MZ세대 분석 서비스

**빅데이터 분석 결과 보고서**

2023.05.25 ~ 2023.06.25

팀명 : DreamCatcher

참여인원 : 윤희선, 조민정, 이준형

**목차**

1. 분석 프로젝트 개요
   1. 개요
   2. 기획의도
   3. 분석 목표
2. 분석 데이터
   1. 원자료 설명
      1. 모집단
      2. 표본추출방식
      3. 조사방법
   2. 데이터 설명
3. 데이터 전처리
   1. 특정 연령대 추출
   2. 결측값 처리
   3. 연령 그룹화
   4. 삶의 만족도 수치
4. 데이터 분석
   1. 다중회귀분석
   2. 로지스틱 회귀 분석
5. 분석 결과
   1. 삶의 만족도 회귀 분석 결과
6. 결론
   1. 분석결과 요약
   2. 분석의 한계
   3. 활용방안
7. 참고문헌
8. 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

“MZ세대”는 밀레니얼 세대와 Z세대를 통칭하는 말이다. 그러나 세대를 가르는 기준은 차이가 있는데 , 일반적으로 현재 20·30대 청년들을 가리킨다.

현재의 MZ세대는 학창 시절부터 시작된 타인과의 경쟁 구도, 취업 난이 더해지며 계속해서 오르는 주택 가격과 등 여러 어려움을 겪으며 살아가고 있다. 한편, 유튜브, SNS 등 다양한 플랫폼이 활성화되면서 자신과 타인의 삶을 비교하는 모습을 많이 볼 수 있다. 이에 따라 상대적인 박탈감으로 의해 MZ세대의 삶에 대한 만족도는 계속해서 낮아지고 있다.

따라서 개인의 특성인 학력, 혼인 상태 등이 삶의 만족도에 어떤 영향을 미치는지를 객관적으로 분석하고자 한다.

* 1. 분석 목표
* 설문 조사 기반의 빅데이터를 활용하여 20, 30대 청년들이 느끼는 삶의 만족도와 삶의 질을 분석
* 상기 분석을 통해 MZ세대의 삶의 만족도에 영향을 미치는 요인을 도출하고 이를 통해 개인의 조건에 따라 만족도를 개선할 수 있는 방향성을 제시함

1. 분석 데이터
   1. 원자료 설명

한국 노동연구원에서 약 1만 3000 여 가구를 24년간 추적조사한 데이터 중 4개년도의 자료를 수집한 데이터

* + 1. 모집단

2018년 11월 현재 대한민국 영토 내의 일반 가구에 거주하고 있는 가구원

* + 1. 표본추출

2016년 인구주택총조사 조사구를 1차 표본 추출 단위로 하고 가구를 2차 표본 추출 단위로 하는 2단계 집락 계통추출법을 사용하였다.

먼저 광역시도, 동 읍면 부 및 주택 유형의 층별로 조사구는 지역 순, 가구 순으로 정렬한 후, 조사구별 가구 수 규모에 비례하는 확률비례추출 (Probability Proportionate Sampling) 방법을 사용하였다. 이렇게 해서 얻어진 표본 조사구에서 가구 명부를 작성해 계통추출법으로 10가구를 추출되었다. 2018년 기준 12,134가구가 선정되었다.

* + 1. 조사방법

면접원이 질문하고 그 응답을 면접원이 기록하는 면접 타계식을 원칙으로 한다. 그러나 심야 귀가 등의 원인으로 면접원이 조사대상자를 만나기 어려운 경우에는 개인용 설문지에만 유치 조사 방법을 사용할 수 있다. 실사 기간 동안 유학, 출장, 여행, 입원, 실종, 행상 등의 이유로 함께 거주치 않아 면접조사에 응할 수 없는 가구 구성원에 대하여는 아래의 원칙에 따른다.

1. 실사 기간에 가구로 일시 또는 완전히 복귀하는 사람은 직접조사를 실시한다.
2. 가구로 복귀하지 않는 사람은 가능한 범위 내에서 전화조사를 시행한다.
3. 전화 면접조사가 어려운 경우, 조사일 현재 경제 활동을 하는 가구원 (예 : 취업으로 외지에 상주하는 가구원, 사업상 장기간 출타 중인 사람, 영내거주 직업군인 또는 경찰 등)에 대한 개인용 설문지는 해당 가구원에 가장 가까운 조사대상자가 대리 응답한다. 또한 정신박약, 허약, 치매, 노환 등 정신적, 신체적 장애로 응답할 수 없는 성인은 가장 가까운 조사대상자가 대리 응답한다.
4. 전화면접조사가 어려울 경우, 조사일 현재 비경제활동 인구에 속하는 가구원(예: 영내 거주 군인 및 전투경찰, 취학으로 외지에 상주하는 가구원, 교도소, 소년원, 요양원 등 사회 보호시설 수용자 등)은 개인용 설문지의 조사 대상에서 제외한다. 단, 이들에 대한 연락처 및 인적 사항은 가구용 설문지에서 질문한다.
   1. 데이터 설명

|  |  |
| --- | --- |
| 데이터명 | 데이터 설명 |
| 한국노동패널조사  설문조사 데이터 | - 한국노동연구원에서 진행한 24회차 중 4개 년도의 설문 결과 자료 (12,962건)  - 37개의 변수 중 만족도와 직, 간접적으로 관계가 있을 것으로 예상되는 13개의 변수만 추출하여 사용   |  |  | | --- | --- | | 컬럼 ID | 컬럼명 | | year | 조사년도 | | p\_age | 나이 | | p\_edu | 교육수준 | | p\_firm\_size | 종업원 규모 | | p\_hours | 주당 평균근로시간 | | p\_jobfam2017 | 직종 | | p\_married | 가구원혼인상태 | | p\_region | 거주지역 | | p\_sex | 성별 | | p\_wage | 월평균 임금 | | p0201 | 취업/미취업 여부 | | p0317 | 정규직/비정규직 여부 | | p6508 | 삶의 만족도 | |

1. 데이터 전처리
   1. 특정 연령층 추출

본 프로젝트의 목표 연령이 20~30대이므로 전 연령이 표본인 원자료에서 연령이 20~39세에 대한 데이터만 추출

* 1. 결측값 처리

패널조사 자료의 특성상 특정 시점에 조사에 응하지 못하여 무응답이 발생하거나 조사에 응한다고 하더라도 일부 항목에 대하여 응답을 거부하거나 여러 가지 이유로 인하여 항목 무응답이 발생하는데 이 중 항목 무응답을 처리하기 위하여 데이터 전처리를 진행함. 결측값이 발생하였을 때 NA로 대체하였으며, 미취업자의 월 임금의 경우 0으로 대체

* 1. 연령 그룹화

20~39세 연령을 3세 단위로 그룹화

20대 초반(20~22), 20대 중반(23~26), 20대 후반(27~29), 30대 초반(30~32), 30대 중반(33~36), 30대 후반(37~39)

* 1. 삶의 만족도 수치

만족도는 최소 1점에서 최대 5점의 범주를 가짐

1. 데이터 분석
   1. 다중 회귀분석

다중회귀분석이란 독립변수가 둘 이상일 때 진행하는 회귀분석을 말한다.

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

Call:

lm(formula = sat\_lifeTotal ~ age + edu + marriage + sex, data = train.set)

Residuals:

Min 1Q Median 3Q Max

-2.6058 -0.6416 -0.1416 0.5646 3.0214

Coefficients:

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

(Intercept) 1.541479 0.047261 32.62 <2e-16 \*\*\*

age 0.075285 0.006631 11.35 <2e-16 \*\*\*

edu 0.369160 0.007624 48.42 <2e-16 \*\*\*

marriage -0.245614 0.024054 -10.21 <2e-16 \*\*\*

sex -0.206197 0.017774 -11.60 <2e-16 \*\*\*

---

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

Residual standard error: 0.8478 on 9707 degrees of freedom

Multiple R-squared: 0.2339, Adjusted R-squared: 0.2336

F-statistic: 740.8 on 4 and 9707 DF, p-value: < 2.2e-16

그 결과 p-value는 0.5보다 낮아 유의미한 값을 보이는 듯하였지만, 결정계수가 0.2339로 너무 낮은 값이 산출되었다. 이는 다중회귀분석은 연속형 독립변수에 적합한 분석 방법인 반면에 현재 분석에 이용하고자 하는 독립변수는 대부분 범주형이었던 것이 원인이라고 판단하였다. 이에 유의미한 회귀가 어렵다고 판단하여 범주형 데이터 분석에 적합한 로지스틱 회귀분석을 진행하였다.

* 1. 다중 로지스틱 회귀분석
  + 로지스틱 회귀분석은 범주형 데이터를 분석하기 위한 분석 방법이며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류기법으로도 볼 수 있는 분석 방법이다. 독립변수가 2개 이상인 점을 고려하여 R에서 제공하는 multinom() 함수를 이용하여 다중 로지스틱 회귀분석을 진행하였다.
* 분석에 앞서 Z-test를 통하여 단일변수-종속변수간 p-value를 산출했고, 이 때 유의미한 결과를 보이는 변수들을 채택하였다. 연령의 경우 p-value가 0.5 이상이지만, 연령과 다른 독립변수간 상관관계가 있어 변수에 포함시켰다.
* 상기 방식으로 선택된 변수들을 아래 표와 같이 범주형 데이터로 설정한 후 로지스틱 회귀분석을 진행하였다.

|  |  |
| --- | --- |
| 종속변수 | 삶의 만족도 (sat\_lifeTotal) : 1점 ~ 5점  만족도를 점수별로 나누어 5개의 범주로 진행 |
| 독립변수 | - 연령(age)  1 = 20대 초반 / 2 = 20대 중반 / 3 = 20대 후반 / 4 = 30대 초반 / 5 = 30대 중반 /6 = 30대 후반  - 성별(sex)  1 = 남자 / 2 = 여자  - 학력(edu)  1 = 무학 / 2 = 고등학교 졸업 미만 / 3 = 고등학교 졸업 / 4 = 대학 재학, 중퇴 / 5 = 2·3년제 대학 졸업 / 6 = 4년제 대학 졸업 이상  - 취업여부(hire)  1 = 취업 / 2 = 미취업  - 혼인상태(marriage)  1 = 미혼 / 2 = 유배우자 |

* 다중 로지스틱 회귀 분석(Multinomial Logistic Regression) 진행

R에서 제공하는 multinom() 함수를 사용하여 로지스틱 회귀 분석을 실시

총 12,962개의 데이터 중 75%인 9,712개를 학습 데이터로 나누어 로지스틱 모델을 학습시켜 각 분류 별로 회귀 모델 도출한 결과

Call:

multinom(formula = sat\_lifeTotal ~ age + hire + marriage + sex +

edu, data = train.set)

Coefficients:

(Intercept) age hire marriage sex edu

2 2.9328753 0.08915888 -0.6188779 -1.402977 0.13586685 0.9434152

3 3.7799499 0.10452629 -1.6718984 -2.353177 -0.04319590 1.3975842

4 0.3499232 0.17639996 -2.3634291 -2.059885 -0.06258468 2.0507193

5 -1.9850527 0.16977247 -2.2859893 -2.300455 -0.14621132 2.3454761

Std. Errors:

(Intercept) age hire marriage sex edu

2 0.7706066 0.08277204 0.2762476 0.3416481 0.2718459 0.1196034

3 0.7768380 0.08411287 0.2789100 0.3460566 0.2739208 0.1211181

4 0.7917008 0.08581562 0.2823217 0.3492683 0.2770621 0.1237764

5 0.8531236 0.09022255 0.2920795 0.3589628 0.2852425 0.1339159

Residual Deviance: 21272.01

AIC: 21320.01

위와 같은 다중 로지스틱 회귀분석의 계수를 구할 수 있다.

계수를 통해 만족도 분류 별로 어떤 변수가 영향을 많이 미치는지 알 수 있다.

또한 회귀 모델에 사용자가 입력한 조건을 변수로 대입하여 predict() 함수를

이용하여 사용자의 삶의 만족도 예측 값을 산출해낼 수 있다.

1. 분석결과

* 삶의 만족도 분석 결과

삶의 만족도에 대하여 분석한 결과 모든 그룹에서 취업여부, 혼인상태, 학력이 가장 많은 영향을 미치는 것을 알 수 있었다.

이 세가지 중에서도 각 그룹별로 가중치의 순위는 다르게 나타났다.

|  |  |
| --- | --- |
| 삶의 만족도 | 중요도 |
| 1점 | reference |
| 2점 | ① 학력 / ② 혼인상태 / ③ 취업여부 |
| 3점 | ① 혼인상태 / ② 취업여부 / ③ 학력 |
| 4점 | ① 취업여부 / ② 학력 / ③ 혼인상태 |
| 5점 | ① 학력 / ② 취업여부 / ③ 혼인상태 |

또한 변수의 값에 따라 삶의 만족도가 달라지는 것을 확인할 수 있었다.

* 연령: 연령이 증가할수록 삶의 만족도 증가
* 성별: 여자에 비해 남자가 삶의 만족도가 높음
* 학력: 고학력자일수록 삶의 만족도가 높음
* 혼인상태: 배우자가 있는 사람이 삶의 만족도가 높음
* 취업여부: 미취업자에 비해 취업자가 삶의 만족도가 높음

1. 결론
   1. 분석 결과 요약

데이터 분석의 목적은 MZ세대 즉 2·30대의 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.

연령 또는 성별과 같이 바꿀 수 없는 요인은 삶의 만족도에 많은 영향을 미치지 못했다. 반대로 학력, 취업 여부와 같은 경제활동에 직결되는 요인들이 많은 영향을 미쳤다. 하지만 월 임금, 종사하고 있는 직종, 근무시간과 같은 직접적인 직무 요인은 삶의 만족도에 대한 영향력이 낮은 것으로 나타났다. 또한 배우자가 있는 사람들이 비교적으로 높고, 고연령일수록 높은 것으로 보아 안정적인 삶을 살아가는 사람들이 만족도가 높다는 것을 알 수 있었다.

* 1. 분석의 한계

한국노동패널데이터를 활용하여 분석하는 과정에서 다음과 같은 한계점이 나타났다.

첫째, 삶의 만족을 중점으로 조사된 데이터가 아니기 때문에 영향을 미치는 변수가 다양하고 사람마다 느끼는 정도가 주관적인 삶의 만족도를 정확하게 예측하는 것은 한계가 존재했다.

둘째, 설문조사로 취합된 데이터이기 때문에 중간 수치의 값으로 선택된 데이터가 많았다. 특히 삶의 만족도에 대한 데이터는 보통으로 선택 된 데이터가 대부분이어서 높고 낮음의 차이를 명확하게 분석하는 데에 한계가 존재하였다.

* 1. 활용방안

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

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.

1. 참고문헌

선우지예(2022), 공공부문 종사자의 세대별 직무만족 비교분석 - 직무만족도 비교 및 직무만족 결정요인 분석, 서울대학교 행정대학원 행정학 석사 학위논문