**2주차 과제: EDA 방법론 세션 내용을 바탕으로, 본인 팀 데이터에 맞는 EDA를 어떻게 진행할 지, 자유 양식의 형태로 서술하여 제출해주세요.**

9기 임선민 (4조: 환경 - 대기오염도)

**A) EDA 일반론이란?**

EDA란 Exploratory Data Analysis로, 탐색적 데이터 분석을 일컫는다. 프로젝트에서 분석 목적에 따라 분석 진행 방향, 그리고 접근법이 달라지는데, EDA 방법론 세션에서는 EDA 일반론 과정을 학습하였다. 과정의 단계는 1) 데이터 수집, 2) 데이터 전처리, 3) 데이터 Scaling, 4) 데이터 시각화, 5) 사후 처리이다. EDA 일반론의 각각의 단계를 4조의 EDA 프로젝트에 어떻게 적용할 수 있을지 생각해보았다.

**B) 4조 EDA 프로젝트 적용 방안**

(1) 데이터 수집

4조의 데이터셋은 총 6개로, 서울시 년도별 평균 대기오염도, 서울시 월별 평균 대기오염도, 서울시 시간 평균 대기오염도 정보, 서울시 지역 구별 측정소 행정코드 정보, 서울시 기간별 일평균 대기환경 정보, 그리고 서울시 기간별 시간평균 대기환경 정보이다. 필요한 데이터는 서울 열린데이터 광장에서 수집을 모두 마쳤고, 시간 평균과 기간별 시간평균 데이터의 경우는 다운받는 일시에 따라 달라진다는 특징이 있다.

분석하는 방향과 서울시 지역 구별 측정소 행정코드 정보는 필요성이 떨어지기에, 선택적으로 꼭 필요한 경우에만 사용하면 좋을 것 같고, 서울시 시간 평균 대기오염도 정보와 서울시 기간별 시간평균 대기환경 정보는 동일한 데이터의 내용과 범위를 다루고 있기에, 상대적으로 최근 또는 결측치가 적은 데이터 셋을 활용하면 좋을 것 같다.

(2) 데이터 전처리

데이터 셋에 대한 결측치 처리 및 이상치 탐색의 과정이 필요하다. 각각의 데이터셋의 결측치처리 방법을 통일하는 과정 또한 필요하다. 이때, 결측치를 포함하는 행 자체를 없애는 것보다, 해당 데이터 종류의 평균값을 이용하여 보간하는 방향이 더 좋을 것 같다.

데이터 셋 분석하는 과정에서 너무 크게 결측치가 나타나거나, 이상치가 관찰이 된다면, 그 이유 및 구별/지역별 관측소에 문제가 있었는지 등에 대한 이유도 분석해볼 필요가 있을 것 같다.

(3) 데이터 Scaling

주어진 6개의 데이터셋의 활용 방안을 생각해보면, 서울시 년도별 평균 대기오염도 정보는 초근 5~10년 정도의 데이터를 활용하여, 구별 대기오염도 농도 및 오염물질 각각의 전체적인 추이를 관찰하기 위해 활용할 수 있다. 또한, 장기간의 변화 및 상관성 추이를 통해, 대기 물질 간의 상관관계를 파악하여, 상관관계가 서로 높아 분석하기 좋은 오염물질을 7개 중 4~5개 선정하는 것이 좋을 것 같다.

서울시 월별 평균 대기오염도 정보는 계절별로 분석해보거나, 역시 전체적인 추이를 관찰하기 위해 사용할 수 있으며, 분석 방향에 따라 1개의 년도만을 특정하여 분석하거나, 5~10년 정도의 장기간 데이터를 이용해볼 수 있을 것 같다. 이러한 분석 결과는 4조의 EDA 프로젝트 아이디어 중 공원 부지를 만들 지역을 선정하는 기준이 될 수도 있고, 벚꽃 한강 피크닉을 가기에 가장 적합한 월을 추천할 때의 근거가 될 수도 있다.

서울시 시간 평균 대기오염도 정보 및 기간별 시간평균 대기환경 정보는 묶어서 함께 분석할 필요가 있을 것 같다. 데이터 셋을 다운로드 받는 일시에 따라 데이터가 계속 바뀌므로, 다운받아 분석하는 기간을 1월 둘째 주 등으로 특정할 필요가 있으며, 한강이 위치한 구의 하루 동안의 대기 오염정도를 분석할 때, 한강 피크닉이 가장 적합한 장소(구) 또는 시간을 추천해주는 근거가 될 수 있다.

데이터의 범위를 조율하는 과정에서 시간의 범위 및 장소의 범위를 조율하는 과정이 필요하다. 24시간 또는 오전/오후 또는 아침/저녁 등 시간의 범위도 묶거나 특정할 필요가 있으며, 장소(구)는 임의로 몇 개를 선정하거나, 서북/서남/동북/동남/도심으로 구역을 5개로 크게 나누어 분석을 진행하는 것이 좋을 것 같다. 주어진 데이터를 그대로 모두 사용하려고 한다면, 구의 개수 및 시간의 범위도 너무 커서 분석을 진행하는 과정이 너무 복잡하고 힘들 것이다.

(4) 데이터 시각화

데이터의 일변량 시각화의 방안으로 QQ-Plot 또는 다변량 시각화 방안으로 Scatter plot 또는 범주별 시각화를 활용하면 좋을 것 같다. 년도별/월별/시간별 대기오염물질 변화 추이를 그래프로 그리거나, 서울시 전체 지도에 구별로 색깔의 진하기로 대기오염물질 농도를 그리거나, 오염물질 간의 상관관계를 히트맵 또는 회귀분석 그래프로 그리는 등의 시각화 방안을 생각해보았다.

5) 사후 처리

프로젝트 마무리 단계로, 추후 분석 방향을 결정하고 Fine Tuning을 진행하는 과정이다. 이는 프로젝트 분석을 마무리하고 한계점을 통해 더 생각해보겠다.

\*\*6개의 데이터 셋 중 서울시 시간 평균 대기오염도 정보를 EDA 일반론의 과정으로 조금 분석해본 자료를 함께 제출합니다. 최종적으로 확실히 결정된 프로젝트의 방향으로 다시 자세히 분석을 진행할 예정입니다.