※ "비정상적으로 튀는 데이터"를 찾아 제거/수정하는 작업)

---

1. 주요 이상치 유형 **&** 처리 방안

항목	이상치 예시	처리 방법	근거
연봉	신입 1억↑, 경력 1000만원↓	직군별 상하위 1% 제거	시장 평균과 현격한 차이
기술 스택	20개↑ 기술 나열	상위 5개 기술만保留	과도한 우대사항은 노이즈
경력	"경력 30년" (오타)	최대 경력 15년으로 제한	현실적인 경력 범위 적용
지역	"해외" (국내 분석 시)	필터링 제외	분석 대상 지역과 불일치
공고 기간	게시일 3년 전	최신 데이터(1년 이내)만保留	구직 시장 현황 반영 필요

---

2. 이상치 탐지 방법

- 수치형 데이터 **(**연봉 등**)**:

```
python
```

Copy

```
# IQR(사분위수) 기준으로 상하위 이상치 탐지

Q1 = df['salary'].quantile(0.25)

Q3 = df['salary'].quantile(0.75)

IQR = Q3 - Q1
```

- `outliers = df[(df['salary'] < Q1 - 1.5*IQR) | (df['salary'] > Q3 + 1.5*  )]`
- 텍스트 데이터 **(**기술 스택 등**)**:
    - 토큰 개수 제한 (예: 기술 10개 초과 → 필터링)
    - 비정상 키워드 제거 (예: "엑셀"이 포함된 AI 개발자 공고)

---

3. 처리 결과 예시

원본 데이터	이상치 판단 근거	처리 후
신입 연봉 1.2억	IT 신입 평균 4,000만원 대비 3배↑	제외
기술: Python,Java,React,...,Excel	기술 15개 나열	상위 5개(Python,Java,React,...)保留
경력: "30년" (오타)	개발 직군 최대 15년 경력 가정	"3년"으로 수정

---

4. 유저 친화적 안내

- 시각화로 설명: 박스 플롯으로 이상치 위치 표시

```
python
Copy
```

```
import seaborn as sns
```

- `sns.boxplot(x=df['salary'])  # 연봉 분포에서 튀는 값` 강조
- 리포트 표기:
  *"연봉 데이터의 **2%**가 이상치로 제거되었습니다 (신입 **1**억↑, 경력 **1000**만원↓)."*

✅ 기획안 반영 팁

- 자동화 룰 설정:
  - "신입 연봉 > 6,000만원 → 자동 검토 대상"
- 유저 참여 기능:
  - "이 데이터를 제외할까요? [예/아니오]" (직접 선택 가능)

⚠️ 주의: 이상치가 진짜 특이 케이스(예: AI 박사 초봉 1억)일 수 있으니, 도메인 지식과 병행 검토 필요!

📌 채용 데이터 이상치 처리 핵심 요약

1. 주요 이상치 유형

- 연봉: 신입 1억↑, 경력 1000만원↓
- 기술 스택: 20개↑ 과도한 기술 요구
- 경력: "30년" (오타)
- 지역: "해외" (국내 분석 시)

2. 처리 방법

- 제거: 시장 평균과 현격히 다른 데이터 (상하위 1%)
- 수정: 오타/비현실적 값 조정 (예: "30년" → "3년")
- 필터링: 분석 대상과 무관한 데이터 (예: 해외 공고)

3. 간단 예시

- 원본: 신입 연봉 1.2억 → 처리: 제외
- 원본: 기술 15개 → 처리: 상위 5개만保留

4. 유저 안내

- "이상치 2% 자동 제외됨"
- 박스 플롯으로 이상치 시각화

💡 기획 시 추가할 내용:

- 자동화 규칙 (예: "신입 연봉 > 6,000만원 → 경고")
- 유저 확인 기능 ("이 데이터를 제외할까요?")

한 줄 요약: "튀는 데이터는 제거/수정해서 분석 신뢰성 높이기!"