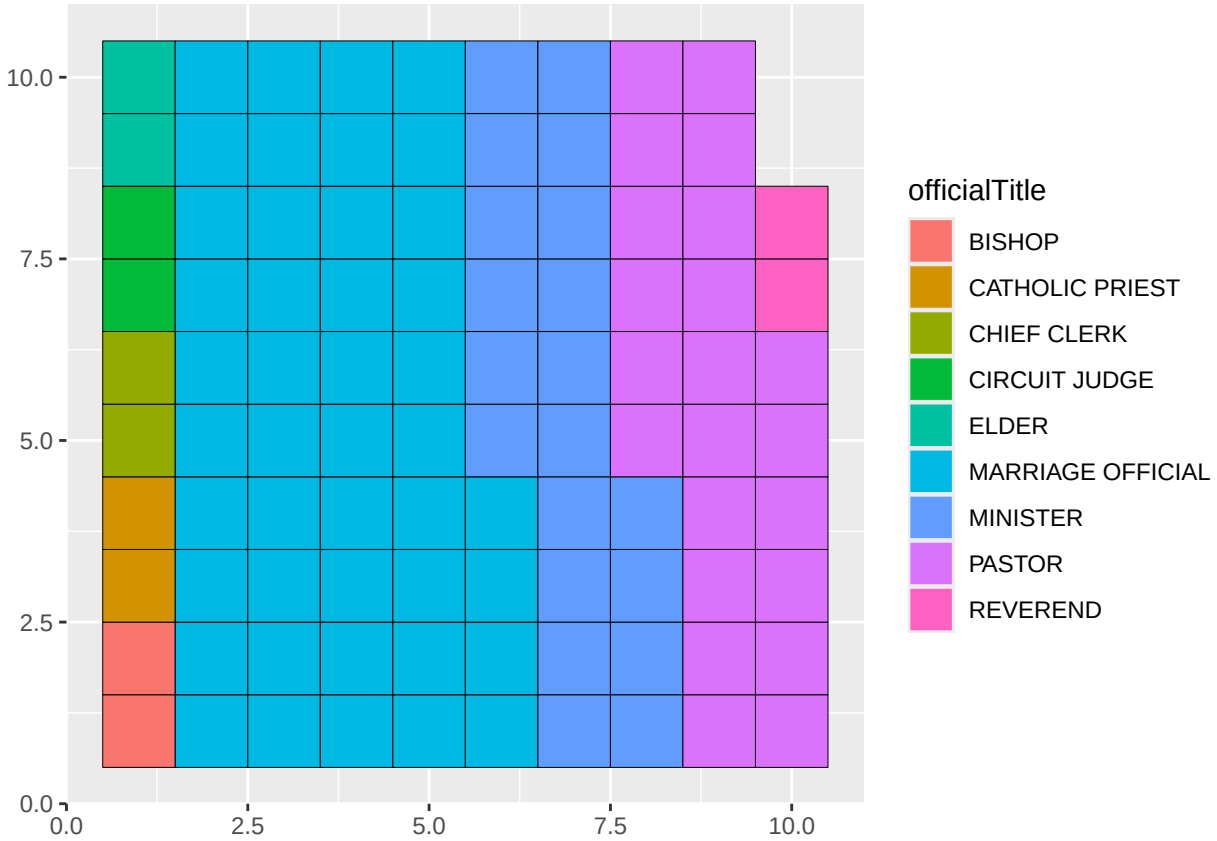


Interpretation: It shows the frequency of marriages conducted by different types of officiates. Larger blocks represent higher frequencies, indicating that "MINISTER" and "MARRIAGE OFFICIAL" conduct most ceremonies. Smaller blocks indicate less frequent officiates.The chart shows that certain officiates, like "MINISTER" and "MARRIAGE OFFICIAL," conduct the majority of the ceremonies, as indicated by the larger block sizes for these categories.

WAFFLE CHART

```
plotdata <- Marriage %>%
  count(officialTitle)
```

```
ggplot(plotdata, aes(fill = officialTitle, values=n)) +
  geom_waffle(na.rm=TRUE)
```



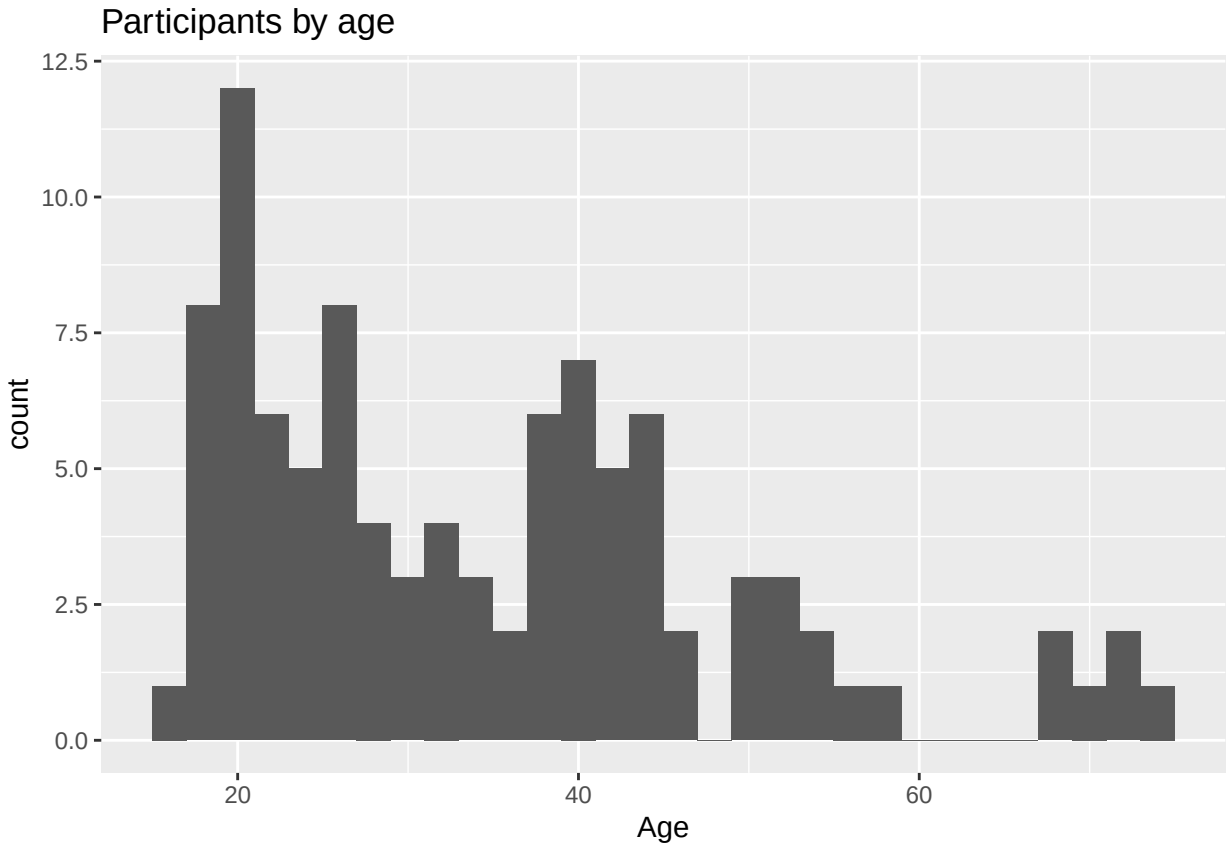
Interpretation: This waffle chart shows the frequency of marriages conducted by various officiates. Each square represents a count, with larger colored sections indicating higher frequencies. "MARRIAGE OFFICIAL" and "MINISTER" have the largest sections, indicating they conduct the most marriages.

QUANTITATIVE

Histogram

```
ggplot(Marriage, aes(x = age)) +
  geom_histogram() +
  labs(title = "Participants by age",
       x = "Age")
```

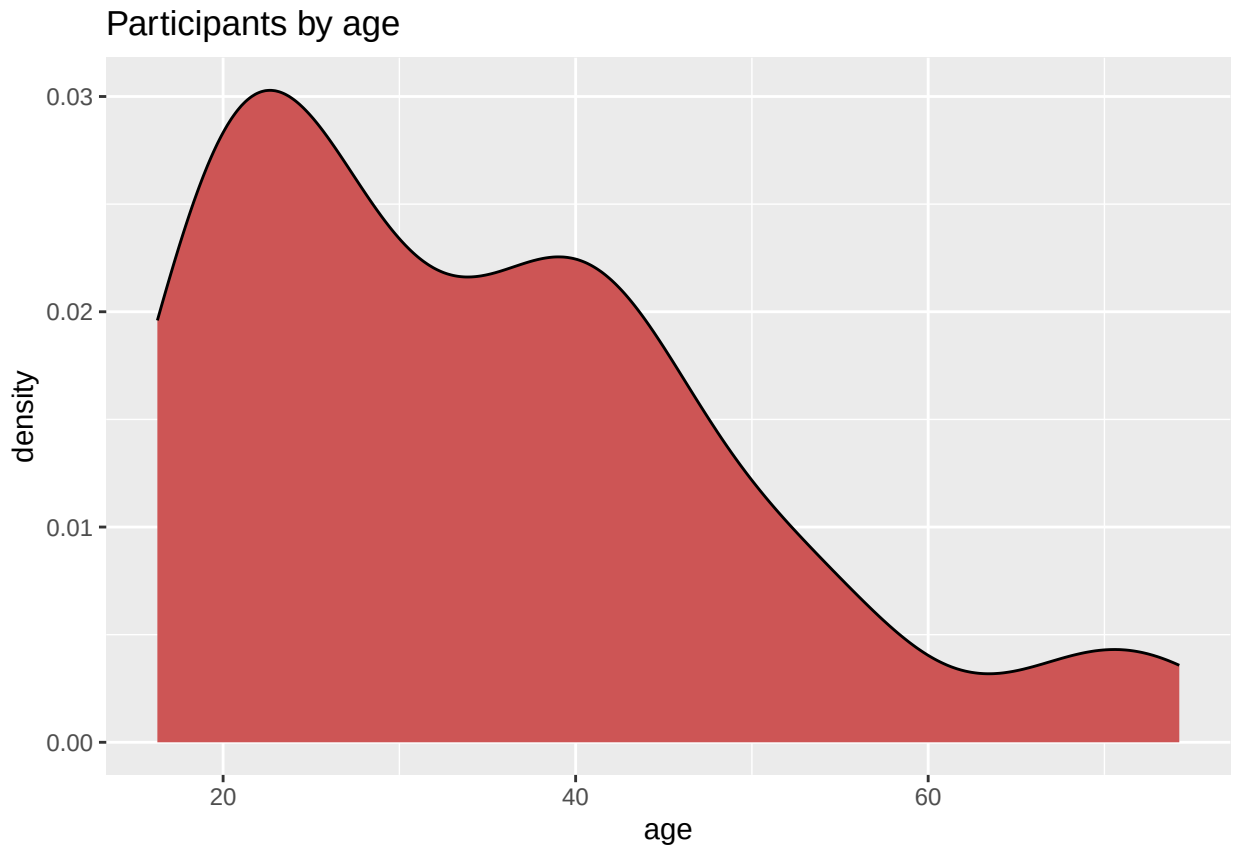
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Interpretation: The histogram displays the distribution of participants by age. Most participants are in their early 20s, with smaller peaks around ages 30, 40, and 55. There are fewer participants in the age ranges 25-30 and 45-50.

