

Introduction to Big Data

Prediction of Biological Age Using a Linear and Multiple Regression with Health Examination Data

Team 10: Soonjun Kwon (21900045), Yubin Ham (21900797),
Ryumin Lee (22200506)

Content

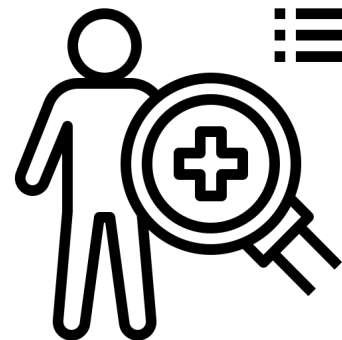
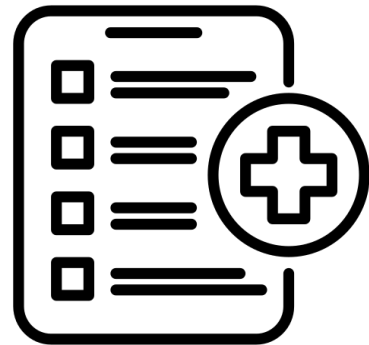
1. Introduction
2. Data preparation
3. Data analysis
4. Conclusion
5. Reference

Keywords: #Health Examination Data #Multiple Regression #Prediction #Biological Age

Introduction

Background: Health examination is important in our life where disease can be prevented

- Health examination (“건강검진”)
 - Conducted **every two years** from the age of 20 (when they **become adults**) in South Korea
 - **Prevent** the occurrence of **disease** and improve their lifestyle to live a **healthy life**
 - Can see obesity, diabetes, dyslipidemia, hypertension and more from results of a health examination
 - As important as **recommended by the country**



Problem Statement: There are a lot of biological terms and numbers that are difficult to understand and interpret

성명		[redacted]		주민등록번호		[redacted]	
사업장명		[redacted]		건강진단장소		[redacted]	
진행	과거병력	진단여부		외상 및 후유증	[redacted]		
	생활습관	[redacted]		일반상태	[redacted]		
구분	검표결과	검사항목		결과	참고치		
개측	비만	신장		cm	[redacted]		
		체중		kg	[redacted]		
		허리둘레		cm	남 90이하/여 85이하		
		체질량지수		kg/m ²	18.5~24.9		
사	시각 이상	시력 (좌/우)		원시/근시	[redacted]		
	청각 이상	청력 (좌/우)		폐쇄음	[redacted]		
	고혈압	혈압(최고/최저)		mmHg	120이하 / 80이하		
요양사	신장질환	요단백		음성	[redacted]		
혈액	빈혈 등	혈색소		g/dL	남 13~18.5 여 12~15.5	남 12~12.9 여 10~11.9	[redacted]
	당뇨병	식전혈당		mg/dL	100 이하	100~125	[redacted]
	고혈압 이상지질혈 증상 악화	총콜레스테롤		mg/dL	200 이하	200~239	[redacted]
		HDL-콜레스테롤		mg/dL	60 이상	40~59	[redacted]
		트리글리세라이드		mg/dL	100~150 미만	150~199	[redacted]
		(LDL-콜레스테롤)		mg/dL	100이하	100~159	[redacted]
	만성신장질환	혈청크레아티닌		mg/dL	1.50 이하	-	[redacted]
	간기능	AST(SGOT)		U/L	400이하	41~50	[redacted]
		ALT(SGPT)		U/L	350이하	36~45	[redacted]
	간장질환	감마지티피프(γ-GTP)		U/L	남 11~63 여 8~35	남 64~77 여 36~45	[redacted]
영상	흉부	[redacted]		[redacted]	정상, 비정상성		
	복부	[redacted]		[redacted]	[redacted]		
*검토의 필요성(연구결과)이 다른 지속적인 혈액 측정과 함께 혈압관리를 위한 생활습관을 유지하기 바랍니다. *시력저하-시력검찰요							
발행처	발행처	[redacted]		발행일	2011-11-21		
판	간주 의뢰 (C)	[redacted]		판정 의뢰	53683		
	환원의 의뢰 (R1)	[redacted]		판정 의뢰	[redacted]		
	환원의 의뢰 (R2)	[redacted]		판정 의뢰	[redacted]		
검진일		2011-11-18		검진기관명		한성의원 (31396925)	

Proteinuria?

Hemoglobin?

Triglyceride?

Creatinine?

Cholesterol?

AST?

ALT?

LDL (Low
density
lipoprotein)

HDL (High
density
lipoprotein)

Fasting blood
glucose (FBG)?

Problem Statement: Many people are facing difficult to understand the results and figures of the medical examination

"직장인 71%, 건강검진 결과지 충분히 이해 못 해"

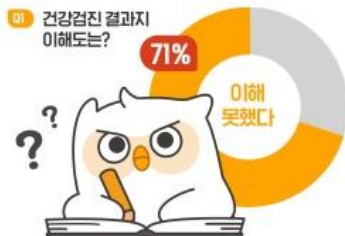
송고 2020.03.25 09:56 | 수정 2020.03.25 09:56
강승혁 기자 (kang0623@ebn.co.kr)

"71% of office workers didn't fully understand the results of the health examination"

GoodRich

건강검진 결과 이해도...
10명 중 7명은
"추가 설명 필요해요"

* 수도권 거주 25~49세 직장인 1000명 대상 조사



(산업경제, 2020)

25~49세 직장인 1000여명에게 물었더니 건강검진 결과지 내용을 충분히 이해하지 못했다는 응답자가 71%였다는 조사결과가 있었다. 수치에 대한 자세한 설명이 없는 데다 어려운 전문용어가 많고 항목이 복잡해 검진 결과의 의미를 정확히 알 수 없다는 것이다. 자주 다니는 병원의 의사가 진료 기록을 보여줘도 뭔 말인지 도통 알 수 없다. 의사가 써주는 처방전도 해석하기 어렵기는 마찬가지다.

