Data visualization with ggplot2

Install packages

```
# remove comment for the first time
# install.packages( "ggplot2" )
# install.packages( "titanic" )
```

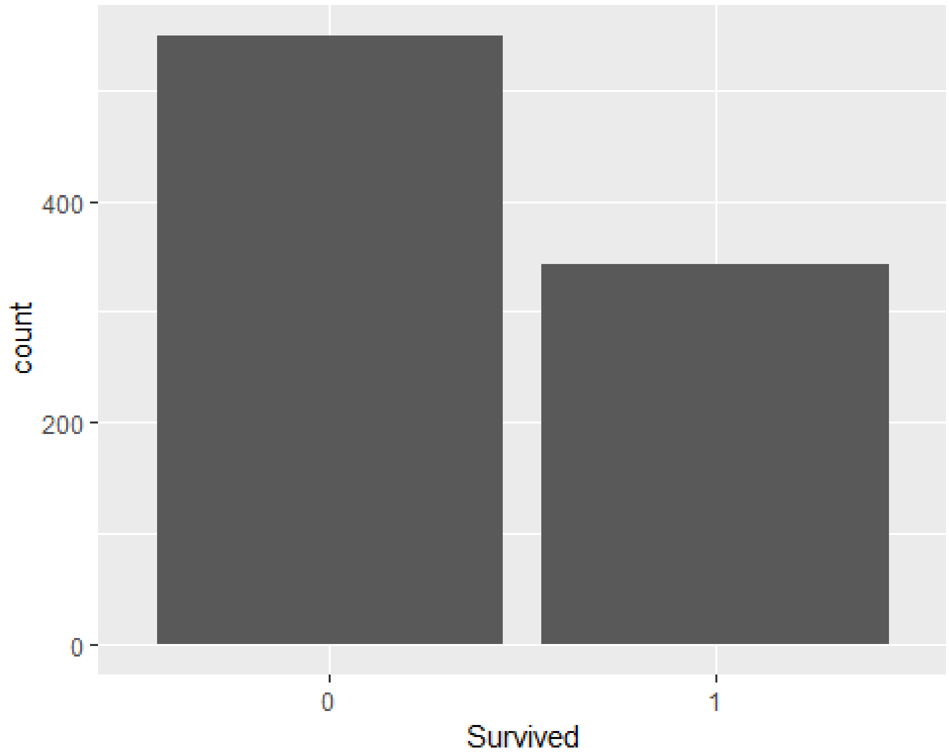
Preparations of the titanic dataset

```
library(titanic)

titanic.df <- titanic_train

View(titanic.df)

#for( n in names(titanic.df) )
#  print( paste( n, class( titanic.df[,n] ) ) )

titanic.df$Survived <- as.factor( titanic.df$Survived )
titanic.df$Pclass <- as.factor( titanic.df$Pclass )
titanic.df$Sex <- as.factor( titanic.df$Sex )
titanic.df$Embarked <- as.factor( titanic.df$Embarked )

titanic.df$NumRelatives <- titanic.df$SibSp + titanic.df$Parch

library(ggplot2)
```

Bar plot

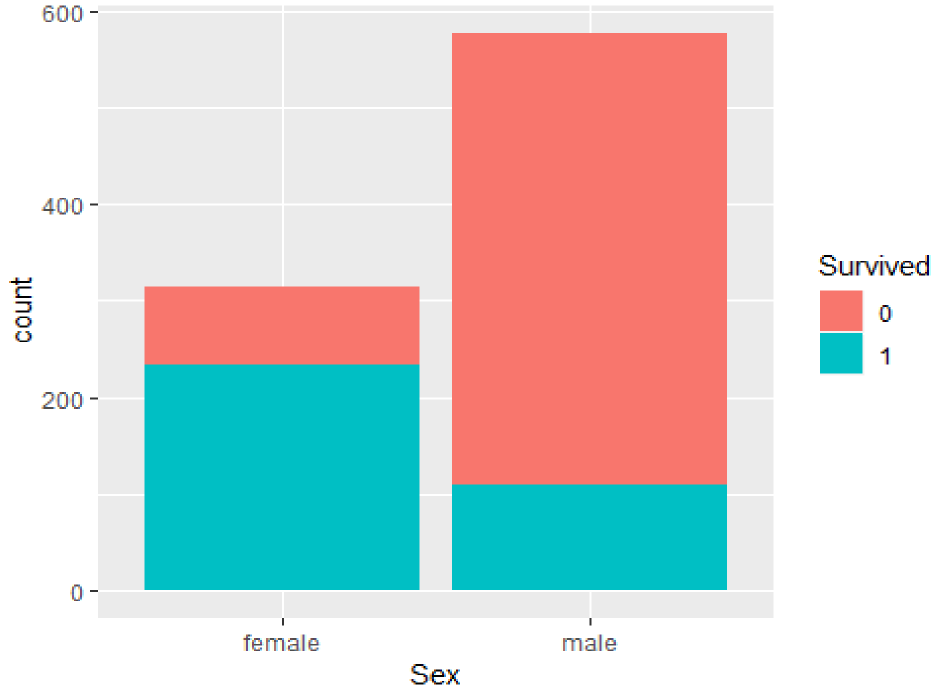
1. Survival rates

```
ggplot( data = titanic.df, aes( x = Survived ) ) +
  geom_bar()
```



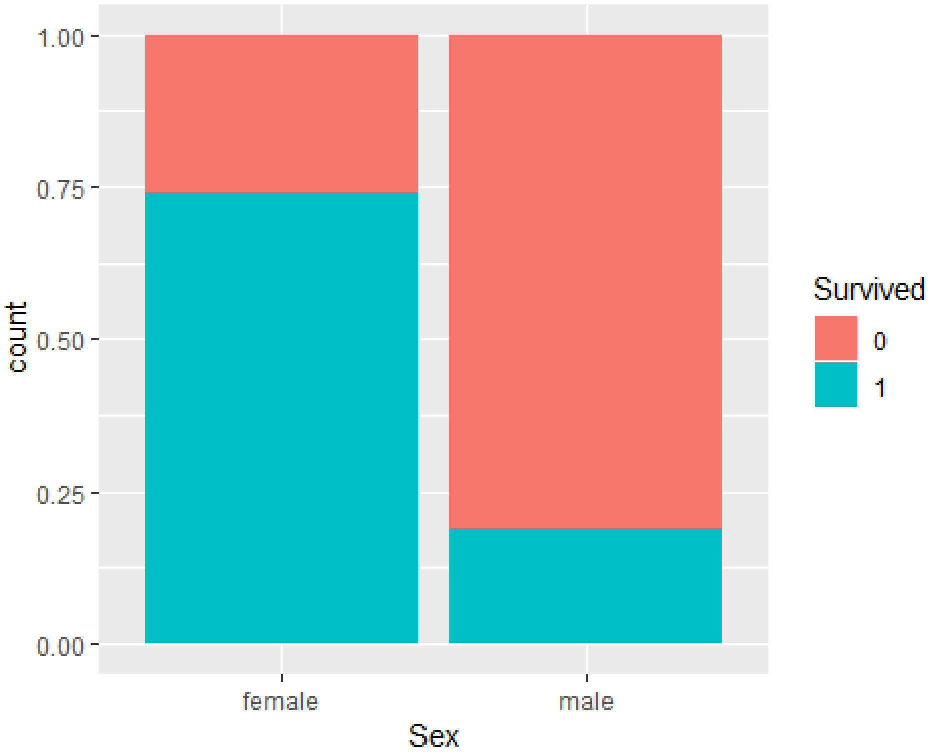
2. Survival rate by gender

```
ggplot( data = titanic.df, aes( x = Sex, fill = Survived ) ) +
  geom_bar()
```



