In [1]:
```
#Name:- Kuldeep Ghorpade
#Div:-B
#Roll No.:-09
#ExperimentN No. & Name:- (09). Data munging and Data Preprocessing
```

In [2]:
```
using CSV
using DataFrames
```

In [3]:
```
train=CSV.read("loan.csv",DataFrame,normalizenames=true)
```

Out[3]: 614×13 DataFrame

Row	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	Coa
	String15	String7?	String3?	String3?	String15	String3?	Int64	Floa
1	LP001002	Male	No	0	Graduate	No	5849	
2	LP001003	Male	Yes	1	Graduate	No	4583	
3	LP001005	Male	Yes	0	Graduate	Yes	3000	
4	LP001006	Male	Yes	0	Not Graduate	No	2583	
5	LP001008	Male	No	0	Graduate	No	6000	
6	LP001011	Male	Yes	2	Graduate	Yes	5417	
7	LP001013	Male	Yes	0	Not Graduate	No	2333	
8	LP001014	Male	Yes	3+	Graduate	No	3036	
9	LP001018	Male	Yes	2	Graduate	No	4006	
10	LP001020	Male	Yes	1	Graduate	No	12841	
11	LP001024	Male	Yes	2	Graduate	No	3200	
12	LP001027	Male	Yes	2	Graduate	*missing*	2500	
13	LP001028	Male	Yes	2	Graduate	No	3073	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
603	LP002953	Male	Yes	3+	Graduate	No	5703	
604	LP002958	Male	No	0	Graduate	No	3676	
605	LP002959	Female	Yes	1	Graduate	No	12000	
606	LP002960	Male	Yes	0	Not Graduate	No	2400	
607	LP002961	Male	Yes	1	Graduate	No	3400	
608	LP002964	Male	Yes	2	Not Graduate	No	3987	
609	LP002974	Male	Yes	0	Graduate	*No*	3232	
610	LP002978	Female	No	0	Graduate	No	2900	
611	LP002979	Male	Yes	3+	Graduate	No	4106	
612	LP002983	Male	Yes	1	Graduate	No	8072	
613	LP002984	*Male*	Yes	2	Graduate	No	7583	
614	LP002990	Female	No	0	Graduate	*Yes*	4583	

In [4]:
```
describe(train)
```

Out[4]: 13×7 DataFrame

Row	variable	mean	min	median	max	nmissing	eltype
	Symbol	Union...	Any	Union...	Any	Int64	Type
1	Loan_ID		LP001002		LP002990	0	String15
2	Gender		Female		Male	13	Union{Missing, String7}
3	Married		No		Yes	3	Union{Missing, String3}
4	Dependents		0		3+	15	Union{Missing, String3}
5	Education		Graduate		Not Graduate	0	String15
6	Self_Employed		No		Yes	32	Union{Missing, String3}
7	ApplicantIncome	5403.46	150	3812.5	81000	0	Int64
8	CoapplicantIncome	1621.25	0.0	1188.5	41667.0	0	Float64
9	LoanAmount	146.412	9	128.0	700	22	Union{Missing, Int64}
10	Loan_Amount_Term	342.0	12	360.0	480	14	Union{Missing, Int64}
11	Credit_History	0.842199	0	1.0	1	50	Union{Missing, Int64}
12	Property_Area		Rural		Urban	0	String15
13	Loan_Status		N		Y	0	String1

In [5]:
```
using Pkg
Pkg.add("StatsBase")
```

```
   Updating registry at `C:\Users\kulde\.julia\registries\General.toml`
  Resolving package versions...
   Updating `C:\Users\kulde\.julia\environments\v1.8\Project.toml`
 [2913bbd2] + StatsBase v0.33.21
 No Changes to `C:\Users\kulde\.julia\environments\v1.8\Manifest.toml`
```

In [9]:
```
using StatsBase
train[!,"LoanAmount"]=[item===missing ? floor(mean(skipmissing(train[!,"LoanAmount"
    for item in train[!,"LoanAmount"]]
```

