1.안녕하세요. 저희들은 불로소득과 부채가 총 생활비에 미치는 영향을 연구주제로 선정하여 데이터 분석을 진행하게 되었고 이번 발표를 진행할 김정훈이라고 합니다. 그럼 시작하도록 하겠습니다.

2.먼저 목차입니다. 목차는 총8단계로 구성되어 있으며 분석 개요, 데이터준비, 이상치 확인 및 정제, 신뢰구간 확인, 정규성 및 상관관계 확인, 주성분 분석, 다중선형회귀분석, 결과로 구성되어 있습니다.

3.분석 개요입니다. 현대사회에서 금융 부채 문제와 소득 양극화가 이슈가 되고 있어 받은 데이터파일을 확인해본 결과, 연구 주제와 관련된 변수를 확인 할 수 있었습니다. 본 연구를 통해 가구별 총생활비를 부채, 불로소득과의 상관관계를 조사하여 도출된 결론을 바탕으로 다양한 분야와 단체의 도움이 되는 참고 자료를 제공하고자 분석을 진행하였습니다. 저희가 설정한 가설은 첫번째로 금융기관 부채가 증가할수록 총생활비가 감소하는지 여부와 재난지원금을 받는 사람들의 경우 이것으로 인한 효과로 인해 전체적인 생활비의 증가가 있는지 확인하려고 합니다.

4.첫번째로 데이터 준비입니다. 저희들 같은 경우에는 총 15개의 변수를 선정하게 되었고요. 이 중에서 총생활비를 종속변수로 하고 그 외에는 독립변수로 선정했습니다. 독립변수 같은 경우에는 처음에 말씀 드린 것처럼 불로소득과 부채 2가지 카테고리로 나눠집니다. 또한 해당 변수들은 확인 결과 전부 연속형 데이터 및 비율척도로 확인되었으며 마지막으로 모든 변수들의 단위는 만원으로 통일 되어있는 점 참고 부탁드립니다.

5.다음으로 변수 선정을 하던 중 저희들이 선정한 재난지원금,공적연금 같은 것도 불로소득에 포함되는지 의문을 갖게 되었습니다. 왜냐하면 연구주제와 상관없는 변수를 포함하여 데이터 분석을 하게 될 경우 잘못된 결과를 도출할 수 있기 때문인데요. 먼저 불로소득에 대해 간단히 설명 드리자면 말 그대로 일을 하지 않고 얻는 소득을 의미합니다. 불로소득에 범주로는 보시는 바와 같이 여러가지가 포함되는데요. 여기서 공적연금과 기초연금은 연금속성으로 간주하고 재난지원금은 복지제도의 혜택으로 간주해서 이번 조사 대상에 같이 포함하게 되었습니다.

6.본격적으로 전처리 과정에 대해 설명 드리겠습니다. 먼저 데이터셋은 총 7865개이며 컬럼 수는 이전에 말씀드린 것처럼 총 15개로 구성되어 있습니다. 형식은 모두 int, float로 수치형 타입으로 확인되었고 결측치 확인 결과 주택관련부채액에서 1079건이 발견되어 부채액이 없는 것으로 간주하고 수치값을 0으로 대체하였습니다.

7.그 다음으로 저희들이 선정한 변수들의 데이터 수치, 즉 결측치가 아닌 0의 빈도수를 확인해보았습니다. 0값이 너무 많으면 정상적인 분석이 어렵기 때문인데 확인 결과 다행히 선정한 컬럼 모두 최소 1000개 이상 유효하여 따로 제외하는 독립변수 없이 진행하기로 결정했습니다.

8.기초통계량을 통해 기본적인 분포를 확인해보았습니다. 조사가구는 7865가구, 종속변수인 총생활비의 평균값은 약 329만원이며 최대값이 114446만원으로 나왔습니다. 총생활비의 단위는 월 단위 및 만원인데 최대값 같은 경우 월 10억을 사용한다는 결론이 나와 결론적으로 이상치가 있을 것으로 판단하였고 시각화를 통해 알아보기로 하였습니다.

