Seminar_1

Ingrid Canelles Campas

Install and load data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(datasauRus)
```

Have a look at the data

```
datasaurus_dozen_wide
```

```
# A tibble: 142 x 26
   away_x away_y bullseye_x bullseye_y circle_x circle_y dino_x dino_y dots_x
    <dbl>  <dbl>      <dbl>      <dbl>    <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
 1   32.3   61.4       51.2       83.3     56.0     79.3   55.4   97.2   51.1
 2   53.4   26.2       59.0       85.5     50.0     79.0   51.5   96.0   50.5
 3   63.9   30.8       51.9       85.8     51.3     82.4   46.2   94.5   50.2
```

```
 4   70.3   82.5       48.2       85.0     51.2     79.2   42.8   91.4   50.1
 5   34.1   45.7       41.7       84.0     44.4     78.2   40.8   88.3   50.6
 6   67.7   37.1       37.9       82.6     45.0     77.9   38.7   84.9   50.3
 7   53.3   97.5       39.5       80.8     48.6     78.8   35.6   79.9   25.6
 8   63.5   25.1       39.6       82.7     42.1     76.9   33.1   77.6   25.5
 9   68.0   81.0       34.8       80.0     41.0     76.4   29.0   74.5   25.4
10   67.4   29.7       27.6       72.8     34.6     72.7   26.2   71.4   25.6
# i 132 more rows
# i 17 more variables: dots_y <dbl>, h_lines_x <dbl>, h_lines_y <dbl>,
#   high_lines_x <dbl>, high_lines_y <dbl>, slant_down_x <dbl>,
#   slant_down_y <dbl>, slant_up_x <dbl>, slant_up_y <dbl>, star_x <dbl>,
#   star_y <dbl>, v_lines_x <dbl>, v_lines_y <dbl>, wide_lines_x <dbl>,
#   wide_lines_y <dbl>, x_shape_x <dbl>, x_shape_y <dbl>
```

```
data("datasaurus_dozen_wide")
datasauRus::twelve_from_slant_wide
```

```
# A tibble: 182 x 24
   bullseye_x bullseye_y circle_x circle_y dots_x dots_y h_lines_x h_lines_y
        <dbl>      <dbl>    <dbl>    <dbl>  <dbl>  <dbl>     <dbl>     <dbl>
 1       66.3       84.2     57.6     76.2   73.5   82.5      66.1     89.8
 2       73.5       47.4     82.4     43.2   72.5   50.0      85.9     49.6
 3       34.9       49.2     27.2     37.1   27.6   48.6      26.0     49.6
 4       56.9       28.8     51.7     19.4   49.5   17.2      49.2     29.6
 5       68.4       60.8     62.9     75.2   52.2   48.7      51.0     49.6
 6       57.0       86.1     61.7     75.5   51.3   82.2      65.4     89.7
 7       39.9       12.1     42.1     21.6   49.9   17.3      38.7      9.56
 8       70.9       13.1     64.9     21.3   72.4   19.5      74.2     29.9
 9       42.8       11.2     44.1     20.9   26.4   17.7      35.4      9.63
10       61.5       30.0     60.5     20.0   50.0   17.0      55.9     29.6
# i 172 more rows
# i 16 more variables: high_lines_x <dbl>, high_lines_y <dbl>, slant_x <dbl>,
#   slant_y <dbl>, slant_down_x <dbl>, slant_down_y <dbl>, slant_up_x <dbl>,
#   slant_up_y <dbl>, star_x <dbl>, star_y <dbl>, v_lines_x <dbl>,
#   v_lines_y <dbl>, wide_lines_x <dbl>, wide_lines_y <dbl>, x_shape_x <dbl>,
#   x_shape_y <dbl>
```

Obtain info

