Bank Churn Prediction

AIML course - Neural Networks

Azin Faghihi
Role: Data Scientist
March 2025

Contents / Agenda

Executive Summary

- Our modeling parameter, **Exited** is an **imbalance class**. Hence, we will try oversampling and undersampling to address this issue.
- By conducting ***exploratory data analysis*** (univariate and bivariate) we explore relationships between all the variables (categorical & numerical)
    - We conclude that Exited is affected by Age, NumOfProducts, Balance, ...
    - Due to the nature of the problem at hand, we chose **Recall** as the score to maximize.
- We try different combinations of Neural Network layers, regularization, optimizers to select a model that does not overfit and also has the highest Recall score.

Business Problem Overview and Solution Approach

- ***Problem Overview:*** Banks must address customer churn, where customers switch to competitors. Identifying key service factors influencing this decision helps management focus on targeted improvements.
- ***Objective:*** Based on the gathered data, we aim to build a neural network based classifier that can determine whether a customer will leave the bank or not in the next 6 months.
- ***Solution Approach:*** We divide the data into three partitions: train, validation, and test. Various models are trained using Neural Networks on the train set, and their performance is evaluated by comparing recall scores between the train and validation sets. The model with the highest recall and minimal overfitting is then selected.

Data

- There are 10,000 (~10K) rows and 13 columns in the dataset.
- The ***memory usage*** is approximately 1015.8 KB.
- There are ***no missing values*** in the data.
- There are ***no duplicated rows*** in the data.

Memory Usage	1015.8 KB

#	
Rows	10000
Columns	13
Null Values	0
Duplicated Rows	0

Data Dictionary

Column	Data Type	Description	# unique
CustomerId	int64	Unique ID which is assigned to each customer	10000
Surname	object	Last name of the customer	2932
CreditScore	int64	It defines the credit history of the customer	460
Geography	object	Customer's location	3
Gender	object	It defines the gender of the customer	2
Age	int64	Age of the customer	70
Tenure	int64	Number of years for which the customer has been with the bank	11
Balance	float64	Account balance	6382
NumOfProducts	int64	Refers to the number of products that a customer has purchased through the bank	4
HasCrCard	int64	It is a categorical variable which decides whether the customer has credit card or not	2
IsActiveMember	int64	Is is a categorical variable which decides whether the customer is active member of the bank or not	2
EstimatedSalary	float64	Estimated salary	9999
Exited	int64	Whether or not the customer left the bank within six month, 0 = No, 1 = Yes	2

Data

- The following tables show the summary information of our variables.
  - ***Categorical***: unique counts, most common value and its corresponding frequency
  - ***Numerical***: mean, median, standard deviation, minimum, maximum, outlier counts, ...

Object/Categorical Column	unique	top	freq
Geography	3	France	5014
Gender	2	Male	5457
NumOfProducts	4	1	5084
HasCrCard	2	1	7055
IsActiveMember	2	1	5151
Exited	2	0	7963

Numerical Column	mean	std	min	25%	50%	75%	max	IQR	# Outliers (Upper)	# Outliers (Lower)	# Outliers	Outliers %
CreditScore	650.5	96.7	350.00000	584.00000	652.000000	718.000000	850.00000	134.000000	0	16	16	0.2
Age	38.9	10.5	18.00000	32.00000	37.000000	44.000000	92.00000	12.000000	411	0	411	4.1
Tenure	5.0	2.9	0.00000	3.00000	5.000000	7.000000	10.00000	4.000000	0	0	0	0.0
Balance (K)	76.5	62.4	0.00000	0.00000	97.198540	127.644240	250.89809	127.644240	0	0	0	0.0
EstimatedSalary (K)	100.1	57.5	0.01158	51.00211	100.193915	149.388247	199.99248	98.386137	0	0	0	0.0

EDA Results
Categorical

- About 80% of the customers have not left the bank within the past 6 months. (**Exited**)
- More than half (54.6%) of the customers are male. (**Gender**)
- Half of the customers live in France. (**Geography**)
- About 70% of the customers have credit cards. (**HasCrCard**)
- More than half (~52%) of the customers are active members of the bank. (**IsActiveMember**)
- About half (~51%) of the customers have 1 product through the bank followed by 46% of them who have 2 products. (**NumOfProducts**)

EDA Results
Numerical

- The customer age is positively skewed with mean of 39 years old and median of 37 years old. (**Age**)
- The non-zero balance is almost normally distributed. The median balance is ~ $98K. (**Balance**)
- Credit score is slightly negatively skewed with median of 652 and it is likely close to normal distribution. (**CreditScore**)
- Estimated Salary is slightly positively skewed with average of $100 K. (**EstimatedSalary**)
- The average number of years the customers have been with the bank is about 5 years. (**Tenure**)

EDA Results

Correlations and Effects - Categorical vs Categorical

- Number of products a customer have could have an effect on their decision to leave the bank. (**NumOfProducts, Exited**)
- Customer's location could have an effect on their decision to leave the bank. (**Geography, Exited**)
- Whether a customer is active or not could have an effect on their decision to leave the bank. (**IsActiveMember, Exited**)
- The gender of the customer could have an effect on their decision to leave the bank. (**Gender, Exited**)
- It seems like the customer's location, gender, and being an active member affect the number of products. (**Geography/Gender/IsActiveMember, NumOfProducts**)
- It seems like how active customers are could vary with their gender. (**Gender, IsActiveMember**).
- It seems like the gender proportions might vary with the customer location. (**Geography, Gender**)

EDA Results

Correlations and Effects - Numerical vs Categorical

(using One-Way ANOVA F-test or Two-Sample T-Test)

- The customer location could have an effect on their account balance. (**Geography, Balance**)
- The customer's account balance could be related to the number of products they have through the bank. (**NumOfProducts, Balance**)
- The customer age and their decision to leave the bank could be related. (**Age, Exited**)
- The customer age could be related to the number of products they own with the bank. (**Age, NumOfProducts**)
- The account balance could be related to the customer's decision to leave the bank. (**Balance, Exited**)
- The age and being an active member could be related. (**Age, IsActiveMember**)
- Customers of different age groups might have different gender proportions. (**Gender, Age**)

EDA Results

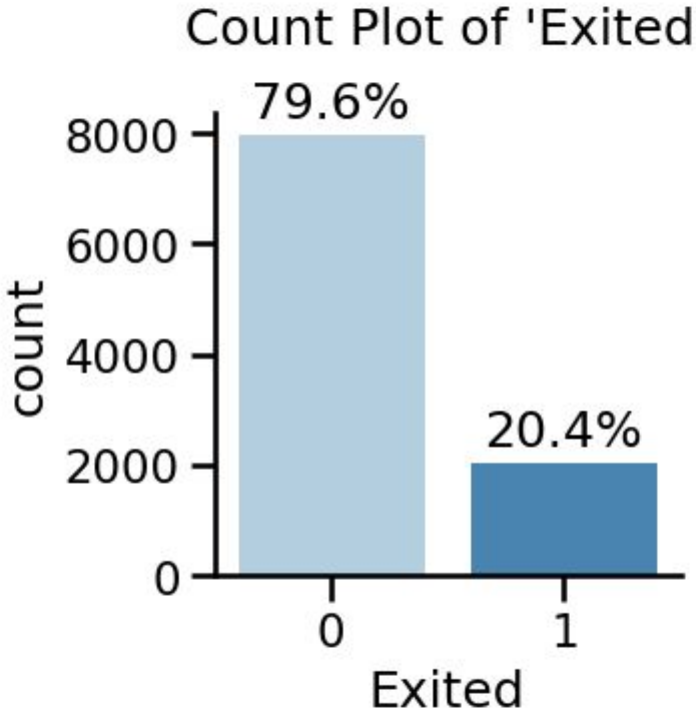
Correlations and Effects - Numerical vs Categorical - continued

(using One-Way ANOVA F-test or Two-Sample T-Test)

- The number of the years a customer has been with the bank could be related to them being active members or not. (**IsActiveMember, Tenure**)
- The credit score and the customer being an active member could be related. (**CreditScore, IsActiveMember**)
- The number of years a customer has been with the bank could be related to whether they have a credit card. (**HasCrCard, Tenure**)
- The customer credit score might be related to their decision to leave the bank or not. (**CreditScore, Exited**)

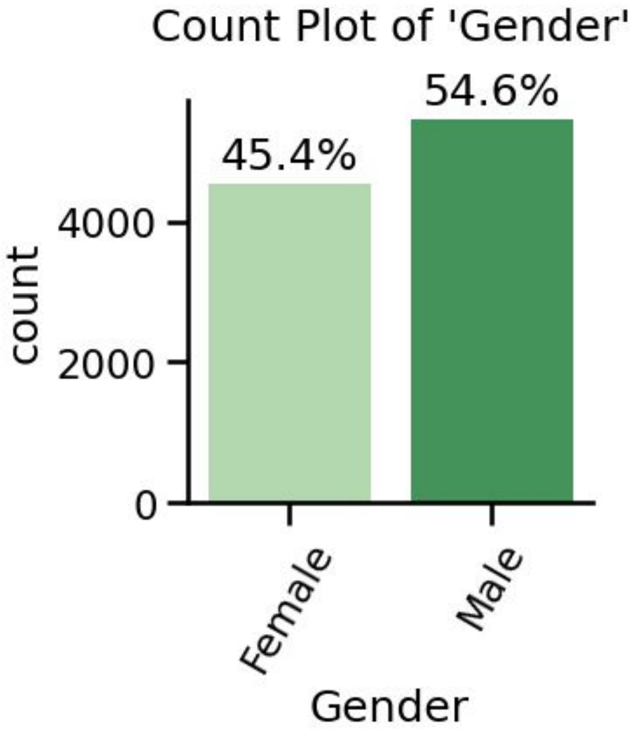
EDA - Univariate - Exited

- About 80% of the customers have not left the bank within the past 6 months.



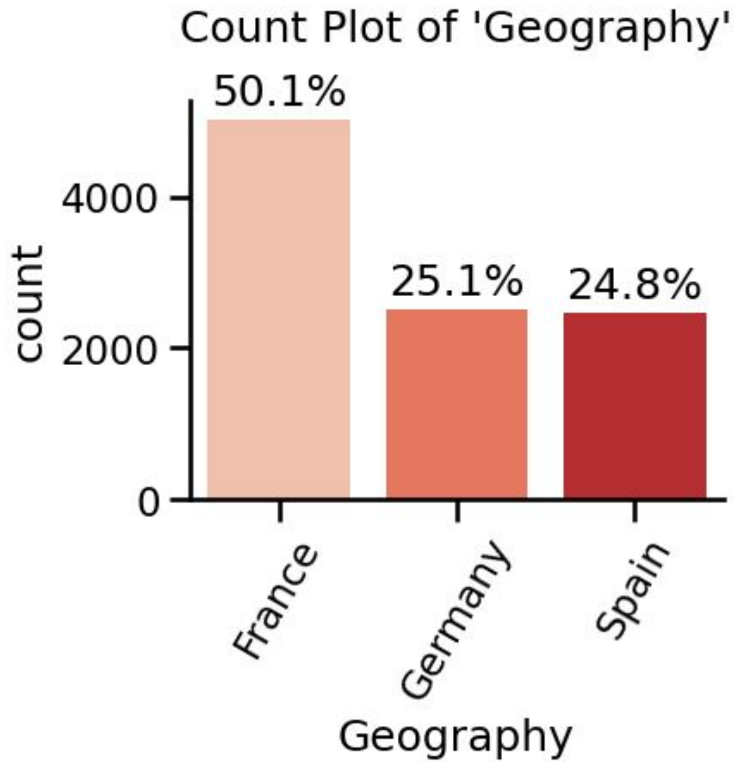
EDA - Univariate - Gender

- More than half (54.6%) of the customers are male.



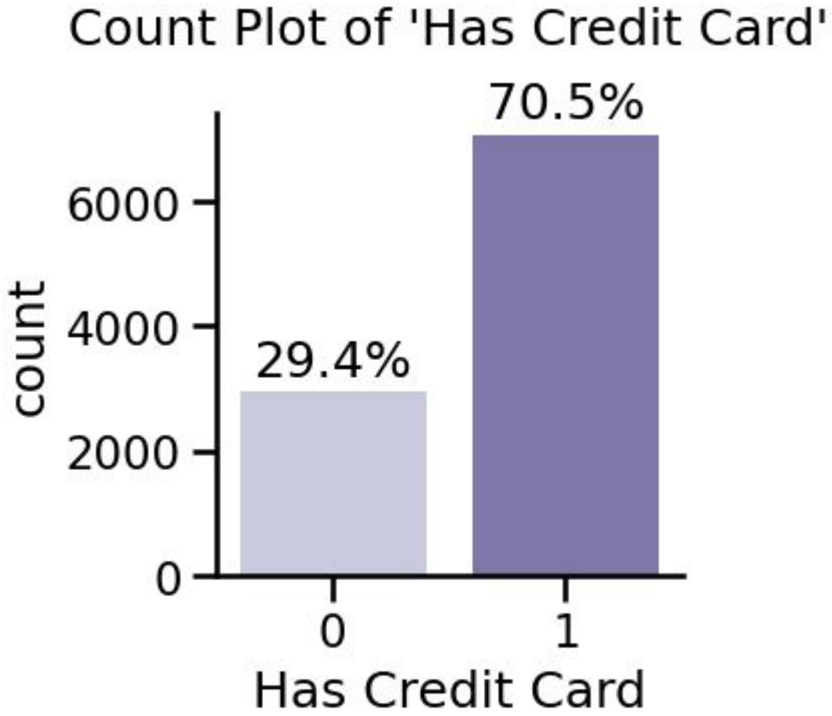
EDA - Univariate - Geography

- Half of the customers live in France.



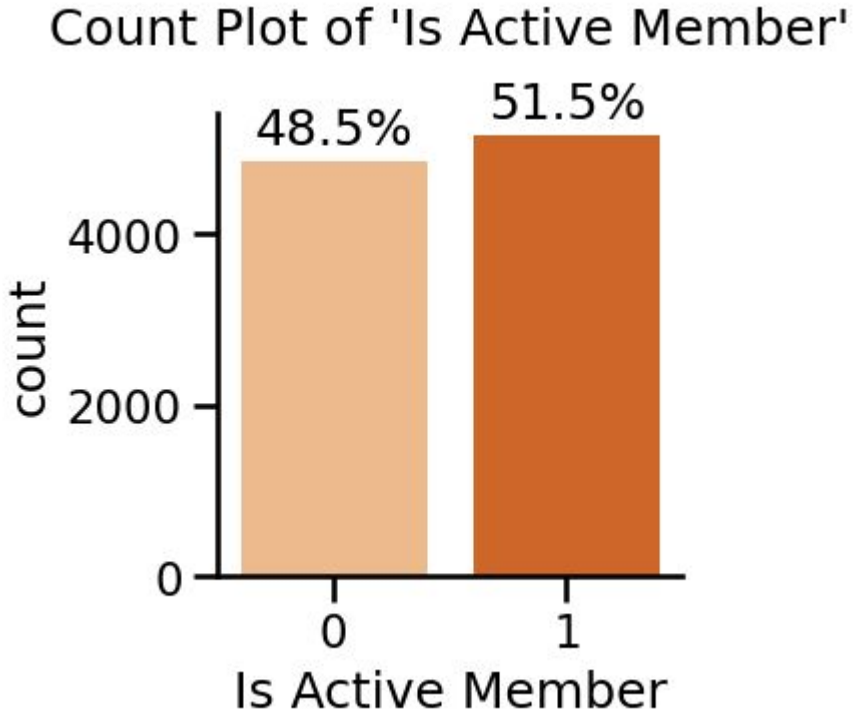
EDA - Univariate - HasCrCard

- About 70% of the customers have credit cards.



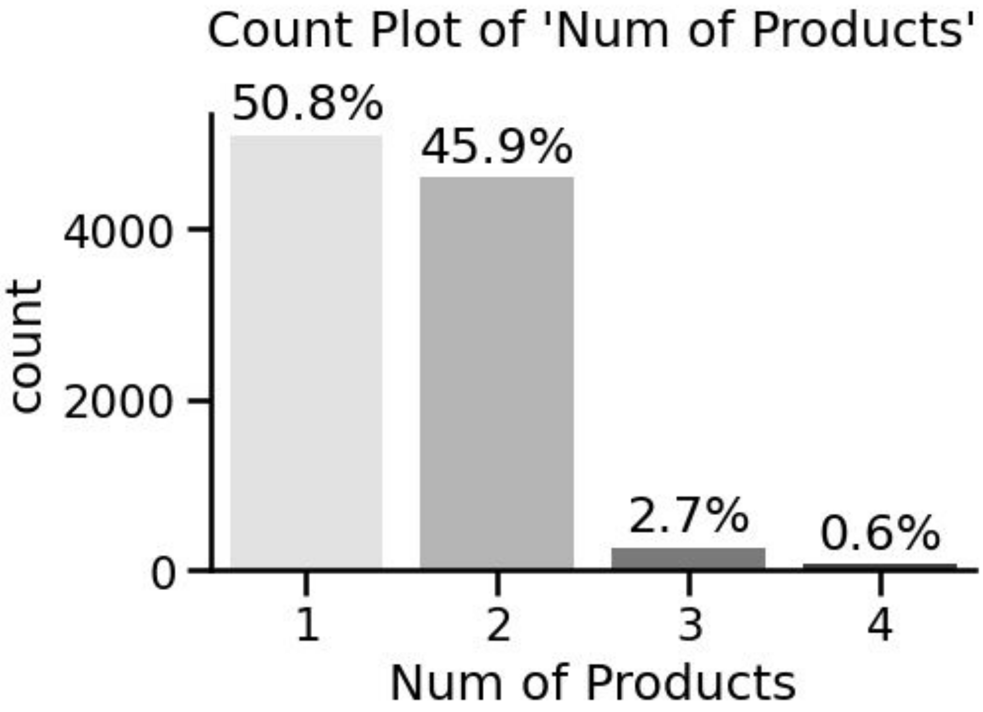
EDA - Univariate - IsActiveMember

- More than half (~52%) of the customers are active members of the bank.



