BANK CHURN

LI WU

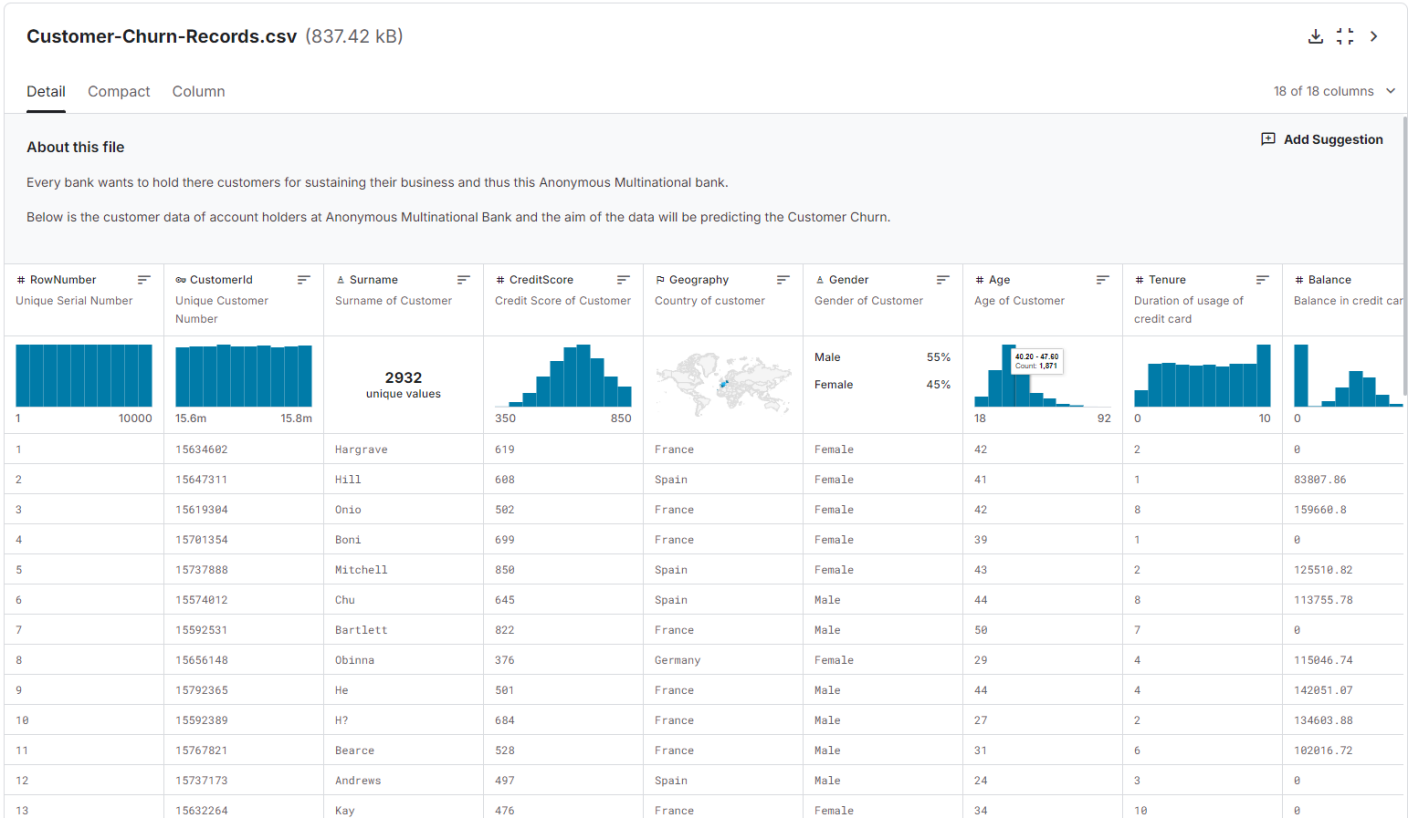
OCT. 23, 2024

AGENDA

- DATASET INTRODUCTION
- UNIVARIATE ANALYSIS
- BIVARIATE ANALYSIS
- CLASSIFICATION MODELS
- REGRESSION MODELS

DATASET INTRODUCTION

- Kaggle: Customer data from European multinational bank (1 year)
- Objectives: predicting the customer churn
- Provide strategies and recommendations for bank marketing department



Customer-Churn-Records.csv (837.42 kB)

Detail | Compact | Column — 18 of 18 columns

About this file — Add Suggestion

Every bank wants to hold there customers for sustaining their business and thus this Anonymous Multinational bank.

Below is the customer data of account holders at Anonymous Multinational Bank and the aim of the data will be predicting the Customer Churn.

RowNumber Unique Serial Number	CustomerId Unique Customer Number	Surname Surname of Customer	CreditScore Credit Score of Customer	Geography Country of customer	Gender Gender of Customer	Age Age of Customer	Tenure Duration of usage of credit card	Balance Balance in credit car
1 — 10000	15.6m — 15.8m	2932 unique values	350 — 850		Male 55% / Female 45%	18 — 92	0 — 10	0
1	15634602	Hargrave	619	France	Female	42	2	0
2	15647311	Hill	608	Spain	Female	41	1	83807.86
3	15619304	Onio	502	France	Female	42	8	159660.8
4	15701354	Boni	699	France	Female	39	1	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82
6	15574012	Chu	645	Spain	Male	44	8	113755.78
7	15592531	Bartlett	822	France	Male	50	7	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74
9	15792365	He	501	France	Male	44	4	142051.07
10	15592389	H?	684	France	Male	27	2	134603.88
11	15767821	Bearce	528	France	Male	31	6	102016.72
12	15737173	Andrews	497	Spain	Male	24	3	0
13	15632264	Kay	476	France	Female	34	10	0

DATASET INTRODUCTION

- 10,000 Observations

- Target Variable
  - Categorial : “Exited”
  - Continuous: “Balance”

