`Capstone 1 Statistical Data Analysis Report

From previous analysis the following trends were noticed:

Customers with balance in a certain range seem to tend to exit?

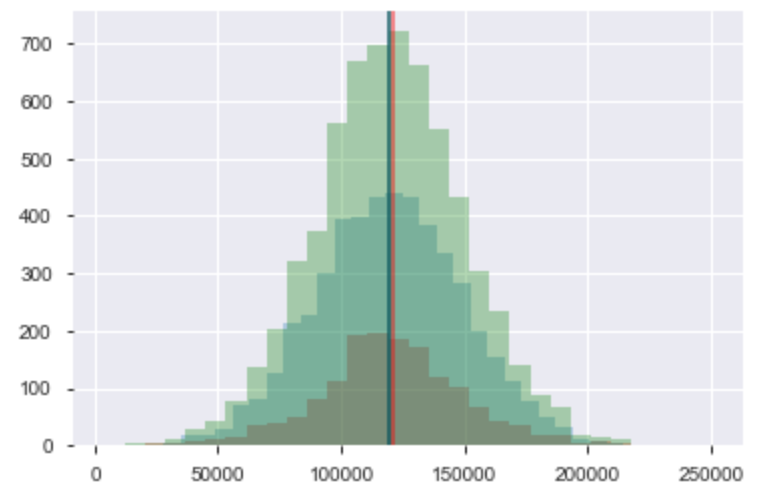
A lot of zero balance customers and do they behave/distribute differently?

Not active after age 45 especially females tend to exit?

Customers who purchased 3-4 products tend to exit?

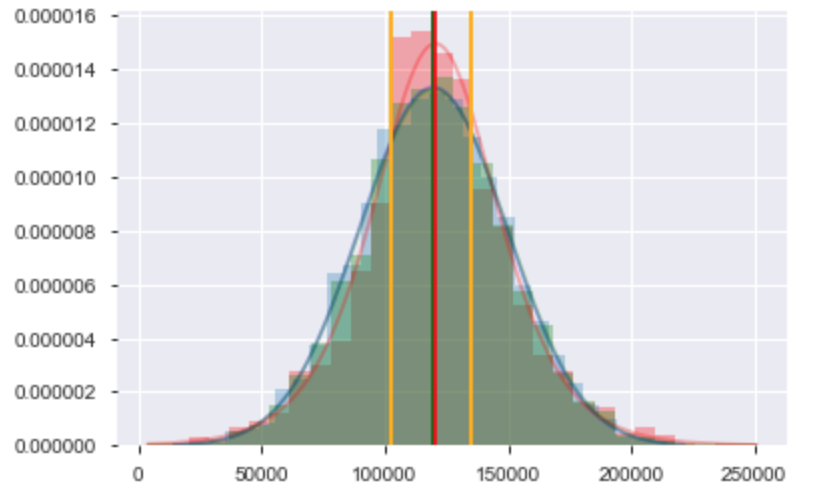
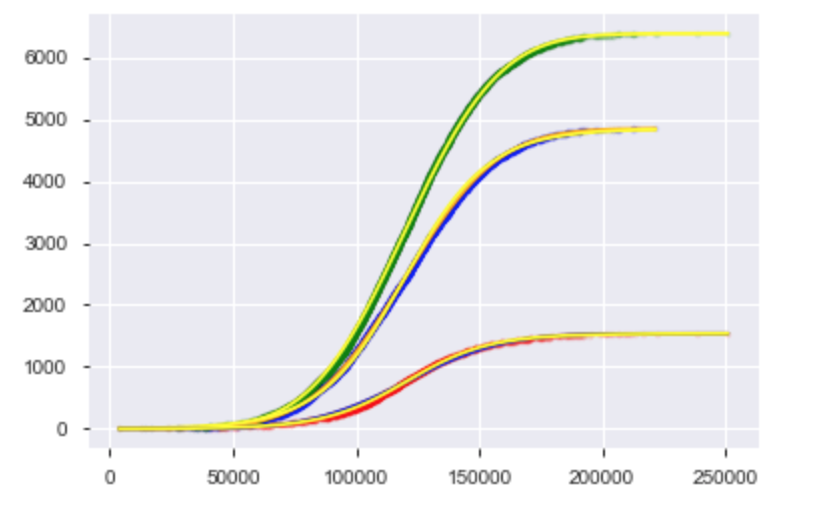
German customers tend to have more balance and also exit?

To further investigate these the following steps were performed:

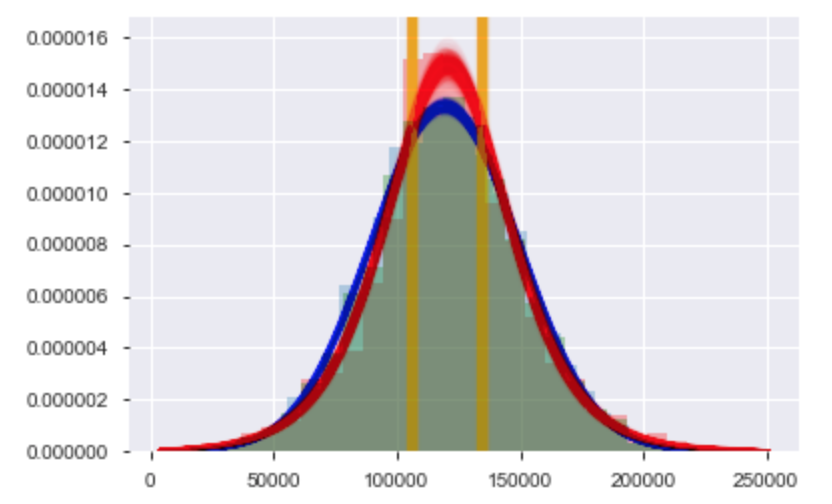


By plotting non zero balance, we can see non zero stayed customers’ balance, non zero exited customers’ balance have very similar distributions.