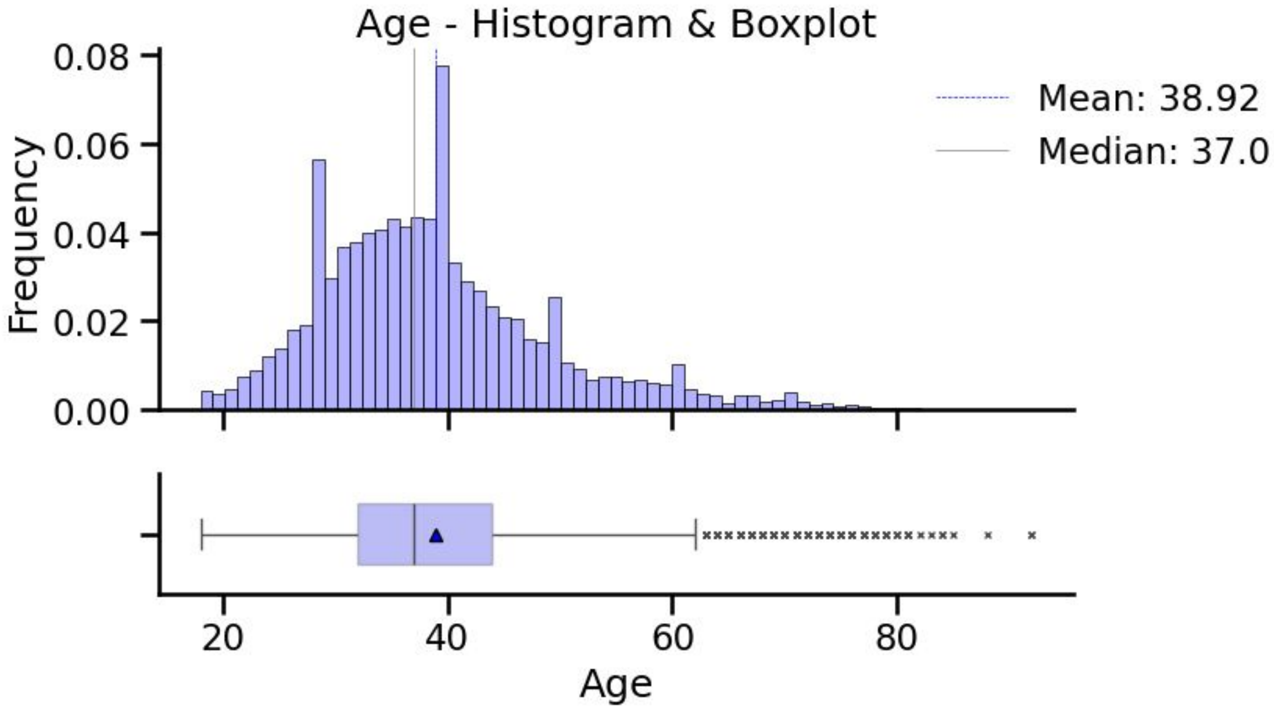
EDA - Univariate - NumOfProducts

- About half (~51%) of the customers have 1 product through the bank followed by 46% of them who have 2 products.



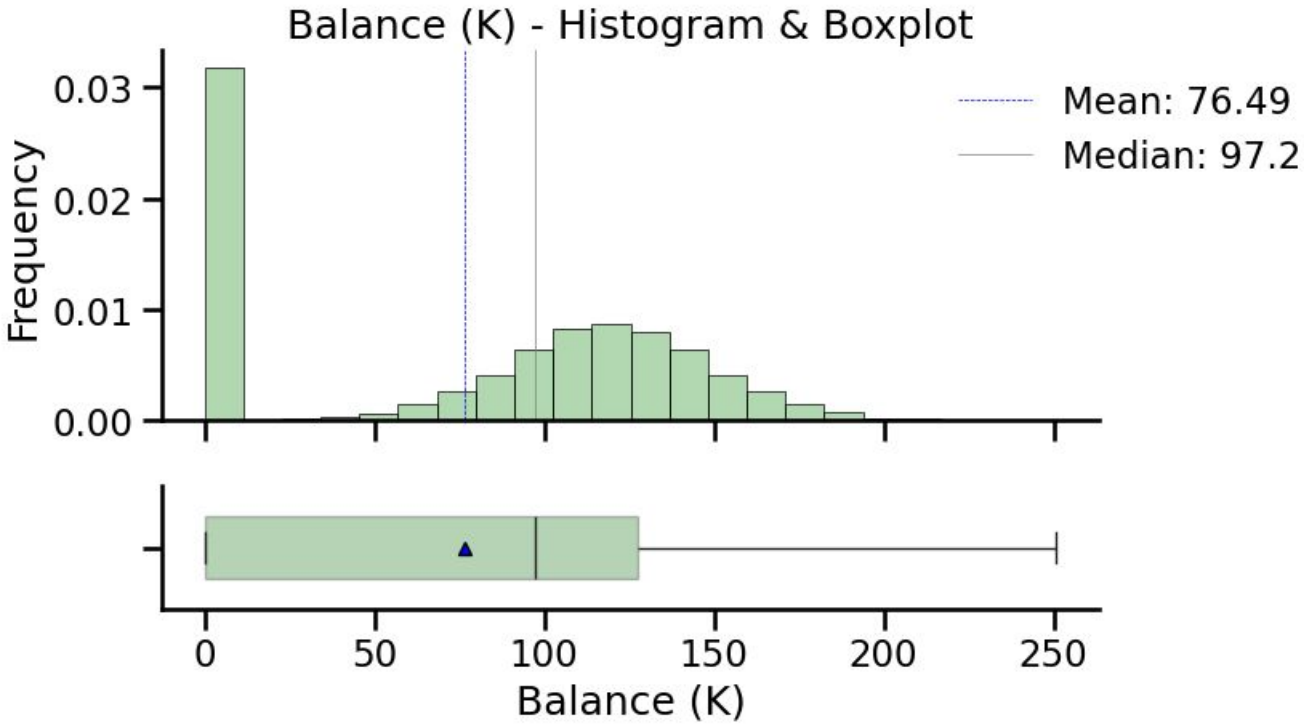
EDA - Univariate - Age

- The customer age is positively skewed with mean of 39 years old and median of 37 years old.



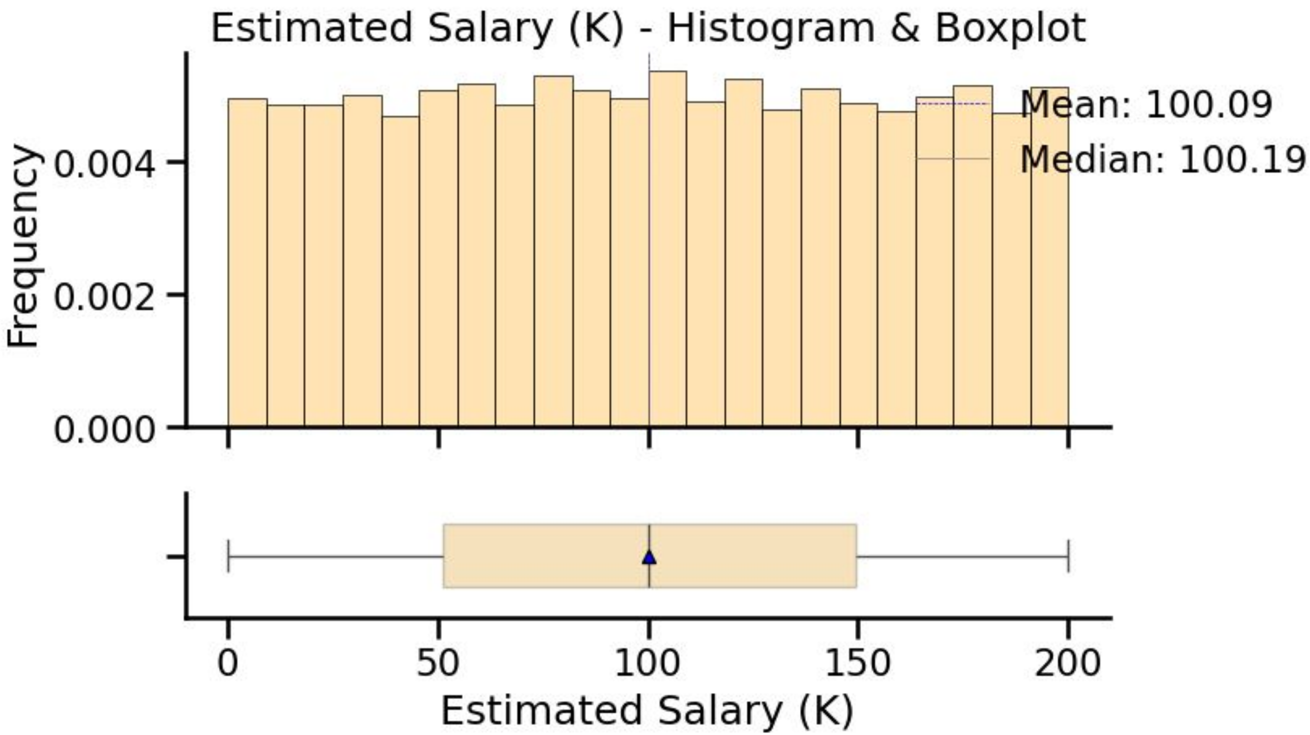
EDA - Univariate - Account Balance

- The non-zero balance is almost normally distributed. The median balance is ~ $98K.



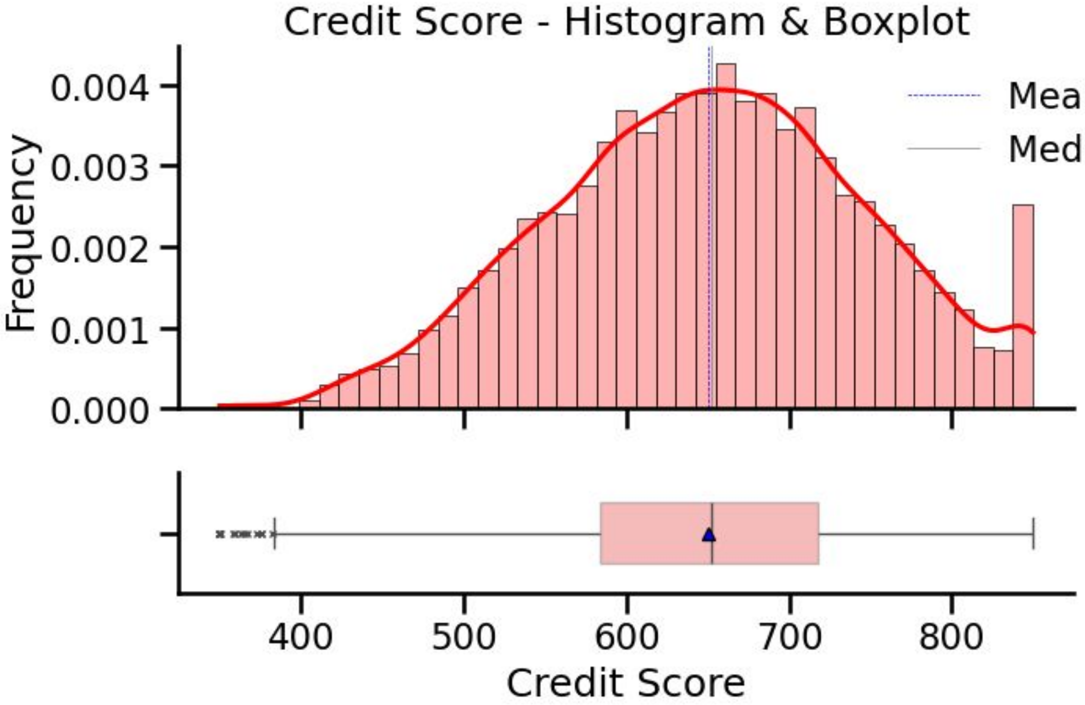
EDA - Univariate - Estimated Salary

- Estimated Salary is slightly positively skewed with average of $100 K.



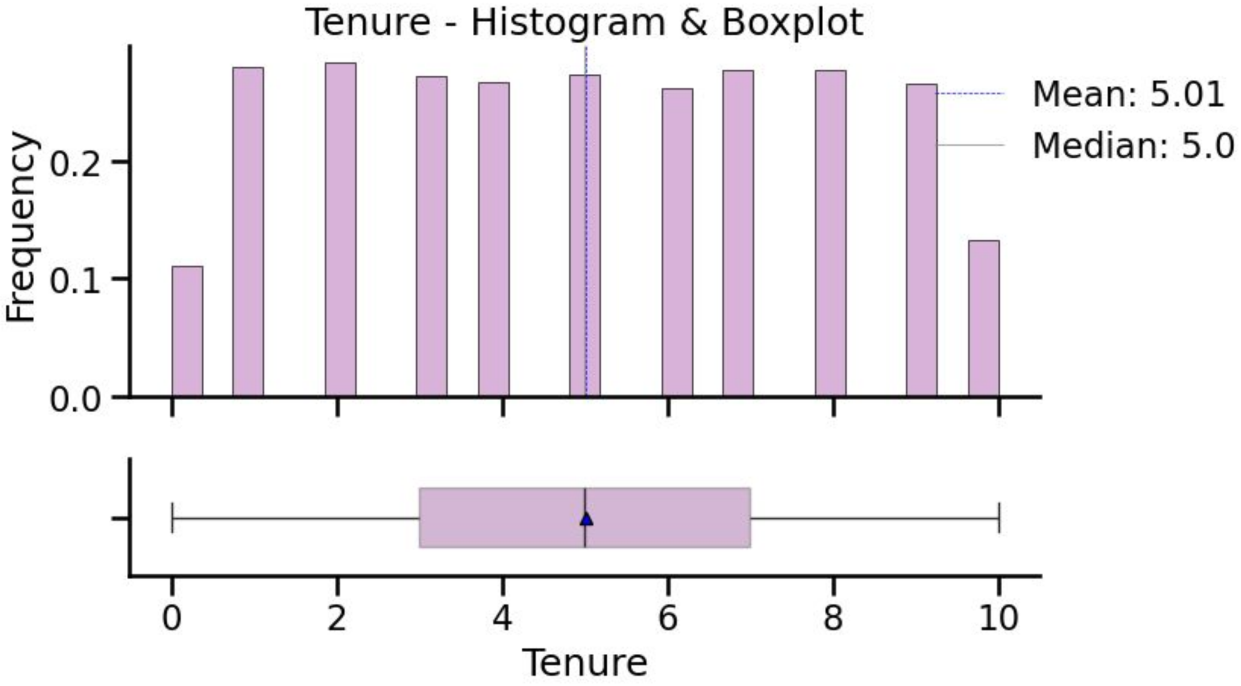
EDA - Univariate - Credit Score

- Credit score is slightly negatively skewed with median of 652 and it is likely close to normal distribution.



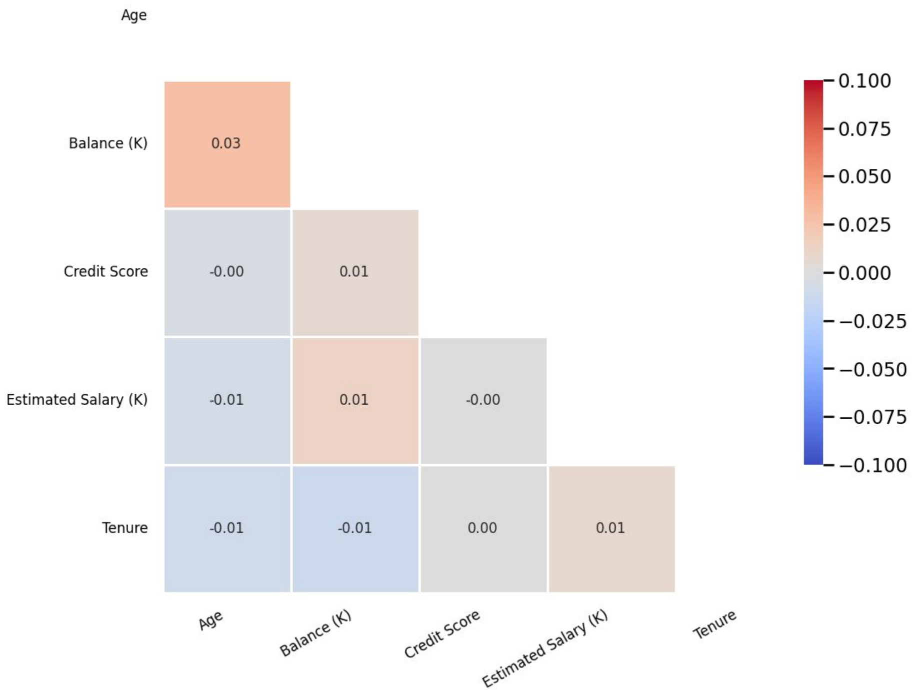
EDA - Univariate - Tenure

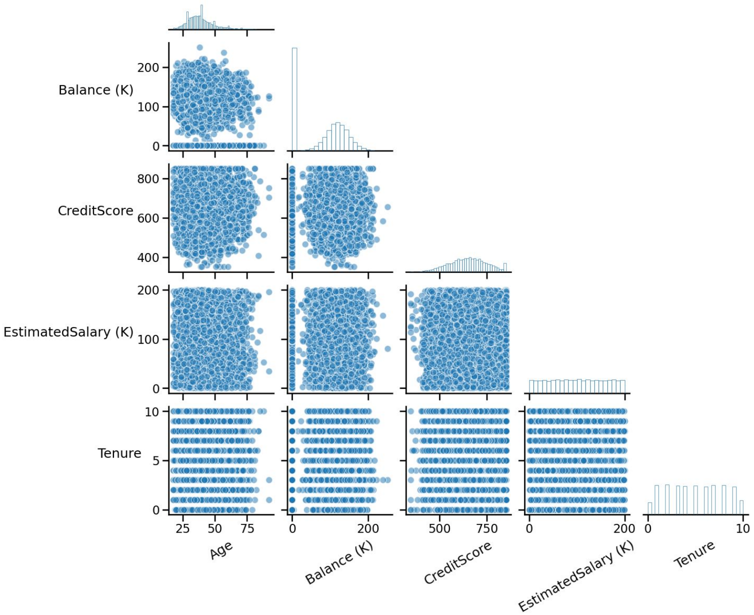
- The average number of years the customers have been with the bank is about 5 years.



EDA - Bivariate - *Numerical Variables*

- We do not observe significant correlation amongst the numerical variables.

