

서울시 상권 데이터를 활용한 창업 추천 서비스

20243053 김찬영, 20243077 황성주

Summary of questions and results

1. 어떤 업종의 매출이 다른 업종의 매출과 연관이 있을까?

어떤 업종의 매출액과 다른 업종의 매출액 간의 상권별 편차를 이용하여 구한 상관계수를 통해 업종들의 매출간 연관성을 분석하였다.

2. 어떤 지역의 특정 업종 매출이 부족하며 어떤 업종을 세워야 매출이 잘 나올까?

상관계수를 통해 구한 가중평균매출액 대비 매출액이 작은 지역은 매출액이 잘 나올 가능성이 있지만 점포의 수가 적다는 뜻이기 때문에 이러한 업종을 추천해 주고자 한다.

3. 매출액 자료를 활용하여 다음 분기 매출액을 예측할 수 있을까?

XGBoost Model을 이용하여 Machine Learning을 통해 매출액 예측을 시도하였지만, 결측치가 나와 결과적으로 실패하였다.

Motivation

업장을 차리기 전 필수적인 상권 조사를 빅데이터 분석을 이용한다면 효율적으로 할 수 있을 것이라는 생각에서 우리 프로젝트는 출발하였다. 특정 업종을 어떤 상권에 세워야 잘 세울 수 있을지, 상권에 영향을 미치는 다른 요인들에는 어떤 것들이 있는지 등을 알려주면 좋을 것 같았다.

Dataset

서울 열린 데이터 광장의 '서울시 상권분석서비스(추정매출-상권배후지)', '서울시 상권분석서비스(직장인구-상권배후지)', '서울시 상권분석서비스(길단위인구-상권배후지)'와 '서울시 상권분석서비스(점포-상권배후지)'를 사용하였다.

<http://data.seoul.go.kr/dataList/OA-15573/S/1/datasetView.do>

<http://data.seoul.go.kr/dataList/OA-15570/S/1/datasetView.do>

<http://data.seoul.go.kr/dataList/OA-15582/S/1/datasetView.do>

<http://data.seoul.go.kr/dataList/OA-15578/S/1/datasetView.do>

Method

1. 데이터 전 처리 (파편적 datasets merge, cleaning etc)
2. 상권 별 업종들의 매출 상관관계수 도출
3. 상관계수를 이용하여 특정 업종의 상권 별 가중평균매출액 구하기
4. 상권 별 가중평균매출액 대비 특정 업종의 매출액으로 특정 업종의 창업 추천 상권 순위 매기기
5. 유동/직장인구의 평균을 구하고 그에 따른 현재 유동/직장인구의 비율 나타내기
6. 매출액 데이터를 이용하여 상권별로 다음 분기의 매출액 예측하기 (ML)
7. 최종적으로 모든 데이터를 합쳐 결과를 하나의 Dataframe으로 나타내기

Results

1. 모든 업종의 매출액에 대한 상관계수를 구함으로써 어떤 업종의 매출과 다른 업종의 매출 사이의 상관관계를 확인할 수 있었다. (업종 A와 업종 B의 매출 상관관계)

대표적으로 한식음식점과 호프간이주점, 노래방과 한식음식점, 분식전문점과 한식음식점 등이 0.6 이상의 높은 상관계수를 보였다. 이는 한 업종의 매출이 높을수록 다른 업종의 매출이 높을 가능성에 대한 수치이다.

서비스_업종_코드_명	노래방	미용실	반찬가게	동산중개	분식전문점	슈퍼마켓	영식음식점	예술학원	육류판매	의약품	반교습학	일반의류	일반의원	자상거래	커피-음료	및주변장	피부관리	한식음식점	프-간이주	화장품
노래방	1	0.35	0.1	0.07	0.41	0.36	0.03	0.14	0.04	0.3	0.03	0.13	0.36	0.03	0.21	-0.02	0.23	0.66	0.56	0.28
미용실	0.35	1	-0.01	0.06	0.54	0.28	0.34	0.34	0	0.18	0.27	0.37	0.5	0.05	0.42	-0.03	0.5	0.48	0.41	0.33
반찬가게	0.1	-0.01	1	0	0.06	0.06	-0.04	0	0.01	0.2	0	-0.02	0.03	0.08	-0.02	0.03	-0.02	0.08	0.03	-0.03
부동산중개업	0.07	0.06	0	1	0.06	-0.02	-0.01	0.14	0.02	0	0.05	-0.01	0.03	0	0.07	-0.01	0.14	0.08	0.08	0.01
분식전문점	0.41	0.54	0.06	0.06	1	0.21	0.29	0.28	-0.01	0.25	0.25	0.24	0.45	0.11	0.55	0.03	0.42	0.61	0.54	0.34
슈퍼마켓	0.36	0.28	0.06	-0.02	0.21	1	0.03	0.15	0.04	0.23	0.02	0.09	0.26	0.03	0.06	-0.01	0.07	0.29	0.25	0.12
영식음식점	0.03	0.34	-0.04	-0.01	0.29	0.03	1	0.1	-0.02	0	0.04	0.34	0.22	0.1	0.52	-0.01	0.32	0.33	0.49	0.26
예술학원	0.14	0.34	0	0.14	0.28	0.15	0.1	1	-0.02	0.09	0.28	0.07	0.21	0.11	0.26	0.01	0.32	0.27	0.2	0.13
육류판매	0.04	0	0.01	0.02	-0.01	0.04	-0.02	-0.02	1	0.01	-0.01	-0.01	0	-0.02	-0.04	-0.01	-0.01	0.02	0.01	-0.02
의약품	0.3	0.18	0.2	0	0.25	0.23	0	0.09	0.01	1	0.17	0.07	0.28	0.04	0.12	0	0.1	0.31	0.2	0.18
일반교습학원	0.03	0.27	0	0.05	0.25	0.02	0.04	0.28	-0.01	0.17	1	0.01	0.08	0	0.12	-0.01	0.1	0.13	0.02	0.04
일반의류	0.13	0.37	-0.02	-0.01	0.24	0.09	0.34	0.07	-0.01	0.07	0.01	1	0.25	0.18	0.26	0.03	0.25	0.27	0.23	0.26
일반의원	0.36	0.5	0.03	0.03	0.45	0.26	0.22	0.21	0	0.28	0.08	0.25	1	0.02	0.36	-0.03	0.47	0.43	0.37	0.25
전자상거래업	0.03	0.05	0.08	0	0.11	0.03	0.1	0.11	-0.02	0.04	0	0.18	0.02	1	0.18	0.13	0.08	0.15	0.12	0.24
커피-음료	0.21	0.42	-0.02	0.07	0.55	0.06	0.52	0.26	-0.04	0.12	0.12	0.26	0.36	0.18	1	0.01	0.41	0.59	0.47	0.32
컴퓨터및주변장치판매	-0.02	-0.03	0.03	-0.01	0.03	-0.01	-0.01	0.01	-0.01	0	-0.01	0.03	-0.03	0.13	0.01	1	-0.01	0.04	0.01	0.06
피부관리	0.23	0.5	-0.02	0.14	0.42	0.07	0.32	0.32	-0.01	0.1	0.1	0.25	0.47	0.08	0.41	-0.01	1	0.44	0.36	0.33
한식음식점	0.66	0.48	0.08	0.08	0.61	0.29	0.33	0.27	0.02	0.31	0.13	0.27	0.43	0.15	0.59	0.04	0.44	1	0.67	0.42
호프-간이주점	0.56	0.41	0.03	0.08	0.54	0.25	0.49	0.2	0.01	0.2	0.02	0.23	0.37	0.12	0.47	0.01	0.36	0.67	1	0.31
화장품	0.28	0.33	-0.03	0.01	0.34	0.12	0.26	0.13	-0.02	0.18	0.04	0.26	0.25	0.24	0.32	0.06	0.33	0.42	0.31	1

[그림 1] 업종간 상관계수 테이블

2. 특정 업종의 가중평균매출액을 구함으로써 창업 추천 상권 추천을 할 수 있었다.

대표적으로 '한식음식점'과 상관계수가 가장 높게 나왔던 '호프-간이주점' 업종의 경우 '가중평균매출액 대비 호프-간이주점 매출 비율' 컬럼을 확인해 보았을 때, 이 수치가 작다는 것은 가중평균 매출액에 비하여 호프-간이주점 업종의 매출이 해당 상권에서는 미진하다는 것을 나타낸다.따라서 이 수치를 reverse=True로 sort 해 보았을 때, 이 수치가 가장 적은 상권들인 마천1치안센터, 남가좌동현대아파트, 구로창업지원센터, 강남세브란스병원미래의학연구센터, 언주역 8번에서 창업을 하였을 때 높은 수익을 얻을 가능성이 다른 상권에 비해 높다는 연구 결과가 나왔다.

W1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																							
1	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마	바	사	아	야	가	나	다	라	마

3. 평균 유동/직장인구 수 대비 특정 상권의 유동/직장인구 수를 구하여 유동/직장인구의 비율을 나타낼 수 있었고, 이를 추천 상권 데이터와 합쳐 나타낼 수 있었다.

가중평균매출액 대비 호프-간이주점 매출 비율										
	A	B	C	D	E	F	G	H	I	J
1	Unnamed: 배후지_코드_명,기준_년분기_코드,당월_매출_금액,다음_분기_총_매출_금액,다음분기_예측_매출_금액									
2	16698	16698	남가좌동호프-간이	574905	74	1573985	0.0004		0.041	0.4121
3	23757	23757	구로창업지호프-간이	2152351	1045	2923707	0.0005		0.5786	0.7656
4	32258	32258	연주역 8번호프-간이	11305884	1199	3640377	0.0014		0.6639	0.9532
5	32869	32869	강남세브란호프-간이	16080446	2628	2836067	0.0016		1.4551	0.7426
6	35449	35449	마천1치안	983146	111	2675579	0.0018		0.0615	0.7006
7										

[그림 3] '호프-간이주점' 업종의 창업 추천 상권들의 평균 대비 직장/유동인구 비율

4. XGBoost Model을 이용하여 2023년1분기 데이터 학습을 통해 상권 별 2023년 2분기와 2023년 3분기의 매출액을 구했고, 예측의 정확도를 높이기 위해 hyper parameter를 튜닝해보는 등 여러가지 시도를 해보았지만 예측 매출이 음수가 나오는 등 예상 밖의 결과치가 나와 결과적으로 실패하였다.

```

상권배후지_코드_명,기준_년분기_코드,당월_매출_금액,다음_분기_총_매출_금액,다음분기_예측_매출_금액
강남세브란스병원미래의학연구센터,20231,16080446,10051376.0,1327028100.0
강남세브란스병원미래의학연구센터,20232,10051376,13681856.0,-301087900.0
강남세브란스병원미래의학연구센터,20233,13681856,25655242.0,-198658740.0
구로창업지원센터,20231,2152351,4230939.0,1327028100.0
구로창업지원센터,20232,4230939,7721042.0,-301087900.0
구로창업지원센터,20233,7721042,6951971.0,-198658740.0
마천1치안센터,20231,983146,31264045.0,1327028100.0
마천1치안센터,20232,31264045,6292135.0,-301087900.0
마천1치안센터,20233,6292135,78946629.0,-198658740.0
연주역 8번,20231,11305884,13598748.0,1327028100.0
연주역 8번,20232,13598748,14202066.0,-301087900.0
연주역 8번,20233,14202066,14666385.0,-198658740.0

```

[그림 4] 다음 분기 매출액을 예측한 데이터

5. 최종적으로 이 데이터들을 모두 하나의 Dataframe으로 나타내었다.

M6	-301087904													
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Unnamed: 배후지_코스_업종_코스_매출_금액	직장_인구	유동인구	비	호프_건	비	직장인	비	유동인	년분기_구	매출_금액	기_총_매	기_예측_매출_금액	
2	23757	구로창업지_호프-간이	2152351	1045	2923707	0.0005	0.5786	0.7656	20231	2152351	4230939	1.33E+09		
3	23757	구로창업지_호프-간이	2152351	1045	2923707	0.0005	0.5786	0.7656	20232	4230939	7721042	-3E+08		
4	23757	구로창업지_호프-간이	2152351	1045	2923707	0.0005	0.5786	0.7656	20233	7721042	6951971	-2E+08		
5	32258	언주역 8번_호프-간이	11305884	1199	3640377	0.0014	0.6639	0.9532	20231	11305884	13598748	1.33E+09		
6	32258	언주역 8번_호프-간이	11305884	1199	3640377	0.0014	0.6639	0.9532	20232	13598748	14202066	-3E+08		
7	32258	언주역 8번_호프-간이	11305884	1199	3640377	0.0014	0.6639	0.9532	20233	14202066	14666385	-2E+08		
8	32869	강남세브란_호프-간이	16080446	2628	2836067	0.0016	1.4551	0.7426	20231	16080446	10051376	1.33E+09		
9	32869	강남세브란_호프-간이	16080446	2628	2836067	0.0016	1.4551	0.7426	20232	10051376	13681856	-3E+08		
10	32869	강남세브란_호프-간이	16080446	2628	2836067	0.0016	1.4551	0.7426	20233	13681856	25655242	-2E+08		
11	35449	마천1치안_호프-간이	983146	111	2675579	0.0018	0.0615	0.7006	20231	983146	31264045	1.33E+09		
12	35449	마천1치안_호프-간이	983146	111	2675579	0.0018	0.0615	0.7006	20232	31264045	6292135	-3E+08		
13	35449	마천1치안_호프-간이	983146	111	2675579	0.0018	0.0615	0.7006	20233	6292135	78946629	-2E+08		

[그림 5] 최종 데이터

Impact and Limitations

특정 업종이 창업하기 좋은 상권을 빅데이터 분석을 통해 알아볼 수 있다. 창업을 위해 이곳저곳 돌아다녀 보면서 상권 분석을 하기보단, 빅데이터 분석을 통해 창업 결정을 보다 효율적으로 해보자는 아이디어에서 출발한 프로젝트였다.

결과적으로 client에게는 다음과 같은 정보를 제공하고 있다.

1. 어떠한 특정 업종과의 매출 상관관계가 높은 업종을 select하여 그를 바탕으로 한 특정 상권의 가중평균매출액 대비 특정 업종의 매출을 확인하여 가중평균매출액 대비 특정 업종의 매출이 미진한 곳을 추천한다.
2. 매출적 데이터 뿐만 아니라 서울시 상권 평균 길단위 유동인구 대비 특정 상권의 유동인구 비율과 평균 직장인구 대비 특정 상권의 직장인구 비율, 해당 업종의 다음 분기 매출 예측액을 제시하였다. 이 데이터들은 client가 창업을 결정할 때 매우 도움이 될 수 치들이다. client에게 각기 다른 타입의 4가지 수치를 제공함으로써 보다 창업 결정에 도움이 되는 수치들을 바탕으로 성공률이 높은 창업 추천 서비스를 구현하였다.

하지만 모든 업종에 대한 일반적인 데이터를 얻지 못했다는 한계가 있다.

(2023년 1분기로 제한, 모든 업종에 대한 데이터 분석을 하지 못하고 호프-간이주점에 대한 분석만 진행함)

또한 상권에 특정 업종이 하나도 존재하지 않을 때의 데이터는 분석에서 제외하였고, 점포수가 10000개 미만인 업종은 분석에서 제외하였기 때문에, 완전히 새롭고, 신선한 업종의 창업을 제안할 수 없다는 단점이 있다.

Challenge Goals

1. Multiple Dataset을 활용한 데이터 분석
2. Machine Learning (다음 분기 매출액 예측)
3. New Library(XGBoost ML library를 활용한 학습, openpyxl을 활용한 xlsx 파일 writing)

초기에는 추정매출액-상권배후지 데이터만 가지고 분석을 진행할 생각이었기 때문에 pre-report에는 Multiple dataset을 기재하지 않았지만, 분석에 사용할 데이터를 더 끌어오는 과정에서 총 4개의 데이터를 사용하게 되었고 그 결과 Challenge Goals에 Multiple Dataset을 추가하게 되었다.

