



Data Collection and Preprocessing Phase

Section	Des	script	ion										
	Rac	ic star	tictics	din	nenci	One	211/	d struc	ture (of the	data		
		data.describ		s, um	1101131	OHS,	, am	a siruc	ture (of the	uata.		•
Data Overview	[22]:		itScore	Age	Tenure	R	talance I	NumOfProducts	HasCrCard	IsActiveMe	mber Estimat	adSalary	Exited
					0000,000000		000000	10000.000000	10000.00000				000.000000
				38.921800	5.012800		889288	1.530200	0.70550			0.239881	0.203700
	1	std 96	653299	10.487806	2.892174	62397.4	405202	0.581654	0.45584	0.49	9797 5751	0.492818	0.402769
		min 350	000000	18.000000	0.000000	0.0	000000	1.000000	0.00000	0.00	0000 1	1.580000	0.000000
		25 % 584	000000	32.000000	3.000000	0.0	000000	1.000000	0.00000	0.00	0000 5100	2.110000	0.000000
		50 % 652	000000	37.000000	5.000000	97198.	540000	1.000000	1.00000	1.00	0000 10019	3.915000	0.000000
		75 % 718	000000	44.000000	7.000000	127644.	240000	2.000000	1.00000	1.00	0000 14938	8.247500	0.000000
		max 850	000000	92.000000	10.000000	250898.6	090000	4.000000	1.00000	1.00	0000 19999	2.480000	1.000000
	[30]:	data.corr()											
	[30];		CreditScore	Geography	Gender	Age	Tenure	Balance Nur	nOfProducts	HasCrCard IsA	ctiveMember	EstimatedSalar	ry Exited
		CreditScor	1.000000	0.007888	-0.002857	-0.003965	0.000842	0.006268	0.012238	-0.005458	0.025651	-0.00138	34 -0.027094
		Geograph	0.007888	1.000000	0.004719	0.022812	0.003739	0.069408	0.003972	-0.008523	0.006724	-0.00136	59 0.035943
		Gende	r -0.002857	0.004719	1.000000	-0.027544	0.014733	0.012087	-0.021859	0.005766	0.022544	-0.00811	12 -0.106512
		Ag				1.000000	-0.009997		-0.030680	-0.011721	0.085472	-0.00720	
	l u	Tenur				-0.009997 0.028308	-0.012254		0.013444 -0.304180	0.022583 -0.014858	-0.028362 -0.010084	0.00778	
		NumOfProduct				-0.030680	0.013444		1.000000	0.003183	0.009612	0.01420	
	l l	HasCrCar				-0.011721	0.022583	-0.014858	0.003183	1.000000	-0.011866	-0.00993	
		IsActiveMembe	r 0.025651	0.006724	0.022544	0.085472	-0.028362	-0.010084	0.009612	-0.011866	1.000000	-0.01142	21 -0.156128
		EstimatedSalar	-0.001384	-0.001369	-0.008112	-0.007201	0.007784	0.012797	0.014204	-0.009933	-0.011421	1.00000	0.012097
		Exite	-0.027094	0.035943	-0.106512	0.285323	-0.014001	0.118533	-0.047820	-0.007138	-0.156128	0.01209	1.000000
Univariate Analysis	Fvr	Morati	on of	indi	vidu	a1 v	arial	oles (n	าคลท	medi	an ma	nde e	etc)
Cilivariate Aliarysis	LA	morali	on or	mui	v Iuu	ii vc	11 IUI	oics (II	ican,	mean	u11, 111N	ouc, t)
Univariate Analysis	Exp	olorati	on of	indi	vidua	al va	arial	oles (n	nean,	medi	an, mo	ode, e	





Bivariate Analysis	Relationships between two variables (correlation, scatter plots).
--------------------	---

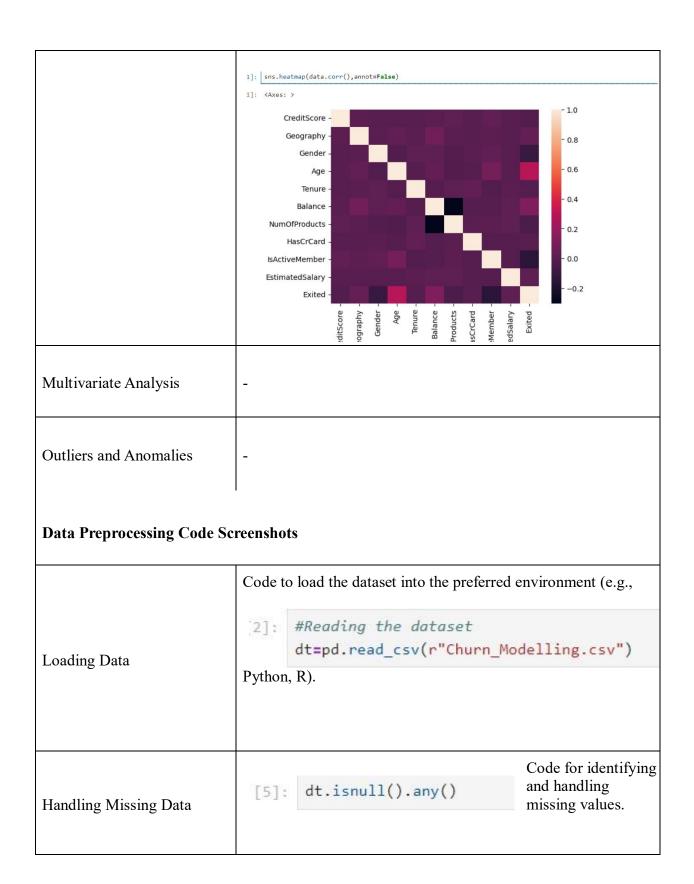
Date	15 July 2024
Team ID	739886
Project Title	Telecom Customer Churn Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.











Data Transformation	[42]: #training and testing the data from sklearn.model_selection import train_test_split x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0) [43]: #Feature Scaling from sklearn.preprocessing import StandardScaler sc=StandardScaler() x_train=sc.fit_transform(x_train) x_test = sc.transform(x_test) Code for transforming variables (scaling, normalization).
Feature Engineering	7]: #Removing columns data=dt.drop(['RowNumber','CustomerId','Surname'],axis=True) Code for creating new features or modifying existing ones.

Coc [23]:	ode to save the cleaned and processed data for future use. 3]: #labelEncoding
[2	4]: from sklearn.preprocessing import LabelEncoder le=LabelEncoder()
Save Processed Data	### ### ### ##########################