DATA STANDARDIZATION STRATEGIES FOR ANALYTICS SOLUTIONS

[분석 솔루션을 위한 데이터 표준화 전략]

NLK – [이주환, 김택관]

Project Goal

01 문제 정의 - 개요

01 문제 정의 - 목적

대회 미션 상품명을 표준화하기

대회 목적 더 나은 표준화 방안을 찾기

01 문제 정의 - 목적

비버의 웍스 에 대한 분석을 통해 문제점 및 성장 방향 찾기

Why? 데이터 표준화가 필요할까?

목표

성장 가능성이 높은 키오스크 산업에서의 **점유율 확보**

전략

빅데이터 기반 다양한 **솔루션** 제공

01 문제 정의 - 목적

더 나은 표준화 방안 찾기
일반화된 **데이터 표준화** 방안 찾기
비버웍스 기업에게 **일반화된 데이터 표준화** 방안을 찾기

01 문제 정의 - 목표

TEAM 문제 정의

어떠한 데이터가 들어와도 **데이터 표준화**가 가능한 **프로세스**

02 EDA - 시각화

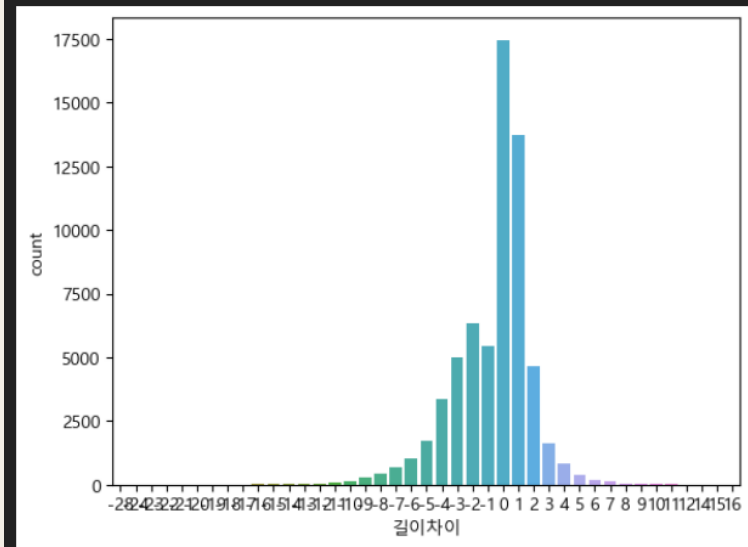
offerTrain.xlsx

	표준상품명	- 상품명	= 차이
길이	**표준상품명_길이**	**상품명_길이**	**길이차이**
요소	**표준상품명_요소**	**상품명_요소**	**요소차이**

02 EDA - 시각화

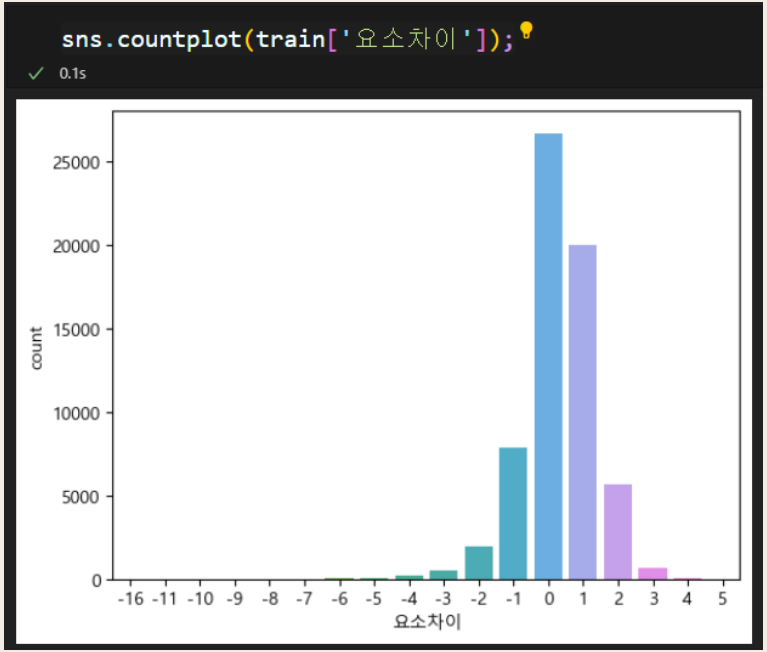
```
sns.countplot(train['길이차이']);
```

✓ 0.2s

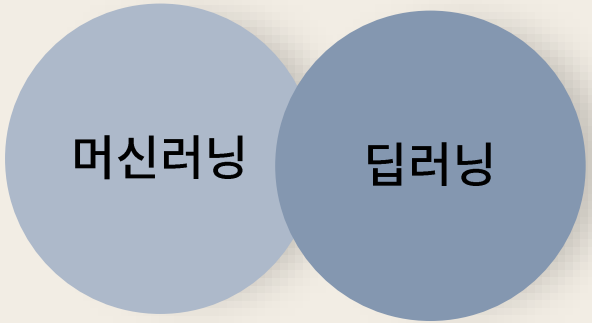



```
sns.countplot(train['요소차이']);
```

