Kent State University

MIS-64060: Fundamentals of Machine Learning

Fall 2021

Final ML Project

Trail Segmentation and Recommendation

By:

Jayasri Maditati

Table of Contents

"A national trail is a gateway into nature's secret beauties, a portal to the past, a way into solitude and community. It is also an inroad to our national character. Our trails are both irresistible and indispensable."

—STEWART UDALL, US Secretary of the Interior (1961–69); 1920–2010

Overview:

The United States is as diverse as it is vast. There is no other country with as many hiking trails as the United States. I go on a lot of hikes. My closest friends and family are avid hikers and I always get dragged along. However, I'll be the first to say that I don't enjoy it sometimes. That doesn't mean that every hike is terrible; I have been on some fantastic ones. After that, I've always thought to myself, why can't all hikes be like that one? The one thing, I really wish all these hiking resources had, was a personalized hiking suggestions based on hikes I've enjoyed in the past (i.e. if I liked hike X, I'll like hike Y because it has similar hiking trail features). Well, 'Wanderlust Bundle' is here to solve that problem! It will recommend hikes to you based on the user specified trail.

Data Exploration and Preparation:

I got my dataset (AllTrails) from Kaggle that provides information on the top/best trails in the United States and Hawaii. I focused only on the United States trails that left me with 3240 trails. Once I cleaned the data by getting rid of the hikes that had missing data, I was left with about 2996 trails.

The following trail attributes were collected, which included a mix of numeric and categorical features:

- **Trail ID** - A Unique Id given for each trail
- **Name** - Name of the trail
- **Area Name** – Name of the National Park in which the trail is located
- **City Name** – City in which the trail is located
- **State Name** – State in which the trail is located
- **Country Name**- Name of the Country
- **X_geoloc** – Details about Latitude and Longitude of the trail
- **Popularity** – Details about how popular the trail is
- **Difficulty Rating** (Easy, Moderate, Hard) – 1,3,5,7
- **Average Rating** – It is given by Reviewers (0 to 5)
- **Number of Reviews** – Number of reviews for the trail
- **Length (Meters)** – Length of the Trail
- **Elevation Gain (Meters)** – Elevation Gain of the Trail
- **Route Type** (Loop, Out & Back, Point to Point)
- **Visitor Usage** – Usage of the trail by visitors – 1,2,3,4
- **Features** – Views, wildlife, forest, lake, paved, river etc.,
- **Activities** – Camping, Backpacking, hiking, biking etc.,

The numerical features would be length, elevation gain, popularity, number of reviews, visitor usage, difficulty rating and average rating.

The remaining features would be categorically tagged with a value of 0 for “No” or 1 for “Yes” depending on if the feature is described a given hike. I would be representing them as ‘Trail Tags’ going forward.

Data Analysis:

During my initial Analysis on the data, I identified 3 different patterns in the dataset.

- **OHV Trails** - Trails mainly designed for Off-Road Driving, Bike-Touring, Road-Biking and Scenic Driving
- **Recreation Trails** – These trails are primarily used for Paddle sports, kayaking, Canoeing, Cross Country skiing, snowshoeing etc., Visitors for these trails seems be more interested in doing these activities rather than hiking or walking.
- **Hiking Trails** – These are mainly used for hiking, walking, backpacking, camping, and enjoying wildlife.

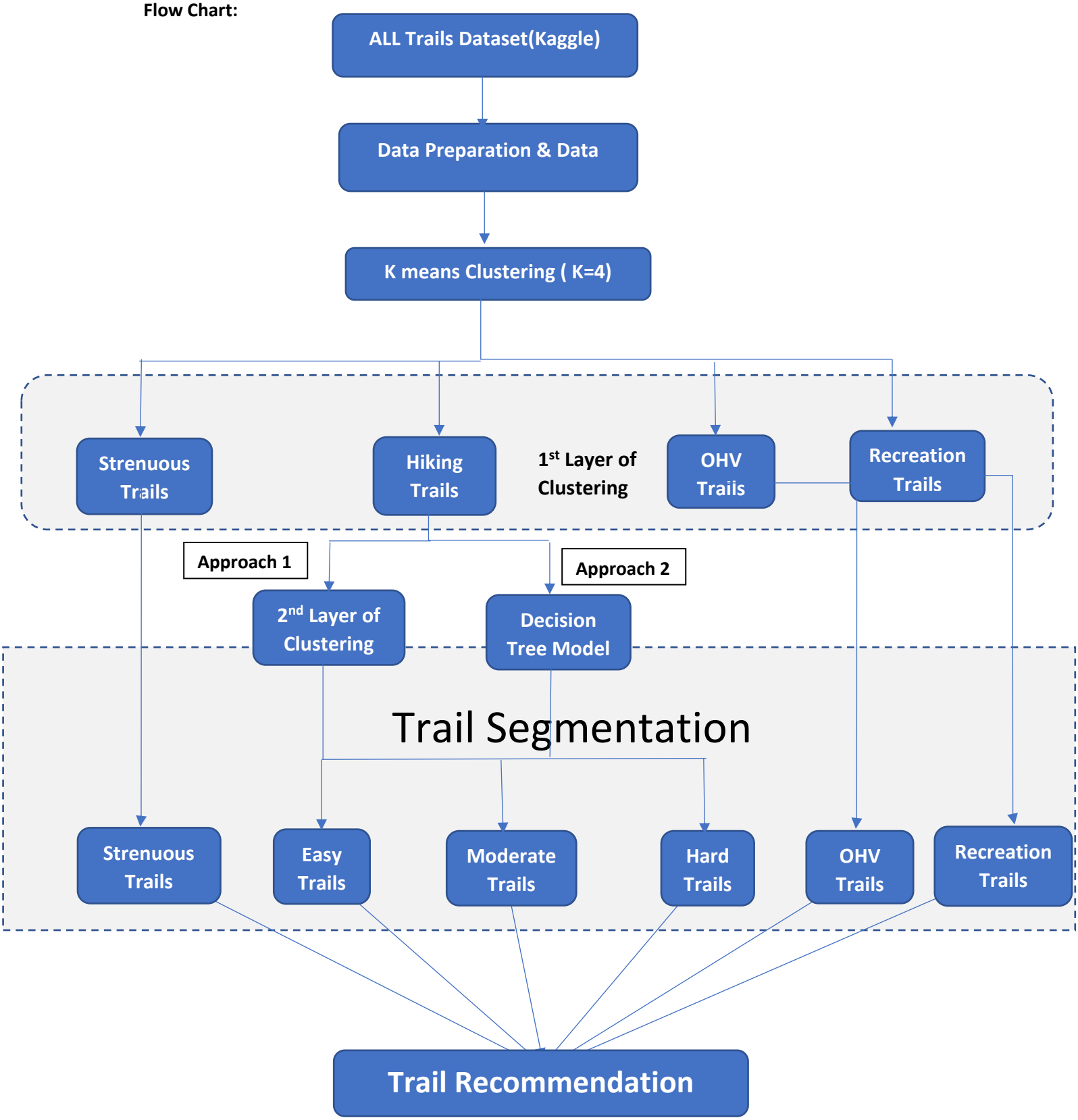
As my objective is more focused on trails for hiking, I would be working mostly on Hiking Trails, which can be further segregated as Easy, Moderate and Hard Trails.

