

# 미세먼지의 사회적 영향력 분석

---

## 냥냥편치

김진욱 이지수 이지원 허새롬



# 목차

---

- 데이터 전처리
- 군집분석
- 다중회귀분석
- 분산분석
- SNS 데이터 분석
- 비즈니스 아이디어

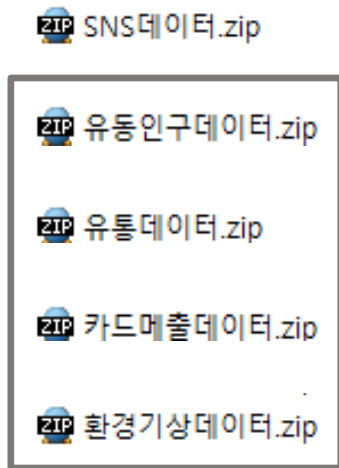
# 데이터 전처리

---



# 데이터 전처리

## 공통 처리사항



행정동 단위로 분류

날짜로 인덱스 설정



# 군집분석

---



# 군집분석

## 전처리

종로구		노원구	
청운효자동	11110515	월계1동	11350560
사직동	11110530	월계3동	11350580
부암동	11110550	공릉1동	11350595
평창동	11110560	공릉2동	11350600
교남동	11110580	하계1동	11350611
가회동	11110600	중계본동	11350619
종로1.2.3.4가동	11110615	중계2.3동	11350625
종로5.6가동	11110630	상계1동	11350630
이화동	11110640	상계2동	11350640
혜화동	11110650	상계3.4동	11350665
창신1동	11110670	상계5동	11350670
창신3동	11110690	상계6.7동	11350695
송인2동	11110710	상계10동	11350720

환경기상 데이터는 총 33개의 행정동 중 26개 동에 대한 데이터만 존재



# 군집분석

## 다중공선성

데이터	독립변수						
환경기상	pm10	noise	temp	humi	pm25		
유통	매출지수	식사	간식	마실거리	홈리빙	헬스뷰티	취미여가활동
	사회활동	임신육아					
유동인구	유동인구						
카드매출	C20	C21	C22	C40	C42	C44	C50
	C52	C62	C70	C71	C80	C81	C92

독립변수가 총 29개로 다중공선성이 발생할 확률 높음  
독립변수들 간의 강한 상관관계



# 군집분석

## 다중공선성 제거

	pm10	noise	temp	humi	pm25	매출지수	식사	간식	...
...									
humi	-0.04828	0.086883	0.059157	1	-0.00673	0.316972	0.472792	0.274613	...
pm25	0.97046	-0.15805	0.128455	-0.00673	1	-0.02711	0.060721	-0.009	...
매출지수	-0.03532	0.357514	0.12154	0.316972	-0.02711	1	0.863653	0.920223	...
식사	0.022109	0.397347	0.143911	0.472792	0.060721	0.863653	1	0.768508	...
간식	-0.06298	0.270185	-0.0194	0.274613	-0.009	0.920223	0.768508	1	...
...	...								

