2020년 공공 빅데이터 청년 인턴십

1차 프로젝트(2020.09 ~ 2020.11): 국가산업단지의 현황 분석 및 변화 예측
2차 프로젝트(2020.12 ~ 2021.02): 공산품 안전·인증 결과의 신뢰성 분석

빅데이터 분석 프로젝트 수행보고서

국가산업단지의 현황 분석 및 변화 예측

박은총, 홍강우

2020. 11. 25

(참여기관) 산업통상자원부
(주관기관) 행정안전부
(전담기관) 한국정보화진흥원
(운영기관) 씨에스리컨소시엄

목 차

1. 분석 개요

가. 분석 목표

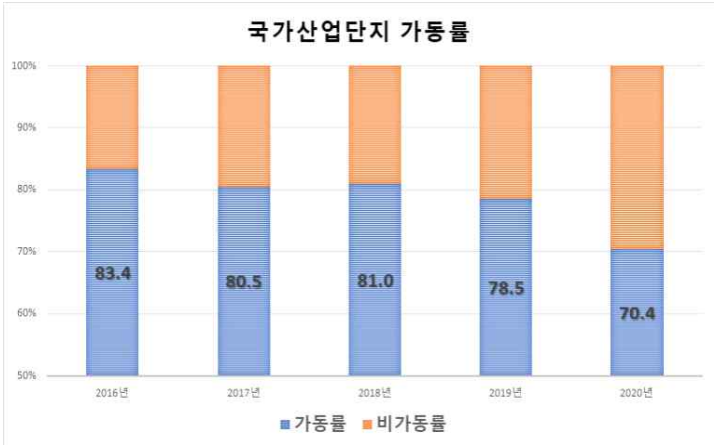
□ 한국산업단지공단에서 관리하고 있는 국가산업단지의 현황을 분석하고 변화를 예측한다.

나. 분석 배경

1) 현황 및 필요성

□ 코로나19 바이러스의 여파로 국가산업단지의 가동률이 역대 최저치를 기록했다. 지난 1998년 국제통화기금(IMF) 사태나 2009년 글로벌 금융위기 직후보다도 낮은 가동률을 보이고 있다. 이와 관련하여 국가산단의 변화를 파악해보고자 한다.

□ IMF 직후인 1998년이나 글로벌 금융위기 이후인 2009년에도 가동률은 70%대 후반을 지켰으나 현재(2020년)의 가동률 70.4%는 처음 기록한 수치이다.



[그림 1-1] 국가산업단지 가동률

□ 전국 산업단지 1,176개 1,402㎢ 중 국가산업단지는 44개 786㎢가 지정되어 있으며, 면적으로 56.1%를 점유하여 대규모로 조성되어있다. (2017년 09월 기준)

□ 전국 산업단지에서 가장 많은 비중을 차지하는 국가산업단지의 현황 분석을 통해 앞으로의 변화를 예측한다.

다. 분석 주제

1) 코로나19 바이러스로 인한 국가산업단지의 변화를 파악한다.
2) 국가산업단지가 전체 산단에서 갖는 영향력을 분석 및 예측한다.
3) 국가산업단지의 자원 사용량을 분석하여 신규 입주 기업의 자원 사용량을 예측한다.

라. 분석 결과물

1) 분석 개요서
   □ 분석 목표, 분석 배경, 분석 주제, 분석 결과물, 분석 일정
2) 원시 데이터셋 및 원시 데이터 목록(설명서)
3) 통합 데이터 가공 계획서
4) 통합 데이터셋 및 통합 데이터 목록(설명서)
5) 분석 수행계획서
   □ 분석 개요
   □ 데이터 수집 및 가공 프로세스
   □ 분석 주제 소개 및 상세 프로세스
   □ 참고자료
   □ 별첨
6) 분석용 데이터 가공 계획서
   □ 문제 및 데이터 이해
   □ 분석용 데이터 가공 방법
7) 분석용 데이터셋 및 분석용 데이터 목록(설명서)
8) 프로그램 코드
   □ 통합 데이터 가공 코드: 1차, 2차, 3차 데이터 가공 코드
   □ 분석용 데이터 가공 코드: 주제별 분석용 데이터 가공 코드
   □ 코드명, 작성 언어, 알고리즘 설명
9) 분석 자료
   □ 분석 결과 등
10) 분석모형 검증 자료
   □ 정확도 등
11) 시각화 자료
   □ 변수 시각화

- 전국산업단지 현황(개수, 위치) / 지도
- 국가산업단지 현황(개수, 위치) / 지도
- 단지별 입주, 생산, 수출, 고용, 가동률 / 꺽은선 그래프
- 업종별 입주, 생산, 수출, 고용, 가동률 / 꺽은선 그래프

□ 분석 시각화

12) 분석 결과보고서

□ 분석 개요

□ 데이터 수집 및 가공 프로세스

□ 분석 주제 소개 및 상세 프로세스

□ 분석 결과

□ 기대 효과 및 활용 방안

□ 참고자료

□ 별첨

마. 분석 일정

작업 구분	작업명	담당자 (주:책임)	결과	계획 시작일	계획 기간 (DAY)	계획 근무주 (WEEK)	계획 종료일
1	요건 분석			2020-09-02			2020-09-08
1.1	분석요건 정리	정종민		2020-09-02	1	1	2020-09-02
1.2	분석목표 설정	정종민		2020-09-03	1	1	2020-09-03
1.3	시나리오 수립	홍강우(주)	분석 개요서	2020-09-04	3	1	2020-09-08
2	데이터 수집 및 정제			2020-09-09			2020-10-06
2.1	필요데이터 정의 및 파악	박은총(주)		2020-09-09	3	1	2020-09-11
2.2	원시데이터 수집	박은총(주)	원시 데이터셋 및 원시 데이터 목록(설명서)	2020-09-14	3	1	2020-09-16
2.3	국가산업단지 데이터 수집 및 점검	정종민		2020-09-17	3	1	2020-09-21
2.4	통합 데이터셋 설계	박은총(주)	통합 데이터 가공 계획서	2020-09-22	6	2	2020-09-29
2.5	통합 데이터셋 구축	박은총(주)	통합 데이터셋 및 통합 데이터 목록(설명서)	2020-10-05	2	1	2020-10-06
3	분석모형 설계 및 개발			2020-10-07			2020-11-06
3.1	분석모형 설계	홍강우(주)	분석 수행계획서	2020-10-07	3	1	2020-10-12
3.2	분석용 데이터셋 설계	박은총(주)	분석용 데이터 가공 계획서	2020-10-13	4	1	2020-10-16
3.3	분석용 데이터셋 구축	박은총(주)	분석용 데이터셋 및 분석용 데이터 목록(설명서)	2020-10-17	2	1	2020-10-20
3.4	분석모형 개발	홍강우(주)	프로그램 코드	2020-10-21	3	1	2020-10-23
3.5	분석	홍강우(주)	분석 자료	2020-10-26	5	1	2020-10-30
3.6	분석모형 검증	박은총(주)	분석모형 검증 자료	2020-11-02	5	1	2020-11-06
4	분석 결과보고서 작성			2020-11-09			2020-11-30
4.1	분석 결과 시각화	홍강우(주)	시각화 자료	2020-11-09	5	1	2020-11-13
4.2	분석 결과보고서 작성	홍강우(주)	분석 결과보고서	2020-11-16	10	2	2020-11-27
4.3	분석 결과 발표	홍강우(주)		2020-11-30	1	1	2020-11-30

[표 1-1] 분석 일정

2. 데이터 수집 및 가공 프로세스

가. 입수 및 활용 데이터

입수 데이터	활용 유무	생성 주기	지역 속성	제공처
산업단지 현황보고서	X	연간	단지주소 (동/면)	공공데이터 포털
국가산업단지 산업동향정보	O	월간	산업단지	공공데이터 포털
전국산업단지 현황통계	O	분기	시도, 시군, 단지명	공공데이터 포털
공장등록 현황 통계정보	X	월간	시도명	공공데이터 포털
공장등록 통계	X	월간 (요청 자료)	시도명	한국산업단지공단
전국등록공장 현황	O	월간 (요청 자료)	공장주소	한국산업단지공단
전국 지식산업센터현황	O	월간 (요청 자료)	공장대표주소 (도로명), 공장대표주소 (지번)	한국산업단지공단
빅데이터 기반의 공장설립 온라인지원시스템 확대 구축을 위한 마스터플랜(ISMP)	X	없음 (요청 자료)	없음	한국산업단지공단

[표 2-1] 입수 및 활용 데이터

나. 활용 데이터 상세 소개

1) 국가산업단지 통계정보

생성주기	기간	시트명	표명	컬럼
월간	2011.01 ~ 2020.06	표1 단지별 입주	-	산업단지, 구분, 입주(당월, 전월), 가동(당월, 전월), 임차
		표2 업종별 입주	-	산업단지, 구분, 음식료, 섬유의복, 목재종이, 석유화학, 비금속, 철강, 기계, 전기전자, 운송장비, 기타, 비제조, 총계
		표3 업종별 가동	-	산업단지, 구분, 음식료, 섬유의복, 목재종이, 석유화학, 비금속, 철강, 기계, 전기전자, 운송장비, 기타, 비제조, 총계
		표4 단지별 생산	-	산업단지, 구분, 당월, 전월, 2020 누계, 증감률(전월대비)
		표5 업종별 생산	-	산업단지, 구분, 음식료, 섬유의복, 목재종이, 석유화학, 비금속, 철강, 기계, 전기전자, 운송장비, 기타, 계
		표6 단지별 수출	-	산업단지, 구분, 당월, 전월, 2020 누계, 증감률(전월대비)
		표7 업종별 수출	-	산업단지, 구분, 음식료, 섬유의복, 목재종이, 석유화학, 비금속, 철강, 기계, 전기전자, 운송장비, 기타, 계
		표8 단지별 고용	-	산업단지, 구분, 당월(계, 남, 여), 전월, 전월대비
		표9 업종별 고용	-	산업단지, 구분, 음식료, 섬유의복, 목재종이, 석유화학, 비금속, 철강, 기계, 전기전자, 운송장비, 기타, 비제조, 계
		표10 단지별 가동률	-	산업단지, 구분, 제조업 가동업체수(개사), '생산(백만원)'(최대생산능력, 당월생산액), '가동률(%)'(당월, 전월, 전월대비(%p))
		표11 가동률 세부내역	-	산업단지, 구분, '가동률(%)'(50인 미만 기업, 50인 이상~300인 미만 기업, 300인 이상 기업, 계), '가동업체수(개사)'(50인 미만 기업, 50인 이상~300인 미만 기업, 300인 이상 기업, 계)
		표12 업종별 가동률	-	산업단지, 구분, 음식료, 섬유의복, 목재종이, 석유화학, 비금속, 철강, 기계, 전기전자, 운송장비, 기타, 계

[표 2-2] 국가산업단지 통계정보 데이터

2) 전국산업단지 현황 통계

생성주기	기간	시트명	표명	컬럼
분기	2011.01 ~ 2020.01	요약	(1) 조성 및 분양(천㎡)	단지유형, 단지수, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률(%))
			(2) 입주 및 고용	단지유형, 입주계약 업체(A), 가동업체(B), 남, 녀, 고용(계)
			(3) 생산 및 수출	단지유형, 생 산(20.1분기, 19.1분기, 증감률(%)), 수출(20.1분기, 19.1분기, 증감률(%))
		시도별	-	구분, 단지수, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, 고용, 누계생산(백만원), 누계수출(천달러)
		전국산업단지 현황	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		국가	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		일반	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		도시첨단	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		농공	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		부록1)신규지정 및 해제현황	2020년 1분기 산업단지 신규지정 현황	유형, 시도, 시군구, 단지명, 지정면적, 산업용지, 지정일자, 비고
			2020년 1분기 산업단지 지정해제 현황	유형, 시도, 시군구, 단지명, 지정면적, 산업용지, 지정일자, 해제일자
		부록2-1)자유무역	-	시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		부록2-2)외국인투자지역	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		부록3)2개시도에 걸친 산단	-	상위단지, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)
		부록4)노후산업단지	-	유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 산업시설구역(전체면적, 분양대상, 분양, 미분양, 분양률), 입주업체, 가동업체, '고용현황(명)'(남, 여, 계), 누계생산(백만원), 누계수출(천달러)

[표 2-3] 전국산업단지 현황 통계 데이터

3) 전국공장현황

생성주기	기간	시트명	표명	컬럼
반기	2011.01 ~ 2020.01	Sheet1	-	기준연도, 시도명, 시군구명, 관리기관, 회사명, 공장구분, 단지명, 설립구분, 입주형태, 보유구분, 최초등록일, 전화번호, 남자종업원, 여자종업원, 외국인남자종업원, 외국인여자종업원, 종업원합계, 생산품, 원자재, 사업시작일, 공장규모, 용도지역, 지목, 용지면적, 제조시설면적, 부대시설면적, 건축면적, 대기오염물질, 수질오염물질, 소음진동, 생활용수, 산업용수, 전력, 석유, 가스, 기타, 외국인투자비율, 외국인투자액, 홈페이지, 지식산업센터명, 대표업종, 업종명, 업종코드, 공장주소우편번호, 공장주소, 공장주소_지번, 자기자본액, 타인자본액

[표 2-4] 전국공장현황 데이터

4) 지식산업센터 현황

생성주기	기간	시트명	표명	컬럼
월간	2020년 07월 기준	NO2_전국_지식산업센터현황_202007	-	시도, 시군구, 지식산업센터명, 입지구분, 회사명, 등록구분, 단지명, 관할기관, 산단구분, 상태, 최초승인일, 승인일, 최초등록일, 등록일, 준공일, 착공예정일, 준공예정일, 사업시작일, 전화번호, 지목, 자기자본액, 타인자본액, 외국인투자비율, 외국인투자금액, 취소일, 취소사유, 홈페이지, 용지면적, 건축면적, 제조면적, 부대면적, 공장대표주소(도로명), 공장대표주소(지번), 분양형태, 건축상태, 용도지역1, 용도지역2, 용도지역3, 설치자, 입주업체, 종업원수

[표 2-5] 지식산업센터 현황 데이터

다. 통합 데이터 가공

1) 1차 가공

가) 국가산업단지

□ 필요 없는 시트 제거 및 통합

- 불필요 시트 제거
- 남은 시트를 전부 변수로 추가해 하나의 시트로 통합
- “날짜” 변수를 추가해서 날짜별로 구분

□ 파일 생성

- 월별로 파일을 생성 (2011.01 ~ 2020.06)
- 파일마다 산업단지, 입주, 가동, 입주_음식료, …, 가동률_기타, 날짜와 같은 컬럼으로 구성

나) 전국산업단지

□ 필요 없는 시트 제거 및 통합

- 전국산업단지 현황시트가 나머지 시트를 포함하고 있는 구조
- 전국산업단지 현황시트에 자유무역, 외국인투자지역, 2개 시도에 걸친 산단, 노후 산업단지를 컬럼으로 추가 후 현재 시트만 활용
- 들여쓰기 되어있는 산업단지들 들여쓰기 제거, 전국산업단지 시트에 파란색으로 쓴 산업단지들 제거
- “날짜” 변수를 추가해서 날짜별로 구분

□ 파일 생성

- 분기별로 파일 생성 (2011.01 ~ 2020.01)
- 파일마다 유형, 시도, 시군, 단지명, 조성상태, 지정면적, 관리면적, 미분양, …, 노후산단, 날짜와 같은 컬럼으로 구성

다) 전국공장현황

□ 필요 없는 컬럼 제거 및 전월대비 변동사항 체크

- 분석에 필요 없는 컬럼(전화번호, 인터넷 주소 등) 제거
- 전월 대비 변동사항 체크 및 컬럼 추가

□ 파일 생성

- 파일 하나당 row 개수가 100만개가 넘고 월별로 되어있어 Resource 문제가 클 것으로 판단됨. 하나의 파일로 만들기는 무리라고 판단되어 하나의 파일로 통합하지 않음
- 파일마다 기준연도, 시도, 시군구명, 회사명, 공장구분, 단지명, 설립구분, 입주형태, 최초등록일, …, 공장주소, 업종명, 대표업종, 전월대비 변동사항과 같은 컬럼으로 구성

2) 2차 가공

가) 국가산업단지

□ 파일 통합

- 월별로 되어있는 파일들을 하나의 파일로 통합
- 기존 국가산업단지 파일을 다 통합한 것이므로 row 개수만 늘어나게 됨

나) 전국산업단지

□ 파일 통합

- 분기별로 되어있는 파일들을 하나로 통합
- 기존 전국산업단지 파일을 다 통합한 것이므로 row 개수만 늘어나게 됨

3) 3차 가공(최종)

가) 국가산업단지

	내용
입주	산업단지별 입주현황
가동	산업단지별 입주현황
입주_음식료	산업단지별 음식료 업종의 입주현황
입주_섬유의복	산업단지별 섬유의복 업종의 입주현황
입주_목재종이	산업단지별 목재종이 업종의 입주현황
입주_석유화학	산업단지별 석유화학 업종의 입주현황
입주_비금속	산업단지별 비금속 업종의 입주현황
입주_철강	산업단지별 철강 업종의 입주현황
입주_기계	산업단지별 음식료 업종의 입주현황
입주_전기전자	산업단지별 기계 업종의 입주현황
입주_운송장비	산업단지별 전기전자 업종의 입주현황
입주_기타	산업단지별 기타 업종의 입주현황
입주_비제조	산업단지별 비제조 업종의 입주현황
가동_음식료	산업단지별 음식료 업종의 가동현황
⋮	⋮
가동률_전기전자	산업단지별 전기전자 업종의 가동률현황
가동률_운송장비	산업단지별 운송장비 업종의 가동률현황
가동률_기타	산업단지별 기타 업종의 가동률현황
날짜	날짜

[표 2-6] 국가산업단지 최종 가공 데이터

□ 파일 개수: 1개

□ 컬럼은 표와 같이 구성되어있고 업종별로 입주, 가동, 생산, 수출, 고용, 가동률로 나뉘어져 있음

나) 전체산업단지

	내용
유형	산업단지의 유형(일반, 국가, 도시첨단, 농공)
시도	행정구역 시/도
시군구	행정구역 시/군/구
단지명	산업단지명
조성상태	산업단지의 조성상황
지정면적	산업단지의 지정면적
관리면적	관리하고 있는 산업단지의 면적
산업시설구역_전체면적	산업단지내에 산업시설구역의 대한 면적
산업시설구역_분양대상	산업시설구역의 대한 조성면적
산업시설구역_분양	산업시설구역의 대한 분양된 면적
산업시설구역_미분양	산업시설구역_분양대상-산업시설구역_분양
산업시설구역_분양률	분양대상 ÷ 분양
자유무역	자유무역 여부
외국인투자지역	외국인투자지역 여부
2개의 시도의 걸친지역	2개의 시도의 걸친지역 여부
노후산업단지	노후산업단지 여부
날짜	날짜

[표 2-7] 전체산업단지 최종 가공 데이터

□ 파일 개수: 1개

다) 전국공장현황

컬럼	내용
기준연도	연도와 월 (예 : 202008)
시도명	시/도명
회사명	회사명
공장구분	계획인지 개별인지 여부
단지명	소속되어있는 산업단지명
설립구분	설립될 때 형태구분
입주형태	산업단지별 석유화학 업종의 입주현황
최초등록일	최초 공장등록일
남자종업원	남 종업원의 수
여자종업원	여 종업원의 수
외국인남자	외국인 남 종업원에 수
외국인여자	외국인 여 종업원의 수
종업원합계	전체 종업원의 수
생산품	공장의 생산 물품
공장규모	공장의 규모(대, 중, 소)
⋮	⋮
외국인투자액	외국인투자액
외국인투자비율	외국인투자비율
업종명	업종구분
공장주소	해당 공장 주소
날짜	날짜

[표 2-8] 전국공장현황 최종 가공 데이터

□ 파일 개수: 20개

□ 컬럼은 기준연도, 시도, 시군구명, 회사명, 공장구분, 단지명, 설립구분, 입주형태, 최초등록일, …, 공장주소, 업종명, 대표업종으로 구성

라. 분석용 데이터 가공

1) 코로나19 바이러스로 인한 국가산업단지의 변화를 파악한다.

