MZ세대 분석 서비스

**빅데이터 분석 결과 보고서**

2023.05.25 ~ 2023.06.25

이준형

**목차**

1. 분석 프로젝트 개요
   1. 개요
   2. 분석 배경 및 기획의도
   3. 분석 목표
2. 분석 데이터
   1. 원자료 설명
      1. 모집단
      2. 표본추출
      3. 조사방법
   2. 데이터 설명
3. 데이터 전처리
   1. 결측값 처리
   2. 연령 그룹화
   3. 범주형 데이터 수치 매핑
4. 데이터 분석
   1. 다중회귀분석
   2. 순서형 로지스틱 회귀 분석
5. 분석 결과
   1. 삶의 만족도 회귀 분석 결과
6. 결론
   1. 분석결과 요약
   2. 분석의 한계
   3. 활용방안
7. 프로젝트 개요
   1. 개요
      * 프로젝트명: MZ세대 삶의 질 분석 서비스
      * 프로젝트 기간: 2023.05.25 ~ 2023.06.25
      * 데이터 분석 언어: R
   2. 분석 배경 및 기획의도

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명최근 많이 사용되는 MZ세대라는 표현은 1980년대 초~1990년대 중반에 태어난 M(밀레니엄)세대와 1990년대 후반~2010년대 초반에 태어난 Z세대를 아울러 이르는 말이다. 다만 학술적으로 정의된 개념인 M세대, Z세대와 다르게, MZ세대라는 표현 자체는 한국의 잡지사로부터 촉발된 신조어에 가깝다. 따라서 사용하는 화자에 따라 MZ세대를 지칭하는 범위가 다른데, 일반적으로는 청년층으로 인식되는 10대 중후반에서 30대 초중반을 가리킨다.

MZ세대는 과열된 학업 경쟁, 높은 청년 실업률과 청년 우울증, 천정부지로 치솟는 주택가격 등 다양한 어려움을 겪으며 살아가고 있다. 한편 유튜브, SNS등 다양한 플랫폼이 활성화되면서 많은 이들이 자신과 타인의 삶을 비교하게 되었는데, 이는 상대적인 박탈감을 유발해 삶에 대한 만족도를 떨어뜨리는 결과를 야기했다. 이에 더해 타인의 삶을 무조건적으로 모방하여 행복을 얻고자 하는 문제도 나타났다.

이에 본 프로젝트는 MZ세대 청년층에게 객관적인 개인의 특성, 가령 학력, 주택 형태, 가구 형태와 같은 개인적인 조건들이 삶의 만족도에 미치는 영향에 대해서 분석하고, 개인에 특성에 맞추어 추천되는 행복을 위한 방안을 제시하고자 한다.

* 1. 분석 목표
* 설문조사 기반의 데이터를 활용하여 MZ 세대 청년들이 느끼는 삶의 만족도와 삶의 질을 분석
* 상기 분석을 통해 사용자 개인의 특성에 맞추어 삶의 만족도에 영향을 미치는 요인을 도출하고 이를 기반으로 사용자 개인에게 적합한 만족도를 개선 방향을 제시

1. 분석 데이터
   1. 원자료 설명

한국 청소년 정책 연구원에서 만 18세부터 만 34세의 청년을 대상으로 실시한 청년 사회 경제 실태 조사 중 2020년 3500여 명과 2021년 2000여 명의 2개년 5500여 명 데이터

* + 1. 모집단

조사 당년 기준 만 18세에서 만34세 사이의 대한민국 거주자

* + 1. 표본추출

2단계 층화추출법을 사용하여 2019년 통계청 인구주택총조사 조사구를 표본으로, 1단계에서 290개 표본조사구를 추출했다. 이후 2단계에서 실제 지역 별 청년 인구 수 비율과 유사한 비율이 되도록 각 조사구의 청년 인구를 추출하였다.

* + 1. 조사방법

면접원이 직접 방문하여 질문하고 그 응답을 면접원이 기록하는 면접타계식을 원칙으로 했다. 그러나 심야 귀가 등의 원인으로 면접원이 조사대상자를 만나기 어려운 경우에는 개인용 설문지에 한하여 유치 조사방법을 사용했다.

* 1. 데이터 설명

|  |  |
| --- | --- |
| 데이터명 | 데이터 설명 |
| 한국청소년정책연구원  청년 사회 경제 실태 설문조사 데이터 | - 한국청소년정책연구원에서 진행한 청년 사회 경제 실태 조사 2개년 데이터 (20년 3520건, 21년 2041건으로 총 5561건)  - 20년 378개, 21년 482개의 변수 중 2개년도에 모두 존재하며 만족도와 직간접적 관계가 있을 것으로 생각되는 변수를 추출   |  |  | | --- | --- | | 컬럼 ID | 컬럼명 | | pid | pk | | sex | 성별 | | age | 만 나이 | | about\_marriage | 결혼관 | | marriage | 결혼 여부 | | car | 자차 여부 | | house | 자가 여부 | | parents | 부모 동거 여부 | | live\_alone | 1인 가구 여부 | | house\_form | 주거 건물 | | house\_rental | 주택 점유 형태 | | health | 건강 | | body | 비만 정도 | | exercise | 주 평균 운동 횟수 | | work\_life | 일-여가 중시 정도 | | real\_dream | 현실-꿈 중시 정도 | | result\_procedure | 결과-과정 중시 정도 | | individual\_group | 개인-집단 중시 정도 | | me\_other | 자기 주관-타인 시선 중시 정도 | | for\_happinese | 행복을 위해 필요한 것 | | edu | 학력 | | int\_family | 가족 교류 빈도 | | int\_friend | 친구 교류 빈도 | | int\_other | 이외 대인 교류 빈도 | | sat\_living | 생활수준 만족도 | | sat\_health | 건강 만족도 | | sat\_achieve | 성취 만족도 | | sat\_human | 대인 관계 만족도 | | sat\_all | 전반적인 만족도 | |

1. 데이터 전처리
   1. 결측값 처리

패널조사 자료의 특성 상 특정 시점에 조사에 응하지 못하여 무응답이 발생하거나 조사에 응한다고 하더라도 일부 항목에 대하여 응답을 거부하거나 여러가지 이유로 인하여 항목 무응답이 발생하는데 이 중 항목 무응답을 처리하기 위하여 데이터 전처리를 진행함. 결측값이 발생한 경우 NA로 대체.

* 1. 연령 그룹화

18~34세 연령을 4개 그룹으로 그룹화함.

10대 후반(18~19), 20대 초반(20~24), 20대 후반(25~29), 30대 초반(30~34)

* 1. 범주형 데이터 수치 매핑

문자형으로 되어있는 범주형 데이터들을 수치 매핑함.

비만 정도, 건강 정도는 ‘매우 양호’부터 ‘매우 나쁨’의 다섯 단계로 되어있어 항목의 경우 ‘매우 양호’부터 1~5의 수치로 매핑했고, ‘예/아니오’로 대답하는 항목에 대해서는 예를 1, 아니오를 2로 설정했다. 다만 ‘자동차를 보유하고 있지 않다’ 항목에 대해서는 아니오가 자차 보유를 뜻하므로, 아니오를 1로, 예를 2로 설정했다. 또한 ‘몇 살에 결혼을 해야 한다고 생각하는가’ 항목에 대해서는 ‘n세에 결혼해야 한다’는 응답을 결혼해야 한다고 생각하는 나이에 상관없이 일괄적으로 1로 설정하여 ‘결혼해야 한다’로 통일했다. 결과적으로 ‘결혼은 해야 한다’를 1, ‘결혼은 해도 안 해도 상관없다’를 2, ‘결혼은 할 필요 없다’를 3으로 설정했다. 이외의 범주형 데이터는 임의로 1부터 수치를 부여하였다.

1. 데이터 분석
   1. 다중 회귀분석

다중 회귀분석이란 독립변수가 둘 이상인 경우 진행하는 회귀분석을 말한다.

본 프로젝트에서는 다수의 독립변수(나이, 학력, 주택 보유 여부 등)와 종속변수(삶의 만족도)간의 상관관계를 알고자 했으므로 R에서 제공하는 lm() 함수로 다중회귀분석을 진행하였다.

Call:

lm(formula = sat\_all ~ ., data = data)

Residuals:

Min 1Q Median 3Q Max

-6.4528 -0.3011 -0.0170 0.3966 4.4932

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.501e+00 2.699e-01 9.265 < 2e-16 \*\*\*

pid -1.448e-05 5.857e-06 -2.472 0.01346 \*

sex 8.687e-02 1.995e-02 4.354 1.36e-05 \*\*\*

age -2.212e-02 1.329e-02 -1.664 0.09609 .

about\_marriage 2.215e-02 1.622e-02 1.365 0.17225

marriage -1.627e-01 5.460e-02 -2.980 0.00290 \*\*

car -2.517e-02 2.245e-02 -1.121 0.26228

house -3.159e-01 4.708e-02 -6.709 2.16e-11 \*\*\*

parents -9.618e-02 3.379e-02 -2.847 0.00444 \*\*

live\_alone -2.622e-01 5.550e-02 -4.725 2.36e-06 \*\*\*

house\_form -2.798e-03 1.222e-02 -0.229 0.81895

house\_rental -1.630e-02 1.939e-02 -0.841 0.40045

health 4.268e-02 1.925e-02 2.218 0.02662 \*

body -7.868e-03 1.389e-02 -0.566 0.57125

exercise 1.992e-02 8.426e-03 2.364 0.01812 \*

work\_life -1.191e-02 6.988e-03 -1.704 0.08850 .

real\_dream -2.593e-02 7.900e-03 -3.282 0.00104 \*\*

result\_procedure 4.138e-02 7.489e-03 5.526 3.43e-08 \*\*\*

individual\_group 1.748e-02 7.692e-03 2.273 0.02307 \*

me\_other -7.807e-02 7.851e-03 -9.944 < 2e-16 \*\*\*

for\_happinese 2.164e-02 3.231e-03 6.696 2.35e-11 \*\*\*

edu 9.867e-02 1.169e-02 8.440 < 2e-16 \*\*\*

int\_family -3.475e-02 1.764e-02 -1.969 0.04895 \*

int\_friend -1.345e-02 1.835e-02 -0.733 0.46378

int\_other -2.502e-02 1.110e-02 -2.255 0.02417 \*

sat\_living 1.182e-01 9.666e-03 12.224 < 2e-16 \*\*\*

sat\_health 6.479e-02 1.033e-02 6.272 3.84e-10 \*\*\*

sat\_achieve 2.722e-01 9.791e-03 27.798 < 2e-16 \*\*\*

sat\_human 3.670e-01 1.332e-02 27.555 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6986 on 5532 degrees of freedom

Multiple R-squared: 0.6293, Adjusted R-squared: 0.6274

F-statistic: 335.4 on 28 and 5532 DF, p-value: < 2.2e-16

우선 전반적인 만족도를 분석한 결과, 결정 계수가 0.6293으로 유의미한 수준의 설명력을 가지고 있었다. 많은 값들이 p-value가 0.5보다 낮아 유의미한 영향력을 보였는데, 이 중 성별, 자가 보유 여부, 1인 가구 여부, 결과-과정 중시 정도, 자기 주관-타인 시선 중시 정도, 행복을 위해 필요한 것, 교육, 생활 수준 만족도, 건강 만족도, 성취 만족도, 대인관계 만족도가 특히 유의미한 영향을 끼치는 것으로 나왔다. 다만 다른 만족도의 경우 전반적인 만족도에 직접적인 영향을 주는 값이기에 특히 p-value가 유의미하게 나온 것으로 생각된다. 따라서 결정 계수가 다른 만족도 값의 영향으로 왜곡되었을 가능성을 고려하여, 다른 만족도 데이터를 제거한 데이터 프레임을 만족도의 수만큼 생성하여 각각 lm() 함수로 분석을 진행해보았다.

Call:

lm(formula = sat\_living ~ ., data = living\_data)

Residuals:

Min 1Q Median 3Q Max

-6.4086 -0.5986 0.1182 0.6817 5.2750

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.348e+00 4.323e-01 10.058 < 2e-16 \*\*\*

pid -4.511e-06 9.606e-06 -0.470 0.63864

sex 8.542e-02 3.267e-02 2.615 0.00896 \*\*

age -1.460e-01 2.160e-02 -6.760 1.52e-11 \*\*\*

about\_marriage -2.146e-01 2.644e-02 -8.117 5.85e-16 \*\*\*

marriage 1.066e+00 8.845e-02 12.051 < 2e-16 \*\*\*

car -7.992e-01 3.510e-02 -22.766 < 2e-16 \*\*\*

house -1.232e+00 7.443e-02 -16.554 < 2e-16 \*\*\*

parents 4.226e-01 5.498e-02 7.686 1.79e-14 \*\*\*

live\_alone 1.576e-01 9.087e-02 1.735 0.08283 .

house\_form 3.788e-02 2.001e-02 1.893 0.05845 .

house\_rental 1.782e-01 3.161e-02 5.640 1.79e-08 \*\*\*

health 3.717e-01 2.408e-02 15.438 < 2e-16 \*\*\*

body -1.575e-02 2.270e-02 -0.694 0.48777

exercise 9.950e-02 1.375e-02 7.235 5.29e-13 \*\*\*

work\_life 2.910e-02 1.138e-02 2.557 0.01057 \*

real\_dream 1.524e-02 1.283e-02 1.187 0.23514

result\_procedure 5.463e-02 1.218e-02 4.486 7.41e-06 \*\*\*

individual\_group 7.777e-02 1.254e-02 6.203 5.95e-10 \*\*\*

me\_other -8.314e-02 1.283e-02 -6.482 9.81e-11 \*\*\*

for\_happinese 4.306e-02 5.261e-03 8.185 3.34e-16 \*\*\*

edu 9.707e-02 1.911e-02 5.079 3.93e-07 \*\*\*

int\_family -4.762e-02 2.889e-02 -1.649 0.09926 .

int\_friend -3.808e-02 2.989e-02 -1.274 0.20272

int\_other -7.716e-02 1.785e-02 -4.322 1.57e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.147 on 5536 degrees of freedom

Multiple R-squared: 0.2867, Adjusted R-squared: 0.2836

F-statistic: 92.7 on 24 and 5536 DF, p-value: < 2.2e-16

가장 먼저 생활수준 만족도를 기준으로 다중 회귀분석을 진행한 결과, 실제로 결정 계수가 0.2867로 모든 만족도를 포함하여 돌렸을 때의 0.6293보다 확연히 낮게 나왔다. 따라서 첫 다중 회귀분석은 다른 만족도 값이 결정 계수에 큰 왜곡을 가한 것으로 사료된다. 추가적으로 다른 세 개의 만족도에 대해서도 각각의 다중 회귀분석을 시행해보았는데, 건강 만족도를 제외한 모든 만족도의 결정 계수가 0.26~0.29 사이로 나왔다. 건강 만족도의 결정 계수만이 0.5 수준으로 비교적 높게 나왔는데, 이는 건강, 비만 정도 등 건강에 직접적으로 연관된 요인들이 존재하는 탓으로 보인다. 추가적으로 다른 만족도 값을 제외한 요인들로 전반적인 만족도를 다중 회귀분석 해보았을 때는 0.29 정도의 결정 계수가 나왔다.

결정 계수가 낮다는 것이 반드시 회귀 모델의 타당성을 부정하는 것은 아니나, 데이터 분석의 신뢰도가 떨어지는 것은 사실이기 때문에 다른 방법을 물색해보았다.

* 1. 순서형 로지스틱 회귀분석

로지스틱 회귀분석은 범주형 데이터를 분석하기 위한 분석 방법이다. 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류기법으로도 볼 수 있는 분석 방법이다. 범주형 데이터가 둘로 나뉘는 이항 로지스틱 회귀분석, 둘 이상으로 나뉘는 다항 로지스틱 회귀분석이 존재한다. 단 범주형 데이터가 어느 정도의 연속성을 가지는 순서형 데이터인 경우 순서형 로지스틱 회귀 분석을 사용하는데, 만족도 데이터의 경우에도 11점으로 이루어진 리커트 척도를 사용했으므로 R에서 제공하는 MASS 패키지의 polr() 함수를 이용하여 순서형 로지스틱 회귀분석을 진행하였고, AER 패키지의 coeftest() 함수를 이용하여 결과를 요약하였다.

z test of coefficients:

Estimate Std. Error z value Pr(>|z|)

pid -3.9367e-05 2.0681e-05 -1.9035 0.056975 .

sex 2.8274e-01 5.6819e-02 4.9762 6.484e-07 \*\*\*

age -7.2792e-02 3.8652e-02 -1.8833 0.059663 .

about\_marriage 8.1128e-02 4.7022e-02 1.7253 0.084468 .

marriage -2.7743e-01 3.0497e-02 -9.0970 < 2.2e-16 \*\*\*

car -1.2223e-01 6.1891e-02 -1.9750 0.048271 \*

house -1.2018e+00 2.4806e-02 -48.4502 < 2.2e-16 \*\*\*

parents -2.4443e-01 4.6300e-02 -5.2793 1.297e-07 \*\*\*

live\_alone -8.3464e-01 2.0520e-02 -40.6747 < 2.2e-16 \*\*\*

house\_form 2.5747e-02 3.4417e-02 0.7481 0.454403

house\_rental -8.4886e-02 4.3249e-02 -1.9627 0.049679 \*

health 1.5145e-01 5.4657e-02 2.7710 0.005588 \*\*

body -3.2610e-02 3.9480e-02 -0.8260 0.408815

exercise 1.0209e-01 2.4819e-02 4.1134 3.898e-05 \*\*\*

work\_life -5.1025e-02 2.0633e-02 -2.4730 0.013400 \*

real\_dream -1.0248e-01 2.3467e-02 -4.3671 1.259e-05 \*\*\*

result\_procedure 1.5329e-01 2.2205e-02 6.9035 5.074e-12 \*\*\*

individual\_group 6.1147e-02 2.2862e-02 2.6746 0.007482 \*\*

me\_other -2.1758e-01 2.2944e-02 -9.4832 < 2.2e-16 \*\*\*

for\_happinese 6.7475e-02 9.6585e-03 6.9860 2.828e-12 \*\*\*

edu 3.6093e-01 3.4757e-02 10.3841 < 2.2e-16 \*\*\*

int\_family -1.0580e-01 4.2330e-02 -2.4995 0.012438 \*

int\_friend 3.1480e-02 5.4409e-02 0.5786 0.562866

int\_other -7.9504e-03 3.3885e-02 -0.2346 0.814497

sat\_living 3.6752e-01 2.9702e-02 12.3736 < 2.2e-16 \*\*\*

sat\_health 1.2721e-01 3.1662e-02 4.0178 5.875e-05 \*\*\*

sat\_achieve 7.7606e-01 3.2893e-02 23.5934 < 2.2e-16 \*\*\*

sat\_human 1.1833e+00 3.7913e-02 31.2099 < 2.2e-16 \*\*\*

0|1 9.7219e-01 3.2782e-03 296.5644 < 2.2e-16 \*\*\*

1|2 1.6735e+00 4.8974e-03 341.7167 < 2.2e-16 \*\*\*

2|3 3.4680e+00 1.6630e-02 208.5407 < 2.2e-16 \*\*\*

3|4 5.3396e+00 3.1132e-02 171.5161 < 2.2e-16 \*\*\*

4|5 6.8539e+00 4.2707e-02 160.4843 < 2.2e-16 \*\*\*

5|6 9.4158e+00 1.1932e-01 78.9096 < 2.2e-16 \*\*\*

6|7 1.1786e+01 1.3361e-01 88.2156 < 2.2e-16 \*\*\*

7|8 1.5850e+01 1.5784e-01 100.4163 < 2.2e-16 \*\*\*

8|9 1.9118e+01 1.9437e-01 98.3561 < 2.2e-16 \*\*\*

9|10 2.1414e+01 1.9286e-01 111.0327 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

우선 생활수준 만족도를 분석한 결과 다음과 같은 결과가 나왔다. 대부분의 값이 지나칠 정도로 유의미한 p-value를 보였다. 이에 각각의 만족도 데이터 및 모든 만족도 데이터가 존재하는 기존 데이터에 순서형 로지스틱 회귀분석을 진행하였는데, 이 또한 대부분의 값이 극도로 낮은 p-value를 나타내었다.

분석 결과가 극단적인 값이 나왔기에 분석 상의 오류가 있을 가능성이 높다고 판단하고, R이 아닌 Python을 사용해 동일한 분석을 진행해보기로 하였다. 우선 순서형 로지스틱 회귀분석을 위해 mord 모듈을 설치하였다. 추가로 머신 러닝에 사용되는 사이킷런 라이브러리를 설치하여 회귀모델의 정확도를 산출해내었다.

우선 모든 만족도가 들어있는 데이터를 사용하여 80%의 데이터를 학습하였고, 20%의 데이터로 정확도를 산출한 결과, 약 0.6765 정도의 수치가 나왔다. 생활수준 만족도의 정확도는 약 0.5902, 건강 만족도의 정확도는 약 0.6199, 성취 만족도의 정확도는 약 0.4806, 대인관계 만족도의 정확도는 약 0.6154가 나왔다. 추가로 상기한 네 종류의 만족도 데이터만으로 전반적인 만족도를 예측했을 경우 0.6118의 정확도를 보였다. 60%에 근접한 세 만족도와 달리, 50% 수준에도 미치지 못한 성취 만족도는 제거하기로 했다.

이후 성취 만족도를 제거하고 다시 산출한 모델의 정확도는 0,6711로, 약 0.3%p 가량 떨어졌으나 유의미한 변화는 아니라고 판단된다. 이후 성취 만족도를 제외한 만족도 예측 모델에서 각 변인이 갖는 계수를 분석해보았다.

Accuracy: 0.6711590296495957

Feature: pid, Coefficient: -0.12359734381830578

Feature: sex, Coefficient: 0.0

Feature: age, Coefficient: 0.4059157189374425

Feature: about\_marriage, Coefficient: 0.16897731135092106

Feature: marriage, Coefficient: 0.28301955072086615

Feature: car, Coefficient: -0.45577929537528195

Feature: house, Coefficient: 0.2382005657706726

Feature: parents, Coefficient: -0.16295902858716163

Feature: live\_alone, Coefficient: -0.23071029745316018

Feature: house\_form, Coefficient: -0.06553923261089206

Feature: house\_rental, Coefficient: -0.48585529134294264

Feature: health, Coefficient: 0.09657902506307364

Feature: body, Coefficient: -0.18142955307859945

Feature: exercise, Coefficient: -0.17029639992601872

Feature: work\_life, Coefficient: 0.04776882190558002

Feature: real\_dream, Coefficient: -0.2171986129003514

Feature: result\_procedure, Coefficient: 0.1387708291511659

Feature: individual\_group, Coefficient: 0.0

Feature: me\_other, Coefficient: -0.6523002496363066

Feature: for\_happinese, Coefficient: -0.012374017647247895

Feature: edu, Coefficient: 0.09640167245055112

Feature: int\_family, Coefficient: 0.6655181087823645

Feature: int\_friend, Coefficient: -0.6196402909628205

Feature: int\_other, Coefficient: -0.6981428120428971

Feature: sat\_living, Coefficient: -0.11537277134616941

Feature: sat\_health, Coefficient: -0.4773135609545594

Feature: sat\_human, Coefficient: -0.11045535476626193

이후 이 회귀 모델에 사용자가 입력한 조건을 변수로 대입하여 predict() 함수를 사용하면 사용자의 각 삶의 만족도 예측 값을 산출해낼 수 있다. 계수의 절댓값이 높은 변수 중 사용자가 현 상황에서 바꿀 수 있는 요소를 추천하여 각 만족도를 높이기 위한 방안을 제시할 수 있을 것으로 보인다.

1. 분석결과

* 삶의 만족도 분석 결과

삶의 만족도에 대하여 분석한 결과 각 만족도 별로 영향력이 높은 변인이 다르게 나타났다.

|  |  |
| --- | --- |
|  | 중요도 |
| 전반적인 만족도 | ① 대인 교류 ② 가족 교류 ③ 자기 주관-타인 시선 중시 정도 |
| 생활수준 만족도 | ① 개인-단체 중시 정도 ② 일-여가 중시 정도 ③ 건강 |
| 건강 만족도 | ① 대인 교류 ② 일-여가 중시 정도 ③ 건강 |
| 대인관계 만족도 | ① 성별 ② 현실-꿈 중시 정도 ③ 건강 |

* 전반적인 만족도의 경우, 대인 교류 및 가족 교류가 가장 큰 영향을 끼치는 것으로 나타났다. 다만 방향성은 반대인데, 대인 교류가 많을수록 전반적인 만족도는 낮은 경향을 보였다. 다른 만족도에서 대인 교류는 대체로 부정적 영향을 주었는데, 대인 교류의 경우 직장에서의 교류를 비롯한 사회 생활을 포함하기 때문에, 사회 생활 스트레스의 영향일 가능성도 고려할 수 있다. 그 외에는 자기 주관-타인 시선 중시 정도가 영향을 주었는데, 대체로 자기 주관을 중시할수록 만족도가 올라가는 추세를 보였다.
* 생활수준 만족도의 경우, 단체 생활보다 개인 생활을 중시할수록 만족도가 높게 나타났다. 그 뒤로는 일-여가 중시 정도인데, 일을 중시하는 경우가 생활수준의 만족도가 높았다. 그 뒤로는 건강으 좋을수록 만족도가 높은 것으로 나타났다.
* 건강 만족도의 경우 건강 상태는 3순위로, 가장 큰 영향은 대인 교류 빈도였다. 2순위는 일-여가 중시도인데, 생활수준과 달리 일보다 여가를 중시할 때 건강 만족도가 높게 나왔다. 건강 정도는 당연히도 건강할수록 건강 만족도가 높았다.
* 대인관계 만족도의 경우에는 특이하게도 성별이 1순위의 영향력을 지녔는데, 여자인 경우가 만족도가 낮았다. 남녀의 친구관계 차이에서 기인한 듯하다. 다른 요소와 달리 그 다음으로는 현실-꿈 중시 정도가 2순위로 나왔는데, 현실을 중시할수록 대인관계 만족도가 높고 꿈을 중시할수록 대인관계 만족도가 낮게 나왔다. 다음으로는 건강 정도가 어느 정도 영향을 주었다.

1. 결론
   1. 분석 결과 요약

데이터 분석의 목적은 MZ세대의 삶의 만족도에 영향을 미치는 요인을 파악하여 만족도가 낮은 사람에게 개선할 수 있는 방법을 제안해주고 결과적으로 삶의 만족을 높이고자 하는데 있다. 각 요인들이 삶의 만족도의 미치는 영향의 정도를 분석한 결과를 살펴보면 다음과 같다.

대인 교류의 경우 생활 수준 만족도를 제외한 모든 만족도에서 높은 영향력을 행사했다. 그러나 대인 교류가 빈번할수록 만족도가 낮은 음의 상관관계를 보였는데, 해당 데이터의 설문 조사에 따르면 대인 교류는 직장에서의 교류를 포함한다. 따라서 직장인이 대부분이면서 조직 내에서 낮은 위치에 존재하는 젊은 세대를 대상으로 한 데이터이기에 이러한 경향이 보이는 것으로 추정된다. 다른 교류의 경우에는 전반적인 만족도에서 가족과의 교류 빈도가 잦을수록 긍정적인 영향을 끼쳤는데, 생활 수준 만족도에서도 가족과의 교류는 어느 정도 양의 관계를 보였다. 다만 그 외의 경우에는 교류 빈도가 일정 이상의 영향력을 행사하지 못했다.

한편 타인의 시선보다 자기 주관을 중시하는 이들이 전체적으로 만족도가 높게 나타났다. 최근 MZ세대 대부분이 sns 과시 문화, 이로 인한 우울증 등 타인의 시선이 부정적으로 작용하기 때문인 듯하다. 그 외에도 일을 우선시하는 것보단 여가를 중요시하는 것이, 꿈보단 현실을 좇는 것이 만족도가 높은 것으로 나타났다. 그 외의 가치관은 특별히 영향을 주지 못했다.

일-여가 중시 정도의 경우, 생활수준에서는 일을 중시하는 경우가 만족도가 높았으나 건강에서는 여가를 중시할수록 만족도가 높게 나타났다. 일을 중시할수록 경제력이 높아져 생활수준에 대한 만족도는 증가하지만, 그만큼 건강에 영향을 주는 것으로 사료된다.

특이하게도 대인관계 만족도에 한정해서 성별이 큰 영향력을 끼쳤는데, 남성이 대인관계 만족도가 높게 나타난다. 정확한 이유는 알 수 없으나, 남성과 여성의 친구 관계에서의 성향 차이인 것으로 생각된다.

한편 데이터 분석에 머신 러닝 기법을 사용하였으나 근본적으로 데이터가 부족한 문제로 인해 높은 정확도를 뽑아내지 못했다. 이로 인해 위에서 서술한 요인들의 영향력이 실제보다 부풀려지거나 축소되었을 가능성도 있고, 프로젝트에 사용함에 있어 사용자의 예상 만족도를 제대로 산출하지 못하거나, 행복을 위한 추천 가이드라인을 잘못 제시할 수 있다는 한계를 가진다.

* 1. 활용방안

분석한 결과를 바탕으로 개인의 조건에 따라 만족도 향상을 위한 방법을 제시하는 서비스를 개발하고자 한다. 이는 MZ세대에 속해 있는 본인의 상황에 대한 척도로서 활용 가능하며 MZ세대가 핵심 타겟인 기업의 마케팅 자료로 활용 가능하다.

또한 사용자의 데이터를 지속적으로 축적, 분석하여 순서형 로지스틱 회귀 분석 및 머신 러닝의 정확도를 높일 것이다. 또한 사용자에게 사용 경험을 설문하고 이를 토대로 각 변인의 영향력을 심층 분석하여, 필요 없는 요인이나 필요한 요인을 추가/제거할 수 있을 것이다.

향후에는 현재 분석 대상인 MZ세대뿐만 아니라 세대를 넓혀 X세대, 은퇴를 바로 앞둔 베이비붐 세대를 위한 삶의 만족도 분석 서비스를 개발하고자 한다.

추가적으로 분석하는 데이터의 범위를 넓혀 취미 관련 빅데이터 분석을 통해 MZ세대의 개인 환경에 따라 선호도가 높을 것으로 예상되는 취미를 추천해주는 서비스를 개발하고자 하는 계획이 있다. 더욱이 각 취미가 만족도에 끼치는 영향을 파악하여 분석 변인으로 추가할 수도 있다.