포트폴리오

정승태

- 목 차 -

1. 개요

데이터를 수집하여 데이터베이스에 저장한 후 SQL 을 활용하여 데이터를 전처리하는 파이프라인을 생성한다. 처리된 데이터로 데이터마트를 생성한 후 데이터를 시각화한다.

2. 데이터분석

가. 분석목적

가설 : 노인일자리 사업은 노인의 인구분포에 따라 제공되는가

분석 목표 : 노인일자리 모집현황과 노인인구분포의 상관관계를 분석하여 인구분포별로 적합한 일자리 제공 수를 예측한다.

나. 데이터 수집/저장

- 데이터 출처

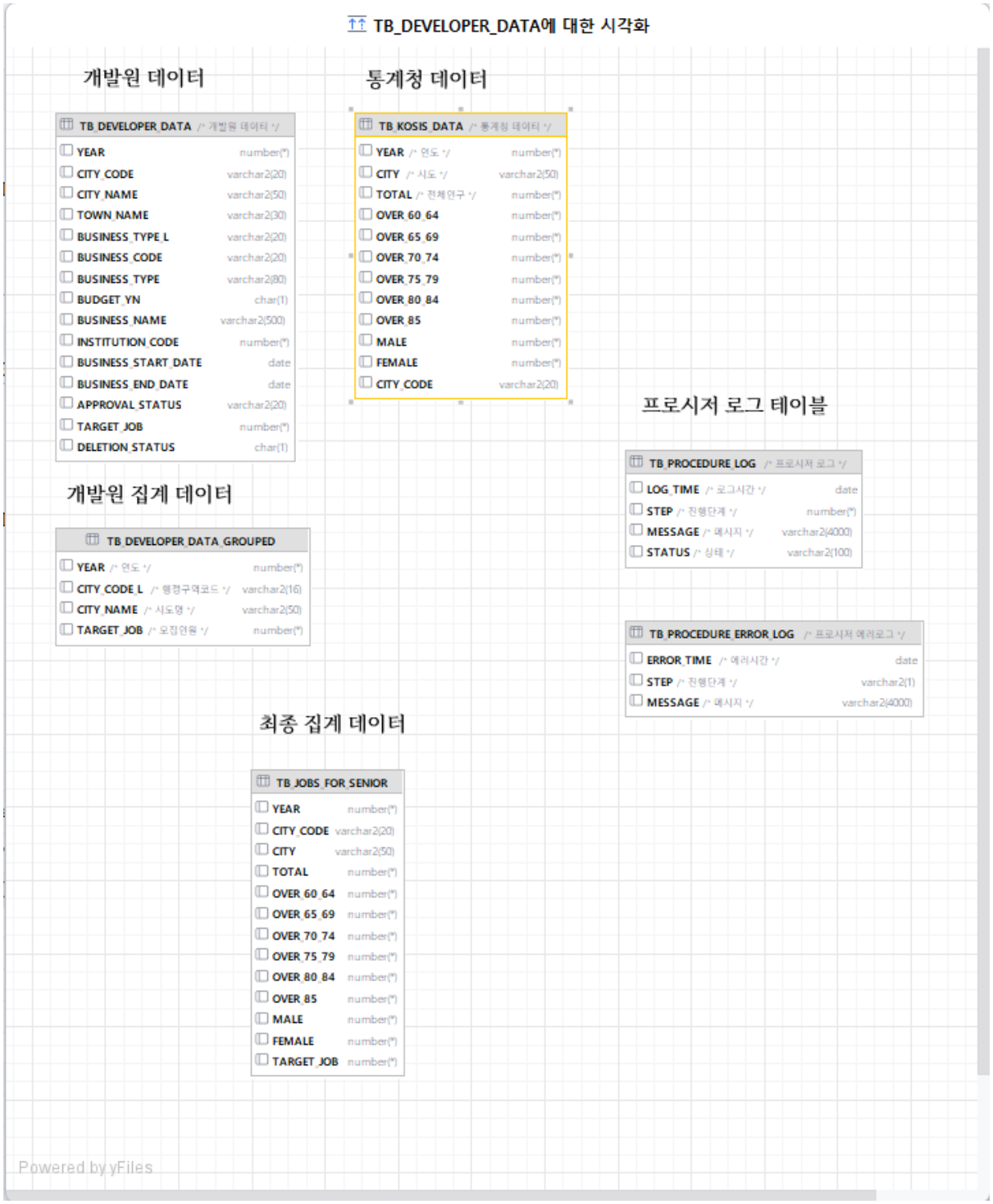
공공데이터 포털 : https://www.data.go.kr/, 노인일자리 사업계획서 현황

통계청 : https://kosis.kr/index/index.do, 인구총조사 – 고령자

출처로부터 수집한 데이터를 데이터베이스에 테이블로 저장합니다. 저장된 데이터는 아래와 같습니다.

