

Sprint 05 - Intro to Data Visualization

The Secret Sauce

- อยากเก่งเรื่องไหน ให้ฝึกเรื่องนั้นเยอะ ๆ
- ดูคลิป ลองทำตามทีสอน เขียนสรุป และลองหา Dataset มาลองเล่น
- แหล่ง Dataset:

data.world | The Cloud-Native Data Catalog

Digital Event: February 9th - How to accelerate your cloud data adoption journey Close Make knowledge your superpower data.world is the enterprise data catalog for the modern data stack. Our cloud-native


 <https://data.world/>







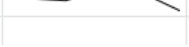
data.world

Data Visualization in Google Sheets

- Exercise File (ให้ทำสำเนาเพื่อทำตามแบบฝึกหัด):
<https://docs.google.com/spreadsheets/d/1DGIjlxrOMXvVv7dSiTmptnbWbK5wimq4LCnKDw-WOo/edit#gid=0>
- My Exercise File: https://docs.google.com/spreadsheets/d/1P-EQGssdtgFOCiAErVDqBICHjbao50eye_DYvgt1l_4/edit#gid=696084048
- เราสามารถสร้าง Spark Line เป็น Line Chart อย่างง่ายได้ด้วย =SPARKLINE() เช่น:











F3		fx	=SPARKLINE(B3:E3)			
	A	B	C	D	E	F
1		Y2020				
2		Q1	Q2	Q3	Q4	
3	Toyota	100	200	250	300	
4	Honda	200	150	120	160	
5	Mazda	80	250	400	490	
6	Ford	50	60	55	40	
7	Nissan	160	150	200	120	
8						

- เราสามารถเพิ่ม Option ให้กับ Spark Line ของเราได้ เช่น สีของเส้น (กำหนดสีด้วย Hex Code ได้) หรือความหนาของเส้น เป็นต้น

F3		fx	=SPARKLINE(B3:E3, {"color","red";"linewidth",3})				
	A	B	C	D	E	F	
1		Y2020					
2		Q1	Q2	Q3	Q4		
3	Toyota	100	200	250	300		
4	Honda	200	150	120	160		
5	Mazda	80	250	400	490		
6	Ford	50	60	55	40		
7	Nissan	160	150	200	120		
8							

-Spark Line จะไม่โชว์แกน x แกน y แต่จะฝังอยู่ในตัว cell

-เราสามารถใช่ =SPARKLINE ในการทำ Bar Chart ได้ เช่น:

D11		fx	=SPARKLINE(B11:C11, {"charttype","bar" ; "color1","red" ; "color2","blue"})				
	A	B	C	D	E	F	
1		Y2020					
2		Q1	Q2	Q3	Q4		
3	Toyota	100	200	250	300		
4	Honda	200	150	120	160		
5	Mazda	80	250	400	490		
6	Ford	50	60	55	40		
7	Nissan	160	150	200	120		
8							
9		Net Promoter Score					
10		0-8	9-10				
11	Toyota	20	80				
12	Honda	25	75				
13	Mazda	12	88				
14	Ford	35	65				
15	Nissan	26	74				

-เราสามารถเลือก Dataset → ไปที่ Format → เลือก Alternating Color เพื่อให้ Table ของเรามีสีได้ เราสามารถเลือกสีได้ตามต้องการที่แถบด้านข้าง ข้อดีคือ ต่อให้เพิ่มข้อมูลแถวใหม่ขึ้นมา Table ก็จะไม่ Update สีให้โดยอัตโนมัติ

A2:E7											
	A	B	C	D	E	F	G	H	I	J	K
1											
2			Q1	Q2	Q3	Q4					
3	Toyota	100	200	250	300						
4	Honda	200	150	120	160						
5	Mazda	80	250	400	490						
6	Ford	50	60	55	40						
7	Nissan	160	150	200	120						
8											
9											
10			Q1	Q2	Q3	Q4					
11	Toyota	100	200	250	300						
12	Honda	200	150	120	160						
13	Mazda	80	250	400	490						
14	Ford	50	60	55	40						
15	Nissan	160	150	200	120						
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											

-เราสามารถทำ Heatmap ได้ด้วย Conditional Formatting เพื่อตั้งเงื่อนไขในการเทสี Cell ได้ เช่น:

D5											
	A	B	C	D	E	F	G	H	I	J	K
1											
2			Q1	Q2	Q3	Q4	Target				
3	Toyota	100	200	250	300		250				
4	Honda	200	150	120	160						
5	Mazda	80	250	400	490						
6	Ford	50	60	55	40						
7	Nissan	160	150	200	120						
8											
9											
10			Q1	Q2	Q3	Q4					
11	Toyota	100	200	250	300						
12	Honda	200	150	120	160						
13	Mazda	80	250	400	490						
14	Ford	50	60	55	40						
15	Nissan	160	150	200	120						
16											
17											
18											
19											
20											
21											
22											
23											
24											

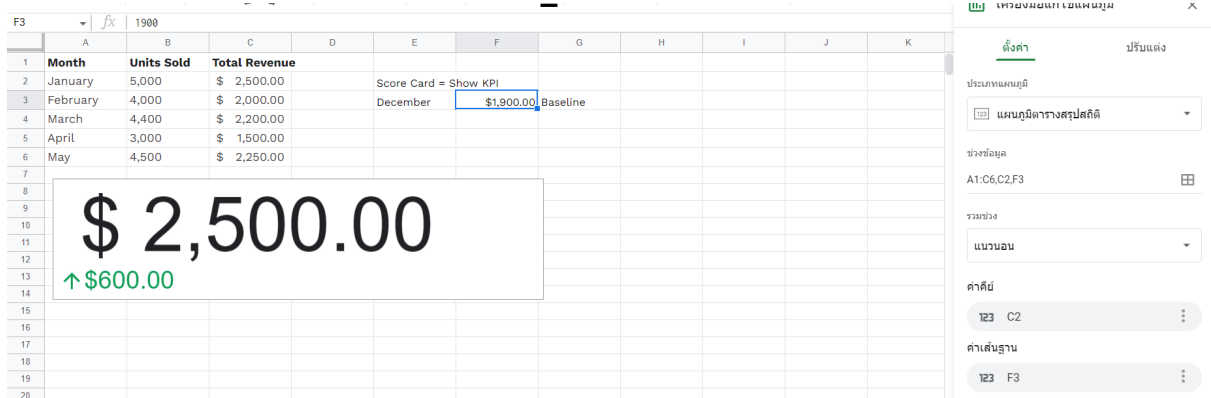
B11:E15											
	A	B	C	D	E	F	G	H	I	J	K
1											
2			Q1	Q2	Q3	Q4	Target				
3	Toyota	100	200	250	300		250				
4	Honda	200	150	120	160						
5	Mazda	80	250	400	490						
6	Ford	50	60	55	40						
7	Nissan	160	150	200	120						
8											
9											
10			Q1	Q2	Q3	Q4					
11	Toyota	100	200	250	300						
12	Honda	200	150	120	160						
13	Mazda	80	250	400	490						
14	Ford	50	60	55	40						
15	Nissan	160	150	200	120						
16											
17											
18											
19											
20											
21											
22											
23											
24											

*ยิ่งใช้สีเยอะ คนดูจะยิ่งงง แนะนำให้ใช้สีเฉดเดียว คนดูจะได้ดูง่าย

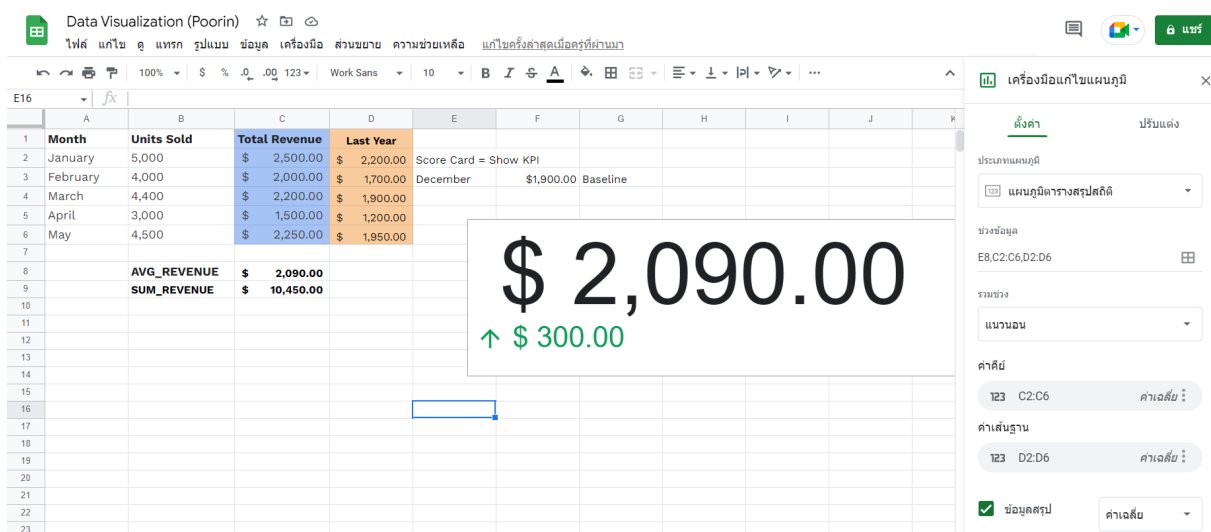
	บริษัท	ปีงบประมาณ 2563			
		10/1	10/2	20/1	12/1
8	Nissan				
9					
10		Y2020			
11	Toyota	100	200	250	300
12	Honda	200	150	120	160
13	Mazda	80	250	400	490
14	Ford	50	60	55	40
15	Nissan	160	150	200	120
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					

-KPI = Key Performance Indicator

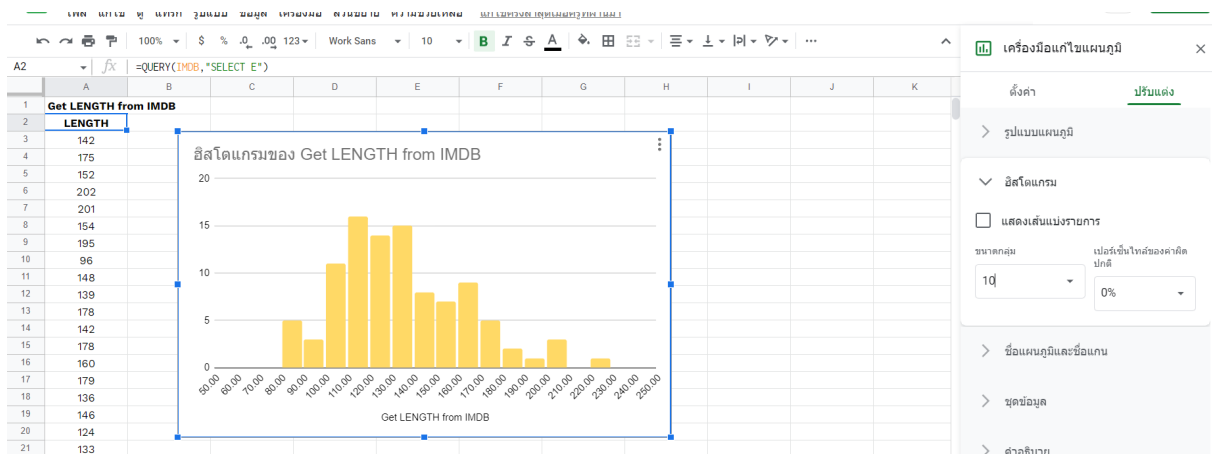
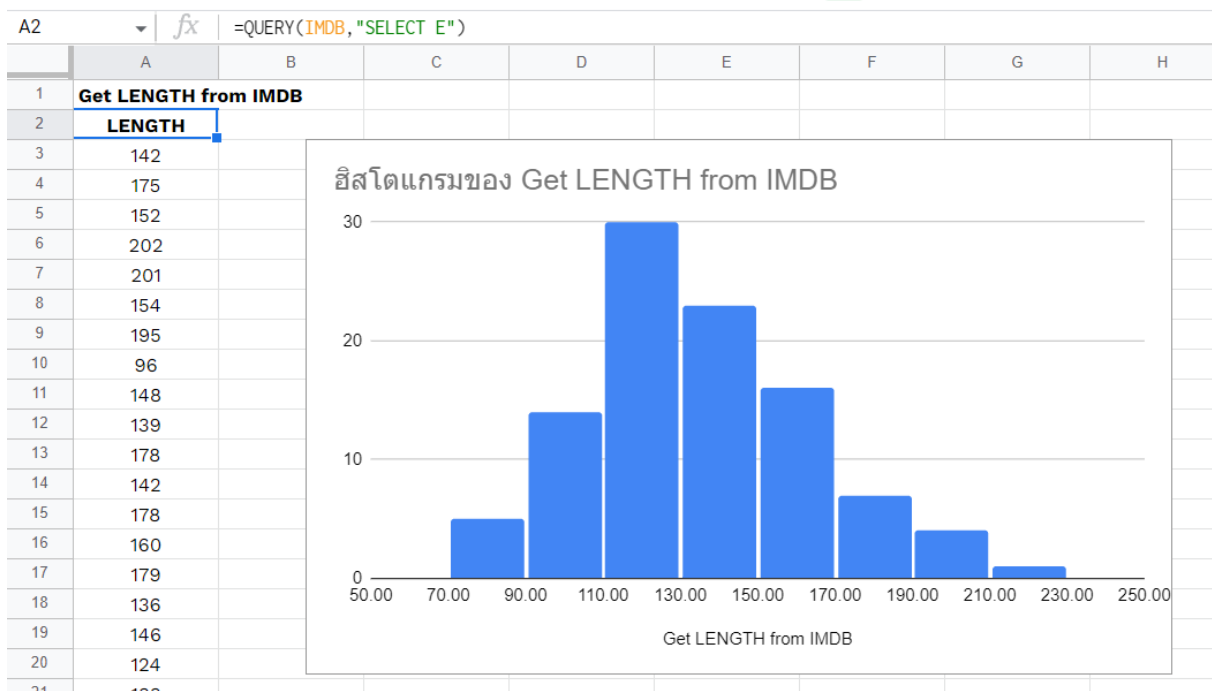
-เราสามารถใช่ Scorecard Chart ในการแสดง KPI ได้ เช่น:



-เราสามารถเลือกช่วง (Range) มาเป็น Key Value ได้เช่นกัน แล้วค่อย Aggregate ได้แบบไม่ต้องพิมพ์สูตร



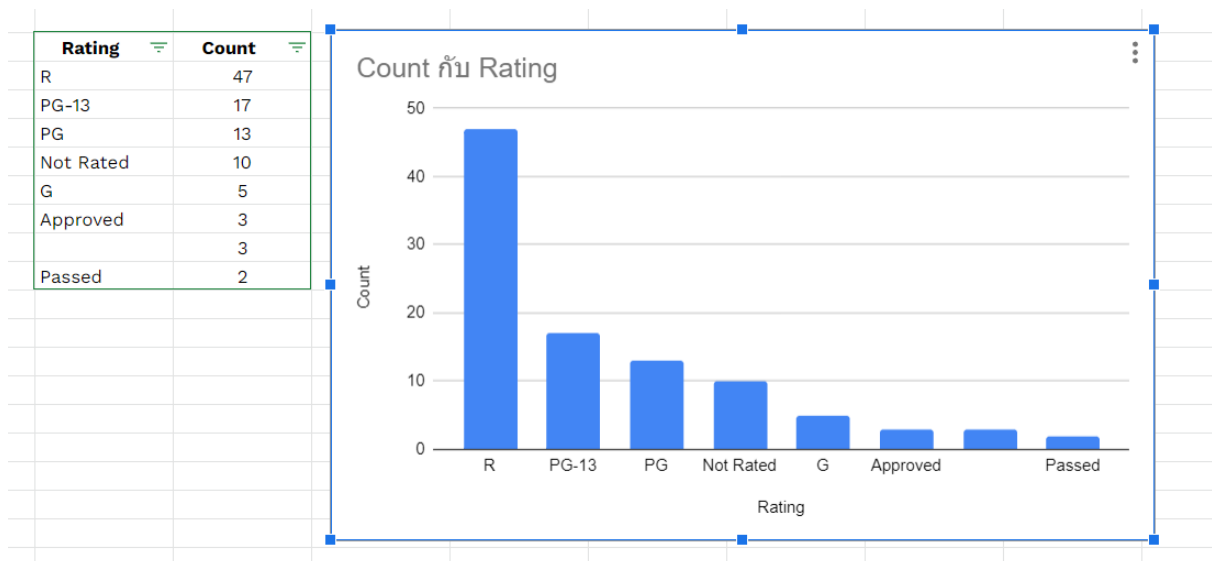
-เราสามารถใช่ Google Sheets ทำกราฟ Histogram ได้ และปรับแต่งลักษณะของกราฟได้ตามต้องการ เช่นสีหรือ Bucket Size