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Geography          10000 non-null  object
 1   Gender             10000 non-null  object
 2   Age                10000 non-null  int64
 3   EstimatedSalary    10000 non-null  float64
 4   Balance            10000 non-null  float64
 5   CardType           10000 non-null  object
 6   Tenure             10000 non-null  int64
 7   CreditScore        10000 non-null  int64
 8   HasCrCard          10000 non-null  int64
 9   PointEarned        10000 non-null  int64
 10  NumOfProducts      10000 non-null  int64
 11  IsActiveMember     10000 non-null  int64
 12  SatisfactionScore  10000 non-null  int64
 13  Complain           10000 non-null  int64
 14  Exited             10000 non-null  int64
dtypes: float64(2), int64(10), object(3)
memory usage: 1.1+ MB
```

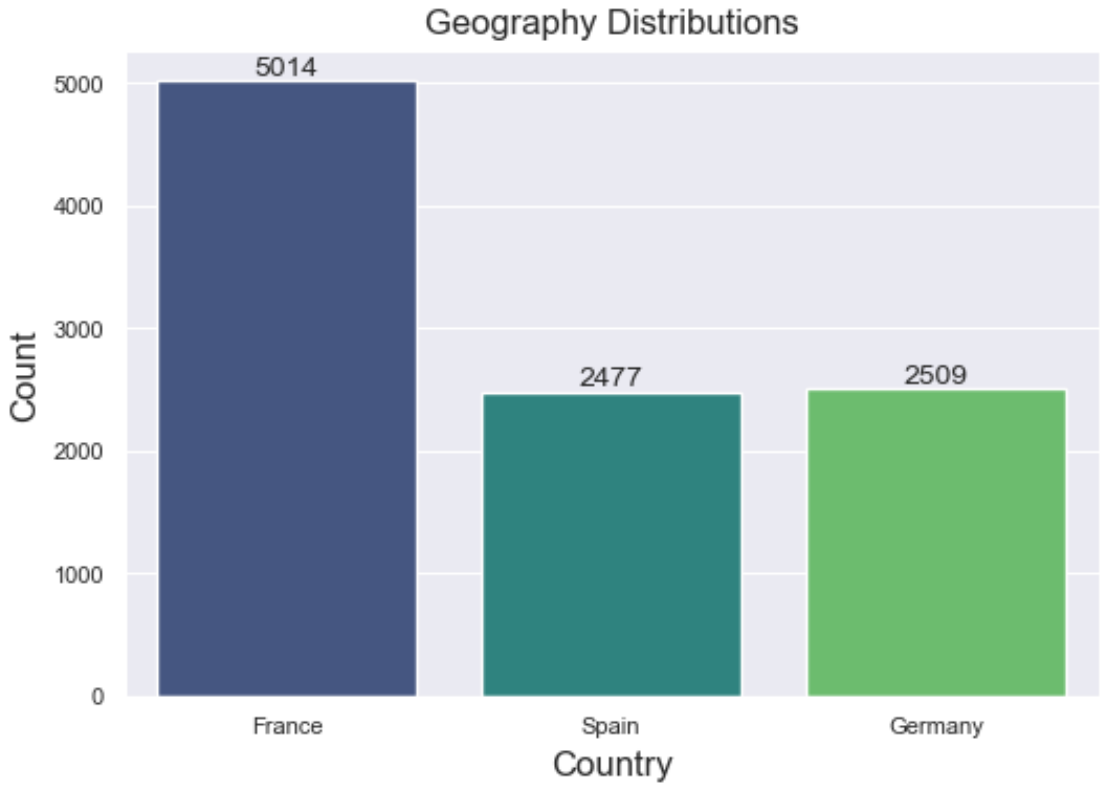
EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

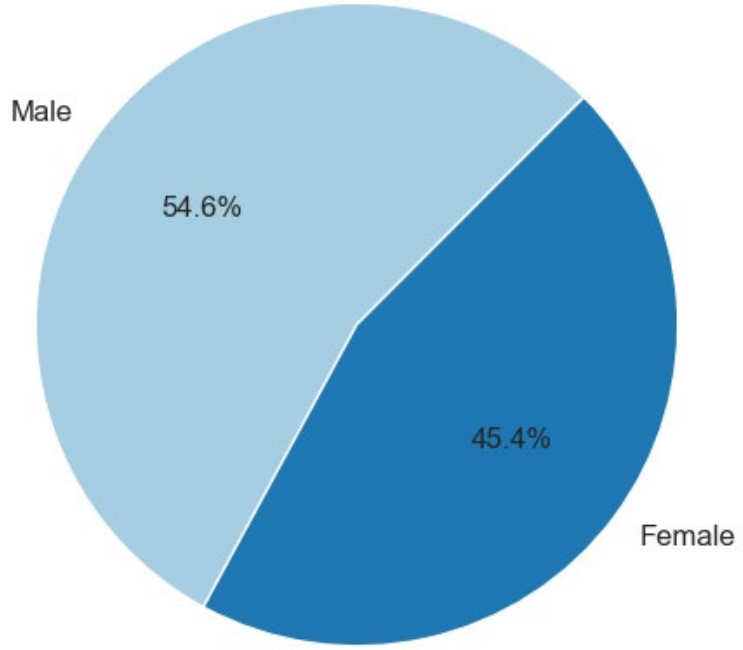
CATEGORICAL / NUMERICAL VARIABLES