각 변수 간 상관계수가 절댓값 0.7 이상인 변수를 제거해  
다중공선성 제거





# 군집분석

## 다중공선성 제거

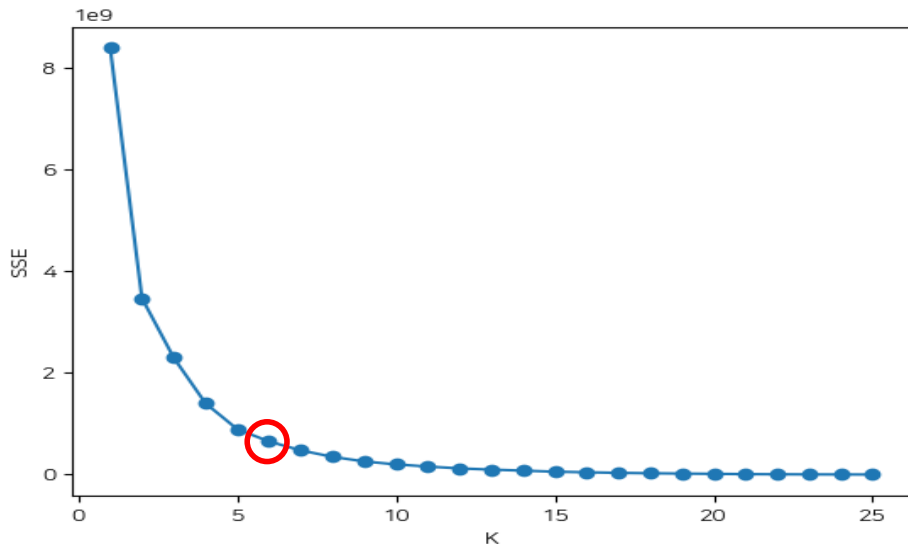
features	VIF	features	VIF
pm10	1.99899	flow	6.309606
noise	2.436092	C20	4.468563
temp	1.743349	C21	3.071286
humi	1.887539	C22	4.651343
매출지수	4.703123	C40	3.297197
홈리빙	4.233787	C52	1.891197
취미여가활동	1.912857	C62	1.925277
임신육아	2.563654	C70	2.111599

분산팽창인자 VIF가 전부 10 미만으로 다중공선성 제거 확인 후  
16개의 변수를 최종변수로 선택



# 군집분석

## 군집 개수 선택



각 군집에 속한 데이터와 군집의 중심 사이의 거리가 작아지는 방향으로  
최적의 군집 개수를 탐색하는 Elbow Method를 사용,  
가장 적합한 군집 수를 6개로 판단



# 군집분석

## K-means Clustering

군집	행정동							
0	창신3동							
1	사직동	월계3동	공릉1동	하계1동	중계2.3동	상계1동	상계2동	
2	부암동	평창동	교남동	가회동	혜화동	창신1동	송인2동	
3	종로1.2.3.4.가동							
4	이화동	상계6.7동						
5	청운효자동	종로5.6가동	월계1동	공릉2동	중계본동	상계3.4동	상계5동	상계10동

알고리즘이 쉽고 직관적인 군집분석 방법인 K-means를 이용하여  
26개의 행정동을 6개의 군집으로 분류



# 군집분석

## Mapping



군집 별로 지도상의 위치 표시



# 군집분석

## Result

군집	pm10	noise	temp	humi	매출지수	홍리빙	취미여가 활동	임신육아	flow	C20	C21	C22	C40	C52	C62	C70
0	1.59	-2.35	0.78	-1.90	-2.00	-2.00	2.42	-0.80	-0.96	1.16	0.96	1.53	-0.76	-0.48	-0.19	-0.63
1	-0.72	-0.16	-0.61	0.08	0.41	-0.10	0.03	0.54	0.13	-0.40	0.15	0.19	1.12	-0.25	0.76	-0.20
2	0.16	0.92	0.22	0.14	0.14	0.54	0.60	-0.94	-0.34	0.14	-0.28	-0.32	-0.57	-0.20	-0.58	-0.32
3	-1.18	0.58	-0.04	-0.17	2.38	1.01	-0.27	-1.09	4.23	3.15	2.74	3.47	1.27	1.99	0.08	1.28
4	0.93	0.33	0.25	0.49	0.40	1.56	-0.60	0.98	0.32	-0.45	-0.05	-0.47	-0.23	-0.25	-1.04	2.80
5	0.20	-0.52	0.18	-0.06	-0.63	-0.65	-0.67	0.34	-0.30	-0.20	-0.34	-0.39	-0.49	0.27	0.12	-0.33

6개의 군집의 변수별 표준화값



# 다중회귀분석

---



# 다중회귀분석

## 분석 목적

종속변수

성별별 카드매출  
유통데이터



독립변수

pm10 temp  
humi flow

독립변수가 종속변수에 미치는 영향을 알아보기 위해  
회귀분석 진행



# 다중회귀분석

## 전처리

25세 이하 여성의 레저업소 카드매출

A: 전성별  
F: 여성  
M: 남성

**F 2 5 C 2 1**

성별 나이코드 업종코드

\* 카드매출데이터 데이터정의서의  
나이코드, 업종코드 참고

	A20C40	A20C70	...	F20C40	F20C70	F25C21	...	M65C40	M65C70
2018-04-01	0.463162		...	-0.40609			...	-0.36758	
2018-04-02	0.051093	1.813224	...	0.875589	1.793695	0.17569	...	-0.13686	-0.13441
2018-04-03	-0.25502	-0.24329	...	-0.72274			...	-0.62576	
2018-04-04	-0.21969	-0.36359	...	0.038729	-0.26123		...	-0.85098	
2018-04-05	-0.53758	-0.30631	...	-1.03939			...	-0.80703	-0.12269
...									

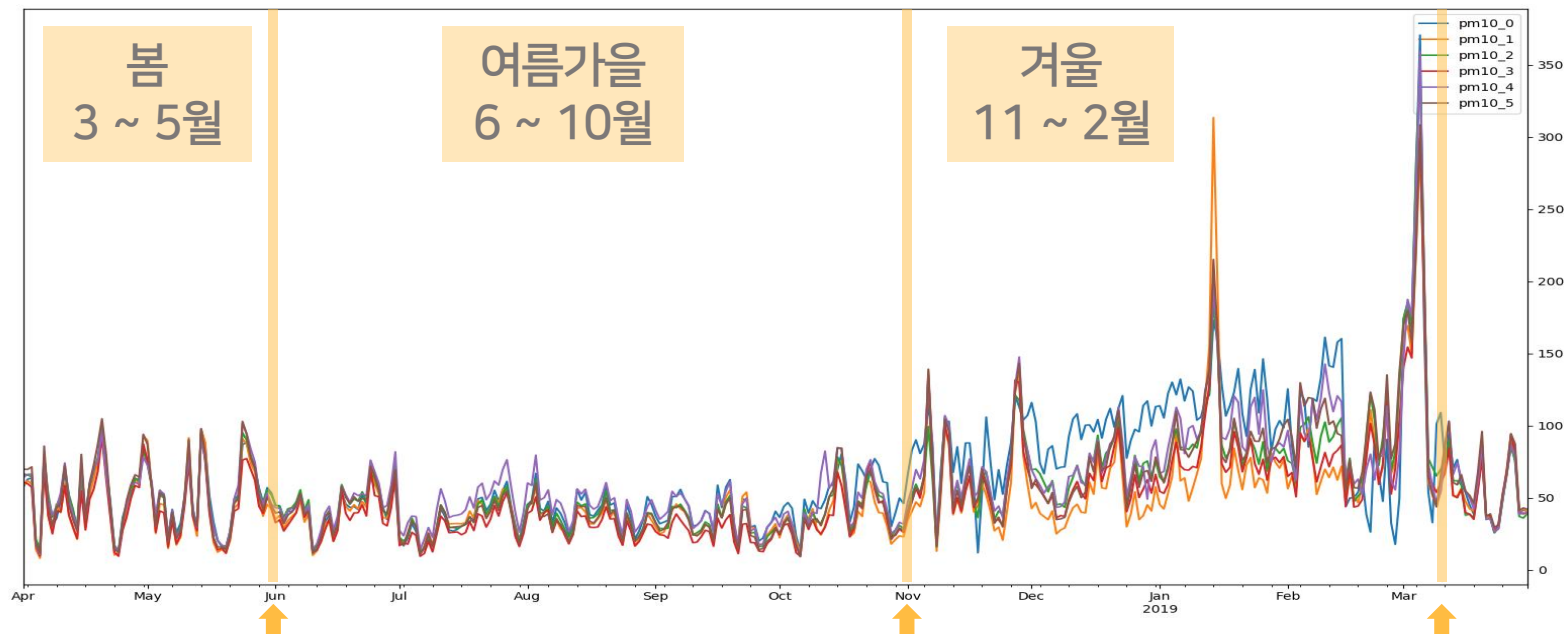
카드매출 데이터의 변수는 성별 및 연령을 구분하여 재조정된 뒤  
변수의 데이터가 20개 미만인 경우 제외





# 다중회귀분석

## 미세먼지의 계절적 요인



미세먼지의 시계열 그래프로 추이를 파악해 계절 구분



# 다중회귀분석

	종속변수	독립변수
1	매출지수 등의 유통데이터   전체 카드매출 데이터	pm10 + temp + humi + Aflow
2	여성 카드매출 데이터	pm10 + temp + humi + Fflow
3	남성 카드매출 데이터	pm10 + temp + humi + Mflow

(Ordinary Least Squares)

가장 기본적인 결정론적 선형 회귀방법인 OLS 사용  
미세먼지의 계절성을 고려하여 계절별로 회귀분석 진행



# 다중회귀분석

OLS Regression Results

Dep. Variable:

Model:

Method:

Date:

Time:

No. Observations:

Df Residuals:

Df Model:

Covariance Type:

Least Squares

Mon, 19 Aug 2019

23:24:33

43

38

4

nonrobust

R-squared:

F-statistic:

Prob (F-statistic):

Log-Likelihood:

AIC:

BIC:

0.219

2.670

0.0467

-10.137

30.27

39.08

	coef	std err	t	P> t	[0.025	0.975]	
pm10	0.1157	0.036	3.198	0.003	0.042		0.189
humi	-0.0099	0.048	-0.204	0.839	-0.108		0.088
temp	-0.0348	0.106	-0.328	0.744	-0.249		0.180

Omnibus:

Prob(Omnibus):

Skew:

Kurtosis:

47.325

0.000

2.745

12.127

Durbin-Watson:

Jarque-Bera (JB):

Prob(JB):

Cond. No.

2.536

203.263

7.28e-45

3.26

모델의 결정력이 0.2이상, pm10의 p-value가 0.05 이하,  
humi와 temp보다 pm10의 영향력이 큰 변수만 추출



# 다중회귀분석

## 분석결과

미세먼지의 영향력이 큰 변수	군집	특징
자동차정비	1군집	아파트 위주 주거지역
레저업소	3군집	관광지, 유동인구 높음
의료기관	1/2/3군집	봄철 환절기와 미세먼지의 시너지
유통	2/4/5군집	겨울 : 추운 날씨가 더해져 밖으로 잘 안 나옴 봄 : 실외활동보다 실내활동 선호
서적	1/3/4/5군집	3월 학기 시작



# 다중회귀분석

## 분석결과

미세먼지의 영향력이 큰 변수	군집	특징
자동차정비	1군집	아파트 위주 주거지역
레저업소	3군집	관광지, 유동인구 높음
의료기관	1/2/3군집	봄철 환절기와 미세먼지의 시너지
유통	2/4/5군집	겨울 : 추운 날씨가 더해져 밖으로 잘 안 나옴 봄 : 실외활동보다 실내활동 선호

미세먼지 변수와 인과관계가 확실하지 않은 서적 변수 삭제  
(새학기라는 특수 시즌의 영향이 더 큼)



# 분산분석

---



# 분산분석

## 분석 목적

종속변수

유동인구  
카드매출



독립변수

mise  $\begin{pmatrix} \text{good} \\ \text{bad} \\ \text{dead} \end{pmatrix}$

미세먼지 수치 정도에 따라 유동인구와 카드매출이  
어느 정도 차이나는지 알아보기 위해 분산분석 시행



# 분산분석

## 전처리

	pm10	mise	Fflow	...	F20C21	F20C35	F20C40	F20C62
2018-04-01	58.49509	bad	5060.331	...	535	87	6872.143	90.5
2018-04-02	62.08959	bad	5999.874	...	305	202.5	8037.143	267
2018-04-03	59.4706	bad	6255.262	...	417.8	237.5	4697.429	125
2018-04-04	14.28357	good	6313.673	...	454.1667	598	6921.714	39
2018-04-05	8.193378	good	5964.006	...	894.5714	46	5708.286	94
...								

