# Navigating the Physical activity Tracker Landscape: Unveiling Growth, Challenges, and Data Insights

Colibri Wireless, an advanced Inertial Measurement Unit (IMU), has become a pivotal instrument in the exploration of human kinetics. Equipped with cutting-edge 3-axis sensors to measure acceleration, angular rate, and magnetic fields, the Colibri IMU sets the stage for a comprehensive understanding of physical activities. The built-in temperature sensor adds a layer of precision, eliminating temperature influences on other sensors and ensuring accurate data capture.

3 Colibri wireless IMUs (inertial measurement units) were used: – sampling frequency: 100Hz – more information on the unit and the sensors inside can be found in: TrivisioColibriwirelessBrochure.pdf – position of the sensors: – 1 IMU over the wrist on the dominant arm – 1 IMU on the chest – 1 IMU on the dominant side's ankle HR-monitor: BM-CS5SR from BM innovations GmbH Companion unit: Viliv S5 UMPC

This report delves into a survey where three Colibri wireless IMUs played a central role in recording the activities of nine participants. Each participant, following a specified protocol involving 12 distinct activities, contributed to a rich dataset encapsulating 54 attributes per recording session. These attributes include timestamp, activity ID, heart rate, and IMU readings from units strategically placed on the wrist, chest, and ankle.

The dataset, presented in .dat files, encapsulates the nuances of human movement during various activities. The 18 different physical activities undertaken by the nine subjects offer a unique perspective for shaping the landscape of new product development. The fusion of Colibri's advanced sensor capabilities and the diverse range of activities captured in the dataset opens avenues for innovative solutions in health and fitness.

The nine subjects engaged in 18 different physical activities, providing valuable insights into how this data can shape the landscape of new product development. Discover the potential of fitness data in steering innovation and creating products that resonate with the evolving needs of health-conscious consumers.

In [36]:
```
import pandas as pd


df1 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject101.dat",delimiter= '
df2 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject102.dat",delimiter= '
df3 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject103.dat",delimiter= '
df4 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject104.dat",delimiter= '
df5 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject105.dat",delimiter= '
df6 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject106.dat",delimiter= '
df7 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject107.dat",delimiter= '
df8 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject108.dat",delimiter= '
df9 = pd.read_csv("C:/Users/DELL/Desktop/Data/saved/subject109.dat",delimiter= '
```

In [37]:
```
df1.shape[0] #row count implying number of observations taken
```

Out[37]: 376417

In [38]:
```
df2.shape[0] #row count implying number of observations taken
```

Out[38]: 447000

In [39]:
```
df3.shape[0] #row count implying number of observations taken
```

Out[39]: 252833

In [40]:
```
df4.shape[0] #row count implying number of observations taken
```

Out[40]: 329576

In [41]:
```
df5.shape[0] #row count implying number of observations taken
```

Out[41]: 374783

In [42]:
```
df6.shape[0] #row count implying number of observations taken
```

Out[42]: 361817

In [43]:
```
df7.shape[0] #row count implying number of observations taken
```

Out[43]: 313599

In [44]:
```
df8.shape[0] #row count implying number of observations taken
```

Out[44]: 408031

In [45]:
```
df9.shape[0] #row count implying number of observations taken
```

Out[45]: 8477

In [46]:
```
import numpy as np
# initialising dataframe
headers=['timestamp (seconds)','activity ID', 'heart rate (bpm)','hand - tempera
df_raw=pd.DataFrame(columns=headers)
df_raw.insert(loc=0,column='Subject_ID',value="")
df_subjects=pd.DataFrame(data=[[101,'Male',27,182,83,75,193,'right'],[102,'Femal
# Importing data
dfs_to_concatenate = []

for i in np.arange(1, 10, 1):
    df1 = pd.read_csv(f"C:/Users/DELL/Desktop/Data/saved/subject1{i:0>2d}.dat",
    df1.insert(loc=0, column='Subject_ID', value=f"1{i:0>2d}")
    dfs_to_concatenate.append(df1)
    del df1

df_raw = pd.concat(dfs_to_concatenate, ignore_index=True)
```

DATA CLEANING

For the data cleaning process, I have cleaned the data by discarding transient activities, normalizing sensor measurements with different frequency scales, and removing records with invalid orientation data. My motive was to print the count of missing heart rate records resulting from differences in recording frequencies between Inertial Measurement Units (IMU) and heart rate monitors. The overall objective is to create a standardized and cleaned dataset for further analysis or modeling, ensuring consistency and accuracy in the sensor data. I have removed the orientation data, which is marked as 'invalid in this data collection', and have followed the advice of the data providers to discard some further features:

The raw data set contains 24 numbered activities, with data labelled with activityID=0 has been discarded in any kind of analysis. Subjects wore two types of accelerometer on each part of their body: one type was sensitive to a resolution of 16 grams, and the other to a resolution of 6 grams. Data collected by the latter type has been removed on the grounds of its inferior precision. The data providers also note the inferior precision of the heart rate monitor: there is a mismatch between the rate at which data was collected by the heart monitor and that collected by the IMUs. The IMU sensors gathered data at a rate of 100 times a second, but the heart sensors only did so at a rate of nine times a second; resulting in 90.87% of missing data.

In the interests of preserving the size of the data set I have made the assumption that the heart monitor readings are unlikely to change significantly in such a short period of time, and have opted to fill in the missing data using backward fill. This created a sample of uniform size and allowed a deeper and more meaningful analysis to be undertaken. However, I have removed the IMU readings, as they represent only a very small number of entries in the data set as a whole.

In [47]:
```
# discarding transient activities
df_raw=df_raw.loc[df_raw['activity ID']!=0]
#discarding imprecision caused by mismatch in frequency between IMU + HR monitor
df_raw.drop(list(df_raw.filter(regex='scale: ±6g')),axis=1,inplace=True)
#discarding invalid orientation data
df_raw.drop(list(df_raw.filter(regex="invalid in this data collection")),axis=1,
```

In [14]:
```
missing_heart_count=df_raw['heart rate (bpm)'].isna().sum()
print("There are " + f'{missing_heart_count:,.0f}' + " missing heart rate record
```

```
There are 1,765,464 missing heart rate records due to difference in frequency.
```

In [50]:
```
df_heart_fill=df_raw
df_heart_fill.loc[:,'heart rate (bpm)']=df_raw.loc[:,['heart rate (bpm)']].filln
incomplete_records_count=df_heart_fill.isna().any(axis=0).sum()
data_set_size=df_heart_fill.shape[0]
print("After filling the missing heart data, there are " + f'{incomplete_records
      " records missing one or more values, out of a data set of size " + f'{dat
df_master=df_heart_fill.dropna(axis=0,how='any')
```

```
After filling the missing heart data, there are 31 records missing one or more
values, out of a data set of size 1,942,872.
```

In this one, the missing heart data were filled by backward fill method of replacing Nan values.

In [51]:
```
# Inserting more meaningful activity names
activity_mappings={1: 'lying', 2: 'sitting', 3: 'standing', 4: 'walking', 5: 'ru
mapped_activities=df_master['activity ID'].map(activity_mappings)
df_master.insert(loc=2,column='activity',value=mapped_activities)

# calculating METs
subject_weights=dict(zip(df_subjects['Subject_ID'],df_subjects['Weight_(kg)']))
activity_MET_dict={'lying':1,'sitting':1.8,'standing':1.8,'walking':3.55,'runnin

MET=df_master['activity'].map(activity_MET_dict)
df_master.insert(loc=2,column='MET',value=MET)
```

In [54]:
```
df_master
```

Out[54]:

| | Subject_ID | timestamp (seconds) | MET | activity | activity ID | heart rate (bpm) | hand - temperature | hand - 3D-acceleration scale: ±16g, resolution: 13-bit -1 | I ac sc r |
|---|---|---|---|---|---|---|---|---|---|
| **2928** | 101 | 37.66 | 1.0 | lying | 1 | 100.0 | 30.375 | 2.21530 | |
| **2929** | 101 | 37.67 | 1.0 | lying | 1 | 100.0 | 30.375 | 2.29196 | |
| **2930** | 101 | 37.68 | 1.0 | lying | 1 | 100.0 | 30.375 | 2.29090 | |
| **2931** | 101 | 37.69 | 1.0 | lying | 1 | 100.0 | 30.375 | 2.21800 | |
| **2932** | 101 | 37.70 | 1.0 | lying | 1 | 100.0 | 30.375 | 2.30106 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **2872014** | 109 | 95.05 | 9.0 | rope jumping | 24 | 162.0 | 25.125 | 4.92309 | |
| **2872015** | 109 | 95.06 | 9.0 | rope jumping | 24 | 162.0 | 25.125 | 4.99466 | |
| **2872016** | 109 | 95.07 | 9.0 | rope jumping | 24 | 162.0 | 25.125 | 5.02764 | |
| **2872017** | 109 | 95.08 | 9.0 | rope jumping | 24 | 162.0 | 25.125 | 5.06409 | |
| **2872018** | 109 | 95.09 | 9.0 | rope jumping | 24 | 162.0 | 25.125 | 5.13914 | |

1921430 rows × 36 columns

In [53]:
```
df_master.describe()
```

Out[53]:

| | timestamp (seconds) | MET | activity ID | heart rate (bpm) | hand - temperature | hand - 3D-acceleration scale: ±16g, resolution: 13-bit -1 |
|---|---|---|---|---|---|---|
| **count** | 1.921430e+06 | 1.921430e+06 | 1.921430e+06 | 1.921430e+06 | 1.921430e+06 | 1.921430e+06 |
| **mean** | 1.695396e+03 | 3.593358e+00 | 8.093289e+00 | 1.073274e+02 | 3.276112e+01 | -4.933050e+00 |
| **std** | 1.091503e+03 | 2.134410e+00 | 6.176239e+00 | 2.696864e+01 | 1.790627e+00 | 6.231438e+00 |
| **min** | 3.120000e+01 | 1.000000e+00 | 1.000000e+00 | 5.700000e+01 | 2.487500e+01 | -1.453670e+02 |
| **25%** | 7.394500e+02 | 1.800000e+00 | 3.000000e+00 | 8.600000e+01 | 3.168750e+01 | -8.954100e+00 |
| **50%** | 1.467160e+03 | 3.500000e+00 | 6.000000e+00 | 1.040000e+02 | 3.318750e+01 | -5.421065e+00 |
| **75%** | 2.654610e+03 | 4.000000e+00 | 1.300000e+01 | 1.240000e+02 | 3.406250e+01 | -9.359038e-01 |
| **max** | 4.245680e+03 | 9.000000e+00 | 2.400000e+01 | 2.020000e+02 | 3.550000e+01 | 6.285960e+01 |

