QUESTION:

WHY AGE ISN’T AFFECTED

AND WHY IS CLUSTER STILL EMPTY AND NOT RELATED

Key Concepts for Interpreting the Correlation Matrix:

# - Correlation Coefficient (r): This is a number between -1 and 1 that quantifies the relationship between two variables.

# - r = 1: Perfect positive correlation. As one variable increases, the other increases proportionally.

# - r = -1: Perfect negative correlation. As one variable increases, the other decreases proportionally.

# - r = 0: No correlation. The variables are independent of each other.

# - Heatmap Color:

# - Red/Orange: Positive correlation.

# - Blue: Negative correlation.

# - White/Gray: Near zero, indicating weak or no correlation.

# Interpreting the Matrix:

# 1. Diagonal Elements (r = 1):

# - The diagonal values in the matrix are all 1, which is expected because any feature is perfectly correlated with itself.

# 2. Age and Other Features:

# - Age vs. Other Features:

# - The empty correlations suggest that age either has no significant variation or doesn’t strongly correlate with other features.

# - Low correlations with weight (-0.10) and height (-0.48), meaning age does not have a significant linear relationship with these features in this dataset.

# 3. Height and Weight (0.71):

# - There is a strong positive correlation between height and weight, which makes intuitive sense. Generally, taller people tend to have a higher weight.

# 4. Steps, Heart Rate, and Calories:

# - Applewatch.Steps\_LE vs. Applewatch.Calories\_LE (0.22): A positive correlation is expected since more steps generally lead to more calories burned,

# but the correlation isn't very strong, suggesting other factors (like intensity or duration) might also play a role in calorie burn.

# - Applewatch.Steps\_LE vs. Applewatch.Heart\_LE (0.03): This weak correlation suggests that steps and heart rate aren't strongly related.

# This could be because people might have a high heart rate without taking many steps (e.g., during stationary exercises).

# - Applewatch.Calories\_LE vs. Applewatch.Distance\_LE (0.66): This is a relatively strong correlation. More distance usually corresponds to more calories burned, as expected.

# 5. Heart Rate and Its Related Features:

# - Applewatch.Heart\_LE vs. RestingApplewatchHeartrate\_LE (0.44): This moderate positive correlation shows that higher average heart rates correspond somewhat

# to higher resting heart rates, but they are distinct metrics.

# - Applewatch.Heart\_LE vs. NormalizedApplewatchHeartrate\_LE (0.90): This very strong correlation indicates that these two features are highly related,

# likely because one is derived from the other or normalized based on similar data.

# - ApplewatchIntensity\_LE vs. Applewatch.Heart\_LE (0.80): A high correlation is expected since intensity should correlate strongly with heart rate.

# 6. Steps and Distance:

# - Applewatch.StepsXDistance\_LE vs. Applewatch.Steps\_LE (0.70) and vs. Applewatch.Distance\_LE (0.90):

# As expected, this feature, which is likely a product of steps and distance, shows strong correlations with both of its constituent parts.

# 7. Entropy Features:

# - EntropyApplewatchStepsPerDay\_LE vs. Steps and Distance:

# - The entropy of steps per day correlates negatively with steps (-0.36) and distance (-0.36).

# Entropy likely captures variability or irregularity in daily activity. So, people who consistently take a similar number of steps every day

# may have low entropy, while those whose step count varies more may have higher entropy.

# - EntropyApplewatchHeartPerDay\_LE vs. Heart Rate (-0.11): This negative correlation indicates that the more regular a person’s heart rate is (low entropy),

# the less their overall heart rate fluctuates.

# 8. The Cluster Column:

# - cluster vs. Other Features: The correlation between the cluster (the cluster assignment from K-Means) and the other features is relatively low,

# suggesting that the clustering was not strongly dependent on any single feature.

# - Negative correlation with Applewatch.Calories\_LE (-0.47) and Distance (-0.47): This indicates that certain clusters may be characterized by

# lower calorie burn and shorter distances, potentially identifying less active groups.

# Insights and Next Steps:

# 1. Focus on Strong Correlations:

# - You can focus your analysis or model-building efforts on features that are strongly correlated, like height and weight, steps and distance,

# and heart rate-related features (Applewatch.Heart\_LE and NormalizedApplewatchHeartrate\_LE).

# 2. Weak or No Correlations: THIS CAN HELP ELIMINATE FEATURES

# - Features with weak or no correlations (e.g., age, steps, and heart rate) suggest that they might not be significant predictors in certain models or analyses.

# Consider dropping these or exploring non-linear relationships.

# 3. Feature Engineering:

# - You might want to further explore the entropy features. Since they show moderate correlations with key metrics like steps and heart rate,

# they might add valuable insights when predicting activity levels.

# 4. Clustering:

# - The cluster variable's low correlations with most features suggest that the current features may not be enough to create distinct, well-separated clusters.

# You might consider adding more features or trying different clustering methods (e.g., DBSCAN or Hierarchical Clustering).

# Conclusion:

# - This correlation matrix provides a good overview of the relationships between the different features in your dataset.

# - Strong correlations (e.g., between calories, steps, and distance) make intuitive sense and align with expectations.

# - Weak correlations indicate areas where features might not be as predictive of one another.

# By focusing on the most important features (based on their correlations) and possibly removing or down-weighting less important ones,

# you can refine your model or analysis for better results.

**### Why Focus on Strong Correlations? next steps for feature selection**

1. \*\*Predictive Power\*\*:

- Features that are strongly correlated often provide similar or related information about the dataset. In machine learning, these features can be \*\*strong predictors\*\* for models if your target variable is influenced by them.

- For example, in a model predicting activity levels, features like \*\*steps\*\* and \*\*distance\*\* are strongly correlated, meaning that one feature can help predict the other, which in turn is likely related to overall activity.

2. \*\*Redundancy in Features\*\*:

- When two features are highly correlated, they may contain \*\*redundant information\*\*. Including both features in a model can sometimes lead to multicollinearity, which can negatively impact certain types of models (like linear regression) by making it difficult to interpret the importance of each feature.

- To handle this, you can either combine correlated features into a \*\*single feature\*\* or prioritize one feature over the other based on context or domain knowledge.

3. \*\*Interpretability and Efficiency\*\*:

- Reducing the number of features by focusing on highly correlated variables makes models more interpretable and computationally efficient. Instead of using a large number of loosely correlated features, focusing on a smaller subset of highly correlated features allows the model to learn more efficiently.

### Strongly Correlated Features in Your Dataset

#### 1. \*\*Height and Weight\*\* (Correlation: \*\*0.71\*\*)

- \*\*Interpretation\*\*: There’s a strong positive correlation between \*\*height\*\* and \*\*weight\*\*. This is expected since taller individuals tend to weigh more, given normal variation.

- \*\*How to Use It\*\*:

- If your goal is to predict a person’s activity level or health metrics, you could include either height or weight (or a combination of both) to represent the physical characteristics of the user.

- However, since both features provide related information, you could consider \*\*normalizing\*\* weight by height (e.g., calculating the Body Mass Index, BMI), which could encapsulate both features into a single, meaningful metric.

\*\*Action\*\*: You might decide to:

- Create a derived feature such as BMI: `BMI = weight / (height)^2`.

- Use this derived feature in place of height and weight in the model to reduce redundancy.

#### 2. \*\*Steps and Distance\*\* (Correlation: \*\*0.66\*\*)

- \*\*Interpretation\*\*: More steps generally mean a greater distance traveled, which is why we see a positive correlation between these two features.

- \*\*How to Use It\*\*:

- In predictive models for activity level, both \*\*steps\*\* and \*\*distance\*\* provide insight into how much physical activity a user is engaging in. These are key indicators of movement.

- Since they’re strongly correlated, you could consider using either of these features or a combination, depending on the context of your model.

- Additionally, you could create a \*\*ratio\*\* feature like `Steps per Distance` to capture the user’s stride length or efficiency of movement, which could help differentiate activity types (e.g., walking vs. running).

\*\*Action\*\*: You might decide to:

- Use both steps and distance in your model, but be mindful that they might contribute similar information.

- Alternatively, you can create a new feature like `Steps/Distance` to add new, potentially useful insights to the model.

#### 3. \*\*Heart Rate and Related Features\*\*

- \*\*`Applewatch.Heart\_LE` vs. `NormalizedApplewatchHeartrate\_LE` (Correlation: \*\*0.90\*\*)\*\*:

- \*\*Interpretation\*\*: The \*\*heart rate\*\* (`Applewatch.Heart\_LE`) and \*\*normalized heart rate\*\* (`NormalizedApplewatchHeartrate\_LE`) are almost perfectly correlated, meaning they are likely different representations of the same underlying metric.

- \*\*How to Use It\*\*:

- In a predictive model, it’s unnecessary to use both features because they provide nearly identical information. Instead, choose one (likely the normalized version, which accounts for individual baseline variations).

- This high correlation suggests that heart rate data plays a significant role in distinguishing between different levels of activity (e.g., low vs. high intensity).

\*\*Action\*\*: You might decide to:

- Use \*\*only one\*\* of the heart rate features (preferably the normalized version, since it accounts for baseline variations).

- Consider excluding the other feature to avoid redundancy.

- \*\*`ApplewatchIntensity\_LE` vs. `Applewatch.Heart\_LE` (Correlation: \*\*0.80\*\*)\*\*:

- \*\*Interpretation\*\*: The strong positive correlation between \*\*intensity\*\* and \*\*heart rate\*\* makes sense because higher heart rates are associated with higher physical activity intensity.

- \*\*How to Use It\*\*:

- Both intensity and heart rate capture a user’s exertion levels. In a model for predicting activity levels, including one or both of these features could provide important signals about whether someone is engaged in low, medium, or high-intensity activity.

- However, given their strong correlation, you may choose to prioritize one over the other, depending on the focus of your analysis (e.g., heart rate for health-related studies, intensity for fitness tracking).

\*\*Action\*\*: You might decide to:

- Choose one of these features to avoid redundancy, or

- Combine them into a new feature (e.g., `Heart Rate × Intensity`), which could capture the combined effect of how hard someone is working during physical activity.

### Practical Examples of Using Strongly Correlated Features

#### \*\*Scenario 1: Predicting User’s Activity Level\*\*

If you are building a model to predict whether a user is engaging in \*\*low, medium, or high activity\*\*, you would want to focus on features that directly capture their physical exertion:

- \*\*Steps\*\* and \*\*distance\*\* are strong indicators of physical activity, so including both in the model makes sense, though a ratio like `Steps/Distance` might provide additional insight.

- \*\*Heart rate\*\* and \*\*intensity\*\* both give direct information about the user’s exertion level. Since they are strongly correlated, you might choose the \*\*intensity\*\* feature for simplicity.

- You could exclude or downplay features like \*\*age\*\* if they are not strongly correlated with physical activity.

#### \*\*Scenario 2: Health Analysis Using Heart Rate\*\*

If you are analyzing heart health, then the strong correlations between the various heart rate features would be important:

- You would likely focus on the \*\*normalized heart rate\*\* (`NormalizedApplewatchHeartrate\_LE`) and \*\*resting heart rate\*\* (`RestingApplewatchHeartrate\_LE`), as they provide important information about overall cardiovascular fitness.

- The high correlation between \*\*heart rate\*\* and \*\*intensity\*\* suggests that users who regularly engage in high-intensity activities will also exhibit elevated heart rates, which might influence the prediction of long-term health outcomes.

### Conclusion:

By focusing on features with strong correlations, you streamline your analysis and model-building by emphasizing the most impactful features and avoiding redundancy. You can also engineer new features (like BMI or Steps/Distance) to capture the relationships between correlated variables more effectively.

Ultimately, these strong correlations provide powerful signals for understanding user behavior and improving model performance. Let me know if you need help applying these insights to your specific task!

PCA