Kernel Density Plot

```
ggplot(Marriage, aes(x = age)) +
  geom_density(fill = "indianred3") +
  labs(title = "Participants by age")
```

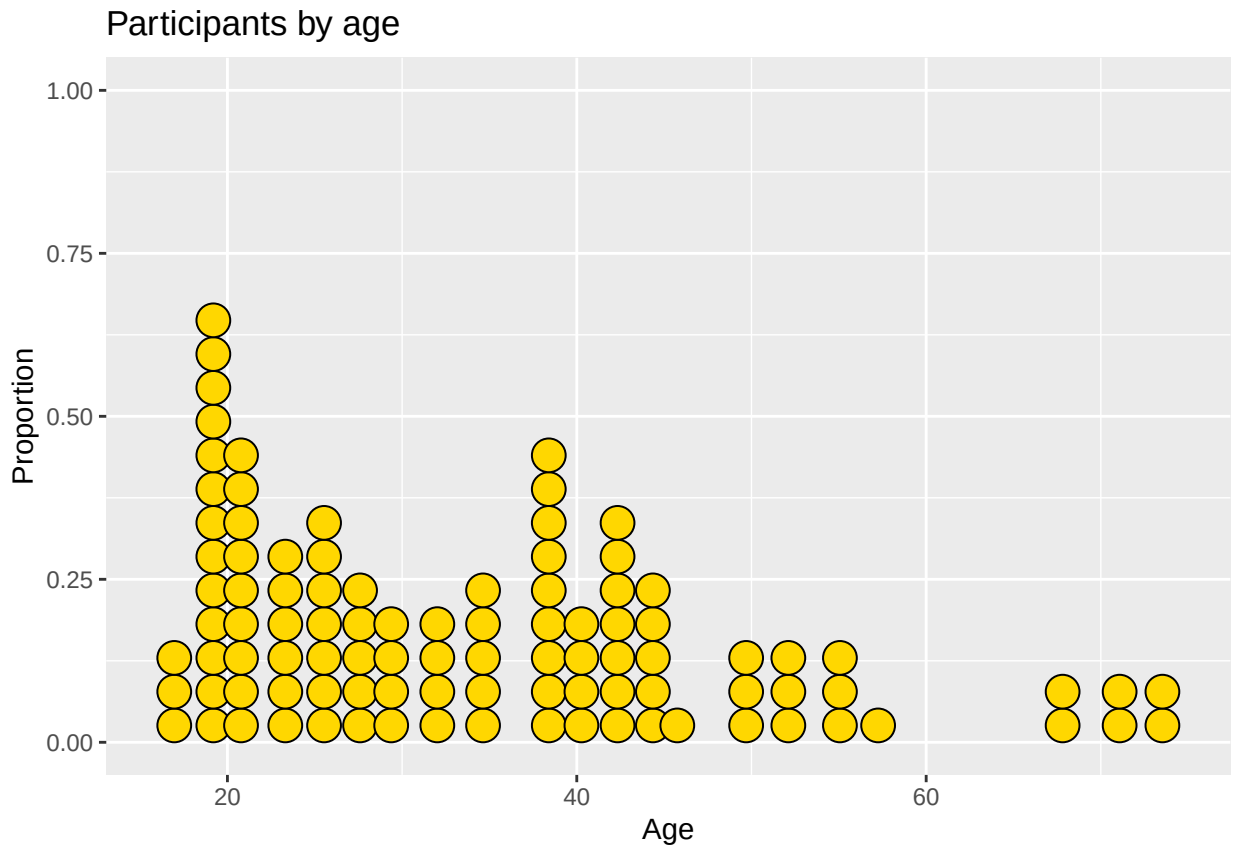


Interpretation: The density plot shows the age distribution of participants, indicating that the highest density of participants is around age 20. The density gradually decreases with age, showing a secondary smaller peak around age 40, and a slight increase again after age of 60.

