Move AI

일상생활 속 작은 움직임

2팀 : 최규철, 윤지현, 한시호, 정서영

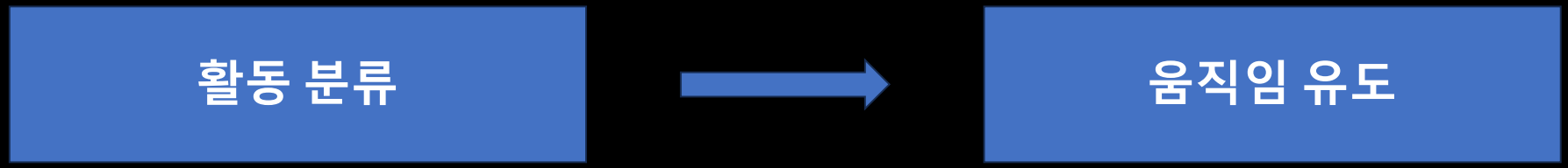
타임라인

날짜	시간	단계	상세 활동 내용
12.15 (월)	12:20 ~ 12:30	데이터 탐색	팀원 각자 데이터 구조 및 특성 탐색
	13:30 ~ 13:40	방향 설정	프로젝트 진행 목적 및 목표(Why) 구체화
	13:40 ~ 14:00	전처리	결측치(Missing Value) 확인 및 데이터 클렌징
	14:00 ~ 15:30	모델링	Clustering 진행, Elbow method로 K값 탐색 및 성능 측정
	15:30 ~ 16:30	논의	성능 지표 바탕 적정 K값 선정 관련 팀 논의
12.16 (화)	09:00 ~ 10:30	의사결정	최종 K값 결정 및 확정
	10:30 ~ 12:30	시각화	확정된 Cluster 시각화 구현
	13:30 ~ 15:00	재분석	Random Forest 지표 확인 및 기존 자료 재분석
	15:00 ~ 15:10	배포	분석 모델의 HTML 변환
	15:10 ~ 18:00	문서화	PPT 제작, 수정 및 발표 리허설/구역 설정

Move AI ?

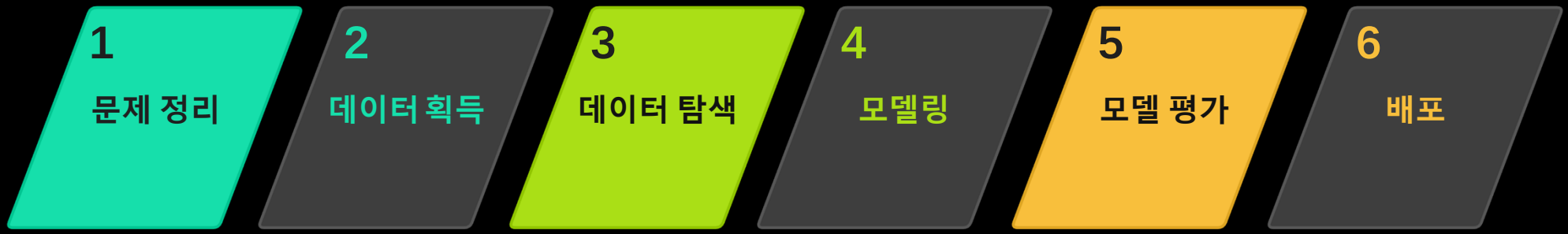
일상 생활 중 활동을 유도하는 AI

사용자의 비활동적인 시간(앉아있거나 누워있는 시간 등)을 파악하고, 활동을 하도록 안내하는 AI



1. 웨어러블 기기 착용자의 활동을 앉아있음, 누워있음 등으로 분류
2. 장시간 활동이 없을 경우, "가장 가까운 창가까지 걸어갔다 오세요"와 같은 **작고 실현 가능한** 활동을 제안

AI 프로젝트 사이클



1. 문제 정리 - UN SDG Goal 3 Target 4

비전염성 질병(NCDs) 현황

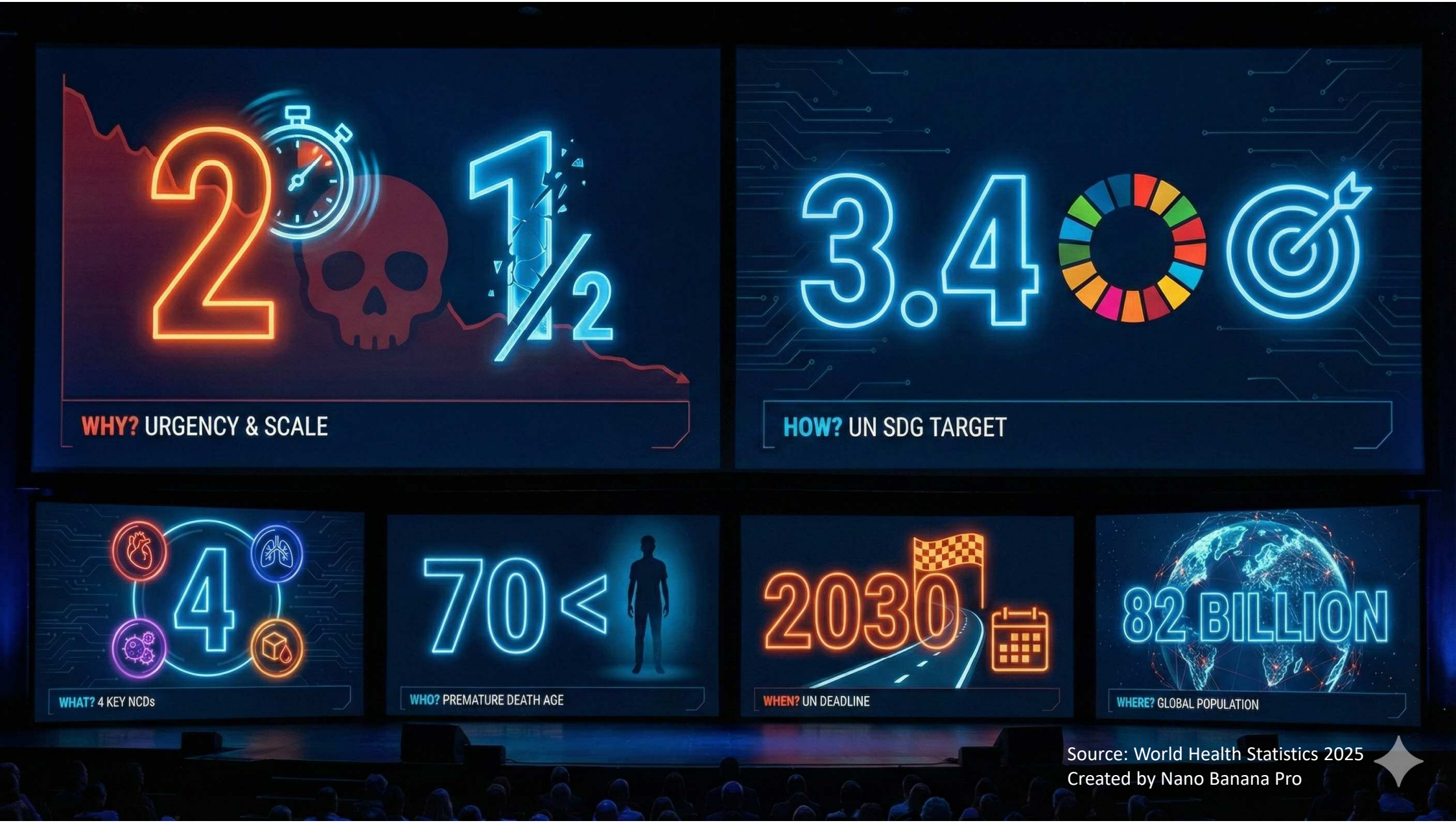
- **2021년 기준**
- 전 세계 **70세 미만 인구 중 약 1,800만명**
- 비전염성 질병(NCDs)으로 사망
- **해당 연령대 사망자의 절반 이상**을 차지

