EDA 개요

5기 한영웅

목차 : 1. EDA 란?
2. 데이터의 특성과 분석 수단
3. 간단한 실습
4. 발표예제

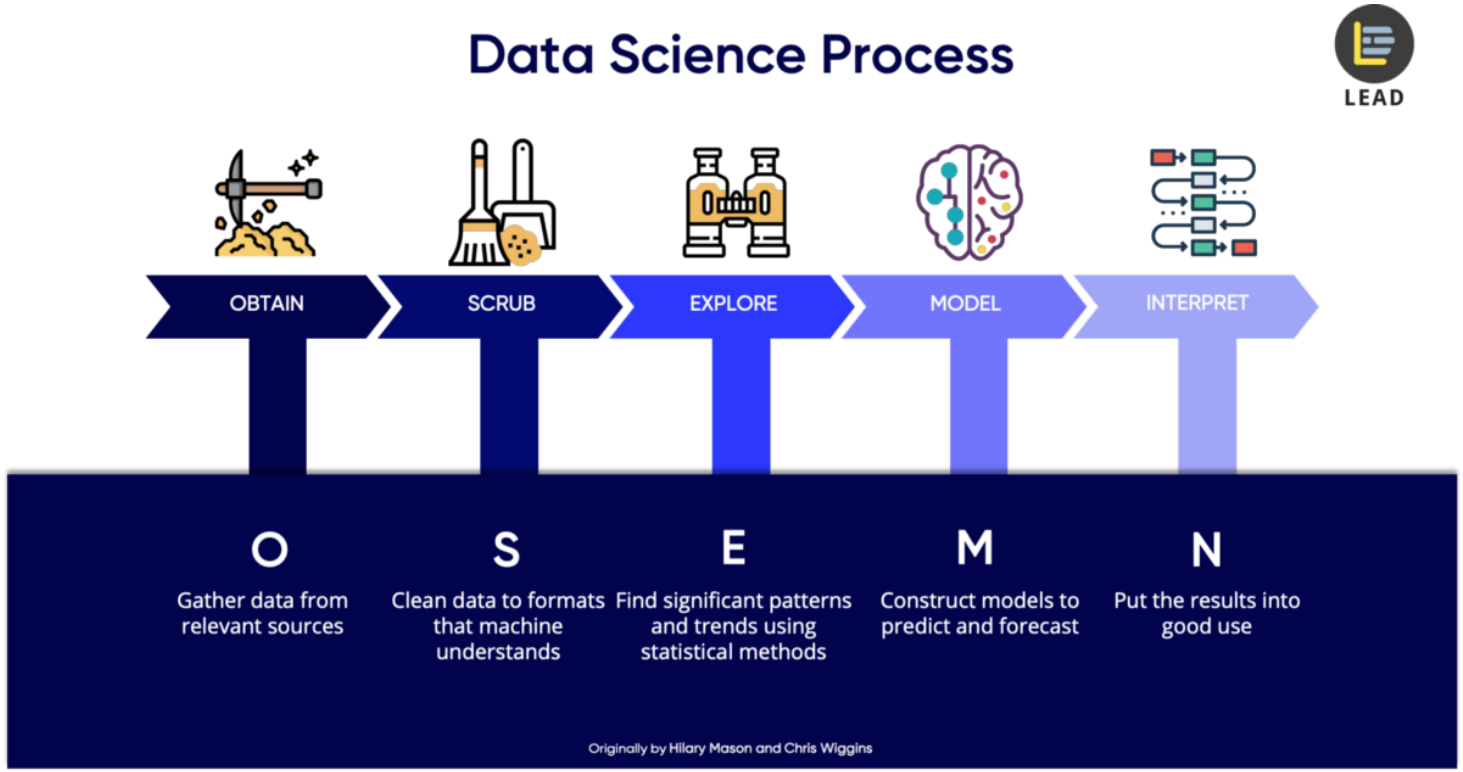
EDA

***E**xploratory*

***D**ata*

***A**nalysis*

시각화와 각종 정보들을 바탕으로 데이터의 패턴을 발견하고 그곳에서 의미있는 정보를 도출해내는 작업.