```
?datasaurus_dozen_wide
```

Part 1

```
summary(datasaurus_dozen_wide)
```

```
     away_x          away_y           bullseye_x      bullseye_y
 Min.   :15.56   Min.   : 0.01512   Min.   :19.29   Min.   : 9.692
 1st Qu.:39.72   1st Qu.:24.62589   1st Qu.:41.63   1st Qu.:26.245
 Median :53.34   Median :47.53527   Median :53.84   Median :47.383
 Mean   :54.27   Mean   :47.83472   Mean   :54.27   Mean   :47.831
 3rd Qu.:69.15   3rd Qu.:71.80315   3rd Qu.:64.80   3rd Qu.:72.533
 Max.   :91.64   Max.   :97.47577   Max.   :91.74   Max.   :85.876
    circle_x        circle_y         dino_x          dino_y
 Min.   :21.86   Min.   :16.33   Min.   :22.31   Min.   : 2.949
 1st Qu.:43.38   1st Qu.:18.35   1st Qu.:44.10   1st Qu.:25.288
 Median :54.02   Median :51.03   Median :53.33   Median :46.026
 Mean   :54.27   Mean   :47.84   Mean   :54.26   Mean   :47.832
 3rd Qu.:64.97   3rd Qu.:77.78   3rd Qu.:64.74   3rd Qu.:68.526
 Max.   :85.66   Max.   :85.58   Max.   :98.21   Max.   :99.487
     dots_x          dots_y        h_lines_x       h_lines_y
 Min.   :25.44   Min.   :15.77   Min.   :22.00   Min.   :10.46
 1st Qu.:50.36   1st Qu.:17.11   1st Qu.:42.29   1st Qu.:30.48
 Median :50.98   Median :51.30   Median :53.07   Median :50.47
 Mean   :54.26   Mean   :47.84   Mean   :54.26   Mean   :47.83
 3rd Qu.:75.20   3rd Qu.:82.88   3rd Qu.:66.77   3rd Qu.:70.35
 Max.   :77.95   Max.   :94.25   Max.   :98.29   Max.   :90.46
  high_lines_x    high_lines_y    slant_down_x    slant_down_y
 Min.   :17.89   Min.   :14.91   Min.   :18.11   Min.   : 0.3039
 1st Qu.:41.54   1st Qu.:22.92   1st Qu.:42.89   1st Qu.:27.8409
 Median :54.17   Median :32.50   Median :53.14   Median :46.4013
 Mean   :54.27   Mean   :47.84   Mean   :54.27   Mean   :47.8359
 3rd Qu.:63.95   3rd Qu.:75.94   3rd Qu.:64.47   3rd Qu.:68.4394
 Max.   :96.08   Max.   :87.15   Max.   :95.59   Max.   :99.6442
   slant_up_x      slant_up_y         star_x          star_y
 Min.   :20.21   Min.   : 5.646   Min.   :27.02   Min.   :14.37
 1st Qu.:42.81   1st Qu.:24.756   1st Qu.:41.03   1st Qu.:20.37
 Median :54.26   Median :45.292   Median :56.53   Median :50.11
 Mean   :54.27   Mean   :47.831   Mean   :54.27   Mean   :47.84
```

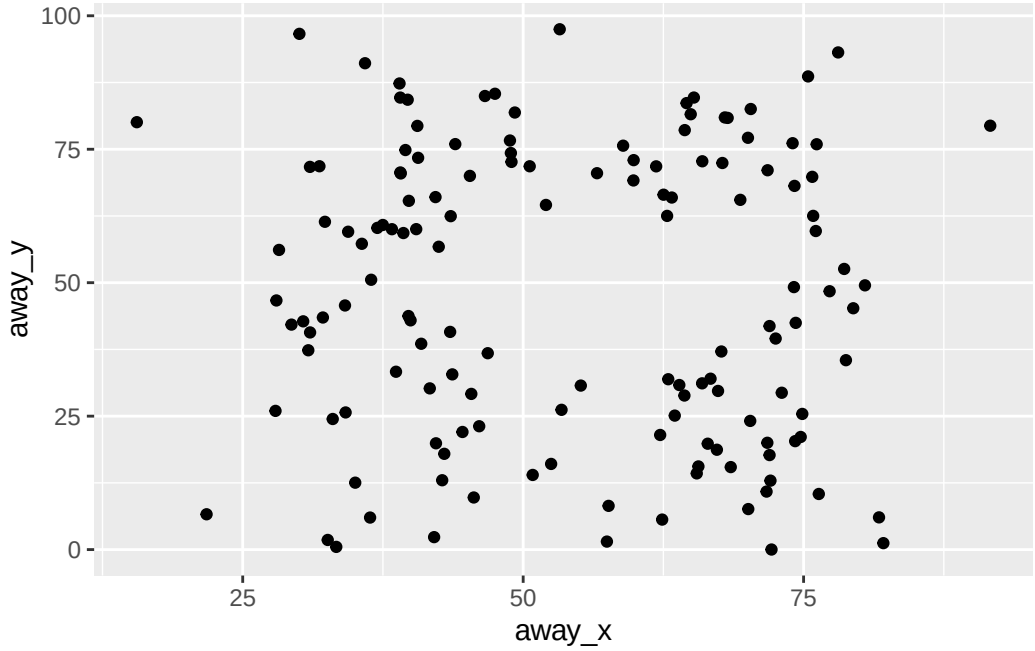
```
 3rd Qu.:64.49   3rd Qu.:70.856   3rd Qu.:68.71   3rd Qu.:63.55
 Max.   :95.26   Max.   :99.580   Max.   :86.44   Max.   :92.21
   v_lines_x       v_lines_y       wide_lines_x    wide_lines_y
 Min.   :30.45   Min.   : 2.735   Min.   :27.44   Min.   : 0.217
 1st Qu.:49.96   1st Qu.:22.753   1st Qu.:35.52   1st Qu.:24.347
 Median :50.36   Median :47.114   Median :64.55   Median :46.279
 Mean   :54.27   Mean   :47.837   Mean   :54.27   Mean   :47.832
 3rd Qu.:69.50   3rd Qu.:65.845   3rd Qu.:67.45   3rd Qu.:67.568
 Max.   :89.50   Max.   :99.695   Max.   :77.92   Max.   :99.284
   x_shape_x       x_shape_y
 Min.   :31.11   Min.   : 4.578
 1st Qu.:40.09   1st Qu.:23.471
 Median :47.14   Median :39.876
 Mean   :54.26   Mean   :47.840
 3rd Qu.:71.86   3rd Qu.:73.610
 Max.   :85.45   Max.   :97.838
```

```
glimpse(datasaurus_dozen_wide)
```

```
Rows: 142
Columns: 26
$ away_x       <dbl> 32.33111, 53.42146, 63.92020, 70.28951, 34.11883, 67.6707~
$ away_y       <dbl> 61.411101, 26.186880, 30.832194, 82.533649, 45.734551, 37~
$ bullseye_x   <dbl> 51.20389, 58.97447, 51.87207, 48.17993, 41.68320, 37.8904~
$ bullseye_y   <dbl> 83.33978, 85.49982, 85.82974, 85.04512, 84.01794, 82.5674~
$ circle_x     <dbl> 55.99303, 50.03225, 51.28846, 51.17054, 44.37791, 45.0102~
$ circle_y     <dbl> 79.27726, 79.01307, 82.43594, 79.16529, 78.16463, 77.8808~
$ dino_x       <dbl> 55.3846, 51.5385, 46.1538, 42.8205, 40.7692, 38.7179, 35.~
$ dino_y       <dbl> 97.1795, 96.0256, 94.4872, 91.4103, 88.3333, 84.8718, 79.~
$ dots_x       <dbl> 51.14792, 50.51713, 50.20748, 50.06948, 50.56285, 50.2885~
$ dots_y       <dbl> 90.86741, 89.10239, 85.46005, 83.05767, 82.93782, 82.9752~
$ h_lines_x    <dbl> 53.36657, 52.80198, 47.05413, 42.44843, 42.70404, 32.3789~
$ h_lines_y    <dbl> 90.20803, 90.08806, 90.45894, 89.50770, 90.44263, 90.1441~
$ high_lines_x <dbl> 57.61323, 51.27439, 50.75390, 37.02118, 42.88176, 37.1557~
$ high_lines_y <dbl> 83.90517, 82.81798, 76.75413, 81.95447, 80.18477, 84.9541~
$ slant_down_x <dbl> 52.87202, 59.01414, 56.37511, 37.83920, 39.88537, 44.0774~
$ slant_down_y <dbl> 97.34322, 93.57487, 96.30515, 94.35944, 90.63466, 84.1258~
$ slant_up_x   <dbl> 47.69520, 44.60998, 43.85638, 41.57893, 49.17742, 42.6522~
$ slant_up_y   <dbl> 95.24119, 93.07584, 94.08587, 90.30357, 96.61053, 90.5606~
$ star_x       <dbl> 58.21361, 58.19605, 58.71823, 57.27837, 58.08202, 57.4894~
$ star_y       <dbl> 91.88189, 92.21499, 90.31053, 89.90761, 92.00815, 88.0852~
$ v_lines_x    <dbl> 50.48151, 50.28241, 50.18670, 50.32691, 50.45621, 30.4648~
```

