

산학연계 SW프로젝트

SNS 정보를 활용한 소상공인 발전 가능성 평가 서비스

TEAM INT



2018204035 한정수



2018204036 윤서안



2018204046 조민경



2018204021 권나현

INDEX

1

연구 주제 주제 변경 이유 및 소개

2

데이터셋 데이터 수집 및 전처리 방법

3

분석 및 모델링 데이터 분석 및 모델 구축

4

최종 결과물 UI 설명 및 시연

1

연구 주제

주제 변경 이유 및 소개

회사 소개 및 팀원별 역할

8PERCENT 주식회사 에잇퍼센트 (8PERCENT, Inc.)



한정수 (팀장)

Python, R 코드 작성
게시글, 좋아요 관련 코드 관리
데이터 수집 및 데이터베이스 관리
데이터 전처리 및 모델링
UI - JS 및 계정별 그래프 제작



좋아요 8,888개

방금



윤서안

Python, R 코드 작성
해시태그 관련 코드 관리
데이터 수집 및 데이터베이스 관리
데이터 전처리 및 모델링
UI - 디자인 및 분포 그래프 제작



좋아요 8,888개

방금



조민경

Python, R 코드 작성
프로필, 게시글 분류 코드 관리
데이터 수집 및 데이터베이스 관리
데이터 전처리 및 모델링
발표 자료 제작



좋아요 8,888개

방금



권나현

Python, R 코드 작성
위치, 날짜 관련 코드 관리
데이터 수집 및 데이터베이스 관리
데이터 전처리 및 모델링
UI - 디자인



좋아요 8,888개

방금

연구 주제 소개



기존 주제

SNS 정보를 활용한
개인 신용 평가 서비스



SNS만을 통한 개인 정보 수집의 한계
신용 평가 관련 데이터를 구할 수 없음



변경된 주제

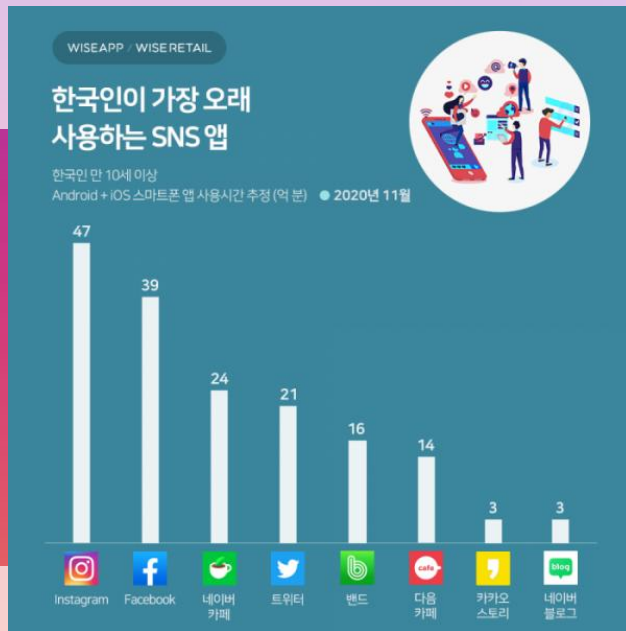
SNS 정보를 활용한
소상공인 발전 가능성 평가 서비스

소상공인 계정 개인 사업자 또는 작은 규모의 사업자가 상품 판매를 위해 운영하는 SNS 계정

신용 평가를 하는 방향으로 진행하기보다는
SNS를 통해 얻을 수 있는 인지도, 신뢰도 등의 데이터를 활용하여
소상공인들의 현재 활동과 추후 발전 가능성을 평가하려고 함

왜 인스타그램을 선택했는가?

해시태그를 이용한 검색이 용이
한국인이 가장 오래 사용하는 SNS 앱 1위를 기록 (지난 11월 기준)
미디어 이용자 수에 관하여 인스타그램이 다양한 연령층의 상위권을 차지
SNS 매체 중 홍보 효과가 탁월



국내 소셜 미디어 연령별 월평균 이용자 수 단위: 명

	10대	20대	30대	40대	50대
1위	221만	493만	440만	502만	544만명
2위	191만	386만	319만	298만	297만명
3위	86만	178만	268만	266만	177만명

※월 평균 이용자 수는 2020년 1분기(1~3월) 내 월별로 발생한 이용자 수의 산술평균값
자료=DMC미디어



SNS중 가장 향후 이용 확대 가능성이 높아 보이고, 기업의 마케팅 전략에 수월하게 이용될 수 있으며,
이용자들 또한 기업 브랜드 광고에 대한 거부감이 덜한 인스타그램을 연구대상으로 삼았다.
(이지영, '2019, 인스타그램 정보원 유형에 따른 지각된 정보원 속성이 광고효과에 미치는 영향')



Instagram

SNS



데이터 수집 및 분석



Markup Language
Content



Style sheet Language
Presentation



Programming Language
Behavior

UI

2

데이터셋

데이터 수집 및 전처리 방법

소상공인 계정 판별

다음과 같은 방법으로 소상공인의 계정 판별


- 1 마켓, 쇼핑몰, 패션, 주문제작 등 인스타 마켓과 관련도가 높은 23개의 해시태그로 게시물 검색
- 2 게시물을 업로드한 계정이 소상공인 계정이 맞는지 아래와 같이 확인

