|  |  |
| --- | --- |
| 교육 제목 | **2021.09.15 Daily Assignment** |
| 교육 일시 | 2021-09-15 |
| 교육 장소 | 자택 |
| **교육 내용** | |
| 오전 | * 1. 9.14에 배운 깃 내용 복습   Git을 이용하여 원격 저장소 생성부터 로컬 저장소와 연결 이후 commit, add, staging 작업을 다시 한번 진행했고 conflict가 발생 했을 때 해결 방법을 공부 했습니다.  이후 branch 생성 및 관리, 충돌 해결을 다시 복습 했습니다.   * 1. 데이터 사이언스와 통계학의 차이   컴퓨터 프로그래밍, 모집단이냐 샘플링이냐  순수한 통계학은 프로그래밍 스킬이 필요 X  현재는 데이터의 사이즈가 매우 커진 빅데이터를 다루며 이에 따라 일반 유저에게도 이를 잘 보여 줘야 하는 시각화 컴퓨팅 비전이 발전했다. 때문에 단순 패키지가 아닌 프로그래밍으로 구현을 해야 한다.   * 1. 기초 통계학   자료의 종류  정형 데이터  연속형 자료  등간척도: 절대 값 0이 존재 X, 곱셈 법칙이 적용X  Ex) 온도: 0도는 물이 어는 지점이지 물리적으로 0인 값이 아니다.  비율척도: 절대 값 0이 존재  Ex) 키, 몸무게, 혈압: 물리적으로 존재 하지 않는 값이 존재할 수 있으며 배수가 적용된다.  범주형 자료  명목척도: 속성을 분류하는 척도, 속성간 순위가 없다.  Ex)남/녀  순서척도: 속성간에 순위 개념이 존재하게 된다.  Ex)상/중/하  비 정형 데이터  그림, 글, 단어(NLP), Vision  자료의 축약  모집단과 표본 집단  모집단   * 이해하고자 하는 집단 전체, Ex) 우리나라 고등학생 평균 키, 고혈압 환자   표본집단   * 모집단에서 추출된 집단, Ex) 추출된 고등학생 100명, 연구에 참여한 고혈압 환자 50명   데이터의 구조와 크기를 보고 분석을 방법을 정하는 것이 중요하다.  범주형 자료는 각 범주가 가진 값의 크기에 관심이 있다면 도수 분포표는 상대적인 비율, 값, 그래프의 모양 자체에 초점이 맞추어진다.  그래프를 통한 자료의 요약의 장단점  장점: 자료를 한눈에 알아 볼 수 있다.  단점: 그림의 모양이 작성자의 주관적 판단에 따라 달라 질 수 있다.  수치를 통한 연속형 자료의 요약   * 표본 평균(Sample Mean)   중심을 나타내는 측도 중에서 가장 많이 사용되는 방법  자료의 무게 주심을 나타냄  자료의 이상치에 영향을 많이 받음   * 중앙 값(Median)   전체 관측 값을 크기 순으로 나열 했을 때 중앙에 위치한 값  자료의 이상치에 영향을 적게 받음   * 최빈 값   관측 값 가운데 가장 자주 나온 값  표본 평균, 중앙 값, 최빈 값의 비교  표본 평균은 이해하기 쉽고 이론적 전개가 용이  표본평균은 전체 관측치를 반영하지만 이상치에 영향을 받음  중앙 값 중앙 부분의 관측치에 영향을 받고 이상을 받지 않 음  이상치들이 있는 경우 표본 평균과 중앙 값을 적절히 사용  퍼짐의 정도  분산과 표준편차 (Variance and Standard deviation, SD)  자료가 중심으로부터 얼마나 퍼져있는지를 표현 수치  사분위수 범위 (IQR): 3사분위수 - 1분위수(Q3-Q1)  상방 기준: Q3 + 1.5\*IQR  하방 기준: Q1 – 1.5\*IQR |
| 오후 | 1. 코딩 테스트 참여로 교재 참고 + 5시 이후 강의 내용만 기록   표본 공간: 한 실험에서 나올 수 있는 결과들의 모임  근원 사건: 표본공간을 구성하는 개개의 결과  사건: 표본 공간의 부분집합으로서 근원 사건들의 집합  확률: 동일한 실험이 무한히 반복했을 때 나타나는 사건의 상대돗수의 비   1. 확률 분포   확률 변수  실험의 결과가 확률적으로 나타나는 실험에서 얻어진 결과를 실수 값에 대응시킨 결과  확률 분포  확률 변수가 갖는 값들과 그에 대응하는 확률 값 |