**Clustering Project**

**1. Explore the data. Understand the variables you have. Run basic descriptions on the variables to confirm that they make sense. Make a list of issues you find.**

**Questions:**

* What does “abDisabledPc” mean? Maybe advertising disabled?
* How to choose the right clustering number based on BIC value?
* Data skew (Some observation have much more volume data compared with others)
* Which variables could be used to profile, which are should be clustered?
* Whether we need to standardize?

**2. Create variables as we discussed in class by aggregating the web data. You may use Python or dplyr/spread (or for extra practice, do it in both!). Submit a list of your variables, give basic descriptive (e.g., n, mean, min, max, sd).**

The variables we used in model:

|  |  |
| --- | --- |
| **Variables take into clusters** | |
| SubscriptionID | The variable indicates the customer id. |
| news | The number of articles the customer viewed relating to news. |
| business | The number of articles the customer viewed relating to business. |
| sports | The number of articles the customer viewed relating to sports. |
| food | The number of articles the customer viewed relating to food. |
| entertainment | The number of articles the customer viewed relating to entertainment. |
| local | The number of articles the customer viewed relating to local. |
| trivia | The number of articles the customer viewed relating to trivia. |
| chronicle vault | The number of articles the customer viewed relating to chronicle vault. |
| travel | The number of articles the customer viewed relating to travel. |
| ion | The number of articles the customer viewed relating to ion. |
| **Variables profile on** | |
| month | The month every visitor happens. |
| time | Accumulated time for each customer. |
| pv | Number of pages the customer viewed. |
| readingDayNum | The number of days the customer viewed in a certain month. |
| sessionNum | Section numbers the customer viewed. |
| mobileUse | The percentage of times when the customer use mobile to view. |
| tabletUse | The percentage of times when the customer use tablet to view. |
| desktopUse | The percentage of times when the customer use desktop to view. |

Based on the raw data, we aggregated the variables grouping by subscription id, regarding each customer as an observation:

|  |  |
| --- | --- |
| Max | ‘month’ |
| Mean | ‘mobileUse’, ‘desktopUse’, ‘tabletUse’, ‘abDisabledPct’, ‘abEnabledPct’,‘abnoResponsePct’ |
| Sum | All the variables relating to article contents, ‘time’, ‘pv’, ‘readingDayNum’, ‘sessionNum’ |

We selected the variables, which has been taken into cluster according to the article content, as the article content should be highlighted in market promotion. Then, we profiled the cluster on devices, time, and number of pages viewed.

**3. Identify 5-12 clusters based on reading behaviors**

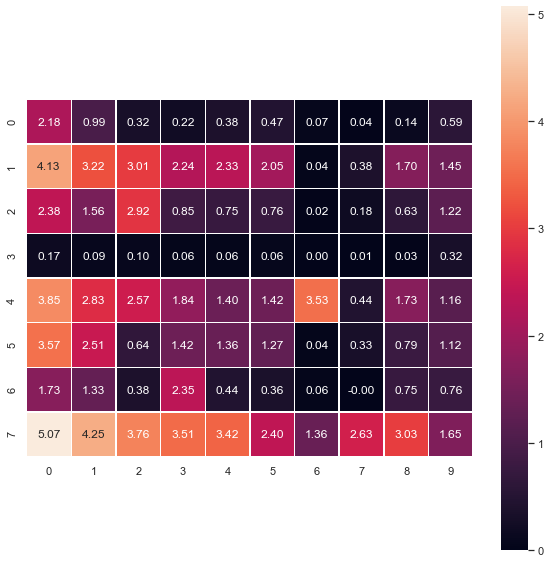
**Submit your clusters with names.**

**List which variables were used to dene the clusters. Give the cluster sizes and means on the variables used for clustering and proling. Discuss how you would recommend using the clusters.**

**I will give you comments and you will have the chance to revise your work.**

Clustering Means

As shown in the following image, on the Y axis are our 8 clusters. On the X axis are our 10 clustering variables. The color of the cells indicate how high or how low a value is.

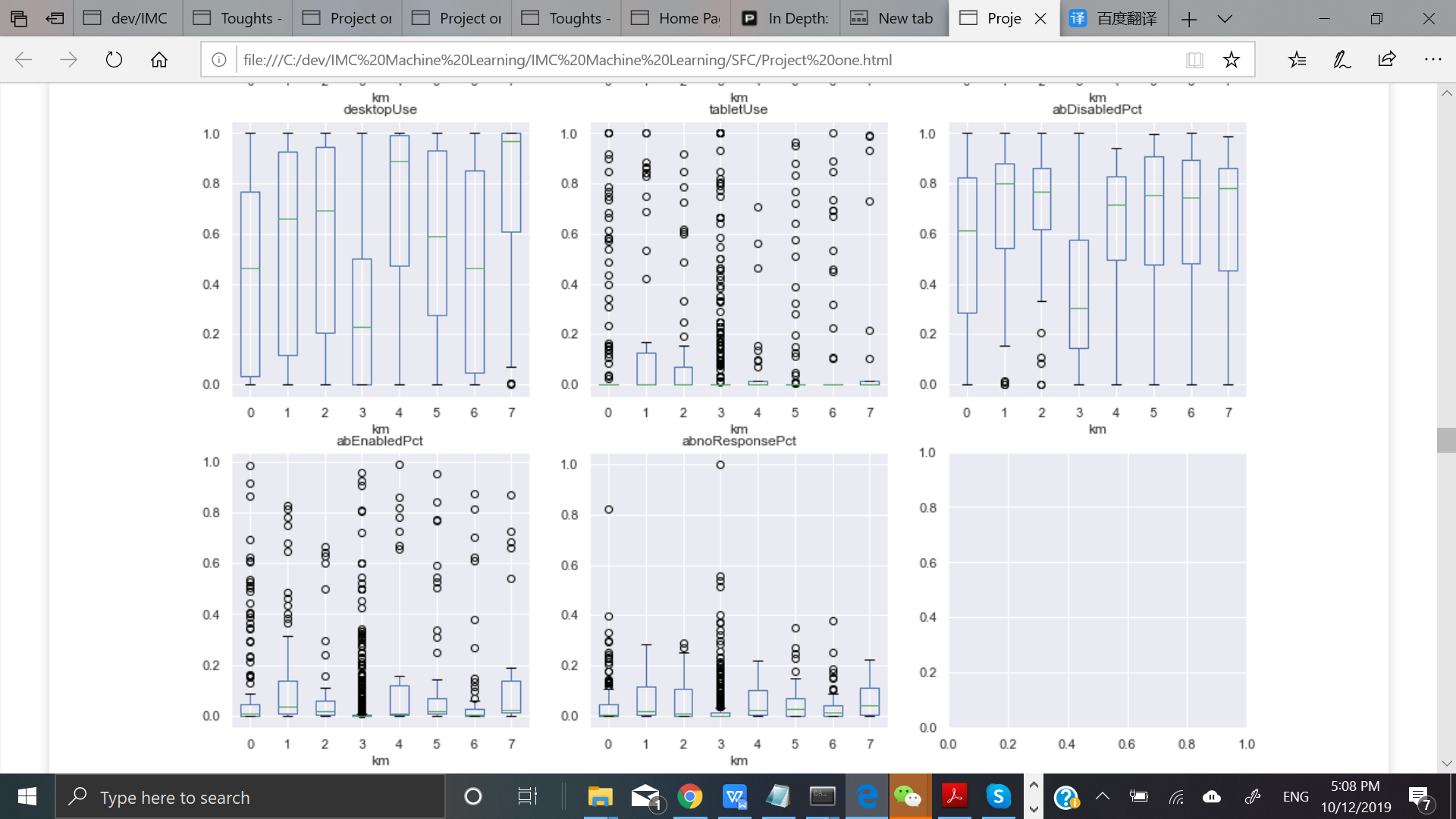


Cluster names:

|  |  |  |
| --- | --- | --- |
| cluster number | cluster name | size |
| 0 | Light users interested in news only | 195 |
| 1 | Medium active users with broad interests | 62 |
| 2 | Sports fans | 57 |
| 3 | Inactive users | 556 |
| 4 | Trivia seekers | 34 |
| 5 | Users interested in business and news (Professionals) | 88 |
| 6 | Food lovers | 67 |
| 7 | Highly active users with broad interests | 27 |

About Profiling





|  |  |  |  |
| --- | --- | --- | --- |
| cluster number | cluster name | size | Profiling/Recommendation |
| 0 | Light users interested in news only | 195 | Offer different kinds of News. |
| 1 | Medium active users with broad interests | 62 | Normally subscribe for 1 year. Offer a variety of contents. |
| 2 | Sports fans | 57 | Tend to use desktop more. Offer sport news mainly on desktop. |
| 3 | Inactive users | 556 | Subscribe for relatively short time. Offer discount for renew subscription. |
| 4 | Trivia seekers | 34 | Mainly use desktop, maybe not too busy. Offer trivia. |
| 5 | Users interested in business and news (Professionals) | 88 | Offer business and news. |
| 6 | Food lovers | 67 | Mainly use mobile. Offer food articles. |
| 7 | Highly active users with broad interests | 27 | Mainly use desktop. Can ask them for survey and referrals. |