Then by plotting the ECDFs and trying different regressions (normal, t, logistic) we find non zero exited customers’ balance seems to be logistically distributed while the other two groups are normally distributed.

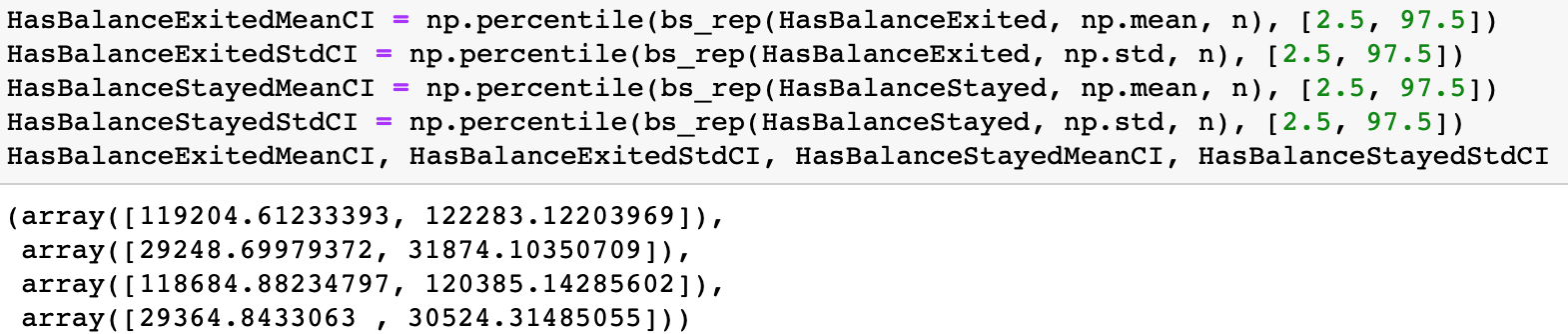


By normalizing the histogram we can tell customers of a certain range of balance (around [25, 70] percentile of exited group) do tend to leave more. And the regression functions fit well with the empirical data.

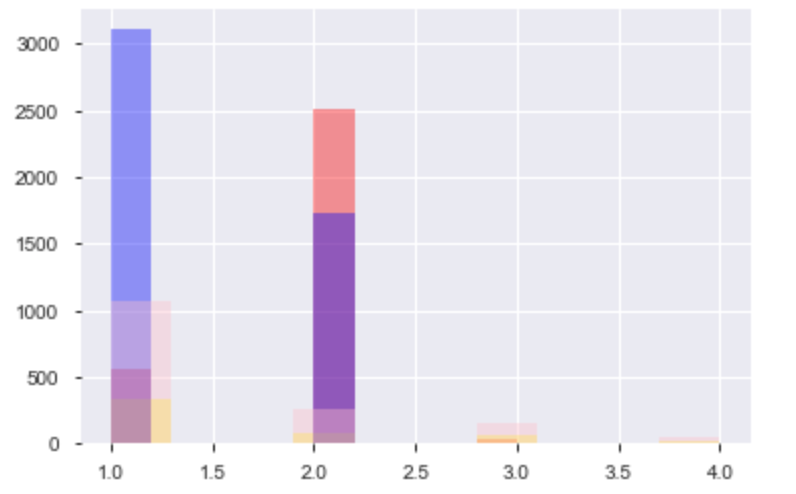


By performing bootstrap sampling and regressions (n=1000) we can get confidence intervals of the regression functions and we can tell around [30, 70] percentile of the exited group is where customers tend to leave.

Confidence intervals for means and stds of the groups (n=10000)