`

- 테이블 정보


TB_DEVELOPER_DATA에 대한 시각화
개발원 데이터
TB_DEVELOPER_DATA /* 개발원 데이터 */
YEAR number(*)
CITY_CODE varchar2(20)
CITY_NAME varchar2(50)
TOWN_NAME varchar2(30)
BUSINESS_TYPE_L varchar2(20)
BUSINESS_CODE varchar2(20)
BUSINESS_TYPE varchar2(80)
BUDGET_YN char(1)
BUSINESS_NAME varchar2(500)
INSTITUTION_CODE number(*)
BUSINESS_START_DATE date
BUSINESS_END_DATE date
APPROVAL_STATUS varchar2(20)
TARGET_JOB number(*)
DELETION_STATUS char(1)
통계청 데이터
TB_KOSIS_DATA /* 통계청 데이터 */
YEAR /* 연도 */ number(*)
CITY /* 시도 */ varchar2(50)
TOTAL /* 전체인구 */ number(*)
OVER_60_64 number(*)
OVER_65_69 number(*)
OVER_70_74 number(*)
OVER_75_79 number(*)
OVER_80_84 number(*)
OVER_85 number(*)
MALE number(*)
FEMALE number(*)
CITY_CODE varchar2(20)
프로시저 로그 테이블
TB_PROCEDURE_LOG /* 프로시저 로그 */
LOG_TIME /* 로그시간 */ date
STEP /* 진행단계 */ number(*)
MESSAGE /* 메시지 */ varchar2(4000)
STATUS /* 상태 */ varchar2(100)
개발원 집계 데이터
TB_DEVELOPER_DATA_GROUPED
YEAR /* 연도 */ number(*)
CITY_CODE_L /* 행정구역코드 */ varchar2(16)
CITY_NAME /* 시도명 */ varchar2(50)
TARGET_JOB /* 모집인원 */ number(*)
TB_PROCEDURE_ERROR_LOG /* 프로시저 에러로그 */
ERROR_TIME /* 에러시간 */ date
STEP /* 진행단계 */ varchar2(1)
MESSAGE /* 메시지 */ varchar2(4000)
최종 집계 데이터
TB_JOBS_FOR_SENIOR
YEAR number(*)
CITY_CODE varchar2(20)
CITY varchar2(50)
TOTAL number(*)
OVER_60_64 number(*)
OVER_65_69 number(*)
OVER_70_74 number(*)
OVER_75_79 number(*)
OVER_80_84 number(*)
OVER_85 number(*)
MALE number(*)
FEMALE number(*)
TARGET_JOB number(*)
Powered by yFiles

각 테이블의 정보는 아래와 같습니다.