큰 회사나 유명인을 거르기 위해  '인증됨' 마크가 있는 계정 제외

이미 성공하여 자리를 잡아 대출 가능성이 별로 없는 계정을 거르기 위 팔로워 수가 100,000명이 넘는 계정 제외

소개글에 문의, 이벤트, 주문, 판매 등 상품 판매 계정이 자주 사용하는 35개의 단어 중 하나를 사용하면 계정 ID 수집

Profile 계정 기본 정보 크롤링



계정 이름

vtg_vivi 팔로우

게시물 / 팔로워 / 팔로잉

게시물 5,350 팔로워 5,472 팔로우 4,046


빈티지샵 (빈티지비비)

- ♥ 빈티지샵 빈티지 비비 입니다 ♥
- ♥ 택배 배송비 일반지역3천원 / 제주 및 산간지역 7천원 ♥
- ♥ 문의는 디엠으로 받고 답장없을때 아래 오픈챗 이용바래요 ♥
- ♥ 5만원 이상 구매시 배송비 무료 ♥

open.kakao.com/o/s3jAr8kc

소개글

Post 해당 계정의 게시물들 크롤링



계정 이름

vtg_vivi • 팔로우

vtg_vivi ♥ 버버리 금장버튼 니트 #빈티지비비구매가능

♥ 사이즈 : L 가슴 : 60 총장 : 60

♥ 상태 : 9

♥ 가격 : 12.0

♥ 문의는 DM 디엠으로만 받습니다

♥ 사진에 오염이 보이는것은 그것을 감안하고 가격을 맞춘것입니다 그래서 추가적인 할인은 없습니다

♥ 모든 사이즈는 실측을 기준으로 하였습니다

3시간

좋아요 1개

3시간 전

게시

해시태그

본문

위치정보

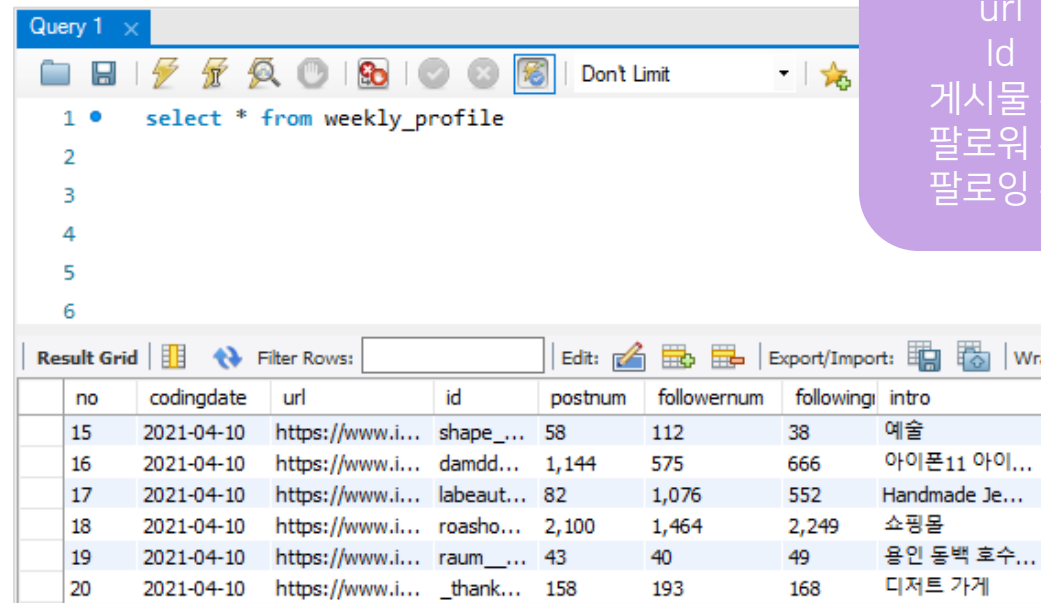
좋아요 / 작성 날짜

데이터 수집 방법

weekly_profile DB

총 3500개의 프로필 데이터

크롤링 날짜
url
id
게시물 수
팔로워 수
팔로잉 수



Query 1 x

1 • `select * from weekly_profile`

2

3

4

5

6

Result Grid

Filter Rows:

Edit: Export/Import: Wrap

	no	codingdate	url	id	postnum	followernum	following	intro
	15	2021-04-10	https://www.i...	shape_...	58	112	38	예술
	16	2021-04-10	https://www.i...	damdd...	1,144	575	666	아이폰11 아이...
	17	2021-04-10	https://www.i...	labeaut...	82	1,076	552	Handmade Je...
	18	2021-04-10	https://www.i...	roasho...	2,100	1,464	2,249	쇼핑몰
	19	2021-04-10	https://www.i...	raum_...	43	40	49	용인 동백 호수...
	20	2021-04-10	https://www.i...	_thank...	158	193	168	디저트 가게

