EDA with R

Junyuan Zheng

2020-12-07

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork) # combining panels
library(ggbeeswarm) # swarmplot
library(corrplot) # plotting correlation matrix
```

```
## corrplot 0.84 loaded
```

Preprocessing

1.1 Data import

```
data_train = read_csv("./dataset/train.csv") # lib(readr)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   PassengerId = col_double(),
##   Survived = col_double(),
##   Pclass = col_double(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   SibSp = col_double(),
##   Parch = col_double(),
##   Ticket = col_character(),
```

```
##   Fare = col_double(),
##   Cabin = col_character(),
##   Embarked = col_character()
## )
```

```
is.data.frame(data_train) # lib(base)
```

```
## [1] TRUE
```

1.2 A Glimpse of the Data

```
head(data_train, 5) # lib(utils)
```

```
## # A tibble: 5 x 12
##   PassengerId Survived Pclass Name  Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1           1        0      3 Brau~ male     22     1     0 A/5 2~  7.25 <NA>
## 2           2        1      1 Cumi~ fema~    38     1     0 PC 17~ 71.3  C85
## 3           3        1      3 Heik~ fema~    26     0     0 STON/~  7.92 <NA>
## 4           4        1      1 Futr~ fema~    35     1     0 113803 53.1  C123
## 5           5        0      3 Alle~ male     35     0     0 373450  8.05 <NA>
## # ... with 1 more variable: Embarked <chr>
```

```
# structure of the data
str(data_train) # lib(utils)
```

```
## tibble [891 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ PassengerId: num [1:891] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : num [1:891] 0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : num [1:891] 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr [1:891] "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Th
##  $ Sex        : chr [1:891] "male" "female" "female" "female" ...
##  $ Age        : num [1:891] 22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : num [1:891] 1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : num [1:891] 0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr [1:891] "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num [1:891] 7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr [1:891] NA "C85" NA "C123" ...
##  $ Embarked   : chr [1:891] "S" "C" "S" "S" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   PassengerId = col_double(),
##   ..   Survived = col_double(),
##   ..   Pclass = col_double(),
##   ..   Name = col_character(),
##   ..   Sex = col_character(),
##   ..   Age = col_double(),
##   ..   SibSp = col_double(),
##   ..   Parch = col_double(),
##   ..   Ticket = col_character(),
```

```
##   ..   Fare = col_double(),
##   ..   Cabin = col_character(),
##   ..   Embarked = col_character()
##   .. )
```

- **Note** that some default data type is not appropriate, such as 'Survived', 'Pclass'.

```
# dimension
dim(data_train) # lib(base)
```

```
## [1] 891  12
```

- **891** rows.
- 12 cols.

```
summary(data_train) # lib(base)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
# check frequency for categorical variables
table(data_train$Survived)
```

```
##
##   0   1
## 549 342
```

```r
# two-way table
surv_sex = table(data_train$Survived, data_train$Sex) # lib(base); -> vector
surv_sex
```

```
##
##     female male
##   0     81  468
##   1    233  109
```

```r
matrix(surv_sex, nrow=2,
       dimnames = list(c("notSurv", "surv"),
                       c("female", "male")))
```

```
##         female male
## notSurv     81  468
## surv       233  109
```

```r
# three-way table
surv_sex = xtabs(~ Survived + Sex + Pclass, data=data_train) # lib(stats) -> matrix, array
surv_sex
```

```
## , , Pclass = 1
##
##         Sex
## Survived female male
##        0      3   77
##        1     91   45
##
## , , Pclass = 2
##
##         Sex
## Survived female male
##        0      6   91
##        1     70   17
##
## , , Pclass = 3
##
##         Sex
## Survived female male
##        0     72  300
##        1     72   47
```

```r
is.array(surv_sex)
```

```
## [1] TRUE
```

1.3 Check for Duplicates

```r
# check duplicates
sum(duplicated(data_train)) # lib(base)
```

```
## [1] 0
```

```r
# check unique
unique(data_train) %>%
  nrow()
```

```
## [1] 891
```

- There is no duplicates in the original data.

```r
# remove duplicates
data_train %>%
  distinct()  # lib(dplyr)
```

```
## # A tibble: 891 x 12
##    PassengerId Survived Pclass Name  Sex     Age SibSp Parch Ticket  Fare Cabin
##          <dbl>    <dbl>  <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
##  1           1        0      3 Brau~ male     22     1     0 A/5 2~  7.25 <NA>
##  2           2        1      1 Cumi~ fema~    38     1     0 PC 17~ 71.3  C85
##  3           3        1      3 Heik~ fema~    26     0     0 STON/~  7.92 <NA>
##  4           4        1      1 Futr~ fema~    35     1     0 113803 53.1  C123
##  5           5        0      3 Alle~ male     35     0     0 373450  8.05 <NA>
##  6           6        0      3 Mora~ male     NA     0     0 330877  8.46 <NA>
##  7           7        0      1 McCa~ male     54     0     0 17463  51.9  E46
##  8           8        0      3 Pals~ male      2     3     1 349909 21.1  <NA>
##  9           9        1      3 John~ fema~    27     0     2 347742 11.1  <NA>
## 10          10        1      2 Nass~ fema~    14     1     0 237736 30.1  <NA>
## # ... with 881 more rows, and 1 more variable: Embarked <chr>
```

```r
# drop any row with NA
data_train %>%
  drop_na()
```

```
## # A tibble: 183 x 12
##    PassengerId Survived Pclass Name  Sex     Age SibSp Parch Ticket  Fare Cabin
##          <dbl>    <dbl>  <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
##  1           2        1      1 Cumi~ fema~    38     1     0 PC 17~  71.3 C85
##  2           4        1      1 Futr~ fema~    35     1     0 113803  53.1 C123
##  3           7        0      1 McCa~ male     54     0     0 17463   51.9 E46
##  4          11        1      3 Sand~ fema~     4     1     1 PP 95~  16.7 G6
##  5          12        1      1 Bonn~ fema~    58     0     0 113783  26.6 C103
##  6          22        1      2 Bees~ male     34     0     0 248698  13   D56
##  7          24        1      1 Slop~ male     28     0     0 113788  35.5 A6
##  8          28        0      1 Fort~ male     19     3     2 19950  263   C23 ~
##  9          53        1      1 Harp~ fema~    49     1     0 PC 17~  76.7 D33
## 10          55        0      1 Ostb~ male     65     0     1 113509  62.0 B30
## # ... with 173 more rows, and 1 more variable: Embarked <chr>
```

1.4 Check for Missing Values

```
is.na(data_train) %>% # lib(base)
  summary()
```

```
##  PassengerId      Survived         Pclass           Name
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:891       FALSE:891       FALSE:891       FALSE:891
##
##     Sex             Age            SibSp           Parch
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:891       FALSE:714       FALSE:891       FALSE:891
##                  TRUE :177
##    Ticket           Fare           Cabin           Embarked
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:891       FALSE:891       FALSE:204       FALSE:889
##                                  TRUE :687       TRUE :2
```

```
is.na(data_train) %>%
  as.data.frame() %>%
  colSums() / nrow(data_train) # lib(base)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
## 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.198653199
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
## 0.000000000 0.000000000 0.000000000 0.000000000 0.771043771 0.002244669
```

- Age, Cabin, Embarked have missing values.