(경향신문, 2023)

"There is no detailed explanation of the numbers in result and have many difficult technical terms. So, the meaning of the examination results cannot be accurately understood."

Purpose & Goal: Predict biological age through the numbers of health examination items from health examination result sheet

건강검진 결과통보서									
성명		[인공]			주민등록번호		[인공]		
사망장소		[인공]			건강진단장소		[인공]		
진단	과거병력	진단여부			진단		진단 결과		
	영양섭취	[인공]			진단		[인공]		
구분		검사항목			결과		참고치		
비만	신장	cm			혈당(A2(당화혈당))		[인공]		
		체중			kg				
		허리둘레			cm				
		체질량지수			kg/m ²				
시각	시력 (右/左)			[인공]			[인공]		
	청각 (右/左)			[인공]			[인공]		
신장	혈당 (공복)			mmHg			120/80 / 80/60		
	혈당 (당고혈당)			mmHg			120/80 / 80/60		
신장	혈액 소	g/dL			M 13-16.5		M 13-12.9 / 16.6-17.5		
		혈액 소			g/dL		0.12-15.5 / 15.5-16.5		
		혈액 소			mg/dL		100-125		
		혈액 소			mg/dL		200-239		
신장	HDL-콜레스테롤	mg/dL			60 이상		40-59		
		LDL-콜레스테롤			100-150 미만		150-199		
		LDL-콜레스테롤			1000 이상		100-159		
		LDL-콜레스테롤			1500 이상		1500 이상		
신장	AST(SGOT)	U/L			400 이하		41-59		
		ALT(SGPT)			350 이하		36-45		
		간기능지표 (γ-GTP)			M 11-63		M 11-63 / 63-77		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
		간기능지표 (γ-GTP)			77 이상		77 이상		
신장	간기능지표 (γ-GTP)	U/L			77 이상		77 이상		

Data Preparation

Data Searching: Health examination result data

Data from National Health Interview Survey (NHIS)

국민건강보험공단_건강검진정보

건강검진정보란 국민건강보험의 직장가입자와 40세 이상의 피부양자, 세대주인 지역가입자와 40세 이상의 지역가입자의 일반건강검진 결과와 이들 일반건강검진 대상자 중에 만40세와 만66세에 도달한 이들이 받게 되는 생애전환기건강진단 수검이력이 있는 각 연도별 수진자 100만 명에 대한 기본정보(성, 연령대, 시도코드 등)와 검진내역(신장, 체중, 총콜레스테롤, 혈색소 등)으로 구성된 개방데이터입니다.

파일내 항목별 코드설명은 "국민건강정보데이터 건강검진정보 사용자 매뉴얼"을 참고해 주시기 바랍니다.
(하단의 "주기성 과거 데이터")

9

1

관심

파일데이터

오픈API

추천데이터

공공데이터활용지원센터는 공공데이터포털에 개방되는 3단계 이상의 오픈 포맷 파일데이터를 오픈 API(RestAPI 기반의 JSON/XML)로 자동변환하여 제공합니다.
오픈 API를 활용하기 위해서는 공공데이터포털 회원 가입 및 활용신청이 필요하며, 활용 관련 문의는 공공데이터활용지원센터로 연락주시기 바라며,
데이터 자체에 대한 문의는 아래 제공기관의 관리부서 전화번호로 연락주시기 바랍니다.
파일데이터는 로그인 없이 다운로드를 통해 이용하실 수 있습니다.

csv 국민건강보험공단_건강검진정보

다운로드

We used data from
2017 to 2021

국민건강보험공단_건강검진정보_2017

국민건강보험공단_건강검진정보_2018

국민건강보험공단_건강검진정보_2019

국민건강보험공단_건강검진정보_2020

국민건강보험공단_건강검진정보_2021

Data Description: Column explanation

Name of Column (name in English) : Explanation

- 기준년도 (HCHK_YEAR) : Year of health examination
- 가입자일련번호 (IDV_ID) : ID
- 성별코드 (SEX) : Gender
- 연령대코드(5세단위) (AGE_GROUP) : Age group
- 시도코드 (SIDO) : Provinces
- 신장(5Cm단위) (HEIGHT) : Height
- 체중(5Kg 단위) (WEIGHT) : Weight
- 허리둘레 (WAIST) : Waist measurement
- 시력(좌) (SIGHT_LEFT) : Sight (left)
- 시력(우) (SIGHT_RIGHT) : Sight (right)
- 청력(좌) (HEAR_LEFT) : Hearing (left)
- 청력(우) (HEAR_RIGHT) : Hearing (right)
- 수축기혈압 (BP_HIGH) : Systolic blood pressure
- 이완기혈압 (BP_LWST) : Diastolic blood pressure
- 식전혈당(공복혈당) (BLDS) : Fasting blood sugar
- 총콜레스테롤 (TOT_CHOLE) : Total cholesterol
- 트리글리세라이드 (TRIGLYCERIDE) : Triglyceride
- HDL콜레스테롤 (HDL_CHOLE) : HDL cholesterol
-
- 제3대구치(사랑니)이상 (WSDM_DIS_YN) : Wisdom teeth
- 치석 (TTR_YN) : Plaque
- 데이터공개일자 (DATA_STD_DT) : Date of data created