미세먼지 수치에 따라 3개의 집단으로 구분한 mise 변수 추가

- 50 이하 : good
- 50 초과 100 이하 : bad
- 100 이상 : dead






# 분산분석

## 전처리

					$\frac{X - \text{good}}{\text{good}}$				
mise	Fflow	F20C62	F25C62	...					
good	5678.057	172.9701	469.5001	...					
bad	5400.645	130.3305	442.3063	...					
dead	5182.968	260.7679	814.0938	...					



mise	Fflow	F20C62	F25C62	...
good	0	0	0	...
bad	-4.88569	-24.6514	-5.79209	...
dead	-8.71934	50.75892	73.39584	...

미세먼지 'good' 을 기준으로 그룹별 증감율 계산



# 분산분석

---

	df	sum_sq	mean_sq	F	PR(>F)
C(mise)	1.0	3.027217e+05	302721.691302	4.546535	0.034603
Residual	151.0	1.005403e+07	66582.949454	NaN	NaN

독립변수의 p-value가 0.05 이하인 변수만 추출



# 분산분석

## 분석결과

미세먼지의 영향력이 큰 변수	군집	특징
자동차정비	1군집	아파트 위주 주거지역
레저업소	3/5군집	3군집 : 관광지, 유동인구 높음 5군집 : 종로구 쪽 관광지 존재
가전	1/2/3군집	봄철 공기청정기 등 가전제품 구매
유통	3군집	실외활동보다 실내활동 선호



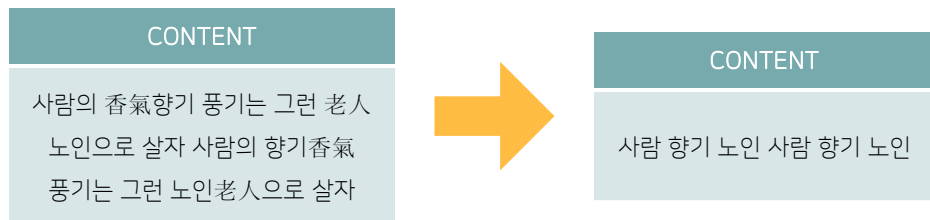
# SNS 데이터 분석

---



# SNS 데이터 분석

## 명사추출



2018-04_블로그.txt	2018-10_블로그.txt
2018-04_카페.txt	2018-10_카페.txt
2018-05_블로그.txt	2018-11_블로그.txt
2018-05_카페.txt	2018-11_카페.txt
2018-06_블로그.txt	2018-12_블로그.txt
2018-06_카페.txt	2018-12_카페.txt
2018-07_블로그.txt	2019-01_블로그.txt
2018-07_카페.txt	2019-01_카페.txt
2018-08_블로그.txt	2019-02_블로그.txt
2018-08_카페.txt	2019-02_카페.txt
2018-09_블로그.txt	2019-03_블로그.txt
2018-09_카페.txt	2019-03_카페.txt

명사추출을 위해 영문, 한자, 특수문자 등은 삭제처리  
한글과 숫자만으로 남긴 뒤 2글자 이상의 명사만 추출 후  
띄어쓰기로 구분해 파일 생성

결과파일



# SNS 데이터 분석

## 빈도분석

CONTENT
사람 향기 노인 사람 향기 노인



이름	Frequency
미세먼지	667227
피부	412327
...	...
닥트간	1
추추엘	1



이름	Frequency
미세먼지	0.916274
피부	0.583875
...	...
닥트간	1.45E-06
추추엘	1.45E-06

Komoran 형태소 분석기를 이용해  
고유명사와 일반명사로만 빈도분석

섹션별 / 계절별 / 미세먼지그룹별  
기준별로 비교를 위해  
상대적 수치로 변경



# SNS 데이터 분석

## Word Cloud - 섹션별



블로그



카페



# SNS 데이터 분석

## 비즈니스 아이디어를 위한 빈도분석

이름	Frequency
미세먼지	0.916274
피부	0.583875
사용	0.573134
제품	0.420413
생각	0.36086
아이	0.339985
오늘	0.31065
사진	0.305563
청소	0.296889
시간	0.291634
사람	0.286551
관리	0.271601
마스크	0.252269
...	...



