**YouView Recruitment Data Science Exercise**

You have been provided with a data set (YouView Recruitment Test Data Set v02.csv) containing weekly minutes of on-demand video consumption from 10,000 YouView set-top-box devices.

**Key to data**

|  |  |
| --- | --- |
| device id | A unique identifier for a consumer set top box device |
| application name | the name of the application which the device played (note that these have been obfuscated to e.g. “Application 1”) |
| weekly minutes of content played | the number of minutes of on-demand video content consumed by the specified device on the specified application within a 7-day period |

Important note:

**\* This data is proprietary and confidential to Youview and should not be shared with any third parties \***

The object of the exercise is to analyse the supplied data set to define a set of clusters based on the on-demand viewing behaviour of the 10,000 set-top-box devices, and to place each device into one of these clusters. You are then asked to use the output of this analysis to draw simple insights to help the business understand customer behaviour. The exercise should take two to three hours to complete, and it is expected that you will complete this work in Python or R.

We are looking to test your ability to prepare data, write code to perform a simple algorithmic analysis, critically analyse the output and relate that output to business-relevant insights.

Please include in your submission:

* Your code and any instructions that we may require to run it
* A brief analysis of your methods, reasoning and output
* How you checked the quality of your approach
* Your conclusions, including the insights you can draw about how different groups of users are consuming on-demand video content on the YouView platform

Tips on completing the exercise:

* Choose an appropriate clustering algorithm, e.g. k-means
* Please provide a graphical representation of your output and any intermediate steps that you feel are relevant
* After analysis of the data set you may conclude that this data cannot be partitioned into clusters in a meaningful way. If so, please provide evidence of how you came to this conclusion.
* Any interesting analysis above and beyond the ask of the exercise is a bonus!

Good luck!