

공공데이터 품질 진단

보고서 전체적인 흐름

데이터, 진단 도구, 진단도구 이용, 진단 도구를 이용해서 찾은 오류들, 진단 도구를 이용해서 얻은 보고서들, 진단 도구를 이용해서 찾은 오류들 수정, 수정 과정, 수정 완료, 요류 수정된 데이터 다시 진단 후 오류 없음 확인, 데이터 값 확인, 완료된 데이터를 목록등록관리 시스 템에 등록, 목록등록관리 시스템 등록 절차

위 내용으로만 써도 분량 충분히 나올거 같음..:)

0824

데이터는 csv 파일로 제공받음

본인이 맡은 데이터 범위 : 분야별 재정지출 / 국토 및 지역개발 분야 ~ 분야별 재정지출 / 환경 분야 까지의 데이터를 담당데이터 파일 대략적인 스크린샷

담당할 데이터 파일 강조

공공데이터포털(data.go.kr)의 개방데이터(파일) 진단 정비 도구를 이용

다운 링크

https://www.data.go.kr/bbs/rcr/selectRecsroom.do?

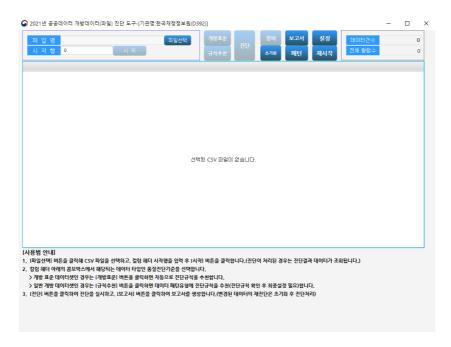
 $\underline{pageIndex=1\&originId=PDS_000000000755+++\&atchFileId=FILE_000000002414717\&searchCondition3=\&searchCondition2=\\ \underline{post=2\&searchKeyword1=\&Laword=PDTY01\&Edc=PDTY02\&Bigdata=PDTY03\&Std=PDTY04\&News=PDTY05\&Etc=PDTY06}$

개방데이터(파일) 진단 정비 도구를 다운받고 설치하는 과정

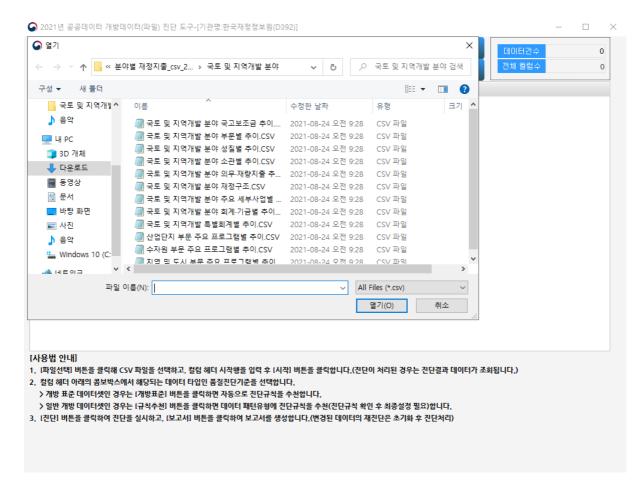
추가 예정

설명서 참고

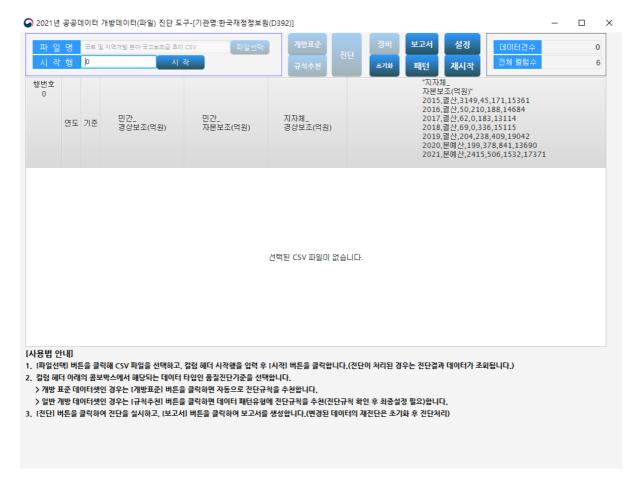
공공데이터 개방데이터(파일) 진단 도구



파일 선택 클릭



품질 진단 대상 데이터 파일 열기

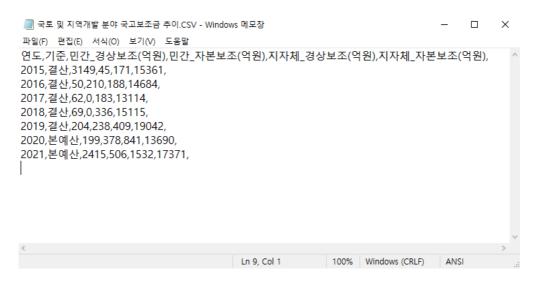


데이터가 올바르게 들어오지 않는 모습

csv 파일 자체에 대한 형식 오류 발견

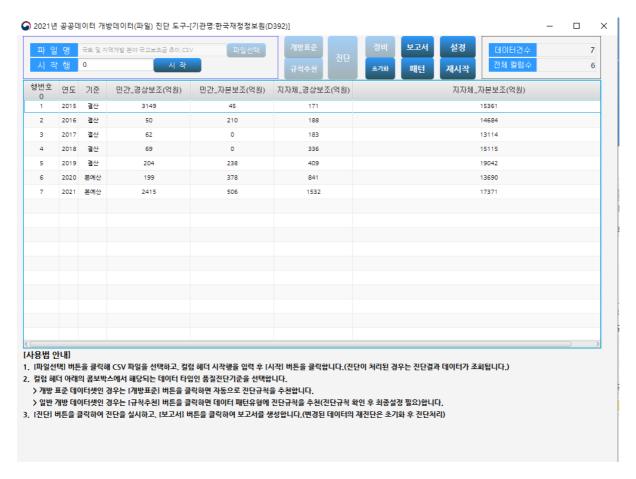


메모장에서 csv 파일 형식에 맞게 수정

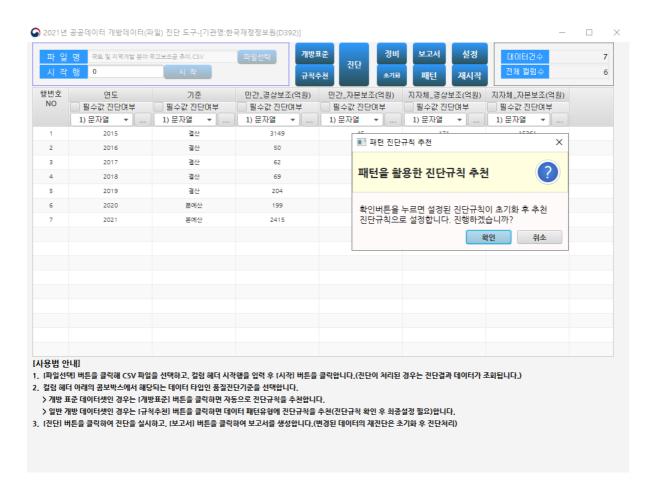


수정 완료

초기화 버튼을 눌러서 수정된 csv 파일을 가져옴



시작 버튼을 누름



규칙추천 버튼을 통해 추천 진단 규칙을 받아볼 수 있음



[사용법 안내]

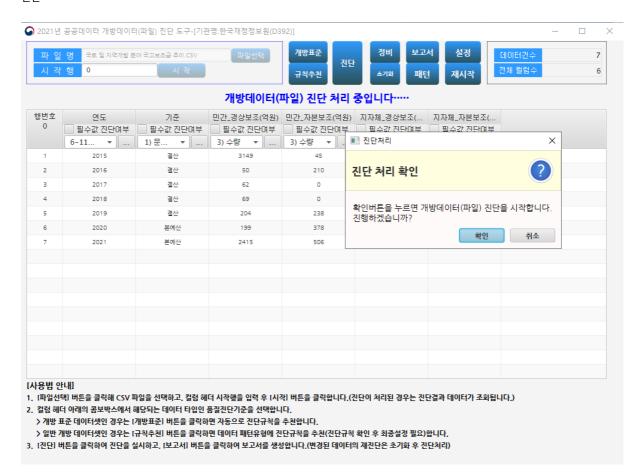
- 1. [파일선택] 버튼을 클릭해 CSV 파일을 선택하고, 컬럼 헤더 시작행을 입력 후 [시작] 버튼을 클릭합니다.(진단이 처리된 경우는 진단결과 데이터가 조회됩니다.)
- 2. 컬럼 헤더 아래의 콤보박스에서 해당되는 데이터 타입인 품질진단기준을 선택합니다.
 - > 개방 표준 데이터셋인 경우는 [개방표준] 버튼을 클릭하면 자동으로 진단규칙을 추천합니다.
- 〉일반 개방 데이터셋인 경우는 [규칙추천] 버튼을 클릭하면 데이터 패턴유형에 진단규칙을 추천(진단규칙 확인 후 최종설정 필요)합니다.
- 3, [진단] 버튼을 클릭하여 진단을 실시하고, [보고서] 버튼을 클릭하여 보고서를 생성합니다.(변경된 데이터의 재진단은 초기화 후 진단처리)



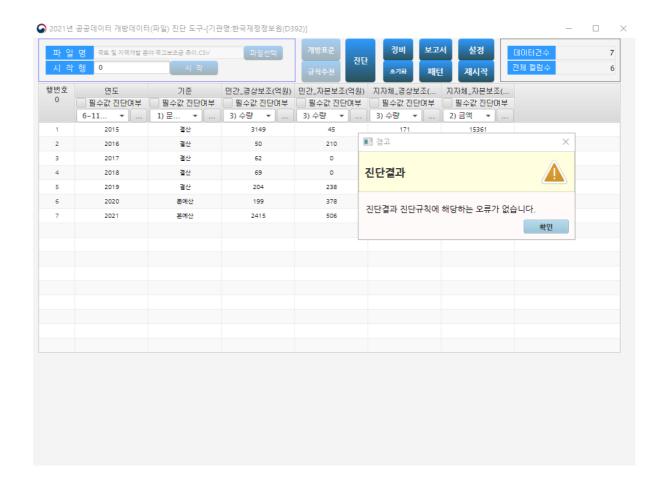
[사용법 안내]

- 1. [파일선택] 버튼을 클릭해 CSV 파일을 선택하고, 컬럼 헤더 시작행을 입력 후 [시작] 버튼을 클릭합니다.(진단이 처리된 경우는 진단결과 데이터가 조회됩니다.)
- 2. 컬럼 헤더 아래의 콤보박스에서 해당되는 데이터 타입인 품질진단기준을 선택합니다.
 - > 개방 표준 데이터셋인 경우는 [개방표준] 버튼을 클릭하면 자동으로 진단규칙을 추천합니다.
 - 〉일반 개방 데이터셋인 경우는 [규칙추천] 버튼을 클릭하면 데이터 패턴유형에 진단규칙을 추천(진단규칙 확인 후 최종설정 필요)합니다.
- 3. [진단] 버튼을 클릭하여 진단을 실시하고, [보고서] 버튼을 클릭하여 보고서를 생성합니다.(변경된 데이터의 재진단은 초기화 후 진단처리)

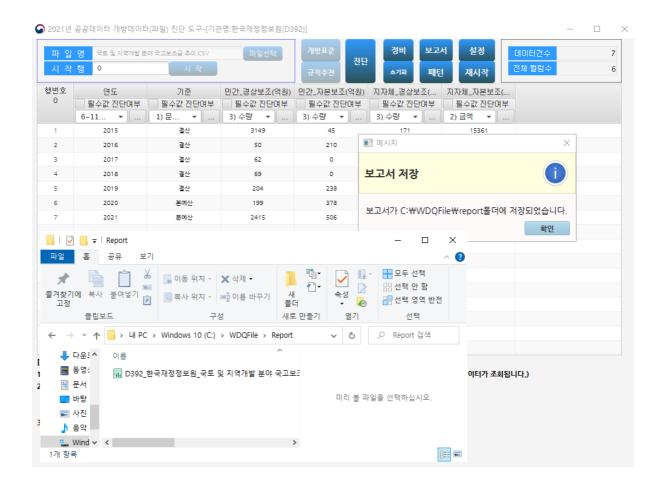
진단



진단 결과



보고서 저장



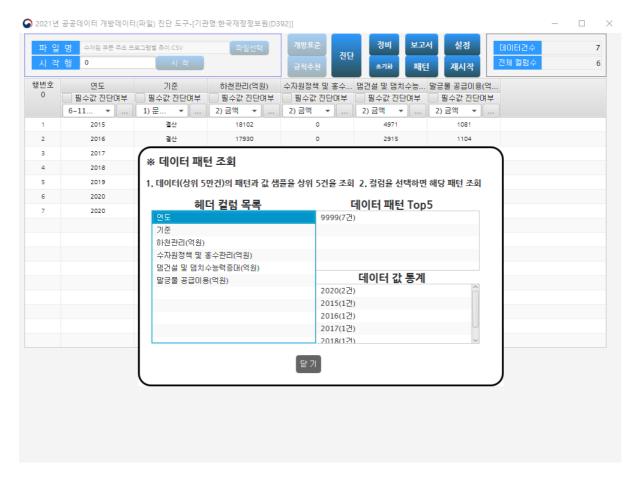
보고서 내용



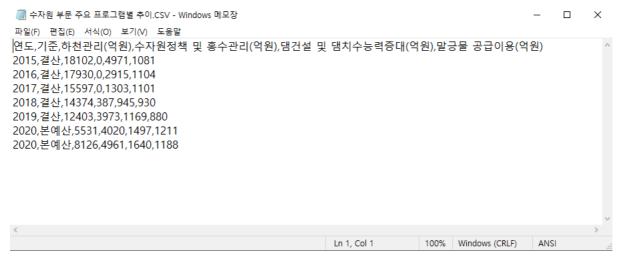
다른 예시



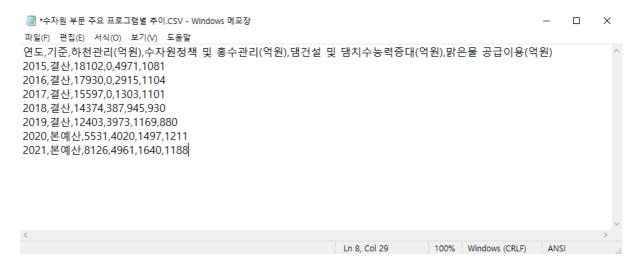
진단 결과에는 오류가 없음. 하지만 데이터 패턴을 확인하는 과정에서 2020년이 2건임을 알아냄



또한, 데이터 칼럼 명에도 오타가 있음을 확인



오류 수정 완료

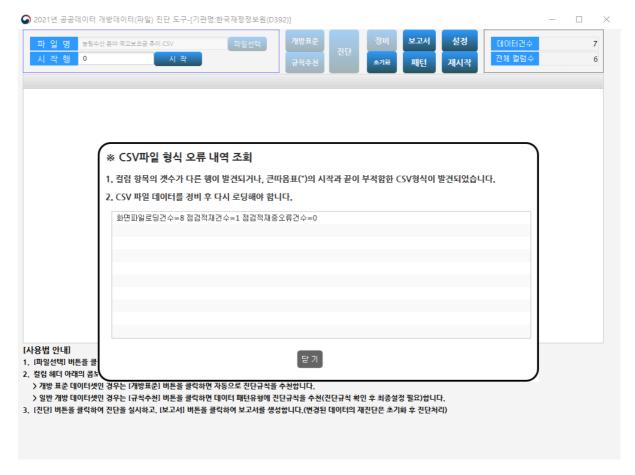


진단도구를 통해 다시 확인

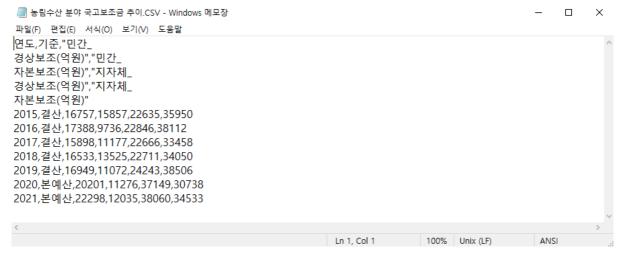


다른 예시

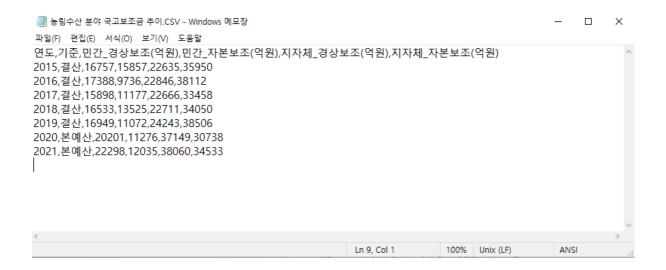
이번 파일은 csv 파일 형식 오류가 발견되어서 파일이 열리지 않음



메모장으로 csv 파일 형식에 맞게 수정해줌



수정 완료



<목록 등록 관리 시스템>

진단 도구로 오류를 모두 수정한 파일들을 공공데이터 목록등록관리 시스템에 등록

데이터 담당자님이 변경 요청하신 건에 맞는 데이터를 등록하고 데이터에 맞는 설명과 기준일을 설정하여 등록

키워드 작성 : 주요재정통계를 재정지출로 변경하고 자동 번역

데이터 설명 작성 : 데이터에 맞게 연도를 변경하고 자동 번역

파일 업로드

데이터 기준일 설정 : 2021-04-21

등록

0824 등록 완료된 파일

- 분야별 재정지출 / 국토 및 지역개발 분야
- 분야별 재정지출 / 농림수산 분야 (14개 파일 중 11개 완료)
- ▼ 보고서 파일

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/96e56c15-5b65-43a4-8e66-dc1e75fea8c0/Report.zip

0825

0826

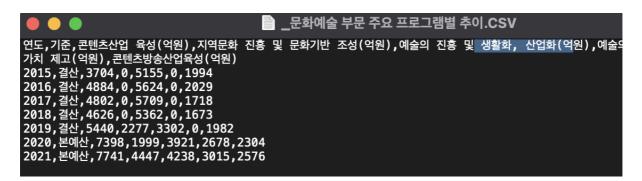
문화 및 관광 분야에서 데이터 2개 미완료 사유



■ 문화 및 관광 분야 주요 세부사업 현황.CS

연도,기준,관광산업 융자지원(융자)(억원),문화재보수정비(총액사업,보조)(억원),국민체육센터 건립지원(억원 자율)(억원),문화예술향유 지원(억원),주최단체지원(억원),문화예술교육 활성화(억원),스포츠산업 금융지원(2020,본예산,5500,4200,2553,-,2309,1433,1568,1293,762,1130 2021,본예산,5990,3970,2694,2581,1853,1671,1510,1319,1192,1148

칼럼명에 ,(쉼표)가 들어가 있어서 csv는 ,(쉼표)를 구분자로 데이터를 읽기때문에 ,(쉼표)를 기준으로 다른 칼럼이라고 인식 따라서 칼럼 안의 쉼표를 다른 문자로 입력해줄 필요 있음 → 이때 공통의 기준을 정해야하지 않을까라는 생각으로 아직 등록하지 않음. 즉, ,(쉼표) 대신 어떤 문자를 써야하는지 연구원님께 물어보고 경민이랑 상의하고 진행할 생각.



이 데이터도 같은 경우

- 이 경우에 해당하는 데이터
- 문화 및 관광 분야 > 문화 및 관광 분야 주요 세부사업 현황.csv
- 문화 및 관광 분야 > 문화예술 부문 주요 프로그램별 추이.csv

현재 파일명 앞에 '_' 를 붙여주어서 미완료된 파일이라는 것을 표시한 상태 등록 전에 ' ' 지우고 등록 예정

0827 인수인계

먼저, 경민아 같이 나가자고 했는데 먼저 나가서 미안해...(한달 조금 넘는 짧은 시간동안 여기 회사, 연구원님들, 너까지 다 좋은 곳이고 좋은 사람들이어서 금방 정들었나봐. 그래서 그런지 뭔가 기쁜 마음보다 아쉬운 마음이 큰거 같아. 그리고 가야할 맛집들이 많았는데.. 같이 가고 싶었는데 못가서 아쉽다ㅠ 너가 가고 싶다고 해서 간 집들 전부 맛있는 곳이어서 앞으로 갈 곳들도 기대가 됐었거든.. (맛있는게 짱이 ㅑ..) 그리고.. 너가 준비하는 것들 너가 하고 싶은 것들 준비 잘해서 꼭 이루길 바래. 항상 너 보면 열심히 공부하고 준비하고 있어서 걱정은 없지만 나는 내 주변에서 같이 공부하던 사람들이 하나 둘씩 자기 자리를 찾아서 떠날 때 많이 불안했었거든.. 그래서 혹시나 내가 먼저 나가서 너한테 내가 전에 받은 감정들을 주지 않았을까 하는 마음이 조금 있긴해.. 보 너가 더 잘 알겠지만 결국 그런 마음들 없어지게하는건 더 열심히 사는거 밖에 없더라구.. 너가 준비하는 것들 모두 잘 되길 응원할게. 너 실력도 있고, 좋은 사람이니까 자신감 가지고 좋은 소식 생기면 말해줘! 기다리고 있을게. 종종 얼굴보자:) 언제 한번 퇴근하고 저녁 먹자구~

파일 현황

데이터 품질 진단 및 수정 후 목록등록관리에 등록까지

분야별 재정지출_csv_2021_0802_수정 : 데이터 파일

국토 및 지역개발 분야, 농림수산 분야는 등록까지는 완료 했는데, 진단하고 오류 수정하고 난 데이터 파일을 남겨두질 못했어 문화 및 관광 분야, 보건 분야는 진단하고 오류 수정한 데이터야, 근데 문화 및 관광 분야 파일 중에서 2개 파일을 오류 수정하고 등 로해야해

고치지 않은 이유랑 오류 내용은 0826에서 확인하면 될거야. (0826 링크도 걸어뒀어)

진단보고서

국토 및 지역개발 분야부터 보건 분야까지 진단한 보고서 모두 모아뒀어.

위에서 언급했던 문화 및 관광분야에서 미완료한 2개는 진단보고서 없어!

캡처사진

이건 내가 보고서 쓸때 오류 사진이랑 오류 해결 사진이랑 해서 보고서 작성할때 쓰려고 했는데 사진 파일 이름을 그 상황에 대한 내용으로 만들어둬서 혹시 사용하게되면 파일 이름보면서 이해하면 도움이 될거야