가) 문제 및 데이터 이해

□ 코로나19의 발생시점부터 현재까지의 월별 데이터 적용 (2011.02 ~ 2020.08, 09월~01월 제외)

□ 입주, 가동, 생산, 수출, 고용 등의 변수 필요

나) 분석용 데이터 가공 방법

□ 날짜 컬럼을 기준으로 02월부터 08월까지의 데이터만 추출

□ 값이 'x'인 것은 업체가 몇 개 없어서 값을 제공 하지 않는 데이터이므로 해당 데이터들은 제거

□ 값이 비어있는 것은 해당 산업에 공장이 없는 것으로 전부 0으로 처리

2) 국가산업단지가 전체 산단에서 갖는 영향력을 분석 및 예측 한다.

가) 문제 및 데이터 이해

□ 전국산업단지 데이터 안에 국가, 일반, 도시첨단, 농공과 같은 산업단지 유형 별로 구분이 되어있다. 분석 대상이 전체산업단지와 국가산업단지이고 컬럼이 입주, 가동, 생산, 수출, 고용 등의 변수가 필요함. 즉, 모든 데이터를 필요로 하므로 통합 데이터를 그대로 사용

나) 분석용 데이터 가공 방법

□ 값이 'x'인 것은 업체가 몇 개 없어서 값을 제공 하지 않는 데이터이므로 해당 데이터들은 제거

□ 지정면적과 관리면적은 데이터 분석에 불필요한 컬럼이므로 제거

3) 국가산업단지의 자원 사용량을 분석하여 신규 입주 기업의 자원 사용량을 예측한다.

가) 문제 및 데이터 이해

□ 시트가 월별로 되어있어서 월별로 되어있는 것들을 하나로 통합

□ 자원 사용량을 파악하기 위해서는 전국공장현황 데이터가 필요

□ 국가산업단지에 입주해있는 공장만 추출

□ 사업장 데이터(업종, 면적, 공장규모 …, 생활용수, 산업용수, 전력, 석유, 가스, 기타 등)가 필요

□ 입주 공장 중 국가산업단지에 입주해있는 공장들만 추출하고 월별을 하나의 데이터로 통합

나) 분석용 데이터 가공 방법

☐ 전국공장현황 데이터에서 국가산업단지에 입주해 있는 공장들만 추출

☐ 공장구분은 불필요한 컬럼이므로 제거

☐ 국가산업단지에 입주해있는 공장들만 추출하고 월별을 하나의 데이터로 통합

4) 전국 공장들의 등록 및 해제 현황을 분석하여 앞으로의 공장 현황을 예측한다.

가) 문제 및 데이터 이해

☐ 등록 및 해제 현황을 보기 위해서는 전월 대비 변동사항 컬럼 필요

☐ 최초등록일, [해제일], 남자종업원, 여자종업원, 면적, 공장규모, 자원사용량 등 컬럼 필요

☐ 빈값들에 대한 해결 방법 필요

나) 분석용 데이터 가공 방법

☐ 빈값들을 전부 0으로 채움

☐ 전월 대비 변동사항 컬럼을 추가하고, 월별로 전월 대비 변동사항을 병합

3. 분석 주제 소개 및 상세 프로세스

가. 주요 분석 프로세스

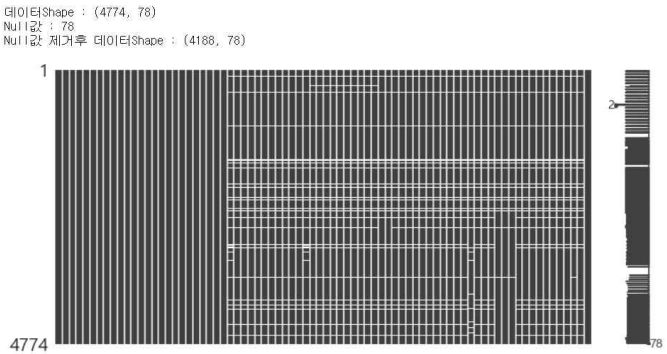
[그림 3-1] 주요 분석 프로세스

나. 주제별 분석 상세 프로세스

1) 주제 1번 - 코로나19 바이러스로 인한 국가산업단지의 변화 파악

[그림 3-2] 데이터 분석 프로세스

가) 데이터 체크



[그림 3-3] 데이터 체크 Null

☐ 데이터를 체크한다. 데이터의 형태, Null의 유무

☐ Null이 있어서 drop시켰다. 분석하는데 크게 지장이 없어보인다.

나) 분석에 필요한 변수 만들기

☐ 코로나 바이러스 전과 후를 비교 할 것이므로 2월~8월을 빼서 코로나와 관련된 날짜로 구성된 새로운 테이블 구성했다.

☐ 산업단지별 가장 주력으로 하는 업종을 찾기 위해 입주현황 칼럼에서 가장 많이 분포하는 업종을 뽑아서 새로운 변수로 구성했다.

다) 월별 공장 입주 동향

☐ 월별로 정리 되어있는 국가산업단지 데이터를 불러와서 전체 국가산업단지에 입주 되어있는 공장 수를 파악

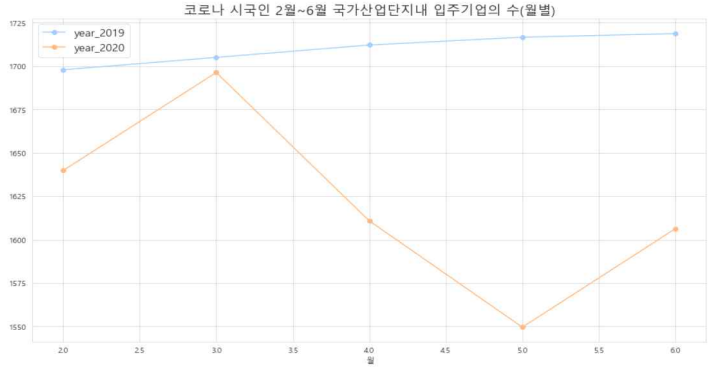
[그림 3-4] 연도별 국가산업단지내 입주기업의 수

☐ 파란색 선은 연도별 1월부터 12월까지의 평균 입주기업의 수이고 주황색 점선은 코로나 시점인 연도별 2월부터 8월까지의 평균 입주기업 수이다.

☐ 1월부터 12월까지 입주수와 2월부터 8월까지의 입주수가 크게 다르지 않다.

☐ 전체적으로 입주기업의 수는 늘고 있다. 하지만 2019년, 2020년에 입주기업의 수가 확 줄었다.

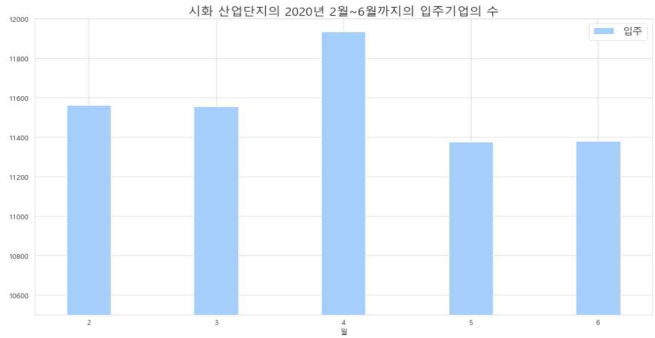
☐ 2019년과 2020년의 입주기업의 수를 비교해봤다.



[그림 3-5] 2019년과 2020년 2월부터 6월까지의 입주기업의 수

☐ 2019년에는 꾸준하게 입주기업이 늘어나고 있다. 반면 2020년에는 3월을 기점으로 계속 입주기업의 수가 줄고 있다. 코로나 시기의 전년(2019년)과는 확연하게 다른 차이가 보인다.

☐ 2월부터 6월까지 입주기업의 수가 가장 많이 변동을 보인 산업단지는 시화 국가산업단지이다. 시화국가산업단지의 경우 2월부터 6월까지 560개의 기업이 줄었다.



[그림 3-6] 시화국가산업단지의 2020년 2월부터 6월까지의 입주기업 수

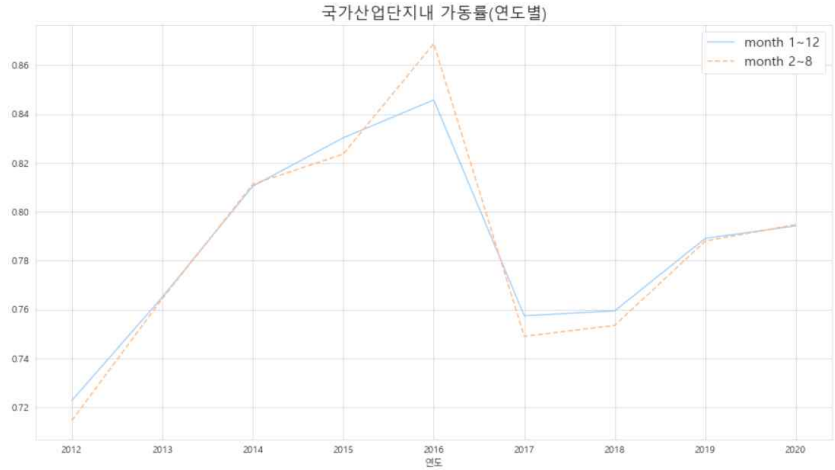
□ 코로나의 대한 타격이 늦게 오는지 4월에 갑자기 입주기업의 수가 늘었다. 하지만 코로나가 한창 진행중일 때 입주기업의 수가 크게 준 것으로 보아 피해가 크다고 볼 수 있다.

산업단지	입주
시화	560.0
서울	220.0
반월	134.0
광주첨단	84.0
장항생태	74.0

[표 3-1] 2020년 2월부터 6월까지 입주기업수가 많이 줄어든 산업단지

□ 코로나로 타격을 많이 입은 산업단지 5개를 뽑아봤는데 시화가 가장 많이 타격을 입고 서울, 반월, 광주첨단, 장항생태 순이다.

라) 월별 가동률 동향

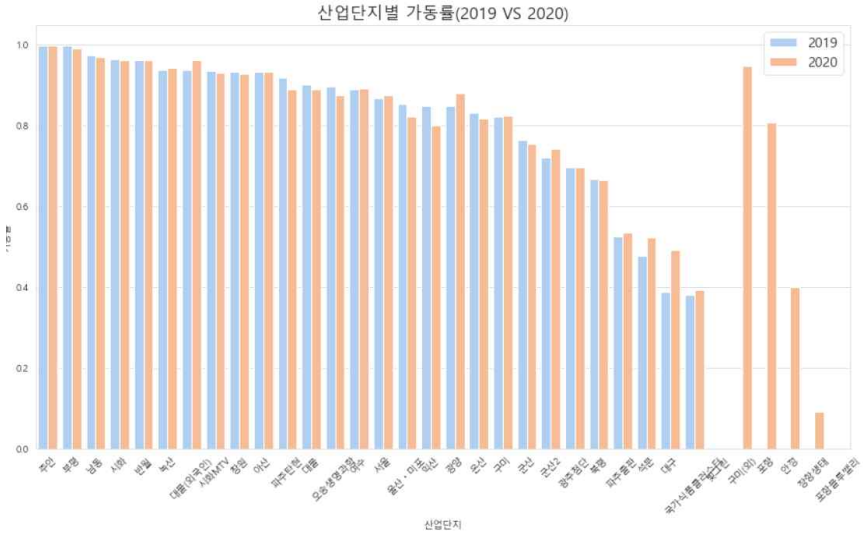


[그림 3-7] 연도별 국가산업단지내 가동률

□ 파란색 선은 연도별 1월부터 12월까지의 평균 가동률이고 주황색 점선은 코로나 시점인 연도별 2월부터 8월까지의 평균 가동률 이다.

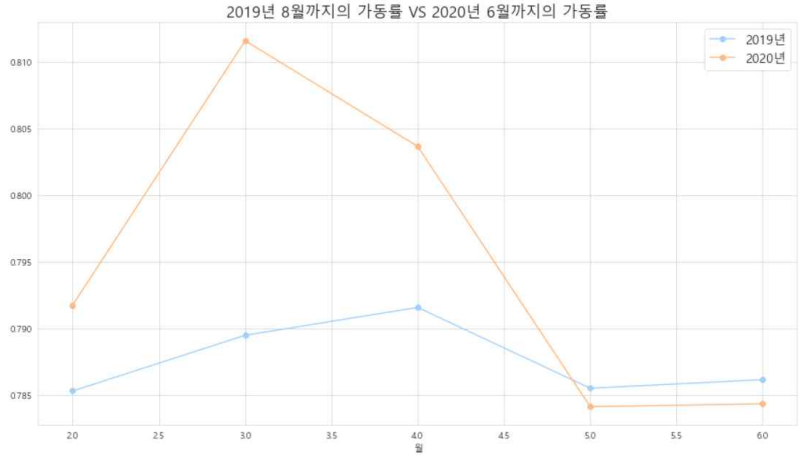
□ 전반적으로 1월부터 12월, 2월부터 8월까지의 가동률은 크게 차이가 나지 않아보인다.

□ 2019년과 2020년 사이에 가동률이 늘었다. 코로나 시국인 2019년 2020년인데 왜 그런것일까 알아볼 필요가 있어보인다.



[그림 3-8] 2019년 2020년 산업단지별 가동률

□ 2019년과 2020년 산업단지별 가동률이다. 2019년 대비 2020년에 갑자기 가동률이 늘어난 것은 2019년에 측정되지 않았던 산업단지들이 2020년에 측정되어서 평균을 높인 것으로 판단되고 전반적으로 2019년보다 2020년에 가동률이 높아졌다.



[그림 3-9] 2019년 2020년 월별 가동률

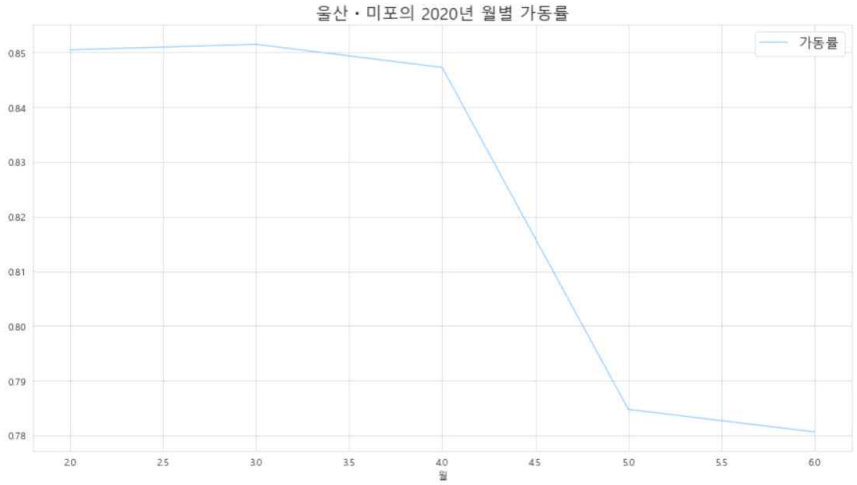
□ 코로나의 여파가 조금 느리게 오는 것은 확실하다. 코로나가 절정이 시기

였던 3월에 가동률은 최고였고 코로나가 조금 잠잠했던 5월에 가동률이 많이 떨어졌다.

