King Saud University

College of Computer and Information Sciences

Information Technology department

IT 326: Data Mining

Course Project

Analysis of The Google Play Store Dataset

Project Report final

Group #: 2 LAB Day-Time: Wed 10-12

Group members:

Name	ID	Section
Maryam AlAli	436202235	
Maha AlMutawaa	438202246	
Shahad alshibry	438201590	
Noura AlSubaiee	438202523	

4/11/2020

Contents

1 Problem 4

2 Data Mining Task 4

3 Data 4

4 Data preprocessing 6

5 Data Mining Technique 10

6 Evaluation and Comparison 11

7 Extra Clustering methods: 17

8 Findings 19

9 Code 19

10 References 27

11 Tasks Distribution 27

1 Problem

We focus on analyzing Google Play store, the largest Android app store that provides a wide collection of data on features (ratings, price and number of downloads). The overall objective of this analysis effort is to provide in-depth insight about real properties of app repositories in general. This allows us to draw a comprehensive picture of current situation of app market in order to help application developers to understand customers desire and attitude and the trend in the market. The availability of this rich source of information in a single software repository provides a unique opportunity to analyze and understand the relations between these sorts of interrelated data.

2 Data Mining Task

We identify clusters of similar app and then examine the association between characteristics of these clusters and some features of interest. For instance, we would like to know if applications placed in the same category are also functionally similar. In order to find answers for these queries, we should construct clusters of similar applications where the similarity is derived from latent topic models extracted from application description.

3 Data

Since we have a huge amount of Data, we have different types of data which is factor(which is categorical variable that can be either numeric or string variable) and numeric .

We convert the factor data to be numeric so that we can analysis and do many operations in our dataset in efficient way, we got our data set from Kaggle.com. We selected a dataset of 10842

rows and 13 columns, it consists 13 attributes which are: App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating, Genres Last Updated, Current Version, Android Version.

data attribute	type
App ,Category, PriceReviews ,Size, Installs ,Type,Content.Rating, Genres, Last.Updated,Current.Ver ,Android.Ver	factor
Rating	numeric

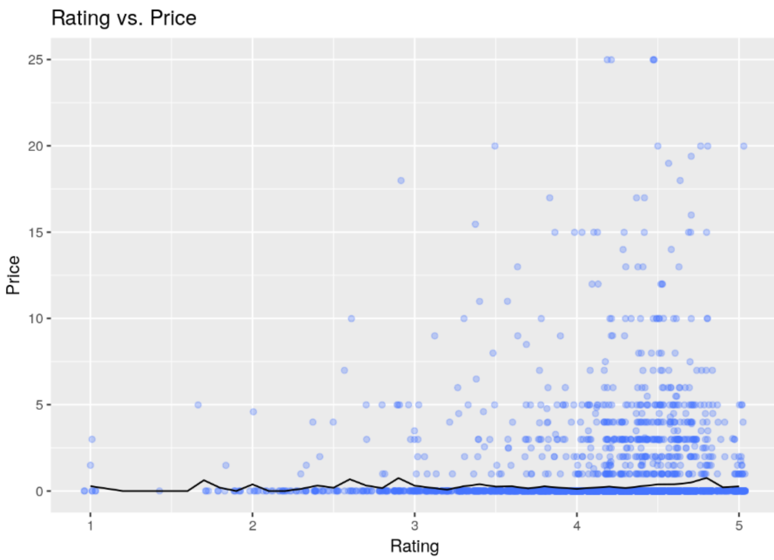


Figure1

In figure 1, Here we see that the mean Price is very close to 0 this is because most of the apps are free. To get a better understanding of the relationship between rating and Price we need to filter by type we're going to do this in the multivariate section with type as the third variable

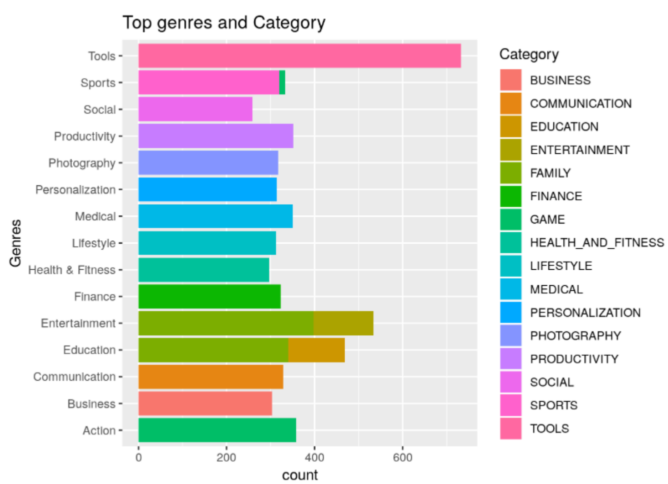
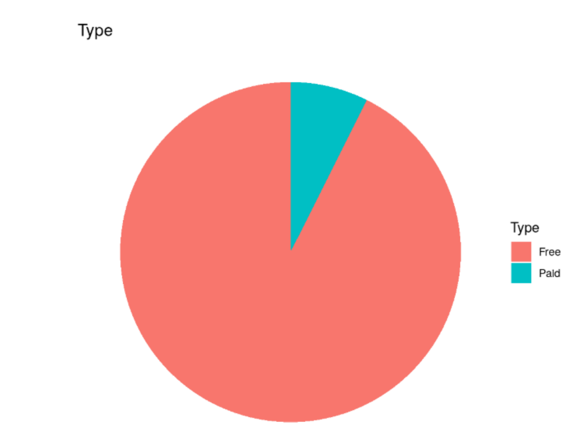


Figure 2

In figure 2, we see that a genre can fall under different categories, for example most of the apps in the entertainment genre are in the family category although there is an entertainment category. We see the same thing with the education genre; most of the apps in the education genre are under the family category although there is a category for educational apps.



