Statistics Bootcamp using R

DAY 2 DATA VISUALIZATION & UNDERSTANDING PATTERN

2.1 DATA VISUALIZATION

GU Zhan (Sam)
Institute of Systems Science
National University of Singapore

issgz@nus.edu.sg

Agenda

Day 2 : Data Visualization & Understanding Pattern

- **Data Visualization**
- Descriptive Statistics & Sampling
- Introduction to Normal Distribution

Learning objectives

- Understand concepts of data exploration using graph
- Understand data exploration/visualization using R

Effective data exploration using graphs

Human Perception

Understanding and communicating patterns in raw data can be difficult…

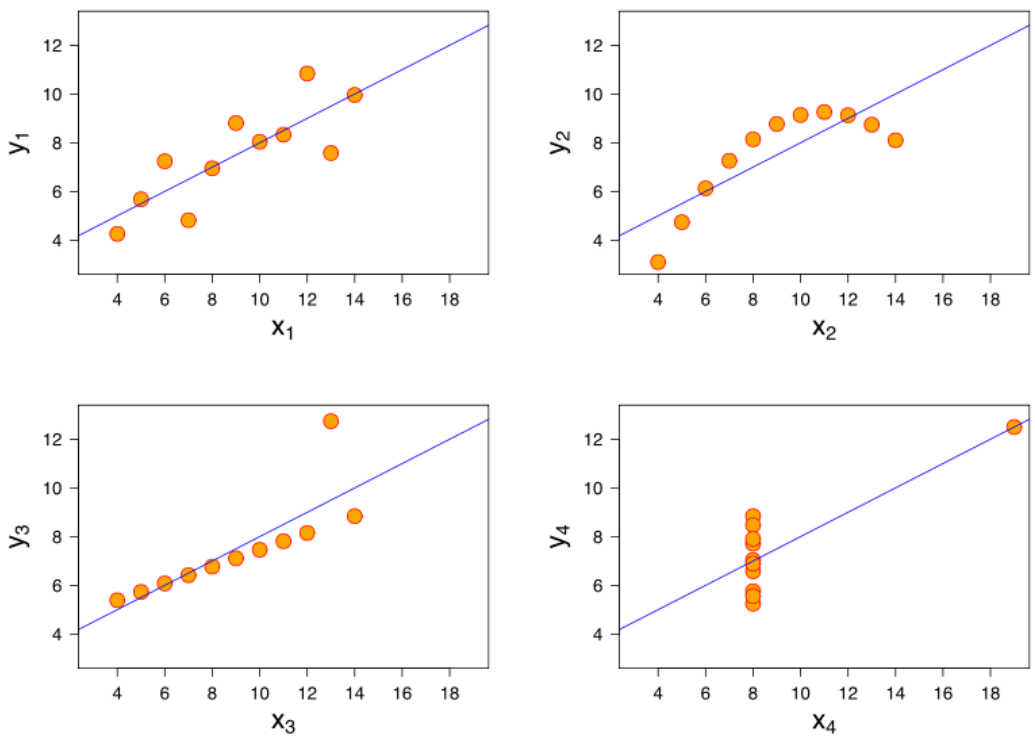
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Human Perception

The same information expressed ‘visually’ is far easier to understand, interpret, and communicate…

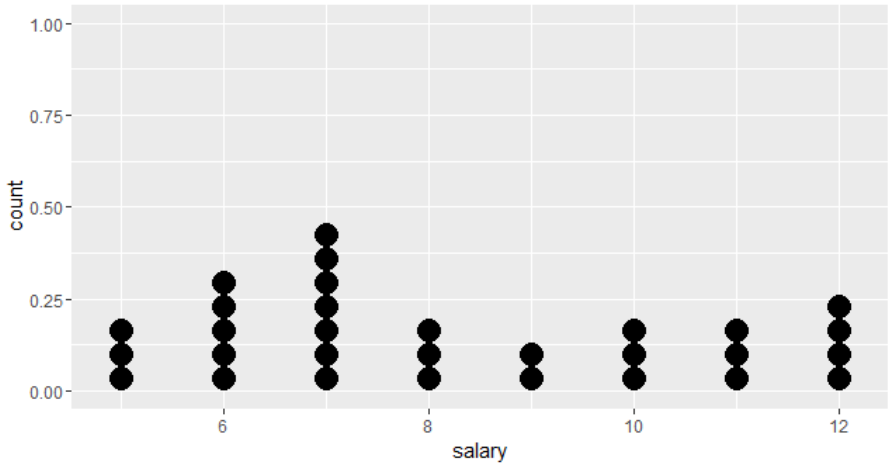
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

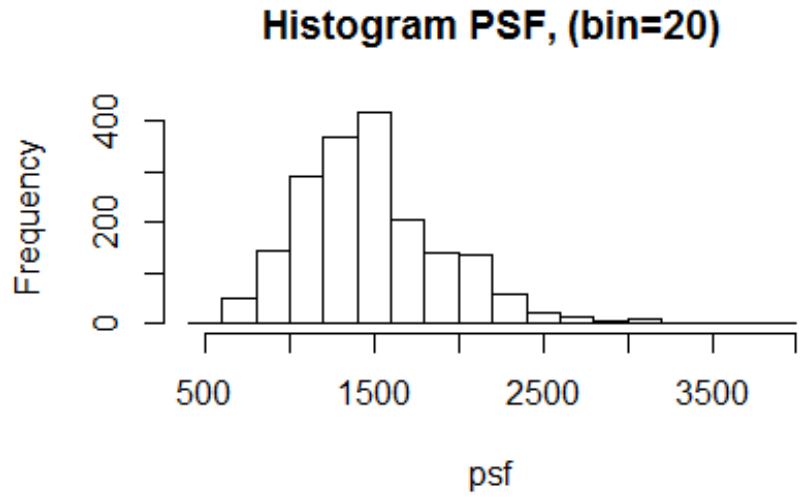


Histogram

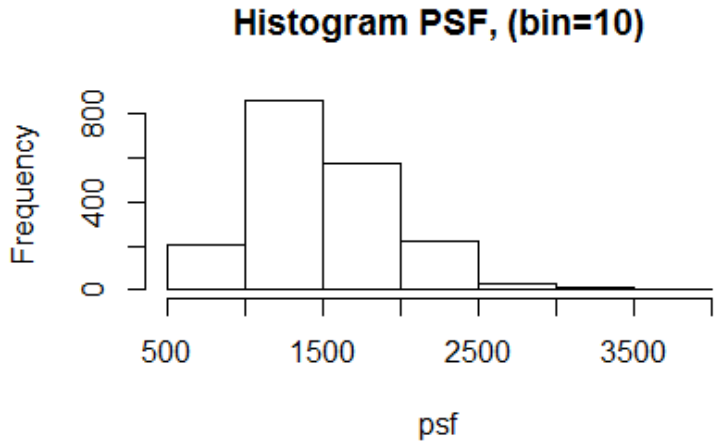
- Used for showing **distribution** of your data and makes categorizing easier.
  - Bin size is important:
    - Too small – you get spurious patterns
    - Too large – you might miss important patterns

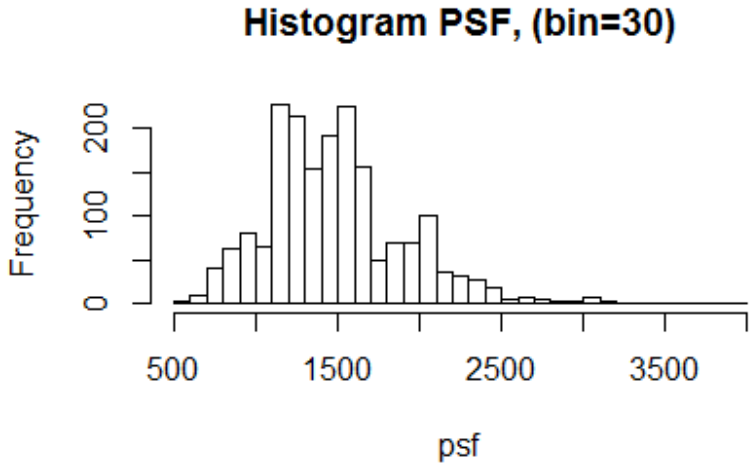
- Histogram vs. Bar Chart
  - Bar Chart is used to display categorical data
  - Histogram is used to display a range of numeric values