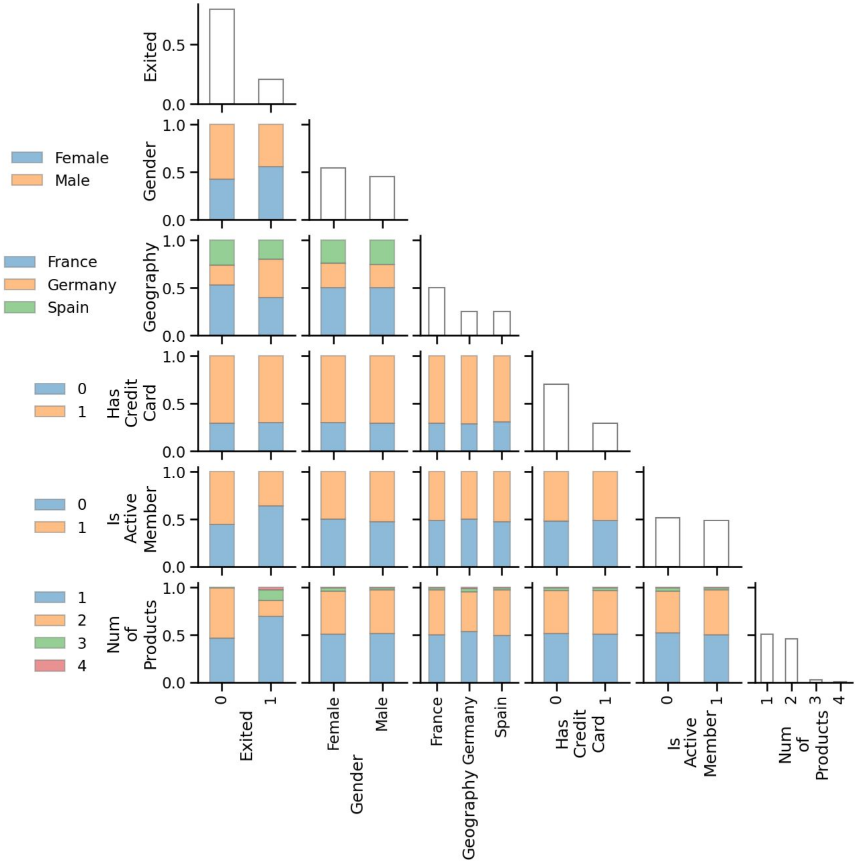




EDA - Bivariate - *Categorical Variables*

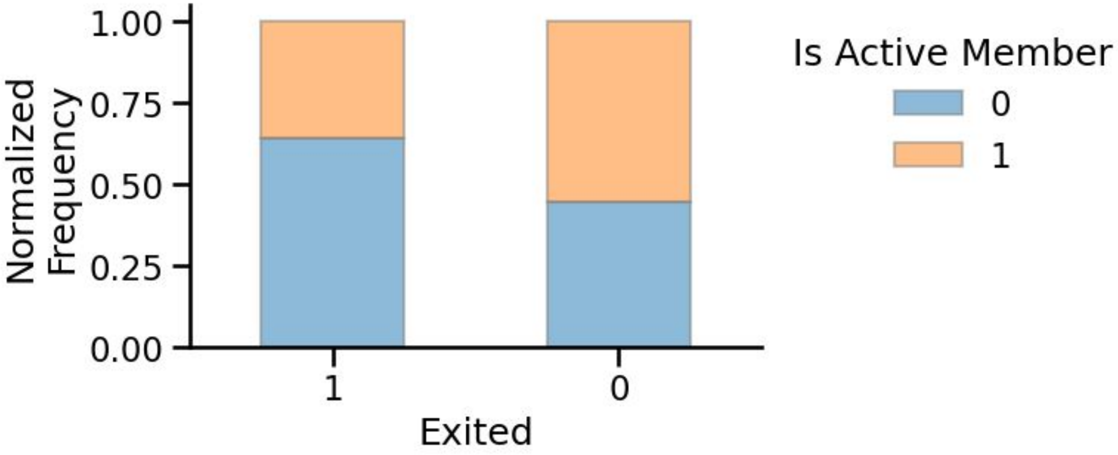
- By conducting χ^2 test of independence we observe that the following variables have effects on each other:

Category 1	Category 2	p-value
Exited	NumOfProducts	0.0e+00
Exited	Geography	3.8e-66
Exited	IsActiveMember	8.8e-55
Exited	Gender	2.2e-26
Geography	NumOfProducts	6.7e-09
Gender	NumOfProducts	1.3e-04
IsActiveMember	NumOfProducts	6.4e-04
Gender	IsActiveMember	2.5e-02
Gender	Geography	3.1e-02



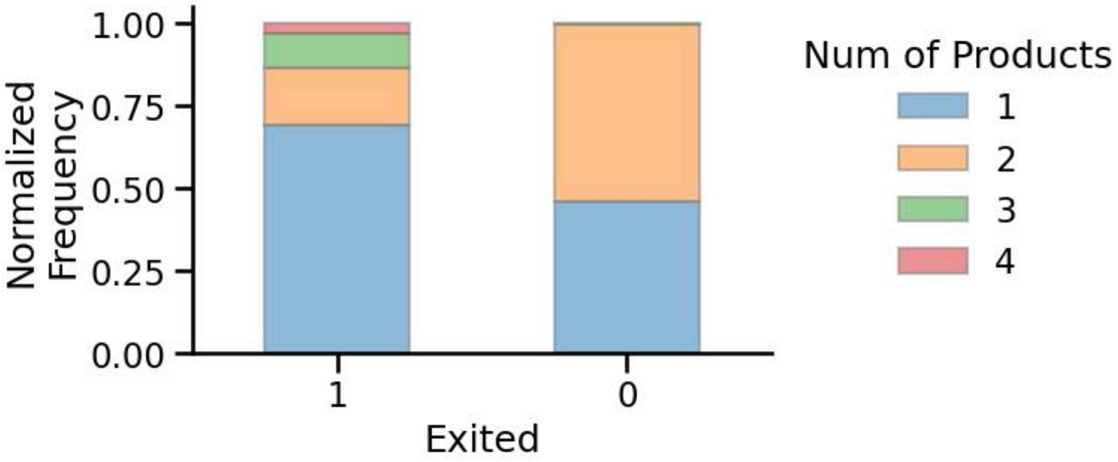
EDA - Bivariate - Exited vs IsActiveMember

- The proportion of non-active members is higher amongst customers who leave the bank.



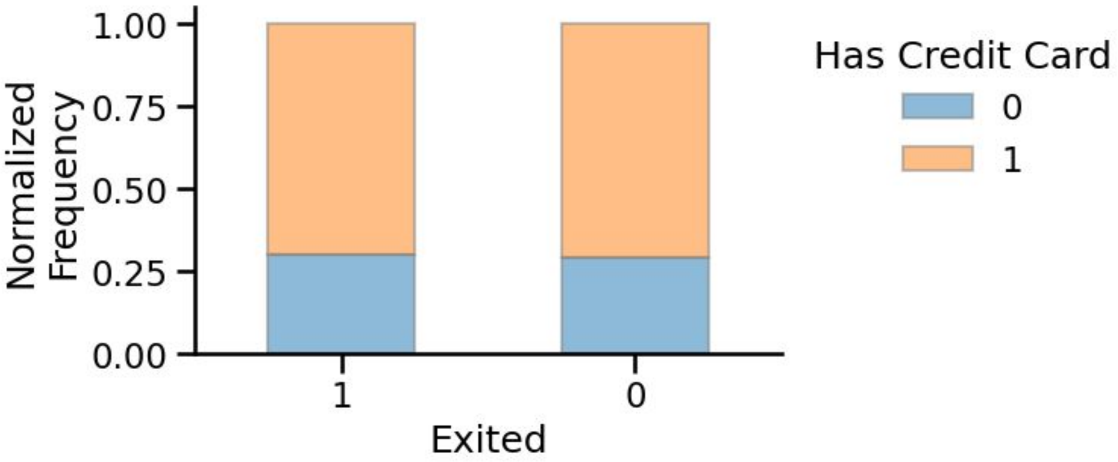
EDA - Bivariate - Exited vs NumOfProducts

- The tendency to leave the bank is different amongst the customers who have 1, 2, 3, or 4 bank products.



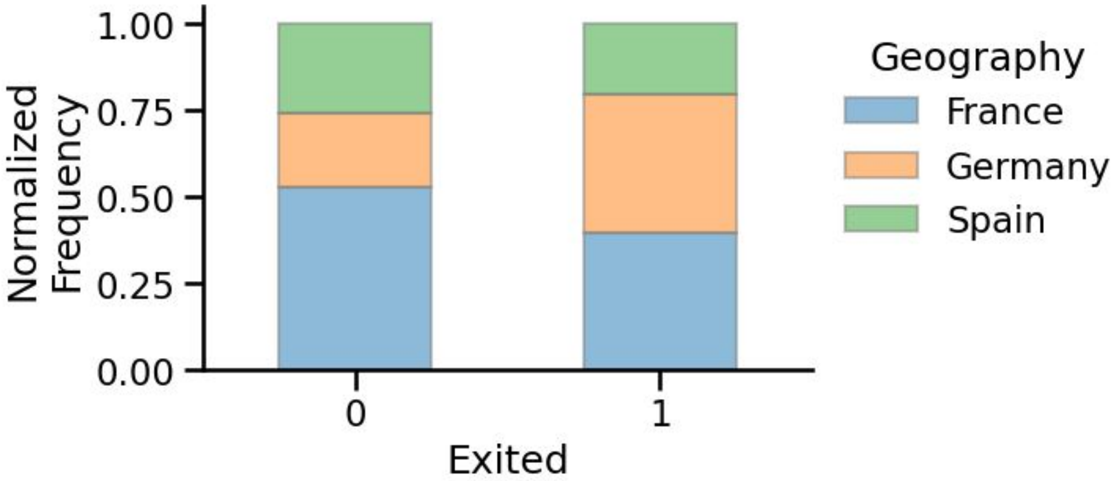
EDA - Bivariate - Exited vs HasCrCard

- Having credit cards or not does not seem to affect the customer's decision to leave the bank or not.



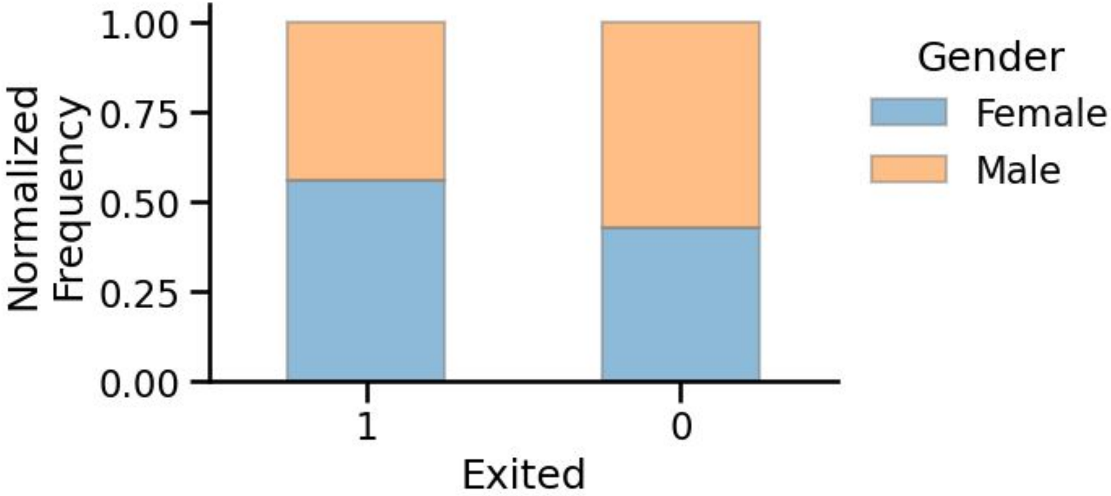
EDA - Bivariate - Exited vs Geography

- The proportions of customers in different locations is different amongst those customers who leave the bank and those who do not.



EDA - Bivariate - Exited vs Gender

- The proportion of female customers is higher among those who leave the bank.



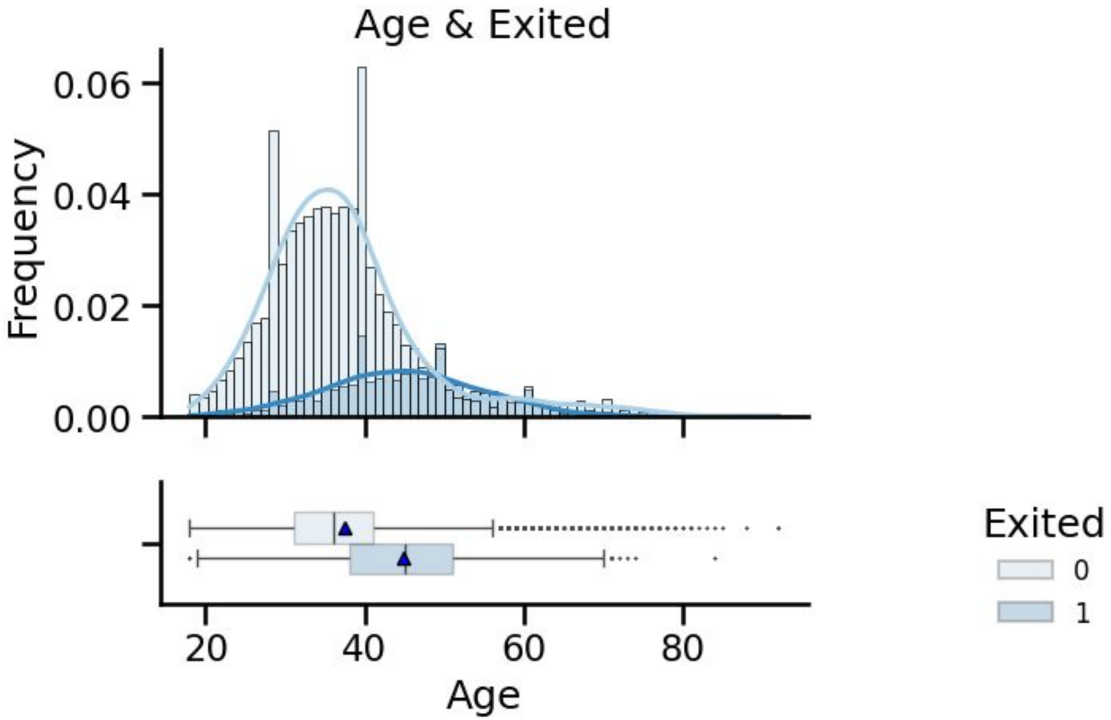
EDA - Bivariate - *Categorical-Numerical Variables*

- By conducting **One-Way ANOVA F-test** we observe that the following variables have effects on each other:
  - In the following slides, we take a look at some of these relationships.

Category	Numerical	p-value
Geography	Balance (K)	0.0e+00
NumOfProducts	Balance (K)	3.7e-315
Exited	Age	1.2e-186
NumOfProducts	Age	5.2e-33
Exited	Balance (K)	1.3e-32
IsActiveMember	Age	1.1e-17
Geography	Age	5.6e-06
IsActiveMember	Tenure	4.6e-03
Gender	Age	5.9e-03
Exited	CreditScore	6.7e-03
IsActiveMember	CreditScore	1.0e-02
HasCrCard	Tenure	2.4e-02

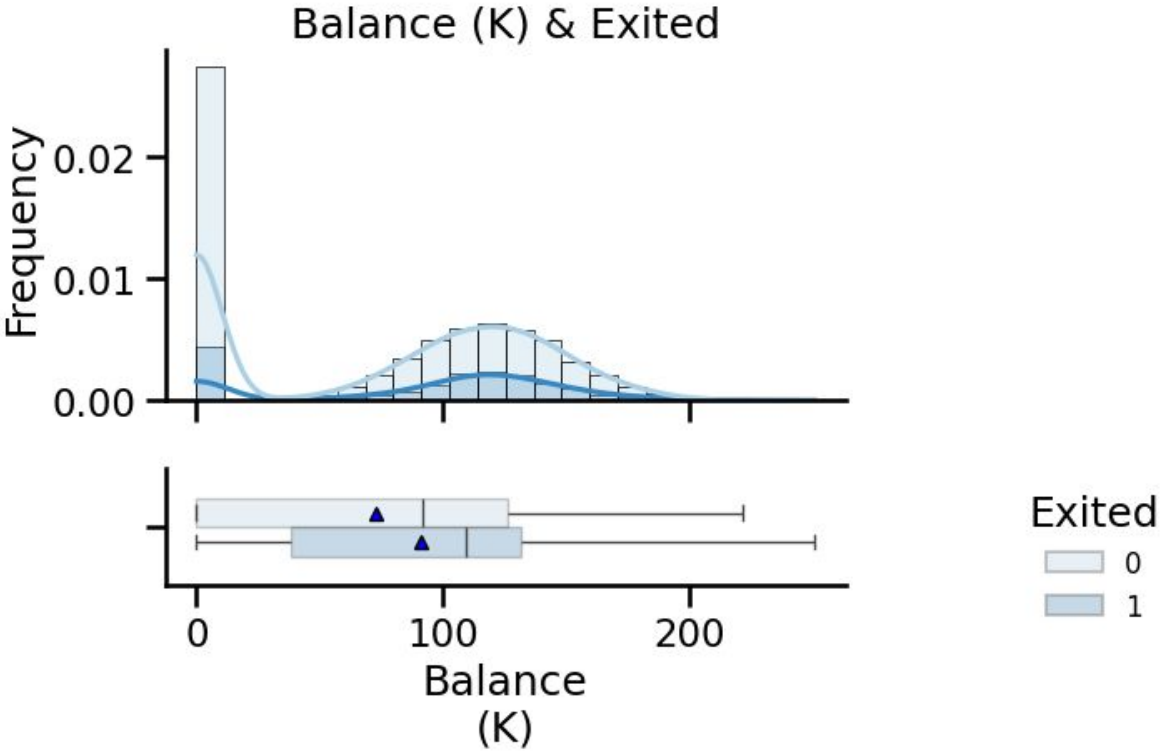
EDA - Bivariate - Exited vs Age

- It looks like older customer have a higher tendency for leaving the bank.



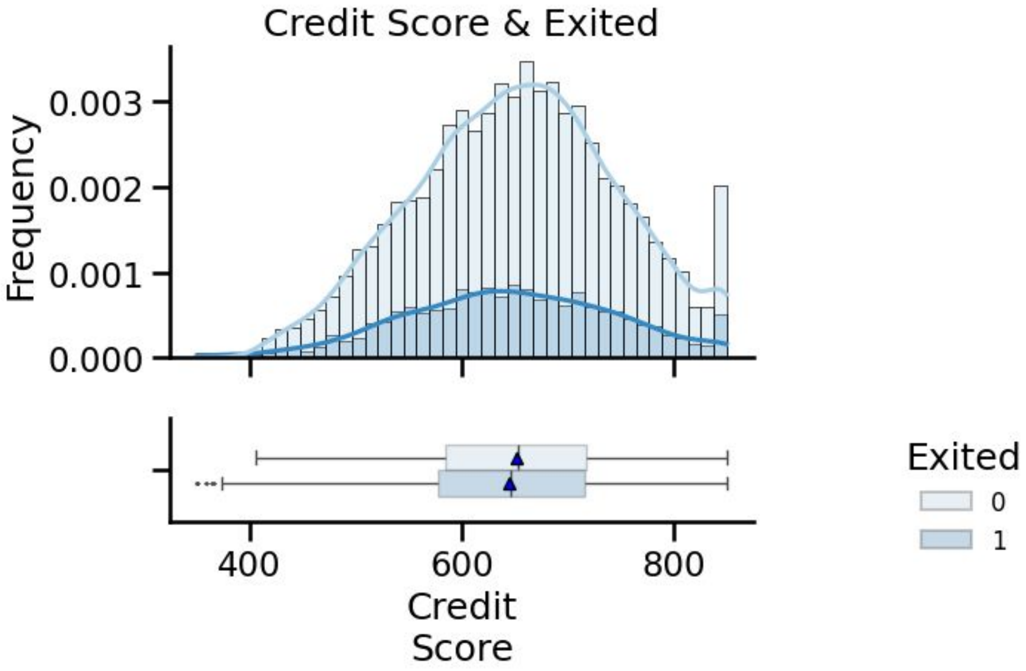
EDA - Bivariate - Exited vs Account Balance

- The customers with higher account balance seem to have more tendency to leave the bank.



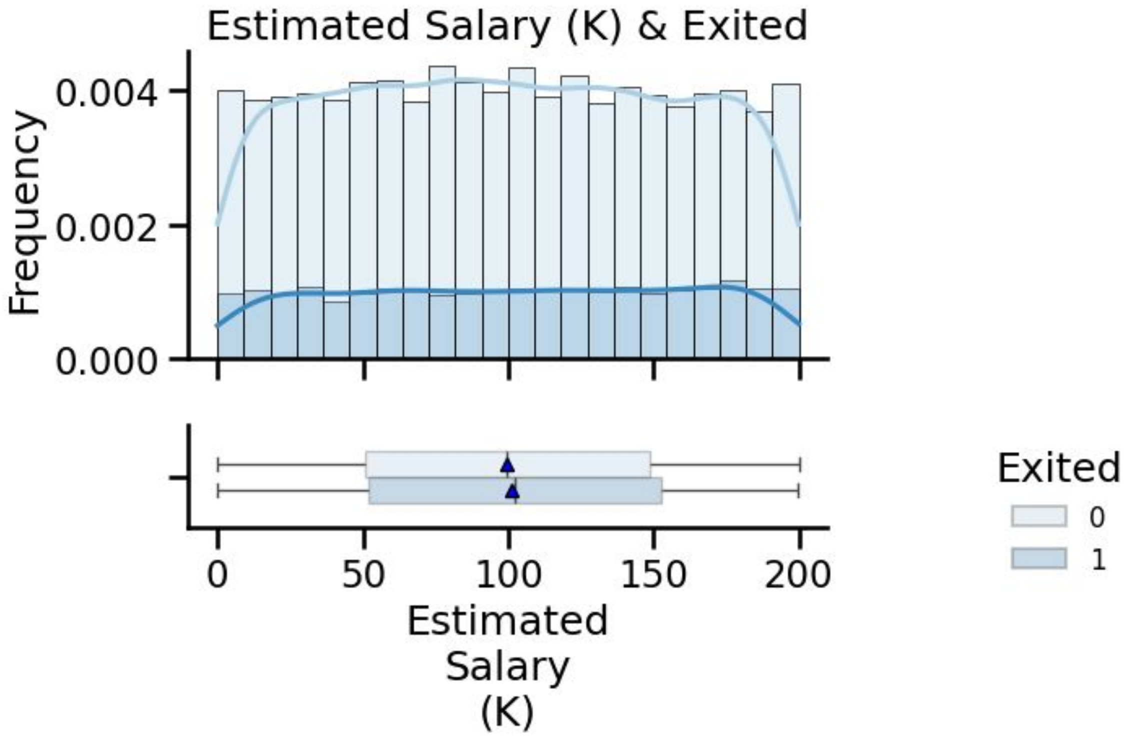
EDA - Bivariate - Exited vs Credit Score

- There's a slight more tendency within the customers with lower credit scores to leave the bank.



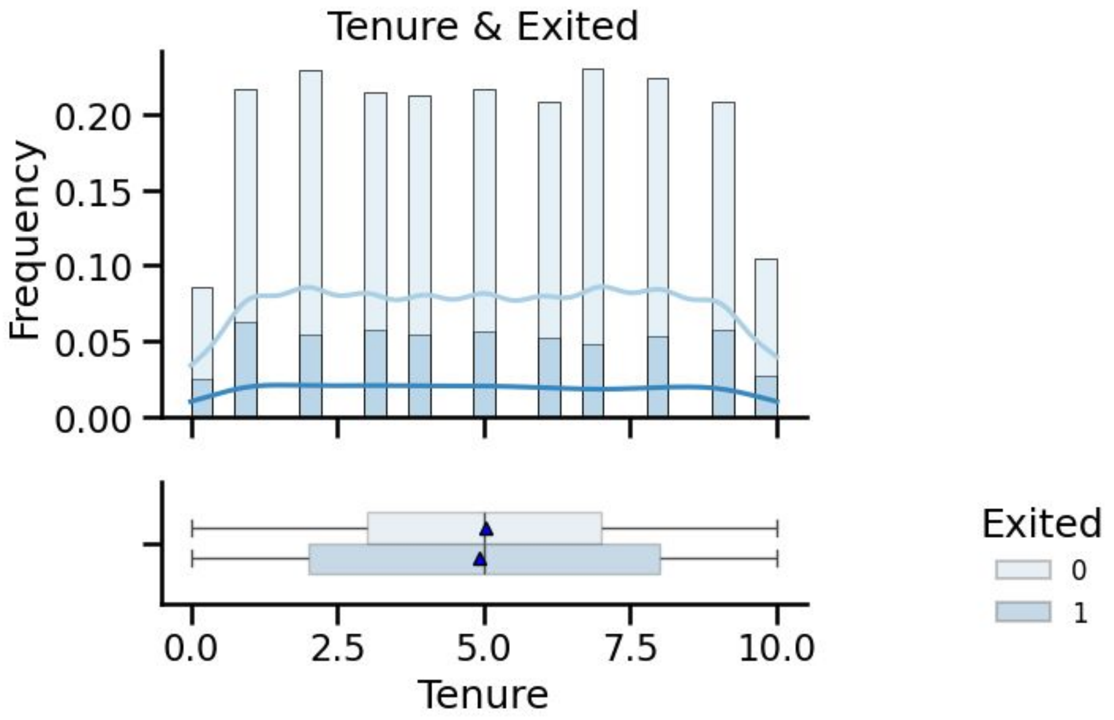
EDA - Bivariate - Exited vs Estimated Salary

- The average salary of the customers who leave the bank is slightly higher.



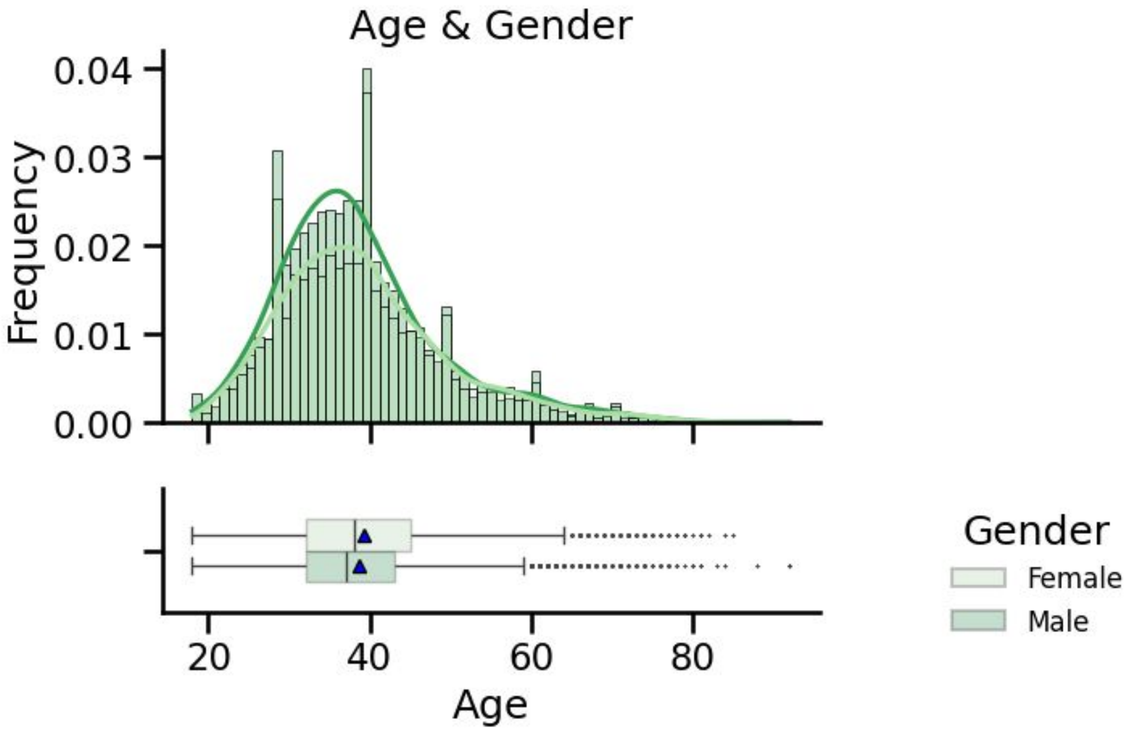
EDA - Bivariate - Exited vs Tenure

- The years at the bank do not seem to have a relationship with the customer's decision to leave the bank.



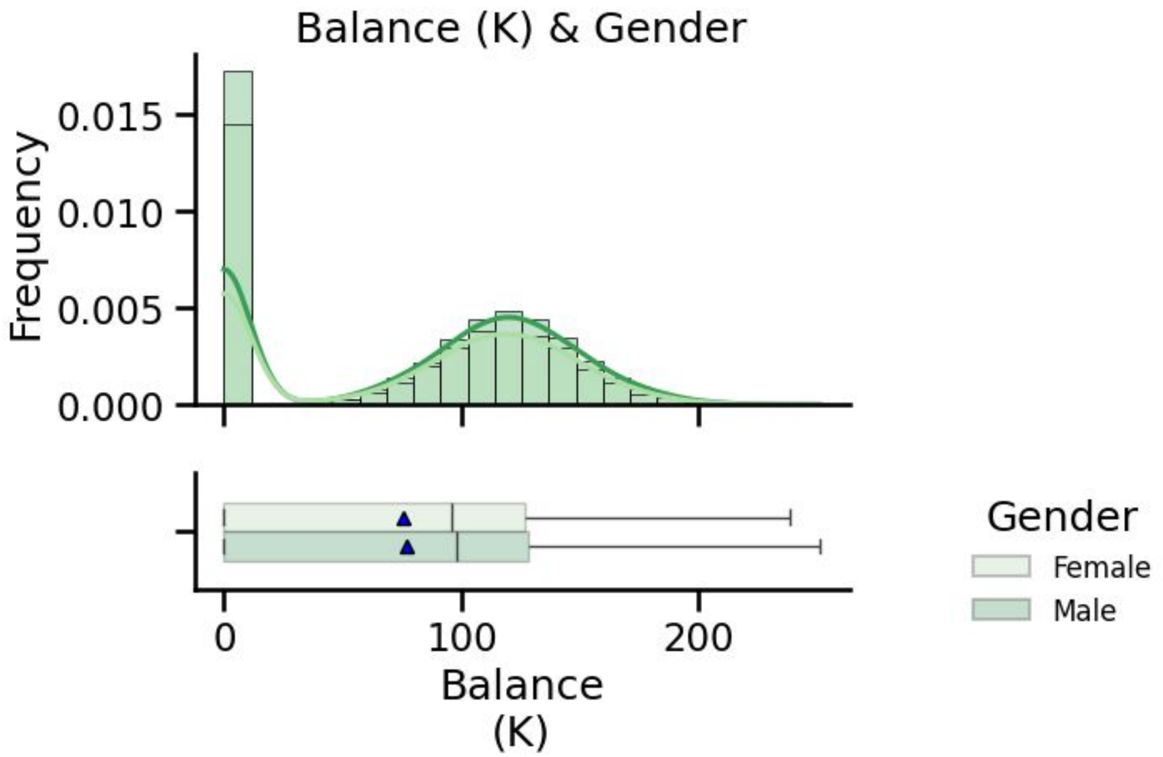
EDA - Bivariate - Age vs Gender

- The mean age range amongst female customers is slightly higher.



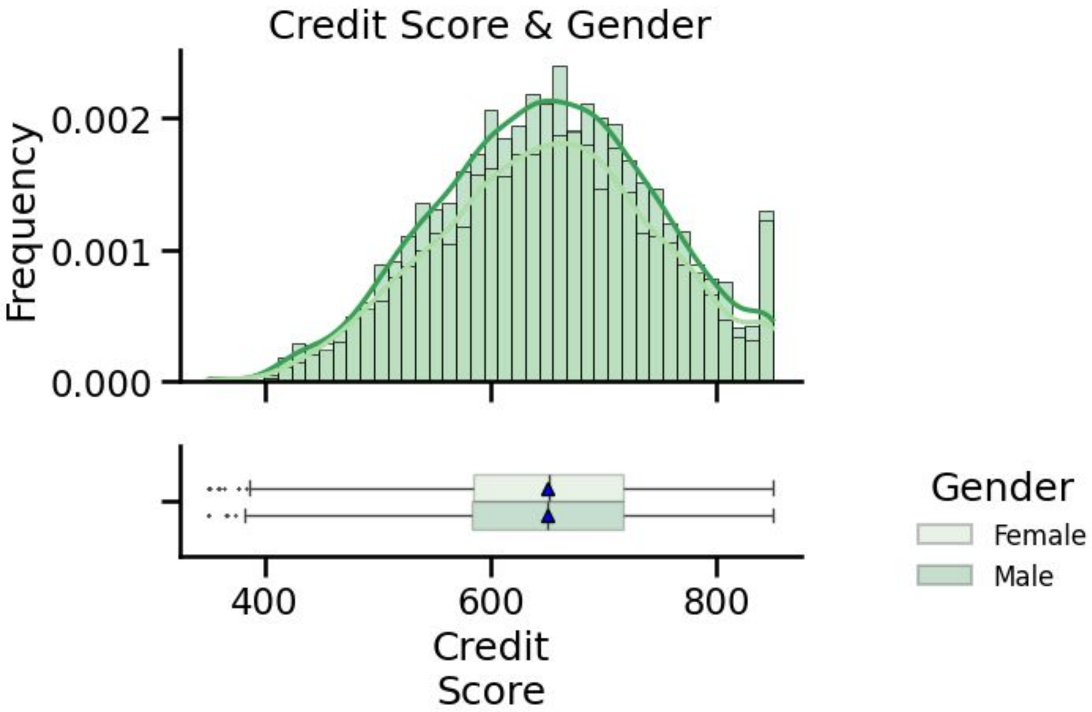
EDA - Bivariate - Balance vs Gender

- The average account balance is very slightly higher amongst male customers.



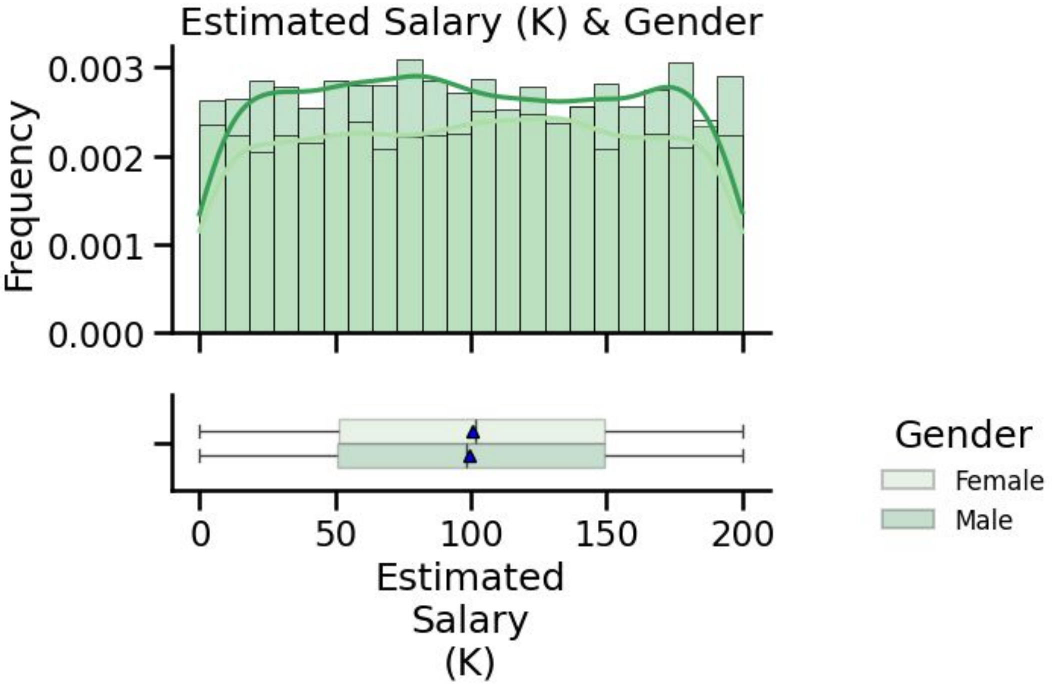
EDA - Bivariate - Credit Score vs Gender

- The credit score distribution seems to be very similar amongst male and female customers.

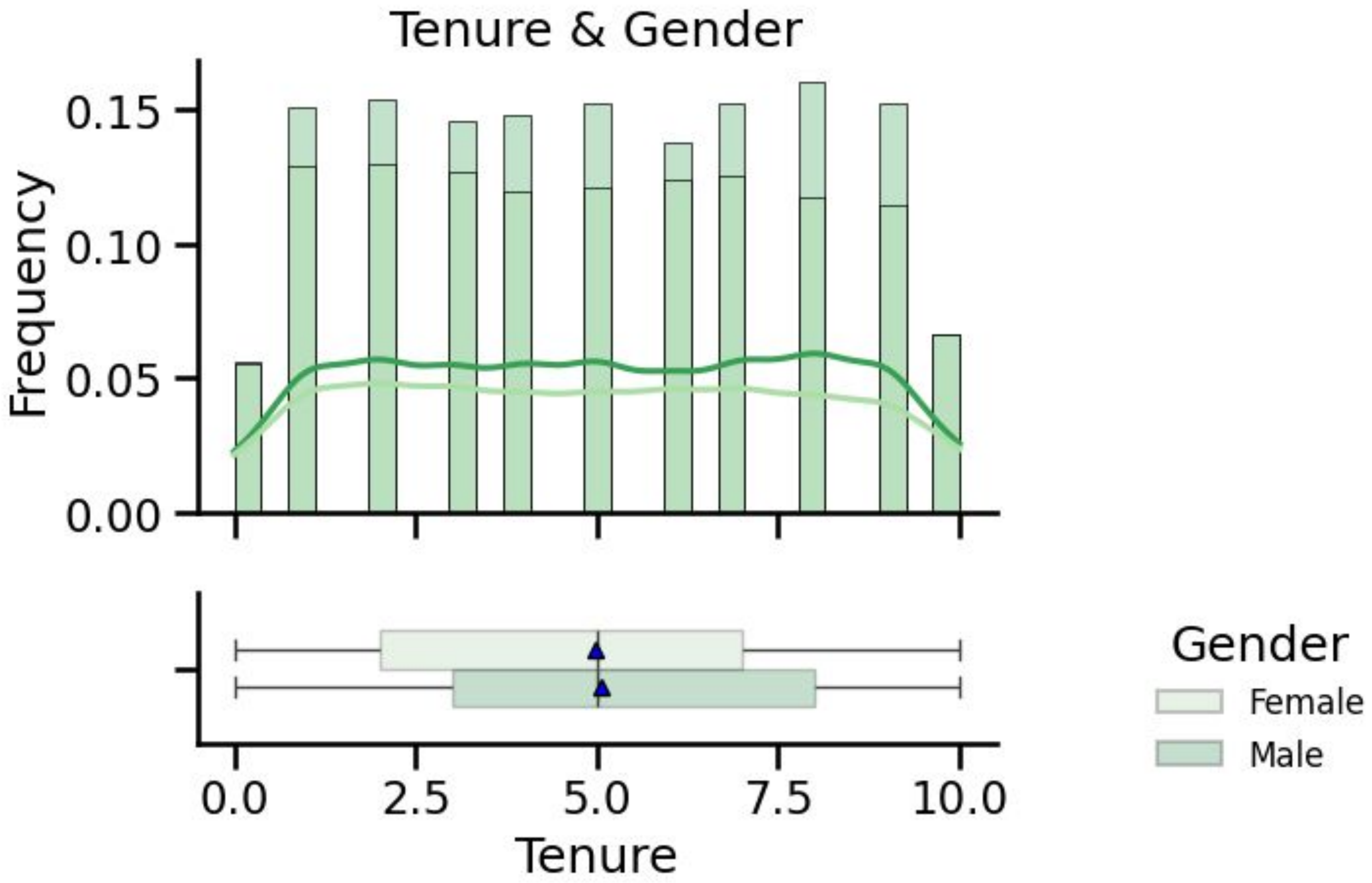


EDA - Bivariate - Salary vs Gender

- The median estimated salary amongst the female customers is slightly higher than the male customers.



