쇼핑 트렌드 데이터를 이용한 시즈널리티 분석

유통 | scipy.stats | Tableau | Python

프로젝트 개요

데이터 개요

- 의류사 고객별 인적 정보(나이, 성별 등), 구매 제품 정보(상품 분류, 사이즈, 색상 등), 구매 지역 및 계절 정보, 결제 정보(결제 수단, 할인 코드 적용 여부 등) 등에 대한 field로 구성됨.

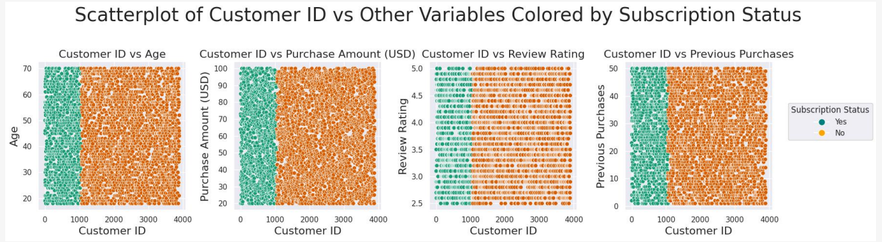
분석 목표

- 지역(U.S. State)별 계절별 구매건수가 지역별로 서로 다른 양상을 보이는 데 착안하여, 이를 외부 데이터인 미국 주별 기후 데이터와 결합하여 **의류 구매 양상의 시즈널리티 분석**을 진행하고자 함.

분석 한계점

- 초반 다양한 방향으로 EDA를 진행했으나 눈에 띄는 트렌드가 보이지 않음.
- 확인 결과 원본 데이터는 ChatGPT로 만들어진 것이었고 이 과정에서 분포가 너무 균일하게 조작된 것으로 보임. (우측 그림 참고)
- 이러한 한계점은 있었지만 해당 프로젝트에서는 다양한 수단을 동원하여(외부 데이터 첨가, 시각화와 수치 분석 병행) 분석 목표를 Drill Down하는데 초점을 맞추어 진행함.

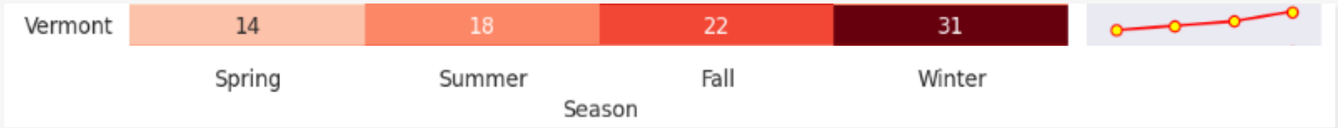
This dataset is a synthetic creation generated using ChatGPT to simulate a realistic customer shopping experience. Its purpose is to

