성균관대학교
SUNG KYUN KWAN UNIVERSITY

합천다목적댐 운영분석 및 유입량 예측

Data Analysis Report

23-1 학기 데이터사이언스캡스톤프로젝트

밤밭조

2019311621 송재현

2019312564 노최유하

2019310847 송예진

2019313321 신다은

Table of Contents

1. Data Observation

1.1. 합천다목적댐 운영정보

본 데이터는 K-Water 에서 관리하는 합천다목적댐의 운영정보로, API 를 통해 제공하는 raw data 이다. 수집한 기간은 2000.01.01.부터 2023.04.03.까지이고, 각 데이터는 1 시간 간격으로 이루어져 있다. 유입량 정보와 관련된 댐의 운영정보를 얻기 위해 활용하였다. K-Water 와의 논의를 통해 기존 유입량 예측값인 '유입량'과, 댐 운영과 밀접한 연관이 있는 '방류량', '유입량', '발전량'에 대해 분석해보았다. 원본은 총 23 개의 feature 로 제공되고, 그 중 '발전량(실적)', '발전량(계획)', '발전량(계획대비)', '전일유입량', '저수위(전년)', '저수량(전년)', '저수위(현재)', '저수량(현재)', '현재저수율', '전일방류량(본댐)', '시간'을 활용하였다.

1.2. 합천군 기상정보

본 데이터는 기상청에서 제공하는 합천군의 종관기상관측자료이다. 수집한 기간은 2000.01.01 부터 2023.04.03.까지이고, 각 데이터는 1 시간 간격으로 이루어져 있다. K-Water 와의 논의를 통해 유입량 예측에 제일 필수적인 요소인 '강수량', 부가적인 날씨요소로 활용될 수 있는 '기온'과 '습도'에 대해 분석해보았다. 원본은 총 36 개의 feature 로 제공되고, 그 중 '일시', '기온(° C)', '강수량(mm)', '습도(%)'를 활용하였다.

1.3. 태양고도정보

본 데이터는 한국천문연구원에서 제공하는 대구광역시 태양고도정보이다. 기간은 2016.01.01 부터 2022.12.31.이고, 각 데이터는 1 일 간격으로 이루어져 있다. K-Water 와의 논의를 통해 계절적인 요소가 유입량의 변화와 연관이 있다는 것을 확인했고, 계절적인 요소로 '남중고도'를 활용하고자 한다. 원본은 총 10 개의 feature 로 제공되고, 그 중 '날짜', '남중고도'를 활용하였다.

2. Imputation

2.1. 합천다목적댐 운영정보

모든 데이터는 수치형 데이터였으므로 그 중 연속형 데이터인 실시간 관측 데이터의 결측치는 선형 보간법을 사용하여 대체하였고, 그 외의 결측치는 이전 행의 값으로 대체하였다.

2.2. 합천군 기상정보

원본 변수에는 범주형 데이터와 연속형(수치형) 데이터가 존재하므로, 결측치를 처리할 때 범주형 데이터와 수치형 데이터를 따로 처리하였다. 수치형 데이터는 결측치가 하나 이상 존재하는 행의 앞 뒤 값에 대한 중간값으로 결측치를 대체하였으며, 범주형 데이터는 결측치가 하나 이상 존재하는 행의 앞의 값으로 결측치를 대체하였다.

3. Feature Engineering

3.1. Datetime Indexing

이후 데이터 분석 및 모델에의 적용의 편의성을 위해 시간 열을 datetime 인덱스로 사용하였다.

