PREPARED BY:
Grace Skelley

DATA MINING

To understand the data, first step is to assess the quality of the data, by checking for missing values, errors, and inconsistencies.

COLUMN NAME	DATA CONSISTENCY CHECK	COMMENTS/REASONS
Row_Number	Column deleted	Irrelevant data points
Customer_ID	Renamed Column to Customer ID	For clarity
Last Name	Column deleted	Removed in adherence to data privacy
Credit Score	3 missing values	blank fields are replaced with "Null" value, kept 1 "NULL" and then converted to "Null" for consistency.
Country	Country format inconsistency: 23 DEs, 118 ESs, 244 FRs	Replaced the abbreviated country with its equivalent full name.
Gender	Gender format inconsistency: 19 Fs & 49 Ms	Converted to full form
Age	Found 11 values with 2 years; 1 Null	Converted 11 values to Null and maintained 1 "NULL" value. Data entry may be entered by mistake however, will keep these data points and will update them later on if needed.
Tenure	No Changes made	Values and Format are consistent
Balance	No Changes made	Values and Format are consistent
NumOfProducts	Renamed Column to No. of Products	For clarity
HasCrCard?	Renamed Column to Credit Card Status Found 700 in 1s value; 291 in 0s value	Renamed column for clarity. Replaced 700 1s value with "With Credit Card" and 291 0s with "No Credit Card"
IsActiveMember	Renamed Column to Membership Status Found 503 in 1s value; 488 in 0s value	Renamed column for clarity. Replaced 503 1s value with "Active" and 488 0s with "Not active"
Estimated Salary	Found 1 NULL; 1 missing value	A blank field is replaced with a "Null" value, kept 1 "NULL" and then converted to "Null" for consistency.
ExitedFromBank?	Renamed Column to Exited Status Found 204 in 1s value; 787 in 0s value	Renamed column for clarity. Replaced 204 1s value with "Left" and 787 0s with "Stayed"

DATA SET DIMENSION:

RAW DATA: *14 col and 992 rows* | CLEAN DATA: *13 col and 992 rows*

No duplicates are found and null/missing values are kept before performing initial descriptive statistics because these data points may provide useful insights.

BASIC DESCRIPTIVE STATISTICS

Performing basic descriptive statistics to understand the data and identify the risk factors contributing to customers leaving the bank.

INITIAL DESCRIPTIVE STATISTICAL ANALYSIS BETWEEN client who STAYED vs. LEFT

STAYED	MIN	MAX	AVERAGE
Credit Score	411	850	652
Age	18	82	37
Tenure	0	10	5
Balance	$0.00	$197,041.80	$74,830.87
No. of Products	1	3	2
Estimated Salary	$371.05	$199,661.50	$98,943.39

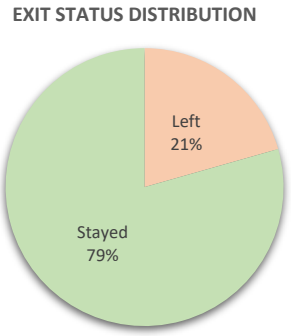
LEFT	MIN	MAX	AVERAGE
Credit Score	376	850	637
Age	22	69	45
Tenure	0	10	0
Balance	$0.00	$213,146.20	$90,239.22
No. of Products	1	4	1
Estimated Salary	$417.41	$199,725.39	$97,155.20

EXPLORING VARIABLES TO IDENTIFY FACTORS THAT LEAD TO CLIENTS LEAVING

RATIO DISTRIBUTION BY EXIT STATUS

Examining the magnitude of distribution based on exit status provides insight into the overall influence.

EXIT STATUS	TOTAL COUNT	DISTRIBUTION (%)
Left	204	21%
Stayed	787	79%
Grand Total	**991**	**100%**



AVERAGE CREDIT SCORE

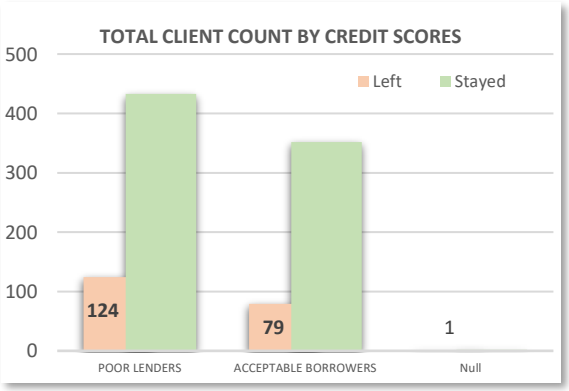
Although the total count of credit scores belonging to poor lenders is higher, analyzing the proportion uncovers intriguing findings.

