


[2023 빅콘테스트][빅데이터플랫폼 활용 분야]

1. 배경 & 목적
2. 주최 & 참가 대상 & 성과
3. 프로젝트 기간(대회 기간)
4. 담당 역할
강수연
정가연
최민
5. 데이터 분석 Process
ch0. 데이터 수집 
ch1. 데이터 전처리
1-1. 이상치 처리 및 결측값 대체
1-2. 파생변수 생성 및 시계열 확장
1-3. 이종 데이터 병합
ch2. 데이터 분석 모형 구축
ch3. 데이터 시각화
3-1. 예측 매출등급 및 상권특성 시각화
ch4. 한계와 의의
4-1. 한계
4-2. 의의

1. 배경 & 목적

- 분석 주제명 : 상업용 부동산 가치 창출을 위한 소상공인 매출등급 예측모형 제작 및 활용 방안 제시
- 분석 배경 : 2023년 초 코로나 사태 종료 후 고물가, 고금리, 고환율 등으로 인해 소상공인 위기감은 계속 증가하여 상업용 부동산 시장이 침체되었습니다. 특히, 창업/폐업 비율이 상대적으로 높은 음식점의 정확한 매출 진단을 통해 폐업 예방 및 상권 활성화 방안 모색 등 상업용 부동산의 가치 창출을 도모하고자 합니다.
- 분석 목적 : 서울특별시 지역상권의 특성 및 세부적인 (소상공인) 물건지의 입지조건을 분석하여 지역상권 경제에 영향을 미치는 소상공인 매출규모를 예측하고자 합니다. 이는 지역경제에 필요한 정보를 제공할 수 있는 기회가 될 것이라고 사료됩니다.

2. 주최 & 참가 대상 & 성과

- 주최: 과학기술정보통신부, 한국지능정보사회진흥원
- 참가 자격 및 팀 인원 제한 사항: 데이터에 관심있는 누구나, 개인 또는 팀(팀장포함 최대 4명)
- 성과: ING

3. 프로젝트 기간(대회 기간)

- 사전 분석계획서 제출 마감: 2023년 9월 15일
- 결과보고서 제출 마감: 2023년 9월 27일
- 1차 서류심사: 2023년 10월 7일 ~ 11월 7일(심사 중)

4. 담당 역할

강수연

- 역할과 책임**
SQL 쿼리문을 통한 서울특별시 소상공인 KCD 신용 데이터·상권특성 공공 데이터의 추출과 병합, Python을 활용한 결측값 대체와 2단계 머신러닝 모형 클래스 구현을 맡았습니다.
- 성장한 경험**
소호 신용 데이터의 익명정보 처리와 분기 데이터를 계절성 지수를 생성하고 결합함으로써 월별로 확장하여 시계열 패턴을 반영하는 전처리를 수행하였습니다.

정가연

- 역할과 책임**
서울특별시 상권 분석 서비스 데이터와 SQL 쿼리문을 통해서 신용 데이터 수집 후 병합을 진행하였습니다.
결측값의 50%는 유사도 기준으로 근접한 변수 간 그룹화해서 처리했습니다.
분기로 이뤄진 상권 데이터를 월별 데이터로 확장해서 이종 데이터의 시계열을 일치시켰습니다. 이를 위해 신용거래정보 기반의 월별 계절성 지수를 파생하여 상관관계가 있는 변수와 1차원 축소된 값을 곱하는 방식으로 확장해보았습니다.
시계열적인 패턴을 고려해서 23년 1월, 2월 매출등급 예측 회귀 모델을 구축한 후 모델링을 진행하였습니다.
- 성장한 경험**
SQL 쿼리문을 통해서 데이터 수집해보았고 200개가 넘는 다량의 데이터를 결합해보았습니다.
이종 데이터의 시계열을 일치시키기 위해서 PCA 기반의 계절성 지수를 만들어서 데이터에 시계열적인 특성을 포함시키는 방법론을 적용해보았습니다.
22년 데이터를 사용해서 23년 1월을 예측하였고, 23년 1월을 포함해서 23년 2월 예측을 진행해보며 시계열 데이터 기반의 회귀분석 예측을 진행해볼 수 있었습니다.

최민

- 역할과 책임**
KCD 신용데이터, 공공데이터를 분석한 결과를 Tableau를 활용해 상권특성 시각화를 진행하였습니다.
<상권의 특징>
1. 카드 / 배달 매출액 변동계수와 매출등급 간의 관계
2. 객단가와 매출등급 간의 관계
3. 손익분기점과 매출등급 간의 관계
4. 부가세 차감 전 영업이익과 매출등급 간의 관계

를 라인차트와 막대그래프를 이용하여 시각화 하고 의미있는 인사이트를 도출하였습니다.

• 성장한 경험

SQL문을 사용하여 데이터를 수집 및 정제하였습니다. 시각화를 진행하는 과정에서 도메인지식과 분석결과를 연결하여 설명하였습니다.

5. 데이터 분석 Process

ch0. 데이터 수집 v

▼ 데이터 수집

- 2022년 ~ 2023년 서울열린데이터광장의 상권특성 관련 공개 데이터셋
 - 유형
 - 서울시 행정동별 지하철 총 승차 승객수 정보
 - 서울시 행정동별 버스 총 승차 승객수 정보
 - 서울시 우리마을가게 상권분석 서비스(행정동별 상권변화지표)
 - 수집방법
 - 서울열린데이터광장 홈페이지에서 다운로드.
- 2022년 ~ 2023년 (주)오아시스비즈니스의 수익형 부동산 관련 공개 데이터셋
 - 유형
 - 상업용 부동산거래량(금액) 대비 유동인구
 - 상업용 부동산의 공실을 대비 매매가, 임대료(이하 '매매가 점수')
 - 수집방법
 - 부동산 빅데이터 플랫폼에서 다운로드.

▼ 활용 데이터

- 서울열린데이터광장의 상권특성 관련 공개 데이터셋

	기준_년_코드	기준_분기_코드	상권_구분_코드	상권_구분_코드_명	상권_코드	상권_코드_명	월_평균_소득_금액	소득_구간_코드	지출_총금액	식료품_지출_총금액	...	기준_월_코드	법정동_코드	STDG_EMD_CD	행정동_코드_명	상권_변화_지표	상권_변화_지표_명	운영_영업_개월_평균	폐업_영업_개월_평균	서울_운영_영업_개월_평균	서울_폐업_영업_개월_평균
0	2023	1	A	관북로상권	3110001	이북5도청사	NaN	NaN	520427635.0000	129126028.0000	...	12.0000	11110182.0000	11110182.0000	평창동	HH	정체	111.0000	58.0000	104.0000	52.0000
1	2023	1	A	관북로상권	3110001	이북5도청사	NaN	NaN	520427635.0000	129126028.0000	...	12.0000	11110183.0000	11110183.0000	평창동	HH	정체	111.0000	58.0000	104.0000	52.0000
6	2023	1	A	관북로상권	3110002	독립문역 1번	NaN	NaN	232465205.0000	54998987.0000	...	12.0000	11110187.0000	11110187.0000	무악동	HH	정체	105.0000	64.0000	104.0000	52.0000
...
141699	2023	1	R	전통시장	3130156	시영2단지 무지개종합상가(중계 무지개2단지아파트상가)	NaN	NaN	NaN	NaN	...	12.0000	11350106.0000	11350106.0000	중계273동	LH	상권확장	97.0000	52.0000	104.0000	52.0000