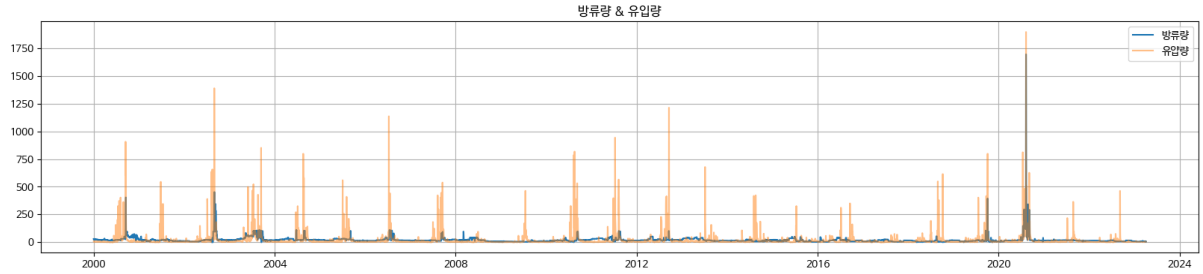
3.2. Daily Data

시간별 데이터를 사용해 모델을 학습시키는 것은 어려울 것이므로 일별 데이터를 사용할 것을 추천한다는 한국수자원공사의 조언을 듣고, 1 시간 간격으로 이루어져 있던 데이터를 1 일 간격으로 평균값을 사용하여 변환하였다.

4. Exploratory Data Analysis

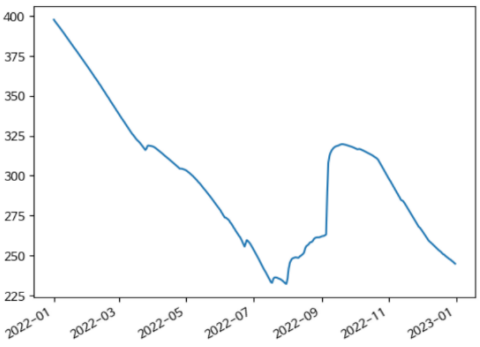
4.1. 합천다목적댐 운영정보

4.1.1. 유입량과 방류량

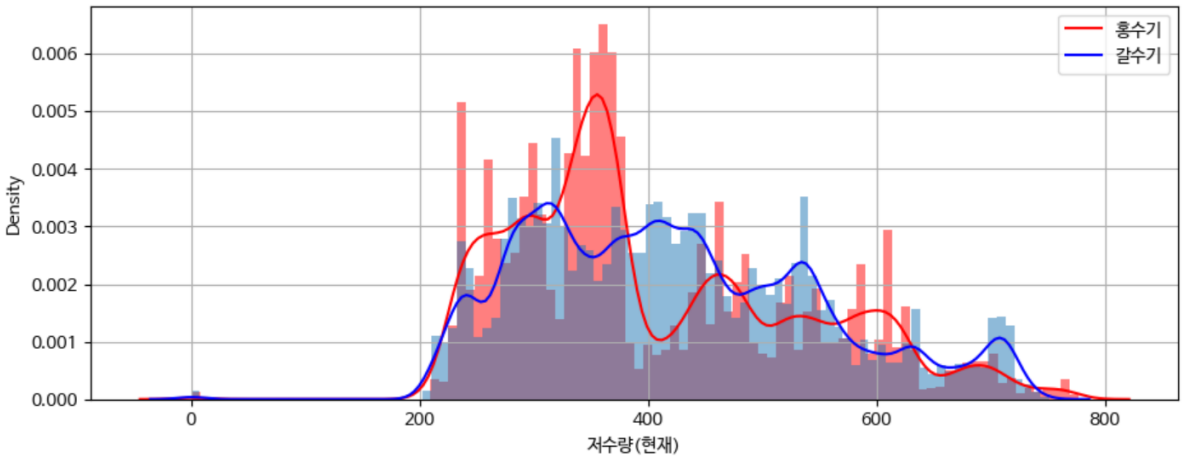


일반적으로 홍수기(6/21~9/20)에 강우량이 늘어 댐 유입량이 늘어나면 그에 따라 방류량이 늘어나는 것을 확인할 수 있다. 2020 년에는 이전과 달리 유입량이 증가함에 따라 방류량도 같은 수준으로 증가했음을 확인할 수 있는데, 댐의 저수위가 홍수위를 초과할 것으로 예상되어 급히 방류를 해야 했던 시점이었으며, 이는 합천댐하류 지역 홍수로 이어져 수많은 피해를 낳았다.

4.1.2. 저수량

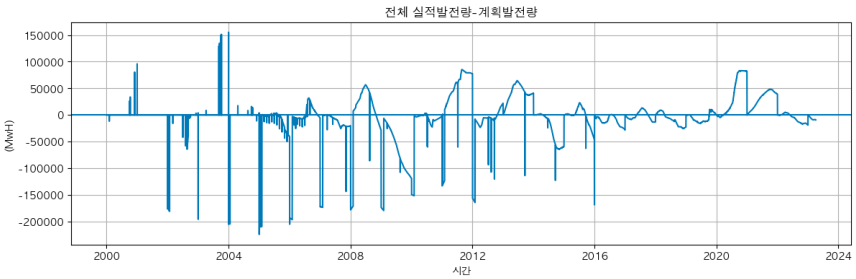


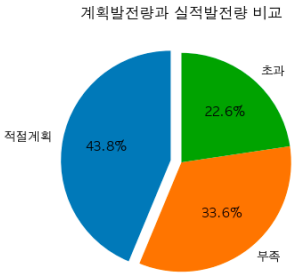
홍수기엔 여름철 호우에 대비하여 저수량이 상대적으로 적었다가, 갈수기에 홍수기 영향 및 용수 확보를 위해 저수량이 증가할 것이라는 가정을 세우고 EDA 를 진행하였다. 가정 확인을 위해 먼저 월별 저수량을 시각화한 결과 왼쪽 그래프와 같았다. 가정과 동일하게 홍수기에 저수량이 적었다가 갈수기에 점차 증가함을 확인하였다.



또한, 홍수기와 갈수기의 저수량 비교를 위해 위와 같이 히스토그램을 overlay 하여 시각화 하였다. 홍수기 저수량이 가장 높은 분포를 보이는 구간은 300 후반대로, 갈수기 저수량의 주요 분포 구간보다 낮음을 확인하였다. 위의 그래프 역시 홍수기에는 갈수기보다 저수량이 적을 것이라는 가정에 부합하는 결과이다.

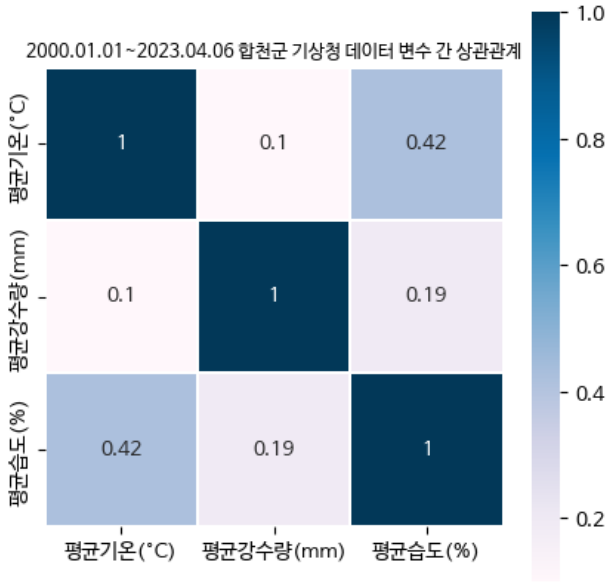
4.1.3. 발전량





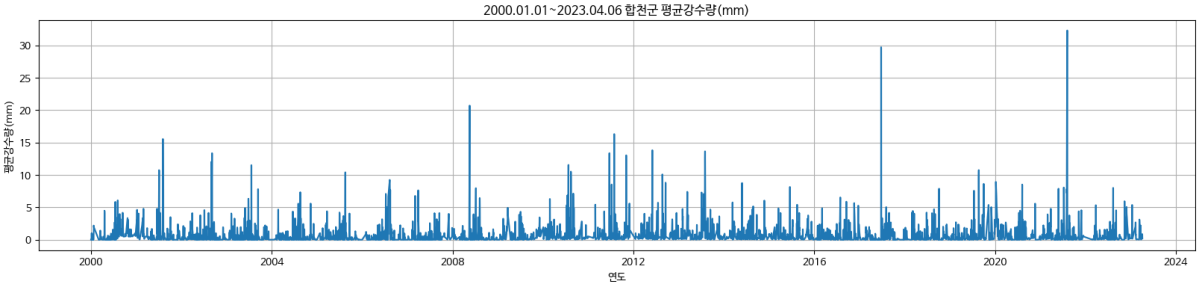
발전량은 댐에서 방류 시, 발전기를 돌려 생산하는 발전량이다. 하지만 실적량과 계획량의 차이를 확인해본 결과, 위와 같이 많은 차이가 존재하는 것을 알 수 있다. 관측기간의 56.2%가 적절한 계획을 하지 못하는 것을 확인할 수 있다. 유입량 예측의 성능이 좋아진다면, 방류량을 적절히 조절해 발전량의 부족/초과 없이 운영될 수 있을 것이라 예상한다.

4.2. 합천군 기상정보



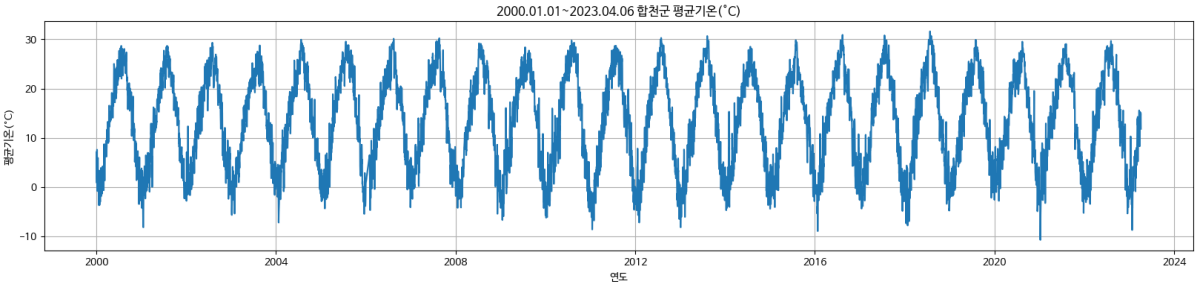
기상 변수로 사용할 일일 평균기온, 평균 강수량, 그리고 평균습도가 서로 상관관계가 있는지, 다중공선성 문제가 발생할 가능성이 있는지 분석하기 위해 기상청 데이터의 세 변수에 대한 상관관계를 계산하여 그래프로 그렸다. 대각선 요소는 같은 변수와의 상관관계를 나타낸 것이며, 세 변수 간 모든 상관관계 값을 분석하였을 때, 0.5 이상의 경우는 존재하지 않아 그대로 사용해도 될 것이라고 판단하였다.

4.2.1. 강수량



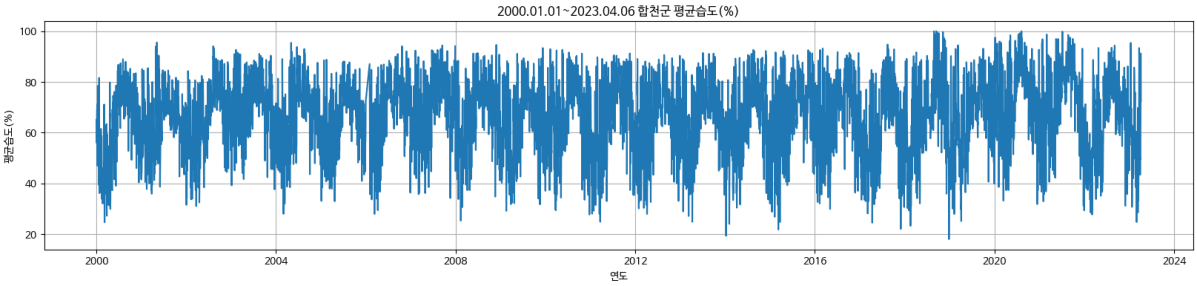
기후와의 연관성을 파악하기 위해, 2000 년부터 2023 년 4 월 6 일까지의 일일 평균 강수량을 그래프로 시각화 하였다. 간혹 강수량이 많아지는 시기를 제외하고 일정한 패턴을 보였다.

4.2.2. 기온



또한 2000 년부터 2023 년 4 월 6 일까지 합천군에서의 일일 평균기온 역시 그래프로 시각화 하였다. 평균기온은 특이사항 없이 매년 일정한 패턴을 보였다.

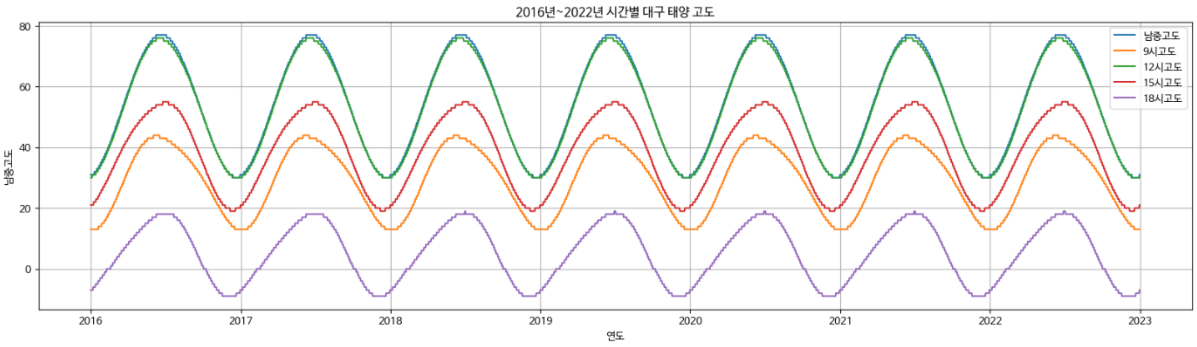
4.2.3. 습도



마지막으로, 2000 년 2023 년 4 월 6 일까지의 일일 평균 습도를 그래프로 시각화 하였다. 평균 습도 역시 일정한 패턴을 보였다.

4.3. 태양고도정보

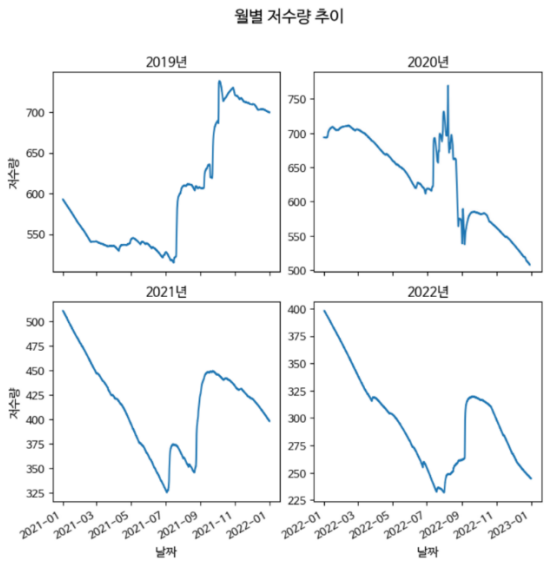
4.3.1. 고도



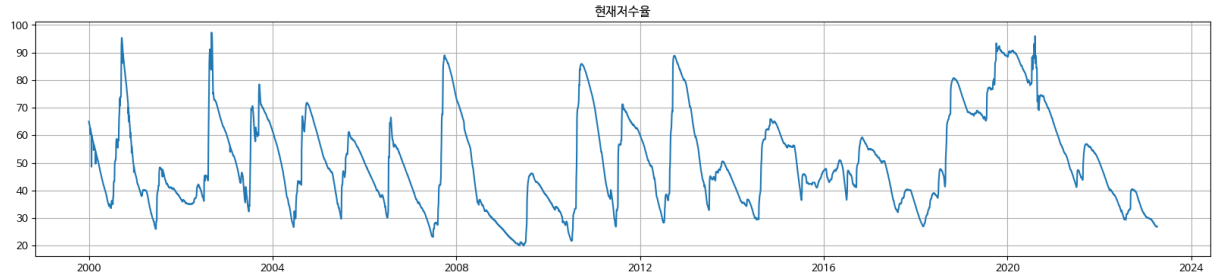
계절적인 요인 중에서는 태양 남중고도를 변수로 사용하기로 결정하였다. 합천군에서 가까운 대구 지역은 2016 년도 데이터부터 존재하였다. 따라서 년도 별로 태양 고도 값이 크게 다르지 않으면 가장 최신 년도인 2022 년 값만 추출하여 사용하기로 하였다. 그래프로 시각화 한 결과, 각 년도 별 태양고도 정보에는 이상치 없이 일정한 정제된 데이터임을 확인했다.

4.4. 특이사항

4.4.1. 2020 년 태풍 바비



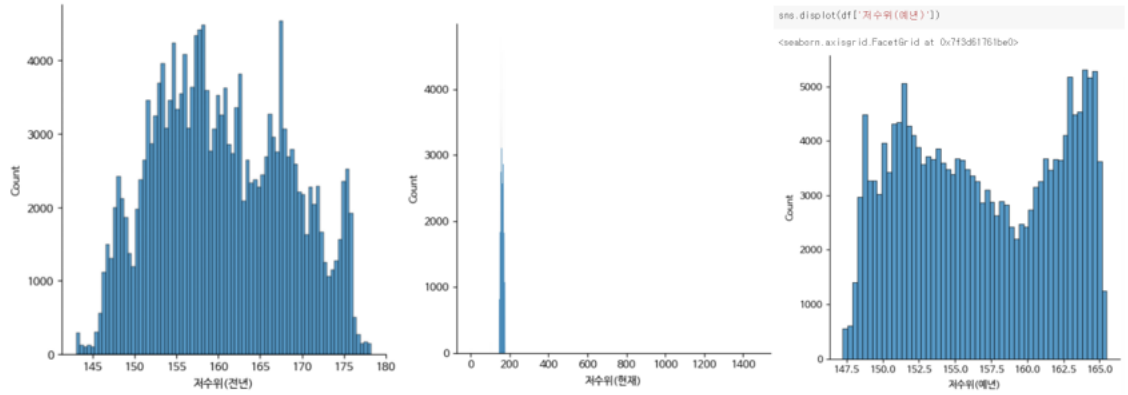
최근 4개년의 월별 저수량 추이를 시각화 하였을 때, 일반적으로 홍수기를 대비하여 6~9월에 저수량이 감소했다가 갈수기가 되어 다시 증가하는 양상을 보이나, 그래프 우측 상단의 2020년도만 다른 추이를 보였다. 홍수기 이후 갈수기를 대비하여 증가해야 할 저수량이 2020년도에만 급격히 감소한 점에 대해 수자원공사에 문의한 결과, 태풍 '바비'의 예상 세력과 실제 세력 간 차이로 인한 것이라는 답변을 받았다. 서해상에서 중심 기압이 950hPa까지 떨어지는 등 강력한 위력을 지닐 것으로 예측되면서 방류량을 늘렸으나, 태풍의 중심이 육지에서 멀리 떨어졌고 빨리 약화가 진행된 탓에 저수량을 다시 확보하지 못했음을 확인하였다. 이러한 요인으로 인해 2020년도에 이상 추이가 나타났음을 파악하였다.



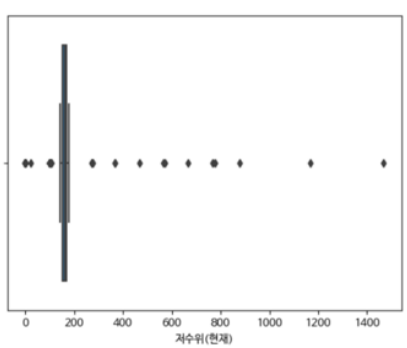
20 년간의 저수율을 그래프로 나타낸 결과는 위와 같다. 2019 년에 예년과 달리 저수율이 충분히 감소하지 않았음을 확인할 수 있으며 이는 2020 년 저수율이 치솟는 부분으로 이어진다. 따라서 저수율이 충분히 줄지 않은 상태에서 홍수기를 맞아 강우량을 수용하지 못하여 홍수가 발생한 것으로 생각된다.

5. Outlier Handling

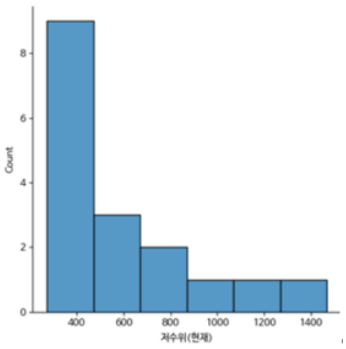
5.1. 저수위 이상치



저수위(전년), 저수위(현재), 저수위(예년)의 분포를 확인해보기 위해 각 변수를 히스토그램으로 시각화했을 때, '저수위(현재)' 변수에만 이상치가 존재함을 확인하였다.