```
# check the proportion of missingness in 'Age' on 'Survived'
data_train %>%
  mutate(age_missing = is.na(Age)) %>%
  xtabs(~ Survived + age_missing, data = .) %>%
  prop.table(1) # lib(base)
```

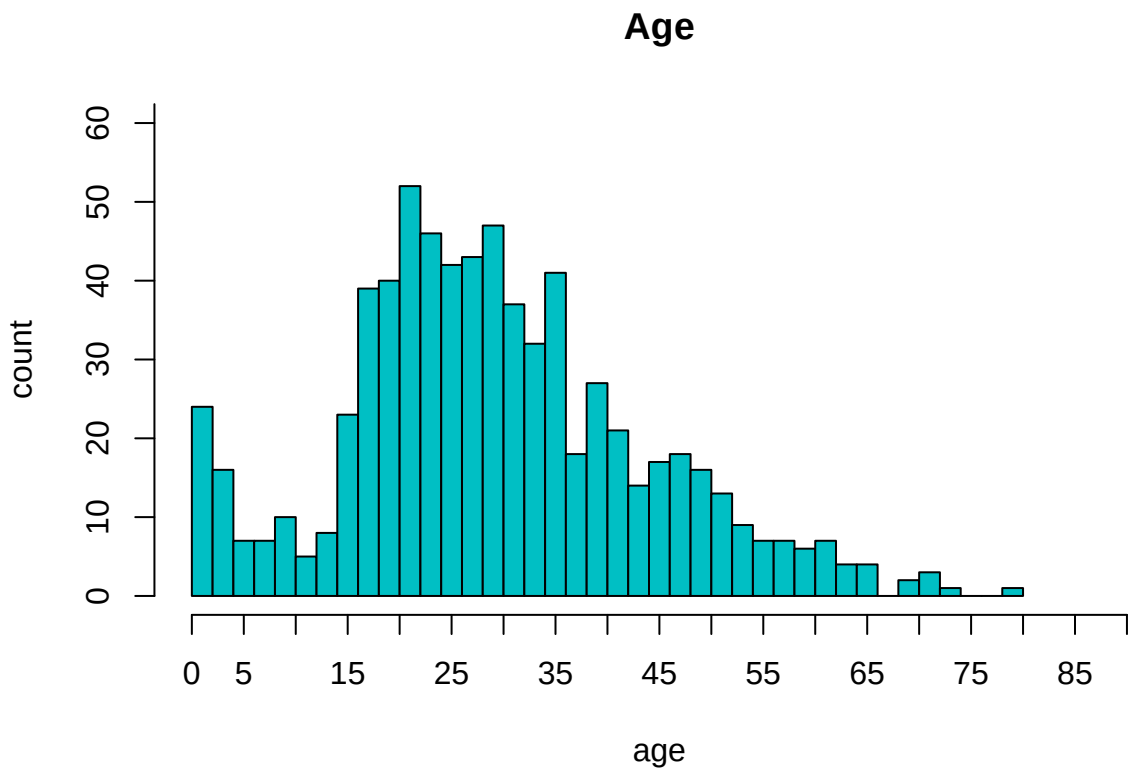
```
##          age_missing
## Survived     FALSE      TRUE
##        0 0.7723133 0.2276867
##        1 0.8479532 0.1520468
```

```
# addmargins() # lib(stat)
```

1.5 Exploratory Analysis

1.5.1 'Age'

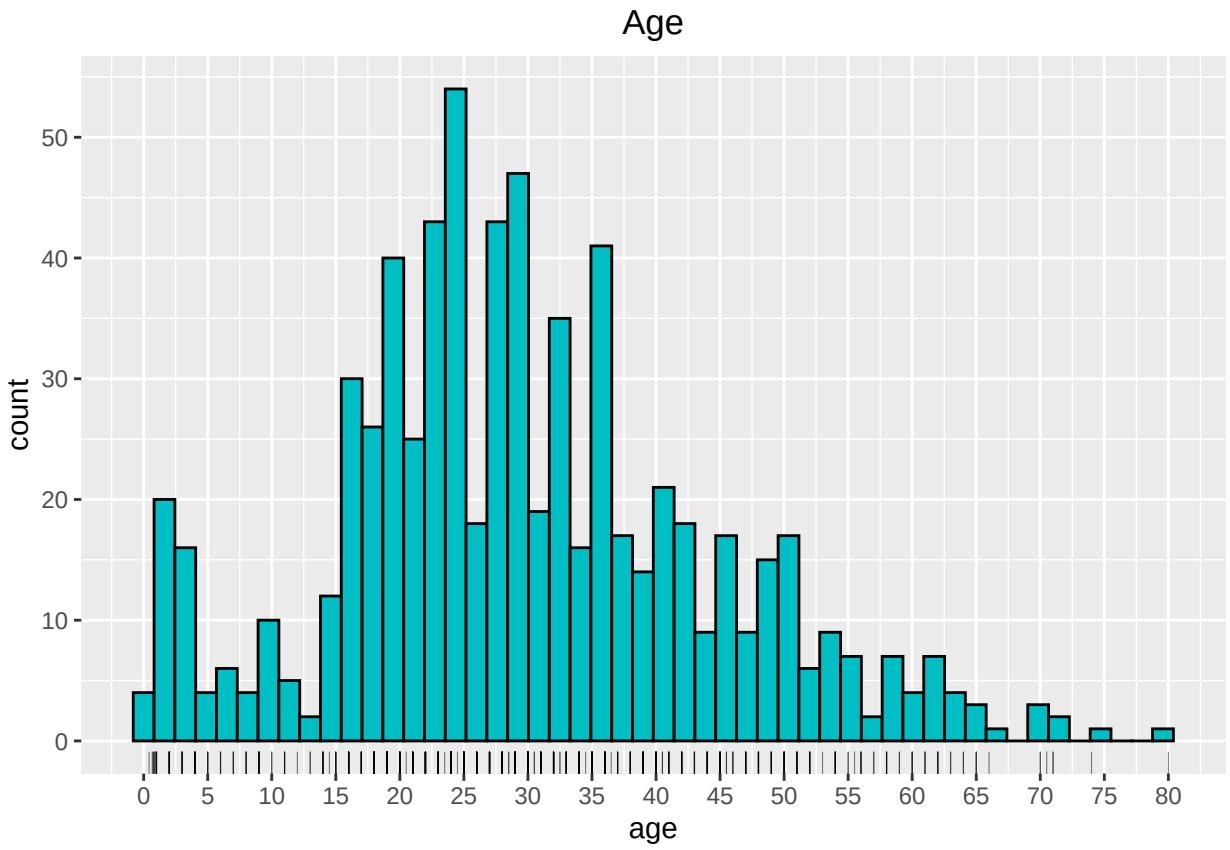
```r
# base R
hist_age =
hist(data_train$Age, # lib(graphics)
     breaks = 50,
     col = "#00BFC4",
     border = "black",
     xlim = c(0, 90),
     xaxp = c(0, 90, 18),
     ylim = c(0, 60),
     yaxp = c(0, 100, 10),
     main = "Age",
     xlab = "age",
     ylab = "count")
```



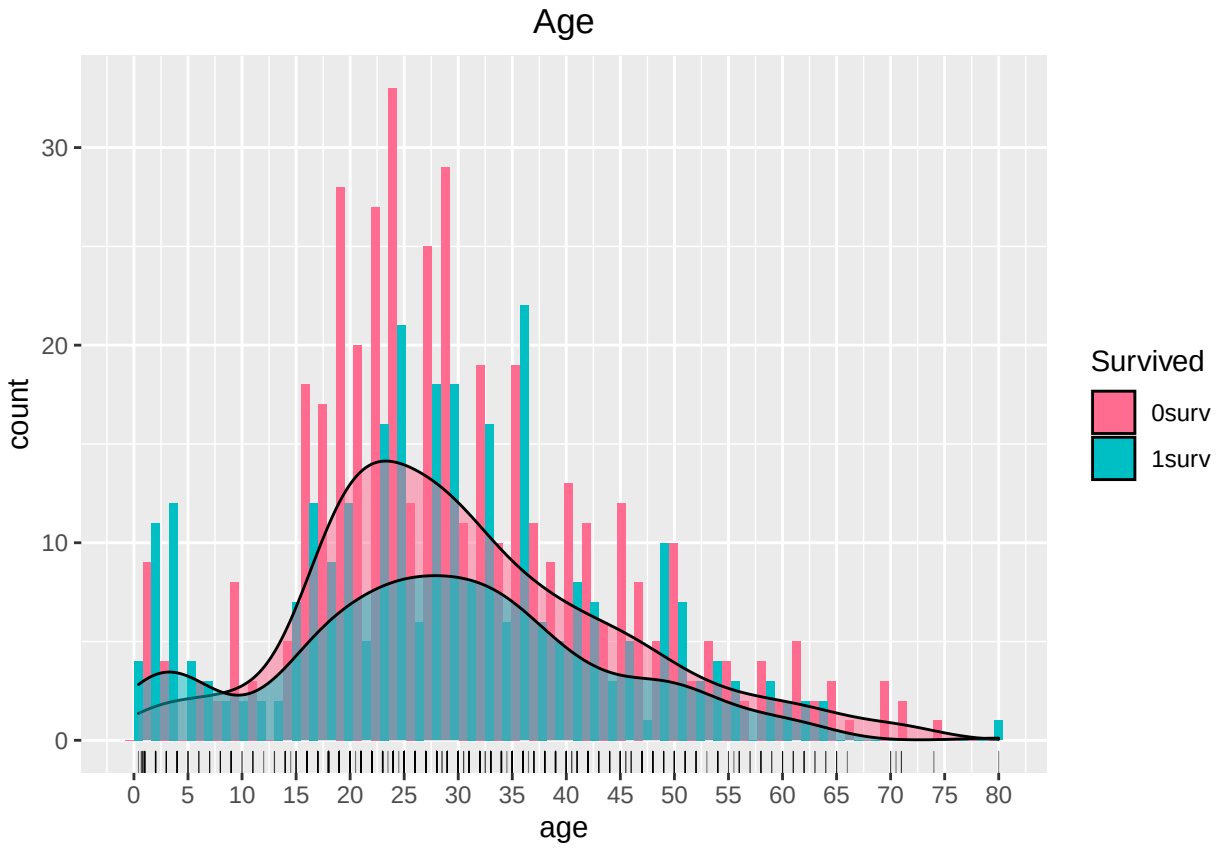

```r
#hist_age
#plot(hist_age)
```

```r
# ggplot2
data_train %>%
  filter(!is.na(Age)) %>% # filter out NA values
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 50, fill = "#00BFC4", color = "black") +
  scale_x_continuous(breaks = seq(0, 90, 5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  labs(
```

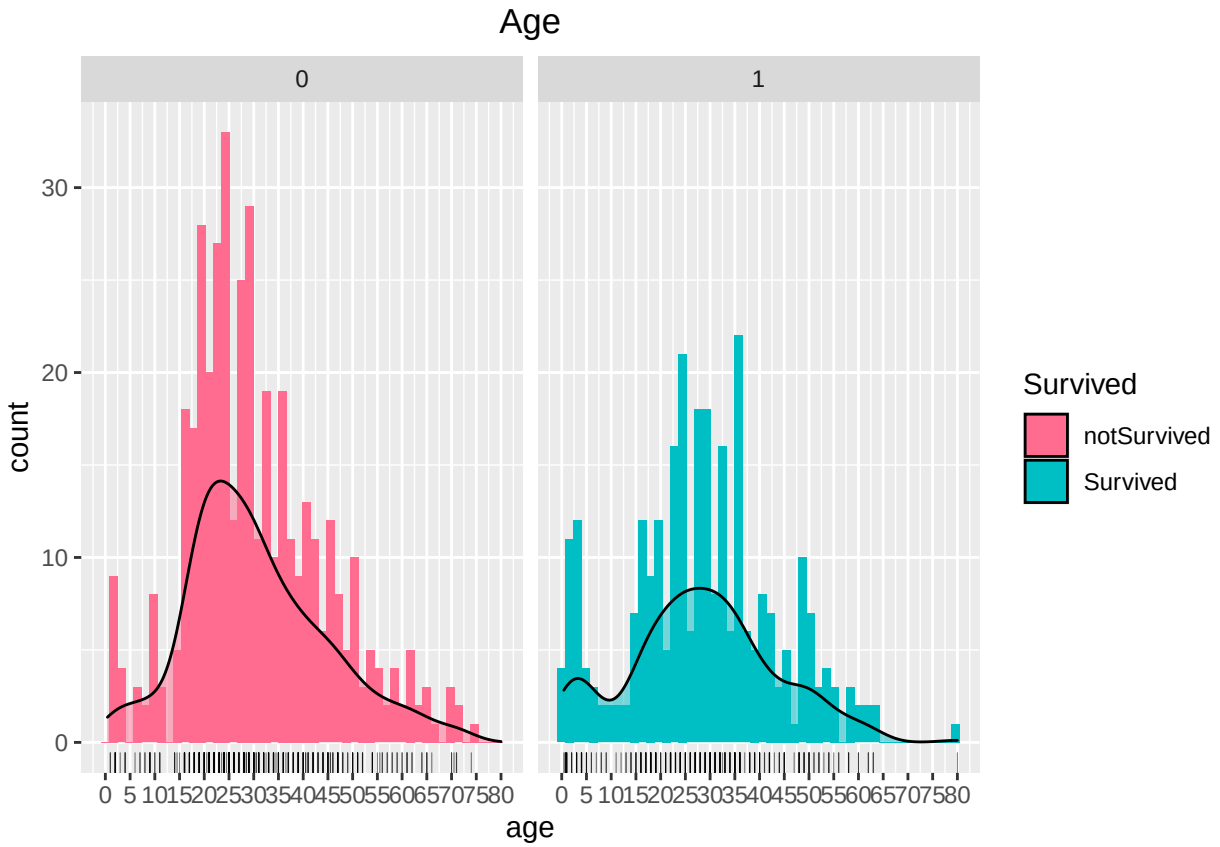
```
    title = "Age",
    x = "age",
    y = "count"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) + # center title
  geom_rug(alpha = 0.5, size = 0.2)
```




```
# 'Age' vs. 'Survived'
data_train %>%
  filter(!is.na(Age)) %>% # filter out NA values
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 50, aes(fill = as.factor(Survived)), position = "dodge") +
  scale_x_continuous(breaks = seq(0, 90, 5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  labs(
    title = "Age",
    x = "age",
    y = "count"
    #fill = "Survived"
  ) +
  scale_fill_manual(name = "Survived", values=c("#FF6C90", "#00BFC4"), labels = c("0surv", "1surv")) +
  theme(plot.title = element_text(hjust = 0.5)) + # center title
  geom_rug(alpha = 0.5, size = 0.2) +
  geom_density(aes(y = ..count.., fill = as.factor(Survived)), alpha = 0.5)
```

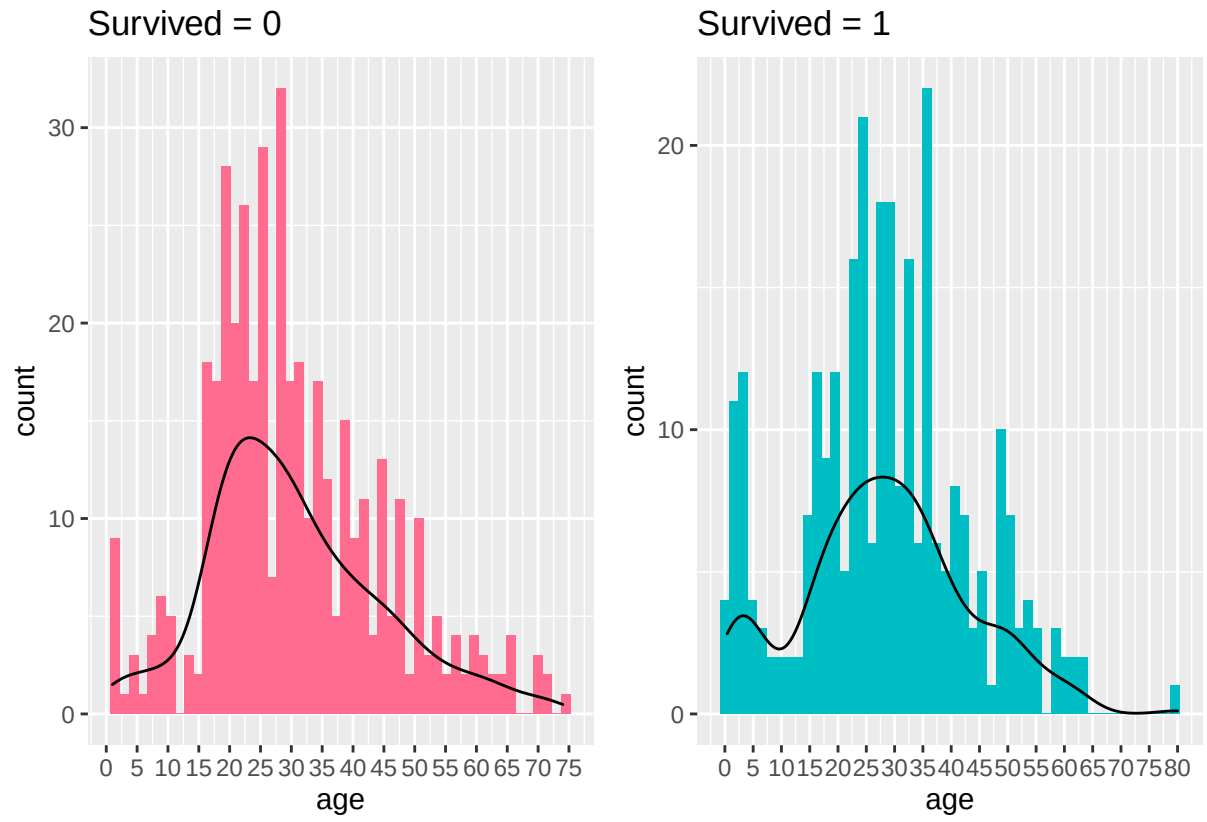