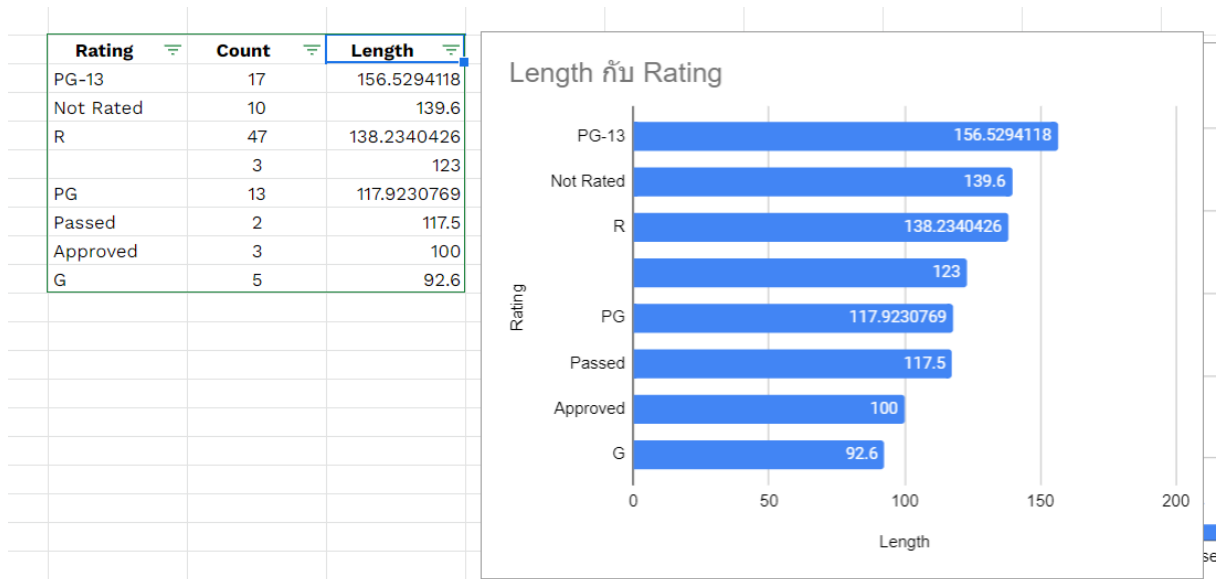
-วิธีการนับ Blank Cell คือการ Copy แล้ว Paste Special ให้เหลือแต่ค่า (ไม่ใช่สูตร) จากนั้น ที่ช่อง Blank ให้พิมพ์ "=" ดังนี้:

D11					
	A	B	C	D	E
1	Get LENGTH from IMDB				
2	RATING	LENGTH	=QUERY(IMDB,"SELECT D, E")		
3	R	142			
4	R	175			
5	PG-13	152		R	47
6	R	202		PG-13	17
7	PG-13	201		Approved	3
8	R	154		PG	13
9	R	195		Not Rated	10
10	Approved	96		G	5
11	PG-13	148			3
12	R	139		Passed	2
13	PG-13	178			
14	PG-13	142			

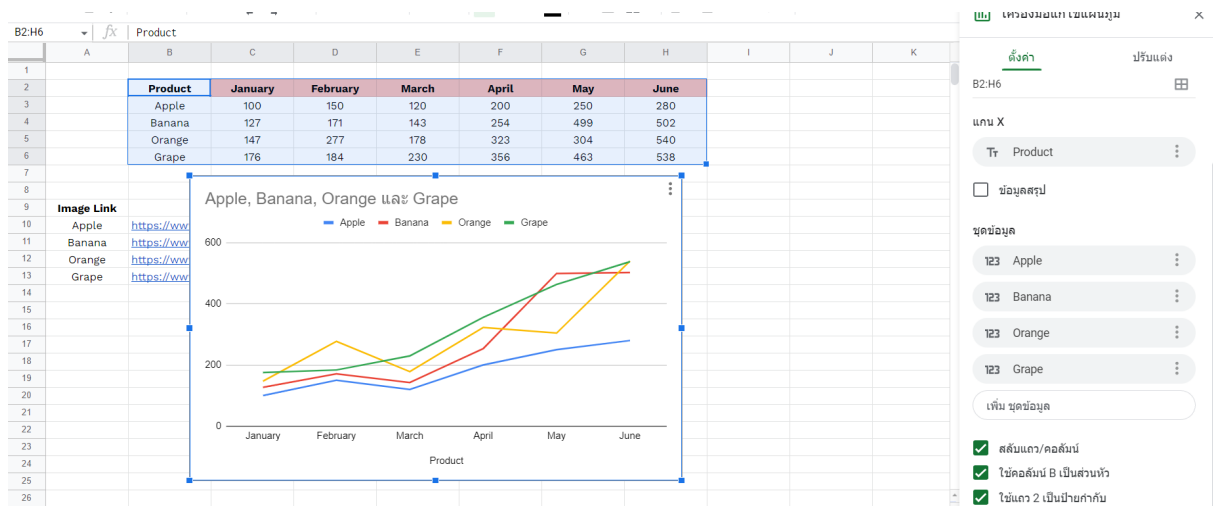
-ข้อดีของการทำ Summary Table มาก่อนคือ เราสามารถ Sort Data จากมากไปน้อยหรือน้อยไปมากได้ตามความต้องการ



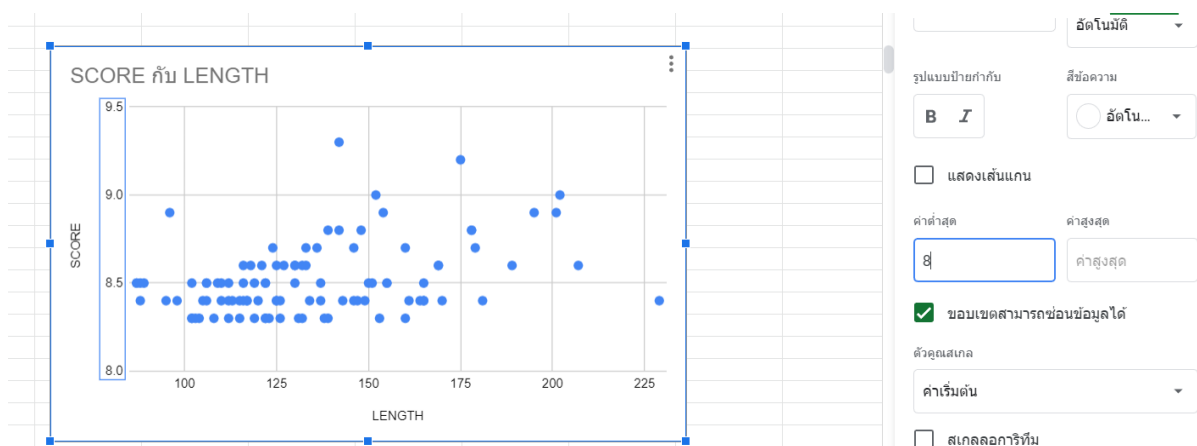
-เราสามารถแสดง Label ใน Bar Chart ได้

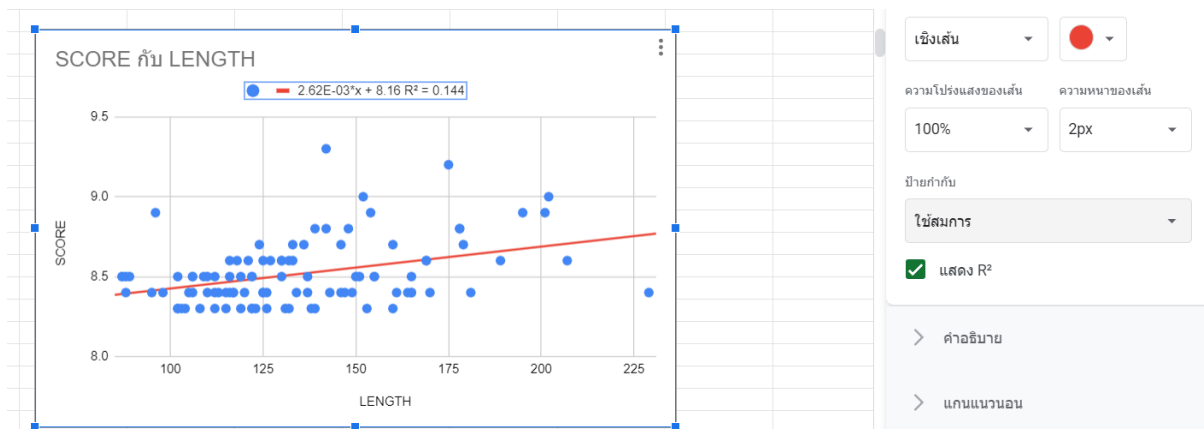


-ข้อมูลที่เรานิยมใช้ใน Line Chart จะเป็น Time Series

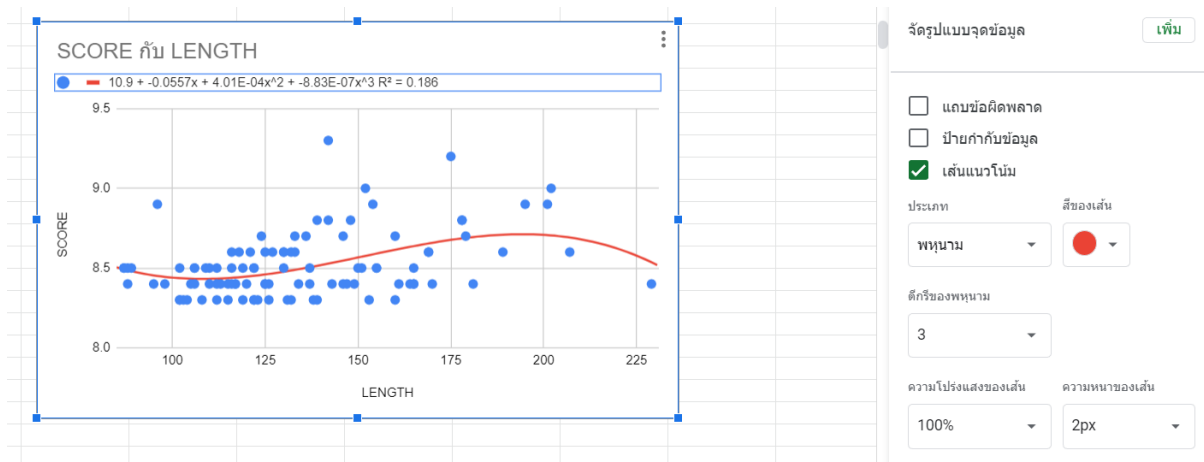


-Scatter Plot นิยมใช้ในการหาความสัมพันธ์ระหว่างสองตัวแปร ใน Google Sheets เราสามารถตั้งค่าให้ค่าต่ำสุดของ Plot เป็นค่าอื่นที่ไม่ใช่ 0 ได้ และเราสามารถแสดง Trend line พร้อมกับสมการประกอบได้



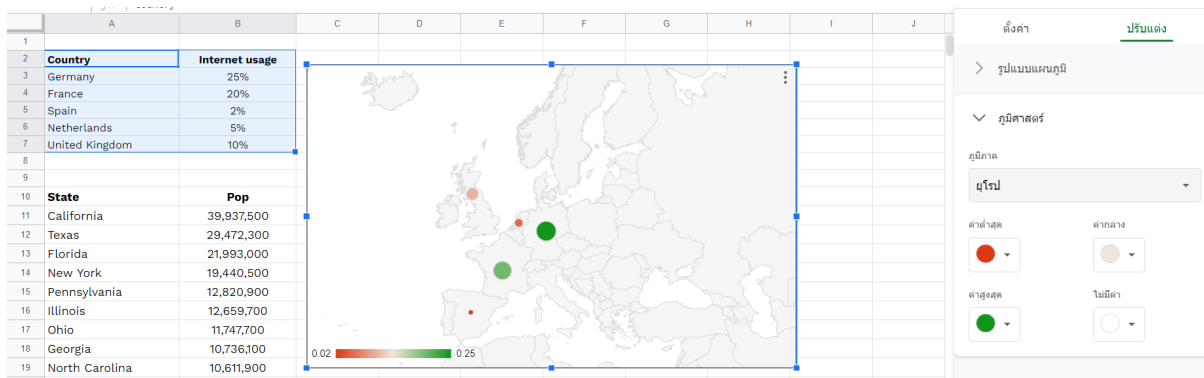


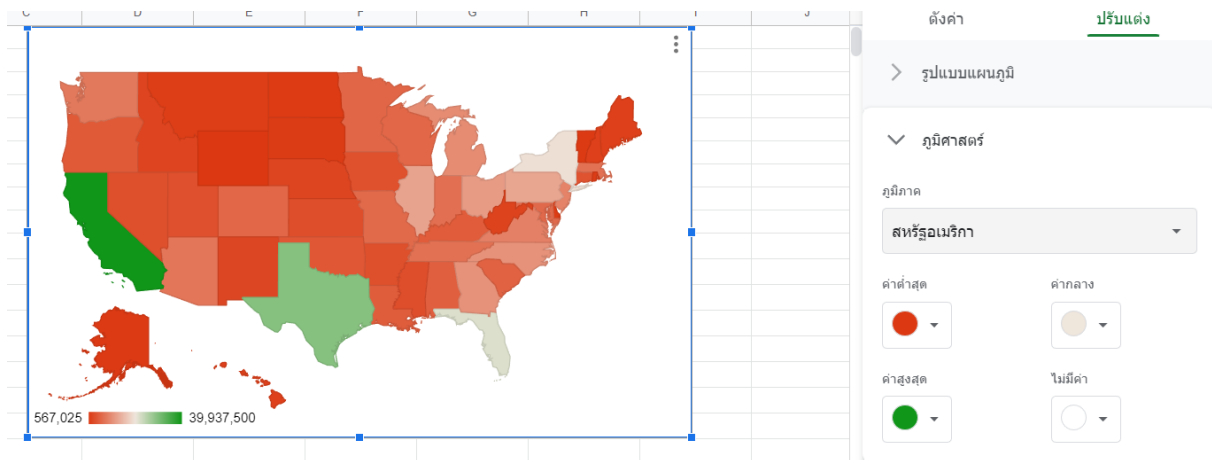
-เราสามารถเปลี่ยนเส้นแนวโน้มเป็นพหุนามยกกำลังได้ ซึ่งการใช้งานจริง เราไม่ต้องการสมการที่ Overfit กับข้อมูลที่มีอยู่จนมากเกินไป ดังนั้น ใช้สมการกำลัง 3 หรือกำลัง 4 ก็เพียงพอต่อการใช้งานจริงแล้ว



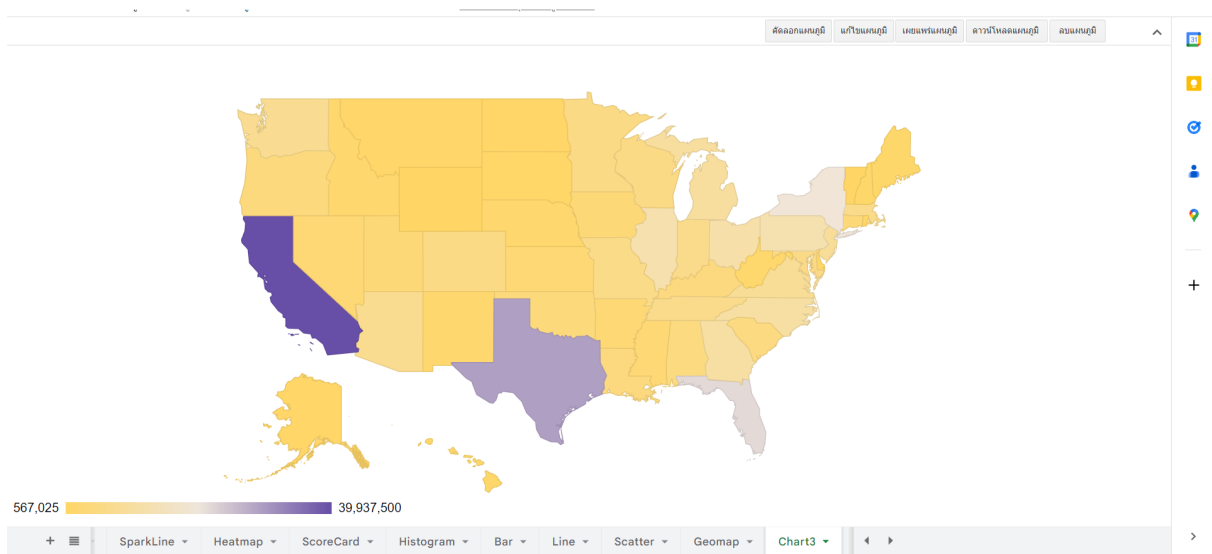
-Scatter Plot ไม่จำเป็นต้อง Normalize

-ใน Google Sheets เราสามารถสร้าง Geomap ได้ และเลือกภูมิภาคเฉพาะในการแสดงผลได้เช่นกัน

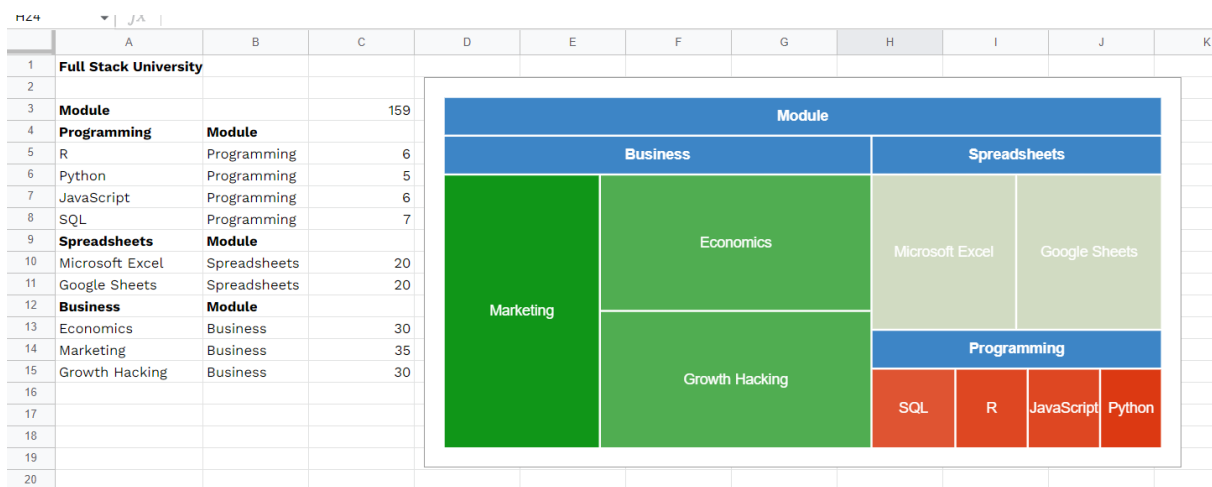


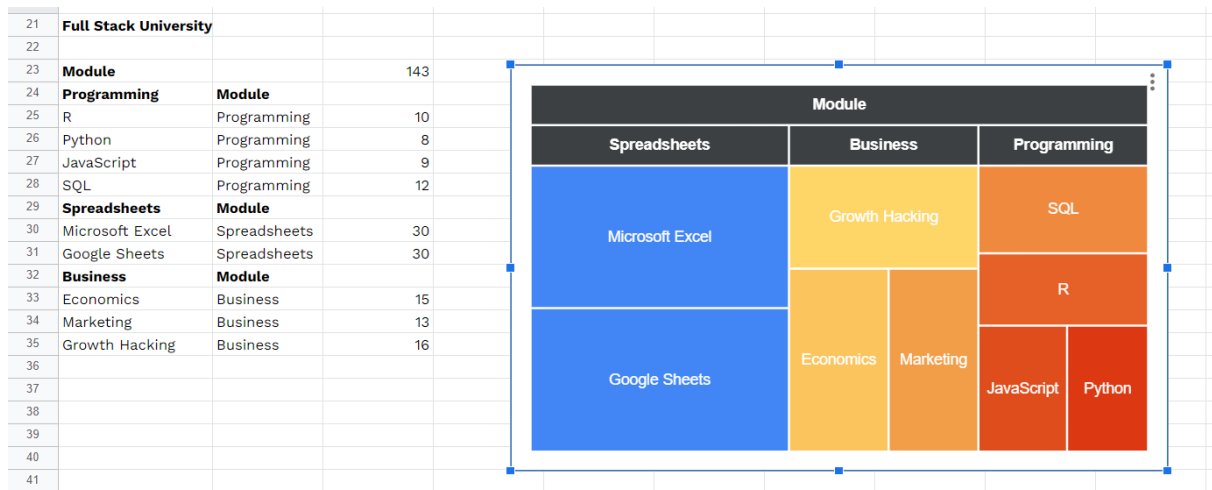


-เราสามารถ Move to own sheet แผนภูมิที่เราต้องการ เพื่อสร้างเป็น Dashboard ใช้ได้



-Treemap จำเป็นต้อง Format ข้อมูลของเราให้ตรงตามเงื่อนไขก่อน ถึงจะออกมาอย่างมีประสิทธิภาพ (Parent-Child Hierachy)





-Treemap ใช้ในการดู Contribution ในแต่ละภาคส่วนได้

Data Visualization in R

[Data Visualization with R.pdf](#)

-ภาษา R เป็นภาษาที่เหมาะสมสำหรับการทำ Data Visualization (ggplot2 ใน tidyverse)

-ข้อดีของ ggplot2 คือ User สามารถสร้าง Template ทำทีเดียวใช้ได้หลายงาน

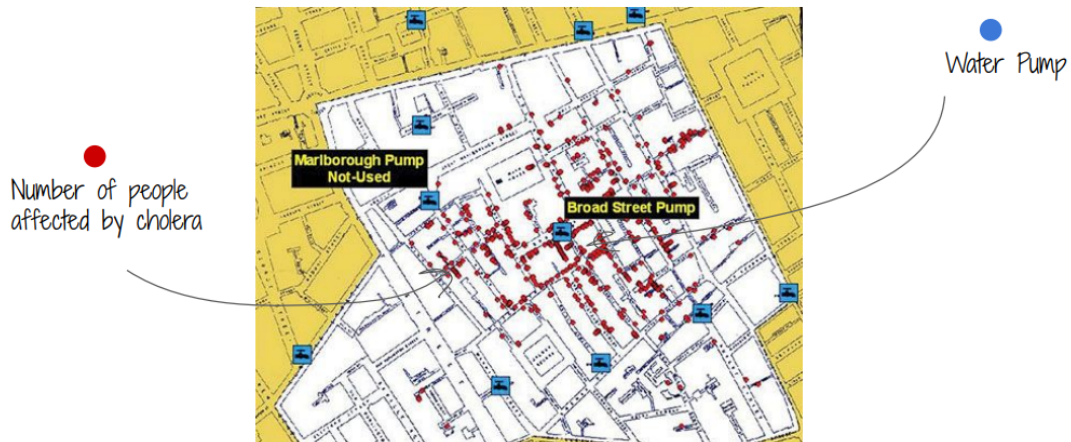
-Data Visualization ทำให้เห็นภาพความสัมพันธ์ของข้อมูลได้ชัดเจนกว่าการมองข้อมูลดิบตรง ๆ และในบางครั้ง ค่าทางสถิติ เช่น ค่าเฉลี่ย S.D. หรือค่า Correlation ก็ไม่ได้บ่งบอกทุกอย่างเสมอไป [Anscombe's Quartet]

"The greatest value of a picture is when it forces us to notice what we never expected to see."