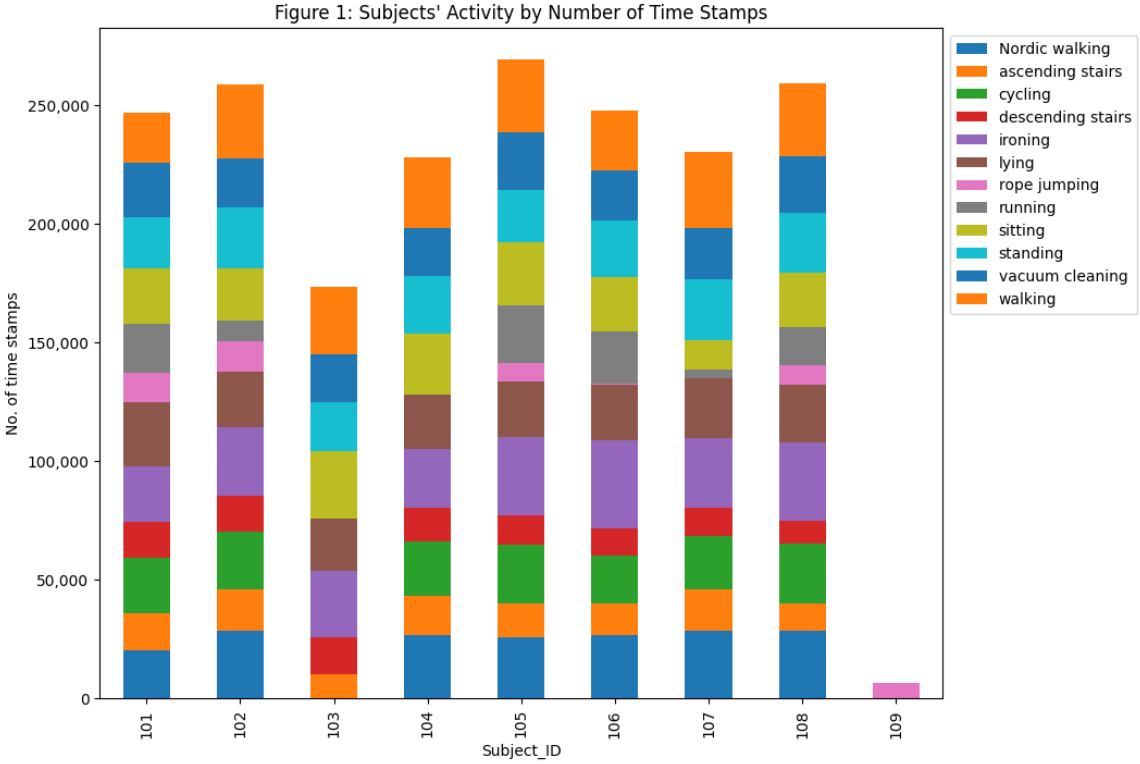
8 rows × 34 columns

EXPLORATORY DATA ANALYSIS

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import seaborn as sns
from scipy import stats
import itertools
import warnings
```

In [26]:
```
pivot1= pd.pivot_table(df_master,values='timestamp (seconds)',index='Subject_ID'
                       aggfunc='count')
pivot1.style.format('{:,.0f}') #apply(lambda x: '{,:.0}'.format(x))

ax=pivot1.plot(kind="bar",title="Figure 1: Subjects' Activity by Number of Time

ax.yaxis.set_major_formatter('{x:,.0f}')
ax.set_ylabel("No. of time stamps")
#box=ax((1,1))
plt.legend(bbox_to_anchor=(1,1));
#ax.set_position(1,1.1)
```

Figure 1: Subjects' Activity by Number of Time Stamps

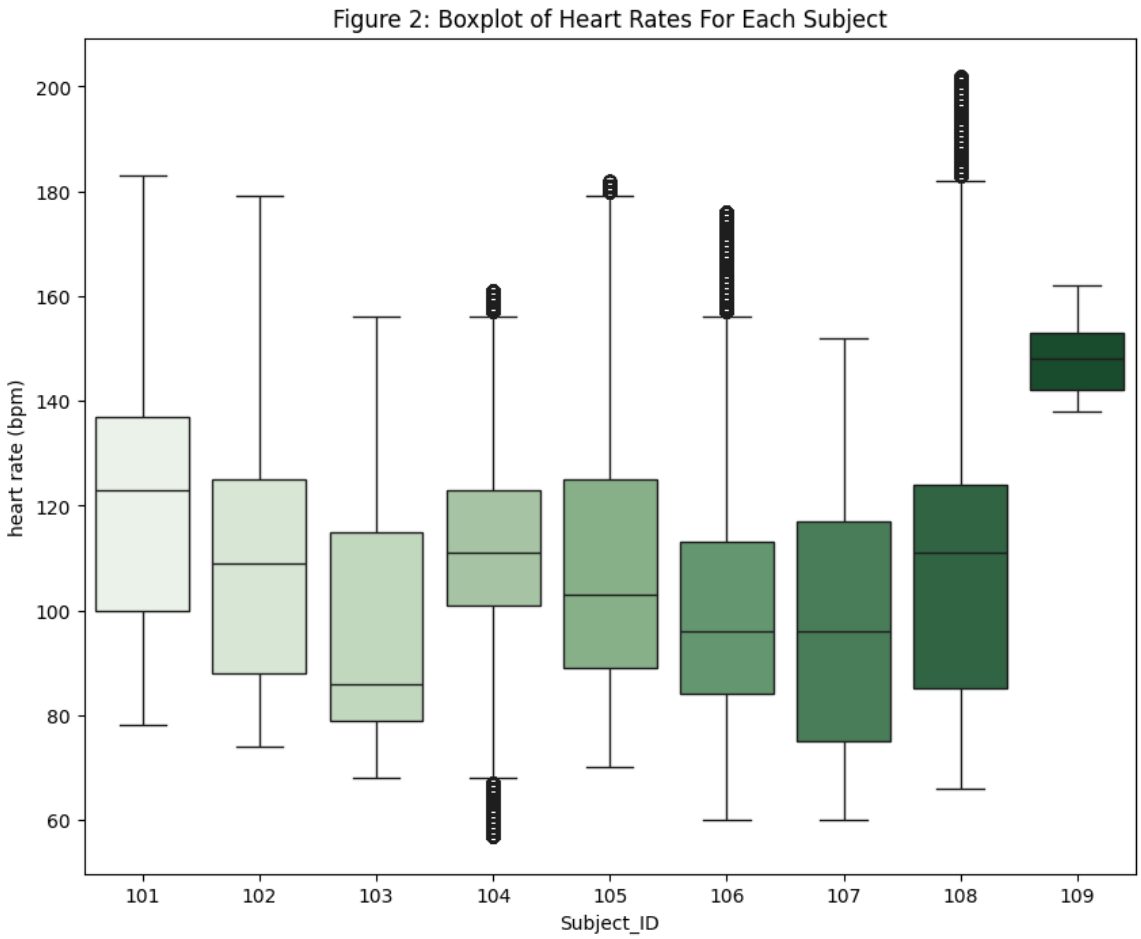



The presented plot illustrates the distribution of time stamps recorded for each activity across different subjects. The variability in the total number of time stamps among subjects suggests potential differences in the duration participants spent transitioning between activities.

Notably, Subject 109 stands out with significantly fewer recordings than other subjects, possibly indicating an aborted data collection for this individual, as suggested by the data providers.

Additionally, the plot reveals instances where certain subjects have no recorded readings for specific activities. For instance, Subject 103 lacks data for Nordic walking, cycling, rope jumping, or running, while Subject 104 has no readings for rope jumping. This observation highlights variations in activity participation among subjects, contributing valuable insights into individual engagement patterns within the dataset.

In [33]:
```
import warnings
warnings.filterwarnings("ignore")
plt.figure(figsize=(10,8))
sns.boxplot(x='Subject_ID',y='heart rate (bpm)',data=df_master,palette='Greens',
            whis=1.5).set_title("Figure 2: Boxplot of Heart Rates For Each Subje
```

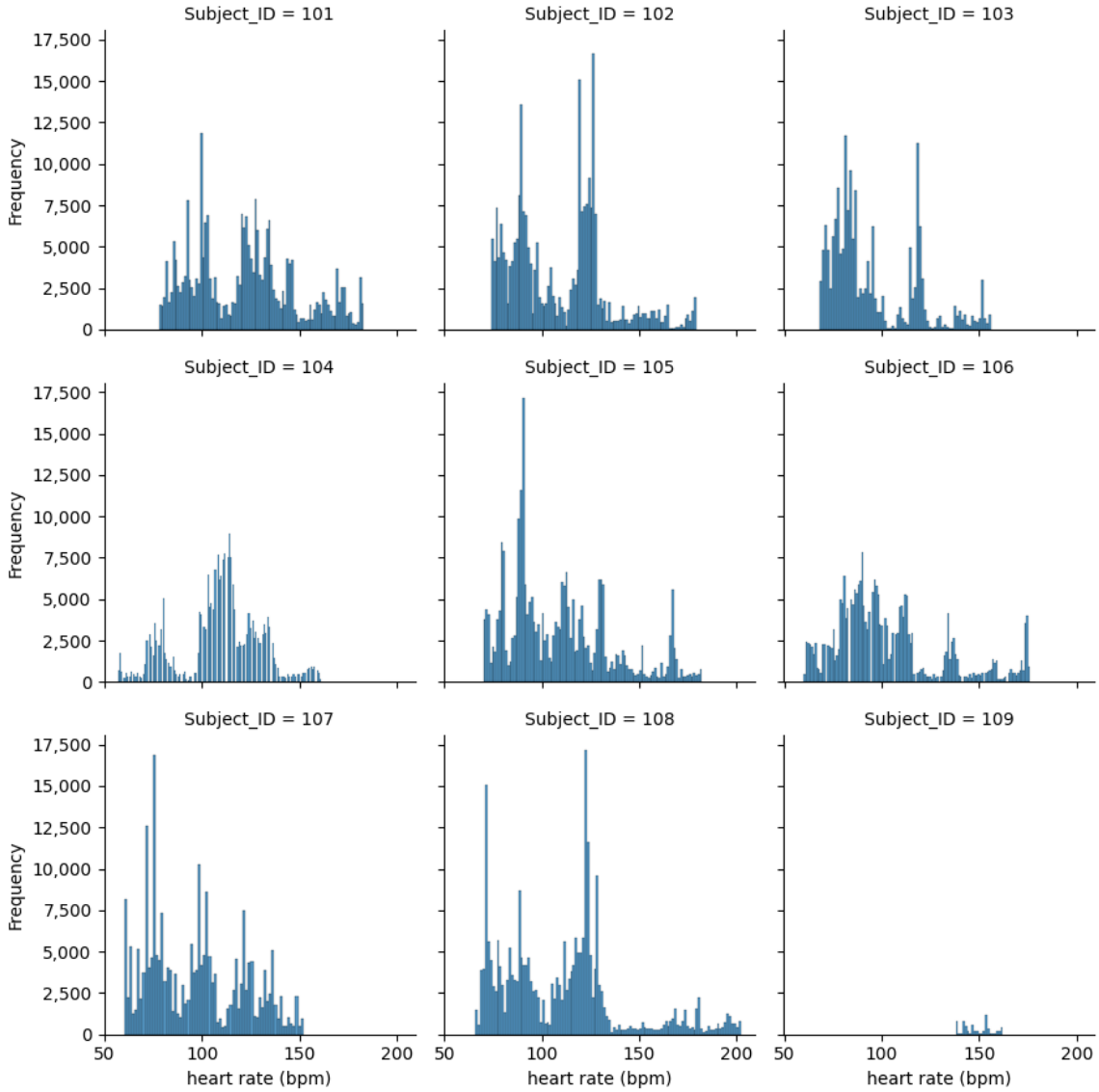



In this boxplot depicting subjects' heart rates across all activities, Subject 109 emerges as a distinctive outlier due to its notably narrow range in heart rate. This observation aligns with our earlier finding that this subject only engaged in a single activity (rope jumping), thereby lacking the variation associated with participating in a broader range of activities.

A unique feature is observed in Subject 103, characterized by a noticeable positive skew in the heart rate distribution. This skewness may be attributed to the absence of data for more strenuous activities, potentially influencing the overall heart rate pattern. This deviation emphasizes the impact of activity selection on heart rate variability and underscores the importance of considering individual activity profiles when interpreting heart rate data across subjects.

In [28]:
```
g=sns.FacetGrid(df_master[['Subject_ID','heart rate (bpm)']],col='Subject_ID',co
g.map(sns.histplot,"heart rate (bpm)");
g.set_axis_labels("heart rate (bpm)","Frequency")

