ChatGPT를 활용한 실무 데이터 분석

- 데이터와 기술통계1 -

presented by A.Iglue

오늘의 학습 목표

- 데이터와 데이터 분석의 의미를 이해할 수 있다.
- 데이터 분석에서 통계의 역할을 이해할 수 있다.
- Pandas 라이브러리를 활용하여 파일을 읽고 쓸 수 있다.

데이터와 데이터 분석

■ 데이터란?

- ✓ 흥미가 있는 대상을 관찰하고 측정하여 얻는 수치, 문자, 기호의 집합
- ✓ 자체적으로는 특정한 의미를 지니지 않음.
- ✓ 그러나 분석하고 해석함으로서 유용한 정보를 제공할 수 있음.

■ 데이터 분석이란?

- ✓ 수집된 데이터를 검토, 해석하여 유용한 정보 및 인사이트를 도출하는 과정
- ✓ 비즈니스 인텔리전스, 통계적 분석, 예측 모델링을 활용하여 데어터에 숨겨져 있는 패턴, 연관성, 추세 발견
- ✓ 데이터 수집, 정리, 변환, 모델링하여 의미있는 인사이트나 지식을 얻기 위해 사용되는 기술과 절차를 포함함.

데이터 분석의 목적

■ 데이터를 요약하는 것(기술적•탐색적 데이터 분석)

- ✓ 데이터의 대표값, 분산, 분포 모양 등을 기술하는 간략한 데이터 분석
- ✓ 가시화 도구를 활용하여 데이터에 대한 이해를 높임.

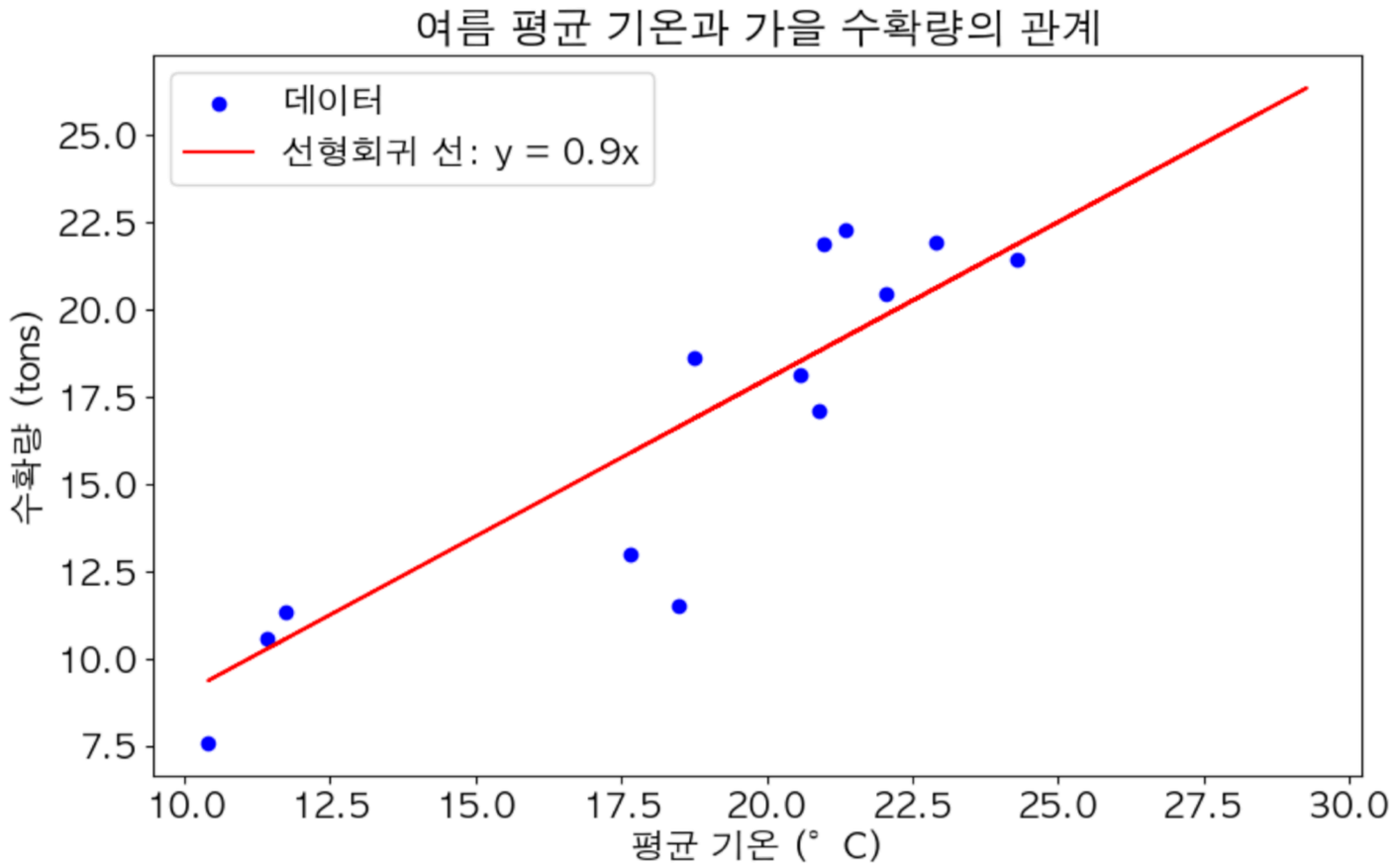
■ 데이터를 설명하는 것(확증적 데이터 분석)