CATEGORICAL VARIABLES

- Geography

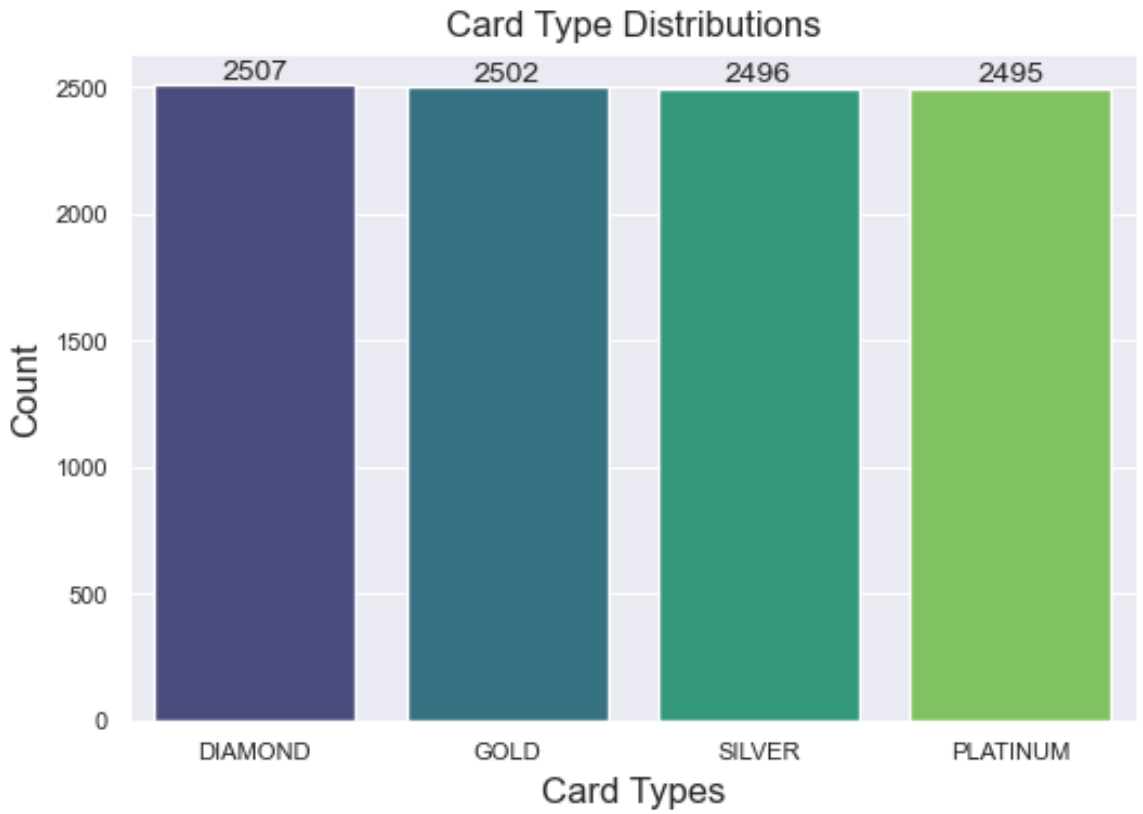


- Gender

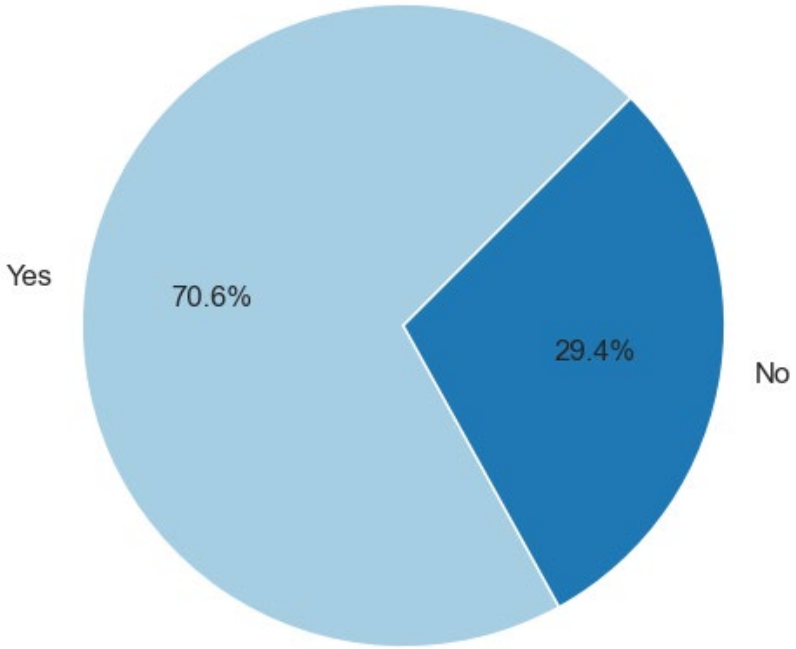


CATEGORICAL VARIABLES

- CardType



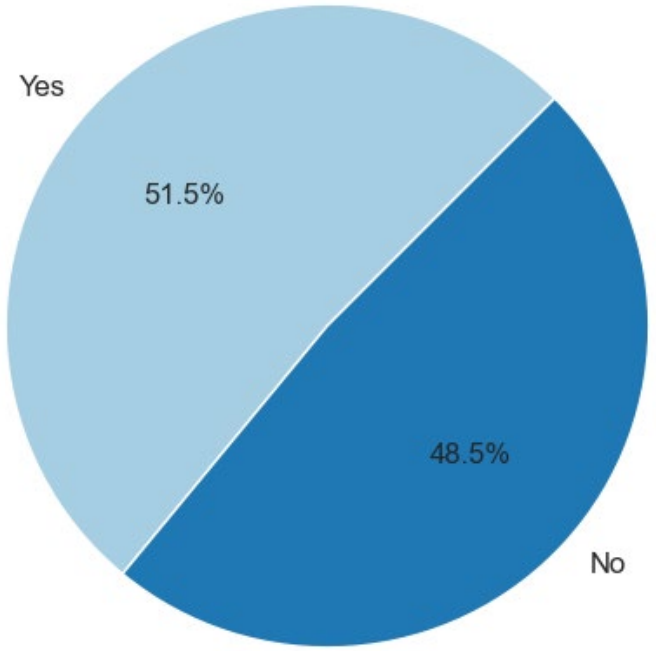
- HasCrCard



CATEGORICAL VARIABLES

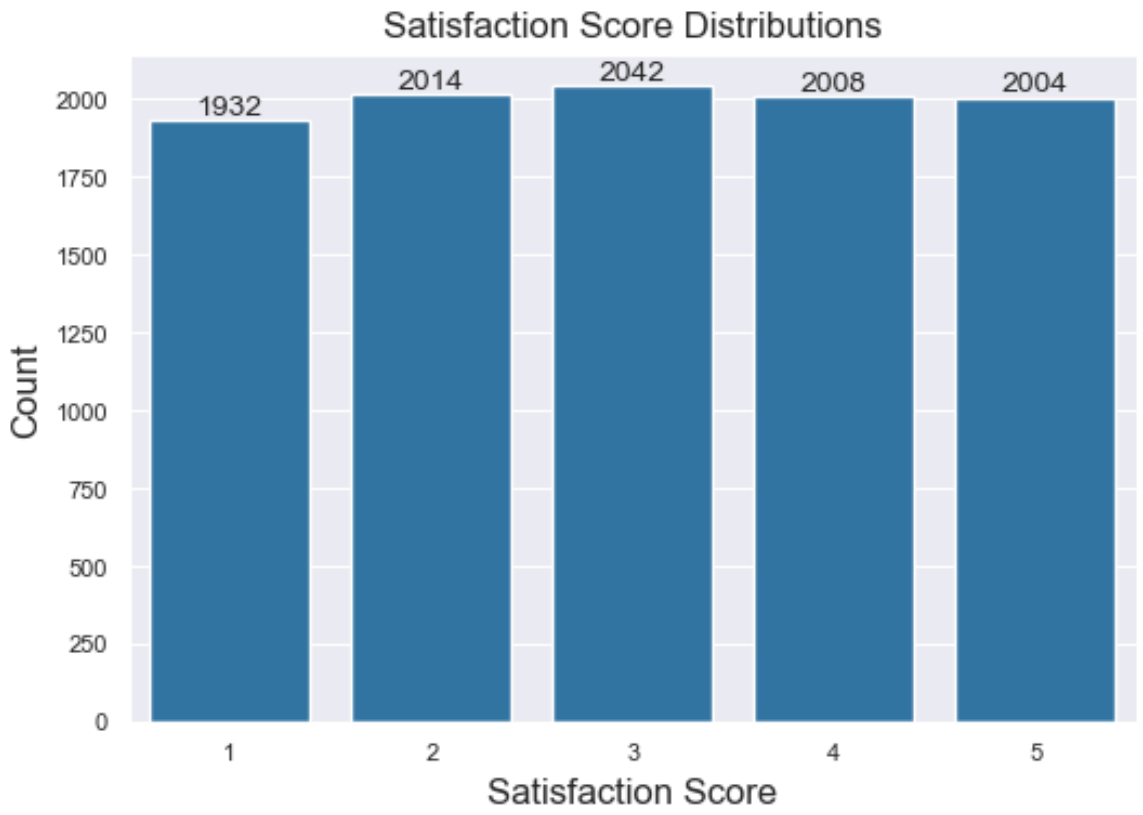
- NumOfProducts

- IsActiveMember

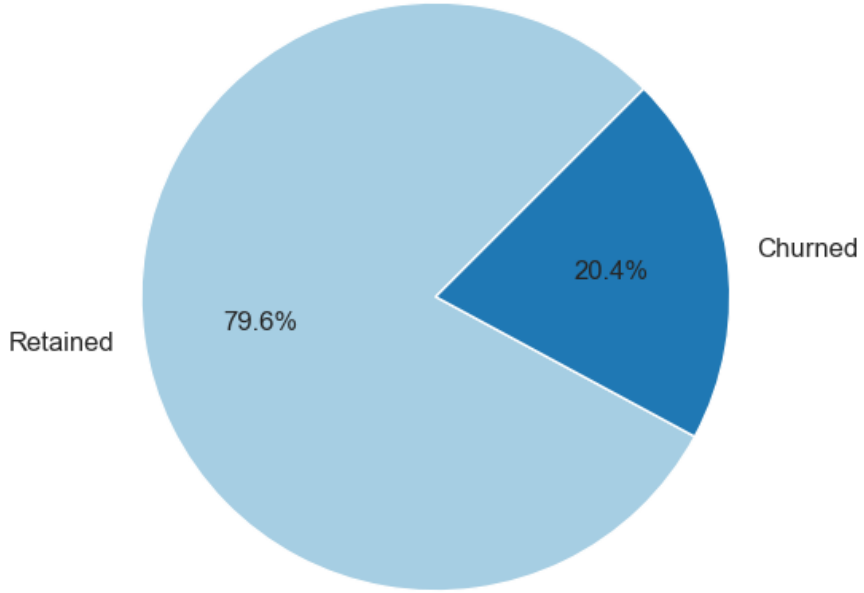


CATEGORICAL VARIABLES

- SatisfactionScore



- **Exited *(Target Variable)***

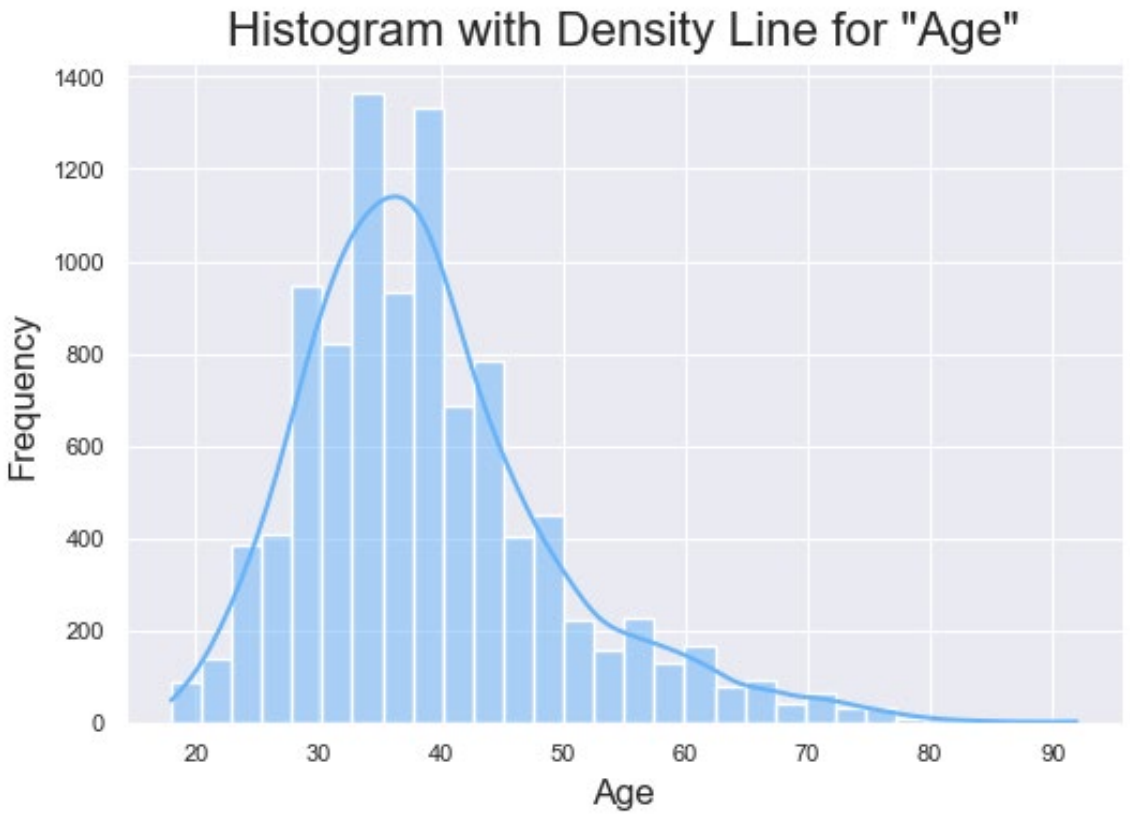


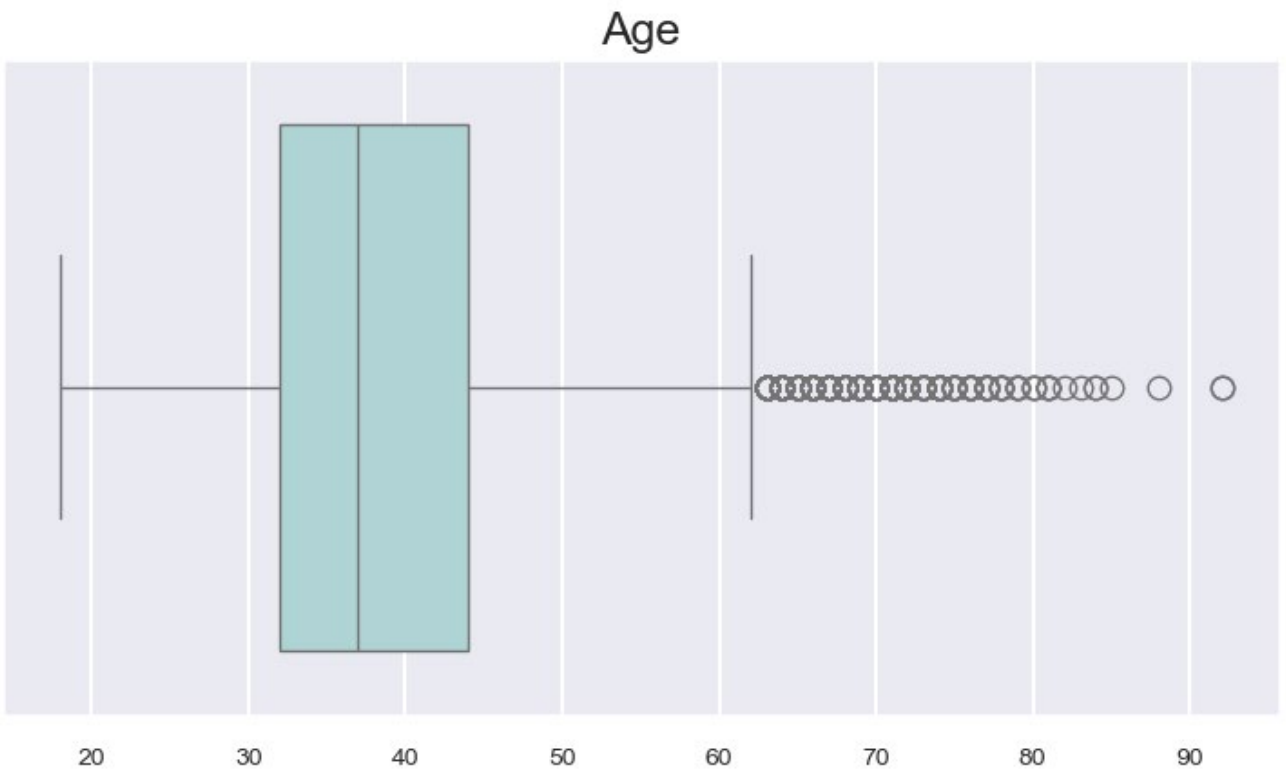
NUMERICAL VARIABLES

	Age	EstimatedSalary	Balance	Tenure	CreditScore	PointEarned
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	38.921800	100090.239881	76485.889288	5.012800	650.528800	606.515100
std	10.487806	57510.492818	62397.405202	2.892174	96.653299	225.924839
min	18.000000	11.580000	0.000000	0.000000	350.000000	119.000000
25%	32.000000	51002.110000	0.000000	3.000000	584.000000	410.000000