- 아 래 -

(1) 개발원 데이터 : 노인일자리 정보

(2) 통계청 데이터 : 노인인구 정보

(3) 개발원 집계 데이터 : 전처리 후 시도별로 일자리 수를 집계한 테이블

(4) 최종 집계 데이터 : 집계 데이터와 통계청 데이터를 결합한 테이블

(5) 프로시저 로그 테이블 : 전처리와 집계처리를 수행하는 프로시저의 로그정보 테이블

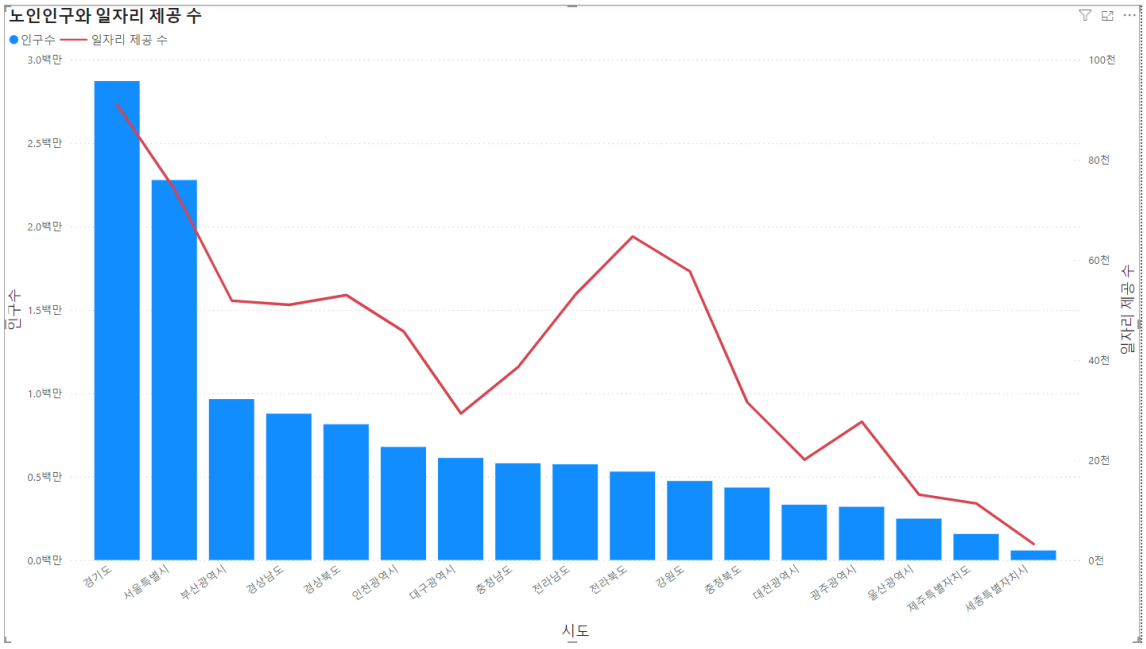
다. 데이터 전처리 및 파이프라인

데이터 전처리 과정은 프로시저를 생성하여 처리합니다. 생성된 프로시저는 스케줄러를 활용하여 정해진 주기마다 실행하여 데이터를 처리하는 파이프라인을 구성하고 데이터마트를 생성합니다. 처리된 데이터는 아래와 같습니다.

	YEAR	CITY_CODE	CITY	TOTAL	OVER_60_64	OVER_65_69	OVER_70_74	OVER_75_79	OVER_80_84
1	2022	2900000000	광주광역시	319736	104414	73621	55281	39738	2
2	2022	2600000000	부산광역시	965387	293439	244948	177466	120231	8
3	2022	4700000000	경상북도	814516	237039	189995	141627	97700	8
4	2022	4100000000	경기도	2872546	992291	691281	448782	331625	24
5	2022	4200000000	강원도	474360	144371	112710	72462	61241	5
6	2022	4400000000	충청남도	580382	170562	129344	97347	70459	6
7	2022	4500000000	전라북도	530632	148203	117035	92476	70132	6
8	2022	4800000000	경상남도	878048	281822	211254	147971	100805	8
9	2022	4600000000	전라남도	574325	155396	119211	99829	79131	7
10	2022	3000000000	대전광역시	332117	110774	81842	55111	37680	2
11	2022	3600000000	세종특별자치시	57387	19979	14308	9136	5837	
12	2022	2800000000	인천광역시	678410	240953	165728	106152	75837	5
13	2022	1100000000	서울특별시	2279066	693030	561673	399408	294504	20
14	2022	2700000000	대구광역시	612655	195849	148586	108016	71817	5
15	2022	4300000000	충청북도	434829	137334	103867	68236	51160	4

- 프로시저 코드 경로 :
https://github.com/jungseungtae/portpolio/blob/master/sql/SP_JOBS_THE_ELDERL.sql

