고려대학교
빅데이터 연구회
KU-BIG

EDA(데이터 시각화)

목차

1 EDA

2 DieTanic 커널 소개

3 시각화 - ggplot 소개

4 Tietanic data 시각화- ggplot 활용

“

‘탐색적 데이터 분석(EDA)’은 우리가 존재한다고 믿는 것들은 물론이고 존재하지 않는다고 믿는 것들을 발견하려는 태도, 유연성, 그리고 자발성이다. - Schutt Rachel(Doing Data Science의 저자)

”

EDA

1.Maximize Insight into a data set

2.Uncover Underlying Structure

3.Extract Important Variables

4.Detect Outliers and Anomalies

5.Test Underlying Assumptions

6.Develop Parsimonious Models

7.Determine Optimal Factor Settings

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

EDA

1. 데이터 분석 목적
2. 데이터 구조 및 변수 확인
3. 변수별 type/분포 확인
4. 변수의 수준별 분포 비교
5. Insight 도출

DieTanic

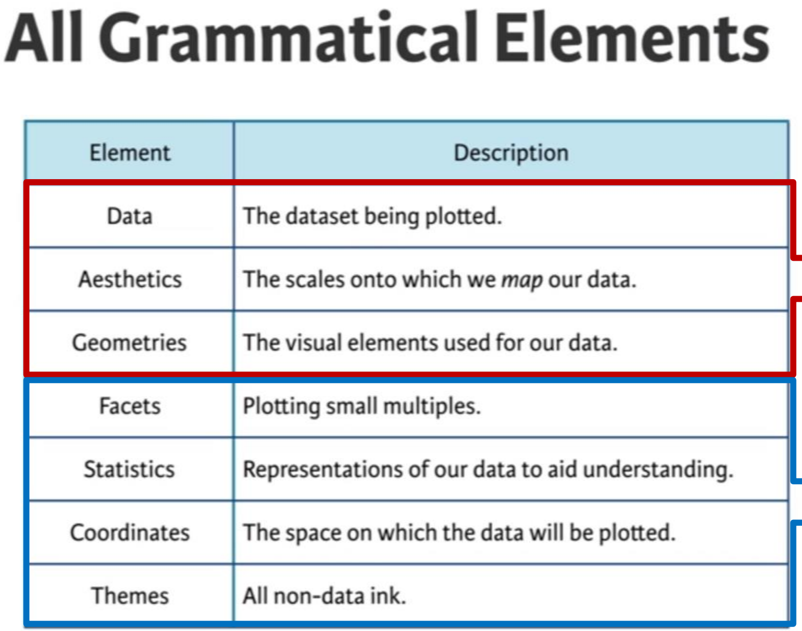
https://www.kaggle.com/ash316/eda-to-prediction-dietanic

Contents of the Notebook:

Part1: Exploratory Data Analysis(EDA):
1)Analysis of the features.
2)Finding any relations or trends considering multiple features.

Part2: Feature Engineering and Data Cleaning:
1)Adding any few features.
2)Removing redundant features.
3)Converting features into suitable form for modeling.

데이터시각화 - ggplot



All Grammatical Elements

Element	Description
Data	The dataset being plotted.
Aesthetics	The scales onto which we *map* our data.
Geometries	The visual elements used for our data.
Facets	Plotting small multiples.
Statistics	Representations of our data to aid understanding.
Coordinates	The space on which the data will be plotted.
Themes	All non-data ink.

3개의 필수 Layer

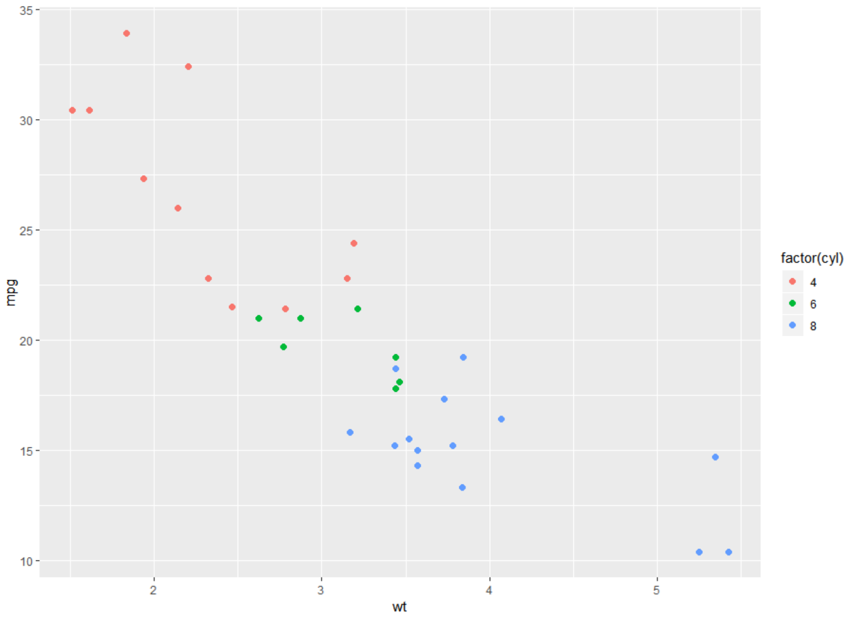
```
ggplot(data, aes()) +
  geom_***()
```