□ 4월까지는 평년도보다 높았지만 갑자기 5월부터 평년보다 가동률이 낮아졌다.

□ 위 그래프에서 전체적인 평균치로 판단해서 5월, 6월의 데이터가 전체 데이터에 크게 영향을 미치지 않은 것 같다.

□ 가동률은 2019년에는 평균적으로 0.6% 줄었고 2020년에는 2.7% 줄었다. 월 평균으로 놓고 보니 4배 이상 차이가 난다. 5월이 가장 많이 가동률이 낮아졌으므로 2020년 5월을 분석해볼 필요가 있다.



[그림 3-10] 2020년 울산、미포 월 별 가동률

□ 4월에서 5월 넘어갈 때 가장 많이 가동률이 하락한 곳은 울산、미포 이다. 6.25% 정도 하락했다.

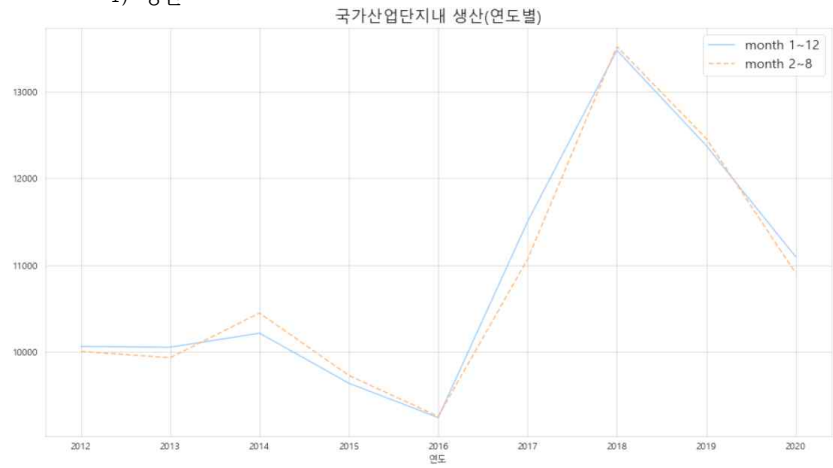
산업단지	가동률(%)
장항생태	18.1818
광주첨단	7.6558
울산、미포	7.0860
파주탄현	5.5233
구미(외)	4.1667

[표 3-2] 2020년 2월부터 6월까지 가동률이 많이 줄어든 산업단지

□ 장항생태가 가동률이 가장 많이 줄어들었다. 18%정도가 가동을 중지했고 광주첨단, 울산미포, 파주탄현, 구미(외) 순으로 가동률이 많이 낮아졌다.

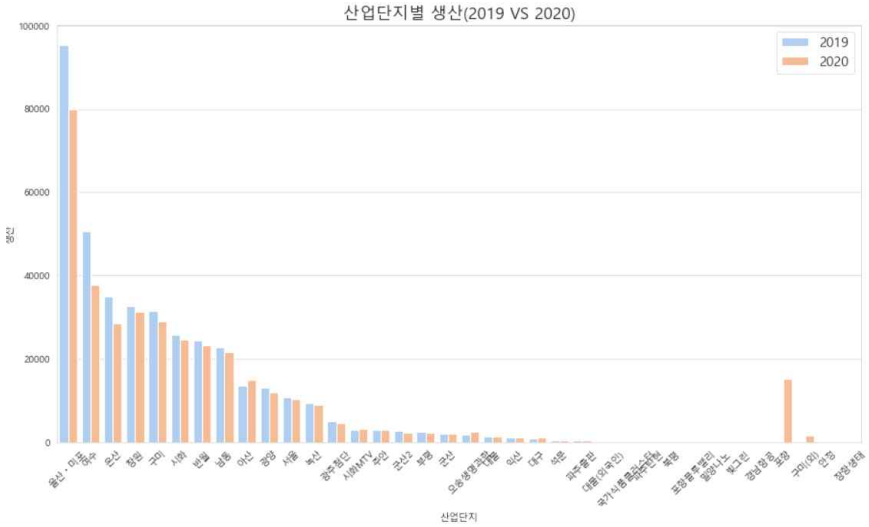
마) 월별 생산/수출 동향

1) 생산



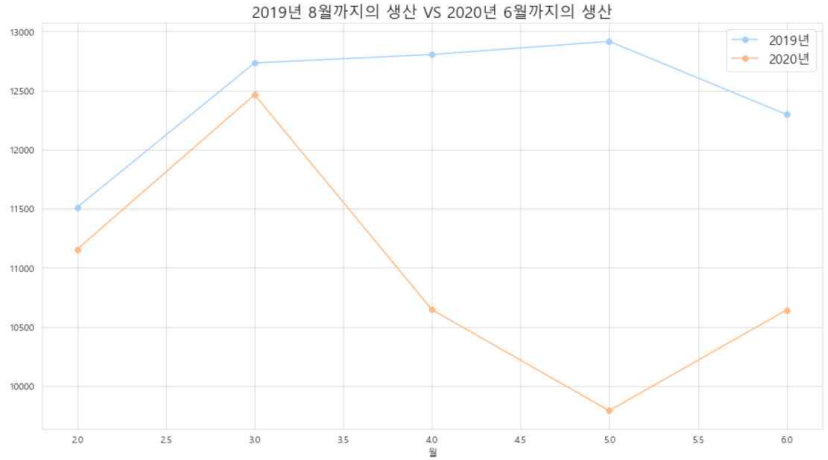
[그림 3-11] 연도별 국가산업단지 내 생산

□ 1월부터 12월, 2월부터 8월까지는 비교적 크게 차이 안난다. 하지만 2018년에서 2020년 사이에 국가산업단지 생산이 확 줄었다.



[그림 3-12] 2019년, 2020년 국가산업단지별 생산

□ 아산, 오송 생명 과학단지와 몇 개를 제외하면 2019년 대비 2020년에 생산이 확 줄었다는 것을 알 수 있다.



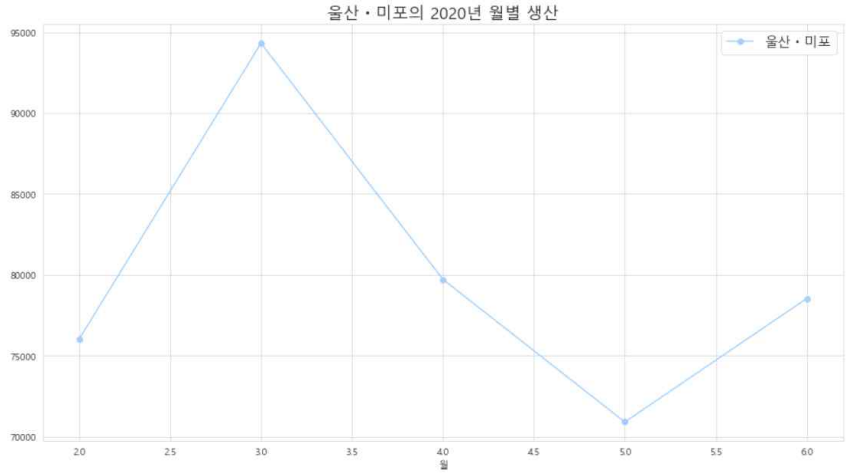
[그림 3-13] 2019년, 2020년 월별 생산

□ 생산 같은 경우 바로 코로나에 타격을 입는 것을 알 수 있다.

□ 3월에 정점을 찍고 계속 하락한다. 5월에서 6월 넘어가는 시점에 조금 생산이 증가하지만 코로나 이전인 2019년과 비교하면 턱없이 부족하다.

□ 2019년에는 월 평균 생산이 1428정도의 차이를 보였지만 2020년에는 2672의 차이를 보인다.

□ 생산이 급격하게 떨어진 2020년 3월에서 5월 사이를 조사해 보았더니 울산、미포 국가산업단지가 생산이 가장 줄었다.



[그림 3-14] 울산、미포 국가산업단지의 2020년 월별 생산

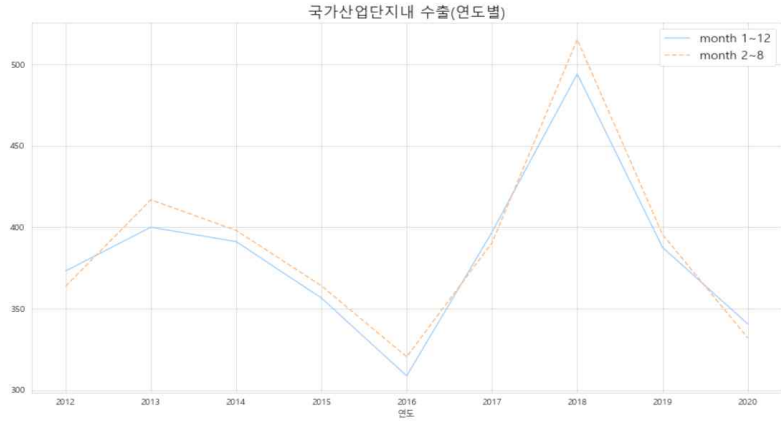
□ 2020년 울산、미포의 생산은 3월부터 쭉 떨어지더니 5월에 회복을 못하다가 6월에 회복하는 모습을 보여준다.

산업단지	생산
울산、미포	23419.36828
여수	13368.13124
구미	8259.76152
아산	6253.38493
온산	5940.96687

[표 3-3] 2020년 2월부터 6월까지 생산이 많이 줄어든 산업단지

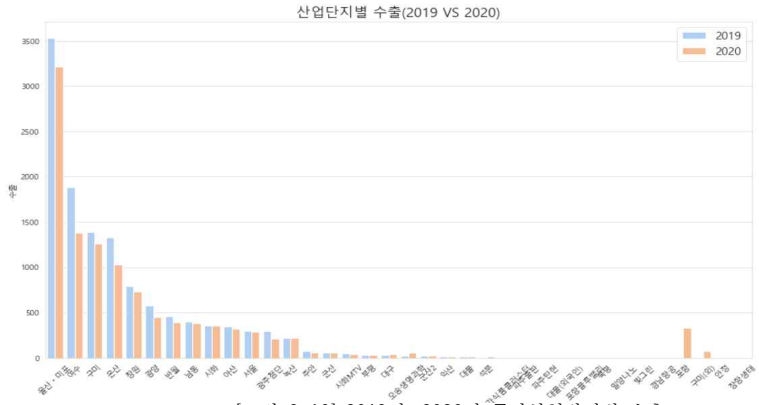
□ 울산、미포가 생산이 가장 많이 줄어들었고 여수, 구미, 아산, 온산 순으로 생산이 많이 줄어들었다.

2) 수출



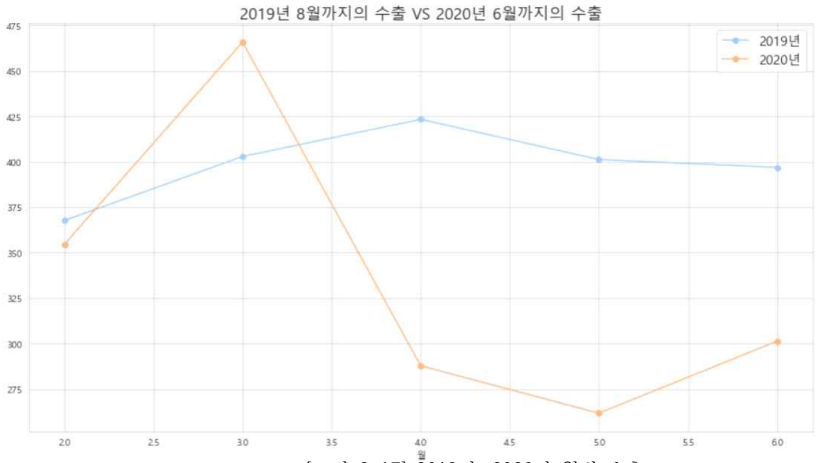
[그림 3-15] 연도별 국가산업단지 내 수출

□ 1월부터 12월, 2월부터 8월까지는 비교적 크게 차이 안난다. 하지만 2018년에서 2020년 사이에 국가산업단지 생산이 확 줄었다.



[그림 3-16] 2019년, 2020년 국가산업단지별 수출

□ 많은 산업단지가 2019년 대비 2020년 전체적으로 수출이 많이 줄었다.

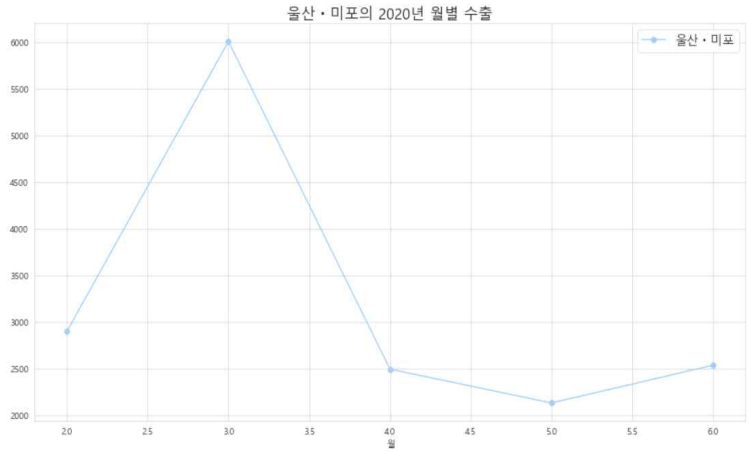


[그림 3-17] 2019년, 2020년 월별 수출

□ 월별로 분리해서 보면 2019년에는 급격하게 수출이 줄어들지 않은 반면 2020년에는 3월을 기준으로 급격하게 수출이 줄어든 것을 확인할 수 있다. 수출에서는 코로나에 대한 영향이 바로 나타나는 것으로 판단된다.

□ 2020년 6월에는 조금 수출이 올라갔지만 전반적으로 전년도에 비하면 수출이 많이 떨어진 상태다.

□ 2019년 월 평균 55 정도 떨어진 반면, 2020년에는 월 평균 204 정도 수출이 떨어졌다.



[그림 3-18] 울산、미포 국가산업단지의 2020 월별 수출

□ 국가산업단지중 울산、미포 국가산업단지의 수출이 코로나로 인해 가장 많은 타격을 입었다.

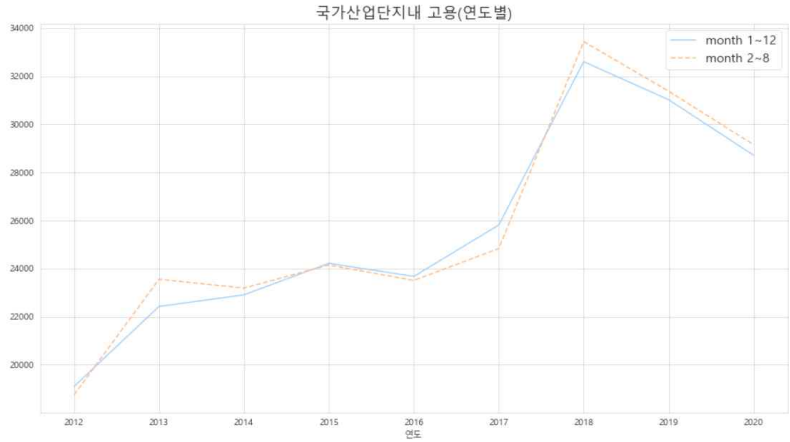
□ 3월을 기점으로 5월까지 수출이 하락했다. 6월에 조금 증가하긴 했지만 전반적으로 하락세가 큰 편이다.

산업단지	수출
울산、미포	3876.739000
여수	582.044318
온산	573.742607
구미	518.937529
광양	211.185000

[표 3-4] 2020년 2월부터 6월까지 수출이 많이 줄어든 산업단지

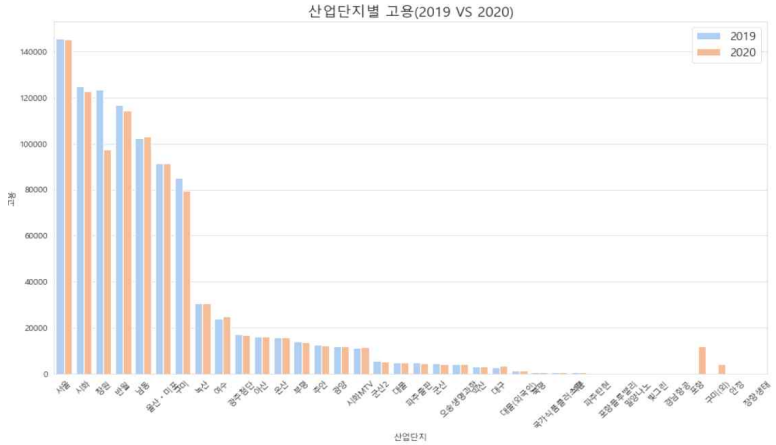
□ 표를 보면 울산、미포가 수출이 가장 많이 줄어들었고 여수, 온산, 구미, 광양 순으로 수출이 많이 줄어들었다.