Dot chart

```
# Plot ages as a dot plot using
# gold dots with black borders
ggplot(Marriage, aes(x = age)) +
  geom_dotplot(fill = "gold",
               color="black") +
  labs(title = "Participants by age",
       y = "Proportion",
       x = "Age")
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```



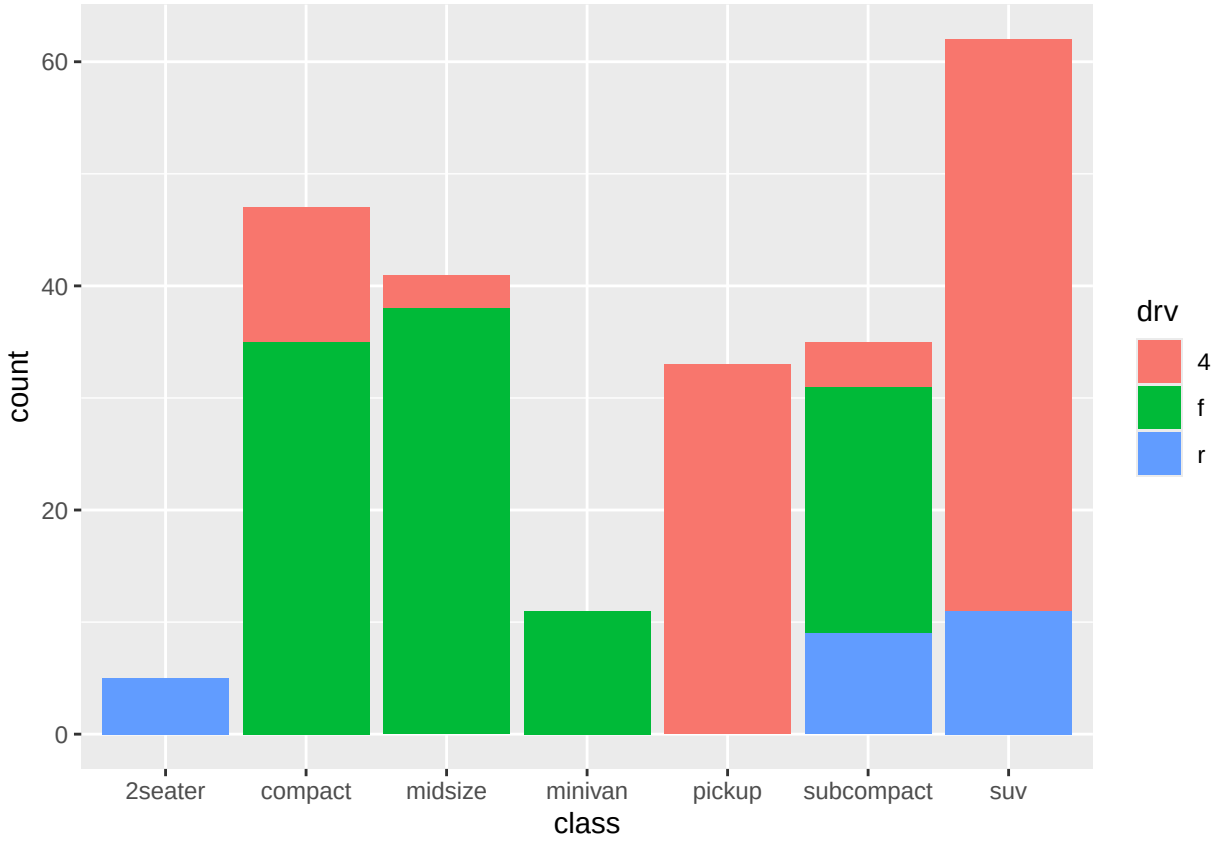
Interpretation: The dot plot shows the proportion of participants by age, with each dot representing a specific proportion. The highest concentration of participants is around age 20, with additional notable clusters around ages 30, 40, and 55. The distribution indicates more participants in the younger age groups, with fewer participants as age increases.

Bivariate Graphs

Categorical vs. Categorical

stacked bar chart

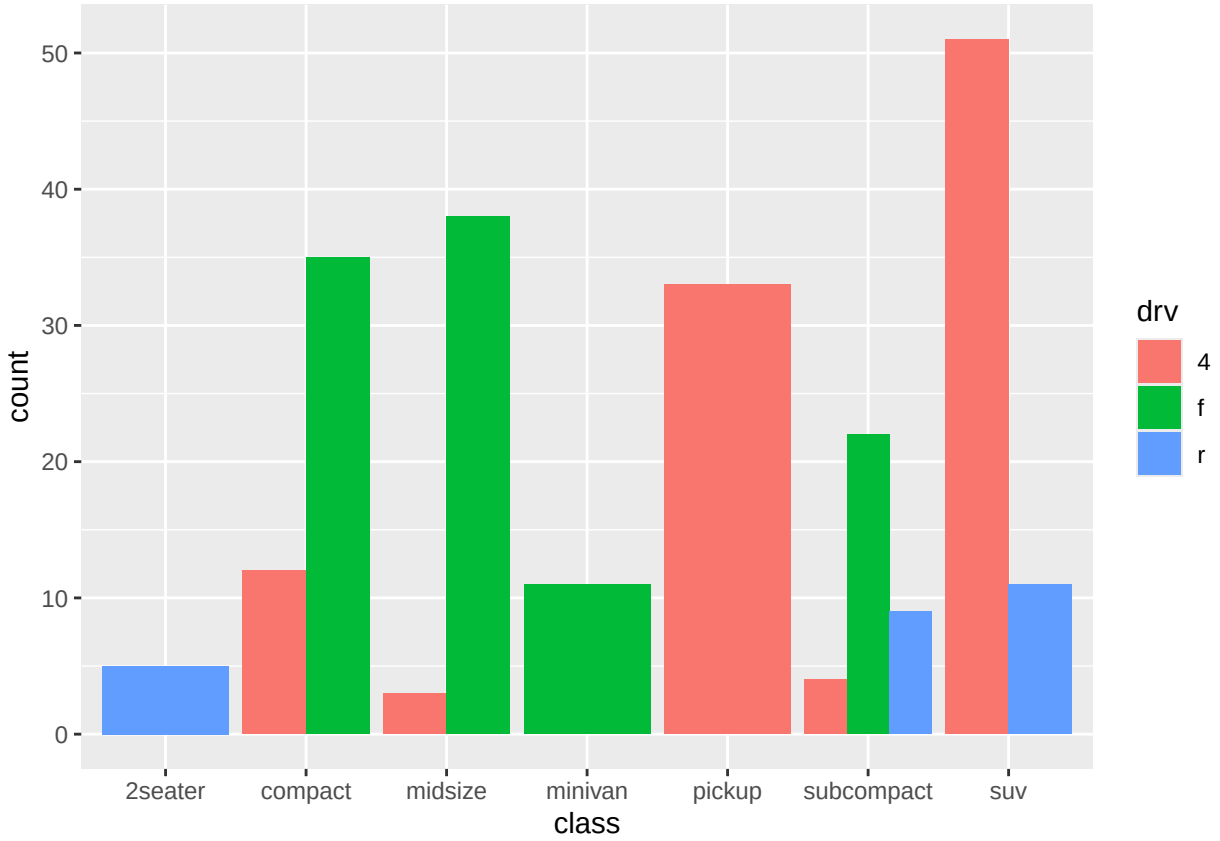
```
ggplot(mpg, aes(x = class, fill = drv)) +
  geom_bar(position = "stack")
```



Interpretation: This bar chart shows that the most common vehicle is the SUV. All 2seater cars are rear wheel drive and pickup trucks are 4-wheel, while most, but not all SUVs are 4-wheel drive and we can see all minivan are real wheel drive where as subcompact can be found in all 3 varient.

Grouped bar chart

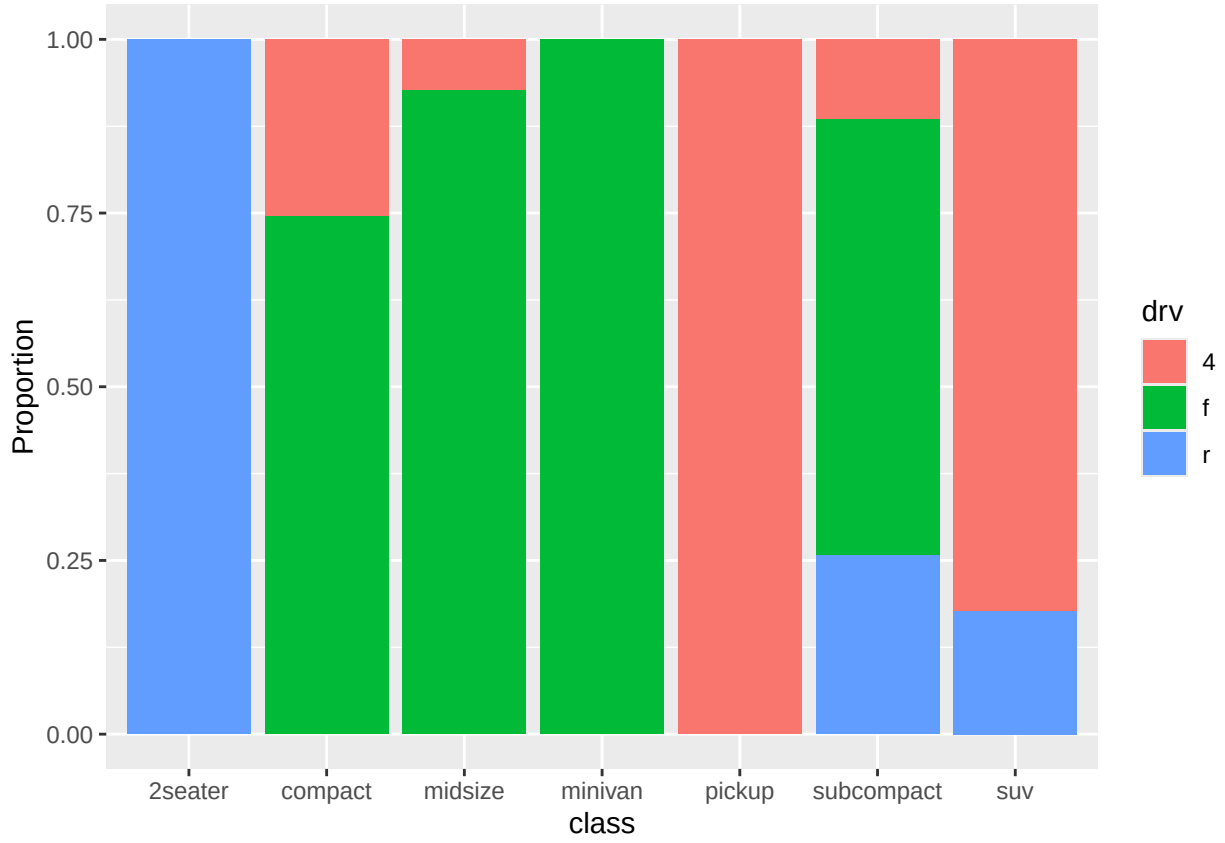
```
ggplot(mpg, aes(x = class, fill = drv)) +
  geom_bar(position = "dodge")
```



Interpretation: This bar chart shoes that all Minivans are front-wheel drive and 2-seater are rear wheel drive . By default, zero count bars are dropped and the remaining bars are made wider.

Segmented bar chart

```
ggplot(mpg, aes(x = class, fill = drv)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion")
```



Improving the color and labeling

```
plotdata <- mpg %>%
  group_by(class, drv) %>%
  summarize(n = n()) %>%
  mutate(pct = n/sum(n),
         lbl = scales::percent(pct))
```

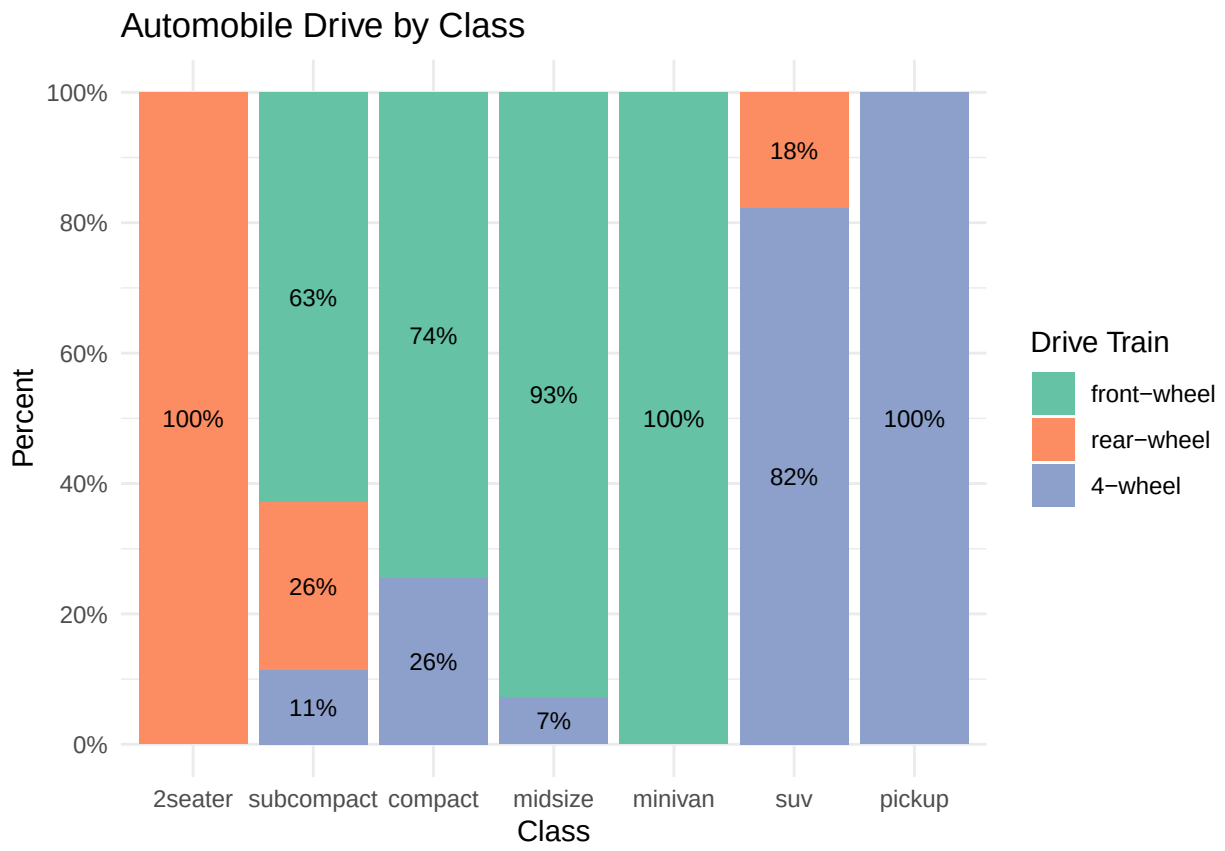
```
## `summarise()` has grouped output by 'class'. You can override using the
## `.groups` argument.
```