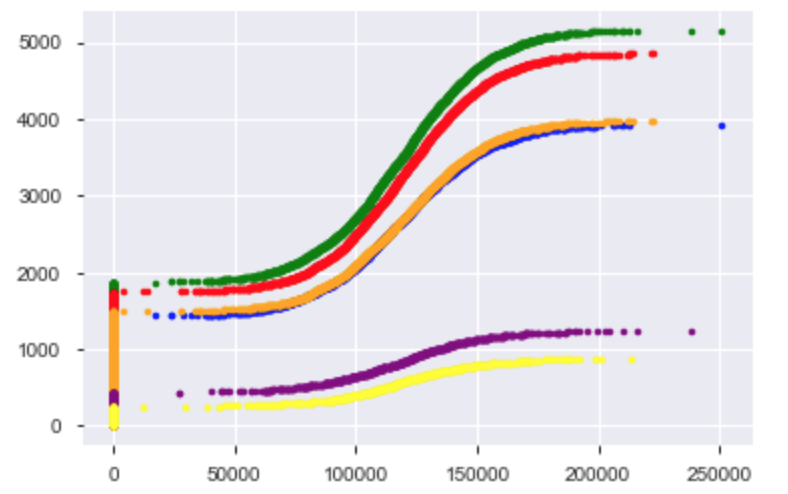
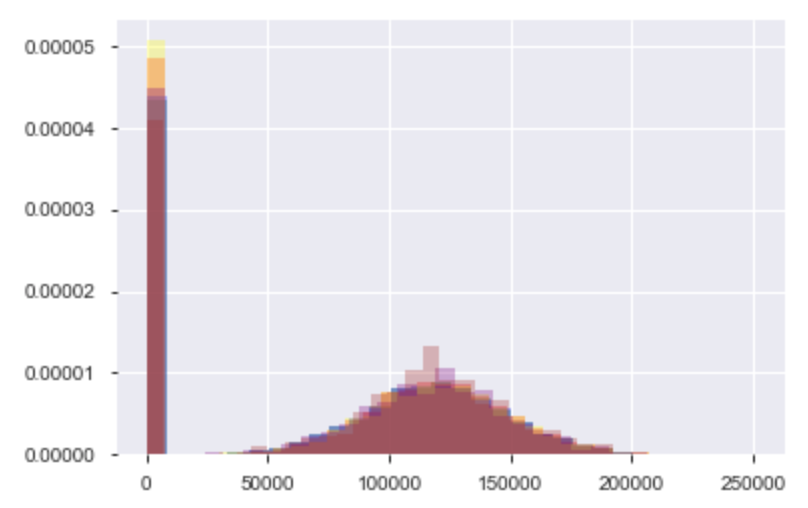
By grouping customers into has balance and zero balance group and comparing distributions on age, credit score, number of products, tenure, estimated salary, credit card and active status, we find zero balance customers only show different distribution on number of products.



Then we divide customers by balance (zero and non zero) and exited (exited and stayed) and see how many products they purchased.

It shows that stayed customers tend to buy only 1-2 products and stayed with zero balance group tends to buy 2 while stayed with positive balance group tends to buy buy 1. But overall more customers tend to buy less products.

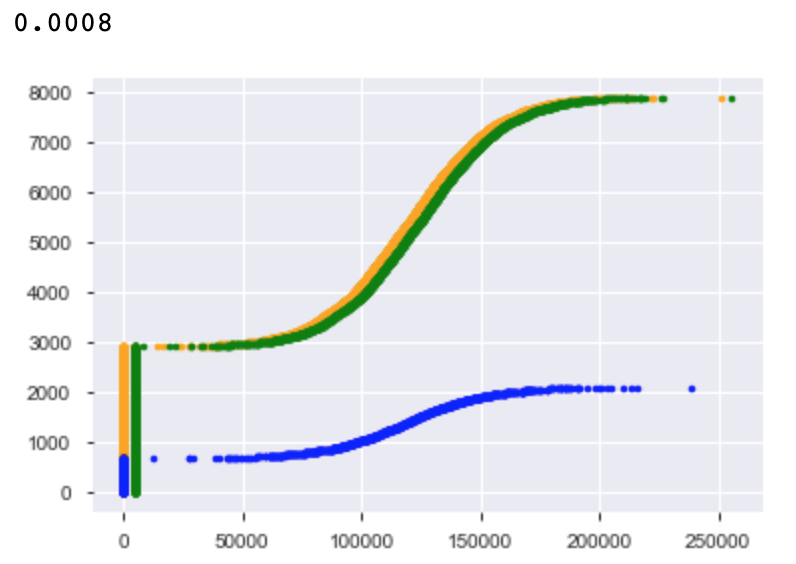
By grouping customers by active status (active and not active) and age (<=45 and >45) and plotting a normalized histogram we notice that they have very similar distributions of non zero balance. Age > 45 not active group has less members with zero balance and more of balance within the range just mentioned. Not active groups tend to have more zero balance members. Active status doesn’t seem to affect the shapes of the distributions.



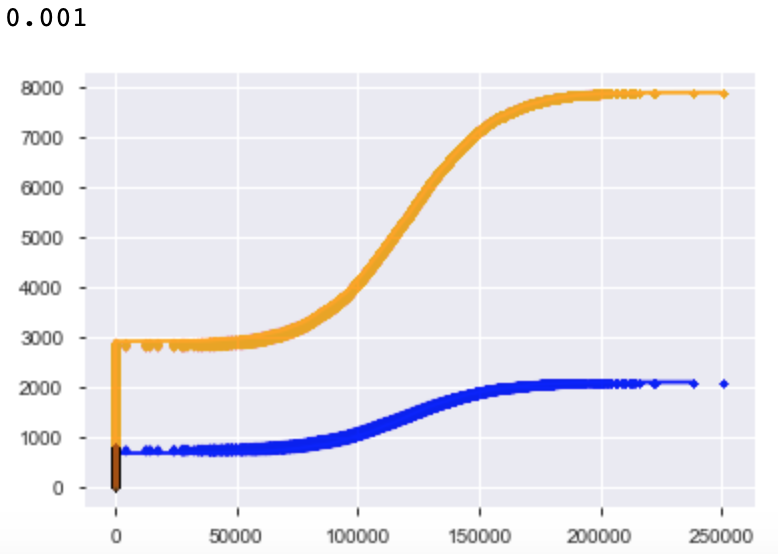
From their ECDFs we see active status doesn’t make a big difference but age seems to be a more important factor.

