



Data Collection and Preprocessing Phase

Date	02 June 2024
Team ID	737568
Project Title	AutoForesight : A Predictive Model for Streamlining Car Loan Repayment Planning
Maximum Marks	6 Marks

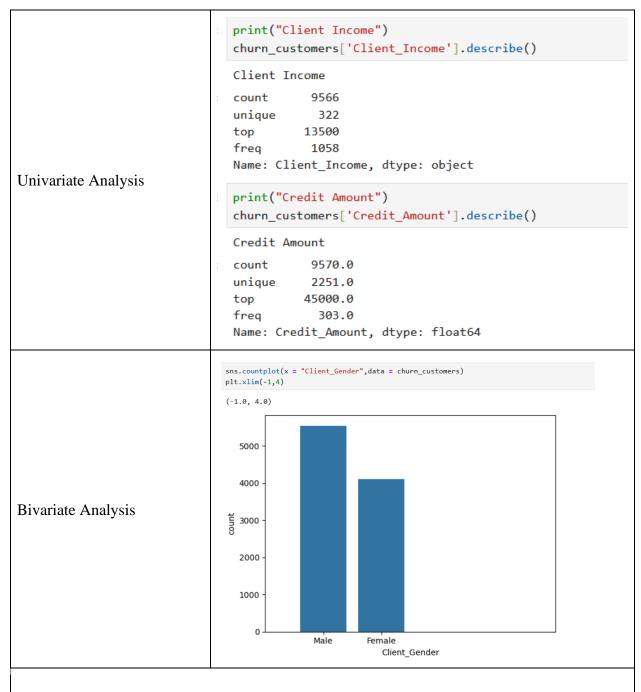
Data Exploration and Preprocessing Template

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

Section	De	escrij	otion									
	: data	set										
	1	ID	Client_Income	Car_Owned	Bike_Owned	Active_Loan	House_Own	Child_Count	Credit_Amount	Loan_Annuity	Accompany_Client	Client_Permanent
		0 12142509	6750	0.0	0.0	1.0	0.0	0.0	61190.55	3416.85	Alone	
		1 12138936	20250	1.0	0.0	1.0	NaN	0.0	15282	1826.55	Alone	***
		2 12181264	18000	0.0	0.0	1.0	0.0	1.0	59527.35	2788.2	Alone	
		3 12188929	15750	0.0	0.0	1.0	1.0	0.0	53870.4	2295.45	Alone	
		4 12133385	33750	1.0	0.0	1.0	0.0	2.0	133988.4	3547.35	Alone	-
Data Overview			-			-		-	-		-	
	1218	351 12207714	29250	0.0	0.0	NaN	1.0	0.0	107820	3165.3	Relative	
	1218	12173765	15750	0.0	1.0	1.0	0.0	0.0	104256	3388.05	Alone	
	1218	12103937	8100	0.0	1.0	0.0	1.0	1.0	55107.9	2989.35	Alone	
	1218	354 12170623	38250	1.0	1.0	0.0	1.0	0.0	45000	2719.35	Alone	
	1210	355 12105610	9000	1.0	1.0	1.0	1.0	1.0	62428.95	4201.65	Alone	
	1218	56 rows × 40 c	olumns									







