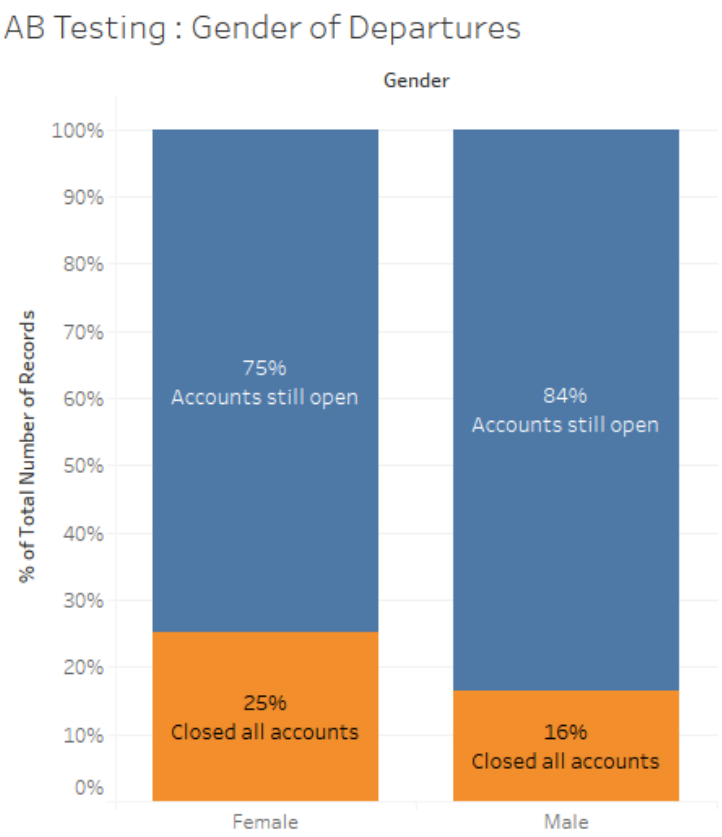
The executives at Sample Bank International are losing customers at an alarming rate. We are going to investigate a 10,000-row, 14-column .xls file (Churn Modeling.xls) to determine possible reasons for the customer departure.

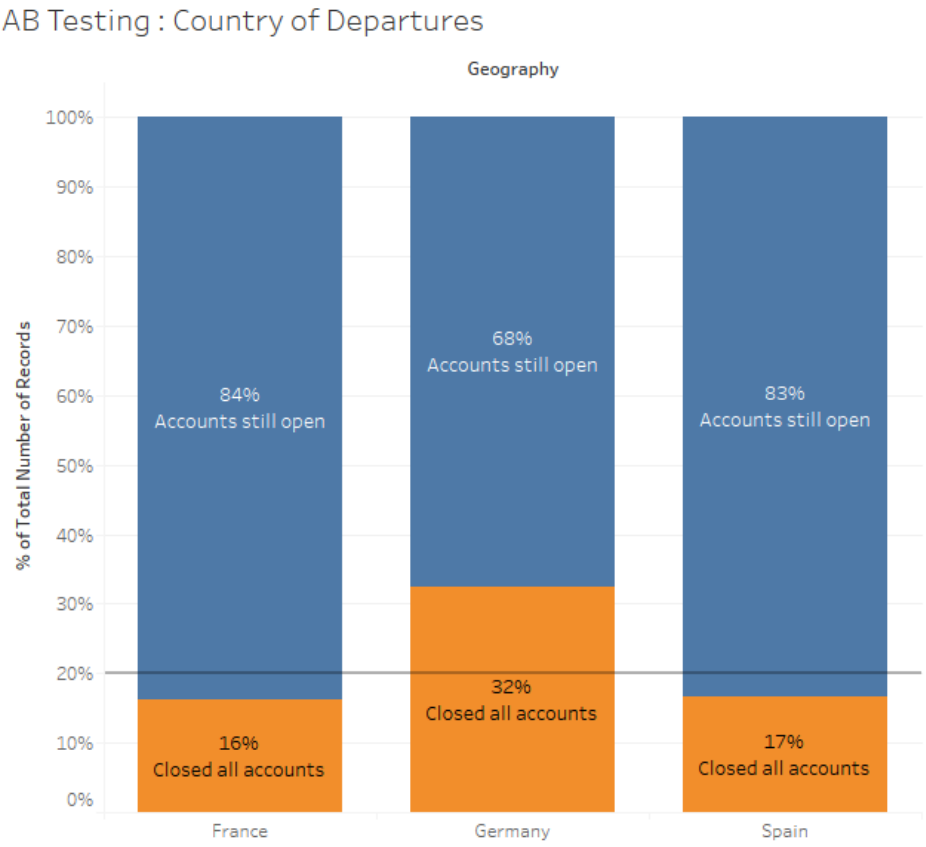
We first practiced with Tableau to prepare an ad-hoc AB test of those who departed and their gender.

In this figure, we can see that there is a higher percentage of females who left the Bank during our study period, but should be tested for statistical significance to determine any substantial conclusions.

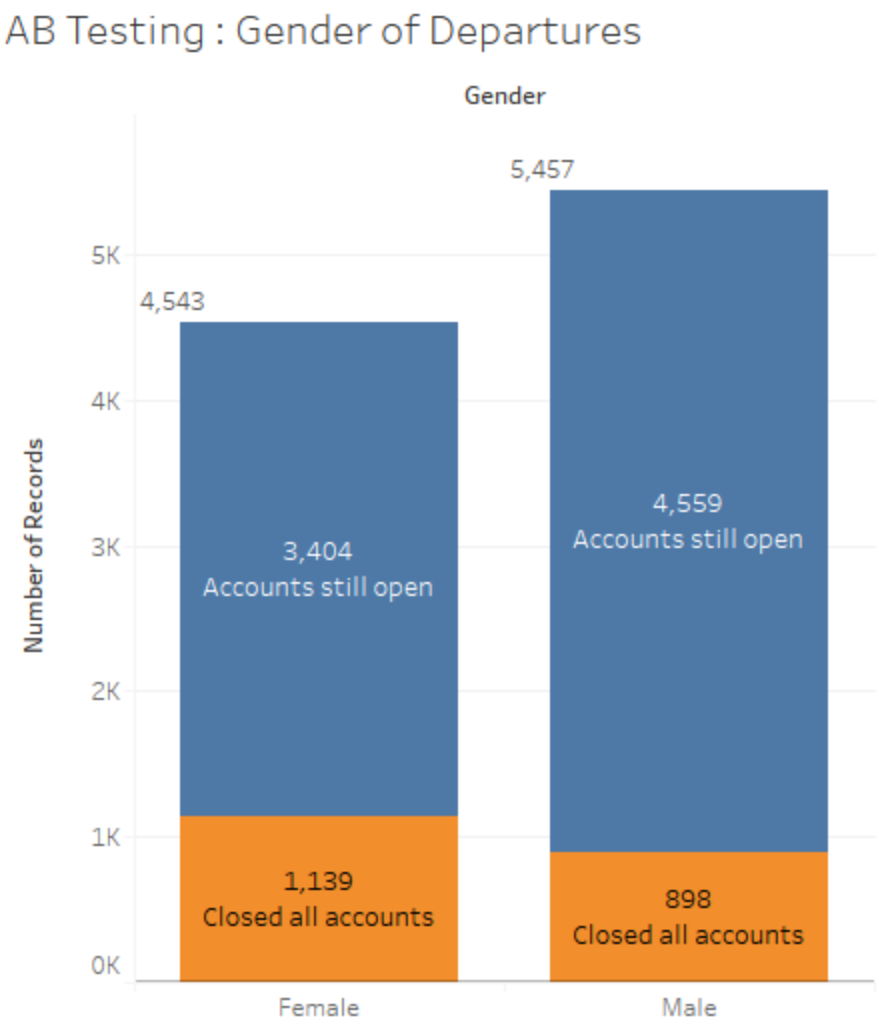


We also tested the country of departures and noticed a rapid rate of departure from Germany, which may be statistically significant. Note this is no longer a pseudo-AB test, because of the number of variables.

