**L.POINT USER 온라인 행동기반 트렌드 분석**

**2020. 4. 29**

|  |
| --- |
| **그리기이(가) 표시된 사진  자동 생성된 설명 훈련과정명 : 데이터베이스기반 빅데이터 분석가 양성과정** |
| **그리기이(가) 표시된 사진  자동 생성된 설명 훈련기간 : 2019-11-26 ~ 2020-04-29 (800시간 / 100일)** |
| 그리기이(가) 표시된 사진  자동 생성된 설명 **팀명 : 노아 (Noah)** |
| **그리기이(가) 표시된 사진  자동 생성된 설명 팀원 : 권나현, 김지현, 김현준, 박윤정, 이승진, 최승예, 황윤재** |

**요약**

본 프로젝트는 롯데 그룹 내의 온라인 계열사를 대상으로 소비자의 상품 군별 온라인 선호지수를 생성하고, 이를 활용한 전략을 제시하는 것을 목적으로 시행되었다. 시계열 분석을 통해 살펴본 바로는, 전반적으로 매출액이 하락하는 추세이며, 이후 예측 또한 하락의 추세를 보였다. 성별, 기기, 연령, 고객등급, 상품 등급에 대한 분산분석을 통해 매출의 구성과 현황을 분석하였으며, 이를 바탕으로 소비자 선호지수를 생성하였다.

소비자 선호지수(CP INDEX)는 대량구매 등의 사유로 인한 오류를 줄이기 위해 소분류 및 상품군별로 판매된 빈도를 나타낸다. 소비자 선호지수와 관련된 변수로 상품데이터인 소분류명 및 소분류별 판매 빈도수를 사용하였고, 구매데이터로는 총 구매액, 온라인 행동데이터로는 클릭 횟수, 시청 페이지 개수, 세션에 머문 시간 데이터를 활용하였다. 해당 변수를 이용하여 소비자 선호지수를 가장 잘 설명할 수 있는 식을 만들기 위해 회귀분석을 실시하여 각 변수의 가중치를 생성하였고, 해당 식을 활용하여 선호경향을 정량적으로 파악하였다.

정량적으로 파악된 선호지수를 통하여 아래의 2가지 서비스를 제안한다.

첫 번째, ‘역시즌 프로모션’이다.

5월 대비 8월에 겨울 상품인 패딩의 선호지수가 큰 폭으로 증가하였는데, 이를 활용하여 가격에민감한 소비자들을 잡고, 기업 입장에서도 재고를 처리할 수 있는 ‘역시즌 프로모션’을 더욱 활성화하여 롯데 온라인몰 이용을 더욱 늘리는 전략을 제안하는 바이다.

두 번째로는, ‘NEW 타겟팅’ 전략이다.

매출에 가장 큰 비중을 차지하고 있는 서울에 거주하는 30대 여성의 종합 선호지수에서 남성티셔츠, 남성캐주얼바지 등 남성패션용품이 상위권에 위치해있음을 발견하였다. 이는 특정 남성제품의 실 구매자는 ‘여성’인 경우가 많다는 것으로 해석되며, 이에 따라 여성에게 선호도가 높은 남성제품들은 여성으로 ‘NEW 타겟팅’ 전략이 필요하다 생각된다.

**목차**

[**1.** **분석배경** 5](#_Toc39047502)

[**1.1.** **연구 목적** 5](#_Toc39047503)

[**1.2.** **연구 범위** 5](#_Toc39047504)

[**2.** **데이터 설명** 5](#_Toc39047505)

[**3.** **데이터 전처리** 7](#_Toc39047506)

[**4.** **분석** 9](#_Toc39047507)

[**4.1.** **기초통계 분석** 9](#_Toc39047508)

[**4.2.** **시계열 분석** 10](#_Toc39047509)

[**4.2.1.** **단순 지수예측 모형** 10](#_Toc39047510)

[**4.2.2.** **자동 ARIMA 모형** 12](#_Toc39047511)

[**4.2.3.** **prophet 패키지 이용한 예측** 14](#_Toc39047512)

[**4.3.** **t-test** 17](#_Toc39047513)

[**4.4.** **상관분석** 19](#_Toc39047514)

[**4.5.** **회귀분석** 21](#_Toc39047515)

[**4.5.1.** **외부데이터 활용** 21](#_Toc39047516)

[**4.5.2.** **회귀분석(롯데 내부데이터)** 24](#_Toc39047517)

[**4.6.** **클러스터링** 27](#_Toc39047518)

[**4.6.1.** **고객 클러스터링 – 비위계적 방법** 27](#_Toc39047519)

[**4.6.2.** **고객 클러스터링 – 위계적 방법** 28](#_Toc39047520)

[**4.6.3.** **제품 클러스터링** 28](#_Toc39047521)

[**4.7.** **분산분석** 29](#_Toc39047522)

[**4.7.1.** **연령대별** 29](#_Toc39047523)

[**4.7.2.** **기기유형별** 32](#_Toc39047524)

[**4.7.3.** **지역별** 35](#_Toc39047525)

[**4.7.4.** **고객등급별** 37](#_Toc39047526)

[**4.7.5.** **상품등급별** 40](#_Toc39047527)

[**5.** **소비자선호지수 생성** 44](#_Toc39047528)

[**5.1.** **선호지수 생성 배경** 44](#_Toc39047529)

[**5.2.** **선호지수 생성 방법** 44](#_Toc39047530)

[**5.3.** **선호지수 생성** 45](#_Toc39047531)

[**6.** **서비스 제안 및 결론** 47](#_Toc39047532)

[**7.** **부록** 49](#_Toc39047533)

# **분석배경**

## **연구 목적**

롯데쇼핑은 2018년 그룹 온라인 사업을 담당하는 롯데닷컴을 인수한 뒤 e커머스 사업 본부를 신설하고 2년간 각 계열사 통합 작업을 진행해왔다. 그 동안 롯데는 계열사별로 온라인 쇼핑몰을 따로 운영해와 계열사 간 연계성이 떨어져 고객 혜택으로 이어지지 않는다는 지적이 계속 나왔다. 오프라인 시장을 지배해왔지만 점포 매출의 지속적인 감소에 위기감이 커진 롯데는 2020년 4월 28일 통합 플랫폼 론칭을 계기로 온라인 시장에서 존재감을 나타내기 위해 “이제 ‘모두를 위한 서비스’를 더 하지 않고, 단 한 사람만을 위한 ‘퍼스널 코디 서비스’를 할 예정”이라며 e커머스 사업에 박차를 가하고 있다.

이에 따라 롯데 온라인 사업에 온라인 행동기반을 통한 트렌드 분석에 대해 긍정적으로 평가되고 있고, 이를 기반으로 현재 롯데 데이터를 분석 후 상품군 선호지수 식을 생성하고 상품군별 선호지수 상승을 위해 인사이트 도출과 창의적인 신규서비스를 제시하고자 한다.

## **연구 범위**

롯데그룹 온라인 계열사(롯데 닷컴, 롯데홈쇼핑 등)를 임의로 선정하여 구매 및 이용 이력과 관련된 데이터 통계기반 분석 기법으로 빅데이터 분석을 진행하고. 외부데이터인 경기지표, 시장 금리, 소비자 심리지수로 상관분석을 진행한다.

L.POINT 멤버십 가입자 소비동향을 보고 소비자 선호지수를 식으로 생성하고 디지털 트렌드 Key Needs & Key Insight를 도출하려 한다.

# **데이터 설명**

* 1. **상품구매 (PRODUCT) 테이블**

상품구매(PRODUCT) 테이블은 분석기간 내 온라인에서 구매된 상품명, 상품금액, 상품건수 등 방문자의 상품 구매정보를 담고 있는 테이블이다.

클라이언트 ID는 PC/모바일에 접속하는 순간 방문자로 인지하여 쿠키에 랜덤으로 부여된 아이디이다. Sessions(세션)은 방문자가 일정한 기간 내에 웹/앱 사이트에 방문하여 활동하는 영역을 의미하며, 세션 시작 시 세션ID가 발급된다. 세션이 변경/만료되는 경우는 사이트간 이동을 하거나, 세션 활동이 없을 경우 30분 후 세션이 만료되고, 자정에도 세션이 만료된다. HITS(조회)는 방문자의 세션 내 최소 행동단위를 의미하며, 페이지 조회, 클릭, 검색, 상품 장바구니 담기, 상품 구매 등 모든 행위를 포함한다.



* 1. **검색어1, 2 (Search1, 2) 테이블**

검색어1 (Search1) 테이블은 분석기간 내 온라인에서 상품을 구매한 방문자가 검색한 검색어의 검색량을 담고 있는 테이블이고, 검색어2 (Search2) 테이블은 분석기간 내 검색된 모든 일별, 검색어별 검색량을 담고 있는 테이블이다.



* 1. **회원 (Customer) 테이블**

회원(Customer) 테이블은 분석기간 내 온라인에서 상품을 구매한 방문자의 성별, 연령 정보를 담고 있는 테이블이다. 단, 미상정보 등은 제외되었다.



* 1. **세션 (Session) 테이블**

세션(Session) 테이블은 분석기간 내 온라인에서 상품을 구매한 방문자의 세션 정보를 담고 있으며, 온라인 행동데이터를 보여준다. 단, 미상정보 등은 제외되었다.

온라인 행동데이터란 고객의 온라인 행동에 대한 추적을 통해 유입부터 전환까지 모든 행동과정을 분석할 수 있는 데이터이다. 온라인 행동 데이터를 활용하여 사이트 내 흐름파악 및 진단, 잠재고객 타겟 맞춤 서비스 제공, 매출 트렌드 분석 및 예측 등 다양한 분석이 가능하다.



* 1. **상품분류 (Master) 테이블**

상품분류(Master) 테이블은 각 상품을 대분류, 중분류, 소분류로 그룹 지어 분류한 이름을 담고 있는 테이블이다.

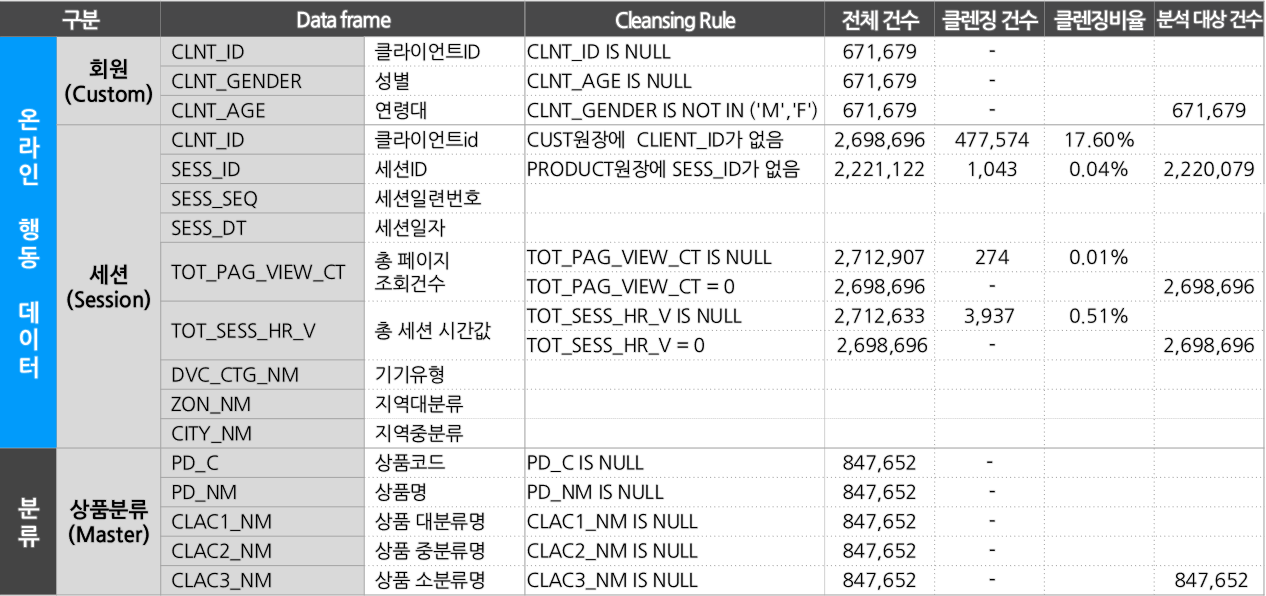


# **데이터 전처리**

롯데 온라인 데이터를 다양한 분석에 활용하기 위해서는 자료 전처리 과정이 필요하다. 롯데에서 제공한 데이터 설명 부분을 참고하여 분석목적에 맞게 데이터를 가공하였다. 아래 결과는 분석에 맞게 데이터를 가공시킨 결과를 정리한 내용이다.

[데이터 전처리 결과]





데이터 전처리 과정은 롯데에서 제공한 총 6개의 테이블별로 진행하였다. 위의 표에서 확인할 수 있듯이, 롯데에서 제공한 데이터 설명과 저희 팀 자체적인 기준을 가지고 분석 목적에 맞게 데이터 셋을 구성하였다. 자세한 전처리(클렌징)내용은 위에 있는 표에 명시하였고, 클렌징된 데이터 건수와 비율 등도 명시하였다.

전체적인 클렌징 작업 이후 SEARCH1과 2을 제외한 나머지 네 개의 테이블을 결합해 하나의 데이터셋을 구성하는 작업을 진행하였다. 상품구매(Product) 테이블의 고객 아이디(CLNT\_ID)를 기준으로 회원(Custom), 세션(Session) 그리고 상품분류(Master) 테이블을 결합시켰다.

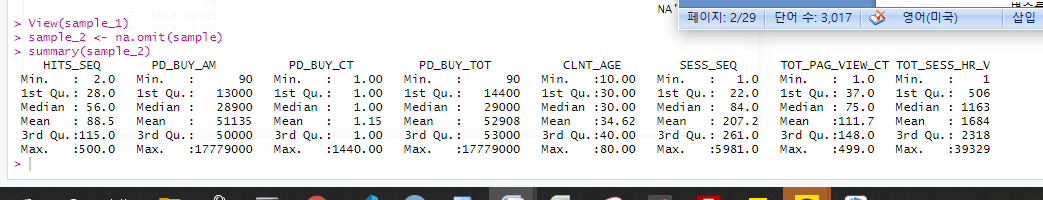
# **분석**

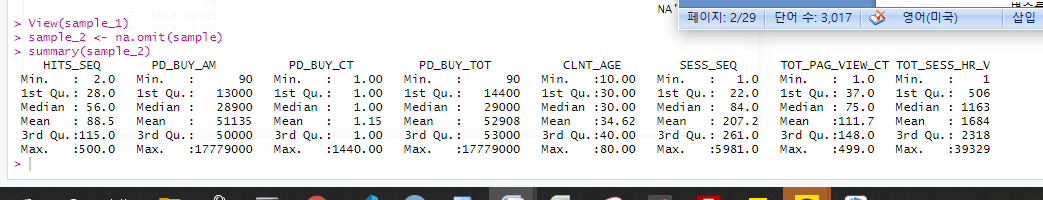
## **기초통계 분석**

롯데 온라인 데이터 분석 진행에 앞서서 개별 변수들의 기초통계량을 확인하는 절차가 필요하다. 통계 프로그램 R과 SAS를 이용해서 변수별로 특성을 파악하였다. 주어진 변수들 중 정량적 변수들을 선택해서 분석을 진행하였다.

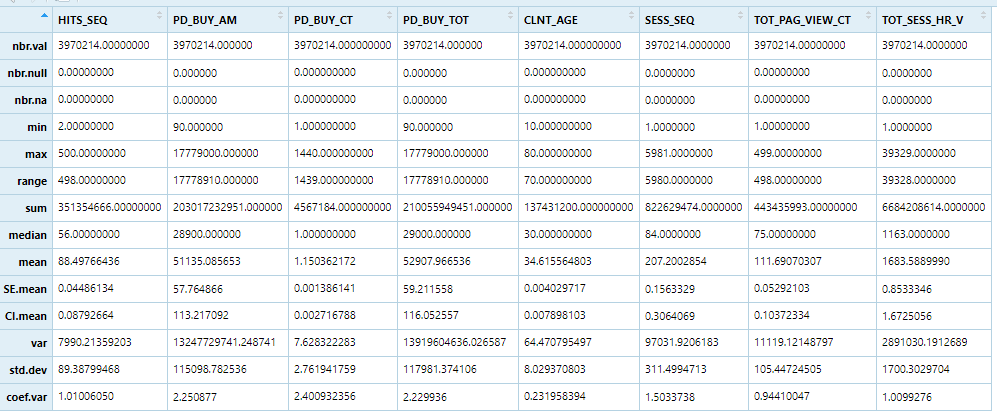
아래 결과는 R프로그램에서 summary{base} 함수와 stat.desc{pastec}를 이용해서 기초통계량 부분을 구하였다.

4.1.1. 기초통계량 ( {base} 라이브러리 summary 함수 )





4.1.2. 기초통계량 ( {pastec} 라이브러리 stat.desc 함수 )



개별 변수에 관한 자세한 분석 결과는 SAS 통계 프로그램을 이용하였다. 왜도와 첨도를 구하였고 추가로 정규성을 검정을 진행하였다. Shapiro-Wilk 특성상 표본의 개수가 많으면 따로 진행하지 않고도 데이터셋의 정규성을 가정한다. 하지만 정확한 수치를 확인할 필요가 있어서 검정을 진행하였으나 결과자료가 너무 큰 관계로 부록에 첨부하였다.

## **시계열 분석**

일, 주, 월별로 총 구매금액의 추이를 살펴보기 위해 시계열 분석을 먼저 진행하였다.

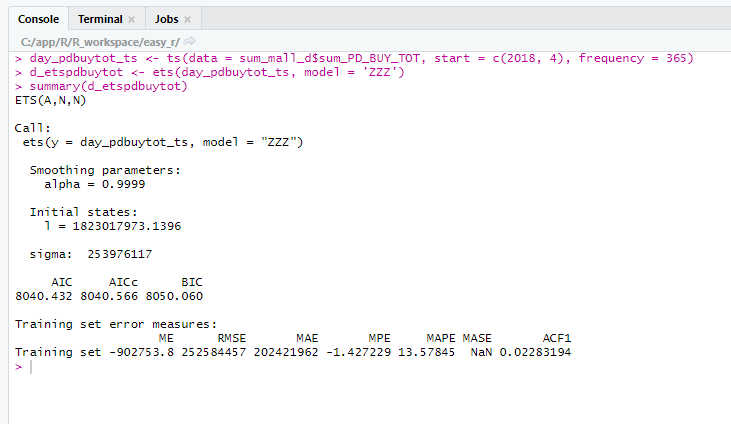
전체 데이터에서 SESS\_DT(날짜), PD\_BUY\_TOT(총 구매금액) 변수를 추출하여 SESS\_DT변수의 NA값을 제거한 후 일별, 주별, 월별로 구매금액을 합산한 sum\_PD\_BUY\_TOT 변수를 생성 후 ts객체로 변환하여 분석하였다.

분석의 원활한 진행을 위하여 forecast, prophet 패키지를 설치하였다.

### **단순 지수예측 모형**

{forecast}의 ets 함수를 이용하여 단순 지수예측모형 분석을 진행하였다.

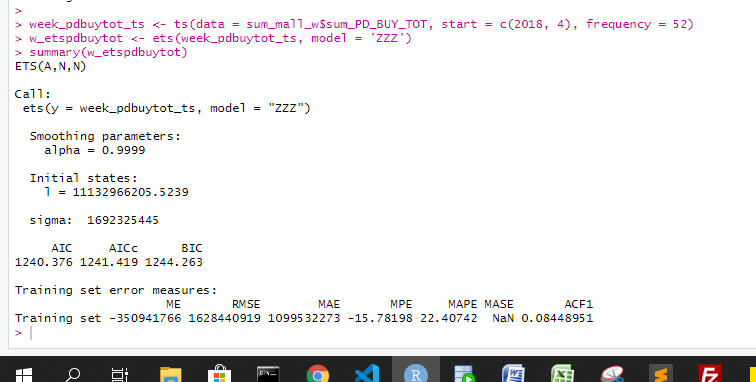
4.2.1.1. 일별 지수예측 모형.



일별 ts객체를 모델을 자동선택으로 지정되게 한 후 ets함수를 이용하여 지수예측모형을 실행한 결과, AIC값은 8040.432, RMSE값은 252584457, MAPE값은 13.578이다. 이때 RMSE값은 예측 대상의 크기에 영향을 받고, 보통 척도에 의존성을 가지기 때문에 모형 간 비교가 쉽고 이상치(outlier)에 대해 민감하게 반응하지 않아 신뢰도가 높은 기준으로 평가받는 MAPE값을 주로 본다.

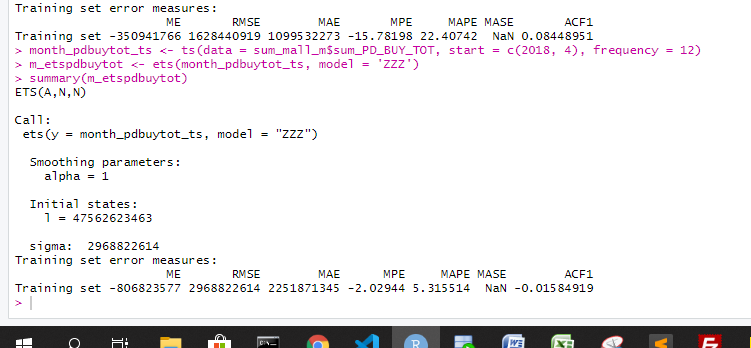
MAPE값이 0% ≦ MAPE＜10%인 경우 매우 정확한 예측을 나타내고, 10% ≦ MAPE＜20%는 비교적 정확한 예측을 나타내며, 20% ≦ MAPE＜50%는 비교적 합리적 예측을 나타내며, MAPE ≧ 50%는 부정확한 예측으로 판단한다. 이 결과에서 MAPE값은 13.578의 값을 가지고 있으므로 비교적 정확한 예측을 나타낸다고 볼 수 있다.

4.2.1.2 주별 지수예측 모형

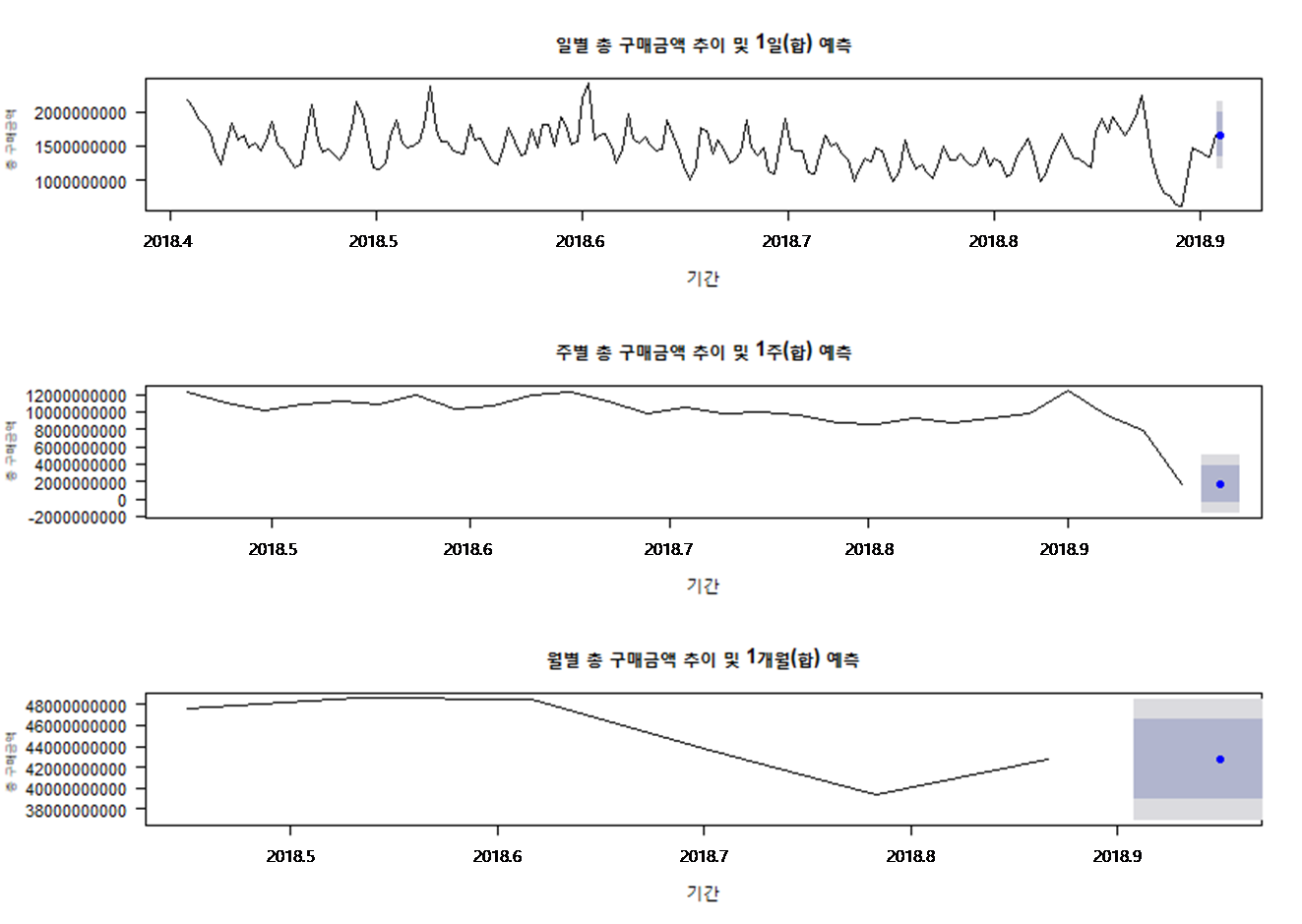


주별 ts객체를 모델을 자동선택으로 지정되게 한 후 ets함수를 이용하여 지수예측모형을 실행한 결과, AIC값은 1240.376, RMSE값은 1628440919, MAPE값은 22.407이다. 일별 지수예측 모형과 비교해보면 AIC값이 현저히 낮아진 것을 볼 수 있으나, MAPE값이 다소 높아진 것을 볼 수 있다. 이 결과에서 MAPE값은 22.407의 값을 가지고 있으므로 비교적 합리적 예측을 나타낸다고 볼 수 있다.

4.2.1.3 월별 지수예측 모형



월별 ts객체를 ets함수를 이용하여 지수예측모형을 실행한 결과, RMSE값은 2968822614, MAPE값은 5.316이다. 일별, 주별 지수예측 모형과 비교해보면 가장 낮은 MAPE값을 가지는 것을 알 수 있다. 이 결과에서 MAPE값은 5.316의 값을 가지고 있으므로 매우 정확한 예측을 나타낸다고 볼 수 있다.



[그림 4-1] 단순 지수예측모형 일, 주, 월별 추이 및 예측 plot

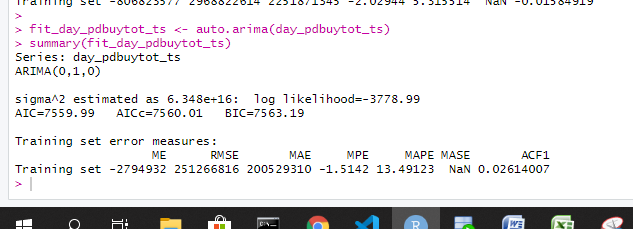
위의 [그림 4-1]은 ets함수를 이용하여 일, 주, 월별 지수예측모형의 결과를 plot으로 나타낸 것이다. 주별 plot의 예측 신뢰구간 값이 0보다 작게 나타나는 것은 마지막 주에 묶인 데이터가 적어서이다. 대체로 총 구매금액은 전일 또는 전월의 추이를 유지할 것으로 예상된다.

수집한 데이터의 한계로 향후 예측결과의 신뢰구간이 넓게 나타나는 것을 확인할 수 있다.

### **자동 ARIMA 모형**

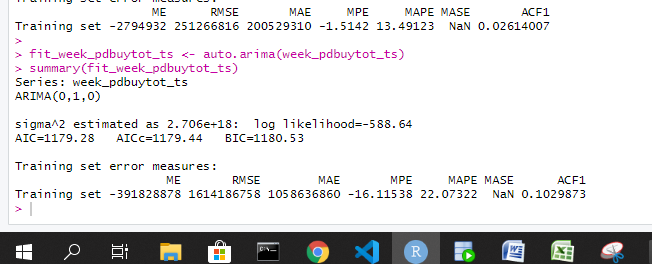
{forecast}의 auto.arima 함수를 이용 ARIMA모형 분석을 진행하였다.