To test if there’s a significant difference in balance between two age groups, bootstrap and permutation tests were used.

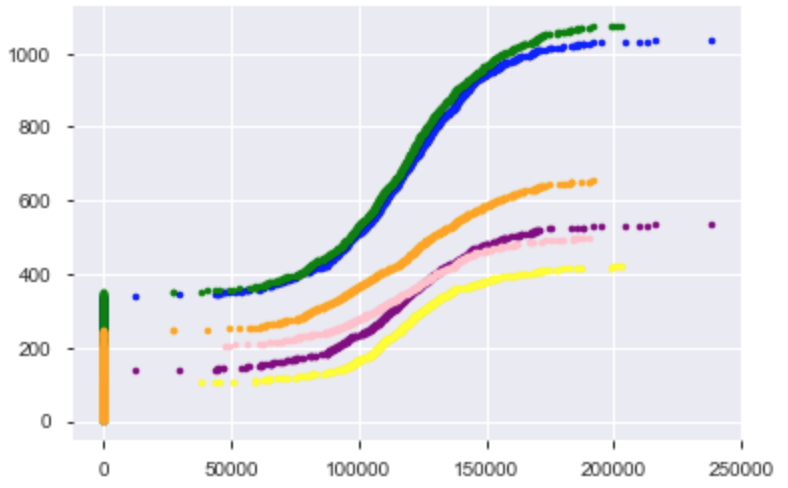
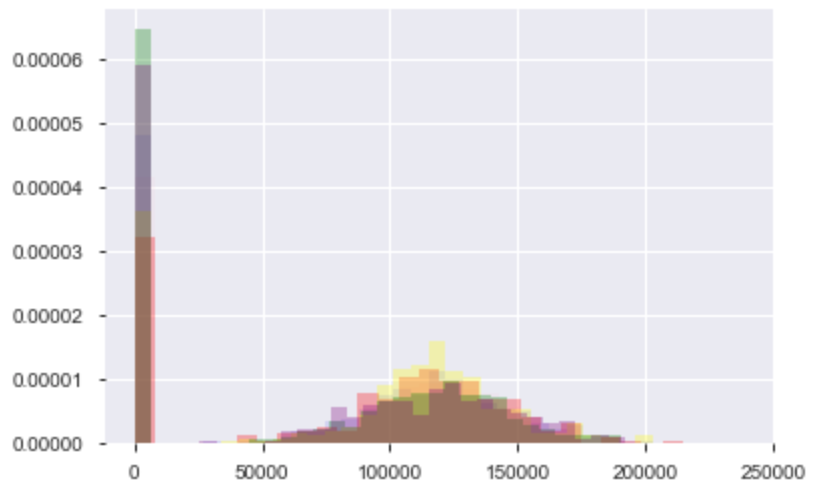
By shifting the mean of one group to the same as the other group and running bootstrap sampling (n = 10000) and comparing the simulated difference of means and empirical difference of means, a p value of 0.0008 was computed, indicating it’s not likely that the two age groups have the same mean.



Permutation (n = 1000) shows that two groups are two parts of the same distribution, a p value of 0.001 indicates that two groups are not likely to have the same mean.

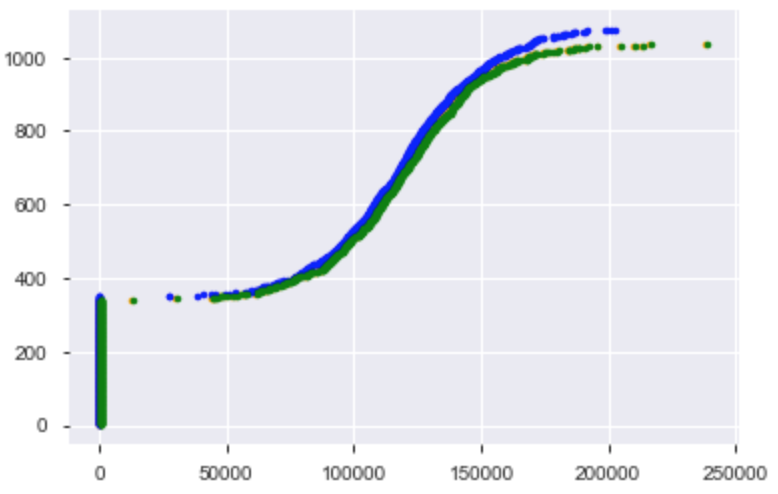


By grouping the age > 45 members by gender and exited and plotting the normalized histogram and ECDFs, we can see the other than the percentage of zero balance members, six groups seem to have similar distribution of balance, the exited groups seem to have more balance in the range mentioned previously. Gender doesn’t seem to be affecting the shapes of the distributions.

