

PORTFOLIO

데이터 분석가

지원자 곽종현

Blog : <https://duckkkk.com>

Github : <https://github.com/duck9667>

목차

01 데이터 수집 및 파싱

- 삼성헬스 데이터 수집 및 파싱(더비타 인턴쉽)

02 탐색적 데이터 분석

- 삼성헬스 데이터 분석(더비타 인턴쉽)

03 데이터 시각화

- 서울시 약취민원 분석 대시보드 구축(협업 프로젝트)

04 데이터 분석

- 건강검진 데이터 분석(더비타 인턴쉽)

- 쇼핑 구매데이터 분석(공모전)

- 동작구 입지 분석(공모전)

05 부록

- 롤 게임데이터 분석(개인 프로젝트)

- 카카오톡 채팅분석(개인 프로젝트)

01

데이터 수집 및 파싱 I 더비타 인턴쉽

삼성 헬스 데이터 수집 및 파싱

프로젝트 목적

삼성 웨어러블 기기를 통해 수집가능한 라이프로그(Life log) 데이터의 종류와 특징을 파악하고 데이터 분석을 통해 개인 맞춤형 건강 관리 솔루션 서비스 기획

맡은 업무

데이터 수집 업무 : 갤럭시 워치 사용자로부터 삼성헬스 앱을 통해 48건의 데이터 수집

데이터 파싱 업무 : 데이터 파싱 및 파서(Parser) 프로그램 제작

데이터 분포 파악 업무 : 건강과 관련된 핵심 변수 3가지 걸음, 심박수, 수면을 대상으로 분석진행

프로젝트 기간

2020년 10월 (약 2주)

데이터

데이터 : 48건

변수 : 78개

제작 환경

PyCharm



01

데이터 수집 및 파싱 | 더비타 인턴쉽

삼성 헬스 데이터 수집 및 파싱

데이터 수집 및 파싱

수집 대상자

삼성 웨어러블 기기 사용자

수집 방법

삼성헬스 어플에 축적된 데이터를

대상자로부터 직접 수령

제작 내용

78개의 데이터별 클래스 생성

해당 데이터의 칼럼을 매소드로 하는

파서(Parser) 제작

기타

클라우드나 데이터 저장소와 같이

데이터가 한곳에 모여있지 않아 일일이

수집하는 과정에서 데이터 확보의

중요성을 깨달음

이름	유형	크기
samsunghealth_1956(F)	파일 폴더	
samsunghealth_1963(M)	파일 폴더	
samsunghealth_1965(F)	파일 폴더	

	A	B	C	D	E	F	G	H	I
1	com.sams	6111001	1						
2	step_coun	active_tim	others_tim	update_tir	create_tim	goal	longest_ac	score	distance
3	5306	3370873	0	50:35.1	35:44.5	60	180000	91	4223.31
4	0	0	0	35:45.1	35:45.1	-1	0	0	0
5	0	0	0	35:45.5	35:45.5	-1	0	0	0
6	0	0	0	35:45.9	35:45.9	-1	0	0	0
7	0	0	0	35:46.5	35:46.5	-1	0	0	0
8	10024	5888230	0	53:42.7	00:00.3	60	1069435	161	7763.677
9	5796	3654006	0	17:09.1	00:00.4	60	300000	100	4491.879
10	0	0	0	35:44.8	35:44.8	-1	0	0	0
11	0	0	0	35:45.2	35:45.2	-1	0	0	0
12	0	0	0	35:44.9	35:44.9	-1	0	0	0
13	0	0	0	35:45.0	35:45.0	-1	0	0	0
14	0	0	0	35:45.2	35:45.2	-1	0	0	0
15	0	0	0	35:45.3	35:45.3	-1	0	0	0
16	0	0	0	35:45.8	35:45.8	-1	0	0	0
17	0	0	0	35:45.7	35:45.7	-1	0	0	0
18	0	0	0	35:44.6	35:44.6	-1	0	0	0
19	0	0	0	35:45.6	35:45.6	-1	0	0	0
20	0	0	0	35:46.1	35:46.1	-1	0	0	0
21	0	0	0	35:46.2	35:46.2	-1	0	0	0
22	7050	4296692	0	48:02.6	56:59.3	60	360000	118	5429.439
23	1515	1001931	0	06:18.1	00:00.8	60	60000	26	1202.18

이름
com.samsung.health.caffeine_intake.202008171151
com.samsung.health.device_profile.202008171151
com.samsung.health.floors_climbed.202008171151

-----	Activity_day_summary Daily Object
Day_Time:	2019-11-21 09:00:00
Goal:	60
Update_Time:	2019-11-21 15:36:51.474000
Pkg_Name:	com.sec.android.app.shealth
Walk_Time:	4680540
Longest_Idle_Time:	23940000
Active_Time:	4743085
Longest_Active_Time:	180000
Calorie:	272.82907
Deviceuuid:	h+LaCemfvc
Run_Time:	62545
Step_Count:	7361
-----	Extra Data Object -----
mActivityList:	None
mAdaptiveGoal:	0
mIsGoalAchieved:	True
mIsMostActiveAchieved:	False
mMostActiveMinutes:	361
mStreakDayCount:	0

02

탐색적 데이터 분석 | 더비타 인턴쉽

삼성 헬스 데이터 분석

프로젝트 목적

삼성 헬스 데이터 분포 파악

프로젝트 기간

2020년 10월 (약 2주)