```
plotdata
```

```
## # A tibble: 12 x 5
## # Groups:   class [7]
##    class      drv       n    pct lbl
##    <chr>      <chr> <int>  <dbl> <chr>
##  1 2seater    r         5 1      100%
##  2 compact    4        12 0.255  26%
##  3 compact    f        35 0.745  74%
##  4 midsize    4         3 0.0732 7%
##  5 midsize    f        38 0.927  93%
##  6 minivan    f        11 1      100%
```

```
##  7 pickup     4        33 1      100%
##  8 subcompact 4         4 0.114  11%
##  9 subcompact f        22 0.629  63%
## 10 subcompact r         9 0.257  26%
## 11 suv        4        51 0.823  82%
## 12 suv        r        11 0.177  18%
```

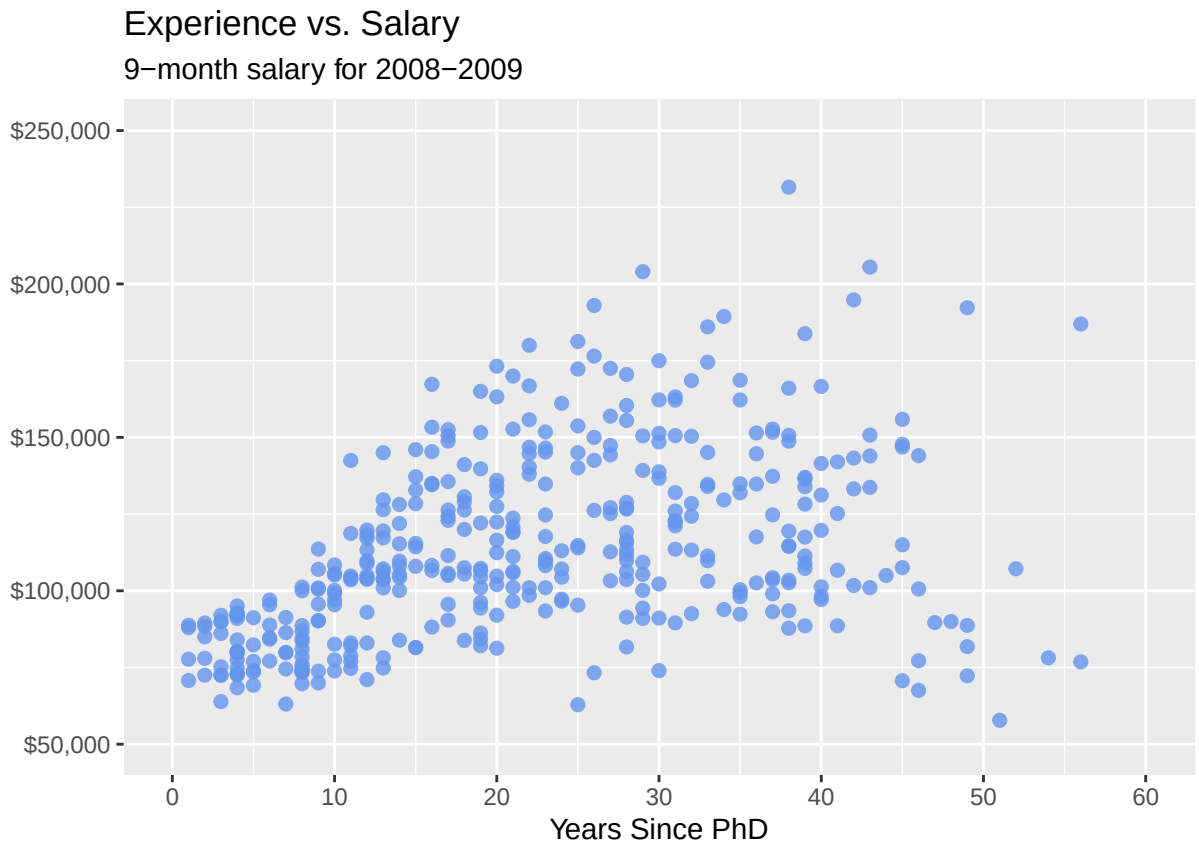
```
ggplot(plotdata,
       aes(x = factor(class,
                      levels = c("2seater", "subcompact",
                                 "compact", "midsize",
                                 "minivan", "suv", "pickup")),
           y = pct,
          fill = factor(drv,
                        levels = c("f", "r", "4"),
                        labels = c("front-wheel",
                                   "rear-wheel",
                                   "4-wheel")))) +
  geom_bar(stat = "identity",
           position = "fill") +
  scale_y_continuous(breaks = seq(0, 1, .2),
                     label = percent) +
  geom_text(aes(label = lbl),
            size = 3,
            position = position_stack(vjust = 0.5)) +
  scale_fill_brewer(palette = "Set2") +
  labs(y = "Percent",
       fill="Drive Train",
       x = "Class",
       title = "Automobile Drive by Class") +
  theme_minimal()
```



Quantitative vs. Quantitative

Scatterplot

```
data(Salaries, package="carData")
ggplot(Salaries,
       aes(x = yrs.since.phd, y = salary)) +
  geom_point(color="cornflowerblue",
             size = 2,
             alpha=.8) +
  scale_y_continuous(label = scales::dollar,
                     limits = c(50000, 250000)) +
  scale_x_continuous(breaks = seq(0, 60, 10),
                     limits=c(0, 60)) +
  labs(x = "Years Since PhD",
       y = "",
       title = "Experience vs. Salary",
       subtitle = "9-month salary for 2008-2009")
```



Interpretation: This scatter plit illustrating the relationship between years since obtaining a PhD and 9-month salary for 2008-2009. It shows that salaries generally increase with experience, but there is substantial variation at each experience level, with more individuals having fewer years of experience and lower salaries.

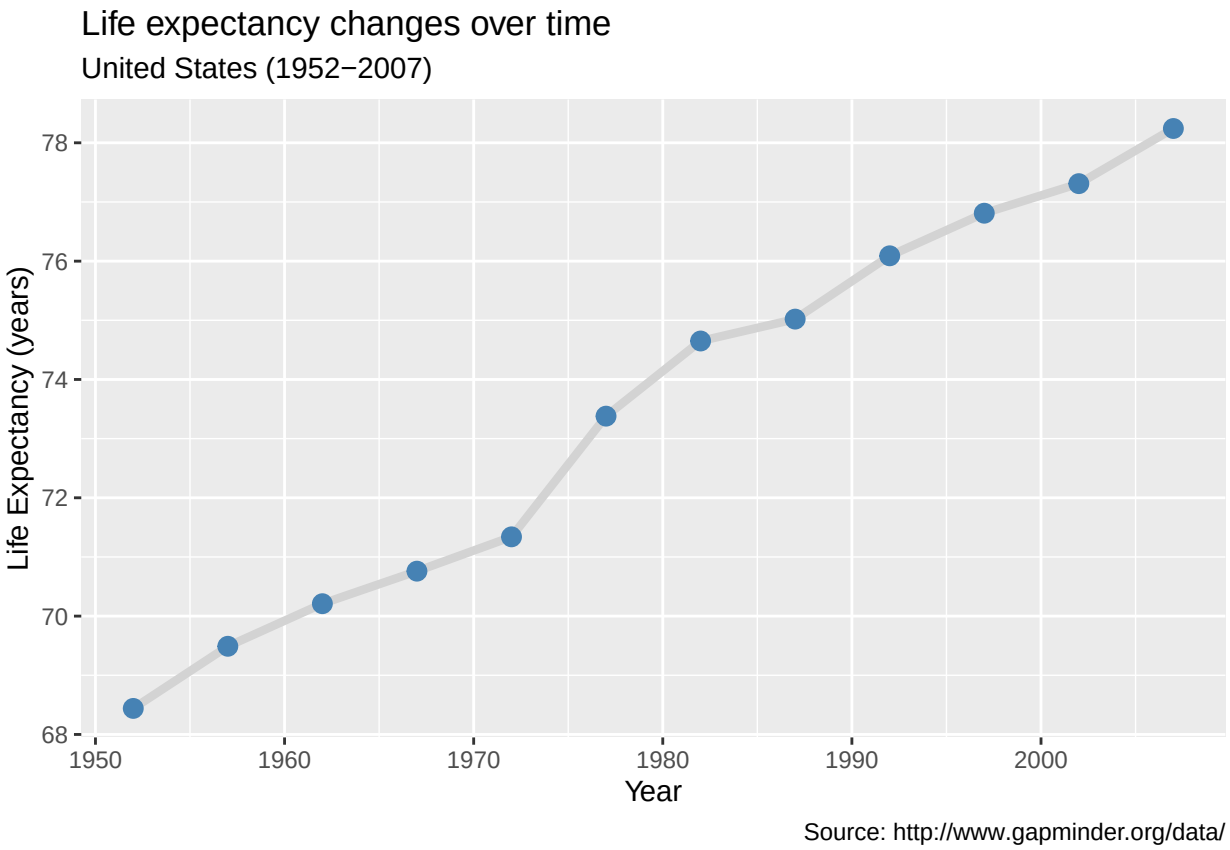
Line plot

```
data(gapminder, package="gapminder")
plotdata <- filter(gapminder, country == "United States")

ggplot(plotdata, aes(x = year, y = lifeExp)) +
  geom_line(size = 1.5,
            color = "lightgrey") +
  geom_point(size = 3,
             color = "steelblue") +
  labs(y = "Life Expectancy (years)",
       x = "Year",
       title = "Life expectancy changes over time",
       subtitle = "United States (1952-2007)",
       caption = "Source: http://www.gapminder.org/data/")
```