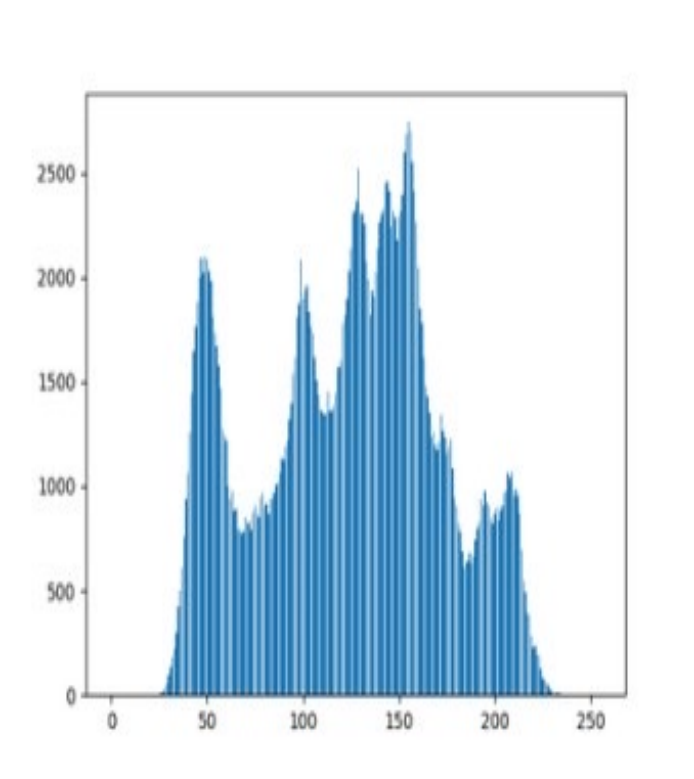
데이터의 특성과 분석 수단

변수의 갯수와 특성에 따른 데이터 분류

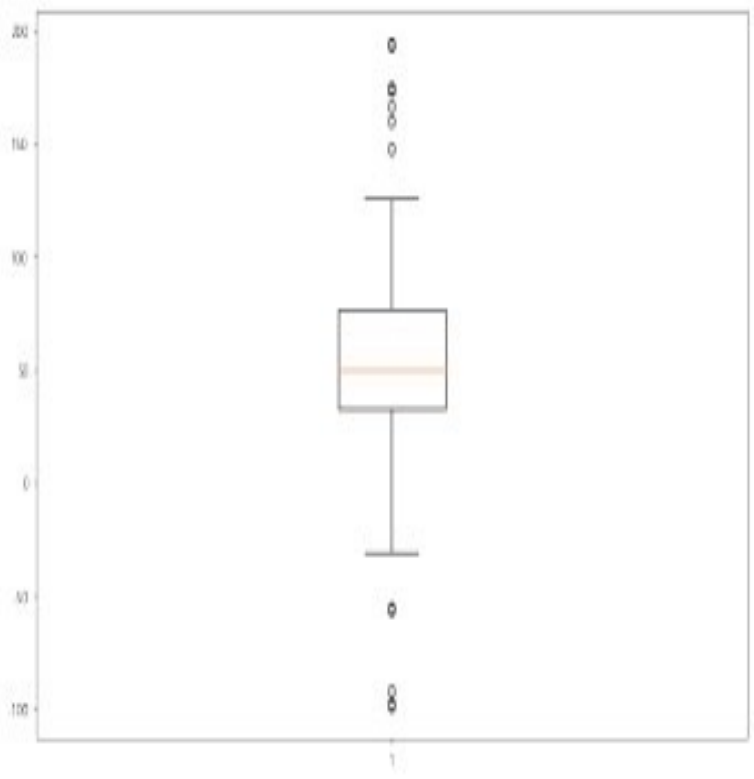
수치형 데이터 - 일변량	요약 통계량, histogram, Boxplot
범주형 데이터 -일변량	Barplot, value_counts
수치형 데이터 -다변량	Correlation, scatter plot
범주형 데이터 -다변량	Side-by-side boxplot 등
비정형 데이터	적합한 분석모델이 필요