3500개 계정의 프로필을 주 1회, 약 8주 동안 크롤링

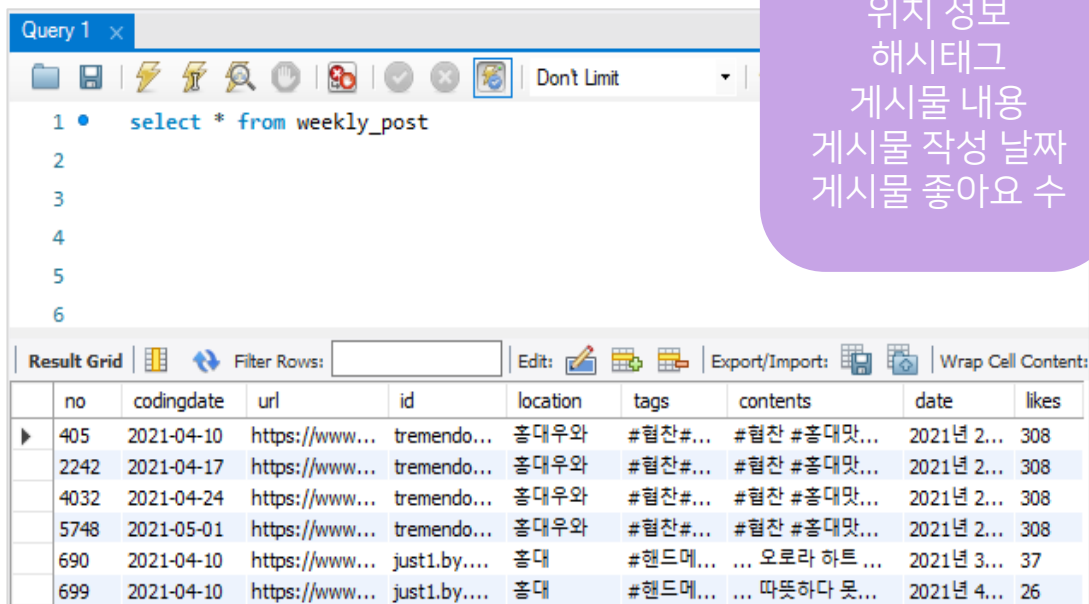
→ 프로필 정보의 변동을 보기 위함

데이터 수집 방법

weekly_post DB

약 28000개의 포스트 데이터

크롤링 날짜
게시물 URL
ID
위치 정보
해시태그
게시물 내용
게시물 작성 날짜
게시물 좋아요 수



Query 1 x

1 • select * from weekly_post

2

3

4

5

6

Result Grid

	no	codingdate	url	id	location	tags	contents	date	likes
▶	405	2021-04-10	https://www...	tremendo...	홍대우와	#협찬#...	#협찬 #홍대맛...	2021년 2...	308
	2242	2021-04-17	https://www...	tremendo...	홍대우와	#협찬#...	#협찬 #홍대맛...	2021년 2...	308
	4032	2021-04-24	https://www...	tremendo...	홍대우와	#협찬#...	#협찬 #홍대맛...	2021년 2...	308
	5748	2021-05-01	https://www...	tremendo...	홍대우와	#협찬#...	#협찬 #홍대맛...	2021년 2...	308
	690	2021-04-10	https://www...	just1.by...	홍대	#핸드메...	... 오로라 하트 ...	2021년 3...	37
	699	2021-04-10	https://www...	just1.by...	홍대	#핸드메...	... 따뜻하다 못...	2021년 4...	26

3500명 중 이상치 제거 진행하고 가계정들을 걸러냄

→ 박스 플롯으로 이상치 제거 후

팔로워 수에 비해 팔로잉 수가 심히 큰 경우 가계정이라 판단하고

팔로워/팔로잉 값이 0.25 미만인 값 제거

이 중 200개의 계정을 랜덤 샘플링하고

주 1회, 4주 동안 게시물 약 30개씩 크롤링

→ 게시물 관련 정보에 대한 데이터를 얻기 위함

새로운 컬럼 추가(R 사용)

like_mean_7	like_mean_all
35.000000	47.043478
135.400000	104.275862
99.750000	108.055556
53.000000	58.222222
11.500000	6.828571

like_mean_7 최근 7일 이내의 게시글의 좋아요 평균

like_mean_all 수집한 게시글 전체의 좋아요 평균

location_num	hashtag_average	post_upload
0	21.227273	9.850000
19	22.727273	14.857143
34	20.172414	3.137931
31	23.166667	1.055556
0	27.200000	4.285714

location num 계정 별 위치정보를 사용한 게시글의 수

hashtag_average 계정 별 게시글 당 사용 해시태그 개수 평균

post_upload 계정 별 게시글 업로드 주기

sellnum	dailynum
23	0
29	0
33	3
34	2
35	0

일상글 / 판매글

게시글 본문 내용 중 특정 단어가 들어갔는지

유무에 따라 구분

새로운 컬럼 추가 및 결측치 제거

all_tags	max_tags	max_tags_num
#에그타르트#취미#홀베이킹#베이킹#おかしづくり#ペー...	#디저트	28
#맛팔#맛팔선팔#맛팔한영#맛팔해요#중반#맛팔그림#맛팔...	#맛팔해요#맛팔좋아요	6
#오름에게#제주도게스트하우스#합덕게스트하우스#김녕게...	#제주도게스트하우스#서점숙소	27
#음식냄새제거#어라운드제이#어라운드제이캔들#어라운드...	#어라운드제이캔들앤숍	33
#블루라밍#instagram#instagood#쇼핑베이비#베이비스타...	#좋아요	36

all_tags 계정 별 전체 게시글의 모든 해시태그

max_tags 계정 별 가장 많이 사용한 해시태그

max_tags_num 계정 별 가장 많이 사용된 해시태그의 사용 횟수

```
#결측치 모두 0으로 처리  
post[is.na(post)] <- 0  
profile[is.na(profile)] <- 0
```