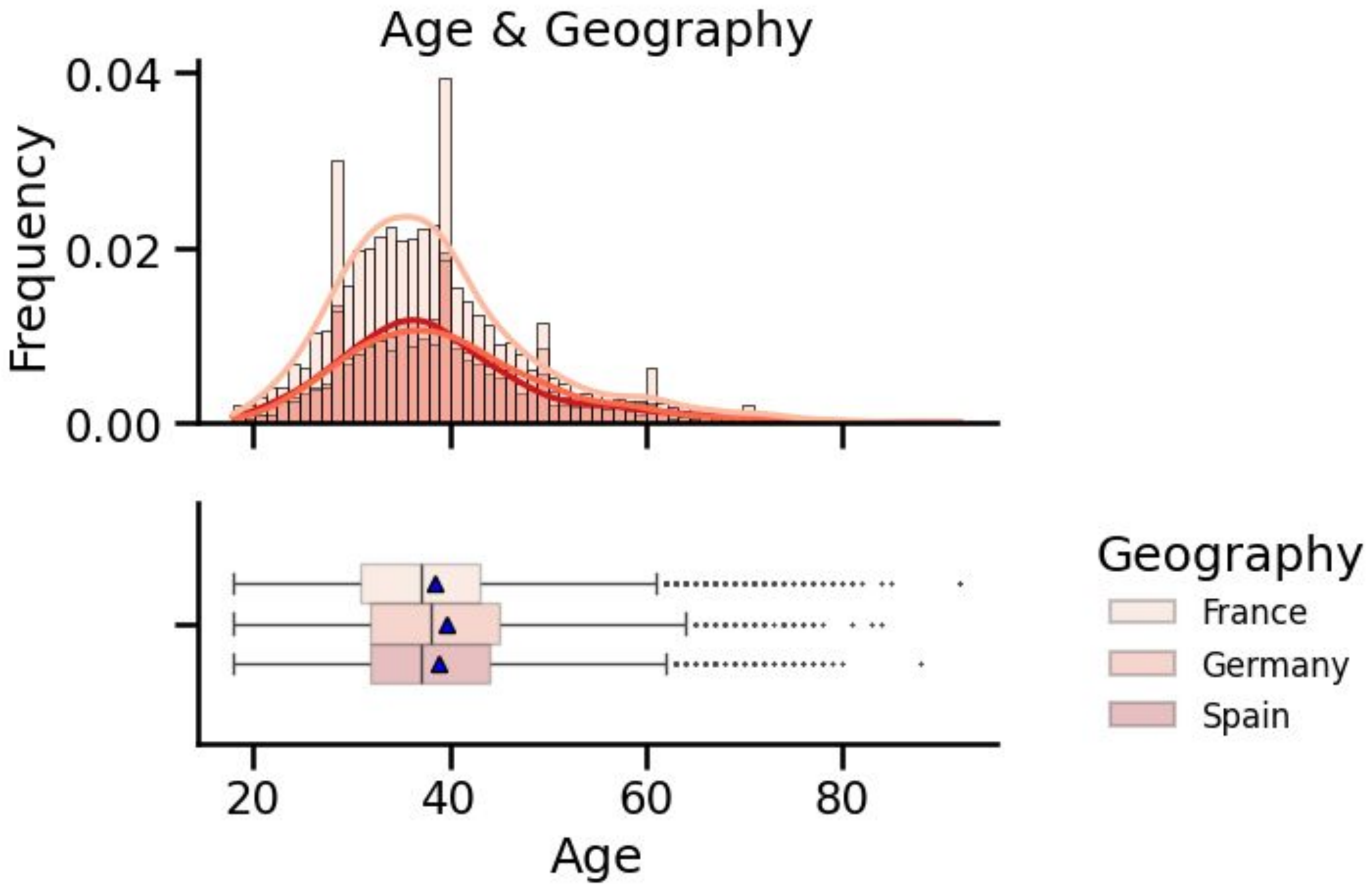
EDA -



Tenure & Gender
Frequency
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
Tenure
Gender
Female
Male

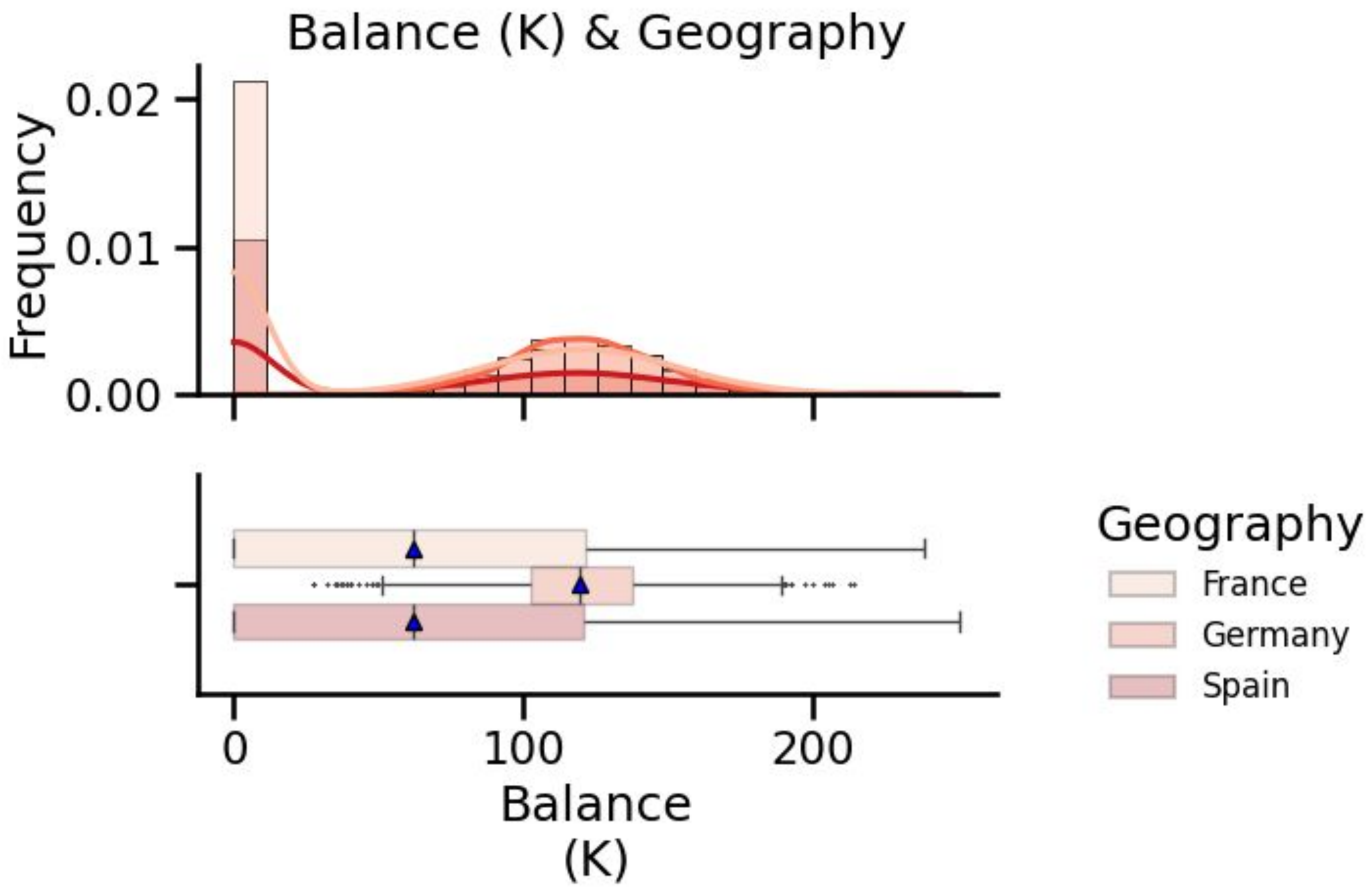
EDA -

•

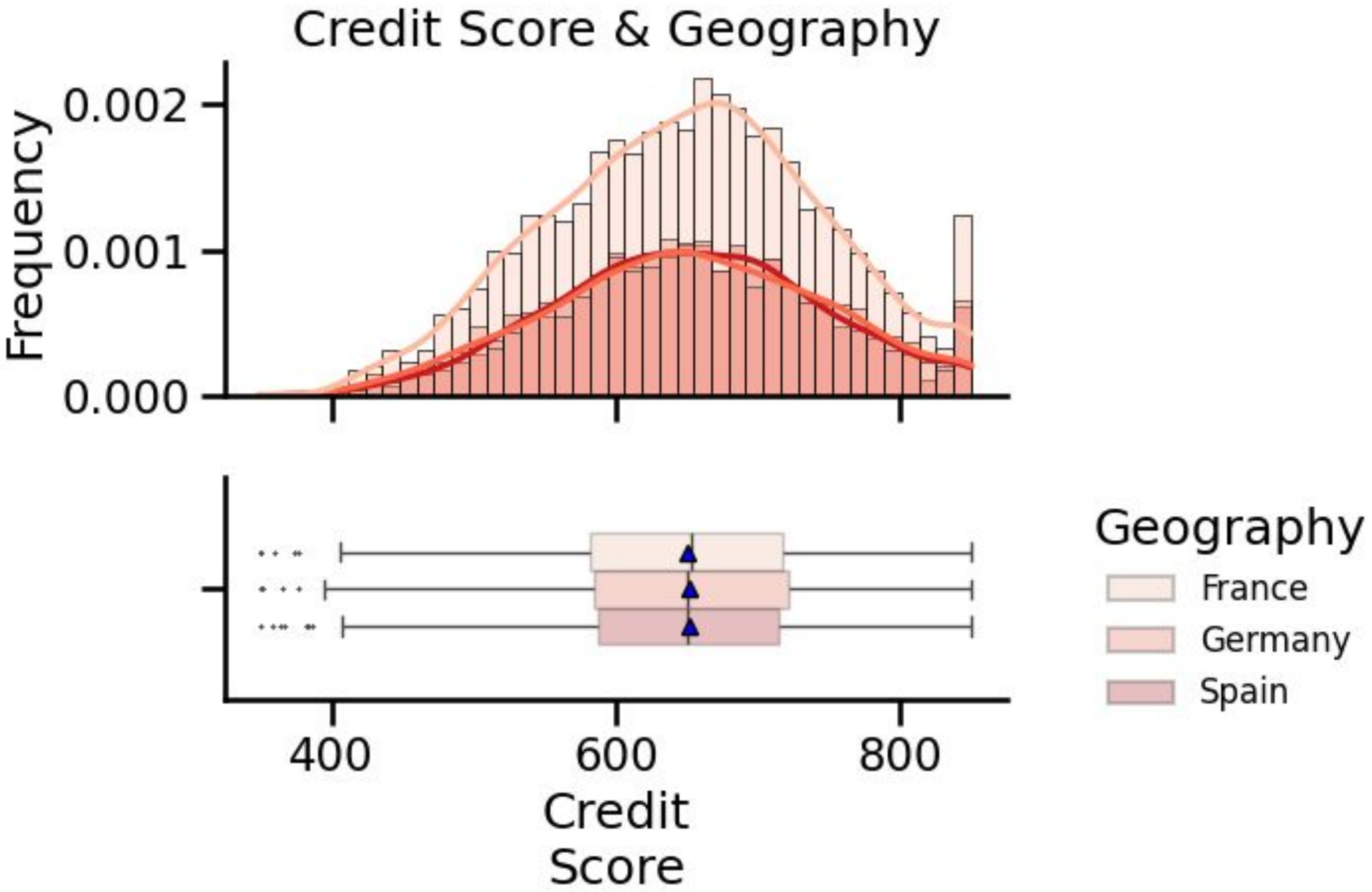

Age & Geography
Frequency
0.04
0.02
0.00
20
40
60
80
Age
Geography
France
Germany
Spain

EDA - Bivariate

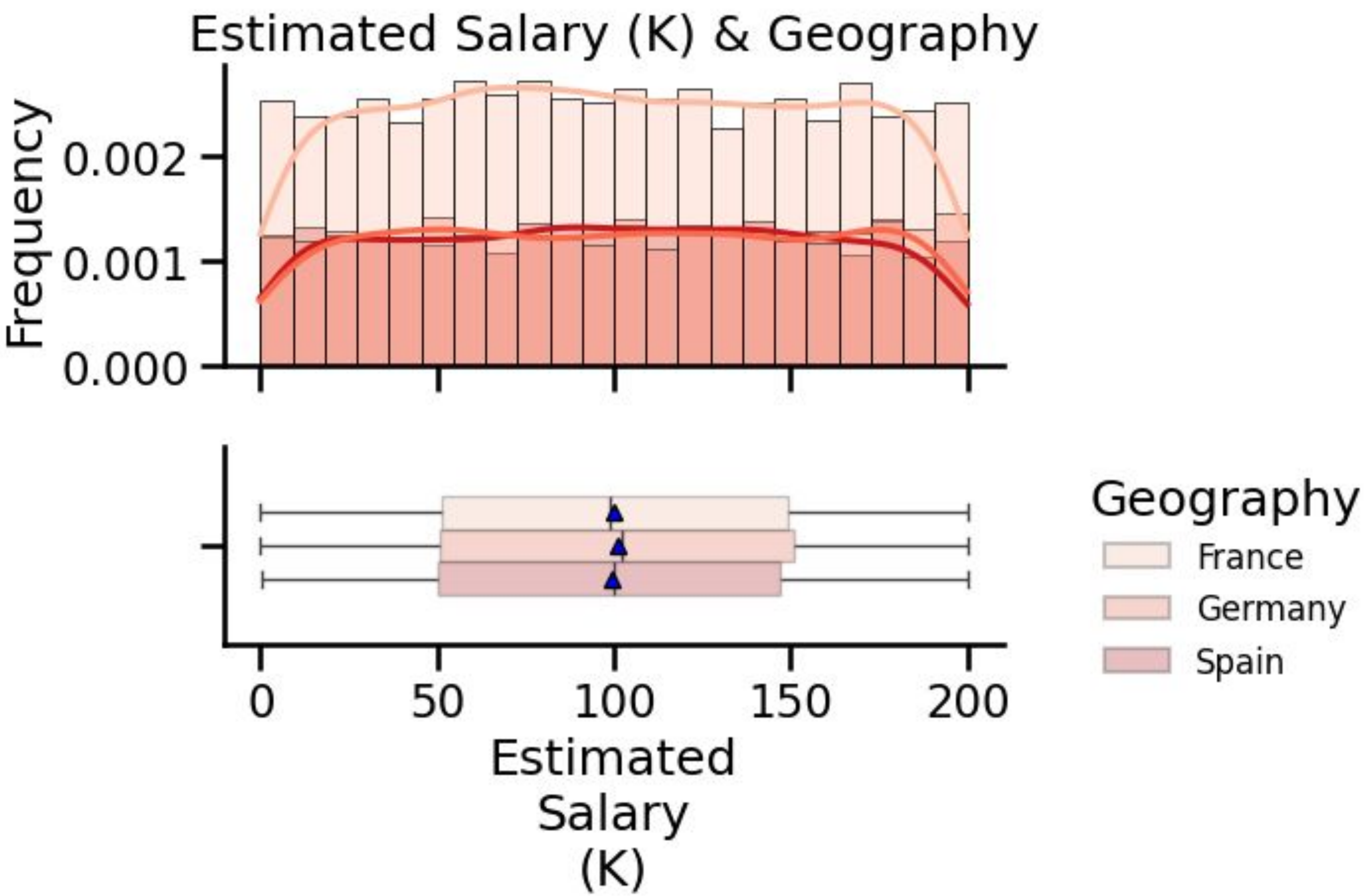
•



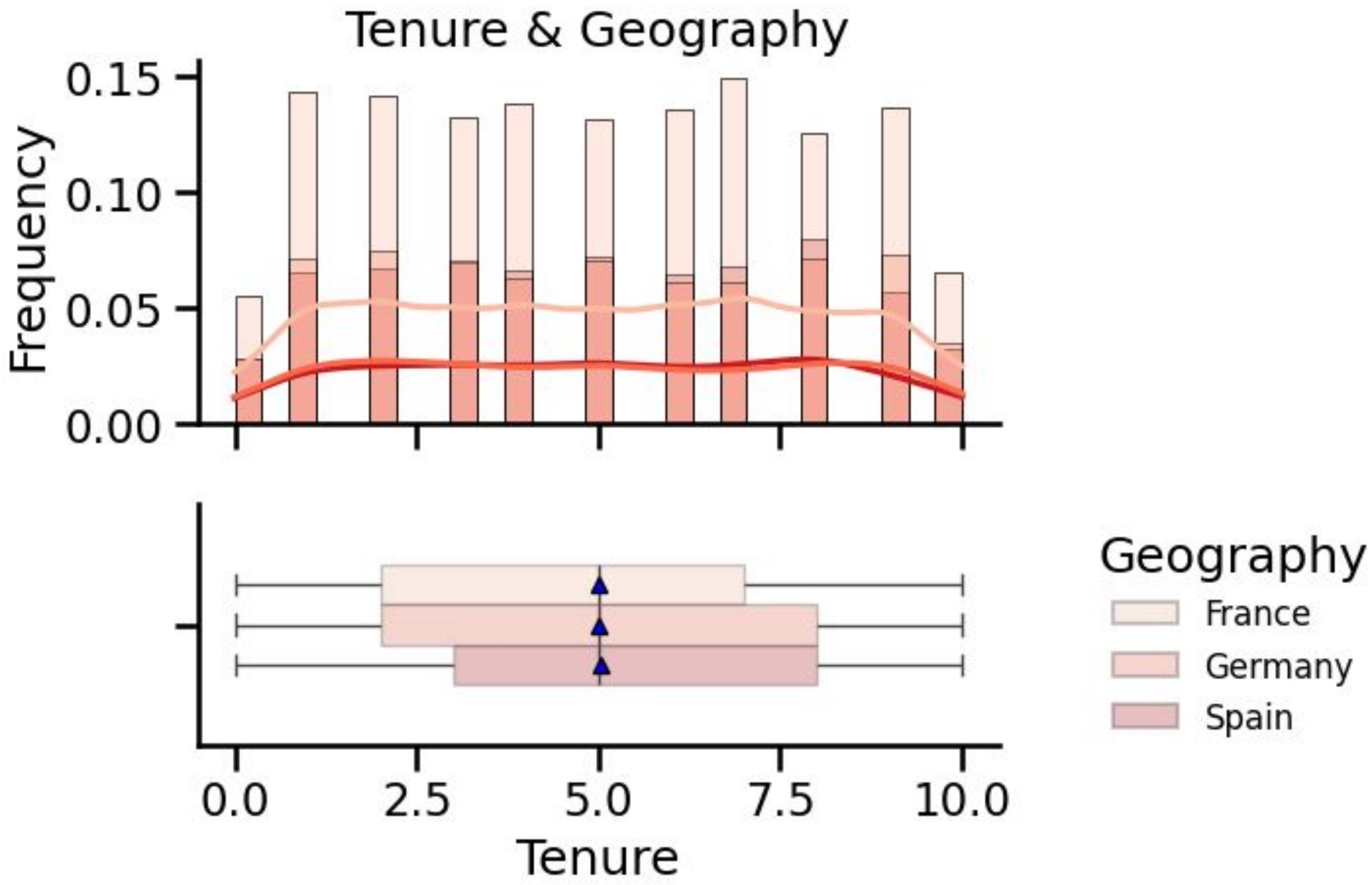
EDA -


Credit Score & Geography
Frequency
0.002
0.001
0.000
400
600
800
Credit
Score
Geography
France
Germany
Spain

EDA -

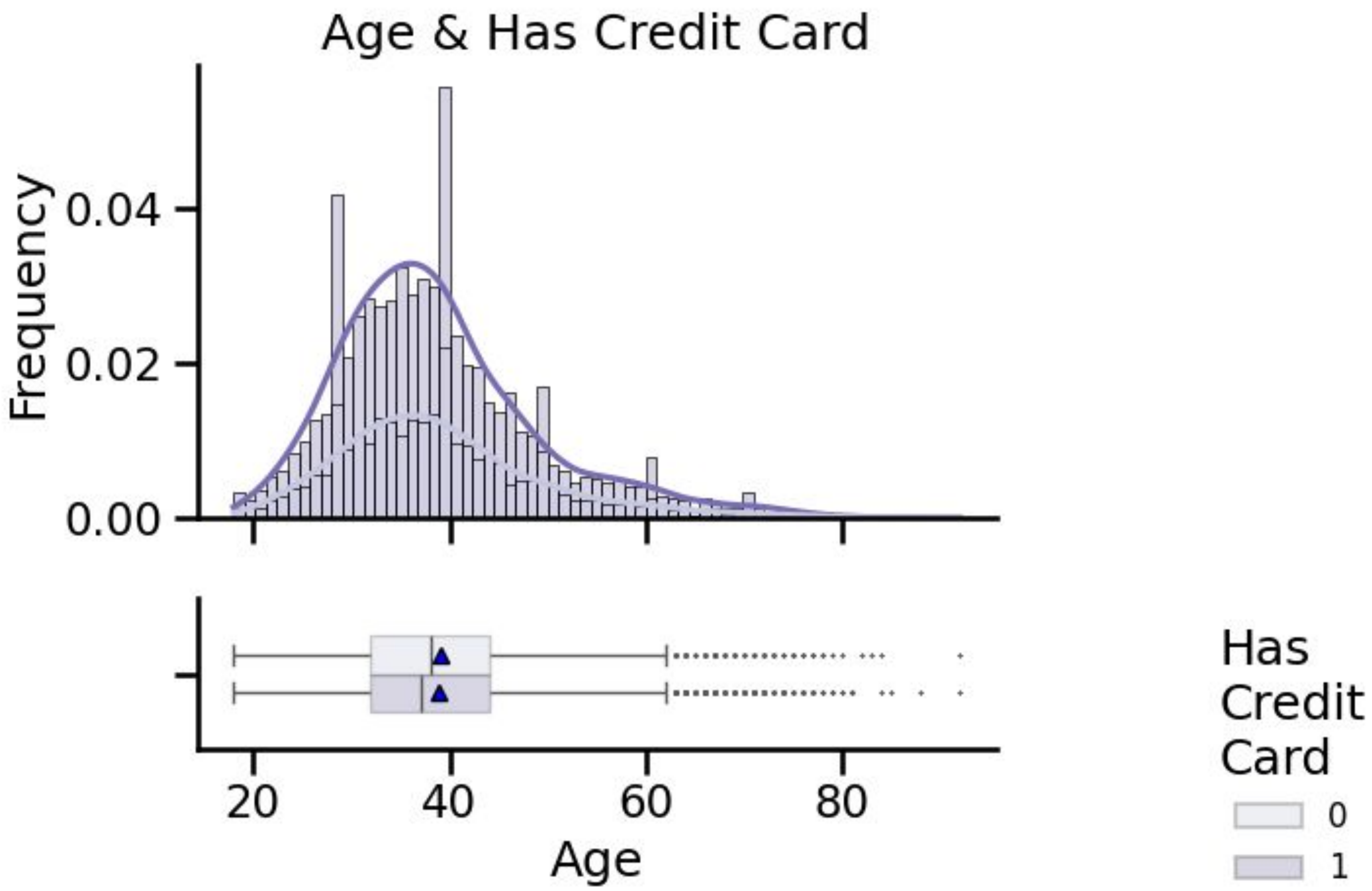
- 
Estimated Salary (K) & Geography
Frequency
0.002
0.001
0.000
0
50
100
150
200
Estimated Salary (K)
Geography
France
Germany
Spain

EDA -


Tenure & Geography
Frequency
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
Tenure
Geography
France
Germany
Spain

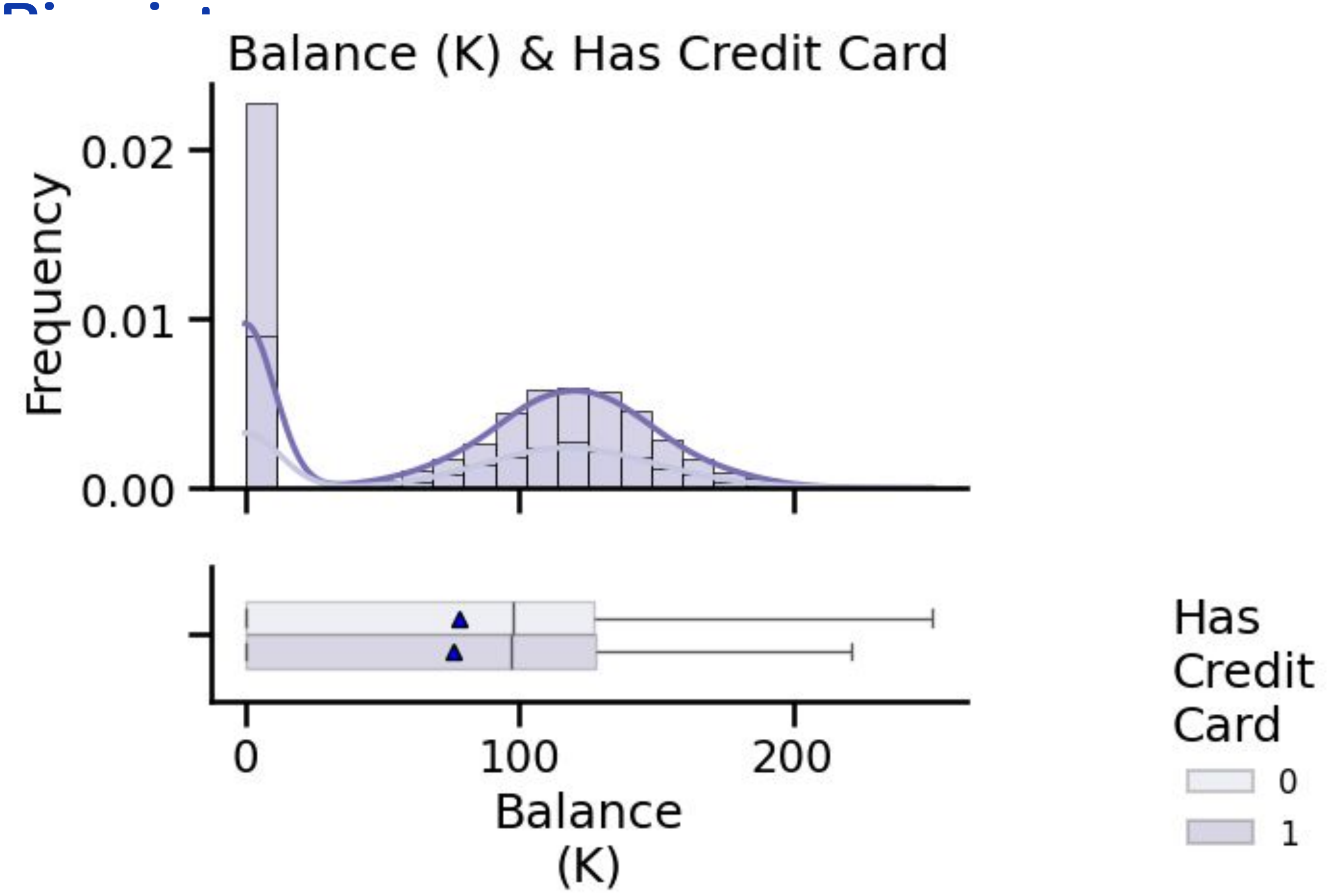
EDA -

•


Age & Has Credit Card
Frequency
0.04
0.02
0.00
20
40
60
80
Age
Has Credit Card
0
1

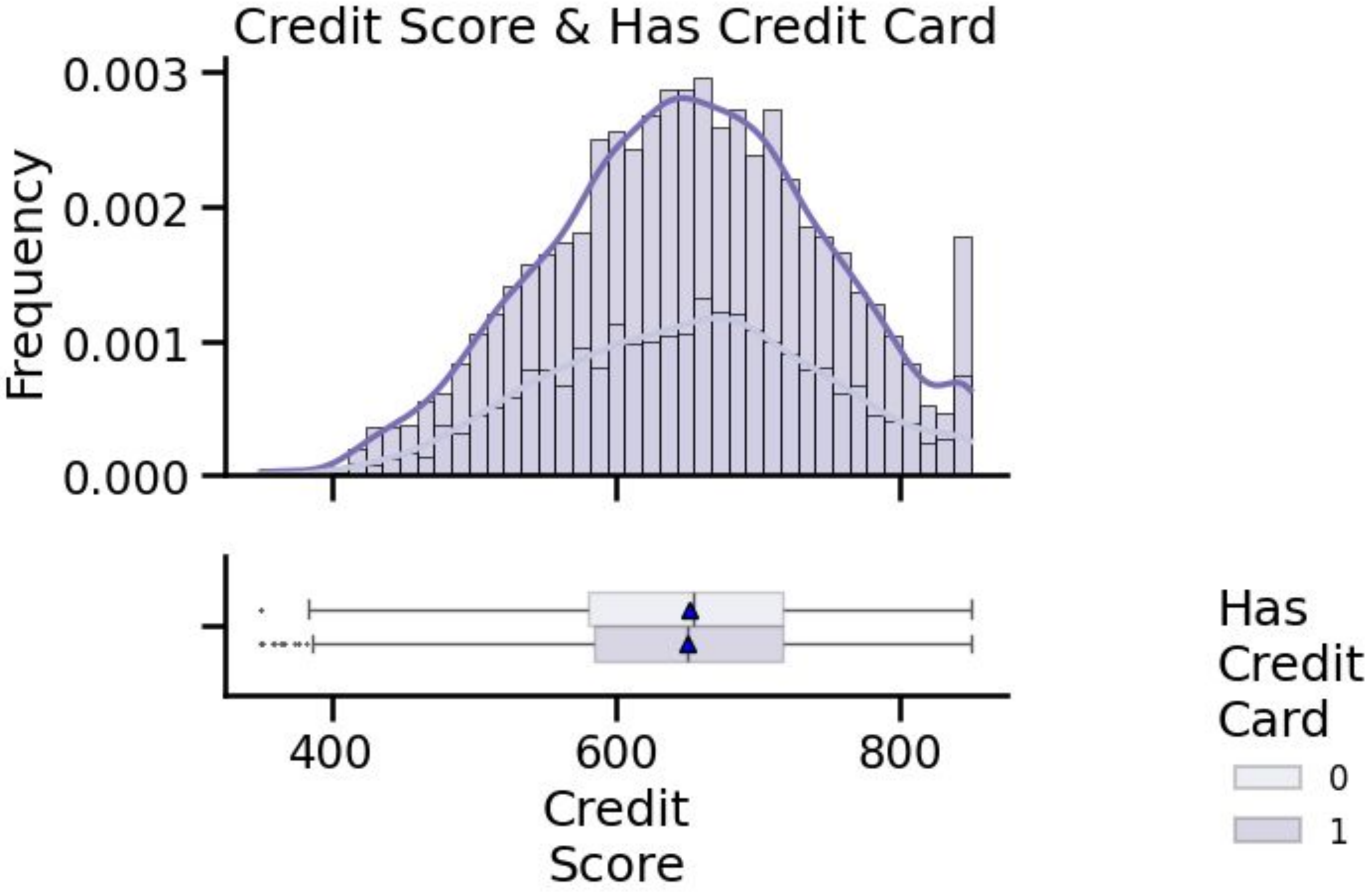
EDA - Bivariate

•



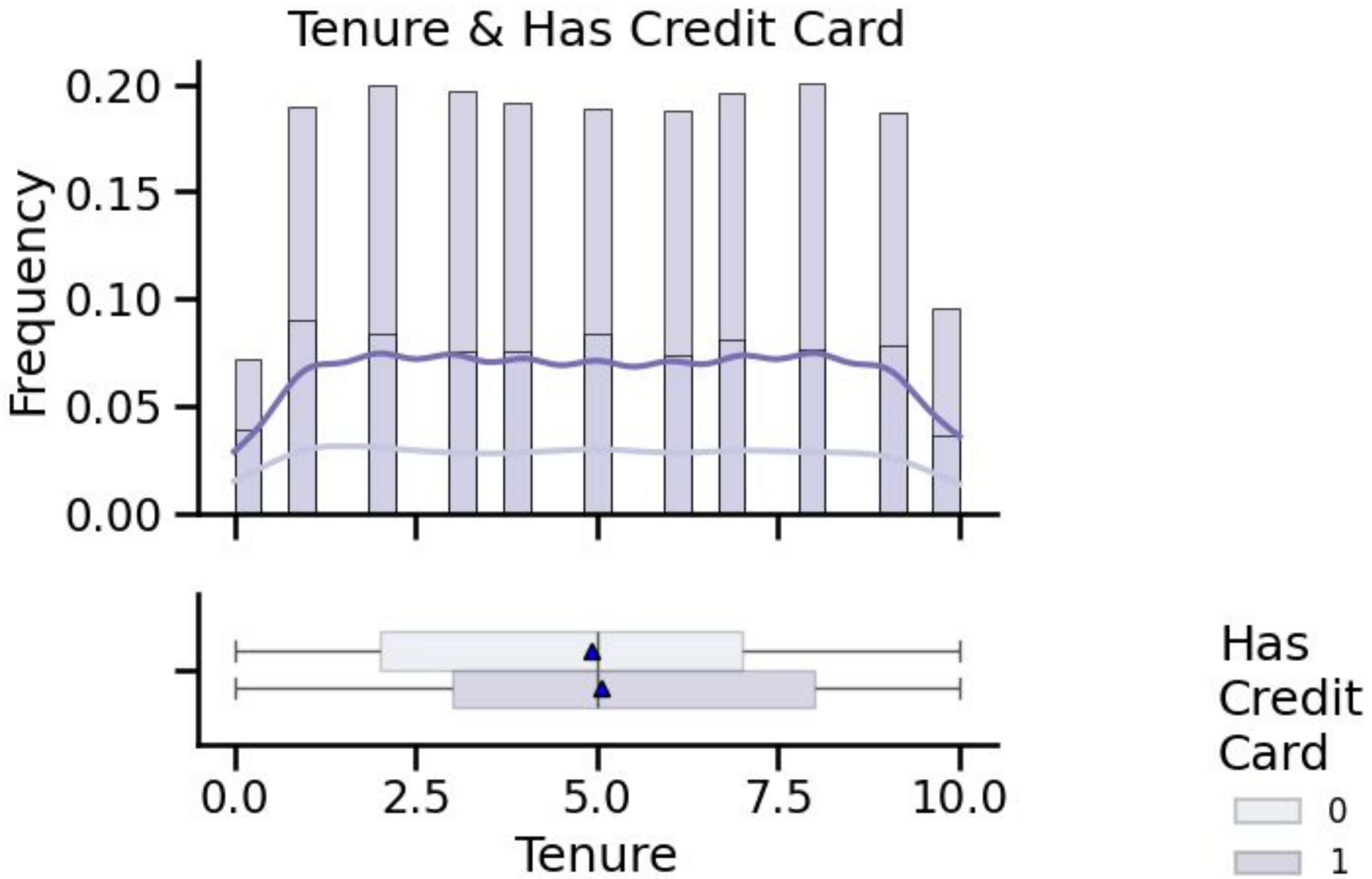
EDA - Bivariate

-



EDA -

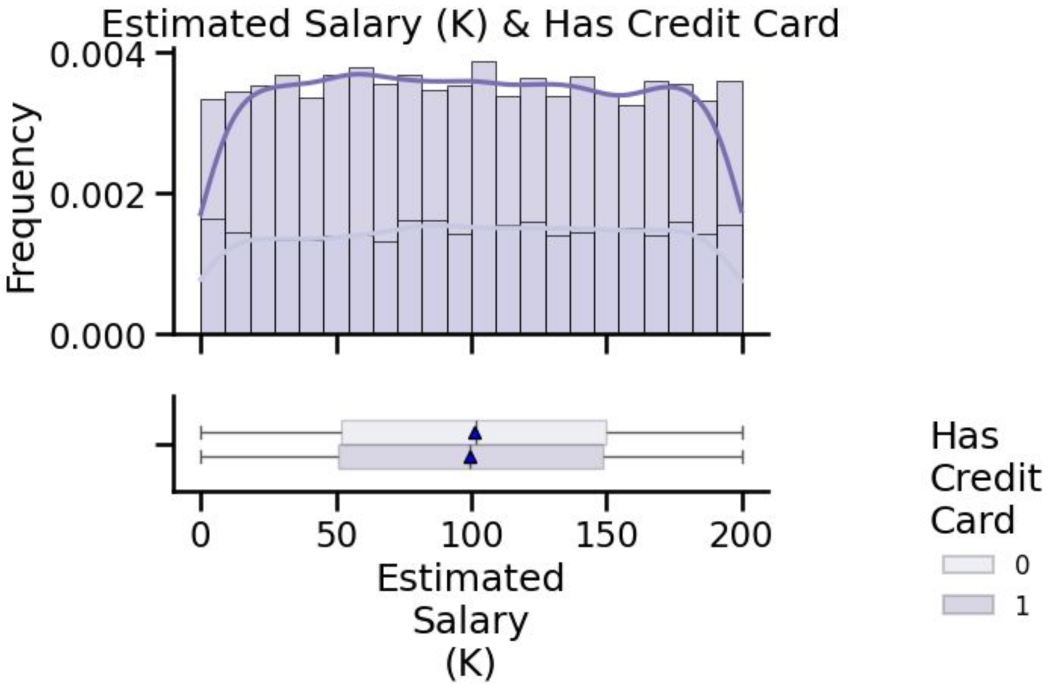
•



Tenure & Has Credit Card
Frequency
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
Tenure
Has Credit Card
0
1

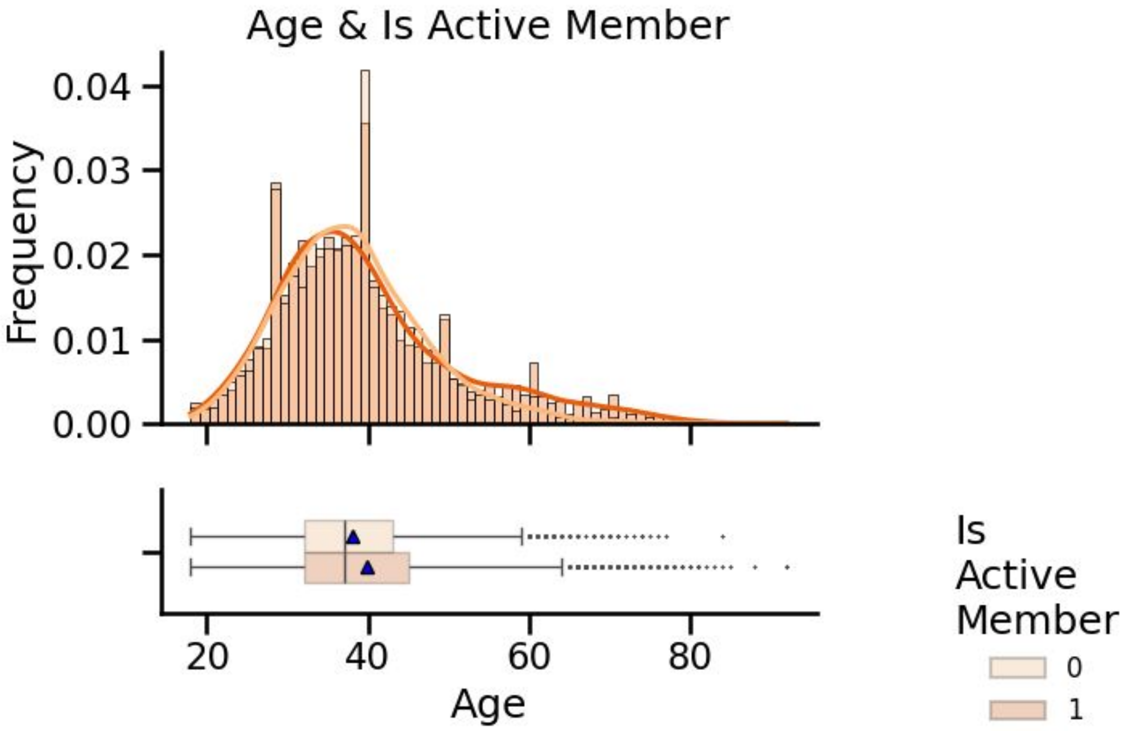
EDA - Bivariate - Salary vs Having Credit Cards

- The salary distribution of customers with credit cards is similar to that of those without. However, the average salary is slightly higher for customers who do not have credit cards.