4개의 부가 Layer

```
ggplot(data, aes()) +
  geom_***() +
  facet_***() +
  stat_***() +
  coord_***() +
  theme_***()
```

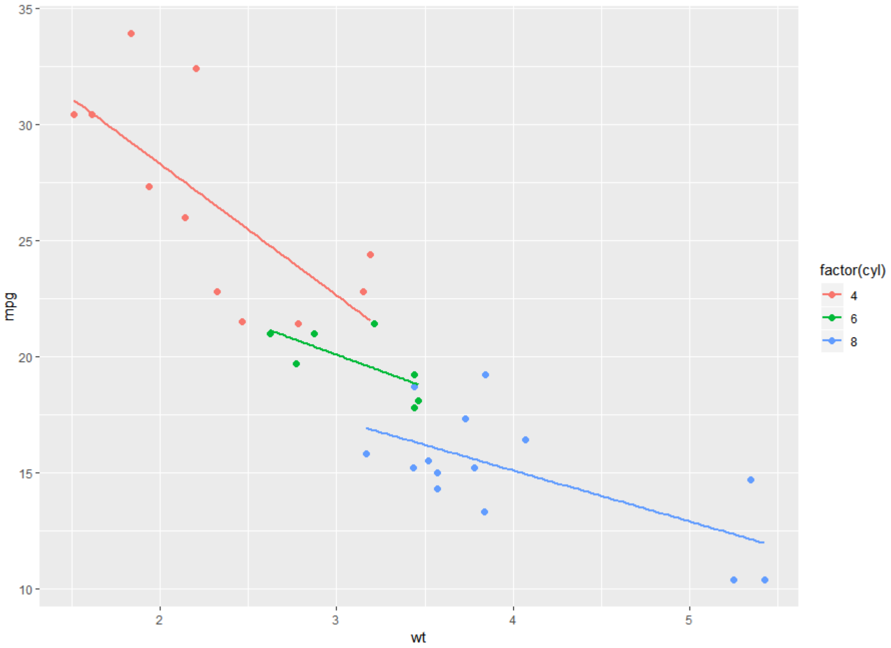
데이터시각화 - ggplot; Data, Aesthetics, Geometries

```
install.packages(“ggplot2)
library(ggplot2)
ggplot(mtcars, aes(x=wt, y=mpg, col=factor(cyl)) +
geom_point() +
geom_smooth(method=“lm”, se = F) +
facet_grid(.~gear) +
```



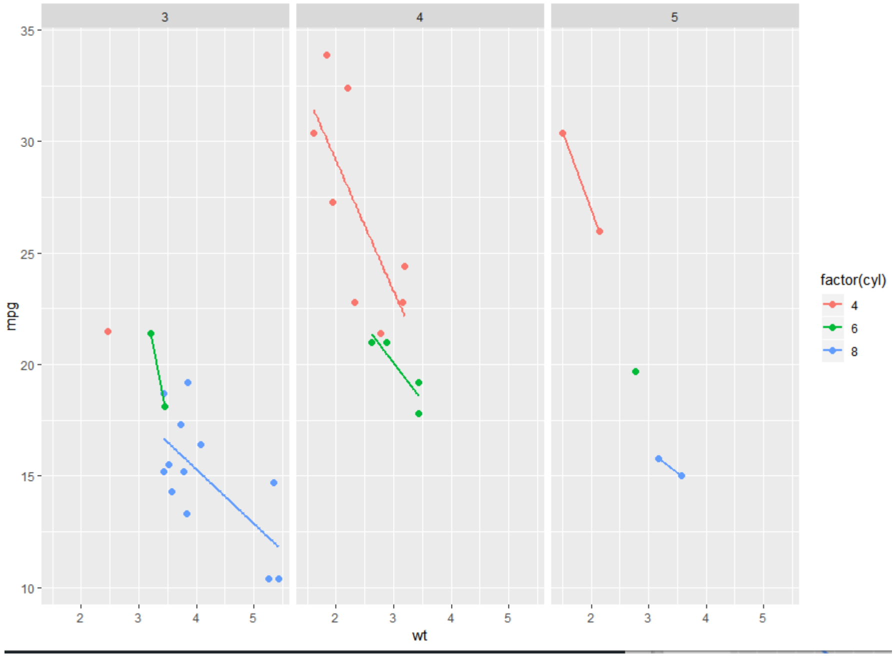
데이터시각화 - ggplot; Geometries

```
install.packages(“ggplot2)
library(ggplot2)
ggplot(mtcars, aes(x=wt, y=mpg, col=factor(cyl)) +
geom_point() +
geom_smooth(method=“lm”, se = F) +
facet_grid(.~gear)
```



