1. 문제를 정의하고 큰 그림 보기

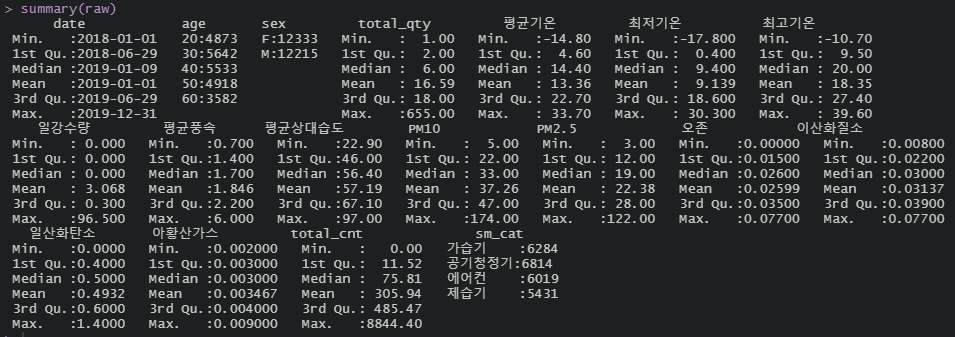
모델을 학습시켜 새로운 데이터가 주어졌을 때 실제로 얼마나 팔릴지 예측하는 것이 목표입니다. 품목은 비교적 날씨와 연관성이 클 것이라고 생각되는 에어컨, 가습기, 제습기, 공기청정기 이며, 각 제품들을 구매할 때 날씨적 특성과 SNS마케팅에 얼마나 영향을 받는지를 파악하여 SNS광고 유무와 어떤 기상상태일 때 얼마나 팔릴지 예측해 보는 것이 목적입니다. 그러므로 이 모델은 레이블이 판매건수인 수치예측 지도학습 모델입니다.

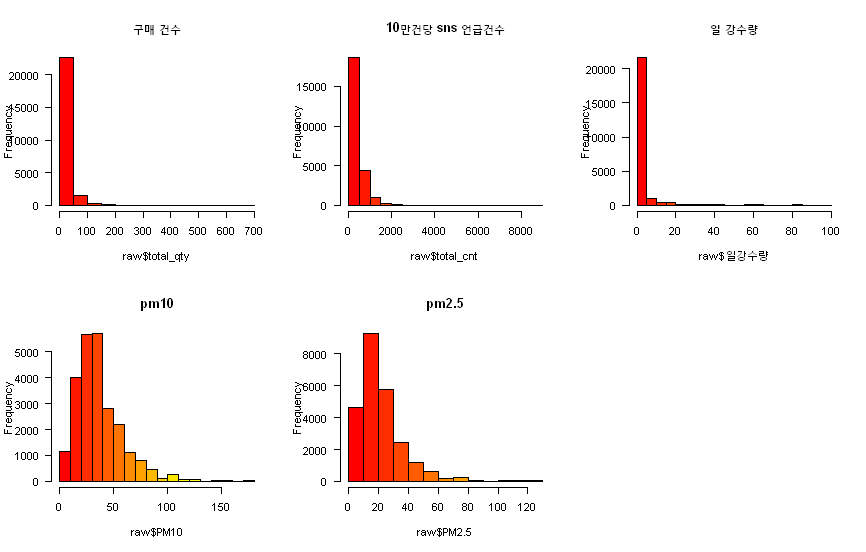
2. 데이터 가져오기

데이터는 날씨마루 공모전, 기상청에서 제공한 데이터로 소셜 데이터, 온라인 구매이력 데이터, 날씨 데이터로 나눠집니다. 모든 데이터는 2018.01.01 ~ 2019.12.31 2년간 수집한 자료입니다.

