

## DFMBA 경영통계 - 2<sup>nd</sup> Assignment (중간고사 대체시험)

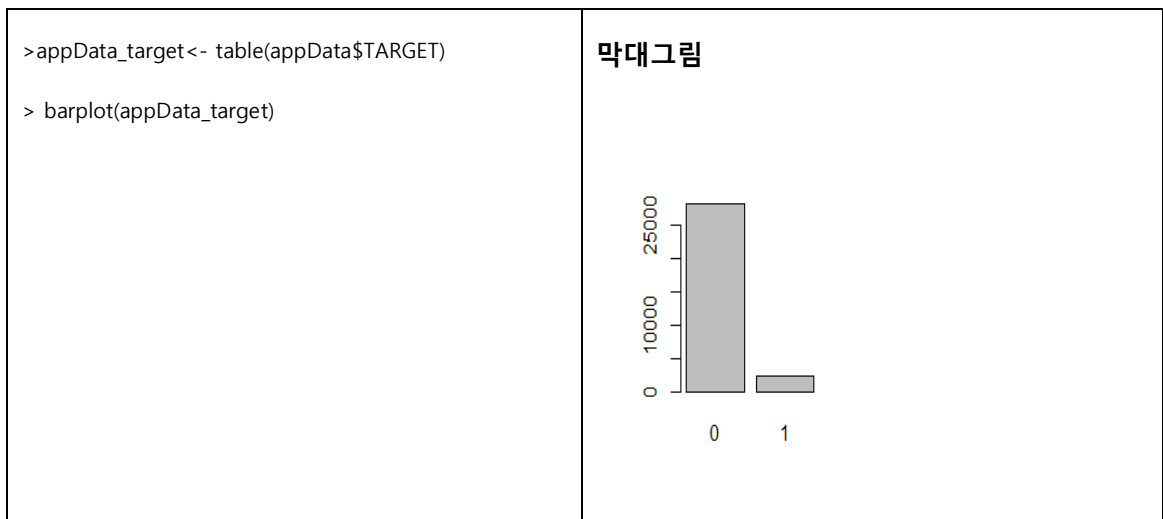
학번 : 20224071

이름 : 오택건

※ 'app.csv' 파일은 어느 금융회사의 대출고객에 대한 정보에 관한 것이다. 첫 컬럼인 IDX는 각 고객에 대한 index이고, CODE\_GENDER는 성별, FLAG\_OWN\_CAR와 FLAG\_OWN\_REALTY는 차 또는 부동산의 소유 여부, AMT\_INCOME\_TOTAL은 소득, AMT\_ANNUITY는 월 대출 지급액, AMT\_CREDIT은 대출금액, DAYS\_BIRTH는 출생시점 (데이터 수집일을 0으로 보았을 때 출생일이 며칠 전인지로 기록), NAME\_EDUCATION\_TYPE은 교육수준, OCCUPATION\_TYPE은 직업군, TARGET은 해당 고객이 default인지 여부(0 : non-default, 1: default )를 나타낸다. 아래 1~7번은 이 자료에 관한 것으로 모두 R을 활용하는 문제이다. 물음에 답하여라. (각 10점)

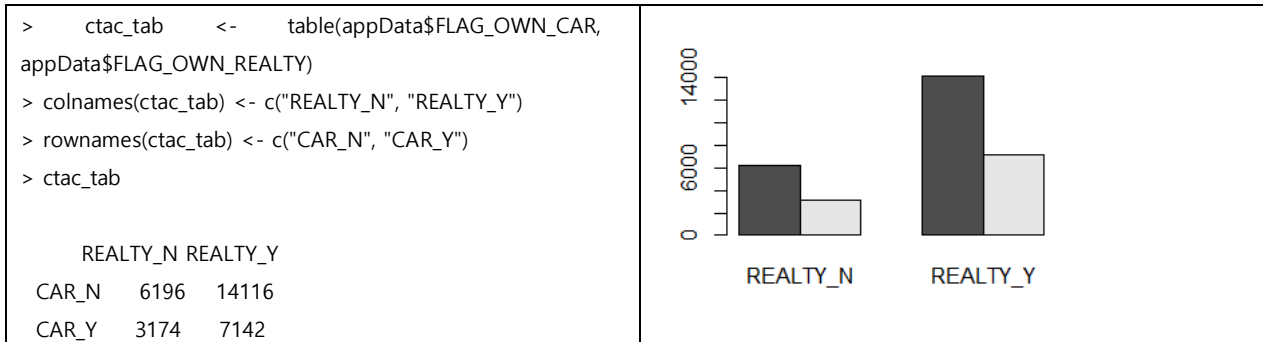
1. 고객의 default 여부를 나머지 변수 정보를 이용하여 예측하고자 한다. 범주를 예측하는 문제에서 목표 변수의 범주 간 비중 차이가 큰 자료를 불균형(imbalanced) 자료라고 하는데, 이 자료가 불균형 자료에 해당할지를 적절한 그래프를 이용하여 파악하여라.

→ 아래 막대그림에 범주간 비중 차이가 큰 것을 보아 불균형 자료임을 알 수 있습니다



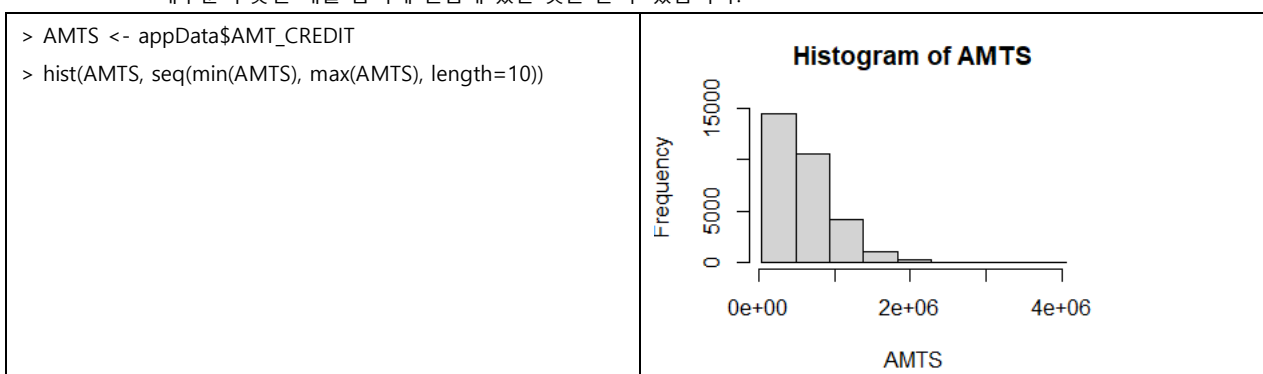
2. 주어진 자료에서 각 고객이 차를 소유했는지 여부와 부동산을 소유했는지 여부 간의 연관성을 파악하고자 한다. 적절한 그래프를 이용하여 표현하고 이를 해석하여라.

→ 아래 두 변수에 관한 요약과 스택트컬럼차트에서 특별한 상관관계를 발견할 수 없습니다



3. 전체 고객의 대출금액에 대한 분포를 적절한 그래프를 이용하여 표현하고, 대칭/치우침 여부와 봉우리의 개수 등을 해석하여라.

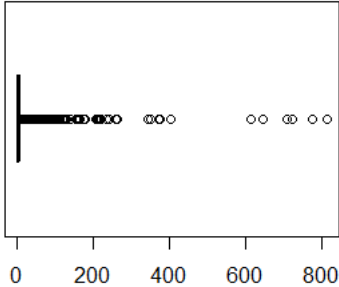
→ 히스토그램에서 오른쪽으로 치우침과 봉우리 개수 1개를 확인할 수 있습니다  
대부분이 낮은 대출 금액에 밀집해 있는 것을 알 수 있습니다.



4. Non-default인 고객과 default인 고객 간에 대출금액에 어떠한 차이가 있는지를 확인하고자 한다.

- (1) 적절한 그래프를 이용하여 non-default 고객과 default 고객의 각 그룹 별 대출금액의 분포를 표현하고, 두 그룹의 중심값, 변동성, 이상치 존재 등의 차이를 해석하여라.

→ 아래 상자수염그림에서 상대적으로 non-default 고객은 중심값이 있고, 변동성과 이상치가 적으며 default 고객은 중심값이 없고, 변동성과 이상치가 큰 것을 확인할 수 있습니다

<pre> &gt; AMTS &lt;- table( appData\$AMT_CREDIT, appData\$TARGET)  &gt; head(AMTS)        0  1 45000 25  3 47970 23  1 49500  6  0 49752  6  0 50940 34  1 52128  8  0  &gt; boxplot(AMTS)  &gt; boxplot(AMTS, horizontal=TRUE) </pre>	
---	--

(2) non-default 고객과 default 고객의 각 그룹 별로 표본평균, 표본표준편차, 20퍼센타일과 80퍼센타일, 사분위간 범위를 계산하여라.

