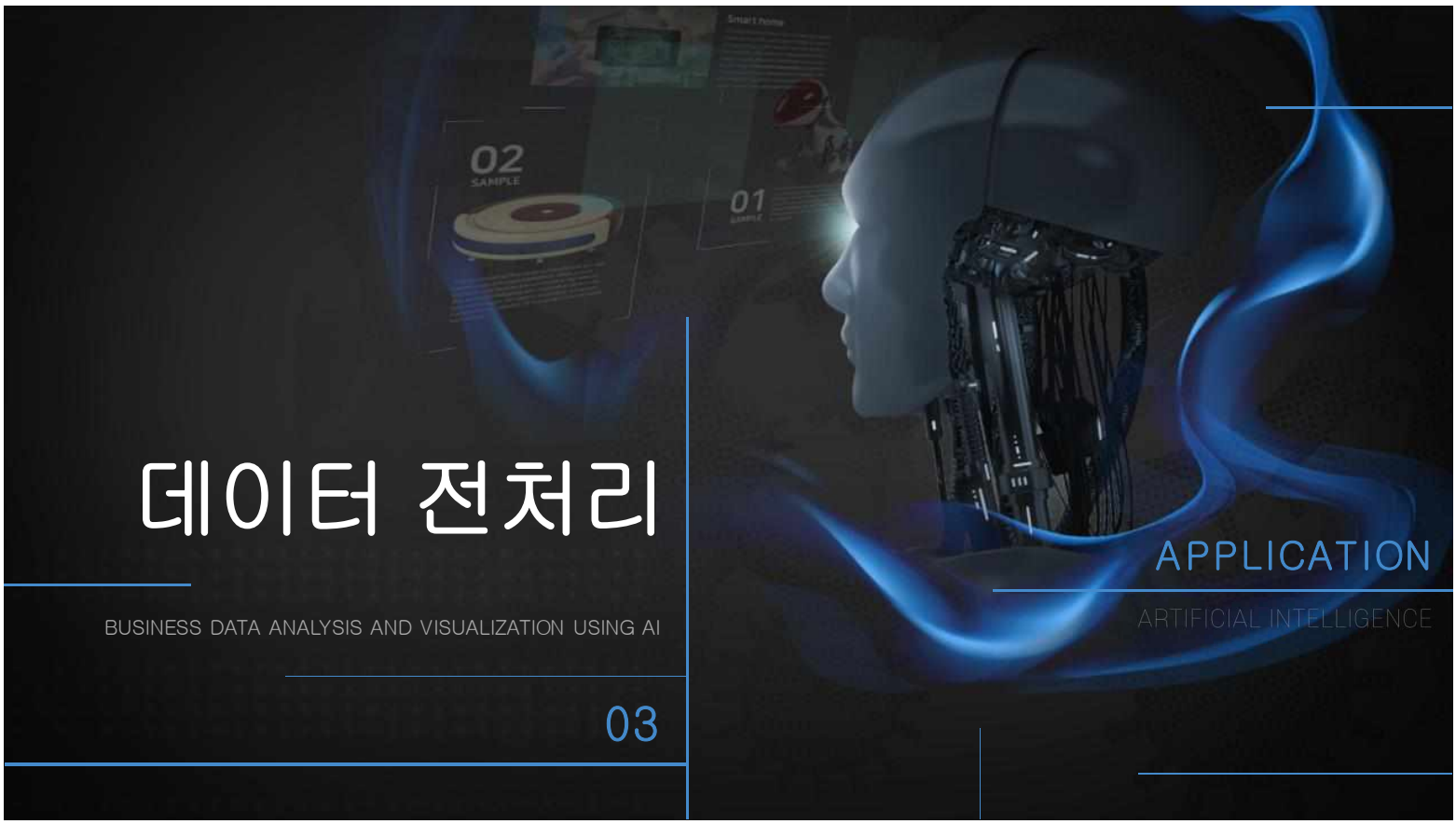


데이터 전처리

BUSINESS DATA ANALYSIS AND VISUALIZATION USING AI

03

APPLICATION

ARTIFICIAL INTELLIGENCE

Notice

학습목표

GOALS

1 데이터 전처리 개념과 과정을 이해할 수 있다.

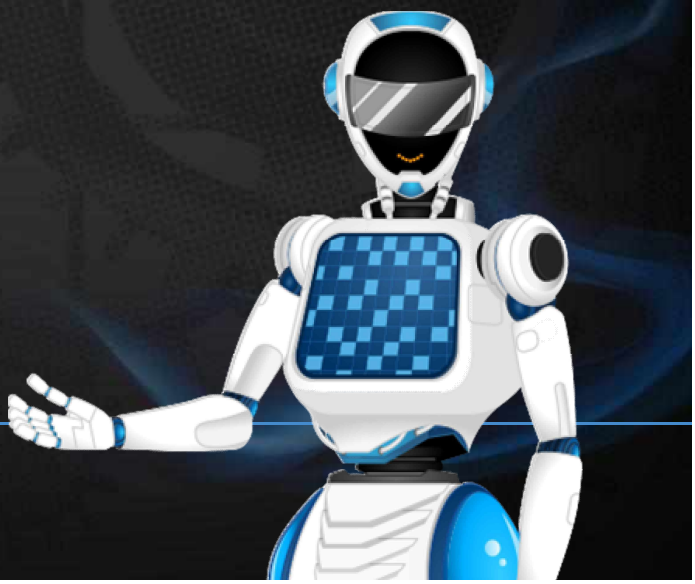
2 탐색적 데이터 분석을 적용할 수 있다.