```
## 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.
```



Interpretation: This line chart shoes the e United States from 1952 to 2007, displaying life expectancy over time. The chart shows a consistent increase in life expectancy from just below 70 years in 1952 to around 78 years in 2007, indicating improvements in healthcare, nutrition, and living conditions.

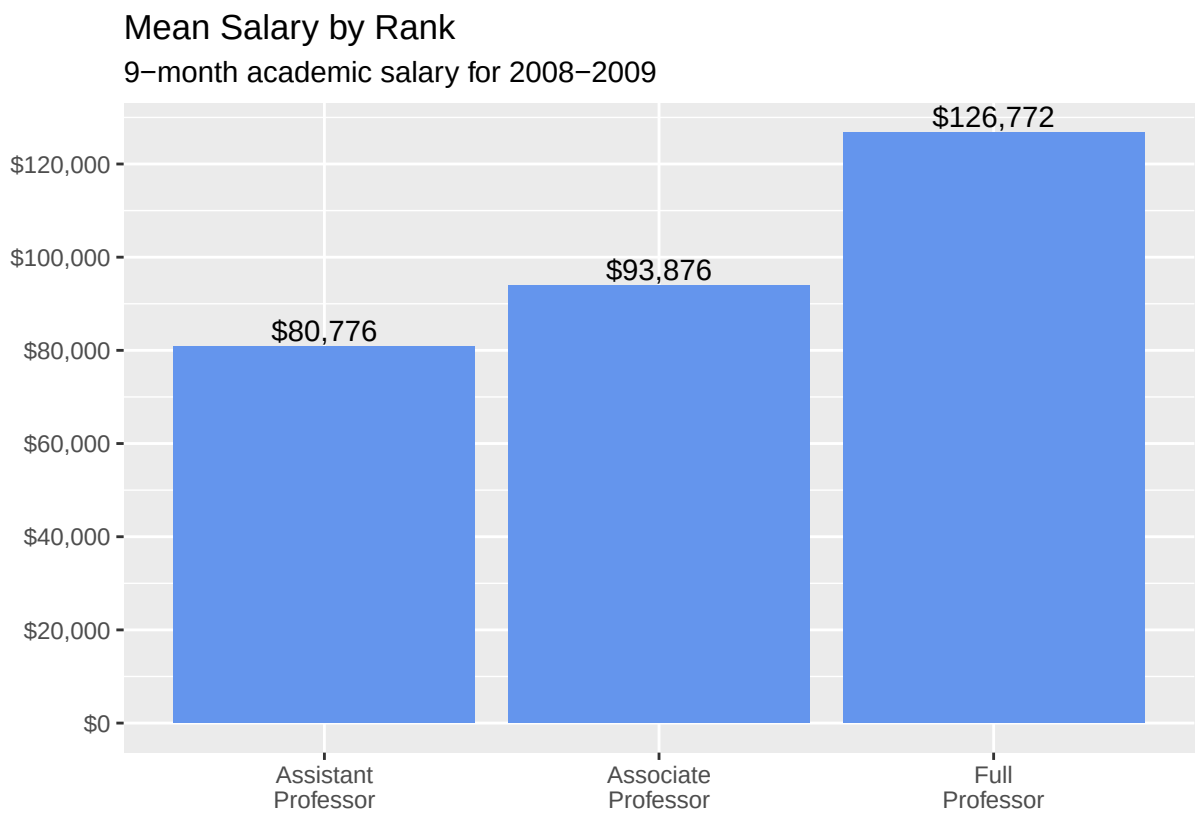
Categorical vs. Quantitative

Bar chart (on summary statistics)

```
data(Salaries, package="carData")
plotdata <- Salaries %>%
  group_by(rank) %>%
  summarize(mean_salary = mean(salary))

ggplot(plotdata,
       aes(x = factor(rank,
                      labels = c("Assistant\nProfessor",
                                 "Associate\nProfessor",
                                 "Full\nProfessor")),
                      y = mean_salary)) +
  geom_bar(stat = "identity",
           fill = "cornflowerblue") +
  geom_text(aes(label = dollar(mean_salary)),
```

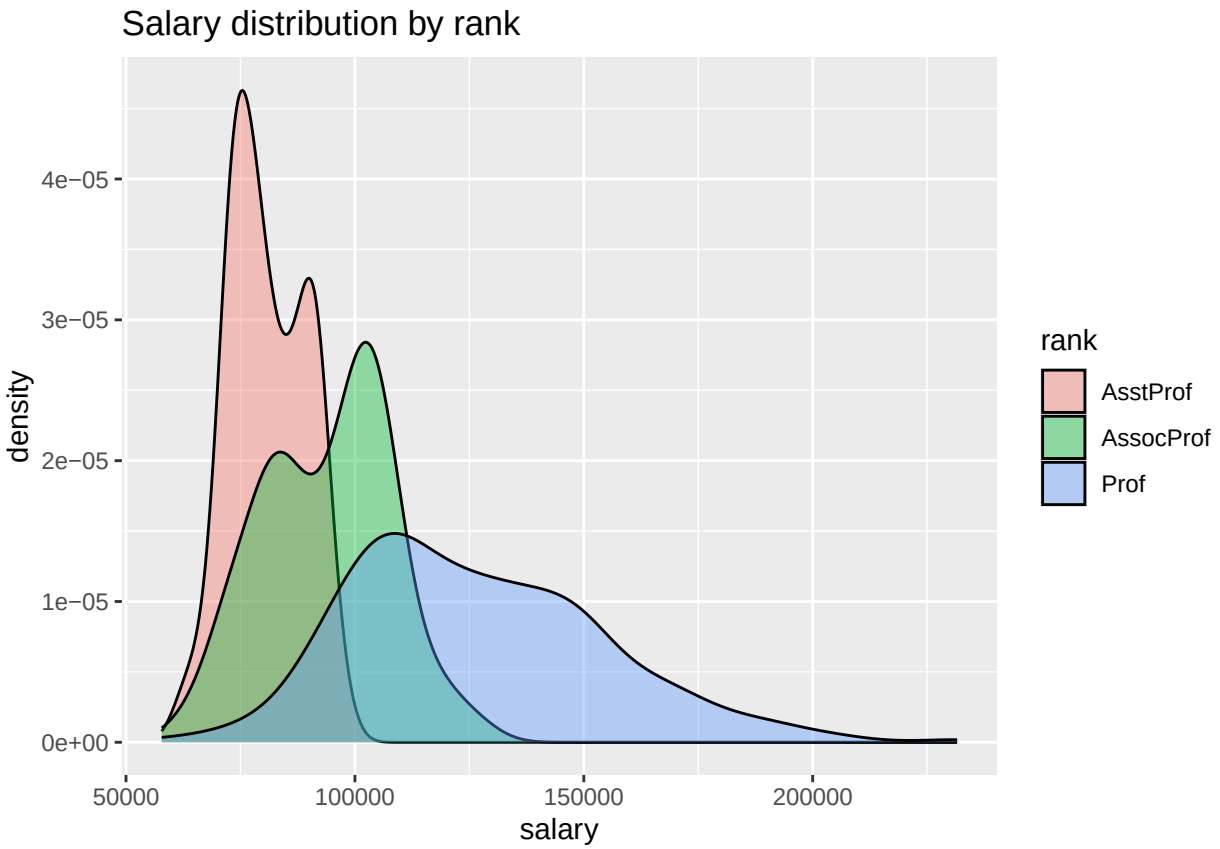
```
            vjust = -0.25) +
  scale_y_continuous(breaks = seq(0, 130000, 20000),
                     label = dollar) +
  labs(title = "Mean Salary by Rank",
       subtitle = "9-month academic salary for 2008-2009",
       x = "",
       y = "")
```



Interpretation: This bar graph shows the mean salaries for different academic ranks (Assistant, Associate, and Full Professor) for the 2008-2009 academic year. The chart reveals a clear upward trend in mean salary with higher academic rank, with the most significant salary increase occurring between Associate Professors and Full Professors

Grouped kernel density plots

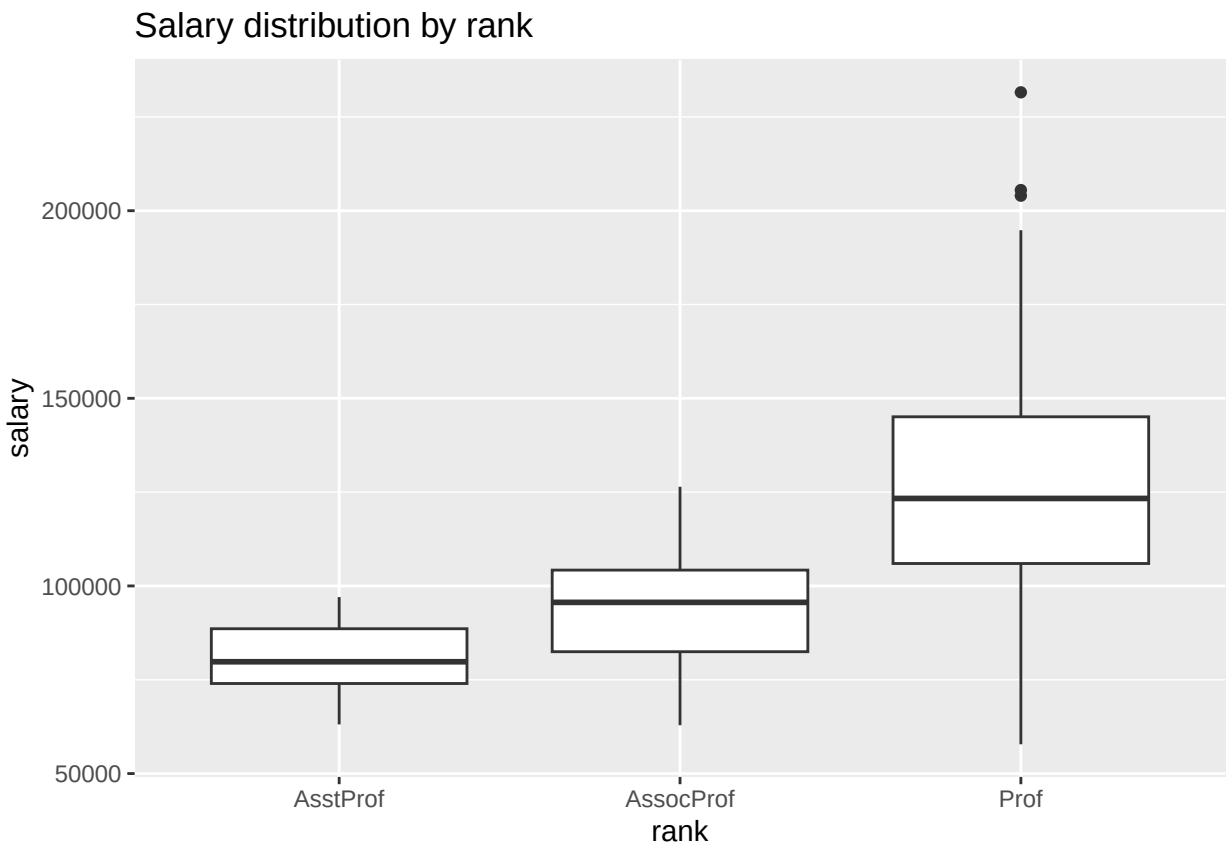
```
ggplot(Salaries, aes(x = salary, fill = rank)) +
  geom_density(alpha = 0.4) +
  labs(title = "Salary distribution by rank")
```



Interpretation: This density plot illustrating the distribution of salaries for Assistant Professors, Associate Professors, and Full Professors. The plot shows how salaries vary across these ranks, with each rank having a distinct salary distribution

Box plots

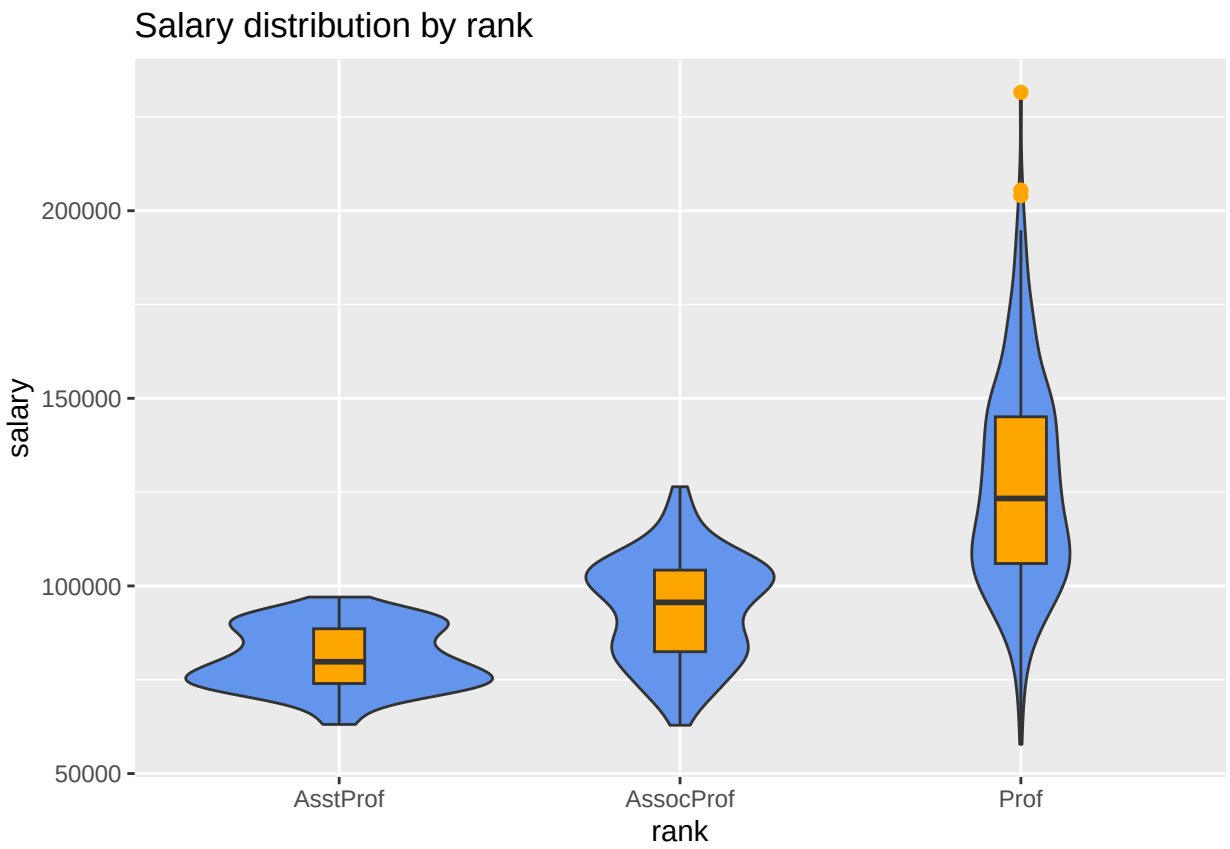
```
# plot the distribution of salaries by rank using boxplots
ggplot(Salaries, aes(x = rank, y = salary)) +
  geom_boxplot() +
  labs(title = "Salary distribution by rank")
```



Interpretation: This box plot shows that salaries increase with rank: Assistant Professors have the lowest median and least variation, Associate Professors earn more with more variability, and Professors have the highest median with the most variability and outliers.

Violin plots