결측치 없음을 확인

```
#결측값 측정  
df.isnull().sum()
```

```
id          0  
postnum     0  
followerum  0  
followingum 0  
sellnum     0  
dailynum    0  
like_mean_7 0  
like_mean_all 0  
location_num 0  
hashtag_average 0  
post_upload 0  
dtype: int64
```

최종 데이터

id	no	codingdate	url	postnum	followerum.x	followingnum	intro	sellnum	dailynum
0_baking_0	215	2021-05-01	https://www.instagram.com/0_baking_0 /	23	75	42	얼렁뚱땅 취미생활과 올레볼레 사진실력 맞팔은 편하게 댓...	23	0
530ee	230	2021-05-01	https://www.instagram.com/530ee /	259	1302	831	예술가	29	0
a.bookhome	232	2021-05-01	https://www.instagram.com/a.bookhome /	656	1765	744	서점	33	3
around_j_candle.soap	257	2021-05-01	https://www.instagram.com/around_j_candle.soap /	1290	1497	500	지역 비즈니스	35	1
blue_raming	227	2021-05-01	https://www.instagram.com/blue_raming /	44	293	606	유아 및 어린이 의류 상점	35	0
bong_pot	262	2021-05-01	https://www.instagram.com/bong_pot /	162	263	254	김포시 김포대로1199 일봉도예 **100% 수제로 만들어지는 ...	35	0
by.eva92	241	2021-05-01	https://www.instagram.com/by.eva92 /	50	1361	773	여성복 상점	32	1

like_mean_7	like_mean_all	location_num	hashtag_average	post_upload	all_tags	max_tags	max_tags_num
35.000000	47.043478	0	21.227273	9.850000	#에그타르트#취미#홈베이킹#베이킹#おかしづくり#ペー...	#디저트	28
142.000000	112.655172	19	23.307692	12.857143	#온더센셋#인스타그램#친스타그램#럽스타그램#울스타그램...	#좋아요#맞팔해요	6
76.500000	109.472222	34	19.866667	3.862069	#제주도게스트하우스#합덕게스트하우스#김녕게스트하우...	#제주도게스트하우스	28
58.909091	60.250000	28	20.861111	1.000000	#불멍#점화식#캔들포장#카네이션캔들#어버이날선물#어...	#어라운드제이캔들앤숍	30
11.500000	6.828571	0	27.200000	4.285714	#블루라밍#instagram#instagood#쇼핑베이비#베이비스타...	#좋아요	36
36.428571	40.200000	0	20.900000	2.235294	#수제화본#다육화본#도자기화본#분재화본#화본#예쁜화...	#일봉도예#봉팻	29
76.333333	88.606061	21	9.850000	11.700000	#대구#대구핫플#수성못#수성못카페#고바순#데이트#유부...	#수성못	8

총 18개의 컬럼

약 200개의 계정을 1달간 주 1회마다 수집

3

분석 및 모델링

데이터 분석 및 모델 구축

미래 팔로워 수를 종속 변수로

논문 발췌

인스타그램 계정 팔로워 수는 해당 계정의 신뢰성과 호감에 유의한 영향을 미치는 것으로 밝혀졌다. 즉, 팔로워 수가 높을수록 이용자에 대한 신뢰감과 호감을 높게 인지하였다. 이는 게시글을 보는 이용자의 게시글 태도와 구매의도에서도 유의미한 영향을 미치는 것으로 나타났다.

소비자 92%는 일반광고보다 인플루언서가 언급한 내용에 신뢰감을 가진다고 한다.

많은 팔로워 수를 보유하고 있는 브랜드는 이용자들의 관심을 받는 만큼 인기 있는 브랜드라는 것이 증명된다. (김린아, 2017; Iconsquare, 2016)



미래(다음주) 팔로워 수를 예측하여 발전가능성을 평가하기로 결정

05/08의 팔로워 수 (= 현재 05/01 기준 다음주 팔로워 수)

이전 3주간의 데이터 (04/10 04/17 04/24)를 이용

학습을 위해 사용할 데이터

이번주 프로필 데이터에 다음주 팔로워 수를 붙임

ex) 4/10의 profile 데이터(현재 데이터) + 4/17 팔로워 수(미래 팔로워 수)

4/17의 profile 데이터(현재 데이터) + 4/24 팔로워 수(미래 팔로워 수)

