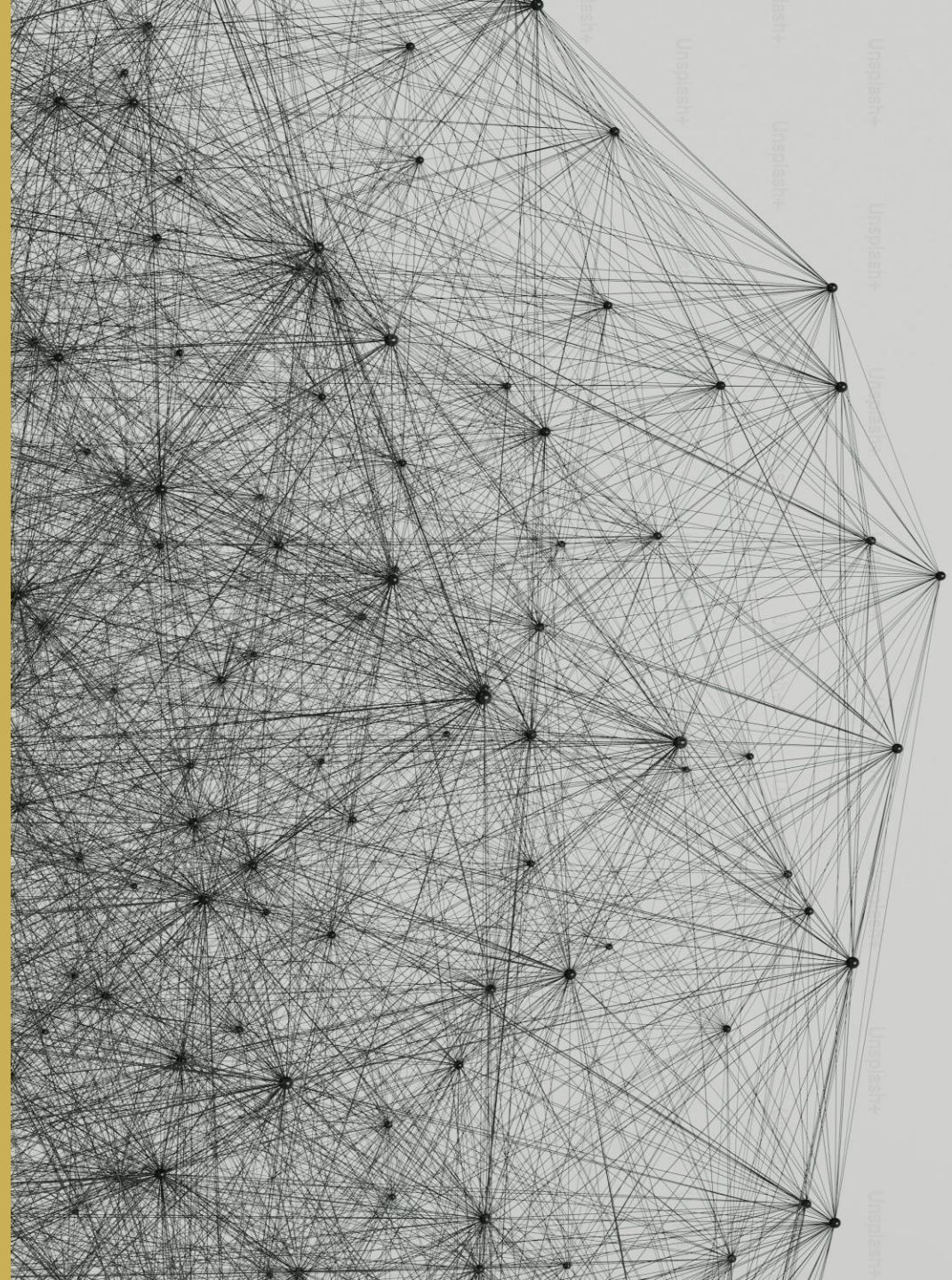
# 브랜드 이름이 포함된 검색어의 CTR 및 전환율 영향 분석

기획: 송은서 (개인 프로젝트)
코드: github
일시: 2024.11 ~ 2024.24
기술 스택: Python

# 목차

1. 프로젝트 개요
2. EDA
3. 가설 설정 및 검정
4.심화 분석
5.결론 및 제안



## 프로젝트 요약

- 브랜드 이름을 포함한 검색어가 클릭율(CTR)과 전환율에 미치는 영향 분석
- 브랜드 이름이 포함된 검색어와 비브랜드 검색어 간의 CTR 차이 검증 + EDA 통해 검색어 특성 파악하여 구매 결정에 미치는 영향 분석
- 이를 바탕으로 브랜드별 성과 평가 및 주요 성과 지표 개선을 위한 전략 제안

· 데이터 출처: Kaggle - Amazon Advertising Performance Metrics

· 기술 스택 및 도구: Python의 pandas, numpy, matplotlib, seaborn, scikit-learn, scipy 라이브러리

## 문제 정의

" 수집된 데이터에서 각 행의 고유함을 결정하는 것은 특정 기간의 검색어이다.

*그렇다면 기간별 검색어 중 무엇이 클릭율, 장바구니 전환율, 구매율의 차이를 유발하는가?"*

## 데이터 출처 및 구조

· 데이터 출처

Kaggle - Amazon Advertising Performance Metrics

· 주요 변수 및 데이터 구조

- search_query: 고객이 아마존에서 제품을 검색할 때 사용한 특정 키워드나 문구
- clk_click_rate: 광고가 표시된 횟수에 대해 클릭된 비율
- cart_add_rate: 광고가 고객의 장바구니에 제품을 추가하는 데 기여한 비율
- pur_purchase_rate : 광고가 고객의 구매로 이어진 비율
- imp_total_count: 고객에게 검색 결과나 제품 페이지에서 광고가 표시된 총 횟수
- clk_total_count: 광고가 클릭된 총 횟수
- cart_total_count: 고객이 광고를 클릭한 후 장바구니에 제품을 추가한 총 횟수
- pur_total_count: 광고를 클릭한 후 고객이 구매한 총 횟수

```
Dataset information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2589 entries, 0 to 2588
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   week                 2589 non-null   object
 1   search_query         2589 non-null   object
 2   search_query_score   2589 non-null   int64
 3   search_quey_volume   2589 non-null   int64
 4   imp_total_count      2589 non-null   int64
 5   imp_ASIN_count       2589 non-null   int64
 6   imp_ASIN_share       2589 non-null   float64
 7   clk_total_count      2589 non-null   int64
 8   clk_click_rate       2589 non-null   float64
 9   clk_ASIN_count       2589 non-null   int64
 10  clk_ASIN_share       2589 non-null   float64
 11  cart_total_count     2589 non-null   int64
 12  cart_add_rate        2589 non-null   float64
 13  cart_ASIN_count      2589 non-null   int64
 14  cart_ASIN_share      2589 non-null   float64
 15  pur_total_count      2589 non-null   int64
 16  pur_purchase_rate    2589 non-null   float64
 17  pur_ASIN_count       2589 non-null   int64
 18  pur_ASIN_share       2575 non-null   object
dtypes: float64(6), int64(10), object(3)
memory usage: 384.4+ KB
None
```

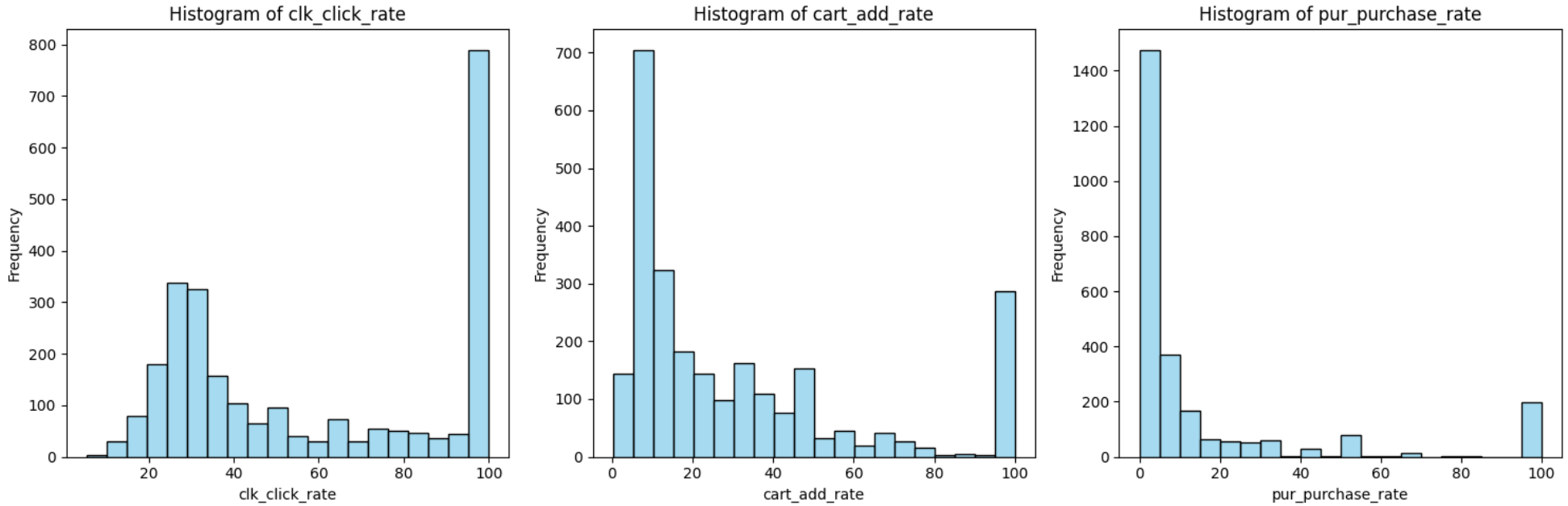
## 데이터 분석

- **결측치 처리**: 0.54%의 결측치 삭제
- **'week'열 처리**
  : 데이터의 고유성을 만드는 기준 중 하나이지만, week3에 대부분 값이 편중되어 있어 주요 분석 변수에서 제외
- **비율 데이터 수정**
  · 데이터 확인 과정에서 다음과 같은 공식 발견
    · clk_click_rate = (clk_total_count / search_query_volume) * 100
    · cart_add_rate = (cart_total_count / search_query_volume) * 100
    · pur_purchase_rate = (pur_total_count / search_query_volume) * 100
  → 모든 데이터가 이 공식 만족하는지 확인후, 50개 미만의 데이터에서 공통적으로 반올림/소수점 자리 표시 오류 있는 것 발견 후 수정
  · 'rate' 열은 비율이기에 0 ~ 100 사이의 값을 가져야 하는데, 20개 미만의 데이터에서 100을 초과하는 데이터들이 발견되어 해당 데이터 값들을 100으로 클리핑
- **ASIN 관련 열 삭제**
  : 아마존 고유 식별 번호(ASIN)의 카운트나 점유율을 나타내는 열은 본 프로젝트의 분석 목적과 관련이 없고, 다른 변수들과 중복될 가능성이 있어 삭제한 후 분석 진행

## 데이터 분석

· 변수 간의 상관관계를 보기 위해 모든 변수 간 히트맵을 생성했지만, 예상된 상관관계(광고 클릭 수와 클릭율, 총 광고 노출 수와 클릭율)만 확인되었고, 추가로 주목할만한 분석 포인트는 발견되지 않음

· 주요 지표들의 분포(clk_click_rate, cart_add_rate, pur_purchase_rate)




→ 모든 주요 비율 변수들의 분포가 한 쪽으로 치우쳐져 있음

## 데이터 분석

· **'search_queries' 열에서의 단어 카운트, 빈도 분석, 중복된 쿼리들**

→ 해당 열에 총 2575개의 단어가 있고, 평균적으로 3.9개의 단어가 존재

· **빈도 상위 10개 단어들**

→ [('toys', 1486), ('sensory', 1056), ('for', 879), ('autism', 452), ('fidget', 340), ('kids', 316), ('autistic', 256), ('stretchy', 251), ('bunmo', 207), ('children', 179), ('textured', 169), ('toy', 161), ('5-7', 136), ('strings', 133), ('adults', 108), ('special', 101), ('needs', 92), ('with', 87), ('chew', 86), ('noodles', 82)]

· **중복된 쿼리들과 그 개수**

Duplicate Query Counts:

| search_query | |
|---|---|
| autism sensory toys | 26 |
| autism sensory products | 26 |
| autism toys | 26 |
| sensory toys for autistic children | 25 |
| bunmo sensory toys | 25 |
| autistic toys for boys 5-7 | 25 |
| toys for autistic children age 5-7 | 25 |
| sensory toys | 24 |
| stretchy toys | 24 |
| sensory toys for kids 5-7 | 23 |

→ 빈도 상위 10개 단어와 중복된 쿼리 리스트를 확인했을 때, 'search_queries'열에서 자폐(autism)과 관련된 쿼리가 대부분인 것에서 수집된 데이터는 해당 필드에  초점이 맞춰져 있음을 알 수 있음

## 데이터 분석

· TF-IDF 방식을 통한 키워드 추출

- 높은 TF-IDF 점수가 나온 13개의 키워드: 'sensory', 'vibe', 'flipazoo', 'b082d7wvt8', 'speks', 'theraputty', 'slinky', 'autism', 'noodlies', 'fidget', 'stretchlerz', 'bunmo', 'figetget'

- TF-IDF 결과 분석

1. 상위 점수 30개를 추출하는 코드를 작성했으나, 13개밖에 나오지 않음
   · 가능성 1: 30개 이하의 고유한 상위 단어들이 있음
   · 가능성 2: 대다수의 키워드가 비슷한 점수를 가지고 있음
   · 가능성 3: 짧은 서치 쿼리
   · 가능성 4: 특정 토픽에 몰려 있는 데이터
   · 가능성 5: 한정된 양의 데이터
2. 'Bunmo', 'Speks','Flipazoo','Theraputty'와 같은 특정 브랜드 이름을 볼 수 있음
3. 'b082d7wvt8' , 'stretchlerz' 과 같은 특정 제품의 제품코드나 이름을 볼 수 있음

· TF-IDF 방식을 통한 키워드들의 주요지표(clk_click_rate, cart_add_rate, pur_purchase_rate) 분석

→ 해당 키워드들의 평균 클릭율, 장바구니 추가율, 구매율을 확인한 결과, '브랜드 이름'을 가진 키워드들은 그 중에서도 성과가 좋다는 것 확인

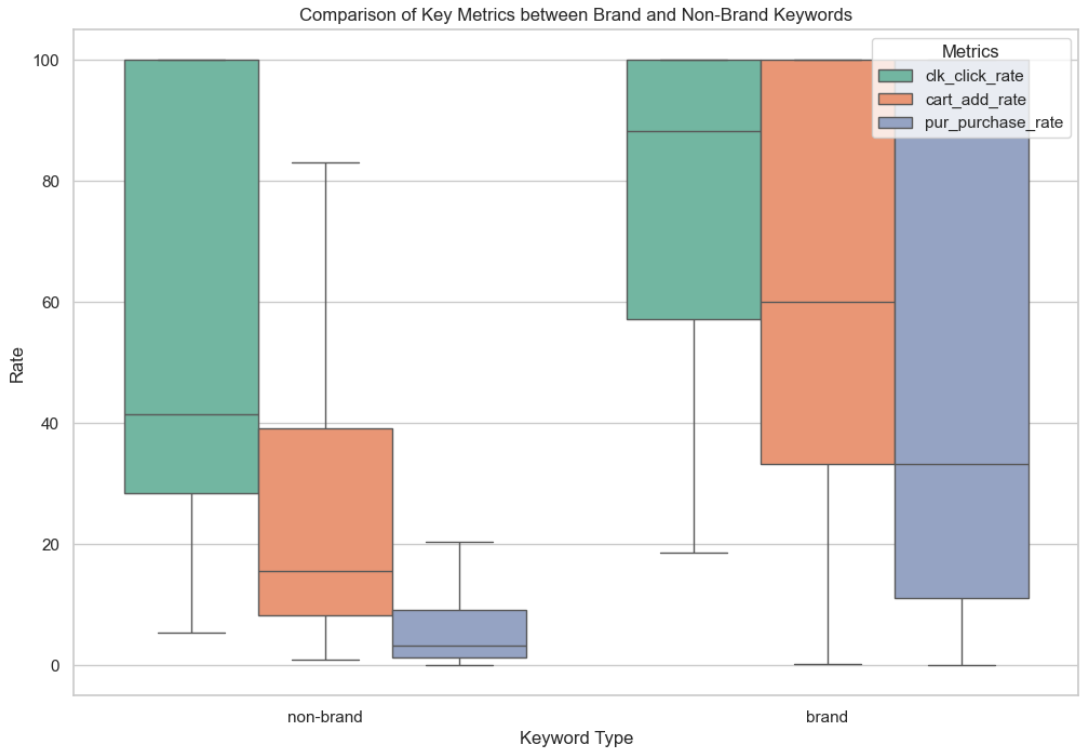
## 데이터 분석

· **브랜드 이름을 포함한 서치쿼리와 그렇지 않은 서치쿼리들의 주요 지표(clk_click_rate, cart_add_rate, pur_purchase_rate) 비교**

- 비교를 위해 'search_queries' 열에서 고유한 단어 추출하고, 생성형 AI 사용해 이 중 브랜드 이름을 추출한 리스트(brand_name) 생성
- 각 행의 'search_queries' 열의 데이터 값과 리스트를 비교하여 브랜드 이름을 포함한 서치쿼리 그룹과 아닌 그룹 나누어 박스플롯 시각화

```
brand_names = [
    'chewigem',
    'fidgetland',
    'specialkids.company',
    'bunmo',
    'speks',
    'z-vibe',
    'moluk',
    'oombee',
    'flipazoo',
    'bunmoo',
    'needoh',
    'tangle'
]
```