In figure 3, There are a lot more free apps in the app store than paid apps. But we need to keep in mind that some of these free apps have in app purchases.

4 Data preprocessing

We choose to replace missing values in rating because we noticed that a-lot of missing value in our dataset , So we decided to replace them to increase the accuracy of the “Data visualization” while in another hand we don’t need to use outliers because we think that it will not add any significant to our Dataset so we decided to remove them. Finally, we encoded and categorized the data so it will be easier to deal with it in the coming phases.

1) Replace the missing value in rating with (0.0)

	App	Category	Rating	Reviews	Size
133	Border Ag & Energy	BUSINESS	NaN	0	12M
134	Ag-Pro Companies	BUSINESS	NaN	0	45M
135	West Central Ag	BUSINESS	NaN	3	1.7M
136	United Ag Cooperative	BUSINESS	NaN	0	4.2M
137	Ag Valley Cooperative	BUSINESS	5.0	6	7 4.2M
138	Ag-Power	BUSINESS	NaN	0	47M
139	i am rich	BUSINESS	3.9	213	2.9M
140	Create apps fast with beautiful design and no code	BUSINESS	3.7	23729	24M
141	Resume Builder Free, 5 Minute CV Maker & Templates	BUSINESS	4.4	72202	6.7M
142	AO-EVENT	BUSINESS	NaN	0	42M
143	AP Mobile 104	BUSINESS	NaN	0	14M

Before □Rating column have (NaN).

	App	Category	Rating	Reviews
42	ElejaOnline DF	BUSINESS	0.0	0
43	DG Monitor	BUSINESS	0.0	1
44	DN Advanced Service Coder	BUSINESS	0.0	0
45	DN Snacks	BUSINESS	0.0	0
46	EG Mantenimiento	BUSINESS	0.0	1
47	EO GSEA	BUSINESS	0.0	1
48	23rd QM BDE EO	BUSINESS	0.0	0
49	EU GDPR RiskCalc	BUSINESS	0.0	1
50	EU Whoiswho	BUSINESS	0.0	0
51	EW Manager	BUSINESS	0.0	0
52	EW Login	BUSINESS	0.0	0

After □Rating column after changing to (0.0).

2) *Remove outliers*

Filter

	App	Category	Rating	Reviews	Size	Ir
1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	
2	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	
3	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	87510	8.7M	
4	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	
5	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	
6	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	
7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	
8	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	
9	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	
10	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	
11	Text on Photo - Fonteee	ART_AND_DESIGN	4.4	13880	28M	

Showing 1 to 14 of 10,841 entries, 13 total columns

Before removing outlier, we had 10,841 entries.

Filter

Category	Rating	Reviews	Size	Installs	Type	Price	Content.Rating	Genres
ART_AND_DESIGN	4.1	1183	19M	10,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	3.9	5924	14M	500,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.7	5681	8.7M	5,000,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.5	1947	25M	50,000,000+	Free	0	Teen	Art & De
ART_AND_DESIGN	4.3	5924	2.8M	100,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.4	1310	5.6M	50,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	3.8	1464	19M	50,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.1	3385	29M	1,000,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.4	816	33M	1,000,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.7	485	3.1M	10,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.4	834	28M	1,000,000+	Free	0	Everyone	Art & De
ART_AND_DESIGN	4.4	5695	12M	1,000,000+	Free	0	Everyone	Art & De

Showing 1 to 14 of 9,367 entries, 13 total columns

After removing outlier, we have 9,367 entries.

3) Encoding categorical data:

Category	Rating	Reviews	Size	Installs	Type	Price	Content.Rating	Genr
BUSINESS	4.2	5467	Varies with device	5,000,000+	Free	0	Everyone	Bus
BUSINESS	4.4	3771	Varies with device	1,000,000+	Free	0	Everyone	Bus
BUSINESS	4.7	2717	39M	10,000,000+	Free	0	Everyone	Bus
BUSINESS	4.5	113	14M	10,000,000+	Free	0	Everyone	Bus
BUSINESS	4.2	4017	19M	1,000,000+	Free	0	Everyone	Bus
BUSINESS	4.7	365	6.8M	100,000+	Paid	$4.99	Everyone	Bus
BUSINESS	4.8	87	39M	100,000+	Paid	$4.99	Everyone	Bus
BUSINESS	4.1	1973	Varies with device	50,000,000+	Free	0	Everyone	Bus
BUSINESS	4.3	11	35M	100,000,000+	Free	0	Everyone	Bus
BUSINESS	4.4	4282	Varies with device	5,000,000+	Free	0	Everyone	Bus
BUSINESS	4.5	3693	29M	10,000+	Free	0	Everyone	Bus
BUSINESS	4.3	2148	41M	1,000,000+	Free	0	Everyone	Bus

Before □ Type was (free, paid, 0).

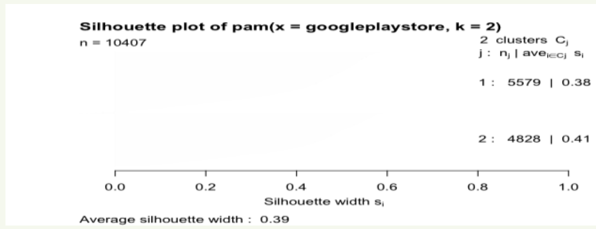
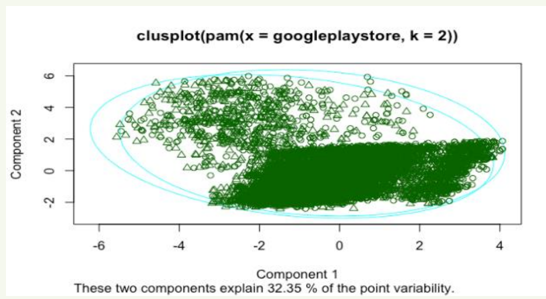
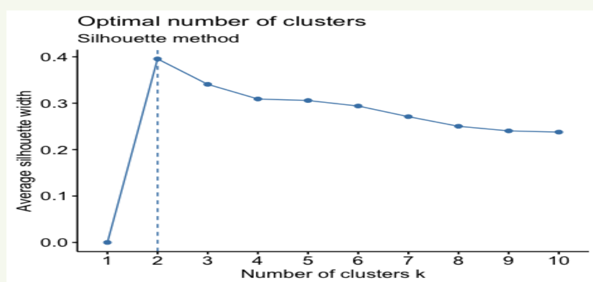
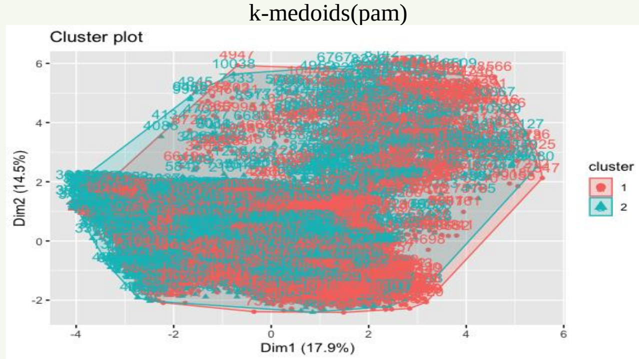
	Category	Rating	Reviews	Size	Installs	Type	Price
	GAME	5.0	1745	16M	1+	2	$0.99
	HEALTH_AND_FITNESS	4.4	3261	2.4M	1,000+	2	$7.99
	GAME	3.8	2372	11M	10,000+	2	$16.99
	FAMILY	4.2	852	9.5M	10,000+	2	$1.20
Grid & ScrapBook	ART_AND_DESIGN	4.1	1183	19M	10,000+	1	0
	ART_AND_DESIGN	3.9	5924	14M	500,000+	1	0
l Themes, Hide Apps	ART_AND_DESIGN	4.7	5681	8.7M	5,000,000+	1	0
	ART_AND_DESIGN	4.5	1947	25M	50,000,000+	1	0
Book	ART_AND_DESIGN	4.3	5924	2.8M	100,000+	1	0
	ART_AND_DESIGN	4.4	1310	5.6M	50,000+	1	0
ke Editor	ART_AND_DESIGN	3.8	1464	19M	50,000+	1	0

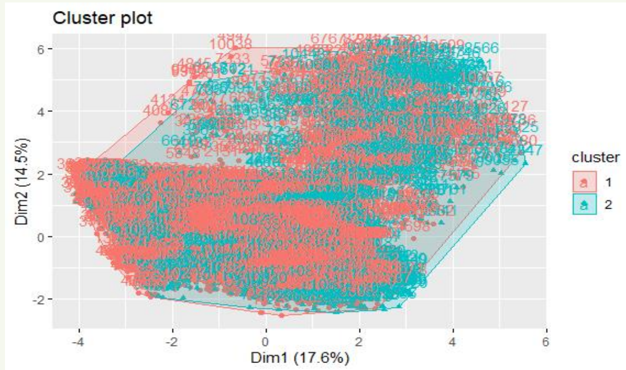
After □ Type is (1, 2, 3).

5 Data Mining Technique

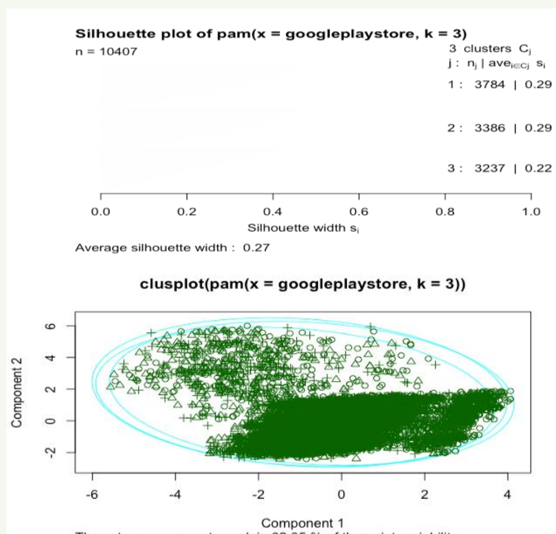
Clustering is an information retrieval technique that puts the items physically together, which are logically similar. In detail, it groups the common characteristics sharing objects together in multi-dimensional space and shows dissimilarity with objects in other clusters with different characteristics at the same time. So, this technique is considered as a tool that can efficiently perform data reduction by creating more manageable subgroups, so that data indexing, filtering, searching, mining and in general, information retrieval becomes easier and faster, we will use the k-means technique because it is the most popular partitioning method, we have to specify the number of clusters using this technique. Using k-means function we need to import these packages: **cluster**, **factoextra** for Determining and Visualizing the Optimal Number of Clusters, **nbClust** and **fpc** to visualizing clustering into our four groups.

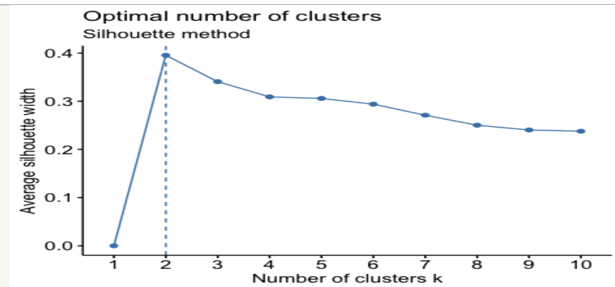
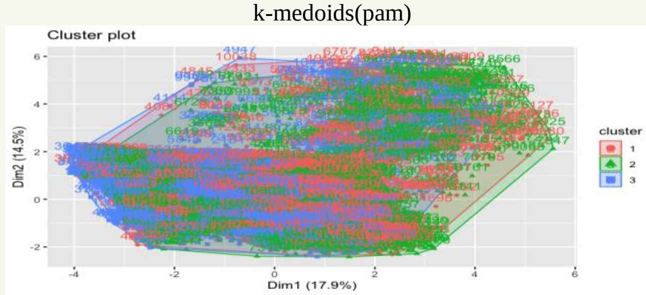
6 Evaluation and Comparison

	Cluster=2
Silhouette width for each cluster	  From these figures, we can see that all the values of the two clusters (0.38, 0.41) are higher and larger than all other clusters, which shows that the corresponding observations are very well clustered than the others.
Silhouette width for all clusters (which shows the perfect cluster =2)	
Visualization pam	k-medoids(pam)  From these figure The association between cluster 1 and 2 appear very strong ,when we get close to (-2,6) we can see a difference in the data from (-2,4)to(0,4)which could be moderately strong

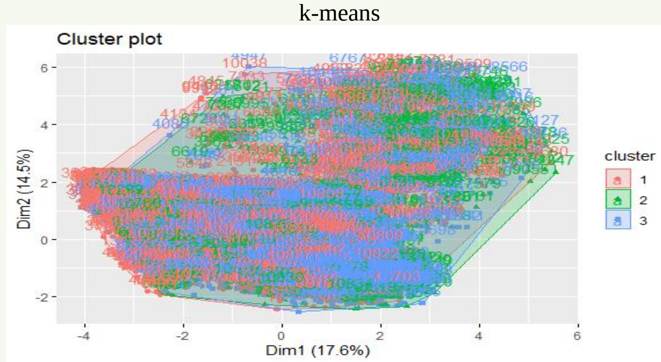
<table>
<tr><td></td><td>K-means

From these figure The association between cluster 1 and 2 appear very strong ,when we get close to (-2,6) we can see a difference in the data from (-2,4)to(0,4)which could be moderately strong

In this cluster as we can see there is no much difference between K-means And K-medoids</td></tr>
</table>

<table>
<tr><td></td><td>Cluster=3</td></tr>
<tr><td>Silhouette width for each cluster</td><td>
From the figure, we can see that every cluster has a small value for the clusters, as long as the value getting close to zero that means the observation lies between two clusters, if the observations has a negative value(which we do not have) are probably placed in the wrong cluster. This is the same in all clusters except cluster 2.</td></tr>
</table>

<table>
<tr>
<td>Silhouette width for all clusters</td>
<td>
</td>
</tr>
<tr>
<td>Visualization</td>
<td>k-medoids(pam)


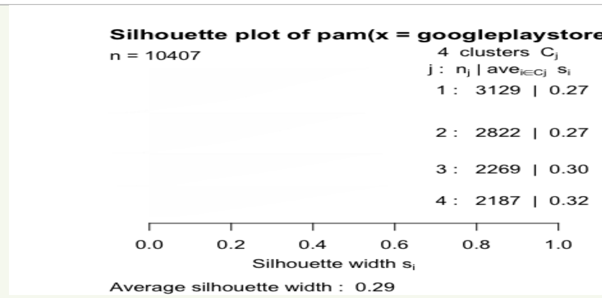
From these figure The association between cluster 1 , 2 and 3 appear very strong from (2.5 , 2.5) and less, and the rest seems moderately strong ,and as we can see there is no unusually data.

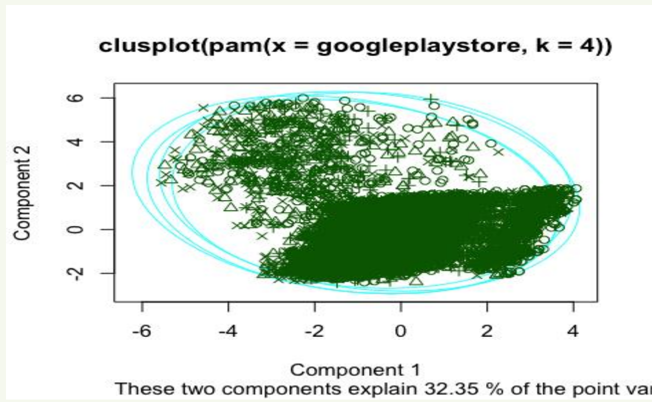
k-means


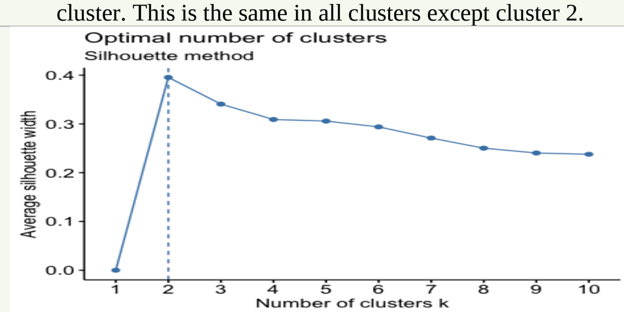
From these figure The association between cluster 1 , 2 and 3 appear very strong from (2.5 , 2.5) and less, and the rest seems moderately strong ,and as we can see there is no unusually data.

In this cluster as we can see there is no much difference between K-means
And K-medoids</td>
</tr>
</table>

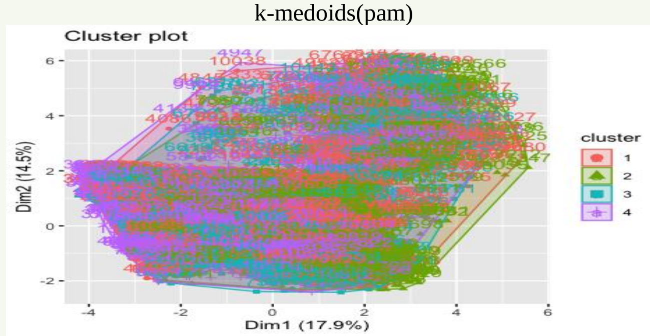
<table>
<tr>
<td></td>
<td>Cluster=4</td>
</tr>
<tr>
<td>Silhouette width for each cluster</td>
<td></td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>




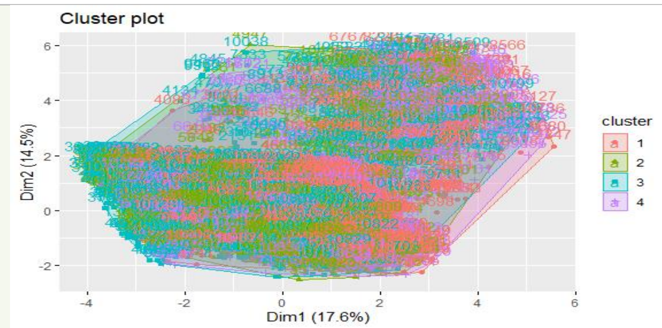
From the figure, we can see that every cluster has a small value for the clusters, as long as the value getting close to zero that means the observation lies between two clusters, if the observations has a negative value(which we do not have) are probably placed in the wrong cluster. This is the same in all clusters except cluster 2.
</td>
</tr>
<tr>
<td>Silhouette width for all clusters</td>
<td>


</td>
</tr>
<tr>
<td>Visualization</td>
<td>
k-medoids(pam)

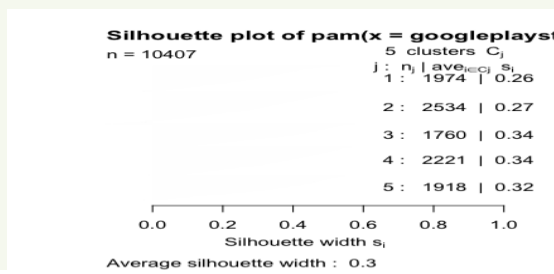


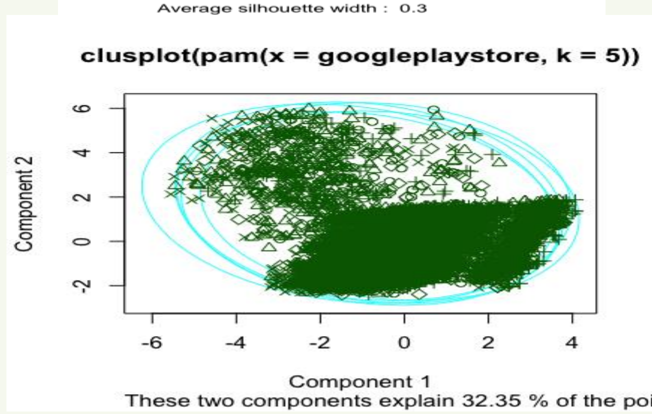
From these figure The association between cluster 1 , 2 , 3and4 appear very strong except the data between x(-3,0)and y(2.5,6) is moderately strong ,and as we can see there is no unusually data.

k-means
</td>
</tr>
</table>

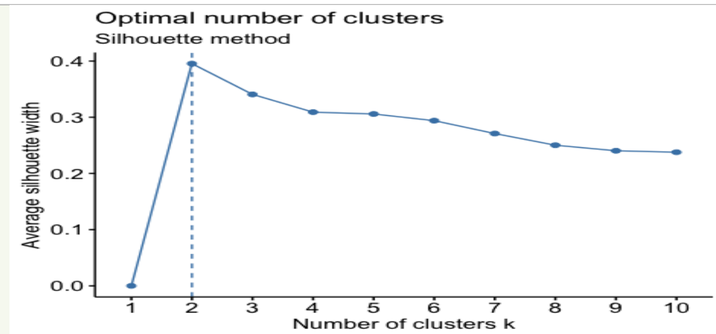
<table>
<tr>
<td></td>
<td>


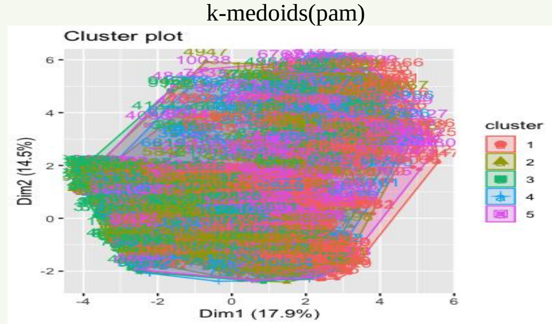
From these figure The association between cluster 1 , 2 , 3and4 appear very strong except the data between x(-3,0)and y(2.5,6) is moderately strong ,and as we can see there is no unusually data.
In this cluster as we can see there is no much difference between K-means
And K-medoids
</td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>Cluster=5</td>
</tr>
<tr>
<td>Silhouette width for each cluster</td>
<td>


