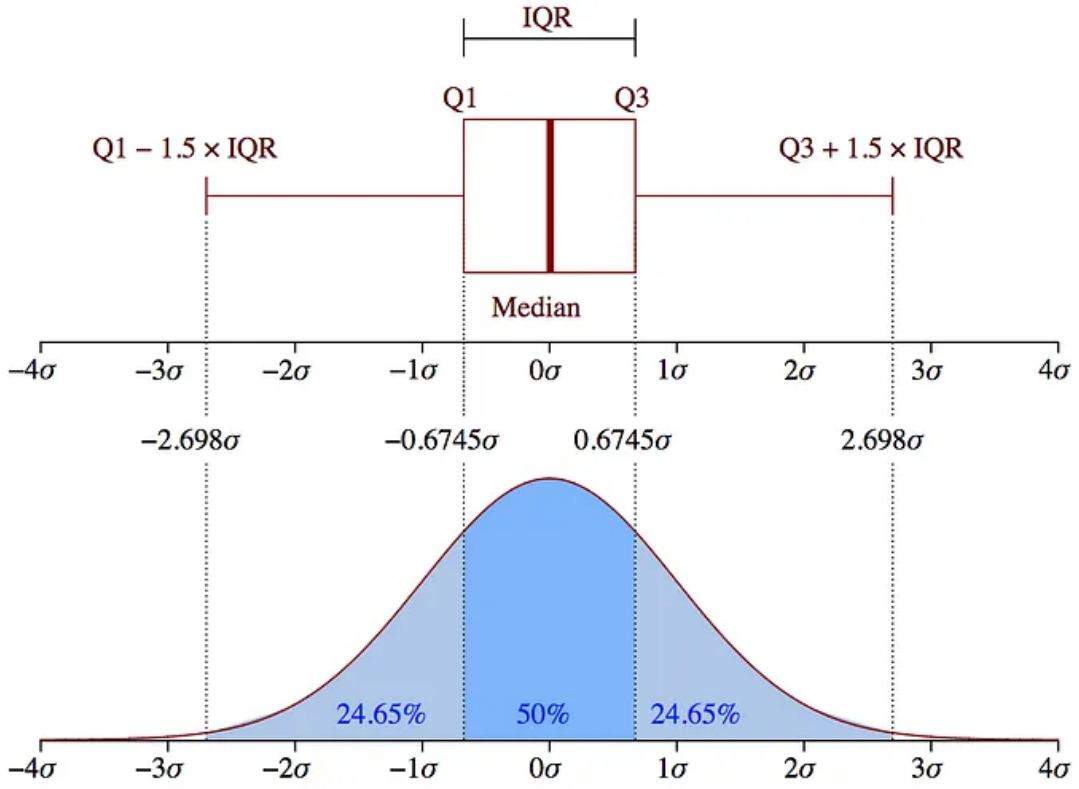
2023-11-30 회의

- 프로젝트 요약

1. 데이터 선정 https://data.seoul.go.kr/dataList/OA-12921/F/1/datasetView.do#
   15년치의 서울 지하철 일별 역별 시간별 승하차 인원 데이터
2. 필요 컬럼 결정 (날짜, 호선명, 역명, 시간) → 승차 하차는 합산해서 시간대별 유동 인구로 표현
3. 필요 컬럼 생성 (각 날짜 별/역 별 유동인구의 총합을 새로운 컬럼으로 생성한다)
4. 프로젝트 목적 : 각 날짜별로 특정 역의 유동 인구를 분석하고, 다른 날짜나 월별, 년도별 평균과 비교하여 특출나게 인원이 많거나 적은 유동 인구를 가진 역을 추출하여 그 원인을 뉴스 페이지 크롤링을 통해 밝혀내는 것.
5. 이상치 감지 기준 (IQR이상치 검증 방식 활용)



5.1. 연 평균 비교 ex) 2023년 12월 20일 홍대입구역의 유동 인구를 15개 년치 12월 20일
홍대입구역 유동 인구 평균과 비교

5.2. 월 평균 비교 ex) 2023년 9월 27일 사당역의 유동 인구를 2023년 9월 사당역 유동 인구 평균과 비교

- 파이프 라인

1. https://data.seoul.go.kr/dataList/OA-12921/F/1/datasetView.do# 에서 데이터 수집
2. wget 으로 hdfs에 수집한 데이터를 업로드
3. 데이터를 제플린 pyspark 를 활용해 데이터 이상치 검증