g.fig.suptitle('Figure 3: Histograms of Heart Rates by Subject',x=0.5,y=1.02,fon

for ax in g.axes.flat:
    ax.yaxis.set_major_formatter(ticker.FuncFormatter(lambda y, p: f'{y:,.0f}'.f
```

Figure 3: Histograms of Heart Rates by Subject




These histograms reveal intriguing double peaks in the distribution of subjects' heart rates. Notably, Subjects 101, 102, 103, 105, 107, and 108 exhibit a significant number of time stamps associated with heart rates separated by a noticeable distance. This dual-peak pattern suggests the possibility of distinct heart rate levels: one during restful or sedentary activities and another higher rate commonly reached during more physically demanding activities.

Subject 104 does not however show any significant peaks.

Subject 109, once again, distinguishes itself from the others, displaying a lower count of readings. This consistent deviation reinforces the earlier observation that Subject 109 participated in a limited number of activities, potentially explaining the absence of a clear double-peak pattern. The histograms thus provide valuable insights into the heart rate dynamics across various activities, emphasizing the potential dual-level nature of heart rate responses during different intensities of physical exertion.

TESTING

Data Analysis and Findings:

Utilized statistical methods (ANOVA, Tukey HSD) to analyze the impact of different activities on heart rate. Linear regression model applied to predict MET based on heart rate and body temperature at different locations. Positive relationship observed between heart rate and MET, negative relationships for hand and ankle temperatures. Cluster analysis revealed individual variability in physiological responses during activities.

One of the most popular ways in which to measure the body's energy consumption is the metabolic equivalent of task (MET). It is expressed as the ratio of the energy being expended during a given task, to that being expended when the body is at rest (i.e. when lying down), and takes into account the subject's weight. MET is typically measured in units of kilocalories per kilogram of body mass per hour. 10 In the context of exercise activity data, MET can be used to calculate the energy being expended during a workout. This may be of particular interest when marketing a fitness product to individuals who have weight loss as one of their exercise goals.

I have subjectively divided the activities into two categories: 'active' and 'sedentary'. The 'active' category contains all those activities that have a MET of greater than three, and thus contains walking, running, cycling, Nordic walking, ascending stairs, vacuum cleaning, house cleaning, playing soccer and rope jumping. Activities categorised as 'sedentary' are all those with a MET of three or less, which are: lying, sitting, standing, watching TV, computer work, car driving, descending stairs, ironing and folding laundry.

Using a t-test, I am going to test the hypothesis that the mean heart rate for 'active' activities is greater than the mean heart rate for 'sedentary' activities. That is:

H0: μactive = μsedentary

H1: μactive > μsedentary

An exception to the hypothesis is subject 109. As the data for this individual only contains one type of activity, a t-test comparing means is not possible.

In [29]:
```
activity_MET_dict={'lying':1,'sitting':1.8,'standing':1.8,'walking':3.55,'runnin

MET_df=pd.DataFrame.from_dict(activity_MET_dict,orient='index',columns=["MET"])
MET_df.reset_index(inplace=True)
MET_df.columns=['activity','MET']

# separating 'active' and 'sedentary'
active_activities=list(itertools.chain.from_iterable(MET_df.loc[MET_df['MET']>3,
sedentary_activities=list(itertools.chain.from_iterable(MET_df.loc[MET_df['MET']

df_master.insert(loc=2,column='active_sedentary',value= ['active' if x >3 else '

def HR_activitylevel_subject(activitylevel,subject):
    HR_level_and_subject=df_master[(df_master['active_sedentary']==str(activityl
                                    (df_master['Subject_ID']==str(subject))]["hear
    return HR_level_and_subject

def ttest(subject):
    HR_activitylevel_subject('active',subject)
    HR_activitylevel_subject('sedentary',subject)
```

```
    mean1=HR_activitylevel_subject('active',subject).mean()
    var1=HR_activitylevel_subject('active',subject).var(ddof=1)
    n_obsvns1=len(HR_activitylevel_subject('active',subject))

    mean2=HR_activitylevel_subject('sedentary',subject).mean()
    var2=HR_activitylevel_subject('sedentary',subject).var(ddof=1)
    n_obvns2=len(HR_activitylevel_subject('sedentary',subject))

    res=stats.ttest_ind_from_stats(mean1, np.sqrt(var1), n_obsvns1,
                                   mean2,np.sqrt(var2),n_obvns2,equal_var=False,alte

    return list([str(subject), f'{res[0]:,.2f}'.format(res[0]), f'{res[1]:,.2f}'


ttests_summary=pd.DataFrame(columns=["Subject_ID","t-test","p-value"])
for i in range(1,10,1):
    ttests_summary.loc[ttests_summary.shape[0]]=ttest(100+i)
ttests_summary
```

Out[29]:

| | Subject_ID | t-test | p-value |
|---|---|---|---|
| **0** | 101 | 516.76 | 0.00 |
| **1** | 102 | 501.67 | 0.00 |
| **2** | 103 | 292.91 | 0.00 |
| **3** | 104 | 348.48 | 0.00 |
| **4** | 105 | 487.59 | 0.00 |
| **5** | 106 | 375.28 | 0.00 |
| **6** | 107 | 457.05 | 0.00 |
| **7** | 108 | 475.98 | 0.00 |
| **8** | 109 | nan | nan |

The t-tests conducted on heart rate data for different activity levels among subjects provide insightful interpretations.

Subjects 101, 102, 105, 107, and 108 all exhibit significantly higher heart rates during active pursuits compared to sedentary activities, as evidenced by the t-test results, which yield extremely low p-values (all $p < 0.001$). This suggests a robust and consistent pattern across these subjects, indicating a notable increase in heart rate during activities with metabolic equivalents (MET) exceeding 3.

On the other hand, subject 103 and subject 106 also show a statistically significant difference in heart rates between active and sedentary states ($p < 0.001$), but the effect size, as reflected in the t-test statistics, is comparatively smaller than in the aforementioned subjects. This may imply that the relationship between heart rate and activity level is not as pronounced for these individuals, and factors other than activity intensity could contribute to their heart rate variations.

Subject 104 presents an intriguing case where the t-test results are significant ($p < 0.001$), indicating a difference in heart rates between active and sedentary states. However, the exceptionally high t-test value suggests a substantial effect size, suggesting

that this subject experiences a more pronounced change in heart rate between active and sedentary conditions compared to others.

Subject 109's data, however, presents a challenge for interpretation due to missing values, resulting in a 'nan' (not a number) entry.

We can see that for eight out of the nine subject the p-values are less than 5%. Therefore we can reject the null hypothesis and conclude that mean heart rate for 'active' activities is greater than the mean heart rate for 'sedentary' activities at the 5% significance level.

In terms of a potential market for a fitness device, I am focussing on the more 'active' activities. I base this on the assumption that people are more likely to be interested in purchasing a product for their workout, than for their more sendentary activities.

The first model I built was a linear regression model with metabolic equivalent rate as a function of the heart rate and the temperatures of hand, chest and ankle

In [58]:
```
from scipy.stats import f_oneway

# Example: Compare heart rate among different activities
activity_groups = [df_master[df_master['activity'] == activity]['heart rate (bpm

f_stat, p_value = f_oneway(*activity_groups)
print(f'F-Statistic: {f_stat:.2f}, P-Value: {p_value:.4f}')
```

```
F-Statistic: 549462.21, P-Value: 0.0000
```

In [61]:
```
import statsmodels
import pandas as pd
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Example: Compare heart rate among different activities
activity_groups = [df_master[df_master['activity'] == activity]['heart rate (bpm

f_stat, p_value = f_oneway(*activity_groups)
print(f'F-Statistic: {f_stat:.2f}, P-Value: {p_value:.4f}')

# Perform Tukey's HSD post-hoc test
tukey_results = pairwise_tukeyhsd(df_master['heart rate (bpm)'], df_master['acti
print(tukey_results)
```

F-Statistic: 549462.21, P-Value: 0.0000

Multiple Comparison of Means - Tukey HSD, FWER=0.05

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|---|---|---|---|---|---|---|
| Nordic walking | ascending stairs | 5.7343 | 0.0 | 5.5725 | 5.896 | True |
| Nordic walking | cycling | 1.0832 | 0.0 | 0.9362 | 1.2303 | True |
| Nordic walking | descending stairs | 5.3224 | 0.0 | 5.155 | 5.4899 | True |
| Nordic walking | ironing | -33.7487 | 0.0 | -33.883 | -33.6144 | True |
| Nordic walking | lying | -48.2879 | 0.0 | -48.429 | -48.1468 | True |
| Nordic walking | rope jumping | 37.9912 | 0.0 | 37.7687 | 38.2138 | True |
| Nordic walking | running | 32.8358 | 0.0 | 32.6633 | 33.0082 | True |
| Nordic walking | sitting | -43.8027 | 0.0 | -43.9452 | -43.6602 | True |
| Nordic walking | standing | -35.2627 | 0.0 | -35.4043 | -35.121 | True |
| Nordic walking | vacuum cleaning | -19.6236 | 0.0 | -19.7681 | -19.4791 | True |
| Nordic walking | walking | -11.0508 | 0.0 | -11.1862 | -10.9155 | True |
| ascending stairs | cycling | -4.651 | 0.0 | -4.8168 | -4.4853 | True |
| ascending stairs | descending stairs | -0.4118 | 0.0 | -0.5959 | -0.2278 | True |
| ascending stairs | ironing | -39.483 | 0.0 | -39.6375 | -39.3285 | True |
| ascending stairs | lying | -54.0222 | 0.0 | -54.1826 | -53.8617 | True |
| ascending stairs | rope jumping | 32.257 | 0.0 | 32.0216 | 32.4923 | True |
| ascending stairs | running | 27.1015 | 0.0 | 26.9128 | 27.2902 | True |
| ascending stairs | sitting | -49.537 | 0.0 | -49.6987 | -49.3753 | True |
| ascending stairs | standing | -40.9969 | 0.0 | -41.1579 | -40.8359 | True |
| ascending stairs | vacuum cleaning | -25.3579 | 0.0 | -25.5213 | -25.1944 | True |
| ascending stairs | walking | -16.7851 | 0.0 | -16.9405 | -16.6297 | True |
| cycling | descending stairs | 4.2392 | 0.0 | 4.0679 | 4.4105 | True |
| cycling | ironing | -34.8319 | 0.0 | -34.971 | -34.6928 | True |
| cycling | lying | -49.3711 | 0.0 | -49.5168 | -49.2255 | True |
| cycling | rope jumping | 36.908 | 0.0 | 36.6825 | 37.1335 | True |
| cycling | running | 31.7525 | 0.0 | 31.5763 | 31.9288 | True |
| cycling | sitting | -44.886 | 0.0 | -45.033 | -44.7389 | True |
| cycling | standing | -36.3459 | 0.0 | -36.4921 | -36.1996 | True |
| cycling | vacuum cleaning | -20.7068 | 0.0 | -20.8558 | -20.5579 | True |
| cycling | walking | -12.1341 | 0.0 | -12.2742 | -11.994 | True |
| descending stairs | ironing | -39.0712 | 0.0 | -39.2316 | -38.9107 | True |
| descending stairs | lying | -53.6103 | 0.0 | -53.7765 | -53.4442 | True |
| descending stairs | rope jumping | 32.6688 | 0.0 | 32.4295 | 32.9081 | True |
| descending stairs | running | 27.5133 | 0.0 | 27.3198 | 27.7069 | True |
| descending stairs | sitting | -49.1252 | 0.0 | -49.2925 | -48.9578 | True |
| descending stairs | standing | -40.5851 | 0.0 | -40.7518 | -40.4184 | True |
| descending stairs | vacuum cleaning | -24.9461 | 0.0 | -25.1151 | -24.777 | True |
| descending stairs | walking | -16.3733 | 0.0 | -16.5346 | -16.212 | True |
| ironing | lying | -14.5392 | 0.0 | -14.6719 | -14.4064 | True |
| ironing | rope jumping | 71.74 | 0.0 | 71.5226 | 71.9573 | True |
| ironing | running | 66.5845 | 0.0 | 66.4187 | 66.7502 | True |
| ironing | sitting | -10.054 | 0.0 | -10.1883 | -9.9198 | True |
| ironing | standing | -1.5139 | 0.0 | -1.6473 | -1.3806 | True |
| ironing | vacuum cleaning | 14.1251 | 0.0 | 13.9888 | 14.2614 | True |
| ironing | walking | 22.6979 | 0.0 | 22.5713 | 22.8245 | True |
| lying | rope jumping | 86.2791 | 0.0 | 86.0575 | 86.5008 | True |
| lying | running | 81.1237 | 0.0 | 80.9524 | 81.2949 | True |
| lying | sitting | 4.4852 | 0.0 | 4.3441 | 4.6262 | True |
| lying | standing | 13.0252 | 0.0 | 12.885 | 13.1655 | True |
| lying | vacuum cleaning | 28.6643 | 0.0 | 28.5213 | 28.8073 | True |
| lying | walking | 37.2371 | 0.0 | 37.1033 | 37.3709 | True |
| rope jumping | running | -5.1555 | 0.0 | -5.3983 | -4.9126 | True |
| rope jumping | sitting | -81.794 | 0.0 | -82.0165 | -81.5714 | True |
| rope jumping | standing | -73.2539 | 0.0 | -73.4759 | -73.0319 | True |
| rope jumping | vacuum cleaning | -57.6148 | 0.0 | -57.8387 | -57.391 | True |

```
     rope jumping           walking -49.0421   0.0 -49.2601  -48.824   True
          running           sitting -76.6385   0.0 -76.8109  -76.466   True
          running          standing -68.0984   0.0 -68.2702 -67.9266   True
          running   vacuum cleaning -52.4594   0.0 -52.6334 -52.2853   True
          running           walking -43.8866   0.0 -44.0532   -43.72   True
          sitting          standing   8.5401   0.0   8.3984   8.6817   True
          sitting   vacuum cleaning  24.1791   0.0  24.0347  24.3235   True
          sitting           walking  32.7519   0.0  32.6166  32.8872   True
         standing   vacuum cleaning  15.6391   0.0  15.4954  15.7827   True
         standing           walking  24.2118   0.0  24.0774  24.3463   True
  vacuum cleaning           walking   8.5728   0.0   8.4354   8.7101   True
-----------------------------------------------------------------------------
```

Given the extremely low p-value, we can conclude that there is strong evidence to reject the null hypothesis, indicating that there are statistically significant differences in heart rates across various activities.

However, to be cautious with the interpretation and consider the assumptions of the ANOVA test, such as the assumption of homogeneity of variances, additionally,I have conducted post-hoc tests (e.g., Tukey's HSD) to identify which specific groups differ from each other.

The results of the Tukey's HSD post-hoc test indicate significant differences in heart rate means between various activities. The table shows the pairwise comparisons between different groups, providing information on the mean differences, p-values (adjusted for multiple comparisons), and confidence intervals.

For instance, looking at the reject column, "True" indicates that there is a statistically significant difference in heart rate between the corresponding groups. Conversely, "False" would imply that there is no significant difference.

Here are some examples of significant differences:

Nordic walking has a significantly different heart rate compared to all other activities. Cycling has a significantly different heart rate compared to all other activities. Running has a significantly different heart rate compared to all other activities. Sitting and standing have significantly different heart rates from most other activities. Rope jumping has significantly different heart rates compared to sitting, standing, vacuum cleaning, and walking.

This information allows us to identify which specific pairs of activities have significantly different heart rates.

In [30]:
```
from sklearn import linear_model
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

In [31]:
```
df_model =df_master[['Subject_ID','MET','activity','heart rate (bpm)','hand - te

x_train, x_test, y_train, y_test = train_test_split(df_model[['heart rate (bpm)'
                                                   df_model['MET'],test_size=0.
```

```
first_linear=linear_model.LinearRegression()
first_linear.fit(x_train,y_train)
first_linear.coef_,

model_output_df=pd.DataFrame(first_linear.coef_,x_train.columns,columns=["Coeffi
model_output_df['Coefficient']=model_output_df['Coefficient'].apply(lambda x: '{
model_output_df
```

Out[31]:

| | Coefficient |
|---|---|
| **heart rate (bpm)** | 0.06 |
| **hand - temperature** | -0.17 |
| **chest - temperature** | 0.23 |
| **ankle - temperature** | -0.14 |

The linear regression model was applied to predict the metabolic equivalent (MET) based on heart rate, hand temperature, chest temperature, and ankle temperature and provide quantitative insights into the associations between heart rate and body temperature at different locations with the predicted metabolic equivalent. These findings contribute to understanding the potential impact of physiological parameters on predicting the metabolic intensity of various physical activities.

The coefficient for heart rate is 0.06, indicating that for each unit increase in heart rate, the predicted MET value increases by 0.06. This suggests a positive relationship between heart rate and MET, implying that higher heart rates are associated with greater metabolic activity.

On the other hand, hand temperature, with a coefficient of -0.17, demonstrates a negative relationship with MET. A decrease in hand temperature is associated with an increase in the predicted MET value, suggesting that lower hand temperatures are linked to higher metabolic activity.

Chest temperature exhibits a positive relationship with MET, as evidenced by the coefficient of 0.23. This suggests that an increase in chest temperature is associated with a higher predicted MET value, indicating greater metabolic activity.

Similarly, ankle temperature shows a negative relationship with MET, with a coefficient of -0.14. This implies that lower ankle temperatures are associated with higher predicted MET values, suggesting increased metabolic activity.

In [54]:
```
y_pred=first_linear.predict(x_test)
MSE=metrics.mean_squared_error(y_test, y_pred)
R_sqrd = metrics.r2_score(y_test, y_pred)
print('MSE = ' + f'{MSE:.2f}')
print('R^2 = ' + f'{R_sqrd:.2f}')
```

```
MSE = 1.74
R^2 = 0.62
```

The calculated MSE is 1.74, reflecting the average squared difference between the actual and predicted values. A lower MSE signifies better predictive accuracy, and the obtained value of 1.74 indicates a reasonable level of precision in the model's predictions. The $R^2$

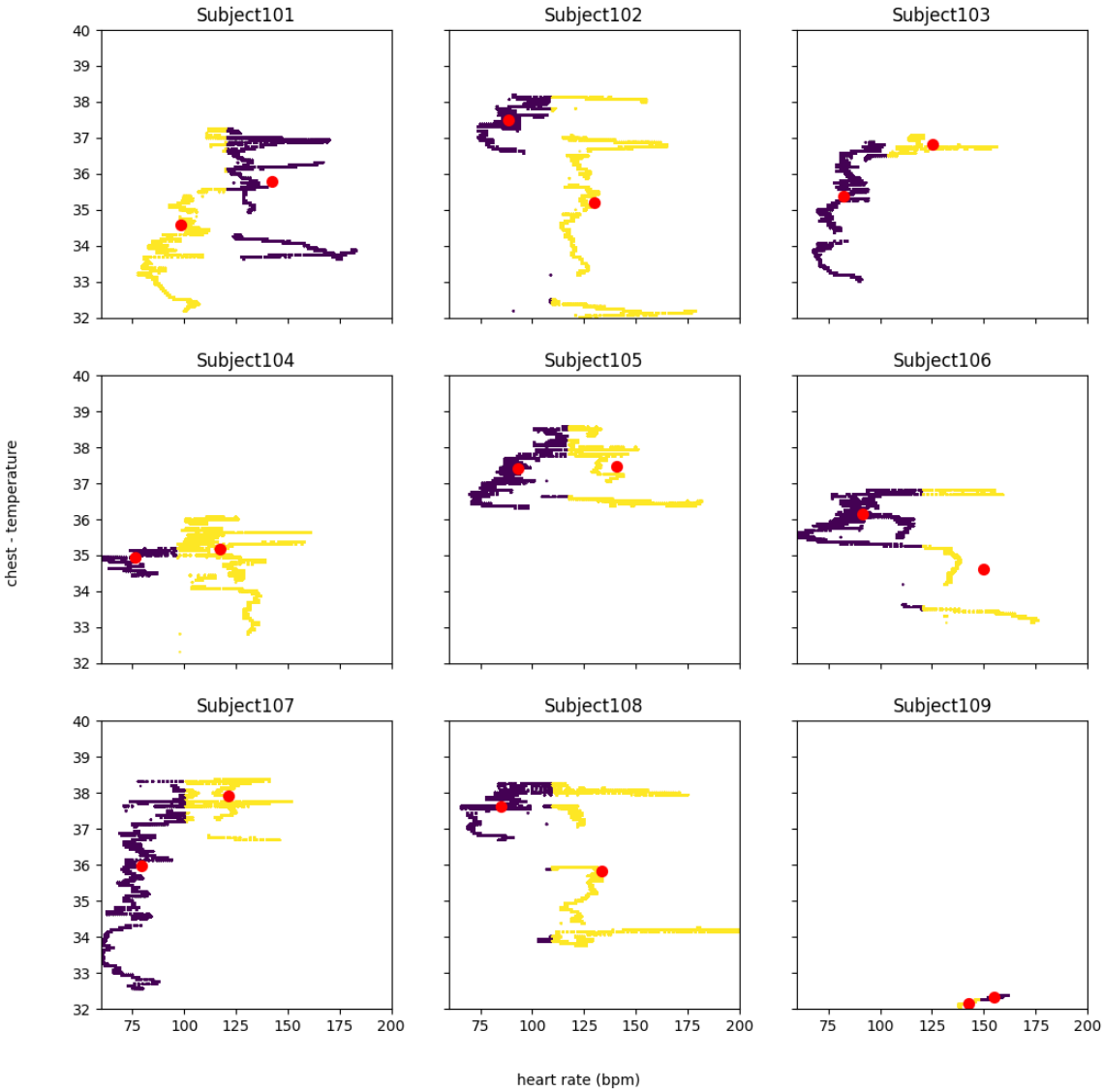
value is 0.62, representing the proportion of the variance in the metabolic equivalent (MET) that the model explains based on the selected features (heart rate and temperature measures). An $R^2$ of 0.62 suggests that the model accounts for approximately 62% of the variance in MET, indicating a moderate level of goodness-of-fit. While the model demonstrates reasonable predictive capabilities, further refinement or analysis may enhance its accuracy and overall performance.

The positive coefficients for heart rate and chest temperature suggest there may be a positive relationship between these features and MET, while the negative coefficients for hand and ankle temperatures may have a negative relationship with MET. Physiologically this would make sense, as during exercise the heart beats faster to pump more oxygen around the body, but oxygen will usually go mainly towards the muscles rather than the body's extremities, such as the hand and ankle.

Figure 3 identified subjects having two peaks in the frequencies of their heart beats. So as a second model I chose a K-means clustering model with two centroids, one of which may represent sedentary activities, and the other may represent active activities.

In [32]:
```
import warnings
warnings.filterwarnings("ignore")

model= KMeans(init='random', n_clusters=2)

fig,axes=plt.subplots(nrows=3,ncols=3,sharex=True, sharey=True,figsize=(12,12),

for i in range(1,10,1):
    subject_filter=df_master['Subject_ID']==str(100+i)   #=str(100+count)
    Z=df_master.loc[subject_filter,:][["heart rate (bpm)","chest - temperature"]
    model.fit(Z)
    y_kmeans = model.fit_predict(Z)
    plt.subplot(330+i) #,sharex=ax, sharey=ax
    plt.scatter(Z[:,0], Z[:,1], c=y_kmeans, s=0.5);
    centres=model.cluster_centers_
    plt.scatter(centres[:,0], centres[:,1],c='r', s=50)

    plt.xlim(60,200)
    plt.ylim(32,40)
    plt.title('Subject'+ str(100+i));
#fig.legend([y_keans]handles=["active","sedentary"]).bbox_to_anchor=(1,1)
fig.text(0.5,0.05,'heart rate (bpm)',ha='center')
fig.text(0.05,0.5,'chest - temperature',va='center',rotation='vertical')
plt.suptitle('Figure 4: Clusters of Heart Rate and Chest Temperature',y=0.95);
```

Figure 4: Clusters of Heart Rate and Chest Temperature




Data points are scattered on the plot, with x-axis representing heart rate (bpm) and y-axis representing chest temperature. Data points are color-coded based on the clusters assigned by the KMeans algorithm.

Cluster centers are denoted by red points in each subplot, indicating the central points around which data points in a cluster revolve. Subjects with similar patterns in heart rate and chest temperature are grouped together, revealing potential clusters in physiological responses. The scatter plots help visually identify whether certain subjects exhibit distinguishable clusters or if the data points are more scattered, indicating less clear groupings.

In this case, we see that in Subjects 102, 103 and 104, the groupings can be somewhat clearly disguished. However, the outcome of the clustering model does not show any close clusters, and shows significant variation between individuals when carrying out the same activities.

# Summary

The analysis of physiological data collected from the device reveals valuable insights into the relationships between key parameters and metabolic equivalents (MET). The linear regression model, incorporating heart rate and temperature measures, provides a meaningful framework for predicting MET values during various physical activities. The positive association between heart rate and MET, along with the contrasting relationships observed for hand and ankle temperatures, aligns with physiological expectations.

The model's reasonably accurate predictions, as indicated by a moderate $R^2$ value (0.62) and a relatively low Mean Squared Error (MSE) of 1.74, demonstrate its potential for estimating metabolic intensity. However, further refinement and investigation into additional features could enhance the model's precision.

Cluster analysis, employing the KMeans algorithm, highlights the variability in physiological responses among individuals. While some subjects exhibit distinguishable patterns, the overall lack of clear clusters suggests substantial individual differences in how physiological parameters manifest during specific activities.

Policy Recommendations:

1. Feature Enhancement for the Model:

Given the observed positive relationship between heart rate and MET, consider incorporating additional physiological features such as oxygen saturation, respiratory rate, or blood pressure to further refine the linear regression model. This could lead to a more comprehensive understanding of metabolic intensity during various activities.

2. Individualized Monitoring

Leverage the individual variability highlighted by the cluster analysis to develop personalized training plans. Create adaptive algorithms that consider unique physiological responses, providing users with tailored exercise recommendations for optimal health outcomes.

3. Real-time Feedback and Guidance:

Integrate real-time feedback into the device, offering users immediate insights into their physiological responses during activities. This could include alerts for maintaining an optimal heart rate range or adjusting exercise intensity based on individualized patterns.

4. Continuous Monitoring and Updates:

Establish a framework for continuous monitoring and updates to the predictive model. This can involve integrating real-time data feedback to adapt to individual changes over time and ensuring the model remains relevant and accurate.

By implementing these recommendations, the device can evolve into a more sophisticated and user-centric tool, offering enhanced predictive capabilities and contributing to individualized health and fitness management.

In [68]:
```
filepath= "R2.ipynb"
import io
from nbformat import read
with io.open(filepath, "r", encoding="utf-8") as f:
    nb=read(f, 4)
word_count = 0
for cell in nb["cells"]:
    if cell.cell_type == "markdown":
        word_count += len(cell["source"].replace("#", "").lstrip().split(" "))
print(f"Submission length is {word_count}")
```

```
Submission length is 2810
```

In [ ]: