# ESAA YB4조 - 방학 플젝 자료 조사

## 칼럼 별 분석

### 1. ID

- 데이터의 각 행을 고유하게 식별하는 샘플 고유 ID
- 형식: 문자열 (예: `TRAIN_000000`, `TRAIN_000001`)

### 2. Month

- 항공편의 출발 월
- 값 범위: 1~12 (1월~12월)
- 데이터 유형: 정수형

### 3. Day_of_Month

- 항공편이 출발한 월의 날짜
- 값 범위: 1~31
- 데이터 유형: 정수형

### 4. Estimated_Departure_Time

- 항공편의 **예상 출발 시간**
- 데이터는 **HHMM** 형식의 숫자로 제공되며, 이를 **HH:MM** 형식으로 변환할 수 있음
  - 예: `740` → `07:40`, `1610` → `16:10`
- 일부 데이터는 결측값(빈 값)으로 표시될 수 있음
- 데이터 유형: 실수형 (float)

### 5. Estimated_Arrival_Time

- 항공편의 **예상 도착 시간**
- 데이터는 **HHMM** 형식의 숫자로 제공되며, 이를 **HH:MM** 형식으로 변환할 수 있음

    - 예: `1024` → `10:24`, `1805` → `18:05`
- 일부 데이터는 결측값(빈 값)으로 표시될 수 있음
- 데이터 유형: 실수형 (float)

## 6. Cancelled

- 항공편이 취소되었는지를 나타내는 이진 변수
    - **0**: 취소되지 않음
    - **1**: 취소됨
- 데이터 유형: 정수형 (0 또는 1)

## 7. Diverted

- 항공편이 경유되었는지를 나타내는 이진 변수
    - **0**: 경유되지 않음
    - **1**: 경유됨
- 데이터 유형: 정수형 (0 또는 1)

## 8. Origin_Airport

- 항공편의 **출발 공항 코드**
- **IATA 공항 코드**로 제공되며, 알파벳 대문자 세 글자로 구성됨
    - 예: `OKC`, `ORD`, `CLT`, `LAX`, `SFO`
- 데이터 유형: 문자열

## 9. Origin_Airport_ID

- 항공편의 **출발 공항 고유 ID**입니다.
- **US DOT ID**로 제공되며, 숫자로 구성됨
    - 예: `13851`, `13930`, `11057`
- 데이터 유형: 정수형

## 10. Origin_State

- 항공편 **출발 공항이 위치한 주**의 이름
  - 예: `Oklahoma`, `Illinois`, `North Carolina`, `California`
- 데이터 유형: 문자열

## 11. Destination_Airport

- 항공편의 **도착 공항 코드**
- **IATA 공항 코드**로 제공되며, 알파벳 대문자 세 글자로 구성됨
  - 예: `HOU`, `SLC`, `LGA`, `EWR`, `ACV`
- 데이터 유형: 문자열

## 12. Destination_Airport_ID

- 항공편의 **도착 공항 고유 ID**
- **US DOT ID**로 제공되며, 숫자로 구성됨
  - 예: `12191`, `14869`, `12953`
- 데이터 유형: 정수형

## 13. Destination_State

- 항공편 **도착 공항이 위치한 주**의 이름
  - 예: `Texas`, `Utah`, `New York`, `New Jersey`, `California`
- 데이터 유형: 문자열

## 14. Distance

- 출발 공항과 도착 공항 간의 **거리**를 나타냄
- 단위는 마일(mile)
  - 예: `419.0`, `1250.0`, `544.0`, `2454.0`, `250.0`
- 데이터 유형: 실수형 (float)

## 15. Airline

- 항공편을 운항하는 **항공사 이름**

    - 예: `Southwest Airlines Co.`, `SkyWest Airlines Inc.`, `American Airlines Inc.`, `United Air Lines Inc.`
- 데이터 유형: 문자열

## 16. Carrier_Code(IATA)

- 항공편을 운항하는 **항공사의 고유 코드**
- **IATA 공항 코드**로 제공되며, 알파벳 두 글자 또는 세 글자로 구성됨
    - 예: `WN`, `UA`, `AA`
- 데이터 유형: 문자열

## 17. Carrier_ID(DOT)

- 항공편을 운항하는 **항공사의 고유 ID**
- **US DOT ID**로 제공되며, 숫자로 구성
    - 예: `19393.0`, `20304.0`, `19805.0`
- 일부 데이터는 결측값(빈 값)으로 표시될 수 있음
- 데이터 유형: 실수형 (float)

