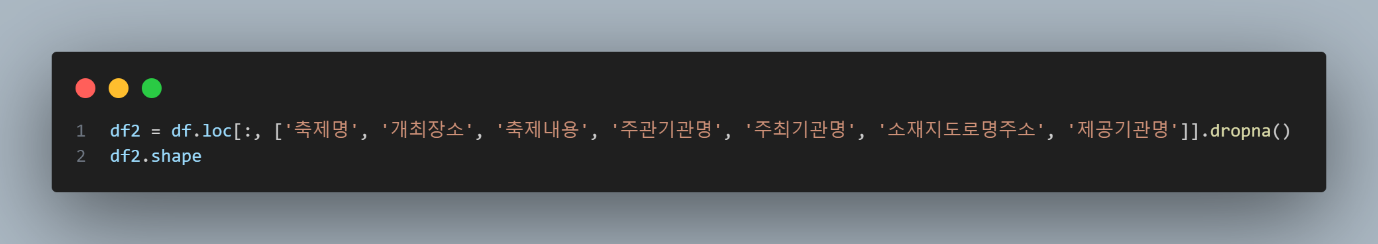
[ 코드 설명 및 구현 ]



데이터프레임 클래스 사용을 위한 pandas 라이브러리와 시각화를 위한 matplotlib 라이브러리를 임포트하고, 그래프 출력 시 한글이 깨지는 현상을 처리하기위해 터미널 명령어를 사용하여 폰트를 설치하는 코드. (주석은 실행 중 이미 처리했기에 주석 처리해두었음)



데이터셋을 불러와 상위 5개의 row만 열람하는 코드. Csv 파일의 데이터들이 한글이기 때문에 마찬가지로 cp949 형식으로 인코딩 됨을 선언해줌으로써 디코딩 문제 해결.



필요없거나 결측치가 너무 많은 column 들을 제외하여 선택하고 결측치를 제거한 후에 크기를 출력해보는 코드. Column에 따라 결측치를 지우면 데이터의 수가 확연히 줄어드는 column들이 존재하여 column마다 실행해본 후 적절하게 판단하여 남겨두었음.



주최기관명을 기준으로 데이터를 그룹으로 나누어보았음. 하지만 주최 기관이 너무 많아 기준을 변경함. 주최기관명을 기준으로 데이터를 나누었을 때, 804개의 그룹으로 나뉘어짐.



마찬가지로 주관기관명을 기준으로 데이터를 그룹으로 나누었을때도 무려 946개의 그룹으로 나뉘어짐. 역시 분석이 불가능하다고 판단하여 기준 탈락.

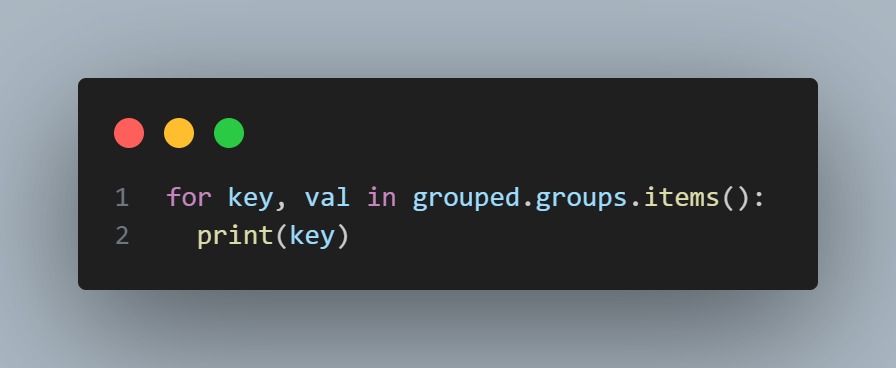
.