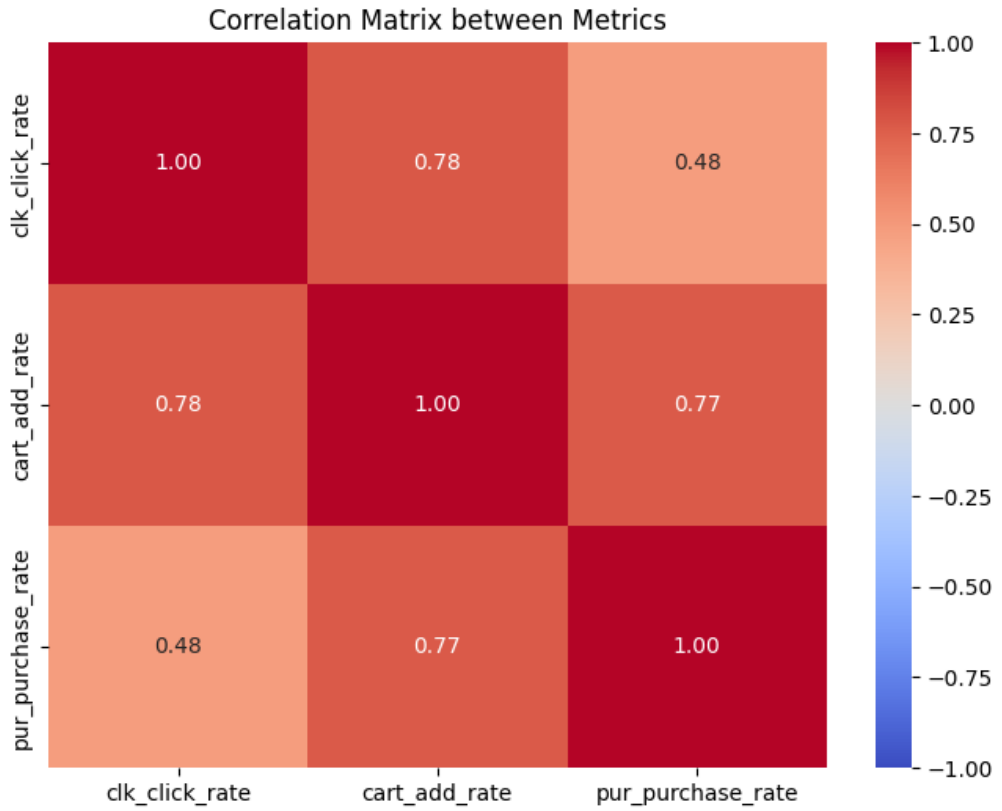
```
# Categorize queries containing brand keywords as 'brand'(queries containing brand name) or 'non-brand'(queries without brand name)
def categorize_query(query):
    for brand in brand_names:
        if brand in query.lower():
            return 'brand'
    return 'non-brand'

df['category'] = df['search_query'].apply(categorize_query)

metrics = df[['category', 'clk_click_rate', 'cart_add_rate', 'pur_purchase_rate']].melt(
    id_vars=['category'], value_vars=['clk_click_rate', 'cart_add_rate', 'pur_purchase_rate'],
    var_name='metric', value_name='rate')

plt.figure(figsize=(12, 8))
sns.boxplot(data=metrics, x='category', y='rate', hue='metric', palette="Set2", showfliers=False)
plt.title('Comparison of Key Metrics between Brand and Non-Brand Keywords')
plt.ylabel('Rate')
plt.xlabel('Keyword Type')
plt.legend(title='Metrics', loc='upper right')
plt.show()
```




→ 박스의 위치와 중간값을 통해, 브랜드 이름이 포함된 서치쿼리가 평균적으로 높은 클릭율과 전환율을 보여준다는 것을 알 수 있음

## 데이터 분석

· clk_click_rate, cart_add_rate, pur_purchase_rate 간의 상관관계




→ 광고 클릭률은 창바구니 추가율과 구매율에 유의미한 양의 상관관계를 보임

→ 클릭률은 사용자 상호작용의 첫 단계이기 때문에 클릭률을 최적화하는 것은 전환율 전반을 개선하는 데 중요한 전략임을 알 수 있음

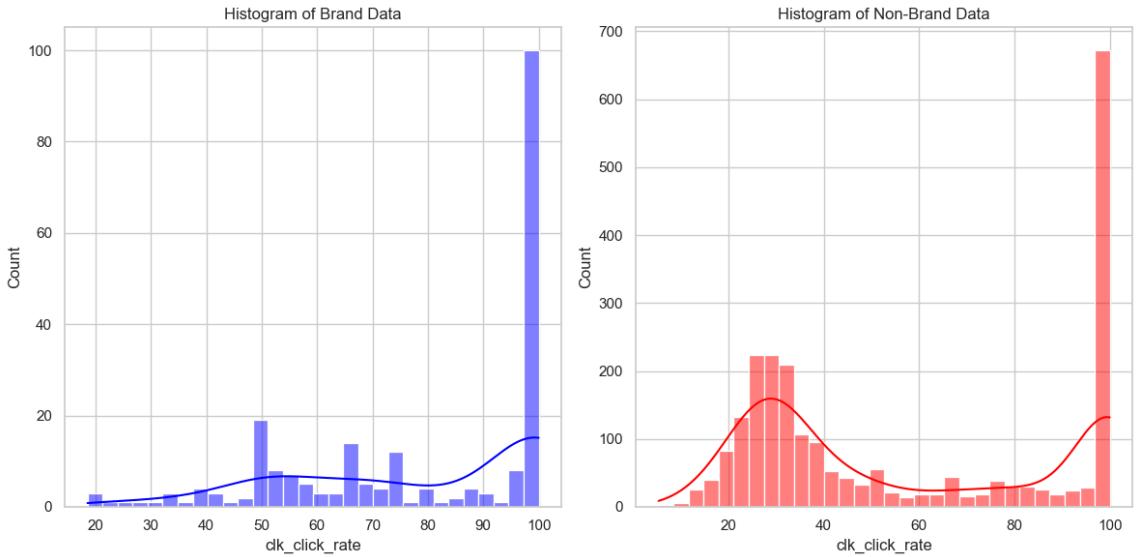
## 가설설정

귀무 가설 (H0): 브랜드 쿼리와 비브랜드 쿼리 간의 평균 CTR에 유의미한 차이가 없다.(H0: μ_brand = μ_non_brand)