✓ 0.1s



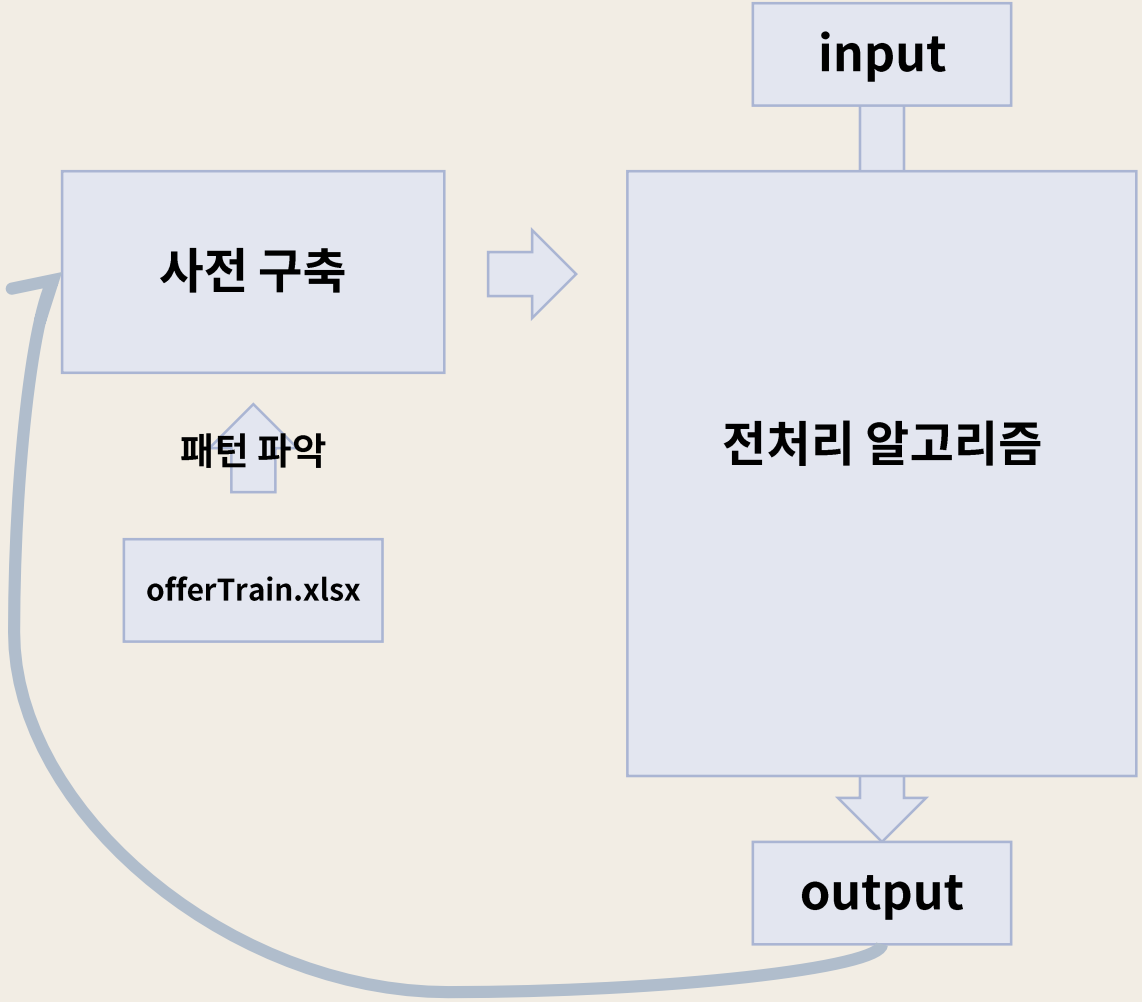
03 모델 – 선정기준



- 단기간에 좋은 결과 가능성 있음
- 시간에 따른 데이터 변화에 따라 모델을 계속 학습시켜야 함
- 문제 정의랑 어울리지 않다고 판단

- 일반화된 데이터 표준화 방안 찾기를 위한 패턴 파악 후 단어 추출
- 추출한 값으로 데이터 표준화
- 표준화 사전 구축

03 모델 – 청사진



input
돼지두부김치찌개2P

Output
메뉴명:김치찌개
맛/재료:돼지,두부
단위:None
세트옵션:None

input
아이스레몬에이드1L

Output
메뉴명:에이드
맛/재료:레몬
단위:1L
세트옵션:None

03 모델 – 사전 구축

- res_df_121
1. 길이 차이= 0, 요소 차이= 0
2. 텍스트.replace(‘ ’, ‘’) 일치

- res_df_131
1. 길이 차이= 0, 요소 차이= 0
2. 텍스트.replace(‘ ’, ‘’) 불일치
3. set(텍스트.split(‘ ’)) 일치

```
res_df_121[['상품명','표준상품명']].sample(5)
```

✓ 0.4s

	상품명	표준상품명
127748	딸기 그릭요거트	딸기 그릭요거트
424780	휘낭시에 시나몬	휘낭시에 시나몬
98559	김치찌개 정식	김치찌개 정식
356770	카페라떼	카페라떼
227060	생강 라떼	생강 라떼

```
res_df_131[['상품명','표준상품명']].sample(5)
```

.3s

	상품명	표준상품명
213	골드메달 스파클링 애플 주스	골드메달 애플 스파클링 주스
874	마카롱 돼지바	돼지바 마카롱
808	머핀 초코칩	초코칩 머핀
377	베이글 어니언	어니언 베이글
345	콜드브루 디카페인 라떼	디카페인 콜드브루 라떼

03 모델 – 사전 구축

- res_df_141
1. 길이 차이= 0, 요소 차이= 0
2. 텍스트.replace(‘ ’, ‘’) 불일치
3. set(텍스트.split(‘ ’)) 불일치
4. 상품명, 표준상품명 요소 개수 =1

- res_df_151
1. 길이 차이= 0, 요소 차이= 0
2. 텍스트.replace(‘ ’, ‘’) 불일치
3. set(텍스트.split(‘ ’)) 불일치
4. 상품명, 표준상품명 요소 개수 != 1
5. 각 요소 세트 차집합 개수 = 1

```
res_df_141[['상품명','표준상품명']].sample(5)
```
✓ 0.7s

	상품명	표준상품명
231764	슬러쉬	셰이크
260658	얼룩소프트	아이스크림
118326	닭도리탕	닭볶음탕
267051	오뎅	어묵
230267	샤인머스켓	샤인머스캣

```
res_df_151[['상품명','표준상품명']].sample(5)
```
✓ 0.4s

	상품명	표준상품명
403205	풀드포크 부리또	풀드포크 부리토
362299	카라멜 다쿠아즈	캐러멜 다쿠아즈
177945	퐁숫가루 프라페	미숫가루 프라페
418631	헤이즐럿 라테	헤이즐넛 라떼
310845	조리퐁 쉐이크	죠리퐁 셰이크

03 모델 – 사전 구축

- res_df_212
1. train['요소차이'] == -1
2. 텍스트.replace(‘ ’, ‘’) 일치
3. 상품명 요소 개수 = 2
4. 요소 세트 차집합 2, 1

```
for x, y in ls[0].unique()[:5]:
    print(x,'\t',y)
```

0.4s

```
생고기 김치찌개    생고기김치찌개
생 자몽티          생자몽티
덮밥 가라아게      가라아게덮밥
가쓰오 장국        가쓰오장국
카츠 가지          가지카츠
```

- res_df_213
1. train['요소차이'] == -1
2. 텍스트.replace(‘ ’, ‘’) 일치
3. 상품명 요소 개수 = 3
4. 요소 세트 차집합 2, 1

```
for x, y in ls[1].unique()[:5]
    print(x,'\t',y)
```

0.3s

```
생 고구마          생고구마
생 딸기            생딸기
망고 생            생망고
생 복숭아          생복숭아
생 블루베리        생블루베리
```

03 모델 – 전처리 알고리즘

	제주흑돼지등심돈가스300g
불용어	제주흑돼지등심돈가스300g
옵션처리	제주흑돼지등심돈가스300g
단위처리	제주흑돼지등심돈가스, 300g
spacing	제주 흑돼지 등심 돈가스
고유명사 제거	흑돼지 등심 돈가스
단어 표준화	돼지 등심 돈까스
데이터 표준화	메뉴명:돈까스, 맛:돼지 등심, 단위:300g

03 모델 – 전처리 알고리즘

사전	알고리즘
-	불용어
-	옵션처리
-	단위처리
Kiwi	spacing
국가, 지역, 브랜드 데이터	고유명사 제거
단어 표준화 사전	단어 표준화
-	데이터 표준화

03 모델 – 전처리 알고리즘

```
preprocessing("돼지김치찌개2인분")
```

✓ 0.3s

```
{'menu': '돼지 김치찌개',
 'std': {'메뉴': '김치찌개', '맛/재료': '돼지', '단위': '2인분', '세트/옵션': None}}
```

04 마무리