Task (Data mining using Tableau)

A bank operating over three countries in Europe has encountered a problem where a lot of customers have been leaving the bank lately. The bank took a random sample of 10,000 customers to represent the population of its customers and kept tracking of who is leaving the bank over six months and gave you the following data to perform data mining and deliver some insights that can help the bank solve the problem.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	RowNumb	CustomerI	Surname	CreditScor	Geography	Gender	Age	Tenure	Balance	NumOfPro	HasCrCar	IsActiveMe	Estimated	Exited
2	1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.9	1
3	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.6	0
4	3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.6	1
5	4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
6	5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	1	1	79084.1	0
7	6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	1	0	149756.7	1
8	7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
9	8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	1	0	119346.9	1
10	9	15792365	He	501	France	Male	44	4	142051.1	2	0	1	74940.5	0
11	10	15592389	H?	684	France	Male	27	2	134603.9	1	1	1	71725.73	0
12	11	15767821	Bearce	528	France	Male	31	6	102016.7	2	0	0	80181.12	0
13	12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0
14	13	15632264	Kay	476	France	Female	34	10	0	2	1	0	26260.98	0
15	14	15691483	Chin	549	France	Female	25	5	0	2	0	0	190857.8	0
16	15	15600882	Scott	635	Spain	Female	35	7	0	2	1	1	65951.65	0
17	16	15643966	Goforth	616	Germany	Male	45	3	143129.4	2	0	1	64327.26	0
18	17	15737452	Romeo	653	Germany	Male	58	1	132602.9	1	1	0	5097.67	1
19	18	15788218	Henderson	549	Spain	Female	24	9	0	2	1	1	14406.41	0
20	19	15661507	Muldrow	587	Spain	Male	45	6	0	1	0	0	158684.8	0
21	20	15568982	Hao	726	France	Female	24	6	0	2	1	1	54724.03	0
22	21	15577657	McDonald	732	France	Male	41	8	0	2	1	1	170886.2	0
23	22	15597945	Dellucci	636	Spain	Female	32	8	0	2	1	0	138555.5	0
24	23	15699309	Gerasimov	510	Spain	Female	38	4	0	1	1	0	118913.5	1
25	24	15725737	Mosman	669	France	Male	46	3	0	2	0	1	8487.75	0
26	25	15625047	Yen	846	France	Female	38	5	0	1	1	1	187616.2	0

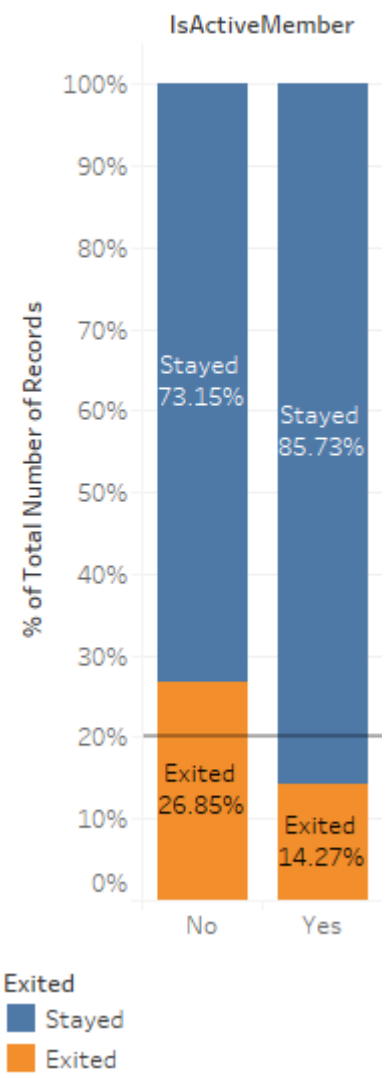
This is a snapshot of the data delivered to you and the full CSV file can be found here:

http://www.superdatascience.com/wp-content/uploads/2015/08/Churn-Modelling-Test-Data.csv