9.첫번째로 히스토그램을 통해 그래프 확인 결과 각 변수의 최대값으로 인해 도수분포가 비정상적인 형태를 보이고 있음을 확인 할 수 있었으며 변수들의 값이 특정 분포에 상당히 치우쳐 있음을 확인 할 수 있습니다.

10.두번째로 레그플롯으로 확인 결과 선형관계를 확인해야 하는데 종속변수인 총생활비의 최대값으로 인해 추세선이 제대로 안 올라오고 눌리는 듯한 모습을 확인 할 수 있었습니다.

11.세번째로 boxplot을 통해 이상치를 확인해보고자 하였습니다. 각 컬럼별 데이터 수치가 거의 1분위수에 몰려 있다 보니 나머지 값들은 전부 이상치로 보여지고 있으나 해당 데이터는 실조사 기반으로 만들어진 데이터이기 때문에 boxplot으로 보이는 모든 이상치를 제거하지 않고 유독 튀는 값들만 제거하기로 팀원들과 논의하였습니다. 그 결과 파이썬 내 query함수를 통해 총 6건의 데이터셋을 제거하기로 결론 및 7859건으로 진행하게 되었습니다.

12. query 적용 후 레그플롯으로 재확인 결과 이상치가 제거되면서 추세선이 확실히 올라오는 것을 확인할 수 있었습니다. 이를 통해 1차적으로 총생활비가 높을수록 대다수 변수들의 수치 또한 오는 양의 상관관계임을 예상할 수 있습니다.

13.query적용 후 히스토그램도 재확인 해보았습니다. 이상치 제거 후 분포도가 개선된 것을 확인 할 수 있었으며 종속변수인 총생활비의 경우 절반 이상의 가구가 월 239만원 이하에 분포되어 있는 것을 확인 할 수 있었습니다.

14.추가적으로 신뢰구간도 구해보았는데 대한민국 전체 가구 모수의 종속변수, 독립변수 추정값 범위를 추측하는 데에 의미가 있다고 판단하여 이렇게 모든 컬럼별로 시각화를 진행해보았습니다.

15.다음으로 페어플롯을 이용하여 모든 컬럼 별 간에 상관관계를 확인해보았습니다. 그래프가 좀 작긴 하지만 몇몇 요인들 간에 상관관계가 있는 것을 확인할 수 있었고 독립변수 중 하나인 기초연금의 경우 대부분의 변수들과 반비례함을 확인할 수 있었습니다.

16.왜도와 첨도를 이용하여 정규분포를 따르는지 확인해보았습니다. 왜도는 자료가 대칭적으로 분포되어 있는지 여부를 확인하는 측도이고, 첨도는 양쪽 꼬리의 두터움을 나타내는 값으로 편차나 이상치가 많을수록 큰 값을 나타내는 게 특징입니다. 보시면 알겠지만 왜도 및 첨도 모두 기초연금을 제외한 나머지 변수들은 정규분포를 따르지 않는 것으로 확인되었습니다. 다만 해당 데이터는 일전에 말씀 드린 것처럼 시민을 대상으로 한 실 데이터 기반이기 때문에 정규분포를 따르지 않을 수 있음을 염두해야 할 것으로 판단됩니다. 우선 해당 데이터는 비선형적이고 정규분포를 따르지 않기 때문에 상관관계 및 유의성을 확인하기 위해 스피어만 상관분석을 이용하기로 결정했습니다.

17.스피어만 상관분석을 통해 확인한 결과 p-value값이 모두 0.05이하로 나와 통계적으로 유의미한 것을 확인 할 수 있었고 이는 변수들 간에 관계가 실제로 존재한다고 말할 수 있을 것입니다. 그리고 부모가 보낸 생활비와 자녀가 보낸 생활비의 경우 음의 관계로 서로 상충관계인 것을 확인 할 수 있습니다.

18.히트맵을 통해 상관관계를 좀 더 자세히 확인해보았습니다. 확인 결과 이전 페어플롯에 그래프로 본 것처럼 몇몇 요인간에 상관관계가 발견됐는데 특히 금융기관대출, 주택관련부채, 주거관련 부채의 이자 변수들의 강한 상관관계를 확인 할 수 있었습니다. 이럴 경우 회귀 분석 시 다중공선성이 발생할 가능성이 있기에 3개의 변수를 순차적으로 제외 후 재확인 하기로 결정했습니다.

19.3개의 변수를 일괄 제외하지 않고 2개씩 짝을 이뤄 제외를 해보았고 이 중 주택관련 부채액과 주거관련 부채의 이자 2개 변수를 제외하니 금융기관대출의 상관계수가 개선되어 금융기관대출은 제외하지 않고 진행하기로 하였습니다.

20.정규성 및 상관관계에 대한 정리본 입니다. 앞서 내용을 전부 설명 드렸기 때문에 넘어갈 수 있도록 하겠습니다.

21.이제 다중회귀분석을 하기 전 주성분 분석을 이용해 종속변수에 대해 미치는 영향력이 적다고 판단되는 변수를 선정해보겠습니다. 먼저 종속변수를 제외한 독립변수들을 동일한 기준으로 맞추기 위해 표준화 진행을 하였습니다.

21.주성분 분석 결과 임대료 연간소득, 주택관련 부채액, 주거관련 부채의 이자가 설명력이 낮은 것으로 나타났습니다. 앞서 히트맵에서 확인된 서로간의 상관관계가 높은 주택관련 부채액과 주거관련 부채의 이자 2개 변수가 확인되어 최종적으로 해당 2개의 변수는 제외 후 1차적으로 다중회귀분석을 진행해보기로 하였습니다.

22.종속변수는 원본 데이터셋으로, 독립변수는 표준화가 된 12개 컬럼의 데이터셋으로 다중회귀분석을 시행한 결과 설명력은 0.019 즉 1.9%의 설명력이 나왔습니다. 이는 여러가지 이유가 있으나 아무래도 변수 간의 수치적인 차이가 심하지 않아 해당 분석은 표준화를 하지 않고 원본 데이터셋으로 진행하는 것이 더 타당하다고 판단하였습니다.

23.2차 다중회귀분석 결과입니다. 설명력은 0.391로 약 39%의 설명력이 나왔으며 이 중에서 유의확률이 0.05이상 나오는 컬럼은 기타소득,기타이자,공적연금 이렇게 3가지가 나왔습니다. 해당 변수들은 통계적으로 유의미한 예측 변인이 아니기 때문에 제외 할 예정이며 케이스별로 전부 비교하여 설명력이 높은 케이스를 찾아보기로 하였습니다.

24.케이스별로 전부 비교를 해본 결과 주성분 분석에서 weak나온 주거관련 부채의 이자 변수를 제외하고 시행한 다중회귀분석의 케이스가 설명력이 f-statistic값이 가장 높게 나와 케이스4번을 최종 분석 모델로 채택하게 되었습니다.

25.케이스4를 이용해 최종 다중회귀분석을 한 결과입니다. 설명력은 0.396으로 약 40%의 설명력을 가지고 있으며 durbin-watson값은 1.878로 2에 근사한 수치를 보여 잔차의 독립성이 가정한다고 판단하였습니다. 또한 모든 변수들의 vif(다중 공선성)값은 10이하로 다중 공선성을 발생하지 않는 것으로 확인되었습니다. 결과적으로 총10개의 독립변수 중 종속변수인 총생활비에 가장 큰 영향을 미치는 변수는 세금환급금, 재난지원금, 기초연금, 민간보조금으로 나타났습니다.