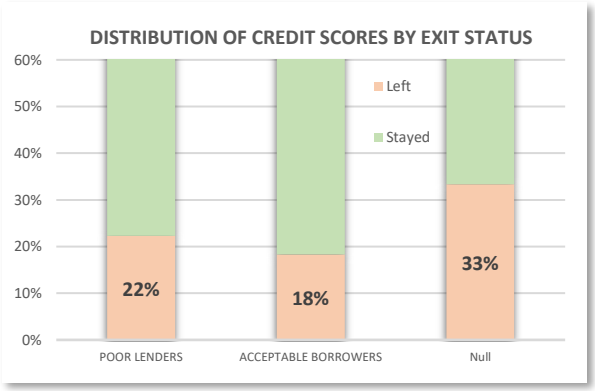
TOTAL COUNT

CREDIT SCORE RANGE	Left	Stayed	Grand Total
POOR LENDERS	124	433	557
ACCEPTABLE BORROWERS	79	352	431
Null	1	2	3
Grand Total	204	787	991

DISTRIBUTION (%)

CREDIT SCORE RANGE	Left	Stayed	Grand Total
POOR LENDERS	22%	78%	100%
ACCEPTABLE BORROWERS	18%	82%	100%
Null	33%	67%	100%
Grand Total	21%	79%	100%



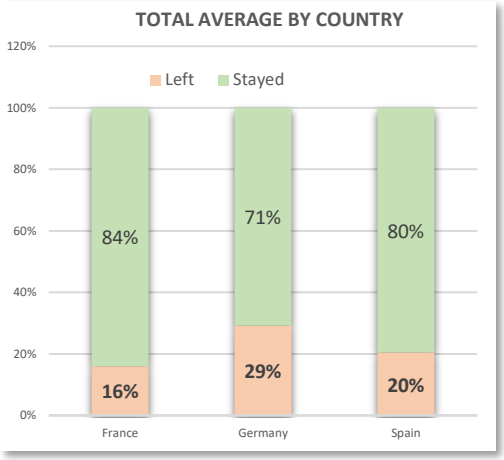


BY COUNTRY

The highest volume of withdrawn accounts is observed among clients residing in Germany, followed by Spain and France.

TOTAL AVG.

COUNTRY	Left	Stayed	Grand Total
France	16%	84%	100%
Germany	29%	71%	100%
Spain	20%	80%	100%
Grand Total	**21%**	**79%**	**100%**



BY GENDER

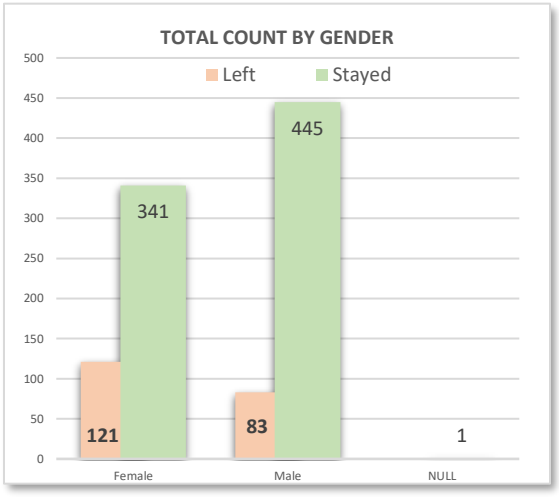
Although it may appear that females exhibit the highest exit rates within the gender division, this variable may need to be excluded as a parameter due to the existence of ***contentious biases associated with gender equality****.*

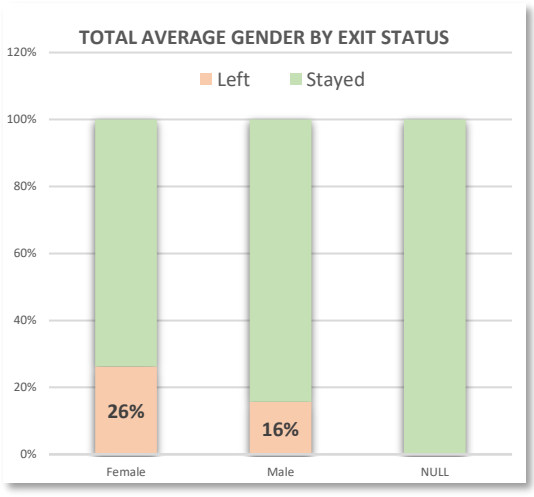
TOTAL COUNT

COUNTRY	Left	Stayed	Grand Total
Female	121	341	462
Male	83	445	528
NULL		1	1
Grand Total	**204**	**787**	**991**

TOTAL AVG.