바) 월별 고용 동향



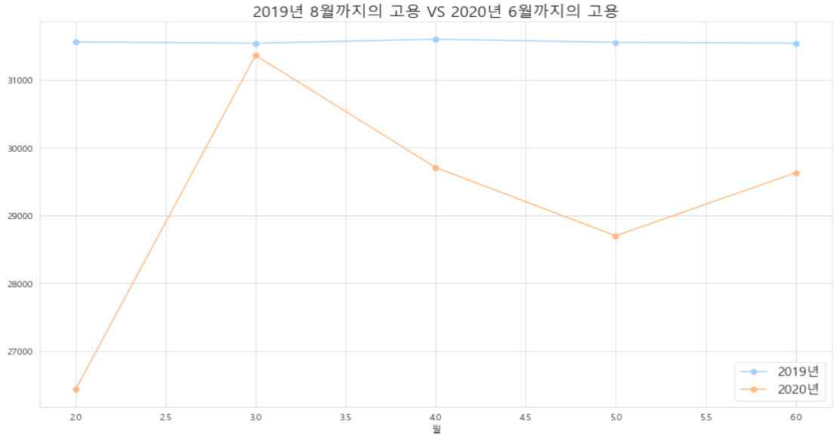
[그림 3-19] 연도별 국가산업단지내 고용

□ 1월부터 12월, 2월부터 8월까지는 비교적 크게 차이가 나지 않는다. 2012년부터 2017년까지 고용은 계속 늘고 있었는데 2018년에 고용이 갑자기 줄어들기 시작했다.



[그림 3-20] 2019년, 2020년 산업단지별 고용

□ 많은 국가산업단지가 2019년과 2020년에 비슷하게 고용을 유지하고 있다.

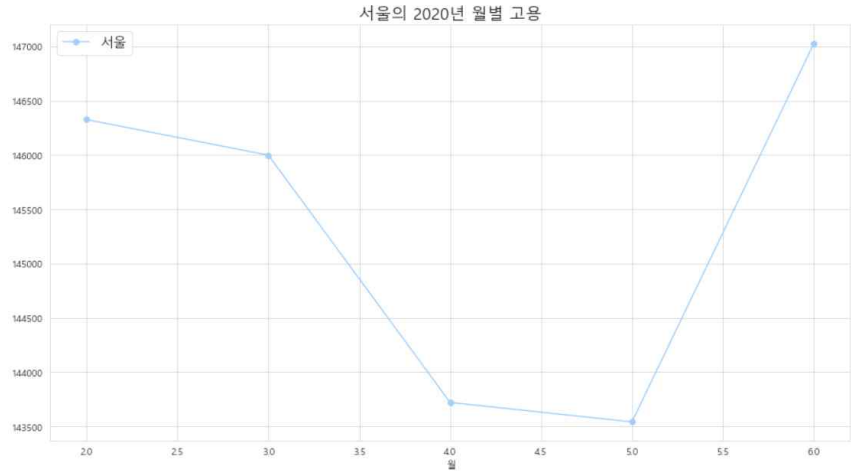


[그림 3-21] 2019년, 2020년 월별 고용

□ 2019년과 2020년 월별 고용현황을 보니 2019년에는 평이하게 지나간 반면 2020년에는 3월부터 5월까지는 고용이 하락하고 6월에는 다시 상승하고 있다.

□ 3월부터 많은 코로나 확진자들이 나왔다. 고용은 3월부터 하락했으니 코로나로 인한 타격이 바로 나타난 것으로 보인다.

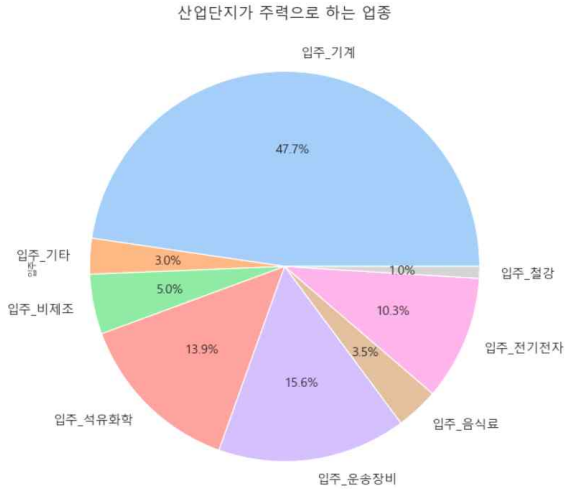
□ 2019년 고용은 월 평균 1149 차이, 2020년 고용은 월 평균 4922 차이난다.



[그림 3-22] 2020년 서울 국가산업단지 월별 고용

□ 가장 많이 차이나는 곳은 서울인데 3월에서 5월까지는 고용이 하락, 6월부터는 고용의 상승폭이 높다.

사) 월별 산업단지별 주력 업종 동향

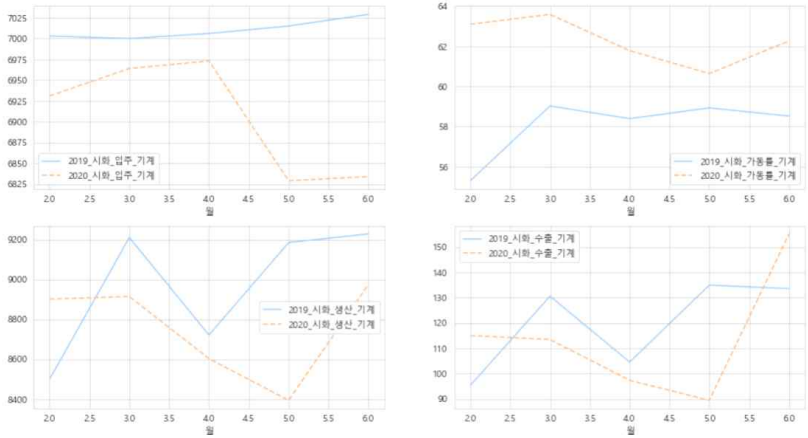


[그림 3-23] 2020년 서울 국가산업단지 월별 고용

□ 전체 국가산업단지에서 주력으로 삼는 산업들을 비교 해보니 절반정도가 기계를 주력으로 삼고 있다. 그 다음 운송장비, 석유화학, 전기전자 순이다. 산업들이 많기 때문에 기계, 운송장비, 석유화학, 전기전자 만 보겠다.

1) 기계

□ 기계를 주력으로 삼는 곳 중 가장 큰 규모인 곳은 시화 국가산업단지다.



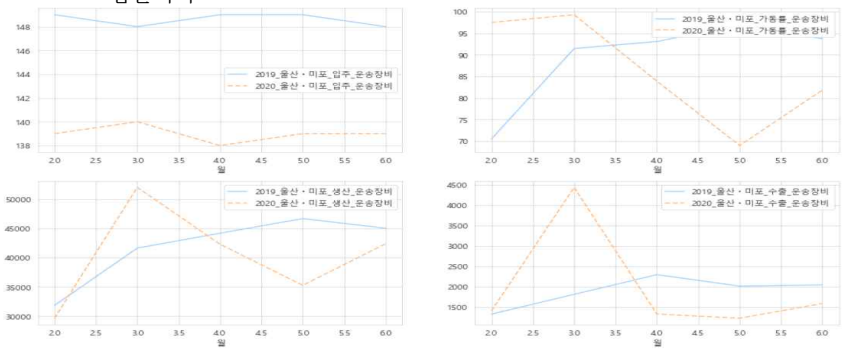
[그림 3-24] 시화 국가산업단지의 2019년, 2020년 기계업종 현황

□ 입주는 2020년 5월을 기준으로 많이 줄었다. 가동률은 60%이상 되었으며 3월 직후 가동률이 줄어들긴 했지만 그래도 전년도보다 높은 가동률을 보이고 있다. 하지만 생산에서는 높은 가동률에 비해 고전을 줄어들고 있다. 전년도보다 많이 떨어진 모습이 보인다. 수출은 5월에 최저점을 찍고 6월에 전년도를 뛰어넘는 성과를 보여주고 있다.

□ 시화산업단지는 코로나의 영향으로 주력산업인 기계에서 5월이후 높은 수출을 보여주며 위기를 극복해 나가고 있다.

2) 운송장비

□ 운송장비를 주력으로 삼는 곳 중 가장 큰 규모인 곳은 울산·미포 국가산업단지다



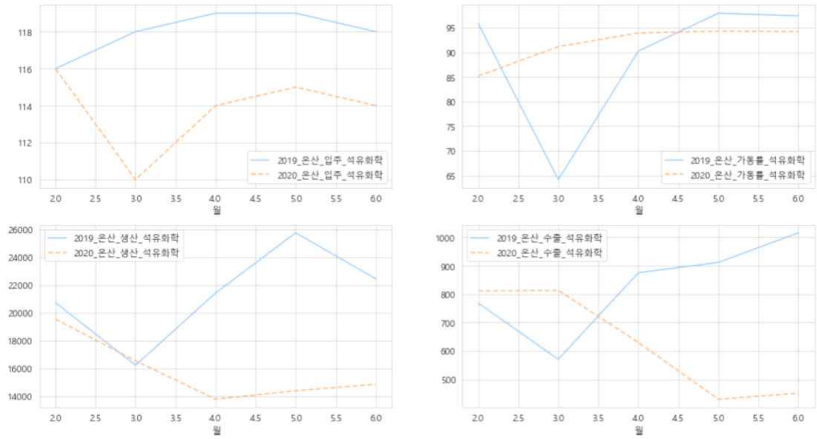
[그림 3-25] 울산.미포 국가산업단지의 2019년, 2020년 운송장비 업종 현황

□ 입주는 2020년에 많이 떨어진 상태로 사직했다. 코로나의 여파로 입주기업이 많이 줄지는 않았다. 하지만 가동률에서는 2019년 2020년 100%에 가까운 가동률을 보여주다 5월에 60%후반까지 내려갔다. 생산에서는 떨어진 가동률과 더불어서 2020년 3월부터 쭉 떨어지고 있다. 5만대에서 3만대 중반으로 떨어지며 많은 타격을 입고있음이 확인된다. 수출 역시 2020년 3월을 이후로 많은 타격을 입고 있다. 2020년 3월에는 전년도와 대비해 3배에 가까운 차이를 냈고 2020년 4월부터는 전년도보다 더 떨어졌다.

□ 울산·미포산업 단지는 코로나의 영향으로 주력산업인 운송장비에서 크게 타격을 입고 있다.

3) 석유화학

□ 석유화학을 주력으로 삼는 곳 중 가장 큰 규모인 곳은 온산 국가산업단지다.



[그림 3-26] 온산 국가산업단지의 2019년, 2020년 석유화학 업종 현황

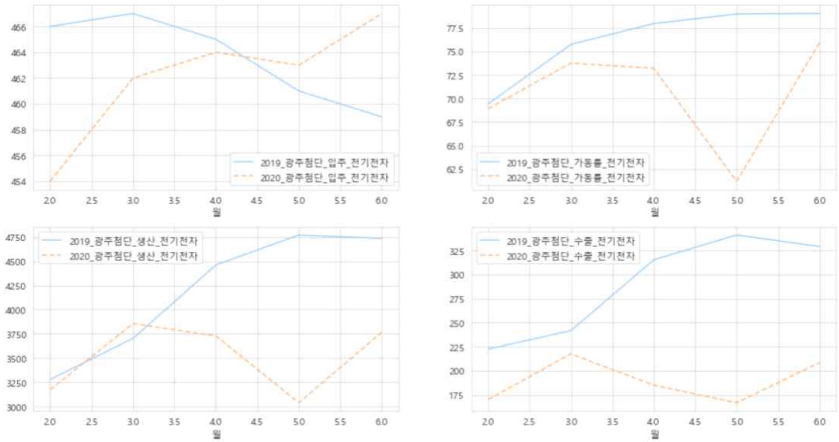
□ 입주는 2020년에 많이 떨어진 상태로 사직했다. 코로나의 여파로 입주기업이 3월에 팍 줄었고 그 뒤로 조금씩 늘어났다. 가동률에서는 준수한 상황이다. 90%대 가동률을 보이고 있다. 생산에서는 다른 산업단지와 마찬가지로 바닥을 치고 있으며 10000이상이 빠졌다. 타격이 심각하다. 수출 역시 절반정도로 줄었으며 6월까지 회복을 못하고 있다.

□ 온산의 경우 가동률은 준수한편이지만 수출과 생산이 많이 떨어졌다.

□ 가동률이 좋은데 왜 생산과, 수출이 많이 떨어졌을까? 해당 문제에 대한 조사가 필요할 듯 하다.

4) 전기전자

□ 전기전자를 주력으로 삼는 곳 중 가장 큰 규모인 곳은 광주첨단 국가산업단지다.



[그림 3-27] 광주첨단 국가산업단지의 2019년, 2020년 전기전자 업종 현황

□ 전기전자를 주력으로 삼는 곳 중 가장 큰 규모인 곳은 광주첨단 국가산업단지다.

□ 입주는 코로나 여파인데도 불구하고 늘어나고 있다. 조성중인 산업단지인지 확인이 필요하다. 가동률에서는 5월에 바닥을 찍었는데 6월에 다시 회복하는 모습을 보여준다. 생산에서는 3월이후로 하락하고 있다. 수출 역시 전년도보다 절반정도의 수준이다.

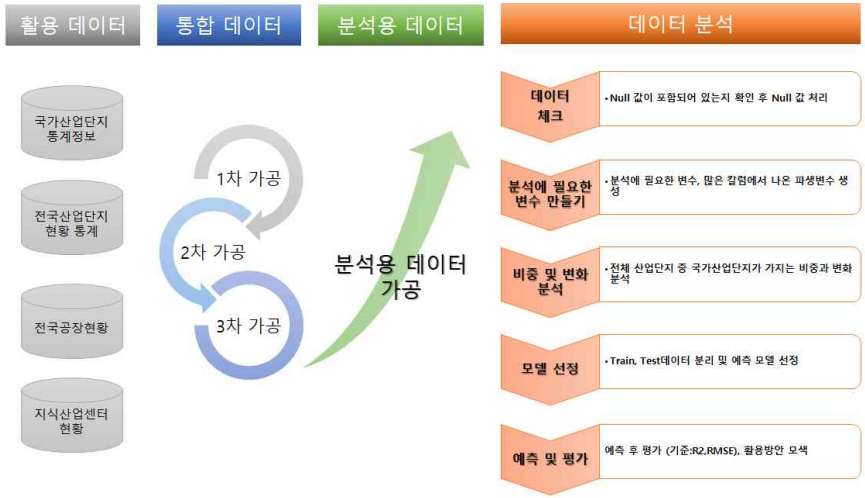
□ 광주첨단의 경우 입주는 주력산업의 입주는 늘어나고 있다. 하지만 생산과 수출역시 코로나로 인한 타격을 입었다.

아) 평가

□ 입주, 가동률, 생산, 수출, 고용을 살펴보았다. 코로나 같은 외부 문제에 바로 영향을 받는 것들은 입주, 고용, 생산, 수출이고 가동률 같은경우에는 조금 영향을 느리게 받는다. 이러한 특성들을 활용해서 다른 외부의 문제가 있을 때 선별적인 지원이 가능하고 판단된다.

□ 코로나 19로 인해서 많은 산업단지가 피해를 보고 있다. 생산과 수출에서 큰 타격을 입고 있는 산업단지가 많았으며 업종을 불문하고 피해를 입고 있다. 특히 해당 산업단지의 핵심산업들이 무너지고 있는데 그런 산업단지들의 피해를 최소화 시키는 방안이 필요하다고 판단된다.

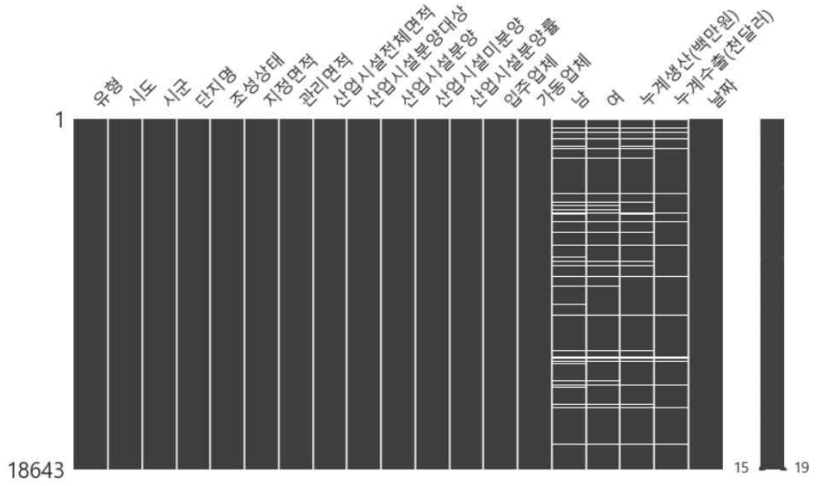
2) 주제 2번 - 국가산업단지가 전체 산단에서 갖는 영향력을 분석 및 예측



[그림 3-28] 데이터 분석 프로세스

가) 데이터 체크

```
데이터Shape : (18643, 19)
Null값 : 19
Null값 제거후 데이터Shape : (16598, 17)
```



[그림 3-29] 전국산업단지 데이터 체크

□ 데이터를 체크한다. 데이터의 형태, Null의 유무

□ Null이 있어서 drop시켰다. 분석하는데 크게 지장이 없어보인다.

