**3주차 과제: EDA 프로젝트에서 통계적으로 의사소통을 해볼 수 있는 부분이 어떤 것들이 있을까 OR 통계적 기법을 활용할 수 있는 부분이 있나?**

9기 임선민 (4조: 환경 - 대기오염도)

**A) 통계학 및 통계적 사고**

통계학이란, 하나의 문제를 어떤 방법으로 해결해야 가장 잘 해결할 수 있을지 탐구하는 학문이다. 3주차 세션에서는 4개의 대표적인 통계적 사고 이론 및 실습을 학습하였으며, 4개의 주제는 바로 t-test, Anova, Regression, Sampling theory이다. 아래는 이러한 통계적 기법을 어떻게 4조의 데이터셋에 적용해 볼 수 있을 지 생각해본 아이디어이다.

**B) 4조 EDA 프로젝트 적용 방안**

1) t-test

우선, 서로 다른 두 표본을 비교를 하기 위해 사용하는 two-sample t-test를 진행해 볼 수 있을 것 같다. 예를 들어, 10년동안의 특정 월 (ex: 벚꽃이 가장 만개하는 4월) 미세먼지 농도의 평균 데이터를 정리하여, 가장 높은 미세먼지 농도를 보이는 구와 가장 낮은 미세먼지 농도를 보이는 구를 두 개의 표본으로 선정하여, 두 표본 간의 농도 차이가 통계적으로 유의미한지, 아니면 그저 우연일 뿐인가 분석해볼 수 있다. 만약 두 개의 구간 미세먼지 월별 농도 차이가 유의미하다는 결과가 나온다면, 미세먼지 농도가 높은 구역에 공원 부지를 만들어야 한다는 주장에 대한 근거 그리고 미세먼지 농도가 낮은 구역에 한강 피크닉을 가기 적합하다는 근거가 될 수 있다.

다음으로, 하나의 표본의 두 시점을 비교하는 paired-sample t-test 또한 진행해 볼 수 있을 것 같다. 예를 들어, 한강 공원이 위치하는 하나의 구를 선정하여, 해당 구의 일별 데이터를 분석하여, 가장 높은 미세먼지 농도를 보이는 시간 데이터와 가장 낮은 미세먼지 농도를 보이는 시간 데이터를 두 개의 표본으로 선정할 수 있을 것이다. 이후, 두 표본 간의 농도 차이가 통계적으로 유의미한지, 아니면 그저 우연일 뿐인지 똑같이 분석해볼 수 있다. 만약 그 차이가 통계적으로 유의미하다는 결과가 나온다면, 미세먼지 농도가 낮은 시간에 한강 피크닉을 가야 한다는 주장에 대한 적절한 근거 및 4월에 한강 피크닉을 가기 좋은 시간대를 추천할 때의 기준이 될 것이다.

2) ANOVA

ANOVA는 Analysis of Variance (분산분석)으로, 특히 One-way ANOVA는 하나의 변수에 대해 세 개 이상의 표본의 평균이 같은지 또는 다른지 비교하기 위해 사용하는 통계적 기법이다. 4조의 데이터 셋에서 ANOVA를 활용할 수 있는 방법으로는 한강공원이 위치하는 대표적인 5개의 구에 대한 4월의 미세먼지 평균농도가 서로 같은지 또는 다른지 분석해보는 적용방안이 있다. 분석결과를 통해, 5개의 구 중 4월에 피크닉을 가기 좋은 구의 한강공원을 추천할 수 있을 것이다. 또한, ANOVA를 활용할 수 있는 다른 방법은 특정 구를 하나 선정하여, 월별 또는 일별 대기오염물질들의 농도의 평균이 같은지 또는 다른지 분석해보는 적용 방안도 생각해볼 수 있다. 분석 결과를 통해 대기오염물질 간의 상관관계를 볼 수 있고, 7개 중 분석할 4개를 선정하는 근거가 될 수 있다.

3) Regression

회귀 분석이란, 독립변수로 종속변수를 예측하는 통계적 방법론이다. 2주차 EDA 방법론 세션 때 학습한 내용을 토대로 서울시 시간별 대기오염물질 농도 정보 데이터를 이용하여 분석해본 결과, 대기오염물질 간에 상관관계를 보임을 알 수 있었다. 오존과 이산화질소, 미세먼지와 초미세먼지 등 두개의 대기오염물질은 서로 밀접한 상관관계를 보이기에, 이러한 상관관계를 나타내는 그래프를 그린 후, 데이터를 설명하는 추세선을 그려봄으로써 통계적 사고를 적용해보면 좋을 것 같다. 이후, t-test를 통해 분속한 추세선의 표준오차를 통해, 나타난 추세선이 유의미한지 또는 우연인지 분석해볼 수 있다. 이러한 분석 결과는 대기오염물질 간의 상관관계를 파악하고, 분석에 사용할 특정 대기오염물질을 선정하는 적절한 기준이 될 수 있다.

\*\*통계적 기법을 이용하여 4조 eda 데이터를 분석해본 ipynb 파일을 같이 첨부합니다. 최종적으로 확실히 결정된 프로젝트의 방향으로 다시 자세히 분석을 진행할 예정입니다.