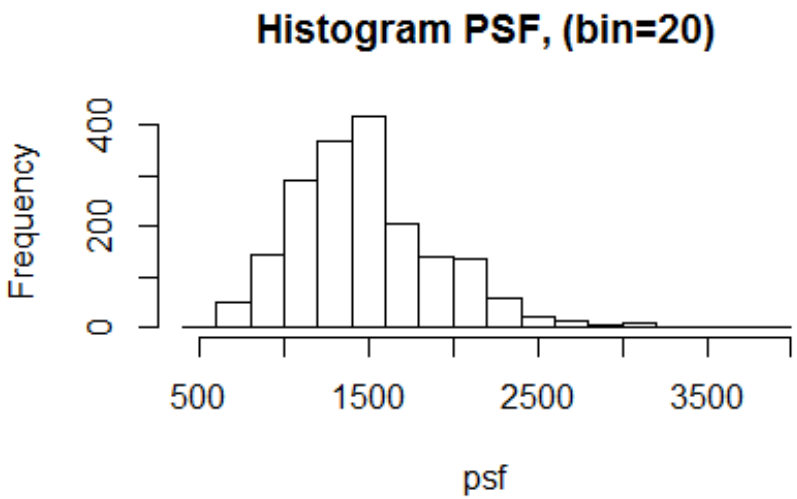


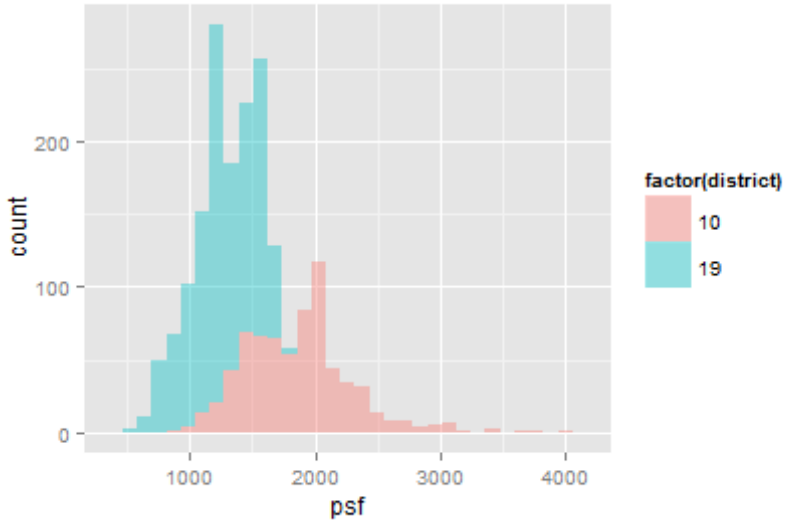


Histogram





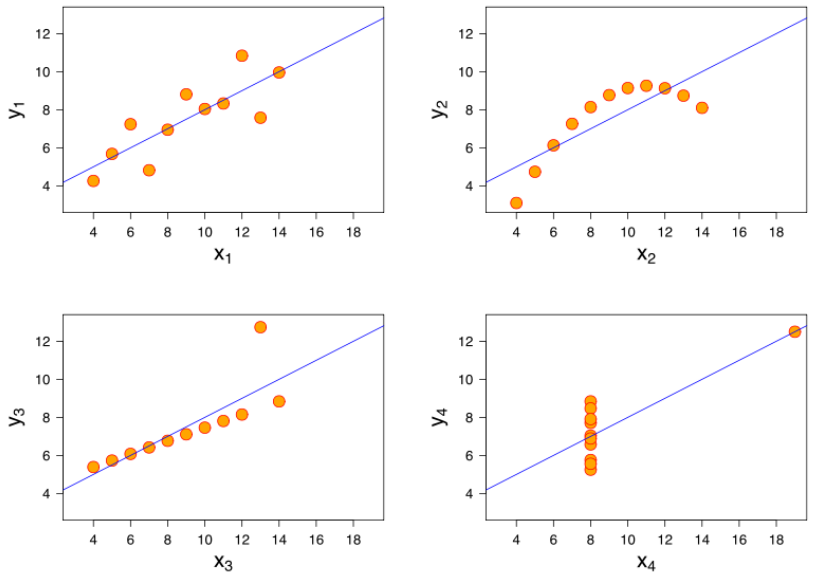




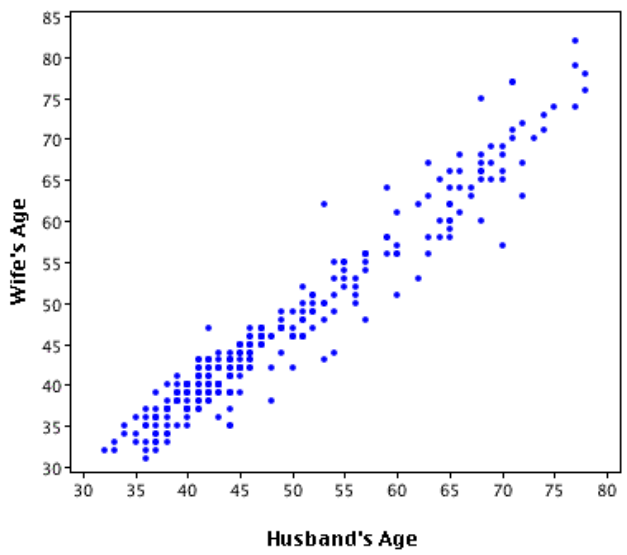
Scatter plots

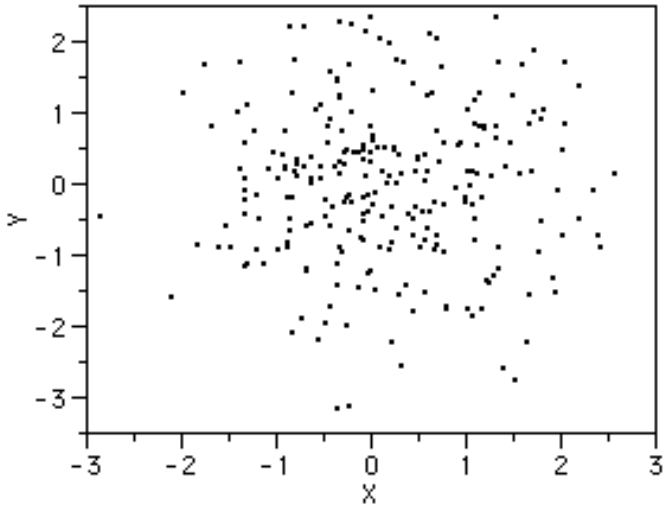
- Used for **analyzing relationships** between two numeric data variables.
  - It is often used to prove or disprove a relationship between the two variables. Example- see how sales and profit relate.
  - Allows a view of skewness of the data distribution. Example – can easily **detect Outliers**.

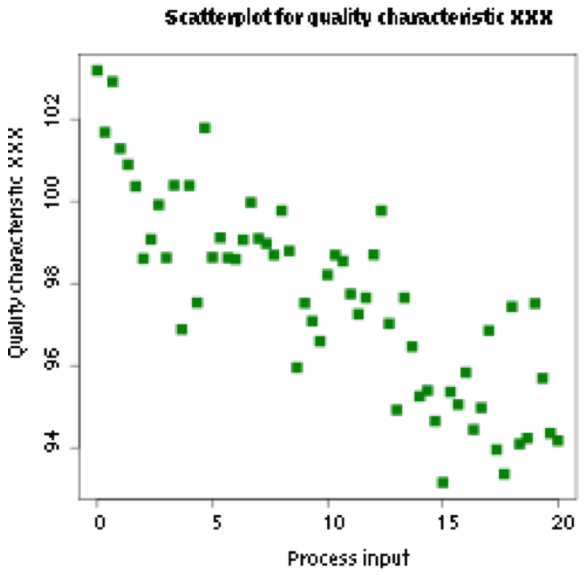
- Scatter plots can be more useful if you draw **trend lines** or **reference line**.



Scatter Plots and Correlation

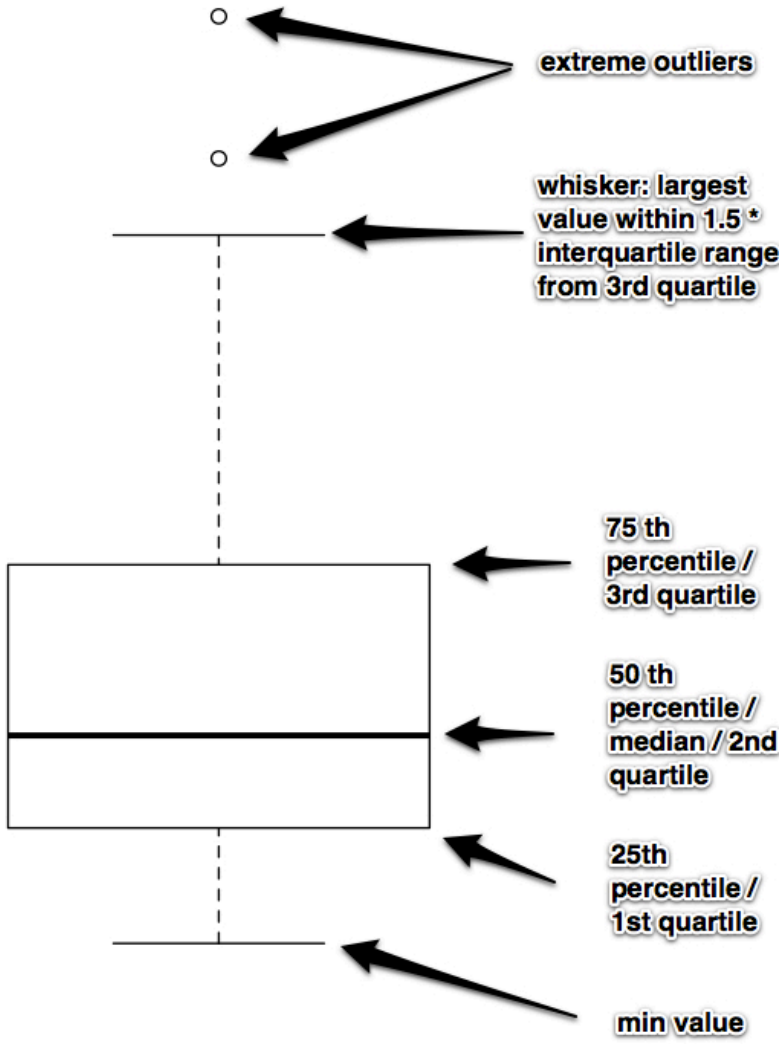






Box Plot

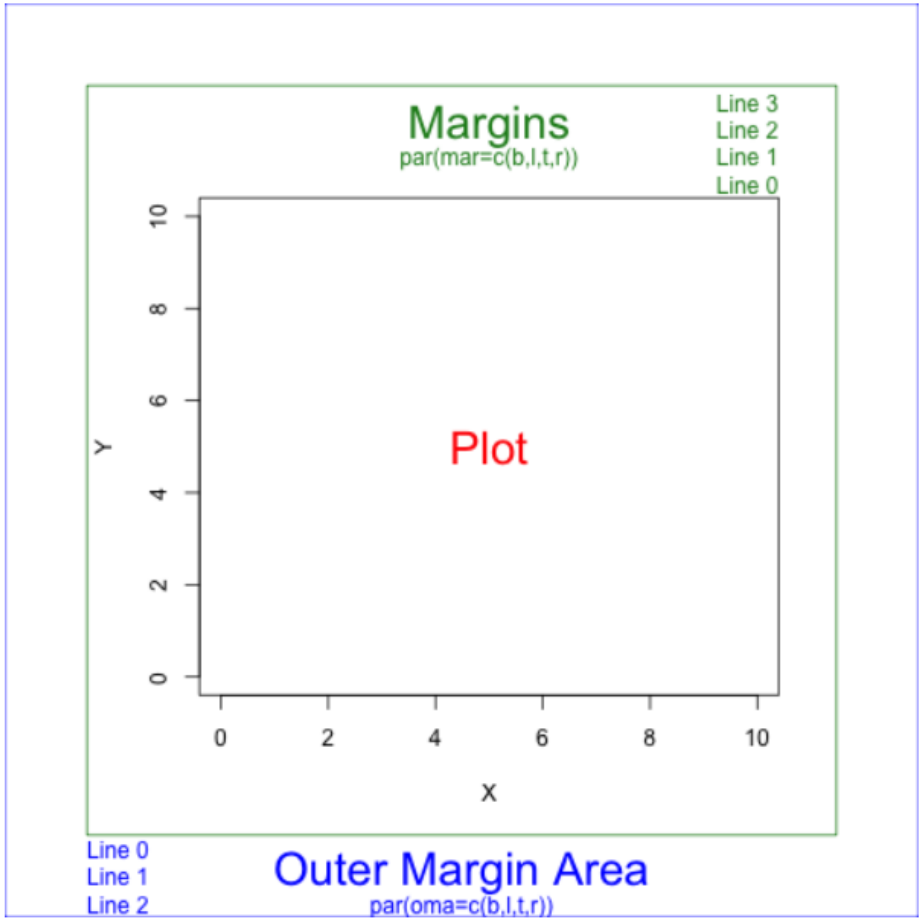
- The five-number summary can be represented graphically using a boxplot (as discussed in day 1).



Charting using ggplot2

- Understand the plotting ideas behind ggplot2
- Use ggplot2 to create graphs
- Why learn ggplot2?

Coordinate System: R plot



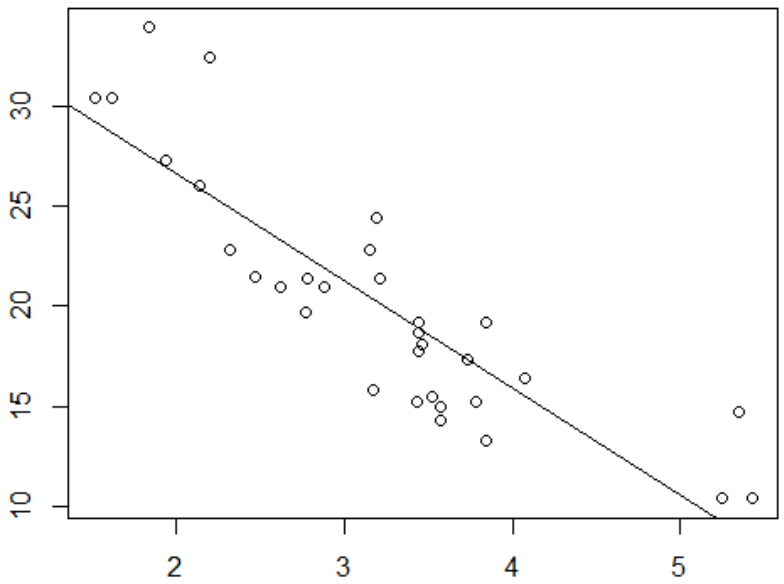
You can control their size calling the par function before your plot and giving the corresponding arguments:

- mar() for margin.
- oma() for outer margin area
- you must give four values [example: *par(mar=c(4,0,0,0))]*

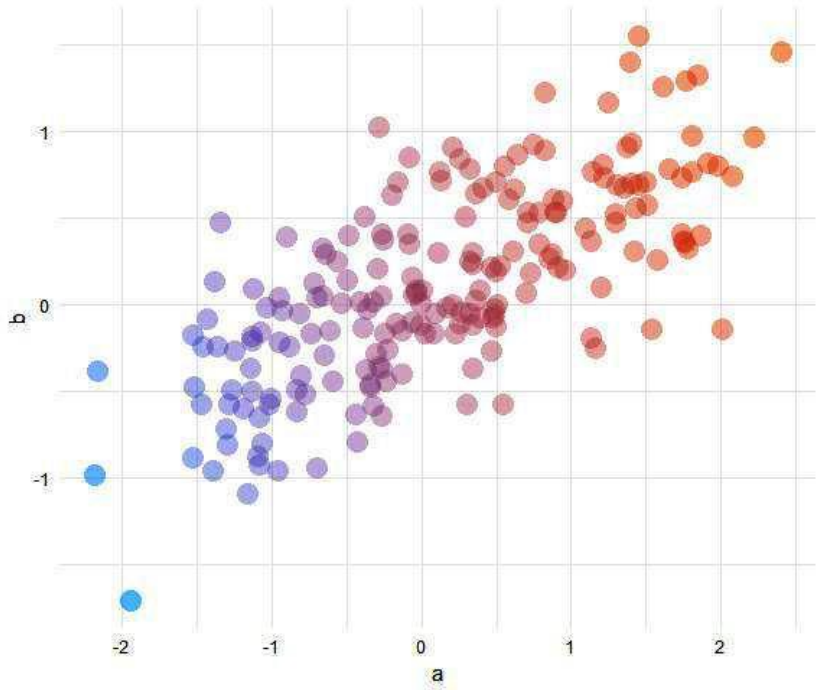
Source https://www.r-graph-gallery.com/74-margin-and-oma-cheatsheet/

Why learn ggplot2?

Orthodox…

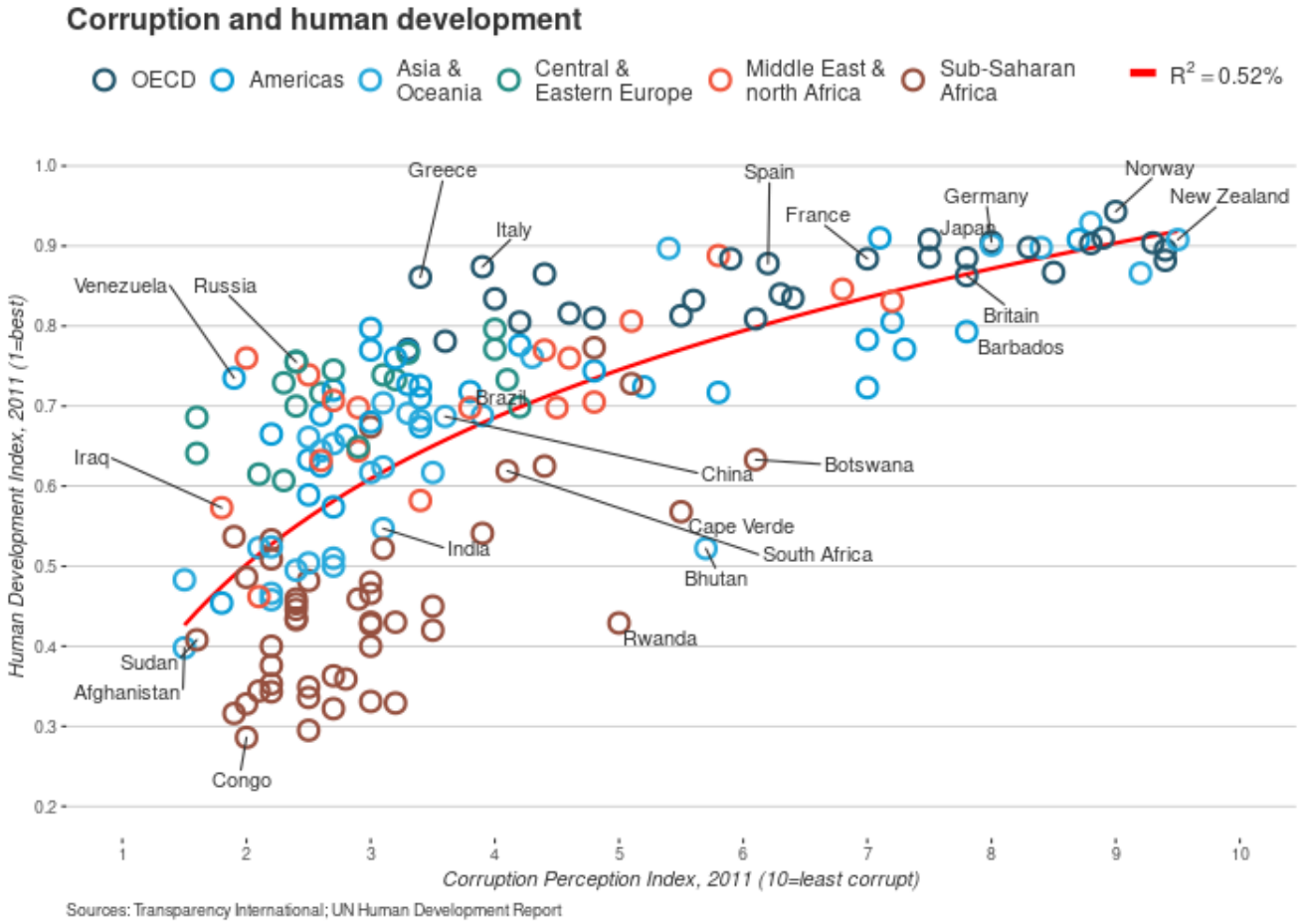


Stylish !