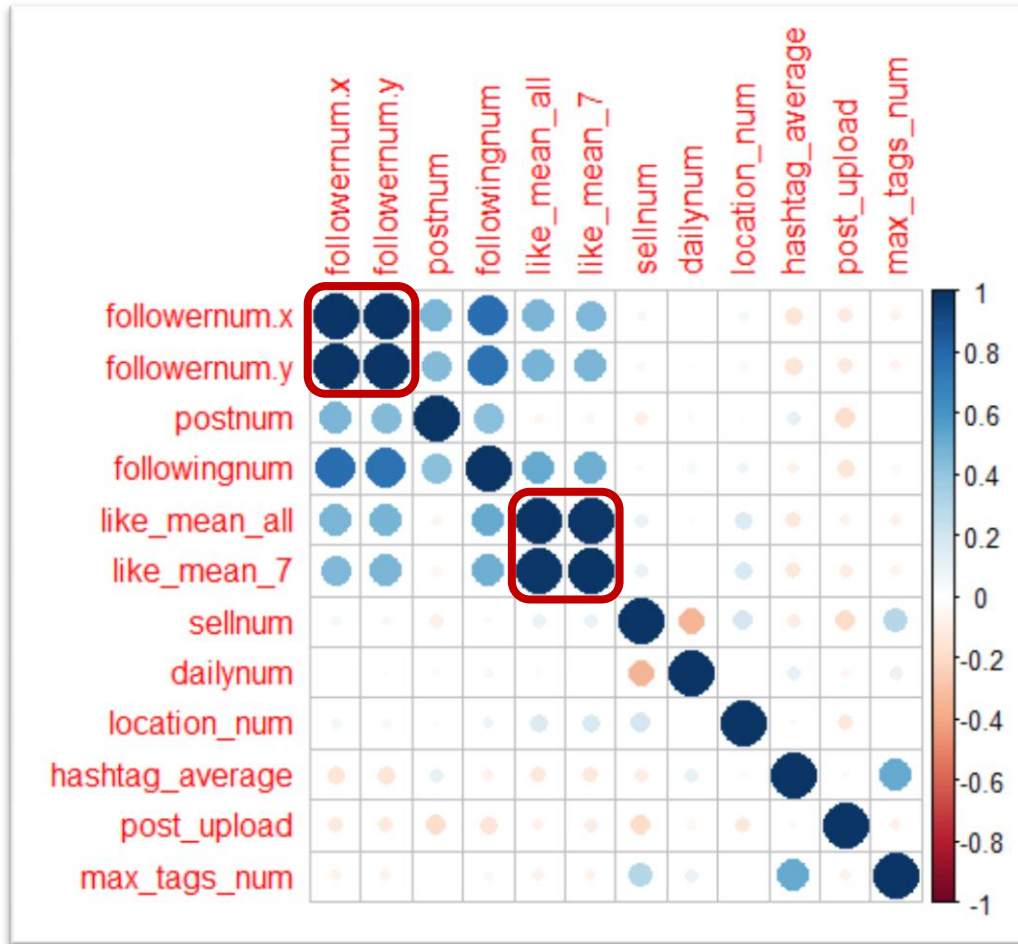
4/24의 profile 데이터(현재 데이터) + 5 / 1 팔로워 수(미래 팔로워 수)

id	no	codingdate	url	postnum	followernum	followingnum	intro	setinum	dailyum	like_mean	like_mean_2	location_num	hashtag_ave	post_upload	all_tags	max_tags	max_tags
kong_mi	23	2021-04-10	https://www	440	872	304	모든문의카	36	0	12.4545455	16.9444444	35	30	2.3	#공자매#원!#공자매#신:		
ramitage_	25	2021-04-10	https://www	354	1090	478	번티지스토	32	0	15.125	21.96875	0	15.7931034	5	#번티지선글#번티지선글		
yesalpha	21	2021-04-10	https://www	540	2373	1883	블로거	36	0	293	143.638889	28	3.22222222	1.03448276	#나인블럭#알파양투아		
irdhand_jew	36	2021-04-10	https://www	139	505	979	서드핸드_든	24	0	42	47.25	24	24.1666667	2.05882353	#아트클레아#영종도		
lyoul.mom	44	2021-04-10	https://www	5686	3972	5268	17년 딸 18년	19	0	17.9333333	22.1052632	0	21.5263158	3	#형찬#광고#이벤트		
ing_star030	7	2021-04-10	https://www	360	998	373	0	32	1	31.4	31.5151515	0	20.4848485	1.38461538	#생후369일#육아스타그		
abybo_91	4	2021-04-10	https://www	661	1397	1135	외류(브랜드	36	0	15.2777778	15.2777778	1	8.78947368	1	#생애첫베아#공구#고민		
rebedudu_	11	2021-04-10	https://www	321	512	264	외류	35	0	19.25	15.9428571	1	7.85714286	1.47368421	#나영원피스#나영원피스		
enign99	31	2021-04-10	https://www	1239	408	308	비나인 카톡	24	0	9.42857143	11.125	0	25.7083333	1.52380952	#반팔셔츠도#ootd#문당		
ingha_net	50	2021-04-10	https://www	36	351	319	시즌2 OPEN	36	0	82	111.111111	0	24.3793103	3.1	#다이어리꾸#다꾸스타그		
lackholej_c	33	2021-04-10	https://www	575	420	273	쇼핑 및 유통	36	0	8.4	8.75	0	27.5833333	1.66666667	#BOTTOM#O#ACC		
o_r10806	38	2021-04-10	https://www	77	770	738	이름:보리	36	0	65.5	54.6388889	0	7	6	#보리#자기#영스타그		
ry_yiso	19	2021-04-10	https://www	693	3922	5184	쇼핑 및 유통	12	0	39.2	23.5	1	29.3333333	3.11111111	#by_yiso#데#데일리룩#		
harming_hc	26	2021-04-10	https://www	388	463	123	패션개미지	35	0	20	14.2571429	34	29.4	3.75	#짐머만원#립스타그		
herrying_01	18	2021-04-10	https://www	272	3595	1096	패션모델	36	0	113.666667	209.638889	0	6.56521739	1.22727273	#형찬#형찬#형찬		
layyoons_	24	2021-04-10	https://www	130	909	666	쇼핑 및 유통	36	0	106.375	101.777778	0	32.5	1.32	#데이윤즈#스타일		
lana_rally	25	2021-04-10	https://www	1077	1107	081	자기	36	0	07	08.1500000	0	0.07142857	1.14285714	#드레가#드레가		



followernum.y
883
1088
2388
507
3993
1028
2261
520
411
356
421
773
3919
464
3618
998
1195