맡은 업무

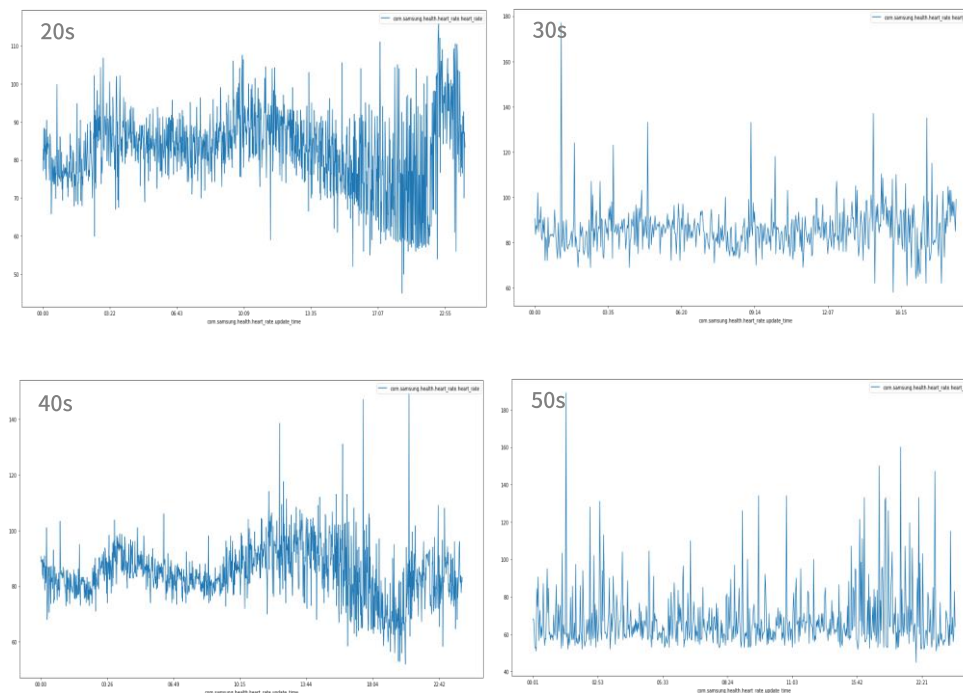
핵심 건강 데이터인 걸음, 심박수, 수면데이터를 기반으로 탐색적 데이터 분석 실시

기타

데이터의 양이 적고 비연속적인 데이터로 인해 분포를 파악하기 어려웠으나 양질의 데이터와 연속적인 데이터의 중요성을 깨달음

심박수

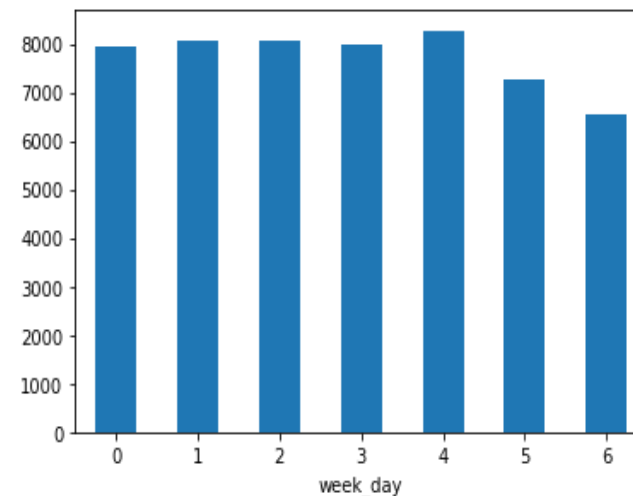
연령별·시간대별 평균 심박수 비교



걸음

성별·요일별 평균 걸음수 비교

	Sex	Count	distance	day_time	step_daily_mean
0	F	5164023.86	3755810.90	729.29	6914.36
1	M	6473699.91	4997627.51	830.67	7880.58



03

데이터 시각화 I 중앙대학교 환경시스템연구실 협업 프로젝트 서울시 악취민원 분석 대시보드 구축

프로젝트 목적

한국환경산업기술원의 지원을 받아 진행한 프로젝트로 서울시 악취민원의 효율적인 종합상태 평가 방안 모색

프로젝트 내용

행정동별 인구와 면적 그리고 하수관의 길이에 따른 악취민원 기술 통계·시각화를 웹 대시보드로 구축

맡은 업무

데이터 전처리·기초 통계·데이터 시각화

데이터

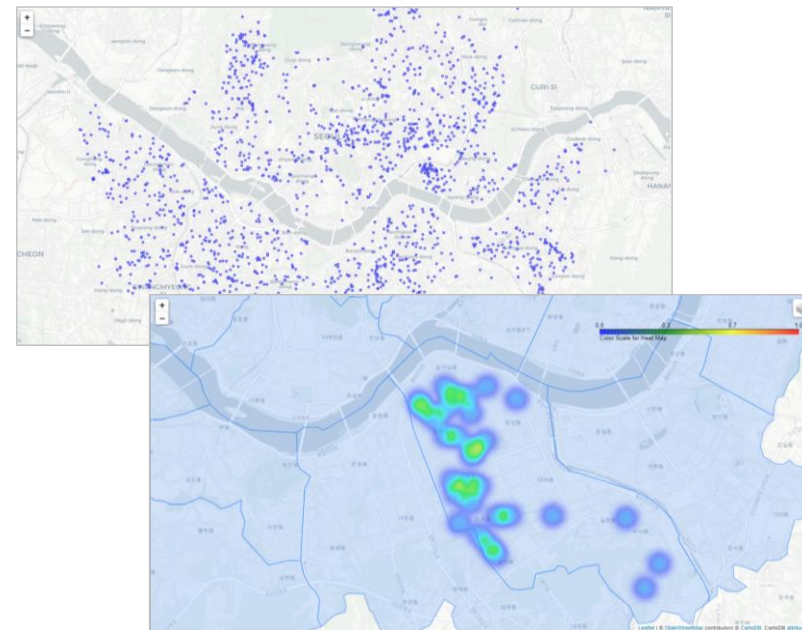
구성 : 2014년 3월 부터 2020년 5월 악취민원 8,636건 및 서울시 행정동별 인구,면적, 하수관 연장길이 데이터

프로젝트 기간

2020년 12월 (약 1개월)