```
$ v_lines_y    <dbl> 93.22270, 97.60998, 99.69468, 90.02205, 89.98741, 82.0892~
$ wide_lines_x <dbl> 65.81554, 65.67227, 39.00272, 37.79530, 35.51390, 39.2194~
$ wide_lines_y <dbl> 95.58837, 91.93340, 92.26184, 93.53246, 89.59919, 83.5434~
$ x_shape_x    <dbl> 38.33776, 35.75187, 32.76722, 33.72961, 37.23825, 36.0272~
$ x_shape_y    <dbl> 92.47272, 94.11677, 88.51829, 88.62227, 83.72493, 82.0407~
```

```r
# Command + shift + m: %>% --> pipe operator
datasaurus_dozen_wide %>%
  ggplot() +
  geom_point(mapping = aes(x = away_x, y = away_y))
```



Tidying Data

column = variable and row = observation

```r
df <- datasaurus_dozen_wide %>%
  mutate(id = row_number()) %>%
  pivot_longer(cols = -id,
               names_to = c("dataset","var_name"),
               values_to = "value",
               names_pattern = "^(.*)_(.*)$") %>%
```

```
  pivot_wider(names_from = var_name, values_from = value) %>%
  select(-id)
```

Summary statistics

```
# names of the dataset
df %>%
  distinct(dataset)
```

```
# A tibble: 13 x 1
   dataset
   <chr>
 1 away
 2 bullseye
 3 circle
 4 dino
 5 dots
 6 h_lines
 7 high_lines
 8 slant_down
 9 slant_up
10 star
11 v_lines
12 wide_lines
13 x_shape
```

```
dataset_list <- c("dino","star","bullseye","away")
```

```
df %>%
  filter(dataset == "dino") %>%
  summarise(mean_x = mean(x),
            sd_x = sd(x))
```

```
# A tibble: 1 x 2
  mean_x  sd_x
   <dbl> <dbl>
1   54.3  16.8
```

```r
df %>%
  filter(dataset %in% dataset_list) %>%
  group_by(dataset) %>%
  summarise(across(c(x,y), list(mean = mean, sd = sd)))
```

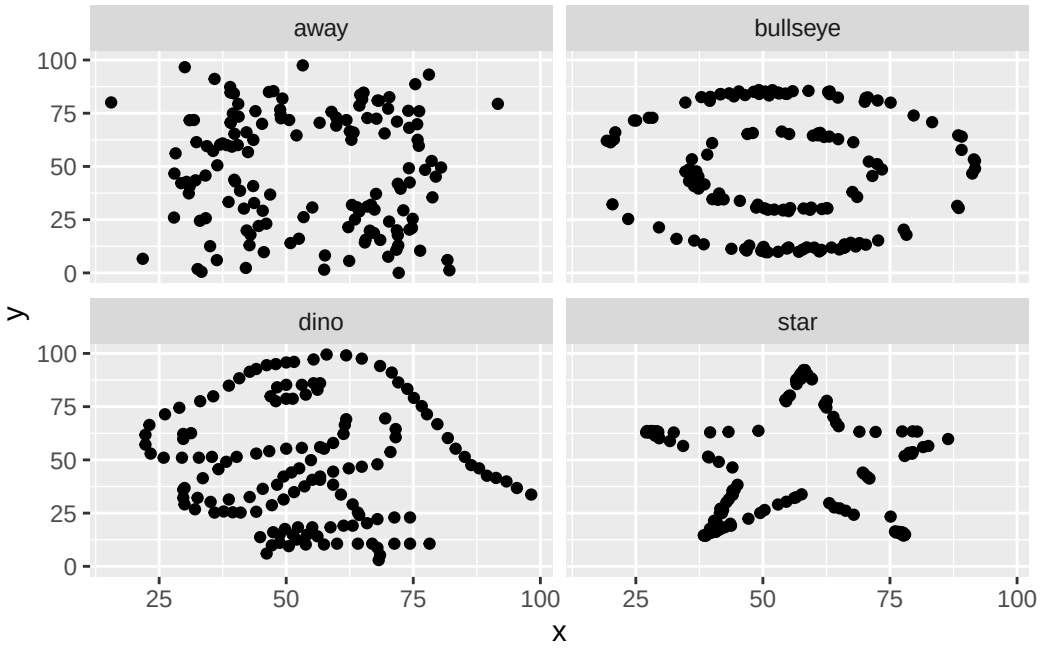
```
# A tibble: 4 x 5
  dataset  x_mean  x_sd y_mean  y_sd
  <chr>     <dbl> <dbl>  <dbl> <dbl>
1 away       54.3  16.8   47.8  26.9
2 bullseye   54.3  16.8   47.8  26.9
3 dino       54.3  16.8   47.8  26.9
4 star       54.3  16.8   47.8  26.9
```

```r
df %>%
  filter(dataset %in% dataset_list) %>%
  group_by(dataset) %>%
  summarise(across(c(x, y), list(mean = mean, sd = sd)),
            slope = coef(lm(y ~ x))[2])
```

```
# A tibble: 4 x 6
  dataset  x_mean  x_sd y_mean  y_sd  slope
  <chr>     <dbl> <dbl>  <dbl> <dbl>  <dbl>
1 away       54.3  16.8   47.8  26.9 -0.103
2 bullseye   54.3  16.8   47.8  26.9 -0.110
3 dino       54.3  16.8   47.8  26.9 -0.104
4 star       54.3  16.8   47.8  26.9 -0.101
```

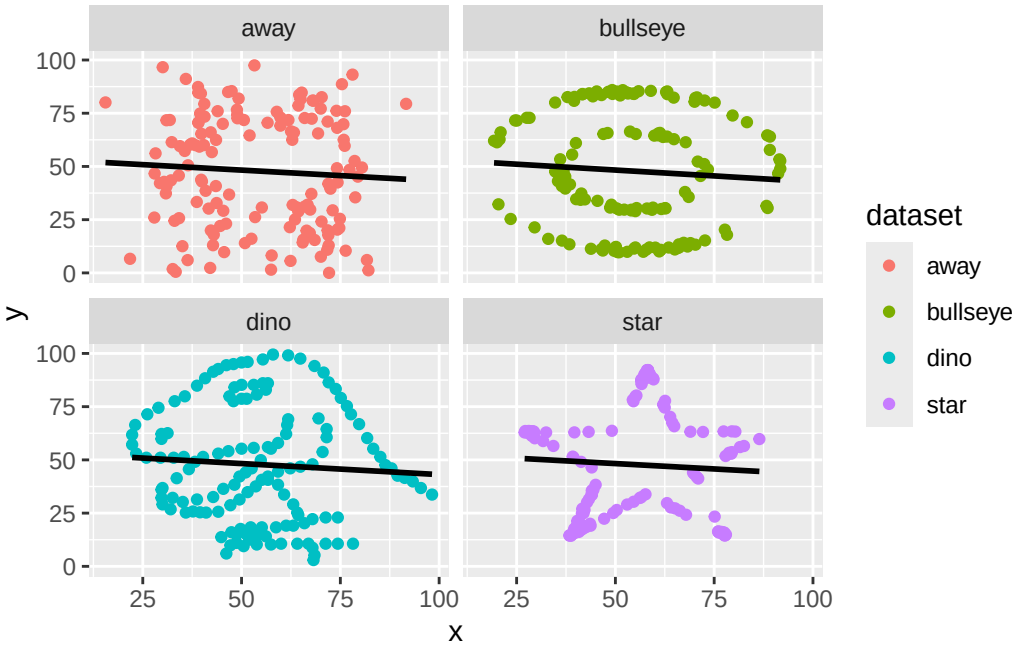
Plotting

```r
df %>%
  filter(dataset %in% dataset_list) %>%
  ggplot() +
  geom_point(aes(x, y)) +
  facet_wrap(vars(dataset))
```




```
df %>%
  filter(dataset %in% dataset_list) %>%
  ggplot(aes(x, y, color = dataset)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  facet_wrap(vars(dataset))
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Part 2: The Titanic Dataset

```
# use assignment side
titanic <- read_csv("/Users/ingridcanelles/Documents/GitHub/statcomp/datasets/titanic.csv")
```