COUNTRY	Left	Stayed	Grand Total
Female	26%	74%	100%
Male	16%	84%	100%
NULL	0%	100%	100%
Grand Total	**21%**	**79%**	**100%**

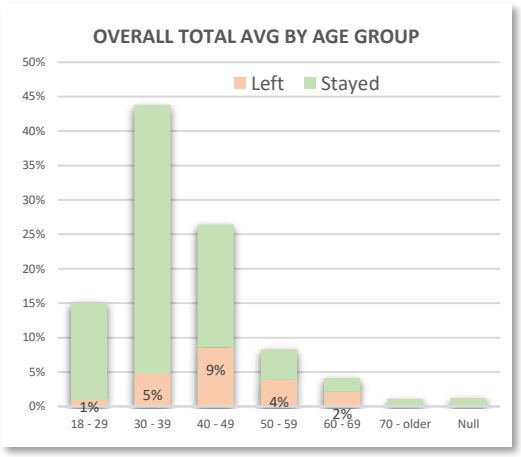




BY AGE

*The 40-49 age range demonstrates a significant risk rate among all the client age groups. Similarly, to gender bias, incorporating age as a parameter may lead to **age discrimination among clients**.*

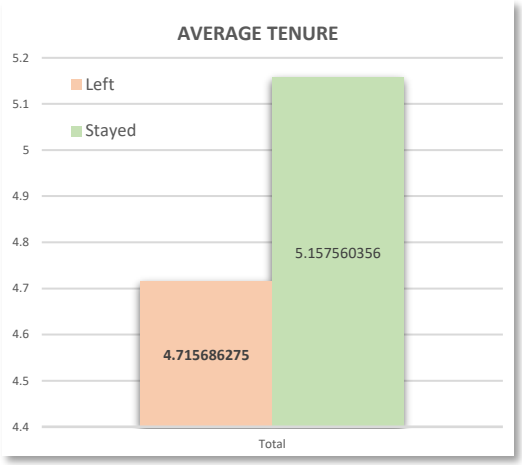
OVERALLL AVERAGE (%)			
AGE GROUP	**Left**	**Stayed**	**Grand Total**
18 - 29	1%	14%	15%
30 - 39	5%	39%	44%
40 - 49	9%	18%	26%
50 - 59	4%	4%	8%
60 - 69	2%	2%	4%
70 - older	0%	1%	1%
Null	0%	1%	1%
Grand Total	**21%**	**79%**	**100%**

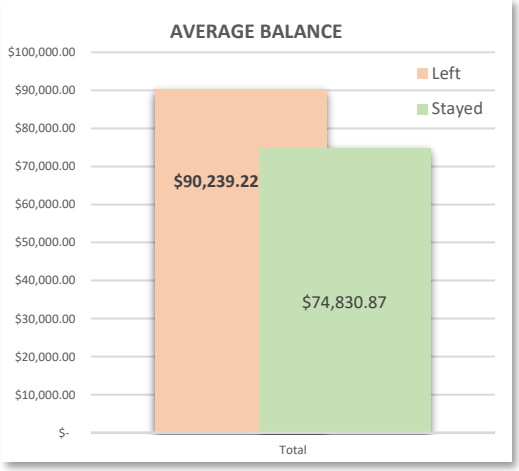


BY TENURE & BY ACCOUNT BALANCE

The average tenure between the clients that closed their accounts and existing ones is 5 years. However, based on the initial descriptive statistical analysis on the graph below, there is insufficient evidence to discern a distinct pattern. Therefore, this variable may not be a reliable measure for defining the algorithm.

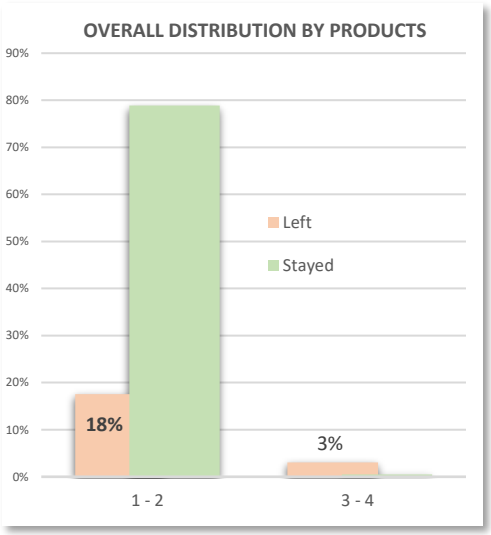
Clients who maintain an average balance of $90,000.00 are more prone to discontinuing the services, while those with lower balances tend to keep their bank accounts open. To comprehend the underlying reasons for this trend, further supporting information is necessary for this variable.