데이터시각화 – ggplot; Facets

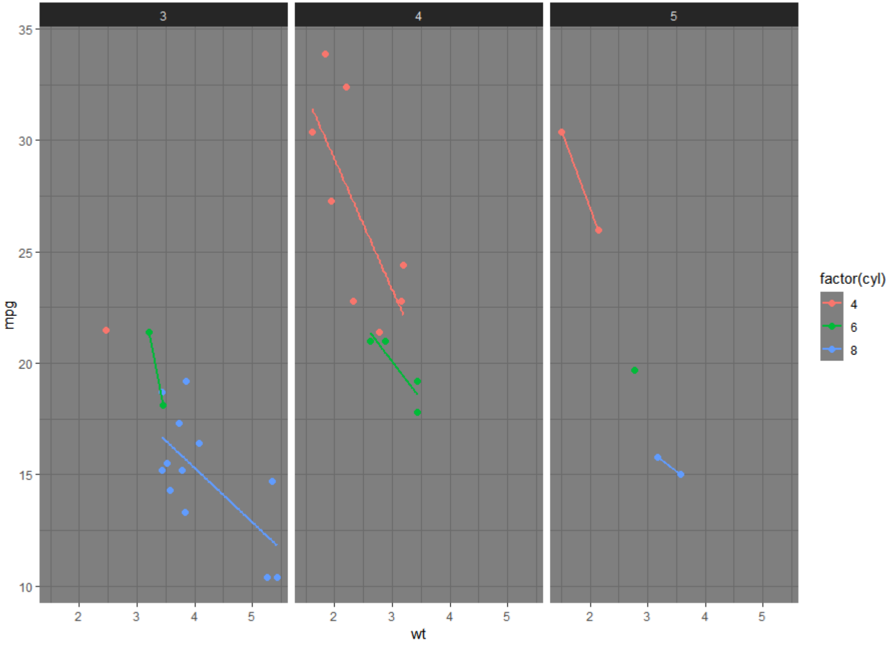
```
install.packages(“ggplot2)
library(ggplot2)
ggplot(mtcars, aes(x=wt, y=mpg, col=factor(cyl)) +
geom_point() +
geom_smooth(method=“lm”, se = F) +
facet_grid(.~gear)
```



데이터시각화 - ggplot; Themes

```
install.packages(“ggplot2)
library(ggplot2)
p = ggplot(mtcars, aes(x=wt, y=mpg, col=factor(cyl)) +
              geom_point() +
              geom_smooth(method=“lm”, se = F) +
              facet_grid(.~gear)
```

```
install.packages(“ggthemes”)
library(ggthemes)
p + theme_dark()
```



데이터시각화 - graph types

Univariate

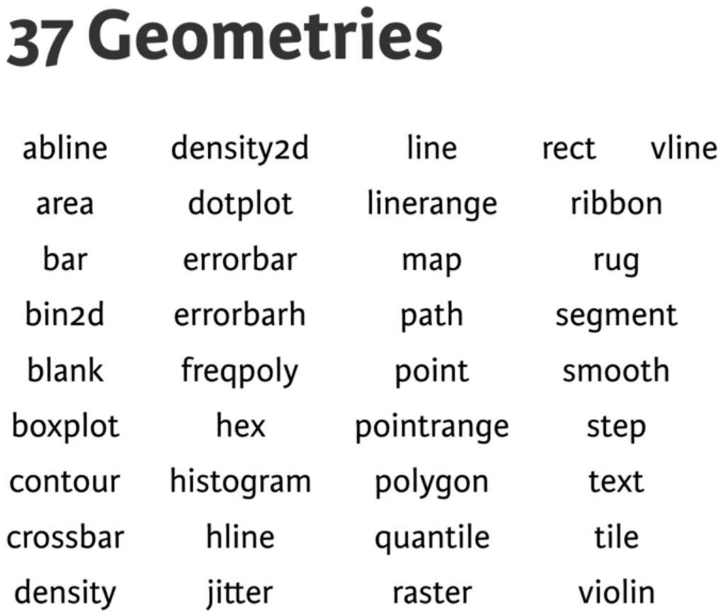
Discrete/Categorical - **bar graph, pie graph**