50%	37.000000	100193.915000	97198.540000	5.000000	652.000000	605.000000
75%	44.000000	149388.247500	127644.240000	7.000000	718.000000	801.000000
max	92.000000	199992.480000	250898.090000	10.000000	850.000000	1000.000000

NUMERICAL VARIABLES

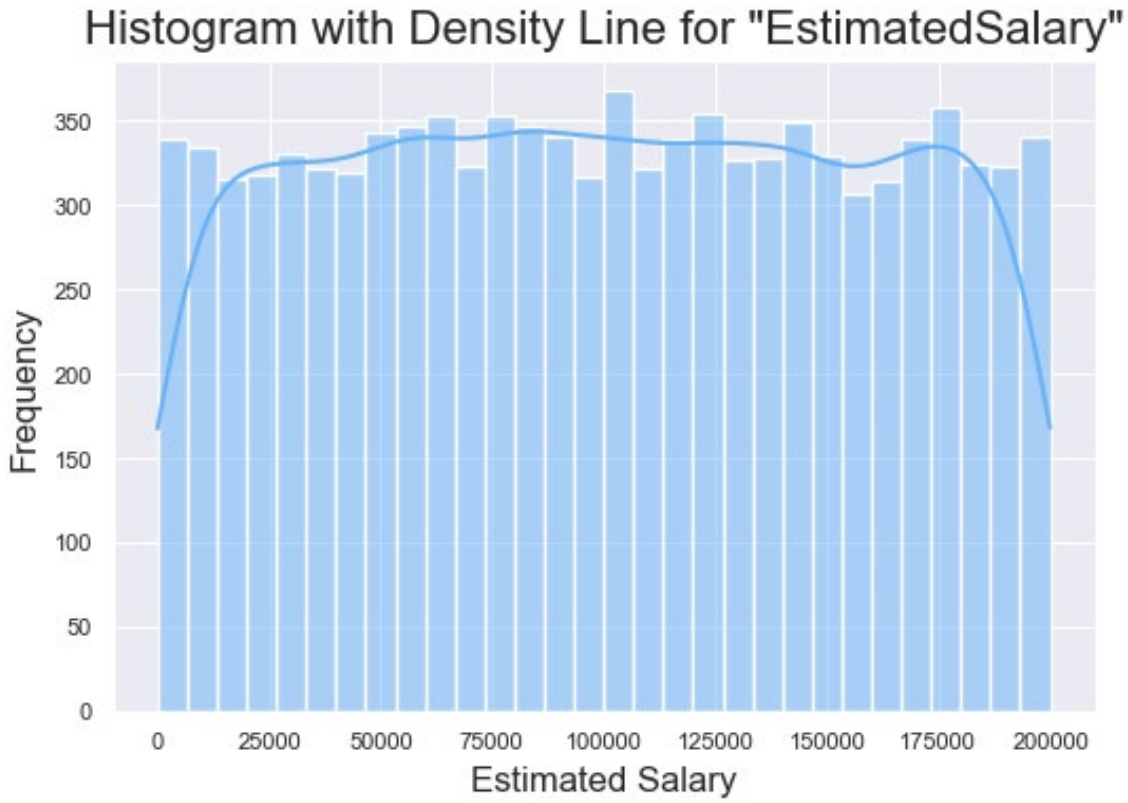
- Age

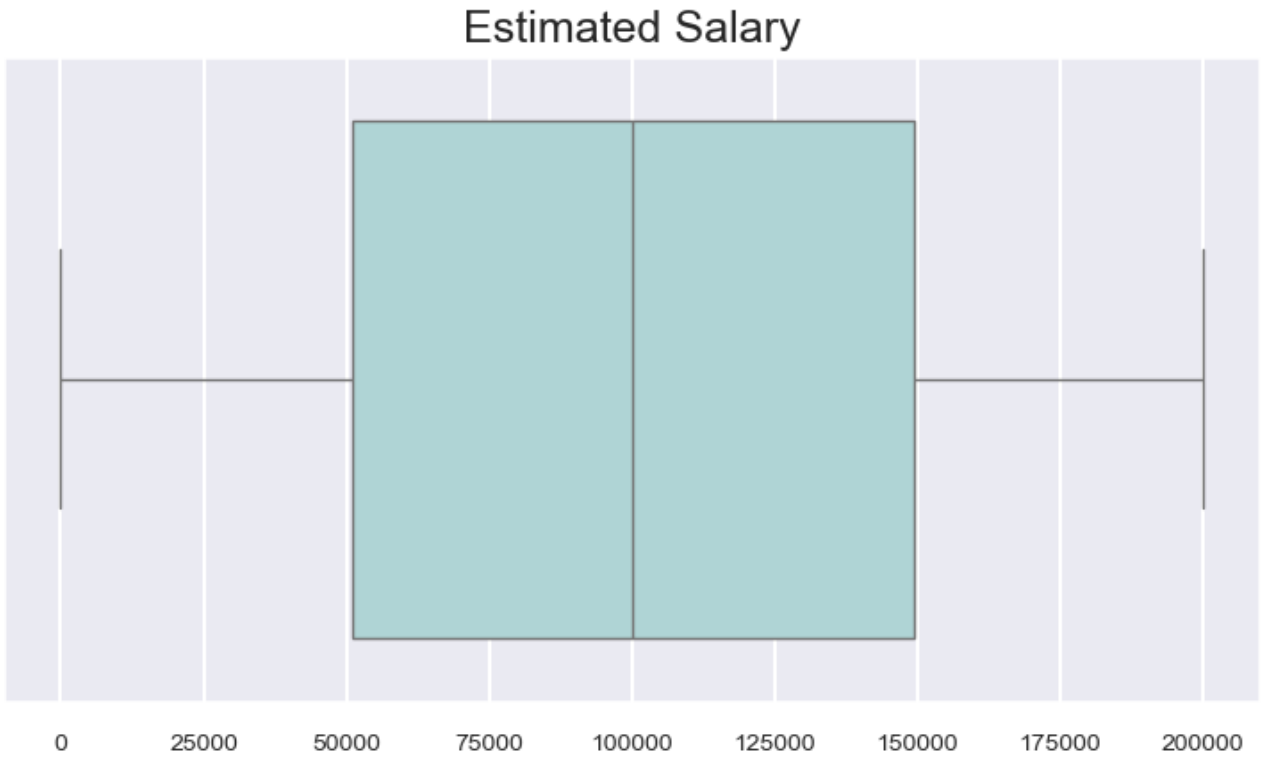




NUMERICAL VARIABLES

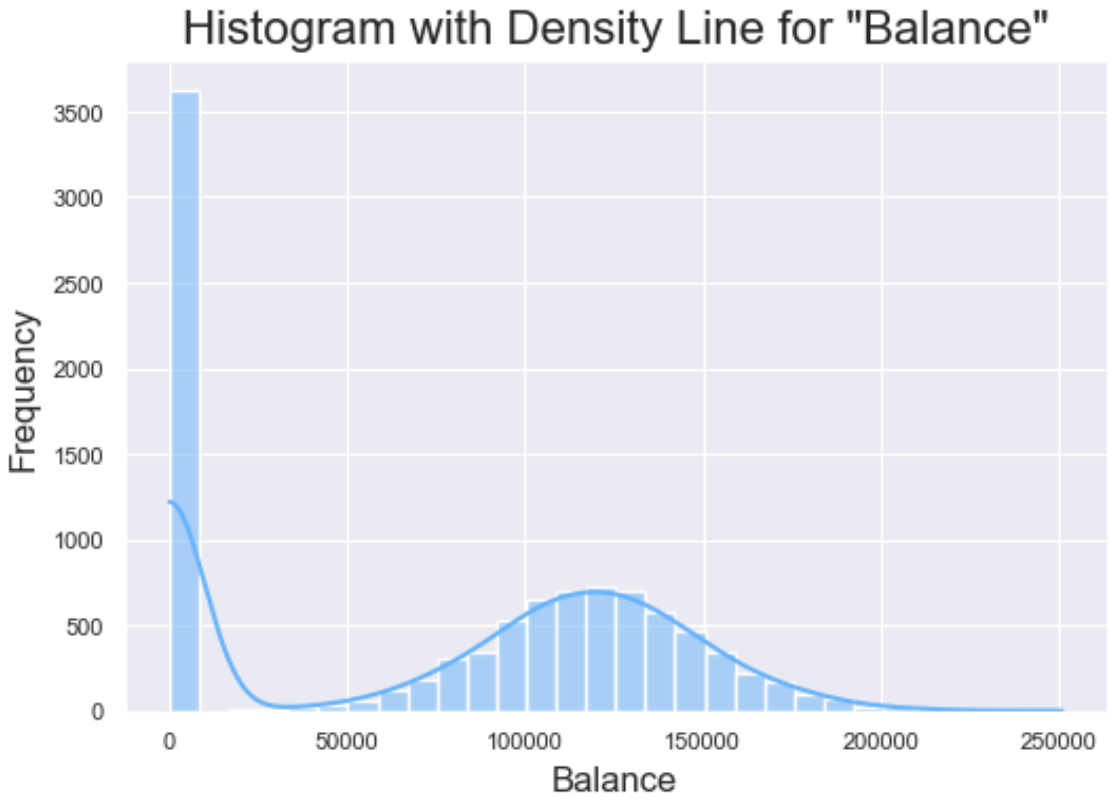
- EstimatedSalary

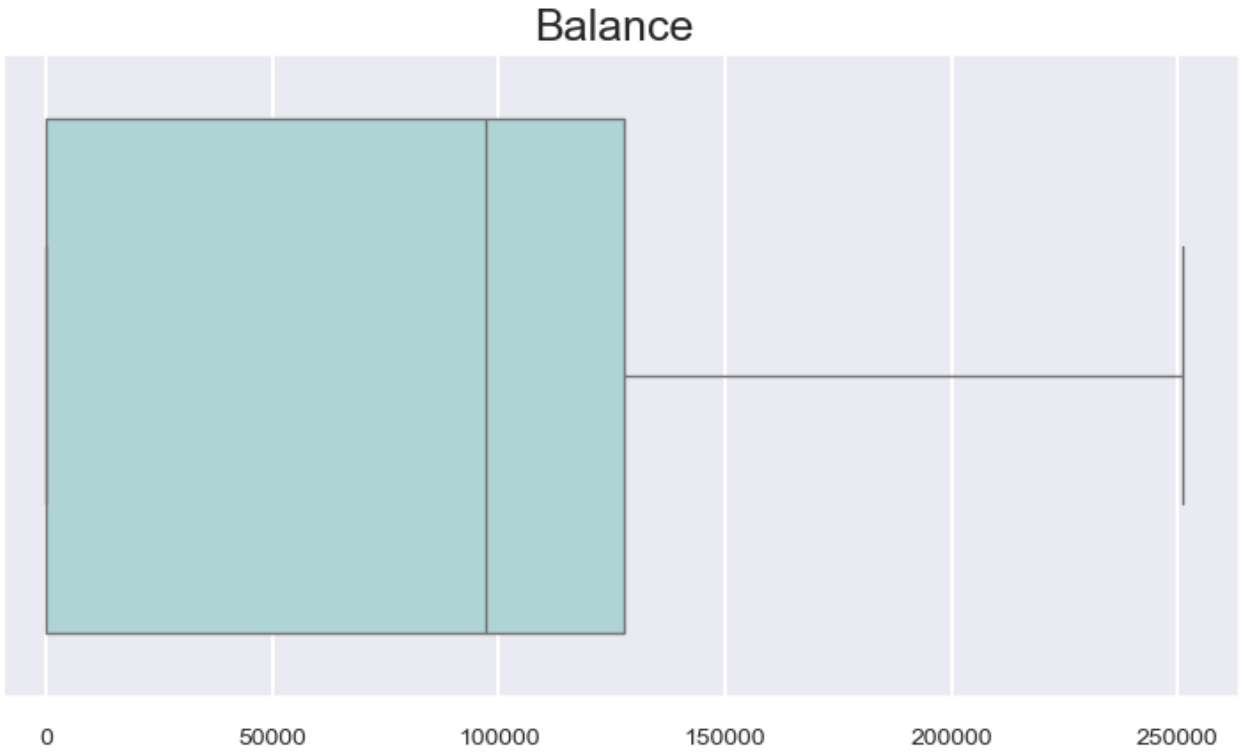




NUMERICAL VARIABLES

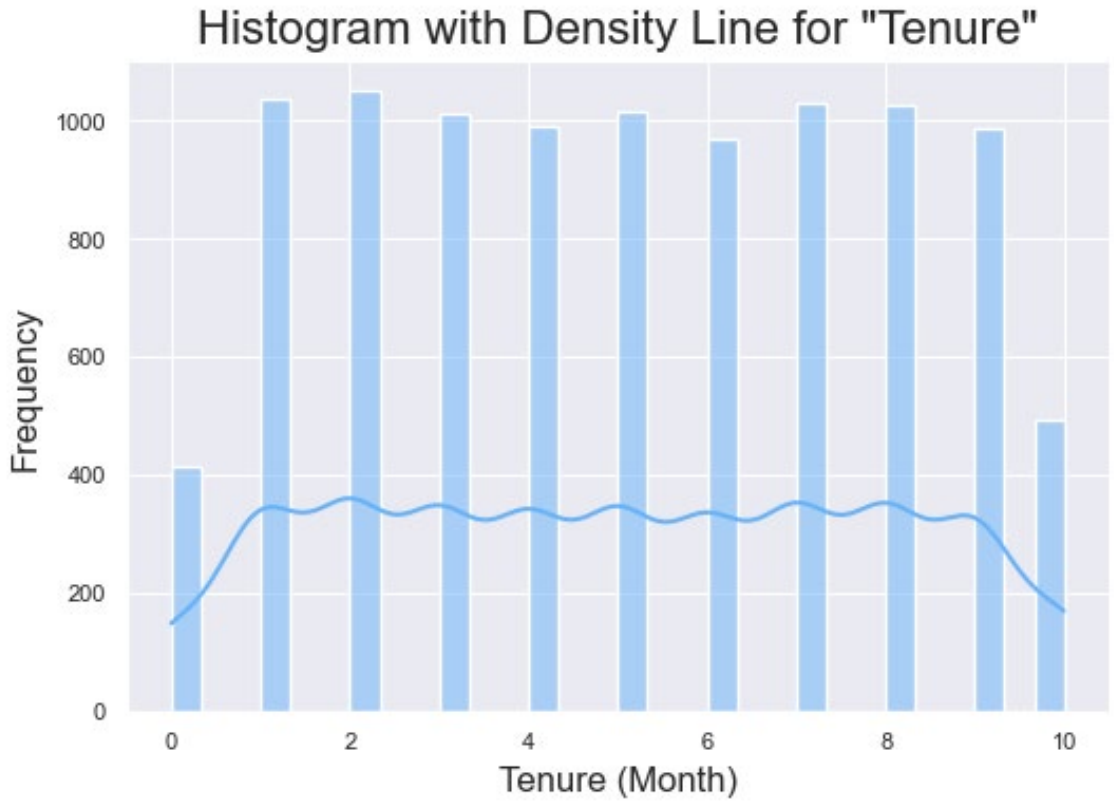
- Balance

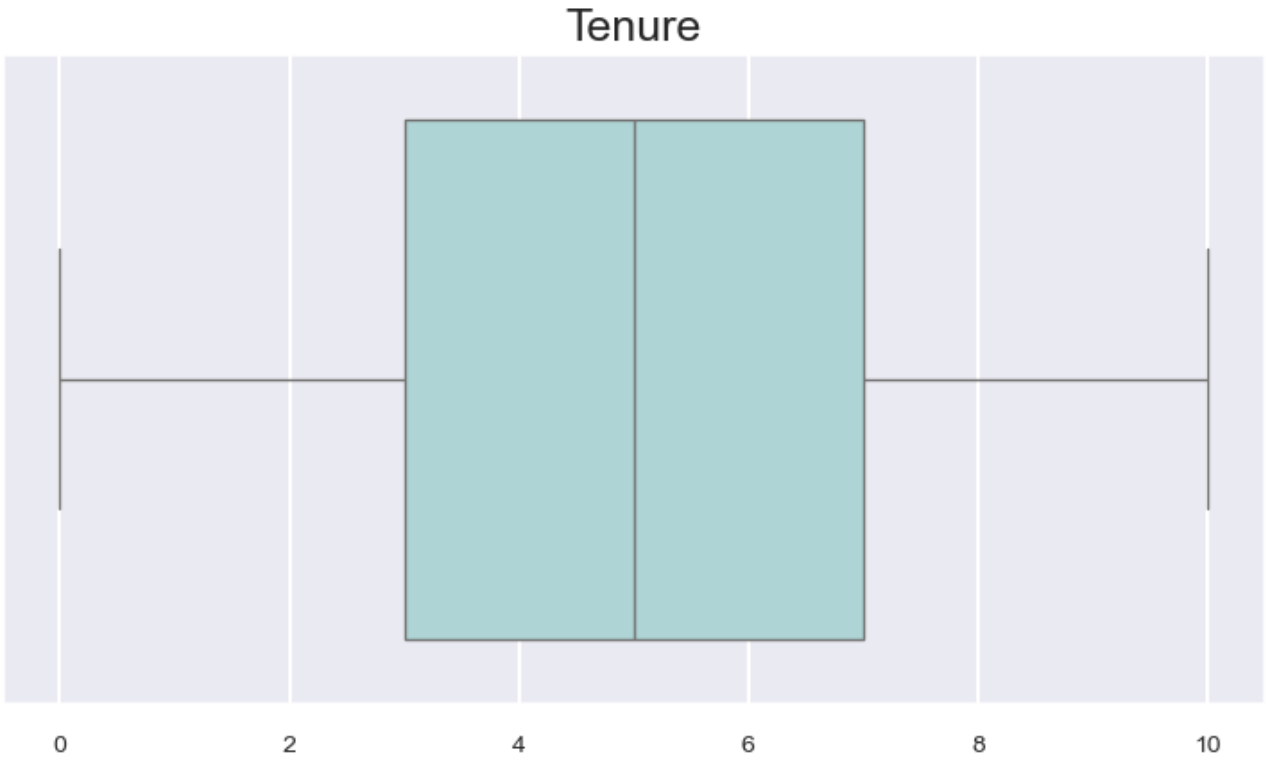




NUMERICAL VARIABLES

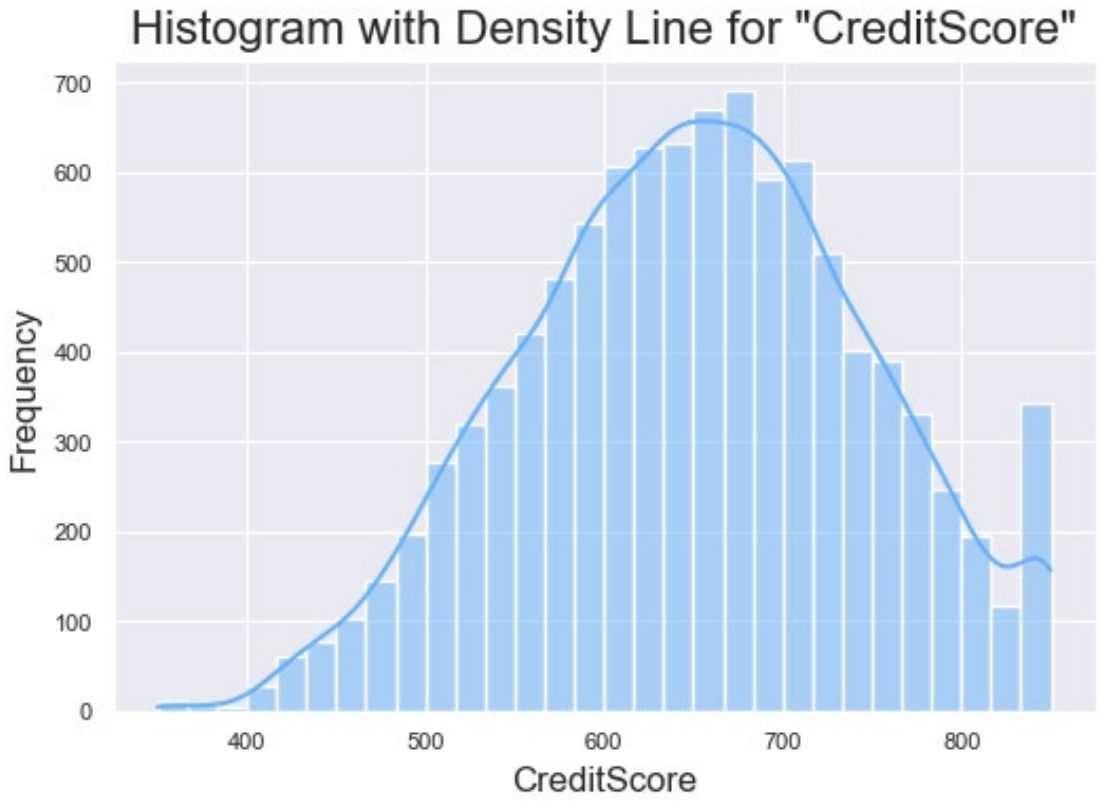
- Tenure (discrete number in month)





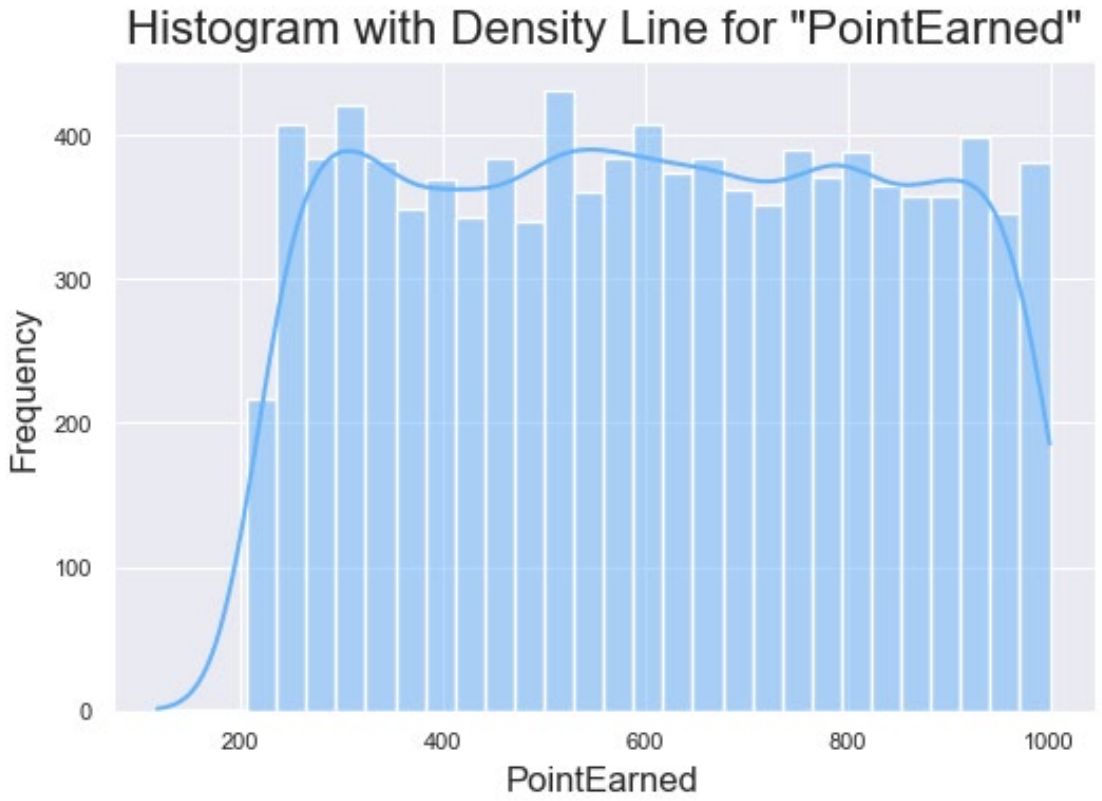
NUMERICAL VARIABLES

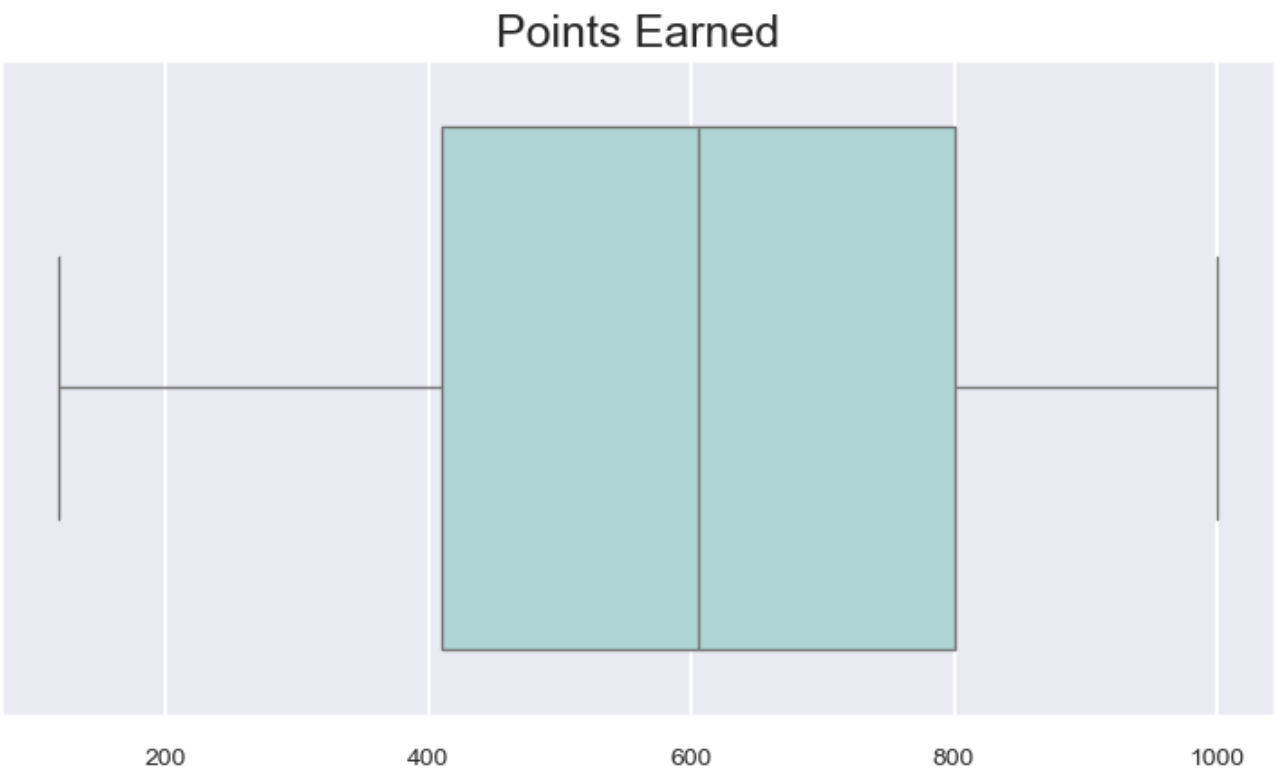
- CreditScore



NUMERICAL VARIABLES

- PointEarned





BIVARIATE ANALYSIS

HYPOTHESIS TESTS

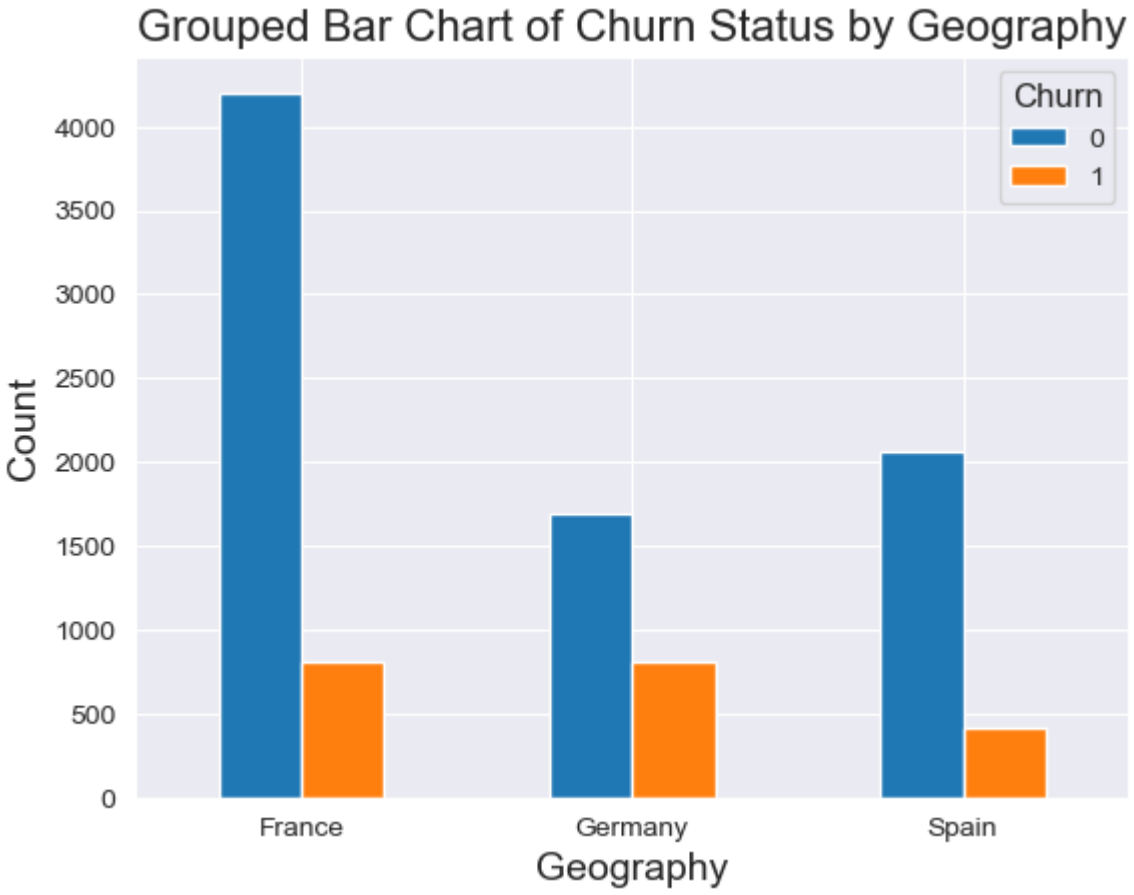
CHI-SQUARE TEST

Categorical vs Categorical

“Geography" vs "Exited"