1 데이터 전처리 이해

2 탐색적 데이터 분석

3 데이터 전처리 실습

CONTENTS



학습내용

APPLICATION

01 데이터 전처리 이해

01 데이터 전처리 기술

빅데이터 전처리 (Pre-processing)

- 빅데이터 분석 목적에 적합한 데이터 목록 작성, 확보 가능여부 점검
- 수집한 데이터를 저장소에 적재하기 위한 작업

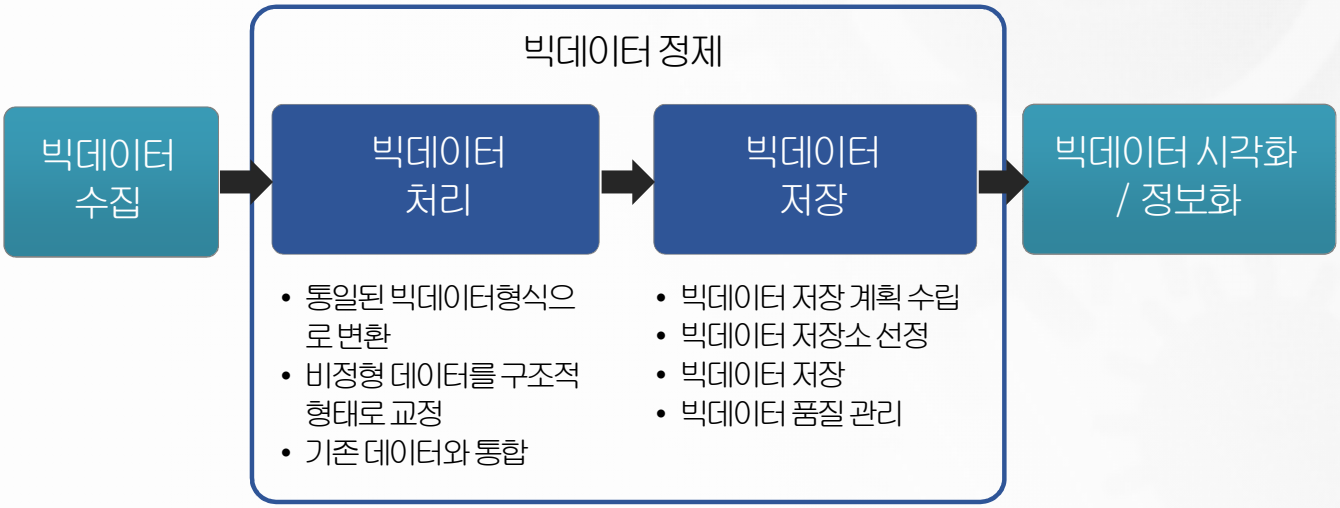
구분	수행 내용
데이터 유형 변환	데이터 유형을 변환하거나 데이터 분석에 용이한 형태로 변환
데이터 여과 (Filtering)	오류발견, 보정, 삭제 및 중복성 확인 등 데이터 품질 향상
데이터 정제 (Cleansing)	결측치 변환, 이상치 제거, 노이즈 데이터 교정 비정형 데이터를 수집할 때 반드시 수행

그 외 데이터 통합/축소/세분화 등

01 데이터 전처리 기술

[빅데이터 전처리 과정]

실무자를 위한 빅데이터 업무절차 및 기술활용 매뉴얼 1.0



01 데이터 전처리 이해

01 데이터 전처리 기술

데이터 변환

다양한 형식으로 수집 된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환

평활화 Smoothing	데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 값들을 변환
집계 Aggregation	다양한 차원의 방법으로 데이터를 요약
일반화 Generalization	특정 구간에 분포하는 값으로 스케일을 변화

01 데이터 전처리 기술

데이터 변환

다양한 형식으로 수집 된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환

정규화
Normalization

데이터에 대한 최소-최대 정규화, Z-Score 정규화, 소수 스케일링 등 통계적 기법 적용

속성 생성

Attribute/feature construction
데이터 통합을 위해 새로운 속성이나 특징을 만드는 방법으로 주어진 여러 데이터 분포를 대표할 수 있는 새로운 속성/특징 활용

01 데이터 전처리 이해

01 데이터 전처리 기술

데이터 필터링 (Filterling)

1. 데이터 중복성, 오류 제거들을 위한 데이터 필터링 기준 설정
2. 실제 사전 테스트를 통하여 오류 발견, 보정, 삭제 및 중복성 검사 등 필터링 과정을 거쳐 필터링 기준을 최적화 하여 활용
3. 비정형 데이터는 데이터 마이닝을 통해 오류, 중복, 저품질 데이터를 처리할 수 있도록 자연어처리 및 기계학습과 같은 추가기술 적용 필요
4. 분석을 위하여 단위 저장소에 파일형태로 저장 할 경우, 데이터 활용 목적에 맞지 않는 정보는 필터링하여 제거해야 분석시간을 단축하고 저장 공간의 효율적 활용 가능

01 데이터 전처리 기술

데이터 정제

1. 수집된 데이터의 불일치성을 교정하기 위한 방식
2. 결측치(Missing Value) 처리, 이상치/잡음(Noise) 처리 기술 활용

01 데이터 전처리 이해

01 데이터 전처리 기술

데이터 정제

[결측값 처리 방법]

방법	설명
해당 레코드 무시	결측치가 적을 경우 효율적 분류에서 클래스 구분 라벨이 빠진 경우
자동으로 채우기	통계값 적용 : 평균값, 중앙값, 같은 클래스 값의 평균값 등 추정치 적용 : 베이지안 확률, 결정트리 모든 값을 '0'으로 적용
전문가 수작업	담당자가 직접 확인하고 적절한 값으로 수정

02 데이터 결측값 처리

삭제

1. 결측 값이 발생한 모든 관측치를 삭제
2. 간편한 반면 데이터의 수가 줄어들어 모델의 유효성이 낮아질 수 있음
3. 부분 삭제는 관리 Cost가 늘어난다
4. 삭제는 결측 값이 무작위로 발생한 경우에 사용

01 데이터 전처리 이해

02 데이터 결측값 처리

통계값 대체

1. 결측값을 다른 관측치의 평균, 최빈값, 중간값 등으로 대체

일괄 대체 방법	모든 결측값을 동일한 통계값으로 대체
유사 유형 대체 방법	범주형 변수를 활용해 유사한 유형의 평균값 등으로 대체

2. 범주형 변수를 유사한 유형으로 선택할 것인지는 자의적으로 선택할 수 있으므로 모델의 왜곡이 가능

02 데이터 결측값 처리

예측값(회귀분석) 삽입

1. 결측값이 없는 관측치를 활용하여 결측값을 예측하는 모델을 만들고 이 모델을 통해 결측값이 있는 관측 데이터의 결측값을 예측
2. Linear Regression, Logistic regression 사용

결측값이 발생한 변수가 많을 경우,
사용 가능한 변수가 적어 적합한 모델을 만들기 어렵고,
많은 모델을 만들어야 하는 단점

01 데이터 전처리 이해

02 데이터 결측값 처리

데이터 정제

[잡음 / 아웃라이어 처리 방법]

방법	설명
구간화	정렬된 데이터를 여러 개의 구간으로 배분 후 구간 대표값으로 대체
회귀값 적용	데이터를 가장 잘 표현하는 추세 함수를 찾아서 적용
군집화	비슷한 성격을 가진 클러스터 단위로 묶은 다음 처리

03 데이터 이상값 처리

이상값

1. 데이터/샘플과 동떨어진 관측치, 모델을 왜곡할 가능성이 있는 관측치

이상값 찾아내기

1. 일반적으로 하나의 변수에 대해서는 Boxplot이나 Histogram을 두개의 변수 간 이상값을 찾기 위해서는 Scatter plot을 사용
2. 두 변수 간 이상값을 찾기 위한 또 다른 방법으로는 두 변수 간 회귀 모형에서 Residual, Studentized residual(혹은 standardized residual), leverage, Cook's D값을 확인

01 데이터 전처리 이해

03 데이터 이상값 처리

이상값 처리

1. 단순 삭제

- 이상값이 Human error에 의해서 발생한 경우에는 해당 관측치를 삭제
- 단순 오타나, 주관식 설문 등의 비현실 적인 응답, 데이터 처리 과정에서의 오류 등의 경우

03 데이터 이상값 처리

이상값 처리

② 통계값 대체

- 삭제의 방법으로 이상치를 제거하면 관측치의 절대량이 작아지는 문제
- 다른 값(평균, 중앙값 등)으로 대체
- 결측값과 유사하게 다른 변수들을 사용해서 예측 모델을 만들고, 이상값을 예측한 후 해당 값으로 대체하는 방법

01 데이터 전처리 이해

03 데이터 이상값 처리

이상값 처리

③ 변수화

- 자연발생적인 이상값의 경우, 바로 삭제하지 말고 좀 더 자세히 이상값에 대해 파악하는 것이 중요

예시 위 이상 값의 경우 의사 등 전문직종에 종사하는 사람이라고 가정할 때, 이럴 경우 전문직종 종사 여부를 로 변수화 하면 이상값을 삭제하지 않고 모델에 포함시킬 수 있음.

03 데이터 이상값 처리

이상값 처리

4 리샘플링

- 해당 이상값을 분리해서 모델을 만드는 방법
- 케이스별로 여러가지 모델을 만드는 방법

04 데이터 전처리 기술

데이터 통합

1. 출처가 다른 상호 연관성이 있는 데이터들을 하나로 결합하는 기술
2. 데이터 통합 시 동일한 데이터가 입력 될 수 있으므로 연관관계 분석 등을 통해 중복 데이터를 검출 할 것
3. 데이터 통합 전·후 수치·통계 등 데이터 값들이 일치 할 수 있도록 검증
4. 통합 대상 entity가 통합 이후에 동일한지 여부를 확인하기 위한 동일성 검사
5. 표현 단위(파운드와 kg, inch와 cm, 시간 등) 등 서로 다른 방식에 대해 표현을 일치할 수 있도록 변환

04 데이터 전처리 기술

데이터 축소

분석에 불필요한 데이터를 축소하여 고유한 특성은 손상되지 않도록 하고 분석에 대한 효율성을 증대

방법	설명
데이터 압축	데이터 인코딩이나 변환을 통해 데이터 포맷 변경
DWT (Discrete wavelet transform)	선형 신호 처리 여러 개의 벡터 중에서 가장 영향력이 큰 벡터를 선택하여 다른 벡터들을 제거
차원 축소 (PCA)	데이터를 가장 잘 표현하고 있는 직교 상의 데이터 벡터를 찾아서 압축
차원 축	데이터를 더 작은 형태로 표현해서 데이터 저장

05 데이터 세분화란?

데이터 세분화

보유한 데이터를 특정 요소를 기반으로 비슷한 데이터를 그룹화하여 나눔

예시 성별, 나이, 산업군

05 데이터 세분화란?

데이터 세분화

데이터 세분화 중요성

① 목표한 타겟에 집중해서 데이터를 형성할 수 있음

② 데이터 베이스에 있는 데이터 분석에 용이하도록 도움

③ 비용 감소

MEMO

APPLICATION

02 탐색적 데이터 분석 (EDA: Exploratory Data Analysis)

01 탐색적 데이터 분석

탐색적 데이터 분석의 목적

- 데이터를 이해하는 것

목적 달성을 위한 가장 쉬운 방법

- 적절한 질문으로 데이터를 표현하는 적합한 모형, 시각화 산출물, 다음 과정을 위한 데이터를 생성

02 탐색적 데이터 기법사용 절차 수립

(1) 개념적 데이터 탐색과정

- 데이터에 포함된 변수에 내재된 변동성(variation) 유형은 어떻게 되나?
- 변수들 간에 공변동(covariation)은 어떻게 되나?

02 탐색적 데이터 분석

02 탐색적 데이터 기법사용 절차 수립

(1) 개념적 데이터 탐색과정

구성요소

- 변수(Variable): 측정할 수 있는 질적 속성, 양적 수량, 속성이 된다.
- 값(Value): 값은 측정할 때 변수의 상태로 정의되어, 변수값은 측정시점 마다 변화하는 특징이 있다.
- 관측점(Observation): 동일한 시점 동일한 객체에 대해 측정된 관측정보 집합. 관측점에는 서로 다른 연관된 변수에 값들이 담겨있다.
- 표로 표현된 데이터(Tabular Data): 연관된 변수와 관측점을 값으로 표현한 집합. 값 각각이 본연의 위치(셀, cell) 에 담겨있고, 칼럼에 변수가 위치하고, 행에 각 관측점이 위치할 때 깔끔한 데이터(tidy data) 라고 정의한다.

02 탐색적 데이터 기법사용 절차 수립

(2) [가정]다중회귀분석

$$Y \leftarrow X1, X2, X3 \cdots\cdots, Xp$$

- 여기서 독립변수는 p 개로 다중회귀분석의 가정인 다중공산성이 없으며 설명력이 높아야 한다. 그러나 흔치 않다.
- 가령 변수 p=7 개를 모두 사용해서 분석을 한 결과 설명력이 Rp2=89% 라고 한다면 일단 그 자체는 좋은 일이다.
- 그런데 만약 p=2 개의 독립변수만으로도 설명력이 Rq2=87% 정도로 나와 준다면 그냥 두 개만 쓰는 게 훨씬 나을 수도 있다.

02 탐색적 데이터 분석

02 탐색적 데이터 기법사용 절차 수립

(2) [가정]다중회귀분석

$$Y \leftarrow X1, X2, X3 \cdots\cdots, Xp$$

- 이 때 원래대로 모든 독립변수 p 개로 얻은 모형을 전체모형 FM (Full Model) 이라 하고 독립변수를 q개만큼 줄여서 얻은 모형을 축소모형 RM (Reduced Model) 이라 한다.
- 독립변수가 하나도 없는 모형을 영모형 NM (Null Model) 이라고 하며, 영모형과 전체모형 사이에 있는 축소모형들 중 가장 적절한 것을 찾아내는 것이 문제다.
- 다행스럽게도, 이 과정을 이해하는 것 자체는 수학적인 지식이 없어도 상관 없다. 쉽게 말하면 그냥 '가능한 경우를 모두 검토' 해보는 것이다.

03 변수 선택 기법 적용 절차

변수 선택 기법 적용 절차

- **전진선택법** Forward Selection Procedure
- **후진 제거법** Backward Elimination Procedure
- **단계적 선택법** Stepwise Method

02 탐색적 데이터 분석

03 변수 선택 기법 적용 절차

변수 선택 기법 적용 절차

전진선택법 Forward Selection Procedure

- 영모형 Y←1 에서 하나씩 변수를 추가해가면서 모형을 선택 한다.
- 가장 작은 t검정 통계량이 1 보다 작으면 변수 추가를 멈추는 식으로 사용하면 간단하면서도 꽤 효과적이다.

03 변수 선택 기법 적용 절차

변수 선택 기법 적용 절차

후진 제거법
Backward Elimination Procedure

- 전체모형 Y←X1,…,Xp 에서 하나씩 변수를 제거해가면서 모형을 선택한다.
- 더 이상 제거할 변수가 없을 때의 모형을 선택

02 탐색적 데이터 분석

03 변수 선택 기법 적용 절차

변수 선택 기법 적용 절차

단계적 선택법
Stepwise Method

- 모든 부분집합을 고려하는 방법으로 최적의 변수를 선택할 수 있다.
- 전진선택법에 따라 변수 추가한 후 기존 변수의 중요도가 낮아질 시 해당 변수를 제거하는 등 추가 또는 제거되는 변수 여부를 판단, 더 이상 없을 때에 중단한다

03 변수 선택 기법 적용 절차

step 1.
데이터 핸들링

- 직접 데이터를 보면서 결측치, 이상치와 변환이 필요한 부분을 찾아낸다.

step 2.
변수 간의 관계 파악

- 모든 변수를 사용해 얻은 전체모형의 모형진단, VIF 를 보고 교호작용이나 다중공선성을 찾아낸다.

02 탐색적 데이터 분석

03 변수 선택 기법 적용 절차

Step 3.
전체모형 결정

- 다중공선성 문제가 해결된 회귀모형에 대해 모형진단을 하고 새로운 전체모형으로 결정한다.

step 4.
변수 선택

- 단계적 선택법을 적용하고 변수 선택 기준에 따라 가장 좋은 축소모형을 선택한다.

03 변수 선택 기법 적용 절차

Step 5. 최적모형 결정

- 선택된 축소모형이 모형진단을 문제 없이 통과하면 이를 최적모형으로 결정한다.

04 표본조사

표본조사

- 연구 문제의 해결을 위하여 필요한 자료를 수집하기 전에 측정도구를 제공할 대상을 선정해야 한다. 즉 전수조사를 할 것인가 혹은 그 일부만을 대상으로 할 것인지를 정해야 한다.
- 통계학에서는 전체를 모집단 일부를 표본이라고 한다.
- 모집단에서 표본을 선택하는 과정을 표본 추출이라 하며 수집된 표본에 대해서 자료를 수집하는 것을 표본조사라고 한다.

04 표본조사

상관성 분석 correlation analysis

- 확률론과 통계학에서 두 변수 간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법
- 두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며 이 때 두 변수 간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)라 한다. 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수 ρ를 사용한다.

02 탐색적 데이터 분석

04 표본조사

상관계수 Correlation coefficient

- 두 변수 간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다.
- 두 변수 간에 원인과 결과의 인과관계가 있는지에 대한 것은 회귀분석을 통해 인과관계의 방향, 정도와 수학적 모델을 확인해볼 수 있다.

06 변수의 상관성 분석방법

(1) 상관 분석의 기본 가정

선형성 | 동변량성 | 두 변인의 정규분포성 | 무선독립표본

선형성

- 두 변인 X와 Y의 관계가 직선적인지를 알아보는 것으로 이 가정은 분포를 나타내는 산점도를 통하여 확인할 수 있다.

02 탐색적 데이터 분석

06 변수의 상관성 분석방법

(1) 상관 분석의 기본 가정

선형성 | 동변량성 | 두 변인의 정규분포성 | 무선독립표본

동변량성

- X의 값에 관계없이 Y의 흩어진 정도가 같은 것을 의미한다. 이분산성이 반대말이다.

06 변수의 상관성 분석방법

(1) 상관 분석의 기본 가정

선형성 | 동변량성 | **두 변인의 정규분포성** | 무선독립표본

두 변인의 정규분포성

- 두 변인의 측정치 분포가 모집단에서 모두 정규분포를 이루는 것이다.

02 탐색적 데이터 분석

06 변수의 상관성 분석방법

(1) 상관 분석의 기본 가정

선형성 | 동변량성 | 두 변인의 정규분포성 | **무선독립표본**

무선독립표본

- 모집단에서 표본을 뽑을 때 표본대상이 확률적으로 선정된다는 것이다.

06 변수의 상관성 분석방법

(2) 상관 분석 방법

1. 단순상관분석

- 단순히 두 개의 변수가 어느 정도 강한 관계에 있는가를 측정

2. 다중상관분석

- 3개 이상의 변수들간의 관계에 대한 강도를 측정

02 탐색적 데이터 분석

06 변수의 상관성 분석방법

(2) 상관 분석 방법

3. 편상관계분석

- partial correlation analysis
- 다중상관분석에서 다른 변수들과의 관계를 고정하고 두 변수 만의 관계에 대한 강도를 나타내는 것
- 이 때 상관관계가 $0<\rho\leq+1$ 이면 양의 상관, $-1\leq\rho<0$ 이면 음의 상관, $\rho=0$이면 무상관이라고 한다.
  하지만 0인 경우 상관이 없다는 것이 아니라 선형의 상관관계가 아니라는 것이다.

06 변수의 상관성 분석방법

(3) 피어슨 상관계수

피어슨 상관계수

◆ 두 변수간의 관련성을 구하기 위해 보편적으로 이용

$$r = \frac{\text{X와 Y가 함께 변하는 정도}}{\text{X와 Y가 각각 변하는 정도}}$$

[결과의 해석]

r 값은 X 와 Y 가 완전히 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일하면 −1 을 가진다. 결정계수(coefficient of determination)는 r^2 로 계산하며 이것은 X 로부터 Y 를 예측할 수 있는 정도를 의미한다.

02 탐색적 데이터 분석

06 변수의 상관성 분석방법

(3) 피어슨 상관계수

r이 −1.0과 −0.7 사이	강한 음적 선형관계
r이 −0.7과 −0.3 사이	뚜렷한 음적 선형관계
r이 −0.3과 −0.1 사이	약한 음적 선형관계
r이 −0.1과 +0.1 사이	거의 무시될 수 있는 선형관계
r이 +0.1과 +0.3 사이	약한 양적 선형관계
r이 +0.3과 +0.7 사이	뚜렷한 양적 선형관계
r이 +0.7과 +1.0 사이	강한 양적 선형관계

06 변수의 상관성 분석방법

(4) 스피어만 상관 계수

스피어만 상관 계수

- 데이터가 서열척도인 경우 즉 자료의 값 대신 순위를 이용하는 경우의 상관계수로서, 데이터를 작은 것부터 차례로 순위를 매겨 서열 순서로 바꾼 뒤 순위를 이용해 상관계수를 구함.
- 두 변수 간의 연관 관계가 있는지 없는지를 밝혀주며 자료에 이상점이 있거나 표본크기가 작을 때 유용하다.
- -1과 1 사이의 값을 가지는데 두 변수 안의 순위가 완전히 일치하면 +1이고, 두 변수의 순위가 완전히 반대이면 -1이 된다.

예시 수학 잘하는 학생이 영어를 잘하는 것과 상관 있는지 없는지를 알아보는데 사용

07 기초 통계량 추출 및 분석

(2) 기초통계량 추출 및 이해

기초통계량 추출 및 이해

- 통계적 분석방법은 변수 타입에 따라 분석 방법이 결정되어 있다.

방법	숫자 요약	그래프 요약
측정형	중앙 척도, 평균 중위값	박스 플롯,히스토그램
범주형	빈도, 상대빈도	파이차트, 바 차트
시계열 데이터	증가율,차분	시간 도표

07 기초 통계량 추출 및 분석

(3) 평균

평균

- 평균은 통계학에서 두 가지 서로 연관된 뜻이 있다.
- 일상에서 평균이라고 부르는 것으로 산술 평균이라고도 한다.
  이 평균은 표본 평균이라고도 한다.

07 기초 통계량 추출 및 분석

(4) 중앙값

중앙값

- 중앙값 또는 중위수는 어떤 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값

예시 1, 2, 100의 세 값이 있을 때, 2가 가장 중앙에 있기 때문에 2가 중앙값

07 기초 통계량 추출 및 분석

(4) 중앙값

중앙값

- 값이 짝수 개일 때에는 중앙값이 유일하지 않고 두 개가 될 수도 있다.
  이 경우 그 두 값의 평균을 취한다.

  예시 1, 10, 90, 200 네 수의 중앙값은 10과 90의 평균인 50

- 중앙값(median)은 전체 데이터 중 가운데에 있는 수이다.
  예와 같이 극단적인 값이 있는 경우 중앙값이 평균값보다 유용하다.

  예시 연봉 평균은 5천만원인데 사장의 연봉이 100억인 경우, 회사 전체의 연봉 평균은 1억 4,851만원이 된다.

02 탐색적 데이터 분석

07 기초 통계량 추출 및 분석

(5) 최빈값

최빈값

- 최빈값 모드는 통계학 용어로, 가장 많이 관측되는 수, 즉 주어진 값 중에서 가장 자주 나오는 값

  예시 {1, 3, 6, 6, 6, 7, 7, 12, 12, 17}의 최빈값은 6이다.

- 최빈값은 산술 평균과 달리 유일한 값이 아닐 수도 있으며 주어진 자료나 관측치의 값이 모두 다른 경우에는 존재하지 않는다.
  주어진 자료에서 평균이나 중앙값을 구하기 어려운 경우에 특히 유용하다.

07 기초 통계량 추출 및 분석

(6) 사분위수

사분위수

- 데이터 표본을 4개의 동일한 부분으로 나눈 값
- 사분위수를 사용하여 데이터 집합의 범위와 중심 위치를 신속하게 알 수 있다.

사분위수	설명
제1 사분위수	데이터의 25%가 이 값보다 작거나 같음
제2 사분위수	데이터의 50%가 이 값보다 작거나 같음
제3 사분위수	데이터의 75%가 이 값보다 작거나 같음
사분위간 범위	제 1 사분위수와 제3 사분위수 간의 거리가 (Q3-Q1) 이므로 데이터의 중간 50%에 대한 범위이다

02 탐색적 데이터 분석

07 기초 통계량 추출 및 분석

(7) 분산

분산

- 확률론과 통계학에서 어떤 확률변수의 분산은 그 확률변수가 기대값으로부터 얼마나 떨어진 곳에 분포하는지를 가늠하는 숫자
- 기대값은 확률변수의 위치를 나타내고 분산은 그것이 얼마나 넓게 퍼져 있는지를 나타낸다. 분산보다는 분산의 제곱근인 표준편차가 더 자주 사용된다.

07 기초 통계량 추출 및 분석

(7) 분산

분산의 계산 방법

1. 분산(variance)은 관측값에서 평균을 뺀 값을 제곱하고, 그것을 모두 더한 후 전체 개수로 나눠서 구한다.
   즉, 차이값의 제곱의 평균이다

2. 관측값에서 평균을 뺀 값인 편차를 모두 더하면 0이 나오므로 제곱해서 더한다.

07 기초 통계량 추출 및 분석

(8) 변동계수

변동계수

- 측정단위에 따라 표준 편차의 값의 크기가 달라지므로 단위가 다른 두 집단을 비교하는 경우 두 표준 편차의 단위를 같게 할 필요가 있다.
- 이를 위하여 표준편차를 평균으로 나눈 값에 100을 곱하는 변동계수라고 하고 상대 변동 개념으로 정의 하고 있다.

07 기초 통계량 추출 및 분석

(9) 분산,표준편차의 해석

분산,표준편차의 해석

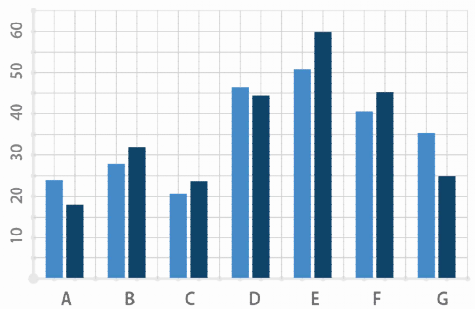
- 표준 편차의 자료의 흩어진 정도를 나타내는 수치이므로 이를 이용하여 측정 자료를 해석하는데 이용할 수 있다.

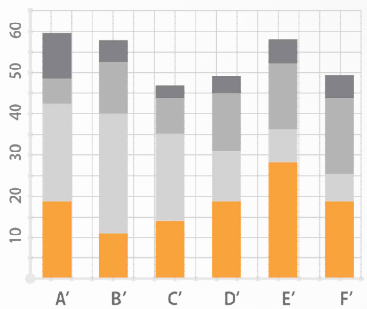
예시 수능 성적이 동일한 학생 100명을 대상으로 새로운 학습 방법을 적용한 후 100문제를 풀게 하여 평균이 80점이고 표준편차가 10이라고 가정하면, 성적의 분포가 종 모양이라면 100명 학생 중 학생 성적은 대부분 77-83점에 분포된다고 생각한다. 만약 표준편차가 10라고 얻어지면 학생들의 점수는 40-100점 사이에 있으므로 학습 능력은 다소 차이가 있다고 볼 수 있다.

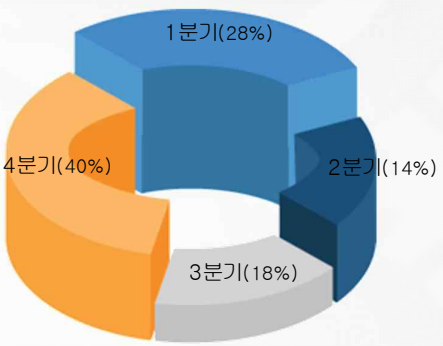
08 시각화를 통한 탐색적 자료분석

(1) 빈도 분석

범주형 변수의 범주의 빈도: 표로 나타내면 bar chart 또는 파이차트가 됨







08 시각화를 통한 탐색적 자료분석

(2) 측정형 변수

- 집단이 하나인 경우

측정형 데이터를 요약하는 값을 통계량이라 하며 요악 통계량은 중앙 위치와 산포를 가장 많이 사용한다.

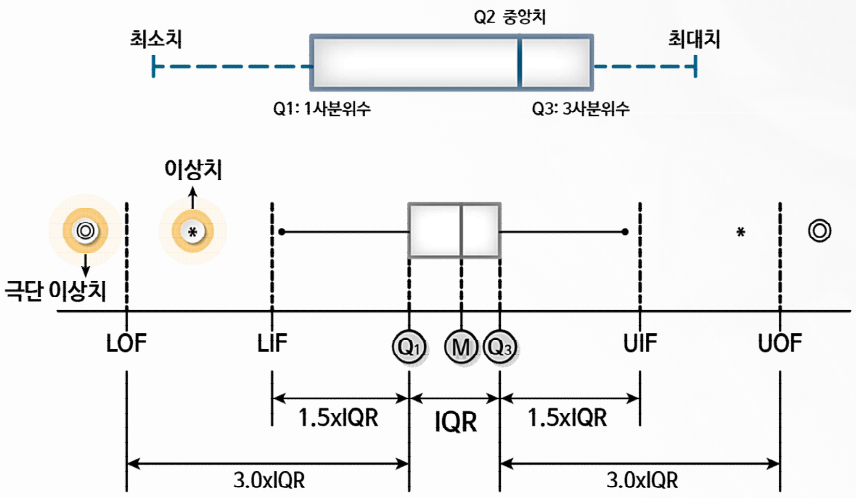
데이터를 크기 순으로 정렬한 통계량인 순서 통계량, 평균, 중위값, 사분위수,표준편차 등을 사용한다.

02 탐색적 데이터 분석

08 시각화를 통한 탐색적 자료분석

(2) 측정형 변수

- 기초통계량을 시각화 하는데 대표적으로 box plot이 있다.



08 시각화를 통한 탐색적 자료분석

(2) 측정형 변수

❶ LIF(Lower Inner Fence)

하 내부 울타리(Q1 − 1.5 *IQR)

❸ LOF(Lower Outer Fence)

하 외부울타리(Q1 − 3.0 *IQR)

❷ UIF(Upper Inner Fence)

상 내부울타리(Q3 + 1.5 *IQR)

❹ UOF(Upper Outer Fence)

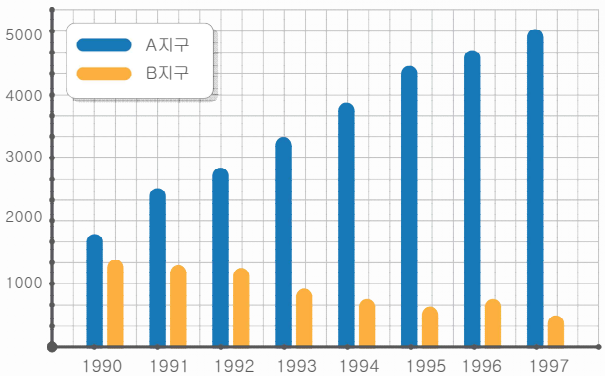
상 외부울타리(Q3 + 3.0 *IQR)

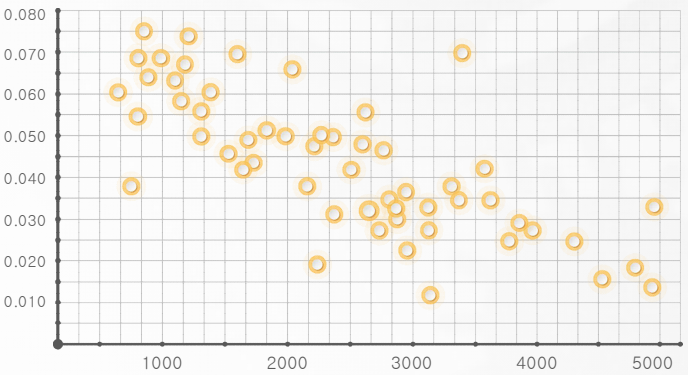
02 탐색적 데이터 분석

08 시각화를 통한 탐색적 자료분석

(2) 측정형 변수

- EDA 위한 파이썬 라이브러리: numpy, pandas, matplotlib, seaborn
  Boxplt, 산점도, 막대그래프, 히스토그램, 히트맵 등





APPLICATION

03 데이터 전처리 실습

01 개요

(1) 데이터 셋

- 기상 데이터 + 편의점 데이터

	편의점 데이터셋	기상 데이터셋
레코드 개수	2,707,786	59,113
기간	2016/01/01 ~ 2018/12/31	2016/01/01 ~ 2018/12/31
컬럼 개수	7개	11개
주요 컬럼	광역시 시군구 일자	광역시 시군구 일자
파일명(크기)	GS25.csv (147,610KB)	기상데이터.csv (3,694KB)

02 데이터 준비 및 전처리

(1) 데이터 준비하기

1. LOAD : 원본 데이터 불러오기
2. Column Subset : 필요한 컬럼 선택하기
3. Row Subset : 특정 지역 데이터만 선택하기
4. 상품의 각 카테고리 데이터를 변수화하기
5. 기상 데이터와 편의점 데이터를 결합하기

(2) 분석 목표

1. 기상 데이터와 편의점 데이터의 상관 관계 분석

03 데이터 전처리 실습

02 데이터 준비 및 전처리

① 데이터 준비하기

```
#Load
# 원본 데이터 불러오기

rawdata = open('기상데이터.csv', 'rb').read()
result = chardet.detect(rawdata)
char = result['encoding'] #인코딩 확인

Df = pd.read_csv( ' 기상데이터.csv ' , encoding = 'EUC-KR')
df2 = pd.read_csv('GS25.csv', encoding = 'EUC-KR')
```

02 데이터 준비 및 전처리

① 데이터 준비하기

```
# 강남구에 대한 관측일,관측번호,법정동코드.최고기온,최대풍속,최소기온,평균기온,
평균 상대습도,평균 풍속, 합계 강수량 데이터 준비

df[df['bigcon_weather.bor_nm'] == '강남구'].describe()

df[df['bigcon_weather.bor_nm'] == '강남구'].head()
```

03 데이터 전처리 실습

02 데이터 준비 및 전처리

② 데이터 전처리

```
# 강남구 요일별 / 성별 / 판매량 / 타겟 그룹핑
dfg2 = dfg.groupby(['date', 'korea_cvs.gen_cd',
'korea_cvs.category']).sum().reset_index()
dfg2 = dfg2[dfg2['korea_cvs.category'] == '과자'].reset_index(drop=True)

# 이상치 처리
dfg2.loc[1015, 'korea_cvs.adj_qty'] = 2558

# 날씨 데이터와 과자 판매 데이터 병합
df_m = df_t.merge(dfg2, on='date')
df_m.head()
```

02 데이터 준비 및 전처리

② 데이터 전처리

```
#Plotly 시각화 기반 EDA : 강남구 지역별 과자가격 상승 대비 온도 상관 분석

fig = make_subplots(specs=[[{"secondary_y": True}]])

fig.add_trace(go.Scatter(
        x=df_m.date,
        y=df_m['bigcon_weather.max_ta'],
        name="Max Temp",
        line_color='deepskyblue',
        opacity=0.9), secondary_y=True)
```

03 데이터 전처리 실습

02 데이터 준비 및 전처리

② 데이터 전처리

```
fig.add_trace(go.Scatter(
        x=df_m.date,
        y=df_m['korea_cvs.adj_qty'],
        name='타겟',
        marker_color='indianred',
        opacity=0.8
))

fig.update_layout(title_text="강남구 타겟 상품 vs 온도")
fig.show()
```

02 데이터 준비 및 전처리

② 데이터 전처리

```
#Plotly 시각화 기반 EDA : 강남구 지역의 요일별 과자판매량 상관분석

 #강남구 요일별 과자 판매량
dfg3 = dfg.groupby(['date', 'korea_cvs.category']).sum()
dfg3.head(12)
```

03 데이터 전처리 실습

02 데이터 준비 및 전처리

② 데이터 전처리

```
# 날씨데이터와 과자판매 항목 데이터 결합 및 히트맵을 통한 EDA 분석
dfg6 = dfg6.rename(columns = {'과자' : 'Snack',
                              '라면':'Ramen',
                              '마스크':'Mask',
                             '맥주': 'Beer',
                             '면도기' : 'Shaving',
                             '생리대':'Sanitary pad',
                             '숙취해소제' : 'Hangover Relief',
                             '스타킹' : 'Stockings',
                             '아이스크림' : 'Ice cream',
                             '우산' : 'Umbrella',
                             '탄산음료' : 'Drinks',
                             '생수':'Water'})
corr = dfg6.corr()
```

02 데이터 준비 및 전처리

② 데이터 전처리

```
# 히트맵 기반 상관관계 EDA분석
sns.heatmap(corr, square=True, vmin=0, vmax=1, annot=True,
cmap='Blues',
      xticklabels=corr.columns,
      yticklabels=corr.columns)
```

03 데이터 전처리 실습

02 데이터 준비 및 전처리

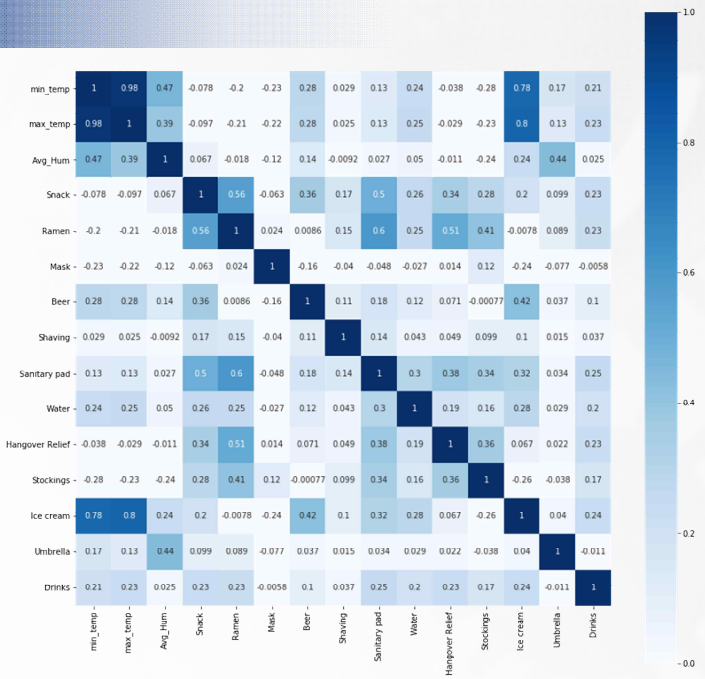
② 데이터 전처리

1. 온도와 상관관계가 가장 높은 상품은?

   아이스크림

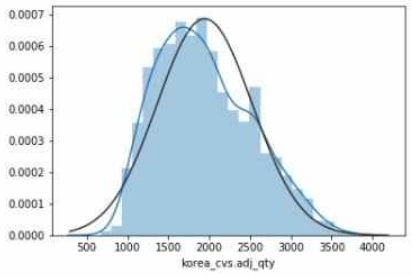
2. 숙취해소제 구매고객이 같이 구매할 확률이 높은 상품은?

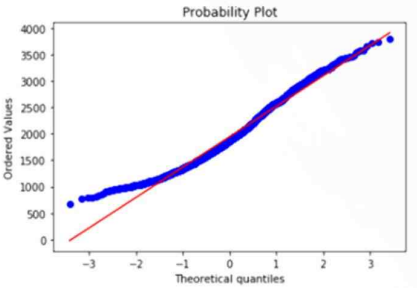
   라면

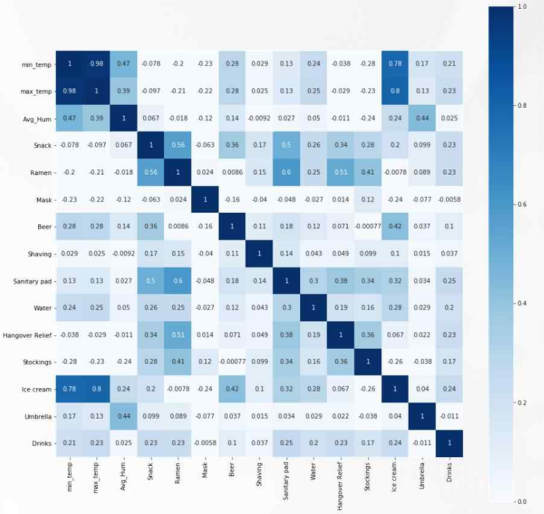


② 데이터 전처리

#EDA 시각화



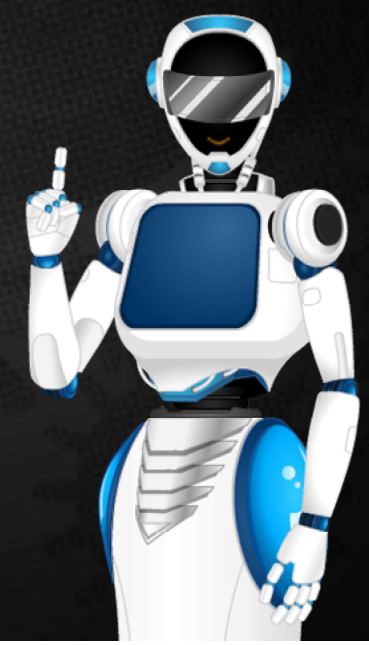




SUMMARY

학습정리

- 데이터 전처리 이해
- 탐색적 자료 분석
- 데이터 전처리 실습



quiz

정리퀴즈

1. 데이터 타입 확인 함수는?

정답 및 해설은 강의에서 확인 할 수 있습니다.

quiz

정리퀴즈

2. 삭제의 관련함수는?

정답 및 해설은 강의에서 확인 할 수 있습니다.

quiz

정리퀴즈

3. 인덱스 핸들링의 관련함수는?

정답 및 해설은 강의에서 확인 할 수 있습니다.

참고 문헌

REFERENCE

- 빅데이터 분석과 활용 :
  4차 산업혁명 현장 전문가가 알려주는
  by 박인근, 홍지후, 강남규, 김성호, 정구범, 제이펍 ,2019

- 데이터 분석 전문가 가이드 (ADP, ADsP)
  by 한국데이터진흥원, 한국데이터진흥원, 2019

서체 출처 : 에스코어드림체-㈜에스코어, 나눔글꼴체-㈜네이버