제작 환경

구름 IDE



중앙대학교

03

데이터 시각화 I 중앙대학교 환경시스템연구실 협업 프로젝트 서울시 악취민원 분석 대시보드 구축

데이터 전처리

카카오 지도 API를 활용하여 위경도에 맞는 행정동 추출

C	D	E	F	G	H
유형_소	신청일시	경도	위도	상태	행정동
도봉구	2014-03-05	127.04183	37.668629	정좌표	방학1동
성북구	2014-03-06	127.04305	37.600927	정좌표	월곡2동
동대문구	2014-03-07	127.04239	37.576022	정좌표	용신동

기초통계 분석

통계 패키지 Pandas를 활용하여 기초 통계 산출

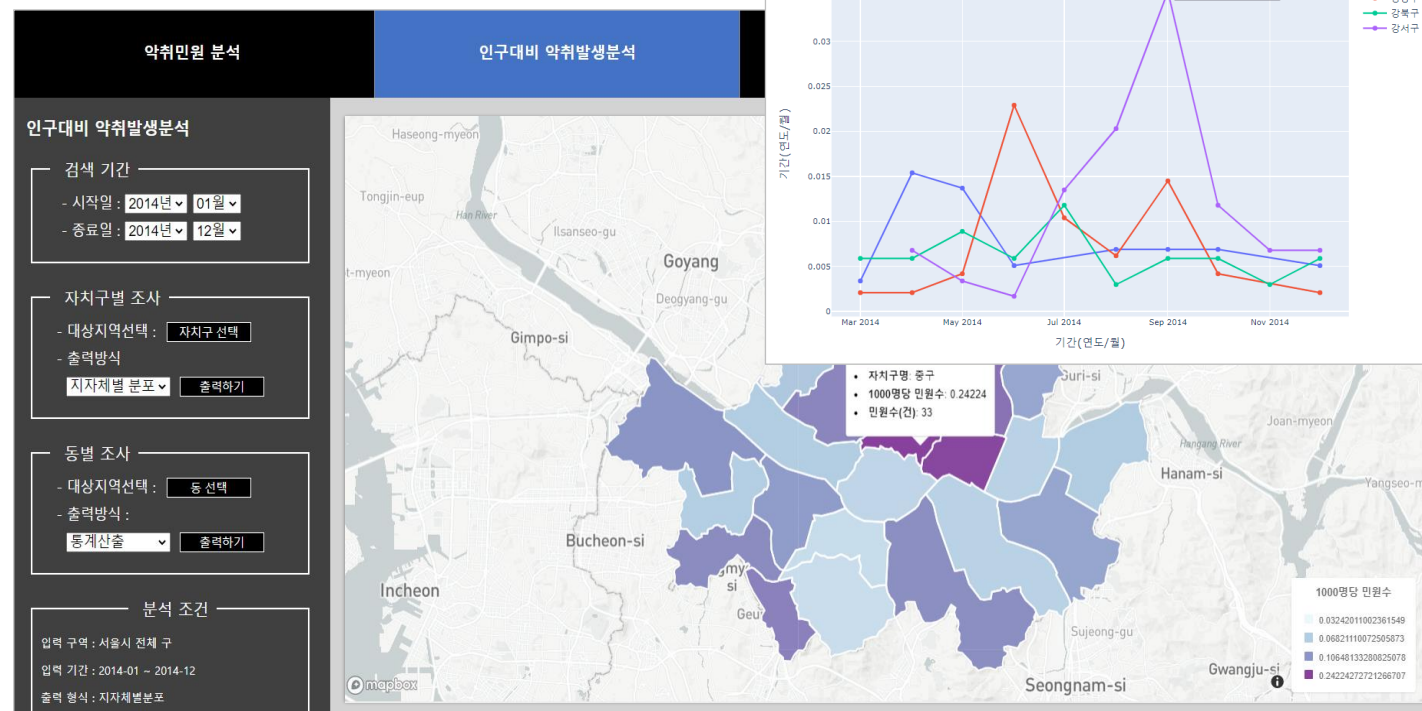
	연월	잠원동	양재1동	내곡동	양재2동
0	2014-03	0.0291	0.0000	0.0000	0.0417
6	2014-04	0.0000	0.0000	0.0000	0.0417

기타

이상치, 중복값 등을 제거하고 오류를 해결하는데 많은 시간이 소요되어 분석 시 사전에 데이터 정합성을 파악하는 것의 중요성을 깨달음

데이터 시각화

1. 동적 시각화 패키지인 “Plotly”를 활용하여 그래프 시각화
2. 지리정보 시각화 패키지인 “Folium”을 활용하여 지도 시각화



03

데이터 분석 I 더비타 인턴쉽 건강검진 데이터 분석

프로젝트 목적

건강보험공단에서 제공하는 건강검진 데이터(공공데이터)를 활용하여 분석을 실시
유의미한 분석결과를 토대로 건강 리포트 서비스 기획에 실마리를 찾고자 함

맡은 업무

데이터 분석, 데이터 분류 성능 평가

데이터

구성 : 총 데이터로 34개의 변수로 구성 (출처-건강보험공단)

샘플링 : 총 100만개의 데이터에서 랜덤 샘플링을 통해 10만개 데이터 선택

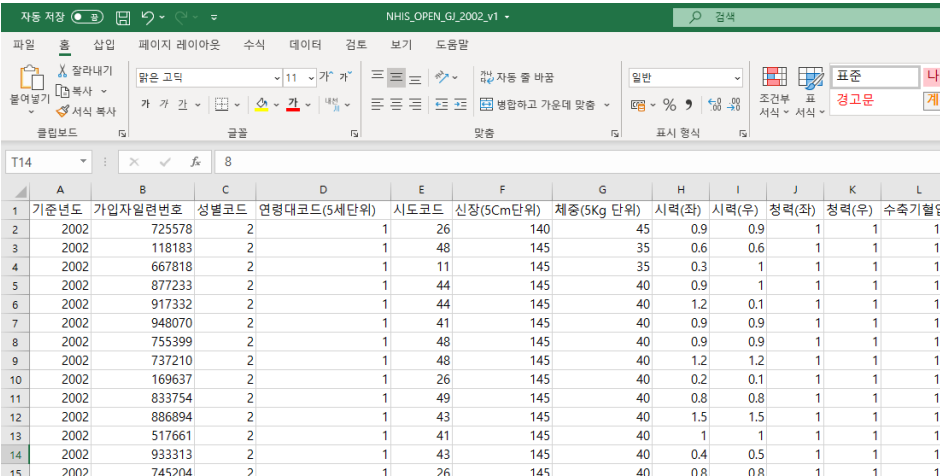
기준년 : 2002년 ~ 2015년 (14개년)동안 “일반 건강검진”, “생애전환기 건강진단”의 결과

프로젝트 기간

2020년 11월 (약 1주)

제작 환경

Jupyter Lab



	A	B	C	D	E	F	G	H	I	J	K	L
1	기준년도	가입자일련번호	성별코드	연령대코드(5세단위)	시도코드	신장(5Cm단위)	체중(5Kg 단위)	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압
2	2002	725578	2	1	26	140	45	0.9	0.9	1	1	100
3	2002	118183	2	1	48	145	35	0.6	0.6	1	1	120
4	2002	667818	2	1	11	145	35	0.3	1	1	1	111
5	2002	877233	2	1	44	145	40	0.9	1	1	1	110
6	2002	917332	2	1	44	145	40	1.2	0.1	1	1	110
7	2002	948070	2	1	41	145	40	0.9	0.9	1	1	110
8	2002	755399	2	1	48	145	40	0.9	0.9	1	1	150
9	2002	737210	2	1	48	145	40	1.2	1.2	1	1	110
10	2002	169637	2	1	26	145	40	0.2	0.1	1	1	120
11	2002	833754	2	1	49	145	40	0.8	0.8	1	1	130
12	2002	886894	2	1	43	145	40	1.5	1.5	1	1	120
13	2002	517661	2	1	41	145	40	1	1	1	1	120
14	2002	933313	2	1	43	145	40	0.4	0.5	1	1	100
15	2002	745204	2	1	26	145	40	0.8	0.8	1	1	100



03

데이터 분석 I 더비타 인턴쉽 건강검진 데이터 분석

분석 개요

데이터 분석 시각화

- 데이터의 복잡성을 확인하기 위해 근접도 그래프(Proximity Graph) 선택
- 특징 중요도가 높은 변수들을 차례대로 제외시킴으로써 분석 실시
- 특징 중요도를 위해 Decision tree, Random forest 등의 알고리즘 활용

분류 성능 평가

- 성능 측정 방법으로 Accuracy, f1-score 활용

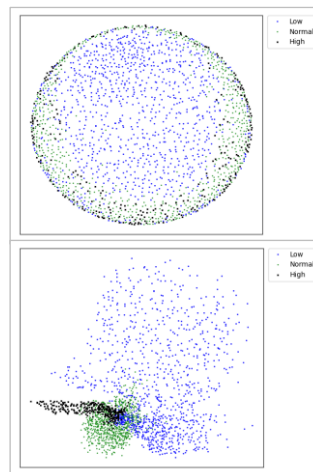
분석 효과

- 선택한 Class와 밀접한 상관관계가 있는 변수 추출
- 건강검진정보 활용방안 마련
- 개인 or 군집별 분류 모델 생성시, 의미 없는 변수 제거

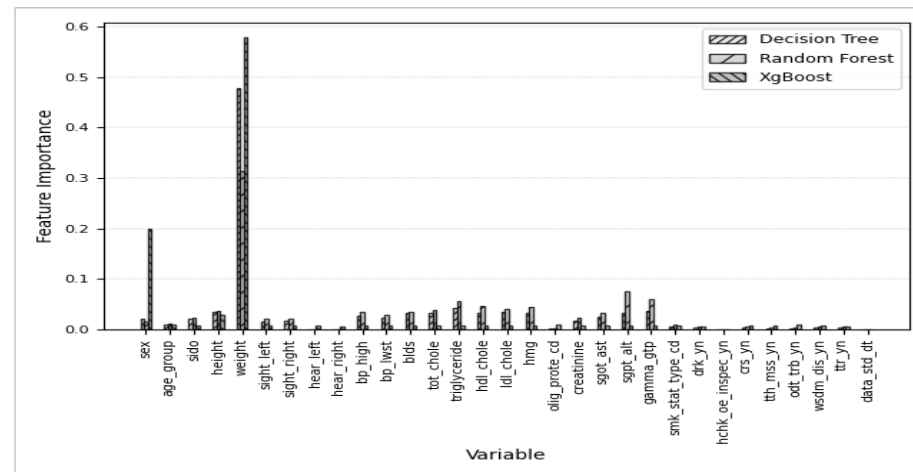
허리둘레 변수를 기준으로 실시한 분석 결과

- 성별에 따라 허리둘레를 3가지 클래스 설정
- 변수 중 기준년도, 가입자 일련번호 제외
- 체중, 혈청지오티 ALT, 혈청지오티 AST 등이 높은 특징 중요도로 분석

Proximity Graph



Feature Importance



데이터 개요

Algorithm	Test Acc.
Low (남: 80미만, 여: 75미만)	51,793
Normal (남: 80~90 / 여: 75~85)	34,412
High (남: 90이상, 여: 85이상)	13,795

분류성능 평가

Algorithm	Test Acc.	Test f1-score
Decision tree	68.00	64.76
Random forest	76.93	73.57
XGBoost	76.52	73.54

04

데이터 분석 I 공모전

유한킴벌리 쇼핑몰 구매 데이터 분석

프로젝트 목적

생활용품 전문 기업 유한킴벌리 자사 쇼핑몰 momQ의 고객 구매 데이터를 기반으로 고객을 세분화하고 이탈가능성이 높은 고객을 충성도 높은 고객으로 전환할 전략 수립