```
chi_square(df,'Geography','Exited')
Actual values:

Exited        0     1    All
Geography
France     4203   811   5014
Germany    1695   814   2509
Spain      2064   413   2477
All        7962  2038  10000
============================================================
Expected values:

[[3992.1468 1021.8532]
 [1997.6658  511.3342]
 [1972.1874  504.8126]]
============================================================

Chi-square is : 300.6264011211942

p_value is : 5.245736109572763e-66

degree of freedom is :2
==========================================================================================================

There is statistically significant association between "Geography" and "Exited" at 0.05 significant leve
```

Exited	0	1	All
Geography			
France	4203	811	5014
Germany	1695	814	2509
Spain	2064	413	2477
All	7962	2038	10000

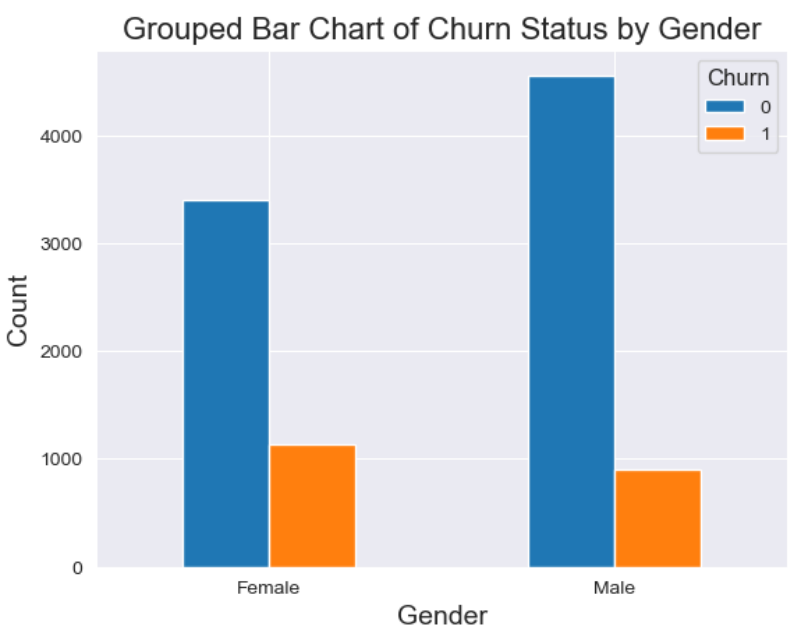


CHI-SQUARE TEST

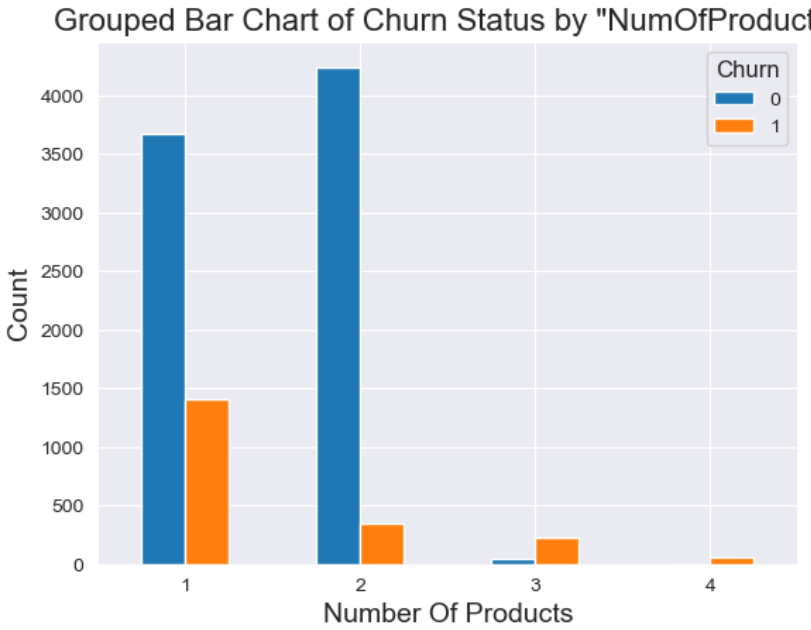
Categorical vs Categorical

ASSOCIATIONS WITH CHURN

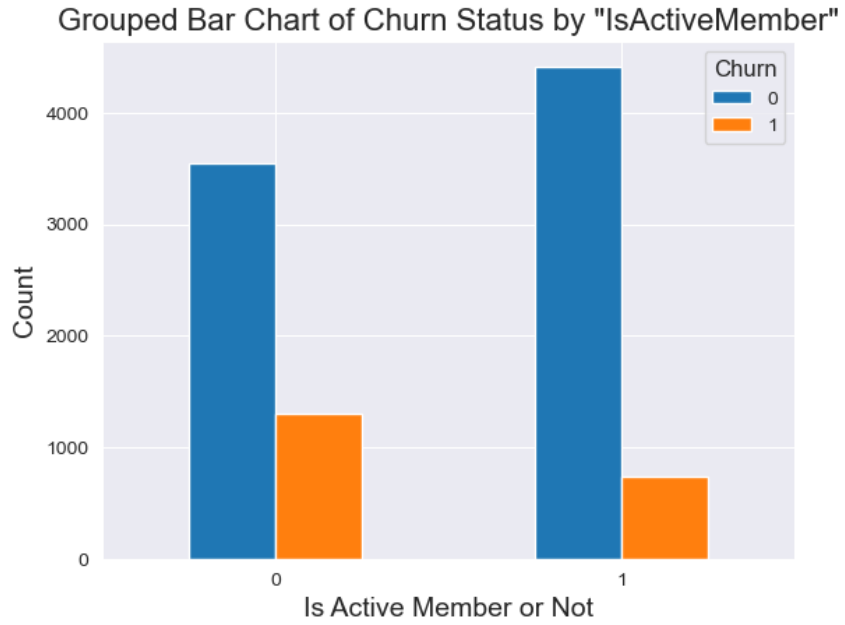
- Gender



- NumOfProducts

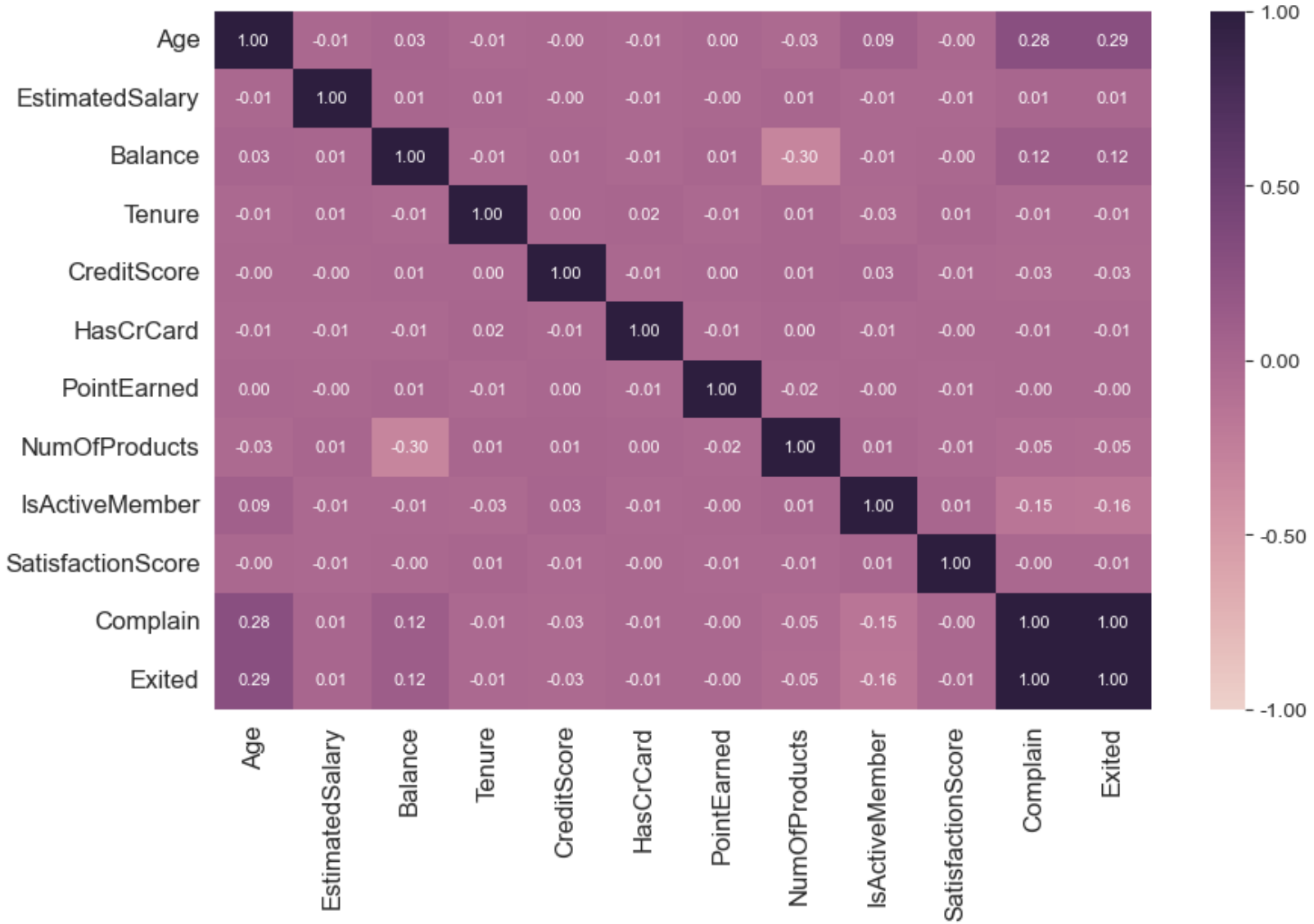


- IsActiveMember



CORRELATION ANALYSIS

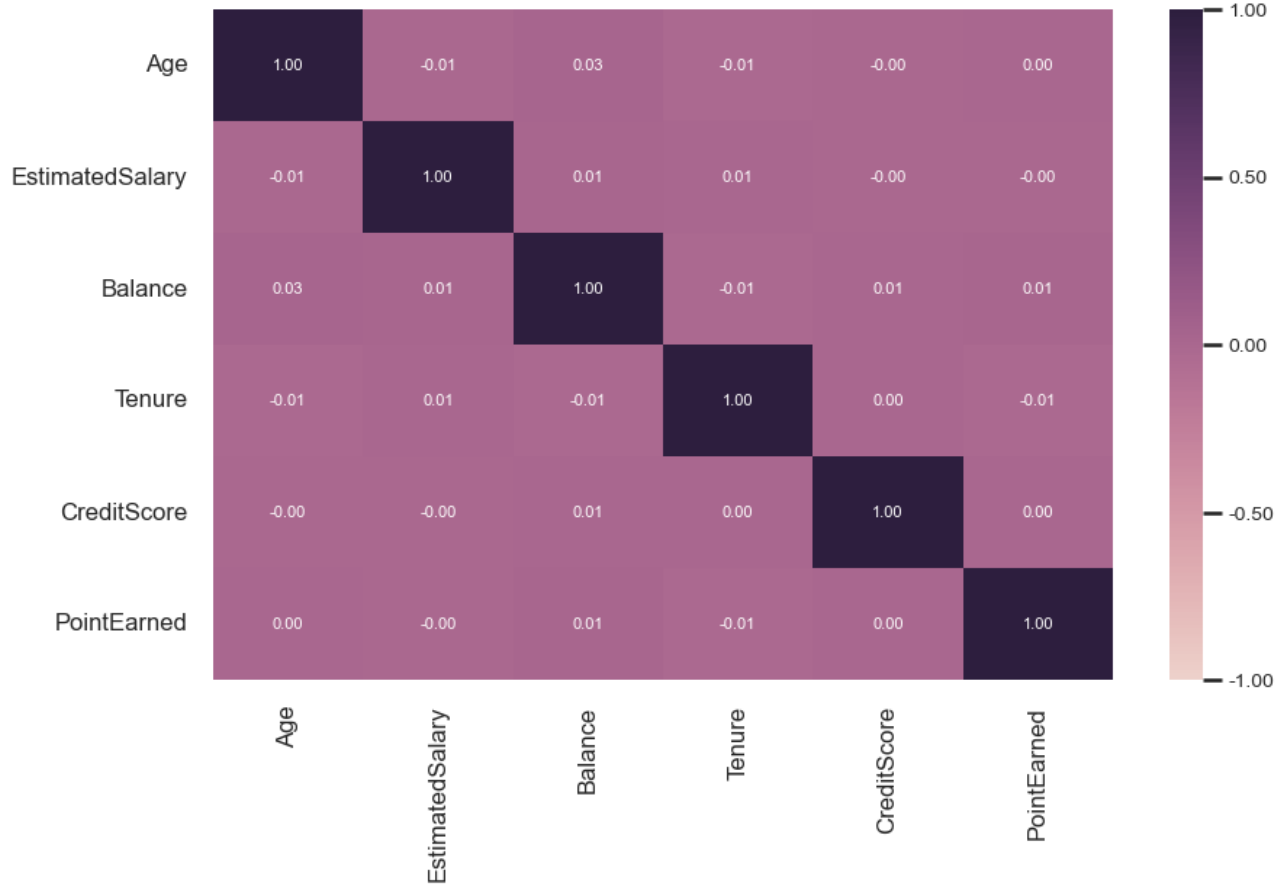
Continuous vs Continuous



CORRELATION ANALYSIS

Continuous vs Continuous

	Age	EstimatedSalary	Balance	Tenure	CreditScore	PointEarned
Age	1.000000	-0.007201	0.028308	-0.009997	-0.003965	0.002222
EstimatedSalary	-0.007201	1.000000	0.012797	0.007784	-0.001384	-0.001515
Balance	0.028308	0.012797	1.000000	-0.012254	0.006268	0.014608
Tenure	-0.009997	0.007784	-0.012254	1.000000	0.000842	-0.010196
CreditScore	-0.003965	-0.001384	0.006268	0.000842	1.000000	0.000077
PointEarned	0.002222	-0.001515	0.014608	-0.010196	0.000077	1.000000



CORRELATION ANALYSIS

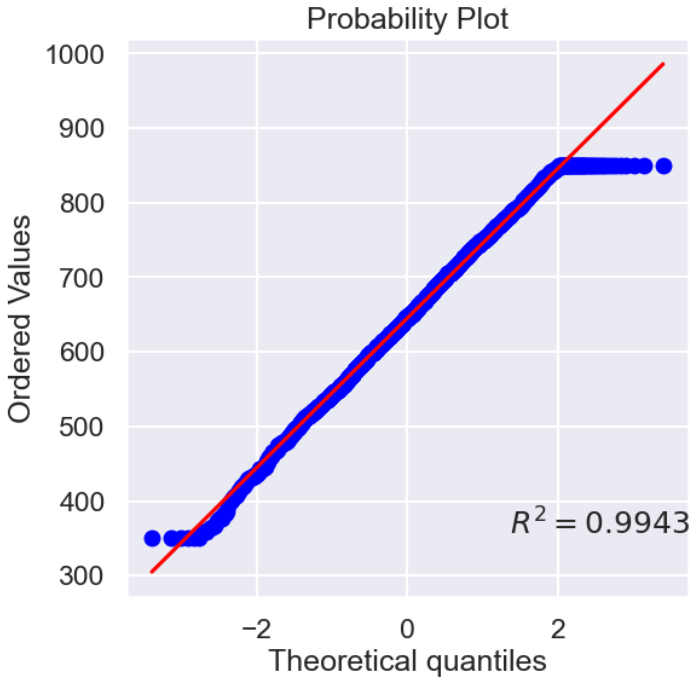
Continuous vs Continuous

T TEST

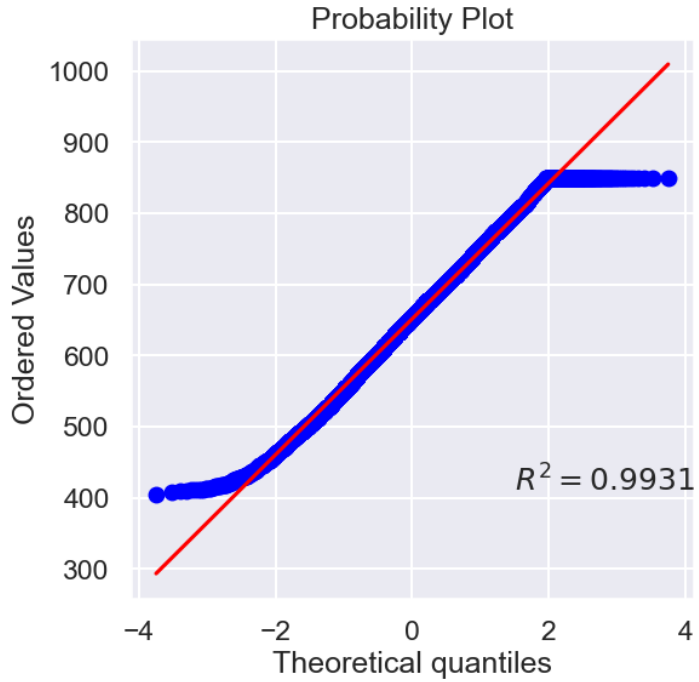
Continuous vs Categorical (2 levels)

"CreditScore" vs "Exited"

- 2 Independent groups (0 and 1)
- Normality Tests (p-value < 0.05):
  - *not normally distributed*
- Visual Evidence from Q-Q Plot
- Large Sample Size (CLT):
  - *sample mean is normally distributed*



"churn" (1)



"retain" (0)

T TEST

Continuous vs Categorical (2 levels)

"CreditScore" vs "Exited"

```
stats.levene(churn, no_churn)
```

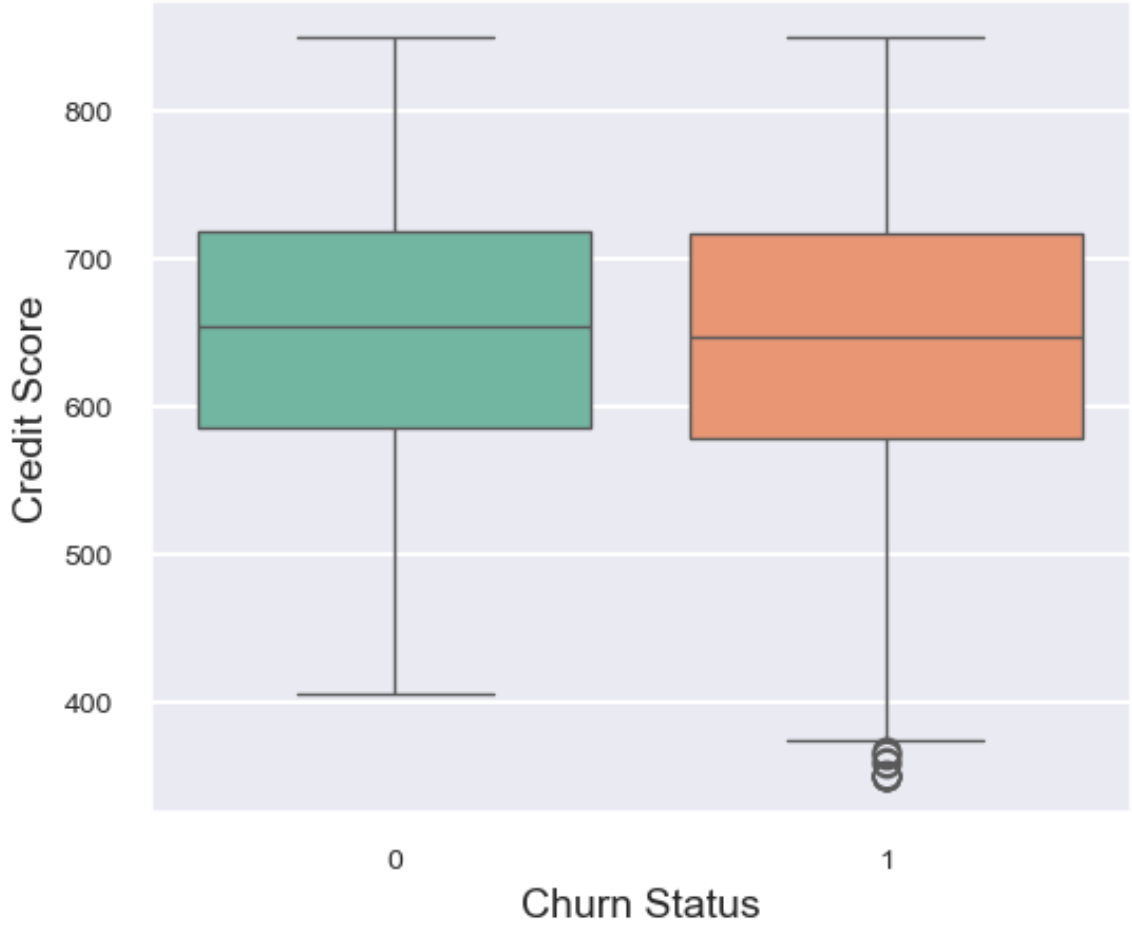
```
LeveneResult(statistic=5.616580271197907, pvalue=0.017810201526065078)
```