<class 'pandas.core.frame.DataFrame'>				31	개업_점포_수	4831 non-null	float64	69	시간대_6_생활인구_수	4831 non-null	float64
Int64Index: 8263 entries, 0 to 141704				32	폐업_률	4831 non-null	float64	70	월요일_생활인구_수	4831 non-null	float64
Data columns (total 93 columns):				33	폐업_점포_수	4831 non-null	float64	71	화요일_생활인구_수	4831 non-null	float64
#	Column	Non-Null Count	Dtype	34	프랜차이즈_점포_수	4831 non-null	float64	72	수요일_생활인구_수	4831 non-null	float64
0	기준_년_코드	8263 non-null	int64	35	집객시설_수	1867 non-null	float64	73	목요일_생활인구_수	4831 non-null	float64
1	기준_분기_코드	8263 non-null	int64	36	관공서_수	1800 non-null	float64	74	금요일_생활인구_수	4831 non-null	float64
2	상권_구분_코드	8263 non-null	object	37	은행_수	1632 non-null	float64	75	토요일_생활인구_수	4831 non-null	float64
3	상권_구분_코드_명	8263 non-null	object	38	종합병원_수	241 non-null	float64	76	일요일_생활인구_수	4831 non-null	float64
4	상권_코드	8263 non-null	int64	39	일반_병원_수	507 non-null	float64	77	기준_년월_코드	8235 non-null	float64
5	상권_코드_명	8263 non-null	object	40	약국_수	1773 non-null	float64	78	엑스와표_값	8235 non-null	float64
6	월_평균_소득_금액	0 non-null	float64	41	유치원_수	929 non-null	float64	79	와이와표_값	8235 non-null	float64
7	소득_구간_코드	0 non-null	float64	42	초등학교_수	1434 non-null	float64	80	시군구_코드	8235 non-null	float64
8	지출_총금액	3317 non-null	float64	43	중학교_수	903 non-null	float64	81	행정동_코드	8235 non-null	float64
9	식료품_지출_총금액	3317 non-null	float64	44	고등학교_수	873 non-null	float64	82	형태정보	0 non-null	float64
10	의류_신발_지출_총금액	3317 non-null	float64	45	대학교_수	731 non-null	float64	83	기준_월_코드	8235 non-null	float64
11	생활용품_지출_총금액	3317 non-null	float64	46	백화점_수	17 non-null	float64	84	법정동_코드	8235 non-null	float64
12	의료비_지출_총금액	3317 non-null	float64	47	슈퍼마켓_수	306 non-null	float64	85	STDG_EMD_CD	8235 non-null	float64
13	교통_지출_총금액	3317 non-null	float64	48	극장_수	3 non-null	float64	86	행정동_코드_명	8235 non-null	object
14	여가_지출_총금액	3317 non-null	float64	49	숙박_시설_수	573 non-null	float64	87	상권_변화_지표	8235 non-null	object
15	문화_지출_총금액	3317 non-null	float64	50	공항_수	20 non-null	float64	88	상권_변화_지표_명	8235 non-null	object
16	교육_지출_총금액	3317 non-null	float64	51	철도_역_수	0 non-null	float64	89	운영_영업_개월_평균	8235 non-null	float64
17	유흥_지출_총금액	3317 non-null	float64	52	버스_터미널_수	7 non-null	float64	90	폐업_영업_개월_평균	8235 non-null	float64
18	아파트_단지_수	3699 non-null	float64	53	지하철_역_수	1061 non-null	float64	91	서울_운영_영업_개월_평균	8235 non-null	float64
19	아파트_가격_1_억_미만_세대_수	2819 non-null	float64	54	버스_정거장_수	1867 non-null	float64	92	서울_폐업_영업_개월_평균	8235 non-null	float64
20	아파트_가격_1_억_세대_수	3699 non-null	float64	55	총_생활인구_수	4831 non-null	float64	dtypes: float64(82), int64(3), object(8)			
21	아파트_가격_2_억_세대_수	2819 non-null	float64	56	남성_생활인구_수	4831 non-null	float64	memory usage: 5.9+ MB			
22	아파트_가격_3_억_세대_수	180 non-null	float64	57	여성_생활인구_수	4831 non-null	float64				
23	아파트_가격_4_억_세대_수	180 non-null	float64	58	연령대_10_생활인구_수	4831 non-null	float64				
24	아파트_가격_5_억_세대_수	180 non-null	float64	59	연령대_20_생활인구_수	4831 non-null	float64				
25	아파트_가격_6_억_이상_세대_수	1060 non-null	float64	60	연령대_30_생활인구_수	4831 non-null	float64				
26	서비스_업종_코드	4831 non-null	object	61	연령대_40_생활인구_수	4831 non-null	float64				
27	서비스_업종_코드_명	4831 non-null	object	62	연령대_50_생활인구_수	4831 non-null	float64				
28	점포_수	4831 non-null	float64	63	연령대_60_이상_생활인구_수	4831 non-null	float64				
29	유사_업종_점포_수	4831 non-null	float64	64	시간대_1_생활인구_수	4831 non-null	float64				
30	개업_율	4831 non-null	float64	65	시간대_2_생활인구_수	4831 non-null	float64				
				66	시간대_3_생활인구_수	4831 non-null	float64				
				67	시간대_4_생활인구_수	4831 non-null	float64				
				68	시간대_5_생활인구_수	4831 non-null	float64				

- (주)오아시스비즈니스의 수익형 부동산 관련 공개 데이터셋

	기 준 연 월	필 지 고 유 번 호	법 정 동 코 드	업 종 코 드	매 매 가 점 수
0	202203	1111010100100040014	11110101	B01	58.6100
1	202203	1111010100100480000	11110101	A01	73.9100
2	202203	1111010100100500031	11110101	C05	65.6400
...
781778	202203	5013032026104230003	50130320	A01	99.2100
781779	202203	5013032026104300003	50130320	C01	99.0900
781780	202203	5013032026104390007	50130320	A03	99.2600

상업용 부동산거래량(금액) 대비 유동인구

	기 준 연 월	필 지 고 유 번 호	법 정 동 코 드	업 종 코 드	유 동 인 구 수
0	202303	1111010100100010000	11110101	C05	41.3400
1	202303	1111010100100030100	11110101	A03	23.3500
2	202303	1111010100100040014	11110101	B01	6.8200
...
1168435	202303	5013032026113660001	50130320	C06	3.2300
1168436	202303	5013032026113660001	50130320	A01	3.9000
1168437	202303	5013032026114300001	50130320	C01	4.0200

상업용 부동산의 매매가 점수

ch1. 데이터 전처리

1-1. 이상치 처리 및 결측값 대체

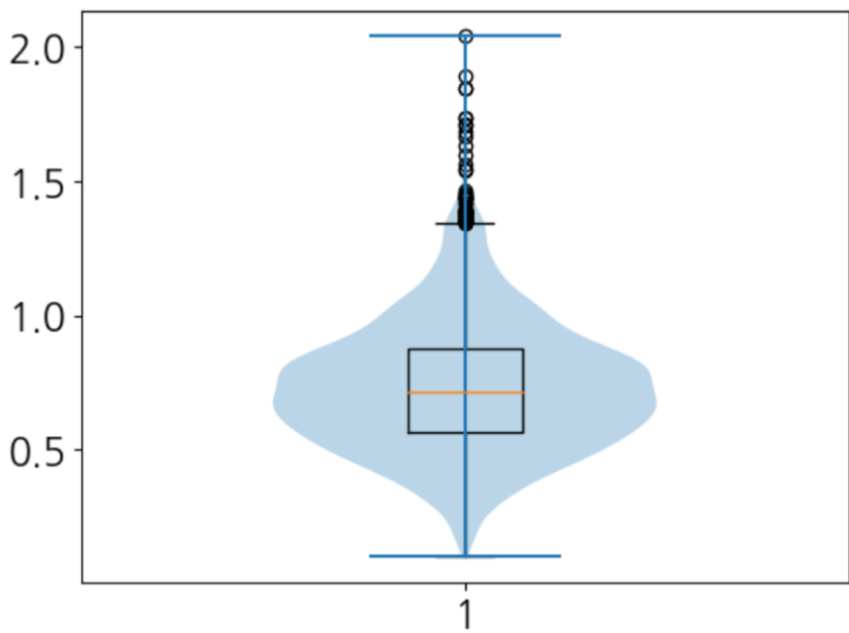
▼ 분석 의도

서울시 상권별 외식업종 소상공인 집단 내 이상치로 인하여 집단별 중심경향치(산술평균, 표준편차)를 활용하여 구한 계절성 지수 중 기준인 1에서 극단적으로 벗어난 값이 있었습니다.

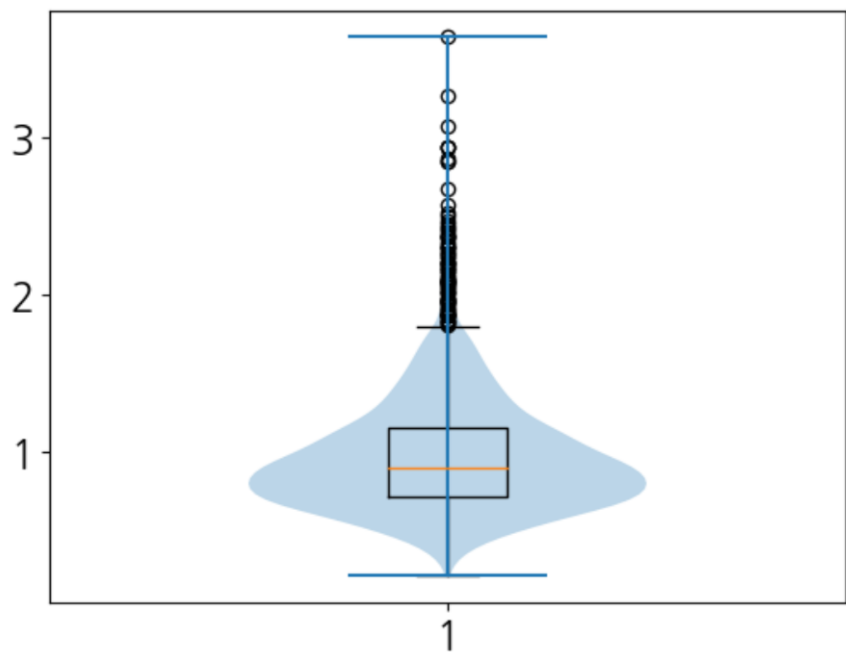
이상치가 반영된 계절성 지수를 활용하여 분기별 데이터의 시계열을 월별로 확장하거나 매출등급을 예측하는 모형을 적합(fitting)할 경우 결과를 신뢰할 수 없기 때문에 후술된 방법으로 전처리를 수행하였습니다.

기준일자	소호 사업_수	경영 위기 사업_총 합	프랜 차이즈 사업_총 합	정규 고용 인원_평 균	사업 장방 문고 객수_평 균	사업 장방 문신 규고 객수_평 균	매출 액_변 동계수	카드 매출 액_변 동계수	배달매 출액_변 동계수	주말 카드 매출 액_변 동계수	주말배 달매출 액_변 동계수	매입 액_변 동계수	매출 총이 익_변 동계수	부가 가치 세_변 동계수	부가 세차 감전 영업 이익_변 동계수	사업 장방 문고 객수_변 동계수	사업 장방 문신 규고 객수_변 동계수	사업 장임 대면 적_변 동계수	월임 대표_변 동계수	임대 보증 금_변 동계수	손익분 기점매 출액_변 동계수	객단 가_변 동계수	신규 고객 단가_변 동계수
202201	0.9528	0.0000	1.0180	0.9591	0.8503	0.8241	1.0216	1.0186	-0.0013	1.0336	-0.0004	0.9594	1.1218	0.8633	1.1300	1.0074	1.0347	0.9881	0.9942	0.9817	3.5917	1.0059	0.9720
202202	0.9538	1.4188	1.0061	0.9671	0.7280	0.6905	1.0428	1.0301	-0.0009	1.0220	0.0016	1.0660	1.1044	0.9850	1.3460	1.0239	1.0477	0.9952	0.9982	0.9847	15.2345	1.0160	0.9852
202203	0.9624	1.2986	1.0286	0.9714	0.8883	0.8567	1.0189	1.0144	0.0085	1.0197	-0.0001	0.9752	1.2202	0.8637	1.1547	1.0111	1.0222	0.9950	0.9973	0.9898	3.3060	1.0119	1.0034
202204	0.9555	1.0421	1.0118	0.9759	1.0398	1.0437	0.9799	0.9868	0.0071	0.9621	0.0005	0.9575	1.0251	0.7983	0.6626	0.9819	0.9882	0.9896	0.9936	0.9928	1.7315	0.9956	0.9876
202205	0.9694	0.8377	1.0061	0.9828	1.1280	1.1568	0.9956	1.0009	0.0143	0.9883	0.0005	0.9176	0.9729	2.3832	0.9948	0.9962	1.0019	1.0055	0.9963	1.0012	1.3886	0.9960	0.9923
202206	0.9776	1.0776	1.0111	0.9744	1.0538	1.0681	0.7842	0.7429	12.9139	0.8169	11.9954	0.9526	0.7913	0.7661	0.8397	0.7756	0.7783	0.9942	0.9914	0.9945	5.7873	1.0378	1.0346
202207	0.9946	1.0053	1.0000	0.9692	1.0902	1.1142	1.0038	1.0054	-0.0078	1.0000	NaN	0.9327	0.9907	0.8585	1.0310	1.0093	1.0124	0.9947	0.9986	0.9868	2.4555	0.9845	0.9875
202208	1.0081	1.0287	1.0110	0.9781	1.0373	1.0416	1.0142	1.0153	-0.0033	1.0169	0.0007	0.8904	0.9955	0.9200	0.6941	1.0143	1.0144	0.9950	1.0007	1.0318	1.9408	0.9874	0.9984
202209	1.0265	1.0889	1.0163	0.9851	1.0248	1.0257	1.0007	1.0047	-0.0022	1.0209	0.0011	0.9506	1.0406	0.5190	1.2496	1.0190	1.0038	1.0126	1.0069	1.0057	4.3873	0.9746	0.9977
202210	1.0423	0.8969	0.9998	0.9699	1.0724	1.0896	1.0122	1.0222	0.0150	1.0126	0.0009	0.9580	0.9686	1.0316	0.9948	1.0234	1.0172	1.0190	0.9970	0.9960	10.6527	0.9821	0.9913
202211	1.0334	1.0902	0.9783	0.9693	1.0464	1.0151	0.9988	1.0061	0.0015	1.0120	0.0005	0.9318	0.9527	0.9636	0.9818	1.0223	1.0034	1.0167	1.0055	1.0059	-39.7562	0.9744	0.9893
202212	1.0338	1.2398	0.9783	0.9670	1.0267	1.0301	1.0265	1.0341	0.0075	1.0140	NaN	0.9426	0.9742	0.9955	0.9950	1.0043	1.0030	1.0093	1.0017	1.0021	-9.2585	1.0326	1.0234
202301	1.0318	0.7551	0.9890	1.0883	0.9679	0.9708	1.0324	1.0418	NaN	1.0472	-0.0009	1.6550	0.8628	1.0004	0.8576	1.0561	1.0484	0.9918	1.0066	1.0081	6.7845	0.9803	0.9861
202302	1.0315	1.1044	0.9673	1.1079	0.9713	0.9897	1.0453	1.0485	NaN	1.0212	0.0002	0.9491	1.0073	1.0643	1.0775	1.0301	1.0193	0.9922	1.0074	1.0102	3.4446	1.0229	1.0275
202303	1.0265	1.1160	0.9782	1.1344	1.0749	1.0833	1.0233	1.0282	0.0477	1.0127	NaN	0.9613	0.9719	0.9874	0.9907	1.0250	1.0051	1.0012	1.0048	1.0088	3.3096	0.9979	1.0236

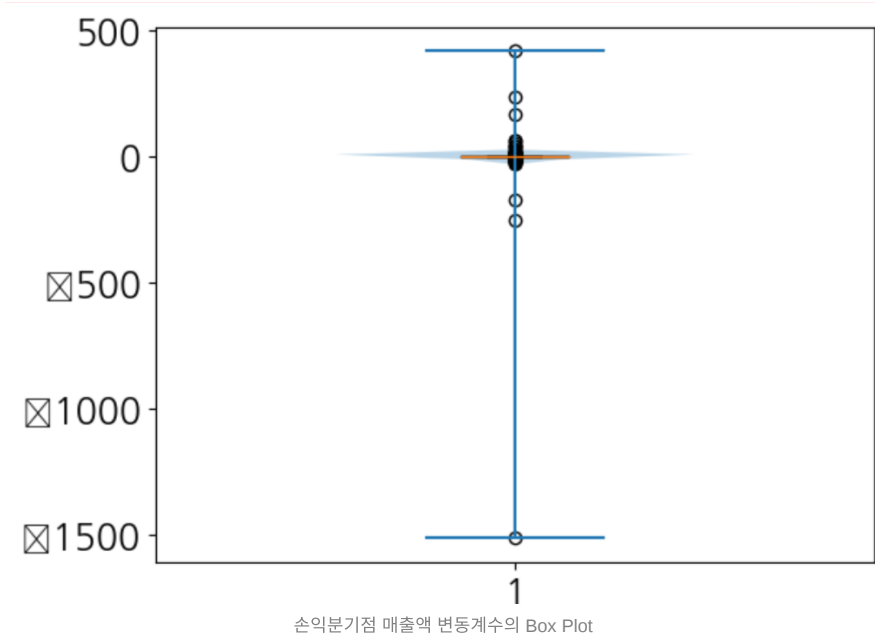
이상치와 결측값이 포함된 계절성 지수



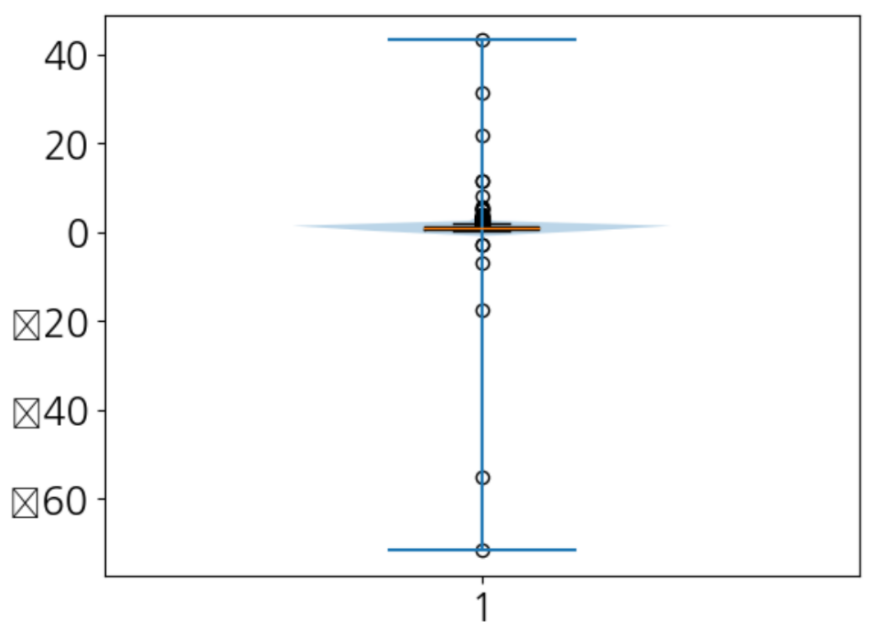
카드매출액 변동계수의 Box Plot



객단가 변동계수의 Box Plot



손익분기점 매출액 변동계수의 Box Plot



부가가치세 차감 전 영업이익 변동계수의 Box Plot

▼ 결과

• 이상치 처리 방법

시계열별로 일관된 기준을 적용하기 위해 1사분위수(하위 25%인 값)과 3사분위수(하위 75%인 값)에서 사분범위(InterQuartile Range, IQR)의 1.5배 이상 벗어난 값을 이상치로 판정하고 결측값으로 처리하였습니다.

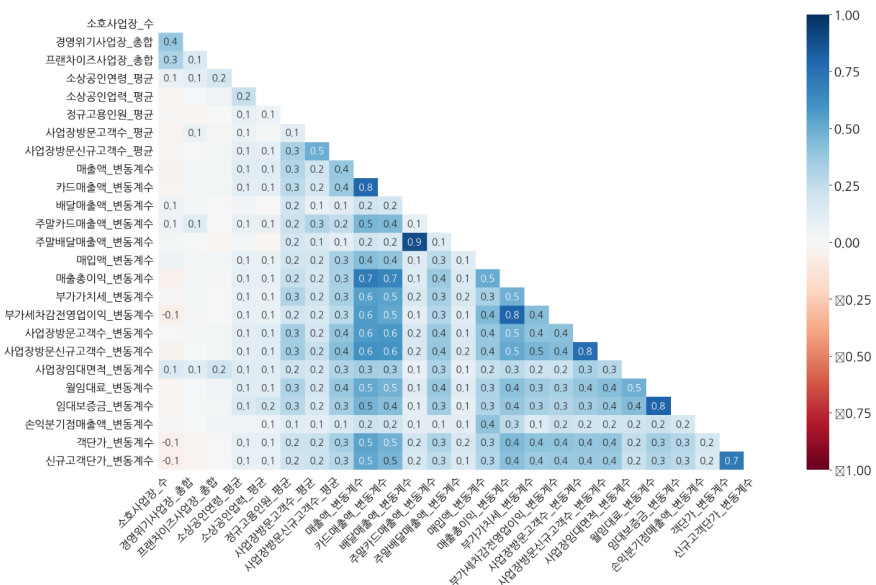
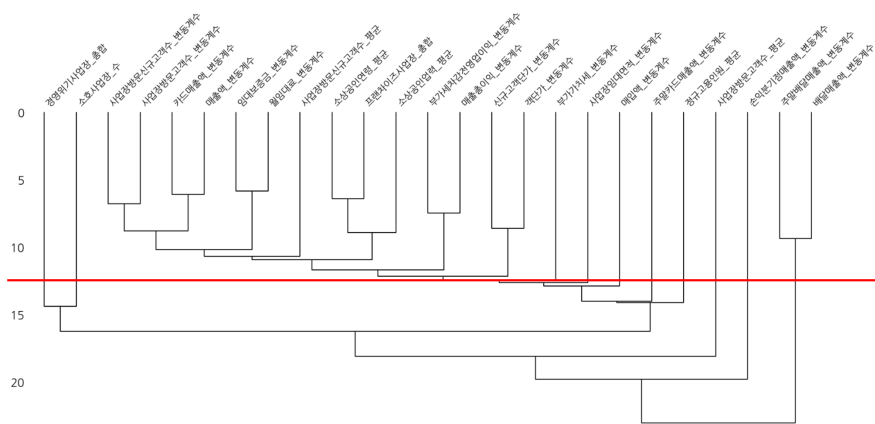
• 결측값 대체 방법

결측값의 경우 KNN Imputer을 사용해서 유사도 기준으로 근접한 변수 간 그룹화해서 이상치를 처리하였습니다. 한국신용데이터(KCD) 소상공인 신용 데이터의 경우 덴드로그램(계층적 군집)을 시각화하였습니다. 0에 가깝게 쪼개지기 시작하는 클러스터 리프노드(특성변수) 간에 유사도가 상대적으로 높은 조합을 찾았습니다. 총 13개의 각 조합별로 결측값에 최근접한 5가지 이웃 데이터의 평균값(K-Nearest Neighbor)으로 대체하였습니다.

• 결측값 모형에 투입할 13개의 특성변수 조합

변수 1개	{‘사업장방문고객수_평균’, {‘정규고용인원_평균’}, {‘주말카드매출액_변동계수’}, {‘매입액_변동계수’}, {‘사업장임대면적_변동계수’}, {‘부가가치세_변동계수’}, {‘손익분기점매출액_변동계수’}
변수 2개인 조합	{‘배달매출액_변동계수’, ‘주말배달매출액_변동계수’}, {‘객단가_변동계수’, ‘신규고객단가_변동계수’}, {‘매출총이익_변동계수’, ‘부가세차감전영업이익_변동계수’}, {‘소호사업장_수’, ‘경영위기사업장_총합’}
변수 3개 이상인 조합	{‘소상공인업력_평균’, ‘프랜차이즈사업장_총합’, ‘소상공인연령_평균’}, {‘사업장방문신규고객수_평균’, ‘월임대료_변동계수’, ‘임대보증금_변동계수’, ‘매출액_변동계수’, ‘카드매출액_변동계수’, ‘사업장방문고객수_변동계수’, ‘사업장방문신규고객수_변동계수’}

• 결측값 시각화



▼ 시행착오

2022년 1월부터 2023년 2월까지 결측값 있는 변수 간 계층적 군집에 대한 14개의 트리를 시각화하여 군집을 나누기 위한 기준을 설정하였습니다.

다만, 14개의 개별적인 기준을 시점별로 적용하여 비어있는 값을 채우는 방식보다는 시계열 패턴이 반영된 소호 사업장의 신용거래정보 변수 조합을 찾아 대체하는 것이 real life를 반영하기 적합한 결론 끝에 전처리 방법을 수정하였습니다.

1-2. 파생변수 생성 및 시계열 확장

▼ 분석 의도

- 첫째, 서울시 상권특성 공공데이터의 경우 분기별 데이터만 있으며, 월별 데이터는 없었습니다. 따라서 월별 계절성 지수를 분기 데이터에 곱하여 시계열을 확장(매핑)하였습니다.
- 둘째, 경영성과(매출등급)를 예측하는 분석 목표를 달성하기 위해 소호 상권분석이라는 비즈니스 기준에 따른 새로운 변수를 파생하였습니다. 예를 들어, ‘손익분기점 매출액’, ‘부가가치세 차감 전 영업이익’, ‘사업장 방문고객 단가(이하 ‘객단가’라 합니다)’, ‘주말 카드[배달] 매출액 비중’ 등을 생성하였습니다.

▼ 결과

• 계절성 지수 생성방법

계절성 패턴을 따르는 분기 데이터에 대하여 월 단위로 확장(매핑)하는 방법은 다음과 같습니다. 먼저, 판매량, 방문고객 수, 매출액 등 시계열에 따라 변동가능성이 있는 변수의 월별 평균값을 계산하였습니다. 각 월별 평균값에서 연간 평균값을 나누어 계절성 지수를 산출하였습니다.

이후 월별 계절성 지수가 1보다 크면 평균 이상의 활동이 있었다고 해석하며, 1보다 작으면 평균에 미달하는 영업활동이 있다고 해석하였습니다. 이러한 계절성 지수를 활용하여 각 분기에 대한 컬럼 값을 월 별 데이터로 조정하였습니다.

- 데이터에 계절성을 반영하기 위한 전처리 완료된 지수

	기준일 자	배달매출액_변 동계수	주말배달매출액_변 동계수	손익분기점매출액_변 동계수	사업장방문고객수_평 균	점구고용인원_평 균	주말카드매출액_변 동계수	매입액_변동 계수	사업장임대면적_변 동계수	부가가치세_변 동계수	객단가_변동 계수	신규고객단가_변 동계수	매출총이익_변 동계수	부가세차감전영업이익_변 동계수	프랜차이즈사업장_총 합	소호사업장_수	경영위기사업장_총 합	사업장방문신규고객수_평 균	월임대료_변동 계수	임대보증금_변 동계수	매출액_변동 계수	카드매출액_변 동계수	사업장방문고객수_변 동계수	사업장방문신규고객수_변 동계수
0	202201	0.9564	0.9403	1.0283	0.8896	0.9341	1.0471	0.9759	0.9857	1.0235	1.0014	0.9747	1.0846	1.1379	1.0259	0.9318	0.0000	0.8081	0.9975	0.9819	1.0283	1.0127	1.0051	1.0189
1	202202	0.9463	0.9294	1.0386	0.7282	0.9521	1.0424	0.9568	0.9954	1.0477	1.0148	0.9817	1.1198	1.2125	1.0131	0.9641	1.4374	0.6889	0.9954	0.9821	1.0434	1.0384	1.0330	1.0535
2	202203	0.9631	0.9463	1.0649	0.8732	0.9560	0.9920	0.9943	0.9919	1.0448	1.0062	0.9964	1.0634	1.0810	1.0083	0.9691	1.3578	0.8323	1.0008	0.9923	1.0342	1.0246	1.0135	1.0121
3	202204	0.9669	0.9482	1.0430	1.0069	0.9614	0.9471	0.9611	0.9860	1.0300	0.9872	0.9823	0.9890	0.9838	0.9937	0.9657	1.0511	1.0550	1.0015	0.9976	0.9859	0.9931	0.9983	0.9950
4	202205	0.9459	0.9293	0.9974	1.1129	1.0043	0.9883	0.9203	0.9944	1.0091	0.9943	0.9881	0.9936	0.9728	0.9853	0.9750	0.9001	1.1702	1.0013	1.0035	0.9941	0.9944	1.0111	1.0115
5	202206	1.2375	1.4159	0.9060	1.0287	0.9557	0.7879	0.8320	0.9901	0.7101	0.9986	1.0347	0.7645	0.7561	0.9998	0.9767	0.9423	1.0849	0.9950	1.0027	0.7577	0.7308	0.7416	0.7380
6	202207	0.9833	0.9703	1.0073	1.0827	0.9546	0.9991	0.9336	0.9935	1.0152	0.9991	1.0049	1.0097	1.0002	0.9850	0.9869	1.0962	1.1216	1.0044	0.9948	1.0087	1.0122	1.0264	1.0288
7	202208	0.9956	0.9880	1.0277	1.0260	0.9774	1.0178	0.9151	0.9747	1.0105	0.9986	0.9932	1.0256	1.0115	0.9967	1.0075	1.0686	1.0470	1.0020	1.0105	1.0196	1.0188	1.0262	1.0272
8	202209	0.9933	0.9908	1.0763	1.0205	0.9845	1.0248	0.9747	1.0368	1.0294	0.9968	1.0071	0.9976	0.9876	1.0053	1.0085	0.9337	1.0380	1.0010	1.0064	1.0039	1.0070	1.0129	1.0001
9	202210	0.9962	0.9878	1.0520	1.0793	0.9694	1.0150	0.9597	1.0259	1.0197	0.9978	1.0012	1.0037	0.9970	0.9908	1.0598	0.9768	1.0832	0.9969	0.9975	1.0164	1.0238	1.0089	1.0211
10	202211	0.9858	0.9803	1.0378	1.0486	0.9773	1.0056	0.9546	1.0215	1.0213	0.9837	0.9947	1.0046	0.9975	0.9958	1.0184	1.1050	1.0347	1.0022	1.0058	1.0066	1.0143	1.0299	1.0135
11	202212	0.9963	0.9784	1.0143	1.0617	0.9876	1.0318	0.9663	1.0068	1.0651	1.0186	1.0175	1.0211	0.9966	0.9652	1.0320	1.0770	1.0378	0.9979	1.0017	1.0340	1.0408	0.9952	1.0080
12	202301	1.0046	0.9925	0.5578	0.9894	1.1016	1.0712	1.6976	0.9899	1.0027	0.9914	0.9860	0.8984	0.8688	1.0047	1.0557	0.8253	0.9536	1.0007	1.0096	1.0272	1.0413	1.0526	1.0488
13	202302	1.0103	0.9948	1.0784	0.9729	1.1283	1.0451	0.9650	1.0020	0.9844	1.0202	1.0235	1.0171	1.0043	1.0100	1.0215	1.1650	0.9784	1.0033	1.0076	1.0290	1.0446	1.0293	1.0207
14	202303	1.0185	1.0078	1.0701	1.0794	1.1556	0.9848	0.9931	1.0053	0.9865	0.9914	1.0141	1.0075	0.9925	1.0205	1.0272	1.0638	1.0664	1.0001	1.0058	1.0111	1.0034	1.0161	1.0028

- 소호 사업장의 경영성과 도메인에 적합한 파생변수 생성

변동계수(coefficient of variation, CV)란 표준편차를 산술평균을 기준으로 표준화(standardization)한 값입니다. 표준편차를 산술평균으로 나눈 값으로 같은 단위를 가지는 표준편차를 평균으로 나누면, 단위가 사라지고 표준화된 수치를 비교하기 위해 아래의 파생변수를 변동계수로 변환하는 익명정보 처리 쿼리를 통해 디사일로 클린룸에서 데이터를 추출하였습니다.

연번	파생변수명	산식
1	매출액	카드매출액 + 배달매출액(변동계수)
2	매입액	카드매입액 + 현금매입액 (변동계수)
3	매출총이익	매출액 - 매입액(단, 기초와 기말재고액 미고려) (변동계수) * 상품매출원가의 원칙적인 산출과정 기초상품재고액 + 당기상품매입액 - (매입환출 + 매입에누리 + 매입할인) - 기말상품재고액 - 매출원가란 판매한 상품 또는 제품에 대한 매입원가 및 제조원가
4	부가가치세	매출세액(매출 세금계산서) - 매입세액(매입세금계산서) * 부가가치세는 매출액의 10%로 잡는 것이 일반적
5	부가세 차감 전 영업이익	매출총이익(매출액 - 매입액) - 월 임대료 (변동계수) * 원칙 손익계산서 내 판매비와 관리비 계정과목, 급여, 감가상각비, 대손상각비(=손상차손) 등을 고려함이 원칙이나 본 분석에서는 미고려
6	영업이익	5.부가세 차감 전 영업이익 - 4.부가가치세 (변동계수)
7	주말 매출액	카드매출액 + 배달매출액 (5.에서 1. 나누어 주말 배후세대의 영향력 관측) (변동계수)
8	객단가 및 신규 고객 단가	사업장방문고객 수 및 신규고객 수(변동계수) (8. 을 1. 로 나누어 객단가 산출)
9	손익분기점 매출액	월임대료 / (1 - ((부가가치세 + 매입액) / 매출액)) (변동계수) * 원칙 고정비용 / 변동비용(= 1 - 변동비용/매출액)

▼ 시행착오

데이터 결합분석 플랫폼(Desilo DCR)에서 추출 행 수를 500개로 제한하여 시점별(월별) · 상권별 · 업종별로 컬럼을 집계할 경우 데이터를 반출할 수 없는 문제가 발생하였습니다.

따라서 2022년 1월부터 2023년 3월까지 시점별로 총계처리된 개별 데이터를 반출한 후 다시 열 기준 병합을 수행하여 분석 데이터 셋을 구축할 수 있었습니다.

- 1) 소호신용 데이터 추출 SQL문

- SOHO_202201_INCOME_230916
2022.01. 한국신용데이터(KCD) 소호 사업장 월별 신용거래정보

```
SELECT
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash),
  STANDARD_DEVIATION_POPULATION(((kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery) - (kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash))),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_sales_invoice - kcd.transactions.sum_purchase_invoice),
  STANDARD_DEVIATION_POPULATION((((kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery) - (kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash)) - kcd.transact
  STANDARD_DEVIATION_POPULATION((((kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery) - (kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash)) - kcd.transac
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_weekend_sales_card + kcd.transactions.sum_weekend_sales_delivery), STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_customer_cnt),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_new_customer_cnt),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_sales_delivery),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_weekend_sales_delivery),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_sales_card),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.sum_weekend_sales_card)
FROM kcd.meta-info
JOIN kcd.transactions
ON kcd.meta-info.b_id = kcd.transactions.b_id
WHERE kcd.transactions.month_id = "2022-01-01" AND kcd.meta-info.class_1_name = "외식업"
GROUP BY kcd.meta-info.trdar_nm, kcd.meta-info.trdar_no, kcd.meta-info.class_2_name
```

- SOHO_202201_BUSINESS_230911
2022.01. 소호 개인사업자 업력, 연령, 경영위기 FLAG, 프랜차이즈(브랜드) TF

```
SELECT
  COUNT(*),
  SUM(kcd.transactions.is_risky),
  SUM(kcd.meta-info.is_franchise),
  AVERAGE(kcd.meta-info.age),
  AVERAGE(kcd.meta-info.duration)
FROM kcd.meta-info
JOIN kcd.transactions
ON kcd.meta-info.b_id = kcd.transactions.b_id
WHERE kcd.meta-info.class_1_name = "외식업" AND kcd.transactions.month_id = "2022-01-01"
GROUP BY kcd.meta-info.trdar_nm, kcd.meta-info.class_2_name
```

- SOHO_202201_METAINFO_230911

2022.01. 소호 임대동향, 사업장 정보

```
SELECT
  AVERAGE(kcd.transactions.regular_employees_count),
  AVERAGE(kcd.transactions.sum_customer_cnt),
  AVERAGE(kcd.transactions.sum_new_customer_cnt),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.business_square_size),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.monthly_rental_fee),
  STANDARD_DEVIATION_POPULATION(kcd.transactions.rental_deposit)
FROM kcd.meta-info
JOIN kcd.transactions
ON kcd.meta-info.b_id = kcd.transactions.b_id
WHERE kcd.meta-info.class_1_name = "외식업" AND kcd.transactions.month_id = "2022-01-01"
GROUP BY kcd.meta-info.trdar_nm, kcd.meta-info.class_2_name
```

- SOHO_202201_STANDARDIZATION_230916

2022.01. 변동계수(coefficient of variation, CV) 산출을 위한 연산

```
SELECT
  AVERAGE(kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery),
  AVERAGE(kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash),
  AVERAGE((kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery) - (kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash)),
  AVERAGE(kcd.transactions.sum_sales_invoice - kcd.transactions.sum_purchase_invoice),
  AVERAGE(((kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery) - (kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash)) - kcd.transactions.monthly_rental_fee),
  AVERAGE(((kcd.transactions.sum_sales_card + kcd.transactions.sum_sales_delivery) - (kcd.transactions.sum_purchase_card + kcd.transactions.sum_purchase_cash)) - kcd.transactions.monthly_rental_fee),
  AVERAGE(kcd.transactions.sum_weekend_sales_card + kcd.transactions.sum_weekend_sales_delivery),
  AVERAGE(kcd.transactions.sum_customer_cnt), AVERAGE(kcd.transactions.sum_new_customer_cnt),
  AVERAGE(kcd.transactions.sum_sales_delivery),
  AVERAGE(kcd.transactions.sum_weekend_sales_delivery),
  AVERAGE(kcd.transactions.sum_sales_card),
  AVERAGE(kcd.transactions.sum_weekend_sales_card),
  AVERAGE(kcd.transactions.business_square_size),
  AVERAGE(kcd.transactions.monthly_rental_fee),
  AVERAGE(kcd.transactions.rental_deposit)
FROM kcd.meta-info
JOIN kcd.transactions
ON kcd.meta-info.b_id = kcd.transactions.b_id
WHERE kcd.transactions.month_id = "2022-01-01" AND kcd.meta-info.class_1_name = "외식업"
GROUP BY kcd.meta-info.trdar_nm, kcd.meta-info.trdar_no, kcd.meta-info.class_2_name
```

▼ 2) 상권특성 데이터 추출 SQL문

- 상권_아파트

```
SELECT
  AVERAGE(seoul.상권특성.아파트_단지_수),
  AVERAGE(seoul.상권특성.아파트_가격_1_억_미만_세대_수),
  AVERAGE(seoul.상권특성.아파트_가격_1_억_세대_수),
  AVERAGE(seoul.상권특성.아파트_가격_2_억_세대_수),
  AVERAGE(seoul.상권특성.아파트_가격_3_억_세대_수),
  AVERAGE(seoul.상권특성.아파트_가격_4_억_세대_수),
  AVERAGE(seoul.상권특성.아파트_가격_5_억_세대_수),
  AVERAGE(seoul.상권특성.아파트_가격_6_억_이상_세대_수),
  COUNT(*), SUM(kcd.meta-info.is_franchise)
FROM kcd.meta-info
JOIN seoul.상권특성
ON kcd.meta-info.trdar_nm = seoul.상권특성.상권_코드_명
WHERE kcd.meta-info.class_1_name = "외식업"
GROUP BY seoul.상권특성.기준_년_코드, seoul.상권특성.기준_분기_코드, kcd.meta-info.class_2_name
```

- 상권_점포

```
SELECT
  AVERAGE(seoul.상권특성.개업_점포_수),
  AVERAGE(seoul.상권특성.유사_업종_점포_수),
  AVERAGE(seoul.상권특성.폐업_점포_수),
  AVERAGE(seoul.상권특성.점포_수),
  AVERAGE(seoul.상권특성.프랜차이즈_점포_수),
  STANDARD_DEVIATION_POPULATION(seoul.상권특성.개업_율),
  STANDARD_DEVIATION_POPULATION(seoul.상권특성.폐업_율),
  COUNT(*), SUM(kcd.meta-info.is_franchise)
FROM kcd.meta-info
JOIN seoul.상권특성
ON kcd.meta-info.trdar_nm = seoul.상권특성.상권_코드_명
WHERE kcd.meta-info.class_1_name = "외식업"
GROUP BY seoul.상권특성.기준_년_코드, seoul.상권특성.기준_분기_코드, kcd.meta-info.class_2_name
```

- 상권_소득소비

```
SELECT
  AVERAGE(seoul.상권특성.교육_지출_총금액),
  AVERAGE(seoul.상권특성.교통_지출_총금액),
  AVERAGE(seoul.상권특성.생활용품_지출_총금액),
  AVERAGE(seoul.상권특성.문화_지출_총금액),
  AVERAGE(seoul.상권특성.유흥_지출_총금액),
  AVERAGE(seoul.상권특성.의료비_지출_총금액),
  AVERAGE(seoul.상권특성.의류_신발_지출_총금액),
  AVERAGE(seoul.상권특성.식품_지출_총금액),
  AVERAGE(seoul.상권특성.여가_지출_총금액),
  COUNT(*), SUM(kcd.meta-info.is_franchise)
FROM kcd.meta-info
JOIN seoul.상권특성
ON kcd.meta-info.trdar_no = seoul.상권특성.상권_코드
WHERE kcd.meta-info.class_1_name = "외식업"
GROUP BY seoul.상권특성.기준_년_코드, seoul.상권특성.기준_분기_코드, kcd.meta-info.class_2_name
```

- 상권_집객인구

```
SELECT
  AVERAGE(seoul.상권특성.집객시설_수),
  AVERAGE(seoul.상권특성.관광서_수),
  AVERAGE(seoul.상권특성.은행_수),
  AVERAGE(seoul.상권특성.종합병원_수),
  AVERAGE(seoul.상권특성.일반_병원_수),
  AVERAGE(seoul.상권특성.약국_수),
  AVERAGE(seoul.상권특성.유치원_수),
  AVERAGE(seoul.상권특성.초등학교_수),
  AVERAGE(seoul.상권특성.고등학교_수),
  AVERAGE(seoul.상권특성.대학교_수),
  AVERAGE(seoul.상권특성.백화점_수),
  AVERAGE(seoul.상권특성.슈퍼마켓_수),
  AVERAGE(seoul.상권특성.숙박_시설_수),
  AVERAGE(seoul.상권특성.버스_터미널_수),
  AVERAGE(seoul.상권특성.지하철_역_수),
  AVERAGE(seoul.상권특성.버스_정거장_수),
  COUNT(*), SUM(kcd.meta-info.is_franchise)
FROM kcd.meta-info
JOIN seoul.상권특성
ON kcd.meta-info.trdar_no = seoul.상권특성.상권_코드
WHERE kcd.meta-info.class_1_name = "외식업"
GROUP BY seoul.상권특성.기준_년_코드, seoul.상권특성.기준_분기_코드, kcd.meta-info.class_2_name
```

- 상권_상주인구

```
SELECT
  AVERAGE(seoul.상권특성.연령대_10_상주인구_수),
  AVERAGE(seoul.상권특성.연령대_20_상주인구_수),
  AVERAGE(seoul.상권특성.연령대_30_상주인구_수),
  AVERAGE(seoul.상권특성.연령대_40_상주인구_수),
  AVERAGE(seoul.상권특성.연령대_50_상주인구_수),
  AVERAGE(seoul.상권특성.연령대_60_이상_상주인구_수),
  AVERAGE(seoul.상권특성.총_가구_수),
  AVERAGE(seoul.상권특성.비_아파트_가구_수),
  AVERAGE(seoul.상권특성.아파트_가구_수),
  AVERAGE(seoul.상권특성.총_상주인구_수),
  AVERAGE(seoul.상권특성.남성_상주인구_수),
  AVERAGE(seoul.상권특성.여성_상주인구_수),
  COUNT(*), SUM(kcd.meta-info.is_franchise)
FROM kcd.meta-info
JOIN seoul.상권특성
ON kcd.meta-info.trdar_no = seoul.상권특성.상권_코드
WHERE kcd.meta-info.class_1_name = "외식업"
GROUP BY seoul.상권특성.기준_년_코드, seoul.상권특성.기준_분기_코드, kcd.meta-info.class_2_name
```

- 상권_직장인구

```
SELECT
  AVERAGE(seoul.상권특성.남성_직장_인구_수),
  AVERAGE(seoul.상권특성.여성_직장_인구_수),
  AVERAGE(seoul.상권특성.연령대_10_직장_인구_수),
  AVERAGE(seoul.상권특성.연령대_20_직장_인구_수),
  AVERAGE(seoul.상권특성.연령대_30_직장_인구_수),
  AVERAGE(seoul.상권특성.연령대_50_직장_인구_수),
  AVERAGE(seoul.상권특성.연령대_40_직장_인구_수),
  AVERAGE(seoul.상권특성.연령대_60_이상_직장_인구_수),
  AVERAGE(seoul.상권특성.총_직장_인구_수),
  COUNT(*), SUM(kcd.meta-info.is_franchise)
FROM kcd.meta-info
JOIN seoul.상권특성
ON kcd.meta-info.trdar_no = seoul.상권특성.상권_코드
WHERE kcd.meta-info.class_1_name = "외식업"
GROUP BY seoul.상권특성.기준_년_코드, seoul.상권특성.기준_분기_코드, kcd.meta-info.class_2_name
```

1-3. 이중 데이터 병합

▼ 분석 의도

▼ 결과

- 대회 데이터 기준으로 소호 사업장 및 상권특성 데이터 병합

분기별 상권데이터와 월별 소호 신용데이터를 오아시스비즈니스 행정동코드에 맞추어 분기별로 변환하였습니다. 월별 KCD데이터 상권코드와 오아시스비즈니스 매출등급 예상필지데이터의 행정동코드를 매핑하였습니다. 분기별 상권데이터를 소호 신용데이터와 계절성 지수를 활용하여 월별로 확장하였습니다.

상권 데이터가 계절성 패턴을 보이는지 소호 신용데이터와 상관분석을 수행하여 상관계수가 최소 15% ~ 최대 70% 이내로 통계적으로 상호 연관성이 있다고 해석할 수 있는 변수로 파생한 계절성 지수의 1차원 축소한 주성분을 곱하여 월별로 확장하였습니다.

▼ 시행착오

서울특별시(열린데이터광장), 신용평가회사(KCD), 부동산플랫폼회사(오아시스비즈니스)의 데이터를 LEFT JOIN하기 위한 기본키(Primary Key)를 상권 코드로 둘 경우 ① 출처 · 수집경로가 달라 관측되지 않는 값(NA)이 400개 행 이상 발생하는 문제와 ② 조인연산 수행 후 중복된 행이 발생하는 문제 등에 직면하였습니다.

2가지 문제를 해결하기 위해 더 넓은 범위의 ID인 행정동 코드를 활용하여 병합하였습니다.

ch2. 데이터 분석 모형 구축

▼ 분석 의도

- 2022년의 소호 사업장 매출등급 회귀모형을 추정하여 2023년 1월, 2월의 등급을 예측하기 위해서 시계열을 고려한 2단계 모형을 고안해보고자 하였습니다.
- 매출등급을 예측할 때 필지고유번호에 해당하는 특성 또한 고려해 주고자 하였습니다. 이를 위해 필지고유번호별 매매가 점수 및 유동인구 수 기반의 매출등급 로지스틱 추정확률을 구해주었고 이를 매출등급의 필지고유번호 특성으로 사용해주었습니다.
- 매출등급 예측값에 로지스틱 추정확률을 곱하여 상권별 업종 매출등급 예측 문제에서 필지 단위별 매출등급 예측 문제로 확장을 함으로써 예측의 정확도를 높이고자 하였습니다.

▼ 결과

- 1) 필지고유번호별 '유동인구 수'와 '매매가 점수' 칼럼을 사용해서 필지의 특징을 반영한 23년 1월과 2월의 매출등급 로짓추정확률을 도출하였습니다. 필지고유번호를 기준으로 부동산 데이터와 대회 데이터를 병합하였고 이때 생기는 결측값은 2차 다항회귀를 사용해서 처리해주었습니다. 종속변수는 매출등급, 독립변수를 유동인구수와 매매가 점수로 지정한 후 로지스틱 회귀 추정을 진행하였습니다. 매출등급별로 나온 회귀 추정 값을 주성분 분석을 통해서 1차원 축소 값을 사용해주었습니다.
- 2) 22년 전체 데이터로 RandomForest 기반의 회귀분석 모델링을 진행해서 23년 1월 매출등급을 예측해주었습니다.
- 3) 예측된 23년 1월 매출등급에 23년 1월 매출등급 로짓추정확률을 곱해줌으로써 필지고유번호 특징이 반영된 매출등급을 최종 예측하였습니다.
- 4) 22년 데이터에 예측한 23년 1월 매출등급 데이터를 더한 후, 이를 바탕으로 23년 2월 매출등급을 예측해줍니다.
- 5) 예측된 23년 2월 매출등급에 23년 2월 매출등급 로짓추정확률을 곱해줌으로써 필지고유번호 특징이 반영된 매출등급을 최종 예측하였습니다.

▼ 시행착오

- 필지고유번호 기반의 부동산 입지 특성을 반영하기 위해서 '한국부동산원'의 필지고유번호별 토지이용 데이터를 가져왔었습니다. 하지만 대회 제공 데이터와 한국부동산원의 데이터의 필지고유번호가 매핑되지 않는 문제가 발생하였습니다.

```
[ ] 1 CONTEST_2023.FIELD = pd.merge(
2     left = CONTEST_2023, right = FIELD_DATA_LIST[0],
3     on = ["필지고유번호"],
4     how = "left"
5 )
6 CONTEST_2023.FIELD.isnull().sum()
```

필지고유번호	0
업종코드	0
분양구분코드	0
2023년1월_예측매출등급	10000
2023년2월_예측매출등급	10000
원도지역명	10000
신축구분명	10000
신축건축확장보도분명주소코드	10000
주방속건물구분코드	10000
건물구조명	10000
건물구조명	10000
대지면적	10000
건축면적	10000
연면적	10000
기온일차	10000
dtype:	int64

- 이를 해결하고자 대회에서 지정해준 '오아시스비즈니스 플랫폼'의 '수익형 부동산거래량(금액) 대비 유동인구 수' 데이터와 '수익형 부동산 공실률 대비 매매가 점수' 데이터를 가져온 후 대회 데이터의 필지고유번호와 결합해주었습니다.
- 대회에서 제공해준 데이터는 22년 데이터였으므로 23년 1월 매출등급과 23년 2월 매출등급을 각각 예측하기 위해서 시계열 특성을 어떻게 부여할지 고민해보았습니다.
- 다양한 시도를 해본 후, 22년의 데이터로 23년 1월 매출 등급을 예측하였고, 예측된 1월 매출 등급을 22년 데이터에 포함시켜서 2월 매출등급을 예측해봄으로써 시계열 특성을 고려한 모델링을 진행할 수 있었습니다.
- Oputna, GridSearchCV 등을 사용해서 하이퍼 파라미터 튜닝을 진행해보고 싶었지만 컴퓨팅 자원과 시간 부족 문제로 인해서 튜닝을 진행하지 못했습니다. 향후 좀 더 효율적인 방법을 찾아서 튜닝을 진행할 계획입니다.

ch3. 데이터 시각화

3-1. 예측 매출등급 및 상권특성 시각화

▼ 분석의도

- 카드와 배달 매출액의 변동계수와 매출등급 간의 관계를 비교함으로써, 배달 여부에 따라 매출등급 예측값의 차이가 있는지 알아보고자 하였습니다.
- 객단가가 높을수록 고객의 해당 가게에 대한 접근성이 낮는데 이가 매출액에 어떤 영향을 미치는지 분석해보고자 해당 시각화를 진행하였습니다.
- 손익분기점이 높을수록 더 높은 매출을 달성해야하므로, 결국 손익분기점이 높다는 점이 매출등급에 악영향을 미치지않을까 하는 궁금증에서 이에 대한 가시화를 진행하였습니다.
- 부가세를 통해 영업이익에 대한 현금흐름을 예상해볼 수 있으므로 이를 시각화하여 매출등급간의 관계를 살펴보고자 시각적 분석을 시도하였습니다.

▼ 결과

- 카드 / 배달 매출액 변동계수와 매출등급 간의 관계, 객단가와 매출등급 간의 관계, 손익분기점과 매출등급 간의 관계, 부가세 차감 전 영업이익과 매출등급 간의 관계 등을 라인차트와 막대차트로 시각화하였습니다.
- 배달매출액의 변동계수가 카드매출액의 변동계수에 비해 매출등급예측이 월등히 높은 결과를 보였습니다. 또한 업종(술집, 한식, 양식, 일식)에 따른 차이는 크지 않았습니다.
- 평균객단가변동계수는 매출등급예측값에 따라 큰 차이는 보이지 않았습니다. 즉, 객단가에 따라 매출등급의 유희리가 발생하지는 않는 시각화 결과를 보였습니다.
- 전반적으로 손익분기점이 높을수록 매출액이 악영향을 받는 시각화 결과가 도출되었습니다. 덧붙여 업종에 따라 추가로 구분하였을때, 한식이 다른 업종에 비해 손익분기점이 월등하게 높은 결과를 보인 것이 특징적이었습니다.
- 부가세 차감 이전 이익 변동계수와 매출등급간의 관계를 막대그래프로 시각화한 결과 매출등급에 따른 차이는 크지 않았습니다. 한편, 술집이 다른 업종에 비해 영업이익 변동계수가 높은 것이 특징적이었습니다. 즉, 술집이 부가세 차감 이전 영업이익이 타업종에 비해 높음을 확인할 수 있었습니다.

▼ 시행착오

- 이미 모델링을 위해 정규화를 위한 스케일링을 완료한 데이터이기에 시각화를 진행했을 때 변수간의 차이가 두드러지지 않아 직관적으로 이해가능한 시각화가 어려웠습니다. 이를 해결하기 위해 가장 높은 값을 가진 변수와 가장 낮은 값을 가진 변수들에만 텍스트를 추가하여 가시화 하였습니다.

ch4. 한계와 의의

4-1. 한계

• 모델링 관점

회귀 모델을 선택하는데 근거가 부족했다고 생각합니다. 모델을 선택했던 기준은 기본 모델인 Linear Regression, DecisionTreeRegressor, RandomForestRegressor 모델의 성능을 비교해본 후 가장 성능이 좋았던 RandomForest Regressor모델을 선택하였습니다. 모델의 구조를 이해한 후 본 데이터에 가장 적합한 모델을 선택하는 검증 과정이 추가로 필요하다고 사료됩니다.

Scaler 적용, 모델 Ensemble, Hyper Parameter Tuning 등과 같은 모델의 성능을 높이기 위한 다양한 시도를 해보지 못했던 점이 아쉬웠습니다. 이와 같은 방법론을 적용한다면 성능 개선을 가져올 수 있을 것이라 생각합니다.

• 시각화 관점

필지고유번호를 반영하여 맵차트를 만들고 그 안에 매출액 예측등급을 시각화하려는 목표를 가지고 있었으나, 필지고유번호와 매출등급 및 상권특성을 함께 매핑해둔 데이터프레임의 부재로 실행하지 못한점이 아쉽습니다. 또한, 막대그래프와 라인차트 뿐만 아니라 파이차트, 도넛차트, 맵차트, 히트맵 등 변수의 특성에 따라 다양한 시각화 기법을 적용해보는 것도 필요합니다.

• 도메인 관점

소호 사업장의 입지(세부적인 물건지) 분석에 그치지 않고 광역적인 상권분석을 위한 공간 데이터를 시각화하여 하위 매출등급(평당 고정비용 대비 매출액 백분위수)에 대한 솔루션 제안이 필요합니다.

4-2. 의의

- 첫째, 출처 · 수집경로 · 시계열이 다른 소상공인 신용 데이터 · 상권특성 데이터를 모형에 투입할 수 있도록 병합하는 전처리 과정의 중요성을 배웠습니다.
- 둘째, 최적화된 매출등급 예측모형을 적합하고 시각화하는 것을 넘어 상권 · 입지분석 도메인에 따른 결과 활용 방법 제안의 필요성을 깨달았습니다.

- 셋째, 매출등급예측모형을 통해 해당 상권 및 업종에서 앞으로의 매출 경향성을 미리 파악해 볼 수 있었습니다. 또한, 해당 상권의 입지 데이터 시각화를 통해 매출액 예측등급에 미치는 요인들을 분석하고 활용 할 수 있습니다.