K-means cluster analysis of tourist destination in special region of Yogyakarta using spatial approach and social network analysis (a case study: post of @explorejogja instagram account in 2016)

N Iswandhani¹, M Muhajir²

¹,²Department of Statistics Islamic University of Indonesia, Sleman, Yogyakarta, Indonesia, 55584.
E-mail: ¹13611066@students.ui.ac.id , ²mmuhajir@uii.ac.id

Abstract. This research was conducted in Department of Statistics Islamic University of Indonesia. The data used are primary data obtained by post @explorejogja instagram account from January until December 2016. In the @explorejogja Instagram account found many tourist destinations that can be visited by tourists both in the country and abroad. Therefore it is necessary to form a cluster of existing tourist destinations based on the number of likes from user Instagram assumed as the most popular. The purpose of this research is to know the most popular distribution of tourist spot, the cluster formation of tourist destinations, and central popularity of tourist destinations based on @explorejogja Instagram account in 2016. Statistical analysis used is descriptive statistics, k-means clustering, and social network analysis. The results of this research were obtained the top 10 most popular destinations in Yogyakarta, map of html-based tourist destination distribution consisting of 121 tourist destination points, formed 3 clusters each consisting of cluster 1 with 52 destinations, cluster 2 with 9 destinations and cluster 3 with 60 destinations, and central popularity of tourist destinations in the special region of Yogyakarta by district.

1. Introduction
The development of internet technology continues to progress very rapidly. With the advent of various kinds of facilities offered the internet makes users more access to information. The development of the internet that is rife among the public is social media.

One of the most popular social media is Instagram. Instagram is a social media containing photo content, information media and references. Instagram users are now more than 50 million inhabitants thus making Instagram occupy the 4th position of most users for social media category [1].

Instagram contains about a variety of account content, ranging from online shop accounts that sell goods and services, campus organization accounts and also Instagram accounts that promote tourism in a country or city.

Tourism is a journey from one place to another, temporary, done by individuals or groups, in an attempt to find balance or harmony and happiness with the environment in the social, cultural, natural and scientific dimension [2].

Not a few tourists who seek ideas through Facebook, Twitter, and other social networking. 65% of travelers are looking for travel ideas through social search. 52% Facebook users are heavily influenced by the photos of friends in his Facebook network to determine the sights. 33% of travelers changed their original plans after viewing the photos [3]. @explorejogja Instagram account serves as a tourism promotion tool for Yogyakarta area, with Follower 510k (Five hundred ten thousand) they are able to promote or introduce new and old tourist spot to people outside Yogyakarta and Yogyakarta society itself.
Popularity comes from the popular word, meaning known and liked by the crowd [4]. In the Indonesian General Dictionary of popularity means the fame of a person [4].

Yogyakarta is a province with a variety of tourist destinations. Ranging from mountains, waterfalls, zoos to complete beaches are available in the region. Newer attractions will soon become booming as many social media users share when visiting the place [5]. With the increasing number of new places, there is a need for facilities that can summarize information from those places.

The tourism sector is a potential sector to be developed as one source of local revenue. For that in need of way for tourist destinations in the province of Yogyakarta can be known by the public, one of them with the media publications and information.

One solution to be able to display information in the right format is a dynamic map that developed a web-based application that is able to manage and display tourist destination information. This app is expected to address some of the limitations of the manual map. Maps used with features owned by Google Map. With the development of the dynamic map is then realized a geographic information system with the theme of the spread of tourist destinations in yogyakarta [6].

In the @explorejogja instagram account found many tourist destinations that can be visited by tourists both in the country and abroad, Therefore it is necessary to form a cluster of existing tourist destinations based on the number of likes from user instagram assumed as the most popular.

Cluster analysis is one type of problem in data mining. Cluster analysis in data mining (also known as clustering) is a method used to divide the data set into clusters based on similarities that have been determined [7]. The object of this research is the number of likes of post account of Instagram of Explorejogja relating to tourist destinations in Yogyakarta province in 2016, to categorize the object of the researcher using K-Means method.

Based on post account instagram @explorejogja get the data name of tourist destinations and also the location of tourist destinations (in this case the city), then will be searched the popularity of tourist destinations that most many get like from instagram users who see posts from account instagram @explorejogja based on city location tourist destination. The existence of this popularity center will be determined through graphs on Social Network Analysis (SNA). SNA is one of the analysis in data mining that connects several interrelated objects through graph [8].

2. Basic theory

2.1. Spatial analysis
Spatial analysis is a visual inference to a map that is a combination of spatial data and attribute data. Spatial data refers to a location or position on the surface of the earth, which is the coordinates, raster or administrative boundaries of the region.

![Figure 1. Differences of spatial data and attribute data](image)

2.2. Data mining
Data mining is the mining or discovery of new information by searching for a particular pattern or rule from a large amount of data [10]. Data mining is also called a series of processes to explore the added value of knowledge that so far can not be known manually from a data collection [11]. Data mining has stages like in figure 2.
2.3 K-means

Clustering is used to create a group (cluster) of the data so that it can easily find the necessary data. Clustering is a classification of similar objects into several different groups, it is usually applied in the analysis of statistical data which can be utilized in various fields, for example, machine learning, data mining, pattern recognition, image analysis and bioinformatics [13].

Clustering including supervised learning types. There are four types of clustering algorithms that have been compared based on performance, such as K-Means, hierarchical clustering, self-organization map (SOM) and expectation maximization (EM Clustering). Based on these test results can be concluded that the k-means algorithm performance and EM better than a hierarchical clustering algorithm. In general, partitioning algorithms such as K-Means and EM highly recommended for use in large-size data. This is different from a hierarchical clustering algorithm that has good performance when they are used in small size data [14].

The method of K-means algorithm as follows [15]:

1) Determine the number of clusters k as in shape. To determine the number of clusters K was done with some consideration as theoretical and conceptual considerations that may be proposed to determine how many clusters.

2) Generate K centroid (the center point of the cluster) beginning at random. Determination of initial centroid done at random from objects provided as K cluster, then to calculate the i cluster centroid next, use the following formula:

\[ v = \frac{\sum_{i=1}^{n} x_i}{n} \]  

\[ (1) \]

v: cluster centroid  
n: the number of objects to members of the cluster  
x_i: the object to-i

3) Calculate the distance of each object to each centroid of each cluster. To calculate the distance between the object with the centroid author using Euclidian Distance.

\[ d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]  

\[ (2) \]

x_i: object x to-i  
y_i: object y to-i  
n: the number of object

4) Allocate each object into the nearest centroid. To perform the allocation of objects into each cluster during the iteration can generally be done in two ways, with a hard K-means, where it is explicitly every object is declared as a member of the cluster by measuring the distance of the proximity of nature towards the center point of the cluster, another way to do with fuzzy C-Means.

5) Do iteration, then specify a new centroid position using equation (1).
6) Repeat step 3 if the new centroid position is not the same.

2.4. Social Network Analysis

Social Network Analysis (SNA) is one of the analysis in data mining that connects several interrelated objects through graph. Objects in SNA called actor terms are the main focus in this analysis.

There are two types of relationships that can be explained in SNA, namely [16]:

1) Directional Relations: the type of relationship "self choices" that the relationship that occurs between actors is the choice of each actor and does not apply to each other in opposite, eg friendship relationship between A and B. If A recognizes B as a friend then not B will recognize A as his friend.

X: friendship sosiomatriks A, B and C. If it is known that A is friends with B (A → B), B be friends with C (B → C), C be friends with A (C → A), dan C be friends with B (C → B), If a relationship is dichotomous then the element of the matrix X (X_{ij}), with i=A,B,C and j=A,B,C are:

\[
X = \begin{bmatrix}
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & -1 \\
\end{bmatrix}
\] (3)

2) Nondirectional Relations: the type of relationship between actors symmetrical each other.

Example of this relationship is the neighboring neighborhood. If A-neighbor is next door with B then it is definitely B after the house again with A. This type of relationship will be denoted by a line (without arrows) on the sosiogram. In the form of matrix notation, this relationship can be described as follows:

X: sosiomatriks next door neighbor A, B and C. If it is known that A next door neighbor with B (A-B and BA), adjacent neighbor B with C (B-C and C-B). If a relationship is dichotomous then the element of the X matrix (X_{ij}), with i=A,B,C and j=A,B,C are:

\[
X = \begin{bmatrix}
-1 & 0 & 0 \\
0 & -1 & 0 \\
0 & 0 & -1 \\
\end{bmatrix}
\] (4)

In SNA, the relationships between actors can be valuable dichotomes and have value. The relationship is dichotomous if the relation exists then it is worth 1 and if there is no relationship will be worth 0. The relation can also be valuable so that each relationship between the actor, the intensity of the relationship or the frequency of the relationship.

3. Research Methods

The first step in conducting this research is data collection, based on previous research conducted by researchers to update data so that get 313 pots of tourism destinations from @explorejogja then get 121 points of tourist destination location by collecting the point of coordinat based on latitude and longitude of each tourist destinations, descriptive statistical analysis, create a map of web-based tourism destination distribution, Cluster k-means analysis, and create Social network analysis graphs.

Data analysis method used in this research is descriptive analysis, tourism destination distribution in the form of web based map, non hierarchy cluster analysis that is K-means method and Social network analysis. The tools used in data analysis in this study are Ms. Excel, QGIS 2.10, R, and gephi.
4. Result

4.1. Distribution of tourist destination

Instagram accounts explorejogja have a fairly large existence as an account that promotes tourist destinations contained in the province of Yogyakarta. These accounts also get a fairly positive response from Instagram users, especially followers (followers) of the explorejogja account.

![Figure 3. Chart the number of posts explorejogja](image)

From the picture above in figure 3, it can be seen the top 10 of most popular tourist destination from January until December 2016. Namely is Tugu, Kaliadem, Siung beach, Suroloyo peak, Sewu temple, Prambanan temple, Mojo Hill, Taman Pelangi, keraton of Yogyakarta and Kalibiru.

Based on 121 tourist destinations obtained, then in the form of map web-based tourist destination with 121 point coordinates in figure 4.

![Figure 4. The map of Tourist Destination Destinations in special region of Yogyakarta](image)

As for the explanation of tourist destinations in accordance with the coordinate point as follows:

| No | Destination Names          | Coordinate point          |
|----|----------------------------|---------------------------|
| 1  | Kedung kandang waterfall  | -7.855807 110.535138      |
| 2  | Kedung Pedut waterfall    | -7.766689 110.120270      |
4.2. **K-means clustering**

The K-means method processes all objects simultaneously where k is the number of groups. So in the K-Means method also determined the number of groups formed, in this study in the form of three groups.

| Cluster | Number of tourist destination |
|---------|-------------------------------|
| 1       | 9                             |
| 2       | 60                            |
| 3       | 52                            |

Table 2. A list of tourist destinations along with point coordinates

| Cluster | Average amount like |
|---------|---------------------|
| 1       | 19683.78            |
| 2       | 14213.23            |
| 3       | 10538.98            |

In table 2 above can be seen that the grouping by using K-means method of popularity of tourist destinations based on the number of like from account Instagram Explorejogja consists of 121 tourist destinations that have been in execution as much as 50 times by using statistical software. There are 3 clusters each consisting of cluster 1 with 9 tourist destinations, cluster 2 consists with 60 tourist destinations and cluster 3 with 52 tourist destinations. Of the three clusters are then performed profiling cluster that obtained the following results.

4.3. **Social network analysis**

Social Network Analysis (SNA) is formed between tourist destinations based on tourist sites (sub-district) seen there are some locations that become a big role as the location that has the most tourist destinations.
In the SNA Graph above in figure 5, it can be concluded that the sections 8.7 and 6 are expressed as the center of popularity based on the location (sub-district) tourist destination. So to be more clearly done the analysis by looking at the percentage of the main popularity by using the gephi program as in the following figure.

As in the picture above in figure 6, there are 16 red dots, each of which is 8 tourist destinations located in Dlingo district and 8 other tourist destinations located in Girimulyo district with each percentage of 6.61%. Then at the blue point there are 28 points, each of which is 7 tourist destinations located in Pakem sub-district, Tepus, Patuk and Tanjungsari with each percentage percentage of 5.79%. And lastly on the yellow point are 6 tourist destinations located in the district of Prambanan.

5. Conclusion

The Distribution of tourist destinations in Yogyakarta province is displayed in the form of a web-based dynamic map in which there are 121 points coordinates that symbolize each tourist destination. Of the 121 destinations are then obtained top 10 most popular destinations based on the number of like. Namely, Tugu yogyakarta, Kaliadem, Siung beach, Suroloyo peak, Sewu temple, Prambanan temple, Mojo hill, Taman Pelangi, and Kalibiru.

Grouping by means of K-means method to classify the popularity of tourist destination based on the number of ig account from explorejogja consists of 121 tourist destinations that have been in execution 50 times using statistical software. There are 3 groups, each of which consists of cluster 1 of 9 tourist destinations, cluster 2 consists of 60 tourist destinations and for cluster 3 as many as 52 tourist destinations.

In the SNA graphic it is found that among the 8 sections present, sections 8, 7 and 6 are expressed as the popularity centers based on the location (sub-district) of the tourist destination. central
popularity of tourist destinations in the province of Yogyakarta based on the sub-district is located in 7
districts namely Dlingo, Girimulyo, Pakem, Tepus, Patuk, Tanjungsari and Prambanan.

Based on the conclusions obtained by the researcher wants the results of this study can be used by
the Yogyakarta government, especially in the field of tourism in order to be used as a reference to
improve quality and publications related tourism that has been considered popular. This is certainly in
order to bring in more tourists-both from within the country and from abroad.

References
[1] Huraerah A 2016 Communication behavior of explorejogja instagram account in Yogyakarta
tourism journey (Yogyakarta: Respositary UMY)
[2] Kodyat H 1998 history of tourism and its development of Indonesia (Jakarta: Grasindo)
[3] Nevita S 2016 optimize social media for tourism promotion
http://m.radarbangka.co.id/rubrik/detail/perspektif/13837/mengoptimalkan-medsos-untuk-
promosi-pariwisata.html
[4] Poerwadarminta 2006 Indonesia dictionary Third edition (Jakarta: Balai pustaka)
[5] Nita V 2015 #explorejogja As Reference Places in Yogyakarta
https://citraceritajogja.wordpress.com/2015/01/12/explorejogja-sebagai-referensi-tempat-
wisata-di-yogyakarta/
[6] Dewanto 2012 Design of Geographic Information System of Web-based Culinary Tour with
Google API.(Siliwangi: Siliwangi University)
[7] Gorunescu F 2011 Data mining: Concepts, Models and Techniques (Verlag Berlin Heidel:
Springer)
[8] Han J, Kamber M 2006 Data Mining Concepts and Techniques Second Edition (San Francisco:
Morgan Kaufmann)
[9] Neteler M and Mitasova H 2004 Open Source GIS: A GRASS GIS Approach (Boston: Kluwer
Academic Publishers)
[10] Davies and Beynon P 2004 Database Systems Third Edition (Palgrave Macmillan: New York)
[11] Pramudiono 2007 Introduction to Data Mining: Mining Gems of Knowledge in Mount Data
http://www.ilmukomputer.org/wpcontent/uploads/2006/08/iko-datamining.zip
[12] Lei Xu C J, Wang J, Yuan J, and Ren Y 2014 Information Security in Big Data : Privacy and
Data Mining pp. 1149–1176
[13] Chauhan A, Mishra G, and Kumar G 2011 Survey on Data Mining Techniques in Intrusion
Detection vol. 2, no. 7, pp. 2–5
[14] Riadi I, Istiyanto J E, and Saleh S 2013 Internet Forensics Framework Based-on Internet
Forensics Framework Based-on Clustering
[15] Eiyanto N S, Mara M N 2013 Characteristics classification by method k-means cluster analysis
Bul. Llm., vol.02.,no.2, pp.133-136,2013.
[16] Wasserman S, and Faust K. 1994. Social Network Analysis: Methods and Application (New
York: Cambridge University Press)