```
stats.ttest_ind(churn, no_churn, equal_var=False)
```

```
TtestResult(statistic=-2.6030372644244175, pvalue=0.009284913465813381, df=3052.604723318706)
```

- Unequal Variances (p-value < 0.05)
- Welch T-Test (p-value < 0.05):

  Statistically significant difference between the means of the churn and no-churn groups

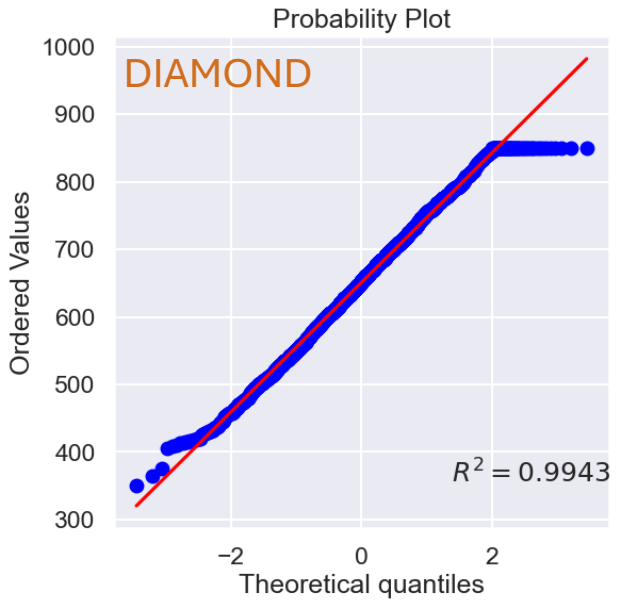


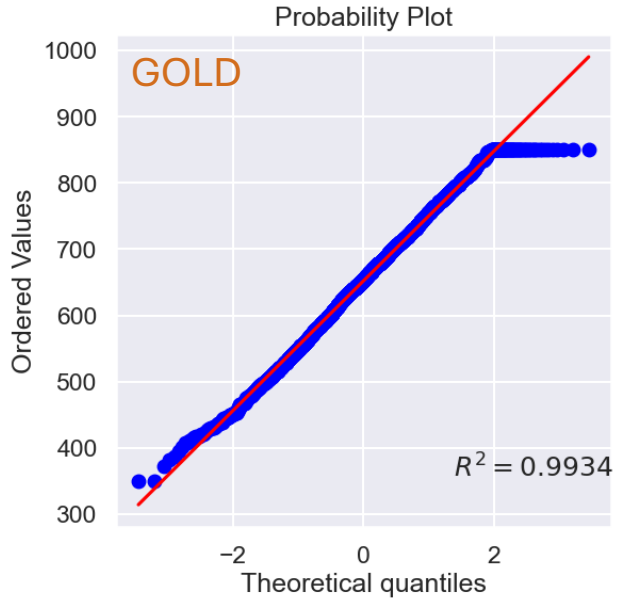
ANOVA TEST

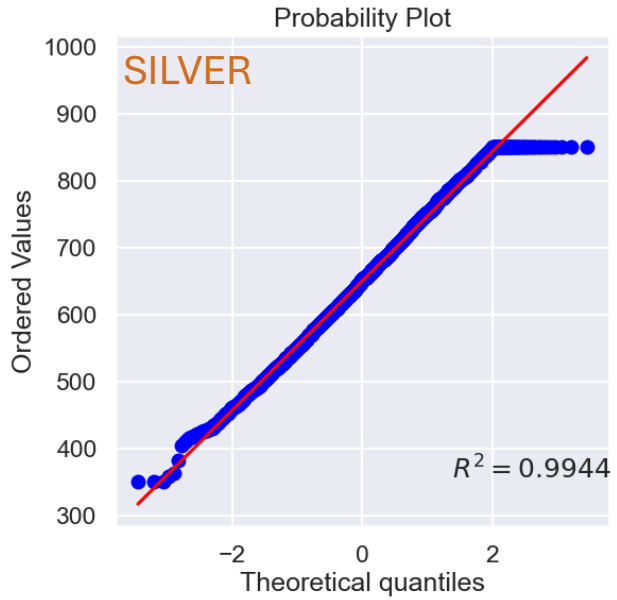
Continuous vs Categorical (>2 levels)

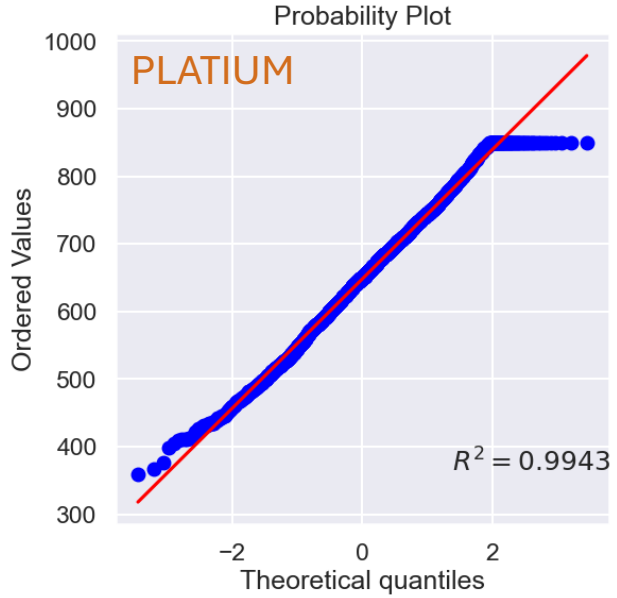
"CreditScore" vs "CardType""

- 4 independent groups
  - *DIAMOND, GOLD, SILVER, PLATIUM*
- Normality Tests (p-value < 0.05):
  - *not normally distributed*
- Visual Evidence from Q-Q Plot
- Large Sample Size (CLT):
  - *sample mean is normally distributed*









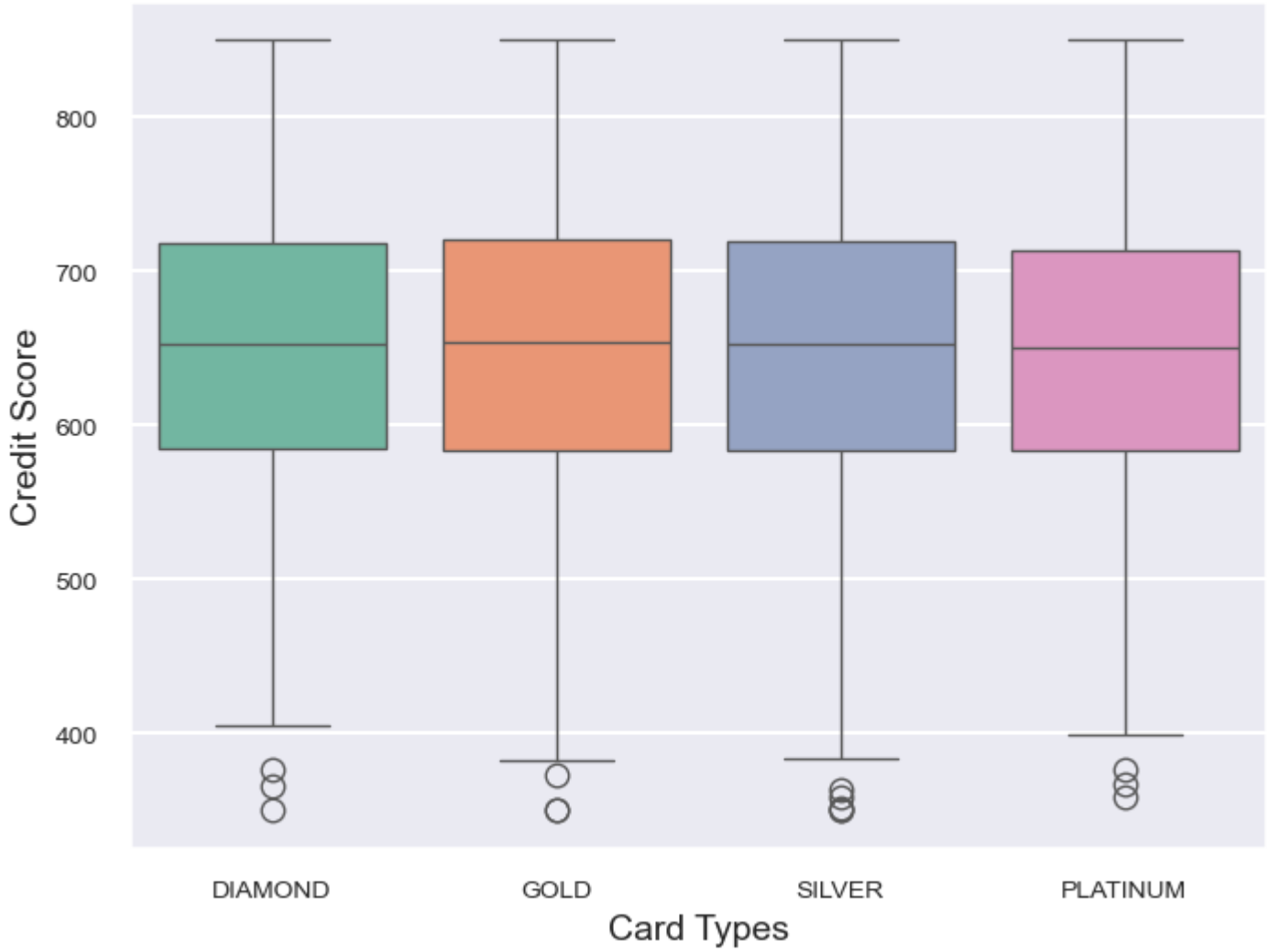
ANOVA TEST

Continuous vs Categorical (>2 levels)

"CreditScore" vs "CardType""

```
stats.levene(diamond,gold,silver,platium)
LeveneResult(statistic=0.6212572449672729, pvalue=0.6011740411463264)
stats.f_oneway(diamond,gold,silver,platium)
F_onewayResult(statistic=0.7810294924693836, pvalue=0.5043463759245638)
```

- Equal Variances (p-value > 0.05)
- One-Way ANOVA Test (p-value > 0.05):
  NO statistically significant difference between the means of the diamond, gold, silver, and platinum groups



CLASSIFICATION MODELS

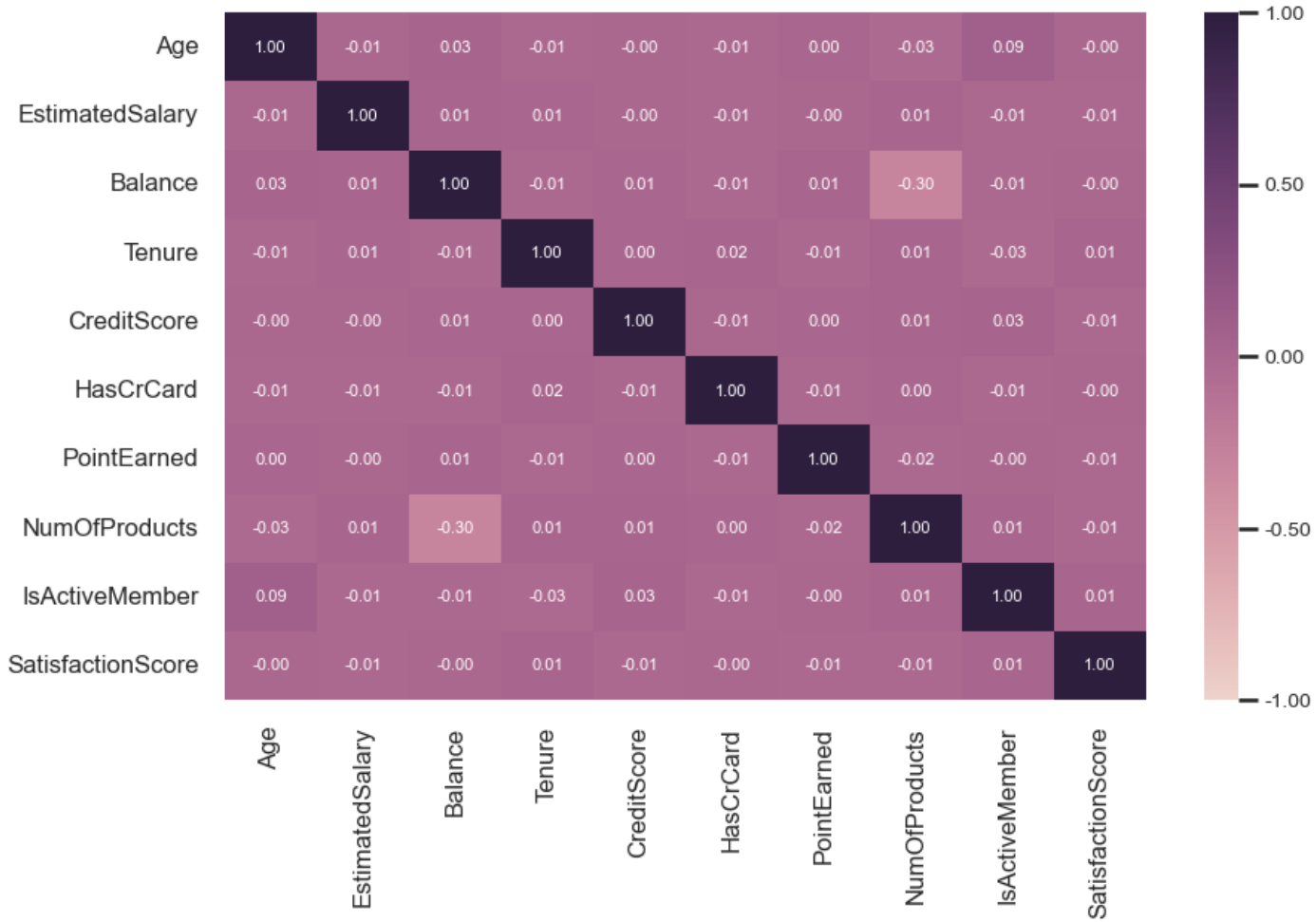
LOGISTIC REGRESSION

TARGET VARIABLE: “EXITED”

COLLINEARITY & MULTI-COLLINEARITY

	Feature	VIF
0	Age	12.646869
1	EstimatedSalary	3.900937
2	Balance	2.645507
3	Tenure	3.884442
4	CreditScore	23.923070
5	HasCrCard	3.298456
6	PointEarned	7.537213
7	NumOfProducts	7.812911
8	IsActiveMember	2.073668
9	SatisfactionScore	5.277849