상관성 분석



follower.x - follower.y

like mean 7 - like mean all

위와 같이 높은 상관관계를 보이는 컬럼들을 함께 두면

과적합이 발생할 수 있음

따라서 해당 컬럼들의 다중공선성 확인이 필요

다중공선성 확인

다중공선성(Multicollinearity)

모형의 일부 예측 변수가 다른 예측 변수와 상관되어 있을 때 발생하는 조건

중대한 다중공선성은 회귀계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들

like_mean_7 - like_mean_all의 다중공선성이 높으므로 둘 중 하나 제거

다중공선성이 모두 낮음을 확인할 수 있음

```
result1 <- lm(followernum.y ~ postnum + followernum.x + followingnum +  
sellnum + dailynum + like_mean_7 + like_mean_all + location_num +  
hashtag_average + post_upload + max_tags_num, data=data)
```

```
vif(result1)  
postnum      followernum.x    followingnum      sellnum      dailynum  
1.701636      2.972341      3.192245      1.597886      1.215484  
like_mean_7   like_mean_all    location_num    hashtag_average    post_upload  
29.516719     31.467250      1.071740      1.692178      1.138432  
max_tags_num  
1.924389
```



```
result2 <- lm(followernum.y ~ postnum + followernum.x + followingnum +  
sellnum + dailynum + like_mean_7 + location_num +  
hashtag_average + post_upload + max_tags_num, data=data)
```

```
vif(result2)  
postnum      followernum.x    followingnum      sellnum      dailynum  
1.646818      2.969070      2.968434      1.590330      1.207347  
like_mean_7   location_num    hashtag_average    post_upload    max_tags_num  
1.639745      1.071632      1.687525      1.128709      1.900614
```

like_mean_all 제거

종속변수 y = 미래(다음주) 팔로워 수

독립변수 x = 다중공선성을 보이는 변수를 제거한 변수 전부

모델 성능 비교

Random Forest

	.metric <chr>	.estimator <chr>	.estimate <dbl>
1	rmse	standard	803.
2	rsq	standard	0.939
3	mae	standard	202.

RMSE = 803

RSQ = 0.939

Boosting

	.metric <chr>	.estimator <chr>	.estimate <dbl>
1	rmse	standard	815.
2	rsq	standard	0.935
3	mae	standard	239.

RMSE = 815

RSQ = 0.935

Multiple Linear Regression

Residual standard error: 0.08438 on 410 degrees of freedom
Multiple R-squared: 0.4507, Adjusted R-squared: 0.4413
F-statistic: 48.06 on 7 and 410 DF, p-value: < 2.2e-16

RSQ = 0.4413

RMSE값은 작을수록, R-Square(결정계수)값은 클수록 좋은 모델
93.9%의 설명력을 가지는 **Random Forest**가 가장 예측력이 좋다고 판단

Random Forest 결과

현재 팔로워 수

계정 아이디

예측된 미래 팔로워 수

followernum.x	id	.pred
592	_e_room_	586.96167
890	_ggomzi	933.34017
901	_kong_mi_	854.83100
237	_maison_de_thread	249.44583
1127	_ramitage_	1062.59883
1209	_s.label_	1210.49217
1069	_sian.7	1127.63467
1025	_tembox	1058.05133
330	_thanks_more	463.18383

Random Forest 결과

follower_Rate_of_change	Rate_rank	rank_percent
1.0000000	146	73.3668342
1.0144231	51	25.6281407
1.0028620	110	55.2763819
1.0027119	111	55.7788945
0.9901316	197	98.9949749
1.0286885	32	16.0804020
1.0022288	120	60.3015075

Rate_rank 전체 계정 중 해당 계정의 순위
rank_percent 순위를 백분율로 표현 (상위 %)

follower_Rate_of_change (팔로워 변화율)

= 미래 팔로워 수 / 현재 팔로워 수

팔로워 변화율 < 1 : 값이 작아질수록 미래 팔로워 수가 줄어듦

팔로워 변화율 = 1 : 팔로워 수 변화 없음

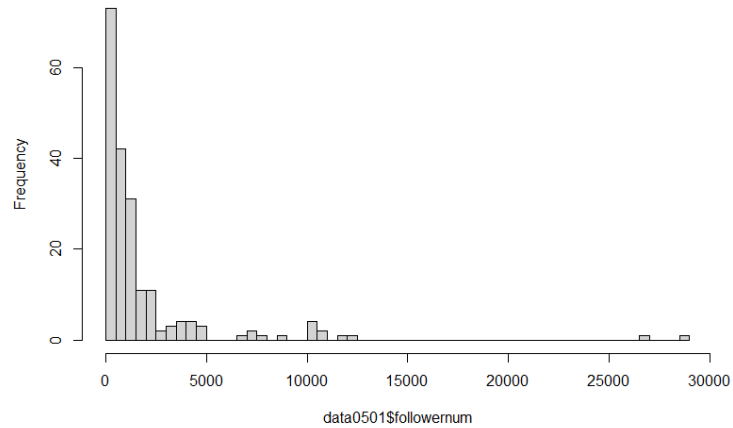
팔로워 변화율 > 1 : 값이 커질수록 미래 팔로워 수가 늘어남

$$\text{발전 가능성} = \left(\frac{\text{미래 팔로워 수}}{\text{현재 팔로워 수}} - 1 \right) \times 100$$