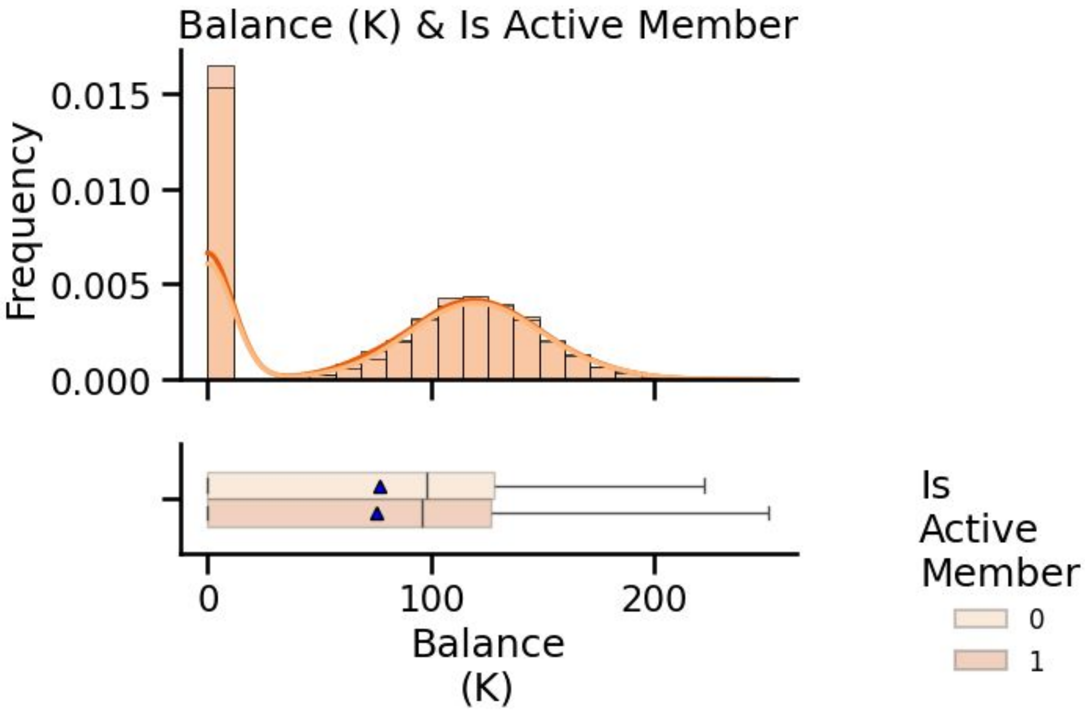
EDA - Bivariate - Age vs IsActiveMember

- The average age of active customers is higher than that of non-active members.



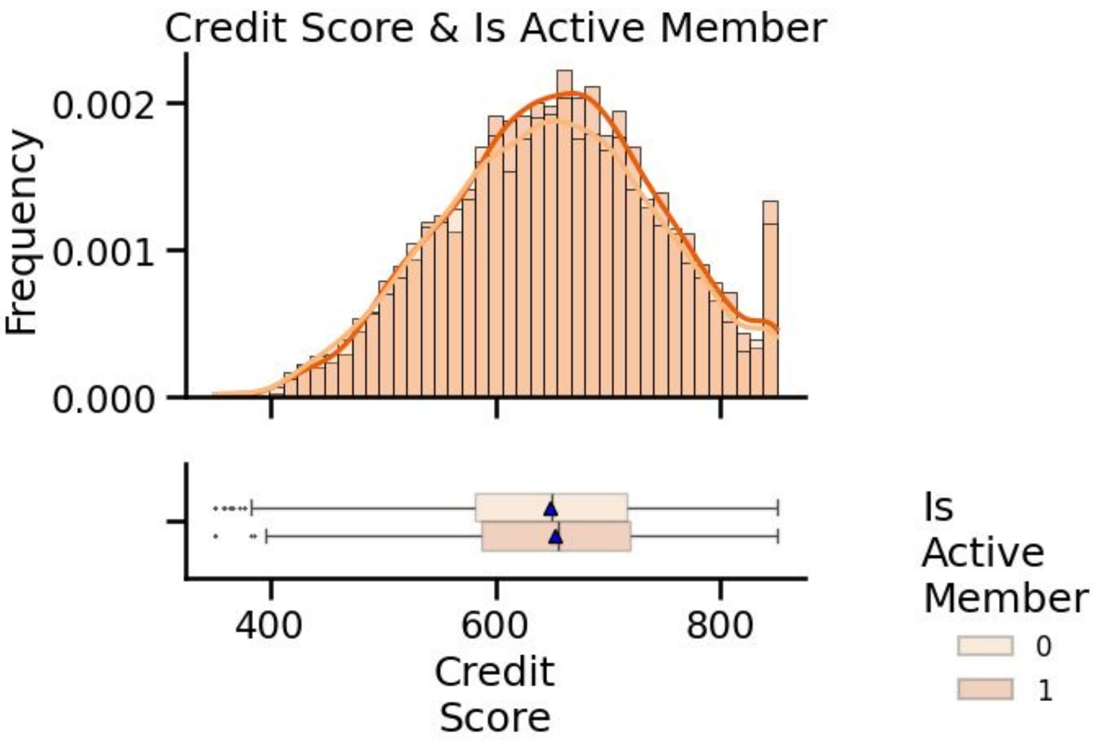
EDA - Bivariate - Balance vs IsActiveMember

- The active and non-active members seem to have similar distributions of account balance.



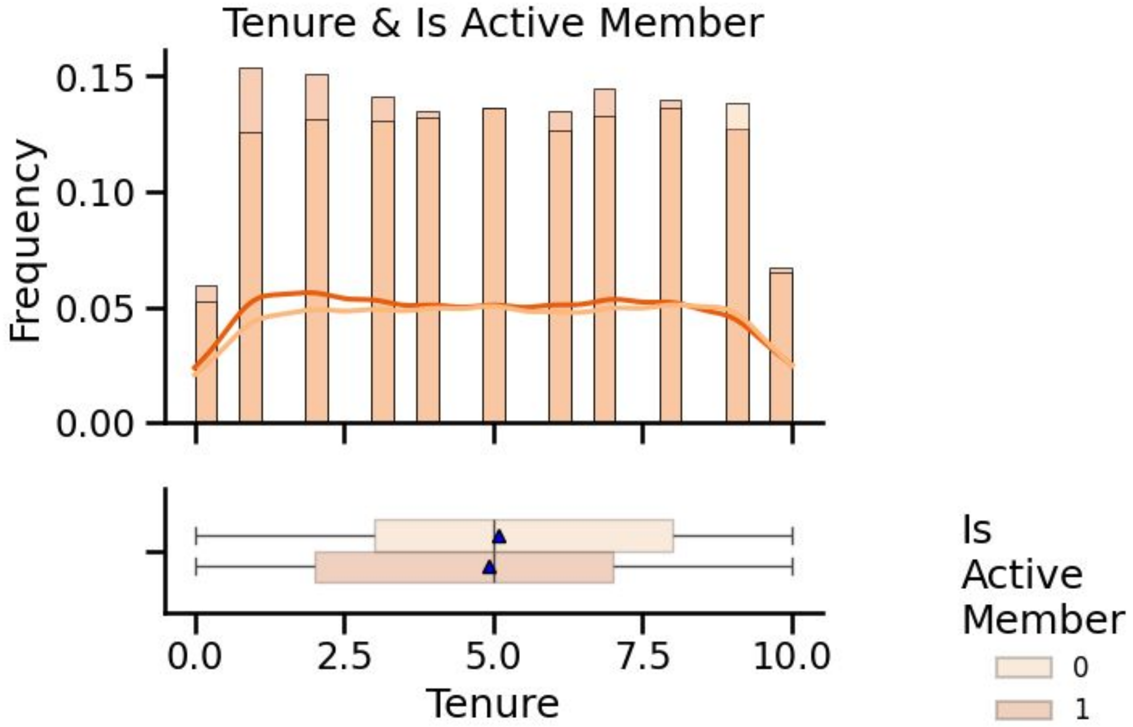
EDA - Bivariate - Credit Score vs IsActiveMember

- The non-active members seem to have slightly lower credit scores on average.



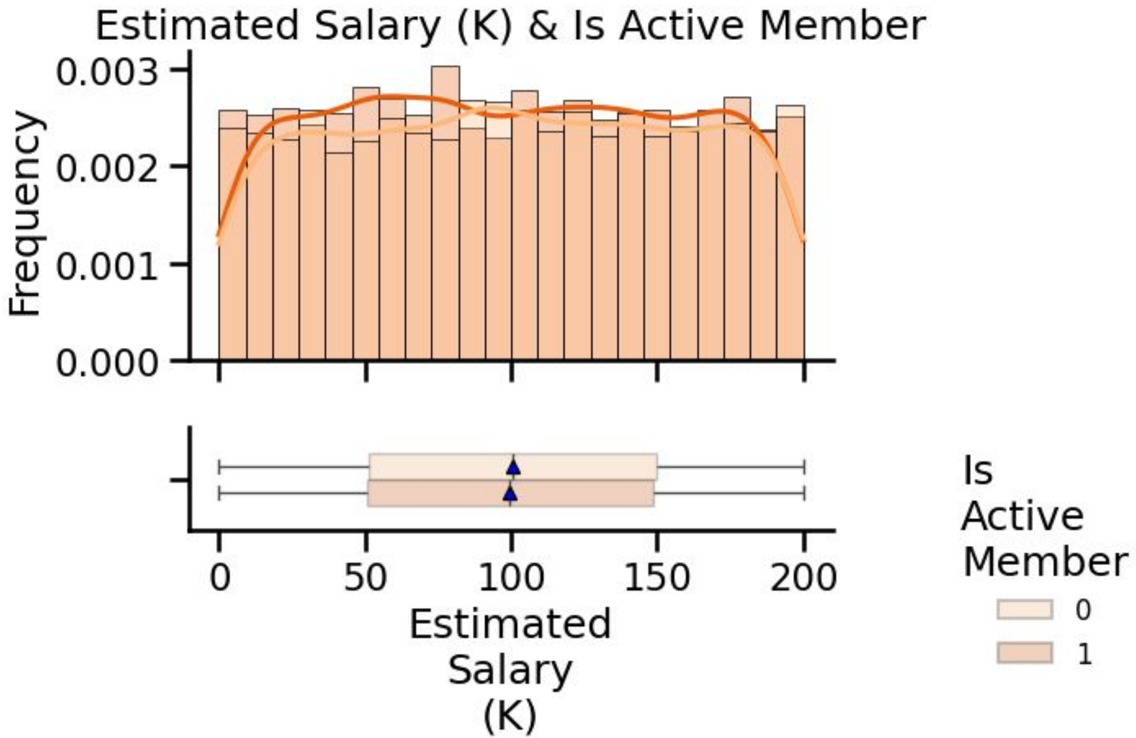
EDA - Bivariate - Tenure vs IsActiveMember

- The customers who on average have stayed longer with the bank tend to be less active members.



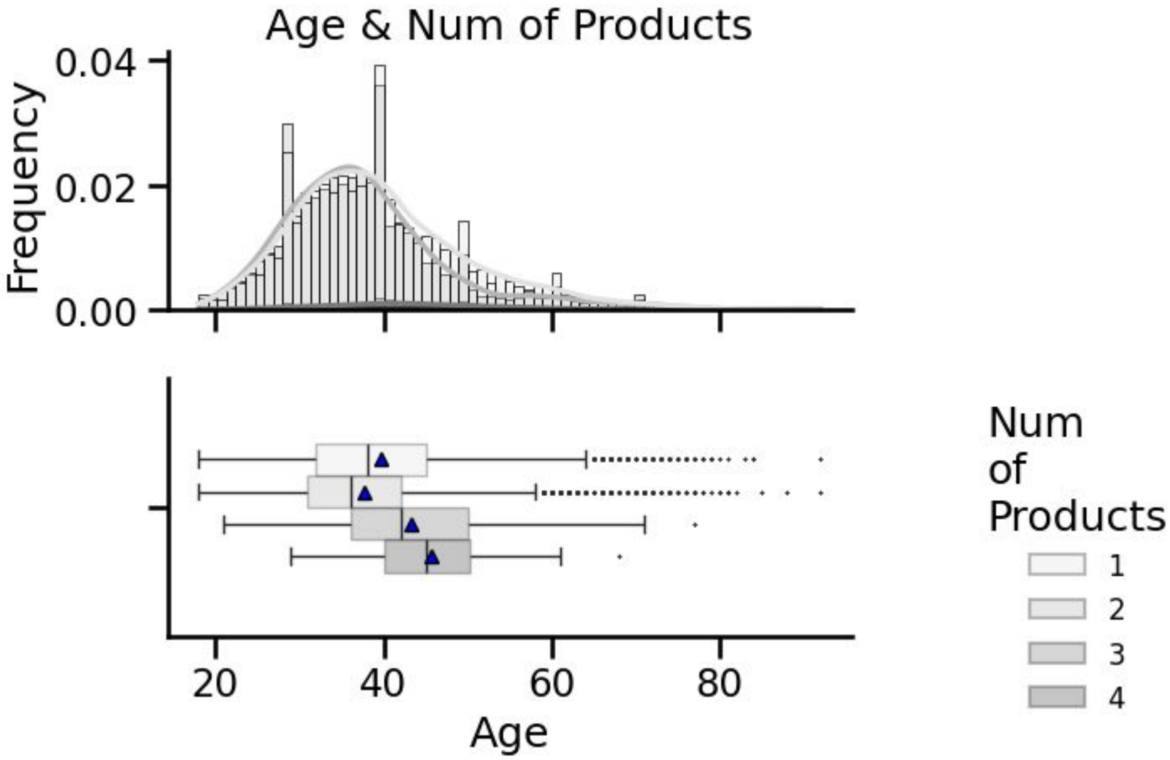
EDA - Bivariate - Salary vs IsActiveMember

- Active and on-active members tend to have similar salaries.



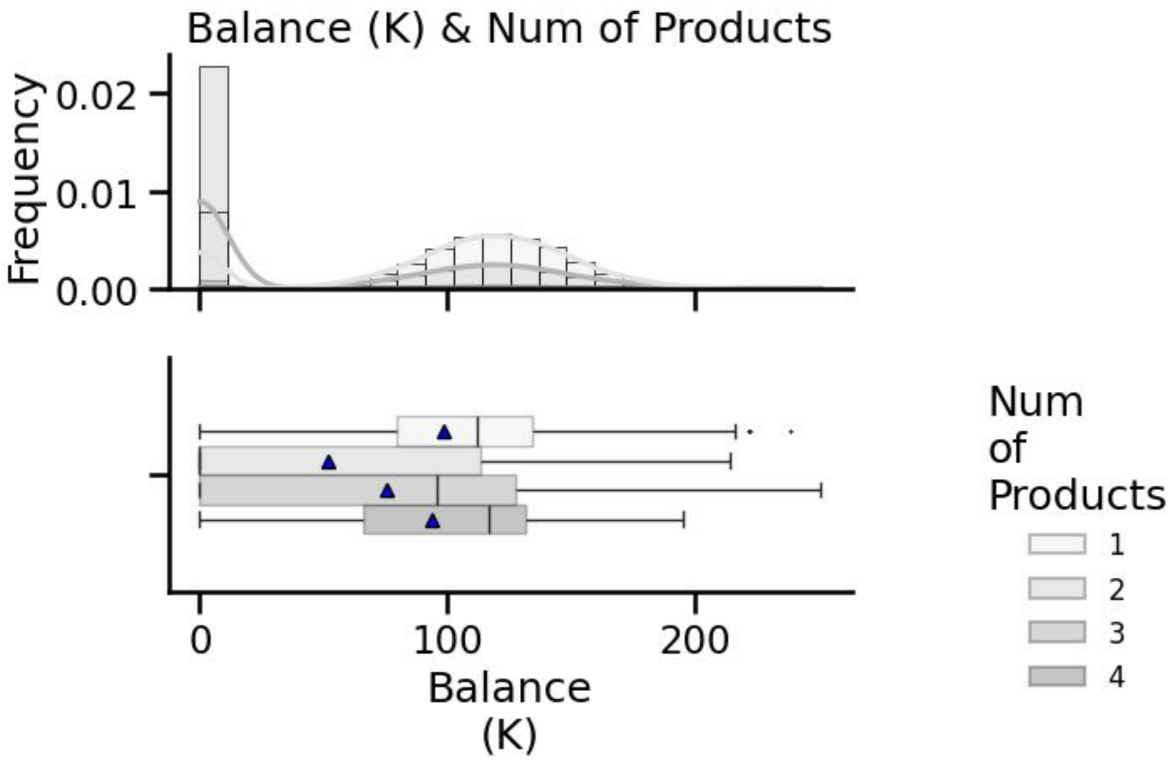
EDA - Bivariate - Age vs # of Products

- The customers who have more than two bank products are on average older.



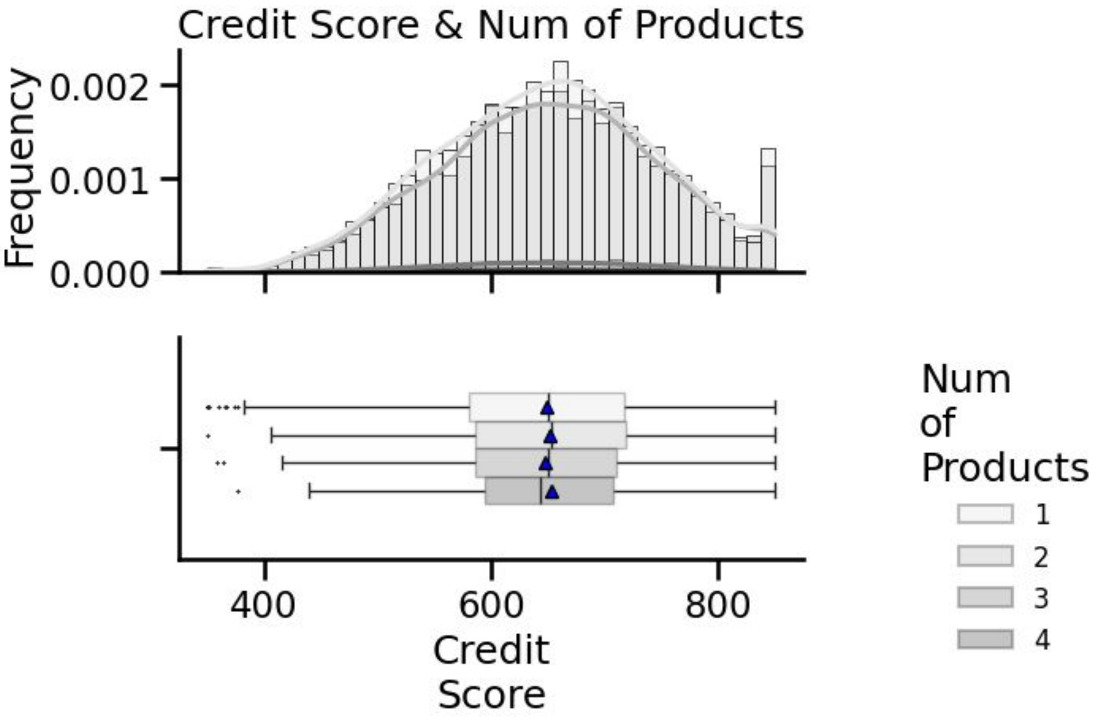
EDA - Bivariate - Account Balance vs # of Products

- The customers who have 4 bank products on average have a higher median account balance.



