# k-means\_ color quantization

# Part 1: k-means clustering

## Overview

This part analyzes the Shopping parking occupancy of Parking Birmingham dataset. The original dataset was downloaded from [Parking Birmingham Data Set](https://archive.ics.uci.edu/ml/datasets/Parking+Birmingham)which contains data collected from car parks in Birmingham that are operated by NCP from Birmingham City Council. The dataset contains 35k instances with 4 variables. The data was placed into a public folder in DropBox and can be accessed [here](https://www.dropbox.com/s/fkal5ownnyy8p28/ParkingBirmingham.csv?dl=1)

The general approach to the analysis was:

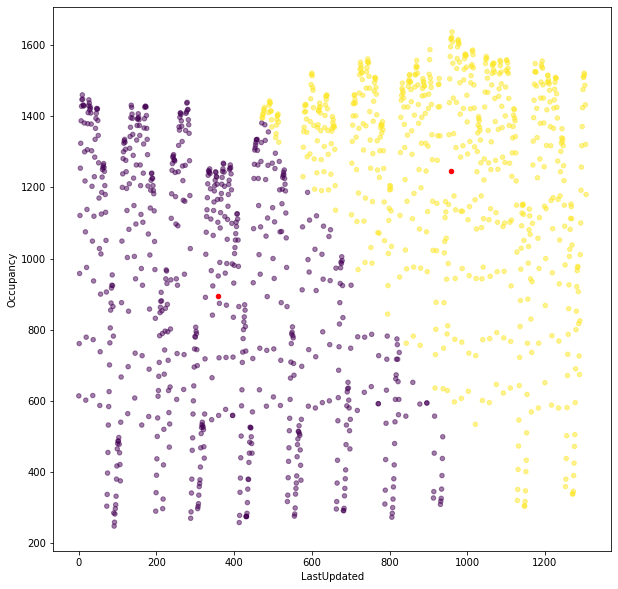
* Clean up data by identifying and correcting missing values and outliers.
* Convert columns with object type to numeric type
* The goal was to use k-means clustering to analyze the Shopping parking occupancy

## Data Clean Up

* Review the dataset to understand the overall structure and content of the data.
* Look for null value in data.
* Check data for outliners. The min occupancy is greater than 0, and the max occupancy is less than the capacity.
* Look at the columns and data types. Converted string data type to numeric type before applying k-means algorithm.

## Applying K-means to data

1. Clustering plot with k-means equal 2



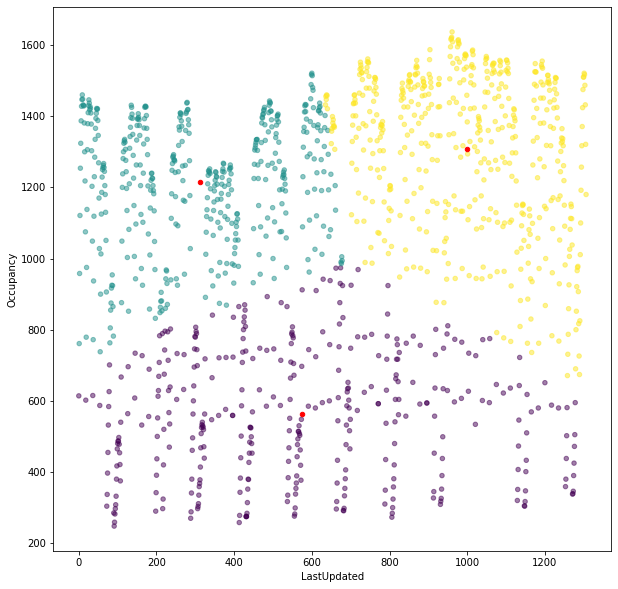
Sum of Squared Error (SSE) 208965492.32715625

The number of iterations required to converge 10

Locations of the centroids [[ 357.31137725 894.46556886]

[ 959.18012422 1245.83850932]]

1. Clustering plot with k-means equal 3



Sum of Squared Error (SSE) 121327071.02792938

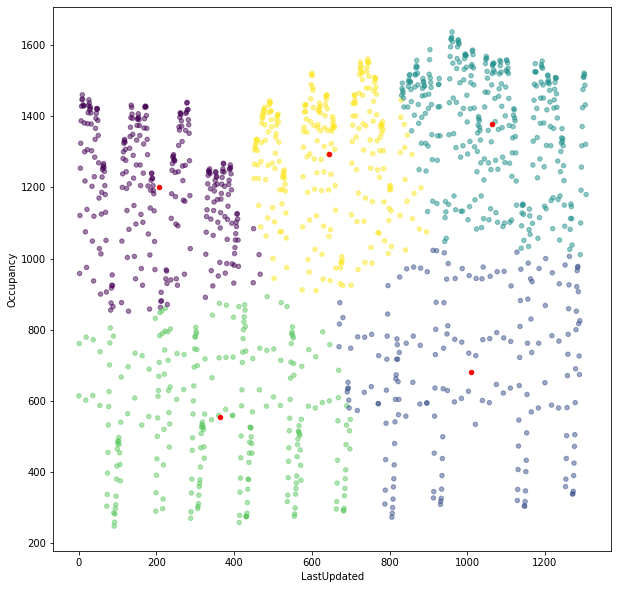
The number of iterations required to converge 7

Locations of the centroids [[ 574.12432432 562.77567568]

[ 311.90300231 1214.95381062]

[ 999.83497053 1307.50687623]]

1. Clustering plot with k-means equal 5



Sum of Squared Error (SSE) 68139854.67740366

The number of iterations required to converge 10

Locations of the centroids [[ 205.58802817 1201.27112676]

[1009.44148936 682.2712766 ]

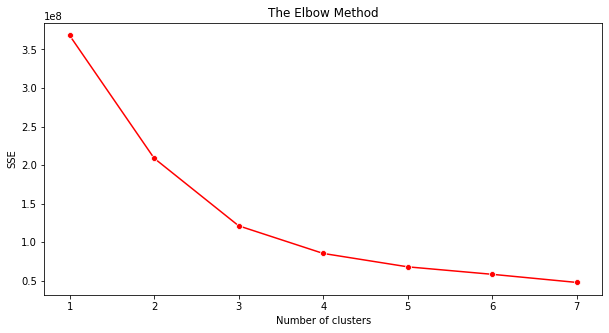
[1064.88650307 1378.63496933]

[ 363.16064257 555.99598394]

[ 643.97735849 1292.51698113]]

## Result Table

|  |  |  |  |
| --- | --- | --- | --- |
| Experiment  Number | Value of k | SSE Value | % Decrease of SSE (Compare to previous SSE) |
| 1 | 1 | 368227591.88 | N/A |
| 2 | 2 | 208965492.32 | 43% |
| 3 | 3 | 121327071.02 | 42% |
| 4 | 4 | 85516064.45 | 30% |
| 5 | 5 | 68139854.67 | 20% |
| 6 | 6 | 58446911.73 | 14% |
| 7 | 7 | 47913754.26 | 18% |



## Conclusion

We used the Elbow method to determine the optimal number of the clusters. From the graph, at x=3, the SSE starts decreasing in a linear fashion. Number of clusters equal 3 seems to be the reasonable trade-off between SSE and the number of clusters. Therefore, we conclude k-means equal 3 is the optimal number of clusters.

# Part II: Color Quantization