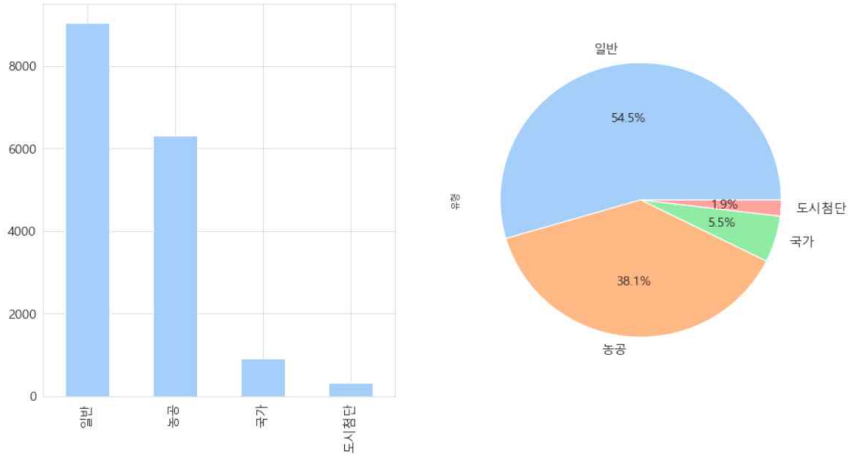
나) 분석에 필요한 변수 만들기

□ 하나로 된 날짜를 연도와 월로 나누었다.

□ 1면적당 생산, 수출 변수 생성

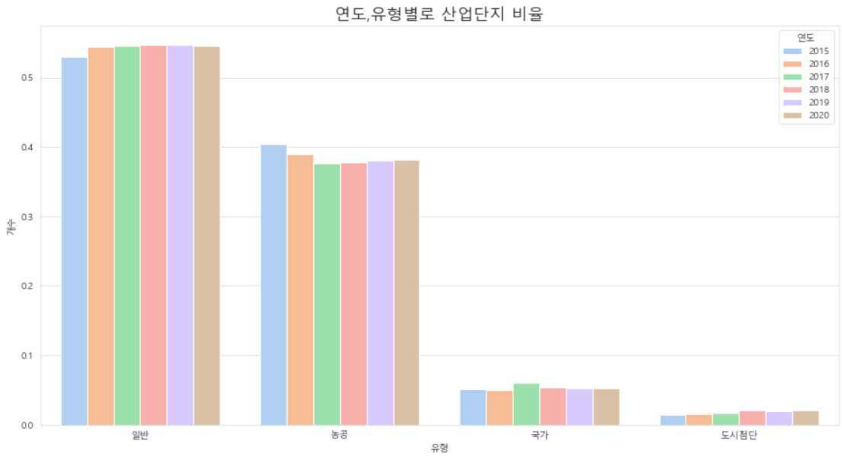
다) 비중 및 변화 분석

1) 유형



[그림 3-30] 유형별 전국산업단지 비중

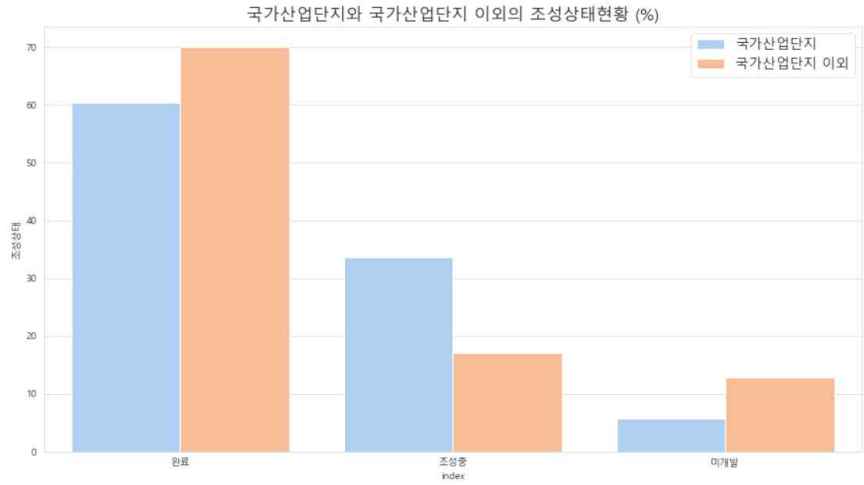
□ 대부분의 산업단지는 일반, 농공이다. 국가산업단지는 5.5% 정도에 불과하다.



[그림 3-31] 유형, 연도별 전국산업단지 비율

□ 국가산업단지는 2015년부터 2020년 중에 2017년이 가장 많은 비중을 차지했고 도시첨단의 경우 지속적으로 비중이 증가하고 있다.

2) 조성상태



[그림 3-32] 국가산업단지와 국가산업단지 이외의 조성상태현황

□ 조성상태는 국가산업단지 같은 경우 조성중인 곳이 30%넘는다. 국가산업단지 이외의 곳은 조성완료, 미개발인 곳들이 많다.

3) 산업시설전체면적

□ 국가산업단지 산업시설전체면적

기초통계량	산업시설전체면적
mean	4694.834802
std	6635.576295
min	0.000000
25%	965.000000
50%	1764.500000
75%	5163.000000
max	34747.000000

[표 3-5]국가산업단지 산업시설전체면적 기초통계량

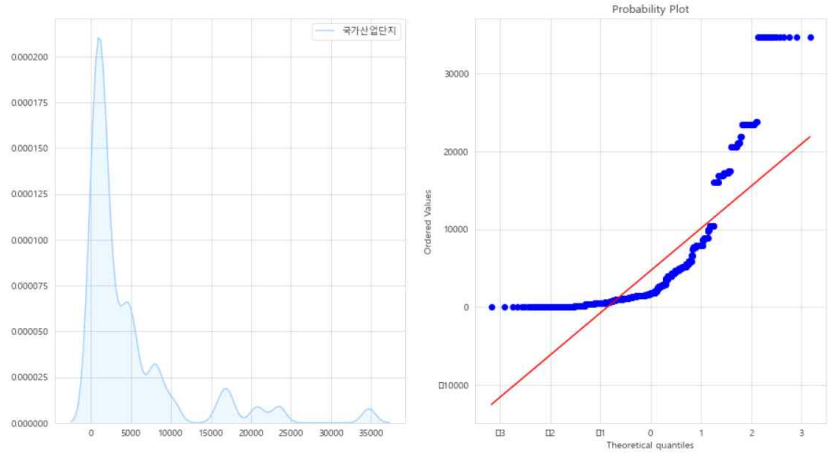
□ 국가산업단지 이외 산업시설전체면적

기초통계량	산업시설전체면적
mean	351.724996
std	603.129904
min	0.000000
25%	88.000000
50%	145.500000
75%	389.000000
max	9787.000000

[표 3-6]국가산업단지 이외 산업시설전체면적 기초통계량

□ 국가산업단지의 산업시설 면적이 다른 산업단지들보다 월등히 높다.

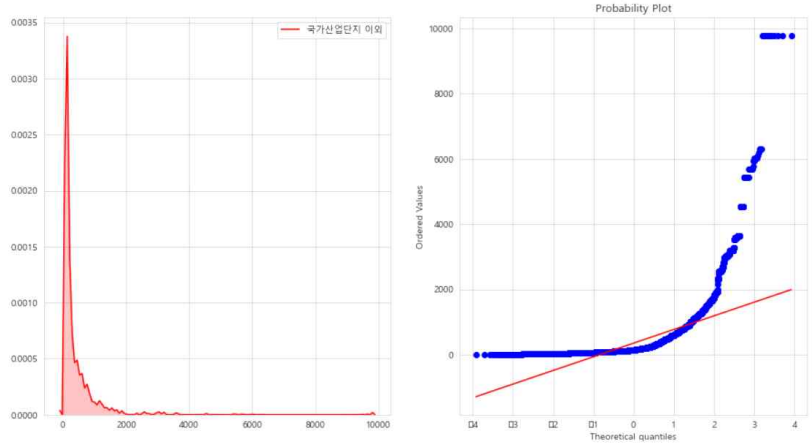
□ 수치형변수이므로 정규성이 파악되는지 확인이 필요할듯하다.



국가산업단지 Skewness: 02.49 Kurtosis: 06.58

[그림 3-33] 국가산업단지 산업시설전체면적 정규성 파악 분포

□ 국가산업단지의 산업시설전체면적에 대한 왜도와 첨도이다. 왜도가 2.49 첨도가 6.58이고 QQ차트로 표현해봐도 빨간색 직선에 데이터 형태가 맞춰지지 않아 정규성이 파악 되지않는다.

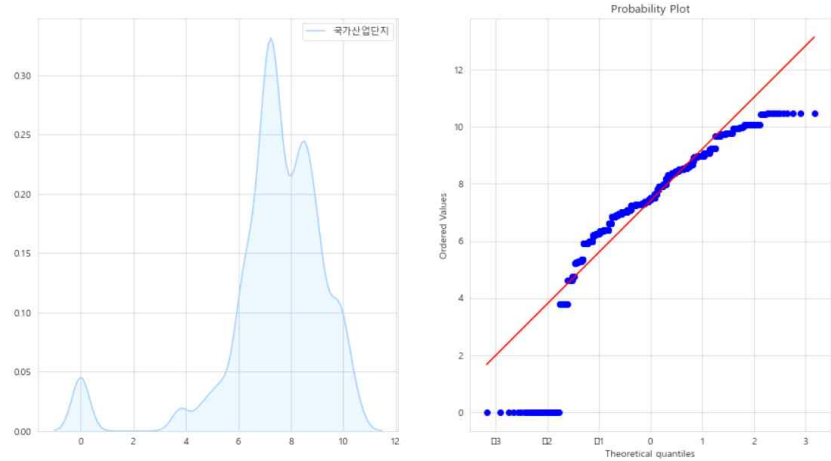


국가산업단지 이외 Skewness: 06.19 Kurtosis: 62.52

[그림 3-34] 국가산업단지 이외 산업시설전체면적 정규성 파악 분포

□ 국가산업단지의 이외 산업시설전체면적에 대한 왜도와 첨도이다. 왜도가 6.19 첨도가 62.52 이고 QQ차트로 표현해봐도 빨간색 직선에 데이터 형태가 맞춰지지 않아 정규성이 파악 되지않는다.

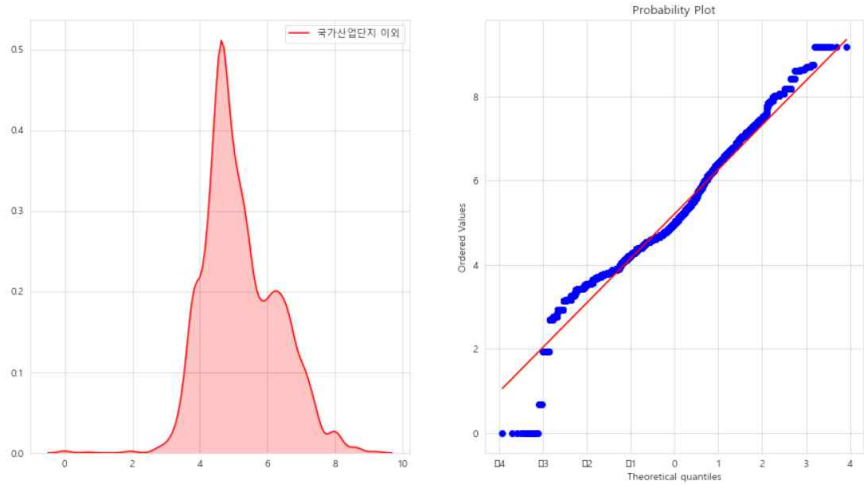
□ 두 개다 정규성이 파악 되지않아서 Log 변환을 통해 정규성을 확보하겠다.



국가산업단지 Skewness: -1.90 Kurtosis: 05.19

[그림 3-35] log변환한 국가산업단지 산업시설전체면적 정규성 파악 분포

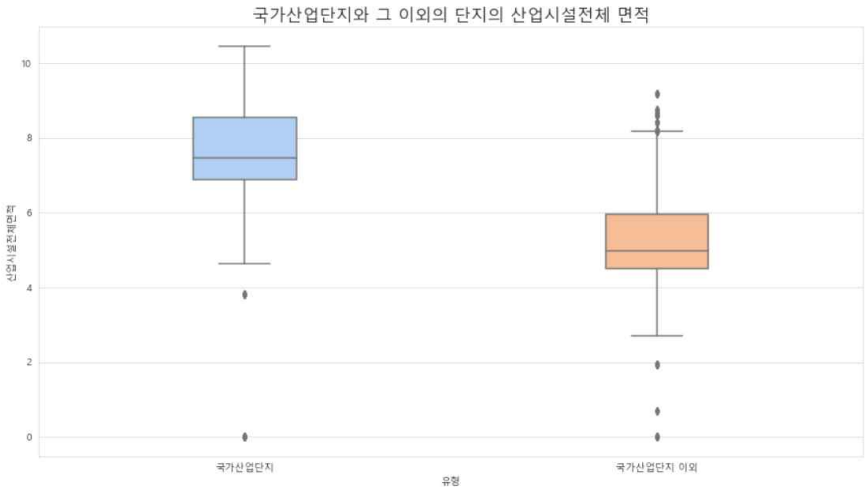
□ 국가산업단지의 산업시설전체면적이 조금 정규성이 맞춰지긴 했지만 여전히 정규분포로 보기에는 무리가 있다. 여러 가지 방법이 있지만 log변환을 많이 했을 경우 다시 되돌릴 때 많은 오차가 발생하므로 한번의 Log변환만 사용하겠다.



국가산업단지 이외 Skewness: 00.43 Kurtosis: 00.46

[그림 3-36] log변환한 국가산업단지 이외 산업시설전체면적 정규성 파악 분포

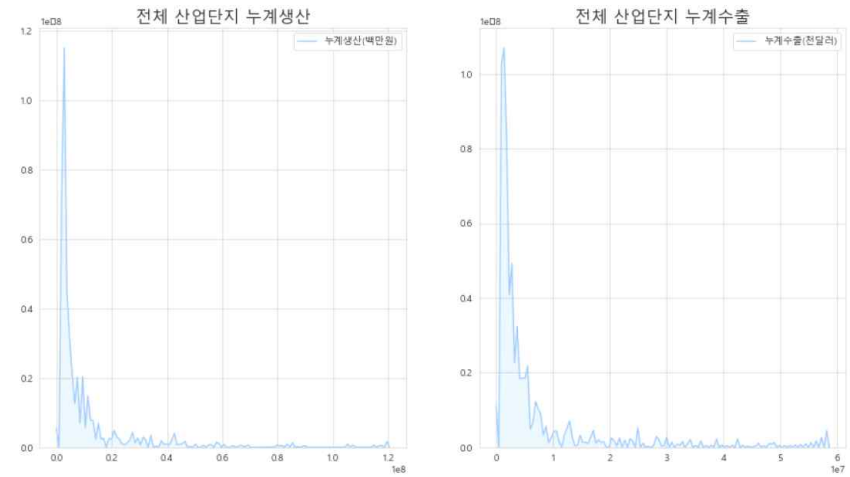
□ 국가산업단지 이외 산업시설전체면적이 정규성이 확보되었다. 여기도 한번의 Log변환으로 마무리 하겠다.



[그림 3-37] 국가산업단지, 국가산업단지 이외의 산업시설전체면적의 차이

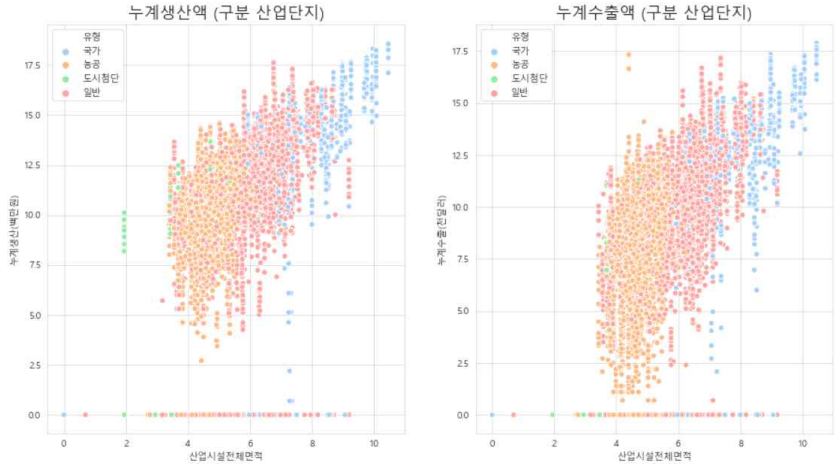
□ 국가산업단지와 그 이외의 산업단지의 면적이 많이 차이난다. 몇 개의 이상치가 보이지만 많지않고 크게 분석에 영향을 미치지 않을거 같다.

4) 누계생산, 누계수출



[그림 3-38] 전체 산업단지 누계생산, 수출 정규성 파악 분포

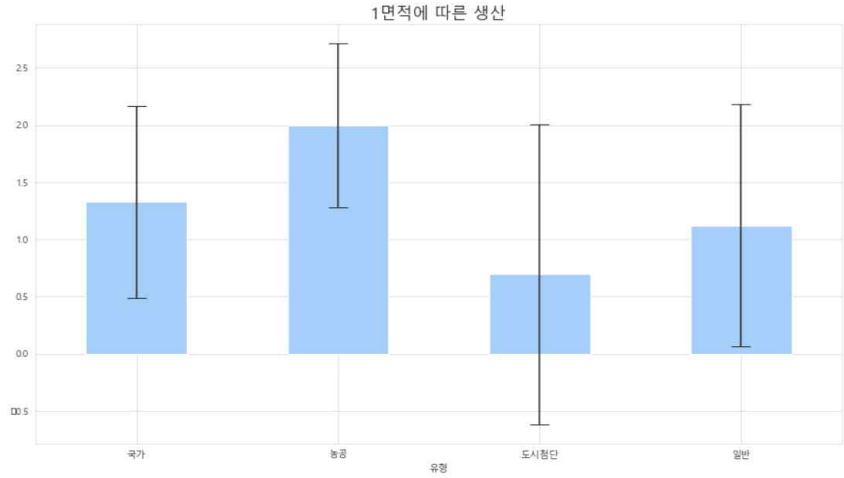
□ 누계생산, 누계수출 두 변수 다 수치형 변수이고 정규성이 파악이 되지않는다. 두 변수 모두 Log변환을 통해 정규성을 파악하겠다.



[그림 3-39] 누계 생산, 누계 수출과 산업시설면적과의 관계(산업단지 구분)

□ 국가산업단지의 경우 면적이 크고 생산액, 수출액 둘 다 다른 산업단지 유형에 비해 높다.

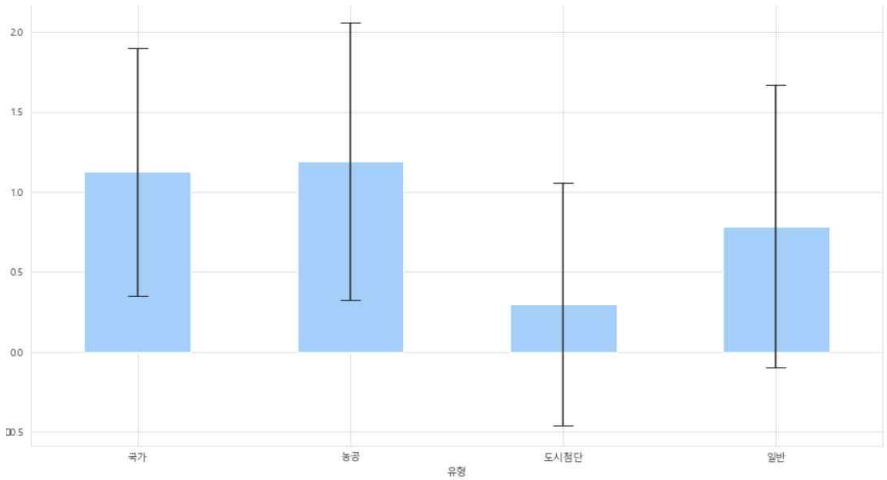
□ 자세히 살펴보면 면적이 유형을 나누는 기준이 될 수 있을거라고 생각된다.



[그림 3-40] 1면적에 따른 누계생산

□ 검은색 막대가 표준편차를 표현한 것이다. 표준편차가 꽤나 크다. 도시첨단 같은 경우는 어마어마하게 크다.

□ 국가산업단지는 면적이 커서 생산량이 많은 것이었다. 1면적당 생산은 농공이 가장 크다.



[그림 3-41] 1면적에 따른 누계수출

□ 수출도 표준편차가 엄청 높다.

□ 농공단지가 1면적당 누계수출이 많고 국가산업단지도 많은 편이다.

5) 산업시설분양률

□ 국가산업단지의 평균 산업시설분양률은 73.14%이고 국가산업단지 이외의 평균 산업시설 분양률은 74.77%이다.

[그림 3-42] 국가산업단지, 국가산업단지 이외 산업시설분양률

□ 0%부터 100%까지 다양하게 분포되어있다.

□ 국가산업단지 이외의 박스의 폭이 좁은걸로 봐서 국가산업단지 이외의 산업시설 분양률이 높다는 것을 확인할 수 있다.

□ 신업시설분양률을 조금 더 세밀하게 파악하기 위해 기초 통계량을 구해보았다.

기초통계량	산업시설분양률
mean	73.145374
std	42.158346
min	0.000000
25%	27.500000
50%	100.000000
75%	100.000000
max	100.000000

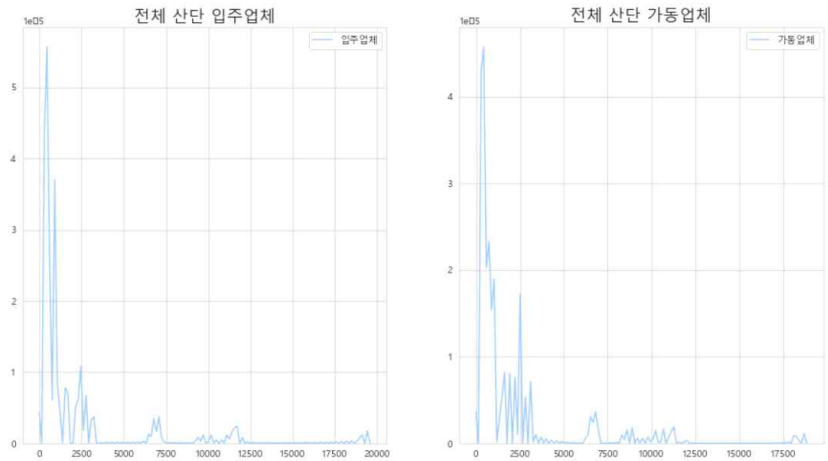
[표 3-7]국가산업단지 산업시설분양률 기초통계량

기초통계량	산업시설분양률
mean	74.772658
std	40.696458
min	0.000000
25%	56.000000
50%	100.000000
75%	100.000000
max	100.000000

[표 3-8]국가산업단지 이외 산업시설분양률 기초통계량

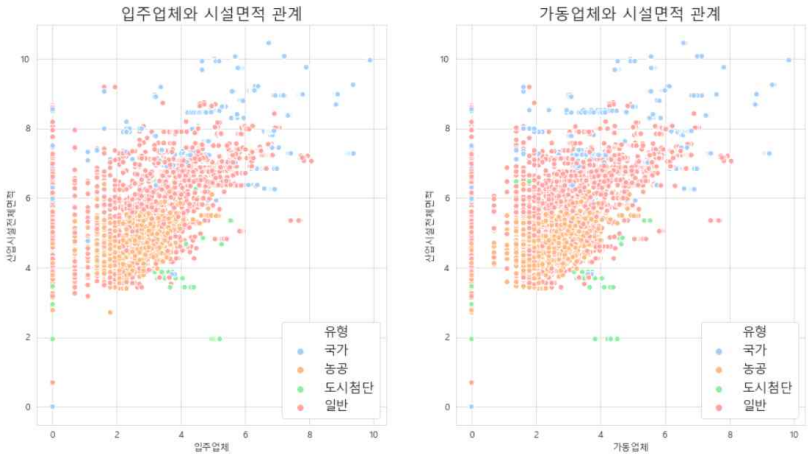
□ 국가산업단지와 국가산업단지 이외의 산업시설분양률 모두 상위 50%가 넘는 순간 100%의 분양률을 기록하고 있다.

6) 입주업체, 가동업체



[그림 3-43] 전체 산업단지 입주업체, 가동업체의 정규성 파악

□ 입주업체, 가동업체 두 변수 다 수치형 변수인데 정규성이 파악 되지않는다. Log변환을 사용해서 정규성을 확보해주겠다.



[그림 3-44] 전체 산업단지 입주업체, 가동업체의 시설면적 간의 관계(산업단지 구분)

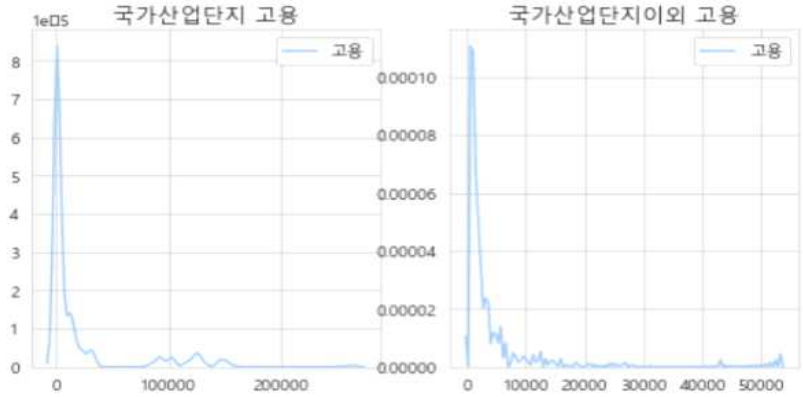
□ 입주업체와 가동업체가 크게 다르지 않아 보인다.

□ 전제적으로 면적이 넓으면 입주업체나 가동업체가 많은 것이 확인되었다.

□ 국가산업단지의 경우 면적이 넓으니 입주와 가동업체가 많았다.

7) 고용

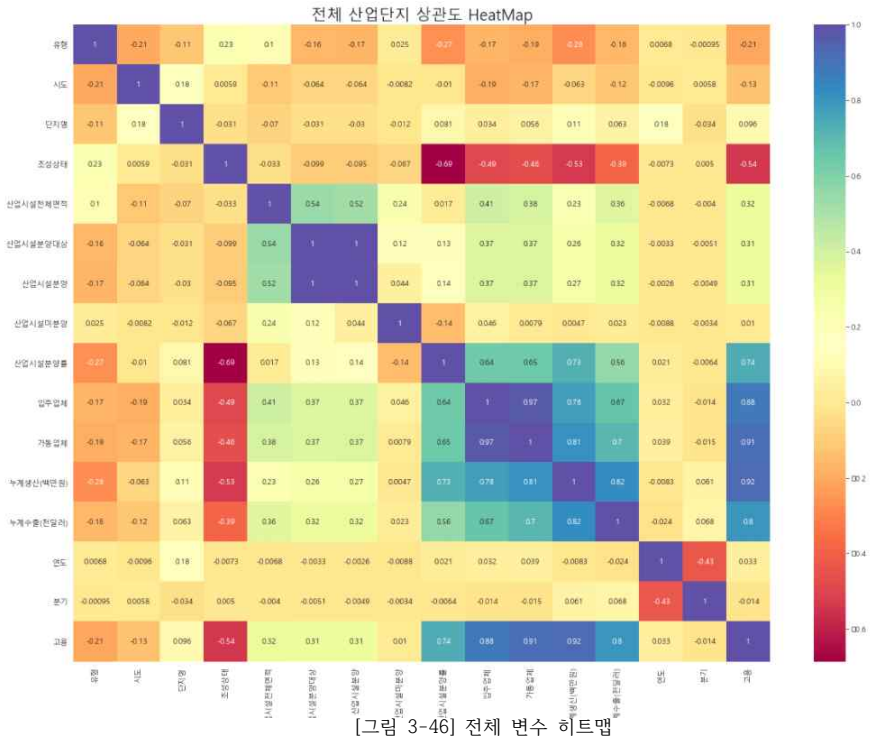
□ 고용부분에 있어서 남성과 여성을 구분 할 필요가 없다고 판단되어 변수 남, 변수 여를 합쳐서 고용이라는 변수를 생성했다.



[그림 3-45] 국가산업단지, 국가산업단지이외 고용

□ 변수 고용은 정규성이 파악이 되지않아 Log변환으로 정규성을 확보해주었다.

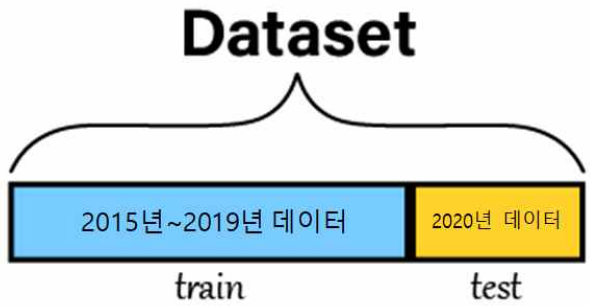
8) 전체 히트맵



[그림 3-46] 전체 변수 히트맵

라) 모델 선정

□ 학습은 2015년부터 2019년의 데이터로 학습하고 2020년 1분기를 예측하겠다. 예측하는 값(target)은 2020년 1분기 누적 생산액이다.

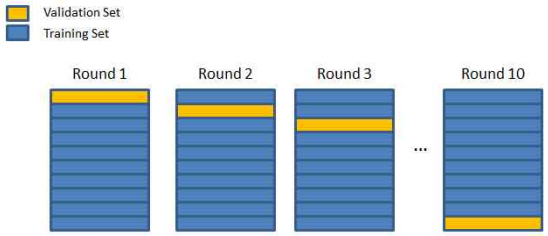


[그림 3-47] Train, Test 분할

□ var로 예측을 시도했지만 예측이 잘되지 않아서 regression(회귀)로 예측 알고리즘을 변경했다.

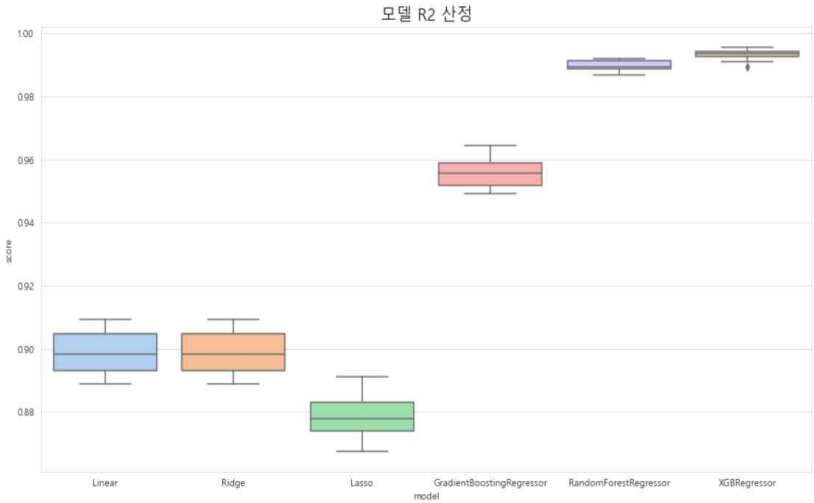
□ 회귀 알고리즘으로는 Linear Regression, Ridge Regression, Lasso Regression, GradientBoostingRegressor, RandomForestRegressor, XGBRegressor를 쓰도록 하겠다.

□ K-Fold 진행을 위한 함수 설계 (K=10)



[그림 3-48] K-Fold 출처 : ResearchGate

□ 모델별로 K-Fold 진행하면서 을 산출



[그림 3-49] 모델의 R^2

□ Linear, Ridge는 거의 차이가 없고 Lasso가 제일 낮다. 앙상블 모형들의 R^2가 높게 확인되었고 그 중 XGBoost가 가장 높다.

□ R^2만으로 모델을 판단 할 수 없으니 RMSE를 측정해 판단해보겠다.

모델	RMSE
Linear Regression	1.6833673211240199
Ridge Regression	1.6833618479782888
Lasso Regression	1.850720099288563
Gradient Boosting Regression	1.2099396917223875
Random Forest Regression	0.7225119797893482
XGBoost Regression	0.6571448412256842

[표 3-9] 모델 평가(RMSE)

□ RMSE가 낮으면 낮을수록 실제 데이터와 오차가 없다는 뜻이다.

□ XGBoost Regression이 RMSE기준으로 수치가 가장 낮다.

□ RMSE와 R^2이 가장 좋은 XGBoost Regression으로 학습을 진행하겠다.

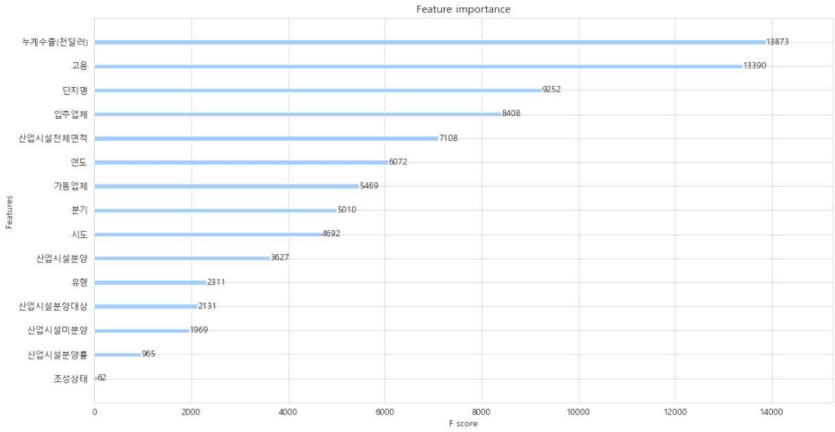
마) 예측 및 평가

[그림 3-50] XGBoost Regression 모델의 예측

□ 파란색 선은 실제 값이고 주황색선은 예측 값이다. 정확하게 예측했다면 파란색부분이 보이지 않아야 한다.

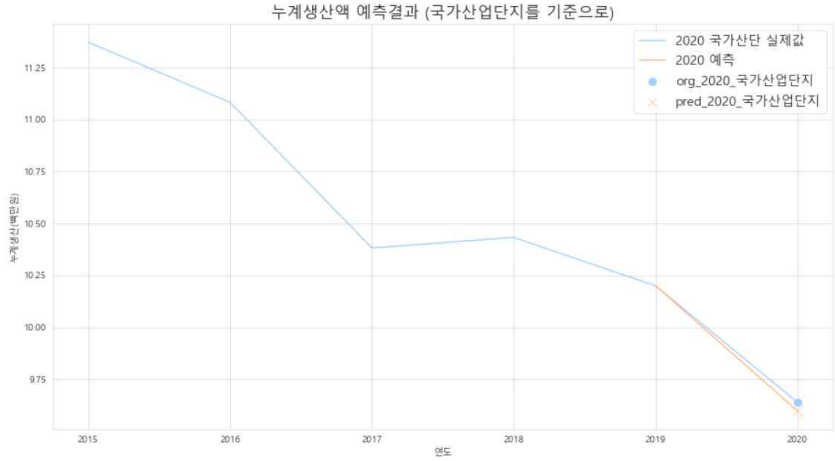
□ 조금의 파란색 선들이 보이긴 하지만 예측이 잘된 것으로 판단이 된다.

□ 어떤 변수가 누적 생산에 많은 영향을 미치나 판단해 볼 필요가 있다.



[그림 3-51] XGBoost Regression Featrue Importance

□ 누계 수출이 가장 많은 영향을 미치고 있고 고용, 단지명(어떤 산업단지인지), 입주업체 등 영향을 미치는 요소가 많이 있는 것으로 파악이 된다.

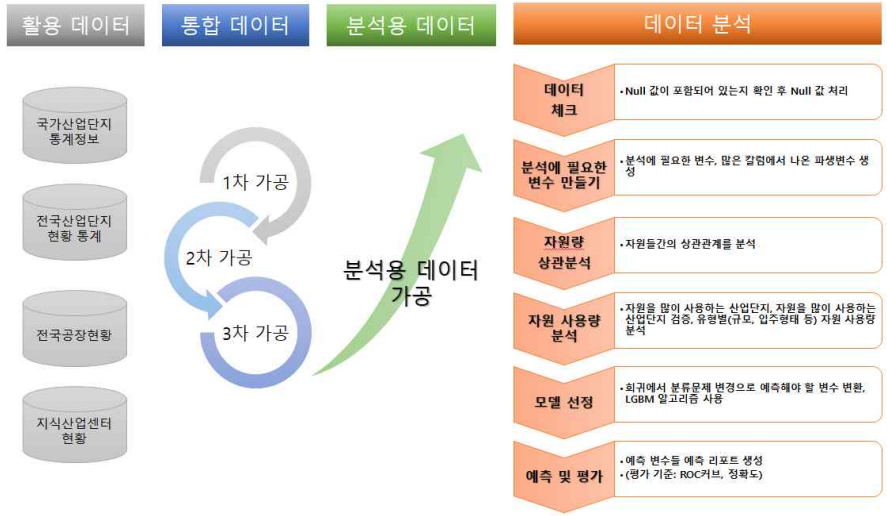


[그림 3-52] 누계 생산액 예측 결과(국가산업단지기준)

□ 파란색 선은 실제 값이고 주황색 선은 예측 값으로 파란색 선과 주황색 선이 많이 차이나지 않는다. 2020년 1분기의 예측이 잘 이루어진 것을 확인할 수 있다.

□ 생산액을 예측해서 향후에 더 지원해야 할 산업단지를 골라서 차등적으로 지원 방안을 마련 할 수 있다.

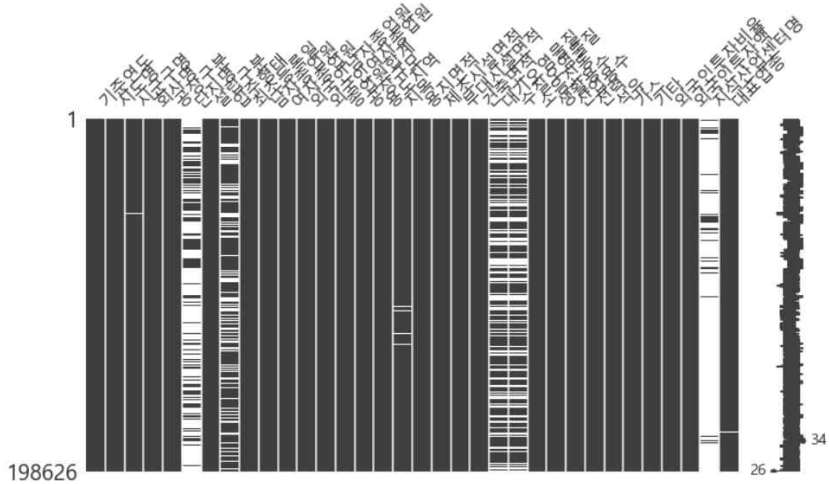
3) 주제 3번 - 국가산업단지의 자원 사용량을 분석하여 신규 입주 기업의 자원 사용량을 예측



[그림 3-53] 데이터 분석 프로세스

가) 데이터 체크

```
데이터Shape : (198626, 34)
Null값 : 34
```



[그림 3-54] 전국공장현황 데이터 체크

□ Null값이 너무 많다. 데이터 설명을 보니 Null값은 해당 칼럼에 해당되지 않는 공장들이라고 한다.

□ Drop하기에는 Null값이 너무 많아서 데이터가 남아있지않을 듯해서 0으로 채우기로 결정했다.

나) 분석에 필요한 변수 정리

□ 전국 공장현황이므로 주제에 사용될 국가산업단지만 추출했다.

□ 우리가 분석하게 될 변수(생활용수, 산업용수, 전력, 석유, 가스, 기타)를 따로 뽑아내서 분석에 용이하게 변형했다.

다) 자원량 상관 분석

	생활용수	산업용수	전력	석유	가스	기타
생활용수	1.000000	-0.000105	0.011805	-0.000034	-0.000088	-0.000066
산업용수	-0.000105	1.000000	0.088608	0.010965	0.027037	0.031020
전력	0.011805	0.088608	1.000000	0.886623	0.025329	0.011061
석유	-0.000034	0.010965	0.886623	1.000000	0.000699	0.000093
가스	-0.000088	0.027037	0.025329	0.000699	1.000000	0.005862
기타	-0.000066	0.031020	0.011061	0.000093	0.005862	1.000000

[그림 3-55] 분석 할 변수간의 상관관계

□ 전력과 석유를 제외하면 분석 할 변수간의 상관관계가 없다.

□ 전력과 석유는 서로 양의 상관관계가 있다.

	생활용수	산업용수	전력	석유	가스	기타
count	3.055900e+04	30559.000000	3.055900e+04	3.055900e+04	3.055900e+04	30559.000000
mean	1.417722e+04	57.474636	7.472420e+03	3.479938e+03	2.534015e+02	14.341228
std	2.476956e+06	2460.730649	3.721565e+05	5.683521e+05	1.540861e+04	1234.073260
min	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
25%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
50%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
75%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
max	4.330000e+08	400000.000000	5.762310e+07	9.933200e+07	2.515755e+06	200000.000000

[그림 3-56] 분석 할 변수의 기초 통계량

□ 우리가 분석해야할 변수들이 전부 75% 이상 0 값이다. 분석할 때 0 값은 제외하고 분석을 진행하겠다.

□ 다들 수치가 다르고 단위가 달라서 어느정도 맞춰줘야 할 필요가 있어보인다.

분석할 변수	비율(%)
생활용수	17.16
산업용수	6.3
전력	24.08
석유	1.15
가스	3.09
기타	0.51

[표 3-10] 분석할 변수 값이 0초과인 값들의 비율

□ 0보다 큰 값이 전력은 상대적으로 많고 기타의 비율을 0.51%로 가장 낮다. 값이 0인 데이터가 많다는 이야기이고 해당변수에 해당하지않는다는 뜻이다.

라) 자원사용량 분석

단지명	생활용수	산업용수	전력	석유	가스	기타
여수국가산업단지	1순위 (4706556)	2순위 (2326.82)	3순위 (93166.09)	2순위 (56422.04)		
포항국가산업단지	2순위 (896.749)	3순위 (2212.17)	2순위 (355414.9)		1순위 (159135)	2순위 (4166)
안정국가산업단지	3순위 (587.5)		1순위 (2561644)			
울산미포국가산업단지		1순위 (4324.486)		3순위 (29158.28)	3순위 (12039.8)	
창원국가산업단지				1순위 (3540147)		
아산국가산업단지					2순위 (20681.1)	
광양국가산업단지						1순위 (44806.4)
시화국가산업단지						3순위 (4000)

[표 3-11] 자원을 많이 사용하는 단지 3개

□ 비어있는 칸은 순위 밖이다.

□ 표 3-11에 겹치는 산업단지들이 많이 있다. 여수, 포항, 울산미포는 1순위부터 3순위까지 3번이상 겹친 산업단지들이다. 규모가 큰 산업단지일 것으로 추측이 된다. 자원을 많이 사용하는 산업단지들을 묶어서 보겠다.

□ 입주형태에 따른 자원 사용량을 확인해 보겠다.

많이 사용하는 자원	입주형태	사용량
생활용수	양수도	238591.54
산업용수	기타	2028.96
전력	기타	158207.54
석유	기타	1535997
가스	분양	12391.83
기타	공장임대	14460.52

[표 3-12] 자원을 많이 사용하는 입주형태와 사용량

□ 표 3-12를 보면 가장 많이 사용하는 자원별로 입주형태와 사용량이 나와

있다.

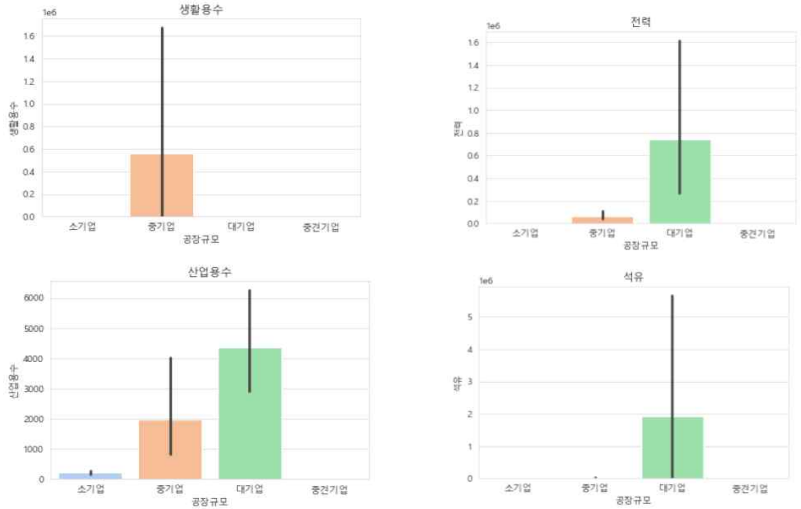
□ 기타가 산업용수, 전력, 석유를 가장 많이 사용하는 것으로 나타나고 있고 산업단지 별로 많이 사용한 자원과 입주형태는 연관이 있을 것이다. 표 3-11과 연관이 있을 것으로 판단된다.

산업단지	가장 많은 입주형태	자원량 많이 사용하는 산업단지 여부
여수국가산업단지	분양, 양수도	생활용수, O
울산미포국가산업단지	기타, 양수도	산업용수, O
안정국가산업단지	분양, 공장임대	전력, X
창원국가산업단지	공장임대, 양수도	석유, X
포항국가산업단지	분양, 양수도	가스, O
광양국가산업단지	공장임대, 분양	기타, O

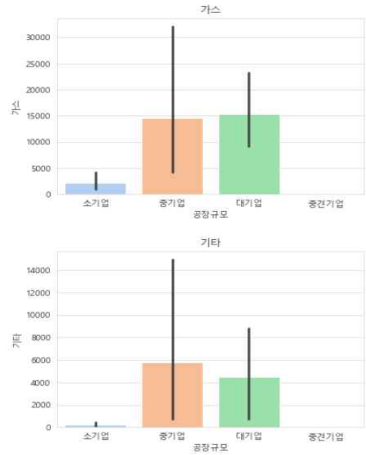
[표 3-13] 자원을 많이 사용하는 산업단지와 입주형태 관계 여부

□ 6개의 국가산업단지를 판단해보았다. 6개중에 4개가 맞게 판단되었다. 안정, 창원 국가산업단지는 한번 확인이 필요하겠지만 전반적으로 자원을 많이 사용하는 산업단지와 자원을 많이 사용하는 입주형태는 관계가 있는 것으로 판단된다.

□ 안정국가산업단지는 입주기업은 적은데 대기업이 있어서 전력사용량이 많은 것으로 판단되었고 창원국가산업단지는 몇 개의 기업이 석유를 많이 사용해서 석유사용량이 많은 것으로 판단되었다.



[그림 3-57] 공장규모별 자원 사용1



[그림 3-58] 공장규모별 자원 사용2

□ 보통 대기업과 중기업이 많은 자원을 쓰는 것으로 확인이 되었다. 기타같은 경우 중기업이 대기업보다 자원을 더 많이 쓰는 것으로 나타났다.

용도지역	생활용수	산업용수	전력	석유	가스	기타
0	0.363636	8.181818	629.545455	0.000000	90.909091	0.045455
관리지역/계획관리지역	22.048980	3.836735	1375.979592	10.204082	549.346939	0.000000
관리지역/관리지역기타	10.417857	8.162857	810.857143	8.214286	1024.392857	0.000000
관리지역/생산관리지역	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
농림지역/준보전산지	6.000000	16.500000	990.000000	0.000000	200.000000	0.000000
도시지역/공업지역	47.372446	123.204392	21851.648332	204.658103	302.695830	17.285476
도시지역/공업지역/일반공업지역	17838.210665	62.144280	7905.547436	4360.407259	157.410617	16.939422
도시지역/공업지역/전용공업지역	6.122807	59.556442	5573.402186	20.036740	4377.068693	2.508085
도시지역/공업지역/준공업지역	0.558623	6.095460	123.661144	39.131651	10.974348	0.017100
도시지역/녹지지역/보전녹지지역	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
도시지역/녹지지역/자연녹지지역	0.120891	1.236634	184.422772	0.000000	5.901386	0.346535
도시지역/상업지역/유통상업지역	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
도시지역/상업지역/일반상업지역	1.333333	0.000000	183.666667	0.000000	5.000000	0.000000
도시지역/주거지역/제2종일반주거지역	6.933333	14.000000	6524.666667	22.333333	256.000000	0.000000
도시지역/주거지역/준주거지역	1.147727	0.136364	246.795455	0.000000	25.004545	0.000000

[그림 3-59] 용도지역별 평균 자원 사용량

□ 그림 3-55를 보면 생활용수는 일반공업지역, 산업용수는 도시/공업지역, 전력은 도시/공업지역, 석유는 일반공업지역, 가스는 전용공업지역, 기타는 도시/공업지역으로 전부 공업지역에서 자원을 많이 사용한다는 것을 알 수 있다.

지목	생활용수	산업용수	전력	석유	가스	기타
0	0.388571	4.657143	285.885714	0.000000	4.285714	0.000000
공장용지	14464.965988	58.079505	7578.285651	3525.183756	257.862273	14.607759
답	7.357143	135.714286	408.571429	14.285714	142.857143	0.000000
대	2.989041	33.264733	2821.794030	1620.597015	38.676205	1.570576
목장용지	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
묘지	0.500000	0.000000	5.000000	0.000000	0.000000	0.000000
양어장	2.790909	2.659091	113.818182	21.590909	30.545455	0.000000
임야	4.183333	9.166667	5301.666667	0.000000	83.333333	0.000000
잡종지	10.000000	0.000000	22.000000	0.000000	0.000000	0.000000
전	1.500000	0.000000	147.750000	0.250000	0.500000	0.000000
제방	1.500000	0.250000	200.000000	0.000000	0.500000	0.000000
하천	2.000000	0.000000	60.000000	0.000000	0.000000	0.000000
학교용지	0.006250	0.000000	11.208333	0.000000	0.000000	0.000000

[그림 3-60] 지목별 평균 자원 사용량

□ 대부분 공장용지에서 자원을 많이 사용하는 것으로 나오지만 흥미로운 점은 임야에서 전력이 많이 사용된다.

마) 모델선정

□ 공장별로 사용되는 자원량을 예측해보겠다. 예측에 필요한 변수만 남기고 다 지우도록 한다.

□ 회귀분석으로 사용자원량을 예측해보려고 시도 해보았지만 0값이 너무 많아서 제대로 된 예측이 되지않아 회귀문제를 분류문제로 바꿔서 예측을 하겠다.

□ 분류문제를 해결하는 많은 알고리즘이 있지만 최근 제일 강력하고, 카테고리 변수를 숫자로 변환하지 않아도 되는 LightGBM 알고리즘을 사용하도록 하겠다.

자원 사용량	전력	생활용수	산업용수	석유	가스	기타
4 (아주 많이 사용)	700 이상	5 이상	100 이상	1000 이상	1500 이상	300 이상
3 (많이 사용)	700	5	100	1000	1500	300
2 (보통)	150	1	10	84.5	200	35
1 (적게 사용)	30	0.4	1.5	10	10	2
0 (사용안함)	0	0	0	0	0	0

[표 3-14] 분류 알고리즘을 위한 자원 사용량 구간화

□ 자원 사용량은 5단계로 구분하겠다. 0이 많은 관계로 0을 따로 부여하고 0을 제외한 나머지는 4분위수를 따라서 자원 사용량 1-> 하위 25%, 2-> 하위 50%, 3-> 하위 75%, 4->하위 75% 초과로 결정했다.

□ class imbalance가 나타날 수 있기 때문에 4분위수로 구간화하는 것이 낫다고 판단했다.

□ 카테고리 변수들(시도명, 단지명, 용도지역, 공장규모, 입주형태, 지목, 대기오염물질, 수질오염물질, 대표업종)은 LightGBM의 카테고리 변수항목에 넣어주겠다.

바) 예측 및 평가

□ 생활용수, 산업용수, 전력, 석유, 가스, 기타 순으로 예측 하겠다. 각 변수마다 0부터 4의 class가 있으므로 Multiclass Classification으로 진행하겠다.

예측 리포트	생활용수	산업용수	전력	석유	가스	기타
Train 정확도(%)	94.53	99.54	92.26	100	100	100
Test 정확도(%)	84.67	93.82	80.60	98.81	96.63	99.43

[표 3-15] 오버피팅 체크를 위한 Train, Test 정확도 체크

□ Train정확도가 Test정확도를 보다 높다. 오버피팅은 일어나지 않은 것으로 확인된다. Test데이터의 정확도가 해당 변수의 정확도다.

□ 각 변수 별로 정확도가 다르지만 평균적으로 92% 정도의 높은 정확도로 나타난다.

	precision	recall	f1-score	support
0	0.89	0.97	0.93	5062
1	0.58	0.23	0.33	291
2	0.40	0.24	0.30	308
3	0.26	0.17	0.20	205
4	0.52	0.93	0.40	246
accuracy			0.85	6112
macro avg	0.53	0.39	0.43	6112
weighted avg	0.82	0.85	0.82	6112

	precision	recall	f1-score	support
0	0.95	0.99	0.97	5717
1	0.24	0.04	0.07	103
2	0.28	0.05	0.08	107
3	0.34	0.12	0.18	85
4	0.66	0.45	0.54	100
accuracy			0.94	6112
macro avg	0.49	0.33	0.37	6112
weighted avg	0.91	0.94	0.92	6112

[그림 3-61] 생활용수(좌), 산업용수(우)의 예측 리포트

	precision	recall	f1-score	support
0	0.88	0.96	0.92	4646
1	0.49	0.24	0.32	379
2	0.33	0.20	0.25	363
3	0.36	0.32	0.34	353
4	0.59	0.53	0.56	371
accuracy			0.81	6112
macro avg	0.53	0.45	0.48	6112
weighted avg	0.78	0.81	0.79	6112

	precision	recall	f1-score	support
0	0.99	1.00	0.99	6042
1	0.00	0.00	0.00	19
2	0.00	0.00	0.00	18
3	0.00	0.00	0.00	19
4	0.25	0.14	0.18	14
accuracy			0.99	6112
macro avg	0.25	0.23	0.24	6112
weighted avg	0.98	0.99	0.98	6112

[그림 3-62] 전력(좌), 석유(우)의 예측 리포트

	precision	recall	f1-score	support
0	0.97	1.00	0.98	5917
1	0.00	0.00	0.00	50
2	0.14	0.02	0.04	44
3	0.13	0.05	0.07	42
4	0.41	0.15	0.22	59
accuracy			0.97	6112
macro avg	0.33	0.24	0.26	6112
weighted avg	0.95	0.97	0.96	6112

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6080
1	0.00	0.00	0.00	8
2	0.00	0.00	0.00	7
3	0.50	0.25	0.33	4
4	0.00	0.00	0.00	13
accuracy			0.99	6112
macro avg	0.30	0.25	0.27	6112
weighted avg	0.99	0.99	0.99	6112

[그림 3-59] 가스(좌), 기타(우)의 예측 리포트

□ 높은 예측률에 비해 class 별로 정확도는 class가 0인 것을 제외하면 그렇게 높지않다. class가 0인 것만 잘 예측한다.

예측 리포트	생활용수	산업용수	전력	석유	가스	기타
ROC커브 Score	0.90	0.91	0.90	0.91	0.91	0.86

[표 3-16] 각 예측 변수별 ROC커브 Score

□ ROC커브 Score은 준수한 상황이다.

순위	생활용수	산업용수	전력	석유	가스	기타
1 순위	최초등록일	최초등록일	최초등록일	종업원합계	최초등록일	종업원합계
2 순위	용지면적	종업원합계	대표업종	최초등록일	종업원합계	최초등록일
3 순위	종업원합계	용지면적	종업원합계	용지면적	용지면적	용지면적
4 순위	대표업종	대표업종	용지면적	대표업종	대표업종	공장규모
5 순위	단지명	단지명	단지명	공장규모	단지명	대표업종

[표 3-17] 중요 변수 추출

□ 변수 중요도를 체크해보니 최초등록일, 용지면적, 대표업종, 종업원합계 정도가 예측하는데 중요한 것들이다.

□ 예측을 회귀로 해서 정확한 값(수치)를 예측했으면 좋았겠지만 0값 다수 포함, 데이터의 불균형으로 회귀의 정확성이 떨어져 분류문제로 변경하여 예측을 진행했다. 정확도는 높았지만 Class가 0인 값만 정확하게 예측하고 다른 Class는 예측 정확도가 낮은 한계가 있다.

□ 새로운 공장들이 산업단지에 입주할 때 자원 사용량을 구분지어 예측해 향후 자원문제를 대비 할 수 있을 것이다.

4. 참고자료

가. 공공데이터 포털

[그림 4-1] 공공데이터 포털

나. 한국산업단지공단

1) 입지지원팀 과장 김정남
2) 입지지원팀 대리 윤현호

다. Rucrazia's Blog

1) [통계] Normalization(정규화) / Standardization(표준화)

라. 공공빅데이터 표준분석모델 매뉴얼[도로관리(포트홀·포장관리·안전시설물)]

[그림 4-2] 공공 빅데이터 표준분석모델 매뉴얼 - 도로관리

마. 두산백과

1) 시계열분석 [time series analysis, 時系列分析]

5. 별첨

가. 분석 방법 및 활용 기술

1) 정규화(Normalization)

□ 데이터를 특정 구간으로 바꾸는 척도법이다 (ex. 0~1 or 0~100).

□ 식 : (측정값 - 최소값) / (최대값 - 최소값)

□ 데이터 군 내에서 특정 데이터가 가지는 위치를 볼 때 사용된다.

□ 주가와 같은 주기를 띄는 데이터의 경우 과거에 비해서 현재 데이터의 위치가 어느 정도인지 파악하기에 좋아진다.

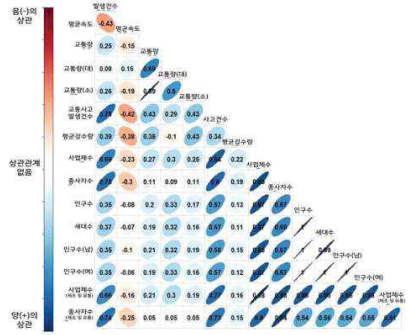
2) 표준화(Standardization)

□ - 데이터를 0을 중심으로 양쪽으로 데이터를 분포시키는 방법이다. 표준화를 하게 되면 각 데이터들은 평균을 기준으로 얼마나 떨어져 있는지를 나타내는 값으로 변환된다.

□ - 식 (Z-score 표준화) : (측정값 - 평균) / 표준편차

□ - 변환된 데이터는 다소 평평하게 만드는 특성을 가진다.(진폭의 감소) 진폭의 감소로 각 데이터의 간격이 감소하게 된다. (ex. 10000의 단위에서 0.1 단위로 감소)

3) 상관관계 분석



[그림 5-1] 상관관계 분석 결과 예시

□ 두 변수 간에 어떤 선형적 관계를 갖고 있는지 분석하는 방법으로 정도를 파악하기 위해 상관계수를 이용한다. 이때 상관계수가 (0,1]이면 양의 상관,

[-1,0)이면 음의 상관, 0이면 무상관이라고 한다. 하지만 0인 경우 상관이 없다는 것이 아니라 선형의 상관관계가 아니라는 것이다. 또한, 상관관계는 인과관계를 설명하는 것은 아니다.

4) 시계열 분석

□ 경기변동 등의 연구에 사용되고 있다. 통계숫자를 시간의 흐름에 따라 일정한 간격마다 기록한 통계계열을 시계열 데이터라고 하며, 이 계열의 시간적 변화에는 여러 원인에 기인한 변동이 포함되어 있다.

□ 예를 들면, 돌연적인 사건을 원인으로 하는 것(우연변동 또는 불규칙변동), 해마다 똑같이 되풀이되는 계절변동, 또한 오랜 세월에 걸쳐 추세적으로 나타나는 구조변동, 1년 이상의 장기간에 걸쳐 규칙적으로 반복되는 순환변동 등이 있는데, 이들 변동이 복잡하게 혼합되어 하나의 시계열 데이터를 이루고 있다.

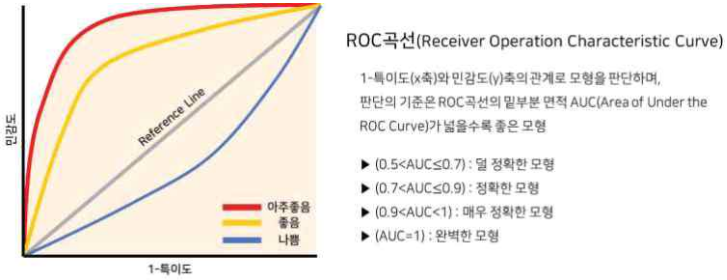
□ 연구목적에 따라 특정한 원인에 의거하여 나타나는 변동부분만을 분리하여 추출하거나 또는 소거하는 일이 필요하게 된다. 이와 같은 통계기술을 사용하는 연구를 시계열분석이라고 한다.

5) 예측모델 평가

		예측결과	
		TRUE	FALSE
실제	Positive	A (True positive)	B (False Negative)
	Negative	C (False Positive)	D (True Negative)

- 정분류율(Accuracy) = $\frac{(A+D)}{(A+B+C+D)} X\ 100$
  전체 중 예측모델이 예측(분류)한 값과 실제 값 동일한 비율
- 민감도 (Sensitivity) = $\frac{(A)}{(A+B)} X\ 100$
  실제 값 Positive 중 예측모델이 얼마나 많은 True를 예측했는지에 대한 비율
- 특이도(Specificity) = $\frac{(D)}{(C+D)} X\ 100$
  실제 값 Negative중 예측모델이 얼마나 많은 False를 예측했는지에 대한 비율

[그림 5-3] 정오분류표와 예측모델 평가 척도

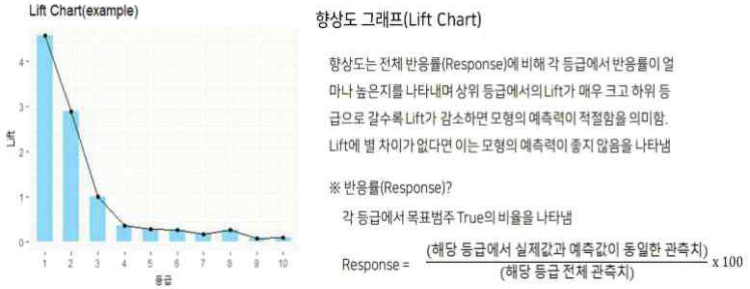


ROC곡선(Receiver Operation Characteristic Curve)

1-특이도(x축)와 민감도(y축)의 관계로 모형을 판단하며, 판단의 기준은 ROC곡선의 밑부분 면적 AUC(Area of Under the ROC Curve)가 넓을수록 좋은 모형

- (0.5<AUC≤0.7) : 덜 정확한 모형
- (0.7<AUC≤0.9) : 정확한 모형
- (0.9<AUC<1) : 매우 정확한 모형
- (AUC=1) : 완벽한 모형

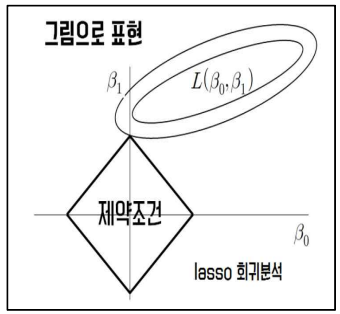
[그림 5-4] ROC곡선



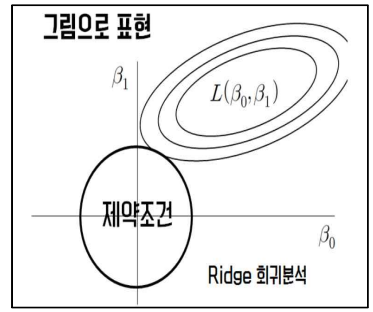
[그림 5-5] 향상도 그래프

□ 예측을 위해 구축된 모형이 '임의의 모형'보다 과연 우수한지 고려된 서로 다른 모형들 중 어느 것이 가장 우수한 예측력을 보유하고 있는지 등을 비교하고 분석하는 과정으로 범주형 예측모형의 평가를 위해 주로 사용되는 척도는 정오분류표를 통해 산출된 정분류율과 정확도이며, ROC곡선 및 리포트 차트를 통해 예측모형의 성능을 확인할 수도 있다.

6) Lasso Regression과 Ridge Regression



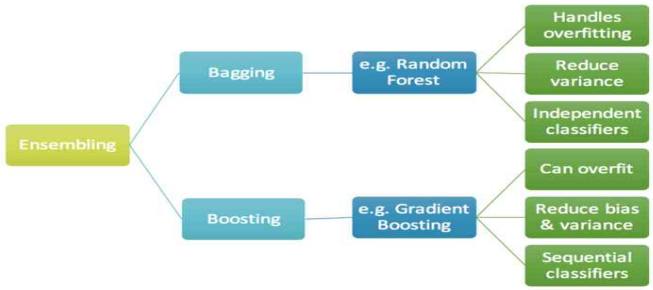
[그림 5-6] Lasso Regression
출처: datamarket 빅데이터 강의



[그림 5-7] Ridge Regression
출처: datamarket 빅데이터 강의

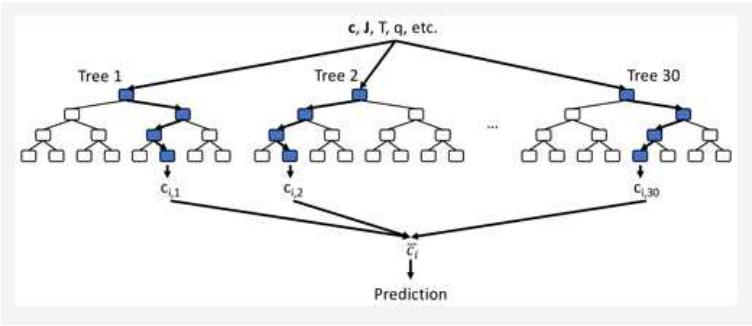
□ Lasso Regression(L1)과 Ridge Regression(L2)은 정규화 모델이다. 둘 다 제약조건에 따라서 변수의 영향도를 줄이는데 Ridge Regression은 변수의 계수를 축소해 모델의 복잡도를 줄인다. Lasso Regression은 몇 개의 변수만 선택하고 나머지 변수의 계수는 0으로 줄여버린다. 즉, Feature Selection의 효과가 있다. 두 모델의 차이는 변수의 계수를 줄이는 정도의 차이이다.

7) Random Forest & Gradient Boosting

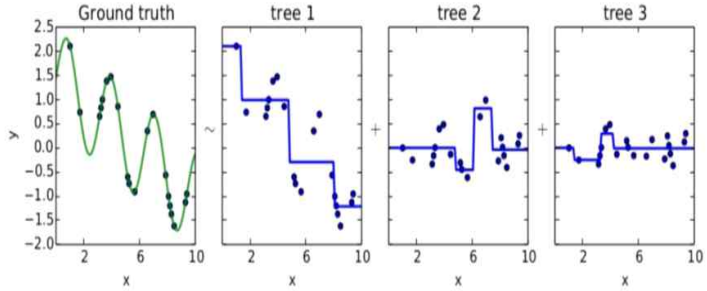


[그림 5-8] 앙상블 개요
출처: Deep Play(네이버 블로그)

□ Random Forest 와 Gradient Boosting은 앙상블 기법이다. 그 중에서 Random Forest Regression은 배깅 기법이고 Gradient Boosting Regression은 부스팅 기법이다. Random Forest Regression은 여러 개의 Decision Tree를 결합해서 예측하는 방법으로 각 Decision Tree 예측값의 평균을 사용한다.



[그림 5-9] Random Forest
출처: datamarket 빅 데이터 강의



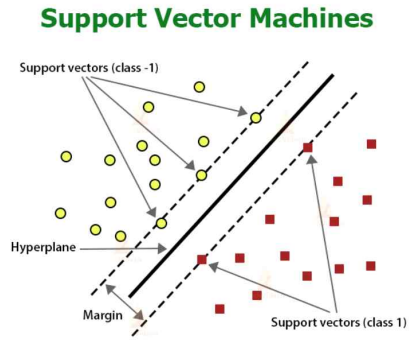
[그림 5-10] GBM
출처: Deep Play(네이버 블로그)

□ Gradient Boosting는 여러 개의 Decision Tree를 결합해서 잔차를 줄여나가는 방법으로 모델의 성능을 높이는데 사용된다.

8) XGBoost

□ XGBoost는 GBM + 정규화 이다. GBM은 해당 학습 data에 잔차를 계속 줄여 overfitting 되기 쉽다는 문제점이 있다. 이를 해결하기 위해 XGBoost는 GBM에 정규화 텀을 추가한 알고리즘이다. 따라서 XGBoost의 정규화 텀은 tree 복잡도가 증가할수록 loss에 패널티를 주는 방식으로 overfitting을 막고 있다.

9) SVM(Support Vector Machine)



[그림 5-11] SVM
출처: techvidvan

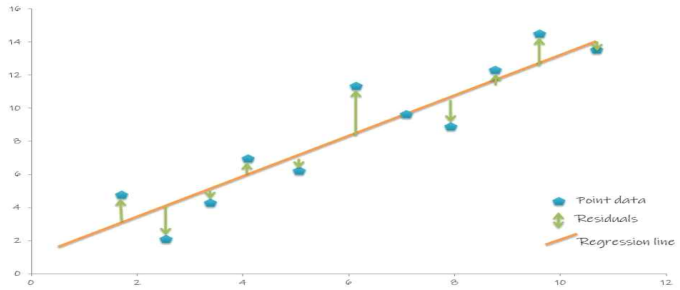
□ SVM은 분류를 하는 결정 경계선을 만드는데 가장 가까이 있는 데이터 포

인트들을 가지고 경계선을 정하고 데이터 포인트들과 경계선 사이의 거리를 최대화 시켜 분류를 시켜준다.

10) RMSE(Root Mean Squared Error)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

[그림 5-12] RMSE
출처: chris-song(brunch)



[그림 5-13] RMSE
출처: hatarilabs(blog)

☐ 회귀모델의 평가방법중 하나로 널리 쓰고있는 방법이다.

☐ MSE는 실제값-예측값을 한 다음에 그것에 제곱한 것에 평균이고 그것의 Root를 씌운 것이 RMSE다. 크기의존적에러가 있다.