- ✓ 데이터(표본)가 가진 성질과 관계성을 명확히 밝히고 이를 이해하는 데이터 분석
- ✓ 데이터를 정량적이고 객관적으로 평가하여 대상(모집단)이 가진 성질과 관계성을 올바르게 추론하고자 하는 시도임.
- ✓ 통계추론(모수 추정), 통계적가설검정 등

■ 새로 얻을 데이터를 추론하는 것(예측적 데이터 분석)

- ✓ 이미 얻은 데이터를 기반으로, 이후 새롭게 얻을 데이터를 예측하는 데이터 분석
- ✓ 의료나 비즈니스 현장의 의사결정에서 중요한 역할을 담당
- ✓ 선형 예측 분석, 기계학습 예측 분석 등

회귀 분석과 예측


여름 평균 기온과 가을 수확량의 관계
데이터
선형회귀 선: y = 0.9x
수확량 (tons)
평균 기온 (° C)
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0

통계학의 쓸모

- 퍼짐(산포, dispersion)이 있는 데이터에 대해 설명이나 예측 가능
- “데이터의 퍼짐”은 대상이 가진 성질이나 관계성의 본모습을 감추고, 정확하게 파악할 수 없도록 함.
- 통계학은 데이터의 퍼짐을 ‘불확실성’이라 평가하고, 정량화하여 통계학의 목적인 “대상의 설명과 예측”을 수행함.
- 통계학은 데이터 퍼짐이나 불확실성을 확률로 표현하며, 수학의 확률론은 데이터 분석의 근간이 됨.

통계학의 청사진

■ 기술통계(Descriptive Statistics)

- ✓ 수집한 데이터를 정리하고 요약하는 통계
- ✓ 데이터 그 자체의 특성이나 경향을 알 수 있음.
- ✓ 확보한 데이터에만 집중하여 자체 성질을 이해하는 것이 목적임.

■ 추론통계(Inferential Statistics)

- ✓ 수집한 데이터(표본)으로부터 데이터의 발생원(모집단)을 추정하는 통계
- ✓ 대상을 이해하거나 미지의 데이터를 예측하기 위해서는, 데이터 자체를 너머 데이터의 발생원에 대해 알 필요가 있음.

