|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ColumnName |  |  |  | Status |
| Loan\_ID | LP001002 | LP001003 | LP001005 | Drop |
| CustomerId | 15634602 | 15647311 | 15619304 | Drop |
| Surname | Hargrave | Hill | Onio | Drop |
| Credit\_History | 1 | 1 | 1 | Drop |
| HasCar | 1 | 0 | 1 | Drop |
| NumOfProducts | 1 | 1 | 3 | Drop |
| IsActiveMember | 1 | 1 | 0 | Drop |
| EstimatedSalary | 101348.88 | 112542.58 | 113931.57 | Drop |
| Exited | 1 | 0 | 1 | Drop |
| CreditScore | 619 | 608 | 502 | Keep it |
| Age | 42 | 41 | 42 | Keep it |
| Tenure | 2 | 1 | 8 | Keep it |
| Balance | 0 | 83807.86 | 159660.8 | Keep it |
| HasCrCard | 1 | 0 | 1 | Keep it |
| Dependents | 0 | 1 | 0 | Keep it |
| ApplicantIncome | 5849 | 4583 | 3000 | Keep it |
| CoapplicantIncome | 0 | 1508 | 0 | Keep it |
| LoanAmount |  | 128 | 66 | Keep it |
| Self\_Employed | No | No | Yes | OneHot |
| Geography | France | Spain | France | OneHot |
| Property\_Area | Urban | Rural | Urban | OneHot |
| Married | No | Divorced | Divorced | OneHot |
| Gender | Female | Female | Female | OneHot |
| Designation | Ma nager | Senior Manager | Manager | Ordinal |
| Education | Graduate | Graduate | Graduate | Ordinal |
| Loan\_Status | Y | N | Y | Y Column |

df =  self.data\_TransformationObj.DropColumns(listOfColumns,df)

df=  self.data\_TransformationObj.ReplaceWithCorrectValue(colName='Dependents',valuesToReplace={'3+':3},df=df)

df =  self.data\_TransformationObj.ReplaceWithCorrectValue(colName='Loan\_Status',valuesToReplace={'Y':1,'N':0},df=df)

df =  self.data\_TransformationObj.ReplaceWithCorrectValue(colName='Designation',valuesToReplace={'Assistant Vice President':'AVP','Ma nager': 'Manager'},df=df)

df =  self.data\_TransformationObj.ReplaceWithCorrectValue(colName='Married',valuesToReplace={'No':'Single'},df=df)

columnName = ['Self\_Employed','Married','Dependents']

df = self.data\_TransformationObj.replaceWithMode(columnName,df) *# Replacing the 'Self\_Employed','Married' with mode*

columnName1 = ['LoanAmount']

df = self.data\_TransformationObj.replaceWithMean(columnName1,df)

df = self.data\_TransformationObj.removeOutlier(colName='ApplicantIncome',df=df,LowerQuantile=.20,HigherQuantile=.80)

df = self.data\_TransformationObj.removeOutlier(colName='NumOfProducts',df=df,LowerQuantile=.10,HigherQuantile=.90)

df.to\_csv(self.cleaned\_DataFile\_Path, index=False)  *# Save as CSV without index*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column |  |  |  | Status | Status |
| Route | BLR → DEL | CCU → IXR → BBI → BLR | DEL → LKO → BOM → COK | Drop | Drop |
| Additional\_Info | No info | No info | No info | Drop | Drop |
| Airline | IndiGo | Air India | Jet Airways | Cleaning | OneHot |
| Source | Banglore | Kolkata | Delhi | Cleaning | OneHot |
| Destination | New Delhi | Banglore | Cochin | Cleaning | OneHot |
| Dep\_Time | 22:20 | 05:50 | 09:25 | Cleaning | keep it |
| Arrival\_Time | 01:10 22 Mar | 13:15 | 04:25 10 Jun | Cleaning | keep it |
| Duration | 2h 50m | 7h 25m | 19h | Cleaning | keep it |
| Total\_Stops | non-stop | 2 stops | 2 stops | Cleaning | keep it |
| Date\_of\_Journey | 24/03/2019 | 1/05/2019 | 9/06/2019 | Cleaning | Cleaning |
| Price | 3897 | 7662 | 13882 | Y Column |  |

column\_names, NumberofColumns, airlineName = self.preprocessingObj.valuesFromSchema()

df = self.preprocessingObj.readingDataSet(self.File\_Path)

df = self.preprocessingObj.removeNullValues(df)

df = self.preprocessingObj.removingUnevenValues(df)

df = self.preprocessingObj.removingOutlier(column\_names,airlineName,df)

df = self.preprocessingObj.convertDateInToDayMonthYear(df)

df = self.preprocessingObj.convertHoursAndMinutesToIndependantColumns(df=df,columName="Dep\_Time")

df = self.preprocessingObj.convertHoursAndMinutesToIndependantColumns(df=df,columName="Arrival\_Time")

df = self.preprocessingObj.convertDurationToMunutes(df)

df = self.preprocessingObj.dropUncessaryColumns(df)

self.preprocessingObj.saveDataToFolder(df)