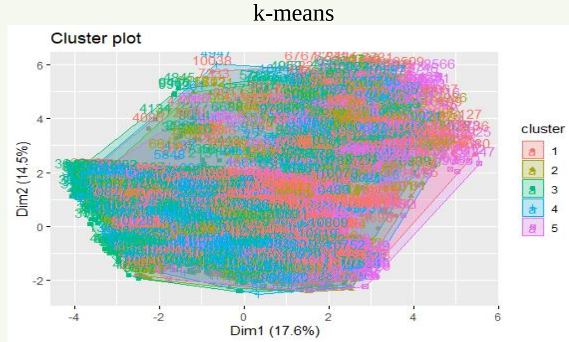


From the figure, we can see that every cluster has a small value for the clusters, as long as the value getting close to zero that means the observation lies between two clusters, if the observations has a negative value(which we do not have) are probably placed in the wrong cluster. This is the same in all clusters except cluster 2.
</td>
</tr>
</table>

<table>
<tr>
<td>Silhouette width for all clusters</td>
<td>


</td>
</tr>
<tr>
<td>Visualization</td>
<td>
k-medoids(pam)


From these figure The association between cluster 1 , 2 , 3,4and 5 appear very strong, and as we can see there is no unusually data.

k-means


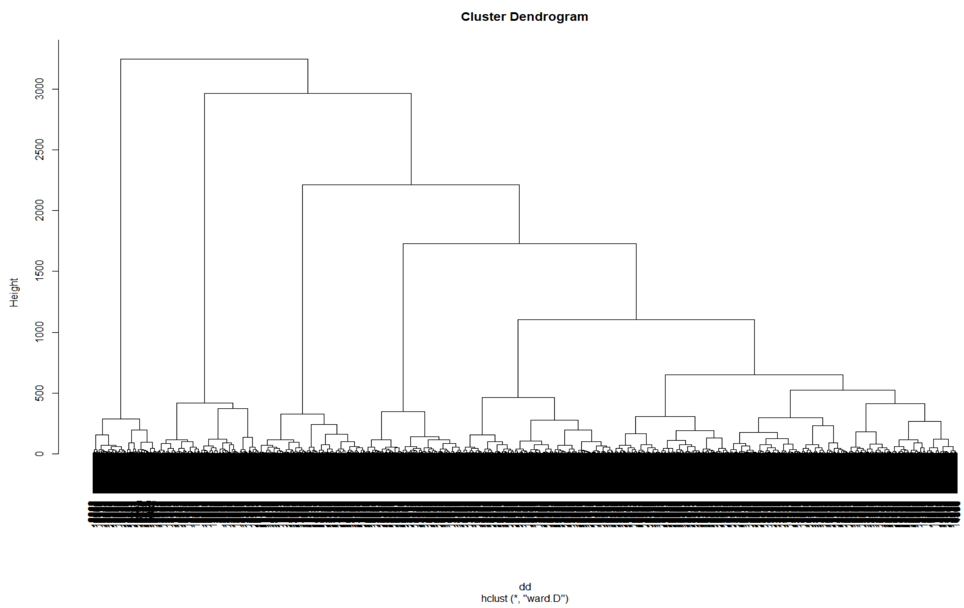
From these figure The association between cluster 1 , 2 , 3,4and 5 appear very strong except the data between x(-3,2.5)and y(2.5,6) is strong ,and as we can see there is no unusually data.
In this cluster as we can see there is no much difference between K-means And K-medoids
</td>
</tr>
</table>

- In all clusters there is no much difference between K-means and K-medoids(pam).

7 Extra Clustering methods

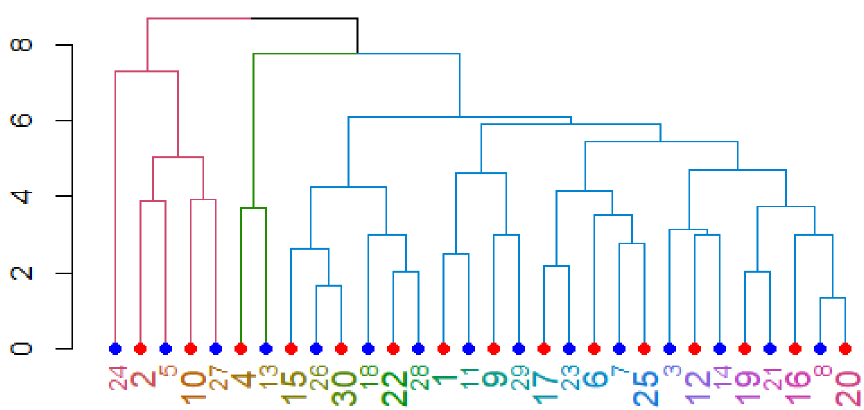
hierarchical cluster:

1) First method



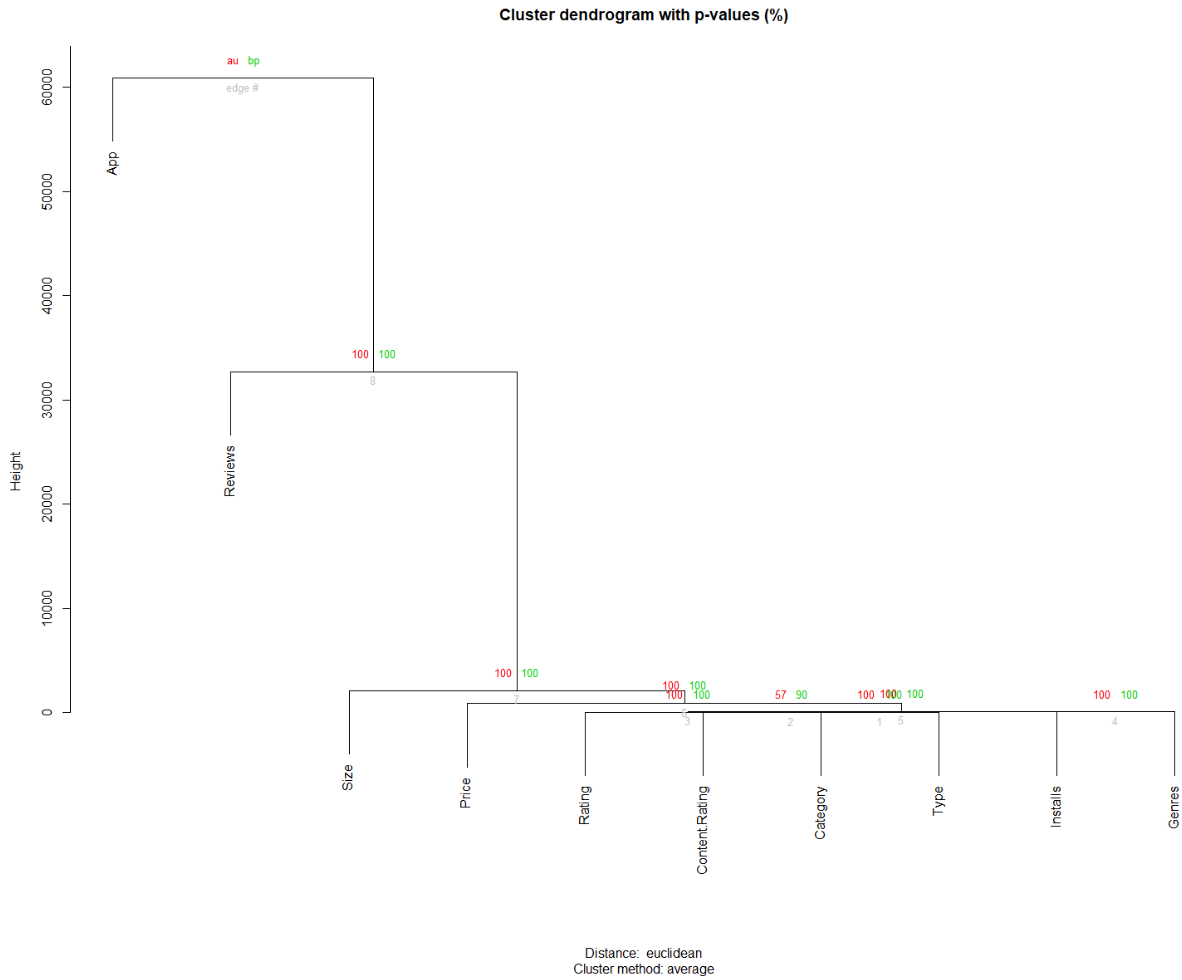
From the figure this hierarchical cluster didn't give us a useful information .

2) Second method



From the figure this hierarchical cluster is better than the previous one but still didn't give us the best visualization.

3) Third method



From the figure this cluster gave us a very useful information through the perfect visualization of our data the figure shows us a strong association between Size ,Price ,Rating,ContentRating ,Category,Type,Installs and Genres when the reviews and apps seems independent from the associated group .

8 Findings

For this project, we took the Google Play Store Data sets and analyzed and processed the data. After the data was transformed into a usable set, we used plots and functions to understand the correlations between features. We then used this knowledge to build the best model we could for all the dataset on the cleaned data set.

We thought finding a decent model would not be too difficult and some values will be true. Instead, we learned that creating a model was not a simple task. We used 4 different clusters to find which one is the best, when the cluster=3 the average of the 3 clusters with these values(0.29,0.29,0.22) was 27% and when the clusters=4 (0.27,0.27,0.30,0.32) the average was 29% and for the cluster 5(0.26,0.27,0.34,0.34,0.32)the average was 30% and for the best cluster, when k=2 the values were (0.38,0.41) and the average was 39%. Since k=2 has the highest average which equals 39% means that it is the best cluster between all the clusters we create. Overall, we can see the average of all clusters not close enough to 1 so our data needs more operations to make it clearer, also we thought we will see different results from what we come up with, that means our result was not clear enough.

9 Code

```
install.packages("dplyr")

install.packages("readr")

install.packages("ggplot2")

install.packages("scales")

install.packages("ggthemes")

install.packages("tidyr")

install.packages("AppliedPredictiveModeling")
```

```
install.packages("outliers")

install.packages("factoextra")

install.packages("cluster")

install.packages("ggfortify")

install.packages("fpc")

install.packages("NbClust")

library(dplyr)

library(readr)

library(ggplot2)

library(scales)

library(ggthemes)

library(tidyr)

library(NbClust)

library(cluster)

library(fpc)

library(outliers)

library(factoextra)

#--------- read -------

help(read.csv)

?read.csv

 googleplaystore<- read.csv(file.choose(),header=TRUE)

#-----------------
```

preprocessing

```
#------------ Rating vs.price scatter plot-------

```{r echo= FALSE, message=FALSE, warning=FALSE, Bivariate_Plots}
```
```

~~~
#plot of rating vs. Price

ggplot(aes(x = Rating , y = Price), data = df)+

  geom_jitter(alpha = 0.3, color = 'royalblue1')+

  ylim(0,25)+

  geom_line(stat = 'summary', fun.y = mean)+  ggtitle('Rating vs. Price')

```

#--------- Plot for top genres vs. category  --------------

ggplot(aes (x = Genres), data = topgenres)+

  geom_bar(aes(fill = Category))+

  coord_flip()+

  ggtitle('Top genres and Category') ```


#--------- Pie chart for Type ---------------


```{r echo= FALSE,message=FALSE, warning=FALSE}

#There are two types paid and unpaid

df_type = subset(df, (Type == 'Free' | Type == 'Paid'))

temp <- df_type%>%

  group_by(Type)%>%

  summarise(n = n())

#pie chart

ggplot(aes(x = '', y = n, fill = Type), data = temp )+

  geom_bar(stat = 'identity')+
~~~

```
  coord_polar('y', start = 0)+

  theme_void()+

  ggtitle('Type')

googleplaystore$Rating[is.nan(googleplaystore$Rating)]<-0.0

outrev = outlier(googleplaystore$Rating, logical = TRUE)

sum(outrev)

Find_outlier = which(outrev ==TRUE, arr.ind = TRUE)

googleplaystore= googleplaystore[-Find_outlier,]

#-----------------

outGen = outlier(googleplaystore$Genres, logical = TRUE)

sum(outGen)

Find_outlier = which(outGen ==TRUE, arr.ind = TRUE)

googleplaystore= googleplaystore[-Find_outlier,]

#------------------

outCon = outlier(googleplaystore$Content.Rating, logical = TRUE)

sum(outCon)

Find_outlier = which(outCon ==TRUE, arr.ind = TRUE)

googleplaystore= googleplaystore[-Find_outlier,]

outcat = outlier(googleplaystore$Category, logical = TRUE)

sum(outcat)

Find_outlier = which(outcat ==TRUE, arr.ind = TRUE)

googleplaystore= googleplaystore[-Find_outlier,]

outrevi = outlier(googleplaystore$Reviews, logical = TRUE)
```

sum(outrevi)

Find_outlier = which(outrevi ==TRUE, arr.ind = TRUE)

googleplaystore= googleplaystore[-Find_outlier,]

##---------------------

##Data mining task

#----------------------

googleplaystore$Category <- sapply(googleplaystore$Category,as.numeric)

googleplaystore$Reviews <- sapply(googleplaystore$Reviews,as.numeric)

googleplaystore$Size <- sapply(googleplaystore$Size,as.numeric)

googleplaystore$Installs <- sapply(googleplaystore$Installs,as.numeric)

googleplaystore$Type <- sapply(googleplaystore$Type,as.numeric)

googleplaystore$Price <- sapply(googleplaystore$Price,as.numeric)

googleplaystore$Content.Rating <- sapply(googleplaystore$Content.Rating,as.numeric)

googleplaystore$Genres <- sapply(googleplaystore$Genres,as.numeric)

googleplaystore$Last.Updated <- sapply(googleplaystore$Last.Updated,as.numeric)

googleplaystore$Current.Ver<- sapply(googleplaystore$Current.Ver,as.numeric)

googleplaystore$Android.Ver<- sapply(googleplaystore$Android.Ver,as.numeric)

googleplaystore$App<- sapply(googleplaystore$App,as.numeric)

#------------------

kmeans2.result <- kmeans(googleplaystore,2)

kmeans2.result

visualize clustering k=2

fviz_cluster(kmeans2.result, data = googleplaystore)

set.seed(8953)

kmeans3.result <- kmeans(googleplaystore, 3)

kmeans3.result

visualize clustering k=3

fviz_cluster(kmeans3.result, data = googleplaystore)

kmeans4.result <- kmeans(googleplaystore, 4)

kmeans4.result

visualize clustering k=4

fviz_cluster(kmeans4.result, data = googleplaystore)

kmeans5.result <- kmeans(googleplaystore, 5)

kmeans5.result

visualize clustering k=5

fviz_cluster(kmeans5.result, data = googleplaystore)

##--------------------------

Evaluation

##--------------------------

group into 4 clusters

pam1.result <- pam(googleplaystore, 2)

visualize clustering k=2

```
fviz_cluster(pam1.result, data = googleplaystore)

plot(pam1.result)


pam2.result <- pam(googleplaystore, 3)

## visualize clustering k=3

fviz_cluster(pam2.result, data = googleplaystore)

plot(pam2.result)


pam3.result <- pam(googleplaystore, 4)

## visualize clustering k=4

fviz_cluster(pam3.result, data = googleplaystore)

plot(pam3.result)


pam4.result <- pam(googleplaystore, 5)

## visualize clustering k=5

fviz_cluster(kmeans2.result, data = googleplaystore)

plot(pam4.result)


##for all clusters

fviz_nbclust(googleplaystore, kmeans, method = "silhouette")+ labs(subtitle = "Silhouette method")
```

Extra Clustering Method

```
Install.package("pvclust")

library(pvclust)


set.seed(1234)

result <- pvclust(googleplaystore[1:100, 1:10], method.dist="euclidean",

                  method.hclust="average", nboot=10)

plot(result)

pvrect(result)


dd <- dist(scale(googleplaystore), method = "euclidean")

hc <- hclust(dd, method = "ward.D")

plot(hc)


dend <- googleplaystore[1:30,-5] %>% scale %>% dist %>%

  hclust %>% as.dendrogram %>%

  set("branches_k_color", k=3) %>% set("branches_lwd", 1.2) %>%

  set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%
```

```
  set("leaves_pch", 19) %>% set("leaves_col", c("blue", "red"))

# plot the dend in usual "base" plotting engine:

plot(dend)
```

10 References

[1] our dataset

https://www.kaggle.com/lava18/google-play-store-apps/activity

[2] (PDF) Mining and analysis of apps in google play. Available from: https://www.researchgate.net/publication/290102532_Mining_and_analysis_of_apps_in_google_play [accessed Nov 24 2019].

11 Tasks Distribution

ID	Name	Responsibilities
Maryam AlAli	436202235	Preprocessing – clustering – presentation
Maha AlMutawaa	438202246	Preprocessing – clustering
Shahad alshabri	438201590	edit phase two- extra Clustering
Nura AlSubaye	438202523	edit phase two- extra Clustering