**최신정보기술**

**8조 최종 보고서**

**-최신정보기술 소논문-**

**컴퓨터학부**

**8조**

**2015100962최예준, 2014335054이종한, 2016100909 이승현**

**-소논문 주제-**

**빅데이터 기반 2019서울시 미세먼지 원인 상관관계 분석 연구**

**-목차-**

**소감**

**서론**

1. 개요  
   1.1 주제 선정 목적 및 이유

1.2 사용 된 기술

**본론**

1. **데이터 분석 과정**  
   2.1 크롤링  
   2.2 워드클라우드   
   2.3 서울시 데이터 분석 및 시각화  
   2.3.1 데이터 수집  
   2.3.2 데이터 처리
   * 1. 데이터 시각화
        1. GoogleVis
        2. Ggmap
        3. 2015~2019 서울시 지역 미세먼지 plot시각화

**제3장 서울시 미세먼지 요인 데이터 분석**

3.1 2018 서울시 사업체 현황

3.2 2019 서울시 대기오염 물질 수치

**결론**

**참조**

**역할, 기여도**

프로젝트 전체 진행과 논문 작성 및 데이터 수집 R프로그래밍 분석,시각화 등 모두 균등하게 다같이 진행하였습니다.

**소감문**

최예준

미세먼지 원인, 지역적 차이, 대기오염 물질과의 관계 등 미세먼지에 대한 전반적인 지식에 대해 습득 할 수 있는 기회가 되었습니다.

미세먼지라는 주제를 선정함에 있어 작년을 기준으로 가장 사회적으로 이슈가 활발 했던 것을 선택하여 자료를 수집하는데 있어서 막힘이 없는 점이 장점이 였습니다.

그 덕에 2019년 일별 미세먼지 수치와 2015~2019년 년 별 평균 수치의 데이터 셋을 구할 수 있었습니다. 하지만 수집한 데이터의 수치의 차이점을 발견 하였습니다.

2015~2019년까지의 데이터와 2019년 일별 평균데이터에서 평균 값을 구했는데 동일한 2019년의 서울시 자치구에 대한 평균 수치가 다른 점을 발견 할 수 있었고 각 데이터 셋이 출처에 따라 데이터가 조금씩 다르다는 걸 알게 되었습니다.

같은 내용의 데이터셋이지만 데이터가 어디서 공유 되는지에 따라 다른 점이 흥미로웠습니다. 그 외 wordcloud를 진행하면서 실제 알고 있는 사실과 wordcloud를 통해 얻은 결과 값의 동일함 과 R프로그래밍을 통해 분석하고 시각화 하는 점에 대해서 빅데이터와 데이터 가공, 수집 등 데이터에 관련하여 큰 관심을 가지게 되었습니다.

그 외 대기오염 물질 수치나, 사업체, 측정소 위치 등 이론상 관계가 있다고 생각이 둔 개념들을 데이터를 분석화, 시각화 함으로서 그 개념을 증명하고 정의하는 것이 흥미로웠습니다. 이번 팀프로젝트를 논문, R프로그래밍 등 기타 여러가지 기술을 이용하여 앞으로의 지식을 습득한 것 같아 좋은 프로젝트 였던 것 같습니다.

이승현

이번 최신정보기술 수업을 통해 데이터 분석을 처음으로 해보았습니다. 빅데이터 분석 연구를 정하기 위해서 데이터 안심구역에도 방문하여 다양한 주제의 데이터셋들이 제공되는 것을 알 수 있었으며 안심구역 뿐만 아니라 공공데이터, 서울 열린 데이터 광장에서도 데이터를 쉽게 구할 수 있었습니다. 특히 저희조는 2019년도 미세먼지 데이터셋을 통해 서울시 자치구별 미세먼지 평균과 미세먼지 요인에 대한 상관관계를 분석하였습니다. R프로그래밍을 통한 데이터 분석으로 쉽게 시각화 할 수 있었기 때문에 그래프,지도,표로 어디 지역의 수치가 가장 높고 낮은지 한 눈에 인식할 수 있었습니다. 처음 저희조는 미세먼지의 주 원인은 석탄,석유, 경유차를 사용하는 사업체가 많은 곳이 가장 높다고 생각했지만 실제 분석해보니 미세먼지가 높게 측정되는 지역과 사업체의 관련성은 달랐습니다. 저희의 예상과 다른 데이터 분석 결과로 저희는 다수의 요인이 있다고 파악했으며 분석의 중요성을 느낄 수 있었습니다. 또한 개인이 아닌 팀프로젝트를 통해 함께 논문 작성법을 익히고 다양한 시각으로 데이터들을 분석할 수 있었습니다. 앞으로도 다양한 분야의 데이터들을 분석하여 연구하고 싶은 계기가 되었던 활동이었던 것 같습니다.

이종한

8조의 주제는 빅데이터 기반 2019서울시 미세먼지 원인 상관관계 분석 연구로 진행하였습니다. 조원들과 주제 선정에 대한 논의를 하던 중 많은 사람들이 관심을 가지고 있는 대기 환경 문제에 대해 주제를 선정하게 되었습니다. 안심구역에 방문하여 자문을 구하고, 관련 데이터 조사를 시작하였습니다. R프로그램을 이용하여 조원들이 제공한 추가 정보들과 분석 데이터, 시각화 된 자료를 보고, 이에 피드백 하였습니다. 팀프로젝트를 통해 최신정보기술 강의에서 배운 것이 실제 분석에서 어떻게 사용되고, 데이터가 관리되는지 가늠할 수 있었던 시간이었습니다. 코로나 사태로 대면 미팅이 다소 부담이되어 온라인 팀프로젝트로 논문 과제를 진행하는 것이 쉽지 않을거라 예상했습니다. 그러나 전체적으로 조원들의 빠른 피드백과 역학 분담의 도움이 컸고, 체계적으로 과제가 완성될 수 있었습니다.

**1.개요 및 목적**

코로나 19 사태로 인해 피해성과 코로나 19에 대한 데이터를 분석하기 위해 시도했지만 현재 데이터 안심구역에서 코로나 19와 관련하여 학생들에게는 정보를 공유해 주지 않아 학생 신분에 어울리는 영화 관람 분석에 대한 내용을 경험했습니다. 또한 영화 관람에 대한 분석 과정에서 데이터의 자료들이 2019년 1월~12월까지인 걸 고려하여 작년에 가장 화두가 됐던 환경 분야의 미세먼지에 대한 데이터를 접근 할 수 있었습니다.

최근 미세먼지로 인해 외부 활동이 자제된 점, 마스크를 착용한 점들이 사회에 어떤 영향을 미쳤는지 생각해보았습니다. 따라서 초기에 계획했던 프로젝트는 미세먼지로 마스크를 착용하게 되면서 기능성 화장품의 소비가 실제로 늘어났는지, 기름기 있는 음식을 먹으면 미세먼지가 씻겨 내려간다는 이야기를 통해 소비자들이 식자재 중 고기 지출이 증가하였거나 외부활동 자제로 영화 관람 관객 수가 감소하였는지에 대해 분석해보고 시각화해보고 싶었습니다.

그러나 공공데이터나, 서울시 데이터를 통해 고객들의 소비 데이터를 구하기엔 데이터들이 제한적이어서 분석하기 어려웠습니다.

따라서 매해 봄철마다 돌아오는 중국발 황사 문제와 더불어 미세먼지 및 초미세먼지 경보, 호흡기 질환 급증, 시정 장애 등과 같은 문제 발생으로 미세먼지에 대한 관심과 관련 연구 필요성이 대두되고 있습니다. 실생활에서는 최근까지 미세먼지로 인해 외부 활동이 자제되었으며 마스크 착용이 생활화되고 있으며 미세먼지 및 초미세먼지로 인한 질병 및 질환으로 사람들의 피해가 증가하여 저희 조는 서울시 미세먼지를 파악하여 어떤 요인과 상관관계를 미치는지 알아보기 위한 프로젝트로 변경하였습니다.

**1.2.사용된 기술**

1. 크롤링 2) 데이터 마이닝 3) 워드클라우드 4) R프로그래밍

**2.데이터 분석 단계**

1. **크롤링**

**기사 분석**

**Crawling**

**미세먼지 관련 지면 기사 크롤링**

**미세먼지 키워드에 대한 관심 추이 분석**

****

텍스트, 신문, 서류이(가) 표시된 사진

자동 생성된 설명

미세먼지의 실시간 반응과 원인과 관련된 키워드를 조사하기 위해 네이버 검색어 “미세먼지 원인” 키워드를 통해 지면뉴스 Title을 기반으로 Word Cloud를 진행하여 주요 키워드를 조사했습니다.

사람들의 관심사를 분석하기 위해 “미세먼지 원인” 관련 지면 기사들을 크롤링 하여 페이지당 10건의 뉴스가 보여지도록 url<-c(url, past0(base\_url, x\*10+1, seq=“”))명령어로 최대한 많은 양의 기사 title을 가져올 수 있도록 url을 설정하고 각 기사의 html\_text()를 이용하여 title 제목만을 추출했습니다.

**WordCloud**

**2) 텍스트마이닝**



테이블이(가) 표시된 사진

자동 생성된 설명R스튜디오를 사용하여 주요 키워드 분석 결과 뉴스 미세먼지 키워드 분석 결과 중국발, 자동차, 경유차, 배출, 석탄 같은 원인에 대한 키워드가 나타난 것을 볼 수 있습니다.

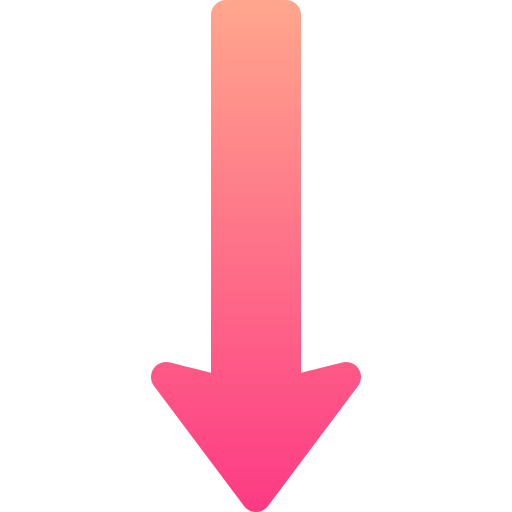
**3) R언어 적용을 통한 년 평균 데이터 도출**

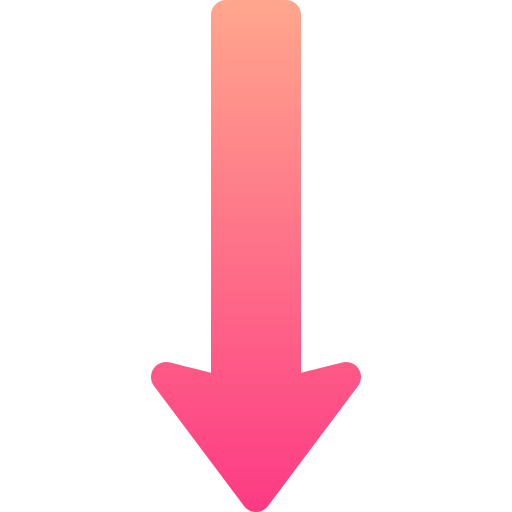
테이블이(가) 표시된 사진

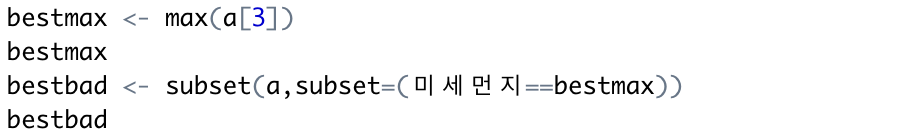
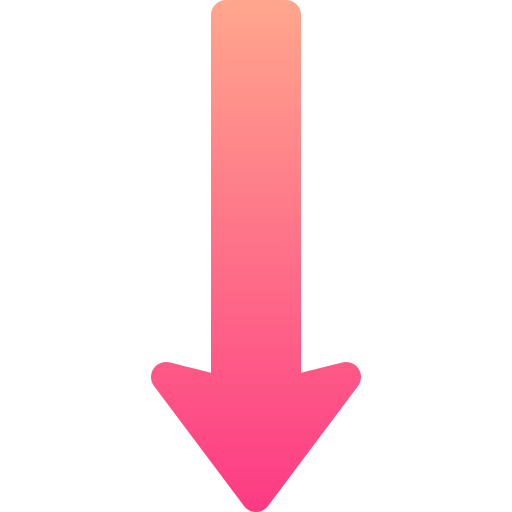
자동 생성된 설명텍스트이(가) 표시된 사진

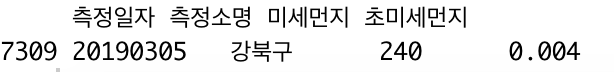
자동 생성된 설명

8조







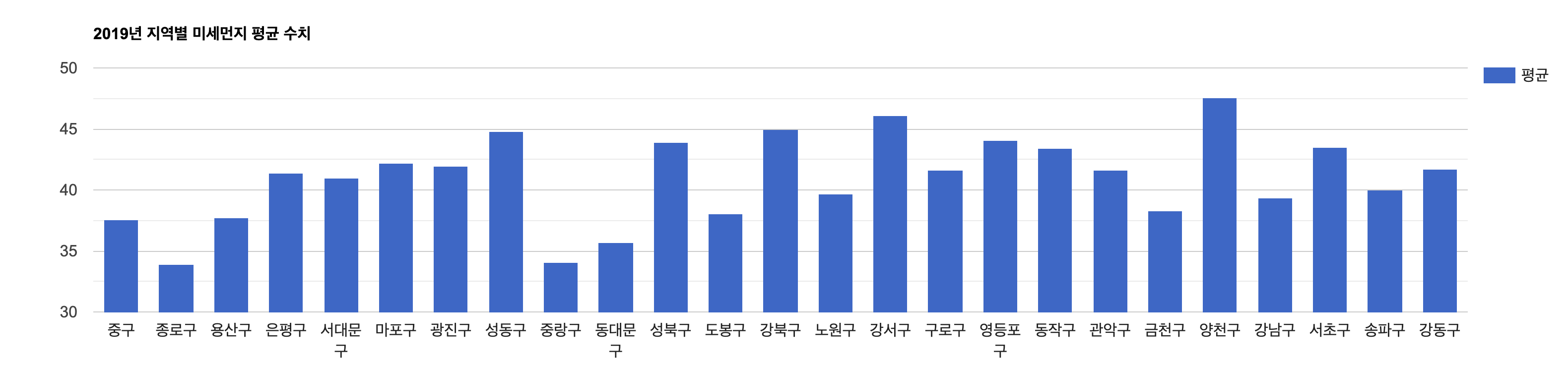


데이터 처리 과정에는 R스튜디오를 사용하였습니다. 2019 기간별 대기환경 데이터에는 측정일자, 권역코드, 권역명, 측정소코드 및 이름, 대기환경 정보가 담겨져있습니다.

그러나 분석할 데이터는 측정일자, 권역명, 측정소명, 미세먼지, 초미세먼지 이므로 이 외의 변수는 삭제하였습니다. 2019 csv파일에서 불필요한 변수는 삭제한 후 R스튜디오 read.csv하여 데이터를 처리했고 read.csv함수를 사용하여 파일을 불러온 후 각 측정소명을 약어로 바꿔주었습니다.

이후 서울시 지역별 미세먼지와 초미세먼지의 년 평균을 구하기 위해 mean() 함수를 사용했으며 그래프로 시각화하기 위해 data.frame로 저장해주었습니다.

**4) 데이터 시각화**

**4.1 googlevis사용**

**4.2 ggmap 사용**

Chart

Description automatically generatedChart

Description automatically generatedA screenshot of a cell phone

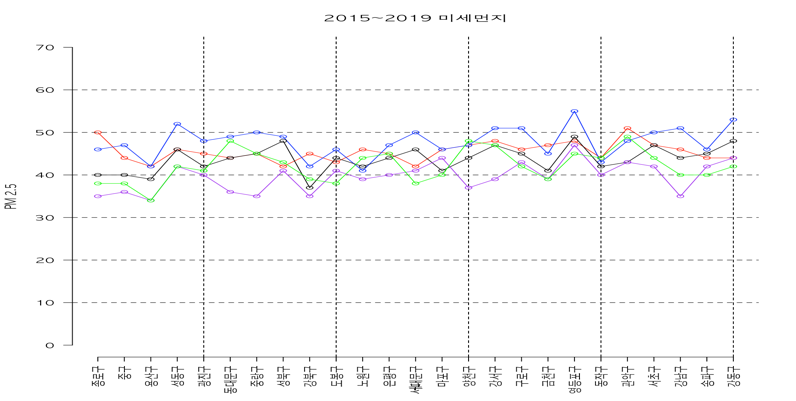
Description automatically generated

**양천구 평균 47.58146으로 가장 높음**

**중랑구 평균 0.004940678으로 가장 높음**

8조

4.3 2015~2019년 지역별 서울시 미세먼지 평균 시각화

테이블이(가) 표시된 사진

자동 생성된 설명

데이터 처리 후 googleVis 라이브러리를 통해 2019년도 서울시 지역별 미세먼지 평균을 height=400, weight=500의 막대그래프를 사용하여 시각화 하였습니다.

또한 막대 그래프 외에 서울시 지역을 지도로 시각화하여 알아보기 위하여 ggplot 라이브러리로 시각화 하였습니다. 이 결과 2019년도 미세먼지 농도 평균 양천구가 가장 높은 것을 알 수 있으며, 이후 강서구 강북구 순으로 높았습니다. 반면 중랑구의 미세먼지 평균이 가장 낮았던 것을 알 수있습니다.

ggplot함수를 사용하여 경도와 위도를 설정하여 구별 경계를 나타낸 후 ggplot을 활용하여 평균에 따라 가장 높은 구는 진한 색상 낮은 구는 연한 색상을 사용하였습니다.

Map의 그림을 보면 지역마다 미세먼지와 초미세먼지의 평균 차이가 존재하는 것을 확인 할 수 있습니다. 그림 8과 같이 양천구가 가장 높으며 중랑구가 낮음을 지도를 통해 시각적으로 알아볼 수 있습니다.

지역별 차이의 존재 여부를 보다 정확하게 확인하기 위해 2019년 데이터뿐만 아니라 2015~2019 서울시 지역별 연평균 데이터를 이용했습니다. 그래프의 형태가 굉장히 불규칙적으로 보이지만 특정 부분에서 유사한 부분을 발견할 수 있습니다. 년도 별 특정 지역의 수치가 매년 상위권인 것을 확인 할 수 있다. 영등포구의 수치가 대체로 높게 나온 것을 확인할 수 있습니다.

**3. 서울시 미세먼지 요인 분석**

**3.1 2018 서울시 사업체 현황**

저희 조는 2019 미세먼지 파악한 결과 양천구 가장 높게 측정, 초미세먼지 중랑구가 가장 높게 측정되었습니다. 미세먼지의 요인 중 하나는 가스, 석탄, 석유 등 연소 시 발생하는 대기오염 물질이라는 결과가 나왔습니다. 따라서 저희는 공장 및 가스, 석탄, 석유 등 연소가 필요한 사업체가 많은 지역의 미세먼지와 초미세먼지가 증가한다고 생각하여 2018년도의 석탄, 석유들의 사업체 지역 현황을 파악하였습니다.

2015년~2019년도의 미세먼지 결과 영등포구, 양천구 등 사업체 수가 가장 많을 것으로 예측하였습니다.

지도이(가) 표시된 사진

자동 생성된 설명

주 원인이 되는 사업체의 수를 보면 서울시의 약 5%에 불과하며 사업체가 가장 높은 곳은 강남에 해당 되었습니다.

영등포구의 미세먼지가 높게 나오는 이유로는 2019년 11월까지 영등포구의 미세먼지 측정 장소는 당산1동 주민센터 옥상이며 당산 1동에는 먹자골목이 형성되어 있어 먹자골목에서 나오는 각종 연기가 미세먼지 측정의 정확성을 떨어트립니다.

그로 인해 뉴스 기사를 바탕으로 보면 작년 11월 우리은행 영등포구청점 건물 옥상으로 측정 위치를 이전하였으며. 그 후 2020년 1분기 총 6월까지의 미세먼지 수치가 하위권으로 조정되었습니다.

**3.2 서울시 대기오염 물질 수치**

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

2019년 미세먼지 수치가 높았던 양천구의 대기오염 물질을 살펴보면 아황산가스 0.005 일산화탄소 0.5 이산화질소 0.032 오존 0.022으로 다른 지역에 비해 대기오염 물질의 수치가 높지 않은 것을 확인 할 수 있습니다.

또한 각 대기오염물질 별 가장 높은 수치를 가지는 지역 또한 미세먼지 수치가 높지 않습니다.

**결론**

결과적으로 미세먼지 지역의 차이를 확인하기 위해 R프로그래밍을 이용하여 위와 같이 시각화, 분석을 진행하여 지역적 차이의 분명함 시각화하였고 WordCloud를 통한 키워드를 분석하여 미세먼지의 주요 원인이 무엇인지 확인할 수 있었습니다.

미세먼지의 주원인이 되는 석유, 석탄, 가스 연소 시 발생하는 대기오염 물질, 그 물질을 발생하는 사업체를 통해 지역별 사업체 수를 조사하여 미세먼지 수치가 높은 지역을 비교하였지만, 미세먼지 수치와는 큰 영향을 작용하지 않는 것으로 나타나고 있습니다.

또한 대기오염 물질의 수치를 파악하여 미세먼지와의 상호 연관성을 조사하였고 그 결과 또한 미세먼지 수치와 크게 상관없는 것으로 나타나고 있습니다.

미세먼지를 측정하는 측정소의 위치에 따라 미세먼지 수치의 정확성을 떨어트리는 외부 요인이 작용하며 그 외 미세먼지의 원인이 되는 교통량, 국외 요인 등 고려해야 할 사항이 많은 것으로 판단하여 다수의 요인이 있을 거라고 생각합니다.

그러므로 서울시 내의 자치구의 차이점은 미세먼지 측정소의 위치로 인한 외부 영향으로 미세먼지 측정의 정확성을 떨어트리는 점과 국내에서 생성되는 미세먼지의 원인이 되는 정확한 요소를 확인하기 어렵기 때문에 국내에서 발생하는 미세먼지의 요인보다 국외의 미세먼지 발생 요인에 대해 더 고려해 볼 필요가 있습니다.

**참조**

1. 데이터안심구역, <https://dsz.kdata.or.kr/svc/data/search.do>
2. 데이터셋(서울열린데이터광장), <https://data.seoul.go.kr/dataList/OA-2220/F/1/datasetView.do;jsessionid=8A20A9F90045C1143FB406FD8ABAA73D.new_portal-svr-11>
3. 서울특별시 대기환경정보, <https://cleanair.seoul.go.kr/2020/>
4. 서울시 행정구역 시군구 정보 <https://data.seoul.go.kr/dataList/OA-11677/S/1/datasetView.do>
5. 대한민국 최신 행정구역 파일 다운로드 <http://www.gisdeveloper.co.kr/?p=2332>
6. 에어코리아

https://www.airkorea.or.kr/web/detailViewDown?pMENU\_NO=12