-eda는 탐색적 데이터분석

eda를 하기 위해서는 데이터가 수집되어 있는 상태에서

이거 할때는 데이터셋이 어떤 타입으로 구성되어 있는지 값의 분포가 어떻게 되어 있는지

특히 정형데이터를 받아올땐 각각의 변수들의 타입 분포등도 알아야 한다.

이상치가 결측치가 있는지도 파악한다. 이럴때는 적당한 기술통계량과 관련된

박스플롯 산점도 를 통해 시각화를 통해 값의 전반적인 분포와 이상치를 파악하는 분석을 탐색적 데이터 분석이라고 한다.

-데이터수집은 오픈API - 그딜이 제공한 url가지고 하는거. 정적 크롤링 API를 써서 오픈 크롤릴ㅇ은 . 동적 크롤링은 어떤때에 계획해야 하는지. 동적크롤링 하려면 필요한 패키지가 무엇인지.

-수치데이터랑 -- 중간값도 알려주고 -- MAX MIN 을 한번에 사용할 수 있는

-데이터수집할때 많이 썼고, 일반적인 데이터 전처리 할때도  특히 수집할때 url문자열이 고정내용이 아니라 일부가 그때그때 바껴야 하는 경우에 -- 반복문을 쓰면서 특정한 함수 기억하기.

-평균이 뭔지 분산이 뭔지 표준편차가 뭔지

-데이터셋의 특징에 대한 문제 2개

벡터, /메트릭스 /데이터 프레임/ 리스트

데이터프레임은 벡터들이 열다뉘로 모여있는거.

특징과 어떤경우에 쓰는지 기억하기

리스트는 데이터들을 원하는대로 놓을 수 있고 심지어는 데이터셋도 집어 넣을 수 있다.

벡터는 저장되는 데이터의 갯수는 무한대지만 데이터타입이 동일해야 한다.  다ㅡㄹ때는 리스트 써야한다.

벡터에서 데이터 인덱싱할때 인덱스를 1부터 시작하는 숫자주는에 양./음 값을 줄 수 있다. 음의 값을 줄때 어떻게 적용되는지/

벡터 만들때 C함수, rep, seq 차이점과 특징 파악하기

-함수 R 함수의 특징

-기본그래프쪽에서 바/파/히스토/바플롯/박스플롯/ 특징

히스토그램

범주형데이터 이산형 데이터 값의 분포 - 데이터갯수를 세서 바그래프나 비율을 적용해서 파이 그래프 사용한다.

연속형데이터의 경우 연속형의 최대 최소 값 2,3 분위 등 값의 분포를 체크하고자 할 때 박스데이터

-회귀분석에서는

독립변수가 뭔지 종속변수가 뭔지 기억하기.

회귀분석에 의해서 만들어진 회귀식 -- 평가할 때 사용하는 값

-ggplot2 패키지 - 레이어링 방식ㅇ로 표현. 밑그림을 그리고 바그래프, 산점도 등을 그려나간다. 그럴때 사용해야 하는 기호 같은거.

-시각화 패키지 공부한거 리플랫, 워드클라우드2 gg맵 패키지 역할