Continuous - **histogram, KDE, box graph**

Multivariate

Discrete/Categorical - **mosaic graph**

Continuous - **scatterplot**

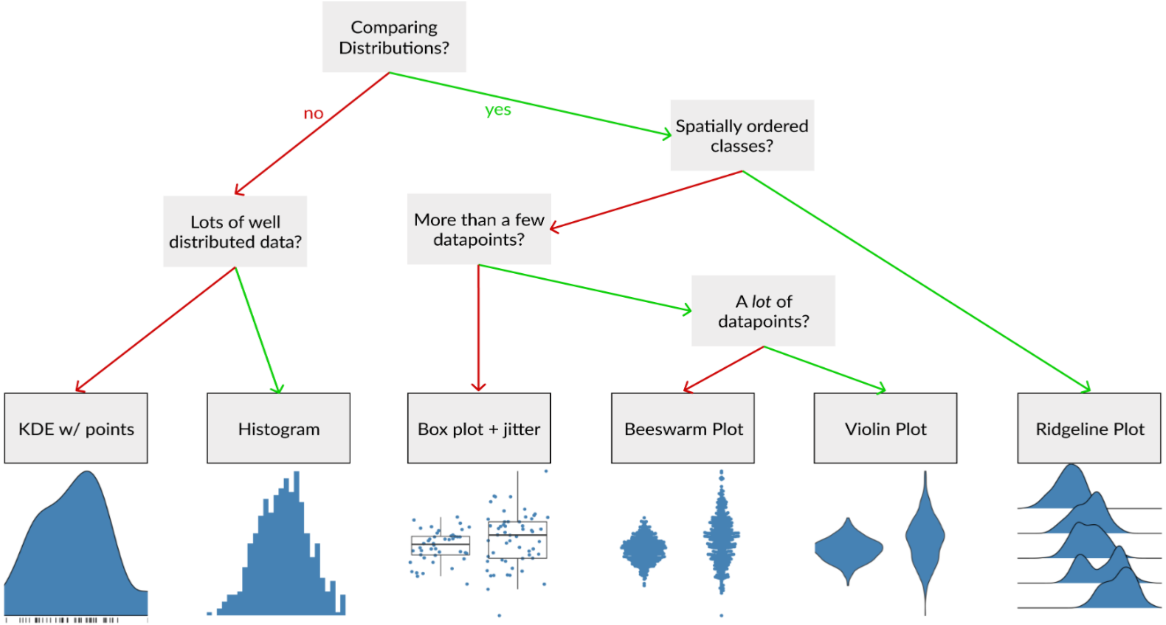


37 Geometries

abline
area
bar
bin2d
blank
boxplot
contour
crossbar
density
density2d
dotplot
errorbar
errorbarh
freqpoly
hex
histogram
hline
jitter
line
linerange
map
path
point
pointrange
polygon
quantile
raster
rect
ribbon
rug
segment
smooth
step
text
tile
violin
vline

데이터시각화 - graph types

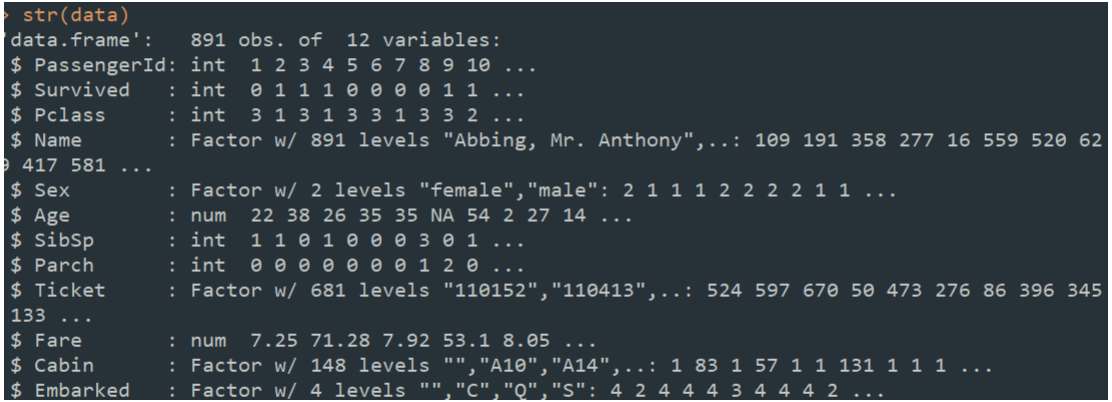
Overview of distribution visualizations



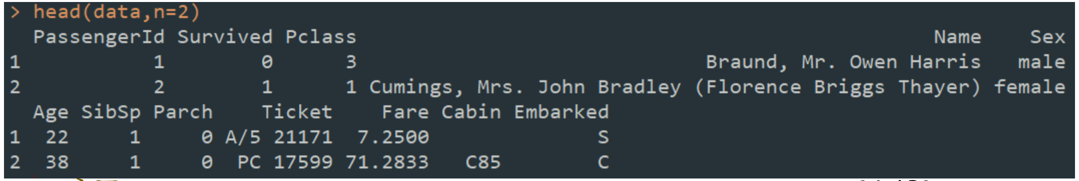
ggplot을 활용한 Tietanic data 시각화

1. 데이터 분석 목적 : Survival Prediction

2. 데이터 구조 및 변수 확인



```
> str(data)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 62
9 417 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345
 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```



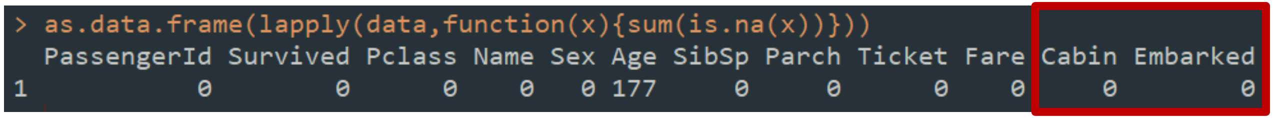
```
> head(data,n=2)
  PassengerId Survived Pclass                                                Name    Sex
1           1        0      3                             Braund, Mr. Owen Harris   male
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
  Age SibSp Parch    Ticket    Fare Cabin Embarked
1  22     1     0 A/5 21171  7.2500              S
2  38     1     0  PC 17599 71.2833   C85        C
```

Variable	Definition	Key
Sibsp	# of Siblings or Spouses aboard the Titanic	
Parch	# of parents or children aboard the Titanic	

Embarked	Port of Embarkation	C = Cherbourg,
		Q = Queenstown,
		S = Southampton

ggplot을 활용한 Tietanic data 시각화

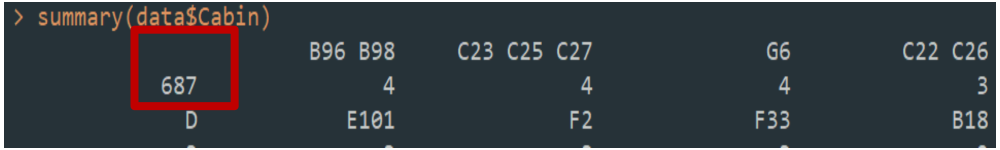
2. 데이터 구조 및 변수 확인 ; 변수별 결측치(이상치) 확인



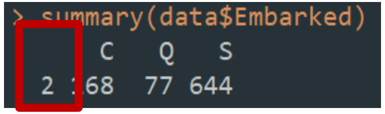
```
> as.data.frame(lapply(data,function(x){sum(is.na(x))}))
  PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare Cabin Embarked
1           0        0      0    0   0 177     0     0      0    0     0        0
```

실제로는 결측치 존재!

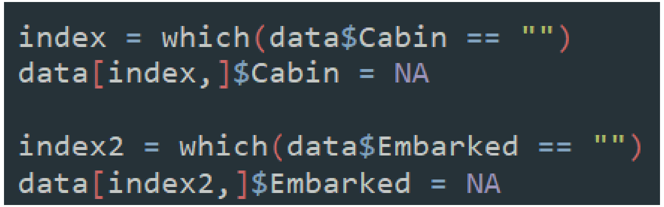
Raw Data도 항상 확인해야함



```
> summary(data$Cabin)
                    B96 B98     C23 C25 C27              G6     C22 C26
            687           4               4               4           3
              D        E101              F2             F33         B18
```



```
> summary(data$Embarked)
      C   Q   S
  2 168  77 644
```



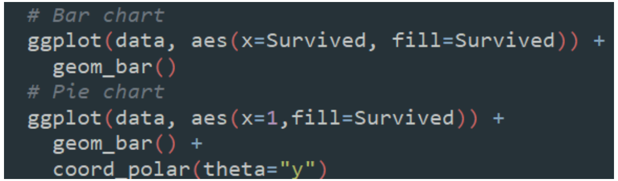
```
index = which(data$Cabin == "")
data[index,]$Cabin = NA

index2 = which(data$Embarked == "")
data[index2,]$Embarked = NA
```

ggplot을 활용한 Tietanic data 시각화

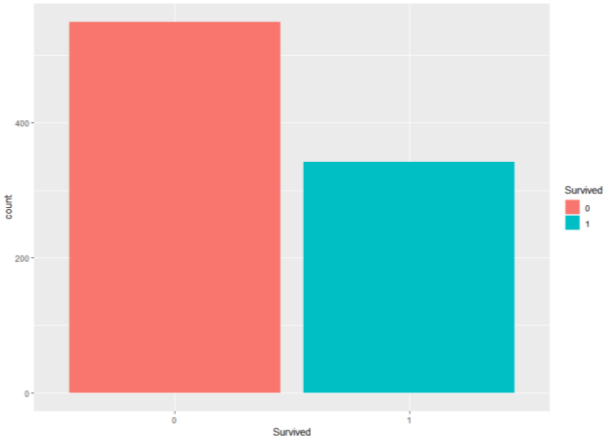
3. 변수별 type/분포 확인 ; Survived (Categorical)

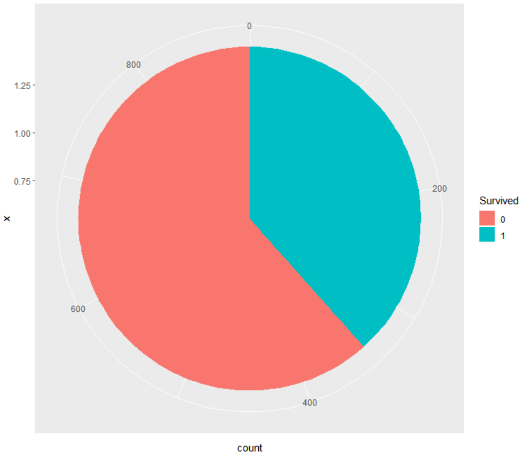
- 수치형(Int) 변수 -> 범주형으로 바꾸기

- Pie Chart, Bar Chart로 분포 확인



```
# Bar chart
ggplot(data, aes(x=Survived, fill=Survived)) +
  geom_bar()
# Pie chart
ggplot(data, aes(x=1,fill=Survived)) +
  geom_bar() +
  coord_polar(theta="y")
```

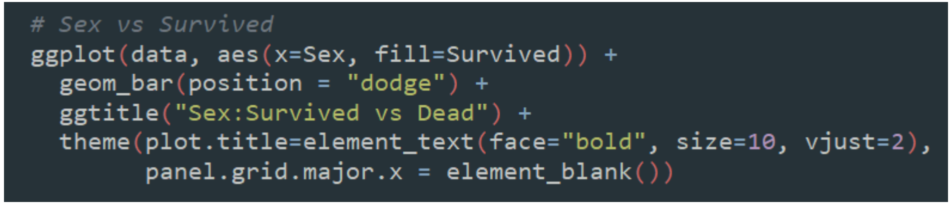




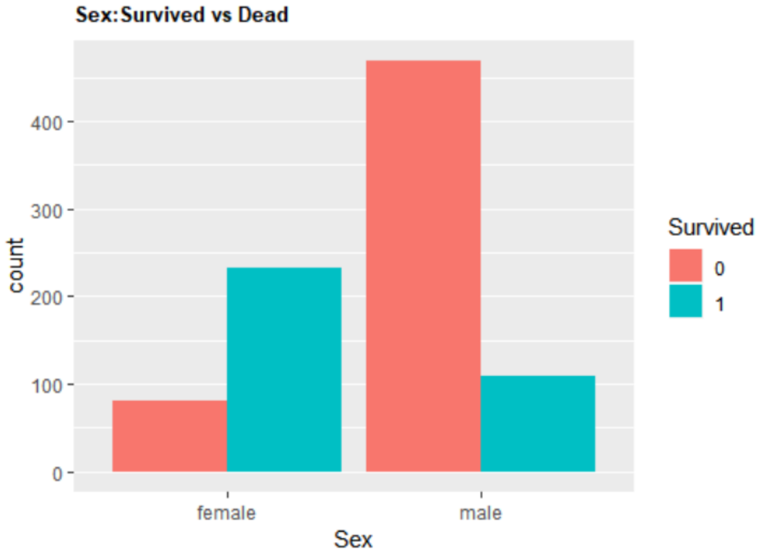
ggplot을 활용한 Tietanic data 시각화

3. 변수별 type/분포 확인 ; Sex (Categorical)

- Sex vs Survived



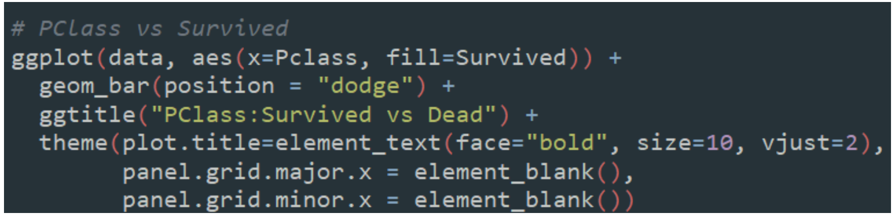
```
# Sex vs Survived
ggplot(data, aes(x=Sex, fill=Survived)) +
  geom_bar(position = "dodge") +
  ggtitle("Sex:Survived vs Dead") +
  theme(plot.title=element_text(face="bold", size=10, vjust=2),
        panel.grid.major.x = element_blank())
```



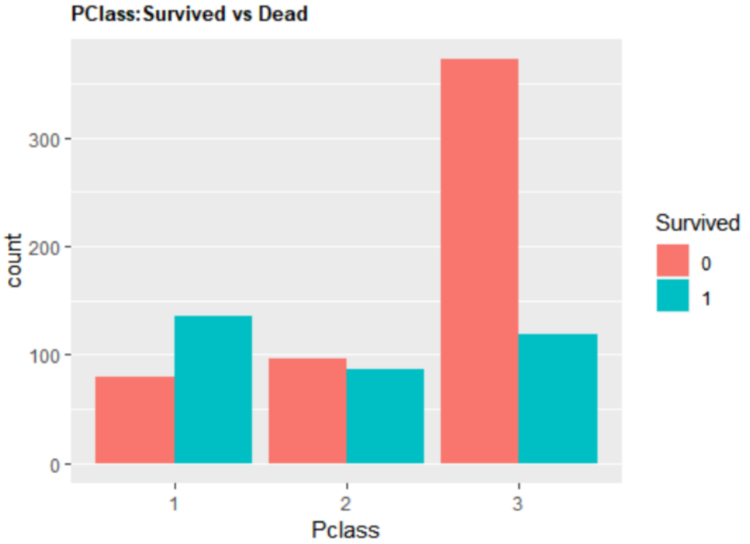
ggplot을 활용한 Tietanic data 시각화

3. 변수별 type/분포 확인 ; Pclass (Ordinal)

- Pcalss vs Survived



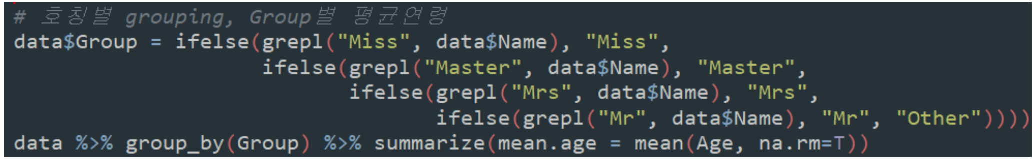
```
# PClass vs Survived
ggplot(data, aes(x=Pclass, fill=Survived)) +
  geom_bar(position = "dodge") +
  ggtitle("PClass:Survived vs Dead") +
  theme(plot.title=element_text(face="bold", size=10, vjust=2),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())
```



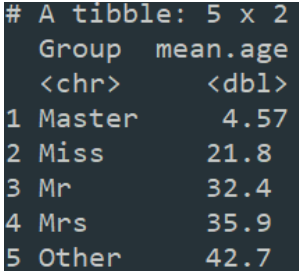
ggplot을 활용한 Tietanic data 시각화

3. 변수별 type/분포 확인 ; Age (Continuous)

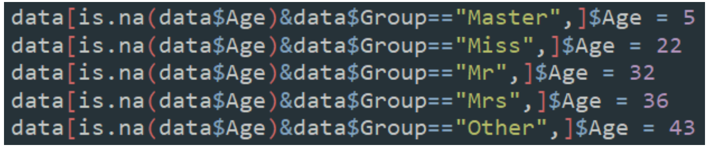
-결측치 -> 호칭별 Group의 평균연령으로 대체



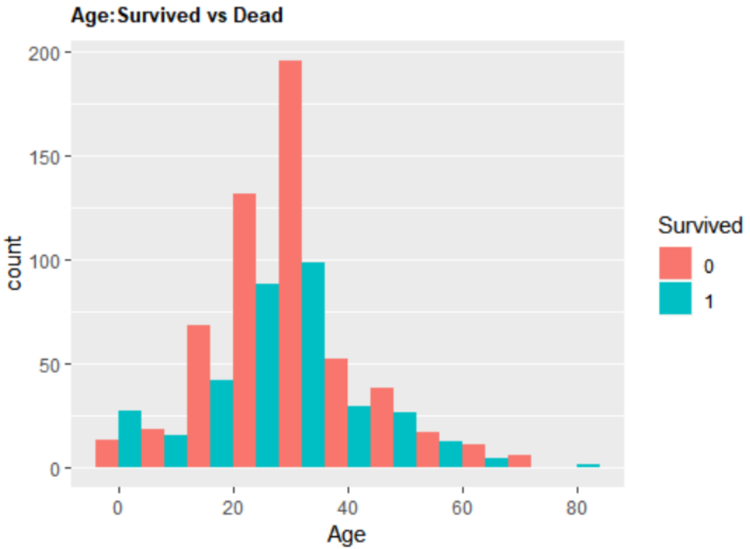
```
# 호칭별 grouping, Group별 평균연령
data$Group = ifelse(grepl("Miss", data$Name), "Miss",
                    ifelse(grepl("Master", data$Name), "Master",
                           ifelse(grepl("Mrs", data$Name), "Mrs",
                                  ifelse(grepl("Mr", data$Name), "Mr", "Other"))))
data %>% group_by(Group) %>% summarize(mean.age = mean(Age, na.rm=T))
```