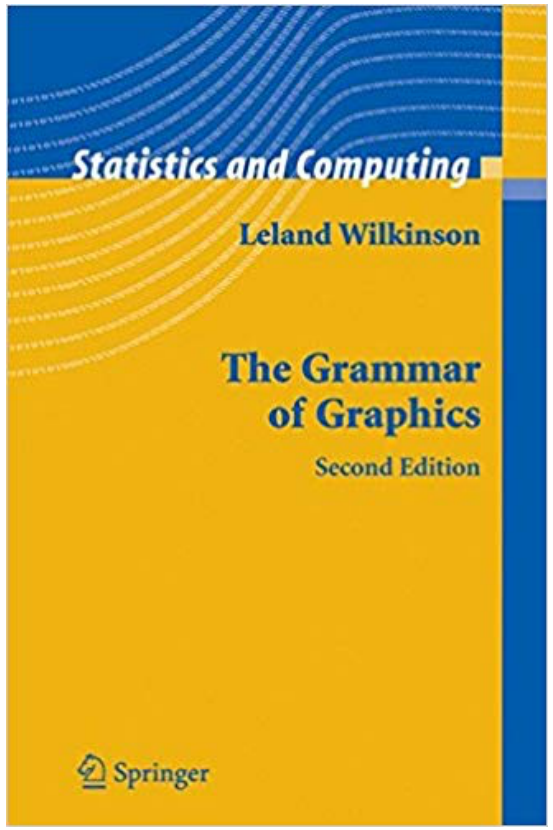
Source: https://drsimonj.svbtle.com/pretty-scatter-plots-with-ggplot2

Everything by code



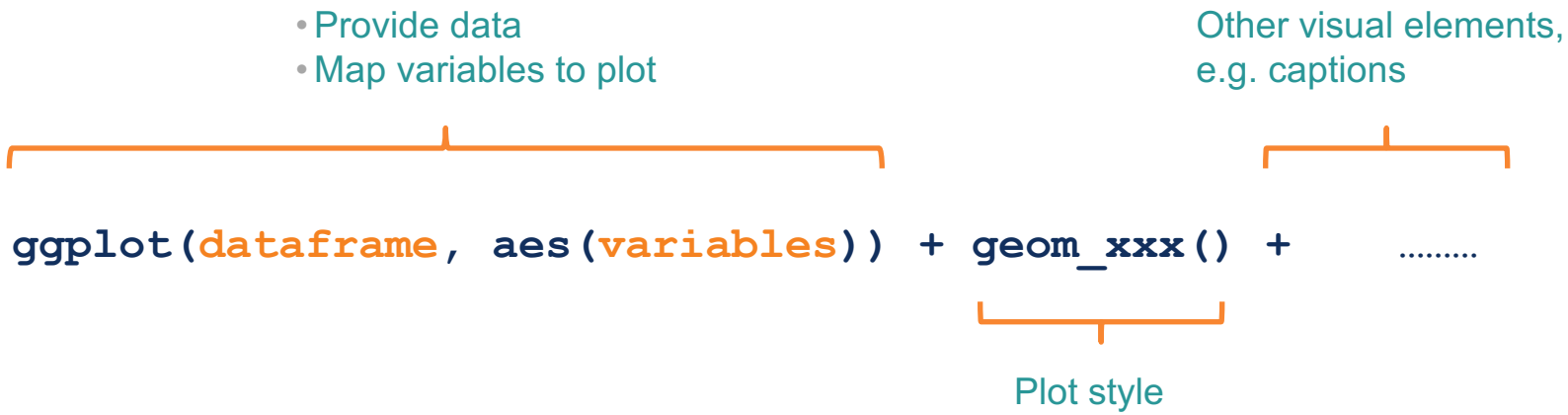
Source: http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

ggplot2

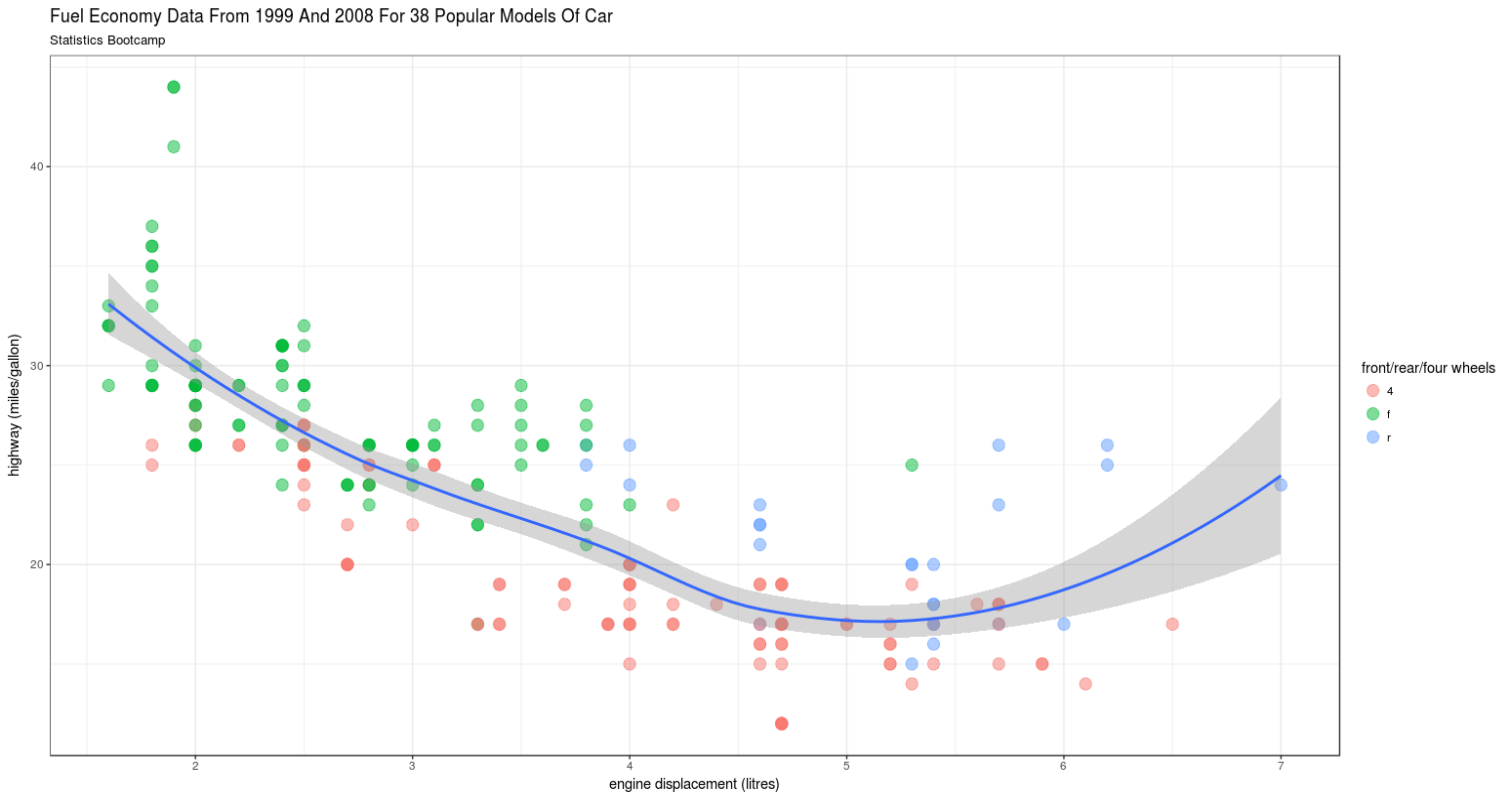


- A system for declaratively creating graphics, based on *The Grammar of Graphics*.
- One of the most popular R packages since 2005
- How to use it:
  - Provide data
  - Tell ggplot2 how to map variables to aesthetics (visual elements)
  - What graphical primitives to use
  - Plot generated