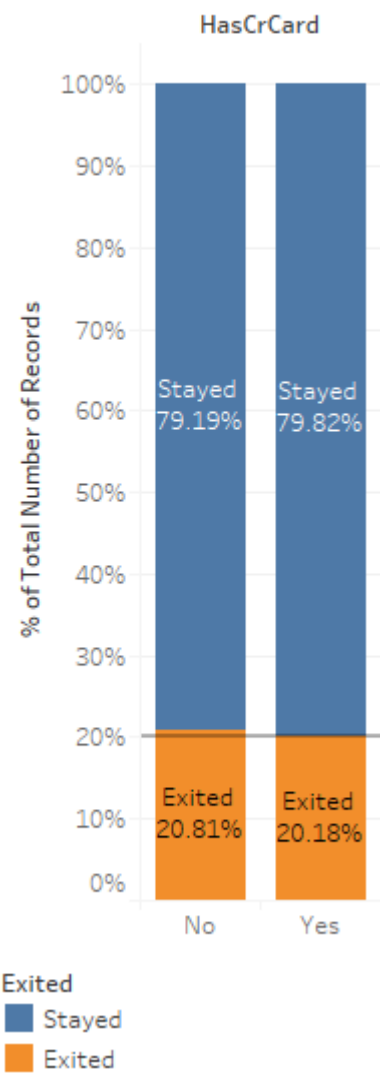
Analysis (Solution)

Ad-Hoc A-B Tests:

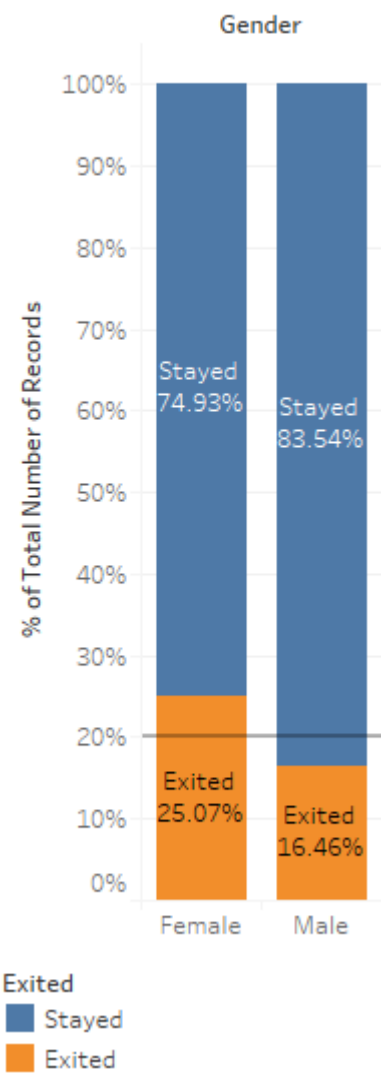
<Active (Last 6 months)>



<Has Credit Card>



Gender



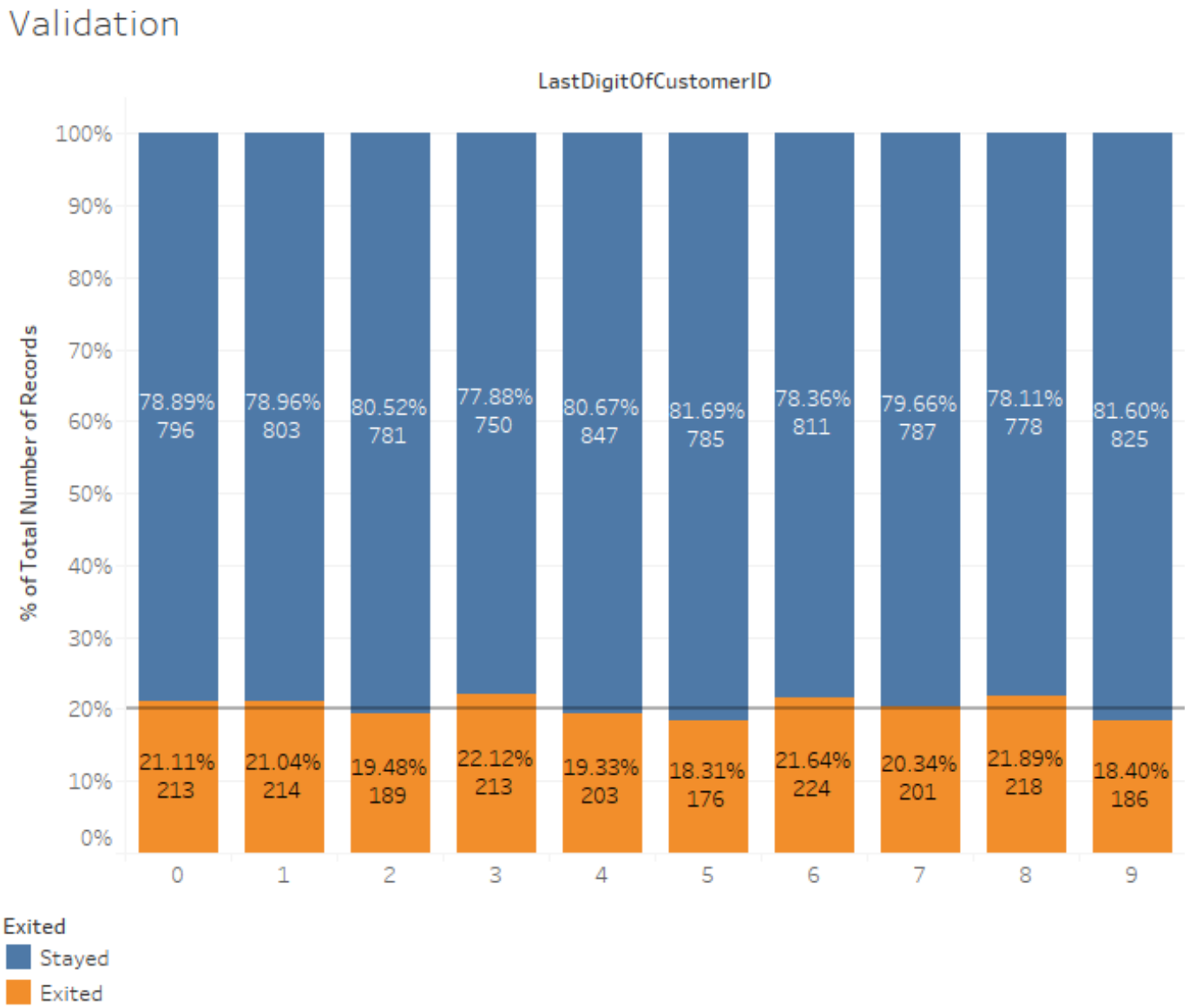
These simple ad-hoc tests can help us know some initial insights as follows:

- Women are more likely to leave the bank than men so the bank has to investigate more about the reason and try to solve it.
- Whether the customer has a credit card or not does not provide any statistical significance in the analysis.

- The member who is not active in the last six months is more likely to leave the bank than the one who has been active.