Data Preprocessing Code Screenshots





	#Reading the datase	t									
	dataset:pd.read_csv(r*D:\Documents\dataset\train.csv*')										
	C:\Users\Sharan\AppOats\iocal\Templipykernel_5604\3481427543.py:1: DtypeWarning: Columns (1,7,8,16,17,18,19,20,35) have mixed types. Specify dtype option on import or set low_memory=false. dataset-pd.read_cx/fo":\Documents\dataset\train.csv")										
	dataset										
	ID C	lient_Income	Car_Owned Bil	ce_Owned Act	ive_Loan Ho	use_Own Ch	ild_Count	Credit_Amount	Loan_Annuity A	Accompany_Client C	lient_Permanent
Loading Data	0 12142509	6750	0.0	0.0	1.0	0.0	0.0	61190.55	3416.85	Alone	
Louding Data	1 12138936	20250	1.0	0.0	1.0	NaN	0.0	15282	1826.55	Alone	
	2 12181264	18000	0.0	0,0	1.0	0.0	1.0	59527.35	2788.2	Alone	
	3 12188929	15750	0.0	0.0	1.0	1.0	0.0	53870.4	2295.45	Alone	
	4 12133385	33750	1.0	0.0	1.0	0.0	2.0	133988.4	3547.35	Alone	
			-	***	-		-				
	121851 12207714	29250	0.0	0.0	NaN	1.0	0.0	107820	3165.3	Relative	
	<pre># [Data Pre-Process dataset= dataset.dr dataset.head()</pre>				ult','Age_D	ys','Employ	ed_Days','S	core_Source_1	','Scone_Sounce_	_2','Score_Source_3','	Registration_
	ID Client_I	ncome Car_O	wned Bike_Ow	ned Active_Lo	an House_O	wn Child_Co	unt Credit	Amount Loan	_Annuity Accomp	pany_Client Client_H	lousing_Type
Handling Missing Data	0 12142509	6750	0.0	0.0	1.0	0.0	0.0	61190.55	3416.85	Alone	Home
Handing Missing Data	1 12138936	20250					0.0	15282	1826.55	Alone	Home
	2 12181264	18000	0.0			0.0		59527.35	2788.2	Alone	Family
	3 12188929	15750	0.0	0.0	1.0	1.0	0.0	53870.4	2295.45	Alone	Home
	4 12133385	33750	1.0	0.0	1.0	0.0	2.0	133988.4	3547.35	Alone	Home
	5 rows × 25 columns # -HandLing Co			numonis (d	atacet['	Client Tr	vromo!]		anso!)		
Data Transformation	# -Handling Co dataset['Clien dataset['Cred dataset['Popu	nt_Income it_Amount lation_Re _l	'] = pd.to '] = pd.to gion_Relat:	_numeric(d ive'] = pd	ataset[' .to_nume	Credit_Am	nount'],	errors='co ulation_Re	erce') gion_Relativ	ve'],errors='coe	erce')
Data Transformation	# -Handling Co	nt_Income it_Amount lation_Re; Annuity'	'] = pd.to. '] = pd.to. gion_Relat: = pd.to_ reating data fi	_numeric(daive'] = pd numeric(da	ataset[' .to_nume taset['L	Credit_Am ric(datas pan_Annui	mount'], set['Pop ity'],er	errors='co ulation_Re rors='coer	erce') gion_Relativ ce')	ve'],errors='coe arvity','Accompany_Cl	÷ 🖵 🖺
Data Transformation	dataset['Clien' dataset['Cred' dataset['Popu' dataset['Loan' # -Filling Missing from mlatend_preprincolumn_namess['1D', #Create dataframes	nt_Income it_Amount lation_Reg Annuity' Values and Coccessing impoorcesive 'Client_Income	'] = pd.to. '] = pd.to. '] = pd.to. = pd.to_r = pd.to_r = rd.to_r = rd	_numeric(danumer	ataset[' .to_nume taset['L	Credit_Am ric(datas pan_Annui	mount'], set['Pop ity'],er	errors='co ulation_Re rors='coer	erce') gion_Relativ ce')	□ ↑ ↓	÷ 🖵 🖺
Data Transformation	dataset['Clien' dataset['Cred' dataset['Popu' dataset['Loan' # -filling Missing from mlatend.prepricolumm_numess['ID',	nt_Income it_Amount lation_Reg Annuity' Values and Coccessing impoorcesive 'Client_Income	'] = pd.to. '] = pd.to. '] = pd.to. = pd.to_r = pd.to_r = rd.to_r = rd	_numeric(danumer	ataset[' .to_nume taset['L	Credit_Am ric(datas pan_Annui	mount'], set['Pop ity'],er	errors='co ulation_Re rors='coer	erce') gion_Relativ ce')	□ ↑ ↓	÷ ₽ i
Data Transformation	dataset['Clien' dataset['Cred' dataset['Popu' dataset['Loan' " - rilling Missing from mixtend.prepricolumm.names=['D') " #Create dostofromes loan_datased.batafi	it_Amount lation_Rep Annuity' Values and O ccessing imposes 'Client_Income] = pd.to.] = pd.to.] = pd.to. = pd.to. = pd.to. = pd.to. = pd.to.	_numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da	ataset[' .to_nume taset['L ,'Active_Lo	Credit_Amric(datas	set['Populity'],eru	errors='co ulation_Re rors='coer	erce') gion_Relativ ce') ce')	□ ↑ ↓	± ♀ ≣ ient','Clie
Data Transformation	dataset['Clien' dataset['Cred' dataset['Popu' dataset['Loan' " - rilling Missing from mixtend.prepricolumm.names=['D') " #Create dostofromes loan_datased.batafi	it_Amount lation_Rep Annuity' Values and Coccessing impenses 'Client_Income 27000.0] = pd.to.] = pd.to. gion_Relat: = pd.to_ peeting data fi t Transactionine', 'Car_Owned Car_Owned 8	_numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da numeric(da	ataset[' .to_nume taset['L ,'Active_Lo	credit_Am ric(datas pan_Annui m','House_On puse_Own_Ch	set['Populativ'],er	errors='co plation_Re prors='coer Count','Credit Credit_Amount 60750.00	gion_Relativ ce') Loan_Annuity Ac 722250	⑥ ↑ ↓ Annuity','Accompany_Cl	± ♀ ∎ ient','Clie
	# -Handling Co	it_Amount lation_Rep _Annuity' _Values and G _Client_Income] = pd.to,] = pd.to, gion_Relat: = pd.to_ recting data fi t Transactioninel, 'Car_Owned Car_Owned 8	_numeric(da numeric(da numeric(da rame Encoder r, Bike_Owned	ataset[' .to_nume taset['L ,'Active_Lo	Credit_Am ric(datas pan_Annui m','House_O	set['Populativ'],erd	errors='co ulation_Re rors='coer Count','Credit	erce') gion_Relativ ce') t_Amount', 'Loan_I	● ↑ ↓ Annuity','Accompany_Cl ccompany_Client Cli	± ♀ ∎ ient','Clie
	# -Handling Co	nt_Income it_Amount lation_Reg Annuity' Values and Co occasing import 'Client_Income 2700000 135000 135000	[] = pd.to. [] = p	_numeric(d ive'] = pd numeric(dan numeric(ataset[' .to_nume taset['L ,'Active_Lo tive_Loan H 1.0 1.0 1.0	Credit_Am ric(datas pan_Annui mn','House_On 1.0 1.0 1.0	set['Populaty'],end	ulation_Re rors='Coer Count','Credit Credit_Amount 60750.00 28440.00 18000.00	gion_Relativ ce') Lean_Annuity A. 722230 1851.30 900.00	Annuity', 'Accompany_Cl ccompany_Cllent Cli 1.0 1.0	± ♀ ∎ ient','Clie
	# -Handling Co	nt_Income it_Amount lation_Rej Annuity' Values and Coccessing impea 'Client_Income 27000.0 13500.0 15750.0	'] = pd.to. '] = pd.to. tion_Relat: = pd.to_i = p	_numeric(ddive'] = pdi numeric(dan numeric	ataset[' .to_nume taset['L ,'Active_Lo tive_Loan H 1.0 1.0 1.0	credit_Am ric(datas pan_Annui un','House_Ou 1.0 1.0 1.0	set['Population ity'],erri	count','Credit Count','Credit Count','Credit Count','Credit Count','Credit Count','Credit Count','Credit Count','Credit	gion_Relativ ce') Loan_Annuity Ar 722250 1851.30 900.00 1746.90	Annuity', 'Accompany_Cl ccompany_Client Clic 1.0 _ 1.0 _ 1.0 _ 1.0 _	± ♀ ∎ ient','Clie
	dataset['Clien' dataset['Cred' dataset['Popu' dataset['Loan' "Filling Missing from alxend_prepricolumn_namess' 'D', "Creet dotaframes loan_dataspd_Ostafrloom_data 1 121966540 2 122071300 3 12131950 4 12214557.0	nt_Income it_Amount lation_Reg Annuity' Values and Co occasing import 'Client_Income 2700000 135000 135000	[] = pd.to. [] = p	_numeric(d ive'] = pd numeric(dan numeric(ataset[' .to_nume taset['L ,'Active_Lo tive_Loan H 1.0 1.0 1.0	Credit_Am ric(datas pan_Annui mn','House_On 1.0 1.0 1.0	set['Populaty'],end	ulation_Re rors='Coer Count','Credit Credit_Amount 60750.00 28440.00 18000.00	gion_Relativ ce') Lean_Annuity A. 722230 1851.30 900.00	@ ↑ ↓ Annuity', 'Accompany_Cl ccompany_Client Cli 1.0 _ 1.0 _ 1.0 _ 1.0 _ 1.0 _	± ♀ ∎ ient','Clie
	dataset['Clien' dataset['Clien' dataset['Popul dataset['Popul dataset['Loan' # -Filling Missing from mixtend_prepriction_namess['10', #Create dataframes loan_data loan_data 1	it_Amount lation_Reg _Annuity' Values and Coccessing impeace **Cilent_Income 27000.0 13500.0 13500.0 13500.0	= pd.to, = pd.to, = pd.to, = pd.to, = pd.to_	_numeric(daive'] = pd numeric(daive'] = pd numeric(daive') = pd numeric(ataset['L .to_nume taset['L ./Active_Lo 1.0 1.0 1.0 1.0 -	Dan_Annui	set['Populativ'],end ity'],end ity'],end iid_Count (credit.Amount 60750.00 28440.00 18900.00 30234.15	erce') gion_Relativ ce') Loan_Annulty Av 7222.50 1851.30 900.00 1746.90 1840.05	Company Client Cli 1.0 -	± ♀ ∎ ient','Clie
	# -Handling Co	nt_Income it_Amount lation_Rej _Annuity' Values and Co cossing imposes c'Client_Income 27000.0 13500.0 13500.0 12150.0	= pd.to, = pd.to, = pd.to, = pd.to, = pd.to_	_numeric(dive'] = pd numeric(danumer	ataset['L'to_nume taset['L','Active_Lo tive_Loan H 1.0 1.0 1.0 1.0 1.0 1.0 1.0	Duse_Own Ch	set['Populativ'],erdity'],erdi	Count', 'Credit	gion_Relativ ce') Loan_Annuity Ac 722250 1851.30 900.00 1746.90 1840.05	CCCOMPANY_CILENT Clin 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 -	± ♀ ∎ ient','Clie
	dataset['Clien' dataset['Clien' dataset['Popul dataset['Popul dataset['Loan' # -Filling Missing from mixtend_prepriction_namess['10', #Create dataframes loan_data loan_data 1	it_Amount lation_Reg _Annuity' Values and Coccessing impeace **Cilent_Income 27000.0 13500.0 13500.0 13500.0	= pd.to, = pd.to, = pd.to, = pd.to, = pd.to_	_numeric(daive'] = pd numeric(daive'] = pd numeric(daive') = pd numeric(ataset['L .to_nume taset['L ./Active_Lo 1.0 1.0 1.0 1.0 -	Dan_Annui	set['Populativ'],end ity'],end ity'],end iid_Count (credit.Amount 60750.00 28440.00 18900.00 30234.15	erce') gion_Relativ ce') Loan_Annulty Av 7222.50 1851.30 900.00 1746.90 1840.05	Company Client Cli 1.0 -	± ♀ ∎ ient','Clie
	# -HandLing Co dataset['Clien dataset['Cred dataset['Popu. dataset['Loan '# -Filling Missing 'from Nakend.prepricolumn_namess['10', '# Create dataspd.DataFr loan_dataspd.DataFr loan_data	nt_Income it_Amount lation_Rej _Annuity' Values and Co cossing impact ('Client_Income 27000.0 13500.0 13500.0 13500.0 12150.0 15750.0	[] = pd.to. [] = p	_numeric(ddive'] = pdive']	ataset['L'	credit_Am ric(datas pan_Annui m','House_On 1.0 1.0 1.0 0.0 0.0 1.0	set['Poppor'], end to the control of	Credit Amount 60750.00 28440.00 18000.00 59301.00 30234.15 78192.00 104256.00	gion_Relativ ce') Loan_Annuity Ar 722250 1851.30 900.00 1746.90 1840.05 2383.65 3388.05 2989.35	Company_Client Clic 1.0 -	± ♀ ∎ ient','Clie
Data Transformation Feature Engineering	# -HandLing Co dataset['Clien' dataset['Popu. dataset['Popu. dataset['Loan, # -filling Missing from mixtend.prepre column_numers=['10', #Create dataframes loan_dataspd.Dutafr loan_data D 0 121320450 1 121966540 2 122017380 3 121311950 4 1224570 — 224017 121364060 224018 121737650	nt_Income it_Amount lation_Rej Annuity' Values and Coccessing import 'Client_Income 27000.0 13500.0 15750.0 12150.0 15750.0 8100.0	= pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ = pd.to_ 	_numeric(daive'] = pd numeric(daive'] = pd numeric(daive') = pd numeric(ataset['L .to_nume taset['L ,'Active_Lo tive_Loan H 10 10 10 10 10 10 10 10	puse_Own_Ch	set['Popole'], end of the control of	Count', Credit	gion_Relativ ce') Lean_Annulty Ar 722230 1851.30 900.00 1746.90 1840.05 2283.65 3388.05	Annuity', 'Accompany_Cl ccompany_Cllent Cli 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 - 1.0 -	± ♀ ∎ ient','Clie