4.2.2.1. 일별 자동 ARIMA 모형



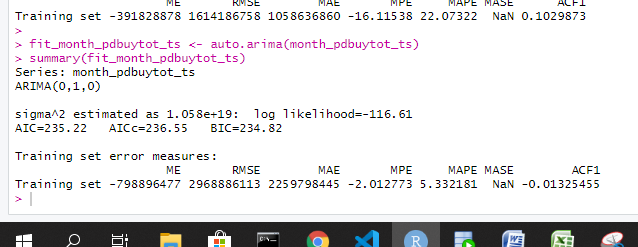
일별 ts객체를 auto.arima함수를 이용하여 자동 ARIMA모형을 실행한 결과, AIC값은7559.99, RMSE값은 251266816, MAPE값은 13.491이다. 이 결과에서 MAPE값은 13.491의 값을 가지고 있으므로 비교적 정확한 예측을 나타낸다고 볼 수 있다.

4.2.2.2 주별 자동 ARIMA 모형



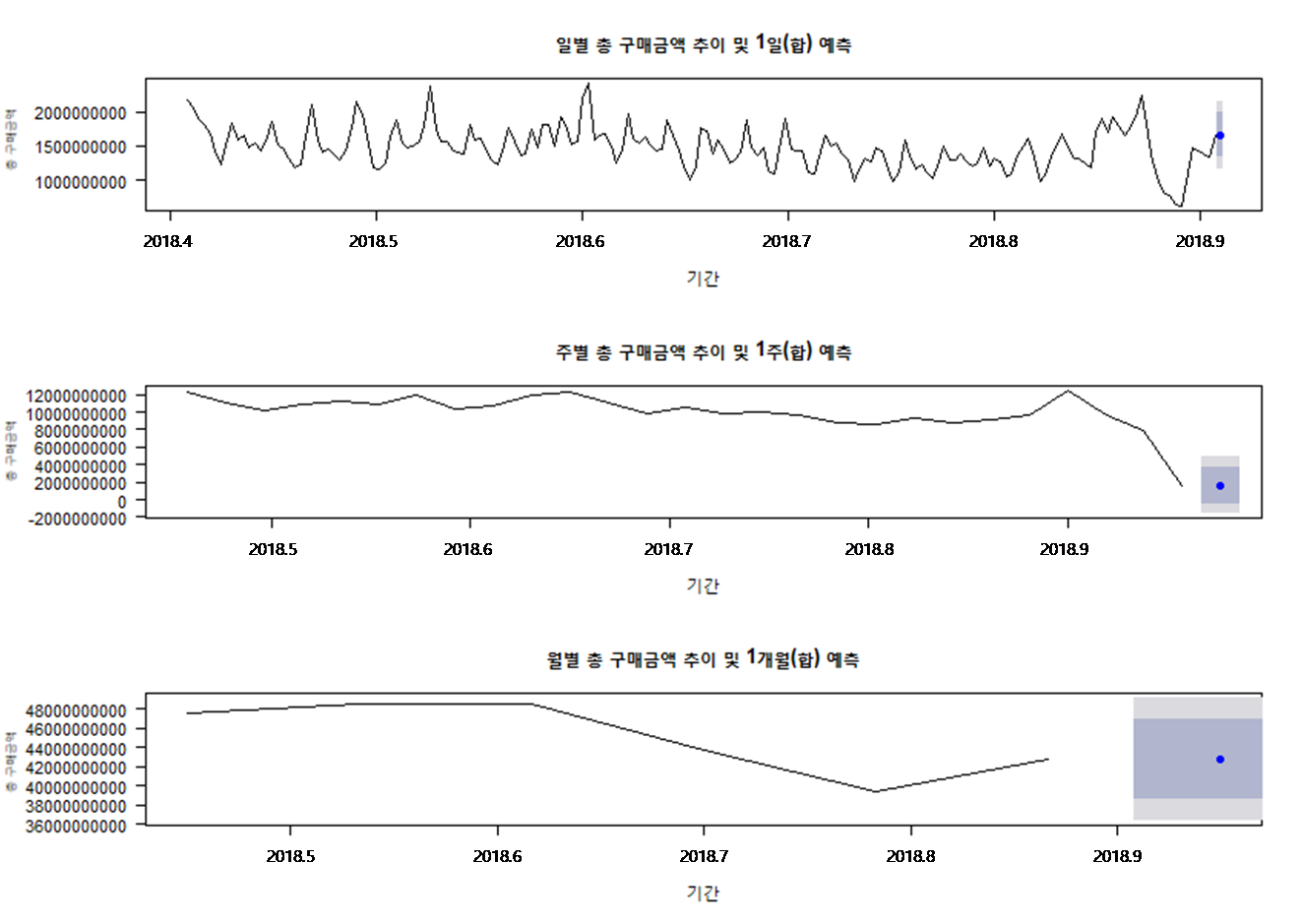
주별 ts객체를 auto.arima함수를 이용하여 자동 ARIMA모형을 실행한 결과, AIC값은 1179.28, RMSE값은 1614186758, MAPE값은 22.073이다. 이 결과에서 MAPE값은 22.073의 값을 가지고 있으므로 비교적 합리적 예측을 나타낸다고 볼 수 있다.

4.2.2.3 월별 자동 ARIMA 모형



월별 ts객체를 auto.arima함수를 이용하여 자동 ARIMA모형을 실행한 결과, AIC값은 235.22, RMSE값은 2968886113, MAPE값은 5.332이다. 이 결과에서 MAPE값은 5.332의 값을 가지고 있으므로 매우 정확한 예측을 나타낸다고 볼 수 있다.

일, 주, 월별 자동 ARIMA 모형의 결과를 비교해본 결과, 월별 자동 ARIMA모형의 AIC값과 MAPE값이 가장 작은 것을 확인할 수 있다.



[그림 4-2] 자동 ARIMA모형 일, 주, 월별 추이 및 예측 plot

위의 [그림 4-2]는 auto.arima함수를 이용하여 자동 ARIMA모형의 결과를 plot으로 나타낸 것이다. 단순 지수예측모형의 plot(그림 4-1)과 매우 유사하게 그려지는 것을 볼 수 있다.

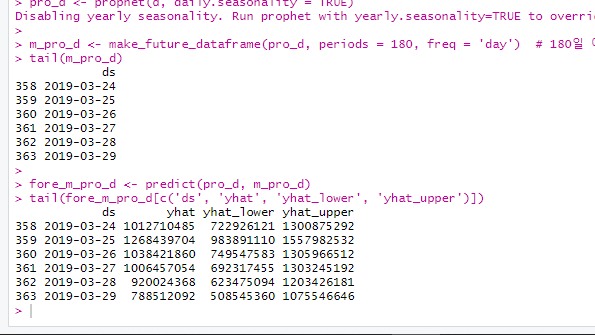
자동 ARIMA모형의 plot 역시 대체로 총 구매금액은 전일 또는 전월의 추이를 유지할 것으로 예상된다. 이 plot 역시 데이터의 한계로 향후 예측결과의 신뢰구간이 넓게 나타나는 것을 확인할 수 있다.

### **prophet 패키지 이용한 예측**

페이스북에서 개발한 시계열 예측 패키지인 {prophet}패키지의 prophet 함수를 사용하여 일별, 주별, 월별 prophet 객체를 생성 후 예측 날짜 구간을 생성하고, 예측을 실행하였다.,

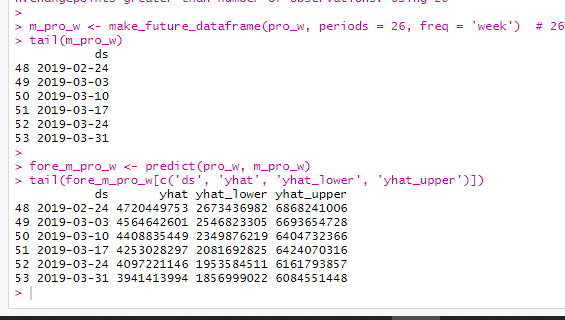
(주의: prophet 클래스 객체를 생성하기 위해서는 컬럼 이름(ds, y)을 꼭 맞춰주어야 한다.)

4.2.3.1 일별



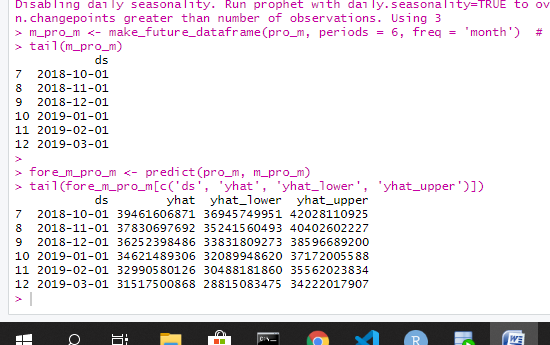
make\_future\_dataframe 함수를 이용하여 periods를 180, freq를 ‘day’로 설정하여 미래 180일을 예상하는 날짜구간을 생성하였다. 이후 predict함수를 이용하여 원본 데이터와 예상 날짜구간에 대하여 신뢰구간을 포함한 예측을 실행하였다. 예측값을 보여주는 plot은 아래에 나타내었다.

4.2.3.2 주별

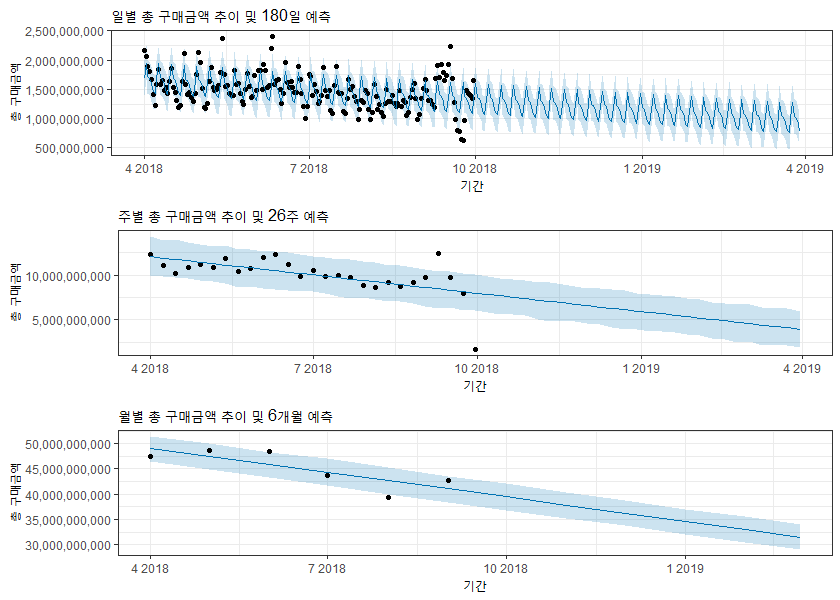


make\_future\_dataframe 함수를 이용하여 periods를 26, freq를 ‘week’로 설정하여 미래 26주를 예상하는 날짜구간을 생성하였다. 이후 predict함수를 이용하여 원본 데이터와 예상 날짜구간에 대하여 신뢰구간을 포함한 예측을 실행하였다. 예측값을 보여주는 plot은 아래에 나타내었다.

4.2.3.3 월별



make\_future\_dataframe 함수를 이용하여 periods를 6, freq를 ‘month’로 설정하여 미래 6개월을 예상하는 날짜구간을 생성하였다. 이후 predict함수를 이용하여 원본 데이터와 예상 날짜구간에 대하여 신뢰구간을 포함한 예측을 실행하였다. 예측값을 보여주는 plot은 아래에 나타내었다.

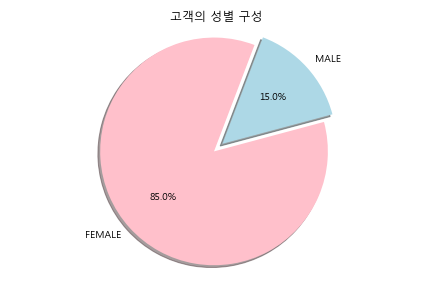


[그림 4-3] prophet 패키지를 이용한 일, 주, 월별 추이 및 예측 plot

위의 [그림 4-3]은 일별, 주별, 월별 각각의 prophet 객체를 예측한 결과를 plot으로 나타낸 것이다. 앞의 두 모형(단순 지수예측모형, 자동 ARIMA모형)에서는 총 구매금액이 어느 정도 유지될 것으로 예상되었으나, prophet 패키지를 이용한 예측결과는 계속해서 총 구매금액이 하락 할 것으로 예측되고 있음을 알 수 있다. prophet 패키지 자체의 특성상 ARIMA와 같은 확률론적이고 이론적인 모형이 아니라 몇 가지 경험적 규칙을 사용하는 단순 회귀모형인 점에서 앞의 두 모형의 예측결과와 다른 것이 아닐까 생각해본다.

## **t-test**

집단간 차이를 분석하는 t-test를 진행하였다. 주어진 데이터 셋에 고객의 성별을 파악할 수 있는 변수(CLNT\_GENDER)을 통해서 나머지 양적인 변수들을 비교할 수 있다.



[그림 4-4] 고객 성별 분포

위 그림에서 확인할 수 있듯이 롯데 온라인 쇼핑몰 전체 이용자 중 85%가 여성 고객이고, 나머지 15%가 남성 고객이다. 일차적으로 시각적인 자료를 확인해봐도 두 집단 사이에 표본(쇼핑몰 이용자) 크기의 차이가 크다는 것을 확인할 수 있다. 이 결과에서 표본 크기가 다른 만큼 온라인 쇼핑의 행태에도 차이가 있을 것으로 가정하였다.

시각적인 결과를 바탕으로, 롯데 온라인 데이터에서 제공하는 모든 양적인 변수들을 성별에 따라 개별적으로 분석을 진행해서 집단 간 차이를 확인하였다. 개별 변수별로 등분산성 검정과 t-test(또는 Wilcoxon signed rank test)을 진행하였고, 추가적으로 박스 플롯으로 두 집단 간 차이를 시각적인 방법으로 확인하였다.

변수1. 총 구매금액 (PD\_BUY\_TOT)

1) 총 구매금액 등분산성 검정

조류이(가) 표시된 사진

자동 생성된 설명

집단 간 차이를 확인하는 t-test을 확인하기 전에 먼저 F-test을 진행해 두 집단(성별)이 등분산성 가정을 만족하는지 확인하였다.

분석결과, 성별에 따른 총 구매금액의 등분산성 가정(귀무가설: 두 집단 간 분산은 동일하다)이 p-value(유의수준) 0.005 하에 기각되었다. 따라서, 등분산성이 만족되지 않는다는 대립가설이 채택되었음을 확인할 수 있다.

등분산성 가정이 만족되지 않아 설명변수의 집단간 차이를 확인할 수 있는 t-test를 진행할 수 없게 되었다. 이 문제를 해결하기 위해서 t-test와 같이 모수적인 방법이 아닌 비모수적인 방법인윌콕슨 검정(Wilcoxon signed rank test)을 이용해 집단 간 차이를 확인하였다.

2) 총 구매금액 윌콕슨 검정

테이블, 방, 쥐고있는, 목재의이(가) 표시된 사진

자동 생성된 설명

비모수적 방법인 윌콕슨 검정으로 총 구매금액에 따른 집단(성별)간 차이를 검정하였다. 위 분석결과에서 확인할 수 있듯이 귀무가설(집단간 차이가 없다)은 유의수준(p-value) 0.005 하에 기각된다는 사실을 확인할 수 있다. 대립가설을 채택함으로써 성별 간에 총 구매금액이 차이가 있다는 사실을 통계적인 수치로 입증하였다.

정량화된 수치 방법 이외에도 시각적인 방법을 통해서 성별 사이에 차이가 있다는 사실을 관찰할 수 있다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

[그림 4-5] 성별에 따른 총 구매금액박스 플롯 분포

박스 플롯을 이용하여 시각적인 방법으로 총 구매금액에 따른 두 집단 사이에 차이를 확인하였다. 총 데이터셋이 너무 커서 전체 데이터의 0.0001%만 무작위로 샘플링하여 시각화를 진행하였다. 일부만 확인해봐도 성별에 따라 차이가 있다는 것을 관찰할 수 있다. 눈에 띄는 점은 두 성별 모두 이상치가 매우 많고, 특히 여성 표본에서 이상치 분포가 크다는 것을 확인할 수 있다.

총 구매금액(PD\_BUY\_TOT) 이외에도 롯데에서 제공한 양적인 변수 6개(상품 단가(PD\_BUY\_AM), 상품 구매수량(PD\_BUY\_CT), 구매하기까지 모든 행위(HITS\_SEQ), 총 세션시간(TOT\_SESS\_HR\_V), 세션 시퀀스(SESS\_SEQ), 총 페이지 뷰(TOT\_PAG\_VIEW\_CT)) 에 관한 집단간 차이를 분석하였다.

성별에 따른 소비행태가 확연히 다르고 그리고 등분산성 가정도 만족되지 않아 모든 설명변수들을 윌콕슨 검정을 통해서 분석하였다. 모든 분석 결과에서 차이가 있는 것이 확인되었고, 절차가 동일한 관계로, 나머지 변수들에 대한 결과는 부록에 첨부하였다.

## **상관분석**

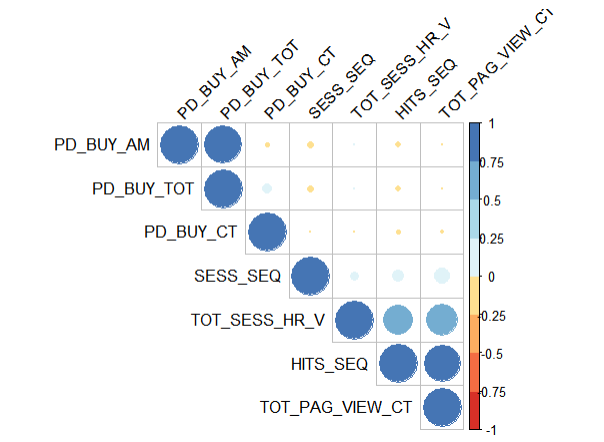
기초통계량을 토대로 개별변수들 사이에 상관분석을 진행하였다. 총 7개의 정량적인 변수를 중심으로 분석을 하였다. 2018년 4월부터 9월까지 총 6개월 구매기간 동안 거래된 고객 아이디(CLNT\_ID)를 따로 합산하지 않고, 일차적으로 상관분석을 진행하였다.

개별 구매 거래 내역 관측치 5,019,730개를 개별 아이디로 보고 변수별로 상관행렬을 만들고, 추가적으로 correlogram을 이용해서 시각적인 방법으로 변수별로 상관관계를 파악하였다.

4.4.1. 상관 분석 (개별 고객 아이디별)

사진, 병, 걸린, 테이블이(가) 표시된 사진

자동 생성된 설명



[그림 4-6] 개별 고객 아이디별 상관계수 시각화

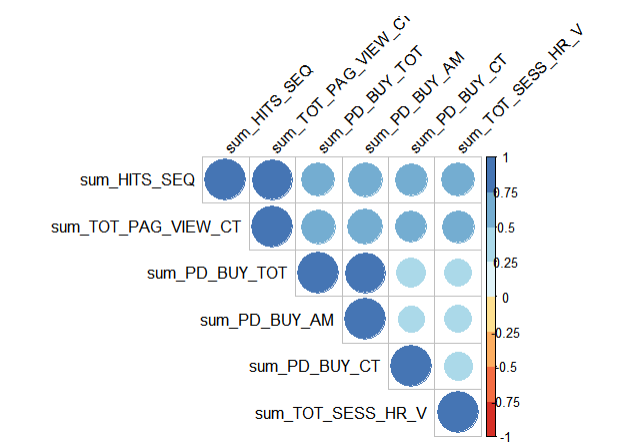
분석결과, 총 구매금액과 상품 단가를 제외하고 유의미한 상관 결과를 도출하지 못하였다. 또한 상품 단가도 총 구매금액을 구성하는 하나의 요소로서 상관 관계가 높은 것이 당연한 결과이다. 따라서, 개별 고객 아이디에 따른 변수별 상관분석은 통계적으로 유의한 결과가 없는 것으로 판단된다.

4.4.2. 상관 분석 (고객 아이디별 합산)

위 분석결과를 토대로 고객 아이디별로 6개월 총 구매내역을 합산하는 것이 필요하다는 결과를 얻었고, 총 912,837명의 고객 데이터로 상관분석을 추가로 진행하였다.

스크린샷이(가) 표시된 사진

자동 생성된 설명



[그림 4-7] 고객 아이디별 합산한 상관계수 시각화

고객 아이디별 합산한 개별 변수들 사이에는 일부 유의미한 상관 관계가 있는 결과가 도출되었다. 특히 고객별 총 구매금액(PD\_BUY\_TOT)에 히트 시퀀스(HITS\_SEQ), 총 페이지뷰(TOT\_PAG\_VIEW\_CT), 그리고 총 세션 시간(TOT\_SESS\_HR\_V) 모두 0.4 이상의 약 상관관계가 있다는 사실을 얻었다. 히트 시퀀스와 총 페이지 뷰 사이에는 0.6 이상의 상관도가 있어서, 추후 진행될 회귀분석 부분에서도 통계적으로 유의미한 회귀 결과가 도출될 것으로 예상된다.

## **회귀분석**

### **외부데이터 활용**

2018년 4월부터 9월까지 롯데 온라인 데이터를 관찰한 결과 전체적으로 매출이 하락하는 것을 관찰할 수 있었다. 매출액이 제일 높았던 5월 대비 8월에 전체 매출액의 12%정도가 하락하였으며, 9월에 약소하게 반등하는 추세였다.

롯데의 전체적인 매출 하락의 원인을 외부적인 요인에 의해 발생한 것이 아닐까 하는 가설을 세웠고, 외부데이터를 활용한 분석을 진행하였다. 매출하락의 원인을 내부적인 자료와 외부적인 데이터 셋을 종합적으로 고려한 회귀분석을 진행해 매출 하락의 원인을 파악하고자 하였다.

온라인 행동 데이터의 총 매출금액(PD\_BUY\_TOT)을 기준으로 외부데이터를 활용한 분석을 진행하였다. 즉, 총 매출금액과 외부 변수들 사이의 상관관계를 파악하고 그 결과를 토대로 추가적인 외부 변수들과 회귀분석을 진행하였다.

4.5.1.1. 총 구매금액과 외부데이터 상관관계 분석

테이블, 많은, 대형이(가) 표시된 사진

자동 생성된 설명

데이터 셋에서 고객 아이디(CLNT\_ID)를 기준으로 총 구매금액(PD\_BUY\_TOT)을 합산하여 고객별 6개월간 총 구매금액(sum\_PD\_BUY\_TOT) 변수를 생성하였다. 외부데이터를 선정하는 과정에서 경기 지표로 대변되는 (경기동행, 선행 지수), 시장금리(콜금리, 3년 국채)와 소비자들의 행태를 반영하지는 지수들을 분석에 포함시켰다. 데이터셋의 한계로 2018년 4월부터 9월까지, 변수별로 총 6개월 기간의 지수들을 수집하였다. 총 구매금액(sum\_PD\_BUY\_TOT)을 기준으로 소비자 물가지수(CPI), 3년 국채(TB3), 콜금리(Call), 경기동행지수(CCI), 경기선행지수(CLI), 소비자심리지수(mind)와 상관관계 분석을 진행하였다.

분석결과, 총 구매금액과 제일 상관관계가 높은 외부 변수는 소비자 심리지수(mind)로 관찰되었다. 상관계수가 0.944로 매우 높은 상관관계를 가지고 있다는 것을 확인할 수 있었다. 소비자 심리지수는 일반 소비자들이 전반적인 시장 상황에 대한 심리를 정량화된 수치로서 보여주는 지표이다. 상관분석 결과로 롯데 온라인 쇼핑몰을 이용하는 고객들은 시장에서 소비자 심리지수가 좋을 때 구매금액이 높았고, 낮을 때 구매를 줄이는 행동을 보여주었다.

소비자 심리지수가 총 구매금액과 상관성이 높다는 결과를 토대로 소비자 심리지수와 추가적인 외부변수들 사이의 회귀 분석을 진행하였다.

4.5.1.2. 소비자 심리지수와 추가 외부데이터 상관분석

소비자 심리지수는 롯데 온라인 데이터의 고객별 총 구매금액과 상관관계가 제일 높은 것으로 확인되었다. 2018년 4월부터 9월까지 총 6개월 기간 동안 두 그래프는 비슷한 패턴으로 움직이는 것이 관찰되었다. 따라서, 전반적인 롯데 기업의 매출 하락의 원인을 파악하기 위한 방법으로 소비자 심리지수를 종속변수로 설정해 외부 설명변수들과 회귀분석을 진행하였다.

외부 데이터를 선택하는 과정에서 종합적인 거시 경제지표들을 선택하였다. 한국은행과 통계청에서 2008년에서 2018년까지 총 10년치 월별 지수들을 확보하였다. 외부데이터로 선택 변수들의 목록은 다음과 같다.

1. 경제 지표 : GDP 성장률, 총 저축률, 국내 총 투자율, 경기선행지수, 경기 동행지수, 공산품 수입물가지수
2. 시장 지표 : 무담보 콜금리(1일), CD유통 수익률(91일), CP유통 수익률(91일), 국고채(1년, 3년), 회사채(장외 3년, AA-등급), 은행 신용카드 개인 이용건수, 은행 신용카드 개인 이용금액, 신용카드 발급장수, KOSPI 시가총액, 거래량, 종가, KOSDAQ 시가총액, 거래량, 종가
3. 거시 경제 지표 : 생산자 물가지수, 소비자 물가지수, 달러 환율, 위안 환율, 엔화 환율

소비자 심리지수를 비롯한 총 26개의 외부 데이터를 일차적인 상관분석의 회귀분석을 선택하였다. 회귀분석을 위해 소비자 심리지수와 상관도가 높고 실제 경제학(금융)이론을 뒷받침하는 변수들을 선택하였다. 상관분석 결과는 26개의 변수가 포함되어 있어 자료가 크기 때문에 부록에 따로 첨부하였다.

4.5.1.3. 소비자 심리지수와 추가 외부데이터 회귀분석

롯데 온라인몰의 매출의 하락원인을 제대로 설명하기 위해서 거시 경제 주기 하나만 포함하기로 결정하였다. 따라서 상관분석에서 선택한 변수들을 2016년부터 2018년까지 2년치 월별 데이터로 축소해서 분석에 사용하였다. 지표 간 단위 격차가 커서 결과를 왜곡할 수 있다고 판단해, 전월비 자료로 변수들을 전처리하였다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

먼저, 소비자 심리지수(CCSI)를 종속변수로 두고 코스피 종가(KOSPI\_Y), 달러환율(USD\_Y), 경제 심리지수(ESI), 국내 총 투자율(Invest\_M)을 설명변수에 포함시켜 회귀분석을 진행하였다.

분석결과, 결정계수()가 0.9411로 매우 높았다. 코스피 종가와 경제 심리지수는 유의수준(p-value) 0.05 하에 귀무가설이 기각되어 통계적으로 매우 유의미하다는 결과를 얻었다. 달러 환율과 국내 총 투자율 또한 유의수준 0.05하에 유의한 변수로 판별되었다.

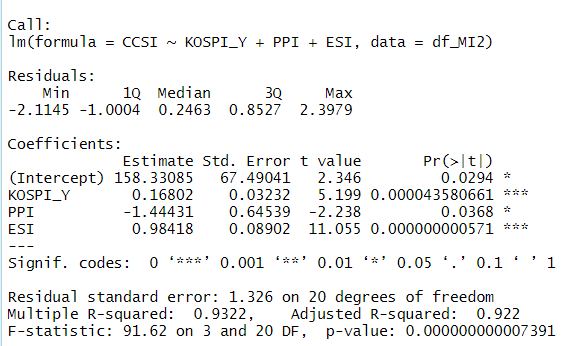
소비자 심리지수에 환율을 제외한 나머지 설명변수들은 양의 영향력을 미치는 것으로 확인되었다. 환율이 오르게 되면 소비자들의 주가가 하락하고 전반적인 경제상황에 대해 부정적으로 판단해 소비를 줄이는 것으로 보인다.

텍스트이(가) 표시된 사진

자동 생성된 설명

다음으로 신용카드 이용건수(Card\_Cnt\_Y)와 소비자 심리지수 사이에 양의 상관관계가 있을 것으로 가설을 세우고 추가적인 분석을 진행하였다. 소비자 심리지수를 종속변수로 두고 신용카드 이용건수, 코스피 종가, 달러 환율, 경제 심리지수, 국내 총 투자율을 설명변수에 포함시켜 이차 회귀분석을 진행하였다. 이번 과정에서는 stepAIC기법을 활용하여 최적의 모델을 선정하여 분석하였다.

분석결과, 결정계수()가 0.9512로 이전에 진행했던 회귀분석 결과보다 조금 더 높다. 하지만 새롭게 추가된 신용카드 이용건수는 ‘통계적으로 유의하지 않다’는 결과가 도출되었다.



