# Tugas Kelompok Data Mining

# Laporan Data Mining



Disusun oleh :

* Sultan Arizal Mahing (3311801016)
* Melina Witri (3311801011)
* Iqbal Afif (3311801019)
* M. Reza Pahlefi (3311601014)

Kelas : Informatika 3 A

## Program Studi D3 Teknik Informatika

## Jurusan Teknik Informatika

## Politeknik Negeri Batam

## 2019

# HALAMAN PENGESAHAN

**DATA MINING DARI DATASET ‘buddymove\_holidayiq’**

**DENGAN MENGGUNAKAN ALGORITMA CLUSTERING**

**Disusun oleh:**

**Sultan Arizal Mahing 3311801016**

**Melina Witri 3311801011**

**Iqbal Afif 3311801019**

**M. Reza Pahlefi 3311601014**

Batam, …………….. 2019

Disetujui dan disahkan oleh:

|  |
| --- |
| Dosen pengajar,  **Muhammad Nasrullah**  **NIK/NIP.** |
|  |

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311801016

Nama : Sultan Arizal Mahing

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET ‘buddymove\_holidayiq’

DENGAN MENGGUNAKAN ALGORITMA CLUSTERING

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

**Sultan Arizal Mahing**

**NIM. 3311801016**

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311801011

Nama : Melina Witri

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET ‘buddymove\_holidayiq’

DENGAN MENGGUNAKAN ALGORITMA CLUSTERING

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

**Melina Witri**

**NIM. 3311801011**

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311801019

Nama : Iqbal Afif

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET ‘buddymove\_holidayiq’

DENGAN MENGGUNAKAN ALGORITMA CLUSTERING

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

**Iqbal Afif**

**NIM. 3311801019**

# HALAMAN PERNYATAAN

Dengan ini, saya:

NIM : 3311601014

Nama : M.Reza Pahlefi

adalah mahasiswa Program Studi Teknik Informatika Politeknik Negeri Batam menyatakan bahwa Tugas Praktikum dengan judul:

DATA MINING DARI DATASET ‘buddymove\_holidayiq’

DENGAN MENGGUNAKAN ALGORITMA CLUSTERING

disusun dengan:

1. tidak melakukan plagiat terhadap naskah karya orang lain
2. tidak melakukan pemalsuan data
3. tidak menggunakan karya orang lain tanpa menyebut sumber asli atau tanpa ijin pemilik

Jika kemudian terbukti terjadi pelanggaran terhadap pernyataan di atas, maka saya bersedia menerima sanksi apapun termasuk pencabutan gelar akademik.

Lembar pernyataan ini juga memberikan hak kepada Politeknik Negeri Batam untuk mempergunakan, mendistribusikan ataupun memproduksi ulang seluruh hasil Tugas Praktikum ini.

