1.

안녕하세요, 데면대면 팀입니다.!

제일 먼저 저희 팀명 작명 비화를 말씀드리겠습니다,,, ㅎㅎ

아시다시피 고정적으로 대면으로 현장에 나오는 수강생은 저와 종원님 두 명 뿐인데요, 그럼에도 불구하고 아직 서로 친하지 않았던 점이 있었어서 ‘데면데면하다’ 와 ‘대면’ 수강생의 의미를 합친 언어유희인 ‘데면대면’이 어떻냐고 아이디어를 냈고, 종원님도 좋다고 해주셔서 팀명이 확정되었습니다.

저희 발표의 제목은 ‘유흥시설 밀집도와 범죄 예방 설치물 대비 청소년 범죄 빈도 분석’ 입니다.

발표 시작하도록 하겠습니다.

2.

목차를 말씀드리겠습니다.

첫번째로는 저희 팀의 연구 주제와 수집한 데이터셋을 간략하게 소개드리도록 하겠습니다.

두번째로는 데이터 전처리 과정을 설명드리도록 하겠습니다.

그 다음으로는 각 주제에 대한 데이터 분석 결과를 보여드리고,

마지막으로 결론과 분석의 한계점에 대해 설명드리도록 하겠습니다.

3.

저희 팀의 연구 주제와 데이터셋을 소개하겠습니다.

4.

저희는 주제를 정하기 전에 먼저 어떤 데이터들을 가져올 수 있는지 먼저 구글링을 해보았습니다!

그러면서 발견한 데이터들로 여러개의 아이디어를 냈고,

그 중에서 채택한 주제는 “ 지역별 유흥시설 밀집도와 범죄 예방 설치물 대비 비행청소년 및 학교폭력 신고 빈도 분석” 이었습니다.

이에 따라 전국의 학교폭력 / 비행청소년 신고 수와 유흥시설의 개수가 나타나있는 데이터 파일을 수집하고,

전국의 비상벨과 스마트가로등의 위치 데이터 셋을 수집하였습니다.

그리고 나중에 전처리 과정에서 전국 법정동 코드 데이터셋을 하나 더 수집하여,

총 3개의 데이터 셋을 사용하여 분석을 진행하였습니다.

이 중에서 첫번째와 세번째는 csv파일이지만, 두번째는 shp 라는 위치 데이터 확장자의 형태로 되어있는 걸 주제를 이미 정한 후에 발견하게 되었는데요, 이것때문에 전처리 과정에서 살짝 수고를 하는 일이 생겼습니다.

여기서 앞으로는 데이터를 수집할 때 확장자도 잘 보고 수집해야한다는 교훈을 얻었습니다.

5.

저희는 총 네가지 주제로 데이터 분석을 진행했는데요,

처음에는 1번과 2번을 주제로 정해놓고 분석하다가,

유의미한 결과가 나오지 않았고, 또 시간이 남기도 해서 나중에 주제 3번과 4번이 추가되었습니다.

주제 관련 설명은 분석 과정에서 다루도록 하겠습니다.

6.

네 두번째로는 데이터 전처리 과정에 대해서 설명드리도록 하겠습니다.

7.

ppt에는 간결하게 적었지만, 사실 저희가 이 프로젝트를 총 12일 진행했는데 전처리 한 과정에만 4일이 걸렸습니다. 그 만큼 전처리가 데이터 분석 과정 중 시간이 많이 소요되는 과정이라는 걸 배울 수 있었는데요,

<STEP 1>

첫번째로는 저희가 수집한 데이터를 정제해주었습니다.

그 과정에서 rawdata의 컬럼명, 즉 변수명을 저희가 알아보기 편하게, 그리고 가공하기 쉽게 바꿔주는 작업을 가장 먼저 진행했습니다.

그리고 파일 확장자 변환을 진행해주었는데요, 앞에서 데이터셋을 소개할 때 말씀드렸던 shp파일 을 구글링을 해서 방법을 찾아서, R에서 데이터를 열고 그걸 다시 csv 파일로 바꿔서 저장했습니다.

그리고 모든 데이터 셋을 R로 불러온 이후에는, 학교폭력 데이터셋의 행정구역은 ‘시군구’ 기준으로 되어있고, 비상벨 데이터셋의 행정구역은 ‘시”와 ‘법정동’ 기준으로 분류가 되어있는 것을 확인했습니다.

또한, 비상벨 데이터 셋 중대한 결함이 있는 것을 발견했는데요, 상위 행정구역명이 “서울시”와 “세종시”를 제외하고는 전부 데이터 가공 회사명으로 추측되는 “이투온”이라는 값으로 들어가 있는 것을 확인했습니다.

따라서 82000여개의 관측값들을 전부 분류해줄 수 없어서, 분석을 “서울특별시”만 진행하기로 결정했습니다.

그리고, 나중에 두 데이터셋을 병합해야하니 행정구역 기준을 정립해줄 필요가 있었습니다.

따라서 기준을 학교폭력 데이터 셋에 맞춰서 ‘시군구’로 잡기로 결정하고, 비상벨 데이터셋에 파생변수를 하나 더 만들어서, 해당 법정동이 어느 ‘시군구’에 속하는지를 알아낸 값을 추가해주기로 했습니다.

<STEP 2>

이를 위해서, 처음에는 서울특별시는 구가 25개 밖에 안되기 때문에 수기로 값을 넣어주려고 했습니다…

하지만 비상벨 데이터는 법정동 기준으로 되어있고, 학교폭력 데이터는 행정동 명도 섞여있어서 정확히 분류하기가 어려웠고, “신사동”이라는 법정동이 강남구에도 있고 은평구에도 존재해서 “신사동”만 보고 이 비상벨이 어느 구에 속하는지 알기 어려웠습니다.

이를 위해 처음에 shp 위치 데이터 파일 원본을 가져와서, 그 데이터 파일은 좌표를 기준으로 해당 데이터가 어디있는지를 알려주는 데이터이기 때문에, R에서 그래프를 이용하여 실제로 “신사동”에 있는 비상벨 데이터를 나타내서 위도와 경도를 파악하고, 네이버에서 은평구 신사동과 강남구 신사동의 위도와 경도를 검색해서, 결국 어느 데이터가 어느 구에 속하는지 분류해주었습니다.

아니 그런데 말입니다…

지금까지는 비상벨 데이터 셋에 좌표라고 생각했었던 8자리 숫자가 법정동 코드라는 걸 이 모든 과정이 지나간 이후에 알게되었습니다..

그래서 이제는 코드를 기준으로 모든 행정구역을 알 수 있게 되었기 때문에

저희는 전처리 방향을 바꿔서, 전국 법정동 코드가 나와있는 데이터를 수집해서,

왼쪽에서 부터 5번째 자리가 ‘구’를 나타내는 숫자임을 알아냈습니다.

그래서 법정동 코드 데이터와 비상벨 데이터의 코드를 문자열로 바꿔서, 앞에서 부터 5자리 숫자까지만 잘라주었습니다.

그럼 이제 다른 데이터셋에 같은 데이터를 포함하고 있는 컬럼이 생겼으니, 이 코드 컬럼을 기준으로 merge 함수를 이용해서 데이터를 병합해 주었습니다!

그리고 이제 전처리가 거의 다 끝나간다고 생각했지만,

지금까지는 빙산의 일각이었습니다.

병합된 데이터를 요약해보니, 저희는 비상벨 데이터를 기준으로 병합을 해주었기 때문에 기존 비상벨 데이터와 새로 병합된 데이터의 행의 개수가 똑같아야 하는데, 딱 1개가 차이나는 것을 발견했습니다.

알고보니, 도봉구가 빠져있었고, 도봉구가 빠진 이유는 서울시와 세종시까지는 똑바로 분류가 되어있는 줄 알았는데, 서울시의 행정구역인 것들도 상위 행정구역명이 “서울시”가 아니라 “이투온”으로 들어가있는 것을 발견했습니다….

그래서 서울특별시만 필터링해줬던 걸 없애고, 다시 rawdata를 가져와서 다시 코드를 정제하고, 데이터를 병합해주었습니다.

그런데 이제 저와 종원님은 이 rawdata를 신뢰할 수 없어졌고, 결국 데이터가 중복되는 것이 있는지부터 위도와 경도를 이용해서 다시 확인해주었습니다.

표본을 뽑아서 확인해 본 결과, 위도와 경도가 둘 다 똑같은 중복데이터는 없다고 추론할 수 있었고, 또한 혹시 위도/경도가 같다고 해도, 비상벨/스마트가로등 객체는 두 개 이상일 수 있음을 확인했습니다.

그래서 프로젝트를 엎고 다시 하지 않고, 계속 진행할 수 있었습니다…

그런데 말입니다…

유흥시설 개수 데이터에 결측치가 너무 많은 것을 확인했습니다.

구글링도 많이 해보고 어떻게 결측치를 처리할지, 고민해 본 결과 너무 애매하다는 결론이 나와서 결국 데이터 제공 측에 결측치에 관련해서 문의해봤고, 0이라는 답변을 받아서, 결측치는 전부 0으로 처리하였습니다.

그리고 마지막으로 비상벨 데이터와 학교폭력 데이터셋을 ‘시군구’를 기준으로 병합하여

분석에 쓸 최종 데이터 셋을 만들어주었습니다.

8.

이렇게 전처리가 끝나고, 분석에 들어가기 전,

저희가 쓰는 종속 변수와 독립 변수에 대해서 소개해드리겠습니다.

종속 변수에는 ~

독립 변수에는 ~ 가 있습니다.

9.

이제부터 분석 과정에 대해 설명드리겠습니다.

10.

주제 1번에 대한 가설은 이러합니다.

“비상벨 개수가 많은 지역일수록 청소년 비행 또는 학교폭력 신고 빈도가 적을 것이다.”

이에 따른 분석 방법으로는 단순 선형 회귀를 선택해주었습니다.

분석 방법에 따른 귀무가설과 대립가설은 비상벨 개수와 비행 또는 학교 폭력 신고 수는 선형 관계가 없다., 그리고 있다 입니다.

이번 주제를 분석하기 위한 전처리로는 최종 데이터 셋에서 서울특별시만 필터링해서 진행해주었습니다.

11.

다음은 독립변수의 잔차에 대한 정규성 검정입니다.

슬라이드 내용 읽기.

12.

다음은 ncvTest를 이용한 잔차의 등분산성 검정입니다.

슬라이드 내용 읽기

13.

단순 선형 회귀 모델 분석 결과는 이러합니다.

슬라이드 내용 읽기

14.

~

이번에는 유흥시설 개수와 청소년 범죄 신고 빈도를 비교해보겠습니다.

가설 : ~

분석 방법 : ~

분석 방법에 따른 귀무가설과 대립가설은 ~

이를 위한 전처리는 ~

15.

저희는 step 함수를 이용해서 단계적 회귀분석을 해주었는데요,

슬라이드 읽기

왼쪽은 모든 독립변수들을 넣은 초기 모델, 오른쪽은 함수의 결과로 나온 회귀 모델입니다.

16.

이에 따른 결과는 이렇게 해석할 수 있습니다.

슬라이드 읽기

17p.

다음으로 전국의 행정구역 별 유흥시설 개수에 따른 비상벨 설치 빈도를 분석했습니다. 이번 주제에서는 기존에 독립변수로 쓰였던 비상벨 설치 빈도를 종속변수로 사용하여 의미 있는 결과를 도출해보고자 했습니다.

18p.

단순 선형 회귀 분석을 통해 ‘유흥 시설 개수가 많은 지역일수록 비상벨의 개수가 많을 것이다’ 라는 가설을 밝히고자 했습니다.

여기서 독립 변수인 유흥 시설의 개수는 유흥 시설 개수들의 합입니다.

전처리 과정에서 유흥시설 중 0인 값이 많았던 무인텔, 생활숙박시설, 여행숙박시설은 제외했습니다.

또한 독립변수들이 각기 다른 단위를 가지는 것으로 판단하여 정규화 작업을 진행했습니다.

이후 apply 함수를 사용하여 전체 유흥시설의 총합을 나타내는 파생변수를 생성했습니다.

19p.

회귀 모델의 유의확률은 0.6으로 유의 수준 0.05에서 유의하지 않으므로, 해당 회귀 분석 모델은 폐기했습니다.

20p.

마지막 주제로 전국 행정구역 별 유흥시설 개수에 따른 청소년 비행 및 학교 폭력 신고 빈도를 분석했습니다.

21p.

‘유흥시설 개수가 많은 지역일수록 청소년 비행 또는 학교 폭력 신고 빈도가 많을 것이다’ 라는 가설을 세우고 다중 선형 회귀를 통해 분석했습니다.

먼저, 독립변수에서 0값이 많은 무인텔, 생활숙박시설, 여행숙박시설의 개수는 제외했습니다.

그리고, 주제 2번에서 분석할 때 서울특별시로 제한했던 범위를 전국으로 넓혀 분석했습니다.

22p.

종속변수가 청소년 비행 신고 빈도인 모델의 경우, step 함수를 적용한 결과, 유의하지 않은 독립변수인 단란주점, 유흥주점 개수가 제거되었습니다.

AIC가 가장 낮은 회귀 모델의 유의확률은 0으로 유의수준 0.05에서 유의한 모형을 갖는다고 할 수 있습니다.

유의한 독립변수는 총 3가지이며, 회귀 모델의 설명력은 0.65로 총 3개의 독립변수들이 종속변수의 약 65%를 설명하고 있습니다.

회귀 계수를 보시면 단란 주점과 클럽의 경우 두 모델 모두 음의 상관관계를 갖고, 나머지 변수들은 양의 상관관계를 갖습니다.

이는 단란 주점과 클럽이 많은 지역은 청소년이 많이 거주하지 않아서 나타나는 현상으로 보입니다.

23p.

종속변수가 학교 폭력 신고 빈도인 모델의 경우, step 함수를 실행할 결과, 모든 독립변수가 유의했습니다.

AIC가 가장 낮은 회귀 모델의 유의확률은 0으로 유의수준 0.05에서 유의합니다.

회귀 모델의 설명력은 0.48로, 총 5개의 독립변수들이 종속변수를 약 48% 설명하고 있다.

24p.

독립 변수끼리의 강한 상관관계가 존재하는지 확인하기 위해 다중공선성 검정을 실시했습니다.

‘청소년 비행 신고 빈도’ 를 종속변수로 하는 모델의 경우, vif 다중공선성 검정 결과, 모든 독립변수의 분산 팽창 요인 값이 5 미만이므로 다중공선성이 없습니다.

‘학교 폭력 신고빈도’ 를 종속변수로 하는 모델의 경우, vif 다중공선성 검정 결과, 모든 독립변수의 분산 팽창 요인 값이 5 미만이므로 다중공선성이 없습니다.

25p.

정규성 검정을 위해 QQ plot을 그렸습니다. 점선에서 많이 벗어난 것으로 보아 정규성을 만족하지 않는 것으로 보입니다.

Shapirpo 테스트 결과 유의 확률이 0으로, 유의 수준 0.05에서 귀무가설을 기각합니다.

따라서 주제 4번의 모델은 정규성을 만족하지 않습니다.

26p.

각각 종속변수가 청소년 비행 신고 수, 학교 폭력 신고 수일 경우 등분산성 검정 그래프입니다. 잔차가 한쪽에 몰려 있는 것으로 보아 등분산성을 만족하지 못하는 것으로 보입니다.

ncvTest 등분산성 검정 결과, 유의확률이 모두 0으로 유의수준 0.05에서 귀무가설을 기각합니다. 따라서 모두 이분산성을 띕니다.

정규성 및 등분산성을 만족하지 못했지만, 데이터의 개수가 충분이 많기 때문에, 각 가정들을 만족한다고 가정하고 분석에 대한 결론을 내렸습니다.

27p.

저희 분석에 대한 결론입니다.

28p.

저희는 유흥 시설은 청소년의 범죄 빈도에 영향을 끼친다는 결론을 내렸습니다.

단란 주점과 클럽이 많은 지역은 청소년이 많이 거주하지 않아, 오히려 청소년 범죄와 음의 상관관계를 갖는 것으로 보입니다.

성인 게임장과 술집, 유흥 주점이 많은 지역은 청소년 범죄와 양의 상관관계를 갖습니다.

따라서 청소년이 밀집되어 있는 지역에는 유흥 시설에 제한을 두고, 유흥 시설 밀집 구역에 청소년이 쉽게 접근할 수 없도록 해야 합니다.

29p.

저희 분석에는 몇 가지 한계점이 있습니다.

먼저, 유흥 시설이 개장 및 폐장되고 어느 정도의 기간이 지나야 청소년 범죄 빈도에 영향을 끼치는지 알 수 없습니다.

예를 들어, 유흥 시설이 개장하고 1년 뒤에 청소년 범죄에 영향을 끼친다고 한다면, 현재 모델로는 결론에 대해 충분한 근거가 되지 않을 수 있습니다.

두 번째로, 학교 폭력 데이터는 2021년 한 해의 신고 통계이고, 비상벨 데이터는 2022년을 기준으로 수집한 데이터입니다.

따라서 학교 폭력 데이터에 영향을 끼치지 않는 비상벨 데이터, 즉, 22년도에 새로 설치된 비상벨에 대한 데이터가 포함되어 있으며, 이 또한 분석 결과에 영향을 끼칠 수 있습니다.

마지막으로, 유의했던 모델이 기본 가정 중 정규성과 등분산성을 만족하지 못했습니다.

관측 값의 개수가 충분하여 만족한다고 가정하고 진행했지만, 다른 분석 방법을 사용하는 등 더 나은 방식으로 풀어나갈 수 있지 않았을까 하는 아쉬움이 있습니다.

30p.

질문 받는 시간을 갖겠습니다. 질문하실 분 계신가요?

31p.

이상으로 데면대면 팀이었습니다. 감사합니다.