맡은 업무(참여도%)

탐색적 데이터 분석 : “Tableau”를 활용한 데이터 시각화

클러스터링 : 고객 생애가치를 기반으로한 클러스터링

프로젝트 기간

2019년 4월부터 6월(약 2개월)

데이터

기준 년 : 2018년

회원 수 : 115,386 명

레코드 수 : 751,419 건

제작 환경

Python R

Tableau Excel



04

데이터 분석 I 공모전

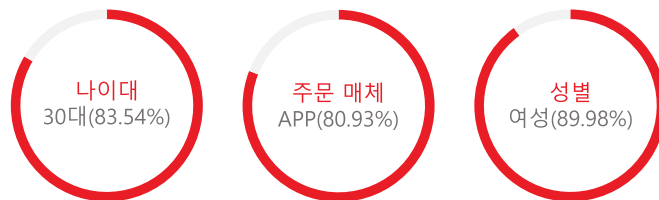
유한킴벌리 쇼핑몰 구매 데이터 분석

데이터 시각화

탐색적 데이터 분석

- 구매 데이터를 기반으로 고객 페르소나 유추
- 코호트 분석을 통해 구매주기 분석
- 구매 고객과 사용자의 차이에 따른 특성 파악
- 제품 카테고리별 수요 파악

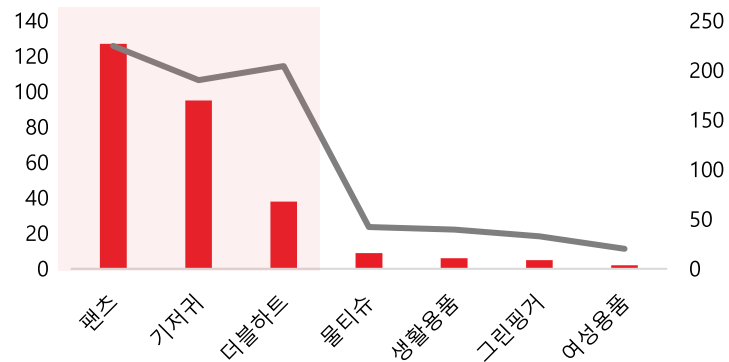
평균 고객은 어떤 사람일까?
페르소나 키워드 : 30대, 여성, APP



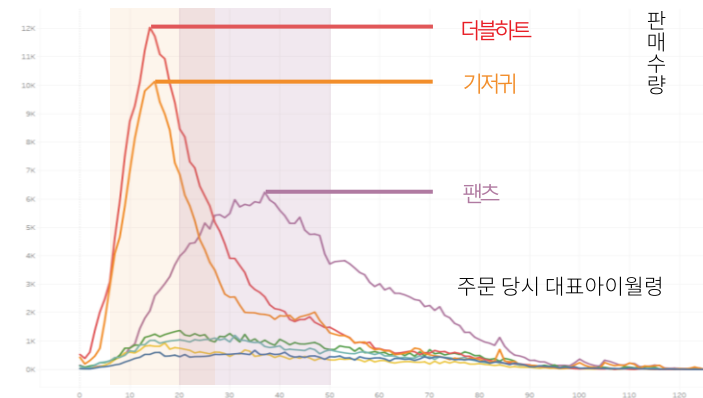
Cohort Analysis [구매주기 분석]
= 3개월 내 재구매가 가장 많이 발생



평균 고객에게 인기 있는 제품 카테고리
판매액 기준 / 단위 억(좌), 개(우)



월령에 따른 인기 제품 차이



04

데이터 분석 I 공모전

유한킴벌리 쇼핑몰 구매 데이터 분석

클러스터링

고객의 생애 가치를 구하는 변인

구매 최근성(Recency), 구매 빈도(Frequency), 매출액(Monetary)

정규화

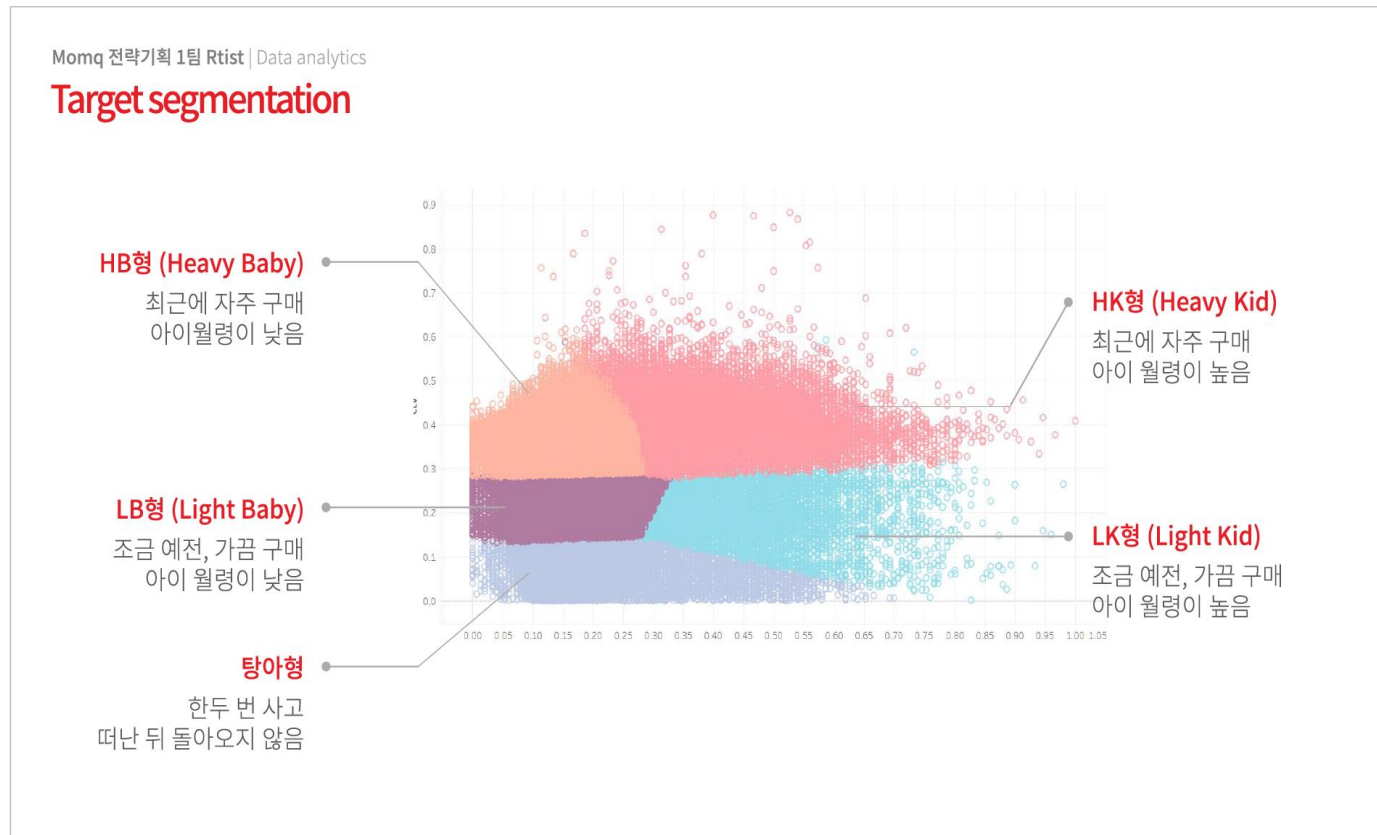
주어진 데이터에서 해당 파생변수를 생성하고 범위와 단위가 다른 파생변수들을 정규화한 후 이를 합산하여 고객별 생애가치 산출

