CS 455 Homework 3

Group members: Mia Lacey, Molly McNamara, Leah Casey

Data source: Kaggle.com

https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset

About the dataset:

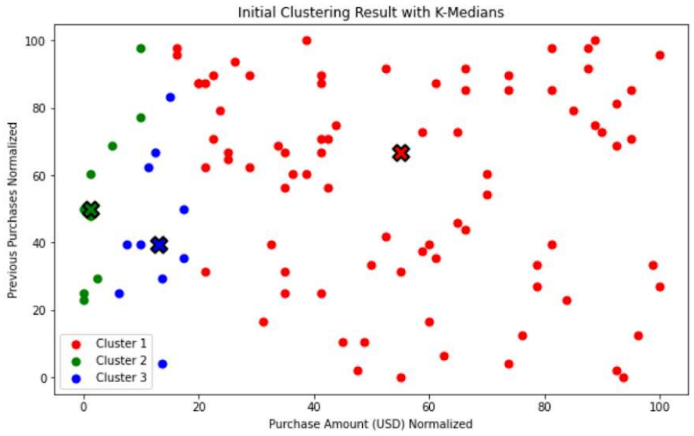
Our original dataset describes the purchases of certain items. The data describes the demographic of the person who bought said item and details about the item. This data was modified to then fit the normalized dataset. This was done by using the MinMaxScaler to normalize the columns within range 0 to 100. The updated dataset includes only the Purchase Amount (USD) and the Previous Purchases columns with these normalized ranges.

The Manhattan distance was then utilized for computation. Here, the SSE was calculated and the clusters, centroids, and iterations were all computed and returned.

The initial clustering was done by printing cluster members (Cluster 1, Cluster 2, Cluster 3) after the first loops. Then the cluster was visualized by creating a scatterplot that

highlighted the different clusters with their respective colors and markers. The SSE was also calculated using the sse_initial and then printed to show them using squared Manhattan distances.

```
In [13]: # Visualization of initial clustering
colors = ['r', 'g', 'b']
plt.figure(figsize=(10, 6))
for i in range(3):
    plt.scatter(X[clusters_initial == i, 0], X[clusters_initial == i, 1], s=50, color=colors[i], label=f'Cluster {i+
    plt.scatter(centroids_initial[i, 0], centroids_initial[i, 1], s=200, color=colors[i], marker='X', edgecolor='k',
plt.title('Initial Clustering Result with K-Medians')
plt.xlabel('Purchase Amount (USD) Normalized')
plt.ylabel('Previous Purchases Normalized')
plt.legend()
plt.show()
```




```
In [14]: print("Initial SSE using squared Manhattan distances:", sse_initial)
Initial SSE using squared Manhattan distances: 232145.22569444447
```

```
Initial Cluster Members:
Cluster 1: [0, 1, 2, 3, 4, 6, 8, 11, 12, 13, 14, 15, 16, 17, 18, 1
9, 20, 21, 22, 23, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39,
40, 41, 42, 43, 44, 46, 47, 48, 51, 54, 56, 57, 58, 59, 60, 61, 63,
64, 65, 66, 67, 68, 69, 71, 73, 74, 75, 76, 77, 78, 79, 80, 81, 83,
85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99]
Cluster 2: [5, 24, 26, 45, 50, 52, 53, 55, 62]
Cluster 3: [7, 9, 10, 25, 30, 49, 70, 72, 82, 84]
```

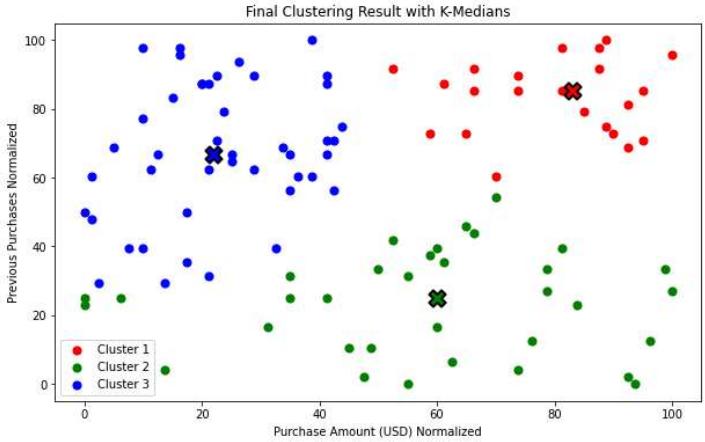
The final clustering was then performed by printing the final cluster members (Cluster 1, CLuster 2, Cluster 3). The clustering was then visualized through a similar technique as for the initial clustering. A scatterplot was created and they highlighted the different clusters and are organized by color and marker. The final SSE was calculated using the squared Manhattan distances. Finally, the total number of iterations were are printed.

In [15]:
```
# Perform final clustering until convergence
clusters_final, centroids_final, sse_final, total_iterations = k_medians(X, k=3, max_it
```

In [16]:
```
# Display final cluster members
print("\nFinal Cluster Members:")
for i in range(3):
    members = np.where(clusters_final == i)[0]
    print(f"Cluster {i+1}: {members.tolist()}")
```

```
Final Cluster Members:
Cluster 1: [3, 6, 12, 19, 23, 28, 31, 32, 34, 36, 40, 54, 56, 63, 73, 78, 79, 81, 93,
95, 96, 97]
Cluster 2: [0, 1, 2, 5, 8, 9, 11, 15, 18, 21, 25, 26, 27, 29, 37, 39, 42, 43, 51, 57,
59, 61, 64, 66, 69, 74, 75, 80, 85, 89, 91, 92, 94, 98]
Cluster 3: [4, 7, 10, 13, 14, 16, 17, 20, 22, 24, 30, 33, 35, 38, 41, 44, 45, 46, 47,
48, 49, 50, 52, 53, 55, 58, 60, 62, 65, 67, 68, 70, 71, 72, 76, 77, 82, 83, 84, 86, 8
7, 88, 90, 99]
```

In [17]:
```
# Visualization of final clustering
plt.figure(figsize=(10, 6))
for i in range(3):
    plt.scatter(X[clusters_final == i, 0], X[clusters_final == i, 1], s=50, color=color
    plt.scatter(centroids_final[i, 0], centroids_final[i, 1], s=200, color=colors[i], m
plt.title('Final Clustering Result with K-Medians')
plt.xlabel('Purchase Amount (USD) Normalized')
plt.ylabel('Previous Purchases Normalized')
plt.legend()
plt.show()
```



In [18]:
```
print("\nFinal SSE using squared Manhattan distances:", sse_final)
```

```
Final SSE using squared Manhattan distances: 94220.92013888886
```


```
Final SSE using squared Manhattan distances: 94220.92013888886
```


```
print("Total number of iterations:", total_iterations)
```

```
Total number of iterations: 6
```