-John Tukey

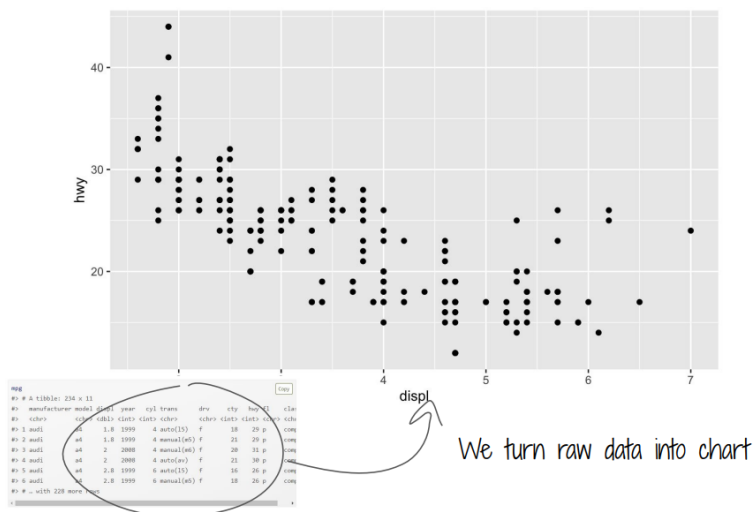
-Case Study: John Snow (จำนวนคนที่เป็นโรคหิวาต์ กับตำแหน่งของบ่อน้ำ)

Case Study - John Snow



-EDA (Exploratory Data Analysis)

Exploratory Data Analysis



EDA can be done in two ways

1. Numerical Method
 - a. summary stats
 - b. basic modeling
2. Graphical Method

-Chart ที่เราทำให้ตัวเองเห็นคนเดียว หน้าตาไม่ต้องสวยมากก็ได้ (เช่น เป็นขาว-ดำ) แต่ถ้าจะทำให้คนอื่นดู ก็ควรทำให้สวยงามหน่อย

-getwd(): ทำให้เราเห็นตำแหน่ง Working Directory ปัจจุบัน

-Code:

```
#Get working directory
getwd()
```

```

#Library tidyverse
library(tidyverse)

#Basic plot in Base R
#Histogram
hist(mtcars$mpg)

#Analyzing horsepower
#Histogram - One Quantitative Variable
hist(mtcars$hp)
mean(mtcars$hp)
median(mtcars$hp)

#
str(mtcars)
mtcars$am <- factor(mtcars$am,
                    levels = c(0,1),
                    labels = c("Auto", "Manual"))

#Bar Plot - One Qualitative Variable
barplot(table(mtcars$am))

#Box Plot
boxplot(mtcars$hp)
fivenum(mtcars$hp)

min(mtcars$hp)
quantile(mtcars$hp, probs = c(0.25, 0.5, 0.75))
max(mtcars$hp)

#Whisker Calculation
Q3 <- quantile(mtcars$hp, probs = 0.75)
Q1 <- quantile(mtcars$hp, probs = 0.25)
IQR_HP <- Q3 - Q1

Q3 + 1.5 * IQR_HP
Q1 + 1.5 * IQR_HP

boxplot.stats(mtcars$hp, coef = 1.5)

#Filter Outliers

```

```

mtcars_no_out <- mtcars %>%
  filter (hp < 335)
boxplot(mtcars_no_out$hp)

#Boxplot 2 Variables
#Qualitative X Quantitative
boxplot(mpg ~ am,
        data = mtcars,
        col = c("gold", "salmon"))

#How to restore dataframe
data(mtcars)

#Scatterplot
#2 X Quantitative
plot(mtcars$hp,
     mtcars$mpg,
     pch = 16,
     col = c("red", "blue"),
     main = "My first scatter plot",
     xlab = "Horsepower",
     ylab = "Miles per Gallon")
cor(mtcars$hp, mtcars$mpg)
lm(mpg ~ hp, data = mtcars)

##ggplot2
#tidyverse
library(tidyverse)

#Our very first plot in ggplot2
ggplot(data = mtcars,
       mapping = aes(x = hp, y = mpg)) +
  geom_point() +
  geom_smooth() +
  geom_rug()

#Simpler:
ggplot(mtcars,
      aes(hp, mpg)) + geom_point(size = 3,
                                col = "blue",
                                alpha = 0.5)

```

```

#Histogram with ggplot2 (default bin = 30)
ggplot(mtcars,
       aes(hp)) + geom_histogram(bins = 10,
                                fill = "salmon",
                                alpha = 0.7)

#boxplot
ggplot(mtcars,
       aes(hp)) + geom_boxplot()

p <- ggplot(mtcars, aes(hp))
p + geom_histogram(bins = 10)
p + geom_density()

#Box plot by groups
diamonds %>% count(cut)
ggplot(diamonds,
       aes(cut)) + geom_bar(fill = "#0366fc")

ggplot(diamonds,
       aes(cut, fill = color)) + geom_bar(position = "fill")

#Scatter Plot
set.seed(99)
smol_diamonds <- sample_n(diamonds, 5000)

ggplot(smol_diamonds,
       aes(carat, price)) + geom_point()

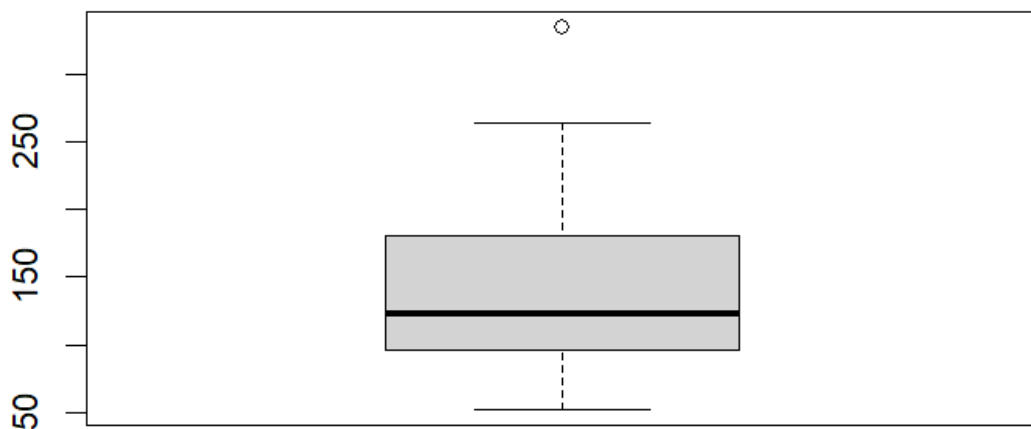
#Facet: Small Multiples
ggplot(smol_diamonds,
       aes(carat, price)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  facet_wrap(~color, ncol = 2) +
  theme_minimal() +
  labs(title = "Relationship between carat and price by colour",
       x = "Carat",
       y = "Price USD",
       caption = "Source: Diamonds from ggplot2 package")

#Final Example:

```

```
ggplot(smol_diamonds,
      aes(carat, price, col = cut)) +
  geom_point() +
  facet_wrap(~color, ncol = 2) +
  theme_minimal()
```

-Box Plot สามารถใช้หา Outlier (Extreme Value) ในข้อมูลของเราได้ จะแสดงออกมาเป็นวงกลมเล็ก ๆ นอกช่วง Box Plot



-fivenum() ใช้หา 5 ค่า ได้แก่:

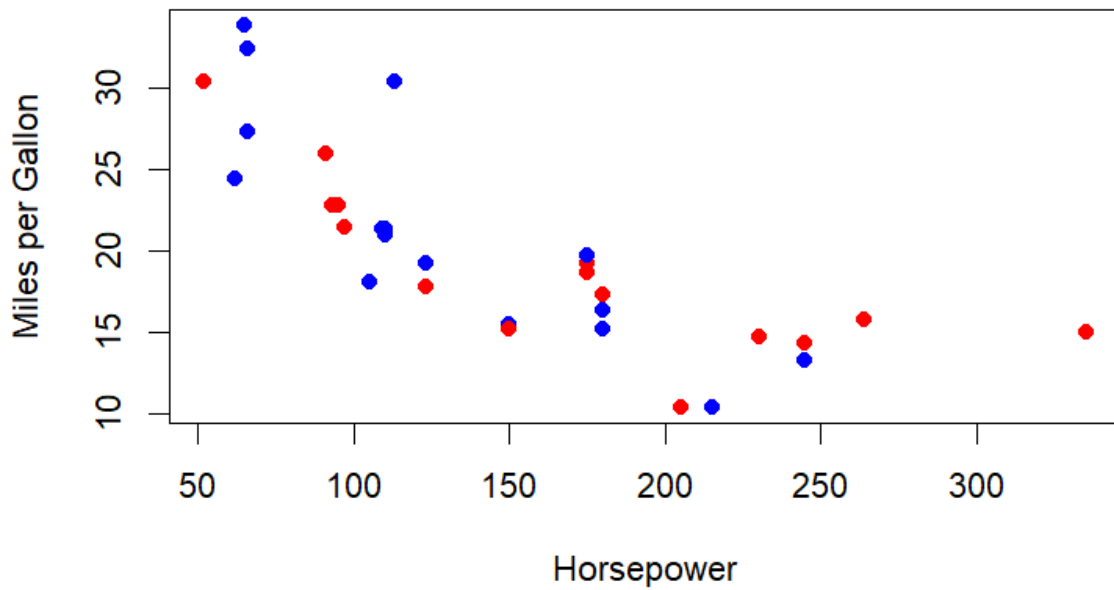
1. ค่าต่ำสุด (Min)
2. ค่า Quartile ที่ 1 (Q1)
3. ค่า Quartile ที่ 2 (Q2/Median)
4. ค่า Quartile ที่ 3 (Q3)
5. ค่าสูงสุด (Max)