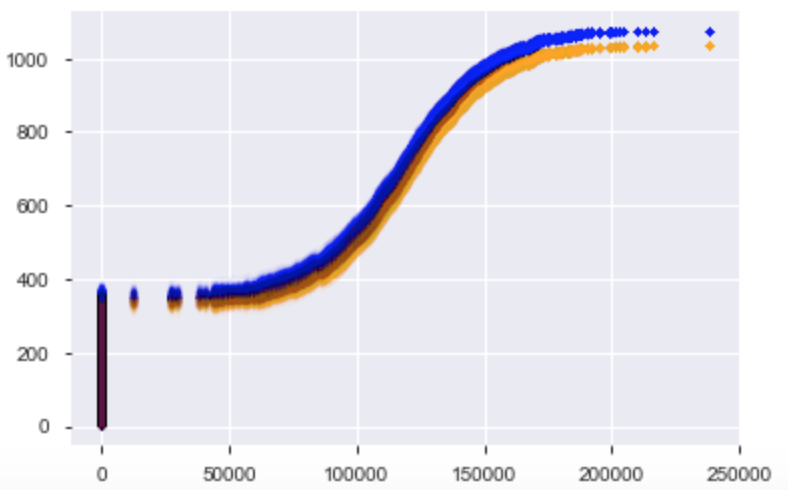


To test if age >45 female and male groups have the same mean and distribution, bootstrap and permutation tests were used.

Same as the bootstrap test before, the mean of one group was shifted to the same as the other group, and after bootstrap sampling (n = 10000), a p value of 0.3859 shows it’s possible the two groups have the same mean.

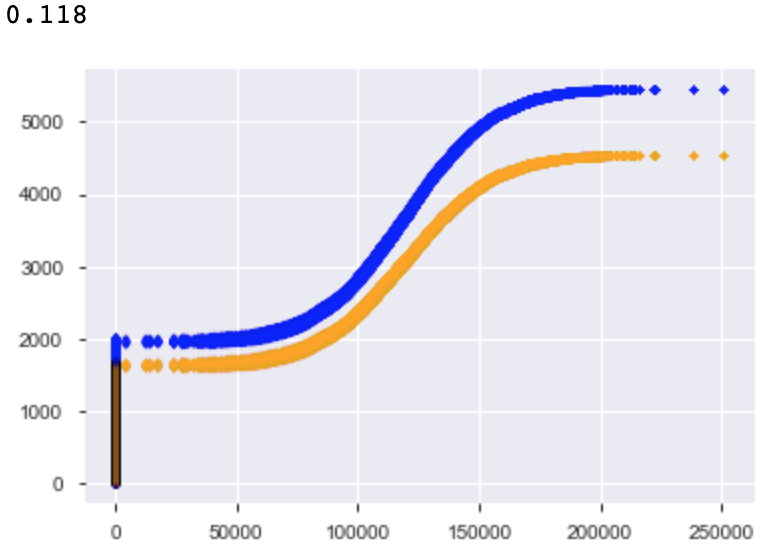
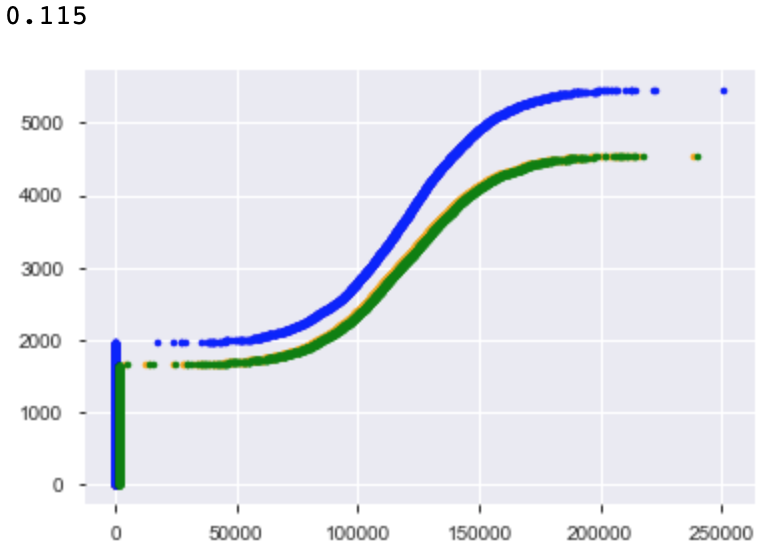


Then from the permutation (n = 1000) test we can see the two groups are from the same distribution and a p value of 0.397 shows that it’s possible they have the same means.

