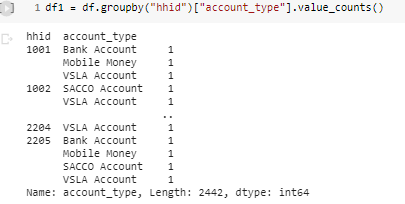
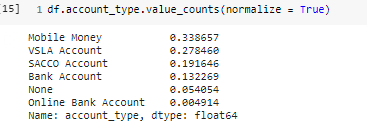
For **\*\*Question 1\*\***, we may want to see display how many products a single ID has. For doing so, we use the function below.



For **Question 2**, I created two dummy variables for each ID: **Financially excluded** and **Digital inclusion**. Both can be seen as boolean expresions but for predicting purposes I rather prefer to transform them into integers so we can later on do some ML predictions

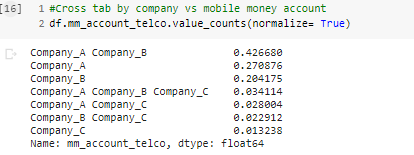
I also consider that from now on, its important to do some EDA so we can have an intuition of which variables may be correlated to the dependant variable, **which is how likely clients will cancel their mobile money account (mm\_account\_cancelled)**.

So as we can see, Mobile Money shares is 33%, so it is the most popular among the remaining. Slightly, VSLA Account also counts for 28% of the share. So from this perspective, not that the Mobile Money will remain the majority of preference among clients. So first off, worth it to explore which target prefer VSLA Account, how they ended up adquiring this product. So far, VSLA Account is the traditional product among the sample of this study, which is majority rural. From Question 1 descriptive, we have seen that clients can have both products. From a microfinance perspective, it is ok to increase the number of products among customers, but mobile money operators only require a mobile and internet to operate, so worth it the make it the preferred one among the clients.



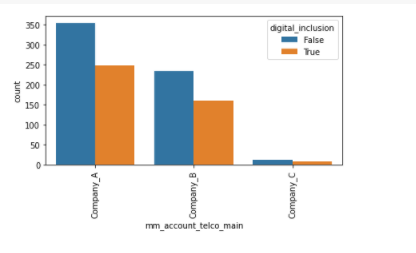
**\*\*Question 3. \*\***

As we can see in the descriptive below, Commpany\_A and Company\_B are the ones that customers prefer to have a mobile money account. So we can do the following actions: stick to Company A and B for promotions and that way increase share, or seeing Company C as a potential brand that need to fuel some promotions and advertising so they can increase market share.

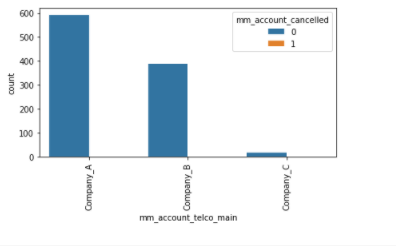




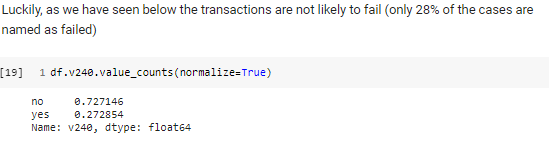
Among all the dataset, Its company A and B indeed, the ones that are preferred among the clients that are digital included.



I also explored, if there is relationship between cancelling a mobile account and the providers for the mobile money account. In the graph below, in the legend, 0 stands for not cancelled and 1 for cancelled. As we can see, none of the companies are related to cancelled in the past. Actually the company that is doing a good performance in this regard is Company A.

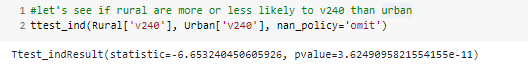


The variable v240 which ask if a transaction ever failed to go through, its important explore and so forth, to clean it up because this feature may cause that clients decide it to keep digital products or to drop them. So that's why I decided to clean it up.



