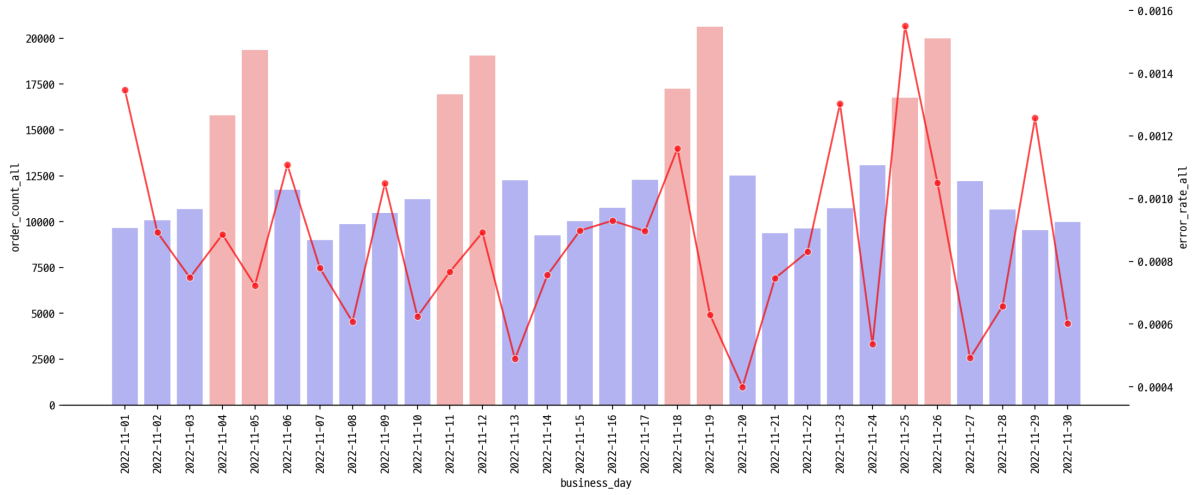
분석 과제 요약 - 신상훈

1. 일별 데이터 그래프 시각화

주문에 대한 기본 정보는 다음과 같습니다.

- 일일 평균 주문 수: 12,691
- 주중 일일 평균 주문 수: 11,500
- 주말 일일 평균 주문 수: 15,968

11월 한달 동안의 주문 추이는 일주일 단위로 패턴을 보이는 것을 확인할 수 있습니다.



빨간색은 주말(금, 토) 파란색은 평일(일~목)을 표현하고 있습니다. 매주 토요일에 주문이 가장 많이 발생했으며, 일요일보다는 금요일에 주문이 많이 발생한 것을 확인할 수 있습니다. 티오더 태블릿이 외식업체에 많은 것으로 미루어봤을 때, 일요일보다는 금요일에 약속을 많이 잡아서 그렇지 않을까 조심스럽게 추측해볼 수 있을 것 같습니다. 일별 에러율은 11월 25일에 크게 증가하는 모습을 보였지만 그 외에는 특별한 추이를 보이지 않습니다.

2. 12월 1일 ~ 3일까지 주문수 예측

주문 수는 시계열 데이터로 표현이 되고 있으며 이를 예측하는 방법은 다양합니다. 단순 평균 계산, 전통적인 시계열 통계 모델(Moving Average, Exponential Smoothing, ARIMA), Prophet 모델, 회귀분석, Seuqence 데이터를 다루는 딥러닝 모델(RNN, LSTM, Transformer)을 사용할 수 있습니다. 이 외에도 여러 예측 방법론이 존재하겠지만 이번 분석에서는 간단하게 몇가지 분석 방법만 사용해보도록 하겠습니다.

1) 단순 평균 계산

11월 주문 수의 추이를 봤을 때는 금요일 ,토요일에 대한 효과가 주기적으로 나타나는 것을 확인할 수 있습니다. 12월 1~3일은 목, 금, 토요일입니다. 11월의 목, 금, 토요일 데이터를 평균하여 대략적인 주문 수를 추측해볼 수 있습니다.

2) 시계열 모델을 활용한 예측

SARIMA(Seasonal Autoregressive Integrated Moving Average)는 계절성을 포함하는 시계열 모델입니다. 계절성이라는 것은 어떤 특정 패턴이 주기적으로 나타나는 것을 의미합니다. 주문 데이터의 경우, 1주를 주기로 패턴이 발생하기 때문에 SARIMA 모델을 사용해서 예측을 진행하였습니다.

3) LSTM을 활용한 예측

딥러닝에서 sequence 데이터를 다룰 때 사용하는 모델인 LSTM을 사용해서도 주문 수를 예측할 수 있습니다. 간단하게 모델링을 하여 예측을 진행하였습니다.

4) 결과

단순 평균, 시계열 모델, 딥러닝 모델을 사용한 예측 결과는 아래와 같습니다.

날짜	단순 평균 모델	시계열 모델	딥러닝 모델
2022-12-01	11,811	13,069	12,317
2022-12-02	16,686	16,756	17,145
2022-12-03	19,756	19,991	19,654

3. 하루 주문 중 주류가 가장 많이 팔렸을끼?

0) 하루 주문 중 주류가 가장 많이 팔렸다의 의미

주류가 가장 많이 팔렸다는 것에 대한 상세한 정의가 필요합니다. 판매 금액이 높은건지, 판매 수량이 많은건지 이러한 기준을 명확하게 하기 위한 정의가 필요합니다. 하루라는 단어도 비슷한 맥락에서 정의가 필요합니다. 이번 분석에서는 판매 금액, 판매 수량을 모두 고려하였으며, 하루라는 의미는 단순하게 주중, 주말로 나눠서 진행하였습니다.

1) 카테고리 분류하기

주류가 가장 많이 팔렸을 것이라는 주장을 데이터로 뒷받침 하기 위해서는 주류와 주류가 아닌 항목들에 대해서 그룹핑을 해서 데이터를 집계할 수 있어야 합니다. 이를 위해서 주문 품목에 대한 카테고리가 있어야합니다.

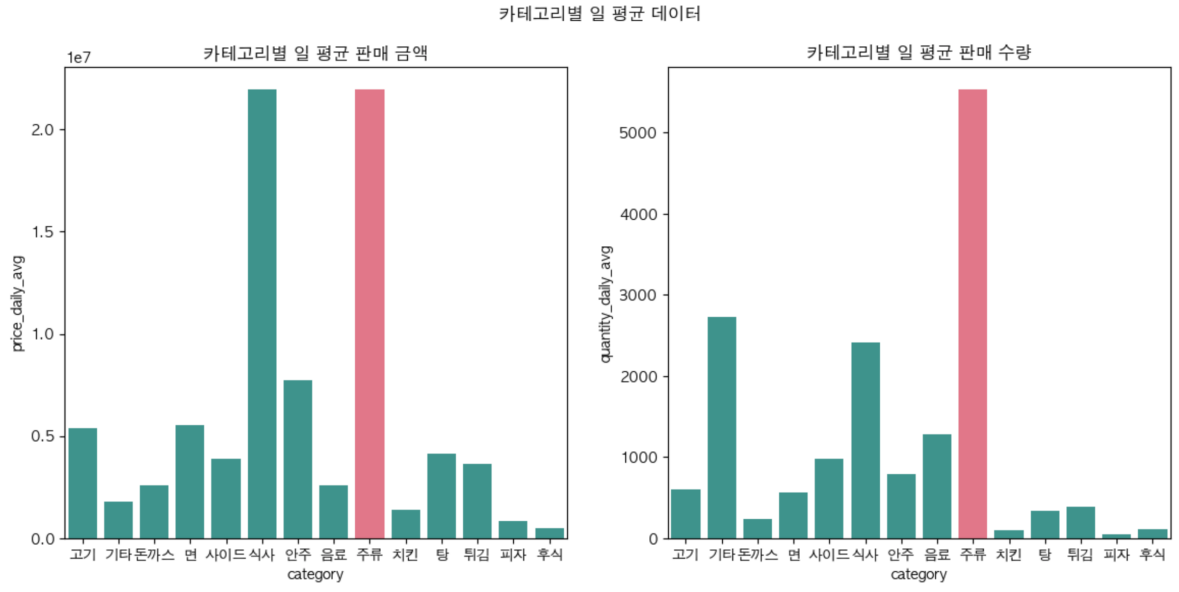
현재 데이터의 `good_category` 컬럼은 명확하게 주류를 묶을 수 있는 컬럼은 아닙니다. 왜냐하면 해당 컬럼에 주류와 관련해서 주류, 맥주, 소주, 맥주&소주, 하이볼, 막걸리 등 여러 값이 존재하기 때문입니다. 주류뿐만 아니라 비주류 항목도 같은 현상을 보이고 있습니다. 주류가 가장 많이 팔렸다는 것을 데이터로 확인하기 위해서는 단순하게 주류, 비주류에 대한 구분으로 끝나서는 안되고 비주류에 대해서도 세세한 카테고리화가 필요합니다. 그래야 다른 품목들에 비해 주류가 많이 팔린다는 것에 대한 적합한 근거가 될 수 있습니다.

