|  |
| --- |
| REPORT |
| **과제 9**  K-Means 군집화 |

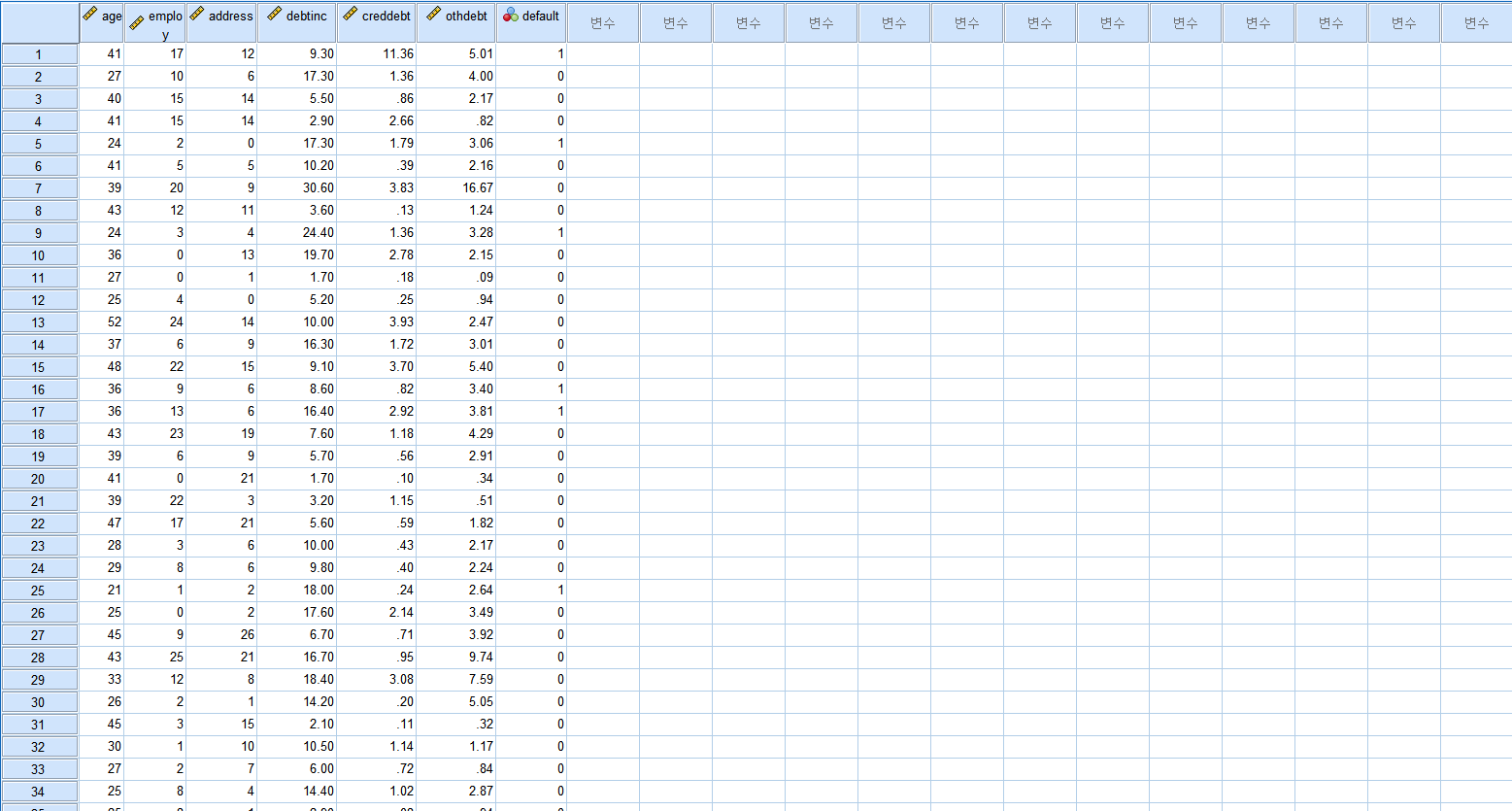




|  |  |  |
| --- | --- | --- |
| 제출일 | 2020. 06. 02 |  |
| 과목명 | 비즈니스애널리틱스개론 |  |
| 담당교수 | 안현철 |  |
| 전공 | 경영정보전공 |  |
| 학번 | 20175288 |  |
| 이름 | 최광삼 |  |

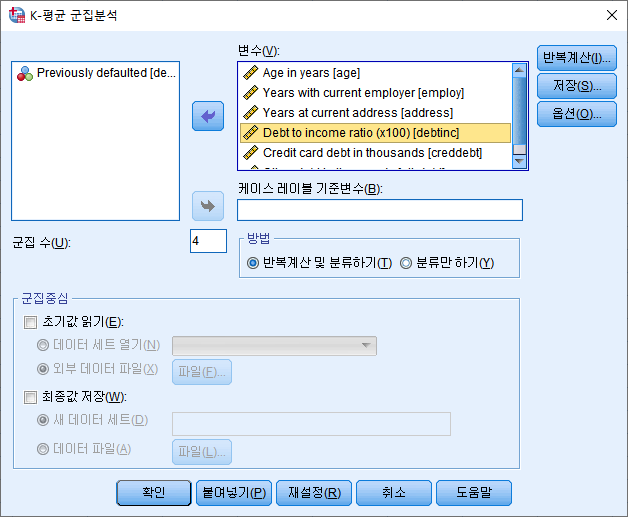
**목차**

1. **bankloan.sav 군집화**
2. K-Means 설정값
3. 군집분석 결과
4. 집단별 특징
5. **Bankloan.sav 군집화**

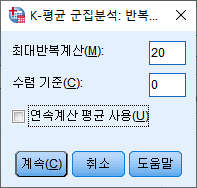
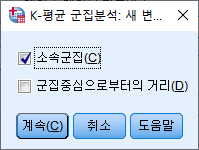


bankloan.sav는 위와 같은 데이터를 가진 SPSS 파일이다. bankloan.sav의 변수로는 7개가 있으며, 여기에서 군집화하는데 유의하게 사용할 수 있는 정량적 변수는 6개로 age는 나이, employ는 현재 직장에서 근무한 연차, address는 현재 주소지에 거주하고 있는 연수, debtinc는 소득 대비 부채비율, creddebt는 신용카드 연체액, othdebt는 기타 체불금액을 의미한다.

* 1. K-Means 설정값



K-Means 군집분석을 위해서 정량적으로 평가될 수 있는 데이터를 가진 6개 변수를 지정하였고, 4개의 집단으로 구분하기 위해 군집 수 설정에 4를 입력하였다.

군집중심점을 구하기 위하여 반복하는 계산 횟수는 최대 20번까지 가능하도록 설정하였고, 수렴 기준에는 변화가 하나도 없을 때까지 군집을 나누기 위해 0의 값으로 설정하였다.

또한 각각의 데이터마다 어느 집단으로 분류되었는지 확인할 수 있도록 소속군집에 대한 정보를 저장할 수 있도록 설정하였다.

나머지 설정은 초기값을 유지하였다.

* 1. 군집분석 결과

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **반복계산과정a** | | | | |
| 반복 | 군집중심의 변화량 | | | |
| 1 | 2 | 3 | 4 |
| 1 | 14.541 | 14.398 | 12.167 | 16.877 |
| 2 | 1.478 | 1.592 | .455 | 4.287 |
| 3 | 1.030 | .962 | .383 | 2.129 |
| 4 | .453 | .883 | .251 | 1.289 |
| 5 | .383 | 1.033 | .251 | 1.290 |
| 6 | .295 | .472 | .222 | .587 |
| 7 | .166 | .152 | .129 | .260 |
| 8 | .126 | .000 | .093 | .311 |
| 9 | .000 | .000 | .000 | .000 |

K-Means 군집분석을 실시한 결과, 해당 데이터는 9번째 계산에서 0의 값이 도출되어 총 9번 반복 계산한 끝에 군집 중심점을 기준으로 온전히 구분된 것을 알 수 있다.

|  |  |  |
| --- | --- | --- |
| **각 군집의 케이스 수** | | |
| 군집 | 1 | 233.000 |
| 2 | 96.000 |
| 3 | 256.000 |
| 4 | 115.000 |
| 유효 | | 700.000 |
| 결측 | | .000 |

총 700건의 데이터 가운데, 집단1에 대항되는 경우는 233건, 집단2에 해당하는 경우는 96건, 집단3에 해당하는 경우는 256건, 집단4에 해당하는 경우는 115건으로 나타났다.

