상명 AI+X 선도인재양성

EDA 분석 프로젝트

만나서 반갑습니다!

현) KT 그룹인재개발실
AI 코치/전문강사

경력 사항

- KT 사내/외 AI 강사/코칭 수행 (데이터분석, AI원팀교육 등)
- KT AI 전문강사
- 현장 AI 300 프로젝트 수행
- AIFB Associate/Professional 출제 위원

자격 사항

- CCNA(Cisco Certified Network Associate)
- Google TensorFlow Certificate Developer
- AIFB Associate 外 다수

학습목표

- EDA 뭔지 알아 보기
- EDA 분석하는 방법 배우기
- EDA 데이터 분석 실습해 보기

학습내용

- EDA 정의
- EDA 기본 개요(속성,관계)
- EDA 분석 방법(시각/비시각)
- 실습데이터로 EDA 분석하기

숙제 다운로드 URL

https://url.kr/wylueh

- 교재, 실습 Jupyter notebook, 지하철 이용승객 데이터, 지하철 노선 정보데이터

EDA
탐색적 데이터 분석

머신러닝 절차에서 EDA 위치

데이터 제일중요하고 이해해야 한다.

데이터 안에 우리가 모르는 많은 정보와 특징이 있다.

- 트위터,페이스북,인스타그램 ,아마존 기업이 사활 걸고 데이터 모운다.
- 수집된 데이터 를 분석, 유지하는데 많은 인력과 비용 투자한다.
- 왜? 여러분 남긴 데이터에서 특징과 트렌트 파악, 인사이트 도출 및 비즈니스 연계
  → 수익 창출

데이터 없이는 머신러닝도 쓸모없다.

- 머신러닝은 데이터를 가지고 학습하기에 데이터가 없으면 학습 불가

Q) 토익성적 데이터에 대한 질문

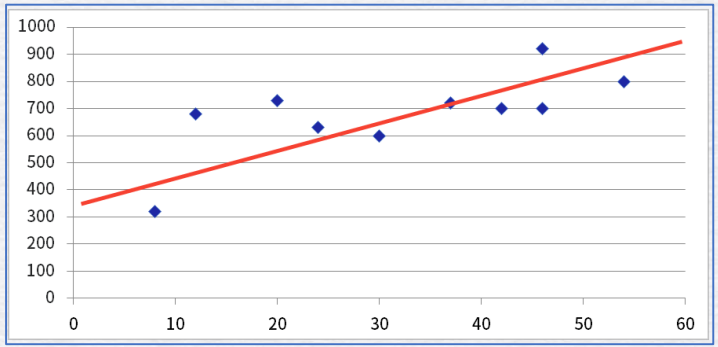
학습시간(시간)	토익성적(점)
54	800
8	320
30	600
24	630
46	700
12	680
20	730
37	720
42	700
46	920

- 1. 학습시간과 토익성적에 대한 평균값, 최소값, 최대값은?
- 2. 가설) 공부시간이 많을수록 토익성적이 잘 나온다?
- 3. 예측) 50시간 공부하면 토익성적 몇점을 맞을까?

A) 토익성적 데이터에 대한 질문

학습시간(시간)	토익성적(점)
54	800
8	320
30	600
24	630
46	700
12	680
20	730
37	720
42	700
46	920

1. 학습시간과 토익성적에 대한 평균값, 최소값, 최대값은?
   - 기술 통계량으로 평균, 최소, 최대 구할수 있다.
2. 가설) 공부시간이 많을수록 토익성적이 잘 나온다?
3. 예측) 50시간 공부하면 토익성적 몇점을 맞을까?



- 2. 산점도
- 3. 선형회귀

EDA 뭐니?

탐색적 데이터 분석(EDA : Exploratory Data Analysis)

- 앞쪽의 토익점수 데이터의 숫자들을 한눈에 파악하기 쉽지 않다. 도와 주는것이 EDA다.
- 즉, EDA는 데이터를 다양한 측면에서 바라보고 이해하는 과정
- 기술 통계적 요약, 분포 파악 및 시각화 등의 기법을 통해 직관적으로 데이터 특성 파악

EDA 제대로 못하면?

탐색적 데이터 분석(EDA) 간과할 경우

- 처음 데이터 분석을 공부를 시작할 때, 도대체 이 EDA 개념이 왜 그렇게 중요한지 모른다.
- '아니 파이썬 코드 한 줄 작성하는 것도 손에 익히기 벅차 죽겠는데, 그냥 빨리 그래프 그리거나 모델링하면 되지.' 하면서 제대로 EDA 분석없이 넘어 갑니다.
- 그런데 데이터로 실제 아웃풋을 만들어 내기 시작하면서 깨닫기 시작합니다.
- 기본적으로 데이터 자체에 대한 해석이 잘못되면, 열심히 한줄 한줄 코드를 짜며 고생해서 만든 그 데이터 프레임과 시각화한 그래프들이 그냥 휴지조각이 되고 만다는 것을 알게되며, 모델링 또한 의미 없는 결과를 도출하게 됩니다.

그럼, EDA 어쩌라는거냐?

데이터에 대한 도메인 전문가 도움 필요하다.

- 사실, 시중에 돌아다니고 있는 데이터, 예를 들어 타이타닉 데이터, 보스톤 집값 데이터등은 **토이 데이터**로 여러분들의 실력향상에 도움을 주기 위한 데이터다.
- 하지만, 실업무에서 사용되는 데이터, 예들 들어 영업 데이터, 회계데이터, 의학 데이터등은 우리가 접해보지 못한 데이터로, 파악하고 이해하기 쉽지 않다.
- 실업무에서 뛰고 있는 담당자가 해당 데이터를 만들고 관리하기에 제일 잘 알고 해당 데이터를 이해하고 있습니다.
- 따라서, 실업무 데이터를 EDA하려면, 해당 도메인 전문가와 협업하여 데이터를 이해하고 파악해야 합니다.
- 만약, 해당 도메인 전문가가 없다면, 데이터 이해하는데 많은 시간과 노력이 필요하며, 제대로 데이터를 이해하지 못할수도 있습니다.

그럼, EDA 왜 하는데?

EDA를 통해 데이터에서 트랜드 파악하고 인사이트 도출

- 데이터 표현하는 현상을 **이해**하고, 다양한 **특성과 패턴**을 발견
- 데이터 각 요소의 속성 파악하고, 데이터간의 관계 파악
- 데이터의 특징과 구조로부터 얻은 정보를 바탕으로 **인사이트** 도출
- 구글,아마존,MS,네이버등에서 수억건의 고객접속/사용이력/구매등의 데이터를 수집,정제,분석(데이터웨어하우스)하여 **인사이트 도출하고 트렌트 파악과 새로운 비즈니스 창출 → 수익 창출**
- **추가로, 기업들이 AI 왜 사활을 걸고 할까요?**

EDA 기본 개요

속성 파악

- 분석 목적 및 개별 변수 속성 파악
- 예) 가격 예측 분석 과제에서 가격 컬럼 유형 및 관측치 범위 확인

관계 파악

- 변수간의 관계 파악 및 가설 검증
- 예) 건물의 건축 연도와 가격 사이에 유의미한 영향 관계 유무 확인
- 가설) 건축연도 오래될수록, 주택가격이 떨어진다?

사전 데이터 분석

데이터 정의 확인

- 정의서 기반 데이터 확인
  - 테이블별 변수 목록, 개수, 설명, 타입 등

테이블명	HOUSE_PRICE_SEOUL				작성일	2022. 4. 1.	…
No.	컬럼명	설명	유형	타입	Null 허용	비고	…
1	PRICE	가격	연속형	DOUBLE	N	개별 공시 가격	…
2	ADDRESS	주소	명목형	STRING	N	도로명 주소	…
3	LON	경도	연속형	DOUBLE	N	경도 좌표	…
4	LTTD	위도	명목형	DOUBLE	N	위도 좌표	…
…	…	…	…	…		…	…

실 데이터 확인

- 실제 데이터 개요, 결측치, 형상 등 확인
  - head, tail, info 기반 확인
- 변수별 정의된 범위 및 분포 등 확인
  - 관측치 범위/분포 등

No.	컬럼명	설명	타입
1	PRICE	가격	DOUBLE
3	LON	경도	DOUBLE
4	LTTD	위도	DOUBLE

양수 범위

위/경도의 유효 범위

EDA 유형 구분

	일변량 (Univariable)	다변량 (Multivarable)
비시각화	빈도표 기술 통계량	교차표 상관계수
시각화	파이차트 막대그래프 히스토그램 박스플롯	모자이크플롯 박스플롯 산점도

일변량 범주형 비시각화

빈도표

› **범주별** 빈도 파악이 목적

No.	Gender	City	Age	…
1	M	Seoul	22	…
2	F	New York	13	…
3	F	London	32	…
4	M	Tokyo	43	…
5	F	Paris	51	…
…	…	…	…	…

빈도표

Gender	Frequency	Ratio
M	132	44.0%
F	160	53.3%
Missing	8	2.7%
sum	300	100%

일변량 연속형 비시각화

주요 통계 지표

› **연속형** 데이터의 대표 특징 확인

- 평균, 분산등의 **기술 통계량**
- 중앙값 등의 **사분위수**

No.	Gender	City	Age	...
1	M	Seoul	22	...
2	F	New York	13	...
3	F	London	32	...
4	M	Tokyo	43	...
5	F	Paris	51	...
...	...	...	...	...

결과

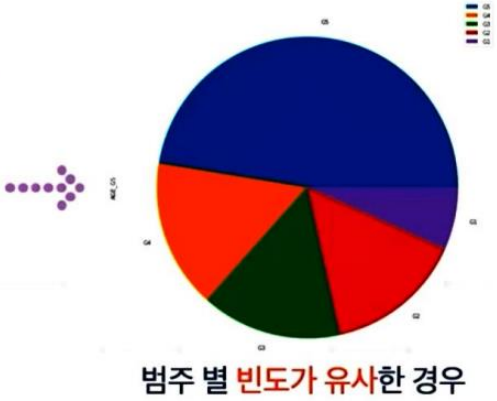
No.	Index	Value
1	Mean	53.4
2	Std	1.29
3	Variance	1.68
4	Median	52
5	Skewness	1.72
6	Kurtosis	5.05
...	...	...

일변량 범주형 시각화

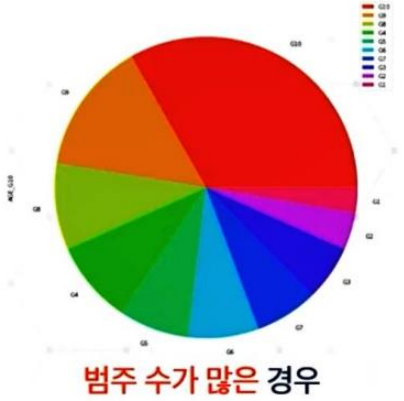
파이차트

빈도표

Variable	Frequency	Ratio
A	132	44.0%
B	160	53.3%
…	…	…
…	…	…



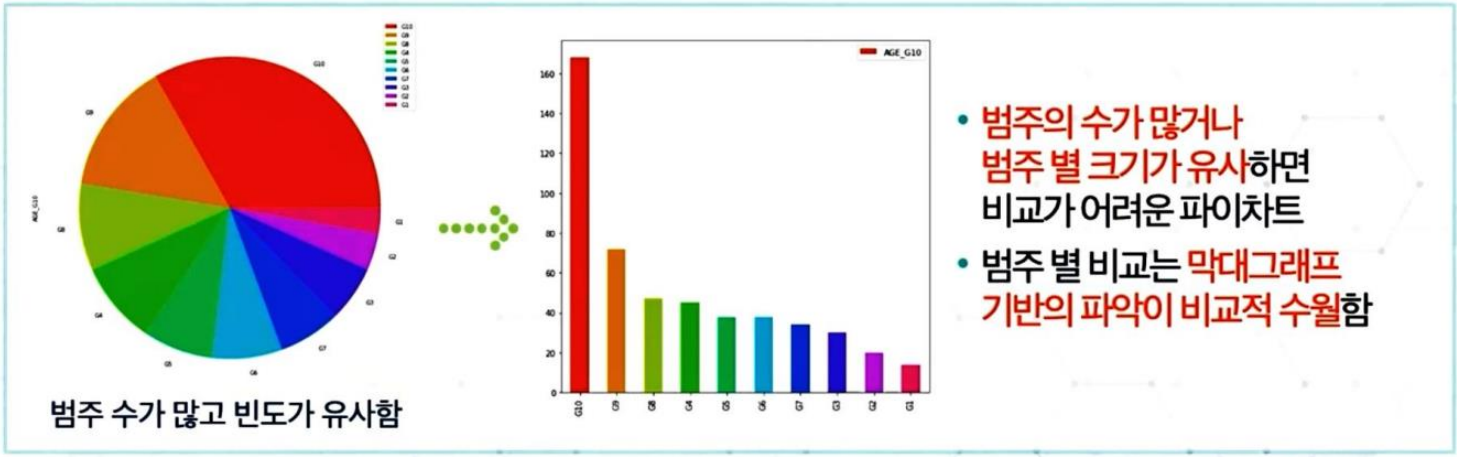
범주 별 빈도가 유사한 경우



범주 수가 많은 경우

일변량 범주형 시각화

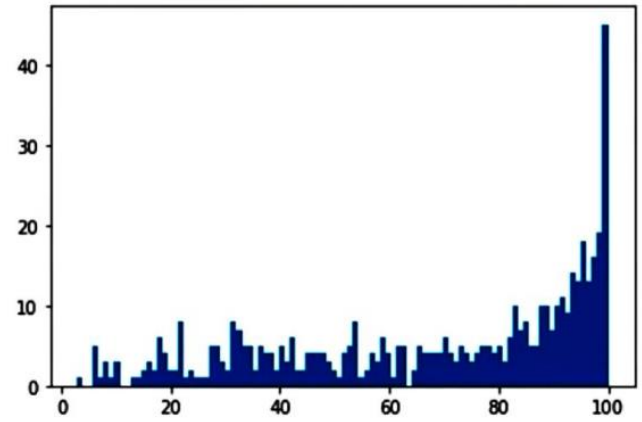
막대 그래프



- 범주의 수가 많거나 범주 별 크기가 유사하면 비교가 어려운 파이차트
- 범주 별 비교는 막대그래프 기반의 파악이 비교적 수월함

일변량 연속형 시각화

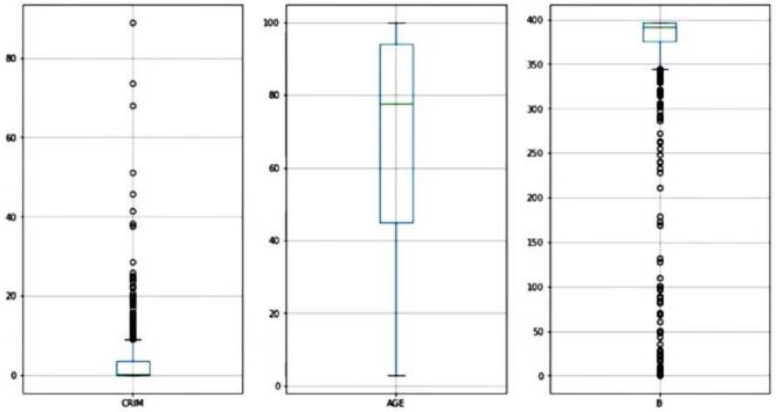
히스토그램



- 관측된 연속형 데이터 값들의 분포 파악 가능
- 구간 별 분포 상태를 쉽게 알아볼 수 있는 그래프
- 구간 내 속하는 자료의 수가 많고 적음을 쉽게 파악
- 막대그래프와 유사한 형태를 보유
  - 히스토그램은 연속형 변수의 빈도 분포를 표현
  - 막대그래프는 범주형 (이산형 포함) 변수의 빈도표 비교 표현

일변량 연속형 시각화

박스플롯



- 연속형 데이터의 양상을 직관적으로 파악할 수 있는 방안으로 5가지 요약치를 기반으로 생성되며 다양한 정보를 손쉽게 표현
  - ①중앙값 ②1분위수
  - ③3분위수 ④최대값 (IQR Value)
  - ⑤최소값 (IQR Value)
- 데이터의 개략적인 흩어짐의 형태 파악 및 IQR 기반의 이상치 판단에 용이함

다변량 비시각화

교차표(Crosstab)

- 범주형-범주형 변수간 연관 관계 파악

No.	ID	차량 보유	소득 범주	거주 지역	연령	연소득	가구 구성원	…
1	001	보유	고	서울	54	23,728	5	…
2	002	보유	중	경기	48	5,143	3	…
3	003	미보유	고	경북	32	10,567	1	…
4	004	미보유	저	전남	23	2,782	2	…
5	005	미보유	저	제주	31	2,987	3	…
…	…	…	…	…	…	…	…	…

		변수 2 (차량 보유)		합계
		보유	미보유	
변수 1 (소득 범주)	고	30 (83.3%)	6 (16.7%)	36
	중	68 (52.3%)	62 (47.7%)	130
	저	70 (29.9%)	164 (70.9%)	234
합계		168(42.0%)	232 (58.0%)	400

고소득 및 저소득 범주에서
자동차 보유 여부 구성이 두드러짐

다변량 비시각화

범주별 요약 통계량(Groupby)

› 범주형-연속형 변수 조합간 범주별 대표 수치 비교

No.	ID	차량 보유	소득 범주	연령	연소득	가구 구성원	…
1	001	보유	고	54	23,728	5	…
2	002	보유	중	48	5,143	3	…
3	003	미보유	고	32	10,567	1	…
4	004	미보유	저	23	2,782	2	…
5	005	미보유	저	31	2,987	3	…
…	…	…	…	…	…	…	…

		변수 2 (연령)			
		평균	차이	중앙값	차이
변수 1 (소득 범주)	고	59.4	10.7	53.4	8.9
	중	48.7		44.5	
	저	27.6	21.1	29.8	14.7

소득 범주 별 평균 연령 차이가
연령 중앙값 대비 두드러짐

다변량 비시각화

상관계수(corr)

› **연속형-연속형** 변수 조합간 관계성 강도 파악

No.	ID	차량 보유	소득 범주	연령	연소득	가구 구성원	…
1	001	보유	고	54	23,728	5	…
2	002	보유	중	48	5,143	3	…
3	003	미보유	고	32	10,567	1	…
4	004	미보유	저	23	2,782	2	…
5	005	미보유	저	31	2,987	3	…
…	…	…	…	…	…	…	…

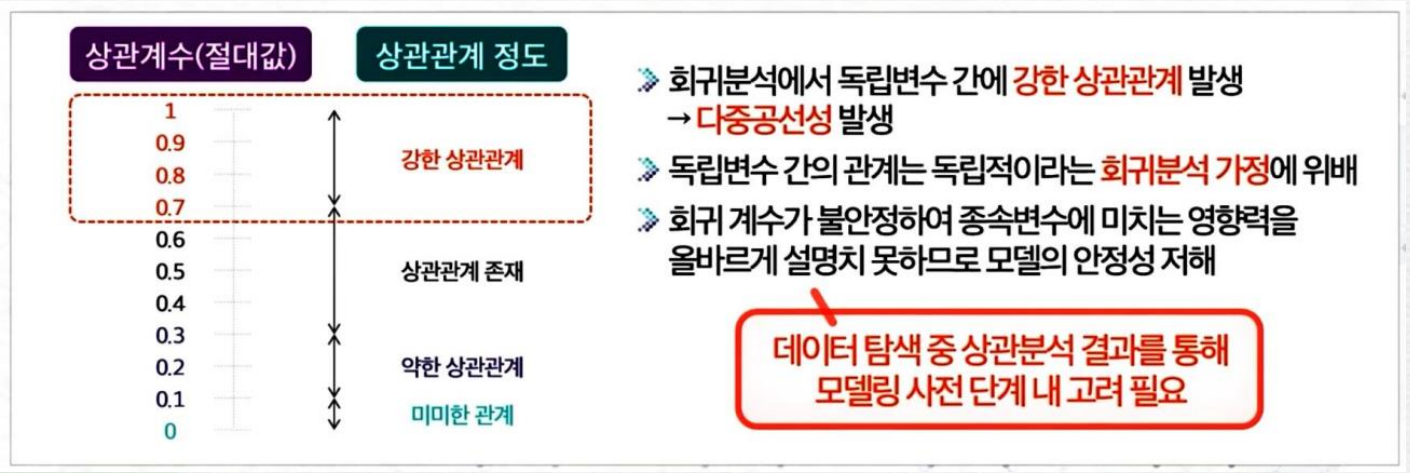
	연령	연소득	가구 구성원
연령	-	0.89	0.35
연소득	0.89	-	0.17
가구 구성원	0.35	0.17	-

연령과 연소득의 상관관계가
다른 관계 대비 강도 높음

다변량 비시각화

상관계수(corr)

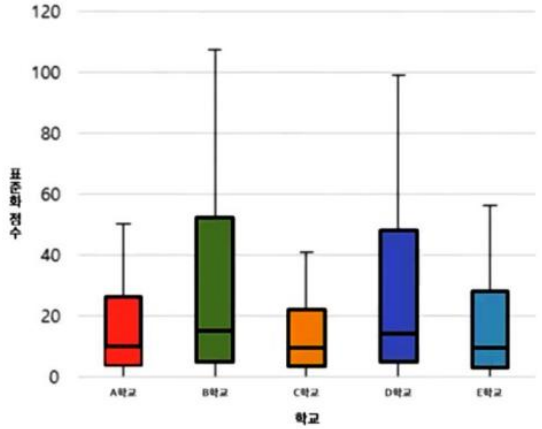
- 높은 상관 계수 : 비슷한 정보를 제공하는 밀접한 관계의 변수들 있을 경우



다변량 시각화

박스플롯

- 범주형-연속형 변수 조합간 전반적 요약통계량 파악

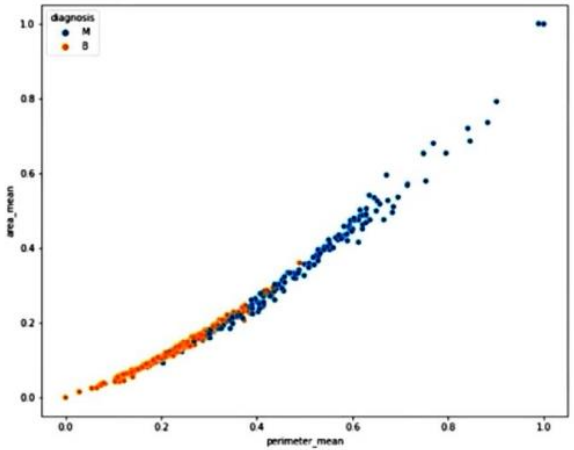


- 많은 데이터를 눈으로 직접 확인하기 어렵고, 대표적 통계 값만으로 파악하기 어려울 때 용이함
- 범주 그룹(범주형 변수) 간 수치(연속형 변수)의 집합 범위와 중앙값, 이상치 등을 빠르게 확인할 수 있음
- 비시각화 기반의 단순 수치값 비교보다 데이터가 설명하는 많은 정보 획득 가능

다변량 시각화

산점도

› **연속형-연속형** 변수 조합간 상관도 파악



- 연속형 데이터 간의 관계를 그래프상으로 어떠한 관계가 있는지 파악하기 위함
- 변수 간 분포를 통해 선형 혹은 비선형 관계 및 음양의 방향 등을 빠르게 파악할 수 있음
- 범주 Label 간 비교가 필요할 경우, 해당 부분의 그룹 정보를 표시하면 변수 간 관계 및 범주 그룹 간 관계를 함께 파악 가능

지하철 이용승객 분석

지하철 이용승객 분석 EDA

데이터

- 지하철 승하자 이용객 데이터 (2019.01.01 ~ 2019.06.30)
- 분석 메인 데이터로 1개월 데이터 6개를 합쳐야 하며
- Feature Engineering : 사용일자 활용해서 '요일', '연월', '월일' 컬럼 추가. 승하차승객수 컬럼 추가

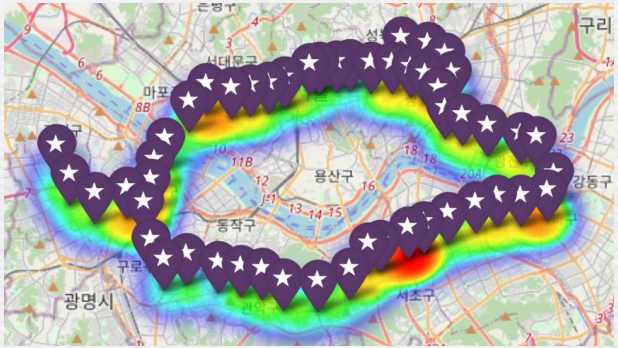
	사용일자	노선명	역명	승차총승객수	하차총승객수	등록일자
0	20190101	2호선	을지로4가	3862.0	3728.0	20190104
1	20190101	3호선	을지로3가	8104.0	7554.0	20190104

지하철 이용승객 분석 EDA

데이터

- 지하철 노선 정보 데이터
- Folium 시각화를 통해 위도와 경도를 이용해서 역 위치를 지도에 표시

	역이름	역지역	위도	경도	호선
0	낙성대	수도권	37.477090	126.963506	2호선
1	구룡	수도권	37.487027	127.059475	분당선



지하철 이용승객 분석 EDA

지하철 이용승객 분석 EDA를 통해 여러분이 배울 기술적 내용들

- 판다스로 데이터 읽고 합치기
- 요일, 연월, 일월, 승하차승객수등 컬럼 추가하기
- 데이터 저장하기
- 판다스 groupby , pivot_table, Boolean index 활용하기
- Seaborn pointplot, heatmap 그래프 그리기
- Folium 지도 시각화

지하철 이용승객 분석 EDA

지하철 이용승객 EDA 분석시 질문 리스트

- Q) 2019.01~06 중에 언제 지하철을 가장 많이 이용했을까? (기준: 승하자총승객수)
- Q, 가설) 1월~6월중에 5월에 지하철 승객수가 많다? (기준 :승하자총승객수)
- Q, 가설) 요일중에서 목요일에 지하철 승객수가 많다? (기준 :승하자총승객수)
- Q) 연월 각각에 대해 일자별(월일별) 승하차총승객수 그래프 그려 볼까요?(pointplot)
- Q) 가장 승객이 많이 타는 승차역은?
- Q) 노선별로 역별/요일별 승차승객수를 비교해 볼수 있을까? (1~9호선, 역별/요일별 heatmap)
- Q) 1호선에서 가장 하자를 많이 하는 역은? (groupby)
- Q) 2호선중에서 어느 역에서 승차가 가장 많이 발생할까? (Folium 역 표시)

정리해 보자

탐색적 데이터 분석(EDA)

- 뭐니뭐니해도 데이터가 제일 중요. 데이터 없으면 AI도 무용지물이다.
- EDA는 데이터 이해하는 과정. 데이터 특성 파악하고 인사이트 도출하는 과정
- EDA 기본 : 변수 속성파악하고 변수간의 관계 파악
- 기술 통계량, 상관관계, 시각화등을 통해 EDA 분석을 할수 있다.
- 지하철 이용 승객 데이터를 통해 EDA 분석 실습을 진행하자.

(실습) 지하철 이용 승객 데이터를 가지고 EDA 분석해 보기

DIGICO KT