ANSWER SHEET DAC 2021

DATA ANALYSIS COMPETITION 2021

TEAM NAME	Dedomena
TEAM ID	ID-21-0131
UNIVERSITY	University Of Brawijaya

Analyse Factors Which Influencing Request Of Residence Specially Type Of Cluster with K-Modes Algorithm

Muhammad Zidan Akmaludin Akbar, Theo Credo Situmorang
Industrial Engineering, University Of Brawijaya
Jl. Veteran, Malang, East Java, Indonesia
Zidanaks_@student.ub.ac.id , theocredositumorang@gmail.com

Abstract

In this research, clustering of property's demand data will be described to determine the characteristics of buyers in making choices to determine property according to their needs. The choice of housing is adjusted to environmental preferences which involves understanding the characteristics of people and their environment. Humans are always faced with choices for something, including choosing the location of the house as a place to live. First, there are 364601 data collected from data from the Indonesian Real Estate Brokers Association (AREBI), then the variables will be extracted and cleaned. This data analysis strategy can be used to complement the results of data mining classification analysis used to identify data patterns. In this article, it will be explained that EDA can assist in enriching the results of data analysis and assist in the preprocessing stage of data mining classification. The clustering method chosen is k-modes because it is suitable for use on categorical data. Based on the graph of the elbow method, the appropriate k value is 37 with a cost value of 513965.0.

Keywords : data mining, clustering, k-modes, regency cluster, property, real estate

Introduction

Property growth in Indonesia is relatively new. The Real Estate Industry has existed since the old order government, which was implemented by local governments. The property industry experienced rapid development in the 1980s, after Indonesia entered the Five-Year Development. Real Estate is property that consists of land and everything in it. These buildings can be in the form of residential houses, shop houses, offices, apartments, malls, and so on. Looking at the type, all that is classified as real estate is an immovable asset and has a fairly high value. However, not all real estate is classified as residential, some of which is only used as a place to conduct business activities (Ruegg & Marshall, 1990).

The house is a dwelling built to meet physiological, psychological, and sociological needs (Maslow, 1971). The house can also express the occupants' lifestyle which is influenced by psychological aspects of social, economic, and aesthetic balance (Campbell et al, 1976). Residential property is vacant land or a plot of land that is developed, used or provided for residence, such as single family houses, apartments, flats, and so on. Property, especially housing, apart from being an investment, is also an asset (The Dictionary of Real Estate Appraisal, 2002: 313).

Just like what happened in other countries, the property industry in Indonesia developed after the government gave special attention to the housing sector with the formation of the Minister of Public Housing. The government is the first party to carry out a housing development program, through a program called Perumahan Nasional (Perumnas). The development of the property industry in Indonesia from the demand sIDe is booming along with the increasing purchasing power of the people. From the supply sIDe, more and more funds are available for the property industry provided by the banking sector, especially after the issuance of the deregulation package in the banking sector. The number of banks in Indonesia has increased, reaching more than 200 banks.

One type of housing that is currently developing and popular in Indonesia according to research conducted by Property Consultants Panangian Simanungkalit and Associates (PSA) is cluster-type housing, namely housing that groups the same architectural style of residential buildings (Hill, 1990), intended for the modern society of middle to upper economic class who tend to have a lifestyle. Modern lifestyle is a life that is stylish, effective, efficient, aesthetic, practical, functional, multipurpose and energy efficient (Pasaribu, 2006). The selection of housing units in cluster type housing can be analyzed by market analysis theory based on decisions due to buyer behavior from James H. Myers (1977). According to Rapoport (1977:81) that people will adjust to their preferences to choose a residential environment. The choice of housing is adjusted to environmental preferences which involves understanding the characteristics of people and their environment. Humans are always faced with choices for something, including choosing the location of the house as a place to live.

The potential for people to be able to own real estate is indeed very large because apart from being able to be used as a residence, real estate ownership is also economically profitable. Therefore, it is not surprising that many companies are becoming increasingly competitive when choosing to become developers and enter the property management business. The development of this industry is indeed very fast because every human being needs space or a place, whether it is a place to live, entertain, or even work. This growth then drives growth in the construction industry. The higher the level of demand from the public, both for housing and investment, the higher the growth of the construction industry (Venclauskiene & Snieska, 2009).

Nowadays, the world's startup enterprise has grown rapidly. Indonesia itself has 230 start-ups, 4 of that have grow to be unicorns. Startups have revolutionized diverse commercial sectors, one in every of that is the belongings enterprise specifically for house. It passed off because, in line with the Vice President of Non-Subsidized Mortgage and Consumer Lending Division of PT Bank Tabungan Negara (Persero) Tbk Suryanti Agustinar, the housing quarter skilled increase of 2% withinside the 1/3 area of 2020 and in 2021 it's far projected that financial increase will attain 5% withinside the housing quarter. This is likewise supported through the assertion of the General Chairperson of the Indonesian Real Estate Brokers Association (AREBI) Lukas Bong who stated that the belongings fashion in 2021 is extra in the direction of landed homes in comparison to apartments. This revolution has many advantages for house sellers and customers. With app developer startups now making plans a difficult holiday. All may be executed in only a hand and a count number of minutes. However, does the consumer effortlessly look for the residence in house that suits with their preference? Of path it's far difficult. Therefore, the utility improvement group created a house advice gadget in line with consumer preferences. The facts is accrued to gain statistics that enables the utility improvement group see purchaser status.

Machine learning is a branch of artificial intelligence Machine Learning is a discipline that includes the design and development of algorithms that allow computers to develop behavior based on empirical data, such as from sensor data databases. The learning system can utilize examples (data) to capture the necessary features of the underlying (unknown) probabilities. The data can be seen as an example that illustrates the relationship between the observed variables. A big focus of machine learning research is how to automatically recognize complex patterns and make intelligent decisions based on data. In 1959, Arthur Samuel defined machine learning as a field of study that provIDes the ability to learn without being explicitly programmed. The dominant learning ability is determined by the ability of the software or its algorithm. Machine learning can function to adapt to a new situation, as well as to detect and predict a pattern. Algorithms in machine learning can be grouped based on the input and output expected from the algorithm, namely directed learning, undirected learning, semi-directed learning, and reinforcement learning.

K-modes clustering was first introduced by Huang (1998) as a clustering method which was developed from the k-means method. Therefore, k-modes are efficient as k-means but are used on categorical data.

Machine learning is a branch of artificial intelligence Machine Learning is a discipline that includes the design and development of algorithms that allow computers to develop behavior based on empirical data, such as from sensor data databases. The learning system can utilize examples (data) to capture the necessary features of the underlying (unknown) probabilities. The data can be seen as an example that illustrates the relationship between the observed variables. A big focus of machine learning research is how to automatically recognize complex patterns and make intelligent decisions based on data. In 1959, Arthur Samuel defined machine learning as a field of study that provIDes the ability to learn without being explicitly programmed. The dominant learning ability is determined by the ability of the software or its algorithm. Machine learning can function to adapt to a new situation, as well as to detect and predict a pattern. Algorithms in machine learning can be grouped based on the input and output expected from the algorithm, namely directed learning, undirected learning, semi-directed learning, and reinforcement learning.

K-Modes clustering was first introduced by Huang (1998) as a clustering method which was developed from the k-means method. Therefore, k-modes are efficient as k-means but are used on categorical data.

The Elbow method is a method to determine the right number of clusters through the percentage of the comparison between the number of clusters that will form an elbow at a point . The elbow method plots the value of the cost function produced by different values of k. The different percentage results from each cluster value can be shown using a graphic as the source of information. The principle of this elbow method is to select the cluster value and then add the cluster value to be used as a data model in determining the best cluster. When the cluster value has the greatest decline and forms an angle, the number of cluster values is saID to be the most appropriate value.

Chapter 1

For the success of property development, all marketing activities carried out must lead to consumer desires and consumer needs. Consumer needs in this case are users of the product being marketed, while consumers are buyers or potential buyers. These consumer needs are the basis for starting the implementation of market activities. Entrepreneurs always try to meet the needs of consumers who are the target for the products they produce.

The consumer's desire is a statement about the level of satisfaction he expects from meeting the needs he faces for a product. Desire is always influenced by the culture of the community where consumers are located, especially in formulating objects that can ensure the fulfillment of needs. In a more developed society, the level of desire of the members of the community will be higher and more varied. Of course, in such a company competes in provIDing satisfaction to consumer desires and companies that are successful in marketing are companies that can create desires for consumers.

Several theories about consumer behavior that need to be studied to find out the motivational processes that underlie and direct consumer behavior in making purchases are theories based on economics, psychology, sociology and anthropology views. Consumer behavior describes the way individuals make decisions to use their available resources (time, money, effort) to buy consumption-related goods. (Schiffman and Kanuk, 2007:6).

