* We need to segment our customers based on their *frequency* and *order* value
* Which segment could be a valuable target group for a Marketing campaign about Coffee (“Breakfast” cuisine\_parent)?

--To be frank, I like BigQuery, but I work much faster on my beloved dBeaver and PostgreSQL.

--First of all, I tried to find out if all those available tables on BigQuery were supposed to be related - I guess not.

--SELECT count(\*), count(distinct shop\_id) FROM `bi-2019-test.ad\_hoc.business\_intelligence\_analyst\_2019\_restaurants` where shop\_id=10881;

--select \* from `bi-2019-test.ad\_hoc.business\_intelligence\_analyst\_2019\_restaurants` where shop\_id=10881

--no shop\_id matches with the ones found in table business\_intelligence\_analyst\_2019\_restaurants

--SELECT count(\*) FROM orders\_jan2021 where shop\_id in (select shop\_id from `bi-2019-test.ad\_hoc.business\_intelligence\_analyst\_2019\_restaurants`)

--SELECT count(\*) FROM orders\_jan2021 where shop\_id in (select \* from `bi-2019-test.ad\_hoc.bi\_dummy\_orders` LIMIT 1000)

--Thus, I took the liberty to download the data, since the dataset was quite small, after all.

-- I ran this on BigQuery:

**SELECT** \* **FROM** `bi-2019-test.ad\_hoc.orders\_jan2021`

--and then exported the csv results to my Google Drive with the SAVE RESULTS button.

--Then, I moved to PostgreSQL, created a table with practically the same datatypes:

**CREATE** **TABLE** public.orders\_jan2021 (

order\_id **int8** **NULL**,

brand **varchar** **NULL**,

submit\_dt **timestamp** **NULL**,

user\_id **int8** **NULL**,

shop\_id **int8** **NULL**,

city **varchar** **NULL**,

cuisine\_parent **varchar** **NULL**,

basket **numeric** **NULL**

);

--and imported the csv data

**SELECT** \* **FROM** orders\_jan2021;

--Then, I started exploring it; e.g. I saw we have max 33 orders per user:

**select** **count**(\*),user\_id **from** orders\_jan2021 **group** **by** user\_id **order** **by** 1 **desc**;

--the only brand is GR-EFOOD:

**select** **count**(\*),brand **from** orders\_jan2021 **group** **by** brand **order** **by** 1 **desc**;

--The highest selling shop got 5733 orders:

**select** **count**(\*),shop\_id **from** orders\_jan2021 **group** **by** shop\_id **order** **by** 1 **desc**;

--And that was a surprise, the shop is in Lesvos! Even though I know Athens and Thessaloniki are excluded...

--I suspect our Army did this. :-) Or maybe immigrants and refugees??

**SELECT** \* **FROM** orders\_jan2021 **where** shop\_id=10881;

--So we have 400k orders

**SELECT** **count**(\*) **FROM** orders\_jan2021;

--Find number of users per order frequency:

**select** **count**(\*),cnt **from** (

**select** **count**(\*) cnt ,user\_id **from** orders\_jan2021 **group** **by** user\_id

) a **group** **by** cnt **order** **by** cnt **desc**;

--Create a new table to save the orders frequency along with the total sales:

**create** **table** clustering **as**

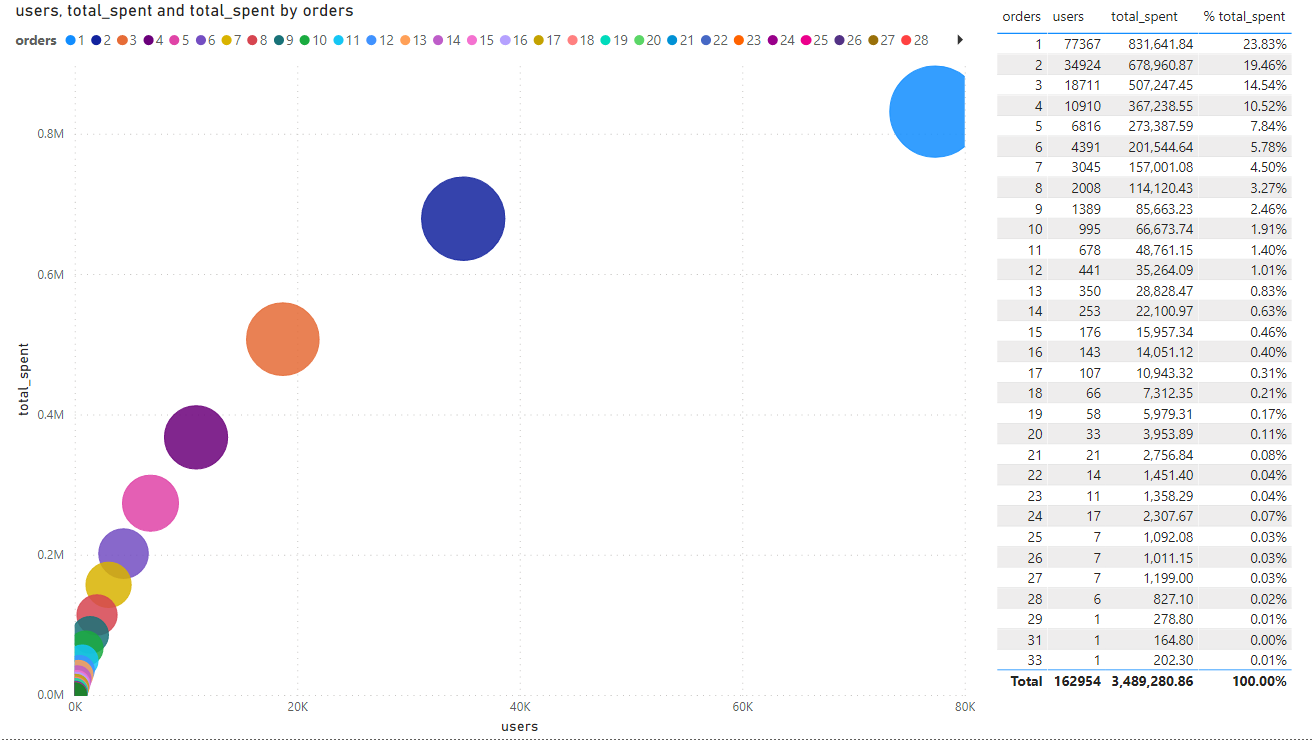
**select** **count**(\*) **as** users,orders,**sum**(total\_spent) total\_spent **from** (