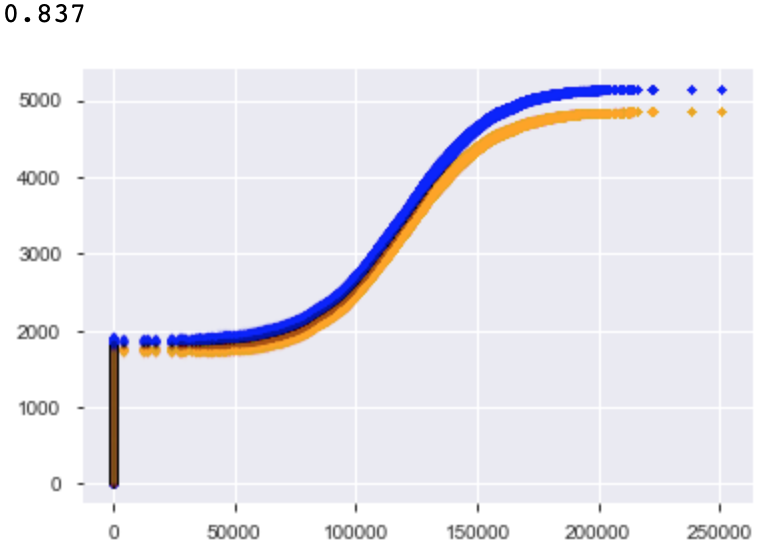
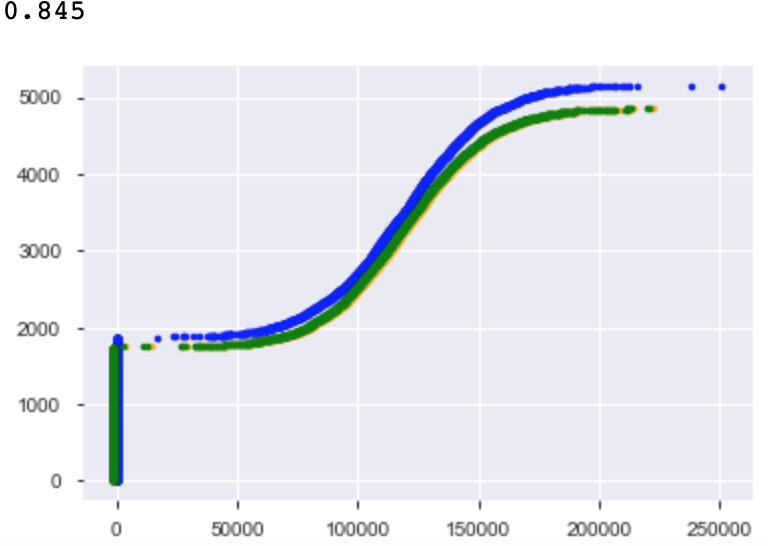


Then after performing the same tests on female and male members, no significant difference in their mean was found, and they are most likely from the same distribution.

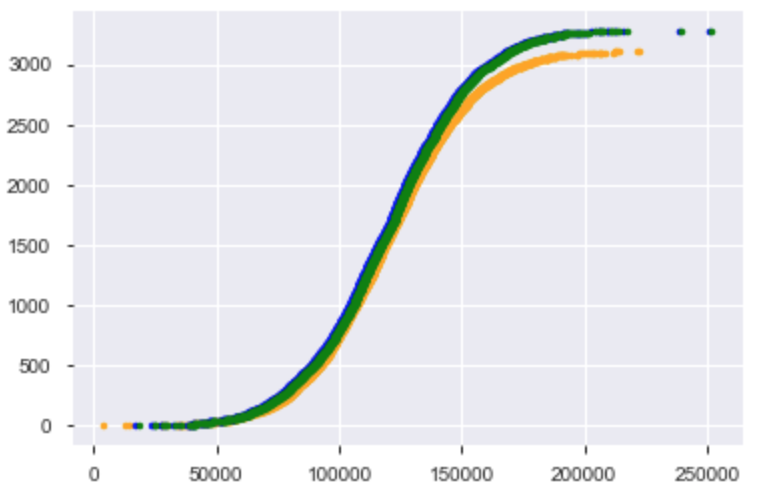
(p value for bootstrap test is 0.115, and 0.118 for permutation test)



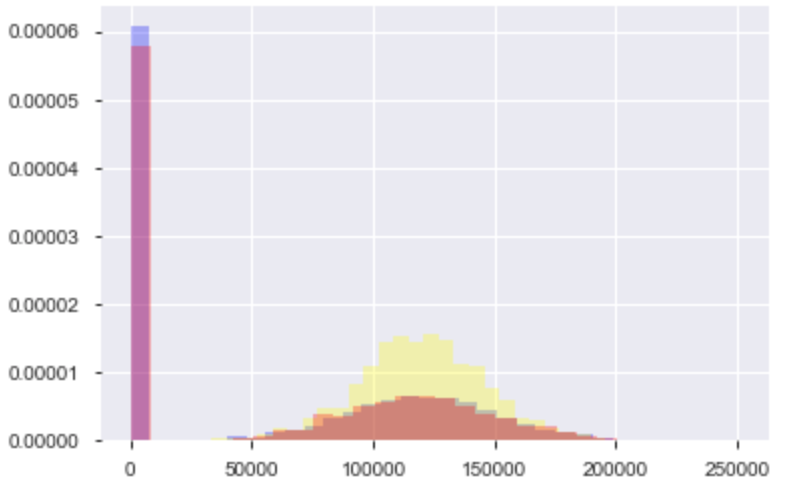
Then to test the effect of active status on balance, same bootstrap and permutation tests were performed on active and inactive groups. A p value of 0.845 from the bootstrap test shows that two groups are likely to have the same mean. And a p value of 0.837 indicates that they are likely to have the same mean and distribution.



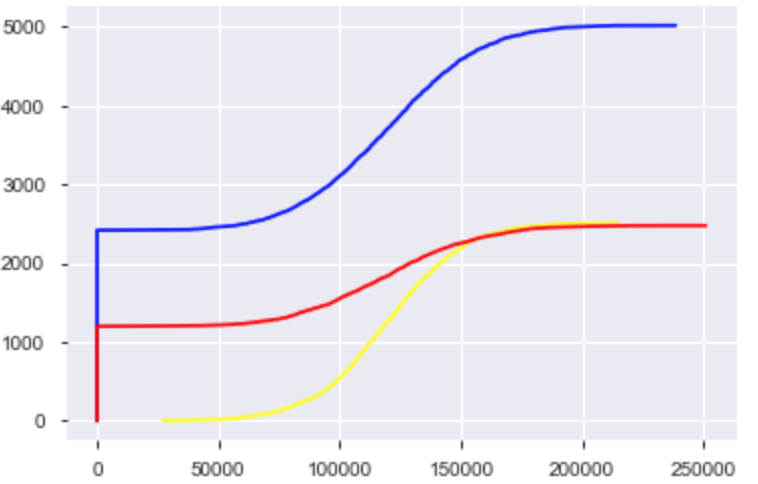
And after removing the zero balance members from the two groups and running the same bootstrap test again, a p value of 0.9492 was computed showing that the non zero balance members of two groups are very likely of the same mean.