Approach

Flow Chart for the Approach:

The below flow chart is for the approach I have followed for solving the project.

Flow Chart:


ALL Trails Dataset(Kaggle)
Data Preparation & Data
K means Clustering (K=4)
Strenuous Trails
Hiking Trails
1st Layer of Clustering
OHV Trails
Recreation Trails
Approach 1
Approach 2
2nd Layer of Clustering
Decision Tree Model
Trail Segmentation
Strenuous Trails
Easy Trails
Moderate Trails
Hard Trails
OHV Trails
Recreation Trails
Trail Recommendation

Trail Segmentation:

Machine Learning Algorithm Used: Clustering Technique (Unsupervised)

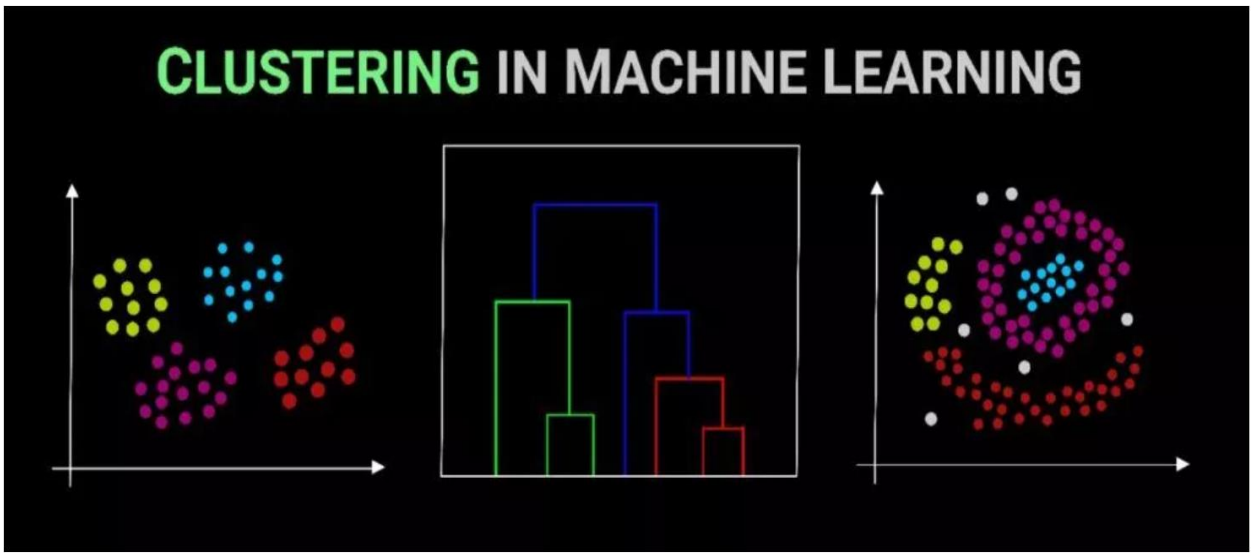
As Trail segmentation is my primary objective, clustering technique is the feasible solution

Clustering:

It is the process of grouping a set of objects into the classes of similar objects. Clustering can be used as a stand – alone tool to gain insight into data distribution and as a preprocessing step of other algorithms in intelligent systems.

Clustering Technique Type used: K- Means Clustering Technique

Types of Clustering:



K-Means Clustering **Hierarchical Clustering** **DBScan Clustering**

There are three types of clustering techniques which we have discussed in the module

- **K-Means Clustering** – (Partitional Clustering Algorithm) – can handle high dimensional and large dataset well. Highly Computational
- **Hierarchical Clustering**- (Hierarchical Clustering Algorithm) – can't handle large dataset computationally
- **DB Scan Clustering** – Density Based Clustering Algorithm – doesn't work well with high dimensional dataset.

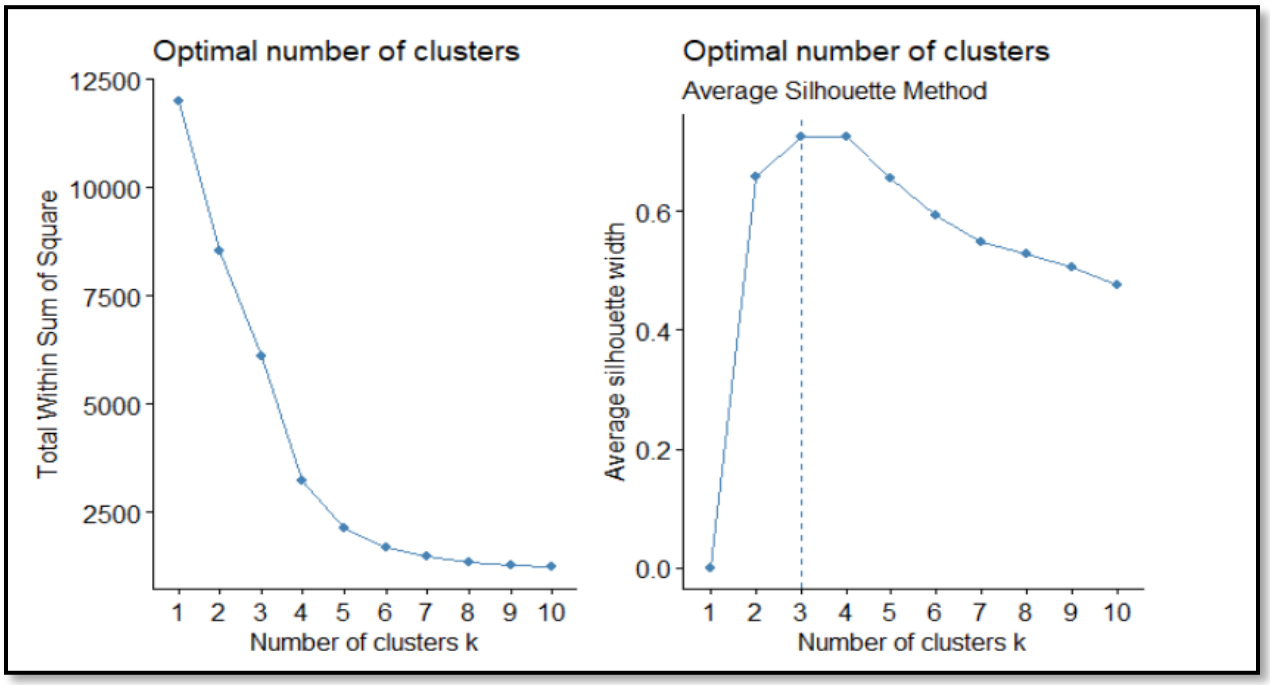
All Trails dataset is a high dimensional and large dataset where K-Means clustering algorithm is the efficient option for my project.

Choose the Optimal K:

The choice of the number of clusters can either be driven by:

- External Considerations (e.g., previous knowledge, practical constraints, etc.)
- The Elbow Method
- The Average Silhouette Method

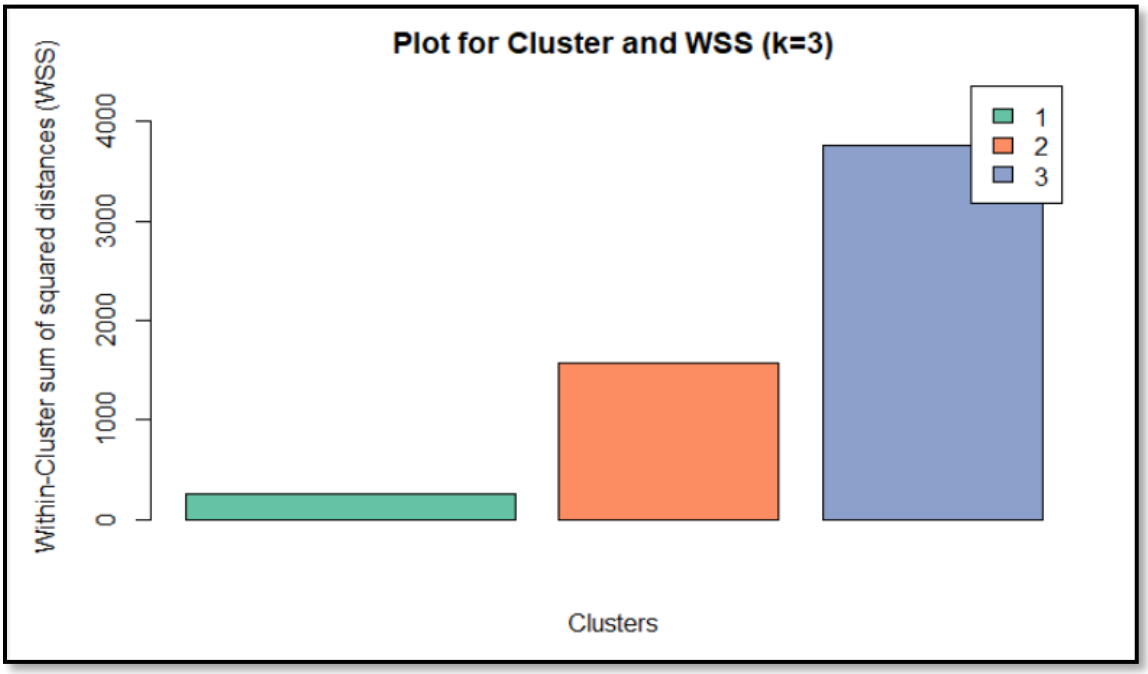
First, Let's use the two data driven methods to check for the value of K.

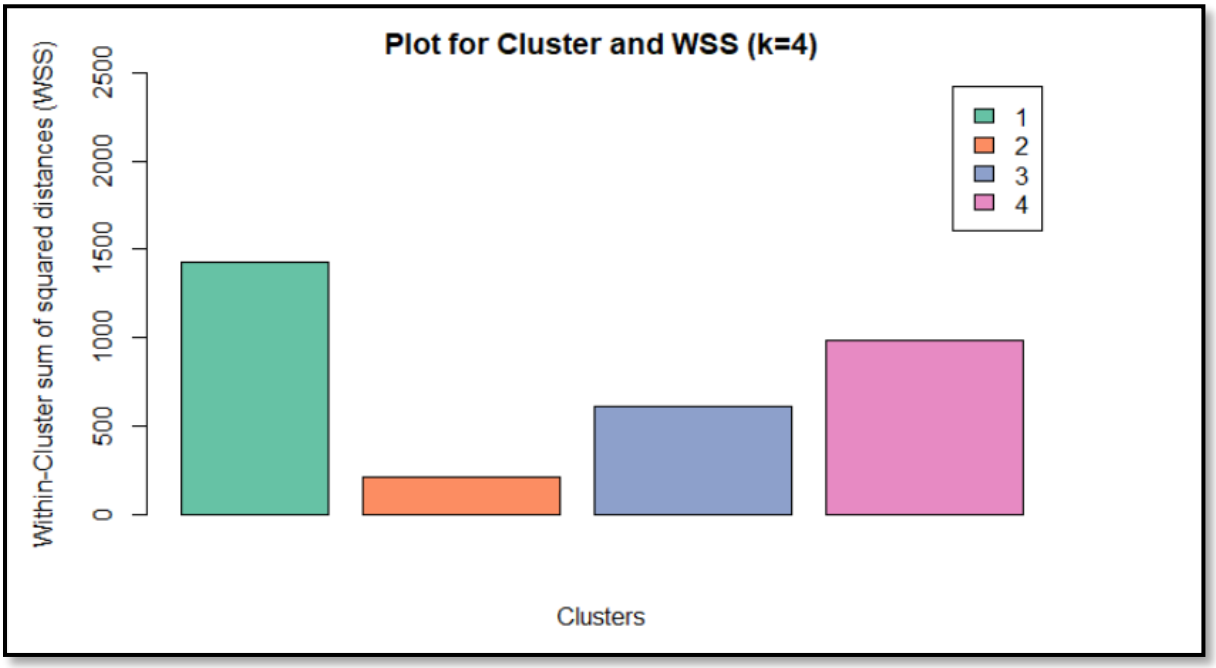


Plot for Elbow Method and Average Silhouette Method

From the two methods plotted, we can see that the Elbow Method gives us the optimal value of k is 4, While the Silhouette Method gives us k=3 as result.