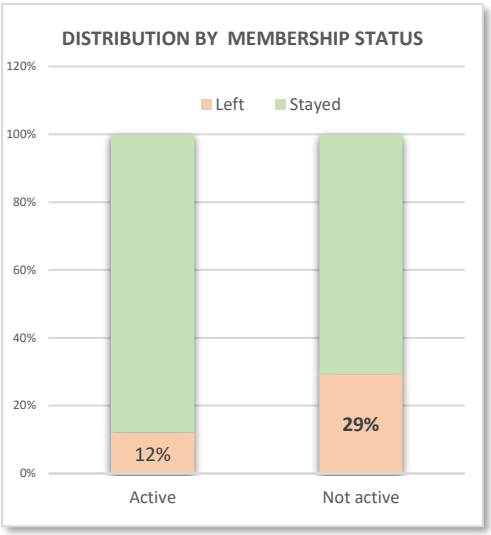


BY PRODUCTS & BY MEMBERSHIP STATUS

Clients who have only 1 or 2 products in their records exhibit a high risk of exiting, while those with more products are more likely to retain their accounts.

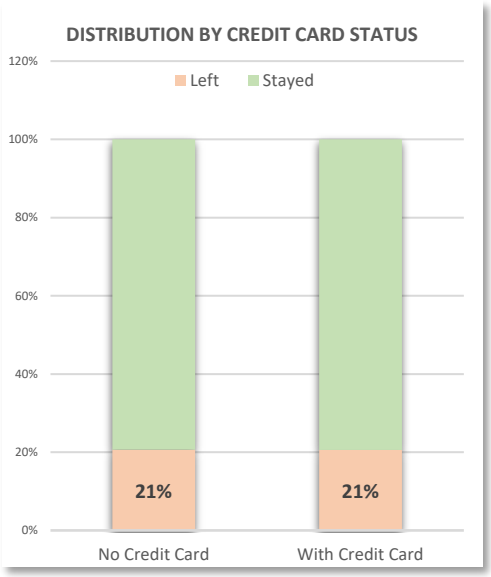


The percentage of inactive members who have left, which stands at 29.30%, is significantly higher compared to the proportion of active individuals who have departed.

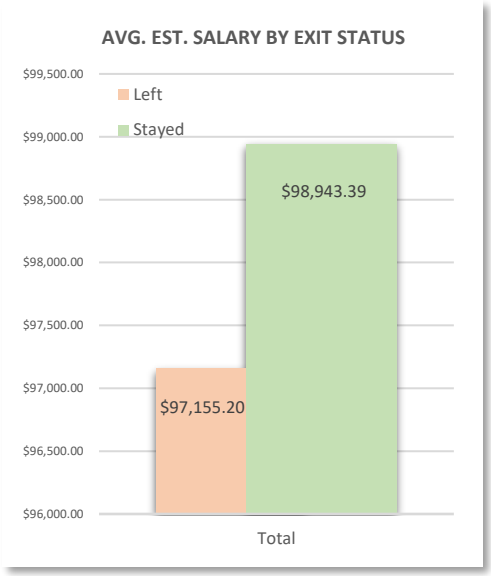


BY CREDIT CARD STATUS & BY ESTIMATED SALARY

There is no substantial distinction between the credit card statuses, indicating that this variable will not be taken into account during the modeling phase.



The average estimated salary for both groups fall within a similar range, making it impractical to consider this variable during the modeling stage.



CONCLUSION & RECOMMENDATIONS

Basing on the current findings, certain factors increase the likelihood of customer churn, such as having clients in Germany and inactive members with 1-2 products. It is advisable to avoid including variables like age and gender to prevent any potential discrimination. Instead, focusing on analyzing the German market and gaining a better understanding of customer preferences can be beneficial in effectively promoting product lines that cater to their interests. By doing so, it is possible to mitigate the decrease in account closures specifically in this region.

Moreover, it is advisable to conduct an investigation into customer behavior trends in order to obtain valuable insights regarding specific demands and prevent customer attrition. This endeavor will yield useful information for evaluating the performance of products and services, leading to the development of innovative solutions and generating greater interest.