```
ggplot(Salaries, aes(x = rank, y = salary)) +
  geom_violin(fill = "cornflowerblue") +
  geom_boxplot(width = .15,
               fill = "orange",
               outlier.color = "orange",
               outlier.size = 2) +
  labs(title = "Salary distribution by rank")
```

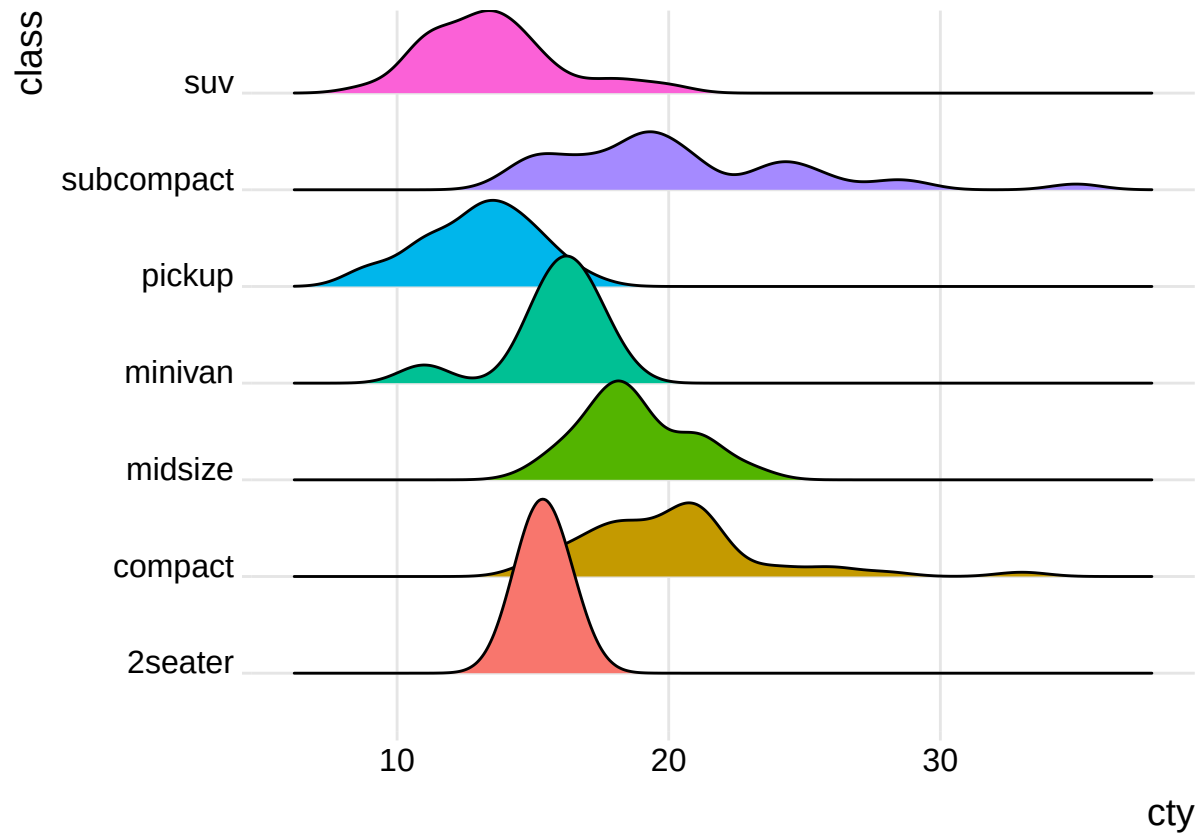


Ridgeline plots

```
library(ggridges)

ggplot(mpg,
       aes(x = cty, y = class, fill = class)) +
  geom_density_ridges() +
  theme_ridges() +
  labs("Highway mileage by auto class") +
  theme(legend.position = "none")
```

```
## Picking joint bandwidth of 0.929
```

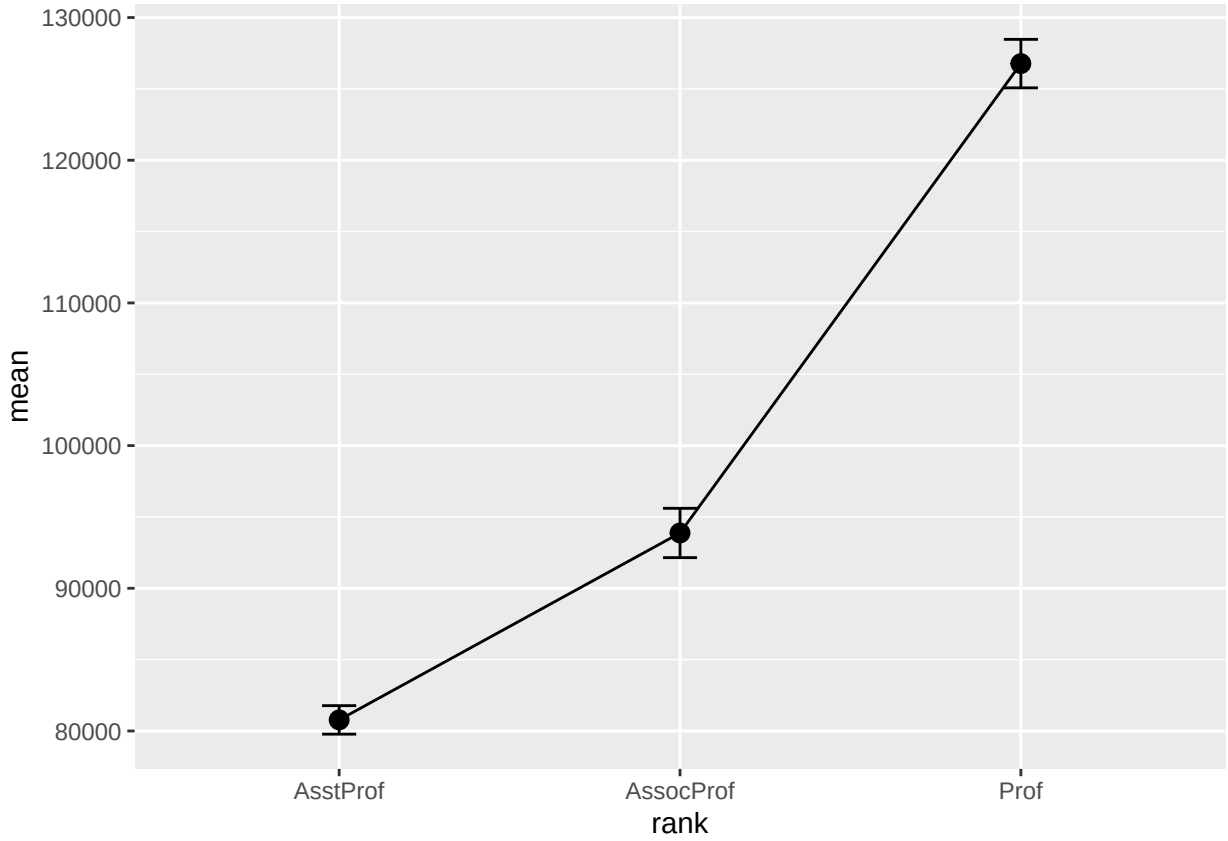


Interpretation:This ridge chart shows the density distributions of city mileage (cty) across different car classes. The distributions indicate how city mileage varies among SUVs, subcompacts, pickups, minivans, midsize cars, compacts, and 2-seaters, with each car class having a distinct mileage pattern.

Mean/SEM plots

```
plotdata <- Salaries %>%
  group_by(rank) %>%
  summarize(n = n(),
          mean = mean(salary),
          sd = sd(salary),
          se = sd / sqrt(n),
          ci = qt(0.975, df = n - 1) * sd / sqrt(n))

ggplot(plotdata,
       aes(x = rank,
           y = mean,
           group = 1)) +
  geom_point(size = 3) +
  geom_line() +
  geom_errorbar(aes(ymin = mean - se,
                    ymax = mean + se),
                width = .1)
```

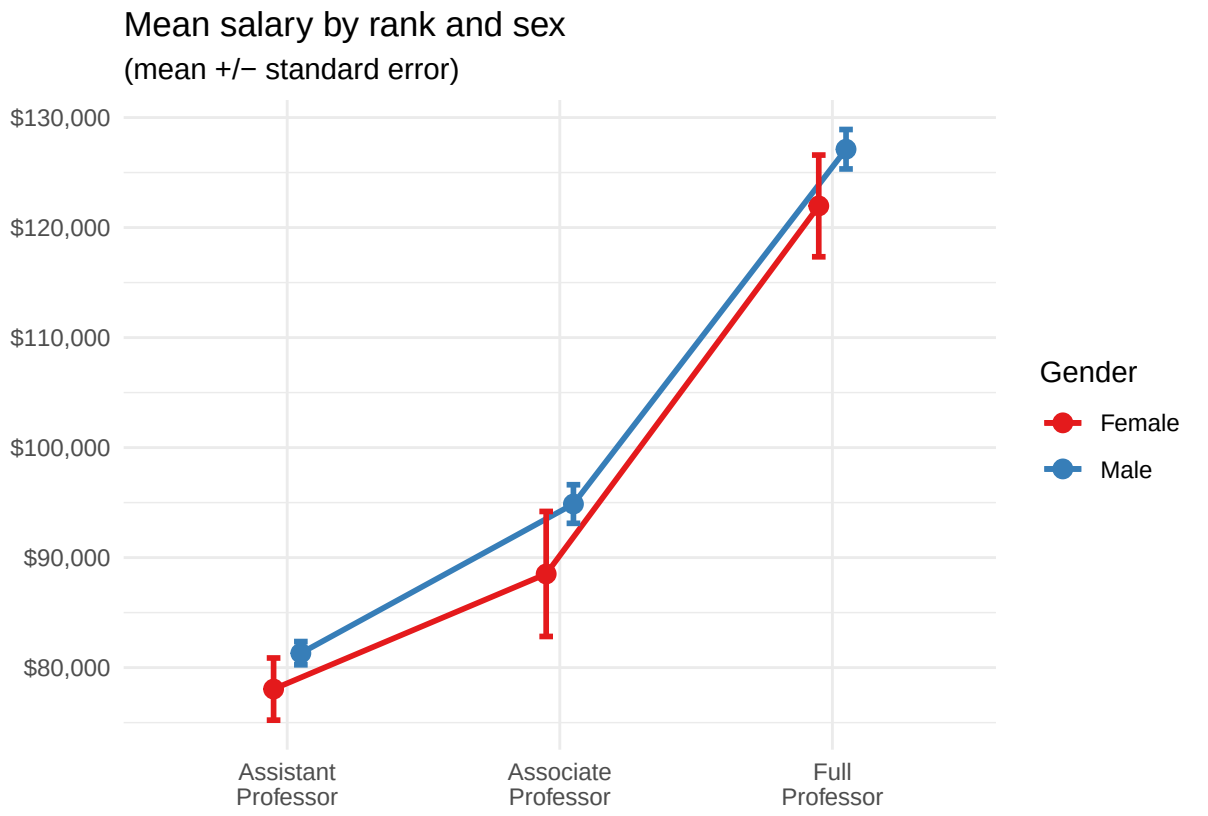



```
plotdata <- Salaries %>%
  group_by(rank, sex) %>%
  summarize(n = n(),
            mean = mean(salary),
            sd = sd(salary),
            se = sd/sqrt(n))
```

```
## `summarise()` has grouped output by 'rank'. You can override using the
## `.groups` argument.
```

```
pd <- position_dodge(0.2)
ggplot(plotdata,
       aes(x = factor(rank,
                      labels = c("Assistant\nProfessor",
                                 "Associate\nProfessor",
                                 "Full\nProfessor")),
                      y = mean, group=sex, color=sex)) +
  geom_point(position=pd,
             size=3) +
  geom_line(position=pd,
            size = 1) +
  geom_errorbar(aes(ymin = mean - se,
                    ymax = mean + se),
                width = .1,
                position=pd,
```

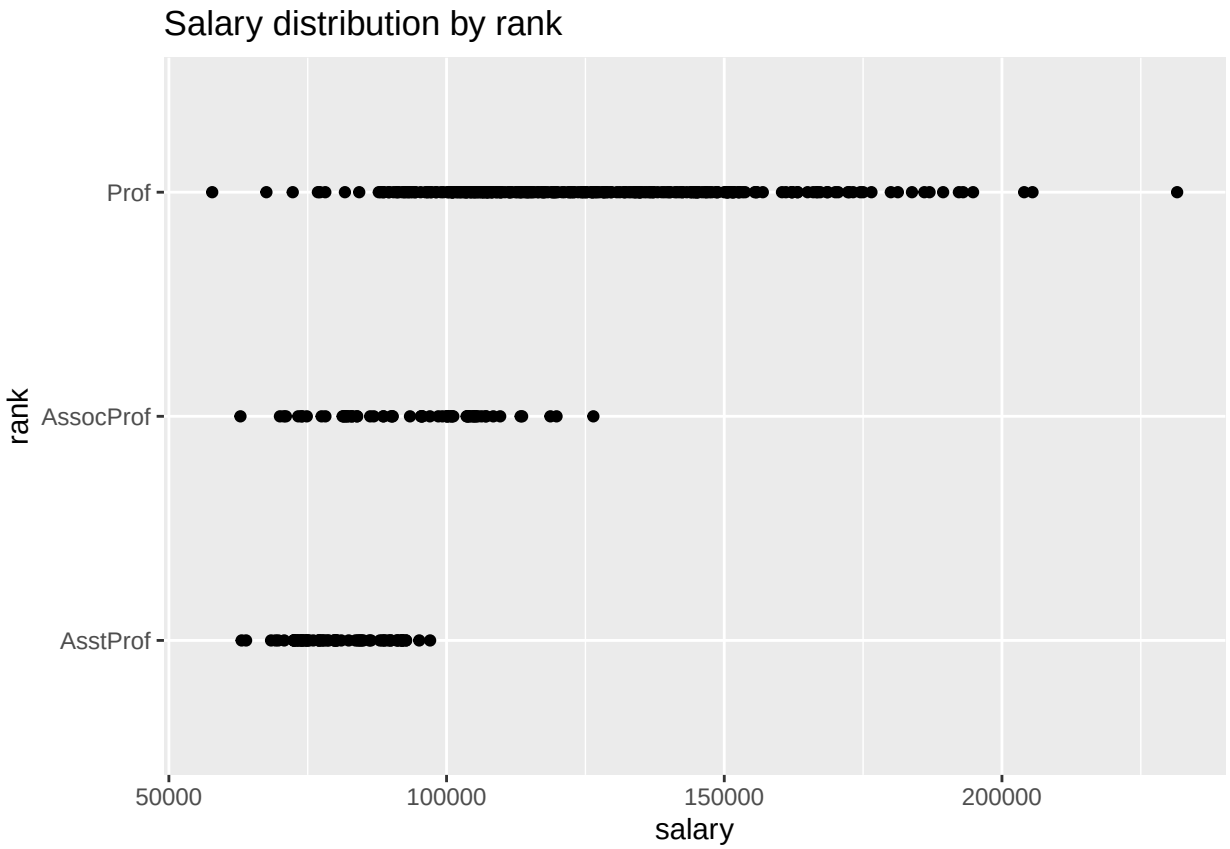
```
                size=1) +
  scale_y_continuous(label = scales::dollar) +
  scale_color_brewer(palette="Set1") +
  theme_minimal() +
  labs(title = "Mean salary by rank and sex",
       subtitle = "(mean +/- standard error)",
       x = "",
       y = "",
       color = "Gender")
```



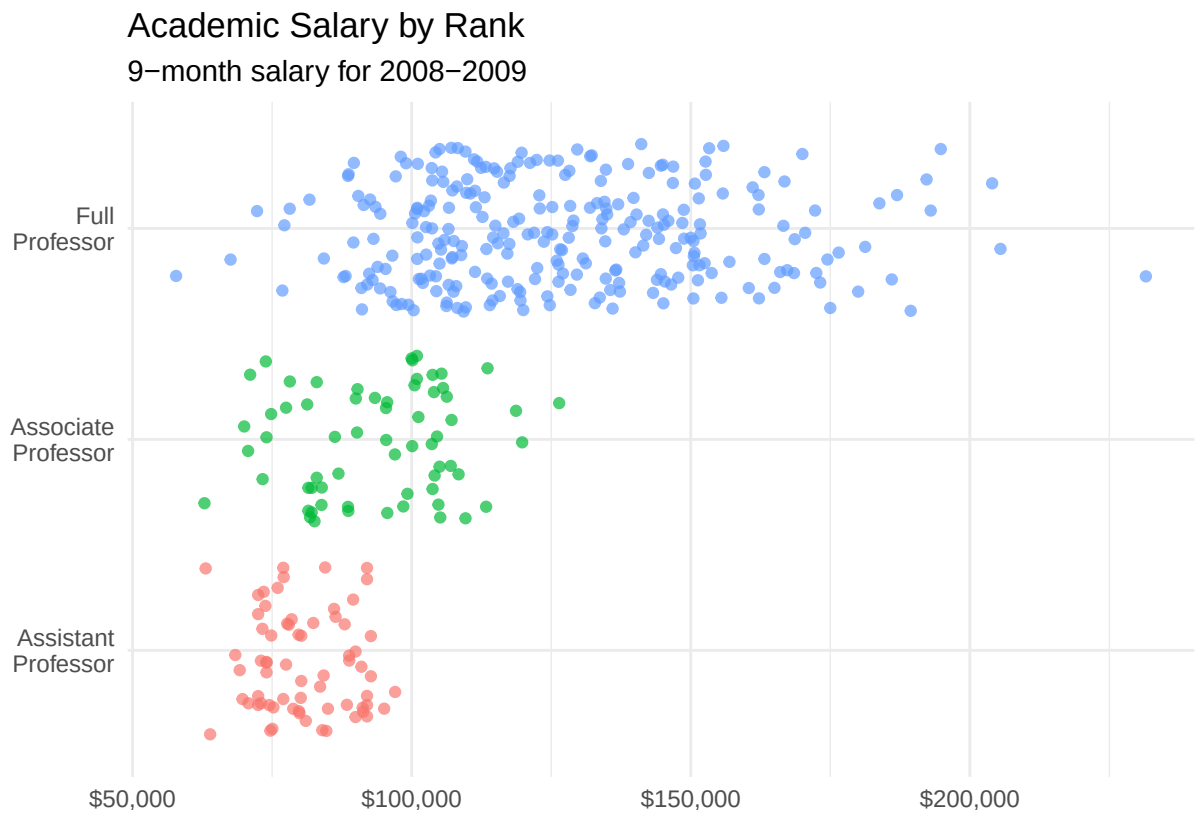
Interpretation: This chart shows the mean salary by academic rank and gender, with error bars representing standard error. For each rank (Assistant Professor, Associate Professor, and Full Professor), males generally earn slightly more than females. Salaries increase with higher academic ranks for both genders, with the gap being most noticeable at the Full Professor level.

Strip plots

```
ggplot(Salaries, aes(y = rank, x = salary)) +
  geom_point() +
  labs(title = "Salary distribution by rank")
```




```
# plot the distribution of salaries
# by rank using jittering
library(scales)
ggplot(Salaries,
       aes(y = factor(rank,
                      labels = c("Assistant\nProfessor",
                                 "Associate\nProfessor",
                                 "Full\nProfessor")),
           x = salary, color = rank)) +
  geom_jitter(alpha = 0.7) +
  scale_x_continuous(label = dollar) +
  labs(title = "Academic Salary by Rank",
       subtitle = "9-month salary for 2008-2009",
       x = "",
       y = "") +
  theme_minimal() +
  theme(legend.position = "none")
```



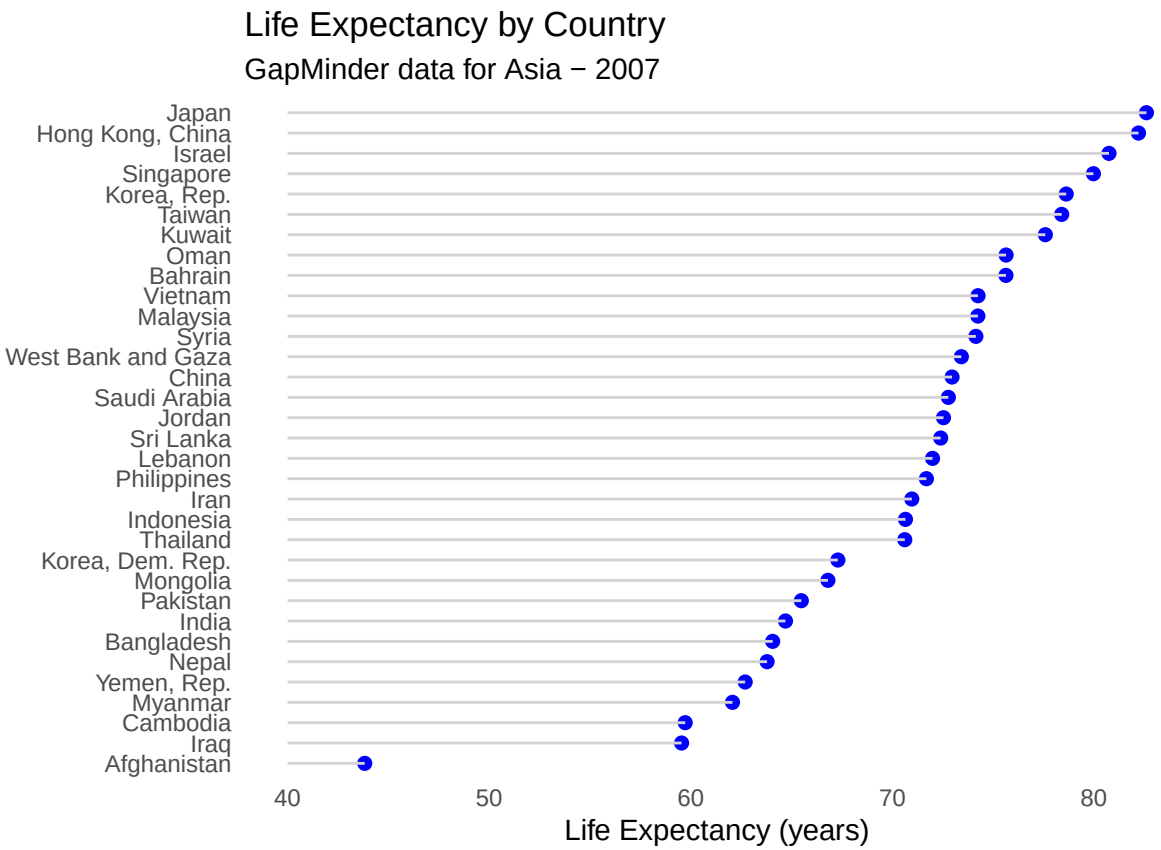
Interpretation: This jitter plot shows the academic salaries by rank for a 9-month period in 2008-2009. Full professors generally earn between $100,000 and over $200,000, associate professors between $75,000 and $150,000, and assistant professors between $50,000 and $100,000. There is a clear upward trend in salary with higher academic rank.

Cleveland Dot Charts

```
data(gapminder, package="gapminder")
plotdata <- gapminder %>%
  filter(continent == "Asia" &
           year == 2007)

ggplot(plotdata, aes(x=lifeExp,
                     y=reorder(country, lifeExp))) +
  geom_point(color="blue", size = 2) +
  geom_segment(aes(x = 40,
                   xend = lifeExp,
                   y = reorder(country, lifeExp),
                   yend = reorder(country, lifeExp)),
                   color = "lightgrey") +
  labs (x = "Life Expectancy (years)",
        y = "",
        title = "Life Expectancy by Country",
        subtitle = "GapMinder data for Asia - 2007") +
  theme_minimal() +
```

```
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



Interpretation: This chart shows life expectancy by country in Asia for the year 2007, according to GapMinder data. Japan and Hong Kong have the highest life expectancy, exceeding 80 years, while Afghanistan has the lowest, below 45 years. The data reflects significant variation in life expectancy across the region.