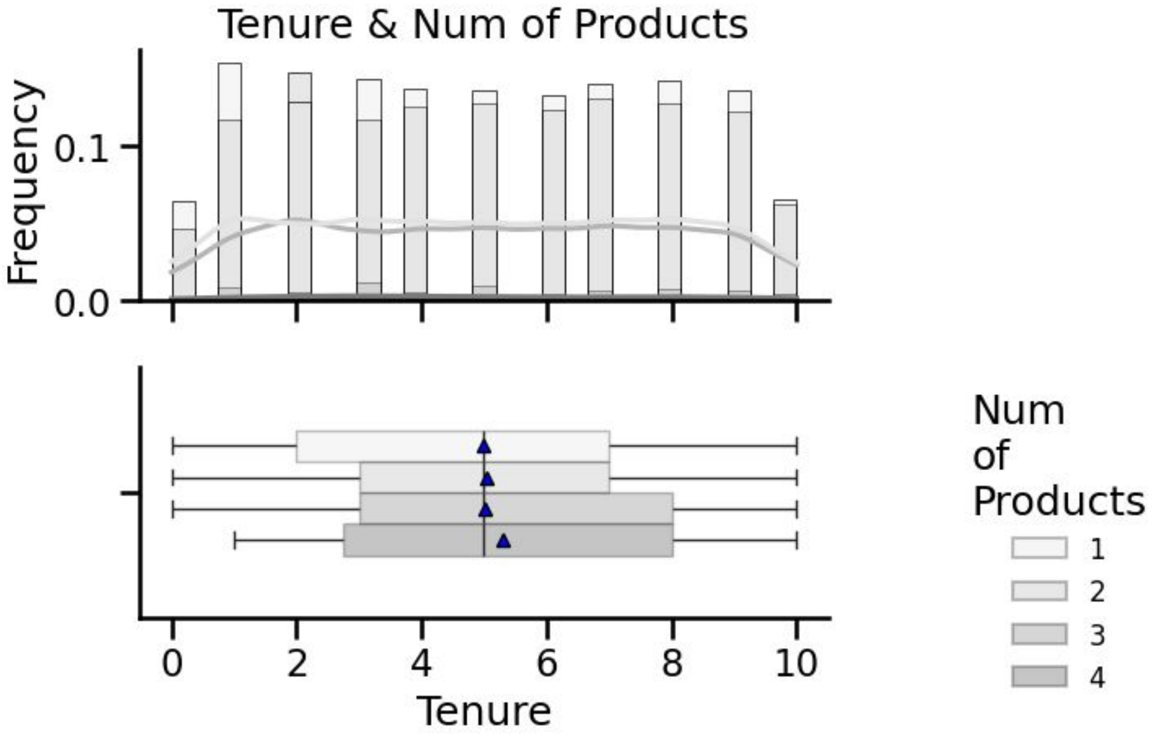
EDA - Bivariate - Credit Score vs # of Products

- The customers who have 4 bank products on average have a higher credit score.



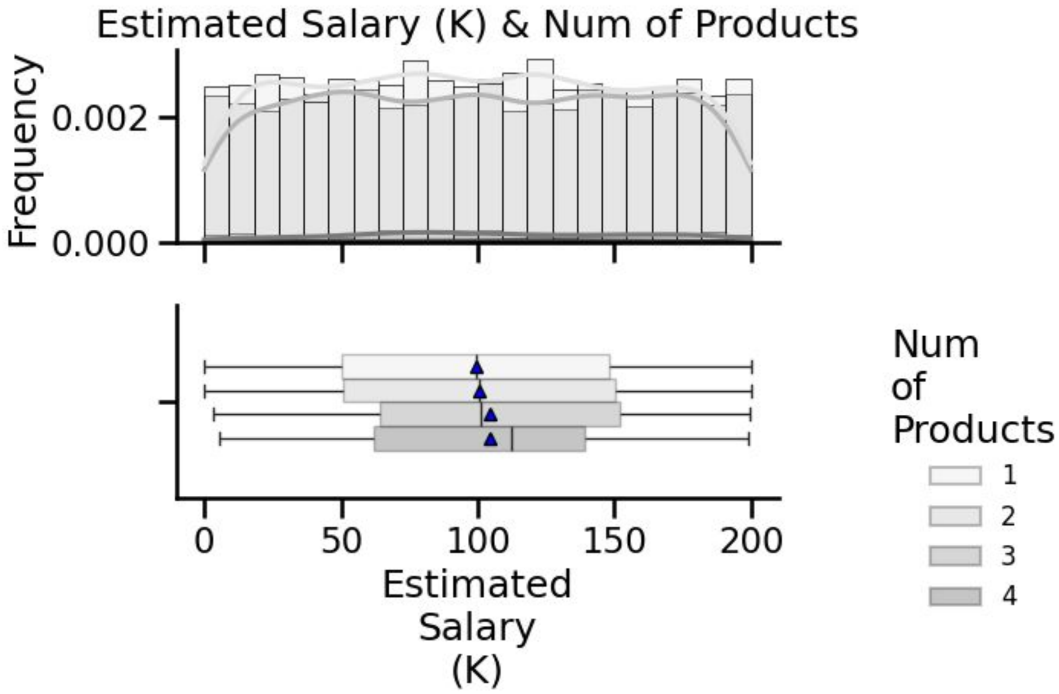
EDA - Bivariate - Tenure vs # of Products

- The customers who have 4 bank products on average have stayed longer with the bank.



EDA - Bivariate - Salary vs # of Products

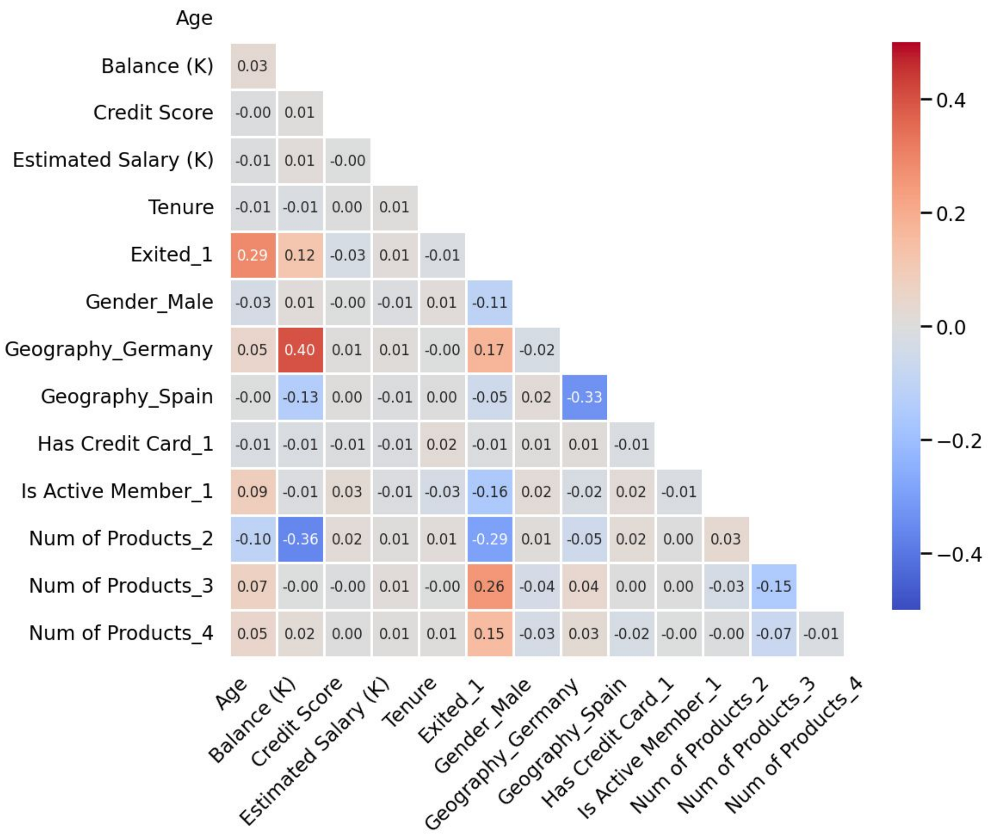
- The customers who have 4 bank products tend to have a higher salary.



EDA - Bivariate - *All the variables*

- We can convert the categorical variables into dummy variables and construct the correlation matrix for all the columns

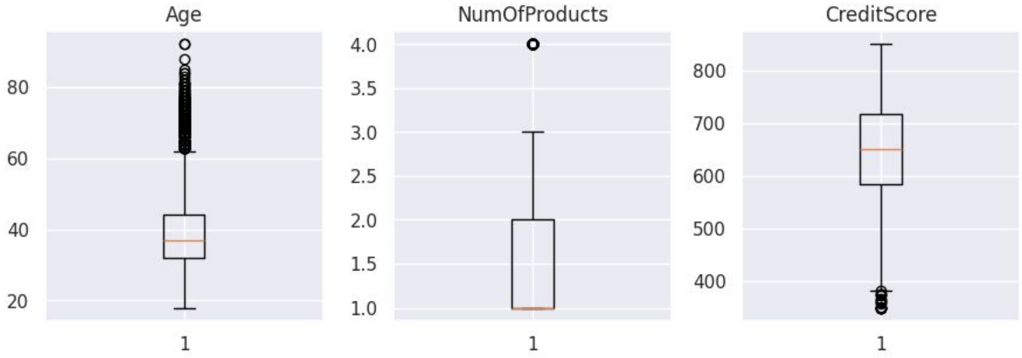
- This confirms our previous observations.



Data Preprocessing

- There are **no duplicates** in the original data.
- **Missing values**: There are **no missing values** in the original data.
- **Outliers:** The outliers are a small percentage and we won't need to treat them.
- **Feature engineering:** An optional data conversion of the dollar values by dividing them by 1000 for simpler visualization.
- The fields *Surname* and *CustomerId* are dropped before modeling.

	Outlier %
Column	
Age	3.6
NumOfProducts	0.6
CreditScore	0.2



Data Preprocessing

- Dummy variables are added for the categorical/string variables
- The numerical values are normalized using `sklearn.preprocessing.StandardScaler`
- Below we have some sample data before we start the modeling process.

Age	Balance (K)	CreditScore	EstimatedSalary (K)	HasCrCard	IsActiveMember	NumOfProducts	Tenure	Gender_Male	Geography_Germany	Geography_Spain
3.516577	0.925750	-0.556600	-1.505216	1.0	1.0	-0.914333	1.380964	0.0	0.0	1.0
0.956282	-1.219862	-0.370472	1.614655	1.0	0.0	0.797901	-1.376312	1.0	0.0	0.0
2.189016	-0.194773	-2.118010	-0.405078	1.0	1.0	-0.914333	-1.376312	1.0	0.0	1.0
0.102850	-1.219862	-1.094304	1.067959	0.0	1.0	-0.914333	0.691645	0.0	0.0	0.0
-0.181627	-1.219862	0.994469	0.756245	1.0	0.0	0.797901	-1.031652	1.0	0.0	0.0

Model Building

Train, Validation, Test split

- The data is split into **train, validation and test** sets.
- Since we have an ***imbalanced class***, we also conduct ***oversampling*** and ***undersampling***.
- Here are the dimensions of the data partitions in original/over-sampled/under-sampled data.

Under-Sampled	Rows	Columns	Class Proportions %
X Train	3,260	11	{0: 50.0, 1: 50.0}
X Validation	406	11	{0: 50.0, 1: 50.0}

Over-Sampled	Rows	Columns	Class Proportions %
X Train	12,740	11	{0: 50.0, 1: 50.0}
X Validation	1,594	11	{0: 50.0, 1: 50.0}

Original	Rows	Columns	Proportion %	Class Proportions %
X Train	8,000	11	80	{0: 79.62, 1: 20.38}
X Validation	1,000	11	10	{0: 79.7, 1: 20.3}
X Test	1,000	11	10	{0: 79.6, 1: 20.4}

Model Building

NN Modeling

We build our NN model sequentially using `keras.models.Sequential`. We try the following combinations:

- Adding **`Dense` layers** with various number of neurons.
- Adding **`Dropout` layers** to the hidden layers for guarding against the overfitting.
- We also try different **optimizers** such as SGD and Adam with various **learning rates** and **momentum**.
- We try different **activation functions** such as ‘ReLu’ and ‘tanh’.
- We compute the recall score for training and validation sets.
- We try modeling on **over-sampled** and **under-sampled** data.

Model Performance Summary

Model Evaluation Criterion

Our model can make wrong predictions in two ways:

- ***False Positive***: Predicting that a customer will not leave the bank when they will.
  - If minimized, it improves the precision.
- ***False Negative***: Predicting that a customer will stay at the bank while they would leave.
  - If minimized, it improves the recall.

In our problem, we are more interested in reducing the false negative and thus minimizing the **Recall**.

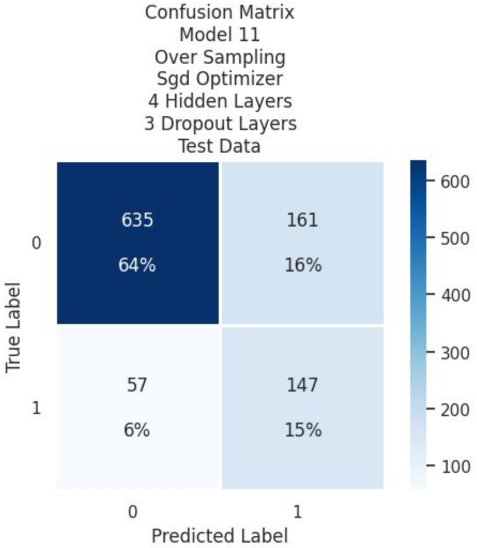
Model Performance Summary

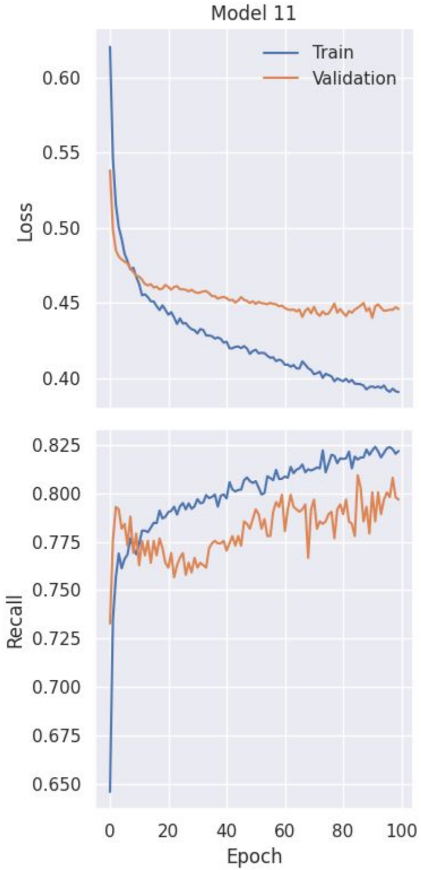
Here we have the summary information of the 14 trained models. We select Model #11 as our best model. It has the best train and valid recall scores while their difference is smaller than the other models. It takes about an hour to train this model. It has [64, 32, 16, 8] neurons in its hidden layers and 3 dense layers. We are using the Stochastic Gradient Descent with momentum .95 and learning rate of 1e-3.

	Sampling	# Neurons	Activation Function	# Dropout Layers	# Epochs	Batch Size	Optimizer	Learning Rate	Momentum	Weight Init.	Reg.	Train Loss	Validation Loss	Train Recall	Valid Recall	|Train - Valid| Recall	Time (min)
0	Original	64, 32, 1	relu, relu, sigmoid	0	120	32	SGD	0.0010	0.0	GlorotUniform		0.371278	0.379614	0.361963	0.334975	0.026988	25.33
1	Original	64, 32, 1	relu, relu, sigmoid	0	120	32	adam	0.0010	-	GlorotUniform		0.241684	0.418014	0.653988	0.487685	0.166303	36.75
2	Original	32, 16, 8, 4, 1	relu, relu, relu, relu, sigmoid	2	100	32	adam	0.0010	-	GlorotUniform	Dropout(0.2), Dropout(0.1)	0.320479	0.347207	0.504294	0.463054	0.041240	53.69
3	Over	32, 16, 8, 1	relu, relu, relu, sigmoid	0	100	32	SGD	0.0100	0.0	GlorotUniform		0.353260	0.496989	0.845683	0.638645	0.207038	31.67
4	Under	32, 16, 8, 1	relu, relu, relu, sigmoid	0	100	32	adam	0.0010	-	GlorotUniform		0.372198	0.522871	0.799387	0.704434	0.094953	14.88
5	Over	32, 16, 8, 1	relu, relu, relu, sigmoid	0	100	32	adam	0.0010	-	GlorotUniform		0.306442	0.554419	0.875824	0.742785	0.133039	61.58
6	Over	32, 16, 8, 1	relu, relu, relu, sigmoid	2	100	32	adam	0.0010	-	GlorotUniform	Dropout(0.2), Dropout(0.1)	0.390816	0.454033	0.827159	0.754078	0.073081	65.15
7	Over	32, 16, 8, 1	relu, relu, relu, sigmoid	2	100	32	adam	0.0010	-	GlorotUniform	Dropout(0.2), Dropout(0.1)	0.390816	0.454033	0.827159	0.754078	0.073081	64.92
8	Over	32, 16, 8, 1	relu, tanh, tanh, sigmoid	2	100	32	adam	0.0010	-	GlorotUniform	Dropout(0.2), Dropout(0.1)	0.391210	0.458709	0.822920	0.772898	0.050022	63.58
9	Over	32, 16, 8, 1	tanh, tanh, tanh, sigmoid	2	150	32	adam	0.0010	-	GlorotUniform	Dropout(0.2), Dropout(0.1)	0.416945	0.457552	0.797488	0.767880	0.029609	91.36
10	Over	64, 32, 16, 8, 1	relu, tanh, tanh, tanh, sigmoid	3	100	32	adam	0.0010	-	GlorotUniform	Dropout(0.2), Dropout(0.1), Dropout(0.1)	0.348860	0.499543	0.861538	0.759097	0.102442	68.68
11	Over	64, 32, 16, 8, 1	relu, tanh, tanh, tanh, sigmoid	3	100	32	SGD	0.0010	0.95	GlorotUniform	Dropout(0.2), Dropout(0.1), Dropout(0.1)	0.390957	0.446063	0.821821	0.796738	0.025083	56.70
12	Over	64, 32, 16, 8, 1	relu, tanh, tanh, tanh, sigmoid	3	100	32	SGD	0.0010	0.9	GlorotUniform	Dropout(0.2), Dropout(0.1), Dropout(0.1)	0.415607	0.446772	0.803768	0.775408	0.028360	61.50
13	Over	64, 32, 16, 8, 1	relu, tanh, tanh, tanh, sigmoid	3	100	32	SGD	0.0001	0.95	GlorotUniform	Dropout(0.3), Dropout(0.2), Dropout(0.1)	0.476278	0.474513	0.770330	0.761606	0.008724	62.40

Model Performance Summary

- For the selected model, we are plotting the loss and recall scores against the number of epochs.
- Here we have the prediction results on the test data in the confusion matrix and also tabulated.
- The recall score on the test data is around .76.





	Accuracy	Recall	Precision	F1
Train	0.841	0.841	0.841	0.841
Validation	0.792	0.792	0.792	0.792
Test	0.782	0.759	0.697	0.714

APPENDIX

Great Learning

Happy Learning !