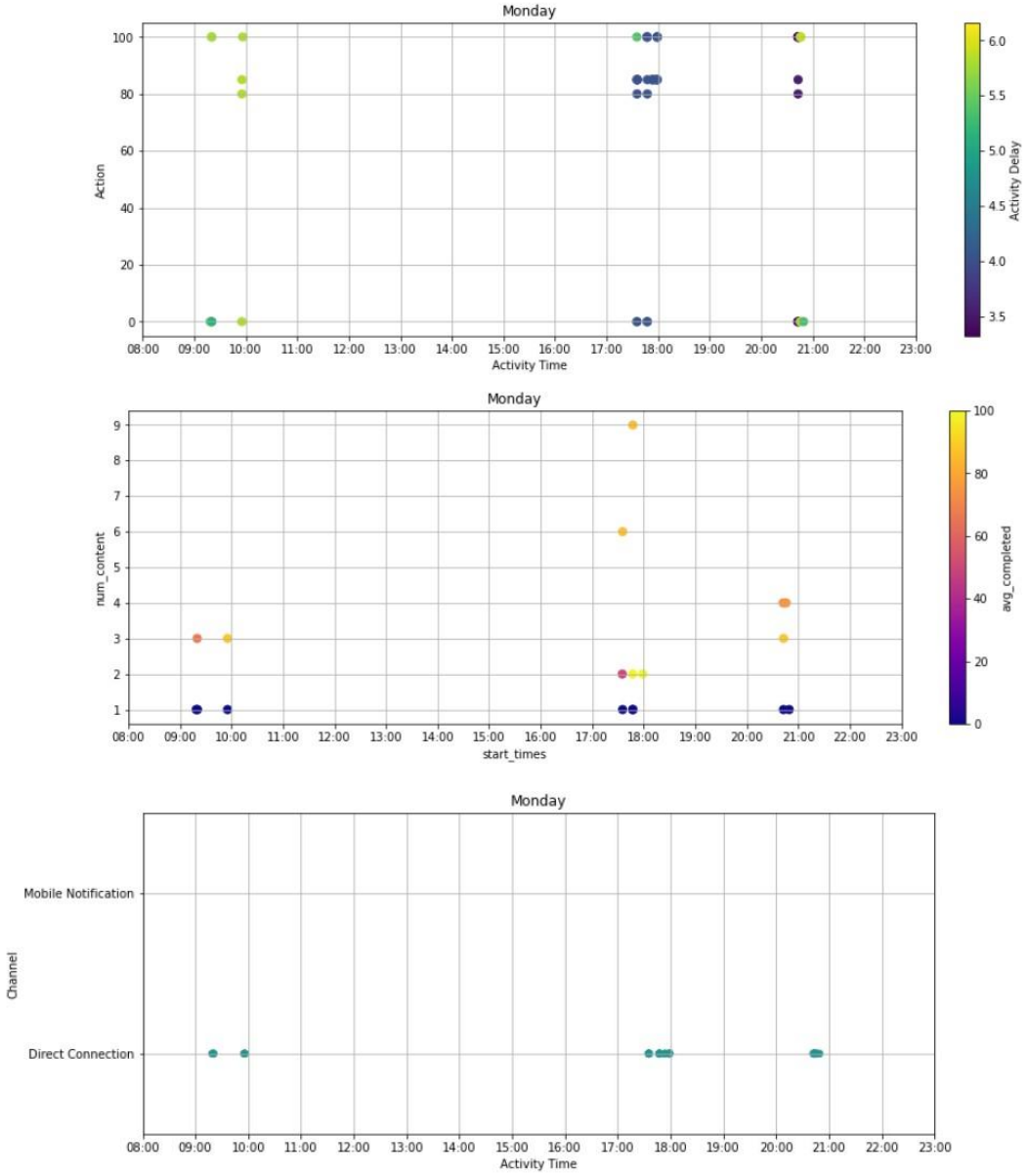
Deliverable 1 Spark Project Vibons

For the first phase of our project, we started by opening and inspecting the data. The dataset was provided by Vibons, the organization we are working with. We did not have to gather any additional data, so our analysis began with cleaning and processing the dataset. This was fairly straightforward. The columns of the dataset include customer Id and User Id, which are just ID numbers for each corporate client and individual employee who are signed up with Vibons. User Created At indicates the time that an individual was entered into the Vibons database. Activation Date and Activity Date are the date and time that the user received a piece of content from Vibons, and the time at which they opened it, respectively. Name is the name of the piece of content received by the user. Content Type is the format of the content, it includes several categories such as infographic, video, and flipbook. Journey Name refers to the subject of the content. Action refers to the percentage of the content completed by the user; it ranges from 0 to 100. Duration represents the amount of time the user spent looking at the content. Device represents the platform on which the user accessed the content, for example, the app on Android or iOS or a mobile browser. Channel refers to whether they accessed the content through a mobile notification or a direct connection. Session Id gives a unique identifier for each session that they begin looking at content from Vibons. Rating contains a value for the rating assigned by the user to the content, although most rows don't have a value in this column.

After loading, examining and doing some like cleaning on the dataset (i.e. dropping certain corrupted rows) we decided to add a few columns of our own to the dataset. First we looked at Activation Date and Activity Date and we went through and added a new column that just represented the day of the week for Activation and Activity. Our motivation for this was that trends for each user in terms of content consumption may be different depending on the day of the week, so we wanted to be able to isolate their trends and habits by day. We also added a column called "Activity Delay" this was the amount of time elapsed between Activation Date and Activity Date, i.e. how much time went by between the content being received by the user and being opened by the user. There were some cases where hundreds or even thousands of hours went by between receiving the opening the content, so we applied a log transformation here to make the scale of the data more manageable (and to visualize it more easily on the graphs that we created; however, this transformation can be easily undone at a later stage).

We also created a list of each user, and the number of rows in which they appeared in the data. Then we calculated the mean action and standard deviation to be able to quickly find the percent of the content that someone was completing. We also broke down the sessions by each user, calculating the start time of each session, the amount of content accessed during each session and the average percentage of content completed by each user during a session

We took this information (particularly the columns that we added to the dataset) and generated some graphs to help visualize the data to see if we could visually ascertain any insights. A couple of sample graphs are as follows (we generated separate graphs for each day of the week



The three graphs we constructed all have time on the x-axis. The first one has each row represented as a datapoint, with Action, or pecent completion on the y-axis. The color of each dot indicates the activity delay with the colorscheme in the bar on the right. The second graph indicates each individual session started by the user (each point on the graph represents one session). The amount of content completed by the user in each session is indiciated by the y-axis, and the color scheme represents the average percent completed for the content in that session. Finally, the third graph shows the timing of the content accessed by either direct connection or mobile connection by the user. When generating these graphs for every day of the week for a certain user, they reveal on which days the user is most likely to access content, and at what times the user is most likely to access the content.