시각화

고객생애 가치에 따라 K-means 클러스터링 기법을 통해 군집화

기타

기업의 실제 데이터를 바탕으로 진행한 첫 데이터 분석 프로젝트로서 다양한 분석기법 습득 및 분석가로 진로를 결정하게된 계기가 되었음



데이터 분석 | 공모전

동작구 내 고등학교 입지 분석

프로젝트 목적

2020 동작구 빅데이터 활용 정책 제안 공모전으로 동작구 내 지리정보와 공공데이터를 활용하여 최적의 고등학교 입지 제안

말은 어머니

데이터 수집 : 동작구 관련 인터넷 커뮤니티의 게시글 크롤링

데이터 전처리 : 공공데이터를 활용하여 치안,교통, 고등학교 수요를 나타내는 지표 생성

클러스터링 : 생성된 지표를 기준으로 군집화 실시·최적의 입지 후보 선정

더|이|터

구성: 커뮤니티 게시글 데이터, 동작구 행정동별 인구 및 고등학교 수 데이터

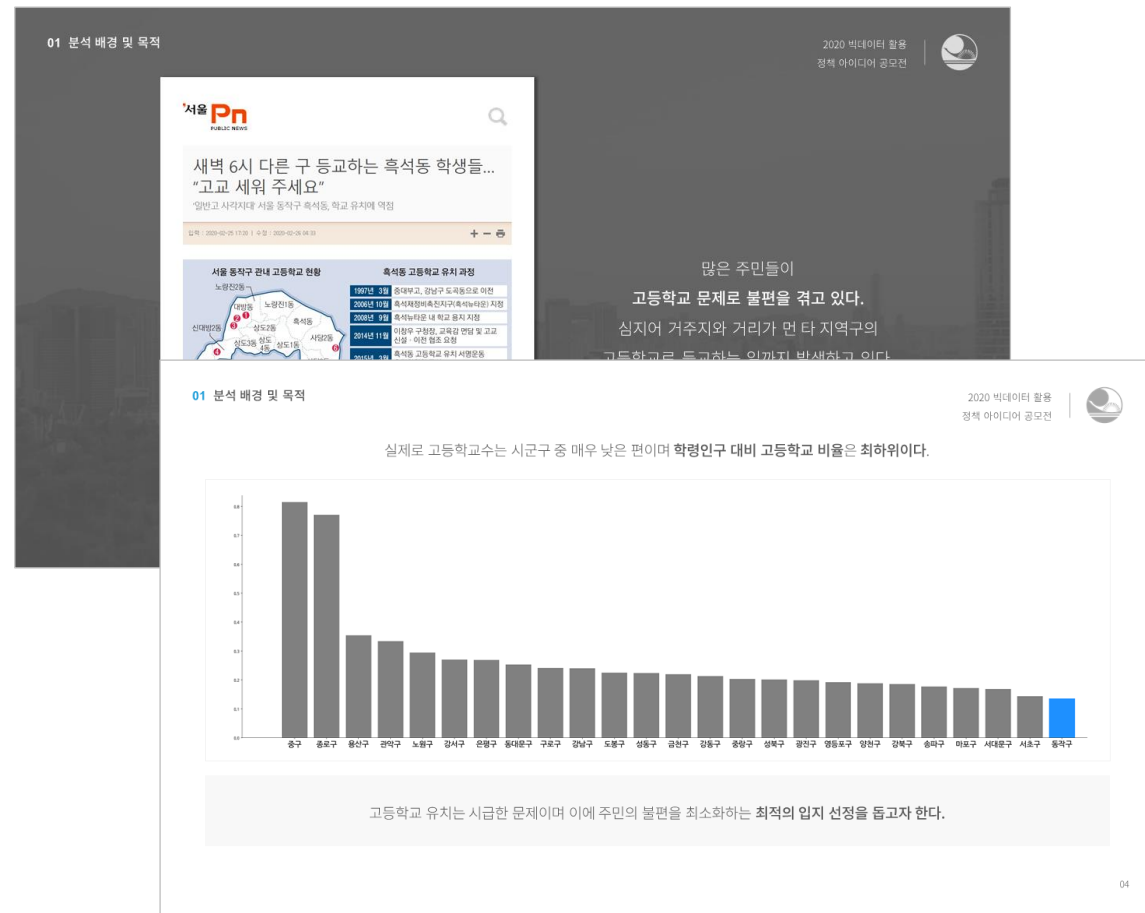
서울시 업종 및 토지 특성 데이터와 같은 공공데이터

프로젝트 기간

2020년 4월 (약 4주)

제작 환경

Jupyter Lab



04

데이터 분석 I 공모전

동작구 내 고등학교 입지 분석

분석개요

분석 주제 선정 과정

동작구 맘카페 파인트리 게시판 글 크롤링
결과, “흑석동”, “동작구”를 제외하고
“고등학교”, “이전”과 같은 단어의 빈도가 높음
따라서 고등학교 입지 선정을 주제로 선정

동작구 내 행정동 군집화 기준

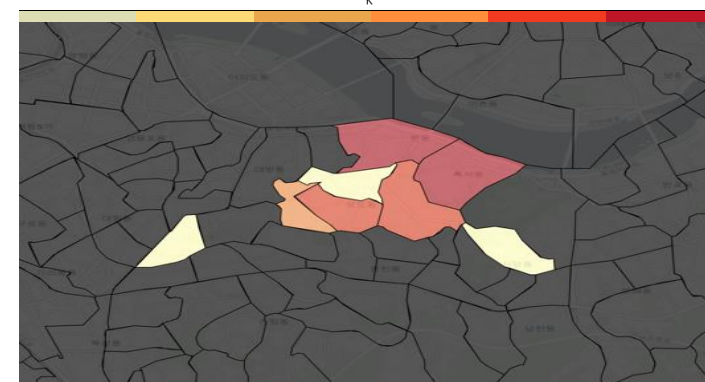
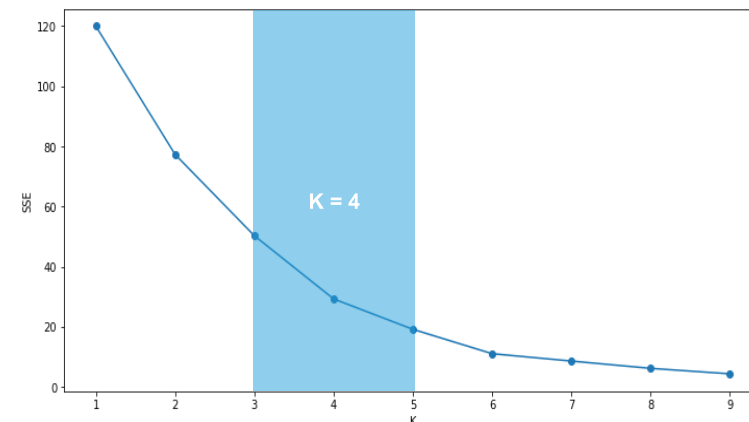
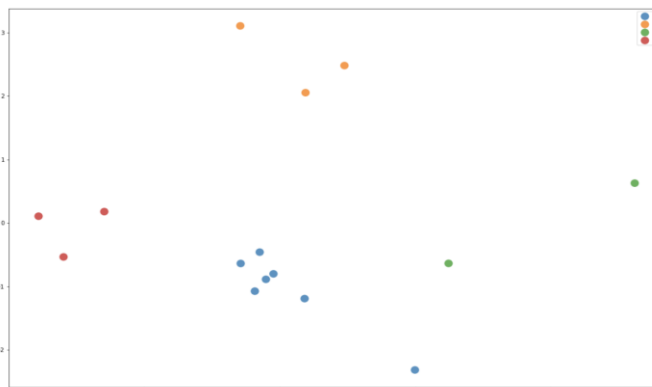
- 접근성 : 통학에 유리한가
- 안정성 : 치안등급이 높은가

추가 가중치

- 인구수 대비 청소년 비율 및 학교 비율

기타

동적·정적 크롤링 및 비정형 데이터 처리에 대한
이해와 호기심 발굴, 데이터 분석 주제를 설정 및
적정한 분석 기법 적용의 중요성을 깨달음



05

부록 I 개인 프로젝트

리그오브레전드 게임데이터 분석

프로젝트 목적

개인 프로젝트로 전략시뮬레이션 게임 리그오브레전드 개발사 라이엇에서 제공하는 API를 활용, 유저의 게임 데이터를 기반으로 승패 예측 분석을 실시

맡은 업무

데이터 수집 : API를 통해 매치 데이터 수집

데이터 분석 : 결측치 처리, Feature engineering, 모델링

데이터

기간 : 2013시즌, 그랜드마스터 레벨 유저 4380건 매치 데이터

프로젝트 기간

2020년 4월 (약 4주)

제작 환경

Jupyter Lab



05

부록 I 개인 프로젝트

리그오브레전드 게임데이터 분석

분석 개요

데이터 분석

- 데이터의 75%는 훈련용, 25%는 테스트용으로 사용하며 결측치를 “0”으로 변환
- 모델링을 위해 논리값을 가지는 변수의 값을 숫자형으로 변환하는 One-Hot Encoding 실시
- 랜덤포레스트 기법을 사용
- 모델링 결과는 정확도0.938로 승패 예측