```
# 'Age' vs. 'Survived'
data_train %>%
  filter(!is.na(Age)) %>% # filter out NA values
  ggplot(aes(x = Age, fill = as.factor(Survived))) +
  geom_histogram(bins = 50) +
  scale_x_continuous(breaks = seq(0, 90, 5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  labs(
    title = "Age",
    x = "age",
    y = "count"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) + # center title
  geom_rug(alpha = 0.5, size = 0.2) +
  facet_grid(. ~ as.factor(Survived)) +
  scale_fill_manual(name = "Survived", values=c("#FF6C90", "#00BFC4"), labels = c("notSurvived", "Survi
  geom_density(aes(y = ..count.., fill = as.factor(Survived)), alpha = 0.5)
```

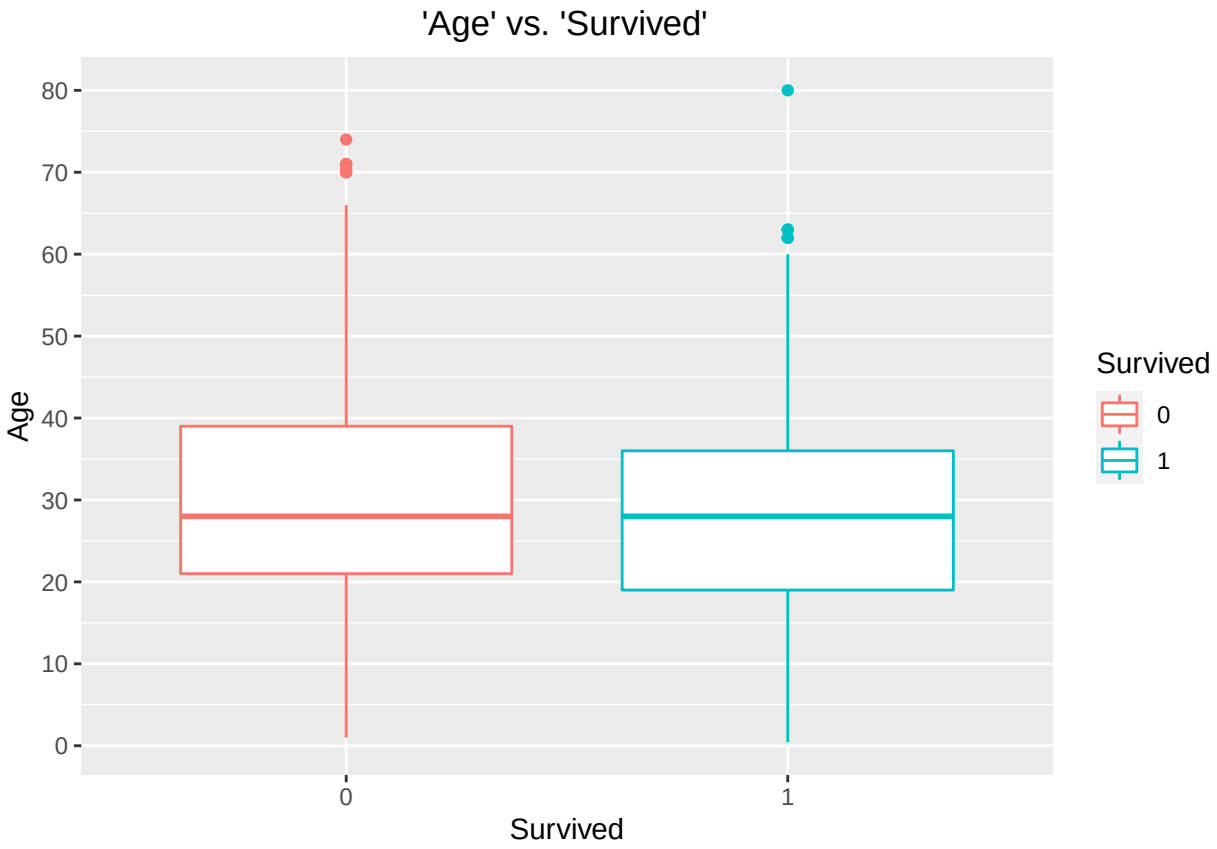



```
# 'Age' vs. 'Survived'
# patchwork
# library(patchwork)
surv0_age =
  data_train %>%
  filter(!is.na(Age)) %>%
  filter(Survived == 0) %>%
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 50, fill = "#FF6C90") +
  scale_x_continuous(breaks = seq(0, 90, 5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  labs(
    title = "Survived = 0",
    x = "age",
    y = "count"
  ) +
  geom_density(aes(y = ..count..))

surv1_age =
  data_train %>%
  filter(!is.na(Age)) %>%
  filter(Survived == 1) %>%
  ggplot(aes(x = Age)) +
  geom_histogram(bins = 50, fill = "#00BFC4") +
  scale_x_continuous(breaks = seq(0, 90, 5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
```

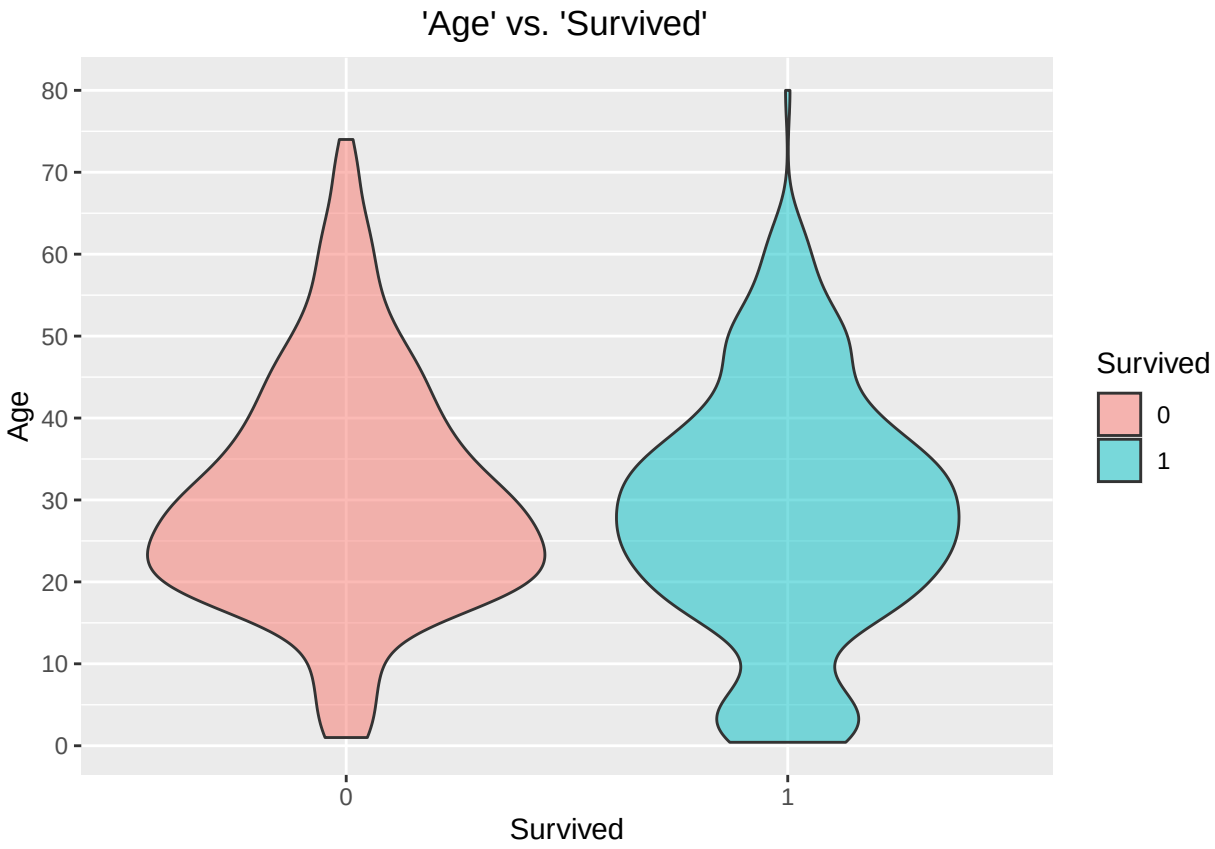
```
  labs(
    title = "Survived = 1",
    x = "age",
    y = "count"
  ) +
  geom_density(aes(y = ..count..))

surv0_age + surv1_age
```




```
# 'Age' vs. 'Survived'
# boxplot
data_train %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = as.factor(Survived), y = Age)) +
  geom_boxplot(aes(color = as.factor(Survived))) +
  labs(
    title = "'Age' vs. 'Survived'",
    x = "Survived"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  scale_color_discrete(name = "Survived")
```

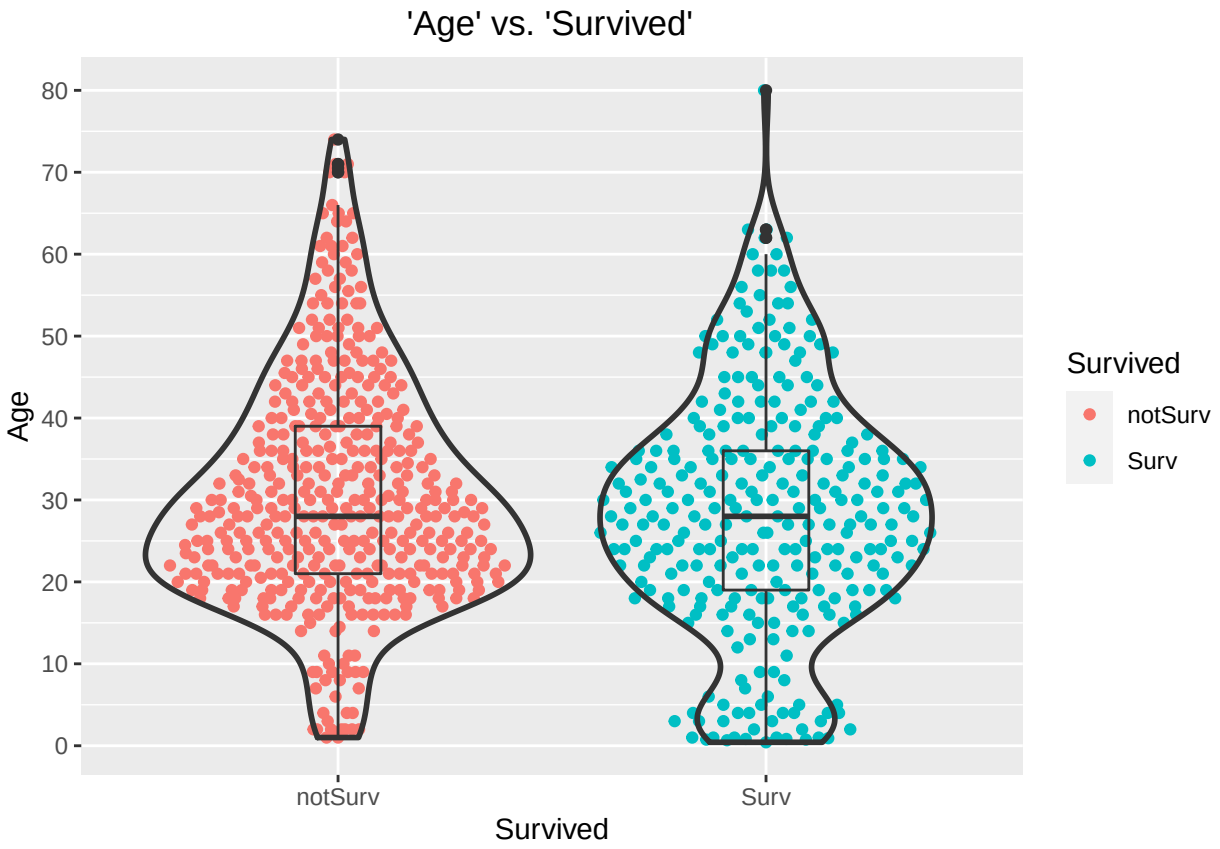



```
# 'Age' vs. 'Survived'
# violin
data_train %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = as.factor(Survived), y = Age)) +
  geom_violin(aes(fill = as.factor(Survived)), alpha = 0.5) +
  labs(
    title = "'Age' vs. 'Survived'",
    x = "Survived"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  scale_fill_discrete(name = "Survived")
```

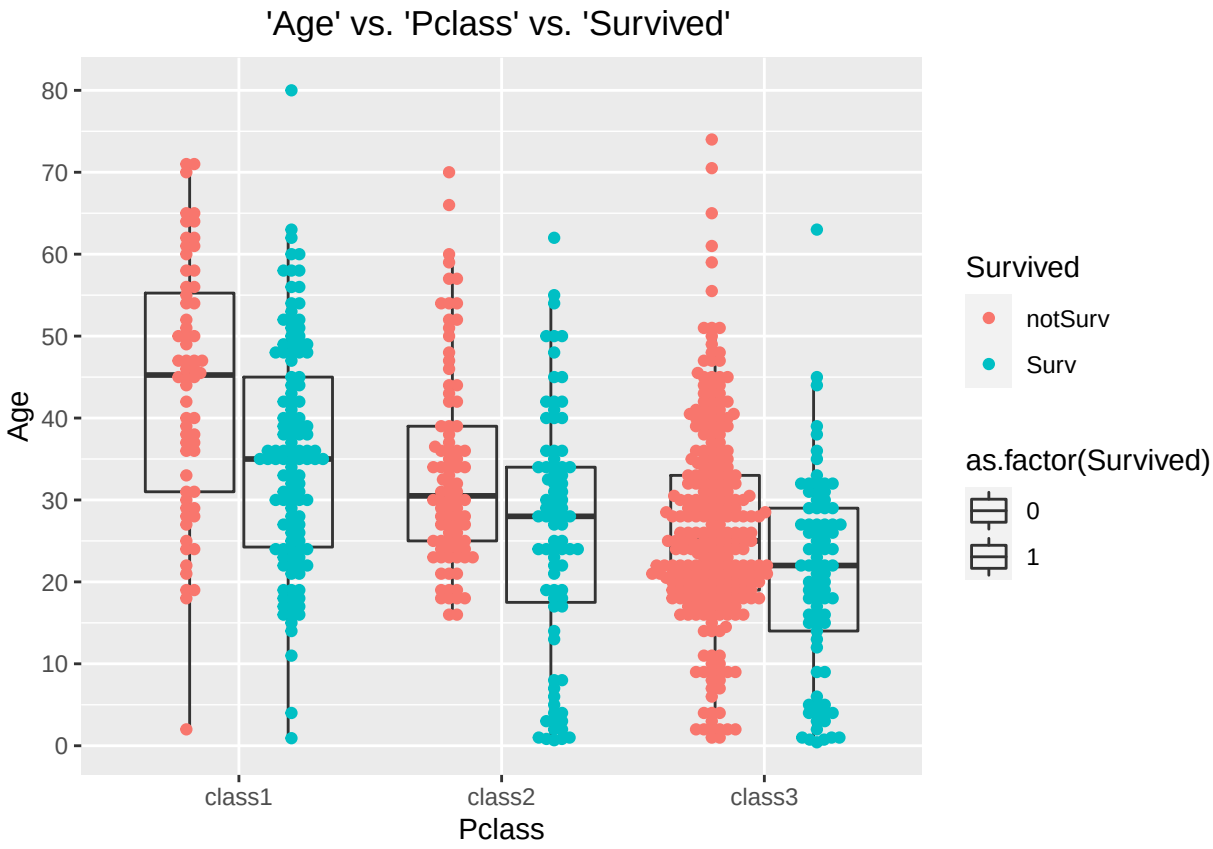



```
# 'Age' vs. 'Survived'
# swarmplot
# library(ggbeeswarm)
data_train %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = as.factor(Survived), y = Age)) +
  geom_quasirandom(aes(color = as.factor(Survived)), alpha = 0.8) +
  labs(
    title = "'Age' vs. 'Survived'",
    x = "Survived"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  scale_color_discrete(name = "Survived")
```


```
# 'Age' vs. 'Survived'
# swarmplot + violin + boxplot
data_train %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = as.factor(Survived), y = Age)) +
  geom_quasirandom(aes(color = as.factor(Survived)), alpha = 1) +
  geom_violin(size = 1, fill = NA) +
  geom_boxplot(fill = NA, width = 0.2, size = 0.5) +
  labs(
    title = "'Age' vs. 'Survived'",
    x = "Survived"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  scale_color_discrete(name = "Survived", labels = c("notSurv", "Surv")) +
  scale_x_discrete(labels = c("notSurv", "Surv"))
```

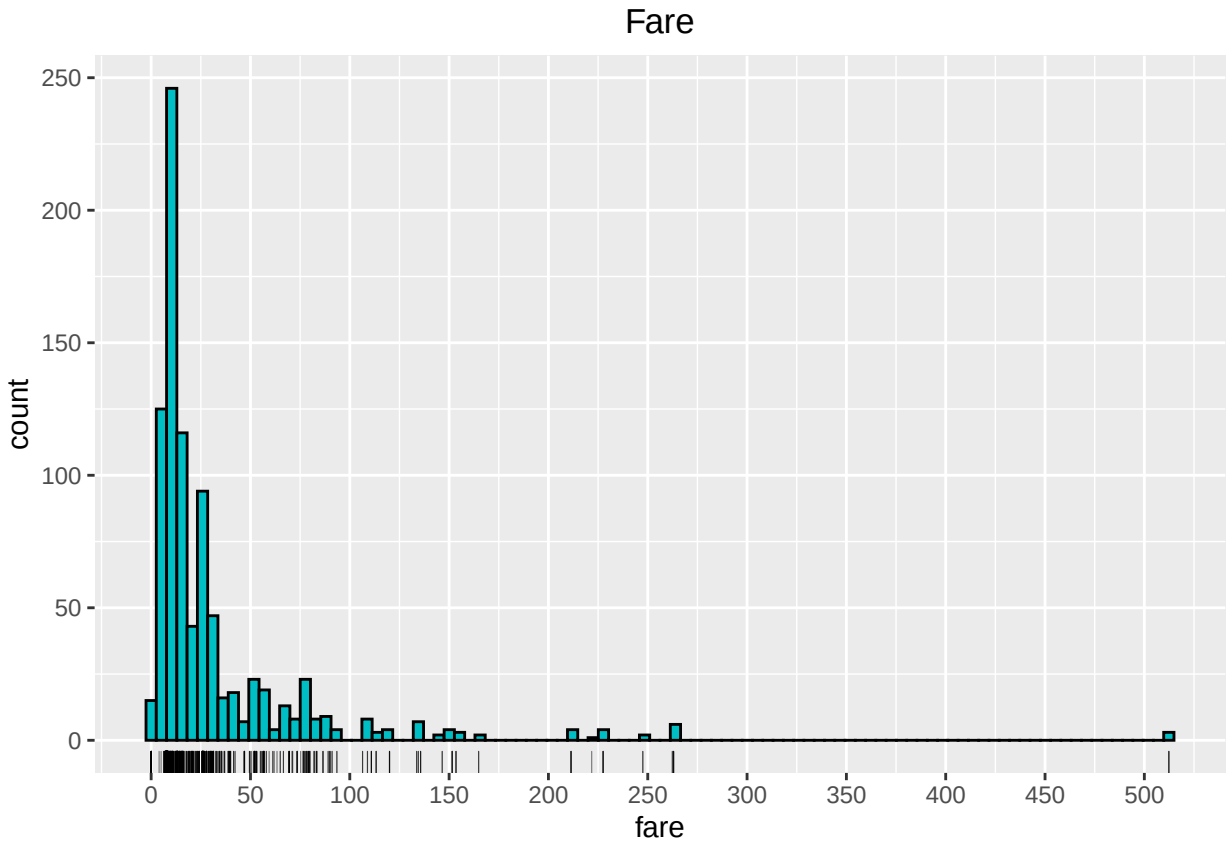



```
# 'Age' vs. 'Survived'
# boxplot
data_train %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = as.factor(Pclass), y = Age)) +
  geom_boxplot(aes(fill = as.factor(Survived)), alpha = 0) +
  geom_beeswarm(aes(color = as.factor(Survived)), alpha = 1, dodge.width = 0.8) +
  labs(
    title = "'Age' vs. 'Pclass' vs. 'Survived'",
    x = "Pclass"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  scale_x_discrete(labels = c("class1", "class2", "class3")) +
  scale_color_discrete(name = "Survived", labels = c("notSurv", "Surv"))
```

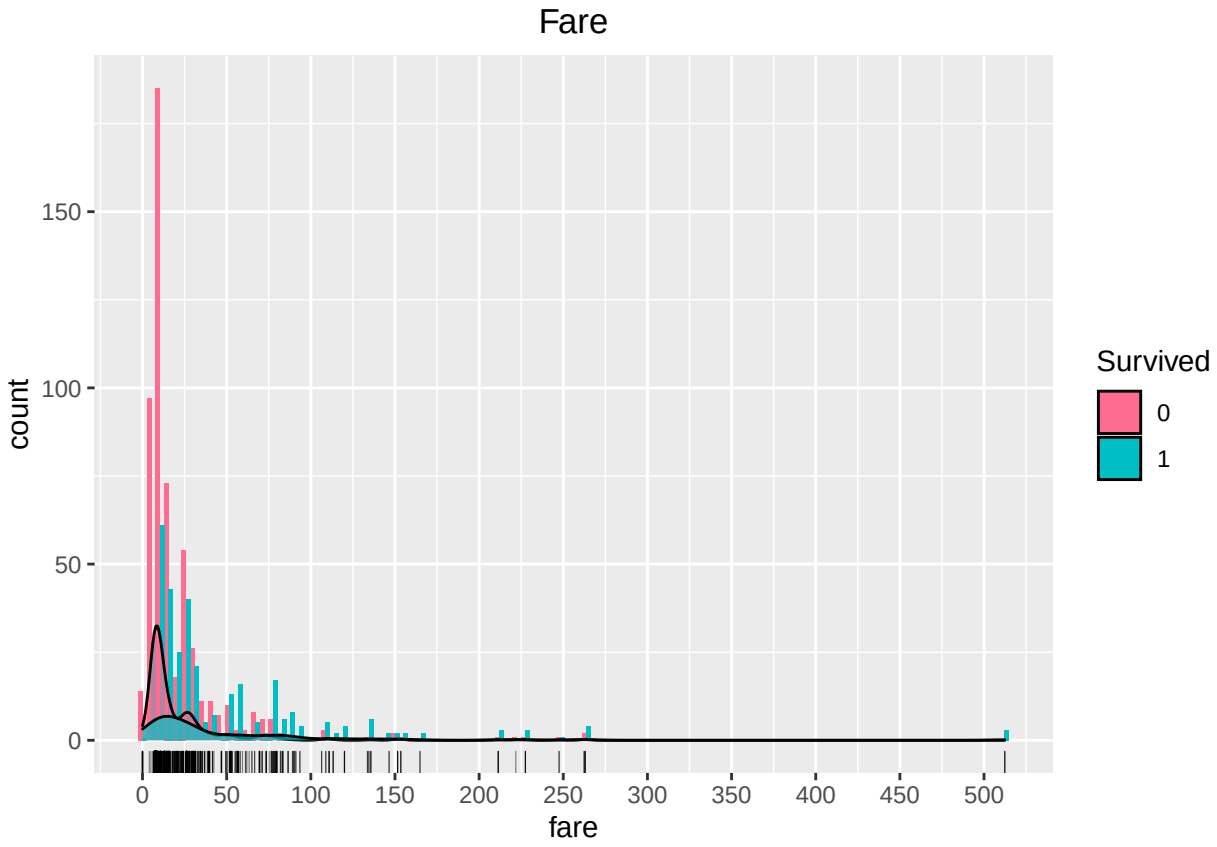


1.5.2 'Fare'

```
# ggplot2
data_train %>%
  ggplot(aes(x = Fare)) +
  geom_histogram(bins = 100, fill = "#00BFC4", color = "black") +
  scale_x_continuous(breaks = seq(0, 600, 50)) +
  scale_y_continuous(breaks = seq(0, 300, 50)) +
  labs(
    title = "Fare",
    x = "fare",
    y = "count"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) + # center title
  geom_rug(alpha = 0.5, size = 0.2)
```




```
# 'Fare' vs. 'Survived'
data_train %>%
  ggplot(aes(x = Fare)) +
  geom_histogram(bins = 100, aes(fill = as.factor(Survived)), position = "dodge") +
  scale_x_continuous(breaks = seq(0, 600, 50)) +
  scale_y_continuous(breaks = seq(0, 300, 50)) +
  labs(
    title = "Fare",
    x = "fare",
    y = "count"
    #fill = "Survived"
  ) +
  scale_fill_manual(name = "Survived", values=c("#FF6C90", "#00BFC4"), labels = c("0", "1")) +
  theme(plot.title = element_text(hjust = 0.5)) + # center title
  geom_rug(alpha = 0.5, size = 0.2) +
  geom_density(aes(y = ..count.., fill = as.factor(Survived)), alpha = 0.7)
```

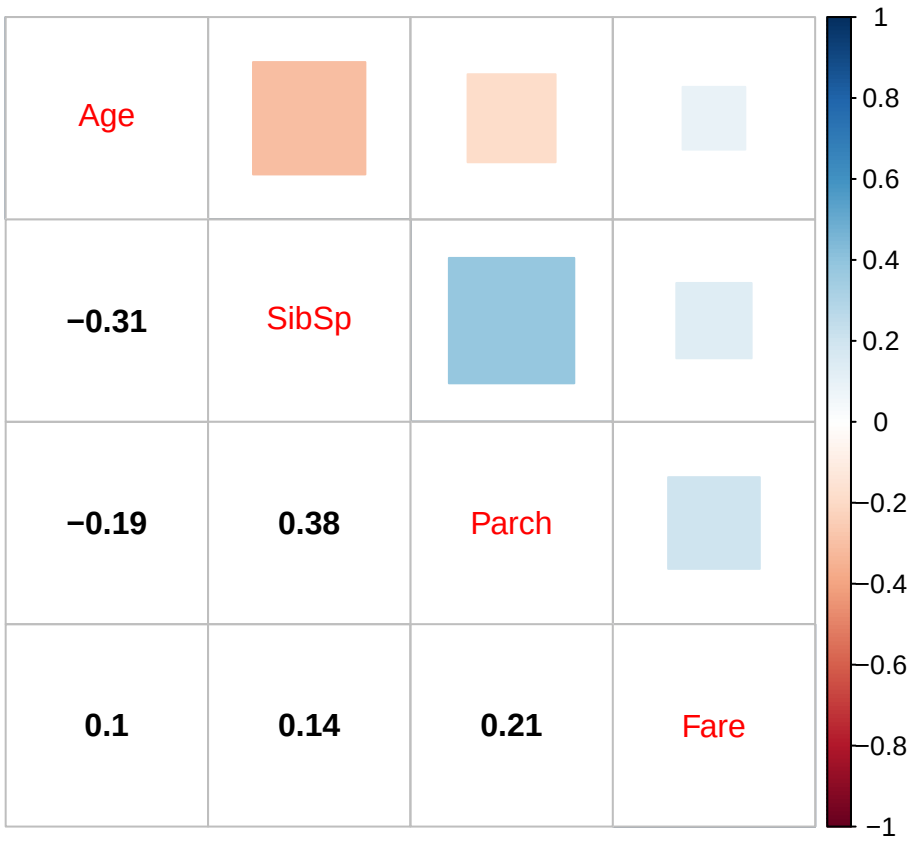


1.5.3 Comprehensive

```
# correlation matrix of continuous variables
data_train %>%
  filter(!is.na(Age)) %>%
  select(Age, SibSp, Parch, Fare) %>%
  cor()
```

```
##                Age      SibSp      Parch       Fare
## Age    1.00000000 -0.3082468 -0.1891193 0.09606669
## SibSp -0.30824676  1.0000000  0.3838199 0.13832879
## Parch -0.18911926  0.3838199  1.0000000 0.20511888
## Fare   0.09606669  0.1383288  0.2051189 1.00000000
```

```
# library(corrplot)
data_train %>%
  filter(!is.na(Age)) %>%
  select(Age, SibSp, Parch, Fare) %>%
  cor() %>%
  corrplot.mixed(lower.col = "black", upper = "square")
```




```
# facet grid
data_train %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = as.factor(Survived), y = Age)) +
  geom_quasirandom(aes(color = as.factor(Survived)), alpha = 0.8) +
  labs(
    title = "'Age' vs. 'Survived' by 'Sex' and 'Pclass'",
    x = "Survived"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks = seq(0, 100, 10)) +
  scale_color_discrete(name = "Survived") +
  facet_grid(Pclass ~ Sex,
             labeller = label_both)
```


'Age' vs. 'Survived' by 'Sex' and 'Pclass'
Sex: female
Sex: male
Pclass: 1
Pclass: 2
Pclass: 3
Age
Survived
0
1
Survived
0
1