26.회귀계수가 높은 상위 4개의 요인을 먼저 확인해보겠습니다. 첫번째로 세금환급금 입니다. 세금환급은 당연히 많이 소비할수록, 소비가 많을수록 내는 세금(소비세,부가가치세 등등)또한 많아지며 보통 연말정산을 통해 세금환급을 받기 때문에 세금환급이 많다 = 총생활비 값 또한 높다고 결론을 지을 수 있으며 그렇기 때문에 영향력 또한 높다고 결론 내렸습니다. 두번째로 재난지원금 입니다. 코로나19 시기에 생계안정을 위해 지급된 불로소득의 한 종류로서 보통 재난지원금은 전자화폐나 지역화폐로 수령 했으며 이는 곧 소비로 이어졌습니다. 그렇기 때문에 생활비 소비량 또한 올라갔다고 볼 수 있겠습니다. 세번째로 기초연금입니다. 기초 연금은 총생활비와 반비례하는 관계로 나타났으며 기초연금의 특성은 만 65세 이상 일정소득 이하에게 지급되는 성격의 불로소득입니다. 따라서 기초연금을 받는 가구는 소득이 없거나 적을 것이며 이로 인해 자연스럽게 생활비 또한 낮을 것으로 판단됩니다. 마지막으로 민간보조금입니다. 민간 보조금의 정의는 정확히 나와있지 않으나 양의 상관관계로 나타났고 이는 재난지원금과 같은 맥락으로 파악 할 수 있다고 결론 내렸습니다.

27.해당 4가지 요인을 시각화한 결과 실데이터를 어느정도 예측하는 것을 확인 할 수 있습니다.

28. 잔차의 선형성 및 정규성을 확인해보았습니다. 선형성의 경우 앞부분을 제외한 대부분은 선형 기준선을 따라 비교적 만족한다고 판단되며 이를 근거로 분석을 통해 내려진 결론 또한 어느정도 유효할 것이라고 생각됩니다. 정규성의 경우 처음과 끝부분이 따르지 않고 이는 변수들 중에서 평균보다 큰 값이 많아 분포의 오른쪽 끝이 길게 뻗어 있음을 알 수 있었습니다.

29.최종 결론입니다. 저희들이 가정한 가설 중 하나인 “금융기관 대출 부채가 증가할수록 총생활비는 감소할 것이다”는 따르지 않는 것으로 확인되었습니다. 부채가 증가하면 가구 생활수준이 낮아져 생활비도 낮아질 것으로 예상했지만 반대로 금융기관 대출을 하는 사람들의 경우 원금 및 이자 상환을 할 수 있는 가구도 반영하기 때문에 높은 대출액은 곧 높은 수입 및 자본력을 의미한다고 볼 수 있습니다. 따라서 금융기관과의 대출 시 상환 능력과 이자율을 신중히 고려하여 대출 부채를 효율적으로 활용하고 투자 또는 자산 확장에 사용한다면 총 생활비 증가에 긍정적으로 작용할 것으로 판단됩니다. 또 다른 가성 중 하나인 “재난지원금을 수령할 경우 총생활비는 증가할 것이다”는 따르는 것으로 확인되었습니다. 재난 지원금 등의 소득 보조 프로그램이 실제로 사람들의 총생활비 증감에 영향을 미친다면 이것은 사회적 안전망 프로그램의 중요성에 대한 인식을 증가시키고 관련된 행동 변화를 촉진하는데 기여할 수 있다고 판단됩니다. 다만 국가 재정에 부담을 줄 영향이 있으므로 규모와 빈도를 적절히 조절할 필요가 있고 정기적인 모니터링 및 평가를 통해 지원금의 효율성을 높여야 할 것입니다. 마지막으로 생활비는 다양한 요인의 상호작용에 의해 결정되므로 단순 금융 요소 이외에도 복합적인 요소를 고려하는 것이 중요합니다. 이상으로 발표를 마치겠습니다. 감사합니다.