Batam, .......... 2019

**M. Reza Pahlefi**

**NIM. 3311601014**

# Link Repository Github

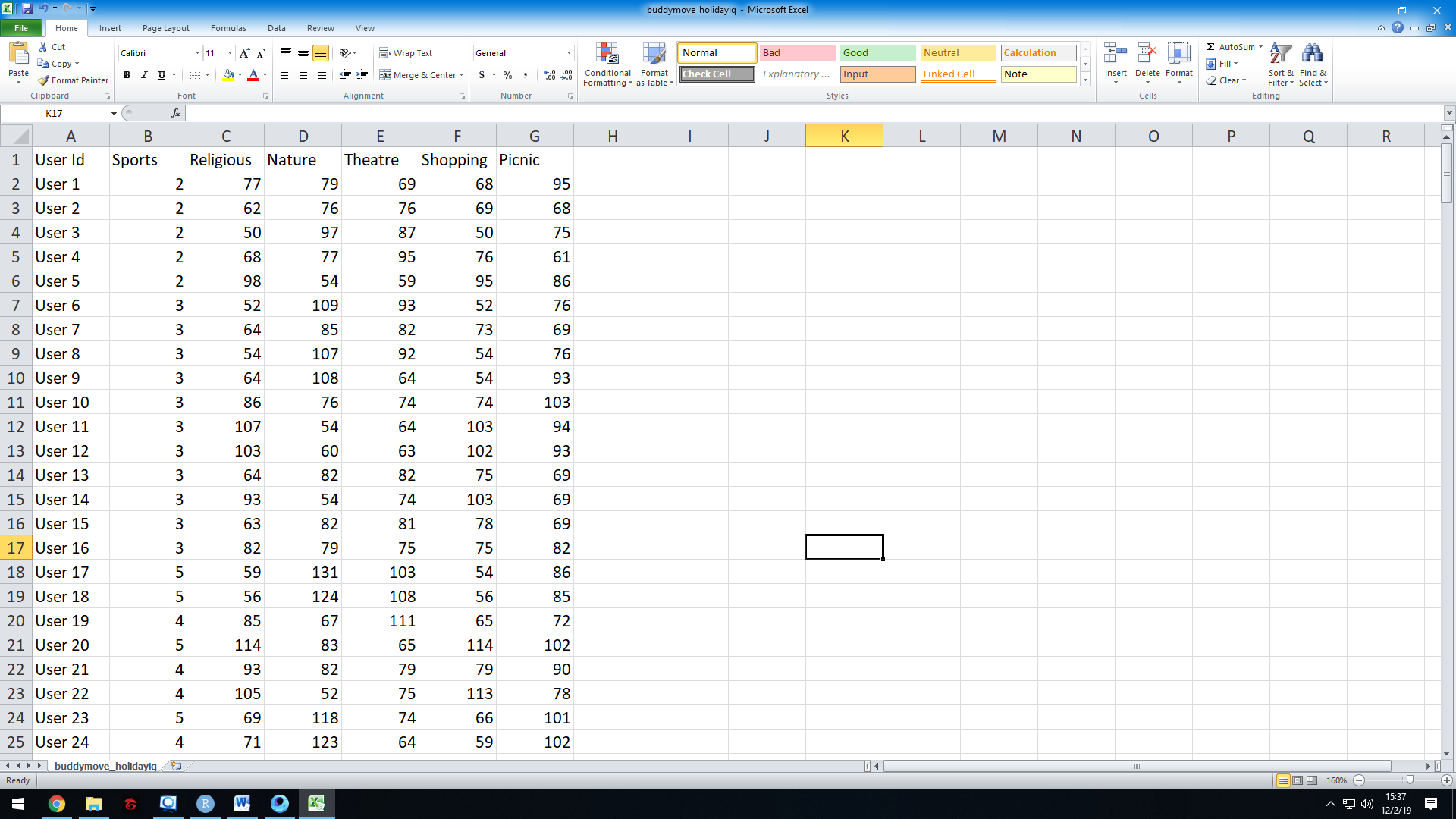
<https://github.com/arizalsultan25/TugasDM-if3a-kel9>

# Link Anggota Kelompok

1. Sultan Arizal Mahing (<https://github.com/arizalsultan25>)
2. Melina Witri (<https://github.com/MelinaWitri>)
3. Iqbal Afif (<https://github.com/iqbal3311801019>)
4. M. Reza Pahlefi (<https://github.com/pahlefi30>)

# Dataset Yang Digunakan

Dataset yang kami gunakan adalah dataset yang bernama “*Buddy Move Dataset”*. Dataset tersebut terdiri dari 249 data. Dengan isi sebagai berikut :



Gambar 01. Isi dari file *buddymove\_holidayiq.csv*

Dataset *Buddy Move* ini memiliki 8, attribut diantaranya :

1. User Id : berisi user id yang unique.
2. Sports : berisi berapa banyak jumlah review yang dilakukan oleh user tersebut

terhadap tempat tempat olah raga.

1. Religious : berisi berapa banyak jumlah review yang dilakukan oleh user tersebut

terhadap Institusi keagamaan.

1. Nature : berisi berapa banyak jumlah review yang dilakukan oleh user tersebut

terhadap pemandangan alam, seperti danau, pantai, dll.

1. Theatre : berisi berapa banyak jumlah review yang dilakukan oleh user tersebut

terhadap bioskop atau tempat hiburan lainnya.

1. Shopping : berisi berapa banyak jumlah review yang dilakukan oleh user tersebut

terhadap mall dan pusat perbelanjaan lainnya.

1. Picnic : berisi berapa banyak jumlah review yang dilakukan oleh user tersebut

terhadap tempat tempat picnic seperti taman / picnic spot lainnya.

Dataset *Buddy Move*  ini merupakan dataset yang sudah *clean*. Dalam artian tidak terdapat data abnomali data seperti, data yang kosong, data tidak konsisten atau data outbound. Sehingga tidak perlu dilakukan proses *cleaning*. Dataset ini dapat di download di halaman : <https://archive.ics.uci.edu/ml/datasets/BuddyMove+Move+Data+Set>

# Algoritma Yang Digunakan (*Clustering*)

Clustering atau klasterisasi adalah metode pengelompokan data. Menurut Tan, 2006 clustering adalah sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum.

Serta tipe data pada dataset yang kami gunakan cocok dengan syarat dilakukannya metode clustering, Berikut adalah syarat-syarat dilakukannya metode clustering menurut Han dan Kamber, 2012, syarat sekaligus tantangan yang harus dipenuhi oleh suatu algoritma *clustering* adalah:

1. Skalabilitas

Suatu metode *clustering* harus mampu menangani data dalam jumlah yang besar. Saat ini data dalam jumlah besar sudah sangat umum digunakan dalam berbagai bidang misalnya saja suatu database. Tidak hanya berisi ratusan objek, suatu database dengan ukuran besar bahkan berisi lebih dari jutaan objek.

1. Kemampuan analisa beragam bentukdata

Algortima klasteriasi harus mampu dimplementasikan pada berbagai macam bentuk data seperti data nominal, ordinal maupun gabungannya.

1. Menemukan*cluster*dengan bentuk yang tidak terduga

Banyak algoritma *clustering* yang menggunakan metode *Euclidean*atau *Manhattan*yang hasilnya berbentuk bulat. Padahal hasil *clustering* dapat berbentuk aneh dan tidak sama antara satu dengan yang lain. Karenanya dibutuhkan kemampuan untuk menganalisa*cluster*dengan bentuk apapun pada suatu algoritma *clustering*.

1. Kemampuan untuk dapat menangani noise

Data tidak selalu dalam keadaan baik. Ada kalanya terdapat data yang rusak, tidak dimengerti atau hilang. Karena system inilah, suatu algortima *clustering* dituntut untuk mampu menangani data yang rusak.

1. Sensitifitas terhadap perubahan input

Perubahan atau penambahan data pada input dapat menyebabkan terjadi perubahan pada*cluster*yang telah ada bahkan bisa menyebabkan perubahan yang mencolok apabila menggunakan algoritma *clustering* yang memiliki tingkat sensitifitas rendah.

1. Mampu melakukan *clustering* untuk data dimensi tinggi

Suatu kelompok data dapat berisi banyak dimensi ataupun atribut. Untuk itu diperlukan algoritma *clustering* yang mampu menangani data dengan dimensi yang jumlahnya tidak sedikit.

1. Interpresasi dan kegunaan

Hasil dari *clustering* harus dapat diinterpretasikan dan berguna.

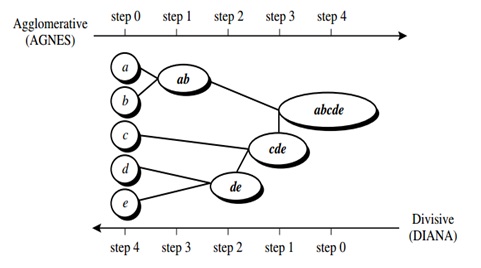
Clustering itu sendiri memiliki bermacam macam metode, diantaranya adalah K-Means dan Agglomerative.

1. **K-means Clustering**

K-means clustering adalah salah satu bentuk atau metode dalam algoritma clustering yang paling saderhana dan umum digunakan. Algoritma K-means mendefinisikan centroid atau pusat cluster dari cluster menjadi rata-rata point dari cluster tersebut.Dalam penerapan algoritma k-Means, jika diberikan sekumpulan data X = {x1, x2, …,xn} dimana xi = (xi1, xi2, …, xin) adalah ystem dalam ruang real Rn, maka algoritma k-Means akan menyusun partisi X dalam sejumlah k cluster (a priori). Setiap cluster memiliki titik tengah (centroid) yang merupakan nilai rata rata (mean) dari data-data dalam cluster tersebut. Tahapan awal, algoritma k-Means adalah memilih secara acak k buah obyek sebagai centroid dalam data. Kemudian, jarak antara obyek dan centroid dihitung menggunakan Euclidian distance. Algoritma k-Means secara iterative meningkatkan variasi nilai dalam dalam tiap tiap cluster dimana obyek selanjutnya ditempatkan dalam kelompok yang terdekat, dihitung dari titik tengah klaster. Titik tengah baru ditentukan bila semua data telah ditempatkan dalam cluster terdekat. Proses penentuan titik tengah dan penempatan data dalam cluster diulangi sampai nilai titik tengah dari semua cluster yang terbentuk tidak berubah lagi.

1. **Agglomerative Hierarchical Clustering**

Pada*hierarchical clustering*data dikelompokkan melalui suatu bagan yang berupa hirarki, dimana terdapat penggabungan dua grup yang terdekat disetiap iterasinya ataupun pembagian dari seluruh set data kedalam *cluster.*

[](https://socs.binus.ac.id/files/2017/03/edy-3.jpg)

Gambar 2. Hierarchical Clustering  
(Sumber:Han dkk, 2012)

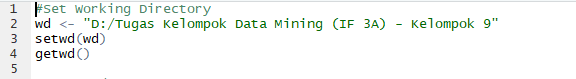
Langkah melakukan *Hierarchical clustering*:

1. Identifikasi *item* dengan jarak terdekat
2. Gabungkan *item* itu kedalam satu*cluster*
3. Hitung jarak antar*cluster*
4. Ulangi dari awal sampai semua terhubung

# Proses Data Mining Menggunakan R Programming

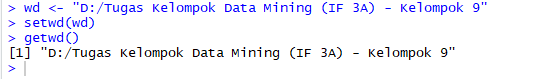
Langkah langkah melakukan data mining menggunakan R programming & Rstudio :

1. **Algoritma K-Means**
2. Setting *working directory :*



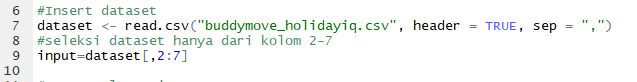
Gambar 3. Source Code langkah 1

Hasil running :



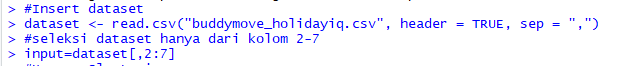
Gambar 4. Hasil Running langkah 1

1. Insert dataset



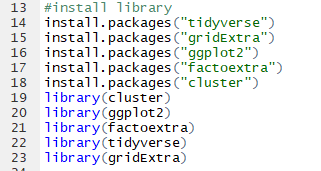
Gambar 5. Source Code langkah 2

Hasil running :



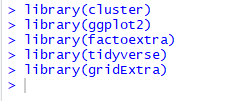
Gambar 6. Hasil Running langkah 2

1. Install package dependency



Gambar 7. Source code langkah 3

Hasil Running :



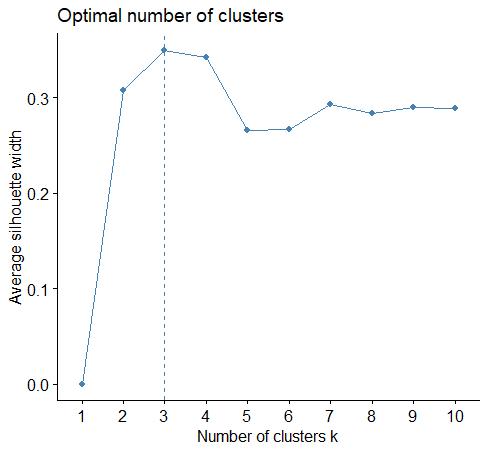
Gambar 8. Hasil Running langkah 3

1. Menentukan jumlah clustering kmeans dengan metode Silhouette :



Gambar 9. Source Code langkah 3

Hasil Running :



Gambar 10. Hasil running langkah 4

1. Membuat variable Kmeans



Gambar 11. Source Code Langkah 5

Hasil Running :



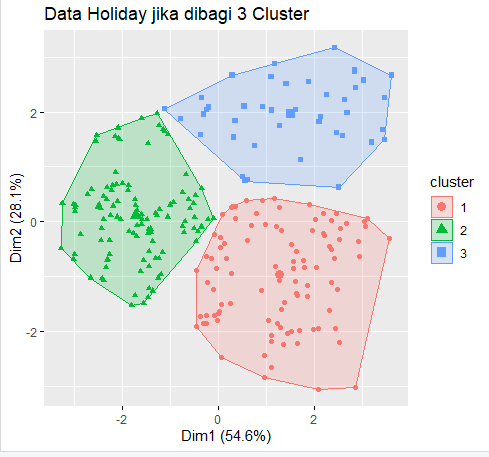
Gambar 12. Hasil running langkah 5

1. Menampilkan hasil cluster



Gambar 13. Source Code langkah 6

Hasil Running :



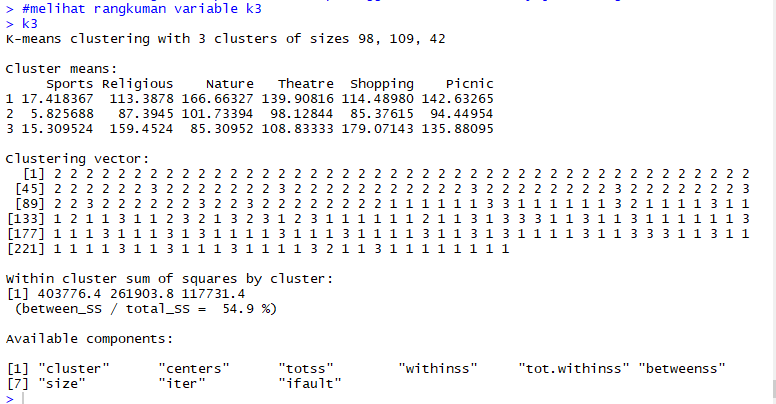
Gambar 14. Hasil running langkah 6

1. Melihat Kesimpilan dari variable k-means



Gambar 15. Source Code langkah 7

Hasil Running :

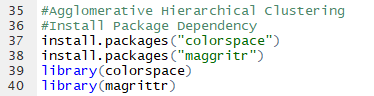


Gambar 16. Hasil Kesimpulan langkah 7

1. **Algoritma Aggloromerative Hierarchical Clustering**

Langkah-langkah melakukan clustering dengan AHC :

1. Install Package Dependency untuk algoritma AHC



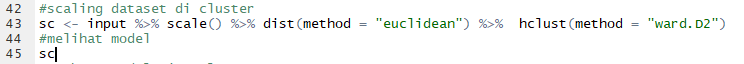
Gambar 17. Source Code langkah 1 AHC

Hasil Running :



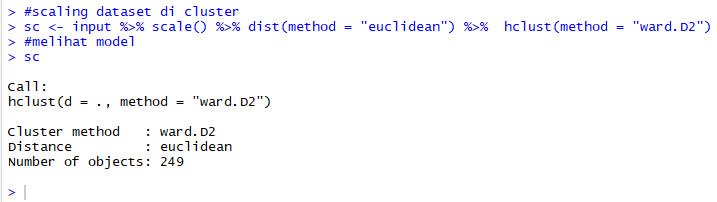
Gambar 18. Hasil running langkah 1 AHC

1. Membuat model dan melihat model tersebut



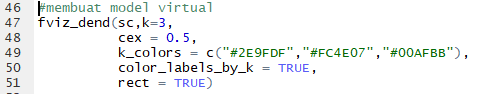
Gambar 19. Source Code langkah 2 AHC

Hasil Running :



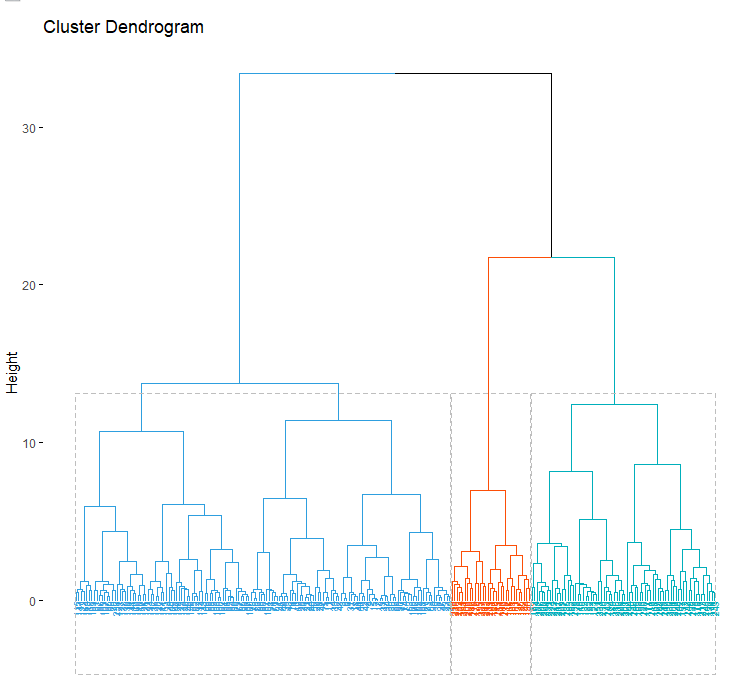
Gambar 20. Hasil running langkah 2 AHC

1. Membuat hierarchy dari model yang telah dibuat



Gambar 21. Source Code Langkah 3 AHC

Hasil Running :



Gambar 22. Hasil Running Langkah 3 AHC

# Kesimpulan

Dataset ‘buddymove\_hollydayiq’ tersebut setelah dilakukan data mining dengan menggunakan metode clustering. Menciptakan 3 buah cluster yang masing masing berisi :

1. Cluster 1 : berisi 98 data
2. Cluster 2 : berisi 102 data
3. Cluster 3 : berisi 42 data



Gambar 23. Ukuran dari cluster yang ada

# Daftar Pustaka

<https://socs.binus.ac.id/2017/03/09/clustering/>

https://medium.com/@16611021/analisis-cluster-non-hirarki-k-means-menggunakan-r-c5ee35de2f14