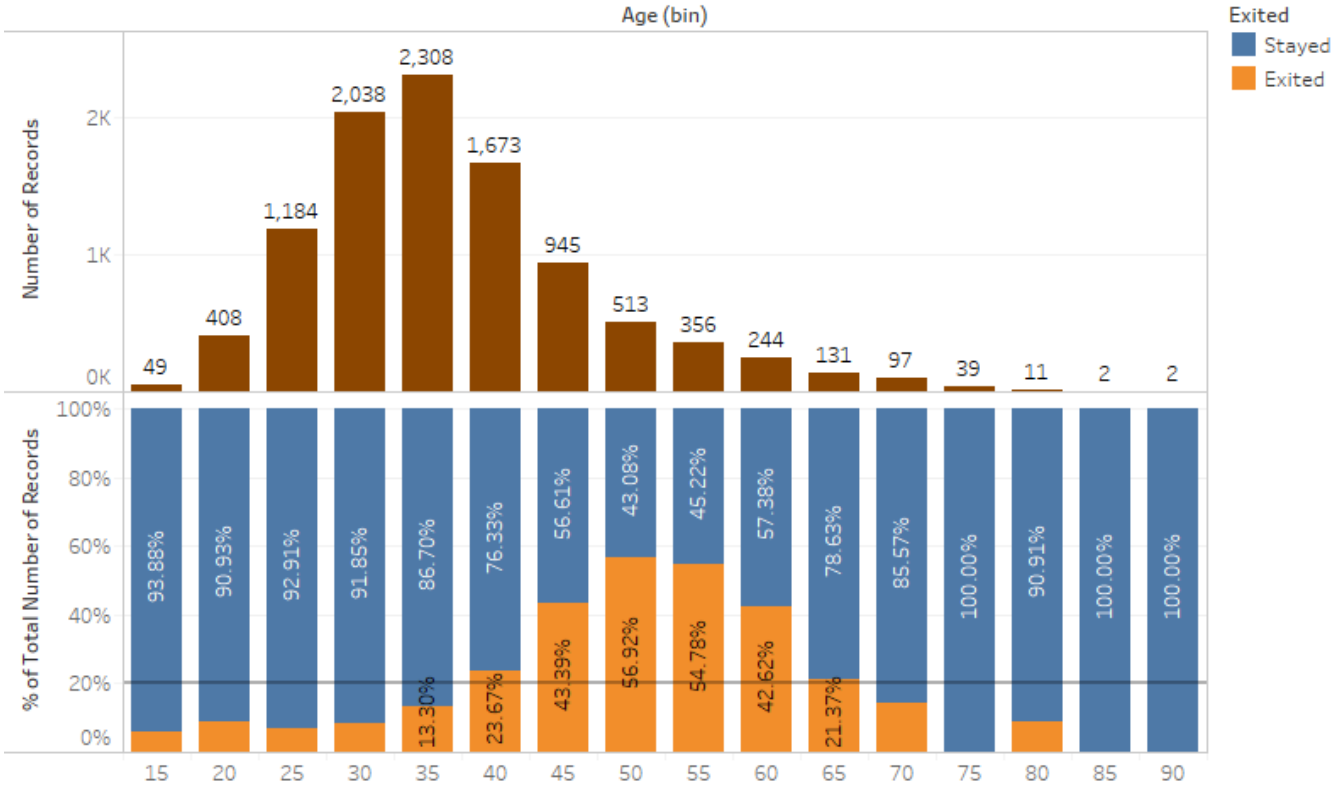
Validating approach and data:

An approach to validate our data is to take the last digit of the customer ID and find the percentages of people leaving the bank and those who don't for each number we get. This is much better than taking the last letter of the customer name because letters are less uniformly distributed. The test shows uniform distribution for all digits which is close to the average number of people leaving the bank in the whole population (20%).



More analysis:

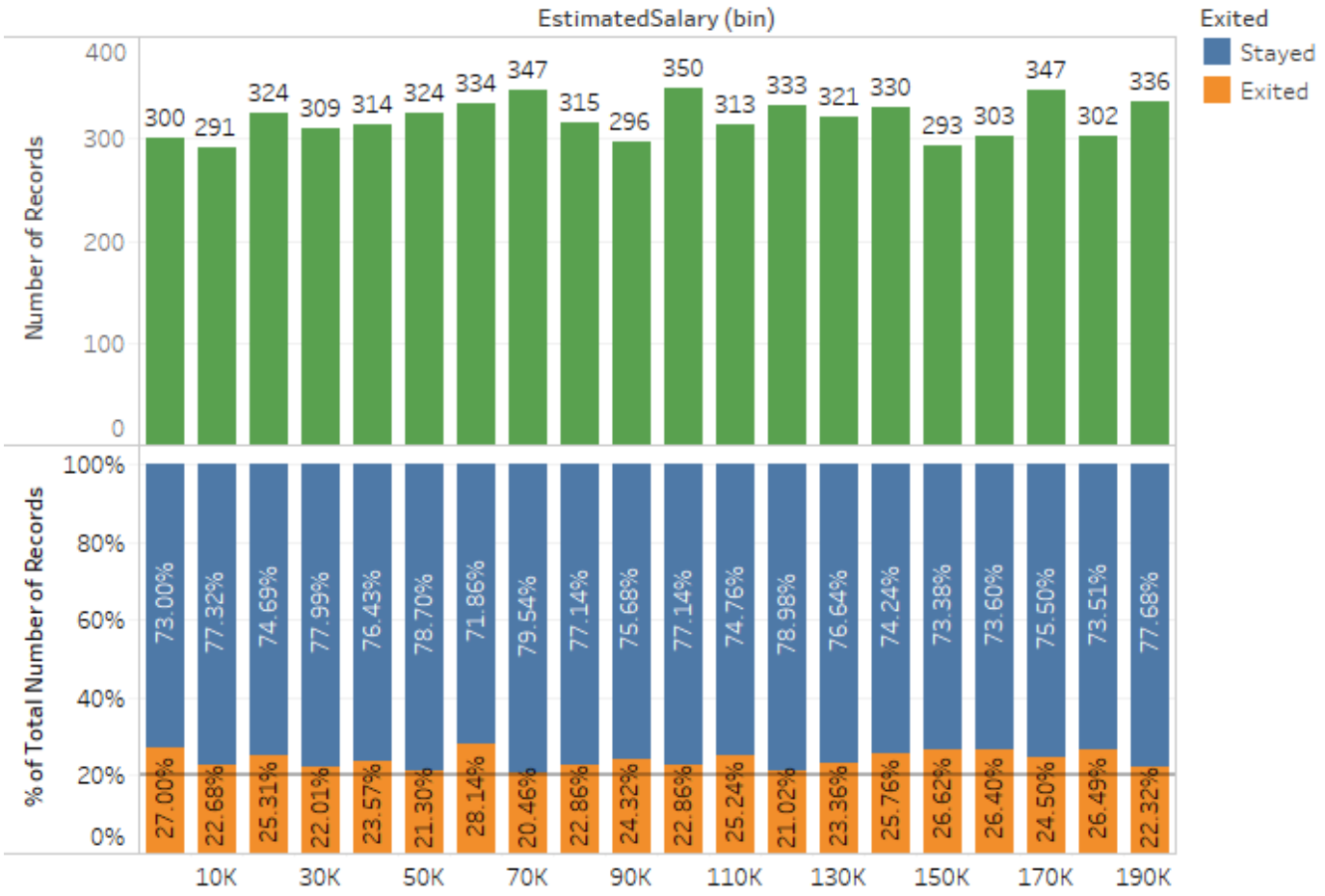
<Age Classification>



Age classification test shows some significant insights that the bank has to investigate. It shows that people between the age of 45 and 60 are most likely to leave the bank for some reason. Thus, the bank has to do further investigations on people in this age range so as to maintain its customers.

It can also be noticed that the age distribution is right skewed. Also customers are mostly between 30 and 45 years old. Furthermore, there are some anomalies for ages more than 70 because the population is very low.

<Estimated Salary Classification>



Using the estimated salary of each customer by the bank as a classifier, it can be noticed that it does not have statistical significance on the number of people exiting the bank.

Another insight that can be delivered is that the bank estimation of each customer's salary appears to be inaccurate or outdated since it has very little variations (nearly uniform) and it does not have any correlation with the customers balance which shows normal distribution.