I tried to run k-means Model with both k=3 and 4 and decided to consider **withinss** to find the optimal k among these two values.

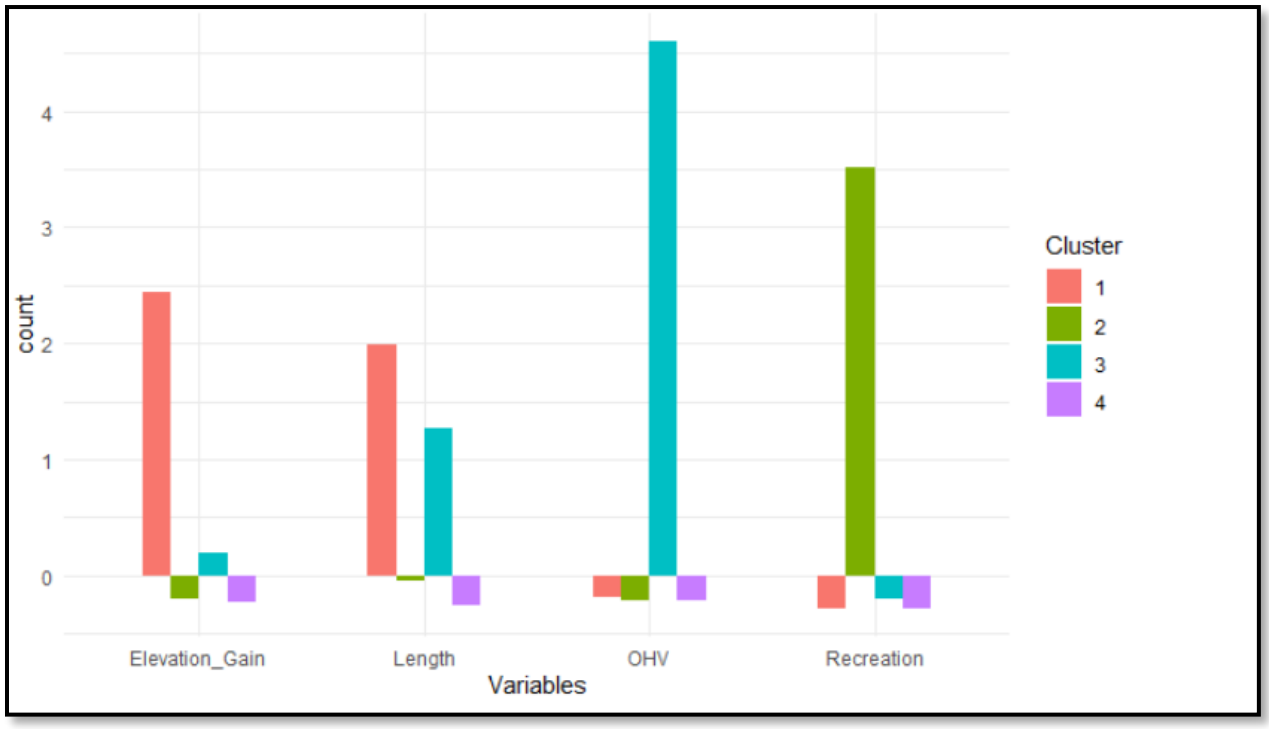




Observations:

- **WithinSS** is a measure of dispersion of the data within the cluster.
- When we take look at the above plots, It is clearly evident that the clusters in the plot where k=3 looks less homogeneous compared to the clusters in the plot where k=4.
- The goal here isn't just to make clusters, but to make good meaningful clusters.

Cluster Interpretation (First Layer of Clustering):



Plot for the average of Length, Elevation Gain and Trail Tags by clusters

- **Cluster1:** Strenuous Trails
- **Cluster2:** Recreation Trails
- **Cluster3:** OHV Trails
- **Cluster4:** Hiking Trails

Segmentation of Hiking Trails:

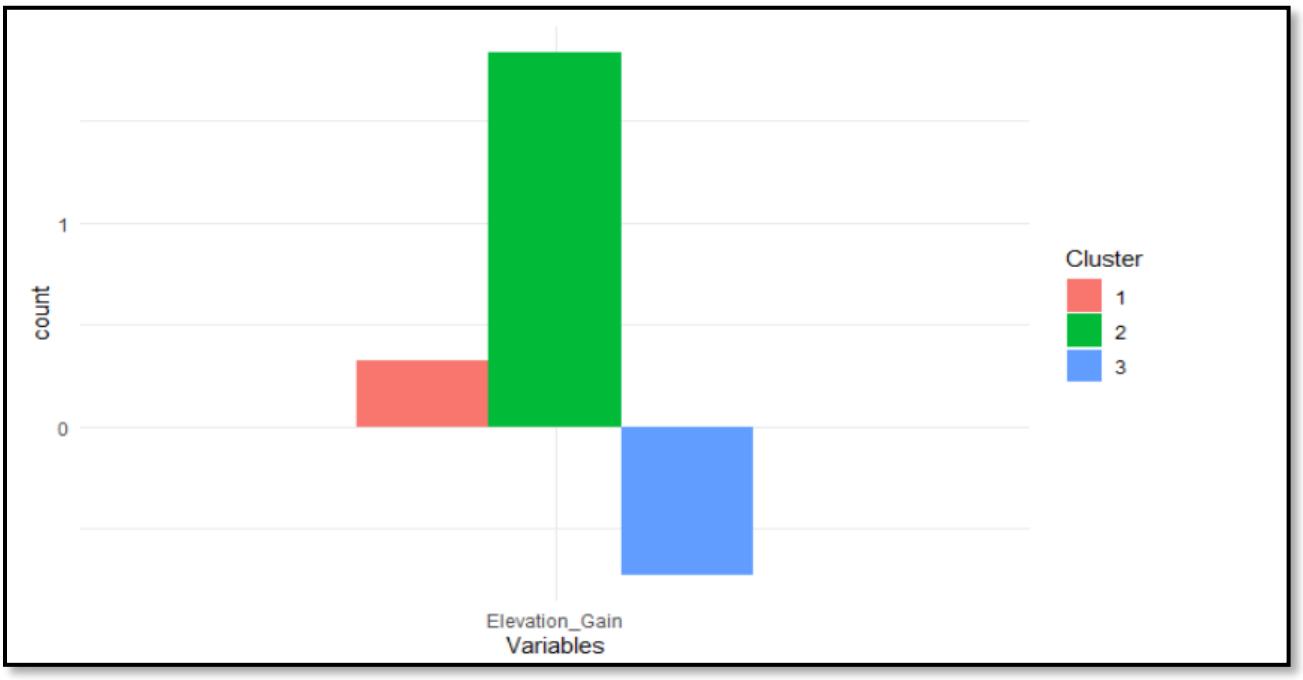
As my project is more focused on hiking trails, I decided to segment the hiking trails into Easy, Moderate and Hard Trails using 2 approaches:

- 2nd Layer of Clustering
- Decision Tree Model

Approach 1: *2nd Layer of Clustering using K means Model*

Performed K means algorithm on the Hiking Trails cluster with k=3 using **Elevation_Gain** as a variable.

Cluster Interpretation of 2nd layer Clustering:



Plot for the average of Elevation Gain by 3 Clusters

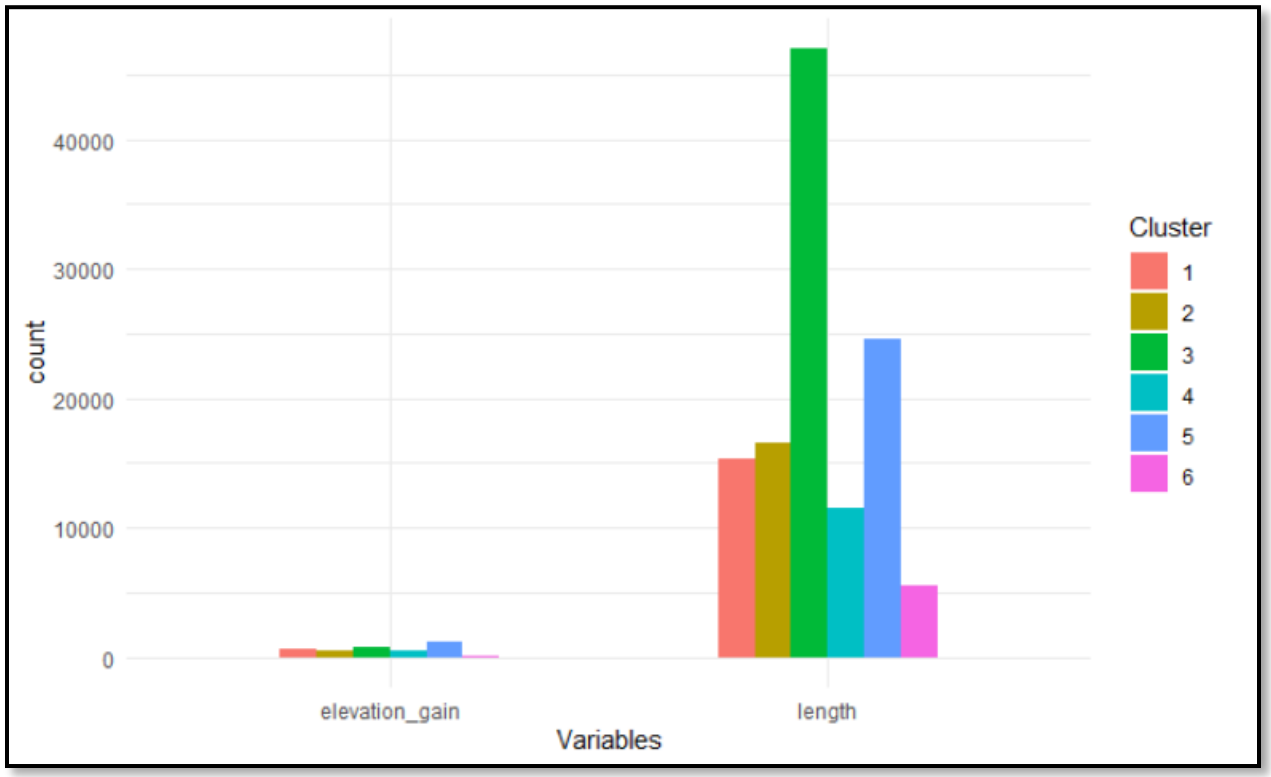
- **Cluster 1:** Moderate Trails
- **Cluster 2:** Hard Trails
- **Cluster 3:** Easy Trails

Cluster Interpretation of 6 Clusters (After 2nd layer of clustering):

After combining the 2 layers of clustering, the below table represents the summary of the mean of length and elevation gain by clusters

Table 1: Summary of Mean of Length and Elevation Gain by Clusters

Group.1 <dbl>	length <dbl>	elevation_gain <dbl>
1	15353.196	579.2975
2	16516.758	464.4012
3	47025.155	804.3399
4	11511.459	441.0636
5	24600.956	1217.3142
6	5518.455	133.3555