분석 결과 시각화

예측을 하는데 있어 어떤 특성이 가장 중요한지 알아보기 위해 변수 중요도를 산출하여 그래프로 시각화

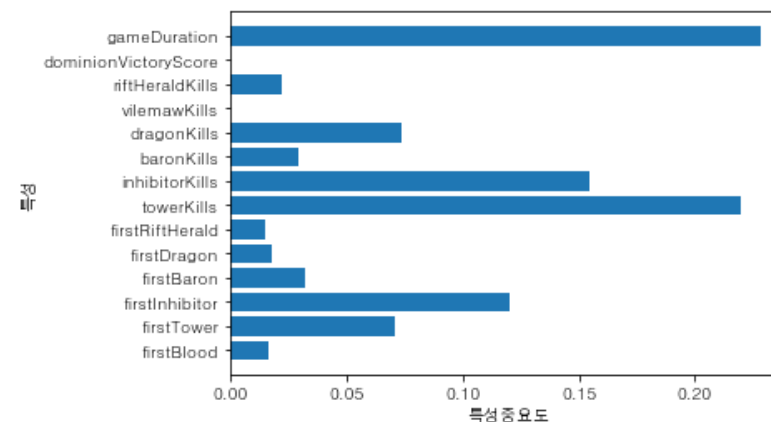
기타

단순 성능만 본다면 높은 확률로 승패를 예측한 것으로 보이나 너무나도 당연한 결과임
유익미한 분석을 위해선 분석 과제 또는 도메인에 대한 이해도가 높아야함을 깨달음

data_team

	teamId	win	firstBlood	firstTower	firstInhibitor	firstBaron	firstDragon	firstRiftHerald	towerKills	inhibitorKills	baronKills	dragonKills	vilemawKills
0	100	Fail	False	False	False	False	False	False	0	0	0	0	0
1	100	Win	True	True	True	True	False	True	9	2	1	2	0
2	100	Win	True	True	False	False	True	False	4	0	0	2	0
3	100	Win	True	True	True	True	False	True	6	1	1	2	0
4	100	Fail	False	True	False	False	True	True	3	0	0	1	0
...
4375	200	Fail	True	False	False	False	True	True	2	0	0	1	0
4376	200	Win	False	True	True	False	False	False	4	2	0	0	0
4377	200	Win	False	False	False	False	False	False	4	1	0	0	0
4378	200	Win	True	True	True	False	False	False	4	2	0	0	0
4379	200	Fail	True	True	False	False	False	False	1	0	0	0	0

4380 rows x 16 columns



05

부록 I 개인 프로젝트

카카오톡 채팅분석

프로젝트 목적

개인 프로젝트로 카카오톡 단체 대화방 데이터를 기반으로 토픽을 추출

맡은 업무

데이터 수집 : 카카오톡 어플의 대화 내보내기 기능을 통해 단체 대화방 데이터 수집

데이터 시각화 및 EDA : 데이터 분포를 시각화를 통해 파악

모델링 : 비정형 텍스트 데이터를 모델링에 적합한 형태로 전처리한 후 토픽 추출

프로젝트 기간

2020년 3월 (약 1주)

데이터

기간 : 2017년 1월 ~ 2020년 2월

대화 : 149,208건

제작 환경

Jupyter Lab



05

부록 I 개인 프로젝트

카카오톡 채팅분석

분석 개요

데이터 시각화 및 EDA

시간별, 월별, 일별 카톡수, 사람별 보낸

카톡수를 시각화하여 대략적 데이터 분포 파악

데이터 전처리

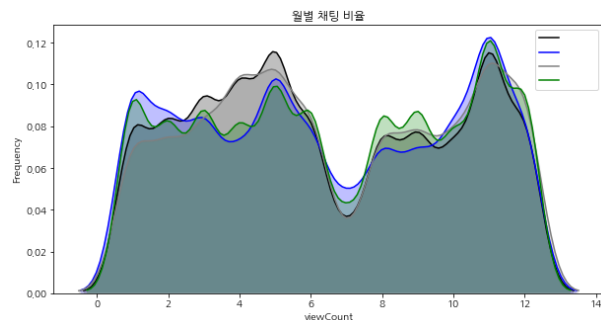
- 불용어 제거(특수문자, 공백, 의미없는 문자 제거)
- Konlpy를 활용 데이터 토큰화
- 명사와 동사만 추출하여 문장형태로 변환

모델링

- LDA 모듈을 통해 모델링을 실시 및 10개 토픽 추출

기타

자연어처리에 대한 이해



	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10
0	친구	신세	한국	미안	저울	보고	문제	사랑	먹고	머리
1	진짜	사람		한다	추천	거기	운동	원래	요요	고생
2	요즘	어디	바로	정도	한번	하고	노래	내일	가서	
3		해도	할게	하는거	안산	관심	하루	가능	되나	아하
4	하노	느낌	이야기	누구	카페	대구	방금	그때	아침	해야
5	오늘	이노	우리	하자	학교	저녁	지금	이름	확원	얼마
6	생각	하나	하면	어제	검색	경찰	공부	리얼	여행	주식
7		하는	해서		다시	언제	들어	시작	맞다	않나
8	사진			요미	그래그래	봤는데	되면	활인	했나	뉴스
9	시간	여기	이번	소리	무슨	연락	아따	하는데	인정	캐나다

최종 결과 데이터프레임

contents	year	month	day	weekday	24time	...	delta	is_indifferent	new	nnew	nouns	morphs	pos	dio_verb	verb	sentence
오 팀 이 만남 받 음	2017	01	19	Thursday	02:16:00	...	00:00:00	False	오 팀 이 만 남 받 음	오 팀 이 만 남 받 음	[오, 팀, 이 만]	[오, 팀, 이 만, 남, 받 음]	[(오, Noun), (팀, Noun), (이만, Noun), (남, Noun), (Josa), ...]	{'오': 'Noun', '팀': 'Noun', '이만': 'Noun', '남': 'Noun', ...}	[받 음]	[오, 팀, 이만, 받 음]
재민 그 동네 좀 좋네	2017	01	19	Thursday	02:16:00	...	00:03:00	False	재 민 그 동 네 좀 좋 네	재 민 그 동 네 좀 좋 네	[그, 동 네, 좀]	[그, 동 네, 좀, 좋 네]	[(그, Determiner), (동네, Noun), (좀, Noun), (좋네, ...)]	{'그': 'Determiner', '동네': 'Noun', '좀': 'Noun', ...}	[]	[동네, 좀]

감사합니다.