```
# A tibble: 5 x 2
  Group   mean.age
  <chr>      <dbl>
1 Master      4.57
2 Miss       21.8
3 Mr         32.4
4 Mrs        35.9
5 Other      42.7
```



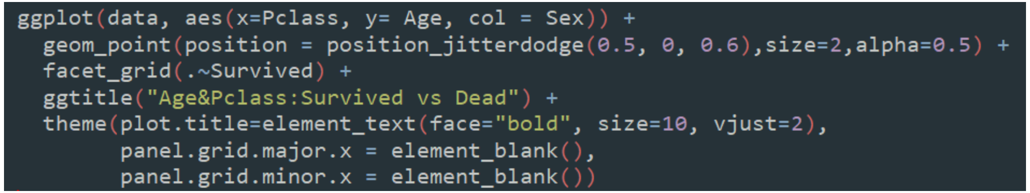
```
data[is.na(data$Age)&data$Group=="Master",]$Age = 5
data[is.na(data$Age)&data$Group=="Miss",]$Age = 22
data[is.na(data$Age)&data$Group=="Mr",]$Age = 32
data[is.na(data$Age)&data$Group=="Mrs",]$Age = 36
data[is.na(data$Age)&data$Group=="Other",]$Age = 43
```



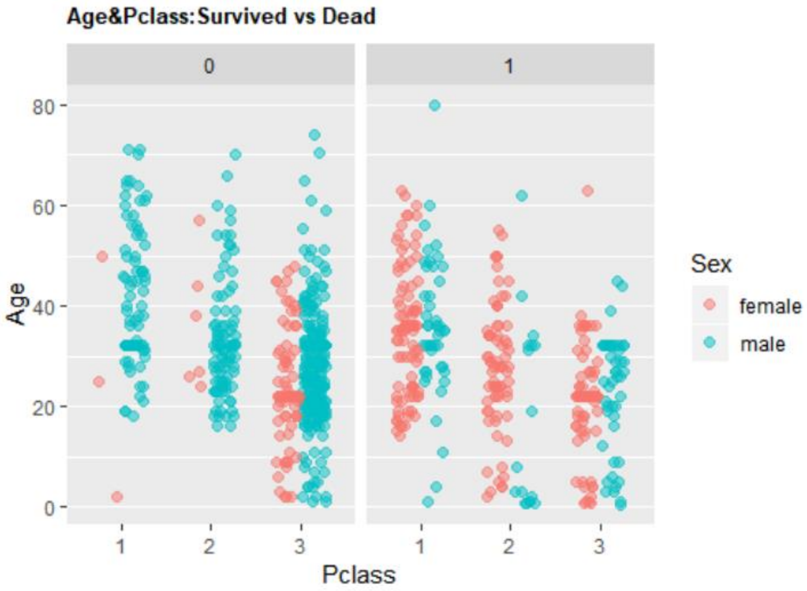
ggplot을 활용한 Tietanic data 시각화

4. 변수의 수준별 분포 비교

-x, y 축, fill, facet까지 4가지 변수를 한 plot에 표현

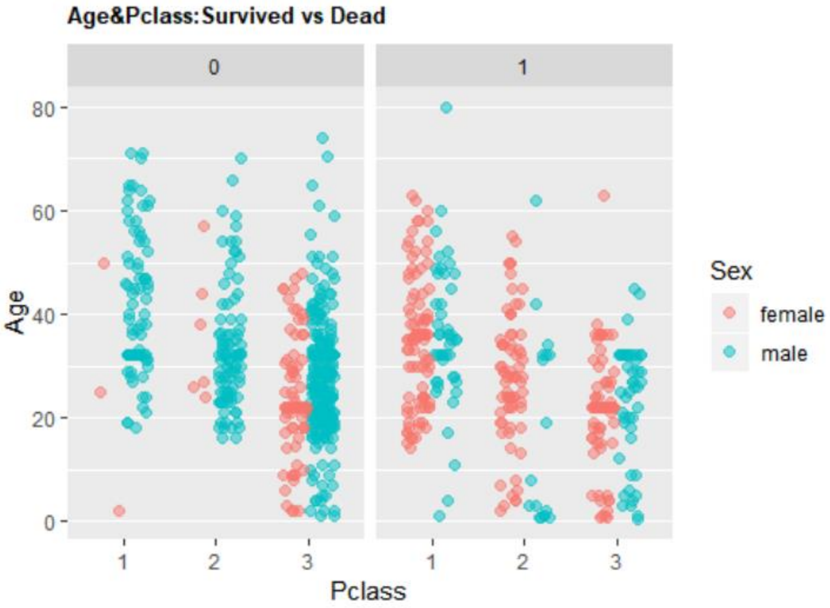


```
ggplot(data, aes(x=Pclass, y= Age, col = Sex)) +
  geom_point(position = position_jitterdodge(0.5, 0, 0.6),size=2,alpha=0.5) +
  facet_grid(.~Survived) +
  ggtitle("Age&Pclass:Survived vs Dead") +
  theme(plot.title=element_text(face="bold", size=10, vjust=2),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())
```



ggplot을 활용한 Tietanic data 시각화

5. Insight 도출



?