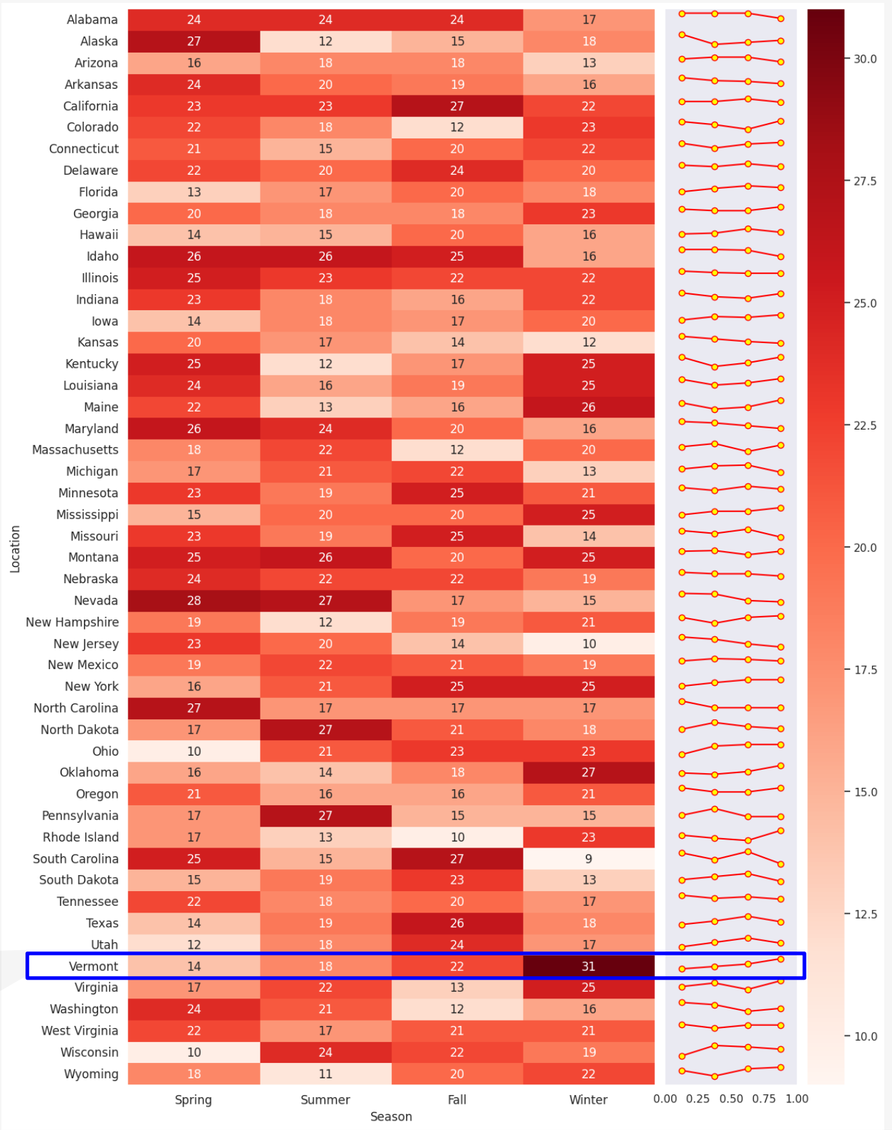


쇼핑 트렌드 데이터를 이용한 시즈널리티 분석

계절별 지역별 구매건수 분석

- 미국 50개 주(State)의 **계절별 구매건수**를 Pivot Table로 집계하고 이를 Heatmap으로 시각화 함. 구매건수의 경향성을 보다 한 눈에 파악하기 위하여 Line Chart도 추가하여 데이터를 다차원적으로 살펴보았음.
- 주별로 상이한 경향성을 보임을 확인할 수 있었음. 그리고 **겨울철 가장 높은 구매건수를 보이는 Vermont 주가 지리적으로 미국 최상단에 위치**한다는 점에 착안하여, 이를 미국의 기후 데이터와 연관지어 분석하면 유의미한 결론에 도달할 수도 있겠다는 아이디어를 도출함.

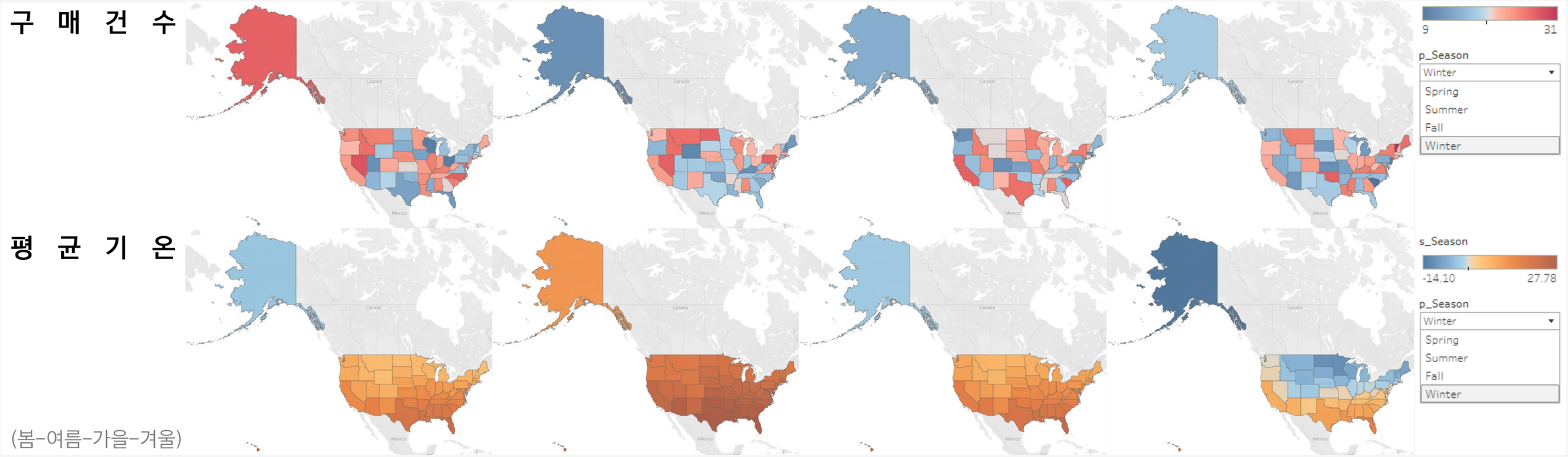




쇼핑 트렌드 데이터를 이용한 시즈널리티 분석

지역별 기후 및 구매건수 분석

- 원본 데이터에는 기후 정보가 없었기에 1991-2020 미국 주(State)별 월별 기온을 기록한 외부 데이터를 수집하여 이를 계절별로 가공함. (집계 방식: 평균) (데이터 원본 출처: NOAA, 미국 해양대기청)
- Tableau를 활용하여 구매건수와 평균기온 트렌드를 Geo Chart로 시각화 함. 이 때 **계절은 매개변수로 지정하여 드롭박스 형태로 선택 가능**하도록 설계함. 그러나 시각적인 정보만으로는 한 눈에 들어오는 상관관계를 파악하기 어려웠고 따라서 상관계수 분석을 결정함.



구매건수 차트 Tableau Public Link: https://public.tableau.com/app/profile/ha.eun.lee2020/viz/us_states_purchase_count_by_season_ver2/PurchaseCount
평균기온 차트 Tableau Public Link: https://public.tableau.com/app/profile/ha.eun.lee2020/viz/us_states_mean_temp_by_season_ver2/MeanTemp

쇼핑 트렌드 데이터를 이용한 시즈널리티 분석

상관계수 분석 및 개선점 도출

상관계수 분석

- 아래 표와 같이 계절별 구매건수와 기후지표 (총 5개) 간 상관계수를 도출함.
- 각 기후지표별로 1개 정도 주에서는 유의미한[1] 상관관계를 보였지만, 미국 전체 주 개수(50개)를 감안하면 해당 지표가 구매건수와 상관관계가 있다고 주장하기는 어려움.

구매건수 vs	State	Correlation	P-value
평균기온	Maine	-0.956043	0.043957
최고기온		null	
최저기온	Maine	-0.965747	0.034253
일교차	Missouri	0.962515	0.037485
기온차[2]	Nevada	0.972993	0.027007

1) 기준: P-value < 0.05
2) (현재 계절 평균기온) – (이전 계절 평균 기온)

분석 한계점 및 개선점 도출

1. 원본 데이터셋의 row 개수는 3,900개로 이를 50개 State와 4개 계절로 나누면 각 구분 당 평균 20개 내외의 데이터만을 가지게 되고, 실제 Pivot Table 집계 결과도 이와 유사한 수치를 보임. → 경향성 분석 및 유의성 판단에는 데이터 개수가 충분하지 않다고 볼 수 있음.
2. 미국 50개 주 중 절반 이상이 남한보다 면적이 넓은 바, 우리나라의 도시별 기후 분석과 같은 개념을 미국의 주별 기후 분석에 적용시키기에는 무리가 있을 가능성이 큼.
3. 활용한 기후 데이터는 1991-2020의 수치를 평균한 것으로, 원본 데이터셋을 수집한 특정 년도의 기후 특성을 제대로 반영하고 있다고 보기 어려움. (원본 데이터셋 수집 년도는 공개되지 않음)

⇒ **1. 더 큰 규모의 데이터셋 활용 2. City별로 세분화된 데이터셋 활용 3. 데이터셋 수집 년도의 기후 데이터 활용** 등의 조건을 갖춘다면 더 유의미한 분석이 가능할 것으로 보임.

주(State)별 – 계절별 20개 내외의 표본

Heatmap of Purchase Count by Season and State

Alabama	24	24	24	17
Alaska	27	12	15	18
Arizona	16	18	18	13
Arkansas	24	20	19	16
California	23	23	27	22
Colorado	22	18	12	23

영토 크기 비교