```
Dropping CreditScore with VIF 23.92
Dropping Age with VIF 10.62
```



- ENCODE CATEGORICAL VARIABLES
- SCALING
- SPLIT DATASET

Geography_Germany	Geography_Spain	Gender_Male	CardType_GOLD	CardType_PLATINUM	CardType_SILVER
False	False	False	False	False	False
False	True	False	False	False	False
False	False	False	False	False	False
False	False	False	True	False	False
False	True	False	True	False	False
...	...	...	...	...	...
False	False	True	False	False	False
False	False	True	False	True	False
False	False	False	False	False	True
True	False	True	True	False	False
False	False	False	False	False	False

LOGISTIC REGRESSION

Coefficients

EstimatedSalary	Balance	Tenure	HasCrCard	PointEarned	NumOfProducts	IsActiveMember	SatisfactionScore
0.024799	0.118303	-0.048266	-0.024232	-0.014383	-0.078063	-0.402038	-0.018859

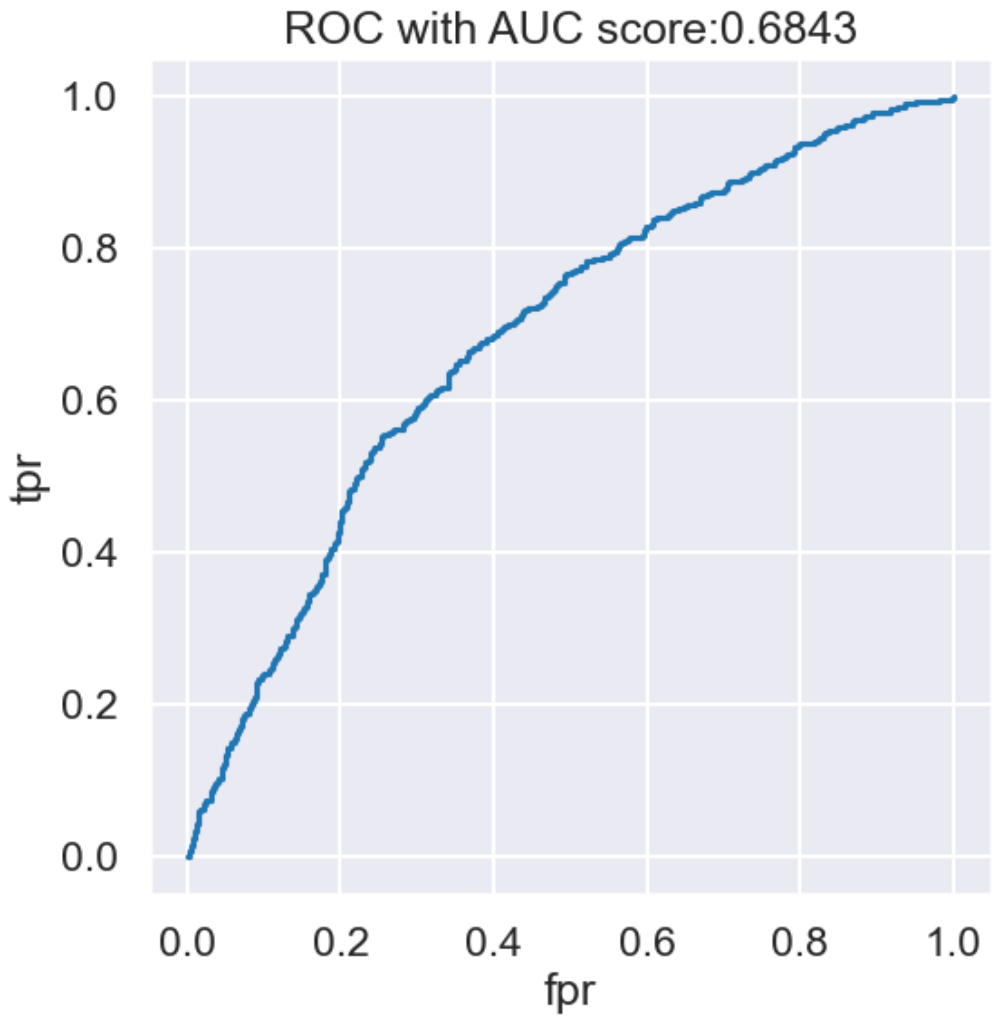
Geography_Germany	Geography_Spain	Gender_Male	CardType_GOLD	CardType_PLATINUM	CardType_SILVER
0.365241	0.03908	-0.262572	-0.03751	-0.02789	-0.009132

- Customers in Germany and those with higher balances are more likely to churn
- Active members and males are less likely to leave the service.

LOGISTIC REGRESSION

```
Accuracy_Score of Training Dataset: 0.7971
Confusion Matrix of Training Dataset:
[[6312   58]
 [1565   65]]
Classification Report of Training Dataset:
              precision    recall  f1-score   support

           0       0.80      0.99      0.89      6370
           1       0.53      0.04      0.07      1630

    accuracy                           0.80      8000
   macro avg       0.66      0.52      0.48      8000
weighted avg       0.75      0.80      0.72      8000
```

```
Accuracy_Score of Test Dataset: 0.7940
Confusion Matrix of Test Dataset:
[[1572   20]
 [ 392   16]]
Classification Report of Test Dataset:
              precision    recall  f1-score   support

           0       0.80      0.99      0.88      1592
           1       0.44      0.04      0.07       408

    accuracy                           0.79      2000
   macro avg       0.62      0.51      0.48      2000
weighted avg       0.73      0.79      0.72      2000
```



- **Cross Validation Score' mean is around 79%**
- **No severe overfitting**

LOGISTIC REGRESSION

Hyperparameter Tuning: *penalty='l1', solver='liblinear', class_weight='balanced'*

```
Accuracy_Score of Test Dataset: 0.6480
[[1037  555]
 [ 149  259]]
              precision    recall  f1-score   support

           0       0.87      0.65      0.75      1592
           1       0.32      0.63      0.42       408

    accuracy                           0.65      2000
   macro avg       0.60      0.64      0.59      2000
weighted avg       0.76      0.65      0.68      2000
```

```
Accuracy_Score of Training Dataset: 0.6464
[[4169 2201]
 [ 628 1002]]
Classification Report of Training Dataset:
              precision    recall  f1-score   support

           0       0.87      0.65      0.75      6370
           1       0.31      0.61      0.41      1630

    accuracy                           0.65      8000
   macro avg       0.59      0.63      0.58      8000
weighted avg       0.76      0.65      0.68      8000
```

- Overall model accuracy dropped
- Churn (class “1”) recall improved

LOGISTIC REGRESSION

Hyperparameter Tuning

Training Model: C=0.3 & l1_ratio=0.3

```
[[1037  555]
 [ 149  259]]
```

	precision	recall	f1-score	support
0	0.87	0.65	0.75	1592
1	0.32	0.63	0.42	408
accuracy			0.65	2000
macro avg	0.60	0.64	0.59	2000
weighted avg	0.76	0.65	0.68	2000

Training Model: C=0.1 & l1_ratio=0.7

```
[[1034  558]
 [ 149  259]]
```

	precision	recall	f1-score	support
0	0.87	0.65	0.75	1592
1	0.32	0.63	0.42	408
accuracy			0.65	2000
macro avg	0.60	0.64	0.58	2000
weighted avg	0.76	0.65	0.68	2000

- Additional hyperparameter tuning did not lead to further improvement

DECISION TREE CLASSIFICATION

TARGET VARIABLE: “EXITED”

DECISION TREE CLASSIFICATION

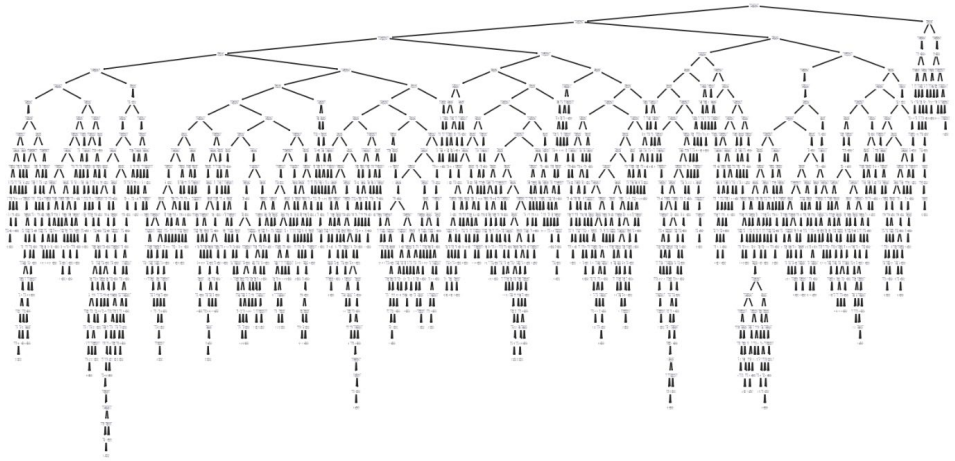
Gini Impurity

- Information Gain:
  - "EstimatedSalary" is the most informative features with lower Gini impurity

```
Gini for Geography is 0.3148
Gini for Gender is 0.3209
Gini for Age is 0.2724
Gini for EstimatedSalary is 0.0000
Gini for Balance is 0.0863
Gini for CardType is 0.3244
Gini for Tenure is 0.3241
Gini for CreditScore is 0.3080
Gini for HasCrCard is 0.3245
Gini for PointEarned is 0.2987
Gini for NumOfProducts is 0.2758
Gini for IsActiveMember is 0.3166
Gini for SatisfactionScore is 0.3244
```

DECISION TREE CLASSIFICATION

- Overfitting
  - the model have learned noise and specific patterns in the training data rather than general trends, resulting in poor generalization



```
Training Dataset: 1.0
Test Dataset: 0.7515
```

	precision	recall	f1-score	support
0	0.85	0.84	0.84	1592
1	0.40	0.42	0.41	408
accuracy			0.75	2000
macro avg	0.62	0.63	0.63	2000
weighted avg	0.76	0.75	0.75	2000

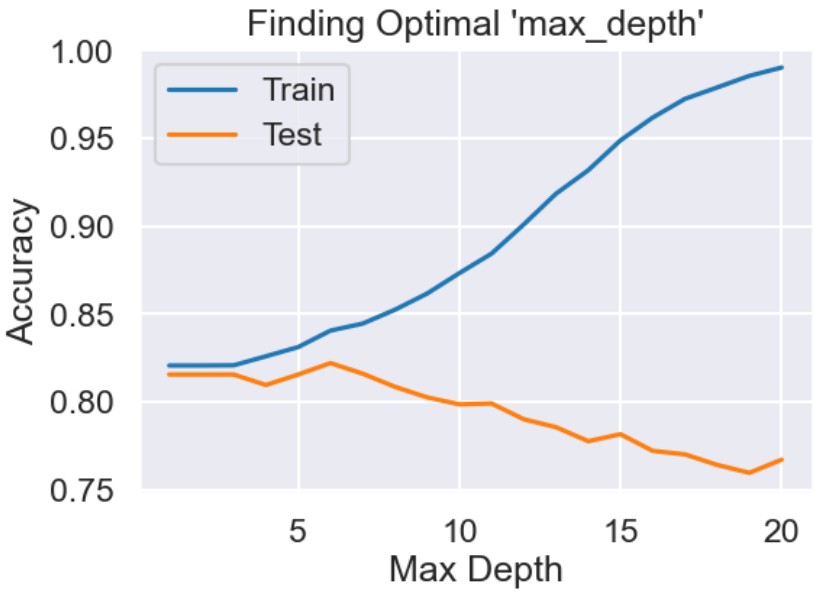
DECISION TREE CLASSIFICATION

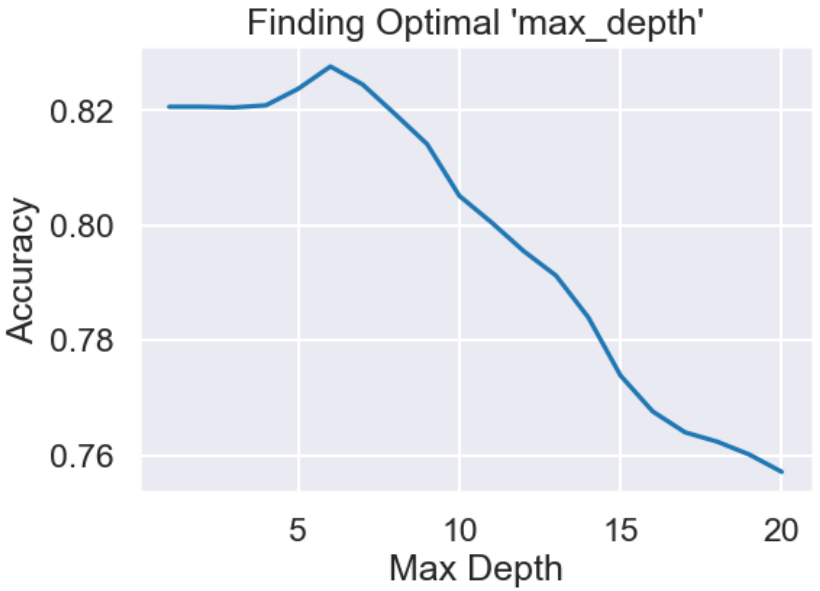
max_depth = 6

```
Training Dataset: 0.8405
Test Dataset: 0.8225
```

	precision	recall	f1-score	support
0	0.85	0.94	0.89	1592
1	0.61	0.35	0.44	408
accuracy			0.82	2000
macro avg	0.73	0.65	0.67	2000
weighted avg	0.80	0.82	0.80	2000

- Overfitting reduced
- Accuracy score improved 7%
- Churn recall dropped 7%
- Overall model generalization improved





CONCLUSIONS:

- Decision Tree model performs better overall with higher accuracy score (82%)
- Logistic Regression model is more effective at identifying churners (63%)
- Decision Tree predicts "Not Churn" (94%) with much higher recall than the Logistic Regression model (65%).

Depending on the business objective, either model could be considered.

REGRESSION MODELS

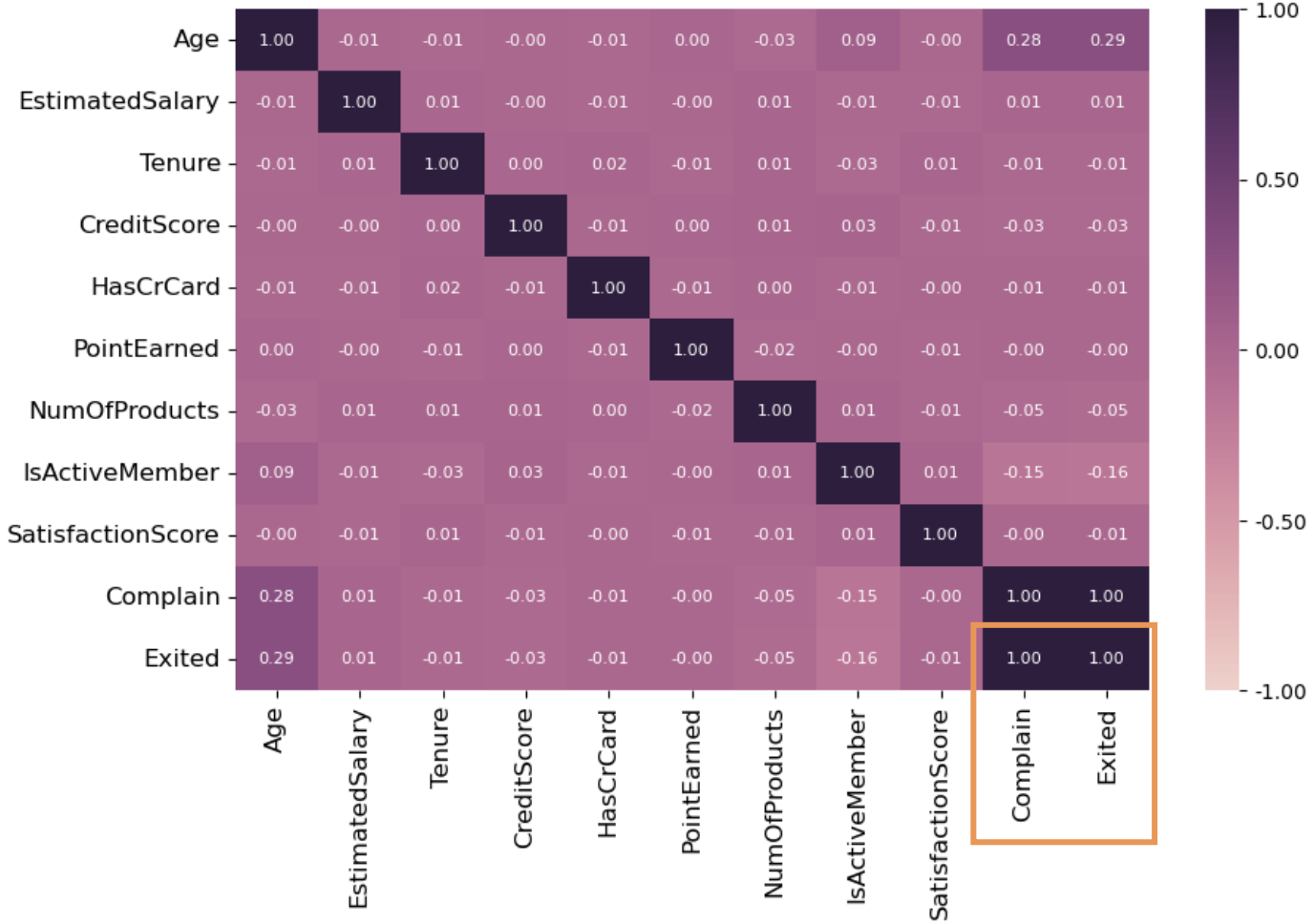
MULTI LINEAR REGRESSION

TARGET VARIABLE: “BALANCE”

COLLINEARITY & MULTI-COLLINEARITY

	Feature	VIF
0	Age	13.905867
1	EstimatedSalary	3.888651
2	Tenure	3.882498
3	CreditScore	22.839344
4	HasCrCard	3.297072
5	PointEarned	7.497384
6	NumOfProducts	7.315640
7	IsActiveMember	2.149608
8	SatisfactionScore	5.267046
9	Complain	1.419533