Data of 2020 health examination result

	기준년 도	가입자 일련 번호	시도코 드	성별코 드	연령대 코드(5 세단위)	신장(50cm 단위)	체중(5Kg 단위)	허리둘 레	시력 (좌)	시력 (우)	...	치아우식증 유무	치 석	데이터 공개 일자
0	2020	1	36	1	9	165	60	72.1	1.2	1.5	...	NaN	NaN	2021-12-29
1	2020	2	27	2	13	150	65	81.0	0.8	0.8	...	NaN	NaN	2021-12-29
2	2020	3	11	2	12	155	55	70.0	0.6	0.7	...	NaN	NaN	2021-12-29
3	2020	4	31	1	13	160	70	90.8	1.0	1.0	...	0.0	0.0	2021-12-29
4	2020	5	41	2	12	155	50	75.2	1.5	1.2	...	NaN	NaN	2021-12-29
...
999995	2020	999996	41	2	13	145	55	81.0	0.9	1.0	...	NaN	NaN	2021-12-29
999996	2020	999997	26	2	12	160	55	76.5	1.5	1.5	...	NaN	NaN	2021-12-29
999997	2020	999998	48	1	9	175	70	85.0	1.2	1.2	...	NaN	NaN	2021-12-29
999998	2020	999999	48	2	9	160	70	91.0	0.6	0.5	...	NaN	NaN	2021-12-29
999999	2020	1000000	28	1	11	160	55	76.1	2.0	2.0	...	NaN	NaN	2021-12-29

1000000 rows x 31 columns

Columns for items that are included in a health examination

Data Description: Explanation of age group (Dependent variable) column

Column of Age Group

연령대 코드(5세 단위)
9
13
12
13
12



Age Code	Age Group
1	0 ~ 4
2	5 ~ 9
3	10 ~ 14
4	15 ~ 19
5	20 ~ 24
6	25 ~ 29
7	30 ~ 34
8	35 ~ 39
9	40 ~ 44

Age Code	Age Group
10	45 ~ 49
11	50 ~ 54
12	55 ~ 59
13	60 ~ 64
14	65 ~ 69
15	70 ~ 74
16	75 ~ 79
17	80 ~ 84
18	85 +



It does not mean
12 years old

Data Preprocessing

1. We have **selected columns** that can be used to predict biological age (through background research)

```
['성별코드', '연령대코드(5세단위)', '신장(50cm단위)',  
 '체중(5Kg단위)', '허리둘레', '시력(좌)', '시력(우)', '청력(좌)', '청력(우)',  
 '수축기혈압', '이완기혈압', '식전혈당(공복혈당)', '총콜레스테롤', '트리글리세라이드',  
 'HDL콜레스테롤', 'LDL콜레스테롤', '혈색소', '요단백', '혈청크레아티닌',  
 '(혈청지오티)AST', '(혈청지오티)ALT', '감마지티피']
```



2. We have **concatenate** all of DataFrame from 2017 to 2021

```
#2017~2021 Combine Data into One  
merged_health_df = pd.concat([health_df_2017, health_df_2018, health_df_2019, health_df_2020, health_df_2021], axis=0)
```

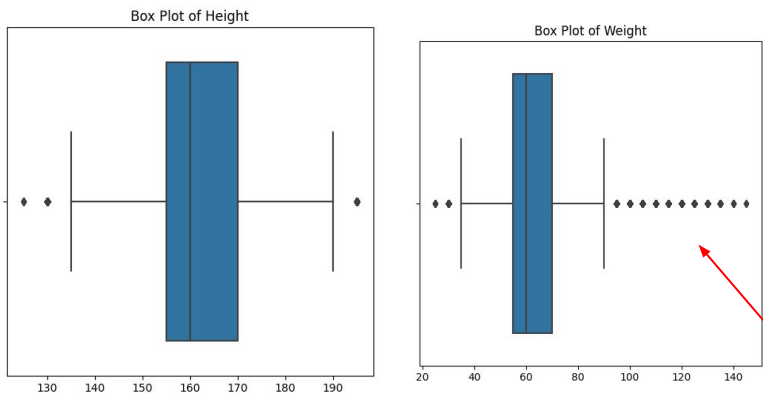
Data Preprocessing

3. Changed column names in English

```
# Rename columns with parentheses
merged_health_df = merged_health_df.rename(columns={
    '기준년 도': 'year',
    '성별코드': 'Gender_Code',
    '연령대 코드(5세 단위)': 'Age_Code',
    '신장(50cm단위)': 'Height',
    '체중(5kg단위)': 'Weight',
    '허리둘레': 'Waist_Circumference',
    '시력(좌)': 'Vision_Left',
    '시력(우)': 'Vision_Right',
    '청력(좌)': 'Hearing_Left',
    '청력(우)': 'Hearing_Right',
    '수축기혈압': 'Systolic_Blood_Pressure',
    '이완기혈압': 'Diastolic_Blood_Pressure',
    '식전혈당(공복혈당)': 'Fasting_Blood_Sugar',
    '총콜레스테롤': 'Total_Cholesterol',
    '트리글리세라이드': 'Triglycerides',
    'HDL콜레스테롤': 'HDL_Cholesterol',
    'LDL콜레스테롤': 'LDL_Cholesterol',
    '혈색소': 'Hemoglobin',
    '요단백': 'Proteinuria',
    '혈청크레아티닌': 'Serum_Creatinine',
    '감마지티피': 'Gamma_GT',
    '(혈청지오티)AST': 'AST',
    '(혈청지오티)ALT': 'ALT',
})
```