Plot for the average of Length and Elevation Gain by clusters

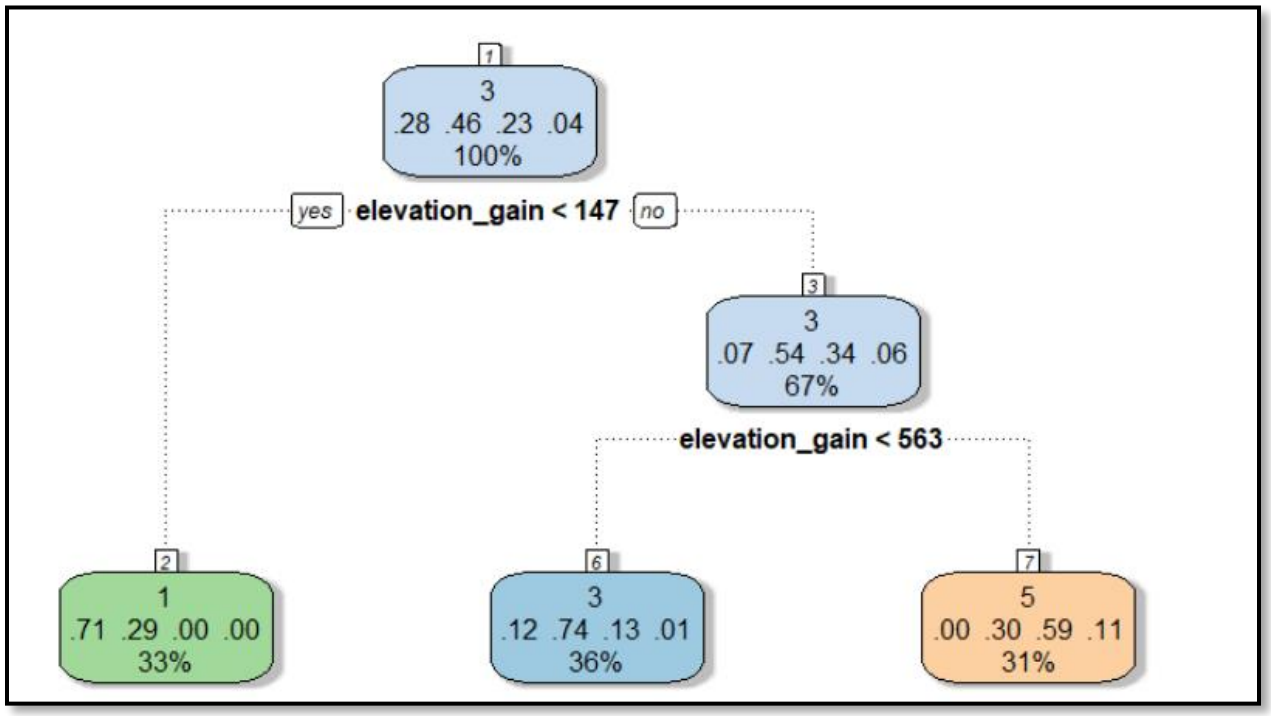
Observations:

We can see that 6 clusters with sizes of different characterizations:

- **Cluster 1**: Strenuous Trails
- **Cluster 2**: Recreation Trails
- **Cluster 3**: OHV Trails
- **Cluster 4**: Moderate Trails
- **Cluster 5**: Hard Trails
- **Cluster 6**: Easy Trails

Approach 2: Using the Decision Tree Model

To find the threshold values of length or Elevation gain in that hiking trails cluster, we are using Decision tree model.



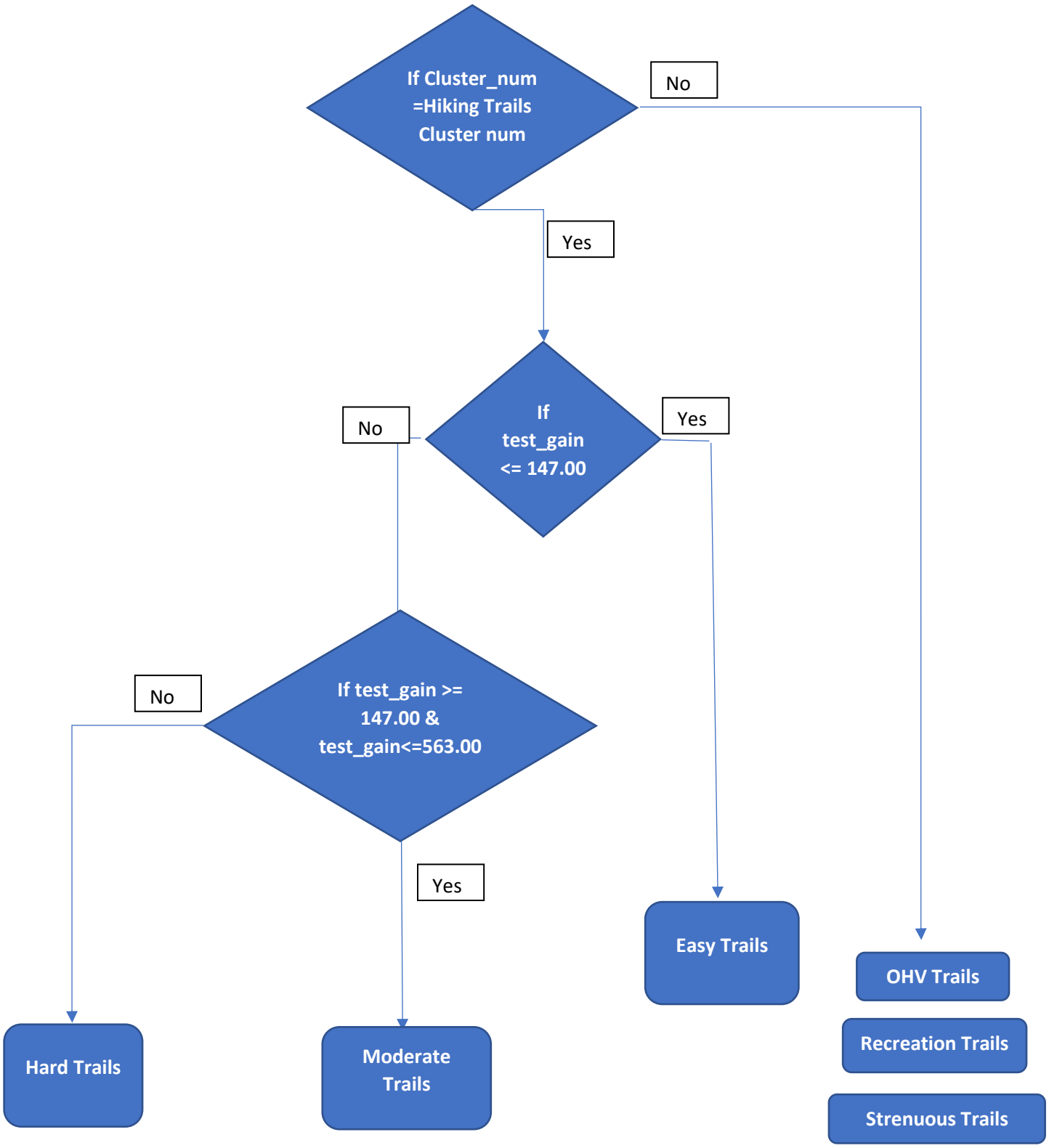
Decision Tree Model Plot

It divided into 3 leaf nodes considering elevation gain as a parameter and these values can be used during trail recommendation.

Trail Recommendation:

Trail Recommendation is one of the use cases of Trail segmentation.

Flow chart of Trail Recommendation using the Decision Tree Model Elevation gain threshold values.