```
Dropping CreditScore with VIF 22.84
Dropping Age with VIF 11.34
```



- ENCODE CATEGORICAL VARIABLES
- SCALING
- SPLIT DATASET

Geography_Germany	Geography_Spain	Gender_Male	CardType_GOLD	CardType_PLATINUM	CardType_SILVER
False	False	False	False	False	False
False	True	False	False	False	False
False	False	False	False	False	False
False	False	False	True	False	False
False	True	False	True	False	False
...	...	...	...	...	...
False	False	True	False	False	False
False	False	True	False	True	False
False	False	False	False	False	True
True	False	True	True	False	False
False	False	False	False	False	False

MULTI LINEAR REGRESSION

Coefficients

Balance = const + NumOfProducts *(-1908e+04) + Complain*(2500.3903) + Geography_Germany*(2.447e+04)

- R-squared (26%)
- Most p-value of coefficients are higher than 5%

Cross Validation Score:

RMSE: 53703.90262768056
R^2: 0.2585967474099884

OLS Regression Results

Dep. Variable:	Balance	R-squared:	0.259
Model:	OLS	Adj. R-squared:	0.259
Method:	Least Squares	F-statistic:	931.1
Date:	Mon, 21 Oct 2024	Prob (F-statistic):	0.00
Time:	23:25:51	Log-Likelihood:	-98480.
No. Observations:	8000	AIC:	1.970e+05
Df Residuals:	7996	BIC:	1.970e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.646e+04	600.440	127.346	0.000	7.53e+04	7.76e+04
NumOfProducts	-1.908e+04	600.942	-31.743	0.000	-2.03e+04	-1.79e+04
Complain	2500.3903	609.597	4.102	0.000	1305.420	3695.360
Geography_Germany	2.447e+04	609.172	40.171	0.000	2.33e+04	2.57e+04

Omnibus:	663.303	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	369.651
Skew:	0.380	Prob(JB):	5.39e-81
Kurtosis:	2.271	Cond. No.	1.19

OLS Regression Results

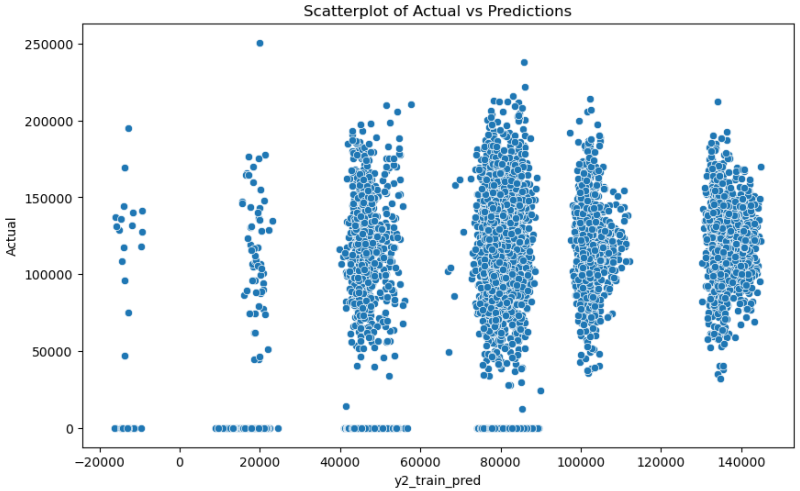
Dep. Variable:	Balance	R-squared:	0.260
Model:	OLS	Adj. R-squared:	0.258
Method:	Least Squares	F-statistic:	200.0
Date:	Mon, 21 Oct 2024	Prob (F-statistic):	0.00
Time:	23:24:06	Log-Likelihood:	-98476.
No. Observations:	8000	AIC:	1.970e+05
Df Residuals:	7985	BIC:	1.971e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.646e+04	600.554	127.322	0.000	7.53e+04	7.76e+04
EstimatedSalary	528.5089	600.952	0.879	0.379	-649.514	1706.532
Tenure	-141.5173	601.364	-0.235	0.814	-1320.349	1037.314
HasCrCard	-905.3534	601.147	-1.506	0.132	-2083.758	273.051
PointEarned	556.9127	601.081	0.927	0.354	-621.362	1735.188
NumOfProducts	-1.904e+04	601.677	-31.644	0.000	-2.02e+04	-1.79e+04
IsActiveMember	309.6614	608.860	0.509	0.611	-883.863	1503.186
SatisfactionScore	84.8345	600.797	0.141	0.888	-1092.884	1262.553
Complain	2623.9893	621.071	4.225	0.000	1406.528	3841.451
Geography_Germany	2.436e+04	645.952	37.708	0.000	2.31e+04	2.56e+04
Geography_Spain	-332.6295	637.586	-0.522	0.602	-1582.464	917.205
Gender_Male	917.9316	604.758	1.518	0.129	-267.552	2103.415
CardType_GOLD	-632.8493	736.077	-0.860	0.390	-2075.753	810.055
CardType_PLATINUM	-424.2279	737.259	-0.575	0.565	-1869.448	1020.992
CardType_SILVER	-771.6610	736.355	-1.048	0.295	-2215.109	671.787

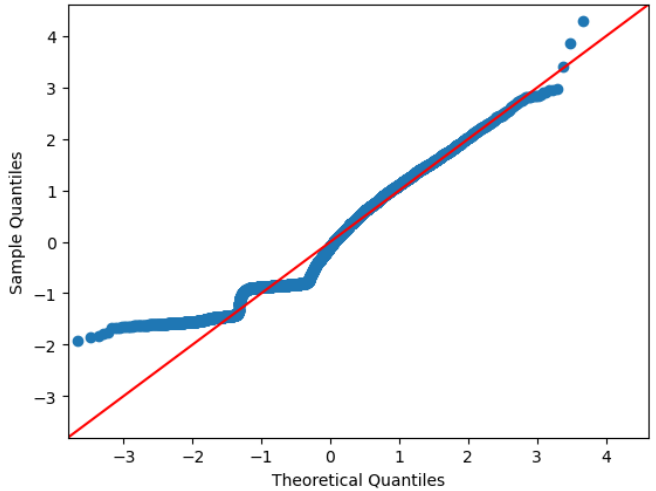
Omnibus:	661.380	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	369.067
Skew:	0.380	Prob(JB):	7.21e-81
Kurtosis:	2.272	Cond. No.	2.08

MULTI LINEAR REGRESSION

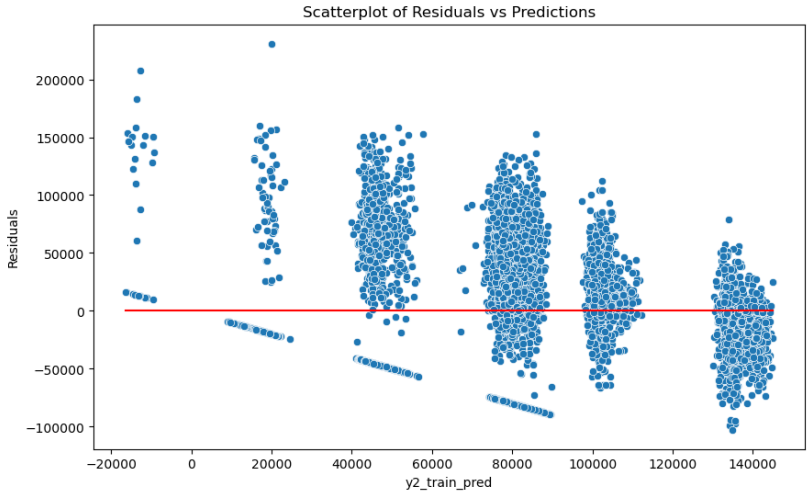
Check Assumptions: LINE



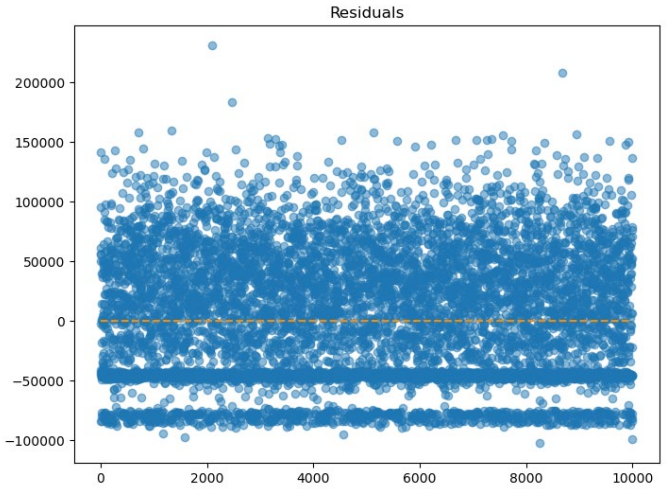
NO Linearity



NOT
Normally Distributed



NOT
Independent
Residuals



Equal Variance

DECISION TREE REGRESSOR

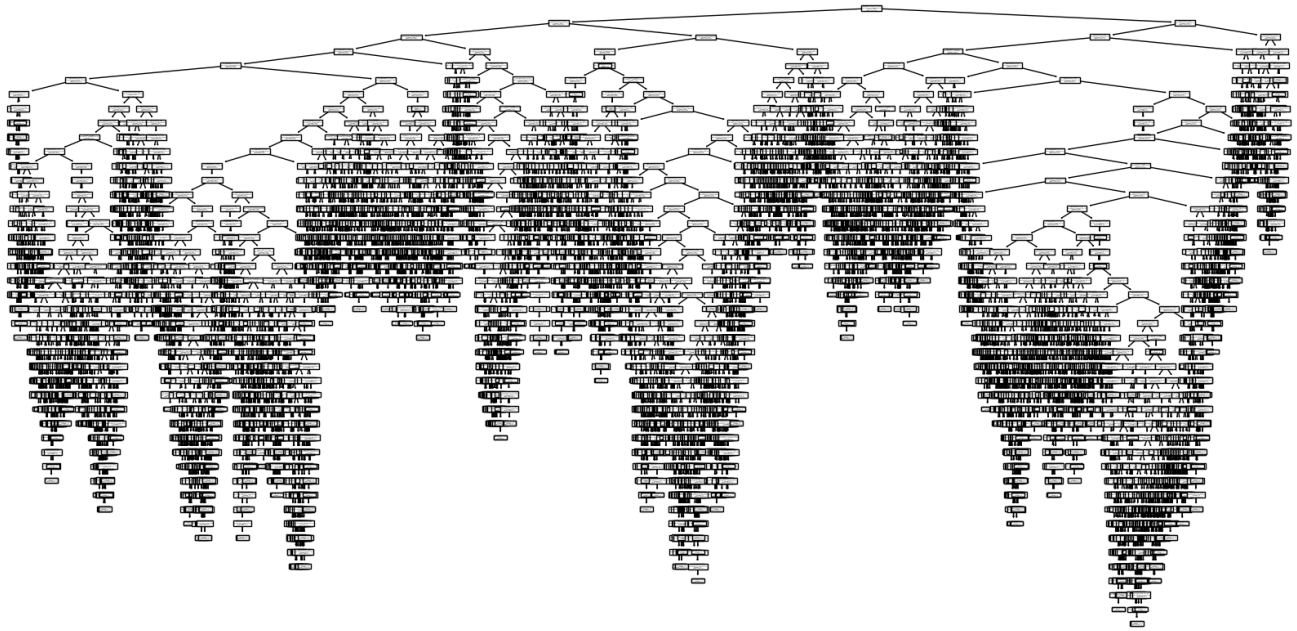
TARGET VARIABLE: “BALANCE”

DECISION TREE REGRESSOR

```
Training Dataset:
RMSE(Mean Squared Error): 0.0
R^2: 1.0
================================================
Test Dataset:
RMSE(Mean Squared Error): 73461.64656807632
R^2: -0.38168071031768824
```

Cross Validation Score:

```
R^2: -0.3422997757934867
RMSE: 71515.35943989789
```



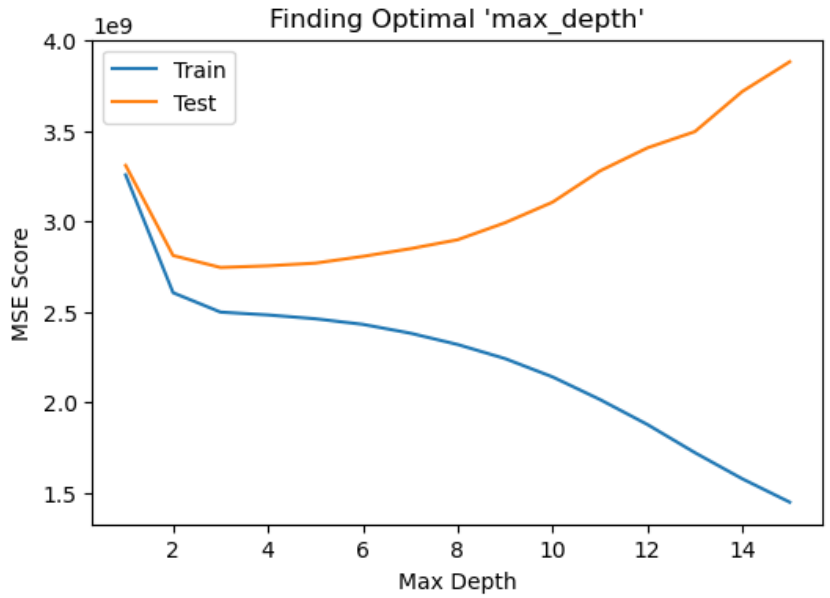
- R-squared is negative
- Overfitting

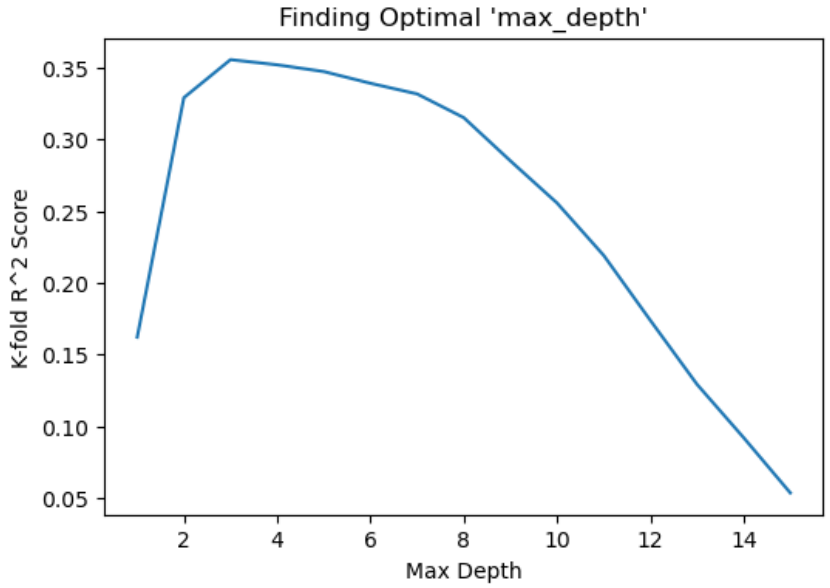
DECISION TREE REGRESSOR

max_depth = 3

```
Training Dataset:
RMSE(Mean Squared Error): 49986.095611323035
R^2: 0.3576587486420212
================================================
Test Dataset:
RMSE(Mean Squared Error): 52393.29245000105
R^2: 0.2971902798735644
```

- Overfitting reduced
- R-squared improved
- Overall model generalization improved





CONCLUSIONS:

Both the Linear Regression and Decision Tree Regressor models struggle to accurately predict "Balance." However, the Decision Tree Regressor demonstrates better performance, with a higher accuracy score and reduced error after pruning. Despite this improvement, both models still leave room for optimization, indicating that additional tuning or feature engineering may be necessary to enhance predictive accuracy.


THANK YOU
Instructor: Amit Kukreja