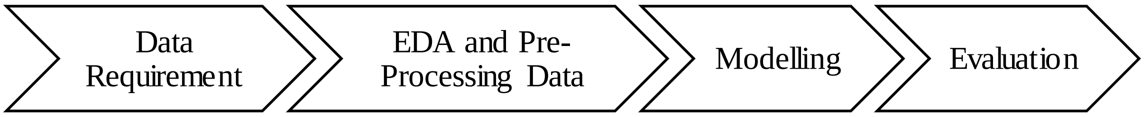
Rahma (2010) conducted research on the analysis of the factors that influence the demand for cluster-type housing (case study of Taman Sari Housing) in Semarang City. Based on the results of the study, it was stated that the influence of price perception variables, facilities, location, environment, income, and substitution prices proved significant to housing demand, this is indicated by the high adjusted coefficient of determination, namely 0.686 or 68.6%. The purchase of consumers who buy and live in Tamansari Semarang Housing is explained by the level of change in price perceptions, facilities, location, environment, income and substitution prices and only 31.4% of the remaining is influenced by other factors not included in the model. Primananda (2010) researched the factors that influence consumers in buying a house with research variables of price, location, building, and environment. The results of the study indicate that the variable that has the largest positive and significant effect on purchasing decisions is the location variable with a coefficient of 0.405. Price variable with a coefficient of 0.276. Environmental variables with a coefficient of 0.228. The building variable with a coefficient of 0.108. Mahardini (2012) analyzed the effect of price, income, location and facilities on the demand for simple houses. The results showed that the effect of price, income, location and facilities had a significant effect on the demand for simple houses.

The variables used to cluster buyer behavior are buyer_city, destination_type, and regency_market.

Buyer_city is The ID of the city the customer is located. This variable shows the specific location of the buyer's residence from continent to city. Then, destination_type is the type of destination the buyer wants, including ; adults, children, room, package, and their environment. And regency_market is regency market of the destination cluster.

Chapter 2

The methodology used to analyze the data in this study will incorporate the EDA stage in the initial data exploration, which will then be modeled based on the data mining classification method. This research went through several stages which can be seen in the following figure. All stages will be implemented using Python language.



Picture 1**Error! No text of specified style in document.**1 Methodology

2.1 Data Requirement

The dataset used in this study is the "Customer's Search History Data" dataset obtained from the Indonesian Real Estate Brokers Association (AREBI). The dataset is the search data recorded by the data storage of application. This dataset has the extension csv, with the following columns:

Table 2.1 Dataset Table

Name of Variable	Description
time_date	Timestamp
site	ID of the site
continent_ID	ID of continent
buyer_country	The ID of the country the customer is located
buyer_region	The ID of the region the customer is located
buyer_city	The ID of the city the customer is located
distance	Physical distance between a regency and a customer at the time of search. A null means the distance could not be calculated
buyer_ID	ID of user
mobile	1 when a user connected from a mobile device, 0 otherwise
package	1 if the click/buying was generated as a part of a package (i.e. combined with a furniture), 0 otherwise
channel_ID	ID of a marketing channel
buying_date	Buying date
dealing_date	Dealing with seller date
adults	The number of adults specified in the room
children	The number of (extra occupancy) children specified in the room
room	The number of rooms specified in the search
destination_ID	ID of the destination where the regency search was performed

destination_type	Type of destination
regency_continent	Regency continent
regency_country	Regency country
regency_market	Regency market
dealing	1 if dealing, 0 if a click
cnt	Number of similar events in the context of the same user session
regency_cluster	ID of regency cluster

2.2 EDA and Pre-Processing Data

At this stage, data exploration is carried out using statistical, mathematical functions, and visualized in the form of graphs. This will make it easier to understand the data and basic data patterns. In addition, this approach is also used to see and find outliers in the data. At this stage, pre-processing of data is also carried out, such as cleaning outlier data and handling empty data, so that it is ready to be processed/analyzed at the stage of making the classification model.

2.3 Modelling Phase

Problems of prediction, forecasting, clustering, anomaly detection can use existing data mining techniques. Each data mining technique is derived from several supervised and unsupervised models. (Prasetyo, 2014) Every problem that exists will be solved by using a model that has been made using the right technique to gain knowledge. However, it is necessary to measure the accuracy of each learning model used, so that it can be concluded the percentage of accuracy in each problem with the techniques used in each model.

Data Clustering is a data mining method that is unsupervised. There are two types of data clustering that are often used in the data grouping process, namely hierarchical data clustering (hierarchical) and non-hierarchical data clustering (non-hierarchical).

Agresti (1996) suggests that categorical variables are one of the measurement scales consisting of a number of categories. Based on the measurement scale, categorical variables are divided into:

1. Nominal scale, namely categorical variables that do not have a sequence of values. For example, the preferred type of music (classical, country, folk, jazz, rock), gender variables, and so on.
2. Ordinal scale, ie categorical variables that have a sequence of values. For example, response to medical care (very good, good, adequate, poor) (Agresti, 1996)

K-modes clustering was first introduced by Huang (1998) as a clustering method which was developed from the k-means method. Therefore, k-modes are efficient as k-means but are used on categorical data. The modifications made to the k-means method are:

1. The distance between two data points X and Y is the number of features in X and Y whose values are different (simple dissimilarity measure), formally formulated as follows:

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \quad (2\text{-}1)$$

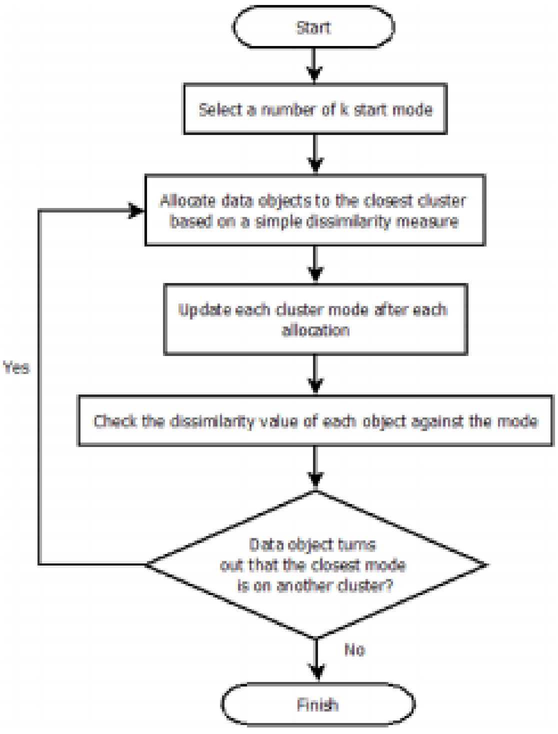
where:

$$\delta(x_j, y_j) = \begin{cases} 0 \, (x_j = y_j) \\ 1 \, (x_j \neq y_j) \end{cases} \quad (2\text{-}2)$$

with:

x_j dan y_j adalah are the j-th feature values of the X dan Y, and m is the number of features.

2. Change the means to modes
3. Using frequency to find the mode (the data value that appears the most). The formation of the centroid is to find the mode of each feature.


Start
Select a number of k start mode
Allocate data objects to the closest cluster based on a simple dissimilarity measure
Update each cluster mode after each allocation
Check the dissimilarity value of each object against the mode
Data object turns out that the closest mode is on another cluster?
Yes
No
Finish

Picture 2.1 K-Modes Flowchart

The following are the steps for k-modes clustering based on (Huang, 2008):

1. Select the starting mode a number of k
2. Allocate data objects to the nearest cluster based on a simple dissimilarity measure. Update each cluster mode after each allocation.
3. After all data objects have been allocated to a cluster, check the dissimilarity value of each object to the mode. If a data object turns out to be the closest mode to be in another cluster, move the object to the appropriate cluster and update the mode of both clusters.
4. Repeat step 3 until no data objects change clusters.

2.4 Evaluation

The evaluation of the classification model made in the previous stage is tested at this stage. The testing phase is carried out using test data. The evaluation of the classification model was analyzed by comparing the overall accuracy value and the time required to conduct model training. .

To determine the optimal number of clusters, the Elbow method Goutte et al (1999) was used but modified using the within cluster difference. From the results of plotting within cluster differences on various values, the principle of the Elbow method takes the value of k at the point when the value does not decrease significantly with the increase in the value of k.

$$y = \sum_{i=1}^{k} \sum_{j=1}^{m} d_1 \left(x_j , x_c\right) \quad (2\text{-}3)$$

with:

y = number within cluster difference
k = number of clusters
m = number of members in each cluster
c = centroid of cluster
d = simple dissimilarity measure according to formula (2-1)
x = data point

Furthermore, cluster evaluation is carried out externally using the purity metric (Tan et al., 2005). Purity of cluster i is:

$$p_i = \max p_{ij} \quad (2\text{-}4)$$

While the overall purity is :

$$Purity = \sum_{j=1}^{m} \frac{m_i}{m} pi \quad (2\text{-}5)$$

2.5 Result and Discussion

This phase is carried out for knowledge discovery (Identification of unexpected and useful relationships) to then be applied to business operations in various purposes, including clustering.

Chapter 3

3.1 Data Requirements

At this stage, data collection was carried out as much as 364601 collected from data from the Indonesian Real Estate Brokers Association (AREBI), then the variables will be extracted and cleaned.

3.2 EDA and Pre-Processing

At this stage, data exploration will be carried out with the aim of finding insights from the dataset that has been given. The following describes EDA and pre-processing of the dataset. The classification of each variable is as follows:

1. Numerical Variable
   a. Continous

      There is only one continuous variable, namely distance. This distance column has a percentage of 57% null values. Due to the very high percentage of null values, we decided not to use this variable for further exploration.

   b. Discrete

      There is room, adults, and children.

2. Categorical Variable
   a. Identifier Variable
      - Cnt

        We decided not to use this variable for further exploration due to inconsistencies between the train dataset and the test dataset.
      - site, continent_id, buyer_id, buyer_country, buyer_region, buyer_city, channel_id, destination_id, destination_type, regency_continent, regency_country, regency_market
      - regency_cluster

        The purpose of this article is to do clustering based on the 3 main variables we chose, namely buyer_city, regency_market, and destination_type. Therefore, we did not use the regency_cluster variable data in this dataset and decided to specify the cluster based on the three selected variables above (based on several clusters that you specify according to your creativity).

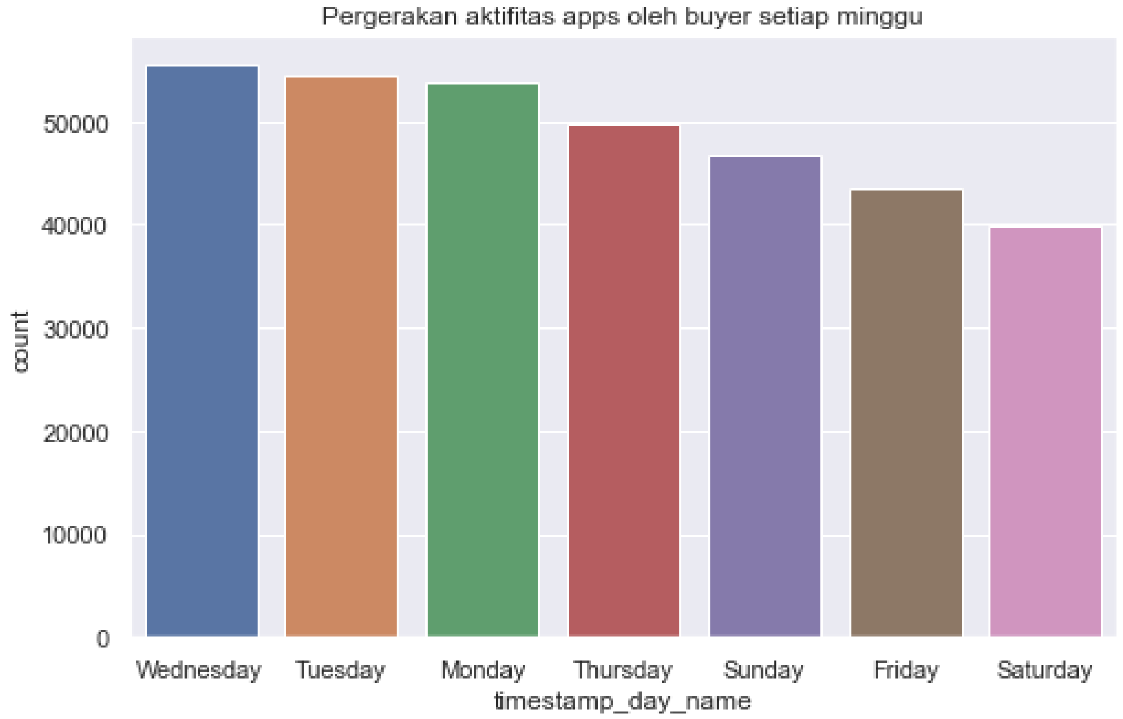
   b. Boolean Variable
      - Dealing

        We decided not to use this variable for further exploration due to inconsistencies between the train dataset and the test dataset.
      - mobile
      - package

3. Time Series Variable
   - time_date
   - buying_date
   - dealing_date

Based on the exploration that has been done using train.csv, the following insights were obtained:

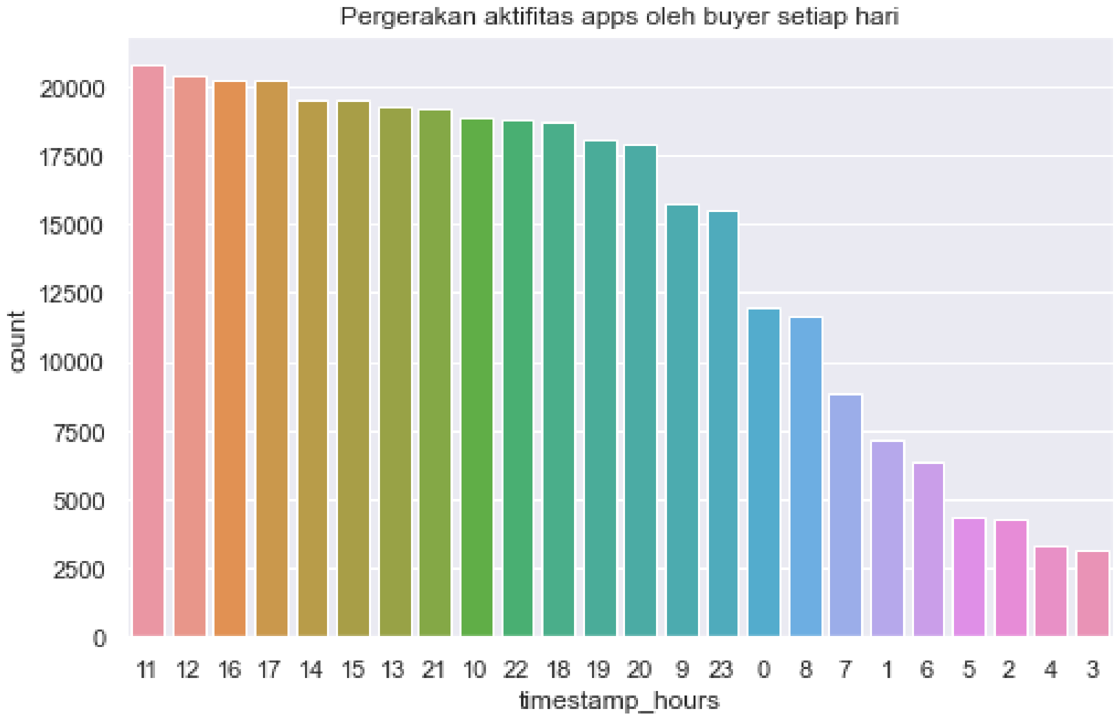
1. Distribution of movement’s buyer activity in apps every week. (extracted from time_date varable)



Picture 3.1 Timestamp Day Barplot

From this it can be concluded that the movement of application activity by buyers every week mostly occurs on Wednesdays.

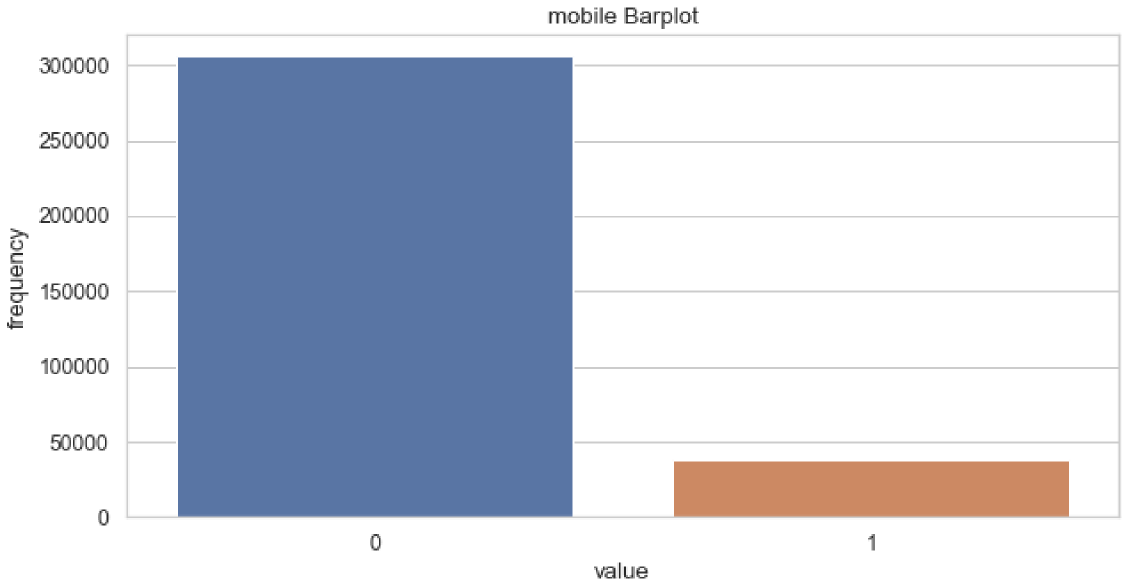
2. Time interval of buyer activity in the application (extracted from the time_date varable)



Picture 3.2 Timestamps Barplot

From this it can be concluded that the time range for the movement of application activity by buyers tends to be intense starting from 10 am to 8 pm.

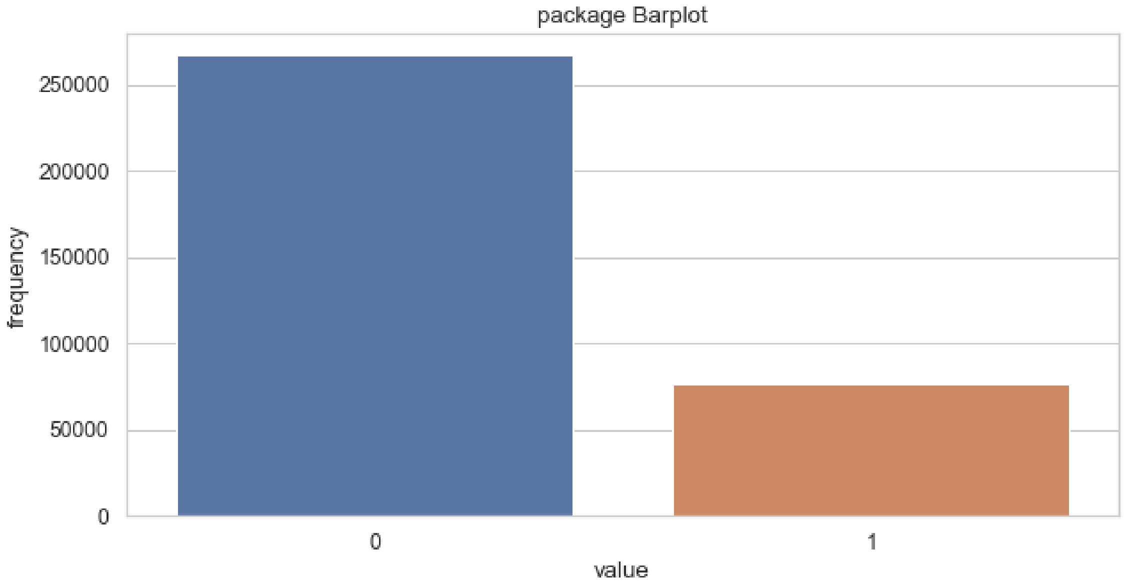
3. The media used by the buyer



Picture 3.3 Mobile Barplot

From this it can be concluded that the majority of buyers use mobile as a device to access applications.

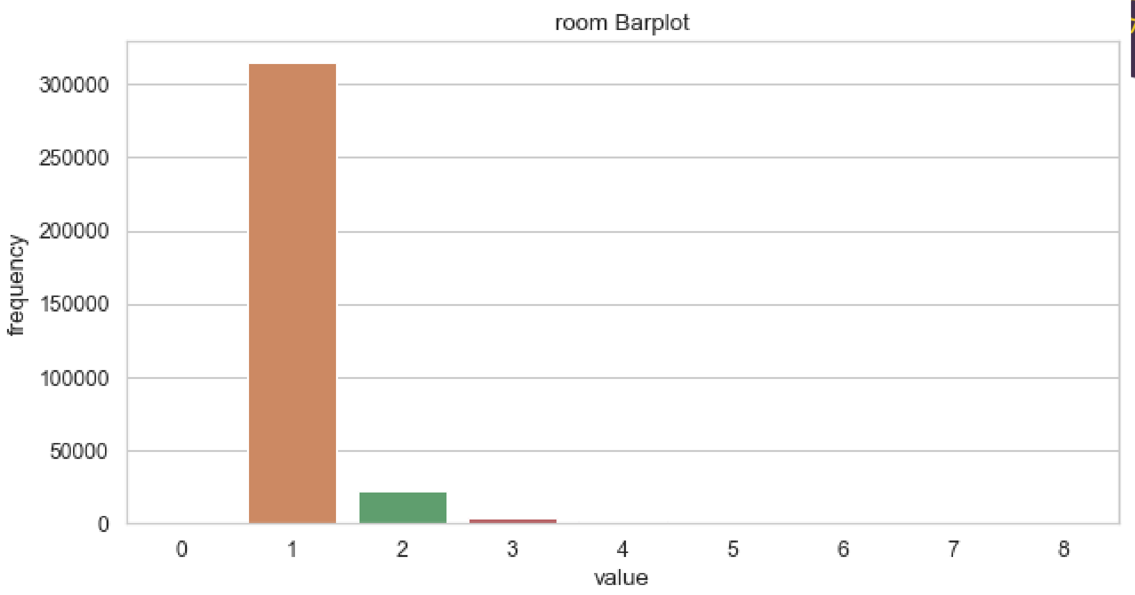
4. Request for additional packages by the buyer



Picture 3.4 Package Barplot

From this it can be concluded that the majority of buyers do not need additional packages (furniture, etc.)

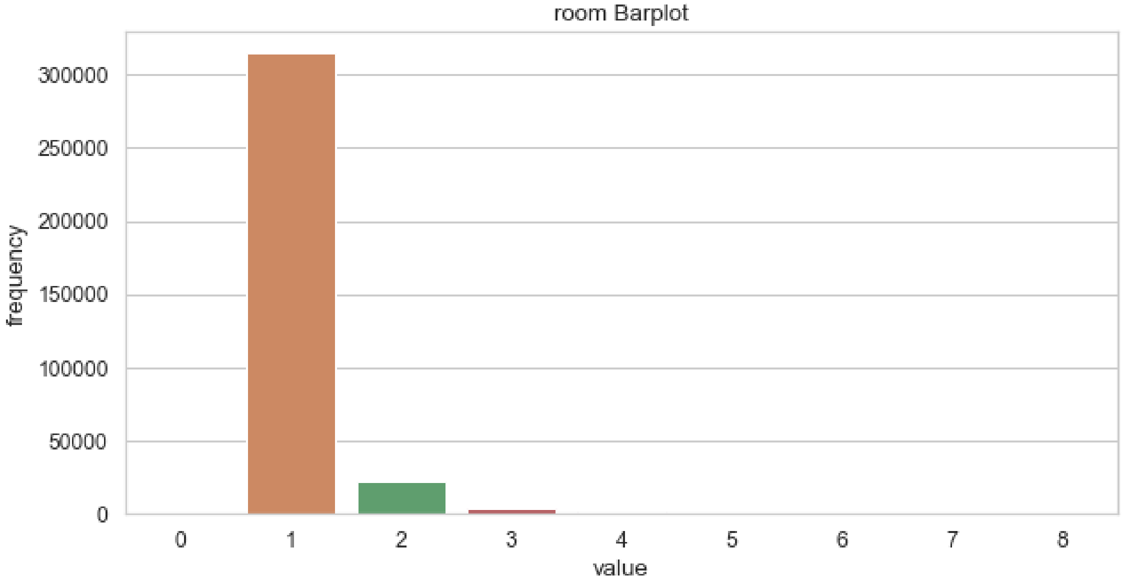
5. The number of rooms required by the buyer



Picture 3.5 Room Barplot

From this it can be concluded that the majority of buyers choose one room.

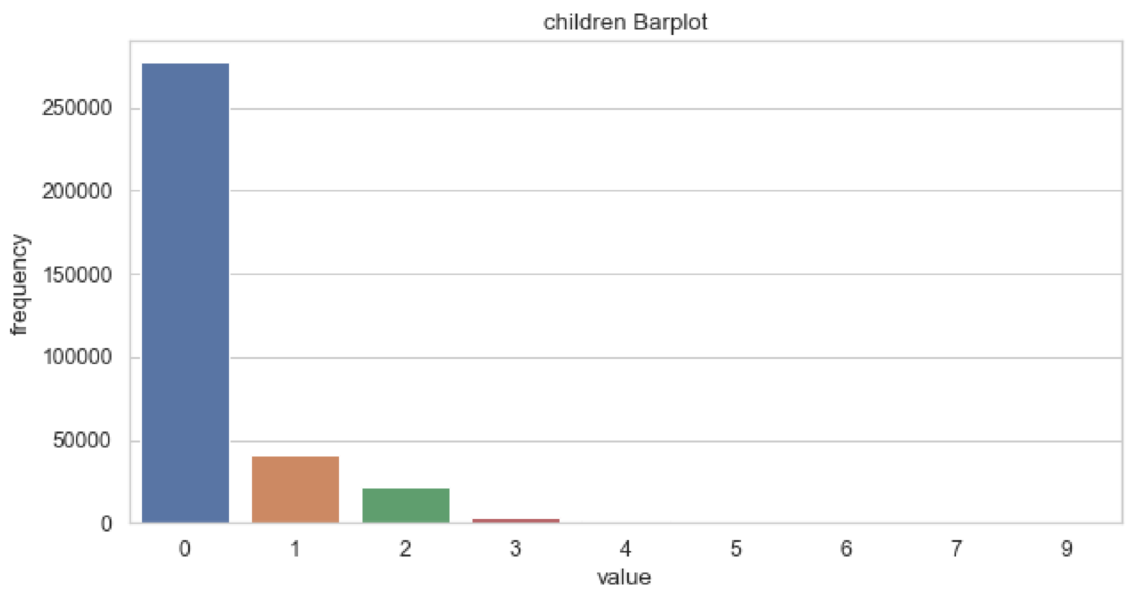
6. Number of adults per order.



Picture 3.6 Room Barplot

From this it can be concluded that the majority of buyers booked a room with two adults.

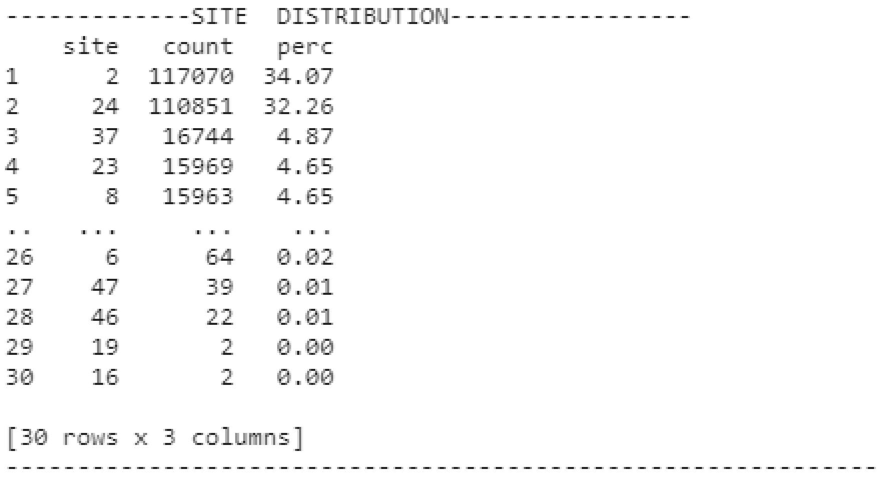
7. Buyer’s Children



Picture 3.7 Children Barplot

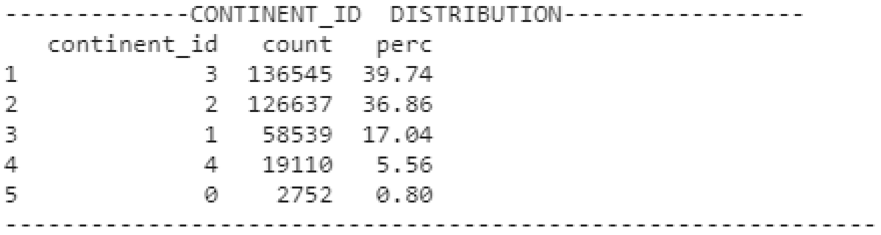
From this it can be concluded that most buyer activities do not require a children's slot in the room they ordered.

8. How is the uniqueness value of each categorical variable (identifier), which ID is the most for each variable?



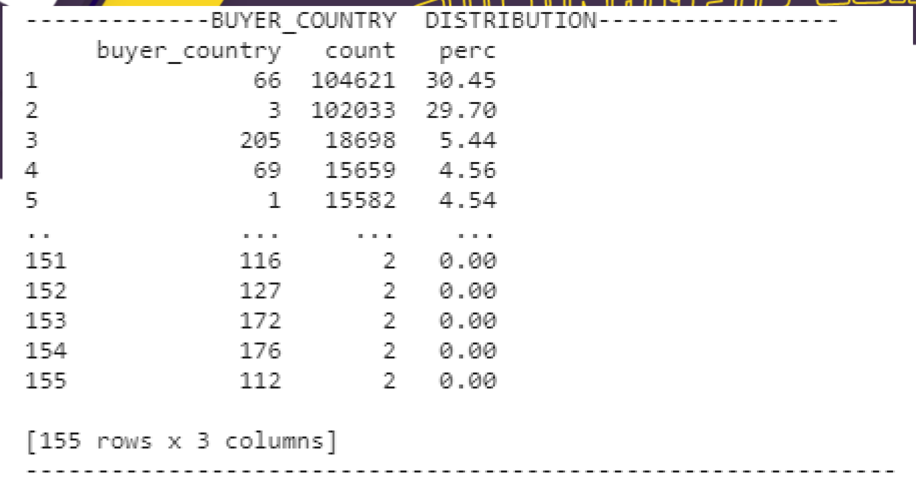
```
-------------SITE  DISTRIBUTION-----------------
    site   count   perc
1      2  117070  34.07
2     24  110851  32.26
3     37   16744   4.87
4     23   15969   4.65
5      8   15963   4.65
..   ...     ...    ...
26     6      64   0.02
27    47      39   0.01
28    46      22   0.01
29    19       2   0.00
30    16       2   0.00

[30 rows x 3 columns]
-------------------------------------------------------------
```

From this it can be concluded that the majority of sites identified in the dataset are sites with ID 2 and 24.



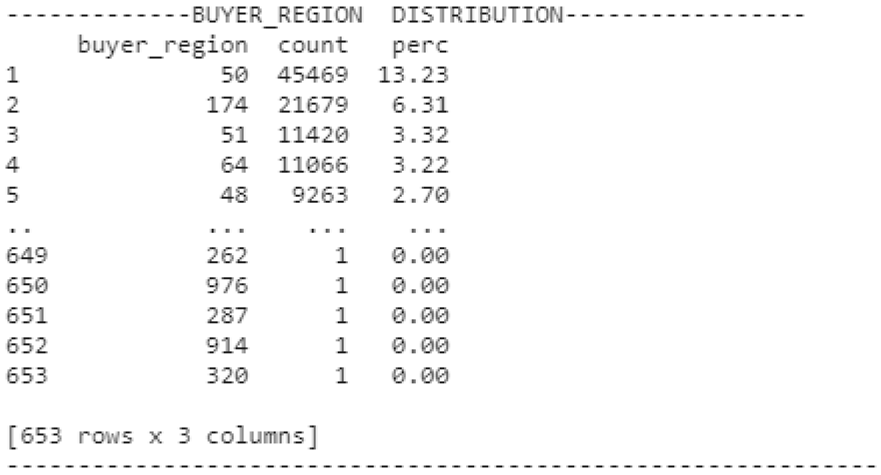
```
-------------CONTINENT_ID  DISTRIBUTION-----------------
   continent_id   count   perc
1             3  136545  39.74
2             2  126637  36.86
3             1   58539  17.04
4             4   19110   5.56
5             0    2752   0.80
-------------------------------------------------------------
```

From this it can be concluded that the majority of continent_id identified in the dataset are continents with id 3 and 2.



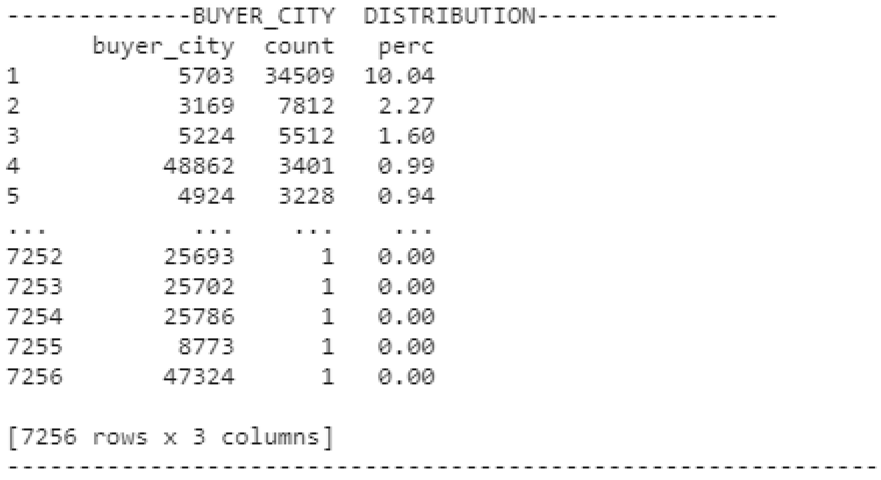
```
-------------BUYER_COUNTRY  DISTRIBUTION----------------
     buyer_country   count   perc
1               66  104621  30.45
2                3  102033  29.70
3              205   18698   5.44
4               69   15659   4.56
5                1   15582   4.54
..             ...     ...    ...
151            116       2   0.00
152            127       2   0.00
153            172       2   0.00
154            176       2   0.00
155            112       2   0.00

[155 rows x 3 columns]
-------------------------------------------------------------
```

From this, it can be concluded that the majority of buyer_country Identified in the dataset are buyer_country with ID 66 and 3.



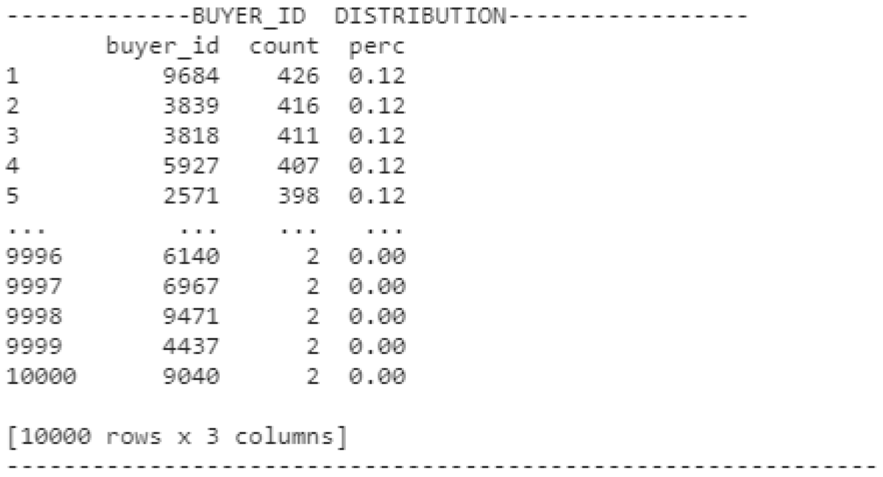
```
-------------BUYER_REGION  DISTRIBUTION----------------
     buyer_region  count   perc
1              50  45469  13.23
2             174  21679   6.31
3              51  11420   3.32
4              64  11066   3.22
5              48   9263   2.70
..            ...    ...    ...
649           262      1   0.00
650           976      1   0.00
651           287      1   0.00
652           914      1   0.00
653           320      1   0.00

[653 rows x 3 columns]
-------------------------------------------------------------
```

From this, it can be concluded that the majority of buyer_regions Identified in the dataset are buyer_regions with ID 50. The number of unique ID for this variable is 653.



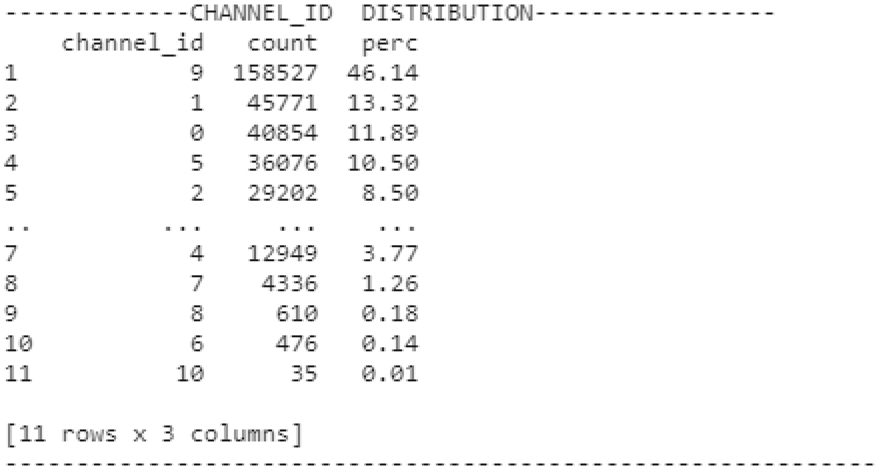
```
-------------BUYER_CITY  DISTRIBUTION----------------
      buyer_city  count   perc
1           5703  34509  10.04
2           3169   7812   2.27
3           5224   5512   1.60
4          48862   3401   0.99
5           4924   3228   0.94
...          ...    ...    ...
7252       25693      1   0.00
7253       25702      1   0.00
7254       25786      1   0.00
7255        8773      1   0.00
7256       47324      1   0.00

[7256 rows x 3 columns]
-------------------------------------------------------------
```

From this it can be concluded that the majority of buyer_city Identified in the dataset is buyer_city with ID 5703. The number of unique ID for this variable is 7256.



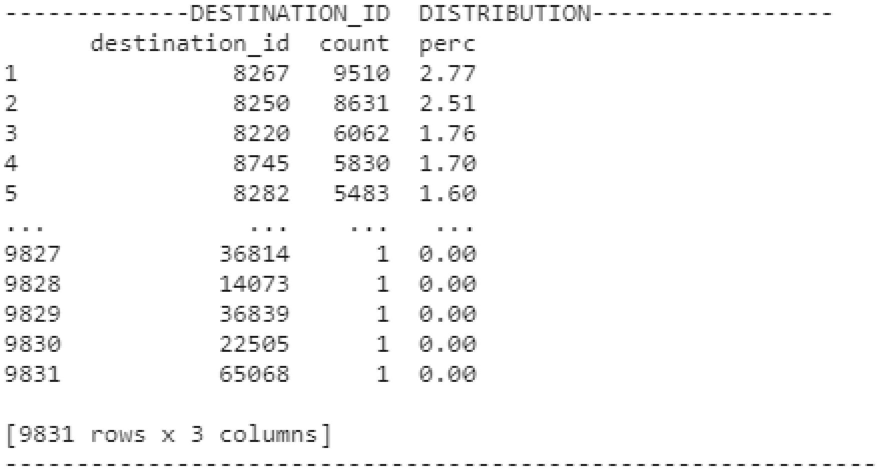
```
-------------BUYER_ID  DISTRIBUTION----------------
       buyer_id  count  perc
1          9684    426  0.12
2          3839    416  0.12
3          3818    411  0.12
4          5927    407  0.12
5          2571    398  0.12
...         ...    ...   ...
9996       6140      2  0.00
9997       6967      2  0.00
9998       9471      2  0.00
9999       4437      2  0.00
10000      9040      2  0.00

[10000 rows x 3 columns]
-------------------------------------------------------------
```

It can be seen that there are 10000 unique ID for each buyer. The buyers who access the application the most are buyers with buyer_ID 9684



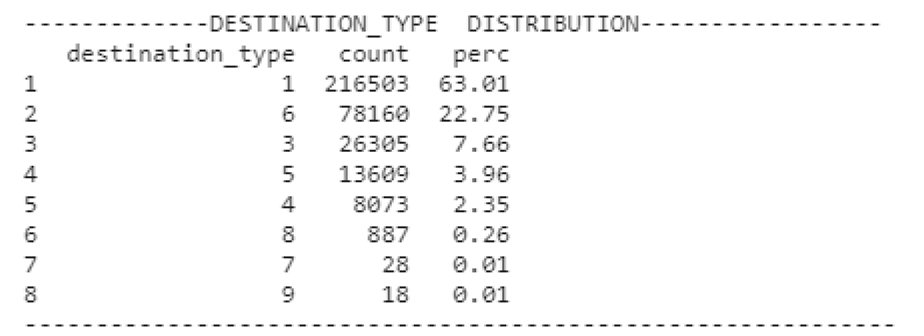
```
-------------CHANNEL_ID  DISTRIBUTION----------------
    channel_id   count   perc
1            9  158527  46.14
2            1   45771  13.32
3            0   40854  11.89
4            5   36076  10.50
5            2   29202   8.50
..         ...     ...    ...
7            4   12949   3.77
8            7    4336   1.26
9            8     610   0.18
10           6     476   0.14
11          10      35   0.01

[11 rows x 3 columns]
-------------------------------------------------------------
```

From this it can be concluded that the majority of channel_ID Identified in the dataset are channel_ID with ID 9. The number of unique ID for this variable is 11.



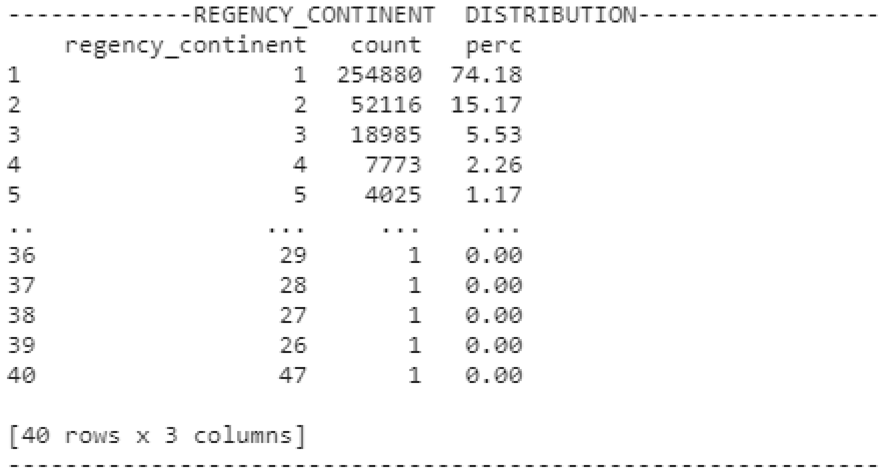
```
-------------DESTINATION_ID  DISTRIBUTION----------------
      destination_id  count  perc
1               8267   9510  2.77
2               8250   8631  2.51
3               8220   6062  1.76
4               8745   5830  1.70
5               8282   5483  1.60
...              ...    ...   ...
9827           36814      1  0.00
9828           14073      1  0.00
9829           36839      1  0.00
9830           22505      1  0.00
9831           65068      1  0.00

[9831 rows x 3 columns]
-------------------------------------------------------------
```

From this, it can be concluded that the most frequently selected destination_ID in the application is destination_ID 8267. The number of unique ID for this variable is 9831.



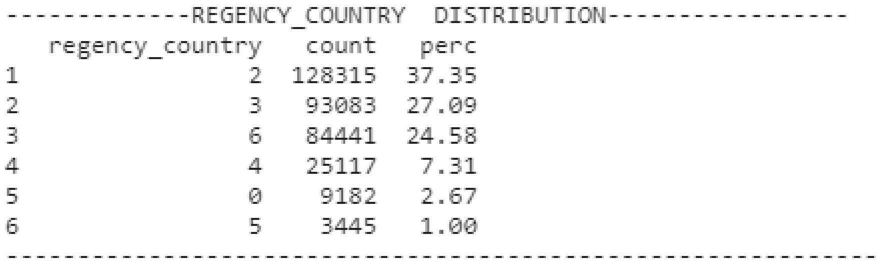
```
-------------DESTINATION_TYPE  DISTRIBUTION----------------
   destination_type   count   perc
1                 1  216503  63.01
2                 6   78160  22.75
3                 3   26305   7.66
4                 5   13609   3.96
5                 4    8073   2.35
6                 8     887   0.26
7                 7      28   0.01
8                 9      18   0.01
-------------------------------------------------------------
```

From this it can be concluded that the majority of destination_type selected in the application are destination types with ID 1. For the number of unique ID this variable is 8.



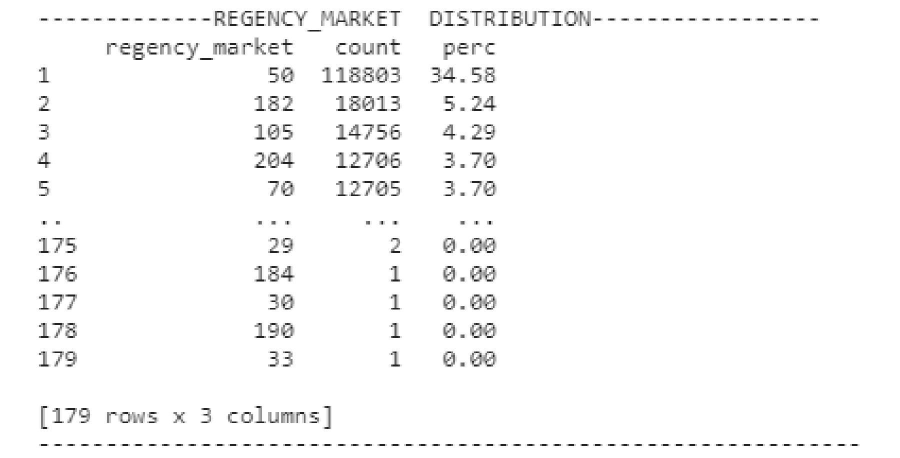
```
-------------REGENCY_CONTINENT  DISTRIBUTION----------------
    regency_continent    count   perc
1                   1   254880  74.18
2                   2    52116  15.17
3                   3    18985   5.53
4                   4     7773   2.26
5                   5     4025   1.17
..                ...      ...    ...
36                 29        1   0.00
37                 28        1   0.00
38                 27        1   0.00
39                 26        1   0.00
40                 47        1   0.00

[40 rows x 3 columns]
-------------------------------------------------------------
```

From this, it can be concluded that the majority of regency_continet Identified in the dataset is regency_continent with ID 1. The number of unique ID for this variable is 40



```
-------------REGENCY_COUNTRY  DISTRIBUTION----------------
   regency_country   count   perc
1                2  128315  37.35
2                3   93083  27.09
3                6   84441  24.58
4                4   25117   7.31
5                0    9182   2.67
6                5    3445   1.00
-------------------------------------------------------------
```

From this, it can be concluded that the majority of regency_continet Identified in the dataset is regency_country with ID 2. The number of unique ID for this variable is 6



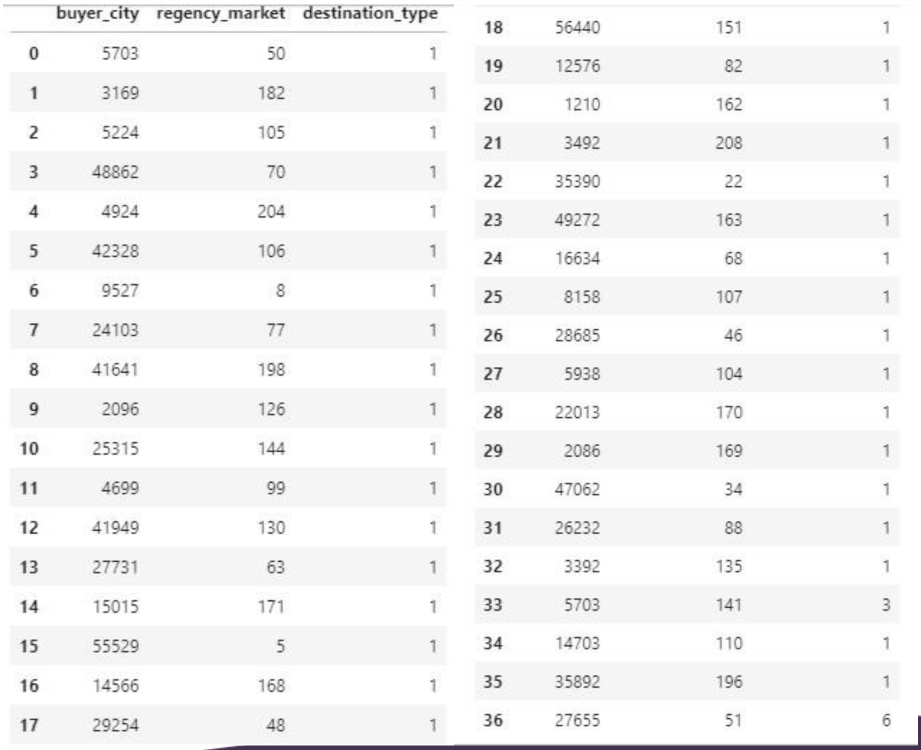
```
------------REGENCY_MARKET  DISTRIBUTION----------------
     regency_market   count   perc
1                50  118803  34.58
2               182   18013   5.24
3               105   14756   4.29
4               204   12706   3.70
5                70   12705   3.70
..              ...     ...    ...
175              29       2   0.00
176             184       1   0.00
177              30       1   0.00
178             190       1   0.00
179              33       1   0.00

[179 rows x 3 columns]
-------------------------------------------------------------
```

From this, it can be concluded that the majority of regency_markets Identified in the dataset are regency_markets with ID 50. The number of unique ID for this variable is 179.

3.3 Modeling Phase and Evaluation

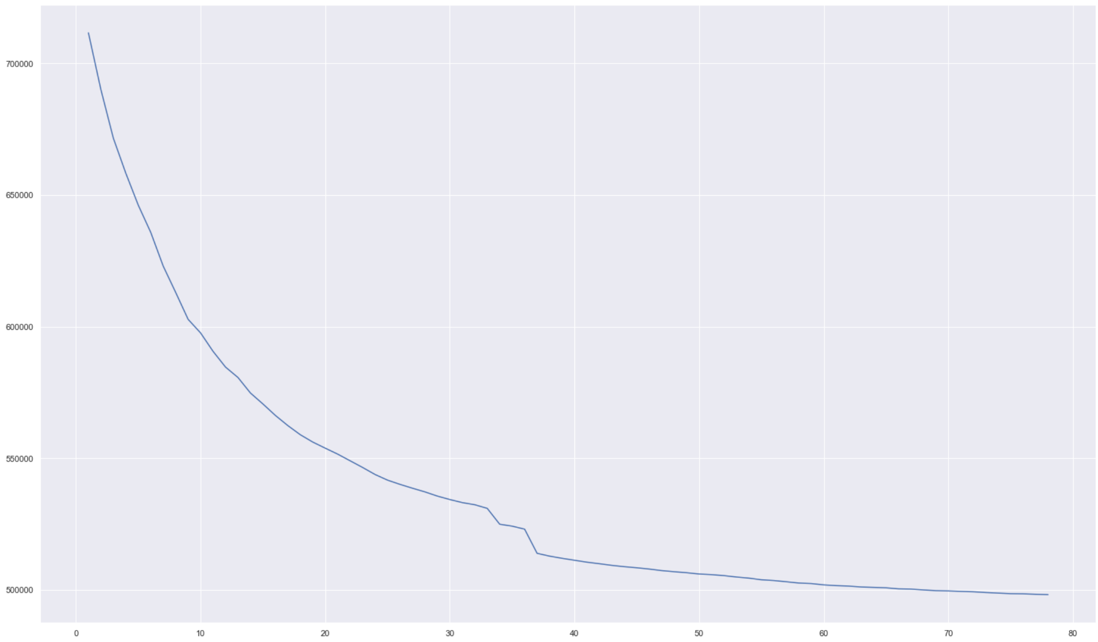
In the modeling stage, the variables used are buyer_city, regency_market, and destination_type. The model used in analyzing this data is the K-Modes Clustering algorithm, because it will be used on categorical data (Huang, 1998). To determine the optimal number of clusters, the Elbow Method Goutte, et al (1999) was used but modified using the within cluster difference. From the results of plotting within cluster differences on various values, the principle of the Elbow method takes the value of k at the point when the value does not decrease significantly with the increase in the value of k.