4대 비전염성 질병 (NCDs)

- 심혈관 질환
- 암
- 만성 호흡기 질환
- 당뇨병
- 2015년 이후 조기 사망 위험은 감소
- but **2030년 NCDs 감소 목표 달성 어려운 상황**

비만과 NCDs의 관계

- 비만: **4대 NCDs 목록에 직접 포함 X**
- but **모든 주요 NCDs와 밀접한 관련**
- WHO는 **NCDs의 핵심 위험 요인(Risk Factor)**으로 규정



Source: World Health Statistics 2025
Created by Nano Banana Pro

2. 데이터 획득

웨어러블 기기 이용

분류

activity = 사람이 수행한 동작 (STANDING, WALKING, SITTING 등)

데이터

가속도계(accelerometer)·자이로스코프(gyroscope) 등의 신호에서 추출

3. 데이터 탐색 - 데이터 정의

1	**tBodyAcc / tGravityAcc / tBodyAccJerk / tBodyGyro / tBodyGyroJerk** XYZ축 / 통계량(mean, std 등)
2	**tBodyAccMag / tGravityAccMag / tBodyAccJerkMag / tBodyGyroMag** 통계량(mean, std 등)
3	**fBodyAcc / fBodyAccJerk / fBodyGyro** XYZ축 / 통계량(mean, std 등) / bandsEnergy
4	**fBodyAccMag / fBodyBodyAccJerkMag / fBodyBodyGyroMag / fBodyBodyGyroJerkMag** 통계량(mean, std 등)
5	**angle.tBody / angle.XYZ** GravityMean

3. 데이터 탐색 - 데이터 전처리

```
df.head(10)
```

	rn	activity	tBodyAcc.mean.X	tBodyAcc.mean.Y	tBodyAcc.mean.Z	tBodyAcc.std.X	tBodyAcc.std.Y	tBodyAcc.std.Z	tBodyAcc.mad.X	tBodyAcc.mad.Y	...	fBody
0	7	STANDING	0.2790	-0.0196	-0.1100	-0.997	-0.967	-0.983	-0.997	-0.966	...	
1	11	STANDING	0.2770	-0.0127	-0.1030	-0.995	-0.973	-0.985	-0.996	-0.974	...	
2	14	STANDING	0.2770	-0.0147	-0.1070	-0.999	-0.991	-0.993	-0.999	-0.991	...	
3	15	STANDING	0.2980	0.0271	-0.0617	-0.989	-0.817	-0.902	-0.989	-0.794	...	
4	20	STANDING	0.2760	-0.0170	-0.1110	-0.998	-0.991	-0.998	-0.998	-0.989	...	
5	21	STANDING	0.2780	-0.0143	-0.1080	-0.998	-0.994	-0.996	-0.998	-0.994	...	
6	22	STANDING	0.2770	-0.0180	-0.1070	-0.998	-0.990	-0.997	-0.998	-0.990	...	
7	24	STANDING	0.2790	-0.0177	-0.1090	-0.998	-0.987	-0.991	-0.999	-0.987	...	
8	31	SITTING	0.2220	0.0341	-0.1240	-0.815	-0.749	-0.572	-0.879	-0.777	...	
9	32	SITTING	-0.0417	0.1750	0.0256	-0.758	-0.587	-0.439	-0.774	-0.555	...	

10 rows × 563 columns

'rn', 'activity' 열 삭제

- 'rn' = 인덱스열
- Activity = Target

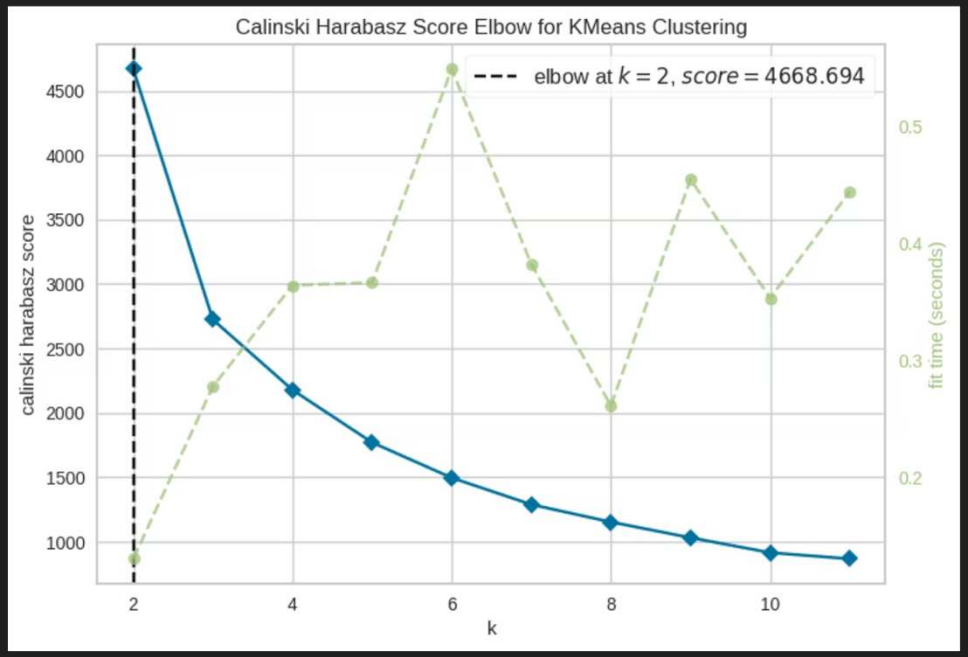
결측치 없음

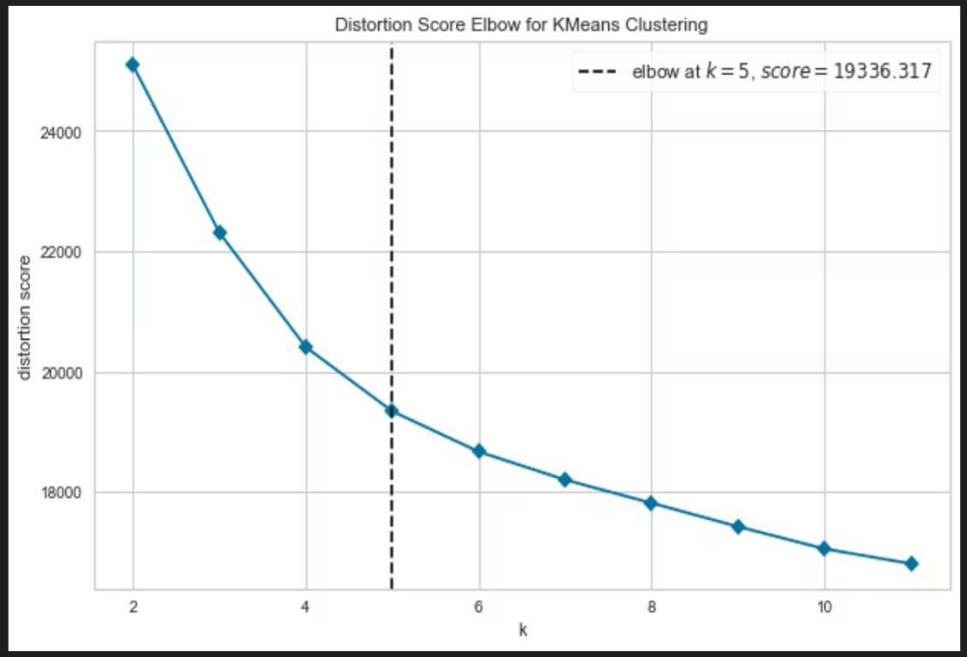
- 결측치에 대한 처리 X

데이터 정규화가 이미 완료됨

- 데이터 분포가 -1과 1 사이

4. 모델링 - Elbow Methods를 이용한 k값 탐색

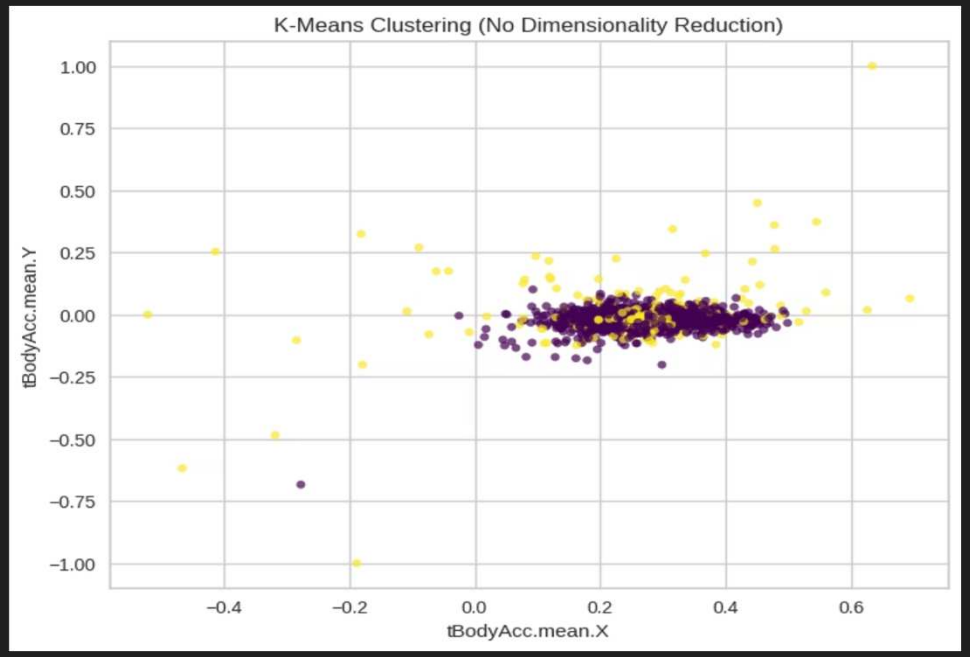


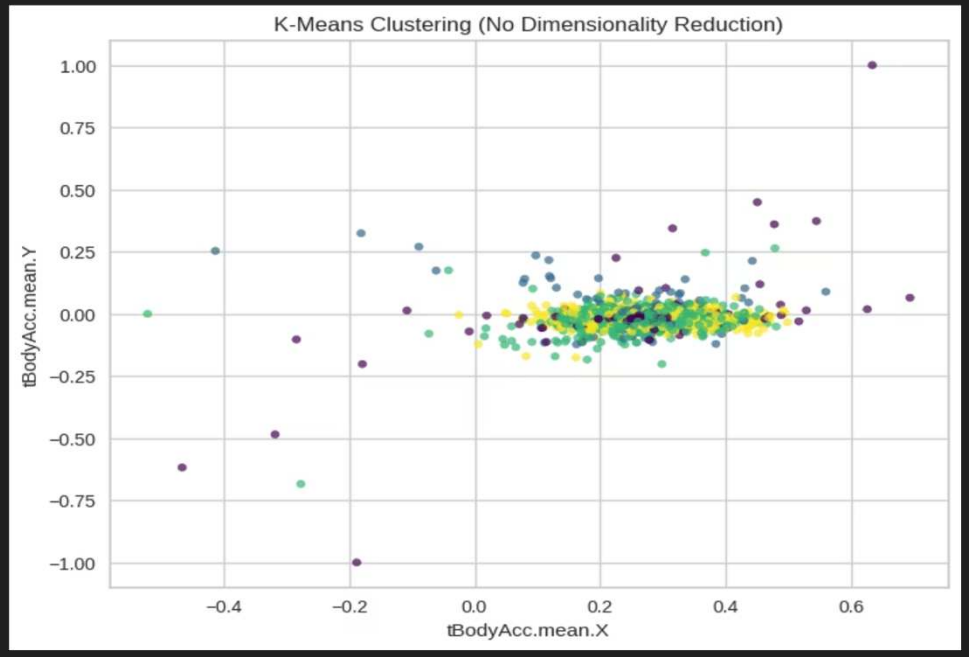


Elbow Method에 따른 적절한 k값은 2 , 5

- 정적인 행동과 동적인 행동으로 나눈 k=2일 때가 가장 클러스터링이 잘 될 것으로 예측
- k=5 이후부터 그래프가 완만해짐
- Silhouette 등의 다른 지표와 교차 검증 후 적절한 k값을 도출

4. 모델링 - 클러스터링 (차원 축소 X)

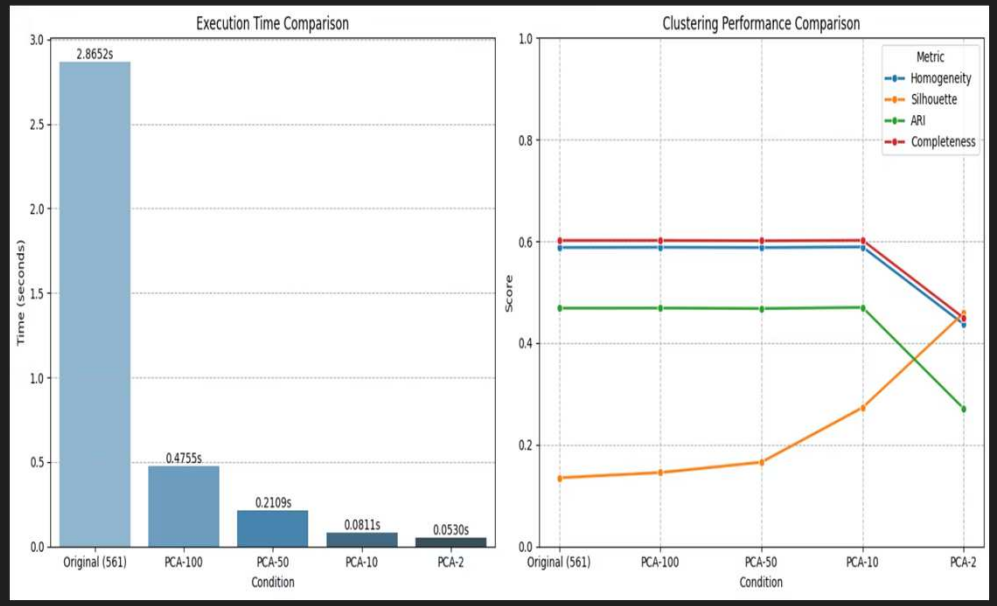


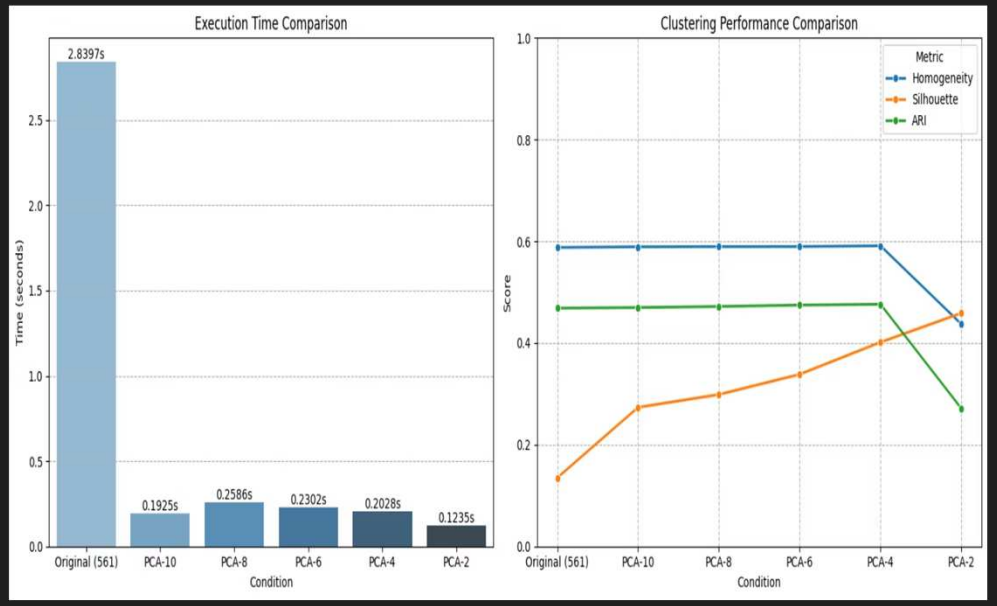


K-MEANS 클러스터링

- k=2일 때, 정적인 활동(앉기, 눕기)과 동적인 활동(걷기, 계단)으로 분류
- k=4일 때, 단순히 움직임 유무를 넘어 더 세밀한 패턴이 드러남
- 단 두 개의 변수로는 데이터 구분이 불가능하여, 차원 축소(PCA, t-SNE)가 필수적임을 보여주는 이미지

4. 모델링 - 차원의 개수에 따른 학습 시간 및 지표 비교

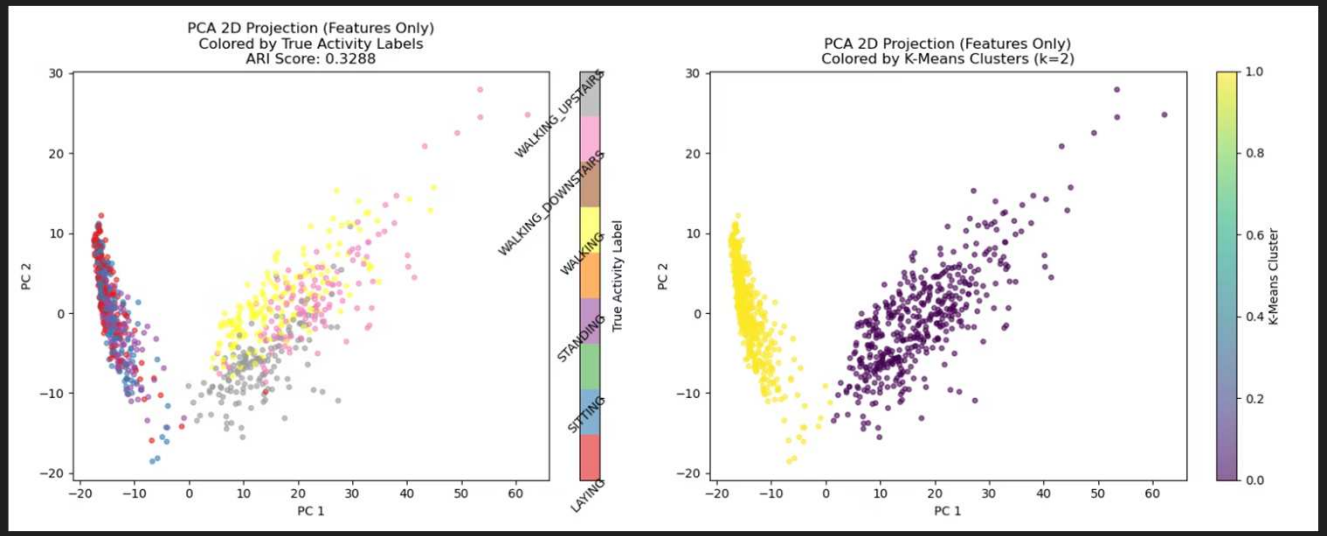




적절한 차원의 개수는?

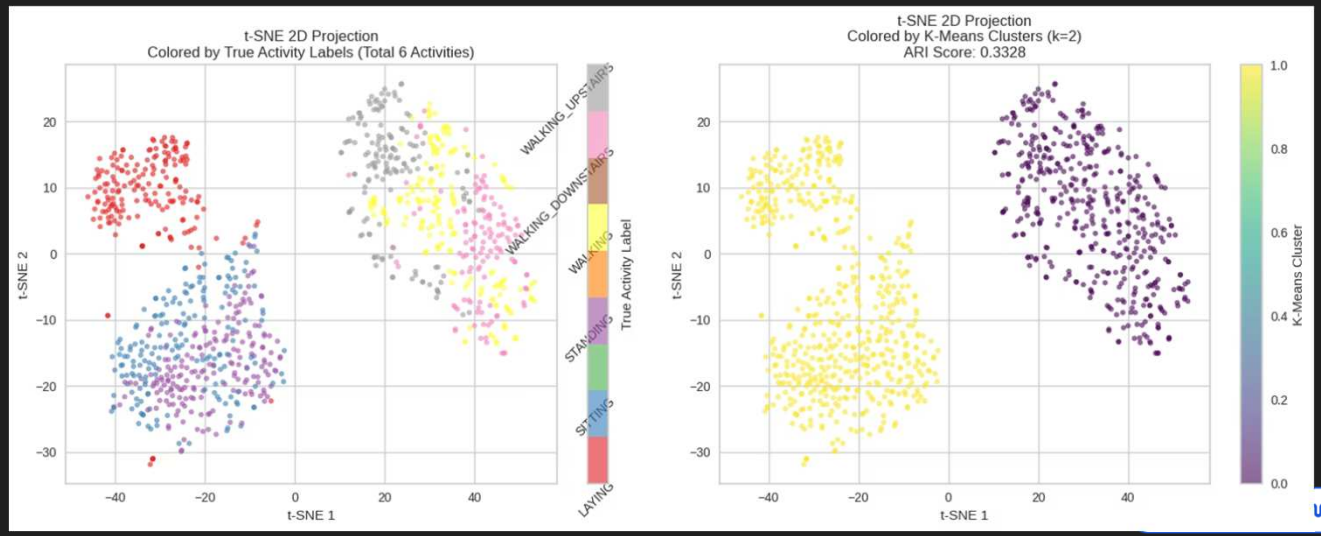
- 차원이 적을수록 학습 시간은 줄어든다.
- 50차원으로 축소했을 때 성능 지표의 차이는 미미하지만, 학습 소요 시간은 대폭 단축

4. 모델링 – 차원축소 후 실제 데이터와 비교(K=2)

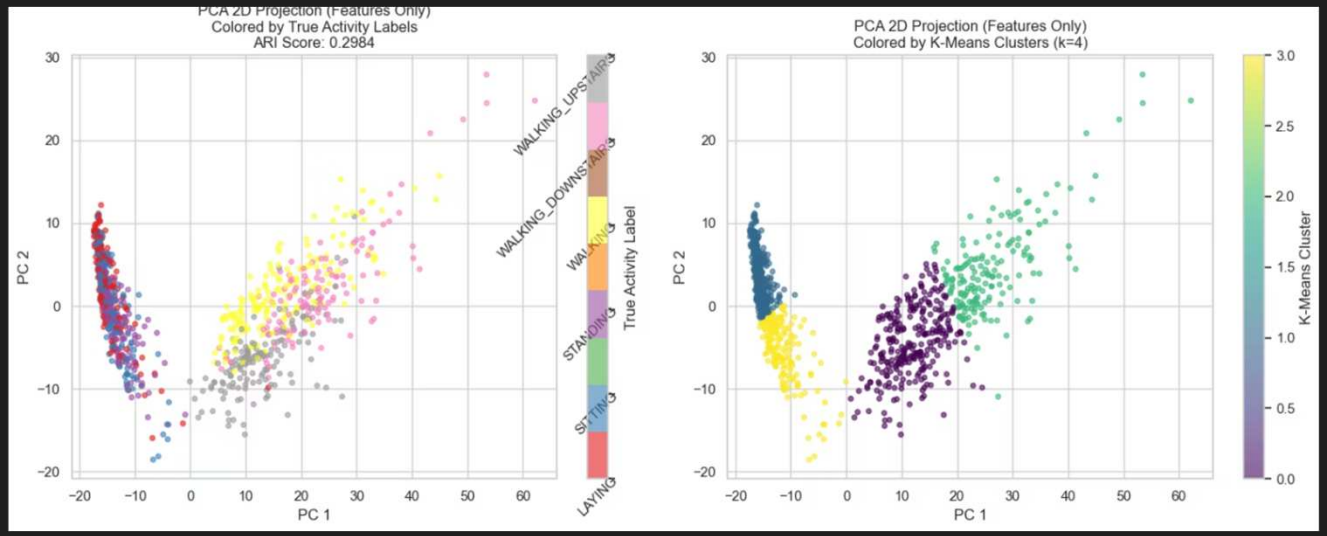


PCA

t-SNE

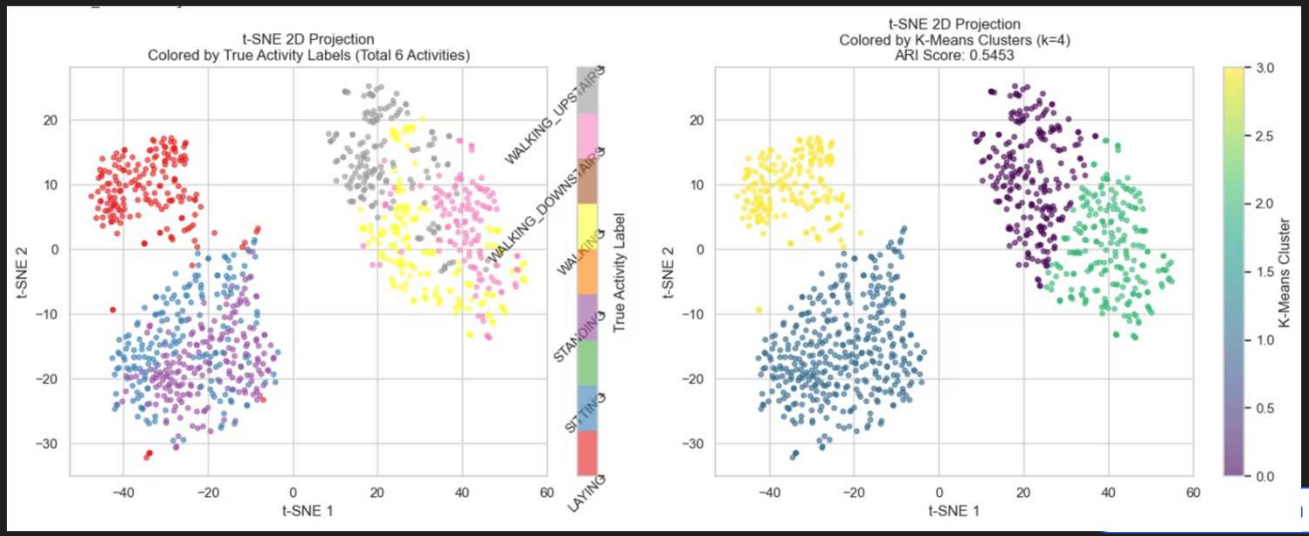


4. 모델링 – 차원축소 후 실제 데이터와 비교(K=4)

