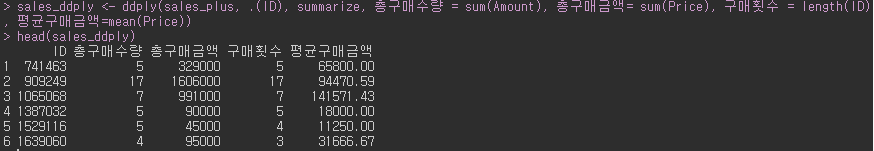
**빅데이터경영MBA U2015046 이 재현**

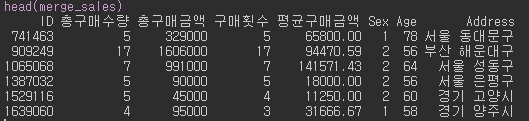
**1. sales.csv 파일은 2014년 1년간 3507명의 구매정보에 관한 데이터이고 demo.csv는 각 고객의 개인정보를 포함하고 있다.**

1. sales.csv 데이터를 사용해 각 고객에 대한 아래의 변수를 생성하고 demo.csv의 데이터와 merge 하시오. (price가 양의 값을 갖는 자료만 사용)

* 총 구매수량
* 총 구매금액
* 구매횟수
* 평균 구매금액



* 나이



1. 산점도, 히스토그램을 살피면서 이상점이 있는지 파악하고 log 변환된 산점도와 비교하여 변수변환의 필요성 여부를 결정하시오.

<구매수량에 따른 총 구매금액>

|  |  |
| --- | --- |
|  |  |

이상값 Id : 348197478



<총 구매수량에 따른 총 구매금액>

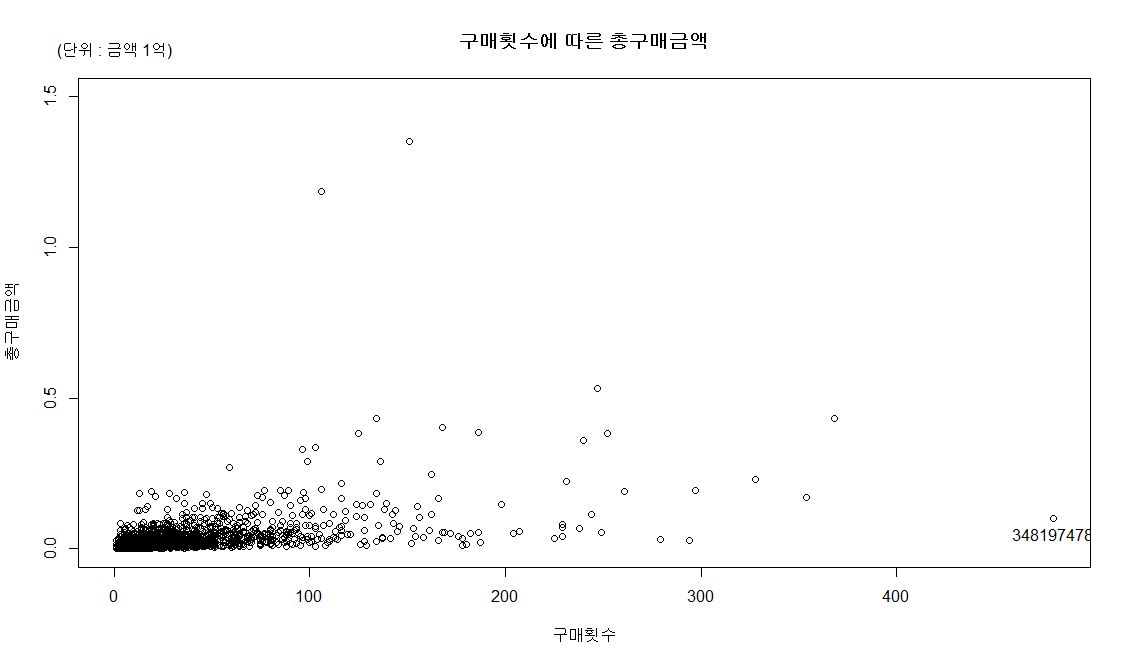
|  |  |
| --- | --- |
|  |  |

이상값 Id : 231374882

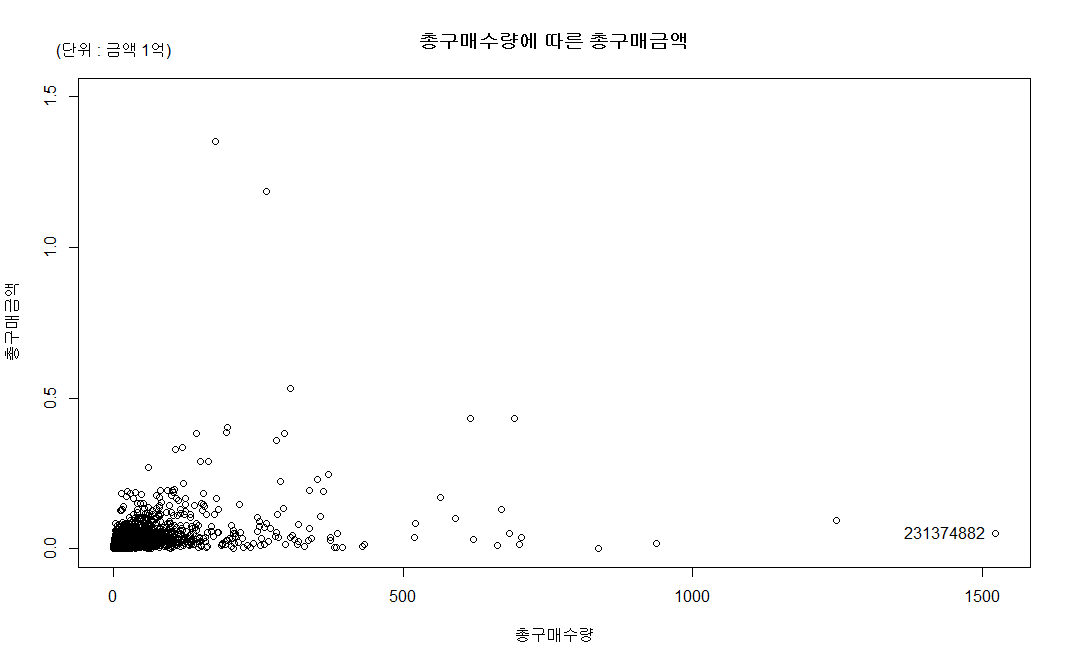


Log 변환 값과 그렇지 않은 값을 그래프를 통해 봤을 때, 변수 변환을 취하더라도 이상치 값은 변함이 없기 때문에 변수변환의 필요성을 느끼지 않음.

1. 변환된 변수 간의 산점도, 히스토그램을 살피면서 변수 간의 관계를 파악하고 해석하시오.



구매횟수가 대체적으로 0~100회 사이에 많이 분포되어 있고, 구매금액은 0~5억 사이에 주로 분포되어 있음을 알 수 있다.



구매수량은 0건에서 약 200건 사이에 많이 분포되어 있음을 알 수 있고, 구매금액은 0~5억 사이에 분포되어 있음을 알 수 있다.

1. 각 고객의 카테고리별 총 구매수량, 총 구매금액, 구매횟수, 평균 구매금액을 계산하고 히스토그램 행렬을 사용하여 카테고리별로 각 변수의 변화를 관찰하고 발견한 점을 서술하시오. (필요시 변수변환된 자료 사용)