일변량분석 – 수치형변수

Histogram



Box plot



요약 통계량:
mean, max , median ,std 등

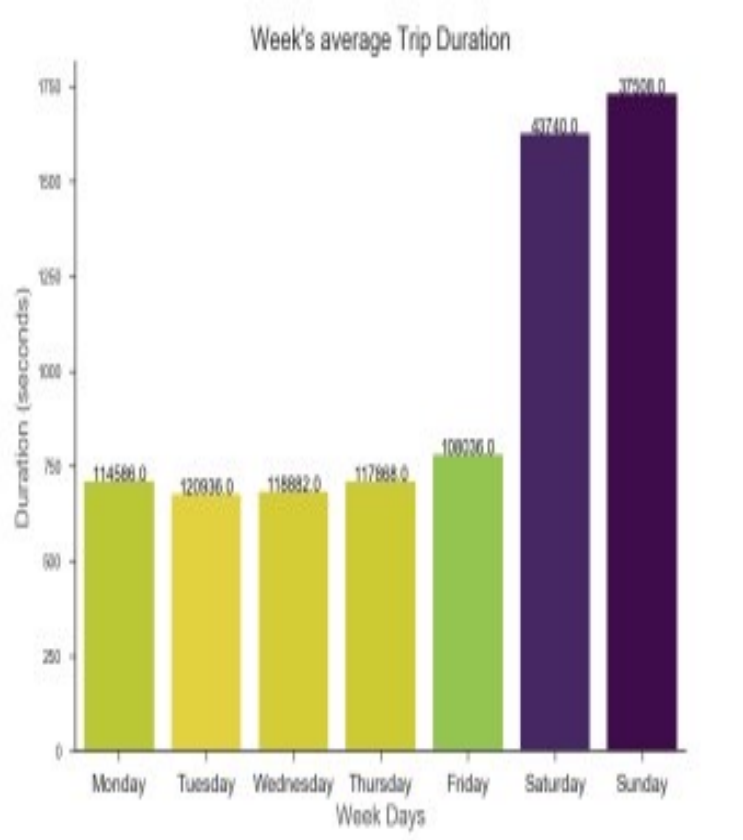
Stem and leaf plot

히스토그램을 통해서 해당 데이터의 분포를 예상해본다.
Ex) Normal , Beta , Gamma 등등

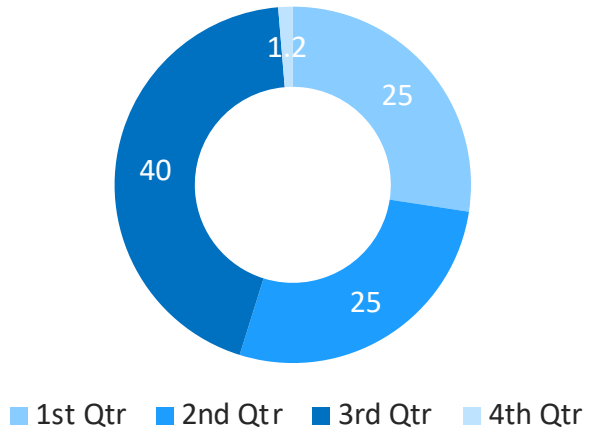
Tip: EDA에서는 Mean보다 Median을 많이 사용.
Why? Median이 Mean보다 이상치에 대한 민감도가 낮기 때문이다.
대표적으로 Boxplot의 box의 선은 median을 가리킨다.

일변량분석 – 범주형 변수

Barplot



Pie plot



Bar plot

Pie plot

.value_counts() 등 이용

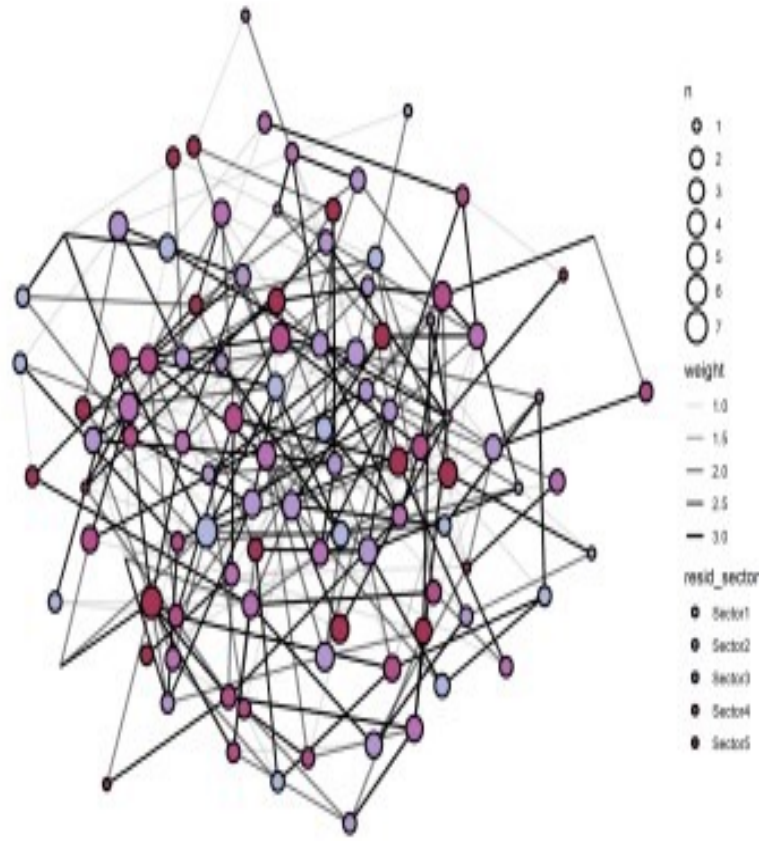
해당 범주의 도수를 이용한 분석!

절대적인 값이 사용되기도 하지만 주로 범주별로 상대적인 값이 사용됩니다.

일변량분석 – 비정형 데이터

Word cloud

Network

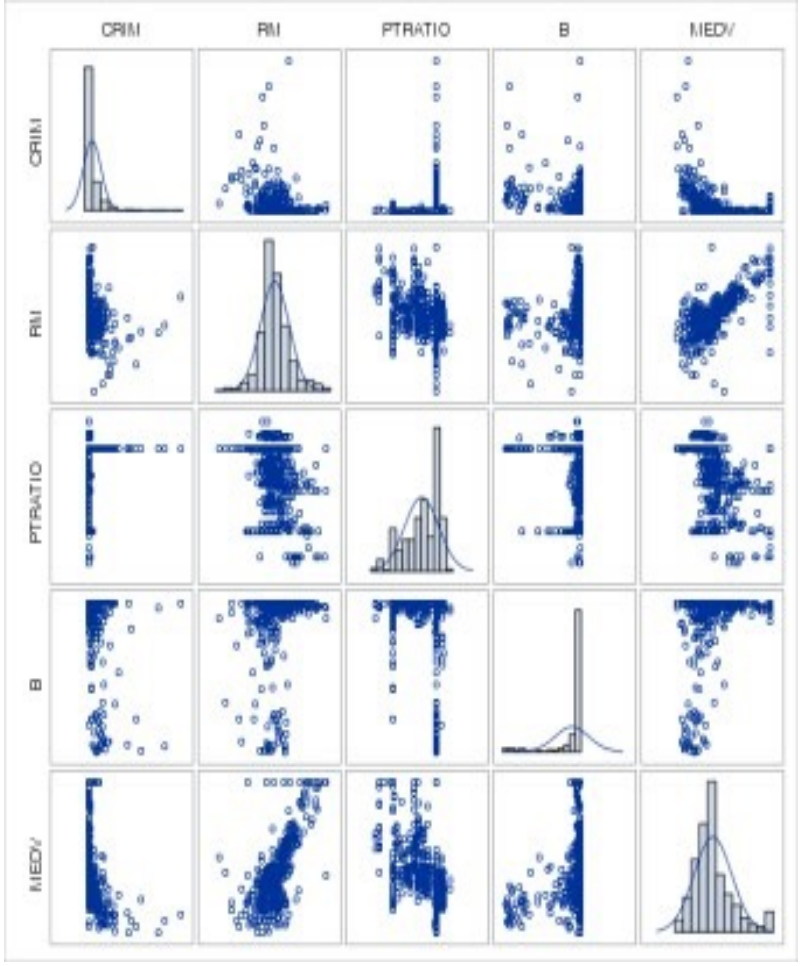


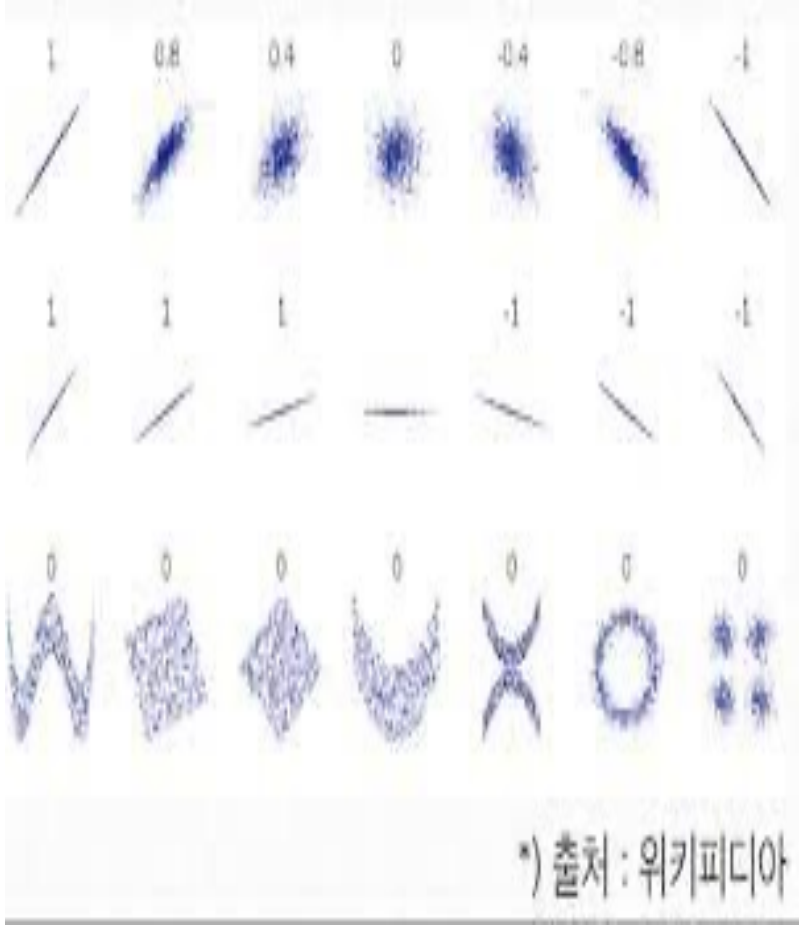
텍스트 데이터, 네트워크 데이터, 이미지 데이터 등!

Data-specific knowledge가 필요

다변량분석 – 수치형 데이터 vs 수치형 데이터

Scatter plot matrix





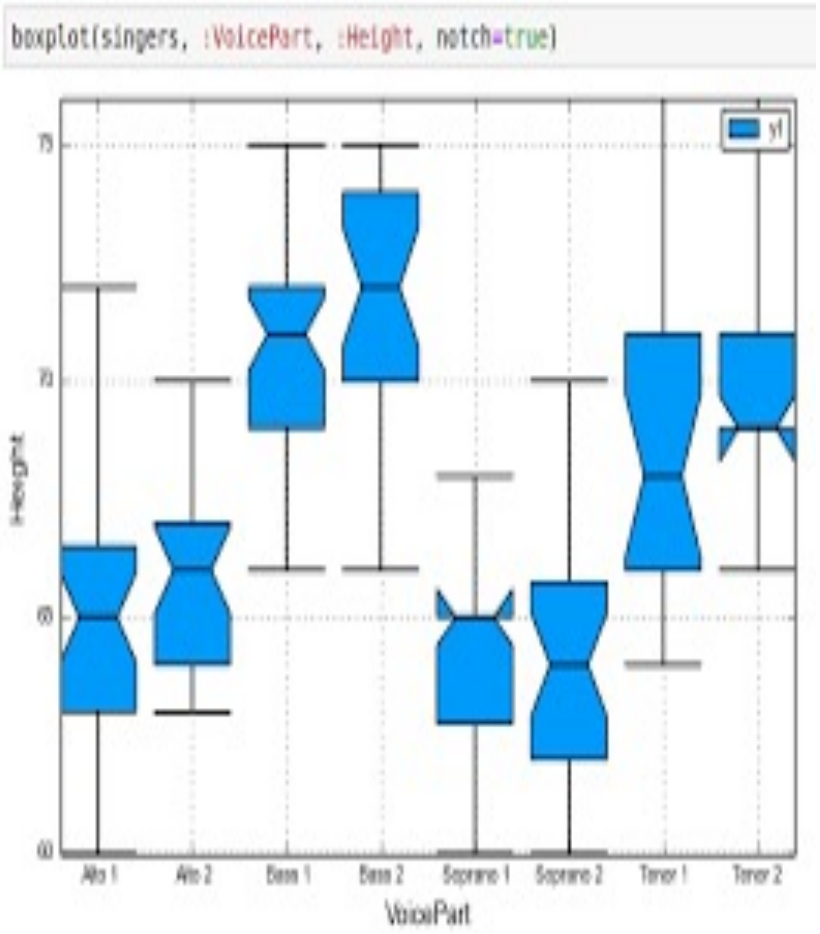
변수가 많을 경우, 어떤 변수쌍이 많은 연관성을 가지고 있는지 판단해야 합니다.

산점도를 꼭 그려서 모든 데이터쌍의 경향성을 눈으로 파악합시다.