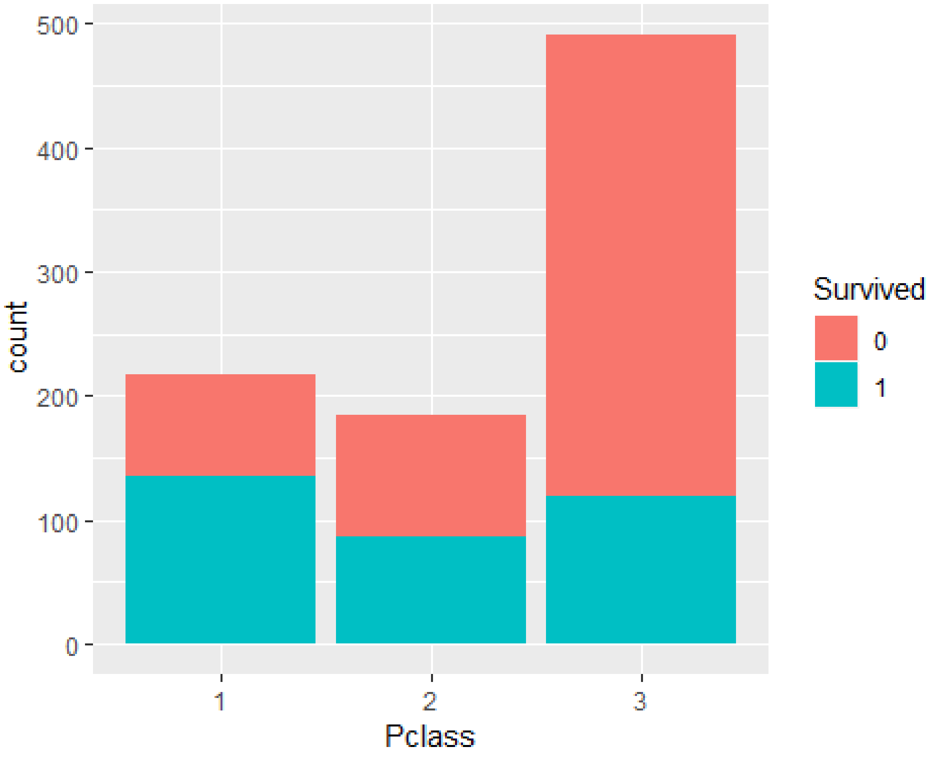
3. Proportions of survival rate by gender

```
ggplot( data = titanic.df, aes( x = Sex, fill = Survived ) ) +
  geom_bar( position = "fill" )
```



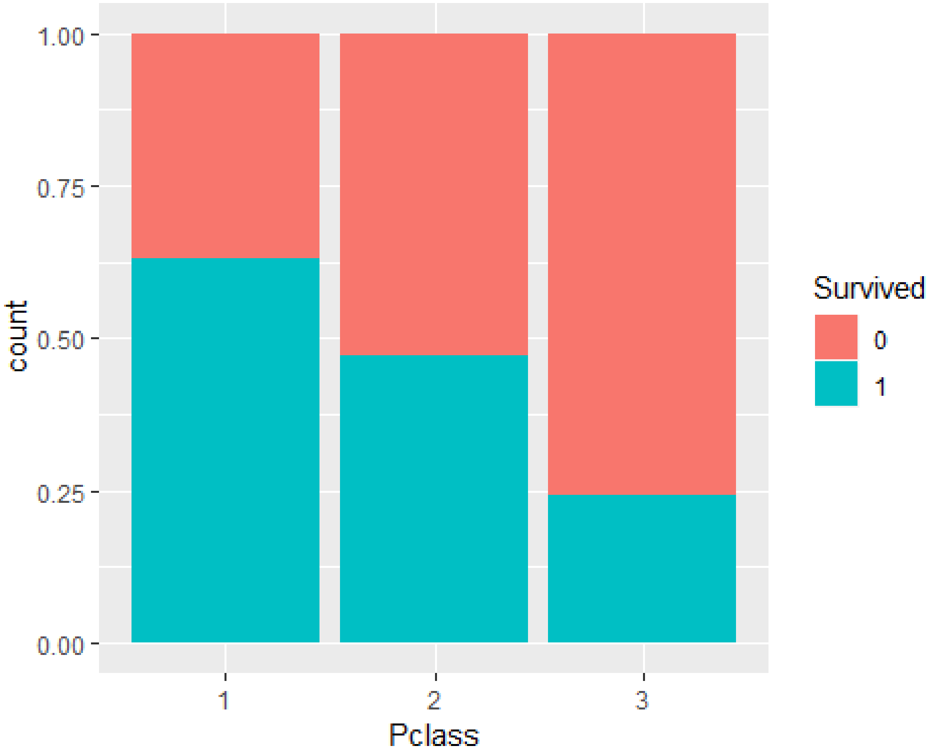
4. Survival rate by Passenger class

```
ggplot( data = titanic.df, aes( x = Pclass, fill = Survived ) ) +
  geom_bar()
```



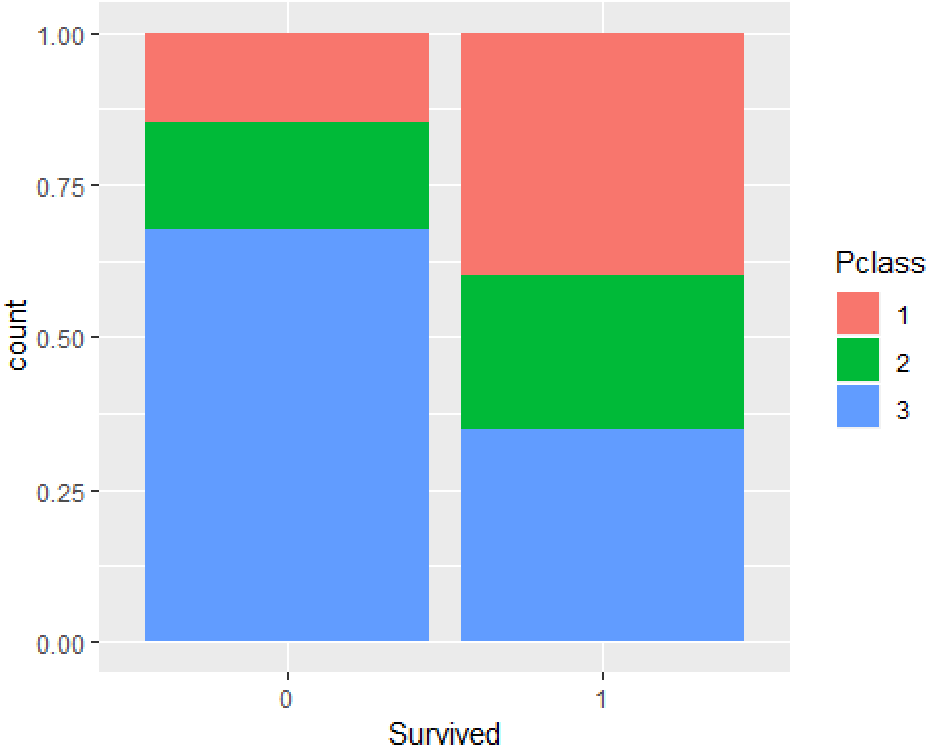
5. Proportions of survival rate by Passenger class

```
ggplot( data = titanic.df, aes( x = Pclass, fill = Survived ) ) +
  geom_bar( position = "fill" )
```



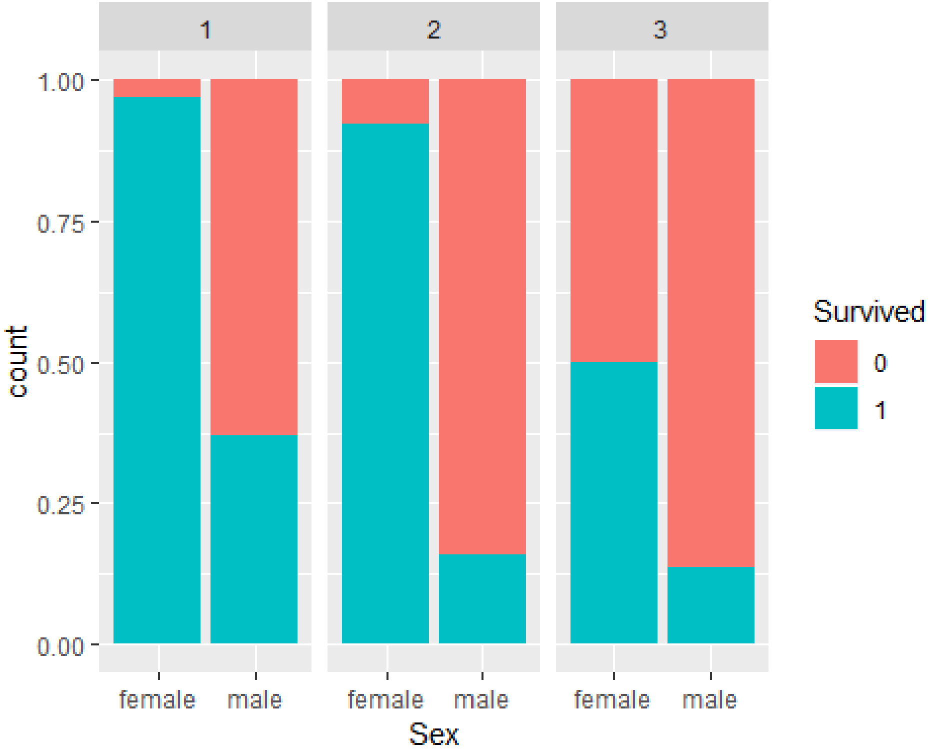
6. Proportions of Passenger class by survival

```
ggplot( data = titanic.df, aes( x = Survived, fill = Pclass ) ) +
  geom_bar( position = "fill" )
```