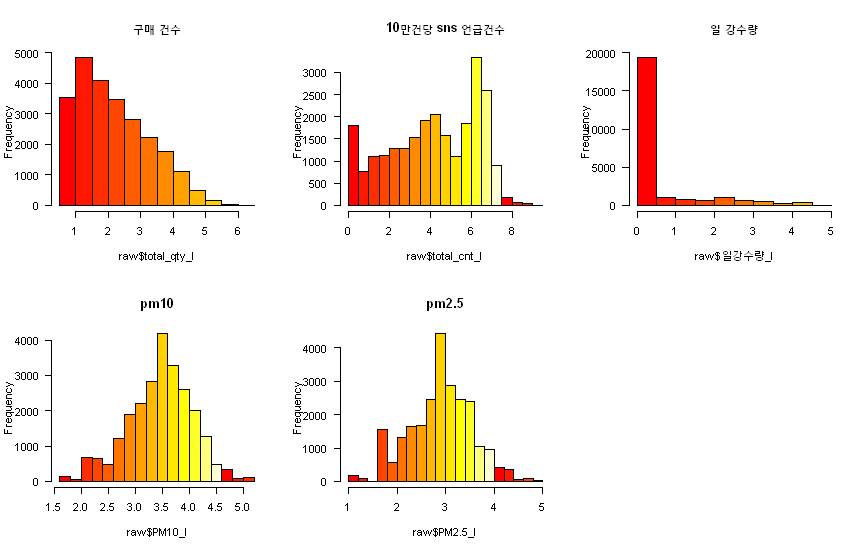
먼저 소셜 데이터는 ㈜바이브컴퍼니에서 제공한 자료로 블로그, 커뮤니티, 인스타그램을 분석하여 각 품목별 10만건당 언급 건수를 변수로 포함하고 있으며 여기서 10만건당 언급건수는 서로 다른 유저 수를 보유한 채널별 문서수의 격차를 보정하기 위해 상대적 문서수로 나타낸 값입니다. 각 채널별의 버즈량을 10만건 대비한 문서수로 살펴봄으로써 채널간의 문서수 차이나 채널안에서 발생하는 전체 문서수의 변동성에 비해 안정적으로 해당 키워드에 해당하는 문서 수를 확인할 수 있습니다. 온라인 구매이력 데이터 구매변수 (구매 날짜, 구매자 성별, 구매자 연령대, 품목별 구매 건수), 날씨 데이터는 일별 날씨변수(평균기온, 최저기온, 최고기온, 일강수량, 평균풍속, 평균상대습도, PM10, PM2.5, 오존, 이산화질소, 일산화탄소, 아황산가스)들을 포함하고 있습니다.

세가지 테이블을 합치는 과정에서 발생한 결측치는 제거하였으며 date변수는 날짜형으로, 구매성별, 구매 나이, 구매 품목은 범주형 변수로 변경 후 summary를 통해 각 변수들의 특성을 확인하였습니다.

 확인해 본 결과, 범주형 변수에서의 심한 데이터 불균형은 발생하지 않은 것으로 보이며, 판매 건수(total\_qty), 일강수량, 10만건당 언급건수(total\_cnt), 미세먼지(PM10, PM2.5)에서 치우침 및 이상치가 발생한 것으로 보입니다. 조금 더 시각적으로 해당 데이터들의 모양을 확인해 보도록 하겠습니다.



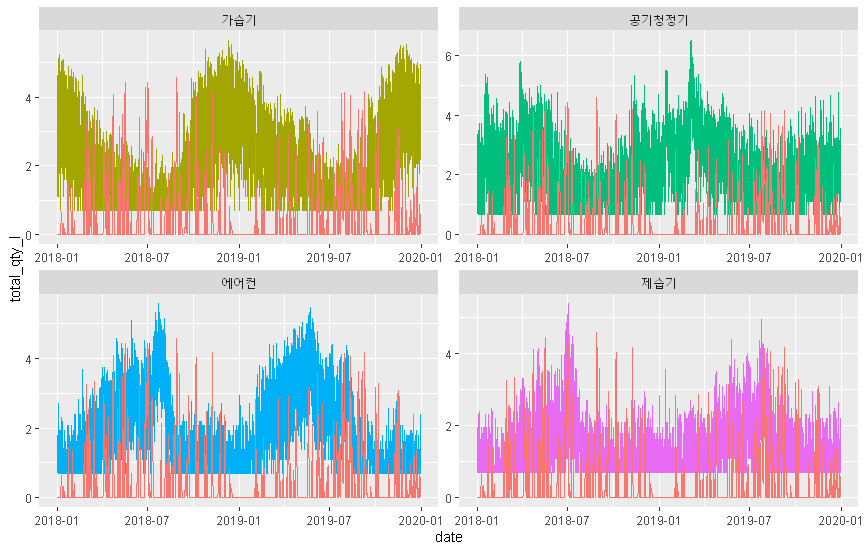
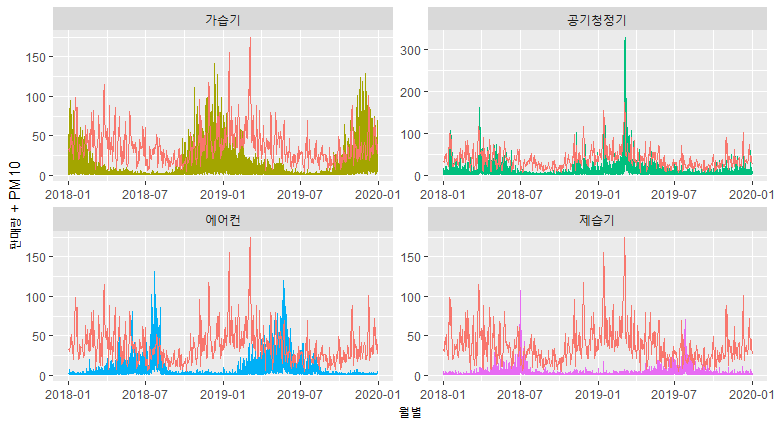
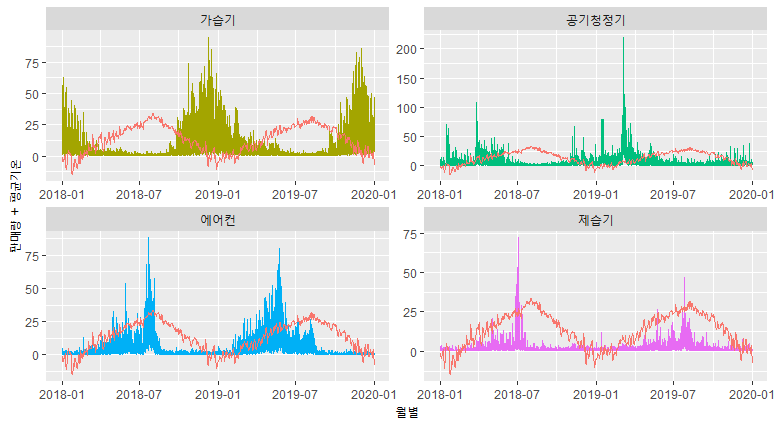
확인 결과 구매건수, SNS언급건수, 일 강수량에서는 심각한 치우침이, 미세먼지에서는 치우침이 발견되었습니다. 머신러닝 알고리즘은 정규분포나 가우시안 분포에서 더욱 쉽게 패턴을 찾을 수 있으므로 해당 변수들에 log를 취하여 분포 모양을 종모양에 가깝게 변형하도록 하겠습니다.



분포가 아까보다는 조금 더 종모양으로 바뀐 것을 확인할 수 있습니다.

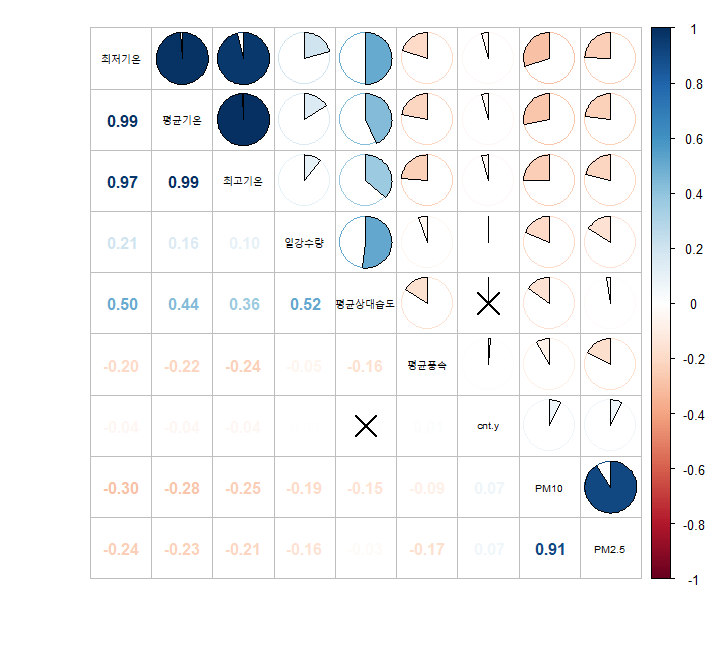
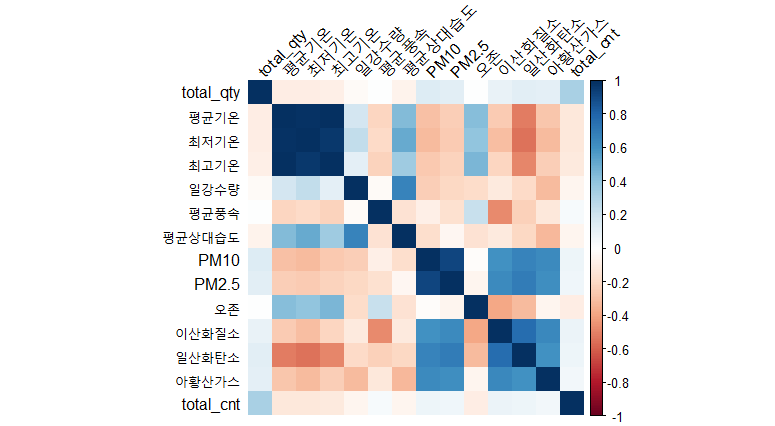
3. 데이터로부터 통찰을 얻기 위해 탐색하고 시각화.