·대립 가설 (H1): 브랜드 쿼리와 비브랜드 쿼리 간의 평균 CTR에 유의미한 차이가 있다.(H1: μ_brand ≠ μ_non_brand)

* 브랜드 쿼리 = 브랜드 이름을 포함하고 있는 서치 쿼리 / 비브랜드 쿼리: 브랜드 이름을 포함하고 있지 않은 쿼리

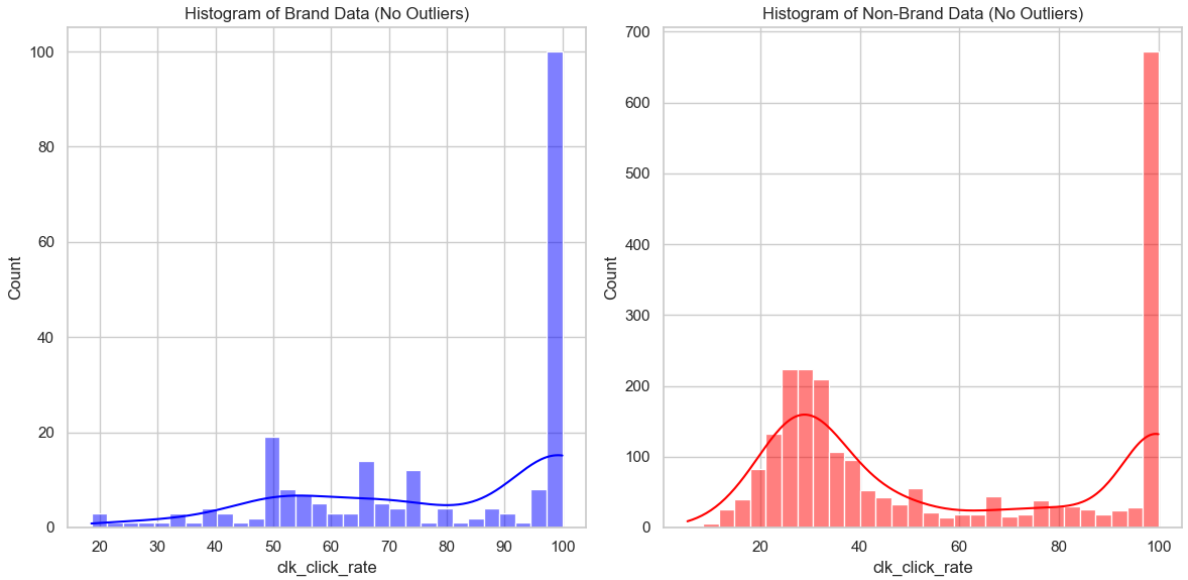
· 초기 방법론: 독립 표본 t-검정을 사용하여 두 그룹의 클릭율(CTR) 비교



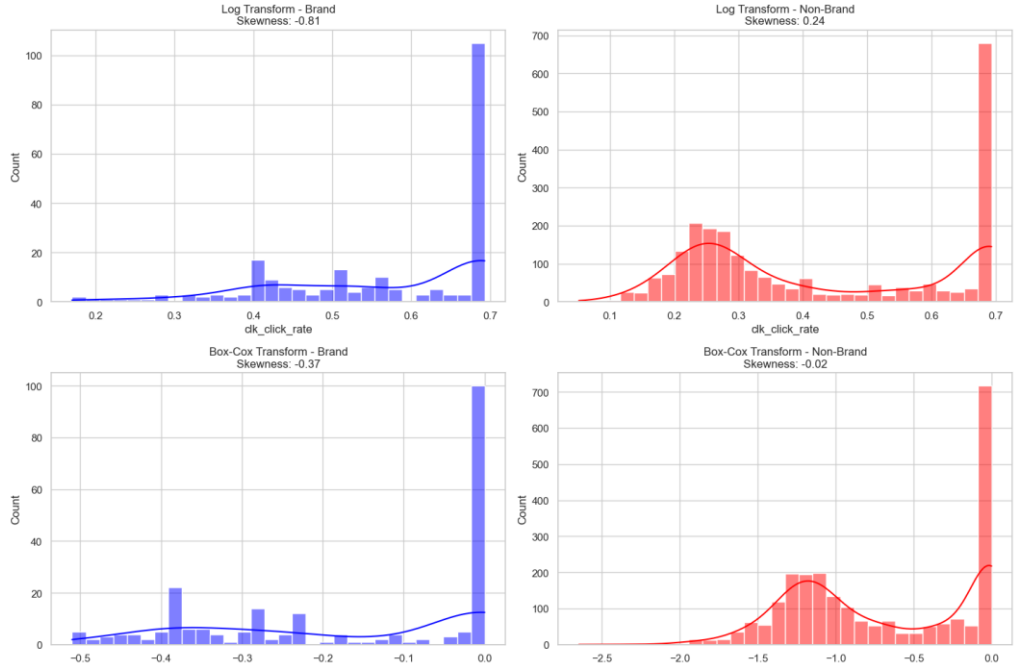

→ 독립 표본 t-검정을 위해 러프하게 정규성을 확인할 수 있게 시각화 했지만, 두 그룹 모두 우측으로 심하게 쏠려있는 것을 알 수 있음

## 정규성 확보를 위한 보정

· 아웃라이어 제거




· log, box-cox 변환




· 정규성 확보하기 위해 아웃라이어를 제거하고 로그 변환과 박스콕스 변환을 시도한 후 정규성 검사를 하였으나 만족되지 않았음
· 오버/언더샘플링도 고려하였지만, 오버샘플링의 경우 노이즈의 우려가, 언더샘플링의 경우 데이터셋 자체가 매우 작아 선택하지 않음
→ 변환을 통해서도 정규성 확보가 어려웠기 때문에, 비모수 검정 방법인 Mann-Whitney U Test를 통해 가설 검정

## 가설 검정

```python
from scipy.stats import mannwhitneyu

# Mann-Whitney U Test
stat, p_value = mannwhitneyu(brand_data_no_outliers, non_brand_data_no_outliers, alternative='two-sided')

print("Mann-Whitney U Test statistic:", stat)
print(f"P-value: {p_value:.50f}")

if p_value < 0.05:
    print("There is a significant difference between the two groups (reject H0).")
else:
    print("There is no significant difference between the two groups (fail to reject H0).")
```

```
Mann-Whitney U Test statistic: 368998.5
P-value: 0.00000000000000000000002587369537637319923878584393
There is a significant difference between the two groups (reject H0).
```

→ 가설검증 결과, 귀무가설을 기각할 수 있어 '브랜드 이름이 포함된 서치쿼리와 그렇지 않은 서치쿼리의 평균 CTR에 유의미한 차이가 있음'을 확인

## 브랜드 이름이 포함된 서치쿼리 데이터 대상 특징 분석

· 가설 검증 결과를 바탕으로, 브랜드 이름이 포함된 서치쿼리 데이터만 필터링하여 특징 분석

| brand | week | search_query_score | search_quey_volume |
|---|---|---|---|
| bunmo | 35.676329 | 37.768116 | 55.763285 |
| chewigem | 33.000000 | 58.000000 | 69.000000 |
| fidgetland | 39.000000 | 52.000000 | 1834.000000 |
| flipazoo | 26.600000 | 55.400000 | 303.800000 |
| moluk | 45.000000 | 75.000000 | 44.000000 |
| needoh | 43.000000 | 51.000000 | 43.000000 |
| oombee | 36.000000 | 55.000000 | 42.000000 |
| specialkids.company | 27.500000 | 53.000000 | 419.000000 |
| speks | 39.750000 | 66.750000 | 8911.500000 |
| tangle | 52.000000 | 100.000000 | 2319.000000 |
| z-vibe | 44.000000 | 32.000000 | 115.000000 |

| brand | imp_total_count | clk_total_count | clk_click_rate |
|---|---|---|---|
| bunmo | 1685.009662 | 33.241546 | 80.725314 |
| chewigem | 5777.000000 | 80.000000 | 100.000000 |
| fidgetland | 40591.000000 | 839.000000 | 45.750000 |
| flipazoo | 12929.200000 | 180.800000 | 76.392000 |
| moluk | 961.000000 | 20.000000 | 45.450000 |
| needoh | 858.000000 | 13.000000 | 30.230000 |
| oombee | 2140.000000 | 48.000000 | 100.000000 |
| specialkids.company | 12937.000000 | 79.000000 | 18.880000 |
| speks | 195551.500000 | 2947.500000 | 47.095000 |
| tangle | 52253.000000 | 871.000000 | 37.560000 |
| z-vibe | 4269.000000 | 49.000000 | 42.610000 |

| brand | cart_total_count | cart_add_rate | pur_total_count |
|---|---|---|---|
| bunmo | 16.227053 | 65.230145 | 6.623188 |
| chewigem | 24.000000 | 34.780000 | 8.000000 |
| fidgetland | 99.000000 | 5.400000 | 18.000000 |
| flipazoo | 29.400000 | 13.044000 | 4.200000 |
| moluk | 9.000000 | 20.450000 | 4.000000 |
| needoh | 5.000000 | 11.630000 | 2.000000 |
| oombee | 27.000000 | 64.290000 | 7.000000 |
| specialkids.company | 5.000000 | 1.135000 | 1.000000 |
| speks | 866.750000 | 12.120000 | 97.000000 |
| tangle | 283.000000 | 12.200000 | 60.000000 |
| z-vibe | 15.000000 | 13.040000 | 6.000000 |

| brand | pur_purchase_rate | word_count | count |
|---|---|---|---|
| bunmo | 47.610628 | 7.599034 | 207 |
| chewigem | 11.590000 | 1.000000 | 1 |
| fidgetland | 0.980000 | 1.000000 | 1 |
| flipazoo | 1.702000 | 1.000000 | 5 |
| moluk | 9.090000 | 3.000000 | 1 |
| needoh | 4.650000 | 2.000000 | 1 |
| oombee | 16.670000 | 2.000000 | 1 |
| specialkids.company | 0.240000 | 1.000000 | 2 |
| speks | 1.550000 | 1.000000 | 4 |
| tangle | 2.590000 | 3.000000 | 1 |
| z-vibe | 5.220000 | 1.000000 | 1 |

## 브랜드 이름이 포함된 서치쿼리 데이터 대상 특징 분석 - 결과

| | |
|---|---|
| Bunmo | · 검색량은 중간수준(55.77회)였지만, 클릭률(80.72%), 장바구니 추가율(65.23%), 구매율(47.61%)에서 최고 성과<br>· 전체적으로 균형 잡힌 퍼널 성과 보이며, 특히 최종 구매 전환율이 매우 높음 |
| Chewigem, Oombee | · Chewigem: 클릭률 100%를 기록했지만, 장바구니 추가율(34.78%)과 구매율(11.59%)이 급격히 감소<br>· Oombee: 클릭률 100%를 기록했지만, 장바구니 추가율(64.29%)과 구매율(16.67%)이 급격히 감소<br>· 두 브랜드 모두 다른 브랜드에 비해 전반적으로 높은 성과를 보였지만, 장바구니 추가와 구매로 이어지지 않음<br>→ 광고 클릭 후 전환 포인트에서 문제가 발생하며, 초기 관심을 최종 구매로 연결할 수 있는 원인 파악과 해결책 필요 |
| Fidgetland, Tangle, Speks | · 검색량이 높은 그룹(Fidgetland, Tangle: 1000 이상, Speks: 8000 이상)이지만, 클릭률은 중간 수준에 그치고 장바구니 추가율과 구매율은 낮음<br>→ 브랜드 인지도나 관심도는 높지만 실제 구매로 이어지지 않으므로,<br>구매 결정을 촉진할 수 있는 세일즈 페이지 최적화 및 제품의 실제 가치를 강조하는 마케팅 전략 개발 필요 |
| Flipazoo, Moluk, Needoh, Specialkids, Z-vibe | · 검색량, 클릭률, 장바구니 추가율, 구매율 모두 낮은 수준을 보임<br>→ 전반적인 마케팅 전략에 대한 포괄적 검토 필요 |

## 결론

**1. 키워드 분석 결과**

· 수집된 키워드 대부분이 '자폐증(autism)'과 관련된 것으로 나타남

· 본 프로젝트를 사용해 Amazon에서 'autism' 키워드 관련 고객에게 타겟팅 가능

**2. CTR 분석 결과(Mann-Whitney U Test)**

· 브랜드 이름을 포함한 서치쿼리와 그렇지 않은 서치쿼리 간 CTR의 유의미한 차이 발견

→ 브랜드 이름을 포함한 서치 쿼리가 사용자의  클릭에 영향을 주는 관심을 더 끌며, 광고 타켓팅 전략에 활용 가능

**3. 브랜드별 퍼널 성과 비교**

· Bunmo: 균형 잡힌 성과 및 높은 전환율

· Chewigem & Oombee: 초기 관심은 높으나 최종 구매율 낮음

· Fidgetland, Tangle, Speks: 브랜드 인지도는 높으나 구매율 저조

· Flipazoo 등: 전반적으로 낮은 성과, 포괄적 전략 검토 필요

→ 브랜드별 퍼널 약점에 맞춘 맞춤형 전략 필요성 강조

## 한계점 및 제안

<한계점>

**1. 데이터 불균형 및 왜곡 문제**

· 어려움: 데이터가 심하게 왜곡되어 있어 변환을 시도해도 정규성을 확보할 수 없었음

· 제안: 데이터의 수가 많으면 정규성을 확보할 수 있는 언더/오버 샘플링이 가능하므로 많은 수의 데이터를 수집하거나.
수집 단계에서 데이터의 불균형을 덜 수 있는 전략 또는 구조 설계

**2. 데이터의 단순성**

· 어려움: 서치쿼리에 따른 행동 카운트를 바탕으로 한 비율이 중심인 단순한 수치 데이터

· 제안: 고객의 인구 정보, 행동 정보 등의 다양한 변수를 포함한 데이터를 추가하여 분석 결과의 신뢰성과 세부 사항 강화

<그 외 분석 고려사항>

· 본 프로젝트는 아마존 고유 식별 번호(ASIN)과 관련된 열을 분석 대상에서 제외했으나, ASIN과 서치쿼리 간의 관계를 분석함으로써 추가적인 인사이트를 도출할 가능성이 있음(e.g., ASIN별로 사용자 관심도가 다를 수 있으며, 이는 서치쿼리와의 상관관계 통해 확인 가능)

· 보완된 데이터셋을 통해, 클릭율과 전환율을 예측할 수 있는 예측모델 만들기 가능

# Thank you