순위	이름	Frequency
1	피부	403328
2	아이	234854
3	청소	205084
4	마스크	174262
5	필터	118033
6	공기청정기	117720
7	차량	108083
8	건강	106533
9	엄마	100960
10	실내	88081
...	...	...

비즈니스 아이디어와 연계해서 사용할 수 있는 단어 위주로 단어 재선택





# SNS 데이터 분석

## 비즈니스 아이디어를 위한 빈도분석 결과 - 섹션별

순위	이름	Frequency
1	피부	0.583875
2	아이	0.339985
3	청소	0.296889
4	마스크	0.252269
5	필터	0.17087
6	공기청정기	0.170417
7	차량	0.156466
8	건강	0.154222
9	엄마	0.146154
10	실내	0.12751
...	...	...

블로그

순위	이름	Frequency
1	아이	0.322006
2	공기청정기	0.239105
3	피부	0.21648
4	청소	0.201622
5	마스크	0.197793
6	필터	0.175705
7	차량	0.166173
8	건강	0.139862
9	실내	0.121823
10	환기	0.111336
...	...	...

카페



# SNS 데이터 분석

## Word2Vector

w2v 파라미터	내용
데이터	계절 별 명사 데이터
크기	500차원의 벡터
주변 단어	앞, 뒤 10개씩 확인
출현빈도 제한	10회 이상
사용방식	Skip-gram, Hierarchical Softmax

미세먼지와 다른 단어간의 유사도를 알아보기 위해  
단어를 벡터로 바꿔주는 알고리즘인 Word2Vector를 사용



# SNS 데이터 분석

## Word2Vector

w2v 파라미터	내용
데이터	계절 별 명사 데이터
크기	500차원의 벡터
주변 단어	앞, 뒤 10개씩 확인
출현빈도 제한	10회 이상
사용방식	Skip-gram, Hierarchical Softmax