A reference line was added at 20% which represents the overall rate of departure.



Running an AB test with chi-squared, our result shows that “Sample 1 is more successful” which means that women indeed have a higher rate of leaving the bank, and is statistically significant with a p-value of .001. Using the exit values of 1139 out of 4543 for women, and 898 out of 5457 for men, we are able to calculate our findings.



Question: Does the rate of success differ across two groups?

	# successes		# trials	Confidence interval	[clear] [link]
Sample 1:	1139	/	4543	23.8% – 26.4%	
Sample 2:	898	/	5457	15.5% – 17.5%	

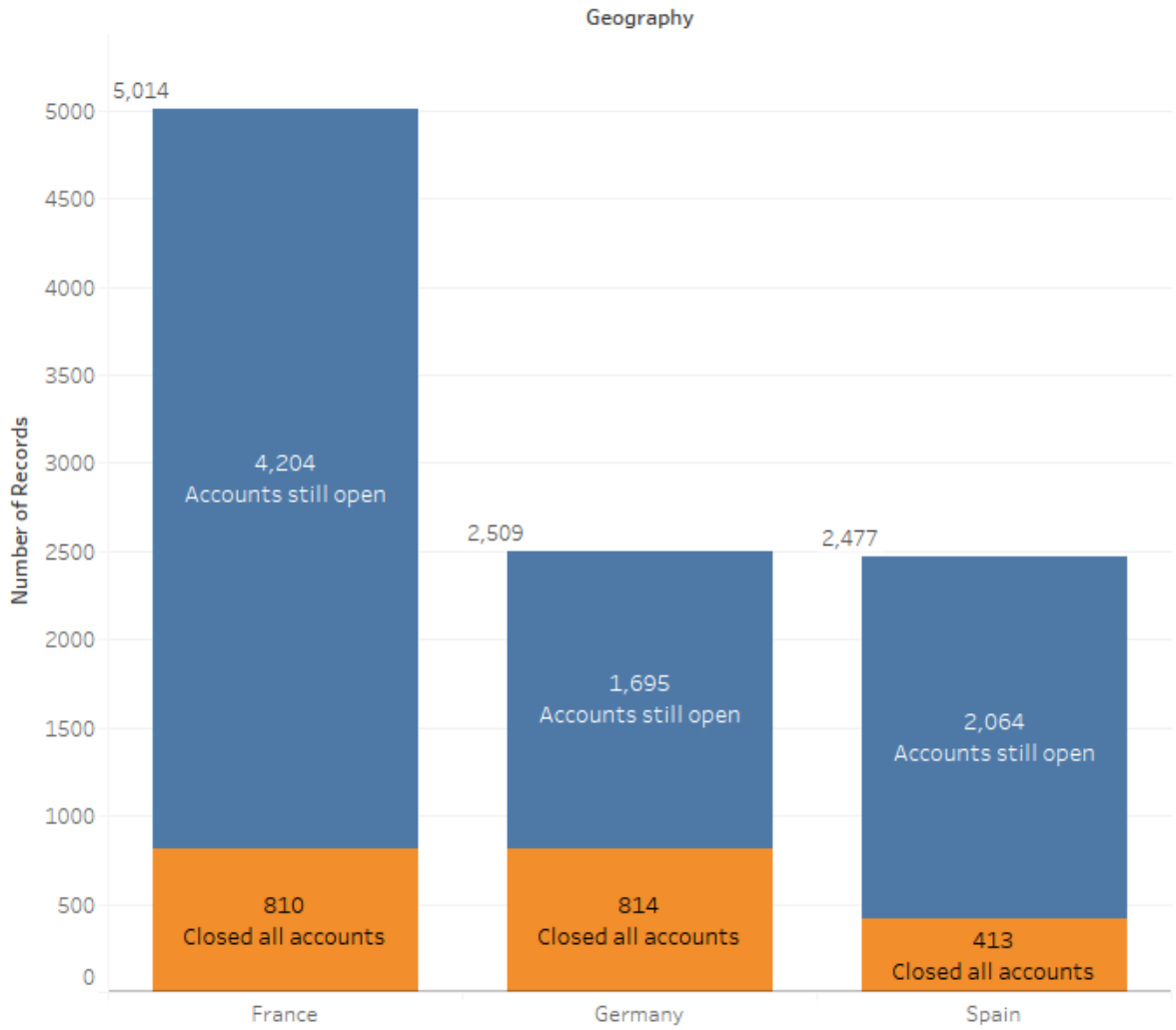
Verdict:

Sample 1 is more successful

($p < 0.001$)

Confidence level: 95%

Country Actuals



Running a similar test, a contingency test, on the geographical significance provides there exists a statistical significance across all samples, and that Germany's 34% departure rate is also significant. Further details are provided below.

Data Entry

	B_1	B_2	B_3	B_4	B_5	Totals
A_1	4204	810	-----	-----	-----	5014
A_2	1695	814	-----	-----	-----	2509
A_3	2064	413	-----	-----	-----	2477
A_4	-----	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----	-----
Totals	7963	2037	-----	-----	-----	10000

Reset Calculate

Chi-Square	df	P
301.26	2	<.0001

Cramer's V = 0.1736

No message for this analysis.

Percentage deviation and *standardized residual* are both measures of the degree to which an observed chi-square cell frequency differs from the value that would be expected on the basis of the null hypothesis.

For each cell, *percentage deviation* is calculated as

$$\frac{\text{observed} - \text{expected}}{\text{expected}} \times 100$$

Thus, a percentage deviation of +15% within a cell indicates that the observed frequency is 15% greater than the expected, while a percentage deviation of -15% indicates that the observed frequency is 15% smaller than the expected.

In the special case of df=1, the calculation of percentage deviation incorporates a correction for continuity:

$$\frac{|\text{observed} - \text{expected}| - 0.5}{\text{expected}} \times 100$$

The resulting value is then given a positive sign if *observed>expected* and a negative sign if *observed<expected.*

The *standardized residual* for a cell in a chi-square table is a version of the standard normal deviate, *z*, calculated as

$$z = \frac{\text{observed} - \text{expected}}{\text{sqrt[expected]}}$$

In the special case of df=1, the calculation of the standardized residual incorporates a correction for continuity:

$$z = \frac{|\text{observed} - \text{expected}| - 0.5}{\text{sqrt[expected]}}$$

The resulting value of *z* is then given a positive sign if *observed>expected* and a negative sign if *observed<expected.*

The chi-square value that results from a chi-square analysis is equal to the sum of the squares of the standardized residuals.

Assuming the null hypothesis to be true, and providing that the expected value for a cell is at least 5, values of the standardized residual belong to a normally distributed sampling distribution with a mean of zero and a standard deviation of ±1.0.

Percentage Deviations

	B_1	B_2	B_3	B_4	B_5
A_1	+5.3%	-20.7%			
A_2	-15.2%	+59.3%			
A_3	+4.6%	-18.1%			
A_4					
A_5					

Standardized Residuals

	B_1	B_2	B_3	B_4	B_5
A_1	+3.34	-6.61	-----	-----	-----
A_2	-6.78	+13.4	-----	-----	-----
A_3	+2.06	-4.08	-----	-----	-----
A_4	-----	-----	-----	-----	-----
A_5	-----	-----	-----	-----	-----

Lambda for predicting		Standard Error	.95 CI Limits	
			Lower	Upper
A from B:	0.0008	0.0163	0	0.0328
B from A:	0			

[Click here for a brief explanation of lambda.]

Estimated Probability of Correct Prediction when Predicting:

A without knowledge of B	0.5014
A from B	0.5018
B without knowledge of A	0.7963
B from A	0.7963