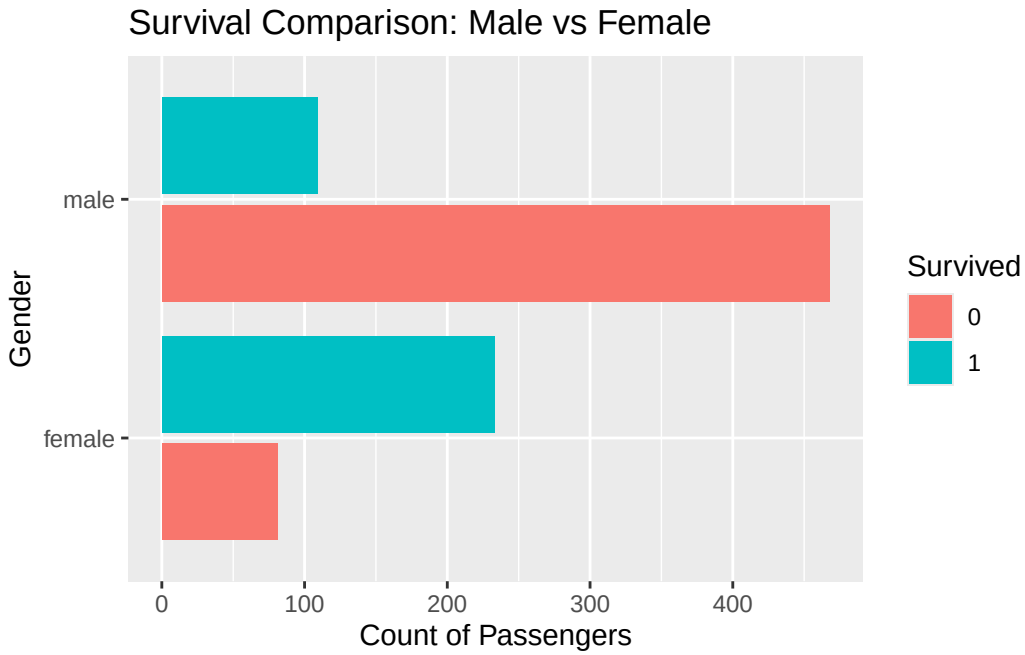
```
Rows: 891 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (5): Name, Sex, Ticket, Cabin, Embarked
dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Gender vs survival

```
titanic = titanic %>%
  mutate(across(c(Survived, Pclass, Sex), as.factor))
```

```r
titanic %>%
  ggplot() +
  geom_bar(aes(y = Sex, fill = Survived), position = "dodge2") +
  labs(title = "Survival Comparison: Male vs Female",
       x = "Count of Passengers",
       y = "Gender")
```

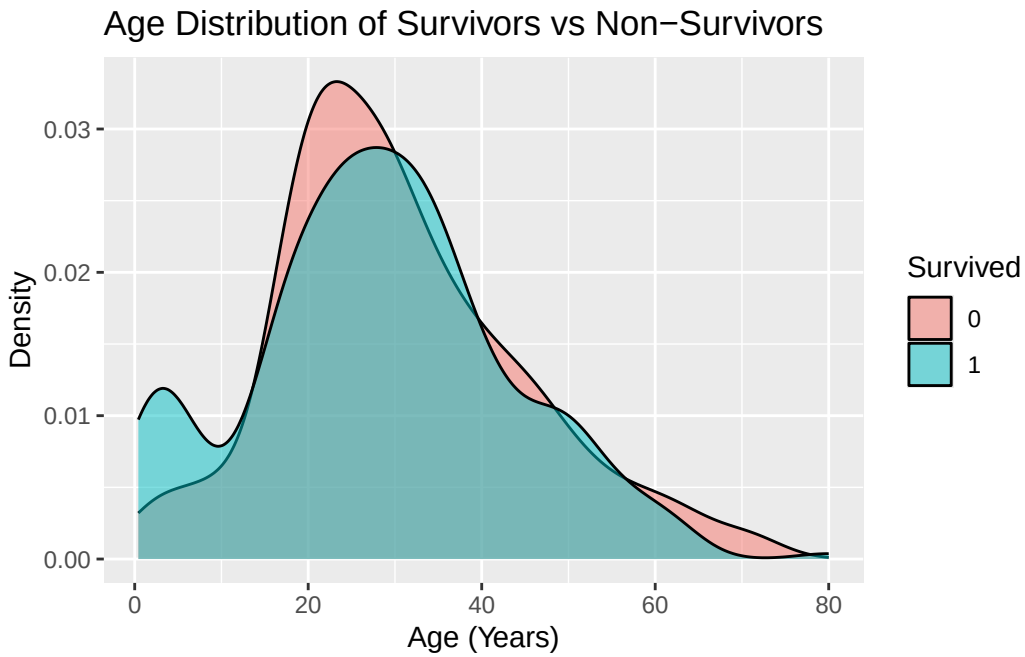


Survival by gender and class