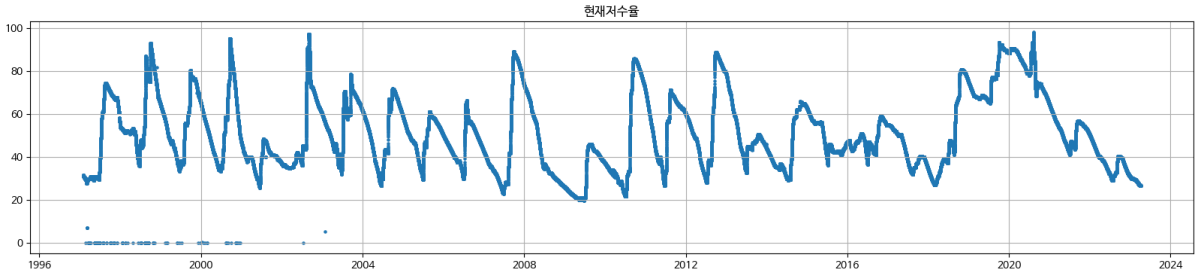

```
df['저수위(현재)'].describe()

count    184969.000000
mean        160.487997
std           9.719948
min           0.000000
25%         154.580000
50%         160.020000
75%         166.440000
max        1468.080000
Name: 저수위(현재), dtype: float64
```



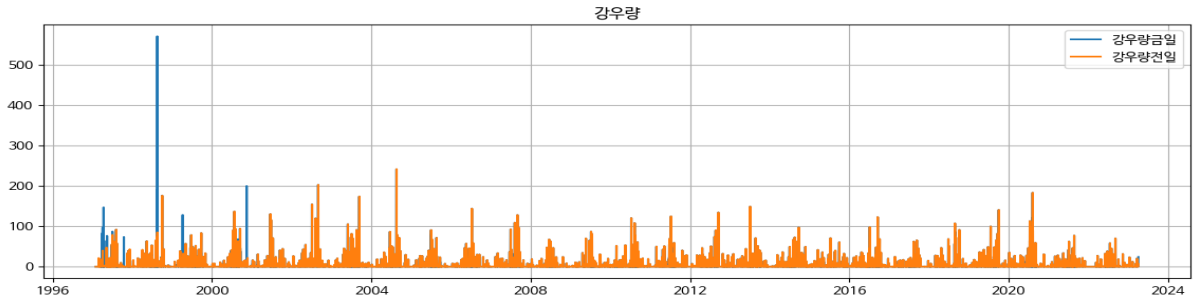
이상치 판단 기준인 Q3+1.5*IQR을 초과하는 값들만 뽑아내어 시각화 해본 결과 우측 그래프와 같이 약 20 개의 이상치가 확인되었고, 1000을 넘어가는 매우 높은 값들도 존재하였다. 이에 수자원공사에 방문하여 문의한 결과, 저수위는 계획홍수위를 넘을 수 없으며, 계획홍수위를 넘는 경우 측정 오류로 간주하면 된다는 답변을 받았다. 이에 이상치를 보이는 기간 앞뒤의 값으로 보간하는 방식을 사용하여 이상치를 처리하기로 결정하였다.

5.2. 과거 이상치

2000 년 이전의 데이터에서 이상치가 자주 발견되어 한국수자원공사에 자문을 구하였고, 1997 년부터 3 년의 데이터를 사용하지 않는다 해서 학습 결과에 큰 영향을 미치지 않을 것 같다는 답변을 얻어 2000 년 이전의 데이터를 삭제하기로 결정하였다. 발견되었던 대표적인 이상치는 아래와 같다.



위의 산점도에서 2000 년 이전에서 0 값이 자주 나타나는 것을 확인할 수 있다. 저수율은 연속형 데이터이므로 중간에 나타나는 0 값은 이상치로 간주된다.



위는 금일 강우량과 전일 강우량 데이터를 함께 나타낸 그래프이다. 전일 강우량은 금일 강우량의 최종값(23 시 누계강우량)과 같으므로 그래프로 나타냈을 때 같은 양상을 보여야 한다. 2000 년 이전에 계속해서 그래프가 일치하지 않는 부분을 확인할 수 있다.