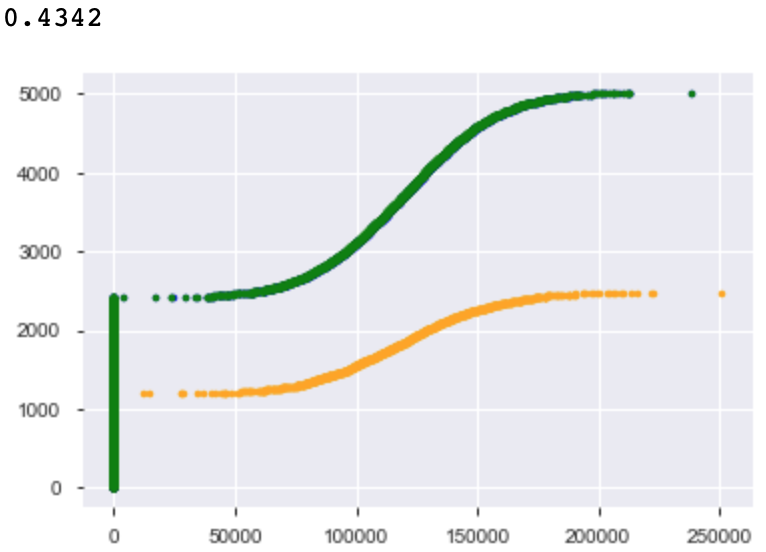
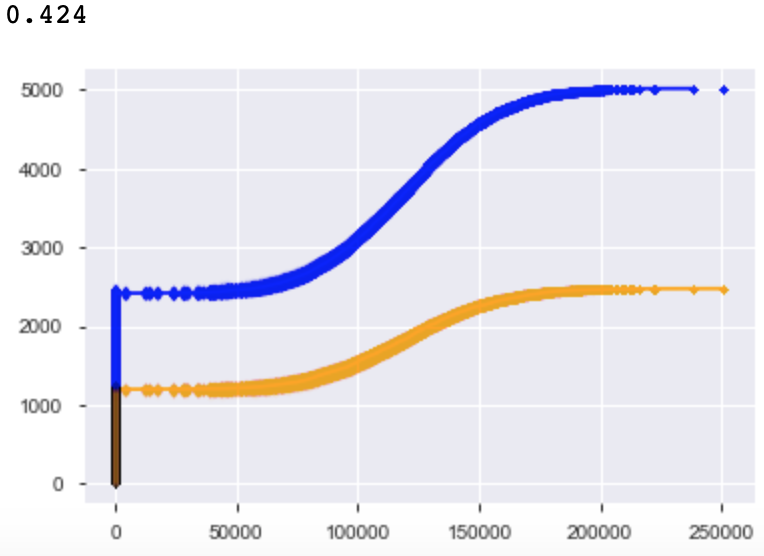
By plotting a normalized histogram of balance by country and the ECDFs, we can tell Germany has a very different distribution pattern compared to two other countries, and the width is narrower with lots of members having balance within the range we found.



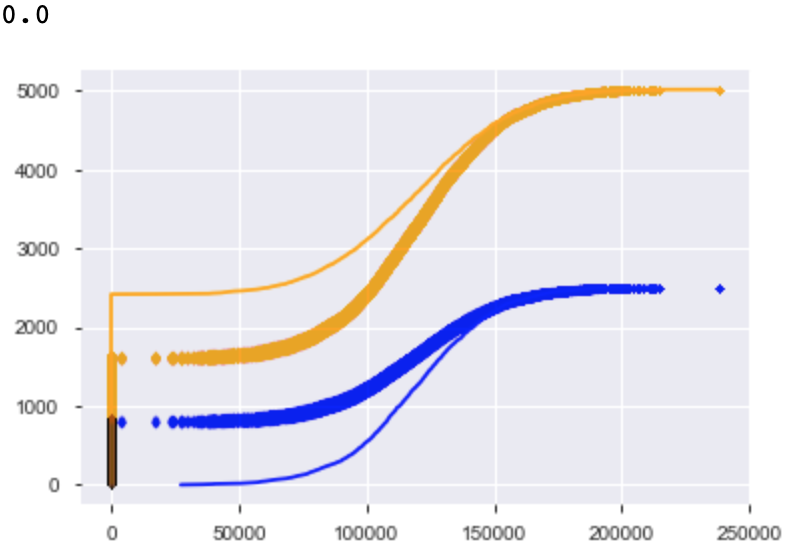
Their ECDFs show that France and Spain are a bit similar, and France’s non zero part looks similar in shape to Germany.