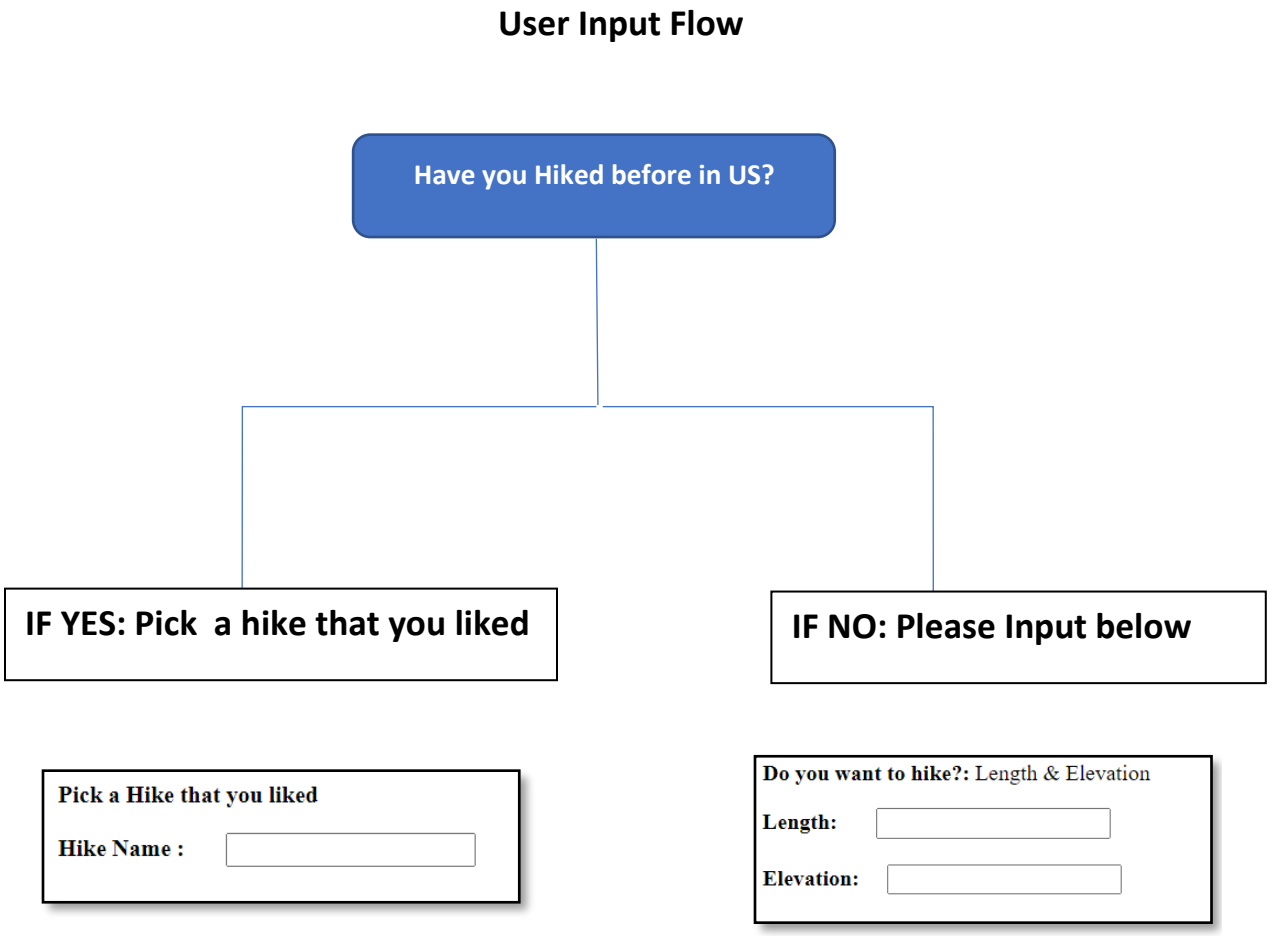


Let's recommend the similar trails using the threshold values from the decision tree as per the flowchart.

Here, a distance matrix is created using the cluster centroids and the user inputs. Once the cluster that is most similar is chosen, another distance matrix is used to determine which hikes in the cluster are closest. The recommended hike is a random choice of the top three most similar hikes.

In the above flow chart, **cluster_num = the similar cluster,** that is chosen from the distance matrix which is created using the cluster centroids and the user inputs

Results:



Scenario 1: OHV Trails

Input: Cades Cove Loop Road

Output:

```
These are OHV trails.
These are mainly for Off-Road Driving,Road-Biking,Scenic Driving and Bike Touring.
Here are some of the similar trails you would like to try.
```

Trail_Name <chr>	State <chr>	Length(mts) <dbl>	Elevation(mts) <dbl>
Cades Cove Loop Road	Tennessee	16254.33	210.9216
Cades Cove Loop Road	Tennessee	16254.33	210.9216
Cactus Forest Loop Drive	Arizona	16576.20	213.9696
Brooklyn Mine OHV Trail	California	16737.14	268.8336
Sovereign OHV Loop	Utah	16898.07	282.8544
Capitol Reef Scenic Drive	Utah	12713.79	215.7984

Scenario 2: Recreation Trails

Input: Ice Lake and Little Gibbon Falls Loop

Output:

```
These are Recreation Trails.
These are primarily used for fun activities such as paddle sports,caneoing,skiing and
kayaking.
Here are some similar trails.
```

Trail_Name <chr>	State <chr>	Length(mts) <dbl>	Elevation(mts) <dbl>
Ice Lake and Little Gibbon Falls Loop	Wyoming	5954.558	82.9056
Ice Lake and Little Gibbon Falls Loop	Wyoming	5954.558	82.9056
Wawona Meadow Loop Trail	California	5793.624	74.9808
Carriage and Around Mountain Road Loop	Maine	6598.294	81.9912
Boston Run Trail	Ohio	4988.954	70.7136
Panther Peak Wash - Roadrunner Loop	Arizona	6759.228	57.9120

Scenario 3: Strenuous Trails

Input: Ice Lake and Little Gibbon Falls Loop

Output:

```
These are Strenous Trails and Here are some Similar trails.
These are designed for hiking,backpacking and camping.
```

Trail_Name <chr>	State <chr>	Length(mts) <dbl>	Elevation(mts) <dbl>
South Kaibab to North Kaibab to Bright Angel Trail	Arizona	65822.01	3578.962
South Kaibab to North Kaibab to Bright Angel Trail	Arizona	65822.01	3578.962
Yosemite Valley East Loop	California	65982.94	3649.980
Kearsarge Pass Trail to John Muir Trail Loop	California	68236.02	3702.710
Lost Pass Primitive Trail	Washington	70650.03	3567.989
Mineral King Loop	California	60350.25	3397.910