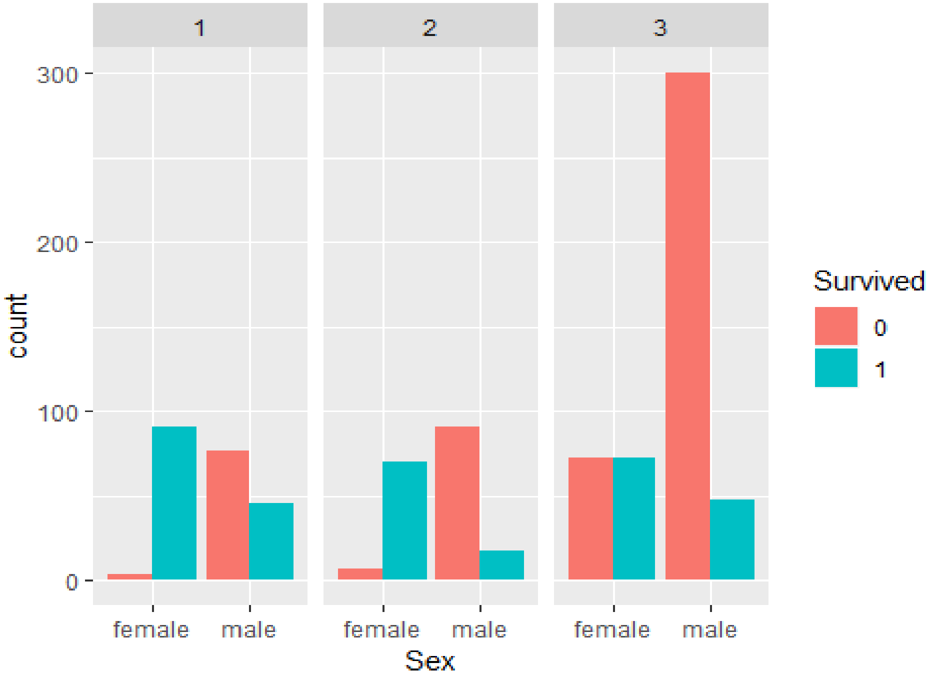
7. Survival rate by gender, but split barplot by Passenger class

```
ggplot( data = titanic.df, aes( x = Sex, fill = Survived ) ) +
  geom_bar( position = "fill" ) +
  facet_wrap( ~ Pclass )
```



8. Survival rate by gender, but split barplot by Passenger class, bars next to each other

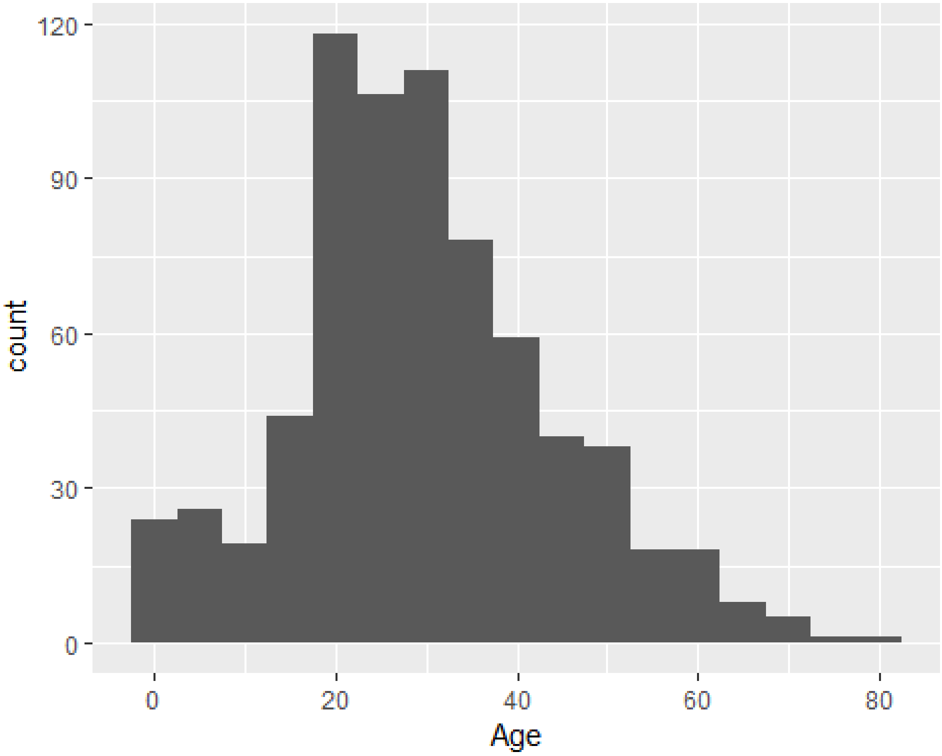
```
ggplot( data = titanic.df, aes( x = Sex, fill = Survived ) ) +
  geom_bar( position = "dodge" ) +
  facet_wrap( ~ Pclass )
```



Histograms, boxplots, density plots

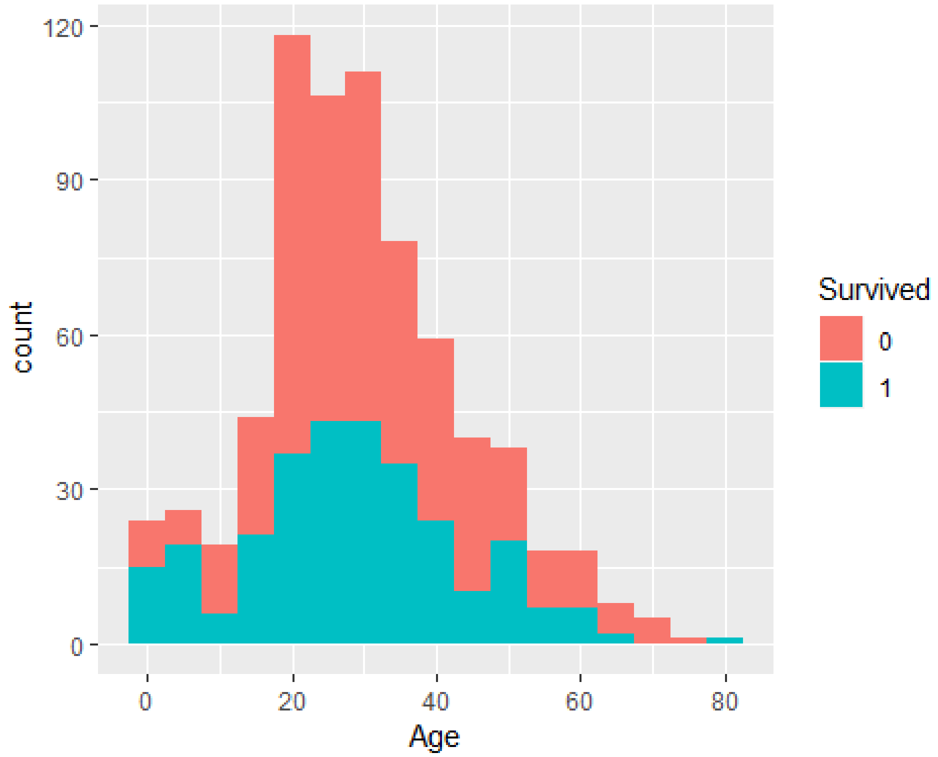
1. Age distribution

```
ggplot( data = titanic.df, aes( x = Age ) ) +
  geom_histogram( binwidth = 5 )

## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



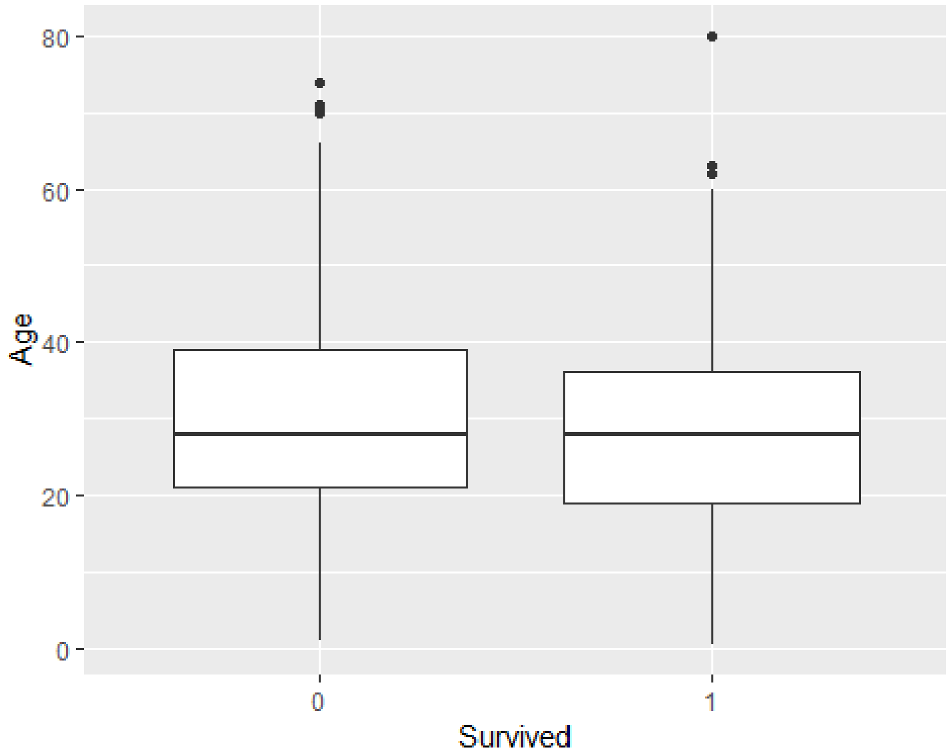
2. Age distribution by survival

```
ggplot( data = titanic.df, aes( x = Age, fill = Survived ) ) +
  geom_histogram( binwidth = 5 )

## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



3. Age distribution by survival using boxplots

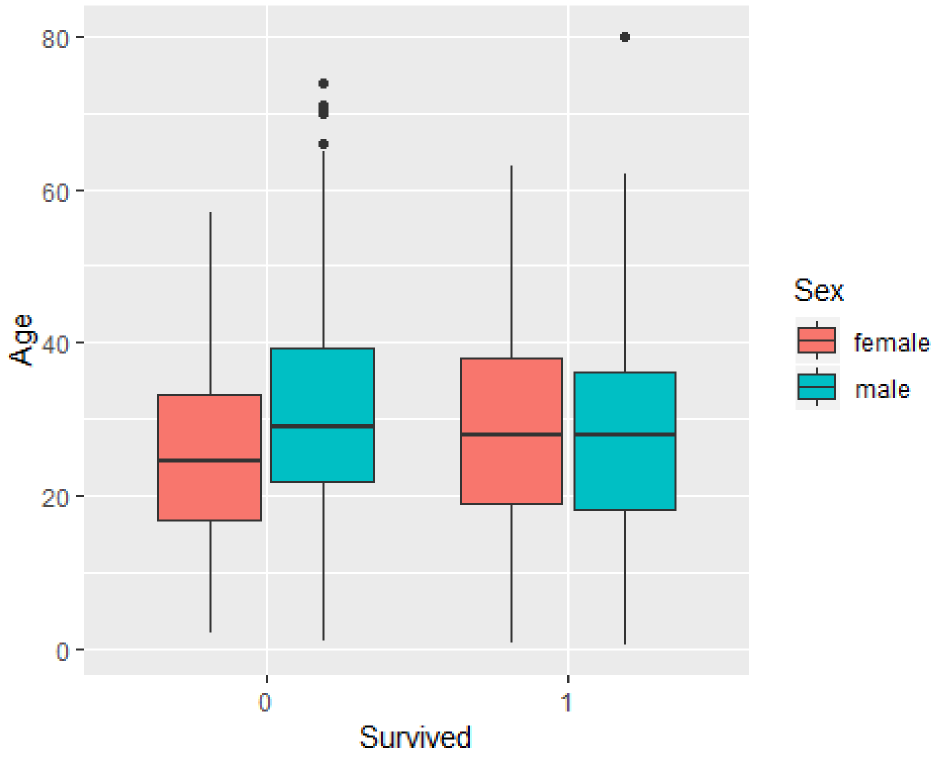
```
ggplot( data = titanic.df, aes( x = Survived, y = Age ) ) +
  geom_boxplot()

## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```



4. Age distribution by survival and gender using boxplots

```
ggplot( data = titanic.df, aes( x = Survived, y = Age, fill = Sex ) ) +
  geom_boxplot()
```

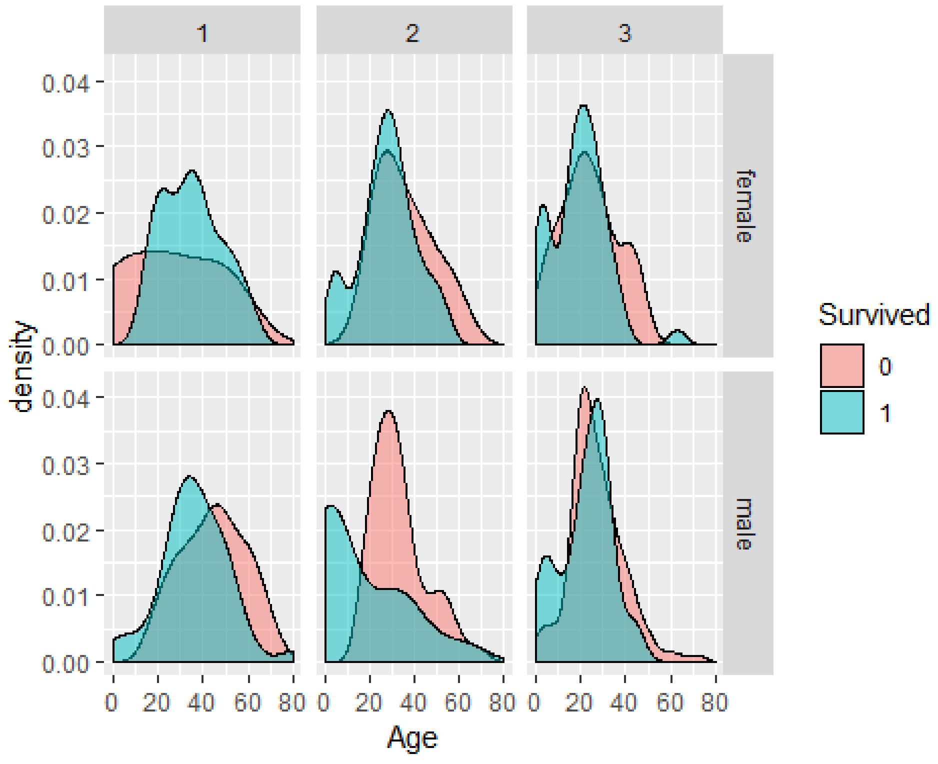
```
## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```



5. Age distribution by survival, gender and passenger class using transparent density plots

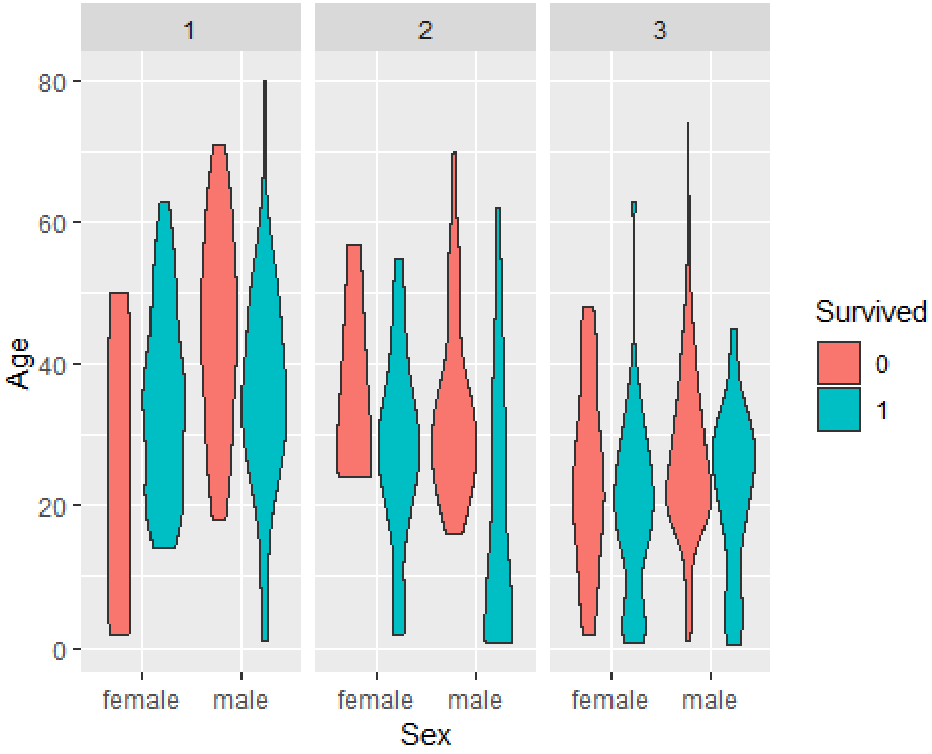
```
ggplot( data = titanic.df, aes( x = Age, fill = Survived ) ) +
  facet_grid( Sex ~ Pclass ) +
  geom_density( alpha = 0.5 )
```

```
## Warning: Removed 177 rows containing non-finite values (stat_density).
```



6. Age distribution by survival, gender and passenger class using violin plots

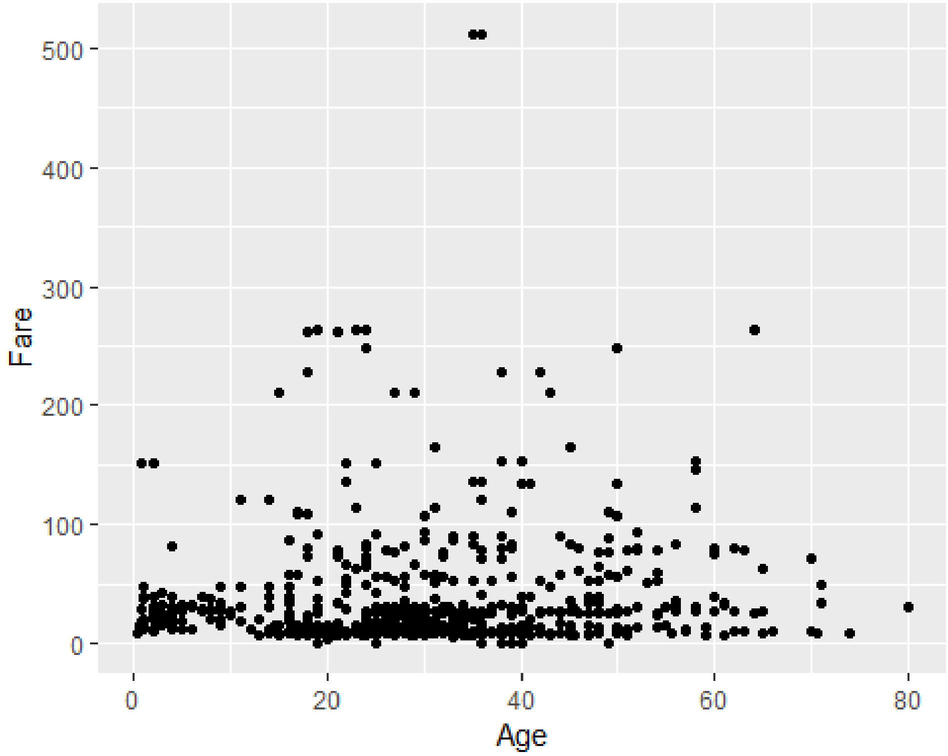
```
ggplot( data = titanic.df, aes( x = Sex, y = Age, fill = Survived ) ) +
  facet_grid( ~ Pclass ) +
  geom_violin()

## Warning: Removed 177 rows containing non-finite values (stat_ydensity).
```



Scatter plots

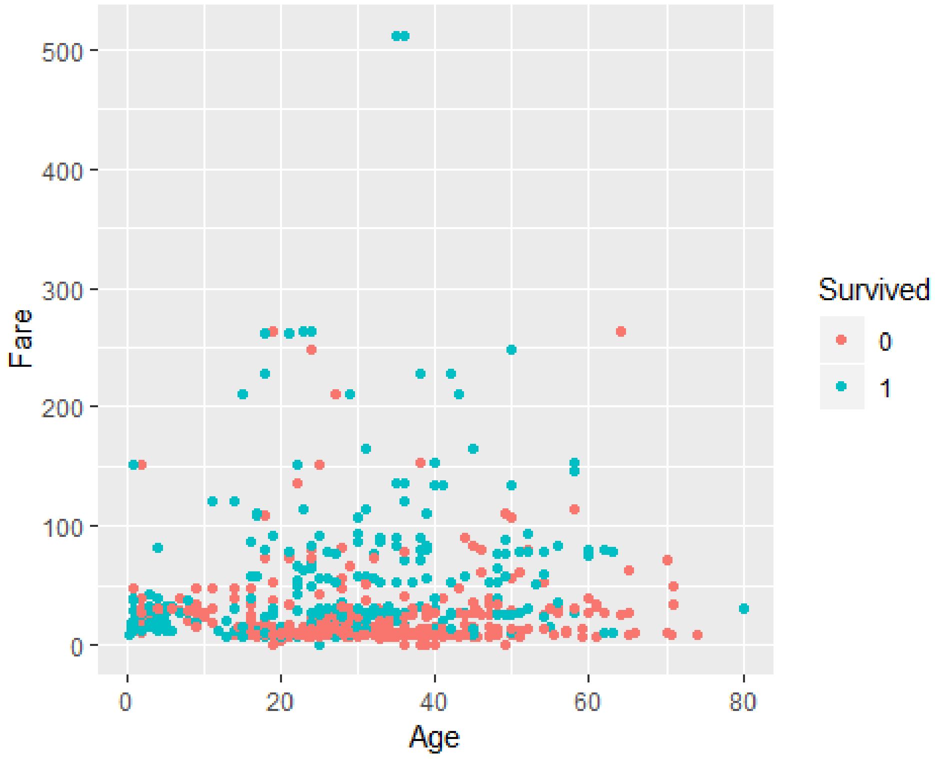
1. Age vs fare

```
ggplot( data = titanic.df, aes( x = Age, y = Fare ) ) +
  geom_point()

## Warning: Removed 177 rows containing missing values (geom_point).
```



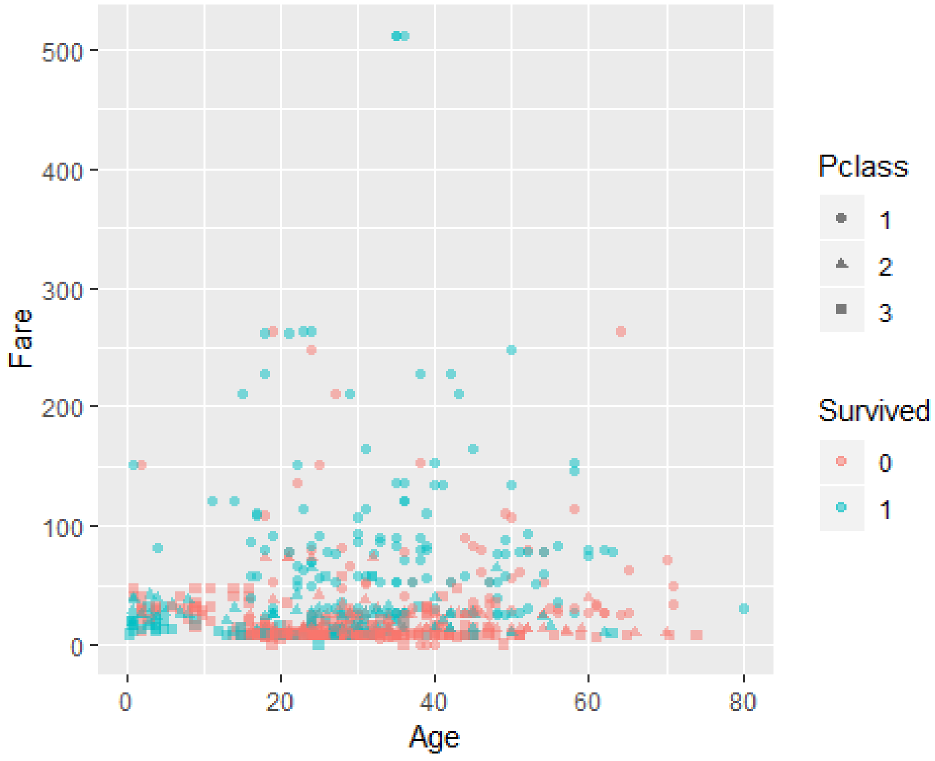
2. Age vs fare, color by survival

```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived ) ) +
  geom_point()

## Warning: Removed 177 rows containing missing values (geom_point).
```



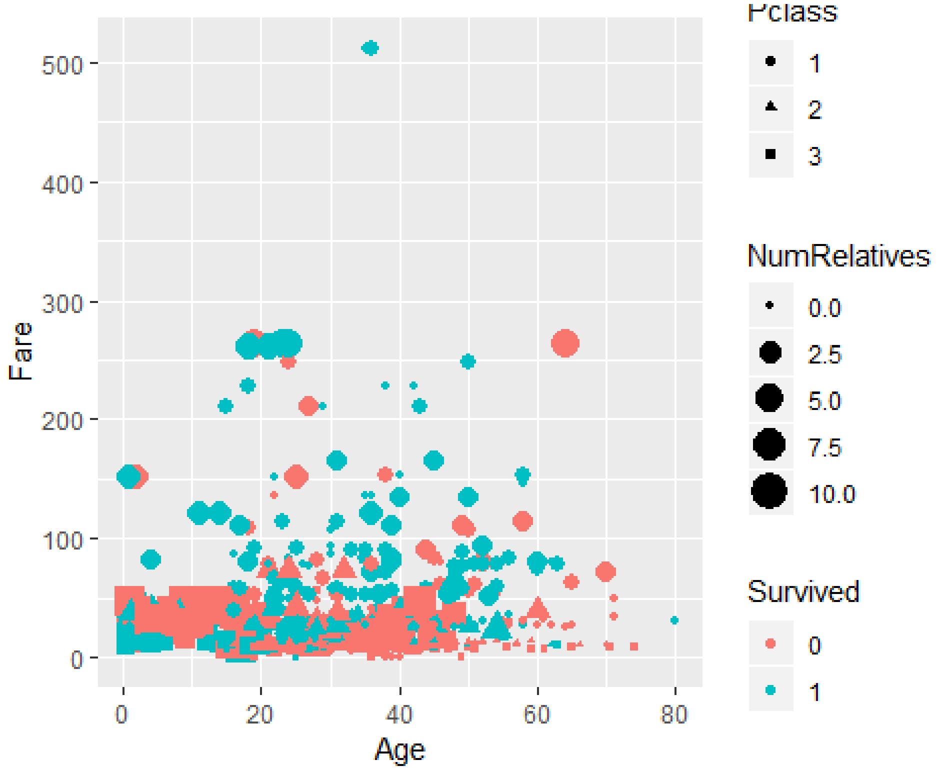
3. Age vs fare, color by survival, shape by Pclass

```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, shape =
Pclass ) ) +
  geom_point( alpha = 0.5 )

## Warning: Removed 177 rows containing missing values (geom_point).
```



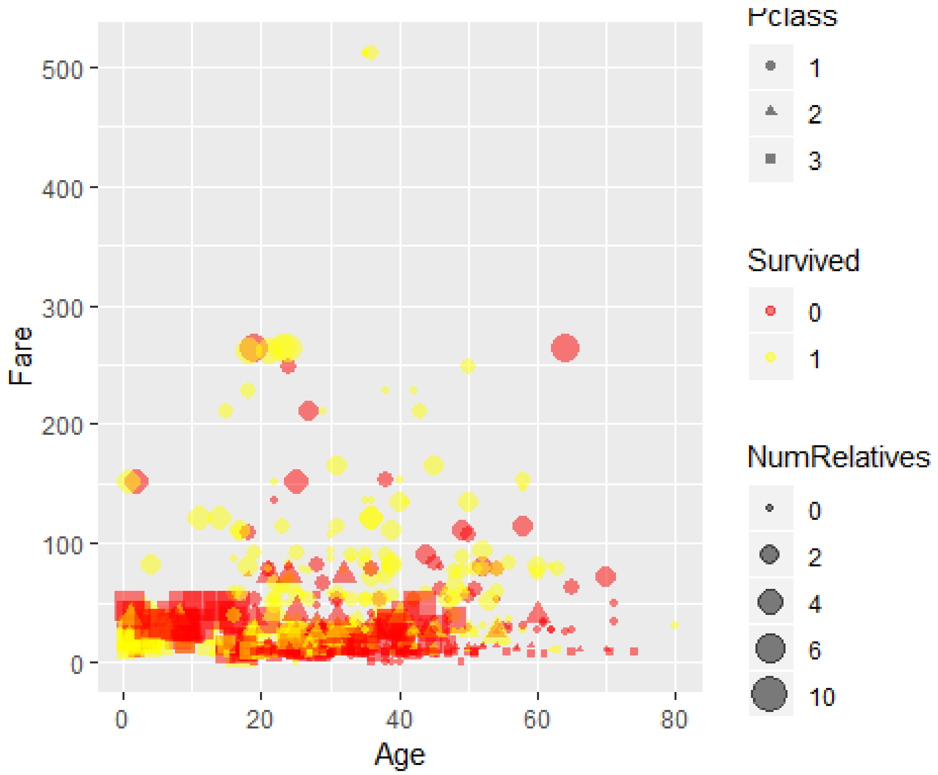
4. Age vs fare, color by survival, shape by Pclass, size by number of relatives

```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, shape =
Pclass, size = NumRelatives ) ) +
  geom_point()

## Warning: Removed 177 rows containing missing values (geom_point).
```



5. Age vs fare, color by survival, shape by Pclass, size by number of relatives, semi-transparent points

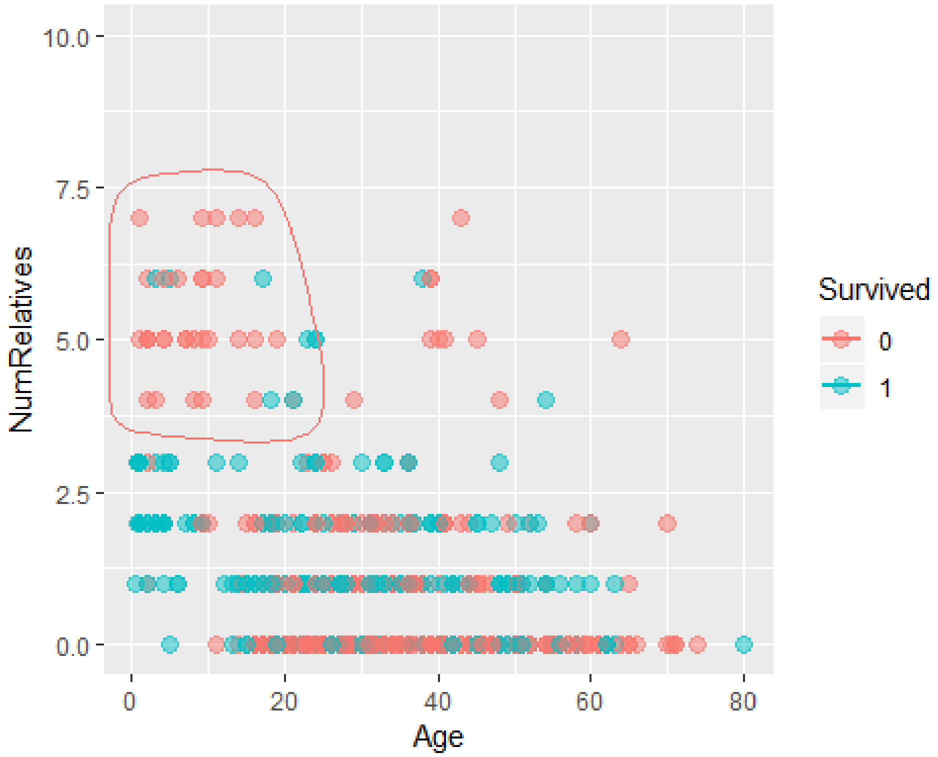
```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, shape =
Pclass, size = NumRelatives ) ) +
  geom_point( alpha = 0.5 ) +
  scale_size_continuous( breaks = c(0,2,4,6,10) ) +
  scale_color_manual(values = c("red","yellow"))

## Warning: Removed 177 rows containing missing values (geom_point).
```



6. Age vs number of relatives, color by survival, semi-transparent points

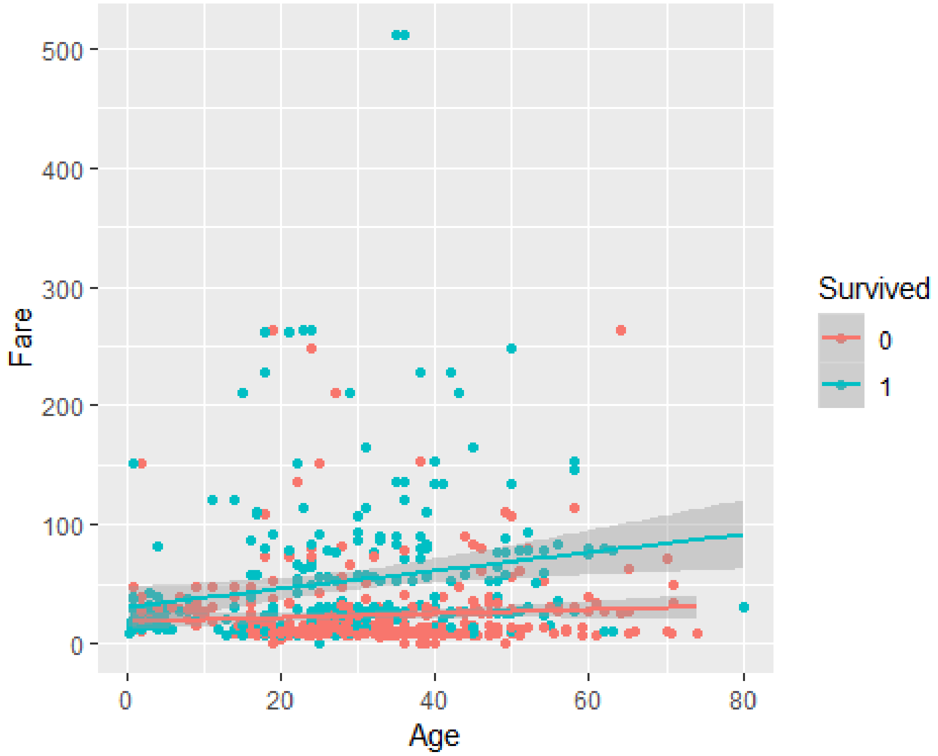
```
# devtools::install_github( "hrbrmstr/ggalt" )
library(ggalt)

ggplot( data = titanic.df, aes( x = Age, y = NumRelatives, color = Survived )
) +
  geom_point( alpha = 0.5, size = 3 ) +
  geom_encircle( data = subset( titanic.df,
                               Age < 25 & NumRelatives >= 4 & Survived == 0
),
                 aes( x = Age, y = NumRelatives ) )
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
```



7. Age vs fare, color by survival, smoothed estimator

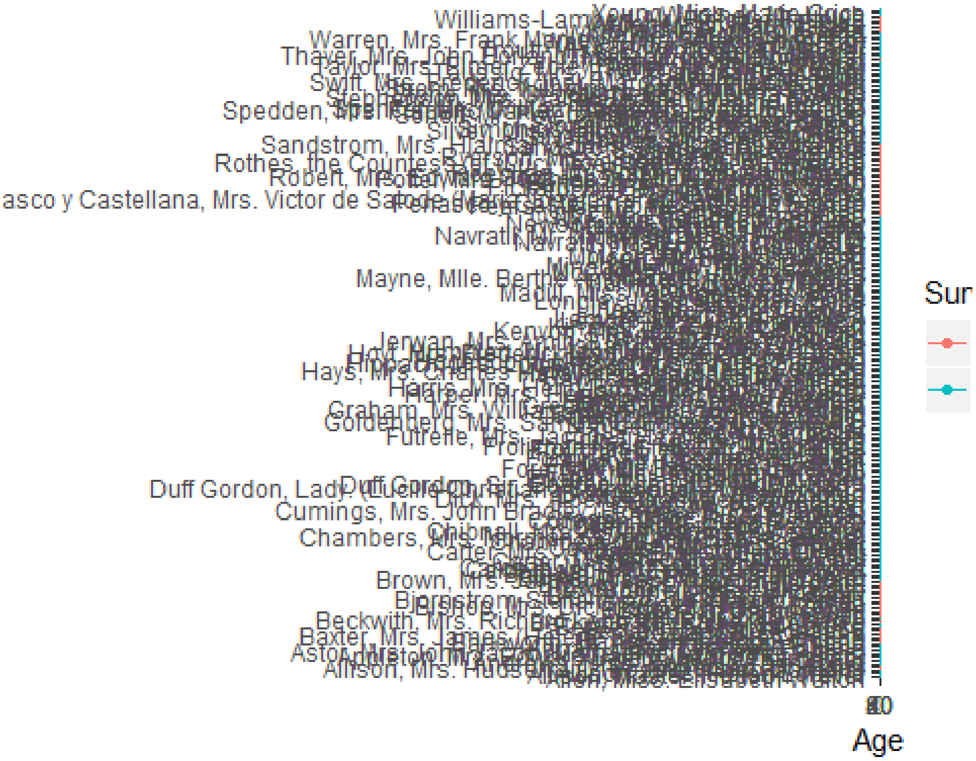
```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived ) ) +
  geom_point() +
  geom_smooth( method="lm")  #

## Warning: Removed 177 rows containing non-finite values (stat_smooth).

## Warning: Removed 177 rows containing missing values (geom_point).
```



Line plots

1. connect people by Cabin numbers

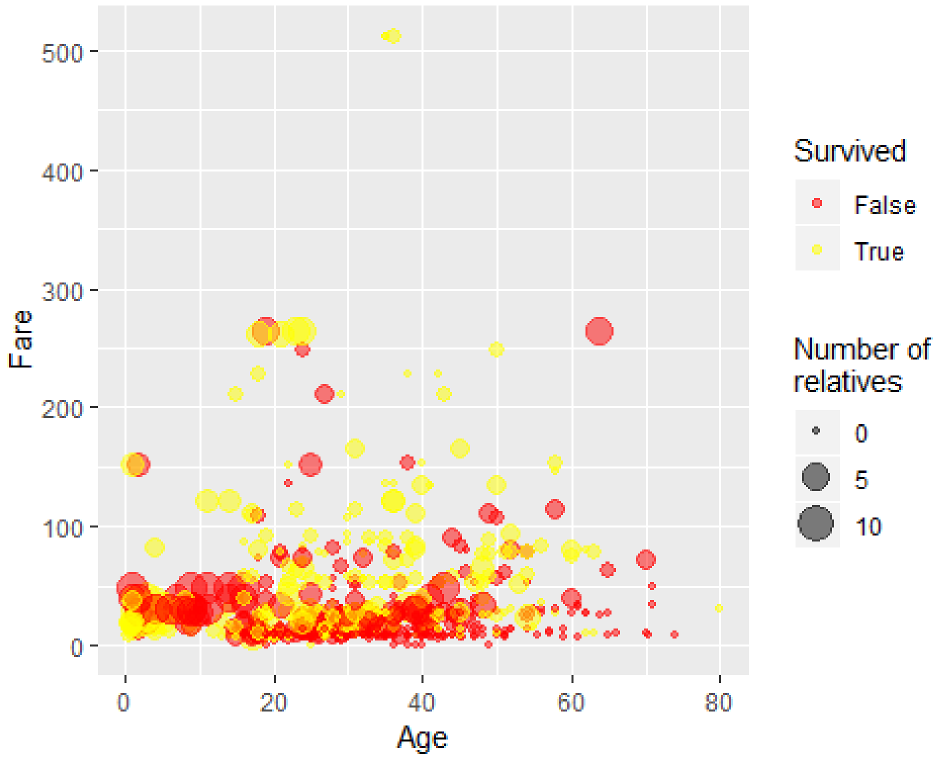
```
ggplot( data = subset( titanic.df, Cabin != "" ),
        aes( x = Age, y = Name, color = Survived ) ) +
  geom_point() +
  geom_line( aes( group = Cabin ) )

## Warning: Removed 19 rows containing missing values (geom_point).

## Warning: Removed 19 rows containing missing values (geom_path).
```



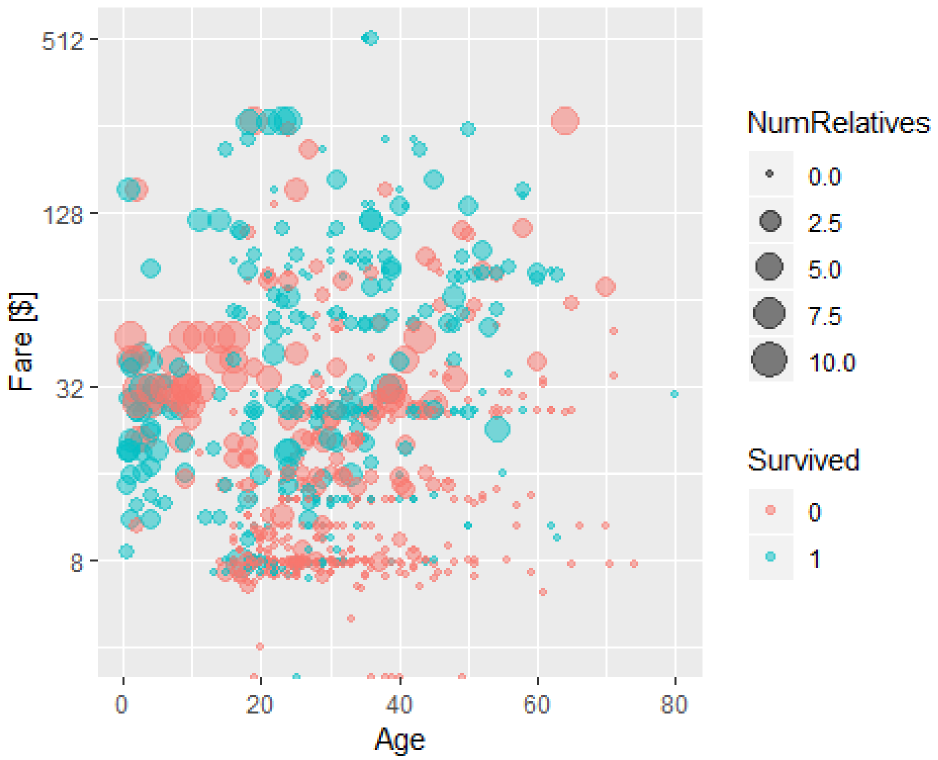
Fine tuning

Scales (color, fill, size, shape, linetype)

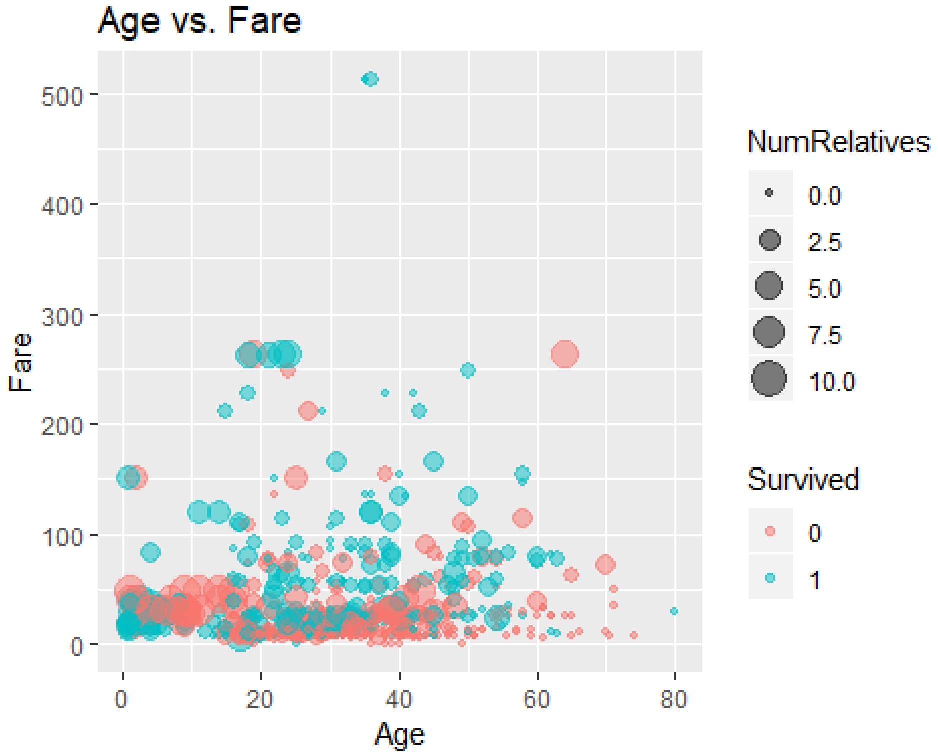
```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, size = N
umRelatives ) ) +
  geom_point( alpha = 0.5 ) +
  scale_size_continuous( breaks = c(0,5,10),
                         name = "Number of\nrelatives" ) +
  scale_color_manual( labels = c("False", "True"), values = c("red","yellow")
)

## Warning: Removed 177 rows containing missing values (geom_point).
```



```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, size = N
umRelatives ) ) +
  geom_point( alpha = 0.5 ) +
  scale_y_continuous( trans = "log2" ) +
  labs( y = "Fare [$]" )

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 177 rows containing missing values (geom_point).
```



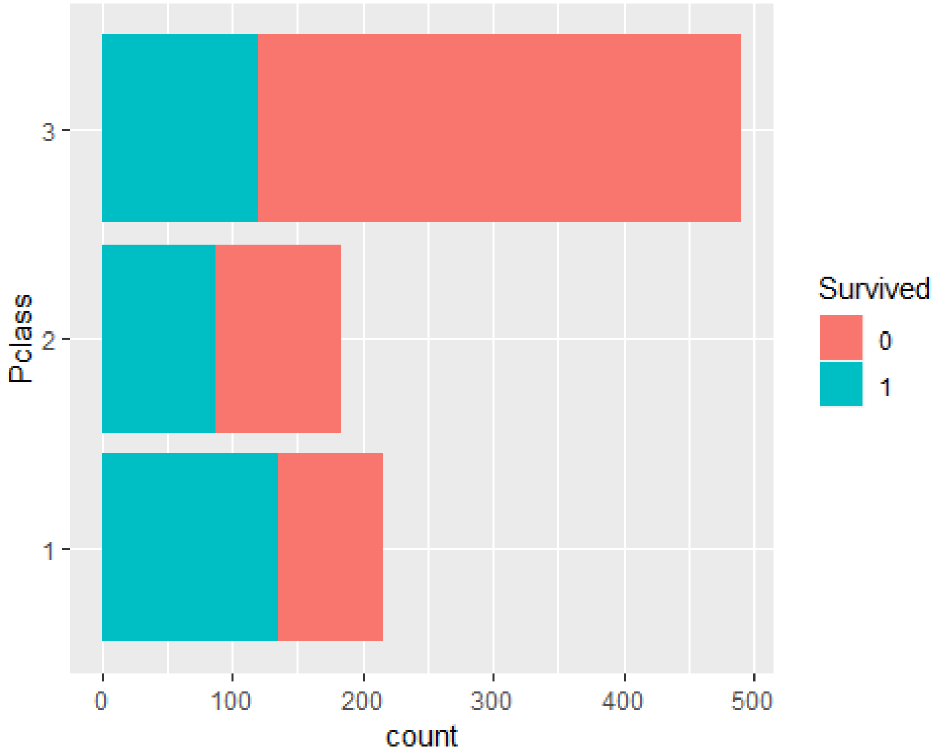
```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, size = N
umRelatives ) ) +
  geom_point( alpha = 0.5 ) +
  labs( title = "Age vs. Fare" )

## Warning: Removed 177 rows containing missing values (geom_point).
```



Coordinates

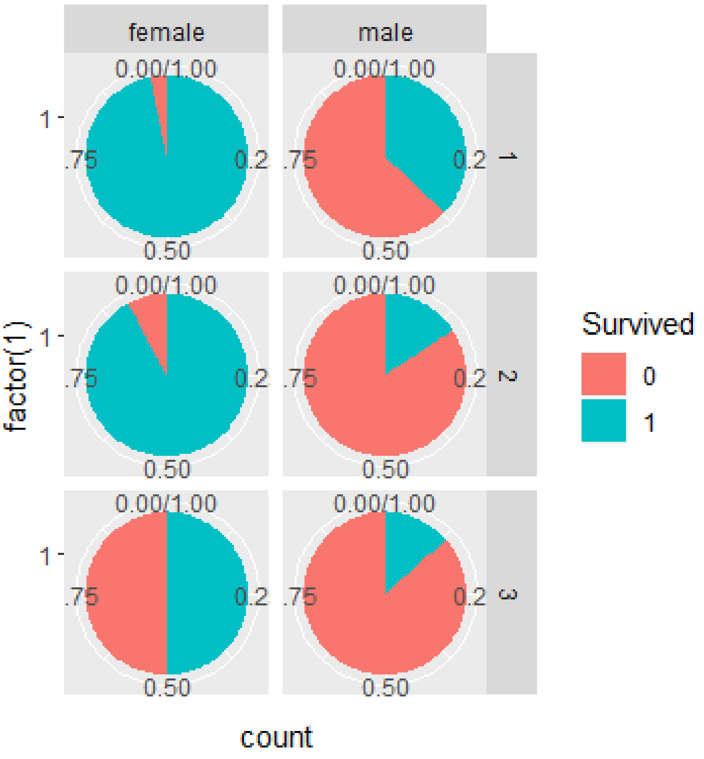
Flip

```
ggplot( data = titanic.df, aes( x = Pclass, fill = Survived ) ) +
  geom_bar() +
  coord_flip()
```



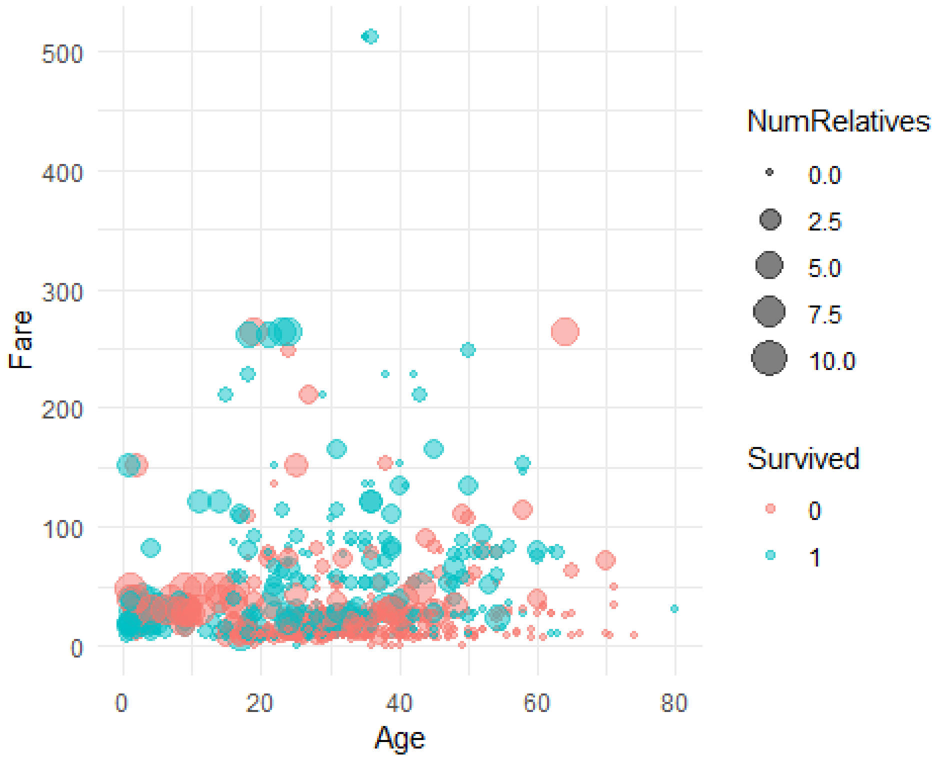
Pie chart

```
ggplot( data = titanic.df, aes( x = factor(1), fill = Survived ) ) +
  geom_bar( position="fill" ) +
  facet_grid( Pclass ~ Sex ) +
  coord_polar( theta = "y" )
```



Themes

```
ggplot( data = titanic.df, aes( x = Age, y = Fare, color = Survived, size = N
umRelatives ) ) +
  geom_point( alpha = 0.5 ) +
  theme_minimal() #

## Warning: Removed 177 rows containing missing values (geom_point).
```


Fare
Age
NumRelatives
Survived