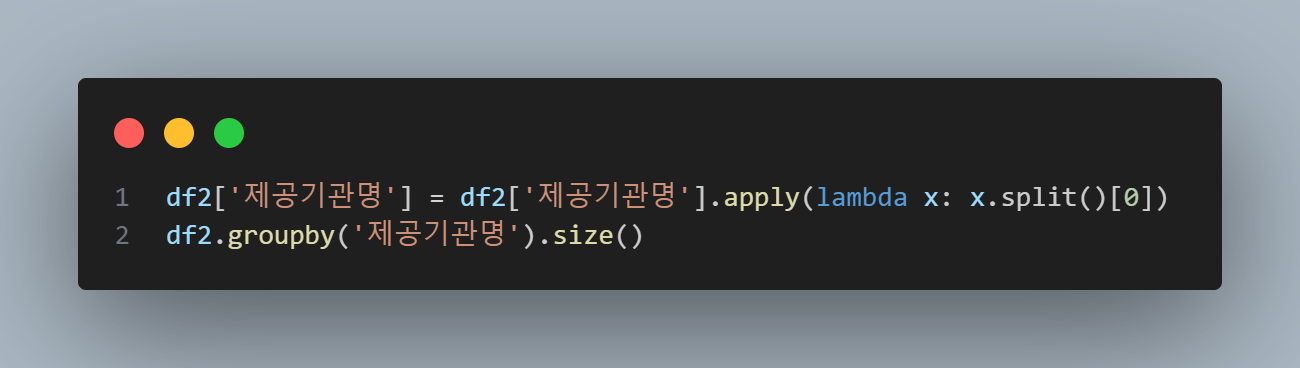
상위데이터 확인 시, 제공 기관명은 ~도 ~시 로 나뉘어져있어서 우선 시도해보았음. 약 234개의 그룹으로 앞 두개의 기준에 비해 현저히 적은 모습을 보여줌.



해당 그룹으로 나눈 데이터 들을 대략적으로 출력해보았으며, 그룹마다 몇 개의 데이터가 있는지 대략적으로 확인이 가능.



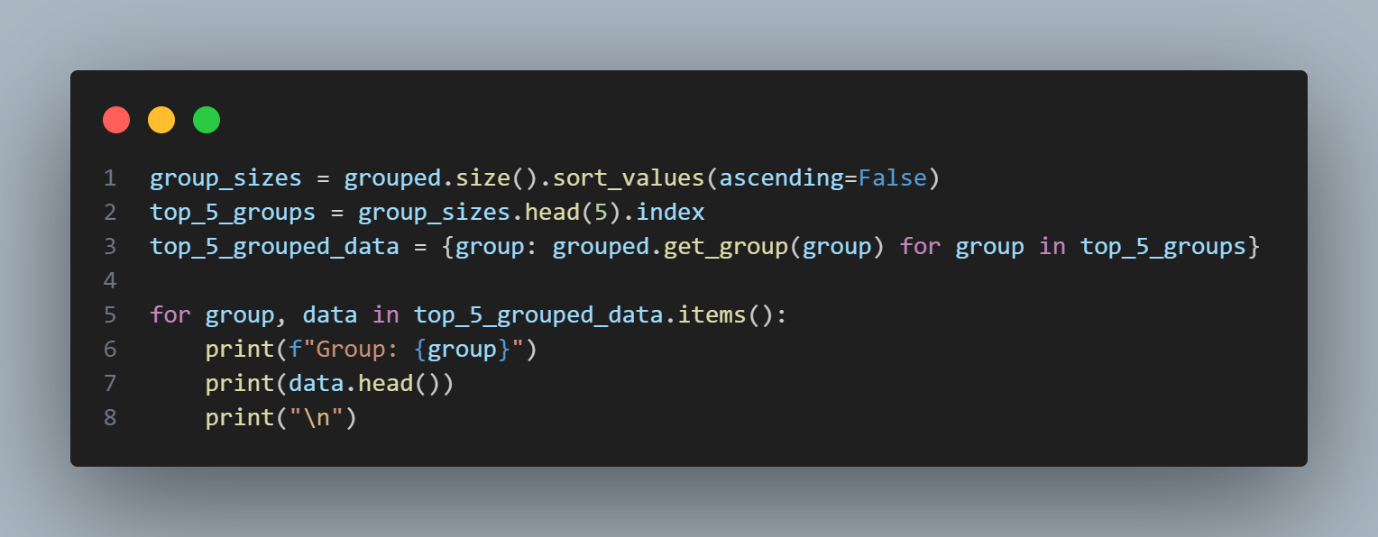
혹시나 존재하는 이상치 존재여부 확인을 위해 그룹의 모든 키값을 출력해보았으나, 이상치를 발견하지는 못하였음. 한국관광공사를 제외한 모든 데이터가 지명.



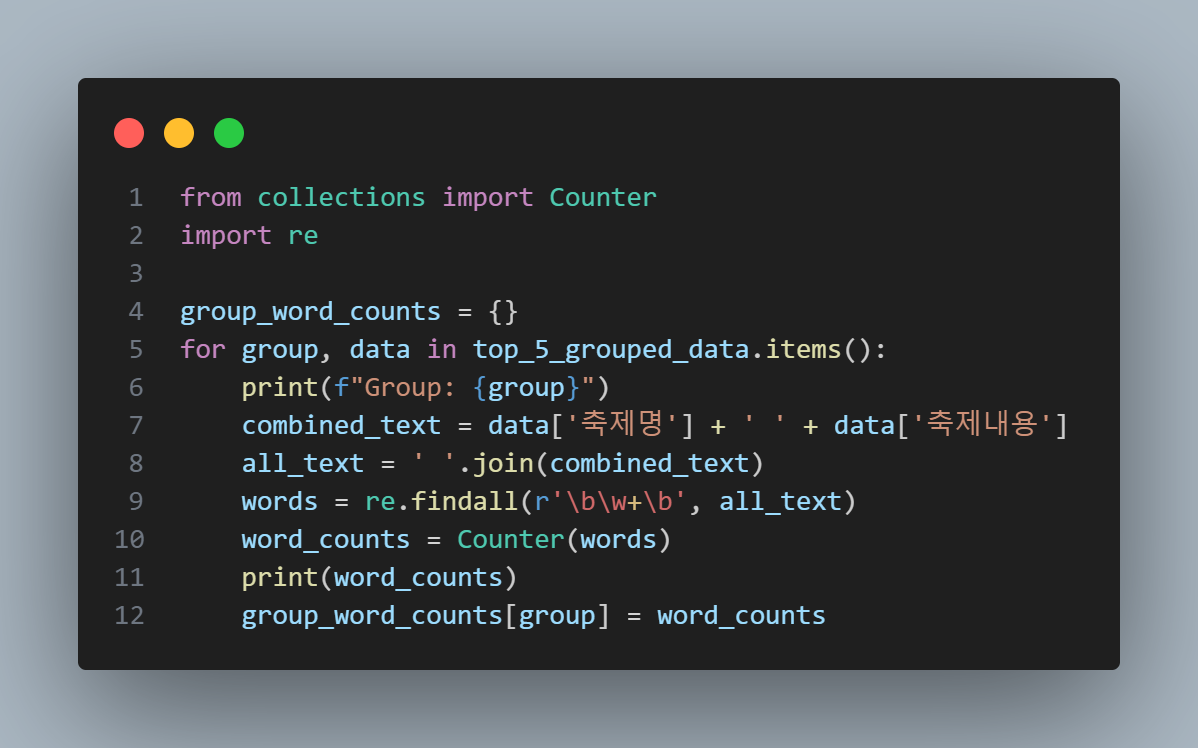
그룹을 더 줄이기 위해 제공기관명을 ~도 ~시 단위에서 ~도 단위로 모두 변경하여 각 도 및 특별 광역시에 따른 행사 횟수를 정리해보았음.



정리한 데이터를 토대로 막대그래프로 나타내서 대략적으로 어떤 도에서 많은 행사를 개최했는지 확인할 수 있음.



정리된 데이터에서 행사 수 상위 5개의 지역을 선정하여 이전과 같은 방식으로 그룹별로 출력을 진행.



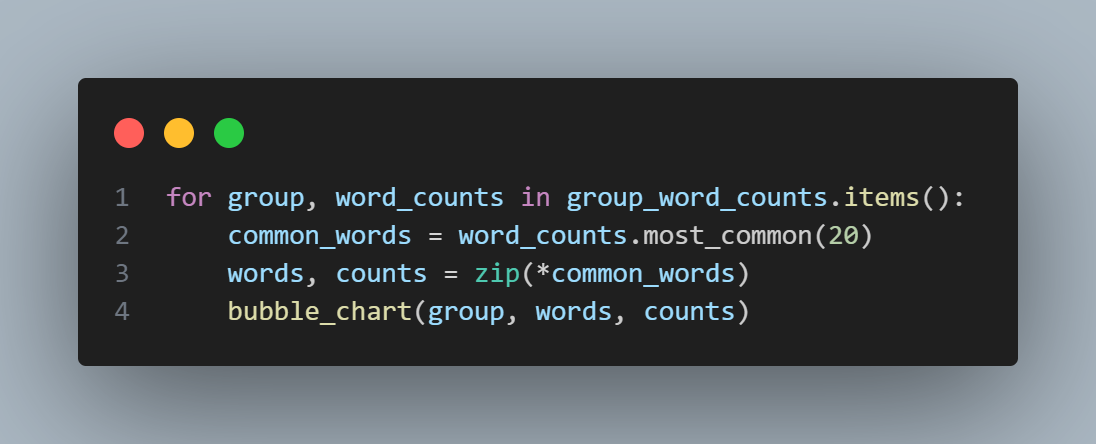
본격적으로 분석을 하기 시작. 어떤 기준으로 분석을 해야 원하는 결과를 얻을 수 있을지 고민한 끝에, 축제명과 축제내용을 토대로 분석을 진행해보고자 함. 축제명과 축제 내용에 포함된 키워드의 빈도수를 측정하여 어떤 종류의 행사와 어떤 형식의 행사가 자주 일어나며 지역에서 인기있는지 파악할 수 있을거라 판단. 파이썬 collections의 빈도를 셀 수 있는 Counter 클래스를 사용하여 그룹별 사용된문장의 각 단어 빈도를 측정하고, 정규식을 통해 각종 문장부호를 제거.



Matplotlib 라이브러리에서는 버블 차트를 제공하지 않아서 matplotlib 라이브러리 홈페이지에서 제공하는 matplotlib 라이브러리를 활용한 버블 차트 구현 클래스 코드를 가져와 사용.



구성된 버블 차트 클래스를 사용하여 하나의 버블 차트를 그려내는 함수. 마찬가지로 matplotlib 라이브러리 홈페이지에서 제공하는 예제를 토대로 수정하여 작성.



앞에서 만든 버블 차트 함수를 활용하여 각 그룹별로 20개의 최빈 단어들을 버블 차트로 출력하여 확인.

[ 데이터 분석 결과 ]

1. 강원도

춘천, 공연, 개인전, 체험, 작품, 전시, 함께 등의 키워드가 많이 사용되었으며, 강원도 지방에서는 다양한 공연과 전시, 즐기고 직접 체험할 수 있는 행사가 많이 열린 것으로 보이며, 2022의 키워드가 2021보다 더욱 많은 걸로 보아 2022년도 당시 다른 해 보다 많은 행사가 열렸을 것으로 짐작된다. 또한 춘천 키워드가 압도적으로 많은 것으로 보아, 대부분 춘천에서 행사가 열리고 있을 것으로 짐작된다.

2. 경기도

축제, 공연, 체험, 청소년, 축하공연, 구리, 안양, 코스모스, 유채꽃 등의 키워드가 많이 등장했다. 마찬가지로 공연, 체험 위주의 행사들이 많은 것으로 보이고, 청소년 키워드가 크게 나왔는데, 청소년을 위한 행사가 많이 열리고 있는 것으로 보이며, 꽃 관련 키워드가 자주 등장한 것으로 보아 자연 관련 행사들도 많이 열린 것으로 보인다. 또한 주요 지역은 안양, 구리로 추정된다.

3. 제주특별자치도

축제, 체험, 공연, 댄스, 음악, 페스티벌, 걷기, 어린이 등의 키워드가 자주 등장했다. 댄스, 음악, 페스티벌 키워드의 빈도를 보아 아마 예술과 관련된 행사들이 더러 진행된 것으로 보이고, 어린이 키워드를 보았을 때, 아이를 동반한 가족 단위의 참가자가 많아 어린이를 위한 행사들도 많이 열린 것으로 예상된다.

4. 대구광역시

전시, 개막식, 부대행사, 동성로축제, 동성로가요제, 대구화교중화문화축제, DJ페스티벌, 중구미술협회, 화랑기획전, 전시행사 등의 키워드가 등장하였고, 동성로가 행사가 이루어지는 주요 도시로 보이며, 부대행사의 키워드가 많은 것을 보았을 때, 군부대에서 개최하는 행사들이 다수 존재했던 것 같으며, 화교민들을 위한 행사도 존재하고, 미술과 관련된 행사가 많이 열리는 것으로 예상된다.

5. 서울특별시

축제, 공연, 체험부스, 체험, 전시, 프로그램, 도봉, 주민, 부대행사, 발표 등 키워드가 자주 등장했다. 키워드를 살펴보았을 때, 행사 개최 수는 5위이나 뚜렷하게 눈이 가는 키워드는 없었다. 다양한 종류의 행사가 골고루 열린다고 생각될 수 있으며, 2023과 2024 키워드도 등장한걸로 보아 현재까지도 다양한 행사를 개최하고 있는 것으로 보인다.

6. 종합

지역별로 중복되는 키워드는 축제, 공연, 체험, 전시 등이 있었으며, 지역에 따라 미술, 음악, 공연, 자연 등 차이를 약간씩 보이는 것을 알 수 있으며, 군부대 행사가 자주 일어나는 지역도 알 수 있었다. 또한 관사나 조사 및, 등 과 같은 단어가 높은 빈도로 등장하였고, 측정되었으며, 이러한 단어들을 제외한다면 더 다양하고 자세한 분석이 가능할 것으로 보인다.