Scenario 4: Easy Trails

Input: Landscape Arch Trail

Output:

```
These are Easy Trails.
Just relax with a small stroll and enjoy bird watching.
Here are some similar Trails where you can enjoy with kids
```

Trail_Name <chr>	State <chr>	Length(mts) <dbl>	Elevation(mts) <dbl>
Landscape Arch Trail	Utah	3057.746	78.9432
Landscape Arch Trail	Utah	3057.746	78.9432
McDonald Creek Via Johns Lake	Montana	3057.746	79.8576
Split Rock Loop Trail	California	3057.746	76.8096
Great Head Trail (Full Loop)	Maine	2896.812	80.7720
Exit Glacier Trail	Alaska	2896.812	81.9912

Scenario 5: Moderate Trails

Input: Length: 5149.9, Elevation Gain= 191.719

Output:

```
You have selected the Moderate Trails.
These are mainly for hiking,bird watching and to enjoy wildlife.Below are some similar
Trails
```

Trail_Name <chr>	State <chr>	Length(mts) <dbl>	Elevation(mts) <dbl>
High Dune Trail	Colorado	4828.020	191.7192
Fort Bottom Ruin Trail	Utah	5471.756	196.9008
Eureka Dunes	California	5632.690	187.7568
Sunset Trail: Sugarloaf Mountain Section	Arkansas	4667.086	183.7944
High Lakes Loop Trail	Washington	5149.888	169.7736
Norumbega Mountain and Hadlock Ponds Loop Trail	Maine	4828.020	214.8840

Scenario 6: Hard Trails

Input: Half Dome Trail

Output:

```
Hard and Challenging Trails.Here are some of the Hard trails.
Dont forget to backpack and carry lots of fluids.
These trails are mainly for Hiking and enjoying wildlife
```

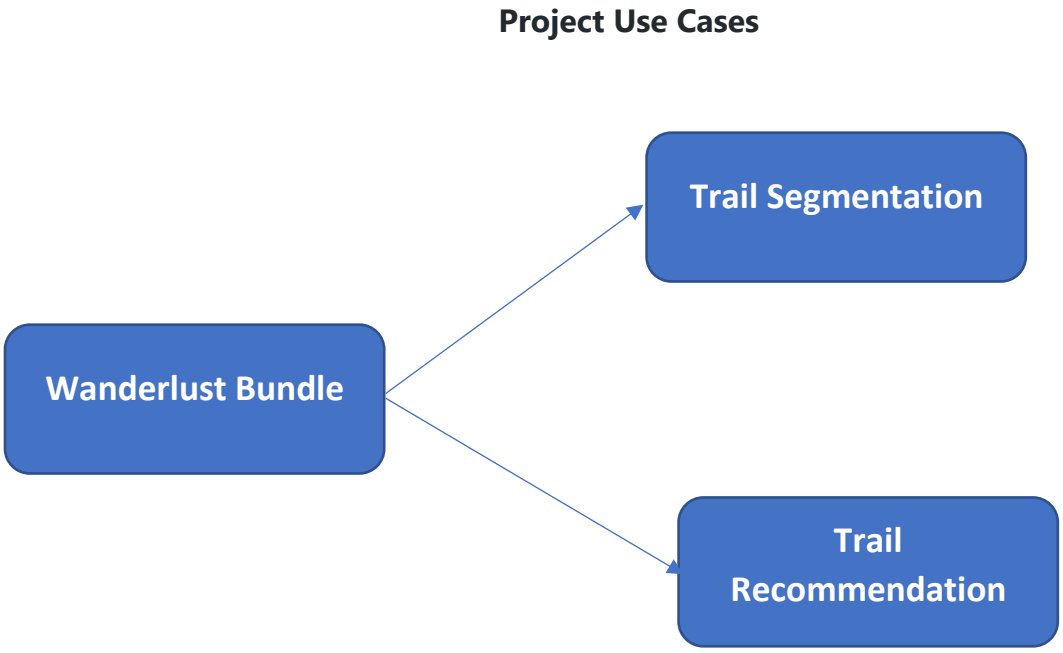
Trail_Name <chr>	State <chr>	Length(mts) <dbl>	Elevation(mts) <dbl>
Half Dome Trail	California	23818.23	1573.987
Half Dome Trail	California	23818.23	1573.987
Appalachian Trail: Fontana Lake to Mount Squires	Tennessee	24622.90	1590.751
The Loop and Garden Wall	Montana	22852.63	1577.950
Hermit Trail	Arizona	24783.84	1567.891
Tonto Trail: New Hance Trail to Grandview Point	Arizona	25105.70	1569.720

Model Accuracy :

Accuracy.xlsx

The above attached Excel file has few trails which are compared with the All Trails website for model accuracy. I have tested manually for 40 trails and accuracy for the corresponding trails is 87.5%

Conclusion:



In a nutshell, I would like to conclude that Wanderlust Bundle will help you in two scenarios

Trail Segmentation: in which the entire dataset has been segmented into 6 clusters, which can be used by the business based on the use cases.

Trail Recommendation: is one such use case of the segmentation where similar trails are recommended along with their designated activities based on the user specified trails which can help people to save lot of time.

Please find the code in the below link

https://github.com/JayasriMaditati/jmaditat_64060.git

Final Thoughts:

So many of our daily decisions are made on the basis of recommendations. The list may go on and on, whether it's food, restaurants, movies, or retail merchandise. It's also crucial to know that what is being recommended to us will be something that we will have a high confidence that we know we will like or enjoy. Wanderlust Bundle is created to help people identify user-friendly paths and to save lot of time.

If I had more time and data to develop Wanderlust Bundle, I could have put some additional data into the project that could have been useful. Trail user information may have been integrated during the data collection procedure. Additional types of recommender models, such as an Item Similarity Recommender, could have been explored if user ratings had been available. Additionally, if I have more data on weather details for each trail, recommendations based on weather and safety suggestions could have been provided and would like to create an interactive website for the user input as a future work.

References:

- Data Mining for Business Analytics – Concepts, techniques, and applications in R – Galit Shmueli
- K-Means Clustering Algorithm Modules.
- https://www.alltrails.com/
- How Machine Learning Can Lower the Search Cost for Finding Better Hikes
- Building California Trail Finder - Tom Weichle