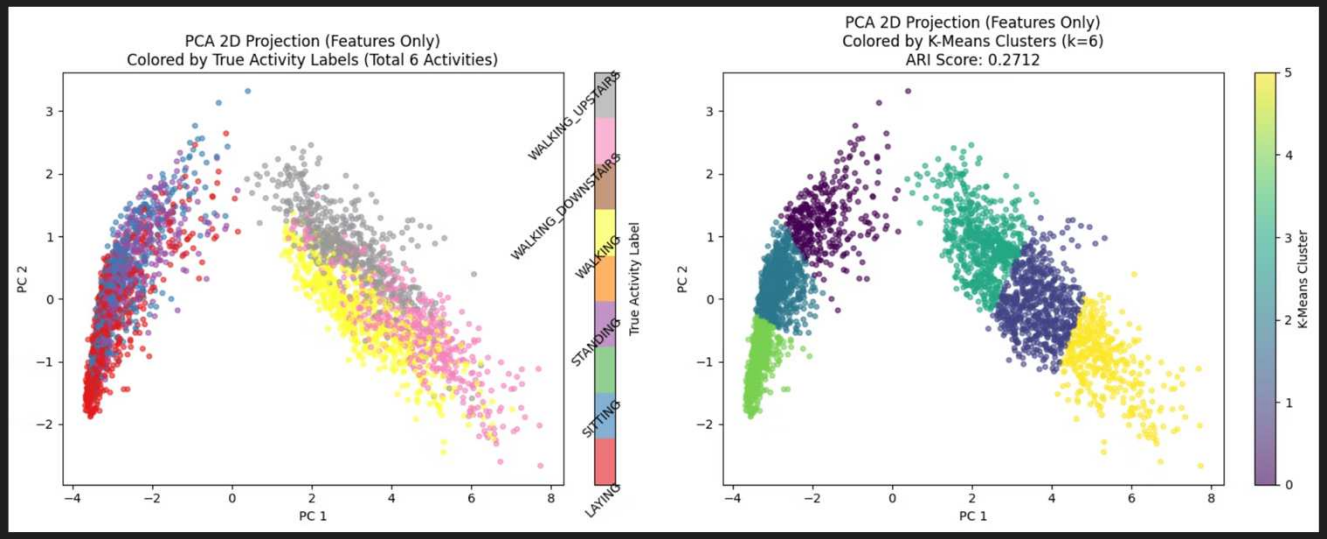


PCA

t-SNE

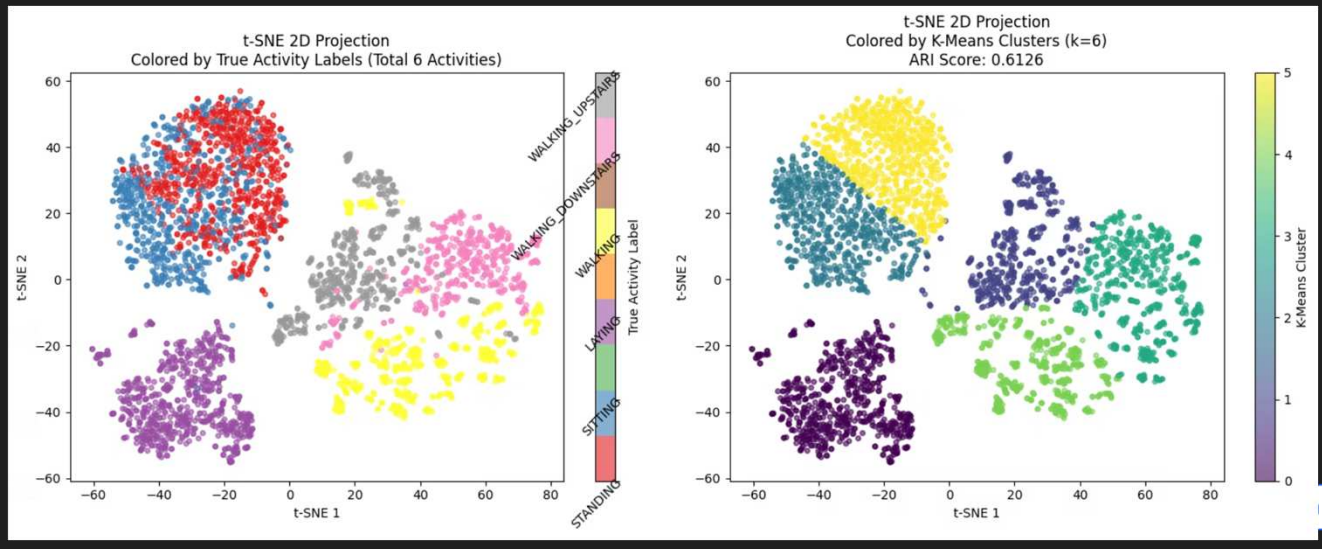


4. 모델링 – 차원축소 후 실제 데이터와 비교(K=6)



PCA

t-SNE



5. 모델 평가 - 지표의 의미

Inertia

- 클러스터 내부 데이터들이 해당 중심점과 얼마나 가까운지를 합한 값
- **작을수록 좋으며**, 클러스터 수가 많아질수록 낮아지는 경향

Homogeneity

- 각 클러스터가 하나의 레이블(정답 클래스)만 포함하는 정도
- **1 → 모든 클러스터가 단일 클래스만 포함**
- **0** → 클러스터 안에 클래스가 뒤섞임

Completeness

- 같은 레이블에 속하는 모든 데이터가 같은 클러스터에 잘 모였는지를 평가
- **1 → 하나의 클래스가 하나의 클러스터에만 존재**
- **0** → 하나의 클래스가 여러 클러스터에 분산
- 클러스터의 수가 작을수록 높아지는 경향

V_measure

- homogeneity와 completeness의 조화 평균
- 클러스터는 순수하면서도 동시에 클래스가 잘 모여 있는가?
- **1 → 완벽한 군집**
- **0** → 완전히 무의미한 군집

5. 모델 평가 - 지표의 의미

Adjusted Rand Index (ARI)

- 클러스터링 결과와 정답 라벨 간의 유사도를 보는 지표
- **1에 가까울 수록 좋음**

Adjusted Mutual Information (AMI)

- 클러스터와 정답 라벨 간의 상호정보량을 기준으로 한 지표
- **1에 가까울 수록 좋음**

Silhouette

- 각 데이터가 자신의 클러스터에 잘 속해 있는지
- **1에 가까울 수록 좋음**

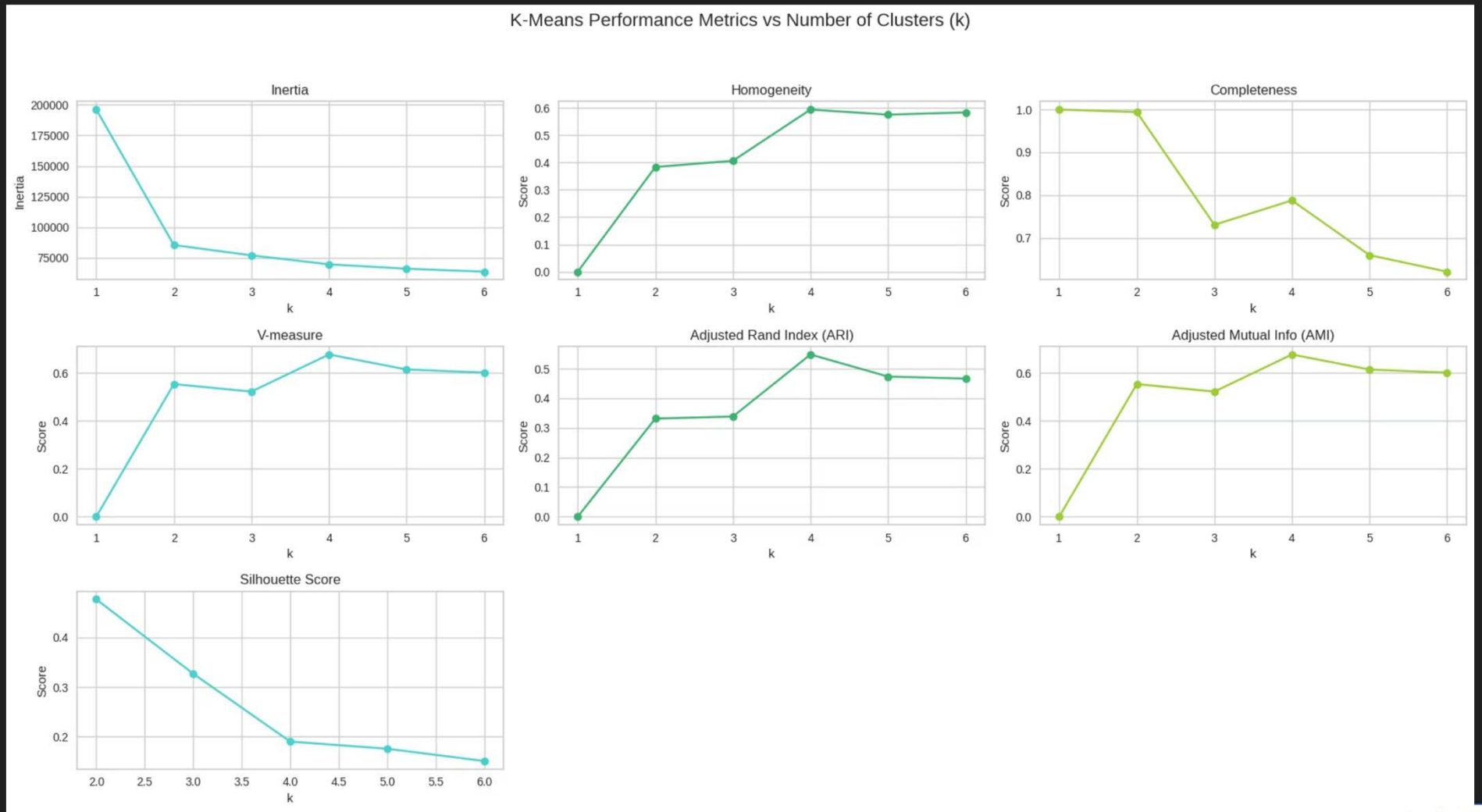
5. 모델 평가 - 지표 분석 (차원 축소 x)

k	inertia	homogeneity	completeness	v_measure	ARI	AMI	silhouette
2	**85475.325967**	**0.383527**	**0.993985**	0.553490	**0.331632**	0.553240	**0.477162**
3	77062.489233	0.406291	0.730459	**0.522154**	0.338558	**0.521676**	0.327017
4	69728.589007	**0.593987**	0.787508	**0.677193**	**0.548489**	**0.676763**	0.190173
5	66150.017443	0.575360	0.659479	0.614554	0.473725	0.613913	0.175676
6	**63732.067058**	0.583456	**0.620791**	0.601545	0.467087	0.600744	**0.150974**

K = 4

- **ARI, V-measure, AMI** 등 주요 분류 평가지표가 모두 **최고점**을 기록
- K=4일 때 데이터 간의 동질성(Homogeneity)이 급격히 상승하며, 모델이 정답 레이블의 구조를 가장 신뢰도 있게 재현
- 군집 응집도가 떨어지는 K=6이나 단순한 K=2보다, 통계적 성능이 극대화되는 **K=4를 최종 모델로 선정**하는 것이 가장 이상적↑

5. 모델 평가 – 그래프 시각화



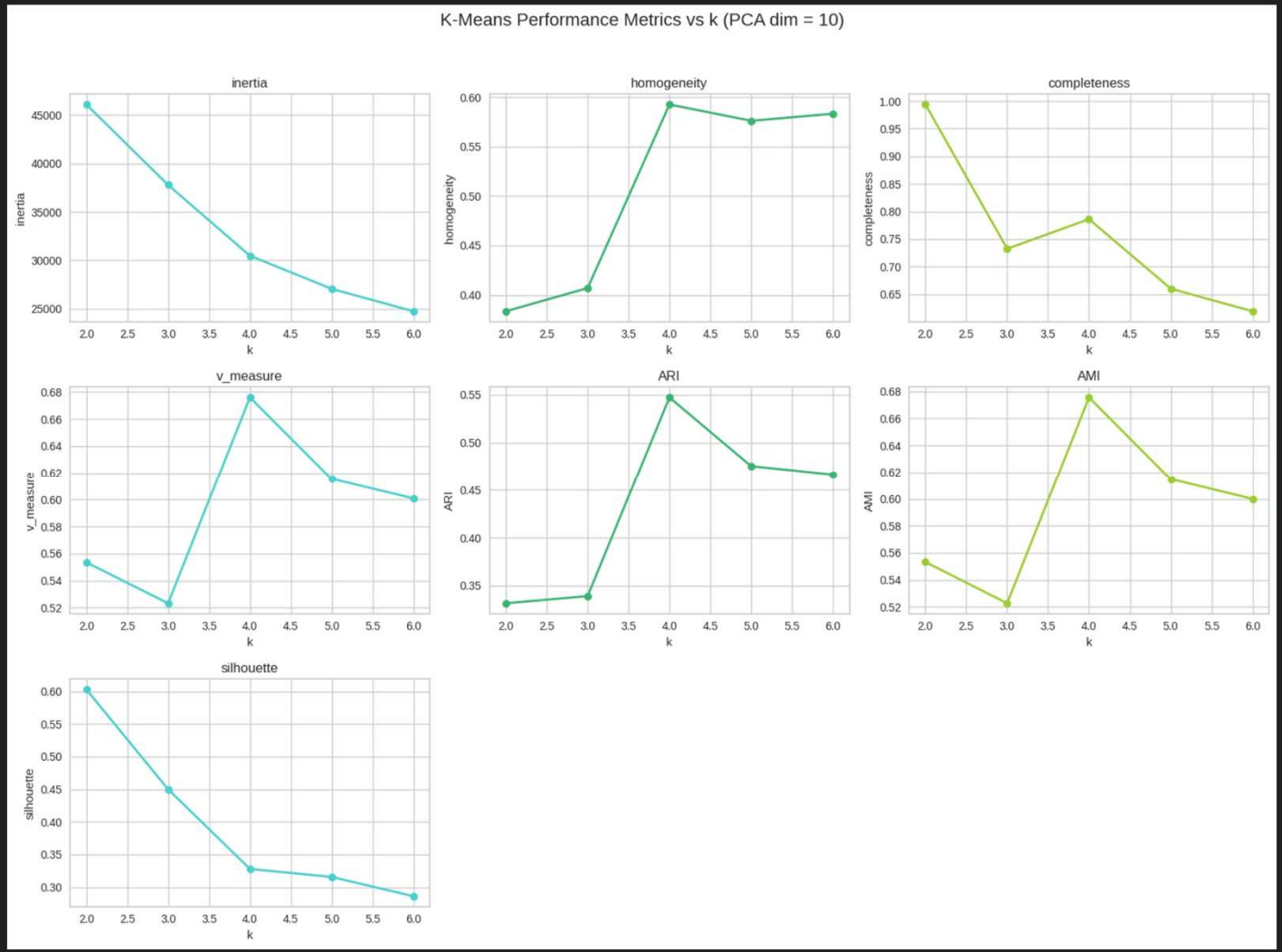
5. 모델 평가 - 지표 분석 (PCA – 10dim)

k	inertia	homogeneity	completeness	v_measure	ARI	AMI	silhouette
2	**46090.328790**	**0.383527**	**0.993985**	0.553490	**0.331632**	**0.553240**	**0.602920**
3	37787.611280	0.406955	0.732476	**0.523217**	0.339161	0.522740	0.449976
4	30481.176099	**0.592726**	0.786185	**0.675885**	**0.547271**	**0.675453**	0.328025
5	27059.203091	0.576206	0.660123	0.615316	0.475416	0.614676	0.315687
6	**24748.931986**	0.583364	**0.619211**	0.600753	0.466622	0.599952	**0.286213**

K = 4

- 실제 정답과의 일치도를 나타내는 **ARI 점수가 0.54로 가장 높게 측정**
- 실제 행동 개수인 6개로 설정할 경우, **유사한 행동(앉기/서기)의 모호한 경계** 때문에 오히려 분류 정확도가 0.50으로 하락
- 무리하게 6개로 세분화하는 것보다, 4개의 세분화된 패턴으로 묶는 것이 통계적으로 가장 신뢰성 ↑

5. 모델 평가 - 그래프 시각화 / PCA



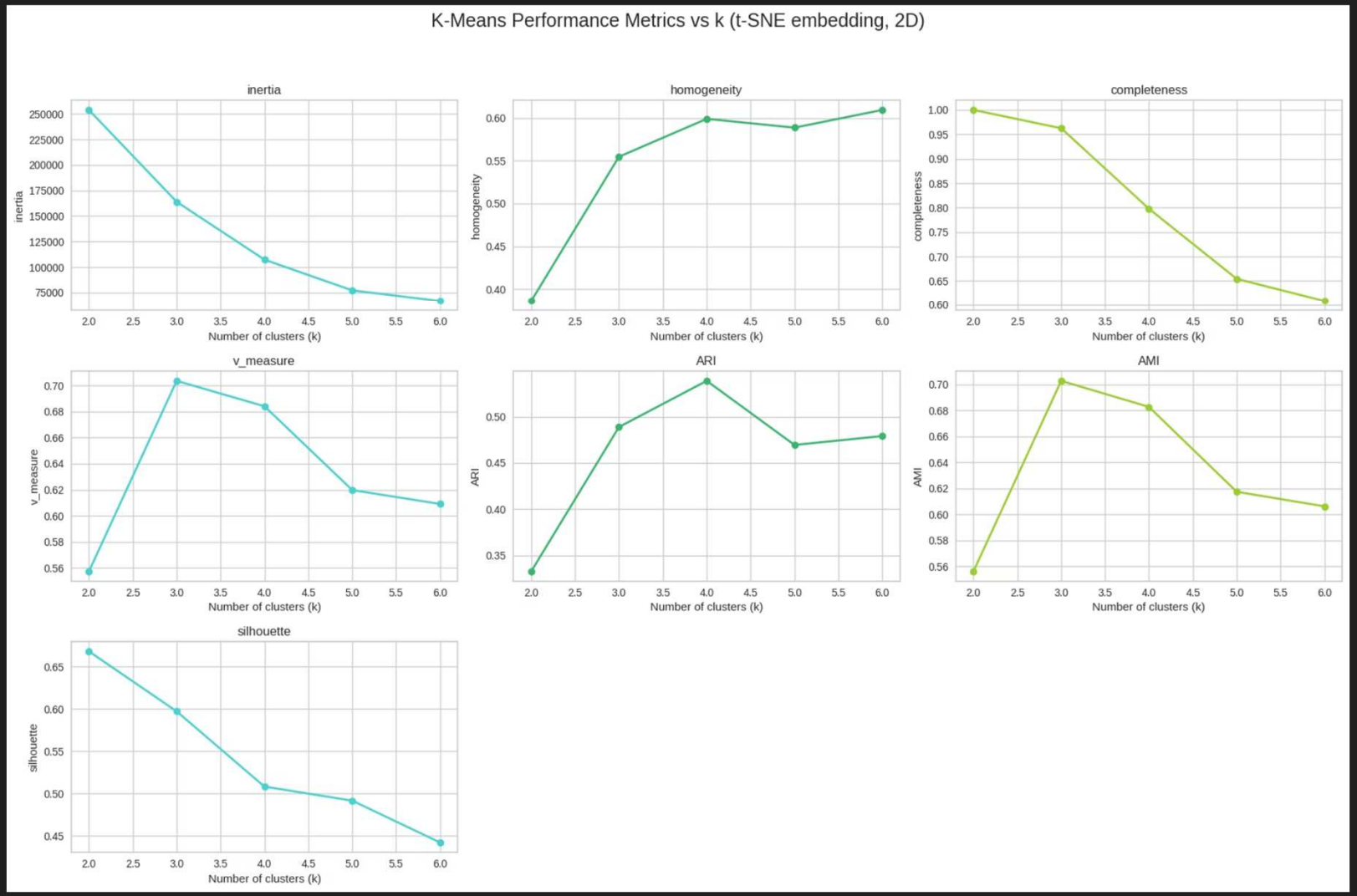
5. 모델 평가 - 지표 분석 (t-SNE)

k	inertia	homogeneity	completeness	v_measure	ARI	AMI	silhouette
2	**253899.50000**	**0.386087**	**1.000000**	0.557089	**0.332751**	**0.556187**	**0.668305**
3	163958.65625	0.554643	0.962499	**0.703748**	0.488767	**0.702684**	0.597231
4	107360.71875	0.598903	0.797800	0.684190	**0.538671**	0.682654	0.508167
5	77241.15625	0.588597	0.654300	0.619712	0.469439	0.617435	0.491534
6	**66746.40625**	**0.609317**	**0.608803**	**0.609060**	0.479050	0.606275	**0.441985**

K = 4

- 실제 정답은 6개지만, 데이터 분석 결과 4개로 분류했을 때 정답지와 가장 높은 일치율(ARI 1위)을 보임
- K=6으로 설정하면 완전성 점수가 급락. 이는 하나의 행동(예: 걷기)을 억지로 여러 그룹으로 찢어놓는 '과잉 분할' 오류가 발생
- 따라서 데이터를 불필요하게 분산시키지 않으면서도 가장 명확하게 패턴을 설명하는 통계적 최적값은 K=4

5. 모델 평가 - 그래프 시각화 / t - SNE



5. 모델 평가 — fit 시간 비교

	Data Type	Total Fit Time (sec)	Average Time per Fit (sec)
0	Raw Features	7.665676	0.153314
1	PCA (10-dim)	0.355113	0.007102
2	PCA (50-dim)	0.441870	0.008837
3	PCA (100-dim)	0.562406	0.011248

fit 50회 진행

- 원본 데이터의 fit 시간이 가장 오래 걸림

5. 모델 평가

적절한 차원의 개수는?

- 50차원에서 원본 데이터와 유사하면서 시간이 단축됨

적절한 차원 축소 방법은?

- PCA 방식은 머신러닝 모델에 적절함
- t-SNE 방식은 시각화 모델에 적절함

적절한 K의 값은?

- K=2일 때 군집화 성능이 좋으므로, 활동/비활동으로 나눌 때는 K=2인 모델 사용
- PCA 방식에서는 K=4일 때가 최적의 모델
- t-SNE 방식에서는 K=3,4일 때가 최적의 모델

번외 - *Random Forest*

활동	정밀도 (Precision)	재현율 (Recall)	F1-점수 (F1-Score)	지원 (Support)
LAYING (눕기)	1.0000	1.0000	1.0000	136
SITTING (앉기)	0.9593	0.9440	0.9516	125
STANDING (서기)	0.9485	0.9627	0.9556	134
WALKING (걷기)	0.9754	0.9835	0.9794	121
WALKING_DOWNSTAIRS (계단 내려가기)	0.9485	0.9388	0.9436	98
WALKING_UPSTAIRS (계단 올라가기)	0.9444	0.9444	0.9444	108
전체 정확도 (Accuracy)	0.9640	0.9640	0.9640	1
Macro Avg	0.9627	0.9622	0.9624	722
Weighted Avg	0.9640	0.9640	0.9640	722

지도 학습 알고리즘

Random Forest는 K-Means와 달리
지도 학습 알고리즘에 속함

96% 정확도

높은 정확도로 행동을 분류

K-Means와 결합하여 활용 가능성

높은 정확도를 가진 만큼
K-Means 모델과 비교하여 활용할 가능성

6. 배포

Introducing Move AI 1.0.0