Clustering is performed on various values. The smallest within cluster difference value for each k value is shown in the following figure. Based on the elbow method, the best alternative number of k is 37, because there is a sloping shape like an elbow.



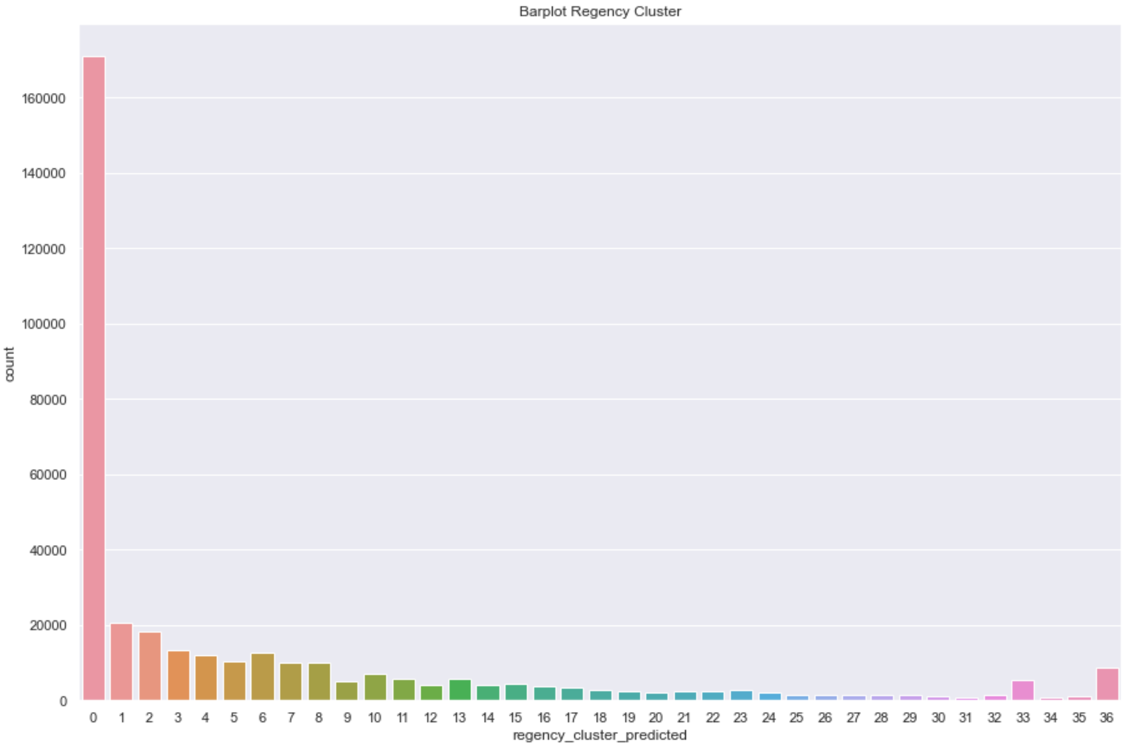
	buyer_city	regency_market	destination_type
0	5703	50	1
1	3169	182	1
2	5224	105	1
3	48862	70	1
4	4924	204	1
5	42328	106	1
6	9527	8	1
7	24103	77	1
8	41641	198	1
9	2096	126	1
10	25315	144	1
11	4699	99	1
12	41949	130	1
13	27731	63	1
14	15015	171	1
15	55529	5	1
16	14566	168	1
17	29254	48	1
18	56440	151	1
19	12576	82	1
20	1210	162	1
21	3492	208	1
22	35390	22	1
23	49272	163	1
24	16634	68	1
25	8158	107	1
26	28685	46	1
27	5938	104	1
28	22013	170	1
29	2086	169	1
30	47062	34	1
31	26232	88	1
32	3392	135	1
33	5703	141	3
34	14703	110	1
35	35892	196	1
36	27655	51	6

Picture 3.8 Clustering’s centroids with k=37



Picture 3.9 Elbow's Method Graph

The results of the cluster with the number of k = 37 produce a group of buyer characteristics when buying property. It can be seen that the center of the majority of buyers comes from cities with ID-5073 by visiting the regency market that is often visited, namely ID-50, and the type of destination that is also the interest of the majority of buyers with ID-1. Thus, this makes the application developer's strategy to increase buyer interest in the type of cluster that is less desirable by analyzing the majority cluster.



Picture 3.10 Regency Cluster Barplot

Chapter 4

4.1 Conclusion

This study uses clustering experiments with the K-Modes method on the distribution of variables in determining the characteristics of buyers in choosing the regency cluster that they want. Data which amounted to 364601 were collected from data from the Indonesian Real Estate Brokers Association (AREBI), which will then be extracted and cleaned. And modeled to get the ideal number of clusters, namely k = 37 to produce the majority group. It can be seen that the center of the majority of buyers comes from cities with ID-5073 by visiting the regency market that is often visited, namely ID-50, and the type of destination that is also the interest of the majority of buyers with ID-1. Thus, this makes the application developer's strategy to increase buyer interest in the type of cluster that is less desirable by analyzing the majority cluster.

4.2 Suggestions

1. The application server can be strengthened starting at 10.00 every day because the movement of application bandwidth begins to be intense at that hour.
2. The UX and UI designs of mobile applications (smartphones) are further improved in quality and aesthetics because the majority of buyers use mobile as a device to access applications.
3. The application provides advertisements with the aim of attracting buyers to add packages (such as furniture, etc.) because the majority of buyers do not choose additional packages.
4. The application developer pays attention to clusters that are still lacking in interest by suggesting to improve their marketing strategy according to the clusters that have been analyzed in this article.

Bibliography

Airea, 2001. The Apprisal of Real Estate 12th edition, Chicago USA.

Ahn, Y., Ahnert, S. E., Bagrow, J. P., & Barabási, A. 2011. Flavor Network And The Principles Of Food Pairing. Sci. Rep. 1. Doi:10.1038/Srep00196.

Appraisal Institute, 1993. The Dictionary of Real Estate Appraisal. Illinois: Appraisal Institute.

Awang Firdaos. 1997. “ Permintaan dan Penawaran Perumahan” Value estate, Vol. 007, Jakarta. Chiara, Joseph De dan Lee E. Koppelman. Site Planning Standards. New York: McGraw-Hill, 1978.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. A., & Hansen, L. K. 1999. On Clustering Fmri Time Series. Neuroimage. 9 (3): 298–310. Doi:10.1006/Nimg.1998.0391.

Huang, Z. 1998. Extensions To The K-Means Algorithm For Clustering Large Data Sets With Categorical Values. Data Mining And Knowledge Discovery, 2(3), 283–304.

Christianto Steven. 2013. Factor Analysis Of Real Estate Investment Trust (REIT) Affecting Use For Property Financing In Surabaya. Institut Teknologi Sepuluh Nopember

Kusrini dan E.T. Luthfi., 2009, Algoritma Data Mining, Andi, Yogyakarta.8

K. Rajalakshmi, S. S. Dhenakaran, and N. Roobini, “Comparative Analysis of K-Means Algorithm in Disease Prediction,” Int. J. Sci. Eng. Technol. Res., vol. 4, no. 7, pp. 2697– 269

Marzuki Muryati. 2002. Restrukturisasi Kredit Sektor Properti dan Real Estate. Jurnal Hukum. No. 19 Vol 9. Februari2002: 64 – 80

Nengsih Warnia. 2017. Analisa Akurasi Permodelan Supervised Dan Unsupervised Learning Menggunakan Data Mining. Politeknik Caltex Riau

Prasetyo, Eko, "Data Mining Mengolah Data menjadi Informasi dengan Matlab," Andi Yogyakarta, 2014.

Witten, et al., 2012, Data Mining Practical Machine Learning Tools and Technique, 2nd Edition, Morgan Kaufmann, San Faransisco.