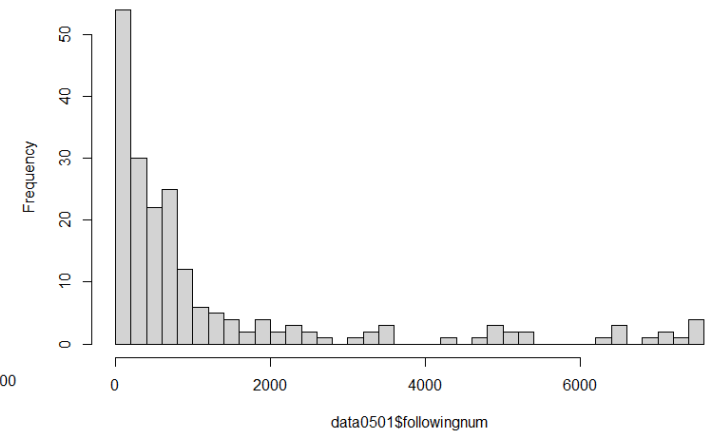
유의한 변수들의 가치와 분포

```
# A tibble: 10 x 2
  variable      imp
  <fct>      <dbl>
1 followernum.x 100
2 followingnum  59.4
3 like_mean_7   26.2
4 postnum       20.8
5 post_upload   13.0
6 hashtag_average 9.87
7 max_tags_num  6.77
8 sellnum        2.80
9 location_num   2.30
10 dailynum      0.837
```