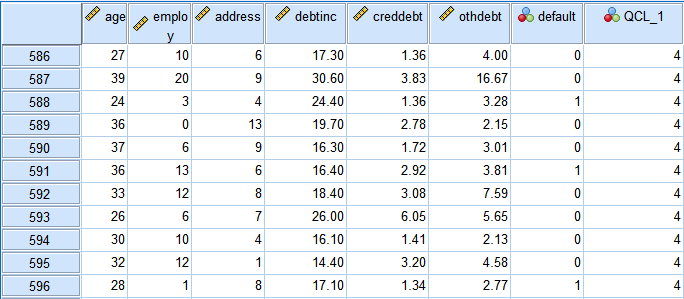
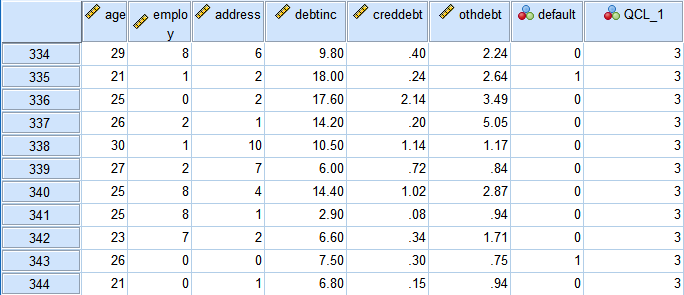
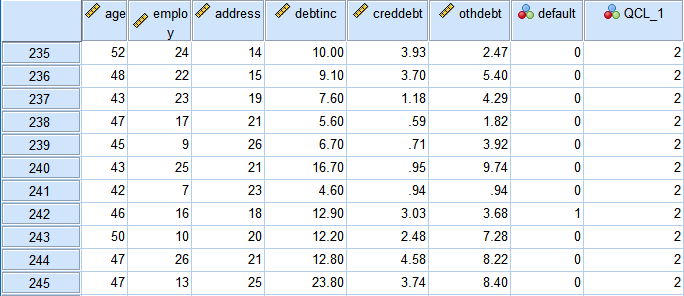
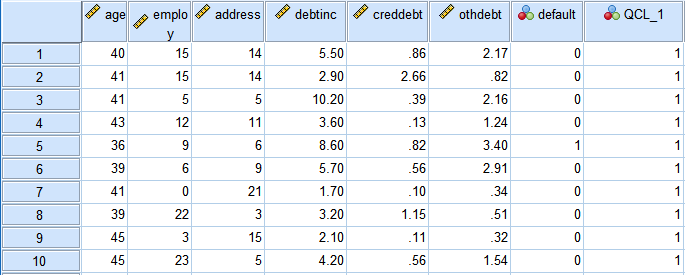
* 1. 집단별 특징

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **최종 군집중심** | | | | |
|  | 군집 | | | |
| 1 | 2 | 3 | 4 |
| Age in years | 39 | 46 | 27 | 34 |
| Years with current employer | 10 | 17 | 4 | 7 |
| Years at current address | 9 | 20 | 4 | 8 |
| Debt to income ratio (x100) | 7.09 | 10.92 | 8.03 | 21.10 |
| Credit card debt in thousands | 1.10 | 3.43 | .66 | 2.90 |
| Other debt in thousands | 2.28 | 5.72 | 1.42 | 6.06 |

각 군집의 중심점을 살펴보면 군집별로 어떠한 특징을 가지고 있는지 살펴볼 수 있다.

먼저 **나이**를 기준으로 집단1은 중심점이 39세, 집단2는 46세, 집단3은 27세, 집단4는 34세로 집단1, 2에 비해 집단3, 4가 젊은 나이 쪽에 속한다. **근속연수, 현 거주지 거주 연수**를 기준으로 중심점의 값을 보면 집단1과 집단3이 집단2와 집단4 보다 빚을 적게 지고 있는 것을 알 수 있다. **소득 대비 부채비율, 신용카드 연체액, 기타 체불금액** 부분을 보면 근속연수와 마찬가지로 집단1과 집단3이 집단2와 집단4 보다 빚을 적게 지고 있는 것을 알 수 있다.

따라서 K-Means 기법을 활용해 구분된 군집의 특징은 다음과 같이 해석할 수 있다. **집단1** 은 나이가 많고 근속연수가 적으며 현 거주지에 상대적으로 적게 거주하였고, 빚이 상대적으로 적은 군집, **집단2**는 나이와 근속연수가 많고 현 거주지에 상대적으로 오래 거주하였으며, 빚이 상대적으로 많은 군집, **집단3**은 나이가 젋고 근속연수가 적으며, 현 거주지에 상대적으로 적게 거주하고 빚이 상대적으로 적은 군집, **집단4**는 나이가 젋고 현 직장과 현 거주지에 상대적으로 오래 머물렀으며 빚이 상대적으로 많은 군집이라고 볼 수 있다.



실제 데이터를 살펴보아도 앞선 특징와 같은 모습을 볼 수 있다. 전체적으로 보았을 때, 집단1과 2는 집단3, 4에 비해 나이가 많고, 집단1과 집단3은 집단2, 4에 비해 상대적으로 빚을 적게 진 것을 확인할 수 있다.