ggplot2 command

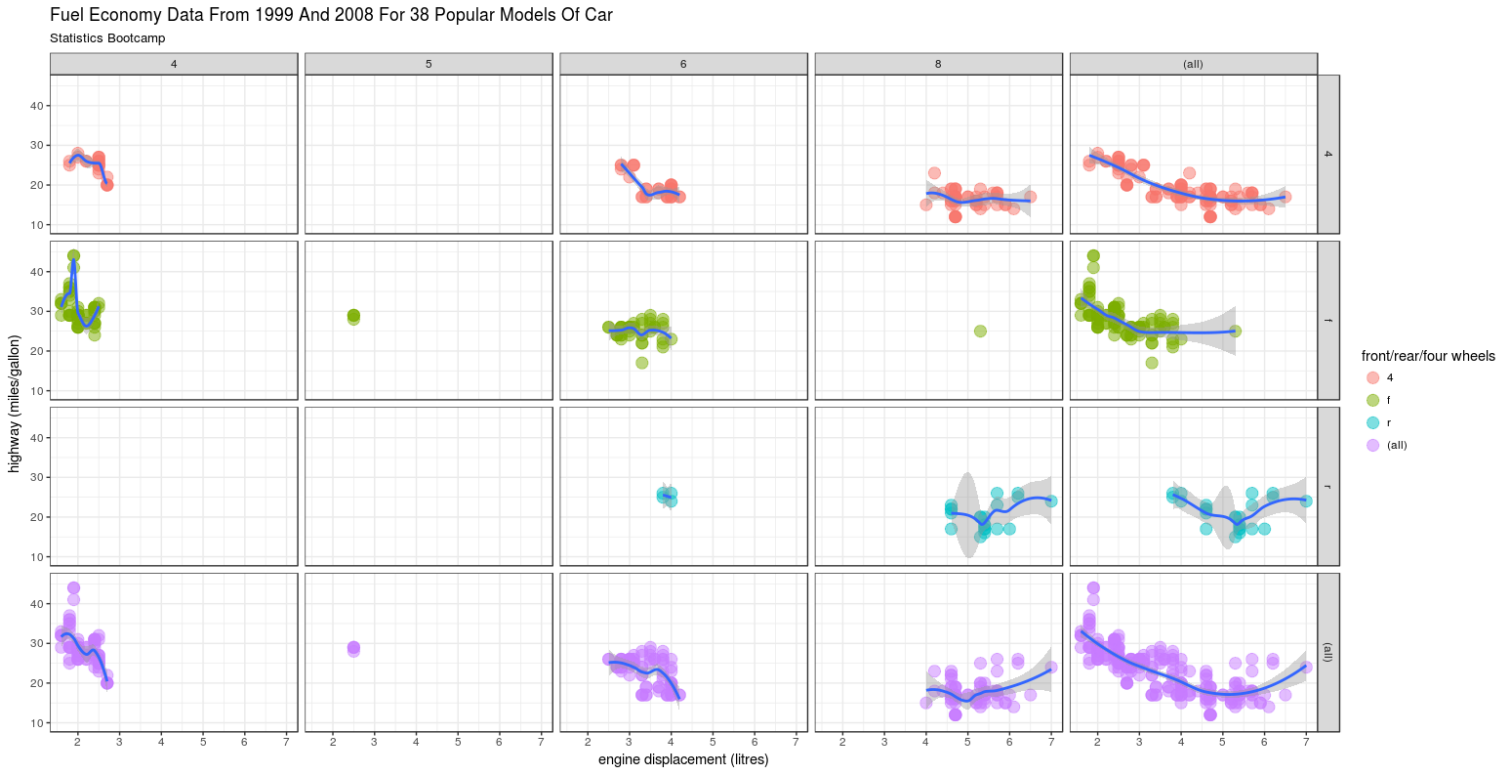


Everything by code




```
g_base = ggplot(mpg, aes(displ, hwy))
g_aes = g_base + geom_point(aes(color = drv),
size = 4, alpha = 0.5) + geom_smooth(size = 1,
linetype = 1, se = TRUE) + theme_bw(base_family
= "Times")
g_aes + labs(x = "engine displacement (litres)",
             y = "highway (miles/gallon)",
             title = "Fuel Economy Data From 1999 And
2008 For 38 Popular Models Of Car",
             subtitle = "Statistics Bootcamp",
             color = "front/rear/four wheels")
```

Everything by code




```
g_base = ggplot(mpg, aes(displ, hwy))
g_aes = g_base + geom_point(aes(color = drv),
size = 4, alpha = 0.5) + geom_smooth(size = 1,
linetype = 1, se = TRUE) + theme_bw(base_family
= "Times")
g_aes + labs(x = "engine displacement (litres)",
             y = "highway (miles/gallon)",
             title = "Fuel Economy Data From 1999 And
2008 For 38 Popular Models Of Car",
             subtitle = "Statistics Bootcamp",
             color = "front/rear/four wheels") +
facet_grid(facets = drv ~ cyl, margins = TRUE)
```

Let's look at a ship

Source: https://medium.com/i-like-big-data-and-i-cannot-lie/how-i-scored-in-the-top-9-of-kaggles-titanic-machine-learning-challenge-243b5f45c8e9

Titanic disaster data

Field	Description	Values
pclass	Passenger Class	1,2,3 for class
survived	Whether the passenger survived or not	1:survived, 0:perished
Name	Name of the passenger	Text string
Sex	Gender of the passenger	male, female
Age	Age of the passenger	Number
Sibsp	Siblings/spouses on board	Number
Parch	??	Number
Ticket	??	Text string
Fare	Fare paid	Number
Cabin	Cabin number	Text string
Embarked	Port of embarkation	Text string
Boat	??	Number
Body	??	Number
Home.dest	Home Destination	Text String

- 14 fields, 1309 observations
- Some values were not available

t

Filter

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1	1	1	Allen, Miss. Elisabeth Walton	female	29.00	0	0	24160	211.3375	B5	S	2	NA	St Louis, MO
2	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	C22 C26	S	11	NA	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	C22 C26	S		NA	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville, ON
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00	1	2	113781	151.5500	C22 C26	S		NA	Montreal, PQ / Chesterville, ON
6	1	1	Anderson, Mr. Harry	male	48.00	0	0	19952	26.5500	E12	S	3	NA	New York, NY
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63.00	1	0	13502	77.9583	D7	S	10	NA	Hudson, NY
8	1	0	Andrews, Mr. Thomas Jr	male	39.00	0	0	112050	0.0000	A36	S		NA	Belfast, NI
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.00	2	0	11769	51.4792	C101	S	D	NA	Bayside, Queens, NY
10	1	0	Artagaveytia, Mr. Ramon	male	71.00	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
11	1	0	Astor, Col. John Jacob	male	47.00	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
12	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18.00	1	0	PC 17757	227.5250	C62 C64	C	4	NA	New York, NY
13	1	1	Aubart, Mme. Leontine Pauline	female	24.00	0	0	PC 17477	69.3000	B35	C	9	NA	Paris, France
14	1	1	Barber, Miss. Ellen "Nellie"	female	26.00	0	0	19877	78.8500		S	6	NA	
15	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.00	0	0	27042	30.0000	A23	S	B	NA	Hessle, Yorks
16	1	0	Baumann, Mr. John D	male	NA	0	0	PC 17318	25.9250		S		NA	New York, NY
17	1	0	Baxter, Mr. Quigg Edmond	male	24.00	0	1	PC 17558	247.5208	B58 B60	C		NA	Montreal, PQ
18	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50.00	0	1	PC 17558	247.5208	B58 B60	C	6	NA	Montreal, PQ
19	1	1	Bazzani, Miss. Albina	female	32.00	0	0	11813	76.2917	D15	C	8	NA	
20	1	0	Beattie, Mr. Thomson	male	36.00	0	0	13050	75.2417	C6	C	A	NA	Winnipeg, MN
21	1	1	Beckwith, Mr. Richard Leonard	male	37.00	1	1	11751	52.5542	D35	S	5	NA	New York, NY

Showing 1 to 23 of 1,309 entries

Preprocessing

1. Install ‘ggplot2’ package
   (*Do this if this package is not available in your environment*)
2. Load the package

3. Load the csv file

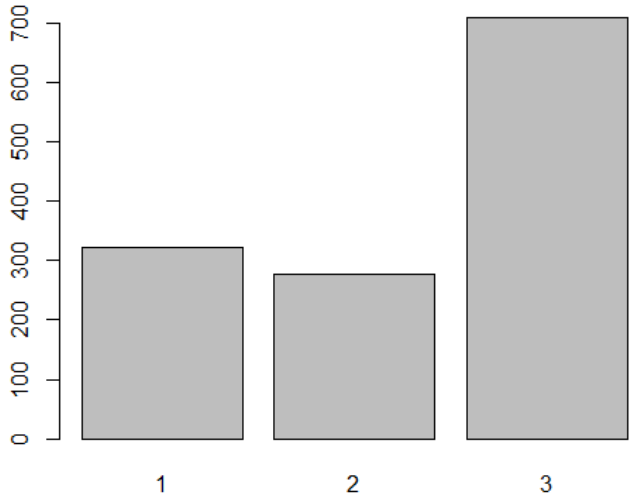
4. Make variables ‘pclass’ and ‘survived’ a factor

5. Recode variable ‘sex’ into a new variable called ‘gender’

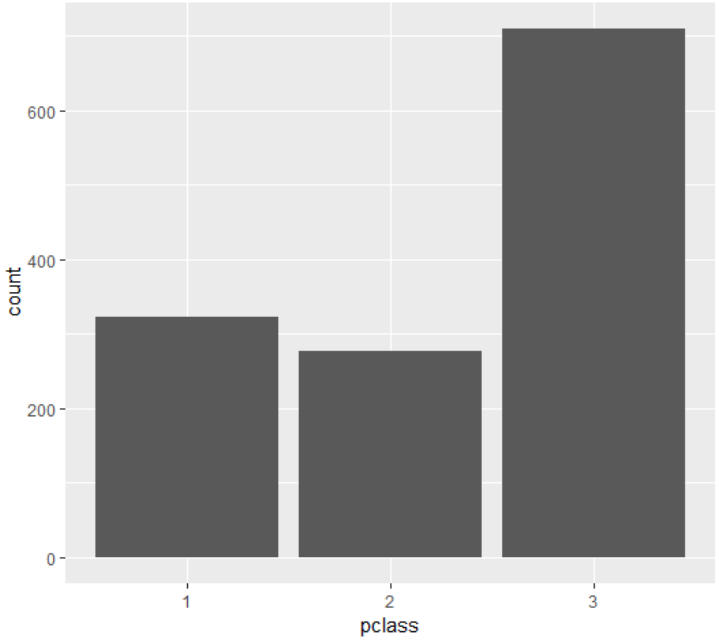
```
# install.packages("ggplot2")    # Do this if the package was not installed

> library(ggplot2)

> t = read.csv('titanic3.csv', stringsAsFactors=FALSE)


> t$pclass = factor(t$pclass)

> t$survived = factor(t$survived)

> t$gender = factor(t$sex)
```

Bar chart

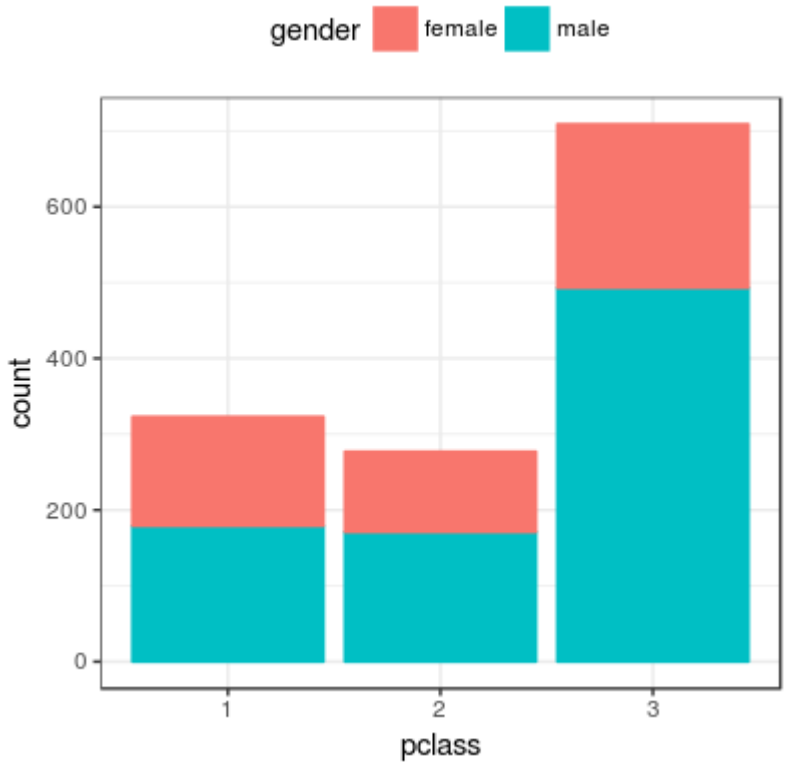
```
# Create bar chart using R graphics
> barplot(table(t$pclass))
```



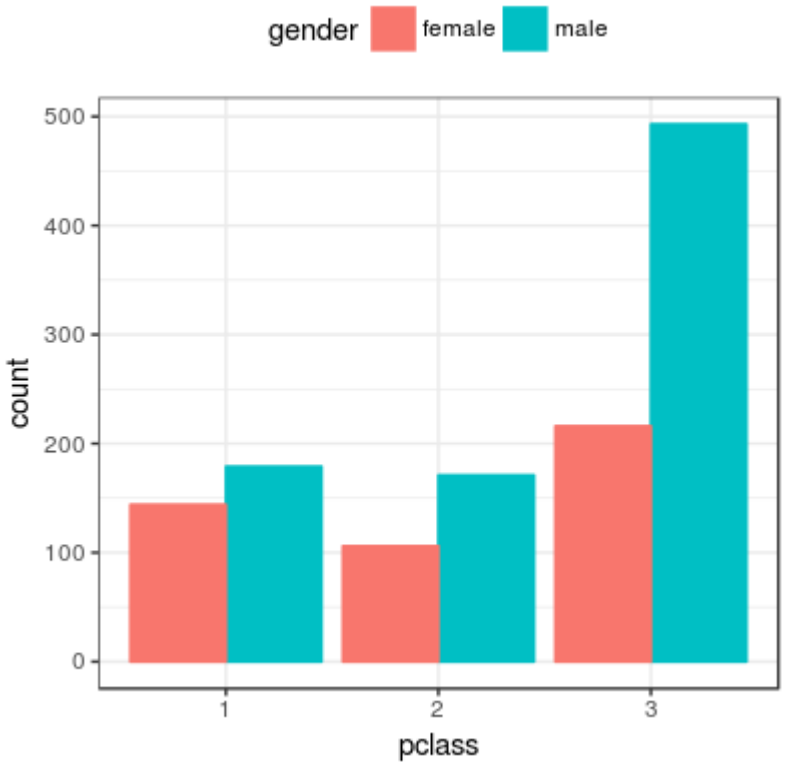

```
# Create bar chart using ggplot2
> ggplot(t, aes(x=pclass)) + geom_bar()
```



Bar chart

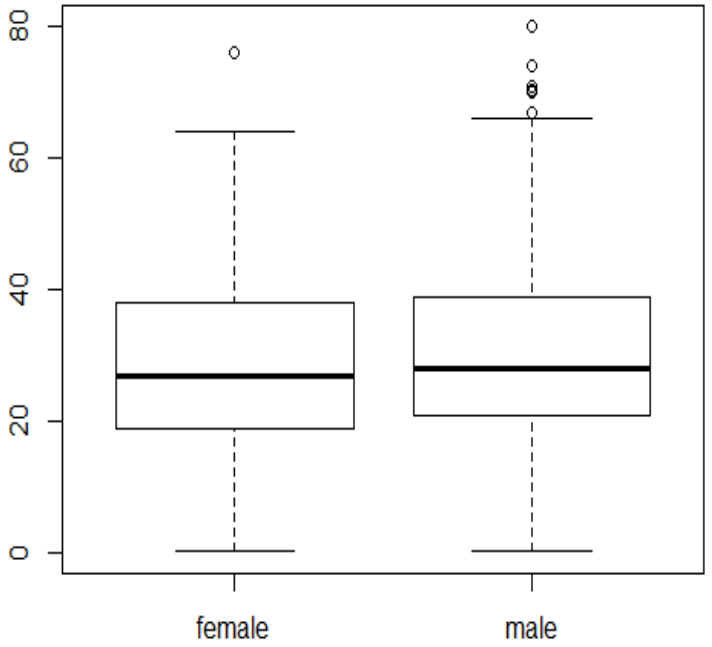


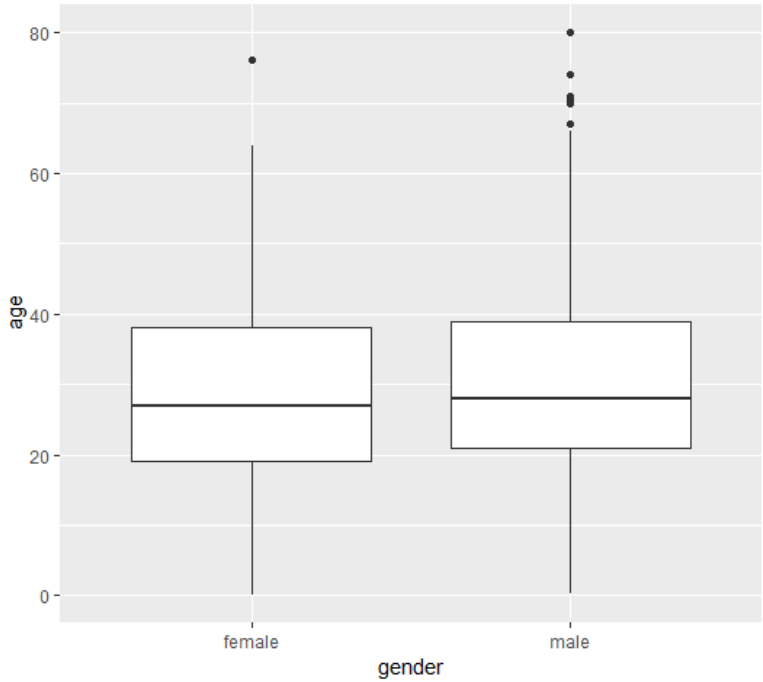

```
> ggplot(t, aes(x=pclass,
color=gender, fill = gender)) +
    geom_bar() +
    theme_bw() +
theme(legend.position="top")
```




```
> ggplot(t, aes(x=pclass,
color=gender, fill = gender)) +
    geom_bar(position="dodge") +
    theme_bw() +
theme(legend.position="top")
```

Box plot

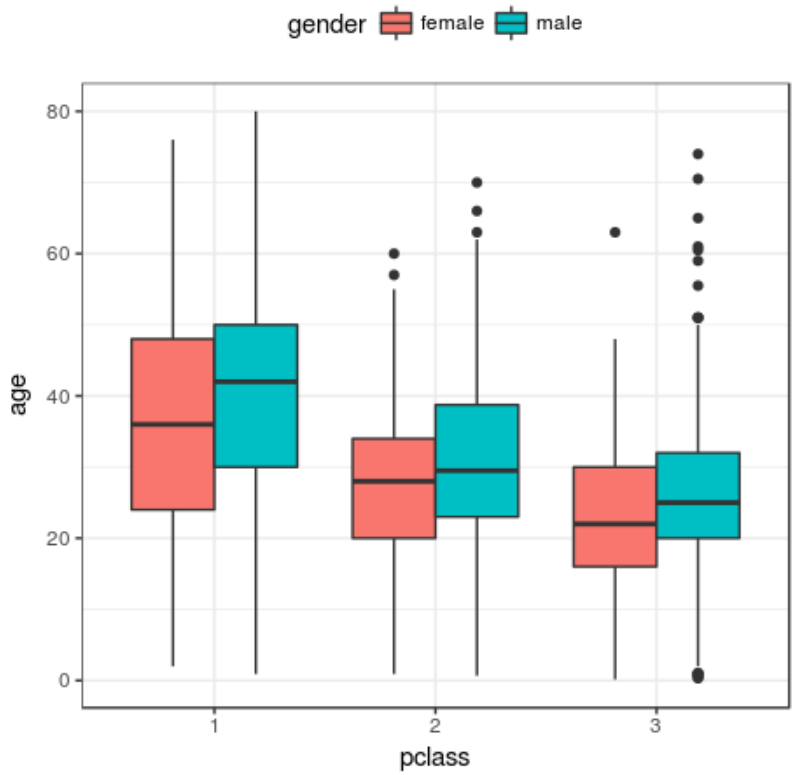




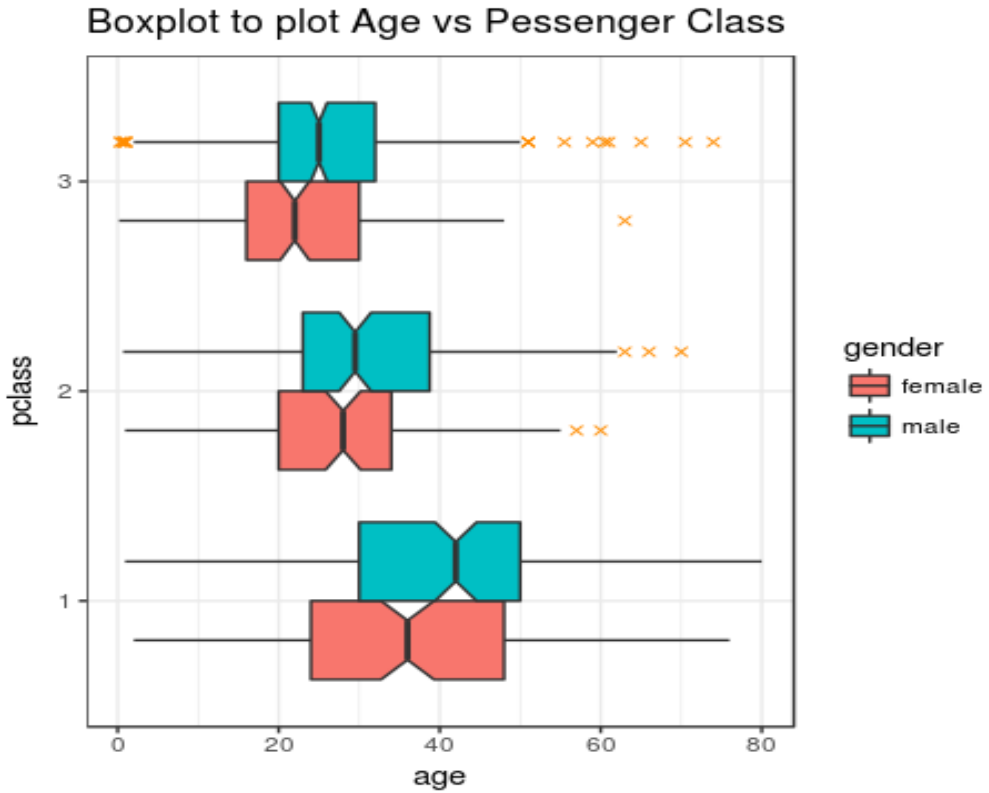

```
# Create box plot using R graphics
> boxplot(t$age ~ t$gender)
```

```
# Create box plot using ggplot2
> ggplot(t, aes(gender, age))+ geom_boxplot()
```

Box plot

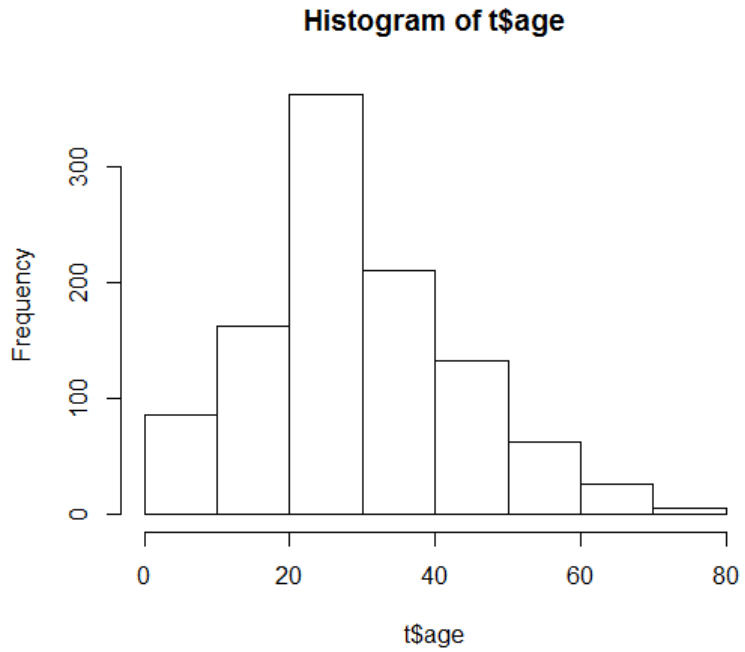



```
> ggplot(t, aes(x=pclass, y=age, fill = gender)) +
    geom_boxplot() + theme_bw() + theme(legend.position="top")
```

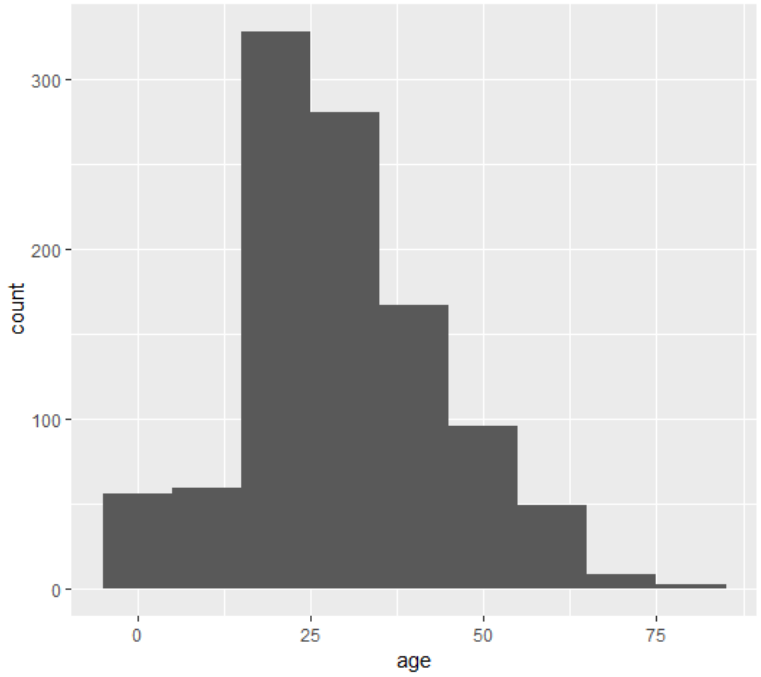



```
> ggplot(t, aes(x=pclass, y=age, fill = gender)) + geom_boxplot(outlier.colour = "dark orange", outlier.shape = 4, notch = TRUE) +
    coord_flip() + theme_bw() + ggtitle("Boxplot to plot Age vs Pessenger Class")
```

Histogram

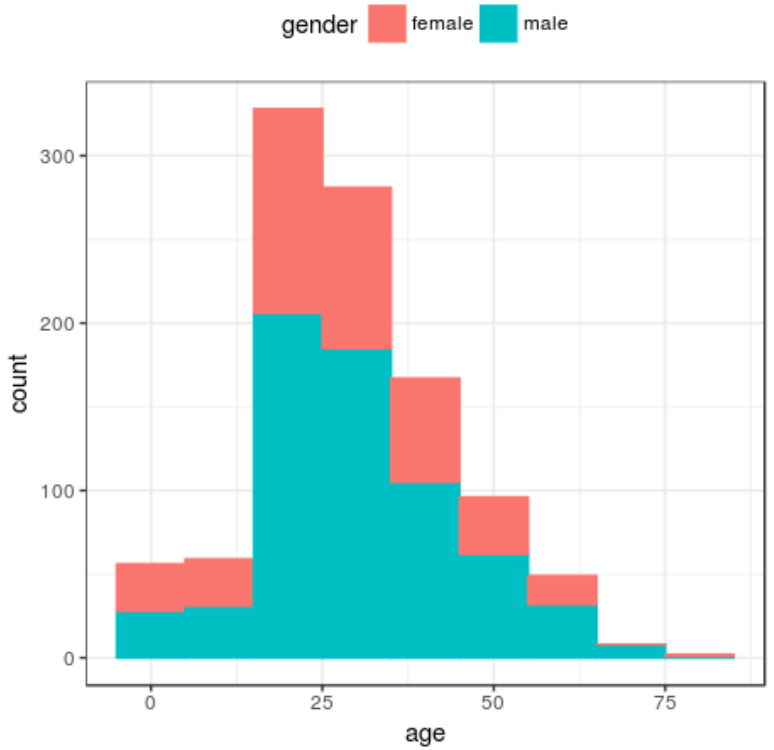



```
# Create histogram using R graphics
> hist(t$age, breaks=10)
```

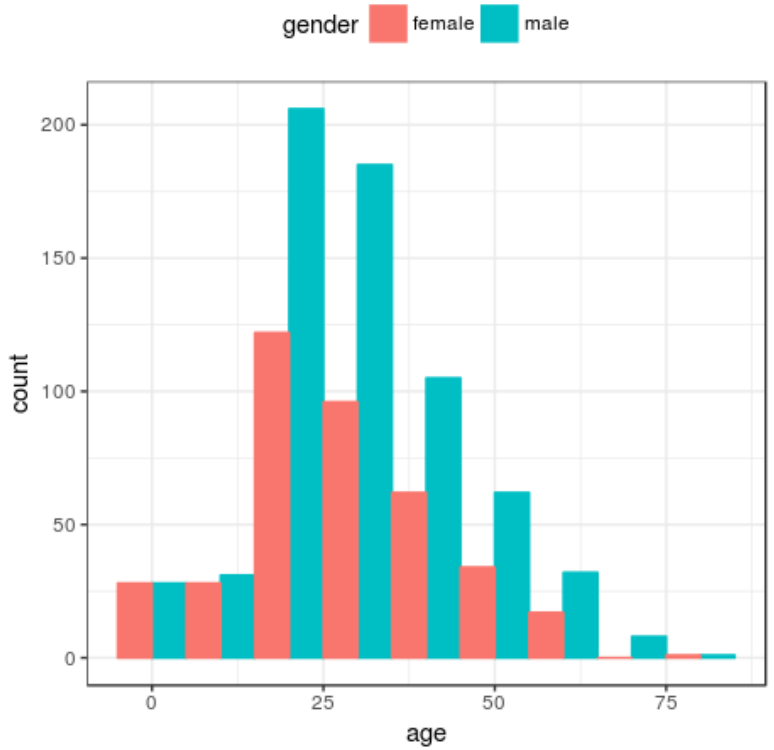



```
# Create histogram using ggplot2
> ggplot(t, aes(x=age)) +
    geom_histogram(binwidth = 10)
```

Histogram

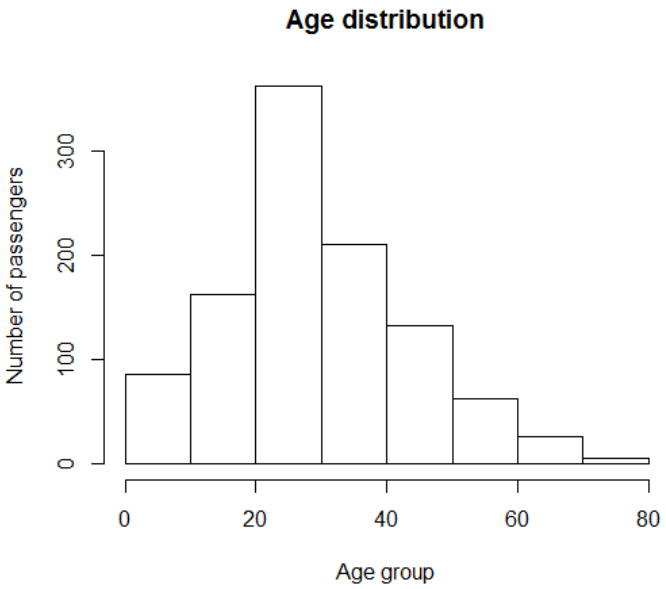



```
> ggplot(t, aes(x=age,
color=gender, fill = gender)) +
    geom_histogram(binwidth = 10) +
    theme_bw() +
theme(legend.position="top")
```

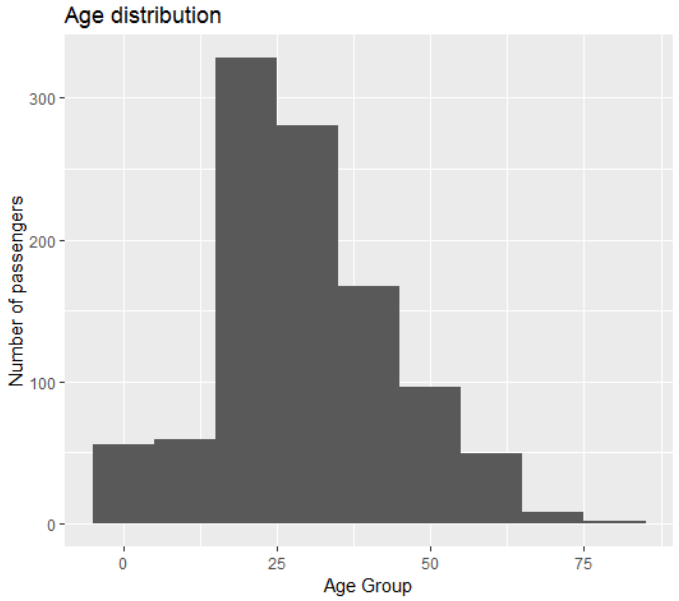



```
> ggplot(t, aes(x=age, color=gender,
fill = gender)) +
    geom_histogram(binwidth = 10,
position="dodge")+ theme_bw() +
theme(legend.position="top")
```

Display captions



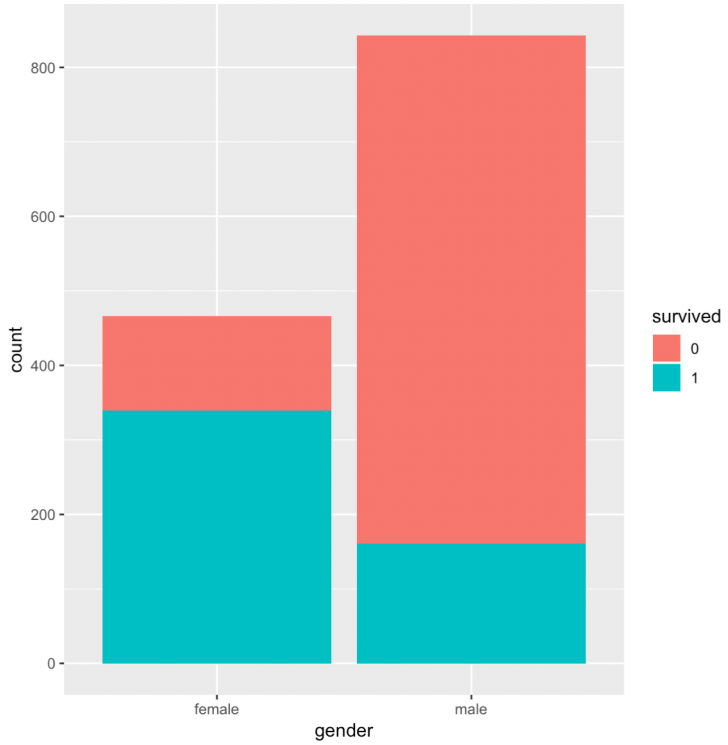

```
# Display caption using R graphics
> hist(t$age,breaks=10,
       xlab="Age group",
       ylab="Number of passengers",
       main="Age distribution")
```



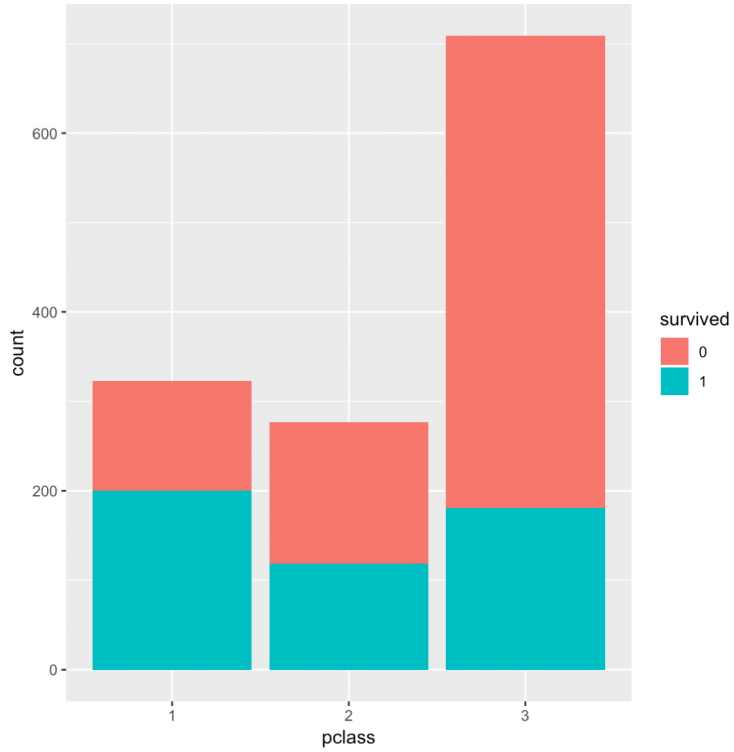

```
# Display caption using ggplot2
> age_hist = ggplot(t, aes(x=age)) +
                 geom_histogram(binwidth = 10)
> captions = labs(x="Age Group",
                  y="Number of passengers",
                  title="Age distribution")
> age_hist + captions
```

Who had survived?

Examination by stacked bar plot



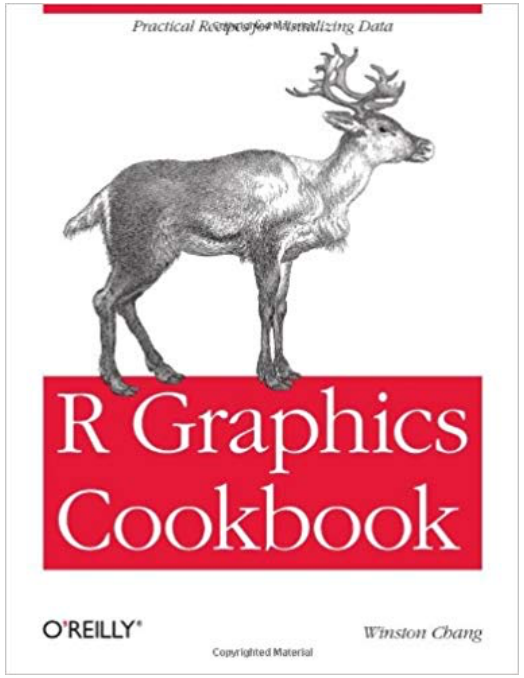

```
# Create stacked barplot using ggplot2
> ggplot(t, aes(x=gender, fill=survived)) + geom_bar()
```

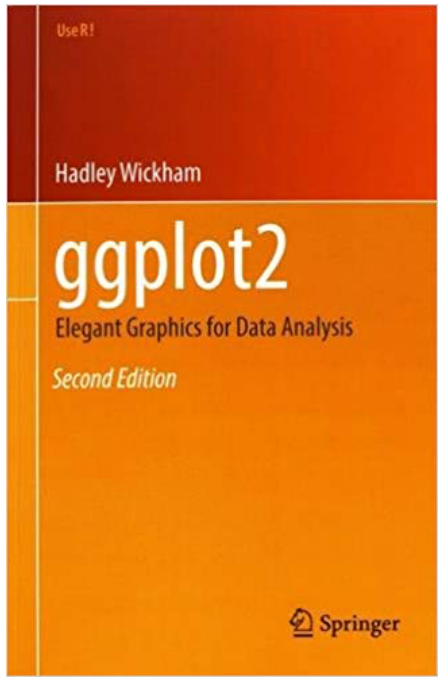


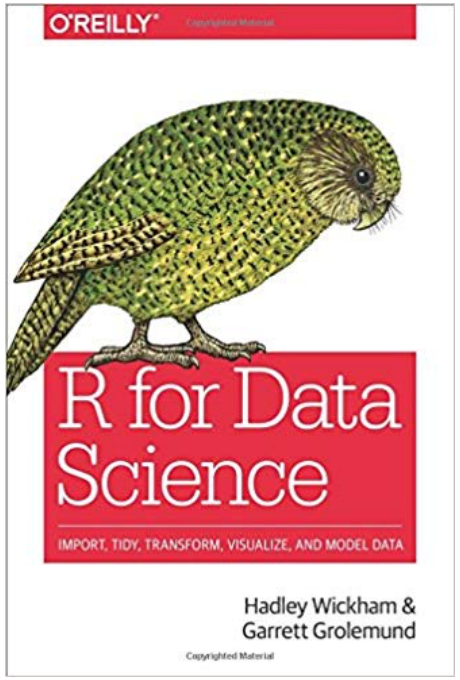

```
# Create stacked barplot using ggplot2
> ggplot(t, aes(x=pclass, fill=survived)) + geom_bar()
```

Other resources

https://ggplot2.tidyverse.org
http://r4ds.had.co.nz/data-visualisation.html
http://r4ds.had.co.nz/graphics-for-communication.html







Data visualization for Story Telling & Call for Action

End of Lecture Notes