Segmentation of Users Based on Usage Patterns

Saif Eldin Mohamed

221001554

Mustafa Sultan

221000432

Farida Hazem Nasrat

221000214

Doaa Mahdy Agha

221000424

Mohanad Ashraf Abdelwahab

222000061

Abstract— *This paper explores mobile application usage patterns in three French cities—Lyon, Marseille, and Montpellier—using data from NetMob 2023 Challenge dataset. We seek to determine when and where individuals interact via popular social media and entertainment apps (Instagram, Twitter and YouTube) to determine the most effective times and locations for marketers to advertise. Preprocessing of the data entails substituting means for missing values and removing outlier days using the Interquartile Range (IQR) approach. Next, we use heat maps to visualize the data by day of the week and determine the average traffic pattern overall. We use the elbow method to find the ideal number of clusters and the K-Means clustering technique to cluster geographic areas based on daily traffic. Our results provide useful information to app owners and advertisers to enhance* their ad pricing strategies and increase marketing reach.

Keywords— Outlier Detection - Clustering – Classification -Data Traffic - Data Analysis - Usage patterns

# Introduction

Nowadays, most of the time people spend is on their phones using social media. This led to a great shift of advertising focus going from TV advertisements to focusing more on social media apps. But to guarantee effective advertising, suitable locations and timings must be selected so that the advertisement is exposed to most of the audience. Another point of view that could be taken is one of the app owners. As an app owner, optimum monetization needs to be achieved so that maximum profit is gained. For that reason, the app owner should know how each time and location fare so that they can monetize it accordingly and how much they should charge for hosting an advertisement. This paper aims to analyze the usage patterns of the users of different cities across multiple apps. Those patterns will be analyzed to determine when the usage rate spikes and when it reaches its low. This will tell an advertiser when and where to advertise while guaranteeing maximum efficiency and give an idea to app owners on what basis they should monetize their apps. Data will first be cleaned to determine the outliers. After that, various clustering techniques will be used to help make the data more digestible. The apps that were picked for sampling were Instagram, Twitter, and YouTube. Those apps were selected due to their popularity and would be most suitable to host advertisements on.

# Literature Review

The widespread use of mobile applications and smartphones has significantly changed the digital landscape, influencing user behavior and advertising methods. This literature review examines previous studies and theories related to evaluating mobile app usage patterns, focusing on user segmentation, advertising efficacy, and clustering methods. Data preprocessing procedures, such as imputation for missing values and IQR for outlier detection, ensure the reliability and accuracy of data analysis. Clustering methods like K-Means are essential for data segmentation, with the elbow approach aiding in determining the ideal number of clusters, thereby improving balance accuracy and processing efficiency. The literature underscores the value of user segmentation, mobile advertising efficacy, and clustering methodologies. This study analyzes app usage in three French cities using K-Means clustering and the elbow method on mobile data from Orange, providing insights for advertisers and app owners to improve their strategies. According to Miklos Radics et. al.(2023) NetMob 2023 data can be useful for the estimation of the day and night population and grid cell level and can explain part of the dynamics of urban mobility.

# Methdology

This research aims to draw analysis and conclusions on the advertising field. The target is to manipulate data from 3 cities in France (Lyon, Marseille, and Montpellier) to reach a statistical and visual analysis to assist people investing in the advertising field.

#### Data Collection

Those cities were selected due to their diversity in population and difference in age demographics. The data used is from NetMob’s 2023 Challenge, which has been collected using open-source geospatial data and measurements from Orange, a major company in the mobile network field. Originally, it was collected utilizing coverage information from a commercial radio-frequency signal propagation tool. Its computation involves probabilities and applying Bayes’ theorem. For each city, there will be an analysis of the following applications: Instagram, Twitter, and YouTube. The apps were specifically chosen because of their excessive use in France. Per each application, there is a statistic of its network traffic every 15 minutes for 77 days (about two and a half months) starting from March 16, 2019.

#### Data Variables

The main variable in the data is the network traffic concerning time. Each application will be analyzed separately and record the observed difference between all of them for each city. After deciding on the peak hours of each application for every city, a comparison between different cities per application will be made.

#### Data Preprocessing

##### Handling null values

The data has been preprocessed to determine and detect outliers before drawing any conclusions. For each given day, scanning has been done to find the null values or non-numerical traffic values that could lead to crashing the program. Then all of them were replaced, with the average traffic for the rest of the day. That was found only on 31 March 2019 for four consecutive slots with a total duration of one hour from 11 pm to midnight.

##### Handling outliers

Further, as there is a possibility to find outlier values that are very different from the rest of the data, the average traffic value for each day’s hour was calculated and then all outliers were excluded using The Interquartile Range (IQR) method. IQR is the range between the first quartile (Q1) and the third quartile (Q3) of a dataset. The first quartile (Q1) is the value below which 25% of the data points lie, and the third quartile (Q3) is the value below which 75% of the data points lie. The IQR measures the spread of the middle 50% of the data.

##### Data Reduction

The data was manipulated using data frames and the NumPy library in Python methods. After removing all values that do not belong in the Interquartile Range and fixing the miswritten data, a new data frame was created. All the remaining days were grouped together, each according to their respective day of the week (e.g., all Mondays together, all Tuesdays together, etc.). Consequently, all days were reduced into a list of seven groups representing the average network traffic of their respective day of the week.

#### Data Visulaization

##### Visualization using Heatmap

After the data cleaning and reduction part, a heat map was created to visualize the average traffic for each day of the week. This assisted in identifying patterns and variations across different days, providing a clear and intuitive understanding of traffic distribution. Furthermore, the average traffic of all days was calculated to represent a single average traffic of each city.

##### Clustring

The K-Means clustering technique was used to group geographic areas based on their daily average traffic patterns. This involves calculating the daily average traffic for each area and then applying the K-Means algorithm to partition the areas into clusters with similar traffic patterns.

K-Means Clustering:

To determine the optimal number of clusters (K) for the K-Means algorithm, the elbow method was used. It plots the sum of squared distances (inertia) from each point to its assigned cluster center for different values of K. The optimal K is the point where the plot shows a sharp bend (elbow), indicating a little decrease in clustering performance with increasing K.

#### Data Segmentaion

Each of the previous methods was done for all three cities applications and the results and findings will be compared and discussed in the analysis section.

# Anaylsis

Our methodology was applied to three different cities, Lyon, Marseille, and Montpellier. We selected three cities to ensure that the data we analyze is representative and consistent, rather than an isolated occurrence. We will first describe our methodology and findings for the initial city, and then apply the same approach to the remaining two cities.

1. Marseille

Notice from these two graphs how the data got cleaner after all the noise caused by the outliers has been removed.

### Classification and Clustering

Below is a representation of the density of data traffic using classification and clustering.

A map of a country with red and blue colors

Description automatically generated

Figure- 1

A blue and white map

Description automatically generated

Figure- 2

### Heat Map

Next, a heat map of times where user activity spikes.

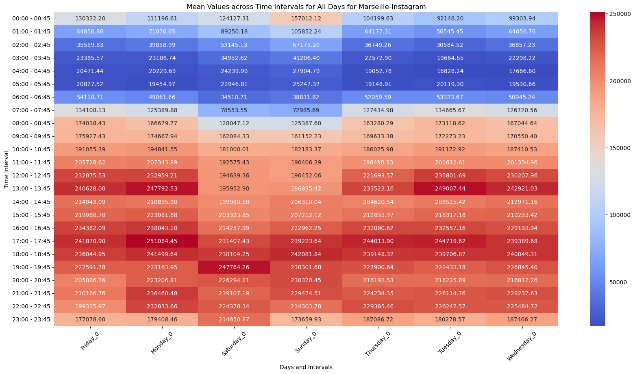


Figure- 3

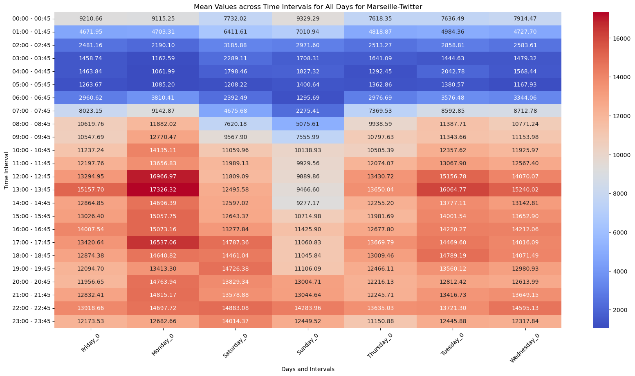


Figure- 4

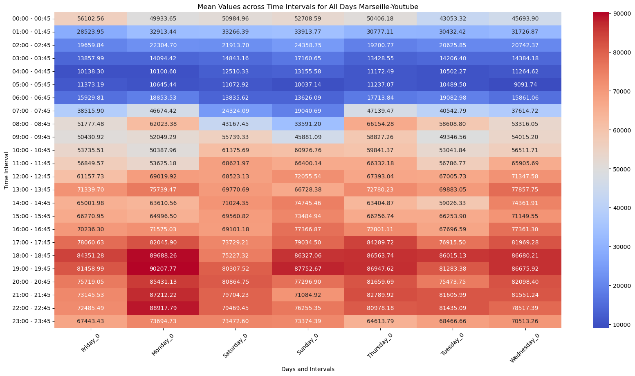


Figure- 5

## Montpellier

### Outliers:

Before:

A group of black and white lines

Description automatically generated

Figure- 6

After:A line drawing of a graph

Description automatically generated with medium confidence

Figure- 7

### Classification And Clustering

A map of a city

Description automatically generated

Figure- 8

### Heat Map

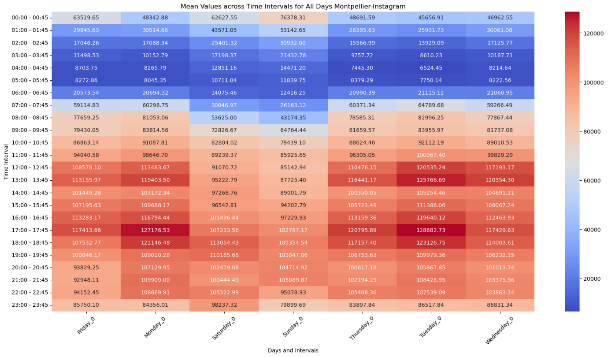


Figure- 9

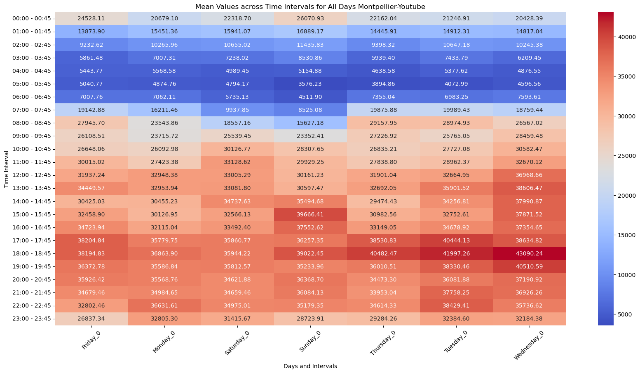


Figure- 10

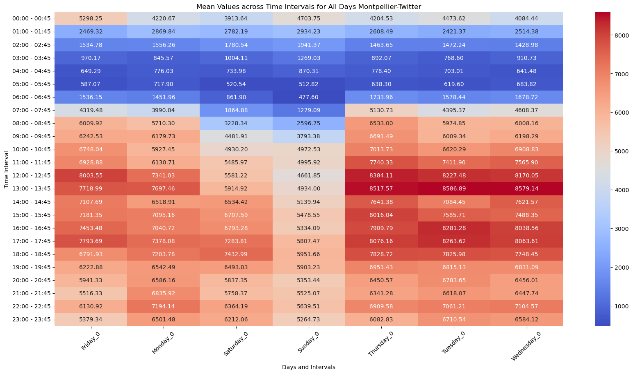


Figure- 11

## Lyon

### Outliers

Before: A black and white image of a graph

Description automatically generated

Figure- 12

After: A black and white image of lines

Description automatically generated

Figure- 13

### Classification And Clustering

A map of a city

Description automatically generated

Figure- 14

A blue and white map

Description automatically generated

Figure- 15

### Heat Map

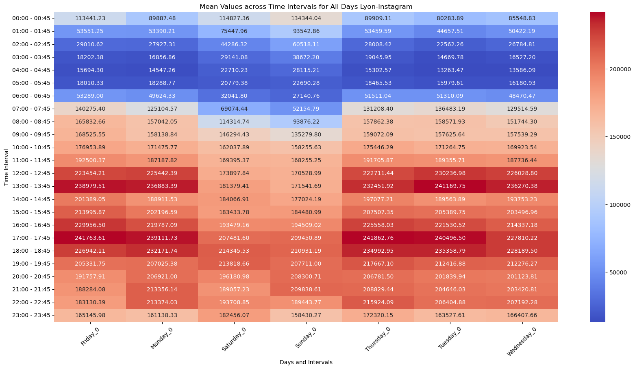


Figure- 16

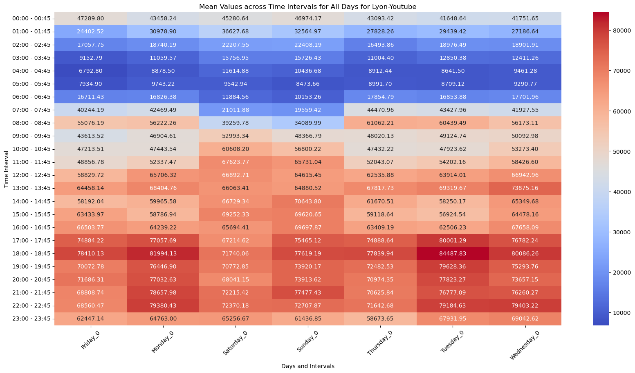


Figure- 17

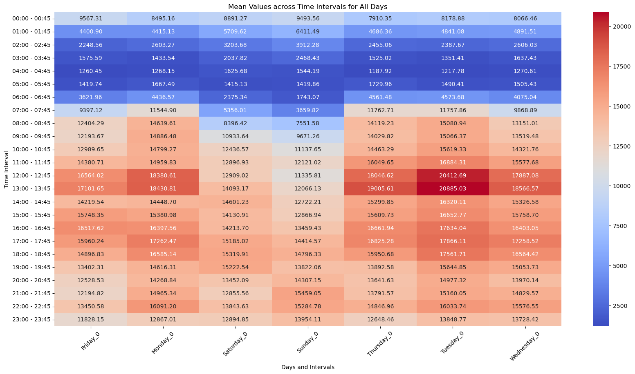


Figure- 18

## Findings

### Outliers

First, notice how much noise the data had before cleaning. Lyon’s Data was the noisiest as it had 33 outliers while Marseille was the least as it had 26 and Montpelier in between at 29 outliers. We deduce from this how stable and consistent each city is. Lyon being the noisiest means it is the most volatile while since Marseille had the least noise it is the most stable.

### Classification And Clustering

Secondly, from the clustering graph, it can be concluded that Marseille has the least spread of all three. This could be a good thing as it means when advertising, it can reach the same efficiency as the other cities while only investing in a small area. It can also be noted that the traffic clustering matches the population density distribution of each city.

### Heat Maps

The heat maps reveal the user activity patterns for each city, highlighting how they vary. In Lyon, people’s activity spikes at 1 pm, which could be due to lunch breaks. The heat maps also indicate a drop in usage on weekends, suggesting that people tend to use social media less on those days. This could indicate that people spend more time outdoors during weekends. Looking at Marseille we see that there are 2 spikes, one at 1 pm, the same as Lyon, and one at 5 pm. These 2 spikes could be great to utilize as they can increase the frequency of the advertisements at those times to reach a larger audience. Finally, looking at Montpellier we can see that it has a spike like Lyon’s but weaker in intensity with no other time that has a noticeable spike. This means that it will not be as efficient to cater the advertisement to time in Montpellier as it will not be a significant increase.

Furthermore, analyzing applications from the same city lends us important insights. For example, in the case of YouTube, traffic activity had its spike in the later hours of the day. This contrasts with Twitter, which tended towards traffic spikes during afternoon lunch hours. Meanwhile, Instagram traffic saw many spikes throughout the day. These differences align with each application’s different usage, with YouTube tending towards longer formatted videos and Twitter for numerous posts which is easier for consumption in a time constraint such as a lunch break. These differences underscore the importance of tailoring advertising strategies to the specific usage patterns of each platform.

# Summary And Conclusions

To summarize, using the analysis of the traffic data, it could be determined what areas had the most user activity, and which areas had the highest traffic density. It could also determine what times sees the most traffic of data and what times are considered a dead zone to avoid. This data is extremely useful to both an advertiser and an app owner as it lets them know how to ensure optimum profit from their business.

# Recommendations

We recommend applying our methodology and analysis to a larger scale as we are limited by time and equipment. This will provide a greater understanding of the patterns of people and their usage, which in turn will lead to a more profitable advertising system.

# Refrences

1. Christidis, Panayotis, et al. Modelling Daily Mobility Using Mobile Data Traffic at Fine Spatiotemporal Scale. Nov. 2023.
2. Christidis, P., Vega-Gonzalo, M., & Radics, M. (2023). Modelling daily mobility using mobile data traffic at fine spatiotemporal scale. <https://arxiv.org/pdf/2311.09683>