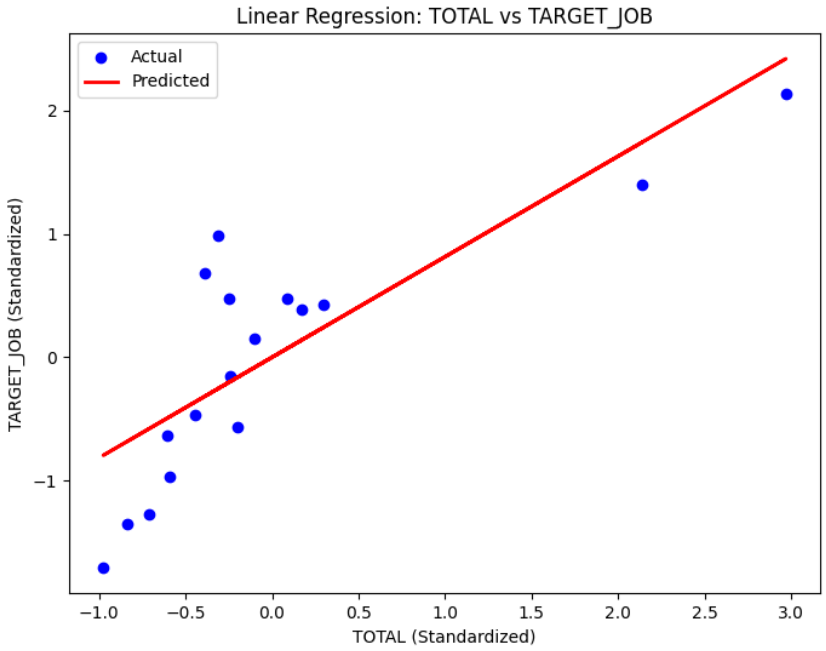
라. 데이터 시각화



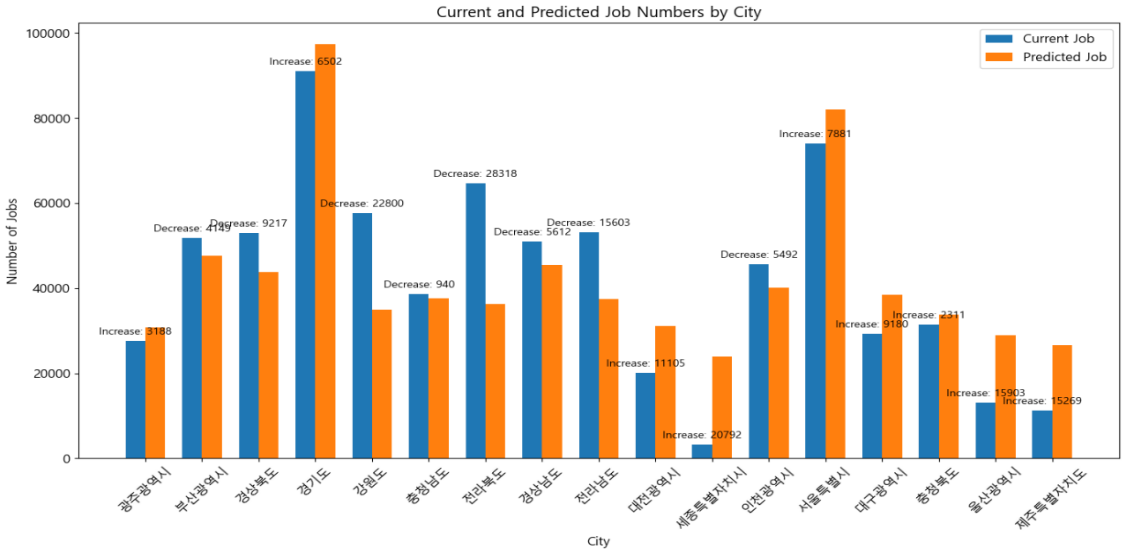
위 이미지를 살펴보면 5 개정도의 시도를 제외하면 노인일자리는 노인인구 분포에 따라 제공되고 있는 것으로 보이며 상관계수는 0.815281 로 높은 양의 상관관계에 있음을 확인할 수 있습니다.

한편 몇몇 시도에서는 인구수 대비 높은 일자리를 제공하고 있습니다. 그렇다면 노인 인구의 분포만을 기준으로 하였을 때 각 행정구역별로 어느 정도의 일자리를 제공하는 것이 적합한지 선형회귀 알고리즘을 통해 알아보겠습니다.

위 수치를 계산하기 위해서 노인 인구수와 일자리 수를 표준화한 후 선형회귀를 적용한 결과는 아래와 같습니다.



인구 분포에 따르는 예측 일자리 수는 그래프의 직선에 해당하며 증감을 구하기 위해 데이터를 다시 역표준화하여 조정해야 하는 수치는 아래와 같습니다.



위 그래프에서 파란색은 현재 제공하고 있는 일자리 수를 의미하고 주황색은 노인 인구분포에 따라 제공되어야 하는 일자리 수의 예측 값을 의미하며 구체적인 수치는 아래와 같습니다.

시도	실제 값	예측 값	증감	증감 수치
전라북도	64,665	36,346	Decrease	-28,319
강원도	57,678	34,877	Decrease	-22,801
세종특별자치시	3,202	23,994	Increase	20,792
울산광역시	13,078	28,981	Increase	15,903
전라남도	53,090	37,486	Decrease	-15,604
제주특별자치도	11,320	26,589	Increase	15,269
대전광역시	20,059	31,165	Increase	11,106
경상북도	52,973	43,755	Decrease	-9,218
대구광역시	29,306	38,487	Increase	9,181
서울특별시	74,098	81,980	Increase	7,882
경기도	90,967	97,470	Increase	6,503
경상남도	51,026	45,414	Decrease	-5,612
인천광역시	45,696	40,203	Decrease	-5,493
부산광역시	51,843	47,693	Decrease	-4,150
광주광역시	27,653	30,842	Increase	3,189
충청북도	31,534	33,846	Increase	2,312
충청남도	38,585	37,645	Decrease	-940

바. 결론

위 분석과정을 통하여 노인인구분포에 따라 제공되어야 하는 적합한 노인일자리 수를 예측하여 보았습니다. 한편 노인일자리는 공익성을 가진 사업으로 수혜자 결정에 있어 소득, 주거환경, 세대구성 등을 함께 고려해야 합니다. 해당 데이터들은 개인정보에 해당하여 수집할 수 없었지만 추가적으로 반영할 수 있다면 모델의 정확성과 신뢰성을 더욱 높일 수 있을 것입니다.

- 코드 경로 :

https://github.com/jungseungtae/portpolio/blob/master/Portfolio/python/senior.py