

# | 데이터 마이닝 및 시각화



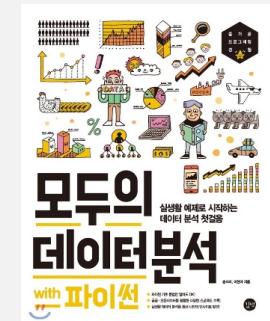
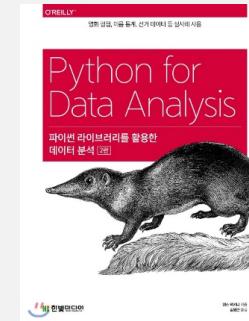
# 강의계획

## ■ 담당교수 및 강의

- 담당교수:정재희(jhjung@mju.ac.kr)
- 상담 : Y5521호 화 13:00-14:00 , 목 13:00-14:00, 사전약속

## ■ 강의교재 : LMS 강의자료 활용

- 강의노트
- 파이썬 데이터 사이언스 핸드북
- 파이썬 라이브러리를 활용한 데이터 분석
- 모두의 데이터 분석 with 파이썬



## 학습목표

---

■ 본 과목에서는 대용량 데이터 마이닝 능력을 함양시키기 위해 데이터의 상관관계를 분석하여 분석된 데이터를 바탕으로 효과적인 시각화 방법에 대해 학습한다. 데이터를 분석하기 위한 데이터의 전처리 과정과 데이터마이닝 모델링 기법을 학습하고, 분석된 데이터의 성질에 따른 효과적 시각화 및 응용 능력을 갖출 수 있도록 교과 내용을 구성한다.

# 평가 및 강의계획

## 강의 방식

- LMS, 강의업로드 및 zoom QnA 활용

## 평가

- 팀 프로젝트 (20%), 기말고사 (40%), 과제 (30%), 출석 및 태도 (10%)
- 평가의 비율은 학업 성취도에 따라 변경 될 수 있음

## 과제 (30%)

- 기한이 지난 과제물 제출 받지 않음
- **과제 복사 및 본인이 하지 않은 것으로 간주시 -300% 점 처리**  
**(예: 10 점 만점 과제 →-30점, 20점 만점과제 →-60점)**

## 시험이나 모든 과제는 추가 시험/과제 없음

- 예외 사항
- 적당한 이유가 있을 경우 시험 전 상담 필히 요청

# 평가 및 강의계획

## ■ 팀프로젝트 (20%)

- 캐글, 데이콘, 공공데이터 이용하여 분석 보고서 작성 및 발표

## ■ 출석 및 태도(10%)

- 결석횟수만큼 점수에 반영
- 수업 일수의 1/5 이상 결석 시 출석 미달로 'F'
- 모든 강의는 출석 필
- 차후 대면 시
  - 예비군 훈련, 징병검사 및 학교의 공식 행사를 제외한 출석 예외는 인정하지 않음 (예, 감기 및 장염 등으로 인한 병원 진단서는 출석에 반영하지 않음)
  - 무단 조퇴는 결석과 같음