	표본 평균	표본표준편차	20퍼센타일	80퍼센타일	사분위간 범위
non-default <pre> &gt; nondefault &lt;- appData[appData\$TARGET==0, ][' AMT_CREDIT'] &gt; summary(nondefault) AMT_CREDIT Min.   : 45000 1st Qu.: 270000 Median : 517500 Mean    : 601224 3rd Qu.: 810000 Max.    :4050000 </pre>	601224	<pre> &gt; nondefault_tmp &lt;- as.numeric(unlist(non default)) &gt; sd(nondefault_tmp) [1] 407256.1 </pre>	<pre> &gt; quantile(nondefa ult_tmp, 0.20) 20% 252000 </pre>	<pre> &gt; quantile(nondefa ult_tmp, 0.80) 80% 9e+05 </pre>	<pre> &gt; IQR(nondefaul t_tmp) [1] 540000 </pre>
default <pre> &gt; default &lt;- appData[appData\$TARGET==1, ][' AMT_CREDIT'] &gt; summary(default) </pre>	543164	<pre> &gt; default_tmp &lt;- as.numeric(unlist(def ault)) &gt; sd(default_tmp + ) </pre>	<pre> &gt; quantile(default_t mp, 0.20) 20% 254700 </pre>	<pre> &gt; quantile(default_t mp, 0.80) 80% 790830 </pre>	<pre> &gt; IQR(default_t mp) [1] 402719.6 </pre>

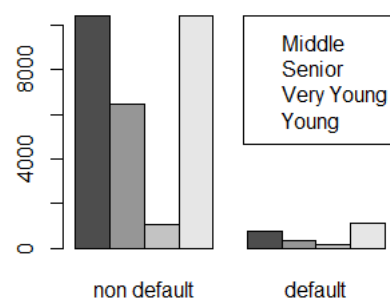
AMT_CREDIT Min. : 45000 1st Qu.: 276278 Median : 473760 Mean : 543164 3rd Qu.: 678997 Max. :2695500		[1] 344497.1			
---	--	--------------	--	--	--

5. DAYS\_BIRTH 값을 365로 나눈 뒤 절대값을 취하여 각 고객의 나이를 구한 뒤, 나이가 0 초과 25 이하인 경우는 "Very Young", 25 초과 40 이하인 경우는 "Young", 40 초과 55 이하인 경우는 "Middle", 55 초과 100 이하인 경우는 "Senior"의 범주로 표현되는 agegrp이라는 변수를 생성하여라. (원 데이터프레임에 agegrp을 새로운 변수로 포함해도 되고 하지 않아도 됨)

```
> appData['DAYS_BIRTH'] <- abs(appData$DAYS_BIRTH)/365
> appData <- transform(appData, agegrp =
+ ifelse(DAYS_BIRTH <= 25, "Very Young",
+ ifelse(DAYS_BIRTH <= 40, "YOUNG",
+ ifelse(DAYS_BIRTH <= 55, "Middle",
+ ifelse(DAYS_BIRTH <= 100, "Senior"))))
+ )
> head(appData$agegrp)
[1] "YOUNG" "YOUNG" "YOUNG" "Middle" "Senior" "Middle"
```

6. 고객의 연령대 (5번에서 만든 agegrp 변수)와 고객의 default 여부가 연관성이 있는지 확인하고자 한다. agegrp 별로 non-default나 default인 고객의 수를 확인할 수 있도록 적절한 그래프로 표현하여라.

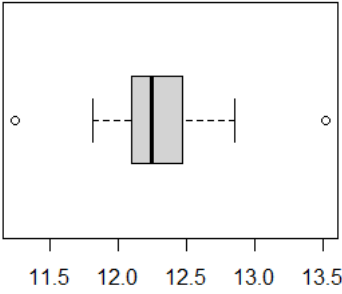
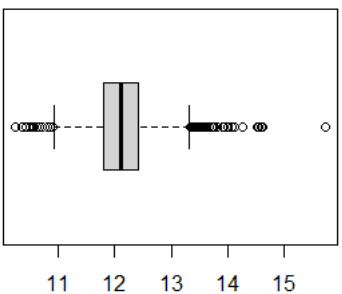
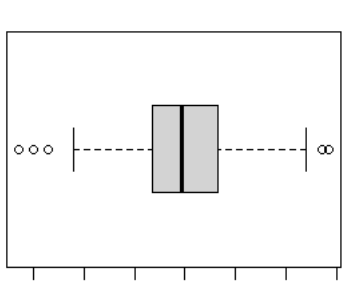
```
> ctac_tab <- table(appData$agegrp, appData$TARGET)
> head(ctac_tab)
      non default default
Middle      10383      770
Senior       6426      360
Very Young   1042      155
YOUNG        10383     1109
> barplot(ctac_tab, beside=TRUE)
> legend("topright", legend=c("Middle", "Senior", "Very
Young", "Young"))
```

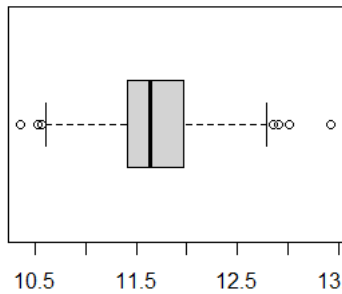
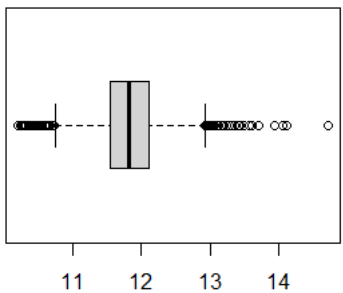


7. 교육수준은 다음의 다섯 범주로 표현되어 있다. 각 교육 수준 별로 소득에 차이가 있는지를 확인하고자 한다. 소득 자료는 양의 왜도가 심하므로, 로그 소득 자료(소득에 log 변환)의 분포를 각 교육수준 그룹별로 도출하여라. 또 이를 이용하여 교육 수준에 따라 로그소득의 중심값, 변동성, 이상치 존재여부 등에서 어떤 차이가 있는지를 해석하여라.

- academic degree : A University or professional institute
- higher education : Community or junior colleges (non-university)
- incomplete higher
- lower secondary : 7th-9th grades
- secondary/ secondary special : 10th-12th grades or vocational education

<pre> &gt; appData['AMT_INCOME_TOTAL'] &lt;- log(appData['AMT_INCOME_TOTAL']) &gt; ad &lt;- appData[appData\$NAME_EDUCATION_TYPE=="Academic degree",]['AMT_INCOME_TOTAL'] &gt; boxplot(as.data.frame(ad), horizontal=TRUE) </pre>	<ul style="list-style-type: none"> <li>&gt; 로그 소득 자료 변환</li> <li>&gt; 교육수준 그룹별 리스트 추출</li> <li>&gt; 상자수염그림 작성</li> </ul>
---	--

<b>academic degree</b> - 상대적으로 변동성과 이상치 데이터가 적음	<b>higher education</b> - 상대적으로 변동성과 이상치 데이터가 많음	<b>incomplete higher</b> - 상대적으로 변동성과 이상치 데이터가 적음
		

<b>lower secondary</b> - 상대적으로 변동성과 이상치 데이터가 적음	<b>secondary/ secondary special</b> - 상대적으로 변동성과 이상치 데이터가 많음
	

8. 다음은 표본 자료의 수치적 특성을 파악하기 위한 기술 통계량들의 특징을 서술한 것이다. 적절한 것은 T, 잘못된 것은 F로 답하여라. (14점)

- (1) 표본 자료에 극단치가 있다면, 자료의 중심 위치를 파악하기 위한 척도로 평균보다 중위수가 더 적절하다. T

- (2) 표본 자료의 산포에 따라 표준편차가 음수가 될 수도 있다. F
- (3) 사분위간 범위는 가운데 50%에 해당하는 표본의 변동성을 나타내는 값이다. T
- (4) 자료값의 단위가 달라져도 변동계수는 바뀌지 않는다. T
- (5) 모든 자료값이 동일한 경우 분산은 0이 된다. T
- (6) 평균이 클수록 표준편차도 클 것이라고 예상할 수 있다. F
- (7) 자료의 분포가 대칭형인 경우 평균과 중위수는 그 값이 비슷하다. T

9. 어느 금융기관에서 하루 중 방문한 고객의 수를 무작위로 400일을 선택한 뒤 기록한 결과, 평균이 125, 분산이 25, 중위수는 140로 나왔다고 한다. 이 자료에 대한 다음의 설명 중 잘못된 것은 무엇인가? 2 (5점)

- ① 하루 방문 고객 수의 표준편차는 5이다.
- ② 하루 방문 고객 수의 분포는 오른 꼬리가 긴 형태일 것이다.
- ③ 하루 방문 고객 수의 변동계수는 0.04이다.
- ④ 전체 표본 자료 중 125보다 큰 값의 비율은 50%를 초과한다.

10. 어느 50개의 표본 자료에서 최대값이 93인데, 실수로 930으로 잘못 기록하였다고 한다.

이렇게 잘못 기록된 상태로 구한 기초통계량과 930을 93으로 다시 정정한 뒤 다시 구한 기초통계량을 비교한다고 할 때, 다음 중 두 경우의 값이 다르지 않은 기초통계량에 해당하는 것은 무엇인가? 3 (5점)

- ① 평균(mean)
- ② 표준편차(standard deviation)
- ③ 3사분위수(Q3)
- ④ 범위(range)

11. 어느  $n$ 개의 표본자료  $x_1, \dots, x_n$ 로 구한 기초통계량 값들에 비해, 각 자료값을 모두 3배한  $3x_1, \dots, 3x_n$ 로 구한 기초통계량 값이 어떻게 달라지는가에 관해 설명한 것이다. 다음 중 잘못된 설명은 무엇인가? 4 (6점)

- ① 중위수(median)는 3배가 된다.
- ② 표준편차(standard deviation)는 3배가 된다.
- ③ 범위(range)는 3배가 된다.
- ④ 변동계수(cv)는 3배가 된다.