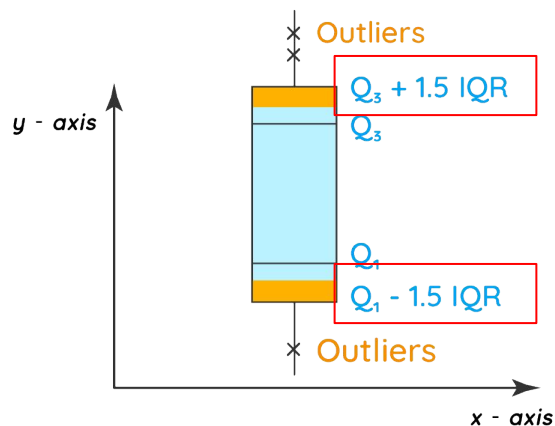


4. Draw box plot to see the outliers of each column values



Data Preprocessing

5. We made a function to **remove outliers** in columns that have outliers.



```
# Define a function to replace outliers with NaN using IQR
def replace_outliers_with_nan_iqr(df):
    Q1 = df.quantile(0.25)
    Q3 = df.quantile(0.75)
    IQR = Q3 - Q1
    outliers_mask = (df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))
    df[outliers_mask] = np.nan
    return df
```

Data Preprocessing

5. We checked and dropped rows with any NaN values

6. Data are preprocessed to use in the model

```
merged_health_df.isnull().sum()
merged_health_df.dropna(inplace=True)
merged_health_df.isnull().sum()
```

year	0	year	0
Gender_Code	0	Gender_Code	0
Age_Code	0	Age_Code	0
Height	0	Height	0
Weight	101285	Weight	0
Waist_Circumference	37663	Waist_Circumference	0
Vision_Left	53315	Vision_Left	0
Vision_Right	48513	Vision_Right	0
Hearing_Left	976	Hearing_Left	0
Hearing_Right	977	Hearing_Right	0
Systolic_Blood_Pressure	91544	Systolic_Blood_Pressure	0
Diastolic_Blood_Pressure	100705	Diastolic_Blood_Pressure	0
Fasting_Blood_Sugar	370721	Fasting_Blood_Sugar	0
Total_Cholesterol	2549273	Total_Cholesterol	0
Triglycerides	2660865	Triglycerides	0
HDL_Cholesterol	2571180	HDL_Cholesterol	0
LDL_Cholesterol	2579245	LDL_Cholesterol	0
Hemoglobin	73599	Hemoglobin	0
Proteinuria	50156	Proteinuria	0
Serum_Creatinine	47348	Serum_Creatinine	0
AST	334012	AST	0
ALT	372728	ALT	0
Gamma_GT	493147	Gamma_GT	0
dtype: int64		dtype: int64	



Before Data Preprocessing

```
print(merged_health_df.shape[0])
```

5048575



After Data Preprocessing

```
merged_health_df.shape[0]
```

1744685

Data analysis

Analytic Method: We used multiple linear regression to predict biological age

Multiple Linear Regression

$$y = (a \times x1) + (b \times x2) + (c \times x3) \dots\dots\dots + (v \times x21) + B$$

Predict
Biological Age

Using independent variables

Dep. Variable:

Age_Code

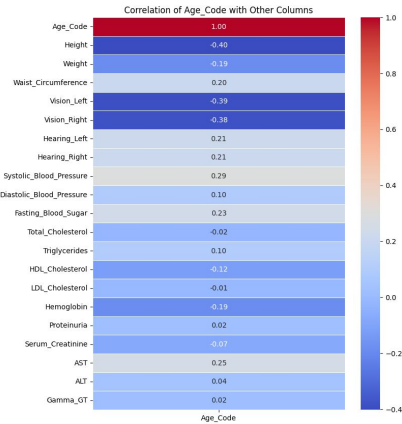
```
independent_variables = [  
    'Gender_Code', 'Height', 'Weight', 'Waist_Circumference', 'Vision_Left',  
    'Vision_Right', 'Hearing_Left', 'Hearing_Right', 'Systolic_Blood_Pressure', 'Diastolic_Blood_Pressure',  
    'Fasting_Blood_Sugar', 'Total_Cholesterol', 'Triglycerides', 'HDL_Cholesterol', 'LDL_Cholesterol',  
    'Hemoglobin', 'Proteinuria', 'Serum_Creatinine', 'AST', 'ALT', 'Gamma_GT'  
]
```

	coef
Gender_Code	-1.3438
Height	-0.0840
Weight	-0.0966
Waist_Circumference	0.1244
Vision_Left	-0.8612
Vision_Right	-0.7724
Hearing_Left	0.8515
Hearing_Right	0.7764
Systolic_Blood_Pressure	0.0406
Diastolic_Blood_Pressure	-0.0231
Fasting_Blood_Sugar	0.0280
Total_Cholesterol	0.0099
Triglycerides	-0.0008
HDL_Cholesterol	-0.0232
LDL_Cholesterol	-0.0082
Hemoglobin	-0.2389
Proteinuria	-0.0107
Serum_Creatinine	0.4939
AST	0.0968
ALT	-0.0376
Gamma_GT	-0.0055

	coef
Intercept	18.5869

Visualization: To understand and evaluate model

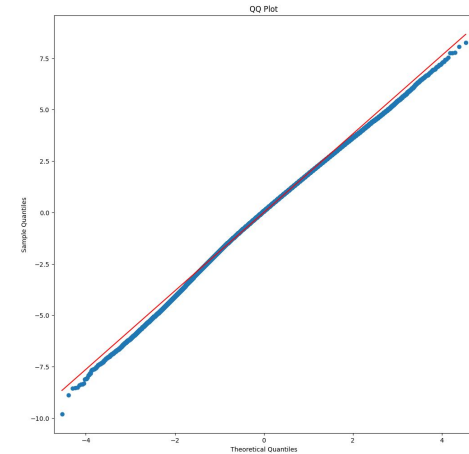
Correlation Heatmap



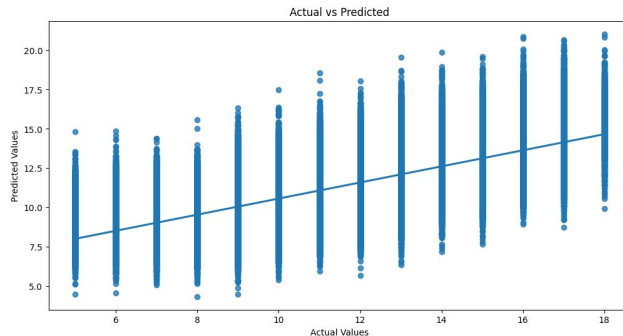
OLS Regression Results

OLS Regression Results						
Dep. Variable:	Age_Code	R-squared:	0.517			
Model:	OLS	Adj. R-squared:	0.517			
Method:	Least Squares	F-statistic:	8.891e+04			
Date:	Tue, 16 Dec 2023	Prob (F-statistic):	0.00			
Time:	15:51:17	Log-Likelihood:	-3.5825e+06			
No. Observations:	1744995	AIC:	7.187e+06			
Df Residuals:	1744993	BIC:	7.187e+06			
Df Model:	21					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	18.5869	0.063	296.474	0.000	18.464	18.710
gender_code	-1.3438	0.005	-266.236	0.000	-1.354	-1.334
Height	-0.0040	0.000	-204.082	0.000	-0.005	-0.003
Weight	-0.0066	0.000	-331.714	0.000	-0.007	-0.006
Waist_Circumference	0.1244	0.000	421.729	0.000	0.124	0.125
Vision_Left	-0.8612	0.005	-161.177	0.000	-0.872	-0.851
Vision_Right	-0.7724	0.005	-144.476	0.000	-0.783	-0.762
Hearing_Left	0.8515	0.009	93.688	0.000	0.834	0.869
Hearing_Right	0.7754	0.009	83.882	0.000	0.756	0.795
Systolic_Blood_Pressure	0.0406	0.000	260.419	0.000	0.040	0.041
Diastolic_Blood_Pressure	-0.0229	0.000	-102.383	0.000	-0.024	-0.022
Fasting_Blood_Sugar	0.0280	0.000	215.180	0.000	0.028	0.028
Total_Cholesterol	0.0099	0.001	16.511	0.000	0.009	0.011
Triglycerides	-0.0068	0.000	-6.680	0.000	-0.007	-0.006
HDL_Cholesterol	-0.0282	0.001	-37.887	0.000	-0.029	-0.027
LDL_Cholesterol	-0.0082	0.001	-13.783	0.000	-0.009	-0.007
Hemoglobin	-0.2389	0.001	-169.776	0.000	-0.242	-0.236
Proteinuria	-0.0107	0.004	-2.655	0.004	-0.019	-0.003
Serum_Creatinine	0.4359	0.005	98.203	0.000	0.424	0.448
AST	0.0889	0.000	305.361	0.000	0.086	0.091
ALT	-0.0376	0.000	-154.697	0.000	-0.038	-0.037
Gamma_GT	-0.0055	0.000	-42.022	0.000	-0.006	-0.005
Intercept	12895.095	Durbin-Watson	1.349			
Prob(omni)	0.000	Jarque-Bera (JB)	12895.095			
Nu	2.148	Prob(JB)	0.000			
h	0.190	Cond. No.	1.65e+04			

QQ Plot



Actual vs Predicted



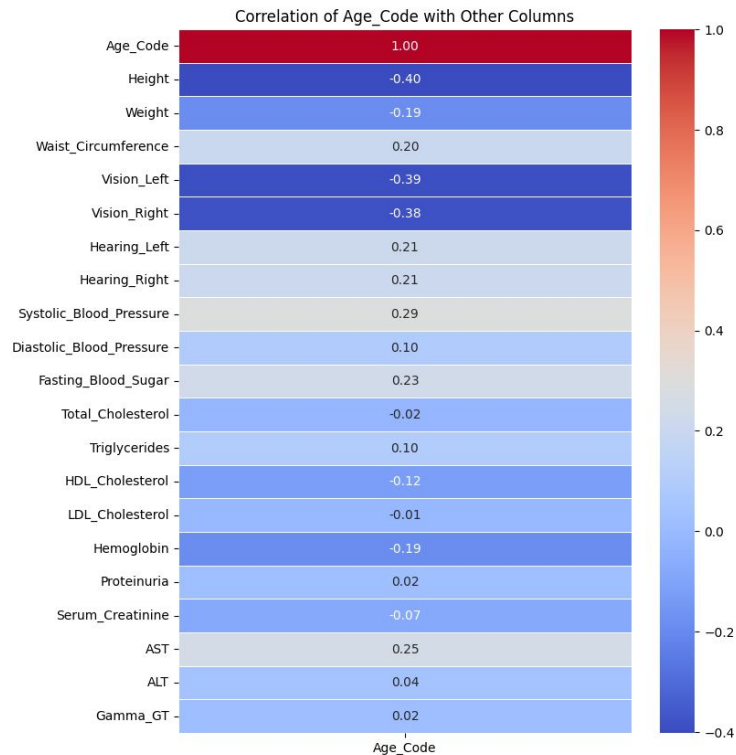
MSE and RMSE

```
[ ] # Evaluating the model
r2_value = r2_score(y_test, y_pred)
mse_value = mean_squared_error(y_test, y_pred)
rmse_value = np.sqrt(mse_value) #RMSE

print(f'R-squared Value: {r2_value}')
print(f'Mean Squared Error: {mse_value}')
print(f'Root of Mean Squared Error: {rmse_value}') #print RMSE
```

Mean Squared Error: 3.6446431769905527
Root of Mean Squared Error: 1.909094858038896

Interpretation: Correlation heatmap on Age group by other columns



**More Positively
Correlated**

**More Negatively
Correlated**

All columns are **negatively
correlated** by Age group

Interpretation: OLS Regression Results & VIF

Coefficient and p-value

OLS Regression Results						
Dep. Variable:	Age_Code	R-squared:	0.517			
Model:	OLS	Adj. R-squared:	0.517			
Method:	Least Squares	F-statistic:	8.891e+04			
Date:	Tue, 05 Dec 2023	Prob (F-statistic):	0.00			
Time:	15:51:17	Log-Likelihood:	-3.5935e+06			
No. Observations:	1744665	AIC:	7.187e+06			
Df Residuals:	1744663	BIC:	7.187e+06			
Df Model:	21					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	18.5869	0.063	296.474	0.000	18.464	18.710
Gender_Code	-1.3438	0.005	-256.236	0.000	-1.354	-1.334
Height	-0.0840	0.000	-284.062	0.000	-0.085	-0.083
Weight	-0.0966	0.000	-331.714	0.000	-0.097	-0.096
Waist_Circumference	0.1244	0.000	421.729	0.000	0.124	0.125
Vision_Left	-0.0612	0.005	-161.177	0.000	-0.071	-0.051
Vision_Right	-0.7724	0.005	-144.476	0.000	-0.783	-0.762
Hearing_Left	0.0515	0.009	93.688	0.000	0.034	0.069
Hearing_Right	0.7764	0.009	83.862	0.000	0.759	0.795
Systolic_Blood_Pressure	0.0406	0.000	260.419	0.000	0.040	0.041
Diastolic_Blood_Pressure	-0.0231	0.000	-102.393	0.000	-0.024	-0.023
Fasting_Blood_Sugar	0.0280	0.000	215.180	0.000	0.028	0.028
Total_Cholesterol	0.0099	0.001	16.511	0.000	0.009	0.011
Triglycerides	-0.0008	0.000	-6.660	0.000	-0.001	-0.001
HDL_Cholesterol	-0.0232	0.001	-37.997	0.000	-0.024	-0.022
LDL_Cholesterol	-0.0062	0.001	-13.763	0.000	-0.009	-0.007
Hemoglobin	-0.2389	0.001	-169.776	0.000	-0.242	-0.236
Proteinuria	-0.0107	0.004	-2.855	0.004	-0.019	-0.003
Serum_Creatinine	0.4999	0.005	98.203	0.000	0.484	0.504
AST	0.0968	0.000	306.361	0.000	0.096	0.097
ALT	-0.0376	0.000	-154.697	0.000	-0.038	-0.037
Gamma_GT	-0.0055	0.000	-42.022	0.000	-0.006	-0.005
Omnibus:	12458.855	Durbin-Watson:	1.949			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12855.058			
Skew:	-0.198	Prob(JB):	0.00			
Kurtosis:	3.140	Cond. No.	1.61e+04			

All less than
5% (Not a
coincidence)

“Statistically
significant”

Multicollinearity

VIF:	Variable	VIF
0	Gender_Code	3.312304
1	Height	3.582309
2	Weight	4.939496
3	Waist_Circumference	3.279338
4	Vision_Left	2.235236
5	Vision_Right	2.228883
6	Hearing_Left	1.441630
7	Hearing_Right	1.439346
8	Systolic_Blood_Pressure	2.136122
9	Diastolic_Blood_Pressure	2.031803
10	Fasting_Blood_Sugar	1.112169
11	Total_Cholesterol	224.483788
12	Triglycerides	20.304785
13	HDL_Cholesterol	31.376813
14	LDL_Cholesterol	187.696656
15	Hemoglobin	2.019023
16	Proteinuria	1.003076
17	Serum_Creatinine	1.381576
18	AST	1.824775
19	ALT	2.241380
20	Gamma_GT	1.493511

Very High

Included
only ‘Total
Cholesterol’
column

Interpretation: OLS Regression Results & VIF

OLS Regression Results

Dep. Variable:	Age_Code	R-squared:	0.512
Model:	OLS	Adj. R-squared:	0.512
Method:	Least Squares	F-statistic:	1.016e+05
Date:	Tue, 05 Dec 2023	Prob (F-statistic):	0.00
Time:	15:52:53	Log-Likelihood:	-3.6026e+06
No. Observations:	1744665	AIC:	7.205e+06
Df Residuals:	1744666	BIC:	7.206e+06
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	18.0726	0.063	287.369	0.000	17.949	18.196
Gender_Code	-1.4253	0.005	-272.832	0.000	-1.436	-1.415
Height	-0.0876	0.000	-295.830	0.000	-0.088	-0.087
Weight	-0.0947	0.000	-323.710	0.000	-0.095	-0.094
Waist_Circumference	0.1235	0.000	440.483	0.000	0.129	0.130
Vision_Left	-0.8753	0.005	-162.997	0.000	-0.886	-0.865
Vision_Right	-0.7834	0.005	-145.802	0.000	-0.794	-0.773
Hearing_Left	0.8619	0.009	94.339	0.000	0.844	0.880
Hearing_Right	0.7890	0.009	84.782	0.000	0.771	0.807
Systolic_Blood_Pressure	0.0406	0.000	259.255	0.000	0.040	0.041
Diastolic_Blood_Pressure	-0.0230	0.000	-101.533	0.000	-0.023	-0.023
Fasting_Blood_Sugar	0.0285	0.000	219.008	0.000	0.028	0.029
Total_Cholesterol	0.0007	4.11e-05	16.539	0.000	0.001	0.001
Hemoglobin	-0.2351	0.001	-166.338	0.000	-0.238	-0.232
Proteinuria	-0.0112	0.004	-2.967	0.003	-0.019	-0.004
Serum_Creatinine	0.5186	0.005	102.655	0.000	0.509	0.529
AST	0.0925	0.000	292.880	0.000	0.092	0.093
ALT	-0.0340	0.000	-140.307	0.000	-0.035	-0.034
Gamma_GT	-0.0053	0.000	-40.655	0.000	-0.006	-0.005

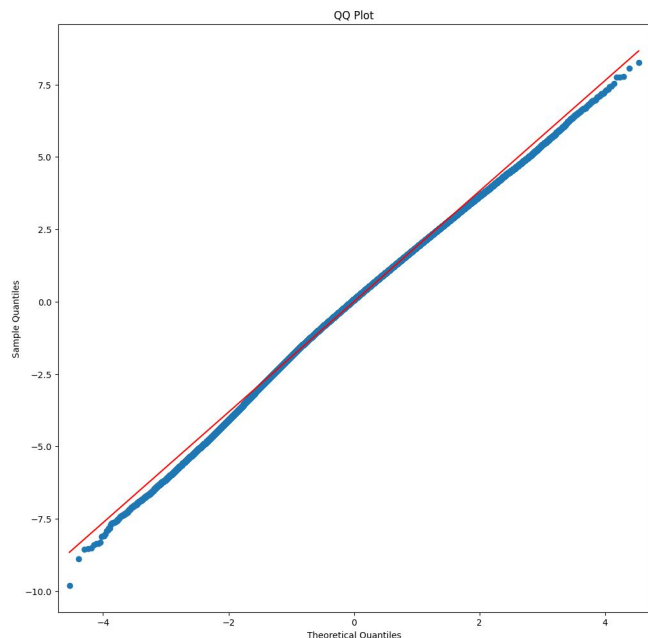
Omnibus:	13495.999	Durbin-Watson:	1.949
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13938.458
Skew:	-0.207	Prob(JB):	0.00
Kurtosis:	3.141	Cond. No.	1.42e+04

p-value
has
decreased
0.004 to
0.003

VIF:		
	Variable	VIF
0	Gender_Code	3.252247
1	Height	3.553476
2	Weight	4.927063
3	Waist_Circumference	3.222488
4	Vision_Left	2.234359
5	Vision_Right	2.228342
6	Hearing_Left	1.441493
7	Hearing_Right	1.439179
8	Systolic_Blood_Pressure	2.133992
9	Diastolic_Blood_Pressure	2.030912
10	Fasting_Blood_Sugar	1.104184
11	Total_Cholesterol	1.050850
12	Hemoglobin	2.014896
13	Proteinuria	1.002936
14	Serum_Creatinine	1.379661
15	AST	1.802978
16	ALT	2.213108
17	Gamma_GT	1.445966

Interpretation: QQ plot & R squared value

Normality test



Points lies on the line well so the model follows normal distribution

R squared value

```
[ ] #import packages to caculate R squared score
    from sklearn.metrics import r2_score

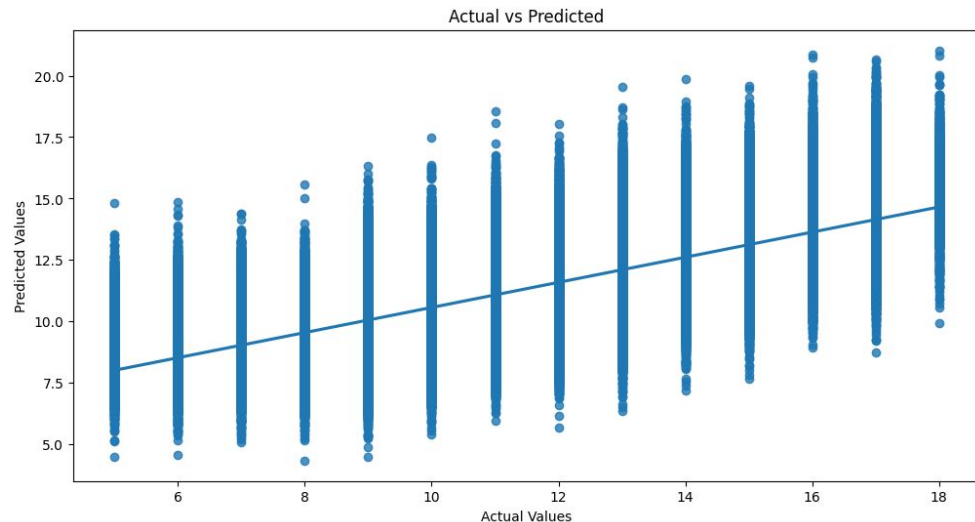
    # Calculate R squared score
    r2_value = r2_score(y_test, y_pred) #caculate R squared score
    print(f'R-squared Value: {r2_value}') #print R squared score

    # Calculate adjusted R squared
    n = len(y_test) #number of validation set
    k = X_test.shape[1] #number of predictors (number of column)
    adjusted_r2_value = 1 - ((1 - r2_value) * (n - 1) / (n - k - 1)) #
    print(f'Adjusted R-squared Value: {adjusted_r2_value}') #print adj
```

R-squared Value: 0.5125971364862545

Adjusted R-squared Value: 0.5125719923218857

Interpretation: Actual by Predicted and MSE & RMSE



MSE & RMSE

```
[ ] # Evaluating the model
r2_value = r2_score(y_test, y_pred)
mse_value = mean_squared_error(y_test, y_pred)
rmse_value = np.sqrt(mse_value) #RMSE

print(f'R-squared Value: {r2_value}')
print(f'Mean Squared Error: {mse_value}')
print(f'Root of Mean Squared Error: {rmse_value}') #print RMSE
```

Mean Squared Error: 3.6446431769905527

Root of Mean Squared Error: 1.909094858038896

Conclusion

Conclusion: We have tested model with input

Randomly selected 10 data from
DataFrame

```
# Select 10 random rows  
random_rows = merged_health_df.sample(n=10)
```

Input for model

```
gender = row['Gender_Code']  
height = row['Height']  
weight = row['Weight']  
waist_circumference = row['Waist_Circumference']  
vision_left = row['Vision_Left']  
vision_right = row['Vision_Right']  
hearing_left = row['Hearing_Left']  
hearing_right = row['Hearing_Right']  
systolic_bp = row['Systolic_Blood_Pressure']  
diastolic_bp = row['Diastolic_Blood_Pressure']  
fasting_blood_sugar = row['Fasting_Blood_Sugar']  
total_cholesterol = row['Total_Cholesterol']  
hemoglobin = row['Hemoglobin']  
proteinuria = row['Proteinuria']  
serum_creatinine = row['Serum_Creatinine']  
ast = row['AST']  
alt = row['ALT']  
gamma_gt = row['Gamma_GT']
```

Prediction of biological age

```
Row 562850:  
Actual Age Code: 9.0  
Predicted Age Code: 10.551680661378985  
-----  
Row 996948:  
Actual Age Code: 10.0  
Predicted Age Code: 10.492293119304025  
-----  
Row 730156:  
Actual Age Code: 9.0  
Predicted Age Code: 8.826687538294419  
-----  
Row 499504:  
Actual Age Code: 17.0  
Predicted Age Code: 16.22515095957945  
-----  
Row 147488:  
Actual Age Code: 11.0  
Predicted Age Code: 12.295394383004297  
-----  
Row 22551:  
Actual Age Code: 11.0  
Predicted Age Code: 11.83068099601828  
-----  
Row 895529:  
Actual Age Code: 10.0  
Predicted Age Code: 11.000846754943328  
-----  
Row 639115:  
Actual Age Code: 9.0  
Predicted Age Code: 8.951380627567692  
-----  
Row 463071:  
Actual Age Code: 9.0  
Predicted Age Code: 9.158815857815979  
-----  
Row 710238:  
Actual Age Code: 9.0  
Predicted Age Code: 10.703892792171349  
-----
```

Conclusion: Multiple linear regression model that could predict the biological age using values in health examination data

- Predict the biological age in 'Age code'
- Provide one of indicator that could make easy to understand about health using values that are hard to understand
- Discussion
 - Predict 'Age Code', not actual age
 - Make big difference (5 years for each code)
 - Have to find way to increase fitness & error
 - If we can, it would be better to get a data with actual age.

Row 562850:
Actual Age Code: 9.0
Predicted Age Code: 10.551680661378995
Row 996948:
Actual Age Code: 10.0
Predicted Age Code: 10.492293119304025
Row 730156:
Actual Age Code: 9.0
Predicted Age Code: 8.826687538294419
Row 499504:
Actual Age Code: 17.0
Predicted Age Code: 16.22515095957945
Row 147488:
Actual Age Code: 11.0
Predicted Age Code: 12.295394383004297
Row 22551:
Actual Age Code: 11.0
Predicted Age Code: 11.83068099601828
Row 895529:
Actual Age Code: 10.0
Predicted Age Code: 11.000846754943328
Row 639115:
Actual Age Code: 9.0
Predicted Age Code: 8.951380627567692
Row 463071:
Actual Age Code: 9.0
Predicted Age Code: 9.158815857815979
Row 710238:
Actual Age Code: 9.0
Predicted Age Code: 10.703892792171349

Age Code	Age Group
1	0 ~ 4
2	5 ~ 9
3	10 ~ 14
4	15 ~ 19
5	20 ~ 24
6	25 ~ 29
7	30 ~ 34
8	35 ~ 39
9	40 ~ 44

Age Code	Age Group
10	45 ~ 49
11	50 ~ 54
12	55 ~ 59
13	60 ~ 64
14	65 ~ 69
15	70 ~ 74
16	75 ~ 79
17	80 ~ 84
18	85 +

Reference

- https://health.chosun.com/site/data/html_dir/2016/11/30/2016113001291.html
- <https://m.khan.co.kr/opinion/yeojeok/article/202303052038005#c2b>
- <https://ebn.co.kr/news/view/1027995>
- <https://www.cuemath.com/outlier-formula/>

Thank you