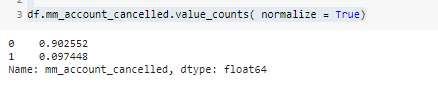
**\*\*\*Question 4 \*\***\*asks for finding statistical differences in the share of customers who have experienced failed mobile money transactions in rural and urban villages. For that we need first to load packages and later, doing some data transformation for establishing the different groups to compare.

For responding this question I used the independent sample t-test among rural vs urban population. So that is why I split the dataset and the ouput below shows that there is not significant differences regarding rural vs urban populations. This means that the kind of geography is not related with a crash in the transations that a client may display. So we may assume that that the companies that operate in those areas are not having difficulties in this regard.



**Question 5** asks for the variables that are good predictors that someone will cancel their mobile money account? Discuss what causes a customer to stop using their mobile money account including how strong the evidence is.

In order to respond this question, still worth it to do some EDA so we can select the potential independent variables that may explain whether a person is likely to cancel a mobile money account or not. I start off with using **\*mm\_account\_cancelled \***as a dependent variable. As we can see 90% of people hasnt cancelled their mobile account. This is a good indicator but for predicting it may causes a problem because we notice that the Dependent variable is imbalance, so it worth to balance with different approaches. One of the most common ones is SMOTE, but given the limitation of time I'm not doing imbalancing for this experiment



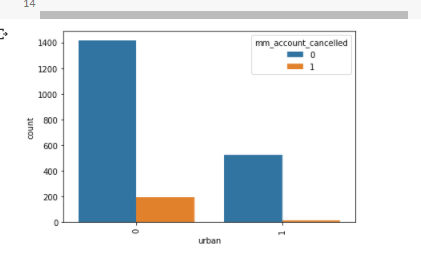
#With 0 being rural, among them, there are few chances of cancel your mobile account(since 0 is no cancelled)

#With 1 being urban, among them, there are few chances of cancel your mobile account.

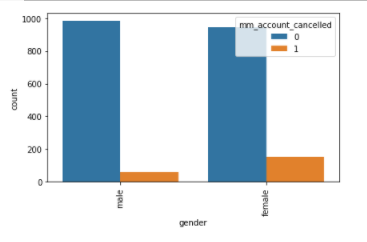
#Still, seems that cancel is more likely to happen in rural areas. But this may looks this way because

#of the imbalanced data.

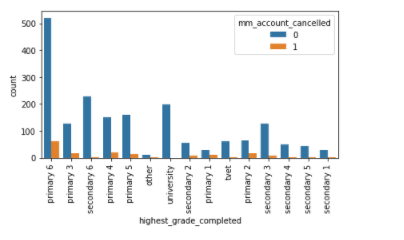
#Overall, seems that region variable does not have an impact whether or not someone cancel



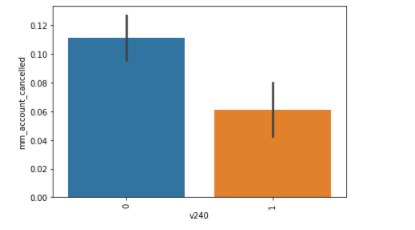
#With gender, it seems that women are more likely to cancel than men. So, Sex may be not a good predictor for canceling or not.



Education seems a good variable to include, so may be worth it to include it.



Has a transaction ever failed to go through? (v240) with 0 being no and 1 being yes.

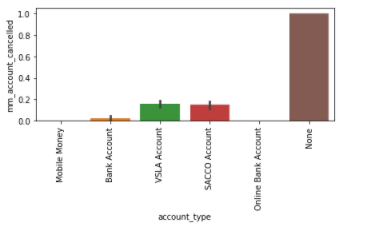


#The graph shows us that both VSLA Account and SACCO Account owners are more likely to cancel their

#mobile account. They may be more comfortable with this kind of accounts.

#may be worth it to address on a qualitative level what are the attributes that they like

#from both accounts to stick with them.



It was the attempt of this project to select the features that better predict the depedent variable. Due to the time that requires pre-process the independent variables and the later on techniques to fined-tuned the features the author of this paper could not implement it for that. The descriptives above however, can tell us before hand which variables seems to be good predictors.