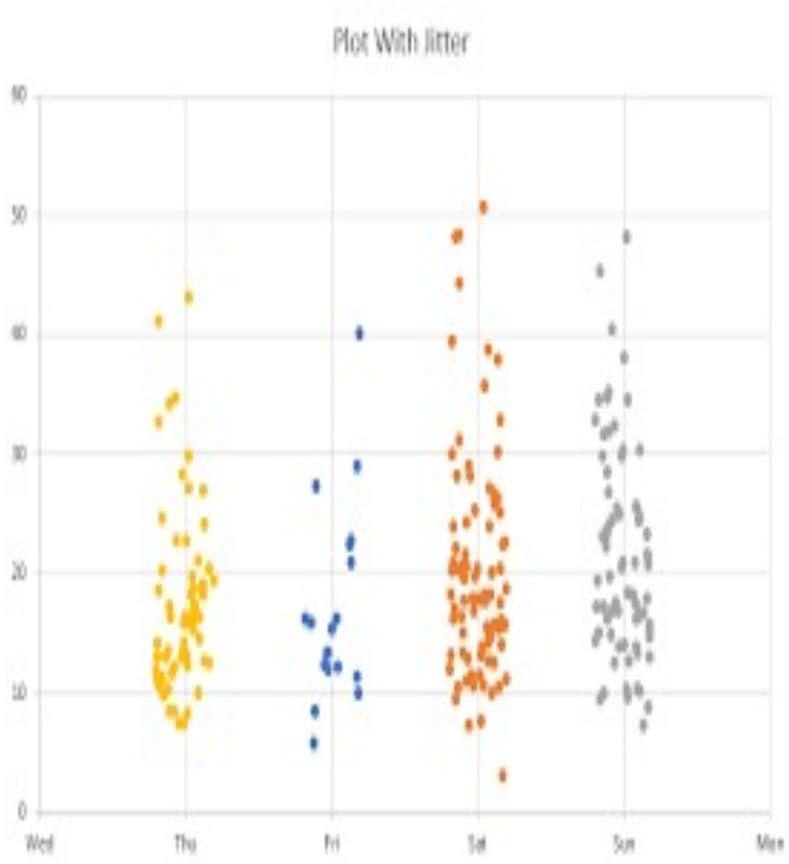
상관계수를 구하되, 상관계수만을 이용해서는 안 됩니다.

다변량분석 – 범주형 데이터 vs 수치형 데이터

Box plot



Jittered scatter plot



각 범주가 가진 데이터가 얼마나 다른지 확인!

Box plot

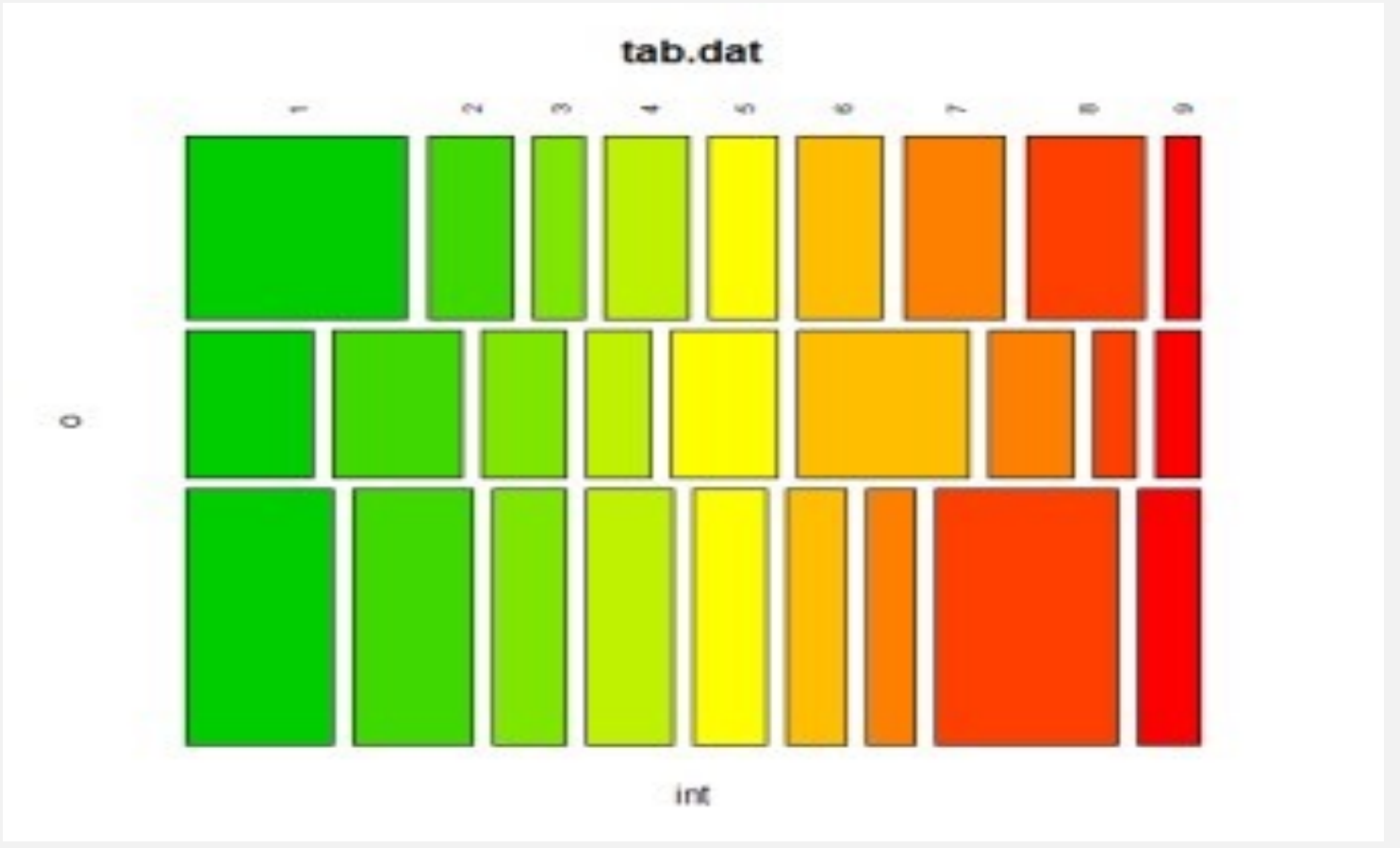
Jittered scatter plot

여러 Boxplot을 그려 각 범주를 비교할 때 notch를 이용하는 것이 좋습니다. !

.groupby() 와 pivot table을 주로 사용합니다

다변량분석 – 범주형 자료 vs 범주형 자료

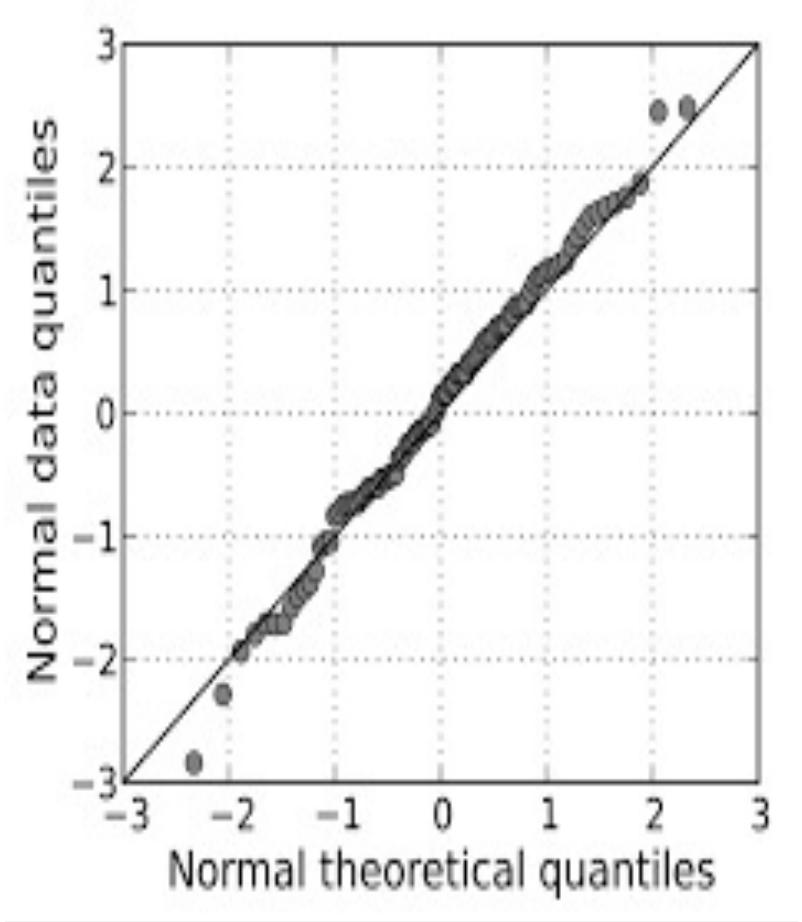
Mosaic Plot을 사용!



