출퇴근 시간대 기반 지하철 혼잡도 분석을 위한 데이터 준비 절차

본 분석에서는 팀원 윤여윤님이 1차 전처리를 완료한 데이터를 기반으로, 추가적인 병합 및 필터링 과정을 통해 최종 분석용 데이터를 구축하였다.

1. total.csv를 기반으로 분석에 필요한 열('날짜', '호선', '역명', '상하구분', '기온', '시간강수량', '혼잡도'만 선별하고, 날짜 정보를 이용해 '년', '월', '일', '시' 파생 컬럼을 생성하였다. 이를 활용하여 '계절' 및 '시간대' 변수도 추가하였다.

2. 강수량 데이터를 바탕으로 '시간강수량_상태' 변수를 생성하였다. 분류 기준은 다음과 같다:

- 강수량 없음 (0mm 미만)
- 약한 비 or 눈 (1~3mm)
- 보통 비 or 눈 (3~15mm)
- 강한 비 or 눈 (15mm 이상)

3. 이상치 및 결측치 처리를 다음 기준에 따라 수행하였다:

- 시간강수량 == -99 → 결측치 처리 후 제거
- 기온, 체감온도는 -30도 ~ 50도 범위를 벗어나는 값 제거
- 혼잡도: 0~100 범위 밖의 값 제거
- 결측치 및 중복 행 제거

4. 출퇴근 시간대 분석을 위해 출근(6~9시) 및 퇴근 (17~20시) 구간만 필터링하고, 계절, 호선, 상하구분, 출퇴근시간, 시간강수량_상태를 기준으로 그룹화하여 평균 혼잡도를 산출하였다.

- **분석에 사용한 주요 컬럼**:
    - 계절 (봄, 여름, 가을, 겨울)
    - 시간대 (시간 단위, 예: 07시, 08시)
    - 출퇴근시간 (출근 / 퇴근)
    - 호선, 역명, 상하구분
    - 기온, 시간강수량, 혼잡도
    - 시간강수량 상태 (강수량 없음 / 비 / 눈)
      → 겨울철 강수량이 없지 않은 경우 '비 or 눈'으로 분류 (Tableau 계산 필드 활용)
    - 평균 혼잡도 (그룹별 평균 혼잡도 수치)