## 18. Tail_Number

- 항공편을 운항하는 **항공기의 고유 등록번호**
    - 예: `N7858A`, `N125SY`, `N103US`
- 일부 데이터는 결측값(빈 값)으로 표시될 수 있음
- 데이터 유형: 문자열

## 19. Delay

- 항공편의 **지연 여부**를 나타내는 타깃 변수
    - `Not_Delayed`: 지연되지 않음
    - `Delayed`: 지연됨
- 이 변수는 예측해야 하는 목표 변수
- 일부 데이터는 레이블이 존재하지 않을 수 있음

- 데이터 유형: 문자열

---

## 변수 그룹화 기준

1. **시간 관련 변수** → 지연 시간과 계절적 패턴 분석에 유리한 그룹
2. **경로 관련 변수** → 출발/도착 공항 및 주(state)별 지연 패턴 분석 그룹
3. **항공사 및 거리 관련 변수** → 항공사별 지연 패턴 및 거리와 지연의 상관관계 분석 그룹
4. **항공편 상태 및 기타 변수** → 취소 여부, 경유 여부와 지연의 관계를 분석할 수 있는 그룹

# 그룹 1: 시간 관련 변수 (계절 및 시간대 패턴 분석)

이 그룹에서는 **출발 시간, 도착 시간, 월, 날짜** 등을 분석하여 지연이 **시간적 패턴**과 어떤 관계가 있는지 살펴볼 수 있습니다.

### 포함 변수:

- `Month` (출발 월)
- `Day_of_Month` (출발 날짜)
- `Estimated_Departure_Time` (출발 시간)
- `Estimated_Arrival_Time` (도착 시간)

### 추천 시각화:

- 월별 지연 비율 막대그래프
- 시간대별 지연 비율 히트맵
- 날짜별 지연 추세 라인차트

---

# 그룹 2: 경로 관련 변수 (지역별 패턴 분석)

이 그룹에서는 **출발지 및 도착지 공항**과 **주(state)** 별로 지연이 발생하는 패턴을 분석할 수 있습니다.

### 포함 변수:

- `Origin_Airport` (출발 공항 코드)

- `Destination_Airport` (도착 공항 코드)
- `Origin_State` (출발 주)
- `Destination_State` (도착 주)

### 추천 시각화:

- 주별 지연 비율 지도 시각화
- 공항별 지연 빈도 막대그래프
- 출발지-도착지 간 지연 비율 히트맵

---

## 그룹 3: 항공사 및 거리 관련 변수 (항공사 및 거리 분석)

이 그룹에서는 **항공사**와 **거리**가 지연에 미치는 영향을 분석할 수 있습니다. 항공사별로 지연 패턴이 다르고, **거리**가 지연에 어떤 영향을 미치는지 확인할 수 있습니다.

### 포함 변수:

- `Airline` (항공사 이름)
- `Carrier_Code(IATA)` (항공사 코드)
- `Distance` (출발지와 도착지 간 거리)
- `Tail_Number` (항공기 고유 번호)

### 추천 시각화:

- 항공사별 지연 비율 막대그래프
- 거리와 지연 시간의 상관관계 산점도
- 항공기별 지연 패턴 히트맵

---

## 그룹 4: 항공편 상태 및 기타 변수 (항공편 상태 분석)

이 그룹에서는 **항공편 취소 여부, 경유 여부**와 지연의 관계를 분석할 수 있습니다.

### 포함 변수:

- `Cancelled` (취소 여부)

- `Diverted` (경유 여부)
- `Carrier_ID(DOT)` (항공사 고유 ID)
- `Delay` (지연 여부 - 타깃 변수)

### 추천 시각화:

- 취소된 항공편과 지연 비율 비교 막대그래프
- 경유 여부와 지연 비율 비교 차트
- 항공사별 지연 비율 비교 그래프

---

## 파생 변수 아이디어

1. **출발 시간대 (Departure_Time_Block)**
   - 출발 시간을 **시간대별**로 나누어 아침, 오후, 저녁, 심야 등으로 분류.
   - 예: `06:00~12:00` → Morning, `12:00~18:00` → Afternoon
2. **출발 요일 (Day_of_Week)**
   - 날짜 정보를 요일로 변환하여 특정 요일에 지연이 더 많이 발생하는지 분석.
3. **거리 범주화 (Distance_Category)**

- 거리를 기준으로 **단거리, 중거리, 장거리**로 분류.
- 예: 0-500마일 → Short, 500-1500마일 → Medium, 1500마일 이상 → Long