-แขนที่เป็นเส้นประใน Box Plot มีชื่อเรียกว่า Whisker

-ใน R เราสามารถ Save รูป Chart ที่เราทำได้หลายนามสกุล เช่น PNG, JPEG, TIFF เป็นต้น

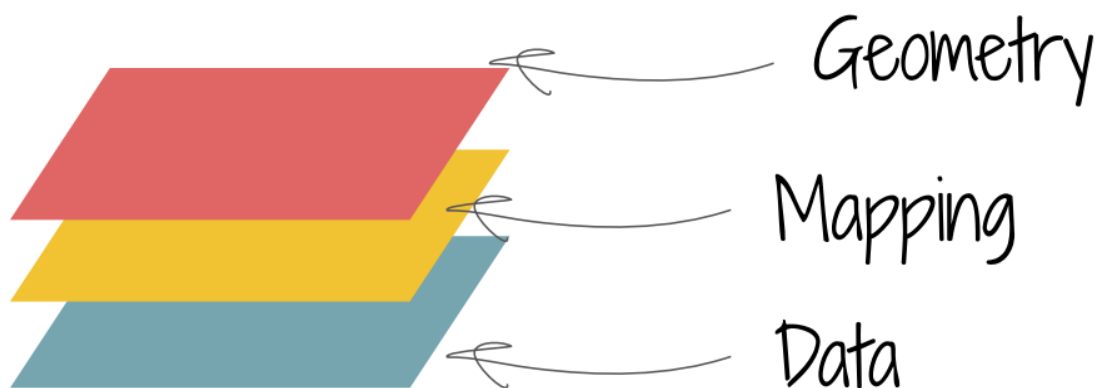
Scatter Plot ด้วย Base R:

My first scatter plot



-ggplot = grammar of graphics plot

Intro to ggplot2



1. Data: ข้อมูล
2. Mapping: การดึง Data ไป Map ที่แกน x แกน y หรือจุดต่าง ๆ

3. Geometry: รูปแบบการ Visualization ของ Chart เช่น Bar Chart

*ggplot template:

```
#This template can generate more than 30 charts  
ggplot(data = ..., mapping = aes(...)) + geom_...()
```

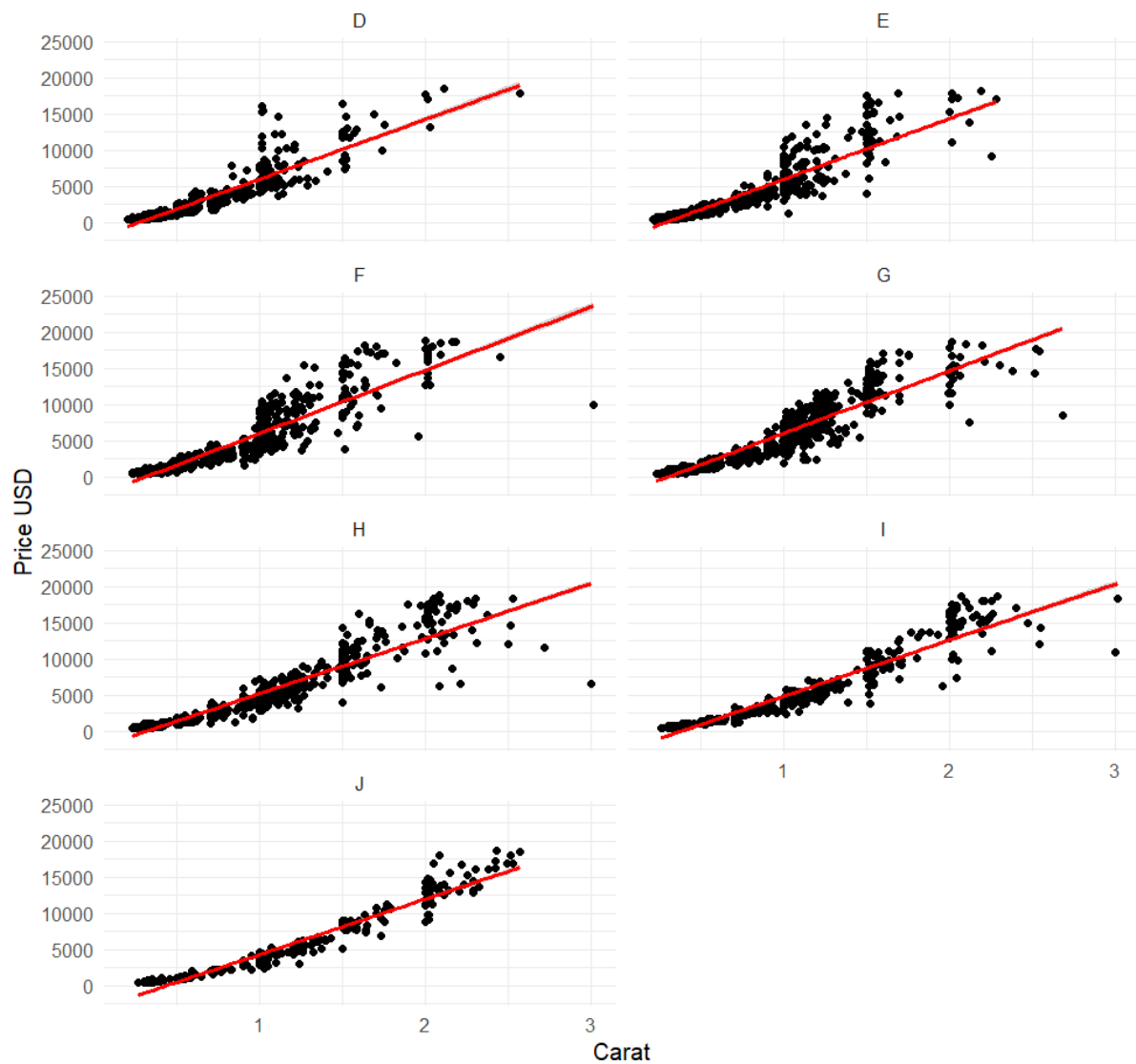
-การเลือก Chart มีกฎ 3 ข้อ

1. ข้อมูลมีกี่ตัวแปร (1, 2, หรือมากกว่านั้น)
2. ประเภทข้อมูลที่เราจะนำไปทำ Chart เป็น Quantitative (Numeric/Continuous) หรือ Qualitative (Categorical/Discrete)
3. สิ่งที่ยากให้คนดูเข้าใจ เป้าหมายของเราในการสร้าง Chart นั้น ๆ (focus on audience)