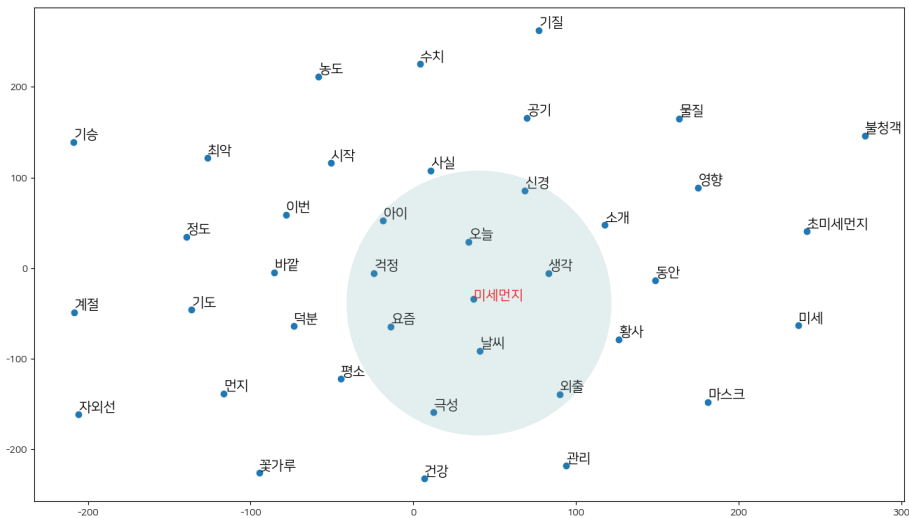
## Skip-gram

중심 단어를 기준으로  
거리가 가까운 주변 단어를  
유추하는 방식

계절별로 Word2Vector 모델 생성



# SNS 데이터 분석

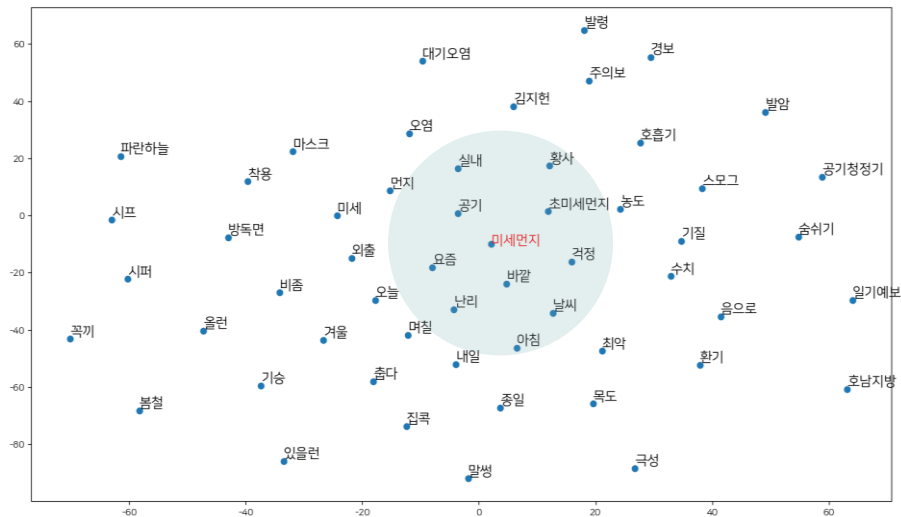


## 블로그 - 봄



# SNS 데이터 분석

## Word2Vector 결과



미세먼지와 유사한 상위 10개 단어

바깥	난리
공기	날씨
요즘	실내
초미세먼지	항사
걱정	아침

카페 - 겨울



# 비즈니스 아이디어

---



# 비즈니스 아이디어 제시

---



미세먼지 어린이 박물관



실내레저업소 추천 플랫폼



# 미세먼지 어린이 박물관

## 선정 과정

### 회귀분석 결과

변수	군집	특징
...	...	...
유통	2/4/5군집	겨울 : 밖으로 잘 안 나옴 봄 : 실외활동보다 실내활동 선호

### 분산분석 결과

변수	군집	특징
...	...	...
유통	3군집	실외활동보다 실내활동 선호

미세먼지 수치 증가 시 실내활동 증가

### Word2Vec 블로그 / 봄 결과

미세먼지와 유사한 상위 10개 단어	
날씨	신경
오늘	아이
생각	극성
요즘	사실
걱정	외출

아이는 미세먼지와 유사한 단어 중 하나





# 미세먼지 어린이 박물관

---

선정 과정

아이들에게 미세먼지에 대한 경각심을 일깨워주고

가족 단위 및 단체로 즐길 수 있는 체험형 실내공간에 대한 필요성 절감



미세먼지 어린이 박물관



# 미세먼지 어린이 박물관

## Business Canvas

<div>핵심파트너 KP</div> <div><ul style="list-style-type: none"><li>- 산림청, 정부기관</li><li>- 장소제공업체</li><li>- 환경관련 전문가</li><li>- 체험기기 대여 업체</li></ul></div>	<div>핵심활동 KA</div> <div><ul style="list-style-type: none"><li>- 전문인력과의 협업을 통한 프로그램 개발</li><li>- 온라인 홍보 및 관련 기관과의 협업</li></ul></div> <div>핵심자원 KR</div> <div><ul style="list-style-type: none"><li>- 환경 전문가</li><li>- 환경친화적 장소</li><li>- 캐릭터와 체험 기술</li></ul></div>	<div>가치제안 VP</div> <div>미세먼지 교육 프로그램</div>	<div>고객관계 CR</div> <div><ul style="list-style-type: none"><li>- 할인쿠폰 발급</li><li>- SNS 리뷰 작성 시 혜택</li><li>- 단체 할인</li></ul></div> <div>유통채널 CH</div> <div><ul style="list-style-type: none"><li>- 오프라인 체험관</li><li>- 홈페이지</li></ul></div>	<div>고객 CS</div> <div>어린이 (48개월 이상) 보호자</div>
<div>비용구조 C\$</div> <div><ul style="list-style-type: none"><li>- 체험 프로그램 개발비</li><li>- 전문인력 고용 인건비</li><li>- 대관, 대여비, 상품 제작, 홍보, 유통비 등 진행비용</li></ul></div>	<div>가치창출 V\$</div> <div><ul style="list-style-type: none"><li>- 미세먼지에 대한 어린이들의 경각심 강화</li><li>- 미세먼지 수치가 높을 때 가족단위로 내부활동이 가능한 장소 및 프로그램 제공</li></ul></div>	<div>수익원 R\$</div> <div><ul style="list-style-type: none"><li>- 관람수익</li><li>- 제휴업체와의 협업을 통한 자금 조달</li><li>- 정부지원</li></ul></div>		



# 실내레저업소 추천 플랫폼

## 선정 과정

### 회귀분석 결과

변수	군집	특징
자동차정비	1군집	아파트 위주 주거지역
레저업소	3군집	관광지, 유동인구 높음
의료기관	1/2/3군집	봄철 환절기와 미세먼지의 시너지
유통	2/4/5군집	겨울 : 밖으로 잘 안 나옴 봄 : 실외활동보다 실내활동 선호

### 분산분석 결과

변수	군집	특징
자동차정비	1군집	아파트 위주 주거지역
레저업소	3/5군집	3군집 : 관광지, 유동인구 높음 5군집 : 종로구 쪽 관광지 존재
가전	1/2/3군집	봄철 공기청정기 등 가전제품 구매
유통	3군집	실외활동보다 실내활동 선호

미세먼지 수치 증가 시 실내활동과 레저업소 매출 증가 확인



# 실내레저업소 추천 플랫폼

---

## 선정 과정

미세먼지 수치 증가 시 실내활동과 레저업소 매출 증가  
실내레저업소를 한 눈에 파악할 수 있는 플랫폼의 부재



위치기반 실내레저업소 검색 및 추천 플랫폼



# 실내레저업소 추천 플랫폼

## Business Canvas

<b>핵심파트너 KP</b>  어플리케이션 개발 업체	<b>핵심활동 KA</b> <ul style="list-style-type: none"><li>- 추천알고리즘 개발을 통한 실내레저업소 추천</li><li>- 리뷰 빅데이터 분석, 실내레저업소 제휴업체 확보</li></ul>	<b>가치제안 VP</b> <ul style="list-style-type: none"><li>- 실내레저업소를 한눈에 파악할 수 있는 플랫폼 마련으로 인한 편의성 제공</li><li>- 어플리케이션에 등록한 실내레저업소의 매출 증대</li></ul>	<b>고객관계 CR</b> <ul style="list-style-type: none"><li>- 가입 축하 선물</li><li>- 지속적인 어플 유지보수와 업데이트</li><li>- 신뢰성 있는 제휴업체 확보</li><li>- 고객 리뷰, 우수업체 선정</li></ul>	<b>고객 CS</b> <ul style="list-style-type: none"><li>- 실내레저업소를 손쉽게 이용하기를 원하는 사람(검색에 질린 사람)</li><li>- 실내레저업소 업체 대표</li></ul>
	<b>핵심자원 KR</b> <ul style="list-style-type: none"><li>- 제휴업소 정보</li><li>- IT 기술개발</li></ul>		<b>유통채널 CH</b>  어플리케이션	
<b>비용구조 C\$</b> <ul style="list-style-type: none"><li>- 어플리케이션 개발 및 운영비</li><li>- 마케팅 비용</li></ul>		<b>가치창출 V\$</b>  (소비자에게 실내레저업소에 대한 신뢰성 있는 정보를 빠르고 간편하게 제공)		<b>수익원 R\$</b> <ul style="list-style-type: none"><li>- 광고 수익</li><li>- 제휴수익</li></ul>



# THANK YOU!