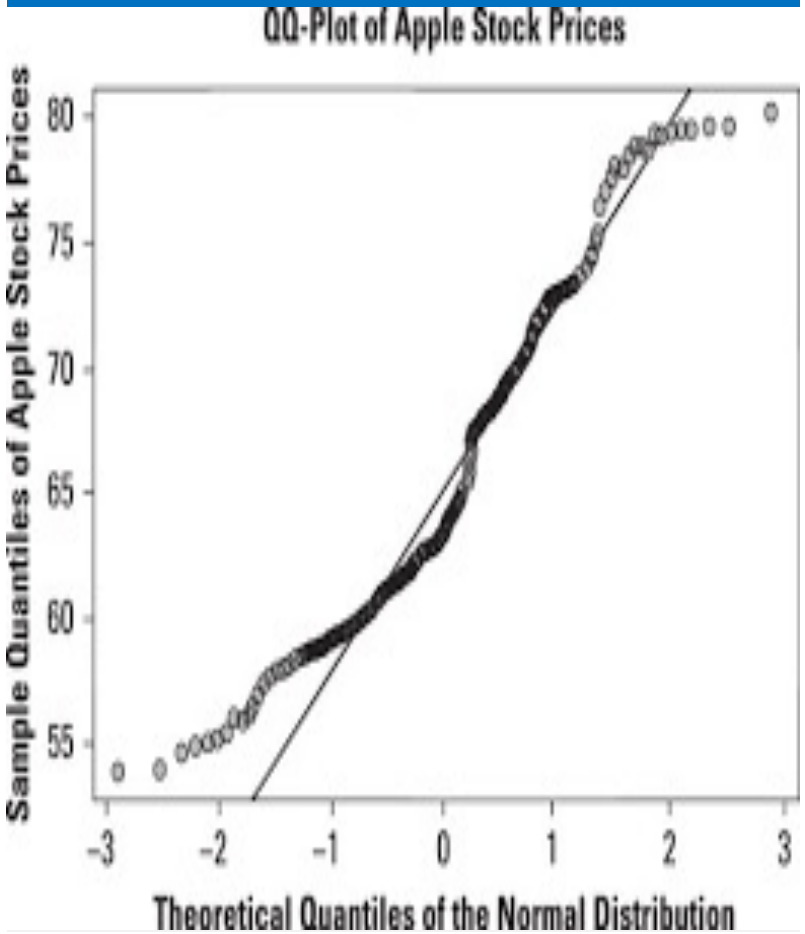
Mosaic Plot을 통해 시각화합니다.
이 때 면적을 통해 비교하면 됩니다.

Probability-plot(분포검정의 EDA버전)

QQplot-1



QQplot-2



우리가 배운 모델들과 데이터가 얼마나 적합되는지 판단하는 데에 도움이 되는 그림입니다.
당연히 모델에 완벽히 들어맞는 데이터는 없겠지요 따라서 이다, 아니다가 아니라 얼마나의 문제이겠지요. (통계학은 불확실성의 논리를 다루는 학문이니까요!)
QQplot이 기울기1, 절편0 직선에 잘 fitting되면 두 데이터의 분포는 같은 것.

Python에서

```
import scipy.stats as stats
stats.probplot(col, plot=plt)
```

R에서

```
library(lattice)
qqmath(x,distribution=qnorm,…)
```

QQ-plot은 probability plot의 일종

EDA Tip

1. 분석방법은 발표 때 굳이 설명하지 않으셔도 됩니다 .

2. 다변량 분석을 지향해주세요!

3. 결과물을 수치 그대로 보여주지 말고 꼭 시각화 해주세요!

4. 데이터를 많이 보고 어떤 데이터를 덧붙일지, 잠재변수에는 무엇이 있을지 충분히 고민해주세요 !

    4-(1). 모든 변수를 어떤 방식으로든 시각화 해보려고 노력하셔야합니다

5. 설명이 필요없는 플롯을 지향합시다! (labeling, legend, 의미있고 유니크한 변수명)

6.데이터 전처리가 90%이상입니다. 데이터 전처리가 분석하는 것보다 힘든 것 같습니다.
 기초세션 때 배우셨던 전처리 방법, 데이터 수집 방법 등을 잘 활용하시길 바랍니다!

7.추가적인 데이터를 찾아보는 걸 적극 추천드립니다.

THANK YOU