Out[9]:
```
614-element Vector{Real}:
 146.0
 128
  66
 120
 141
 267
  95
 158
 168
 349
  70
 109
 200
   ⋮
 128
 172
 496
 146.0
 173
 157
 108
  71
  40
 253
 187
 133
```

In [11]:
```
train[!,"LoanAmount"]=[item===0.0 ? floor(mean(skipmissing(train[!,"LoanAmount"])))
    for item in train[!,"LoanAmount"]]
```

Out[11]:
```
614-element Vector{Real}:
 146.0
 128
  66
 120
 141
 267
  95
 158
 168
 349
  70
 109
 200
   ⋮
 128
 172
 496
 146.0
 173
 157
 108
  71
  40
 253
 187
 133
```

In [12]:
```
train[!,"Gender"]=[ismissing(item) ? mode(skipmissing(train[!,"Gender"])) : item
 for item in train[!,"Gender"]]
```

Out[12]:
```
614-element Vector{String7}:
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 "Male"
 ⋮
 "Male"
 "Male"
 "Female"
 "Male"
 "Male"
 "Male"
 "Male"
 "Female"
 "Male"
 "Male"
 "Male"
 "Female"
```

In [13]:
```
train[!,"Married"]=[ismissing(item) ? mode(skipmissing(train[!,"Married"])) : item
 for item in train[!,"Married"]]
```

Out[13]:
```
614-element Vector{String3}:
 "No"
 "Yes"
 "Yes"
 "Yes"
 "No"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 ⋮
 "Yes"
 "No"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 "Yes"
 "No"
 "Yes"
 "Yes"
 "Yes"
 "No"
```

In [16]:
```
train[!,"Dependents"]=[ismissing(item) ? mode(skipmissing(train[!,"Dependents"])) :
    for item in train[!,"Dependents"]]
```

Out[16]:
```
614-element Vector{String3}:
 "0"
 "1"
 "0"
 "0"
 "0"
 "2"
 "0"
 "3+"
 "2"
 "1"
 "2"
 "2"
 "2"
 ⋮
 "3+"
 "0"
 "1"
 "0"
 "1"
 "2"
 "0"
 "0"
 "3+"
 "1"
 "2"
 "0"
```

In [17]:
```
train[!,"Self_Employed"]=[ismissing(item) ? mode(skipmissing(train[!,"Self_Employed
    for item in train[!,"Self_Employed"]]
```

Out[17]:
```
614-element Vector{String3}:
 "No"
 "No"
 "Yes"
 "No"
 "No"
 "Yes"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 ⋮
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "No"
 "Yes"
```

In [18]:
```
train[!,"Loan_Amount_Term"]=[ismissing(item) ? mode(skipmissing(train[!,"Loan_Amoun
    for item in train[!,"Loan_Amount_Term"]]
```

Out[18]:
```
614-element Vector{Int64}:
 360
 360
 360
 360
 360
 360
 360
 360
 360
 360
 360
 360
 360
   ⋮
 360
 360
 360
 180
 360
 360
 360
 360
 360
 180
 360
 360
 360
```

In [19]:
```
train[!,"Credit_History"]=[ismissing(item) ? mode(skipmissing(train[!,"Credit_Histo
    for item in train[!,"Credit_History"]]
```

Out[19]:
```
614-element Vector{Int64}:
 1
 1
 1
 1
 1
 1
 1
 0
 1
 1
 1
 1
 1
 ⋮
 1
 1
 1
 1
 1
 1
 1
 1
 1
 1
 1
 0
```

In [20]:
```
Pkg.add("ScikitLearn")
```

```
   Resolving package versions...
   Installed ScikitLearnBase ─ v0.5.0
   Installed ScikitLearn ──── v0.6.5
    Updating `C:\Users\kulde\.julia\environments\v1.8\Project.toml`
  [3646fa90] + ScikitLearn v0.6.5
    Updating `C:\Users\kulde\.julia\environments\v1.8\Manifest.toml`
  [3646fa90] + ScikitLearn v0.6.5
  [6e75b9c4] + ScikitLearnBase v0.5.0
Precompiling project...
  ✓ ScikitLearnBase
  ✓ ScikitLearn
  2 dependencies successfully precompiled in 13 seconds. 215 already precompiled.
```