■ 통계적 추론(Statistical Inference)

- ✓ 데이터(표본)에서 가정한 확률 모형의 성질을 추정하는 방법으로 모수에 초점을 둠.

■ 가설검정(Hypothesis Test)

- ✓ 알고 싶은 것을 가설로 세우고, 데이터를 통해 도출한 객관적인 수치를 이용하여 가설의 채택 여부를 판별하는 방법으로 표본 통계량에 초점을 둠.

확률 모형이란?

- 데이터는 대상을 관찰함으로 얻지만, 대상의 성질 자체는 직접 관찰할 수 없으며 다루기도 어려움.
- 데이터는 비교적 단순한 활률장치에서 생성되었다고 가정하고, 관찰된 데이터를 통해 확률장치의 성질을 추론할 수 있는데, 이 확률장치를 확률모형이라고 함.
- 데이터와 확률 모형의 관계 .vs. 표본과 모집단의 관계
- ex) 주사위를 던졌을 때, 1이 나올 확률은? 만약 확률분포가 없다면…

변수

■ 변수(variables)란?

- ✓ 변수란 데이터 중 공통의 측정 방법으로 얻은 같은 성질의 값의 집단
- ✓ 변수의 개수에 따라, 단일변수(1개), 다변수(2개 이상)로 나눔.
- ✓ 통계학에서 변수의 개수를 ‘차원’이라고 표현함.

데이터 유형

■ 데이터 유형

✓ 변수의 유형에 따라 데이터 분석 방법이 달라지기 때문에, 데이터를 수집할 때나 분석할 때, 변수가 어떤 유형인지 주의 깊게 살펴야 함.

✓ 수치형(양적) 변수와 양적 변수로 나눔.

■ 수치형(양적) 변수

✓ 이산형 변수(discrete variables): 셀 수 있는 값 또는 개별적이고 분리된 값을 가지는 변수

- ex) 주사위 눈의 값, 빈도, 횟수

✓ 연속형 변수(continuous variables): 어떤 범위 내에서 연속된 어떤 값을 취할 수 있는 변수

■ 범주형(질적) 변수

✓ 숫자가 아닌 범주로 표현되는 변수

✓ ex) 설문조사의 예/아니오, 동전의 앞/뒤, 맑음/흐림/눈/비

Pandas란?

■ Pandas란?

✓ 고수준의 데이터 분석 도구를 제공하는 오픈 소스 라이브러리

✓ Series, DataFrame이라는 구조를 통해, csv, excel 파일 등 다양한 형태의 데이터를 쉽게 읽고, 처리 및 분석할 수 있는 라이브러리

✓ NumPy의 고성능 배열 계산 기능을 활용할 수 있고, 데이터 과학 및 분석 파이프라인에서 중요한 역할을 함.

■ Pandas의 기능

✓ 데이터 조작 - 데이터 정제, 변환, 결합, 정렬, 집계 등을 쉽게 수행함.

✓ 데이터 분석 - statsmodels, scipy.stats 등과 함께 데이터 분석에 활용됨.

✓ 데이터 입출력 - 다양한 파일 형식(csv, Excel, SQL DB, JSON)으로 읽고 쓸 수 있음.

✓ 시계열 데이터 처리 - 날짜 범위 생성, 날짜 쉬프팅 등 시계열 분석을 위한 특수 기능 제공

오늘의 실습

- Colab 사용법 알아보기
- Jupyterlab 둘러보기
- Pandas library 설치하기
- Pandas로 기술적 데이터 분석하기
  - ✓ 데이터출처: https://www.kaggle.com/datasets/amirmotefaker/supply-chain-dataset
  - ✓ 데이터 csv 파일 불러오기
  - ✓ 데이터 정보 확인하기
  - ✓ 변수 목록 확인하기
  - ✓ 데이터(수치형, 범주형) 기술 통계량 확인하기
  - ✓ 기술 통계량 csv 파일로 저장하기