마지막으로 코스피 종가, 생산자 물가지수(PPI), 경제 심리지수를 설명변수로 하는 회귀분석을 진행하였고, 이전 과정과 마찬가지로 stepAIC방법을 이용하여 최적의 분석모델을 선정하여 분석하였다.

분석결과, 결정계수()는0.9332로 이전 두 모델에 비해 다소 떨어졌으나, 충분히 높은 설명력을 가지고 있다고 볼 수 있다. 코스피 종가와 경제 심리지수는 유의수준(p-value) 0.05하에 매우 유의하였고, 생산자 물가지수도 0.05 수준에서 유의한 것으로 판명되었다.

지금까지 진행했던 회귀분석을 종합해보면, 소비자 심리지수는 코스피 종가와 경제 심리지수와 매우 영향이 높다는 것을 확인할 수 있었다. 이는 소비자들은 경제 심리가 좋거나 코스피 지수가 상승할 경우에 소비를 늘린다고 볼 수 있다. 반면, 경제 상황이 좋지 못하다고 판단이 되거나 주가가 환율상승 등의 원인으로 하락할 경우 소비 지출을 줄이는 것으로 확인되었다.

외부데이터를 이용한 회귀분석을 진행해 온라인 롯데 매출 하락의 원인을 파악하고자 하였다. 고객별 총 구매금액은 소비자 심리지수가 상관도가 높았고, 이것은 코스피 종가와 경제 심리지수에 가장 큰 영향을 받는 것으로 확인되었다.

외부데이터를 활용한 분석은 롯데 자체만의 문제를 설명하지 못한다는 단점이 있다. 롯데 매출액은 타 경쟁사(SSG, 쿠팡, 마켓컬리 등)와 비교했을 때 매출 하락의 정도가 매우 컸다. 비록, 외부데이터를 통해 2018년 4월 ~ 9월까지 경제가 하향 추세를 보인다는 게 사실이더라도 롯데의 경우 하락의 폭이 더 심각하였다. 롯데 온라인 쇼핑의 하락을 파악하기 위해서는 롯데 매출의 하락을 설명할 수 있는 자체 외부데이터를 활용하는 방법이 필요하다고 생각된다. 하지만 자료 수집에 한계가 있어, 외부데이터를 활용한 분석(상관분석, 회귀분석)은 여기까지 진행하는 것으로 선택하였다.

### **회귀분석(롯데 내부데이터)**

앞에서 진행한 상관분석 결과를 토대로 회귀분석을 진행하였다. 전체 6개월 기간 동안 롯데 온라인 쇼핑에서 구매한 고객 아이디(CLNT\_ID)별로 따로 합산하지 않은 상태에서 일차적으로 진행하였다.

데이터 셋을 확인해본 결과, Null 값을 제외하고 총 5,019,730개의 주문 내역이 존재하였다.

종속변수를 총 구매금액(PD\_BUY\_TOT)으로 두고 히트 시퀀스(HITS\_SEQ), 총 페이지 뷰(TOT\_PAG\_VIEW\_CT), 총 세션시간(TOT\_SESS\_HR\_V)을 설명변수로 포함시켰다. 상품 단가(PD\_BUY\_AM)와 총 상품구매 개수(PD\_BUY\_CT)는 총 구매금액(PD\_BUY\_TOT)에 다중공선성 문제가 발생할 것으로 예상돼 설명변수로 포함하지 않았다.

텍스트이(가) 표시된 사진

자동 생성된 설명

회귀분석 결과, 세 가지 설명변수가 유의수준(p-value) 0.05 하에 모두 유의한 것으로 확인되었으나 결정계수()이 매우 낮게 나온 것으로 관찰되었다. 이 결과는 아무리 설명변수가 통계적으로 유의성이 높다고 하더라도, 분석 모델 자체를 사용할 수 없는 것으로 보인다. 따라서, 다른 접근 방식으로 회귀분석을 진행하는 것이 바람직하다고 판단되었다.

위에서 발생한 문제를 해결하기 위해 6개월(2018년 4월 ~ 9월까지) 총 구매내역을 고객 아이디(CLNT\_ID)별로 합산하는 방식을 채택하였다. Null 값을 제외하고 총 912,837개의 고객 아이디가 있는 것으로 확인되었다. 고객 아이디별로 이전 분석에서 사용한 설명변수들을 개별적으로 합산하였다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

외부데이터를 이용한 회귀분석에서 사용한 stepAIC방식을 활용해서 분석에 최적의 모델을 선택하였다. 추가적으로 후진/전진 두 가지 방법을 이용해서 AIC 값이 25060760 값이 나오는 모델을 선택하였다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

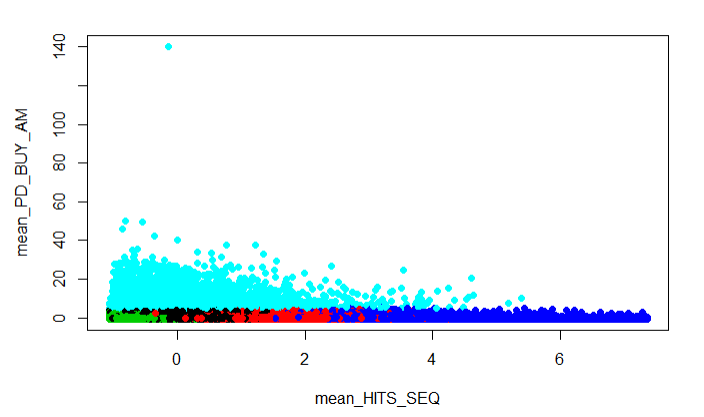
분석결과, 고객 아이디별로 합산한 개별 변수들 모두 유의수준(p-value) 0.05 하에 통계적으로 매우 유의한 것으로 나타났다. 또한, 결정계수()도 0.5372로 이전 모델에 비해 매우 높은 설명력을 가지는 것으로 관찰되었다.

온라인 쇼핑 고객들은 페이지를 많이 보고 세션에 머무르는 시간이 길어질수록 총 구매금액이 높아지는 것으로 확인되었다. 반면, 구매하기까지 행동(히트 시퀀스)이 많을수록 총 구매금액이 낮아지는 것을 볼 수 있다.

## **클러스터링**

클러스터링의 목적은 분류를 통해 금액, 소분류와 관련한 고객의 특성을 찾고자 함이다. 따라서,비위계적 방법의 고객 클러스터링, 위계적 방법의 고객 클러스터링, 제품별 클러스터링을 수행하였다.

### **고객 클러스터링 – 비위계적 방법**

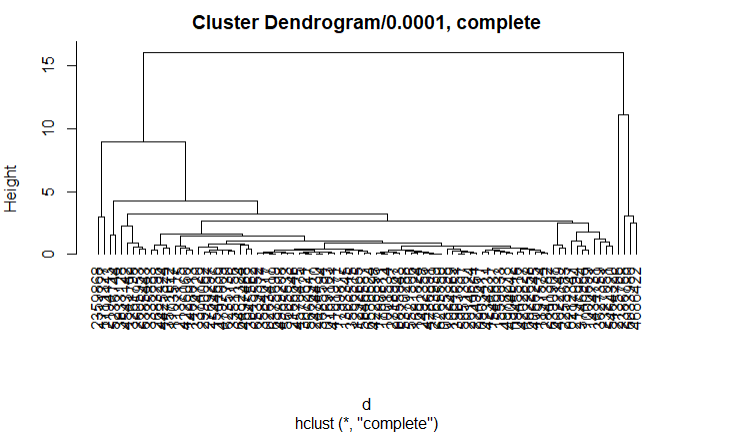


[그림 4-8] 비위계적 방법을 이용한 고객 클러스터링

위의 그림은 비위계적 방법으로 고객별 클러스터링을 수행한 결과이다. 고객별 클러스터링을 수행하기 위해 먼저 고객 아이디(CLNT\_ID)별로 그룹화를 수행하였고, NA값을 제거 후, 관련된 연속형 변수 6개(클릭 수(HITS\_SEQ), 상품 단가(PD\_BUY\_AM), 상품 구매수량(PD\_BUY\_CT), 상품 총 구매금액(PD\_BUY\_TOT), 총 시청 페이지수(TOT\_PAG\_VIEW\_CT), 총 세션시간(TOT\_SESS\_HR\_V)) 는 모두 mean() 함수로 처리하였다. 고객 아이디별로 그룹화된 변수는 표준화를 수행 후 k-means 방식과 5개의 군집으로 클러스터링을 수행하였다.

해당 클러스터링 결과로 유의미해보이는 클러스터링 집단이 나뉘었으나, 클러스터 번호 별로 subset을 생성하여 살펴본 결과, 대분류와 관련한 뚜렷한 고객의 특성을 찾기 힘들다고 판단하여 사용하지 않기로 결정하였다.

### **고객 클러스터링 – 위계적 방법**



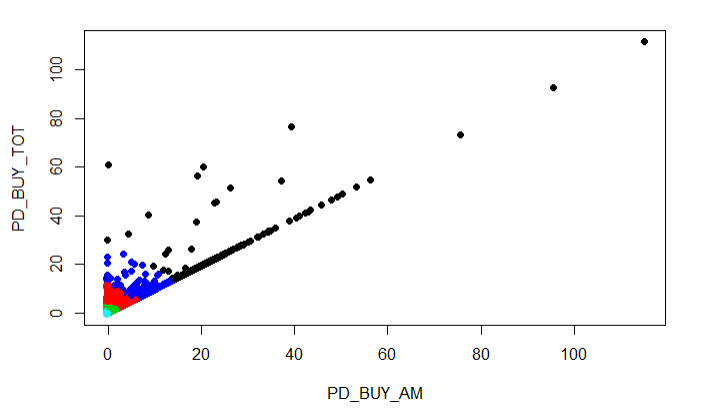
[그림 4-9] 위계적 방법을 이용한 고객 클러스터링

위계적 방법으로도 고객을 클러스터링 하기 위해 역시 연속형 변수 6개를 추출하여 새 변수에 담은 후 NA값을 제거하였다. 이후, 고객 아이디(CLNT\_ID)별로 그룹화를 수행한 후, 수형도에 CLNT\_ID 값이 표시되도록 행이름을 고객코드로 변경하였다. 전체 데이터 셋이 너무 크기 때문에 0.05%의 샘플 데이터를 비복원추출 방식으로 생성하였다. 그 다음으로 변수의 척도화 수행 후 유클리디안 거리척도를 기준으로 유사성행렬을 생성하고 이를 매개변수로 지정하여 클러스터링을 수행하였다.

위계적 고객 클러스터링의 결과를 시각화 했을 때, 금액 또는 분류와 관련한 고객의 특성을 찾기 어려웠고, 따라서 해당 클러스터링 결과도 유의미하지 않다고 판단하였다.

### **제품 클러스터링**

고객 클러스터링을 통해 클러스터링의 목적에 맞는 유의미한 결과를 찾기 어렵다고 판단하였다. 따라서, 다음으로는 제품 클러스터링을 먼저 수행한 후, 제품 클러스터 별 구매 고객의 특성을 판단하고자 제품 클러스터링을 수행하였다.



[그림 4-10] 제품 클러스터링

제품을 클러스터링 하기 위해 먼저 제품과 관련 변수인 상품코드(PD\_C), 상품단가(PD\_BUY\_AM), 총 구매금액(PD\_BUY\_TOT)을 추출하여 새 변수에 담았다. 그 다음으로 상품코드(PD\_C)별로 그룹화를 수행하였고, 상품단가와 총 구매금액은 mean() 함수를 사용하여 처리하였다. 변수의 척도화 수행 후, k-means 방식, 5개의 집단으로 클러스터링을 수행하였다.

제품별 클러스터링 결과, 어느 정도 유의미해보이는 결과가 나왔지만, 제품과 관련한 변수가 가격 데이터밖에 없어 관련 변수가 부족했다. 또한, 클러스터링의 결과가 분류와 관련한 고객의 유의미한 특성을 보여주지 못한다고 판단하여 사용하지 않기로 하였다.

## **분산분석**

### **연령대별**

4.7.1.1. 연령대별 총 구매금액의 평균 차이에 대한 분산분석

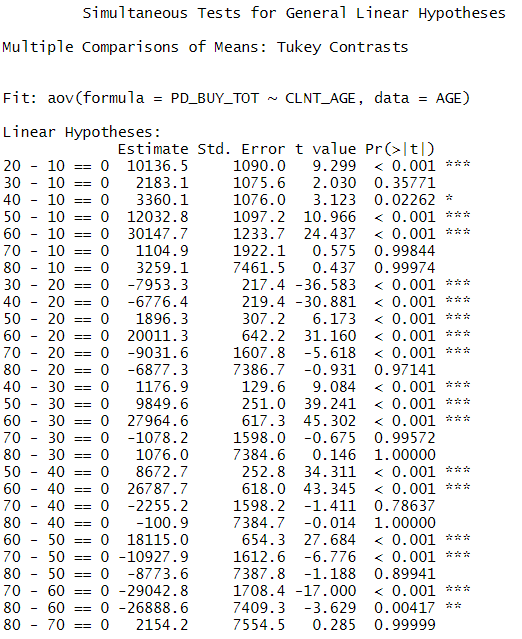
방, 쥐고있는, 테이블, 화면이(가) 표시된 사진

자동 생성된 설명

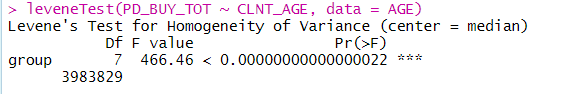
연령대별 총 구매금액에 대한 분산분석 결과, 유의수준(p-value) 0.05 하에 귀무가설을 기각하고 대립가설을 채택하는 것으로 확인하였다. 따라서, 통계적으로 연령대별로 총 구매금액에 대한 평균에 차이가 존재한다고 볼 수 있다.

4.7.1.2. 연령대별 분산분석에 대한 Tukey 사후 검정

위에서 진행한 분산분석에 이어서 수준(Level)의 집단 간에 어디서 평균의 차이가 생겼는지 검증하는 차원에서 Tukey’s HSD(honestly significant difference) test를 진행하였다. 분석결과를 확인해 본 결과, 롯데 온라인 쇼핑몰의 주 고객층인 20, 30, 40대 연령대와 그 외 연령대 사이에서 평균의 차가 뚜렷한 것으로 확인되었다.



4.7.1.3. 연령대별 등분산성 검사



Levene의 등분산성 검정을 이용하여 연령대별 등분산성 여부를 확인하였다. 분석 결과, 연령대별로 등분산을 이룬다는 귀무가설이 기각되었고 대립가설이 채택되었다. 등분산성 가정이 위배되어, 추가적인 ANOVA검정으로 연령별 평균의 차이에 대한 분석을 진행하였다.

4.7.1.4. 연령대별 Welch One-way test

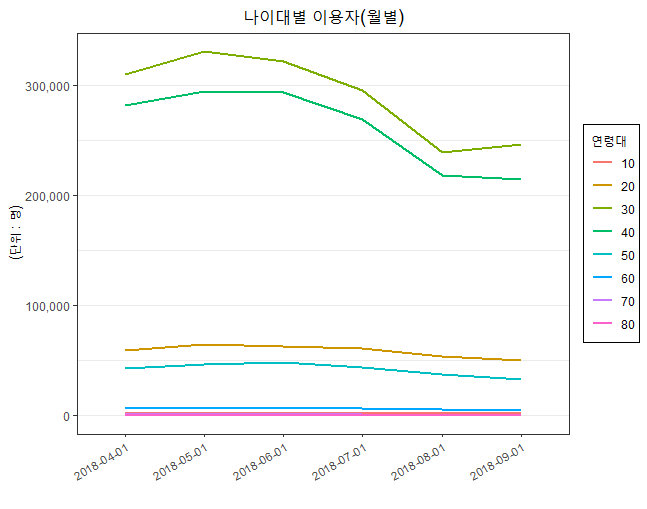
쥐고있는이(가) 표시된 사진

자동 생성된 설명

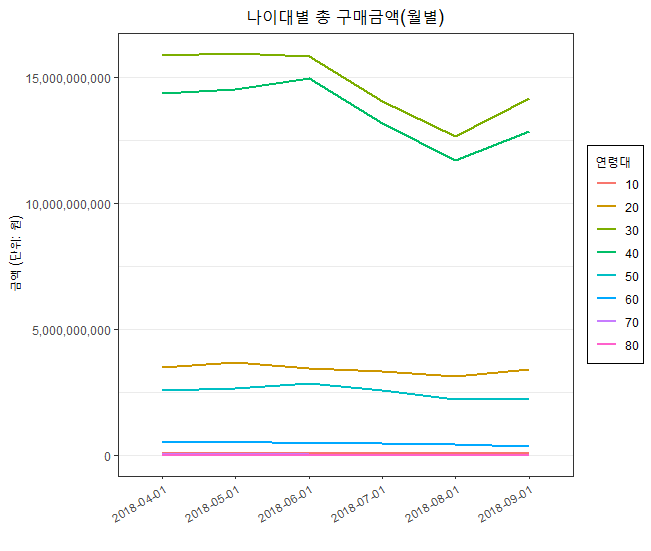
등분산성 가정이 위배되어 추가적인 분석방법으로 연령대별 평균의 차이를 확인이 필요하였다. Welch’s One-Way test는 등분산성 가정이 충족되지 못한 상태에서도 분산분석을 수행시켜준다.

분석결과, 수준별로 평균이 차이가 없다는 귀무가설이 기각되고 대립가설이 채택된다는 사실을 확인할 수 있다.

따라서, 두 번의 분산분석 결과 총 구매금액에 대한 연령대의 집단간에 평균에 차이가 있다는 결과를 입증하였다.



[그림 4-11] 2018년4월 ~ 9월 연령대별 이용자수



[그림 4-12] 2018년4월 ~ 9월 연령대별 총 구매금액

### **기기유형별**

4.7.2.1. 기기유형별 총 구매금액의 평균 차이에 대한 분산분석

테이블, 쥐고있는, 사람들, 방이(가) 표시된 사진

자동 생성된 설명

기기유형별 총 구매금액에 따른 분산분석 결과, 유의수준(p-value) 0.05 하에 귀무가설을 기각하는 것을 확인하였다. 대립가설을 선택하고, 기기유형별로 총 구매금액에 대한 평균에 차이가 통계적으로 존재한다고 볼 수 있다.

4.7.2.2. 기기유형별 분산분석에 대한 Tukey 사후 검정

병, 테이블이(가) 표시된 사진

자동 생성된 설명

연령별 분산분석 이후 진행한 사후분석과 마찬가지로, 디바이스 수준(Level)의 집단 간에 어디서 평균의 차이가 생겼는지 검증하는 차원에서 Tukey’s HSD(honestly significant difference) test를 진행하였다. 분석결과, 롯데 온라인 쇼핑몰을 테블릿으로 이용하는 고객과 모바일(휴대폰)과 데스크톱으로 이용하는 고객 사이에 총 구매금액에 대한 평균의 차이가 명확한 것으로 확인되었다.

4.7.2.3. 기기유형별 등분산성 검사

실내이(가) 표시된 사진

자동 생성된 설명

기기유형에 따른 총 매출금액이 등분산성 가정을 충족하는지 여부를 확인하기 위해 Levene의 등분산성 검사를 진행하였다. 분석결과, 기기유형별로 등분산을 이룬다는 귀무가설이 유의수준(p-value) 0.001에서 기각되므로, 등분산성 가정이 위배되고 추가적인 ANOVA검정으로 기기유형별 평균의 차이에 대한 분석을 진행하였다.

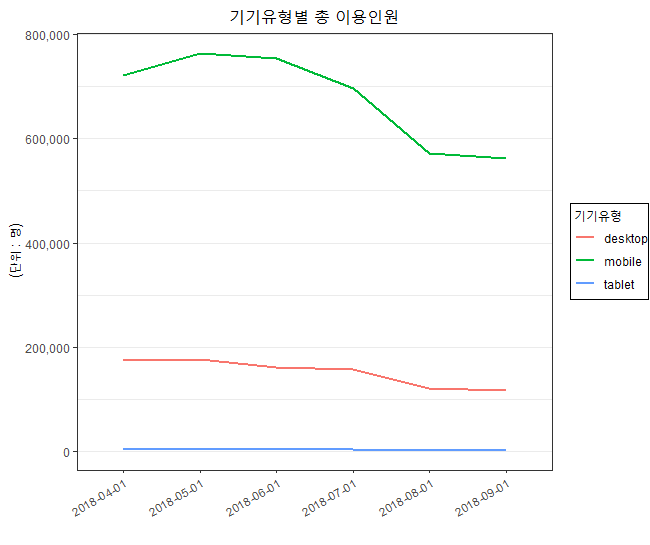
4.7.2.4. 기기유형별 Welch One-way test

화면, 쥐고있는, 테이블, 방이(가) 표시된 사진

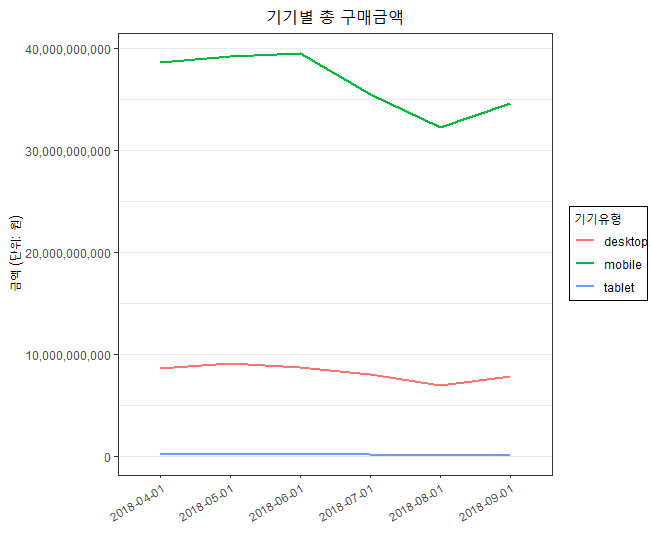
자동 생성된 설명

등분산성 검정 결과 등분산 가정이 위배되어 추가적인 분석방법으로 기기유형별 평균의 차이를 확인하였다. 등분산성 가정이 충족되지 못한 상황에서 Welch’s One-Way test를 수행하여 수준별로 평균이 차이가 유무를 확인하였다.

분석결과, 평균이 같다는 귀무가설이 기각되고 대립가설이 채택되었다. 따라서, 분산분석 결과 총 구매금액에 대한 기기유형의 집단간에 평균에 차이가 있다는 결과를 입증하였다.



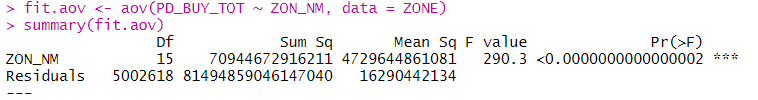
[그림 4-13] 2018년4월 ~ 9월 기기유형별 총 이용자



[그림 4-14] 2018년4월 ~ 9월 기기유형별 총 구매금액

### **지역별**

4.7.3.1. 지역별 총 구매금액의 평균 차이에 대한 분산분석



지역별 총 구매금액에 따른 평균의 차이가 있는지 확인하기 위해서 분산분석 결과를 진행하였다. 유의수준(p-value) 0.05 하에 귀무가설이 기각되어 대립가설을 선택하였다. 따라서, 지역별로 총 구매금액에 대한 평균에 차이가 통계적으로 존재한다고 볼 수 있다.

4.7.3.2. 지역별 분산분석에 대한 Tukey 사후 검정

분산분석 결과에서 귀무가설이 기각된 것을 확인하고, 추가로 수준(Level)의 집단 간 평균의 차이를 확인하는 차원에서 Tukey’s HSD(honestly significant difference) test를 진행하였다.

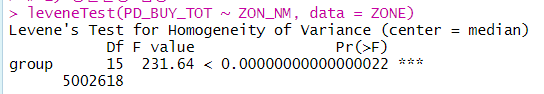
결과를 확인해보니, 서울 지역이 매출이 제일 높았고 그 뒤로 경기도와 부산이 뒤따랐다. 롯데 온라인 쇼핑몰의 고객들의 60% 이상은 서울에 거주하였으며, 다른 지역의 고객들과 총 구매금액의 평균이 달랐다.

텍스트이(가) 표시된 사진

자동 생성된 설명

특히 서울, 경기도, 부산에 사는 고객들과 다른 지역의 고객들 사이에 구매금액 평균의 차이가 통계적으로 유의미하다는 결과를 얻을 수 있었다. 위에 있는 지역간 평균의 차이 검정 결과는 내용이 많은 관계로 자세한 결과는 부록 부분에 남겼다.

4.7.3.3. 지역별 등분산성 검사



Tuckey 사후 검정 이후, 지역별 총 구매금액에 차이가 있다는 사실을 확인할 수 있었다. 다음 분석을 진행하기 전에, 위에서 진행한 분산분석 결과가 등분산성 가정을 충족된 상태로 진행되었는지 판별하는 절차가 필요하였다.

Levene의 등분산성 검사를 진행하였고, 지역별로 등분산을 이룬다는 귀무가설이 유의수준(p-value) 0.001에서 기각된다. 따라서, 등분산성 가정이 위배되는 대립가설이 채택되었고, 추가적으로 다른 방식의 ANOVA검정으로 기기유형별 평균의 차이에 대한 분석을 진행하였다.

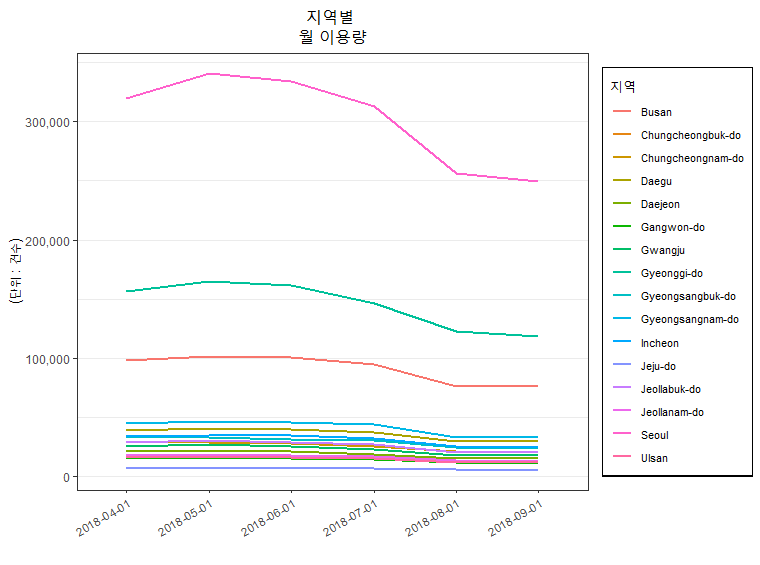
4.7.3.4. 지역별 Welch One-way test

쥐고있는, 손, 화면, 방이(가) 표시된 사진

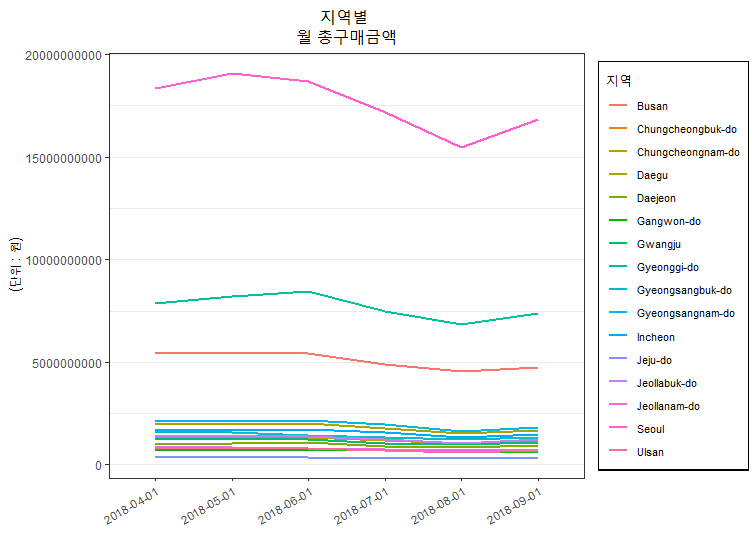
자동 생성된 설명

등분산성 검정 결과 기본 가정이 위배되어 Welch’s One-Way test를 수행하여 지역별로 평균의 차이가 있는지 확인하였다. Welch’s 분산분석 검정은 등분산성 가정이 충족되지 않는 경우에도 진행이 가능한 방법이다.

분석결과, 귀무가설(수준별로 평균이 같다)은 기각되고 대립가설이 채택되었다. 따라서 지역 간에 총 구매금액에 대한 평균의 차이가 있다는 결과를 입증하였다.



[그림 4-15] 2018년 4월 ~ 9월 지역별 총 이용자



[그림 4-16] 2018년 4월 ~ 9월 지역별 총 구매금액

### **고객등급별**

롯데 온라인 데이터에서 추가적인 파생 변수를 생성해 분산분석을 진행하였다. 고객 아이디(CLNT\_ID)별로 총 구매금액이 상이하다는 사실을 확인하였고, 이것을 의미 있는 집단으로 그룹화하는 작업을 진행하였다.

롯데 그룹 계열사 중 롯데 하이마트와 롯데 마트를 기준으로 고객 등급을 총 구매금액별로 구분하였다. 특히 롯데 하이마트가 제공하는 고객 등급표를 참고하여, 6개월동안 1000만원 이상 소비하는 고객에게 VVIP 등급을 부여하였고, 250만원에서 1000만원 미만 소비하는 고객에게는 VIP 그리고 그 이하를 소비하는 고객에게 CUST라는 등급을 부여하였다.

고객 아이디(CLNT\_ID)별로 총 922,618명의 고객에게 위 같은 기준으로 등급을 부여하였다. 전체 인원 중 472(0.0005%) 명이 VVIP 등급을 받았고, VIP는 9908(0.1%)명 그리고 나머지 912238명의 고객이 CUST 일반 등급을 받았다.

4.7.4.1. 고객등급별 총 구매금액의 평균 차이에 대한 분산분석

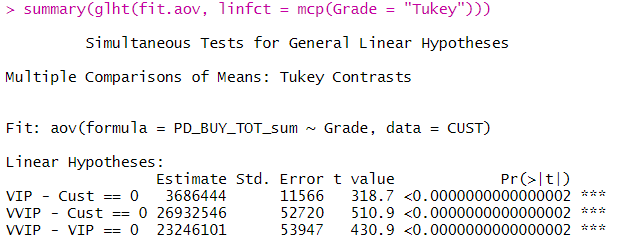
테이블이(가) 표시된 사진

자동 생성된 설명

자체적으로 만든 고객등급 파생변수를 분산분석 방법을 이용하여 총 구매금액에 따른 평균의 차이를 확인하였다.

총 세 개의 고객등급별로 ANOVA test를 진행하였고, 유의수준(p-value) 0.05 하에 귀무가설을 기각하였다. 이는 대립가설을 채택하는 결과로, 통계적으로 고객등급별 총 구매금액에 평균의 차이가 존재한다고 볼 수 있다.

4.7.4.2. 고객등급별 분산분석에 대한 Tukey 사후 검정



고객등급에 따른 분산분석 결과에서 대립가설이 채택되는 것을 확인하고, 수준(Level)의 집단 간 평균의 차이를 확인하는 Tukey’s HSD(honestly significant difference) 사후검정을 진행하였다.

분석결과, 세 등급 모두 총 구매금액에 따른 평균의 차이가 나타나는 것을 확인할 수 있었다. 위에 제시된 검정결과를 확인해 보면, VVIP와 나머지 고객 등급 사이에 예상 표준편차의 크기가 크다는 것을 확인할 수 있다.

4.7.4.3. 고객등급별 등분산성 검사

테이블이(가) 표시된 사진

자동 생성된 설명

고객등급별 분산분석 결과에서 평균에 차이가 있다는 사실을 확인하였고, 이 분석 결과가 등분산성 가정을 충족시킨 상태에서 진행되었는지 여부를 확인하는 절차가 필요하였다.

위 결과를 확인해본 결과, 귀무가설(등분산성 가정이 만족)이 기각되고 대립가설이 채택되었다는 것을 확인할 수 있다. 이것은 등분산성 기본 가정을 위배한 것으로 4.7.4.1. 에서 진행했던 분산 분석의 결과에 오류가 있을 수 있다는 것을 의미한다. 따라서, 다른 분석기법을 활용해 분산분석을 진행하는 게 바람직하다.

4.7.4.4. 고객등급별 Welch One-way test

쥐고있는, 방이(가) 표시된 사진

자동 생성된 설명

Welch’s One-Way test를 통해서 등분산성이 만족되지 못한 데이터셋의 분산분석을 진행하였다. 고객 등급별로 평균의 차이가 있기에 1차 ANOVA test로 확인하였고, Welch’s 분산분석 검정을 통해 추가적인 확인 절차를 거쳤다.

위 결과에서 알 수 있듯이 귀무가설(수준별로 평균이 같다)이 기각되고 대립가설이 채택된다는 사실을 입증할 수 있다. 즉, 고객등급 간에 총 구매금액에 대한 평균값의 차이가 있다고 볼 수 있다.

텍스트, 지도이(가) 표시된 사진

자동 생성된 설명

[그림 4-17] 2018년4월 ~ 9월 고객등급별 총 이용자

텍스트, 지도이(가) 표시된 사진

자동 생성된 설명

[그림 4-18] 2018년4월 ~ 9월 고객등급별 총 구매금액

### **상품등급별**

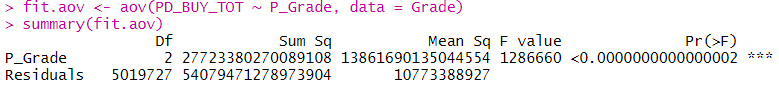
고객등급과 같은 방법으로 추가적인 파생변수를 생성하였다. 상품의 단가를 기준으로 박스 플롯(Boxplot)을 생성하였고 IQR을 기준으로 상품을 구분하였다.총 세 가지 수준(level)로 상품을 분류했고, 분산분석을 진행하였다.

L.POINT 데이터에서 제공하는 상품이 롯데 하이마트를 비롯한 롯데마트, 백화점 그리고 홈쇼핑 등 다양한 채널에서 가져온 것을 확인하였다. 가전기기부터 일반 생필품까지, 상품 분포가 매우 크고 다양하다는 사실에서 분류의 필요성을 느꼈다.

박스 플롯(Boxplot)을 참고해서 상품 단가가 131,550원 이상인 제품을 High, 29,000원 이상 131,550원 미만 제품은 Middle, 29,000원 미만의 제품을 Low라는 등급을 부여하였다.

총 5,019,730개의 상품을 위에서 제시한 기준으로 상품 등급을 부여하였다. 전체 상품 중 High 등급은 323,795개(6.4%), Middle등급은 2,067,440개(41%), 그리고 나머지 2,628,495개(52%) 상품이 Low 등급으로 분류되었다..

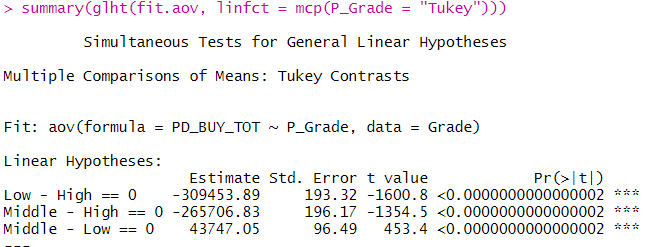
4.7.5.1. 상품등급별 총 구매금액의 평균 차이에 대한 분산분석



상품 단가에 따른 상품등급 파생변수를 One-Way ANOVA 분산분석을 진행하여 상품등급별 총구매금액의 평균의 차이를 확인하였다.

분석결과, 유의수준(p-value) 0.05 하에 귀무가설이 기각되었다. 통계적으로 데이터 셋에서 상품등급에 따른 집단 간 평균의 차이가 존재한다고 볼 수 있다.

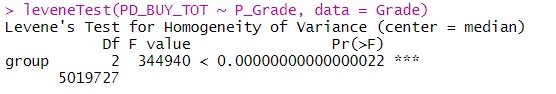
4.7.5.2. 상품등급별 분산분석에 대한 Tukey 사후 검정



분산분석 결과에서 대립가설이 채택되어 상품등급별로 평균에 차이가 존재한다는 것을 확인하였고, Tukey’s HSD(honestly significant difference) 사후검정으로 구체적으로 수준(Level)의 집단 간 차이를 확인하였다.

사후분석 결과, 상품 등급 세 분류 모두 총 구매금액에 따른 평균의 차이가 나타나는 것이 관찰되었다. 검정결과를 확인해 보면, High 등급을 상품과 나머지 Middle과 Low 상품에서 예상 표준편차의 음수 값이 나온다는 것을 확인할 수 있다.

4.7.5.3. 상품등급별 등분산성 검사



위에서 진행했던 분석 절차와 마찬가지로, 상품등급에서 등분산성 가정이 충족되었는지 여부를 판단하기 위해서 Levene test를 진행하였다.

등분산성 검정을 진행해본 결과, 등분산성 가정(귀무가설)이 기각되고 대립가설이 채택되었다는 것을 확인할 수 있다. 따라서, 등분산성 기준이 완화된 다른 분석기법을 활용해 분산분석을 진행하는 게 필요하다고 판단된다.

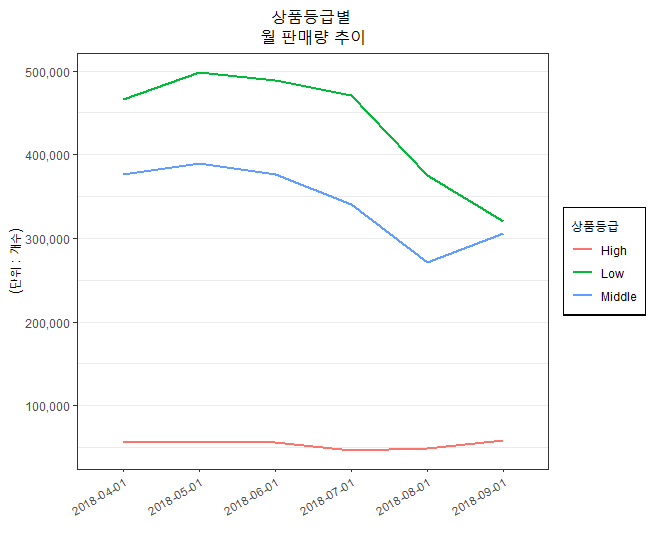
4.7.5.4. 고객등급별 Kruskal-Wallis One-way analysis of variance

스크린샷이(가) 표시된 사진

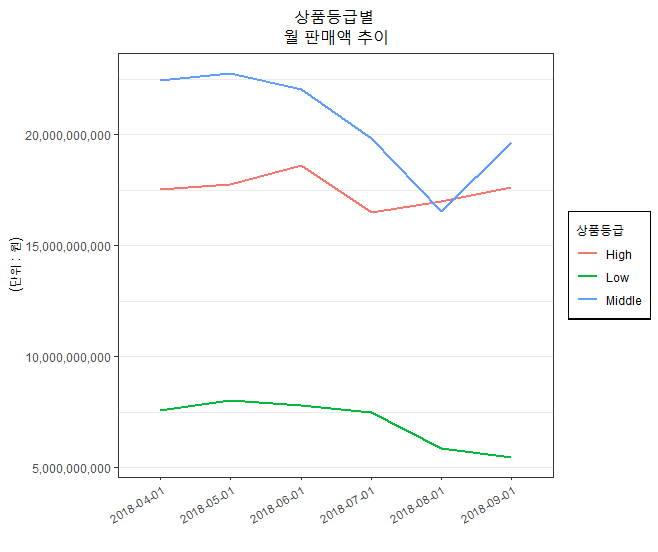
자동 생성된 설명

Kruskal-Wallis One-way analysis of variance 검정을 통해 분산분석을 하였다. 이 분석 방법은 등분산성 가정이 충족되지 못한 데이터셋에, 비모수적인 분석 기법을 이용해 분산분석을 진행할 수 있다.

분석결과, p-value 값이 0.05 수준에서 귀무가설이(수준별로 평균이 같다) 기각되고 대립가설이 채택된다는 사실을 입증할 수 있다. 위에서 진행했던 Welch test와 마찬가지로 등분산성 가정이 완화된 상태에서 상품등급 간에 총 구매금액에 대한 평균값의 차이가 있다고 결과를 확인할 수 있다.



[그림 4-19] 2018년 4월 ~ 9월 상품등급별 총 판매량



[그림 4-20] 2018년 4월 ~ 9월 상품등급별 총 구매금액

# **소비자선호지수 생성**

## **선호지수 생성 배경**

시계열 분석, 회귀분석, 분산분석으로 전반적인 롯데 온라인 쇼핑의 구매행태를 조사해 봤다. 분석결과, 분류별 그리고 기간별로(주, 월)매출에 차이가 있다는 것을 관찰할 수 있었다. 특히 시계열 분석의 세 가지 예측 방법을 이용해서, 롯데 온라인 쇼핑몰의 총 매출금액은 하향 추세를 보이고 있다는 점에서 반등을 위한 대책이 마련되어야 한다고 생각되었다.

추가적인 분석을 진행하면서, 매출이 제일 높았던 5월 대비 매출이 낮았던 8월의 상품 군(소분류)의 판매빈도가 다르다는 사실을 확인하였고, 이에 따라 상품 군별 판매빈도의 차이를 정량화된 수치로 바꿔서 구매 경향을 파악해, 롯데 온라인 쇼핑의 매출 하락을 제고하기 위한 마케팅 방안으로 전환하는 것이 필요하다고 생각하였다.

정량화된 도구로서 “선호지수”을 개발해 상품군(소분류)을 분류와 기간별 구매 경향이 높은 상품군을 판별할 수 있는 도구로 활용할 수 있다는 점에서 개발의 필요성을 느꼈다. 상품군에 따른 선호지수를 파악해 회사입장에서 어떤 판매전략을 세우는 중요한 지표로서 활용될 수 있고, 차별화된 마케팅 방법을 마련하는데 도움될 것으로 판단된다.

## **선호지수 생성 방법**

상품 군에 따른 선호지수를 생성하는데, 어떤 분류를 선택할 것인지 선행되어야 한다.

롯데 온라인 데이터에서는 상품의 대/중/소 세 분류를 제공한다. 차별화된 마케팅 방법으로 활용되기 위해서는 상품 소분류별로 선호지수를 생성하는 것이 필요하다고 생각했고, 상품 소분류에 따른 선호지수를 생성하였다.

우선, 전체 롯데 데이터셋에서 상품 소분류 변수와 정량적인 변수를 선별했다. 총 898개의 상품 소분류에 따라 개별 변수들을 합산하였다. 추가적으로 데이터 가공 과정에서 개별 상품 군별로 판매빈도 파생변수를 생성했다.

새로운 파생변수(판매빈도)를 종속변수로 하고 총 페이지뷰(TOT\_PAG\_VIEW\_CT), 총 구매금액(PD\_BUY\_TOT), 총 세션시간(TOT\_SESS\_HR\_V), 세션시퀀스(SESS\_SEQ), 히트 시퀀스(HITS\_SEQ)를 설명변수로 두고 회귀분석을 진행하였다. 분석을 실행하기 앞서, 변수별로 값들의 범위 격차가 클 것으로 예상되어 표준화 작업을 선행하고 분석을 진행하였다.

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

회귀분석 결과 결정계수()는 0.9909로 높게 나왔지만, 선호지수를 생성할 때 변수 사이의 가중치를 설정하기 위해서 분석을 진행하였다. 분석결과를 참고해서 회귀계수 값이 제일 높은 총 세션시간(TOT\_SESS\_HR\_V)을 기준으로 나머지 네 변수의 비율 값으로 선호지수 가중치의 값을 선택하였다.

## **선호지수 생성**

위에서 진행한 회귀분석을 바탕으로 소비자 선호지수를 생성하였다. 베타(β)는 개별 변수에 대한 가중치이고, X는 선호지수 생성에 사용한 변수를 의미한다. 회귀분석 결과를 토대로, 총 세션시간은 1을 기준으로 하고, 총 구매금액(PD\_BUY\_TOT)의 가중치는 0.04, 히트 시퀀스(HITS\_SEQ)의 가중치는 0.4, 총 페이지 뷰(TOT\_PAG\_VIEW\_CT)는 -0.55로 설정한 후 이 가중치 값을 선호지수 생성에 반영하였다. 아래 표는 2018년 4월부터 9월까지 총 6개월 전체 고객이 구매한 상품에 선호지수를 반영한 결과이다. 대표적으로 상/하위 각각 15개만을 첨부하였다.



[그림5-1] 2018년 4월 ~ 9월 기간 선호지수 상위, 하위 15개 상품

추가로, 매출이 제일 높았던 5월과 가장 낮았던 8월의 선호되는 상품군의 선호지수를 반영하여 결과를 도출하였다.



[그림 5-2] 5월과 8월의 선호지수 상위 15개 상품

두 기간을 분석해본 결과, 상품 선호간 변동이 있다는 사실을 관찰할 수 있다. 5월은 남성티셔츠 선호지수가 4418.445로 가장 높았던 반면, 8월에는 여성원피스 선호지수가 3618.456점으로 제일 높다. 전반적으로 8월의 상품 선호지수가 5월에 비해서 낮다는 특징을 확인할 수 있다. 선호지수를 생성할 때 총 구매금액이 포함되었다는 점에서, 전체적으로 매출이 하락한 8월에 총 구매금액이 낮음과 동시에 그 결과 선호지수가 전체적으로 낮아졌다는 인사이트를 도출할 수 있다.

# **서비스 제안 및 결론**

Turn A 전략을 제시하여 두 가지 서비스를 제안하고자 한다.

Turn A 전략의 의미는 A 는 ALL을 뜻하고 합집합 기호 ∀는 for all을 의미하며, 모두를 위해 관점을 바꾸어 새로운 전략을 제공하고자 한다.

* 1. **역시즌 프로모션**

최근 SNS와 유튜브 등 온라인 매체가 활성화되면서 온라인 소비자들의 구매패턴이 변하고 있다. 이번 롯데 온라인 행동 데이터에서도 이러한 경향이 확인되었다.

30대 여성 월별 데이터 셋을 이용해서 상품 군별 선호지수를 도출한 결과, 소비자들의 독특한 구매패턴을 관찰할 수 있었다. 롯데 온라인 쇼핑에서 여성 고객이 85%정도 비중을 차지하고, 30대 연령대가 주 구매층이며 그리고 롯데 온라인 쇼핑 이용 고객 중 서울에서 구매하는 사람이 많다는 점에서 분석의 대상으로 선정하였다.



[표] 5월 선호지수 대비 8월 선호지수 상승한 상품 15개

우선, 매출이 제일 높았던 5월과 제일 낮았던 8월의 상품 군별 선호지수를 구하였다. 그리고 같은 상품 소분류별 선호지수 격차가 큰 순서대로 위와 같은 표를 작성하였다. 선호지수 격차가 높을수록 5월에 특정 상품 군을 구매한 빈도가 8월에는 확연하게 상승했다는 것을 의미한다.

가장 특징적인 부분은 8월 판매된 상품 군들 중에서 패딩 제품들이 상위부분에 분포해 있는 것을 확인할 수 있다. 패딩이 대체로 겨울에 많이 구매하고 이용한다는 제품임을 감안하면, 굉장히 독특한 구매 패턴이라고 볼 수 있다.

따라서 역시즌 프로모션을 활성화하는 서비스를 제안한다. 이미 온라인 상으로 역시즌 프로모션은 많이 진행되고 있다. 소비자들은 여름에 저렴한(할인) 가격으로 겨울 제품을 구매하기를 선호한다. 따라서, 롯데 기업 입장에서 이러한 소비자 구매 행태를 인지하고, 더욱 역시즌을 활성한다면 8월 매출 향상에 도움이 될 것으로 판단된다. 겨울 용품을 판매하는 브랜드를 선정해서 8월 기간동안 세일 및 판매 프로모션을 촉진시켜 매출 하락을 반등하는 경영 전략을 마련할 필요가 있다고 판단된다.

* 1. **새로운 타겟팅**

롯데 상품군들 중, 일부 남성 제품들이 성별에 상관없이 많이 판매되는 것이 관찰된다. 특히, 판매 빈도가 제일 높은 남성티셔츠의 경우 여성 30대 고객층에서도 많이 구하는 것으로 보인다.



[표] 30대 여성 4월, 8월 선호지수상위 상품 15개

매출의 가장 높은 비율을 차지하는 서울에서 구매하는 30대 여성 고객의 4월과 8월 선호지수 상위 15개 상품 군을 표로 정리하였다. 여성 고객을 대상으로 데이터를 가공하여 선호지수를 구했지만, 놀랍게도 남성 제품들이 상위권에 분포해 있다는 것을 관찰할 수 있다.

이를 통해 여성 고객들도 남성 제품을 많이 구매한다는 사실에서, 여성들을 대상으로도 남성 상품에 대해 소개하는 ‘NEW 타겟팅’ 프로모션을 제안하고자 한다. 전략의 예시로 남성이 선호하는 제품을 추천 또는 남/녀가 맞춰 입을 수 있는 코디상품, 여성이 보았을 때 남성에게 꾸며줄 수 있는 상품 등을 어플리케이션 내에 적합한 남성의 모델을 생성한 후 구매하고자 하는 상품을 입혀봄으로써 상품구매에 도움을 줄 수 있다. 이를 통해 거래에 대한 확실성과 관심을 유발하여 구매에 이어질 수 있도록 제공할 수 있을 것이다.

# **부록**

* 1. **변수별 기초 통계량 SAS 출력물**

1. PD\_BUY\_AM (상품 단가)

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 3970214 | **Sum Weights** | 3970214 |
| **Mean** | 51135.0857 | **Sum Observations** | 2.03017E11 |
| **Std Deviation** | 115098.783 | **Variance** | 1.32477E10 |
| **Skewness** | 15.070338 | **Kurtosis** | 556.966908 |
| **Uncorrected SS** | 6.29776E16 | **Corrected SS** | 5.25963E16 |
| **Coeff Variation** | 225.087689 | **Std Error Mean** | 57.7648662 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.328732 | **Pr> D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 139006.6 | **Pr> W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 704473.2 | **Pr> A-Sq** | <0.0050 |

1. PD\_BUY\_CT (상품 구매수량)

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 3970214 | **Sum Weights** | 3970214 |
| **Mean** | 1.15036217 | **Sum Observations** | 4567184 |
| **Std Deviation** | 2.76194176 | **Variance** | 7.62832228 |
| **Skewness** | 186.189229 | **Kurtosis** | 59080.5032 |
| **Uncorrected SS** | 35539980 | **Corrected SS** | 30286064.3 |
| **Coeff Variation** | 240.093236 | **Std Error Mean** | 0.00138614 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.478292 | **Pr> D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 293948.9 | **Pr> W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 1372027 | **Pr> A-Sq** | <0.0050 |

1. PD\_BUY\_TOT (상품 총 구매금액)

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 3970214 | **Sum Weights** | 3970214 |
| **Mean** | 52907.9665 | **Sum Observations** | 2.10056E11 |
| **Std Deviation** | 117981.374 | **Variance** | 1.39196E10 |
| **Skewness** | 15.7932164 | **Kurtosis** | 634.455931 |
| **Uncorrected SS** | 6.63774E16 | **Corrected SS** | 5.52638E16 |
| **Coeff Variation** | 222.99359 | **Std Error Mean** | 59.2115584 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.327904 | **Pr> D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 139308.1 | **Pr> W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 705613.3 | **Pr> A-Sq** | <0.0050 |

1. HITS\_SEQ (구매하기까지 모든 행위)

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 3970214 | **Sum Weights** | 3970214 |
| **Mean** | 88.4976644 | **Sum Observations** | 351354666 |
| **Std Deviation** | 89.3879947 | **Variance** | 7990.21359 |
| **Skewness** | 1.93188052 | **Kurtosis** | 3.87441131 |
| **Uncorrected SS** | 6.28169E10 | **Corrected SS** | 3.17228E10 |
| **Coeff Variation** | 101.00605 | **Std Error Mean** | 0.04486134 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.170844 | **Pr> D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 47625.16 | **Pr> W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 268567.5 | **Pr> A-Sq** | <0.0050 |

1. TOT\_SESS\_HR\_V (총 세션에 머무른 시간(단위: 초))

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 3970214 | **Sum Weights** | 3970214 |
| **Mean** | 1683.589 | **Sum Observations** | 6684208614 |
| **Std Deviation** | 1700.30297 | **Variance** | 2891030.19 |
| **Skewness** | 3.19590191 | **Kurtosis** | 29.5718377 |
| **Uncorrected SS** | 2.27315E13 | **Corrected SS** | 1.1478E13 |
| **Coeff Variation** | 100.992758 | **Std Error Mean** | 0.8533346 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.164238 | **Pr> D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 33670.52 | **Pr> W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 194930.3 | **Pr> A-Sq** | <0.0050 |

1. TOT\_PAG\_VIEW\_CT (총 시청 페이지 수)

| **Moments** | | | |
| --- | --- | --- | --- |
| **N** | 3970214 | **Sum Weights** | 3970214 |
| **Mean** | 111.690703 | **Sum Observations** | 443435993 |
| **Std Deviation** | 105.447245 | **Variance** | 11119.1215 |
| **Skewness** | 1.69090858 | **Kurtosis** | 2.63333709 |
| **Uncorrected SS** | 9.3673E10 | **Corrected SS** | 4.41453E10 |
| **Coeff Variation** | 94.4100468 | **Std Error Mean** | 0.05292103 |

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Kolmogorov-Smirnov** | **D** | 0.157769 | **Pr> D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 39974.23 | **Pr> W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 231168.7 | **Pr> A-Sq** | <0.0050 |

* 1. **t-test (성별 간 비교)**

1. PD\_BUY\_AM (상품 단가)

* 등분산성 검정

조류이(가) 표시된 사진

자동 생성된 설명

* 윌콕슨 검정

테이블, 방, 조류이(가) 표시된 사진

자동 생성된 설명

* 박스 플롯

스크린샷이(가) 표시된 사진

자동 생성된 설명

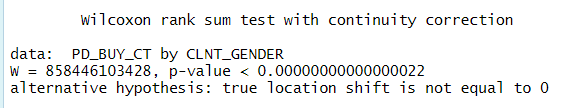
1. PD\_BUY\_CT (상품 구매수량)

* 등분산성 검정

조류이(가) 표시된 사진

자동 생성된 설명

* 윌콕슨 검정



* 박스 플롯

스크린샷이(가) 표시된 사진

자동 생성된 설명

1. HITS\_SEQ (구매하기까지 모든 행위)

* 등분산성 검정

조류이(가) 표시된 사진

자동 생성된 설명

* 윌콕슨 검정

테이블, 쥐고있는, 목재의, 방이(가) 표시된 사진

자동 생성된 설명

* 박스 플롯

스크린샷이(가) 표시된 사진

자동 생성된 설명

1. SESS\_SEQ

* 등분산성 검정

조류이(가) 표시된 사진

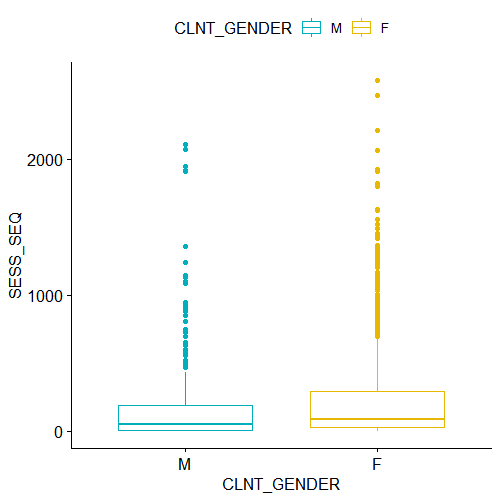
자동 생성된 설명

* 윌콕슨 검정

방, 테이블, 오렌지, 목재의이(가) 표시된 사진

자동 생성된 설명

* 박스 플롯



1. TOT\_SESS\_HR\_V (총 세션에 머무른 시간(단위: 초))

* 등분산성 검정

조류이(가) 표시된 사진

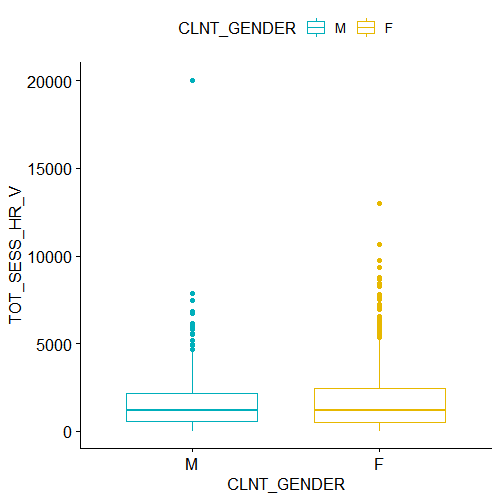
자동 생성된 설명

* 윌콕슨 검정

방, 테이블, 쥐고있는, 목재의이(가) 표시된 사진

자동 생성된 설명

* 박스 플롯



1. TOT\_PAG\_VIEW\_CT (총 시청 페이지 수)

* 등분산성 검정

조류이(가) 표시된 사진

자동 생성된 설명

* 윌콕슨 검정

목재의, 테이블, 검은색, 오렌지이(가) 표시된 사진

자동 생성된 설명

* 박스 플롯

스크린샷이(가) 표시된 사진

자동 생성된 설명