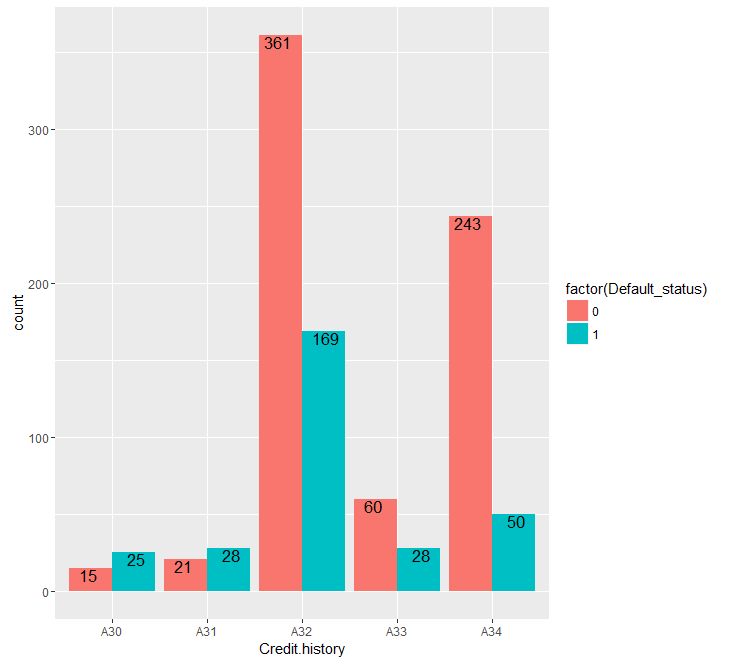
LOGISTIC REGRESSION SUBMISSION

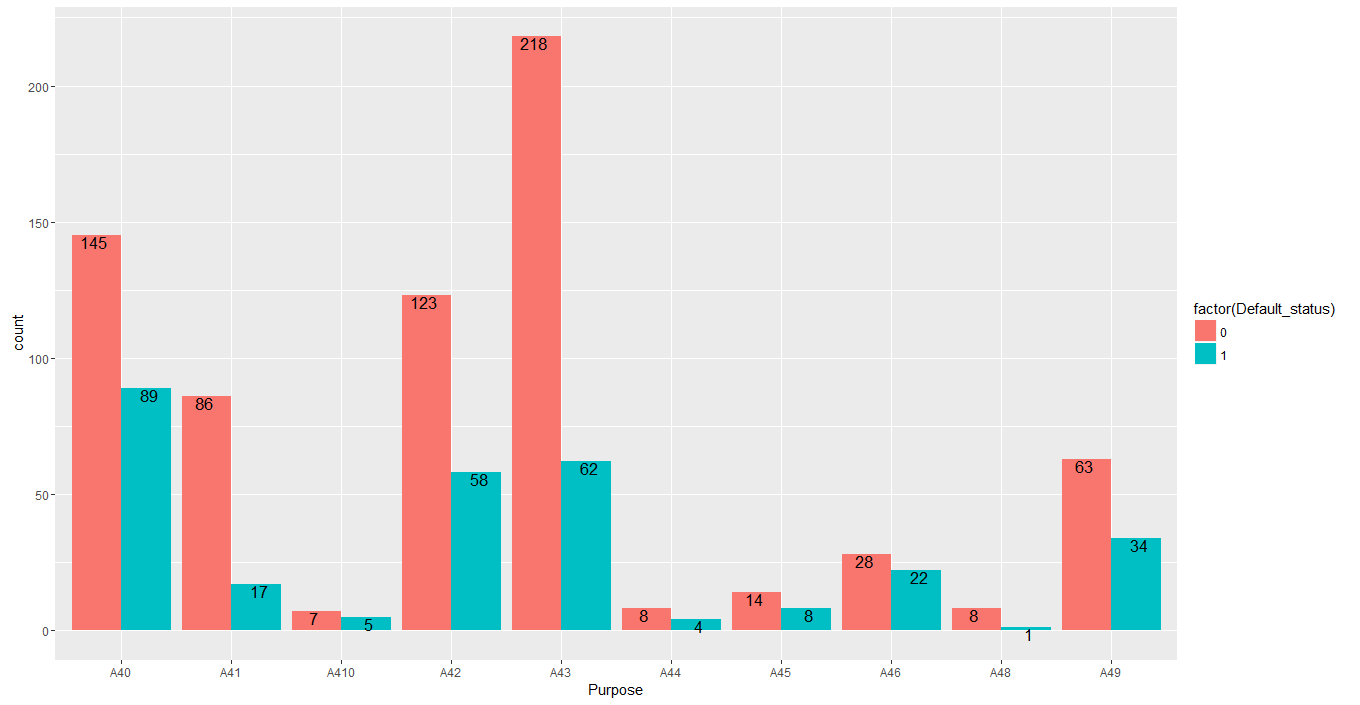
# Checkpoint-1: Data Understanding and Data Exploration

* This plot shows the credit history of the customers, we can see the more number defaulters and

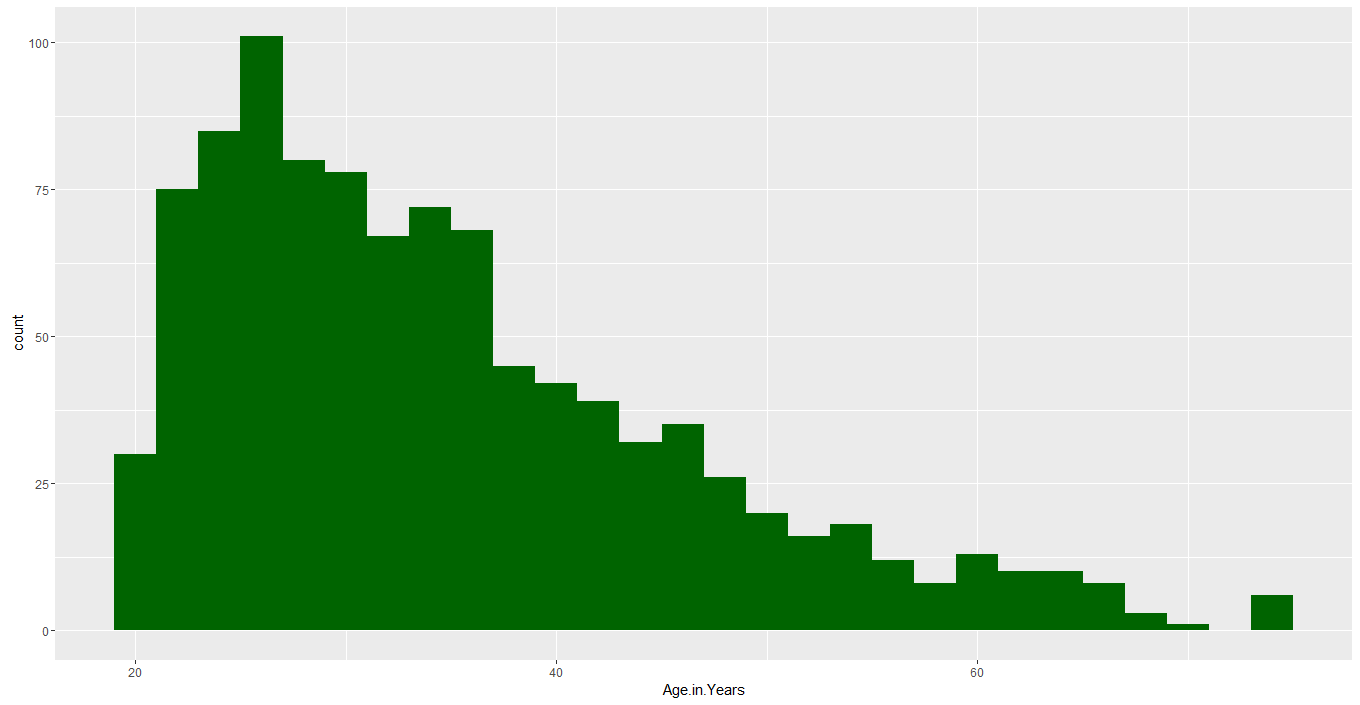
Non-defaulters existing credits paid back duly till now.



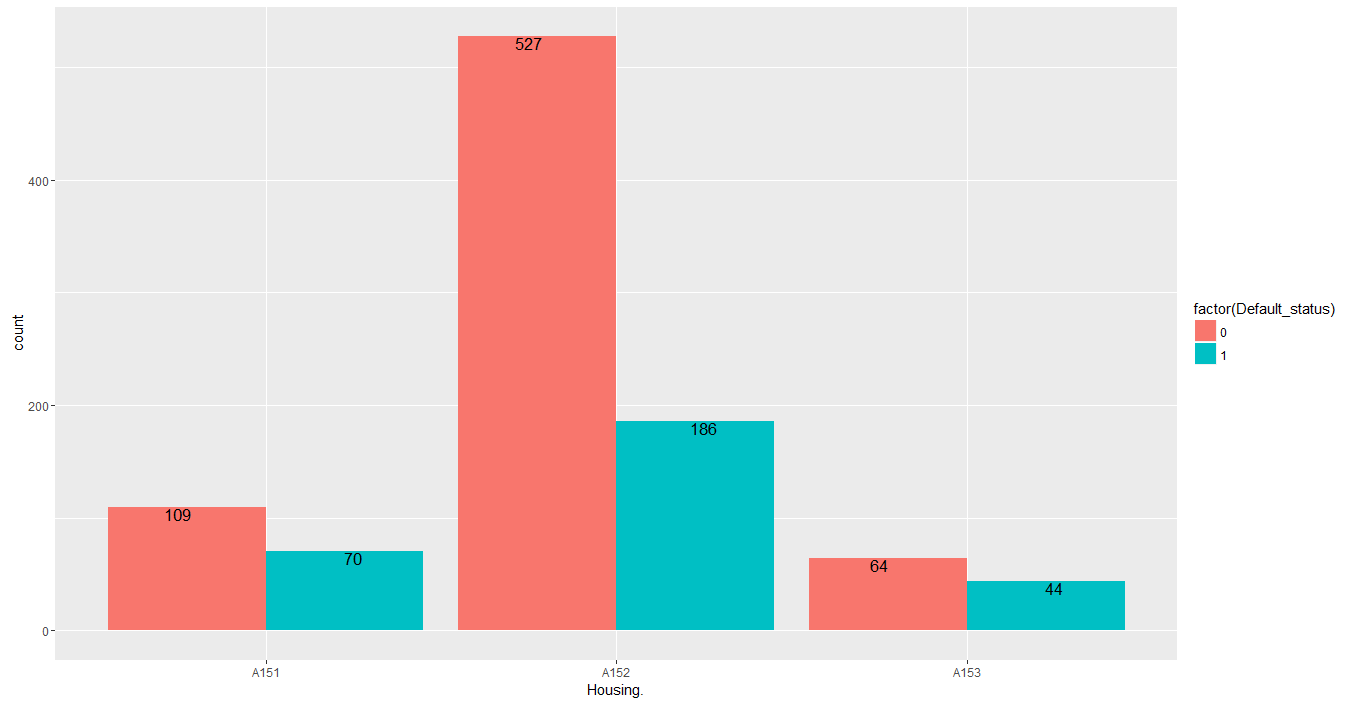
This plot depicts the most number of the credits purpose is radio/television followed by new cars



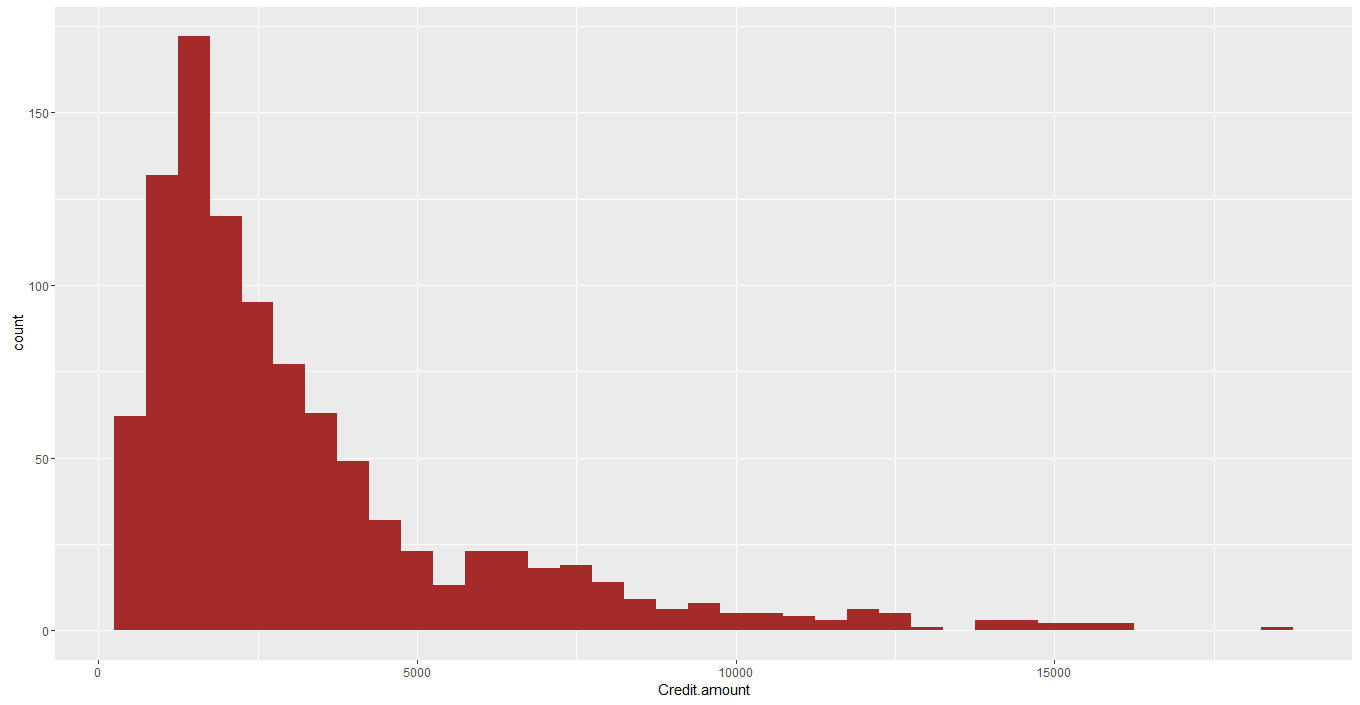
The age in years variable is positively skewed and also has heavy tail. It indicates as the more number of customers age is in between 20 and 60 and very few customers after 60



This plot shows that the more Number of customers are having own house



The credit amount variable is positively skewed and also has heavy tail. It indicates as the more number of customers credit amount is less than 5000



# Checkpoint 2: Data Cleaning and Transformation

* Explain the methodology of Missing value treatment and additionally fill the below table:

|  |  |
| --- | --- |
| **Questions** | **Results(Numeric)** |
| Total number of observations in the dataset | 1000 |
| Total number of variables in the dataset | 21 |
| Total missing values in the dataset | 0 |

* Explain the methodology of Outlier treatment and fill the below table:

I have used the capping method to treat the outliers. Capping method is done by looking at the percentile distribution of the variable and capping the value where there is abrupt change.

* Explain the methodology of how did you created dummy variables

I have used dummy encoding to create dummy variables.n-1 dummy variables are created for a variable having n categories. By using model.matrix function in R created dummy variables

* If binning for numerical variables done explain why it was required?
* I haven’t used any binning technique for numerical variables and used scaling

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Operations performed** | **Variable Name** |
| Outlier treatment | Duration.in.month,Credit.amount,Age.in.Years |
| Dummy creation | Status.of.existing.checking.account, Credit.history, Purpose, Savings.account.bonds, Present.employment.since., Installment.rate.in.percentage.of.disposable.income, Personal.status.and.sex, Other.debtors...guarantors, Present.residence.since, Property, Other.installment.plans,Housing, Number.of.existing.credits.at.this.bank., Job\_status, Number.of.people.being.liable.to.provide.maintenance.for. Telephone.,foreign.worker |
| Binning of variables | NA |

# Checkpoint 3: Splitting the Dataset into train and test

For splitting data set in to train and test,used sample function in R to split the data in to 70 :30,i.e,

Train data is 70 % and test data is 30%

# Checkpoint 4: Modelling

* Explain the methodology of building the model? In the final model, interpret what the coefficients of the variable imply. Check if the coefficients make business sense

I have used the generalised linear regression model for this case study as the business objective is the determine the customer is default or not which is binary outcome.

First step is data preparation where we take care of outliers an missing values

Converted all variables to numeric format by creating dummy variables

Used StepAIC function to perform stepwise variable selection process

Used VIF value to check the multi collinearity and eliminated the variables which are greater than threshold value

After which removed insignificant variables based on P value

Additionally, fill the below table:

|  |  |
| --- | --- |
| **Significant variables in final model (add more rows if requires)** | **Coefficients value (Numeric)** |
| Duration.in.month | 0.535404 |
| Status.of.existing.checking.accountA12 | -0.6711921 |
| Status.of.existing.checking.accountA13 | -1.3173636 |
| Status.of.existing.checking.accountA14 | -1.9827653 |
| Credit.historyA34 | -1.0226318 |
| PurposeA43 | -0.463483 |
| PurposeA43 | 1.1917008 |
| Savings.account.bondsA65 | -0.5815464 |
| Present.employment.since.A74 | -0.5905108 |
| Personal.status.and.sexA93 | -0.5080305 |
| Other.debtors...guarantorsA102 | 1.0221285 |
| Other.debtors...guarantorsA103 | -1.3638155 |
| Present.residence.since2 | 0.4679660 |
| Other.installment.plansA143 | -0.6621720 |

|  |  |
| --- | --- |
| **Final model metrics** | **Values (Numeric)** |
| AIC value | 710.19 |
| Null deviance | 883.54 on 699 degrees of freedom |
| Residual Deviance | 680.19 on 685 degrees of freedom |

# Checkpoint 5: Model Evaluation

* Calculate c-statistic and KS-statistic. What can you tell about the model based on their values?

Additionally, fill the below table:

**Note**: Write the numeric value of c-statistic and KS-statistic after applying your final model to the train dataset and test dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Dataset** | | **Test Dataset** | |
| C-statistic | 8.150786e-01 | C-statistic | 7.495736e-01 |
| KS-statistic | 0.5096 | KS-statistic | 0.4510 |
| Model Evaluation (write Accept or Reject) | | We can accept the model as we got high sensitivity | |

# Checkpoint 6: Threshold value

* Select an appropriate threshold value and calculate the confusion matrix and overall accuracy, sensitivity and specificity
* 0.3 selected as threshold value to minimize false negatives

Additionally, fill the below table:

For Train Data

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 73.43 |
| Sensitivity | 80.26 |
| Specificity | 70.13 |

For Test Data:

|  |  |
| --- | --- |
| **Threshold value** | **Values (Numeric)** |
| Overall Accuracy | 68.33 |
| Sensitivity | 75.00 |
| Specificity | 66.23 |