Work Plan Evaluation

1. 데이터 전처리를 통해 일정 점포 수 이상의 업종만 남기고, 분석에 필요한 column만 추출 (2 hours)

-> 예상했던 것 보다 필요했던 dataset이 많아 merge 및 cleaning 작업에서 꽤나 시간을 많이 썼다. (약 10시간 이상)

2. 업종별 매출 상관관계 분석을 위해 상관계수 를 구한다. 일정 이상의 상관계수를 가진 업종쌍 을 추출 (4 hours)

-> 이는 numpy 내장 함수를 사용할 수 있었기에, 시간이 단축되었다. (2 hours)

3. 상관계수를 가중치로 활용하여 업종 당 상권 별 가중평균매출액을 구하고 가중평균매출액 대 비 상권 별 특정 업종의 매출 비율을 구한다. (3 hours)

-> 이도 무리없이 잘 진행되었다. (3 hours)

4. 가중평균매출액 대비 상권 별 특정 업종의 매출 비율이 가장 낮은 상권을 추천하여 최종적으로 창업 추천. (1 hours)

-> 최종적으로 창업 추천이라기 보단, 이를 창업 추천에 활용하였다. (2 hours)

5. 머신러닝을 위한 데이터 전처리 (Decision Tree)를 통해 Feature selection + 인구밀도 속성 사용 여부 결정 + 6 . 창업 위험지수를 머신러닝으로 예측하기. (3 hours)

-> 프로젝트 방향이 틀어져 매출을 예측하는 방향으로 진행했기 때문에 Feature Selection은 진행하지 못했다. 인구 밀도 데이터도 사용하기엔 프로젝트가 지나치게 heavy해질 것 같아 사용하지 않았다.

7. 문제가 생겼을 때 : 구글을 찾아보고, 선배님들께 여쭙보고, 최종적으로 교수님께 질문한다. (INF hours)

-> 구글을 많이 참고했다. (?? Hours)

이번 프로젝트 work plan 그대로 구현하기에는 우리의 실력이 약간 부족하기도 했고, 프로젝트 볼륨이 너무 커질 것 같아 프로젝트 진행 중 머신러닝 파트 계획을 단순 매출 예측으로 수정했다. 또한 머신러닝으로 예측하는 파트를 단순 참고수치로 추가시킴으로써 우리가 단방향적으로 client에게 창업을 제시해주는 그림보단, 우리가 창업 상권을 추천해주고 그 상권의 추가적 데이터를 제시해 client가 직접 판단을 내릴 수 있는 유의미한 데이터들을 client에게 제시하는 방향으로 project를 진행하였다.

가중평균매출액을 구하는 과정까지는 예측과 맞게 데이터가 구해졌지만, 머신 러닝을 하는 과정에서 feature가 부족하다 보니 예측치가 벗어난 것 같다.

위의 내용들은 분명히 아쉬운 점들이었지만, 가중평균매출액을 바탕으로 수치적 해석을 통해 창업 추천을 했다는 점은 분명히 뜻깊은 성과임이 틀림 없다.

Testing

특정 업종의 모든 상권에 대한 가중평균매출액을 구하였으나, 20개의 업종을 모두 작업하기에는 상관관계가 특히 유의미하다고 생각이 드는 상권이 몇 개 없다고 판단을 했다, (심지어는 상관계수가 음수인 업종 쌍도 있었다.) 이에 따라 우리는 한 개의 업종에 대해서만 가중평균 매출액 도출 작업을 진행하였다. 각각 가중평균매출액을 구하는 원리는 똑같기 때문에, 전체 다 구하는 것 보단 호프-간이주점을 대표 업종으로 선정하여 깔끔하게 나타내는 작업을 진행하였다. 상권이 너무 많기도 했고, 상관계수도 유의미한 쌍이 몇 개 없어서 모든 업종에 대한 객관적인 Testing을 진행하는 것은 시간적으로도, 프로젝트 결과적으로도 좋지 않다는 판단 하에 특정 업종에 대해서라도 적어도 이것 만큼은 깔끔하게 만들어보자 라는 생각으로 작업을 진행하였다. 실제로 계산기를 통해 호프-간이주점의 가중평균매출액을 계산해보았을 때, 얻은 결과값이 같다는 것을 확인함으로써 우리의 valid 테스트를 마쳤다. 유동인구 비율과 직장인구 비율 또한 마찬가지로였다.

Machine Learning 작업을 할 때에는 매출액 예측 값에 음수가 나오는 등 결측치가 나왔다.

이를 해결하기 위해 hyper parameter를 튜닝하는 등 갖가지 노력들을 해보았지만, 근본적으로 우리가 생각했던 것보다 feature가 훨씬 적었던 것이 원인이 되었다고 추측하고 있다.

Collaboration

<https://github.com/JSSJS/Capstone-Design-Analysis-of-trade-area>