**select** **count**(\*) orders , **sum**(basket) **as** total\_spent, user\_id **from** orders\_jan2021 **group** **by** user\_id

) a **group** **by** orders **order** **by** orders **desc**;

|  |  |  |
| --- | --- | --- |
| **users** | **orders** | **total\_spent** |
| 77,367 | 1 | 831,641.84 |
| 34,924 | 2 | 678,960.87 |
| 18,711 | 3 | 507,247.45 |
| 10,910 | 4 | 367,238.55 |
| 6,816 | 5 | 273,387.59 |
| 4,391 | 6 | 201,544.64 |
| 3,045 | 7 | 157,001.08 |
| 2,008 | 8 | 114,120.43 |
| 1,389 | 9 | 85,663.23 |
| 995 | 10 | 66,673.74 |
| 678 | 11 | 48,761.15 |
| 441 | 12 | 35,264.09 |
| 350 | 13 | 28,828.47 |
| 253 | 14 | 22,100.97 |
| 176 | 15 | 15,957.34 |
| 143 | 16 | 14,051.12 |
| 107 | 17 | 10,943.32 |
| 66 | 18 | 7,312.35 |
| 58 | 19 | 5,979.31 |
| 33 | 20 | 3,953.89 |
| 21 | 21 | 2,756.84 |
| 17 | 24 | 2,307.67 |
| 14 | 22 | 1,451.40 |
| 11 | 23 | 1,358.29 |
| 7 | 27 | 1,199 |
| 7 | 26 | 1,011.15 |
| 7 | 25 | 1,092.08 |
| 6 | 28 | 827.1 |
| 1 | 33 | 202.3 |
| 1 | 31 | 164.8 |
| 1 | 29 | 278.8 |

--After loading the data to PowerBI, I decided to split the userbase to five segments, mainly because of this scatter plot, based on which I believe the reasonably separated areas are five:



--Then I saved each user's segment along with the original data:

**create** **table** orders\_jan2021\_cl **as**

**select** oj.\*,

**case** **when** orders=1 **then** 'SINGLE-ORDER'

**when** orders=2 **then** 'TWO ORDERS'

**when** orders=3 **then** 'THREE ORDERS'

**when** orders=4 **then** 'FOUR ORDERS'

**when** orders>4 **then** 'FIVE+ ORDERS' **end** **as** **cluster**

**from** orders\_jan2021 oj **left** **join**

(**select** **count**(\*) orders, user\_id **from** orders\_jan2021 **group** **by** user\_id) b

**on** oj.user\_id = b.user\_id;

--First of all, Breakfast is the most popular cuisine, even though it comes second in revenue:

**select** **count**(\*) orders,**sum**(basket) sales,cuisine\_parent **from** orders\_jan2021 **group** **by** cuisine\_parent **order** **by** 1 **desc**;

--This shows that the FIVE+ ORDERS segment is the best to use for a campaign about Coffee (cuisine\_parent='Breakfast')

--because they have spent the highest % of their orders in that category:

**select**

**sum**(basket) **as** total\_spent,

**sum**(**case** **when** cuisine\_parent='Breakfast' **then** basket **else** **null** **end**) **as** total\_spent\_breakfast,

**round**(**sum**(**case** **when** cuisine\_parent='Breakfast' **then** basket **else** **null** **end**)/**sum**(basket)\*100,2) **as** **percent**,

**cluster** **from** orders\_jan2021\_cl **group** **by** **cluster** **order** **by** 3 **desc**;

|  |  |  |  |
| --- | --- | --- | --- |
| **total\_spent** | **total\_spent\_breakfast** | **percent** | **cluster** |
| 1,104,192.15 | 334,210.54 | 30.27 | FIVE+ ORDERS |
| 367,238.55 | 83,801.66 | 22.82 | FOUR ORDERS |
| 507,247.45 | 101,318.46 | 19.97 | THREE ORDERS |
| 678,960.87 | 112,742.66 | 16.61 | TWO ORDERS |
| 831,641.84 | 113,579.15 | 13.66 | SINGLE-ORDER |

--Here is another piece of proof that, virtually regardless of the time the order is placed, the 5+ cluster is the absolute winner:

**select** **case** **when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 5 **and** 11 **then** 'MORNING'

**when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 12 **and** 17 **then** 'NOON'

**when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 18 **and** 23 **then** 'NIGHT'

**when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 0 **and** 4 **then** 'AM' **end** **as** time\_of\_day,

**sum**(basket) **as** total\_spent,

**sum**(**case** **when** cuisine\_parent='Breakfast' **then** basket **else** **null** **end**) **as** total\_spent\_breakfast,

**round**(**sum**(**case** **when** cuisine\_parent='Breakfast' **then** basket **else** **null** **end**)/**sum**(basket)\*100,2) **as** **percent**,

**cluster**

**from** orders\_jan2021\_cl **group** **by**

**case** **when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 5 **and** 11 **then** 'MORNING'

**when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 12 **and** 17 **then** 'NOON'

**when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 18 **and** 23 **then** 'NIGHT'

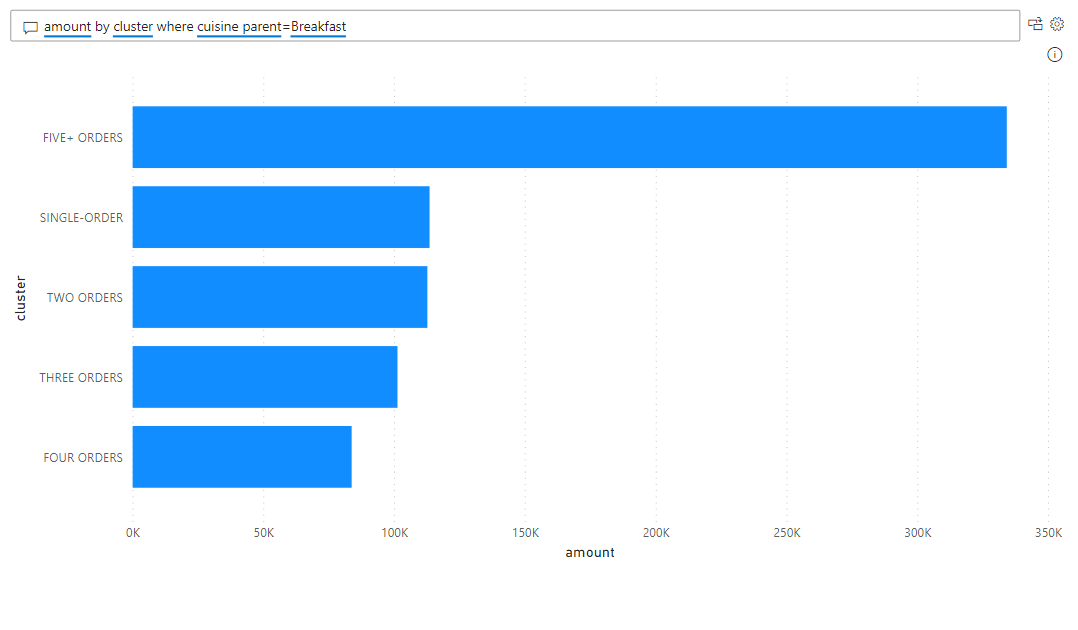
**when** **to\_char**(submit\_dt,'HH24')::**integer** **between** 0 **and** 4 **then** 'AM' **end**,

**cluster**

**order** **by** 3 **desc limit** 10;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **time\_of\_day** | **total\_spent** | **total\_spent\_breakfast** | **percent** | **cluster** |
| MORNING | 152,895.75 | 144,287.04 | 94.37 | FIVE+ ORDERS |
| NOON | 361,368.15 | 132,335.86 | 36.62 | FIVE+ ORDERS |
| NIGHT | 550,545.68 | 52,721.50 | 9.58 | FIVE+ ORDERS |
| MORNING | 48,226.67 | 44,930.07 | 93.16 | TWO ORDERS |
| NOON | 227,970.83 | 43,452.74 | 19.06 | SINGLE-ORDER |
| NOON | 191,809.36 | 43,361.62 | 22.61 | TWO ORDERS |
| MORNING | 45,184.80 | 41,437.65 | 91.71 | SINGLE-ORDER |
| MORNING | 43,245.88 | 40,787.38 | 94.32 | THREE ORDERS |
| NOON | 148,426.16 | 39,331.32 | 26.5 | THREE ORDERS |
| MORNING | 37,503.56 | 35,263.65 | 94.03 | FOUR ORDERS |

--And tried to verify the segmentation by asking PowerBI a simple question:



--Finally, I visualized the best selling areas - maybe it would be a good idea to do some A-n testing for our marketing campaigns along the top 5 cities:

