DATA MINING TERM PROJECT

**HUFSURANCE**

****

201400944 김홍범

201601785 신지해

201601220 박상아

**목 차**

**.**

**1. Data Preprocessing 전처리**

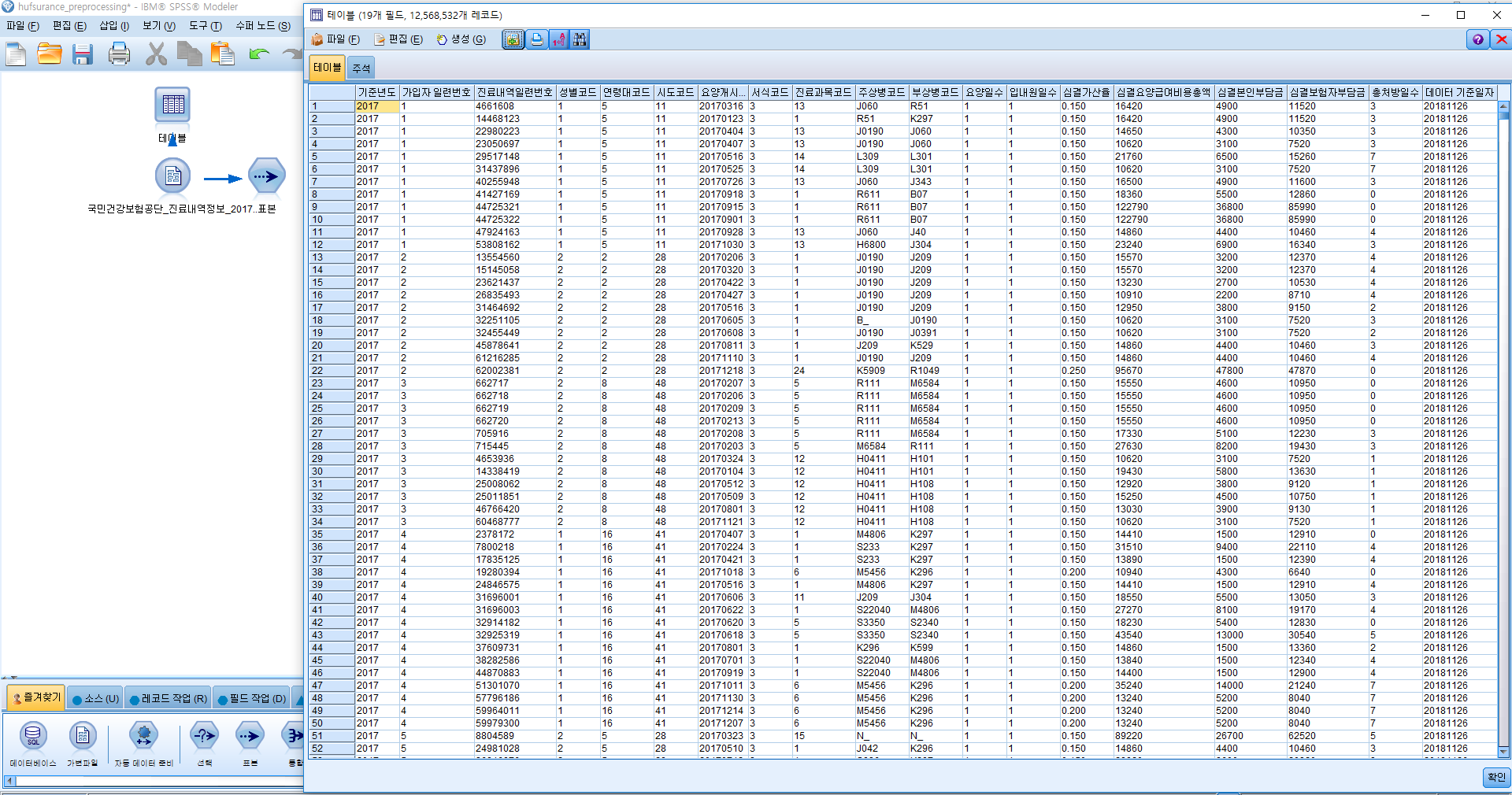
**2. 의사결정나무 분석**

**3. 군집 분석(clustering)**

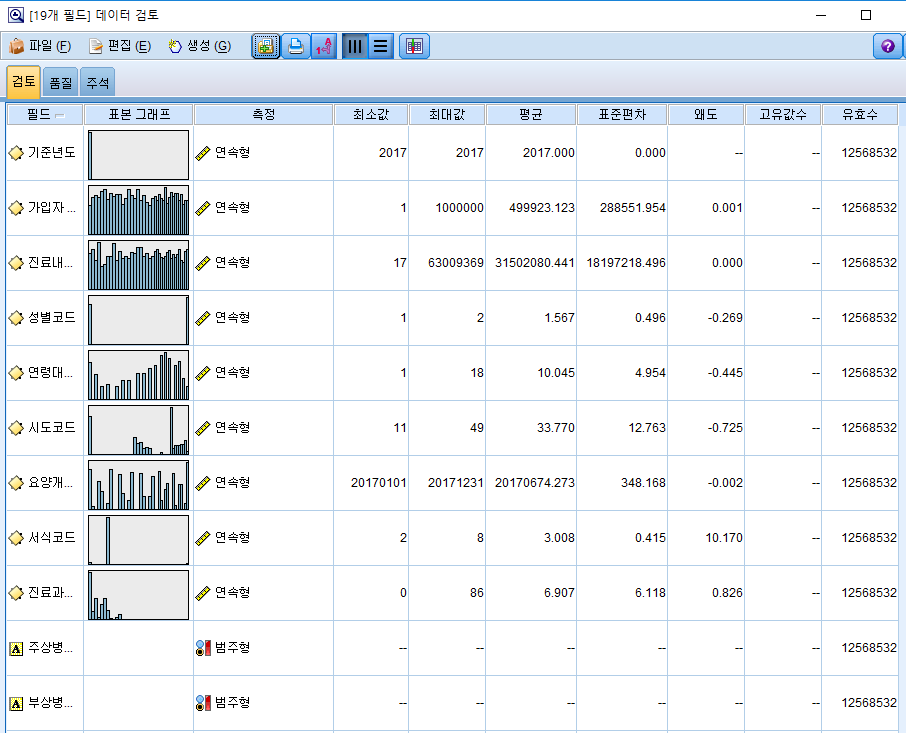
**4. 연관성 분석**

1. Data Preprocessing 전처리

1) row data : 19개 필드, 12,568,532개 레코드



2) row data의 데이터 검토 : 모든 attribute의 유효수가 12,568,532이므로 missing value가 없다.





데이터 검토 결과를 통해 각 attribute를 점검하면 다음과 같다.

1. 첫번째와 마지막 attribute인 *기준년도*와 *데이터 기준일자*는 무의미하므로 삭제한다.

2. *가입자 일련번호* : 총 1,000,000명의 환자에 대한 데이터이다.

3. *연령대코드* : 특정 연령대에 대한 진료 빈도수에 유의미한 분포를 보인다. 대략적으로 0-9세의 영유아, 미취학 아동 및 초등 저학년의 진료빈도가 높고, 청장년층으로 갈수록 빈도가 감소하다가 중장년층 및 노년층을 향하며 빈도가 증가한다.

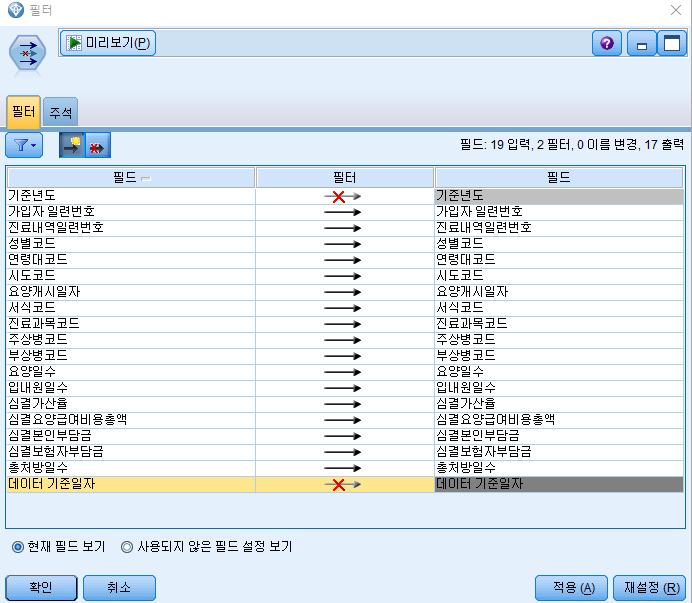
4. *요양개시일자* : 요양기관에 내원한 일자인 요양개시일자의 분포는 일정한 cycle을 형성한다.

