데이터로 이야기하기(실습)

> 데이터로 **질문을 만들고, 시각화로 답을 찾는** 실습

GOAL

- 데이터를 기반으로 스스로 질문을 만들고 답할 수 있다.
- 분석 결과를 시각적으로 정리하고 설명할 수 있다.

“데이터로 이야기한다”는 건?

- 숫자를 나열하는 것이 아니라 **메시지를 전달**하는 것
- 예시:
  - “혈당이 높을수록 당뇨 위험이 커진다.”
  - “BMI가 높을수록 Outcome=1 비율이 높다.”

분석 과정 한눈에 보기

단계	내용	도구
1 문제 정의	궁금한 질문 만들기	-
2 데이터 불러오기	CSV, TSV, Excel	`pd.read_csv()`
3 탐색 (EDA)	결측치, 통계, 분포	`describe()`, `isna()`
4 분석	변수 간 관계 보기	`corr()`, `groupby()`
5 시각화	패턴을 시각적으로 전달	Matplotlib / Seaborn
6 결론	인사이트 정리	발표 슬라이드/보고서

문제 정의 (Problem Definition)

“어떤 요인이 당뇨병(Outcome)에 가장 큰 영향을 미칠까?”

- 가설 1 : 혈당(Glucose)이 높을수록 당뇨병일 확률이 높다.
- 가설 2 : BMI(비만도)가 높으면 당뇨병 발병 가능성이 높다.
- 가설 3 : 나이가 많을수록 당뇨병 위험이 높을 수 있다.

Feature에 대한 이해

상황	해석 포인트
Glucose ↑	인슐린 저항성 → 당뇨 위험 급상승
BMI ↑	비만으로 인슐린 감수성 저하
Age ↑	대사율 저하 + 생활습관 누적
Pregnancies ↑	임신성 당뇨 이력 가능성
Insulin = 0	실제로는 미측정 → 결측 처리 필요
SkinThickness ↑	지방 두께 증가 → 대사 이상 징후
DPF ↑	가족력 존재 시 유전적 위험

데이터 불러오기 (Loading Data)

```
import pandas as pd
df = pd.read_csv("diabetes.csv")
df.head()
```

탐색 (EDA: Exploratory Data Analysis)

- 탐색적 데이터 분석
- 데이터를 분석하기 전에 데이터의 성격을 알아보는 과정

EDA 주요단계

단계	설명	예시 코드
1 데이터 구조 확인	행/열, 자료형, 컬럼 이름	`df.info()`, `df.head()`
2 결측치 탐색	비어있는 값 확인	`df.isnull().sum()`
3 기초 통계 확인	평균, 표준편차, 분포	`df.describe()`
4 시각화 탐색	분포, 관계, 이상치	`sns.histplot()`, `sns.boxplot()`
5 상관관계 분석	변수 간 연관성 파악	`df.corr()`, `sns.heatmap()`

왜 중요할까?

이유	설명
1. 데이터 품질 확인	결측, 이상치, 중복을 미리 발견
2. 인사이트 도출	“어떤 변수가 결과에 영향을 주는가?” 감 잡기
3. 모델 설계 기반	어떤 변수를 쓸지, 스케일링이 필요한지 판단
4. 잘못된 데이터 방지	모델을 오염시키는 데이터를 걸러냄

탐색 (EDA: Exploratory Data Analysis)

(1) 결측치 확인 및 채우기

→ 0은 실제 결측으로 간주해야 하는 컬럼이 있음

```
cols_zero_as_na = ["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]
df[cols_zero_as_na] = df[cols_zero_as_na].replace(0, None)
df.isna().mean().sort_values(ascending=False)

df["Glucose"].fillna(df["Glucose"].median(), inplace=True)
```

(2) 기초 통계 요약

```
df.describe()
```

분석 (변수 간 관계 보기)

(1) 당뇨병 여부(Outcome)별 평균 비교

```
df.groupby("Outcome")[["Glucose", "BMI", "Age"]].mean()
```

✅ 당뇨 그룹은 평균적으로 혈당, BMI, 나이가 모두 높다!

(2) 상관관계 분석

```
corr = df.corr(numeric_only=True)
corr["Outcome"].sort_values(ascending=False)
```

시각화 (Visualization)

(1) 혈당 분포 (단변량)

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.histplot(df["Glucose"].dropna(), kde=True)
plt.title("혈당 분포 (Glucose)")
plt.show()
```

(2) Outcome별 혈당 차이 시각화 (이변량)

```
sns.boxplot(x="Outcome", y="Glucose", data=df)
plt.title("Outcome별 혈당 차이")
plt.show()
```

✅ 당뇨(Outcome=1) 그룹의 중앙값이 훨씬 높음

(3) Glucose별 BMI (이변량)

```
sns.scatterplot(data=df, x='Glucose', y='BMI', hue='Outcome')
```

(4) 전체 상관관계 Heatmap

```
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Feature 상관관계 Heatmap")
plt.show()
```

결론 (Insight)

- 어떤 변수가 Outcome과 가장 관련이 큰가?
- BMI도 의미 있나?
- 나이의 영향은?
- 추가로 보면 좋을 변수는?

실습

1. 1~2개의 가설 세우기
2. 데이터 탐색 및 시각화
3. 인사이트 도출