|  |
| --- |
| REPORT |
| **과제 1**  데이터 전처리 및 SPSS 통계분석 |

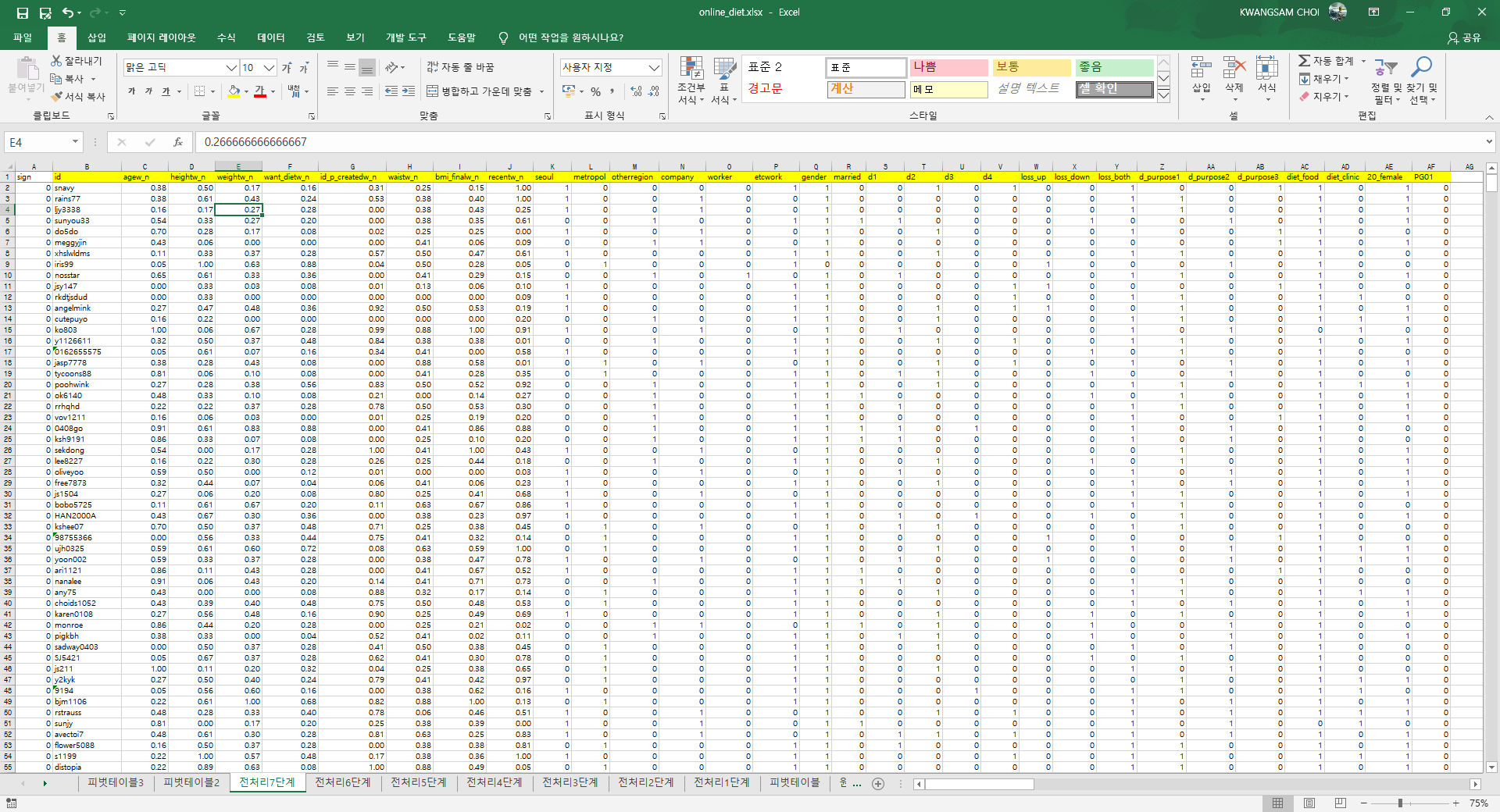




|  |  |  |
| --- | --- | --- |
| 제출일 | 2020. 04. 12 |  |
| 과목명 | 비즈니스애널리틱스개론 |  |
| 담당교수 | 안현철 |  |
| 전공 | 경영정보전공 |  |
| 학번 | 20175288 |  |
| 이름 | 최광삼 |  |

**목차**

1. **데이터 전처리**
2. 원본 데이터
3. 피벗테이블
4. 전처리1단계
5. 전처리2단계
6. 전처리3단계
7. 전처리4단계
8. 전처리5단계
9. 전처리6단계
10. 전처리7단계
11. 피벗테이블2
12. 피벗테이블3
13. 참고
14. **SPSS 통계분석**
15. 빈도분석
16. 카이제곱 검정
17. 독립표본t검정
18. 일원배치 분산분석
19. **데이터 전처리**



online\_diet.xlsx 파일에 대하여 다음과 같은 전처리 과정을 거쳤다. 각각의 시트별로 어떠한 과정을 걸쳐 전처리가 진행되었는지 서술하였으며, 상기 이미지는 최종적으로 전처리가 완료된 모습이다.

1. 원본 데이터

* 엑셀의 ‘중복된 항목 제거’ 기능을 이용해 제품에 관련된 데이터(seq\_num, order\_id, pfid, p\_name, p\_price, p\_created, p\_quantity)를 통합하여 고객 중심의 유의미한 데이터를 추려내었다. (중복된 값 1144개 제거, 남은 데이터 3314개)
* 중복 제거한 데이터에서 pfid는 LEFT 함수를 이용하여 유의미한 앞 7자리 문자만 추출하였고 이를 이용하여 IF, OR 함수로 상품별로 구분하였다. 다이어트 클리닉에 포함되는 제품은 1, 나머지는 0으로 처리하였다. 제품 거래일자를 유의미하게 변형하기 위해 p\_created에 VALUE, LEFT 함수를 이용해 날짜에 해당하는 숫자만 추출하였다.

1. 피벗테이블

* 원본 데이터에서 ID 별로 전략상품 구매 횟수 및 최근 거래일자를 요약해 보기 위해 피벗 테이블을 사용하였다.

1. 전처리1단계

* 위 과정으로 인해 통합되어 의미가 없어진 데이터(seq\_num, order\_id, pfid, p\_name, p\_price, p\_created, p\_quantity)는 ‘전처리1단계’라는 별도의 시트로 분리하고 삭제하였다.
* 시트에 p\_created, PG01 열을 추가하고 VLOOKUP 함수를 이용해 피벗 테이블에서 도출한 데이터(전략제품 구매 여부 및 최근 거래일자)를 1단계 전처리된 데이터에 맞게 넣었다.
* 각각의 열에 대해 COUNTBLANK 함수를 사용하여 빈 값이 얼마나 있는지 파악하였고 이를 통해 공실률(빈 값/전체 데이터)을 계산하였음. 공실률이 높고(약 58%) 무의미하다고 판단한 heap 열은 삭제 처리하였다.
* 행에 대해서도 동일하게 COUNTBLANK 함수를 이용하여 빈 값이 10개 이상인 데이터 2건에 대해서는 무의미하다고 판단해 삭제 처리하였다. (남은 데이터 3312개)

1. 전처리2단계

* 텍스트 처리된 숫자 값을 VALUE 함수를 이용하여 숫자 값으로 바꿔주었다.
* social\_id1 데이터 중 오류로 추정되는 데이터 값 3건에 대해서 삭제 처리하였다.
* (남은 데이터 3309개)
* zip\_code, height, weight 데이터 중 무의미한 값(0, 000-000 등)으로 처리된 데이터 14건은 NULL 값으로 처리하였다.
* waist 데이터에 대해서 잘못 처리된 값은 정상적으로 표출되도록 수정하고 범위 값(예. 28~29, 28-29 등)에 대해서는 중간값으로 처리하였다. 또한 50을 초과하여 단위 오류로 추정되는 값에 대해서는 인치로 변환하였다. waist 값 중 과도하게 큰 값(228, 230) 2건에 대해서는 22.8, 23인치로 처리하였다. NULL 값에 대해서는 평균값(28.24)에서 올림 처리하여 28.3으로 입력하였다.

1. 전처리3단계

* social\_id1 데이터는 10000으로 나누고 INT 함수를 사용하여 생년을 추출하였고, 101에서 생년을 빼어 데이터 생성 당시(2001년)의 나이를 추출하였다.
* zip\_code는 우편번호 체계에 따라 크게 3개(서울, 수도권, 기타지역)으로 구분하였다. NULL 값에는 가장 많이 거주하고 있는 서울(1280건)에 거주한다고 가정하여 전처리 하였다.
* sex 값은 남성과 여성에 따라 구분하기 위해 원래 값에 1씩 빼어 0이면 남성, 1이면 여성으로 처리하였다.
* dise 열은 엑셀의 ‘텍스트 나누기’ 기능을 이용하여 쉼표를 기준으로 각각의 열로 나눈 다음, d1, d2, d3, d4 열을 생성하고 MATCH, ISNA, IF 함수를 이용하여 질병 경험 유무에 따라 1과 0으로 구분하였다.
* Height의 NULL 값 1건에 대해서는 평균값을 적용하였고 비정상적으로 큰 1건(1965)에 대해서는 몸무게 값(60)의 평균값(171)을 적용하여 수정하였다. weight의 비정상적인 값(505.5)에 대해서는 BMI 지수를 역으로 계산하여 50.5로 수정하였다. BMI 열에 대해서도 동일하게 키와 몸무게를 계산하여 수정하였다.
* diet\_purpose 열은 크게 3개(미용목적 1, 2 / 건강목적 3, 4 / 기타 5)로 분류하여 d\_purpose1, 2, 3에 각각 1과 0으로 처리하였다.
* recent 열을 추가하고 가장 최근에 구매한 날짜로부터 얼마나 이전에 구매한 적이 있는지 확인한다.

1. 전처리4단계

* 연속형 변수에 해당하는 데이터(age, height, weight, want\_diet, waist, bmi\_final, recent)에 대해서는 데이터 분석의 원활한 진행을 위해 winsorization 과정을 거쳐 상/하위 5% 값보다 크거나 작은 값에 대해서 보정하였다.
* id생성일자로부터 첫 구매까지 며칠이 걸렸는지를 알아보기 위한 id\_p\_created 열을 생성하고 winsorization 과정을 거쳐 보정하였다.

1. 전처리5단계

* occupation 데이터에 대해서는 IF, OR 함수를 사용하여 크게 3개 분류(직장인[company]: 1, 2, 3, 4, 11, 12 / 개인사업자[worker]: 5, 6, 7, 8, 10 / 기타[etcwork]: 9, 13, 14, 15, 16, 17)로 나누어 1과 0으로 처리하였다.
* loss\_part 데이터에 대해서 엑셀의 ‘텍스트 나누기’ 기능을 이용하여 쉼표를 기준으로 각각의 열로 나눈 다음, 2개(상체: 4, 5, 7, 8, 9 / 하체: 1, 2, 3, 6)로 분류하도록 하고 MATCH, ISNA, IF 함수를 이용하여 구분하였다. 그 다음에는 IF와 AND 함수를 사용하여, 상체살만 빼기를 원하는 사람, 하체살만 빼기를 원하는 사람, 상하체살 모두 빼기를 원하는 사람별로 1와 0으로 구분하였다.
* diet 데이터는 NULL 값에 대해서 COUNTIF 함수를 이용하여 가장 많은 선택(704회)을 받은 1번을 임의로 입력하였다. 또한 ‘텍스트 나누기’ 기능을 이용하여 쉼표를 기준으로 각각의 열로 나누었고, MATCH, ISNA, IF 함수를 이용하여 1단계로 분류한 다음 IF, OR 함수를 이용하여 식이요법(1, 2, 3, 4, 5, 10)과 전문 클리닉(6, 7, 8, 9)의 2가지 분류로 나누었다.

1. 전처리6단계

* 연속형 변수 agew, heightw, weightw, want\_dietw, id\_p\_created, waistw, bmi\_finalw, recentw에 대해서 min-max 정규화 과정을 거쳤다.

1. 전처리7단계

* Random sampling 과정을 거치기 위해 피벗 테이블을 생성하고, 전략 제품을 구매한 사람 295명에 맞추어 샘플을 추출하고 그 중에서 전체의 80%에 해당하는 472명은 학습용 데이터로 사용하고 나머지 20%에 대항하는 118명은 검증용 데이터로 사용하였다.
* Random 열을 생성하여 RAND 함수를 이용하여 랜덤 난수를 생성하고 엑셀의 ‘정렬’ 기능을 이용하여 전략 상품과 랜덤 난수를 오름차순으로 정렬하였다.
* sign 열을 만들어 데이터 분석에 사용할 랜덤한 변수를 추출하고 쓰이지 않는 데이터는 필터를 적용하여 삭제하였다.
* bmi\_finalw\_n을 4등분으로 나누기 위해 IF 함수를 사용하여 0.25 이하는 1, 0.5 이하는 2, 0.75 이하는 3, 나머지는 4로 하고 bmi\_1 열에 데이터 값을 삽입하였다.
* want\_dietw\_n을 4등분으로 나누기 위해 IF 함수를 사용하여 0.25 이하는 1, 0.5 이하는 2, 0.75 이하는 3, 나머지는 4로 하고 want\_diet\_1 열에 데이터 값을 삽입하였다.

1. 피벗테이블2

* id별 전략제품 구매 여부를 알아보기 위한 피벗 테이블로 Random sampling을 하기 위한 데이터 개수 및 비율을 계산하기 위해 사용되었다.

1. 피벗테이블3

* sign 변수를 이용하여 추출된 샘플들이 적절하게 분배되어 있는지 확인하였다.

1. 참고

데이터 전처리 후 사용된 변수 리스트 (TRUE : 0, FALSE : 1)

|  |  |  |  |
| --- | --- | --- | --- |
| 정규화 변수 | | | |
| agew\_n | 나이 | waistw\_n | 허리둘레 |
| heightw\_n | 키 | bmi\_finalw\_n | BMI지수 |
| weightw\_n | 몸무게 | bmi\_1 | BMI지수(4등분) |
| want\_dietw\_n | 희망체중 | id\_p\_createdw\_n | 아이디 생성일로부터 첫 구매가 이루어진 날짜 |
| want\_diet\_1 | 희망체중(4등분) | recentw\_n | 최근 구매일 |

|  |  |  |
| --- | --- | --- |
| 거주지 | | |
| seoul 서울 | metropol 수도권 | otherregion 기타 지역 |

|  |  |  |
| --- | --- | --- |
| 직업 | | |
| company 직장인 | worker 개인사업자 | etcwork 기타 직업 |

|  |  |  |
| --- | --- | --- |
| 다이어트 목적 | | |
| d\_purpose1 미용목적 | d\_purpose2 건강목적 | d\_purpose3 기타목적 |

|  |  |  |
| --- | --- | --- |
| 살 빼고 싶은 부위 | | |
| loss\_up 상체 | loss\_down 하체 | loss\_both 상하체 모두 |

|  |  |  |  |
| --- | --- | --- | --- |
| 질병 여부 | | | |
| d1 부종 | d2 변비 | d3 성인병 | d4 빈혈 및 영양실조 |

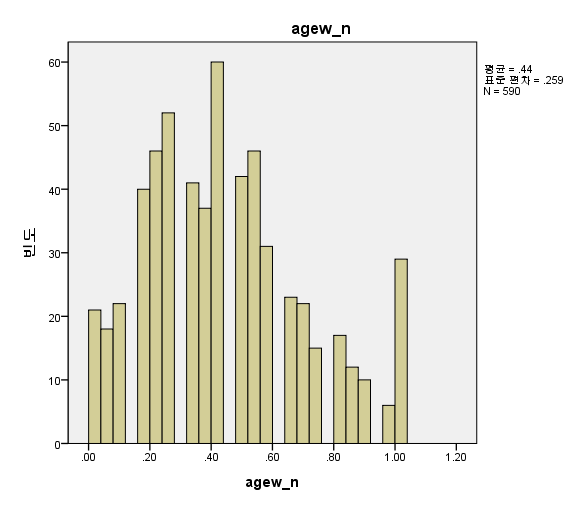
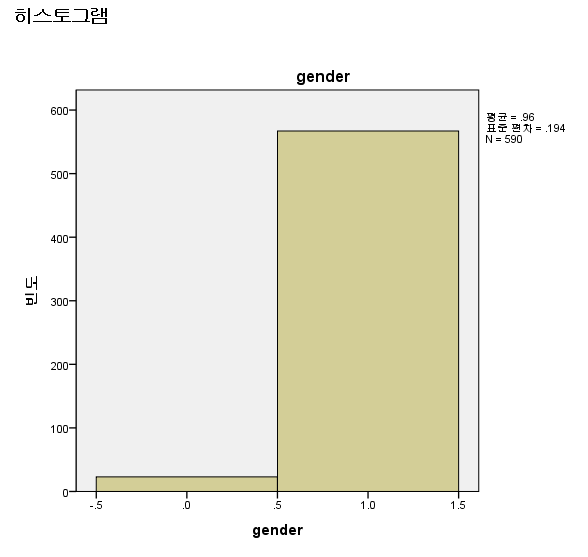
|  |  |
| --- | --- |
| 다이어트 경험 여부 | |
| diet\_food 식이요법 | diet\_clinic 전문 클리닉 |

|  |  |
| --- | --- |
| 기타 변수 | |
| Gender 성별 | 20\_female 20대 여성 여부 |
| married 결혼여부 | PG01 전략상품 구매여부 |

**II. SPSS 통계분석**

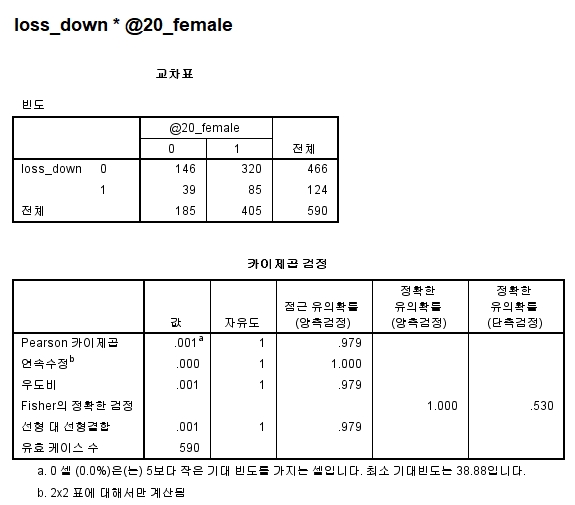
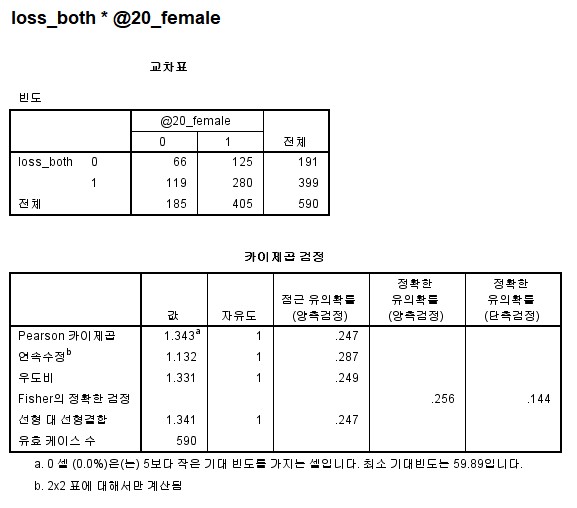
1. 빈도분석

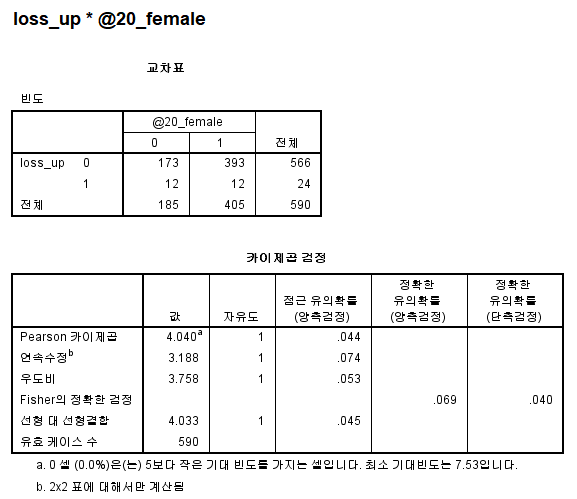
먼저 회원 중 어느 성별과 어느 나이대의 회원이 많은지 알아보기 위해 빈도분석을 실시하였다.



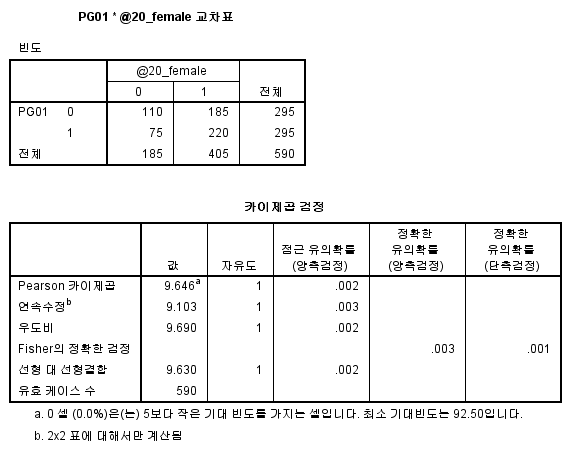
빈도분석 결과 여성 회원이 96.1%를 차지하고 있어, 이 사이트에 가입되어 있는 대다수는 여성임을 알 수 있다. 또한 나이에 대한 결과를 보면 20대 회원이 대다수를 차지하고 있는 것을 알 수 있다. 이를 통해 다이어트에 관심을 가지고 있는 사람은 주로 20대 여성임을 추정해볼 수 있다.

1. 카이제곱 검정

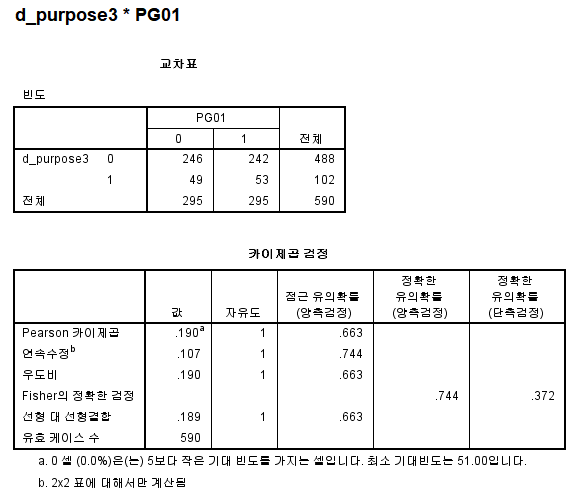
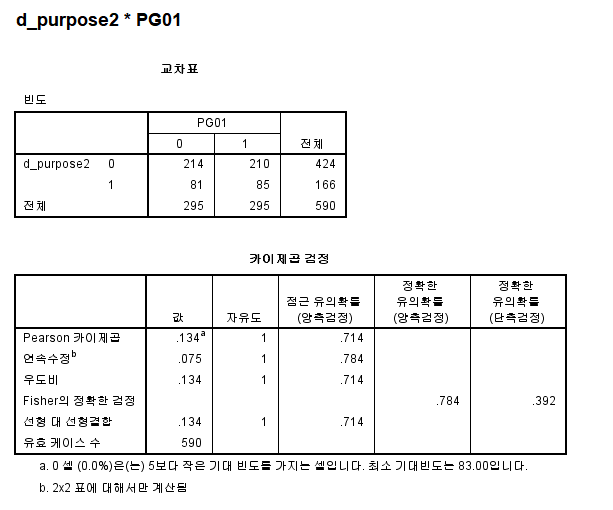
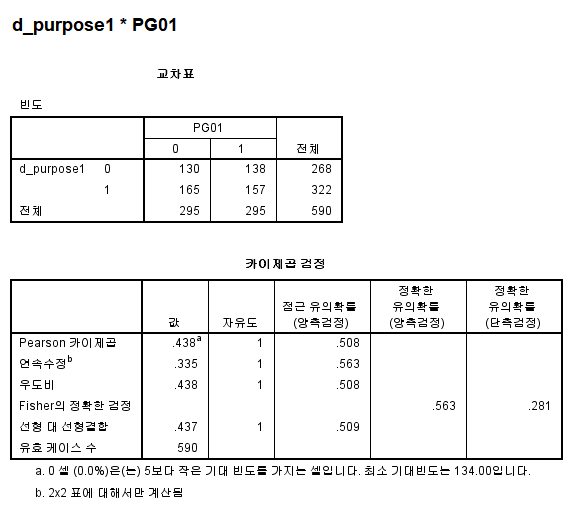




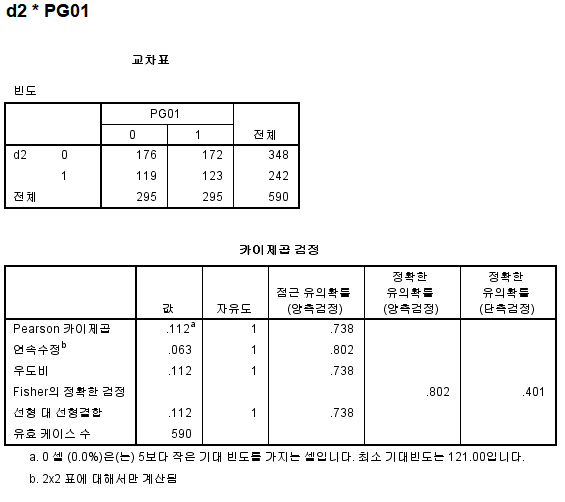
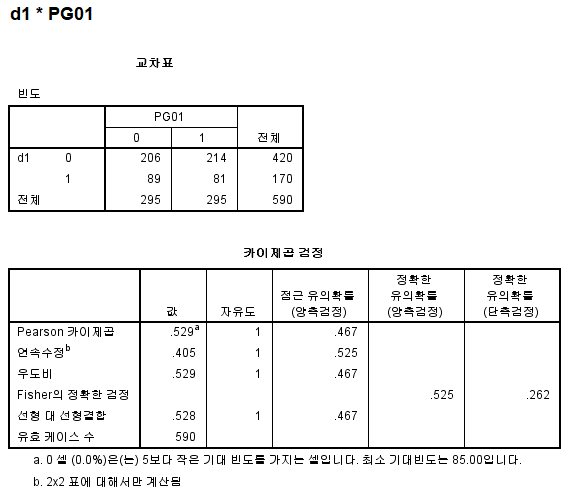
또한 20대 여성이 주로 어느 부위의 살을 빼고 싶어 하는지에 대한 상관관계를 분석한 결과, 95% 신뢰수준 하에서 상체와의 연관성이 있는 것으로 나타났다. 하체와 상하체 모두는 큰 연관성이 없었다.

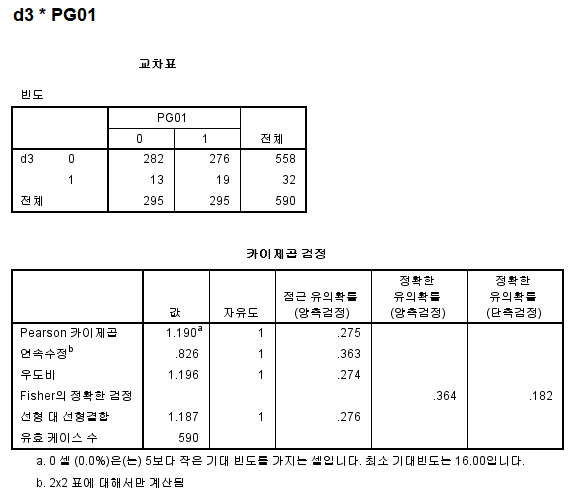
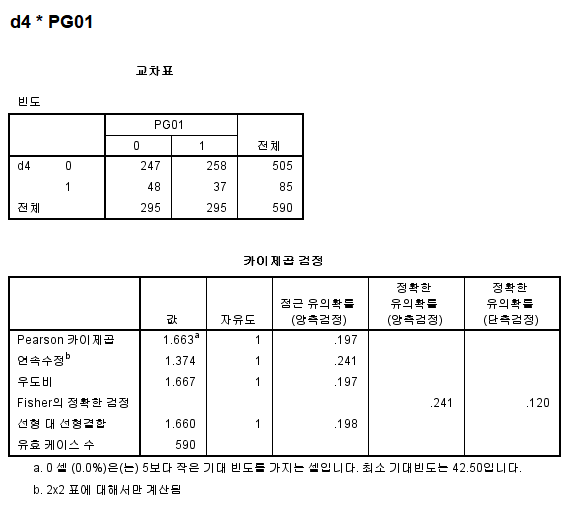


20대 여성이 전략제품을 구매했을 경우를 카이제곱 검정으로 분석하였을 때, 점근 유의확률이 0.002로 상당한 관련이 있는 것으로 나타났다.

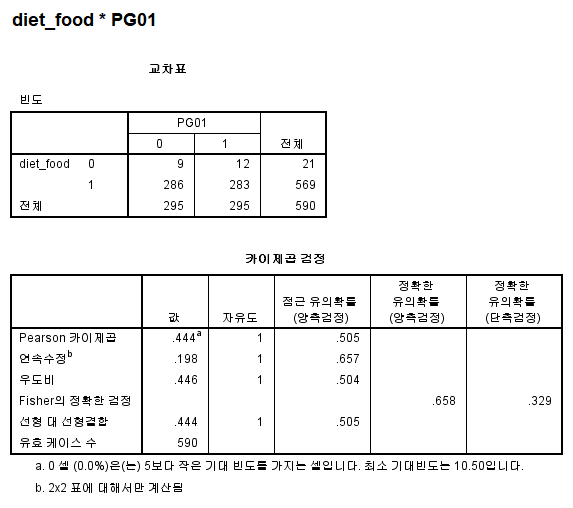
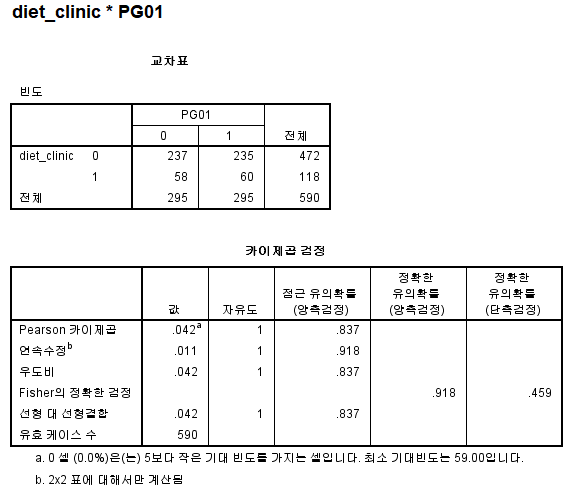


전략제품 구매여부와 다이어트 목적에 상관관계를 카이제곱 검정으로 분석한 결과, 점근 유의확률이 0.5, 0.7, 0.6을 상회하여 다이어트 목적과 전략제품 구매여부는 상관관계가 없는 것으로 나타났다.

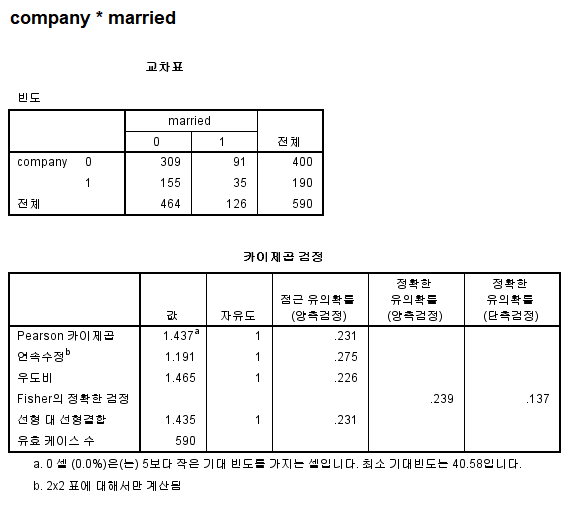


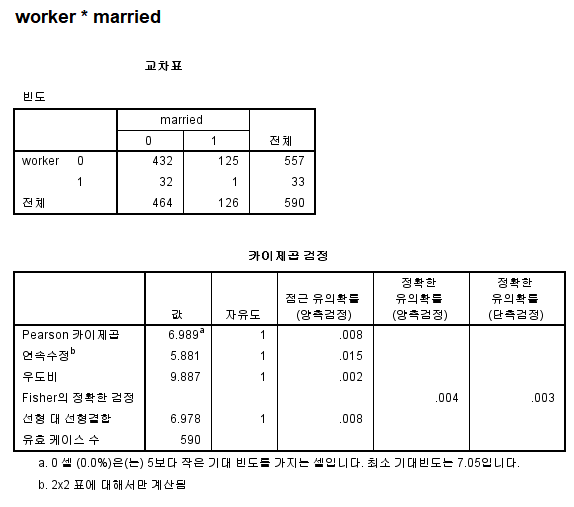
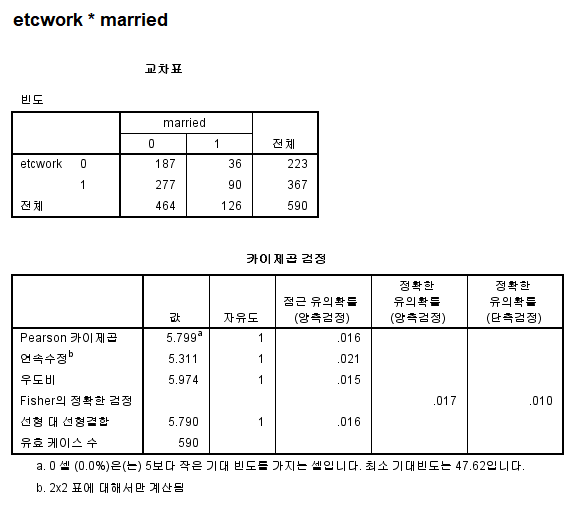


동일하게 카이제곱 검정으로 앓았던 질병과 전략제품 구매여부의 상관관계를 분석하였는데, 앓고 있던 질병과 전략제품 구매여부와의 상관 관계 역시 큰 연관성을 찾아볼 수 없었다.

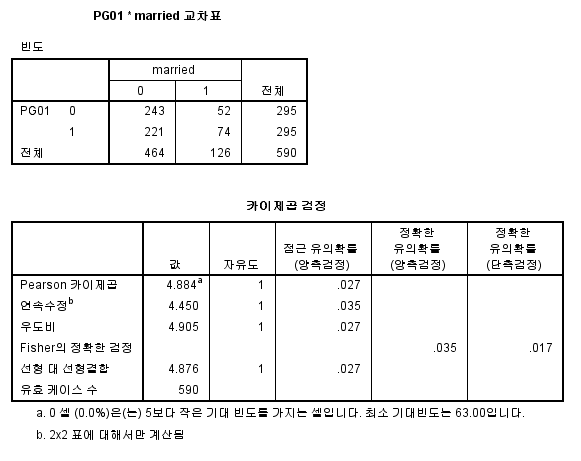


기존에 경험했던 다이어트 방식과 전략제품 구매여부 역시 카이제곱 검정 결과, 연관성을 찾을 수 없었다.



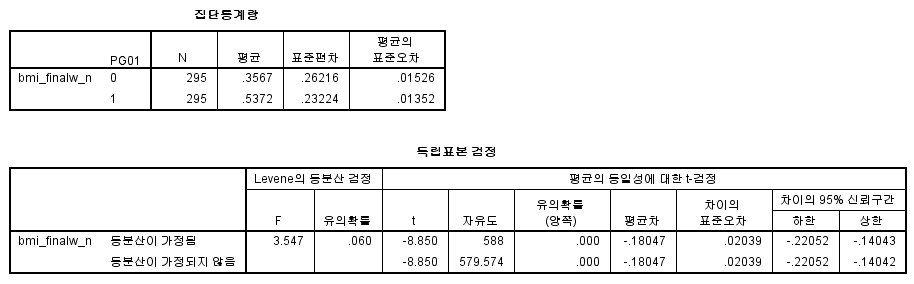


직업과 결혼 여부의 상관관계를 살펴보기 위해 카이제곱 검정을 진행한 결과, 직장인인 사람의 결혼여부는 점근 유의확률이 0.231로 큰 상관관계가 없는 것으로 분석되었다. 반면 개인사업자 및 기타 직업을 가진 사람과 결혼 여부에 대한 분석 결과에서 각각 점근 유의확률이 0.016, 0.08로 계산되어 95% 신뢰수준 하에서 직업이 결혼여부와 관련이 있는 것으로 나타났다.

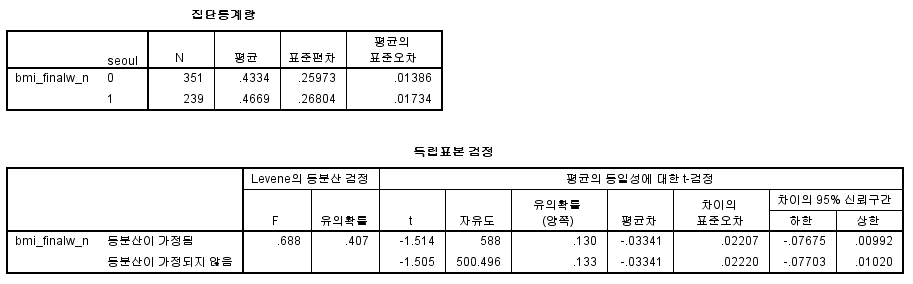


결혼 여부와 전략제품 구매의 상관관계를 알아보기 위해 카이제곱 검정을 진행한 결과, 점근 유의확률의 값이 0.27로 나타났으며, 결혼 여부와 전략제품 구매여부는 95% 신뢰수준 하에서 관련이 있는 것으로 나타났다.

1. 독립표본 t 검정

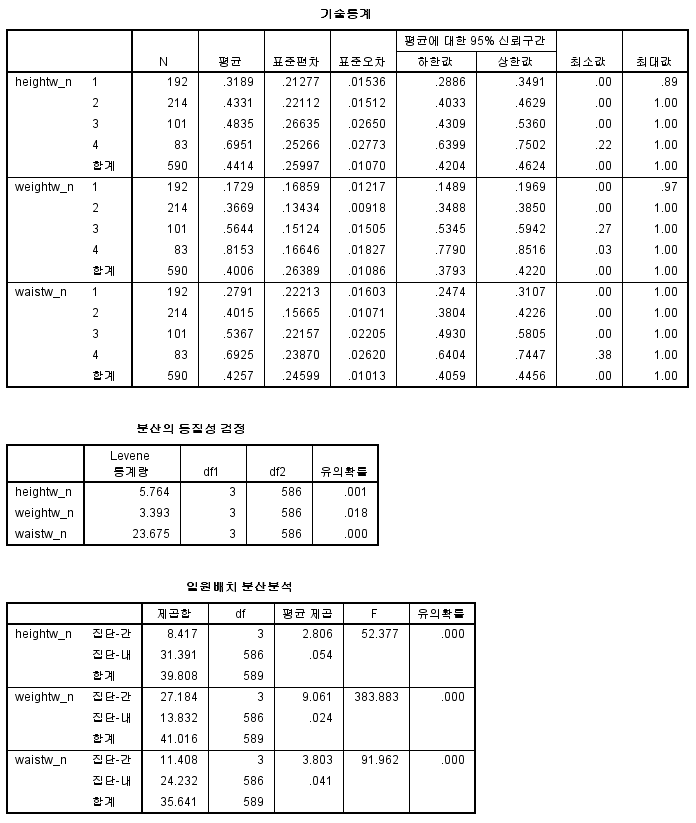


전략제품 구매여부에 따른 BMI 지수의 차이가 있는지를 알아보기 위하여 독립표본t검정을 사용하여 분석을 진행하였다. Levene 등분산 검정 값이 0.06, t검정 유의확률이 0으로 전략제품을 구매한 집단과 구매하지 않은 집단 사이에 BMI 지수의 명확한 차이가 있는 것으로 추정해볼 수 있다.

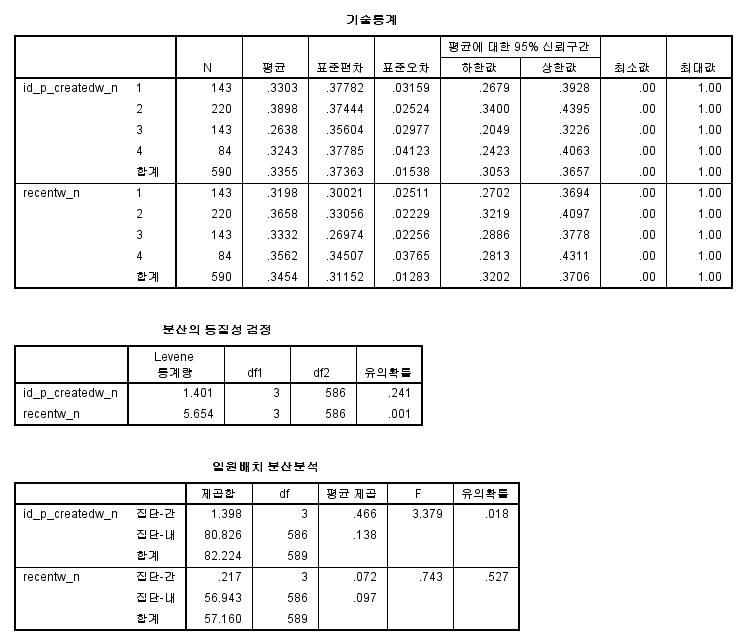


그러나 거주 지역에 따른 BMI 지수의 유의미한 차이는 없는 것으로 예측되었다.

1. 일원배치 분산검정



체중감량 목표에 따른 키, 몸무게, 허리둘레의 연관성을 비교하기 위하여 일원배치 분산검정으로 분석한 결과, 모두의 사례에서 동질성 검정에서의 유의확률이 각각 0.001, 0.018, 0.000으로 나타나 분산의 동질성이 고르게 나타나지 않는 것으로 판단되어 연관성을 알아볼 수 없었다.



BMI 지수에 따른 아이디 생성일로부터 첫 구매일까지의 기간, 그리고 최근 구매일과의 연관성을 일원배치 분산분석으로 분석한 결과, 아이디 생성일로부터 첫 구매일까지의 기간은 동질성 검정에서 유의확률이 0.241로 나타나 분산이 상대적으로 고르게 나타나 있는 것으로 나타났다. 한편 최근 구매일과 BMI 지수의 연관성은 유의확률이 0.001로 분산이 동질하지 못한 것으로 나타났다.

따라서 일원배치 분산분석을 통해서 아이디 생성일로부터 첫 구매까지의 걸린 시간과 BMI 지수에 연관성을 해석해볼 수 있는데, 여기에서 유의확률이 0.018로 나타나 최종적으로 BMI 지수와 아이디 생성일로부터 첫 구매까지의 걸린 시간은 95% 신뢰수준 하에서 연관성이 있다고 판단할 수 있다.