5. *시도코드, 서식코드, 진료과목코드, 심결가산율* : 다음의 네 가지 attribute는 공통적으로 특정 항목의 빈도가 월등히 높다. 이후 범주형으로 분류된다.

6. *요양일수, 입내원일수, 심결요양급여비용총액, 심결본인부담금, 심결보험자부담금, 총처방일수*

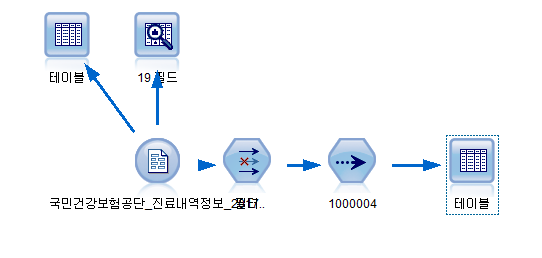
: 다음 항목들은 공통적으로 왜도가 높다. Outlier 판별과 데이터 분리를 전제로 한 분석이 요구된다.

3) 분석에 사용되는 attribute목록



4) 약 백 만개의 표본 추출(가입자 일련번호 79818번까지)

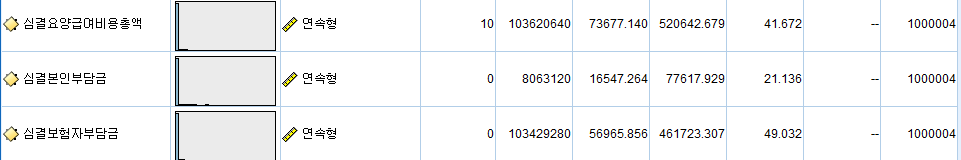
: row data의 용량이 커 원활한 프로그램 작동이 어렵기 때문에, 처음부터 백 만개의 표본을 추출하여 이후 분석을 진행한다.



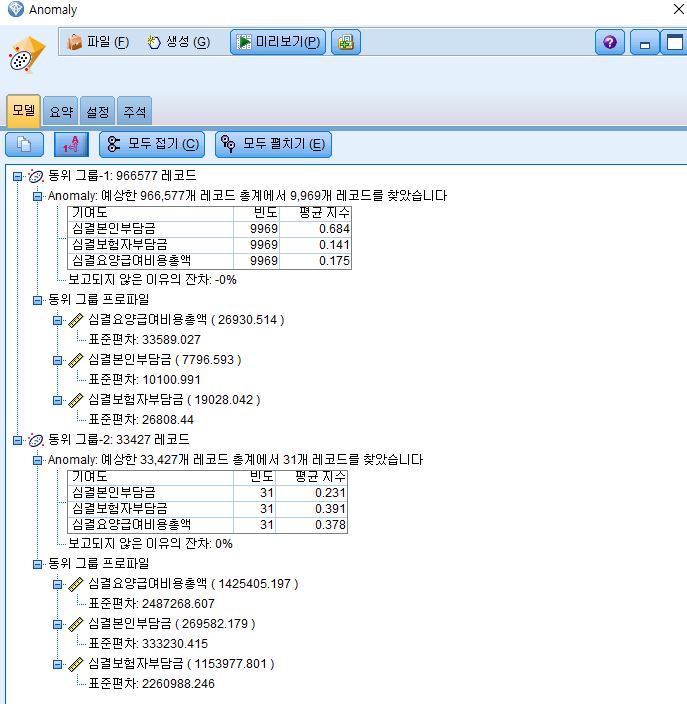
2. 의사결정나무 분석

1) 목표 : 심결요양급여비용총액(심결본인부담금, 심결보험자부담금)을 범주화하여 목표변수(Targeted Variable)로 설정하고, 성별, 나이, 시도, 진료과목(주상별, 부상별), 입내원일수 등을 의사결정 규칙 형성에 대한 변수로 한다. 결과를 통해 진료비 산정에 영향을 주는 요인들을 분석한다.

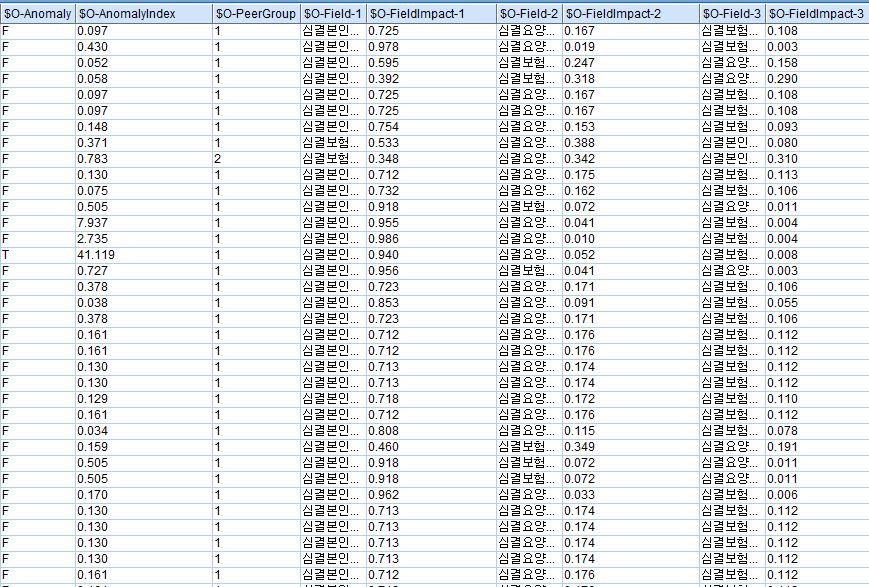
2) 심결요양급여비용총액(심결본인부담금, 심결보험자부담금)의 범주화



백 만개의 표본에 대한 데이터 검토의 결과이며, 데이터 분포가 과도하게 치우친 양상을 보인다. 따라서 노이즈를 제거하고 전체 데이터를 저가형-보통형-고가형으로 나누는 작업이 필요하다.

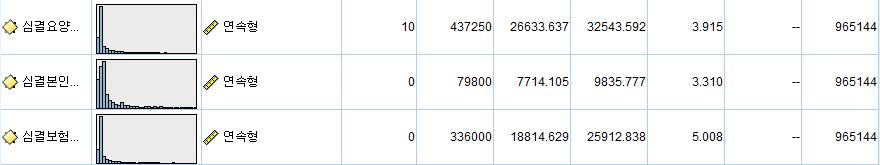


심결요양급여비용총액, 심결본인부담금, 심결보험자부담금에 대한 이상 분석을 진행한다.

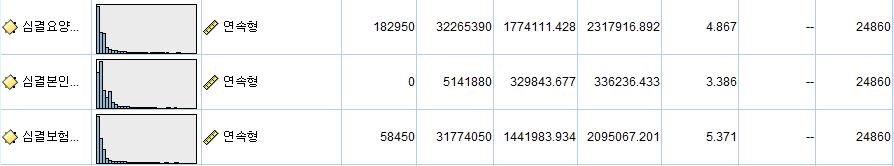




모델링 결과 총 9개의 attribute가 생성되었다. $O-PeerGroup은 1 혹은 2로 구성되는 동위 그룹을 나타내고, $O-Anomaly는 T(true) 혹은 F(false)로 True인 row는 노이즈이다. 이 결과를 바탕으로 $O-Anomaly가 true인 10,000개(1%)의 노이즈를 제거한다.