```r
titanic %>%
  ggplot() +
  geom_bar(aes(y = Sex, fill = Survived), position = "dodge2") +
  facet_wrap(~Pclass) +
  theme_minimal() +
  labs(title = "Survival by Gender across Passenger Classes",
       x = "Count of Passengers",
       y = "Gender")
```

Age distribution

```
titanic %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = Age, fill = Survived)) +
  geom_density(alpha = 0.5) +
  labs(title = "Age Distribution of Survivors vs Non-Survivors",
       x = "Age (Years)",
       y = "Density")
```

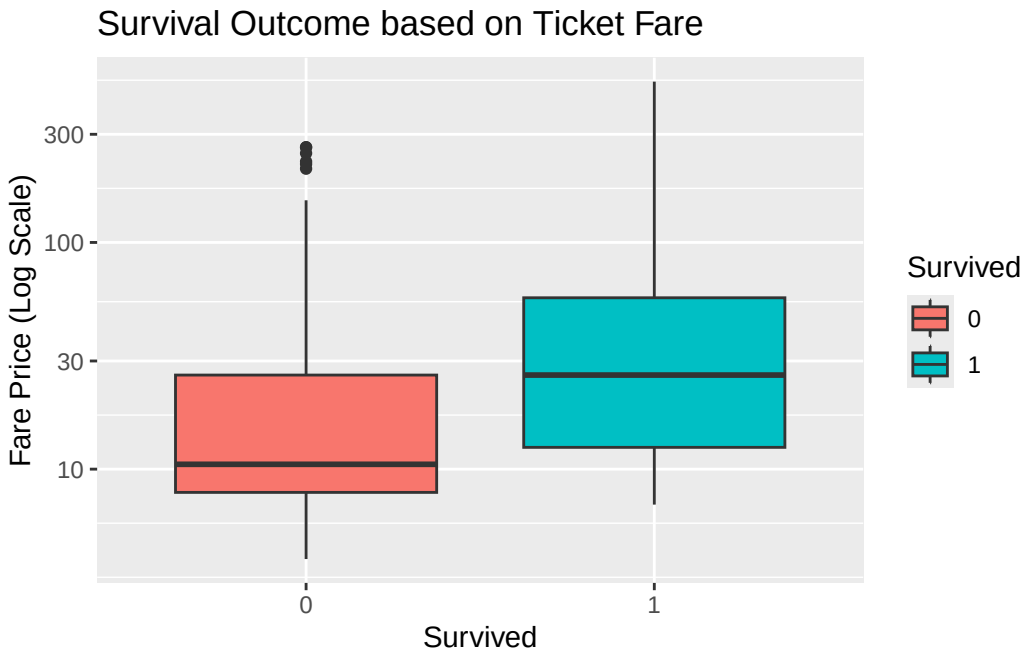


Fare and survival

```
titanic %>%
  ggplot(aes(x = Survived, y = Fare, fill = Survived)) +
  geom_boxplot() +
  scale_y_log10() +
  labs(title = "Survival Outcome based on Ticket Fare",
       y = "Fare Price (Log Scale)",
       x = "Survived")
```

```
Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
Warning: Removed 15 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Interpretation

- Women were much more likely to survive than men.
- Passengers in higher classes showed higher survival rates.
- Higher ticket fares were associated with a greater chance of survival.
- Third class passengers, especially men, had the lowest survival probability.
- Overall, gender and social class strongly influenced survival.