카테고리를 하는 작업을 위해서 몇 가지 방법이 시도될 수 있습니다. 예를 들면, 아래와 같은 방법들이 있습니다.

1. 직접 카테고리를 분류한다.
2. 임베딩된 모델을 사용해서 벡터화를 시킨 후 클러스터링을 한 뒤 라벨링을 한다.

현재 과제에서는 시간 관계상 직접 카테고리를 입력하기로 하였습니다. 단, 전체 데이터를 대상으로 하기엔 시간이 많이 소요되기 때문에 일부만 추출하여 진행하였습니다. 기존 주문 상품 정보 테이블의 전체 데이터 수는 622,465개이고 유니크한 카테고리 수는 3,564개 입니다. 상위 100개 카테고리에 대해서만 수기로 카테고리 작업을 진행하였습니다. 상위 100개의 카테고리는 398,040개이므로 전체 데이터의 약 64%를 차지합니다.

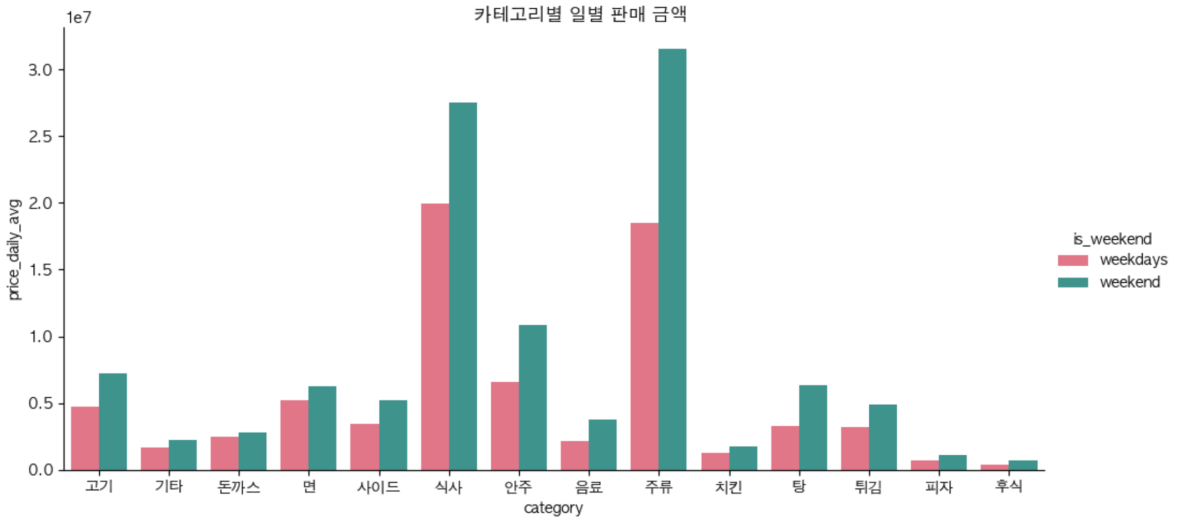
2) 일별 데이터 확인하기

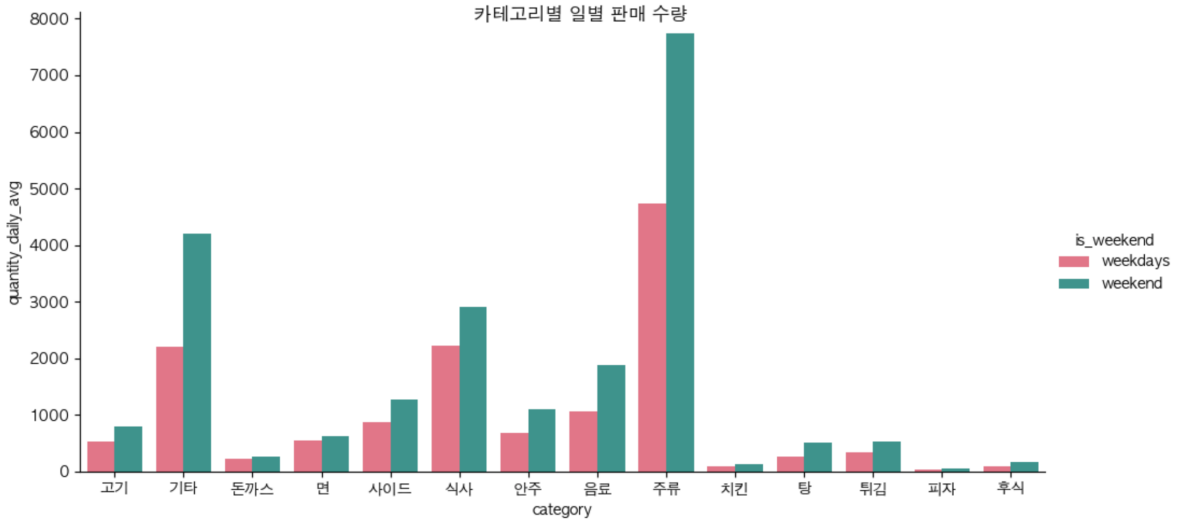


order_contents_information 데이터와 수기로 작성한 category 데이터를 합쳐서 일별 판매 금액, 판매수량을 확인해보았습니다. 일 평균 판매 금액과 판매 수량 모두 주류가 다른 카테

고리에 비해 가장 높은 값을 보인 것을 확인할 수 있었습니다. 판매 금액 기준으로 식사 카테고리가 주류와 비슷하게 많이 팔린 점도 확인할 수 있습니다.

하지만 1번 일별 주문 추이 그래프를 다시 생각해보면, 주문은 주중(일~목), 주말(금,토) 확연하게 차이가 난다는 것을 확인했었습니다. 그렇다면 여기서 한 걸음 더 나아가서 주중과 주말의 일별 데이터를 확인하면 조금 더 의미있는 결과를 얻을 수도 있을 것 같다는 생각이 들었습니다.





예상처럼 주중, 주말 모두 주류가 매우 높은 값을 보여주었습니다. 그리고 몇 가지 특이점도 확인할 수 있었습니다.

- 식사류의 경우 주중 일 평균 판매 금액이 가장 높게 나타난다.
- 주류의 경우, 주말과 주중의 판매 금액 차이가 많이 나는 편이다.