[0112]_통계적사고_9기_박서연_과제제출

과제 2. EDA 프로젝트에서 통계적으로 의사소통하거나 통계적 기법 활용할 수 있는 부분

1) t-test를 이용해 대기오염물질의 계절성/일변화 경향성 검증

앞서 시계열 그래프를 그리는 시각화를 통해 각각의 대기오염물질별로 계절성과 일변화 경향성이 있으리라는 추측을 할 수 있었다. 이산화질소, 일산화탄소, 아황산가스는 주로 겨울에 높고 여름에 낮으며, 오존은 주로 여름에 높고 겨울에 낮다. 이산화질소, 일산화탄소, 아황산가스는 주로 밤에 높고 낮에 낮으며, 오존은 주로 낮에 높고 밤에 낮다. 이러한 사실들은 사실 통계적으로 검증한 것이 아니라 시각화를 통해 눈으로만 파악한 것이므로, 데이터와 숫자를 통한 근거를 파악하는 편이 좋을 것이다. 이때 두 표본의 평균이 같은지, 다른지를 비교할 수 있도록 해주는 t-test를 적용해볼 수 있을 거라는 생각이 들었다.

i) 계절성 파악

모집단 1: 전체 측정 기간 중 여름(6~8월) 동안의 특정 대기오염물질 데이터
모집단 2: 전체 측정 기간 중 겨울(12~2월) 동안의 특정 대기오염물질 데이터
표본 1: 모집단 1에서 추출한 데이터
표본 2: 모집단 2에서 추출한 데이터

귀무가설: 여름 표본의 평균과 겨울 표본의 평균은 같다
대립가설: 여름 표본의 평균과 겨울 표본의 평균은 다르다

이산화질소를 예로 들어보겠다. 시각화 결과를 봤을 때 이산화질소는 여름에 낮고 겨울에 높은 계절성을 띤다는 것을 유추할 수 있었다. 이때 t-test를 이용하면 실제로 이산화질소의 여름 농도, 겨울 농도가 유의미하게 차이 나는지를 통계적으로 검정할 수 있을 것이다.

전체 측정 기간 중 여름 동안의 이산화질소 데이터에서 표본 1을 추출하여 표본 평균을 구하고, 마찬가지로 전체 측정 기간 중 겨울 동안의 이산화질소 데이터에서 표본 2를 추출하여 표본 평균을 구한 뒤, 둘의 차이가 유의한지를 t-test를 이용해 검정해볼 수 있다. 만일 이산화질소가 실제로 여름에는 낮고, 겨울에는 높은 계절성을 띤다면 여름 표본의 평균과 겨울 표본의 평균 사이에 유의한 차이가 발생할 것이고, 귀무가설이 기각될 것이다.

ii) 일변화 경향성 파악

모집단 1: 전체 측정 기간 중 낮(오전 10시~오후 2시)[1] 동안의 특정 대기오염물질 데이터
모집단 2: 전체 측정 기간 중 밤(오후 9시~오전 1시) 동안의 특정 대기오염물질 데이터
표본 1: 모집단 1에서 추출한 데이터
표본 2: 모집단 2에서 추출한 데이터

---

1) 낮과 밤의 기준이 모호할 수 있으므로, 그에 대해서는 더 고민할 필요가 있어 보인다.

귀무가설: 낮 표본의 평균과 밤 표본의 평균은 같다
대립가설: 낮 표본의 평균과 밤 표본의 평균은 다르다

이번에도 이산화질소를 예로 들어보겠다. 시각화 결과를 봤을 때 이산화질소는 낮에 낮고 밤에 높은 일변화 경향성을 띤다는 것을 유추할 수 있었다. 이때 t-test를 이용하면 실제로 이산화질소의 낮 농도, 밤 농도가 유의미하게 차이 나는지를 통계적으로 검정할 수 있을 것이다.

전체 측정 기간 중 낮 동안의 이산화질소 데이터에서 표본 1을 추출하여 표본 평균을 구하고, 마찬가지로 전체 측정 기간 중 밤 동안의 이산화질소 데이터에서 표본 2를 추출하여 표본 평균을 구한 뒤, 둘의 차이가 유의한지를 t-test를 이용해 검정해볼 수 있다. 만일 이산화질소가 실제로 낮에는 낮고, 밤에는 높은 일변화 경향성을 띤다면 낮 표본의 평균과 밤 표본의 평균 사이에 유의한 차이가 발생할 것이고, 귀무가설이 기각될 것이다.

2) 상관분석을 이용하여 어떤 대기오염물질과 다른 대기오염물질, 대기오염물질과 기상 인자 사이의 상관성 탐구

어떤 대기오염물질과 다른 대기오염물질(ex. 이산화질소와 오존, 오존과 일산화탄소...) 사이에 상관관계가 있는지, 대기오염물질과 기상 인자(ex. 미세먼지와 강수량, 오존과 일사량) 사이에 상관관계가 있는지를 통계 기법 중 상관분석을 이용해 알아볼 수 있다.

+ 대기오염물질과 기상 인자 사이의 상관성은 어떻게 활용할 수 있을까?

-> 우리는 일기예보는 자주 확인하지만, 대기오염물질 예보[2]에 대해서는 크게 관심을 기울이지 않는다. 그나마 최근 미세먼지와 초미세먼지에 대한 위험성이 조명되면서 그에 대한 관심은 다른 대기오염물질과 비교해봤을 때 크게 증가하였으나, 우리의 데이터셋에서 다루고 있는 다른 대기오염물질(이산화질소, 오존, 아황산가스, 일산화탄소)에 대해서는 대부분의 사람들이 굳이 찾아보고자 하는 노력을 하지 않는 것이 현실이다. 그렇지만 많은 사람들이 하루하루의 일기예보(기온, 강수 여부)는 매일 확인하기 때문에, 접근성이 좋은 기상 인자들을 토대로 대기오염물질에 대한 위험도를 유추할 수 있는 가이드라인(이를테면 일사량이 강한 날은 오존 농도도 높을 테니 주의하라고 안내하는)을 제공한다면 대기오염물질에 대한 위험성을 미리 알고 대비하는 데 도움을 줄 수 있을 것이다.

2) 사실 예보 자체가 제대로 이루어지지 않고 있는 것이 현실이다. 실제로 서울특별시 대기환경정보 사이트(https://cleanair.seoul.go.kr/forecast/airForecast)에서는 대체로 미세먼지에 대한 예보만 발행하고 있으며(오존도 예보 대상에 포함되어 있긴 하나 데이터가 뜨지 않을 때가 많다), 대기오염물질에 대한 정보는 주로 예보보다는 실황에 집중되어 있다.

**** EX. 추가 데이터와 엮어서 분석 2**

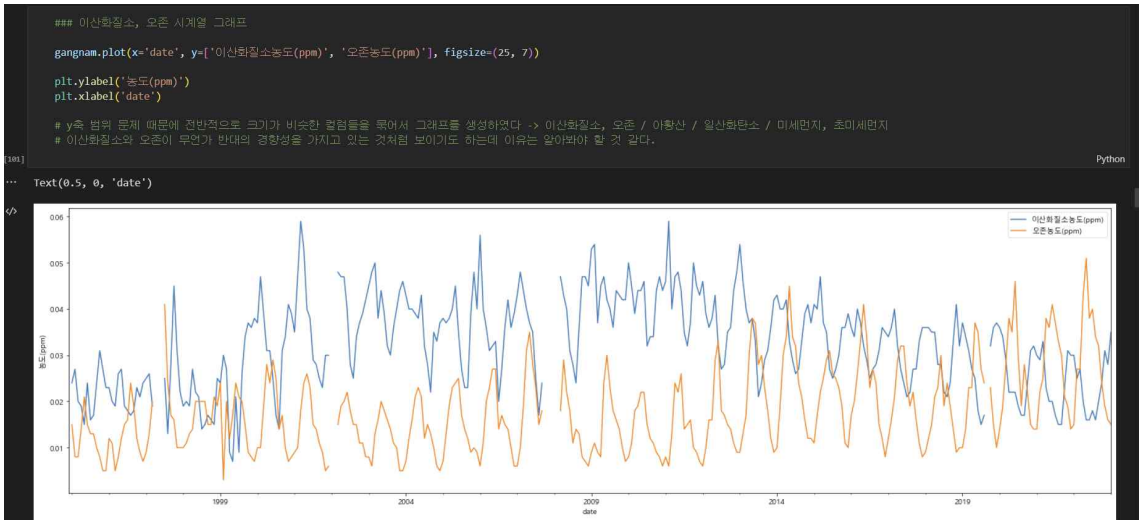

그림 1 전체 범위(1994~2023년) 강남구의 이산화질소 및 오존 시계열 그래프

그림 2 20201~2022년 강남구의 이산화질소 및 오존 시계열 그래프

이산화질소와 오존의 시계열 그래프를 함께 그려봤을 때 마치 반비례 관계에 있는 듯한 상관성이 포착되었는데, 논문을 찾아본 결과[3] 이는 사실인 것으로 확인되었다. 오존은 1년 단위로 봤을 때는 여름, 하루 단위로 봤을 때는 낮에 그 농도가 높아지고, 겨울 및 밤에는 농도가 낮아지는데, 이는 기상 인자 중 일사량과 관계가 있다. 자외선이 가해질 경우 질소산화물 및 휘발성유기화합물(VOCs)이 화학 반응을 일으켜 오존이 탄생하기 때문에, 햇빛을 강하게 받는

3) 관련 논문
- 서울시 대기중 오존오염도의 연도별 변화와 그 영향인자에 관한 연구
https://ir.ymlib.yonsei.ac.kr/handle/22282913/124403

- 대기오염농도와 기상인자의 관련성 연구
https://koreascience.kr/article/JAKO199211919855307.page

[0112]_통계적사고_9기_박서연_과제제출

여름이나 낮에는 오존의 농도가 높아지는 반면 질소산화물(우리의 데이터셋의 경우 이산화질소)의 경우 그 농도가 낮아진다.
이렇듯, 논문을 통해 각 대기오염물질이 다른 대기오염물질뿐 아니라 기상 인자와도 상관관계를 맺을 수 있음을 알게 되어, 통계 기법 중 상관분석을 활용해 실제로 그러한 상관관계가 있는지를 알아보는 것도 하나의 길일 것 같다.

+ 추가로 활용할 수 있는 데이터는 무엇이 더 있을까?

1) 기상 데이터

기상 데이터 다운로드 받을 수 있는 사이트
기상청 날씨데이터 서비스, 기상자료개방포털: https://data.kma.go.kr/cmmn/main.do

-> 아직 기상 인자와 연관 지어 분석을 할지 정해지지 않은 상태이기도 하고, 워낙 데이터의 종류나 범위가 방대하다보니 분석을 진행할 지역구 및 시간대가 정해지면 그에 맞는 자료를 내려받는 것이 좋을 것 같아 아직 다운로드는 하지 않은 상태이다.

2) 오존 데이터: MERRA2 재분석장

-> 전 지구를 대상으로 한 방대한 데이터가 포함되어 있는 자료라서 분석한다면 중위도 지역 전체를 대상으로 오존의 계절성을 직접 파악해보기에 좋을 듯하다.

+ 어떤 대기오염물질을 주로 분석할까?

- 미세먼지, 초미세먼지 포함
사람들의 관심도가 크고, 웬만하면 대기환경기준을 넘지 않는 여타 대기오염물질(이산화질소, 오존, 아황산가스, 일산화탄소)들과 달리 미세먼지, 초미세먼지의 경우 나쁨, 매우나쁨 일수가 꽤 되기 때문에 대기오염에 대해 이야기할 때 빼놓을 수 없으리라 생각한다.

- 이산화질소, 오존 포함
이산화질소와 오존은 서로 간의 상관성이 비교적 크고, 미세먼지, 초미세먼지를 제외한 다른 대기오염물질 사이에서는 비교적 그 농도가 높은 편이라 그만큼 우리 생활에 미치는 영향이 보다 클 수 있으므로, 이 둘도 포함시키는 것이 좋을 듯하다.

- 제외한다면: 아황산가스, 일산화탄소
아황산가스와 일산화탄소도 오존 등 여타 대기오염물질과 상호작용을 하긴 하나 그 상관성이 크지 않고 전반적인 농도도 우리의 데이터셋에 포함된 다른 대기오염물질만큼 높지 않기 때문에 컬럼을 제외한다면 이 둘을 제하는 것이 보다 합리적일 것 같다.

[0112]_통계적사고_9기_박서연_과제제출

+ 어느 지역으로 통일해서 분석할까?

1) 공원 부지 추천

서울시 공원 통계
https://data.seoul.go.kr/dataList/10052/S/2/datasetView.do

서울시에서 공원이 가장 적은 구 Top 5
- 금천구(55개소)
- 광진구(68개소)
- 중구(76개소)
- 도봉구(80개소)
- 강북구(84개소)

서울시에서 공원 면적이 가장 작은 구 Top 5
- 동대문구(1,216.0 천m²)
- 용산구(1,775.5 천m²)
- 금천구(2,774.4 천m²)
- 영등포구(3,009.1 천m²)
- 양천구(3,049.6 천m²)

-> 우리의 목표는 대기오염물질과 공원 부지 추천을 연결 지으려는 것이므로, 대기오염 저감 효과를 누리려는 목적으로 미루어봤을 때 공원 개소보다는 면적을 보다 우선적으로 고려하는 것이 좋을 듯하다. 여기에 더해 각각의 구의 대기오염물질 농도, 인구수를 함께 고려하여 공원 부지를 추천하면 좋을 것 같다.

2) 한강 피크닉 추천

한강공원이 위치해 있는 구: 이용객 도합
- 송파구(광나루 한강공원, 잠실 한강공원): 7,756,024
- 광진구(뚝섬 한강공원): 15,121,427
- 서초구(잠원 한강공원, 반포 한강공원): 8,536,129
- 용산구(이촌 한강공원): 1,981,496
- 영등포구(여의도 한강공원, 양화 한강공원): 15,230,778
- 마포구(망원 한강공원, 난지 한강공원): 5,247,545
- 강서구(강서 한강공원): 3,756,296

-> 한강공원 이용자 수 Top 5 구: 영등포구, 광진구, 서초구, 송파구, 마포구

서울시 한강공원 이용객 현황 통계
https://data.seoul.go.kr/dataList/10798/S/2/datasetView.do

[0112]_통계적사고_9기_박서연_과제제출

2021년 한 해 동안 이용객이 많은 공원 Top 5
- 뚝섬(15,121,427)
- 여의도(11,404,192)
- 잠원(4,957,543)
- 잠실(4,766,845)
- 양화(3,826,586)