-alpha ค่ายิ่งใกล้เลข 1 จะยิ่งทึบ ยิ่งใกล้เลข 0 จะยิ่งจาง

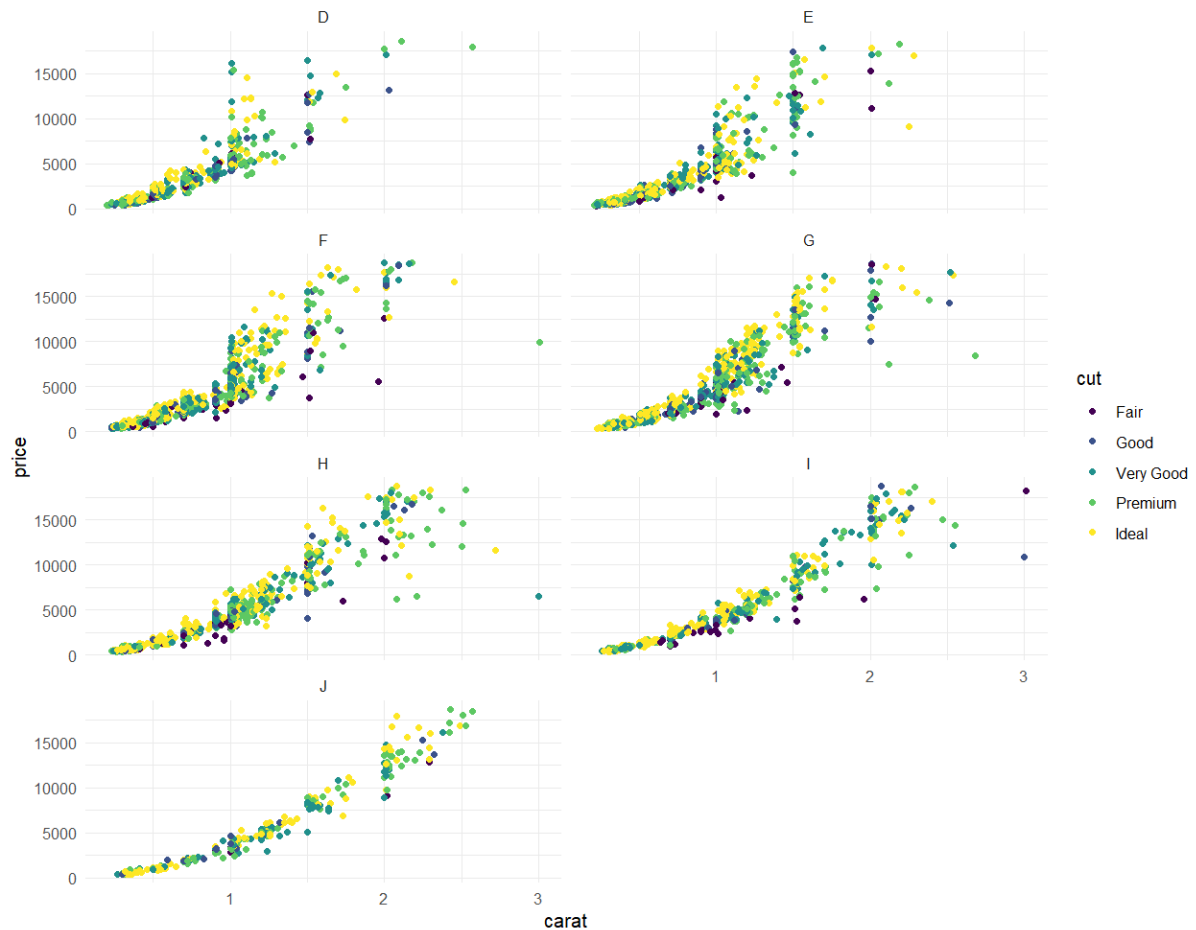
-การซูมตัวอย่างทำให้การ Plot Chart ง่ายขึ้น

Relationship between carat and price by colour



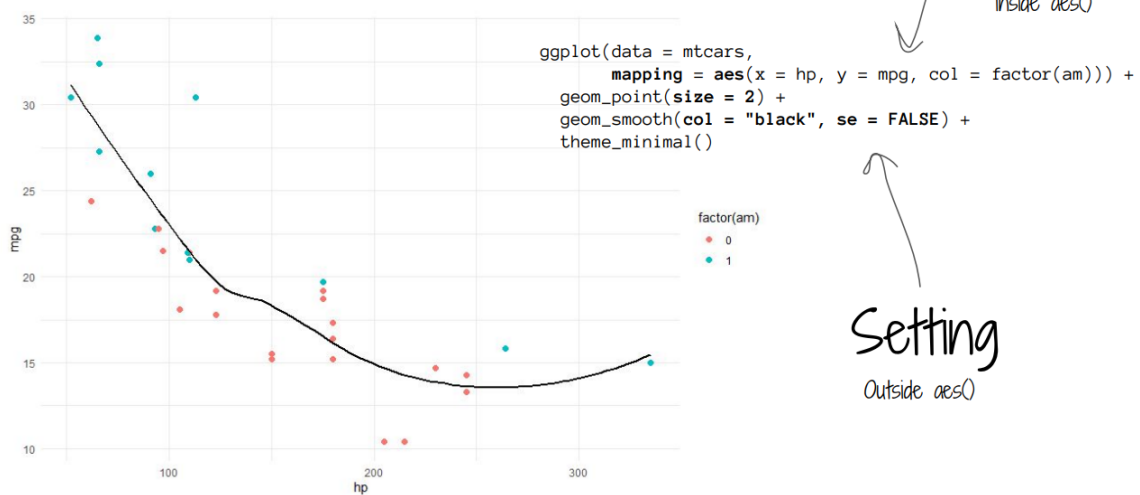
Source: Diamonds from ggplot2 package

-Map Variable กับสี



-Mapping VS Setting:

Setting vs. Mapping



*ตัวทวน (tilde): ~

[ggplot2 cheat sheet]

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/66f9e9ee-4fc2-4100-a31a-6b987d2fd869/data-visualization-ggplot2.pdf>

```
## get working directory
getwd()

## library tidyverse
library(tidyverse)

## basic plots (base R)

hist(mtcars$mpg)

## Analyzing horse power
## Histogram - One quantitative Variable
hist(mtcars$hp)
mean(mtcars$hp)
median(mtcars$hp)

str(mtcars)
mtcars$am <- factor(mtcars$am,
                    levels = c(0,1),
                    labels = c("Auto", "Manual"))

## Bar Plot - One Quantitative Variable
barplot(table(mtcars$am))

## Box Plot
boxplot(mtcars$hp)
fivenum(mtcars$hp)
min(mtcars$hp)
quantile(mtcars$hp, probs = c(.25, .5, .75))
max(mtcars$hp)

## Whisker Calculation
Q3 <- quantile(mtcars$hp, probs = .75)
```

```

Q1 <- quantile(mtcars$hp, probs = .25)
IQR_hp <- Q3 - Q1

Q3 + 1.5*IQR_hp
Q1 - 1.5*IQR_hp

boxplot.stats(mtcars$hp, coef= 1.5)

## filter out outliers
mtcars_no_out <- mtcars %>%
  filter(hp < 335)

boxplot(mtcars_no_out$hp)

## Boxplot 2 variables
## Qualitative x Quantitative
data(mtcars)
mtcars$am <- factor(mtcars$am,
                    levels = c(0,1),
                    labels = c("Auto", "Manual"))
boxplot(mpg ~ am, data = mtcars,
        col = c("gold", "salmon"))

## Scatter Plot
## 2 x Quantitative
plot(mtcars$hp, mtcars$mpg, pch = 16,
     col="blue",
     main = "Relationship between HP and MPG",
     xlab = "Horse Power",
     ylab = "Miles Per Gallon")

cor(mtcars$hp, mtcars$mpg)
lm(mpg ~ hp, data = mtcars)

#####
## ggplot2
##
library(tidyverse)

## First Plot
ggplot(data = mtcars, mapping = aes(x = hp, y = mpg)) +

```

```

    geom_point() +
    geom_smooth() +
    geom_rug()

## ย้ํา code
ggplot(mtcars, aes(hp, mpg)) +
  geom_point()

## Change size
## alpha 0-1
ggplot(mtcars, aes(hp, mpg)) +
  geom_point(size = 3, col = "blue", alpha = 0.2)

ggplot(mtcars, aes(hp)) +
  geom_histogram(bins = 10, fill = "red", alpha = 0.5)

ggplot(mtcars, aes(hp)) +
  geom_boxplot()

p <- ggplot(mtcars, aes(hp))
p + geom_histogram(bins = 10)
p + geom_density()
p + geom_boxplot()

## box plot by groups
diamonds %>%
  count(cut)

ggplot(diamonds, aes(cut)) +
  geom_bar(fill = "#0366fc")

## stack เป็นค่า Default
ggplot(diamonds, mapping = aes(cut, fill = color)) +
  geom_bar(position = "fill")

#### SCATTER PLOT

ggplot(diamonds, aes(carat, price)) +
  geom_point()

##### set.seed lock การสุ่ม
##### สุ่มตัวอย่างใช้ sample_n จาก 50k สุ่มมา 5k

```

```

set.seed(99)
small_diamonds <- sample_n(diamonds, 5000)

ggplot(small_diamonds, aes(carat, price)) +
  geom_point()

#### Facet : small ,multiples
ggplot(small_diamonds, aes(carat, price)) +
  geom_point() +
  geom_smooth(method = "lm", col="red") +
  facet_wrap(~color, ncol=2) +
  theme_minimal() +
  labs(title = "Relationship between carat and price by color",
        x = "Carat",
        y = "Price USD",
        caption = "Source : Diamonds from ggplot2 package")

### Final Example
ggplot(small_diamonds, aes(carat, price, col=cut)) +
  geom_point(size = 3, alpha = 0.2) +
  facet_wrap(~color, ncol=2) +
  theme_minimal()

```