In [21]:
```
using ScikitLearn
@sk_import preprocessing: LabelEncoder
labelencoder = LabelEncoder()
categories = [2 3 4 5 6 12 13]
```

```
┌ Info: Installing sklearn via the Conda scikit-learn package...
└ @ PyCall C:\Users\kulde\.julia\packages\PyCall\ygXW2\src\PyCall.jl:719
┌ Info: Running `conda install -y scikit-learn` in root environment
└ @ Conda C:\Users\kulde\.julia\packages\Conda\x2UxR\src\Conda.jl:127
```

```
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

## Package Plan ##

  environment location: C:\Users\kulde\.julia\conda\3

  added / updated specs:
    - scikit-learn


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    joblib-1.2.0               |     pyhd8ed1ab_0         205 KB  conda-forge
    scikit-learn-1.1.3         |  py310had3394f_1         7.6 MB  conda-forge
    scipy-1.9.3                |  py310h578b7cb_2        28.2 MB  conda-forge
    threadpoolctl-3.1.0        |     pyh8a188c0_0          18 KB  conda-forge
    ------------------------------------------------------------
                                           Total:        36.0 MB

The following NEW packages will be INSTALLED:

  joblib             conda-forge/noarch::joblib-1.2.0-pyhd8ed1ab_0 None
  scikit-learn       conda-forge/win-64::scikit-learn-1.1.3-py310had3394f_1 None
  scipy              conda-forge/win-64::scipy-1.9.3-py310h578b7cb_2 None
  threadpoolctl      conda-forge/noarch::threadpoolctl-3.1.0-pyh8a188c0_0 None



Downloading and Extracting Packages
scikit-learn-1.1.3   | 7.6 MB    | ########## | 100%
joblib-1.2.0         | 205 KB    | ########## | 100%
threadpoolctl-3.1.0  | 18 KB     | ########## | 100%
scipy-1.9.3          | 28.2 MB   | ########## | 100%
Preparing transaction: ...working... done
Verifying transaction: ...working... done
Executing transaction: ...working... done
Retrieving notices: ...working... done
```

Out[21]:
```
1×7 Matrix{Int64}:
 2  3  4  5  6  12  13
```

In [22]:
```
for col in categories
 train[!,col] = fit_transform!(labelencoder, train[!,col])
end
```

In [23]:
```
train
```

Out[23]: 614×13 DataFrame

Row	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	Coapp
	String15	Int64	Int64	Int64	Int64	Int64	Int64	Float(
1	LP001002	1	0	0	0	0	5849	
2	LP001003	1	1	1	0	0	4583	
3	LP001005	1	1	0	0	1	3000	
4	LP001006	1	1	0	1	0	2583	
5	LP001008	1	0	0	0	0	6000	
6	LP001011	1	1	2	0	1	5417	
7	LP001013	1	1	0	1	0	2333	
8	LP001014	1	1	3	0	0	3036	
9	LP001018	1	1	2	0	0	4006	
10	LP001020	1	1	1	0	0	12841	
11	LP001024	1	1	2	0	0	3200	
12	LP001027	1	1	2	0	0	2500	
13	LP001028	1	1	2	0	0	3073	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
603	LP002953	1	1	3	0	0	5703	
604	LP002958	1	0	0	0	0	3676	
605	LP002959	0	1	1	0	0	12000	
606	LP002960	1	1	0	1	0	2400	
607	LP002961	1	1	1	0	0	3400	
608	LP002964	1	1	2	1	0	3987	
609	LP002974	1	1	0	0	0	3232	
610	LP002978	0	0	0	0	0	2900	
611	LP002979	1	1	3	0	0	4106	
612	LP002983	1	1	1	0	0	8072	
613	LP002984	1	1	2	0	0	7583	
614	LP002990	0	0	0	0	1	4583	

In [24]:
```
using CSV
CSV.write("ProcessedData.csv", train)
```

Out[24]: "ProcessedData.csv"

In [ ]: