

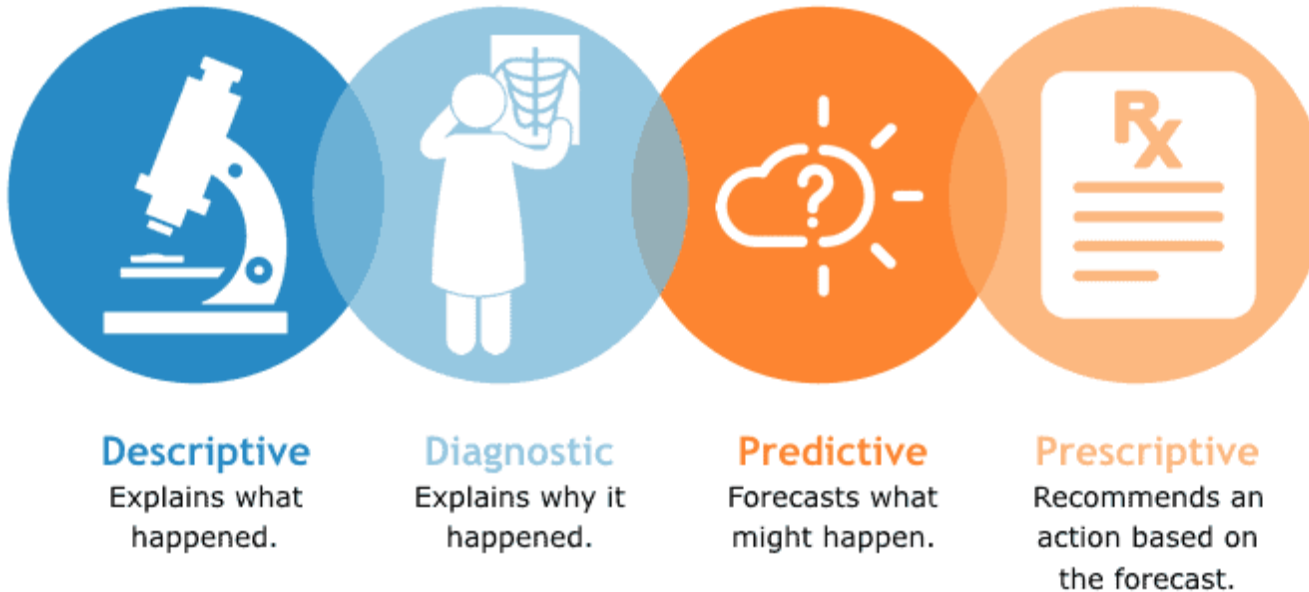
중간고사_과제 제출



- 문제 정의
- 데이터 현황분석
- 파생 변수 만들기(Feature Engineering)
- 예측

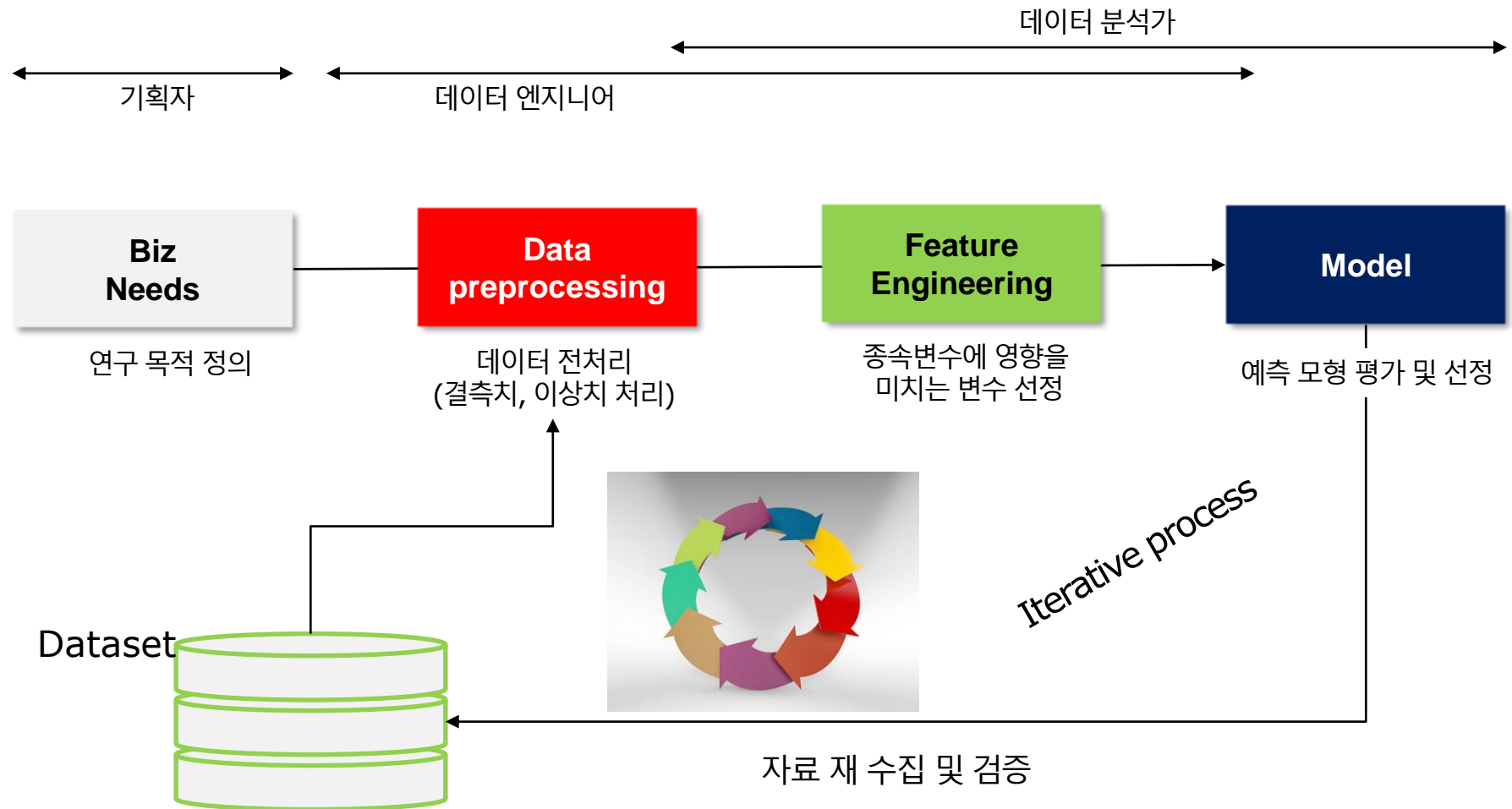
Data driven 의사결정 유형

데이터에 존재하는 규칙성(regularities)을 발견하기 위한 데이터분석(Data analytic)의 유형은 Descriptive, Diagnostic, Predictive, Prescriptive 등 4가지



<https://www.analyticsinsight.net/four-types-of-business-analytics-to-know/>

머신러닝 수행 절차



서울시 공공자전거 데이터

서울시 공공자전거 데이터




- 데이터는 어떻게 찾을 것인가?
- 데이터를 처음 받으면 무엇을 볼 것인가?
- 관련 정보는 어디서 찾을 것인가?
- 어떤 궁금한 점들이 있나(혹은 생기나)?
 - 탐색적 분석?
 - 명시적으로 표현되지 않은 변수들은 어떻게 찾나 혹은 만드나?
 - 문제를 정의하고 해결할 것인가? 예측?



서울시 공공자전거 데이터


<https://data.seoul.go.kr/dataList/literacyList.do>

서울특별시

» Finding your NEXT! 넥스트로컬 전문가 포럼 개최

서울소식 응답소 정보공개 분야별정보


로그인 | 회원가입 | 사이트맵

서울 열린데이터 광장









공공데이터 통계 서울빅데이터 소식&참여 이용안내

인기그룹데이터

Home > 공공데이터 > 인기그룹데이터



인기그룹데이터는 서울시 열린데이터광장에서 제공하는 개방데이터 중 시민 활용이 가장 높은 데이터를 20종의 그룹으로 구분하여 사용자가 손쉽게 관련 데이터를 확인할 수 있는 데이터 소개 콘텐츠입니다.

 코로나19	 건설공사정보	 골목상권분석정보	 공공서비스예약정보
 공공자전거이용정보	 공영주차장정보	 교통카드이용정보	 기상관측정보

서

	A	B	C	D	E	F	G	H	I
1	자전거번호	대여일시	대여 대여소번호	대여 대여소명	대여거치대	반납일시	반납대여소번호	반납대여소명	반납거치대
2	SPB-44695	2022-12-01 0:00	1933	개봉푸르지오아파트 상가	0	2022-12-01 0:00	1933	개봉푸르지오아파트 상가	
3	SPB-31562	2022-12-01 0:00	3007	MBC 앞	0	2022-12-01 0:00	3007	MBC 앞	
4	SPB-56324	2022-12-01 0:01	4468	가락1동주민센터	0	2022-12-01 0:01	4468	가락1동주민센터	
5	SPB-30175	2022-12-01 0:03	652	답십리 래미안엘라인아파트 입구	0	2022-12-01 0:03	652	답십리 래미안엘라인아파트 입	
6	SPB-37639	2022-12-01 0:00	1047	강동 한신휴플러스	0	2022-12-01 0:03	1075	천동초교 삼거리	
7	SPB-39909	2022-12-01 0:03	4508	유원목동아파트 103동 앞	0	2022-12-01 0:03	4508	유원목동아파트 103동 앞	
8	SPB-31850	2022-12-01 0:00	646	장한평역 1번출구	0	2022-12-01 0:04	668	서울축산농협(장안지점)	
9	SPB-36226	2022-12-01 0:00	1175	대한항공 인력개발센터	0	2022-12-01 0:04	1118	중미역 3번출구뒤(등준두산위	
10	SPB-53863	2022-12-01 0:01	1210	롯데월드타워(잠실역2번출구 쪽)	0	2022-12-01 0:04	2608	송파구청	
11	SPB-30127	2022-12-01 0:01	203	국회의사당역 3번출구 옆	0	2022-12-01 0:04	201	진미파라곤 앞	
12	SPB-55046	2022-12-01 0:00	1344	아리랑시네센터 앞	0	2022-12-01 0:04	1386	동선고가차도 초입	
13	SPB-50449	2022-12-01 0:00	498	연남동주민센터 앞	0	2022-12-01 0:04	113	홍대입구역 2번출구 앞	
14	SPB-52881	2022-12-01 0:00	5306	동신아파트 후문 옆	0	2022-12-01 0:04	2915	서울과학기술대학교(여학교육	
15	SPB-50627	2022-12-01 0:00	1351	안암2교 옆	0	2022-12-01 0:04	3416	동묘앞역 6번출구	
16	SPB-61436	2022-12-01 0:00	191	서우빌딩	0	2022-12-01 0:05	4255	래미안 루센티아 아파트 108동	
17	SPB-61011	2022-12-01 0:03	2524	반포소핑타운 8동 앞	0	2022-12-01 0:05	2523	뉴코아 김스클럽 앞	
18	SPB-59271	2022-12-01 0:00	5064	양천향교역8번출구	0	2022-12-01 0:05	3783	마곡금호어울림아파트	
19	SPB-59269	2022-12-01 0:05	1297	석촌호수교차로(동호 팔각정 앞)	0	2022-12-01 0:05	1297	석촌호수교차로(동호 팔각정 앞)	
20	SPB-43585	2022-12-01 0:01	1735	도봉역 1,2번 출구사이 건너편	0	2022-12-01 0:05	1727	서울도봉초등학교인근	
21	SPB-65898	2022-12-01 0:02	1531	미아사거리 1번 출구	0	2022-12-01 0:05	1560	도봉세무서 앞	
22	SPB-32179	2022-12-01 0:00	3007	MBC 앞	0	2022-12-01 0:05	4238	상암한샘빌딩앞	

◆ 개방데이터 정보

서울시 공공자전거 대여
서울시 공공자전거 이용
서울시 공공자전거 이용
서울시 공공자전거 신규
서울시 공공자전거 대여
서울시 공공자전거 이용
서울시 공공자전거 외국
서울시 공공자전거 외국
서울시 공공자전거 대여이력 정보 >
서울시 공공자전거 실시간 대여정보 >

3	데이터	서울특별시 공공자전거 대여이력 정보_2210.csv	309.78	2023.01.17.	다운로드
4	데이터	서울특별시 공공자전거 대여이력 정보_2209.csv	673.48	2023.01.17.	다운로드
5	데이터	서울특별시 공공자전거 대여이력 정보_22081.csv	787.85	2023.01.17.	다운로드
6	데이터	서울특별시 공공자전거 대여이력 정보_2207.csv	842.03	2023.01.17.	다운로드
7	데이터	서울특별시 공공자전거 대여이력 정보_22.06.csv	641.15	2023.01.17.	다운로드
			724.31	2023.01.17.	다운로드
			559.48	2022.07.11.	다운로드

서울시 공공자전거 데이터

실습 내용 - 기본 정보 확인하기

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	자전거번호	대여일시	대여 대여소번호	대여 대여소명	대여거치대	반납일시	반납대여소번호	반납대여소명	반납거치대	이용시간(분)	이용거리(M)	생년	성별	이용자종류	대여대여소ID	반납대여소ID
2	SPB-44695	2022-12-01 0:00	1933	개봉푸르지오아파트 상가	0	2022-12-01 0:00	1933	개봉푸르지오아파트 상가	0	0	0	WN	M	내국인	ST-678	ST-678
3	SPB-31562	2022-12-01 0:00	3007	MBC 앞	0	2022-12-01 0:00	3007	MBC 앞	0	0	111.2	1995	F	내국인	ST-2165	ST-2165
4	SPB-56324	2022-12-01 0:01	4468	가락1동주민센터	0	2022-12-01 0:01	4468	가락1동주민센터	0	0	0	1994	M	내국인	ST-2583	ST-2583
5	SPB-30175	2022-12-01 0:03	652	답십리 래미안엘파인아파트 입구	0	2022-12-01 0:03	652	답십리 래미안엘파인아파트 입	0	0	0	1981	M	내국인	ST-1447	ST-1447
6	SPB-37639	2022-12-01 0:00	1047	강동 한신휴플러스	0	2022-12-01 0:03	1075	천동초교 삼거리	0	3	492.34	1989	M	내국인	ST-1369	ST-1836
7	SPB-39909	2022-12-01 0:03	4508	유원목동아파트 103동 앞	0	2022-12-01 0:03	4508	유원목동아파트 103동 앞	0	0	0	1985	M	내국인	ST-2760	ST-2760
8	SPB-31850	2022-12-01 0:00	646	장한평역 1번출구	0	2022-12-01 0:04	668	서울측산농협(장안지점)	0	3	911.71	1992	WN	내국인	ST-1298	ST-2268
9	SPB-36226	2022-12-01 0:00	1175	대한항공 인력개발센터	0	2022-12-01 0:04	1118	중미역 3번출구뒤(동춘두산위)	0	4	660.88	1970	F	내국인	ST-1512	ST-516
10	SPB-53863	2022-12-01 0:01	1210	롯데월드타워(잠실역2번출구 쪽)	0	2022-12-01 0:04	2608	송파구청	0	3	613.24	2000	F	내국인	ST-891	ST-1681
11	SPB-30127	2022-12-01 0:01	203	국회의사당역 3번출구 옆	0	2022-12-01 0:04	201	진미파라곤 앞	0	2	469.46	1967	WN	내국인	ST-51	ST-46
12	SPB-55046	2022-12-01 0:00	1344	아리랑시네센터 앞	0	2022-12-01 0:04	1386	동선고가차도 초입	0	4	913.6	1959	M	내국인	ST-1208	ST-2227
13	SPB-50449	2022-12-01 0:00	498	연남동주민센터 앞	0	2022-12-01 0:04	113	홍대입구역 2번출구 앞	0	4	905.64	1984	M	내국인	ST-2157	ST-18
14	SPB-52881	2022-12-01 0:00	5306	동신아파트 후문 옆	0	2022-12-01 0:04	2915	서울과학기술대학교(아하교육)	0	4	791.01	1999	M	내국인	ST-3060	ST-2081
15	SPB-50627	2022-12-01 0:00	1351	안암2교 옆	0	2022-12-01 0:04	3416	동묘앞역 6번출구	0	4	1249.23	1997	F	내국인	ST-1215	ST-1813
16	SPB-61436	2022-12-01 0:00	191	서우빌딩	0	2022-12-01 0:05	4255	래미안 루센티아 아파트 108동	0	4	1390	1993	M	내국인	ST-347	ST-2897
17	SPB-61011	2022-12-01 0:03	2524	반포소평타운 8동 앞	0	2022-12-01 0:05	2523	뉴코아 킴스클럽 앞	0	1	0	2000	M	내국인	ST-1909	ST-1908
18	SPB-59271	2022-12-01 0:00	5064	양천학교역8번출구	0	2022-12-01 0:05	3783	마곡금호어울림아파트	0	4	670	WN	M	내국인	ST-2943	ST-2490
19	SPB-59269	2022-12-01 0:05	1297	석촌호수교차로(동호 팔각정 앞)	0	2022-12-01 0:05	1297	석촌호수교차로(동호 팔각정 앞)	0	0	0	1981	M	내국인	ST-1586	ST-1586
20	SPB-43585	2022-12-01 0:01	1735	도봉역 1,2번 출구사이 건너편	0	2022-12-01 0:05	1727	서울도봉초등학교인근	0	3	495.4	1978	M	내국인	ST-1467	ST-1297
21	SPB-65898	2022-12-01 0:02	1531	미아사거리 1번 출구	0	2022-12-01 0:05	1560	도봉세무서 앞	0	3	660	1997	WN	내국인	ST-1105	ST-2132
22	SPB-32179	2022-12-01 0:00	3007	MBC 앞	0	2022-12-01 0:05	4238	상암한샘빌딩앞	0	4	695.74	1995	F	내국인	ST-2165	ST-2981

서울시 공공자전거 데이터

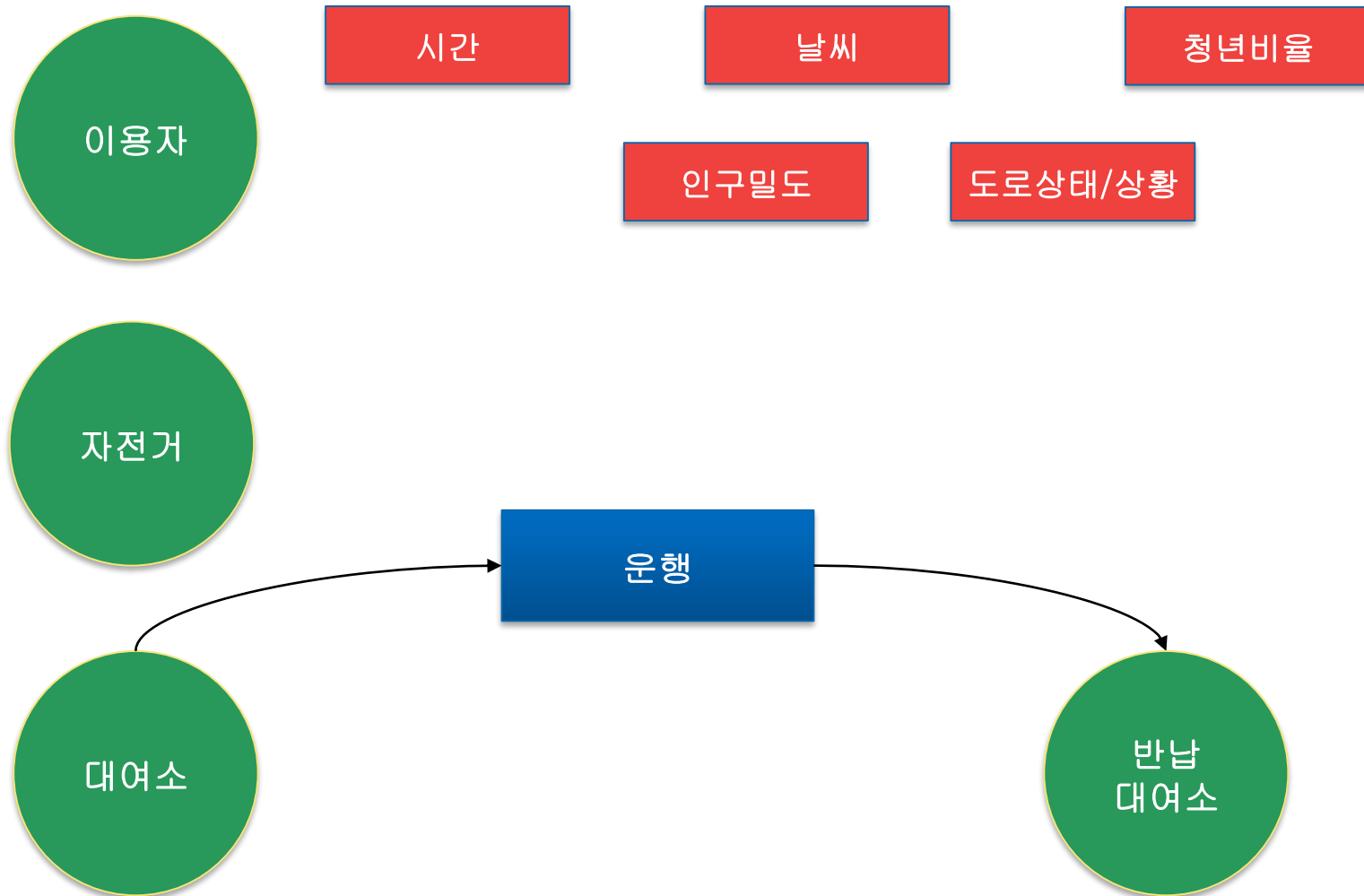
실습 내용 - 기본 정보 확인하기

	데이터형태(dtypes)	비 결측치 수(notnull)	결측치 수(null)	고유값 수(nunique)
자전거번호	object	1048575	0	27431
대여일시	float64	1048575	0	614223
대여 대여소번호	int64	1048575	0	2687
대여 대여소명	object	1048575	0	2684
대여거치대	int64	1048575	0	2
반납일시	float64	1048575	0	614448
반납대여소번호	object	1048575	0	2694
반납대여소명	object	1048575	0	2691
반납거치대	object	1048575	0	3
이용시간(분)	int64	1048575	0	559
이용거리(M)	float64	1048575	0	253209
생년	object	1048575	0	164
성별	object	1014450	34125	5
이용자종류	object	1048575	0	3
대여대여소ID	object	1048575	0	2687
반납대여소ID	object	1048575	0	2694

서울시 공공자전거 데이터 - 궁금한 점 예시

- 어떤 데이터인가? 무엇을 보면 좋을까?
- 데이터에 이상한 점은 없나?
- 있다면 어떻게 수정할 것인가?
- 대여소명에서 대여소번호 제거하기
- 얼마나 운행할까 시간/거리
- 이용시간 초단위로 바꾸기
- 운행 속도 계산
- 운행 속도 분포는?
- binning할 때 어떤 함수를 쓸 수 있을까?
- 역 근처여부 변수 만들기
- 역 이외 다른 어떤 변수는 무엇이 있는가?
- 해당 변수를 어떻게 만들면 좋을까?
- <https://exceljet.net/formula/cell-contains-one-of-many-things>
- 어느 대여소가 가장 많이 이용되는가?
- 어느 대여소를 이전하는 것이 좋을까?
- 대여소번호로 행정구 연결하기
- 어느 구에서 많이 사용하나?
- 어떤 이유로 많이 사용할까?
- 행정구를 넘는 비율 계산하기
- 대여와 반납 위치가 같은 곳은? 비율은
- 어떤 곳인가? 왜 그럴까?
- 거치대 바깥쪽 자전거를 많이 이용할까
- 시간대별로 언제 많이 이용할까?
- 무슨 목적이었을까?
- 요일별로 차이가 있을까? 패턴의 차이가 있나?
- 또 어떤 영향요소가 있을까?
- 그 요소가 영향을 미쳤는지 어떻게 확인할 수 있을까?
- 추가 데이터를 생각해 볼 수 있을까?

서울시 공공자전거 데이터 이해



서울시 공공자전거 데이터 - 궁금한 점 예시

기본정보 확인

- 어떤 데이터인가? 무엇을 보면 좋을까?
- 어떤 상황에서 발생한 정보인가?
- 각 컬럼들이 의미하는 것은 무엇인가?
- 어떤 데이터 종류이고 어떤 값들을 가질 수 있나?
- null값은 없나?
- unique값의 개수는?

이상치 발견

- 데이터에 이상한 점은 없나?
- 있다면 어떻게 수정할 것인가?

데이터 클리닝

- 대여소명에서 대여소번호 제거하기

파생변수 도출

- 이용시간 초단위로 바꾸기
- 운행 속도 계산
- 운행 속도 분포는?
- binning할 때 어떤 함수를 쓸 수 있을까?
- 역 근처여부 변수 만들기
- 역 이외 다른 어떤 변수는 무엇이 있는가?
- 해당 변수를 어떻게 만들면 좋을까?
- <https://exceljet.net/formula/cell-contains-one-of-many-things>

서울시 공공자전거 데이터 - 궁금한 점 예시

서울시 공공자전거 데이터 - EDA(Explanatory Data Analysis)

탐색적 분석

- 얼마나 운행할까 시간/거리
- 어느 대여소가 가장 많이 이용되는가?
- 어느 대여소를 이전하는 것이 좋을까?
- 어느 구에서 많이 사용하나?
 - 행정구 어떻게 찾아 넣을까
- 어떤 이유로 많이 사용할까?
- 대여와 반납 위치
 - 행정구가 다른 경우가 있나?
 - 대여와 반납 위치가 같은 곳은
 - 그 비율은 얼마나 되나
- 어떤 곳인가? 왜 그럴까?
- 거치대 바깥쪽 자전거를 많이 이용할까
- 시간대별로 언제 많이 이용할까?
- 왜 그럴까? 무슨 목적이었을까?
- 요일별로 차이가 있을까? 패턴의 차이가 있나?
- 또 어떤 영향요소가 있을까?
- 그 요소가 영향을 미쳤는지 어떻게 확인할 수 있을까?
- 추가 데이터를 생각해 볼 수 있을까?

서울시 공공자전거 데이터 확보

같이 생각해 볼 문제



- 운행속도
- 역주변 여부
- 학교/공원 주변 여부
- 요일
- 휴일여부
- 시간대
- 바깥거치대 여부 등의 정보가 필요한데 어떻게 추출할 것인가?

Input



함수



Output

서울시 공공자전거 데이터 확보

같이 생각해 볼 문제



- 얼마나 운행할까
 - 시간기준/거리기준
 - 어떤 그래프를 그리는 것이 좋을까?
- 이용시간 초단위로 바꾸기
- 운행 속도 계산
- 운행 속도 분포는?
- 운행속도를 10km/h 단위로 binning한다면?
- 역 근처여부 변수 만들기
- 학교나 공원이 포함된 대여소 여부를 변수로 만들기
- 시간대별로 언제 많이 이용할까?
- 왜 그럴까? 무슨 목적이었을까?
- 요일별로 차이가 있을까? 패턴의 차이가 있나? =1

서울시 공공자전거 데이터 확보

같이 생각해 볼 문제



- 어느 대여소가 가장 많이 이용되는가?
- 어느 대여소를 이전하는 것이 좋을까?
- 어느 구에서 많이 사용하나?
 - 행정구 어떻게 찾아 넣을까
 - 어떤 이유로 많이 사용할까?
- 대여와 반납 위치
 - 행정구가 다른 경우가 있나?
 - 대여와 반납 위치가 같은 곳은
 - 그 비율은 얼마나 되나
 - 어떤 곳인가? 왜 그럴까?
- 거치대는 안쪽을 선호할까 아니면 바깥쪽을 선호할까?
- 대여소 매칭이 안되는 것은 어떤 것들이 있는가?
 - 어떻게 하는 것이 좋을까?

* 기상자료: 종관기상관측(ASOS) - 파일셋

- <https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36&tabNo=1>

서울시 공공자전거 데이터 확보

역 / 학교 / 공원이 포함된 대여소

	A	B	C	D	E	F	G	H	I	J	K
	대여대여소명	역_여부	출구_여부	역+출구	공원_여부	학교_여부	고교_여부	여고_여부	공고_여부	공원또는학교	공원또는학교
2	녹번역 4번출구	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	압구정역 2번 출구 옆	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	CJ 드림시티	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	문래동자이아파트 앞	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	홍대입구역 2번출구 앞	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	KT&G 관악지점	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	서울시립대 정문 앞	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	미성아파트 A동 앞	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
0	한진아파트 입구	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1	신길역3번출구	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	서울시립대 정문 앞	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	당현천근린공원	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
4	봉림교 교통섬	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	영등포구 (양재천) 그리고 앞	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE