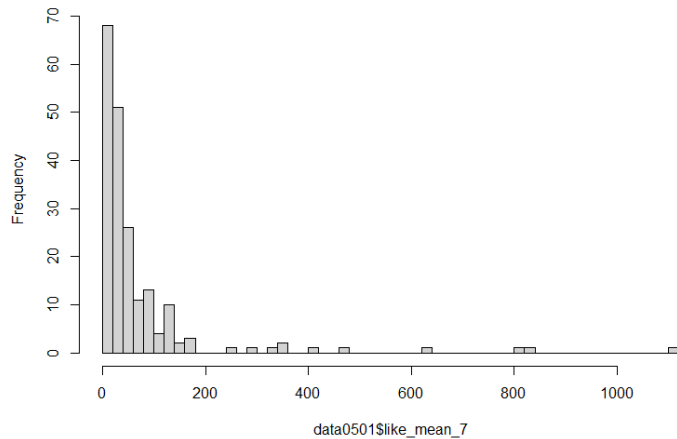
Histogram of data0501\$followernum



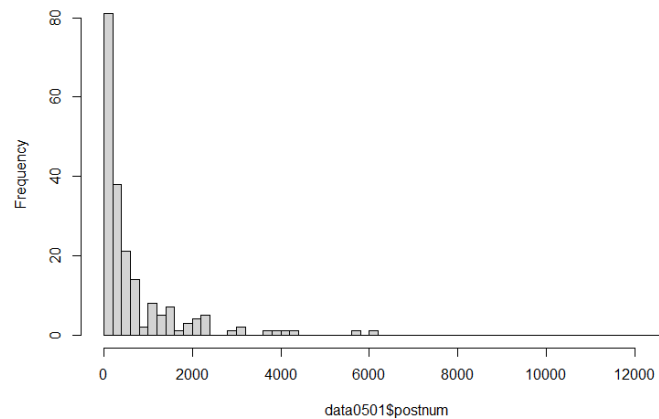
Histogram of data0501\$followingnum



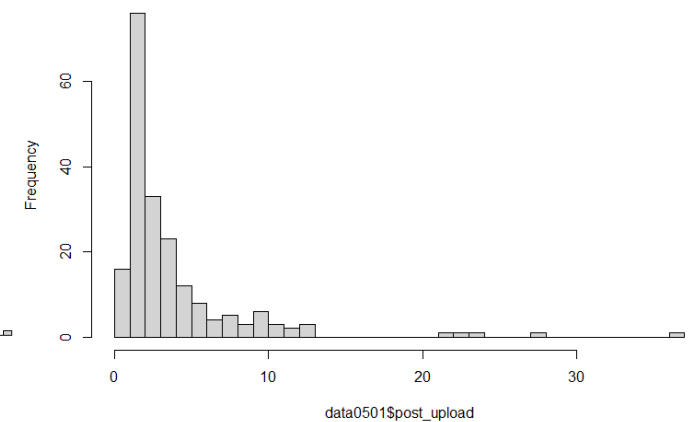
Histogram of data0501\$like_mean_7



Histogram of data0501\$postnum



Histogram of data0501\$post_upload

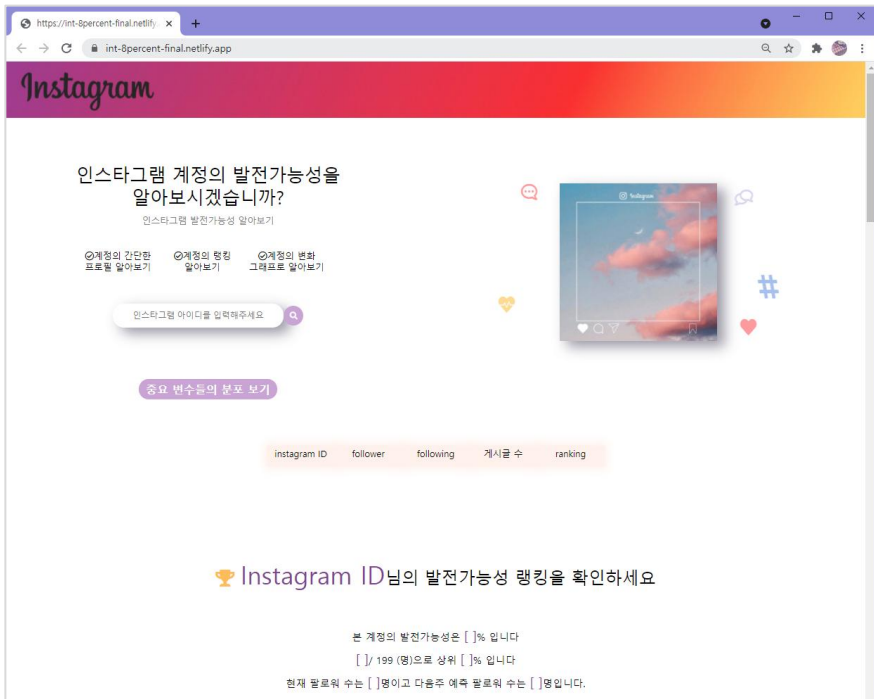


4

최종 결과물

UI 설명 및 시연

메인 페이지



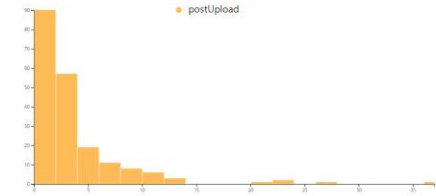
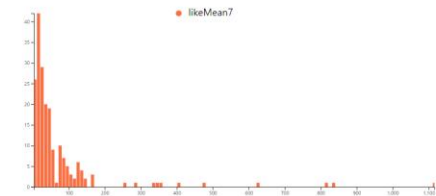
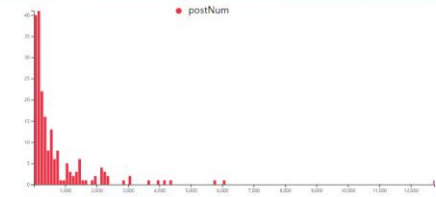
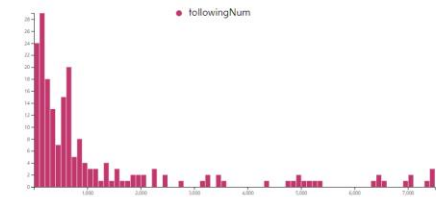
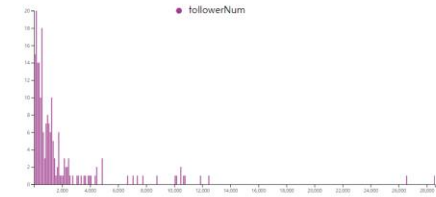
스크롤 다운



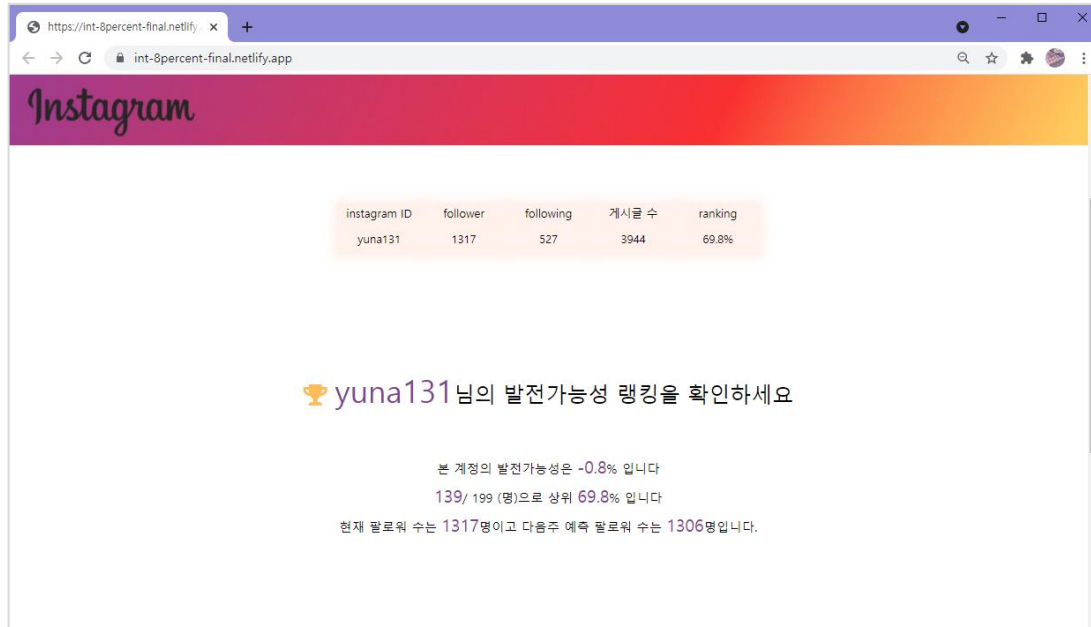
전체 데이터(계정)의
주요 Factor 분포

Instagram

<중요 변수들의 분포>



계정 ID 입력 후



스크롤 다운



하나의 계정을 입력했을 때의
데이터 분석

Instagram



UI 시연

<https://int-8percent-final.netlify.app/>

Thank You