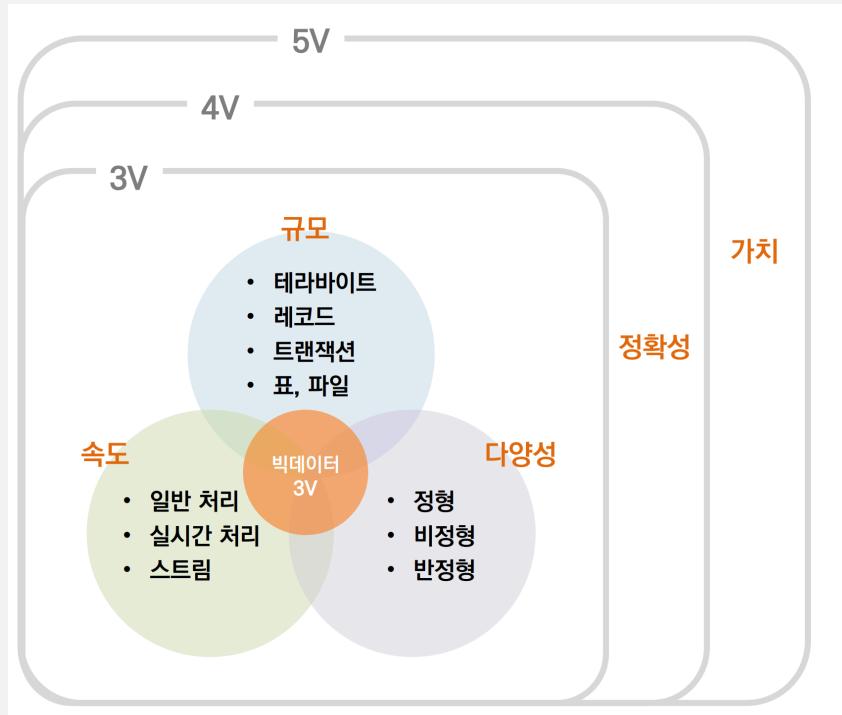
# 평가 및 강의계획

## ■ 평가

- 두 분반 합하여 평가
- 완화된 상대평가
  - 최대 A학점 30-50%,
  - 최대 B학점 20-30%
  - C~F 20~50%.
  - F 학점 가능
    - › 시험을 비롯한 숙제 부정행위 시
    - › 총 점수 100점 만점 환산 시, 30점 미만
    - › 사전 통보 없이 중간 또는 기말고사 불참
    - › 수업시간 1/5 이상 결석 (6번 이상 결석 시)

# 빅데이터

■ 빅데이터의 속성은 3V, 규모(Volume), 다양성(Variety), 속도(Velocity)로 정의하며, 최근 정확성(Veracity)과 가치(Value)를 포함하여 5V 정의

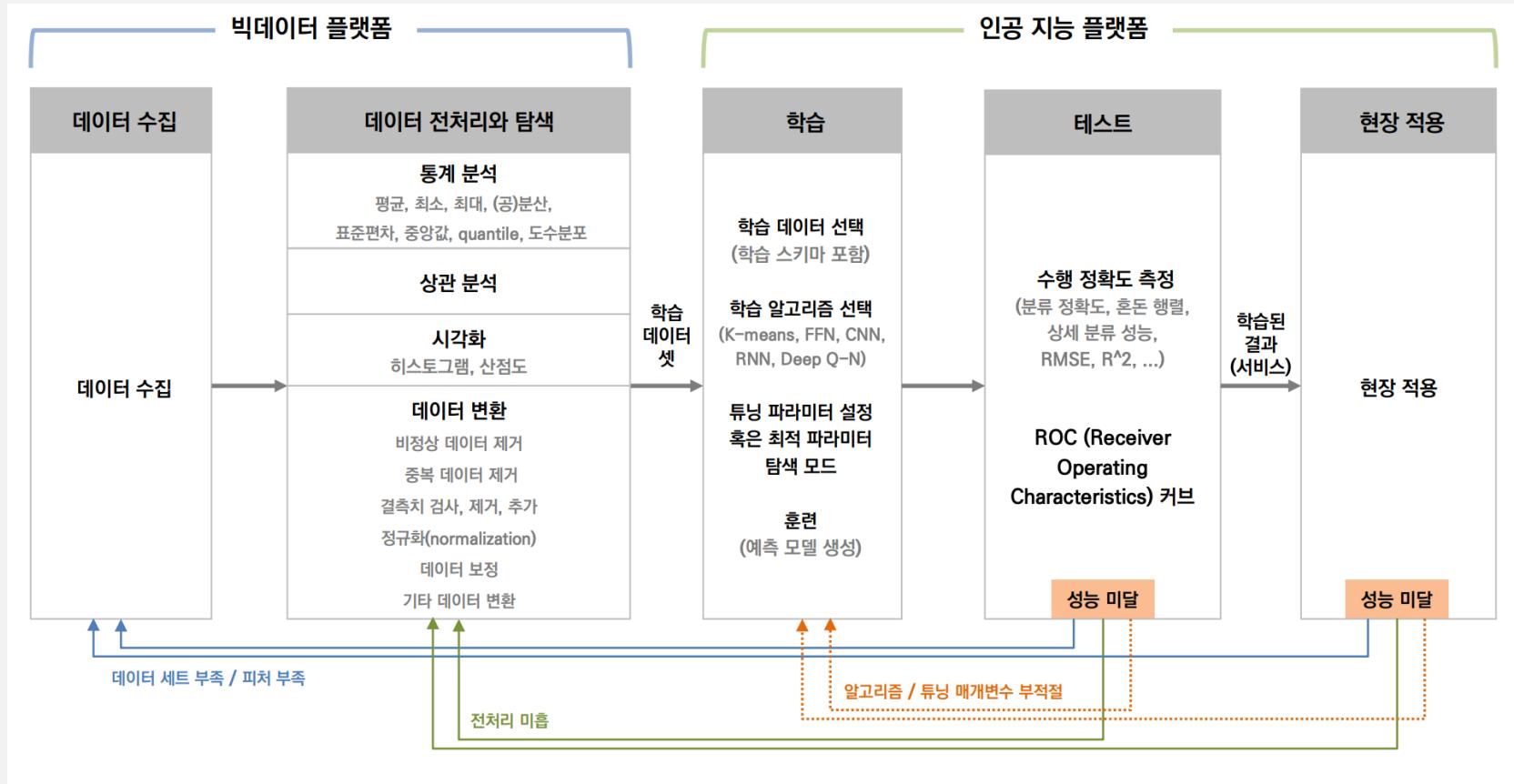


2012년 가트너는 기존 정의를 다음과 같이 개정했다. “빅데이터는 큰 용량, 빠른 속도, 다양성이 높은 정보 자산이다. 이것으로 의사 결정 및 통찰 발견, 프로세스 최적화를 향상시키려면 새로운 형태의 처리 방식이 필요하다.”

IBM은 여기에 정확성(Veracity) 요소를 더해 4V로 정의했고, 최근에는 가치(Value)를 포함하여 5V로 정의하기도 한다.

# 빅데이터 처리과정





# 데이터마이닝

## ■ 다양한 관점에서 데이터를 분석해 의미 도출

### ■ 연관(association)

- 주어진 데이터 세트에서 자주 발생하는 속성 값들을 연결해 주는 연관 규칙을 발견하는 일이다.
  - 예를 들면 고객이 구매한 쇼핑 카트 내의 개별 상품간의 상관관계를 식별하는 경우

### ■ 회귀(regression)

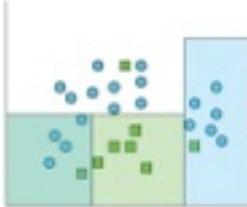
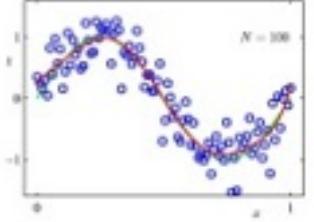
- 독립 변수 분석을 통해 종속 변수가 무엇인지 밝혀내는 일에 사용된다.
  - 예를 들면 어떤 상품의 예상판매실적을 주요 고객들의 소득 수준과 상품의 판매가격과의 상관관계로부터 예측

### ■ 분류(classification)

- 개체들을 여러 등급으로 나누는 모델

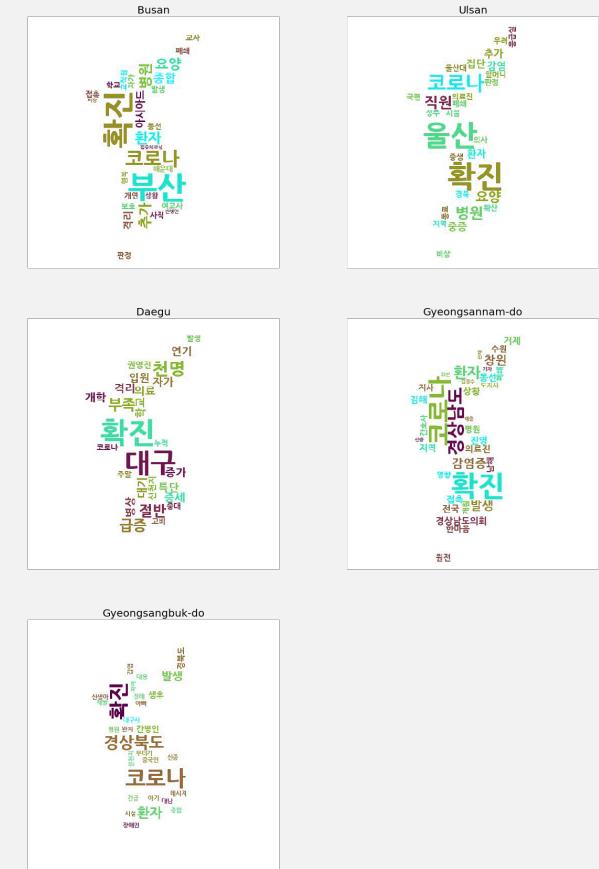
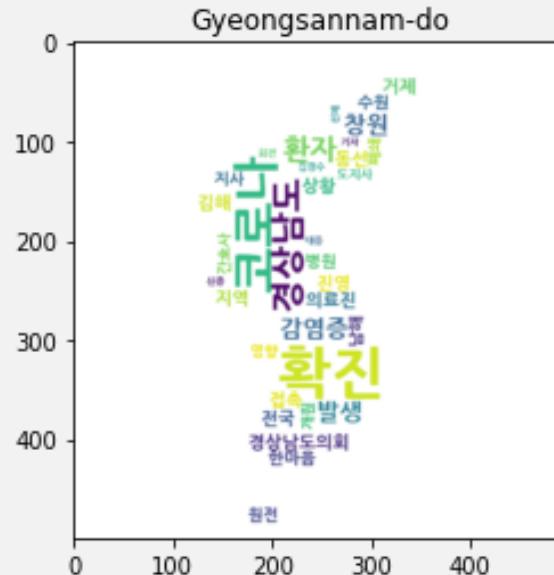
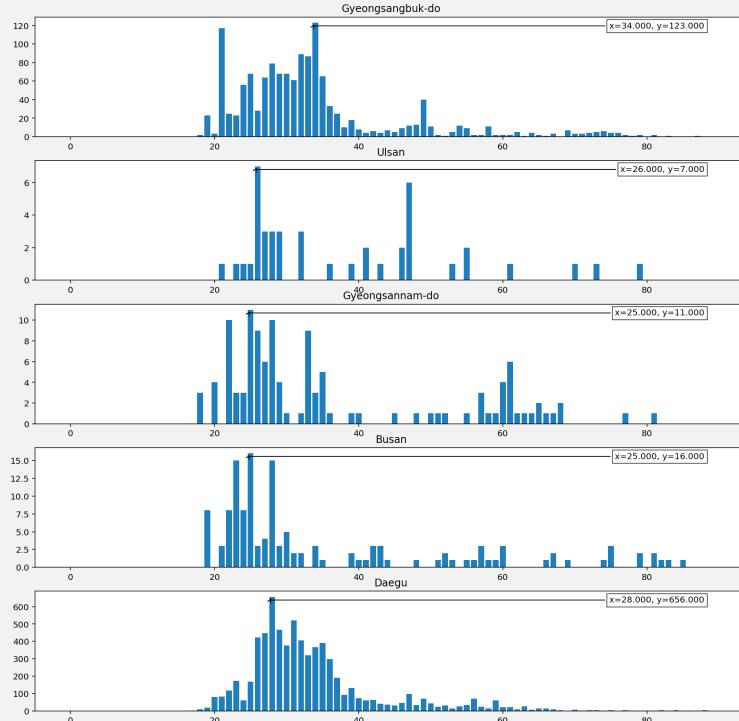
# 데이터마이닝

## Data mining methods

Predictive methods	Descriptive methods
<b>Classification</b>  <p>Learns a method for predicting the instance class from pre-labeled (classified) instances</p>	<b>Clustering</b>  <p>Finds "natural" grouping of instances given un-labeled data</p>
<b>Regression</b>  <p>An attempt to predict a continuous attribute</p>	<b>Association Rules</b>  <p>Method for discovering interesting relations between variables in large DBs</p>

# 데이터 분석 및 시각화 예- 코로나 데이터 시각화 AI 경진대회

## 코로나 데이터 시각화 AI 경진대회



<https://dacon.io/competitions/official/235590/data/>

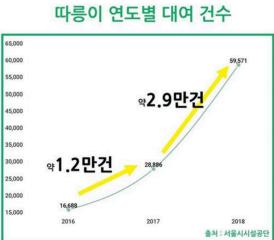
# | 데이터 마이닝 및 시각화

1주차 - 02



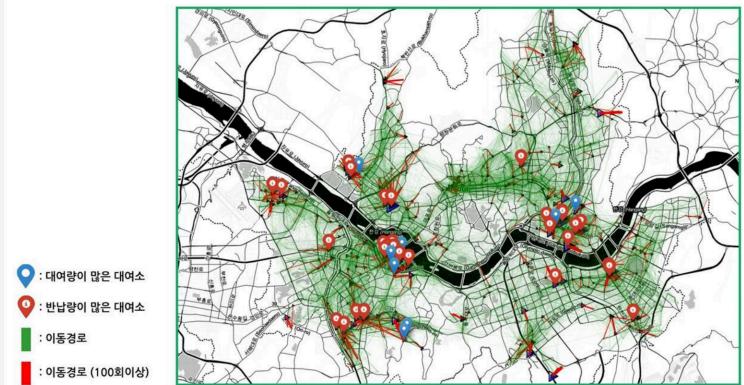
# 데이터 분석 및 시각화 예- 서울시 공공자전거 효율적인 운영을 위한 빅데이터 분석

## I. 분석 배경



따릉이의 연도별 대여건수가 늘어남에 따라  
연도별 고장건수도 급속하게 늘어나는 추세

## II. 재배치 분석 - 현황 및 방향성 설정

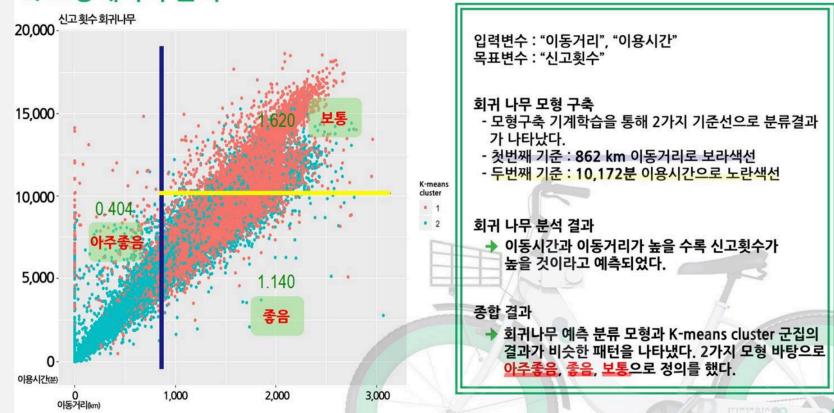


## I. 분석 방향

- ✓ 이용 대수가 특정 시간 및 특정 장소에 집중되는 현상으로 인한 관리 운영의 어려움
  - ▶ 계절별 요일별 시간대별로 데이터를 분석하여 집중되는 패턴을 파악
  - ▶ 분석데이터를 기반으로 과수요 군집을 분류하여 재배치 시간과 방법을 제시
- ✓ 매년 이용객 증가로 인해 서울시 공공자전거 고장을 증가하는 추세
  - ▶ 대여 이력 데이터를 분석하여 고장 자전거 이용패턴 파악
  - ▶ 고장 신고 및 정비 내역을 분석하여 자전거 상태를 진단, 군집화 적용



## II. 고장데이터 분석

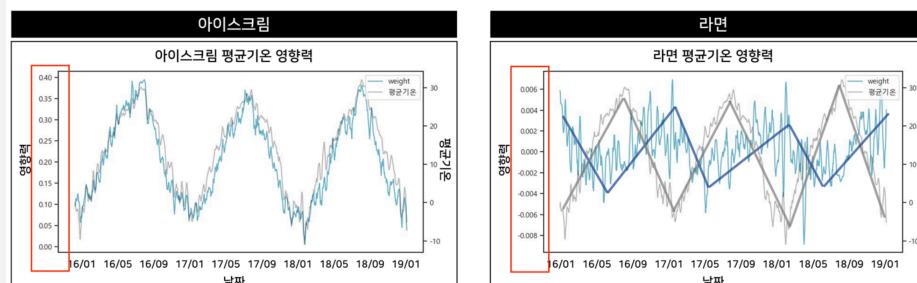
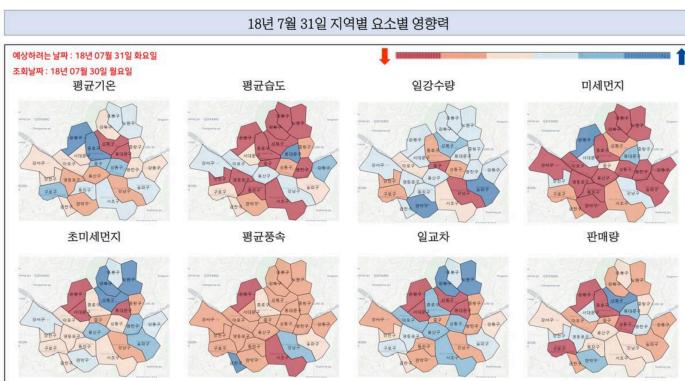


출처 [https://bigdata.seoul.go.kr/noti/selectNoti.do?r\\_id=P260&bbs\\_seq=334&ac\\_type=A1&sch\\_type=&sch\\_text=&currentPage=1](https://bigdata.seoul.go.kr/noti/selectNoti.do?r_id=P260&bbs_seq=334&ac_type=A1&sch_type=&sch_text=&currentPage=1)

# 데이터 분석 및 시각화 예- 날씨가 편의점 소비에 미치는 영향 분석 및 분석 모델 개발



편의점 소비에 대한 기상요소들의 영향력이 지역(구)별로 다르다.



- ✓ 아이스크림에 대한 평균기온의 기여 수치는 라면에 비해 더 큼. 기온은 라면보다 아이스크림 판매량에 더 큰 영향을 미치는 것을 의미하며, 라면보다 아이스크림 판매량을 예측하는데 더 중요한 변수임을 알 수 있다.
- ✓ 아이스크림은 평균기온 관측치(회색선)와 영향력(파란선)의 추이가 유사하다. 기온이 높아질수록 판매량에 대한 기온의 영향력이 더 커짐을 알 수 있다.
- ✓ 라면은 평균기온 관측치(회색선)와 영향력(파란선)의 추이가 반대이다. 아이스크림과는 반대로 기온은 낮아질수록 영향력이 커진다는 의미이다.

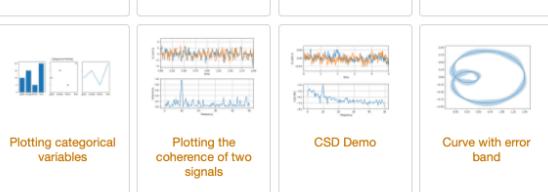
31

출처 <https://bd.kma.go.kr/contest/downloadFile.do?fileCd=FILE022>

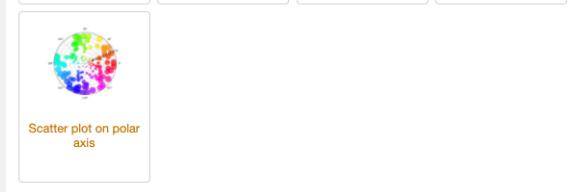
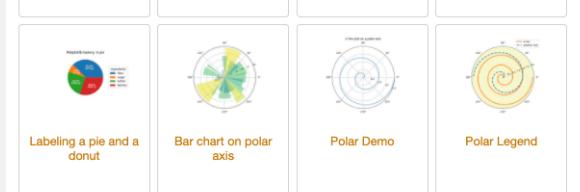
# 시각화 표현

## matplotlib

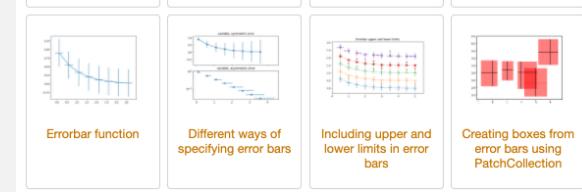
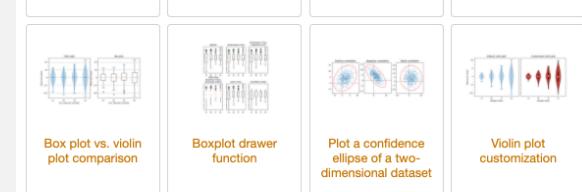
Lines, bars and markers



Pie and polar charts



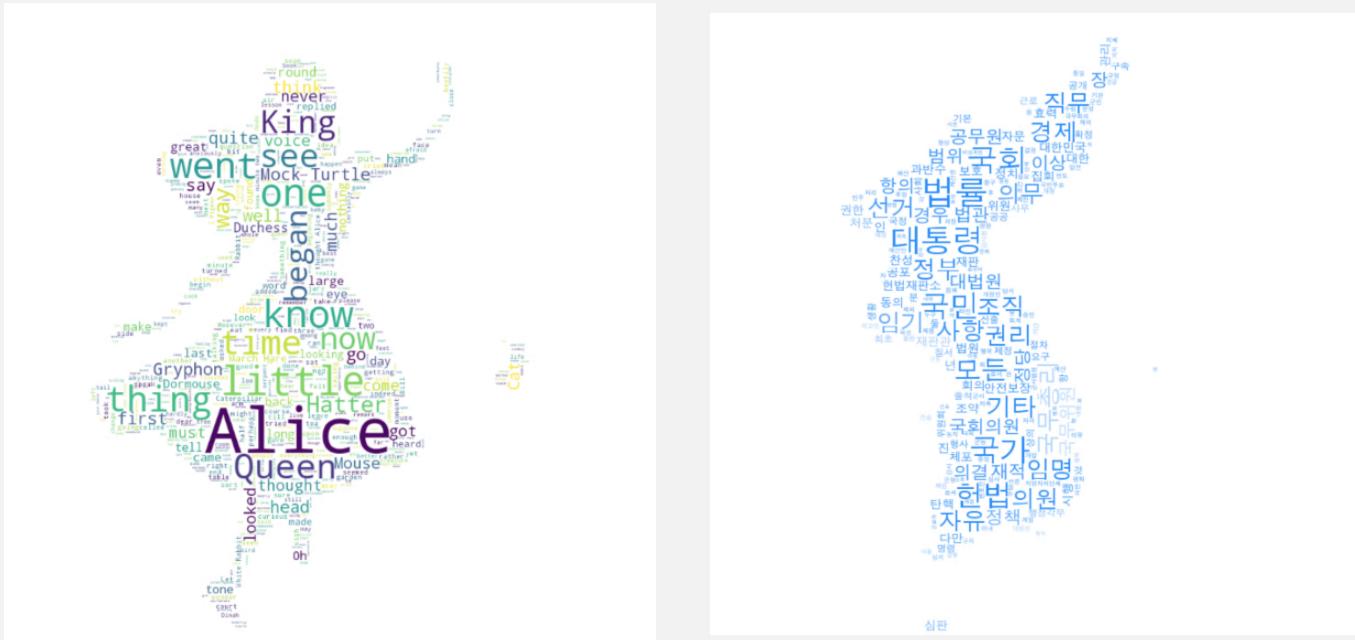
Statistics



<https://matplotlib.org/gallery/index.html>

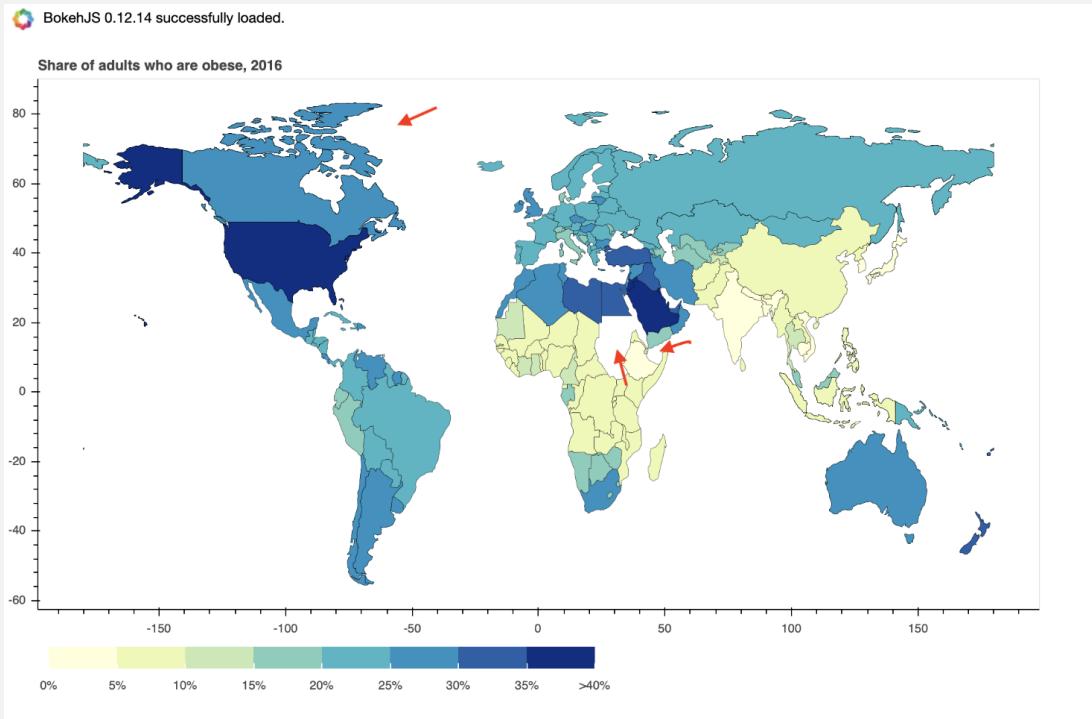
# 시각화 표현

## 워드 클라우드 wordcloud



# 시각화 표현

## 지도 시각화 bokeh, geopandas



# 시각화 표현



<https://datavizproject.com>

## Tableau

The screenshot shows a Tableau dashboard titled "Campaigns and Retention" owned by Emily Chen. The dashboard includes several visualizations:

- Facebook Content Performance: A bubble chart showing posts by type and date.
- Email Performance Overview: A bar chart showing email performance metrics.
- Google Analytics: A map showing website traffic trends.
- Historic Trends: A line chart showing historical performance trends.
- Performance by Week: A stacked bar chart showing weekly performance.
- Renewals by Region: A choropleth map showing renewals by region.

The left sidebar shows navigation options like Home, Explore, Favorites, Recents, Users, Groups, Schedules, Tasks, Site Status, and Settings.

<https://www.tableau.com/ko-kr/products/desktop>

## 참고 사이트

- 지도 : <https://geopandas.org/>
- 보케 : <https://docs.bokeh.org/en/latest/index.html>
- 멧플롯라이브러리 : <https://matplotlib.org/>
- <https://datavizproject.com>
- 타블로 퍼블릭 : <https://www.tableau.com/ko-kr/products/desktop>
- Scipy : <https://docs.scipy.org/doc/scipy/reference/>
- Sckit-learn <https://scikit-learn.org/stable/>

# 데이터 수집

## 2019년 문화, 관광 빅데이터 분석

- <http://www.tourbigdata.kr/award.asp>

## 공공데이터

- <https://www.data.go.kr/>

## 데이콘

- <https://dacon.io/>

## 캐글

- <https://www.kaggle.com/competitions>

## 기상청 빅데이터

- <https://bd.kma.go.kr/contest/>

## 교통데이터

- <http://data.ex.co.kr/bbs/view>

## 빅콘테스트

- <https://www.bigcontest.or.kr/points/content.php>

## L-Point 빅데이터 컴피티션

- <https://competition.lpoint.com/front/Guideline.tran>

# 공공데이터 활용 분석 사례 및 창업경진대회

## ■ 서울시 빅데이터

- [https://bigdata.seoul.go.kr/noti/selectPageListTabNoti.do?r\\_id=P260](https://bigdata.seoul.go.kr/noti/selectPageListTabNoti.do?r_id=P260)

## ■ 범정부 공공 데이터 활용 창업 경진대회

- <http://www.startupidea.kr/preliminary/>

# 강의 계획

## 교과 과정

주	내용
1주	Introduction to DataMining and visualization and Install Python
2주	Python Basic builtin - functions
3주	Numpy basics : data types, array creation, indexing, broadcasting,structured arrays
4주	Pandas series : Shaping, Time-series-related
5주	Pandas Dataframe : Reindexing, Reshaping, Plotting
6주	Pandas Array : Datetime, Sparse data, Text datag
7주	Scipy : interpolation, Linear algebra
8주	Scipy : Statistics, Spatial data structures and algorithms
9주	Seaborn: : Lines, Bar, Maskers, subplot, figures, axes
10주	Matplotlib : Image, Contour, Vector Field
11주	Matplotlib : Text, Labels, Annotation, 3DPlotting
12주	Bokeh : Interactive visualization
13주	Bokeh : Categorical Data and Network graphs
14주	geoPandas Mapping Geo Data
15주	WordCloud : crawling
16주	프로젝트 발표

# Types of Data

■ List of top, most popular best Python libraries for developers in 2020



NumPy



SciPy



Natural Language Analysis  
with Python NLTK



TensorFlow



PyTorch



pandas



Scrapy

Seaborn



SM



bokeh



plotly