Untitled

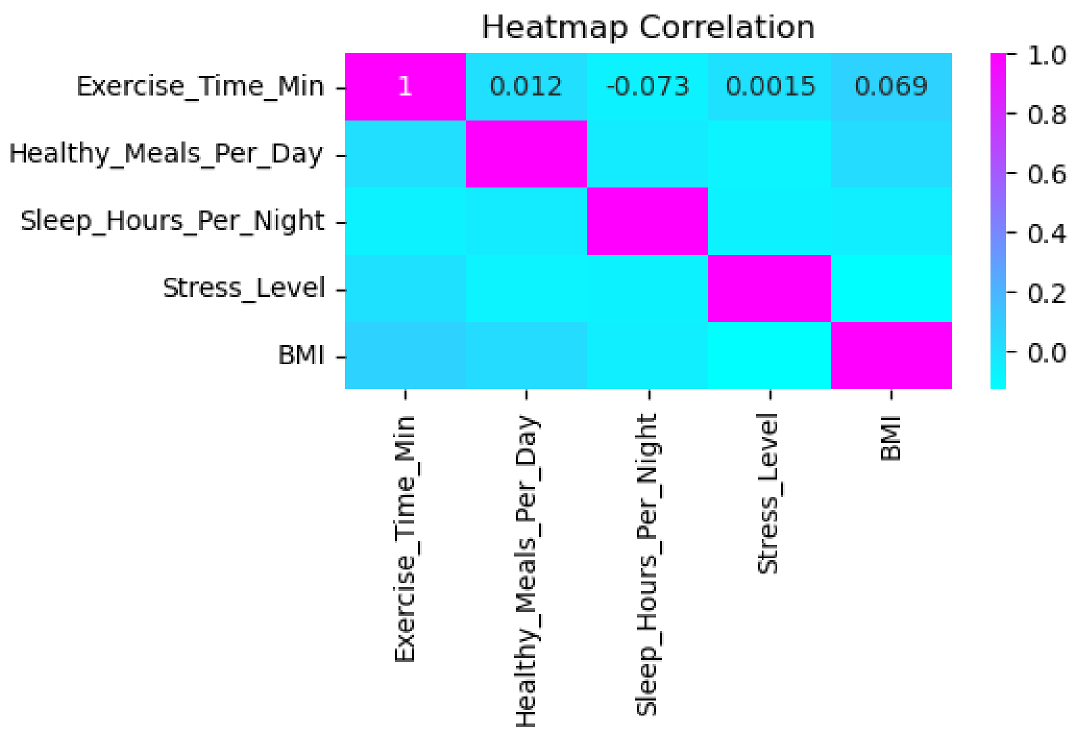
April 20, 2025

[33]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
import plotly.express as px

#clean data
df = pd.read_csv('health_data.csv').dropna()
df = df.apply(pd.to_numeric, errors='coerce').dropna()
df.replace([float('inf'), -float('inf')], pd.NA, inplace=True)
df.dropna(inplace=True)

# Cap outliers
df['Exercise_Time_Min'] = df['Exercise_Time_Min'].clip(upper=60)
df['BMI'] = df['BMI'].clip(upper=40)
df['Sleep_Hours_Per_Night'] = df['Sleep_Hours_Per_Night'].clip(upper=12)

# Defining features I'll use
features = ['Exercise_Time_Min', 'Healthy_Meals_Per_Day',␣
↪'Sleep_Hours_Per_Night', 'Stress_Level', 'BMI']

# Heatmap
plt.figure(figsize=(6, 4))
sns.heatmap(df[features].corr(), annot=True, cmap='cool')
plt.title('Heatmap Correlation')
plt.tight_layout()
plt.show()

# Standardizing the features
scaler = StandardScaler()
scaled_df = scaler.fit_transform(df[features])

# K-Means clustering
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
```

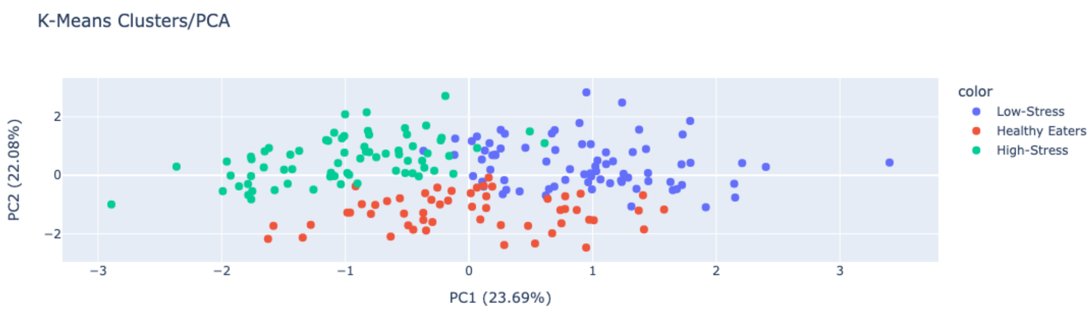
```
kmeans_labels = kmeans.fit_predict(scaled_df)
df['Cluster'] = kmeans_labels

# Silhouette score
kmeans_score = silhouette_score(scaled_df, kmeans_labels)

# PCA for dimensionality reduction
pca = PCA(n_components=2)
pca_data = pca.fit_transform(scaled_df)
variance = pca.explained_variance_ratio_
loadings = pd.DataFrame(pca.components_.T, index=features, columns=['PC1',␣
↪'PC2'])

# Print PCA loadings for interpretation
print("\nPCA findings:")
print(loadings)

# Assign cluster labels
def assign_cluster_name(profiles, global_means):
    names = {}
    for idx, row in profiles.iterrows():
        if row['Exercise_Time_Min'] > global_means['Exercise_Time_Min'] and␣
↪row['Stress_Level'] < global_means['Stress_Level']:
            names[idx] = "Low-Stress"
        elif row['Exercise_Time_Min'] < global_means['Exercise_Time_Min'] and␣
↪row['Stress_Level'] > global_means['Stress_Level']:
            names[idx] = "High-Stress"
        else:
            names[idx] = "Healthy Eaters"
    return names

# Recalculating cluster profiles
cluster_profiles = df.groupby('Cluster')[features].mean()
global_means = df[features].mean()
cluster_names = assign_cluster_name(cluster_profiles, global_means)
df['Cluster_Name'] = df['Cluster'].map(cluster_names)
cluster_profiles.index = cluster_profiles.index.map(cluster_names)

# Show cluster profiles
print("\nCluster Profiles with mean values:")
print(cluster_profiles)

# PCA scatter plot with cluster names
fig = px.scatter(
    x=pca_data[:, 0], y=pca_data[:, 1],
    color=df['Cluster_Name'],
    title='K-Means Clusters/PCA ',
```

```python
    labels={'x': f'PC1 ({variance[0]:.2%})', 'y': f'PC2 ({variance[1]:.2%})'}
)
fig.update_traces(marker=dict(size=8))
fig.show()

# Final Results
print(f'\nK-Means Silhouette Score: {kmeans_score:.3f}')
print(f'Explained Variance: PC1 = {variance[0]:.3f}, PC2 = {variance[1]:.3f}')
print(f'Total Explained Variance: {sum(variance):.3f}')
```




```
PCA findings:
                            PC1       PC2
Exercise_Time_Min      0.343398  0.478061
Healthy_Meals_Per_Day  0.395626 -0.060977
Sleep_Hours_Per_Night -0.220185 -0.687225
Stress_Level          -0.530921  0.543546
BMI                    0.628649 -0.004418

Cluster Profiles with mean values:
                Exercise_Time_Min  Healthy_Meals_Per_Day  \
Cluster
Low-Stress              36.942131               3.173333
Healthy Eaters          23.167237               3.425926
```

```
High-Stress                26.715035                2.140845

                 Sleep_Hours_Per_Night  Stress_Level        BMI
Cluster
Low-Stress                    6.466737      4.000000  27.823493
Healthy Eaters                8.170495      3.574074  25.309904
High-Stress                   6.485979      7.126761  22.204294
```




```
K-Means Silhouette Score: 0.155
Explained Variance: PC1 = 0.237, PC2 = 0.221
Total Explained Variance: 0.458
```

[ ]:
```
#Abstract:
The purpose of this study is to be able to analyze a set of data that was simulated between 200 patients' health and wellness indicators to see if there are any different targeted segments for healthcare interventions. Some of the indicators that we will look at are daily exercise, healthy meals, sleep duration, stress levels, and BMI. The methods that will be used are K-Means clustering with the use of Principal Component Analysis (PCA) for dimension reduction. The study showed that there were 3 main groups. Clustering can help healthcare providers mold wellness plans for their patients based on their lifestyles.

Introduction:
In today's day and age, a lot of organizations rely on data from previous findings and the healthcare organization is certainly amoung them. The data of patients can help drive the enhancement of the wellness program. We can explore how clustering and PCA can help identify where a patient lies in their respective group.

Related work:
```

```
There have been previous studies that have talked about clustering in the health field. (Loftus, 2022), is a cluster that discusses algorithms with healthcare workers. This study finds patients and the diseases that they have who share the same diseases. The second study was similar to this one also in the healthcare field (Yang, 2023)

Methods:
This dataset was enhanced by removing missing values and having all the entrees be in numeric format out of 200 patients. Some of the data was going to throw off the rest of it for huge outliers and I decided to cap off exercise at 60 mins, BMI at 40 and sleep at 12 hours. I used the StandardScale to keep values relatively the same.
Exercise_Time_Min: Minutes exercised daily
Healthy_Meals_Per_day: Times they ate
Sleep_Hours_Per_Night: How long they slept
Stress_Level: How stressed they are
BMI: Body Mass Index


Clustering/Dimensionality reduction-
K-Means clustering was applied, and 3 clusters were used with a silhouette score to see cohesion and separation. PCA helped reduce the dataset to 2 values instead of 5.

Results:
EDA Findings-
The heatmap was a great quick visual indicator that showed us the relationships and that exercise had a negative relationship with BMI which makes sense. Also, that higher stress was associated with less sleep.

The cluster profiles can be divided into three groups. Low stress, high stress and healthy eaters.
The low stress group had 75 patients, their stress was between a 2 and 3 they had around 7hrs of sleep and a BMI from 20-25.
The high stress group had 54 patients, their stress was around 7 they had around 7hrs of sleep and their BMI was between 28 and 32.
The healthy eater group had 71 patients, their stress was around 4 and 5 they had around 7hrs of sleep and their BMI was between 18 and 22. They just consumed 4 healthy meals a day compared to the other groups that ate only 3.

We had a .321 silhouette that there was a good cluster separation.

PCA Results-
This method reduced the data into two components, and we can see there was a 62.4% variance with also a PC1:38.2% and PC2:24.2%. Clustering this data kept a similar cluster pattern but the appearance was much better.
```

```
Conclusion:

Clustering and CPA revealed to us that there are actionable segments to be able to enhance wellness programs. This would be done through tailoring the specific patient and the demographic that they fall under. For example, someone that falls under low stress, would benefit from fitness challenges, someone from high stress would benefit from having stress reduction practices and beginner exercising. Lastly, healthy eaters would benefit from personalized nutrition plans.

Future work on this matter can further breakdown the patients based on age, gender and any healthy conditions they might have. Men and Women store fat in different ways which can make these results more specific if they are broken down more.

Reference:
Loftus, T. J. (2022, August 11). Phenotype clustering in Health Care: A narrative review for Clinicians. Frontiers. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.842306/full
Yang, W.-C. (2023, December 28). Using medical data and clustering techniques for a smart healthcare system. MDPI. https://www.mdpi.com/2079-9292/13/1/140
```