<그룹1($O-PeerGroup)의 데이터 검토 결과>



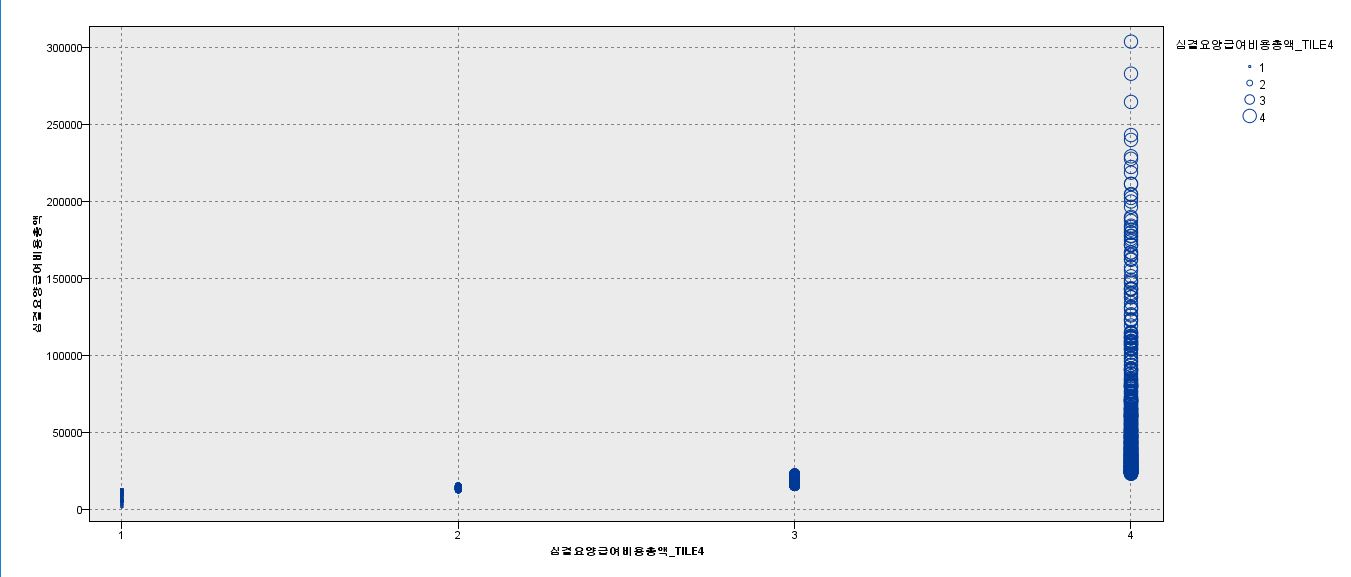
<그룹2($O-PeerGroup)의 데이터 검토 결과>

다음 결과로 보아 ‘저가형’ 진료가 ‘고가형’ 진료보다 월등히 많다. (약38.8배)

따라서 ‘저가형’, ‘고가형’의 분석을 각각 진행하며, 각 유형의 빈도 분포 역시 치우쳐있으므로 세부적인 조정이 필요하다.

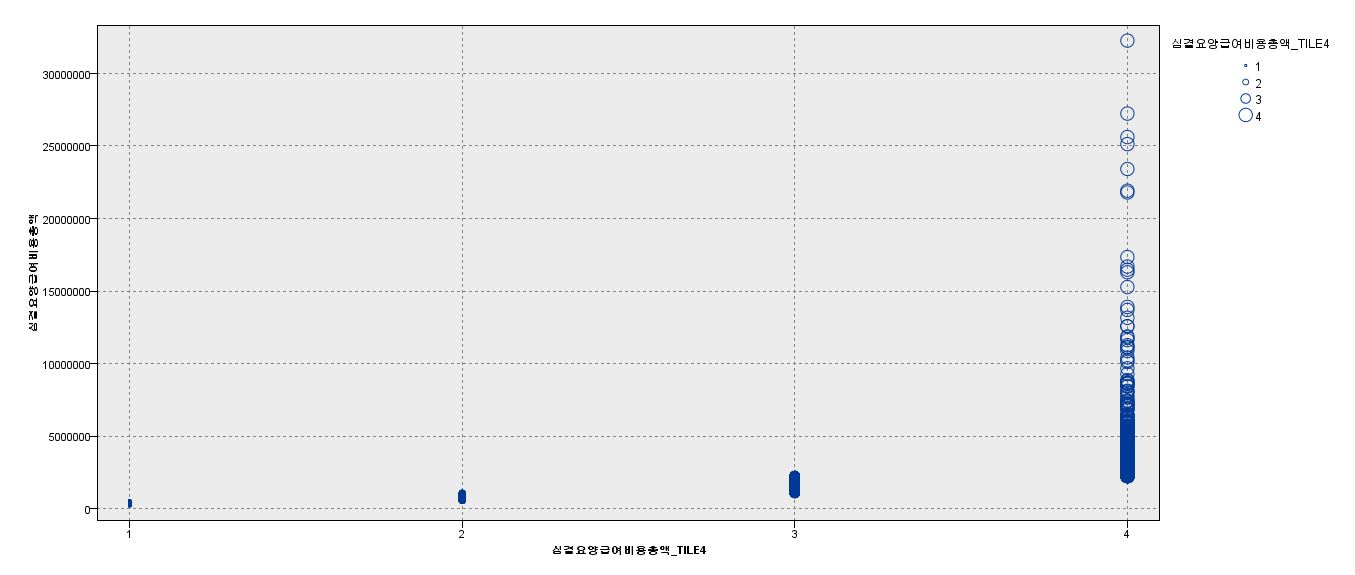
심결요양급여비용총액은 심결본인부담금+심결보험자부담금이므로 앞으로 심결요양급여비용총액기준으로 분석한다.

그룹1(저가형), 그룹2(고가형)을 4분위수로 그룹화 한 뒤, 각 분위를 x축, 심결요양급여비용총액을 y축으로 한 bubble 차트를 생성하면 다음과 같다.



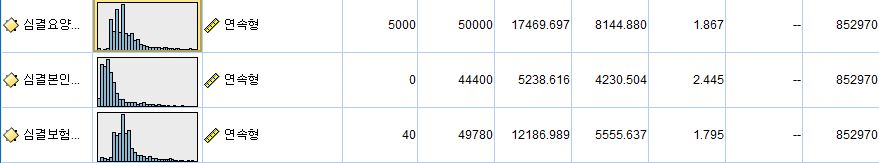
<그룹1의 bubble차트>

75%이상의 표본이 속하는 5,000~50,000원에 대한 분석을 ‘저가형’ 분석으로 한다.

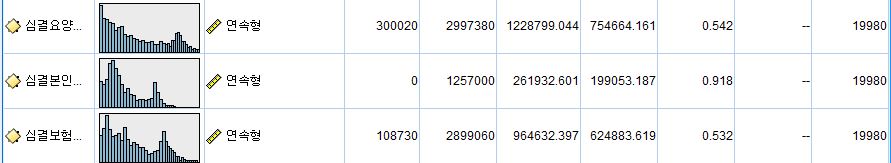


<그룹2의 bubble차트>

75%이상의 표본이 속하는 300,000~3,000,000원에 대한 분석을 ‘고가형’ 분석으로 한다.

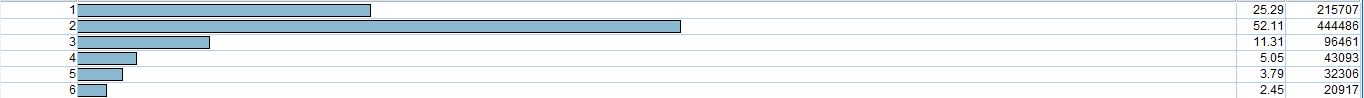


<저가형, 데이터 검토 결과>

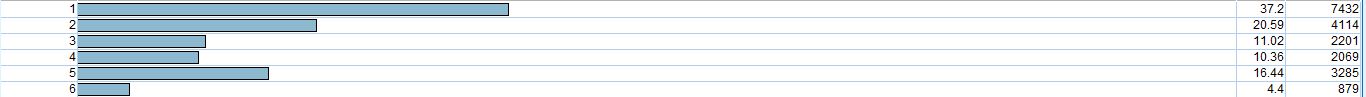


<고가형, 데이터 검토 결과>

저가형, 고가형은 각각 6개의 범주로 구간화한다. 심결요양급여비용총액\_BIN attribute가 생성되고 그 구성은 (1-하하, 2-하상, 3-중하, 4-중상, 5-상하, 6-상상)로 한다.



<저가형, 구간화 분포>



<고가형, 구간화 분포>

3) 의사결정나무 분석

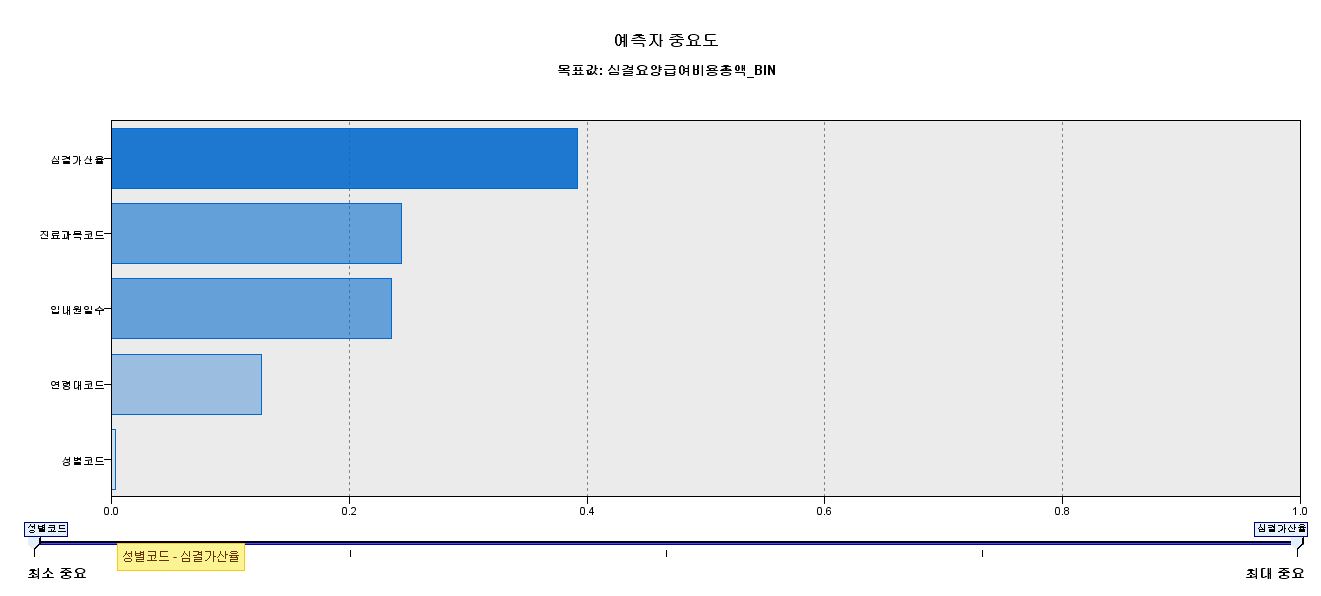
\*70%의 training set과 30%의 test set으로 진행한다.

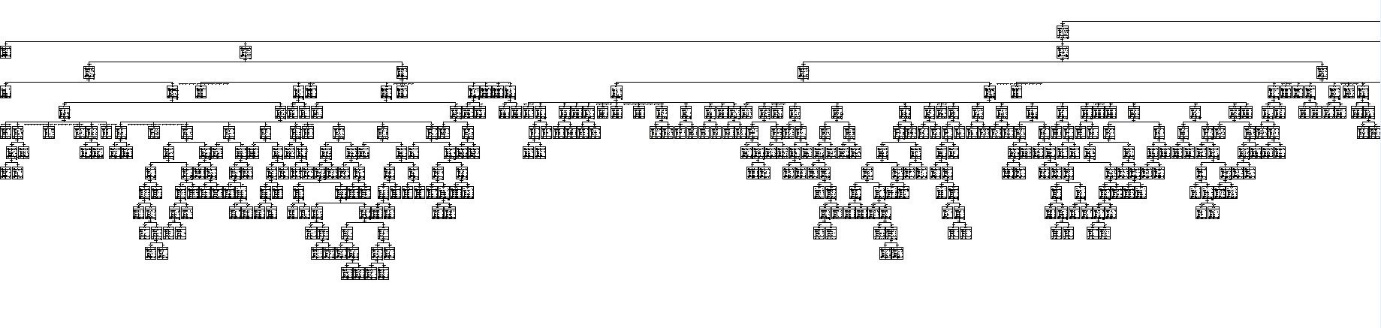
\*목표필드는 "심결요양급여비용총액\_BIN"

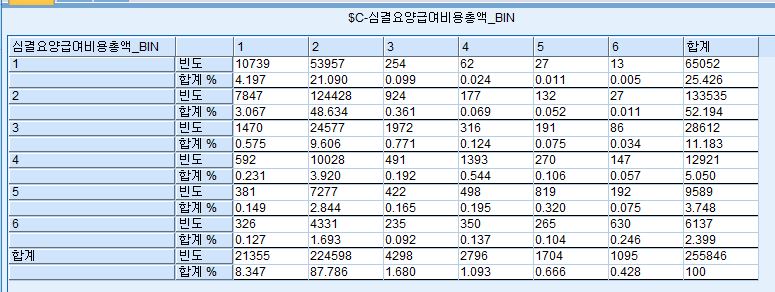
입력필드는 "성별코드", "연령대코드", "진료과목코드", "입내원일수", "심결가산율"이다.

\* CART /C5.0 / CHAID 세 가지 알고리즘을 ‘저가형’, ‘고가형’을 구분하여 적용한다.

\* CART 알고리즘을 ‘저가형’에 적용한 결과를 대표로하여 아래에 첨부한다.

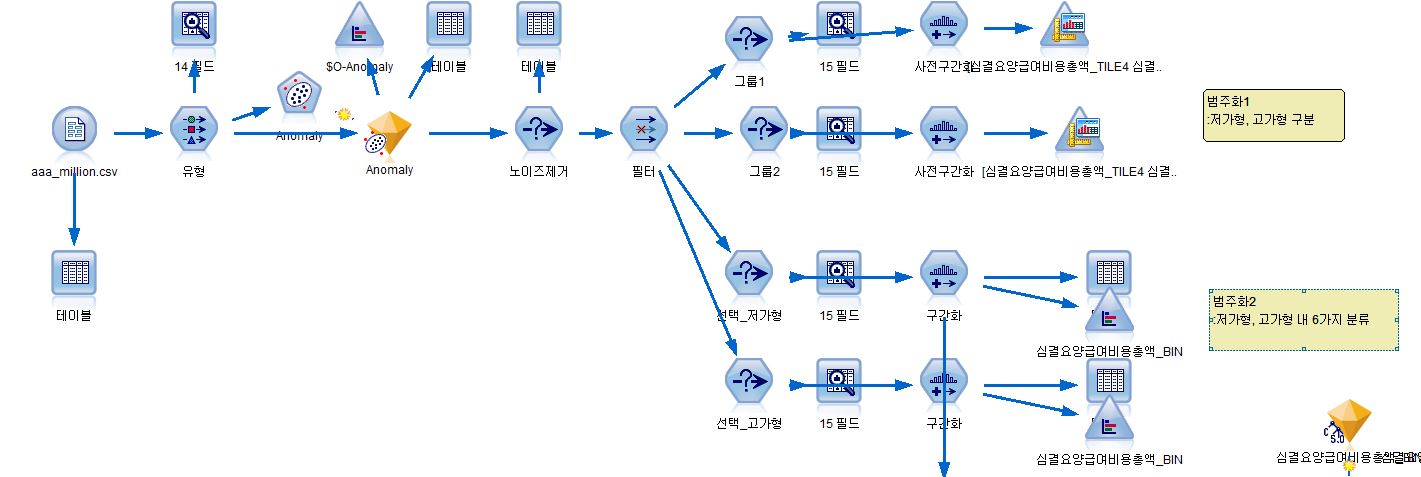


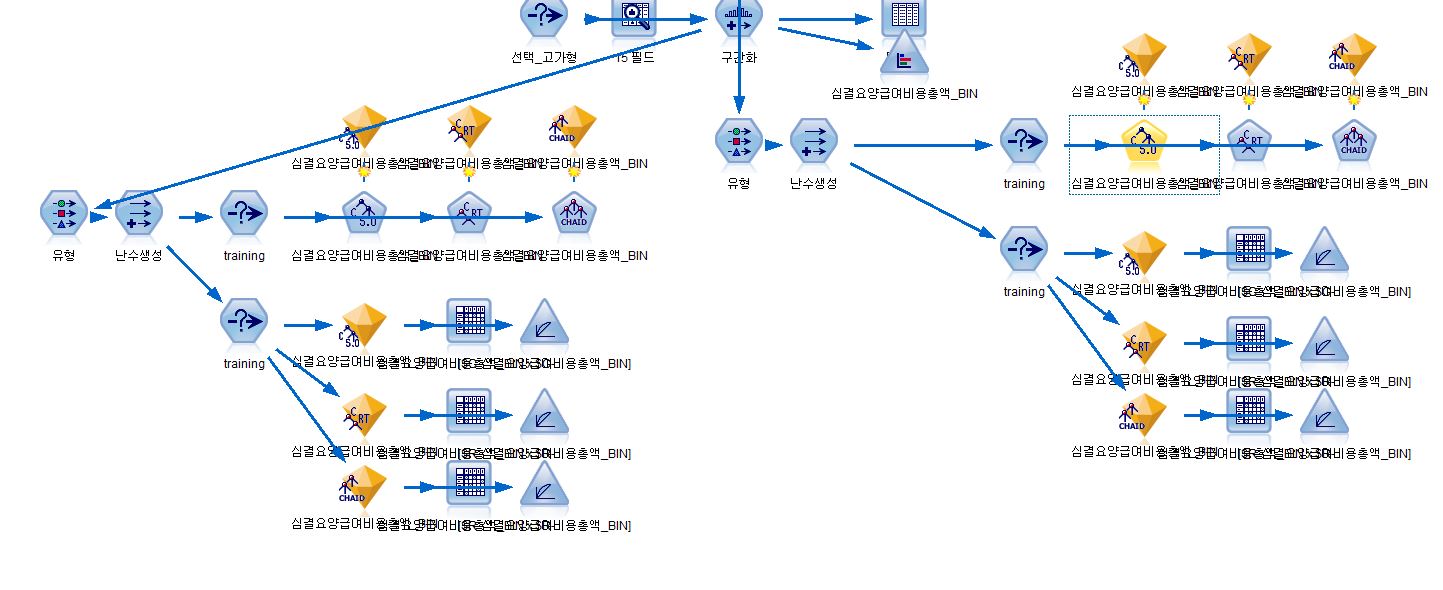




Accuracy는 (10739+124428+1972+1393+819+630) / 255846 \*100 = **54.71**%이다.

4)전체 스트림 구조





5) 개선할 점

a. 모델의 accuracy 높이기 및 의사결정나무 가지치기

-예상원인 ) 동일한 환자의 진료내역 데이터의 산재, 합치는 작업 필요, 보다 유연한 조건 부여 필요

b. 심결요양급여비용총액과 하위 금액과의 관계성 밝히기

c. 심결요양급여비용총액의 범주화에 미포함 된 구간 다루는 법 고안

3. 군집 분석(clustering)

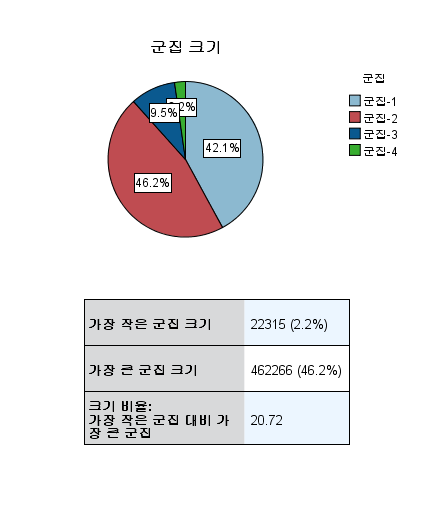
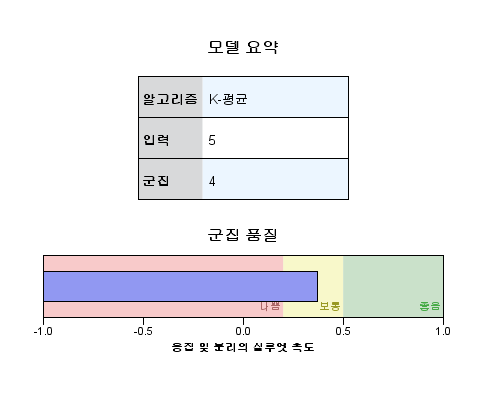
1) 목표 : 모든 변수들을 요인화하여 군집화시키는 것이 아닌 앞의 데이터 분석을 통한 목표 변수들(Targeted Variables)을 설정 하였다. 성별 같은 이분형 변수부터 심결요양급여비용 같은 연속형 변수 그리고 연령대, 시도, 서식, 진료과목 범주 변수까지 총 6가지 변수들을 통하여 요소들 간의 유사성을 관찰하고 이를 보험비 산출에 적용을 목표로 한다.

2) K-means, Twostep, kohonen 3가지 클러스터링 방법의 비교를 통해 가장 적합한 방법을 선택하고 이를 적용할 것이다.

**3) K-means method**

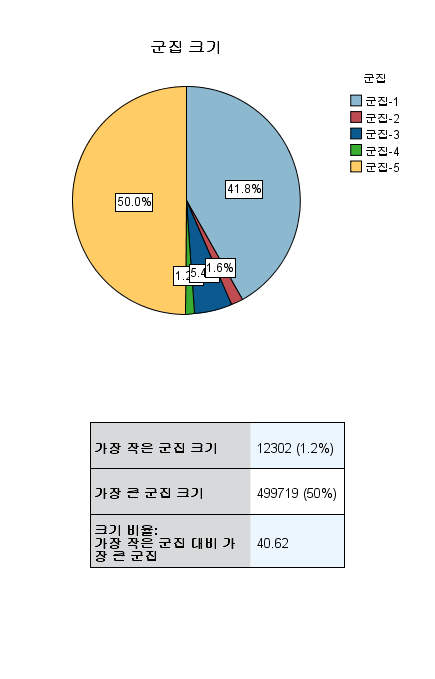
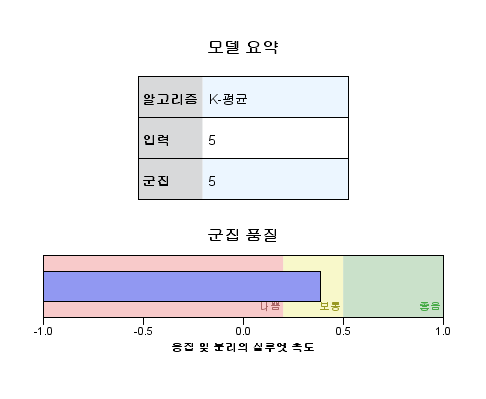
K-means 방법은 직접 군집수를 적용할 수 있다. 따라서 4~6가지의 군집수를 적용해보았고 이중 가장 높은 품질을 가진 군집수를 선택하였다.

K-means -method (군집 수 : 4개)



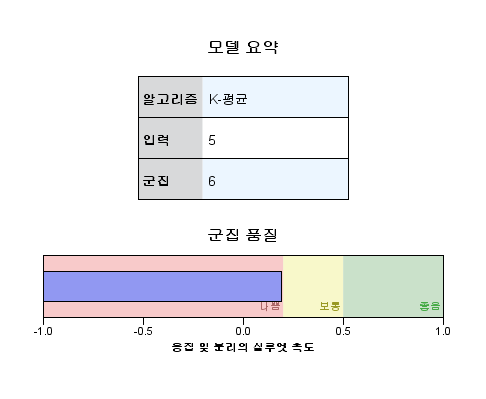
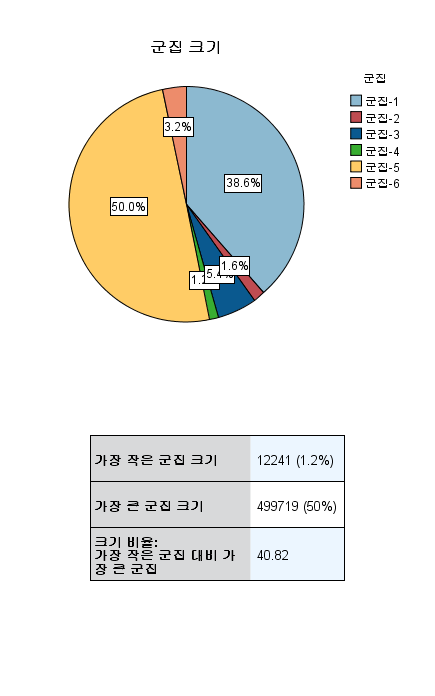
군집 품질이 보통으로 나쁘지 않게 나왔다.

k-means -method (군집 수 : 5개)



군집 품질이 군집 수 4개와 비교하여 같은 보통으로 나왔다.

k-means -method (군집 수 : 6개)

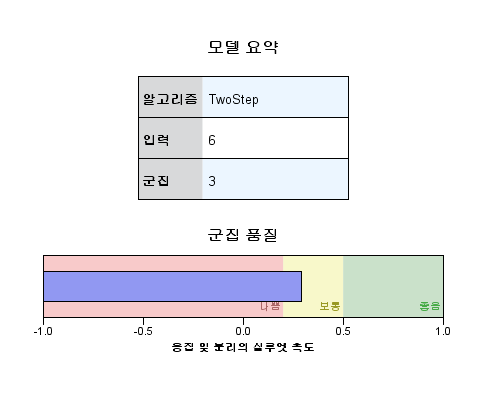
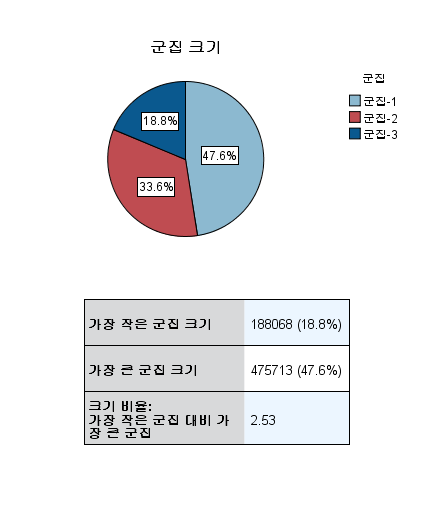
군집 수는 1개 늘어났지만 군집 품질은 오히려 낮아졌음을 확인할 수 있다. 따라서 이를 채택할 수는 없다.

위의 결과를 통해 군집 품질을 5개로 하는게 가장 좋은 결과임을 확인 할 수 있다 군집수를 5개 선택한후 교차표를 통해 군집의 분포를 확인해 보았다.

**4) Two-step method**

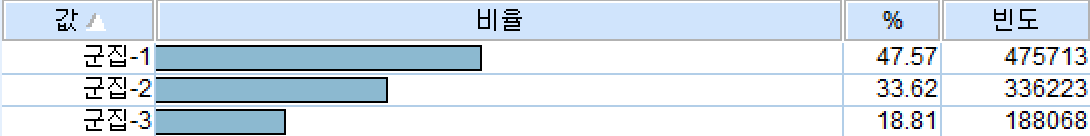
- 개체 수가 큰 자료의 군집화에 매우 효율적인 Two-step method를 사용해 보았다.

- 거리 측도 : 로그 우도, 군집 기준 : 슈바르츠 Bayesian(BIC)을 기준으로 하였다.

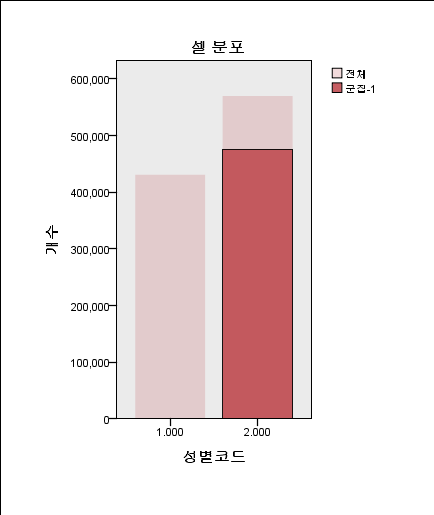
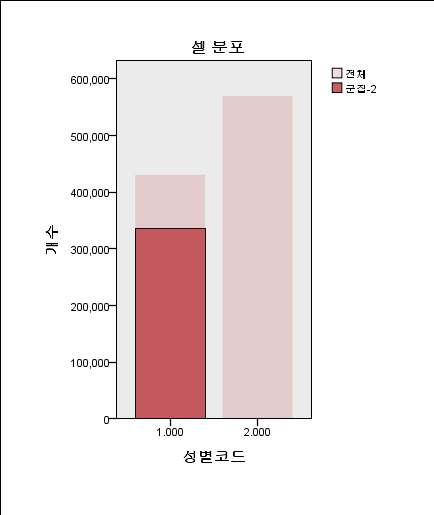
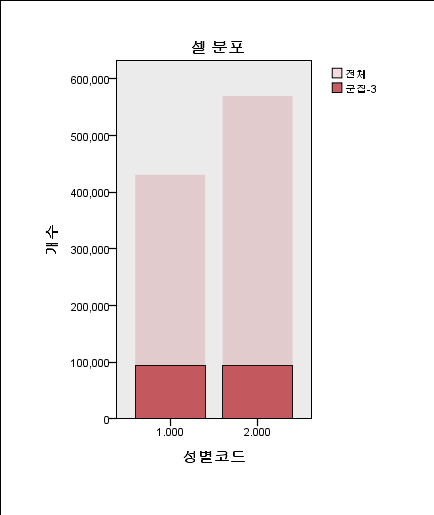
 

그 결과 적정한 군집수는 3개고 군집 품질은 보통으로 선정되었다. 거리 측도를 유클리디안으로 설정하고 분석한 결과 역시 큰 차이는 발견되지 않고 군집 품질 : 보통, 군집크기는 3이 설정되어 분석 되었다. 마찬가지로 군집 기준을 아케이케 정보 기준(AIC)으로 각각 다르게 설정하여도 항상 같은 품질, 같은 군집수의 결과값이 도출되었다.

이를 도표로 나타내보면 다음과 같다.



군집 1의 비율이 거의 50%에 육박함을 알 수 있었다. 또 다른 군집의 특성을 보기 위하여 셀의 성별 분포를 살펴 보았다.

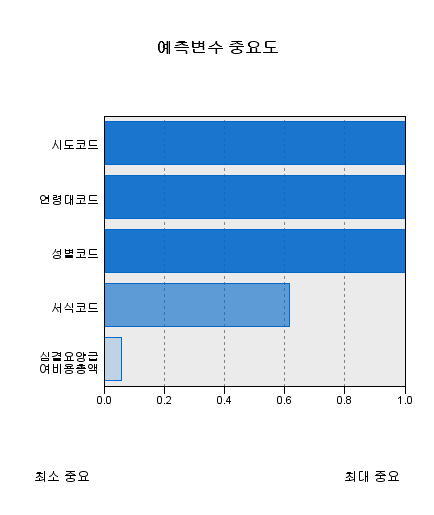
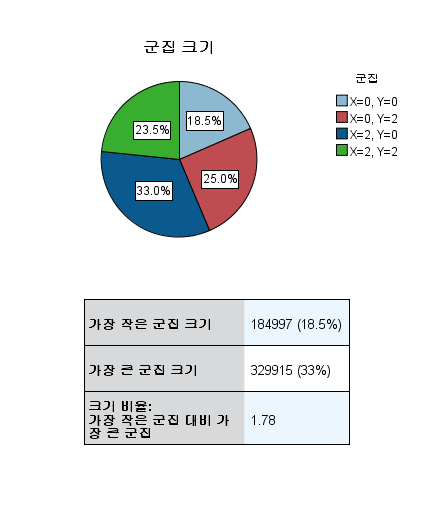
  

각각 군집 1, 군집 2, 군집 3의 분포인데 이를 보면 군집 1= 여자, 군집 2 = 남자, 군집 3= 남자,여자 같은 비율로 분포가 되어 있음을 볼 수 있다.

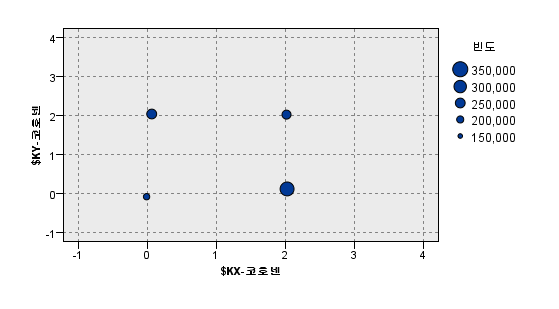
**5) Cohonen-som method**

타 알고리즘에 비해 연산시간이 많이 소요되는 코호넨 알고리즘을 사용해 보았다. 군집 수 결정이 어렵지만 시각화(visualization)가 가능하고 군집간 관계를 파악하는게 쉽다.

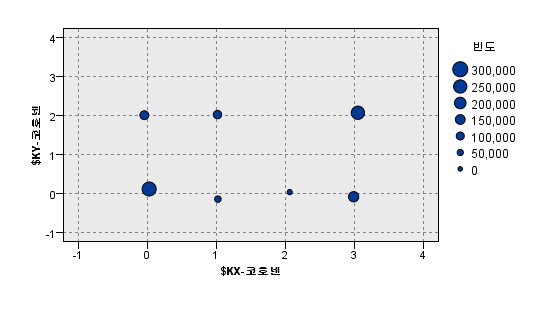
크기 : x=3, y=3 학습결과



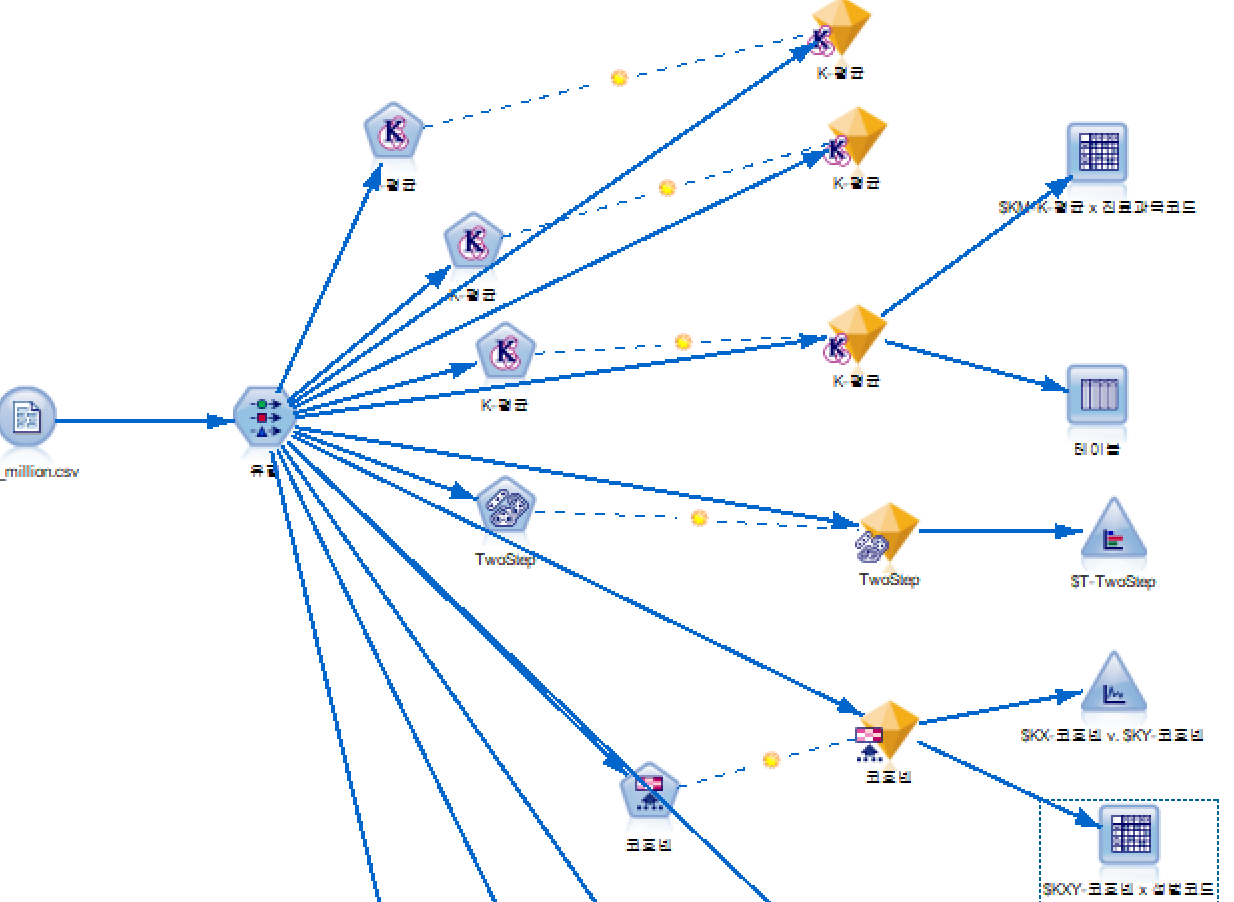
아래는 도표를 이용하여 결과치를 시각화한 그래프이다.



비슷한 방법으로 4 by 3 으로 학습한 그래프이다.



클러스팅 스트림 구조 (부분)

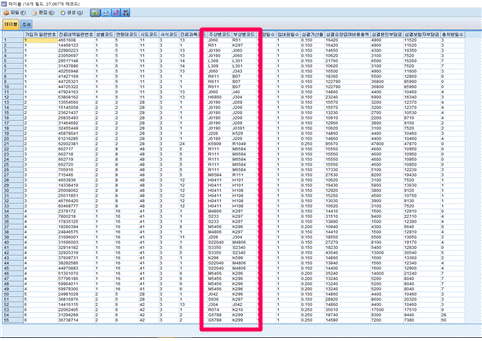


6) 개선할 점

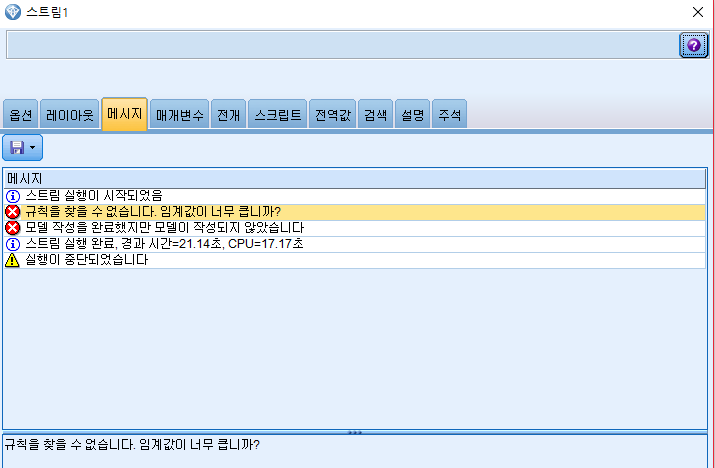
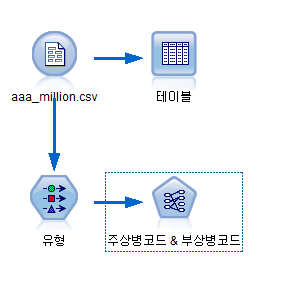
아직 군집화에 따른 각각 군집의 특성을 다 파악하지 못했다. 추후 이 군집의 특성을 보험비와 어떻게 연관시킬 것인지 생각해보아야 할 것 같다. 뿐만 아니라 군집안의 요소 쏠림 현상이 성별 등에서 발생하고 있는데 이를 처리하는 방법에 대해 생각해보아야 할 것 같다.

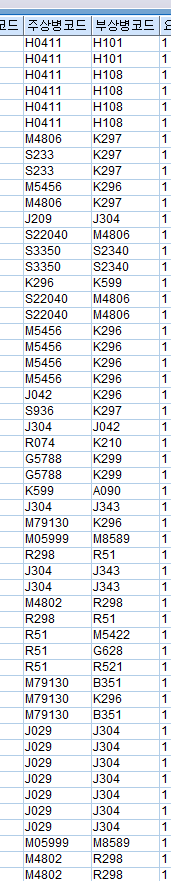
4. 연관성 분석

1) 주상병 코드와 부상병 코드의 연관성 분석으로 통하여 공통적으로 나타나는 유사 질병을 파악하고자 한다.