먼저 데이터에 대략적인 모양을 보고 상관관계를 파악하고 필요한 파생변수를 추가하도록 하겠습니다. 전체적인 데이터의 모양은 변수간 모양은 아래와 같습니다.

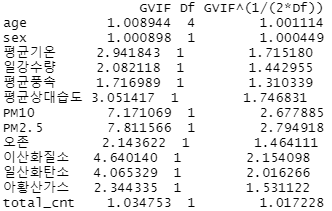
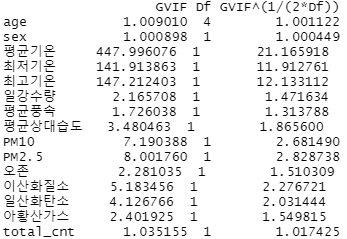


데이터들을 살펴본 결과 SNS언급량과 구매량 사이에 밀접한 연관이 있어 보입니다. 또, 계절의 영향이 있다고 판단 season (봄: 3,4,5월, 여름 : 6,7,8월, 가을: 9,10,11월, 겨울:12,1,2월)변수를 추가했습니다. 일별로 예측하기는 무리가 있다고 판단 월로 묶은 month변수를 추가하고 혹시 요일의 영향이 있는지를 보기 위해 weekday변수를 추가 하겠습니다.

그 다음 연속형 변수간 상관관계를 파악하기 위해 상관관계 그래프를 그려보았습니다.



확인 결과 미세먼지와, 기온들 사이에서 상관관계가 0.8이상(왼쪽 그림의 하단부 수치가 스피어만 상관계수)으로 나옵니다. 나오므로 다중공선성이 의심됩니다. 다중공선성 발생 여부를 수치적으로 확인해 보겠습니다.



모델을 생성하여 VIF를 확인해보니 기온과 미세먼지 관련 변수에서 10이 넘는 숫자가 나옵니다. 평균기온, 최저기온, 최고기온 중 평균을 선택하고 나머지 두 변수를 삭제합니다. 미세먼지의 경우 VIF는 10 이하지만 5이상이며 스피어만 상관계수가 0.8이상이므로 PM2.5를 삭제하도록 하겠습니다.

또한 데이터에는 다양한 범주형 변수들이 있으므로 카이제곱 검정으로 각 변수들의 독립성 검정을 실시해 범주형 변수들간의 상관관계를 파악합니다. 귀무가설은 “두 집단은 서로 독립이다.” 이고 대립가설은 “독립이 아니다.”입니다. 아래는 카이제곱 검정을 진행한 결과 나온 p-value입니다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Age(나이)** | **Sex(성별)** | **Weekday(요일)** | **Month(구매월)** | **Season(계절)** |
| **Age(나이)** | - | **0.000431** | 0.9125 | 0.1752 | **0.002303** |
| **Sex(성별)** |  | - | 0.799 | 1 | 0.9889 |
| **Weekday(요일)** |  |  | - | **5.98E-15** | 0.1684 |
| **Month(구매월)** |  |  |  | - | **2.20E-16** |

위 표를 해석해보면, 나이와 성별, 계절은 상관이 있으며, 요일과 구매 월, 구매 월과 계절은 상관이 있습니다. 집단의 특성을 생각해보면, 구매 월과 계절은 비슷한 속성을 가지므로 구매 월 변수는 삭제하였습니다.

그러므로 최종적으로 사용될 독립변수는 나이, 성별, 구매건수, 평균기온, 일강수량, 평균풍속, 평균상대습도, PM10, 오존, 일산화탄소, 아황산가스, 10만건당 언급건수, 요일, 계절입니다. 설명의 편의성을 위해 앞으로는 구매건수 변수를 판매건수라고 지칭하겠습니다.

4. 머신러닝 알고리즘을 위한 데이터 준비

결측치 처리는 이미 진행하였기에 결측치는 신경쓰지 않도록 합니다. 범주형 변수들은 ONE-HOT-ENCODING을 통하여 더미변수를 만들어 머신러닝 알고리즘에 사용할 수 있도록 변경하고 다양한 변수탐색 기법을 통해 각 품목별 가장 성능이 좋은 변수를 선별하여 모델에 적용해 보도록 하겠습니다.

변수선별에 사용한 기법은 선형 회귀법의 단계별 변수탐색법, Random Forest의 importance, XGBOOST이며, 예측모델을 생성할 품목이 4개이므로 각 품목별로 변수탐색을 진행하였습니다.