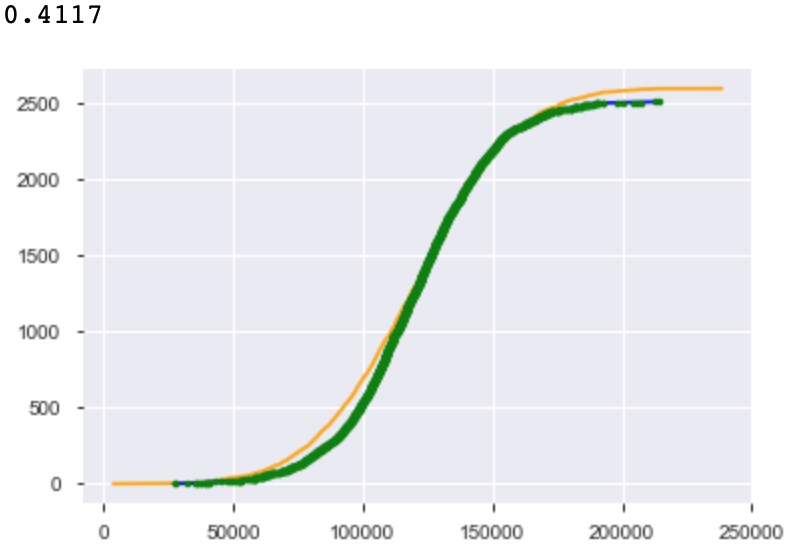
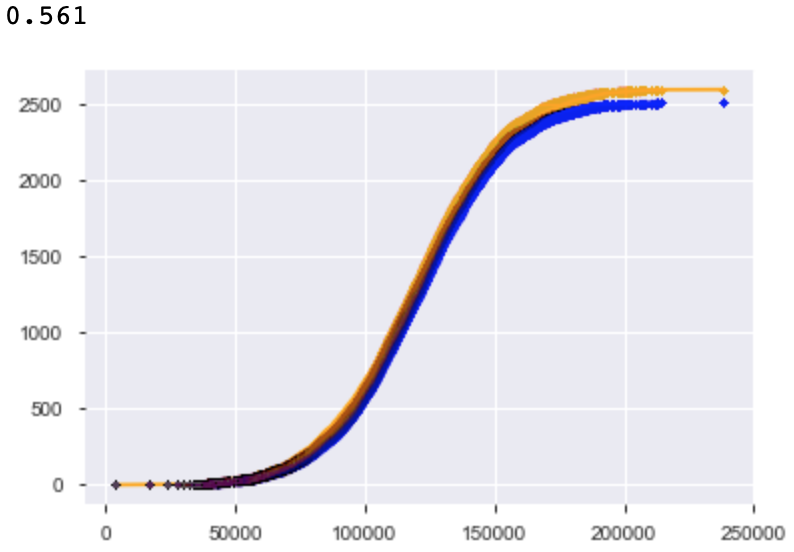
To test if France and Spain have the same mean and distribution, permutation test (n = 1000) was performed and a p value of 0.424 was generated, indicating that the two countries could have the same mean and they are likely two parts of a same distribution. A bootstrap test (n = 10000) then generated a p value of 0.4342, showing the two groups could have the same mean.



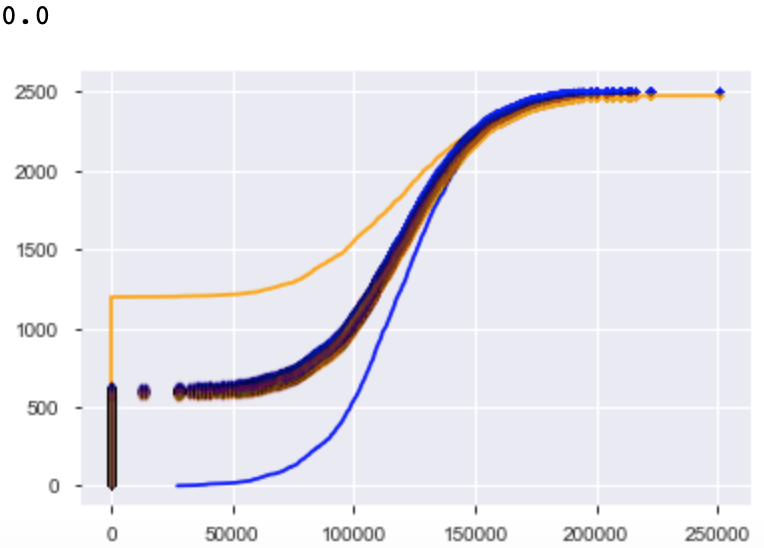
To compare France and Germany, we include the zero balance members first. And the permutation test (n = 1000) shows that the two groups are not likely to have the same mean or distribution (p value is 0.0).



Then after removing the zero balance members in France and running the same permutation test again, we get a p value of 0.561 and they seem to be from the same distribution. And the bootstrap test generated a p value of 0.4117, indicating it is possible non zero balance France and Germany have the same mean and distribution.



The permutation test on Spain and Germany shows that the two countries are not likely from the same distribution.



So in conclusion, customers of a certain range of balance tend to have a higher exit rate; zero balance customers only show different distribution when purchasing products and most customers tend to purchase 1-2 products, those who have 3-4 tend to leave; active status and gender do not significantly affect balance, age > 45 group tend to have a higher exit rate; It’s possible that France and Spain can have the same mean and come from the same distribution, and Germany and France without zero balance members can have the same mean and distribution.