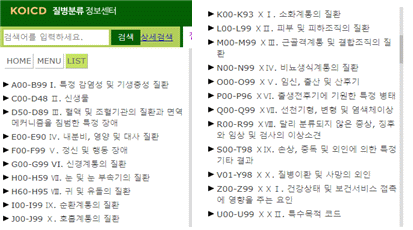


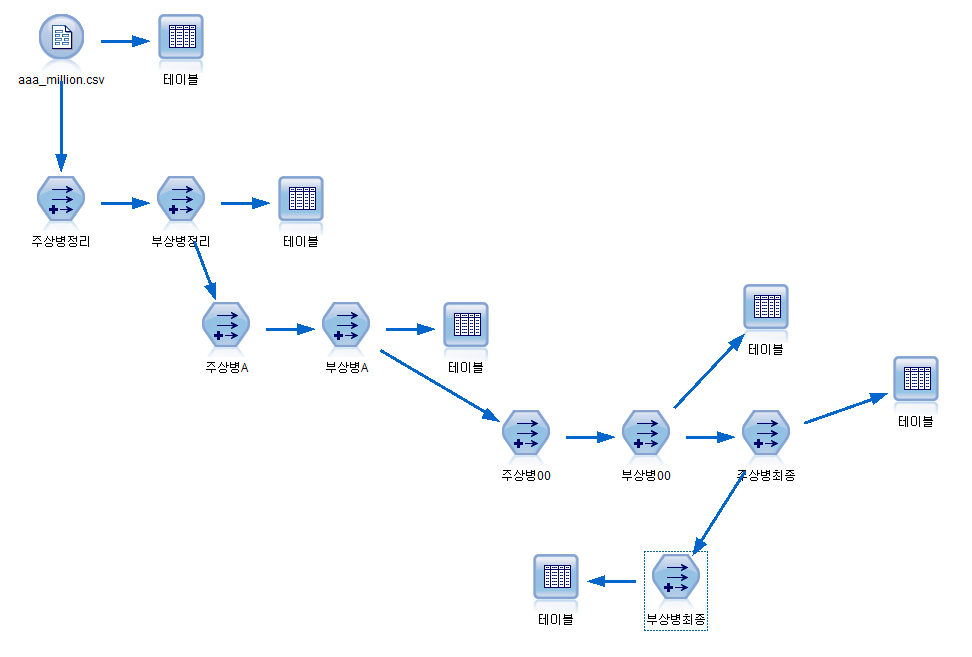
2) 연관성 분석에는 Apriori를 사용하였고, 주상병 코드와 부상병 코드를 변수로 지정하였다. 기본 값으로 지지도 10, 신뢰도 75, 전항값 수 5를 조건으로 찾아 보았을 때, 연관성을 찾을 수 없었다. 따라서 지지도와 신뢰도를 낮춰가며 연관성을 찾아보려 하였다. 결과적으로는 성공적인 결과를 얻지 못하였다. 어느 정도의 결과를 얻기 위해서는 지지도와 신뢰도를 너무 낮춰야 했고, 유의미한 자료를 얻기에는 신뢰도가 너무 까다로웠다.



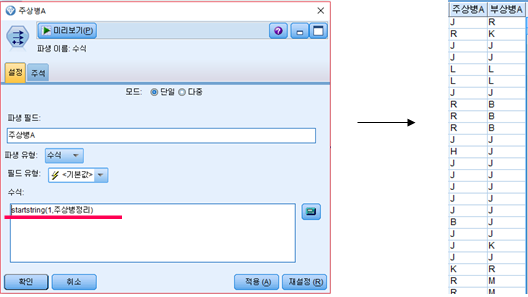


분석해본 이유에 따르면, 주상병 코드와 부상병 코드는 각각 알파벳 1개와 숫자 3~4개의 조합으로 이루어져 있다. 각각 알파벳은 질병 기관별로 분류 한 것이고, 숫자는 구체적인 질병 명에 따라 부여된 번호이다. 알파벳, 즉 질병이 발생한 기관이 비슷하더라도 질병 명이 다르다면 연관성 분석이 이뤄지지 않는 것이다. 따라서 다시 한 번 분류하여 연관성 분석을 실시했다.

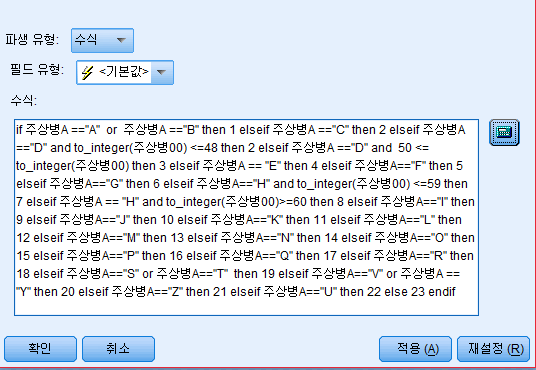
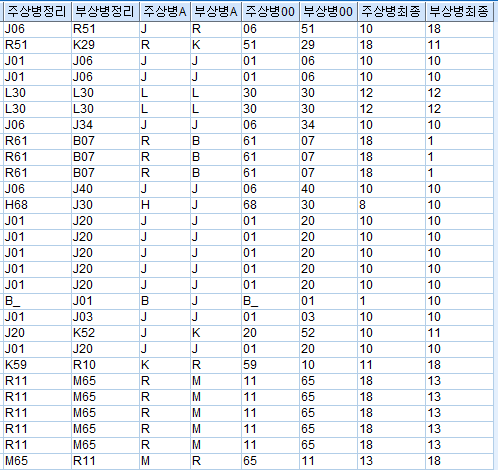
3) 한국질병분류 정보센터에는 다음과 같이 질병 코드를 정리한다.

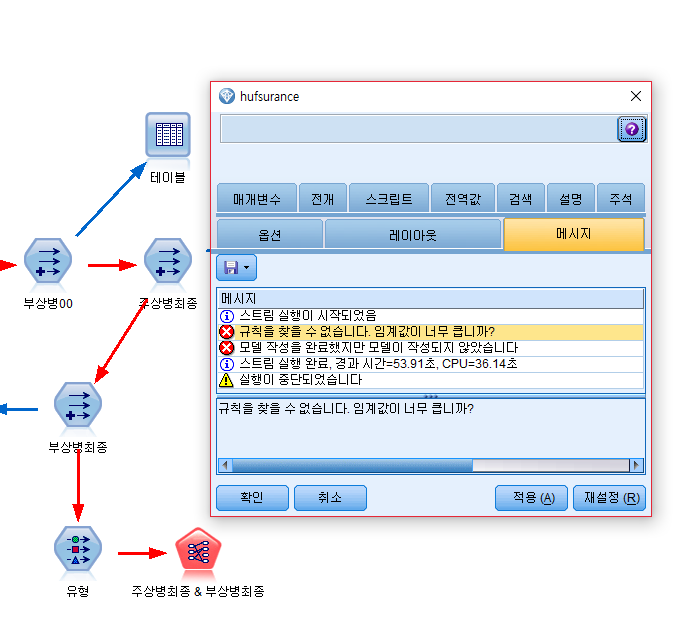
주ㆍ부상병 코드를 위와 같이 정리하기 위해서 다음과 같은 과정을 걸쳤다.

먼저 주상병 정리와 부상병 정리의 파생 노드를 통하여 각각의 코드를 3자리로 정리하였다.

주상병 A와 부상병 A의 파생 노드를 통하여 첫 글자, 즉 알파벳으로 정리하였다.

이후 주상병00과 부상병00 의 파생노드에서 endstring(2,-)를 통하여 다음과 같은 데이터를 얻었다.

4) 이후 파생 노드를 통하여 주상병A와 주상병00 필드를 이용한 수식을 작성, **3)**에서 언급한 바에 따라 각각 1~22의 번호를 부여하였다. 부상병도 마찬가지로 진행하였다.

5) 다시 정리한 필드를 바탕으로 연관성 분석을 진행하였다,

하지만 연관성을 찾기에 힘들었다. 역시나 신뢰도를 높이면 모델링이 나오지 않았고 모델링이 작성된다면 신뢰도가 지나치게 낮았다.

6) 개선점

지금까지의 방식으로는 주상병 코드와 부상병 코드 사이의 연관성을 찾기 어려웠다. 다른 방식 또는 다른 항목으로 그들 사이의 연관성을 파악하기 위해 정보를 좀 더 파악하는 것이 필요할 것이다.