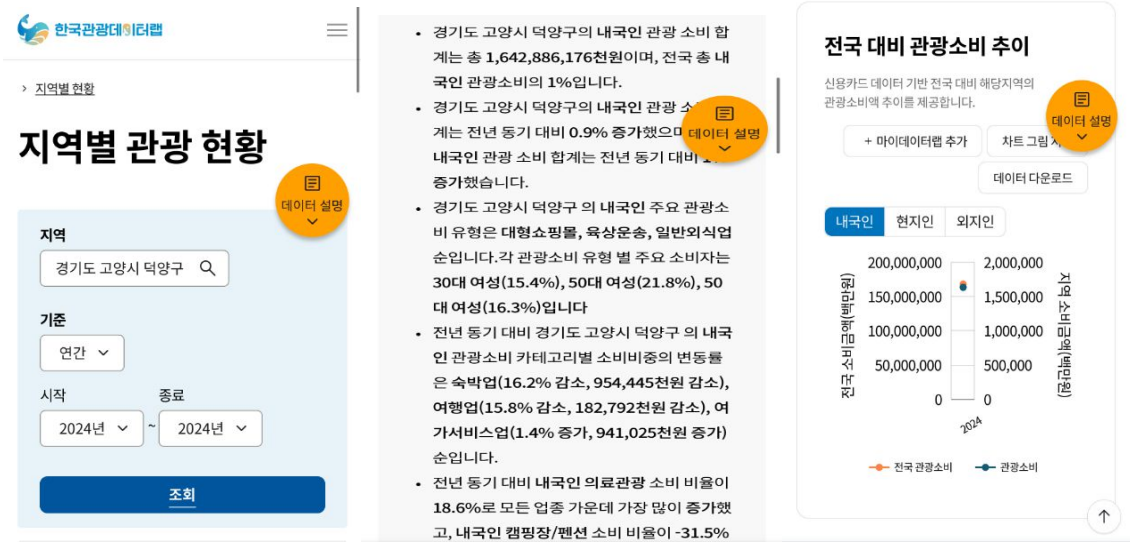
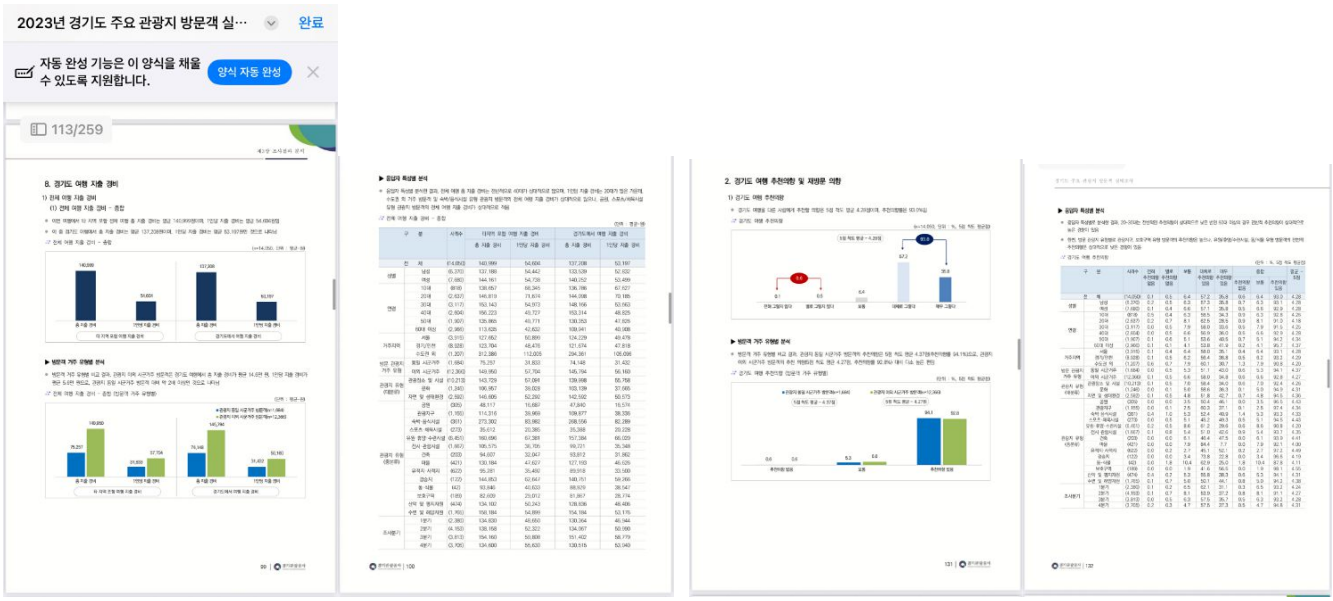
상관관계 분석 주제: 지역별 관광지출액과 재방문율의 관계 분석

[데이터 수집]

<한국관광 데이터랩 (한국관광공사운영 관광 빅데이터 플랫폼)>
- 지역별 관광 현황: '관광소비(내국인 기준)' 지표에서 지역별 관광지출액 관련 데이터 (신용카드 데이터 등을 활용한 추정 방식 가능)



+ 2023년 경기도 주요 관광지 방문객 실태조사 결과보고서
  (경기도 여행 지출 경비/ 경기도 여행 추천의향 및 재방문 의향)



**데이터 수집해서 활용시, 두 변수가 직접적으로 연관되는 데이터셋을 찾기 어려울 수 있음. 이 경우, 서로 다른 출처의 지역/기간 단위 데이터를 수집해서 '지역 코드'나 '시간(분기/연도)'를 기준으로 결합해서 사용 시도

[분석 및 발표 방향] Rscript 활용

1. "변수 표준화/정규화" 데이터 결합
2. "시각화" 변수별 통계(평균/표준편차), 산점도 그려서 두 변수 관계 시각화
3. "상관관계 분석" 피어슨 상관계수(r) 계산 및 유의성 검증(p-value)
4. "결과 해석" 상관관계의 크기/부호를 통해 두 변수의 선형적 관계 해석
5. "예측 모형 제시" 지출액 변화에 따른 재방문 변화 예측 모형 제시 가능

확률분포 분석 주제: 대학생들의 시험기간 '카페 체류 시간' 분포 분석

[데이터 수집]

〈설문〉
1) 온라인 설문조사 : 대규모 데이터 확보 & 시간/비용 효율적이지만, 정확성(기억에 의존)문제와 체류 시간의 구체적인 분포 분석 한계
2) 현장 관찰/설문(소규모): 정확성 확보 & 분포 형태 정확하게 파악 가능, 시간/인력 많이 소요됨. 표본 크기 작음

+ 필요한 수집 변수
1. 주요 변수(연속형): 카페 체류 시간 (시간 또는 분 단위)
2. 보조 변수(분류형): 성별 / 학년 / 전공 유행 / 주로 공부하는 장소

[프로그램] Rscript

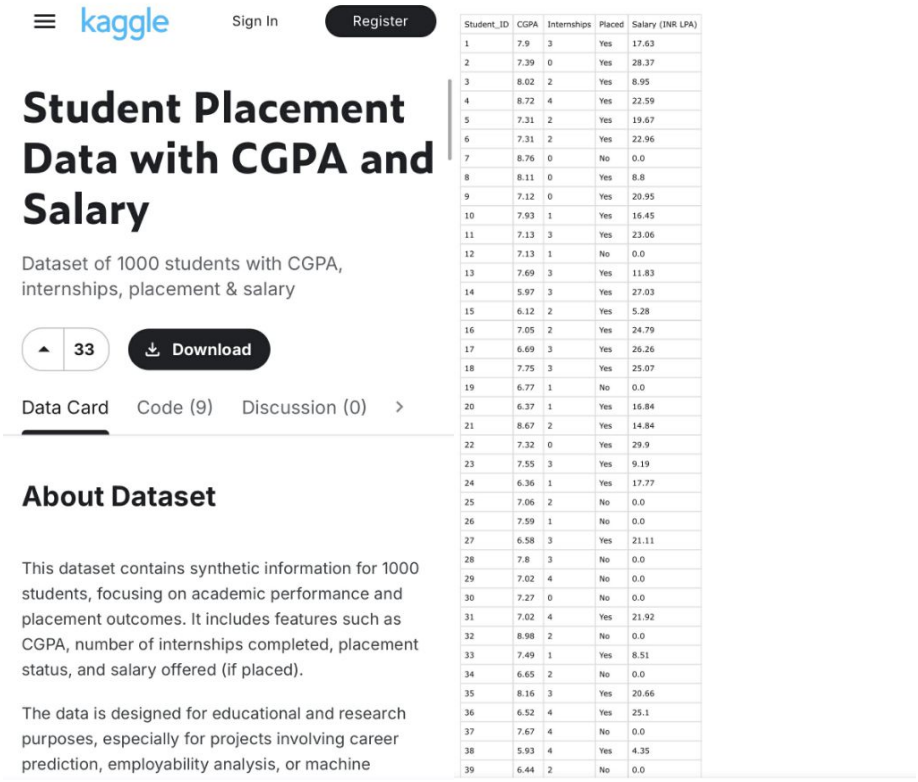
목표: 수집된 체류 시간 데이터가 어떤 확률분포를 따르는지 확인하고 모델링

1. "체류 시간의 확률적 모델링" 시험 기간 카페 이용 현황 파악 목표 제시
2. "데이터 수집 및 처리" 설문 데이터 수집 방법 설명, 표본 크기, 체류 시간의 기본 통계치
3. "확률분포 모델링" 히스토그램으로 분포 형태 제시 (감마분포 선정 및 모수(parameter) 추정 결과 제시)
4. "적합도 검정 결과" 통계적 검정 결과(p-value)를 통해 해당 분포가 데이터를 잘 설명함을 입증
5. "결론 및 응용" 해당 분포 모델을 통해 특정 시간 체류할 확률 계산하여 제시. 이후, 이 모델이 카페 운영 전략이나 학습 코칭에 어떻게 활용될 수 있는지 실질적인 응용 방안 제시.

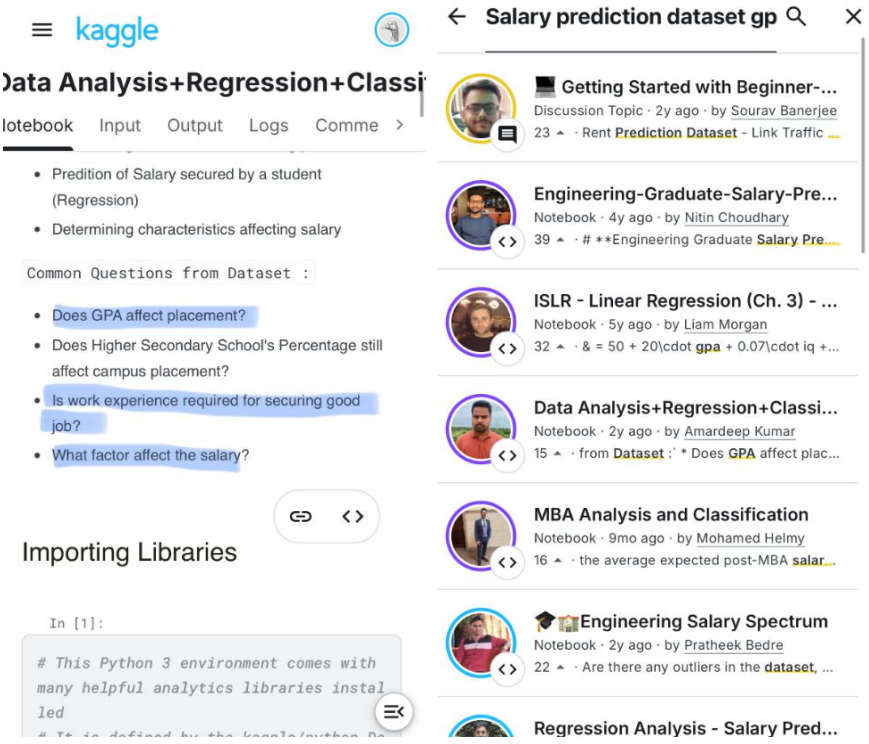
예측 모형 모델링 주제: 대학생 졸업 후 연봉 예측

[데이터 수집]

〈Kaggle〉
- Student Placement Data with CGPA and Salary 을 통해
졸업생의 채용 및 초기 연봉 정보가 포함한 데이터 수집
- Salary Prediction Dataset 등 일반적인 연봉 예측 데이터셋 중 '학력(Education Level)' 및 '경력(Years of Experience, O년 근접 데이터)' 변수가 포함된 것을 활용. (이때 데이터 종류 보고 '졸업 후 초봉에 초점을 맞춰 분석 범위 한정도 가능)



Student_ID	CGPA	Internships	Placed	Salary (INR LPA)
1	7.9	3	Yes	17.63
2	7.39	0	Yes	28.37
3	8.02	2	Yes	8.95
4	8.72	4	Yes	22.59
5	7.31	2	Yes	19.67
6	7.31	2	Yes	22.96
7	8.76	0	No	0.0
8	8.11	0	Yes	8.8
9	7.12	0	Yes	20.95
10	7.93	1	Yes	16.45
11	7.13	3	Yes	23.06
12	7.13	1	No	0.0
13	7.69	3	Yes	11.83
14	5.97	3	Yes	27.03
15	6.12	2	Yes	5.28
16	7.05	2	Yes	24.79
17	6.69	3	Yes	26.26
18	7.75	3	Yes	25.07
19	6.77	1	No	0.0
20	6.37	1	Yes	16.84
21	8.67	2	Yes	14.84
22	7.32	0	Yes	29.9
23	7.55	3	Yes	9.19
24	6.36	1	Yes	17.77
25	7.06	2	No	0.0
26	7.59	1	No	0.0
27	6.58	3	Yes	21.11
28	7.8	3	No	0.0
29	7.02	4	No	0.0
30	7.27	0	No	0.0
31	7.02	4	Yes	21.92
32	8.98	2	No	0.0
33	7.49	1	Yes	8.51
34	6.65	2	No	0.0
35	8.16	3	Yes	20.66
36	6.52	4	Yes	25.1
37	7.67	4	No	0.0
38	5.93	4	Yes	4.35
39	6.44	2	No	0.0



(다양한 전공/기준 등으로 분류한 데이터 검색됨)

+ 주요 변수 (예상)
1. 졸업 후 연봉 (Salary) : 예측의 목표(연속형 변수)
2. 학업: 최종 학위/전공(Educational Level), 학점/GPA, 인턴십 횟수 : 학업 성과가 초봉에 미치는 영향을 분석하기
3. 인적/경력: 성별(Gender), 경력 연수(Years of Experience) (초봉 예측이면 0~2년 사이의 데이터로 좁히면 될듯), 직무/직책(Job title) : 개인 배경 및 직무에 따른 차이를 분석하기
4. 기타: 지역(Location), 회사 규모(Company Size) 등 : 환경적 요인 있음을 인지 (데이터에 따라 상이)

[프로그램] Rscript (dplyr 전처리, ggplot2 시각화, caret/tidymodels 모델링 등 활용)

회귀 문제

1. "데이터 전처리" (일단 이상치 상의해서 제거 필요할듯) 범주형 변수(전공, 직무)를 데이터 변수 변환 (프로그래밍 방법을 잘 모르겠음..)
2. "기술통계 데이터 분석, 시각화" 연봉 분포 확인. Box Plot/ Scatter Plot 등을 통해 주요 독립변수와 연봉 간의 관계 (경향성) 시각화
3. "모델 선택 및 학습" 데이터를 훈련 세트와 테스트 세트로 분리해서 다양한 회귀 모델을 적용해 비교. (ex: 선형 회귀, 릿지/라쏘 회귀, 램덤 포레스트 회귀)
4. "평가" 예측 성능 평가 (MAE 평균 절대 오차, MSE/RMSE 평균 제곱 오차, R제곱- 결정 계수)
5. "최적 모델 선정" 평가 지표가 가장 우수한 모델을 최종 선택하고, 해당 모델의 변수 중요도를 해석

+ 발표 구성

1. 배경 및 목표 어떤 요인이 졸업 후 연봉에 가장 큰 영향을 미치는지 예측 모델 통해 확인
2. 데이터 설명: 사용한 Kaggle 데이터셋의 출처 및 주요 변수 설명
3. 모델링 과정: 데이터 전처리, 훈련/테스트 분리, 사용된 예측 모델 설명
4. 결과 분석: 평가 지표 비교, 최적 모델의 예측력, 변수 중요도를 학술적으로 해석

+ 예상 시사점
1. 선형성 VS 비선형성: 연봉 결정 요인이 단순하지 않으며, 복잡하고 비선형적 관계를 가짐을 시사
2. 실용적 조언: 분석 결과를 통해 대학생들에게 실질적인 조언 제시 가능